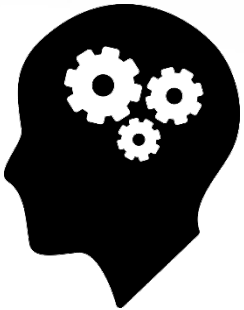
Projet 7 :

Implémentez un modèle de scoring

Lecerf Defer Amandine

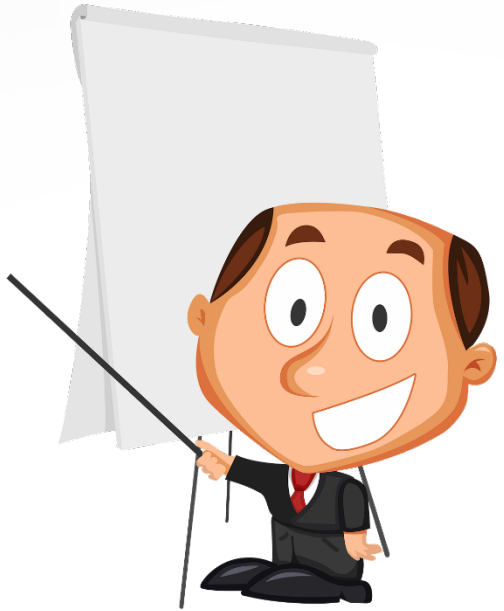
Compétences évaluées

- Présenter son travail de modélisation à l'oral
- Déployer un modèle via une API dans le Web
- Utiliser un logiciel de version de code pour assurer l’intégration du modèle
- Rédiger une note méthodologique afin de communiquer sa démarche de modélisation
- Réaliser un dashboard pour présenter son travail de modélisation



Plan

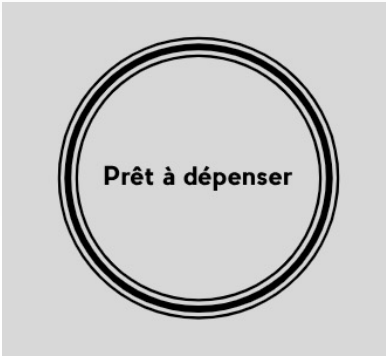
I. Contexte + Données disponibles
II. Traitement des données
III. Méthodologie : entraînement de modèles
IV. Choix d'un algorithme adapté
    a) Métriques + fonction coût
    b) API
V. Interprétabilité
    a) Importance des Features
    b) Dashboard
VI. Limites et améliorations



Problématique + données

Contexte

- Data Scientist pour « Prêt à dépenser »
  - Crédits à la consommation
  - personnes avec peu ou pas d'historique de prêt



- Souhait :
  - Développer modèle de scoring : probabilité défaut de paiement
  - Informations générales : identité, données comportementales, données provenant d'autres institutions bancaires

Objectifs

- **Objectifs :**
  - **Analyse d'un jeu de données :**
    - Préprocessing
    - Détermination algorithme optimal
    - Métrique bancaire
  - **API :**
    - Probabilité défaut paiement
  - **Dashboard interactif**
    - Visualisation score + interprétation
    - Visualisation informations descriptives d'un clients précis
    - Interprétation prédiction modèle

