Project Title: Default risk analysis

# Introduction

Default risk assessment is of utmost importance for the financial institutions to avoid lending money to financially unstable and credit-unworthy borrowers. Using risk analysis one can evaluate the capacity of a borrower to reach the financial obligations. Through this analysis a risk rating is assigned by estimating the probability of default by the borrower at a given confidence level, and by estimating the amount of loss that the lender would suffer in case of default. A detailed analysis includes ratio and trend analysis, detailed analysis of cash flows, creation of projections, examination of collateral and other sources of repayment as well as credit history and management ability. A bank will look into all this factors before approving any loan to a borrower.

Home Credit is a financial organization which provides loans to people with little to no credit history. While this is a great step towards including a large un-banked population into safer loan practice, it is also important to predict their repayment ability to avoid greater financial loss. To do this Home Credit hosted a default risk prediction contest on Kaggle. Numerous historical information were provided as input features including amount of the installment, the annuity, the total credit amount, and various categorical features like what was the loan for, demographic information about the customers, gender, their job type, their income, ratings about their house etc. The objective is to employ machine learning algorithms using all those input features in order to predict if a customer will default on their loan in the future. The previous works showed that neural networks are more accurate, flexible and robust than conventional statistical methods for the assessment of credit risk. However, in this study I will be using random forest, and LightGBM algorithms to predict the default risk based on clients’ attributes, and compare their prediction accuracy. The performance of each method will be compared in terms of their ability to predict the default risk using receiver operating characteristic (ROC) area under the curve (AUC). The aim is to build ML algorithms to achieve an ROC AUC value of atleast 0.79.

# Data Availability

This project was initially organized by Home Credit group and posted on [Kaggle](https://www.kaggle.com/c/home-credit-default-risk/overview) as a contest. The complete dataset required for this project can be obtained from the [Data](https://www.kaggle.com/c/home-credit-default-risk/data) page of the site. The files included are as follows:

* 1. application train(/test).csv: This is the main table with train and test samples. The data includes, loan type (cash or revolving), gender and age of client, if the client has car, hous or flat, number of client’s children, client’s income, income type, living status and family status, credit amount of loan, loan annuity, price of good the loan approved for, client’s education etc.
  2. bureau.csv: It contains application data from previous loans that client received from other financial institutions as reported to credit bureau. The data includes status of the credit bureau reported credits, type of credit bureau credit, current credit amount for credit bureau, current debt on credit bureau credit, time scale of application activities, etc.
  3. bureau\_balance.csv: This file contains data on status of credit bureau loan during the month, month of balance relative to application, etc.
  4. POS\_CASH\_balance.csv: This file contains info on month of balance relative to application date, term of previous credit, installments left to pay on the previous credit, contract status during the month, etc.
  5. credit\_card\_balance.csv: This contains data mainly from during the month of previous credit like month of balance relative to application date, balance, credit card limit, number and amount of drawing (at ATM from others), Minimal installments, total payment received (and receivable) from client, contract status and product type, etc.
  6. previous\_application.csv: This contains information about past applications such as annuity of previous application, final credit amount applied for and finally approved, down payment amount and rate, goods price and type, interest rate, purpose of cash loan, payment method, reason for rejection in previous loans, client’s age at that time, etc
  7. installments\_payments.csv: This file contains information about installment payments for previous credits.

# Data preparation

There are 8 data tables with applicant data, i.e., applicant\_train, applicant\_test, bureau, bureau\_balance, credit\_card\_balance, installment\_payments, POS\_CASH\_balance, and previous application. I went through several steps to examine each feature in those tables.

1. variable type: float/integer/categorical, and unique value for integer/categorical
2. % missing data and dealing with them
3. Assess frequency distribution
4. Create statistical summary
5. Find anomalies if any and dealing with them
6. Assess correlation with target variable
7. Merging with train table for features from other tables
8. Saving memory by converting the data type, object type with category, float64 with float32 and int64 with int32. Also the features are deleted from memory after they are done working with.

For categorical variables with 2 unique categories, we will use label encoding, and with more than 2 unique categories, we will use one-hot encoding. The purpose is to keep both the values in one column when unique category is 2. Thus the number of column is same here.

At first we will look into the train and test dataset.

## Train and Test dataset

These tables contain information about loan application at Home Credit. The train table has 122 features and 307511 observations. The test table on the other hand has 121 features and 48744 observations. Train table has one extra target feature with ‘0’ and ‘1’ as outputs, where ‘0’ means loan repaid without difficulty and ‘1’ means difficulty in loan repayment. Among 307511 applicants, 282686 (92%) repaid the loan on time. 24825 (8%) had difficulties in repaying. Among 122 variables in train dataset, 67 variables have missing values. Among them 57 variables have more than 10% value missing, 50 variables have more than 30% values missing, and 41 variables have more than 50% values missing. The variable types are float (65), integer (41) and categorical (16). Among them 4 categorical variables have 2 unique values. How we deal with the variables depends on the feature itself. Variable description, type, % missing and the measure to replace the missing values in train and test dataset are summarized in table 1.

