Améliorer la détection de la fraude à l'assurance automobile en Inde grâce à des modèles d'apprentissage statistique

Mémoire d'actuariat présenté pour l'obtention du

Master droit économie et gestion

Mention actuariat

Et l'admission à **l'Institut des actuaires**

Mémoire présenté par Mme Cécile Rougier

Caractère confidentiel : non

Signature du responsable entreprise :

Président du Jury :

Membres du jury :

Signature du candidat :

Remerciements

Ce mémoire n'aurait jamais pu voir le jour sans le soutien de Fabrice Besnard. J'ai beaucoup appris au sein de vos équipes, autant personnellement que professionnellement. Un grand merci pour votre accueil et votre confiance !

Je voudrais aussi remercier Pierre Martelly pour sa relecture et Paolo Ogno pour son accompagnement.

Merci à mon tuteur, Oliver Desmettre, pour ses relectures attentives et son accompagnement bienveillant.

Une pensée pour l'équipe « Cnam un jour Cnam toujours » : merci pour Boutayna pour les relectures et les soirées actuariat ! @ Aurore et Olga : on attend vos mémoires !

@ Alexandra : merci pour le bureau refuge !

Tendres remerciements à ma famille : le constant soutien de mes parents pour tous mes projets, la gentillesse et la géniale calculatrice de mon frère. Je souhaite aussi remercier mon compagnon, mari, ami, partenaire de mes plus belles aventures.

A mon père, parti trop tôt.

A mes fils, Gaëtan et Aurélien.

Résumé

La fraude met en péril le fonctionnement même de l'assurance en augmentant artificiellement le montant des primes à collecter et les décorrélant du risque réel assuré. La fraude fait l'objet de nombreux dispositifs de lutte et de surveillance au sein des marchés européens de l'assurance automobile. Dans le contexte indien, la priorité a longtemps été donnée au développement du chiffre d'affaires, dans un climat de croissance forte. Et en même temps, la concurrence entre acteurs entraîne aussi une pression pour un règlement rapide des sinistres- ce délai est en effet une information qui a un impact fort sur le marché de l'assurance automobile en Inde.

L'objectif de ce mémoire est d'identifier si des modèles d'apprentissage statistiques peuvent améliorer la détection de la fraude aux sinistres automobiles en Inde. Ils apportent en effet une réponse pertinente en termes de processus : elles permettent une analyse rapide des demandes d'indemnisation, ainsi que l'amélioration de l'efficacité des investigations sans en augmenter drastiquement le coût.

Cependant l'apprentissage statistique peut être rendu difficile par la fréquence relativement faible de la fraude actuellement identifiée. Et l'imagination des fraudeurs complexifie d'autant la détection de la fraude. Malgré le développement rapide des outils d'apprentissage statistique, leur efficacité dans ce contexte reste à valider sur le terrain.

Ce mémoire s'attache dans un premier temps, à décrire ce contexte si particulier qu'est le marché de l'assurance automobile en Inde, en se basant sur des comparaisons avec le marché français des années 2019/2020 pré-Covid ou passé (années 1950-1980). La notion de fraude sera ensuite définie et explicitée. Puis nous expliquerons les fondements des modèles d'apprentissage statistique utilisés, leur mise en œuvre et leurs résultats par rapport aux méthodes actuelles, basées sur l'expertise métier. Cette étude confirme en outre l'intérêt des forêts aléatoires XGBoost dans la détection de la fraude.

Mots clés : fraude, assurance automobile, Inde, apprentissage statistique

Abstract

Fraud jeopardizes the very functioning of insurance by artificially increasing the amount of premiums to be collected and decorrelating them from the actual insured risk. Fraud fighting measures have been widely implemented within the European automobile insurance markets. In the Indian context, priority has long been given to developing turnover, in a climate of robust growth. And at the same time, competition between players also leads to pressure for rapid settlement of claims - this delay is indeed information that has a strong impact on the automobile insurance market in India.

The objective of this thesis is to identify whether Machine Learning models can improve the detection of automobile claims fraud in India. Machine Learning models indeed provide a relevant response in terms of processes: they allow fast analysis of claims, as well as improving the efficiency of investigations without drastically increasing the cost.

However, machine learning is based on data and the low frequency of detected fraud to train the models is critical. Moreover, the imagination of fraudsters makes fraud detection even more complex. Despite the rapid development of statistical learning tools, their effectiveness in this context must be assessed on the ground.

This thesis firstly aims at describing the very specific context of Indian car insurance market, based on comparisons with the French market in the years 2019/2020 pre-Covid or past (1950s-1980s). The concept of fraud will then be defined and explained. Finally, we will explain the foundations of the statistical learning models used, their implementation and their results compared to current methods, based on business expertise. This study further confirms the value of XGBoost random forests in fraud detection.

Keywords: fraud, car insurance, India, Machine learning

Introduction du mémoire

La fraude met en péril le fonctionnement même de l'assurance en augmentant artificiellement le montant des primes à collecter et les décorrélant du risque réel assuré. La fraude est d'autant plus dangereuse pour la viabilité d'un assureur qu'elle est difficilement quantifiable : les données chiffrées dont nous disposons ne correspondent qu'à la fraude détectée, laissant une marge difficile à évaluer avec la réalité du phénomène.

La fraude fait l'objet de nombreux dispositifs de lutte et de surveillance au sein des marchés européens de l'assurance automobile. Dans le contexte indien, la priorité a longtemps été donnée au développement du chiffre d'affaires, dans un climat de croissance forte (croissance de 8.91% entre 2018 et 2019 selon le régulateur local[1]).

La concurrence entre acteurs entraîne aussi une pression pour un règlement rapide des sinistres- ce délai est en effet une information qui a un impact fort sur le marché de l'assurance automobile en Inde. Par exemple, l'IBAI (International Broker Association of India) publie chaque année une étude comparant le taux de règlement des sinistres sous 3 mois des grands acteurs de l'assurance, au même titre que leur ratio de solvabilité.[2] Ainsi IFFCO Tokio, le quatrième assureur privé en Inde[3], propose un paiement des demandes d'indemnisation en moins de 30 minutes sur le compte de l'assuré (dans la limite de Rs 20 000 par véhicule). Tous ces éléments ont été des facteurs limitant le développement de la lutte contre la fraude.

Les modèles de Machine Learning apportent une réponse pertinente en termes de processus : elles permettent une analyse rapide des demandes d'indemnisation, ainsi que l'amélioration de l'efficacité des investigations sans en augmenter drastiquement le coût.

L'objectif de ce mémoire est d'identifier si des modèles de Machine Learning peuvent améliorer la détection de la fraude aux sinistres automobiles en Inde.

Cette tache peut être rendu difficile par la fréquence relativement faible de la fraude actuellement identifiée. Dans le cas de l'assureur que nous allons étudier, la fraude représentait, avant nos travaux, 0,2% des demandes d'indemnisation. Et l'imagination des fraudeurs complexifie d'autant la détection de la fraude. Ainsi l'un des schémas de fraude le plus fréquent en Inde est la fausse déclaration d'accident pour une personne décédée de causes naturelles.[4] Malgré le développement rapide des outils d'apprentissage statistique, leur efficacité dans ce contexte reste à valider sur le terrain.

Dans un premier temps, nous décrirons ce contexte si particulier qu'est le marché de l'assurance automobile en Inde, en se basant sur des comparaisons avec le marché français des années 2019/2020 pré-Covid ou passé (années 1950-1980). La notion de fraude sera ensuite définie et explicitée. Puis nous expliquerons les fondements des modèles d'apprentissage statistique utilisés, leur mise en œuvre et leurs résultats par rapport aux méthodes actuelles, basées sur l'expertise métier.

[1] https://www.forbes.com/advisor/in/car-insurance/motor-insurance-has-a-new-hope-in-digital/(croissance CA en Inde)
[2] https://www.moneycontrol.com/news/business/companies/highest-claims-settlement-record-is-your-insurer-on-this-list-4827211.html
[3] Au premier trimestre 2019, basé sur le montant de primes brut
https://www.crisil.com/mnt/winshare/Ratings/RatingList/RatingDocs/IFFCO_Tokio_General_Insurance_Company_Limited_August_23_2018_RR.html#:~:text=IFFCO%20Tokio%2C%20one%20of%20India's,first%20quarter%20of%20fiscal%202019.
[4] https://economictimes.indiatimes.com/news/economy/finance/how-insurance-firms-are-dealing-with-fraud-claims/articleshow/59448276.cms

Sommaire

Introduction du mémoire 5
1ère partie : L'assurance automobile en Inde, en regard des marchés français pré-covid ou passé..... 9
1- Les clients de l'assurance 9
a. Quelques ordres de grandeur 9
b. Des hommes et des voitures 15
c. Le besoin d'assurance 22
2- Le marché de l'assurance 26
a. Des dynamiques économiques présentant des similarités 26
b. Le développement du marché de l'assurance 30
3- L'assurance automobile 35
a. L'assurance dommage (avant Covid) 35
b. Les principaux acteurs du marché 36
c. Les produits et leurs tarifications 37
d. La distribution 38
e. La sinistralité 39
f. Les frais d'exploitation 40
g. Les résultats financiers 41
h. Synthèse 42
2ème partie : La fraude à l'assurance automobile, en France et en Inde 44
1. Définition et impacts théoriques 44
a. Définition de la fraude en France et en Inde 44
b. Les types de fraudes 46
c. Les impacts de la fraude 47
2. Description du phénomène en France et en Inde 48
a. En France 48
b. En Inde 49
3. La lutte contre la fraude au niveau des institutions 50
a. Cadre légal 50
b. Cadre règlementaire 53
c. Réponses collectives de la profession 54
d. Réponse au sein de chaque assureur 55
e. Les limites de la lutte contre la fraude 56
4. Le dispositif initial de lutte contre la fraude au sein de Generali Inde 57
a. Vue générale du dispositif 57
b. Processus de gestion du sinistre et lutte anti-fraude 57

c. Règles métiers au début de l'étude 58
3ème partie : Présentation des modèles d'apprentissage statistique 60
1. Définition de l'apprentissage statistique 60
a. Historique et définitions 60
b. Enjeux 61
2. Présentation des modèles d'apprentissage 62
a. Introduction 62
b. Les arbres de décision 63
c. Les forêts aléatoires 68
d. Les K-mean clustering (ou méthode des centres mobiles) 71
e. Les classifier naïf bayésien 74
3. Evaluation des résultats des modèles 76
a. Capacité de généralisation d'un modèle, entre biais et variance 76
b. Définir les données utilisées pour évaluer les performances 78
c. Choisir les mesures de performance adaptées pour les modèles supervisés 82
d. Choisir les mesures de performance adaptées pour les modèles non supervisés 90
4. Revue de la littérature scientifique sur l'utilisation de l'apprentissage statistique dans la détection de fraude en assurance automobile 91
4ème partie : Application des modèles d'apprentissage statistiques 93
1. Analyse des données 93
a. Description des données 93
b. Présentation du portefeuille et des demandes d'indemnisation 95
2. Préparation des données 96
a. Eliminer les doublons 96
b. Gérer les valeurs manquantes 97
c. Identifier les valeurs aberrantes 99
d. Normaliser les valeurs 101
e. Analyse des corrélations entre variables et réduction des dimensions 103
f. Présentation de la base de données après retraitement 104
3. Modélisation 105
a. Forêt aléatoire 105
b. Forêt aléatoire XG Boost 106
c. K-mean clustering 106
5ème partie : Analyse des résultats et perspectives 107
1. Le projet au sein de Generali Inde 107
a. Preuve du concept (POC) 107

b. Premier calibrage en production 107
2. Evaluation 108
a. Résultats des Modèles supervisés 108
b. Résultats des Modèles non supervisés 109
c. Facteurs de risques identifiés 109
3. Déploiement 110
a. Détection à la déclaration du sinistre 110
b. Détection lors du traitement du sinistre 110
4. Perspectives et limites 111
a. Critiques et amélioration des modèles 111
b. Perspectives 112
Conclusion 113
Bibliographie 116
Annexes 120

1ère partie : L'assurance automobile en Inde, en regard des marchés français pré-covid ou passé

L'objectif de cette première partie est de rendre sensible les caractéristiques du marché de l'assurance automobile indien, en établissant des parallèles éclairants avec le marché en France actuelle ou passé (années 1950-1980).

1- Les clients de l'assurance

« En automobile comme dans d'autres branches d'assurance, les pratiques n'étaient pas identiques d'un pays à l'autre. Elles variaient avec le niveau de développement du pays, avec le niveau de vie des habitants et l'importance de leur parc automobile et avec leur degré d'avancement technologique »

Jacques-Henri Gougenheim,

Président du Conseil de surveillance d'UAP International de 1987 à 1992, [5]

a. Quelques ordres de grandeur

i. Population

La première dimension à considérer est celle de la population. En 1980, la France compte 55,2 millions d'habitants et en 2019 67,4 millions d'habitants[6]. **L'Inde compte plus de 20 fois plus d'habitants** avec une population estimée à 1 380 millions[7] au 1er janvier 2019[8].

Même si l'Inde a amorcé sa transition démographique depuis 1950, elle n'est pas terminée avec une moyenne de 2.2 enfants par femme[9], alors que la France est déjà à une moyenne de 1,95 enfant par femme au début des années 80[10], et 1,87 en 2018.[11] Donc la population indienne ne connaîtra pas à court terme les problématiques de vieillissement, que rencontrent les pays européens.

[5] « L'évolution du marché de l'assurance automobile en France depuis la guerre et le rôle des mutuelles sans intermédiaires », Entreprises et histoire, 2013/3 (n° 72), p. 91-112. DOI : 10.3917/eh.072.0091. URL : https://www.cairn.info/revue-entreprises-et-histoire-2013-3-page-91.htm page 112
[6] https://data.worldbank.org/indicator/SP.POP.TOTL?locations=FR
[7] https://data.worldbank.org/indicator/SP.POP.TOTL?locations=IN
[8] (Par rapport à la population française de 2019).
[9] https://www.macrotrends.net/countries/IND/india/fertility-rate
[10] https://www.insee.fr/fr/statistiques/1280932
[11] https://www.lemonde.fr/societe/article/2019/01/15/le-nombre-de-naissances-continue-de-baisser-en-france-qui-reste-malgre-tout-le-pays-le-plus-fecond-de-l-ue_5409337_3224.html

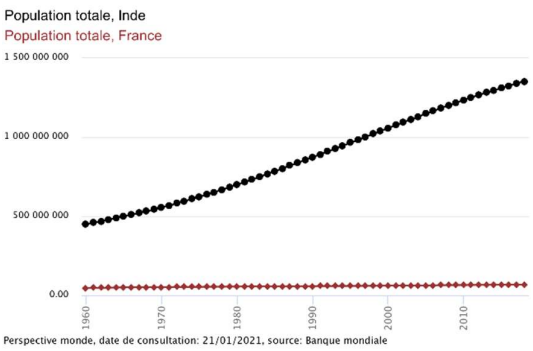


Tableau 1 : Evolution de la population française et indienne de 1960 à 2020

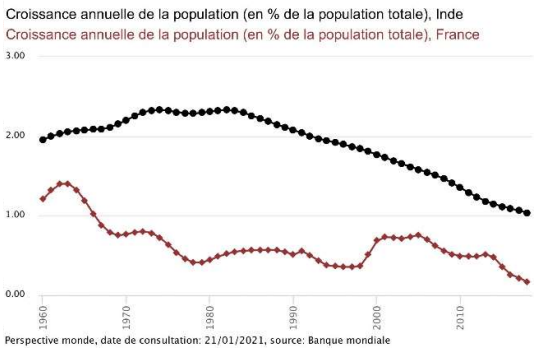


Tableau 2 : Croissance annuelle de la population française et indienne de 1960 à 2020

Au-delà du nombre, la population est donc aussi fondamentalement différente par sa moyenne d'âge : en Inde, 27% de la population avait moins de 14 ans, **67% de la population a entre 15 et 64 ans**, 6% de la population est plus âgée[12]**. En France dans les années 80**, on comptait 30,6% de la population entre 0 et 20 ans, **52,4% de la population entre 20 et 59 ans**, mais déjà 17% plus de 60 ans.[13] Le vieillissement est encore plus net dans les années 2010 en France avec 26% de la population âgée de plus de 60 ans, 24% entre 0 et 19 ans 50% de la population a entre 20 et 59 ans.

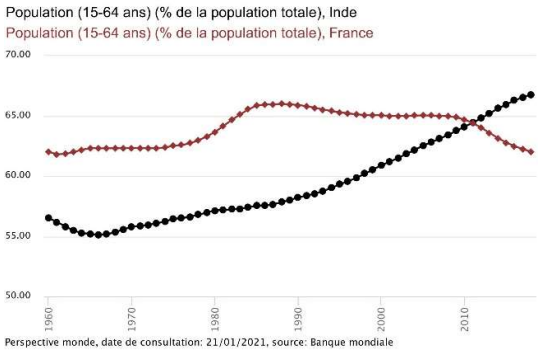


Tableau 3 : Evolution de % des 15-64 ans dans la population totale, en France et en Inde de 1960 à 2020

Les populations française et indienne sont aussi très différentes en termes de géographie : la population rurale en France en 1985 ne représente déjà qu'environ 26% de la population totale[14]. Ce chiffre tombe à 19% en 2019[15]. En 2019, **la population rurale est encore majoritaire en Inde avec 65.5% de la population**[16]

ii. Niveau de développement

Avec une telle différence de population, quel est l'écart en termes de richesse créée ?

Le Produit Intérieur Brut (PIB)[17] **de l'Inde dépasse celui de la France en 2019 :** en Inde il s'élève à 2,87 billions US Dollars[18], contre 2,73 billions US Dollars[19].

[12] https://www.thehindubusinessline.com/news/indias-population-grew-at-12-average-annual-rate-between-2010-and-2019-un/article26803237.ece
[13] https://www.insee.fr/fr/statistiques/1906664?sommaire=1906743
[14] https://donnees.banquemondiale.org/indicateur/SP.RUR.TOTL.ZS?locations=FR
[15] https://donnees.banquemondiale.org/indicateur/SP.RUR.TOTL.ZS?locations=FR
[16] https://tradingeconomics.com/india/rural-population-percent-of-total-population-wb-data.html#:~:text=Rural%20population%20(%25%20of%20total,compiled%20from%20officially%20recognized%20sources.
[17] Le produit intérieur brut est un agrégat de la comptabilité nationale mesurant le résultat final de l'activité de production des unités productrices résidentes sur le territoire national.
(https://www.larousse.fr/encyclopedie/divers/produit_int%C3%A9rieur_brut_PIB/83371)
[18]https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=IN.
[19]https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=IN-FR

Mais si on parle du PIB par habitant, le fossé reste très net : le PIB par habitant en Inde est estimé à 2 101 US Dollar en 2019[20], alors qu'en France, il s'élevait déjà à 12 713 US Dollar en 1980 et atteint 40 579 US Dollar en 2019.

Tableau 4 : Evolution du Produit Intérieur Brut en US Dollar courant, de la France et de l'Inde, de 1960 à nos jours

Tableau 5 : Evolution du Produit Intérieur Brut par habitant en US Dollar courant de la France et de l'Inde, de 1960 à nos jours

L'écart entre la France et l'Inde est particulièrement sensible si on considère les autres **indicateurs utilisés pour calculer l'Indice de Développement Humain** (PIB par habitant, espérance de vie à la naissance et niveau d'éducation des enfants de 17 ans et plus[21]).

- En termes *d'espérance de vie à la naissance*, l'Inde reste nettement en dessous des indicateurs français : en 2017, elle est de 67.8 ans pour un homme et 70.4 ans pour une femme[22]. En 1980, elle était déjà de 70,2 ans pour un homme et de 78,4 ans pour une femme.[23] En 2018, l'espérance de vie française atteint 85,3 ans pour les femmes et de 79,4 ans pour les hommes.[24]

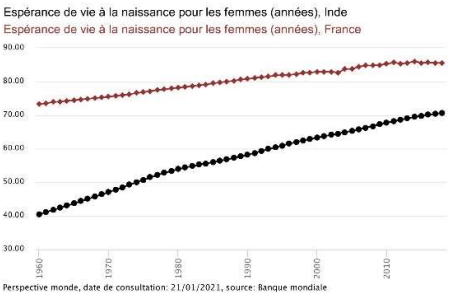


Tableau 6 : Esperance de vie des femmes à la naissance en Inde et en France de 1960 à nos jours

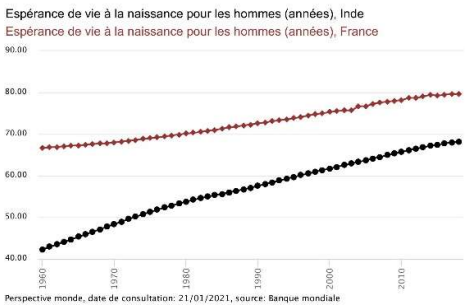


Tableau 7 : Esperance de vie des hommes à la naissance en Inde et en France de 1960 à nos jours

[20] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=IN-FR
[21] https://fr.wikipedia.org/wiki/Indice_de_d%C3%A9veloppement_humain
[22] https://science.thewire.in/health/india-life-expectancy-birth-srs-2013-2017-gender/#:~:text=For%20a%20male%20child%2C%20the,%E2%80%93%20a%20difference%20of%202.9%25.
[23] https://www.insee.fr/fr/statistiques/1288328?sommaire=1288404
[24] https://www.insee.fr/fr/statistiques/3676610?sommaire=3696937

Le nombre de *médecins* pour 1000 habitants est de 0,857 en Inde en 2018[25], alors qu'en France on comptait en 1985 déjà 2,7 médecins pour 1000 habitants et 3,267 en 2018. [26]

- La divergence est aussi très prégnante sur le niveau d'*éducation*. Le taux d'alphabétisation en 2018 en Inde est de 74.37%[27]. En France dans les années 1981/1982, le taux d'analphabète est estimé à 0.90%[28] et le taux d'alphabétisation à 99% en 2019[29]. La durée moyenne de scolarisation en Inde est de 5 ans (versus 8 ans en Chine)[30], alors qu'en France en 1985, l'espérance de scolarisation est de 17 ans[31] . Elle a continué à progresser pour atteindre en 2015, 18,3 ans en France[32].

Même si, en nombre d'étudiants, l'Inde compte plus de trente fois celle de la France en 2013 ! (En 2019 en Inde 37,4 millions d'étudiants en 2019[33], alors que la France des années 80 comptait 1,2 millions d'étudiants en 1980[34] et jusqu'à 2,4 millions d'étudiants en 2013[35])

En synthèse, l'IDH de l'Inde en 2019 est de 0,645 (131ème place du classement), alors que celui de la France est de 0,901 (26ème place du classement) …

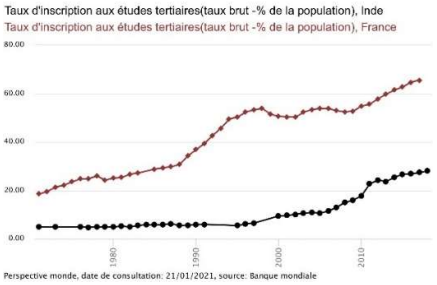


Tableau 8 : Taux d'inscription aux études tertiaires en Inde et en France de 1960 à nos jours

iii. Structure de l'économie

Comment est structurée l'économie ? Quelle est la place des services ?

- **En Inde, l'agriculture emploie encore un peu moins de la moitié de la population active** (42% de l'ensemble des emplois[36]) et un quart du PIB en 2019 (16% du PIB [37]). **En France, elle n'occupait déjà que 8,2% des emplois en 1982**[38] (avec 3,3% de la valeur ajoutée mesurée en % du PIB en 1985) et en 2019 ce chiffre tombe à 2,4% des emplois en 2019[39] (et 1,6 % de la valeur ajoutée du PIB en 2019[40])

[25] https://donnees.banquemondiale.org/indicateur/SH.MED.PHYS.ZS?locations=IN
[26] https://donnees.banquemondiale.org/indicateur/SH.MED.PHYS.ZS?locations=FR
[27]https://perspective.usherbrooke.ca/bilan/servlet/BMTendanceStatPays?codeTheme=4&codeStat=SE.ADT.LITR.ZS&codePays=IND&optionsPeriodes=Aucune&codeTheme2=4&codeStat2=SE.ADT.LITR.ZS&codePays2=IND&optionsDetPeriodes=avecNomP
[28] Dumazedier Joffre, De Gisors Hélène. Français analphabètes ou illettrés ? In : *Revue française de pédagogie*, volume 69, 1984. pp. 15 DOI : https://doi.org/10.3406/rfp.1984.1556
[29] https://www.lemonde.fr/les-decodeurs/article/2014/09/18/qui-sont-les-illettres-en-france_4490014_4355770.html#:~:text=A%20titre%20de%20comparaison%2C%20seulement,seulement%2010%20%25%20sont%20au%20ch%C3%B4mage.
[30] https://www.ficsum.com/dire-archives/ete-2013-3/societe-le-grand-chantier-de-leducation-en-inde-une-visite-dans-une-ecole-de-new-delhi/#_edn23
[31] Insee Références, édition 2018 - Fiches - L'école et ses sortants, page 71
[32] Insee Références, édition 2018 - Fiches - L'école et ses sortants, page 71
[33] https://www.ibef.org/industry/education-sector-india.aspx
[34] https://publication.enseignementsup-recherche.gouv.fr/eesr/8/EESR8_ES_08-les_evolutions_de_l_enseignement_superieur_depuis_50_ans_croissance_et_diversification.php#ILL_EESR8_ES_08_01
[35] https://publication.enseignementsup-recherche.gouv.fr/eesr/8/EESR8_ES_08-les_evolutions_de_l_enseignement_superieur_depuis_50_ans_croissance_et_diversification.php#ILL_EESR8_ES_08_01
[36] https://www.theglobaleconomy.com/India/Employment_in_agriculture/
[37] https://www.theglobaleconomy.com/India/Share_of_agriculture/
[38] https://www.lemonde.fr/les-decodeurs/article/2018/05/27/1968-2018-chomage-conges-payes-secteurs-d-activite-ce-qui-a-change-dans-le-monde-du-travail_5305298_4355770.htm
[39] https://donnees.banquemondiale.org/indicateur/SL.AGR.EMPL.ZS?locations=FR
[40] https://donnees.banquemondiale.org/indicateur/NV.AGR.TOTL.ZS?locations=FR

- L'**industrie** en Inde représente un peu plus **d'un quart des emplois** (26.18% des emplois en 2020[41]) et un peu moins en termes de PIB (23.52% du PIB[42]). Elle occupe une place légèrement plus importante en termes d'emploi (+ 1.64 points en 6 ans[43]) mais **sa contribution au PIB est décroissante** (28,4% du PIB en 2013[44]) **et inférieure à la moyenne des 170 pays** analysés par la banque mondiale (25,06% du PIB en moyenne[45]). Cette tendance pourrait s'expliquer par la structure des entreprises industrielles en Inde, dominée par d'un coté de très grandes entreprises ou très petites- ces dernières ayant peu d'accès au capital et un niveau de compétence faible[46]. Une analyse des résultats de 1 000 fabricants en 2018 montre des résultats économiques inférieurs au coût du capital cette même année[47].

La France commence, elle, à connaître la **désindustrialisation au début des années 80** : en 1984, l'industrie compte pour 26% du PIB[48] et 25,9% des emplois en 1982[49]. La construction en 1982 fournit 8,3% des emplois en France.[50] 35 ans plus tard, l'Industrie représente 17% du PIB et 20% des emplois.

- Quant **aux services**, ils représentent en Inde 32% des emplois[51], mais **50%[52] du PIB, avec une nette croissance.**

Le secteur des services est déjà particulièrement **prépondérant en France dans les années 80 avec 57,6% des emplois en 1982**[53] et 45,7% du PIB qui proviennent des services marchands en 1980, 19,5% des services non marchands[54]. Le déclin de l'agriculture et l'industrie au profit des services a commencé dès 1968[55]. En 2019, en France, les services représentent même 56% du PIB pour les services marchands, 23% pour les services non marchands en 2016[56].

iv. Inégalités et niveau de vie par habitant

Pour mieux illustrer **l'ampleur des inégalités au sein de la population indienne**, nous nous proposons d'étudier 3 indicateurs.

- Le premier est le *taux de pauvreté*, défini par l'OCDE comme la proportion de personnes dont le revenu est inférieur au seuil de pauvreté.

En Inde, de nombreux débats existent sur la définition de la notion de pauvreté[57], avec de fortes variations quant à la dimension de la population vivant sous le seuil de pauvreté.

[41] https://tradingeconomics.com/india/employment-in-industry-percent-of-total-employment-wb-data.html
[42] https://www.theglobaleconomy.com/India/share_of_industry/
[43] https://www.indexmundi.com/facts/india/indicator/SL.IND.EMPL.ZS
[44] https://www.theglobaleconomy.com/India/share_of_industry/
[45] https://www.theglobaleconomy.com/India/share_of_industry/
[46] https://www.financialexpress.com/opinion/employment-growth-indias-jobs-paradigm-needs-to-shift/2495562/
[47] https://www.mckinsey.com/industries/advanced-electronics/our-insights/a-new-growth-formula-for-manufacturing-in-india
[48] https://donnees.banquemondiale.org/indicateur/NV.IND.TOTL.ZS?locations=FR
[49] https://www.lemonde.fr/les-decodeurs/article/2018/05/27/1968-2018-chomage-conges-payes-secteurs-d-activite-ce-qui-a-change-dans-le-monde-du-travail_5305298_4355770.html
[50] https://www.lemonde.fr/les-decodeurs/article/2018/05/27/1968-2018-chomage-conges-payes-secteurs-d-activite-ce-qui-a-change-dans-le-monde-du-travail_5305298_4355770.html
[51] https://www.statista.com/statistics/271320/distribution-of-the-workforce-across-economic-sectors-in-india/
[52] https://www.statista.com/statistics/271329/distribution-of-gross-domestic-product-gdp-across-economic-sectors-in-india/
[53] https://www.lemonde.fr/les-decodeurs/article/2018/05/27/1968-2018-chomage-conges-payes-secteurs-d-activite-ce-qui-a-change-dans-le-monde-du-travail_5305298_4355770.html
[54] https://www.lafinancepourtous.com/wp-content/uploads/2019/02/fiche_repere_pib.pdf
[55] https://www.lemonde.fr/les-decodeurs/article/2018/05/27/1968-2018-chomage-conges-payes-secteurs-d-activite-ce-qui-a-change-dans-le-monde-du-travail_5305298_4355770.html
[56] https://www.lafinancepourtous.com/html/IMG/pdf/fiches_reperes/Fiche_repere_production_francaise.pdf
[57] POVERTY MEASUREMENT IN INDIA: A STATUS UPDATE, Ministry of Rural Development, Dr. SEEMA GAUR & Dr. N SRINIVASA RAO, September 2020

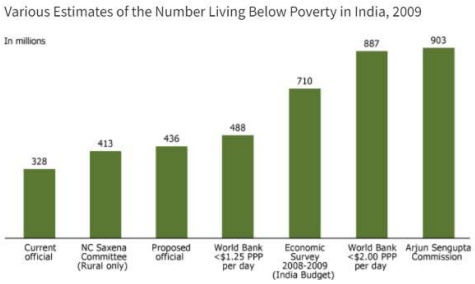


Tableau 9 : Différentes estimations du nombre d'Indiens vivant sous le seuil de pauvreté, en 2019

Si on se base sur les mesures internationales de 1,90 dollars par jour, 13,4% de la population vivait en 2015 sous le seuil de pauvreté[58].

En France avec un seuil de 1000€ par mois, le taux de pauvreté s'élevait à 13,5% en 1984[59] et 13% en 2016[60].

- Le deuxième indicateur des disparités de richesse au sein de la société est la *part des revenus des 10% et 1% les plus riches.*

En 2014, 1% des plus riches indiens possédaient 49% de la richesse du pays[61], alors qu'en France en 1984 cela représentait 7,5% de l'ensemble des revenus du pays[62], et en 2017 11,2 % de l'ensemble des revenus du pays.[63] Toujours en 2014, 10% des Indiens les plus riches possédaient 74% de la richesse du pays[64], alors qu'en France en 1989 cela correspondait à 24,7%.[65] des revenus du pays, et en 2017 de 25,8%.[66]

- Le troisième indicateur des disparités de richesse est le *coefficient de Gini*, qui mesure la dispersion des revenus sur une population[67.]

Si cet indicateur vaut zéro, une personne reçoit tous les revenus de la population, montrant une parfaite inégalité. Généralement il varie entre 24% jusqu'à 63%. En France en 2018 il était de 32,4%, alors qu'en Inde en 2011 il était de 35.7% (versus 38,2% pour la Chine)[68]

Il est donc d'autant plus difficile de décrire un niveau de vie « moyen » par habitant. Une illustration par un exemple concret semble plus pertinente, comme celui de l'équipement des foyers en réfrigérateurs par exemple.

En 2016, 29% de la population indienne possède un réfrigérateur[69]. Si on compare avec la France, le premier équipement en réfrigérateur a eu lieu entre 1950 et 1970 ; en 1983, seuls 4% de la population n'ont pas de réfrigérateur.[70] Les années 80 donnent lieu en France au développement de la multiplicité d'équipement : 4/5ème des ménages possèdent un réfrigérateur, une machine à laver, téléviseur et 2/3 une automobile. Et depuis 2010, le taux de pénétration des réfrigérateurs est considéré de 100%[71]

[58] Global_POVEQ_IND.pdf (worldbank.org)
[59] file:///C:/Users/ceciler/Downloads/HISTO14_h_Flot2_emploi.pdf
[60] https://www.institutmontaigne.org/blog/en-chiffres-pauvrete-dans-le-monde-pauvrete-en-france
[61] https://www.thehindu.com/data/indias-staggering-wealth-gap-in-five-charts/article10935670.ece
[62] https://www.inegalites.fr/Un-siecle-d-inegalites-de-revenus-les-super-riches-regagnent-le-terrain-perdu?id_theme=26#:~:text=En%20Allemagne%20et%20en%20France,de%20la%20crise%20de%202008.
[63] https://www.inegalites.fr/Un-siecle-d-inegalites-de-reventus-les-super-riches-regagnent-le-terrain-perdu?id_theme=26#:~:text=En%20Allemagne%20et%20en%20France,de%20la%20crise%20de%202008.
[64] https://www.thehindu.com/data/indias-staggering-wealth-gap-in-five-charts/article10935670.ece
[65] https://donnees.banquemondiale.org/indicateur/SI.DST.10TH.10?locations=FR
[66] https://donnees.banquemondiale.org/indicateur/SI.DST.10TH.10?locations=FR
[67] Gini Coefficient - Definition, Principles and Limitations (corporatefinanceinstitute.com)
[68] Gini index | Data (worldbank.org)
[69] https://economictimes.indiatimes.com/blogs/et-commentary/the-hidden-truth-behind-indias-low-refrigerator-ownership/
[70] Renaudat Evelyne. La consommation domestique de 1950 à 1980. In : Recherches et Prévisions, n°18-19, Décembre – mars 1989. Activité féminine - Vie familiale. pp. 23-25.
[71] http://www.newsroom-publicismedia.fr/les-taux-dequipement-en-biens-durables-en-france-selon-linsee/

Afin de comprendre l'assurance automobile en Inde, il vaut donc mieux comprendre quelle part de la population a accès à l'automobile et dans quelles conditions.

b. Des hommes et des voitures

« L'activité de l'assurance automobile est conditionnée par intensité de la circulation » [72]

Wolkowitsch Maurice

i. Le parc automobile en France et en Inde

Quel est la taille du parc automobile en France et en Inde ? En se basant sur la définition des automobiles comme des « véhicules terrestres légers, à moteur, constitués d'un châssis généralement sur quatre roues et utilisé principalement pour le transport des personnes »[73], le parc automobile comprend donc les voitures particulières, les véhicules utilitaires légers, poids lourds et autobus[74].
La France comptait 24,1 millions d'automobiles en 1985[75], à 86% constituées de voitures particulières. En 2019, on dénombrait 39,4 millions d'automobiles utilisées en France (versus 52 millions de voitures immatriculées)[76]
Le parc automobile immatriculé en Inde en 2019 correspond à un peu plus de 1,3 fois celui de la France des années 1985 (24,1 millions[77] versus 32,4 millions automobiles immatriculées en Inde en 2019[78]), nettement inférieur au parc des automobiles françaises de 2010.

Si on considère maintenant l'ensemble des véhicules, donc y compris les deux et trois roues, les ordres de grandeur changent considérablement. Le nombre total des véhicules immatriculés en Inde fin mars 2019 était de 296 millions[79].
Les **ventes de véhicules domestiques en Inde en 2018/2019** comptaient 3.38 millions de véhicules passagers, 1.01 million de véhicules commerciaux, 0,7 millions de 3 roues et surtout … **21.2 millions de deux roues**[80].
En comparaison le parc total des deux roues en France en circulation est estimé à 2,9 millions en 2019[81].

Si on rapporte le nombre d'automobile par habitants, l'écart dans l'équipement entre France et Inde est encore plus visible. En France en 1980, on comptait 446 véhicules pour 1000 habitants[82]. En 2017, 478 voitures étaient immatriculées pour 1000 habitants en 2017[83].

[72] Wolkowitsch Maurice. L'assurance en France. In : Annales de Géographie, t. 98, n°550, 1989. pp. 655
[73] Définitions : automobile, auto - Dictionnaire de français Larousse
[74] Données sur le parc automobile français au 1er janvier 2021 | Données et études statistiques (developpement-durable.gouv.fr)
[75] FU4.xls (ccfa.fr)
[76] https://www.lesechos.fr/industrie-services/automobile/les-vrais-chiffres-du-parc-automobile-francais-1168701#:~:text=La%20France%20compte%2052%20millions,avec%20ceux%20%C3%A9manant%20du%20secteur.
[77] https://fr.wikipedia.org/wiki/Parc_automobile_fran%C3%A7ais
[78] http://www.knowindia.net/auto.html
[79] Indian Motor Vehicles | knowIndia.net
[80] Indian Motor Vehicles | knowIndia.net
[81] Le parc deux-roues motorisés des ménages | French road safety observatory (securite-routiere.gouv.fr)
[82] https://www.alternatives-economiques.fr/nombre-de-voitures-de-vehicules-utilitaires-1000-habitants-1er-janvier-0112200974192.html
[83] https://www.bfmtv.com/auto/contrairement-au-reste-des-europeens-les-francais-ont-de-moins-en-moins-de-voitures_AN-201909070025.html

En Inde en 2018, il n'y avait que 22 voitures pour 1000 habitants[84] (versus 3 véhicules en Inde en 1985 et 14 en 2008, toujours pour 1000 habitants)[85].

Comment **évolue** le parc automobile en France et en Inde ? Est-ce qu'un rattrapage dans l'équipement des ménages indiens se dessine ?

Pour le marché français, le boom d'équipement des ménages en voiture a eu lieu dans les années 60 (parc multiplié par 2,2 en dix ans[86]) et la croissance du parc automobile reste élevée jusqu'au milieu des années 70 (50% dans les années 70 puis 20% dans les années 80)[87] Rapportée en nombre de véhicules, la plus forte croissance sur 10 ans représente environ 10 millions de véhicules[88]. **Depuis 1976, le nombre d'immatriculation oscille en moyenne entre « 1,9 et 2,0 millions d'immatriculations chaque année** ».[89] Depuis le milieu des années 70, l'accroissement du parc automobile provient de la multi-motorisation des ménages : en 2011, 30,7% des ménages possèdent deux véhicules, soit deux fois plus qu'en 1980[90].
Pour l'année 2020, on comptait 2,1 millions de voitures particulières immatriculées en France[91].

En 2018, on comptabilisait 3,3 millions de nouvelles immatriculations en Inde[92], ce qui faisait de l'Inde le cinquième marché mondial. **Elle n'est pas suffisante pour atteindre les taux d'équipement des pays matures.** On se souvient notamment que la Chine, avec une population aussi très élevée (1,07 fois la population Indienne[93]) a vu ses ventes atteindre 28,1 millions de véhicules en 2018[94]. La généralisation de la voiture dans les foyers est souvent liée au développement d'une classe moyenne.[95]

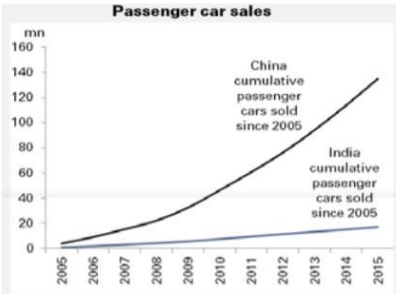


Tableau 10 : Comparaison des ventes cumulées d'automobiles en Inde et en Chine, de 2005 à 2015

De plus, la **croissance du parc indien est très volatile** : les ventes ont par exemple plongé de -3% en 2019, à la suite de la réduction des prêts accordés aux particuliers et à la mise en place de nouvelles normes anti-pollution pour les véhicules. De nombreux paramètres interviennent, des « paramètres parfois contradictoires »[96] rendant ce marché imprévisible

[84] https://auto.economictimes.indiatimes.com/news/passenger-vehicle/cars/india-has-22-cars-per-1000-individuals-amitabh-kant/67059021
[85] https://www.lesechos.fr/industrie-services/automobile/automobile-la-berezina-inattendue-de-leldorado-indien-1134914
[86] https://www.persee.fr/doc/geo_0003-4010_1997_num_106_593_20775
[87] http://isidoredd.documentation.developpement-durable.gouv.fr/documents/CETTEXST005344/29Automobile.pdf
[88] 13, 7 millions en 1970 versus 6,2 millions en 1960. https://www.persee.fr/doc/geo_0003-4010_1997_num_106_593_20775
[89] https://www.persee.fr/doc/rei_0154-3229_1982_num_19_1_2031
[90] https://www.autoplus.fr/societe-economie/depuis-30-ans-les-francais-roulent-moins-mais-ont-plus-de-voitures-252422.html
[91] https://www.insee.fr/fr/statistiques/2012702#tableau-TCRD_075_tab1_departements
[92] https://www.lesechos.fr/industrie-services/automobile/automobile-la-berezina-inattendue-de-leldorado-indien-1134914
[93] 1,43 millions d'habitants en Chine, 1,37 en Inde en 2020 https://lepetitjournal.com/chennai/comprendre-inde/en-2030-la-population-de-linde-depasserait-celle-de-la-chine-264652#:~:text=Aujourd'hui%2C%20la%20Chine%20est,femme%20en%20moyenne%2C%20contre%201
[94] https://www.cnbc.com/2019/01/03/china-annual-auto-sales-fall-for-first-time-in-about-two-decades.html
[95] https://www.goldmansachs.com/insights/pages/macroeconomic-insights-folder/rise-of-the-india-consumer/report.pdf, page 12
[96] https://www.latribune.fr/entreprises-finance/industrie/automobile/l-inde-une-hyper-puissance-automobile-en-sommeil-593295.html#:~:text=Non%2C%20au%20contraire%2C%20une%20v%C3%A9ritable,un%20immense%20potentiel%20de%20croissance.

