Capstone Project Final Report

Home Credit Default Risk: Predicting how likely each applicant is of repaying a loan?

1. Project Overview

The goal of this capstone project is to predict loan repayment ability of clients, namely the probability of default of clients. The reason behind the project is that many people struggle to get loans due to insufficient or non-existent credit histories. As a result, this population is often taken advantage by untrustworthy lenders. Hence, if lenders could predict the repayment ability better, unbanked population could have a better loan experience.

The objective in this project is to use historical data to predict whether or not an applicant will be able to repay a loan. This is a standard supervised classification problem, where the intended result is to be able to predict if a consumer applying for loan will be able to repay it on time (binary variable is zero) or will delay the repayment (binary variable is 1). We will apply machine learning models to learn from the available data and predict on new unseen data of consumers. Our desired output of the model is a probability of a client belonging to one of two classes – one which is likely to repay the loan on time and the other potentially delaying payments.

2. Dataset

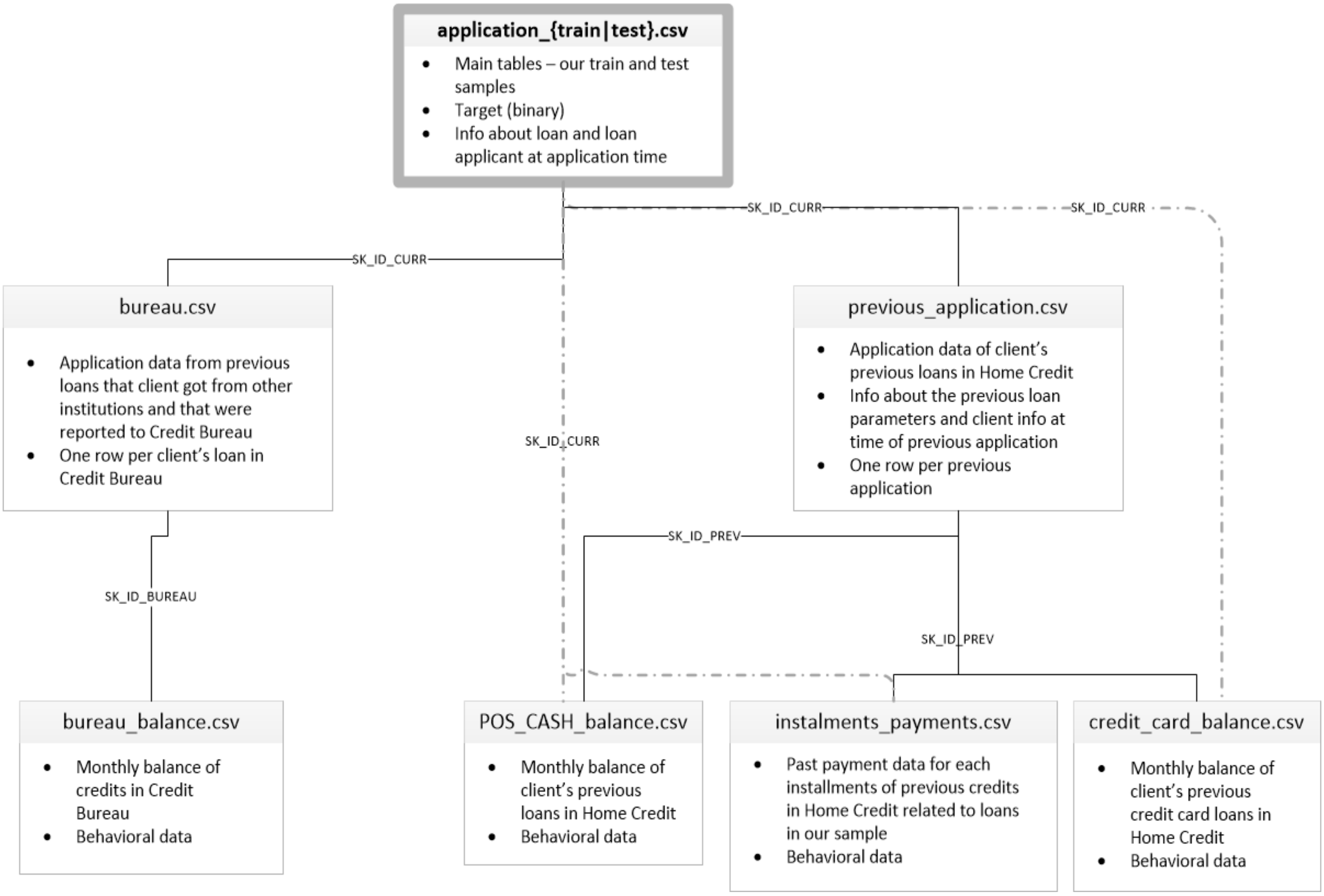
The dataset that we use for this project comes from publicly available Kaggle Competition dataset provided by Home Credit. Home Credit has provided a variety of datasets to predict their clients' repayment abilities. There are seven datasets and the relationshio between the different datasets is ilustrated by the diagram on the next page.

There are seven separate *.csv files were used as part of this project:

- **Application_train:** static data for all applications. One row represents one loan in our data sample. The training set consists of a label, TARGET, where 1 represents client with payment difficulties and 0 for those without. We have 120 features in this dataset, with 307,511 observations, each with a unique ID. This ID is used later to join information from the other datasets to the main dataset. The feature space covers essential key information for loan applications such as the amount of credit for the loan, annual income of the client, annuity of the loan. It also includes detailed information about the client's housing, the documents provided by the client during the application etc
- **Bureau:** All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample). For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- **Bureau Balance**: The bureau balance dataset gives us more details about each of the loans in the bureau dataset. This dataset gives us the monthly balance for the loans specified by the SK_ID_BUREAU in the bureau dataset and the status of the loan with categorical values such as Active, Closed, unknown or the binned amount of days past due.
- **POS Cash Balance:** Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit. This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample.
- **Credit Card Balance:** Monthly balance of clients’ all previous credit cards in Home Credit.