Table 1. Train/test data summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Description | Data type | % Missing | Replacement |
| AMT\_ANNUITY | Loan annuity | float | 0.0039 | median |
| AMT\_CREDIT | Credit amount of the loan | float | 0.0000 | NA |
| AMT\_GOODS\_PRICE | For consumer loans it is the price of the goods for which the loan is given | float | 0.0904 | median |
| AMT\_INCOME\_TOTAL | Income of the client | float | 0.0000 | NA |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | Number of enquiries to Credit Bureau about the client one day before application | float | 13.5016 | 0.00 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | Number of enquiries to Credit Bureau about the client one hour before application | float | 13.5016 | 0.00 |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | Number of enquiries to Credit Bureau about the client one month before application | float | 13.5016 | 0.00 |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | Number of enquiries to Credit Bureau about the client 3 month before application | float | 13.5016 | 0.00 |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | Number of enquiries to Credit Bureau about the client one week before application | float | 13.5016 | 0.00 |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | Number of enquiries to Credit Bureau about the client one day year | float | 13.5016 | 0.00 |
| APARTMENTS\_AVG | Normalized information about building where the client lives | float | 50.7497 | 0.00 |
| APARTMENTS\_MEDI | Normalized information about building where the client lives | float | 50.7497 | 0.00 |
| APARTMENTS\_MODE | Normalized information about building where the client lives | float | 50.7497 | 0.00 |
| BASEMENTAREA\_AVG | Normalized information about building where the client lives | float | 58.5160 | 0.00 |
| BASEMENTAREA\_MEDI | Normalized information about building where the client lives | float | 58.5160 | 0.00 |
| BASEMENTAREA\_MODE | Normalized information about building where the client lives | float | 58.5160 | 0.00 |
| CNT\_FAM\_MEMBERS | How many family members does client have | float | 0.0007 | median |
| COMMONAREA\_AVG | Normalized information about building where the client lives | float | 69.8723 | 0.00 |
| COMMONAREA\_MEDI | Normalized information about building where the client lives | float | 69.8723 | 0.00 |
| COMMONAREA\_MODE | Normalized information about building where the client lives | float | 69.8723 | 0.00 |
| DAYS\_LAST\_PHONE\_CHANGE | How many days before application did client change phone | float | 0.0003 | median |
| DAYS\_REGISTRATION | How many days before the application did client change his registration | float | 0.0000 | NA |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings defaulted on 30 DPD (days past due) | float | 0.3320 | 0.00 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings defaulted on 60 (days past due) DPD | float | 0.3320 | 0.00 |
| ELEVATORS\_AVG | Normalized information about building where the client lives | float | 53.2960 | 0.00 |
| ELEVATORS\_MEDI | Normalized information about building where the client lives | float | 53.2960 | 0.00 |
| ELEVATORS\_MODE | Normalized information about building where the client lives | float | 53.2960 | 0.00 |
| ENTRANCES\_AVG | Normalized information about building where the client lives | float | 50.3488 | 0.00 |
| ENTRANCES\_MEDI | Normalized information about building where the client lives | float | 50.3488 | 0.00 |
| ENTRANCES\_MODE | Normalized information about building where the client lives | float | 50.3488 | 0.00 |
| EXT\_SOURCE\_1 | Normalized score from external data source | float | 56.3811 | 0.00 |
| EXT\_SOURCE\_2 | Normalized score from external data source | float | 0.2146 | 0.00 |
| EXT\_SOURCE\_3 | Normalized score from external data source | float | 19.8253 | 0.00 |
| FLOORSMAX\_AVG | Normalized information about building where the client lives | float | 49.7608 | 0.00 |
| FLOORSMAX\_MEDI | Normalized information about building where the client lives | float | 49.7608 | 0.00 |
| FLOORSMAX\_MODE | Normalized information about building where the client lives | float | 49.7608 | 0.00 |
| FLOORSMIN\_AVG | Normalized information about building where the client lives | float | 67.8486 | 0.00 |
| FLOORSMIN\_MEDI | Normalized information about building where the client lives | float | 67.8486 | 0.00 |
| FLOORSMIN\_MODE | Normalized information about building where the client lives | float | 67.8486 | 0.00 |
| LANDAREA\_AVG | Normalized information about building where the client lives | float | 59.3767 | 0.00 |
| LANDAREA\_MEDI | Normalized information about building where the client lives | float | 59.3767 | 0.00 |
| LANDAREA\_MODE | Normalized information about building where the client lives | float | 59.3767 | 0.00 |
| LIVINGAPARTMENTS\_AVG | Normalized information about building where the client lives | float | 68.3550 | 0.00 |
| LIVINGAPARTMENTS\_MEDI | Normalized information about building where the client lives | float | 68.3550 | 0.00 |
| LIVINGAPARTMENTS\_MODE | Normalized information about building where the client lives | float | 68.3550 | 0.00 |
| LIVINGAREA\_AVG | Normalized information about building where the client lives | float | 50.1933 | 0.00 |
| LIVINGAREA\_MEDI | Normalized information about building where the client lives | float | 50.1933 | 0.00 |
| LIVINGAREA\_MODE | Normalized information about building where the client lives | float | 50.1933 | 0.00 |
| NONLIVINGAPARTMENTS\_AVG | Normalized information about building where the client lives | float | 69.4330 | 0.00 |
| NONLIVINGAPARTMENTS\_MEDI | Normalized information about building where the client lives | float | 69.4330 | 0.00 |
| NONLIVINGAPARTMENTS\_MODE | Normalized information about building where the client lives | float | 69.4330 | 0.00 |
| NONLIVINGAREA\_AVG | Normalized information about building where the client lives | float | 55.1792 | 0.00 |
| NONLIVINGAREA\_MEDI | Normalized information about building where the client lives | float | 55.1792 | 0.00 |
| NONLIVINGAREA\_MODE | Normalized information about building where the client lives | float | 55.1792 | 0.00 |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings with observable 30 DPD (days past due) default | float | 0.3320 | 0.00 |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings with observable 60 DPD (days past due) default | float | 0.3320 | 0.00 |
| OWN\_CAR\_AGE | Age of client's car | float | 65.9908 | 0.00 |
| REGION\_POPULATION\_RELATIVE | Normalized population of region where client lives (higher number means the client lives in more populated region) | float | 0.0000 | NA |
