

Implémenter un modèle de Scoring

Parcours Data Scientist

ASSOIL AYAD

13 Octobre 2023

Sommaire

Introduction

Préparation des données

Exploration des données

Modélisation

Présentation du Dashboard

Conclusion

Contexte et problématique

Contexte:

L'entreprise "Prêt à dépenser" souhaite mettre en œuvre un outil de "scoring crédit" pour calculer la probabilité qu'un client rembourse son crédit.

Mission:

- Construire un modèle de scoring prédictif sur la probabilité de faillite d'un client.
- Développer un dashboard interactif pour les chargés de relation client, offrant transparence et accès aux informations personnelles du client.
- Mettre en production le modèle et le dashboard via une API.

Présentation des données

	lignes	Colonnes
application_test.csv	48744	121
application_train.csv	307511	122
bureau.csv	1716428	17
bureau_balance.csv	27299925	3
credit_card_balance.csv	3840312	23
HomeCredit_columns_description.csv	219	5
installments_payments.csv	13605401	8
POS_CASH_balance.csv	10001358	8
previous_application.csv	1670214	37
sample_submission.csv	48744	2

Présentation des données

- Inspiration pour le traitement des données : Kaggle : Home Credit Default Risk
- Dataset final :
 - Nombre de clients : 307511
 - Nombre de variables : 122 (incluant âge, sexe, revenus, historique de crédits, etc.)
- Variable cible, 'TARGET' : Représente la présence ou non d'un défaut de paiement par le client.

Préparation des données

Enrichissement des features par les ratios:

- Ratio des jours d'emploi par rapport aux jours de naissance.
- Ratio entre revenu total et montant du crédit.
- Revenu moyen par personne dans le ménage.
- Ratio entre l'annuité du prêt et le revenu total.
- Ratio entre l'annuité du prêt et le montant du crédit.
- Produit du ratio montant du crédit sur annuité par l'âge en jours.
- Ratio entre revenu total et nombre de membres de la famille.

Préparation des données

Traitements appliqués:

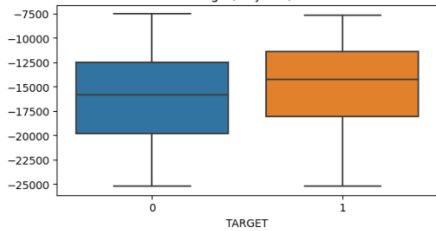
- Imputation par la médiane des variables numériques.
- Imputation "most frequent" pour les variables catégorielles.
- Encodage des variables catégorielles.
- Standardisation des features numériques.

Dimensions des données:

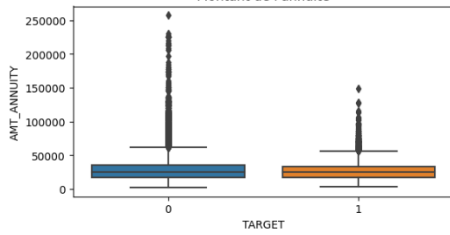
- Ensemble total : (307505, 261)
- Ensemble d'entraînement : (245999, 260)
- Ensemble de test : (61500, 260)

Analyse exploratoire des données

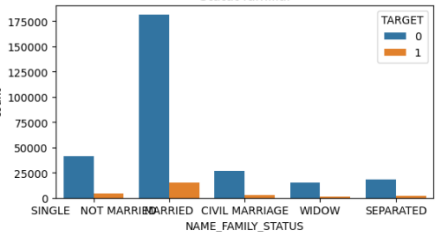
Âge (en jours)



