Prediction of Loan Repayment for customers

Shubham Gupta
190020107
Department of Mechanical Engineering
Email ID : 190020107@iitb.ac.in

Pranamya Kulkarni
190040072
Department of Electrical Engineering
Email ID : 190040072@iitb.ac.in

***Abstract*—Financial institutions such as banks play a major role in the financial system and the economy of a country. Banks have a major responsibility to safeguard people's money and at the same time, support people and businesses by giving out loans. However, credit risk is one of the biggest risks that banks face while lending money. Credit risk is the risk of default, that is, when borrowers fail to repay the principal or interest on the loan. Loans are the biggest assets for a bank. Due to this nature of their business model, banks can never be fully protected from a credit risk. However, they can lower this risk in several ways. One way is to use an algorithm to predict the possibility of a default while giving out loans. This report proposes a logistic regression model to predict loan defaulters using supervised learning. A Deep learning model was further used to give better and more accurate predictions. Data model Evaluation is done on training set and based on the performance parameters, final prediction is done on the Test set.**

***Index Terms*—Machine learning, logistic regression, credit risk, prediction**

I. Introduction

Credit risk plays a major role in the banking domain. Gauging a customer's loan repayment ability is a critical business need to mitigate credit risk. But how does the business do that with insufficient or non-existent credit histories? Home Credit is a financial institution that faces this challenge of predicting loan default because the customers' credit histories are not enough. The objective of this project is to efficiently predict customer's loan repayment ability and minimize credit risk for Home Credit.

This study also aims to identify the factors contributing towards loan defaults, delay in repayments as well as the characteristics of a borrower who will honor all the obligations of a loan. The results enable us to determine the relationship between loan and customer characteristics and the probability to default. The results may also be used to appraise and monitor credit risk at the time of loan approval and during the currency of the loan. The analysis in this area has been done before but currently not many analyses use deep learning techniques.

Additionally, some of our other objectives are to find which categories of people tend to take more loans and also tend to default more, if the amount of credit is related to the default rate, what types of loans people tend to take more, to what degree does the credit history of an applicant, whether he/she has defaulted on a previous credit application, affect the chances of the current application being defaulted..

II. Datasets

We collected our data from Kaggle, an online community that allows data science practitioners access to public data and post their solutions. There are a total of seven tables provided with the problem. This complete data is provide 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

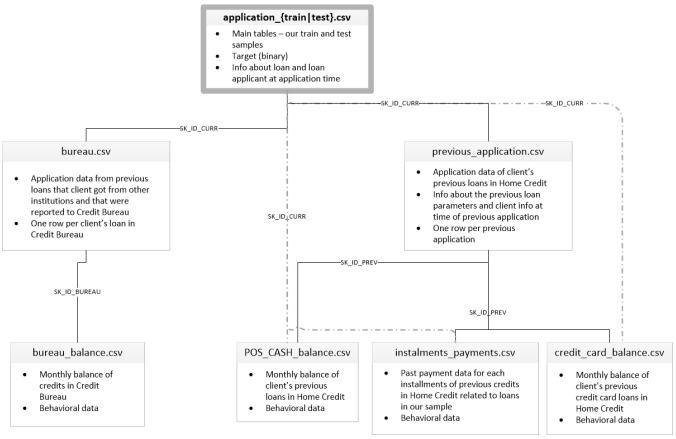


Fig. 1. Relationships between tables

1. Application train: This is the main table broken into two files : the train file and the test file. Since this was taken from a kaggle competition, the train file contains additional column TARGET which tells us the customer has repaid loan or not (1 is default, 0 is paid). The test dataset does not have a TARGET column. 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 contains client's demographic and important information about client which includes whether the applicant owns a house/car, number of family members etc complied in 122 columns.

2. Bureau: This table basically has the credit history of the applicant related to the credits not taken from Home Credit.

It has 17 columns having some basic data of the application like the amount, end date, if it is active right now etc. This helps us profile the applicant based on the credit from other institutions and help us create a broader picture.

3. Bureau Balance: This table contains monthly balance of previous bureau of credits. This table has one row for each month of history of every previous credit reported to Credit Bureau.

4. Previous Application: This table has the data of the previous credit applications applied in the Home Credit by the current applicant. This gives the credit history of client in Home Credit and helps us profile the applicant based on his past with Home Credit. This table has 37 columns having data on the previous submitted applications.