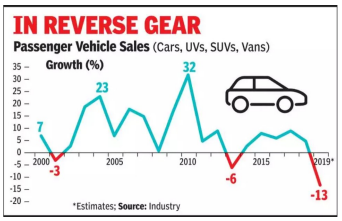


Tableau 11 : Evolution annuelle des ventes de véhicules passagers en Inde entre 2000 et 2019.

Source : La Tribune[97]

ii. Le Budget automobile

Quel budget est alloué à la consommation automobile en France et en Inde ? Nous examinerons dans un premier temps le poids de l'achat d'un véhicule par rapport aux revenus de la population. Puis, nous analyserons les budgets moyens des automobilistes en France et en Inde.

Dans un premier temps, considérons l'investissement initial pour disposer d'un véhicule en France. Selon une étude de « L'argus » publiée en 2018, « le prix moyen dépensé par les Français dépasse les 26.000 euros pour l'achat d'une voiture neuve et les 15.000 euros pour un modèle d'occasion. ».[98] En 1980, la Renault 5 coûtait 25 800 francs, soit avec l'inflation 10 868 euros.[99] Et les **coûts d'acquisition du véhicule sont passés de 44% en 1984 à 34% en 2011 de la dépense automobile [100] grâce à l'essor du recours du marché de l'occasion**.[101]
Ces chiffres ne peuvent se lire sans les rapporter aux revenus moyens. Ainsi en 1981, il fallait 6,3 mois du salaire moyen pour acheter une Renault 5L (3P), alors qu'en 2011, il faut 4,9 mois pour acheter une Renault Clio 1,2 l 75 Alizé (3 P).[102]. « Le prix de la R5 équivalait à un peu plus de 12 Smic mensuels (2 120 francs) en 1980 alors qu'il en faut un peu plus de 10 aujourd'hui (1 498 euros) pour se payer une Clio. »[103] Même calcul pour une 4CV : « S'il faut aujourd'hui l'équivalent de 17 mois de smic pour acheter une voiture à 25.000 euros, en 1953 l'acquisition d'une 4CV, qui valait l'équivalent de 8 300 euros, nécessitait 44 mois de Smic de l'époque. »[104]

Quel est l'investissement alloué à une voiture en Inde ? Les ordres de grandeurs sont là aussi très différents :« **70% des voitures se vendent sous les 8.300 euros** »[105] **La voiture la plus vendue en 2019 est la Maruri Suzuki Dzire[106],** qui coûte entre 5.98 et 7.73 Lakh[107] soit entre 7 075 et 9 146 euros[108] (ou encore 7 996 USD/ 10 337 USD)
Il faut aussi le rapporter au niveau des salaires locaux, et l'exercice est plus délicat, étant donnée l'importance des écarts de niveau de vie au sein de la population (cf. Inégalité et niveau de vie par

[97] https://www.latribune.fr/entreprises-finance/industrie/automobile/l-inde-une-hyper-puissance-automobile-en-sommeil-593295.html#:~:text=Non%2C%20au%20contraire%2C%20une%20v%C3%A9ritable,un%20immense%20potentiel%20de%20croissance.
[98] https://photo.capital.fr/ces-aides-meconnues-pour-acheter-une-voiture-auxquelles-vous-ne-pensez-pas-toujours-36026#7-aides-auxquelles-vous-avez-droit-625130
[99] https://www.bfmtv.com/economie/consommation/non-la-baguette-le-lait-ou-l-essence-ne-coutent-pas-plus-cher-a-un-smicard-qu-en-1980_AV-201811280006.html#:~:text=En%201980%2C%20la%20Renault%205,vite%20que%20l'inflation%20g%C3%A9n%C3%A9rale.
[100] https://metropolitiques.eu/IMG/pdf/met-demoli2.pdf, page 2
[101] « En 1990, le ratio des immatriculations entre les voitures d'occasion et les voitures neuves était de 2,5, contre 4,3 en 2004.
[102] https://abcmoteur.fr/societe/cout-achat-voiture-evolution-prix-en-30-ans/
[103] https://www.bfmtv.com/economie/consommation/non-la-baguette-le-lait-ou-l-essence-ne-coutent-pas-plus-cher-a-un-smicard-qu-en-1980_AV-201811280006.html#:~:text=En%201980%2C%20la%20Renault%205,vite%20que%20l'inflation%20g%C3%A9n%C3%A9rale.
[104] https://www.lefigaro.fr/conso/2017/04/27/20010-20170427ARTFIG00090-une-voiture-neuve-coute-en-moyenne-l-equivalent-de-17-smic.php#:~:text=S'il%20faut%20aujourd'hui,retour%20%C3%A0%20des%20prix%20prohibitifs.
[105] https://www.latribune.fr/entreprises-finance/industrie/automobile/l-inde-une-hyper-puissance-automobile-en-sommeil-593295.html#:~:text=Non%2C%20au%20contraire%2C%20une%20v%C3%A9ritable,un%20immense%20potentiel%20de%20croissance.
[106] https://www.carandbike.com/news/highest-selling-cars-in-2019-2157192#:~:text=The%20Maruti%20Suzuki%20Dzire%20is,success%20in%20the%20B%2B%20segment.
[107] https://www.carwale.com/maruti-suzuki-cars/dzire/
[108] 1 lakh = 100 000 roupies et taux de change du 26/01/2021

habitant). Si on se base sur la répartition des salaires de l'étude de Goldman Sachs[109], les 3 classes les plus représentées sont :

- Les « rural labourers », qui représentent 27% de la population active et qui gagnent en moyenne 810 USD par an
- Les « rural landowners », qui constituent 23% de la population active et qui perçoivent 2,159 USD par an
- Les « urban blue collar/migrant workers" qui rassemblent 19% de la population et gagnent 2,500 USD par an

Une voiture de 8 000 USD représente donc plusieurs années de salaires pour la grande majorité des Indiens : 38 mois de salaire pour un col bleu urbain ou un fermier propriétaire de sa terre. L'achat d'une voiture en Inde reste donc un investissement important, avec des multiples plus comparables à la France de 1953 que de la France de 1980.

Le prix étant le facteur décisif à l'achat, les constructeurs automobiles ont lancé des voitures neuves à partir de 200 000 roupies soit 2 365 euros[110], ce qui représente plus d'une année de salaire pour un « rural landowner » ou « urban blue collar/migrant workers ».

Mais l'achat d'un véhicule ne peut se résumer à la seule contrainte budgétaire, comme l'échec de la Tata Nano a pu le montrer. En effet « la voiture la moins chère du monde »[111], a été jugée « trop chère pour les classes populaires », mais elle était en même temps « trop humble pour les classes moyennes émergentes ».

On remarque que les modèles de Renault, destinés au marché indien, « évoque[nt] immanquablement le mélange d'un véhicule simple comme dans les années 60 avec la modernité du design et des équipements ».[112]

Pour les automobilistes, quel est le budget moyen qui lui est consacré ?

En France, plusieurs sources s'accordent pour montrer une **baisse du budget automobile entre les années 80 et les années 2010**. Les enquêtes « Budget de famille » de 1984 à 2011 évaluent ainsi le budget automobile, à « **16% des dépenses en 1984 » versus « 12% en 2006** »[113]. L'Insee confirme que c'est en 1982, que « la part des dépenses de consommation en automobile dans le budget des ménages est à son apogée à 13,2 % » [114].

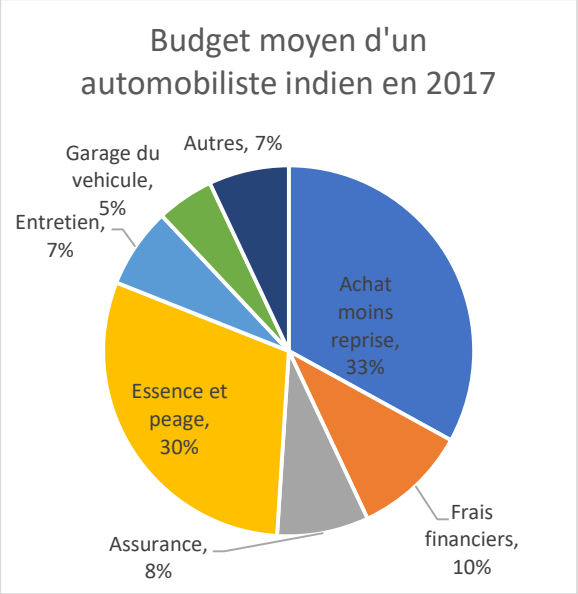
D'après l'ACA (Automobile Club Association, Budget de l'Automobiliste – mars 2018), en 2017, **les automobilistes ont dépensé en moyenne 6 063€ par an pour leur véhicule**.[115]

Une des évolutions les plus marquantes, dans un parc automobile vieillissant est *l'augmentation du prix des réparations* : l'organisme SRA (Sécurité et Réparations automobiles) a en effet mis en lumière une augmentation du prix des pièces de son panier type de 36,1 % en dix ans. Sur la même période, les prix à la consommation établis par l'INSEE ont en effet progressé de « seulement » 13,3 % soit trois fois moins que les prix de pièces de carrosserie.[116]

[109] https://www.goldmansachs.com/insights/pages/macroeconomic-insights-folder/rise-of-the-india-consumer/report.pdf, page 7
[110] https://www.usinenouvelle.com/article/l-inde-nouvelle-frontiere-du-low-cost-pour-l-industrie-automobile.N238907
[111] https://www.lepoint.fr/automobile/tata-nano-la-fin-du-reve-pour-la-voiture-a-1-500-euros-27-01-2019-2289134_646.php
[112] https://www.usinenouvelle.com/article/l-inde-nouvelle-frontiere-du-low-cost-pour-l-industrie-automobile.N238907
[113] https://metropolitiques.eu/IMG/pdf/met-demoli2.pdf, page 2
[114] https://www.insee.fr/fr/statistiques/1281339#consulter
[115] https://www.lelynx.fr/assurance-auto/autour-auto/achat/budget/budget-auto-francais/
[116] https://pro.largus.fr/actualites/en-10-ans-le-prix-des-pieces-de-carrosserie-augmente-de-plus-de-35-5625392.html

Comparons à présent la répartition des dépenses entre la France en 2018 et en Inde en 2017. Les données sont extraites du Budget de l'Automobile club pour la France[117] et le site web Mycarhelpline pour l'Inde[118].

Le premier poste de dépense est l'acquisition du véhicule dans les 2 cas, même si proportionnellement elle occupe une plus grosse part du budget pour l'automobiliste français (43% versus 33%). **L'essence et les péages sont la deuxième charge la plus importante en Inde** (30%), alors que son poids est beaucoup plus limité pour les Français (15%). Le **financement** est aussi beaucoup plus lourd proportionnellement en Inde (10% du budget en Inde versus 5% en France). **L'assurance arrive en 4ème place sur le budget Indien, comme le budget français avec respectivement 8% et 11% du budget**. L'entretien occupe une part beaucoup plus importante pour l'automobiliste français. (17%)



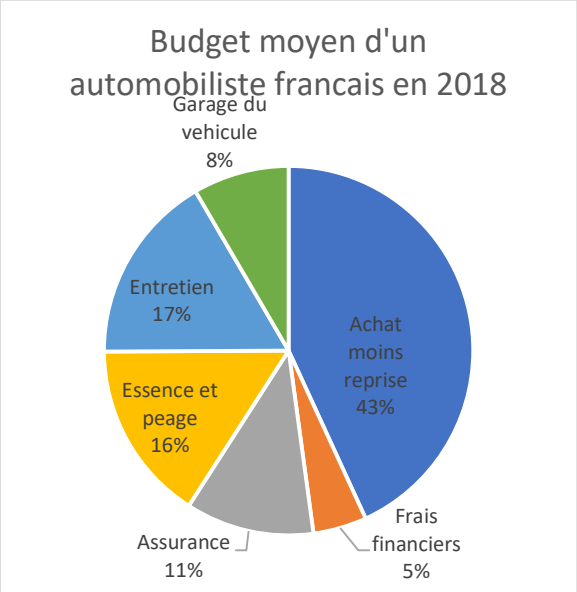


Tableau 12 & 13 : Répartition du budget moyen d'un automobiliste indien et français en 2017.

iii. Sur les routes de France et d'Inde : Portrait-robot des automobilistes et de leurs véhicules en France et en Inde

A quoi ressemblent les véhicules présents sur les routes en France et en Inde ?

La première différence est l'âge des véhicules. En France, le vieillissement du parc automobile s'est accentué entre 1980 et 2020 : l'âge moyen d'un véhicule est de 5,8 ans en 1980[119] et 10,6 ans en 2020[120]. Seuls 28% des voitures particulières en France ont moins de 5 ans en 2021[121]. En Inde, 50.3%

[117] https://www.automobile-club.org/assets/doc/Budget_de_lAutomobiliste_2019.pdf, page 16
[118] https://www.mycarhelpline.com/index.php?option=com_easyblog&view=entry&id=496&Itemid=91
[119] https://www.autoplus.fr/societe-economie/deplacements-menages-1980-2010-ccfa-180081
[120] https://pro.largus.fr/actualites/lage-du-parc-roulant-automobile-atteint-106-ans-en-france-10233401.html
[121] Données sur le parc automobile français au 1er janvier 2021 | Données et études statistiques (developpement-durable.gouv.fr)

des véhicules ont moins de 5 ans et 80% moins de 10 ans[122]. **Le parc automobile indien est donc très récent**. De plus, les conducteurs indiens changent de plus en plus souvent de voitures : « avant c'était entre 7-8 ans et maintenant entre 4-5 ans », selon Umang Kumar, le CEO du portail auto de CarDekho.com[123]

Les marques des véhicules ne se ressemblent pas, avec des marchés automobiles verrouillés par deux groupes dominants dans les deux cas. En France, les trois modèles les plus vendus, que ce soit dans les années 80 ou en 2019 sont françaises. Dans les années 80, on parlait des Renault 5, Peugeot 205, Renault Supercinq[124] et en 2019, Peugeot 208, Renault Clio IV, Citroën C3[125]. Les groupes PSA et Renault détenaient la majorité du marché français avec 57% de part de marché[126], même si les groupes étrangers progressent (3,2% en 2019)

Sur le marché indien, aucun des quatre premiers constructeurs mondiaux ne dispose de plus de 5% du marché[127]. A la fin 2020, 68% du marché était contrôlé par Maruti Suzuki et Hyndai Motors[128]. Cela peut s'expliquer par des taxes sur les véhicules importés particulièrement élevés : 100% de taxes sur les voitures dont la valeur est supérieure à 40 000 dollars et 60% pour celles dont la valeur est inférieure à ce seuil[129].

Qui achète et conduit le véhicule ?

L'âge de l'acheteur d'un véhicule suit des évolutions contraires. En France « en 1981, l'âge moyen d'un acquéreur de véhicule neuf était de 41,7 ans contre 44,7 ans en 1993 **»[130]** Cette tendance se poursuit toujours et l'âge moyen est passé à 51 ans en 2010 et 55 ans en 2018**[131].**

A l'inverse en Inde, les acquéreurs de nouvelles voitures sont de plus en plus jeunes. Selon Rakesh Srivastava, senior vice-président de Hyundai Motors India, **l'âge du premier achat de véhicule est passé « de 30-35 ans à 20-25 ans sur la dernière dizaine d'années »[132]**

Autre différence entre la France et l'Inde : la place des femmes dans l'achat d'un véhicule.

En France, en 2014, « Selon un sondage en ligne mené par le site de petites annonces Vivastreet sur 2199 participants, hommes et femmes de France auraient pratiquement atteint l'égalité concernant le choix d'une nouvelle voiture. »[133]

En Inde, les femmes représentent entre 10% et 12% des achats de voitures[134] Pour Maruti Suzuki, c'est passé de 7% à 12% en 5 ans.

[122] Urban Transport in India: Issues, Challenges, and the Way Forward Sanjay Kumar Singh 1∗ 1 Associate Professor of Economics, Indian Institute of Management Lucknow, Lucknow – 226013, India. page 12 table 7

[123] https://www.livemint.com/Money/G7ugDvBLF2e4rqBvFM1VtN/Car-buyers-in-India-getting-younger-changing-cars-more-freq.html

[124] https://fr.wikipedia.org/wiki/Voitures_les_plus_vendues_en_France

[125] https://www.turbo.fr/photos/top-30-des-voitures-les-plus-vendues-en-france-depuis-le-debut-de-lannee-2019

[126] https://pro.largus.fr/actualites/marche-automobile-une-belle-annee-2019-pour-les-marques-importees-10207399.html

[127] https://www.latribune.fr/entreprises-finance/industrie/automobile/l-inde-une-hyper-puissance-automobile-en-sommeil-593295.html#:~:text=Non%2C%20au%20contraire%2C%20une%20v%C3%A9ritable,un%20immense%20potentiel%20de%20croissance.

[128] https://www.livemint.com/companies/news/maruti-and-hyundai-may-increase-their-market-dominance-amid-current-slowdown-11594035603712.html

[129] Import duty on cars 'outrageous' : Mercedes-Benz - Times of India

[130] https://metropolitiques.eu/Les-jeunes-et-la-voiture-un-desir-contrarie.html

[131] https://www.autoplus.fr/societe-economie/en-france-les-seniors-portent-le-marche-de-la-voiture-neuve-215673

[132] https://economictimes.indiatimes.com/industry/auto/buying-a-car-indians-are-big-on-online-research-find-out-seven-new-habits-of-buyers/articleshow/54873163.cms?from=mdr

[133] https://www.autoplus.fr/societe-economie/egalite-homme-femme-pour-choisir-une-voiture-192273

[134] https://auto.economictimes.indiatimes.com/news/passenger-vehicle/cars/young-affluent-women-too-drive-car-sales-now/74019874

A quoi sert le véhicule ? Combien de kilomètres sont parcourus ?

La tendance en France est à une stabilisation des kilomètres parcourus chaque année en voiture. En 1980, une voiture parcourait en moyenne **12 610km**[135] En 2018, le service des données et études statistique estimait le nombre moyen de kilomètres à 13 117 kilomètres[136], versus 13 356 kilomètres en 1990.

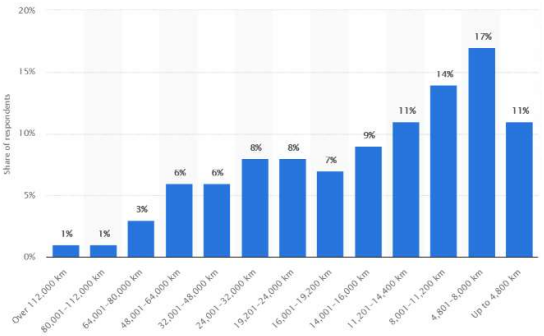


Tableau 14 : Distance parcourues en moyenne par les automobilistes indiens, Source : Statistica

En Inde, de grandes variations existent. Cependant 31% des conducteurs selon cette étude conduirait entre 4 801 et 11 200km, 11% entre 11 201 et 14 400km[137]. Une autre étude conclut que la plupart des **véhicules particuliers sont utilisés en moyenne pour parcourir entre 500 à 700km par mois**[138] (soit entre 6 000 et 8 400 km annuellement).

En France, **la voiture reste « le mode principal de déplacement des actifs pour rejoindre leur lieu de travail** »[139]et le « recours à la voiture augmente avec la distance domicile-travail ». Le recours aux transports en commun pour aller travailler est très spécifique à l'aire d'attraction de Paris.

En Inde, le **développement de l'automobile est lié à la croissance rapide de la population urbaine, qui dépasse de loin les capacités des infrastructures de transport public**[140].

Une tendance commune émerge des 2 marchés (et qui représentent un défi pour les assureurs), c'est la **désaffection des jeunes générations pour l'objet voiture**.

En France, on parle d'un désintérêt de la génération Y (née dans les années 1980 et 1990, appelés aussi milléniale) : selon l'Automobile Club Association, 74% des sondes âgés de 18 à 24 ans « se disent prêts à préférer toute solution alternative à la voiture ».[141]

En Inde, les « *millenials*[142] » sont pointés du doigt pour leurs préférences pour les applications Uber et Ola[143]. Les Gen Z (10/25 ans) et les milléniales sont très intéressés par la location ou des formules d'abonnement[144]. Le débat reste cependant ouvert si cela correspond aux goûts d'une génération ou à une réaction face à des contraintes économiques.

[135] « Les ménages roulent plus, les voitures roulent moins » par Eric Join-Lambert, 1981, p45
[136] https://detours.canal.fr/au-fait-combien-de-kilometres-un-francais-fait-il-tous-les-ans-avec-sa-voiture/
[137] https://www.statista.com/statistics/611800/passenger-passenger-car-annual-mileage/
[138] https://timesofindia.indiatimes.com/business/india-business/indian-commuters-travel-35-km/day-says-survey/articleshow/63140954.cms
[139] https://www.insee.fr/fr/statistiques/5013868#titre-bloc-1
[140] https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=1348&context=jpt
[141]https://books.google.com.hk/books?id=v6lnDQAAQBAJ&pg=PA276&lpg=PA276&dq=BUDGET+AUTOMOBILE+CLUB+1980&source=bl&ots=EieAEdPwre&sig=ACfU3U3ghvW_pr-fjQkBu0Xd-bs2VjmV3w&hl=fr&sa=X&ved=2ahUKEwiHn7LC56_xAhWKMZQKHYFyD_0Q6AF6BAgEEAM#v=onepage&q=BUDGET%20AUTOMOBILE%20CLUB%201980&f=false
[142] Traduction de Milleniales en anglais
[143] https://theprint.in/opinion/newsmaker-of-the-week/dont-hate-millennials-for-not-buying-cars-they-will-be-driving-indias-economy-soon/291351/
[144] https://timesofindia.indiatimes.com/blogs/voices/are-gen-y-and-gen-z-the-leasing-generation (Gen Z)

c. Le besoin d'assurance

i. L'assurance obligatoire

En Inde, la règlementation du trafic autoroutier commence dès 1914 avec le premier « Indian Motor vehicules Act »[145]. Il a fait l'objet de nombreux amendements, en 1920, 1939, 1988, 2017 et 2019.

Mais c'est en 1988 que les articles 146 à 164 du Motor Vehicule Act[146] rendent obligatoire l'assurance au tiers sur les routes indiennes, c'est-à-dire que les dommages causés par le conducteur à un tiers doivent être couverts par une police d'assurance. Les conducteurs de véhicules doivent garder à tout moment les documents attestant de leur assurance.

Plusieurs amendements ont été ajoutés depuis, en 2017 et 2019[147]. Ce dernier[148] augmente les amendes et peines de prisons en cas de conduite sans permis, sans assurance, conduite dangereuse, excès de vitesse ou en état d'ivresse.

En septembre 2018, l'IRDA (Insurance Regulatory and Développement Authority) avait rendu obligatoire, pour tout nouveau véhicule, le paiement lors de l'achat, d'une police de tiers de 3 ans pour les voitures et pour les deux roues d'une police d'assurance de 5 ans[149]. Ce règlement a été retiré le 1er Aout 2020[150].

Malgré cela le défaut d'assurance pour un véhicule est endémique en Inde : en mars 2019, 57% des véhicules sur les routes n'étaient pas assurés selon le rapport annuel de Insurance Information Bureau of India[151] Ils n'étaient que 54% en mars 2018, selon la même source. Cependant cette réalité concerne majoritairement les deux roues (60% de cette catégorie est non assurée), et est plus limitée pour les voitures. **Les voitures non assurées représenteraient que 10 % des voitures sur les routes**. Dans la majorité de ces cas (52%), la police d'assurance n'est pas renouvelée après la première année.

En France, **l'assurance responsabilité civile a été rendue obligatoire par la loi du 27 février 1958[152], soit 30 ans avant l'Inde**. Cependant il resterait **800 000 conducteurs** en 2020 à rouler sans assurance en France, [153] soit 2% environ des véhicules utilisés, un chiffre **en hausse de « 42% depuis 2014** ».

[145] « The first enactment relating to motor vehicles in India was the Indian Motor Vehicles Act, 1914, which was subsequently replaced by the Motor Vehicles Act, 1939" Legislative note, http://164.100.47.193/Refinput/New_Reference_Notes/English/The_motor_vehicles.pdf
[146] https://www.lawyered.in/legal-disrupt/articles/motor-vehicle-act-india-insurance-policiesclaimsaccidental-claims-abhimanyu-batra/
[147] https://www.business-standard.com/about/what-is-motor-vehicles-act
[148] https://www.thehindubusinessline.com/opinion/columns/slate/all-you-wanted-to-know-about-the-new-motor-vehicles-act/article29376425.ece
[149] https://www.livemint.com/insurance/news/why-is-third-party-motor-insurance-cover-mandatory-1561477016906.html
[150] https://www.hindustantimes.com/brand-post/new-vehicle-insurance-rules-cars-and-bikes-are-now-less-expensive/story-MHu3hWsgCZbfUwsOJIhdkO.html
[151] https://indianexpress.com/article/india/over-50-vehicles-on-road-uninsured-report-7103787/
[152] https://www.groupe-sma.fr/SGM/jcms/jizhprod_65176/fr/automobile-et-assurance-obligatoire
[153] https://www.bfmtv.com/auto/qui-sont-ces-800-000-conducteurs-qui-roulent-sans-assurance_AN-202009260026.html

ii. Les principaux risques assurés

1. Les risques d'accident de la route

Cette catégorie est marquée par des montants particulièrement élevés des indemnités versées pour des sinistres corporels tiers, bien qu'elle inclût également d'autres risques comme les dommages aux biens.

Selon les statistiques du World Road Statistics 2018 (sur les données 2016)[154], l'Inde a compté 480 652 accidents, qui ont entraîné 150 785 décès et 494 624 blessés. **Avec 1% des véhicules roulant dans le monde, l'Inde compte pour 11% des morts d'accidents de la route**.[155] C'est la huitième cause de mortalité en Inde. Selon la Banque mondiale, les décès liés aux accidents de la route sont trois fois plus importants dans les pays à faible revenus par rapport aux pays à revenus élevés[156]. Les pauvres, en particulier, ont plus de risque d'être impliqués dans un accident de la route. Car le niveau de revenu a un impact direct sur le mode de transport utilisé. Il existe un surrisque pour les hommes pauvres utilisateurs de la route et au début de leurs années de travail (entre 18-45 ans). Selon le ministère, 76.2% des personnes décédés dans des crashs ont entre 18 et 45 ans. Il existe un acronyme anglais pour désigner cette partie de la population : les VRU, Vulnerable Road Users.

En France, en 1980 on a compté 242 973 accidents corporels de la circulation[157], avec 12 514 décès dans les 6 jours après leur accident. En 1989, on dénombrait 10 528 tués (toujours dans les 6 jours suivant leur accident), alors que le volume de trafic a continué de croître. **En cas d'accident, le risque de décéder est donc sept fois moins élevé en France dans les années 80 qu'en Inde**.

La mortalité sur les routes a continué à décroître pour atteindre en 2018, 3 488 décès[158]. La population en surrisque en France en 2018 est toujours les 18-24 ans (97 tués par million d'habitants), les 75 ans et plus (85 tués par million d'habitants), puis les 25-34 ans (66 tués par million d'habitants). En 2018 on a dénombré 58 353 accidents de la route en France, soit 4 fois moins qu'en 1980 et 8 fois moins qu'en Inde.

Après un accident routier, se pose la question du paiement des soins des éventuels blessés en Inde. 75% des ménages[159] pauvres en Inde connaissent un déclin de leurs revenus : pour les plus pauvres cela représente plus de 7 mois de revenus du ménage ; alors que pour les riches, cela représente moins d'un mois de salaire. En 2018, 70% des frais de santé étaient payés en Inde par des particuliers, **poussant chaque année 7% de la population sous le seuil de pauvreté**.[160] 86% de la population rurale et 82% de la population urbaine n'est couvert par aucune protection financière sur les dépenses de santé.

En **France** c'est le régime de sécurité sociale et / ou l'assurance qui prend en charge les coûts.

[154] https://timesofindia.indiatimes.com/india/this-is-what-makes-indian-roads-deadliest/articleshow/78870630.cms
[155] https://www.thehindubusinessline.com/news/india-accounts-for-11-per-cent-of-global-death-in-road-accidents-world-bank/article33834556.ece
[156] https://www.thehindubusinessline.com/news/india-accounts-for-11-per-cent-of-global-death-in-road-accidents-world-bank/article33834556.ece
[157] https://www.onisr.securite-routiere.gouv.fr/etat-de-l-insecurite-routiere/bilans-annuels-de-la-securite-routiere/bilans-annuels-de-securite-routiere-de-1980-a-1989
[158] https://www.securite-routiere.gouv.fr/actualites/bilan-definitif-de-laccidentalite-routiere-2018
[159] https://www.hindustantimes.com/india-news/burden-of-road-crashes-borne-by-poor-suffer-higher-post-crash-fatality-report-101613222006656.html
[160] https://thewire.in/health/who-is-paying-for-indias-healthcare

2. Les risques d'incendie, de vol, vandalisme

- Le risque de vol

En Inde, le vol de voitures concernerait chaque année environ **2% des voitures assurées**.[161]Les modèles préférés des voleurs sont les Hyundai Creta, Maruti Suzuki NSE Brezza et Mahindra Scorpio. Les voleurs ciblent les SUV, les voitures diesel et automatiques, qui ont de bonnes valeurs de revente[162]. La fréquence des vols est plus importante au nord, qu'au sud et ouest du pays. Hyundai a lancé un système spécifique de localisation, notification et immobilisation du véhicule[163]. Le vol de véhicule constitue le crime le moins résolu par la police en 2018, alors que le nombre de cas a augmenté de 12.9% par rapport à 2017. Seuls 19.6% des cas étaient résolus, selon les données de la police[164]. L'analyse de la police montre que la plupart des véhicules volés étaient garés dans un endroit non gardé et mal éclairé. La capitale New Dehli a le record national de vols par habitant, plus de 5 véhicules par heure ont été volés en 2018.[165]

Cela s'expliquerait par un nombre élevé de véhicules sans espace adapté pour se garer, *l'existence de gangs actifs dans différents états et le business très rentable de revente des voitures volés*[166]. La résolution de ces affaires ne serait pas une priorité pour la police car les propriétaires ne donnent souvent pas suite à leur plainte, étant remboursés par l'assurance.

En France, 140 200 voitures ont été volées en 2019 (et 121 000 en 2020) sur un parc de voitures en circulation de 39,4 millions de véhicules[167], soit environ **0.36% du parc**.

- Le risque d'incendie

En Inde, selon des experts locaux, la plupart des feux sur les véhicules proviennent d'une maintenance inadaptée[168] : le véhicule roule sans liquide de refroidissement ou le garage a remplacé les fils électriques avec du matériel de moins bonne qualité. Les rats peuvent aussi avoir mangé des fils électriques, qui peuvent alors rentrer en contact avec des substances combustibles. Les véhicules commerciaux sont particulièrement affectés, étant fréquemment modifiés pour accepter une surcharge, et entrainant des incendies au niveau des freins dont le refroidissement n'a pas été adapté. Le Motor Vehicule Act exige qu'un extincteur de feu soit installé dans tous les véhicules commerciaux, mais ce n'est pas toujours le cas à cause d'une mauvaise application de la loi[169]

En France, on constate environ **300 000 incendies** de véhicules chaque année selon l'Argus de l'Assurance[170]. D'après la Fédération Française de l'assurance (FFA), les incendies et les vols de véhicules pesaient pour 6,5% des sinistres portés par les assureurs en 2015[171]. Ces incendies peuvent

[161] https://economictimes.indiatimes.com/news/politics-and-nation/grand-theft-auto-car-thieves-love-for-suvs-is-giving-sleepless-nights-to-insurance-cos/articleshow/71120982.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst
[162] https://zascogps.store/grand-theft-auto-car-thieves-love-for-suvs-is-giving-sleepless-nights-to-insurance-cos/
[163] https://www.hyundai.com/content/dam/hyundai/in/en/data/hyundai-story/Technology/blue-link/Blue-Link-Brochure.pdf
[164] https://www.thehindu.com/news/cities/Delhi/police-data-show-motor-vehicle-theft-the-least-solved-crime/article25954331.ece

[165] https://www.financialexpress.com/auto/car-news/more-then-five-vehicles-stolen-every-hour-in-delhi-in-2018/1439451/
[166] https://www.hindustantimes.com/cities/delhi-news/city-of-cars-delhi-reports-maximum-vehicle-thefts-95-stolen-everyday-ncrb-report-101631730601005.html
[167] https://www.lesechos.fr/industrie-services/automobile/les-vrais-chiffres-du-parc-automobile-francais-1168701#:~:text=La%20France%20compte%2052%20millions,avec%20ceux%20%C3%A9manant%20du%20secteur.
[168] https://www.pressreader.com/india/deccan-chronicle/20181213/281930249071080
[169] https://www.deccanchronicle.com/nation/current-affairs/131218/hyderabad-motor-vehicle-act-on-fire-extinguishers-ignored.html
[170] https://www.argusdelassurance.com/acteurs/incendie-et-vehicule-regime-juridique-et-garanties.133754
[171] https://www.argusdelassurance.com/acteurs/incendie-et-vehicule-regime-juridique-et-garanties.133754

avoir de multiples causes : les défaillances mécaniques et électriques sont les causes les plus fréquentes[172], mais aussi le manque d'entretien et de soins, le vandalisme...

- Le risque de vandalisme

En Inde, le vandalisme est défini dans le Code Pénal Indien sous 2 sections :

- La Section 425 caractérise un "mischief"[173]: Quiconque avec intention de causer, ou sachant qu'il est probable que cela cause, des pertes injustifiées ou des dommages à une personne ou au public, entraînant la destruction de n'importe quelle propriété, ou n'importe quel changement dans la propriété ou n'importe quelle situation qui diminue sa valeur ou son utilité, ou affecte son industrie, commet un « mischief »[174]. Cette infraction est punie d'une peine de prison de 3 mois et/ou une amende.

- La Section 403 définit le « criminal misappropriation » : Qui s'approprie de façon malhonnête ou convertit à son usage toute propriété mobile, devrait être puni jusqu'à deux ans d'emprisonnement, ou une amende ou les deux[175].

A noter que certains états comme le Uttar Pradesh et Haryana, ont renforcé leurs législations pour éviter le vandalisme lors de manifestations. Ils permettent notamment la création d'un tribunal spécial dédié, qui peut décider d'amendes jusqu'à deux fois le montant des dommages[176].

En **France**, en 2019, plus d'un million d'automobilistes ont été victime d'un acte de vandalisme selon l'Argus[177].

3. Les risques liés aux catastrophes (y compris tempête)

A cause de son positionnement géographique (zone de convergence intertropicale) et sa géologie (rencontre de la plaque indienne et asiatique), l'Inde est exposé à de nombreuses catastrophes naturelles :

- **les inondations**, le plus fréquemment avec des pluies de mousson frappant la côte Est et remontant le détroit du Gange, mais aussi à l'Ouest (inondations à Mumbai en 2021, 2020). Chaque année les inondations touchent en moyenne 8 millions d'hectares[178] C'est la catastrophe naturelle la plus fréquente en Inde, le deuxième pays le plus exposé au monde après la Chine[179].

- **les sécheresses**. A cause de la nature aléatoire de la mousson, des régions avec des précipitations très limitées peuvent connaître des sécheresses. Sur les vingt dernières années, l'Inde a subi 2 des plus

[172] Qu'est-ce qui peut provoquer un feu de voiture (et comment réagir) ? - La Voix du Nord
[173] https://lawctopus.com/clatalogue/property-offences-under-indian-penal-code/
[174] Whoever with intent to cause, or knowing that he is likely to cause, wrongful loss or damage to the public or any person, causes the destruction of any property, or any such change in any property or the situation thereof as destroys or diminishes its value or utility, or affects it industriously, commits "mischief"

[175] Whoever dishonestly misappropriates or converts to his use any movable property, shall be punished with imprisonment of either description for a term which may extend to two years, or with fine, or with both.

[176] https://www.firstpost.com/india/madhya-pradesh-cabinet-clears-law-to-recover-damages-to-properties-from-rioters-all-you-need-to-know-about-legislation-10218111.html
[177] https://www.largus.fr/actualite-automobile/vols-et-vandalisme-de-voitures-en-2019-des-chiffres-en-legere-baisse-10428405.html
[178] https://www.adrc.asia/countryreport/IND/INDeng98/index.pdf
[179] https://www.atlas-mag.net/en/article/natural-disasters-risk-in-india

sévères sécheresses au monde, en 2002 et 2015, affectant 300 et 330 millions d'indiens respectivement.

- **les cyclones**. Les côtes indiennes s'étendent sur 8 000 km et subissent en moyenne 10% de tous les cyclones tropicaux du monde. Par an, on compte environ 5 à 6 cyclones tropicaux qui touchent les côtes indiennes, dont 2 à 3 peuvent être sévères. Il y a 2 saisons de cyclones en Inde : la période pré-mousson (Mai-Juin) et post-mousson (Octobre-Novembre). Les dégâts matériels restent limités aux côtes, avec un maximum de 100km dans les terres.

- **les tremblement de terre**. Des événements moins fréquents et concentrés dans les Etats moins développés des contreforts de l'Himalaya : seuls 4 tremblements de terre ont dépassé l'échelle 8 de Richter en 53 ans : à Assam en 1987 et 1950, à Kangra en 1905 et à Bihar-Népal en 1935. Cependant, 60% du territoire est considéré comme ayant un risque sismique au moins modéré[180]. Les tremblements de terre peuvent ainsi affecter aussi des états économiquement plus développés (comme le Gujarat, en 2021, 6.9 sur l'échelle de Richter).

- **les glissement de terrain et les avalanches** dans les vallées montagneuses.

Selon un rapport de l'ONU, **les catastrophes naturelles ont coûté à l'Inde 80** milliards de dollars en 20 ans (entre 1998 et 2017)[181] soit en moyenne 4 milliards de dollars par an.

La France est exposée, par ordre de fréquence, à des risques de tempêtes et épisodes de sécheresse et, dans une moindre mesure, submersion marine et tremblements de terre. Ils représentaient déjà une facture estimée entre **1,35 et 1,65 milliard d'euros de sinistres en 2020**[182], avec un épisode de sécheresse chiffré entre 1 et 1,2 milliard d'euros et entre 350 et 450 millions d'euros pour les inondations.

2- Le marché de l'assurance

Pour mieux comprendre les marchés de l'assurance en Inde et en France, nous étudierons d'abord les dynamiques économiques dans lesquelles évoluent les assureurs, puis le niveau de développement du marché de l'assurance, pour ensuite se concentrer sur le marché de l'assurance automobile.

a. Des dynamiques économiques présentant des similarités

Les assureurs vont répondre à un besoin de la part d'une clientèle, mais leur offre sera aussi modelée par l'économie et sa dynamique, dans laquelle l'assurance évolue. Dans ce paragraphe, nous nous attacherons à comparer le cadre économique de la France des années 2019/2020 (pré-covid) ou passé par rapport à celui de l'Inde des années 2019/2020 (pré-covid)

[180] https://www.atlas-mag.net/en/article/natural-disasters-risk-in-india
[181] https://timesofindia.indiatimes.com/india/natural-disasters-cost-india-80-billion-in-20-years-un-report/articleshow/66156074.cms
[182] https://www.bfmtv.com/economie/economie-social/france/catastrophes-naturelles-entre-1-35-et-1-65-milliard-d-euros-de-sinistres-en-france-en-2020_AD-202106160346.html

i. Croissance du PIB

Le premier indicateur à prendre en compte est celui de la croissance du PIB. **Elle est de 2,37% en moyenne dans les années 80 en France** et elle s'établit à 1,72% en 2018. La croissance du PIB en **Inde** est depuis 2003 supérieure à 3% tous les ans, allant jusqu'à **8.5% en 2010.** Il faut remonter aux années 70 en France pour retrouver des chiffres de croissance supérieur à 5%.

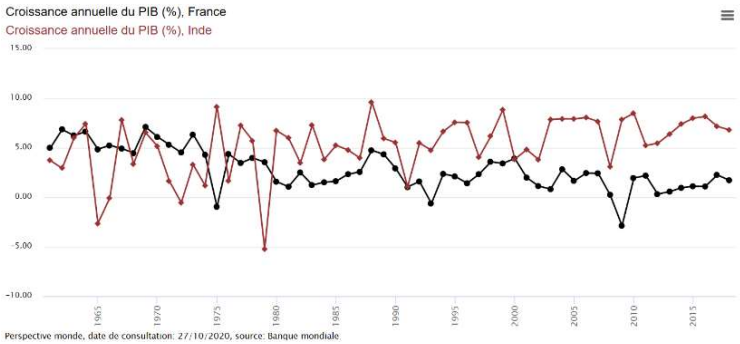


Tableau 15 : Comparaison des croissances annuelles du PIB en France et en Inde de 1960 à 2019

ii. Inflation

L'inflation est souvent un corollaire aux périodes de forte croissance. Et effectivement dans le cas de **l'Inde en 2019, elle a culminé à 12,31% en 2009**. Depuis, elle connaît une décroissance régulière jusqu'à atteindre 4.54% en 2019.[183] **La France des années 80 a connu aussi un pic d'inflation, qui atteint 13,6% en 1980**, mais retombe rapidement dès 1986 en dessous des 4%. A l'inverse, en 2019, l'inflation en France n'a atteint que 1.1%.[184]

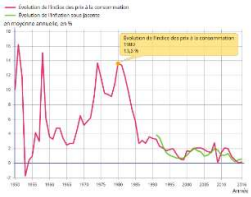


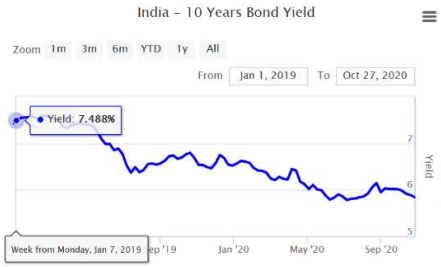
Tableau 16 : Evolution de l'indice des prix à la consommation et l'inflation de 1950 à 2016 en France

Tableau 17 : Evolution du taux d'inflation de 1984 à 2021 en Inde

[183] https://www.statista.com/statistics/271322/inflation-rate-in-india/
[184] https://www.insee.fr/fr/statistiques/4268033

iii. Taux d'intérêt

De la même façon, la situation de l'Inde en 2018 se rapproche plus de l'économie française des années 80 que de celle de 2010 concernant les taux d'intérêt. En effet le taux des obligations d'état à 10 ans pour l'Inde était de 7,5% en janvier 2019[185] et le taux de la banque centrale indienne était de 4.40%, après de nombreuses années au-dessus de 6%[186]



India - Interest Rate Data

	2015	2016	2017	2018	2019
Policy Interest Rate (%)	6.75	6.25	6.00	6.25	4.40

Tableau 18 : Evolution des taux d'intérêt à 10 ans des obligations indiennes au 7 janvier 2019

La France des années 80 connaît des taux d'intérêts particulièrement élevés : le taux moyen des emprunts à 10 ans de l'état français dépasse même plusieurs années les 10% !