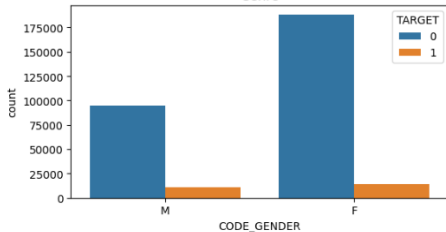
Montant de l'annuité



Statut familial



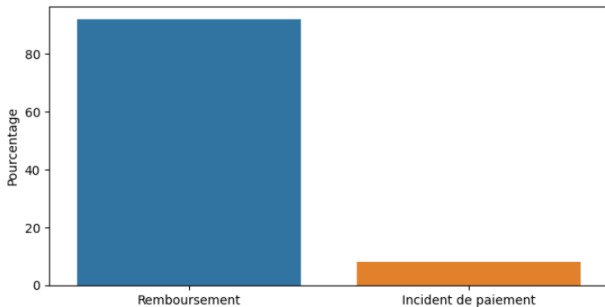
Genre



Choix des Modèles et Problème Identifié

Modèles explorés: Dummy Regression (Base de référence) - Régression Logistique - RandomForestClassifier - GradientBoosting Classifier - XGBClassifier .

Problème rencontré: La modélisation standard n'est pas adaptée à cause du déséquilibre des classes pour la variable 'Target'.



Approches pour les Ensembles Déséquilibrés et Choix de Métrique

Techniques employées:

- **Class-Weight:** Poids plus élevé à la classe sous-représentée.
- **Under-sampling:** Réduction de la classe sur-représentée.
- **Over-Sampling:** Augmentation de la classe sous-représentée.

Métrique de Performance: Une métrique adaptée au métier.

Modèle Final et Évaluation

Modèle final choisi:

- XGBClassifier avec paramètres: {'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100} et undersampling.

	AUC	Precision	Recall	F1-score	Custom Score
class_weight	0.747	0.173	0.496	0.256	-37126
Over_sampling	0.776	1.000	0.278	0.435	-36150
Under_sampling	0.752	0.154	0.592	0.244	-36728

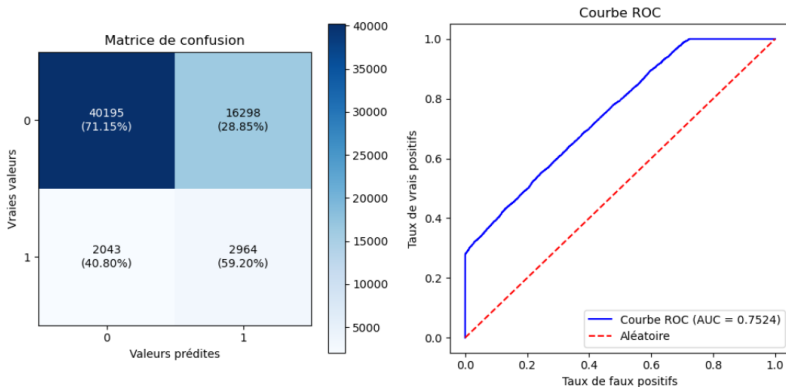
Modèle Final et Évaluation

	TN	FP	FN	TP
class_weight	44597	11896	2523	2484
Over_sampling	56493	0	3615	1392
Under_sampling	40195	16298	2043	2964

Métrique d'évaluation: Le coût d'un FN (Faux Négatif) est dix fois supérieur au coût d'un FP (Faux Positif).

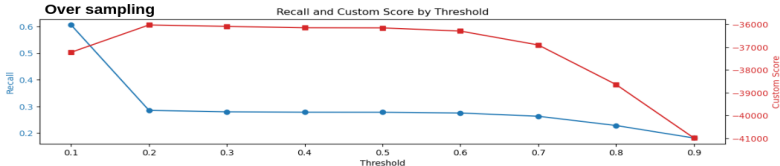
Modèle Final et Évaluation

Modèle XGBClassifier avec rééquilibrage "UnderSampling"

