Prêt à dépenser

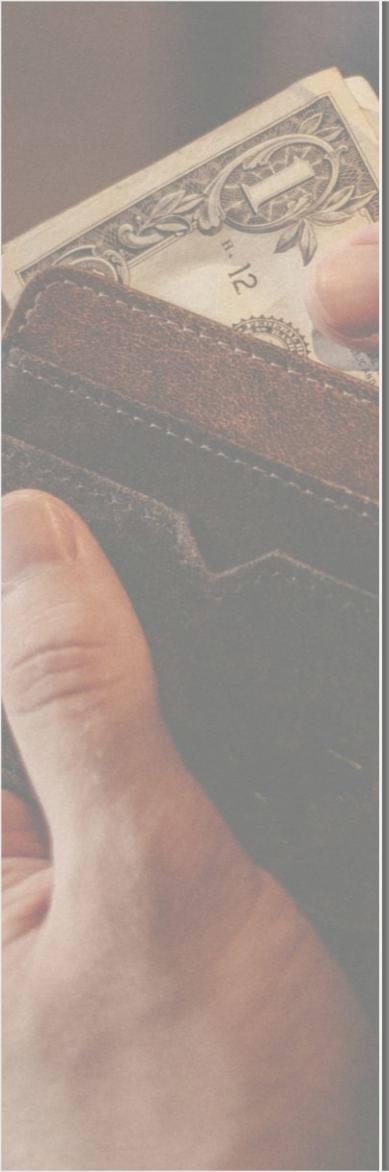
Projet 7 – Raphaël Koudache--Louvet

Analyse de la solvabilité des demandeurs de crédits immobiliers

Classification binaire

Réalisation et déploiement d'un dashboard à destination du banquier

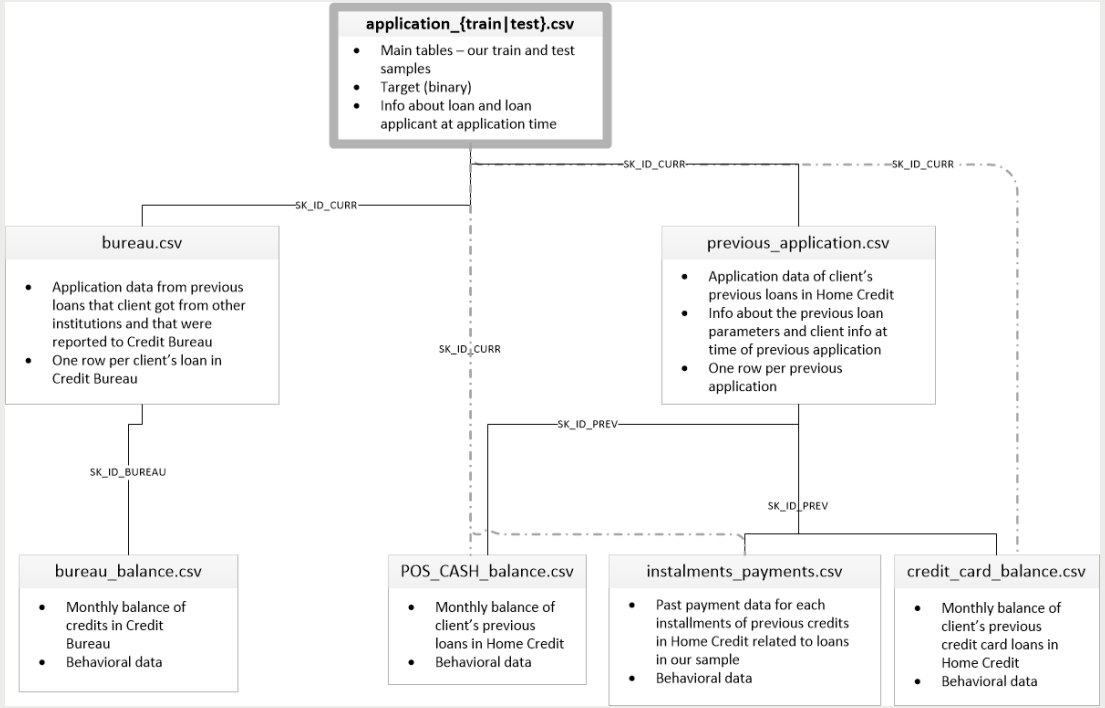
I - Analyse Exploratoire des données (EDA)

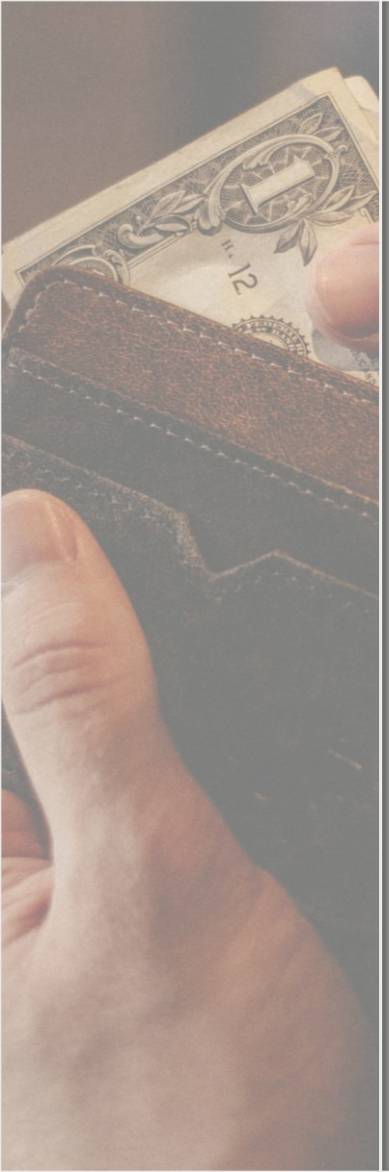


Vue d'ensemble

Données fragmentées (plusieurs .csv)

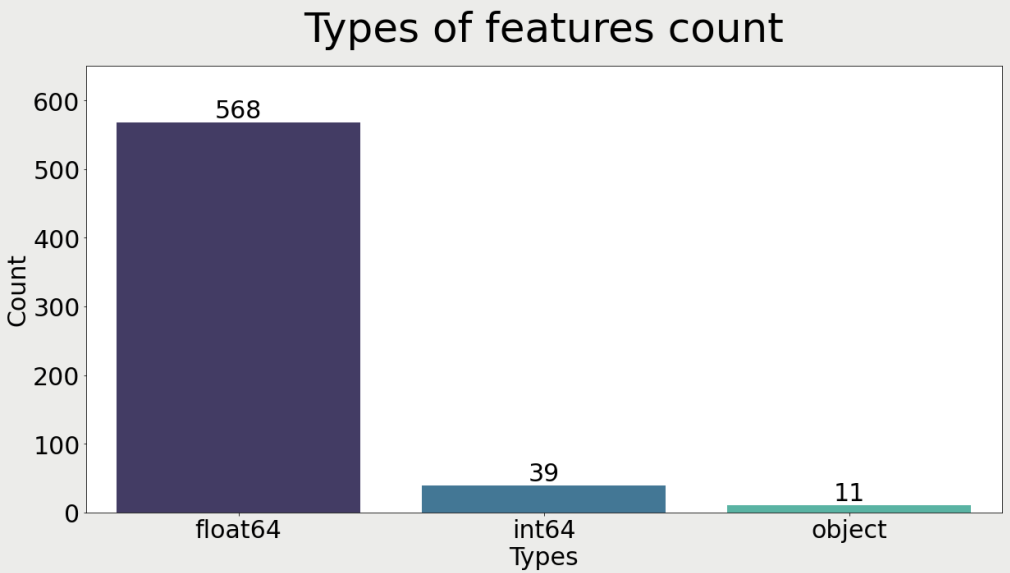
Kernel Kaggle (Imputations, Aggrégations, Jointure)

