

**Mémoire présenté devant l'ENSAE Paris  
pour l'obtention du diplôme de la filière Actuariat  
et l'admission à l'Institut des Actuaires  
le 09/11/2022**

Par : **Fouzia ELAOUNI**

Titre : **Modélisation de l'élasticité au prix des affaires  
nouvelles du produit Parc dénommé**

Confidentialité :  NON       OUI (Durée :  1 an     2 ans)

*Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus*

*Membres présents du jury de la filière*

*Entreprise : AXA France* 

*Nom : Christian-Yann ROBERT*

*Signature :* 

*Membres présents du jury de l'Institut  
des Actuaires*

*Directeur du mémoire en entreprise :*

*Nom : Amanda MARLET*

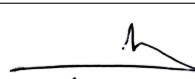
*Signature :*

**Autorisation de publication et de  
mise en ligne sur un site de  
diffusion de documents actuariels  
(après expiration de l'éventuel délai de  
confidentialité)**

*Signature du responsable entreprise*



*Signature du candidat*



Secrétariat :

Bibliothèque :

# Table des matières

<b>Note de confidentialité</b>	iii
<b>Remerciement</b>	iv
<b>Résumé</b>	v
<b>Abstract</b>	vi
<b>Note de synthèse</b>	vii
<b>Executive Summary</b>	xiii
<b>Introduction générale</b>	1
<b>1 Mise en contexte</b>	3
1.1 Présentation du périmètre de l'étude . . . . .	3
1.1.1 L'assurance de flotte automobile d'entreprise . . . . .	3
1.1.2 Le produit Parc dénommé d'AXA France . . . . .	5
1.2 Le phénomène d'antisélection . . . . .	7
1.2.1 Antisélection en assurance automobile . . . . .	7
1.2.2 Stratégie tarifaire et antisélection . . . . .	9
1.2.3 Contexte de l'inflation et risque d'antisélection . . . . .	11
<b>2 Présentation des données</b>	13
2.1 Construction de la base de données . . . . .	13
2.1.1 Présentation des bases utilisées . . . . .	13
2.1.2 Traitement des anomalies . . . . .	15
2.1.2.1 Les lignes anormales . . . . .	15
2.1.2.2 Devis doublons . . . . .	16
2.1.2.3 Devis de remplacement . . . . .	17
2.1.3 Création des variables . . . . .	17
2.1.4 Présentation de la base finale . . . . .	19
2.2 Étude exploratoire . . . . .	20
2.2.1 Analyse du taux de transformation . . . . .	20
2.2.2 L'analyse des corrélations . . . . .	23
<b>3 Modélisation du taux de transformation</b>	29
3.1 Cadre théorique . . . . .	29
3.1.1 Modèle linéaire généralisé (GLM) . . . . .	30
3.1.2 Le modèle Xgboost . . . . .	33
3.1.3 Métriques d'évaluation des modèles . . . . .	36
3.1.4 Sélection de variables à travers l'algorithme RFE . . . . .	38

3.1.5	Mesure de l'influence des variables . . . . .	39
3.2	Application pratique . . . . .	42
3.2.1	Préparation des données pour la modélisation . . . . .	42
3.2.2	Optimisation des modèles . . . . .	43
3.2.3	Résultats . . . . .	46
3.2.3.1	Évaluation des modèles . . . . .	46
3.2.3.2	Résultat du modèle GLM . . . . .	46
3.2.3.3	Résultat du modèle Xgboost . . . . .	48
3.2.3.4	Validation des modèles . . . . .	49
<b>4</b>	<b>Modélisation de l'élasticité au prix</b>	<b>52</b>
4.1	L'élasticité au prix . . . . .	52
4.2	Cadre théorique . . . . .	54
4.2.1	Biais de confusion . . . . .	54
4.2.2	Effet causal du traitement multidose . . . . .	55
4.2.3	Méthode de la pondération sur le score de propension . . . . .	57
4.2.4	Estimation de l'effet du traitement . . . . .	60
4.2.5	Le modèle linéaire généralisé pondéré . . . . .	62
4.3	Application pratique . . . . .	63
4.3.1	Discrétisation du taux de rabais . . . . .	63
4.3.2	Modélisation de l'élasticité au prix . . . . .	66
4.3.2.1	Estimation du score de propension (PS) . . . . .	67
4.3.2.2	Évaluation de l'overlap et de l'équilibre des facteurs de confusion	69
4.3.2.3	Le modèle global . . . . .	70
4.3.3	Résultats et discussions . . . . .	72
<b>5</b>	<b>Application concrète : étude de l'impact de l'inflation sur la rentabilité des affaires nouvelles</b>	<b>77</b>
5.1	La fonction de demande . . . . .	77
5.2	La rentabilité d'une affaire nouvelle . . . . .	78
5.3	Étude de l'impact des différents scénarios de l'inflation . . . . .	81
<b>Conclusion générale</b>		<b>84</b>
<b>Liste des acronymes</b>		<b>xix</b>
<b>Table des figures</b>		<b>xxii</b>
<b>Annexes</b>		<b>xxiv</b>
Annexe 1 : Notions mathématiques . . . . .		xxiv
Annexe 2 : Sorties logiciel R . . . . .		xxv
<b>Bibliographie</b>		<b>xxxii</b>

# Note de confidentialité

Pour des raisons de confidentialité, les données ont été transformées. Plusieurs observations ont été supprimées de la base initiale dans le but de ne pas avoir l'estimation exacte du taux de transformation. Certains graphes ont été modifiés, notamment dans la partie statistique descriptive, analyse de l'élasticité au prix par profils de risque et l'étude de la rentabilité des affaires nouvelles. Les modifications effectuées n'affectent en aucun cas la compréhension des méthodes et des outils statistiques utilisés pour répondre à la problématique de l'étude.

# Remerciement

Je remercie tout d'abord Veronique MARPILLAT la responsable du service Actuariat Produit et Data Science pour son accueil et pour m'avoir permis de réaliser ce mémoire autour d'une étude fort intéressante.

Merci à tous les membres du service Actuariat Produit et Data Science d'AXA France, particulièrement l'équipe auto et transport d'avoir cru en moi et m'avoir confié l'ensemble des tâches réalisées, ainsi que pour l'accueil et leur disponibilité.

Je remercie particulièrement Amanda MARLET, Julien CHATEL et Lilian CHAVENEAU pour leur encadrement sur les sujets traités, leur confiance, leurs conseils et pour tout ce que j'ai pu apprendre à leurs côtés.

Un grand merci à Julien CHATEL pour sa relecture, sa disponibilité au quotidien et pour les remarques pertinentes pour élaborer ce mémoire.

Merci à ma tutrice Amanda MARLET pour sa relecture, son accompagnement, son soutien pour finir ce mémoire ainsi que pour la confiance qu'elle m'a accordée tout au long de mon alternance.

Je remercie mon tuteur académique Christian-Yann ROBERT pour ses remarques et son encadrement. Je remercie également le corps enseignant de l'ENSAE Paris pour leurs enseignements tout au long de mes deux ans à l'ENSAE Paris et qui m'ont permis d'acquérir les connaissances nécessaires à la bonne réussite de ce mémoire.

# Résumé

Dans un marché concurrentiel comme le marché de l’assurance des flottes automobiles d’entreprise, l’antisélection constitue un risque considérable, notamment dans le contexte actuel d’accélération d’inflation qui amplifie les charges des assureurs. Une revue à la hausse de la stratégie tarifaire pour compenser l’inflation risque d’affecter négativement la rentabilité et la demande des affaires nouvelles. Par conséquent, il paraît nécessaire de déterminer l’impact d’une variation du prix sur la demande du Produit Parc dénommé proposé par AXA France. La fonction de demande est composée de deux parties : (i) une partie statique qui représente la probabilité de transformation du devis. Pour modéliser cette partie, deux modèles ont été challengés, le modèle GLM (Generalized Linear Model) et un modèle de machine learning Xgboost (eXtreme Gradient Boosting) qui a été implémenté pour améliorer les performances prédictives et interprété par la théorie de Shap values et les graphiques de dépendances partielles (PDP); (ii) une partie dynamique qui représente l’élasticité au prix de la demande statique. Proposer une estimation robuste de cette partie constitue l’axe principal de l’étude. La difficulté de l’estimation de cette partie réside dans le biais de confusion présent dans les données historiques de l’assurance qui ne permet pas d’avoir une estimation robuste de l’élasticité directement à partir d’un modèle GLM. Pour remédier à ce problème, une méthodologie inspirée de la technique de la pondération sur le score de propension a été proposée. L’avantage de cette méthodologie est qu’elle permet d’avoir une formule opérationnelle de l’élasticité au prix en combinant le critère linéaire du modèle GLM et la capacité du modèle Xgboost de capturer les associations non linéaires entre les variables.

**Mots clés :** taux de transformation, élasticité au prix, biais de confusion, pondération sur le score de propension, demande, marge.

# Abstract

In a competitive market such as the corporate fleet insurance market, adverse selection is a considerable risk, especially in a context of the return of the inflation that amplifies insurers' expenses. An upward review of the pricing strategy to cover inflation risks negatively impacts profitability and new business demand. Therefore, it seems necessary to predict the demand for the Parc Product offered by AXA France. The demand function is composed by two parts : (i) a static part that represents the probability of a conversion of the quote. To model this part, two models have been challenged. A GLM (Generalized Linear Model) model followed by a machine learning model Xgboost (eXtreme Gradient Boosting) has been implemented to improve the forecasting performances and which has been interpreted by Shap values theory and partial dependence plots (PDP); (ii) a dynamic part that represents the price elasticity of the static demand. The main focus of the study is about to propose a robust estimation of this part. The difficulty lies in the confusion bias present in the historical insurance data, which does not allow for a robust estimate of the elasticity directly from the GLM model. A methodology inspired by the propensity score weighting technique has been proposed to remedy this problem. It allows having an operational formula of the price elasticity by combining the linear criterion of the GLM model and the capacity of the Xgboost model to capture the non-linear associations between the variables.

**Keywords :** conversion rate, price elasticity, confounding bias, propensity score weighting, demand, margin.

# Note de synthèse

## A. Contexte

En raison de l'obligation de l'assurance responsabilité civile automobile, tout le parc automobile français est assuré. En cela, le marché de l'assurance de flotte automobile peut s'avérer concurrentiel. Sur ce marché, la souscription se fait d'une manière quasi-privée et l'antisélection<sup>1</sup> constitue un risque inévitable qui s'amplifie avec le contexte actuel du retour de l'inflation. La forte reprise économique après la crise de la Covid-19 en 2020 a été accompagnée par des records successifs d'inflation. Cette accélération d'inflation se répercute sur les coûts et les charges des assureurs, notamment les coûts des sinistres via la hausse des prix de réparation et des pièces détachées. L'inflation affecte aussi la responsabilité civile corporelle via la hausse du coût horaire de l'aide humaine et les frais généraux des assureurs, notamment : les salaires, le loyer, l'expertise, etc. Par conséquent, face à la hausse des charges, les assureurs devront augmenter leurs tarifs lors de la souscription des contrats afin de rester solvables. Cette augmentation du tarif peut conduire à une sélection des profils déficitaires sur le marché.

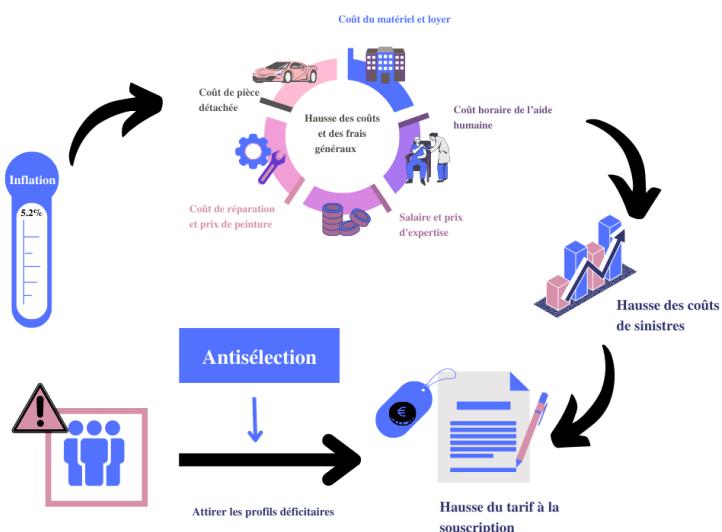


Figure 1 : Inflation et antisélection

1. Sélection des profils déficitaires sur le marché

Le produit Parc dénommé est un contrat d'assurance destiné aux flottes automobiles de 5 à 50 véhicules terrestres d'usage professionnel. Il constitue une part importante de la branche automobile d'entreprise d'AXA France avec environ 46% du chiffre d'affaires de cette branche en 2020. Ainsi, assurer la rentabilité de ce produit constitue un enjeu important. C'est dans ce contexte que le présent mémoire a pris pour objectif d'étudier l'élasticité au prix des affaires nouvelles du produit Parc dénommé dans le but de prédire la demande et d'étudier l'impact de l'inflation sur la rentabilité des affaires nouvelles.

## B. Méthodologie et résultats

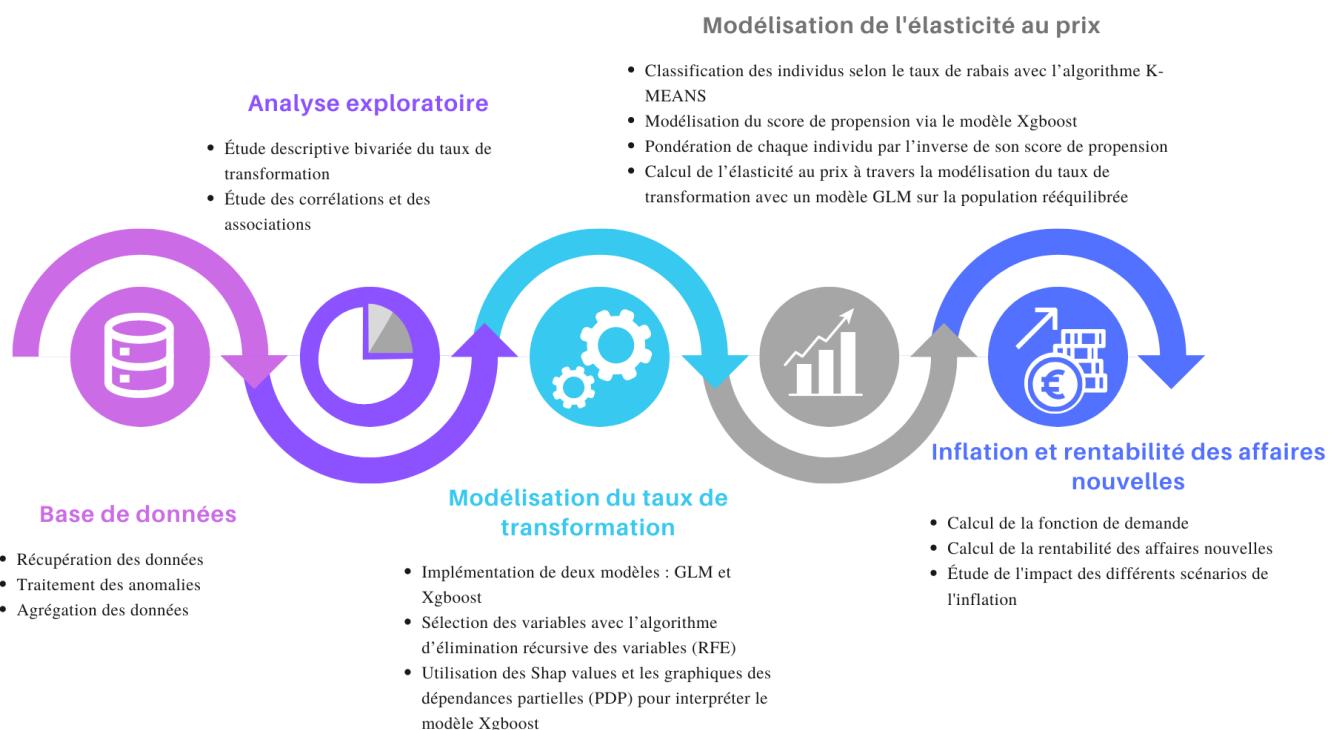


Figure 2 : Méthodologie du travail

### Construction et analyse exploratoire des données :

L'étude est faite sur les données vues au 30 juin 2021 et concerne les devis sur une durée de 5 ans et 6 mois (du 01 janvier 2016 jusqu'au 30 juin 2021). Le périmètre d'étude concerne tous les contrats du produit Parc dénommé. Nous avons commencé tout d'abord par la construction de la base de données par l'agrégation d'un ensemble de bases externes et internes d'AXA France. Ensuite, nous avons traité les anomalies de la base de données et nous avons recodé certaines variables. Enfin, nous avons effectué une analyse exploratoire des données à partir des statistiques bivariées du taux de transformation et des variables explicatives ainsi que l'étude des corrélations et des associations.

Le but de cette partie a été de comprendre les différentes interactions qui existent entre les variables et de connaître les propriétés et les limites des données. L'exploration des données a montré que la prime a un impact important sur le taux de transformation. Une tendance décroissante du taux de transformation à mesure que la prime augmente.

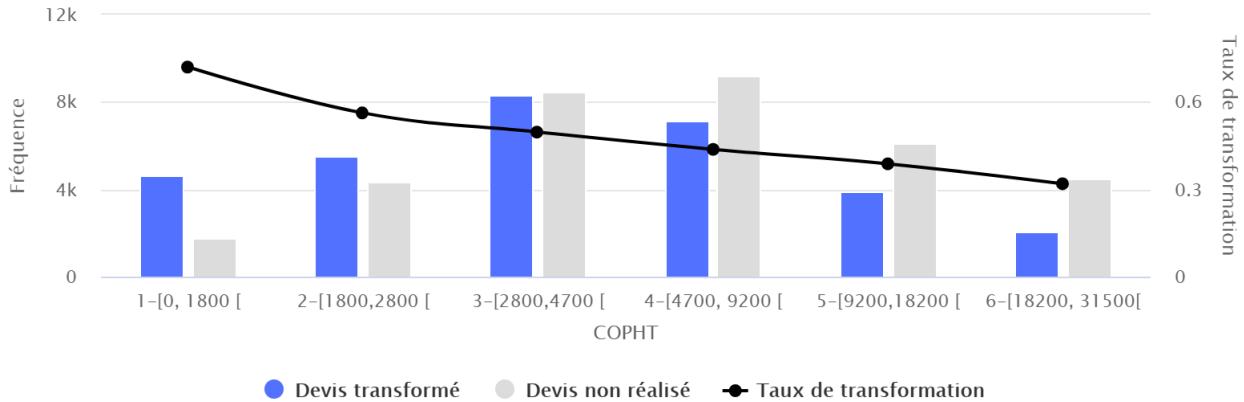


Figure 3 : Taux de transformation par segment de prime

### Modélisation du taux de transformation :

La modélisation du taux de transformation permet de déterminer les segments de clients qui transforment les devis en contrats. Cette problématique a été traitée sous deux approches : la première est statistique en utilisant le modèle GLM (Generalized Linear Model) et la seconde est algorithmique à travers un modèle de machine learning qui est l'Xgboost (eXtreme Gradient Boosting). Pour optimiser les modèles, nous avons utilisé la méthode de recherche par grille et l'algorithme de sélection récursive de variables (RFE). L'objectif principal a été de proposer un meilleur prédicteur de l'acte de transformation, ainsi pour comparer les deux modèles, nous avons utilisé les mêmes ensembles d'apprentissages et de tests. Nous avons aussi évalué la qualité prédictive des deux modèles avec les mêmes métriques d'évaluation (AUC et ACC). L'interprétation du modèle GLM a été faite via l'étude des coefficients et du modèle Xgboost à travers la théorie des Shap values et les graphiques de dépendances partielles (PDP). En comparant les résultats, le modèle GLM a fait preuve de faible capacité de généralisation des résultats de prédiction (AUC = 66.02% sur le jeu du test). En revanche, le modèle Xgboost malgré sa bonne performance prédictive (AUC = 75.87% sur le jeu du test), présente une limite non négligeable due à sa complexité le rendant non opérationnel pour la prédiction du taux de transformation.

### Modélisation de l'élasticité au prix :

L'élasticité au prix permet de déterminer la variation du taux de transformation suite à une variation future de la prime. Cette partie constitue l'axe principal du présent mémoire. L'étude de l'élasticité au prix analytique consiste à modéliser le taux de transformation en utilisant la prime et l'ensemble des

variables explicatives à disposition de l'assureur à travers le modèle GLM et à utiliser le coefficient de la prime pour calculer l'élasticité relative. Cette approche n'est pas fiable en raison du biais de confusion présent dans les données observées. En effet, les bases de données en assurance sont des données observées basées sur l'expérience de l'assureur. Ces dernières ne permettent pas de déterminer l'effet causal (le vrai effet) de la variation du taux de transformation par rapport à la variation de la prime en raison des corrélations qui existent entre la prime et les autres variables explicatives (appelées facteurs de confusion).

Pour éliminer le biais de confusion, nous avons élaboré la problématique de l'élasticité au prix dans le cadre de la littérature sur l'inférence causale qui est un champ de recherche de plusieurs disciplines : sciences sociales, machine learning, médecine et statistiques. L'objectif principal des différentes recherches de l'inférence causale est de démontrer que certains éléments représentent les causes des effets d'un phénomène observé toutes choses égales par ailleurs. La question principale de notre étude va dans le même sens : comment changerait le taux de transformation (la demande) si l'on avait proposé à un client une prime  $i$  au lieu d'une prime  $j$  toutes choses égales par ailleurs ? Pour répondre à cette problématique, nous avons proposé une méthodologie inspirée principalement des travaux de (McCaffrey et al., 2013) et (Imbens, 2000) sur l'étude de l'effet du traitement multidose en utilisant la pondération sur le score de propension. Nous avons considéré le taux de rabais de la prime comme un traitement multidose et nous avons utilisé le score de propension pour équilibrer les différents groupes du traitement. Tout d'abord, nous avons modélisé le score de propension avec le modèle Xgboost, un modèle de machine learning ayant fait ses preuves de performance et qui permet de capter tout type d'association existant entre le taux de rabais et les facteurs de confusion. Ensuite, nous avons pondéré chaque individu par l'inverse de son score de propension et nous avons évalué l'équilibre des facteurs de confusion entre les classes du taux de rabais en utilisant comme métrique la différence absolue des moyennes standardisées (ASMD). Enfin, à partir de la population des groupes équilibrés, nous avons modélisé le taux de transformation à l'aide d'une régression doublement robuste (Robins et al., 1995), appelée dans notre étude modèle global.

L'avantage de l'approche proposée est qu'elle permet de rééquilibrer les données et d'obtenir une formule opérationnelle de l'élasticité en combinant le critère linéaire du modèle GLM et la capacité du modèle Xgboost de capter les associations non linéaires entre les variables explicatives. Les résultats obtenus sont cohérents avec la réalité du marché, l'élasticité au prix est croissante en fonction du taux de rabais. Cette élasticité varie de 0.93 à 7.58. Une élasticité égale 7.58 signifie qu'une augmentation de la prime de 1% entraîne une baisse du taux de transformation de 7.58%. Une élasticité inférieure à 1 signifie que le devis est peu élastique. En revanche, une élasticité supérieure à 1 signifie que le devis est élastique. De ce fait, les devis qui ont bénéficié de rabais minimaux sont peu élastiques et les devis

qui n'ont pas bénéficié de rabais, voire de majoration de prime sont très élastiques à une augmentation ou baisse future de la prime.

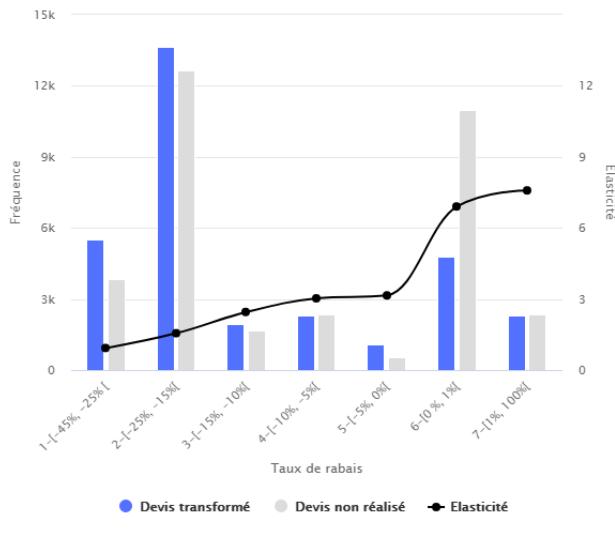


Figure 4 : L'élasticité par taux de rabais

### Inflation et rentabilité des affaires nouvelles :

Pour mesurer l'impact de l'inflation, nous avons introduit deux mesures : la fonction de demande et la marge.

- La fonction de demande représente l'indicateur de production, elle donne la probabilité qu'un client accepte d'acheter les garanties du produit Parc dénommé en fonction du prix. Elle est composée de deux parties : (i) la partie statique représente la probabilité qu'un client accepte d'acheter les garanties du produit Parc dénommé avec le prix actuel (le taux de transformation) ; (ii) la partie dynamique représente l'élasticité au prix de la partie statique.
- La marge représente l'indicateur de rentabilité. La marge d'une affaire nouvelle mesure la capacité de la prime commerciale à payer les sinistres futurs, elle est calculée par l'écart entre la prime pure (l'estimateur du coût moyen futur des sinistres) et la prime commerciale. Pour un ensemble de devis  $L$ , la marge espérée d'un changement tarifaire  $\epsilon\%$  est calculée à partir de la fonction de demande comme suit :

$$marge\% = \frac{\sum_{i \in L} demande_i(\epsilon\%) * (Prime commerciale * (1 + \epsilon\%) - Prime pure_i)}{\sum_{i \in L} demande_i(\epsilon\%) * Prime commerciale * (1 + \epsilon\%)}$$

L'inflation future impacte d'une façon directe la rentabilité d'une affaire nouvelle via la hausse du coût moyen futur des sinistres (la prime pure). La figure ci-dessous donne la fonction de demande et la marge espérée selon les différents scénarios sans et avec ajustement tarifaire pour compenser l'inflation.

Scénarios d'inflation	Marge avec ajustement	Marge sans ajustement	Demande avec ajustement	Demande sans ajustement
0%	23%	23%	47%	47%
1%	23%	22%	46%	47%
2%	23%	22%	45%	47%
3%	23%	21%	43%	47%
4%	22%	20%	42%	47%
5%	22%	19%	41%	47%
6%	22%	19%	39%	47%

Figure 5 : La fonction de demande et la marge selon les différents scénarios d'inflation

Les résultats permettent de constater que l'inflation détériore la rentabilité des affaires nouvelles, un scénario d'inflation de 5% introduit une baisse de la marge de 4 points. Dans le cas de l'ajustement de la prime commerciale avec le même taux d'inflation, le scénario de 5% introduit une baisse d'un point de la marge espérée. En conséquence, cet ajustement a permis de gagner 3 points en termes de rentabilité. En revanche, il a introduit une perte de 6 points de la demande. Ceci suppose que les stratégies des concurrents sont maintenues constantes. En effet, une anticipation semblable à celles des concurrents permet de conserver la position actuelle du produit Parc dénommé sur le marché. En revanche, une anticipation inférieure à celles des concurrents permet de croître la production du produit Parc dénommée et détériorer la rentabilité et inversement.

## C. Conclusion

La fonction de demande et la marge représentent des outils de pilotage importants pour anticiper les variations de la production et de la rentabilité des affaires nouvelles suite à des stratégies de changement tarifaire du produit Parc dénommé. Elles sont calculées à partir du taux de transformation et de la fonction de l'élasticité au prix, une production élevée permet d'avoir un grand nombre d'affaires nouvelles avec une rentabilité inférieure et inversement. Pour une bonne prise de décision, il est important de croiser ces deux quantités et choisir la position souhaitée entre les deux. Une revue à la hausse de la stratégie tarifaire pour compenser l'inflation doit prendre en considération la position choisie entre les deux et elle doit être segmentée, selon les différents profils de risque.

# Executive Summary

## A. Context

Today, the French fleet insurance market is saturated. Given the obligation liability insurance, the entire French fleet is covered. This makes the fleet insurance market competitive. In this market, underwriting is done in a quasi-private way and adverse selection is an inevitable risk amplified with the rise of inflation. The strong economic recovery after the Covid-19 crisis in 2020 was accompanied by successive inflation records. This acceleration of inflation is having an impact on insurers' costs and expenses, particularly on claims costs due to the increase in repair and spare parts prices. Inflation also impacts bodily injury liability via the increase in the hourly cost of human assistance and insurers' overheads, in particular : salaries, rent, expertise, and so on. However, faced with rising costs, insurers will be obliged to increase their premiums when subscribing to contracts. This rate increase may lead to a selection of loss-making profiles on the market.

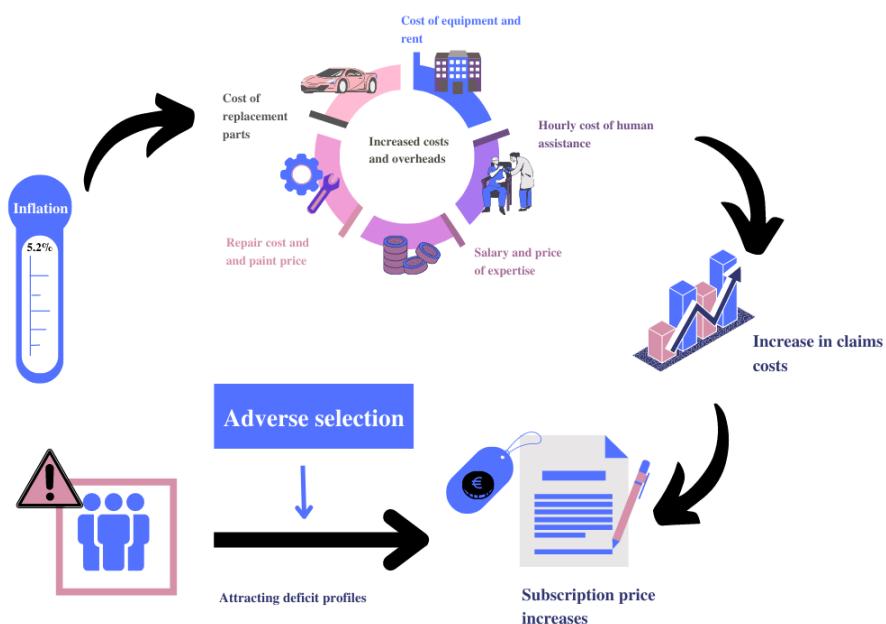


Figure 1 : Inflation & adverse selection

The Parc Product is an insurance contract for automobile fleets of 5 to 50 vehicles for professional use. It represents a significant portion of AXA France's corporate motor business, accounting for approximately 46% of the business line's revenues in 2020. Ensuring the profitability of this product is a major challenge. It is in this context, the objective of this paper is to study the price elasticity of new business to predict demand and to study the impact of inflation on the profitability of new business.

## B. Methodology & Results

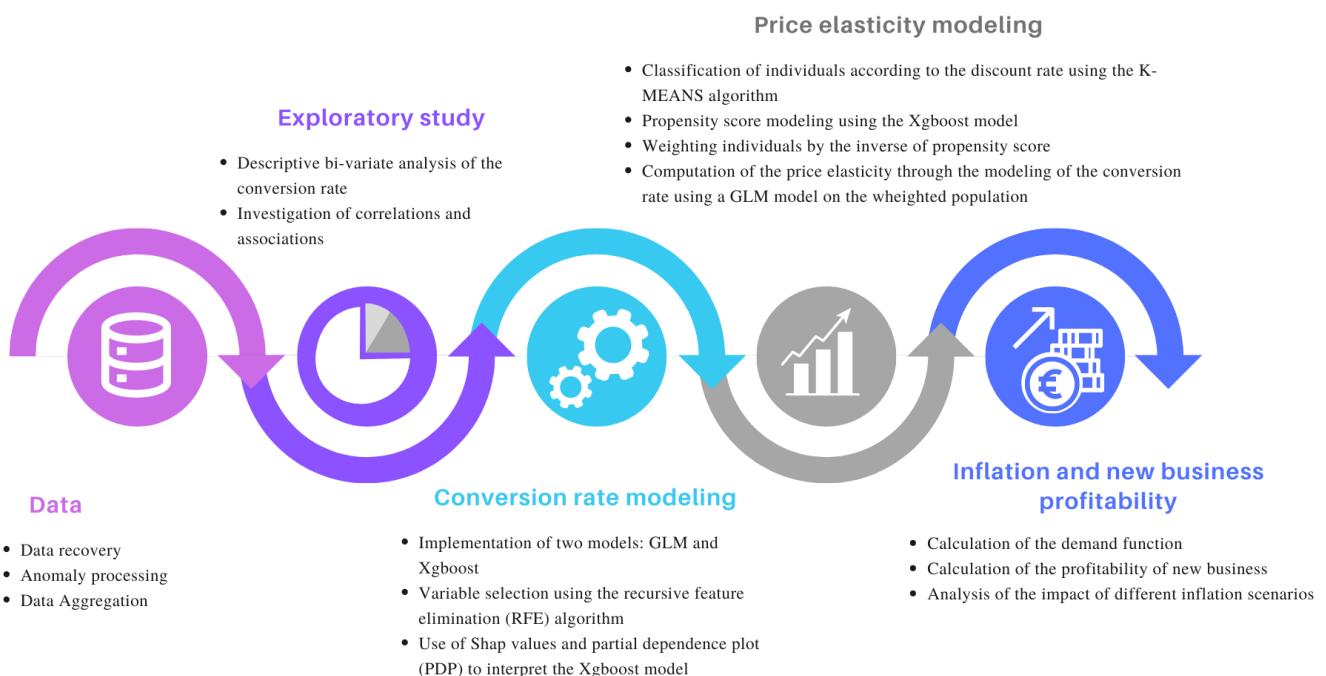


Figure 2 : Methodology

### Data and exploratory study :

The study is made on the data seen on June 30, 2021, which concerns the quotations on a duration of 5 years and 6 months (from January 01, 2016, until June 30, 2021). The scope of the study concerns all the contracts of the Parc product. We started by building the database by aggregating a set of external and internal databases of AXA France. Then, we treated the anomalies of the database and decoded some variables. Finally, we performed an exploratory analysis based on bi-variate statistics of the conversion rate and study of correlations and associations. We wanted to understand the different interactions between the variables and learn about the properties and limitations of the data. The exploration of the data showed that the premium has a negative impact on the conversion rate.

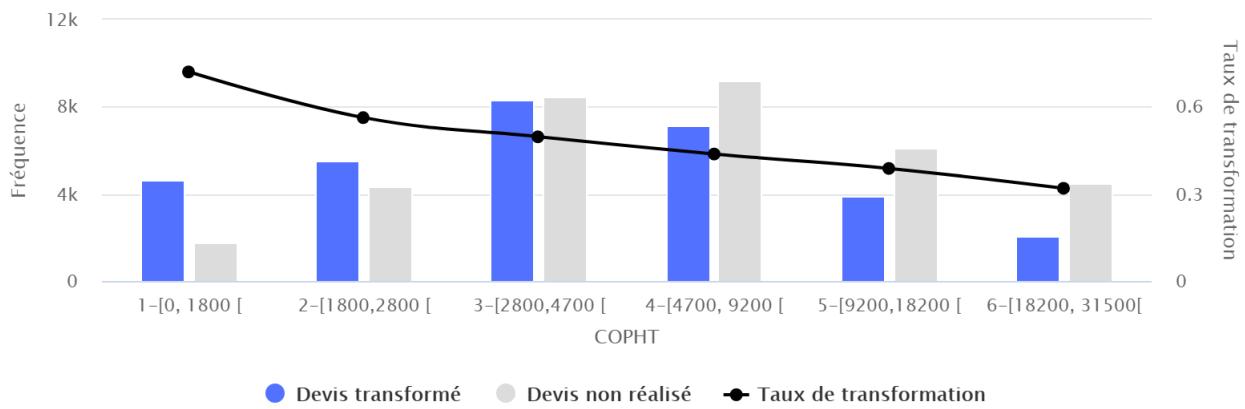


Figure 3 : Conversion rate by premium segment

Interpretation guide : The graph shows the frequency of each premium segment on the left axis. The blue bars represent the converted quotes and the grey bars the unrealized quotes. With the right axis the conversion rate. The black curve represents the conversion rate.

#### Conversion rate modeling :

The modeling of the conversion rate makes it possible to determine the segments of customers who convert more the quote into contract. This problem has been addressed from two angles : a statistical view with GLM (Generalized Linear Model) and an algorithmic view through a machine learning model : Xgboost (eXtreme Gradient Boosting). For their optimization, we used a grid search method and the recursive feature elimination algorithm (RFE). The main objective was to propose a better predictor of the conversion rate. To compare the two models, we used the same training and test sets and evaluated the forecast quality of both models using the same evaluation metrics (AUC and ACC). The interpretation of the GLM model was done via the study of the coefficients and the Xgboost model through the Shap values theory and the partial dependence plots (PDP). By comparing the results, the GLM model showed up poor generalizability of the prediction results (AUC = 66.02% on the test set). On the other hand, the Xgboost model, despite its good forecasting performance (AUC = 75.87% on the test set), presents a non-negligible limitation due to its complexity making it non-operational for conversion rate prediction.

#### Price elasticity modeling :

Price elasticity is used to determine the change in the conversion rate following a future change in the premium. This part is the main focus of this paper. The study of analytical price elasticity consists of modeling the conversion rate using the premium and all the explanatory variables available to the insurer through the GLM model and using the coefficient on the premium to calculate the relative elasticity. This approach is not reliable because of the confounding bias present in the observed data. Indeed, insurance databases are observed data based on the insurer's experience. These data do not

allow us to determine the causal effect (the true effect) of the change in conversion rate versus the change in premium because of the correlations that exist between the premium and the other explanatory variables (called confounding factors). To eliminate the confounding bias, we developed the price elasticity problem within the framework of the causal inference literature, which is a field of research from several disciplines : social sciences, machine learning, medicine and statistics. The main objective of these various research studies on causal inference is to demonstrate that certain elements represent the causes of the effects of an observed phenomenon, all other things being equal. The main question of our study is along the same lines : how would the conversion rate (demand) change if a customer had been offered a premium  $i$  instead of a premium  $j$  all other things being equal ? To address this issue, we proposed a methodology inspired primarily by the work of (McCaffrey and al., 2013) and (Imbens, 2000) on studying the multiple treatment effect using propensity score weighting. We considered the premium discount rate as a multiple treatment and used the propensity score to balance the different treatment groups. First, we modeled the propensity score using the Xgboost model, a machine learning model with proven performance that captures any type of association between the discount rate and confounders. Then, we weighted each individual by the inverse of their propensity score and assessed the balance of confounding factors across classes of the discount rate using the absolute difference of standardized means (ASMD) as a metric. Finally, from the population of balanced groups, we modeled the conversion rate using a doubly robust regression (Robins and al., 1995), referred to in our study as the global model.

The advantage of the proposed approach is that it allows us to rebalance the data and obtain an operational formula for the elasticity by combining the linear criterion of the GLM model and the ability of the Xgboost model to capture the non-linear associations between the explanatory variables. The results obtained are consistent with the market reality, the price elasticity is increasing with the discount rate. This elasticity varies from 0.93 to 7.58. An elasticity equal to 7.58 means that an increase in the premium of 1% leads to a decrease in the conversion rate of 7.58%. An elasticity less than 1 means that the quote is inelastic. On the other hand, an elasticity greater than 1 means that the quote is elastic. Thus, quotes that received minimal discounts are inelastic and quotes that did not receive a discount or a premium increase are highly elastic to a future premium increase or decrease.

