**Introduction**

La data science ouvrent aux entreprises un champ des possibles inédit. Elle confère aujourd'hui aux entreprises qui l’utilisent un avantage compétitif. Les cas d’usages de la data science au sein des entreprises sont légions. On peut en citer comme exemple l’amélioration de la connaissance client, la détection de la fraude, la maintenance préventive.

Définit comme la science des données, la data science est un mélange disciplinaire entre la data inférence, le développement d’algorithme et la technologie, dont l’objectif est la résolution de problèmes analytiques complexes. En effet, la data science va utiliser les données à sa disposition, de façon créative, pour générer et apporter des réponses aux problématiques des entreprises. L’une des problématiques courantes au sein des entreprises qui est résolue par le biais de la data science est l’estimation du bénéfice net annuel attendu par client. Au regard de l’importance de que revêt le bénéfice net annuel pour la valorisation des entreprises et le calcul de leurs rentabilités, une question majeure subsiste : Comment estimer le bénéfice net annuel attendu par client sur un contrat d’assurance automobile en fonction des caractéristiques du client ?

Pour répondre à cette préoccupation, il sera question dans ce projet de construire un modèle de prédiction du bénéfice net annuel attendu par client sur un contrat d’assurance automobile à partir des caractéristiques du client.

Pour atteindre cet objectif , la méthodologie adoptée dans le cadre de ce projet sera la méthodologie CRISP-DM. Cette démarche se décompose en 6 étapes allant de la compréhension du problème métier au déploiement et la mise en production. Ces étapes sont les suivantes :

**La compréhension du problème métier** : cette première étape consiste à bien comprendre les éléments métiers et problématiques que l’on vise à résoudre ou à améliorer.

**La compréhension des données :** Cette phase vise à déterminer précisément les données à analyser, à identifier la qualité des données disponibles et à faire le lien entre les données et leur signification d’un point de vue métier.

**La construction du Data Hub :** Cette phase de préparation des données regroupe les activités liées à la construction de l’ensemble précis des données à analyser, faite à partir des données brutes. Elle inclut ainsi le classement des données en fonction de critères choisis, le nettoyage des données, et surtout leur recodage pour les rendre compatibles avec les algorithmes qui seront utilisés.

**La modélisation :** cette étape de modélisation comprend le choix, le paramétrage et le test de différents algorithmes ainsi que leur enchaînement, qui constitue un modèle.

**L’évaluation :** L’évaluation vise à vérifier le(s) modèle(s) ou les connaissances obtenues afin de s’assurer qu’ils répondent aux objectifs formulés au début du processus. Elle contribue aussi à la décision de déploiement du modèle ou, si besoin est, à son amélioration. A ce stade, on teste notamment la robustesse et la précision des modèles obtenus.

**Le déploiement :** Elle consiste en une mise en production pour les utilisateurs finaux des modèles obtenus. Son objectif : mettre la connaissance obtenue par la modélisation, dans une forme adaptée, et l’intégrer au processus de prise de décision.

1. **La compréhension du problème métier**

**Le marché de l’assurance automobile est l’un des marchés des plus importants pour les assureurs. Selon la federation francaise de l’assurance FFA**

**(**[**https://www.ffa-assurance.fr/etudes-et-chiffres-cles?f%5B0%5D=field\_categorie\_chiffre\_cle%253Aparents\_all%3A43**](https://www.ffa-assurance.fr/etudes-et-chiffres-cles?f%5B0%5D=field_categorie_chiffre_cle%253Aparents_all%3A43)**)**

**en 2015, l**e chiffre d’affaires de l’assurance automobile représente 39 % de l’ensemble des cotisations des assurances de dommages aux biens et de responsabilité civile en france. L’automobile des particuliers, avec un chiffre d’affaires de 18,5 milliards d’euros en 2015, représente 56 % des assurances souscrites par les particuliers en IARD (incendies, accidents et risques divers) et progresse de + 1,7 %.

Pour augmenter leur chiffre d’affaire et mieux piloter leur rentabilités , les compagnies d’assurances ont bésoin d’ameliorer leurs connaissances clients et d’avoir une idée du bénefice net annuel attendu pour chaque client. La data science apparait comme un outil salvateur qui permet de satisfaire ces bésoins. En effet, une analyse approprié des données dont dispose les compagnies d’assurances sur leur clientèle permettra d’ameliorer la connaissance Client.

