**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. It will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

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.

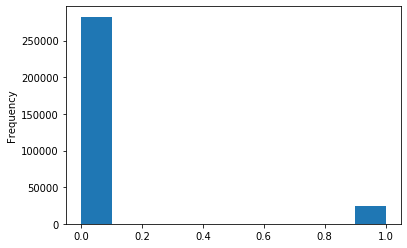


Fig 1. Distribution of TARGET Column

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

* 1. *Selected Columns Exploration*

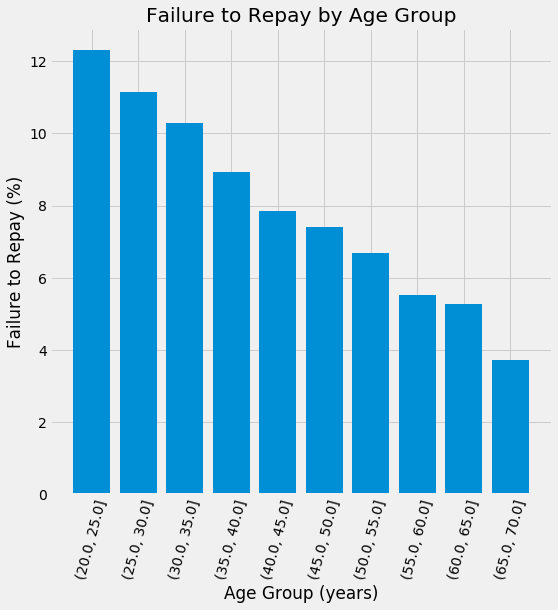


Fig 2. Failure to Repay by Age Group

It is clear to observe from figure 2 that the senior customers are more likely to repay the loan.

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.

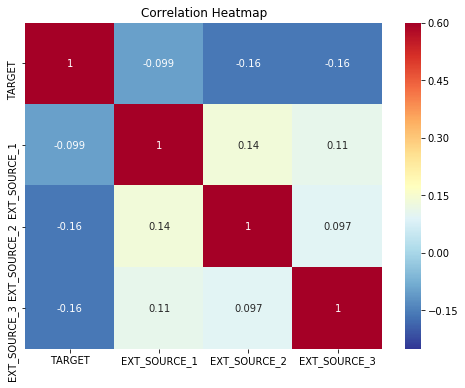


Fig 3. Correlation Heatmap

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.