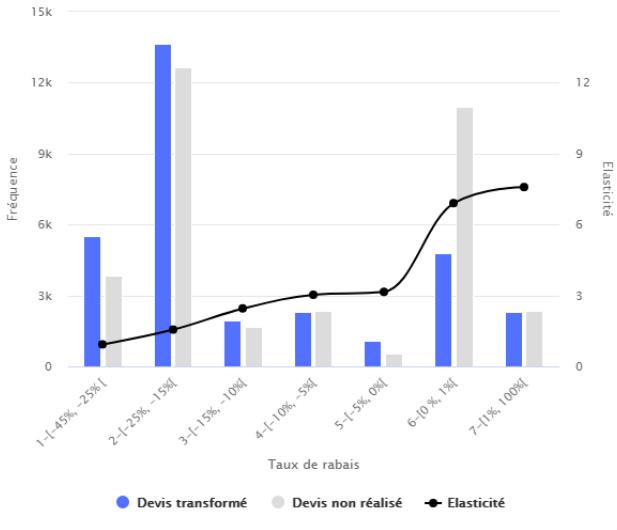


Figure 4 : The elasticity by discount rate

Interpretation guide : The graph shows on the left axis the frequency of each segment of the discount rate. The blue bars represent the transformed quotes and the grey bars the unrealized quotes. With the right axis the price elasticity. The black curve represents the price elasticity.

#### Inflation and new business profitability :

To assess the impact of inflation, we have introduced two measures : the demand function and the margin.

- The demand function is the indicator of production. It gives the probability that a customer will buy the insurance as a function of price. It is composed by two parts : (i) a static part which represents the probability that a customer will agree to purchase the insurance with the current price (the conversion rate); (ii) a dynamic part which represents the price elasticity of the static part.
- The margin is the profitability indicator. The new business margin measures the ability of the commercial premium to pay future claims and is calculated as the difference between the pure premium and the commercial premium. For a set of quotes L, the expected margin of a rate change  $\epsilon\%$  is calculated from the demand function as below :

$$\text{margin}\% = \frac{\sum_{i \in L} \text{demand}_i(\epsilon\%) * (\text{Commercial premium} * (1 + \epsilon\%) - \text{Pure premium}_i)}{\sum_{i \in L} \text{demand}_i(\epsilon\%) * \text{Commercial premium} * (1 + \epsilon\%)}$$

Future inflation directly affects the profitability of new business through the rising future average cost of claims (the pure premium). The figure below shows the demand function and the anticipated margin under the different scenarios without and with price adjustment to cover inflation.

Inflation scenarios	Margin with adjustment	Margin without adjustment	Demand with adjustment	Demand without adjustment
0%	23%	23%	47%	47%
1%	23%	22%	46%	47%
2%	23%	22%	45%	47%
3%	23%	21%	43%	47%
4%	22%	20%	42%	47%
5%	22%	19%	41%	47%
6%	22%	19%	39%	47%

Figure 5 : The demand function and margin for different inflation scenarios

The analysis shows that inflation degrades the profitability of new business, with an inflation scenario of 5% leading to a 4-point drop in margin. When the commercial premium is adjusted with the same inflation rate, the 5% scenario introduces a one point decrease in the expected margin. As a result, this adjustment resulted in a 3-point gain in profitability. On the other hand, it introduced a 6-point loss in demand. This assumes that competitors' strategies are kept constant. Indeed, an anticipation similar to that of competitors allows the current market position of the Parc product to be maintained. On the other hand, an expectation that is lower than that of the competitors allows for an increase in the production of the Parc product and a deterioration in profitability, and vice versa.

## C. Conclusion

The demand function and the margin are essential management tools for anticipating changes in production and profitability of new business. They are calculated using the conversion rate and the price elasticity. High production allows for numerous new business with lower profitability and vice versa. For good decision-making, it is important to cross-reference these two quantities and choose the desired position between the two. An upward revision of the pricing strategy to cover inflation must consider the position chosen between the two, and should be segmented according to different risk profiles.

# Introduction générale

Avec un marché de 2,4 milliards d'euros, les contrats d'assurance des flottes automobiles d'entreprise représentent 10,3 % de l'ensemble des cotisations automobiles en France en 2020<sup>2</sup>. Pour satisfaire la demande importante, plusieurs entreprises d'assurance partagent ce marché. En cela, le marché de l'assurance des flottes automobiles est concurrentiel, et chaque compagnie d'assurance doit développer des stratégies tarifaires à différentes échelles pour garantir un positionnement de choix dans la scène assurantielle.

La particularité de ce marché est l'information incomplète sur le tarif moins cher du marché pour chaque profil de risque. En effet, la souscription se fait d'une manière quasi-privée et le taux de transformation historique représente le seul indicateur de la position de chaque assureur sur le marché. Ce manque d'information accentue le risque de sélectionner les profils de risque non rentables, notamment dans un contexte d'accélération d'inflation qui amplifie les charges des assureurs. Par conséquent, pour couvrir ce risque de hausse des charges et assurer sa solvabilité, l'assureur sera contraint de mettre en place une nouvelle structure tarifaire pour les affaires nouvelles. Ce changement de tarif doit être fait d'une manière stratégique et optimisée et il doit aussi prendre en considération l'élasticité de la probabilité de souscription de chaque profil de risque.

C'est dans ce contexte que nous étudions la sensibilité au prix lors de l'acte de souscription du produit Parc dénommé proposé par AXA France. Ce produit représente une part importante du chiffre d'affaires de la branche automobile d'entreprise d'AXA France. Par conséquent, l'amélioration de la rentabilité de ce produit constitue un enjeu important. Ainsi, il est crucial de suivre et d'analyser le taux de transformation. Ceci permet de suivre la position du produit sur le marché et d'en tirer l'impact de ce changement tarifaire.

Pour répondre à ces problématiques, nous présenterons tout d'abord le contexte de l'étude, les données, les divers retraitements et une analyse exploratoire des données.

Ensuite, nous allons caractériser les profils de risque qui transforment les devis par l'analyse et la modélisation du taux de transformation. L'objectif de cette partie est de comprendre quel type de

---

2. <https://www.franceassureurs.fr/wp-content/uploads/VF-Donnees-cles-2020.pdf>

clients nous attirons par la stratégie tarifaire actuelle. Deux types de modèle seront challengés : le modèle GLM (Generalized Linear Model) et un modèle de machine learning Xgboost (eXtreme Gradient Boosting) qui sera implémenté pour améliorer les performances prédictives et interprété par la théorie de Shap values et les graphiques de dépendances partielles (PDP).

Ensuite, nous menons une étude pour modéliser l'élasticité au prix. Cette partie représente l'axe principal de l'étude. Elle a pour objectif de déterminer l'effet marginal d'un changement tarifaire futur sur le taux de transformation. La difficulté de l'étude de l'élasticité au prix réside dans la nature des données historiques à notre disposition et les corrélations qui existent entre la prime et les variables explicatives et qui apportent une certaine confusion sur l'estimation de cet effet marginal à partir du modèle GLM. Pour remédier à ce problème, nous avons recours aux méthodes de la littérature de l'inférence causale, particulièrement à la méthode de la pondération sur le score de propension (IPTW), une méthode très connue en recherche médicale. La méthodologie que nous proposons pour estimer l'élasticité au prix commence par la discrétisation du taux de rabais de la prime par l'algorithme K-means qui permet de préserver le maximum de l'inertie expliquée par cette variable. Elle se poursuit par la modélisation du score de propension avec l'Xgboost, un modèle qui permet de capter tout type d'association existant entre les facteurs de confusion (les variables explicatives) et le taux de rabais. Elle finit par la pondération de chaque observation par l'inverse de son score de propension. Ceci dans le but d'éliminer les corrélations qui existent entre le taux de rabais et les variables explicatives, de modéliser le taux de transformation à l'aide d'un GLM doublement robuste et de calculer la fonction élasticité.

Enfin, nous utiliserons le taux de transformation et la fonction de l'élasticité au prix pour calculer la fonction de demande et la marge espérée, deux indicateurs de pilotage du produit Parc dénommé qui permettent d'anticiper les variations futures de la production et de la rentabilité selon les stratégies de changement tarifaire et les scénarios d'inflation.

# **Chapitre 1**

## **Mise en contexte**

### **Préambule**

Ce premier chapitre du mémoire a pour objectif de cerner le contexte de l'étude et les différents objectifs. Tout d'abord, il présentera le marché de l'assurance des flottes automobiles d'entreprise et le périmètre d'étude. Ensuite, il se focalisera sur le phénomène d'antisélection comme un dysfonctionnement du marché d'assurance concurrentiel. Enfin, il exposera la problématique de l'étude.

### **1.1 Présentation du périmètre de l'étude**

L'étude s'intéresse au produit Parc dénommé. Il s'agit du contrat d'assurance de flotte automobile de 5 à 50 véhicules terrestres d'usage professionnel. La tarification de ce produit se fait par garantie et par véhicule assuré. Les véhicules peuvent être de toute catégorie, de ce fait le périmètre d'étude contiendra toutes les catégories de véhicules.

#### **1.1.1 L'assurance de flotte automobile d'entreprise**

L'assurance de flotte automobile d'entreprise :

L'assurance de flotte automobile d'entreprise permet de couvrir le risque de plusieurs véhicules dans un même contrat. L'ensemble de ces véhicules appartient à une seule personne morale. La flotte assurée peut contenir différents types de véhicules : voitures de fonction, véhicules utilitaires, poids lourds, remorques et semi-remorques, engins agricoles, engins de chantier, etc. Les garanties du contrat permettent de couvrir le risque du souscripteur du contrat et de l'ensemble des véhicules.

Toutefois, le représentant de la personne morale n'a aucune obligation de mentionner ni les conducteurs des véhicules de la flotte ni l'historique de leurs sinistres. En effet, le souscripteur du risque est le seul responsable des dommages causés par les conducteurs des véhicules. Ceci est dû au fait que la personne morale est civilement responsable des conducteurs en vertu de l'article L121-2 du code des assurances. De ce fait, le contrat d'assurance de flotte automobile n'englobe pas le bonus-malus. Par ailleurs, l'ensemble des sinistres annuels doit être déclaré par le souscripteur du contrat et enregistré par la suite dans un fichier national géré par l'association pour la gestion des informations sur le risque en assurance (AGIRA). L'historique des sinistres déclarés va être utilisé par la suite pour calculer la prime d'une affaire nouvelle. L'absence du système de bonus malus pour ce type d'assurance donne une grande marge de négociation aux assurés lors de la souscription et du renouvellement du contrat. Par conséquent, il n'est pas évident pour les assureurs d'identifier les tarifs proposés par les concurrents. En revanche, l'assureur peut identifier sa position sur le marché à travers l'analyse de l'ensemble des devis résiliés et des devis concrétisés.

#### Le marché de l'assurance de flotte automobile d'entreprise :

S'agissant du marché de flotte automobile d'entreprise, les immatriculations neuves des véhicules de flottes automobiles constituent plus que 43% de l'ensemble des véhicules immatriculés en 2020<sup>1</sup>. Cette proportion importante se répercute sur la demande de l'assurance de flotte automobile. En effet, d'après France Assureur, le nombre de véhicules assurés en contrat flotte automobile d'entreprise a connu une hausse de 5% entre les années 2019 et 2020, il est passé de 4.4 millions véhicules à 4.6 millions. Pour la même période, les cotisations ont enregistré une croissance de 5,7%. Par ailleurs, le ratio sinistre à prime a connu une amélioration entre les années 2019 et 2020, il est passé de 73% à 86%. Le graphe ci-dessous représente l'évolution du ratio sinistre à prime d'une part de 90% du marché de l'assurance de flotte automobile d'entreprise. Nous remarquons une tendance à la hausse avant l'année 2020 qui a été marquée par la crise de la pandémie, et donc une baisse de l'activité économique, ce qui a eu un impact positif sur la charge sinistre. En revanche, la forte reprise économique après la crise et les records successifs de l'inflation peuvent avoir un fort impact négatif sur le ratio sinistre à prime. Par conséquent, assurer la rentabilité dans ce contexte d'incertitudes constitue un enjeu majeur pour les assureurs.

---

1. <https://www.flotauto.com/marche-automobile-2021-bas-20220103.html>

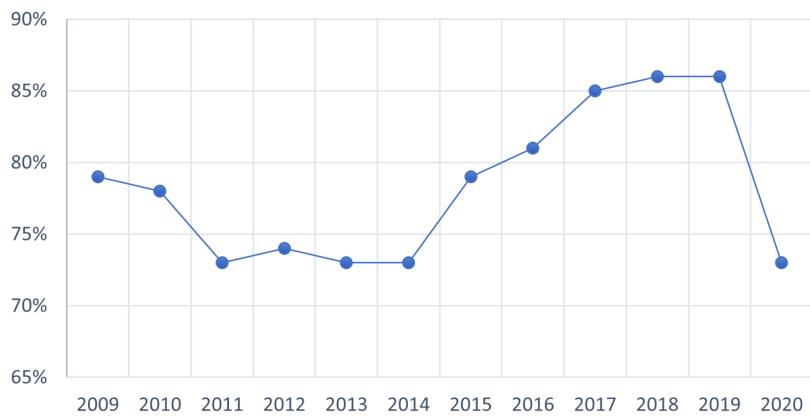


Figure 1.1 – Ratio sinistre à prime du marché français de flotte d’entreprise. Source : France Assureurs

### 1.1.2 Le produit Parc dénommé d’AXA France

L’assurance automobile d’entreprise d’AXA France :

AXA France propose différents produits d’assurance automobile d’entreprise. Ces produits se différencient par la taille de la flotte à assurer et l’activité de la personne morale. La tarification diffère d’un produit à l’autre et les produits permettent de couvrir toute catégorie de véhicules.

- Produit Mono : il s’agit d’un contrat à un seul véhicule. Ce produit est proposé généralement aux petites entreprises. Puisqu’il n’y a pas un grand nombre de véhicules et pour bien estimer le risque, la tarification se fait selon les caractéristiques de chaque véhicule et de chaque conducteur.
- Produit Parc dénommé : ce produit fait l’objet de l’étude de ce mémoire. Il s’agit des flottes fermées de 5 à 50 véhicules. La tarification de ce produit se fait véhicule par véhicule en se basant sur les caractéristiques spécifiques à chaque véhicule. Cependant, il n’existe pas de tarification selon les caractéristiques du conducteur pour ce produit, car les conducteurs ne sont pas toujours affectés à un véhicule précis.
- Produit Flotte ouverte : il s’agit des grandes flottes de 50 véhicules et plus. La tarification de ce produit ne se fait pas véhicule par véhicule du fait de l’absence de déclaration des véhicules présents dans la flotte. Généralement, les clients de ce produit sont des grandes entreprises ou des collectivités locales.
- Produit Garages et Concessions : ce produit concerne principalement les professionnels de l’automobile, à savoir ceux exerçant les activités de : garagistes, concessionnaires, ou encore contrôle technique.

L'assurance automobile occupe une place importante dans le portefeuille AXA IARD Entreprise, à fin 2020, l'assurance automobile constituait 28% de l'ensemble du chiffre d'affaires d'AXA IARD Entreprise (2.8 Md €). Par ailleurs, le produit Parc dénommé représentait une part de 46% de l'ensemble du chiffre d'affaires de l'assurance automobile d'entreprise comme le montre le graphe ci-dessous :

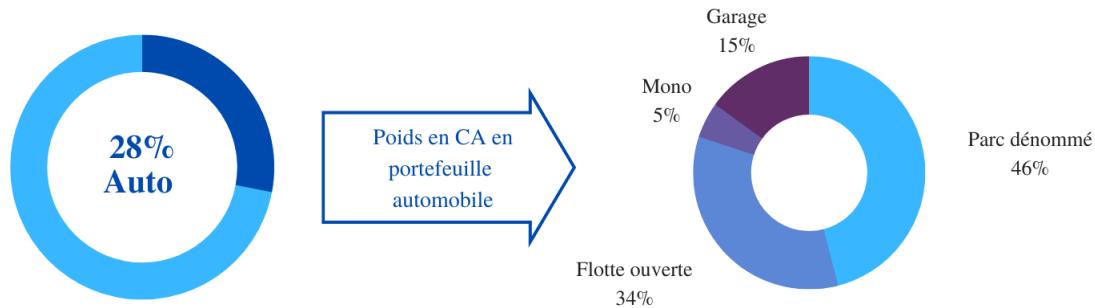


Figure 1.2 – Poids en CA en portefeuille AXA IARD Entreprise en 2020

#### Le produit Parc dénommé :

Le produit Parc dénommé propose différentes garanties selon la catégorie du véhicule et l'activité de l'entreprise. Il existe 4 catégories de garanties :

- Les garanties de la responsabilité civile : la catégorie minimale demandée par l'assuré. Il s'agit d'une garantie obligatoire en assurance automobile selon les articles L211-1 et R211-5 du code des assurances.
- Les garanties dommages : cette catégorie contient l'ensemble des garanties qui couvrent les dommages matériels des véhicules (garantie dommage, vol, incendie, bris de glace).
- Les garanties optionnelles : ce type de garantie couvre les matières transportées (effets personnels, marchandises...), la sécurité du conducteur et les pertes financières de la flotte assurée.
- Les garanties d'assistance : contiennent l'ensemble des garanties d'assistance médicale et de véhicule ainsi que les services de dépannage. L'ensemble des services garantis sont assurés à l'aide du partenaire AXA assistance.

Il est important de noter que les garanties ne sont pas éligibles à tout type de véhicule. En outre, la tarification se fait pour chaque garantie et selon la catégorie de véhicule, ainsi pour ce produit, nous distinguons 7 catégories de véhicules données dans la table (1.1).

Le produit Parc dénommé représente 13% de l'ensemble du chiffre d'affaires AXA IARD Entreprise en 2020. Assurer la rentabilité de ce périmètre constitue un enjeu important. De ce fait, une mise en œuvre d'un nouveau tarif a été effectuée en 2021. L'objectif principal était d'actualiser les coefficients de l'ancien tarif. Le nouveau tarif concerne essentiellement les affaires nouvelles qui

représentent sur une année 10% du chiffre d'affaires du portefeuille Parc dénommé. Une étude de l'impact de ce changement de tarif sur le nombre de devis contractés s'avère importante.

Catégorie 1	Les véhicules légers, utilitaires, voiturettes et véhicules de sports
Catégorie 2	Les camions, tracteurs, remorque de plus de 3.5 tonnes et les semi-remorques
Catégorie 3	Les autobus et autocars
Catégorie 4	Les engins agricoles
Catégorie 5	Les engins de chantier
Catégorie 6	Les deux-roues
Catégorie 7	Les remorques de moins de 3.5 tonnes

Table 1.1 – Les catégories du Parc dénommé

## 1.2 Le phénomène d'antisélection

L'assurance est caractérisée par l'inversion du cycle de production, c'est-à-dire la ressource précède la dépense. Ceci met une certaine incertitude sur les dépenses futures de l'assureur. En effet, lors de la souscription du contrat, l'assureur propose une prime établie à partir de l'historique passé. Celle-ci ne reflète pas exactement le risque couvert, puisque l'information à disposition de l'assureur ne traduit pas d'une manière exacte le comportement futur de ses nouveaux clients en matière de risque. Le manque d'information conduit à un dysfonctionnement général du marché de l'assurance. L'étude présente se focalise sur le comportement des assurés lors de l'acte de souscription. Ainsi, elle s'intéresse à l'asymétrie de l'information qui apparaît lors de la souscription du contrat. L'antisélection (adverse selection) est la principale forme d'asymétrie d'information qui puisse apparaître avant la signature du contrat. La section courante discutera en détail ce risque.

### 1.2.1 Antisélection en assurance automobile

Définition d'antisélection :

Aujourd'hui, le marché d'assurance de flotte automobile français est saturé. En raison de l'obligation de l'assurance responsabilité civile automobile, tout le parc automobile français est assuré. Sur un tel marché, un nouvel assuré a une multitude d'offres et de choix d'assurance. Ainsi, le nouvel assuré choisit le contrat le plus attrayant en termes de prix. Par ailleurs, l'assureur qui n'arrive pas à distinguer les degrés de risque finit par proposer un tarif unique pour les différents types de risque sur

le marché. Ce type de situation désavantage les bas risques puisqu'ils coûtent moins cher que le prix proposé par l'assureur. Ainsi, ils choisissent de refuser le contrat proposé par ce dernier et cherchent un autre contrat moins cher sur le marché (Rothschild et Stiglitz, 1976) [Jos76]. De ce fait, les assurés disposent davantage d'information que l'assureur concerné. Cette asymétrie d'information entre les deux parties est appelée antisélection.

#### Théorie de la sélection :

La théorie de la sélection en assurance est apparue pour la première fois dans les années 70. Notamment, avec les travaux fondamentaux de Rothschild et Stiglitz (1976) [Jos76] qui permettent de décrire le fonctionnement du marché concurrentiel d'assurance dans le cadre d'antisélection. Les auteurs considèrent un marché d'assurance concurrentiel, comme exemple le marché de l'assurance automobile d'entreprise, il est caractérisé par l'hétérogénéité des profils de risque (par exemple : les véhicules légers et les véhicules de poids lourd). Sur ce type de marché, les assurés se distinguent par la probabilité de survenance de sinistre et ne peuvent pas agir sur cette probabilité. Toutefois, l'assureur n'observe pas la probabilité de survenance. Elle est considérée exogène et ne dépend pas des paramètres du modèle de Rothschild et Stiglitz [Jos76]. Les deux auteurs supposent l'existence de deux types d'assuré averses au risque : les bas risques avec une faible probabilité de survenance de sinistre et les hauts risques avec une probabilité de survenance élevée. Puisque l'assureur ne dispose pas de l'information nécessaire pour différencier les deux types, il propose un seul tarif pour les deux types de risque [Jos76].

L'équilibre de Rothschild-Stiglitz (RS) [Jos76] sur ce type de marché est réalisable si chaque contrat présent sur le marché apporte un profit positif et aucun nouveau contrat arrivant peut faire un profit strictement positif en dehors des contrats déjà présents. Dans le cadre d'antisélection, cet équilibre peut ne pas exister. S'il existe, il est séparateur. Dans ce cas, les hauts risques vont bénéficier d'une couverture totale à coût élevé, tandis que les bas risques vont accepter une assurance partielle avec le prix le moins cher sur le marché.

#### L'antisélection et la segmentation du risque :

Pour limiter les effets d'antisélection, l'assureur fait recours à la segmentation du tarif. En effet, la segmentation correspond à la capacité de l'assureur de regrouper les assurés dans des groupes homogènes et de proposer à chaque segment le tarif adéquat afin d'en tirer des profits positifs. En revanche, chaque groupe doit contenir un nombre suffisant d'individus pour éviter la volatilité de la sinistralité

et pour répartir le coût de réalisation d'un sinistre sur l'ensemble du groupe homogène. Nous parlons dans ce cas du principe de la mutualisation du risque. De ce fait, l'assureur doit chercher un compromis entre la mutualisation et la segmentation du risque. Pour avoir une bonne position entre les deux principes, l'assureur utilise son expérience et un ensemble de variables tarifaires à sa disposition.

### 1.2.2 Stratégie tarifaire et antisélection

La stratégie tarifaire dépend tout d'abord des objectifs de l'assureur et des informations à sa disposition sur le marché. Pour illustrer cela, imaginons un assureur A qui détient le monopole sur un marché et qui doit couvrir un capital de 800 € pour deux groupes d'assurés : les hauts risques avec probabilité de survenance de sinistre égale à 1/2, les bas risques avec probabilité de survenance égale à 1/4. Pour des raisons de simplification, nous supposons que chaque groupe contient 6 assurés. L'assureur vise un rapport sinistre à prime (S/P) égal à 100%. Pour atteindre cet objectif, l'assureur propose la prime mutualisée, c'est-à-dire l'espérance des sinistres futurs.

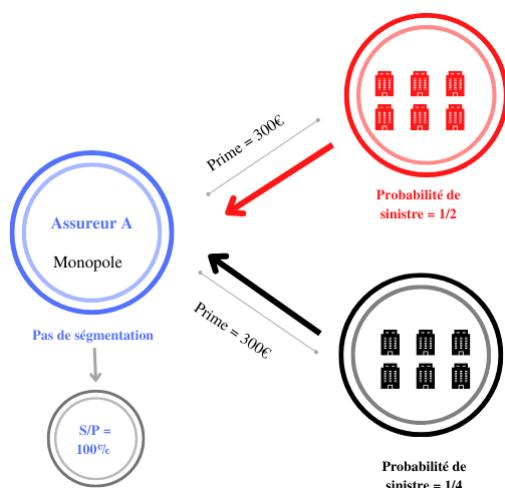


Figure 1.3 – Antisélection : marché monopolistique

Les deux groupes d'assurés vont accepter le prix moyen ( $300 = \frac{6 \cdot \frac{1}{4} \cdot 800 + 6 \cdot \frac{1}{2} \cdot 800}{12}$ ) [Jos76]. En effet, dans ce type de situation, les hauts risques (rouges) vont être assurés moins cher que leur prix réel  $400 = 800 * \frac{1}{2}$  et les bas risques vont être couverts plus cher que la valeur réelle de leur risque  $200 = 800 * \frac{1}{4}$ . En revanche, l'assureur va bénéficier d'une mutualisation parfaite du risque qui va lui assurer un ratio sinistre à prime (S/P) égal à 100%.

Imaginons l'arrivée d'un second assureur B, qui fait une sélection de risque. Il assure les bas risques avec 200 € et il refuse d'assurer les hauts risques avec moins de 400 €. L'assureur B est en concurrence avec l'assureur A qui choisit de maintenir sa stratégie tarifaire initiale. Dans ce type de situation, les bas risques vont résilier leurs contrats avec l'assureur A pour acheter la couverture de

l'assureur B. Les hauts risques vont choisir de rester avec l'assureur A puisqu'il leur propose le contrat le plus attrayant.

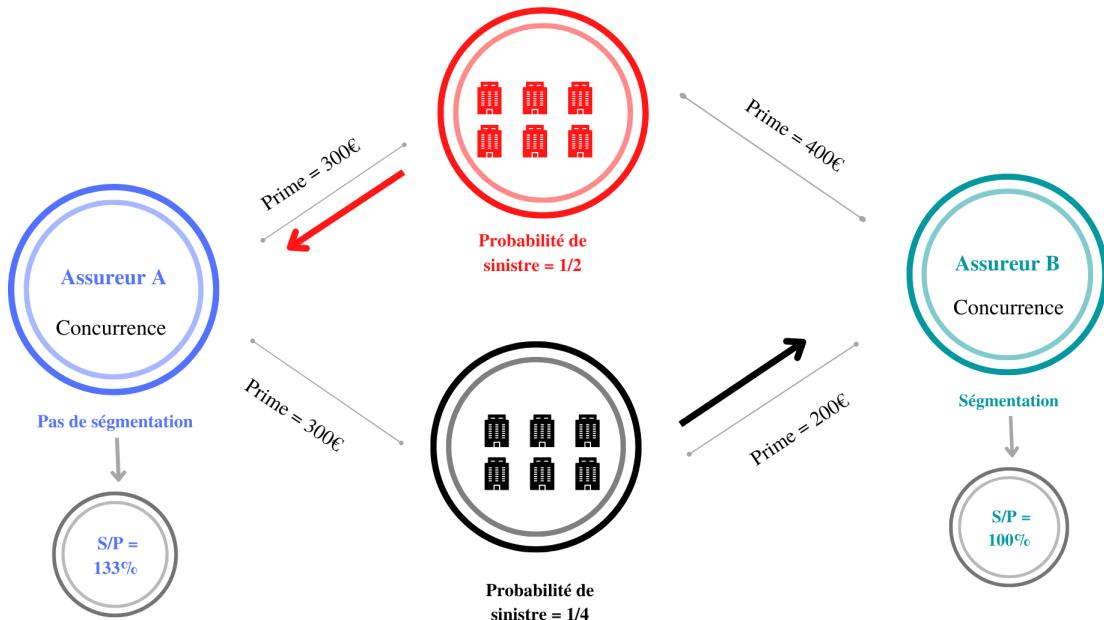


Figure 1.4 – Antisélection : marché avec deux concurrents

L'assureur B finit avec les bas risques dans son portefeuille, ce qui va lui garantir un ratio sinistre à prime égal à 100% (200/200). Tandis que l'assureur A se retrouve avec les hauts risques et un S/P égal à 133% (400/300). La stratégie tarifaire de l'assureur A n'était pas optimale, cet exemple permet de montrer que la sélection dans un marché de deux concurrents est efficace pour assurer sa rentabilité, l'assureur a besoin d'une bonne segmentation et sélection du risque.

#### Que se passe-t-il dans un marché de plusieurs concurrents ?

Imaginons un marché de trois concurrents A, B et C, chacun de ces assureurs vise un S/P égal à 90%. Nous supposons l'existence de 4 profils de risque. Ces profils sont construits à partir de deux variables tarifaires (type d'assuré et type de capital). Chaque profil de risque a un montant de capital différent d'autres profils. Chaque assureur mène une stratégie tarifaire différente : l'assureur A propose la prime mutualiste, l'assureur B propose une prime segmentée selon le type d'assuré et l'assureur C propose une prime segmentée selon le type d'assuré et le type de capital. Pour atteindre le S/P cible, les trois assureurs proposent une prime commerciale qui permet de garantir un S/P égal à 90% calculée comme suit :

$$\text{Prime commerciale} = \frac{\text{Prime pure}}{90\%}$$

Nous supposons que les frais et les charges des trois assureurs sont nuls. Les montants des capitaux proposés sont donnés à titre indicatif, ainsi que les probabilités de survenance de sinistre. Les assurés comparent les différentes offres sur le marché, ils choisissent l'offre la plus attrayante.

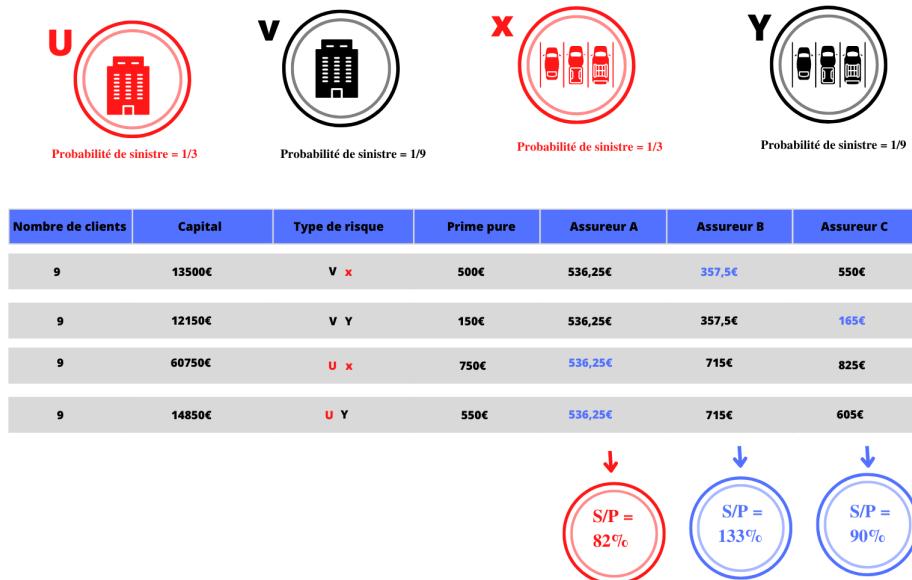


Figure 1.5 – Antisélection : marché avec trois concurrents

Ainsi le calcul permet de trouver que l'assureur A est le plus performant sur le marché avec un S/P égal à 82%. En effet, avec sa stratégie tarifaire, l'assureur A a pu avoir une grande part du marché des profils rentables. Ceci représente un contre-exemple de l'efficacité de la segmentation dans un marché de plusieurs concurrents. De ce fait, proposer un tarif bien segmenté sur un marché concurrentiel ne garantit pas toujours l'optimalité. L'assureur doit faire une sélection non seulement établie sur le risque, mais aussi une sélection établie sur la demande et sur le ciblage de la catégorie la plus rentable du marché.

### 1.2.3 Contexte de l'inflation et risque d'antisélection

L'année 2021 a été marquée par une forte reprise économique après la crise de la Covid-19 en 2020. Cette reprise a été accompagnée par des records successifs d'inflation, en mai 2022 les prix de consommation augmentent de 0,7% en un mois et 5,2% en un an selon l'INSEE<sup>2</sup>. Cette accélération d'inflation se répercute sur les coûts et les charges des assureurs. Selon France Assureurs, en assurance automobile, le coût moyen de sinistre croît avec une tendance annuelle de 3,5 % en matériel durant la dernière décennie. Cette tendance a pour raison principale, la hausse des prix de réparation et des

2. <https://www.insee.fr/fr/statistiques/6455413#tableau-ipc-g1-fr>

pièces détachées. En effet, d'après la SRA (Sécurité et Réparation de l'Automobile)<sup>3</sup>, le coût total de réparation de l'automobile a progressé de +4,1% au premier trimestre 2022 et le coût moyen de la pièce de +25% de 2017 à 2021. L'inflation affecte aussi responsabilité civile corporelle via la hausse du coût horaire de l'aide humaine et les frais généraux des assureurs, notamment : les salaires, le loyer, l'expertise, etc. Par conséquent, face à la hausse des charges, les assureurs devront augmenter leurs tarifs lors du renouvellement et de la souscription des contrats afin de rester solvables. Cette augmentation du tarif peut conduire à une sélection des profils déficitaires sur le marché.

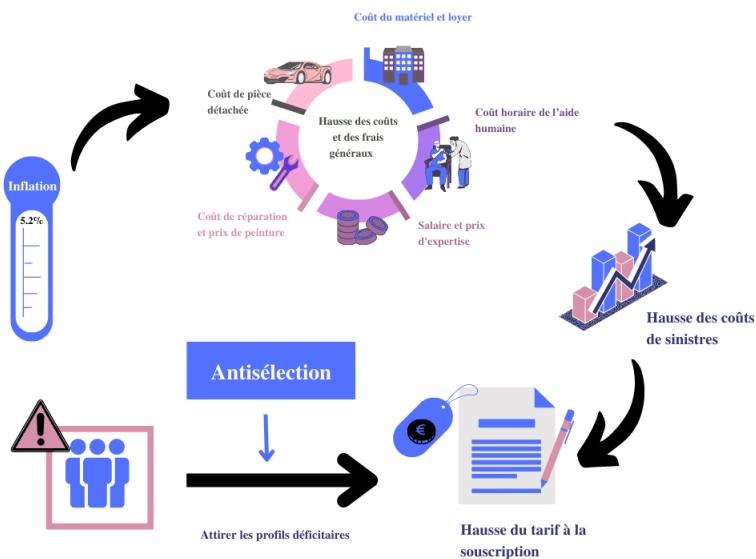


Figure 1.6 – Inflation et risque d'antisélection

## Conclusion partielle

L'antisélection sur le marché concurrentiel est inévitable. La bonne segmentation du tarif ne garantit pas toujours la performance et la rentabilité, notamment dans un marché où nous ignorons les stratégies tarifaires des concurrents. L'inflation est un risque considérable pour les assureurs. Il s'amplifie de plus en plus et accentue le risque d'antisélection. Face à ce contexte d'incertitudes, il est important de déterminer : les profils de risque que l'on attire par la stratégie tarifaire actuelle pour le produit Parc dénommé à travers l'étude du taux de transformation et l'impact d'un changement tarifaire futur sur la structure des profils de risque attirés via l'étude de l'élasticité au prix. Ceci dans le but de prédire la demande future du produit Parc dénommé et d'étudier l'impact de l'inflation sur la rentabilité des affaires nouvelles.

3. <https://www.sra.asso.fr/statistiques/com-statistique-avril-2022/panorama-t1-2022>

# Chapitre 2

## Présentation des données

### Préambule

Ce deuxième chapitre a pour vocation de présenter les données de l'étude. Tout d'abord, il présentera les bases de données utilisées. Ensuite, les traitements effectués et les étapes de construction de la base d'étude. Enfin, il exposera une analyse exploratoire des données.

### 2.1 Construction de la base de données

Les bases de données du produit Parc dénommé sont alimentées mensuellement par les distributeurs (agents et courtiers) à partir d'un outil de souscription nommé OSE Parc. Elles sont stockées sur les serveurs d'AXA France par différents périmètres afin d'optimiser la capacité de stockage. Un serveur SAS permet d'accéder à ces données et de les manipuler. De ce fait, tous les traitements réalisés pour construire la base de données ont été effectués à l'aide du langage SAS.

#### 2.1.1 Présentation des bases utilisées

**Base des devis :** base principale qui contient l'ensemble des informations de chaque devis<sup>1</sup> (statut du devis : transformé ou non, activité de l'entreprise, prime proposée, tonnage, coefficient de réduction, coefficient de sinistralité...). Elle est mise à jour mensuellement. Cette étude est faite sur les données vues au 30 juin 2021 qui concerne les devis sur une durée de 5 ans et 6 mois (du 01 janvier 2016 jusqu'au 30 juin 2021). Celle-ci représente la période où l'ancien tarif était actif. En effet, nous ne pouvons pas mener une étude sur les devis du nouveau tarif du fait d'un manque de volumétrie de don-

1. Un devis est un document non contractuel qui informe l'assuré du tarif des différentes garanties, des franchises et des risques portés par l'assureur. Juridiquement, le devis représente une offre de contrat. L'assuré peut accepter ou refuser cette offre.

nées. Normalement, il faut laisser un temps de vieillissement au devis pour vérifier s'il est concrétisé en contrat, mais dans le cas du produit Parc, on remarque que les devis sont concrétisés rapidement, donc nous n'avons pas besoin d'un vieillissement important. En raison du manque de vieillissement, nous ne considérons que les devis avec une date d'effet inférieure au 30 juin 2021.

**Base des véhicules** : cette base nous permet de récupérer les caractéristiques de l'ensemble des véhicules de devis : date de mise en circulation, type de véhicule (léger, utilitaire...), type de financement (crédit, crédit-bail, autre), garanties choisies... En effet, pour le produit Parc la tarification se fait véhicule par véhicule. La récupération de l'ensemble des informations des véhicules est indispensable pour l'étude. Ces informations permettent d'avoir une vision sur la composition de chaque flotte assurée et également de la bonne définition des profils de risque.

**Base des saisines** : cette base de données contient toutes les informations d'un historique d'un an sur les saisines reçues de la part des distributeurs. La base des saisines est mise à jour mensuellement. Une saisine est un devis qui n'a pas forcément un tarif affiché. C'est-à-dire l'assuré n'a pas renseigné toutes les informations nécessaires pour que l'outil de souscription OSE Parc lui génère le tarif. La base saisine est une base brute qui contient plusieurs colonnes. L'utilisation de cette base nous permet de récupérer certaines variables qui ne sont pas renseignées dans la base devis. Par exemple : numéro de client, date d'émission de devis. La récupération des variables nécessite la jointure de la base des devis avec plusieurs bases saisines qui sont annuelles afin de récupérer les informations sur les 5 ans.

**Base des contrats** : cette base de données contient les informations sur les contrats déjà présents dans le portefeuille, y compris les affaires nouvelles. Cette base est utilisée dans notre étude pour faire les tests de détection des anomalies présentes dans la base devis.