La réalisation de ce projet s’inscris dans ce cadre. Le pojet est appliqué uniquement dans le domaine de l’assurance Auto. La question centrale à laquelle il tente de répondre est la suivante : Comment prédire le bénéfice net annuel attendu par client sur un contrat d’assurance automobile à partir des caractéristiques du client.

L’objectif principal de ce projet est d’estimer le bénéfice net annuel attendu par client sur un contrat d’assurance automobile en fonction des caractéristiques du client.

De cet objectif principal découlent les objectifs spécifiques suivant :

- Construction de modèle de prédiction du beneficie net annuel attendu par client

- Choix des meilleurs modèles de prédiction

- Evaluation et compréhension des résultats obtenus - Explications des résultats obtenus

1. LA COMPREHENSION DES DONNEES

Pour réaliser cette étude nous disposons de deux Data sets. Un data set pour l’entrainement du modèle et un second pour la validation du modèle. Le data set d’entrainement est constitué de 1000 observations et de 13 variables. Parmi les 13 variables il y’a 12 variables explicatives et une variable expliquée qui le bénéfice net annuel. Le data set pour la prédiction est quant à lui constitué de 300 observations et de 12 variables. Il contient l’ensemble des variables figurant dans le data set d’entrainement excepté la variable bénéfice net annuel. Le tableau nous donné une description des différentes variables et de leur types.

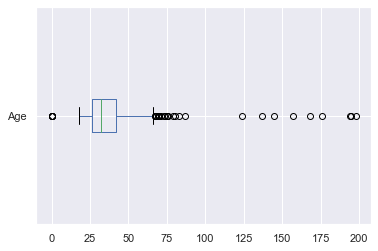
|  |  |
| --- | --- |
| Type de Variable | Noms des varaibles |
| Alphanumeriques (3) | Marque, Categorie socio professionnelle, Type de Vehicule |
| Numeriques  (10) | Age,Prime mensuelle,Kilometres parcourus par mois, Coefficient bonus malus, Score CRM, Niveau de Vie, Salaire annuel, Score credit, Cout entretien annuel,Beneficie net annuel |

Ainsi une analyse descriptive de ses données nous permet d’obtenir le tableau ci-dessous

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Prime mensuelle | Categorie  socio professionnelle | Kilometres parcourus par mois | Coefficient bonus malus | | | Type de Vehicule | Score CRM | | Niveau de Vie | | Marque | | Salaire annuel | Score credit | Cout entretien annuel | Beneficie net annuel |
| Total effectif | 983 | 989 | 1000 | 1000 | | 1000 | 1000 | | | 1000 | | 1000 | | 948 | 1000 | 1000 | 1000 | 1000 |
| Moyenne/ Top | 35,95 | 18,75 | CADRE | 605,675 | | 100 | 5 portes | | | 134,72 | | 4207,239 | | Peugeat | 29553,8 | 493,93 | 792,99 | 16,975 |
| Ecart type | 17,98 | 13,16 | \*\*\*\*\*\*\*\*\*\*\* | 232,39 | | 14,85 | \*\*\*\*\*\*\*\*\* | | | 20,38 | | 2665,07 | | \*\*\*\*\*\*\* | 18054,9 | 288,95 | 99,42 | 19,43 |
| NA (Non renseigné) | 17 | 11 | 1000 | 1000 | | 1000 | 1000 | | | 1000 | | 1000 | | 52 | 1000 | 1000 | 1000 | 1000 |

Comme l’indique l’indique le tableau , il existe des données manquantes au niveau des variables Age, Prime mensuelle.

L’analyse détaillée de cette variable à mis en évidence la presence de données atypiques . C’est le cas de la variable Age,



Il servira de base d’entrainement pour le modèle. une pour la compréhion des

Dans la suite de notre analyse, nous avons procéder a une recodage des données alphabétiques

Elle consista à recoder les modalités avec des nombres de telle sorte que toute la base de données devient de types numériques.

Pour ce faire, j’ai codé la fonction function\_nettoyage qui prend en entrée la base de données et les paramètres (caractéristiques de tendances centrales « moyenne, mode ») de chaque variable.

Les mentions au sein de chaque variable alphanumérique sont récupérées ordonnées et numérotées et stockées de sortes à séries à remplacer ses codes avec ses codes.

