Projet de Machine Learning

Sujet : Construction du Véhiculier en Assurance Auto

Réalisé par :
HAMMAMI Wejdene
MBAYE Serigne Bara
MENNESSIER Lyzie

Supervisé par :
Mr. KEZHAN Shi

Niveau : M2 Actuariat
Année Universitaire : 2024-2025

Table des matières

1 Introduction Générale

Dans un marché concurrentiel tel que celui de l'assurance automobile, il est impératif de disposer d'une tarification adéquate qui permet à l'assureur de se positionner efficacement sur le marché.
Un véhiculier fiable joue un rôle crucial dans cette démarche. Il lui permet d'évaluer les risques de manière précise, d'ajuster ses tarifs et de rester compétitif tout en garantissant une gestion optimale de son portefeuille.
Dans ce cadre, nous développons, lors de ce projet, un modèle prédictif qui nous permet de construire un véhiculier à dire machine. Ce véhiculier, nous sert à classer les véhicules du parc automobile en des groupes de risques homogènes en nous basant sur la probabilité de survenance de sinistres. Pour atteindre cet objectif, nous avons tout d'abord entrepris une étape cruciale de compréhension des données. Ensuite, nous avons élaboré deux modèles de Gradient Boosting Machine (GBM) ; le 1er vise à prédire le nombre de sinistre hors effet véhicules et le 2ème à prédire le nombre de sinistres résiduels (l'écart entre le nombre observé et le nombre prédit). L'étape suivante a consisté à construire un véhiculier en regroupant les prédictions du 2ème GBM en classes de risques homogènes via K-means. Enfin, nous avons procédé à l'évaluation de la performance du modèle. Cette évaluation nous a fourni une idée sur sa robustesse et son efficacité dans la classification des risques associés aux véhicules du parc automobile.

2 Méthodologie

2.1 Analyse des Données

Notre base de données est constituée de 105555 observations et de 30 variables décrivant à la fois la sinistralité et les caractéristiques des véhicules. Chaque véhicule est identifié par un contrat repéré par un "ID".
En analysant les données, nous constatons la présence de valeurs manquantes au niveau de certaines variables, comme l'en atteste la figure ci-dessous.

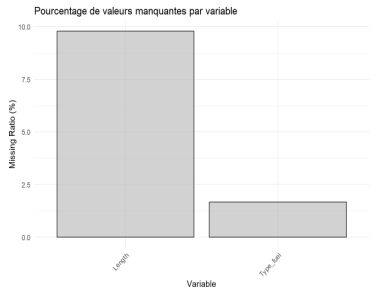

FIGURE 1 – Taux de Valeurs Manquantes par variable

Le taux de NA est faible, nous gardons alors toutes les variables.
Partant du fait que la classification des véhicules est basée sur la fréquence (nb de sinistres/exposition), et par soucis de simplicité, nous supposons que l'exposition est égale à 1 pour tous les contrats.

En faisant un zoom sur la variable N_claims_year, nous remarqueons que le nombre de sinistre observé varie entre 0 et 25.

Dans une première étape, nous isolons l'effet véhicule. Ceci consiste à modéliser le nombre de sinistres en fonction des variables qui ne caractérisent pas le véhicule. L'effet de véhicule sera donc incorporé dans les résidus du modèle.

2.2 Modélisation du Nombre de Sinistres hors Effet Véhicule par un GBM

Généralement, le nombre de sinistre est modélisé par une loi de poisson en utilisant des GLM. Cependant, l'évaluation de ce modèle révèle d'une surdispersion. Dans ce cas, le modèle de Poisson n'est pas approprié. Nous avons alors envisager d'autres alternatives en testant des modèles plus complexes comme le GBM.

Principe du modèle GBM :

Les modèles GBM sont des modèles d'apprentissage automatique ensemblistes qui fonctionnent de manière séquentielle en construisant un ensemble de modèles de prédictions faibles, qui se combinent pour former un modèle plus puissant.

