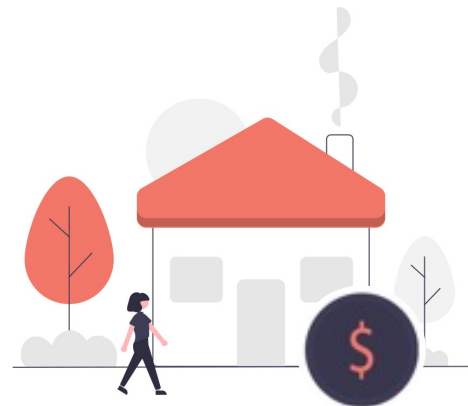


Projet 7



Implémenter un modèle de scoring dans le secteur bancaire

Home Credit Default Risk

Claire-Marie BESNIER

9 août 2021

Implémenter un modèle de scoring dans le secteur bancaire



CONTEXTE

Une entreprise souhaite développer un **modèle de scoring** pour prédire la probabilité de **défaut de paiement** d'un client à partir d'informations diverses. Ce modèle servira d'aide à la décision pour **octroyer ou non un crédit**.

L'entreprise souhaite aussi créer un **dashboard interactif** afin de communiquer de manière transparente avec ses clients.



DONNÉES

On dispose d'un jeu de données comportant plusieurs tables avec des informations sur plus de 307k clients et comprenant des informations très diverses : données comportementales, données provenant d'autres institutions financières, etc.

<https://www.kaggle.com/c/home-credit-default-risk/data>

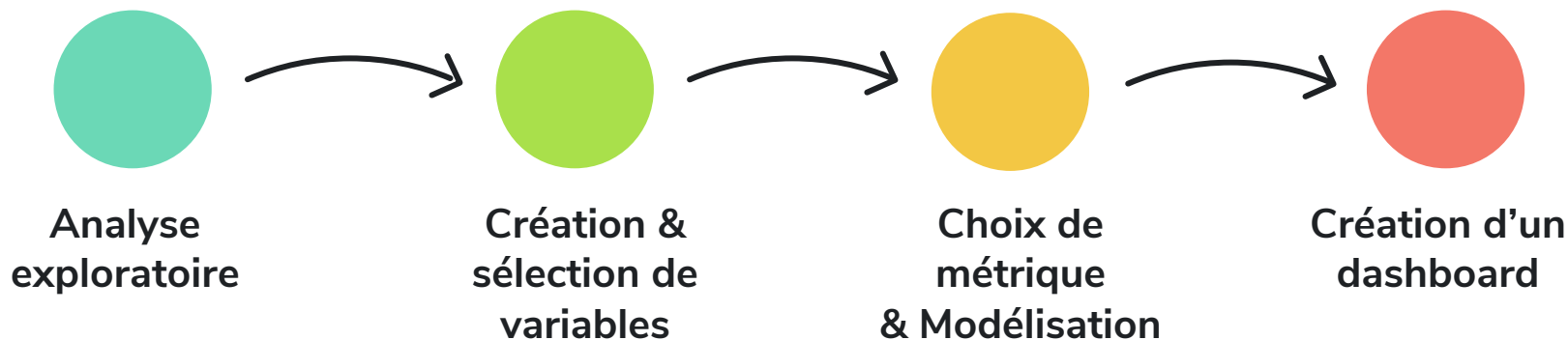


OBJECTIF

Construire un modèle de scoring permettant de prédire la probabilité de faillite d'un client

Construire un dashboard interactif à destination des gestionnaires clients permettant d'interpréter les prédictions du modèle et d'améliorer la connaissance du profil client.

Démarche

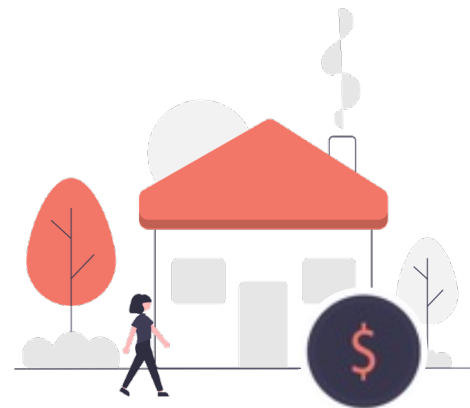


Remarques :

- **Analyse exploratoire et FE :** Sélection de kernel Kaggle pour faciliter la compréhension des données et l'analyse
- **Pre-processing, sélection et optimisation du modèle :** utilisation de la librairie Pycaret
- **Dashboard :** Utilisation de la librairie Streamlit

Plan de la présentation

1. Analyse exploratoire
2. Création et Sélection de variables
3. Modélisation
 - 3.1 Pre-processing
 - 3.2 Métriques
 - 3.3 Résultats
 - 3.4 Interprétation
 - 3.5 Optimisation du seuil
4. Dashboard
5. Conclusion



