Adrien MICHEL
28/10/2020

Repport of project

Part1 (Building ML Models with coding best pratices)

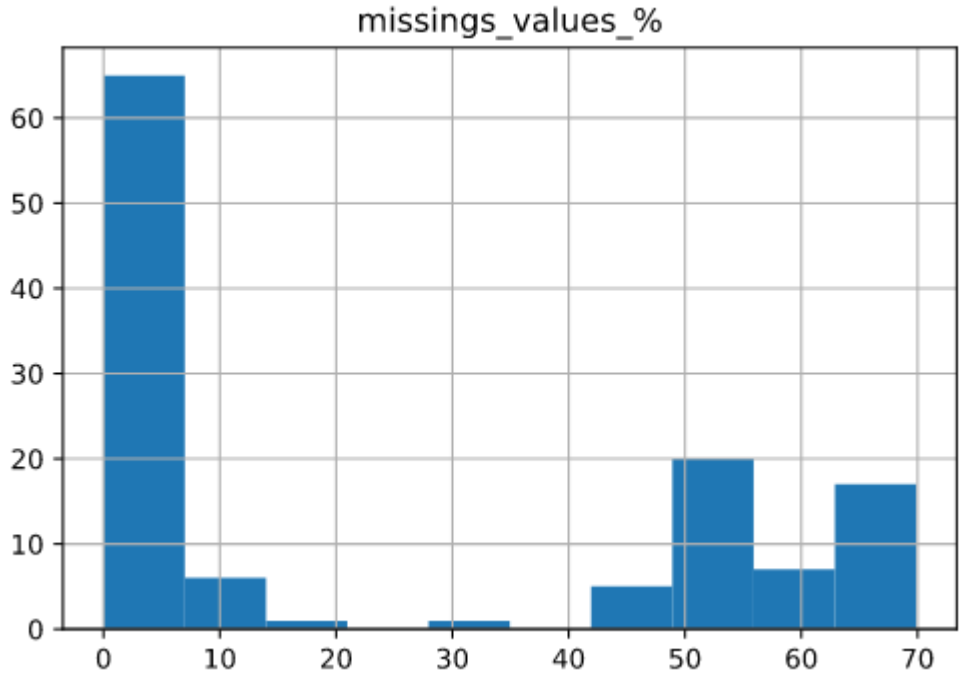
Data exploration

Our dataset dataset contains 307511 rows and 121 features. \

The target output is composed of 2 classes with 282686 entries for class 0, and 24825 entries for class 1

```
>>> df.groupby('TARGET')['TARGET'].count()
>>> TARGET
0    282686
1     24825
Name: TARGET, dtype: int64
```

There is a lot of features containing a lot of null values



Feature engineering

So, we delete features with more than 20% of null values, because they will not be significant for our models.

Next, we delete rows with at least one null value in the categorical features because we lost few data (1292)

We encode categorical features with a label encoder for features with only 2 values and a one hot encoding for the others

We keep numerical features intact

Model building

For this project, we choosed 3 models.
Our problem being a classifying problem, we chose 4 metrics to evaluate our models' performance :

- accuracy
- precision
- recall
- f1 score

Random forest classifier

We chose only two parameters :

- max_depth
- min_samples_split

Gradient boosting classifier

We chose only three parameters :

- learning_rate"
- max_depth
- min_samples_split.

XGBoost classifier

We chose only three parameters :

- learning_rate"
- max_depth
- min_child_weigth

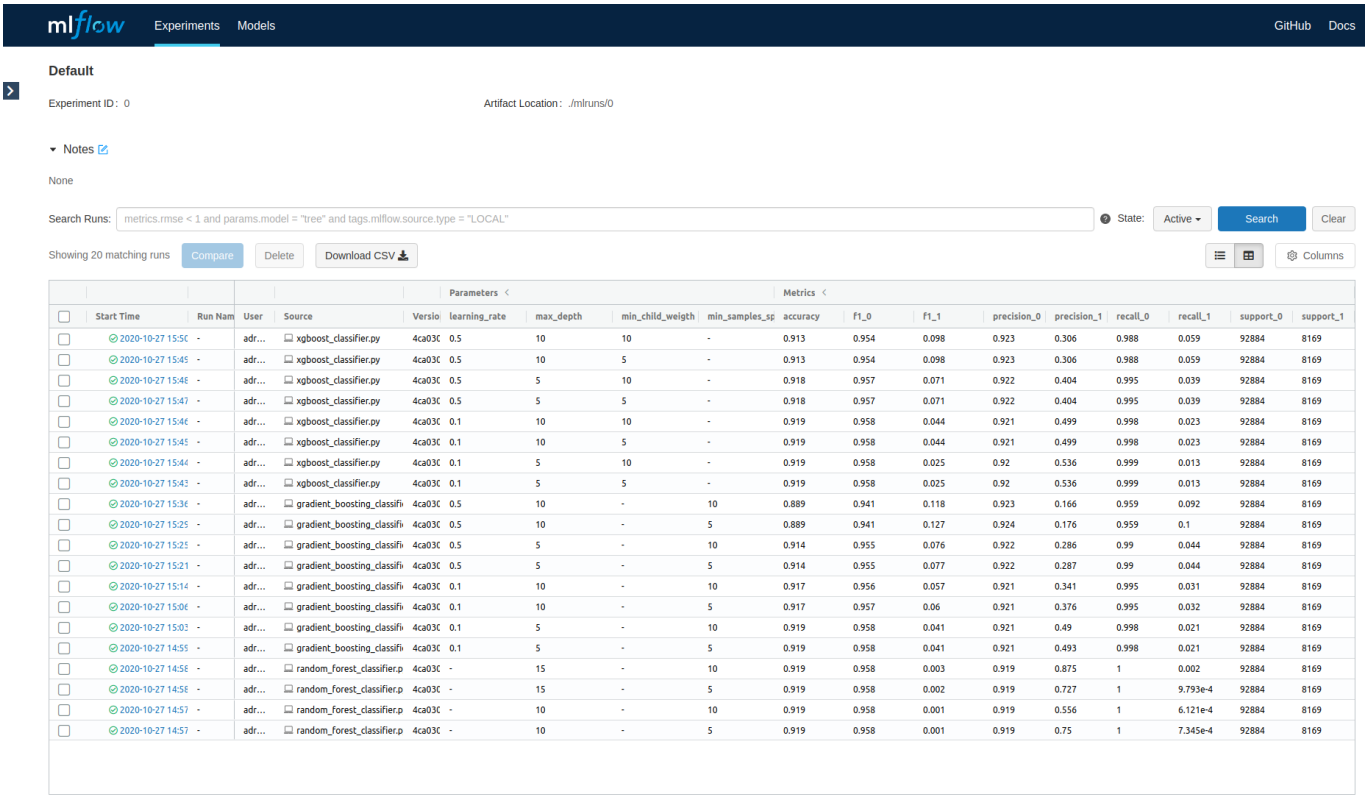
Part 2 (Using MLFLow)