- **Previous Application:** All previous applications for Home Credit loans of clients who have loans in our sample. There is one row for each previous application related to loans in our data sample.
- **Installments Payments:** Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample. There is a) one row for every payment that was made plus b) one row each for missed payment. One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.

The following diagram provides a brief summary of the relationship between the datasets:



application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv

- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv

- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv

- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv

- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv

- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

credit_card_balance.csv

- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

The dataset contains personal and financial information belonging to 356,255 individuals who had previously been recipients of loans from Home Credit. These individuals are divided into training and testing sets:

- Training dataset: 307,511 Records, 122 Columns
- Test dataset: 48,744 Records, 121 Columns.

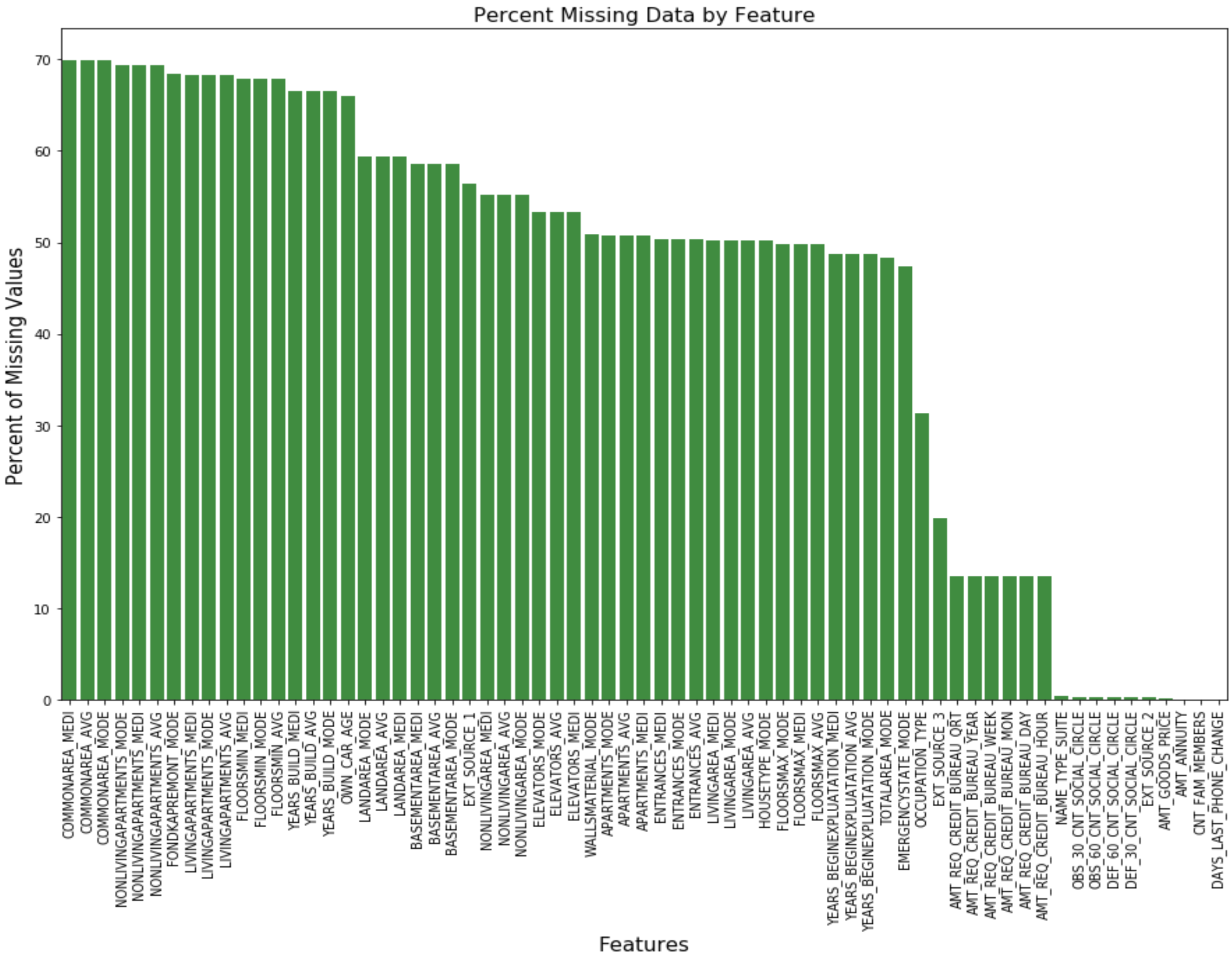
respectively. The loan ID is a unique identifier assigned to each borrower. A target value of 1 indicates that the borrower eventually made at least one late loan payment. A target value of 0 indicates that the borrower always paid on time.

The remaining 120 columns are explanatory variables that contain information about borrowers, including history of past payment, marital status, age, type of housing, region of residence, job type, education level, amount of bill statement.

Summary of the Datasets: The following table summarizes the total number of features and their data types across all the datasets:

Dataset	Observations	Number of Features	Categorical	Numerical
application_train	307,511	120	16	104
bureau	1,716,428	15	3	12
bureau_balance	27,299,925	2	1	1
Pos_Cash_balance	10,001,358	6	1	5
Credit_card_balance	3,840,312	20	1	19
Installments_payments	13,605,401	6	0	6
Previous_application	1,670,214	35	16	19

Missing values in the datasets: There are missing values in the datasets. Over 67 of the 120 total features have one or more entries of NaN entries. The graph below shows the features has missing value and precentage of missing value avaiable in training dataset.



Anomalies in the datasets: We looked at the specific columns, namely days_birth and days_employed to detect anomalies in these columns.