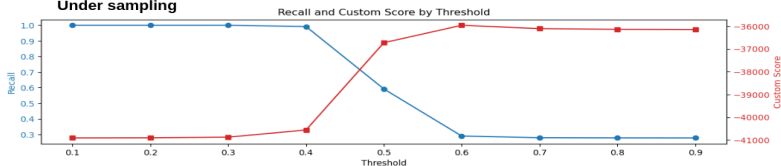


Modification du seuil

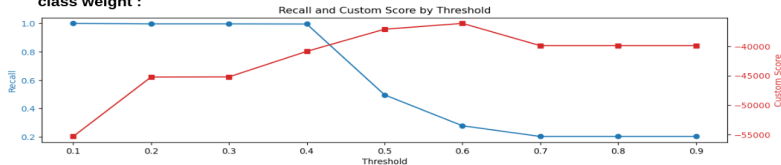
Over sampling



Under sampling

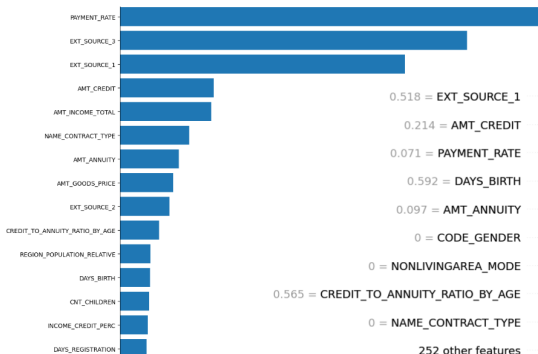


class weight :

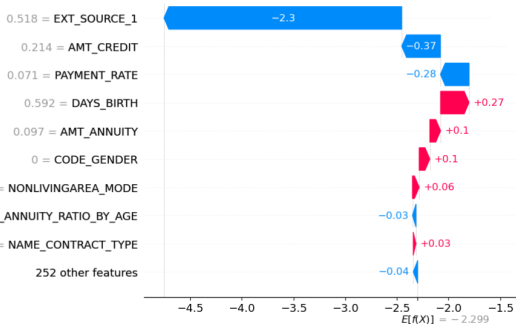


Features importance






Importance globale



Importance locale pour le client 1



Outils utilisés

Outil	Utilisation
 git	Système de contrôle et de suivi de version pour le code du projet.
 Streamlit	Développer le dashboard interactif.
 Flask	API qui renvoie les résultats de la prédiction
 aws	Héberger et déployer l'API et le dashboard.
 EVIDENTLY AI	Détecter le Data Drift en production.

Déploiement en local

Servir le modèle avec Mlflow

```

ayad@ayad-lenovo: ~/projet7

ayad@ayad-lenovo: ~/projet7
(base) ayad@ayad-lenovo: $ conda activate projet7
(projet7) ayad@ayad-lenovo: $ cd projet7
(projet7) ayad@ayad-lenovo:~/projet7$ mlflow models serve -m mlflow_model
Downloading artifacts: 100%|██████████| 1/1 [00:00<00:00, 9868.95it/s]
2023/10/04 16:29:18 INFO mlflow.models.flavor_backend_registry: Selected backend for flavor 'python_function'
2023/10/04 16:29:18 INFO mlflow.utils.virtualenv: Installing python 3.8.18 if it does not exist
2023/10/04 16:29:18 INFO mlflow.utils.environment: Environment /home/ayad/.mlflow/envs/mlflow-d530668635e0baeb09de804a637875e0
2023/10/04 16:29:18 INFO mlflow.utils.environment: == Running command '['bash', '-c', 'source /home/ayad/.mlflow/envs/mlflow
2023/10/04 16:29:18 INFO mlflow.utils.environment: == Running command '['bash', '-c', 'source /home/ayad/.mlflow/envs/mlflow
ut=60 -b 127.0.0.1:5000 -w 1 $(Gunicorn CMD_ARGS) -- mlflow.pyfunc.scoring_server.wsgi:app']
[2023-10-04 16:29:18 +0200] [405845] [INFO] Starting gunicorn 21.2.0
[2023-10-04 16:29:18 +0200] [405845] [INFO] Listening at: http://127.0.0.1:5000 (405845)
[2023-10-04 16:29:18 +0200] [405845] [INFO] Using worker: sync
[2023-10-04 16:29:18 +0200] [405849] [INFO] Booting worker with pid: 405849
/home/ayad/.mlflow/envs/mlflow-d530668635e0baeb09de804a637875e058b2a7fc/lib/python3.8/site-packages/mlflow/models/utils.py:49
ult of calling 'frame.insert' many times, which has poor performance. Consider joining all columns at once using pd.concat(a
new_pf_input[x] = _enforce_mlflow_datatype(x, pf_input[x], input_types[x])
/home/ayad/.mlflow/envs/mlflow-d530668635e0baeb09de804a637875e058b2a7fc/lib/python3.8/site-packages/mlflow/models/utils.py:49