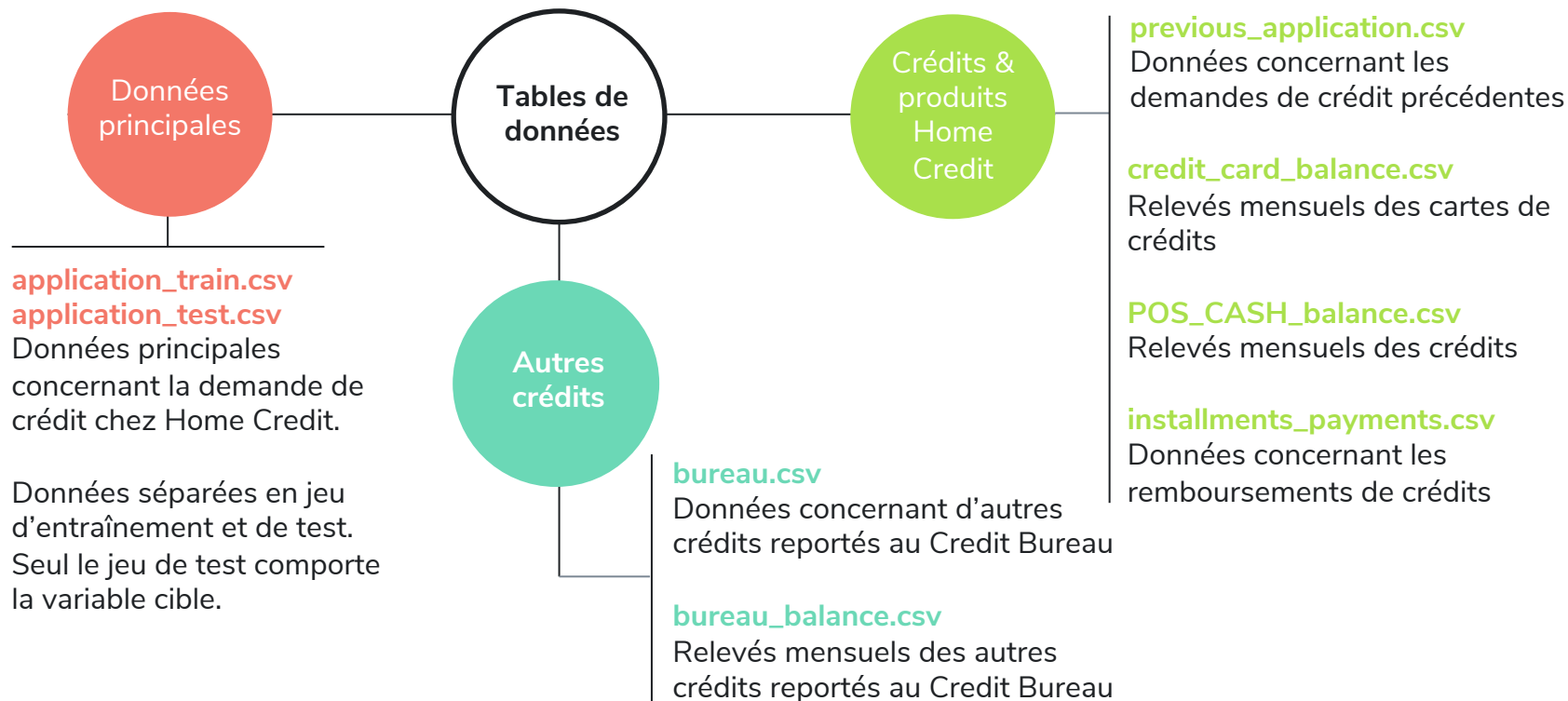









1.

Analyse exploratoire

Objectif : Mieux comprendre le jeu de données
et déterminer les variables les plus liées à la variable cible

Présentation des données



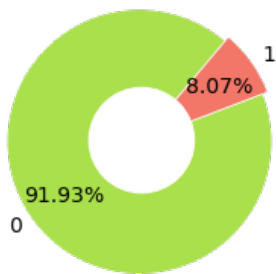
	application_train	credit_card_balance	Installment_payment	POS_CASH_balance	previous_application	bureau	bureau_balance
Lignes	307 511	3 840 312	13 605 401	10 001 358	1 670 214	1 716 428	27 299 925
Colonnes	122	23	8	8	37	17	3
Var. quanti	65	22	8	7	21	14	2
Var. quali	57	1	0	1	16	3	1
Col. Nan	67/122	9/23	2/8	2/7	16/37	7/17	0/3
Taux max Nan	70%	20%	<1%	<1%	100%	70%	-
% clients sur 307 511 clients dans application (SK_ID_CURR)	 100%	 28 %	 95 %	 94 %	 95 %	 86 %	 30 %

application_train

- Variable cible

0 : Non defaulter

1 : Defaulter

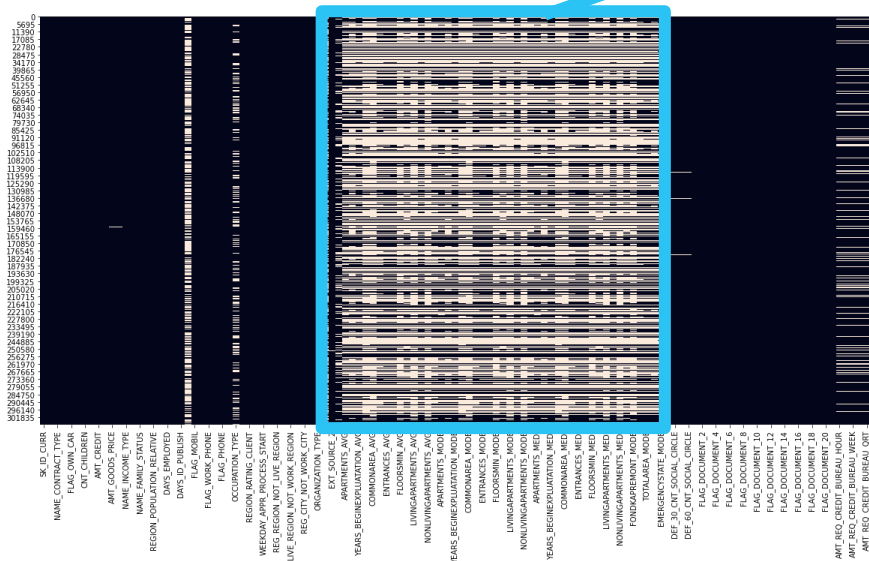


TARGET

→ Déséquilibre des classes

- Valeurs manquantes

67 colonnes concernées
avec taux max de 70%



Données normalisées
concernant le logement
du client : différentes
surfaces, nombre d'accès,
d'ascenseurs, etc.

