PROJECT 1

APPLICATIONS OF BIG DATA

Medina HADJEM

Hanaa LOUATIB - Anna FISCHMEISTER

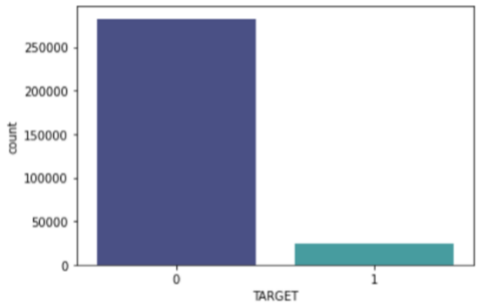
For this project we have to processed a Machine Learning project using concepts and tools seen during courses.
This project is divided in 3 parts :

- Building the Machine Learning project respecting coding best practices
- Integrating MLFlow to our project
- Integrating Machine Learning Interpretability with Shap to our project

Machine Learning project

We have to build a machine learning project for Home Credit Risk Classification based on the given Dataset which is available on Kaggle: application_train.csv, application_test.csv. The Home Credit service is dedicated to provided lines of credit to the unbanked population. So, predicting if a client will repay a loan, the variable in the train is call the TARGET if the person repays the statut is 1 and if not it's 0. The others are related to the global financial situation of the person, like his age, if he is employed or not, how long he is employed, if he is married…different parameters that could affect the budget.

In our dataset we have more people that didn't repay their credit (85 % of the data), we can easily identify person who won't pay that why our model won't have a good accuracy.



We also identified that we have in our data several variables having more than 50 % of their rows with no value. To this problem we decided to divide the problem.

First for quantitative variables, we complete null value by the mean of the variable. For qualitative variables only 5 had null value and we decided to drop them from the data because they aren't correlated to the target variable.

For the other qualitative variables, we need to encode them to see if and how they are correlated to the target.

Finally, we have in our train 307 511 rows with 202 features. We decided to apply these steps to test file to have same features in both files.

Machine Learning Models

According to your suggestion, we had implemented three machine learning models:

- RandomForest, and we calculated the Accuracy score
- XGBoost, giving the Loss and Accuracy score
- GradientBoost, that we took the Accuracy score also.

We will see all this metrics in the MLFlow web server.

Integrating MLFlow to our project

After the cleaning and feature engineering, we implement three Machine Learning models and deploy them with MLflow.
We started with MLFlow Tracking how is a component of MLflow that logs and tracks your training run metrics and model artifacts. We used anaconda as experiment environment.

We used MLflow Tracking to store our training metrics and models.
With MLflow Tracking, you track an experiment's run metrics and store model artifacts.

Prerequisites

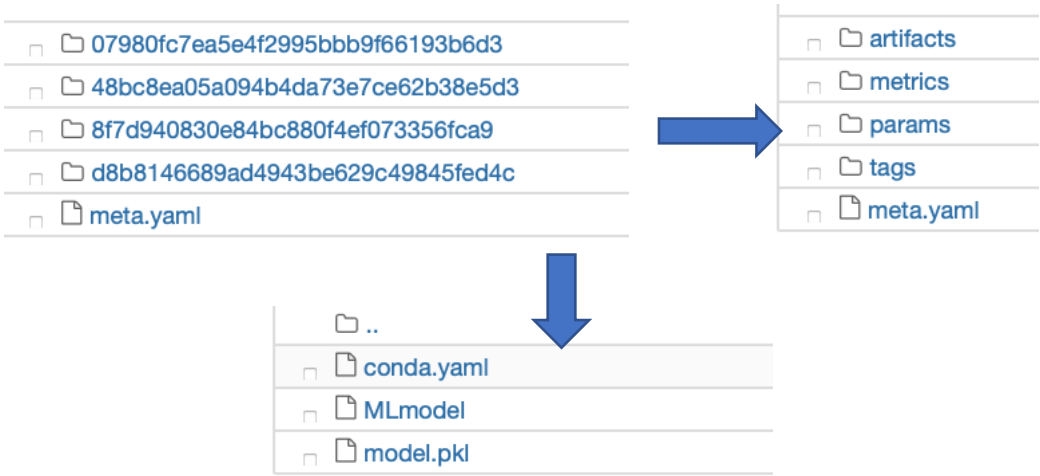
- We installed all the necessary packages and libraries.
- Create environment

Track local runs

To run the MLFlow, we build script for each machine learning model with some hyperparameters and metrics.

Each time we run the MLFlow Tracking codes we generate the metrics and artifacts that are saved to a working directory in newly created /mlruns subdirectory. So, the metrics are logged automatically to our workspace.

In artifacts folder we found the model file that align with the MLFlow model scheme (conda.yaml, MLmodel, model.pkl).

