Debt Payback Prediction Using Machine Learning Algorithms

Ankit Yadav, Sumit Bhong, Tamoghno Kandar and Naresh Balamurugan

***Abstract*—In the lending industry, investors (lenders) provide loans to borrowers in exchange for the promise of repayment with interest. If the borrower repays the loan, then the lender would make profit from the interest. However, if the borrower fails to repay the loan, then the lender loses money. Therefore, lenders face the problem of predicting the risk of a borrower being unable to repay a loan. In this study, the data from Home Credit is used to train several Machine Learning models to determine if the borrower has the ability to repay its loan. In addition, we would analyze the performance of the models (Random Forest and Logistic Regression). As a result, Logistic Regression model is found as the optimal predictive model.**

I. Introduction

The Banking sector emphasizes on the risk on credits lent by the bank. This credit is referred to as debt of a client. It is of high importance to understand the ability of the client to payback the debt. The sole knowledge and data of the past credit status of a client is not enough. This project aims to predict the ability of debt payback and hence reduce risk of Home Credit.

This project evaluates the various parameters signifying inability to payback or delayed payback. Considering factors like client characteristics also plays a vital role. During the consideration of loan sanctioning for a client, this project's results will help classify the client considering the various factors to predict their category whether payback of the loan lent is expected or not. We have also tried to understand through our analysis of the category of people that acquire more debts and also consequently are unable to payback.

II. Data Description

Data is collected from Kaggle, an online community that allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Total seven tables are provided with the problem statement. Complete data is provided by Home Credit, an international non-bank financial institution. This dataset is arranged similar to a database with seven relations having some foreign key constraints. This key were used to later aggragate the data and the relations were subsequently 'joined' to produce the final dataset

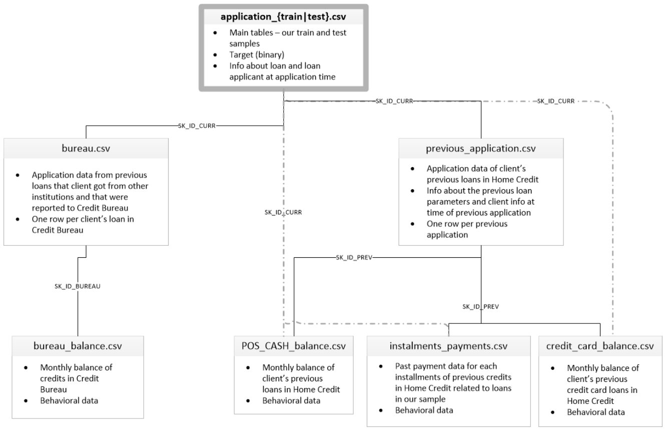

application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
SK_ID_CURR
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
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
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

Fig. 1. Depection of relationship among 7 tables

A. *Application* $\{train|test\}$

This is the primary table, which is categorized into two files: the train and test files. The train file has an additional column 'TARGET', which tells us if the client has paid back the debt or not, as this was collected from a kaggle competition (Unpaid — 1, Paid back — 0) is provided. There does not exist a 'TARGET' column in the test dataset. For the purpose of this machine learning project, we will use the application dataset labeled as "train" to build our model and test our results. This table also provides customer demographics and vital information, such as whether the applicant has a housing, the family size, and so on, all of which are compiled in 100+ columns.

B. *Previous Application*

This table contains information on the current client's past credit applications for Home Credit. This provides the client's credit history with Home Credit and allows us to profile the application based on his Home Credit history. This table contains data from prior filed applications in 37 columns.

C. Bureau

This table essentially contains all client's previous credits provided by other financial institutions (not obtained through Home Credit) that were reported to Credit Bureau. It comprises 17 columns that contain some basic information about the application, such as the amount, the end date, and if it is currently active. This allows us to develop a complete idea of the applicant by profiling him or her based on credit from other institutions.

D. Bureau Balance

This table shows the monthly balances of previous credits in Credit Bureau. This table includes one row for each month of credit history of every previous credit submitted to the Credit Bureau– i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.