+ statistiques associées :
AVG, MEDI, MODE

50-70% Nan

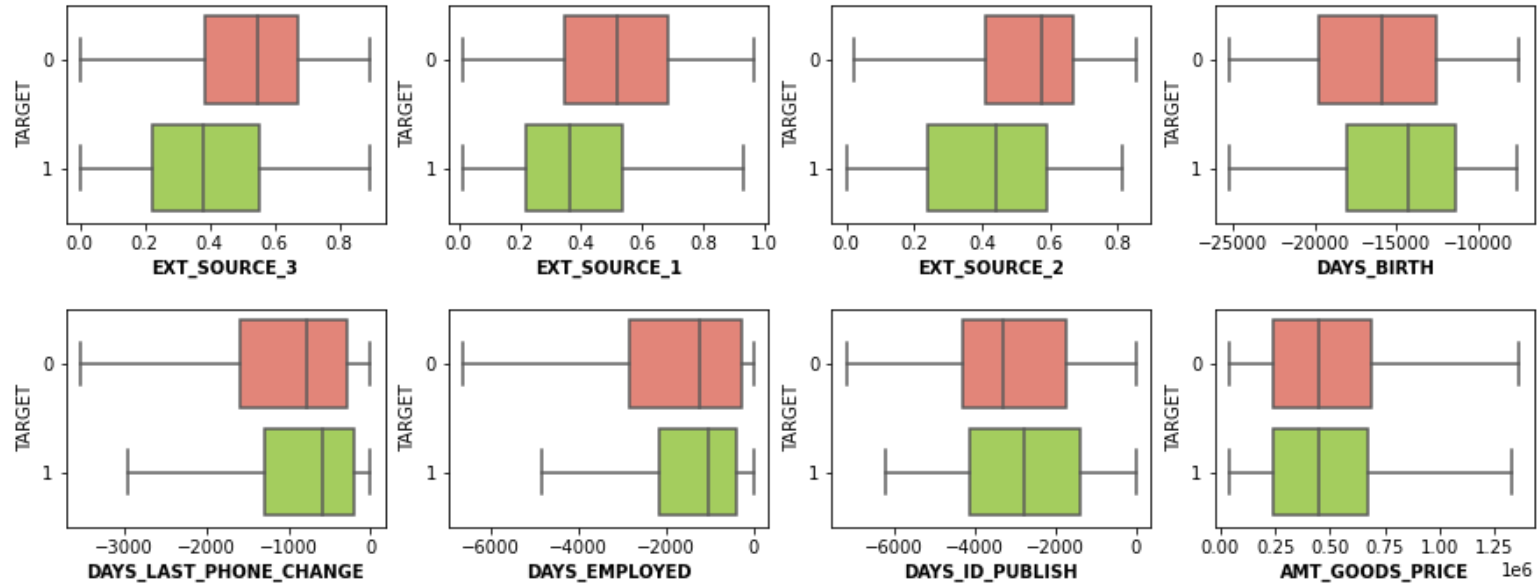
application_train

Variables quantitatives les plus corrélées à la variable cible

Variable	Corr.	Définition
EXT_SOURCE_3	0.247	Scores normalisés de sources externes
EXT_SOURCE_1	0.217	
EXT_SOURCE_2	0.213	
DAYS_BIRTH	0.102	âge du client (jours)
DAYS_LAST_PHONE_CHANGE	0.073	Nombre de jours depuis changement de téléphone
DAYS_EMPLOYED	0.072	Nombre de jours au poste actuel
DAYS_ID_PUBLISH	0.067	Nombre de jours au poste actuel
AMT_GOODS_PRICE	0.059	Montant du bien pour lequel le crédit est attribué

application_train

Variables quantitatives les plus corrélées à la variable cible



application_train

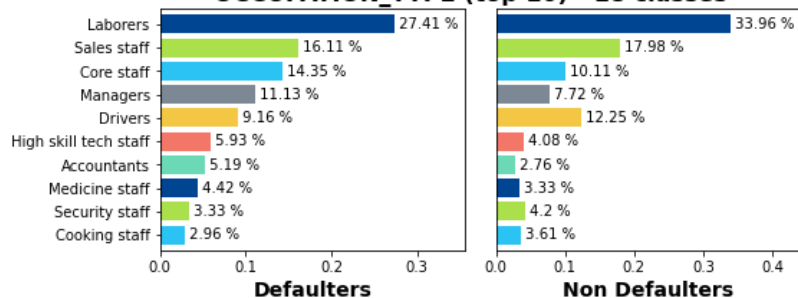
Variables qualitatives les plus corrélées à la variable cible

Variable	Corr.	Définition
OCCUPATION_TYPE	0.102	Type de profession exercée par le client
ORGANIZATION_TYPE	0.089	Type d'organismes où travaille le client
NAME_INCOME_TYPE	0.084	Type de revenus (employé, retraité, étudiant, etc.)
REG_CITY_NOT_WORK_CITY	0.079	1 si l'adresse permanente et professionnelle pas dans la même ville
FLAG_EMP_PHONE	0.072	1 si le client a fourni un numéro de téléphone professionnel, 0 sinon
REG_CITY_NOT_LIVE_CITY	0.069	1 si l'adresse permanente et l'adresse de contact sont différentes
FLAG_DOCUMENT_3	0.069	1 si le client a fourni le document 3
NAME_FAMILY_STATUS	0.056	Status familial (célibataire, marié, etc.)

