# 

# **Vehicle Loan Default Prediction**

# Anshu Yadav

# December 2020

**Problem Statement**

Vehicle Loan default prediction is a common problem for various financial companies and a well-defined and known problem in data science. This is the type of problem banks, credit card companies, micro-credit banks or FinTech companies face whenever customers ask for a loan. Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased vehicle loan rejection rates. The need for a better credit risk scoring model is also raised by these institutions. This warrants a study to estimate the determinants of vehicle loan default. A financial institution has like to accurately predict the probability of loanee/borrower defaulting on a vehicle loan in the first EMI (Equated Monthly Instalments) on the due date.

Applying machine learning to loan default predictions showcase a useful application of this branch of artificial intelligence to solve real-world and business problems. We will try to build this model with the most transparency possible as to mirror the conditions in which financial institutions must disclose this process.

**Data Wrangling:**

The source data for vehicle default loan dataset was chosen from Kaggle. Following Information regarding the loan and loanee are provided:

1. Loanee Information (Demographic data like age, income, Identity proof etc.)
2. Loan Information (Disbursal details, amount, EMI, loan to value ratio etc.)
3. Bureau data & history (Bureau score, number of active accounts, the status of other loans, credit history etc.)

The below list of cleansing and cleanup tasks was performed on the data:

* Dataset came into two sets (train and test set). Training dataset has 233154 rows and 41 columns whereas the test dataset has 112392 rows and 40 columns. All the columns between train and test dataset were similar except the Loan Default column which we want to predict.
* Data wrangling and cleaning was done on Training data set.
* None of the features contain null values except “Employment Type”. Employment type contains 7661 null values, leaving those as nan.
* Columns having data in months and years were converted to months only.
* On the basis of class distribution, it was found 182543 people don't have any default loan, only 50611 have loan default problems.
* Outliers were reviewed to determine if they are true outliers or bad data that should be removed
* Outliers on were imputed with mean.
* 'STATE\_ID', 'EMPLOYEE\_CODE\_ID', 'SUPPLIER\_ID', 'MANUFACTURER\_ID', 'CURRENT\_PINCODE\_ID','BRANCH\_ID' were dropped and rest was included during feature selection.

The final shape of the cleaned data set was 233154 records and 25 parameters.

**Exploratory Data Analysis:**

The dataset has 233154 rows and 41 columns. The data dictionary file provided with the dataset, indicates that columns are information about the borrower and the outcome of their loan repayment. This data is from L&T Financial Services & Analytics Vidhya, headquartered in Mumbai, LTFS is one of India’s most respected & leading NBFCs providing finance for two wheeler, farm equipment, housing, infra & microfinance. This means that we are trying to model future outcomes either vehicle loan is default or not.

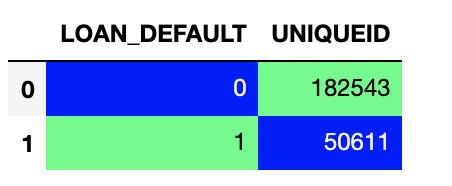


Figure 1: Representation of number of loan default on the basis of uniquid

First, I wanted to see if any of the columns has the missing values. To visualize that I create the heatmap. The figure below displays that ‘Employment Type’ has missing values. There were 7661 missing values which are 3.3% of total values. This was taken care of by imputing “missing” at the missing value place.