**Base INSEE** : cette base externe permet d'enrichir la base de données avec la variable densité de population de la commune concernée. Cette variable contient 4 modalités :

1. Commune densément peuplée
2. Commune de densité intermédiaire
3. Commune peu dense
4. Commune très peu dense

La jointure de cette base de données avec la base devis se fait par code INSEE.

**L'agrégation des différentes bases de données est effectuée selon les différentes clés communes comme suit :**

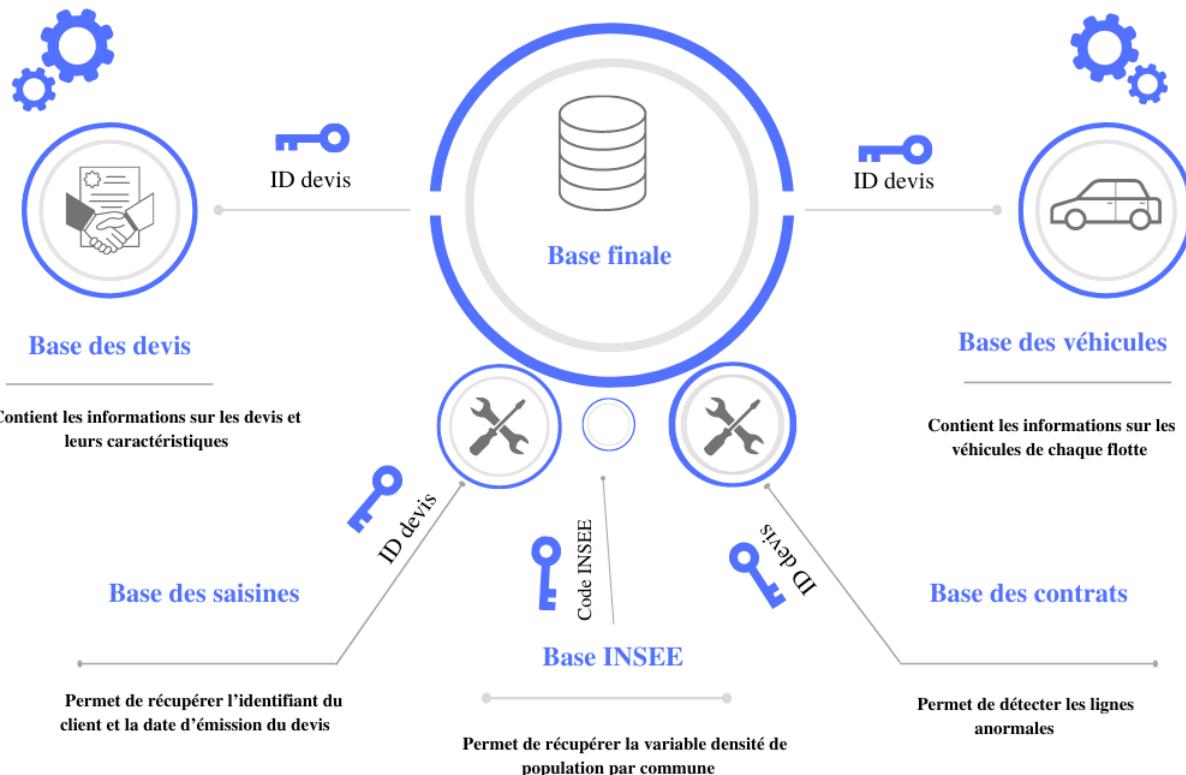


Figure 2.1 – Agrégation des bases de données

## 2.1.2 Traitement des anomalies

La construction de la base de données représente une phase importante de l'étude. En effet pour avoir des résultats exhaustifs et cohérents, il faudra traiter toutes les anomalies de la base. Une base de données non traitée peut mener à des résultats et des conclusions erronés, notamment au moment de la modélisation. La détection des lignes anormales représente la première étape du traitement de la base, c'est une phase qui nécessite l'agrégation de différentes bases.

### 2.1.2.1 Les lignes anormales

Les premières lignes anormales sont les lignes en doublon. Pour la base des devis, la détection de ces lignes se fait par numéro de devis puisqu'il est unique pour chaque prospect. S'agissant de la base des véhicules, la détection des doublons se fait par une clé composée de : numéro de devis, immatricu-

lation du véhicule, marque et date de mise en circulation du véhicule. Normalement, l'immatriculation du véhicule est suffisante pour détecter les doublons de la base véhicule, mais elle n'est pas renseignée d'une manière correcte pour toutes les lignes.

La base principale des devis contient l'ensemble des devis générés par l'outil de souscription OSE Parc. C'est une base brute, il est important de la croiser avec d'autres bases déjà traitées et propres pour fiabiliser les données. La jointure de la base des devis avec la base des contrats, une base propre nous permet de détecter un ensemble de lignes anormales. Il s'agit des devis transformés, mais ils ne sont pas présents dans la base contrat de la même version. Ce problème vient du fait que ces lignes représentent des contrats déjà résiliés. Il s'agit aussi des devis non réalisés, néanmoins ils sont présents dans la base des contrats. En effet, ce problème vient du fait que la durée de vieillissement n'est pas suffisante pour observer leur transformation ou d'une erreur de saisie manuelle.

### 2.1.2.2 Devis doublons

L'assuré peut demander plusieurs devis pour la même flotte. Dans la base des devis, chacun des devis demandés a un identifiant unique. La non-suppression des devis doublons biaise l'estimation du taux de transformation. Par exemple, nous supposons que nous avons dans la base de données 3 lignes comme suit :

ID Devis	Assuré	Flotte	Statut devis
1	A	X	Transformé
2	A	X	Non Transformé
3	B	Y	Non Transformé

En effet, le taux de transformation non biaisé est égal à 1/2 (parmi les deux assurés A et B, seul l'assuré A accepte de transformer son devis en contrat). Tandis que le taux de transformation biaisé est égal à 1/3, ce taux représente la proportion des lignes transformées dans la base de données. Pour supprimer les devis doublons, nous avons besoin d'une variable qui identifie chaque client d'une manière unique. La base des devis utilisée ne contient pas cette information. Nous avons récupéré l'identifiant du client à partir de la base des saisines. Une fois détecté les devis doublons, il est important de déterminer quel devis garder dans la base de données parmi ces devis doublons. Il existe deux possibilités : garder le devis le plus récent demandé par l'assuré ou garder le devis transformé s'il existe et le plus récent sinon. Nous avons opté pour la seconde option parce qu'elle traduit d'une manière fiable le comportement de l'assuré. En effet, nous nous intéressons au comportement des assurés et leur décision d'accepter de souscrire le contrat du produit Parc dénommé. Ainsi l'assuré qui demande plusieurs

devis et n'en transforme aucun, n'est pas intéressé par l'ensemble des tarifs affichés des différents devis. Nous nous intéressons par conséquent au tarif le plus récent, parce qu'il représente le dernier tarif qui a conduit l'assuré à abandonner l'achat des garanties du produit Parc. Concernant l'assuré qui accepte de transformer un devis parmi les devis demandés, nous nous intéressons uniquement au devis transformé puisqu'il représente le devis le plus attrayant pour l'assuré.

#### **2.1.2.3 Devis de remplacement**

Après la signature du contrat, l'assuré doit déclarer tout changement au niveau de sa flotte assurée, notamment la vente ou l'achat d'un nouveau véhicule. Certaines déclarations nécessitent la génération d'un nouveau contrat qui sera précédé par un devis de remplacement. Ce dernier n'est pas considéré comme une affaire nouvelle. Il n'est pas tarifié au tarif affaire nouvelle. Nous devons supprimer ces devis de la base de données parce qu'ils ne concernent pas la version tarifaire actuelle. La détection des devis de remplacement non réalisés n'est pas évidente. Il n'existe pas de variable qui indique le type de devis. Ainsi pour récupérer l'information sur les devis de remplacement, nous avons filtré tout d'abord les devis non réalisés. Ensuite, nous avons croisé la base des devis non réalisés avec la base contrat en utilisant le numéro de contrat. En effet, dans le cas de remplacement, le client préserve le même numéro de contrat qui sera identique au numéro de son devis de remplacement. S'agissant des devis transformés, il existe une variable qui indique si le contrat est une affaire nouvelle ou un remplacement. Par ailleurs, il existe un autre type de devis de remplacement appelé les fausses affaires nouvelles. Il s'agit des devis des clients qui étaient déjà présents dans le portefeuille. Mais, à l'issue d'une déclaration, certains distributeurs proposent au client de résilier l'ancien contrat et de souscrire un nouveau contrat établi à partir de la version tarifaire actuelle. Ceci est dans le but de proposer le tarif le moins cher au client. Le nouveau contrat aura un identifiant différent de l'ancien contrat résilié. Dans cette étude, nous gardons ce type de devis puisqu'il concerne le tarif des affaires nouvelles.

#### **2.1.3 Crédation des variables**

##### **Base des véhicules :**

Tout d'abord, nous avons créé la variable dommage tout accident qui est une variable indicatrice qui prend 1 si le véhicule possède la garantie dommage tout accident et 0 sinon. Ensuite, nous avons créé la variable âge du véhicule à partir de deux variables, la date d'effet de devis et la date de mise en circulation du véhicule et nous avons découpé cette variable en 8 tranches : 1 an, 2 ans, 3 ans, 4 ans, 5 ans, 6 ans, 7-8 ans et 9 ans et plus. Ce découpage est établi sur le fait que nous aimerais avoir un nombre suffisant d'observations dans chaque tranche. Ensuite, nous avons transformé chaque tranche

ainsi que les variables : véhicule de luxe, véhicule de qualité basse, financement par crédit, véhicule léger et véhicule utilitaire en une variable indicatrice de la même manière que la variable dommage tout accident. Enfin, nous avons calculé la proportion de chacune des variables selon le numéro de devis. Le graphe ci-dessous montre les étapes de ces transformations.

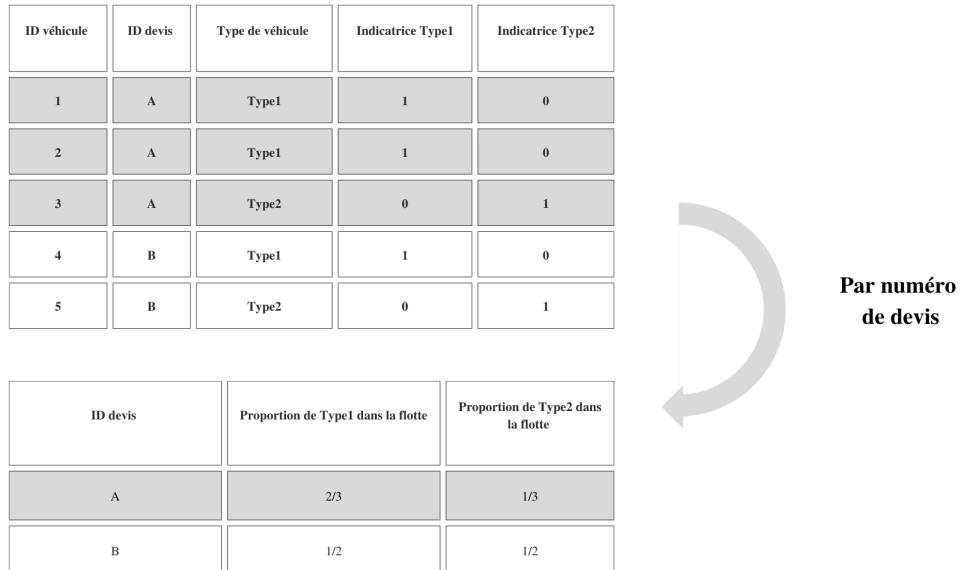


Figure 2.2 – Processus de transformation des variables de la base véhicule

Nous avons récupéré de cette base la prime pour chaque devis en sommant les primes de l’ensemble des véhicules du devis et le nombre de véhicules de chaque devis et la prime moyenne par véhicule, c'est-à-dire la prime sur le nombre de véhicules.

#### Base des devis :

Nous avons créé tout d’abord la variable ancienneté de devis à partir de deux variables : date d’émission (récupérée à partir de la base des saisines) et date d’effet du devis. Ensuite, nous avons regroupé les modalités de certaines variables pour pouvoir observer facilement la tendance des données et éviter les problèmes de corrélation entre les modalités de faible effectif. Par exemple, la variable NAF (Nomenclature d’Activités Française) contient plusieurs modalités avec une proportion inférieure à 5% (Figure 2.3). Nous avons recodé cette variable en trois modalités. Ce codage est basé sur les similitudes en termes de nature d’activité.

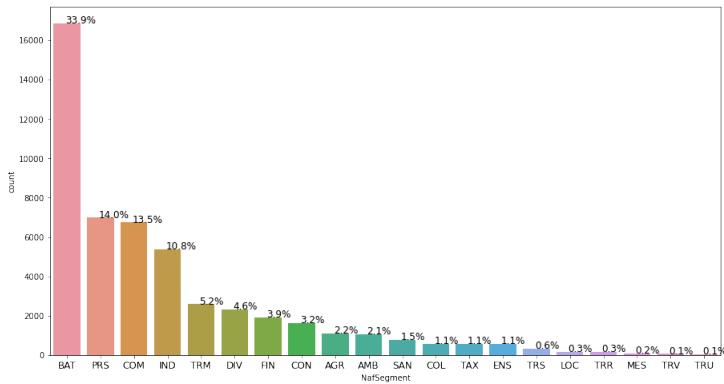


Figure 2.3 – Activité d’entreprise

## 2.1.4 Présentation de la base finale

La base finale est composée de 65949 lignes et 51 variables. L’ensemble des variables ont été sélectionnées sur plusieurs critères. Nous avons éliminé toutes les variables avec une proportion de valeurs manquantes supérieure à 50%, il s’agit principalement des variables qui portent des informations sur les anciens assureurs et les antécédents de sinistres. Il ne s’agit pas de variables tarifaires, c’est la raison pour laquelle la plupart des assurés ne déclarent pas ces informations au moment de la souscription. Nous avons aussi éliminé les variables qui ne fournissent aucune information discriminante. Par exemple, 99,16% des véhicules ont la garantie catastrophe naturelle. Ensuite, nous avons éliminé les variables qui portent une information redondante. Par exemple, le coefficient de sinistralité pour la garantie dommage est calculé à partir du coefficient de sinistralité de la garantie responsabilité civile. Les 51 variables sélectionnées représentent l’ensemble des informations du Parc qui peuvent être regroupées comme suit :

- **Nument** : cette variable représente l’identifiant du devis, elle définit chaque devis d’une manière unique.
- **Caractéristiques de la flotte** : le but est de déterminer le comportement de chaque segment du produit Parc dénommé, selon le type de l’activité de l’entreprise, la région de l’entreprise, la composition de la flotte en termes de type de véhicule.
- **État du devis** : notre variable d’intérêt, elle est composée de 2 modalités : PAN ; projet affaire nouvelle ou devis non réalisé, CAN ; contrat affaire nouvelle ou devis transformé. Pour faire nos analyses, une nouvelle variable **tnf** a été créée, cette dernière est composée de 2 modalités : 0 si PAN et 1 si CAN.
- **Caractéristiques du distributeur** : il s’agit essentiellement de deux variables. La variable type de distributeur est composée de deux modalités agent et courtier et la notation du dis-

tributeur, cette variable est calculée à partir de plusieurs paramètres, notamment le chiffre d'affaires des distributeurs. Elle permet de classifier les distributeurs selon leur performance et leur rentabilité. Cette variable est composée de 3 modalités qui représentent 3 notations.

- **Prime et coefficient technique** : la prime représente le déterminant principal de la transformation du devis. En effet, un assuré peut accepter la transformation s'il estime que la prime proposée est la plus attrayante sur le marché. Nous ajoutons à la prime, le coefficient technique. C'est un coefficient d'augmentation/réduction appliqué par les agents et les courtiers. En effet, on peut s'attendre à ce qu'un client bénéficiant d'une forte réduction tarifaire transforme plus son devis en contrat qu'un client payant le prix fort.
- **Coefficient sinistralité** : l'historique de sinistralité de l'assuré peut impacter la probabilité de transformation. En effet, les assurés qui ont un niveau de sinistralité élevé sont plus motivés pour transformer leur devis. De ce fait, nous calculons le coefficient de sinistralité en comparant la fréquence réelle du Parc et la fréquence modélisée, avec un facteur de crédibilité en fonction de la taille de la flotte.

Nous faisons la remarque que toutes les variables de la base d'étude ne seront pas utilisées dans les modélisations, certaines ont été supprimées suite à l'étude des corrélations.

## 2.2 Étude exploratoire

### 2.2.1 Analyse du taux de transformation

L'objectif est de comprendre les relations qui existent entre le taux de transformation et les variables explicatives ainsi que les liaisons qui existent entre les variables explicatives. Nous définissons le taux de transformation comme suit :

$$\text{taux de transformation} = \frac{\text{nombre de contrats}}{\text{nombre de devis} + \text{nombre de contrats}}$$

En d'autres termes, le taux de transformation représente la probabilité de souscrire un contrat. Il est égal à 47,2 % sur l'ensemble de la base de données. Par ailleurs, il varie selon les variables explicatives.

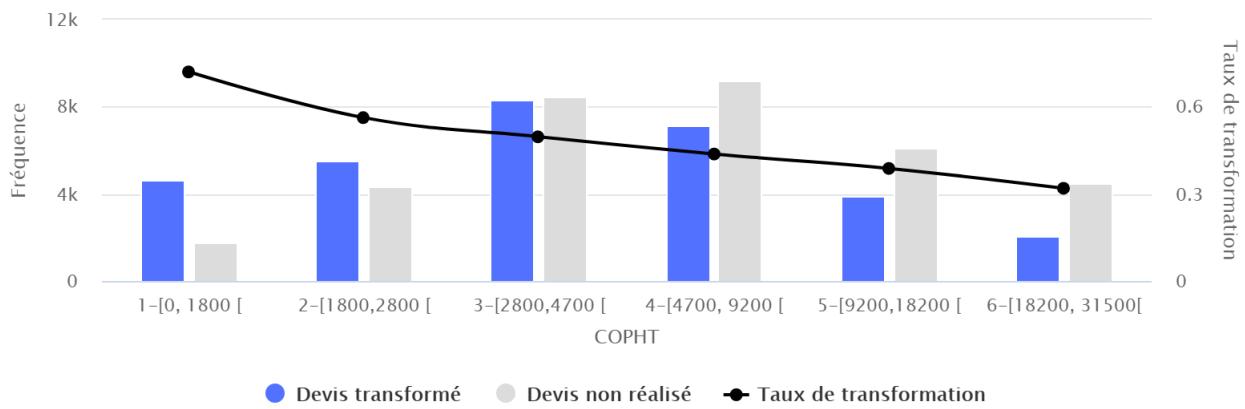


Figure 2.4 – Taux de transformation par segment de prime

Pour la prime (figures 2.4), nous observons une tendance à la baisse : plus la prime augmente, plus le taux de transformation diminue. Cependant, le tarif est un facteur fortement corrélé à la probabilité de souscrire le contrat.

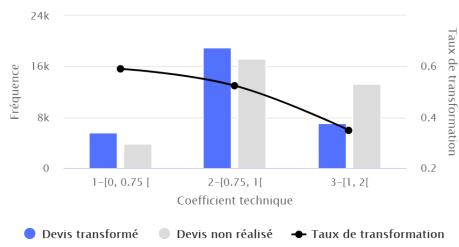


Figure 2.5 – Taux de transformation par coefficient technique

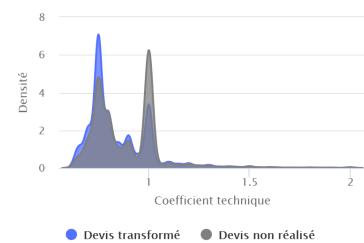


Figure 2.6 – Densité coefficient technique

À côté de la prime, on trouve le coefficient technique (figures 2.5 et 2.6), on remarque que le taux de transformation est décroissant en fonction des classes du coefficient technique. Concernant la distribution du coefficient technique, nous remarquons un pique autour de 0,75. Il s'agit du coefficient minimum recommandé aux agents et aux courtiers par AXA France. Nous remarquons aussi qu'il existe beaucoup de rabais pour les devis transformés en contrat (la courbe bleue) comparativement aux devis non réalisés. Ceci peut être expliqué par le fait que le client est plus intéressé par les devis avec un rabais commercial.

S'agissant des distributeurs (figure 2.7), nous constatons que le taux de transformation est du réseau des agents est supérieur à celui du réseau des courtiers. Ce qui peut être expliqué par le fait que les courtiers proposent une offre variée au client.

S'agissant de la variable région (figure 2.8), la région Ouest est la première en termes de nombre d'affaire nouvelle et du taux de transformation avec 52,38% et DROM est la dernière avec un taux égal à 27,33%. Cette région contient un faible effectif d'observations (environ 0,2%). Pour modéliser le taux de transformation, cette modalité a été regroupée avec la modalité IDF (Île-de-France).

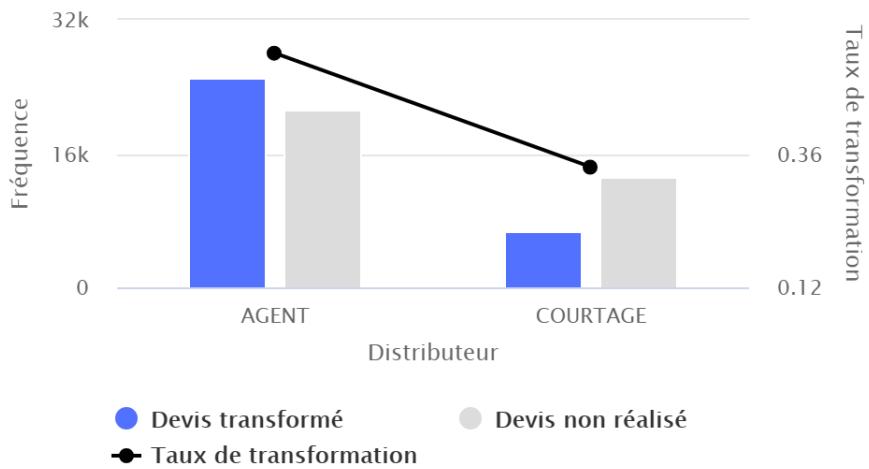


Figure 2.7 – Taux de transformation par réseau

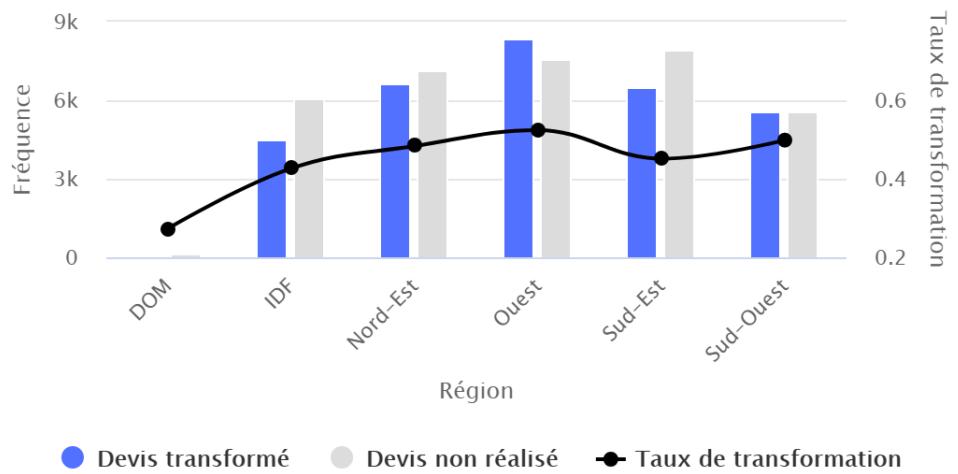


Figure 2.8 – Taux de transformation par région

L'année d'effet représente l'année à laquelle les garanties du devis entrent en vigueur si sa transformation a eu lieu. L'année 2016 a un faible effectif d'observations par rapport aux autres années. Ceci s'explique par le fait que l'outil de souscription OSE Parc a été lancé à la moitié de l'année 2016 et même s'il existait des devis émis à travers OSE Parc, la plupart de ces devis avaient une date d'effet en 2017. En outre, la cohabitation durant la seconde moitié de 2016 entre cet outil et l'ancien outil a réduit le nombre de devis émis par l'outil OSE Parc. Concernant la période de la crise de covid-19, le nombre de devis à date d'effet en 2020 est inférieur au nombre de devis à date d'effet en 2019. En revanche, le taux de transformation est resté constant entre les deux années. Concernant l'année 2021, il existe moins de devis à date d'effet en cette année par rapport à l'année 2020. Ceci s'explique par le fait qu'il existe une partie de ces devis émises en 2020 durant la crise de covid-19 et que la version des données de l'étude s'arrête au 30 juin 2021.

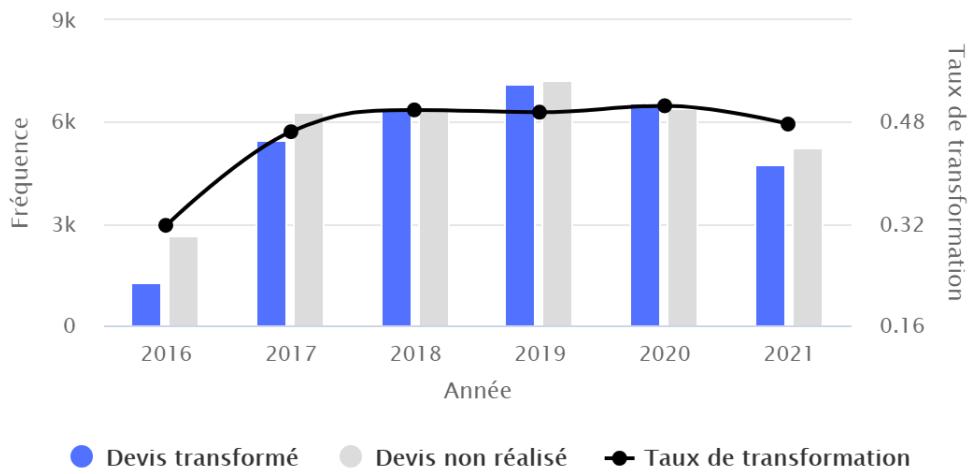


Figure 2.9 – Taux de transformation par année d’effet

## 2.2.2 L’analyse des corrélations

Corrélation entre les variables quantitatives :

La corrélation linéaire représente la liaison linéaire qui existe entre les variables et qui contredit leur indépendance. Ainsi, l’analyse de corrélation permet de détecter les variables qui apportent une information redondante. Ceci, dans le but de réduire la dimension du jeu de données. La corrélation linéaire se détecte à travers le calcul de différentes métriques. Dans notre cas, nous utilisons le coefficient de corrélation de Pearson. Le coefficient de corrélation de Pearson  $\varphi_{i,j}$  entre deux variables  $X_i$  et  $X_j$  d’écart-types finis  $\sigma_i$  (respectivement  $\sigma_j$ ) est défini par :

$$\varphi_{i,j} = \frac{\text{cov}(X_i, X_j)}{\sigma_i \sigma_j}$$

Ainsi, la matrice de corrélation des variables  $X_1, X_2, \dots, X_n$  est définie par le terme  $(\varphi_{i,j})_{1 \leq i, j \leq n}$ . Cette matrice est symétrique et ses éléments diagonaux sont égaux à 1.

Nous constatons (figure 2.10) qu’il y a une corrélation positive entre la prime et le coefficient technique (0,38). Cette corrélation est liée au fait que la prime est calculée à partir du coefficient technique. Par ailleurs, nous remarquons une corrélation négative (-0,47) entre les véhicules légers et les véhicules utilitaires ce qui s’explique par le fait que ces 2 variables représentent la composition du Parc et sont donc complémentaires. Nous constatons une corrélation positive entre les véhicules de moins 1 an et la garantie dommage, ce qui s’explique par le fait que plus le véhicule est récent, plus l’assuré souhaite couvrir au maximum son véhicule, et donc acheter la garantie dommage tout accident. Pour la suite de l’étude, nous ne supprimons aucune variable puisque toutes les corrélations observées sont inférieures à 60%.

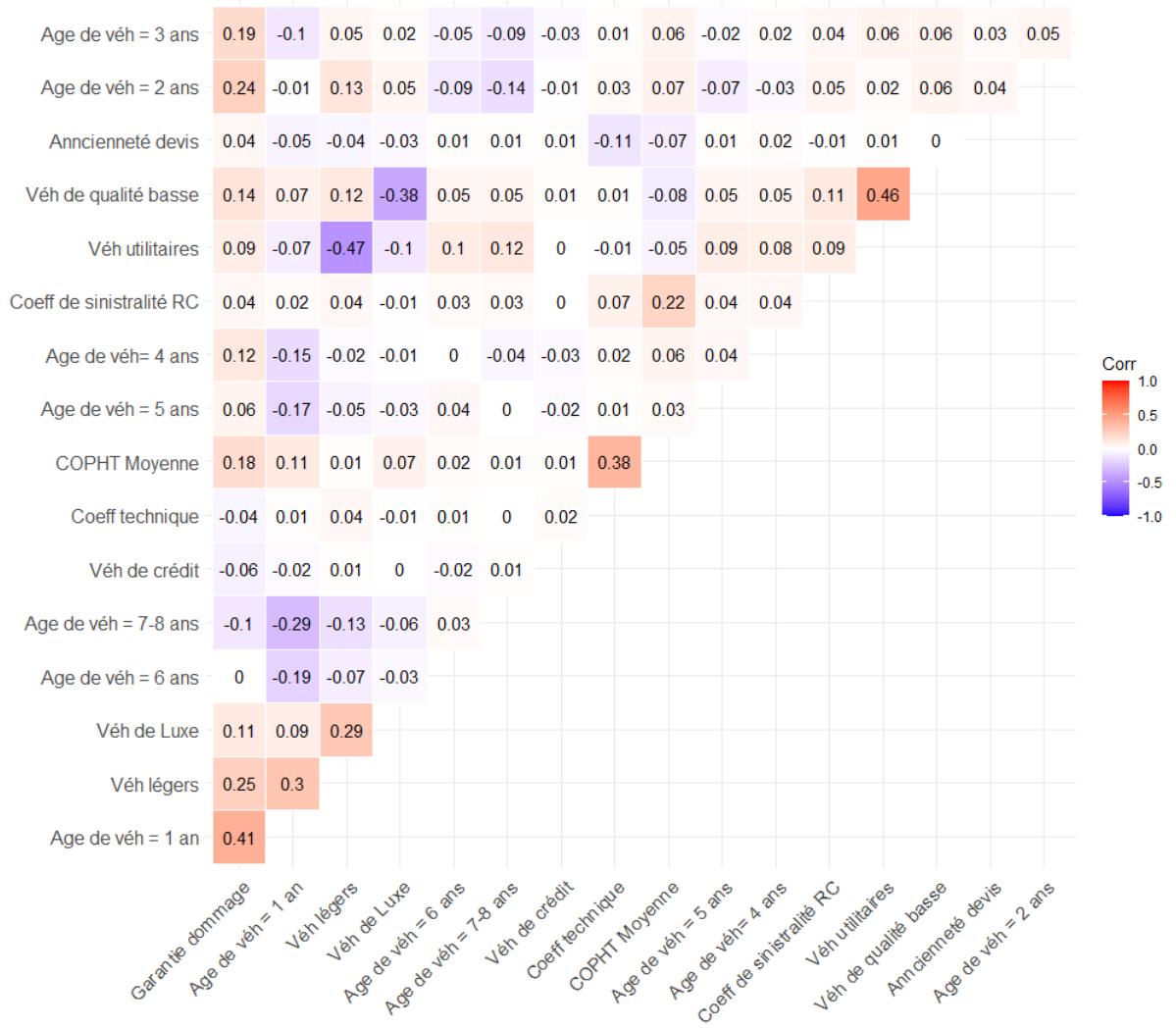


Figure 2.10 – Corrélation entre les variables numériques

### Association entre les variables qualitatives :

L’association est la liaison qui existe entre deux variables catégorielles. L’étude des associations est indispensable pour éliminer les variables qui contiennent des informations redondantes. Le test d’indépendance de  $\chi^2$  permet de prendre une décision sur l’association de deux variables catégorielles. L’hypothèse nulle de ce test suppose qu’il existe indépendance entre les variables en question. Pour deux variables  $U$  et  $V$  de nombre de modalités respectivement  $I$  et  $J$ , la statistique qui permet

d'effectuer ce test est donnée comme suit :

$$D_{U,V} = N \sum_{i=1}^I \sum_{j=1}^J \frac{(O_{i,j} - T_{i,j})^2}{T_{i,j}}$$

Avec :

- $O_{i,j} = \frac{1}{N} \sum_{n=1}^N \mathbb{1}_{\{U_n=u_i, V_n=v_j\}}$ ,  $N$  le nombre d'observations,  $u_i$  la  $i^{\text{ème}}$  modalité de la variable  $U$  et  $v_j$  la  $j^{\text{ème}}$  modalité de la variable  $V$ .
- $T_{i,j} = \frac{1}{N^2} \sum_{n=1}^N \mathbb{1}_{\{U_n=u_i\}} \sum_{n=1}^N \mathbb{1}_{\{V_n=v_j\}}$ .

La statistique  $D_{U,V}$  suit la loi  $\chi^2_{(I-1, J-1)}$ . De ce fait, si  $D_{U,V}$  dépasse le quantile  $q_{\chi^2_{(I-1, J-1)}}^\alpha$ , l'hypothèse nulle est rejetée avec un risque de  $100 * \alpha\%$ . Pour mesurer et tester l'intensité des associations, nous calculons le coefficient de  $V - cramer$ , il s'agit d'une version normalisée de  $D_{U,V}$  :

$$V_{U,V} = \sqrt{\frac{D_{U,V}/N}{\min(I-1, J-1)}}$$

Ce coefficient est interprété comme suit :

- $V_{U,V}$  proche de 0 signifie qu'il n'existe pas d'association entre  $U$  et  $V$ .
- $V_{U,V}$  proche de 1 signifie qu'il existe une forte association entre  $U$  et  $V$ .

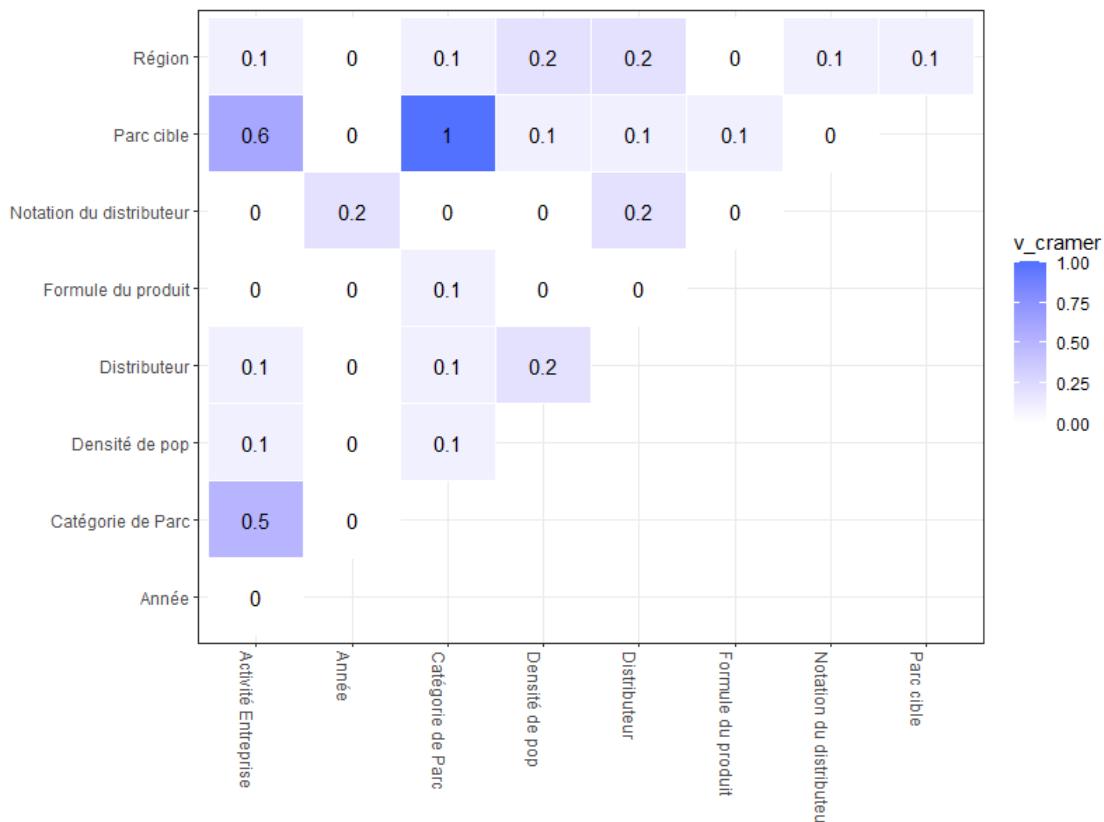


Figure 2.11 – Association entre les variables catégorielles

La figure (2.11) représente le résultat de ce coefficient pour les paires de variables catégorielles de la base de données. Nous observons qu'il existe une forte association entre la catégorie de Parc et la variable Parc cible. Les catégories cibles sont les catégories du Parc rentables sur lesquelles AXA France veut se développer. Ces catégories bénéficient de rabais importants à la souscription relativement aux catégories non cibles. Pour la partie modélisation et étude de l'élasticité au prix, nous supprimons la variable catégorie de Parc. Concernant l'activité de l'entreprise, il existe une association (0,5) avec la variable catégorie de Parc, ce qui s'explique par le fait que certaines catégories de Parc ne concernent que quelques activités, par exemple, les engins agricoles concernent généralement les entreprises qui exercent une activité agricole.

#### Diagnostic de la forte colinéarité :

La forte colinéarité est un problème qui survient lorsqu'une variable peut être écrite sous forme d'une combinaison linéaire des autres variables explicatives. Ceci peut représenter un problème dans le modèle GLM (en particulier la régression logistique). En effet, la forte colinéarité peut augmenter la variance des coefficients de régression et les rendre instables et difficiles à interpréter et par la suite biaiser les résultats du modèle. Elle peut donner des coefficients avec des signes qui ne reflètent pas la réalité. L'absence de la colinéarité est une condition indispensable pour estimer le modèle GLM. Il est important de noter que l'existence d'une forte corrélation ou association implique forcément l'existence de la colinéarité entre les variables, mais la réciproque n'est pas toujours vraie. Pour détecter la colinéarité entre les variables (numériques et catégorielles), une méthode supplémentaire sera introduite qui est l'Analyse Factorielle de Données Mixtes (AFDM).

L'AFMD (en anglais FAMD) tire ses origines des travaux de (Escoffier, 1979)[ESC79] c'est une méthode destinée à l'analyse des jeux de données contenant à la fois les variables numériques et catégorielles. C'est une méthode mixte entre l'analyse de la composante principale (ACP) et l'analyse des correspondances multiples (ACM). Le lecteur peut se référer à (Escoffier et Pagès, 2008) [J P08] pour l'ACP et l'ACM. L'idée de l'AFDM est d'utiliser à la fois l'ACP et l'ACM pour compresser l'information portée par les variables numériques et catégorielles dans un nombre réduit de facteurs d'une manière itérative en réduisant à chaque étape la variance résiduelle.

