Projet 7 : Implémentez un modèle de scoring

Lancelot LECLERCQ

5 mai 2022

Sommaire

1. Introduction

2. Analyse et traitement des données

3. Optimisation du modèle

Introduction

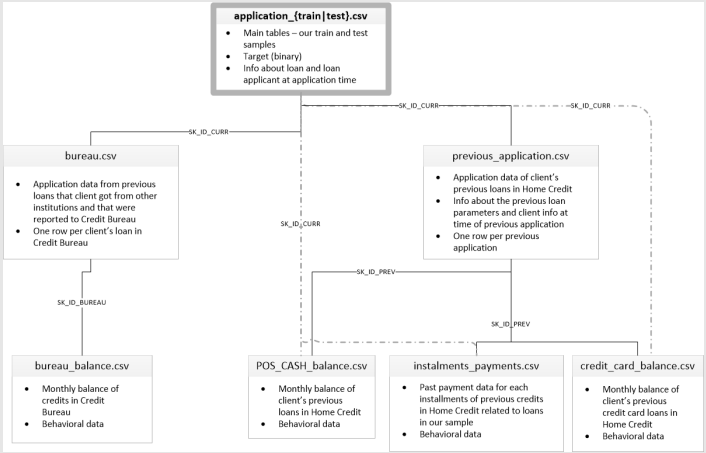
Problématique

- L'entreprise Prêt à dépenser est une société financière qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt

- Objectifs
  - mettre en œuvre un outil de "scoring crédit" pour calculer la probabilité qu'un client rembourse son crédit
  - classifier la demande en crédit accordé ou refusé
    - développer un algorithme de classification en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.)
  - Développer un dashboard interactif
    - expliquer de façon la plus transparente possible les décisions d'octroi de crédit,
    - permettre aux clients de disposer de leurs informations personnelles et de les explorer facilement

Données

- Principal fichier utilisé application_{train‖test}.csv

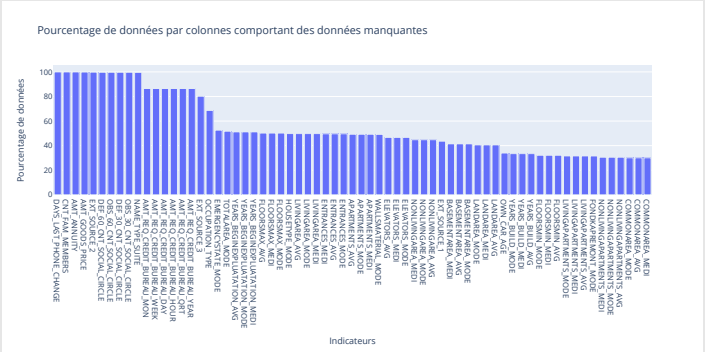


Analyse et traitement des données

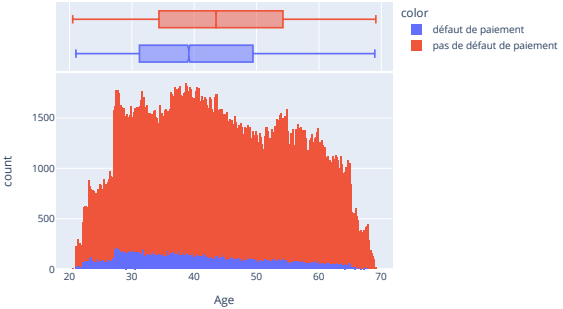
Analyse et traitement des données
Exploration du jeu de données

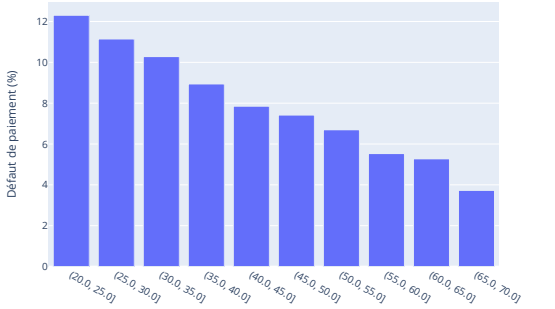
- Certaines colonnes comportent un grand nombre de données manquantes
  - Nous utiliserons des modèles résistants à ces données manquantes comme XGBoost et LightGBM

- Encodage des variables catégorielles
  - par LabelEncoder pour les variables ayant 2 catégories
  - par pandas.get_dummies() pour les varibles ayant plus de 2 catégories