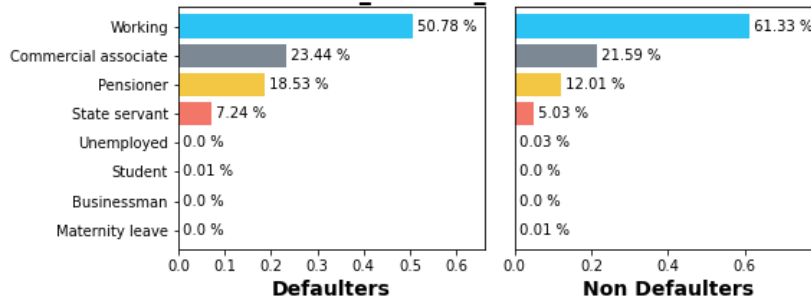
application_train

Variables qualitatives les plus corrélées à la variable cible

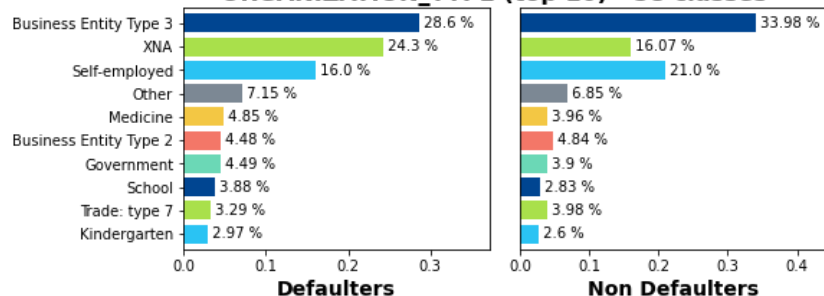
OCCUPATION_TYPE (top 10) - 18 classes



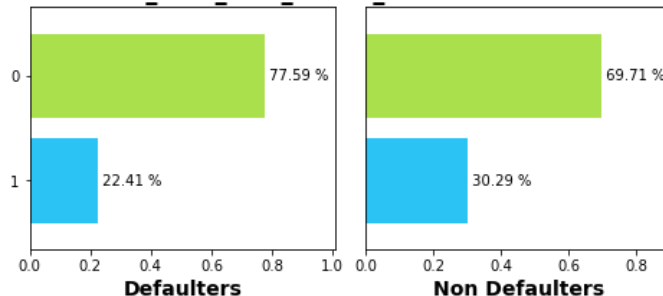
NAME_INCOME_TYPE - 8 classes



ORGANIZATION_TYPE (top 10) - 58 classes



REG_CITY_NOT_WORK_CITY - 2 classes



2.

Création & Sélection de variables

Objectif : Créer des variables supplémentaires permettant de mieux interpréter le comportement des clients Defaulters, puis sélectionner les meilleures variables



Nettoyage & Création de variables



application_train

Nettoyage

- Suppression de 47 variables concernant le logement avec 50-70% de valeurs manquantes
- CODE_GENDER : Suppression de 4 observation avec XNA
- DAYS_EMPLOYED : Remplacement de valeurs aberrantes 365243 par Nan

Création de variables

- Durée du crédit
- Ratio entre les principales variables du prêt
- Différence entre les principales variables du prêt
- Somme des différentes alertes (flags)

13
nouvelles
variables

CREDIT_TERM
CREDIT_INCOME_RATIO
ANNUITY_INCOME_RATIO
INCOME_ANNUITY_DIFF
CREDIT_GOODS_RATIO
CREDIT_GOODS_DIFF
DAYS_EMPLOYED_RATIO
FLAG_CONTACTS_SUM
CNT_NON_CHILDREN
CHILDREN_INCOME_RATIO
PER_CAPITA_INCOME
FLAG_REGIONS
SUM_FLAGS_DOCUMENTS

Nettoyage & Création de variables



bureau

Nettoyage

- Suppression des crédits vieux de plus de 50 ans
- Encodage des statuts de crédits (0/1)

Création de variables

Agrégation des données par identifiant client

- Nombre de compte et de types de comptes
- % de comptes actifs

- Somme des dettes, des crédits, des retards de paiement
- Nombre de jours moyens entre les différents crédits
- Ratio entre les différentes variables créées

14
nouvelles
variables

BUREAU_COUNT
BUREAU_TYPES_COUNT
BUREAU_ACTIVE_LOANS_PCT
BUREAU_PAST_DUE_LOANS_PCT
BUREAU_TOTAL_DEBT
BUREAU_TOTAL_CREDIT
BUREAU_TOTAL_OVERDUE
BUREAU_CREDIT_PROL_AVG
BUREAU_OVERDUE_COUNT
BUREAU_DAYS_DIFF_AVG
BUREAU_AVG_TYPES_COUNT
BUREAU_RATIO_DEBT_CREDIT
BUREAU_RATIO_OVERDUE_DEBT