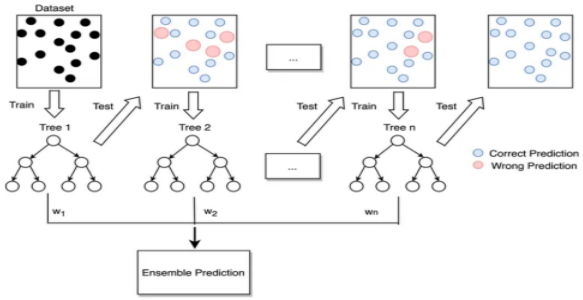

FIGURE 2 – Principe du Modèle GBM

Pour éviter le surajustement, ce modèle utilise souvent des techniques de régularisation telles que l'élagage des arbres, la réduction de leurs profondeurs, la diminution du taux d'apprentissage, etc. Ces techniques sont déterminées lors du process du réglage des hyperparamètres.

Nous avons sélectionné les variables non descriptives du véhicule et nous avons implémenté le modèle.

En observant le diagramme de l'importance des variables ci-dessous, nous remarquons que celles les plus discriminantes sont "Policies in force" correspondant au nombre total de polices détenues par l'assuré auprès de l'entité d'assurance pendant la période, "Premium", "Payment" représentant le mode de paiement.

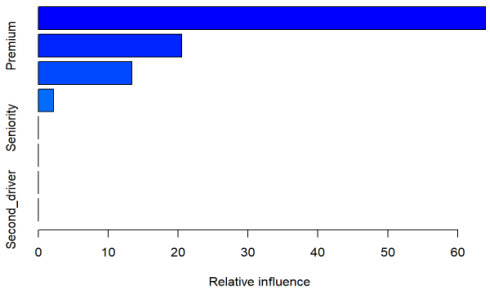

FIGURE 3 – Importance des variables

D'après les statistiques descriptives relatives à cette prédiction, nous notons que le nombre de sinistres prédit varie entre 0.3 et 1.43.

Min	**Q1**	**Q2**	**Mean**	**Q3**	**Max**
0.30	0.31	0.33	0.38	0.40	1.43

TABLEAU 1 – Statistiques descriptives du nb de sinistres prédits

La mesure de déviance, pour ce modèle, confirme l'absence de surdispersion. Une fois le nombre de sinistre estimé, nous calculons le nombre de sinistre résiduel tel que :

$$\text{Res\_diff} = \text{Nb sinistres observé} - \text{Nb sinistres modélisé HEV}$$

Cette variable est ajouté, par la suite, à la base de données.

Min	**Q1**	**Q2**	**Mean**	**Q3**	**Max**
-1.34	-0.39	-0.33	0.01	-0.30	24.31

TABLEAU 2 – Statistiques descriptives du nb de sinistres résiduel

Les statistiques récaptulés dans le tableau ci-dessus donnent un aperçu de la distribution de la variable `Res_diff`. La moyenne est différente de la médiane, cela suggère la présence de valeurs extrêmes dans la distribution. De plus, le 3ème quantile est inférieur à la moyenne : la distribution est étirée vers la droite par les valeurs extrêmes élevées.
Le boxplot, ci-dessous, illustre bien la présence d'outliers.

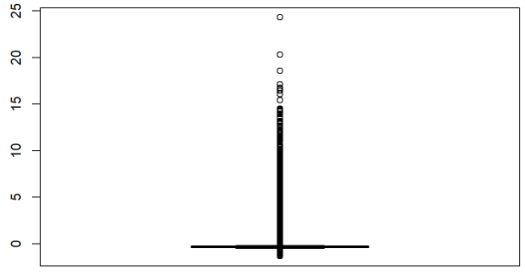

FIGURE 4 – Boxplot du nb de sinistre résiduel

De telles observations peuvent biaiser les résultats, il est donc primordial de les supprimer ou éventuellement de réduire leur taux.
Remarque : dans la partie modélisation, nous comptons tester le modèles avant et après suppression de ces valeurs.

2.3 Détection et Suppression des Outliers

Méthode 1 : IQR criterion

Ce critère consiste à déterminer l'écart entre le 1er et le 3ème quantile, noté IQR. Les observations considérées comme des valeurs aberrantes potentielles sont celles qui se situent en dehors de l'intervalle I=[Q1-1.5IQR ; Q3+1.5IQR].

D'après cette méthode, nous avons environ 24% des observations sont considérées comme outliers. Ce taux est élevé, nous testerons alors une autre approche.