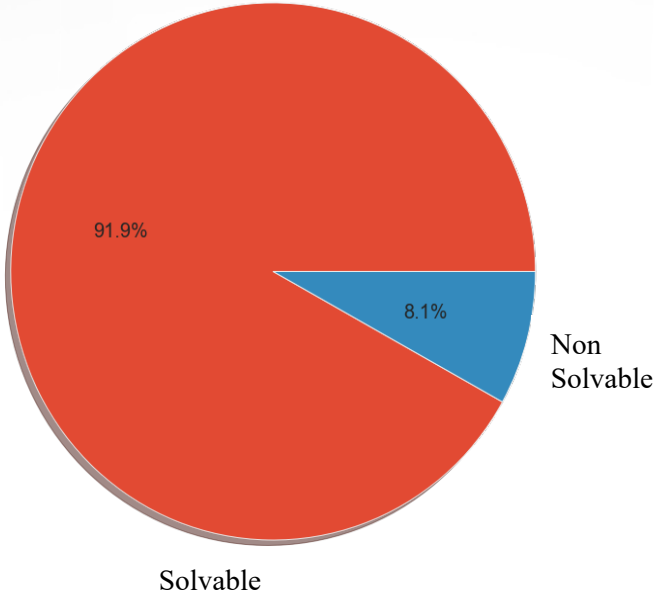
Données Disponibles

- Service « Home Credit » : fourniture de crédits à la population non bancarisée

- 9 fichiers :
  - 307 511 clients pour 122 Indicateurs
  - Client : informations générales + informa sur les prêts précédents
  - Je me base sur les informations générales (âge, revenus, crédit en cours, …)
  - Pour chaque client : capacité ou non de payer crédit (solvabilité)

- Fort déséquilibre entre les personnes solvables et les personnes non solvables.

Déséquilibre de la Target

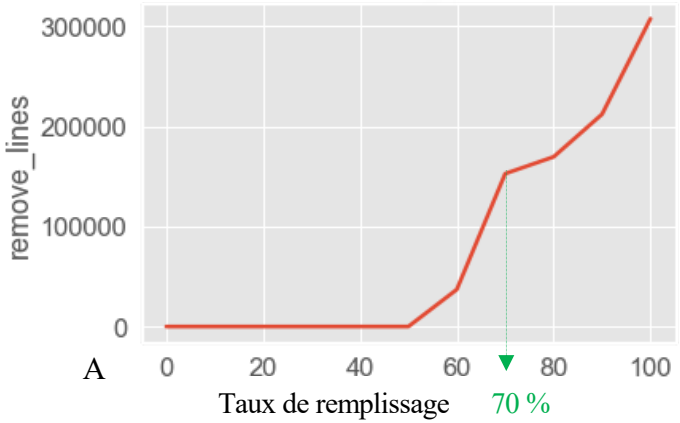


Traitement des données

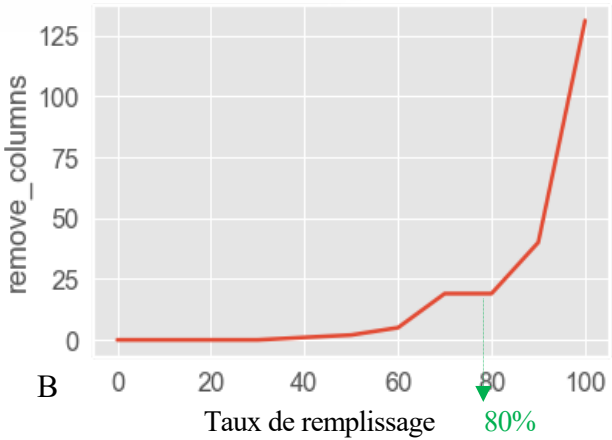
Preprocessing

- Enrichir par de nouvelles variables : pourcentage du montant des crédits sur la totalité des revenus, la durée totale du paiement des crédits, …
- Gestion des données manquantes : suppression lignes et colonnes

Nombre de lignes (A) et de colonnes supprimées (B) en fonction d'un taux de remplissage



B



- Imputation Nan, standardisation, encodage des variables catégorielles

Méthodologie : entraînement de modèles

Choix d'algorithmes

- But :
  - Problème de classification
  - Déterminer un algorithme optimal

- Comparaison de modèles :
  - Régression Logistique
  - Random Forest
  - XGBoost
  - Light-GBM (LGBM Classifier)

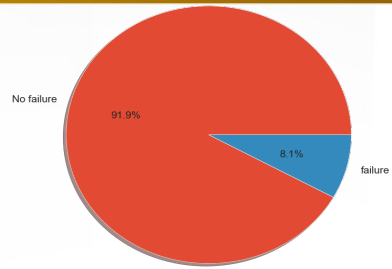
Méthode

- Division de la base de données:
  - Train (80%)
  - Test (20%)

- Optimisation hyperparamètres
  - Modèle adapté aux données
  - Train : Validation croisée
  - Test : Evaluation du modèle

Déséquilibre

- Déséquilibre entre les clients solvables et non solvables :
  - Impact performance + Prédiction faussée