Soit  $N$  le nombre de variables numériques et  $C$  le nombre de variables catégorielles. L'AFMD cherche à maximiser l'inertie d'un espace de dimension  $P$  (le nombre de variables explicatives  $C+N$ ) projetée à un espace de dimension inférieure  $R$ . Soit  $F_1$  le premier facteur créé par L'AFMD. Ce facteur est représenté par le vecteur unitaire  $u$ , le vecteur propre de la matrice des liaisons des  $P$  variables. Ce vecteur est associé à la valeur propre  $VP_1$  donnée comme suit :

$$VP_1 = \sum_{i=1}^N r^2(F_1, X_i) \sum_{i=N+1}^P \eta^2(F_1, X_i)$$

Avec :

- $r^2(F_1, X_i)$  : le coefficient de corrélation au carré entre  $F_1$  et la variable numérique  $X_i$
- $\eta^2(F_1, X_i)^2$  : le rapport de corrélation au carré entre  $F_1$  et la variable catégorielle  $X_i$

La mesure de l'inertie projetée sur chaque axe factoriel se fait à travers le rapport de la variance expliquée et la variance totale des données, elle peut se faire aussi à travers l'étude des valeurs propres. La contribution de chaque variable explicative dans la construction de chaque axe factoriel est donnée comme suit :

$$CTB_l = \frac{c_{il}^2}{\sum_{p=1}^P c_{pl}^2}$$

Avec :

- $l$  représente l'axe factoriel et  $p$  la variable explicative en question.
- $c_{il}^2$  représente la distance au carré entre la variable explicative  $i$  et l'axe factoriel  $l$ .

Il existe une forte colinéarité lorsque l'AFMD arrive à réduire la dimension initiale à un espace de dimension très inférieure tout en gardant une grande partie de l'inertie expliquée. Ceci peut être traduit par l'existence de plusieurs axes factoriels avec des valeurs propres proches de 0.

Nous avons implémenté l'AFMD sous R à l'aide du package *FactoMineR*. En observant la figure (2.13), nous remarquons que les 5 premiers axes préservent seulement 26% de l'inertie expliquée. L'analyse des valeurs propres donne que tous les axes ont des valeurs propres supérieures à 0.8 et uniquement 7 variables ont des valeurs propres inférieures à 1, ce qui montre l'absence de l'existence de la colinéarité parfaite. La figure (2.12) donne la contribution en pourcentage des variables explicatives dans la construction des 5 premiers axes. Pour le premier axe, nous remarquons qu'il existe une forte association entre la variable Parc cible et catégorie du Parc, ce qui confirme les résultats du coefficient de v-cramer. Il existe aussi une corrélation entre ces deux variables et la prime moyenne, ce qui peut s'expliquer par le fait que la tarification se fait par chaque catégorie de véhicule. Puisqu'il existe des corrélations significatives entre certaines variables, pour modéliser le taux de transformation, nous allons utiliser un algorithme de sélection de variables pour éliminer les variables qui n'apportent pas d'information discriminante supplémentaire et prendre la combinaison des variables qui maximise le pouvoir prédictif.

---

2.  $\eta^2(x, y) = \frac{SCE_{inter}}{SCE_{totale}}$ .  $SCE_{inter} = \sum_{i=1}^n \sum_{j \in J_i} J(\bar{y}_{ij} - \bar{y})^2$  et  $SCE_{totale} = \sum_{i=1}^n \sum_{j \in J_i} (\bar{y}_{ij} - \bar{y})^2$   
 -  $x$  représente la variable catégorielle de  $J$  modalités et  $y$  la variable numérique.  
 -  $\bar{y}$  représente la moyenne de  $y$ .  
 -  $\bar{y}_{ij}$  représente la moyenne de  $y$  des individus de la modalité  $j$  de  $x$ .  
 -  $\bar{y}_{ij}$  représente l'observation  $i$  de la variable  $y$  si elle appartient à la modalité  $j$  de  $x$ .

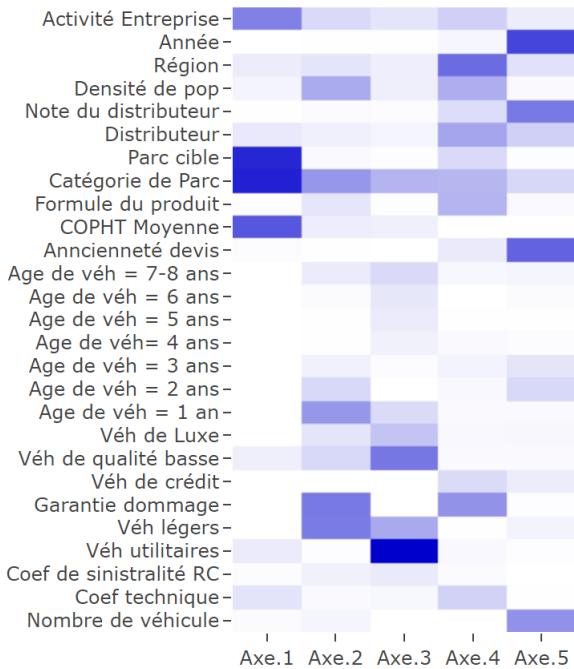


Figure 2.12 – Contribution des variables explicatives aux axes de l'AFDM

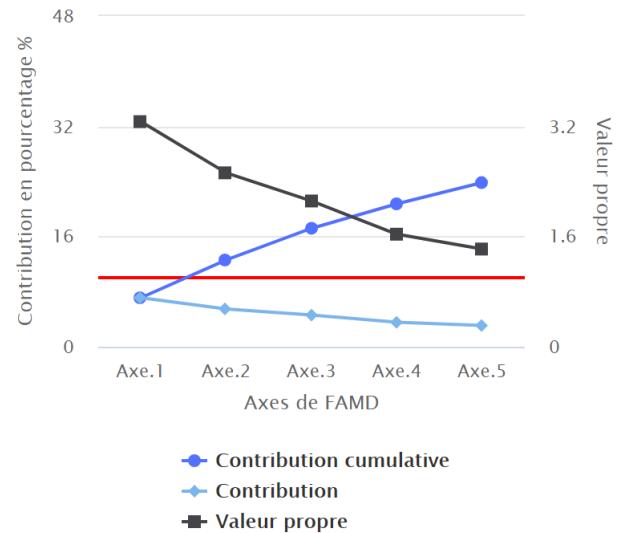


Figure 2.13 – Contribution de chaque axe à l'inertie totale

## Conclusion partielle

Les statistiques descriptives montrent que la prime a un impact important sur le taux de transformation, une tendance croissante du taux de transformation à mesure que la prime baisse. Toutefois, le taux de transformation décroît selon les segments du coefficient technique. L'analyse des corrélations et le diagnostic de la colinéarité permettent d'observer qu'il existe des corrélations entre les variables explicatives ce qui nécessite une sélection de variables pour optimiser la modélisation du taux de transformation.

# Chapitre 3

## Modélisation du taux de transformation

### Préambule

L'objectif de ce chapitre est de proposer un modèle pour prédire l'acte de la transformation d'un devis en contrat. Cette modélisation permet de déterminer les segments des clients qui transforment plus<sup>1</sup> et les segments des clients qui transforment moins. Il s'agit d'une problématique de classification binaire classique. La variable d'intérêt  $y$  est une variable binaire qui prend 1 si le client a accepté de souscrire le contrat du produit Parc dénommé et 0 sinon et les variables explicatives représentent l'ensemble des caractéristiques du contrat. Nous traitons cette problématique sous deux approches : la première est statistique via le modèle GLM et la seconde est algorithmique à travers un modèle de machine learning qui est l'Xgboost. Pour optimiser les modèles, nous utilisons la méthode de recherche par grille et l'algorithme de sélection récursive de variables. L'objectif principal est de proposer un meilleur prédicteur de l'acte de transformation. Ainsi, pour comparer les deux modèles, nous utilisons les mêmes ensembles d'apprentissage et de test et nous évaluons la qualité prédictive des deux modèles avec les mêmes métriques d'évaluation. L'interprétation du modèle GLM se fait via l'étude des coefficients et du modèle Xgboost à travers la théorie des Shap values et le graphique de dépendance partielle (PDP).

### 3.1 Cadre théorique

L'objectif de cette section est de cerner le contexte théorique des modélisations effectuées ainsi que les différentes méthodes d'optimisation et d'interprétation des modèles. Ceci dans le but de comprendre le fonctionnement des outils utilisés et leur pertinence pour modéliser le taux de transforma-

---

1. c'est-à-dire les segments avec une grande probabilité de souscrire au contrat Parc dénommé

tion.

### 3.1.1 Modèle linéaire généralisé (GLM)

L'intérêt d'utiliser le modèle GLM pour modéliser le taux de transformation est sa structure linéaire facile à interpréter. Ce modèle a été formulé par John Nelder et Robert Wedderburn (1972). En raison de son opérationnalité, il est largement utilisé pour répondre aux différentes problématiques actuarielles.

Le modèle linéaire généralisé :

Le modèle linéaire généralisé, abrégé en anglais GLM, est une extension de la régression linéaire multiple dans le cas où la variable dépendante n'est pas forcément quantitative continue. En effet, il permet de mettre une similitude entre la variable dépendante aléatoire et une ou plusieurs variables explicatives déterministes qui peuvent être quantitatives ou qualitatives à travers une fonction lien. Celle-ci peut appartenir aux différentes classes de distribution comme celle de la loi Bernoulli, Poisson, Gamma et même normale. Ainsi, nous définissons le modèle linéaire généralisé par les trois composantes suivantes [Wik] :

- **La composante aléatoire** : la loi de probabilité de la variable dépendante  $y$ . Cette loi appartient à la famille exponentielle, c'est-à-dire sa densité associée se présente comme suit :

$$f(y_i, \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - v(\theta_i)}{u(\phi)} + w(y_i, \phi)\right)$$

Avec  $\phi$  le paramètre de dispersion (connu) et  $\theta$  le paramètre canonique (inconnu).  $u$ ,  $v$  et  $w$  sont des fonctions connues et déterminées à partir de la famille exponentielle du modèle.

- **La composante déterministe** : supposons que nous avons  $p$  variables indépendantes et déterministes que nous classons dans la matrice  $X$ . Alors, la composante déterministe  $X\beta$  un vecteur de dimension  $n$ . Avec  $n$  le nombre d'observations et  $\beta$  un vecteur composé des paramètres estimés  $\beta_j$  du modèle.
- **La fonction lien** : c'est une relation fonctionnelle entre la composante aléatoire du modèle et la composante déterministe, elle est donnée comme suit :

$$g(\mu_i) = \eta_i = X'_i \beta \quad i = 1, \dots, n$$

Avec :

- $\mu_i = E(Y_i)$  la moyenne de la variable  $Y_i$ .
- $g$  la fonction lien qui est monotone et différentiable.

Ainsi, le modèle linéaire généralisé est un modèle qui met en relation chacune de ces moyennes et la matrice des variables déterministes. Il est défini comme suit :

$$g(\mu_i) = X'_i \beta \quad i = 1, \dots, n$$

$\beta$  est estimé via la maximisation de la fonction de vraisemblance suivante :

$$L(y_1, y_2, \dots, y_n, \beta) = \prod_{i=1}^n f(y_i, \beta)$$

Avec  $f$  qui représente la fonction de densité et qui appartient à la famille exponentielle. La maximisation de la vraisemblance est équivalente à la résolution des équations de score suivantes [Ala17] :

$$\bar{U}(\beta) = \sum_{i=1}^n x_i \frac{1}{g'(\mu_i) V(\mu_i)} (y_i - \mu_i(\beta)) = \sum_{i=1}^n U_i(\beta) = 0$$

L'estimateur  $\hat{\beta}$  solution de ces équations satisfait de très bonnes propriétés statistiques classiques (consistance, normalité asymptotique et efficacité).

### Régression logistique :

Le modèle de régression logistique est un cas particulier du modèle linéaire généralisé où la fonction de lien est la densité d'une loi de Bernoulli. La régression logistique est utilisée, généralement, pour modéliser la survenance d'un événement aléatoire. Supposant que l'on soit dans le même cadre que présenté ci-dessus, la distribution de  $y$  suit une loi de Bernoulli signifie que  $y$  prend la valeur 1 lorsque l'événement aléatoire se produit, et 0 sinon. Ainsi, la régression logistique qui permet d'expliquer la survenance de cet événement est donnée par le modèle linéaire de la forme suivante :

$$g(X) = \text{Logit}(\pi(X)) = \beta_0 + \sum_{i=1}^p \beta_i x_i$$

Avec :

- $\pi(X) = P(Y = 1 | x_1, \dots, x_p)$  la probabilité de survenance de l'événement conditionnellement aux variables explicatives ;
- $g(X) = \text{Logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$  ;

—  $(\beta_0, \dots, \beta_p)$  les coefficients de la régression ;

— La relation entre  $\pi(X)$  et  $g(X)$  est donnée comme suit :  $\pi(X) = \frac{e^{g(X)}}{1+e^{g(X)}}$ .

Le rapport  $\frac{\pi(x)}{1-\pi(x)}$  est appelé aussi l'odds ratio (OR). Ce rapport donne la probabilité de survenance d'évènement sur la probabilité de non survenance. Il est positif, il peut être interprété comme suit :

- $OR = 1$  ou  $\log(OR) = 0$  : signifie que la probabilité de survenance est égale à 50%.
- $OR > 1$  ou  $\log(OR) > 0$  : signifie que la probabilité de survenance est supérieure à 50%.
- $OR < 1$  ou  $\log(OR) < 0$  : signifie que la probabilité de survenance est inférieure à 50%.

### Régression régularisée (Lasso) :

La régression régularisée est une version optimisée du modèle linéaire généralisé. Il s'agit d'estimer convenablement les coefficients par l'introduction d'une pénalité  $\lambda$ , dans le but d'éviter le problème de sur-apprentissage et d'améliorer les performances prédictives. Par ailleurs, l'estimation des coefficients de la régression se fait par la résolution d'un problème d'optimisation. Il s'agit de déterminer les coefficients qui minimisent une fonction de perte  $E(\beta)$ . Celle-ci s'écrit sous la forme d'une somme de deux termes : (i) la log-vraisemblance qui minimise le biais ; (ii) le terme de pénalité qui permet de minimiser la variance. Pour le cas du lasso, la fonction de perte s'exprime comme suit :

$$E(\beta) = -\ln(L(\beta)) + \lambda \|\beta\|_1$$

Avec :

—  $\ln(L(\beta)) = \frac{1}{n} \sum_{i=1}^n [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$  la logvraisemblance avec  $y_i \in [0, 1]$

$$— p_i = \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_{ij}}}$$

$$— \|\beta\|_1 = \sum_{i=1}^n |\beta_i|$$

La pénalité  $\lambda$  peut être considérée comme une pondération qui classifie les variables selon l'ordre d'importance. Ainsi, à travers  $\lambda$  le Lasso peut arbitrairement supprimer les variables qui contiennent une information redondante. Ceci peut causer un problème dans la mesure où le lasso peut supprimer certaines variables que l'on considère économiquement importantes.

### 3.1.2 Le modèle Xgboost

L'une des limites du modèle GLM est sa structure linéaire qui ne permet pas de capter les interactions non linéaires qui peuvent exister entre les variables explicatives. Une alternative à ce modèle est l'Xgboost qui est un modèle d'apprentissage automatique, apparu pour la première fois en 2016 dans un article de Carlos Guestrin et Tianqi Chen [Tia16]. Depuis son introduction, cet algorithme a connu une large utilisation par le comité scientifique. Il présente plusieurs avantages, notamment sa capacité de généraliser les résultats de prédiction en rajoutant des pénalités lors de l'optimisation et sa rapidité de calcul.

Apprentissage automatique supervisé :

Dans le cadre d'apprentissage automatique supervisé, l'échantillon de l'étude  $D_n = (Z_i)_{i \in \{1, 2, \dots, n\}}$  est découpé en deux éléments :  $X_i \in \mathbb{R}^P$  représente les variables d'entrée ou les features de la  $i^{\text{ème}}$  observation et  $Y_i \in \mathbb{R}$  représente le label ou l'étiquette. L'objectif d'un modèle d'apprentissage automatique est de prédire le label d'une nouvelle entrée  $X_{n+1}$  à partir de l'échantillon  $D_n$ . En apprentissage automatique, la variable  $Z$  est supposée suivre une loi  $\mathbb{P}$ . La prédiction du label d'une nouvelle entrée se fait via la fonction de prédiction  $f$  qui est supposée être une application mesurable de l'ensemble des entrées vers l'ensemble des labels. Le but principal d'un algorithme d'apprentissage automatique est de trouver la meilleure fonction de prédiction  $f^*$  qui minimise l'écart entre le label réel  $Y_{n+1}$  et le label prédit  $\hat{Y}_{n+1}$ . L'écart entre les deux labels est mesuré à travers une fonction à valeurs réelles appelée fonction de perte  $l$ . En pratique, la détermination du meilleur prédicteur avec l'échantillon  $D_n$  se fait via la minimisation du risque empirique :

$$\hat{R}_{D_n}(f) = \frac{1}{n} \sum_{i=1}^n l(Y_i, f(X_i))$$

Ce qui revient à déterminer le prédicteur qui minimise l'espérance empirique de la fonction perte. Dans le cas de la classification, pour une perte binaire ( $l(\hat{Y}_{n+1}, Y_{n+1}) = \mathbb{1}_{\{\hat{Y}_{n+1} \neq Y_{n+1}\}}$ ) le meilleur prédicteur qui minimise le risque empirique est appelé prédicteur de Bayes :

$$f_{Bayes} = \mathbb{1}_{\{\mathbb{E}_{\mathbb{P}}[Y|X] \geq 1/2\}}$$

Avec  $\mathbb{E}_{\mathbb{P}}[Y|X]$  représente l'espérance conditionnelle de la variable  $Y$  sachant les variables d'entrée  $X$ . Dans le cadre de la classification, cette espérance est égale à :  $\mathbb{P}[Y = 1|X]$ .

## Les arbres de décision CART :

CART (Classification and Regression Tree) est un ensemble d'arbres de décision permettant de résoudre les problématiques de régression et de classification. La construction de CART passe par deux étapes :

1. La première étape permet de découper l'espace des variables explicatives de dimension  $p$  à un espace de dimension supérieure  $r$ . Ceci à partir d'un processus récursif qui commence par un seul nœud composé de  $n$  observations et finit par un ensemble de nœuds ou classes. Ces nœuds sont construits à chaque étape du processus en divisant l'ensemble de ces observations sur les deux nœuds suivants appelés fils gauche et fils droit. Ce découpage se fait de telle manière à réduire un critère d'impureté.
2. La seconde étape permet de faire une prédiction à partir des nœuds ou des classes finales, via l'attribution d'une étiquette à chaque classe finale. Dans le cas de la classification, l'étiquette de chaque classe se détermine à partir du principe de vote majoritaire. C'est-à-dire, nous attribuons à chaque classe la modalité la plus récurrente dans les observations composantes de cette classe.

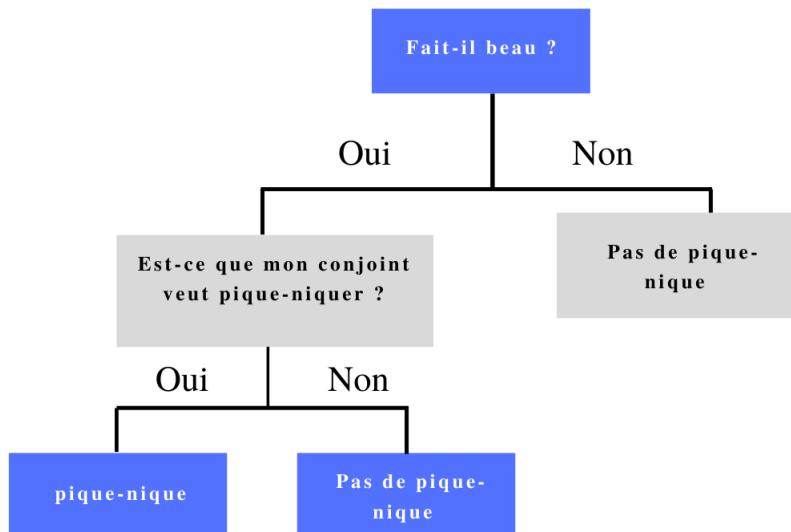


Figure 3.1 – Exemple d'un arbre de décision CART

La figure (3.1) donne un exemple d'arbre de décision de classification binaire (pique-nique et pas de pique-nique) à partir de deux variables catégorielles : "*Fait-il beau ?*" qui prend deux modalités (oui et non) et "*Est-ce que mon conjoint veut pique-niquer ?*" qui prend deux modalités (oui et non). La problématique principale d'un arbre CART est de déterminer l'architecture optimale qui permet de généraliser les résultats de prédiction. L'utilisation de CART seul pour la prédiction présente plusieurs

inconvénients, notamment le sur-apprentissage. Ceci peut être résolu par la méthode de l'élagage qui réduit le nombre de feuilles de l'arbre. Cette méthode nécessite une bonne stratégie de calibration ce qui n'est pas évident en pratique. L'arbre CART est très sensible aux données d'entraînement. Le rajout ou la suppression d'une variable peut changer complètement la structure de l'arbre. CART est aussi sensible aux valeurs aberrantes.

### Le modèle Xgboost :

L'*Xgboost* (eXtreme Gradient Boosting) est un modèle d'apprentissage non supervisé, il est utilisé pour les problématiques de régression et de classification. Il fait partie de la famille des algorithmes de boosting. Le boosting est une méthode qui construit d'une manière successive des prédicteurs faibles en termes de performance et les agrégent pour construire un seul prédicteur plus performant et plus puissant. L'idée est d'améliorer la qualité du prédicteur dans chaque étape en apprenant de l'erreur commise dans l'étape précédente, dans le but de réduire la variance et d'augmenter la stabilité du prédicteur. Le gradient boosting constitue un cas particulier du boosting où les erreurs sont minimisées à partir de l'algorithme de descente de gradient. L'*Xgboost* est une version optimisée du gradient boosting qui permet de contrôler le sur-apprentissage et de réduire le temps de calcul.

Soit :

- $D_n = (X_i, Y_i)_{i \in \{1, 2, \dots, n\}}$  un échantillon.
- Un ensemble d'arbres de décision CART  $\{f^m = w_{\eta(m)}, m \in \{1, 2, \dots, M\}\}$ , avec  $\eta(m) : \mathbb{R}^p \rightarrow T$ , la structure de l'arbre  $m$ ,  $T$  le nombre de feuilles,  $w \in \mathbb{R}^T$  le vecteur des poids de chacune de ces feuilles et  $M$  le nombre d'arbres ou le nombre d'itérations.
- $\lambda$  et  $\gamma$  les paramètres de régularisation.
- $l()$  fonction de perte qui est différentiable et convexe.

La fonction objective de l'optimisation de l'algorithme *Xgboost* peut s'écrire comme suit [Tia16] :

$$Obj(f, D_n) = \sum_{i \in D_n} l(\hat{Y}_i, Y_i) + \sum_{m=1}^M \Omega(f^m)$$

$$\hat{Y}_i = \phi(X_i) = \sum_{m=1}^M f^m(X_i)$$

$$\Omega(f^m) = \gamma T + \frac{1}{2} \lambda \|w_{\eta(m)}\|^2$$

$\Omega(f^m)$  est la fonction de régularisation, elle pénalise via le coefficient  $\gamma$  le nombre de feuilles de

l’arbre en question et  $\lambda$  pénalise les poids des feuilles à travers la norme mathématique  $L_2$ . À l’étape  $m$  du boosting la fonction objective peut se réécrire comme suit [Tia16] :

$$Obj^m(f, D_n) = \sum_{i \in D_n} l(\hat{Y}^{(m-1)} + f^m(X_i), Y_i) + \Omega(f^m)$$

La fonction objective est minimisée en utilisant l’algorithme de descente de gradient. Le gradient peut être approximé par le développement de Taylor au second ordre. Le détail de cette minimisation est expliqué par (Guestrin et Chen, 2016 ) [Tia16].

### 3.1.3 Métriques d’évaluation des modèles

Pour mesurer la capacité de généralisation sur de nouvelles données et comparer la qualité prédictive des deux modèles Xgboost et GLM, nous introduisons deux métriques d’évaluation.

Le taux de bonne classification (ACC) :

L’ACC est calculée à partir des labels observés  $(y_1, y_2, \dots, y_n)$  et des labels prédits par le modèle  $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ , via une matrice qui permet de compter les labels bien prédits et les labels mal prédits. Cette matrice est appelée matrice de confusion, elle est donnée comme suit :

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	Nombre de vrais positifs (VP)	Nombre de faux positifs (FP)
$y = 0$	Nombre de faux négatifs (FN)	Nombre de vrais négatifs (VN)

Table 3.1 – Matrice de confusion

L’ACC est égale au taux de bonnes prédictions, elle est calculée comme suit :

$$ACC = \frac{VP + VN}{n}$$

Plus le modèle de classification est performant, plus l’ACC est proche de 1. Elle peut être interprétée à travers la table (3.2) [Rak17].

L’ACC s’intéresse seulement aux labels prédits  $\hat{y}$  et non à la probabilité de survenance de ce label  $\mathbb{P}[y = 1|X]$ . La définition des labels prédits nécessite la définition d’un seuil de prédiction unique  $s : \hat{y} = \mathbb{1}_{\{\mathbb{P}[y=1|X] \geq s\}}$ . Pour un prédicteur de Bayes, ce seuil est égal à 1/2. Dans la présente étude, nous nous intéressons à la probabilité de survenance de  $y = 1$  qui représente la probabilité de

transformation, ainsi la définition d'un seuil de prédiction n'est pas indispensable. Cette métrique doit être utilisée avec précaution, notamment dans le cas de présence de données déséquilibrées<sup>2</sup>.

	Interprétation
$0.8 \leq ACC < 1$	Classification exceptionnelle
$0.6 \leq ACC < 0.8$	Classification excellente
$0.4 \leq ACC < 0.6$	Classification acceptable
$0.2 \leq ACC < 0.4$	Classification médiocre
$0 \leq ACC < 0.2$	Échec de classification

Table 3.2 – Interprétation ACC

#### La courbe ROC et l'AUC :

La courbe ROC est un outil qui permet de comparer les deux modèles et qui ne dépend pas d'un seul seuil de prédiction. Il s'agit d'une représentation graphique calculée à partir du couple  $(TVN_s, TVP_s)_{s \in [0,1]}$ , avec :

- $TVP_s = \frac{VP_s}{VP_s + FP_s}$  représente le taux de vrais positifs calculé à travers la matrice de confusion de  $y$  et  $\hat{y} = \mathbb{1}_{\{\mathbb{P}[y=1|X] \geq s\}}$ .
- $TVN_s = \frac{VN_s}{VN_s + FN_s}$  représente le taux de vrais négatifs calculé via la même matrice de confusion de  $TVP_s$ .

	Interprétation
$0.9 \leq AUC < 1$	Classification exceptionnelle
$0.8 \leq AUC < 0.9$	Classification excellente
$0.7 \leq AUC < 0.8$	Classification acceptable
$0.6 \leq AUC < 0.7$	Classification médiocre
$0.5 \leq AUC < 0.6$	Échec de classification

Table 3.3 – Interprétation AUC

La visualisation de la courbe ROC pour la comparaison des modèles reste limitée, notamment dans le cas d'intersection des courbes ou dans le cas de courbes très proches. Pour remédier à ce problème, une métrique supplémentaire sera utilisée qui est l'AUC, cette métrique représente l'aire sous la courbe

---

2. On parle des données déséquilibrées lorsque la proportion de 0 n'est pas égale à la proportion de 1 dans la variable d'intérêt  $y$ .

ROC. Ainsi, plus l'AUC est grande, plus le modèle est performant. Elle peut être interprétée à partir de la table (3.3) [Rak17].

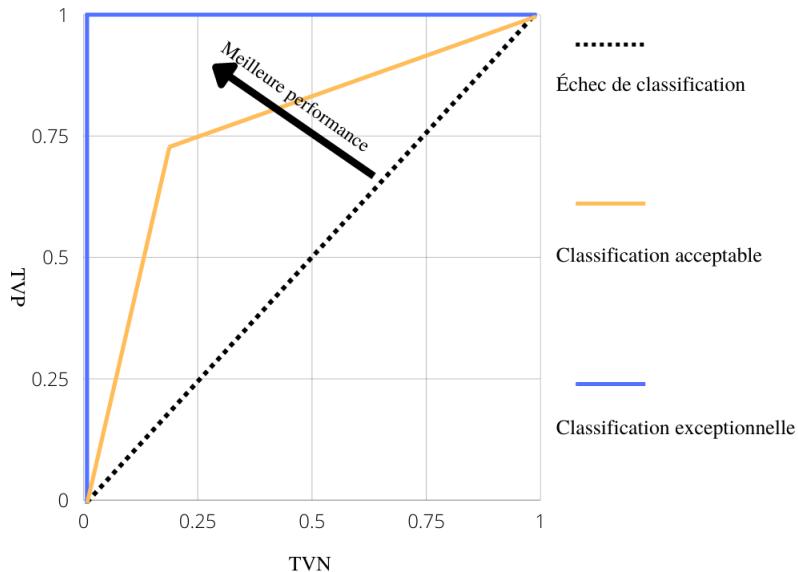


Figure 3.2 – Courbe ROC

### 3.1.4 Sélection de variables à travers l'algorithme RFE

La méthode de l'élimination récursive des variables, abrégée RFE en anglais, est un algorithme de sélection en arrière de variables selon une métrique définie en amont. Elle tire ses racines de l'article de (Guyon et al., 2002 ) [V V02]. Il s'agit de choisir les variables les plus pertinentes pour mieux prédire la variable à expliquer. L'algorithme commence par le calcul de l'importance de chacune des variables dans le modèle initial (Figure 3.3). Ensuite, il classe les variables selon l'ordre d'influence et en fonction de la métrique d'évaluation du modèle et il supprime la variable la moins importante. L'algorithme refait les 2 étapes jusqu'à l'élimination de toutes les variables ou jusqu'à ce que l'amélioration de la métrique d'étude ne soit plus significative.

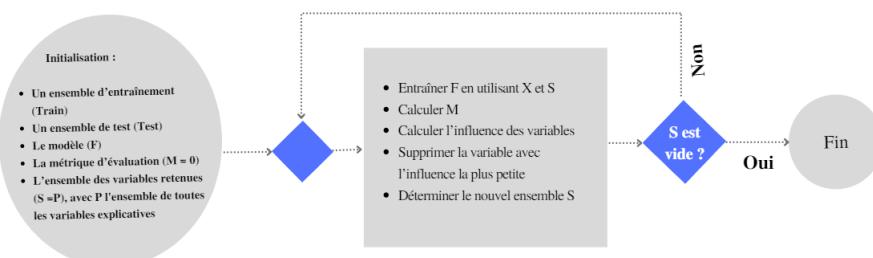


Figure 3.3 – L'algorithme RFE

L'avantage de l'utilisation de cette méthode pour notre étude est qu'elle puisse être implémentée avec différents modèles, notamment la régression logistique et le modèle Xgboost. Elle est basée sur la comparaison de l'influence des variables. Elle permet de suivre la métrique d'étude au fur et à mesure de l'élimination des variables, ce qui permet de réduire la complexité du modèle en choisissant la combinaison optimale des variables.

### 3.1.5 Mesure de l'influence des variables

La contribution de chacune des variables explicatives à la construction du modèle GLM se fait à travers les coefficients de régression. La structure linéaire du modèle GLM permet de récupérer ces coefficients et les classer selon l'ordre d'importance. Dans le cas du modèle Xgboost, la récupération de la contribution des variables n'est pas évidente, il n'existe pas de formule linéaire qui indique la contribution de chacune des variables. Le modèle Xgboost généralement est un modèle de bonne qualité prédictive, mais sa complexité le rend inexploitable pour des problématiques concrètes, notamment dans le domaine de l'assurance où le régulateur exige la transparence des modèles et des outils utilisés pour la mesure du risque.

#### Graphique de dépendance partielle (PDP) :

L'analyse PDP (Partial Dependence Plot) représente la plus ancienne méthode utilisée pour interpréter les modèles complexes tel que le modèle Xgboost. Elle a été introduite pour la première fois par (Friedman, 2001) [Fri01]. Il s'agit d'une méthode graphique qui permet de visualiser l'effet marginal d'une variable sur le modèle.

Nous nous plaçons dans le cadre d'apprentissage automatique supervisé présenté précédemment. Nous considérons un prédicteur  $\hat{f}$  et un échantillon  $D_n = (Z^{(i)})_{i \in \{1, 2, \dots, n\}}$  découpé en deux éléments :  $x^{(i)} \in \mathbb{R}^P$ ; les variables d'entrée de la  $i^{\text{ème}}$  observation et  $y^{(i)} \in \mathbb{R}$ ; le label. La variable Z est supposée suivre une loi théorique  $\mathbb{P}$  et les variables d'entrée  $X$  peuvent être découpées en deux :  $X_s$ ; l'ensemble des variables dont on souhaite mesurer l'effet marginal et  $X_C$ ; les autres variables. La fonction de dépendance partielle de  $X_s$  peut être définie comme suit [Ali20] :

$$\hat{f}_{x_s}(x_s) = \mathbb{E}_{X_c}(\hat{f}(X_c, x_s)) = \int \hat{f}(x_c, x_s) d\mathbb{P}_{X_c}(x_c)$$

L'estimation de  $\hat{f}_{x_s}$  est obtenue via l'échantillon  $D_n$  et la méthode de Monte Carlo [Ali20] :

$$\hat{f}_{x_s}(x_s) \simeq \frac{1}{n} \sum_{i=1}^n \hat{f}(x_c^{(i)}, x_s)$$

L'algorithme PDP permet de calculer la dépendance de la réponse  $Y$  aux variables d'entrée  $X_s$  en marginalisant les autres variables d'entrée  $X_c$ . La figure (3.4) ci-dessous donne un exemple simplifié du calcul de la dépendance partielle en utilisant le PDP [Ali20].

x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	f <sub>S</sub> <sup>(i)</sup> (x <sub>1</sub> )
1	4	5	f <sub>S</sub> <sup>(i)</sup> (1)
1	6	7	f <sub>S</sub> <sup>(i)</sup> (1)
x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	f <sub>S</sub> (x <sub>1</sub> )
1	4	5	1/n * sum(f <sub>S</sub> <sup>(i)</sup> (1))
2	6	7	1/n * sum(f <sub>S</sub> <sup>(i)</sup> (2))

Figure 3.4 – Exemple simplifié du calcul du PDP, source : [Ali20]

L'avantage de cette méthode est qu'elle est simple à implémenter, permet de visualiser l'effet marginal d'une variable sur la réponse pour les modèles complexes. En revanche, cette méthode seule ne permet pas d'interpréter un modèle complexe parce qu'elle est établie sur une hypothèse forte de non-corrélation de  $X_S$  et  $X_C$ . Cette hypothèse n'est pas toujours vérifiée en réalité.

### Théorie des Shap values :

Lundberg et Lee (2017) dans leur article intitulé "*A Unified Approach to interpreting Model Predictions*" ont fait la remarque que les modèles complexes sont inexploitables pour résoudre des problématiques concrètes. L'objectif de cet article est de proposer une méthode pour interpréter les modèles complexes, appelés modèles de boîte noire, il s'agit de la méthode SHAP (SHapley Additive exPlanations). SHAP tire ses origines de la mesure shapley de la théorie de jeux et des méthodes de contribution additive des variables [Su-17].

La valeur de Shapley est apparue pour la première fois avec Lloyd. S (1952) dans son l'article "*A Value for N-person Games*" [Sha52]. La valeur Shapley représente la manière la plus **juste** pour répartir le payoff  $V$  d'un jeu gagné par  $N$  joueurs en collaborant. Dans un tel jeu collaboratif, les joueurs ne contribuent pas de la même manière à la victoire. La répartition du payoff la plus juste est obtenue à travers les payoffs  $V_S$  qui auraient été obtenus pour le même jeu suivant les différentes coalitions  $S$  formées à partir des  $N$  joueurs. Pour le calcul de la valeur de Shapley le lecteur peut se référer à (Lloyd, 1952) [Sha52].

L'analogie avec un modèle de prédiction vient d'une manière naturelle : pour une observation  $x$ , les

$N$  joueurs représentent les variables explicatives, le payoff  $V$  représente la réponse, le jeu représente la prédiction [ $f(x) = V$ ],  $S$  représente une combinaison des variables explicatives et  $V_S$  représente la dépendance partielle de  $V$  à la combinaison  $S$  [ $f(x_S) = V_S$ ]. La valeur du gain représentant de la contribution d'un joueur  $j$  à la victoire est l'importance de la contribution d'une variable  $j$  à la réponse  $V$  pour l'observation  $x$ . L'idée de la valeur de Shapley est d'observer que la contribution de  $j$  peut s'obtenir par le calcul de la différence de  $V$  dans le cas de la contribution et la non contribution de  $j$  à la prédiction de la réponse :

$$\Delta_j(x) = V_{\text{avec } j} - V_{\text{sans } j}$$

Ceci est valide sous l'hypothèse d'indépendance entre les variables explicatives. Cette condition n'est pas toujours vérifiée,  $j$  peut être corrélée à d'autres variables explicatives, retirer cette variable peut augmenter l'importance de la contribution des autres variables. Pour remédier à ce problème, la contribution de  $j$  à la prédiction de  $V$  est calculée en utilisant les différentes combinaisons  $S$  (les coalitions en théorie de jeux), avec  $S$  appartient à  $P \setminus \{X_j\}$  l'ensemble des combinaisons possibles des autres variables explicatives hors la variable  $j$ . La contribution marginale de  $j$  au payoff  $V_S$  est donnée alors comme suit :

$$\Delta_j^S(x) = V_{S \text{ avec } j} - V_S$$

La contribution finale ou la valeur de Shapley de  $j$  à la prédiction de  $x$  est donnée par :

$$\Phi_j(\Delta^x) = \sum_{S \subseteq P \setminus \{X_j\}} \frac{|S|!(N - |S| - 1)!}{N!} * \Delta_j^S(x)$$

Ainsi la contribution de  $j$  à la construction du prédicteur  $f$  est donnée à travers le calcul de la moyenne des valeurs de Shapley de la base d'entraînement :

$$I_j = \frac{1}{n} \sum_{i=1}^n |\Phi_j(\Delta^{x_i})|$$

Il est aussi possible de calculer l'effet des interactions entre les variables à travers les indices d'interaction de Shapley, mais nous ne nous attardons pas sur cette notion dans le cadre de cette étude. La valeur de Shapley vérifie 3 propriétés :

1. **Additivité** : au voisinage d'une observation  $x$ , la prédiction où la réponse peut être approximée par la somme des effets des variables comme suit :

$$f(x) = \Phi_0 + \sum_{i=1}^N \Phi_i(\Delta^x) z'_i$$

Avec  $\Phi_0$  la valeur de base du modèle et  $z' \in \{0, 1\}^N$  le vecteur simplifié du vecteur des observations  $x$ .  $z'_i$  prend 0 si  $x_i$  est manquante et 1 sinon.

2. **Variables nulles sans effet :** si la variable est manquante dans l'approximation locale de  $x$  alors, elle n'a pas d'effet sur l'approximation de  $x$  :

$$z'_i = 0 \Rightarrow \Phi_i(\Delta^x) = 0$$

3. **Cohérence :** si le prédicteur  $f$  change vers un autre modèle  $f'$  tel que l'effet  $\Phi_j^f(\Delta^x)$  d'une variable  $j$  est plus important dans le modèle  $f$  alors :  $\Phi_j^f(\Delta^x) \leq \Phi_j^{f'}(\Delta^x)$ .