Tableau 19 : Evolution des taux d'intérêt à 10 ans des obligations françaises depuis 1970

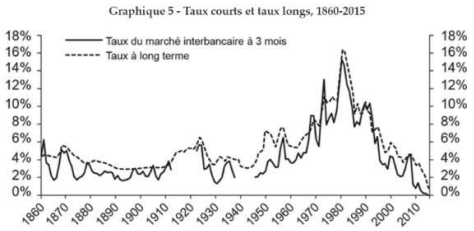


Tableau 20 : Evolution des taux d'intérêt courts et longs en France entre 1860 et 2015

Sources : https://france-inflation.com/taux_10ans.php et https://www.aef.asso.fr/publications/revue-d-economie-financiere/121-les-defis-d-une-economie-a-taux-zero/3392-les-taux-d-interet-en-france-une-perspective-historique

A l'inverse, les taux sont extrêmement réduits en France au début 2019 : 0,6450 pour les OAT 10 ans français fin janvier.[187]

	22/01/2019	23/01/2019	24/01/2019
1 mois	-0,5758	-0,5777	-0,5624
3 mois	-0,5588	-0,5590	-0,5598
6 mois	-0,5633	-0,5590	-0,5594
9 mois	-0,5601	-0,5518	-0,5605
1 an	-0,5450	-0,5393	-0,5394
2 ans	-0,4840	-0,4800	-0,4740
5 ans	-0,0410	-0,0420	-0,0700
10 ans	0,6450	0,6360	0,5890
30 ans	1,6050	1,5810	1,5510

Tableau 21 : Taux des OAT français au 22, 23 et 24 janvier 2019

[185] http://www.worldgovernmentbonds.com/bond-historical-data/india/10-years/
[186] https://www.focus-economics.com/country-indicator/india/interest-rate
[187] https://www.banque-france.fr/statistiques/taux-indicatifs-des-bons-du-tresor-et-oat-28-janv-2019

iv. Solde commercial

En termes de solde commercial, l'Inde et la France des années 2010 partagent un solde négatif. En 2018, la balance commerciale indienne était déficitaire de 106 074 millions de US dollars et la France de 21 484 millions de US dollars.[188] La France des années 80 était légèrement déficitaire, le maximum étant atteint en 1982, avec 14 359 millions de US dollars.

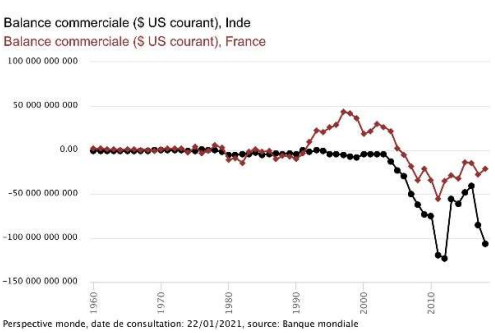


Tableau 22 : Evolution de la balance commerciale en US dollar courant en France et en Inde de 1960 à 2021

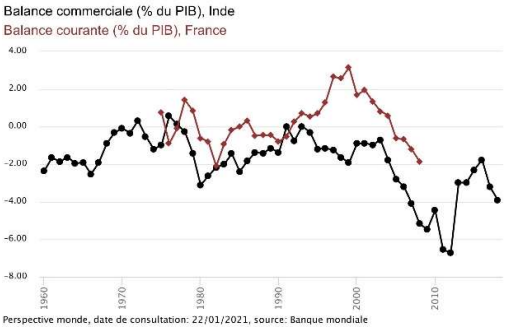


Tableau 23 : Evolution de la balance commerciale en % du PIB en France et en Inde de 1960 à 2021

v. Dette de l'économie

L'endettement est beaucoup plus important au sein de l'économie française des années 2019 qu'en Inde au même moment. En effet, « la dette publique en France a atteint 98,1% du produit intérieur brut (PIB) en fin d'année 2019, soit 2380,1 milliards d'euros. »[189] et « l'endettement du secteur privé non financier a atteint 133,2 % du PIB dans l'Hexagone »[190] à la fin du 1er trimestre 2019.

En Inde **l'endettement du secteur privé et public étaient respectivement de 61% et 69% du PIB** à la fin du 1er trimestre 2020.[191]

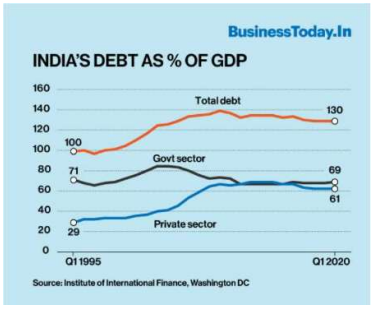


Tableau 24 : Evolution de la dette de l'économie indienne en % du PIB de Q1 1995 à Q1 2020

Pourtant **l'endettement français public brut passé ne représentait que « de 21 à 58,1 points de produit intérieur brut (PIB)** entre 1980 et 1997 »[192].

[188] https://perspective.usherbrooke.ca/bilan/servlet/BMTendanceStatPays?codeTheme=7&codeStat=NE.RSB.GNFS.CD&codePays=IND&optionsPeriodes=Aucune&codeTheme2=7&codeStat2=NE.RSB.GNFS.CD&codePays2=FRA&optionsDetPeriodes=avecNomP

[189] https://www.lefigaro.fr/conjoncture/la-dette-publique-de-la-france-a-atteint-98-1-du-pib-fin-2019-20200325#:~:text=%C3%80%2098%2C1%25%20du%20PIB,tablait%20sur%2098%2C8%25.&text=La%20dette%20publique%20en%20France,a%20rapport%C3%A9%20mercredi%20l'Insee.

[190] https://www.usinenouvelle.com/article/l-endettement-prive-bat-un-record-en-france.N879790

[191] https://www.businesstoday.in/opinion/columns/indian-economy-npas-why-indian-corporates-are-debt-driven-non-performing-assets/story/412124.html

[192] https://www.senat.fr/rap/r98-413/r98-413_mono.html

vi. Taux de chômage

La France des années 80 connaît une forte croissance du chômage, jusqu'à 10,7% de la population active en 1987[193]. Le chômage dépasse les 8 % depuis 2009 en France, atteignant 8,41% de la population active en 2019. **L'Inde connaît depuis 2010 une croissance du chômage**, estimé à 5,33% en 2018.[194]Il existe une forte disparité entre villes et campagnes : il atteint les 9,6% en ville versus 7,8% à la campagne en août 2019[195]

vii. Dépenses des ménages

Concernant les dépenses des ménages, **l'Inde a dépassé en 2019 le niveau de la France** avec 1,671 milliards de US dollars versus 1,594 milliards[196] (et entre 798 et 992 milliards pour la France des années 80) [197]. **Mais en termes de dynamique, l'Inde se rapproche de la France des années 80** avec un taux annuel de croissance des dépenses de consommation finale des ménages supérieur à 4% depuis 2003[198] : en France elle atteint un pic en 1986 à 3,659%, alors qu'elle est en dessous des 3% depuis les années 2000[199]

viii. Revenus fiscaux

Enfin, les revenus fiscaux révèlent un **poids de l'Etat qui reste très différent entre France et Inde** : ils atteignent 12% environ[200] du PIB en Inde (en forte croissance), versus dans la fourchette des 18/20% dans la France des années 80[201] et supérieur à 20% depuis 1996[202]

L'Inde pré-covid partage une même dynamique que de la France des années 80 de part une dynamique de croissance (particulièrement forte pour l'Inde), couplée à une forte inflation et des taux d'intérêt élevés, une dette économique limitée, un chômage qui se développe. **Mais les ordres de grandeurs, notamment en termes de PIB ou de montant de dépenses des ménages se rapprochent de la France des années 2019/2020** pré-covid.

b. Le développement du marché de l'assurance

Dans ce cadre, est ce que l'activité d'assurance a la même dimension dans la France et l'Inde des années pré-covid ? Quel est sa place dans l'activité économique globale du pays ?

i. Dimensions

En 2019, le **chiffre d'affaires des assureurs indiens était de 97 milliards USD**[203]. La même année, le chiffre d'affaires des assureurs sur le marché français s'élevait à 228 milliards d'euros (soit l'équivalent de 255 milliards USD[204]), représentant le 5ème marché mondial de l'assurance[205]. **Le**

[193] https://donnees.banquemondiale.org/indicateur/SL.UEM.TOTL.NE.ZS?locations=FR
[194] https://donnees.banquemondiale.org/indicateur/SL.UEM.TOTL.NE.ZS?locations=IN
[195] https://www.businesstoday.in/sectors/jobs/unemployment-rate-hits-three-year-high-of-84-in-august-cmie/story/377306.html
[196] https://donnees.banquemondiale.org/indicator/NE.CON.PRVT.KD?locations=FR
[197] https://donnees.banquemondiale.org/indicator/NE.CON.PRVT.KD?locations=FR
[198] https://donnees.banquemondiale.org/indicator/NE.CON.PRVT.KD.ZG?locations=IN
[199] https://donnees.banquemondiale.org/indicator/NE.CON.PRVT.KD.ZG?locations=FR
[200] https://donnees.banquemondiale.org/indicateur/GC.TAX.TOTL.GD.ZS?locations=IN
[201] https://donnees.banquemondiale.org/indicateur/GC.TAX.TOTL.GD.ZS?locations=FR
[202] https://donnees.banquemondiale.org/indicateur/GC.TAX.TOTL.GD.ZS?locations=FR
[203] https://www.atlas-mag.net/article/marche-indien-de-l-assurance
[204] Taux de change EUR to DOLLAR en 2019 : 1.1199 USD source : https://www.exchangerates.org.uk/EUR-USD-spot-exchange-rates-history-2019.html
[205] https://www.atlas-mag.net/article/top-10-du-marche-mondial-de-l-assurance-en-2019

marché de l'assurance en France en 2019 pesait donc encore un peu plus de 2,3 fois le marché indien en termes de chiffres d'affaires.

En 1982, le chiffre d'affaires de l'assurance en France s'élevait à 186 milliards de francs et en 1987, 347 milliards de francs.[206] En utilisant le convertisseur de l'Insee[207] et le taux moyen de change en 2019 entre le Dollar et l'Euro[208], le chiffre d'affaires de l'assurance en France respectivement en 1982 et 1987 correspondrait environ à 71,26 milliards USD et à 100,8 milliards USD en 2019. **Le marché de l'assurance en Inde en 2019 se rapproche en termes de dimension à celui de la fin des années 80 en France.**

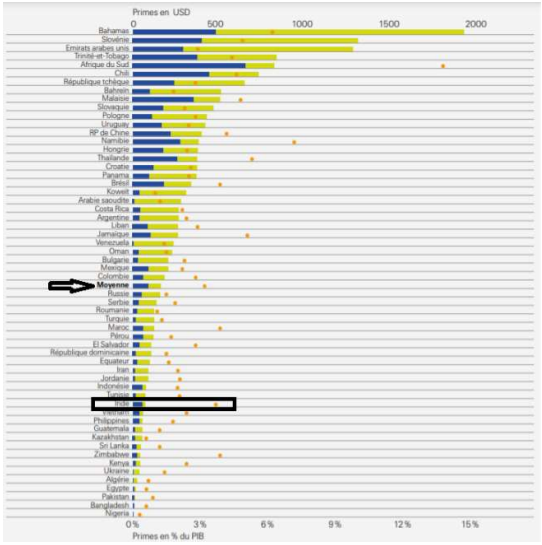


Tableau 25 : Primes en % du PIB au niveau mondial en 2018
Source Sigma 3/2019

Deux autres indicateurs permettent de mieux comprendre le développement des marchés d'assurance : la densité d'assurance (primes par habitant) et la pénétration de l'assurance (part des primes d'assurance dans le PIB) [209].

Dans le cas de l'Inde, ces 2 indicateurs sont relativement bas : la pénétration de l'assurance et la densité sont respectivement à 3.69% et USD 73 pour l'année fiscale 2017-2018 FY2017- 18 (IRDAI, 2019)[210]. En France la pénétration de l'assurance atteignait en 1985, 4,5% [211] et 10,6% en 2018[212]. La densité d'assurance non-vie en France était en 1985, de 496 USD[213], en 2019 de 978 USD[214] versus en Inde 19 USD[215].

ii. Acteurs

En **Inde**, on comptait au 31 Mars 2019**, 70 opérateurs dans le secteur de l'assurance**, qui se répartissent en[216] :

- 27 compagnies d'assurance non-vie
- 24 compagnies d'assurance vie,
- 7 assureurs spécialisés dans la santé (Stand- Alone Health Insurer),
- Un réassureur national (GIC Re),
- 11 filiales de réassurance étranger.

Au niveau du secteur de l'assurance, LIC of India, la seule assurance vie appartenant à l'Etat, compte pour presque 50% de toutes les primes du marché[217].

[206] https://www.senat.fr/rap/r98-045/r98-0454.html
[207] https://www.insee.fr/fr/information/2417794
[208] Taux de change EUR to DOLLAR en 2019 : 1.1199 USD source : https://www.exchangerates.org.uk/EUR-USD-spot-exchange-rates-history-2019.html
[209] https://www.senat.fr/rap/r98-045/r98-045_mono.html#toc8
[210] https://icrier.org/pdf/Working_Paper_394.pdf, page 5
[211] Dickinson Gerard M. Marchés de l'assurance communautaires : Perspective 1992. In : *Revue d'économie financière*, n°4, 1988. Articles divers. pp. 5
[212] https://stats.oecd.org/Index.aspx?QueryId=25444&Lang=fr
[213] Dickinson Gerard M. Marchés de l'assurance communautaires : Perspective 1992. In : *Revue d'économie financière*, n°4, 1988. Articles divers. pp.5
[214] https://www.atlas-mag.net/article/le-marche-francais-de-l-assurance
[215] https://indianexpress.com/article/business/economy/amid-covid-worries-insurance-net-spread-further-both-life-and-health-7691736/
[216] https://www.atlas-mag.net/en/article/indian-insurance-market-structure
[217] https://www.atlas-mag.net/en/article/indian-insurance-market-structure

En **France**, dans les années 80, **le marché de l'assurance est beaucoup plus fragmenté** avec 529 entreprises avec des statuts très différents [218]: 10 sociétés nationalisées en 1986 (les Mutuelles Générales Françaises ont été privatisées le 1er juillet 1987), 173 sociétés anonymes, 180 sociétés mutuelles ou forme mutuelle, 157 succursales de sociétés étrangères et la Caisse Nationale de Prévoyance.

En 1986, les encaissements se répartissent pour **38,7% dans le secteur privé, pour 30% dans le secteur nationalisé, pour 27% dans le secteur mutualiste** et 4,3% pour la Caisse Nationale de Prévoyance. Et le contexte est particulièrement dynamique avec la création de 29 sociétés en 1985, 18 en 1986 et en même temps, 10 disparitions en 1986.

En **2019, l'ACPR recense 695 organismes agréés** ou dispensés d'agrément, dont 349 non-vies, 244 vie et mixtes et 97 mutuelles[219].

iii. Dynamiques communes

L'Inde a connu une croissance régulière **entre 12% et 15% par an sur les 5 dernières années**.[220] Certains analystes affirment que c'est la sous pénétration du secteur de l'assurance en Inde qui en serait le moteur.

En France dans les années 80, l'assurance connaît aussi un fort développement de l'activité, notamment de l'assurance dommage, avec une croissance annuelle moyenne du chiffre d'affaires (primes perçues) de **9,7% en francs courants** et 3,35% en francs 1980 (1,6% en moyenne pour le PIB en francs 1980)[221]. Le secteur connaît depuis une croissance bien plus réduite, avec en 2019 une hausse de **3,8%** du chiffre d'affaires par rapport à 2018.[222]

Même si les dimensions diffèrent, on peut retrouver des thématiques communes au cœur de la croissance connue par l'Inde avant le covid et en France dans les années 80.

1. Une ouverture en cours au secteur privé

En **Inde**, l'assurance vie avait été nationalisée en 1956 et la non-vie en 1972[223]. La **libéralisation de l'assurance a commencé à partir de 1999**[224], sur les recommandations du comité Malhotra. En août 2000, le secteur est ouvert aux opérateurs privés et étrangers. Les compagnies étrangères sont autorisées à détenir jusqu'à 26% d'un assureur. Elle a apporté une amélioration des capacités de distribution et d'efficacité opérationnelle[225],[226]. En 2002, l'autorité de régulation des assureurs permet aux banques de vendre des produits d'assurance vie et non vie[227]. **En 2015, la limite de participation au sein d'un assureur indien pour les groupes étrangers passe de 26% à 49%[228], et en mai 2021, monte jusqu'à 74%.**

---

[218] Wolkowitsch Maurice. L'assurance en France. In : Annales de Géographie, t. 98, n°550, 1989. pp. 647

[219] Les chiffres du marché français de l'assurance, onglet I1

[220] https://home.kpmg/in/en/blogs/home/posts/2022/01/insurance-ecosystems-outlook-growth-mantra.html

[221] https://www.epsilon.insee.fr/jspui/bitstream/1/9852/1/ip244.pdf, page 1

[222] https://www.atlas-mag.net/article/marche-francais-des-assurances-resultats-2019#:~:text=Selon%20la%20F%C3%A9d%C3%A9ration%20Fran%C3%A7aise%20de,milliards%20USD)%20enregistr%C3%A9s%20en%202018.

[223] http://icrier.org/pdf/Liberalizing_Indias_Insurance_Sector_Challenges_andOpportunities.pdf

[224] https://www.oecd.org/finance/insurance/49674941.pdf

[225] https://www.investindia.gov.in/sector/bfsi-insurance

[226] During the first decade after liberalisation, the insurance sector reported an increase in insurance penetration from 2.71 per cent in 2001-02 to 5.20 per cent in 2009-10. Thereafter, the level of insurance penetration declined until 2014-15 due to a decline in life insurance penetration. Source : https://indianexpress.com/article/business/economy/amid-covid-worries-insurance-net-spread-further-both-life-and-health-7691736/

[227] https://icrier.org/pdf/Working_Paper_394.pdf

[228] https://www.business-standard.com/article/economy-policy/dpiit-notifies-74-fdi-cap-in-insurance-sector-under-automatic-route-121061501426_1.html

La **santé financière fragile des assureurs**, en particulier de National Insurance Co[229], Oriental Insurance[230] ou United India Insurance[231] inquiète. Leur fusion annoncée devrait à terme permettre des rationalisations et optimisations de ressources.

En **France**, 34 sociétés avaient été nationalisées par la loi du 25 avril 1946, représentant 50% environ de l'encaissement national en 1945[232]. En 1968, le ministre de l'Economie et des Finances, Michel Debré, regroupe le domaine public de l'assurance en trois groupes : UAP, AGF et GAN.
L'ouverture commence en 1973 par la création d'une holding centrale détenant la totalité des actions des sociétés d'assurance, et en 1977 AGF, GAN et UAP sont introduits en bourse**. La loi du 31 décembre 1987 permet « aux sociétés centrales de devenir des sociétés anonymes à part entière, les actions des sociétés nationales deviennent accessibles à n'importe quel investisseur étranger ou national** : l'interdiction par les sociétés centrales de céder des actions de leur filiales d'assurance est levée [233]»
La privatisation de l'UAP n'aura lieu qu'en mai 1994, celle des AGF au printemps 1996. Et Gan sera racheté par Groupama en juin 1998[234].

2. Une cconcentration des acteurs

Le marché de l'assurance en Inde connaît une phase de consolidation depuis les années 2000, qui s'accélère. On peut citer notamment la fusion de Bharti AXA General Insurance et ICICI lombard en juillet 2020 pour 614,5 millions USD, l'acquisition en février 2019 de 40% de Royal Sundaram General Insurance par le belge Ageas. Gopal Balachandran, CFO[235] de ICICI Lombard General Insurance résume l'évolution de l'assurance en Inde ainsi : « Cela fait deux décennies depuis l'ouverture du secteur de l'assurance. **Alors que dans une première phase, l'accent était mis sur la croissance, l'industrie va se diriger vers une phase de consolidation comme il devient plus mature et les entreprises commencent aussi à chercher la profitabilité et non seulement des chiffres en croissance** »[236]
En assurance dommage en France, en 1987, la première entreprise détient 12,2% de part de marché et les cinq premières 40,65%[237]. L'ouverture au secteur privé a aussi créée une dynamique concurrentielle, une pression sur les prix, qui s'est accompagnée d'une concentration des acteurs depuis les années 1980.

[229] https://timesofindia.indiatimes.com/business/india-business/national-insurance-co-widens-fy20-loss-to-rs-4100-crore/articleshow/78403496.cms
[230] https://www.financialexpress.com/industry/oriental-insurance-requires-rs-3000-crore-for-growth-solvency/1884490/
[231] https://bfsi.economictimes.indiatimes.com/news/insurance/united-india-insurance-solvency-at-0-74-less-than-stipulated-ratio/88434955
[232] *https://www.senat.fr/rap/r98-0452/r98-045210.html*

[233] *https://www.senat.fr/rap/r98-0452/r98-045210.html*
[234] *https://www.senat.fr/rap/r98-0452/r98-045210.html*
[235] Chief Financial Officer
[236] "It has been two decades since the opening up of the insurance industry. While during earlier phases the emphasis was on growth, the industry will move towards consolidation as it matures as companies also start looking at profitability and not only growth numbers," https://economictimes.indiatimes.com/industry/banking/finance/insure/non-life-insurance-to-witness-consolidation-in-2022/articleshow/88643687.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

[237] Dessal René. Panorama de la distribution de l'assurance en Europe. In : Revue d'économie financière, n°11, 1989. L'assurance européenne : la grande transformation. pp. 261-275.
Doi : 10.3406/ecofi.1989.1660
http://www.persee.fr/doc/ecofi_0987-3368_1989_num_11_3_1660, page 269

iv. Des défis et perspectives spécifiques

1. Le marché interieur

En Inde, la croissance de l'assurance est souvent expliquée par le développement rapide d'une **classe moyenne jeune urbaine**[238] et la sous pénétration de l'assurance dans la population. Une part importante de la population n'a **aucun accès aux produits d'assurance**, en particulier les zones rurales[239]. 96% de la population n'a jamais acheté aucune couverture d'assurance[240].

La micro-assurance pourrait être une solution pour adresser cette population. La micro-assurance est définie par l'autorité de régulation des assurances indienne, l'IRDAI, comme une couverture dont le capital assuré en vie, maladie, mortalité du bétail et individuelle accident est compris entre 5 000 INR (70 USD) et 50 000 (700 USD)[241]

La digitalisation des services, comme proposée par PhonePe, peut aussi permettre un meilleur accès à l'assurance, en dehors des grandes villes[242].

Ces problématiques sont loin de la France des années 80. Si on peut associer le développement d'une classe moyenne aux Trente Glorieuses, le début des années 80 voit le retour d'une « polarisation » de la société en deux groupes distincts, très éloignés, chacun à une extrémité de l'échelle sociale. »[243]

2. Ouverture sur l'international

En France, l'ouverture du marché de l'assurance a commencé avec une série de textes visant la libre circulation des prestations et le libre établissement des opérateurs d'assurance en Europe depuis 1964.[244] En 1973, le portefeuille de l'UAP IARD hors métropole compte près de 850 000 contrats[245]

Dans les années 1980, les activités internationales des groupes d'assurance française sont déjà significatives : les activités exportatrices représentent 15,5% du chiffre d'affaires total des sociétés présentes à l'internationale en 1986, plaçant la France en 3ème position derrière la Suisse et le Royaume-Uni[246].

[238] https://www.oecd.org/finance/insurance/49674941.pdf
[239] https://icrier.org/pdf/Working_Paper_394.pdf, page 23
[240] https://www.insurancebusinessmag.com/asia/features/interviews/indian-market-undergoing-wave-of-concentration-218251.aspx
[241] https://www.atlas-mag.net/article/les-specificites-du-marche-indien-de-l-assurance
[242] https://www.cnbctv18.com/finance/busting-myths-about-indian-insurance-market-12164222.htm
[243] https://cours.unjf.fr/file.php/193/Cours/03_item/globalprintcom.htm
[244] https://www.senat.fr/rap/r98-045/r98-045_mono.html
[245] L'EVOLUTION DU MARCHE DE L'ASSURANCE AUTOMOBILE EN FRANCE DEPUIS LA GUERRE ET LE ROLE DES MUTUELLES SANS INTERMEDIAIRES, Jacques-Henri Gougenheim, Jacques Vandier, André Straus, ESKA | « Entreprises et histoire » 2013/3 n° 72 | pages 91 à 112, ISSN 1161-2770 ISBN 9782747222174
[246] Wolkowitsch Maurice. L'assurance en France. In : Annales de Géographie, t. 98, n°550, 1989. pp. 646-657 ;
Doi : https://doi.org/10.3406/geo.1989.20934
https://www.persee.fr/doc/geo_0003-4010_1989_num_98_550_20934

3- L'assurance automobile

a. L'assurance dommage (avant Covid)

En **Inde**, **l'assurance dommage représente environ 30%** de l'ensemble des primes d'assurance en 2019, avec une croissance de 17% en 2019[247]. La pénétration de l'assurance non-vie n'a atteint l'étape importante des 1% qu'en 2021[248]. L'assurance automobile, la santé et l'assurance des récoltes sont les moteurs de la croissance de l'assurance dommage en Inde.

L'**assurance automobile** en Inde comptait pour **36.5% des primes d'assurance non-vie** en Inde en 2018 (versus 27% pour la santé et 7,5% pour les récoltes)

En **France**, la part de l'assurance dommage a cru jusqu'en 1983 dans les primes d'assurance, sa décroissance étant ensuite expliquée par la désinflation[249], et surtout par une croissance de l'activité vie plus forte[250]. **En 1980, l'assurance dommage représente 64% des primes** d'assurance ; en 1988, ce chiffre tombe à 39%[251]. Ce plafonnement des primes dommage peut aussi s'expliquer par la moindre croissance du pouvoir d'achat, l'augmentation du coût de l'énergie, la baisse du nombre de sinistres et la baisse proportionnelle du nombre de jeunes conducteurs.[252]

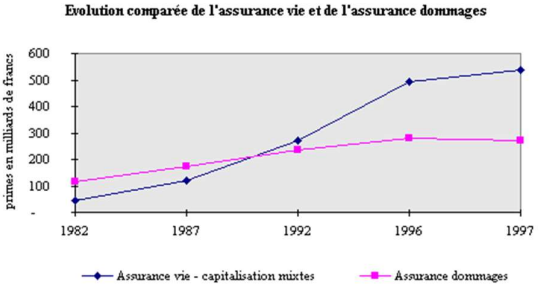


Tableau 26 : Evolution des primes de l'assurance vie et assurance dommage en France entre 1982 et 1997, Source : senat.fr

Depuis 2010, la part du marché de l'assurance dommage est de nouveau en croissance : elle est passée de **22% à 30% de l'ensemble des encaissements des assureurs**[253].

Au sein de l'assurance dommages, l'assurance automobile a culminé à la fin des années 1960, en dépassant les 40% du chiffre d'affaires total du secteur.[254] Le fort ralentissement intervient entre 1974 et 1980, la croissance du secteur se rapproche alors de celle du PIB. Ce recul relatif de l'automobile dans l'assurance-dommages[255] peut s'expliquer par la très forte concurrence entre assureurs sur cette branche, la saturation du marché de l'automobile dont la croissance annuelle du parc n'est plus que de 1,4 % sur la dernière décennie et la diminution de la fréquence des accidents[256]. En 2020, l'assurance automobile représente encore l'essentiel de l'activité non-vie, mais reste en 3ème position en termes de prime au niveau de l'assurance, avec **11,5% des encaissements**[257].

[247] https://indianexpress.com/article/business/economy/amid-covid-worries-insurance-net-spread-further-both-life-and-health-7691736/
[248] https://www.insurancebusinessmag.com/asia/news/breaking-news/pandemic-drives-insurance-penetration-in-india-303124.aspx
249 https://www.persee.fr/doc/ecofi_0987-3368_1989_num_11_3_1633, page 25
[250] https://www.senat.fr/rap/r98-045/r98-0454.html
[251] https://www.persee.fr/doc/ecofi_0987-3368_1989_num_11_3_1634, page 40
[252] Les assurances dans les années quatre-vingt : l'envolée Danielle Besson, Division Synthèse générale des Comptes Nationaux, Insee, Février 1993
[253] https://www.atlas-mag.net/article/le-marche-d-assurance-dommages-en-france#:~:text=Exprim%C3%A9e%20en%20montant%2C%20la%20charge,pertes%20d'exploitation%20et%20annulation.
[254] https://www.persee.fr/doc/rei_0154-3229_1994_num_70_1_1545
[255] Les assurances dans les années quatre-vingt : l'envolée Danielle Besson, Division Synthèse générale des Comptes Nationaux, Insee
[256]Les assurances dans les années quatre-vingt : l'envolée Danielle Besson, Division Synthèse générale des Comptes Nationaux, Insee
[257] https://www.atlas-mag.net/article/le-marche-d-assurance-dommages-en-france#:~:text=Exprim%C3%A9e%20en%20montant%2C%20la%20charge,pertes%20d'exploitation%20et%20annulation.

b. Les principaux acteurs du marché

Quels sont les acteurs de l'assurance automobile ? Comment est-ce reparti entre privé et public ?

Pour l'assurance automobile en Inde, les principaux acteurs sont[258] :

- **Pour le secteur public avec 37% des primes d'assurance automobile encaissées en 2019** (versus 41% en 2018/2019) avec respectivement :
    - The New India Assurance Co. Ltd., 13% des primes
    - United India Insurance Co. Ltd. 9% des primes
    - National Insurance Co. Ltd., 8% des primes
    - The Oriental Insurance Co. Ltd. 6% des primes
- Pour le secteur privé avec 63% des primes encaissées en 2019 (versus 59% en 2018/2019) avec respectivement :
    - **ICICI Lombard General Insurance Co. Ltd., 10% des primes**
    - Bajaj Allianz General Insurance Co. Ltd., 8% des primes
    - Tata AIG General Insurance Co. Ltd., 6% des primes
    - Cholamandalam MS General Insurance Co. Ltd., 5% des primes
    - HDFC ERGO General Insurance Co. Ltd, 5% des primes
    - IFFCO Tokio General Insurance Co. Ltd., 5% des primes
    - Reliance General Insurance Co. Ltd., 5% des primes
    - **Future Generali India Insurance Co. Ltd., 2% des primes**

La **compétition est particulièrement forte** entre les acteurs car ils se concentrent tous sur certains secteurs, comme l'assurance automobile et sur une population urbaine[259].

Pour l'assurance automobile en France, les principaux assureurs en 2021[260] sont, classés par chiffre d'affaires décroissant :

- Covéa
- Axa
- Groupe MACIF
- Groupama
- Allianz
- Groupe Maif
- Crédit agricole Assurance
- Groupe des assurances du Crédit Mutuel
- Generali
- Groupe Matmut

Sur l'assurance non-vie, l'ACPR recense plus de 349 entreprises[261]. Une des caractéristiques de l'assurance dommage en France est **l'importance des mutuelles sans intermédiaires** dans l'assurance dommage. En 1989, leur part de marché est passée de 28% en 1989 à 32% en 1997 dans

[258] Handbook of the regulator, part 2, onglet 50
[259] https://www.ideasforindia.in/topics/money-finance/india-s-insurance-sector-challenges-and-opportunities.html
[260] L'Argus des assurances
[261] Les chiffres du marché de l'assurance, onglet G2

l'assurance dommage[262]. En 1983. La MACIF est la première société automobile de France, vingt-deux ans après sa création[263].

La **concurrence est particulièrement forte** sur « les catégories à déroulement court (automobile, dommages aux biens des particuliers et dommages corporels), qui présentent des barrières à l'entrée moins élevées » [264]. La concentration des acteurs a continué à augmenter depuis les années 90 : en 1994, les 10 principaux acteurs représentaient 57% des parts de marché, en 2014 ils comptaient pour 60,8%[265] en assurance automobile.

c. Les produits et leurs tarifications

Quels sont les couvertures les plus vendues ? Est-ce que les tarifs sont soumis à des contraintes règlementaires ? Quel est le prix moyen d'un produit vendu ? Comment évoluent les tarifs ?

En **Inde**, le régulateur a autorisé en 2018[266] la vente de 3 types d'offres d'assurance : un package d'assurance, une police long terme au tiers ou une police longue terme tous risques[267]. Par long terme, on entend une couverture de 3 ans pour une voiture et 5 ans pour un deux roues. Avant, seules les polices au tiers étaient disponibles séparément, pas les polices tous risques. **Un automobiliste peut maintenant acheter ces 2 types de polices auprès de 2 assureurs différents.**

La proportion de polices d'assurance **au tiers** était dans une tendance **haussière jusqu'à 2018**, passant de 101 milliards de roupies en brut en 2010 pour atteindre 278[268]. En 2018, les primes étaient à 59% pour des **couvertures au tiers**, elles sont passées à **62**% en 2019[269].

La tarification des assurances a été dérégulée, progressivement de 1994 à 2007, à l'exception du tarif des polices d'assurance automobile au tiers. Les remises ne sont plus plafonnées depuis janvier 2008, sous réserve d'une demande auprès du régulateur. La **liberté sur les prix pour les couvertures tous dommages a entraîné une réduction des prix, tandis que les garanties au tiers souffrent d'un tarif insuffisant**[270].

Cylindrée voiture particulier	Tarif FY 2019-2020 - INR[271]	EUR[272]
En dessous de 1000cc	2,072	26
Entre 1000 et 1500 cc	3,221	41
Au-dessus de 1500cc	7,890	100

[262] https://www.senat.fr/rap/r98-045/r98-045_mono.html

[263] L'évolution du marché de l'assurance automobile en France depuis la guerre et le rôle des mutuelles sans intermédiaires, page 99

[264] https://www.institutdesactuaires.com/magazine/article/aux-sources-de-la-rentabilite/2441

[265] https://acpr.banque-france.fr/sites/default/files/medias/documents/201605-as64-elements-analyse-cycles-assurance-non-vie.pdf,page 8

[266] https://economictimes.indiatimes.com/wealth/insure/motor-insurance/how-new-vehicle-owners-can-buy-renew-1-year-own-damage-cover-with-long-term-third-party-insurance/articleshow/70024568.cms?from=mdr

[267] https://www.moneycontrol.com/news/business/personal-finance/why-a-long-term-own-damage-motor-insurance-policy-is-a-must-have-5559161.html Depuis ces offres long termes ont été interdites

[268] https://www.statista.com/statistics/1075225/india-gross-weighted-premiums-motor-own-damage-insurance/

[269] Handbook 2019-2020 part II, onglet 51

[270] http://www.actuariesindia.org/downloads/gcadata/11thGCA/Detariffing%20Presentation.pdf

[271] https://www.policybachat.com/articles/car-insurance-price-list#:~:text=THIRD%2DPARTY%20INSURANCE%20PREMIUM&text=The%20average%20cost%20of%20car,Rs%204000%2D5000%20per%20year.

[272] Taux moyen 2019 INR-EUR : 78.84 INR = 1 EUR Source : https://www.exchangerates.org.uk/EUR-INR-spot-exchange-rates-history-2019.html#:~:text=Average%20exchange%20rate%20in%202019%3A%2078.8393%20INR.

En **France, dans les années 80**, la concurrence tarifaire est forte entre les différents acteurs. Les mutuelles professionnelles et sans intermédiaires, dont la MACIF fondée en 1959, bouleverse le marché en proposant des **tarifs fortement réduits. Cela est rendu possible grâce** à **une vente directe de leurs produits** (sans passer par des réseaux d'agents généraux ou des intermédiaires indépendants), ce qui permet de réduire leurs frais de gestion et en pratiquant **la segmentation et la sélection des risques**[273].

Presque **40 ans plus tard**, la sélection des risques est pratiquée par tous les acteurs du marché. Les prix sont régulièrement à la hausse[274] : le montant moyen de la prime d'assurance a augmenté de 12 % entre 2015 et 2019, à cause d'événements climatiques exceptionnels de plus en plus fréquents (comme les tempêtes Carmen et Eleanor durant l'hiver 2017/2018) et la **hausse du coût des réparations.** Les prix des pièces détachées auraient ainsi augmenté de 6% par an depuis 2013[275].
Malgré cette hausse de prix, **47% des Français optaient pour une couverture tous risques**[276], et la formule au tiers était choisie que par 29% des Français. En 2020, le prix moyen d'une assurance au tiers était de 528 euros, 596 euros pour une formule intermédiaire et 715 euros pour une couverture tous risques[277].

d. La distribution

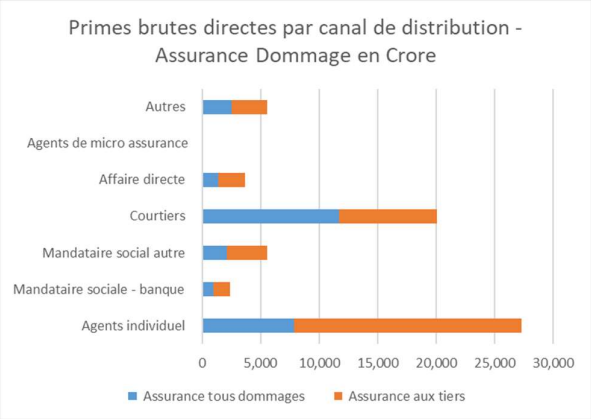


Tableau 28 : Répartition des primes brutes par canal de distribution en Inde de l'assurance dommage en Crore

En **Inde**, la libéralisation de l'assurance a permis le développement de nouveaux canaux de distribution, distincts des agents ou des employés : les courtiers et les mandataires sociaux (dont la banque assurance). Les **agents individuels représentent encore 46% des primes perçues** sur l'année financière 2018/2019, **suivis des courtiers avec 34%**[278]. Les mandataires sociaux (« corporate agent ») contribuent pour 13% des primes - dont 4% pour la banque, alors que la vente en direct atteint 6% des primes. La tendance est à une **diminution de la part des agents individuels** (-2% en 2019/2020) **au profit des affaires directes et des autres canaux de distributions** (« Others »).
En termes de nombre de points de distribution, les entreprises publiques sont encore majoritaires pour l'assurance non-vie : 8 150[279] offices de représentations appartiennent à des entreprises publiques sur un total de 10 695, malgré la création de 416 représentations pour des entreprises privées en 2019.

En **France**, le marché des assureurs dans les **années 80** est bousculé par l'arrivée de nouveaux acteurs : **les sociétés sans intermédiaires** (SSI), qui incluent les filiales de banque. Ils vont bénéficier

[273] https://www.chaire-pari.fr/wp-content/uploads/2020/04/WP-19-Julien-Caranton.pdf, page 3
[274] https://www.largus.fr/actualite-automobile/assurances-12-daugmentation-en-5-ans-10337397.html
[275] https://www.turbo.fr/actualite-automobile/le-prix-des-pieces-detachees-aurait-augmente-de-44-depuis-2013-182480
[276] https://www.largus.fr/actualite-automobile/assurances-12-daugmentation-en-5-ans-10337397.html
[277] https://www.lelynx.fr/assurance-auto/comparaison/pas-cher/tarif/#:~:text=En%202020%2C%20il%20fallait%20compter,pour%20une%20formule%20tous%20risques.
[278] Handbook 2019-20202 part II – onglet 51.
[279] https://www.atlas-mag.net/en/article/indian-insurance-market-structure

de l'accroissement du marché automobile : « entre 1970 et 1983, le parc de véhicules assurés a augmenté de 41% dont 6% pour les sociétés traditionnelles et 113% pour les sociétés à gestion directe. »[280] Au cours des années 80 et 90, la **bancassurance** s'affirme progressivement dans l'assurance vie et se positionne un peu plus tardivement sur l'assurance dommage[281].
Les mutuelles arrivent à des parts de marchés substantielles dans l'assurance auto en 1980 : les mutuelles dites de Niort (MACIF, MAIF, MAAF, MATMUT, GMF) comptent pour 36% des encaissements du marché de l'assurance automobile[282].
Leurs tarifs avantageux sont liés à la sélection des risques, mais aussi à leur **distribution** : en non-vie, leurs coûts de distributions sont en moyenne 7 points en dessous de ceux des agents et leurs coûts de gestion presque 6 points en dessous[283].
En 1985, on compte 211 100 emplois dans l'assurance, à 82% salariés, avec une population de 22 500 agents généraux[284].

Presque 40 ans plus tard, les mutuelles ont pris une place prépondérante dans l'assurance automobile : la vente directe représente environ 35% des primes d'assurance non-vie, alors que les agents ne participent plus que pour 31%. Les courtiers maintiennent leur part à 17%[285]. La bancassurance est dans une nette dynamique de croissance[286] à 15%.

e. La sinistralité

En **Inde, pour 100 véhicules assurés, des sinistres seront déclarés pour entre 35 à 40 véhicules**[287]. La moyenne du **taux de sinistre à prime net du marché est supérieure à 80 %** : en 2018/2019, il s'élevait à 90.6% avant de redescendre à 85.6% en 2019/2020[288]. En particulier dans l'assurance non-vie, le taux de sinistres était croissant avant le Covid. **Pour rester profitable, les assureurs devenaient de plus en plus dépendants des produits financiers de leurs investissements**[289].

En **France, la sinistralité a fortement diminué depuis les années 80** : en 1979, on compte environ 152 accidents pour 1 000 automobiles[290] En 1985, on compte 191 132 blessés (-26% sur 10 ans) et 10 432 morts (en baisse de 20% sur 10 ans) sur les routes pour 24 millions de véhicules.[291]Les résultats techniques se sont améliorés depuis 1981, « en raison de la baisse remarquable de la fréquence des sinistres »[292] Entre 1981 et 1990, la croissance moyenne annuelle des indemnités est de 9,7 % dans l'assurance-dommages. Pour le risque automobile, elles augmentent régulièrement, mais à un rythme plus modéré de 6 % l'an.[293]

[280] https://www.persee.fr/doc/rei_0154-3229_1994_num_70_1_1545, page 25
[281] file:///C:/Users/ceciler/Downloads/Les_assurances_et_les_banques_fin_de_partie.pdf, page 15
[282] https://www.senat.fr/rap/r98-0452/r98-045285.html
[283] https://www.senat.fr/rap/r98-0452/r98-045285.html
[284] https://www.persee.fr/doc/geo_0003-4010_1989_num_98_550_20934?q=Wolkowitsch+Maurice.+L%27assurance+en+France, page 648
[285] https://fr.statista.com/statistiques/491070/canaux-distribution-primes-assurance-non-vie-france/
[286] https://www.latribune.fr/entreprises-finance/banques-finance/assurance/les-bancassureurs-bousculent-les-leaders-de-l-assurance-dommages-aux-particuliers-876012.html
[287] https://economictimes.indiatimes.com/news/politics-and-nation/grand-theft-auto-car-thieves-love-for-suvs-is-giving-sleepless-nights-to-insurance-cos/articleshow/71120982.cms?from=mdr
[288] Handbook 2019/2020 onglet 54
[289] https://www.econstor.eu/bitstream/10419/242873/1/Working-Paper-394.pdf, page 28
[290] Thiry Gilbert. Les risques d'accidents de la route. In : Economie et statistique, n°117, Décembre 1979. Emploi agricole / Les risques d'accidents de la route / Téléphone / Les revenus en 1975 / La circulation de l'information. pp. 13-22 ;
[291] https://fr.wikipedia.org/wiki/Parc_automobile_fran%C3%A7ais
[292] Teyssier Robert. Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45. Doi : 10.3406/ecofi.1988.1545
http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545, Page 34
[293] Les assurances dans les années quatre-vingt : l'envolée Danielle Besson, Division Synthèse générale des Comptes Nationaux, Insee

La désinflation à partir de 1986, va permettre aussi une meilleure couverture des risques par les provisions techniques[294].

