**Home Loan Credibility Assessment**

**Final Report**

**Table of Contents**

|  |  |
| --- | --- |
| 1. Business Problem | 2 |
| 1. Business Objective | 2 |
| 1. Data Acquisition | 2 |
| 1. Exploratory Data Analysis (EDA) | 2 |
| 4.1. Examine the Distribution of the Target Column | 3 |
| 4.2. Missing value treatment | 3 |
| 4.3. Encoding Categorical Variables | 3 |
| 4.4. Outlier Treatment | 3 |
| 4.5. Correlations | 4 |
| 4.6. Pairs Plot | 7 |
| 5. Machine Learning Analysis | 11 |
| 5.1. Data Preparation | 11 |
| 5.2. Model Evaluation | 12 |
| 5.3. Feature Importance | 14 |
| 6. Conclusion | 16 |

**1. Business Problem:**

Mainstream banks and financial institutions check traditional credit score models, which include demographic characteristics, historical payment data, credit bureau data and application data, to determine repayment success. However, many un-banked individuals do not have sufficient credit scores due to their past mistakes of unavoidable circumstances. Therefore, they have to deal with unconventional means such as loan sharks when borrowing money. Moreover, most of these individuals are hard-working and should get a chance to borrow money safely. It is important to identify these individuals from financial records to provide a positive and safe borrowing experience.

**2. Business Objective:**

The primary object of this project is to build a model from the financial data to predict the likelihood that an applicant will experience difficulty in repaying their loan. The output of the proposed model is the probability that determines an applicant in terms of having at least one late payment when repaying their loan.

**3. Data Acquisition:**

The data is acquired from seven different sources. The first dataset, application\_train/ application\_test, is the main training and testing data with information about each loan application. Each row is identified by the feature SK\_ID\_CURR. The TARGET feature in the training data represents load repaid by 0 and not repaid by 1. The second data source bureau provides the client's previous credits from other financial institutions. The third one is bureau\_balance, which provides monthly balances of previous credits in the Credit Bureau. POS\_CASH\_balance provides monthly balance snapshots of the previous point of sales and cash loans that the applicant had with Home Credit. The fifth data source credit\_card\_balance presents the monthly balance snapshots of previous credit cards that the applicant has with Home Credit. All previous applications for Home Credit loans of clients who have loans are mentioned in the previous\_application. Repayment history for the previous loans is provided in the seventh data source installments\_payment.

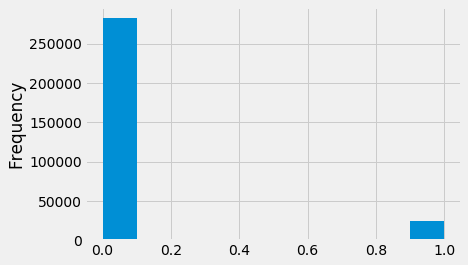
The data is available on <https://www.kaggle.com/c/home-credit-default-risk/data>.

**4. Exploratory Data Analysis (EDA):**

In this section, we calculate statistics and use visualization methods to find trends, anomalies, patterns, or relationships within the data.

*4.1. Examine the Distribution of the Target Column:*

The target column contains two values: 0 for the loan was repaid on time and 1 indicating the client had payment difficulties. The target represents an imbalanced class problem, where far more loans that were repaid on time than loans that were not repaid. This problem can be solved by using machine learning models.

  
Fig. 1. Distribution of the Target column

*4.2. Missing value treatment:*

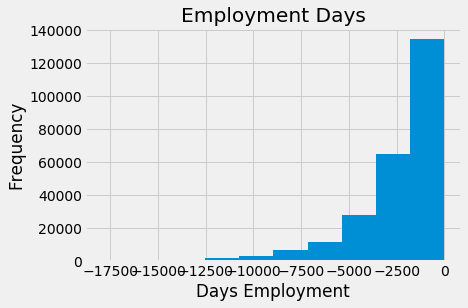
Missing data can lead to a wrong prediction or classification. There are 122 columns in the training data and 67 of them contain missing values. Most of the missing value columns contain more than 50% of missing values. Deleting the missing values is not a good option for this case. We can use imputation to fill in the missing values. XGBoost, LightGBM or another algorithm can be used to deal with the missing values.