Méthode 2 : Percentiles

Selon ce critère, les observations à l'extérieur de l'intervalle constitué par les deux percentiles P1 et P2 sont considérées comme aberrantes. Pour le choix de percentiles, nous pouvons tester plusieurs alternatives : les percentiles 2.5-97.5 ou les percentiles 1-99, etc. . .

Les résultats fournis par cette approche sont présentés dans le tableau ci-dessous :

	[P2.5 ; P97.5]	**[P1 ; P99]**
valeurs aberrantes (%)	4.99	1.78
Res_diff conservé	[-0.57 ; 3.25]	[-0.68 ; 4.66]

TABLEAU 3 – Récap des résultats de l'approche des percentiles

Pour les percentiles P2.5-P97.5, 5% des valeurs sont considérées comme aberrantes, et le nombre de sinistres résiduel varie entre -0.5 et 3.3, alors que pour les percentiles P1-P99, nous avons moins de valeurs aberrantes du coup l'intervalle du nombre de sinistres résiduel qui sera conservé est plus étendu. En faisant un compromis entre valeurs de seuils (bornes inf et sup de l'intervalle des observations non aberrantes) et taux de valeurs aberrantes, nous favorisons les résultats des percentiles P2.5-P97.5

Pour valider ce choix, nous nous appyons sur une méthode graphique issue de la TVE. Il s'agit de l'estimateur de Hill dont le seuil correspond au moment à partir duquel, cet estimateur ce stabilise.

Méthode 3 : HillPlot

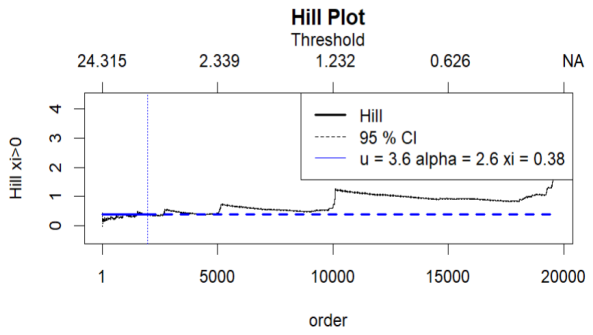

FIGURE 5 – Hillplot

D'après le Hillplot, le seuil est fixé à 3.6, cad au dela de cette valeur les observations sont considérées comme extremes. Cette valeur est plus proche de la borne sup du 1er intervalle (les percentiles 2.5 et 97.5, ce qui confirme le choix effectué.

2.4 Modélisation du Nombre de Sinistres Résiduels par un GBM

Ces modèles consistent à modéliser le nombre de sinistres résiduels en fonction des caractéristiques du véhicules et en passant lexposition en offset. Dans cette partie, nous présentons les résultats des modèles implémentés : des modèles dans lesquels nous intégrons la totalité des observations et d'autres suite à la suppression des valeurs extrêmes.

Avant Suppression des valuers aberrantes :

Les résultats issus de ce modèle réalisé en intégrant toutes les observations sont présentés dans le tableau ci-dessous :

ntrees	n_int_trees	size	min_dep	max_dep	mean_dep	min_leav	max_leav	mean_leav
50	50	20103	5	5	5	10	32	27.28

TABLEAU 4 – Caractéristiques du Modèle - Avant supression des valeurs extrêmes

Ce modèle construit à partir de 50 arbres, de profondeurs maximales 5 et admettant un nombre de feuilles entre 10 et 32.

En ce qui concerne les perforamances du modèle, le tableau ci-dessous montre que les erreurs train et test sont élevées, donc le modèle est incapable de capter la structure des données et de généraliser.

RMSE Train	RMSE Test
0.62	0.65

TABLEAU 5 – RMSE Train & Test - Avant supression des valeurs extrêmes

Ce problème de sous-apprentissage peut être dû à plusieurs facteurs à savoir la présence de valeurs extrêmes, le mauvais choix des hyperparamètres, etc...

Après Suppression des valuers aberrantes :

Le même modèle a été implémenté mais en éliminant les observations considérées comme extrêmes. Le nouveau modèle présente les caractéristiques suivantes :

ntrees	n_int_trees	size	min_dep	max_dep	mean_dep	min_leav	max_leav	mean_leav
50	50	20528	5	5	5	10	32	27.98

TABLEAU 6 – Caractéristiques du Modèle - Après supression des valeurs extrêmes

Ces caractéristiques sont légèrement différentes de celles du modèle précédent.