{"Peugeot": 1, "Renault": 2, "Toyota": 3, "Opel": 4,"Citroen":5,"Volkswagen":6 }

{"Etudiant": 1, "Ouvrier": 2, "Cadre": 3, "Sans emploi": 4,"Travailleur non salarie" :5}

{"SUV": 1, "5 portes": 2, "3 portes": 3, "Utilitaire": 4}

**Correction des données.**

Afin de corriger les données afin d’être prise en compte dans la modélisation, j’ai préfère utilisée une méthode d’imputation de données manquantes avec une mémoire de ses valeurs. Il faut noter qu’il existe plusieurs méthodes cependant l’une les plus robustes que nous aurions aimerez implémenter était l’Algorithme EM (Esperance-maximisation).

Les données manquantes numeriques ont été remplacer par la moyenne , et celle qui était au depart alphanimeriques ont été remplacer par la valeur du mode de celle-ci.

**Les Tests Statistiques**

1. **Coefficient de correlation lineaire**

Pour les variables initialement numériques, nous avons déterminer le niveau de corrélation entre les varaibles y compris la variable Benefice annuel . Ainsi le résultat obtenu la figure ci-dessous

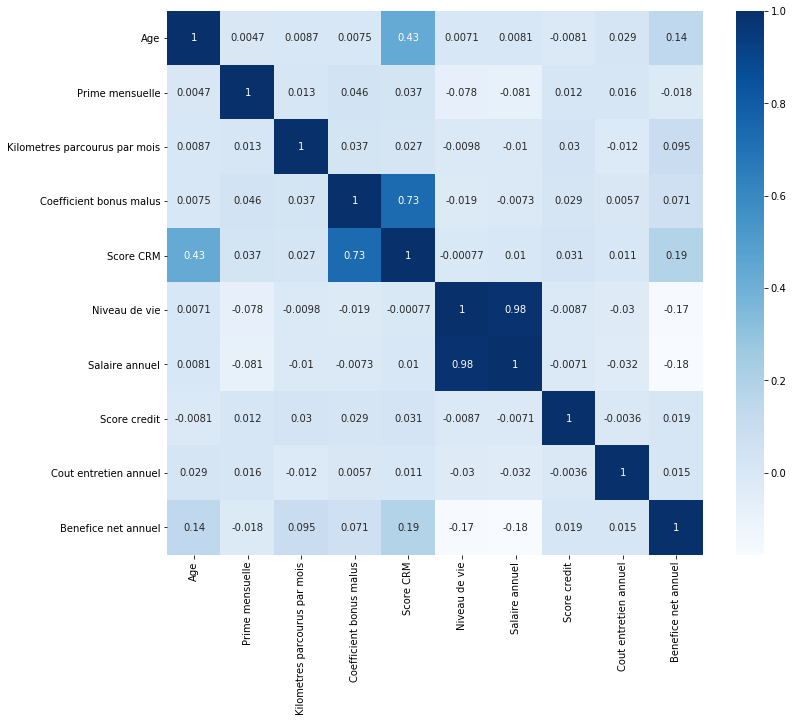


Figure : Coefficient de corrélation entre les variables numériques

Ainsi un certain nombre de variable semble fortement corrèle avec le cas du **Score CRM au Coefficient bonus malus** (0.71)**.**

**Le Salaire annuelle et le niveau de vie** semble être également très corrélé (0.98)

**Conclusion 1 : Pour l’analyse du Benefice net annuel, nous pouvons ainsi privilégier la variable Score CRM idem pour la variable salaire annuelle et le niveau de vie le choix peut être porter sur la variable salaire annuelle ainsi nous pouvons réduire le choix de variable explicatif de 12 à 10 varaibles pour les modèles .**

1. **Tests statistiques**

Pour analyser l’impact de chaque variable y compris les variables alphabétiques recodées, nous utiliserons un test statistique.

Etant donné que nous ignorerons si la variable respecte les conditions de normalité (si la variable suit une loi normale), nous préférons utiliser un test statistique non paramétrique celui du kruskal-Wallis avec ici l’hypothèse nulle étant le faite qu’il n’ait pas de différence significatif entre deux variables étudiés .

Ici cette analyse permet ainsi mettre la lumière sur les variables ayant un lien important. Le degré de significativité généralement utilisée est alpha  inf à 0.05.