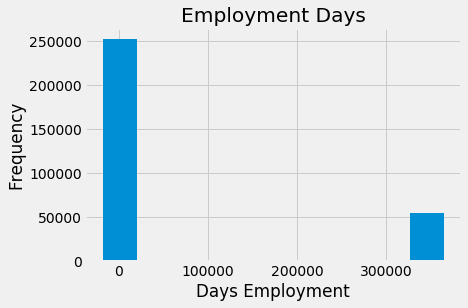
*4.3. Encoding Categorical Variables:*

Most of the machine learning models cannot deal with categorical variables. Therefore, we will encode the categorical variables as numbers. We will use Label Encoding, which does not create new columns, for those categorical variables which have only two categories. In our training data set we have only 3 such categorical variables. The rest of the categorical variables are encoded using One-Hot Encoding that creates a new column for each unique category.

*4.4. Outlier Treatment:*

The days of employment column contains outliers. The maximum employment days can not be 1000 years. All the outliers in days of employment column have the exact same value. Therefore, we will fill in the anomalous values with not a number (np.nan) and then create a new boolean column indicating whether or not the value was anomalous.

  
Fig. 3. Outlier treatment

  
Fig. 2. Outlier detection

*4.5. Correlations:*

The Pearson correlation coefficient gives the correlations between a variable and the target. It gives us an idea of possible relationships within the data. The most positive and negative top 15 correlation are presented below:

Most Positive Correlations:

OCCUPATION\_TYPE\_Laborers: 0.043019

FLAG\_DOCUMENT\_3: 0.044346

REG\_CITY\_NOT\_LIVE\_CITY: 0.044395

FLAG\_EMP\_PHONE: 0.045982

NAME\_EDUCATION\_TYPE\_Secondary / secondary special: 0.049824

REG\_CITY\_NOT\_WORK\_CITY: 0.050994

DAYS\_ID\_PUBLISH: 0.051457

CODE\_GENDER\_M: 0.054713

DAYS\_LAST\_PHONE\_CHANGE: 0.055218

NAME\_INCOME\_TYPE\_Working: 0.057481

REGION\_RATING\_CLIENT: 0.058899

REGION\_RATING\_CLIENT\_W\_CITY: 0.060893

DAYS\_EMPLOYED: 0.074958

DAYS\_BIRTH: 0.078239

Most Negative Correlations:

EXT\_SOURCE\_3: -0.178919

EXT\_SOURCE\_2: -0.160472

EXT\_SOURCE\_1: -0.155317

NAME\_EDUCATION\_TYPE\_Higher education: -0.056593

CODE\_GENDER\_F: -0.054704

NAME\_INCOME\_TYPE\_Pensioner: -0.046209

DAYS\_EMPLOYED\_ANOM: -0.045987

ORGANIZATION\_TYPE\_XNA: -0.045987

FLOORSMAX\_AVG: -0.044003

FLOORSMAX\_MEDI: -0.043768

FLOORSMAX\_MODE: -0.043226

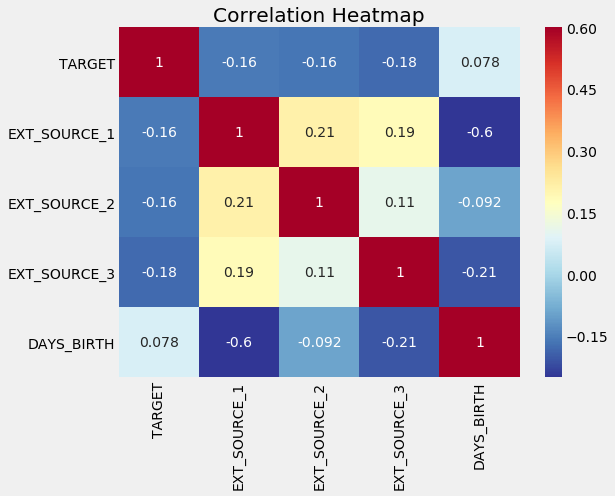
EMERGENCYSTATE\_MODE\_No: -0.042201

HOUSETYPE\_MODE\_block of flats: -0.040594

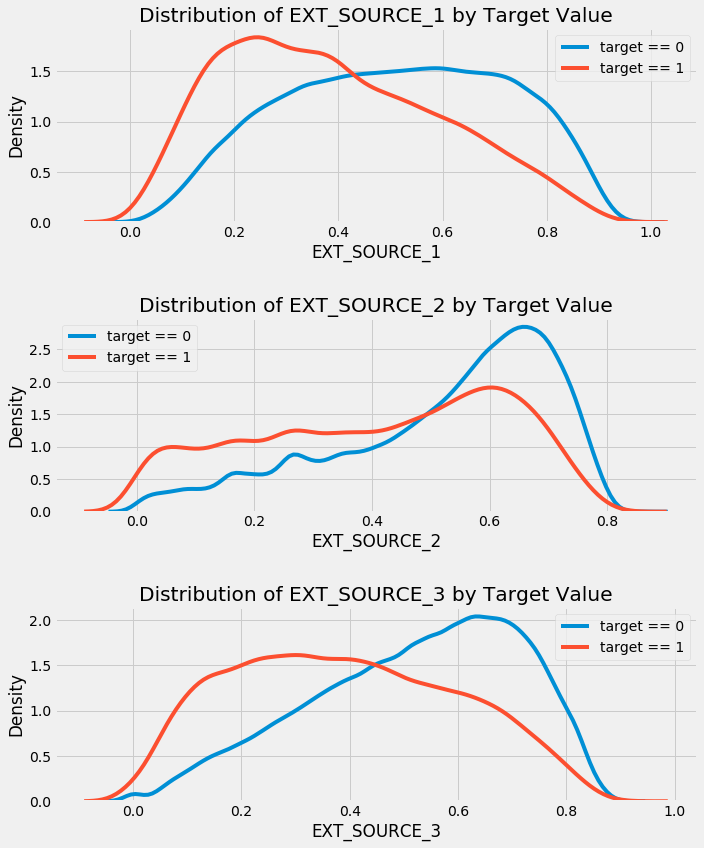
AMT\_GOODS\_PRICE: -0.039645

REGION\_POPULATION\_RELATIVE: -0.037227

EXT\_SOURCE\_3, EXT\_SOURCE\_2, and EXT\_SOURCE\_1 are the variables that have the strongest negative correlations, and DAYS\_BIRTH has the strongest positive correlations with the TARGET variable.