Depuis **2010**, la **fréquence des accidents corporels et matériels ne cesse de décroître, mais leur coût est de plus en plus élevé**. Sur une base de 100 en 2010, on compte en 2018, plus que 77 accidents matériels et 88 accidents corporels[295]. En particulier pour les dommages corporels, l'indemnité moyenne versée a été multipliée par trois depuis 1970 par rapport à 2006[296].

Le **ratio sinistres/ primes brutes en 2018 et 2019 s'élève à 69,7% et 71,9% pour l'assurance dommages**[297].

f. Les frais d'exploitation

Comment impactent les coûts d'acquisition et d'administration sur la rentabilité de l'assurance ? Prenons le ratio combiné[298] pour comparer leurs impacts.

En **Inde**, en 2018, **les commissions et frais du management correspondaient à 30% des primes acquises nettes en moyenne pour les assureurs non-vie**. Ce ratio monte à 31% en 2019[299].

On retrouve ce même ordre de grandeur chez les compagnies traditionnelles en automobile dans les **années 80** : « les frais de gestion étaient de l'ordre de 33% à 35 %, notamment 20 % de commission pour les agents plus 15% de frais de siège, de frais de structures » [300] **Jacques Vandier, dirigeant de la MACIF, affirme que pour son entreprise, ces frais étaient réduits à « 15%, avec aucun frais de production**, uniquement les frais de structure, les salaires des rédacteurs. » Grâce au **développement du constat à l'amiable et au recours forfaitaire**, les coûts de gestion des sinistres sont aussi optimisés : « 98 % des sinistres passent dans la zone couverte par le forfait. Seuls 2 % des sinistres sont en dehors de la zone et ceux-là sont réglés au cas par cas. »

VALEURS MOYENNES POUR LES ASSURANCES IARD
(PÉRIODE 1982-1987)

Forme juridique	Nombre	Coût	Primes brutes (milliers de francs à prix de 1980)	Primes nettes	Prestations	Indicateur de distribution	Indicateur de réassurance	Coûts moyens		
								Primes brutes	Primes nettes	Prestations
Sociétés nationales	4	1 976 963	6 203 745	5 158 687	4 222 336	0,516	0,161	0,319	0,381	0,466
Mut. avec intermédiaire	63	79 881	233 315	187 020	145 809	0,401	0,436	0,372	1,041	4,808
Mut. sans intermédiaire	16	188 264	859 209	797 008	721 201	0,004	0,246	0,265	0,418	0,693
Mut. locales et professionnelles	16	29 379	141 276	97 788	108 884	0,019	0,497	0,231	0,586	1,300
Mut. agricoles	15	64 635	479 376	204 110	167 065	0,143	0,439	0,226	0,406	0,820
Soc. anonymes françaises	100	120 762	349 915	260 512	192 548	0,395	0,349	0,470	1,137	3,419
Sociétés CEE	87	12 162	34 605	21 508	16 635	0,726	0,343	0,328	0,693	1,194
Sociétés hors CEE	30	41 312	119 580	80 301	58 859	0,719	0,327	0,345	0,613	0,812
TOTAL	331	95 969	315 111	242 075	194 871	0,465	0,366	0,368	0,851	2,482

Tableau 29 : Indicateurs moyens des sociétés d'assurance IARD en France entre 1982 et 1987, Source : Revue d'économie financière n.11

En 2019, au niveau de l'assurance dommage, le **ratio des frais de gestion moyen est de 24,80%**, en hausse légère mais régulière[301].

[294] Teyssier Robert. Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45. Doi : 10.3406/ecofi.1988.1545 http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545, page 37
[295] https://www.ffa-assurance.fr/conference-presse-mars-2019/pdf/20190326_FFA_conference_presse_mars_2019.pdf, page 22
[296] https://www.lesechos.fr/2006/11/les-dommages-corporels-principal-facteur-de-derive-des-couts-586665
[297] https://www.atlas-mag.net/article/le-marche-d-assurance-dommages-en-france#:~:text=Exprim%C3%A9e%20en%20montant%2C%20la%20charge,pertes%20d'exploitation%20et%20annulation.
[298] Ratio combine : rapport entre les décaissements (frais de gestion, commissions versées, provisions pour sinistres, et remboursement des sinistres), et les encaissements (primes et cotisations encaissées).
[299] Hand Book 2019 2020 Part 2, onglet 56
[300] L'Evolution du marché de L'assurance automobile en France depuis la guerre et le rôle des mutuelles sans Intermédiaires, André Straus, Page 95
[301] https://www.atlas-mag.net/article/le-marche-d-assurance-dommages-en-france#:~:text=Exprim%C3%A9e%20en%20montant%2C%20la%20charge,pertes%20d'exploitation%20et%20annulation.

En termes de **ratios combinés**, **en Inde, la moyenne du marché est passée de 106% à 107% entre 2018 et 2019**[302]. En 2019, seuls 3 assureurs avaient des ratios inférieurs à 100% : SBI General (98%), Shriram (91%) et Universal Sompo (99%) en assurance non-vie.

En France, dans les années 80, les primes ne couvrent pas non plus les charges des sinistres et charges d'exploitation. En effet, les résultats techniques (qui comprennent les frais d'acquisition, variation du montant des frais d'acquisition reportes, frais d'administration, et commissions reçues des réassureurs) sont fortement négatifs et **le déficit se creuse au cours des années 80**, comme le montre le tableau suivant[303].

LES RENTABILITÉS D'EXPLOITATION DE L'ASSURANCE
(en % des primes brutes)

DOMMAGES	1975	1980	1981	1982	1983	1984	1985	1986
Résultat technique	− 6,6 %	− 10,1 %	− 11,5 %	− 11,7 %	− 10,7 %	− 11,2 %	− 11,4 %	− 11,6 %
Résultat financier	8,2 %	9,9 %	10,5 %	10,8 %	11,0 %	11,4 %	11,4 %	11,4 %
Résultat d'exploitation	2,0 %	0,3 %	− 0,5 %	− 0,3 %	1,4 %	2,1 %	1,9 %	1,8 %
VIE	1975	1980	1981	1982	1983	1984	1985	1986
Résultat technique	− 22,5 %	− 28,5 %	− 29,7 %	− 33,0 %	− 39,0 %	− 36,0 %	− 35,5 %	− 34,8 %
Résultat financier	26,1 %	30,8 %	31,7 %	35,1 %	40,1 %	36,5 %	35,9 %	33,7 %
Résultat d'exploitation	3,5 %	2,4 %	2,2 %	1,6 %	1,8 %	1,0 %	0,7 %	− 0,8 %

Tableau 30 : Les rentabilités d'exploitation des assureurs français en % des primes brutes, Source : Revue d'économie financière n.7, 1988

Dans les années 2010, le ratio combiné brut est resté supérieure à 100% tous les ans à l'exception de 2018 et 2020[304], avec une moyenne de 99%.

g. Les résultats financiers

Avec des résultats techniques négatifs, les résultats financiers sont alors la clé de la rentabilité.

En **France, dans les années 80**, « La somme des produits des placements financiers et des plus-values sur cession d'élément d'actifs est passée, d'environ 15% des primes pour la branche dommages au début des années 1980, à plus de 25% dans les années 1990. **En 1993, les seules plus-values réalisées représentaient 8% du chiffre d'affaires brut** et le déficit d'exploitation 3% ; sur les vingt premiers groupes d'assurance non-vie, quinze étaient déficitaires. »[305]

PLACEMENTS DES SOCIÉTÉS D'ASSURANCES
(encours comptables au 31 décembre)

	1981		1982		1983		1984		1985		1986		1987	
Obligations	122	47,5 %	146,3	49,1 %	174,8	49,8 %	201,5	49,1 %	243,4	49,8 %	291,2	48,3 %	350,7	48,9 %
SICAV et FCP obligataires	3,3	1,3 %	5	1,7 %	7,2	2,1 %	13,2	3,2 %	23,6	4,8 %	34,5	5,7 %	42,9	6,0 %
Actions	33,4	13,0 %	34,2	11,5 %	39,4	11,2 %	45,4	11,1 %	48,4	9,9 %	70	11,6 %	39,7	13,8 %
SICAV et FCP diversifiés	6,3	2,5 %	8,3	2,8 %	12,6	3,6 %	16,8	4,1 %	28	5,7 %	43,3	7,2 %	54,3	7,6 %
Immeubles et parts de SI non cotées	53	20,7 %	60,3	20,2 %	68,2	19,4 %	75,7	18,4 %	83	17,0 %	96,1	16,0 %	103,7	14,5 %
Prêts	16,7	6,5 %	20,9	7,0 %	24,3	6,9 %	26,6	6,5 %	27,5	5,6 %	27,7	4,6 %	24,4	3,4 %
Billets et certificats	1,5	0,6 %	1,1	0,4 %	0,9	0,3 %	1,5	0,4 %	2	0,4 %	3,3	0,5 %	9	1,3 %
Bons et dépôts	11,3	4,4 %	13	4,4 %	13,4	3,8 %	18,1	4,4 %	20,6	4,2 %	21,7	3,6 %	12,9	1,8 %
Placements non affectables	9,1	3,5 %	8,9	3,0 %	10,2	2,9 %	11,8	2,9 %	12,7	2,6 %	14,7	2,4 %	19,9	2,8 %
Total	256,6	100,0 %	298	100,0 %	351	100,0 %	410,6	100,0 %	489,2	100,0 %	602,5	100,0 %	716,5	100,0 %

Tableau 31 : Placement des sociétés d'assurances français, au 31 décembre, Source : Teyssier Robert, Réflexions sur les rentabilités de l'assurance

Le portefeuille est dominé par les obligations (de 47,5% des placements en 1981 à 48,9% en 1987), les immeubles et parts de SI non cotés (de 20,7% à 14,5%) et les actions (de 13% à 13,8%). Les immeubles étaient alors comptabilisés en valeur historique pour des biens datant parfois du début du siècle[306].

[302] Handbook part 2, onglet 56
[303] Teyssier Robert. Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45. Doi : 10.3406/ecofi.1988.1545 http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545
[304] https://www.atlas-mag.net/article/le-marche-d-assurance-dommages-en-france#:~:text=Exprim%C3%A9e%20en%20montant%2C%20la%20charge,pertes%20d'exploitation%20et%20annulation.
[305] Revue d'économie industrielle, N.70 4eme trimestre 1994, Concentration, contestabilité et concurrence sur le marché français de l'assurance, Armand-Denis SCHOR, Page 18
[306] https://www.lesechos.fr/1993/03/une-grande-solidite-financiere-900329

A noter que l'ajustement boursier de l'automne 1987[307] a eu un impact limité à court terme sur le bilan des assureurs grâce à cette composition de portefeuille[308].

Presque 40 après, **les produits financiers représentent environ 4% des primes acquises**, en légère décroissance jusqu'à 3,87% en 2019[309].

Les assureurs non-vie placent majoritairement leurs investissements en actions et titres de créances long termes (42% des encours), en titres de capital (29% des encours, particulièrement en actions non cotées) et OPC non monétaires (15% des encours)*[310]*

Placements financiers des sociétés d'assurance sur 2019

Source : Banque de France

(encours au prix de marché, variation, flux de souscription, valorisation, en milliards d'euros)

	Assurances Vies et Mixtes			Assurance Non-Vie			Total Assurances		
	Flux	Valorisation	Encours	Flux	Valorisation	Encours	Flux	Valorisation	Encours
Monnaie et dépôts	2,9	0,0	32	1,9	0,0	13	4,8	0,0	46
Titres de créance									
CT (maturité d'origine)	-10,5	0,0	14	0,1	0,0	0	-10,3	0,0	15
LT (maturité d'origine)	54,8	44,3	1 481	-3,4	2,8	114	51,4	47,0	1 595
Prêts	3,3	0,0	45	0,1	0,0	9	3,4	0,0	55
Titres de capital									
Actions cotées	-1,6	13,9	78	-0,6	3,1	15	-2,2	17,0	93
Actions non cotées	4,3	2,3	60	1,7	0,2	30	6,0	2,5	90
Autes	3,1	0,8	23	0,8	2,3	27	3,9	3,2	50
OPC									
Fonds monétaires	1,7	-0,4	78	1,1	0,0	8	2,9	-0,4	85
Fonds non monétaires	9,8	58,7	633	-0,5	2,4	37	9,3	61,1	670
Dérivés	-0,1	0,0	5	0,1	0,0	0	-0,1	0,0	5
Total	**67.8**	**119.6**	**2 450**	**1,3**	**10,7**	**253**	**69,1**	**130.4**	**2 703**

Tableau 32 : Placements financier des sociétés d'assurance français en 2019, source : Banque de France

En **Inde**, si on considère les 3 principaux assureurs non-vie public, **le ratio des résultats d'investissements sur les primes nettes acquises[311] est assez stable depuis 2015 à 20%,** avec des pics à 23% en 2017 et 27% en 2019[312] sur l'activité non-vie.

h. Synthèse

Pierre-Charles Pradier affirme que « Les assurances de dommages sont à l'origine de l'assurance : le prêt à la grosse pour les caravanes commerciales, puis les opérations maritimes, existe depuis l'antiquité (Pradier, 2010), les assurances incendie depuis le grand incendie de Londres (1666). **Il semble naturel d'interpréter la relation entre croissance économique et demande d'assurance comme une extension du désir de protection aux biens acquis par la croissance** : logement, automobile, biens capitaux nécessaires à la production »[313] Et effectivement les croissances des assureurs en Inde et en France au début des années 80 sont **nées à la suite de périodes de forte croissance et à un équipement automobile qui se démocratise.**

Cette croissance, même si plus réduite en France, s'accompagne **de taux d'intérêt particulièrement élevés, et de résultats financiers importants**. En même temps, **la concurrence** est extrêmement forte entre les acteurs. Ce contexte a permis que **la rentabilité technique soit « moins considérée comme primordiale**, tant que l'existence de plus-values latentes permettait aux résultats financiers de venir combler les pertes du compte d'exploitation »[314]. En Inde, la rentabilité devient de plus en plus dépendante des produits financiers avec une augmentation des pertes de souscription[315] et des ratios combinés élevés. Il en suivra en Inde, comme cela a pu être le cas en France, **« un assujettissement plus fort que par le passé des entreprises d'assurances aux performances**

[307] https://www.lefigaro.fr/flash-eco/2012/10/19/97002-20121019FILWWW00596-le-krach-de-1987-c-etait-il-y-a-25-ans.php
[308] Teyssier Robert. Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45. Doi : 10.3406/ecofi.1988.1545 http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545, page 37

[309] Chiffres du marché de l'assurance 2019, onglet T10
[310] https://www.banque-france.fr/statistiques/placements-des-assurances-2019t4
[311] (profit/loss on sale/redemption of investment + other income (net of other expenses) + Interest, Dividend & rent- gross)/ premium earned (Net)
[312] Hand Book part 2, onglet 60
[313] La libéralisation des assurances, Pierre-Charles Pradier, https://www.aef.asso.fr/publications/revue-d-economie-financiere/137-40-ans-de-liberalisation-financiere/3727-la-liberalisation-des-assurances
[314] file:///C:/Users/ceciler/Desktop/References/EVOLUTION%20MARCHE%20AUTO%20FR.pdf, p109
[315] https://indianexpress.com/article/business/insurers-underwriting-losses-rise-to-rs-22859-crore-three-psu-players-among-worst-hit-7074521/

techniques de la gestion des risques » ; (…) marquée par son aptitude à concurrencer et à compléter les services rendus par les banques et les établissements financiers[316]

[316] Teyssier Robert. Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45. Doi : 10.3406/ecofi.1988.1545, http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545

2ème partie : La fraude à l'assurance automobile, en France et en Inde

L'objectif de ce chapitre est de mieux comprendre la notion de fraude à l'assurance. Le point de départ de cette analyse sera la définition juridique de la fraude, pour ensuite son importance économique, et enfin lister les mécanismes de lutte autant au niveau légal, règlementaire, qu'au sein de la profession. Ces travaux seront conclus par une description du dispositif de lutte anti-fraude au sein de l'entité étudiée, au début de cette étude.

1. Définition et impacts théoriques

« La fraude est l'hommage que la force rend à la raison. »

Charles P. Curtis, issu du livre « Un Livre bien ordinaire »

a. Définition de la fraude en France et en Inde

En France, le mot même de fraude aurait été emprunté au latin *fraus, fraudis*, « tromperie ; dommage ; délit » au 13eme siècle. Il est défini par le dictionnaire de l'académie comme une « action accomplie de mauvaise foi au préjudice d'une personne ou d'une collectivité. »[317].

Mais il n'existe **pas à proprement parler de définition légale ou réglementaire de la fraude**. Le Code des assurances mentionne la fraude à 5 reprises, en explicitant les sanctions en cas de fraude[318] (Article L172-2, L172-6, L172-7, L172-8, Article L172-9 en date du 20 Mai 2021). Il n'y a pas d'incrimination spécifique dans le code pénal non plus. *Seul le Code du Commerce* précise qu'il « s'agit d'un acte qui a été réalisé en utilisant des moyens déloyaux destinés à surprendre un consentement, à obtenir un avantage matériel ou moral indu ou réalisé avec l'intention d'échapper à l'application d'une loi impérative ou prohibitive » (Article L. 650-1 du code de commerce[319])

Si on se base sur la définition de professionnels, on peut citer la définition de l'ALFA (Agence de Lutte contre la Fraude à l'assurance[320]), qui le spécifie comme **« un acte intentionnel, réalisé par une personne morale ou physique, afin d'obtenir indûment un profit du contrat d'assurance** ».
Plus précisément, Insurance Europe caractérise la fraude quand au moins une de ces conditions est présente :

- Des informations incomplètes ou mensongères sont fournies lors de la souscription ou comme réponse à un formulaire de proposition d'assurance
- La déclaration d'un sinistre basée sur des circonstances mensongères ou fallacieuses, y compris exagérant un véritable sinistre

[317] https://cnrtl.fr/definition/academie9/fraude

[318] https://www.legifrance.gouv.fr/codes/id/LEGISCTA000006174132/

[319] Dictionnaire du Droit privé par Serge Braudo

[320] https://www.alfa.asso.fr/fraude-a-lassurance/#definition

- Sinon toute autre utilisation d'informations mensongères ou fallacieuses à un assureur avec l'intention d'en obtenir un bénéfice utilisant le contrat d'assurance[321]

Généralement les pratiques métier définissent 4 éléments pour caractériser une fraude :
- prouver l'intention et la connaissance
- montrer que la fausse déclaration a un impact direct sur la suite à donner à la demande d'indemnisation
- montrer que le réclamant pourra bénéficier d'un gain pécunier ou a l'intention d'en avoir
- être capable d'identifier clairement le coupable de la fraude.

En Inde, **le cadre légal permettant de définir une fraude est encore plus limité**. En effet, la fraude n'est pas mentionnée dans le Indian Insurance Act[322], il n'est pas non plus défini dans le Indian Contract Act de 1872. Seul le *code pénal indien datant de 1860* mentionne la contrefaçon et les actes frauduleux. Mais selon Business Today, sa définition de la fraude est restrictive, et ne permet pas, par exemple, de poursuivre les fraudeurs dans le cas de la fraude organisée[323]. Entre le retard chronique des cours de justice en Inde, l'absence de textes juridiques permettant de sanctionner la fraude sans aller au pénal, l'investissement en temps et énergie nécessaires devant les tribunaux, les compagnies d'assurance sont découragées de poursuivre en justice les fraudeurs.

Les assureurs indiens peuvent cependant s'appuyer sur la définition de l'International Association of Insurance Supervisors, reprise par l'Insurance Regulatory and Development Authority, comme « un acte ou une omission fait dans le but de gagner de façon malhonnête ou illégale un avantage pour l'intervenant qui commet la fraude ou d'autres intervenants qui lui sont liés[324] »
La chambre de commerce et d'industrie indienne définit la fraude comme « l'acte de faire une fausse déclaration pour induire une tierce partie en erreur et lui faire établir un contrat ou payer une demande d'indemnisation. Cet acte doit être volontaire et délibéré, impliquant des gains financiers, grâce à des faux semblants, et est illégal »[325].

Le système juridique indien est fortement influencé par la Common law, héritée de la colonisation britannique. Il est donc à noter que **l'Insurance Act, révisé en 2015**, a permis d'affiner le cadre juridique de la fraude.
Une *obligation de juste présentation* du risque a été introduite[326]. Il n'est donc plus nécessaire de prouver la mauvaise foi pour qualifier le contrat comme une fraude. La fausse déclaration du risque suffit pour enclencher des poursuites.
En cas de présentation du risque déformée, l'assureur peut garder la prime et reste redevable de tous les sinistres avant la fraude. Il a l'option d'arrêter le contrat à la date de la fraude.[327]
Ce nouveau règlement a été adopté en Australie, Nouvelle Zélande et Singapour, on peut donc se demander si elle sera aussi à terme utilisée en Inde.

[321] https://www.insuranceeurope.eu/sites/default/files/attachments/Insurance%20fraud%20-%20not%20a%20victimless%20crime_0.pdf
[322] https://blog.ipleaders.in/false-insurance-claims/
[323] https://www.businesstoday.in/opinion/columns/insurance-frauds-control-act-an-urgent-need-in-india-fraudulent-claims-indian-penal-code/story/400212.html
[324] "An act or omission intended to gain dishonest or unlawful advantage for a party committing the fraud or for other related parties."
[325] The act of making a statement known to be false and used to induce another party to issue a contract or pay a claim. This act must be wilful and deliberate, involve financial gain, done under false pretences and is illegal."
[326] https://kennedyslaw.com/media/2352/kennedys-the-insurance-act-2015.pdf
[327] "If the insured makes a fraudulent claim, the entire claim is forfeit (including any honest part of the claim) and the insurer can keep the premium. The insurer will remain liable for any genuine claims before the fraud but has the option to terminate the policy from the date of the fraud."

b. Les types de fraudes

Ce mémoire porte sur la détection de la **fraude externe, c'est-à-dire commise par des clients, bénéficiaires ou autres personnes assurées**. La fraude interne, c'est-à-dire perpétrée par des collaborateurs de la société d'assurance (agents, ...) ne sera pas traitée dans cette étude, comme celle des intermédiaires externes (courtiers, …) ou les prestataires/fournisseurs (les garages).

Généralement on distingue la fraude en fonction du moment où elle se produit dans la vie du contrat :

A la souscription	- Fausse déclaration : o Risque présenté minoré, avec un usage non exact (mauvais conducteur principal) o Demande de couverture pour un risque déjà existant au jour de la souscription
Durant la vie du contrat	- Fraude à la déclaration du retrait du permis de conduire - Non déclaration de sinistre : "l'assuré est obligé de donner avis à l'assureur […] de tout sinistre de nature à entraîner la garantie de l'assureur", d'après l'article L113-2 du Code des assurances. C'est-à-dire qu'un sinistre même très réduit doit être déclaré à l'assureur[328]
Après la survenance du sinistre	- Déclaration d'un faux sinistre : Déclaration d'un sinistre qui n'a pas eu lieu, déclaration d'un sinistre perpétré par le client contre ses propres intérêts assurés, avec ou sans l'aide de tierces personnes, … o Faux vol, o Faux accident o Faux incendie o … - Inexactitude dans la déclaration du sinistre o Déclaration d'un sinistre supérieur à celui réel : Demande de remboursement d'un montant supérieur aux dommages réels o Déclaration de dommages qui n'ont pas de lien de causalité avec le sinistre déclaré (dommage préexistant par exemple) o Déclaration d'un autre conducteur au volant lors du sinistre o Responsabilité attribuée à tort à un tiers o …

Nous nous concentrerons sur la fraude après survenance d'un sinistre, en nous basant sur l'analyse des déclarations de sinistres.

[328] https://www.lci.fr/conso-argent/etude-fraude-a-l-assurance-11-des-francais-avouent-s-y-etre-adonne-quels-sont-les-risques-lelynx-fr-csa-2136887.html

c. Les impacts de la fraude

Quels sont les impacts de la fraude à l'assurance ?

La fraude impacte le **positionnement d'une société d'assurance par rapport à ses concurrents**. En détruisant la notion de mutualité, la fraude affecte à la fois la tarification et le provisionnement d'une compagnie d'assurance.
En effet, si la fraude n'est pas détectée, les *prix des produits d'assurance seront plus élevés pour les assurés*. La démonstration est effectuée par Jonathan Karsenty dans son mémoire « La détection des fraudes à l'assurance »[329] . La fraude augmente non seulement le prix moyen des sinistres, mais ajoute aussi des sinistres supplémentaires. La prime que va payer l'assuré honnête va donc être majorée. Cela signifie que la compagnie d'assurance proposera des tarifs plus onéreux que la concurrence, risquant de créer une *antisélection* parmi ses clients.
La fraude va aussi impacter le *provisionnement* : l'assureur va projeter plus que sa réelle sinistralité et sa rentabilité en sera limitée.

La fraude peut affecter le secteur **de l'assurance dans sa globalité**, en limitant sa capacité d'investissement et sa clientèle. Car si la fraude est perçue comme généralisée ou dominante par les clients, cela peut *affecter la volonté de s'assurer de la population (au-delà des minimums légaux).*

Or l'assureur a un impact important au niveau de la société, apportant notamment une sécurité financière nécessaire au développement.
En Inde en particulier, la perte d'un être cher lors d'un accident de voiture peut signifier, pour une famille, la plongée dans la *précarité financière*. Il n'existe pas de compensation financière dictée par la loi avec un délai donné, et les frais médicaux peuvent être particulièrement élevés[330]. L'assurance permet de réduire ces impacts et contribue aux initiatives pour rendre les routes plus sûres (initiatives de prévention, surcoût en cas de conduite dangereuse, …).
L'assurance permet aussi de créer une *sécurité, indispensable à la création de richesse*. Ainsi le Rapport sur le Développement dans le Monde 2000-2001 identifiait 3 domaines clés pour réduire la pauvreté : les opportunités, la prise de pouvoir et la sécurité[331]. En protégeant des aléas de la vie, l'assureur contribue à une société plus sûre, qui permet la prise de risque nécessaire à l'innovation et à la création de richesse.

Enfin la fraude peut aussi devenir une **source de revenus pour des entreprises de crimes organisés**.[332] Et l'impact humain peut être aussi particulièrement important, par exemple dans les accidents mis en scène[333], en mettant d'autres personnes en danger ('cash for crash').

[329] Page 14
[330] https://blogs.worldbank.org/endpovertyinsouthasia/how-do-poor-cope-road-crashes-india
[331] https://www.cairn.info/revue-d-economie-du-developpement-2006-4-page-5.htm
[332] https://www.laa.lv/wp-content/uploads/2020/01/Insurance-fraud-not-a-victimless-crime.pdf
[333] https://www.laa.lv/wp-content/uploads/2020/01/Insurance-fraud-not-a-victimless-crime.pdf

2. Description du phénomène en France et en Inde

a. En France

Par essence, la fraude est un phénomène difficile à quantifier. On peut citer les chiffres de l'ALFA (Agence pour la Lutte contre la Fraude à l'assurance), qui a calculé que la fraude identifiée en IARD s'élève à près d'un demi-milliard d'euros[334]. Seuls 300 millions d'euros sont effectivement détectés et récupérés[335]. Sachant que les prestations IARD en 2018 représentaient 39,1 milliards d'euros, cela signifie que la fraude identifiée représenterait environ 1,3% des règlements de sinistres[336]. Generali France détecterait 300 cas de fraudes sur 125 000 sinistres déclarés par an, soit 0,24%[337].
Pour les estimations de l'ensemble des fraudes, y compris non détectée, l'exercice est plus délicat. Insurance Europe estimait que 10 à 15% des prestations versées dans les pays occidentaux découlent de déclarations frauduleuses[338]. On peut noter que dans le cadre de l'étude du Le Lynx/CS, 11% des Français interrogés ont admis avoir fraudé leur assurance. [339] La population jeune est particulièrement concernée : 25% des jeunes de 18 à 35 ans de cette étude ont avoué avoir fraudé.

L'ALFA distingue dans l'assurance automobile deux types de fraudes[340] :
- la fraude de fréquence, qui joue sur la répétition de schémas de fraude avec des montants réduits. Les faux accidents en stationnement représentaient en 2017 la plus grande partie des dossiers frauduleux (48%) mais avec des impacts financiers relatifs (15% des enjeux financiers)
- la fraude d'intensité, avec des montants particulièrement élevés, notamment les sinistres corporels.

En termes de tendance, avant Covid, on pouvait noter une augmentation des cas[341] (+16% en 2009), notamment sur des sinistres portants sur de faibles montants. Les modes opératoires deviennent plus complexes, en parallèle de la professionnalisation des fraudeurs et du développement depuis les années 200 des fraudes en bandes organisées. Les fraudeurs exploitent aussi de plus en plus les failles des systèmes d'information, avec le développement du vol d'identité[342] : le fraudeur utilise des données personnelles pour usurper l'identité d'un assuré innocent et souscrire une police d'assurance en son nom.

[334] https://www.alfa.asso.fr/fraude-a-lassurance/
[335] https://www.journaldunet.com/solutions/dsi/1493997-lutte-contre-la-fraude-les-assureurs-misent-sur-la-technologie/
[336] https://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurances-de-biens-et-de-responsabilite-donnees-cles-par-annee
[337] https://www.journaldunet.com/solutions/dsi/1493997-lutte-contre-la-fraude-les-assureurs-misent-sur-la-technologie/
[338] https://www.atlas-mag.net/article/fraude-a-l-assurance-le-cout-et-la-lutte-contre-ce-fleau
[339] https://www.lci.fr/conso-argent/etude-fraude-a-l-assurance-11-des-francais-avouent-s-y-etre-adonne-quels-sont-les-risques-lelynx-fr-csa-2136887.html
[340] https://www.argusdelassurance.com/reglementation/jurisprudence/des-modes-operatoires-plus-sophistiques-qui-renvoient-a-une-professionnalisation-des-fraudeurs.117819
[341] https://www.argusdelassurance.com/dossier-ja/la-detection-de-la-fraude-en-vol-automobile-un-enjeu-important.48108
[342] https://www.tribune-assurance.fr/depeches/202006221623FOJ_GALL_468291-l-acpr-victime-d-une-usurpation-d-identite !

b. En Inde

Les chiffres manquent aussi pour évaluer l'ampleur du phénomène en Inde. La fourchette basse place la fraude à 6 % des demandes d'indemnisation[343]. Outlook India estimait que la fraude représentait en octobre 2019, 8,5% du chiffre d'affaires de l'assurance en Inde[344]. Dans le « Journal of Insurance Institute of India », Srinidhi Venkatesh évalue la fraude entre 8% ou 9% des primes du secteur de l'assurance en Inde[345]. Le General Insurance Council affirmait en 2017 qu'entre 10% à 12% des demandes d'indemnisations étaient suspectées d'être frauduleuses. Un article de Business Today[346] estime que la fraude à l'assurance concerne entre 10% et 15% de chaque ligne d'activité de l'assurance. Deux hauts responsables d'assureurs indiens, Ergo General Insurance et New India Chairman, s'accordent sur une estimation de 10% de demandes d'indemnisation frauduleuses dans un article du « Times of India »[347]

La grande majorité de ces fraudes seraient liées à des demandes financières supérieures à la réalité, en ajoutant des dommages, des blessures ou même des passagers fictifs[348]. Seuls 10% des fraudes seraient de faux accidents mis en scènes.

La fraude serait particulièrement importante dans les zones rurales ou semi-urbaines où les assureurs n'ont pas les infrastructures pour faire des inspections et où les autorités locales, qui sont censées attester de la véracité des évènements, pourraient être corrompues[349]. Les paiements en cash aux autorités locales dans des zones rurales seraient particulièrement répandus et admis dans les mœurs, même pour des services basiques, comme l'accès à l'eau ou aux sanitaires[350]. Dans de nombreux cas, des personnes décédées de causes naturelles sont enregistrées comme décédées à la suite d'un accident de la route. Des demandes d'indemnisations peuvent aussi être remplies au nom de personnes qui n'ont jamais existées[351].

[343] [343] Insurance Frauds Across the Globe: Lessons Learnt, Srinidhi Venkatesh, THE JOURNAL OF INSURANCE INSTITUTE OF INDIA, JULY - SEPTEMBER 2020, page 27

[344] https://outlookmoney.com/insurance/frauds-in-insurance-sector-3706

[345] Insurance Frauds Across the Globe: Lessons Learnt, Srinidhi Venkatesh, THE JOURNAL OF INSURANCE INSTITUTE OF INDIA, JULY - SEPTEMBER 2020, page 27

[346] https://www.businesstoday.in/opinion/columns/insurance-frauds-control-act-an-urgent-need-in-india-fraudulent-claims-indian-penal-code/story/400212.html

[347] https://timesofindia.indiatimes.com/business/india-business/1-in-10-insurance-claims-turns-out-to-be-a-fraud/articleshow/50034905.cms

[348] https://www.businesstoday.in/opinion/columns/insurance-frauds-control-act-an-urgent-need-in-india-fraudulent-claims-indian-penal-code/story/400212.html

[349] https://economictimes.indiatimes.com/news/economy/finance/how-insurance-firms-are-dealing-with-fraud-claims/articleshow/59448276.cms?from=mdr

[350] Deconstructing Corruption: A Study of Cash for Favours in Rural India, March 2016Journal of South Asian Development 11(1), DOI:10.1177/0973174115624773

[351] https://economictimes.indiatimes.com/news/economy/finance/how-insurance-firms-are-dealing-with-fraud-claims/articleshow/59448276.cms?from=mdr

3. La lutte contre la fraude au niveau des institutions

a. Cadre légal

i. En France

Sur la base du Code des Assurances et la jurisprudence, on peut lister les mesures qui peuvent être prises par l'assureur.

Dans le cas d'une fraude lors de la déclaration des risques, le code des assurances prévoit à l'article L. 113-8 que « le contrat d'assurance est nul en cas de réticences ou de fausses déclarations intentionnelles de l'assuré, lorsque celles-ci ont pour conséquence de changer l'objet du risque ou d'en diminuer l'opinion pour l'assureur. »[352]. Ces dispositions concernent la fausse déclaration du risque en lui-même. Cette règle **annule le contrat** et oblige les parties à restituer ce qu'elles ont reçu. L'assureur peut cependant « conserver les cotisations payées et exiger le règlement des cotisations échues, à titre de dommages et intérêts ».[353]
Si la mauvaise foi ne peut être établie, l'article L. 113-9 précise que « l'omission ou la déclaration inexacte de la part de l'assuré dont la mauvaise foi n'est pas établie – lorsqu'elle est constatée après sinistre – conduit à l'application de la **règle proportionnelle** »[354]. L'indemnité est alors réduite en proportion du taux des primes payées et de l'indemnité qui aurait été versée si le risque avait été correctement déclaré.

Dans le cas d'une fraude durant la vie du contrat, on distingue notamment la conduite sans permis et la non-déclaration de l'aggravation du risque.[355]
Dans le premier cas, **l'exclusion de garantie** pour absence de permis régulier est parfaitement licite (art. R.211-10, al. 1,1°) ; mais inopposable aux victimes ou à leurs ayants-droits (art. R.211-13, al. 1, 4°), de sorte que l'assureur est tenu de les indemniser pour le compte du responsable (art. R.211-13, al. 3).
Pour le second, on doit se référer à l'article Article L113-2 du Code des Assurances qui prévoit que l'assuré a obligation « de déclarer, en cours de contrat, les circonstances nouvelles qui ont pour conséquence soit d'aggraver les risques, soit d'en créer de nouveaux et rendent de ce fait inexactes ou caduques les réponses faites à l'assureur »[356]

Dans le cas d'une fraude lors de la déclaration du sinistre, l'assureur peut opposer **un refus de garantie ou prononcer une déchéance de garantie**. La déchéance d'une garantie et l'exclusion d'une garantie signifient tous deux que l'assureur n'indemnisera pas l'assuré sur son sinistre. Mais la déchéance est liée à un sinistre alors que l'exclusion de garantie peut se produire avant ou après un sinistre.
Afin de prendre cette mesure, l'assureur doit être en mesure de rassembler différents éléments :

[352] https://www.argusdelassurance.com/juriscope/analyses-d-experts/clause-de-decheance-une-affaire-de-mauvaise-foi.140690
[353] https://www.alfa.asso.fr/fraude-a-lassurance/
[354] https://www.argusdelassurance.com/juriscope/analyses-d-experts/clause-de-decheance-une-affaire-de-mauvaise-foi.140690
[355] Paragraphe "Types de fraude" précédent
[356]https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000035731302/#:~:text=L'assur%C3%A9%20est%20oblig%C3%A9%20%3A&text=Lorsqu'elle%20est%20pr%C3%A9vue%20par,lui%20a%20caus%C3%A9%20un%20pr%C3%A9judice.

- l'assuré doit avoir signé un document contenant une clause de déchéance et cette dernière doit être opposable à l'assuré[357].
- l'assureur doit prouver la mauvaise foi de l'assuré selon la jurisprudence.[358] L'article L. 113-1 du code des assurances précise que : « L'assureur ne répond pas des pertes et dommages provenant d'une faute intentionnelle ou dolosive de l'assuré », lesquelles ôtent au contrat d'assurance son caractère aléatoire.
- « la production de documents inexacts ou l'emploi de moyens frauduleux comme l'intention, par ces moyens, de faire supporter à l'assureur une indemnité indue » doivent aussi être documentés par l'assureur[359]

Enfin, si l'indemnité a été payée avant que l'assureur ne s'aperçoive de la fraude, celui-ci peut exercer une action en remboursement des sommes versées.[360] En assurance automobile, ce type de clause contractuelle ne sera valable que pour les garanties facultatives dites de « dommages » ; en effet, les exclusions de garantie en responsabilité civile automobile sont limitativement prévues par le code des assurances.

Enfin, les articles L113-2 et L113-4 du Code des Assurances permettent l'application de la surprime légale ou la résiliation des autres contrats du présumé fraudeur. [361]

Au niveau pénal, la fraude à l'assurance n'est pas définie comme telle dans le Code pénal mais l'assureur peut porter plainte contre le fraudeur selon les délits commis :
- L'escroquerie constitue un délit passible de 5 ans de prison et 375 000€ d'amende
- La production et l'usage de faux documents sont punis de 3 ans d'emprisonnement et de 45 000€ d'amende (article 313-1 du Code pénal)
- La conduite sans permis de conduire valable est punie d'une amende de 4500€ et d'une peine de prison de 2 ans. Le véhicule est confisqué. (Loi n°2011-267 du 14 mars 2011 - art. 70)

ii. En Inde

Au niveau du droit des contrats, les assureurs indiens peuvent s'appuyer sur 3 sections différentes du Indian Contract Act de 1872.

Dans la section 17[362], la fraude signifie et inclut tous les actes suivants perpétrés par une partie d'un contrat, ou avec ses connivences, ou par ses agents, avec intention de tromper l'autre partie ou son agent, ou l'induire pour le faire participer à un contrat.
(1) La suggestion, comme un fait, qui n'est pas vrai, par quelqu'un qui ne croit pas que ce soit vrai
(2) La dissimulation active de faits par un sachant ou convaincu de ce fait
(3) Une promesse faite sans intention de l'honorer
(4) Tous autres faits pour tromper
(5) Tous les types d'actes ou d'omission que la loi déclare spécifiquement frauduleux

Dans la section 18[363], la notion de fausse déclaration est définie comme :

[357] https://neraudau-avocats.fr/content/blog/déchéance-garantie-cass-civ-2ème-14-septembre-2017-pourvoi-n°16-21674
[358] https://www.argusdelassurance.com/juriscope/analyses-d-experts/clause-de-decheance-une-affaire-de-mauvaise-foi.140690
[359] https://www.radier-associes.fr/non-classe/exageration-des-dommages/
[360] https://www.alfa.asso.fr/fraude-a-lassurance/
[361] https://www.lefigaro.fr/assurance/2013/07/30/05005-20130730ARTFIG00365-auto-les-sanctions-en-cas-de-fraude-a-l-assurance.php
[362] https://www.taxmanagementindia.com/visitor/detail_act.asp?ID=7125
[363] (1) the positive assertion, in a manner not warranted by the information of the person making it, of that which is not true, though he believes it to be true.

(1) L'affirmation explicite, non garantie par des informations de la personne qui le fait, qui n'est pas vrai, même s'il croit que c'est vrai
(2) Tout manquement à un devoir qui, sans intention de tromperie, permet à la personne qui le commet ou toute autre personne réclamant en son nom de tirer profit en trompant quelqu'un
(3) Entrainant, même si cela est involontaire, une partie, en signant le contrat, à commettre une erreur sur la substance des choses sujet du contrat

La section 20 permet d'annuler le contrat en cas de fraude : « ou les deux parties d'un accord se trompent sur un sujet essentiel à l'accord, l'accord est nul »[364]

La demande d'indemnisation peut être rejetée devant les tribunaux si la fraude a été prouvée[365]. La fraude peut mener à l'annulation des polices d'assurance, mais pas dans le cas de fausse déclaration ou abus.

Au niveau pénal, le code civil indien contient 4 articles qui peuvent être utilisés en cas de fraude.
La section 205 concerne l'usurpation d'identité, qui peut être punie jusqu'à 3 ans d'emprisonnement et une amende. La section 420 concerne la tromperie et la transmission malhonnête de biens, qui est punie d'emprisonnement jusqu'à 7 ans ainsi qu'une amende. La section 464 porte sur la création de faux documents. La section 405 concerne les abus de confiance criminel.
Une tribune a même été rédigée en avril 2020 par S K Sethi[366], CEO of the Insurance Foundation of India, pour demander un cadre législatif spécifique et plus sévère. Par exemple les Etats Unis et le Royaume Unis ont classés la fraude à l'assurance comme un crime[367].

(2) any breach of duty which, without an intent to deceive, gains an advantage to the person committing it, or anyone claiming under him, by misleading another to his prejudice or to the prejudice of anyone claiming under him.
(3) causing, however innocently, a party to an agreement to make a mistake as to the substance of the thing which is the subject of the agreement.