| TOTALAREA\_MODE | Normalized information about building where the client lives | float | 48.2685 | 0.00 |
| YEARS\_BEGINEXPLUATATION\_AVG | Normalized information about building where the client lives | float | 48.7810 | 0.00 |
| YEARS\_BEGINEXPLUATATION\_MEDI | Normalized information about building where the client lives | float | 48.7810 | 0.00 |
| YEARS\_BEGINEXPLUATATION\_MODE | Normalized information about building where the client lives | float | 48.7810 | 0.00 |
| YEARS\_BUILD\_AVG | Normalized information about building where the client lives | float | 66.4978 | 0.00 |
| YEARS\_BUILD\_MEDI | Normalized information about building where the client lives | float | 66.4978 | 0.00 |
| YEARS\_BUILD\_MODE | Normalized information about building where the client lives | float | 66.4978 | 0.00 |
| CNT\_CHILDREN | Number of children the client has | integer | 0.0000 | NA |
| DAYS\_BIRTH | Client's age in days at the time of application | integer | 0.0000 | NA |
| DAYS\_EMPLOYED | How many days before the application the person started current employment | integer | 0.0000 | median |
| DAYS\_ID\_PUBLISH | How many days before the application did client change the identity document with which he applied for the loan | integer | 0.0000 | NA |
| FLAG\_CONT\_MOBILE | Was mobile phone reachable (1=YES, 0=NO) | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_10 | Did client provide document 10 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_11 | Did client provide document 11 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_12 | Did client provide document 12 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_13 | Did client provide document 13 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_14 | Did client provide document 14 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_15 | Did client provide document 15 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_16 | Did client provide document 16 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_17 | Did client provide document 17 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_18 | Did client provide document 18 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_19 | Did client provide document 19 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_2 | Did client provide document 2 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_20 | Did client provide document 20 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_21 | Did client provide document 21 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_3 | Did client provide document 3 | integer | 0.0000 | NA |
| FLAG\_DOCUMENT\_4 | Did client provide document 4 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_5 | Did client provide document 5 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_6 | Did client provide document 6 | integer | 0.0000 | NA |
| FLAG\_DOCUMENT\_7 | Did client provide document 7 | integer | 0.0000 | delete col |
| FLAG\_DOCUMENT\_8 | Did client provide document 8 | integer | 0.0000 | NA |
| FLAG\_DOCUMENT\_9 | Did client provide document 9 | integer | 0.0000 | delete col |
| FLAG\_EMAIL | Did client provide email (1=YES, 0=NO) | integer | 0.0000 | delete col |
| FLAG\_EMP\_PHONE | Did client provide work phone (1=YES, 0=NO) | integer | 0.0000 | NA |
| FLAG\_MOBIL | Did client provide mobile phone (1=YES, 0=NO) | integer | 0.0000 | delete col |
| FLAG\_PHONE | Did client provide home phone (1=YES, 0=NO) | integer | 0.0000 | NA |
| FLAG\_WORK\_PHONE | Did client provide home phone (1=YES, 0=NO) | integer | 0.0000 | NA |
| HOUR\_APPR\_PROCESS\_START | Approximately at what hour did the client apply for the loan | integer | 0.0000 | NA |
| LIVE\_CITY\_NOT\_WORK\_CITY | Flag if client's contact address does not match work address (1=different, 0=same, at city level) | integer | 0.0000 | NA |
| LIVE\_REGION\_NOT\_WORK\_REGION | Flag if client's contact address does not match work address (1=different, 0=same, at region level) | integer | 0.0000 | delete col |
| REG\_CITY\_NOT\_LIVE\_CITY | Flag if client's permanent address does not match contact address (1=different, 0=same, at city level) | integer | 0.0000 | NA |
| REG\_CITY\_NOT\_WORK\_CITY | Flag if client's permanent address does not match work address (1=different, 0=same, at city level) | integer | 0.0000 | NA |
| REG\_REGION\_NOT\_LIVE\_REGION | Flag if client's permanent address does not match contact address (1=different, 0=same, at region level) | integer | 0.0000 | delete col |
| REG\_REGION\_NOT\_WORK\_REGION | Flag if client's permanent address does not match work address (1=different, 0=same, at region level) | integer | 0.0000 | delete col |
| REGION\_RATING\_CLIENT | Our rating of the region where client lives (1,2,3) | integer | 0.0000 | NA |
| REGION\_RATING\_CLIENT\_W\_CITY | Our rating of the region where client lives with taking city into account (1,2,3) | integer | 0.0000 | NA |
| CODE\_GENDER | Gender of the client | object | 0.0000 | NA |
| EMERGENCYSTATE\_MODE | Normalized information about building where the client lives | object | 47.3983 | new category |
| FLAG\_OWN\_CAR | Flag if the client owns a car | object | 0.0000 | NA |
| FLAG\_OWN\_REALTY | Flag if client owns a house or flat | object | 0.0000 | NA |
| FONDKAPREMONT\_MODE | Normalized information about building where the client lives | object | 68.3862 | new category |
| HOUSETYPE\_MODE | Normalized information about building where the client lives | object | 50.1761 | new category |
| NAME\_CONTRACT\_TYPE | Identification if loan is cash or revolving | object | 0.0000 | NA |
| NAME\_EDUCATION\_TYPE | Level of highest education the client achieved | object | 0.0000 | NA |
| NAME\_FAMILY\_STATUS | Family status of the client | object | 0.0000 | NA |
| NAME\_HOUSING\_TYPE | What is the housing situation of the client (renting, living with parents, ...) | object | 0.0000 | NA |
| NAME\_INCOME\_TYPE | Clients income type (businessman, working, maternity leave,…) | object | 0.0000 | NA |
| NAME\_TYPE\_SUITE | Who was accompanying client when he was applying for the loan | object | 0.4201 | NA |
| OCCUPATION\_TYPE | What kind of occupation does the client have | object | 31.3455 | new category |
| ORGANIZATION\_TYPE | Type of organization where client works | object | 0.0000 | NA |
| WALLSMATERIAL\_MODE | Normalized information about building where the client lives | object | 50.8408 | new category |
| WEEKDAY\_APPR\_PROCESS\_START | On which day of the week did the client apply for the loan | object | 0.0000 | delete col |

