



PROJET BIG DATA

Rapport

RESUME

Identification de phénomène d'anti sélection en assurance automobile, par l'utilisation de modèles de Machine Learning sur deux bases de données normalisées. L'une venant d'un portefeuille d'assurance et l'autre venant du gouvernement.

Soline BLOCH

Table des matières

I- Introduction	1
II- Le portefeuille d'assurance automobile.....	2
Présentation de la base de données	2
Etude préliminaires sur le portefeuille d'assurance.....	2
Machine Learning sur le portefeuille	4
Les K Nearest Neighbors.....	4
L'arbre de décision	5
La régression logistique	6
L'importance de la variable âge	6
III- Croisement des données.....	6
Présentation et retraitement de la base de données véhicules 2021	7
Comparabilité des bases et utilisation	7
Seuil accident grave à 1300 euros	8
Seuil en pourcentage des accidents graves dans les données gouvernementales.....	8
IV- Conclusion.....	10
Annexe.....	11

I- Introduction

L'assurance automobile est un produit d'assurance très encadré et obligatoire dans la plupart des pays du monde. Cette assurance permet de se protéger notamment en responsabilité civile automobile, c'est-à-dire se protéger en cas d'accident où un tiers serait victime. La responsabilité civile automobile en général et surtout en France causent les sinistres les plus coûteux pour l'assureur et pour cause l'assureur est tenu d'indemniser intégralement la victime depuis la loi Badinter de 1985 (rente s'il n'est plus capable de travailler, tierce personne...).

Les assureurs ont une vision des événements et des risques via leur portefeuille. Cette vision des risques peut être biaisé par plusieurs phénomènes et notamment par le risque d'antisélection. L'antisélection pourrait être résumer comme le fait d'attirer de mauvais risques à cause d'une tarification qui ne reflète pas suffisamment le niveau de risque de l'individu. C'est pour cela que croiser ses données avec des données plus globale est un exercice intéressant qui peut aider à savoir si l'assurance est sujette d'antisélection.

Couramment en assurance automobile la variable affectant le plus la tarification est l'âge, c'est pour cela que nous nous proposons dans ce projet d'étudier un portefeuille d'assurance auto via des techniques de Machine Learning, puis de croiser nos données avec des données du gouvernement pour vérifier si ce portefeuille subit un phénomène d'antisélection.

