**Home Credit Default Risk**

**Final Report**

**Table of Contents**

1. Business Problem 2
2. Business Objective 2
3. Data Acquisition 2
4. Exploratory Data Analysis (EDA) 2

4.1 Distribution of Target Column 2

4.2 Missing Values Treatment 3

4.3 Information Value Analysis 3

4.4 Selected Columns Exploration 5

1. Machine Learning Analysis 8

5.1 Data Preparation 8

5.2 Model Evaluation 8

5.3 In-Depth Analysis 10

1. Conclusions x
2. **Business Problem**

These days, most people are living with loans through their whole lives. Whole bunches of different loan products are becoming available to cover every aspect of people’s daily life. In the meantime, there will be questions arising by banks or credit unions, should be the loan approved or not? How to determine if the applicant is capable of repaying a loan? What is the default risk of each approved loan? Depending on individual’s complex situations and records, the question whether the loan should be approved is the problem resolved here.

1. **Business Objective**

The primary objective of this project is to build a machine learning model to predict the likelihood if the applicant is capable to repay the loan. Depending not only on the financial records, but also personal situations and individual backgrounds, the possibility of repaying the loan will be delivered from the machine learning model.

1. **Data Acquisition**

There are 10 datasets from the original data source: the first two application\_train and application\_test are the two main datasets with the information about each application, sorted by the feature SK\_ID\_CURR and identified by the feature TARGET representing repaid by 0 and not repaid by 1. The bureau dataset displays the applicants’ previous credits from other financial institutions. The bureau\_balance exhibits monthly balances of previous credits in the Credit Bureau. POS\_CASH\_balance shows monthly balance snapshots of the previous point of sales and cash loans that the applicant had with Home Credit. The previous\_application tells all previous application records for Home Credit loans. And all repayment history for previous loans are recorded in installments\_payment dataset. The rest are one file for explaining each column and one submission example file

The original data is available on <https://www.kaggle.com/c/home-credit-default-risk/data>.

1. **Exploratory Data Analysis (EDA)**

In this section, the main dataset will be analyzed with multiple visualization methods and their main characteristics will be summarized.

* 1. *Distribution of Target Column*

There are two values exhibited in TARGET column, representing repaid by 0 and not repaid by 1, respectively.

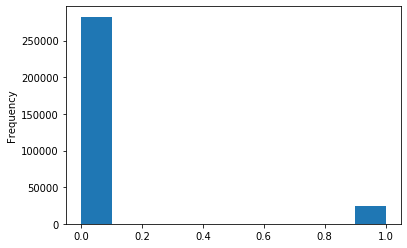


Fig 1. Distribution of TARGET Column

The TARGET feature exhibits imbalanced class problem. Much more loans were repaid, while only a small percentage of loans were not.

* 1. *Missing Values Treatment*

One of the most common problems faced in Data Cleaning/Exploratory Analysis is handling the missing values. There are several causes of missing values: sometimes values are missing because they do not exist, or because of improper collection of data or poor data entry. In that case, various filling strategies are required to operate for different situations. Here use three different filling strategies in this project:

1. Filling in with mean;
2. Filling in with most frequent value/item;
3. Filling in with 0;
   1. *Information Value Analysis*

Information value analysis helps determine which columns in a data set have predictive power or influence on the value of a TARGET variable. After calculation IV score for each column, columns with IV score between 0.01 to 0.8 are listed below:

VAR\_NAME IV

YEARS\_BUILD\_MEDI 0.010338

YEARS\_BUILD\_AVG 0.010419

BASEMENTAREA\_MEDI 0.012495

FLAG\_DOCUMENT\_6 0.012562

BASEMENTAREA\_AVG 0.012700

ENTRANCES\_AVG 0.012872

ENTRANCES\_MEDI 0.012937

LIVE\_CITY\_NOT\_WORK\_CITY 0.012981

FLOORSMIN\_MEDI 0.013600

FLOORSMIN\_AVG 0.014065

NAME\_HOUSING\_TYPE 0.014813

APARTMENTS\_MEDI 0.014901

APARTMENTS\_AVG 0.015408

DAYS\_EMPLOYED 0.015420

NAME\_CONTRACT\_TYPE 0.015777

LIVINGAREA\_MEDI 0.018547

LIVINGAREA\_AVG 0.018766

YEARS\_BEGINEXPLUATATION\_MEDI 0.020777

YEARS\_BEGINEXPLUATATION\_AVG 0.021042

YEARS\_BEGINEXPLUATATION\_MODE 0.021197

DAYS\_REGISTRATION 0.022391

EMERGENCYSTATE\_MODE 0.022947

NAME\_FAMILY\_STATUS 0.023022

REG\_CITY\_NOT\_LIVE\_CITY 0.023249

ELEVATORS\_MEDI 0.023500

ELEVATORS\_AVG 0.023766

TOTALAREA\_MODE 0.024528

FLOORSMAX\_MEDI 0.024818

FLOORSMAX\_AVG 0.024960

REGION\_RATING\_CLIENT 0.027193

FLAG\_DOCUMENT\_3 0.029107

REGION\_RATING\_CLIENT\_W\_CITY 0.029793

REG\_CITY\_NOT\_WORK\_CITY 0.031340

FLAG\_EMP\_PHONE 0.032759

AMT\_GOODS\_PRICE 0.036997

CODE\_GENDER 0.038769

DAYS\_ID\_PUBLISH 0.039190

DAYS\_LAST\_PHONE\_CHANGE 0.040036

OCCUPATION\_TYPE 0.048366

NAME\_EDUCATION\_TYPE 0.049008

NAME\_INCOME\_TYPE 0.056732

ORGANIZATION\_TYPE 0.075998

EXT\_SOURCE\_1 0.085415

DAYS\_BIRTH 0.085549

EXT\_SOURCE\_2 0.313063

EXT\_SOURCE\_3 0.313502

EXT\_SOURCE\_1, EXT\_SOURCE\_2, DAYS\_BIRTH and EXT\_SOURCE\_3 have the highest IV scores; Explore EXT\_SOURCE\_1, EXT\_SOURCE\_2 and EXT\_SOURCE\_3 first, check DAYS\_BIRTH later.

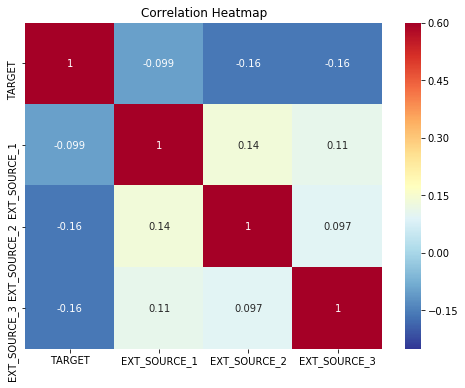


Fig 2. Correlation Heatmap

The variables EXT\_SOURCEs have the most negative correlations with the TARGET variable. When the value of the EXT\_SOURCE increases, it is more likely to repay the loan.

* 1. *Selected Columns Exploration*

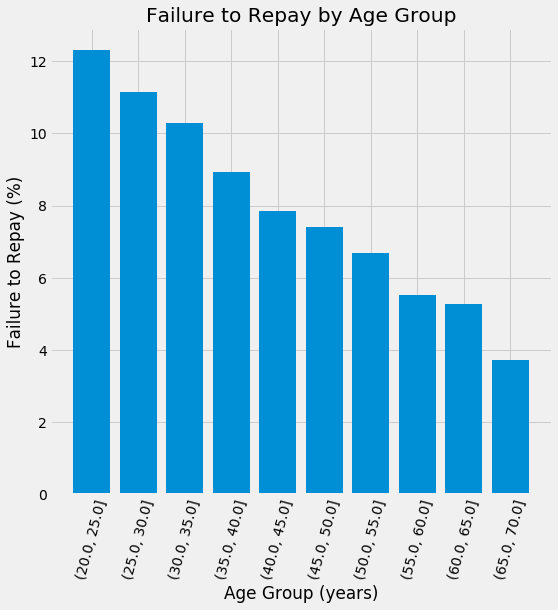


Fig 3. Failure to Repay by Age Group

It is clear to observe from figure 3 that the senior customers are more likely to repay the loan.

