AI Engineer PATH

Project 4

Develop a scoring Model.

Axel Favreul

Contents

01

Exploratory Analysis

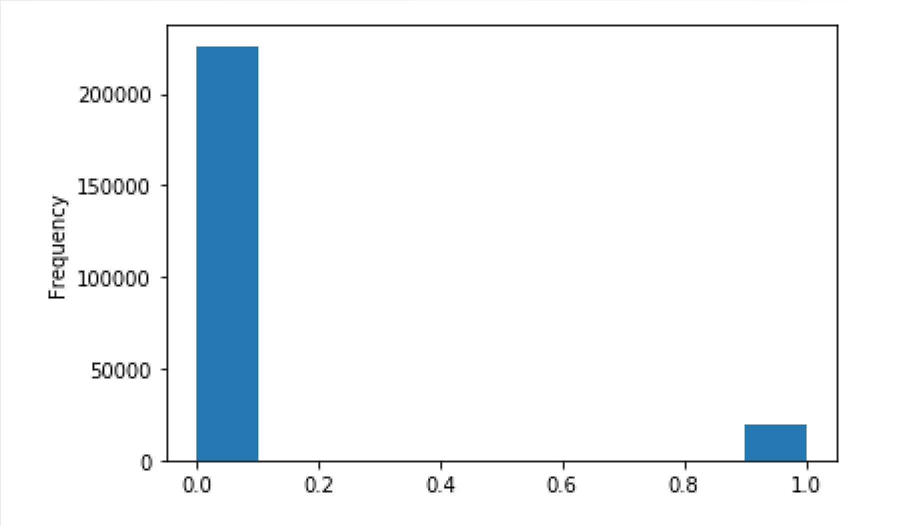
Dataset Overview

Multiple dataset related to client information:

Main dataset concerning client itself.

Additional dataset concerning previous transaction

Target distribution :



Imbalance class problem

Missing value :

66 column over 121 contain MV

Up to 70% for some column.

Column types :

float64	65
int64	41
object	16

Dataset Overview

- **Label encoding for cat**
  columns with two differ
  columns)
- **One-hot for the of the c**
  columns.

Outlier Method :

To detect: zscore thres
Inter quartile range.

21 Found

	Name	zscore	iqr
1	CNT_CHILDREN	3364	3364
2	AMT_INCOME_TOTAL	214	11226
3	AMT_CREDIT	2609	5235
6	REGION_POPULATION_RELATIVE	6745	6745
8	DAYS_EMPLOYED	0	57846
9	DAYS_REGISTRATION	580	505
12	FLAG_MOBIL	1	1
13	FLAG_EMP_PHONE	0	44409
14	FLAG_WORK_PHONE	0	48934
15	FLAG_CONT_MOBILE	456	456
17	FLAG_EMAIL	13978	13978
18	CNT_FAM_MEMBERS	3157	3157
19	REGION_RATING_CLIENT	0	64626
20	REGION_RATING_CLIENT_W_CITY	0	62594
21	HOUR_APPR_PROCESS_START	506	1816

e outlier max = 117 000 000
of the loan.
: NSTR
omalies. see next section
7 years quant75 = 20.5 years. def= how many days
lient change his registration, time relative to the

Dataset Overview

Outlier

AMT Income totale :

High income default on 5,61%

Low income default on 8,20%

Days Employment:

66 columns over 121 contain MV

Up to 70% for some columns.

365243	44397
-200	126
-212	123
-230	122
-196	116
	...
-12367	1
-13904	1
-11725	1
-13648	1
0	1