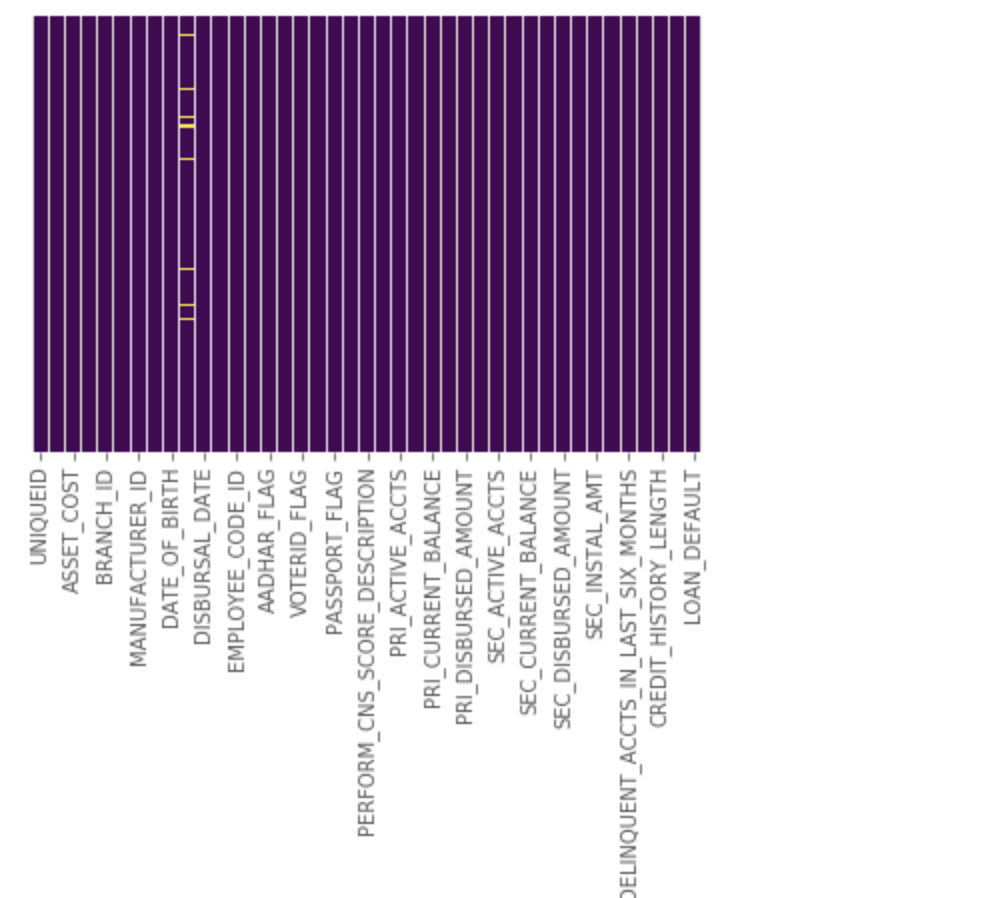


Figure 2: Heatmap of training dataset of 21 columns

Histogram of default vs disbursal date was plotted. It was found that the number of loans given with no default was way more than loans undergo default which means that the model has a good chance of getting it right by predicting all good outcomes. We need to be more careful while building the model. There is high risk of overfitting.