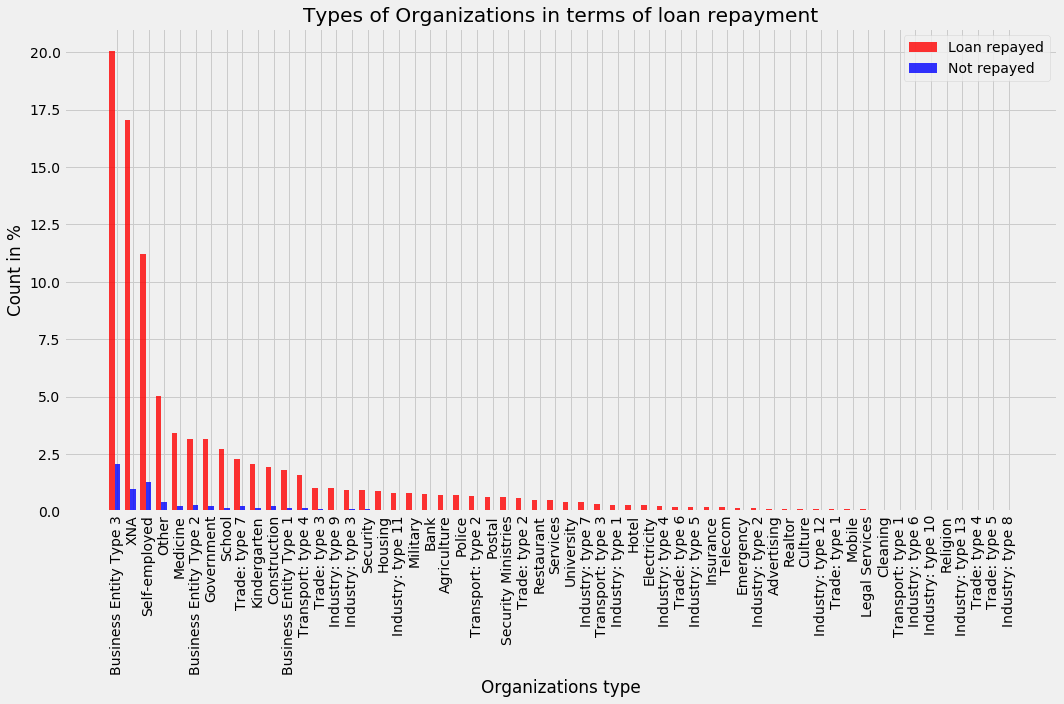


Fig 4. Types of Organizations in Terms of Loan Repayment

From Figure 4, there is not much correlations between types of organizations and loan repayment. The only conclusion from the plot is "Business Entity Type 3" tends to repay the loan comparing to other organizations.

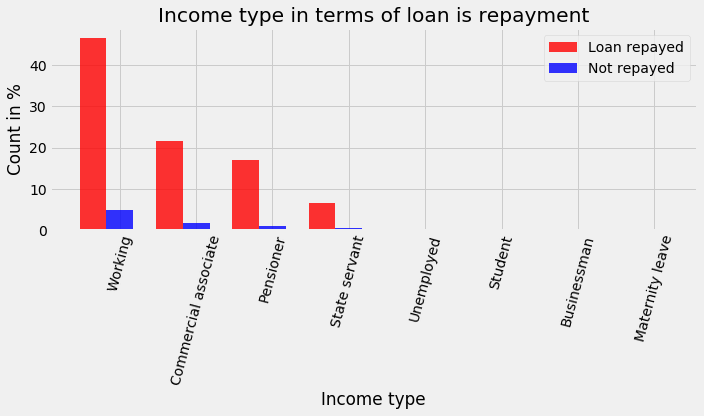


Fig 5. Income Type in Term of Loan Repayment

According to observation in Figure 5, working group tend to repay the loan comparing to other groups.