II- Le portefeuille d'assurance automobile

Présentation de la base de données

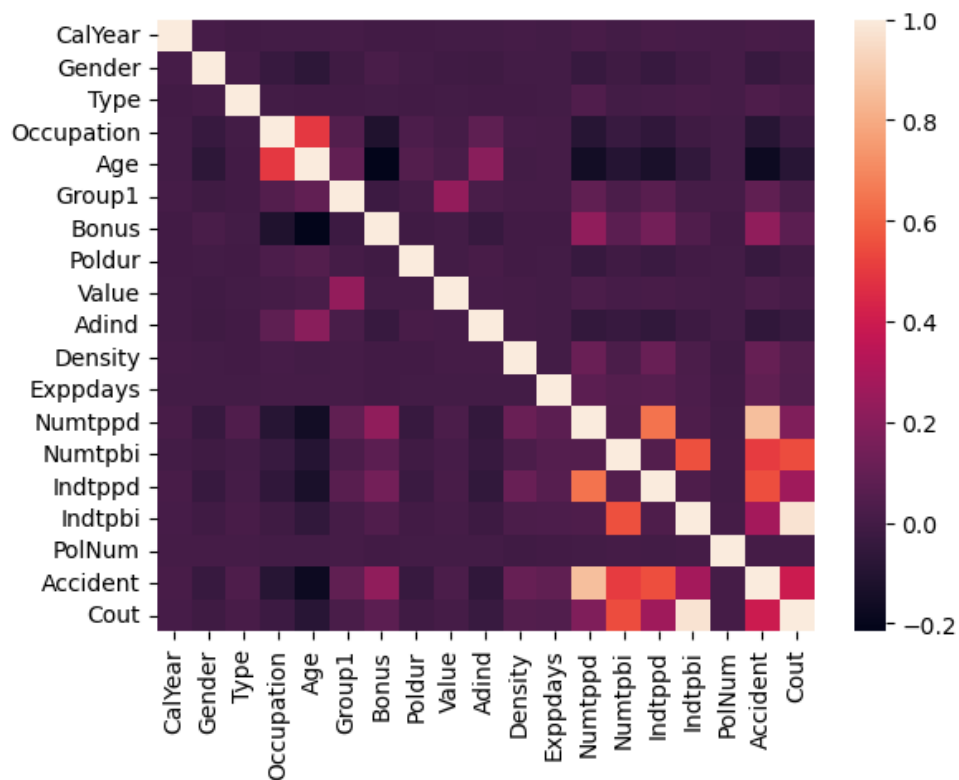
Le portefeuille d'assurance que nous avons a été trouvé sur Kaggle (https://www.kaggle.com/competitions/machinelearningesilv/overview?fbclid=IwAR0yWRKxBYGtQC YsrdDI4m2s9kDoNU9SxuFRRbMt-XRROrYj8He_PmQ-POQ) et nous savons qu'il a été donné par le professeur Arthur Charpentier. Cependant nous ne sommes pas sûr de quel pays les données viennent mais étant donné que les assurés ont tous au moins 18 ans nous supposons que ce sont des données européennes. Cette base de données contient diverses informations sur 80 000 assurés qui ont souscrit en 2009 et 2010 :

- Sur l'assuré (âge, genre, occupation dans la vie et densité de population dans sa ville)
- La voiture (Type (allant de A à F et décrivant citadine, mini...), la catégorie qui reprend la variable type en moins précisé avec juste 3 modalités (Large, Medium, Small) ainsi que la valeur du véhicule)
- Des informations sur le contrat (nombre de jour d'exposition, année de souscription, bonus ou malus, l'ajout d'une garantie dommage matériel)
- Des informations sur les sinistres : nous savons si un sinistre à un tiers à eu lieu corporel ou matériel ainsi que son coût

Etude préliminaires sur le portefeuille d'assurance

Afin de repérer un lien entre nos données nous réalisons une matrice de corrélation en ayant préalablement retraité nos variables catégorielles en variables quantitatives (par exemple pour le genre nous avons les modalités « F » ou « M » nous les remplaçons respectivement par 1 et 0). Dans notre base de données nous avons les variables Numtppd et Numtpbi qui sont respectivement le nombre de sinistre matériel au tiers et corporel au tiers ainsi que les variables Indtppd et Indtpbi qui représentent leur coût. Nous sommes donc ces variables pour obtenir le nombre d'accident globale que nous appellerons « Accident » et le cout des accidents que nous appellerons « Cout ».

Matrice de corrélation de nos données de portefeuille

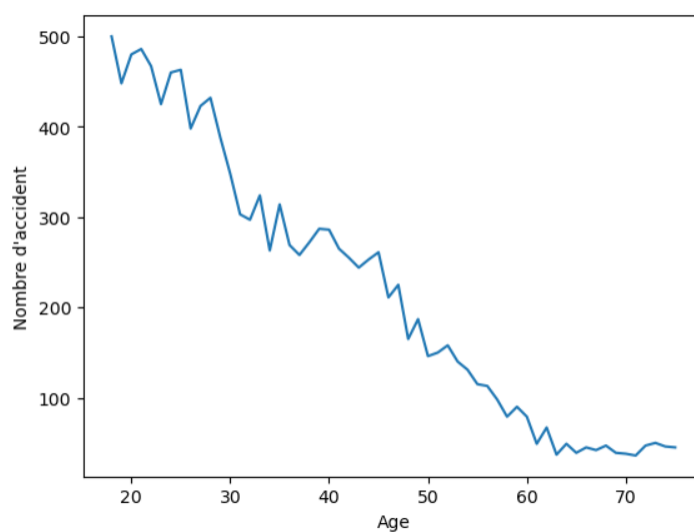


Nous obtenons la matrice de corrélation suivante, les corrélations que nous voyons en bas à droite sont normales puisque les variables Accident et Cout en sommant els autres variable. Nous remarquons une corrélation entre l'âge et l'occupation dans la vie qui est logique au vu des modalités que prend la variable occupation (les jeunes sont le plus souvent étudiants et les personnes âgées sont souvent retraités).