- Days of birth: Ages looks reasonable, max and min ages do not show anomaly:

```
count          307511
mean        43.936973
std         11.956133
min         20.517808
25%         34.008219
50%         43.150685
75%         53.923288
max         69.120548
Name: DAYS_BIRTH, dtype: float64
```

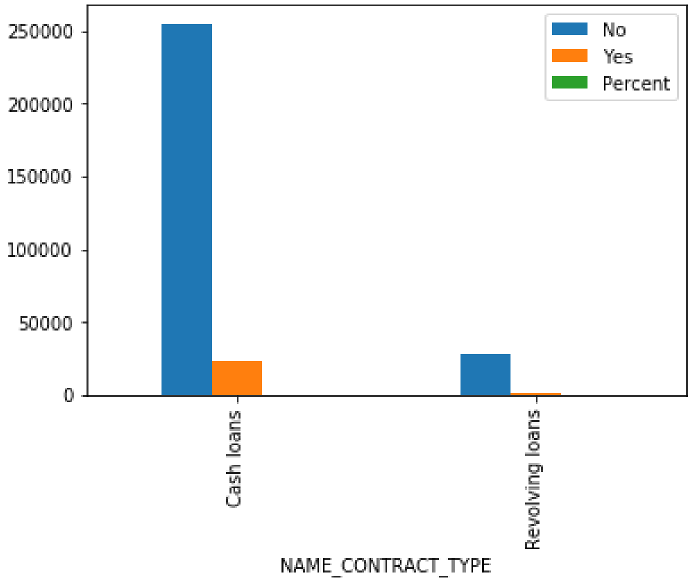
- Days employed: There is a problem in the maximum value, maximum value can not be 1000 years. These abnormal values were replaced with nan, to be treated as missing values.

```
count    307511
mean      63815.045904
std      141275.766519
min      -17912.000000
25%       -2760.000000
50%       -1213.000000
75%        -289.000000
max      365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```

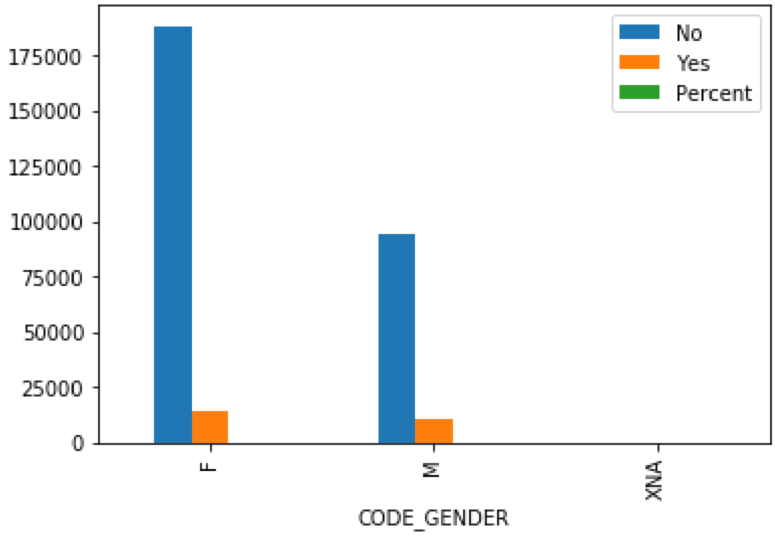
3. Exploratory Data Analysis (EDA)

We use exploratory data to summarize data sets' main characteristics with visual methods. We answer following questions by visually inspecting data:

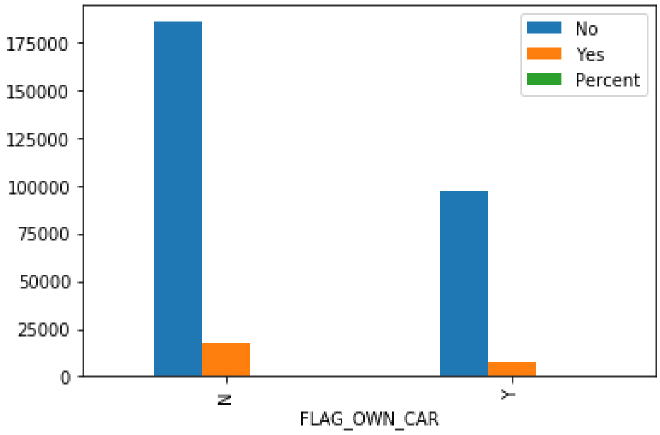
1. **Is there difference in default rate depending on loan type (cash vs revolving)?** Revolving loans are just a small fraction from the total number of loan. Default among revolving loans is lower, as well (cash loans: 8.3% vs revolving loans: 5.5%).



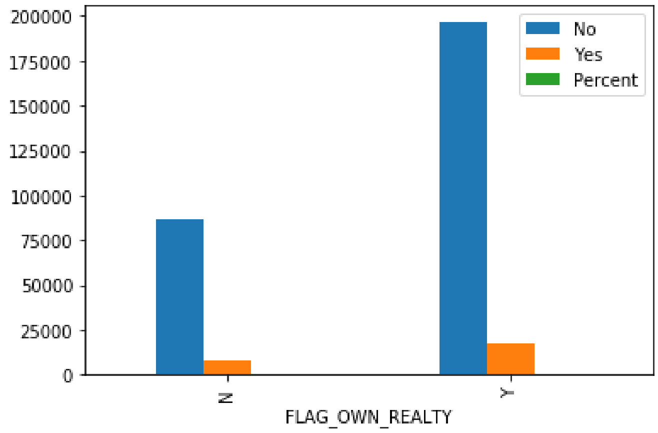
2. **Is there difference in default rate depending on gender?**
   Females have relatively more difficulties in repaying the loan back. The number of female clients is almost double the number of male clients. Looking to the percent of defaulted credits, males have a higher chance of not returning their loans (10.1%), comparing with women (7%). This could be because of the general larger population of female applicants as opposed to male applicants.



3. **If a customer owns a car, would it affect payment of loan?**
   The clients with a car are almost half of the ones that doesn't have one. The clients that owns a car are less likely to not repay loan. Both categories have close default rates (no car: 8.5% vs car owner: 7.2%).

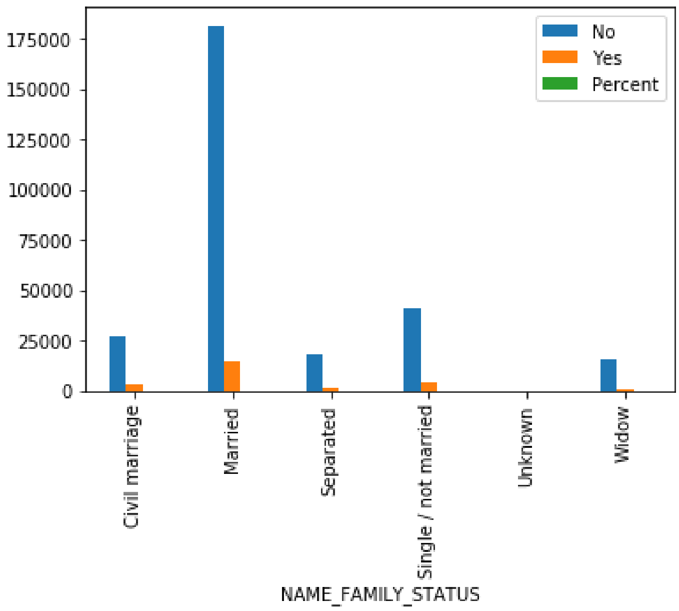


4. **If a customer owns a real estate, would it affect payment of loan?**
   The clients that own real estate are more than double of the ones that doesn't own. Both categories (owning real estate or not owning) have default rates close to 8%.