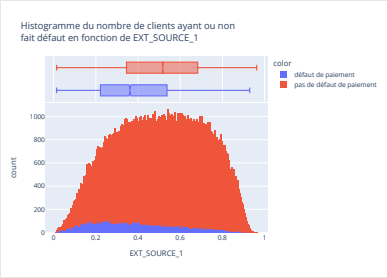
Analyse et traitement des données
Exploration du jeu de données

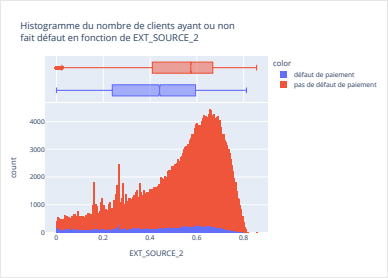


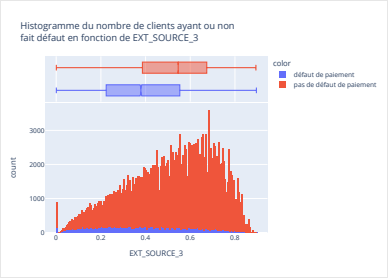


- L'age des clients semble avoir un impact sur le fait que le client fasse défaut ou non

Analyse et traitement des données
Exploration du jeu de données



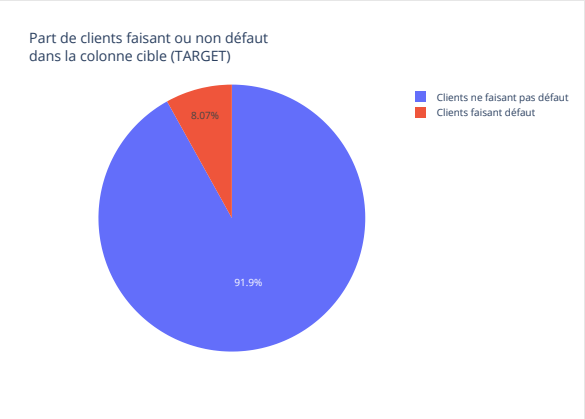




- Les données EXT_SOURCE semblent aussi avoir une certaines corrélation avec le fait que le client fasse défaut

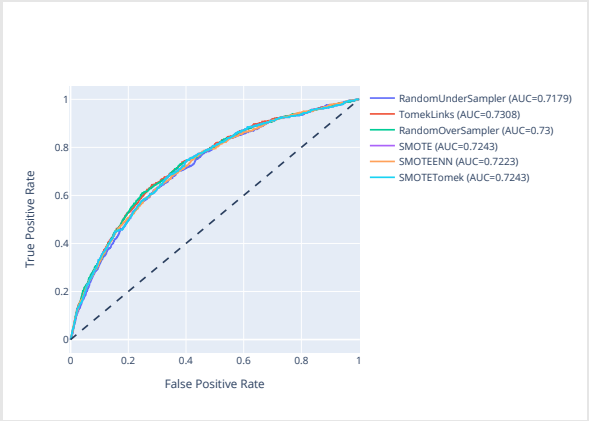
Analyse et traitement des données
Exploration du jeu de données

- Comme on a pu le voir les données sont déséqulibrées du fait que les clients faisant défauts sont peu nombreux par rapport à ceux ne faisant pas défaut
- Classer tous les clients comme ne faisant pas défaut permettrait d'avoir un score honnorable avec seulement 8% d'erreurs
- Nous avons donc utilisé la librairie imblearn qui permet de rééchantillonner notre jeu de données

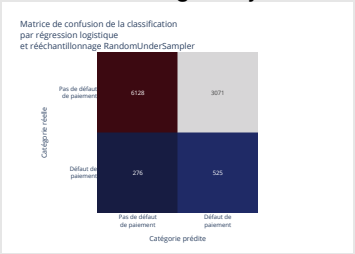


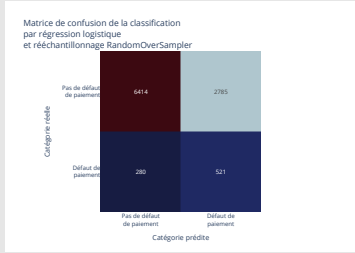
Analyse et traitement des données
Rééchantillonnage du jeux de données

