Approve/Reject Loan Applications based on applicant records

Authors: Nguyễn Thị Hoài Linh

ID:11192942

**1. INTRODUCTION**

It is difficult for banks to lend to people because their credit history lacks a lot of data or they do not have data that exists in the past. Therefore, many consumers take advantage of this to become eligible borrowers.

When banks receive a request for a loan from a customer, the bank will have to rely on the customer's loan record to decide whether to lend this object or not. This can pose some risk if a valid lender is misidentified. In the first case, the customer can repay the loan, but the bank does not accept the loan, which causes the bank's business to lose money and lose potential customers. In the second case, the customer is not able to repay the loan, but the bank approves the loan, which can pose a big risk of financial loss.

This case study is intended to help banks have an appropriate model to identify eligible customers who are able to repay loans. Eliminate and minimize credit risk due to customers' high probability of default causing a lot of losses to the bank.

**2. THEORETICAL BACKGROUND**

In this study, I use 3 commonly used methods in machine learning, namely Logistic Regression, Decision Tree, and Random Forest to build a predictive model for 'TARGET'. In which the Random Forest model with max\_depth = 25 is the most effective with macro F1 score, Accuracy,weighted-averaged F1 score =0.98 and roc\_auc\_score =0.99

Logistic regression analysis is a statistical technique to examine the relationship between an independent variable (variable or categorical) and a dependent variable which is a binary variable (the dependent variable 'y' has only two states: 1 and 0).

A decision tree is a structured hierarchical tree used to classify objects based on sequences of rules. The properties of the object can be of different data types such as Binary, Identical (Nominal), Ordinal (Ordinal), Quantitative while the subclass property must have a data type. be Binary or Ordinal. It can be understood simply that given data about objects including attributes and their classes, the decision tree will generate rules to predict the class of the unknown data.

Random Forests is a supervised learning algorithm. It can be used for both classification and regression. It is also the most flexible and easy-to-use algorithm. Random forests generate a decision tree on randomly selected data samples, are predicted from each tree and choose the best solution by voting. It also provides a pretty good indicator of feature importance

**3. DATA**

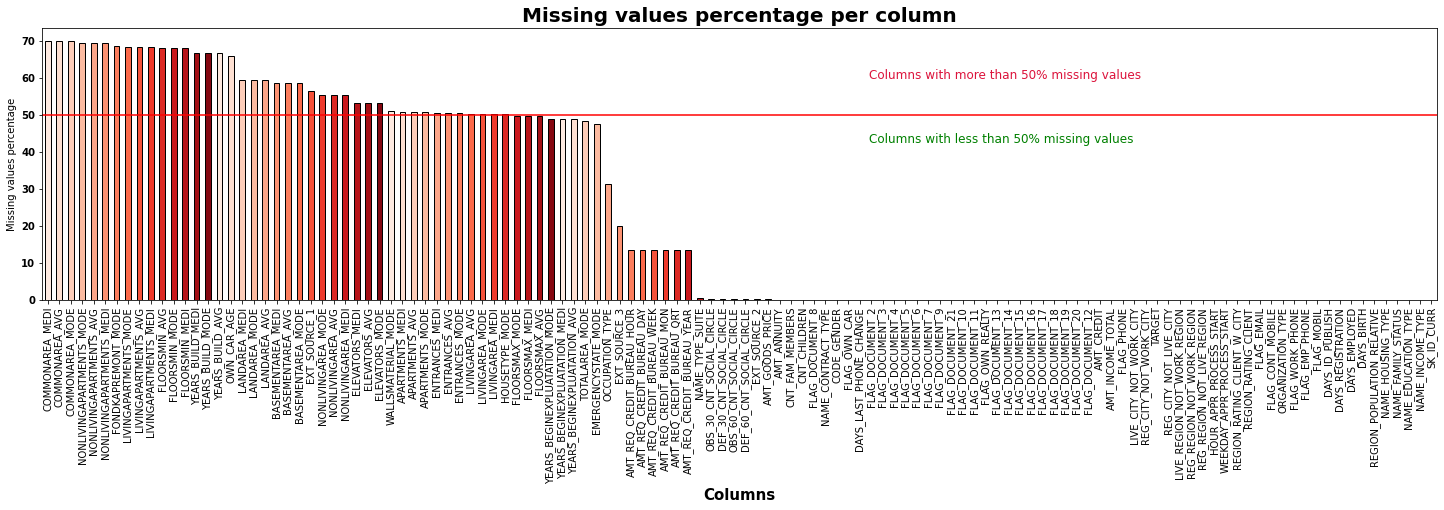
**a) Data :** <https://www.kaggle.com/datasets/arkapravasen/bank-loan-default>