5. POS Cash Balance: This table contains monthly balance of previous bureau of credits. 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.

6. Credit Card Balance: contains monthly balance snapshot of previous credit cards. 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.

7. Installment Payments: This tells us how many times the applicant was unable to pay the installments in time in the past and how many times he did pay on the credits taken from the Home Credit.

III. Analysis Pipeline

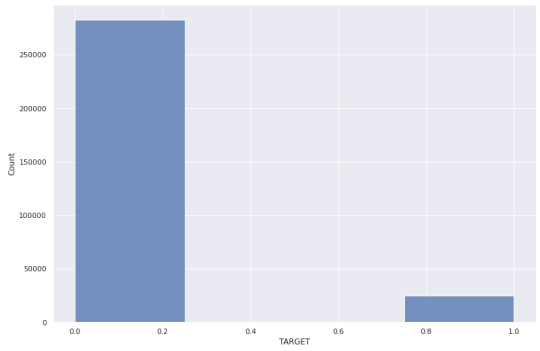
A. Exploratory Data Analysis on primary dataset

After importing all the required packages and downloading the dataset into a data-frame, we performed an initial study of the data, the results of which have been discussed in the later sections. The preliminary analysis of the data revealed a few constraints.

Some basic graphs were plotted which revealed some insights on the data. General analysis like age distribution of applicants and the type of jobs they do are analysed. After this we plot various variables to the target variable and see the amount of default. What are plotted are graphs showing distribution of variables under default and no-default conditions. This helps us get variables which help us get some info on the application at a first glance which maybe of help.

Missing Values – For some of the attributes, the percentage of missing values is higher. Some of the ways to handle missing values are dropping the columns with high number of missing values, dropping records with missing values, or imputation. We decided to drop those columns having more than 55% null values. The number of columns came down to 73 from 122. Next, we chose imputation as a method to deal with missing values in the remaining columns to avoid loss of information. The threshold was set at 14%, that is, we performed imputation for columns where null values were less than 14%.

Imbalanced data – The application data contains less than 10% records where target variable is equal to 1, that is, less than 10% of the sample has defaulted in the past.



Since the dataset has a large number of columns , we obtained the top 15 columns which have the most positive correlation and most negative correlation with the target variable .(This was done after making the values in the 4 days column positive, they were negative as it was measured with respect to the date of application).

B. Exploratory Data Analysis on the secondary and tertiary datasets

In the next step of our our analysis we explored the secondary and tertiary tables in our data. The tables explored are previous_applications, bureau, POS_CASH_balance and installment_payments.

previous_application: In this dataset what we tried to find was which kind of applications are more prone rejection and what type of applicants ask for more amount. The type of applicants refer to him/her being a repeater/new/refreshed applicant.
Bar plots and Box plots were plotted for this purpose. A cross tabs was also made to find the rejection rate for each type.

bureau: In this dataset the objective pursued was do clients have overdue on loans with higher amount? so we plotted some graphs between the days overdue and amount of credit and amount of annuity. We also tried to answer if applicants like to take higher amount of debt credit? Distribution plots were made for both Debt amount and application amount.

installments_payments: in this data how many installments are defaulted or had late payments was made and do people

always pay the exact amount or do they pay more or less? scatterplots were made to supplement the objective.

POS_CASH_balance: This data helped us answer if longer loans(having high number of installments) have more tendency to have dues or late payments? A simple scatterplot showed us if this was true or not?
The other two data tables bureau_balance and credit_card_balance were not used in analysis due to time constraints, a more detailed analysis will also cover these two along with some more and probably well crafted objectives.

C. ML Model training

To increase our accuracy and predictive power, we have added data from other tables to the application dataset to train our model. In the secondary and tertiary tables certain aggregates were taken grouped along the SK_ID_CURR. These aggregates were taken based on the EA conducted as well as some intuition. This aggregation and merging helped our data to account for the credit history of the applicant as well.
Since not all applicants need have a credit history the null vlaues thus occuring were imputed using median imputation and corresponding flag columns were also created in the table which will help the model to account for the 'nullness' of the value.
Since the data is imbalanced as we saw earlier, when building model, we have to design it in a way so that the algorithm will not be biased towards the category with high number of observations (in this case, target=0). The dataset is divided into two subsets, data0, having the paid application records (i.e., target=0), and data1, having the defaulted application records (i.e., target=1). Each of the subsets is further divided into train and test datasets in the ratio (90:10). Lastly, the corresponding train and test datasets for both data0 and data1 are merged together. The columns are then mean normalized and PCA is applied to reduce the number of columns to 150.