E. Installments Payments

This indicates how many times the applicant has been unable to make the payments on time in the past, as well as how many times he has paid on the Home Credit credits– i.e there is a) one row for every payment that was made plus b) one row each for missed payment.

F. Credit Card Balance

This table provides monthly balance snapshots of previous credit cards that the applicant has 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 – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.

G. POS Cash Balance

This table 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 – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.

III. Methodology and Analysis

A. Exploratory Data Analysis on Dataset

In this report, exploratory data analysis is applied to check and handle the missing values, and necessary data transformations is conducted to process the data. The success of classification learning is heavily dependent on the quality of the data provided for training. Firstly we imported the packages available with python to ease our analysis. Moreover, imported the data in a standard python pandas dataframe form. In-depth analysis was conducted though we found a few issues in the basic data segmentation provided.

The usual plotting of graphs were carried out which provided understanding on the nature of the parameters. There is a general examination of the age distribution of candidates and the types of work they do. After that, we plot several variables against the target variable to check how much default exists on a categorical basis. The plots give an overview on the info that the variables can provided on debt payback.

A lot of columns have high percentage of missing values, this will affect our model so we deal this with omitting such columns. Missing values with lower percentages of missing values are handled by imputing. Also, the number of entries with our ”TARGET” variable being 1 turns out to be a small percentage of the total number of entries. This proves that our data is imbalanced.

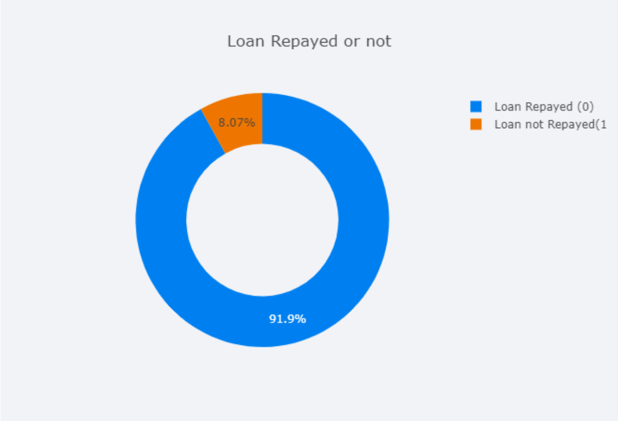


Fig. 2. Percentage split between ”TARGET” variables

B. Machine Learning Models

We've added data from additional tables to the application train dataset to train our model in order to improve accuracy and predictive power. We merge the columns from other datasets to gain more information. Our data was able to account for the applicant's credit history thanks to this aggregation and merging. As all clients not really have a credit history, like the new clients. Such missing values have been replaced with 0. Because the data is unbalanced, as we saw before, we must design the model in such a manner that the algorithm is not biased towards the category with the most observations. Divided the data into train, test and validation in the ratio 70:15:15 respectively. Two machine learning models are used namely:

1. Logistic regression
2. Random Forest

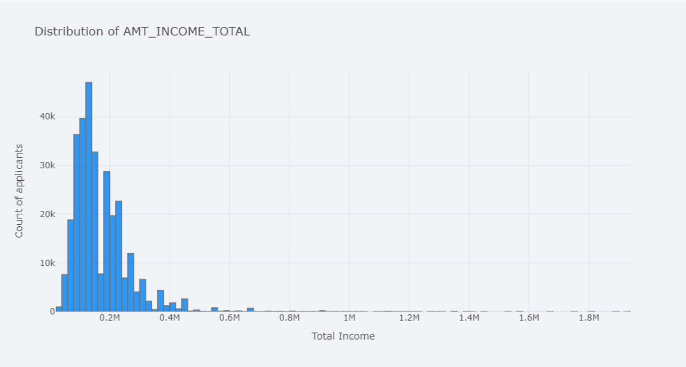
In order to compare the 2 models we are using ROC AUC to understand the accuracy of prediction