(The data has 307511 rows and 122 columns however I am only using 2/3 of the rows from row 100000 to row 300000)

**b) Columns description:**

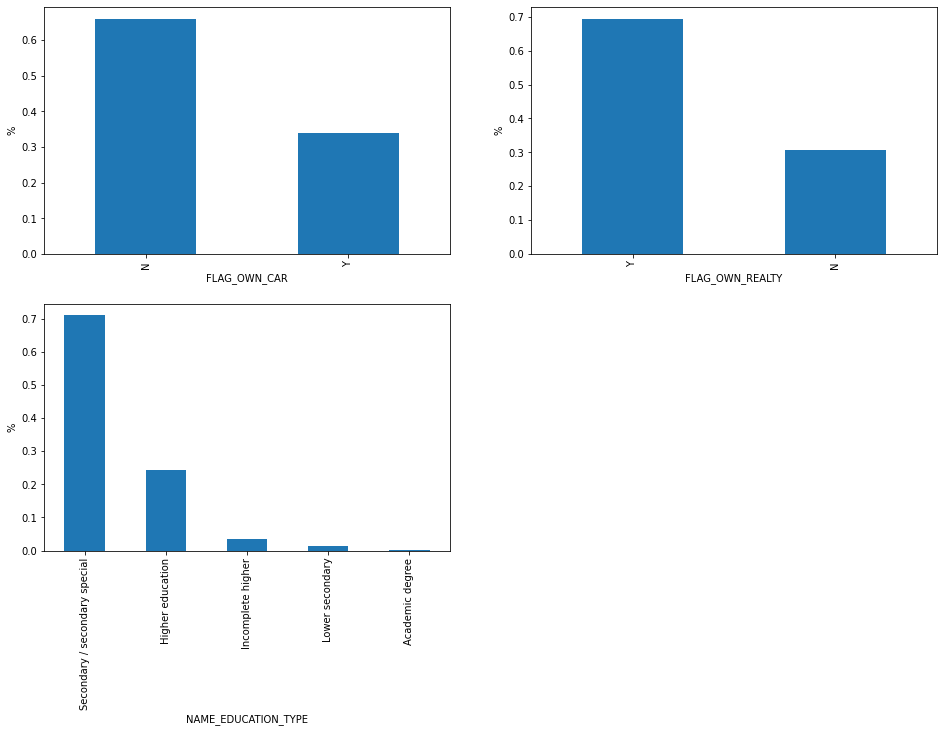
|  |  |
| --- | --- |
| **Row** | **Description** |
| SK\_ID\_CURR | ID of loan in our sample |
| TARGET | Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases) |
| NAME\_CONTRACT\_TYPE | Identification if loan is cash or revolving |
| CODE\_GENDER | Gender of the client |
| FLAG\_OWN\_CAR | Flag if the client owns a car |
| FLAG\_OWN\_REALTY | Flag if client owns a house or flat |
| CNT\_CHILDREN | Number of children the client has |
| AMT\_INCOME\_TOTAL | Income of the client |
| AMT\_CREDIT | Credit amount of the loan |
| AMT\_ANNUITY | Loan annuity |
| AMT\_GOODS\_PRICE | For consumer loans it is the price of the goods for which the loan is given |
| NAME\_TYPE\_SUITE | Who was accompanying client when he was applying for the loan |
| NAME\_INCOME\_TYPE | Clients income type (businessman, working, maternity leave,…) |
| NAME\_EDUCATION\_TYPE | Level of highest education the client achieved |
| NAME\_FAMILY\_STATUS | Family status of the client |
| NAME\_HOUSING\_TYPE | What is the housing situation of the client (renting, living with parents, ...) |
| REGION\_POPULATION\_RELATIVE | Normalized population of region where client lives (higher number means the client lives in more populated region) |
| DAYS\_BIRTH | Client's age in days at the time of application |
| DAYS\_EMPLOYED | How many days before the application the person started current employment |
| DAYS\_REGISTRATION | How many days before the application did client change his registration |
| DAYS\_ID\_PUBLISH | How many days before the application did client change the identity document with which he applied for the loan |
| OWN\_CAR\_AGE | Age of client's car |
| FLAG\_MOBIL | Did client provide mobile phone (1=YES, 0=NO) |
| FLAG\_EMP\_PHONE | Did client provide work phone (1=YES, 0=NO) |
| FLAG\_WORK\_PHONE | Did client provide home phone (1=YES, 0=NO) |
| FLAG\_CONT\_MOBILE | Was mobile phone reachable (1=YES, 0=NO) |
| FLAG\_PHONE | Did client provide home phone (1=YES, 0=NO) |
| FLAG\_EMAIL | Did client provide email (1=YES, 0=NO) |
| OCCUPATION\_TYPE | What kind of occupation does the client have |
| CNT\_FAM\_MEMBERS | How many family members does client have |
| REGION\_RATING\_CLIENT | Our rating of the region where client lives (1,2,3) |
| REGION\_RATING\_CLIENT\_W\_CITY | Our rating of the region where client lives with taking city into account (1,2,3) |
| WEEKDAY\_APPR\_PROCESS\_START | On which day of the week did the client apply for the loan |
| HOUR\_APPR\_PROCESS\_START | Approximately at what hour did the client apply for the loan |
| REG\_REGION\_NOT\_LIVE\_REGION | Flag if client's permanent address does not match contact address (1=different, 0=same, at region level) |
| REG\_REGION\_NOT\_WORK\_REGION | Flag if client's permanent address does not match work address (1=different, 0=same, at region level) |
| LIVE\_REGION\_NOT\_WORK\_REGION | Flag if client's contact address does not match work address (1=different, 0=same, at region level) |
| REG\_CITY\_NOT\_LIVE\_CITY | Flag if client's permanent address does not match contact address (1=different, 0=same, at city level) |
| REG\_CITY\_NOT\_WORK\_CITY | Flag if client's permanent address does not match work address (1=different, 0=same, at city level) |
| LIVE\_CITY\_NOT\_WORK\_CITY | Flag if client's contact address does not match work address (1=different, 0=same, at city level) |
| ORGANIZATION\_TYPE | Type of organization where client works |
| EXT\_SOURCE\_1 | Normalized score from external data source |
| EXT\_SOURCE\_2 | Normalized score from external data source |
| EXT\_SOURCE\_3 | Normalized score from external data source |