Ces trois propriétés rendent la valeur de Shapley légitime pour interpréter le modèle  $f$ . En revanche, cette valeur présente de nombreux inconvénients, notamment la complexité et le temps de calcul. Par exemple, pour une variable  $j$ , la valeur de Shapley nécessite l'implémentation de  $2^{N-1}$  prédicteurs. La valeur de Shapley dépend des observations de la base d'entraînement, la suppression ou le rajout d'une nouvelle observation nécessite la refonte de tout le calcul. (Lundberg et al., 2017) [Su-17] dans leur article intitulé "*Consistent Individualized Feature Attribution for Tree Ensembles*" proposent un algorithme efficace pour le calcul de la valeur de Shapley pour les modèles d'arbres de décision tel que le modèle Xgboost implémenté dans le cadre de cette étude. nous ne nous attarderons pas sur cet algorithme, le lecteur peut se référer à cet article [Su-17]. L'implémentation de l'algorithme est disponible en open source sous Python via le package *shap*.

## 3.2 Application pratique

### 3.2.1 Préparation des données pour la modélisation

S'agissant du traitement des variables, nous transformons les variables qualitatives en des variables indicatrices par la méthode de la dichotomie, dans le but que ces variables soient correctement traitées dans les modèles. La méthode de dichotomie transpose les modalités de chaque variable catégorielle en colonne et leur assigne la valeur 1 lorsque la modalité apparaît, et 0 sinon. Pour éviter le problème lié aux corrélations parfaites, une modalité est supprimée. Pour pouvoir évaluer les performances de généralisation des modèles, nous séparons le jeu de données en deux parties : (i) un jeu de données d'entraînement (80 % du jeu de données initial) qui servira à estimer les deux modèles ; (ii) un jeu de données de test (20 % du jeu de données initial). Il est important de noter que les deux modèles sont implémentés et testés sur les mêmes échantillons d'entraînement et de test, ceci dans le but de les comparer d'une manière juste. Les données à notre disposition sont équilibrées, c'est-à-dire la

proportion des devis transformés est proche de la proportion des devis non réalisés. Ainsi, les données ne nécessitent pas de traitement particulier. Concernant le calcul de l'ACC, nous utilisons le seuil de prédiction de 1/2 ce qui correspond au prédicteur de Bayes.

	Nombre d'observations	Devis transformés %	Devis non réalisés %
Entraînement	52759	48	52
Test	13190	47.5	52.5

Table 3.4 – Découpage des données en jeu d’entraînement et jeu de test

### 3.2.2 Optimisation des modèles

#### Optimisation des hyperparamètres :

Un hyperparamètre est un paramètre réglable du modèle. L’optimisation des hyperparamètres est le processus qui détermine le paramétrage optimal du modèle performant. La recherche par grille (Grid Search) est un algorithme qui cherche le paramétrage optimal via le test des différentes combinaisons des hyperparamètres. À chaque test d’une combinaison des hyperparamètres, Grid Search calcule la métrique d’évaluation à partir d’une validation croisée. Après avoir parcouru toutes les combinaisons, Grid Search retourne la combinaison qui maximise la métrique d’évaluation.

La validation croisée permet d’entraîner et de tester l’algorithme sur différentes parties de l’échantillon d’apprentissage. Le but de l’utilisation de la validation croisée est de réduire l’erreur associée à un unique apprentissage sur l’échantillon d’intérêt. Pour calculer une métrique de performance du modèle (dans notre cas, l’ACC et l’AUC), nous appliquons la validation croisée sur l’échantillon d’intérêt et nous prenons la moyenne des métriques calculées à partir des différents échantillons d’apprentissage.

S’agissant du modèle GLM, les hyperparamètres concernent essentiellement le coefficient de pénalisation du Lasso  $\lambda$ . Avant de lancer la Grid Search, nous allons choisir l’intervalle de la pénalité  $\lambda$ . La pénalité ne doit être ni trop petite ni trop grande. En effet, lorsque la pénalité est petite, c’est-à-dire proche de zéro, nous obtenons les coefficients de la régression logistique sans pénalisation, cependant si la pénalité est trop grande nous obtenons, des coefficients nuls. Ainsi, la définition de l’intervalle de la pénalité  $\lambda$  est une étape importante dans le Lasso. Le graphe (3.5) du chemin de régularisation donne les différents scénarios possibles de la pénalité et le nombre de coefficients nuls dans le modèle.

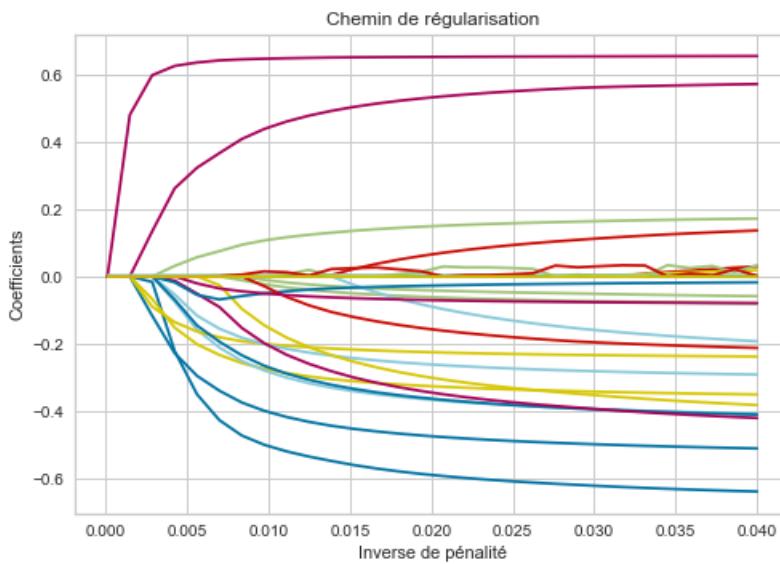


Figure 3.5 – Chemin de régularisation du Lasso

Pour  $1/\lambda = 0.001$ , un seul coefficient non nul ce qui correspond au cas de présence d'une seule variable explicative dans le modèle. Ainsi, le graphe donne une idée générale sur la variation de la pénalité et l'état de présence des variables dans le modèle. Pour trouver exactement la pénalité optimale qui minimise la fonction de perte, nous procédons à une Grid Search par validation croisée. Nous entraînons la validation croisée sur 5 sous-échantillons. Nous choisissons comme critère d'optimisation l'AUC et nous balayons  $1/\lambda$  entre 0.005 et 0.015. Nous obtenons ainsi un coefficient de pénalité  $1/\lambda = 0.01$ .

S'agissant du modèle Xgboost, les hyperparamètres à fixer pour la Grid Search sont donnés comme suit :

- **n\_estimators** : le nombre d'arbres ou le nombre d'itérations de l'algorithme du boosting. Ce paramètre représente la complexité du modèle, plus il est grand plus, le modèle Xgboost perd sa capacité de généralisation. Il prend les valeurs [80, 100, 200].
- **gamma** : le paramètre d'élagage de l'arbre CART prenant les valeurs [0, 0.1, 0.2].
- **subsample** : la part du jeu de données utilisée pour l'entraînement prenant les valeurs [0.7, 0.8, 0.9, 1].
- **max\_depth** : la profondeur maximale d'un arbre CART. Plus l'arbre est profond, plus le modèle est complexe. Il prend les valeurs [3, 4, 5].
- **learning\_rate** : le taux d'apprentissage de l'algorithme prenant les valeurs [0.1, 0.2, 0.3].

Nous lançons une Grid Search avec validation croisée, nous trouvons que le paramétrage optimal du modèle Xgboost est donné comme suit : **n\_estimators** ; 80, **gamma** ; 0.5, **subsample** ; 0.7, **max\_depth** ; 5, **learning\_rate** ; 0.2 .

Pour tester le sur-apprentissage du modèle Xgboost, nous construisons la courbe d'apprentissage. Cette courbe donne à chaque itération du boosting la métrique d'évaluation pour le jeu de test (test) et le jeu d'entraînement (train). Les figures (3.6) et (3.7) donnent la courbe d'apprentissage (Learning Curve) selon les métriques AUC et 1-ACC. Nous remarquons au niveau de l'itération 80, un écart de 2% pour l'AUC et un écart de 2,5% pour l'ACC. De ce fait, nous retenons le nombre d'itérations choisi par la Grid Search.

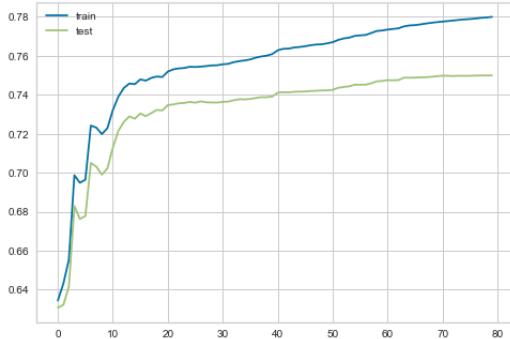


Figure 3.6 – Learning Curve (AUC)

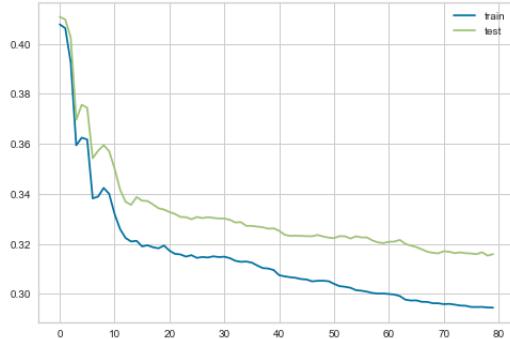


Figure 3.7 – Learning Curve (1-ACC)

### Sélection de variables :

Nous utilisons l'algorithme RFE pour éliminer les variables qui apportent une information redondante au modèle. Cet algorithme nécessite un temps de calcul important, la complexité du calcul augmente avec le nombre de variables explicatives et la taille de l'échantillon. Pour remédier à ce problème, nous réglons le nombre de variables à éliminer dans chaque étape à 3. Les figures (3.8) et (3.9) donnent le résultat de l'implémentation du RFE pour le modèle GLM et pour le modèle Xgboot. Ces figures mettent en relation le nombre de variables sélectionnées par le RFE et le niveau d'AUC. Nous remarquons que le nombre optimal de variables sélectionnées pour le GLM est 34. En revanche, l'AUC de 30 à 34 n'évolue pas d'une manière significative, ainsi pour la suite de l'étude, nous choisissons de retenir les 30 premières variables sélectionnées par le RFE. Concernant le modèle Xgboost, nous constatons après la sélection de 16 variables que la métrique AUC n'évolue plus d'une manière significative. Afin de réduire la complexité du modèle Xgboost, nous ne gardons que les 16 premières variables sélectionnées par le RFE.

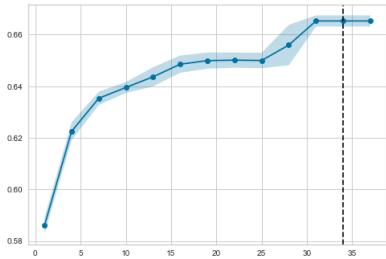


Figure 3.8 – Sélection de variables (GLM)

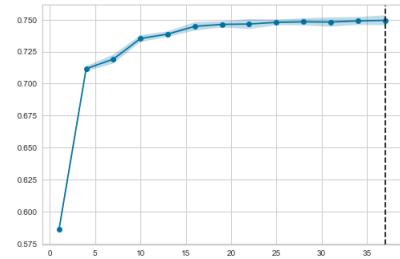


Figure 3.9 – Sélection de variables (Xgboost)

### 3.2.3 Résultats

#### 3.2.3.1 Évaluation des modèles

L’analyse des figures (3.10) et (3.11) révèle que les performances prédictives sur les données de test (Test) et d’entraînement (Train) du modèle Xgboost sont bonnes et meilleures que celles du modèle GLM. En revanche, en comparaison avec le modèle Xgboost, nous constatons que le GLM a des écarts très petits entre les résultats du test et les résultats de l’entraînement, ce qui signifie que le GLM est très stable par rapport au modèle Xgboost.

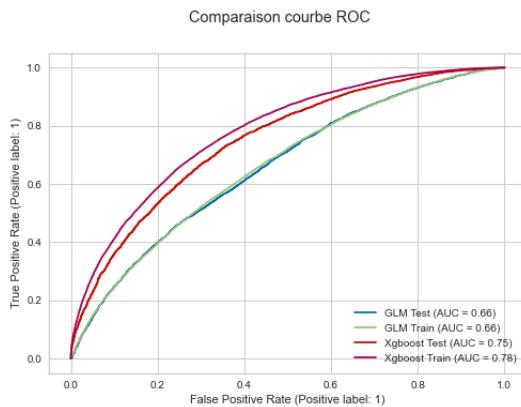


Figure 3.10 – Courbe ROC

	GLM	Xgboost
AUC (Train)	66.11%	78.14%
AUC (Test)	66.02%	75.87%
ACC (Train)	61.37%	69.79 %
ACC (Test)	61.25%	67.46%

Figure 3.11 – Évaluation des modèles

#### 3.2.3.2 Résultat du modèle GLM

En termes d’importance de variables, nous remarquons d’après la figure (3.12) que la prime, la modalité Parc cible et Agent sont parmi les 5 premières variables les plus importantes. La prime et le coefficient technique dépendent négativement de la probabilité de souscription. Ceci semble intuitif, car le tarif est un élément important dans le processus d’attraction des clients. Par ailleurs, les devis émis par les agents ont une probabilité de transformation supérieure. Toutefois, les catégories du Parc cible ont une probabilité de transformation inférieure au Parc non cible et plus la proportion de véhicules d’âge égal à 1 an est élevée dans le devis Parc, plus la probabilité de souscription est

grande. Ainsi, lorsque le client possède une proportion importante de véhicules récents, il a tendance plus à accepter facilement de souscrire le contrat.

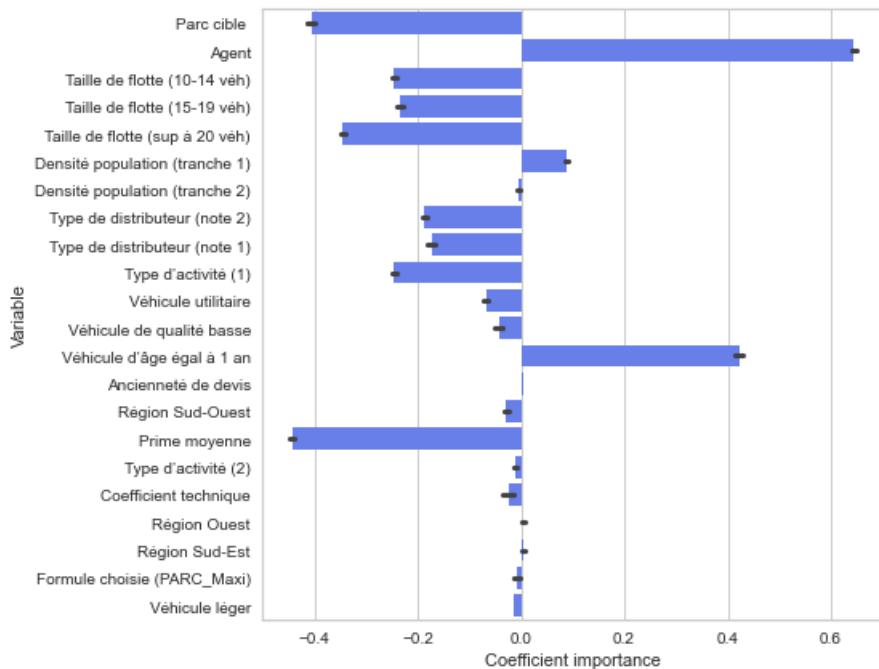


Figure 3.12 – Influence des variables du modèle GLM

Les graphiques de dépendance partielle (3.13) et (3.14) sont cohérents avec les résultats du GLM. La courbe verte représente la dépendance partielle du taux de transformation à la prime et au coefficient technique, les barres en bleues représentent l'exposition et la courbe horizontale représente le seuil 50% à partir duquel le devis est considéré comme un devis transformé. Nous pouvons remarquer que la pente de la courbe de la prime moyenne est plus importante que celle du coefficient technique et que la probabilité de transformation est décroissante en fonction de ces deux variables.

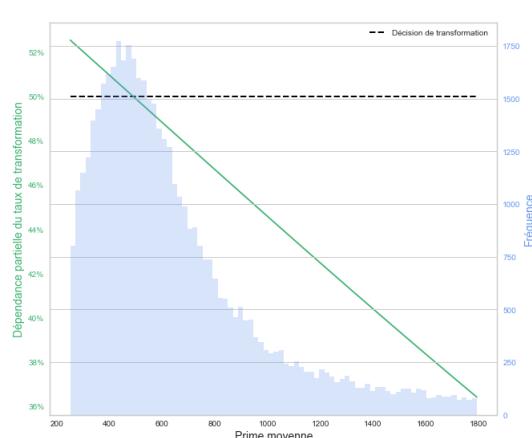


Figure 3.13 – PDP de la prime moyenne (GLM)

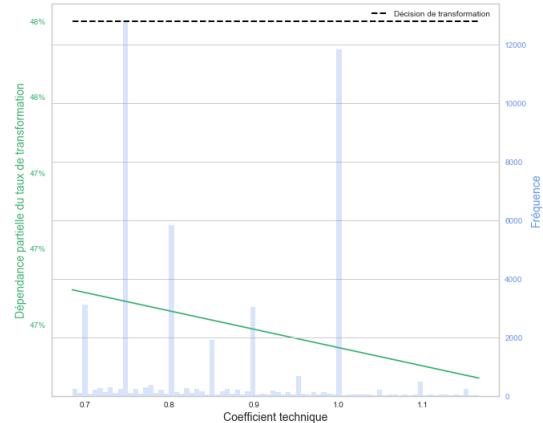


Figure 3.14 – PDP du coefficient technique (GLM)

### 3.2.3.3 Résultat du modèle Xgboost

La figure (3.15) donne la contribution des variables explicatives selon l'ordre d'importance dans la construction du modèle Xgboost. Ce graphe est construit via le calcul et la comparaison de la moyenne des Shap values des variables explicatives. Le classement des variables en termes d'importance est différent de celui du modèle GLM. Le coefficient technique et l'ancienneté de devis ont remonté dans le classement. En revanche, la prime moyenne et la modalité Agent ont perdu deux places par rapport au classement du GLM.

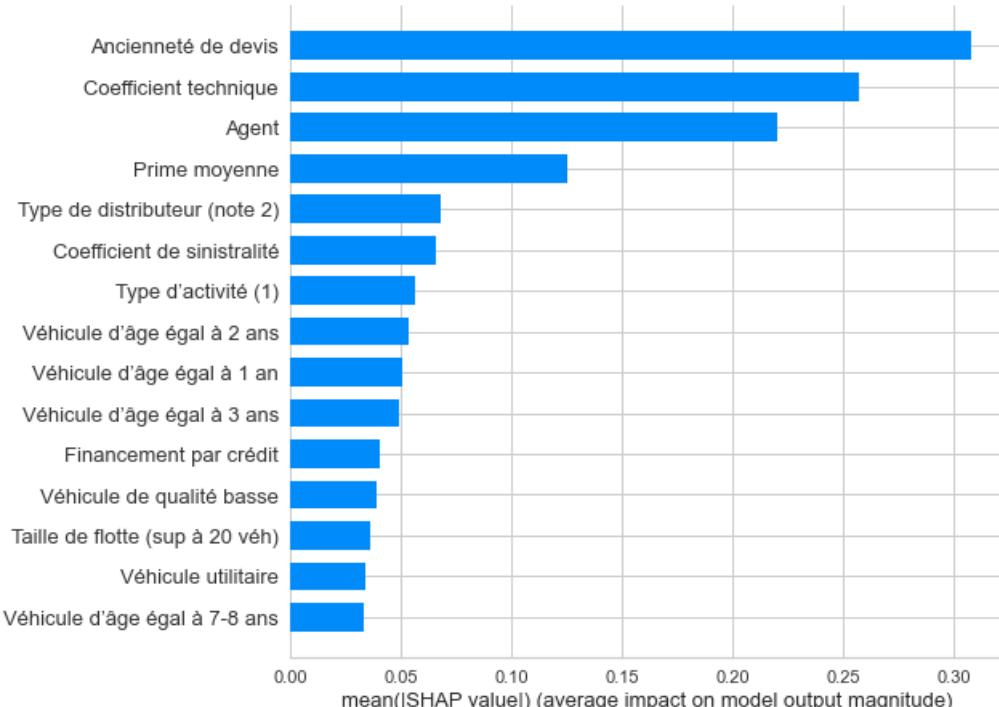


Figure 3.15 – Importance des variables du modèle Xgboost

Pour mesurer l'influence des variables explicatives, nous utilisons la figure (3.16). Cette figure donne le nuage des Shap values calculé pour chaque variable explicative. Le nuage rouge correspond aux valeurs élevées de la variable et le nuage bleu correspond aux valeurs faibles de la variable. En abscisse, nous trouvons les Shap values, une Shap value positive signifie que la variable contribue positivement à la probabilité de transformation et inversement. Nous constatons à travers la figure (3.16) que le nuage des Shap values de la prime est rouge pour les valeurs de Shap négatives. Il devient de plus en plus bleu pour les Shap values positives ce qui signifie que la probabilité de transformation est décroissante en fonction de la prime moyenne, le graphe du PDP (figure 3.17) confirme ce résultat. Toutefois, les devis émis par les agents ont une probabilité de transformation supérieure au devis émis par les courtiers et les devis émis par les distributeurs de note égale à 2 ont une probabilité de transformation inférieure au devis émis par les distributeurs de note égale à 3. Concernant l'ancienneté

de devis et le coefficient technique, le nuage de points des Shap values ne permet pas de détecter une tendance simple ce qui peut être confirmé par le graphe PDP (3.18) du coefficient technique. Ce graphe permet d'observer une tendance complexe de la dépendance partielle en fonction du coefficient technique.

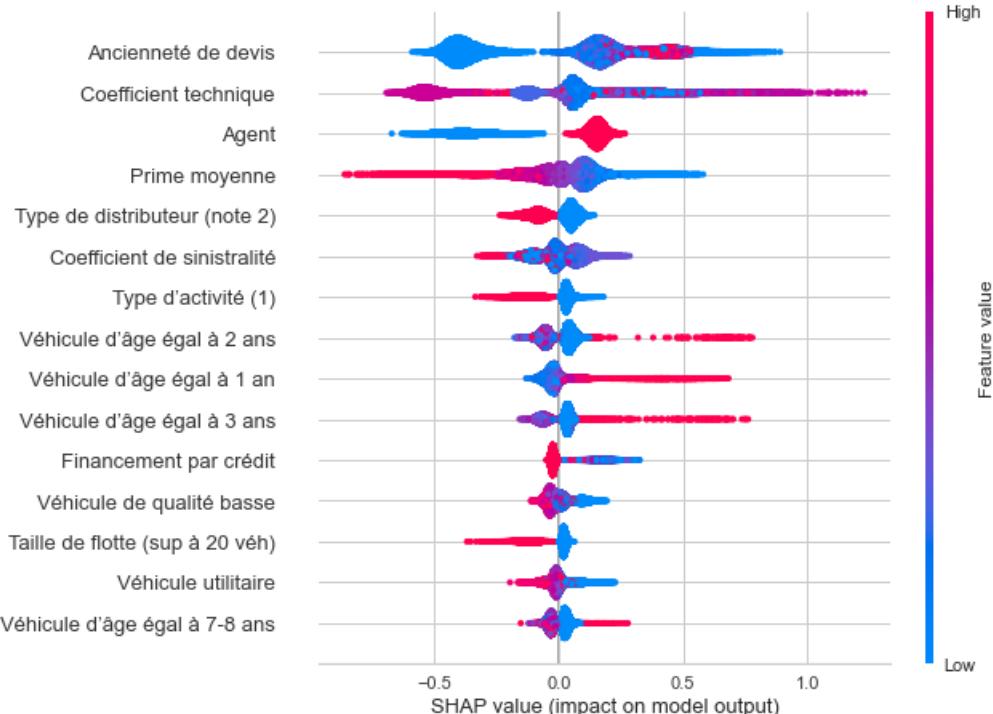


Figure 3.16 – Influence des variables du modèle Xgboost

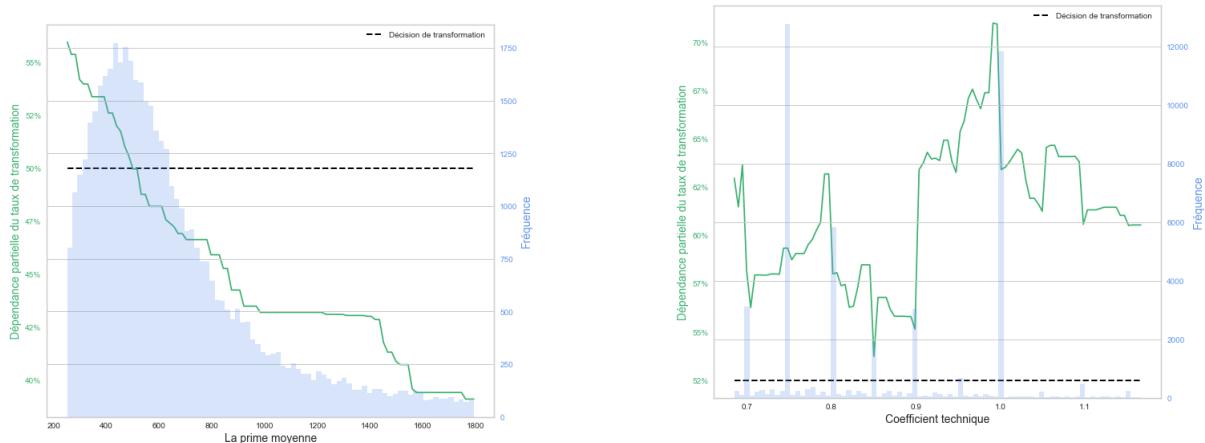


Figure 3.17 – PDP de la prime moyenne (Xgboost)

Figure 3.18 – PDP du coefficient technique (Xgboost)

### 3.2.3.4 Validation des modèles

Pour mesurer la capacité de généralisation des modèles, nous calculons les moyennes des prédictions du taux de transformation en fonction des variables explicatives pour le jeu de test. Les figures

(3.19) et (3.20) représentent les moyennes du taux de transformation prédites du modèle Xgboost, du GLM et la moyenne observée du taux de transformation. En comparaison avec le modèle GLM, nous remarquons que les variables prime moyenne et coefficient technique sont bien prédites par le modèle Xgboost. S’agissant de la prime moyenne, le modèle GLM a surestimé le taux de transformation pour les segments inférieurs à 600-800 €. Concernant le coefficient technique, nous remarquons que le modèle GLM a sous-estimé la probabilité de transformation pour les segments supérieurs à 0.85.

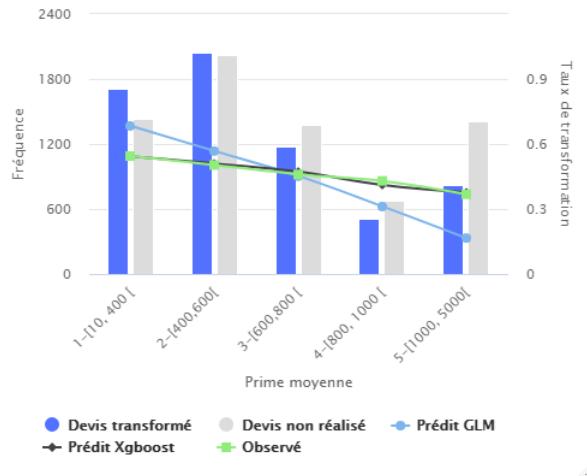


Figure 3.19 – Prédiction du taux de transformation selon la prime moyenne

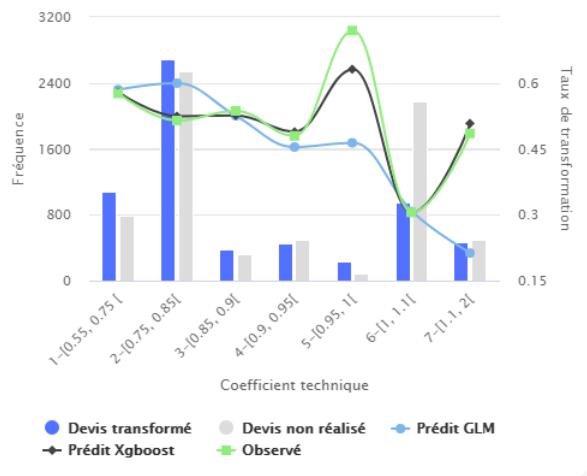


Figure 3.20 – Prédiction du taux de transformation selon le coefficient technique

S’agissant des variables distributeur et catégorie de Parc (cible ou non cible), nous remarquons que le modèle GLM a surestimé la probabilité de transformation pour les segments cibles et pour la modalité agent. En revanche, le modèle Xgboost a bien prédit les deux variables. Concernant la variable région, le modèle GLM a sous-estimé la probabilité de transformation pour la région IDF (Île-de-France) et surestimé la probabilité de transformation pour la région OUEST.

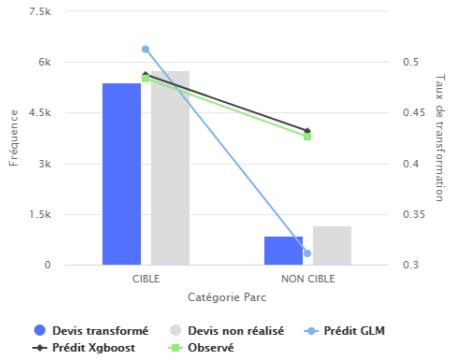


Figure 3.21 – Prédiction du taux de transformation selon la catégorie de Parc

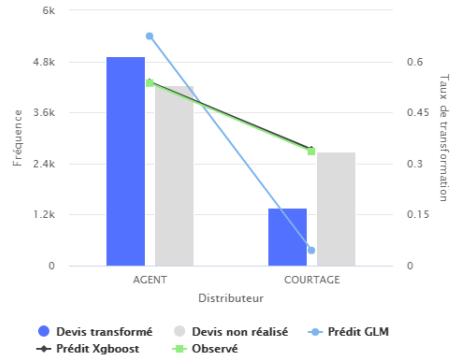


Figure 3.22 – Prédiction du taux de transformation selon le réseau de distribution

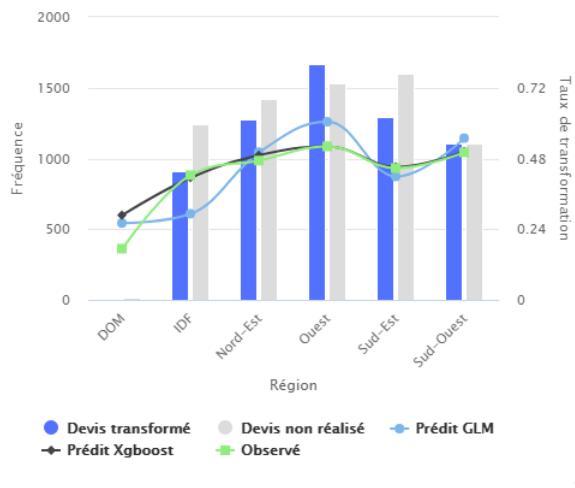


Figure 3.23 – Prédiction du taux de transformation par région

## Conclusion partielle

Les modèles implémentés au sein de ce chapitre permettent de caractériser les clients qui transforment plus les devis du produit Parc dénommé. À l’appui des résultats obtenus, nous savons quels sont les segments des clients susceptibles d’accepter de souscrire le contrat du produit Parc. Les devis du réseau des agents ont une probabilité de transformation supérieure au réseau des courtiers. Les deux modèles donnent que la probabilité de transformation est décroissante en fonction de la prime. Par ailleurs, il est possible de déterminer à priori la probabilité de transformation pour chaque devis à travers l’un des deux modèles. En comparaison avec le modèle Xgboost, le modèle GLM a fait preuve de faible capacité de généralisation des résultats de prédiction. En revanche, le modèle Xgboost malgré sa bonne performance prédictive présente une limite non négligeable due à sa complexité le rendant non opérationnel pour la prédiction du taux de transformation.

# Chapitre 4

## Modélisation de l'élasticité au prix

### Préambule

L'objectif de ce chapitre est de proposer une méthodologie pour modéliser l'élasticité au prix du produit Parc dénommé. Cette modélisation nous permet de répondre à la question suivante : comment changerait le taux de transformation (la demande) si l'on propose à un client une prime  $i$  au lieu d'une prime  $j$  toutes choses égales par ailleurs ? Pour répondre à cette problématique, nous proposons dans ce chapitre une méthodologie inspirée de la littérature de l'inférence causale. Nous considérons le taux de rabais de la prime comme un traitement multidose et nous utilisons le score de propension pour équilibrer les différents groupes du traitement. Tout d'abord, nous modélisons le score de propension avec le modèle Xgboost, un modèle qui permet de capter les associations linéaires et non linéaires existant entre le taux de rabais et les variables explicatives (facteurs de confusion). Ensuite, nous pondérons chaque individu par l'inverse de son score de propension. Nous évaluons aussi l'équilibre des facteurs de confusion entre les classes du taux de rabais avec l'ASMD (Absolute Standardized Mean Difference). Enfin, à partir de la population des groupes équilibrés, nous modélisons le taux de transformation à l'aide d'un GLM doublement robuste appelé modèle global, et nous calculons la fonction élasticité.

### 4.1 L'élasticité au prix

L'étude de l'élasticité au prix au niveau individuel nécessite la définition d'une fonction de demande. La fonction de demande en théorie de la consommation<sup>1</sup> représente la relation entre le prix et la quantité optimale du bien demandé par le consommateur qui satisfait ses préférences et ses contraintes

---

1. La théorie de la consommation est une théorie de la microéconomie néo-classique visant à connaître les comportements des consommateurs afin de satisfaire leurs besoins.

budgétaires. Pour le cas des biens où le consommateur a le droit à un seul bien, la demande au niveau individuel devient binaire et prend 1 si le consommateur accepte d'acheter le bien et 0 sinon. Ainsi, la fonction de demande devient la probabilité d'accepter d'acheter le bien. L'assurance représente un cas particulier de ce bien et le taux de transformation représente la fonction de demande de l'assurance au niveau individuel. L'élasticité au prix permet de mesurer le comportement de l'assuré face à un changement tarifaire futur, en d'autres termes, elle permet de répondre à la question : comment changerait le taux de transformation (la demande) si l'on propose à un client une prime  $t$  au lieu d'une prime  $t'$  toutes choses égales par ailleurs ? Notons  $\hat{f}$  le taux de transformation estimé d'un individu  $i$  et  $P_i$  la prime proposée, alors l'élasticité au prix peut être définie comme suit :

$$e(P_i) = -\frac{\partial \hat{f}(P_i)}{\partial P_i} \frac{P_i}{\hat{f}(P_i)}$$

$e(P_i)$  représente la variation relative du taux de transformation par rapport à la variation relative de la prime, une  $e(P_i) = 4$  signifie qu'une augmentation de 1% de la prime baissa la probabilité de transformation de l'individu  $i$  de 4%. Le taux de transformation est décroissant en fonction de la prime ce qui peut être confirmé par le nuage de la valeur de Shap du modèle Xgboost, il est nul lorsque la prime est infinie et égal à 1 lorsque la prime est nulle. De ce fait, on s'attend à une élasticité positive.

Pour le produit Parc dénommé, la mise à jour de la prime pure nécessite un investissement important. Par conséquent, elle est supposée constante dans la présente étude. Les ajustements tarifaires annuels pour les affaires nouvelles se font à travers deux coefficients : le coefficient technique et le coefficient de sinistralité pour la garantie responsabilité civile. Cette étude se concentre seulement sur l'étude de la sensibilité de la probabilité de transformation suite à la variation du coefficient technique. Notons  $COPHT$  la prime proposée à l'assuré,  $P$  la prime avant l'application du coefficient technique, c'est-à-dire avant le rabais, elle est supposée fixe et  $CT$  le coefficient technique variable, alors nous pouvons remarquer la relation suivante :

$$\frac{\partial CT}{CT} = \frac{P * \partial CT}{CT * P} = \frac{\partial(CT * P)}{CT * P} = \frac{\partial COPHT}{COPHT}$$

L'élasticité au prix peut se réécrire sous la forme suivante :

$$e(COPHT) = e(CT) = -\frac{\partial \hat{f}(CT)}{\partial CT} \frac{CT}{\hat{f}(CT)}$$

De ce fait, étudier l'élasticité au prix au niveau individuel revient à étudier l'élasticité de la probabilité de transformation par rapport au coefficient de rabais. La détermination de  $e(COPHT)$  consiste

à calculer la dérivée :

$$\frac{\partial \hat{f}(CT)}{\partial CT}$$

Le calcul de cette dérivée est compliqué en réalité à cause des associations qui existent entre le taux de rabais et les autres caractéristiques du Parc, notamment la prime, le réseau de distribution et le type du Parc, les segments cibles pour AXA France dépendent de rabais importants par rapport aux segments non cibles. De ce fait, la dérivée peut ne pas exister et même si elle existe, elle peut donner des formes très complexes. Les sections suivantes présenteront la méthodologie de la pondération sur le score de propension utilisée pour détecter ces associations, les éliminer et calculer l'élasticité au prix. Dans la suite de l'étude, nous appelons le coefficient technique centré autour de 0 par le taux de rabais :

$$\text{taux de rabais} = \text{coefficient technique} - 1$$

## 4.2 Cadre théorique

Cette section vise à présenter le cadre théorique de la méthodologie utilisée. Elle commencera par la définition du biais de confusion. Ensuite, elle présentera l'effet causal du traitement et la méthode de la pondération sur le score de propension. Enfin, elle exposera la modélisation doublement robuste.

### 4.2.1 Biais de confusion

La difficulté à mesurer l'élasticité au prix au niveau individuel réside dans la nature des données à notre disposition. En effet, les bases de données en assurance sont des données observées établies à partir de l'expérience de l'assureur. Ces données ne permettent pas de déterminer l'effet causal (le vrai effet) de la variation du taux de transformation par rapport à la variation du taux de rabais.

« *Quand on est malade, il ne faut surtout pas aller à l'hôpital : la probabilité de mourir dans un lit d'hôpital est 10 fois plus grande que dans son lit à la maison* »<sup>2</sup>

Cette proposition présente un exemple classique de la différence entre la causalité et la corrélation, il existe certes une corrélation entre le fait de dormir à l'hôpital et l'évènement de mourir. Ceci ne signifie pas que dormir à l'hôpital est dangereux. Il existe une variable omise qui est l'état de santé qui met un biais de confusion sur la détection de la vraie relation entre mourir et dormir à l'hôpital. Par analogie, les corrélations entre le taux de rabais et certaines variables explicatives apportent une

2. <https://www.lafinancepourtous.com/juniors/lyceens/l-instant-maths/correlation-nest-pas-causalite/>

certaine confusion sur l'estimation de l'effet de la variation du coefficient de rabais sur le taux de transformation.

