Projet Machine Learning - Groupe 2

Léo Dutertre, Axel Gardahaut, Guy Tsang

Résumé
Cet article présente les enjeux du projet de Machine Learning et les données mises à disposition. Dans un premier temps, le fonctionnement de certains modèles (régression logistique pénalisée et Extreme Gradient Boosting) sont explicités (intuition, mécanisme, avantages et inconvénients) puis dans un second temps, la démarche du projet est détaillée dans les grandes lignes afin de justifier les choix effectués dans le traitement et la construction des modèles de prédiction. Une partie s'attardant sur les résultats conclura cet article.

1 Introduction

Une société d'assurance souhaite affiner sa capacité à tarifer ses contrats automobiles avec ses clients. L'objectif est de faire payer à chaque assuré son « juste prix ». Ainsi, à partir d'un jeu de données sur ses clients, le but sera de construire un modèle qui prédit, pour chaque assuré, sa probabilité de déposer une réclamation au cours de la prochaine année. Plus cette probabilité est élevée, plus la tarification sera élevée pour l'assuré.

Ceci constitue un problème classique de modélisation aboutissant à un score. La métrique d'évaluation des prédictions devra être pertinente à cette problématique, sachant que la base de données fournie présente un déséquilibre de la cible.

Le but de cet article scientifique est de présenter les données, la démarche du projet, les modèles de prédiction utilisés et enfin, les résultats obtenus.

2 Données

Les données sont fournies par la société d'assurance, comportant une base d'apprentissage et de test. L'apprentissage et la validation des modèles doit se faire sur la première base tandis que la seconde base doit servir uniquement à tester la performance du modèle retenu. Des valeurs manquantes sont présentes dans les deux échantillons.

Parmi les prédicteurs, on retrouve :

- des variables concernant l'assuré en personne ("ind"),
- des variables concernant la région de l'assuré ("reg"),
- des variables concernant la voiture de l'assuré ("car"),
- des variables calculées ("calc").

La variable cible ("target" dans la base) est binaire et indique si une réclamation a été déposée ("1") par l'assuré ou non ("0"). Cette variable est déséquilibrée, avec moins de 4% de labels positifs. Chaque ligne de la base de données correspond à un assuré automobile. Les intitulés des colonnes sont anonymisées.

3 Démarche du projet

3.1 Définition du cadre d'analyse

L'objectif de ce problème de classification est de réaliser un scoring des clients. Le score correspond à la probabilité pour chaque client de déposer une réclamation dans l'année à venir. La métrique retenue pour évaluer la performance prédictive des modèles est celle du Coefficient Normalisé de Gini. Le coefficient de Gini permet de comparer la proportion cumulée des labels positifs prédits avec la proportion théorique. De plus, ce coefficient est directement lié à la courbe de Lorentz, souvent utilisée en économie pour évaluer les inégalités salariales. Si on transpose ce concept à une société d'assurance (figure 1), on identifie donc l'intérêt du coefficient de Gini dans notre contexte.

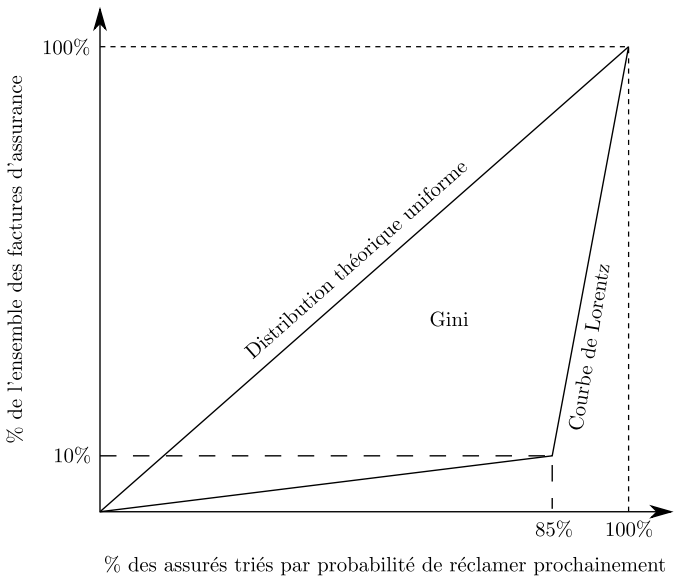

Figure 1 – Coefficient de Gini et Courbe de Lorentz

Une lecture de la courbe de Lorentz (figure 1) serait « les 15% des assurés les plus risqués paient 90% de l'ensemble des factures d'assurance ». Ceci correspond bien au principe de mutualisation des risques qui correspond à la base des sys-

tèmes d'assurance. Ainsi, plus le coefficient de Gini est élevé, plus la capacité à individualiser les tarifs sera bonne. La normalisation de ce coefficient permet de s'assurer que la valeur prise par la métrique va de 0 (aucune capacité prédictive) à 1 (capacité prédictive parfaite).