We have used the following models for prediction:
1. Logistic Regression
2. Light GBM
We have used ROC AUC as the primary performance metric for both the models.

1. Logistic Regression
We ran the logistic regression model on the above training data set. We chose the following hyperparameters for hyperparameter tuning using grid search:
Penalty type (l1, l2) - Used to specify the norm used in the penalization
Regularization parameter C - Inverse of regularization strength

2. Light GBM
We ran the lightGBM model on the above training data set. We chose the following hyperparameters for hyperparameter tuning using randomised search:
n_estimators – Number of boosted trees to fit.
max_depth – Maximum tree depth for base learners
reg_lambda – L2 regularization term on weights
num_leaves – Maximum tree leaves for base learners

Once the hyperparameter tuning is done, the metric score for both the models is compared. The model having a greater metric score is used for testing.

D. Deep Learning

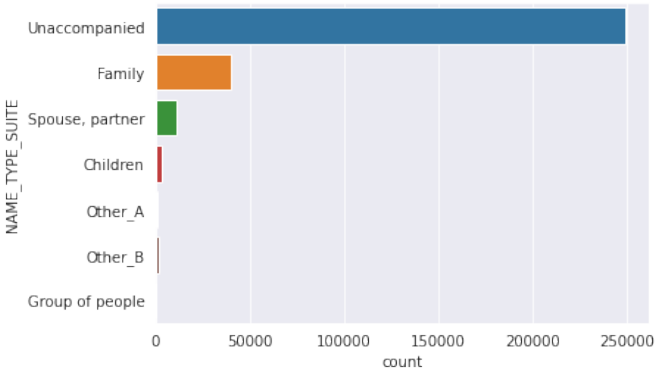
We used a deep neural network and trained it on the data for doing predictive analysis. A 5 layer neural network having 3 hidden layers was used and trained on a GPU, the number of neurons in each layer is varied and finally the best combination is selected which is trained and the best learning rate is found. All this is done using a hold-out validation set. Finally the the network with best hyperparameters is trained and finally tested on the test set.The train-hold-test split was done approximatly 80:10:10.

IV. Results

A. Exploratory Data Analysis on primary dataset

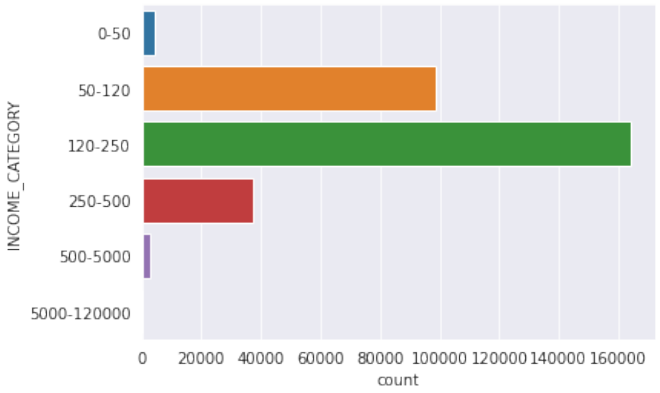
The exploratory data analysis shows the following observations:

Distribution of type of suite



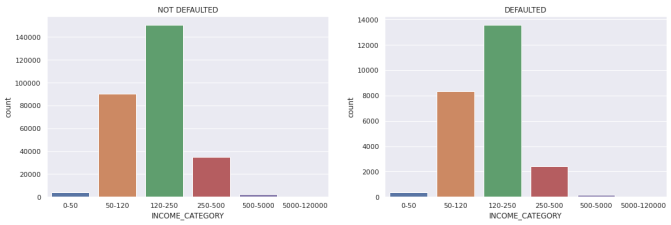
As we can see, the highest number of clients were unaccompanied when they came to apply for the loan.