  
Fig. 4. Correlation heatmap

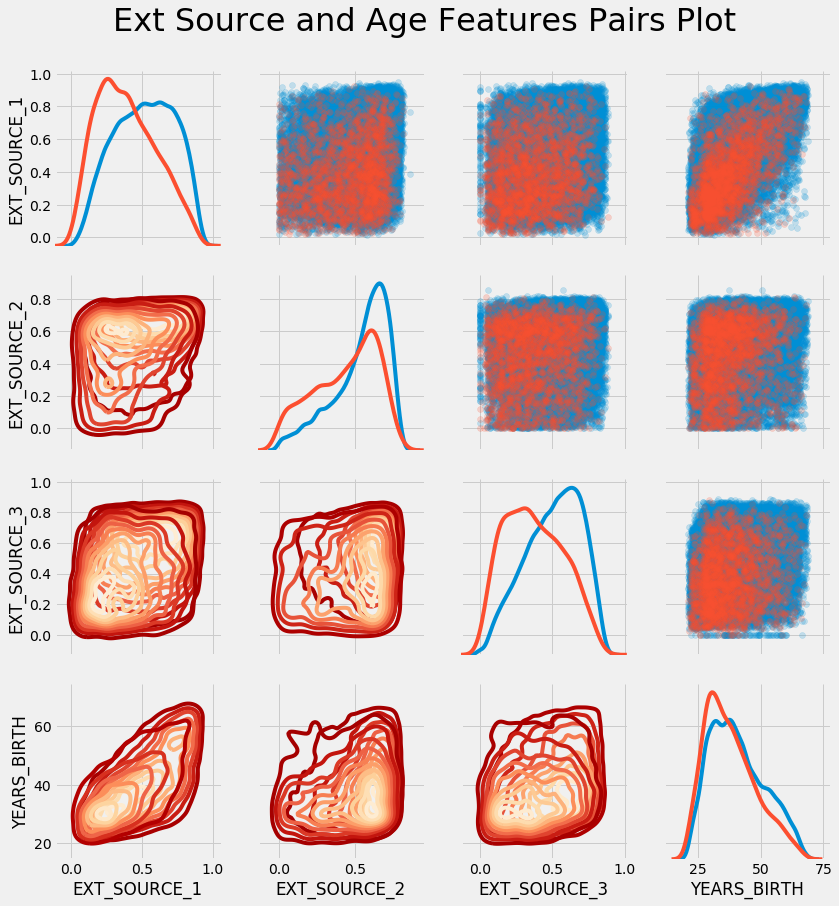
The variables EXT\_SOURCE have negative correlations with the TARGET variable. The cent is more likely to repay the loan when the value of the EXT\_SOURCE increases.

  
Fig. 5. Distribution of Exterior Sources by Target Value

From Fig. 5, we can observe that the variable EXT\_SOURCE\_3 shows the greatest difference between the values of the target. The client is more likely to repay the loan if the value of the EXT\_SOURCE\_3 variable is higher.

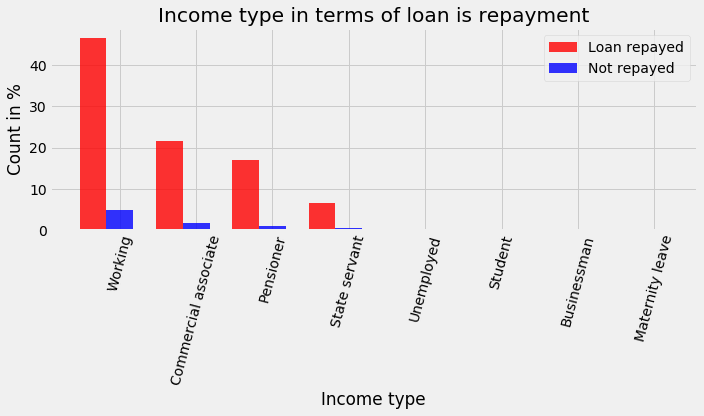
*4.6. Pairs Plot:*

The Pairs Plot lets us see relationships between multiple pairs of variables as well as distributions of single variables.

  
Fig. 6. Exterior Sources and Age Features Pairs Plot

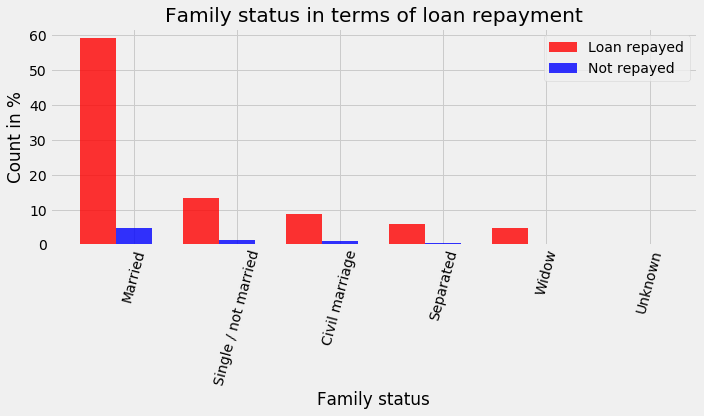
In this plot, the red indicates loans that were not repaid and the blue are loans that are paid. There is a moderate positive linear relationship between the variables EXT\_SOURCE\_1 and the DAYS\_BIRTH.