En statistique, nous parlons du biais de confusion (confounding bias) [Kle10] lorsqu'un facteur tiers met une confusion sur l'estimation de l'effet de la variable d'exposition sur la variable d'intérêt. Pour qu'un facteur entraîne un biais de confusion, il doit vérifier trois conditions :

- L'existence de l'association entre l'exposition de l'étude et le facteur tiers.
- L'existence de l'association entre la variable d'intérêt et le facteur tiers.
- Il existe un lien direct entre la variable d'exposition et la variable d'intérêt. C'est-à-dire, le facteur tiers ne se situe pas sur le chemin causal de la variable d'exposition sur la variable d'intérêt.

Par exemple, dans le cadre de notre étude, les caractéristiques de la flotte représentent des facteurs tiers. Elles sont corrélées simultanément avec le taux de rabais et le taux de transformation et elles ne se situent pas sur le chemin causal du taux de rabais sur le taux de transformation.

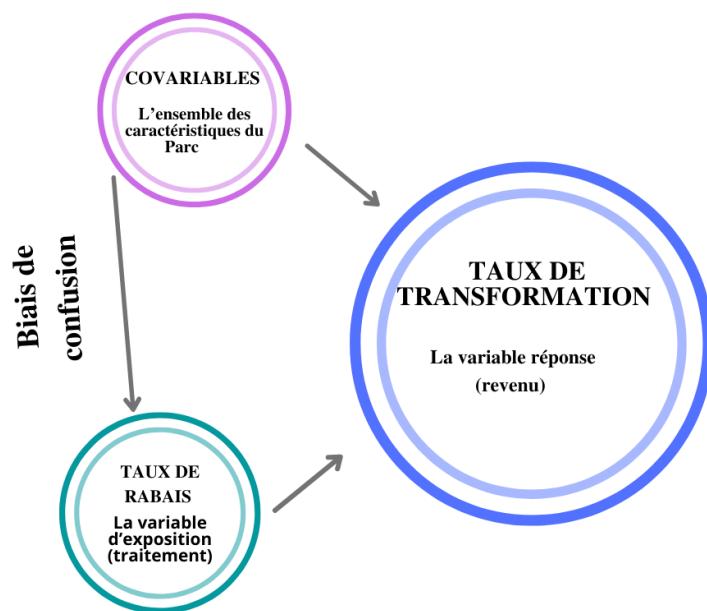


Figure 4.1 – Biais de confusion

#### 4.2.2 Effet causal du traitement multidose

Afin d'éliminer le biais de confusion, nous élaborons la problématique de l'élasticité au prix dans le cadre du modèle de Rubin [RUB74] et de l'inférence causale [Lan13] [Imb00] [D B83] [Li18]. Le lexique de ce modèle est tiré de la recherche médicale. Rubin désigne par **traitement** la variable dont on souhaite mesurer l'effet et par **revenu** la variable d'intérêt. Ainsi, pour la suite de l'étude, nous

utilisons d'une manière interchangeable les termes : traitement et taux de rabais ; revenu et taux de transformation.

Considérons une base de données d'assurance composée de  $N$  devis. Notons  $T$  la variable du traitement (le taux de rabais), elle est composée de  $M$  doses (tranches du taux de rabais). Pour chaque devis  $i \in \{1, 2, \dots, N\}$  et chaque dose  $t \in \{1, 2, \dots, M\}$ ,  $X$  l'ensemble des covariables (facteurs de confusion) et  $K$  le nombre de facteurs de confusion. Nous désignons par  $Y_i[t]$  le revenu potentiel de l'individu  $i$  s'il avait reçu la dose ou le traitement  $t$ . La question principale est de déterminer la variation nette (gain net ou perte nette) du revenu si l'individu  $i$  était exposé au traitement  $t$  au lieu du traitement  $t'$  toutes choses égales par ailleurs (*ceteris paribus*). Sous la même notation, cette variation est donnée par :

$$\Delta_i(t, t') = Y_i[t] - Y_i[t']$$

$\Delta_i(t, t')$  s'appelle l'effet du traitement  $t$  relativement à  $t'$  de l'individu  $i$ . Par ailleurs, nous nous intéressons plus à un abrégé de l'effet individuel du traitement pour toute la population étudiée ou seulement pour une sous-population (un segment de clients). Nous introduisons ainsi l'**ATE** (Average Treatment Effect).  $ATE(t, t')$ , ou l'effet moyen du traitement  $t$  relativement au traitement  $t'$ , représente la différence entre la moyenne du revenu si toute la population avait reçu le traitement  $t$  et la moyenne du revenu si elle avait reçu le traitement  $t'$ . Un traitement de  $M$  doses nécessite l'estimation de  $M(M - 1)/2$  ATE. L'ATE du traitement  $t$  relativement au traitement  $t'$  est calculé comme suit :

$$E(\Delta[t, t']) = E(Y[t] - Y[t']) = E(Y[t]) - E(Y[t']) = \mu_t - \mu_{t'}$$

Le problème fondamental de l'inférence causale [Hol86] est que nous ne pouvons pas observer  $Y_i[t]$  et  $Y_i[t']$  simultanément pour un individu  $i$ . Par exemple, il est difficile de savoir si un client aurait accepté de souscrire au contrat Parc dénommé ou non si on lui avait proposé un autre taux de rabais que celui déjà renseigné dans la base. De ce fait, le problème de l'inférence causale est en quelque sorte un problème de données manquantes puisque nous n'observons pas les revenus sous toutes les doses du traitement [Mon13]. Par ailleurs, la mesure de l' $ATE(t, t')$  nécessite un certain équilibre entre les groupes  $t$  et  $t'$ , c'est-à-dire des groupes identiques ou avec des caractéristiques très proches [RUB74]. Ceci n'est réalisable qu'avec les essais randomisés contrôlés. D'ailleurs, cette méthode est la seule qui permet d'assurer l'équilibre entre les groupes du traitement [Mar10], puisqu'elle est fondée sur un tirage au sort avec remise où chaque individu de la population a une probabilité de  $1/M$  d'appartenir à un des groupes.

### 4.2.3 Méthode de la pondération sur le score de propension

Pour le produit Parc dénommé, il est difficile en pratique, voire impossible de faire des essais randomisés contrôlés pour la mesure de l'élasticité au prix (appelé en actuariat **Price Test**). Dans la présente étude, nous nous basons seulement sur les données observées à notre disposition. L'étude de l'élasticité au prix consiste à modéliser le taux de transformation en utilisant le taux de rabais et l'ensemble des variables explicatives à disposition de l'assureur à l'aide d'un modèle GLM. Dans le but de pouvoir calculer l'élasticité au prix relative pour chaque individu en utilisant le coefficient associé au taux de rabais. Cette approche conduit à des résultats non robustes à cause du biais de confusion présenté dans la sous-section précédente. En effet, le coefficient du taux de rabais issu de cette modélisation ne traduit pas d'une manière exacte l'effet causal, c'est-à-dire la variation nette du taux de transformation par rapport à la variation du taux de rabais. Ceci est à cause des corrélations linéaires et non linéaires qui chevauchent l'estimation et biaise ce coefficient. En revanche, le biais de confusion n'affecte en aucun cas la qualité prédictive du modèle GLM. Toutefois, il est possible d'optimiser le modèle GLM en rajoutant les interactions entre le taux de rabais et les autres variables explicatives et en rajoutant d'autres transformations sur le taux de rabais. Cette optimisation reste limitée dans la mesure où il est difficile de capter tout type d'interaction et d'association entre les variables.

Les modèles d'apprentissage automatique constituent une alternative au modèle GLM. En effet, la particularité de ces modèles est qu'ils permettent de rallier toutes les associations possibles, y compris les associations non linéaires entre les variables explicatives. Le seul inconvénient de ces modèles est la complexité et le grand espace de paramètres à estimer. Le calcul analytique de l'élasticité au prix dans ce cas est très compliqué, voire impossible.

L'actuaire a besoin de modèles parcimonieux et opérationnels pour qu'il puisse les présenter avec un degré de vulgarisation suffisamment adapté à d'autres disciplines et à d'autres parties prenantes non spécialisées de la matière. De ce fait, nous proposons une méthode basée à la fois sur un modèle d'apprentissage automatique (**Xgboost**) et sur un modèle linéaire opérationnel (**GLM**). Notre approche est inspirée de la littérature sur l'inférence causale, un champ de recherche de plusieurs disciplines : médecine, épidémiologie, science sociale et statistique. Elle tire ses racines des travaux de (Gelman et Guillén, 2013) [Mon13] sur l'étude l'élasticité au prix en assurance automobile et de (McCaffrey et al., 2013) [Lan13] et (Imbens, 2000) [Imb00] sur l'étude de l'effet du traitement multidose avec la pondération sur le score de propension.

Le problème principal est l'existence des facteurs de confusion (les caractéristiques du Parc), qui influencent autant le taux de rabais que le taux de transformation et chevauchent l'estimation de l'effet

du traitement. La pondération sur le score de propension **IPTW (Inverse Probability of Treatment Weighting)**[Lan13] que nous allons présenter par la suite est l'une des méthodes qui permettent d'éliminer le biais de confusion. Cette méthode est fondée sur deux principales hypothèses :

- *Hypothèse 1 : Indépendance conditionnelle à des caractéristiques observables (conditional independence assumption – CIA)*

$$(Y_i[t_1], Y_i[t_2], \dots, Y_i[t_M]) \perp T_i | X_i \quad \text{pour } i \in \{1, 2, \dots, N\}$$

Cette hypothèse signifie que le taux de rabais ne dépend que des variables observées  $X$  et non pas des autres variables non observées. En inférence causale, cette hypothèse est non testable, dépend de l'expertise métier et de la compréhension du phénomène étudié [Mon13]. Imbens montre que cette condition est suffisante pour avoir une estimation consistante de  $\mu_t$  [Imb00]. Dans le cadre de notre étude, cette condition est vérifiée. En effet, toutes les variables qui impactent le taux de rabais (les facteurs de confusion) sont observées et présentes dans la base de données.

- *Hypothèse 2 : Support commun (overlap)*

$$0 < \pi_t(X_i) = P(T_i = t | X_i) < 1 \quad \text{pour } t \in \{1, 2, \dots, M\} \text{ et } i \in \{1, 2, \dots, N\}$$

Cela signifie que chaque individu a une chance non nulle d'appartenir à chaque groupe du traitement. Cette condition implique qu'un profil de risque, par exemple les véhicules utilitaires, est présent dans toutes les tranches du taux de rabais.

$\pi_t(X_i)$  appelé score de propension (PS) [Imb00] [Li18] représente la probabilité pour un individu  $i$  d'avoir le traitement  $t$  conditionnellement aux covariables  $X$ . Dans les essais contrôlés randomisés, cette probabilité est égale à  $1/M$ . En revanche, dans le cadre de notre étude, les profils de risque ne sont pas répartis d'une manière équiprobable, ainsi le PS de chaque individu n'est pas forcément égal à  $1/M$ . Ce score résume l'ensemble de l'information porté par les covariables, deux profils de risque semblables ont forcément la même probabilité de propension. Nous parlons de la *propriété d'équilibrage du score de propension*. Cette propriété est primordiale pour l'estimation de  $\mu_t$  et de l' $ATE$ , elle est formulée mathématiquement comme suit :

$$T \perp X | \pi(X)$$

En d'autres termes, si le score de propension  $\pi(X)$  est connu, l'ensemble des facteurs de confusion  $X$  ne peut fournir aucune information supplémentaire sur le traitement  $T$ . La *propriété d'équilibrage*

*du score de propension* nous permet ainsi de réduire la dimension du problème en remplaçant l'ensemble des covariables  $X$  par une seule variable  $\pi(X)$ . De ce fait, sous cette propriété, les hypothèses 1 et 2 deviennent :

- $H1 : (Y_i[t_1], Y_i[t_2], \dots, Y_i[t_M]) \perp T_i | \pi_t(X_i)$  pour  $i \in \{1, 2, \dots, N\}$
- $H2 : \pi_t(X_i) > 0$  pour  $t \in \{1, 2, \dots, M\}$  et  $i \in \{1, 2, \dots, N\}$

Le score de propension a une grande popularité dans la revue de littérature sur l'inférence causale. Il est souvent utilisé pour l'appariement (Matching) [Mon13] [D B83]. L'appariement dans le cas d'un traitement binaire consiste à ne garder que les individus proches en termes du score de propension dans le groupe du traitement et le groupe du contrôle. Gary King et Richard Neilson dans leur article intitulé "*Why Propensity Scores Should Not Be Used for Matching*" [Ric19] montrent l'existence de plusieurs problèmes de l'appariement sur le score de propension (Propensity Score Matching-PSM). Selon l'article, le PSM abouti souvent au contraire de l'objectif visé par l'étude. Il augmente les déséquilibres entre les groupes du traitement, l'inefficacité et les biais statistiques. En revanche, les auteurs affirment que leur raisonnement n'implique pas les autres méthodes qui utilisent le score de propension, notamment l'approche **IPTW** utilisée dans le cadre de cette étude.

L' **IPTW** consiste à créer  $M$  populations fictives ou pseudo-populations à partir de la pondération de la population de base en utilisant le score de propension. Ceci permet d'avoir des pseudo-populations semblables à la population de base. Sans perdre de généralités, nous supposons qu'on est dans le cadre du traitement binaire  $T$ , nous avons ainsi deux groupes : un groupe de traitement et un groupe de contrôle. Nous considérons l'existence d'un seul facteur de confusion binaire  $X$  qui divise la population en deux groupes : les rouges qui représentent les hauts risques ; les noirs qui représentent les bas risques. Le schéma ci-dessous (figure 4.2) donne un exemple simplifié de la pondération sur le score de propension dans ce cadre.

$P(T = 1|X = R)$  représente la probabilité qu'un individu appartient au groupe du traitement sachant que c'est un haut risque, elle est égale à  $2/3$  ce qui représente la proportion des hauts risques dans le groupe du traitement. Nous multiplions chaque individu rouge par l'inverse de son score de propension  $3/2$  et comme il existe 2 individus dans le groupe du traitement, nous aurons  $3 = 2 * 3/2$  individus dans la pseudo-population, ce qui est équivalent au nombre d'individus rouges dans la population globale. Nous constatons à partir de cet exemple que la pondération de chaque individu par l'inverse de son score de propension permet d'avoir des pseudo-populations équilibrées et semblables des groupes du traitement et du contrôle. Il est facile ainsi de calculer l'*ATE* et le revenu potentiel de chaque profil de risque. Cet exemple présente une illustration de la logique derrière l'**IPTW**, la sous-section suivante présentera l'estimation de l'effet de traitement (*ATE*) dans le cadre du traitement multidose.

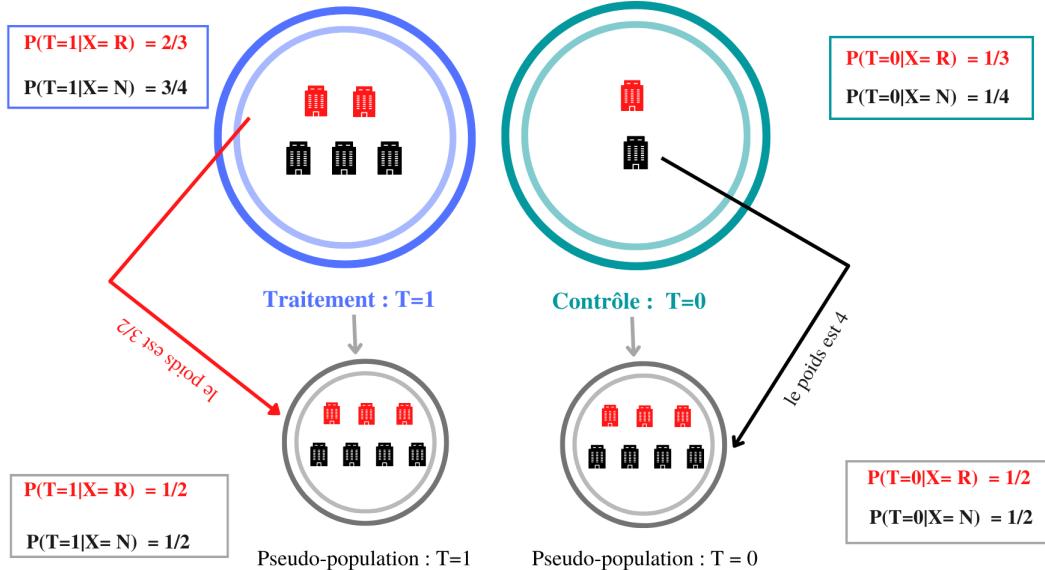


Figure 4.2 – Pondération sur le score de propension

#### 4.2.4 Estimation de l'effet du traitement

À partir des pseudo-populations créées par l'**IPTW** pour les différents groupes du traitement, nous pouvons calculer le revenu moyen du traitement  $t$  toutes choses égales par ailleurs [Lan13]. Ceci revient tout simplement à calculer la moyenne pondérée par l'inverse du score de propension de chaque individu comme suit :

$$\hat{\mu}_t = \frac{\sum_{i=1}^N D_{i,t} w_i[t] Y_i}{\sum_{i=1}^N D_{i,t} w_i[t]} \text{ pour } t \in \{1, 2, \dots, M\}$$

Avec  $w_i[t] = \frac{1}{\pi_t(X_i)}$  et  $D_{i,t}$  une indicatrice qui prend 1 si l'individu  $i$  appartient au groupe  $t$ . Nous pouvons aussi calculer l'effet moyen du traitement d'un groupe  $t$  relativement à un groupe  $t'$  comme suit :

$$ATE(t, t') = \hat{\mu}_t - \hat{\mu}_{t'}$$

Par ailleurs, nous pouvons calculer  $\hat{\mu}_t$  et  $ATE(t, t')$  pour chaque segment de la population en effectuant la moyenne pondérée des individus du segment concerné. Ces estimateurs sont robustes sous l'hypothèse que les pseudo-populations obtenues après la pondération sur le score de propension sont équilibrées, c'est-à-dire chaque groupe du traitement a une distribution analogue à la population de base des facteurs de confusion observés. Il existe plusieurs métriques pour mesurer l'équilibre des facteurs de confusion entre les différents groupes du traitement. La métrique la plus utilisée en revue de

littérature est la différence absolue des moyennes standardisées [Lan13]. Dans le cadre du traitement multidose, elle correspond à :

$$ASMD_{tk} = \frac{|\bar{X}_{tk} - \bar{X}_{pk}|}{\hat{\sigma}_{pk}} \quad \text{pour } t \in \{1, 2, \dots, M\} \text{ et } k \in \{1, 2, \dots, K\}$$

Avec :

- $\bar{X}_{tk} = \frac{\sum_{i=1}^N D_{i,t} w_i[t] X_{tk}^i}{\sum_{i=1}^N D_{i,t} w_i[t]}$  : la moyenne pondérée de la covariable  $k$  et des observations ayant déjà reçu le traitement  $t$ .
- $\bar{X}_{pk}$  : la moyenne non pondérée de la covariable  $k$  pour toute la population.
- $\hat{\sigma}_{pk}$  : l'écart-type non pondéré de la covariable  $k$  pour toute la population.

Intuitivement, cette métrique permet de mesurer si la moyenne standardisée de la covariable  $k$  de la pseudo-population associée au traitement  $t$  est proche de la moyenne standardisée de la population de base. L'avantage de cette métrique est qu'elle ne dépend pas de la taille de l'échantillon et de l'échelle de la covariable quantitative. Concernant les variables qualitatives, nous mesurons la différence entre la proportion pondérée du groupe du traitement  $t$  et la proportion non pondérée de toute la population pour chaque modalité de la variable qualitative. Pour mesurer l'équilibre global de toutes les covariables pour un traitement  $t$ , nous calculons la moyenne de  $ASMD_{tk}$  :  $ASMD_t = \frac{1}{K} \sum_{k=1}^K ASMD_{tk}$ . Il est possible de mesurer l'équilibre d'une covariable  $k$  entre deux groupes du traitement  $t$  et  $t'$  en calculant :  $ASMD_k(t, t') = \frac{|\bar{X}_{tk} - \bar{X}_{t'k}|}{\hat{\sigma}_k}$  avec  $\hat{\sigma}_k$  représentant l'écart type de la covariable  $k$  pour l'échantillon des deux groupes  $t$  et  $t'$ . Une fois l' $ASMD$  calculée, la question naturelle à poser est : à partir de quel seuil de l' $ASMD$  pouvons-nous juger qu'une covariable est équilibrée entre les différents groupes du traitement ? Il est fréquent en revue de littérature sur l'inférence causale de considérer une  $ASDM > 0,25$  comme signe de présence de biais résiduels de confusion dans les pseudo-populations. (McCaffrey et al., 2013) [Lan13] définissent quatre seuils :  $ASDM < 0,2$ ; biais de confusion résiduel négligeable ,  $0,2 < ASDM < 0,4$ ; biais de confusion résiduel modéré,  $0,4 < ASDM < 0,6$ ; présence élevée de biais de confusion résiduel,  $0,6 < ASDM$ ; présence très élevée de biais de confusion résiduel.

En pratique, il est récurrent de trouver des covariables non équilibrées après la pondération sur le score de propension. Ainsi, (Robin et al., 1995) [Jam20] proposent une approche doublement robuste pour estimer l'effet du traitement. Il est important de noter que les estimateurs  $\hat{\mu}_t$  et  $ATE(t, t')$  peuvent aussi être obtenus en effectuant dans le cas d'un revenu continu la régression linéaire pondérée suivante :

$$Y_i = \mu_1 + \sum_{t=2}^M ATE(t, 1) \mathbb{1}_{\{T_i=t\}} + \epsilon_i \quad \text{pour } i \in \{1, \dots, N\}$$

L'approche doublement robuste propose de rajouter à cette régression les facteurs de confusion non équilibrés. L'idée derrière cette approche est de réduire le biais résiduel de confusion en rajoutant les variables non équilibrées. Certains auteurs [Sch07] proposent d'inclure tous les facteurs de confusion de telle sorte à avoir une protection totale contre une mauvaise spécification du score de propension. L'estimateur doublement robuste de l'effet du traitement est convergent et présente une variance plus petite que l'estimateur présenté ci-dessus [GUI01] si la régression du revenu sur le traitement et les autres variables de confusion ou le modèle du score de propension est bien spécifié. Pour la suite de l'étude, nous appelons le modèle de la régression pondérée du revenu sur le traitement **le modèle global**.

#### 4.2.5 Le modèle linéaire généralisé pondéré

Le modèle linéaire généralisé pondéré est utilisé dans le cas où l'échantillon de l'étude n'est pas représentatif de la population globale à laquelle on souhaite généraliser les résultats de la régression. Ce modèle a été développé pour la première fois par Fuller en 1975 pour la régression linéaire et généralisé pour d'autres types de régression par (Binder, 1983) [Ala17].

Nous nous plaçons dans le cadre du modèle GLM présenté dans le troisième chapitre. Nous considérons une population composée de  $n$  individus et un échantillon de cette population composé de  $l$  individus, nous attribuons à chaque individu  $i$  de l'échantillon le poids  $w_i$ . Ceci signifie que cet individu représente  $w_i$  individus de la population globale. Dans les plans d'échantillonnage probabiliste, ce poids représente l'inverse de la probabilité de sélection  $\pi_i$  de l'individu  $i$  [Ala17]. Dans ce cadre, l'estimateur  $\hat{\beta}_l$  du vecteur des coefficients  $\beta$  (de la population globale) à partir des données de l'échantillon est solution des équations de score suivantes [FUL09] [Ala17] :

$$\hat{U}(\beta) = \sum_{i=1}^l w_i x_i \frac{1}{g'(\mu_i) V(\mu_i)} (y_i - \mu_i(\beta)) = \sum_{i=1}^l \hat{U}_i(\beta) = 0$$

$\hat{\beta}_l$  est non biaisé sous hypothèse que  $E_\pi(w_i) = 1$  [FUL09]. La variance de cet estimateur (Binder, 1983) est donnée comme suit :

$$A^{-1} B A^{-1}$$

Avec :

- $A = \sum_{i=1}^l \frac{\partial \hat{U}_i(\beta)}{\partial \beta} \Big|_{\beta=\hat{\beta}_l}$
- $B = \widehat{Var}_\pi \left( \sum_{i=1}^l \hat{U}_i(\beta) \right)$

Nous remarquons que la variance de  $\hat{\beta}_l$  dépend des poids individuels. La présence de poids très

élevés augmente la variance de cet estimateur. Une méthode qui permet de remédier à ce problème est de stabiliser les poids en divisant par un facteur  $h(X)$  qui ne dépend que de la composante déterministe  $X$ . Cette stabilisation des poids ne biaise pas l'estimation de  $\beta$  par l'échantillon en question [FUL09] et permet de rendre l'estimateur  $\hat{\beta}_l$  plus efficient.

La variance est utilisée pour tester la significativité des coefficients à travers le test de significativité de Wald. Pour un coefficient  $\beta_1$  le test de Wald permet de tester l'hypothèse nulle  $\hat{\beta}_1 = 0$  contre l'hypothèse alternative  $\hat{\beta}_1 \neq 0$ , la statistique du test est donnée comme suit :

$$\hat{\beta}_1^T v\hat{a}r(\hat{\beta}_1)^{-1} \hat{\beta}_1$$

Cette statistique a pour distribution asymptotique la loi de  $\chi_2^q$  avec  $q$  le degré de liberté du coefficient  $\hat{\beta}_1$ .

## 4.3 Application pratique

Dans cette sous-section, nous présentons l'application pratique de l'IPTW et les résultats obtenus. Nous commençons tout d'abord par la présentation de l'algorithme K-means qui permet de découper le taux de rabais en classes homogènes. Ensuite, nous présentons la méthodologie de la modélisation de l'élasticité au prix. Enfin, nous commentons les résultats obtenus.

### 4.3.1 Discréétisation du taux de rabais

Le but de cette sous-section est de proposer une méthode qui permet de discréétiser le taux de rabais en nombre réduit de classes tout en préservant une grande proportion de la variance. Ce type de discréétisation permet de ne pas perdre une grande partie de l'information portée par le taux de rabais et d'affecter les résultats de l'étude de l'élasticité au prix. L'algorithme K-means est un outil efficace pour faire une telle discréétisation. Il s'agit d'un algorithme d'apprentissage automatique non supervisé<sup>3</sup> qui permet de regrouper les individus d'un jeu de données en  $k$  classes distinctes en utilisant une mesure de similarité. Dans le cadre de notre étude, nous utilisons la distance euclidienne. Elle est donnée comme suit pour deux individus  $x$  et  $y$  :  $d(x, y) = (TxRabais_x - TxRabais_y)^2$ .

Choisir le nombre de classes optimal  $k$  n'est pas toujours évident dans le cas d'un jeu de données très grand. En effet, un  $k$  très grand ne permet pas d'avoir les similitudes qui existent entre les individus. En revanche, un nombre de classes très petit conduira à des classes qui préservent peu de variance

---

3. L'apprentissage automatique non supervisé est un apprentissage sur des données non étiquetées, l'objectif commun des algorithmes de l'apprentissage non supervisé est de donner une meilleure classification des données non étiquetées

et qui contient beaucoup d'individus. La méthode classique de déterminer  $k$  est de lancer K-means pour plusieurs valeurs de  $k$  et de calculer à chaque fois la proportion de la variance expliquée donnée comme suit :

$$Proportion_{inertie} = \frac{Variance_{inter}}{Variance}$$

Avec :

- $Variance = \sum_j \sum_i d(x_i - x_j)$  et  $x_i$  représente l'individu i.
- $Variance_{inter} = \sum_d \sum_{d'} d(centre_d - centre'_{d'})$  et  $centre_d$  représente le centre de la classe d (la médiane dans notre cas).

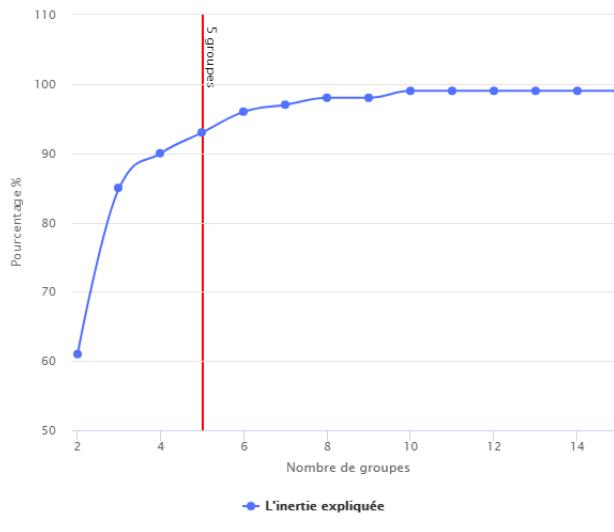


Figure 4.3 – Nombre de classes optimal du K-means

La figure (4.3) permet de remarquer que le nombre de classes optimal est 5 (le coude de la courbe). Ce nombre préserve 94% de l'inertie expliquée. C'est le point à partir duquel la proportion de l'inertie expliquée ne croît plus d'une manière significative.

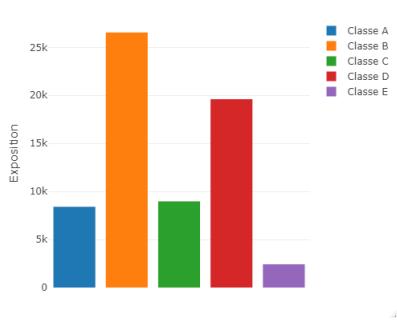


Figure 4.4 – Distribution du taux de rabais

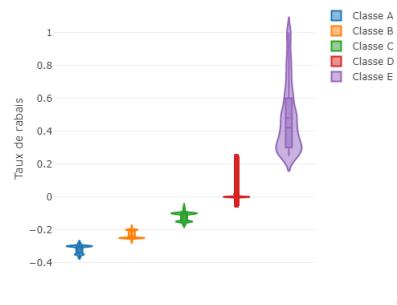


Figure 4.5 – Répartition des observations entre les classes du taux de rabais

L'algorithme K-means a construit 5 classes du taux de rabais : Classe A ; taux de rabais très élevé, Classe B ; taux de rabais élevé, Classe C ; taux de rabais moyen, Classe D ; taux de rabais autour de

zéro, Classe E ; taux de rabais positif.

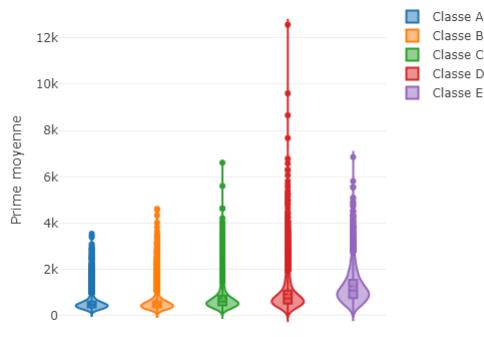


Figure 4.6 – Distribution de la prime moyenne par taux de rabais

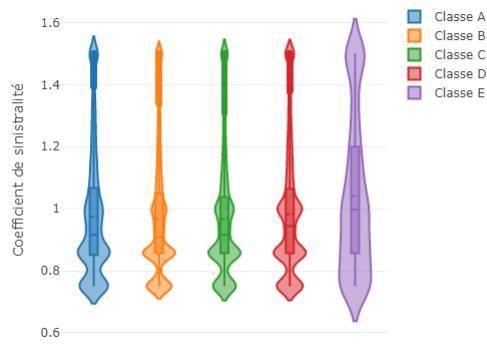


Figure 4.7 – Distribution du coefficient de sinistralité par taux de rabais

Pour détecter les déséquilibres des variables explicatives entre les classes du taux de rabais, nous visualisons la distribution des variables pour chaque classe du taux de rabais. La répartition du coefficient de sinistralité par classe du taux de rabais (figure 4.7) permet d'observer des distributions similaires pour toutes les classes sauf pour la classe E où nous remarquons une distribution différente, le box-plot est décalé en haut ce qui peut être expliqué par le fait que cette classe contient des profils de risque qui ont des grands coefficients de sinistralité. Concernant le graphe de la prime (figure 4.6), nous remarquons que la distribution des petites primes est similaire pour toutes les classes du taux de rabais. En revanche, pour les grandes primes, nous remarquons un déséquilibre entre les classes du taux de rabais, ce qui peut être expliqué par le fait que les grosses affaires ont une grande capacité de négociation.

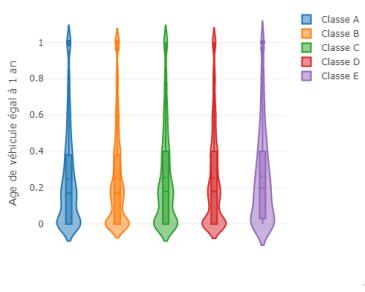


Figure 4.8 – Distribution de la variable âge de véhicule égal à 1 an par taux de rabais

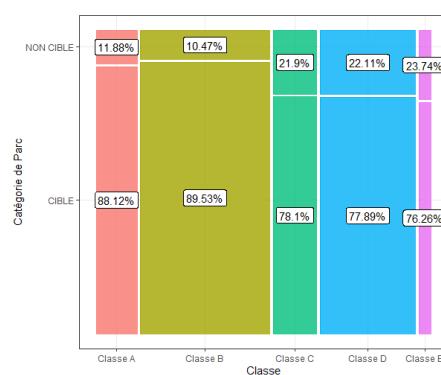


Figure 4.9 – Distribution du taux de rabais par catégorie de Parc

Concernant la proportion de véhicules d'âge égal à un an (figure 4.8), la distribution diffère d'une classe à une autre, ce qui peut être observé à travers la distribution du violon-plot. La répartition de la variable catégorie de Parc (figure 4.9) permet d'observer une grande présence des catégories cibles dans toutes les classes du taux de rabais avec des proportions différentes, ce qui peut être expliqué par

le fait que nous mettons moins de rabais sur les non cibles, car ils sont plus risqués et nous préférons de ne pas diminuer leur prime pour ne pas les avoir en portefeuille.

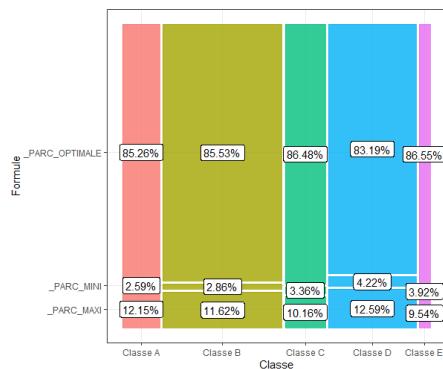


Figure 4.10 – Distribution du taux de rabais par formule du Parc

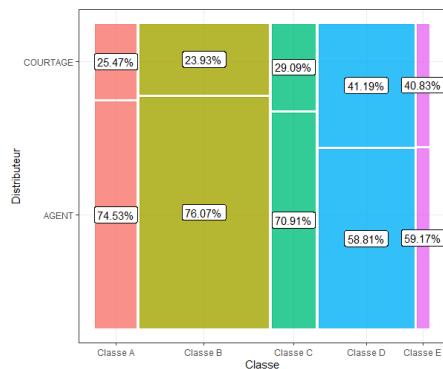


Figure 4.11 – Distribution du taux de rabais par réseau de distribution

La répartition de la variable réseau de distribution (figure 4.11) permet d'observer des proportions différentes de la modalité agent pour les différentes classes, cette proportion est grande pour la classe B (taux de rabais élevé) ce qui peut être expliqué par le fait que les agents proposent de rabais importants et devient de plus en plus petite pour les autres classes de rabais supérieures. La variable formule du Parc donne l'information sur la formule choisie par le client. La formule minimale (\_PARC\_MINI) contient la garantie responsabilité civile automobile obligatoire par le régulateur, la formule optimale (\_PARC\_OPTIMALE) contient uniquement les garanties nécessaires (à savoir : la responsabilité civile, la garantie dommage tout accident, garantie catastrophes naturelles...), la formule maximale (\_PARC\_MAXI) contient toutes les garanties proposées par le produit y compris les garanties accessoires. Environ 80% des clients choisissent la formule optimale. La répartition du taux de rabais par formule choisie (figure 4.10) permet d'observer des proportions proches de la formule optimale pour les différentes classes du taux de rabais. Pour les deux autres formules, il existe des petites différences entre les différentes classes de rabais.

### 4.3.2 Modélisation de l'élasticité au prix

Dans cette sous-section, nous présentons la méthodologie utilisée pour modéliser l'élasticité au prix ainsi que les interprétations des sorties R et la validation de la méthodologie. Tout d'abord, nous commençons par l'estimation du score de propension (PS) à l'aide du modèle Xgboost. Ensuite, nous calculons les poids individuels et nous évaluons l'équilibre des covariables entre les différentes classes à l'aide de la métrique ASMD présentée dans la section précédente et nous vérifions l'Overlap à l'aide de la visualisation des Boxplots des PS. Enfin, nous implémentons le modèle global du taux de transformation à l'aide d'une régression logistique

#### 4.3.2.1 Estimation du score de propension (PS)

Le but de l'estimation du PS est de proposer un score qui permet de compresser toute l'information portée par les facteurs de confusion concernant le taux de rabais. Ainsi, il est inutile d'inclure dans l'estimation du PS les variables qui ne sont pas corrélées avec le taux de rabais. Inclure ces variables peut rajouter du bruit au modèle et nuire à l'effort du PS de supprimer le biais de confusion. À ce stade, il est important de signaler que nous ne cherchons pas à avoir une meilleure qualité prédictive du modèle PS. Mais, nous cherchons un modèle qui permet d'équilibrer les facteurs de confusion entre les classes du taux de rabais afin de satisfaire la *propriété d'équilibrage du score de propension*. De ce fait, l'estimation du PS n'est pas un problème de prédiction. Le modèle classique utilisé pour estimer le PS en littérature de l'inférence causale est la régression logistique multinomiale (voir annexe.1). L'inconvénient de cette approche est que la régression logistique est un modèle paramétrique basé sur l'hypothèse que le taux de rabais discrétisé est une variable aléatoire qui suit une loi multinomiale où la somme des probabilités d'appartenance à chacune des 5 classes est égale à 1. En effet, avoir une somme des probabilités égale à 1 n'est pas indispensable dans notre cas, puisqu'on pondère chaque classe du taux de rabais par l'inverse de son propre score de propension. Notre objectif principal de l'estimation du PS est d'avoir de bons poids qui permettent de supprimer le biais de confusion. De ce fait, nous aimeraisons que le modèle PS capte tout type d'association linéaire et non linéaire et d'interaction existant entre les facteurs de confusion. (McCaffrey et al., 2013) [Lan13] proposent une méthodologie basée sur les modèles GBM (generalized boosted models), Nous résumons cette approche en 3 étapes :

1. Créer  $M$  variables indicatrices définies comme suit :  $D_t^i = 1$  si l'individu  $i$  a reçu le traitement  $t$  et 0 sinon.
2. Pour chaque variable indicatrice  $D_t$  calibrer un modèle GBM à travers la minimisation d'une métrique d'équilibre choisie, par exemple, l' $ASMD_t$ .
3. Pour chaque traitement  $t$ , estimer le score de propension  $SP_t$  à travers le modèle GBM calibré et calculer les poids individuels via la formule suivante :

$$\sum_{t=1}^M D_t^i \frac{1}{SP_t^i}$$

Dans le cadre de cette étude, nous faisons recours à la même approche pour estimer le score de propension. Concrètement, nous utilisons le modèle Xgboost pour estimer le score de propension et la métrique  $ASMD_t$  pour optimiser le modèle. L'implémentation de cette approche est disponible en

open source sous R à travers le package *twang*<sup>4</sup> (Toolkit for Weighting and Analysis of Nonequivalent Groups). Elle est accessible en faisant appel à une seule fonction *mnp()* qui nécessite la spécification d'un ensemble de paramètres, nous en citons ici quelques-uns :

- *formula* : la formule du modèle, elle est donnée comme suit :

$$Tx \text{ rabais discretise} = Covariable_1 + Covariable_2 + \dots + Covariable_K$$

- *estimand* : l'estimateur causal d'intérêt, dans notre cas, nous allons calculer l'*ATE*.

