

## **Project**

### **Goals**

The goal of this project is to apply some concepts & tools seen in the 3 parts of this course, this project is organized into 3 parts :

- Part 1 : Building Classical ML projects with respect to basic ML Coding best practices
- Part 2 : Integrate MLFlow to your project
- Part 3 : Integrate ML Interpretability to your project

### **DataSet (Finance use case)**

DataSet of Home Credit Risk Classification:

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

you'll not use all the datasets available on Kaggle, only the main data set :

application\_train.csv

application\_test.csv

You can also use a part of this dataset (an example is given for this Lab)

### **Requirements**

An IDE with Python 3.6 or later (use anaconda environment for example)

### **Part 1**

Build an ML Project for Home Credit Risk Classification based on the given Dataset with respect to coding best practice for production ready code :

- Separate your ML projects workflow into different scripts (data preparation, feature engineering, models training, predict)
- Use a documentation library (sphinx)
- Use a conda environment for all your libraries
- Use GIT for code & model versioning
- Optionally, you can use a template cookie cutter (Example : <https://drivendata.github.io/cookiecutter-data-science/> ) or define your own

For this project, train & test the models : Xgboost, Random Forest and Gradient Boosting

## **Part 2**

Introduce MLFlow Library to your Project :

- Install MLFlow in your conda environment
- Track parameters of a model of your choice and display the results in your local mlflow UI (one model is enough with multiple runs)
- Deploy the model into a local REST server that will enable you to score predictions

## **Part 3**

Introduce SHAP Library to your Project :

- Install SHAP in your conda environment
- Use it to explain your XGboost model predictions :
  - Build a TreeExplainer and compute Shaplay Values
  - Visualize explanations for a specific point of your data set,
  - Visualize explanations for all points of your data set at once,
  - Visualize a summary plot for each class on the whole dataset.

## **Project conditions & evaluation**

-Work in teams of 2 or 3 max

-Delivery due date : **21/11/2021**

The evaluation of your project will be based on :

-Synthetic Report of your project

-Project code (notebooks, scripts, etc. )

-Project Outputs (predictions on test dataset, MLflow outputs, SHAP Output)