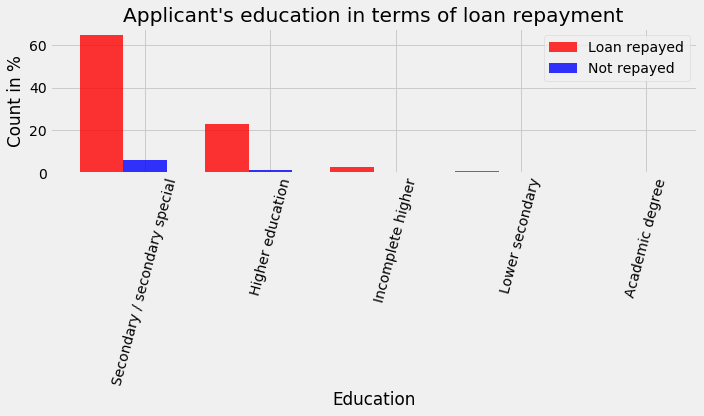


Fig 6. Applicant’s Education in Terms of Loan Repayment

From Figure 6, it is clear that secondary/secondary special tend to repay loan comparing to other groups.

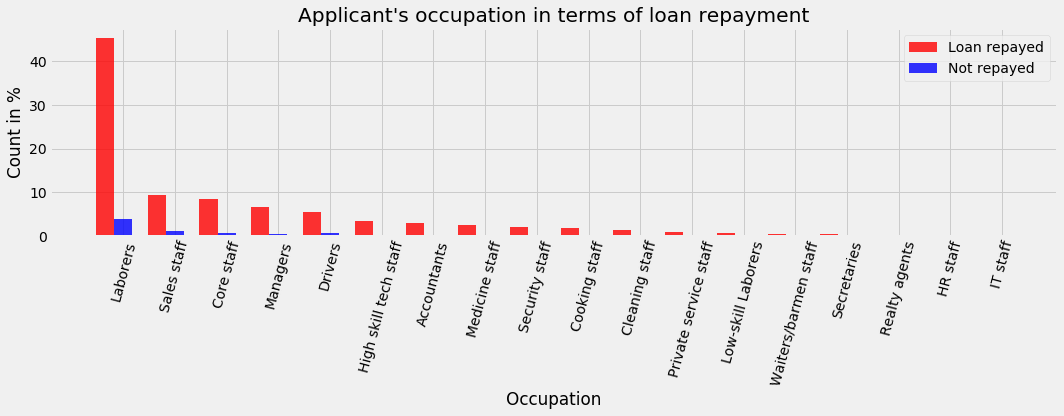


Fig 7. Applicant’s Occupation in Terms of Loan Repayment

In Figure 7, there are only a few occupations tend to repay the loan: "laborers", "Sales staff", "Core staff", "Managers", "Drivers", however, the trend is not that obvious comparing to other occupations.

1. **Machine Learning Analysis**

In this section, machine learning is utilized to build the model to predict the possibility of loan repayment.

* 1. *Data Preparation*

There are total 122 features and 67 of them have the missing values in the main dataset-“application\_train”; Most of the missing values are more than 50%, less than 80%, keep them and fill in with various strategies is wise not lose information. As mentioned early, three different filling strategies are applied in this project: 1). filling in with mean; 2). filling in with most frequent value/item; 3). filling in with 0.

Since most of machine learning models cannot handle categorical variables, encoding is necessary to perform in this section. Here, One-hot coding is applied to transform all categorical variables into numerical variables, and all original categorical columns are dropped.

Here, three different approaches are performed, in order to achieve the best machine learning model and avoid the over-fitting as well.