[364] https://www.aaptaxlaw.com/contract-act/section-20-indian-contract-act-agreement-void-where-both-parties-are-under-mistake-as-to-matter-of-fact-section-20-of-indian-contract-act-1872.html
[365] https://blog.ipleaders.in/false-insurance-claims/
[366] https://www.businesstoday.in/opinion/columns/insurance-frauds-control-act-an-urgent-need-in-india-fraudulent-claims-indian-penal-code/story/400212.html
[367] Fraud Act de 2006 au Royaume-Uni

b. Cadre règlementaire

i. En France

Solvency 2 a contribué à donner aux assureurs un cadre de lutte contre la fraude : la fraude étant considérée comme un risque opérationnel ; à ce titre, elle doit être inventoriée, et disposer d'un provisionnement adéquat et de mesures destinées à réduire ce risque. Solvency 2 a aussi défini une fonction clé de gestion des risques. [368]

La lutte contre le blanchiment des capitaux et le financement du terrorisme a aussi fait progresser les dispositifs de lutte contre la fraude, même si les produits d'assurance non-vie sont considérés comme un risque « faible »[369]. Les assureurs doivent se conformer aux obligations de LCB-FT[370] : ils doivent identifier, connaître leurs clients et avoir une vigilance constante[371].

ii. En Inde

Le régulateur de l'assurance en Inde, l'IRDAI (Insurance Regulatory and Development Authority of India) a prévu plusieurs dispositifs pour contrer la fraude.

Depuis 2013, chaque assureur opérant en Inde doit **disposer obligatoirement d'un cadre de surveillance de la fraude**. Cela signifie des mesures pour détecter, prévenir, protéger et diminuer le risque de fraude des assurés ou demandeurs, employés ou intermédiaires de l'assurance. Cette fonction doit être indépendante afin d'assurer l'efficacité de la mise en place des politiques anti-fraude. Elle est responsable de la création d'un plan de reporting interne des différentes entités du groupe, de l'amélioration des compétences des intermédiaires, des assurés et employés pour identifier et prévenir les fraudes[372].

La mise à disposition d'informations en ligne par le régulateur facilite le travail des investigateurs[373]. En effet, **le registre national des véhicules est accessible en ligne** par les banquiers et les assureurs, sur abonnement payant. Vahan répertorie l'ensemble des 291 millions de véhicules immatriculés en Inde. Les décisions des **cours de justice sont accessibles sur Internet**, permettant des recherches par immatriculation[374].

Les gouvernements régionaux peuvent aussi être mobilisés par les tribunaux : Ainsi la haute cour du Rajasthan a demandé au gouvernement régional de constituer un groupe d'investigation dédié, un SIT (Special Investigation Team)[375], au vu de l'augmentation des cas de fraudes présentés devant les tribunaux, notamment les fraudes à l'assurance automobile.

---

[368] https://acpr.banque-france.fr/autoriser/procedures-secteur-assurance/gouvernance-en-assurance/dirigeants-et-responsables-de-fonctions-cles

[369] https://acpr.banque-france.fr/sites/default/files/medias/documents/20191218_asr_lcbft.pdf#page=2

[370] Lutte contre le Blanchiment et le Financement du Terrorisme

[371] https://www.ffa-assurance.fr/infos-assures/lutte-contre-le-blanchiment-des-capitaux-et-le-financement-du-terrorisme-les

[372] The Evolution of Fraud in Insurance Industry, Vinta Aggarwal, Saurabh Mishra, file:///C:/Users/ceciler/Desktop/References/Motor%20fraud%20in%20India/July-Sept%202020_Final.pdf

[373] Insurance and Fraud Management: Trends, Opportunities and Challenges 2020, Dr. C. K. Buttan, Sanjiv Dwivedi, THE JOURNAL OF INSURANCE INSTITUTE OF INDIA, JULY - SEPTEMBER 2020, page37

[374] https://vcourts.gov.in/virtualcourt/

[375] https://www.hindustantimes.com/jaipur/sit-to-look-into-fake-motor-insurance-claims-hc/story-agSp2EHzcLpj5lcoA1oqWI.html

Le régulateur a autorisé la création d'un compte eIA[376] ou Compte d'assurance électronique : un individu peut gérer l'ensemble de ses polices d'assurance disponible à tout moment et partout dans le monde. L'objectif est de protéger contre le vol ou la perte de produits d'assurance ou la fausse souscription, étant donné qu'un seul compte eiA n'est autorisé par personne, après vérification d'identité. Ce dispositif n'a cependant pas fait l'objet de publicité par le régulateur ou même les assureurs eux même, et reste peu utilisé.

L'IRDAI propose depuis 2019 le dispositif de Sandbox[377] ou bac à sable pour permettre aux assureurs de développer des produits innovants, comme les polices d'assurance Pay as you go, qui leur permettent de suivre leurs clients en temps réels et limiter ainsi la fraude.

c. Réponses collectives de la profession

i. En France

Les assureurs français se sont dotés de différentes instances pour unir leurs efforts et partager des informations dans la lutte contre la fraude.

Dès 1989, a été créée l'**ALFA**, l'Agence pour la Lutte contre la Fraude, une association loi de 1901 qui regroupe actuellement 320 adhérents. Elle intervient à 3 niveaux différents :
- sur la prévention de la fraude, en mettant à disposition des formations[378], une méthodologie, des notes d'information ;
- sur la détection de la fraude, en facilitant les échanges entre assureurs ;
- sur le traitement de la fraude, en assurant la liaison avec les officiers de polices judiciaires et grâce à un appui juridique et conformité

L'ALFA permet des échanges avec les services publics et les administrations, dont le ministère de l'intérieur. Il existe aussi des initiatives plus spécifiques.

Pour lutter contre la **fraude lors de la souscription**, **un fichier listant les résiliations des assurances automobiles et leurs causes** est consolidé par L'AGIRA ou Association pour la Gestion des Informations sur le Risque en Assurance[379] Concrètement, à chaque résiliation d'un contrat d'assurance automobile, l'assureur adhérent communique à l'AGIRA les motifs et date de résiliation, ainsi que les sinistres au cours des 5 années qui ont précédé la résiliation. Ces données sont accessibles pour les assureurs membres de l'AGIRA 2 ans sauf en cas de résiliation après un sinistre (délai de 5 ans) ou résiliation pour non-paiement des cotisations (jusqu'à paiement des sommes dues). Ce fichier permet de lutter contre les fraudes lors de la souscription.[380]

Afin de **mieux détecter la fraude lors des sinistres**, les assureurs peuvent s'appuyer aussi sur deux autres organismes.

[376] Insurance Frauds: A Closer Look, Ankit Jain, THE JOURNAL OF INSURANCE INSTITUTE OF INDIA, JULY - SEPTEMBER 2020, page 48
[377] Insurance Frauds: A Closer Look, Ankit Jain, THE JOURNAL OF INSURANCE INSTITUTE OF INDIA, JULY - SEPTEMBER 2020, page 50
[378] https://www.argusdelassurance.com/les-assureurs/fraude-a-l-assurance-homogeneiser-les-pratiques-est-un-gage-d-efficacite.154634
[379] www.agira.asso.fr/content/resiliations-automobile
[380] A noter que l'AGIRA propose aussi Fichier des Véhicules assurés, partagés depuis le 1er janvier 2019 avec les forces de l'ordre et avec le Fonds de Garantie des Assurances Obligatoires de dommages. L'objectif est de lutter contre la conduite sans assurance.

L'ARGOS est un organisme professionnel de l'assurance à but non lucratif, fondé en 1984, dont l'objectif est d'aider à la recherche, l'identification et la récupération « des véhicules et autres biens mobiliers déclarés volés. »[381] Il est le garant d'une base de données automobiles consolidant les informations provenant des assureurs, pouvoirs publics, constructeurs, loueurs, sociétés de gravage[382] et fourrières. L'ARGOS dispose aussi d'un pôle d'identification et de gestion des rapatriements des véhicules volés et de techniciens sur le terrain.
L'ANEA ou Alliance Nationale des Experts en Automobiles met aussi à disposition des experts, des formations et un process pour prouver la fraude[383]. Une commission Fraude veille et enrichit ces connaissances régulièrement[384].

La lutte antifraude au niveau de la profession passe donc par des bases de données informatiques, des échanges d'informations, le renseignement, la formation des enquêteurs, la sensibilisation des consommateurs et de l'opinion.

ii. En Inde

Les efforts de place en Inde ne sont pas aussi variés et les informations partagées entre assureurs restent limitées. Une initiative forte est toutefois en cours de développement par le General Insurance Council : il souhaite mettre en place une base de données regroupant les cas de fraude avérées et les cas suspicieux. Une première phase de ce projet a été menée à bien en 2016[385] avec l'aide de LexisNexis Risk Solutions : en collectant 2 ans de déclaration de sinistres auprès de 10 assureurs avec 10 millions de polices d'assurance santé et 8 millions de police d'assurance automobile. L'étude a essayé d'identifier les cases de fraude sur la base de 100 éléments. La prochaine étape est la création d'un référentiel des cas de fraude partagé. Ce projet était fin mai 2020 à l'étape d'appel d'offres[386], mais en mai 2023 il n'est toujours pas listé dans les projets en cours de mise en œuvre[387].

d. Réponse au sein de chaque assureur

La réponse à la fraude se fait aussi au sein de chaque assureur, par la mise en place d'une **structure et de ressources dédiées**.

L'objectif d'un service antifraude est de « prévenir, détecter, limiter ou stopper les fraudes, si possible avant de décaisser »[388] Il doit mener aussi une veille sur les pratiques.
Il prend en charge les sinistres identifiés comme douteux et lance des investigations pour confirmer ou infirmer les soupçons. Les employés peuvent être amenés à faire appel à des agences de détectives spécialisés.
La détection en amont peut se faire par des règles métiers, des gestionnaires et/ou des outils informatiques. Etant donnée l'augmentation des cas ces dernières années, l'industrialisation de sa détection et son intégration dans la digitalisation des process est un enjeu fort pour les assureurs.

[381] https://www.gieargos.org/qui-sommes-nous/notre-gie
[382] Gravage = marquage des vitres antivol
[383] http://anea.fr/ifor2a/actualites-ifor2a-37.html
[384] http://anea.fr/actualites-451.html
[385] https://www.icicilombard.com/blog/home-insurance/hoi/general-insurers-collaborate-to-fight-fraud
[386] https://www.gicouncil.in/rfp-and-compliance-matrix/fraud-analytics-rfp-for-data-repository/
[387] https://www.gicouncil.in/projects/frmp/
[388] https://www.generali.fr/dossier/fraude-assurance-consequences/

Afin de répondre à une fraude multiforme et toujours innovante, un dispositif anti-fraude ne **doit pas se limiter à une structure autonome mais doit diffuser ses pratiques**[389] au sein de l'organisation. La fraude doit faire l'objet d'une attention particulière au sein des différents services de l'assurance :

- l'informatique : avec des dispositifs de détection et de suivi
- la communication interne et externe afin de développer une culture anti-fraude[390]
- le service juridique : pour sécuriser les contrats et s'assurer un cadre juridique aux poursuites, aussi bien interne qu'externe
- l'organisation : qui doit s'assurer des procédures, des contrôles et du dispositif de surveillance
- De même la prévention de la fraude ne doit pas se limiter à la gestion des sinistres, mais doit être intégrée dès la souscription, par exemple avec l'analyse des pièces justificatives.

e. Les limites de la lutte contre la fraude

Que ce soit en France ou en Inde, la lutte contre la fraude rencontre de nombreux facteurs limitants.

Les investigations doivent tout d'abord **respecter le droit à la vie privée**, en France notamment. Si l'assureur doit prouver la mauvaise intention de son client, il ne peut le faire que dans un cadre limité par des règles. Ainsi « la Cour de cassation rappelle (…) que ces mesures de détection et d'investigation ne doivent pas présenter un caractère disproportionné au regard du but poursuivi »[391]

En termes de droit d'accès et d'utilisation des données des utilisateurs, les assureurs français doivent prendre en compte notamment la réglementation RGPD. Ils peuvent s'appuyer cependant sur le Pack Assurance défini par la CNIL sur la lutte contre la Fraude.[392]

En Inde, le premier facteur limitant est le **coût de la détection et des investigations**, au vu des recouvrements très réduits. L'autre problématique dans un marché de l'assurance en pleine croissance est la pression pour réduire les délais de paiement, afin de gagner des parts de marchés. Les investigations ralentissent le process de règlement et peuvent avoir un impact fort en termes d'image et mécontenter les réseaux de distribution.

[389] https://www.optimind.com/medias/documents/295/Optimind-Winter-Publication-Fraude-assurance-avril-2015.pdf
[390] Combating Insurance Frauds Through Analytics. Prof. Venkatesh Ganapathy. THE JOURNAL OF INSURANCE INSTITUTE OF INDIA. JULY - SEPTEMBER 2020 page 41
[391] https://www.argusdelassurance.com/juriscope/jurisprudence/le-terrain-de-jeu-privilegie-des-fraudeurs.85924
[392] https://www.cnil.fr/sites/default/files/typo/document/FICHE_05_assurance.pdf

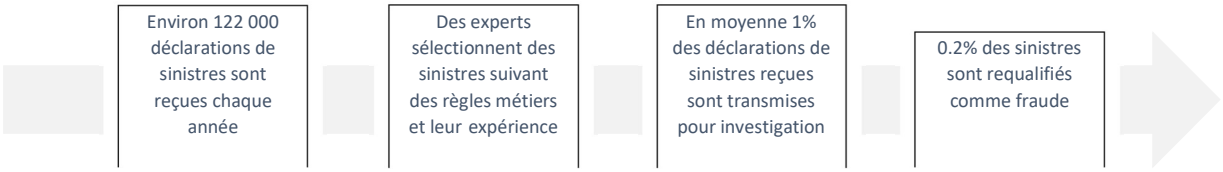
4. Le dispositif initial de lutte contre la fraude au sein de Generali Inde

"Detecting fraud is art, not science".

R. Goldfarb[393]

a. Vue générale du dispositif

Dans le processus actuel, toutes les déclarations de sinistres automobiles ne peuvent faire l'objet d'une analyse manuelle pour détecter la fraude, car leur volumétrie est trop importante (plus de 122 000 dossiers par an) : le coût et les délais seraient trop élevés. Actuellement les experts se basent sur des **règles métiers et leur expérience** pour sélectionner des dossiers, les analyser et si nécessaire, les transférer à des enquêteurs privés.

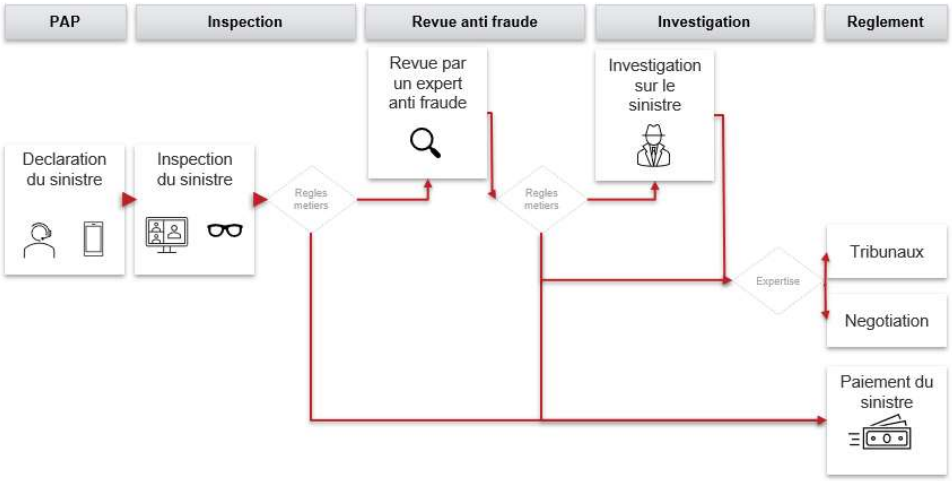


b. Processus de gestion du sinistre et lutte anti-fraude

Un sinistre peut passer par 4 étapes différentes au sein de FGII (Future Generali India Insurance Company) :

- Premier avis de perte (PAP) : l'assuré prévient du sinistre soit par un appel soit par une déclaration en ligne
- Inspection : chaque sinistre doit fait l'objet d'une inspection physique ou en vidéo
- Revue anti-fraude : en fonction de règles métiers, des experts revoient des dossiers pour estimer s'ils sont potentiellement frauduleux
- Investigation : En cas de suspicion de fraude, une investigation est commanditée auprès d'inspecteurs privés
- Règlement : une déclaration de sinistre peut ensuite déboucher sur un paiement, une négociation ou encore une démarche auprès d'un tribunal.

[393] *"Bernie Madoff and the Honey Trap", foodliabilitylaw.com, 2013*



c. Règles métiers au début de l'étude

En France, les indicateurs classiques de fraude sont :

- des paiements en cash
- une demande d'indemnisation moins d'un mois après le début de couverture du contrat
- le véhicule était garé dans la rue alors qu'une protection comme un garage privé et/ou sécurisé était disponible.
- imprécisions ou contradiction des déclarations de l'assuré sur les circonstances de l'accident
- le nombre de sinistres déclarés par l'assuré
- la souscription d'assurances multiples
- la répétition de sinistres similaires
- un sinistre parfaitement documenté et préparé

La co-entreprise de Generali en Inde utilise une liste de 25 facteurs déclenchant une analyse par un expert anti-fraude.

N.	Elément déclenchant une revue dans la lutte anti-fraude	Etape dans la gestion du sinistre
1	Signatures différentes	1er avis de perte
2	Délai dans la déclaration du sinistre	1er avis de perte
3	Proximité entre la date de survenance du sinistre et le début de la police d'assurance	1er avis de perte
4	Rupture dans le contrat	1er avis de perte
5	Conducteur différent du propriétaire du véhicule et sans relation entre les deux	1er avis de perte
6	Documents originaux non disponibles pour vérification	
7	Lors de l'inspection, les dommages semblent anciens et/ou multiples	Inspection du véhicule
8	Causes du dommage non cohérents avec les dommages constatés	Inspection du véhicule
9	Perte totale du véhicule	1er avis de perte
10	Atelier de réparation sous surveillance	

11	Dommages mécaniques qui semblent frauduleux	Inspection du véhicule
12	Date et heure de l'accident	1er avis de perte
13	Parc de véhicules professionnels	1er avis de perte
14	Véhicule en stationnement et des dommages supérieurs à 20% de la valeur du véhicule	1er avis de perte
15	Nombre de kilomètre du véhicule élevé étant donne l'âge du véhicule	Inspection du véhicule
16	Autocollant sur le véhicule accidenté pour des calls centers, compagnies de taxis, auto-école	Inspection du véhicule
17	Absence de preuve du trajet comme des reçus, carnets de route	
18	Présence de sang, de fluides corporels ou de cheveux sur le pare-brise	Inspection du véhicule
19	Incendie du véhicule	1er avis de perte
20	Absence de certificat médicaux	
21	Nom du conducteur non confirmé dans le rapport de police	
22	Lieu de l'accident par rapport au lieu de résidence/travail de l'assure	1er avis de perte
23	Permis de conduit expiré	1er avis de perte
24	Cas de vol	1er avis de perte
25	Proximité de la date de vol et de l'émission du permis de conduire	1er avis de perte

3ème partie : Présentation des modèles d'apprentissage statistique

1. Définition de l'apprentissage statistique

a. Historique et définitions

Qu'est-ce que l'apprentissage statistique ? En quoi cette approche diffère des méthodes de statistique classique ?

La **statistique** peut se définir comme « une méthode mathématique d'analyse qui permet d'étudier avec le maximum de précision les phénomènes incomplètement connus »[394], selon les propos de J. Torrens-Ibern de 1956**.**[395] Historiquement la statistique a développé « des modèles incorporant d'une part des variables explicatives ou prédictives et, d'autre part, une composante aléatoire ou bruit »[396]. L'objectif est **d'estimer les paramètres du modèle en se basant sur les observations, tout en vérifiant le comportement de la partie aléatoire**.

Les techniques de Machine Learning ou apprentissage automatique en français sont plus récentes. La première définition, et la plus couramment admise, est celle d'Arthur Samuel en 1959 "the field of study that gives computers the ability to learn without being explicitly programmed".[397] Plus précisément dans le Cambridge Dictionary, il est défini comme « **un processus d'ordinateurs qui adaptent leur façon de réaliser une tache en apprenant de nouvelles données, sans qu'un être humain n'ait besoin de lui donner d'instruction sous forme d'un programme** »[398]

L'objectif est le même que pour la statistique, à savoir décrire un phénomène inconnu en se basant sur des observations. **Dans les deux cas, on cherche aussi à construire un modèle, qui servira à réaliser des prédictions. Mais, les méthodes comme le vocabulaire diffèrent**.

La statistique se base sur des modèles paramétriques, alors que le machine Learning part des données et construit des modèles non paramétriques. La première étudie les propriétés de la variable aléatoire à prévoir, tant que le second s'intéresse plus aux propriétés du modèle optimal, selon l'article de A. Charpentier, E. Flachaire et A. Ly, « Économétrie & Machine Learning » [2017]. La statistique se concentre sur des modèles, alors que l'apprentissage machine est centré sur les algorithmes.

L'apprentissage statistique est défini par Fabrice Rossi en 2008 comme la somme de l'apprentissage automatique et de la statistique.[399] Cette notion est plus récente, car elle a été introduite par Vapnik[400] et popularisée par Hastie dans son ouvrage « The Elements of Statistical Learning »[401] en 2007. La statistique mathématique et l'apprentissage machine sont « deux cultures »[402], qui ont

[394] J. TORRENS-IBERN Qu'est-ce que la statistique ? Journal de la société statistique de Paris, tome 97 (1956), p. 289-293 <http://www.numdam.org/item?id=JSFS_1956__97__289_0>
[395] Il rappelle aussi " la boutade d'un homme d'État anglais qui disait que la statistique est une des formes les plus raffinées que l'homme a inventé pour mentir »
[396] https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-Intro-ApprentStat.pdf
[397] Plus précisement, 'Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort." "Some studies in machine learning using the game of checkers", In Feigenbaum, E. and Feldman, J. Eds., Computers and Thought. McGraw-Hill, New York, 1963
[398] "the process of computers changing the way they carry out tasks by learning from new data, without a human being needing to give instructions in the form of a program" https://dictionary.cambridge.org/dictionary/english/machine-learning
[399] Introduction à l'apprentissage statistique, Fabrice Rossi, Janvier 2008, page 8
[400] Vapnik, Vladimir Naumovič (1936-) The Nature of Statistical learning theory. N.Y : Springer, 1995.314 p.
[401] Disponible gratuitement ! https://web.stanford.edu/~hastie/Papers/ESLII.pdf
[402] Statistical Modeling: The Two Cultures Leo Breiman, Statistical Science 2001, Vol. 16, No. 3, 199–231

grandi parallèlement. Mais selon A. Charpentier, « les données massives » ont imposé « de créer des passerelles entre les deux approches »[403].

De l'apprentissage automatique sont réalisées une observation d'un phénomène, la construction d'un modèle, la réalisation de prévision et d'analyse grâce à ce modèle, le tout automatiquement. Mais la statistique apporte selon Fabrice Rossi[404], la formalisation du processus et une attention à la qualité des résultats.

b. Enjeux

Qu'est-ce que peux apporter l'apprentissage statistique ?

Par rapport aux méthodes de statistiques traditionnelles, elle ne nécessite **pas d'hypothèse sur la structure ou la distribution des données**. La seule hypothèse qui est faite est que les données à prédire sont générées de façon identiques et indépendantes par un processus, à partir de variables explicatives.

Elle permet aussi **d'exploiter les nombreuses données disponibles**, en se basant sur les capacités machine qui ont fortement augmenté ces dernières années.

L'apprentissage statistique permet aussi de détecter des **phénomènes et des structures de dépendance complexes**, que des outils de régression classique ne peuvent détecter. Le succès de ces algorithmes est particulièrement visible dans le domaine de la génétique, ou de l'écologie, par exemple.[405] En particulier, les anomalies sont assez mal détectées par les modèles statistiques classiques.[406]

Un autre avantage notable de l'apprentissage statistique est sa capacité à **s'adapter et à détecter de nouveaux schémas.**

Cependant tous les algorithmes de machine Learning ne sont **pas interprétables**, l'impact de chaque variable n'est pas toujours évalué. On parle notamment d'effet boite noire de certains algorithmes.

Dans le cas de la fraude à l'assurance, l'apprentissage statistique présente donc de nombreux avantages. Le processus de détection de la fraude de FGII est actuellement basé sur des règles métiers. Cela nécessite beaucoup d'analyses réalisées par des agents.

L'apprentissage statistique permettrait une **détection plus rapide**, **avec plus de précision**. En réduisant les faux positifs[407], on permet aussi de diminuer notablement le coût du processus. L'apprentissage statistique permettrait aussi de s'adapter plus rapidement à des schémas de fraude complexes et qui peuvent évoluer très vite. Enfin il serait plus en mesure de détecter des fraudes qui n'ont jamais été détectées par les experts.

Cependant, les algorithmes doivent prendre aussi en compte les contraintes opérationnelles de l'assurance : les modèles doivent être, dans une certaine mesure, interprétables, transparents et

[403] A. Charpentier, E. Flachaire et A. Ly, « Économétrie & Machine Learning » [2017]

[404] Introduction à l'apprentissage statistique, Fabrice Rossi, Janvier 2008, page 8

[405] Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique, Antoine Paglia, Martial V. Phelippe-Guinvarc'h, Euro-Institut d'Actuariat – Groupama, page 50

[406] http://www.galea-associes.eu/2018/05/traquer-la-fraude-grace-aux-methodes-data-science/

[407] Faux positifs : fausses alertes = cas identifié comme suspect mais qui se révèle non frauduleux

auditables. Les cas de fraude suspectée doivent pouvoir être transmis aux enquêteurs avec les éléments concrets qui le rendent suspectes.

2. Présentation des modèles d'apprentissage

a. Introduction

Classiquement on sépare les modèles d'apprentissage en deux catégories : **supervisés** et non supervisés. Dans le premier cas, **on dispose de ce que devraient être les valeurs de sorties pour notre échantillon.** L'objectif d'un apprentissage supervisé est donc d'apprendre de ces données et de reproduire au mieux la relation entre valeurs d'entrée et sortie disponible sur une partie au moins de l'échantillon. Autrement dit, la base de données peut être considérée comme un ensemble de valeurs indépendantes $\begin{cases} \{X_i, Y_i\} \\ 1 \leq i \leq n \end{cases}$ d'un vecteur aléatoire (X, Y), ou X est le vecteur des variables explicatives et Y la variable réponse. L'objectif du modèle est de construire, à partir des données disponibles, une règle qui pour toute nouvelle observation X_{n+1} associe la réponse Y_{n+1}

A l'inverse, **l'apprentissage non supervisé** n'utilise pas de données étiquetées, son objectif est de **modéliser les observations pour mieux les comprendre**. Donc l'analyse porte seulement sur les observations X_i que l'on souhaite discriminer ou trier.

Au sein des modèles **supervisés**, on peut distinguer deux types de modèles différents, selon la nature de la variable à prédire. On parle de **classification quand la variable à prédire est une catégorie. Si cette valeur est numérique, on parlera alors de régression.** Dans le cadre de ce mémoire, nous étudierons les modèles supervisés suivants :

- régression logistique,

- arbre de décision,

- forêt aléatoire,

- Naïve Bayes,

Les **modèles non supervisés** peuvent se repartir en deux catégories : **le regroupement, appelé aussi partitionnement (*clustering* en anglais) et l'association.**

La logique du premier est de séparer un ensemble de données en différents groupes qui présentent des traits similaires. Dans le cadre de ce mémoire, nous effectuerons un K-mean clustering sur nos données.

Les modèles d'association consistent à découvrir des relations entre les attributs de ces données. (Comme les Réseaux de Neurones)

Deux autres types d'apprentissages peuvent être mentionnés, mais ne seront pas développés : les modèles d'apprentissage semi-supervisés et par renforcement. Les modèles semi-supervisés

concernent des jeux de données partiellement étiquetés. L'apprentissage par renforcement intègre une récompense, positive si l'action est un bon choix, négative en cas d'erreur.

Pour présenter les différents modèles possibles, on peut donc se baser sur les données disponibles, comme le fait la distinction entre modèles supervisés et non supervisés. Il est aussi possible de partir de la notion à la base du modèle, comme décrit par la segmentation créée par John D. Kelleher, Brian Mac Namee, Aoife D'Arcy dans leur livre "*Fundamentals of Machine Learning for Predictive Data Analytics : Algorithms, Worked Examples, and Case Studies*".

Ils distinguent 4 types de modèles :

- Les modèles basés sur l'information. Dans notre mémoire, nous étudierons notamment les arbres de décision et les forêts aléatoires
- Les modèles basés sur les similarités. Les K-mean clustering seront en particulier décrits dans ce mémoire.
- Les modèles basés sur les probabilités. Un modèle Naïf Bayes sera utilisé.
- Les modèles basés sur la notion d'erreur. La régression logistique en est un exemple.

b. Les arbres de décision

Les arbres de décision sont des modèles basés sur l'information : ils cherchent **à déterminer quelle(s) caractéristique(s) apportent le plus d'informations sur la donnée à prédire**. Ensuite, il suffit de tester ces caractéristiques sur toutes les nouvelles observations pour pouvoir prédire leurs valeurs ou leurs catégories.

i. Définitions et Principe

Les arbres de décision classent les observations en se basant sur la valeur de chacune de leurs caractéristiques. L'objectif est **d'obtenir des groupes d'observations les plus homogènes possibles pour effectuer une prédiction.**

Les arbres de décision peuvent être définis comme une « représentation hiérarchique de la structure des données sous forme de séquences de décision (tests) en vue de la prédiction d'un résultat ou d'une classe »[408]. Ils ont été créés dans les années 60 et de nombreuses approches ont été développées depuis. L'algorithme CART (pour Classification And Regression Tree), définie par Breiman en 1984, fait encore référence aujourd'hui et sera utilisée pour nos travaux. Cet algorithme peut aussi

[408] Apprentissage, réseaux de neurones et modèles graphiques (RCP209) Arbres de décision, Marin FERECATU, http://cedric.cnam.fr/vertigo/Cours/ml2/docs/coursArbresDecision-1x2.pdf

être interprétée comme une partition de l'espace des données en sous régions homogènes en termes de classe.[409]

On peut distinguer 2 différents types d'arbre de décision, selon la nature de la variable à prédire.

- Si c'est une classe, on parlera d'un arbre de classification.

- En cas de donnée à prédire d'ordre numérique, il sera appelé arbre de régression.

ii. Construction d'un arbre de décision

Un arbre de décision se construit par division successive de la base d'apprentissage. Il consiste d'un nœud principal, de nœuds intérieurs et des nœuds finaux, des feuilles, qui sont connectées par des branches.

- Chaque nœud spécifie un test qui doit être mené sur une caractéristique. Le nombre de valeurs possibles que cette caractéristique peut prendre, détermine le nombre de branches sous-jacentes. La division peut être binaire (2 sous-ensembles sont créés) ou multiples (autant de partition que de valeurs distinctes). Le test effectué à chaque nœud est unique.

- Chaque branche correspond donc à une modalité de l'attribut.

- Une feuille correspond à un nœud terminal. Les feuilles sont les valeurs de la variable cible ou le résultat du test (par exemple si le test est >=x ou <x. Le parcours d'un chemin qui va, de la racine à une feuille, est une combinaison de modalité d'attribut menant à une valeur cible. L'objectif est donc d'attribuer à chaque feuille une modalité à prédire.

Il s'arrête quand les éléments d'un nœud ont la même variable cible (et sont donc homogènes) ou en fonction d'un critère d'arrêt prédéfini. (Par exemple, pas de séparateur si moins de x observations).

Lors de la construction de l'arbre, quatre questions se posent qui vont le structurer :

1. **Comment sélectionner la meilleure division possible parmi toutes celles possibles ? Quel critère pourrait permettre de déterminer la meilleure division ?**
2. **Est-ce que l'arbre doit être arrêté avant que chaque feuille ne contienne plus qu'un élément ? Quelle règle mettre en place pour décider d'arrêter l'arbre ?**
3. **Si on décide d'arrêter l'arbre, comment attribuer les observations quand une feuille contient différentes modalités de la donnée à prédire ?**
4. **Comment estimer l'erreur de prédiction ?**

La deuxième question est la plus délicate à résoudre.

[409] https://people.minesparis.psl.eu/fabien.moutarde/ES_MachineLearning/Slides/coursFM_AD-RF.pdf

- Critère de séparation d'un arbre de classification

Il existe différentes méthodes pour hiérarchiser les caractéristiques d'une observation en termes d'apport d'information sur la donnée à prédire. R. Rakotomalala précise que la « quasi-totalité de […] expérimentations ont montré que, dès lors que les mesures utilisées possèdent de bonne qualité de spécialisation, c'est-à-dire tendent à mettre en avant les partitions avec des feuilles pures, elles ne jouent pas un rôle majeur dans la qualité de la prédiction »[410]

Dans l'algorithme CART[411], que nous allons utiliser, les divisions sont binaires et utilisent comme critère de séparation, l'indice d'impureté de Gini. Cette fonction mesure « le degré de désordre dans les différentes classes »[412] en mesurant la pureté du nœud.

Cet indice mesure l'inégalité de répartition de valeurs distinctes (classe de la valeur cible) au sein d'une distribution (modalités de notre attribut). Cela mesure à quelle fréquence des observations choisies aléatoirement seraient mal étiquetées si elles l'étaient suivant la distribution des données.

Un nœud est pur si le segment qui lui est associé ne contient que des descriptions d'éléments d'une même classe. L'objectif est de réduire l'impureté des nœuds donc choisir les nœuds dont le découpage apporte en moyenne les nœuds les plus purs possibles.

$$Gini(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} \times \left[1 - \left(\left(\frac{p_k}{p_k+n_k}\right)^2 + \left(\frac{n_k}{p_k+n_k}\right)^2\right)\right]$$

Avec $p_k =$ nombre de valeurs OUI dans le nœud *k* et $n_k =$ nombre de valeurs NON dans le nœud *k*

Si on prend un arbre bimodal, à chaque nœud, le nombre de NON est égal au nombre total du nœud moins le nombre de oui. Si on a un seul nœud, on peut écrire :

$$n_1 + p_1 = (n + p)$$

Et alors Gini(A) = $1 - \left(\left(\frac{p_1}{p_1+n_1}\right)^2 + \left(\frac{n_1}{p_1+n_1}\right)^2\right)$

Le critère de Gini prend ses valeurs dans [0 ;1]. Il s'annule quand les différents types d'éléments sont chacun dans des classes différentes. On cherche donc à avoir un critère de Gini le plus bas possible, signe de la plus forte homogénéité.

D'autre indicateurs peuvent être utilisés, comme la fonction d'entropie dans les arbres C4.5 et C5.

[410] « Arbres de décision » Ricco RAKOTOMALALA, UNIVERSITE Lumière Lyon 2, https://www.rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-rakotomalala-33/rakotomalala-33-tutorial.pdf, page 168

[411] Classification and Regression Tree, Breiman et col. 1984

[412] Apprentissage Artificiel et fouille de données Arbres de décision Jamal Atif, Université Paris Dauphine D'après Céline Hudelot (ECP), d'après Tan, Steinbach, Kumar M2R ISI Université Paris-Dauphine 2015-2016

- Elagage d'un arbre de décision (pruning)

Un arbre de décision peut être de grande taille avec beaucoup de feuilles qui contiennent juste un élément. Si un arbre a trop de nœuds, il devient complexe et risque d'être trop adapté à l'échantillon et donc manquer de capacité de généralisation. Il est alors nécessaire d'élaguer l'arbre construit pour garder un pouvoir de généralisation. (Cf. le paragraphe 3. Evaluation des résultats des modèles, a. Capacité de généralisation d'un modèle, entre biais et variance)

Pour déterminer à quel niveau l'élagage doit être fait, il est possible de calculer pour différentes tailles d'arbre le taux d'erreur de prédiction par validation croisée : on arrête l'arbre au niveau de la taille qui permet de minimiser cette erreur. L'objectif est de d'utiliser la profondeur optimale qui équilibre au mieux variance et biais. Si l'arbre est profond, le biais est faible mais la variance est forte. A l'inverse, un arbre court entraine un biais fort, mais une faible variance. Les données de validation sont utilisées pour déterminer le niveau de feuille optimum.

- Règle d'attribution des observations en cas de feuille avec plusieurs modalités

Lorsque plusieurs modalités de la variable à prédire sont présentes dans une feuille, une règle d'attribution doit être définie pour arbitrer quelle modalité sera retenue.

Dans le cas des classes, c'est le plus souvent la règle de la majorité qui est retenue, c'est-à-dire qu'on attribue à la feuille la modalité à prédire qui est la plus nombreuse.

- Estimation de l'erreur de prévision

L'erreur de prévision est évaluée par la probabilité qu'une observation soit mal classée. Il existe plusieurs estimateurs possibles selon les données utilisées pour le calculer. Si on utilise les données d'apprentissage par re-substitution, l'estimateur est peu représentatif. On peut soit utiliser l'échantillon de validation si on dispose de nombreuses données soit la validation croisée.

iii. Avantage et limites des arbres de décision

Les arbres de décision restent une référence pour de nombreuses raisons :

- ils sont très **flexibles** : les variables explicatives peuvent être de nature qualitative, quantitative ou mixte. Les valeurs manquantes sont directement traitées lors de la construction de l'arbre par la création d'une modalité qui rassemble les valeurs manquantes. Les arbres de décision sont robustes face à des données redondantes ou aberrantes. Ils restent consistants même en cas de données atypiques et conservent les interactions complexes au sein des données.

- ils sont **facilement interprétables** : chaque chemin peut être considéré comme une règle. L'aspect graphique des arbres de décision les rend particulièrement compréhensibles. Chaque étape du modèle est connue et peut être si besoin paramétrée. Les experts peuvent intervenir au sein du modèle et cela facile grandement l'appropriation de l'outil au sein des équipes. La probabilité d'appartenance à un groupe peut être déduit de la proportion dans la feuille.

- ils fonctionnent particulièrement bien avec un grand nombre de données car ils sélectionnent les données pas à pas et donc réduisent les dimensions de la problématique.

Dans notre cas, avec une **base de données déséquilibrée**, les arbres de décision restent très souvent utilisés.[413]

Cependant, les résultats d'un arbre de décision **sont très sensibles à l'échantillon[414], avec une forte variance.** Des arbres construits sur une même base de données mais avec des échantillons différents peuvent être très divergents. Ce problème est particulièrement sensible pour les petites bases de données. **Le risque de surapprentissage** est assez important aussi.

- Il est difficile d'avoir une vue d'ensemble : l'arbre sélectionnant les variables les unes après les autres, on n'a pas de visibilité globale sur les effets des variables.

- Un arbre de décision ne donne pas de pourcentage de risque supplémentaire d'avoir un sinistre frauduleux en fonction de critères.

[413] Learning from imbalanced data: an application to bank fraud detection Guillaume Metzler
"Decision trees are often used in the context of imbalanced data (Cieslak and Chawla, 2008; Parvin et al., 2013; Correa Bahnsen et al., 2017) because they are (i) easier to learn and so suitable to deal with large datasets, and (ii) interpretable (each path of the tree can be seen as a knowledge rule)."

[414] Robin Genuer, Jean-Michel Poggi. Arbres CART et Forêts aléatoires, Importance et sélection de variables. 2017. ffhal-01387654v2f, page 5 « Mais l'un des défauts majeurs, intrinsèque, demeure : l'instabilité »

c. Les forêts aléatoires

i. Principes

L'un des principaux défauts des arbres de décision est sa sensibilité à l'échantillon utilisé pour son apprentissage : en fonction de l'échantillon choisi, les résultats peuvent fortement varier et donc entraîner une plus faible capacité de généralisation. Un modèle de **forêt aléatoire** évite cet écueil **en se basant sur de multiples arbres de décision, les plus décorrélés possibles, entraînés sur des sous ensemble de données différentes**. Pour cela, il s'appuie sur 2 techniques : **le *tree bagging* et le *feature sampling*** (ou sélection d'attribut en français).

Le ***tree bagging*** est la conjonction d'un *bootstrap* et d'une agrégation. Il a été introduit en 1994 par Breiman. L'objectif est de réduire la variance et le surajustement du modèle au données. C'est une technique « ensembliste », c'est-à-dire que le bagging « fait coopérer »[415] : plusieurs modèles d'apprentissage pour améliorer les prédictions.

- Le *Bootstrap* renforce la solidité du modèle : cela veut dire que les **données sont re-échantillonnées avec remplacement. Une même observation peut être utilisée plusieurs fois**.

- L**'agrégation** est utilisée car plusieurs modèles avec des résultats moyens auront de meilleurs résultats qu'un seul très bon modèle, notamment car il aura une plus faible volatilité. Pour un modèle de classification, on utilise les classes les plus utilisées pour constituer l'arbre aléatoire. Pour une régression, on prend la valeur moyenne des arbres pour constituer la forêt. En agrégant des modèles avec du bruit mais non biaisés, on réduit la variance.

Les forêts aléatoires reposent aussi sur la **sélection des attributs**, c'est-à-dire que, à chaque branche, un sous ensemble de Co variables est tiré aléatoirement. Chaque branche s'appuie sur un ensemble différent de Co variables. La variabilité entre les différents arbres en est amplifiée, et donc la forêt est composée d'arbres moins corrélés entre eux.

ii. Construction d'une forêt aléatoire

Une forêt aléatoire est créée en 4 étapes :

1. Création d'un ensemble *K* de données avec un tirage aléatoire avec remise
2. Pour chaque ensemble de données, à chaque nœud, sélection de façon aléatoire de *m* attributs
3. Réaliser *K* arbres de décision, un par ensemble de données
4. Pour faire une prévision sur de nouvelles données, il faut l'utiliser dans chacun des *K* arbres et prendre la classe majoritaire

[415] https://eric.univ-lyon2.fr/~ricco/cours/slides/bagging_boosting.pdf

Une forêt aléatoire est définie par plusieurs paramètres :

1. Le nombre d'arbres de la forêt. Plus ce paramètre est élevé plus les prédictions seront stables
2. Le nombre de variables choisies pour le découpage de chaque nœud. C'est un paramètre très important sur la performance de la forêt. On se limite généralement à des arbres peu profonds, car plus l'arbre est développé plus le risque de surapprentissage augmente. L'agrégation des arbres compense la taille des arbres qui est réduite.[416]
3. Le nombre minimum d'observations en dessous duquel l'arbre n'est plus découpé
4. Le nombre d'observations dans chaque échantillon Bootstrap

a. Reduction de la variance

Prenons l'exemple d'une base d'apprentissage *D* constituée de *n* observations et *p* attributs.

On tire de façon aléatoire et avec remise *m* échantillons D_i de taille *n*. Ensuite pour chaque *Bootstrap*, on crée un arbre de décision avec la règle suivante : **à chaque création de branche pour un nœud, on choisit l'attribut dans un sous-ensemble des *p*-attributs** tirés aléatoirement. Généralement on utilise $p' = \sqrt{d}$ pour la classification et $p' = \frac{p}{3}$ pour la régression. Il faut ensuite exécuter les *m* modèles sur les nouvelles données. On agrège les modèles obtenus par moyenne pour la régression ou par classe majoritaire pour la classification.