- Approches pour pallier à ce déséquilibre :
  - Class weights : pénaliser les poids associés aux observations de la classe sur-représentée
  - Over-sampling : dupliquer aléatoirement des données existantes de la classe sous-représentée
  - SMOTE : créer de nouvelles données à partir des données déjà existantes pour la classe sous-représentée
  - Under-sampling : sélectionner des données de la classe sur-représentée = prendre un sous ensemble
- Chaque modèle est entraîné selon ces méthodes

Fonction coût

- classes prédites VS classes réelles
  - matrice de confusion : représentation des erreurs

		Classe réelle	
		Negatif (0)	Positif (1)
Classe prédite	Negatif (0)	TN	FP
	Positif (1)	FN	TP

- Classe 0 = Solvabilité -> négatif au refus de l'accord de prêt
- Classe 1= Difficulté de paiement -> positif au refus de l'accord de prêt

Fonction coût

		Classe réelle	
		Negatif (0)	Positif (1)
Classe prédite	Negatif (0)	TN	FP
	Positif (1)	FN	TP

- FN : Accorder un crédit à un client ne pouvant pas le rembourser par la suite = perte
- TN : Accorder un crédit à un client qui le remboursera par la suite = gain
- TP : Ne pas accorder le prêt à un client qui ne pourra pas le rembourser= ni perte, ni gain.
- FP : Ne pas accorder le prêt alors que le client pouvait rembourser = perte de client donc d'argent

Métriques d'évaluation

- Efficacité modèles :

$$\text{Accuracy} = \frac{Vrai\ positif + Vrai\ négatif}{Total}$$

$$\text{Precision} = \frac{Vrai\ positif}{Vrai\ positif + Faux\ positif}$$

$$\text{Recall} = \frac{Vrai\ positif}{Vrai\ positif + faux\ négatif}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- AUC = Aire sous la courbe ROC : capacité d'un classificateur à distinguer les classes

A Maximiser	A Minimiser
AUC, F1, Recall, TP	FN

Fonction coût

- Société de crédit = cherche à maximiser son gain d'argent

- Pénaliser impact des erreurs décision d'octroi de crédit:
  - Coefficient négatif : FN et FP
  - Coefficient positif : TN

		Classe réelle	
		Negatif (0)	Positif (1)
Classe prédite	Negatif (0)	TN	FP
	Positif (1)	FN	TP

- $gain\_total = (TN * coeff\_tn + FP * coeff\_fp + FN * coeff\_fn + TP * coeff\_tp)$

Fonction coût

- $gain = \frac{(gain\_total - gain\_min)}{(gain\_max - gain\_min)}$
- Avec :
  - $gain\_min = (TN + FP) * coeff\_fp + (TP + FN) * coeff\_fn$
  - $gain\_\max = (TN + FP) * coeff\_tn + (TP + FN) * coeff\_tp$

- Coefficient arbitraire :
  - FN ==> -100
  - TP ==> 0
  - TN ==> +10
  - FP ==> -1

		Classe réelle	
		Negatif (0)	Positif (1)
Classe prédite	Negatif (0)	TN	FP
	Positif (1)	FN	TP