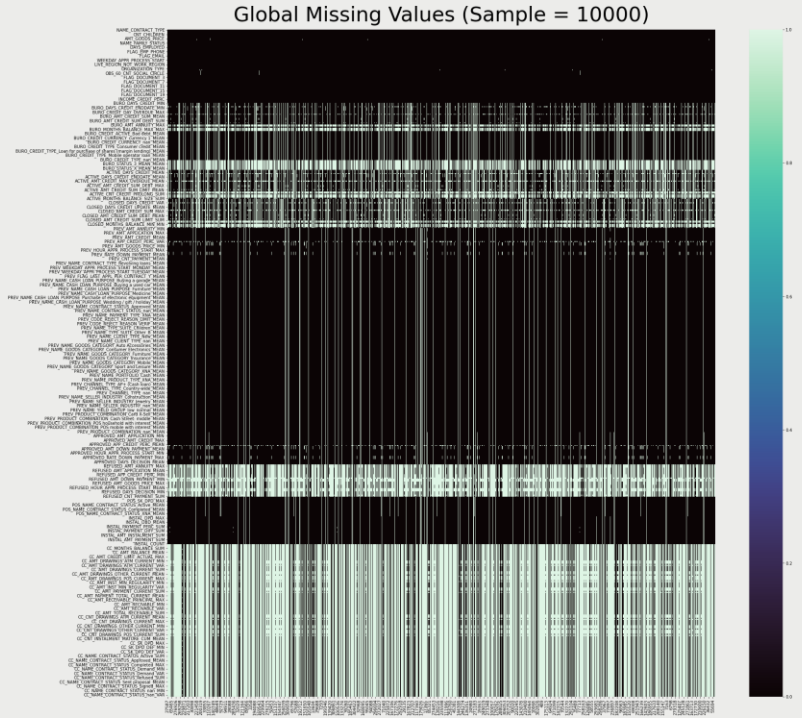


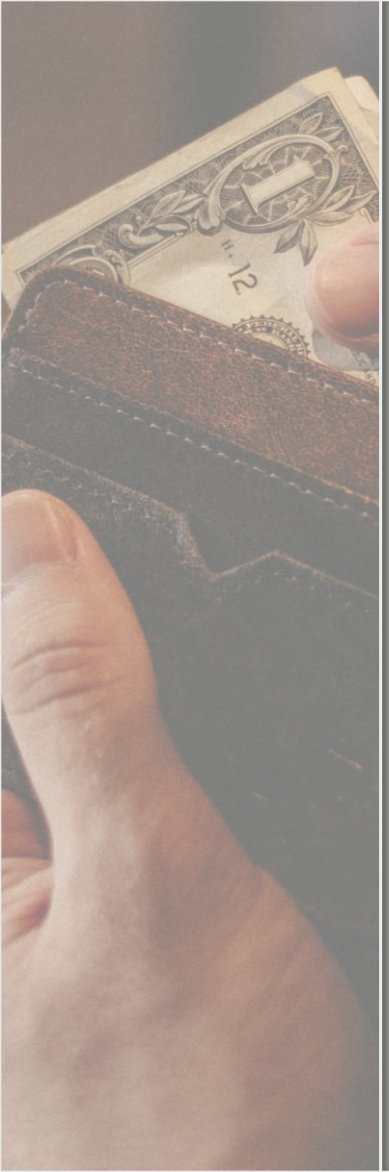


Vue d’ensemble

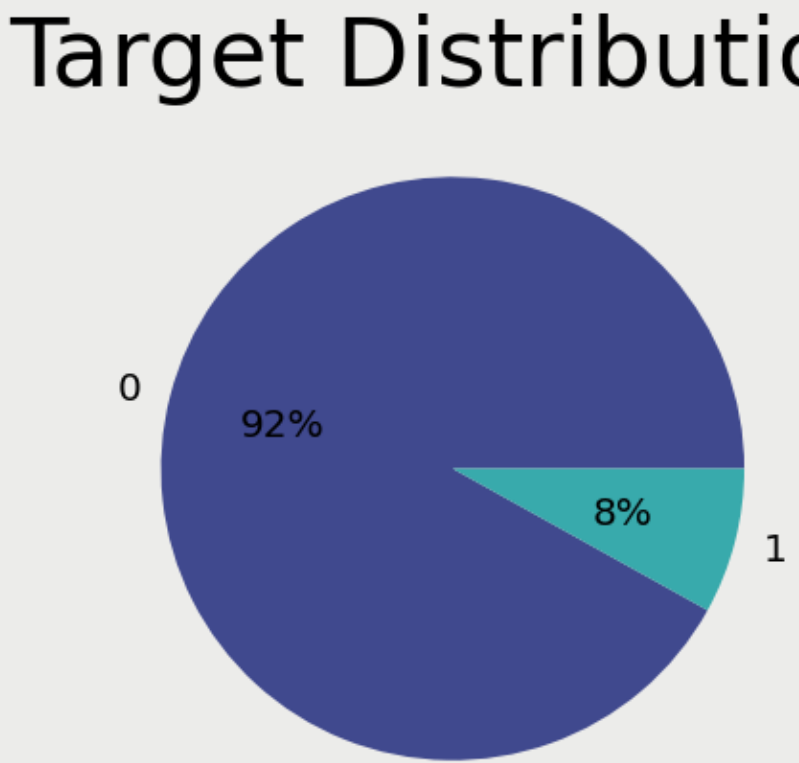
Shape (307511,620)

