PROJET 7 – « IMPLÉMENTEZ UN MODÈLE DE SCORING »

Soutenance de projet – parcours Data Scientist

31 Mars 2020

Sommaire

I. Rappel de la problématique et présentation du jeu de données

II. Explication de l’approche de modélisation

III. Présentation du dashboard métier

I - PROBLÉMATIQUE

Rappel de la problématique

Présentation du jeu de données

Appliquer un scoring crédit aux client

- **Contexte** : « Prêt à Dépenser» : société de crédits à la consommation
- **Objectifs** :
  - développer un modèle de **scoring de la probabilité de défaut de paiement du client** *(avec pas ou peu d'historique de prêt)*
  - Développer un **dashboard interactif** pour assurer une transparence sur les décisions d'octroi de crédit

Jeu de données

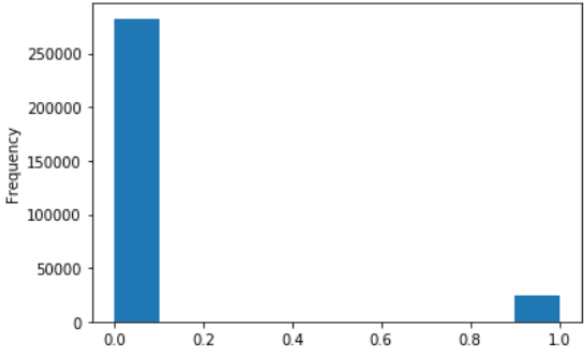
- 7 sources de données (relatives aux clients et à la société : précédentes demandes de crédit, balance de crédit, cash, etc.)
- Base de données principale :
  - 307 000 clients
  - 121 features : âge, sexe, emploi, logement, revenus, informations relatives au crédit, etc.
  - Labels cible : défaut de crédit / pas de défaut de crédit

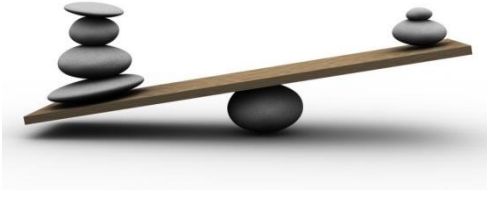
Feature engineering

- Utilisation d'un notebook issu de Kaggle
- Processus:
  - One hot encoding
  - Détection des outliers / anomalies
  - Création de features métier :
    - Ratio du montant du crédit ramené au revenu
    - Ratio des annuités raménées au revenu
    - Durée du crédit
    - Pourcentage de temps employé (relatif à l'âge du client)
  - Imputation des valeurs manquantes
- Obtention d'un jeu de données de 244 features

Un jeu de données déséquilibré

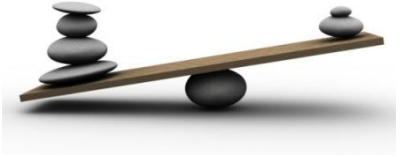
- 91 % des clients réguliers
- 9 % des clients avec des défauts de paiement





- Modèle Naif : classe sans défaut pour tous les cas : accuracy élevée
- Surreprésentation de la classe majoritaire dans la prédiction

Comment réduire les conséquences de se déséquilibre?



- Collecter plus de données sur la classe minoritaire
- Under-sampling (réduire le nombre d'individus de la classe majoritaire)
- Dupliquer des individus sous représentés
- **Choix d'une métrique de performance adaptée**
- **Création d'invididus « synthétiques »**
- **Pondération des observations dans le training**

II – EXPLICATION DE L'APPROCHE DE MODÉLISATION

Métrique de performance

Méthodologie

Modèle retenu

Quel scoring adapté au problème métier?

- **Problématique :**
  - Les clients à risque font perdre de l'argent à la société
  - La société ne doit pas se priver des potentiels clients qui ne présentent pas de risque

  ►► Optimum à déterminer

- **Postulats:**
  - Les clients à risque non identifiés représentent une dépense effective importante pour la société (frais de recouvrement, sommes non recouvrées)
  - Les clients peu risqués identifiés à tort comme risqués font perdre à la société un chiffre d'affaire potentiel (coût d'opportunité)

Fonction de scoring

- **Transposition du problème**
  - Limiter le nombre de faux négatifs
  - Limiter dans une moindre mesure le nombre de faux positifs