| APARTMENTS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| BASEMENTAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BEGINEXPLUATATION\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BUILD\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| COMMONAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ELEVATORS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ENTRANCES\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMAX\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMIN\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LANDAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAPARTMENTS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAPARTMENTS\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAREA\_AVG | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| APARTMENTS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| BASEMENTAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BEGINEXPLUATATION\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BUILD\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| COMMONAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ELEVATORS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ENTRANCES\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMAX\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMIN\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LANDAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAPARTMENTS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAPARTMENTS\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| APARTMENTS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| BASEMENTAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BEGINEXPLUATATION\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| YEARS\_BUILD\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| COMMONAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ELEVATORS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| ENTRANCES\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMAX\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FLOORSMIN\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LANDAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAPARTMENTS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| LIVINGAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAPARTMENTS\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| NONLIVINGAREA\_MEDI | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| FONDKAPREMONT\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| HOUSETYPE\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| TOTALAREA\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| WALLSMATERIAL\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| EMERGENCYSTATE\_MODE | Normalized information about building where the client lives, What is average (\_AVG suffix), modus (\_MODE suffix), median (\_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings with observable 30 DPD (days past due) default |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings defaulted on 30 DPD (days past due) |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings with observable 60 DPD (days past due) default |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | How many observation of client's social surroundings defaulted on 60 (days past due) DPD |
| DAYS\_LAST\_PHONE\_CHANGE | How many days before application did client change phone |
| FLAG\_DOCUMENT\_2 | Did client provide document 2 |
| FLAG\_DOCUMENT\_3 | Did client provide document 3 |
| FLAG\_DOCUMENT\_4 | Did client provide document 4 |
| FLAG\_DOCUMENT\_5 | Did client provide document 5 |
| FLAG\_DOCUMENT\_6 | Did client provide document 6 |
| FLAG\_DOCUMENT\_7 | Did client provide document 7 |
| FLAG\_DOCUMENT\_8 | Did client provide document 8 |
| FLAG\_DOCUMENT\_9 | Did client provide document 9 |
| FLAG\_DOCUMENT\_10 | Did client provide document 10 |
| FLAG\_DOCUMENT\_11 | Did client provide document 11 |
| FLAG\_DOCUMENT\_12 | Did client provide document 12 |
| FLAG\_DOCUMENT\_13 | Did client provide document 13 |
| FLAG\_DOCUMENT\_14 | Did client provide document 14 |
| FLAG\_DOCUMENT\_15 | Did client provide document 15 |
| FLAG\_DOCUMENT\_16 | Did client provide document 16 |
| FLAG\_DOCUMENT\_17 | Did client provide document 17 |
| FLAG\_DOCUMENT\_18 | Did client provide document 18 |
| FLAG\_DOCUMENT\_19 | Did client provide document 19 |
| FLAG\_DOCUMENT\_20 | Did client provide document 20 |
| FLAG\_DOCUMENT\_21 | Did client provide document 21 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | Number of enquiries to Credit Bureau about the client one hour before application |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application) |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application) |