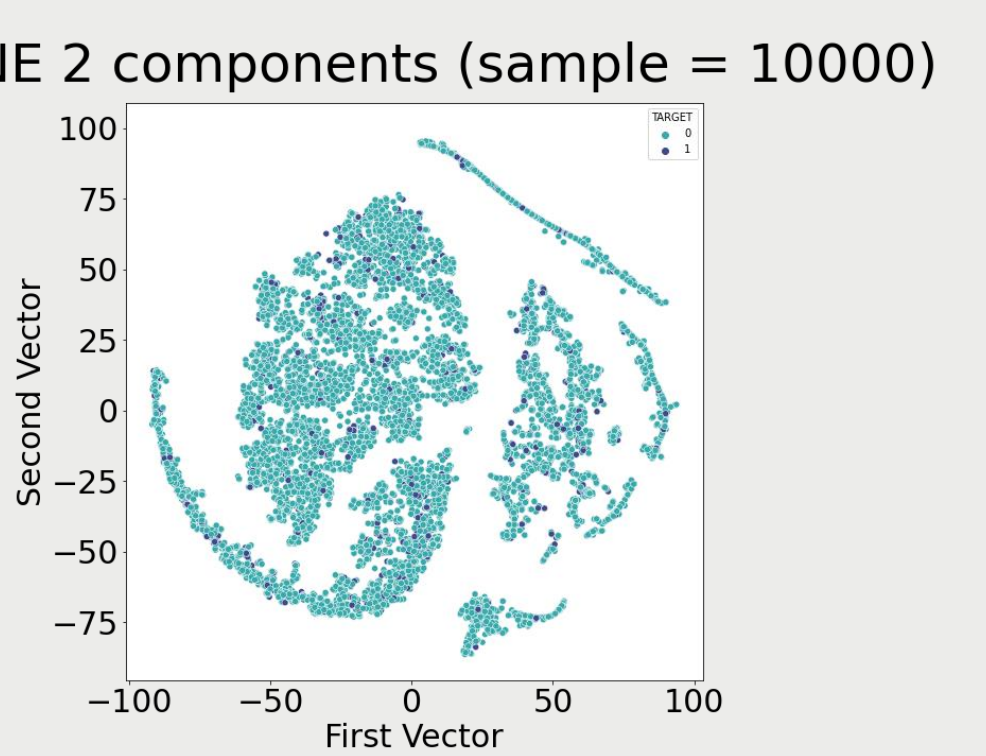


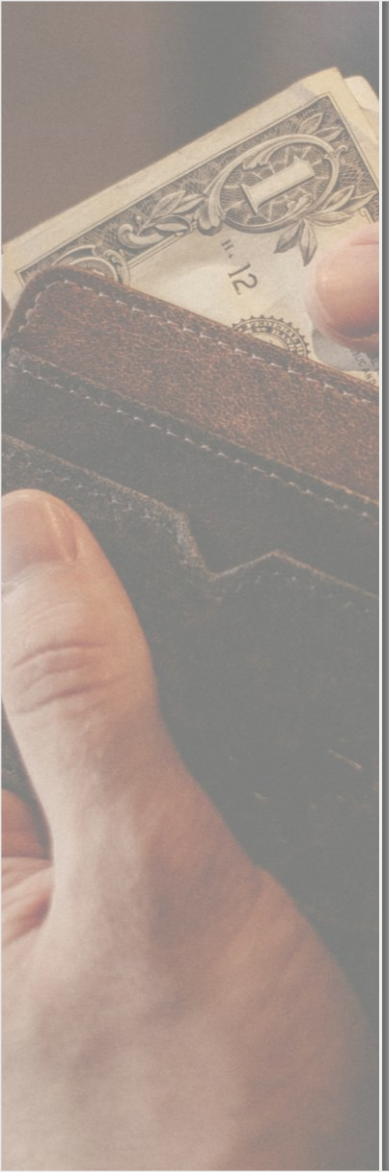




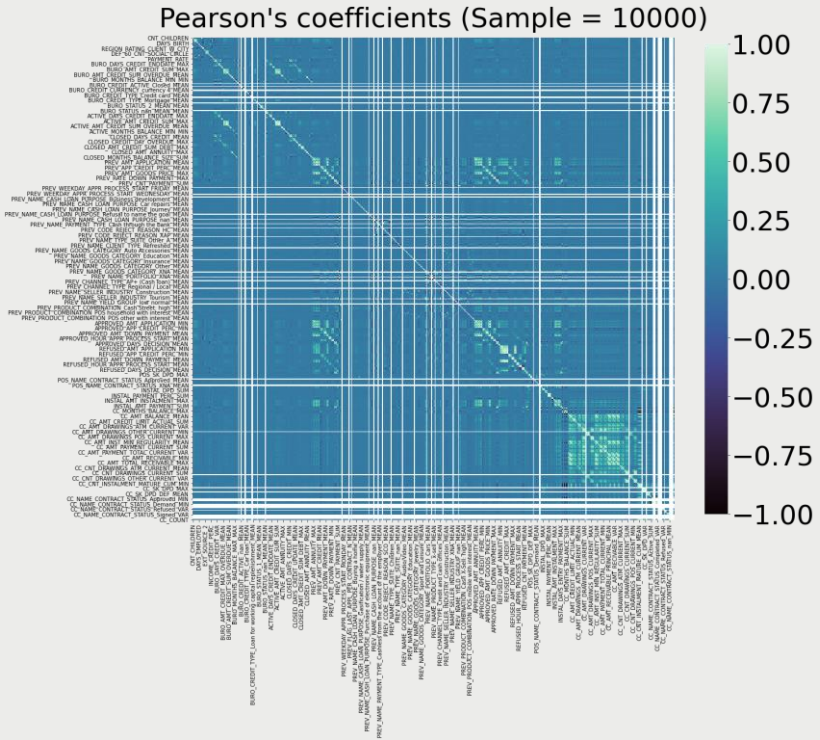
Des cibles éparses et non balancées

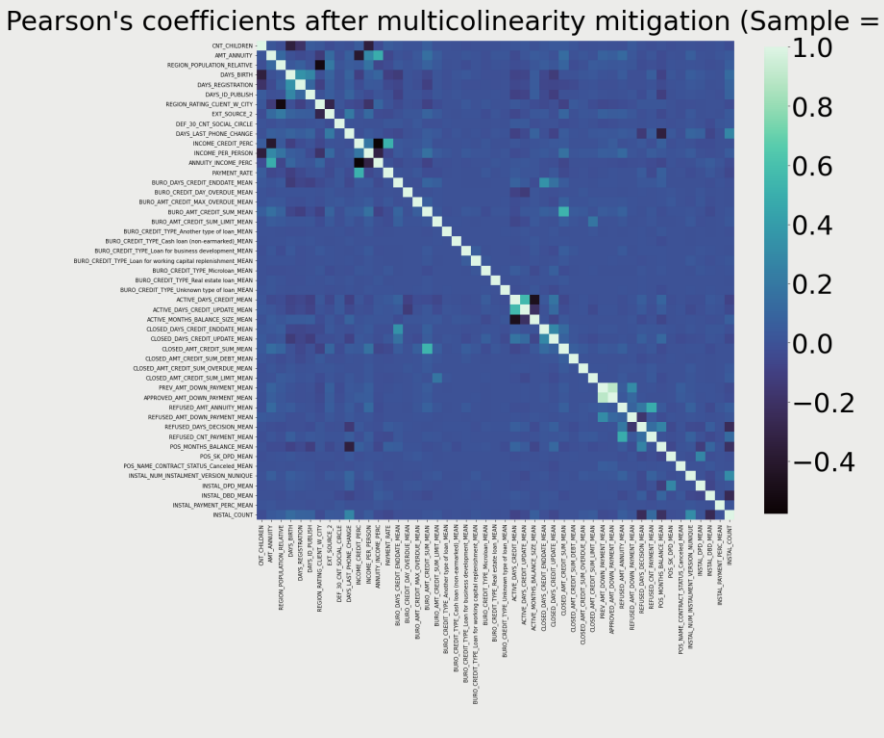






Une Multicolinéarité forte à traiter





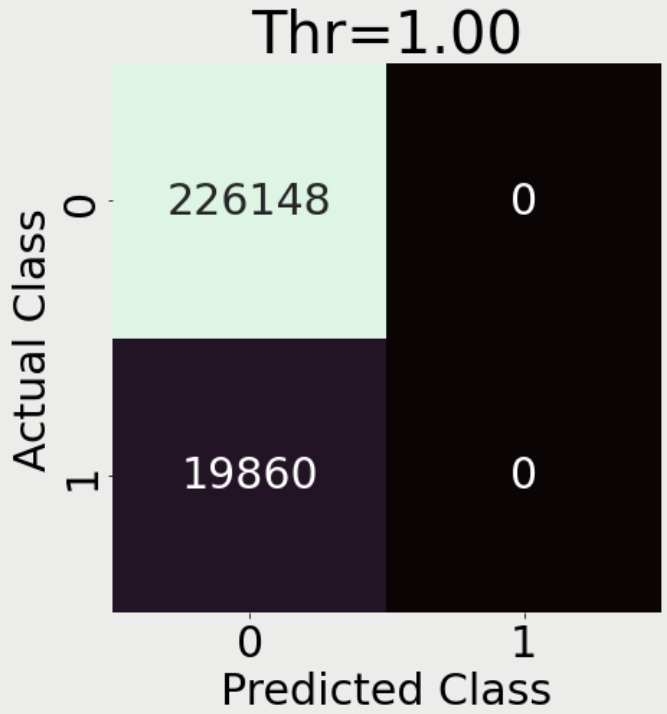
Variance Inflation Factor, Tests Chi deux, ANOVA, Kruskal-Wallis