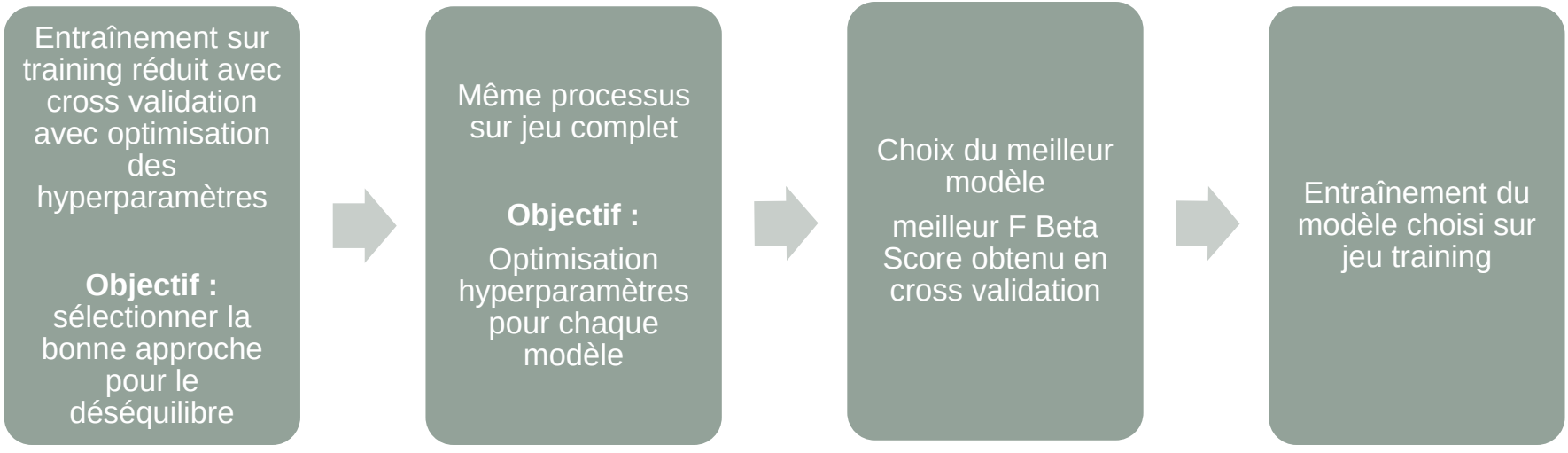
	prédit en défaut	**prédit sans défaut**
réel en défaut	vrais positifs	faux négatifs
réel sans défaut	faux positifs	vrai négatifs

  - Equilibre à trouver entre **Précision et Recall**
    - **Précision = TP /(TP + FP)**
    - **Recall = TP / (TP+FN)**
  - F Beta Score : permet d'identifier un compromis entre les 2 métriques
  - Beta : importance relative du recall par rapport à la précision