**c)** **Data preprocessing:**

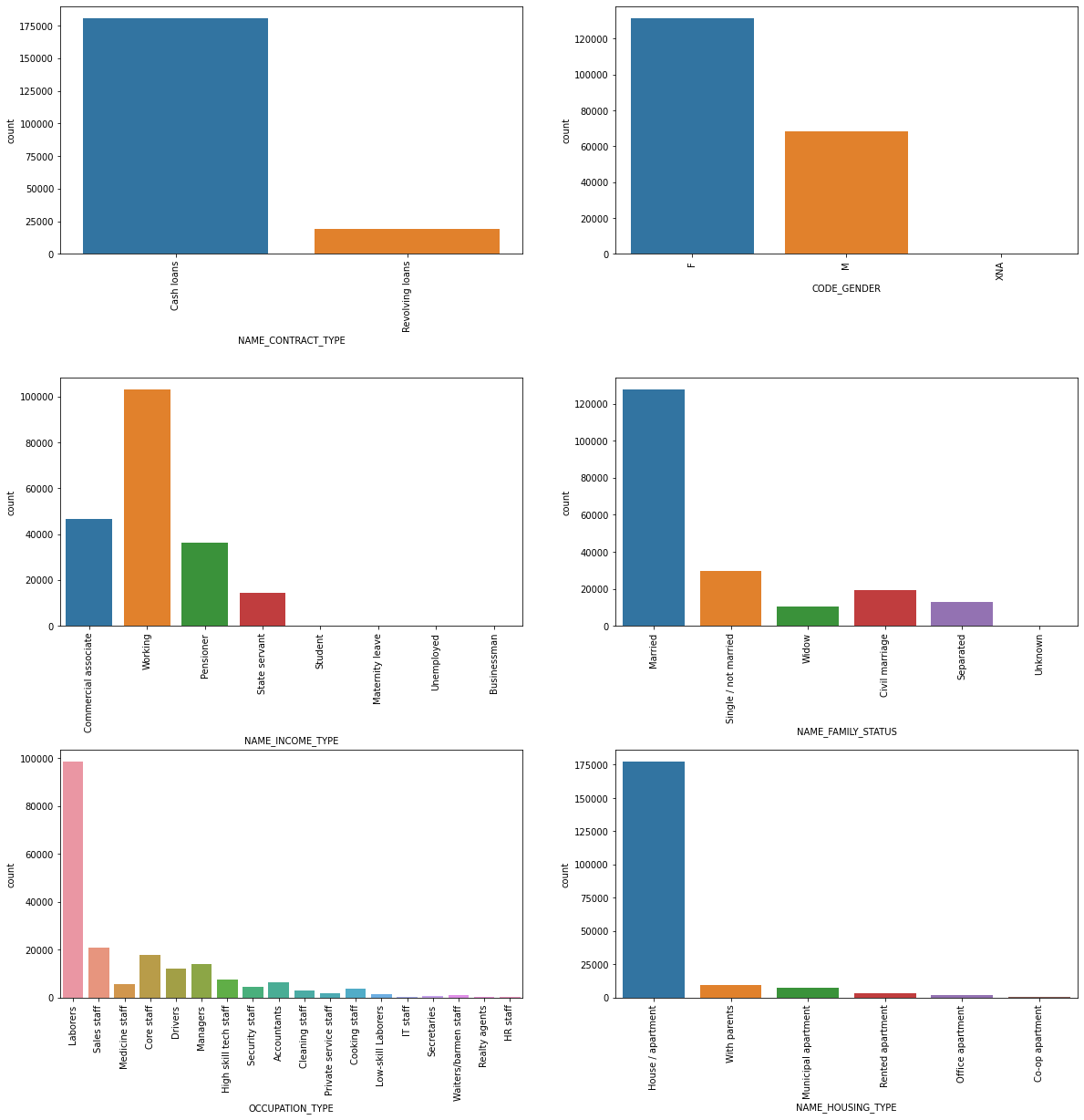
**-Missing value:** About half of the features contain missing values. 41 features containing more than 50% of data have missing values so these columns will be removed. 