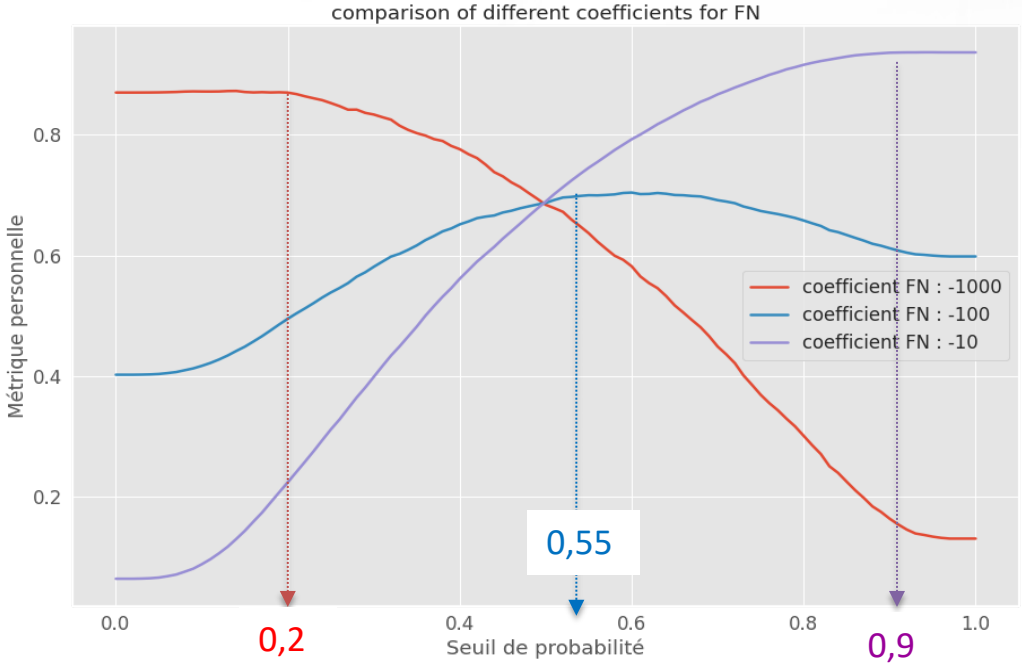
Modèle Optimum

- Meilleurs compromis entre
  - maximiser le gain
  - contraintes des métriques d'évaluation

	Modele	Accuracy	AUC	Recall class 1	F1	TP	Precision	FN	score Gain	time
0	Baseline - XGBoost	0.931432	0.742908	0.013615	0.026593	29	0.568627	2101	0.603242	1.080430
1	Class Weight - LGBMClassifier	0.741167	0.760122	0.632394	0.251588	1347	0.157030	783	0.702272	1.237642
2	RandomUnderSampler - LGBMClassifier	0.689426	0.754453	0.683099	0.232317	1455	0.139958	675	0.687163	1.307599
3	RandomOverSampler - LGBMClassifier	0.744299	0.759012	0.629108	0.252902	1340	0.158261	790	0.703110	1.275609
4	SMOTE - LGBMClassifier	0.931497	0.754088	0.016432	0.031949	35	0.573770	2095	0.604291	1.475769

Coefficient optimal

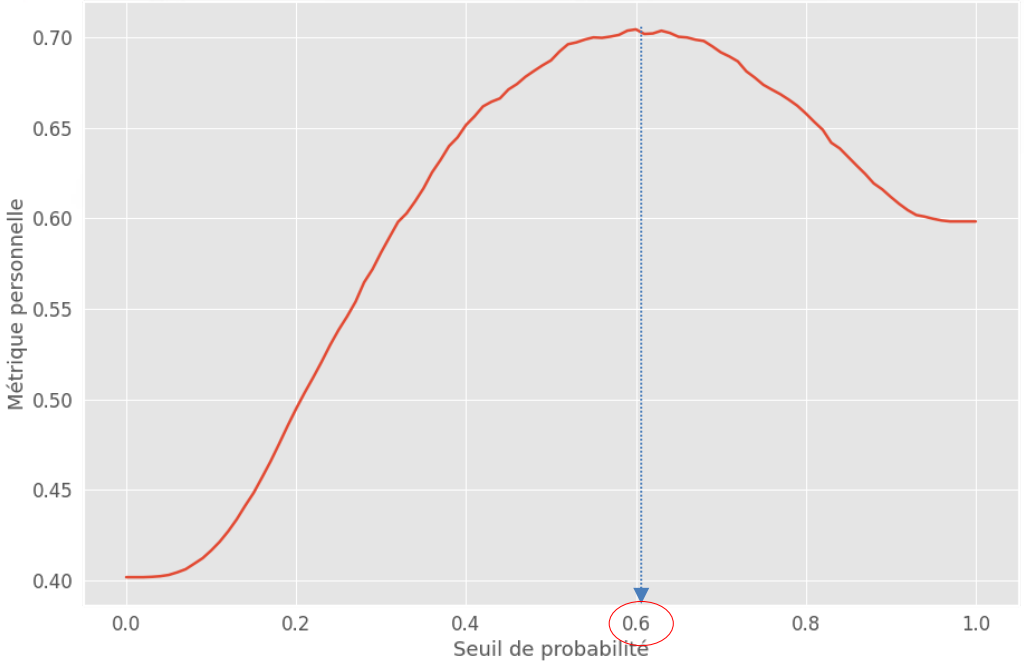
- Fonction coût : coefficients arbitraires = optimums ?
- Coefficient FN : Bon ordre de grandeur ?