Un chiffre intéressant est qu'il y a 15,79% de notre portefeuille qui a eu un accident

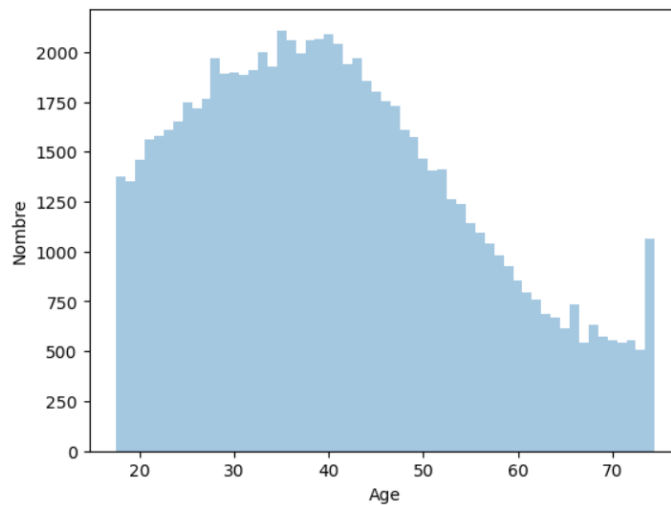
Nous voulions vérifier si l'âge avait réellement un impact sur le nombre d'accident

Graphique du nombre d'accident en fonction de l'âge



Sur le graphique suivant nous voyons clairement que les âges jeunes sont plus accidentogène que les âges plus élevés. Comparons par rapport à la répartition des âges dans notre portefeuille

Histogramme de la répartition des âges dans le portefeuille



Nous remarquons que nous n'avons pas particulièrement beaucoup de jeune dans le portefeuille, en effet le pic des âges est plutôt atteint aux-alentours des 40 ans. Nous pouvons donc dire que l'âge semble jouer un rôle dans le fait d'avoir des accidents.

Machine Learning sur le portefeuille

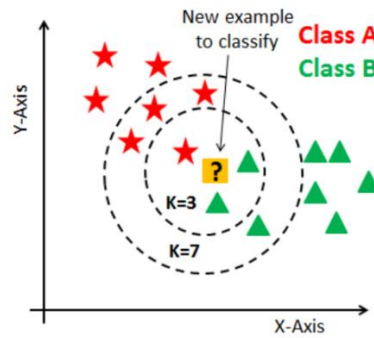
Afin de vérifier nos points précédents nous allons tenter de développer des techniques de Machine Learning pour classer les assurés qui aurait un accident et peut-être comprendre l'importance de l'âge dans les accidents. Nous avons décidé de se placer en situation où nous aurions un potentiel client qui nous appellerait et nous donnerait très peu d'information et que nous souhaiterions savoir si c'est un bon ou un mauvais risque. Les informations qu'il nous donnerait serait les suivantes : son âge, son genre, le type de voiture qu'il a (de A à F) et son occupation dans la vie.

Dans tous les modèles qui seront présentés nous aurons séparer notre portefeuille en deux. 80% de notre portefeuille nous permettra de calibrer notre modèle et 20% de le tester. C'est-à-dire que pour les 20% de tests nous demandons à notre modèle d'indiquer si la personne a eu un accident ou non et nous comparons par rapport au vrai résultat que nous connaissons.