Distribution of income range



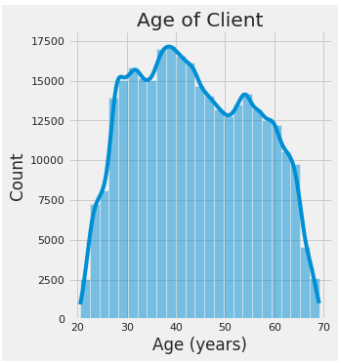
We have a lot of crowding of loan application from people having income between 120k to 250k . Outliers belong to the category 5000k to 120000k(Only 5 belonging to that category in contrast to 164387 in 120k-250k category).

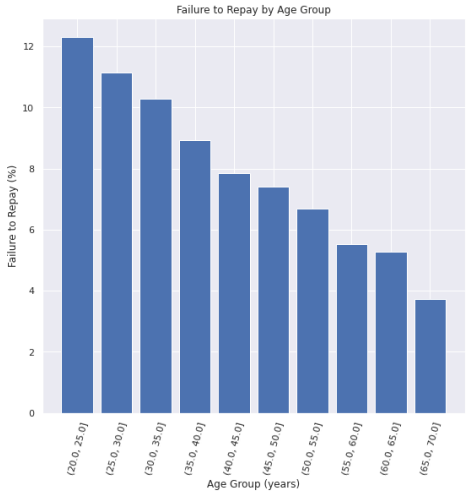
Distribution of income category wrt target variable



As we can see from the above graphs, both the defaulters and non-defaulters are highest in the income category of 120k-250k, also the spreads are very similar for both the target values.

Distribution of range of ages



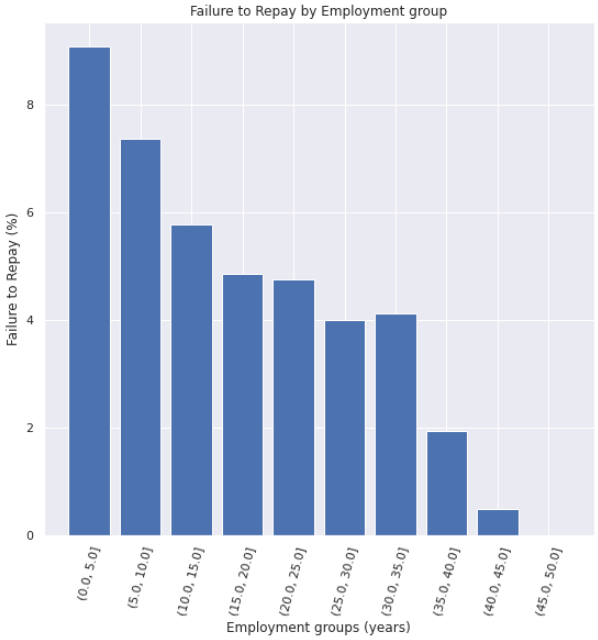


As we can see, the highest number of the loan applications are from customers belonging to the age group of 30-40.There is a clear trend: younger applicants are more likely to not repay the loan. The rate of failure to repay is above 10% for the youngest three age groups and below 5% for the oldest age group.

Distribution of number of days/years employed

This column has 1 outlier. The number of loan applications are the highest for those who have been employed between 1 to 5 years. Post that, we see that as the number of years of employment increases and an individual approaches old age and retirement age, the number of loan applications decreases.

Distribution of employment years group wrt target variable

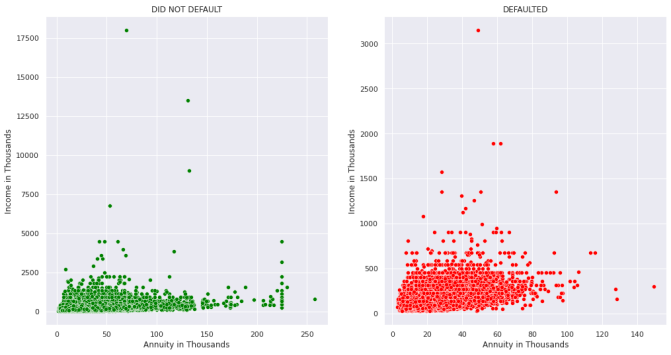


Again there is a clear trend similar to what we saw in age groups: Less employment years leads to higher failure