We started a tracking server on local at http://127.0.0.1:5000 to monitor our models' trainings

```
(appOfBd_project_env) adrien@adrien-Swift-SF515-
51T:~/Documents/Ecole/EFREI/M2-EFREI/Applications of Big Data/appBdProject$
mlflow server
[2020-10-27 10:38:38 +0100] [12022] [INFO] Starting gunicorn 20.0.4
[2020-10-27 10:38:38 +0100] [12022] [INFO] Listening at:
http://127.0.0.1:5000 (12022)
[2020-10-27 10:38:38 +0100] [12022] [INFO] Using worker: sync
[2020-10-27 10:38:38 +0100] [12024] [INFO] Booting worker with pid: 12024
```

```
[2020-10-27 10:38:38 +0100] [12032] [INFO] Booting worker with pid: 12032
[2020-10-27 10:38:38 +0100] [12033] [INFO] Booting worker with pid: 12033
[2020-10-27 10:38:38 +0100] [12034] [INFO] Booting worker with pid: 12034
```

After running multiple models with different parameters, we have obtained this tracking window



To choose the best model, we filter in MLflow the models which obtained the best scores for accuracy and precision with this following command :\

```
metrics.precision_1 > 0.7 and metrics.precision_0 > 0.9 and
metrics.accuracy > 0.9
```

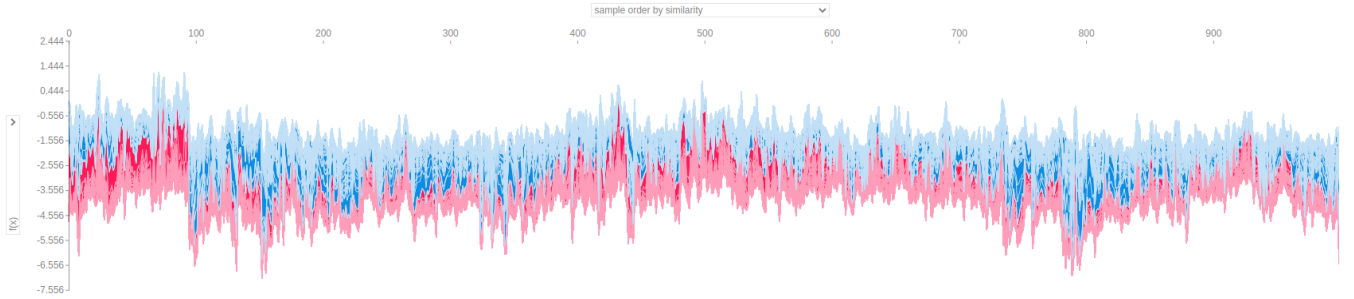
Run ID:	91a2cd2c588746b6beaf62113978…	38bf3a0d2d8b48b5adac49e77888…	898254262e19483d8d3d9555cb4…	e9c10dead6c44405b8e3a85a99fa…	ca30c53662e646bb93ee7e7d66a0…	31ff212f34b74af48ae098de4b9dd…
Run Name:						
Start Time:	2020-10-27 15:44:36	2020-10-27 15:43:57	2020-10-27 14:58:51	2020-10-27 14:58:18	2020-10-27 14:57:44	2020-10-27 14:57:08
Parameters						
learning_rate	0.1	0.1				
max_depth	5	5	15	15	10	10
min_child_weigth	10	5				
min_samples_split			10	5	10	5
Metrics						
accuracy	0.919	0.919	0.919	0.919	0.919	0.919
f1_0	0.958	0.958	0.958	0.958	0.958	0.958
f1_1	0.025	0.025	0.003	0.002	0.001	0.001
precision_0	0.92	0.92	0.919	0.919	0.919	0.919
precision_1	0.536	0.536	0.875	0.727	0.556	0.75
recall_0	0.999	0.999	1	1	1	1
recall_1	0.013	0.013	0.002	9.793e-4	6.121e-4	7.345e-4
support_0	92884	92884	92884	92884	92884	92884
support_1	8169	8169	8169	8169	8169	8169

The best model is the random forest classifier with *max_depth=10* and *min_samples_split=10*

Part 3 (XAI with SHAP Method)

Explanations for a specific point of data set

Explanations for all points of data set at once



Summary plot