Mais le fait d’utiliser plusieurs fois le meme tests dans une boucle afin de tester les variables dans le cas des comparaisons multiples influences enormement le seuil initial de 0.05 alors des correcteurs existent et permerttent de modifier ce biais indu par la repetition des tests statistiques un exemple de correcteur est **le correcteur de bonferroni.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Varaible 1** | **Varaible 2** | **Pvalue** | **Pvalue Corrigée** |
| 'Benefice net annuel | Age' | 1.33 10-110 | 1.23 10-110 |
| 'Benefice net annuel | Prime mensuelle' | 0.02086 | 0.02086 |
| 'Benefice net annuel | Categorie socio professionnelle' | 6.47 10-121 | 5.73 10-121 |
| 'Benefice net annuel | Kilometres parcourus par mois' | 0.0 | 0.0 |
| 'Benefice net annuel | Coefficient bonus malus' | 0.0 | 0.0 |
| 'Benefice net annuel | Type de vehicule' | 1.783 10-124 | 1.55 10-124 |
| 'Benefice net annuel | Score CRM' | 0.0 | 0.0 |
| 'Benefice net annuel | Niveau de vie' | 0.0 | 0.0 |
| 'Benefice net annuel | Marque' | 1.28 10-114 | 1.16 10-114 |
| 'Benefice net annuel | Salaire annuel' | 0.0 | 0.0 |
| 'Benefice net annuel | Score credit' | 1.16 10-302 | 8.3 10-303 |
| 'Benefice net annuel | Cout entretien annuel' | 0.0 | 0.0 |

En se focalisant sur le seuil de 0.05 nous n’avons pas de variable a éliminer de notre modèle vue que les Value et celle corrigées des variables par rapport à la variable Benefice net annuel sont inférieur au seuil. Cependant nous pouvons voir que la variable Prime mensuelle et le Benefice net annuel ne sont pas liées.

**Conclusion 2 : La variable prime pourrait ne pas être prise en compte dans notre modélisation.**

Autres méthodes statistiques pourrait-être utiliser pour réduire la dimension de notre problème cependant ici nous avons des données numériques a prédire sur la base de données connue donc le cas qui colle le mieux est une régression cependant les autres modèles de réduction de dimensions

**La selection des variables**

Elle se base sur les deux methodes precedentes afin d’augmenter le coefficient de correlation de la regression ainsi l’objectif etant le R max , nous optons pour une methode descendante visant a accroite le score de prediction en fonction des variables choisies

1. **Modélisation**

Afin de pouvoir predire nos données dans nore cas , des données continues discrets nous avons ainsi decider de tester les differents algortimes de machines learning et dans garder que celui qui aurait un score au moins supérieur a 0.90 en terme d’accruacy.

Les metriques initalement transmissent dans le projet sont RMSE et l’indice de gini.

Le lien étroit entre ses métriques existent car une RMSE faible sugere une accuracy Forte.

Les modèles implementées sont :

'MLPregression','svm\_m','BayesianRidge\_m','linear\_model\_OLS'

1 Le AdaBoost :

2 Le Random Forest

3 Les arbres de decisions

4 Le Percepetron MultiCouche ou Reseaux de neuronnées

Le SVM Support a vecteurs Machines

Regression Lineaire

7 Les K plus Proches Voisins

Un grand nombre de algorithmes de prédiction peuvent être transformer en algorithme de régression bien que initialement construire pour des cas de classifications telles est le cas des SVM. La bibliothèques Sckit learn offre ainsi ses possibilités que nous ne nous sommes pas privées d’utilisées.

l’une des difficultes qui se posaient fut le chois des parametres de chaque modèle afin d’obtenir les meilleurs scores.

Ainsi pour pouvoir résoudre ce problème nous avons ainsi construire un simulateur afin de lister toute les combinaisons possibles des 12 variables afin de palier au problème de sélection de variable car l’utilisation d’un ACP donné des résultats médiocres dans notre cas lors de son application. Nous avons ainsi 4094 combinaisons possibles sans repetitions que nous avons listés pour reduire la dimensions au meilleur. Cela augment la complexité de notre problème en temps.

Passer ce stage , le problème qui ce pose est plus interne a chaque modèle

tels que le perceptron multicouche le type de fonction d’activation utilisée est ce une fonction tangante , une sigmoidée, etc.. ainsi que la taille des couches cachées .