Les K Nearest Neighbors

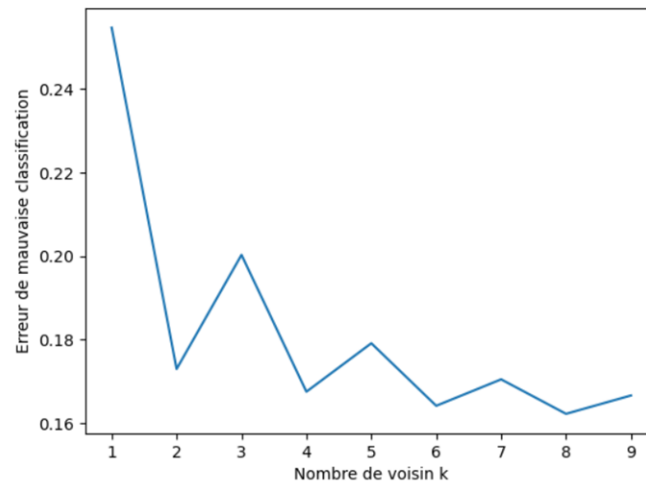
Le KNN ou K Nearest Neighbors est une technique de Machine Learning supervisé où l'algorithme va assigner une classe à chaque individu du test en regardant la classe de ces k voisins les plus proches et en assignant la classe majoritaire de ces voisins.

Schéma explicatif KNN



En fonction du nombre de voisin choisit le résultat peut varier dans le schéma ci-dessus nous remarquons qu'en choisissant $k=3$ la classe majoritaire sera la classe B alors qu'en choisissant $k=7$ la classe majoritaire sera la classe A. Pour bien choisir notre nombre de voisin nous avons réalisé une boucle qui teste pour différents k et nous choisissons le k qui minimise l'erreur de prédiction

Graphique de l'erreur de classification en fonction du nombre de voisin considéré

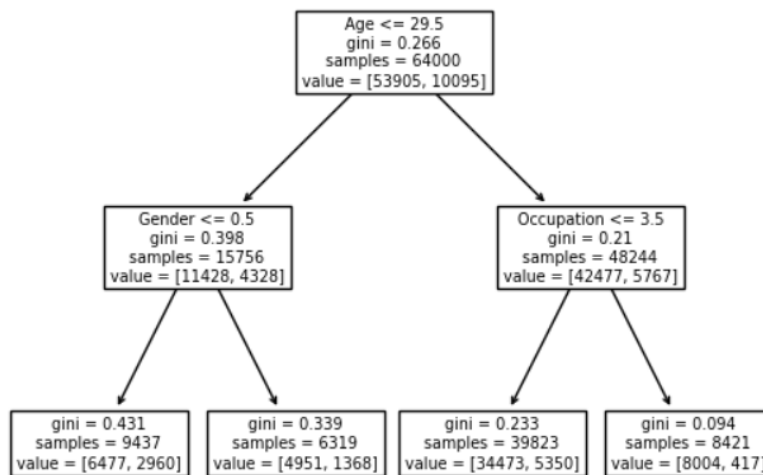


Le nombre idéal de voisin dans notre cas est donc de 8 et ce modèle nous donne un score juste dans 83,7% des cas.

L'arbre de décision

L'arbre de décision est une technique de machine Learning qui permet de classer en plusieurs catégories des données. L'arbre de décision choisit à chaque fois un critère le plus discriminant pour séparer les données et hiérarchise ces critères, cependant avec cette technique le risque de surapprentissage, c'est-à-dire le fait d'être trop adapté aux données d'entraînement et avoir une précision moyenne sur les données test, est élevé.

Arbre de décision sur les données du portefeuille



L'arbre de décision que nous avons réalisé ci-dessus sur les données d'entraînement nous indique que le premier nœud, critère discriminant est l'âge. Nous remarquons cependant qu'avec cette profondeur de 2 nœuds nous n'arrivons pas à avoir de feuille pure c'est-à-dire à avoir une feuille qui contient seulement des personnes accidentées ou non accidentées. Cela montre qu'il reste toujours une part de hasard sans le fait d'avoir un accident. L'indice Gini mesure la probabilité qu'un individu choisit au hasard dans le nœud soit mal classé. Plus l'indice est de zéro plus cela signifie que les individus sont bien classés.