In general the missing values are replaced with '0' / median for numerical variables based on their appropriateness. Several features are removed (e.g., FLAG\_DOCUMENT\_2 - FLAG\_DOCUMENT\_21 ) which have no effect on the target variable as apparent from the frequency distribution. On average, more than 50% values are missing in features presenting normalized information about the building where the client lives. If those variable types are float, then the missing values are replaced with '0', assuming that applicant doesn't own a house. If those are categorical, then a new category with 'missing' type is created as is done for other categorical variables.

From frequency distribution and summary statistics anomaly is detected in ‘DAYS\_EMPLOYED’ feature in train and test data. The anomalies are replaced with median of the feature.

### 3.1.1 Correlation with target variable:

The features with strongest correlation with the target variables are listed in Table 2.

Table 2. Correlation coefficient of train dataset features with target variable.

|  |  |
| --- | --- |
| Variable | Correlation coefficient |
| EXT\_SOURCE\_3 | -0.1789 |
| EXT\_SOURCE\_2 | -0.1605 |
| EXT\_SOURCE\_1 | -0.1553 |
| NAME\_EDUCATION\_TYPE\_Higher education | -0.0566 |
| CODE\_GENDER\_F | -0.0547 |
| DAYS\_LAST\_PHONE\_CHANGE | 0.0552 |
| NAME\_INCOME\_TYPE\_Working | 0.0575 |
| REGION\_RATING\_CLIENT | 0.0589 |
| REGION\_RATING\_CLIENT\_W\_CITY | 0.0609 |
| DAYS\_BIRTH | 0.0782 |

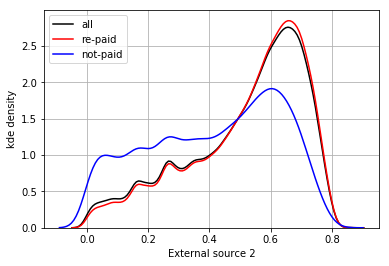
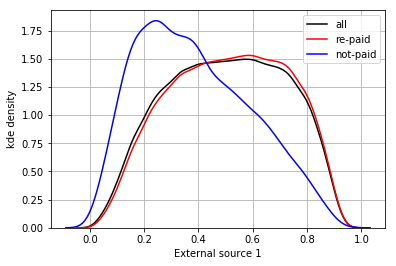


Figure 1. Kernel density estimate plot of External source 1, 2 for target = 0 (no difficulty in repayment), and target = 1 (difficulty in repayment).

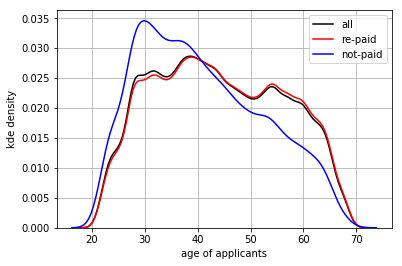
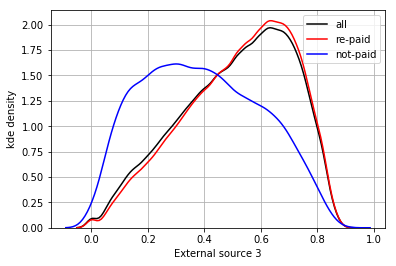


Figure 2. Kernel density estimate plot of External source 3 and applicant age for target = 0 (no difficulty in repayment), and target = 1 (difficulty in repayment).

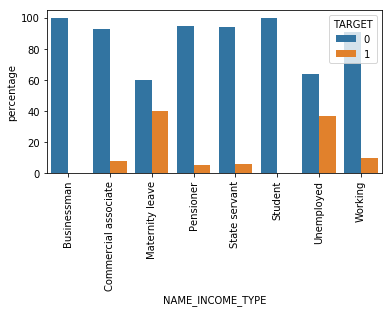
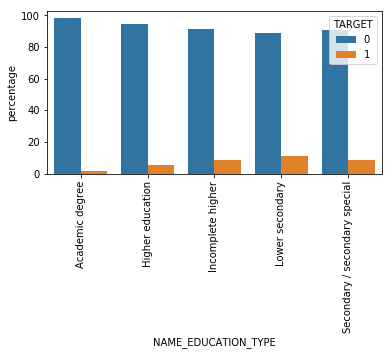
 

Figure 3. Bar plot for categorical features

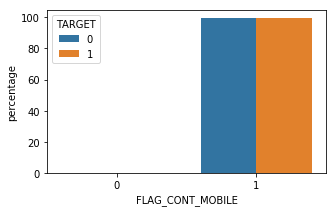
 

Figure 4. Deleted features with no effect on the target

## 3.2 Bureau

This table contains information regarding client’s previous credit from other financial institution. It has 17 variables, including 8 float, 6 integer and 3 categorical variables. The primary id of bureau is SK\_ID\_BUREAU and using the secondary key SK\_ID\_CURR we can merge this table with train/test dataset.

From frequency distribution, anomaly was detected in DAYS\_CREDIT\_ENDDATE feature, which was corrected by clipping the anomalies within 95 percentile as threshold.