*Approach 1:* Model evaluations are run right after missing value filling and categorical variables encoding. The result is the baseline for model evaluation.

*Approach 2*: In order to generate more features, supplementary datasets are merged into the main dataset.

*Approach 3*: The in-depth analysis, including information value analysis and multicollinearity are applied in this approach to avoid over-fitting.

For all the approaches above, dataset are split into two parts, 70% for training and 30% for test.

* 1. *Model Evaluation*

For each approach, four machine learning models are performed, including Light Gradient Boosting, Random Forest, XGBoosting and CatBoosting. The ROC-AUC score for each model under three approaches are shown in Table 1.

Table 1. Model Comparison

|  |  |  |
| --- | --- | --- |
| **Approach** | **Model** | **ROC-AUC** |
| **1** | Lightgbm | 0.7522257459682484 |
|  | Random Forest | 0.5001905604737972 |
|  | XGBoost | 0.5083663311963755 |
|  | CatBoost | 0.5127446516061206 |
| **2** | Lightgbm | 0.7697960074520930 |
|  | Random Forest | 0.5004482826475651 |
|  | XGBoost | 0.5130236655780216 |
|  | CatBoost | 0.5217810129101755 |
| **3** | Lightgbm | 0.7146383322781984 |
|  | Random Forest | 0.5008240960706977 |
|  | XGBoost | 0.5020686407435064 |
|  | CatBoost | 0.5055747336202197 |

According to comparison in Table 1, Lightgbm model delivers the best score. From all three approaches, the confusion matrix and the ROC-AUC score using the Light Gradient Boosting model are listed below.

Approach 1

Confusion\_matrix:

[[83788 1053]

[ 6685 728]]

Classification report:

precision recall f1-score support

0 0.93 0.99 0.96 84841

1 0.41 0.10 0.16 7413

ROC-AUC score of the model: 0.7522257459682484

Approach 2

Confusion\_matrix:

[[83708 1035]

[ 6680 831]]

Classification report:

precision recall f1-score support

0 0.93 0.99 0.96 84743

1 0.45 0.11 0.18 7511

ROC-AUC score of the model: 0.769796007452093

Approach 3

Confusion\_matrix:

[[84404 339]

[ 7287 224]]

Classification report:

precision recall f1-score support

0 0.92 1.00 0.96 84743

1 0.40 0.03 0.06 7511

ROC-AUC score of the model: 0.7146383322781984

* 1. *In-Depth Analysis*

In this section, in-depth analysis, including information value analysis and multicollinearity are applied. Information value analysis helps determine which columns in a data set have predictive power or influence on the value of a TARGET variable, and multicollinearity will lead to over-fitting.

Here is the list after information value analysis, only the columns with the score between 0.01 and 0.8 are kept:

VAR\_NAME IV

2 AMT\_GOODS\_PRICE 0.033286

10 APARTMENTS\_AVG 0.015511

11 APARTMENTS\_MEDI 0.015017

13 BASEMENTAREA\_AVG 0.011850

14 BASEMENTAREA\_MEDI 0.011192

20 CODE\_GENDER\_F 0.039044

21 CODE\_GENDER\_M 0.039057

26 DAYS\_BIRTH 0.083387

27 DAYS\_EMPLOYED 0.014968

28 DAYS\_ID\_PUBLISH 0.036111

29 DAYS\_LAST\_PHONE\_CHANGE 0.041489

30 DAYS\_REGISTRATION 0.021001

33 ELEVATORS\_AVG 0.023385

34 ELEVATORS\_MEDI 0.022748

36 EMERGENCYSTATE\_MODE\_No 0.023611

38 ENTRANCES\_AVG 0.014656

39 ENTRANCES\_MEDI 0.014576

41 EXT\_SOURCE\_1 0.081742

42 EXT\_SOURCE\_2 0.311860

43 EXT\_SOURCE\_3 0.308531

58 FLAG\_DOCUMENT\_3 0.027809

61 FLAG\_DOCUMENT\_6 0.014431

66 FLAG\_EMP\_PHONE 0.034048

73 FLAG\_WORK\_PHONE 0.010034

74 FLOORSMAX\_AVG 0.024231

75 FLOORSMAX\_MEDI 0.023997

77 FLOORSMIN\_AVG 0.013064

78 FLOORSMIN\_MEDI 0.012924

85 HOUSETYPE\_MODE\_block\_of\_flats 0.022311

91 LIVE\_CITY\_NOT\_WORK\_CITY 0.013531

96 LIVINGAREA\_AVG 0.018486

97 LIVINGAREA\_MEDI 0.018302

99 MONTHS\_BALANCE 0.022486

109 NAME\_CONTRACT\_TYPE\_Cash\_loans 0.014626

110 NAME\_CONTRACT\_TYPE\_Revolving\_loans 0.014626

112 NAME\_EDUCATION\_TYPE\_Higher\_education 0.047423

115 NAME\_EDUCATION\_TYPE\_Secondary\_\_\_secondary\_special 0.033856

119 NAME\_FAMILY\_STATUS\_Single\_\_\_not\_married 0.010416

123 NAME\_HOUSING\_TYPE\_House\_\_\_apartment 0.010920

127 NAME\_HOUSING\_TYPE\_With\_parents 0.010418

131 NAME\_INCOME\_TYPE\_Pensioner 0.034384

132 NAME\_INCOME\_TYPE\_State\_servant 0.010515

135 NAME\_INCOME\_TYPE\_Working 0.046252

155 OCCUPATION\_TYPE\_Drivers 0.010323

159 OCCUPATION\_TYPE\_Laborers 0.022614

226 ORGANIZATION\_TYPE\_XNA 0.034081

228 REGION\_POPULATION\_RELATIVE 0.015727

229 REGION\_RATING\_CLIENT 0.028044

230 REGION\_RATING\_CLIENT\_W\_CITY 0.029695

231 REG\_CITY\_NOT\_LIVE\_CITY 0.023641

232 REG\_CITY\_NOT\_WORK\_CITY 0.033331

236 SK\_DPD\_DEF 0.013524

238 SK\_ID\_PREV 0.027075

239 TOTALAREA\_MODE 0.025267

244 WALLSMATERIAL\_MODE\_Panel 0.014294

254 YEARS\_BEGINEXPLUATATION\_AVG 0.020032