IV. Results

A. Exploratory Data Analysis

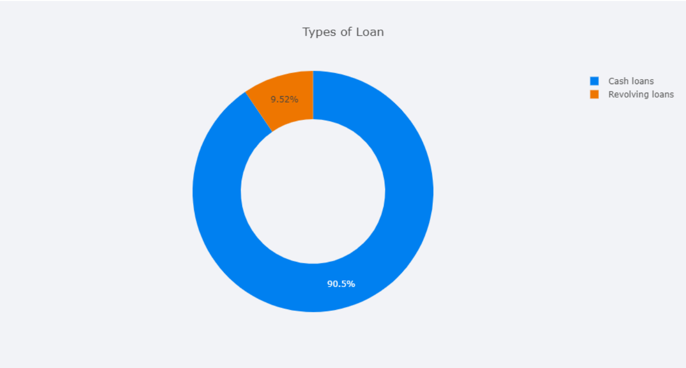
The exploratory data analysis show the following observations:

Distribution of Total Income:



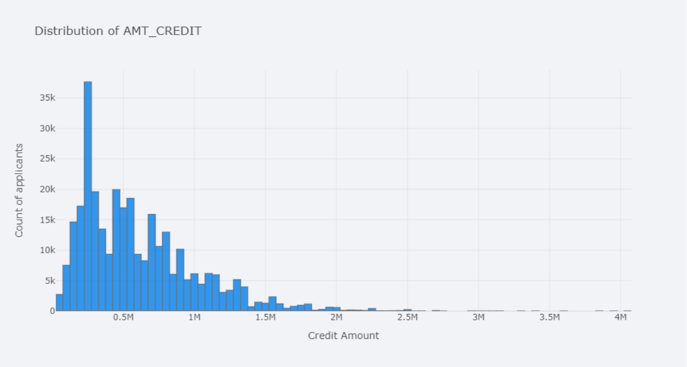
The distribution is right skewed and there are extreme values, we can apply log distribution.People with high income(>1000000) are likely to repay the loan.

Types of loan available

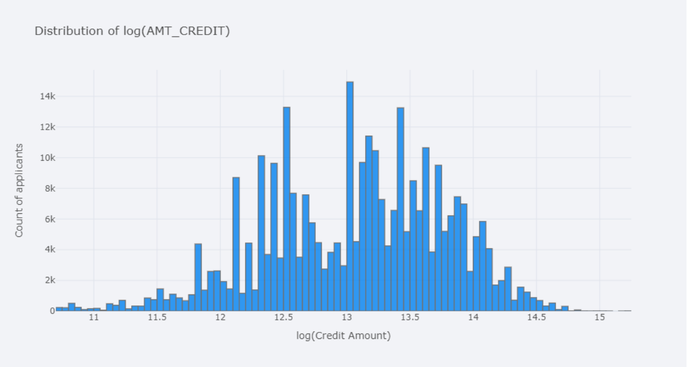


Many people are willing to take cash loan than revolving loan.

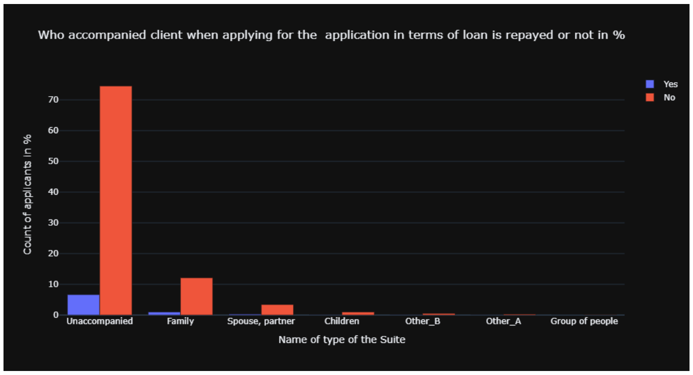
Distribution of Credit Amount



People who are taking credit for large amount are very likely to repay the loan. Originally the distribution is right skewed, we used log transformation to make it normal distributed.