Coefficient FN : -1000 = Trop Strict
Coefficient FN : -10 = Trop Laxiste
Coefficient FN : -100 = Correct

Seuil de probabilité

- Généralement seuil de probabilité : 0,5 = pas optimal



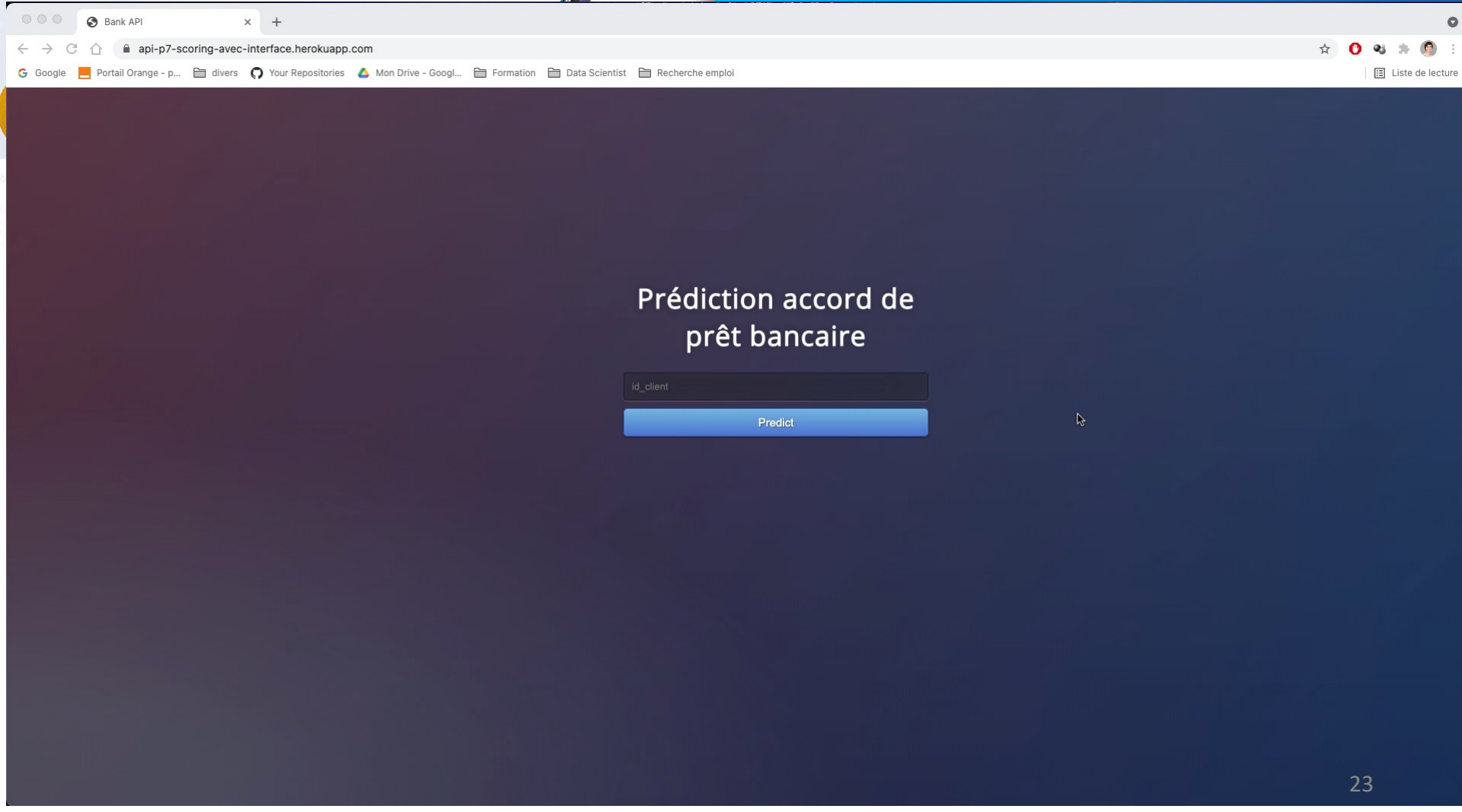
API

- Disponible à l'URL :
  - Avec interface graphique

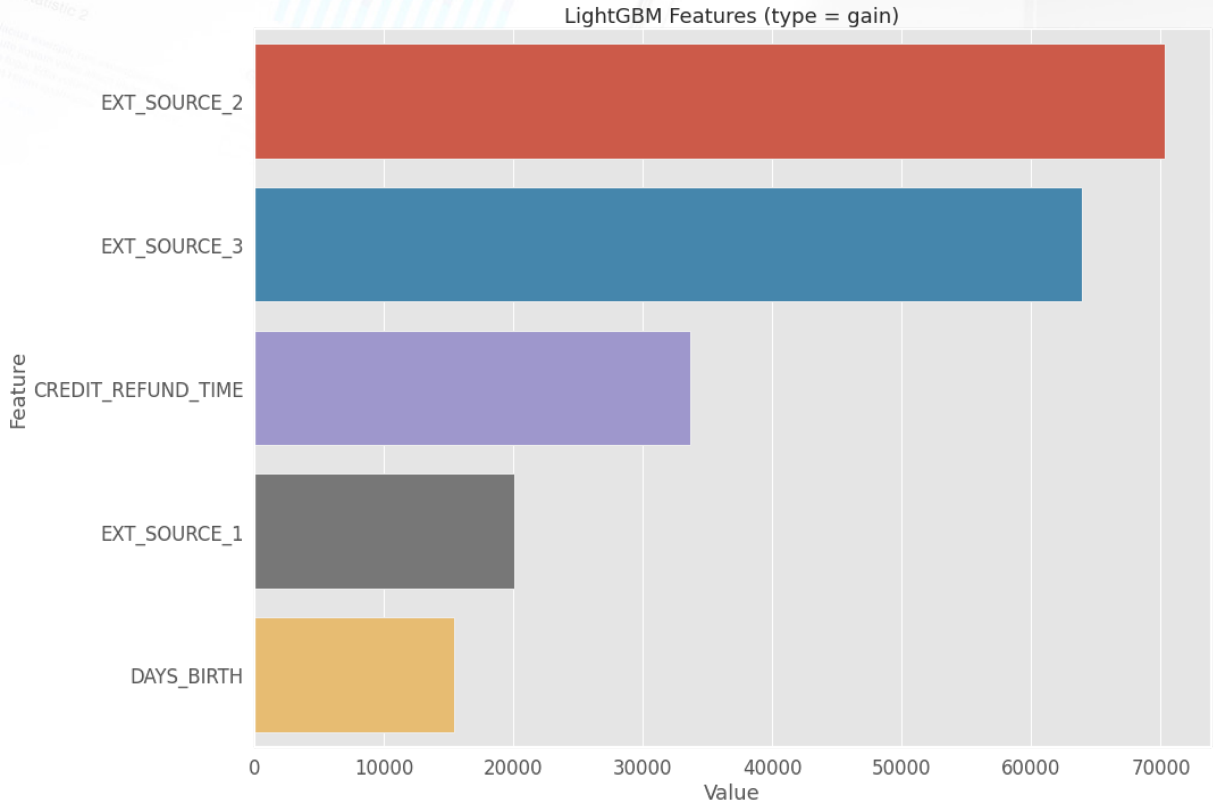
https://api-p7-scoring-avec-interface.herokuapp.com/

  - Sans interface graphique

https://api-p7-scoring-sans-interface.herokuapp.com/


Bank API
api-p7-scoring-avec-interface.herokuapp.com
Google
Portail Orange - p...
divers
Your Repositories
Mon Drive - Googl...
Formation
Data Scientist
Recherche emploi
Liste de lecture
Prédiction accord de prêt bancaire
id_client
Predict