L'arbre de décision nous donne un score de 84,15% c'est légèrement plus élevé que le KNN mais cela reste le même ordre de grandeur, nous constatons que dans le premier nœud qui prend comme critère l'âge nous avons un indice gini de 0,266 cela signifie que rien qu'avec l'âge nous sommes capables de prédire si la personne aura un accident.

La régression logistique

La régression logistique est une technique de base en statistiques et en Machine Learning, elle est utilisée lorsque la variable cible est binaire, une fonction sigmoïde est créée à partir de coefficient des variables explicatives. C'est une méthode plutôt simple qui ne demande pas trop de puissance de calcul. Nous avons obtenu un score de 84% également.

L'importance de la variable âge

Suite aux observations que nous avons pu faire grâce au graphique du nombre d'accident en fonction de l'âge et à l'arbre de décision nous avons décidé de tester les modèles en ne gardant que l'âge et le genre comme variables explicatives. Les résultats obtenus avec l'arbre de décision et la régression logistique dans ce cas sont très semblables au cas où nous avons toutes les variables explicatives.

III- Croisement des données

Un portefeuille d'assurance nous donne un certain nombre d'information mais il peut être biaisé par différents phénomène comme l'antisélection que nous avons expliqué plus haut. Il se pourrait que le portefeuille que nous étudions ne mettent pas des tarifs suffisamment élevés pour les jeunes

conducteurs et attirent donc des mauvais risques. Pour avoir une vision globale des accidents de la route nous nous sommes rendus sur le site data.gouv (<https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2021/>)

Présentation et retraitement de la base de données véhicules 2021

La base de données utilisé pour croiser nos données provient de data.gouv c'est la base usager.2021 qui concentre des informations sur les accidents relevé par les forces de l'ordre en 2021 et remplissant au moins une des trois conditions suivantes :

- Implique au moins une victime
- Survient sur une voie publique ou privée, ouverte à la circulation
- Implique au moins un véhicule

La base usager comporte une ligne par individu et chaque individu est relié à l'accident par le numéro d'accident et par l'identifiant du véhicule.

Numero accident	Identifiant véhicule	Rôle dans l'accident (conducteur/ passager/ piéton)	Age	Sexe	Gravité
-----------------	----------------------	---	-----	------	---------

Nous transformons cette base afin d'avoir la vue suivante avec une ligne non pas par individus concernés par l'accident mais pour avoir une ligne par véhicule. Pour la gravité de l'accident nous prenons la gravité maximale des individus concernés par l'accident. La gravité de l'accident est une variable qui prend ses valeurs de 1 à 4 (1- Indemne, 2- Blessé léger, 3- Blessé hospitalisé, 4- Tué). Pour déterminer la gravité globale de l'accident nous prenons la gravité maximale cela nous indiquera si le sinistre sera onéreux.

Identifiant véhicule	Nombre de conducteur (toujours égale à 1)	Age du conducteur	Sexe du conducteur	Nombre de passager	Nombre de piéton	Gravité de l'accident
----------------------	---	-------------------	--------------------	--------------------	------------------	-----------------------

Comparabilité des bases et utilisation

Dans la base du gouvernement nous avons seulement des individus accidentés et des détails sur les accidents ainsi que leur niveau de gravité. Dans la base de notre portefeuille nous avons des individus accidentés et non accidentés nous ne sélectionnons que les individus accidentés il faut également noter que ce ne sont que les accidents qui ont concerné un tiers qui sont reportés. C'est pour cela, que nous avons gardé dans la base du gouvernement que les accidents où il y avait au moins un passager ou un piéton concerné.