II - Modélisation

Choix d’une métrique

Le paradoxe de l’accuracy

92 %



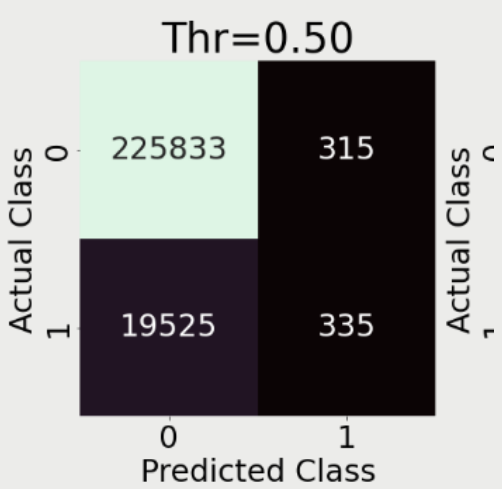
Choix d'une métrique

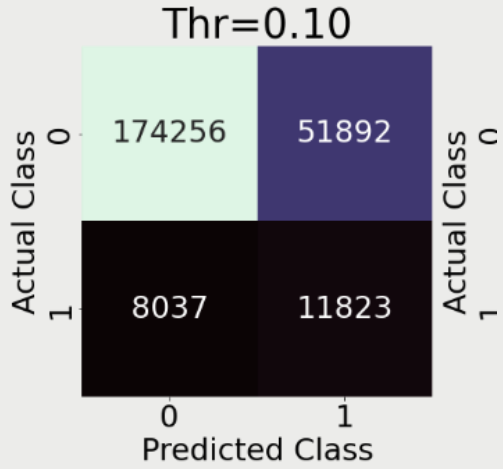
Precision

Recall

F1 Score

F bêta Score





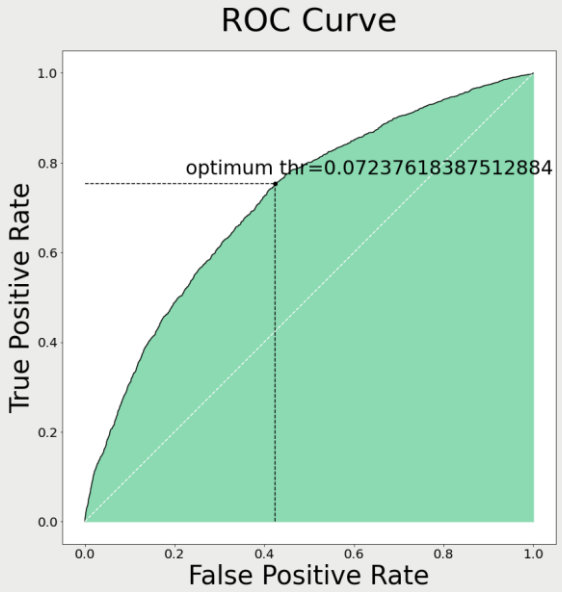
Dilemme

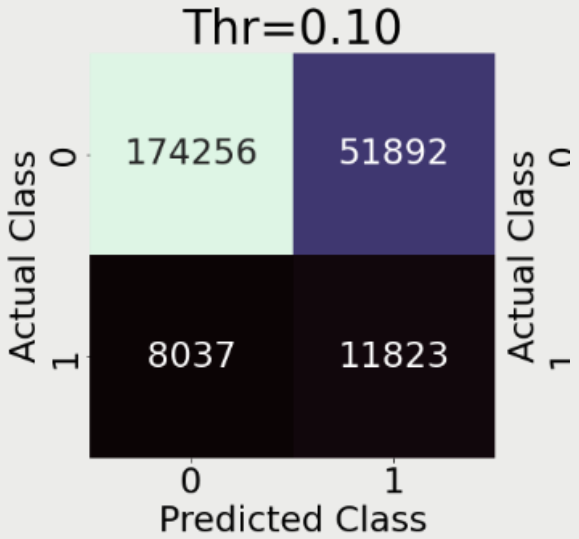
Accorder assez de crédits pour générer de l'argent

Avec assez de précaution pour ne trop en perdre

Choix d'une métrique

ROC-AUC



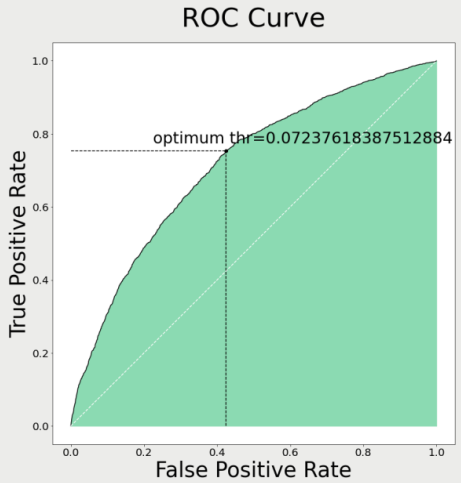


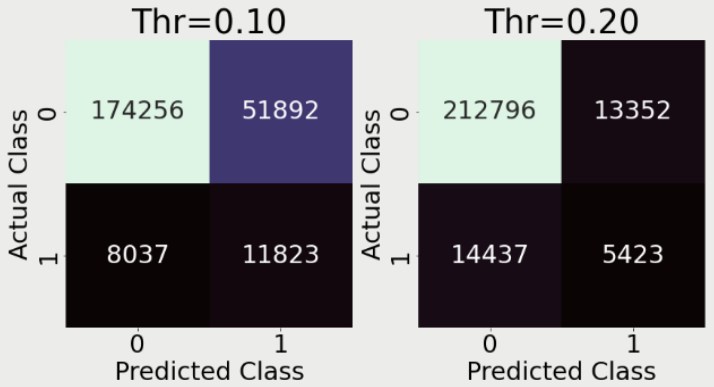
Choix d'une métrique

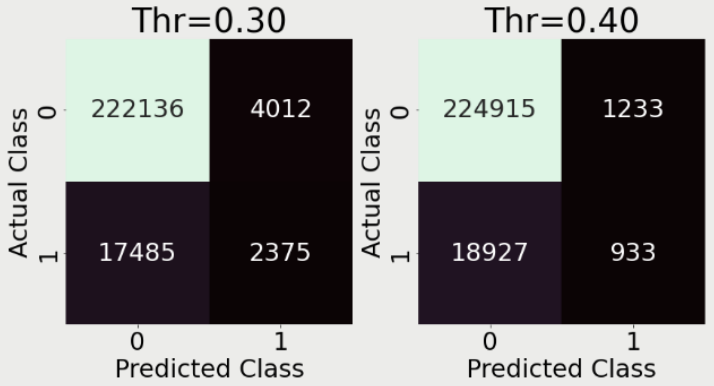
ROC-AUC

L'importance de la threshold

Predict_proba()





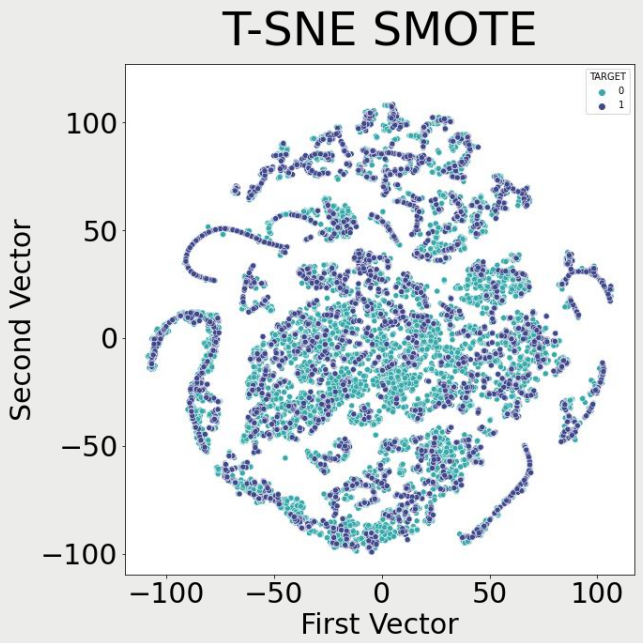


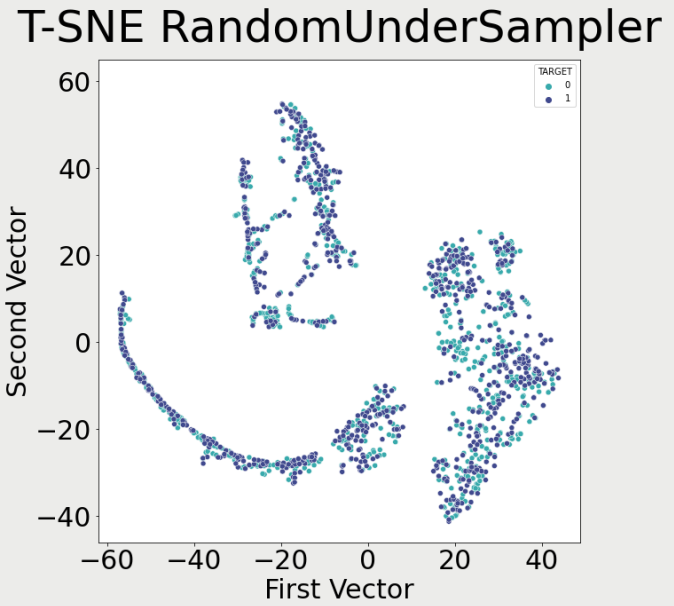
Des cibles non balancées

Imblearn

SMOTE,
ADASYN,
RandomUnderSampler,
ClusterUnderSampler

Class weight





Pipelines & Fine Tuning

Simple Imputer

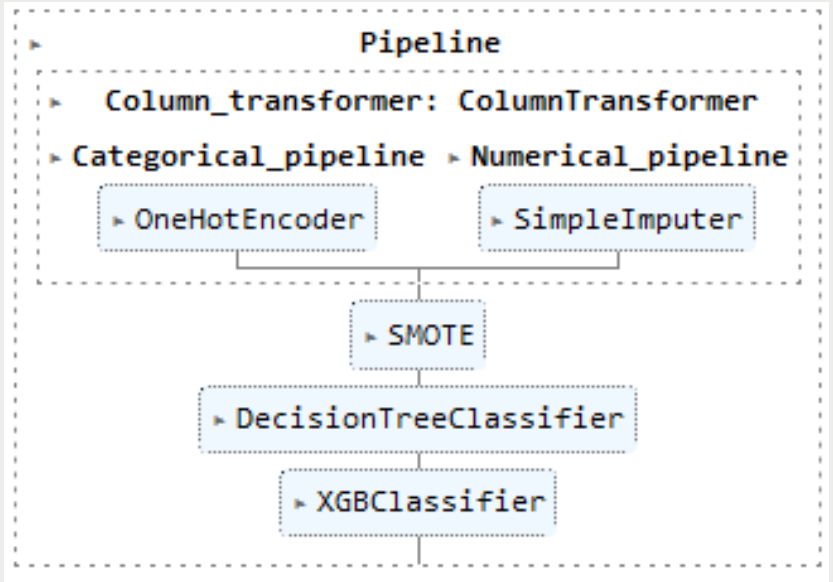
One Hot Ecoder

SelectFromModel

Standardisations

Sampling Method (Imblearn)