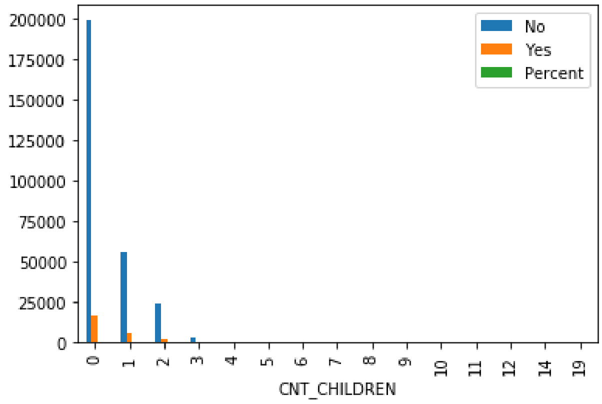
5. **Does the family status of client affects payment?**

Most of clients are married, followed by Single/not married and civil marriage. In terms of percentage of not repayment of loan, Civil marriage has the highest percent of default rate (10%) and Widow has the lowest.

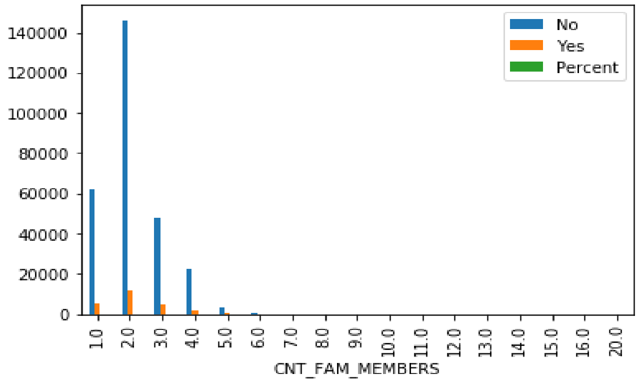


6. **Is having children affects payment?**

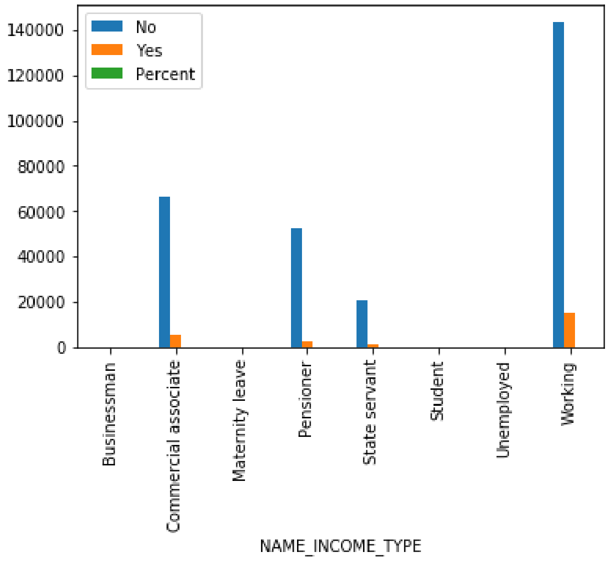
Most of the clients taking a loan have no children. As for repayment, clients with no children, 1, 2, 3, and 5 children have percents of default rate around the average (10%).



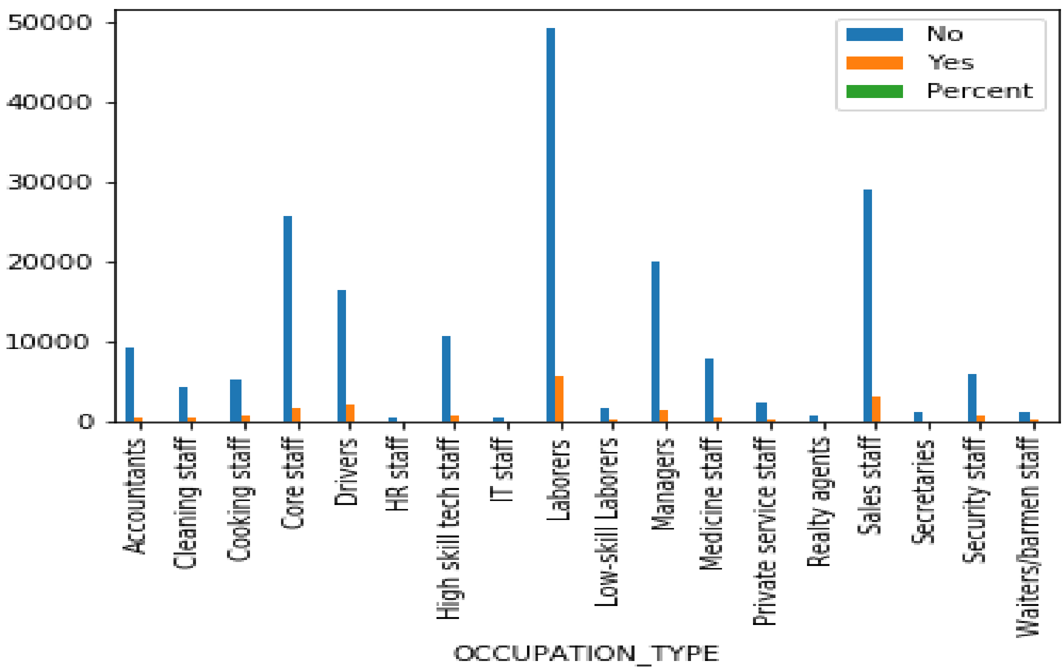
7. **Does family size have an affect on repayment?**
Clients with family members of 2 are most numerous, followed by 1 (single persons), 3 (families with one child) and 4. As for repayment, clients with up until 5 members is below 10%. Clients with family size of 11 and 13 have 100% not repayment rate. Clients with families size of 10 or 8 members have default rate over 30%.



8. **Does income type have an affect on repayment?**
Most of applicants have Working job status, followed by Commercial associate, Pensioner and State servant. The applicants with the type of income Maternity and Unemployed have highest level of default. The restare under the average of 10% not paying loans.

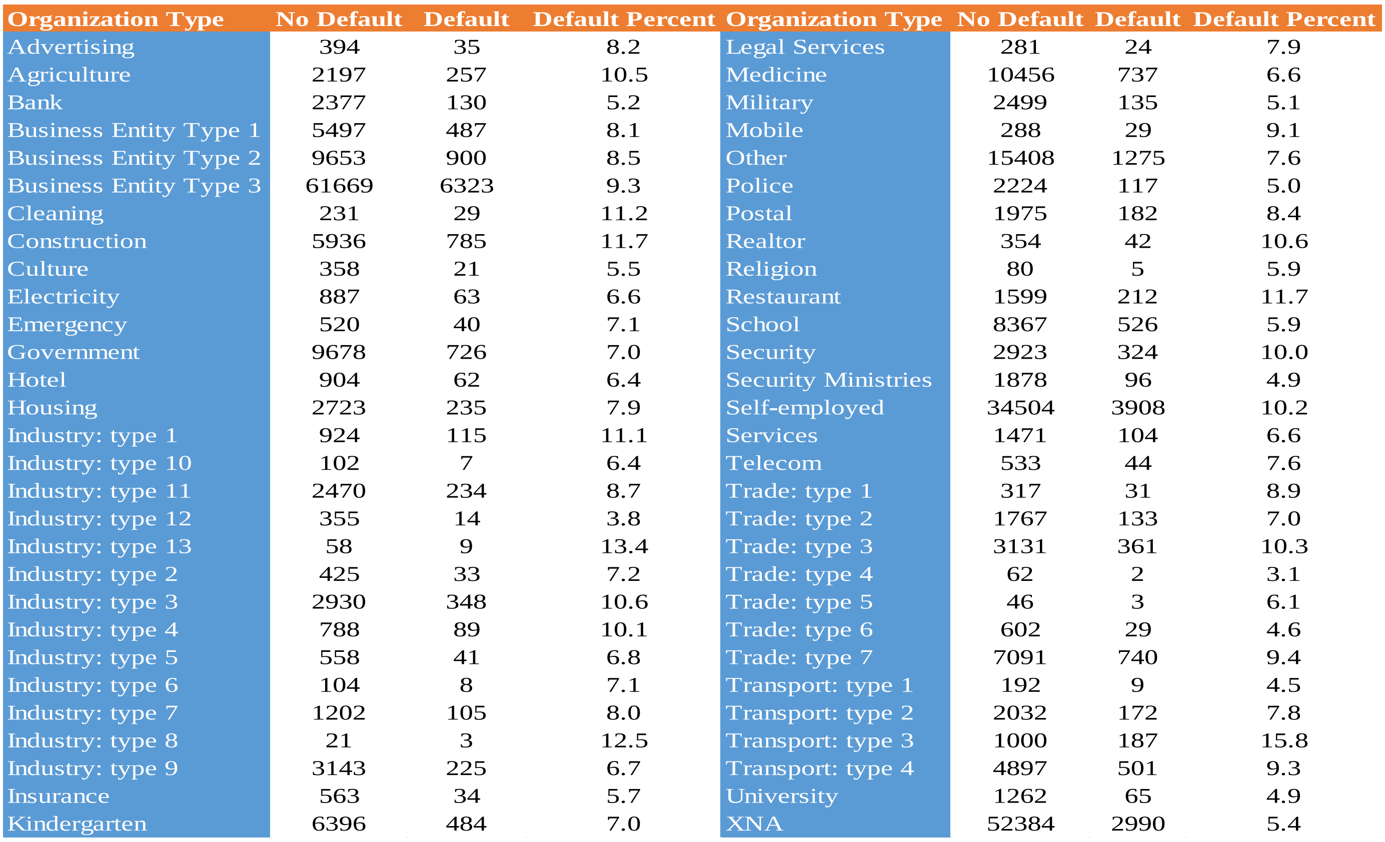


9. **Does the occupation have an affect on repayment?**
Most of the loans are taken by laborers, Sales staff. IT staff take the lowest amount of loans. Low-skill Laborers have the highest default rate (above 17%), followed by Drivers, Waiters/Barmen, Security staff, Laborers and Cooking staff.



10. Does the organization type has an affect on repayment?

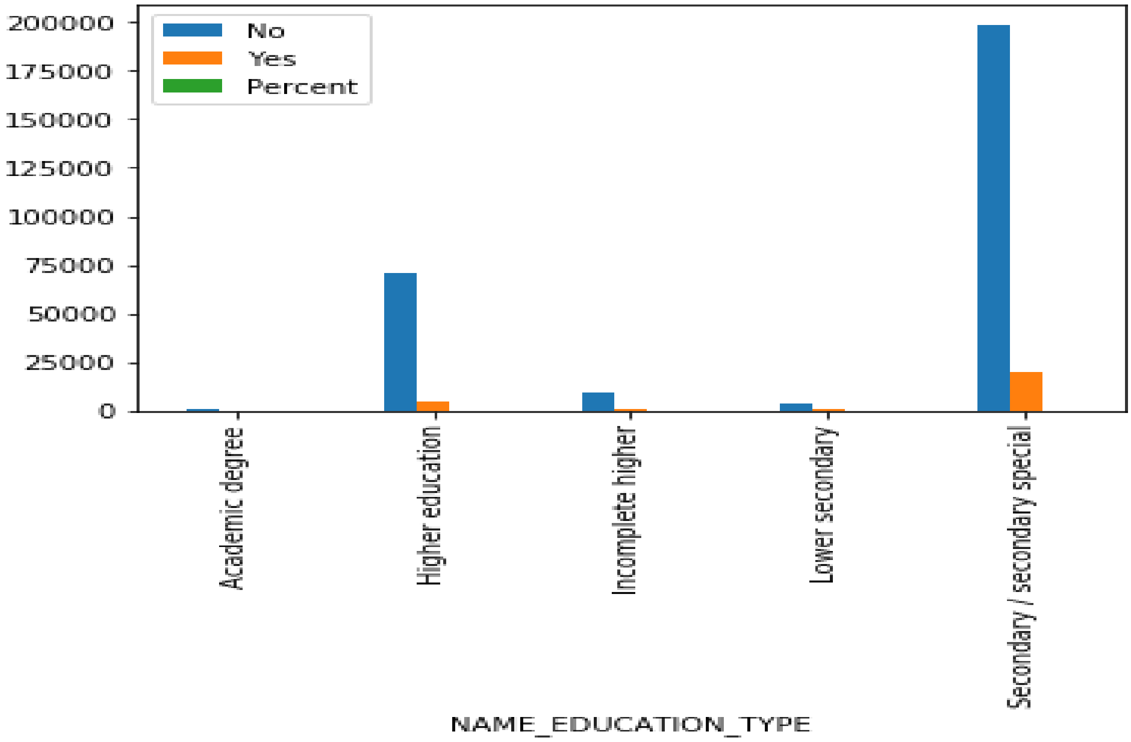
Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).



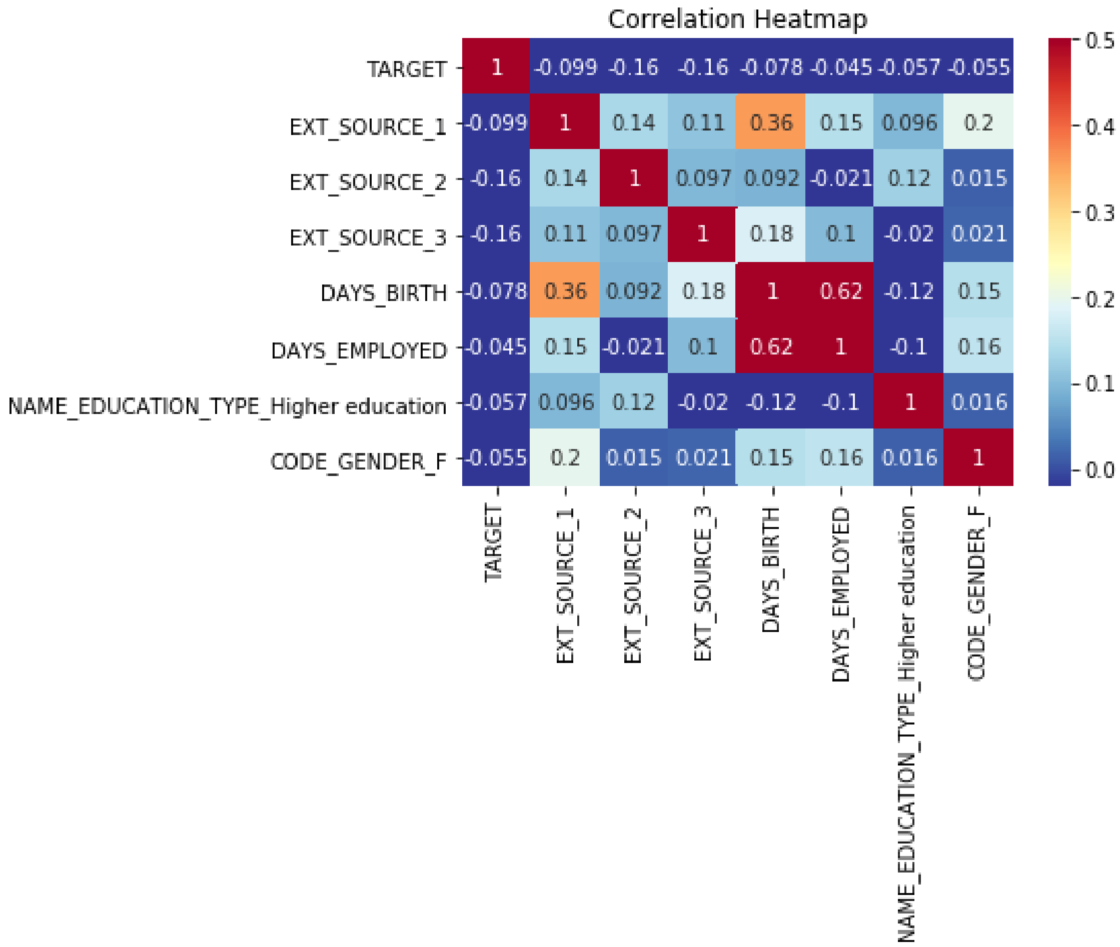
Organization Type	No Default	Default	Default Percent	Organization Type	No Default	Default	Default Percent
Advertising	394	35	8.2	Legal Services	281	24	7.9
Agriculture	2197	257	10.5	Medicine	10456	737	6.6
Bank	2377	130	5.2	Military	2499	135	5.1
Business Entity Type 1	5497	487	8.1	Mobile	288	29	9.1