Si on fait l'hypothèse que les *m* modèles ne sont pas corrélés, la variance de notre forêt aléatoire est :

Var Bagging = Var (Bagging (Modèle)) = $\frac{Var(Modele)}{m}$ [417]

Dans les faits, les modèles sont corrélés suivant un coefficient de corrélation ρ, on aura donc :

$$VarBagging = \rho \times Var(Modele) + \frac{1-\rho}{m} \times Var(Modele)$$

Le bagging permet de réduire la variance sur le terme $\frac{1-\rho}{m}$. Alors que la sélection d'attributs permet de diminuer la variance du modèle lui-même.

[416] http://cedric.cnam.fr/vertigo/cours/ml2/coursForetsAleatoires.html

[417] La variance de la moyenne empirique de n variables aléatoires de même variance σ^2 s'écrit :

$$v(\bar{x}) = \frac{\sigma^2}{n}$$

iii. Avantages et limites

Les modèles de forêt aléatoire sont, comme pour les arbres de décisions, **faciles à visualiser et à appliquer avec un faible nombre de paramètres** et les **interactions complexes** entre données sont conservées. Mais **leur variance est réduite** par rapport à un arbre de décision.

Les forêts aléatoires sont souvent utilisées comme une référence dans le cadre de la détection de la fraude[418]

Cependant les modèles de forêt aléatoire ont souvent des **temps d'apprentissage plus longs**[419]. Ils restent **exposés au risque de surapprentissage**. Ils sont aussi **plus difficiles à interpréter** que les arbres de décision. Si le nombre de variables pertinentes est très faible, les arbres risquent de ne pas être performant et la création d'une forêt aléatoire ne permettra d'obtenir des résultats satisfaisants.

iv. Extrême gradient boosting

GBM (Gradient Boosting Machines ou Generalized Boosted Models) ressemblent aux forêts aléatoires car elles créent des arbres de décision uniques. Ce modèle diffère des forêts aléatoires sur un point majeur : au lieu de préparer les arbres indépendamment, les GBM créent chaque arbre successivement en fonction de l'erreur résiduelle des tous les arbres précédents combinés. C'est particulièrement efficace, étant donné que le modèle concentre pour chaque itération sur les exemples les plus difficiles à prédire.

Etant donnée leur nature itérative, les GBM tendent presque toujours au surapprentissage. C'est pourquoi il faut particulièrement surveiller le nombre d'arbres que le modèle est autorisé à utiliser. On peut utiliser une partie des données d'entrainement pour le déterminer : par exemple, 90% des données d'entrainement sont utilisées pour l'entrainement proprement dit et 10% pour déterminer le nombre d'arbres optimum.

Extrême Gradient Boosting (XGBoost) est une version efficace et très similaire au GBM, mais dont l'algorithme a été optimisée pour réduire les temps de traitement et améliorer la précision.

[418] file:///C:/Users/ceciler/Desktop/References/Machine%20learning/1-%20Decrire%20les%20modeles%20utilises/supervised%20unsupervised%20unbalanced%20data.pdf
[419] https://people.minesparis.psl.eu/fabien.moutarde/ES_MachineLearning/Slides/coursFM_AD-RF.pdf

d. Les K-mean clustering (ou méthode des centres mobiles)

i. Présentation des principes

Les k-mean clustering sont des modèles bases sur les similarités. L'objectif de ce type de modèle est de grouper les données similaires et de découvrir les schémas sous-jacents. Pour cela, un k-mean regarde un nombre défini de groupes. Un groupe correspond à un ensemble de points regroupés à cause de leurs similarités. On définit un nombre cible de *k*, qui définit le nombre de groupes que l'on veut dans la base de données. Un centroïde est l'emplacement réel ou imaginaire du centre du groupe. Chaque point est alloué au groupe le plus proche, tout en essayant de garder les groupes les plus petits possibles. La moyenne dans un K-mean fait référence à la moyenne des données, c'est-à-dire trouver le centroïde.

A la base de ces modèles vient l'idée que la meilleure façon de faire une prédiction, c'est de regarder ce qui a fonctionné dans le passé et de prédire la même chose. On doit donc construire un système basé sur des espaces caractéristiques et des mesures de similarités. L'approche de base est l'algorithme des plus proches voisins ou *« nearest neighbors »* en anglais. Il existe de nombreuses variations, comme les k- plus proches voisins, qui permet de gérer des données avec du bruit. Si les données n'ont pas de label, on parle alors d'une modèle K-mean.

ii. Définitions

On représente les individus comme des points dans un espace ayant comme coordonnées ces mesures. On cherche à regrouper autant que possible les individus les plus semblables, tout en séparant les groupes le plus possible.

La méthode des plus proches voisins : on considère un ensemble de *n* observations, c'est-à-dire des paires (x_i, y_i), étant donnée une nouvelle observation $x \in \mathbb{R}^p$ et avec les observations ordonnées en fonction de la distance entre les x_i avec $\Delta(x_1, x) \leq \Delta(x_n, x)$, un score peut être calculé comme la moyenne y_i des k plus proches voisins.

$$m_k(x) = \frac{1}{k}\sum_{i=1}^{k} y_i$$

On doit donc être capable de calculer une mesure mathématique de la similarité entre observations. Cela peut se faire en considérant la similarité comme une forme de mesure de la distance.

On peut représenter des données dans l'espace de ces caractéristiques en prenant chaque caractéristique commune axe dans un système de coordonnées. Chaque observation est placée en fonction de la valeur de chacune de ses caractéristiques sur chaque axe. Une caractéristique devient une dimension.

KNN signifie en anglais K-Nearest Neighbors. Un seul paramètre doit être déterminé : le nombre de k de voisins à prendre en compte.

Soit T un nouveau point dont on cherche à déterminer la classe. On dispose de *n* points d'apprentissage P_i appartenant chacun a une classe C_i avec i compris entre 1 et *n*. On calcule la

distance entre chaque nouveau point avec les observations existantes. Pour les *k* points les plus proches de la nouvelle observation, on compte le nombre d'occurrence de chaque classe et on attribue à T la classe la plus fréquente.

a. Mesure de la distance

Trois mesures de distance sont majoritairement utilisées. Soit 2 points X, Y de coordonnées respectives $(x_1, \dots x_n)$, $(y_1, \dots y_n)$

- la distance euclidienne

$d_{X,Y}$=$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Pour repartir les points d'un nuage en 2 groupes, et séparer les points qui sont les plus proches d'un point C_1^0 , de ceux qui sont plus proche d'un point C_2^0 en utilisant la distance euclidienne, il suffit de tracer la médiatrice du segment $\left[C_1^0 C_2^0\right]$ (c'est-à-dire la médiatrice en dimension d). La distance euclidienne est généralement utilisée quand toutes les variables ont la même échelle de mesure, car si une variable a une variance plus forte, la distance euclidienne va donner plus d'importance à la différence entre les individus sur cette variable qu'à la différence entre les autres variables.

- la distance de Manhattan

$$d_{X,Y} = \sum_{i\,=\,1}^{n} |x_i - y_i|$$

- la distance de Mahalanobis[420]

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1}}(\vec{x} - \vec{y})$$

Avec $\vec{x}$ qui représente les coordonnées du point et S^{-1} est la matrice de covariance de $\vec{x}, \vec{y}$. Contrairement aux distances euclidiennes, la distance de Mahalanobis tient compte des corrélations entre variables. Dans les calculs de distance euclidienne, des informations sont redondantes comme les variables sont corrélées. L'autre avantage de ce calcul est qu'on n'a pas besoin de standardiser les données comme pour les distances euclidiennes, la matrice de covariance s'en occupe.[421]

b. Choix du paramètre k

Comment fonctionne l'algorithme des k-mean ?

Appelons k le nombre de classes que l'on veut obtenir[422].

1- On tire au hasard *k* individus au hasard dans la population. Ils seront les centres initiaux, les points de départ pour chaque groupe (« cluster ») On choisit aussi une distance entre individus. Les étapes suivantes seront répétées de multiples fois.

2- Constitution d'une classe : On repartit l'ensemble des individus en *k* classes en regroupant autour de chaque centre C_i^0pour $i = 1,2 \dots, n$ l'ensemble des individus qui sont plus proches du centre C_i^0que des autres centres C_j^0pour $j \neq i$

3- Calcul des nouveaux centres : on détermine les centres de gravité $G_1, G_2 \dots, G_k$ des k classes ainsi obtenues et on appelle les nouveaux centres $C_1^1 = G_1$, $C_2^1 = G_2$ …

[420] https://waterprogramming.wordpress.com/2018/07/23/multivariate-distances-mahalanobis-vs-euclidean/
[421] https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2002WR001769
[422] https://math.unice.fr/~diener/MpB2010-2011/COURS9MpB.pdf

4- Les calculs s'arrêtent quand le nombre défini d'itérations a été atteint ou que le découpage en classe n'est presque plus modifié par une itération supplémentaire.

Dans le cas des Centres mobiles (Forgy), on itère l'affectation des individus à une classe et le calcul du barycentre jusqu'à ce que l'inertie intra classe atteigne un minimum local.[423] Pour les k-means (Mc Queen), les barycentres sont recalculés à la fin de chaque allocation d'un individu à une classe et non à la fin des affectations. Les calculs sont plus rapides, mais sont aussi plus dépendant de l'ordre d'apparition des individus.

c. Critère d'évaluation d'une partition

Si on considère K classes d'un ensemble de classes non vides $(C_1, \dots, C_k \dots, C_K)$, 2 à 2 disjointes et dont la réunion forme Ω.

On définit $\mu_k = \sum_{i \in c_k} \omega_i$ le poids de C_k

$g_k = \frac{1}{\mu_k} \Sigma_{i \, \epsilon C_k}^{w_i x_i}$, le centre de gravité de C_k

$d^2(x_i, x_{i'}) = \Sigma \left(x_{ij} - x_{i'j}\right)^2$, la distance entre deux individus.

L'inertie totale T du nuage des *n* individus est[424] :

$T = \sum_{i=1}^{n} \omega_i \, d^2\left(x_{i'g}\right)$, ou g est le centre de gravité du nuage des individus

L'inertie permet de quantifier la dispersion du nuage : elle sera grande pour un nuage très dispersé, et petite si les points du nuage sont rapprochés.

L'inertie interclasse B de la partition est l'inertie des centres de gravité des classes pondérées par μ_k et mesure la **séparation des classes.**

$$B = \sum_{k=1}^{K} \mu_k \, d^2(g_k, g)$$

L'inertie intra-classe W est la somme des inerties des classes et mesure **l'hétérogénéité des classes**.

$$W = \sum_{k=1}^{K} I(C_k)$$

Avec $I(C_k) = \sum_{i \in c} \omega_i \, d^2\left(x_i \ , g_k\right)$

L'inertie totale est égale à l'inertie intra- classe plus l'inertie interclasse[425], c'est-à-dire :

[423] http://iml.univ-mrs.fr/~reboul/ADD4-MAB.pdf, page 12
[424] http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/Classif-L3.pdf
[425] Cette relation se déduit du théorème de Huygues

$$T = W + B$$

iii. Avantages et limites

Pour les données catégoriques, il existe une méthode appelée k-mode, introduite par Z. Huang en 1998. Les variables sont encodées et la mesure de distance entre 2 observations utilise la distance de Hamming.[426]

Les algorithmes k-mean peuvent traiter de très grandes bases, en gardant juste en mémoire les vecteurs des moyennes.

Cependant ce modèle est très sensible aux données : de petits changements peut entrainer une forte variation dans le modèle[427].

La répartition initiale en classes dépend du choix initial des centres.

Le nombre de classes doit être fixé par avance.

iv. Modèle de détection d'anomalie utilisant la distance Mahalanobis

On peut utiliser la distance de Mahalanobis pour identifier les anomalies d'un groupe de données. La distance est nulle si la donnée est à la moyenne et augmente avec la distance du point à la moyenne.

e. Les classifier naïf bayésien

Ces modèles sont bas*é*s sur la notion de probabilité.

i. Théorie

C'est un modèle supervisé qui est basé sur le théorème de Bayes.

Si on définit Y la variable à prédire (avec K modalités) et $X = (x_1, x_2, x_3 \dots x_J)$ les variables descriptives

On cherche à maximiser la probabilité d'appartenance aux classes, ce qui peut s'écrire :

$\overline{y_k} = \ argmax_k \ P(Y = y_k / x_1, x_2, x_3 \dots x_J)$

On peut aussi écrire cette probabilité conditionnelle ainsi :

$P(Y = y_k / x_1, x_2, x_3 \dots x_J) = \frac{P(Y=y) \times P(x_1, x_2, x_3 \dots x_J / Y=y_k)}{P(x_1, x_2, x_3 \dots x_J)}$

Comme on essaie de maximiser $\bar{y}$ et que le dénominateur ne dépend pas de Y, on peut écrire :

$\overline{y_k} = \ argmax_k \ P(Y = y_k) \times P(x_1, x_2, x_3 \dots x_J / Y = y_k)$

[426] file:///C:/Users/ceciler/Desktop/References/Machine%20learning/1-%20Decrire%20les%20modeles%20utilises/Memoire%20scor.pdf, page 24

[427] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Le premier terme $P(Y = y_k)$ peut être estimé facilement à partir d'un échantillon d'observation : c'est la probabilité de chaque modalité de la variable cible. Quand les effectifs sont réduits, on utilise une estimation de probabilité m, qui permet de « lisser » les estimations :

$$\bar{P}(y = y_k) \; = \; p_k = \frac{n_k + m}{n + m \times K}$$

Avec $n_k \; = \; nombre\ d'observations\ de\ la\ modalité\ y_k$ et n le nombre d'observations dans l'échantillon

On fixe souvent $m = 1$

Pour le deuxième terme, nous sommes obligés de faire l'hypothèse que les variables descriptives sont 2 à 2 indépendantes par rapport aux valeurs de la variable cible. On peut alors écrire que :

$P(x_1, x_2, x_3 \ldots x_J / Y = y_k) = \prod_{j=1}^{J} P(x_1, x_2, x_3 \ldots x_j / Y = y_k)$

Afin de faciliter les calculs, nous passons aux logarithmes, en particulier quand le nombre de variables descriptives est élevé.

La règle d'affectation devient alors

$$\overline{y_k} = \; argmax_k \left(lnP(Y = y_k) \; + \sum_{j=1}^{J} ln\, P(x_1, x_2, x_3 \ldots x_j / Y = y_k) \right)$$

ii. Avantages et limites

La littérature retient la grande facilité d'utilisation et de mise en œuvre. Un modèle bayésien naïf peut être utilisé sur de grandes bases de données, mais en même temps nécessite que peu de données pour la phase d'apprentissage. La condition d'indépendance des variables descriptives peut même ne pas être respectées.[428]

Les limites portent principalement encore sur l'interprétation des résultats : le poids de chaque variable explicative sur la décision n'est pas toujours disponible.

[428] "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss" PEDRO DOMINGOS MICHAEL PAZZANI http://engr.case.edu/ray_soumya/mlrg/optimality_of_nb.pdf

3. Evaluation des résultats des modèles

"Essentially all models are wrong, but some are useful."[429]

George Box

a. Capacité de généralisation d'un modèle, entre biais et variance

Pourquoi chercher à évaluer les modèles obtenus ? L'objectif n'est ici pas seulement de pouvoir aider dans le **choix d'un modèle** et de ses paramètres (dans le cas d'un modèle supervisé), mais aussi de pouvoir donner une **mesure de la qualité ou du niveau de confiance que on peut accorder à cette prévision**. En effet, nous n'avons jusqu'à présent aucune information sur la qualité des estimations obtenues.

Les méthodes d'apprentissage permettent de construire un modèle à partir de données initiales. Une des qualités clés d'un modèle est sa capacité à prédire au mieux de nouvelles données, qui n'ont pas été utilisées pour le construire. En informatique, on parle de sa capacité de généralisation, une notion présente pour les modèles supervisés et non supervisés.[430] D'un point de vue statistique, on parle **d'erreur de prévision ou d'erreur de test comme étant « l'erreur moyenne commise par une méthode d'apprentissage statistique pour prédire une réponse sur une nouvelle observation, qui n'a pas été utilisée pour ajuster le modèle** »[431]

Dans le cas d'un modèle supervisé, on peut décomposer l'erreur de prédiction en 3 éléments.

Si on fait l'hypothèse qu'il existe une fonction f_w qui à chaque entrée x associe une sortie t

On peut étudier $t_i = f_w(x_i) + \varepsilon$ ou ε est le bruit (avec une moyenne nulle et une variance σ^2)

On veut estimer $\overline{f_\omega}$ avec les données d'entraînement, en minimisant $E\left[\left(t - \bar{f}_w[x]\right)^2\right]$

- Le biais : c'est la **différence entre l'espérance de la fonction estimée et la fonction réelle**. Si l'espérance de la fonction estimée est égale à la fonction réelle, on peut alors parler d'estimation non biaisée. On peut l'écrire ainsi :

Biais $[\overline{f_\omega}] = E\left[\bar{f}\right] - f$

En d'autres termes, le biais c'est la distance entre les données d'apprentissage et le phénomène générateur qu'on essaie d'approcher. On peut aussi parler d'erreur d'apprentissage. Mais considérant

[429] https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2010.00442.x
[430] Hansen, L. K., & Larsen, J. (1996). Unsupervised Learning and Generalization. In Proceedings of IEEE International Conference on Neural Networks (pp. 25-30). IEEE. https://backend.orbit.dtu.dk/ws/portalfiles/portal/4537864/Hansen.pdf
[431] https://masedki.github.io/enseignements/pdfs/notes.pdf

que chaque phénomène naturel contient du bruit dans ses mesures, on peut aussi le considérer comme des données qu'on a choisi d'ignorer afin d'éviter de prendre le bruit.

- La variance : c'est **l'espérance de la différence aux carrés entre la fonction estimée et l'espérance de la fonction estimée.**

On peut donc écrire :

Var $[\overline{f_\omega}] = E\left[\left(\bar{f} - E[\bar{f}]\right)^2\right]$

On peut aussi considérer la variance comme la mesure du mouvement du modèle en réponse aux données d'apprentissage.

Il a été démontré [432]que l'erreur de généralisation est égale :

$Err = E\left[\left(t - \bar{f}_w[x]\right)^2\right] =$ Biais $[\overline{f_\omega}]^2$ + Var $[\overline{f_\omega}] + \sigma^2$

Concrètement cela veut dire que réduire le biais augmentera la variance et réciproquement, en fonction de la complexité du modèle. Une faible variance et un fort biais auront tendance à sous apprendre des données ('underfitting'), alors qu'un modèle avec un biais plus faible et une forte variance risque le surapprentissage ('overfitting'). **Le modèle optimal doit avoir une variance et un biais faibles.**

Le **surapprentissage** peut être aussi interprété comme le résultat d'un modèle **tellement ajusté aux données utilisées lors de son entraînement qu'il n'est plus en mesure de prédire de nouvelles données** : le modèle a intégré lors de son apprentissage des bruits et lorsqu'il reçoit de nouvelles données, il ne peut généraliser les prédictions. Les modèles les plus adaptables, c'est-à-dire non paramétriques et non linéaires, sont plus sensibles aux problématiques de surapprentissage. Le surapprentissage ou « overfitting » est visible quand les performances sur les données d'entraînement sont bonnes, mais la généralisation donne de mauvais résultats.

Le **sous apprentissage** fait référence à un modèle qui ne peut **ni prédire efficacement les données d'apprentissage ni les généraliser à de nouvelles données, car il n'est pas assez ajusté** aux données. Il est assez facile à détecter. Cela arrive quand les données sont insuffisantes ou quand on essaie d'entraîner un modèle linéaire sur des données non linéaires. Le sous apprentissage ou « underfitting » ou sous apprentissage apparait quand les indicateurs de performance sont mauvais sur les données d'entrainement et leur généralisation. Le sous apprentissage peut être évité avec suffisamment de données et un type de modèle adapté aux données.

[432] https://allenkunle.me/bias-variance-decomposition ou encore
https://ena.etsmtl.ca/pluginfile.php/550493/mod_resource/content/1/Cours2.pdf; slide 19

Cet équilibre est bien illustré par le schéma classique suivant[433] :

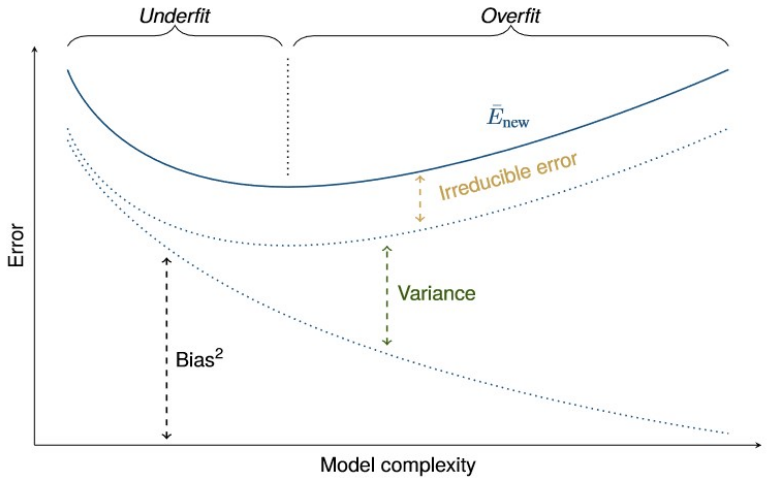


Tableau 33 : Evolution du biais et de la variance d'un modèle en fonction de sa complexité et de ses erreurs de prédictions. Source: Lecture 5 – Cross-validation and the bias-variance trade-of, de Johan Wågberg

b. Définir les données utilisées pour évaluer les performances

Le taux d'erreur théorique est défini comme « la probabilité de mal classer un individu dans la population »[434] . Ne pouvant accéder à toute la population, nous devons créer une estimation. L'objectif est que cet estimateur soit le moins biaisé (que sa moyenne soit la même que le taux d'erreur théorique) et avec le moins de variation possible (faible variance autour du taux d'erreur théorique).

Un estimateur du taux d'erreur théorique est le taux d'erreur en re substitution. On utilise le résultat du modèle sur l'échantillon de données qui a entrainé le modèle, pour calculer le taux d'erreur empirique.

Mais cette technique sous-estime fortement le taux d'erreur théorique : les données utilisées servent à la fois à construire le modèle puis à évaluer sa performance. On parle de biais d'optimisme. Ce biais est d'autant plus fort qu'une observation a un poids important dans la prédiction.

De nombreuses solutions existent dont la validation croisée, early stopping, ou encore régularisation.

La **séparation des observations entre celles qui servent pour l'apprentissage et celles pour l'évaluation des performances est la solution la plus utilisée** pour éviter le biais d'optimisme du taux d'erreur. Différentes méthodes existent, à choisir en fonction des données disponibles, de la capacité de calcul et de la complexité des algorithmes.

[433] Statistical Machine Learning Lecture5–Cross-validation and the bias – variance trade-off
Johan Wågberg, http://www.it.uu.se/edu/course/homepage/sml/lectures/lecture5_handout.pdf
[434] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Resampling_Error_Estimation.pdf

➢ La méthode de l'échantillon test ou « Hold out » :

La base de données est découpée en une **base d'apprentissage et un échantillon test**. Le modèle est construit sur la base d'apprentissage et on l'utilise ensuite pour prédire sur l'échantillon test. En comparant prédiction et observation sur cet échantillon test, on obtient une approximation de l'erreur de test.

Cette méthode est simple, mais **toutes les données ne sont pas utilisées pour l'apprentissage**. De plus, elle n'est vraiment possible que si on dispose de nombreuses observations car il faut que les observations dédiées à l'apprentissage soient suffisamment nombreuses pour construire le modèle, et celles de test, suffisamment importantes pour être capable d'évaluer le modèle avec un degré de confiance suffisante. De plus, la qualité de cette estimation est très variable selon le choix de l'échantillon de test. L'erreur de test dépend en effet alors du découpage des données et peut donc varier significativement.

➢ La validation croisée:

Les données sont aussi **séparées en base d'apprentissage et de test, chaque donnée étant uniquement attribuée à un sous ensemble**. On construit l'estimateur sur l'échantillon d'apprentissage et on utilise l'échantillon test pour calculer le risque de prédiction. Mais cette technique est **répétée plusieurs fois et on moyenne les taux d'erreur ainsi calculés**. Pour chaque sous-ensemble, l'algorithme apprend sur tous les données de tous les autres sous-ensembles. La moyenne du taux d'erreur de chaque sous-ensemble est un proxy du taux d'erreur de l'algorithme.

Des variantes de cette technique existent, avec des **définitions différentes de l'échantillon de test** :

- Méthodes **exhaustives** : **tous les partitionnements possibles** avec un effectif donné sont utilisés. Si les sous-ensembles sont de taille n_e,
    - Le *« leave-p-out »* (LPO) avec $n_e = N - p$

Cela signifie que $N - p$ données sont utilisées pour l'apprentissage et p données sont nécessaires à la validation. Il faut donc construire C_N^p modèles, ce qui peut être très **coûteux en termes de temps de calculs**

    - Le « *leave-one-out* » (LOO) avec $n_e = n - 1$

n-1 données servent à l'apprentissage et une seule donnée pour la validation. Cela signifie que n modèles différents sont créés. Cette technique est possible pour des petits échantillons.[435]Le biais de l'estimateur est très faible, mais la variance est très importante, étant donné qu'une seule donnée est utilisée à chaque fois pour tester le modèle.

- Méthodes **non exhaustives** :

[435] http://cedric.cnam.fr/vertigo/cours/ml2/coursIntroductionApprentissageSupervise.html

- Validation croisée « *K-fold* » **: partition en K- sous-ensembles** des données. Chaque sous-ensemble sert une fois d'échantillon test et le reste d'échantillon d'apprentissage. On a donc K modèles différents, chacun évalué sur un échantillon différent. Dans le cas de K =5, ce schéma de Rebecca Patro illustre bien les calculs[436] :

Tableau 34 : Validation croisée K-fold, illustration de Rebecca Patro,

Une des difficultés est de choisir le nombre K de sous ensemble pour la validation croisée K-fold. En effet, si le nombre de sous ensemble est important, la variance sera plus faible et le biais plus fort. Inversement, avec un nombre de sous ensemble faible, on sera plus en mesure de capter la variance, mais le biais sera plus élevé. Le choix de K doit aussi prendre en compte l'équilibre entre biais et variance.[437] Ce problème d'optimisation « nécessite généralement trop d'observations pour être pratiqué »[438]. Par défaut on choisit une valeur entre 5 et 10[439], le plus souvent 10 (le choix par défaut des logiciels).

- **Échantillonnage répété** (*shuffle and split ou Montecarlo Cross validation*) : **on tire de façon aléatoire *n* données qui seront utilisées pour la validation**, le reste étant utilisé pour l'apprentissage du modèle. On répète ce tirage *k* fois pour avoir *k* modèles. L'inconvénient de cette technique est que certaines données peuvent être utilisées plusieurs fois pour la validation et **certaines ne seront pas utilisées**. Le nombre possible de partition est supérieur à celui d'une validation k-fold. Quand la taille de l'échantillon est trop réduite, cette technique permet d'améliorer la précision[440]. Le même schéma pour cette technique[441] :

Tableau 35 : Monte Carlo Cross Validation, illustration de Rebecca Patro,

[436] https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b
[437] https://masedki.github.io/enseignements/pdfs/notes.pdf, page 24
[438] Qualité de prévision et risque, https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-risque.pdf, page 7
[439] https://www.imo.universite-paris-saclay.fr/~poursat/STA212/2-CV.pdf, page 11
[440] https://juliescholler.gitlab.io/publication/m1dm-1819/M1-Datamining-CM-Partie2.pdf
[441] https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

La méthode du K- fold CV permet d'obtenir un estimateur avec un biais réduit mais une forte variance (il se base sur un nombre limité d'échantillon). Monte Carlo CV peut fournir un estimateur avec moins de variance, mais plus de biais.[442]

En synthèse, quels sont les avantages et inconvénients de la validation croisée ?

Cette technique utilise beaucoup de capacités de calculs mais est aussi l'une des plus précises.

La validation croisée est très utilisée car elle est simple et robuste tant que les données sont indépendantes et identiquement distribuées.[443]

Cependant la validation croisée ne peut être utilisée telle quelle sur une série temporelle non stationnaire, sur des données aberrantes ou dans une collection exponentielle[444]. S**i une classe est déséquilibrée, le rapport entre les classes doit être conservé dans tous les découpages.**[445]

Pour construire des estimations minimisant les biais, il existe aussi la **méthode de la pénalisation de l'erreur d'apprentissage**. L'objectif est de corriger le biais de l'estimateur de l'erreur, en ajoutant une pénalité qui sera liée à la variance et à la complexité du modèle. On peut citer les 3 méthodes les plus connues que sont **le c_ρ de Mallows**, le **critère d'information d'Akaike** (AIC) ou encore le **critère bayésien d'information** (BIC).

Un des inconvénients principaux de cette méthode[446] est qu'elle nécessite une situation paramétrique[447].Dans le cas d'échantillons avec une même variance, le c_ρ de Mallows est même plus performante que la validation croisée.

Nous mentionnerons aussi le **bootstrap**, qui est utilisé principalement sur les combinaisons de modèles[448] Par rapport à la validation croisée, le bootstrap a une variance plus faible mais plus biaisée.

Le choix de la technique doit se faire selon les critères suivants :

- la capacité de calcul disponible : « dans la plupart des cas, le temps de calcul est proportionnel au nombre de découpage »[449]

- la situation paramétrique ou non, homoscédasticité des données, indépendance des données

- la complexité des algorithmes

[442] https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b
[443] « The strongest argument for CV is its quasi-universality: Provided data are i.i.d., CV yields good model selection performances in (almost) any framework." A survey of cross-validation procedures for model selection, Sylvain ArlotCNRS; Willow Project-Team, Laboratoire d'Informatique de l'Ecole Normale Superieure
[444] https://hal.archives-ouvertes.fr/hal-01485508/file/hal_JES_validation-croisee.pdf, page 30
[445] http://cedric.cnam.fr/vertigo/cours/ml2/coursIntroductionApprentissageSupervise.html
[446] https://juliescholler.files.wordpress.com/2018/02/m1-cm-errpred-1718-slides.pdf, page 7
[447] https://juliescholler.gitlab.io/publication/m1dm-1819/M1-Datamining-CM-Partie2.pdf
[448] https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-risque.pdf
[449] Sylvain Arlot. Validation croisée. Myriam Maumy-Bertrand ; Gilbert Saporta ; Christine ThomasAgnan. Apprentissage statistique et données massives, Editions Technip, 2018, page 28

- l'indépendance des données

Dans notre cas, les observations ont été divisées en 3 sous-ensembles, selon une procédure classique[450] :

1. Des données d'apprentissage : algorithme d'apprentissage
2. Des données de validation : choix du modèle, sélection des variables et optimisation des paramètres
3. Des données de test : calcul de l'erreur de prédiction finale

Nous avons isolé les dossiers de sinistres reçus en juillet, août et septembre 2021 pour tester l'erreur de prédiction finale. Les données d'apprentissage ont représenté 80% des données, le dernier 20% étant conservé pour la validation. Comme nous sommes dans le cas d'une base de données déséquilibrée, nous avons veillé à ce que le rapport entre les classes soit conservé lors du découpage en 3 bases.

c. Choisir les mesures de performance adaptées pour les modèles supervisés

Dans le cas des modèles supervisés, les mesures de performance doivent permettre, non seulement d'identifier les modèles les plus pertinents, mais aussi **d'optimiser le choix des paramètres**. Les indicateurs sont nombreux ; mais en fonction de leur construction et des hypothèses sur lesquels ils se basent, ils peuvent apporter une vision erronée des performances du modèle.

Cette problématique est d'autant plus sensible que, dans le cas d'une distribution déséquilibrée, on a besoin de plus d'informations sur la classe minoritaire – alors que justement on a peu d'observations sur ces cas-là. Donc les indicateurs de performance doivent être en mesure de capter précisément les résultats du modèle sur cette classe, en se basant sur un nombre limité d'observations. C'est pourquoi son choix a fait l'objet d'une attention particulière et sera détaillé dans les paragraphes suivants.

Afin de donner une vue d'ensemble des indicateurs possibles, je me suis basée sur la nomenclature définie par l'article de C. Ferri, J. Hernandez et R. Modroiu[451], qui sépare les mesures de performance en 3 familles :

- **Les indicateurs basés sur un seuil et une compréhension qualitative de l'erreur**. Ils sont particulièrement utiles pour minimiser le nombre d'erreurs. Nous décrirons ici la matrice de confusion, accuracy et taux d'erreur, précision, sensitivité et spécificité, F1 score.
- **Les indicateurs basés sur une vision probabiliste de l'erreur**, c'est-à-dire une mesure de la déviation par rapport à la probabilité réelle. Ils permettent de mesurer la fiabilité de

[450] Cours de Marie Anne Poursat, Master 1 Mathématiques Appliquées Université Paris-Saclay, 4 avril 2019, Validation croisée (cross-validation), page 12, https://www.imo.universite-paris-saclay.fr/~poursat/STA212/2-CV.pdf
[451] An experimental comparison of performance measures for classification C. Ferri *, J. Hernández-Orallo, R. Modroiu Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, València, 4 December 2007, https://www.math.ucdavis.edu/~saito/data/roc/ferri-class-perf-metrics.pdf

l'indicateur : pas seulement s'il a prédit la bonne classe, mais de combien il a sélectionné la mauvaise classe… Nous mentionnerons ici le Logloss

- **Les indicateurs basés sur la capacité de classement d'une observation** : la capacité d'un indicateur à séparer les observations. Nous expliciterons ici la courbe ROC AUC et le coefficient de corrélation de Matthews.

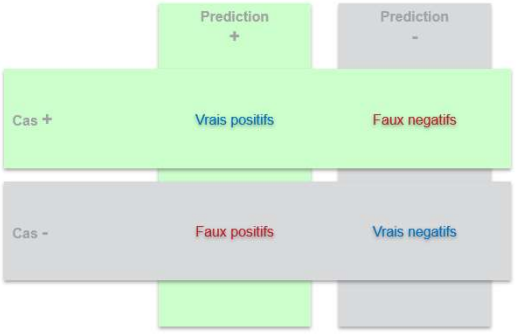
Dans chacune de ces familles d'indicateurs, nous essaierons de distinguer ceux qui sont adaptés à notre problématique.

1. Les indicateurs basés sur un seuil et une compréhension qualitative de l'erreur

a. Matrice de confusion

Dans notre cas, une observation ne peut avoir que 2 natures : soit c'est une fraude, soit c'est une demande d'indemnisation classique. On peut donc estimer les performances d'un modèle en comptabilisant les cas que le modèle classe ou non dans la bonne catégorie.

Si la classe minoritaire a bien été prédit, on parle d'un vrai positif sinon c'est un faux négatif. Quand on parle de la classe majoritaire, le bon classement est un vrai négatif, une erreur est un faux positif.

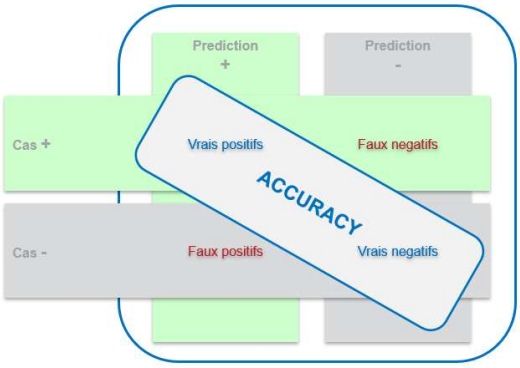


Si le modèle ne donne qu'une probabilité d'appartenance à une catégorie, on définit un seuil, à partir duquel on classe les observations dans une catégorie. Cet outil permet donc de comparer les résultats de nombreux modèles.

Ces nombres d'observations sont transformées en différents indicateurs, mettant en valeur certaines performances du modèle.

b. Accuracy et taux d'erreur

La première mesure utilisée est l'**Accuracy ou Exactitude**. Elle correspond **au nombre d'observations, positives ou négatives, qui ont bien été classées parmi toutes les observations**.



Elle peut s'écrire ainsi :

$m = nombre\ d'observations$[452],

$$Accuracy = \frac{TP + TN}{m}$$

Avec TP = True Positive ou Vrai Positif et TN = True Negative ou Vrai negatif

On peut définir aussi son complémentaire, le **taux d'erreur** comme :

$$Taux\ d'erreur = \frac{FN + FP}{m}$$

Avec FN = False Negative ou Faux Negatif et FP= False Positive ou Faux Positif

Cet indicateur n'est pas pertinent quand on souhaite prédire une classe minoritaire : les faux positifs et les faux négatifs sont mis sur le même plan. Si toutes les observations sont classées dans la classe majoritaire, le taux d'accuracy peut être élevé alors que le modèle n'est pas pertinent.

c. Précision

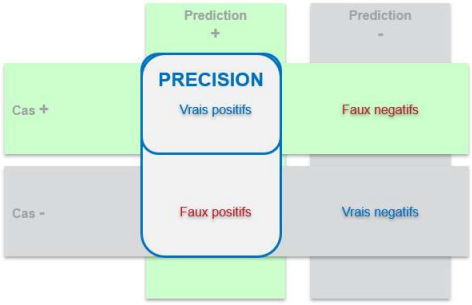
Dans le cas d'une base de données déséquilibrées, on doit s'attacher non seulement à savoir si l'algorithme a commis une erreur ou non, mais aussi sur quelle classe cette erreur a été faite.[453] Ce qui nous intéresse, c'est la **capacité d'un modèle à identifier la classe minoritaire**. On a besoin de mesurer dans quelle mesure les observations positives sont bien identifiées.

La **Précision** (ou Positive predictive value) est le **ratio des observations effectivement positives sur le total des observations prédites comme positives**.

$$Precision = \frac{TP}{TP + FP}$$

[452] https://www.lri.fr/~antoine/Courses/Master-ISI/Tr-evaluation-M-ISIx4.pdf
[453] Guillaume Metzler. Learning from imbalanced data: an application to bank fraud detection. Cryptography and Security [cs.CR]. Université de Lyon, 2019. English.

Avec TP = True Positive ou Vrai Positif et FP = False Positive ou Faux Positif



d. Sensitivité et spécificité

La **sensitivité** mesure le pourcentage de valeurs positives qui ont été correctement identifiées comme positives. **Quand une observation est positive, à quelle fréquence le modèle le prédit effectivement ?**

$Sensitivite = \frac{TP}{TP+F}$ *= Recall= True Positive Rate*

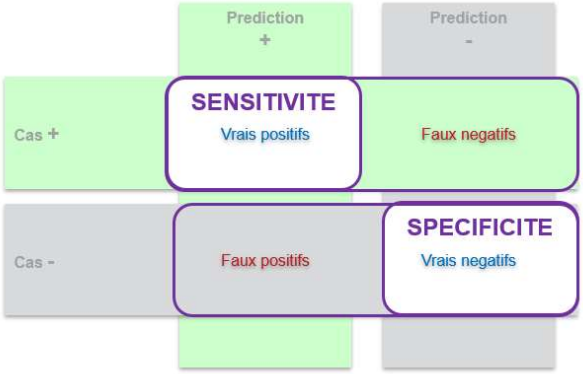
Avec TP = True Positive ou Vrai Positif et FN= False Negative ou Faux Negatif

On peut définir la même chose pour les observations négatives, c'est-à-dire : quand une observation est négative, à quelle fréquence le modèle est-il capable de le prédire ? La **spécificité** (en anglais specificity ou Recall) est le pourcentage d'observations négatives correctement identifiées.

$$Specificite = \ Taux\ de\ Vrai\ Negatif\ = True\ Negative\ Rate = \frac{TN}{TN + FP}$$

Avec TN = True Negative ou Vrai Negatif et FP= False Positive ou Faux Positif

On peut résumer spécificité et sensibilité par le schéma suivant :



e. Indicateurs composites

Entre "faux positifs" et événements non détectés, un équilibre doit être trouvé, dépendant du coût de la détection et de l'événement lui-même. C'est pourquoi de nombreux indicateurs combinent les 2 notions. Par exemple, le G-mean ou le CWA associent sensitivité et spécificité[454].

Dans notre cas, nous avons choisi d'utiliser la moyenne harmonique entre précision et sensibilité, c'est-à-dire la **F-mesure**, qui est un indicateur plus connu :

$$F_\beta = \frac{(1+\beta^2)Precision\ X\ Specificity}{\beta^2 Precision + Specificity} = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}$$

Pour $\beta = 1$, on a :

$$F_1 = 2x\frac{Precision \times Specificity}{Precision + Specificity} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Le F1 score compare les prédictions positives correctes aux erreurs faites par le modèle.

Le meilleur score de F1 est 1. Le modèle est totalement inutile si le F1 score est 0. Faire diminuer les faux négatifs d'un modèle aura un impact sur les faux positifs.

Le F-score présente 2 inconvénients majeurs :

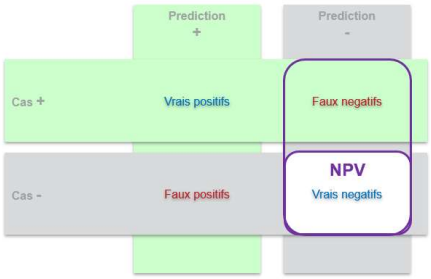
- il est non linéaire : il est difficile de généraliser les valeurs de cette mesure

- il est non convexe : il peut être difficile à optimiser, s'il existe des optima locaux. Il existe cependant des algorithmes pour l'optimiser.

f. Autres indicateurs

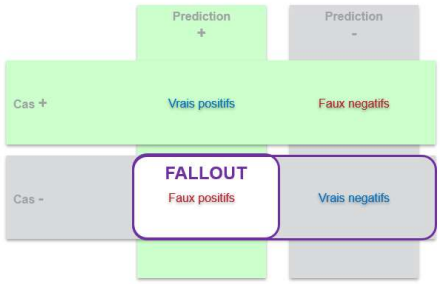
D'autres mesures complémentaires sont possibles comme :

$$\boldsymbol{Taux\ de\ Valeurs\ Negatives\ predites} = Negative\ Predictive\ Value = NPV = \frac{TN}{TN + FN}$$



[454] G-mean = $\sqrt{Specificity\ x\ Sensitivity}$
Pour tout $\alpha \in [0,1]$, CWA = $\alpha\ x\ Sensitivity + (1-\alpha)x\ Specificity$

$$\boldsymbol{Fallout} = Taux\ de\ Faux\ Positif = False\ Positive\ Rate = \frac{FP}{FP+TN}$$ = FPR



g. Limites des indicateurs dérivés de la matrice de confusion

Ces indicateurs ont deux limites principalement :

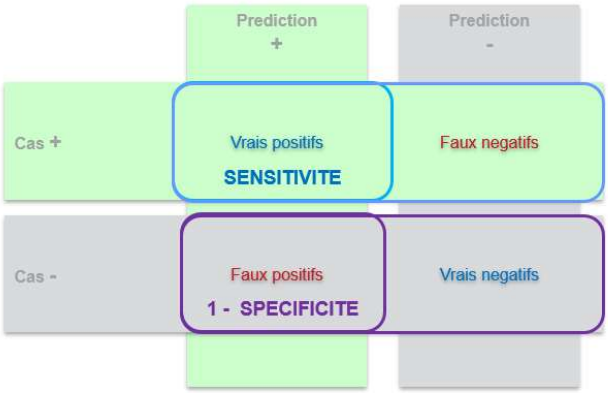
- Ils sont **dépendants d'un seuil de probabilité qui doit être choisi.**

- Ils ne donnent **que des points d'estimations**. On peut être intéressé par avoir une vision plus détaillée de la performance du modèle, sur tout le périmètre de test.