Réduction Dimensionnalité VS Interprétabilité

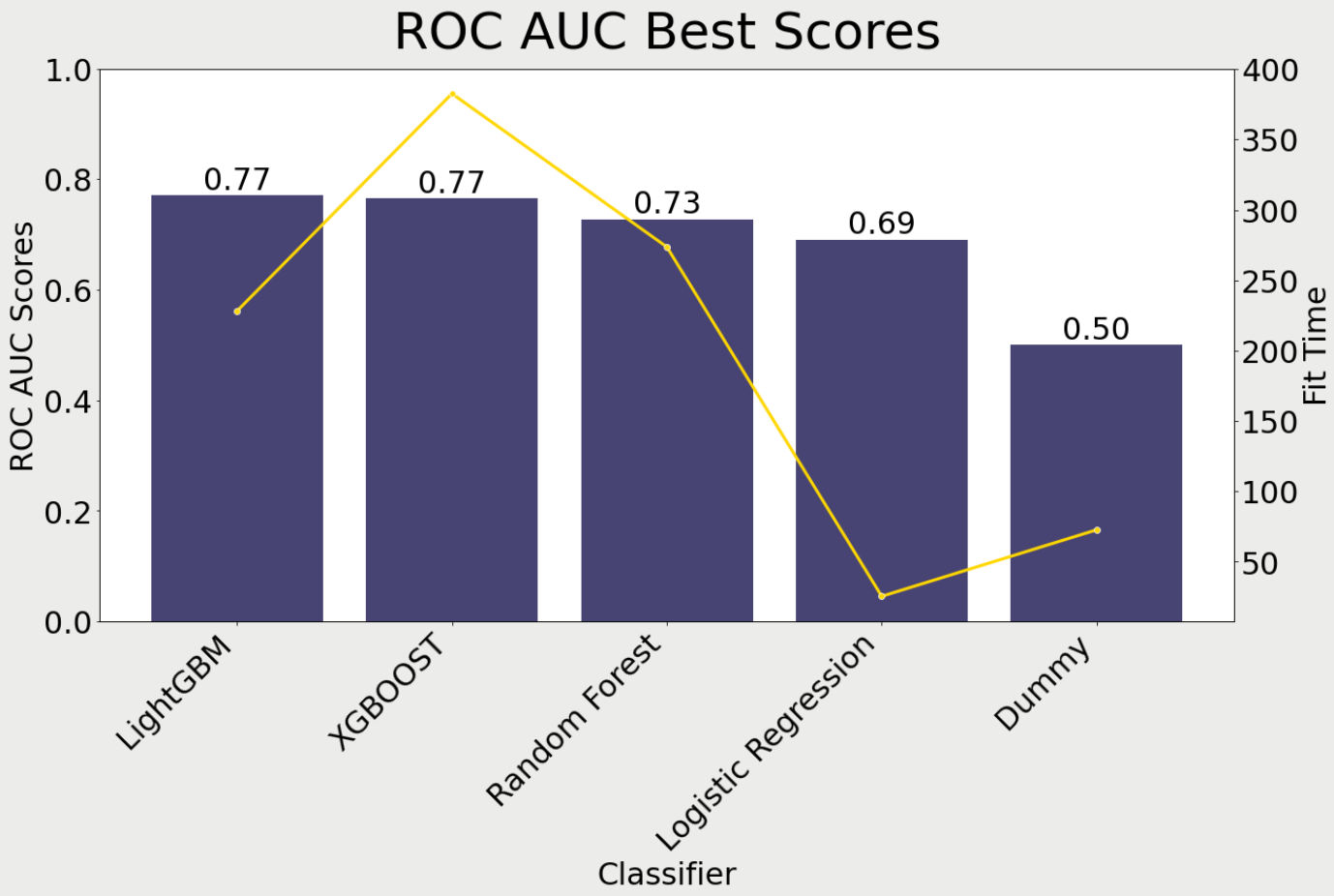


Pipeline Imblearn ~~sklearn~~, Cross-validation (5folds), Stratification, GirdSearch

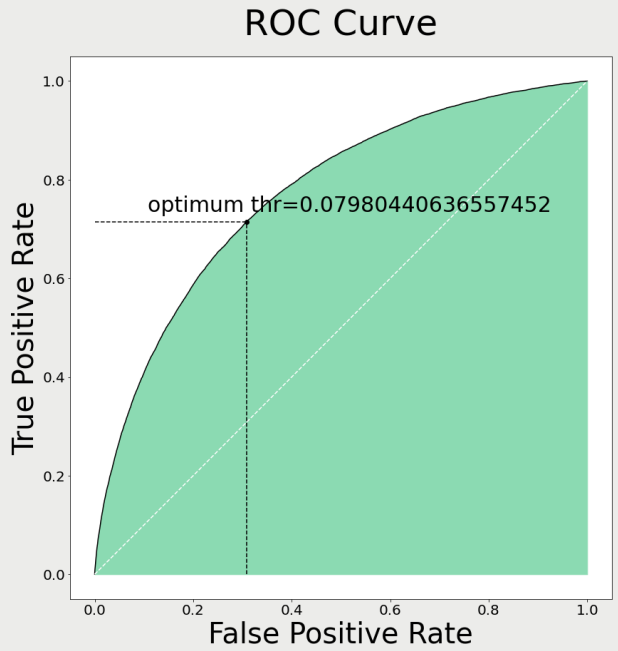
Modèles Entrainés

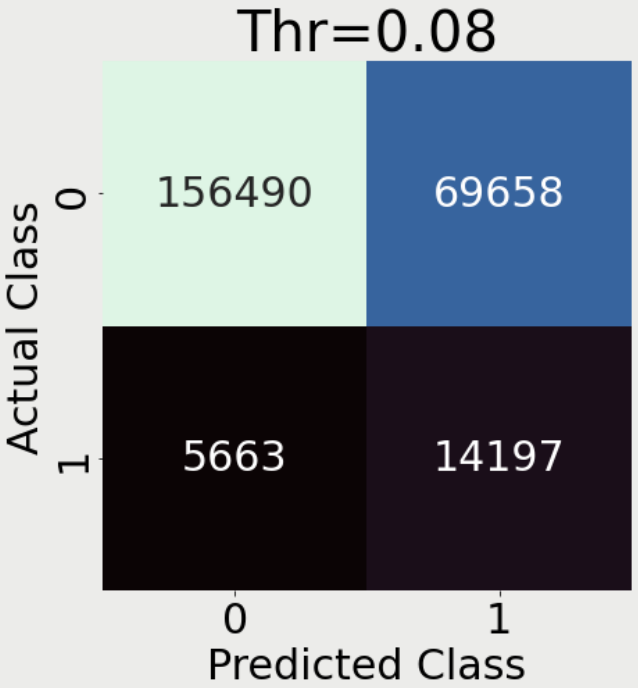
- Linéaire
  - Régression Logistique
- Arbres
  - Random Forest
  - Gradient Boosting
    - XGBoost
    - LightGBM