Days employed value count

Non-Anomalies default on 8,66%

Anomalies on 5,46%

➔ Flagged the line and replace by Nan value

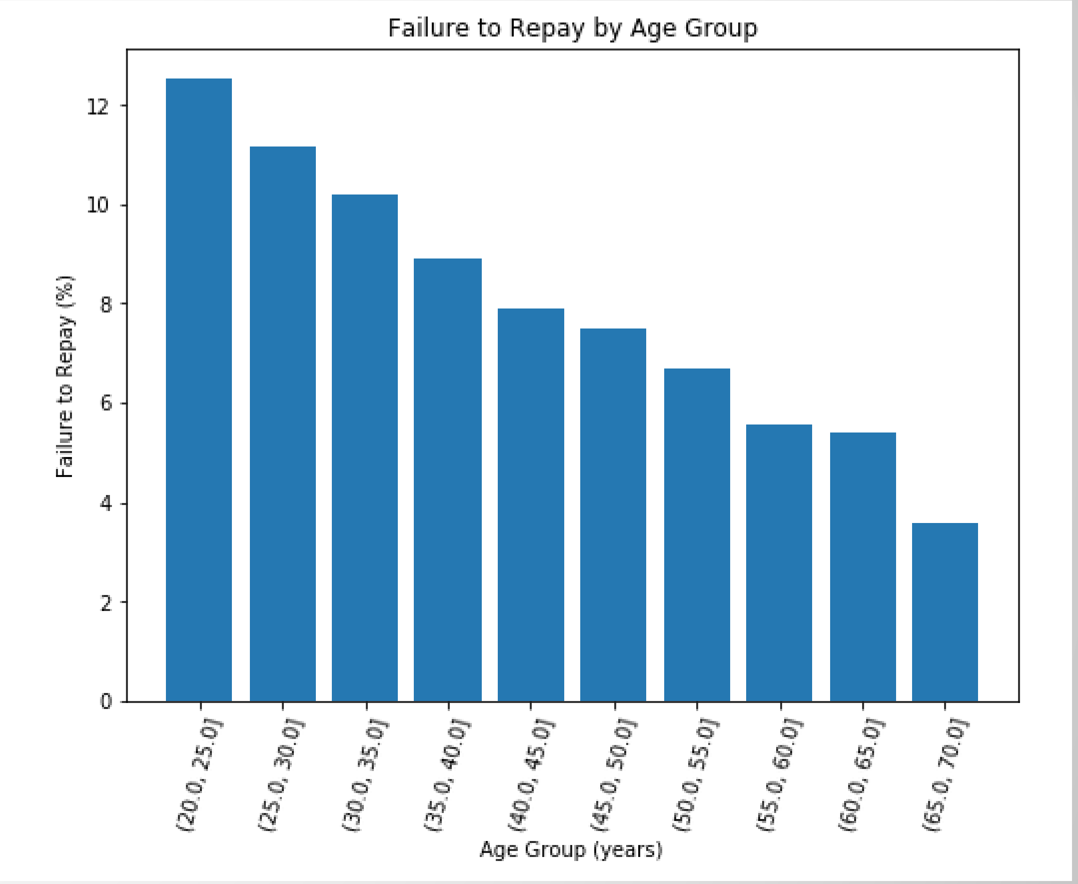
Dataset Overview

```
Most Positive Correlations:
 OCCUPATION_TYPE_Laborers                             0.042102
FLAG_DOCUMENT_3                                      0.043879
FLAG_EMP_PHONE                                       0.045066
REG_CITY_NOT_LIVE_CITY                               0.046282
NAME_EDUCATION_TYPE_Secondary / secondary special    0.048568
REG_CITY_NOT_WORK_CITY                               0.051325
DAYS_ID_PUBLISH                                      0.051918
CODE_GENDER_M                                        0.053534
DAYS_LAST_PHONE_CHANGE                               0.055383
NAME_INCOME_TYPE_Working                             0.057175
REGION_RATING_CLIENT                                 0.059217
REGION_RATING_CLIENT_W_CITY                          0.061117
DAYS_EMPLOYED                                        0.073386
DAYS_BIRTH                                           0.077571
TARGET                                               1.000000
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                            -0.178579
EXT_SOURCE_2                            -0.160777
EXT_SOURCE_1                            -0.155615
NAME_EDUCATION_TYPE_Higher education    -0.055840
CODE_GENDER_F                           -0.053528
NAME_INCOME_TYPE_Pensioner              -0.045358
ORGANIZATION_TYPE_XNA                   -0.045072
DAYS_EMPLOYED_ANOM                      -0.045072
FLOORSMAX_AVG                           -0.042968
EMERGENCYSTATE_MODE_No                  -0.042813
FLOORSMAX_MEDI                          -0.042620
FLOORSMAX_MODE                          -0.042244
HOUSETYPE_MODE_block of flats           -0.041717
AMT_GOODS_PRICE                         -0.040782
REGION_POPULATION_RELATIVE              -0.036493
Name: TARGET, dtype: float64
```