+The remaining columns are divided into two categories, "cate" containing categorical data and "num" containing numerical data.

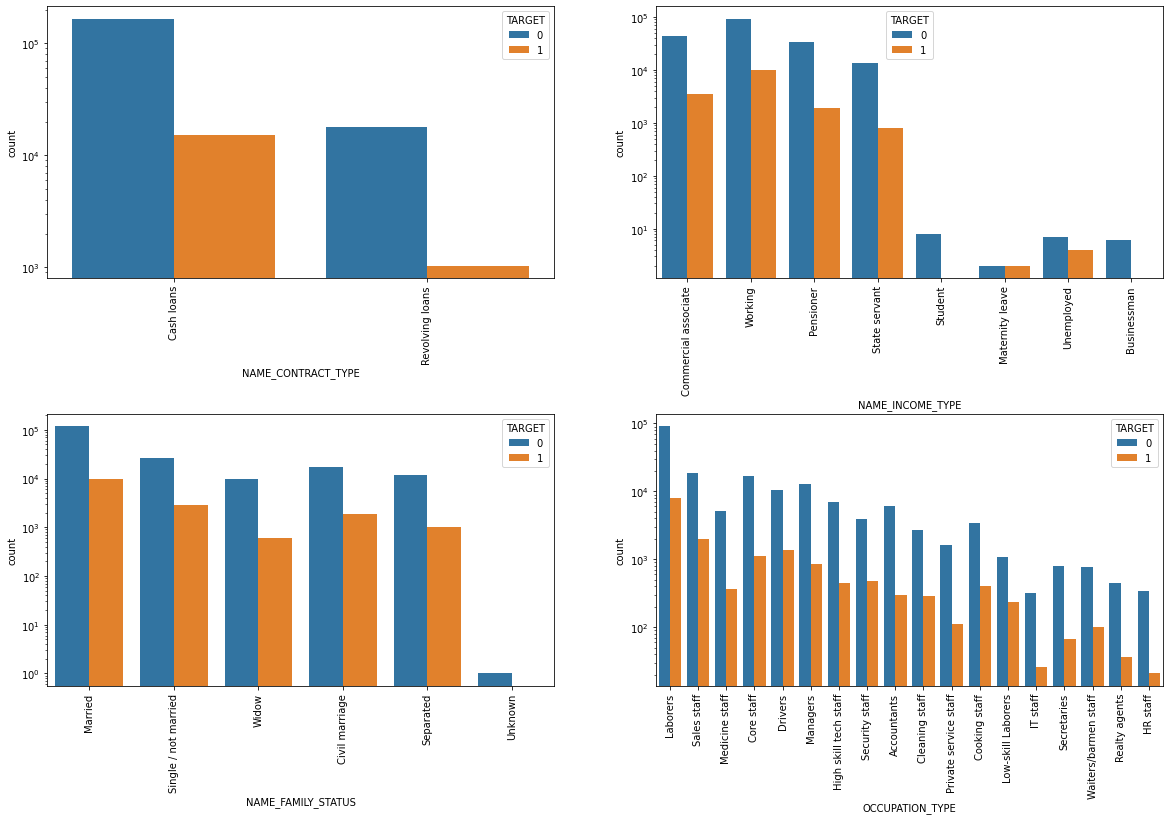
+ The missing value in "cate" will be replaced by mode while the missing value in "num" will be replaced by the mean value