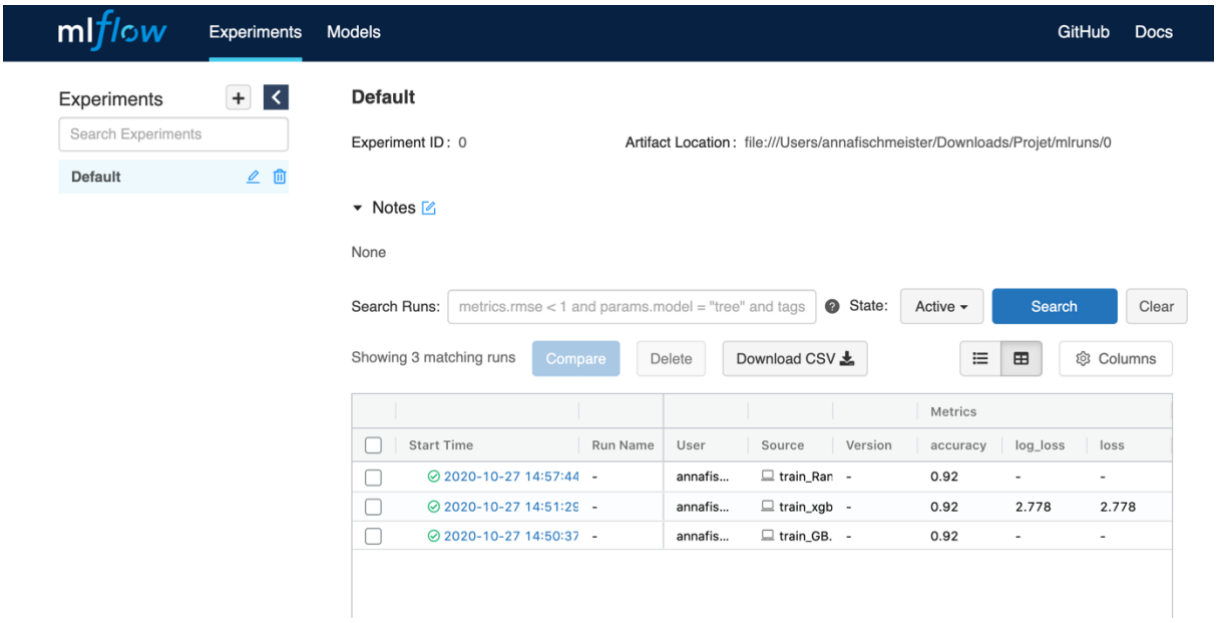


And in metrics folder we have the metrics as Accuracy, Loss …

MLFlow UI

To inspect the recorded data aesthetically, stay in the same directory and run mlflow ui.

```
Entrée [*]: !mlflow ui
[2020-10-27 15:47:47 +0100] [6669] [INFO] Starting gunicorn 20.0.4
[2020-10-27 15:47:47 +0100] [6669] [INFO] Listening at: http://127.0.0.1:5000 (6669)
[2020-10-27 15:47:47 +0100] [6669] [INFO] Using worker: sync
[2020-10-27 15:47:47 +0100] [6671] [INFO] Booting worker with pid: 6671
[2020-10-27 15:48:28 +0100] [6669] [CRITICAL] WORKER TIMEOUT (pid:6671)
[2020-10-27 15:48:28 +0100] [6671] [INFO] Worker exiting (pid: 6671)
```



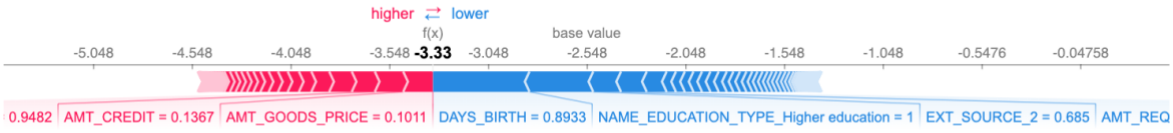
Integrating Machine Learning Interpretability with Shap to our project

Shap is a theoretical approach to explain the output of a machine learning model. We decided to work with out Xgboost model.

First we have to calculate the Shapley value, it explains the prediction of an instance by computing the contribution of each feature. So, in our case which feature affects the repay of a loan and how.

Here, we chose the first shape value and the first raw of our train.

```
]: shap.force_plot(explainer.expected_value, shap_values[0], X_train[0], feature_names=list(X.columns.values))
```



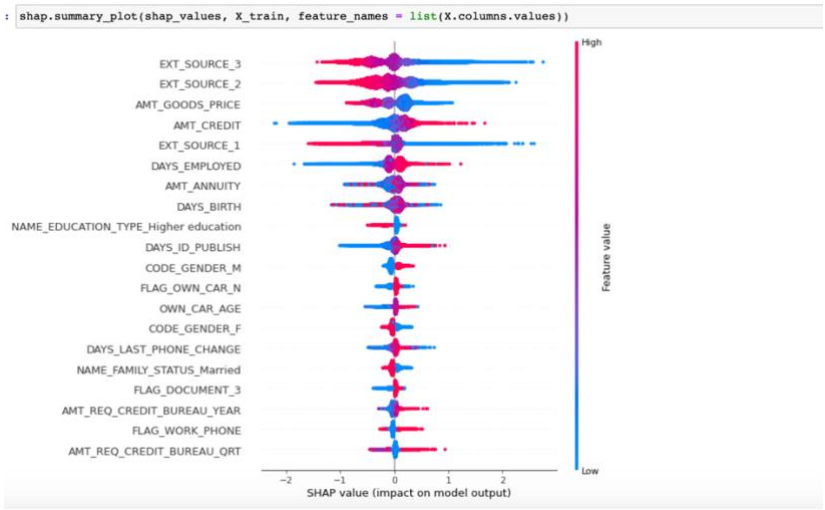
Explanation of the impact of features for one example
in red features impacting positively and in blue which impact negatively

We can see here that if we increase AMT_CREDIT or AMT_GOODS_PRICE the output value (-3,33) will move to the base value (-2,548) on the opposite if we decrease the value of DAYS_BIRTH the output value will continue to decrease.
As we have many targets with 0 so our base value isn't representative here.

We have a huge dataset, so for having a global representation we choose randomly 100 rows with their shap value.

Explanation of the impact of features for 100 examples

The blue line represents the base value. We can see that a lot of parameters impact negatively to the output value of each row.



Representation of the importance of features using shap value

Finally, we have a general representation. The shap value is calculated for each example of the dataset so we can represent them by point. Here, the most important is the EXT_SOURCE_3, but thanks to these representation we can say that this variable has a negative impact if the value is high.

We have worked with a MacOS environment and we had some troubles with the ‘conda’ command. Due to that we cannot export our general conda environment as a yml file.