2. *Les indicateurs basés sur la capacité de classement d'une observation*

a. La courbe ROC AUC

La courbe ROC (« Receiver Operating Characteristic ») est définie par les performances d'un modèle, avec en ordonnée **la sensibilité du modèle (c'est-à-dire le ratio de vrais positifs) et en abscisse (1 - spécificité) ou le taux de faux positifs.** Les différentes valeurs sont obtenues en variant le seuil de probabilité qui définit la classe minoritaire, dans notre cas la fraude.



En augmentant le seuil de probabilité, le nombre d'observation se réduit, et donc le nombre de faux positifs et vrais positifs diminue aussi. **Le modèle est d'autant plus performant que de grandes ordonnées sont associées à de petites abscisses**. Ce qui veut dire que plus l'aire entre la courbe et l'axe des abscisses est important, meilleures sont les prédictions du modèle.

Cette aire, autrement appelée le ROC AUC (« Area Under Curve »), est comprise entre 0 et 1. En répartissant les valeurs de façon aléatoire, on aurait un score d'environ 0.5. **Un modèle est intéressant que si son score est supérieur à 0.5 et le plus près possible de 1**. Généralement au-dessus de 0.8 est considéré comme un bon score. A 0, les prédictions des classes sont inversées par rapport à la réalité.

Le ROC AUC est un indicateur du degré de séparation/distinction ou imbrication/chevauchement entre les prédictions de 2 classes. Plus le score est élevé, plus la séparation est marquée et plus faible est le chevauchement des prédictions entre les deux classes.

Par rapport aux autres indicateurs, le ROC AUC permet de visualiser la performance du modèle sur tous les seuils possibles de probabilité. Il peut aussi être utilisé pour définir le meilleur seuil pour une tache donnée ou pour les régressions logistiques[455].

Cependant une étude a montré qu'il peut être optimiste en particulier quand le déséquilibre entre les 2 classes est important[456]. Un petit nombre d'observations peuvent créer d'important changements sur la courbe ROC[457] et biaiser cet indicateur.

Il convient donc d'ajouter à cet indicateur d'autres mesures pour s'assurer d'une bonne compréhension des performances du modèle.

b. Coefficient de corrélation de Matthews

La MCC représente la corrélation entre la classe prédite et la classe réelle. Cette métrique pourrait être résistante au déséquilibre d'observations des différentes classes car elle prend en compte les vrais et faux positifs, autant que les vrais et faux négatifs.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+F\ )\,.\,(TP\ \ \ \ )\,.\,(TN+F\ )\,.\,(TN+F\ )}}$$

Sa valeur est comprise entre 1 et -1 : +1 étant le modèle parfait, 0 étant le modèle aléatoire et -1 une erreur sur toutes les prévisions. Cet indicateur est donc facilement interprétable. C'est le coefficient de corrélation entre les vraies valeurs et les prédictions.

Cependant une étude publiée en 2020 remet en cause sa pertinence pour les bases de données déséquilibrées[458].

[455] "roc_auc_score" Can Be Calculated Also for Regression Models | by Samuele Mazzanti | Towards Data Science
[456] L. A. Jeni, J. F. Cohn and F. De La Torre, "Facing Imbalanced Data--Recommendations for the Use of Performance Metrics," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 245-251, doi: 10.1109/ACII.2013.47.
[457] "Although ROC graphs are widely used to evaluate classifiers under presence of class imbalance, it has a drawback: under class rarity, that is, when the problem of class imbalance is associated to the presence of a low sample size of minority instances, as the estimates can be unreliable." Page 55, Learning from Imbalanced Data Sets, 2018.
[458] https://www.sciencedirect.com/science/article/abs/pii/S016786552030115X

3. *Les indicateurs basés sur une vision probabiliste de l'erreur c'est-à-dire une mesure de la déviation par rapport à la probabilité réelle = le Logloss*

Le Logloss est un indicateur qui mesure la distance de chaque prédiction par rapport à sa classe effective, en se basant sur la probabilité prédite. La différence entre Log-loss et le ROC AUC est qu'il **indique quelle est la distance**, alors que le ROC-AUC montre le chevauchement des prédictions. Le Logloss permet d'indiquer de combien la probabilité prédite est proche de la vraie valeur.

Logloss = $-\frac{1}{N}\sum_{i=1}^{N} y_i log\, p_i + (1 - y_i) \log(1 - p_i)$

Avec *N* = nombre d'observations, *i* une observation donnée, p_i la probabilité prédite, y_i le résultat de la -ème observation.

On peut aussi dire que le **Logloss mesure l'incertitude** : plus il est bas, moins l'incertitude est importante. Plus la probabilité prédite diverge de la vraie valeur, plus élevée est la valeur du log-loss. Donc, utilisé sur un même échantillon, un modèle ayant un Logloss plus petit sera préférable à un autre modèle avec un score plus élevé.

Une baisse du Log-loss ne signifie pas forcément un changement du ROC AUC, tant que le niveau de séparation entre les 2 classes reste le même. Mais une augmentation du ROC AUC garanti une baisse du Logloss, car les données sont plus séparées, donc plus proches de leurs classes respectives. Le ROC AUC ne prend pas en compte la distance entre la prédiction et la classe actuelle, il prend uniquement en compte le niveau d'imbrication des valeurs prédites par rapport à leurs classes respectives.

Le seul inconvénient de cette métrique est sa difficulté d'interprétation. En effet, son niveau varie fortement en fonction du nombre de classes prédites et de la prévalence de la classe minoritaire.

Dans le cadre de la détection de la fraude, notre priorité est de **détecter les cas réellement frauduleux (Vrais Positifs) tout en limitant le nombre de fausses alertes (Faux Positifs), qui peuvent être coûteux en temps et en investigations. En même temps, on souhaite aussi limiter les fraudes non détectées (Faux Négatifs).**

- **ROC AUC :** Cette mesure nous donne une première idée de l'efficacité du modèle à distinguer les Fraudes (Vrais Positifs) sur les fausses alertes (Faux Positifs)
- **F1 score :** Cet indicateur permet de comprendre le ratio de vraies fraudes (Vrais Positifs) par rapport aux fausses alertes (Faux Positifs) et les fraudes non détectées (Faux Négatifs)
- **Logloss :** Cette mesure donne une vision plus fine de la distance entre les prédictions du modèle et la réalité.

d. Choisir les mesures de performance adaptées pour les modèles non supervisés

1. AUC Synthétique

Pour les modèles de détection d'anomalies, les précédents indicateurs ne peuvent être utilisés. Dans cette étude, nous allons utiliser un AUC Synthétique[459]. Il est calculé en **générant deux bases de données à partir de la base de validation- un normal et un autre avec les anomalies. Un score en termes d'anomalie est calculé pour les deux échantillons et les données sont libellées en conséquence**. Le ROC AUC est ensuite calculé en se basant sur ces derniers labels.

2. Score Silhouette

Pour évaluer des modèles de regroupement ou partitionnement (*clustering* en anglais), on peut utiliser le Score ou Coefficient Silhouette[460]. Cet indicateur combine une vue de la **cohésion au sein du clusters** (la distance moyenne entre chaque point au sein d'un cluster notée *a*), **comme du niveau de séparation entre les clusters** (la distance moyenne entre les clusters notée *b*). Si tous les points d'un cluster sont proches, ils ont un fort degré de similarité.

Il varie de -1 à 1. **Plus il se rapproche de 1, meilleure est la séparation entre clusters et les clusters sont bien denses.** En cas de valeur proche de 0, des clusters peuvent se chevaucher ou être très proches entre clusters voisins. Une valeur négative peut signifier des valeurs affectées à de mauvais clusters.

Avec $C(i)$ le cluster affecté au i -ème point de données et

$|C(i)|$ Le nombre de points de données dans ce dernier cluster

On note $a(i)$=$\frac{1}{|C(i)|-1}\sum_{c(i),i\neq j} d(i,j)$ et note $b(i)$=$min_{i\neq j}$ ($\frac{1}{|C(j)|}\sum_{j\epsilon C(j)} d(i,j)$

Alors le score Silhouette s'écrit :

$$s_{(i)} = \frac{b_i - a_i}{\max(b_i, a_i)}$$

[459] Anomaly detection: Data Robot docs

[460] Peter J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Volume 20, 1987, Pages 53-65, ISSN 0377-0427, https://doi.org/10.1016/0377-0427(87)90125-7.
(https://www.sciencedirect.com/science/article/pii/0377042787901257)

4. Revue de la littérature scientifique sur l'utilisation de l'apprentissage statistique dans la détection de fraude en assurance automobile

N.	Titre	Auteur	Algorithmes utilisées	Conclusion
1	Modélisation de la fraude sur les contrats Auto du particulier par apprentissage statistique	Cédric Decourtieux	- Régression logistique - Arbre de décision - Réseaux de neurones	« Les meilleurs résultats ont été obtenus par le **random forest pour la base brute** et le **boosting pour la base ajustée** avec un AUC respectif de 0.92 et 0.95 »
2	Modélisation de la fraude en assurance IARD : Application à l'assurance automobile	Khouloud Barhoumi	- Régression logistique - Forêt aléatoire - Gradient boosting	« La performance des trois modèles est globalement satisfaisante (AUC est de l'ordre de 70%). **Les résultats sont proches** » Les meilleurs résultats sont obtenus par les forêts aléatoires et le gradient boosting.
3	Comparative Analysis on Classification Algorithms of Auto-Insurance Fraud Detection based on Feature Selection Algorithms (Sept 2018)	Sapna Panigrahi 1*, Bhakti Palkar2	- Forêt aléatoire - Naif Bayes - K-Nearest Neighbor - Arbre de decision	**Les forêts aléatoires ont de meilleures performances en termes de accuracy et precision, les arbres de décision en termes de recall.**
4	A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance	G. Ganesh Sundarkumara,b, Vadlamani Ravia,n	- k Reverse Nearest Neighborhood - One Class support vector machine (OCSVM) - Arbre de Decision - Support Vector Machine - Regression logistique - Probabilistic Neural Network (PNN), - Group Method of Data Handling (GMDH), - MultiLayer Perceptron (MLP).	Les Arbres de décision et les Support Vector Machine atteignent de hauts niveaux de sensibilité, respectivement de 90.74% and 91.89% sur la base de données d'assurance. Il n'y a pas de différence significative entre Support Vector Machine et Arbres de décision, mais comme les **Arbres de décision permettent de définir des règles « et si », ils sont préférables.** Ce papier démontre **l'importance d'éliminer les valeurs extrêmes (bruits) et données redondantes de la classe majoritaire en cas de base de données fortement déséquilibrée**.
5	A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis	Stijn Viaene, Richard A. Derrig, et	- Naïf Bayesien - AdaBoosted naif Bayesien scoring de Ridgeway et al. - AdaBoosted weights of evidence	Sur des données de la vie réelle, AdaBoosted weights of evidence obtiennent de meilleurs résultats que les autres algorithmes sur toutes les mesures de performance.

		Guido Dedene		
7	Automobile Insurance Fraud Detection	Mark Anthony Caruana et Liam Grech	- Réseau de neurones artificiels - Classifier Naif Bayésien.	"**Les performances des 2 méthodes de classification sont comparables**. Le Classifier Naif bayésien surperforme légèrement en termes de classification des nouveaux sinistres correctement ; cependant, il a aussi plus de faux positifs, ce qui n'est pas souhaitable, étant donné le possible coût d'investigation des sinistres. **Le sous échantillonnage semble effectif**. »
8	TGANs with Machine Learning Models in Automobile Insurance Fraud Detection and Comparative Study with Other Data Imbalance Techniques	Rohan Yashraj Gupta, Satya Sai Mudigonda, Pallav Kumar Baruah	- Arbre de décision - Forêt aléatoire - Support Vector Machine - XGBoost - LightGBM - Gradient Boosting Machine (GMB).	Cette étude analyse les impacts des techniques de gestion des bases de données déséquilibrées (SMOTE, MWMOTE, ADASYN et TGAN) sur les performances des principaux algorithmes de classification. **Les meilleures performances sont obtenues avec un TGAN et un GBM**.

4ème partie : Application des modèles d'apprentissage statistiques

Un projet d'apprentissage statistique comporte généralement 6 étapes[461]:

1. Compréhension du besoin du métier
2. Compréhension des données
3. Préparation des données
4. Modélisation
5. Evaluation
6. Déploiement

Cependant dans la pratique, des allers retours peuvent exister entre préparation des données avec la sélection des variables, modélisation et évaluation :

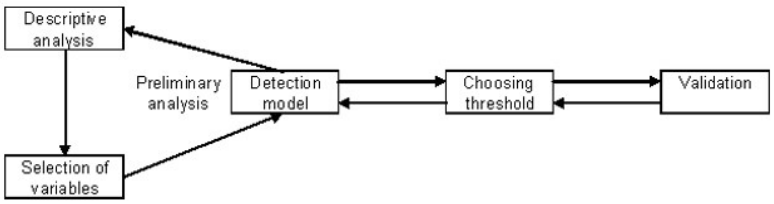


Tableau 36 : Etapes d'un projet d'apprentissage statistique, Source : Automatic Fraud detection – does it work ? By I. Lxgreid (2007)

1. Analyse des données

"In God we trust, all others bring data."

William Edwards Deming (1900-1993)

a. Description des données

i. Définition et typologie des variables

Les demandes d'indemnisations sont décrites à travers **94 caractéristiques**, qui peuvent être regroupées en 3 types de variables :

- Les données liées à l'assuré : AssuredClientCode, Insured courriel, …
- Les données liées au contrat d'assurance : policy ID, Previous Claims history
- Les données liées à l'accident et aux réparations : Incident Date, Location of Accident, Request Channel, BodilyInjury, …

Avant retraitement, ces caractéristiques sont de 3 natures différentes :

a. Variables qualitatives[462]

[461] Fundamentals of Machine learning for predictive data analytics, J. D. Kelleher, B. M. Namee & A. D'Arcy, 2015, MIT

[462] Les données sont qualifiées de variables qualitatives si elles expriment une qualité ou une classe et de quantitatives si elles peuvent être décrites par des nombres. Au sein des variables qualitatives, on distingue si une hiérarchie peut être faite entre les classes (ordinale) ou non (nominale). Si 2 modalités sont seulement possibles, sans qu'elles puissent être hiérarchisées, on parle alors de variable binaire.

- Ordinales
- Nominales
- Variables binaires

b. Variables quantitatives
   - Continues
   - Discrètes

ii. Analyse univariée sur la base de données

Les techniques utilisées pour décrire une variable dépend de sa nature qualitative ou quantitative.

a. Variables qualitatives

On cherche dans ce cas à **décrire les effectifs et les proportions des différentes modalités**.[463] Pour cela, nous avons pu disposer, pour chaque variable qualitative,

- Des effectifs avec un diagramme en barre
- Un tableau de fréquence
- Du nombre de valeurs uniques et de données manquantes.

Nous n'avons pas procédé à des tests plus avancés de normalité étant donné que les modèles d'apprentissage ne nécessitent **aucune hypothèse sur la distribution des données**. La revue des données a pour objectif de comprendre les données et s'assurer de leur qualité.

b. Variables quantitatives

Pour les variables quantitatives, nous nous sommes basés sur :

- Une représentation graphique de la distribution des variables incluant le % de fraudes pour chaque quantile
- Un tableau de fréquence des valeurs
- Les indicateurs de position (minimum, moyenne, médiane, maximum)
- Un indicateur de dispersion (écart type)
- Les indicateurs de forme (coefficient d'asymétrie, coefficient d'aplatissement)
- Le nombre de valeurs uniques et de données manquantes.

Parmi les variables quantitatives, on sépare les variables continues (qui peuvent avoir un nombre infini de valeur), des variables discrètes (qui ne peut avoir que des valeurs limitées comme les entiers par exemple).

[463] Statistique descriptive Pr. Nicolas MEYER, Laboratoire de Biostatistique et Informatique Médicale Fac. De Médecine de Strasbourg, novembre 2010 http://passthrough.fw-notify.net/download/828027/http://udsmed.u-strasbg.fr/labiostat/IMG/pdf/P2_3.pdf, page 15

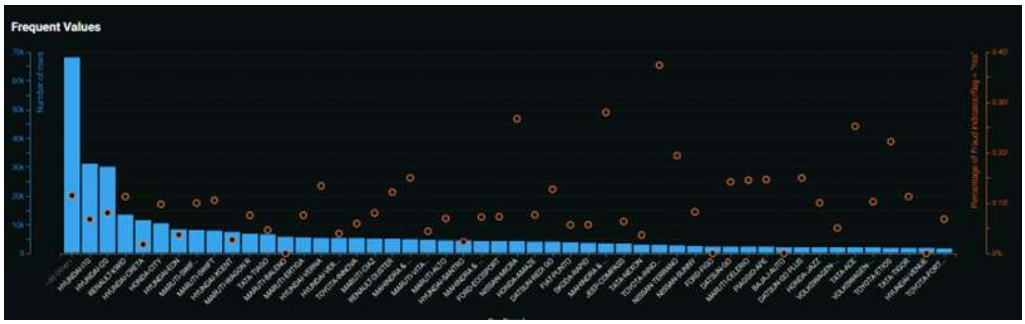
b. Présentation du portefeuille et des demandes d'indemnisation

Future Generali propose en Inde 5 types de produits : couverture santé, assurance voyage, assurance « Lifestyle » (assurance pour les golfeurs, l'art, le sport, mariage…), assurance habitation et assurance automobile. Au moment où cette étude a été réalisée, le portefeuille assurance automobile de Generali était constitué majoritairement de **véhicules à 4 roues, localisés dans le nord du pays.**

En synthèse, dans la base de gestion des indemnisations :

- Les sinistrés :
  - **76% d'hommes** versus 14% de femmes[464]
  - **Moyenne d'âge : trentaine**
  - 20% nouveau business, 80% renouvellement
  - Dans 99% des cas, c'est une première demande d'indemnisation.
  - Les véhicules sont majoritairement immatriculés dans le nord de l'Inde : New Dehli, Uttarakhand, Punjab, Chandigarh.
- Le type de véhicules :
  - **44% de véhicules compactes, 26% de taille moyennes**, 11% de SUV, 6% de voitures haut de gamme et 1.6% de rickshaw.
  - **50% pétrole versus 48% diesel**
  - Les marques les plus représentées sont : **Hyundai, Honda,** Renault et Tata**.**

Tableau 37 : Répartition des marques de voitures sinistrées



- Les sinistres :
  - Un accident à 98%, vol (0,8%) et dégâts des eaux (0.5%)
  - 88% un véhicule en mouvement (12% véhicule en stationnement)
  - 99% sans blessure corporelle, 2 076 avec blessures

- Le processus
  - L'accident a été déclaré dans 98% via un call center.
  - 81% des demandes ont fait l'objet d'une inspection : 80% par le garage et 19% par un inspecteur extérieur.
  - Le coût moyen de la réparation est entre 24 et 26k INR, soit entre 258 et 280 euros.
  - Parmi les 91% des demandes qui ont été clôturées, 80% des paiements ont été faits au garage et seulement 13% aux assurés directement.

[464] La donnée est manquante pour 10% des sinistré(e)s.

- Les succursales en charge du traitement sont Mumbai, West Dehli, Bangalore, Hyperabad.

2. Préparation des données

"How to Lie with Bad Data" est le titre d'un article de Richard D. De Veaux and David J. Hand pour Statistical Science[465]

La base de données a fait l'objet de pré-traitements, afin de faciliter son exploitation et optimiser les résultats des modèles d'apprentissage statistique. A noter que cette étape, très consommatrice en temps et en énergie, est aussi maintenant considérée comme **clé dans les travaux de Machine Learning**, avec parfois un impact aussi déterminant que la technique de modélisation elle-même. [466]

« En réalité, le point essentiel dans la performance d'un modèle est la qualité de la préparation des données »[467]

« La mauvaise qualité des données est l'ennemi numéro un de la diffusion et de l'utilisation profitable du Machine Learning » [468]

a. Eliminer les doublons

La première étape souvent suggérée est la **suppression des observations identiques** dans la base de données[469].

D'un point de vue théorique, cette action est justifiée par ces 4 arguments :

1. La présence de données dupliquées augmente les risques de surapprentissage. [470] Le surapprentissage apparait quand un modèle généralise trop l'information qu'il a appris de la base de données d'apprentissage, et qu'il ne peut plus prédire efficacement sur un nouvel échantillon de valeurs.
2. Les données dupliquées représentent un cas d'échantillon non aléatoire et son biais se retrouvera dans le modèle crée sur cette base. [471]

[465] Article in Statistical Science · August 2005 DOI : 10.1214/088342305000000269 · Source : OAI
[466]« Les performances d'un modèle dépendent souvent moins de la technique de modélisation que de la nature et de la qualité des données » Améliorer les performances d'un modèle prédictif : perspectives et réalité, Stéphane TUFFERY
[467] Améliorer les performances d'un modèle prédictif : perspectives et réalité, Stéphane TUFFERY
[468] « Poor data quality is enemy number one to the widespread, profitable use of machine learning» https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless
[469] Data Wrangling with Python, 2016, — Page 173 *"… if you have used raw data that may have duplicate entries, removing duplicate data will be an important step in ensuring your data can be accurately used"*
[470] « Duplicate data increases the chances of overfitting." https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d
[471] "Duplicates are an extreme case of nonrandom sampling, and they bias your fitted model. Including them will essentially lead to the model overfitting this subset of points." https://stats.stackexchange.com/questions/428191/should-i-remove-duplicates-from-my-dataset-for-my-machine-learning-problem

3. Cela améliorera artificiellement les performances du modèle qui ne pourra avoir les mêmes niveaux de performance avec des données réellement aléatoires.[472]
4. Il a été montré que cela impactait en particulier les performances des modèles de classifications lors de l'étude de A. Kołcz et al. (2003)[473]

Cependant certains auteurs ne partagent pas ce point de vue, arguant que les observations en doublon peuvent être aussi une représentation réellement aléatoire de la distribution de la variable. La valeur apparaissant en doublon peut avoir une probabilité d'occurrence juste plus élevée et la retirer entraînerait une perte d'information importante.[474]

Donc la question importante est **de s'assurer de la représentativité de la base de données par rapport au phénomène étudié**. Un doublon peut être conservé s'il correspond bien à une observation de la variable aléatoire analysée, et non un bruit venant de l'observation même du phénomène.

Il existe aussi un autre cas ou le dédoublonnage n'est pas forcément nécessaire : en cas de base de données déséquilibrées. Dans ce cas, la présence de doublons d'observations peut, sous réserve de conserver un caractère aléatoire et porter sur la classe déséquilibrée, avoir les effets d'un oversampling ou suréchantillonnage de cette classe manquante.

Dans notre cas, la base de données est effectivement déséquilibrée, mais **les doublons ont été enlevés, identifiés par le numéro de demande d'indemnisation**.

b. Gérer les valeurs manquantes

Les modèles d'apprentissage classiques s'appuient sur des bases de données complètes, sans envisager que des valeurs puissent manquer[475]. Ignorer cette problématique pourrait amener à de fortes pertes de précision, ou même à des biais dans les modèles d'analyse[476].
La réalité ne permet pas toujours de disposer d'une telle base de données. Quelles stratégies peuvent s'appliquer ?

On doit distinguer trois types de **valeurs manquantes** selon leur caractère plus ou moins aléatoire, selon D. Rubin[477].

1. Les données qui manquent de façon complètement aléatoire (ou Missing completely at random, MCAR). La probabilité d'être manquante est la même pour toutes les données.

[472] From an algorithm evaluation perspective, duplicate rows will result in misleading performance. This will result in an optimistically biased estimate of performance on unseen data. https://machinelearningmastery.com/basic-data-cleaning-for-machine-learning/
[473] The results obtained indicate that the presence of duplicates in the learning sample does indeed pose a problem, even if they cover a diverse range of content. This stresses the importance of performing data deduplication which, even if not perfect, can at least ensure that the level of duplicate contamination is reduced. This seems to be essential, since our results indicate that the loss of classification accuracy can be strongly correlated with the contamination level. Disponible sur Researchgate :
https://www.researchgate.net/publication/228547112_Data_duplication_an_imbalance_problem/citations
[474]because if it is a random sample, that represents the underlying distribution in the data, that tells you that this value has a higher probability. Removing duplicates will just render the dataset useless.
https://stackoverflow.com/questions/34935940/are-duplicates-useful-in-data-sets/34937252
[475] "Statistical techniques do not enable the analysis of data in which there are missing cases, consequently different approaches for dealing with the issue of missing values have evolved." Missing Data: The Importance and Impact of Missing Data from Clinical Research, C.R. Padgett and al (2014)
[476] "Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions." The prevention and handling of the missing data, Korean J Anesthesiol. 2013 May; 64(5): 402–406.
[477] Inference and missing data, BY DONALD B. RUBIN, Biometrika (1976), 63, 3, pp. 581-92

Cela signifie que la cause d'absence de cette donnée n'est pas reliée aux données en elle-même. Cette situation n'est malheureusement que rarement le cas dans les faits.

2. Les données qui manquent de façon aléatoire (ou Missing at Random, MAR) La probabilité d'être manquante est la même pour un groupe de données défini par la donnée aléatoire observée.
3. Les données dont on peut dire qu'elles ne manquent pas de façon aléatoire (ou Missing Non at random ou MNAR)

Seule une expertise métier permet de se faire une idée de la nature du processus des données manquantes, les données elles même ne sont pas suffisantes. Et dans le cas des MNAR, il est difficile de décrire le processus à l'origine des données manquantes.

De nombreuses stratégies existent, avec des impacts différents selon la nature des données manquantes et leur fréquence. La littérature est particulièrement développée, mais ce n'est pas le cœur de notre étude. Nous avons donc choisi de présenter les principales solutions brièvement, afin de choisir la plus adaptée à nos données.

La solution, qui peut **paraître la plus rapide et la plus pragmatique est la suppression des observations manquantes**. Cette solution peut s'envisager quand plus de 50% des valeurs sont manquantes sur la variable, que l'on dispose d'une base de données assez importante (en millions d'observations), et si le % de données manquantes sur la base de données au global n'est pas trop importante.

Mais cette solution part du principe que les données manquantes le sont de façon complètement aléatoire (MCAR). Si cette hypothèse n'est pas vérifiée, on introduit un biais dans l'échantillon[478]. Même si cette hypothèse est vérifiée, la réduction de la taille de l'échantillon signifie aussi une diminution des capacités d'apprentissage des algorithmes.

Si la suppression n'est pas pertinente, se pose la question de **l'imputation des données manquantes**. La solution la plus simple serait de remplacer ces données manquantes par une seule et même valeur, permettant ainsi de conserver les observations. Par exemple, pour des données numériques, la moyenne peut être utilisée pour remplacer la valeur manquante. Dans le cas des données catégoriques, on substitue la valeur manquante par la catégorie la plus fréquente. Cette technique permet de réduire la perte d'information, et est assez facile à mettre en place.

Elle entraine des biais dans l'estimation des paramètres : la variance des estimateurs est particulièrement sous-estimée.[479] Les indicateurs de corrélation et de covariance sont aussi fortement atténués.[480] Cette technique permet certes de garder la taille de l'échantillon, mais entraîne une sous-estimation de l'erreur[481].

On peut aussi essayer de prédire les données manquantes, par exemple avec des régressions ou des modèles de classification. On construit un modèle par variable, en utilisant les données complétées pour l'apprentissage du modèle. Cela permet de prendre en compte les interactions entre variables, même si cela peut prendre du temps, en particulier sur de grosses bases de données.

Enfin, on peut aussi utiliser des algorithmes qui supportent les données manquantes, comme les XGBoost. Dans le cas de cette dernière, elle va même extraire de l'information[482].

[478] Applied Missing Data Analysis, Enders, C.K. (2010). Applied missing data analysis (1st ed.). New York: The Guildford Press, page 39
[479] https://perso.univ-rennes1.fr/valerie.monbet/doc/cours/IntroDM/Chapitre4.pdf, page 26
[480] Applied Missing Data Analysis, Enders, C.K. (2010). Applied missing data analysis (1st ed.). New York: The Guildford Press Page 43
[481] "Consequently, in no situation is mean imputation defensible, and you should absolutely avoid this approach" Applied Missing Data Analysis, Enders, C.K. (2010). Applied missing data analysis (1st ed.). New York: The Guildford Press Page 43
[482] XGBoost: A Scalable Tree Boosting System, Tianqi Chen et Carlos Guestrin, 10 Juin 2016, Page 4

Dans le cas de notre étude, les données manquantes ont été flaggées et **les catégories ayant plus de 50% de données manquantes ont été supprimées. Pour les modèles linéaires, la valeur médiane a été mise à la place de la donnée manquante. Pour les arbres de décision, la valeur -9999 a été utilisée**.

c. Identifier les valeurs aberrantes

1. Définition

Les valeurs aberrantes ou « outliers » sont définies par Hawkins comme des observations qui diffèrent tellement des autres observations qu'on peut soupçonner que **ces données soient générées par un mécanisme différent du reste des données**.[483] Jason W. Osborne and Amy Overbay identifient 5 causes possibles de valeurs aberrantes dans leur article[484]:

- Les erreurs de données lors de la collecte, l'enregistrement ou l'entrée de la donnée
- Un mauvais reporting intentionnel ou non
- Une erreur d'échantillonnage
- Une erreur de standardisation
- Une erreur d'hypothèse sur la distribution

Pourquoi il est important de les identifier ?

- Selon Jason W. Osborne and Amy Overbay, les valeurs aberrantes peuvent avoir des effets délétères sur les analyses statistiques. **Ils augmentent la variance de l'erreur, et réduit le pouvoir des tests statistiques.**
- Si ces données aberrantes ne suivent une loi normale, elles peuvent réduire la 'normalité' des données.
- Enfin, elles peuvent biaiser les estimateurs.

Comment les identifier ?

De nombreuses méthodes existent. Pour des raisons de praticité**, on a choisi de se baser sur un seuil et non sur un score**. On définit un seuil minimum et maximum en dessous et au-dessous duquel la valeur est définie comme aberrante.

1- **Se baser sur l'écart type** : on pose alors :

Seuil min = moyenne – 3 x écart type

Seuil max = moyenne + 3 x écart type

Cette solution peut convenir pour des **données suivant une loi normale**, ce qui n'est pas toujours le cas, ou à minima symétrique et uni-modale. Elle est inutile sur les distributions bimodales[485]

2- Utiliser les quartiles :

Inter-Quartile Range (IQR) = 75eme percentile – 25eme percentile

Seuil min = 25eme percentile – 1,5 x IQR

Seuil max = 75eme percentile +1,5 x IQR

[483] Hawkins (1980): "An observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".

[484] Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research, 9(6) Online at http://pareonline.net/getvn.asp?v=9&n=6

[485] https://www.math.univ-toulouse.fr/~epauwels/LearningM2SID/outlierDetection.pdf, page 33

3- Utiliser la méthode Ueda[486]

Cette technique se concentre sur la symétrie de la distribution pour identifier les valeurs aberrantes. Elle est facile à calculer, ne nécessite pas de déterminer un nombre de valeurs aberrantes à l'avance et peut être utilisée sur de larges échantillons. Cette méthode est capable de détecter l'absence de valeurs aberrantes, comme leur multiplicité.

Elle se base sur le calcul de l'AIC ou Akaike Information Criterion pour identifier les valeurs aberrantes. Posons X une variable aléatoire avec une distribution de probabilité $f(x)$ et $(x_1 \dots x_n, \dots x_N)$ un échantillon aléatoire de taille $N = n + s$ de cette distribution avec n et $s = (1,2,3, \dots)$ le nombre d'observation extrêmes

$$AIC = 2k - 2\ln(L)$$

Où

k Est le nombre de paramètres indépendants

L Est la vraisemblance maximum pour n nombre d'observations normales

En cas de distribution normale, l'AIC peut s'écrire[487] :

$$AIC = 2n\, ln\, \hat{o} - 2\ln n! + 2s$$

Avec o l'écart type et s le nombre de valeurs extrêmes

Le test statistique proposé est donné par :

$$U_t = \frac{1}{2} AIC$$

$$\approx n\, ln\, \hat{\sigma} - \sqrt{2}\,.\,s\,.\frac{\ln(n!)}{n}$$

$$= (N - s)\, ln\, \hat{\sigma} - \sqrt{2}.\,s\,.\frac{\ln(n!)}{n}$$

Pour calculer U_t,

- il faut ordonner l'échantillon pour avoir $x_{ord} = x_{(1)}, x_{(2)}, \dots x_{(N)}$ avec $x_{(i)}$ étant le i -ème chiffre dans l'ordre statistique
- calculer le z-score pour chaque $x_{(i)}$ avec

$$z_{(i)} = \frac{x_i - \bar{x}}{\bar{s}} \text{ Avec } \bar{s} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})}{N-1}}$$

L'objectif est de regarder la réduction de l'écart type à mesure qu'on retire les valeurs possiblement aberrantes.

4 – Utiliser une Isolation Forest

Une Isolation Forest[488] est une méthode non supervisée qui va isoler les valeurs aberrantes, en se basant sur des algorithmes d'arbre de décision. L'isolation forest permet d'isoler les anomalies sans se baser sur des mesures de densité ou de distance. Elle se base sur deux propriétés des anomalies : ils sont minoritaires avec quelques observations, et ils ont des caractéristiques qui sont très différentes des observations normales.

Une isolation forest, c'est la construction d'un ensemble d'arbres binaires appelées iTree pour une base de données. 3 paramètres sont utilisés : le nombre d'arbres à construire, la taille des échantillons, la taille limite de l'arbre pendant l'évaluation. Parce que les anomalies sont plus susceptibles d'être isolées, elles sont plus proches de la racine de l'arbre, alors que des points normaux seront isolées à

[486] A simple method for the detection of outliers, Taichiro Ueda, Mitsubishi Electric Computer Systems Corporation, Tokyo, 2009 http://siba-ese.unisalento.it/index.php/ejasa/article/view/988/823

[487] https://www.slideshare.net/POOJAPATIL211/outlier-detection-by-uedas-method

[488] Isolation-based Anomaly Detection, Fei Tony Liu and Kai Ming Tin https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/tkdd11.pdf

la fin de l'arbre. Dans l'article « Isolation-based anomaly detection », Fai Tony Liu et Kai Ming Ting montrent que la précision de la forêt converge rapidement avec un petit nombre d'arbres[489].

Une fois ces valeurs identifiées, que faire avec ces données ? Doivent-elles être retirées ? Si oui, définitivement ou être traitée comme une donnée manquante et remplacée ? Les valeurs extrêmes peuvent particulièrement impacter les arbres de décision, en changeant leur taille et précision[490].

Dans le papier « The power of outliers (and why researchers should always check for them"[491], Jason W. Osborne et Amy Overbay précisent que la démarche à suivre devrait dépendre de l'origine de ces valeurs extrêmes. Si ces données ne sont pas légitimes (erreurs de saisie, erreur dans l'échantillon, erreur dans la standardisation de la donnée, erreur sur l'hypothèse de distribution…), elles peuvent être enlevées de la base de données. Sinon elles peuvent apparaitre comme inhérente à la variabilité propre à la donnée, et il serait inadéquat de la supprimer. Une possibilité est de la conserver en la transformant, comme une troncature, ou les scores extrêmes sont recodes en valeurs raisonnables.

Dans notre cas, nous avons choisi de **revoir ces valeurs avec le métier pour identifier si ces données sont possibles ou sont liées à des erreurs de saisie, en utilisant la méthode Ueda d'identification.**

d. Normaliser les valeurs

La nature des variables peut nécessiter des transformations[492]:.

Si les caractéristiques ont **des échelles de valeurs très différentes**, il peut être intéressant de **réduire l'intervalle de variations de ces données**[493]. Plusieurs méthodes sont possibles, parmi les plus utilisées, on trouve[494]:

- la standardisation des caractéristiques : cela signifie soustraire la moyenne à chaque valeur et ensuite la diviser par l'écart type. Cela permet de centrer les valeurs autour de zéro. Mais cette technique ne devrait être utilisée que pour les données suivant une loi normale.

- la normalisation min-max, consiste à mettre à l'échelle la variable entre 0 et 1, c'est à dire en enlevant le minimum à chaque valeur et en divisant par la plage de valeur. La plage de valeur est calculée en faisant la différence entre la valeur maximum et la valeur minimum de la caractéristique. Cette technique est une bonne alternative quand les données ont un biais.

- l'échelonnage par le maximum implique de rééchelonner les valeurs d'une caractéristique entre 0 et 1, en divisant chaque valeur par la valeur maximum.

- un ré-échelonnage robuste implique retirer la valeur médiane de chaque donnée et le diviser ensuite par la gamme interquartile, qui se calcule en soustrayant le quartile 75 et 25. Cette méthode est adaptée aux données très biaisées.

[489] https://blog.paperspace.com/anomaly-detection-isolation-forest/
[490] https://www.inderscience.com/info/inarticle.php?artid=107073
[491] Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why Copyright © 2004. researchers should ALWAYS check for them). Practical Assessment, Research, all rights reserved and Evaluation, 9(6) Online at http://pareonline.net/getvn.asp?v=9&n=6
[492] Sélection des Echantillons pour le Problème de classification en Distributions déséquilibrées, CHAIRI IKRAM, Soutenue le 25 juillet 2014
[493] http://eric.univ-lyon2.fr/~ricco/cours/slides/discretisation.pdf
[494] https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

Il peut être intéressant de **supprimer les caractéristiques ayant une seule valeur, ou très peu de valeurs, ou une très faible variance**[495]

Les variables continues peuvent devoir être transformées en variables discrètes dans les cas suivants :

- pour certains modèles de classification[496]

- quand ces variables ont des distributions de probabilités non standards, très asymétriques ou très dispersées[497]

- quand le type de variables est très hétérogènes.[498]

- la prise en compte des valeurs manquantes (dont l'imputation est toujours délicate) qui se trouvent regroupées dans une modalité spécifique ; [499]

- la neutralisation des valeurs extrêmes (individus hors norme) qui se trouvent intégrés à la première ou la dernière modalité ;

- la prise en compte des ratios dont le numérateur et le dénominateur peuvent être tous deux > 0 ou < 0 (on trouve de tels ratios en analyse financière) ;

- une plus grande robustesse (on constate souvent que 2 ou 3 modalités permettent d'augmenter l'aire sous la courbe ROC, du moins sur l'échantillon de test, par rapport à 4 ou 5 modalités) ;

Dougherty et al. (1995) classent les méthodes de discrétisation selon trois critères : méthodes supervisées versus non supervisées, globales versus locales et dynamiques versus statiques.
Nous décrirons une des méthodes possibles **: la discrétisation à égale distance**[500] **(« equal-width discretization »)**. On divise l'ensemble des valeurs possibles en N ensemble de même taille. Sa taille est calculée en divisant par N la différence entre la valeur maximum et minimum. Cette méthode ne permet pas d'améliorer la dispersion des valeurs mais gère les valeurs extrêmes.

Les variables catégoriques peuvent aussi être transformées en variables numériques, cela s'appelle l'encodage[501]. La plupart des algorithmes d'apprentissage supervisé nécessitent des données d'entrée numérique, comme les réseaux de neurones par exemple.

On peut citer 3 méthodes usuelles : encodage étiqueté (« label encoding »), encodage à points (« one hot encoding »), et encodage de fréquence (« frequency encoding »)[502]

- **Encodage étiqueté** : Les variables catégoriques avec *n* valeurs sont transformées en une variable numérique prenant *n* valeurs différentes. L'avantage de cette technique est d'alléger la mémoire. Le principal inconvénient est la transformation des données catégoriques en données ordonnées. Cela peut être un problème pour des modèles comme les SVM, les réseaux de neurones. Cela n'affecte pas

[495] https://machinelearningmastery.com/basic-data-cleaning-for-machine-learning/
[496] Page 296, Data Mining: Practical Machine Learning Tools and Techniques, 4th edition, 2016.
[497] https://machinelearningmastery.com/discretization-transforms-for-machine-learning/
[498] http://eric.univ-lyon2.fr/~ricco/cours/slides/discretisation.pdf, page 5
[499] Améliorer les performances d'un modèle prédictif : perspectives et réalité
[500] https://heartbeat.comet.ml/hands-on-with-feature-engineering-techniques-variable-discretization-7deb6a5c6e27
[501] Yvan Lucas. Credit card fraud detection using machine learning with integration of contextual knowledge. Artificial Intelligence [cs.AI]. Université de Lyon ; Universität Passau (Deutscheland), 2019. English. NNT: 2019LYSEI110. tel-02951477
[502] https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-encoding-categorical-variables-be4bc0715394

les classifications usant les arbres de décisions, puisqu'ils séparent la base de données en différents sous-ensembles.

- **Encodage à points** : Les variables catégoriques avec *n* valeurs sont transformées en *n-1* variables binaires décrivant la valeur de la variable catégorique. Cette méthode peut être utilisée avec des classifications car il ne met pas d'ordre dans les catégories : toutes sont équidistants.

- **Encodage de fréquence** : la value des catégories est remplacée par le pourcentage de la classe minoritaire.

Dans notre étude, nous avons utilisé :

- Pour les forêts aléatoires (Gini et GBoost) : encodage étiqueté avec des chiffres ordinaux (*ordinal encoding*)
- Pour la classification naïve bayésien, la régression logistique, k-mean clustering : encodage à points (one hot encoding)

e. Analyse des corrélations entre variables et réduction des dimensions

Avant de commencer les travaux d'apprentissage statistique proprement dits, il peut aussi être intéressant d'effectuer une sélection des variables.

Enlever les variables qui n'apportent pas d'information ou sont redondantes[503] permet de **réduire le temps de calcul de la modélisation**, et dans certains cas, améliorer les performances du modèle. La pratique montre que « Non seulement un petit nombre de variables suffit à bien expliquer un phénomène, mais (cela) n**uirait à la qualité du modèle en introduisant plus de variables explicatives** »[504]

Une autre raison de réduire les variables explicatives, est que, « plus les variables sont nombreuses dans un modèle, plus le risque est grand de voir l'une de ces variables provoquer une défaillance du modèle car elle aura été un jour mal alimentée par les programmes informatiques de constitution de la base de données. Il en va des modèles comme des mécaniques : plus elles sont complexes, plus elles risquent de tomber en panne. » [505]

De nombreuses méthodes existent comme des modèles de régression avec une variable ou encore des algorithmes de sélection des variables[506] : univariate feature selection, L1 based feature selection, tree-based feature selection. Dans cette étude, nous nous sommes basés sur **le test V de Cramer** pour identifier les corrélations.