Features Importance

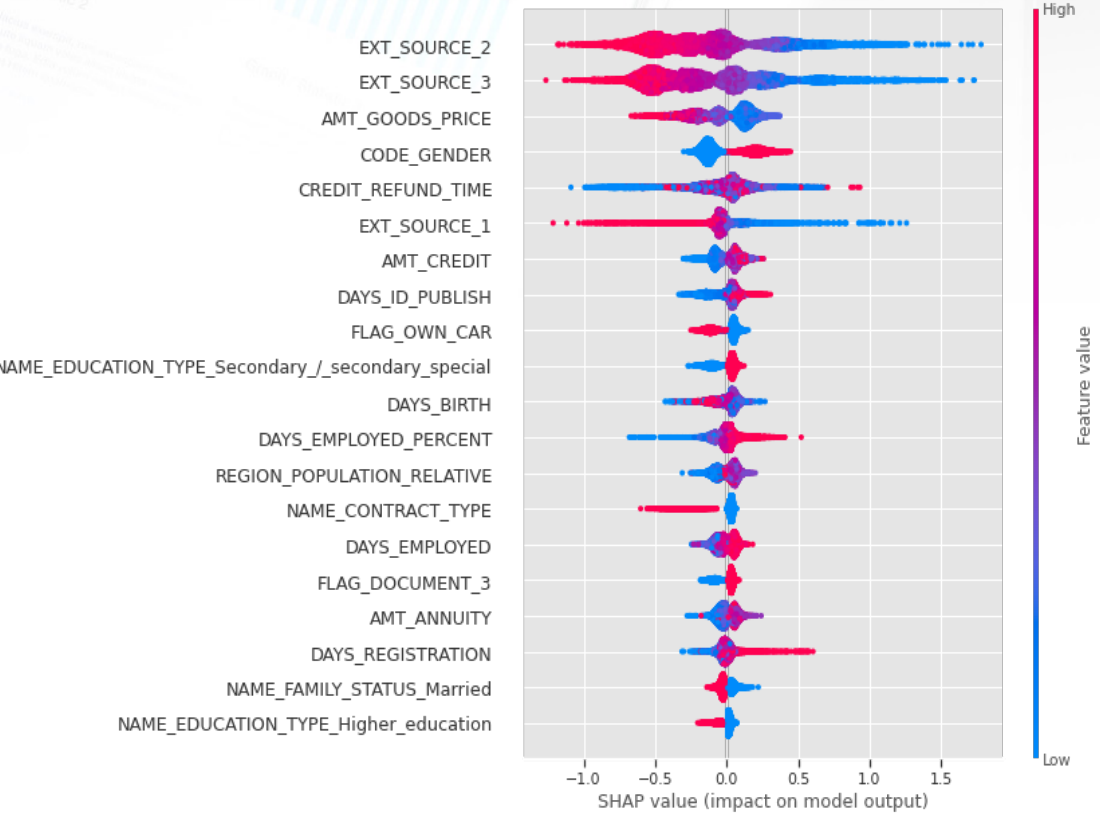


EXT_SOURCE : sources normalisées créées à partir de sources de données externes

CREDIT_REFUND_TIME : durée que va mettre un client à rembourser un prêt en années

DAYS_BIRTH : nombre de jours depuis la naissance des clients donc l'âge

Features Importance



EXT_SOURCEs : plus les valeurs sont faibles, plus la probabilité de défaut de paiement augmente donc plus il y a de chance que le prêt ne soit pas accordé

CREDIT_REFUND_TIME : plus le temps de remboursement des crédits est grand plus il y a de chance que le prêt ne soit pas accordé

Dashboard

- Disponible à l'URL :
  - https://dashboard-p7.herokuapp.com/

Analyse générale

Analyses possibles

Quelle variable voulez-vous voir ?

Accord prêt bancaire: Analyse détaillée

Cette application prédit la probabilité qu'un client de la banque "Prêt à dépenser" ne rembourse pas son prêt.

La probabilité maximale de défaut de remboursement autorisée par la banque est de : 0.6

Pour information : Liste des identifiants possibles

100001

Veuillez entrer l'Identifiant d'un client

Made with Streamlit

Limites et améliorations

- Définir plus précisément ces coefficients
- Modèle tend à être éthique :
  - Toutes les variables discriminantes n'ont pas été enlevées
  - Comparaison modèles avec et sans ces variables
  - Perte de précision des prédictions ? Perte de rentabilité pour la banque?
- Traitement des données superficiel
  - Intégrer d'autres informations sur historique de prêt
  - Créer de nouvelles variables
- Adaptation du Dashboard aux souhaits de la banque

Git

- Lien git général :
  - https://github.com/AmandineLecerfDefer/P7_Implementing_Scoring

Fin de la présentation

Merci pour votre attention