3.2 Statistiques Descriptives

Les statistiques descriptives sont faites pour vérifier le type de données traitées, la présence de valeurs manquantes, la présence de valeurs aberrantes et les distributions. De plus, on repère les types de variables (continues, catégorielles, ordinales, etc.) pour spécifier le traitement au cas par cas.

L'inférence faite à partir de modalités rares (moins de 5% en général) n'est pas robuste et peut conduire à du surapprentissage. Les statistiques descriptives sur les variables catégorielles permettent de détecter ces modalités.

Enfin, l'étude des variables continues permet de prévoir l'utilité de transformations (normalisation ou standardisation).

3.3 Benchmark d'ouverture

Il est intéressant de placer un benchmark afin d'avoir une valeur de la métrique objectif (Coefficient Normalisé de Gini) à dépasser. Le score obtenu par un Light GBM [1] est généralement élevé et rapidement obtenu. Celui-ci est de : 0.2719 par validation croisée sur 5 blocs (5fCV). Bien que le score obtenu est sujet à fluctuer, il est intéressant de refaire un Light GBM après chaque étape de traitement pour voir si les modifications apportées augmente ou diminue le score.

3.4 Pré-traitement des données

Le pré-traitement des données consiste essentiellement à la gestion des valeurs manquantes. Plusieurs variables ont été identifiées dans les statistiques descriptives (figure 2). Selon le taux de valeurs absentes, la manipulation appliquée sera différente.

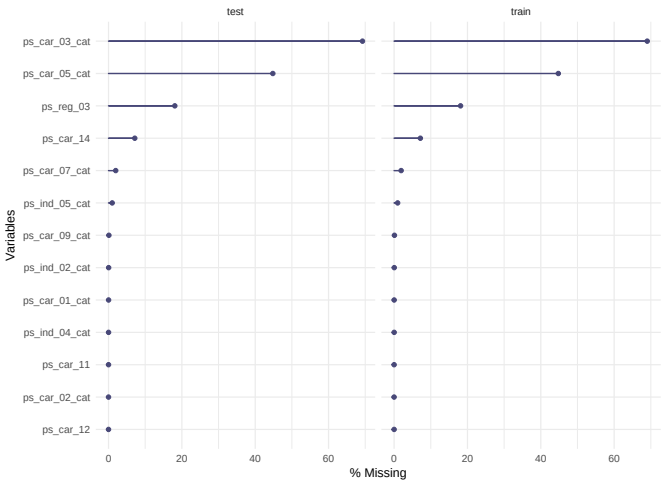

Figure 2 – Valeurs manquantes par variable

1. Gradient Boosting Machine

Deux colonnes présentent trop de valeurs manquantes (plus de 40%) pour une imputation sophistiquée, ainsi, elles sont remplacées par une valeur spéciale (-999 par exemple) pour conserver l'information issue de l'absence de valeur. En effet, en croisant ces variables avec la cible, il est noté que le taux croisé d'effectif est significativement différent à ceux des valeurs normales.

Pour le reste des variables, une imputation par forêt aléatoire est faite. Le procédé d'imputation est le suivant :

- Utiliser uniquement l'échantillon d'apprentissage pour construire la forêt aléatoire.
- Modéliser la variable étudiée en s'assurant qu'il s'agisse du bon type d'arbres (classification ou régression) en utilisant les autres variables comme régresseurs.
- Prédire les valeurs de la variable étudiée pour l'ensemble des observations manquantes des deux échantillons (apprentissage et test).
- Remplacer les valeurs manquantes et s'assurer qu'il n'y a plus de valeurs manquantes pour la colonne traitée.

Suite aux imputations faites, le Light GBM de référence renvoie un Gini de 0.2721, toujours par validation croisée sur 5 folds. On note une augmentation par rapport au benchmark bien que la différence soit trop faible pour en tirer des conclusions.

3.5 Sélection des variables

Le Light GBM effectué après les imputations permet également d'identifier l'importance des variables dans le pouvoir prédictif du modèle. L'importance d'une variable peut être mesurée de différentes manières. Celle adoptée ici est le « Gain », i.e. la contribution de la variable au modèle. Ainsi, plus une variable contribue au pouvoir prédictif d'un modèle, plus son « Gain » associé sera élevé. Arbitrairement, on garde la première moitié des variables les plus contributives au modèle.