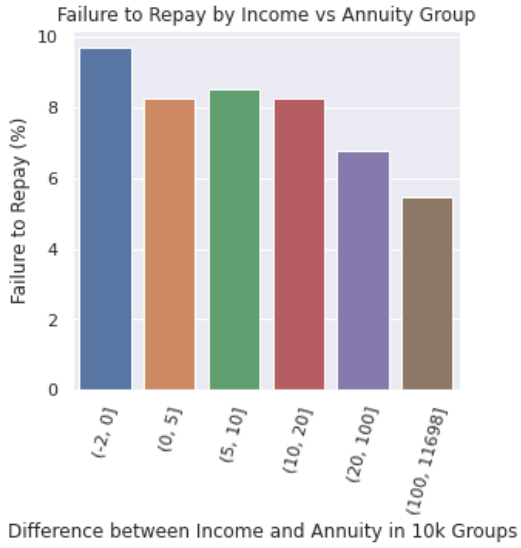
of repayment.

Instead of analysing columns AMT_ANNUITY and AMT_GOODS_PRICE individually, we can compare these columns with AMT_INCOME against the target variable.

Income vs Annuity against the target variable



We can observe that high income earners who get loans for higher annuity values end up defaulting much lesser than those in the low income bracket, even when they opt for lower annuity.



Here we can see that failure to repay is almost 10% (which is highest) when Annuity is more than Income . It decreases as gap between income and annuity increases.

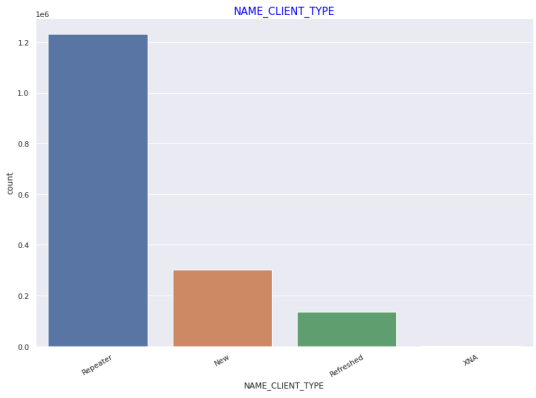
Income vs Goods Price against the target variable

Here we observe that low income earning applicants who seek to buy high value goods are more likely to default on payments.

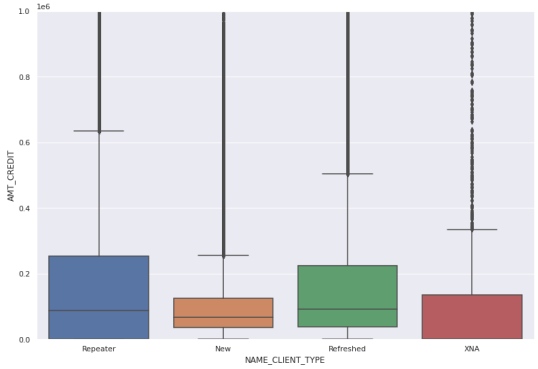
We can conclude from the exploratory data analysis that as the age of a customer increases and the numbers of employment years increases, the failure to repay decreases. Additionally, if the income amount is less than the annuity amount, then the customer is more likely to default.

B. Exploratory Data Analysis on secondary and tertiary dataset

The EDA of other data tables also gave some meaningful results.
We were able to ascertain that at Home Credit most of the applicants are repeaters followed by new ones.

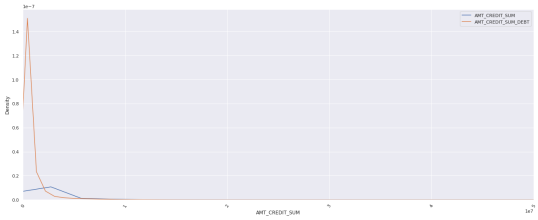


We can see that the repeater applications tend to have higher credit amount on average than the other applications, but refreshed applications also have higher amounts.