**- Data Analysis: **

More than 60% of customers do not own a car. Nearly 70% of customers do not own real estate. About 70% of customers have secondary/ secondary special, more than 20% belong to higher education group. Very few customers belong to the remaining 3 educational groups.

****

The number of people in the cash loan group is much larger than the Revolving loan group. The number of male customers is only half that of female customers. Most customers have working income type and double that of customers with commercial associate income type. The majority of loans come from laborers, married and have a house/apartment

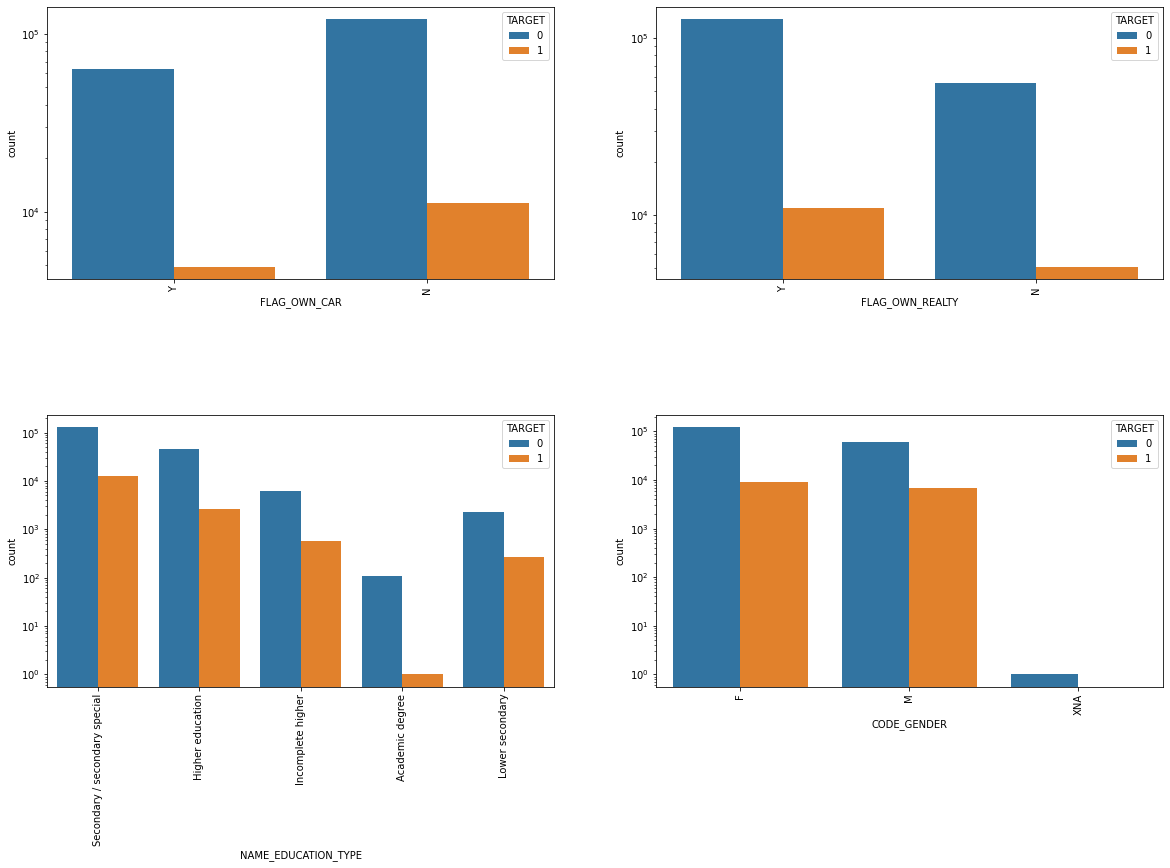
****

NAME\_CONTRACT\_TYPE: Revolving loan type has less number of defaulters.