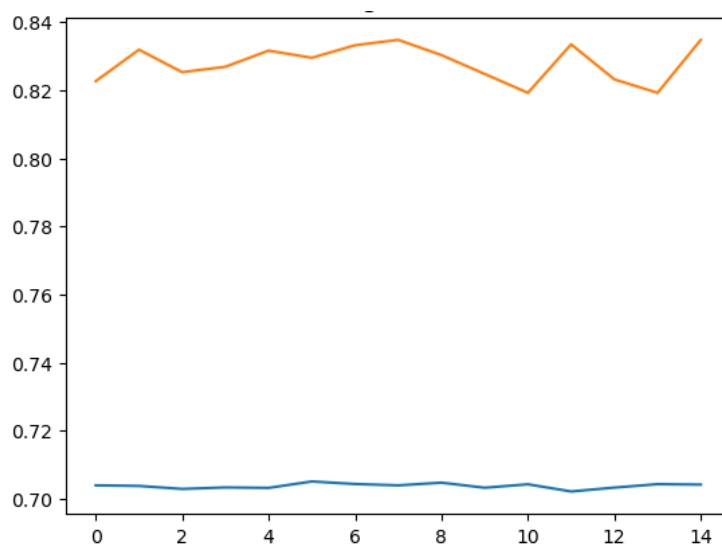
Etant donné que nos bases ne sont pas homogènes (une contient que des accidentés, et l'autre des accidentés et des non accidentés) nous décidons que nous nous concentrerons sur les sinistres graves qui sont ceux qui sont onéreux. Pour la base du gouvernement un sinistre est grave si au moins une des victimes a été blessé et hospitalisé (gravité niveau 3). Pour la base du portefeuille trouver cette limite est plus délicat nous pourrions procéder des deux manières différentes :

- Considérer qu'un sinistre est grave lorsque le coût associé dépasse le coût de 24h hospitalisation (le coût d'une nuit d'hospitalisation est d'environ 1300 euros en France nous choisirons donc ce seuil)
- Se référer au pourcentage d'accidents graves dans la base du gouvernement (ceux qui sont en gravité 3 ou 4) et prendre comme seuil sur la base du portefeuille le plus petit montant de ce pourcentage des accidents les plus onéreux.

Le but ici est non pas comme avant de développer un modèle permettant de prédire si un individu aurait un accident (puisque nos bases ne contiennent plus que des individus accidentés)

Seuil accident grave à 1300 euros

*Graphique des scores de précision de la régression logistique
orange : données test portefeuille, bleu : données gouvernementales*



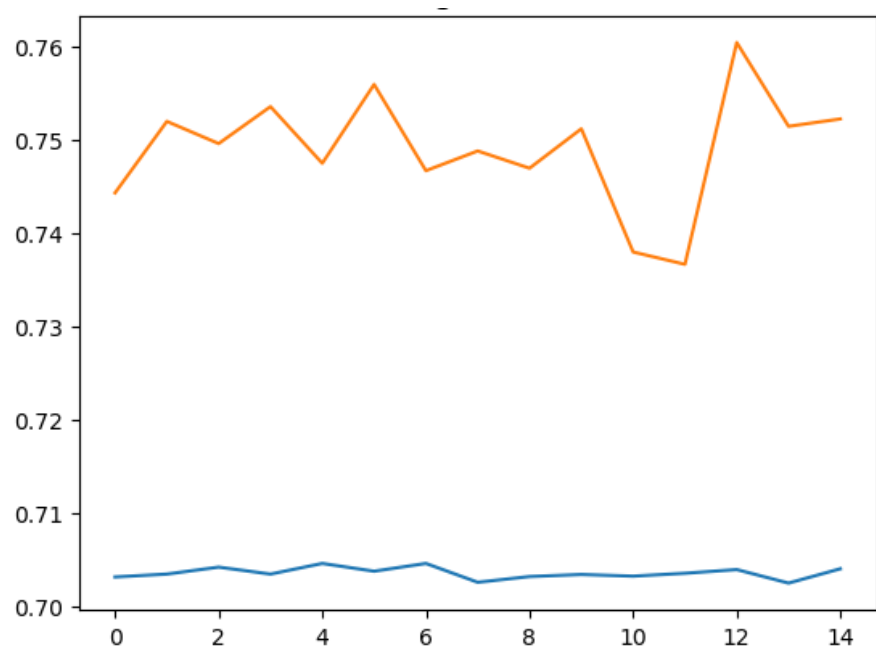
Nous prenons un seuil d'accident grave à 1300 euros et les identifions dans la base portefeuille. Puis nous divisons ces données pour avoir un set d'entraînement et un set de test. Nous entraînons le modèle sur les données d'entraînement de la base d'assurance puis nous testons notre modèle sur les données test du portefeuille et sur les données du gouvernement.