Le V de Cramer permet de mesurer la force du lien entre 2 variables. Il se calcule ainsi :

[503] "Feature selection is primarily focused on removing non-informative or redundant predictors from the model." Page 488, Applied Predictive Modeling, 2013.
[504] Améliorer les performances d'un modèle prédictif : perspectives et réalité, Stéphane TUFFERY
[505] Améliorer les performances d'un modèle prédictif : perspectives et réalité, Stéphane TUFFERY
[506] file:///C:/Users/ceciler/Desktop/References/Machine%20learning/1-%20Decrire%20et%20preparer%20les%20donnees/Feature%20selection%20SapnaPanigragi_MTech2018.pdf

Soient X_1 et X_2 deux variables avec respectivement les modalités K_1 et K_2 et n le nombre d'observations

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n.\min(K_1-1, K_2-1)}}$$

Sa valeur est comprise entre 0 et 1, avec V=1 signifiant une totale dépendance des deux variables. On considère que le lien entre les variables est fort quand le V de Cramer est supérieur à 0.4[507] et il y a colinéarité à partir de 0.8.

Les colinéarités suivantes ont été détectées dans notre étude :

feature1	feature2	statistic
Payment To	insured_account_missing_flag	1.00
AssuredClientCode	Customer ID	0.98
Previous Claims history	Previous claims settled	0.97
latest surveyor code	latest surveyor name	0.96
latest surveyor name	latest surveyor inhouse yesno	0.95
Previous claims settled	Total Claim Amount	0.93
Car Brand	Car Type	0.93
Previous Claims history	Total Claim Amount	0.91
Cause of Loss	VehicleParkedYesNo	0.84
Accident Pincode	ServicingBranchZone	0.82
ServicingBranchZone	latest surveyor name	0.81

Tableau 38 : Liste des colinéarités détectées entre les variables des demandes d'indemnisation

f. Présentation de la base de données après retraitement

i. Séparation en 3 différentes bases selon l'étape du processus

Lors de la gestion d'un sinistre, les informations du dossier sont complétées au fur et à mesure de l'évolution du dossier. Des modèles différents ont donc été bâtis en fonction de l'avancement du processus, des données disponibles et du types de modélisation :

- Première étape : **enregistrement du dossier de sinistre sans l'information du** garage en charge du sinistre
  - Modèle 1.1 supervisé
  - Modèle 1.2 non supervise

- Deuxième étape : **enregistrement du dossier de sinistre avec l'information du garage** en charge du sinistre

- Troisième modèle : Traitement du sinistre
  - Modèle 3.1 supervisé
  - Modèle 3.2 non supervisé

ii. Présentation des 3 bases de données utilisées

La base de données initiale comportait 94 variables. A la fin de cette phase de préparation, entre 54 et 69 variables ont été conservées.

[507] Le V de Cramér – lemakistatheux (wordpress.com)

	Nb de caracteristiques
Base de donnees initiale	94
Base de donnees declaration sans garage ni inspecteur	54
Base de donnees declaration avec garage et inspecteur	57
Base de donnees sinistre en cours de traitement	69

Tableau 39 : Evolution du nombre de variables avant et après traitement

3. Modélisation

Pour déterminer les paramètres optimums, nous avons travaillé sur la base de données des sinistres en cours de traitement et sur les **données de validation.** (Cf. b. Définir les données utilisées pour évaluer les performances)

a. Forêt aléatoire

Dans la 3ème partie, nous avons défini les 4 principaux paramètres qui permettent de définir une forêt aléatoire que sont :

1. Critère de séparation d'un arbre de classification
2. Elagage d'un arbre de décision (pruning)
3. Règle d'attribution des observations en cas de feuille avec plusieurs modalités
4. Estimation de l'erreur de prévision

Dans notre étude, nous avons choisi **l'indice d'impureté de Gini** comme critère de séparation d'un arbre de classification. Et la règle d'attribution des observations en cas de feuilles avec plusieurs modalités est la **règle de la majorité**, c'est-à-dire qu'on attribue à la feuille la modalité à prédire qui est la plus nombreuse. L'estimation de l'erreur de prévision se fera par validation croisée.

Nous avons souhaité déterminer **le nombre optimum d'arbres et la profondeur maximum de chaque arbre**, pour obtenir des estimations les plus pertinentes possibles.
Vous trouverez ci-dessous les résultats obtenus, n_estimators étant le nombre d'arbres de la forêt et max_depth la profondeur maximale d'un arbre.

Random Forest classifier (Gini)

n_estimators	max_depth	AUC (CV)
50	10	0.8307
100	10	0.8356
500	10	0.8451
1000	10	**0.8464**
1500	10	**0.8487**
2000	10	**0.8492**
1000	1	0.7966
100	100	0.7516
1000	100	0.8309
500	100	0.8309

Tableau 40 : Variation de l'AUC de la Forêt aléatoire en fonction du nombre et de la taille des arbres

Sur les paramètres testés, la qualité des prévisions a augmenté avec le nombre d'arbres dans la forêt, jusqu'à 2000 arbres. Une profondeur de la forêt trop importante impacte négativement les prévisions. **Si un arbre a trop de feuilles, il risque de sur-adapter aux données d'apprentissage.**

b. Forêt aléatoire XG Boost

Lors de l'implémentation des forêts aléatoires XG Boost, nous avons essayé de déterminer les 3 paramètres suivants :

1. Le **nombre optimum d'arbres XG boost** (n_estimator)
2. La limite à partir de laquelle on vérifie les performances de la forêt aléatoire avant de la continuer (smooth_interval)
3. L'intervalle de vérification des performances de l'arbre (interval)

Ci-joint les valeurs testées et retenues :

Extreme gradient boosted tres classifier with early stopping			
n-estimators	smooth_interval	interval	AUC (CV)
5,000	200	10	0.8451
10,000	200	10	**0.8484**
20,000	200	10	0.8484
100	200	10	0.756
150	200	10	0.8222
230	200	10	0.8477
300	200	10	0.8481

Tableau 41 : Variation de l'AUC du XGBoost en fonction du nombre d'arbres

c. K-mean clustering

Pour déterminer le nombre de cluster à choisir, nous avons évalué les performances des k-mean suivants :

	Silhouette score	
k-mean clustering	# clusters	Validation
	5	**0.1587**
	7	0.1404
	10	0.1467
	20	0.1433
	40	0.1312
	60	0.1178
	80	0.1177
	100	0.1056

Tableau 42 : Variation du score du k-mean en fonction du nombre de clusters

5ème partie : Analyse des résultats et perspectives

1. Le projet au sein de Generali Inde

Le projet a connu plusieurs phases :

1. Une première phase de « Preuve de concept » ou POC en anglais
2. Un déploiement en production, un premier calibrage qui a été jugé non satisfaisant en production
3. Après recalibrage, une présentation des résultats à la direction Asie de Generali

a. Preuve du concept (POC)

Ce projet a commencé par une phase de POC. Ce travail s'est basé sur une extraction des sinistres de janvier 2017 à décembre 2020 (au total 237 510 déclarations de sinistres), dont **147 fraudes (0.06%).** L'objectif était de prédire les fraudes sur les sinistres déclarés entre janvier et juillet 2021, soit 64 198 déclarations de sinistres.

Après une première préparation des données, une forêt aléatoire XG Boost est utilisée et permet d'obtenir un score pour chaque sinistre. Afin d'évaluer concrètement sa pertinence, les 120 sinistres avec les scores les plus élevés sont revus par des experts pour vérifier s'ils correspondent effectivement à des cas de fraude. En novembre 2020, les résultats suivants sont présentés à la direction Asie de Generali, sur cet échantillon de 120 sinistres :

		Prédiction	
		Fraude	Non Fraude
Réalité	Fraude	35	1
	Non Fraude	7	77

Tableau 43 : Revue des 120 sinistres du POC par des experts métiers

Sur 120 sinistres, les gestionnaires n'avaient initialement identifié que 7 fraudes ; la détection de la fraude est clairement améliorée sans pour autant augmenter les faux positifs. La direction Generali Asie confirme son intérêt pour cet outil et le déploiement en production est décidé.

b. Premier calibrage en production

Lors du déploiement en production, on garde dans un premier temps les paramètres obtenus lors du POC et on essaie de détecter la fraude sur les sinistres de janvier 2021 à septembre 2021, soit 99 252 sinistres. Les premiers résultats suivants sont obtenus :

		Prédiction	
		Fraude	Non Fraude
Réalité	Fraude	148	287
	Non Fraude	15 146	83 671

Tableau 44 : Revue des résultats après le 1er calibrage en production

Ces résultats sont jugés **insatisfaisants**. Une nouvelle modélisation est réalisée **en élargissant la base de données utilisée pour l'entrainement** : on utilise les sinistres de janvier 2017 à juin 2021 soit 332 449 sinistres avec 0.09% de fraudes. On cherche à prédire les cas de fraudes entre juillet et septembre 2021, soit 43 888 sinistres.

2. Evaluation

Ces évaluations de l'erreur de prédiction ont été faites sur les données de test, non utilisé pour l'entrainement des modèles ni pour le choix des paramètres.

a. Résultats des Modèles supervisés

		Donnees de test (Holdout data)				
	Modeles	AUC	Logloss	F1 Score	Precision	Sensitivite
Enregistrement **sans** le garage et l'inspecteur	Foret aleatoire (Gini)	0.7348	0.0071	0.0163	0.011	0.0309
	Foret aleatoire XGBoost	**0.7694**	**0.00692**	**0.0261**	**0.0155**	0.0825
	Modele de classification naif bayesien	0.5946	0.00737	0.0045	0.0023	0.2268
Enregistrement **avec** le garage et inspecteur	Foret aleatoire (Gini)	0.764	0.00696	0.0251	0.0211	0.0309
	Foret aleatoire XGBoost	**0.8083**	**0.00671**	**0.0532**	**0.0422**	0.0722
	Modele de classification naif bayesien	0.6121	0.00732	0.0058	0.0029	0.268
En cours de traitement	Foret aleatoire (Gini)	0.8305	0.00662	0.0396	0.0245	0.1031
	Foret aleatoire XGBoost	**0.8668**	**0.00632**	**0.0722**	**0.0722**	0.0722
	Modele de classification naif bayesien	0.69	0.0071	0.0106	0.0054	0.3711

Tableau 45 : Evaluation des performances des modèles en fonction de la base de données à prédire

Les **forêts aléatoires XGBoost** donnent les meilleurs résultats, quelle que soit l'étape du process et les indicateurs de performance utilisés.

Globalement les résultats sont meilleurs au fur et à mesure du processus, plus on dispose d'informations.

b. Résultats des Modèles non supervisés

		Silhouette score	
	Nb de clusters	Training	Validation
k-mean clustering	5	0.1619	0.1587

Tableau 46 : Evaluation des performances du k-mean clustering sur les données de validation

Les résultats du k-mean clustering ne sont **pas satisfaisants**, étant donné que leur score Silhouette se rapproche plus de 0 que de 1.

	Modeles	AUC synthetique
Enregistrement **sans** le garage et l'inspecteur	Modele de detection d'anomalie (Mahalanobis)	0.6911
Enregistrement **avec** le garage et inspecteur	Modele de detection d'anomalie (Mahalanobis)	0.6228
En cours de traitement	Modele de detection d'anomalie (Mahalanobis)	0.6234

Tableau 47 : Evaluation des performances du modèle de détection des anomalies (Mahalanobis)

Le modèle de détection d'anomalie utilisant les distances de Mahalanobis présente un **résultat modeste** mais supérieur à un modèle juste aléatoire.

c. Facteurs de risques identifiés

Au-delà des résultats des modèles, il est aussi intéressant de comprendre quelles sont les variables identifiées comme augmentant la probabilité d'une fraude.

Au sein de Generali, d'autres facteurs peuvent déclencher une alerte fraude comme :

- Le client demande peu de temps auparavant une question sur un dommage et a un sinistre similaire peu de temps après.
- Le même sinistre est déjà présent dans les sinistres précédents.
- Police d'assurance avec des sinistres annulés
- Fort taux de sinistres dans la zone
- Pertes avec un montant proche de celui de la limite maximale
- Police d'assurance avec des sinistres récents

Les variables suivantes sont aussi ressorties lors de la modélisation (cf. détails en annexe) :

- L'atelier de dernière réparation (*latest_workshop_code*)
- La **description de la réclamation** (*Claims_description*)
- Lieu de l'accident (*Accident_Pincode*)
- Moment de la journée lors du signalement (*Reported_date_time*)

- Historique des sinistres (*Previous_claims_history, Previous_claims_closed_without payment/withdrawn)*
- Date de facturation ou de livraison post réparation *(Invoice_date, delivery_order_date)*
- Date de début ou de fin de la police d'assurance ou nombre de jours avant renouvellement *(Policy_Start_Date, Policy_End_Date, Next_renewal_date)*

3. Déploiement

Les forêts aléatoires XGBoost et le modèle de détection d'anomalie utilisant les distances de Mahalanobis ont été déployés sur la déclaration et le traitement des sinistres.

Sur les 43 888 sinistres, on sait que 214 sont réellement frauduleux (0,49%).

a. Détection à la déclaration du sinistre

Tableau 48: Résultats du modèle de détection d'anomalie (Mahalanobis)

		Détection	
		Fraude	Non Fraude
Réalité	Fraude	1	213
	Non Fraude	83	43 591

Tableau 49: Résultats du Forêt aléatoire XG Boost

		Détection	
		Fraude	Non Fraude
Réalité	Fraude	4	210
	Non Fraude	1 341	42 333

b. Détection lors du traitement du sinistre

Tableau 50: Résultats du modèle de détection d'anomalie (Mahalanobis)

		Détection	
		Fraude	Non Fraude
Réalité	Fraude	40	174
	Non Fraude	1 341	42 333

Tableau 51: Résultats du Forêt aléatoire XG Boost

		Détection	
		Fraude	Non Fraude
Réalité	Fraude	201	13
	Non Fraude	4 044	39 630

Seule la forêt aléatoire XGBoost obtient des résultats intéressants sur les sinistres en cours de traitement : en analysant 9,7% des sinistres, elle détecte 94% de la fraude.

Ces résultats ont été présentés à la direction de Generali Asie. Le coût de déploiement et de maintenance de ces modèles étant trop élevés par rapport aux économies réalisées (réduction du temps consacré à la détection de la fraude par les experts), il a été décidé **d'améliorer les processus existants et la sensibilisation des équipes.**

4. Perspectives et limites

a. Critiques et amélioration des modèles

Cette étude a été **limitée par la taille et la qualité de la base de données d'apprentissage**. Une des limites des modèles supervisés est bien de dépendre d'une base de données suffisamment renseignée pour permettre cet apprentissage. Or la fréquence des fraudes détectées dans la base de données d'apprentissage est assez basse (moins de 1%), nécessitant une base de données d'autant plus volumineuse. La **qualité de la base de données** et ses nombreuses données incomplètes ont rendu cette tache encore plus complexe.

Un **travail d'échantillonnage** aurait pu aider à rééquilibrer la base de données. Il est possible lors d'un *oversampling* ou sur-échantillonnage d'augmenter le nombre d'observations de la classe minoritaire afin d'obtenir un ratio entre les 2 classes plus équilibré. Le sous-échantillonnage ou *undersampling* diminue, lui, le nombre d'observations de la classe majoritaire. L'étude de Mark Antony Crauana et Liam Grech montre en effet que le sous-échantillonnage peut améliorer les performances des modèles de classification bayésien et des réseaux de neurones dans le cas de la détection de la fraude en assurance[508].

Deux autres modèles supervisés auraient aussi pu être ajoutés à cette étude : **les Support Vecteur Machine et les réseaux de neurones**. Le premier modèle produit de meilleurs résultats que la forêt aléatoire XGBoost sur la déclaration de sinistres avec garage et inspecteur, mais pas sur les sinistres en cours de traitement :

[508] Caruana, Mark & Grech, Liam. (2021). Automobile Insurance Fraud Detection. Communications in Statistics Case Studies Data Analysis and Applications. 10.1080/23737484.2021.1986169?tab=permissions&scroll=top.

	Modeles	Holdout data				
		AUC	Logloss	F1 Score	Precision	Sensitivity
Enregistrement avec garage et inspecteur	support vector machin	**0.8158**	**0.00673**	**0.0568**	**0.0633**	0.0515
	Keras Deep Residual Neural network classifier using Training schedule (3 layers : 512, 64, 64 units)	0.7337	0.00631	0.0615	0.1212	0.0412
En cours de traitement	support vector machin	0.8564	0.00651	0.0379	0.0299	0.0515
	Keras Deep Residual Neural network classifier using Training schedule (3 layers : 512, 64, 64 units)	0.5753	0.0114	0.0246	0.0142	0.0928

Tableau 52: Evaluation des performances des modèles de SVM et reseaux de neurons

b. Perspectives

Dans le cas de la détection de la fraude en assurance automobile, on pourrait aussi utiliser un **modèle de classification prenant en compte des pénalités** en fonction du type d'erreur de prédiction (*cost-sensitive models*). Le but de ce modèle n'est plus de réduire toutes les erreurs de la même façon, mais de prioriser les erreurs dont l'impact est le plus fort.

Cette étude porte sur le marché indien de l'assurance. Si on veut comparer avec son concurrent chinois, la détection de la fraude est une priorité pour les autorités, notamment à Hong Kong. Chaque compagnie d'assurance se doit d'avoir un dispositif de lutte contre la fraude. Sa détection reste cependant basée sur des systèmes de scoring des agents et des clients. En cas de suspicion de fraude, une analyse plus détaillée est effectuée par des agents et le processus automatique de paiement est levé.

Conclusion

Même si les PIB de l'Inde et de la France ont atteint des niveaux similaires dans les années 2019/2020, leurs marchés d'assurance automobile diffèrent fondamentalement par leurs dimensions : la **pénétration et la densité d'assurance** sont particulièrement basses pour l'Inde à cause de profondes inégalités au sein de sa population. La société indienne reste plus rurale et bien **moins équipée proportionnellement** que la France des années 80. On ne peut retrouver un boom de l'équipement automobile comme en France entre les années 60-70, mais on constate **une croissance très volatile de son équipement automobile**. Les **risques assurés ont des fréquences bien plus élevés**. Même si on retrouve dans les deux cas des marchés dominés par les constructeurs nationaux, la flotte automobile indienne comme les assurés sont bien plus jeunes que sur le marché français.

On peut cependant rapprocher les marchés de l'assurance indiens et français des années 80 sur de nombreux points :

- Le parc automobile indien actuel est 1,3 celui de la France de 1985.
- On retrouve un **cadre économique similaire** : une dynamique de croissance, des taux d'intérêts élevés, un chômage qui se développe, des montants de dépenses des particuliers en croissance.
- **Le marché de l'assurance** en Inde de 2019 se rapproche en termes de **dimension et de dynamique à celui de la fin des années 80 en France**. Le secteur est, dans les deux cas, animé par une forte croissance, une concentration des acteurs et une ouverture au secteur privé. Les frais de gestion sont élevés et les primes ne couvrent pas le montant des sinistres.

Dans ce contexte de forte compétition, les sociétés d'assurance ne peuvent ignorer la fraude, qui, en impactant tarification et provisionnement, peut accélérer l'antisélection parmi ses clients.

Mais une **définition légale adaptée manque** que ce soit pour la France ou l'Inde. Les assureurs peuvent s'appuyer uniquement sur des définitions de professionnels, le Code des assurances pour la France ou le code pénal pour les Indiens. Face à la fraude, les assureurs peuvent annuler le contrat, rejeter l'indemnisation et poursuivre l'auteur au pénal. Mais pour les assureurs indiens, les poursuites judiciaires au pénal sont rares tant l'investissement en temps et en énergie sont importantes.

Dans les deux cas, l'évaluation de la fraude est difficile et **oscille le plus souvent entre 10% à 15% des prestations versées.**

Face à ce risque, les assureurs français se sont dotés de nombreux organismes de lutte contre la fraude (ALFA, ARGOS, ANEA) ou encore de fichiers partagés listant les résiliations et leurs causes. Cette coopération entre acteurs fait défaut en Inde. Au sein de Generali Inde, la détection de la fraude se fait lors des demandes d'indemnisation, sur la base des règles métiers et l'expérience des gestionnaires de sinistres.

L'objectif de ce mémoire était de mettre en place **un dispositif d'apprentissage statistique** pour améliorer la détection de la fraude. L'apprentissage statistique se base à la fois sur l'apprentissage automatique pour la construction d'un modèle capable de prévision à partir des données, tout en gardant l'attention à la qualité des prévisions de la statistique.

L'avantage de l'apprentissage statistique est qu'elle ne nécessite **pas d'hypothèse sur la structure ou la distribution des données** et elle permet de détecter des phénomènes et des structures de **dépendances complexes**. Elle pourrait en outre **s'adapter** et détecter de nouveaux schémas.

Généralement on distingue apprentissage supervisé, pour lequel on dispose de valeurs de sortie, versus apprentissage non supervisés, dans lequel on modélise toutes les observations pour mieux les comprendre.

L'apprentissage statistique cherche à construire des modèles qui, bâtis sur des données d'entrainement, sont capables d'effectuer des prédictions sur de nouvelles données. Cela veut dire que le modèle construit ne doit pas être trop ajusté aux données d'entrainement (**surapprentissage** ou overfitting) ni insuffisamment ajusté (**sous apprentissage** ou underfitting) - le deuxième cas étant assez facile à détecter.

Afin d'évaluer les performances des modèles utilisés, les données ont été divisées en 3 sous-ensembles : données d'apprentissage, données de validation (pour sélectionner variables et optimiser les paramètres) et données de test.

Dans le cas de la détection de la fraude, les **données sont particulièrement déséquilibrées** entre la classe majorité (sans fraude) et la classe très minoritaire à détecter (fraude). On doit donc avoir le nombre le plus élevé de Vrais Positifs, et le plus bas de Faux Positifs (coût supplémentaire d'investigation) et Faux Négatifs (fraude non détectée). Pour évaluer les réelles performances d'un modèle, on peut utiliser :

- **ROC AUC :** Cette mesure nous donne une première idée de l'efficacité du modèle à distinguer les Fraudes (taux de Vrais Positifs) sur les fausses alertes (taux de Faux Positifs)
- **F1 score :** Cet indicateur permet de comprendre le ratio de vraies fraudes (Vrais Positifs) par rapport aux fausses alertes (Faux Positifs) et les fraudes non détectées (Faux Négatifs)
- **Logloss :** Cette mesure donne une vision plus fine de la distance entre les prédictions du modèle et la réalité.

Même si les **arbres de décisions** fonctionnent bien sur de grosses bases de données et sont flexibles sur la qualité des données, ils connaissent de fortes variances en fonction de l'échantillon utilisé et le risque de surapprentissage est particulièrement important. Les **forêts aléatoires** permettent de réduire la variance liée à l'échantillon, tout en gardant cette capacité à détecter des interactions complexes entre les données. Il demeure un risque de surapprentissage et une interprétation plus difficile que pour les arbres de décision. Les **k-mean** sont aussi sensibles aux données utilisées pour leur apprentissage, plus que les **modèles naïfs bayésiens**, mais ces derniers sont moins facilement interprétables.

Les **arbres de décisions, forêts aléatoires et XG Boost** apparaissent les modèles les plus adaptés à la détection de la fraude.

Ce mémoire se différencie des écrits précédents sur le sujet par son application non seulement par le contexte du marché indien d'assurance, mais surtout par **des données initiales de faible qualité**. Les données ont dû faire l'objet de nombreux traitements et discussions avec les équipes en charge du traitement des sinistres : revue des doublons, des valeurs manquantes, des valeurs extrêmes et une normalisation des valeurs.

Apres implémentation, seules les **forêts aléatoires présentent un intérêt métier, et ce, uniquement sur les données des sinistres en cours de traitement.** Les modèles non supervisés utilisés n'ont pu permettre d'améliorer la détection de la fraude sur nos données.

Depuis la fin de ces travaux, les ressources disponibles en intelligence artificielle générative ont fortement augmenté mais sont aussi devenues accessible à un plus large public, notamment grâce à **Chatgpt,** lancé en novembre 2022. Même si cet outil n'est pas disponible en Chine, il est accessible en Inde et dans de nombreux pays. Pour la lutte contre la fraude, cette dynamique est très ambivalente : **les fraudeurs ont maintenant accès à des outils leur permettant de réaliser de fausses preuves très difficiles à détecter** pour un humain[509]. En même temps, les assureurs peuvent utiliser les capacités de ces nouveaux outils pour **mieux détecter la fraude**, en analysant textes et photos, mais aussi en accélérant les vérifications sur les réseaux sociaux[510].

[509] Artificial Intelligence and Insurance Fraud : Four dangers and four opportunities | RGA
[510] How developments in AI are impacting claims fraud : Clyde & Co

Bibliographie

Partie 1

BESSON, D. (1993) Les assurances dans les années quatre-vingt : l'envolée, Division Synthèse générale des Comptes Nationaux, Insee, Février 1993

DESSAL, R. (1989) Panorama de la distribution de l'assurance en Europe. In : Revue d'économie financière, n°11, 1989. L'assurance européenne : la grande transformation. pp. 261-275.doi : 10.3406/ecofi.1989.1660

DICKINSON, G. (1988) Marchés de l'assurance communautaires : Perspective 1992. In : *Revue d'économie financière*, n°4. Articles divers

DUMAZEDIER J., DE GISOR H. (1984) Français analphabètes ou illettrés ? *Revue française de pédagogie*, volume 69, pp. 15 DOI : https://doi.org/10.3406/rfp.1984.1556

INSEE Références (2018) - Fiches - L'école et ses sortants, page 71

GAUR S., RAO N. S. (2020), Poverty Measurement in India: a status update, Ministry of Rural Development.

GOUGENHAIM JH., VANDIER J. et STAUSS A. (2013) « L'évolution du marché de l'assurance automobile en France depuis la guerre et le rôle des mutuelles sans intermédiaires », *Entreprises et histoire*, 2013/3 (n° 72), p. 91-112. DOI : 10.3917/eh.072.0091.

PRADIER, P.C. (2020). La libéralisation des assurances, *Revue d'économie financière*, N. 137

RENAUDAT, E. (1989) La consommation domestique de 1950 à 1980. In : Recherches et Prévisions, n°18-19, Décembre – mars 1989. Activité féminine - Vie familiale. pp. 23-25

SCHOR, A.D. (1994) Revue d'économie industrielle, N.70 4eme trimestre 1994, Concentration, contestabilité et concurrence sur le marché français de l'assurance, Page 18

SINGH, S. K. (2012) Urban Transport in India: Issues, Challenges, and the Way Forward, Indian Institute of Management Lucknow, Lucknow – 226013, India. page 12 table 7

TEYSSIER, R. (1988) Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45.
Doi : 10.3406/ecofi.1988.1545 http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545, page 37

THIRY, G. (1979) Les risques d'accidents de la route. In : Economie et statistique, n°117, Décembre 1979. Emploi agricole / Les risques d'accidents de la route / Téléphone / Les revenus en 1975 / La circulation de l'information. pp. 13-22

WOLKOWITSCH, M. (1989) L'assurance en France. *Annales de Géographie*, t. 98, n°550, 1989. pp. 655

Partie 2

Articles

BOROOAH, V. (2016), Deconstructing Corruption: A Study of Cash for Favours in Rural India, March 2016, *Journal of South Asian Development 11(1)*, DOI:10.1177/0973174115624773

BUTTAN C. K., DWIVEDI S. (2020), Insurance and Fraud Management: Trends, Opportunities and Challenges, Dr. C. K. Buttan, Sanjiv Dwivedi, *The Journal of insurance institute of India*, July – September 2020, page 37

GANAPATHY, V. (2020) Combating Insurance Frauds Through Analytics, *The Journal of insurance institute of India,* July – September 2020, page 41

GOLDFARB, R. (2013) Bernie Madoff and the Honey Trap (Revised), foodliabilitylaw.com, 27 septembre 2013

JAIN, A. (2020), Insurance Frauds: A Closer Look, *The Journal of insurance institute of India*, July – September 2020, page 48

TEYSSIER, R. (1988) Réflexions sur les rentabilités de l'assurance. In : Revue d'économie financière, n°7, 1988. Articles divers. pp. 29-45.
Doi : 10.3406/ecofi.1988.1545 http://www.persee.fr/doc/ecofi_0987-3368_1988_num_7_3_1545, page 37

VENKATESH, S. (2020) Insurance Frauds Across the Globe: Lessons Learnt, *The Journal of Insurance Institute of India*. July – septembre 2020, page 27

Mémoire

KARSENTY, J. (2016) La détection des fraudes à l'assurance, *Mémoire d'actuariat*, ENSAE ParisTech

Partie 3

ARLOT, S. (2017) Validation croisée.

BARHOUMI, K. (2020) Modélisation de la fraude en assurance IARD : Application `a l'assurance automobile, Mémoire d'actuariat, ISFA

BREIMAN L., FRIEDMAN J., OLSHEN R.A. et STONE C. J. (1984) Classification and Regression Tree, New York

CARUANA M. A., GRECH L. (2021) Automobile Insurance Fraud Detection, *Communications in Statistics Case Studies Data Analysis and Applications*, Octobre 2021

CHARPENTIER A., FLACHAIRE E. et LY A. (2017) « Économétrie & Machine Learning »

GENUER R., POGGI J-M. (2017) Arbres CART et Forêts aléatoires, Importance et sélection de variables, page 5

GUPTA R. Y., MUDIGONDA S. S. et BARUAH P. K. (2021) TGANs with Machine Learning Models in Automobile Insurance Fraud Detection and Comparative Study with Other Data Imbalance Techniques, *International Journal of Recent Technology and Engineering (IJRTE)*, Février 2021

HANSEN L. K., LARSEN J. (1996). Unsupervised Learning and Generalization. In Proceedings of IEEE International Conference on Neural Networks (pp. 25-30). IEEE.

JENI L. A., COHN J. F. et DE LA TORRE F. (2013) Facing Imbalanced Data--Recommendations for the Use of Performance Metrics," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 245-251, Doi: 10.1109/ACII.2013.47.

KELLEHER J., NAMEE B. M. et D'ARCY A. (2015) Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies, MIT Edition

MAUMY-BERTRAND M., SAPORTA G. et THOMAS-AGNAN C. (2018) Apprentissage statistique et données massives, Editions Technip, 2018, page 28

METZLER G. (2019) Learning from imbalanced data: an application to bank fraud detection. *Cryptography and Security* [cs.CR]. Université de Lyon, 2019.

PANIGRAHI S., PALKAR B. (2018) Comparative Analysis on Classification Algorithms of Auto-Insurance Fraud Detection based on Feature Selection Algorithms. *International Journal of Computer Sciences and Engineering*, septembre 2018

ROSSI, F. (2008) Introduction à l'apprentissage statistique, Janvier 2008, page 8

ROUSSEEUW, P. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics,* Volume 20, 1987, Pages 53-65, ISSN 0377-0427,

TORRENS-IBERN, J. (1956) Qu'est-ce que la statistique ? *Journal de la société statistique de Paris*, tome 97, p. 289-293

SUNDARKUMAR G. G., VADLAMANI R. (2015) A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence*, N.37

VAPNIK, V. N. (1995) The Nature of Statistical learning theory N.Y: Springer,314 p.
BREIMAN, L. (2001) Statistical Modeling: The Two Cultures, *Statistical Science 2001*, Vol. 16, No. 3, 199–231

VIAENE S., DERRIG R. A. & DEDENE G. (2004) A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis, *IEEE Transactions on Knowledge and Data Engineering, June 2004*

Mémoire
DECOURTIEUX C. (2016) Modélisation de la fraude sur les contrats Auto du particulier par apprentissage statistique, *Mémoire d'actuariat*, EURIA

Partie 4

ANESTHESIOL, K. J. (2013) The prevention and handling of the missing data: pp402–406.

CHEN, T.& GUESTRIN, C. (2016) XGBoost: A Scalable Tree Boosting System, 10 Juin 2016, Page 4

DE VEAUX R., HAND D. J. (2005) How to Lie with Bad Data, *Statistical Science,* August 2005

ENDERS, C. K. (2010), Applied Missing Data Analysis, Applied missing data analysis (1st ed.). New York: The Guildford Press, page 39

JOHNSON, K. & KUHN, M. (2013) Applied Predictive Modeling, Page 488, 2013.

KAZIL J., JARMUL K. (2016) Data Wrangling with Python, O'Reilly media.

KELLEHER J. D., NAMEE B. M. & D'ARCY A. (2015) Fundamentals of Machine learning for predictive data analytics, MIT Editions

LIU F. T., TIN K. M. & ZHOU Z. H. (2012) Isolation-based Anomaly Detection, *CM Transactions on Knowledge Discovery from Data*, March 2012

LUCAS, Y. (2019) Credit card fraud detection using machine learning with integration of contextual knowledge. Artificial Intelligence [cs.AI]. Université de Lyon ; Universität Passau (Deutscheland), 2019.

MEYER, N. (2010) Statistique descriptive, *Laboratoire de Biostatistique et Informatique Médicale*, Fac. De Médecine de Strasbourg, novembre 2010

OSBORNE, J. W. & OVERBAY A. (2004). The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research, 9(6)

RUBIN, D. B. (1976) Inference and missing data, *Biometrika*, pp. 581-92

TUFFERY, S. (2005) Améliorer les performances d'un modèle prédictif : perspectives et réalité.

UEDA, T. (2009) A simple method for the detection of outliers, Mitsubishi Electric Computer Systems Corporation, Tokyo, 2009

WITTEN I. H., FRANK E., HALL M. A., PAL J. C. (2016) Data Mining: Practical Machine Learning Tools and Techniques, 4th edition, Morgan Kaufmann, Page 296

Mémoire

IKRAM, C. (2014) Sélection des Echantillons pour le Problème de classification en Distributions déséquilibrées, Ecole nationale des sciences appliquées de Tanger, *Thèse de doctorat en science et techniques de l'ingénieur*, soutenue le 25 juillet 2014

Partie 5

CARUANA M. A., GRECH L. (2021) Automobile Insurance Fraud Detection, *Communications in Statistics Case Studies Data Analysis and Applications*, Octobre 2021

PARKIN, N. (2023) Artificial Intelligence and Insurance Fraud : Four dangers and four opportunities, RGA

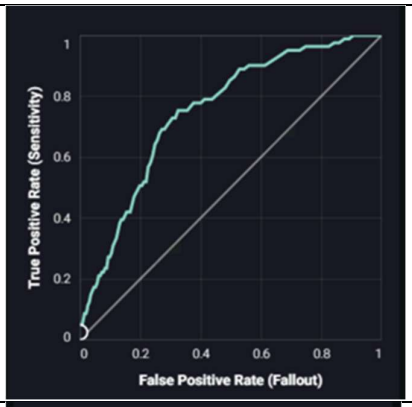
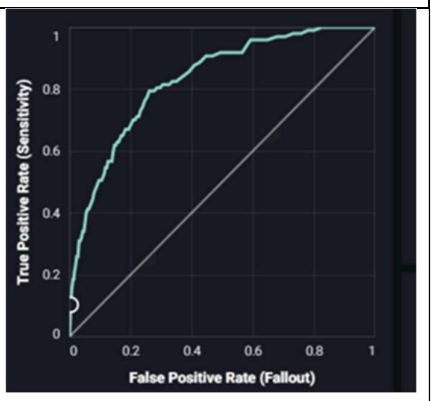
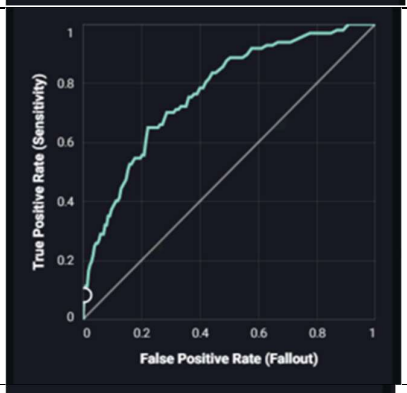
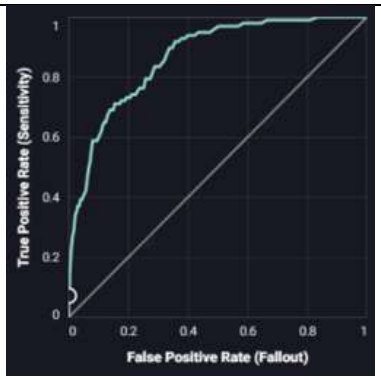
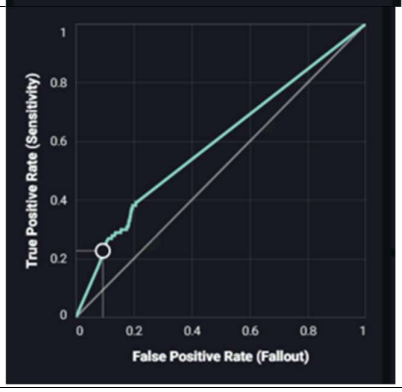
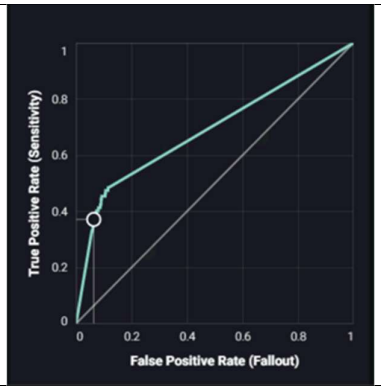
ROURKE, D. (2024) How developments in AI are impacting claims fraud, *Claims magazine*, Clyde & Co

Annexes

1. Courbes ROC

2. Impact des caractéristiques

1. Courbes ROC

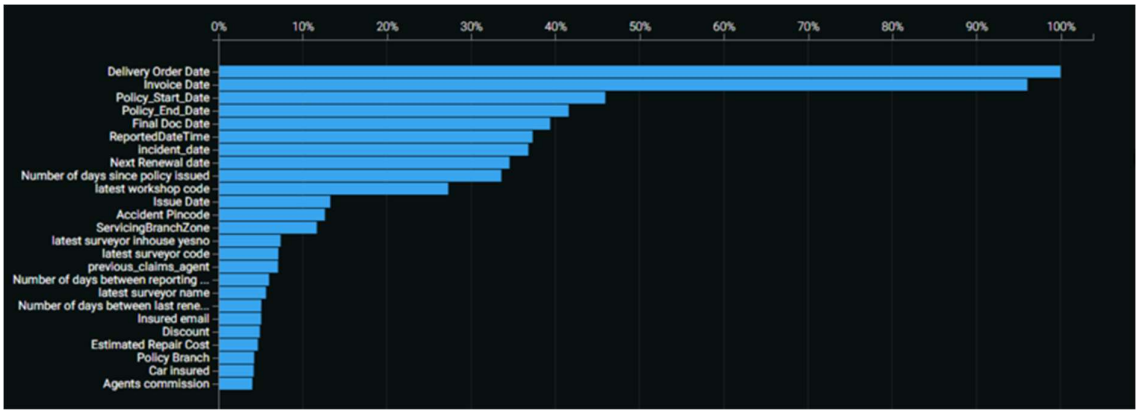
	Courbes ROC des enregistrements sans l'inspecteur et le garage	Courbes ROC des enregistrements durant le traitement du sinistre
Forêt aléatoire		
Forêt aléatoire XG Boost		
Naïf bayésien		

2. Impact des caractéristiques

a. Forêt aléatoire XGBoost

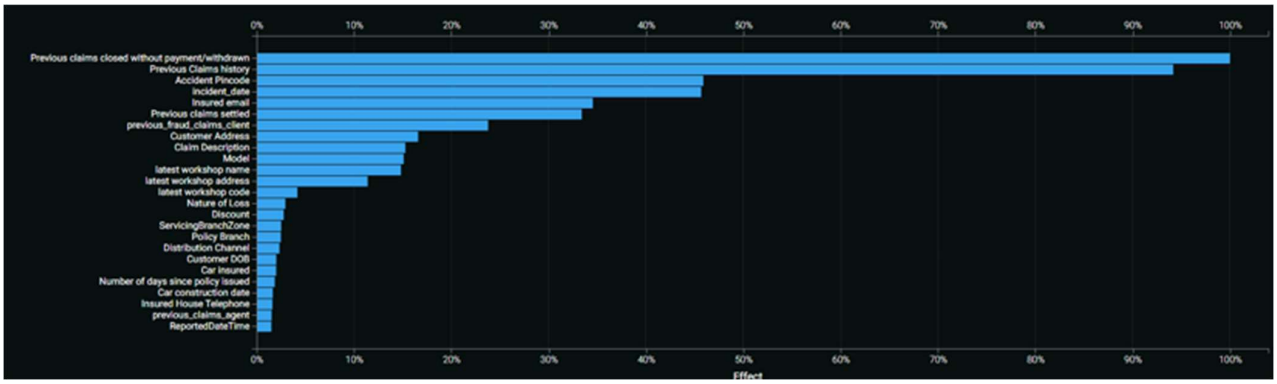
Avec inspecteur et le garage

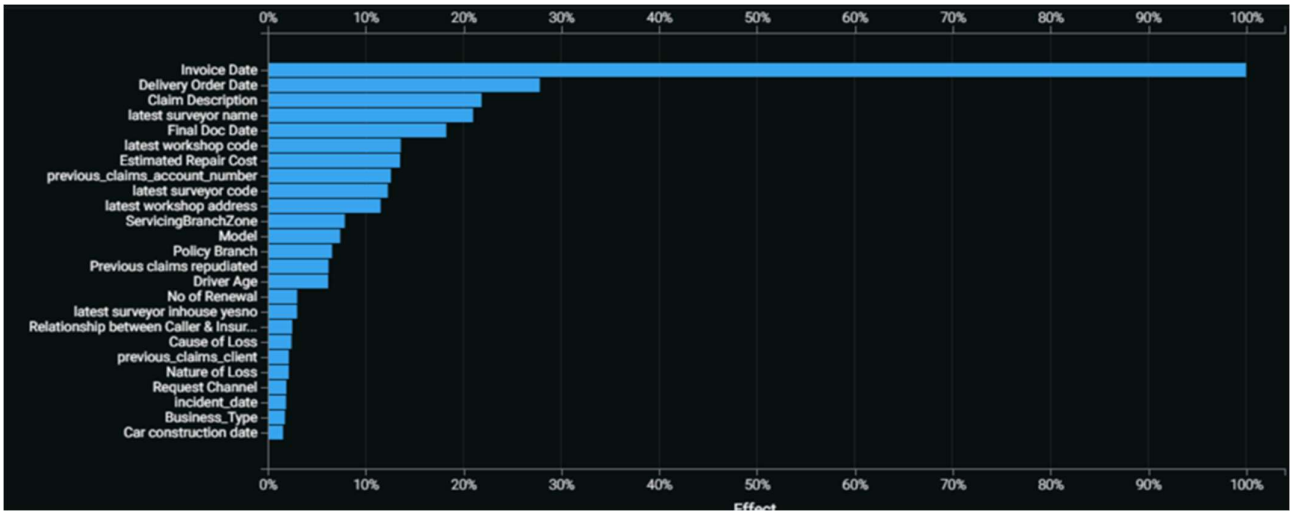
Durant le traitement de la déclaration de sinistres



b. Mahalanobis

Durant le traitement de la déclaration de sinistres



c. Naïf bayésien

Durant le traitement de la déclaration de sinistres