Observe the income type of the client's in terms of the loan is re-payed or not:

  
Fig. 7. Income type in terms of loan is repayment

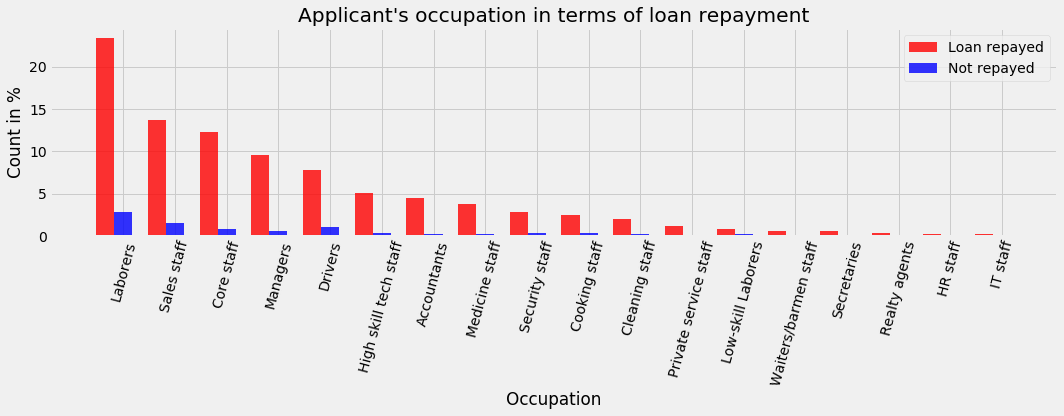
In Fig. 7, we have observed the Income type of the clients in terms of the loan is repaid or not in percentage. The client's with income type "working" tends to repay loans.

*Family status in terms of the loan is re-payed or not:*

  
Fig. 8. Family status in terms of loan repayment

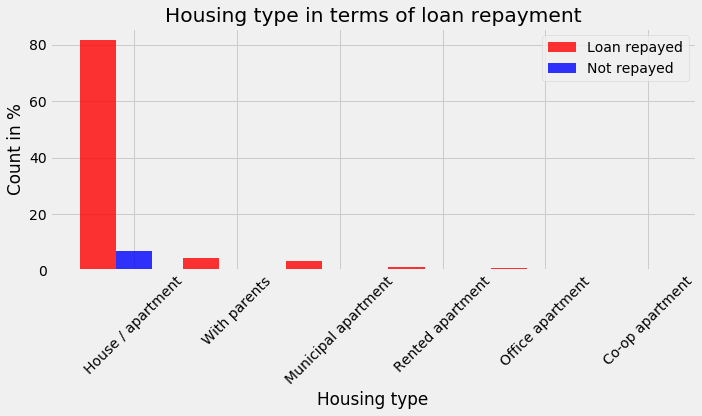
We can observe from Fig. 8 that the percentage of the client with family status as "Married" tends to repay loans.