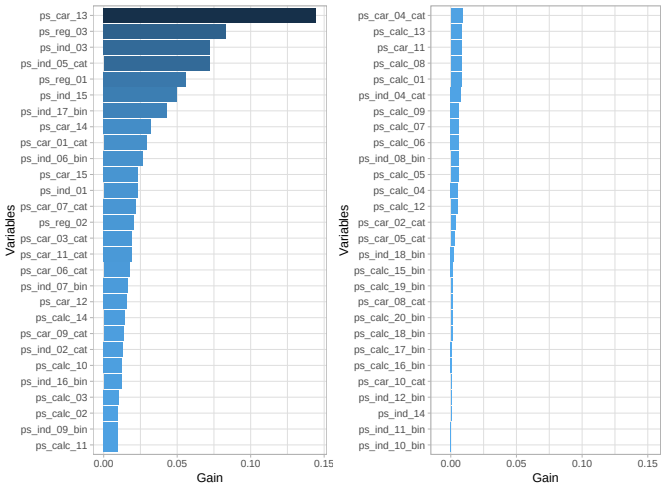

Figure 3 – Importance des variables

En plus des importances issues du Light GBM, une forêt aléatoire (non paramétrée de façon optimale) est construite pour réaliser une seconde liste d'importance en termes de

« permutation » (notion équivalente au « Gain »). De même, cette seconde liste est issue de la première moitié des variables les plus importantes au modèle. L'union de ces deux listes constitue la liste finale des variables retenues pour la suite. Au total, ce sont 33 régresseurs sur les 57 initiaux qui sont retenus.

3.6 Traitement des données

Lors de la phase des statistiques descriptives, plusieurs variables présentaient des modalités rares (effectif inférieur à 5%). Celles-ci doivent être fusionnées pour pouvoir inférer des résultats robustes (AFDM ou tableaux de contingence).

La stratégie de fusion peut passer par une projection des modalités d'une variable sur le premier plan factoriel d'une AFDM. Chaque modalité rare sera alors associée à la modalité fréquente la plus proche d'elle (et si possible projetée dans la même direction), ou un ensemble de modalités rares peuvent se regrouper pour former un groupe supplémentaire (clusters).

La stratégie de fusion peut également passer par des tableaux de contingence (figure 4) qui croisent les variables problématiques avec la cible. Les modalités rares sont fusionnées soit entre elles si elles se comportent de façon similaire, soit avec une modalité fréquente si le comportement s'y rapproche.

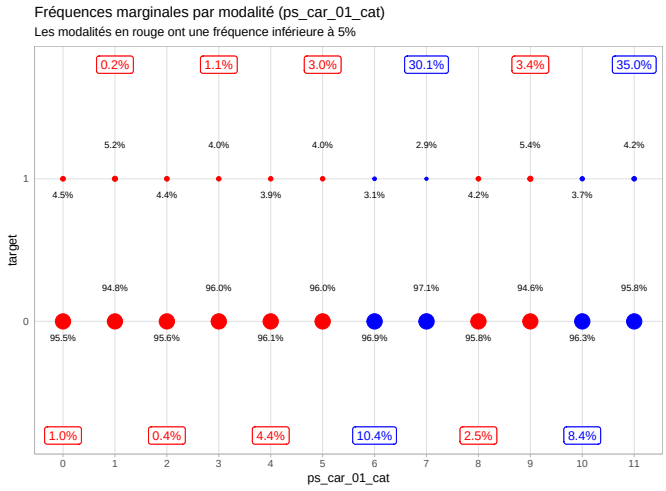

Figure 4 – Exemple de projection de modalités par croisement

* * *

Une variable présente un grand nombre de modalités (104). Plutôt que de fusionner ces modalités, on peut s'en servir comme régresseur unique d'une régression logistique avec la variable cible comme variable à expliquer. Les modalités seront alors remplacées par les coefficients estimés et la variable deviendra alors numérique et continue.

Le Light GBM fournit un score de 0.2729 (5fCV) pour la stratégie par AFDM et de 0.2757 (5fCV) pour la stratégie par tableaux croisés. Le score ayant fortement augmenté pour la seconde stratégie, elle sera retenue pour la suite.

3.6.1 Normalisation des données

Certaines variables continues ont une distribution qui se rapprochent d'une distribution normale. Il serait intéressant de tenter de les normaliser à travers différentes méthodes (box-cox, logarithme, racine carré).