A total of 7 variables have missing values as shown in the table below

Table 3. Bureau data summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Variable description | Data type | % missing | Replace with |
| AMT\_ANNUITY | Annuity of the Credit Bureau credit | float | 71.47 | 0.00 |
| AMT\_CREDIT\_MAX\_OVERDUE | Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample) | float | 65.51 | 0.00 |
| AMT\_CREDIT\_SUM | Current credit amount for the Credit Bureau credit | float | 0.00 | median |
| AMT\_CREDIT\_SUM\_DEBT | Current debt on Credit Bureau credit | float | 15.01 | median |
| AMT\_CREDIT\_SUM\_LIMIT | Current credit limit of credit card reported in Credit Bureau | float | 34.48 | median |
| AMT\_CREDIT\_SUM\_OVERDUE | Current amount overdue on Credit Bureau credit | float | 0.00 | NA |
| CNT\_CREDIT\_PROLONG | How many times was the Credit Bureau credit prolonged | integer | 0.00 | NA |
| CREDIT\_ACTIVE | Status of the Credit Bureau (CB) reported credits | object | 0.00 | NA |
| CREDIT\_CURRENCY | Recoded currency of the Credit Bureau credit | object | 0.00 | NA |
| CREDIT\_DAY\_OVERDUE | Number of days past due on CB credit at the time of application for related loan in our sample | integer | 0.00 | NA |
| CREDIT\_TYPE | Type of Credit Bureau credit (Car, cash,...) | object | 0.00 | NA |
| DAYS\_CREDIT | How many days before current application did client apply for Credit Bureau credit | integer | 0.00 | NA |
| DAYS\_CREDIT\_ENDDATE | Remaining duration of CB credit (in days) at the time of application in Home Credit | float | 6.15 | 0.00 |
| DAYS\_CREDIT\_UPDATE | How many days before loan application did last information about the Credit Bureau credit come | integer | 0.00 | NA |
| DAYS\_ENDDATE\_FACT | Days since CB credit ended at the time of application in Home Credit (only for closed credit) | float | 36.92 | 0.00 |

After the missing values are replaced, the features in bureau were grouped by SK\_ID\_CURR, which is unique to train/test dataset but not so in bureau suggesting people had multiple loan history against their name (SK\_ID\_CURR). New features created through aggregation are previous\_loan\_counts, and count, mean, max, min and sum on each numeric variables totaling 61. Aggregation applied on categorical variables are sum and mean which are count of that category and normalized count. Both aggregated numerical and categorical features are then merged with train and test dataset based on SK\_ID\_CURR. The correlation of this new features with target are as below.

Table 4. Correlation with target

|  |  |
| --- | --- |
| Variable | correlation coefficient |
| CREDIT\_ACTIVE\_Closed\_count\_norm | -0.0794 |
| CREDIT\_ACTIVE\_Active\_count\_norm | 0.0774 |
| CREDIT\_ACTIVE\_Active\_count | 0.0671 |
| bureau\_DAYS\_CREDIT\_mean | 0.0897 |
| bureau\_DAYS\_ENDDATE\_FACT\_mean | 0.0780 |
| bureau\_DAYS\_CREDIT\_min | 0.0752 |
| bureau\_DAYS\_CREDIT\_ENDDATE\_sum | 0.0703 |

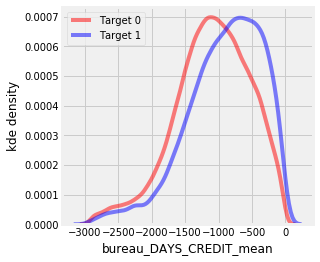
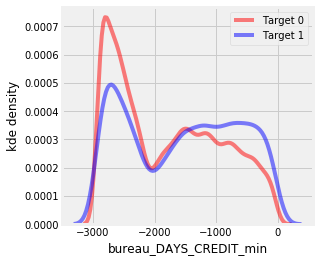
 

Figure 5. Kernel density function distribution of bureau\_DAYS\_CREDIT\_mean and bureau\_DAYS\_CREDIT\_min

## 3.3 Bureau balance

It has monthly data about the previous credits in bureau. This dataset has two variables, MONTHS\_BALANCE (integer) and STATUS (categorical). The categorical variable was aggregated using sum and mean and the numerical variable was aggregated using count, mean, max, min and sum functions as before. The new features were first merged with SK\_ID\_CURR and SK\_ID\_BUREAU from bureau using SK\_ID\_BUREAU and then with train/test dataset using SK\_ID\_CURR.

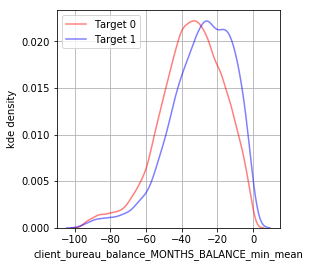
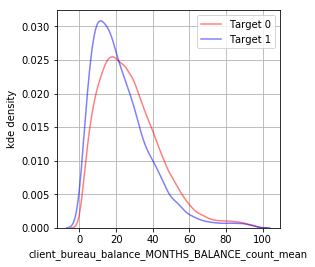
 

Figure 6. Kernel density function distribution of bureau\_balance\_MONTHS\_BALANCE

Table 5. Correlation of Bureau\_balance variables with target

|  |  |
| --- | --- |
| Bureau\_balance variable | Corrcoef |
| client\_bureau\_balance\_MONTHS\_BALANCE\_min\_mean | 0.089 |
| client\_bureau\_balance\_MONTHS\_BALANCE\_count\_mean | -0.0802 |
| client\_bureau\_balance\_MONTHS\_BALANCE\_mean\_mean | 0.0764 |
| client\_bureau\_balance\_MONTHS\_BALANCE\_min\_min | 0.0732 |
| client\_bureau\_balance\_MONTHS\_BALANCE\_sum\_mean | 0.0726 |
| client\_bureau\_balance\_MONTHS\_BALANCE\_count\_max | -0.0688 |
| client\_bureau\_balance\_MONTHS\_BALANCE\_sum\_min | 0.0681 |
| client\_bureau\_balance\_STATUS\_C\_count\_mean | -0.063 |
| client\_bureau\_balance\_STATUS\_1\_count\_norm\_mean | 0.0612 |
| client\_bureau\_balance\_STATUS\_1\_count\_norm\_max | 0.0611 |

## 3.4 Previous\_ application

This table contains previous applications for loans at Home Credit of clients who have loans in the application data Dataset. It has 37 variables, including 16 variables with missing values.