Nettoyage & Création de variables



previous_application

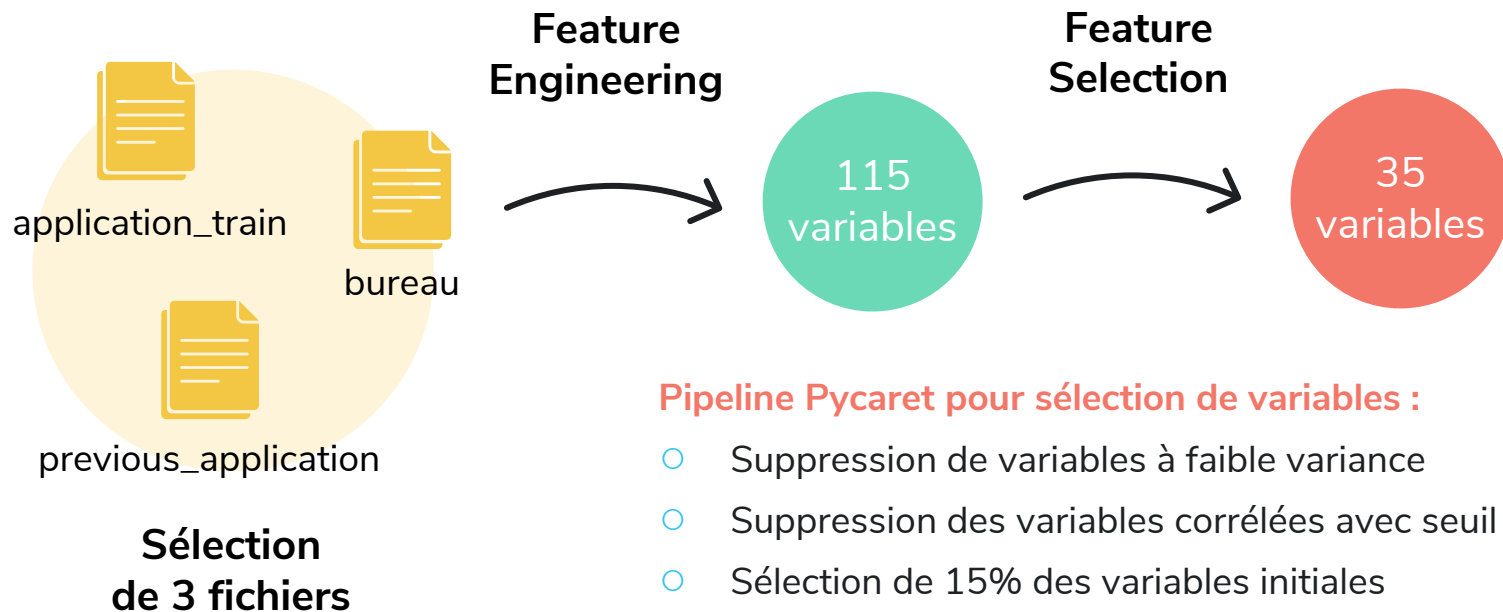
Création de variables

- Nombre de demande
Création de variables et sélection de la valeur correspondante à la dernière demande de crédit
- Données sur le crédit :
montant emprunté, annuité, montant du bien, etc.
- Ratios et différences entre les variables créées