Correlations

Age :



Gender : Male default on 10,11% and female on 7,03%

02

Model testing

Initial performance

+ Impute and scale

Imputation : median
Scaler : MinMax

+ Model tested:

LogisticRegression

Random Forest

XG Boost

+ Initial result

Score measured with AUC-ROC

	Features	Score
0	Log reg	0.689468
1	Random forest	0.708514
2	xgboost	0.752845

03

Feature engineering

Polynomial features

Choose of features

Most strong correlation value :
EXT_SOURCE (1,2 and 3)
DAYS_BIRTH

New score XG Boost

	Features	Score
0	Log reg	0.689468
1	Random forest	0.708514
2	xgboost	0.752845
3	xg_poly	0.753731

New correlation

```
EXT_SOURCE_2 EXT_SOURCE_3                  -0.193755
EXT_SOURCE_1 EXT_SOURCE_2 EXT_SOURCE_3     -0.189492
EXT_SOURCE_2^2 EXT_SOURCE_3                -0.176281
EXT_SOURCE_2 EXT_SOURCE_3^2                -0.172162
EXT_SOURCE_1 EXT_SOURCE_2                  -0.166753
EXT_SOURCE_1 EXT_SOURCE_3                  -0.164042
EXT_SOURCE_2                               -0.160619
EXT_SOURCE_1 EXT_SOURCE_2^2                -0.156908
EXT_SOURCE_3                               -0.155518
EXT_SOURCE_1 EXT_SOURCE_3^2                -0.151002
Name: TARGET, dtype: float64
EXT_SOURCE_1 EXT_SOURCE_2 DAYS_BIRTH      0.155983
EXT_SOURCE_2 DAYS_BIRTH                   0.156999
EXT_SOURCE_2 EXT_SOURCE_3 DAYS_BIRTH      0.180994
TARGET                                    1.000000
1                                              NaN
Name: TARGET, dtype: float64
```

Domain knowledge

Bureau dataset

Number of previous loan

	SK_ID_CURR	previous_loan_counts
0	100001	7
1	100002	8
2	100003	4
3	100004	2
4	100005	3

Numerical information

	SK_ID_CURR	DAYS_CREDIT					CREDIT_DAY_OVERDUE				...	DAYS_CREDIT_UPDATE					
		count	mean	max	min	sum	count	mean	max	min	...	count	mean	max	min	sum	count
0	100001	7	-735.000000	-49	-1572	-5145	7	0.0	0	0	...	7	-93.142857	-6	-155	-652	7
1	100002	8	-874.000000	-103	-1437	-6992	8	0.0	0	0	...	8	-499.875000	-7	-1185	-3999	7
2	100003	4	-1400.750000	-606	-2586	-5603	4	0.0	0	0	...	4	-816.000000	-43	-2131	-3264	0
3	100004	2	-867.000000	-408	-1326	-1734	2	0.0	0	0	...	2	-532.000000	-382	-682	-1064	0
4	100005	3	-190.666667	-62	-373	-572	3	0.0	0	0	...	3	-54.333333	-11	-121	-163	3

Domain knowledge

Score with new features

	Features	Score
0	Log reg	0.689468
1	Random forest	0.708514
2	xgboost	0.752845
3	xg_poly	0.753731
4	xg_burr	0.756839
5	random_f_burr	0.714592

04

Hyperparameters

Hyperparameters

Automated tuning

Using RandomizedSearchCV
To gain computing time.