Table 6. Previous\_application data summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Variable description | Data type | % missing | Replacement |
| AMT\_ANNUITY | Annuity of previous application | float | 22.2867 | 0.00 |
| AMT\_APPLICATION | For how much credit did client ask on the previous application | float | 0 | NA |
| AMT\_CREDIT | Final credit amount on the previous application. This differs from AMT\_APPLICATION in a way that the AMT\_APPLICATION is the amount for which the client initially applied for, but during our approval process he could have received different amount - AMT\_CREDIT | Float | 0.0001 | 0.00 |
| AMT\_DOWN\_PAYMENT | Down payment on the previous application | Float | 53.6365 | 0.00 |
| AMT\_GOODS\_PRICE | Goods price of good that client asked for (if applicable) on the previous application | Float | 23.0818 | 0.00 |
| CHANNEL\_TYPE | Through which channel we acquired the client on the previous application | object | 0 | NA |
| CNT\_PAYMENT | Term of previous credit at application of the previous application | integer | 22.2864 | 0.00 |
| CODE\_REJECT\_REASON | Why was the previous application rejected | object | 0 | NA |
| DAYS\_DECISION | Relative to current application when was the decision about previous application made | integer | 0 | NA |
| DAYS\_FIRST\_DRAWING | Relative to application date of current application when was the first disbursement of the previous application | Integer | 40.298129/ anomaly | delete |
| DAYS\_FIRST\_DUE | Relative to application date of current application when was the first due supposed to be of the previous application | integer | 40.298129/ anomaly | 0 /max |
| DAYS\_LAST\_DUE | Relative to application date of current application when was the last due date of the previous application | integer | 40.298129/ anomaly | 0 /max |
| DAYS\_LAST\_DUE\_1ST\_VERSION | Relative to application date of current application when was the first due of the previous application | integer | 40.298129/ anomaly | 0 /max |
| DAYS\_TERMINATION | Relative to application date of current application when was the expected termination of the previous application | integer | 40.298129/ anomaly | 0 /max |
| FLAG\_LAST\_APPL\_PER\_CONTRACT | Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract | object | 0 | NA |
| HOUR\_APPR\_PROCESS\_START | Approximately at what day hour did the client apply for the previous application | integer | 0 | NA |
| NAME\_CASH\_LOAN\_PURPOSE | Purpose of the cash loan | object | 0 | NA |
| NAME\_CLIENT\_TYPE | Was the client old or new client when applying for the previous application | object | 0 | NA |
| NAME\_CONTRACT\_STATUS | Contract status (approved, cancelled, ...) of previous application | object | 0 | NA |
| NAME\_CONTRACT\_TYPE | Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application | object | 0 | NA |
| NAME\_GOODS\_CATEGORY | What kind of goods did the client apply for in the previous application | object | 0 | NA |
| NAME\_PAYMENT\_TYPE | Payment method that client chose to pay for the previous application | object | 0 | NA |
| NAME\_PORTFOLIO | Was the previous application for CASH, POS, CAR, … | object | 0 | NA |
| NAME\_PRODUCT\_TYPE | Was the previous application x-sell o walk-in | object | 0 | NA |
| NAME\_SELLER\_INDUSTRY | The industry of the seller | object | 0 | NA |
| NAME\_TYPE\_SUITE | Who accompanied client when applying for the previous application | object | 49.1198 | Unaccompanied |
| NAME\_YIELD\_GROUP | Grouped interest rate into small medium and high of the previous application | object | 0 | NA |
| NFLAG\_INSURED\_ON\_APPROVAL | Did the client requested insurance during the previous application | integer | 40.2981 | new value |
| NFLAG\_LAST\_APPL\_IN\_DAY | Flag if the application was the last application per day of the client. Sometimes clients apply for more applications a day. Rarely it could also be error in our system that one application is in the database twice | integer | 0 | NA |
| PRODUCT\_COMBINATION | Detailed product combination of the previous application | object | 0.0207 | new category |
| RATE\_DOWN\_PAYMENT | Down payment rate normalized on previous credit | float | 53.6365 | 0.00 |
| RATE\_INTEREST\_PRIMARY | Interest rate normalized on previous credit | float | 99.6437 | delete |
| RATE\_INTEREST\_PRIVILEGED | Interest rate normalized on previous credit | float | 99.6437 | delete |
| SELLERPLACE\_AREA | Selling area of seller place of the previous application | integer | 0 | NA |
| WEEKDAY\_APPR\_PROCESS\_START | On which day of the week did the client apply for previous application | object | 0 | delete |

Since more than 99% values are missing in RATE\_INTEREST\_PRIVILEGED and RATE\_INTEREST\_PRIMARY, those features are removed. Also WEEKDAY\_APPR\_PROCESS\_START is removed as it does not affect the outcome. Several features, e.g., DAYS\_TERMINATION, has missing values as well as anomalies. The missing values are replaced with ‘0’, while the anomalies are clipped by a maximum threshold. As before, the categorical and numerical features were aggregated separately on SK\_ID\_CURR and then merged with train/test dataset.

### 3.4.1 Correlation with target

The variables having higher correlation with target are shown in Table.