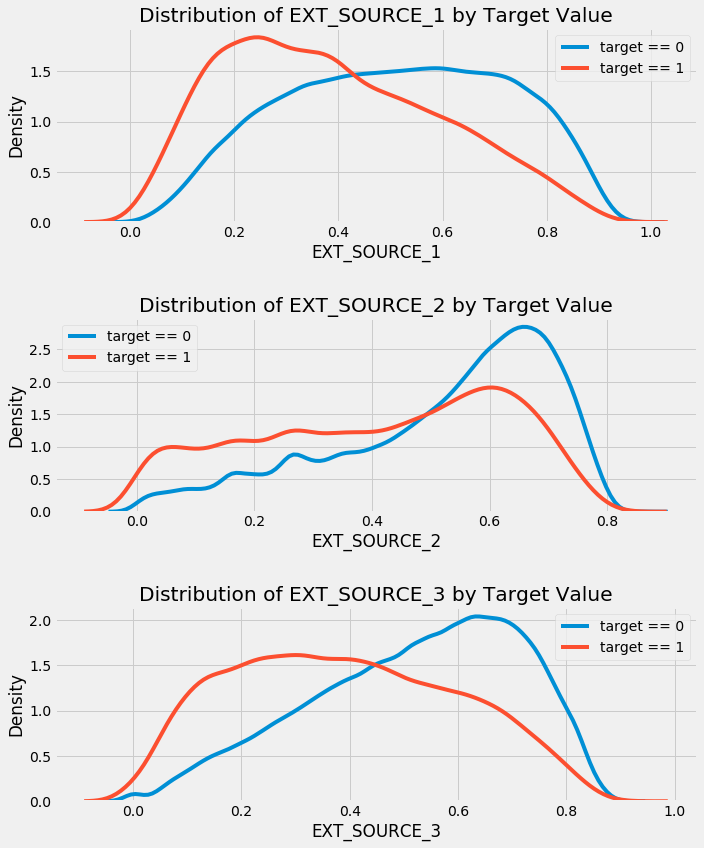


Fig 4. Pair Plots

From the plots above, EXT\_SOURCE\_3 shows the greatest difference with TARGET feature that applicants tend to repay the loan if EXT\_SOURCE\_3 is higher.

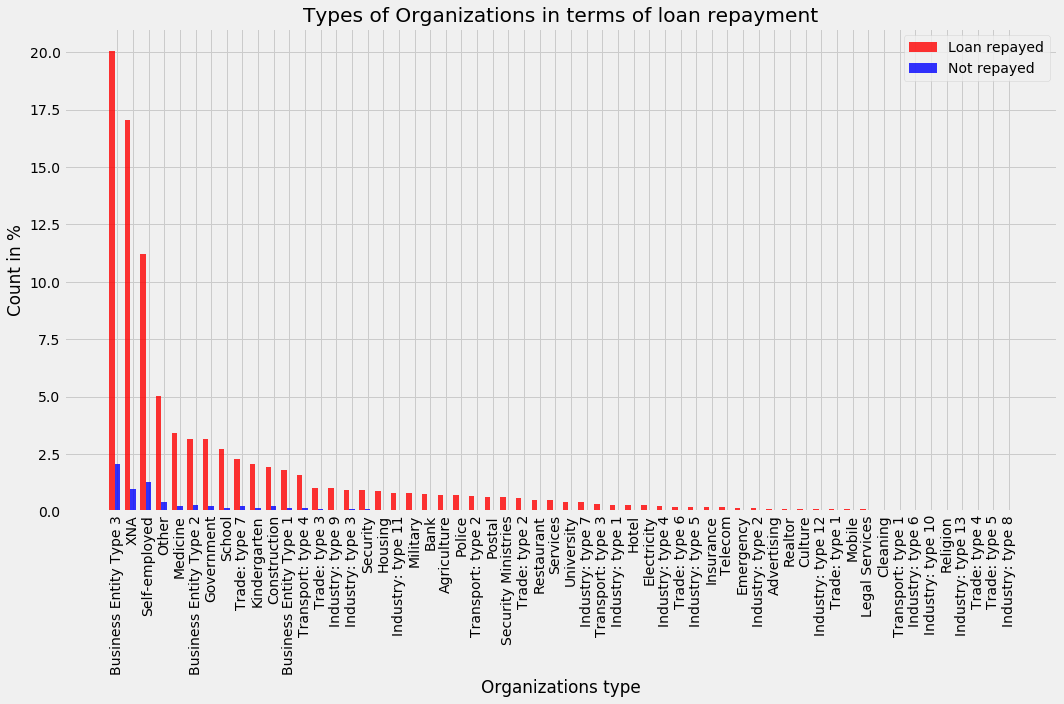


Fig 5. Types of Organizations in Terms of Loan Repayment

From Figure 5, 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.

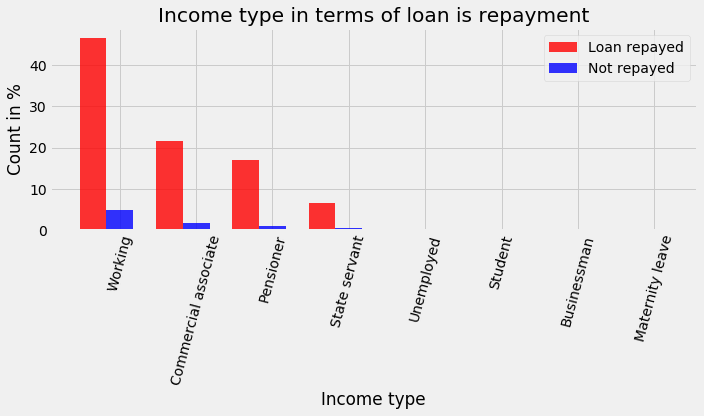


Fig 6. Income Type in Term of Loan Repayment

According to observation in Figure 6, working group tend to repay the loan comparing to other groups.