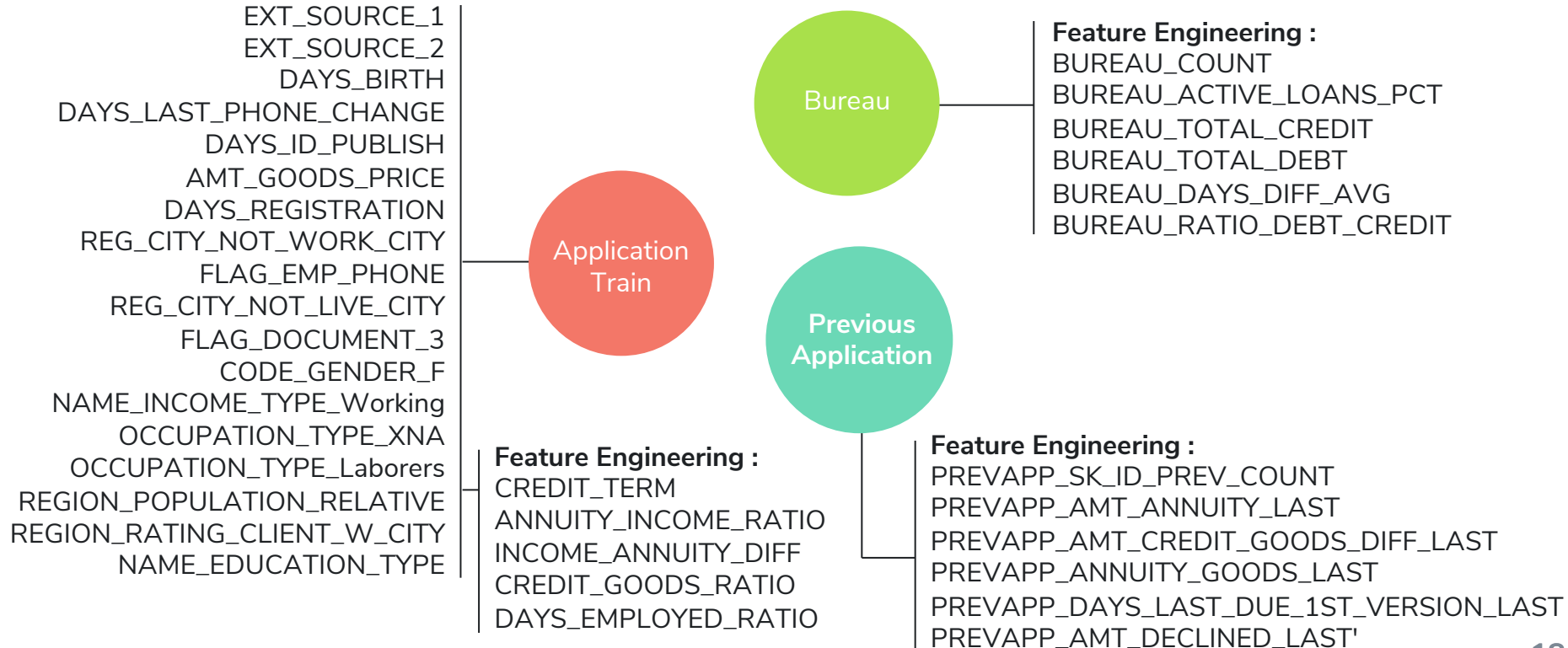
19
nouvelles
variables

PREVAPP_SK_ID_PREV_COUNT
PREVAPP_AMT_ANNUITY_LAST
PREVAPP_AMT_APPLICATION_LAST
PREVAPP_AMT_CREDIT_LAST
PREVAPP_AMT_DOWN_PAYMENT_LAST,
PREVAPP_AMT_GOODS_PRICE_LAST
PREVAPP_FLAG_LAST_APPL_PER_CONTRACT_LAST
PREVAPP_DAYS_FIRST_DUE_LAST
PREVAPP_DAYS_LAST_DUE_1ST_VERSION_LAST
PREVAPP_DAYS_LAST_DUE_LAST
PREVAPP_AMT_DECLINED_LAST
PREVAPP_AMT_CREDIT_GOODS_RATIO_LAST
PREVAPP_AMT_CREDIT_GOODS_DIFF_LAST
PREVAPP_AMT_CREDIT_APPLICATION_RATIO_LAST
PREVAPP_CREDIT_DOWNPAYMENT_RATIO_LAST,
PREVAPP_GOOD_DOWNPAYMET_RATIO_LAST
PREVAPP_ANNUITY_LAST
PREVAPP_ANNUITY_GOODS_LAST

Sélection de variables



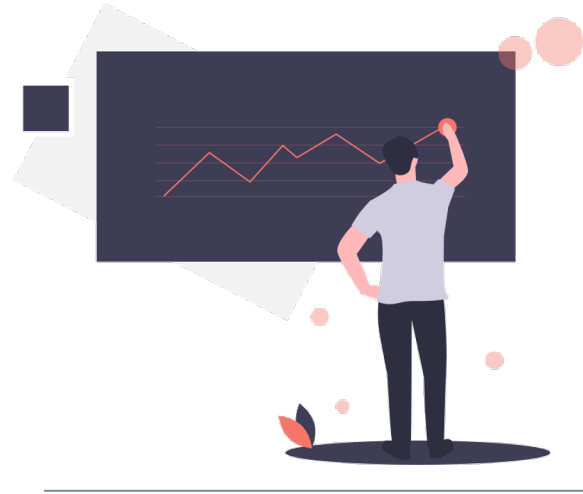
Variables retenues



3.

Pre-processing & Modélisation

Objectif : pré-traiter les données pour optimiser la modélisation, définir la métrique, entraîner et optimiser des modèles puis retenir le meilleur modèle



Preprocessing

Pipeline Pycaret

Essai de plusieurs paramétrages

- Séparation train/test
70/30, 80/20 → 80/20
- Imputation
moyenne/médiane → médiane ('XNA' pour variables qualitatives)
- Normalisation
Standard / MinMax / MaxAbs / Robust Scaler → Standard Scaler
- Transformation
PowerTransformer / Quantile Transformer → PowerTransformer
- Déséquilibre des classe
Oversampling (SMOTE), Pondération → Pondération

Métriques

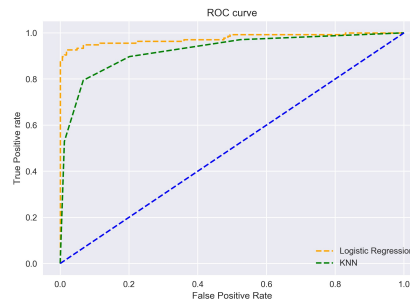
Certaines métriques classiques ne sont pas adaptées à un jeu de données présentant un déséquilibre. D'autres métriques sont plus intéressantes :

- **accuracy** non adaptée
- **précision** : taux d'observations positives parmi les observations prédites positives
- **recall** : taux de vrais positifs, qui permet de limiter les FN
- **ROC-AUC** : permet de maximiser l'aire sous la courbe ROC (sensibilité / spécificité)

TN	FP
FN	TP

Matrice de confusion

$$\text{precision} = \frac{TP}{TP+FP}$$



Courbe ROC

$$\text{recall} = \frac{TP}{TP+FN}$$

Fonction coût

Définition d'une métrique spécifique

- **Hypothèses coût**

	0	1
0	TN + 25 000 <i>Remboursement moyen</i>	FP - 250 <i>Manque à gagner (pénalité)</i>
1	FN -250 000 <i>50% montant moyen emprunté</i>	TP 0 <i>Aucun coût associé</i>

- **Coût**

$$\text{cost} = 25\,000\,TN \\ - 250\,FP - 250\,000\,FN$$

- **Score**

$$\frac{\text{cost} - \text{baseline}}{\text{best} - \text{baseline}}$$

best : coût associé à un modèle parfait, sans FN et FP

baseline : coût associé à un modèle classant tous les clients comme Non Defaulters

Modèles

Optimisation des hyper-paramètres par validation croisée (folds = 10)