Modèles Résultats



Light GBM Global





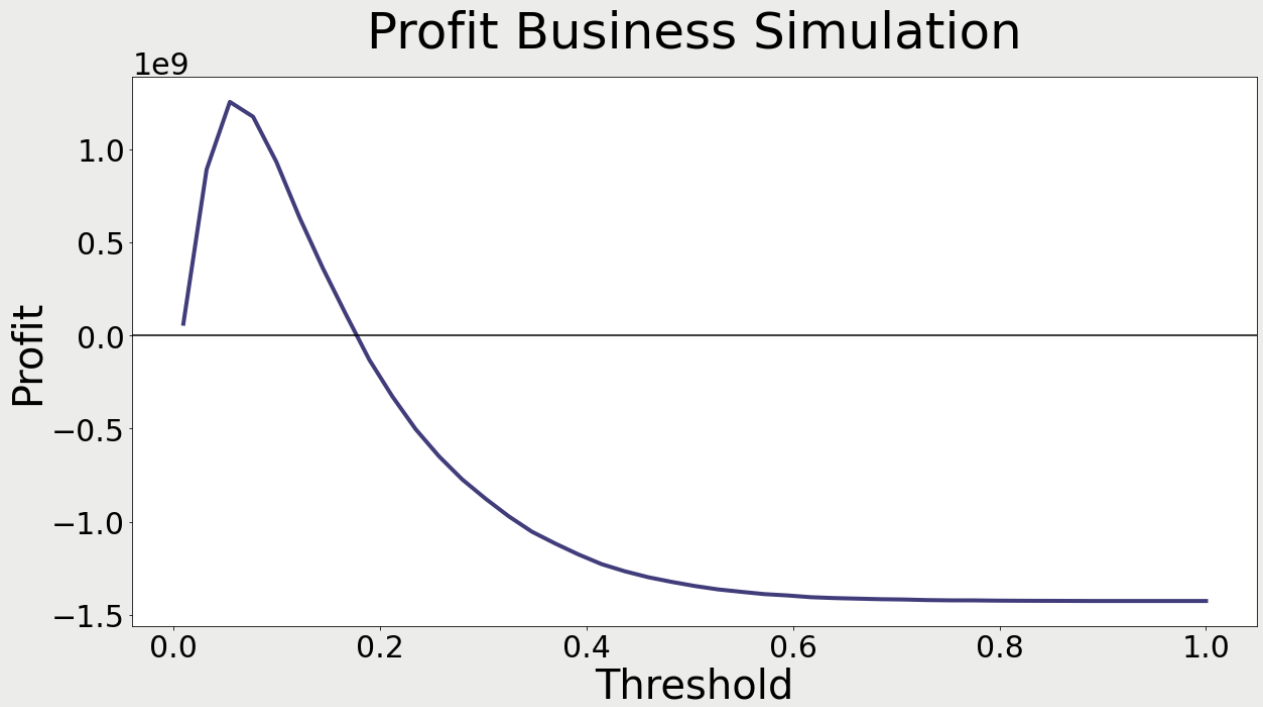
Threshold = 0,08

Fonction Métier

Hypothèses:

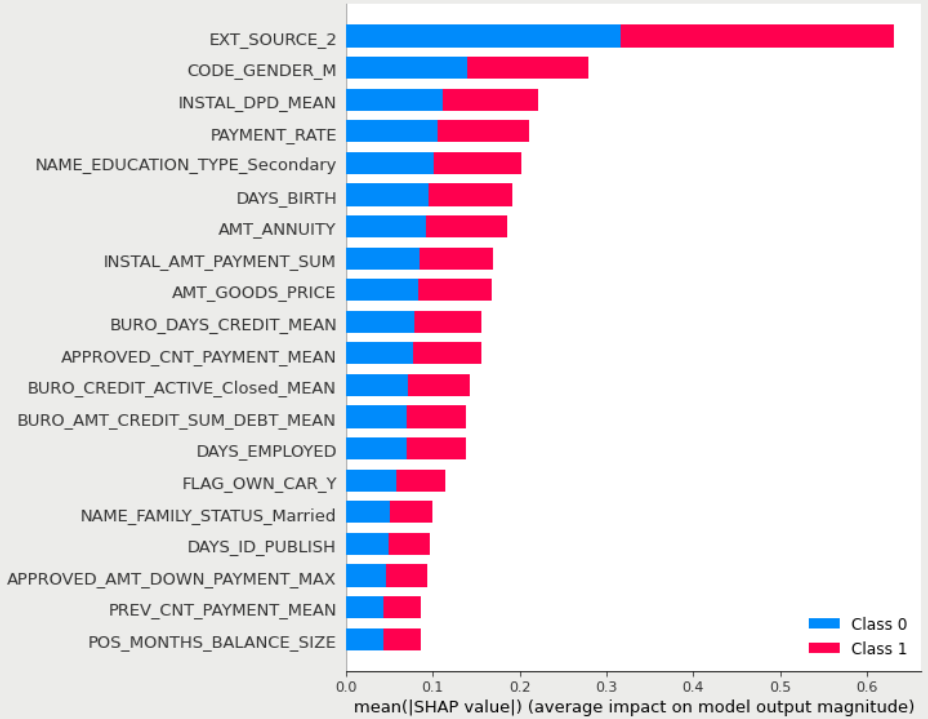
Gain : 1,03
Perte 0,5

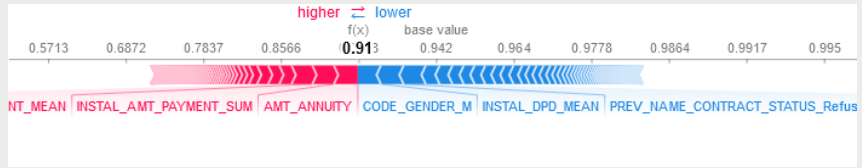
Threshold optimale correspond threshold optimale ROC-AUC

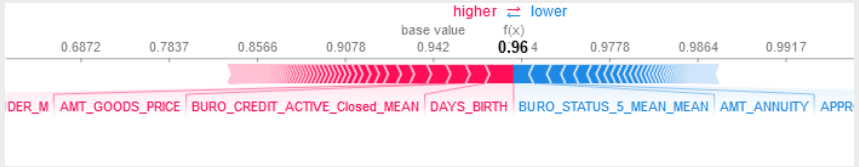


Light GBM Local

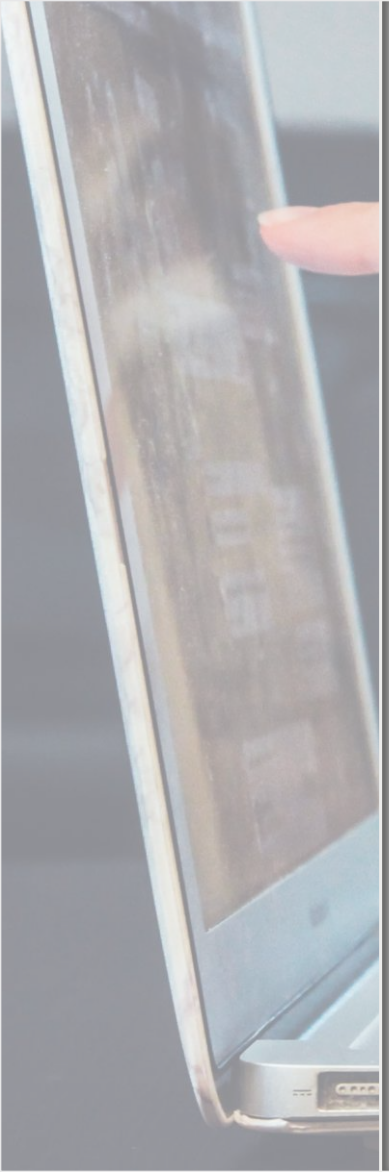
Théorie des jeux coopératifs







III - Dashboard



API

Flask

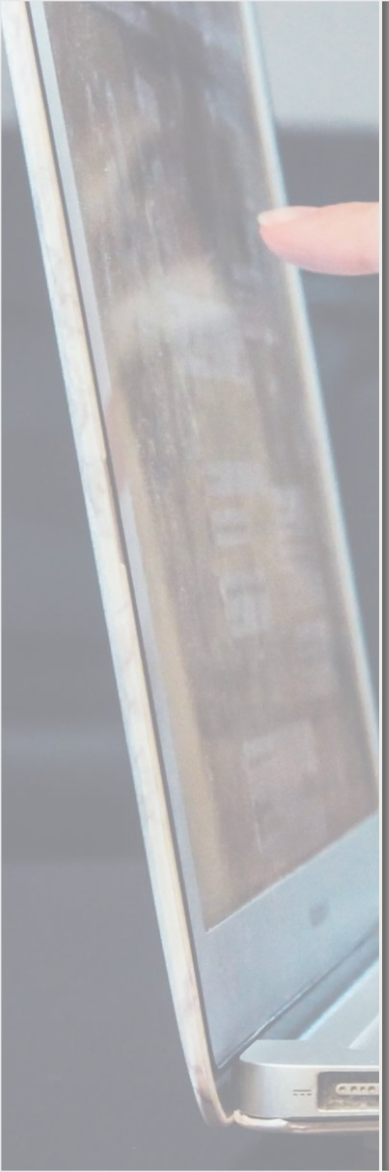
Clients.csv
Modèle.pkl
TreeExplainer

Routing
Logiques métier

Protocole HTTP
(GET, POST)