255 YEARS\_BEGINEXPLUATATION\_MEDI 0.019848

256 YEARS\_BEGINEXPLUATATION\_MODE 0.022206

257 YEARS\_BUILD\_AVG 0.010904

258 YEARS\_BUILD\_MEDI 0.010801

263 b\_AMT\_CREDIT\_SUM\_DEBT 0.032270

264 b\_AMT\_CREDIT\_SUM\_LIMIT 0.026124

265 b\_AMT\_CREDIT\_SUM\_OVERDUE 0.018536

267 b\_CREDIT\_ACTIVE\_Active 0.068332

269 b\_CREDIT\_ACTIVE\_Closed 0.069470

275 b\_CREDIT\_DAY\_OVERDUE 0.017212

279 b\_CREDIT\_TYPE\_Consumer\_credit 0.011726

280 b\_CREDIT\_TYPE\_Credit\_card 0.014842

281 b\_CREDIT\_TYPE\_Interbank\_credit 0.012166

287 b\_CREDIT\_TYPE\_Mobile\_operator\_loan 0.012166

291 b\_DAYS\_CREDIT 0.086050

292 b\_DAYS\_CREDIT\_ENDDATE 0.059518

293 b\_DAYS\_CREDIT\_UPDATE 0.069492

297 cc\_bal\_AMT\_BALANCE 0.029014

299 cc\_bal\_AMT\_DRAWINGS\_ATM\_CURRENT 0.029532

300 cc\_bal\_AMT\_DRAWINGS\_CURRENT 0.027768

303 cc\_bal\_AMT\_INST\_MIN\_REGULARITY 0.019826

306 cc\_bal\_AMT\_RECEIVABLE\_PRINCIPAL 0.028854

307 cc\_bal\_AMT\_RECIVABLE 0.029241

308 cc\_bal\_AMT\_TOTAL\_RECEIVABLE 0.029241

309 cc\_bal\_CNT\_DRAWINGS\_ATM\_CURRENT 0.043244

310 cc\_bal\_CNT\_DRAWINGS\_CURRENT 0.037935

312 cc\_bal\_CNT\_DRAWINGS\_POS\_CURRENT 0.021465

314 cc\_bal\_MONTHS\_BALANCE 0.011852

325 i\_AMT\_INSTALMENT 0.014360

326 i\_AMT\_PAYMENT 0.023405

327 i\_DAYS\_ENTRY\_PAYMENT 0.035098

328 i\_DAYS\_INSTALMENT 0.033975

332 p\_AMT\_ANNUITY 0.013158

335 p\_AMT\_DOWN\_PAYMENT 0.027883

355 p\_DAYS\_DECISION 0.042035

359 p\_DAYS\_LAST\_DUE\_1ST\_VERSION 0.013879

363 p\_HOUR\_APPR\_PROCESS\_START 0.015188

393 p\_NAME\_CONTRACT\_STATUS\_Approved 0.021915

428 p\_NAME\_GOODS\_CATEGORY\_XNA 0.014339

432 p\_NAME\_PAYMENT\_TYPE\_XNA 0.015405

436 p\_NAME\_PORTFOLIO\_POS 0.018143

445 p\_NAME\_SELLER\_INDUSTRY\_Consumer\_electronics 0.012638

462 p\_NAME\_YIELD\_GROUP\_low\_normal 0.027175

483 p\_RATE\_DOWN\_PAYMENT 0.014024

486 p\_SELLERPLACE\_AREA 0.023408

After performing multicollinearity, another 54 columns with high correlation are removed from the final dataset for model evaluation.

