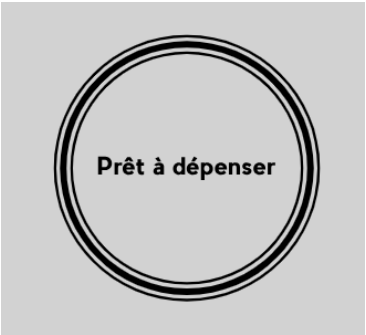
LECLERCQ Lancelot Mai 2022

Note méthodologique



Sommaire

1 Introduction

L'entreprise Prêt à dépenser souhaite utiliser un outil de "scoring" afin de calculer la probabilité qu'un client fasse ou non défaut lors du remboursement de son crédit. Pour cela nous devons entrainer un modèle de classification sur des données variées (comportementales, autres institutions financières, etc).

2 Classification

2.1 Jeu de donées

Afin de mieux comprendre les données nous avons procédé à une analyse exploratoire des données sur le jeu application_train.csv. Pour l'entrainement du modèle nous utiliserons uniquement ces données car l'utilisation des fichiers supplémentaires demande beaucoup de ressources tant en temps d'analyse et d'exploration des données qu'en capacitées de calculs. Nous utiliserons le fichier application_test.csv pour le dashboard.

2.2 "Feature engeneering"

Calcul de variables polynomiales à partir des meilleurs variables (EXT_SOURCES). Il n'y a pas d'amélioration significative des résultats (Fig. 1)

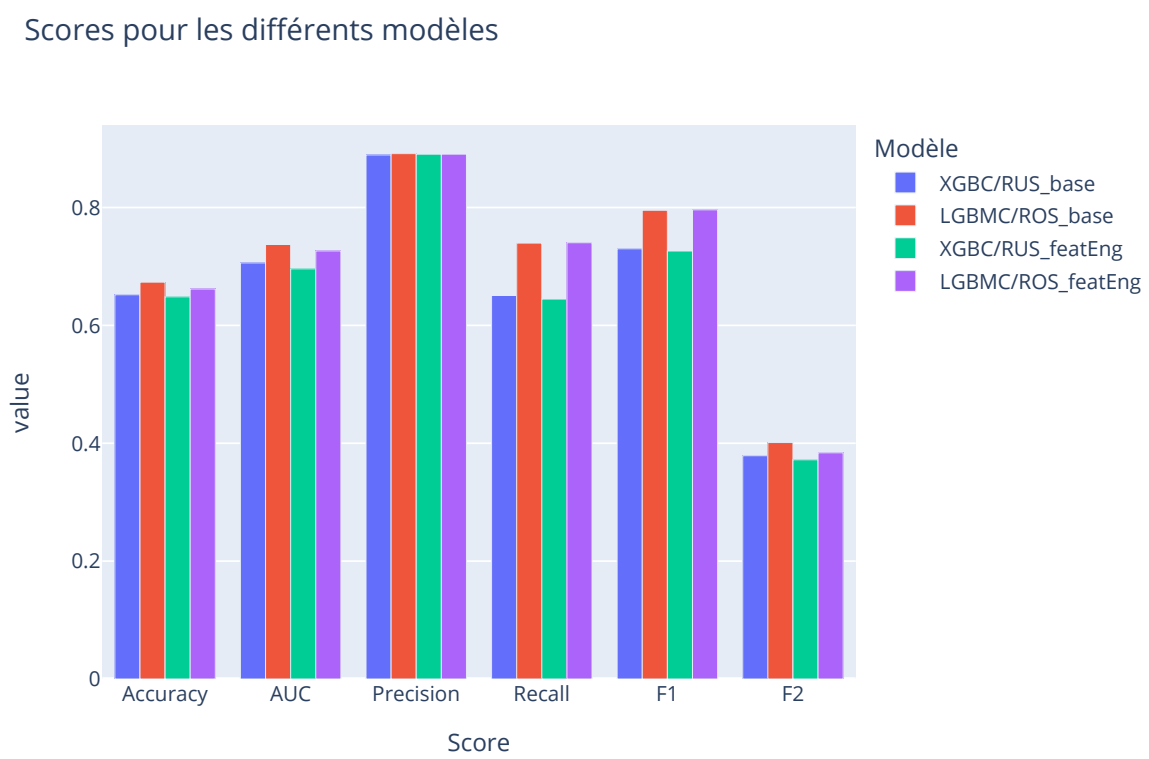


FIG. 1 : Comparaisons des différents scores avec (_featEng) et sans (_base) variables polynomiales

2.3 Déséquilibre des valeurs cible

Du fait d'un déséquilibre dans les valeurs cible il est difficile pour le modèle de classer efficacement les clients (Fig. 2).