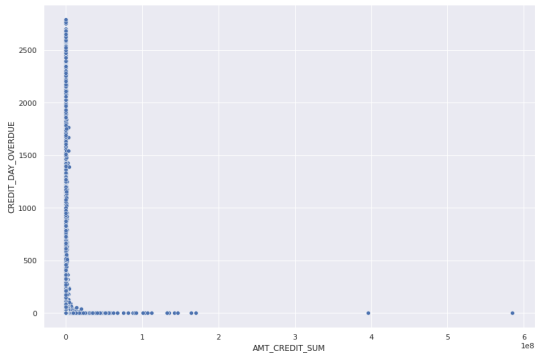


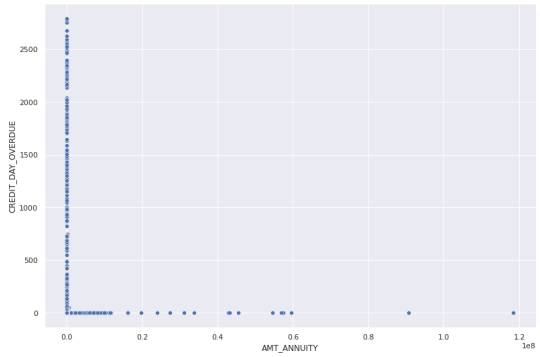
The following crosstab shows that almost all the new applications are approved, while almost half of the repeater applications are not approved.

NAME_CONTRACT_STATUS	Approved	Canceled	Refused	Unused offer
NAME_CLIENT_TYPE				
New	0.933290	0.011773	0.047886	0.007051
Refreshed	0.715818	0.144218	0.110823	0.029141
Repeater	0.534285	0.237344	0.211864	0.016507

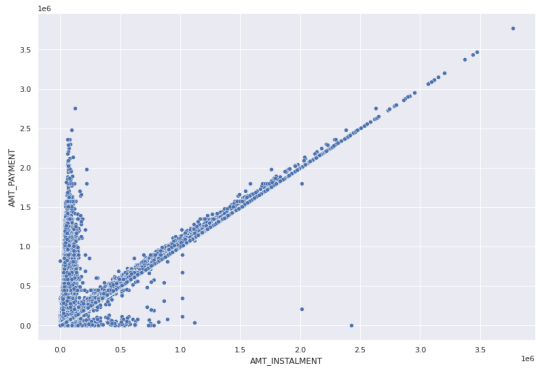


The density of debt amounts are more to lower amounts. Even credit amounts have higher density at lower amounts but not as much. This shows that clients tend to take higher amount credit not as debt but via some other means.

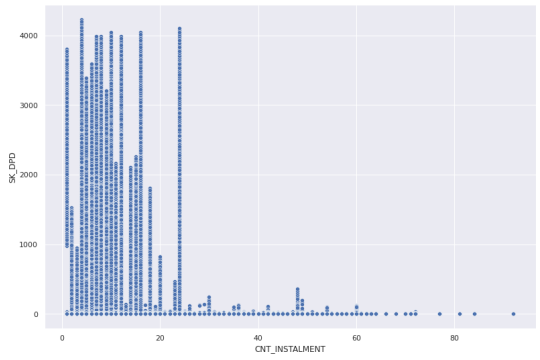




These two graphs clearly shows that more credit overdues happen at lower credit amounts contrary to the general idea that higher amount will result in more overdue.



The graph clearly depicts that most people pay same as the installment amount, and the trends grows stronger with larger installment amounts.



Higher the installment amount lower the DPD(days past due) a result similar to how people tend to keep less overdue on higher credit amounts. The probable reasons for these results as stated below in the discussion section.

C. ML model training

The PCA done had a RMSE reconstruction error of 0.2 which was 3% of the standard deviation of the norm of the columns in the dataset and hence it was deemed acceptable Hyperparameters selected for logistic regression model are: C = 1.57 and penalty = l2.

From model training, we found that the Logistic Regression model has the mean cross validated score of best estimator of 0.746
Hyperparameters selected for LGBM model are:
n_estimators = 500, max_depth = 5, reg_lambda = 0.1, and num_leaves = 10.
The LGBM model's cross validated score is found to be 0.741.

From this result, we can see that the Logistic Regression Model has a better performance for this dataset than the Light Gradient Boosting Machine (LGBM) Model. Hence, the logistic regression model was chosen for testing. The final ROC-AUC score for the model is found to be 0.505.

D. Deep Learning

The model selected was having (64,16,4) as the number of neurons in the hidden layers. the model trained using a learning rate of 0.1 using a SGD optimizer gave a 0.765 area under ROC curve for the validation set. On the test set the area under ROC curve came out to be 0.766. Hence the model is not overfit.