- *n.trees* : nombre d'itérations maximal du modèle *Xgboost*, dans notre cas, nous prenons 2000.

Plus le nombre d'itérations est grand, plus le modèle est complexe.

- *stop.method* : la métrique d'optimisation qui mesure l'équilibre global des covariables, dans notre cas nous utilisons *ASMD<sub>t</sub>*.

- *version* : la version du package *twang* à utiliser, dans notre cas nous utilisons, la version optimisée qui intègre le modèle *Xgboost*.

Concernant les covariables, nous utilisons toutes les variables tarifaires et les variables qui impactent à la fois le taux de transformation et le taux de rabais, à savoir : l'âge du véhicule, la prime avant le rabais, le coefficient de sinistralité, la région, la marque du véhicule, le type de la flotte, l'activité de l'entreprise et d'autres caractéristiques de la flotte. En revanche, nous éliminons les autres variables telles que l'ancienneté de devis et la densité de population. Nous fixons les paramètres de la fonction *mnp()* comme indiqué ci-dessus et nous lançons la fonction. Dans le cadre de notre étude, la fonction *mnp()* calibre 5 modèles d'*Xgboost*, chaque modèle correspond à une classe du taux de rabais. Pour chaque classe du taux de rabais, *mnp()* retourne l'estimation du score de propension à travers le modèle qui minimise l'*ASMD<sub>t</sub>*. Les graphes (4.12) et (4.13) montrent la Learning Curve du modèle *Xgboost* de la classe A et de la classe E.

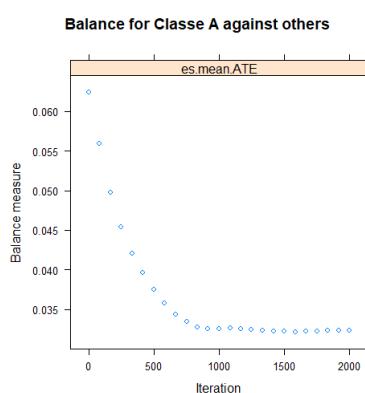


Figure 4.12 – Learning Curve classe A

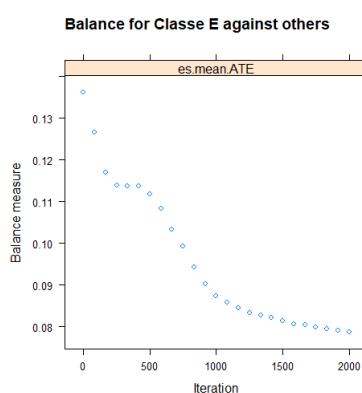


Figure 4.13 – Learning Curve classe E

---

4. 'twang' est un package qui contient un ensemble de fonctions et de procédures pour estimer, analyser le score de propension et calculer les poids du score de propension. Il est disponible via : Stata, SAS et R. Il était développé en 2004 par les chercheurs de Rand Corporation. Il existe plusieurs tutoriels de ce package sur le site de Rand Corporation. Le package est aussi disponible sur le CRAN : <https://cran.r-project.org/web/packages/twang/twang.pdf>

*es.mean* fait référence à *absolute standardized effect size mean*, il s'agit de l' $ASMD_t$  définie dans la section précédente. La Learning Curve donne l' $ASMD_t$  pour chaque niveau de complexité, elle permet de savoir le nombre d'itérations optimal du modèle final et l' $ASMD_t$  pour chaque classe. Par exemple pour la classe A nous remarquons que le nombre d'itérations optimal est proche de 1500 qui correspond approximativement à une  $ASMD_t$  égale à 0,03.

#### 4.3.2.2 Évaluation de l'overlap et de l'équilibre des facteurs de confusion

Après la modélisation du score de propension, il est important de vérifier si toutes les covariables sont équilibrées entre les classes du taux de rabais. Ceci dans le but de mesurer l'équilibre global et de détecter les covariables déséquilibrées. Pour mesurer l'équilibre d'une covariable entre un couple  $t$  et  $t'$  de classes, nous utilisons la métrique  $ASMD_k(t, t')$ .

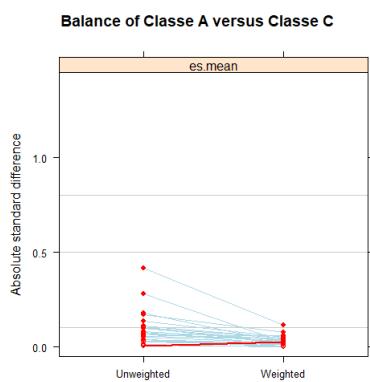


Figure 4.14 – Équilibre entre A et C

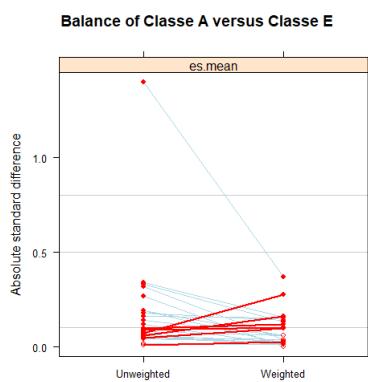


Figure 4.15 – Équilibre entre A et E

En analysant les figures (4.14) et (4.15) et l'équilibre des covariables entre les couples de classes deux à deux, nous constatons que la pondération a réduit les déséquilibres des covariables entre les différentes classes du taux de rabais. Nous remarquons aussi que l' $ASMD_k(t, t')$  est généralement inférieure à 0,4 pour tous les couples, ce qui signifie que l'on a des biais de confusions résiduels petits. Sauf pour le couple (Classe A et Classe E) où nous observons une covariable non équilibrée avec  $ASMD_k(t, t') > 0,4$ . Ceci s'explique par le fait que la classe A (taux de rabais très élevé) est très loin de la classe E (taux de rabais positif) et elles contiennent des profils de risque très distincts. Pour le même couple, nous observons que la pondération a fait augmenter le déséquilibre entre certaines variables (les lignes rouges).

L'hypothèse de l'*Overlap* (support commun ou positivité) est importante pour avoir une estimation robuste de l'effet du traitement. Cette hypothèse permet d'éviter le problème des grands poids. En effet, une observation avec un poids très élevé (un score de propension proche de 0) influence d'une manière significative les estimations puisqu'elle est prise en compte dans les calculs autant de fois que

son poids. De même une observation avec un score de propension proche de 1 influence peu le résultat de l'estimation puisqu'elle est prise une seule fois dans le calcul. Ce problème augmente la variance des estimations effectuées puisque la précision des estimations sera très dépendante de l'échantillon observé. En outre, une observation avec un score de propension égal à 1 pour une classe  $t$  signifie qu'il est impossible de calculer son taux de transformation pour une autre classe que  $t$ . L'analyse des *box-plots* ci-dessous et en annexe (voir annexe.2) de chaque score de propension par classe du taux de rabais permet d'observer l'existence des observations avec des scores de propension proches de 0 et de 1.

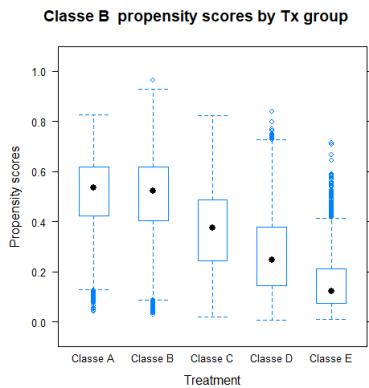


Figure 4.16 – Overlap classe B

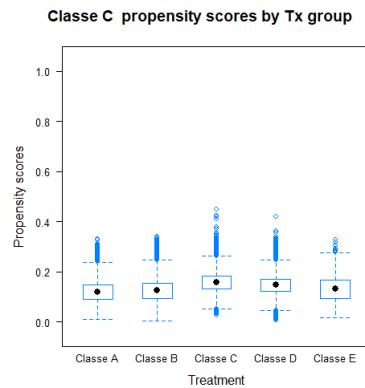


Figure 4.17 – Overlap classe C

En revue de littérature de l'inférence causale, il n'existe pas de seuil à partir duquel nous pouvons juger que l'hypothèse de *Overlap* est rejetée. Mais, il existe plusieurs méthodes pour remédier à ce problème. La méthode la plus facile est de supprimer les observations avec un score de propension inférieur ou supérieur à un seuil choisi. Cette méthode n'est pas efficace puisqu'elle mène à une perte importante d'observations. Il est aussi possible de faire une troncation des données, c'est-à-dire remplacer les observations avec un score de propension inférieur ou supérieur à un seuil par le seuil en question. Toutefois, avoir des scores de propension proches de 0 et de 1 peut être un problème de spécification du modèle de score de propension. Pour avoir une protection contre ces problèmes, nous optons pour une estimation doublement robuste à travers la régression pondérée (modèle global). Nous détaillerons ce modèle dans la sous-section suivante.

#### 4.3.2.3 Le modèle global

Le modèle global nécessite une bonne spécification. De ce fait, nous incluons toutes les variables qui peuvent impacter le taux de transformation. Pour estimer ce modèle, nous utilisons la régression logistique pondérée. Intuitivement, nous pouvons présenter la régression logistique pondérée comme une régression logistique effectuée sur une population fictive où toutes les classes du taux de rabais

contiennent la même répartition des facteurs de confusion qui est semblable à la répartition de la population de base. Pour remédier au problème des poids élevés présenté dans la sous-section précédente, nous utilisons les poids stabilisés (stabilized weights) [Bab00] donnés comme suit :

$$sw_i = \sum_{t=1}^5 D_t^i \frac{P(T_i = t)}{\pi_t(X_i)}$$

Le numérateur [ $P(T_i = t)$ ] est obtenu à travers une régression logistique multinomiale sans variables explicatives du taux de rabais, le lecteur peut se référer à l'annexe 1 pour la régression logistique multinomiale, le dénominateur [ $\pi_t(X_i)$ ] représente le score de propension obtenu via le modèle *xgboost* implémenté précédemment. Cette stabilisation ne biaise pas les résultats de la régression [Bab00] [Ala17]. Cependant, elle permet d'avoir des estimateurs plus efficents et plus stables [Bab00].

Rappelons que notre objectif est de calculer la dérivée :

$$\frac{\partial \hat{f}(CT)}{\partial CT}$$

De ce fait, nous préférons utiliser le coefficient technique continu au lieu du coefficient technique discrétisé, ceci suppose la validité de l'implication suivante :

$$CT_{discrétisé} \perp X | \pi(X) \Rightarrow CT_{continu} \perp X | \pi(X)$$

Avec :

- $X$  les facteurs de confusion.
- $\pi(X)$  le score de propension.

Cette implication suppose que conditionnellement au score de propension, les facteurs de confusion ne peuvent fournir aucune information supplémentaire sur le coefficient technique continu. En réalité cette implication n'est pas toujours vérifiée parce que le coefficient technique discrétisé ne contient pas toute l'information portée par le coefficient technique continu, le découpage du coefficient technique par l'algorithme K-means a introduit une perte de 6% de l'inertie expliquée par le coefficient technique continu. Pour capter les corrélations linéaires et non linéaires du coefficient technique, nous rajoutons les transformations polynomiales d'ordre 2 et d'ordre 3 du coefficient technique. Le rajout de plusieurs transformations polynomiales peut conduire au problème de sur-apprentissage<sup>5</sup>,

---

5. [https://pageperso.lis-lab.fr/~alexis.nasr/Ens/MASCO\\_AA/intro\\_regression.pdf](https://pageperso.lis-lab.fr/~alexis.nasr/Ens/MASCO_AA/intro_regression.pdf)

nous nous arrêtons ainsi au polynôme d'ordre 3. Le modèle global peut être présenté comme suit :

$$\text{logit}(f(CT, X)) = \beta_0 + \beta_1 CT + \beta_2 CT^2 + \beta_3 CT^3 + \sum_{i=4}^P \beta_i x_i$$

Pour implémenter le modèle global, nous supposons que les données sont individuelles, c'est-à-dire chaque client est présent une seule fois dans le jeu de données pour une unique affaire. Nous supposons aussi que toutes les polices d'affaires nouvelles sont indépendantes, déterministes pour les variables explicatives et aléatoirement distribuées suivant une loi de Bernoulli dont le paramètre peut être approché par une combinaison linéaire de l'ensemble des variables explicatives pour la variable d'intérêt qui est la décision de transformation.

### 4.3.3 Résultats et discussions

Estimation non paramétrique de l'élasticité :

Il est possible de déterminer  $\hat{\mu}_t$  pour chaque classe du taux de rabais et de calculer l' $ATE(t, t')$  en utilisant la moyenne pondérée par l'inverse du score de propension. Le calcul analytique de la variance de l'estimateur  $\hat{\mu}_t$  est compliqué en réalité. Cet estimateur dépend des poids calculés à travers l'estimation du score de propension par le modèle Xgboost. La variance peut être calculée d'une manière empirique à travers la méthode bootstrap en faisant plusieurs fois un tirage avec remise de 1000 observations et en calculant l'estimateur  $\hat{\mu}_t$  pour chaque échantillon tiré [QUE18]. Le graphe (4.18) donne l'estimation de  $\hat{\mu}_t$  pour chaque classe du taux de rabais.

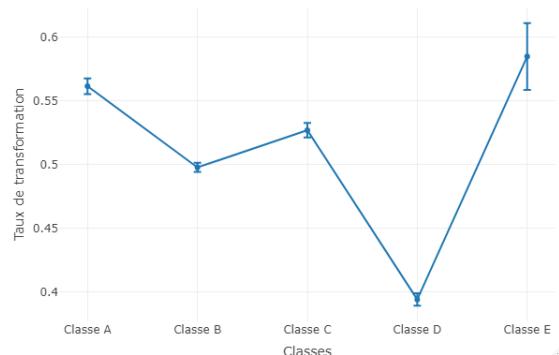


Figure 4.18 – Courbe ATE

Le graphe (4.18) peut être interprété comme suit : par exemple, le passage de la classe C (rabais moyen) à la classe D (rabais autour de zéro) fait baisser en moyenne le taux de transformation de 12 points toutes choses égales par ailleurs. Il est important de remarquer que le passage de n'importe

quelle classe à la classe E (rabais positif) fait augmenter le taux de transformation, ce qui semble non intuitif. Ceci peut être expliqué par le fait qu'il existe encore des covariables non équilibrées, la variance  $\hat{\mu}_E$  est très grande ainsi cet estimateur semble non stable. Toutefois, il est possible d'estimer  $\hat{\mu}_t$  pour un segment spécifique de clients et calculer le gain ou la perte nette en termes du taux de transformation du passage d'une classe à une autre d'une manière non paramétrique.

### Résultats du modèle global :

Nous implémentons le modèle global sous R avec la fonction `svyglm()` du package *survey*<sup>6</sup>. Cette fonction permet d'estimer le modèle GLM pour les données des enquêtes complexes et pour les données pondérées par l'inverse du score de propension.

Les résultats du modèle global (voir annexe.2) et le test de significativité de *Wald* donnent que le coefficient de l'ordre 1 du coefficient technique est non significatif au seuil de 5%, ainsi ce coefficient n'est pas pris en considération dans le calcul de l'élasticité. En revanche, les coefficients des deux transformations polynomiales du coefficient technique sont significatifs au seuil de 5%. L'élasticité au prix relative est calculée à partir du modèle global comme suit :

$$\begin{aligned}
 \text{logit}(\hat{f}(CT)) &= \hat{\beta}_0 + \hat{\beta}_2 CT^2 + \hat{\beta}_3 CT^3 + \sum_{i=4}^P \hat{\beta}_i x_i \\
 \Rightarrow \frac{\partial \text{logit}(\hat{f}(CT))}{\partial CT} &= 2\hat{\beta}_2 CT + 3\hat{\beta}_3 CT^2 \\
 \iff \frac{\partial \text{log}(\hat{f}(CT))}{\partial CT} - \frac{\partial \text{log}(1 - \hat{f}(CT))}{\partial CT} &= 2\hat{\beta}_2 CT + 3\hat{\beta}_3 CT^2 \\
 \iff \frac{\partial \hat{f}(CT)}{\partial CT} \frac{1}{\hat{f}(CT)} + \frac{\partial \hat{f}(CT)}{\partial CT} \frac{1}{1 - \hat{f}(CT)} &= 2\hat{\beta}_2 CT + 3\hat{\beta}_3 CT^2 \\
 \iff \frac{\partial \hat{f}(CT)}{\partial CT} \frac{1}{\hat{f}(CT)(1 - \hat{f}(CT))} &= 2\hat{\beta}_2 CT + 3\hat{\beta}_3 CT^2 \\
 \iff \frac{\partial \hat{f}(CT)}{\partial CT} &= (2\hat{\beta}_2 CT + 3\hat{\beta}_3 CT^2) \hat{f}(CT)(1 - \hat{f}(CT)) \\
 \iff e(CT) &= -(2\hat{\beta}_2 CT + 3\hat{\beta}_3 CT^2) CT(1 - \hat{f}(CT))
 \end{aligned}$$

---

6. 'survey' est un package qui contient un ensemble de fonctions et de procédures pour analyser les données des enquêtes complexes. Il était développé en 2010 par Thomas Lumley un professeur en biostatistique à l'Université d'Auckland en Nouvelle-Zélande.

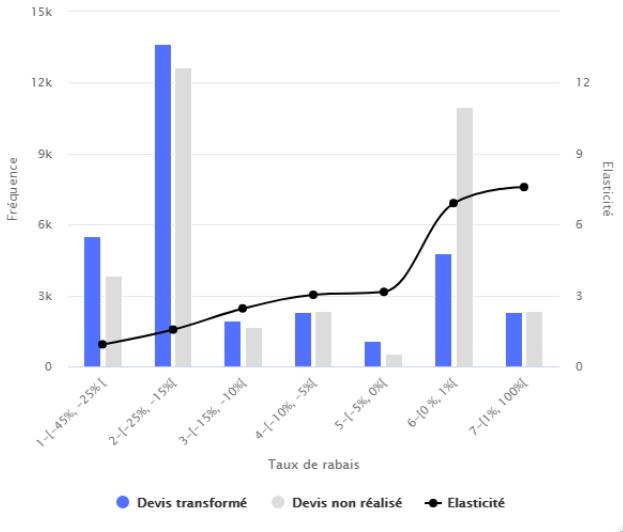


Figure 4.19 – L'élasticité par taux de rabais

La figure (4.19) permet d'observer l'élasticité au prix selon les tranches du taux de rabais, nous remarquons que l'élasticité est croissante en fonction du taux de rabais. Cette élasticité varie de 0.93 à 7.58. Rappelons qu'une élasticité égale 7.58 signifie qu'une augmentation de la prime de 1% entraîne une baisse du taux de transformation de 7.58%. Une élasticité inférieure à 1 signifie que le devis est peu élastique. En revanche, une élasticité supérieure à 1 signifie que le devis est élastique. De ce fait, les devis qui ont bénéficié de rabais minimaux sont peu élastiques et les devis qui n'ont pas bénéficié de rabais, voire de majoration de prime sont très élastiques à une augmentation ou baisse future de la prime.

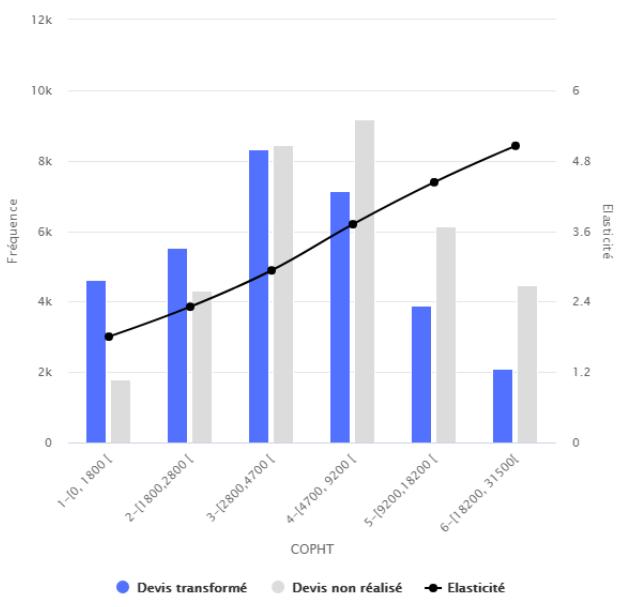


Figure 4.20 – L'élasticité par prime

S'agissant de la prime (figure 4.20), l'élasticité au prix varie entre 1.08 et 5, elle est croissante

en fonction des segments de la prime, ce qui apparaît intuitif puisque les gros devis ont une grande capacité de négociation sur le marché. Toutes fois, les clients raisonnent en termes d'échelle et non pas en termes de pourcentage, augmenter de 1% une prime de 100 € n'est pas équivalent à 1% une prime de 10 000 €.

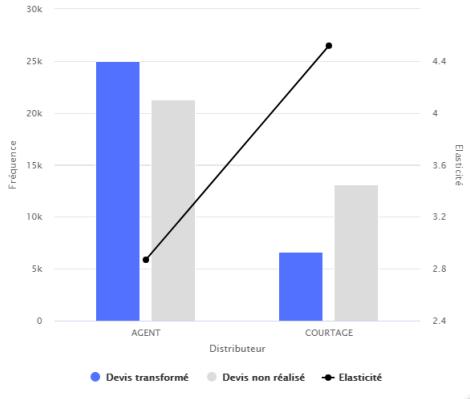


Figure 4.21 – L'élasticité par distributeur

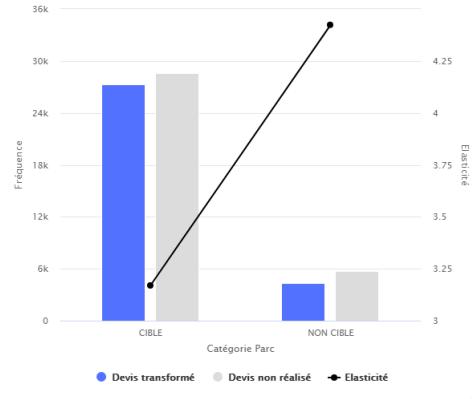


Figure 4.22 – L'élasticité par catégorie du Parc

La figure (4.21) permet d'observer l'élasticité selon le réseau de distribution, le réseau des agents a une faible élasticité (2.28) par rapport au réseau de courtage (4.5) ce qui apparaît intuitif puisque chez un courtier le client a une multitude d'offres à comparer. En revanche, chez un agent, le client a une seule offre, celle d'AXA France. Concernant la variable catégorie du Parc (figure 4.22), nous remarquons que les catégories cibles sont moins élastiques que les catégories non cibles, ce qui peut être expliqué par le fait que les catégories rentables (cibles) bénéficient de rabais importants et par le fait que la plupart des devis non cibles sont émis par les courtiers.

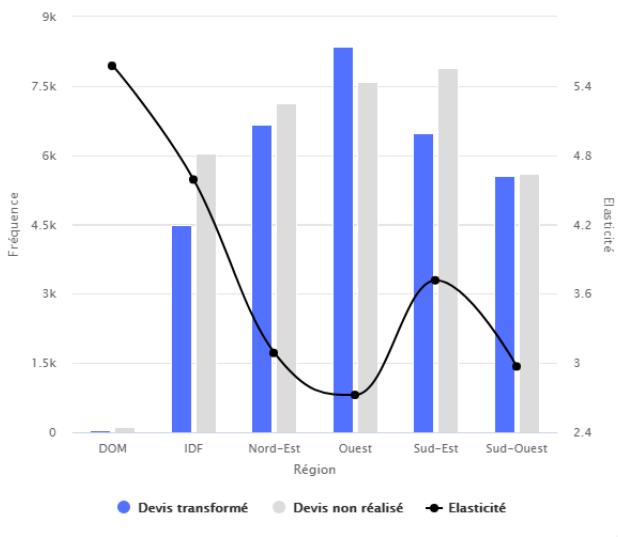


Figure 4.23 – L'élasticité par région

S'agissant de la variable région (figure 4.23), l'élasticité varie d'une région à une autre, en premier rang, nous trouvons la région DROM ceci peut être expliqué par le fait que l'estimation de l'élasticité

au prix pour cette région est influencée par le faible effectif d'observations. En second rang, nous trouvons la région Île-de-France ce qui peut être expliqué par la large présence des courtiers dans cette région et en dernier rang, nous trouvons la région OUEST ce qui peut être expliqué par une prédominance du réseau des agents dans cette région.

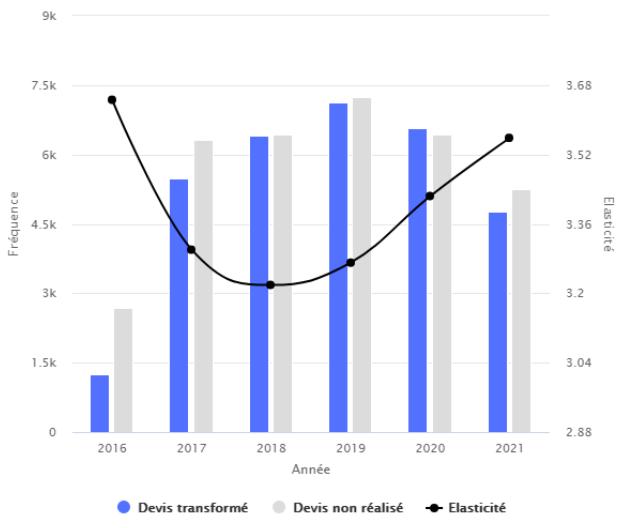


Figure 4.24 – L'élasticité par année d'effet

La figure (4.24) donne l'élasticité au prix par année d'effet du devis, les années 2016 et 2021 ne sont pas prises en compte pour la comparaison de l'élasticité à cause du manque d'observations. L'élasticité par année d'effet varie entre 3.3 et 3.4, elle atteint son minimum en 2018 et son maximum en 2020. La variation de l'élasticité est relativement faible d'une année à une autre.

## Conclusion partielle

Ce chapitre a permis de proposer une méthodologie pour estimer l'élasticité au prix. Cette élasticité représente un outil important pour étudier l'impact d'une stratégie de changement tarifaire sur le taux de transformation. L'élasticité au prix est croissante en fonction de la prime et du taux de rabais. L'analyse de l'élasticité par profil de risque permet de caractériser les profils les plus élastiques. Les profils cibles sont moins élastiques puisqu'ils bénéficient de rabais importants. Le réseau de courtage est plus élastique en raison de l'offre variée proposée par ce réseau. La région Île-de-France représente la région la plus élastique en raison de la prédominance du réseau de courtage dans cette région.

# Chapitre 5

## Application concrète : étude de l'impact de l'inflation sur la rentabilité des affaires nouvelles

### Préambule

Le taux de transformation représente un outil de suivi de la demande du produit Parc dénommé, plus le taux de transformation est élevé meilleure est la demande. Toutefois, l'élasticité au prix permet d'anticiper la variation future de la demande suite à un changement tarifaire. En général, il existe un compromis entre la demande et la rentabilité. Les prospects déficitaires ont généralement une forte probabilité de transformation de devis et les prospects rentables ont une faible probabilité de transformation. Ce chapitre a pour vocation d'étudier l'effet des différents scénarios de l'inflation sur la demande et la rentabilité des affaires nouvelles à court terme. Il commencera par la présentation de la fonction de demande, ensuite présentera la marge comme mesure de rentabilité prospective et enfin discutera l'effet de l'inflation sur la rentabilité.

### 5.1 La fonction de demande

La fonction de demande donne la probabilité qu'un client accepte d'acheter les garanties du produit Parc dénommé en fonction du prix. Elle est composée de deux parties :

- **Partie statique** : la probabilité qu'un client accepte d'acheter les garanties du produit Parc dénommé avec le prix actuel  $P_0$  (le prix affiché sur son devis). Elle est calculée à partir du modèle GLM du taux de transformation  $Tt(P_0)$ .

- **Partie dynamique :** la variation de cette probabilité suite à une variation  $\epsilon\%$  de la prime. Elle est calculée à partir de la fonction de l'élasticité au prix comme suit :  $Tt(P_0)e(P_0)\epsilon\%$ .

Ainsi la fonction de demande peut s'écrire comme suit :

$$Demande(\epsilon\%) = Tt(P_0)(1 + e(P_0)\epsilon\%)$$

En d'autres termes, cette fonction représente le taux de transformation espéré si l'on propose au client la prime  $P_0(1 + \epsilon\%)$  au lieu de la prime  $P_0$ . La figure (5.1) donne la fonction de demande en fonction du réseau de distribution, catégorie du Parc et selon les variations de la prime. Si les stratégies tarifaires des concurrents sont maintenues constantes, une revue à la hausse de la prime de +5% fait baisser la demande totale (le taux de transformation espéré) de 6 points. Ceci apparaît intuitif puisque le marché de l'assurance des flottes automobiles est très concurrentiel.

Variation de la prime	AGENT			COURTAGE			Total général
	CIBLE	NON CIBLE	Total	CIBLE	NON CIBLE	Total	
0%	53%	48%	53%	36%	32%	35%	47%
1%	52%	47%	52%	35%	31%	34%	46%
2%	51%	45%	50%	34%	30%	33%	45%
3%	49%	43%	48%	33%	29%	32%	43%
4%	48%	42%	47%	31%	27%	31%	42%
5%	46%	40%	45%	30%	26%	29%	41%
6%	45%	38%	44%	29%	25%	28%	39%
7%	43%	36%	42%	28%	23%	27%	38%
8%	41%	35%	41%	26%	22%	25%	36%
9%	40%	33%	39%	25%	21%	24%	35%

Figure 5.1 – La fonction de demande par variation de prime

En conséquence, une revue à la hausse du tarif pour compenser l'effet de l'inflation au-dessus des concurrents peut détériorer la demande et la production du produit Parc dénommé. En termes de rentabilité, une revue à la hausse de la prime de +5% fait baisser la demande des catégories rentables (cibles) de 7 points pour le réseau des agents et de 6 points pour le réseau des courtiers. En contrepartie, la même revue à la hausse entraîne pour les non cibles une baisse de la demande de 8 points pour les agents et de 6 points pour le réseau du courtage.

## 5.2 La rentabilité d'une affaire nouvelle

La notion de rentabilité nécessite la définition d'une mesure de rentabilité prospective. Au moment du renouvellement, la rentabilité de chaque affaire est calculée par le rapport entre la charge des sinistres et les primes acquises (S/C). Par ailleurs, au moment de la souscription, la rentabilité nécessite le calcul de la charge des sinistres. Cette quantité n'est connue qu'à l'échéance du contrat. Ainsi, une

estimation de la charge des sinistres est nécessaire pour le calcul de la rentabilité attendue de chaque affaire. La charge totale des sinistres de chaque contrat est donnée comme suit :

$$S = \sum_{i=1}^N Y_i$$

Avec  $N$  la fréquence des sinistres et  $Y_i$  le coût du sinistre  $i$ .  $(Y_i)_{i \in \{1, 2, \dots, N\}}$  sont des variables aléatoires indépendantes et identiquement distribuées et  $N$  est une variable aléatoire indépendante des coûts de sinistres. La prime pure  $PP$  représente la charge des sinistres espérée, elle est calculée à partir de la charge totale comme suit :

$$PP = E(S) = E\left(\sum_{i=1}^N Y_i\right) = E(E\left(\sum_{i=1}^N Y_i/N\right)) = E(Y_1 N) = E(Y_1)E(N)$$

La prime pure est estimée par le produit de la fréquence moyenne et le coût moyen des sinistres modélisés avec les modèles linéaires généralisés (GLM). Toutefois, la prime pure modélisée ( $\hat{PP}$ ) représente une prévision de la charge moyenne annuelle des sinistres de chaque prospect. Nous allons l'utiliser dans la suite de l'étude pour estimer la charge moyenne prospective des sinistres afin de calculer la marge, l'indicateur de rentabilité attendue donné comme suit :

$$\text{marge} = COPHT - \hat{PP}$$

Avec  $\hat{PP}$  représente la prime pure annuelle modélisée et la *COPHT* (la Cotisation Potentielle Hors Taxes) représente la prime commerciale hors taxes annualisée payée par l'assuré. La marge permet de calculer la différence entre le coût du risque (la prime pure) et la prime payée par l'assuré, elle permet de mesurer la capacité de la prime commerciale de payer les sinistres futurs d'une affaire nouvelle sur une vision d'un an. L'inflation affecte directement la rentabilité via le coût moyen futur des sinistres. Les sinistres futurs ne constituent pas les seuls éléments qui affectent la rentabilité prospective d'une affaire nouvelle, il existe d'autres facteurs tels que les commissions, coûts de la compagnie, les fluctuations des taux et la rémunération des actionnaires. Pour pouvoir comparer les affaires nouvelles, nous préférions utiliser la marge en pourcentage de la prime commerciale donnée comme suit :

$$\text{marge\%} = \frac{COPHT - \hat{PP}}{COPHT}$$

Une marge égale à 20% signifie que 80% de la prime permettra de payer les sinistres futurs de l'affaire nouvelle. Une marge inférieure à 0% signifie que l'affaire est déficitaire. En revanche, une

marge supérieure à 0% signifie que l'affaire est rentable. Nous constatons à partir de la figure (5.2) que les devis transformés et les devis non réalisés ont la même distribution de la marge. Cependant, nous remarquons que les devis non réalisés sont plus rentables que les devis transformés (la courbe en bleue) ce qui signifie que nous attirons plus de clients déficitaires que de clients rentables.

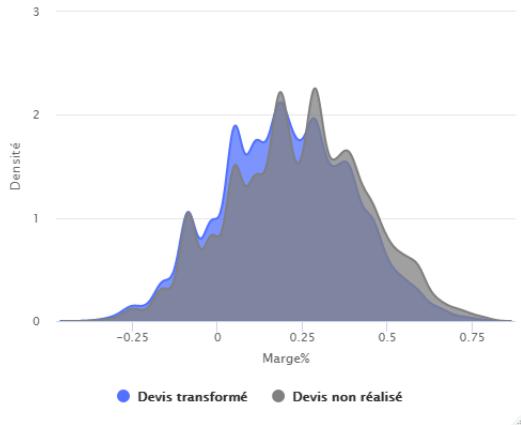


Figure 5.2 – Densité de la marge en pourcentage

Pour un ensemble de devis  $L$ , la marge espérée est calculée à partir de la fonction de demande comme suit :

$$marge = \sum_{i \in L} demande_i(\epsilon\%) * (COPHT_i * (1 + \epsilon\%) - \hat{PP}_i)$$

et la marge en pourcentage :

$$marge\% = \frac{\sum_{i \in L} demande_i(\epsilon\%) * (COPHT_i * (1 + \epsilon\%) - \hat{PP}_i)}{\sum_{i \in L} demande_i(\epsilon\%) * COPHT_i * (1 + \epsilon\%)}$$

La figure (5.3) donne la marge espérée en fonction du réseau de distribution et catégorie du Parc selon les variations de la prime. À l'opposé de la fonction de demande, la marge est croissante en fonction des variations de prime, si les stratégies tarifaires des concurrents sont maintenues constantes, une revue à la hausse de la prime de +5% fait croître de 3 points la marge.

Variation de la prime	AGENT			COURTAGE			Total général
	CIBLE	NON CIBLE	Total	CIBLE	NON CIBLE	Total	
0%	23%	20%	23%	28%	25%	26%	23%
1%	23%	20%	23%	28%	25%	27%	24%
2%	24%	20%	23%	29%	25%	27%	24%
3%	25%	20%	23%	29%	26%	28%	25%
4%	25%	21%	24%	30%	26%	28%	25%
5%	25%	21%	24%	30%	27%	29%	26%
6%	26%	22%	25%	30%	27%	29%	26%
7%	26%	22%	25%	31%	28%	30%	26%
8%	27%	22%	25%	31%	28%	30%	27%
9%	27%	23%	26%	31%	28%	30%	27%

Figure 5.3 – La marge par variation de prime

Pour cette raison, une revue à la hausse du tarif pour compenser l'effet de l'inflation au-dessus des concurrents peut améliorer la rentabilité, mais pas autant que la baisse de la production. En termes du réseau de distribution, une revue à la hausse de la prime de +5% fait augmenter la marge du réseau agent d'un point et de 3 points du réseau courtier. Toutefois, le réseau de courtage est plus rentable que le réseau des agents pour les catégories cibles.

### 5.3 Étude de l'impact des différents scénarios de l'inflation

L'inflation future affecte d'une façon directe le coût moyen futur des sinistres (la prime pure) à travers la hausse des prix de réparation et des pièces détachées. Supposons que nous n'anticipons pas l'inflation, alors la marge espérée d'un ensemble de devis  $L$  selon les différents scénarios de l'inflation est donnée comme suit :

$$marge\% = \frac{\sum_{i \in L} demande_i(0\%) * (COPHT_i - \hat{PP}_i * (1 + \epsilon\%))}{\sum_{i \in L} demande_i(0\%) * COPHT_i}$$

La figure (5.4) permet d'observer la marge anticipée en fonction des différents scénarios d'inflation. Nous constatons que l'inflation détériore la rentabilité des affaires nouvelles, un scénario d'inflation de 5%<sup>1</sup> introduit une baisse de la rentabilité de 4 points. En termes du réseau de distribution, la rentabilité du réseau de courtage reste supérieure à celle du réseau des agents pour les différentes catégories du Parc et scénarios d'inflation.

Scénarios d'inflation	AGENT			COURTAGE			Total général
	CIBLE	NON CIBLE	Total	CIBLE	NON CIBLE	Total	
0%	23%	20%	23%	28%	25%	26%	23%
1%	22%	18%	21%	27%	23%	26%	22%
2%	21%	17%	20%	26%	23%	25%	22%
3%	21%	16%	19%	26%	22%	24%	21%
4%	20%	16%	19%	25%	21%	23%	20%
5%	19%	15%	18%	24%	20%	23%	19%
6%	18%	14%	17%	23%	20%	22%	19%
7%	18%	13%	16%	23%	19%	21%	18%
8%	17%	12%	16%	22%	18%	20%	17%
9%	16%	11%	15%	21%	17%	20%	16%

Figure 5.4 – La marge par scénario d'inflation sans ajustement tarifaire

Supposons maintenant que nous anticipons l'inflation et nous ajustons la prime commerciale avec le même taux d'inflation  $\epsilon\%$ . Cette anticipation affecte négativement la fonction de demande puisqu'il s'agit d'une revue à la hausse du tarif, ainsi la marge en pourcentage devient :

$$marge\% = \frac{\sum_{i \in L} demande_i(\epsilon\%) * (COPHT_i * (1 + \epsilon\%) - \hat{PP}_i * (1 + \epsilon\%))}{\sum_{i \in L} demande_i(\epsilon\%) * COPHT_i * (1 + \epsilon\%)}$$

1. En mai 2022, les prix de consommation augmentent de 0,7% en un mois et 5,2% en un an selon l'INSEE.

$$\Leftrightarrow marge\% = \frac{\sum_{i \in L} demande_i(\epsilon\%) * (COPHT_i - \hat{PP}_i)}{\sum_{i \in L} demande_i(\epsilon\%) * COPHT_i}$$

La figure (5.5) donne la marge en fonction des différents scénarios dans le cas de l'anticipation de l'inflation. Le scénario de 5% introduit une baisse d'un point de la marge. En conséquent, cet ajustement a permis de gagner 3 points en termes de rentabilité. Ceci suppose que les stratégies des concurrents sont maintenues constantes. En effet, une anticipation semblable à celles des concurrents permet de conserver la position actuelle du produit Parc dénommé sur le marché. En revanche, une anticipation inférieure à celles des concurrents permet de croître la production du produit Parc dénommée et détériorer la rentabilité et vice versa.