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

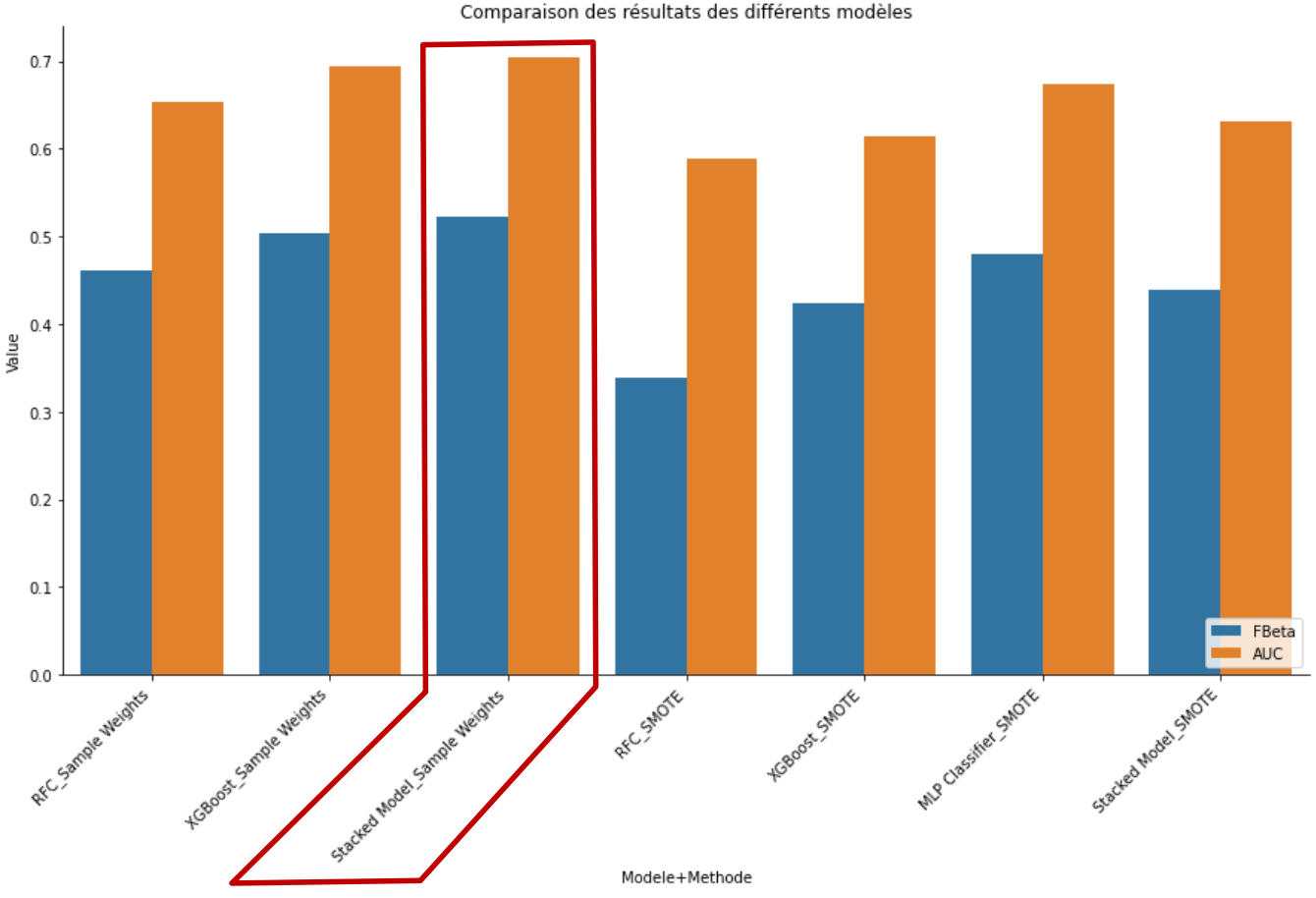
  - Hypothèses de dépense de recouvrement vs coût d'opportunité aboutissent à **Beta = 2,75**
  - F beta score compris entre 0 et 1 : 1 étant le classifieur parfait

Méthodologie



- Algorithmes :
  - Random Forest Classifier
  - XGBoost Classifier
  - MLP Classifier
  - Stacking (XGBoost)
- Approches du déséquilibre:
  - Sample Weights
  - Smote

Comparaison des modèles après cross validation



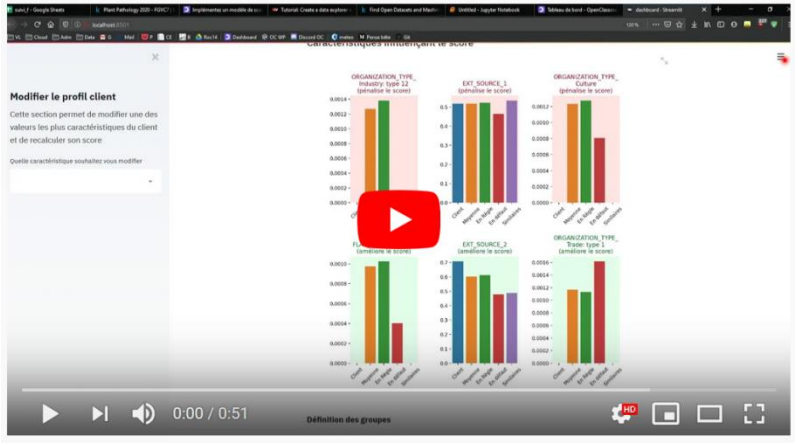
- **Meilleur modèle : Stacking Sample Weights (XGBoost)**
  - Fß = 0.522
  - AUC = 0.704

III – PRÉSENTATION DU DASHBOARD

Outils utilisés

Solution	Description
Flask — web development, one drop at a time	API permettant d'appeler la prédiction à partir de l'ID du client 127.0.0.1:5000/credit/413394 ▶ prediction: 1 proba: 0.456686794757843
Streamlit.	Tableau de bord : Front
GitLab	Versioning
LIME	Explicabilité de la prédiction

Tableau de bord interactif



Commencer la démonstration

Voir la vidéo

CONCLUSION

Aller plus loin

- **Un modèle plus performant**
  - Une métrique d'évaluation basée sur des hypothèses métier confirmées
  - Feature engineering plus poussé

- **Améliorer le dashboard**
  - Explicabilité plus précise (notamment avec variables du one hot encoding)
  - Graphes interactifs
  - Faire évoluer les scoring extérieur en même temps que les features sont modifiées

MERCI DE VOTRE ATTENTION