Pour le 2eme modèle, la suppression d'outliers a un impact sur les erreurs.

En effet, d'après le graphique ci-dessous, représentant les erreurs en fonction du nombre d'arbres, nous remarquons que le problème de sous-apprentissage a été controlé : les erreurs train et test ont considérablement diminué (0.27 sur le train et 0.28 sur le test.)

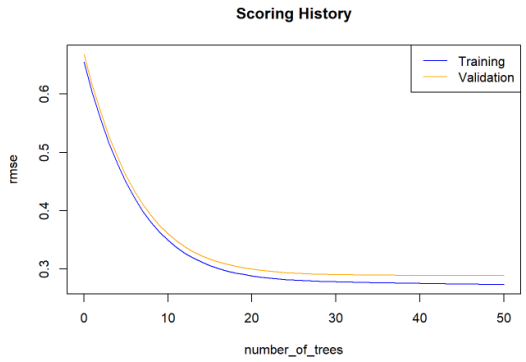

FIGURE 6 – RMSE Train vs RMSE Test du Modèle - Après Suppression des Valeurs Extrêmes

La comparaison entre les RMSE Train et Test des modèles avant et après suppression des valeurs aberrantes prouve bien l'apport de cette suppression : le modèle a bien appris les motifs présents dans les données d'entraînement et a une bonne capacité de généralisation.
Remarque : Pour le tuning du nombre d'arbres, nous choisissons, idéalement le nombre qui correspond au point de croisement des courbes des erreurs : c'est le seuil de passage de l'underfitting à l'overfitting. Dans notre cas, puisque nous n'arrivons pas à visualiser ce point, nous choisissons ntrees=20.
L'optimisation des hyperparamètres (tuning) se fait à l'aide d'une grille qui prend plusieurs valeurs pour chaque paramètre et tente de trouver la combinaison la plus optimale qui minimise les erreurs.
Les hyperparamètres optimaux pour ce modèle sont récapitulés dans le tableau suivant :

col_sample_rate	learn_rate	max_depth	sample_rate	ntrees
1.0	0.1	5	0.8	20

TABLEAU 7 – Hyperparamètres Optimaux du Modèle

Les erreurs train et test sont légèrement modifiés, ces mesures de performances, montrent que le tuning nous permet de calibrer un bon modèle qui s'ajuste bien aux données.
La commande h2o.varimp nous permet d'extraire la liste des variables les plus significatives.

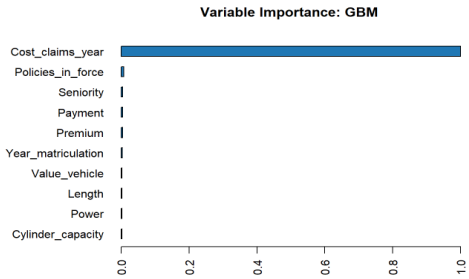

FIGURE 7 – Importance des variables

L'analyse de l'importance des variables du modèle tuné nous permet d'avoir un aperçu sur les variables les plus discriminantes. Parmi ces variables, nous pouvons citer le cout des sinistres, le nb de polices détenues, l'ancienneté de l'assuré,..

2.5 Construction du véhiculer

Dans cette partie, nous aurons recours à l'un des algorithmes de classification non supervisé qui est le kmeans pour regrouper les prédictions en classes de risques homogènes.

2.5.1 Principe de l'Algorithme de Kmeans

L'idée principale du K-means est de trouver des centres de cluster, encore centroïdes, de manière à minimiser la distance entre ces centres et les points de données attribués à chaque cluster. La classification issue de cet algorithme nous permet d'identifier des classes homogènes regroupant des observations en k clusters suivant leurs degrés de similarité repérée par la distance euclidienne ou la distance de Manhattan. Cet algorithme est sensible au nombre de clusters initialisé à k et peut converger vers différents résultats en fonction des centroïdes initiaux. Par conséquent, il est souvent recommandé d'optimiser ce choix de k. La méthode la plus reconnue pour ce choix est la méthode du coude : il suffit d'appliquer l'algorithme avec plusieurs valeurs de k et de calculer la variance des différents clusters. Le point du coude représente le nombre de clusters à partir duquel la variance ne se réduit plus significativement.