Scénarios d'inflation	AGENT			COURTAGE			Total général
	CIBLE	NON CIBLE	Total	CIBLE	NON CIBLE	Total	
0%	23%	20%	23%	28%	25%	26%	23%
1%	23%	19%	22%	28%	24%	26%	23%
2%	23%	18%	21%	27%	24%	26%	23%
3%	22%	18%	21%	27%	24%	26%	23%
4%	22%	18%	21%	27%	23%	26%	22%
5%	22%	17%	21%	27%	23%	25%	22%
6%	21%	17%	20%	26%	23%	25%	22%
7%	21%	17%	20%	26%	23%	25%	21%
8%	21%	16%	19%	26%	22%	24%	21%
9%	20%	16%	19%	25%	22%	24%	20%

Figure 5.5 – La marge par scénario d'inflation avec ajustement tarifaire

Pour illustrer cela, imaginons que nous disposons de 400 devis et nous ne connaissons pas s'ils sont transformés ou non, la prime commerciale actuelle de chaque devis est égale à 1000€. Nous aimeraisons savoir la rentabilité future de ces devis dans le cas de l'anticipation d'une inflation de 5% avec et sans ajustement tarifaire. Les 400 devis sont équirépartis sur les catégories cibles et non cibles et sur le réseau de distribution. La figure (5.6) donne le nombre de devis transformés et la marge espérée sans et avec ajustement tarifaire de 5%. Nous remarquons que l'ajustement tarifaire a introduit une perte de 28 clients et un gain de marge de 10474,45€.

Le nombre de clients perdus pour le réseau des agents est égal à 15, il est supérieur à celui des courtiers 12 et le gain en termes de marge est presque le même pour les deux réseaux. Ainsi, il paraît intéressant d'effectuer des ajustements segmentés, c'est-à-dire faire des ajustements moins agressifs pour le réseau des agents. Ceci le but de ne pas perdre une part importante de la production et aussi parce que nous préférons avoir une production élevée pour le réseau des agents en raison de la rentabilité à long terme de ce réseau. En effet, les affaires nouvelles du réseau des agents ne sont pas rentables la première année. Cependant, les années suivantes, elles deviennent rentables compte tenu du faible taux de résiliation pour ce réseau. En termes de rentabilité, l'ajustement tarifaire de 5% a introduit la même perte du nombre de clients des catégories cibles et non cibles pour le réseau de courtage. De

		AGENT			COURTAGE			Total général
		CIBLE	NON CIBLE	Total	CIBLE	NON CIBLE	Total	
<b>Sans ajustement</b>	Marge en €	19092,27	14740,42	<b>33832,69</b>	24148,31	20447,96	<b>44596,26</b>	<b>78428,95</b>
	Nombre de devis transformés	53	48	<b>101</b>	36	32	<b>69</b>	<b>170</b>
<b>Avec ajustement</b>	Marge en €	21708,75	17376,90	<b>39085,66</b>	26620,90	23196,81	<b>49817,74</b>	<b>88903,40</b>
	Nombre de devis transformés	46	40	<b>86</b>	30	26	<b>56</b>	<b>142</b>
<b>Différence</b>	Marge en €	2616,48	2636,48	<b>5252,97</b>	2472,59	2748,85	<b>5221,48</b>	<b>10474,45</b>
	Nombre de devis transformés	-7	-8	<b>-15</b>	-6	-6	<b>-12</b>	<b>-28</b>

Figure 5.6 – Scénarios d’inflation sans et avec ajustement tarifaire

ce fait, un ajustement segmenté selon les catégories cibles et non cibles pour ce réseau va permettre de ne pas avoir plusieurs affaires non cibles dans le portefeuille. Toutefois, pour le réseau des agents, l’ajustement tarifaire de 5% a introduit une perte de 7 clients cibles et de 8 clients non cibles. Ceci représente une motivation de faire un ajustement plus agressif pour les non cibles afin de pouvoir diminuer leur nombre dans le portefeuille.

## Conclusion partielle

La fonction de demande et la marge représentent des outils de pilotage importants pour anticiper les variations de la production et de la rentabilité des affaires nouvelles suite à des stratégies de changement tarifaire. Elles sont calculées à partir du modèle de taux de transformation et de la fonction de l’élasticité au prix. Une production élevée permet d’avoir un grand nombre d’affaires nouvelles avec une rentabilité inférieure et inversement. Pour une bonne prise de décision, il est important de croiser ces deux quantités et choisir la position souhaitée entre les deux. Une revue à la hausse de la stratégie tarifaire pour compenser l’inflation doit prendre en considération la position choisie entre les deux quantités par réseau de distribution et pour les segments cibles et non cibles.

# Conclusion générale

Arrivant à la fin de cette étude, dont l'objectif est de proposer une méthode robuste pour estimer l'élasticité au prix afin de pouvoir anticiper la demande et la rentabilité des affaires nouvelles du produit Parc dénommé, nous pouvons récapituler l'étude en quatre étapes :

- Étape 1 : dans un premier temps, un soin particulier a été consacré à la construction de la base de données, à la correction de toutes les anomalies et à l'exploration des données. Le but de cette étape a été de connaître les propriétés et les limites des données. L'exploration des données a montré que la prime a un impact important sur le taux de transformation, une tendance décroissante du taux de transformation à mesure que la prime augmente. Toutefois, le taux de transformation décroît selon les segments du coefficient technique.
- Étape 2 : le but de cette étape a été de construire un modèle avec de bonnes qualités prédictives pour le taux de transformation afin d'estimer la demande pour la stratégie tarifaire actuelle. Deux types de modèles ont été challengés : le modèle GLM et le modèle Xgboost qui a été implémenté pour améliorer les performances prédictives et interprété par la théorie de Shap values et les graphiques de dépendances partielles (PDP). Le modèle GLM a fait preuve de faible capacité de généralisation des résultats de prédiction. En revanche, le modèle Xgboost malgré sa bonne performance prédictive, présente une limite non négligeable due à sa complexité le rendant non opérationnel pour la prédiction du taux de transformation.
- Étape 3 : les données historiques de l'assurance ne permettent pas d'avoir une estimation robuste de l'élasticité directement à partir du modèle GLM, pour remédier à ce problème, nous avons proposé une méthodologie inspirée de la littérature de l'inférence causale. L'avantage de la méthodologie proposée est qu'elle permet d'avoir une formule opérationnelle de l'élasticité au prix en combinant le critère opérationnel du modèle GLM et la capacité du modèle Xgboost de capter les associations non linéaires. L'analyse de l'élasticité par profil de risque a permis de caractériser les profils les plus élastiques. Les profils cibles sont moins élastiques puisqu'ils bénéficient de rabais importants. Le réseau de courtage est plus élastique en raison de l'offre variée proposée par ce réseau. La région Île-de-France représente la région la plus élastique

compte tenu de la prédominance du réseau de courtage dans cette région.

- Étape 4 : la fonction de demande et la marge représentent des outils de pilotage importants pour anticiper les variations de la production et de la rentabilité des affaires nouvelles suite à des stratégies de changement tarifaire. Elles ont été calculées à partir du modèle GLM du taux de transformation et de la fonction de l'élasticité au prix. Une production élevée permet d'avoir un grand nombre d'affaires nouvelles avec une rentabilité inférieure et inversement. Pour une bonne prise de décision, il est important de croiser ces deux quantités et choisir la position souhaitée entre les deux.

Les résultats obtenus dans le cadre de cette étude sont encourageants : l'élasticité au prix calculée est en adéquation avec la réalité du marché de l'assurance des flottes automobiles d'entreprise. Il peut être envisagé, en guise d'ouverture, d'effectuer une optimisation tarifaire pour calculer la prime commerciale optimale à partir de la fonction de demande et la mesure de rentabilité. Cette optimisation peut aussi prendre en considération d'autres indicateurs de rentabilité du long terme comme la valeur client ou la valeur contrat.

# Liste des acronymes

<b>ACC</b>	ACCuracy (Taux des bonnes prédictions)
<b>AFDM</b>	Analyse Factorielle de Données Mixtes
<b>AUC</b>	Area Under the Curve (L'aire sous la courbe ROC)
<b>ATE</b>	Average Treatment Effect (Effet moyen du traitement)
<b>ASMD</b>	Absolute Standardized Mean Difference (Différence absolue des moyennes standardisées)
<b>CA</b>	Chiffre d’Affaires
<b>CART</b>	Classification and Regression Tree
<b>COPHT</b>	Cotisation Potentielle Hors Taxes
<b>CT</b>	Coefficient Technique
<b>ID</b>	IDentifiant
<b>IARD</b>	Incendie, Accidents, Risques Divers
<b>IPTW</b>	Inverse Probability of Treatment Weighting (Pondération sur le score de propension)
<b>GLM</b>	Generalized Linear Model (Modèle linéaire généralisé)
<b>NAF</b>	Nomenclature d’Activités Française
<b>PDP</b>	Partial Dependence Plot (Graphique de dépendance partielle)
<b>PP</b>	Prime Pure
<b>PS</b>	Propensity Score (Score de propension)
<b>Tt</b>	Taux de transformation
<b>RC</b>	Responsabilité Civile
<b>RFE</b>	Recursive Feature Elimination
<b>ROC</b>	Receiver Operating Characteristic
<b>Xgboost</b>	eXtreme Gradient Boosting

# Table des figures

1.1	Ratio sinistre à prime du marché français de flotte d'entreprise. Source : France Assureurs . . . . .	5
1.2	Poids en CA en portefeuille AXA IARD Entreprise en 2020 . . . . .	6
1.3	Antisélection : marché monopolistique . . . . .	9
1.4	Antisélection : marché avec deux concurrents . . . . .	10
1.5	Antisélection : marché avec trois concurrents . . . . .	11
1.6	Inflation et risque d'antisélection . . . . .	12
2.1	Agrégation des bases de données . . . . .	15
2.2	Processus de transformation des variables de la base véhicule . . . . .	18
2.3	Activité d'entreprise . . . . .	19
2.4	Taux de transformation par segment de prime . . . . .	21
2.5	Taux de transformation par coefficient technique . . . . .	21
2.6	Densité coefficient technique . . . . .	21
2.7	Taux de transformation par réseau . . . . .	22
2.8	Taux de transformation par région . . . . .	22
2.9	Taux de transformation par année d'effet . . . . .	23
2.10	Corrélation entre les variables numériques . . . . .	24
2.11	Association entre les variables catégorielles . . . . .	25
2.12	Contribution des variables explicatives aux axes de l'AFDM . . . . .	28
2.13	Contribution de chaque axe à l'inertie totale . . . . .	28
3.1	Exemple d'un arbre de décision CART . . . . .	34
3.2	Courbe ROC . . . . .	38
3.3	L'algorithme RFE . . . . .	38
3.4	Exemple simplifié du calcul du PDP, source : [Ali20] . . . . .	40
3.5	Chemin de régularisation du Lasso . . . . .	44

3.6	Lurning Curve (AUC) . . . . .	45
3.7	Lurning Curve (1-ACC) . . . . .	45
3.8	Sélection de variables (GLM) . . . . .	46
3.9	Sélection de variables (Xgboost) . . . . .	46
3.10	Courbe ROC . . . . .	46
3.11	Évaluation des modèles . . . . .	46
3.12	Influence des variables du modèle GLM . . . . .	47
3.13	PDP de la prime moyenne (GLM) . . . . .	47
3.14	PDP du coefficient technique (GLM) . . . . .	47
3.15	Importance des variables du modèle Xgboost . . . . .	48
3.16	Influence des variables du modèle Xgboost . . . . .	49
3.17	PDP de la prime moyenne (Xgboost) . . . . .	49
3.18	PDP du coefficient technique (Xgboost) . . . . .	49
3.19	Prédiction du taux de transformation selon la prime moyenne . . . . .	50
3.20	Prédiction du taux de transformation selon le coefficient technique . . . . .	50
3.21	Prédiction du taux de transformation selon la catégorie de Parc . . . . .	51
3.22	Prédiction du taux de transformation selon le réseau de distribution . . . . .	51
3.23	Prédiction du taux de transformation par région . . . . .	51
4.1	Biais de confusion . . . . .	55
4.2	Pondération sur le score de propension . . . . .	60
4.3	Nombre de classes optimal du K-means . . . . .	64
4.4	Distribution du taux de rabais . . . . .	64
4.5	Répartition des observations entre les classes du taux de rabais . . . . .	64
4.6	Distribution de la prime moyenne par taux de rabais . . . . .	65
4.7	Distribution du coefficient de sinistralité par taux de rabais . . . . .	65
4.8	Distribution de la variable âge de véhicule égal à 1 an par taux de rabais . . . . .	65
4.9	Distribution du taux de rabais par catégorie de Parc . . . . .	65
4.10	Distribution du taux de rabais par formule du Parc . . . . .	66
4.11	Distribution du taux de rabais par réseau de distribution . . . . .	66
4.12	Learning Curve classe A . . . . .	68
4.13	Learning Curve classe E . . . . .	68
4.14	Équilibre entre A et C . . . . .	69
4.15	Équilibre entre A et E . . . . .	69

4.16 Overlap classe B . . . . .	70
4.17 Overlap classe C . . . . .	70
4.18 Courbe ATE . . . . .	72
4.19 L'élasticité par taux de rabais . . . . .	74
4.20 L'élasticité par prime . . . . .	74
4.21 L'élasticité par distributeur . . . . .	75
4.22 L'élasticité par catégorie du Parc . . . . .	75
4.23 L'élasticité par région . . . . .	75
4.24 L'élasticité par année d'effet . . . . .	76
5.1 La fonction de demande par variation de prime . . . . .	78
5.2 Densité de la marge en pourcentage . . . . .	80
5.3 La marge par variation de prime . . . . .	80
5.4 La marge par scénario d'inflation sans ajustement tarifaire . . . . .	81
5.5 La marge par scénario d'inflation avec ajustement tarifaire . . . . .	82
5.6 Scénarios d'inflation sans et avec ajustement tarifaire . . . . .	83

# **Annexes**

# Annexe 1 : Notions mathématiques

## A. La régression logistique multinomiale

La régression logistique multinomiale est un cas particulier de la régression logistique dans le cas où le nombre de modalités de la variable d'intérêt  $y$  est supérieur strictement à 2. Les modalités peuvent être nominales comme la région ou ordinaires comme la satisfaction client (très satisfait, satisfait, peu satisfait, pas du tout satisfait). Dans ce cas, la variable  $Y$  est supposée suivre une distribution multinomiale qui est une généralisation de la loi de Bernoulli. Soit  $K$  le nombre de modalités de la variable  $Y$ . Nous prenons comme modalité de référence la modalité  $y_K$ , alors la probabilité de survenance de l'évènement  $y_k$  pour  $k \in \{1, 2, ..K - 1\}$  est donnée par :

$$g_k(X) = \log\left(\frac{\pi_k(X)}{\pi_K(X)}\right) = \beta_0^k + \sum_{i=1}^p \beta_i^k x_i$$

Avec :

—  $\pi_k(X) = P(Y = y_k | x_1, \dots, x_p)$  pour  $k \in \{1, 2, ..K\}$  la probabilité de survenance de l'évènement  $k$  conditionnellement aux variables explicatives ;

— La relation entre  $\pi_k(X)$  et  $g_k(X)$  est donnée comme suit :  $\pi_k(X) = \frac{e^{g_k(X)}}{1 + \sum_{k=1}^{K-1} e^{g_k(X)}}$   $0 < k < K$ .

Pour chaque probabilité de survenance, nous devons estimer  $p + 1$  paramètres. C'est-à-dire, le modèle de la régression multinomiale estime  $(K - 1) * (P + 1)$  paramètres. Ceci est fait à travers la maximisation de la fonction de vraisemblance suivante :

$$L(y_1, y_2, \dots, y_n, \beta^1, \beta^2 \dots \beta^K) = \prod_{i=1}^n \pi_1^{1_{y_i=1}}(X_i) \pi_2^{1_{y_i=2}}(X_i) \dots \pi_K^{1_{y_i=K}}(X_i)$$

## B. L'algorithme K-means

Soit un échantillon de  $n$  observations et  $p$  variables et  $k$  le nombre de classes que l'on souhaite construire, l'algorithme K-means peut être présenté comme suit :

1. Étape 1 : choix aléatoire de centres initiaux  $c_1, c_2, \dots, c_k$ .
2. Étape 2 : repartir les observations sur les  $k$  classes en prenant en considération la distance minimale entre les observations et les centres des classes.
3. Étape 3 : calculer les bycentres des groupes construits.
4. Étape 4 : si les bycentres restent les mêmes alors sortir les groupes sinon refaire les étapes 2,3 et 4.

## Annexe 2 : Sorties logiciel R

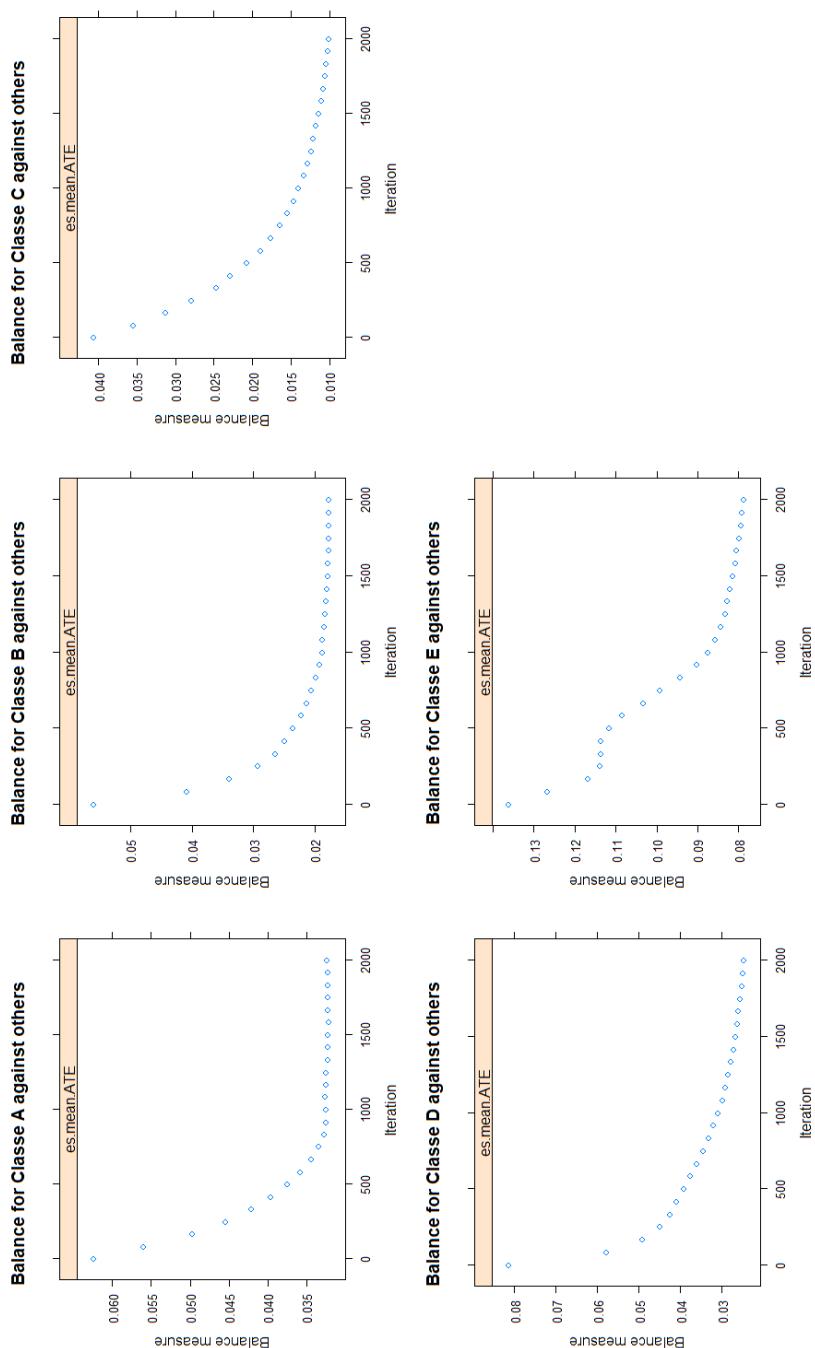


Figure 1 : Learning Curve

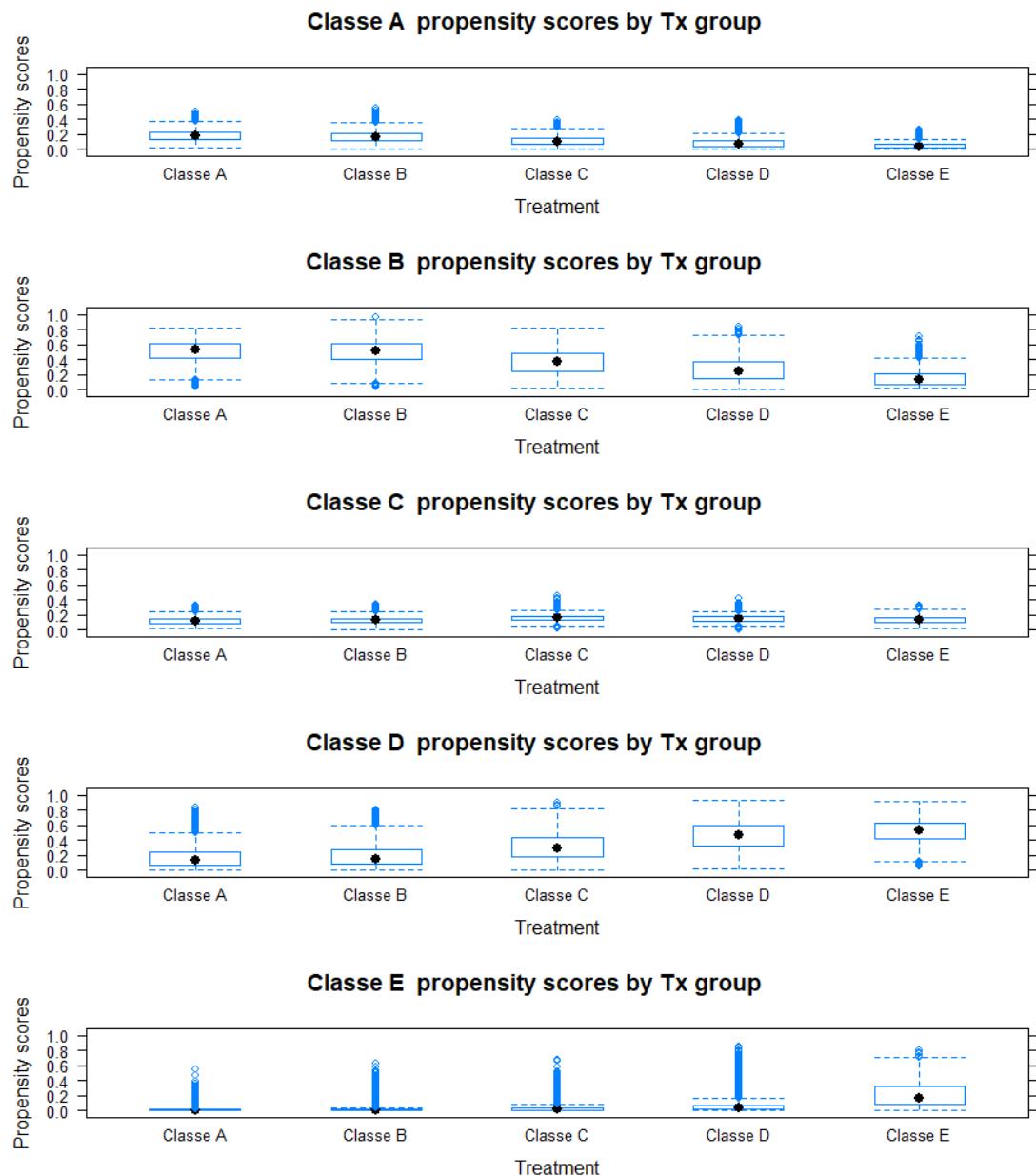


Figure 2 : Vérification de l'hypothèse de la positivité

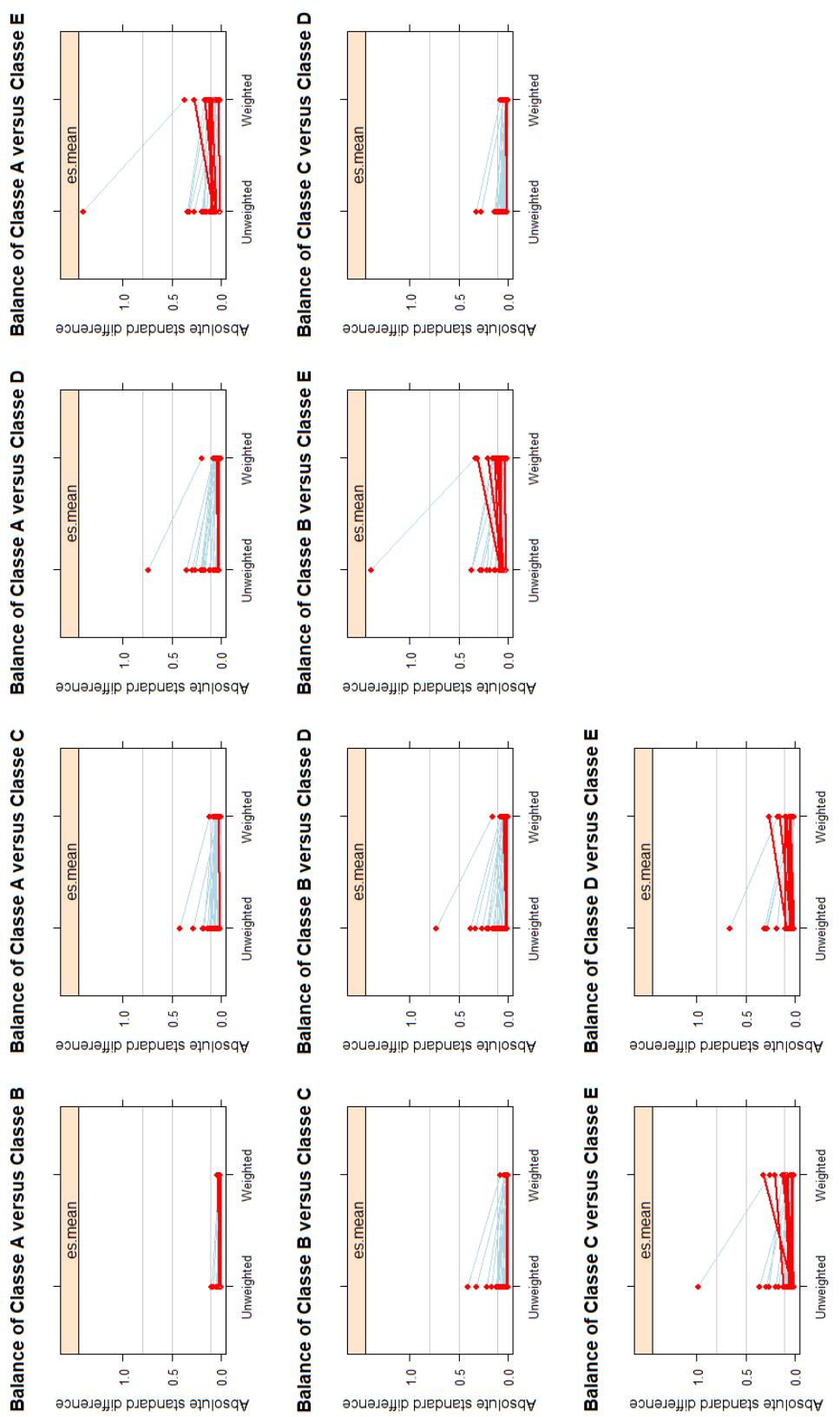


Figure 3 : Évaluation d'équilibre des covariables

Variables	Model 1
(Constante )	-0.22 * (0.09)
Coefficient technique	-8.98 (5.50)
Coefficient technique^2	28.32 *** (5.26)
Coefficient technique^3	-28.69 *** (4.86)
Coefficient de simstralité	-0.01 (0.01)
Véhicule utilitaire	-0.13 *** (0.03)
Véhicule léger	-0.10 ** (0.03)
Garantie dommage	-0.02 (0.02)
Financement par crédit	-0.00 (0.03)
Véhicule de qualité basse	-0.03 (0.03)
Véhicule de luxe	-0.03 (0.02)
Véhicule d'âge égal à 1 an	0.17 *** (0.03)
Véhicule d'âge égal à 2 ans	-0.00 (0.02)
Véhicule d'âge égal à 3 ans	0.01 (0.02)
Véhicule d'âge égal à 3 ans	0.01 (0.02)
Véhicule d'âge égal à 5 ans	0.00 (0.02)
Véhicule d'âge égal à 6 ans	0.03 (0.02)
Véhicule d'âge égal à 7-8 ans	-0.00 (0.02)
Ancienneté de devis	0.24 *** (0.02)
La prime avant rabais	-0.23 *** (0.03)

Figure 4 : Résultats modèle global

		0.37 ***
		(0.11)
Formule choisie (PARC_MINI)		0.15 *
		(0.06)
Formule choisie (PARC_OPTIMALE)		0.63 ***
		(0.07)
Parc non cible		-0.73 ***
		(0.04)
Courtier		-0.29 ***
		(0.04)
Taille de flotte (10-14 véh.)		-0.36 ***
		(0.05)
Taille de flotte (15-19 véh.)		-0.55 ***
		(0.05)
Taille de flotte (sup à 20 véh.)		-0.05
		(0.06)
Type de distributeur (note 2)		0.10 **
		(0.03)
Type de distributeur (note 3)		1.12 ***
		(0.09)
Type de distributeur (note 1)		-0.13 **
		(0.04)
Densité population (tranche 2)		-0.13 **
		(0.05)
Densité population (tranche 3)		-0.19 *
		(0.07)
Densité population (tranche 4)		-0.08
		(0.06)
Région (65)		0.00
		(0.06)
Région (66)		-0.12 *
		(0.06)
Région (67)		-0.04
		(0.06)
Région (68)		0.43 ***
		(0.06)
Type d'activité (2)		0.53 ***
		(0.05)
Type d'activité (3)		65949
		All continuous predictors are mean-centered and scaled by 1 standard deviation. *** p < 0.001; ** p < 0.01; * p < 0.05.

Figure 5 : Résultats modèle global

# Bibliographie

- [Ala17] Thomas Lumley et Alastair Scott. *Fitting Regression Models to Survey Data*. <<https://projecteuclid.org/journals/statistical-science/volume-32/issue-2/Fitting-Regression-Models-to-Survey-Data/10.1214/16-STS605.full>>. [Online ; accessed 13-August-2022]. 2017.
- [Ali20] Dimitri Delcaillau; Antoine Ly; Franck Vermet et Alizé Papp. *Interprétabilité des modèles : état des lieux des méthodes et application à l'assurance*. <<https://arxiv.org/pdf/2007.12919.pdf>>. [Online ; accessed 03-September-2022]. 2020.
- [Bab00] James M. Robins ; Miguel Angel Hernan et Babette Brumback. *Marginal Structural Models and Causal Inference in Epidemiology*. <[https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal\\_Structural\\_Models\\_and\\_Causal\\_Inference\\_in.11.aspx#JCL1-1](https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal_Structural_Models_and_Causal_Inference_in.11.aspx#JCL1-1)>. [Online ; accessed 15-August-2022]. 2000.
- [D B83] P. R. ROSENBAUM et D. B. RUBIN. *The central role of the propensity score in observational studies for causal effects*, *Biometrika*, Volume 70, Issue 1, April 1983, Pages 41–55. <<https://academic.oup.com/biomet/article/70/1/41/240879>>. [Online ; accessed 01-August-2022]. 1983.
- [ESC79] B. ESCOFIER. *Traitemen simultané de variables qualitatives et quantitatives en analyse factorielle*. <[http://archive.numdam.org/item/CAD\\_1979\\_\\_4\\_2\\_137\\_0.pdf](http://archive.numdam.org/item/CAD_1979__4_2_137_0.pdf)>. [Online ; accessed 24-August-2022]. 1979.
- [Fri01] Jerome H. Friedman. *Greedy function approximation : A gradient boosting machine*. <<https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full>>. [Online ; accessed 03-September-2022]. 2001.
- [FUL09] W. A. FULLER. *Sampling Statistics*. Wiley, Hoboken, NJ., 2009.

- [GUI01] KEISUKE HIRANO et GUIDO W. IMBENS. *Estimation of Causal Effects using Propensity Score Weighting : An Application to Data on Right Heart Catheterization*. Health Services and Outcomes Research Methodology 2 :259–278, 2001. <[https://oconnell.fas.harvard.edu/files/imbens/files/estimation\\_of\\_causal\\_effects\\_using\\_propensity\\_score\\_weighting\\_an\\_application\\_to\\_data\\_on\\_right\\_hear\\_catherization.pdf](https://oconnell.fas.harvard.edu/files/imbens/files/estimation_of_causal_effects_using_propensity_score_weighting_an_application_to_data_on_right_hear_catherization.pdf)>. [Online ; accessed 07-August-2022]. 2001.
- [Hol86] Paul W. Holland. *Statistics and Causal Inference*. Published by : Taylor and Francis, Ltd. on behalf of the American Statistical Association <<http://www.jstor.org/stable/2289064>>. [Online ; accessed 10-August-2022]. 1986.
- [Imb00] Guido W. Imbens. *The Role of the Propensity Score in Estimating Dose-Response Functions, Biometrika Vol. 87, No. 3 (Sep., 2000), pp. 706-710 (5 pages)*). Oxford University Press, 2000.
- [Su-17] Scott M. Lundberg ; Gabriel G. Erion et Su-In Lee. *Consistent Individualized Feature Attribution for Tree Ensembles*. <<https://arxiv.org/abs/1802.03888>>. [Online ; accessed 31-August-2022]. 2017.
- [J P08] B. Escofier et J. Pagès. *Analyses factorielles simples et multiples*. Paris : Dunod (ISBN 978-2-10-051932-3), 2008.
- [Jam20] Heejung Bang et James M. Robins. *Doubly Robust Estimation in Missing Data and Causal Inference Models, Biometrics Vol. 61, No. 4 (Dec., 2005), pp. 962-972 (11 pages)*. <<https://www.jstor.org/stable/3695907>>. [Online ; accessed 01-August-2022]. 2020.
- [Jos76] Michael Rothschild et Joseph Stiglitz. *Equilibrium in Competitive Insurance Markets : An Essay on the Economics of Imperfect Information, The Quarterly Journal of Economics, Vol. 90, No. 4 (Nov., 1976), pp. 629-649*. by Oxford University Press : <<https://www.jstor.org/stable/1885326>>. [Online ; accessed 02-August-2022]. 1976.
- [Kle10] Peter Kleist. *Les biais dans les études d'observation*. <[https://swissethics.ch/assets/Fortbildung/Publikationen/kleist\\_p\\_bias\\_in\\_beobachtungsstudien\\_2010\\_f.pdf](https://swissethics.ch/assets/Fortbildung/Publikationen/kleist_p_bias_in_beobachtungsstudien_2010_f.pdf)>. [Online ; accessed 01-August-2022]. 2010.
- [Lan13] Daniel F. McCaffrey ; Beth Ann Griffin ; Daniel Almirall ; Mary Ellen Slaughter ; Rajeev Ramchandb et Lane F. Burgette. *Statistics in medecine, A tutorial on propensity score estimation for multiple treatments using generalized boosted models*. <<https://api.istex.fr/ark:/67375/WNG-7PH25F9M-S/fulltext.pdf?sid=focus>>. [Online ; accessed 01-August-2022]. 2013.

- [Li18] Fan Li. *PROPENSITY SCORE WEIGHTING FOR CAUSAL INFERENCE WITH MULTIPLE TREATMENTS*. Yale University and Duke University : <<https://arxiv.org/pdf/1808.05339.pdf>>. [Online ; accessed 04-August-2022]. 2018.
- [Mar10] Shenyang Y. Guo et Mark W. Fraser. *Propensity Score Analysis : Statistical Methods and Applications (Advanced Quantitative Techniques in the Social Sciences)*. Sage Publishing, 2010.
- [Mon13] Leo Guelman et Montserrat Guillén. *Expert Systems with Applications, A causal inference approach to measure price elasticity in Automobile Insurance*. <<https://www.sciencedirect.com/science/article/abs/pii/S0957417413005460?via%3Dihub>>. [Online ; accessed 01-August-2022]. 2013.
- [QUE18] Simon QUENTIN. *Méthodologie statistique, Estimation avec le score de propension sous R*. <<https://www.insee.fr/fr/statistiques/3546202>>. [Online ; accessed 01-August-2022]. 2018.
- [Rak17] Ricco Rakotomalala. *Pratique de la Régression Logistique Régression Logistique Binaire et Polytomique*. <[http://www.ressources-actuarielles.net/EXT/ISFA/fpisfa.nsf/1bebb4baec15bba8c12580e90064b202/69dec6b0bcef0009c1257f990073b7cf\(FILE/pratique\\_regression\\_logistique.pdf](http://www.ressources-actuarielles.net/EXT/ISFA/fpisfa.nsf/1bebb4baec15bba8c12580e90064b202/69dec6b0bcef0009c1257f990073b7cf(FILE/pratique_regression_logistique.pdf)>. [Online ; accessed 28-August-2022]. 2017.
- [Ric19] Gary King et Richard Nielsen. *Why Propensity Scores Should Not Be Used for Matching*. Political Analysis, 27, 4, Pp. 435-454. <<https://tinyurl.com/y5b5yjxo>>. [Online ; accessed 06-August-2022]. 2019.
- [RUB74] DONALD B. RUBIN. *ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES*. Published by : Journal of Educational Psychology 1974, Vol. 66, No. 5, 688-701 <[http://www.fsb.muohio.edu/lij14/420\\_paper\\_Rubin74.pdf](http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf)>. [Online ; accessed 10-August-2022]. 1974.
- [Sch07] Kang JDY et Schafer JL. *Demystifying double robustness : a comparison of alternative strategies for estimating a population mean from incomplete data*. Statistical Science 2007; 22(4) :523 –539. <<https://arxiv.org/pdf/0804.2958.pdf>>. [Online ; accessed 13-August-2022]. 2007.
- [Sha52] Lloyd S. Shapley. *A Value for N-Person Games*. <<https://www.rand.org/pubs/papers/P295.html>>. [Online ; accessed 29-August-2022]. 1952.

- [Tia16] Carlos Guestrin et Tianqi Chen. *XGBoost : A Scalable Tree Boosting System*. <<https://www.math.mcgill.ca/dstephens/PSMMA/Articles/HIrano-Imbens-2004.pdf>>. [Online ; accessed 26-August-2022]. 2016.
- [V V02] Guyon I; J Weston; S Barnhill et V Vapnik. *Gene Selection for Cancer Classification using Support Vector Machines*. <<https://link.springer.com/content/pdf/10.1023/A:1012487302797.pdf>>. [Online ; accessed 29-August-2022]. 2002.
- [Wik] WikiStat. *Introduction au modèle linéaire généralisé*. <<https://www.math.univ-toulouse.fr>>. [Online ; accessed 03-September-2022].