Business Entity Type 2	9653	900	8.5	Other	15408	1275	7.6
Business Entity Type 3	61669	6323	9.3	Police	2224	117	5.0
Cleaning	231	29	11.2	Postal	1975	182	8.4
Construction	5936	785	11.7	Realtor	354	42	10.6
Culture	358	21	5.5	Religion	80	5	5.9
Electricity	887	63	6.6	Restaurant	1599	212	11.7
Emergency	520	40	7.1	School	8367	526	5.9
Government	9678	726	7.0	Security	2923	324	10.0
Hotel	904	62	6.4	Security Ministries	1878	96	4.9
Housing	2723	235	7.9	Self-employed	34504	3908	10.2
Industry: type 1	924	115	11.1	Services	1471	104	6.6
Industry: type 10	102	7	6.4	Telecom	533	44	7.6
Industry: type 11	2470	234	8.7	Trade: type 1	317	31	8.9
Industry: type 12	355	14	3.8	Trade: type 2	1767	133	7.0
Industry: type 13	58	9	13.4	Trade: type 3	3131	361	10.3
Industry: type 2	425	33	7.2	Trade: type 4	62	2	3.1
Industry: type 3	2930	348	10.6	Trade: type 5	46	3	6.1
Industry: type 4	788	89	10.1	Trade: type 6	602	29	4.6
Industry: type 5	558	41	6.8	Trade: type 7	7091	740	9.4
Industry: type 6	104	8	7.1	Transport: type 1	192	9	4.5
Industry: type 7	1202	105	8.0	Transport: type 2	2032	172	7.8
Industry: type 8	21	3	12.5	Transport: type 3	1000	187	15.8
Industry: type 9	3143	225	6.7	Transport: type 4	4897	501	9.3
Insurance	563	34	5.7	University	1262	65	4.9
Kindergarten	6396	484	7.0	XNA	52384	2990	5.4

11. How the education of client affects repayment rate?

Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. The Lower secondary category have the highest repayment rate (11%). The people with Academic degree have less than 2% not repayment rate.



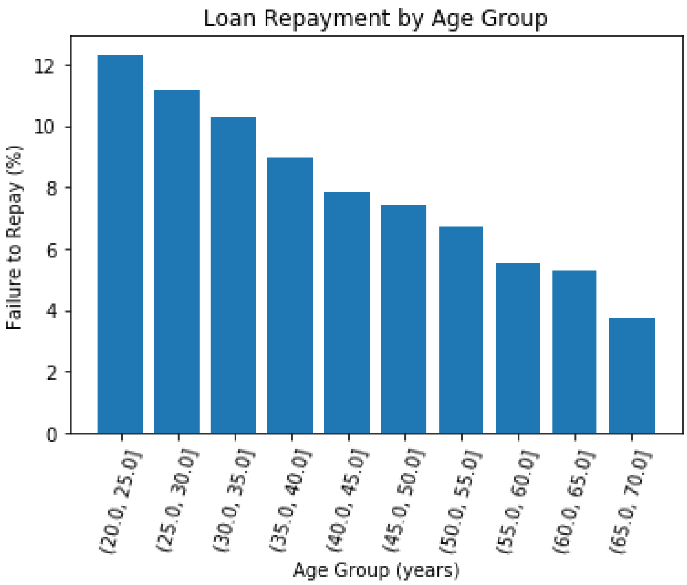
12. Other Exploratory Visualization and Correlations



EXT_SOURCE_1, EXT_SOURCE_2 and EXT_SOURCE_3 show inverse influence on risk repayment ability of the applicant based on above correlation heatmap.

All three EXT_SOURCE featureshave negative correlations with the target, indicating that as the value of the EXT_SOURCE increases, the client is more likely to repay the loan. We can also see that DAYS_BIRTH is positively correlated with EXT_SOURCE_1 indicating that maybe one of the factors in this score is the client age.

The bar plots below show the percentage of loan repayment failure in terms of age groups and years employed groups. Younger clients, in the age group of 20 to 25, appears to default the most on their loans. The risk of defaulting decreases gradually with age.