*Occupation of the applicant's in terms of the loan is re-payed or not*:

  
Fig. 9. Applicant's occupation in terms of loan repayment

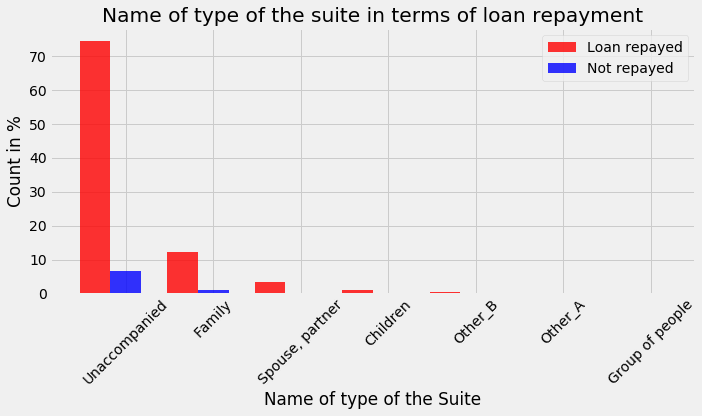
Clients with occupation "Laborers" are better at loan repayment followed by "Sales staff", "core staff", "Managers", and "Drivers". However, the difference in percentage is not significant.

*Applicant's Housing type in terms of the loan is re-payed or not:*

  
Fig. 10. Housing type in terms of loan repayment

The applicants who live in the "House/apartment" are more likely to repay the loan.

*Name of the type of the suite of the applicants in terms of the loan is re-payed or not:*

  
Fig. 11. Name of type of the suite in terms of loan repayment

From Fig. 11, one can observe that the applicants who are "Unaccompanied" are more likely to repay the loan.

**5. Machine Learning Analysis:**

After analyzing the data, we will find out the machine learning model that works best for our project in this section.

*5.1. Data Preparation:*

We have found out that the data shows an imbalanced class problem, that contains far more records of the loan paid instances than loans not repaid. Therefore, we have explored the under-sampling method on the loan paid samples to prepare the data for the machine learning models.

Out of 122 features, 67 of them have the missing values, and most of them have more than 50%. Therefore, I have used three methods to deal with them, one, delete all the rows with missing values, two, use SimpleImputer to fill the missing values, and three, use machine learning models to deal with the missing values.

Most of the machine learning model cannot deal with categorical variables. Therefore, we will encode these variables using the One-Hot Encoding, as numbers before handing them off to the model.

To find out the best machine learning model for our project, we have explored three different data preparation approach, they are:

Method 1: We prepared the data by deleting rows with the mission values and under-sampling the loan repaid data.

Method 2: We prepared the data by deleting rows with mission values, under-sampling the loan repaid, and dropping columns using multicollinearity analysis.

Method 3: We prepared the data by using the imputation to handle the mission values.

We have split the data into two parts, 70% for training and 30% for testing.

*5.2. Model Evaluation:*

First, we will evaluate the three data preparation methods with the confusion matrix and the ROC-AUC score using the Light Gradient Boosting model.

Method 1:

Confusion\_matrix:

[[314 126]

[ 66 94]]

Classification report:

precision recall f1-score support

0 0.83 0.71 0.77 440

1 0.43 0.59 0.49 160

ROC-AUC score of the model: 0.7112357954545454

Method 2:

Confusion\_matrix:

[[312 128]

[ 67 93]]

Classification report:

precision recall f1-score support

0 0.82 0.71 0.76 440

1 0.42 0.58 0.49 160

ROC-AUC score of the model: 0.7150710227272726

Method 3:

Confusion\_matrix:

[[70550 14291]

[ 3664 3749]]

Classification report:

precision recall f1-score support

0 0.95 0.83 0.89 84841

1 0.21 0.51 0.29 7413

ROC-AUC score of the model: 0.7532392541750991

From our experiments, we have observed that Method 3 produces the best accuracy. It does not employ under-sampling, multicollinearity, or deletion of rows with missing values. Therefore, we will use Method 3 to build our machine learning model. Now, we will find out the machine learning model that works best for our project.