```

Lancer le Dashboard

```

ayad@ayad-lenovo: ~/projet7

ayad@ayad-lenovo:~/projet7$ conda activate projet7
(projet7) ayad@ayad-lenovo:~/projet7$ streamlit run dashboard_new.py

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.1.31:8501

2023-10-04 16:33:26.590 Uncaught app exception
Traceback (most recent call last):
  File "/home/ayad/anaconda3/envs/projet7/lib/python3.8/site-packages/pandas/core/indexes/base.py", line 3653, in get_loc
    return self._engine.get_loc(casted_key)
  File "pandas/_libs/index.pyx", line 147, in pandas._libs.index.IndexEngine.get_loc
  File "pandas/_libs/index.pyx", line 176, in pandas._libs.index.IndexEngine.get_loc
  File "pandas/_libs/hashtable_class_helper.pxi", line 7080, in pandas._libs.hashtable.PyObjectHashTable.get_item
  File "pandas/_libs/hashtable_class_helper.pxi", line 7088, in pandas._libs.hashtable.PyObjectHashTable.get_item
KeyError: 'SK_ID_CURR'

```

- Sélection par ID client pour prédiction.
- Résultat : Crédit octroyé ou non.
- Affichage des variables explicatives de la prédiction.

localhost:8501

Credit Default Prediction

Select Client ID

100057

Predict

Félicitations! Votre demande de crédit a été acceptée pour les raisons suivantes:

	6
EXT_SOURCE_3	-2.0339
EXT_SOURCE_1	0.9596
PAYMENT_RATE	-0.547
AMT_REQ_CREDIT_BUREAU_YEAR	-0.1661
AMT_CREDIT	-0.1293
CREDIT_TO_ANNUITY_RATIO_BY_AGE	-0.1167
AMT_ANNUITY	0.116
DAYS_BIRTH	0.0996
INCOME_CREDIT_PERC	-0.0869
AMT_GOODS_PRICE	-0.037

localhost:8501

Credit Default Prediction

Select Client ID

100057

Predict

Votre demande de crédit a été refusée en raison des raisons suivantes:

	2
EXT_SOURCE_1	-2.6204
DAYS_BIRTH	-0.4906
EXT_SOURCE_3	0.2187
PAYMENT_RATE	0.1808
ORGANIZATION_TYPE_Tradetype3	0.1348
AMT_ANNUITY	-0.1037
NAME_CONTRACT_TYPE	0.0644
CODE_GENDER	0.0504
AMT_GOODS_PRICE	0.0167
AMT_REQ_CREDIT_BUREAU_YEAR	0.0158

Déploiement du Dashboard sur AWS EC2

1. Création d'une instance EC2 :

Choix d'un type d'instance adapté à l'application, 't2.micro'.

2. Configuration des groupes de sécurité :

3. Connexion SSH à EC2 :

Utilisation de la clé '.pem' pour se connecter à l'instance EC2.

4. Installation des dépendances :

Installation de Python et des bibliothèques associées.

5. Transfert de l'application :

Utilisation de 'scp' pour transférer l'application du local vers EC2.

6. Lancement de l'application :

7. Accès via le navigateur :

Utilisation de l'adresse IP publique de l'instance EC2 suivie du port d'écoute.

Conclusion

Thématiques abordées :

- Découverte de l'environnement Mlflow.
- Prise en main de AWS
- Prise de conscience des règles de sécurité.

Difficultés rencontrées :

- Difficulté d'interprétation des variables internes.
- Modélisation en fonction du temps de calcul.
- Choix non évident du prestataire Cloud.
- Configuration complexe de l'environnement du travail.
- Frais élevés.

Questions

Merci de votre attention !
Des questions ?