4. Algorithms and Techniques

The objective is to use historical loan application data to predict whether or not an applicant will be able to repay a loan. This is a standard supervised classification task. It is supervised learning problem as the labels are included in the training data and the goal is to train a model to learn to predict the labels from the features. It is considered classification problem since the label is a binary variable, 0 (will repay loan on time) or 1 (will have difficulty repaying loan). Since this is a classification problem, a naive benchmark score would be would be 0.5. Two models were picked to be tested as our baseline: logistic regression and random forest.

4.1. Logistic Regression:

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome and the outcome is measured with a dichotomous variable in which there are only two possible outcomes. The underlying algorithm of Maximum Likelihood Estimation (MLE) determines the regression coefficient for the model that accurately predicts the probability of the binary dependent variable. The algorithm stops when the convergence criterion is met or maximum number of iterations are reached. Since the probability of any event lies between 0 and 1 (or 0% to 100%), when we plot the probability of dependent variable by independent factors, it will demonstrate an 'S' shape curve.

We use LogisticRegression from Scikit-Learn. We will lower the regularization parameter, C, which controls the amount of overfitting (a lower value should decrease overfitting). In this way, we will get better results than the default LogisticRegression. In Scikit-Learn, we first create the model, then we train the model using .fit and then we make predictions on the testing data using .predict_proba.

4.2. Random Forest:

A random forest is an ensemble of decision trees. A large number of decision trees is created by sampling individuals and variables in the training dataset. A key difference from the decision tree is that each node is split by the best of a random subset of variables, rather than the best of all the variables. Each individual is classified by each tree and the most common outcome is used as the final classification