V. Discussion

In our initial data analysis, some of the observations were obvious. For instance, the default rate of the customer goes on decreasing as his/her age or the number of employment years increases. Similarly, people having an income greater than the annuity are more likely to repay the loan. A surprising observation, however, was that the income category of the applicant does not seem to affect the loan repayment.

Some of the results in that come of the analysis seem as if they are not what we should expect, but some amount of thought can help sort them out. The major one in these are that applications having higher credit and instalment amounts tend to have less overdues and payments days past due(DPD). This is not what one will expect as higher amount may attract difficulty in payments.SO how can this be explained? This result become much more intuitive if combine our results of application rejection and what type of application are submitted the most. Since most applications are repeaters and these also have the highest amount and they are also rejected the most. Hence these applicants have a credit history. So the company gives credit of higher amount only to trusted people or people having good credit history hence there is less chance of default.

We chose ROC-AUC as the performance metric for the training models. When we measure a classifier according to the ROC-AUC and predict the probability of default (between 0 and 1). The ROC curve is developed by plotting the ratio of True-Positive-Rate and False-Positive-Rate (TPR / FPR) for different values of the threshold from 0 to 1.

For training and testing, while logistic regression is faster and easier to implement, its major disadvantage is that assumes a linearity between dependent and independent variable. Hence, we have used LGBM as an alternative model and compared the performance of these two ML models. LGBM belongs to a family of boosting algorithms and converts weak learners into strong learners. A weak learner is one which is better than random guessing. Boosting is a sequential process; i.e., trees are grown using the information from a previously grown tree one after the other. This process slowly learns from data and tries to improve its prediction in subsequent iterations. Another model that could have been used here was Random Forest classifier.

The results of the ML models trained were not very good on the test set. a 0.504 area under ROC curve is almost same is random guessing and hence these models were not very good at prediction. This can be because of the many factors like non linearity and complexity of the problem taken. This alos showed why this type of models are not very prevelant in usage. A more exhaustive hyperparameter tuning might start to give better results but nothing can be said much

Further down the line when Deep Learning was implemented it yielded very good results owing to the non-linearities being better handled in the DL frameworks. Hence we can see a rise of Deep Learning in almost all the fields due to such a vast capability of adaptation to problem and why it is replacing traditional machine learning models. Probably this can be further improved if number of layers are increased and if we train longer but these are limited by the amount of resources available to us at the moment.

Acknowledgment

We would like to express our gratitude to Home Credit for providing the data that was used in the analysis.

We are highly indebted to Prof. S. Sudarshan, Prof. Amit Sethi, Prof. Manjesh Hanawal, and Prof. Sunita Sarawagi for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project.
We would like to express our gratitude towards our TA's Drumil and Arijit Jain for their kind co-operation and encouragement which helped us in completion of this project.
Our thanks and appreciations also go to our colleagues and friends who have willingly helped us out with their abilities in developing the project.

References

[1] Aziz, Hafiz Ilyas Tariq Sohail, Asim Aslam, Uzair Batcha, Nowshath. (2019). Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA). Journal of Computational and Theoretical Nanoscience. 16. 3489-3503. 10.1166/jctn.2019.8313.

[2] Aslam, Uzair Aziz, Hafiz Ilyas Tariq Sohail, Asim Batcha, Nowshath. (2019). An Empirical Study on Loan Default Prediction Models. Journal of Computational and Theoretical Nanoscience. 16. 3483-3488. 10.1166/jctn.2019.8312.

[3] K. Ulaga Priya1, S. Pushpa, K. Kalaivani, A. Sartiha. Exploratory analysis on prediction of loan privilege for customers using random forest.International Journal of Engineering Technology, 7 (2.21) (2018) 339-341

APPENDIX

Many of the graphs plotted are not included in the main report but can be seen form these notebooks with inferences given below them appropriately. Most of them do not give much of meaningful results and as such are not important to the analysis and hence are not added in the report.//

Links to the notebooks/code:

EDA :
https://colab.research.google.com/drive/1vtpgAAoSb3E9AU4lYAxkXaYHJLU374sI?usp=sharing

ML :
https://colab.research.google.com/drive/1N07jp8OQDf3tx1cZbX6cvDju0YA_qO-R?usp=sharing

DL :
https://colab.research.google.com/drive/1jXFBelf7iGeVjmY_6oOFzerUdKKD6bPc?usp=sharing