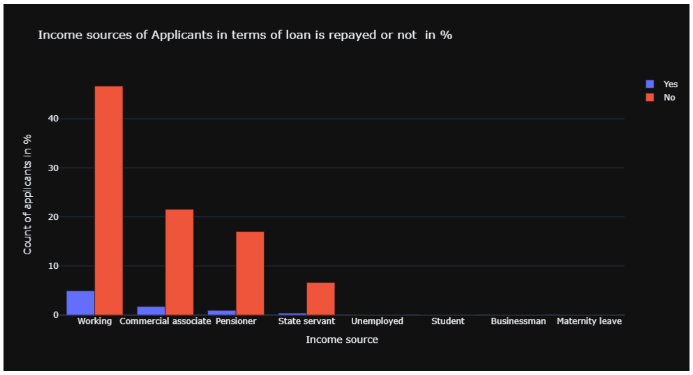


Type of the Suite in terms of loan is repayed or not :



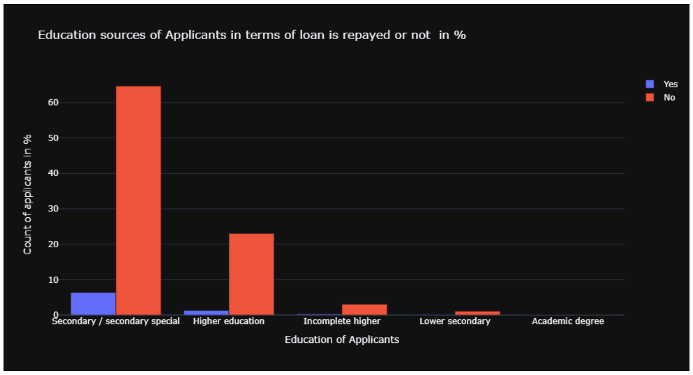
As we can see , most of the clients were unaccompanied when they came to apply for the loan.

Income sources in terms of loan is repayed or not :



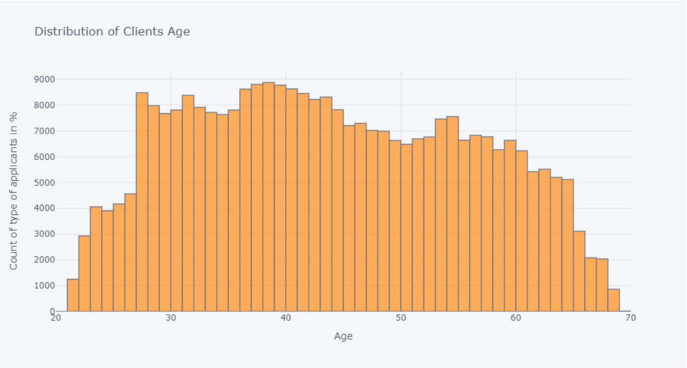
All the Students and Businessman are repaying loan

Education in terms of loan is repayed or not



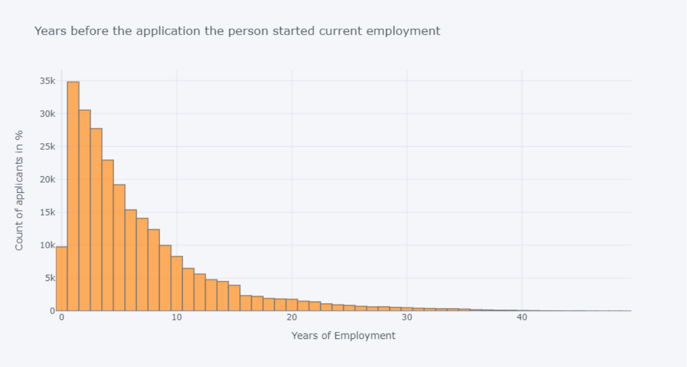
People with Academic Degree are more likely to repay the loan(Out of 164, only 3 applicants are not able to repay)

Distribution of Clients Age



Highest number of the loan applications are from clients belonging to the age group of 30-40.