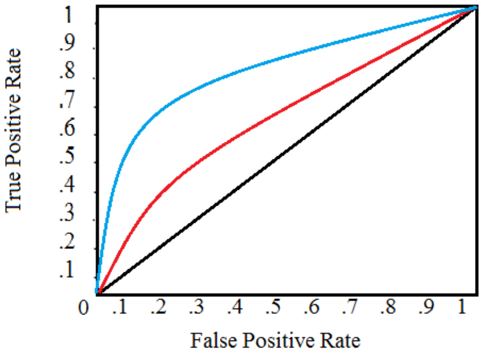
In order to improve our model, we will try Random Forest on the same training data and compare the performance. We will use 100 trees in the random forest model. We will compare our results in both models by using ROC AUC metric which is explained in the next section.

5. Results

In order to check the performance, we will use ROC AUC, Confusion matrix and F1 statistic.

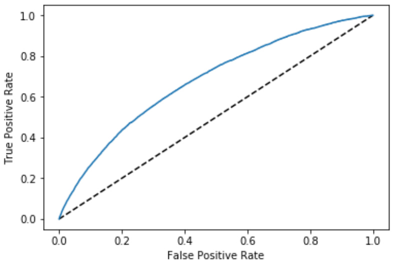
5.1. ROC AUC (Receiver Operating Characteristic Area Under the Curve):

First, we will use ROC AUC (Receiver Operating Characteristic Area Under the Curve) metric to judge the results. ROC curve graphs the true positive rate versus the false positive rate. True positive rate is on y-axis and false positive rate is on the x-axis. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.



Black line on the graph indicates the curve for a single model, and movement along a line indicates changing the threshold used for classifying a positive instance. The threshold starts at 0 in the upper right to and goes to 1 to the left. A curve that is to the left and above another curve indicates a better model. In the graph, the blue model is better than the red model, which is better than the black diagonal line which indicates a naive random guessing model. The Area Under the Curve (AUC) is simply the area under the ROC curve. (This is the integral of the curve.) This metric is between 0 and 1 with a better model scoring higher. A model that simply guesses at random will have an ROC AUC of 0.5.

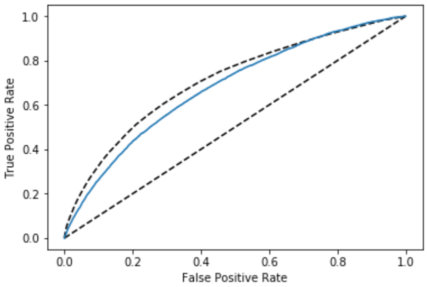
Logistic Regresion Model:



When we use Logistic Regression Model, ROC AUC score of model is : 0.678 and it is better than baseline model which has ROC AUC score of 0.5.

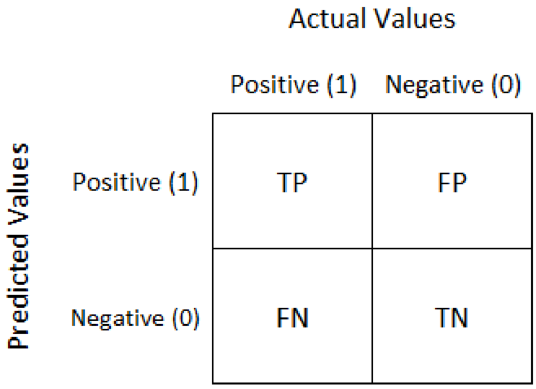
Random Forest Model:

When we fit Random forest model to the data, model improves compared to Logistic Regression as ROC AUC score increases to 0.708.



5.2. Confusion Matrix

Another metric for machine learning classification problem where output can be two or more classes is Confusion Matrix. It is a table with 4 different combinations of predicted and actual values.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Logistic Regresion Model:

```
[[93362     0]”
 [ 8117     0]]
```

In Logistic Regresion, looking at the confusion matrix recall rate (out of all non-default clients)is 92%, which means 92% of not defaulting clients predicted correctly. Precision is high at 100% as there are no false prediction in non-default clients.

Random Forest Model:

```
[[93359     3]
 [ 8112     5]]
```

In Random Forest, matrix recall rate (out of all non-default clients)is 92%, which means 92% of not d efaulting clients predicted correctly. Precision is close to 100% as there are only 3 false prediction in non-default clients.

5.3. F1 Score

F1 is an overall measure of a model's accuracy that combines precision and recall, in that weird way that addition and multiplication just mix two ingredients to make a separate dish altogether. That is, a good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0. In our case, F1 score is in both Logistic Regression and Random Forest models close to each other at 0.88.

6. Conclusion and Recommendations

In this project, we analyzed dataset provided by Kaggle to decide whether a customer would default or not to pay his/her loan. We used two models, namely Logistic Regression and Random Forest Model to predict whether the customer defaults or not. There were missing values, inconsistent values in the data, we overcame this problem by deleting inconsistent values and imputing missing values by mean.

In order to judge the performance of model, we looked at AUC ROC, Confusion matrix and calculated F1 score. Looking at the precision, recall rate calculated from Confusion matrix, both models perform similarly. F1 score of models also close to each other. On the other hand, AUC ROC metric indicates that Random Forest model performs better than Logistic Regression Model as ROC AUC score increases to 0.708 in Random Forest Model. In the extension of this project, other machine learning algorithms, such as Cross Validation Model. To reduce variability, in most machine learning methods multiple rounds of cross-validation are performed using different partitions, and the validation are averaged at the end. This method is known as k-fold cross validation. Increasing the splits would make predictions better but it may increase the time of processing.