réinventons / notre métier

Graduate Program d'AXA 2017 – Study Case Data Science

But de l'Exercice

Le principe général de l'exercice est semblable à une compétition de type « Kaggle » : l'objectif est de construire un modèle prédictif à partir d'un jeu données, ici , il s'agit d'estimer le bénéfice net annuel attendu par client sur un contrat d'assurance automobile en fonction des caractéristiques du client. Ce modèle sera ensuite appliqué à un autre jeu de données afin de générer des prédictions qui seront évaluées.

Consignes

Vous avez normalement reçu 2 fichiers csv en plus de ces instructions.

Labeled_dataset.csv contient 1 ligne de noms de variables, et 1000 lignes de données. Chacune de ces 1000 lignes contient :

- 1 index de 0 à 999, représentant de façon unique chaque client
- 12 variables (age, salaire, coefficient bonus-malus…) caractéristiques du client
- 1 « label », benefice, indiquant le bénéfice net annuel réalisé pour ce client

Scoring_dataset.csv contient 1 ligne de noms de variables, et 300 lignes de données. Chacune de ces lignes contient :

- 1 index, de 1000 à 1299
- 12 variables (*age, salaire, coefficient bonus-malus…)*

Vous devez utiliser les données de ***labeled_dataset*** pour construire un modèle prédictif. Vous pourrez ensuite appliquer ce modèle sur les données de ***scoring_dataset*** afin de générer vos prédictions.
Ces prédictions seront évaluées par nos équipes de data science, qui disposent du dataset étiqueté contenant les « vrais » bénéfices nets réalisés. La métrique utilisée pour comparer vos résultats avec les réponses sera le RMSE (Root Mean Squared Error).

NOTE IMPORTANTE CONCERNANT LE JEU DE DONNEES :

Ce dataset pour ce test, a été intégralement généré par nos équipes de data science. Nous avons essayé de lui donner une connotation métier en lien avec l'assurance, mais toutes les données sont purement fictives.

Plus spécifiquement, les âges, salaires, catégories socio-professionnelles etc… ne sont pas représentatifs de nos clients ou d'une quelconque population. Le lien entre les caractéristiques d'un client et le bénéfice réalisé est également fictif. La précision atteignable en termes de performances du modèle n'est pas non plus nécessairement réaliste.

Restrictions technologiques

Il n'y a pas de liste spécifique de technologies autorisées. Un ensemble non exhaustif de langages et librairies pertinents peut inclure Python, Scikit-learn, R, Apache Spark, Java, Scala, C++… Il est recommandé au candidat d'utiliser une technologie avec laquelle il est familier, la qualité du code faisant partie des critères évalués.

Il est cependant impératif que les équipes d'évaluation soient en mesure d'exécuter le code fourni ; **cela implique qu'aucun des langages ou librairies utilisés ne soit sous licence payante (SAS, SPSS…).** Il est de façon générale préférable de privilégier des technologies open source.

Il est enfin à noter que les différentes équipes de data science d'AXA peuvent être plus ou moins appétentes à certaines technologies en fonction de leurs problématiques. Un choix de technologie pertinent vis-à-vis de l'équipe dans laquelle vous postulez sera certainement apprécié.

Livrables attendus

- Un fichier csv contenant vos prédictions : ce fichier doit être composé de 300 lignes, chacune reprenant les index de *scoring_dataset.csv* ainsi que votre prédiction pour cet index.
- Le code ayant permis la génération du modèle et les instructions permettant d'exécuter le code le cas échéant.
- Un descriptif de la méthodologie mise en place ainsi que des résultats obtenus ; le choix de la forme de ce descriptif est laissé au candidat (fichier pdf, présentation PowerPoint…).

Critères d'évaluation

Comme indiqué précédemment, un score quantifiant la précision de vos prédictions sera calculé à partir du dataset de test, en se basant sur la métrique RMSE. Il ne s'agit cependant pas du seul critère d'évaluation.

La méthodologie mise en place est au moins aussi importante que le score lui-même. Il est ainsi absolument fondamental de rendre votre démarche la plus claire et transparente possible. Notamment, n'hésitez pas à mentionner des méthodes que vous auriez testée y compris si celles-ci qui ne se sont pas révélées pertinentes dans le cas présent. Un code lisible et commenté est de plus fortement valorisé.

Les capacités de restitution seront également évaluées.

Précisions supplémentaires

En cas de question sur le test technique, vous pouvez contacter Marine BIANCONI (marine.bianconi@axa.fr) en mettant Yannick LE CACHEUX (yannick.lecacheux@axa.fr) en copie.

Bonne chance !