NAME\_INCOME\_TYPE: Business man and students groups have higher chances of repaying the loans. Chances that an unemployed become a defaulter is more

NAME\_FAMILY\_STATUS: Widow category has less chances of becoming defaulter whereas married has high chances. Civil marriage category have less defaulters compared to the single

OCCUPATION\_TYPE: IT staff, HR staff and Realty agents have less defaulters whereas sales staff, laborers and drivers have more defaulters

****

Customers with cars are less likely to default, but the number of customers in this group is very small. Customers who do not yet own a home or apartment have lower default rates than those who do. Academic degree group is less likely to default (this group also has very few customers)

**- Encoding data:** Categorical data is a type of data that accepts only a finite number of fixed values, usually represented as text, and is divided into two types nominal and ordinal data. The use of categorical data in machine learning problems has some difficulties such as many machine learning models usually only accept numerical values as input to use these models, categorical data must be converted to numbers; Computers do not perceive data as categorical and the relationships between them the way humans perceive; data can include a very large number of different values, where each value occurs only very few times. Therefore, we must find a way to convert these categories into numerical form so that the computer can process it.

+The columns contain only 2 types of values like NAME\_CONTRACT\_TYPE (Cash loans', 'Revolving loans') , CODE\_GENDER ('M', 'F'), FLAG\_OWN\_CAR and FLAG\_OWN\_REALTY ("Y","N' ") will be replaced by two values 0 and 1

+ Column WEEKDAY\_APPR\_PROCESS\_START contains data about the days of the week to be replaced by corresponding numbers

+ The remaining columns will be handled by Dummy Encoding (code in modeling).

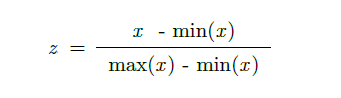
**-Correlation:** A model is stable when the variance must be low. If the variance of the weights is high, the model will be very sensitive to the data. It means that the model may not perform well with the test data. When the success is high, it makes the direction high. Therefore, it is necessary to remove features with high correlation. In addition, this also helps the algorithm run faster. I decided to remove the features with correlation > 0.**8**

**- Imbalanced data:**

TARGET column in dataset is unbalanced with 91.9285% data 0 and 8.0715 is 1. The amount of data is disparate between group 0 and group 1 by a very large amount. This greatly affects the prediction model, when such a large imbalance will make the forecasting model less accurate for the minority group because most of the forecast results are often biased towards the group with more data. Therefore, I used oversampling technique to bring data back to balance.

**- Scaling:**

The data consists of many features (columns), and each feature has different units and magnitudes. This affects the efficiency of many algorithms. Therefore, it is common to adjust the data so that the characteristics have the same data scaling. And usually leave the properties in the range [0, 1]



To be able to normalize data using scikit-learn library with MinMaxScaler, the MinMaxScaler scaler will return the variables to the [0, 1] value domain.

**4. FINDINGS AND DISCUSSION**

**Dividing the data into two parts :**

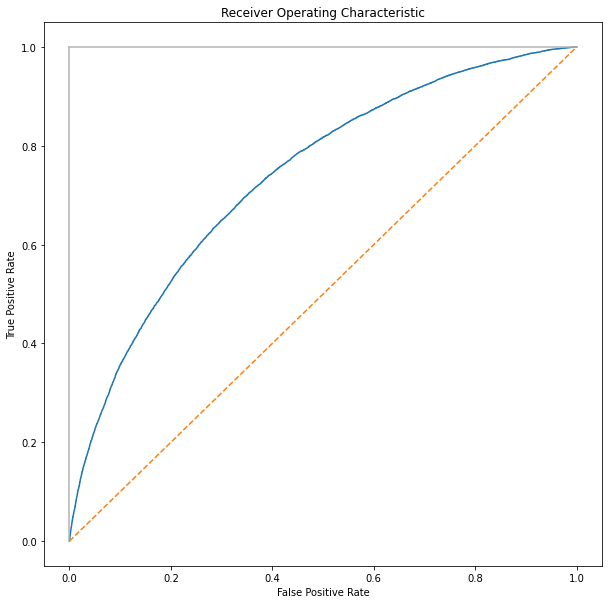
+In-sample results: 70% data for training set