Years before the application the person started current employment:



As we can see number of applicants decreases with increase in employment years.

B. Machine Learning Models

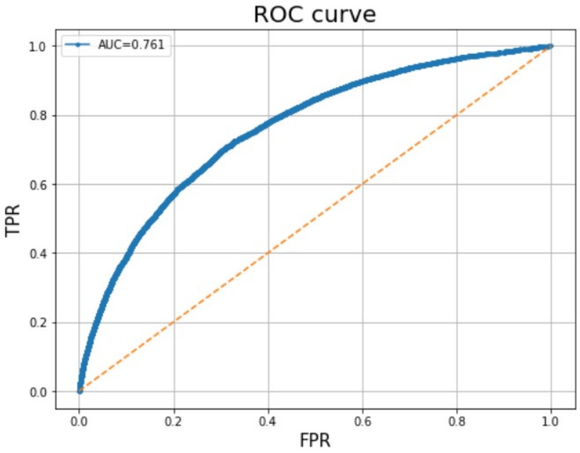


Fig. 3. ROC curve for Logistic Regression

TABLE I
AUC Score

Models	Train AUC	Validation AUC	Test AUC
Logistic Regression	0.763	0.757	0.761
Random Forests	0.844	0.751	0.750

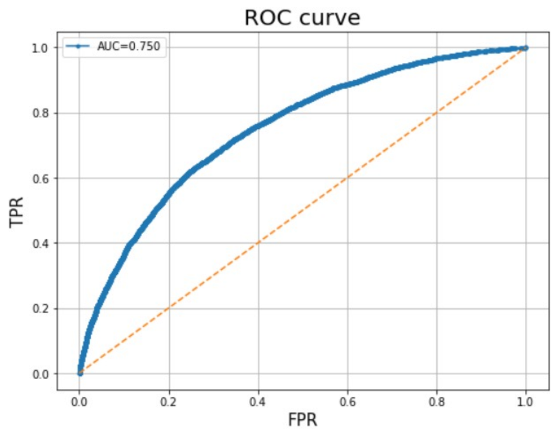


Fig. 4. ROC curve for RF

V. Learning, Conclusions, and Future Work

A. Future enhancement

In this study, there are several enhancements that we could make in the future. For example, outlier problem is not considered in the exploratory data analysis. if there are outliers in the dataset, the results of the predictive model will not be as valid as they are. In addition, the deep learning algorithm method should also be implemented when predicting for the loan the repayment status. In addition, more Machine learning models (Support Vector Machine, K-Nearest Neighbors and LightGBM) should also be implemented when predicting for the loan the repayment status. Also, the dataset should have been more accurate in terms of missing values and misentries which would provide us with a better training data which could increase our model's accuracy score.

B. Conclusion

Nowadays, the loan business becomes more and popular, and many people apply for loans for various reasons. However, there are cases where people do not repay the bulk of the loan amount to the bank which results in huge financial loss. Hence, if there is a way that can efficiently classify the loaners in advance, it would greatly prevent the financial loss.

In this problem, the data was imbalanced. So we couldn't use accuracy as a error metric. When data is imbalanced we can use Log loss, F1-score and AUC. ROC-AUC was chosen as the performance metric for the training models.An ROC curve is the most commonly used way to visualize the performance of a binary classifier, and AUC is (arguably) the best way to summarize its performance in a single number.

VI. References

- https://escholarship.org/uc/item/9cc4t85b
- https://www.kaggle.com/c/home-credit-default-risk
- https://www.youtube.com/watch?v=OAl6eAyP-yo
- https://www.analyticsvidhya.com/blog/2017/06/ which-algorithm-takes-the-crown-light-gbm-vs-xgboost/
- http://mlexplained.com/2018/01/05/ lightgbm-and-xgboost-explained/

VII. Contribution

- Ankit Yadav - Report making and helped in exploratory data analysis.
- Naresh Balamurugan - Report making, Video and Random forest.
- Sumit Bhong - Exploratory data analysis and video making
- Tamoghno Kandar - ML models and Video making.