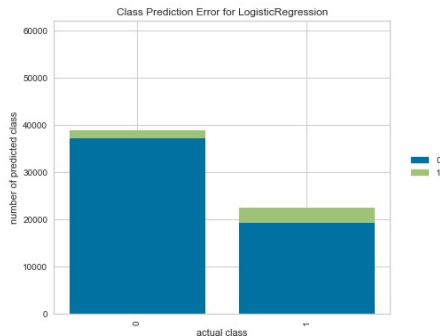
- Régression Logistique

Après optimisation

Score : 0,2735

ROC-AUC : 0,7229

37 275	19 272
1 631	3 324



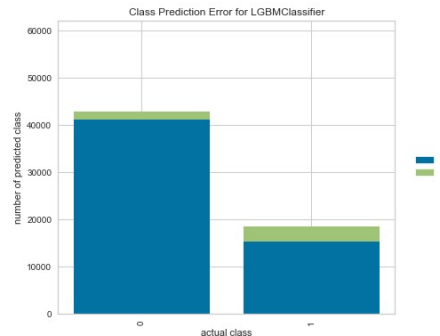
- LightGBM

Après optimisation

Score : 0,3299

ROC-AUC : 0,7520

41 167	15 380
1 767	3 188



Essai Blending et Stacking

Régression Logistique

37 275	19 272
1 631	3 324

Score : 0,2735
ROC-AUC : 0,7229

Light GBM

41 167	15 380
1 767	3 188

Score : 0,3299
ROC-AUC : 0,7520

Blending

- Méthode = soft

39 736	16 811
1 663	3 292

Score : 0,3217
ROC-AUC : 0,7468

Stacking

- Meta model = lr

38 471	18 076
1 516	3 439

Score : 0,3256
ROC-AUC : 0,7520

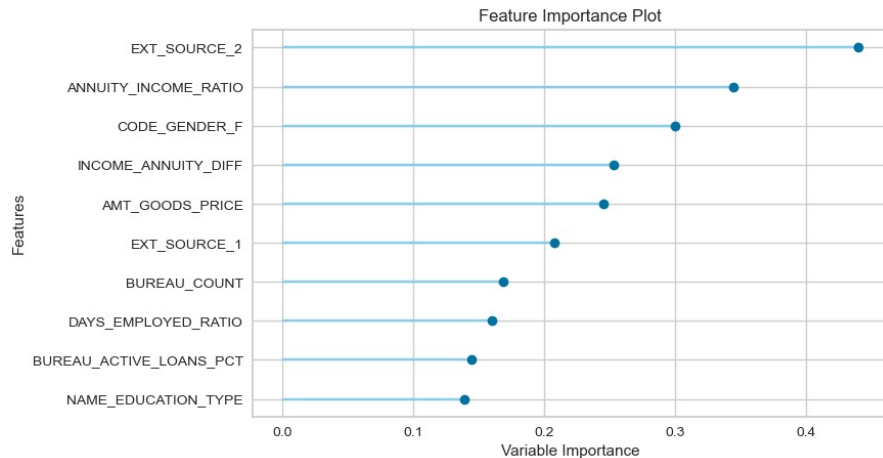
- Meta model = lgbm

40 694	15 853
1 734	3 221

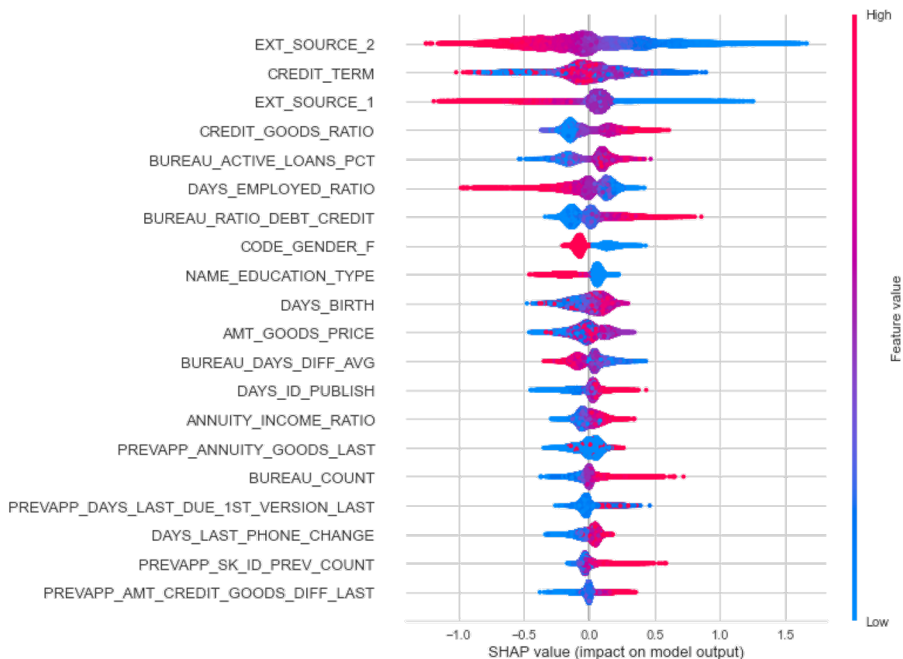
Score : 0,3269
ROC-AUC : 0,7505

Interprétation

- Régression Logistique



- LightGBM



Optimisation du seuil

Choix du modèle

- Light GBM

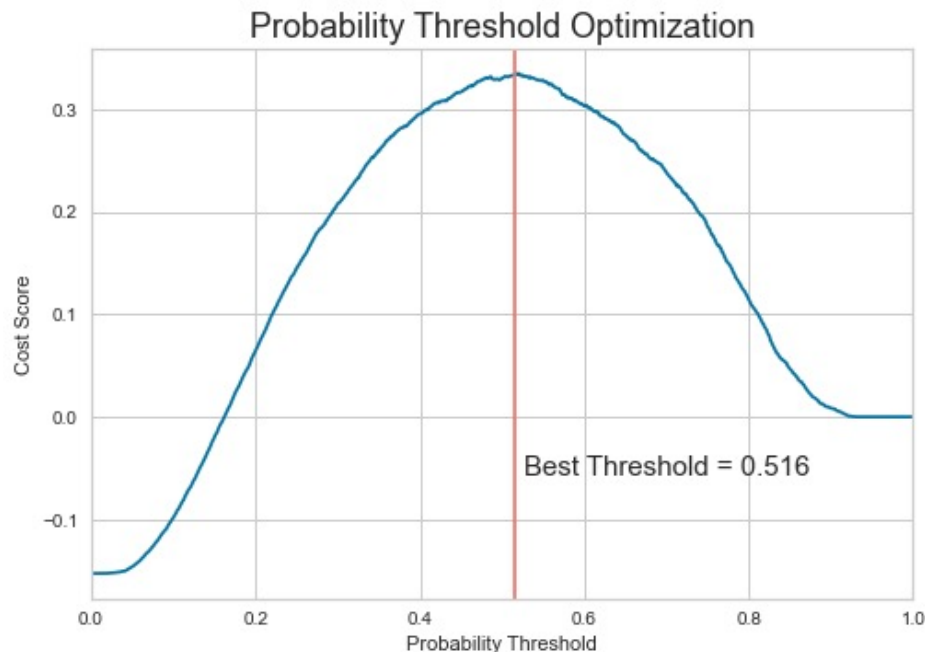
Meilleur seuil

- 0,516

42 248	14 299
1 852	3 103

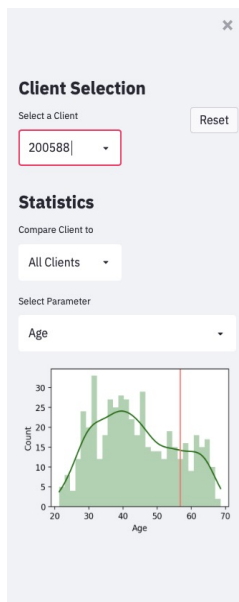
Score : 0,3348

ROC-AUC : 0,7520



Dashboard

- Sauvegarde du pre-processing et du meilleur modèle
- Création d'un dashboard interactif avec Streamlit
 - Caractéristique client
 - Prédiction de défaut de paiement
 - Comparaison des variables principales avec les autres clients



Bank Loan Dashboard

Credit Default Risk Prediction

Client Details

Income Type

Pensioner

Education Type

Secondary /
secondary special

Family Status

Married

Nb Children

0

Bureau Credit Nb
Loans

12

Bureau Credit %
Active Loans

0

Nb Previous
Application

4

Total Income

112500

Goods Price

139500

Credit Amount

205789

Annuity

10638

Scoring

CREDIT GRANTED

Threshold

0.52

Default Risk : 43 %

0.00 1.00