Table 7. Correlation with target variable

|  |  |
| --- | --- |
| Variable | Correlation coefficient |
| NFLAG\_INSURED\_ON\_APPROVAL\_sum | -0.0579 |
| DAYS\_LAST\_DUE\_1ST\_VERSION\_sum | 0.0568 |
| DAYS\_FIRST\_DUE\_mean | 0.0556 |
| DAYS\_DECISION\_min | 0.0534 |
| DAYS\_LAST\_DUE\_1ST\_VERSION\_min | 0.0533 |
| DAYS\_FIRST\_DUE\_min | 0.0532 |

## 3.5 POS\_CASH\_balance

This table contains monthly data about previous point of sale or cash loans clients had with Home Credit. This dataset has 8 variables. There are 2 features with missing values. The missing values are replaced with ‘0’. As usual, the features are aggregated based on datatype and merged with train/test dataset.

Table 8. Pos\_Cash\_balance data summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Variable description | Data  type | % missing  value | Replace  with |
| MONTHS\_BALANCE | Month of balance relative to application date (-1 means the information to the freshest monthly snapshot, 0 means the information at application - often it will be the same as -1 as many banks are not updating the information to Credit Bureau regularly ) | integer | 0 | NA |
| CNT\_INSTALMENT | Term of previous credit (can change over time) | integer | 0.26 | 0 |
| CNT\_INSTALMENT\_FUTURE | Installments left to pay on the previous credit | integer | 0.26 | 0 |
| NAME\_CONTRACT\_STATUS | Contract status during the month | object | 0 | NA |
| SK\_DPD | DPD (days past due) during the month of previous credit | integer | 0 | NA |
| SK\_DPD\_DEF | DPD during the month with tolerance (debts with low loan amounts are ignored) of the previous credit | integer | 0 | NA |

3.6 Credit\_card\_balance

This table has monthly data about previous credit cards clients had with Home Credit. It has a total of 23 variables, among which number of variables with missing values are 9 As can be seen in the table.

Table 9. Credit card balance data summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Variable description | Data  type | % missing | Replace with |
| AMT\_BALANCE | Balance during the month of previous credit | float | 0 | NA |
| AMT\_CREDIT\_LIMIT\_ACTUAL | Credit card limit during the month of the previous credit | float | 0 | NA |
| AMT\_DRAWINGS\_ATM\_CURRENT | Amount drawing at ATM during the month of the previous credit | float | 19.5249 | 0.00 |
| AMT\_DRAWINGS\_CURRENT | Amount drawing during the month of the previous credit | Float | 0 | NA |
| AMT\_DRAWINGS\_OTHER\_CURRENT | Amount of other drawings during the month of the previous credit | float | 19.5249 | 0.00 |
| AMT\_DRAWINGS\_POS\_CURRENT | Amount drawing or buying goods during the month of the previous credit | float | 19.5249 | 0.00 |
| AMT\_INST\_MIN\_REGULARITY | Minimal installment for this month of the previous credit | float | 7.9482 | 0.00 |
| AMT\_PAYMENT\_CURRENT | How much did the client pay during the month on the previous credit | float | 19.9981 | 0.00 |
| AMT\_PAYMENT\_TOTAL\_CURRENT | How much did the client pay during the month in total on the previous credit | float | 0 | NA |
| AMT\_RECEIVABLE\_PRINCIPAL | Amount receivable for principal on the previous credit | float | 0 | NA |
| AMT\_RECIVABLE | Amount receivable on the previous credit | float | 0 | NA |
| AMT\_TOTAL\_RECEIVABLE | Total amount receivable on the previous credit | float | 0 | NA |
| CNT\_DRAWINGS\_ATM\_CURRENT | Number of drawings at ATM during this month on the previous credit | integer | 19.5249 | 0.00 |
| CNT\_DRAWINGS\_CURRENT | Number of drawings during this month on the previous credit | integer | 0 | NA |
| CNT\_DRAWINGS\_OTHER\_CURRENT | Number of other drawings during this month on the previous credit | integer | 19.5249 | 0.00 |
| CNT\_DRAWINGS\_POS\_CURRENT | Number of drawings for goods during this month on the previous credit | integer | 19.5249 | 0.00 |
| CNT\_INSTALMENT\_MATURE\_CUM | Number of paid installments on the previous credit | integer | 7.9482 | 0.00 |
| MONTHS\_BALANCE | Month of balance relative to application date (-1 means the freshest balance date) | integer | 0 | NA |
| NAME\_CONTRACT\_STATUS | Contract status (active signed,...) on the previous credit | object | 0 | NA |
| SK\_DPD | DPD (Days past due) during the month on the previous credit | integer | 0 | NA |
| SK\_DPD\_DEF | DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored) of the previous credit | integer | 0 | NA |

## 3.7 Installments\_payments

This table contains payment history for previous loans at Home Credit. It has 8 variables, among which number of variables with missing variables are 2 as shown in the Table.

Table 10. Installments\_payments data summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Variable description | Data type | % missing value | Replace with |
| NUM\_INSTALMENT\_VERSION | Version of installment calendar (0 is for credit card) of previous credit. Change of installment version from month to month signifies that some parameter of payment calendar has changed | integer | 0.00 | NA |
| NUM\_INSTALMENT\_NUMBER | On which installment we observe payment | integer | 0.00 | NA |
| DAYS\_INSTALMENT | When the installment of previous credit was supposed to be paid (relative to application date of current loan) | integer | 0.00 | NA |
| DAYS\_ENTRY\_PAYMENT | When was the installments of previous credit paid actually (relative to application date of current loan) | integer | 0.2135 | 0.00 |
| AMT\_INSTALMENT | What was the prescribed installment amount of previous credit on this installment | float | 0.00 | NA |
| AMT\_PAYMENT | What the client actually paid on previous credit on this installment | float | 0.2135 | 0.00 |

Following the steps stated earlier, the categorical and numerical features were aggregated and merged with train/test data.