Table 1. Model Comparison

|  |  |
| --- | --- |
| **Model** | **ROC-AUC score** |
| Light Gradient Boosting | 0.7572148211583262 |
| Gradient Boosting | 0.7541132339818492 |
| Ridge Regression | 0.746597202421708 |
| Linear Regression | 0.7465839697165486 |
| Adaptive Boost Regressor | 0.7114294585068995 |
| Decision Tree Regression | 0.7114294585068995 |
| Random Forest | 0.6387905135104288 |
| Lasso Regression | 0.6315649732588609 |
| Logistic Regression | 0.627274997423726 |

In this project, we will use the Light Gradient Boosting to build our machine learning model, which gives the best score.

*5.3. Feature Importance:*

We have studied the importance of the features to improve our model. The following new features are created:

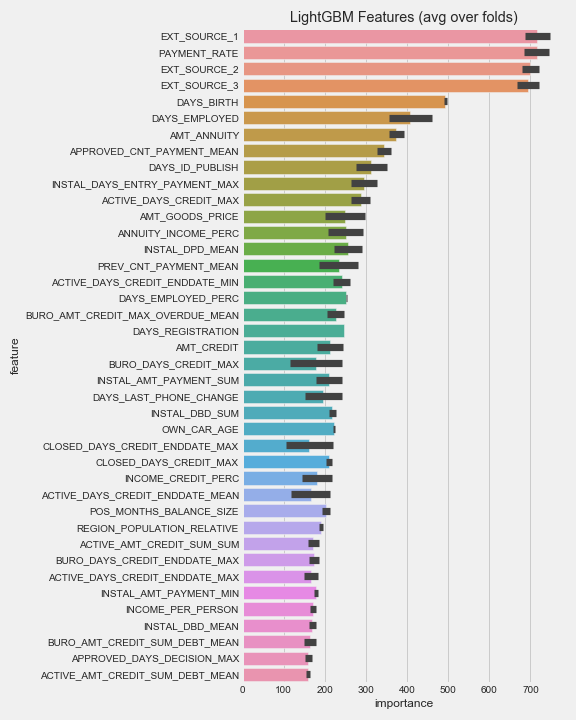
* The 'DAYS\_EMPLOYED\_PERC' feature is derived by dividing 'DAYS\_EMPLOYED' by 'DAYS\_BIRTH'.
* The 'INCOME\_CREDIT\_PERC' feature is derived by dividing 'AMT\_INCOME\_TOTAL' by 'AMT\_CREDIT'.
* The 'INCOME\_PER\_PERSON' feature is derived by dividing 'AMT\_INCOME\_TOTAL' by 'CNT\_FAM\_MEMBERS'.
* The 'ANNUITY\_INCOME\_PERC' feature is derived by dividing 'AMT\_ANNUITY' by 'AMT\_INCOME\_TOTAL'.
* The 'PAYMENT\_RATE' feature is derived by dividing 'AMT\_ANNUITY' by 'AMT\_CREDIT'.

Fig. 12, illustrates the importance of the features. In the following table, we will explore the ROC\_AUC score with the best features.

Table 2. ROC\_AUC score with best features

|  |  |
| --- | --- |
| **Best Features** | **ROC\_AUC Score** |
| 10 | 0.749384 |
| 25 | 0.756939 |
| 50 | 0.757502 |
| 60 | 0.757414 |
| 70 | 0.757372 |
| 80 | 0.757528 |
| 90 | 0.757972 |
| 100 | 0.757626 |
| 150 | 0.757246 |
| 200 | 0.757291 |
| 240 | 0.757362 |
| 245 | 0.757636 |

We can observe from the result that the model with the top 90 features gives us the best accuracy.

  
Fig. 12. Feature Importance

6. Conclusion:

From our analysis, we have studied that the Light Gradient Boosting machine learning model obtains the best accuracy of 75.72%. We have encoded the categorical variables and performed outlier treatment on the data. However, we have not employed under-sampling, multicollinearity, or deletion of rows with missing values to increase the accuracy. The model is evaluated using the confusion matrix and the ROC-AUC score to find the best machine learning model for our project. We have created five new features and the analysis shows that the new features reside in the top important features.