Iteration # 1

Removing b\_CREDIT\_TYPE\_Interbank\_credit with VIF of inf

Iteration # 2

Removing cc\_bal\_AMT\_RECIVABLE with VIF of 591358.400000

Iteration # 3

Removing NAME\_CONTRACT\_TYPE\_Cash\_loans with VIF of 91035.100000

Iteration # 4

Removing cc\_bal\_AMT\_TOTAL\_RECEIVABLE with VIF of 49878.600000

Iteration # 5

Removing YEARS\_BUILD\_MEDI with VIF of 48447.500000

Iteration # 6

Removing YEARS\_BEGINEXPLUATATION\_AVG with VIF of 47058.500000

Iteration # 7

Removing FLAG\_EMP\_PHONE with VIF of 16180.100000

Iteration # 8

Removing cc\_bal\_AMT\_BALANCE with VIF of 9678.600000

Iteration # 9

Removing ORGANIZATION\_TYPE\_XNA with VIF of 8113.600000

Iteration # 10

Removing i\_DAYS\_ENTRY\_PAYMENT with VIF of 6851.200000

Iteration # 11

Removing FLOORSMIN\_AVG with VIF of 1715.300000

Iteration # 12

Removing ENTRANCES\_AVG with VIF of 1473.200000

Iteration # 13

Removing NAME\_INCOME\_TYPE\_Pensioner with VIF of 1256.900000

Iteration # 14

Removing CODE\_GENDER\_F with VIF of 1203.900000

Iteration # 15

Removing FLOORSMAX\_AVG with VIF of 1093.100000

Iteration # 16

Removing LIVINGAREA\_AVG with VIF of 610.800000

Iteration # 17

Removing APARTMENTS\_MEDI with VIF of 496.100000

Iteration # 18

Removing YEARS\_BEGINEXPLUATATION\_MEDI with VIF of 429.200000

Iteration # 19

Removing BASEMENTAREA\_MEDI with VIF of 326.800000

Iteration # 20

Removing ELEVATORS\_AVG with VIF of 226.100000

Iteration # 21

Removing REGION\_RATING\_CLIENT with VIF of 189.600000

Iteration # 22

Removing YEARS\_BUILD\_AVG with VIF of 157.400000

Iteration # 23

Removing b\_CREDIT\_ACTIVE\_Closed with VIF of 124.800000

Iteration # 24

Removing cc\_bal\_AMT\_INST\_MIN\_REGULARITY with VIF of 77.800000

Iteration # 25

Removing p\_NAME\_PORTFOLIO\_POS with VIF of 51.300000

Iteration # 26

Removing i\_AMT\_INSTALMENT with VIF of 39.400000

Iteration # 27

Removing YEARS\_BEGINEXPLUATATION\_MODE with VIF of 35.700000

Iteration # 28

Removing b\_CREDIT\_TYPE\_Consumer\_credit with VIF of 34.800000

Iteration # 29

Removing DAYS\_BIRTH with VIF of 33.300000

Iteration # 30

Removing MONTHS\_BALANCE with VIF of 32.100000

Iteration # 31

Removing cc\_bal\_CNT\_DRAWINGS\_CURRENT with VIF of 27.300000

Iteration # 32

Removing LIVINGAREA\_MEDI with VIF of 23.700000

Iteration # 33

Removing p\_HOUR\_APPR\_PROCESS\_START with VIF of 23.200000

Iteration # 34

Removing REGION\_RATING\_CLIENT\_W\_CITY with VIF of 21.000000

Iteration # 35

Removing APARTMENTS\_AVG with VIF of 19.800000

Iteration # 36

Removing p\_NAME\_CONTRACT\_STATUS\_Approved with VIF of 15.700000

Iteration # 37

Removing FLOORSMAX\_MEDI with VIF of 15.200000

Iteration # 38

Removing EXT\_SOURCE\_1 with VIF of 15.100000

Iteration # 39

Removing NAME\_EDUCATION\_TYPE\_Secondary\_\_\_secondary\_special with VIF of 14.300000

Iteration # 40

Removing EMERGENCYSTATE\_MODE\_No with VIF of 14.200000

Iteration # 41

Removing b\_DAYS\_CREDIT with VIF of 13.700000

Iteration # 42

Removing NAME\_HOUSING\_TYPE\_House\_\_\_apartment with VIF of 12.900000

Iteration # 43

Removing p\_DAYS\_DECISION with VIF of 12.600000

Iteration # 44

Removing EXT\_SOURCE\_3 with VIF of 10.400000

Iteration # 45

Removing ENTRANCES\_MEDI with VIF of 9.500000

Iteration # 46

Removing REG\_CITY\_NOT\_WORK\_CITY with VIF of 9.300000

Iteration # 47

Removing EXT\_SOURCE\_2 with VIF of 8.800000

Iteration # 48

Removing FLOORSMIN\_MEDI with VIF of 8.400000

Iteration # 49

Removing cc\_bal\_AMT\_DRAWINGS\_ATM\_CURRENT with VIF of 7.600000

Iteration # 50

Removing cc\_bal\_MONTHS\_BALANCE with VIF of 7.200000

Iteration # 51

Removing FLAG\_DOCUMENT\_3 with VIF of 6.400000

Iteration # 52

Removing p\_AMT\_ANNUITY with VIF of 6.100000

Iteration # 53

Removing TOTALAREA\_MODE with VIF of 5.800000

Iteration # 54

Removing DAYS\_ID\_PUBLISH with VIF of 5.200000

1. **Conclusions**

Three approaches for data analysis are performed and four machine learning models were evaluated for each approach, respectively. Light Gradient Boosting Machine Learning model shows the best scores from 71.46% to 76.98%. The model is evaluated using the confusion matrix and the ROC-AUC score to find the best machine learning model for this project. Combining with other supplementary datasets for feature engineering was also performed, information value analysis and multicollinearity was employed to increase the model accuracy as well.