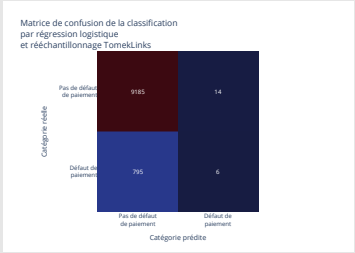
- Pour essayer les différentes méthodes de rééchantillonnage nous avons réaliser une régression logistique sur les données rééchantiollonnées avec les différents outils
  - RandomUnderSampler et TomekLinks : méthodes de sous-échantillonnages
    - on conserve le même nombre de clients ne faisant pas défaut que de client faisant défaut
    - RandomUnderSampler choisi ces dernier au hasard
    - TomekLinks conserve un certains nombre de clients par groupe de clients similaire (repose sur les KNN)
  - RandomOverSampler et SMOTE : méthodes de sur-échantillonnages
    - on multiplie le nombre de clients faisant défaut
    - RandomOverSampler dédouble des clients faisant défaut au hasard
    - SMOTE créé de nouveaux clients à partir de groupe de clients similaires
  - SMOTEENN et SMOTETomek sont des méthodes combinant le sur- et le sous-échantillonnage
- Les scores AUC semblent plutôt bon ($>0{,}7$)

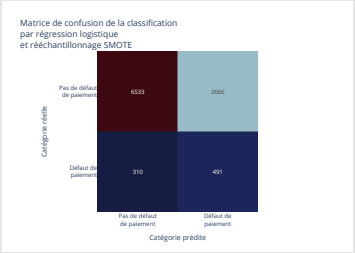


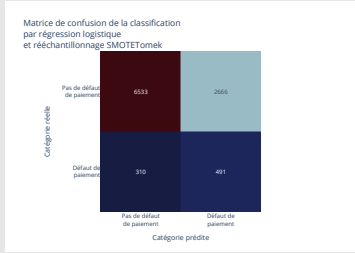
Analyse et traitement des données
Rééchantillonnage du jeux de données

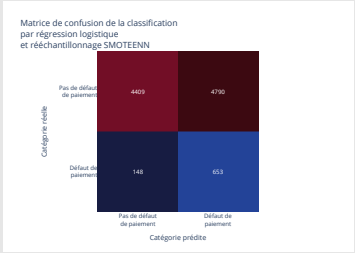












- Les 2 cas de droites ne sont pas intéressant car le premier classe tout les clients comme ne faisant pas défaut et le second classe 50/50

Analyse et traitement des données
Création de variables polynomiales

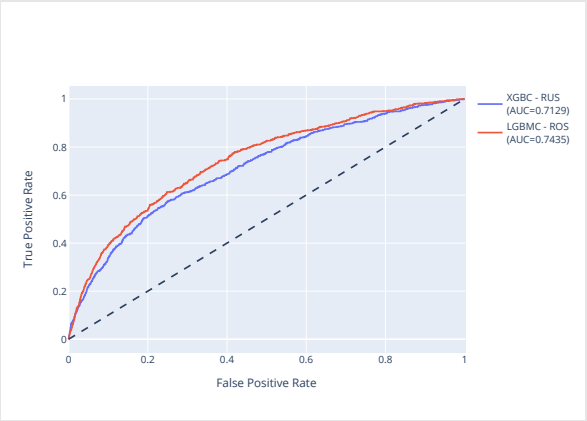
- Afin d'améliorer les scores des modèles nous avons essayé de créer des variables polynomiales à partir des colonnes les plus corrélées avec la cible

- L'amélioration n'est pas pertinente nous n'avons donc pas conservé ces variables pour notre modèle final

Optimisation du modèle

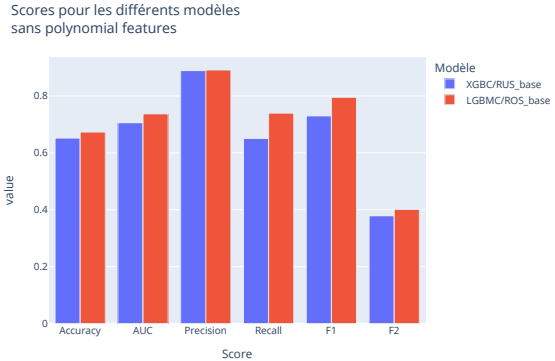
Optimisation du modèle
Courbe ROC

-



Optimisation du modèle
Différentes métriques utilisées

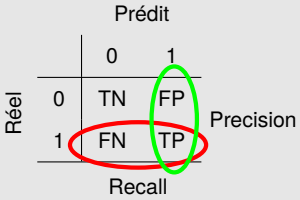
-



Optimisation du modèle
Métrique métier

- But
  - Diminuer le nombre de faux négatifs (prédit 0, réel 1) afin d'éviter de manquer des clients qui pourraient potentiellement faire défaut

  → Améliorer le recall



- Outil
  - Utilisation du F_β-score qui permet d'ajouter du poid respectivement au recall lorsque le facteur β est >1 ou à la precision lorsque le facteur β est <1
  - Utilisation de β=2