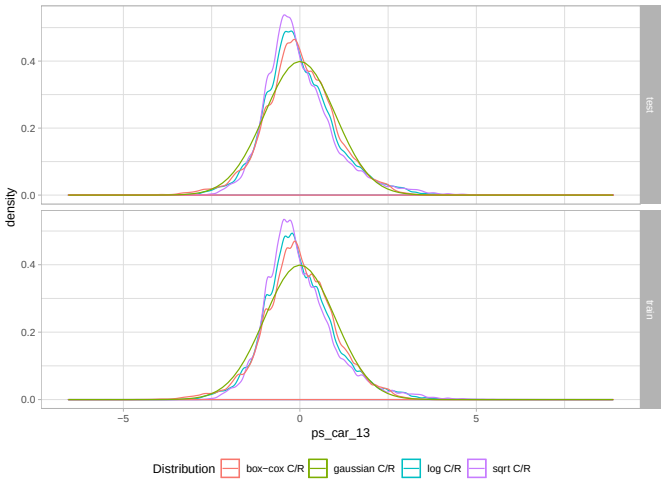

Figure 5 – Exemple de transformation

Il est notable que la transformation de box-cox est la meilleure pour normaliser cette distribution. Cette étude est menée pour les variables continues.

En réutilisant un Light GBM pour mesurer les conséquences des normalisations faites, il est noté que la métrique diminue, suite à chacune des deux stratégies de fusion de modalités. On passe respectivement à un Gini de 0.2729 (AFDM) et 0.2752 (tableaux croisés). Ces diminutions sont minimes mais la normalisation des variables n'est pas obligatoire. Par conséquent, cette étape est laissée de côté pour la suite.

À la fin de cette étape, le traitement se résume à la fusion de modalités en utilisant des tableaux de contingence.

3.7 Étude des corrélations et dépendances

L'objectif de cette étape est de diminuer à nouveau le nombre de variables explicatives. La stratégie est différente : il s'agit d'étudier les corrélations et les dépendances entre variables puis avec la cible. Ceci permet de détecter des problèmes de colinéarité et ou dépendance entre prédicteurs :

- les liens entre variables continues se font grâce aux corrélations,
- les liens entre les variables continues et la cible se font grâce à un test d'ANOVA,
- les liens entre variables catégorielles (et la cible) se font grâce aux V de Cramer [2].

Les variables qui dépassent les seuils arbitrairement fixés ou qui ne passent pas le test d'ANOVA sont laissées de côté. Le nombre de prédicteurs chute alors à 29.

2. Le V de Cramer est la statistique du χ^2 standardisée de façon à ce qu'elle varie entre 0 et 1

3.8 Paramétrisation des modèles

Une dizaine d'algorithmes sont testés pour tenter d'avoir le meilleur pouvoir prédictif. On y trouve des modèles de base (LDA, KNN, régression logistique (pénalisée), SVM), puis des modèles de bagging (RF) et de boosting (ADABoost, GBM, XGBM, LGBM).

L'ensemble des modèles ensemblistes et la régression logistique pénalisée sont optimisés selon leurs paramètres de pré-apprentissage [3].

Pour chaque modèle, la question du ré-échantillonnage est traitée. Pour l'ensemble des modèles, les stratégies de down-sampling, up-sampling et SMOTE sont testées en plus du cas sans sampling.

4 Résultats

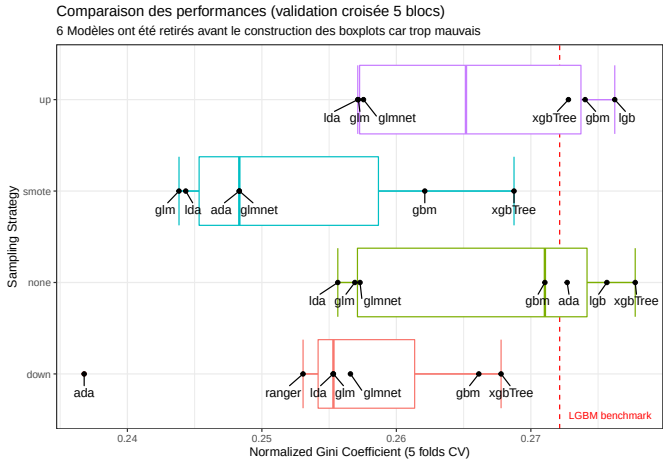

Figure 6 – Performance de validation des modèles

Références

[1] Abu-Rmileh, A. The multiple faces of 'feature importance' in xgboost. Towards data science, 8 Février 2019.

[2] Gandhi, R. Boosting algorithms : Adaboost, gradient boosting and xgboost. Hacknernoon, 5 Mai 2018.

[3] Hale, J. Smarter ways to encode categorical data for machine learning. Towards data science, 11 Septembre 2018.

[4] Juhi. Gini coefficient and lorenz curve explained. Towards data science, 6 Mars 2019.

[5] Rstats on pi : predict/infer. Be aware of bias in rf variable importance metrics. R-bloggers, 19 Juin 2018.

3. Les algorithmes KNN et SVM ne sont pas optimisés faute du temps d'exécution nécessaire au calcul de matrices immenses de distance