2.5.2 Mise en Apllication

La première étape consiste à estimer les prédictions en utlisant la commande h2o.predict prenant comme input le meilleur modèle issu du tuning et la base. Par la suite, nous faisons varier le nombre de clusters k entre 2 et 15 et nous traçons la courbe représentative de la variance intra-classe en fonction de k, nous obtenons alors la courbe d'Elbow représentée dans le graphe ci-dessous :

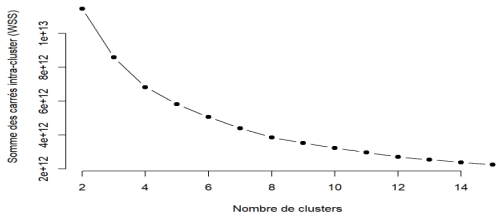

FIGURE 8 – Courbe d'Elbow pour le choix de k

Partant de ce résultat, nous appliquons l'algorithme K-means avec k=7 pour construire le véhiculier.
En analysant les résultats présentés au niveau du graphique ci-dessous, nous notons que la bande la plus étendue est la bande 2 : ce cluster représente le nombre de sinistres résiduels total le plus élevé, sauf qu'il est négatif puisque l'écart total entre les sinistres observés et ceux modélisés hors effet véhicule est négatif.

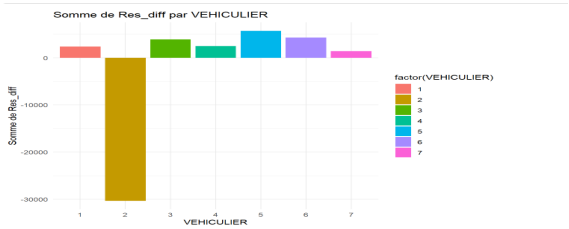

FIGURE 9 – Nb de sinistre résiduels total par cluster

3 Conclusion

En conclusion, ce projet visait à développer un modèle de véhiculier construit à partir des données réelles d'une assurance automobile. Ce véhiculier a été développé dans le but de classer les véhicules du parc automobile en classes de risques homogènes, en fonction de leur probabilité d'avoir un sinistre.
Dans le cadre de ce projet, deux modèles GBM ont été construits :
Le premier a été entraîné en prenant comme variable réponse le nombre de sinistres observés et en utilisant les données descriptives hors véhicule comme variables explicatives. Il a été utilisé pour prédire le nombre de sinistres de l'ensemble des données. Par la suite, le nombre de sinistres résiduels a été calculé.
Quant au deuxième modèle, il a été entraîné en prenant comme variable réponse le nombre de sinistres résiduels et en utilisant l'exposition comme variable offset. Les données descriptives du véhicule ont été utilisées comme variables explicatives. Ce modèle a été utilisé pour prédire le nombre de sinistres résiduels de l'ensemble des données.
Pour ce second modèle, une étude supplémentaire a été réalisée pour prouver l'apport de la suppression des valeurs aberrantes. Suite à ces études, un algorithme de Kmeans a été utilisé pour classer les prédictions en classes de risques homogènes, créant ainsi un véhiculier.

De telle étude permet aux compagnies d'assurance automobile, de mieux segmenter leurs portefeuille et de quantifier le risque associé à chaque véhicule assuré. En résumé, ce projet a permis de développer des outils efficaces pour la tarification des assurances automobiles, contribuant ainsi à une gestion plus précise et efficace du risque.

4 Références

[1] Adrien CONDAMIN. Construction de véhiculier et mise en perspective dans le cadre de tarification d'assurance automobile : Application sur les garanties bris-de-glace et vol. Master's thesis, 2020.

[2] Leslie GNANSOUNOU. Construction d'un véhiculier en assurance automobile à partir de méthodes de machine learning. Master's thesis, 2019.

[3] Julie LAVENU. Les méthodes de machine learning peuvent-elles être plus performantes que l'avis d'experts pour classer les véhicules par risque homogène ? Master's thesis, 2019.

[4] Matthieu QUILFEN. Classification des véhicules en assurance automobile. Master's thesis, 2017.

[5] Magali RUIMY. Elaboration d'un véhiculier en assurance automobile. Master's thesis, 2016.