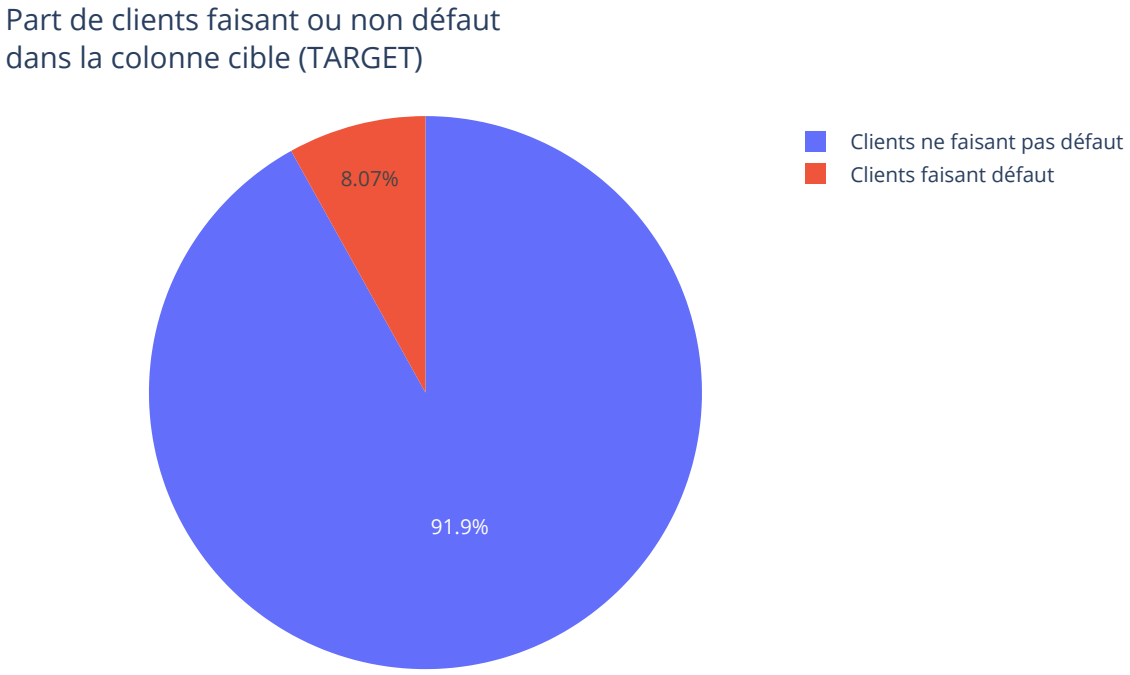


FIG. 2 : Diagramme circulaire illustrant le déséquilibre des types de clients dans la colonne cible

Lors de l'optimisation du score AUC le modèle va préférer classer tout les clients comme ne faisant pas défaut car cela amériorera sont score. Nous avons donc dû rééquilibrer la part des valeurs cible grace à la librairie imblearn.

2.4 Modèles

3 Évaluation, optimisation et coût métier

3.1 Courbe ROC et aire sous la courbe (AUC)

3.2 GridSearch

3.3 Coût métier

4 Interprétabilité

Dans un soucis de transparence nous souhaitons pouvoir expliquer comment fonctionnent nos modèles

4.1 Globale

L'étude des principales variables utilisées par le modèle permet de mieux comprendre son fonctionnement global (Fig. 3)

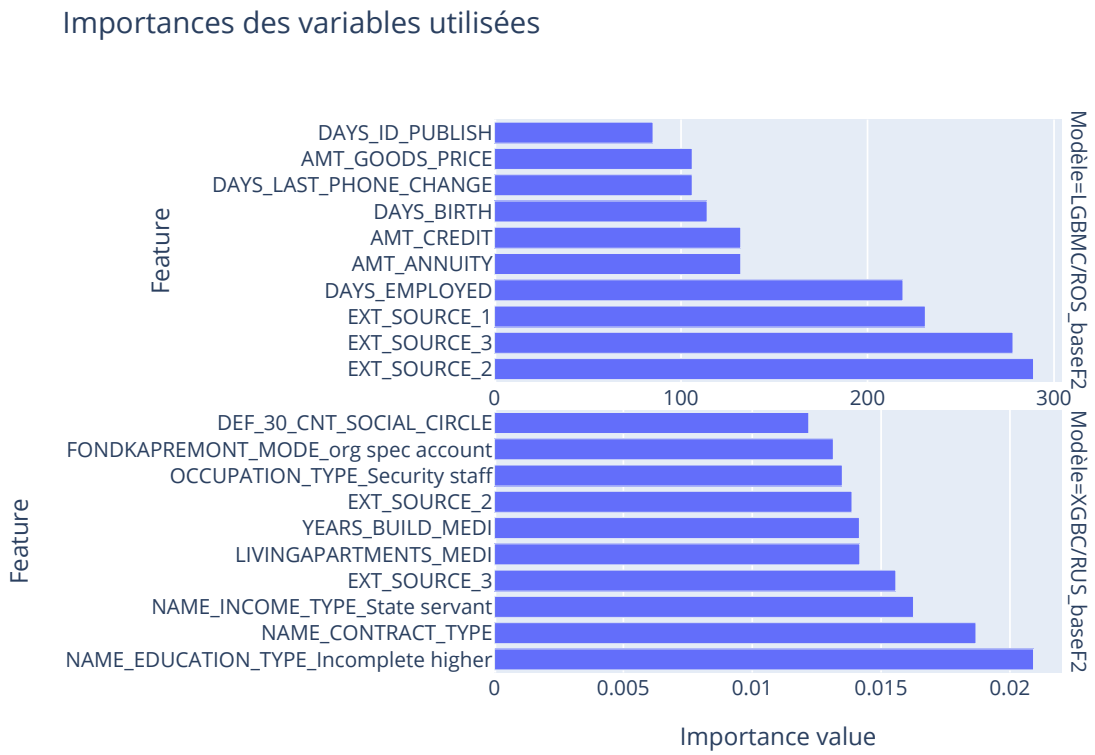


FIG. 3 : Principales variables selon leur importance dans l'entrainement des modèles

4.2 Locale

La librairie SHAP nous permet d'en apprendre un peu plus sur la part des variables dans le classement d'un client en particulier.

5 Limites et améliorations

6 Conclusion