Json

```python
@app.route('/get_setup_infos')
def get_setup_infos():

    """Returns all features in the dataframe as well as all the categorical features only"""

    categories_list = ['EVERY_CLIENTS'] + dataframe.select_dtypes('object').columns.tolist()

    return {'all_features':dataframe.columns.tolist()[1:],
            'all_categories':categories_list}

@app.route('/get_prediction_proba/', methods=['POST'])
def get_prediction_proba():

    """Returns the predicted proba of a given customer id"""

    request_data = request.get_json()

    # Retrieves the row of a user given an id
    user_data = dataframe[dataframe['SK_ID_CURR'] == int(request_data['customer_id'])].iloc[:,1:]

    if len(user_data) == 0:
        abort(404)
    else :
        # Uses the model to predict proba
        customer_prediction = lgbm_model.predict_proba(user_data).tolist()[0][0]

        return {'prediction': customer_prediction}
```



Client

Streamlit

Composants faciles d’utilisation

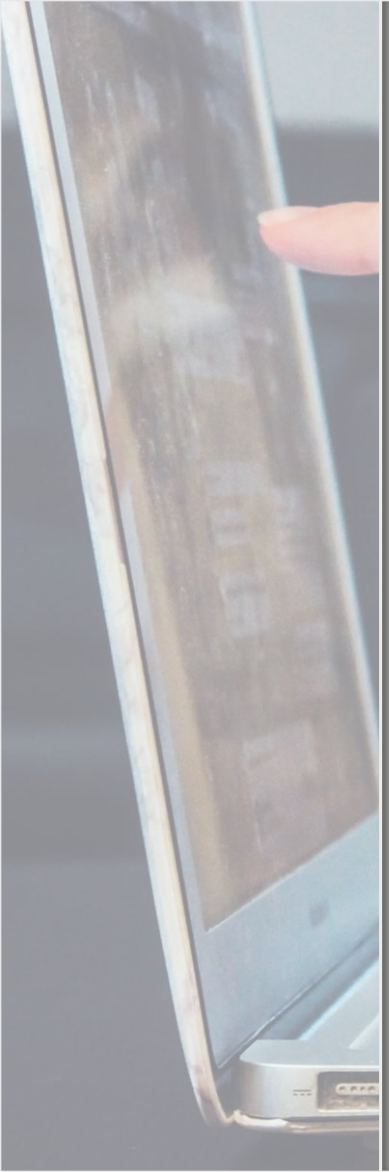
Axé DataScience

Matplotlib compatible, Plotly

React

Limites : CSS, HTML compliqué (iframes), Markdown, attention injection HTML et JS

```
fig = px.box(
            pd.DataFrame(
                feature_group_value['values_list'],
                columns=['SK_ID_CURR',selected_feature]),
            y=selected_feature,
            hover_data=['SK_ID_CURR',selected_feature],
            points="all",
            labels={
                    "variable": re.sub('_', ' ', selected_feature.lower())
            },
            title='{} representation inside {} category = {}'.format(re.sub('_', ' ', selected_feature.lower()),
                                                                     re.sub('_', ' ', selected_category.lower()),
                                                                     category_customer_value['feature_customer_value']))
fig.update_xaxes(tickvals=[""])
fig.add_hline(y=feature_customer_value['feature_customer_value'],
              annotation_text="You are here")
st.plotly_chart(fig)
```



Déploiement dans le cloud

En local Environnement Virtuels (Venv)

Pip freeze > requirements.txt

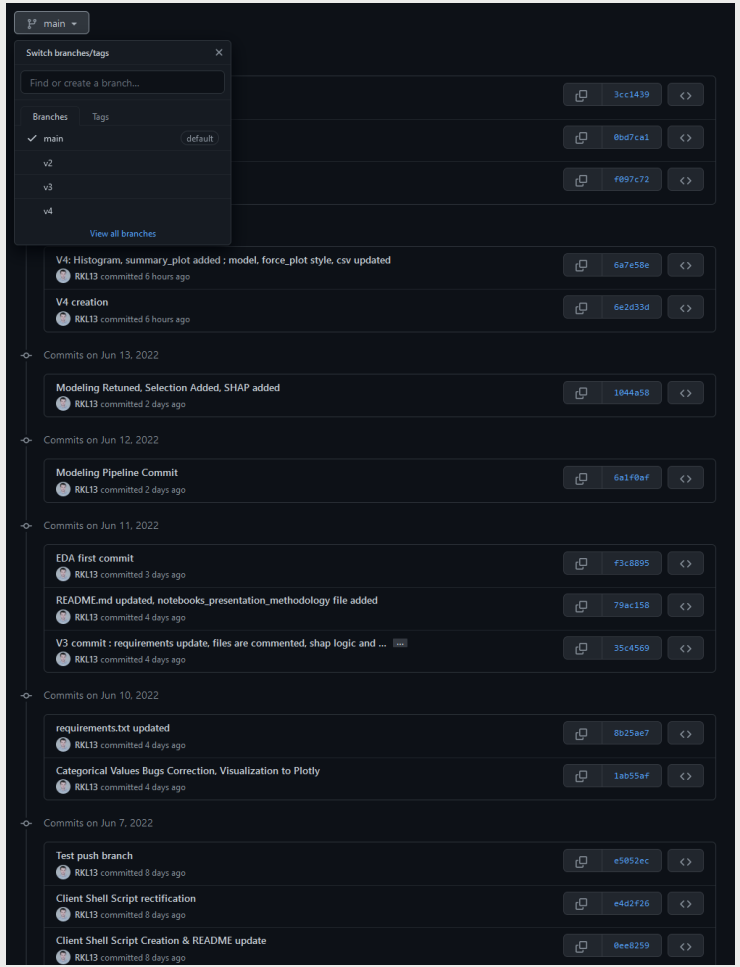
Versioning (4 Versions, 4 Branches)
Git, GitHub