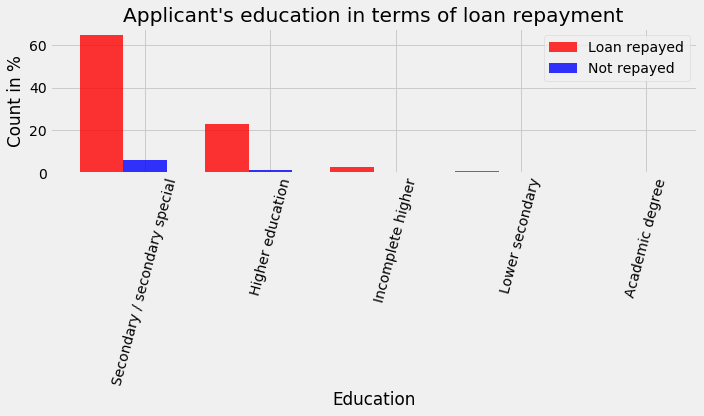


Fig 7. Applicant’s Education in Terms of Loan Repayment

From Figure 7, it is clear that secondary/secondary special tend to repay loan comparing to other groups.

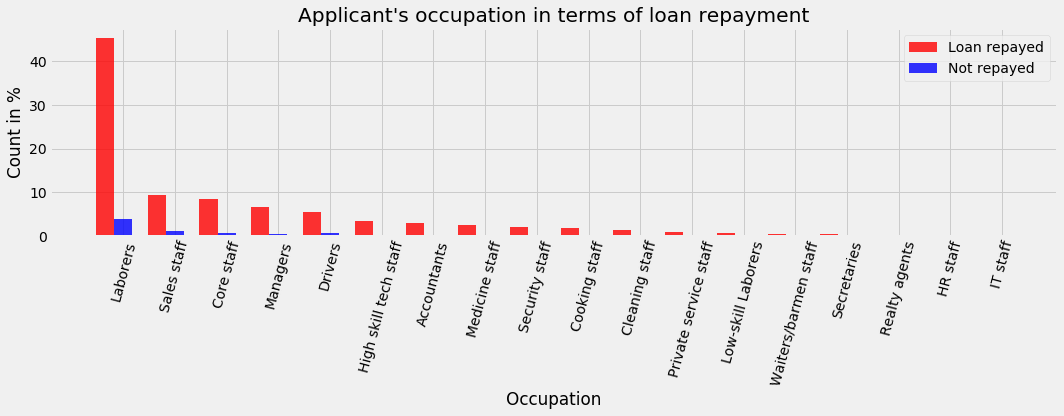


Fig 8. Applicant’s Occupation in Terms of Loan Repayment

In Figure 8, 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: Bureau, POS\_CASH\_balance, previous\_application, installments\_payments and credit\_card\_balance, 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. Only the columns with the score between 0.01 and 0.8 are kept:

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

The best ROC\_AUC score is from Light Gradient Boosting Model that shows 0.7146.

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% for each approach. Approach 1 shows the best accuracy with the highest score and Approach 3 shows the best robustness of the model. The application of three approaches will depend on the banks’ requirements, leaning to model accuracy to achieve more clients, or leaning to robustness to maintain the lowest risk.

Models are 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 robustness as well.

The machine learning models can be used for banks to determine the likelihood that applicants are capable to repay loans. It will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.