Parameters

```
params = {
    #min sum of weight of all observations required. control overfitting
        'min_child_weight': [1, 5, 10],
    #A node is split only when the resulting split gives a positive reduction
    #in the loss function. Gamma specifies the minimum loss reduction required
    #to make a split.
        'gamma': [0.5, 1, 1.5, 2, 5],
    #Denotes the fraction of observations to be randomly samples for each tree
        'subsample': [0.6, 0.8, 1.0],
    #Denotes the fraction of observations to be randomly samples for each tree
        'colsample_bytree': [0.6, 0.8, 1.0],
        'max_depth': [3, 4, 5]
        }
```

Output

Best Parameters

```
print(random_search.best_params_)
```

```
{'subsample': 0.8, 'min_child_weight': 5, 'max_depth': 5, 'gamma': 1, 'colsample_bytree': 0.8}
```

```
random_search.best_estimator_
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.8, gamma=1,
              learning_rate=0.1, max_delta_step=0, max_depth=5,
              min_child_weight=5, missing=None, n_estimators=100, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=0.8, verbosity=1)
```

Hyperparameters

Final score

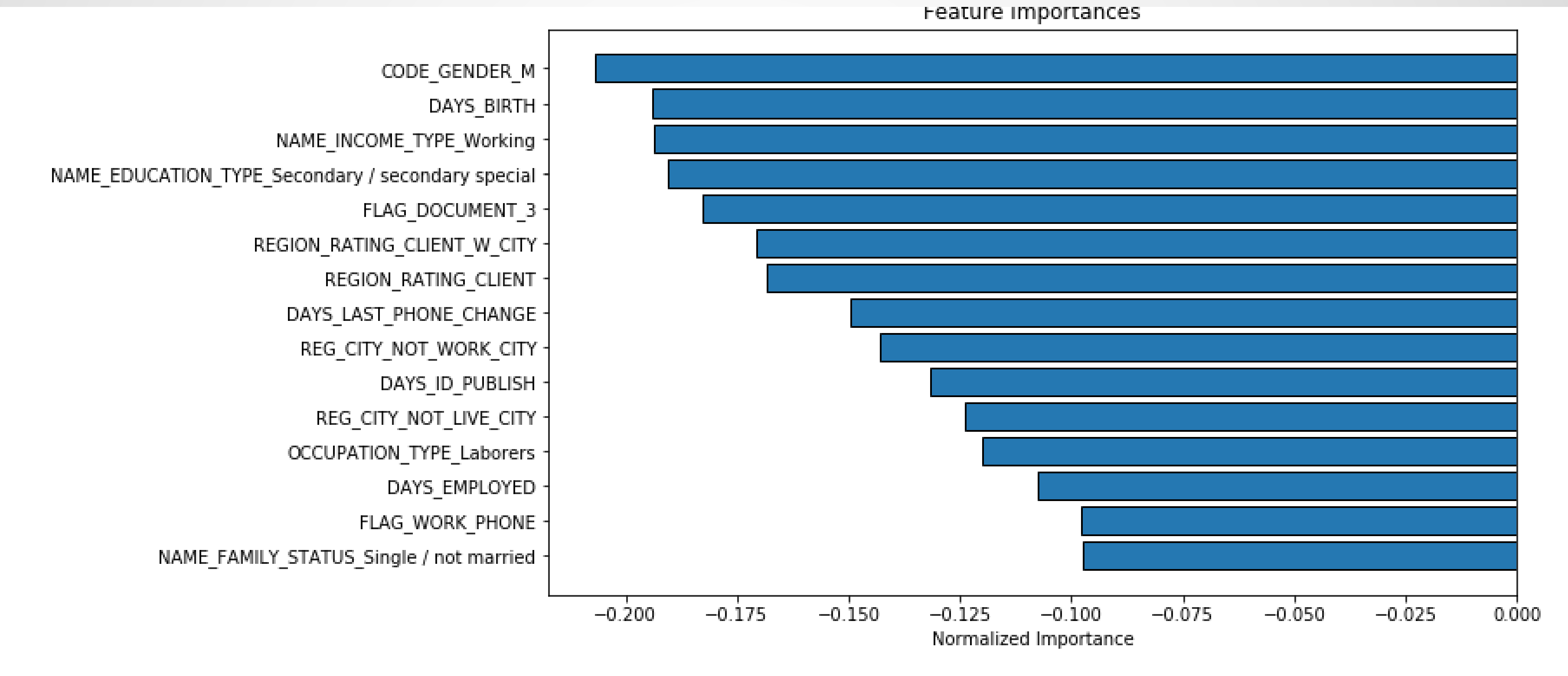
	Features	Score
0	Log reg	0.689468
1	Random forest	0.708514
2	xgboost	0.752845
3	xg_poly	0.753731
4	xg_burr	0.756839
5	random_f_burr	0.714592
6	xg_hp_burr	0.763355