2 Serveurs AWS

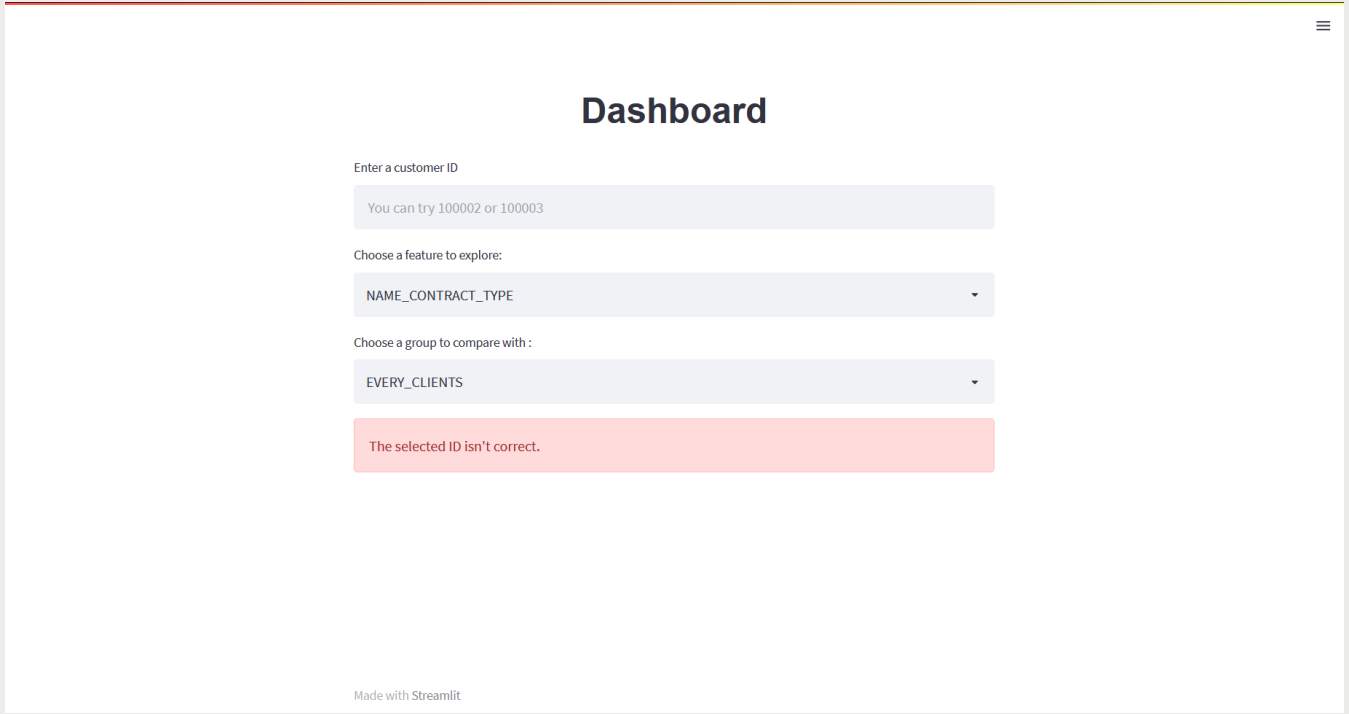
Linux distribution Ubuntu

Tmux

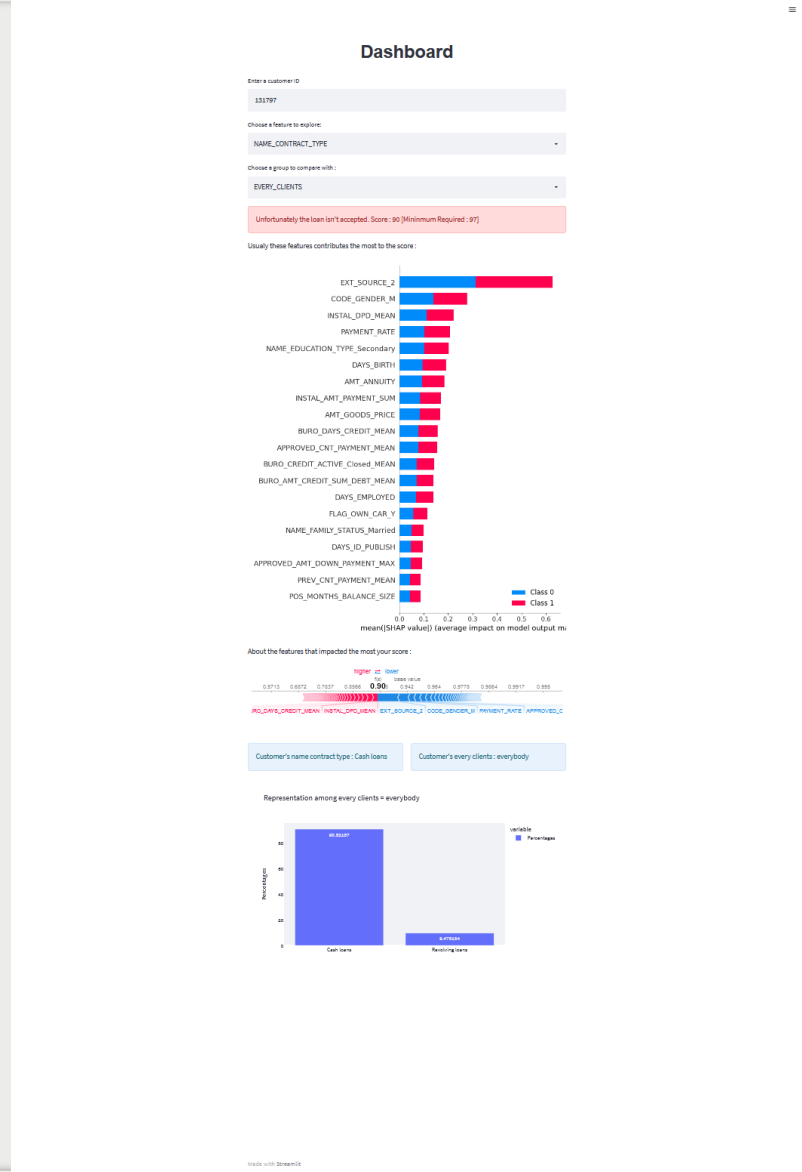
Script Shell pour le déploiement



UX

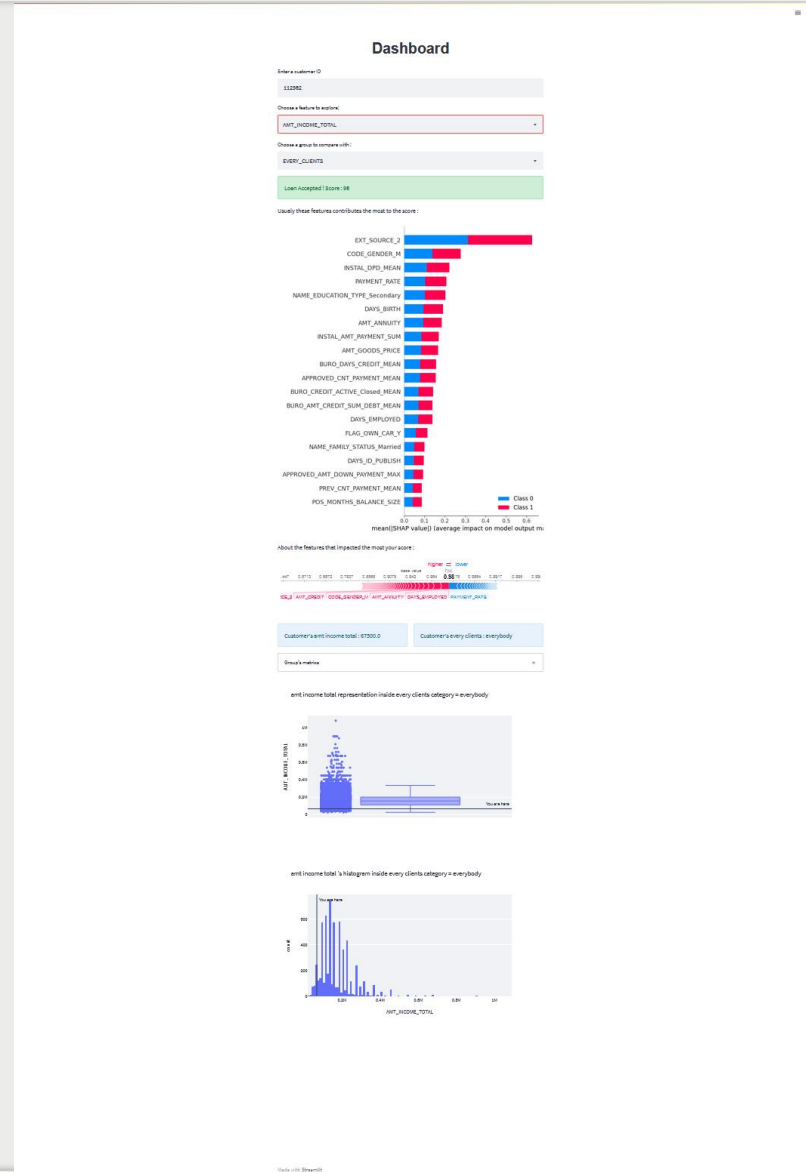


http://18.233.144.105:8501/



UX

http://18.233.144.105:8501/



IV - Limites et Améliorations

Limites et améliorations

Plus de samples de clients non solvables

Voting Classifier

Base de données, CRUD, Simulations

Docker

Score, Vulgarisation plus pousée, customer centric, lean management, feedbacks

Merci pour votre attention !