Sur les données du portefeuille le score est aux alentours de 82% de précision, alors que pour les données du gouvernement le score est aux alentours de 70% de précision. Cet écart de plus de 10% est significatif il pourrait être le signe d'anti sélection dans notre portefeuille.

Seuil en pourcentage des accidents graves dans les données gouvernementales

Nous avons testé cette méthode mais elle semble peu adéquate. Pour se faire nous avons calculer le ratio des accidents en catégorie 3 ou 4 dans la base gouvernementale sur l'ensemble des accidents. 29% des accidents de la base gouvernementale sont graves. Dans notre portefeuille nous rangeons par ordre croissant les coûts des accidents et nous sélectionnons la plus petite valeur faisant partie des 29% des accidents les plus graves. La valeur que nous avons trouvée était de 130. Nous retestons nos données en considérant comme grave sur la base du portefeuille les accidents ayant couté plus de 130 euros. Cependant cela semble incohérent puisque 130 euros n'est pas le montant d'un accident grave.

*Graphique des scores de précision de la régression logistique
orange : données test portefeuille, bleu : données gouvernementales*



L'écart est moins important, mais nous retenons que la méthode semble incohérente.

IV- Conclusion

Pour conclure ce projet avait pour but d'utiliser le Machine Learning afin de détecter l'anti sélection dans des portefeuilles d'assurance automobile. Lors de l'étude de notre portefeuille nous avons remarqué que l'âge jouait un rôle plus qu'essentiel dans la prédiction des accidents. Au point qu'en retirant les variables explicatives « Occupation » et « Type » de véhicule pour ne garder que les variable « Age » et « Genre » nous conservions le même niveau de précision du modèle. Nous avons donc décidé de croiser nos données avec celle du gouvernement en France trouvées sur data. Gouv. Nous nous intéressions plus au fait d'avoir un accident où non mais au fait d'avoir un accident grave.

La méthode où nous fixons un seuil à 1300 euros semble plus pertinente et réaliste que la seconde méthode. Nous trouvons un écart de plus de 10% ce qui serait le signe d'anti sélection.

Cependant cette étude comporte des limites et des axes d'amélioration. Au niveau de la base concernant le portefeuille d'assurance nous n'avons de détail sur la provenance des données et leur pays d'origine. Au niveau des données du gouvernement nous avons l'ensemble des accidents mais nous ne savons si le conducteur était toujours responsable de l'accident nous avons donc supposer qu'il l'était.

Une dernière approximation est le seuil à partir duquel on identifie un accident grave nous avons choisis 1300 euros qui est le coût d'une nuit d'hospitalisation cependant nous ne connaissons pas le coût des accidents en gravité 3 et 4 de la base gouvernementale et nous ne savons pas quel est le coût d'une nuit d'hospitalisation pour un assureur en moyenne sachant que la Sécurité Sociale prend en charge une partie.

Annexe

Vous retrouverez mon Jupyter notebook ainsi que mes bases de données utilisées pour ce rapport :

<https://github.com/SolineBI/Anti-s-lection-Machine-Learning>