05

Features importance

Feature importance

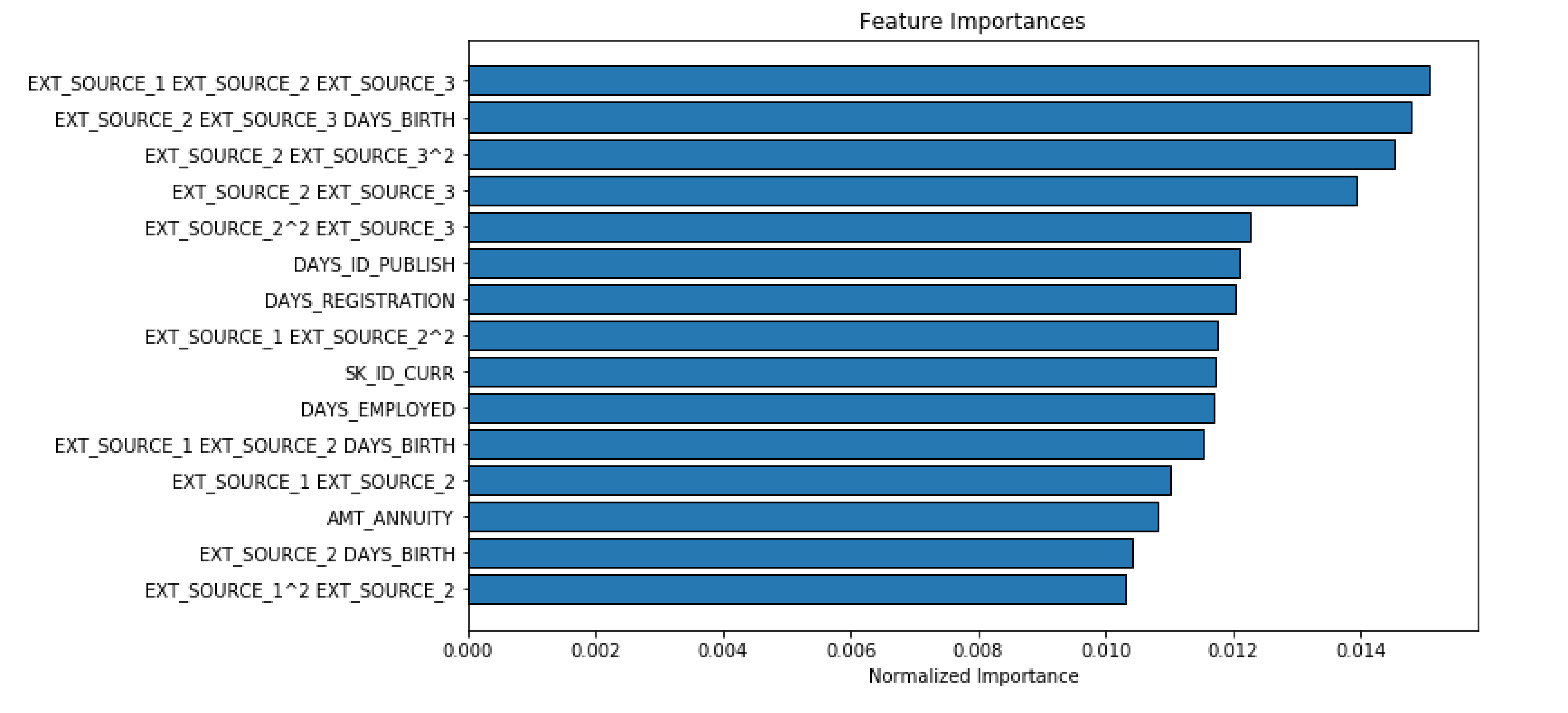
Logistic regression



Analysis
ANOA

Feature importance

Random forest

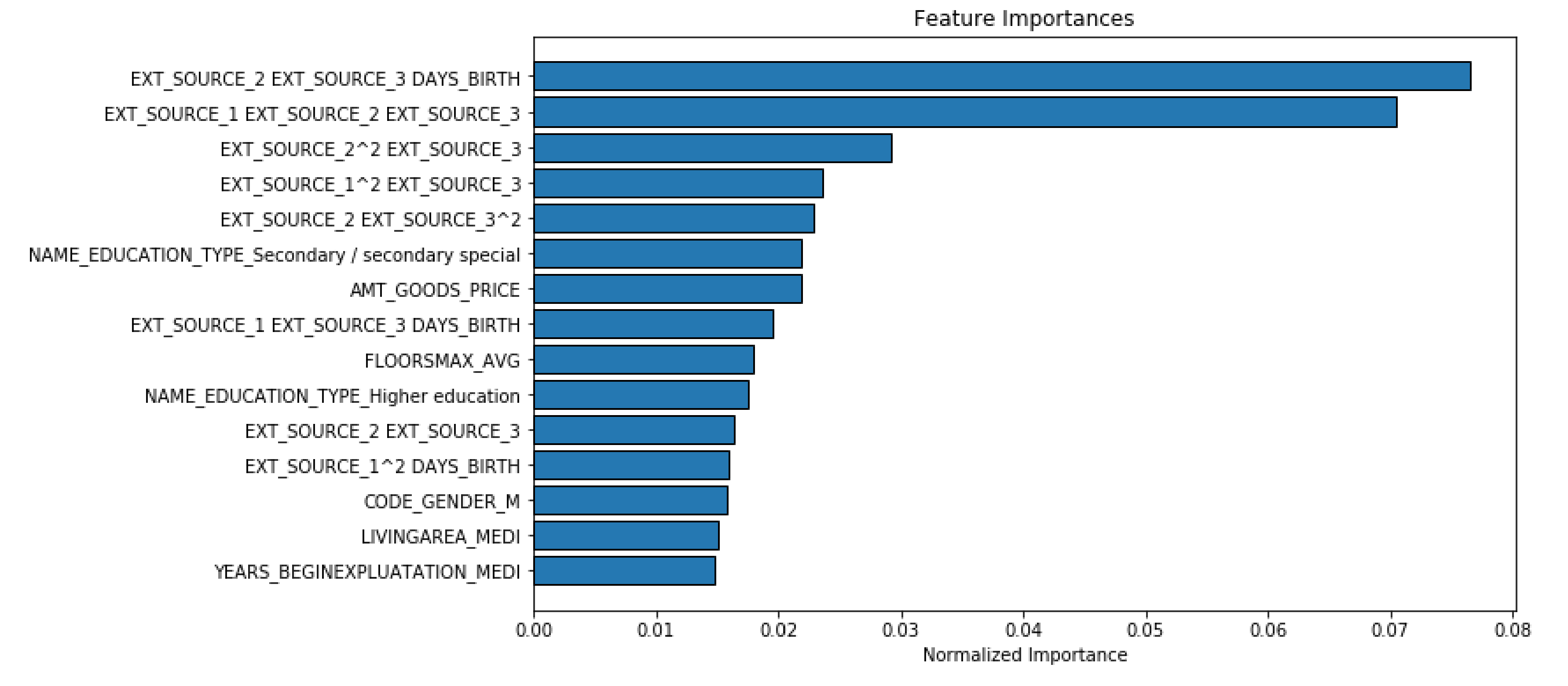


Analysis

ANOA

Feature importance

Xg boost



04

Conclusion

Conclusion

General comments

We manage to improve the score.

Best scoring model : XG Boost.

Scoring might be improved using the rest of the dataset.

Improve data preparation

Solve imbalance issue. (under/over sampling) additionaly with Crossvalidation.

Feature importance

Top feature still involved EXT SOURCE : indicator from the bank on which we have no information.

And other expected features. (age, education, etc)

Way ahead

Model could be changed to optimize to improve Precision (maximizing precision to reduce the probability to miss a default)

QUESTION ?

OpenClassRooms: Project 4

A x e l   F a v r e u l