## 3.8. New features from feature engineering

A total of 4 new features are created from the given features, e.g., CREDIT\_INCOME\_PERCENT, ANNUITY\_INCOME\_PERCENT, CREDIT\_TERM and DAYS\_EMPLOYED\_PERCENT. Feature CREDIT\_INCOME\_PERCENT is the ratio of AMT\_CREDIT and AMT\_INCOME\_TOTAL (loan to income ratio). ANNUITY\_INCOME\_PERCENT is estimated from the ratio of AMT\_ANNUITY and AMT\_INCOME\_TOTAL. The CREDIT\_TERM is created from ratio of AMT\_ANNUITY and AMT\_CREDIT and DAYS\_EMPLOYED\_PERCENT is created as a ratio between DAYS\_EMPLOYED and DAYS\_BIRTH (employment period to age ratio). Also features with stronger correlation with the target variables (i.e., EXT\_SOURCE\_1, EXT\_SOURCE\_2, EXT\_SOURCE\_3, DAYS\_BIRTH) were combined to create new polynomial features using PolynomialFeatures function with degree 3. These new features are then merged with train/test dataset. The final number of the variables in train dataset is 1422.

# Model prediction

The performance metric is a classification metric called receiver operating characteristic area under the curve or ROC AUC. ROC is crucial in deciding the best threshold value. ROC is created by plotting True positive rate (True positive/ (True positive + False negative)) against False positive rate (False positive/ (False positive + True negative)) from the confusion matrix. While True positive rate implies proportion of people correctly classified, False positive rate indicates misclassified proportion. Therefore, if the points on the ROC curve are on a diagonal line o 0 to 1 scale it indicates the proportion of correct classification is equal to the proportion of wrongly classified. If a model predict randomly the area under the curve would be 0.5. The higher the area is within the scale of 0 to 1, better the threshold works for correctly classifying the samples and, more efficient the model is.

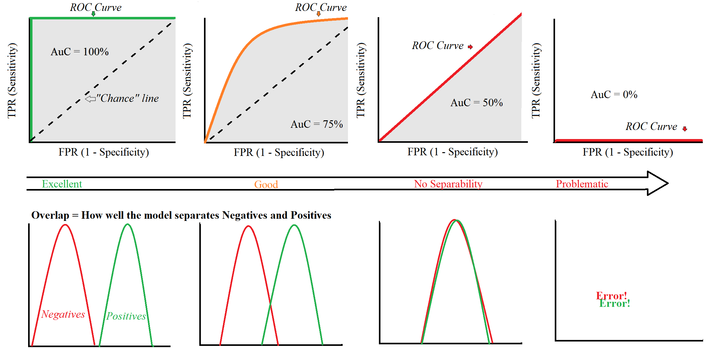


Figure 7. Associating classification performance with ROC curve.

## 4.1 Logistic regression

Initially the model prediction was done using the features from original application\_train and application\_test dataset. The number of variables excluding target were 121 and 240, before and after transforming the categorical variables, respectively. The first model implemented is logistic regression, simplest among the ML classifier models. From the training dataset 30% was kept aside for validation purpose. The model was thus trained on 70% dataset and then performance tested on the validation dataset. The ROC AUC score on the validation dataset using logistic regression was only 0.501, suggesting this is a slightly better model than random classification.

## 4.2 Random Forest Classifier

Next Random Forest Classifier was used with n\_estimator as 100, and random\_state as 50. As before, 30% of the training dataset was kept aside for validation. The ROC AUC score this time came up as 0.706 which is a significant improvement from the previous one. The most significant 3 features in this model are EXT\_SOURCE\_2, EXT\_SOURCE\_3, and DAYS\_BIRTH.

## 4.3 LightGBM

Following I used LightGBM algorithm, which is a gradient boosting framework using tree based learning algorithms. Here I used 5 fold cross validation for performance validation. The n\_estimators was set as 10,000, the evaluation metric was set as ‘auc’, the early stopping round was set as 100 in case the score doesn’t improve, and the iteration with the best performance was recorded for using on the test dataset. The average estimated AUC ROC score from 5 fold validation was 0.758 suggesting this as even better model for classification. The most important features in this model are, EXT\_SOURCE 1, 3 and 2.

Next I included 39 engineered featured stated earlier in the train/test dataset. The AUC ROC score immediately improved from 0.758 to 0.766. The 3 most significant features here are CREDIT\_TERM (engineered feature), DAYS\_ID\_PUBLISH and AMT\_ANNUITY.

Finally, I employed all the features combined together from all 8 tables along with the engineered features and re-run the model. This time the highest AUC ROC score is 0.788. The most significant features are CREDIT\_TERM, AMT\_ANNUITY, and client\_installments\_AMT\_PAYMENT\_min\_sum.

Table 11. ML model performance sumamry

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML model | Dataset used | # of features | Parameters | ROC AUC score | Significant features |
| Logistic regression | application\_train/test  w/ 30% validation | 240 | c =  0.0001 | 0.501 |  |
| Random forest | application\_train/test  w/ 30% validation | 240 | n\_estimator =  100 | 0.706 | EXT\_SOURCE\_2, EXT\_SOURCE\_3, DAYS\_BIRTH |
| LightGBM | application\_train/test  w/ 5-fold CV | 240 | n\_estimator = 10000 | 0.758 | EXT\_SOURCE\_1, EXT\_SOURCE\_3, EXT\_SOURCE\_2 |
| application\_train/test  w/ 5-fold CV | 240+39 engineered features | n\_estimator = 10000 | 0.766 | CREDIT\_TERM (engineered feature), DAYS\_ID\_PUBLISH, AMT\_ANNUITY |
| all 8 dataset  w/ 5-fold CV | 1422 | n\_estimator = 10000 | 0.788 | CREDIT\_TERM, AMT\_ANNUITY, client\_installments\_AMT\_PAYMENT\_min\_sum |

# Conclusion

Overall, appropriate data cleaning, replacing missing values and anomalies, inclusion of relevant features, feature engineering and employing better classifier have improved the classification significantly. The target metric was also achieved.