+Out-of-sample results: 30% data for test set

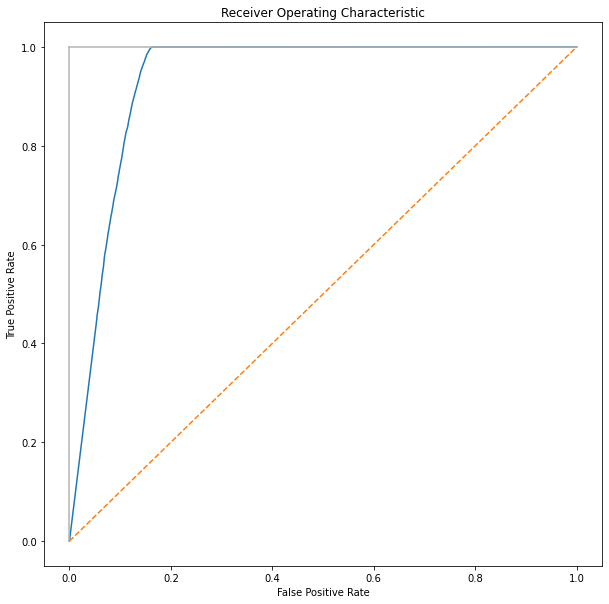
To evaluate the quality of the model I use classification\_report in sklearn , the return will give classification metrics: precision, recall, F1 score.The reported averages include macro average , weighted average and accuracy.

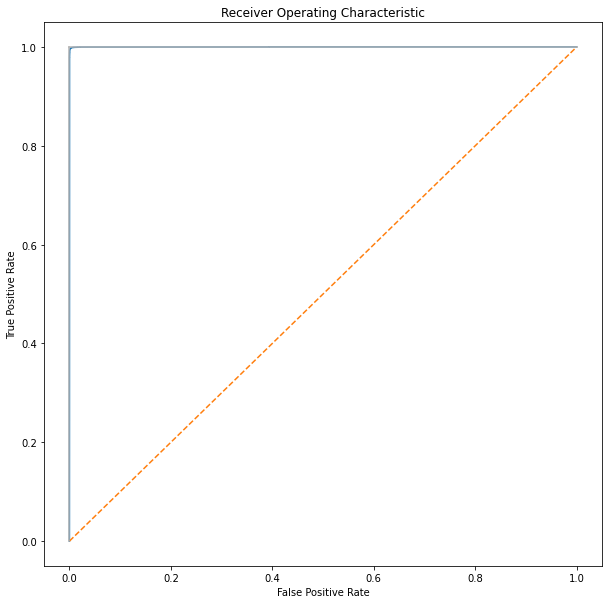
In addition, I also use AUC\_ROC to evaluate the performance of the model, ROC is a probability curve and AUC represents the classification level of the model. AUC\_ROC, also known as AUROC, is the probability that a randomly selected positive sample will rank higher than a randomly selected negative sample. The higher the AUROC, the more accurate the model.

After applying 3 corresponding models of logistic regression, decision tree and random forest, the results show that logistic regression has accuracy, macro avg f1 score, weighted avg f1 score are all equal to 0.67 and AUROC = 0.738. This shows that the classification ability of the Logistic regression model is fair.



Decision tree model with max\_depth=25 gives results with accuracy, macro avg f1 score, weighted avg f1 are all equal to 0.92 and AUROC = 0.93. This is a pretty good model that can be used to assess customer default.



Random Forest model with max\_depth = 25 gives the best value with accuracy, macro F1 score and weighted-averaged F1 score both. has a value of 0.98 In addition, the AUROC value = 0.99 (approximately 1) shows that the random forest model is almost completely accurate. This is the best model of the 3 models.  Banks can refer to this model to determine if potential customers have the ability to repay, as well as reject customers who are at high risk of default and are difficult to pay.