Dashboard : <https://share.streamlit.io/cmbesnier/credit-dashboard/main/main.py>

Github : <https://github.com/cmbesnier/credit-dashboard>

Conclusion

Synthèse

- Sélection de 35 variables (3 fichiers)
- Définition d'un score
- Optimisation de 2 modèles
- Essais de Blending et Stacking
- Choix d'un seuil
- Dashboard interactif

Modèle final

- LightGBM
- Seuil : 0,516

42 248	14 299
1 852	3 103

Matrice de confusion

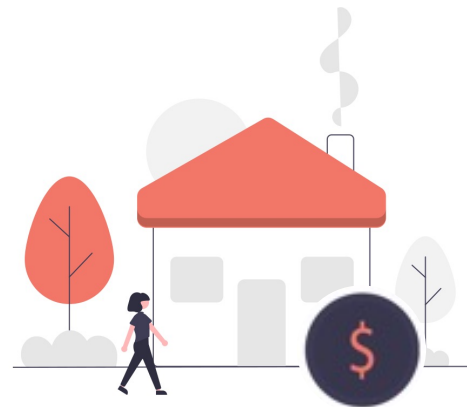
Score : 0,3348

ROC-AUC : 0,7520

Pistes d'amélioration

- Inclure un plus grand nombre de variables (plus de fichiers, FE, ...)
- Optimiser la gestion et l'imputation des valeurs manquantes
- Optimiser la fonction coût et le score avec un expert métier (coûts des FN, TP)

ANNEXES



Régression Logistique

- Hyper-paramètres retenus

LogisticRegression(C=0.472, class_weight='balanced', dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=1000, multi_class='auto', n_jobs=None, penalty='l2', random_state=123, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False)

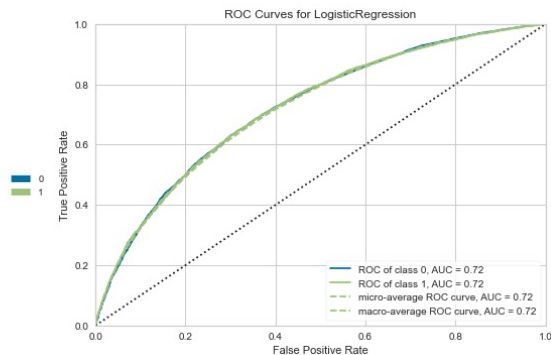
37 275	19 272
1 631	3 324

Score : 0,2735

ROC-AUC : 0,7229

Recall : 0.6664

Precision : 0,1463



Light GBM

- **Hyper-paramètres retenus**

LGBMClassifier(bagging_fraction=0.7, bagging_freq=6, boosting_type='gbdt', class_weight='balanced', colsample_bytree=1.0, feature_fraction=0.5, importance_type='split', learning_rate=0.1, max_depth=-1, min_child_samples=66, min_child_weight=0.001, min_split_gain=0.4, n_estimators=90, n_jobs=-1, num_leaves=90, objective=None, random_state=123, reg_alpha=0.0005, reg_lambda=0.1, silent=True, subsample=1.0, subsample_for_bin=200000, subsample_freq=0)

41 167	15 380
1 767	3 188

Score : 0,3299

ROC-AUC : 0,7520

Recall : 0,6434

Precision : 0,1717