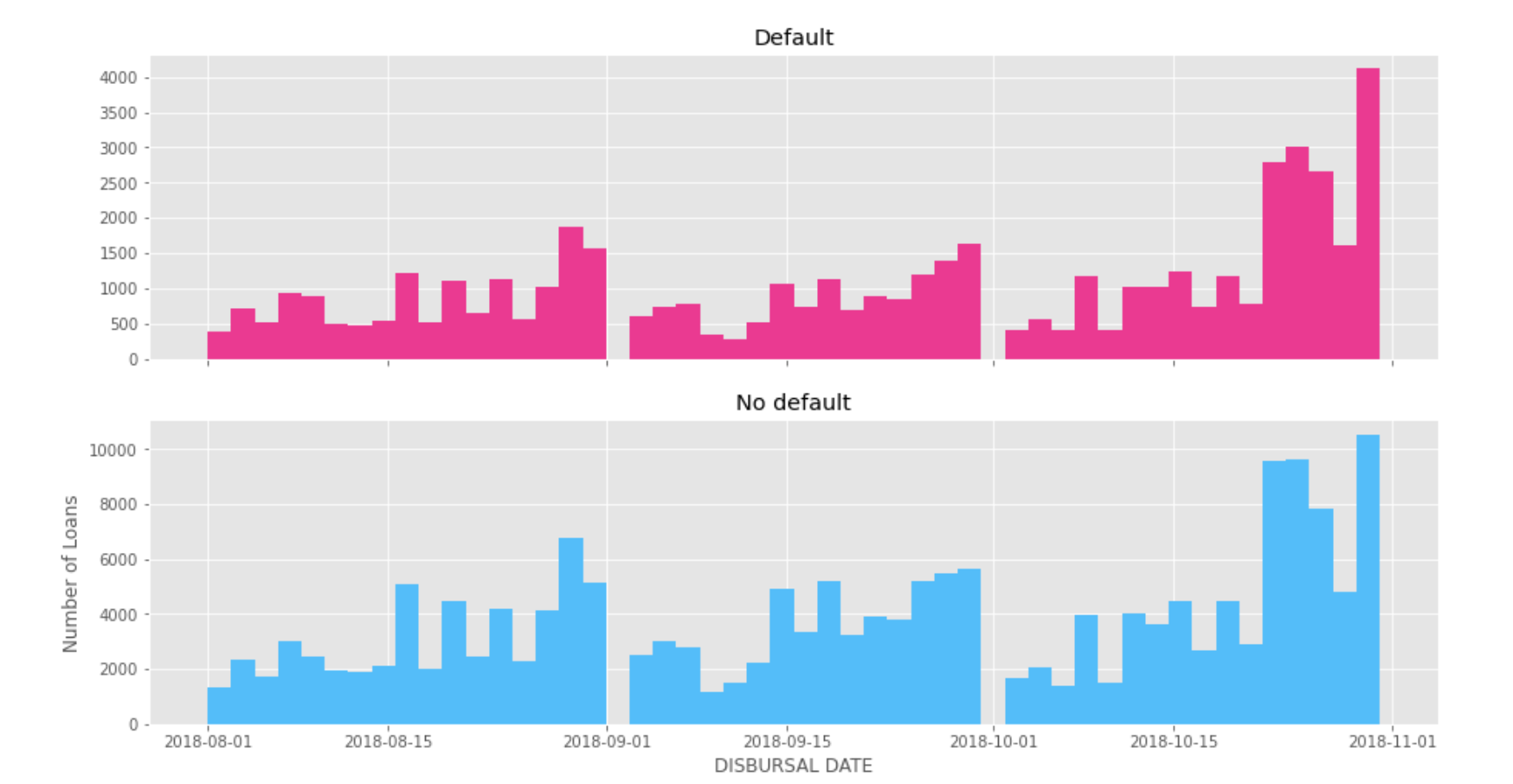
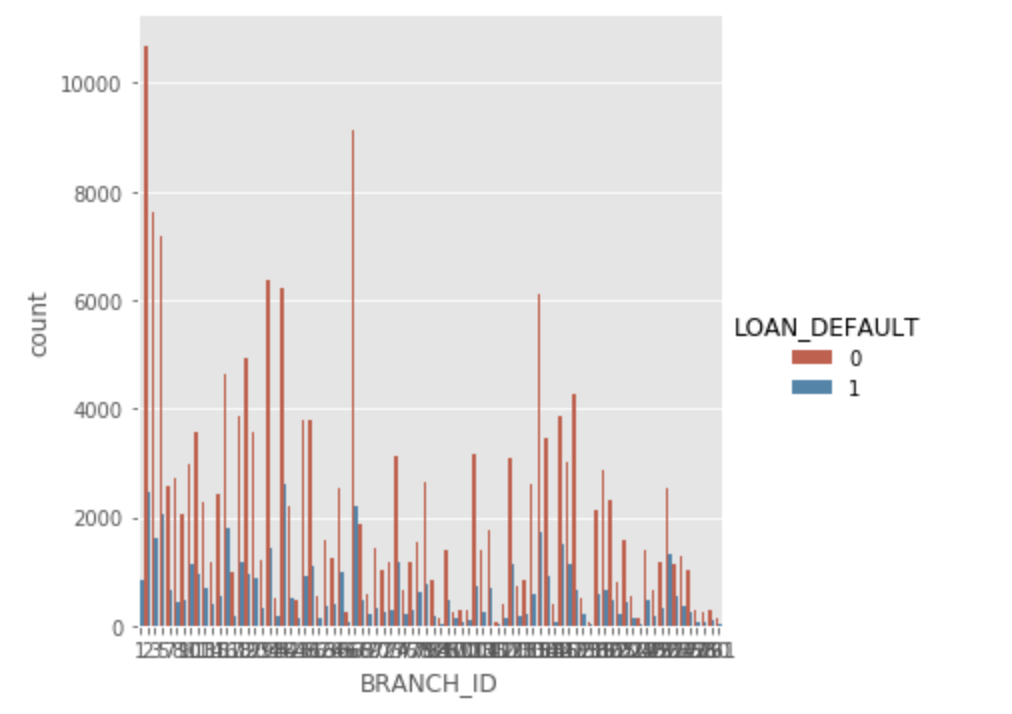


Figure 3: Histogram of number of loans vs Disbursal date

Similarly, I explored loan default cases on the basis of branch Id to see if it is some particular branch issue. But it is found that data shows there are a few numbers of default cases irrespective of branch.

 Figure 4: Count plot of branch id on the basis of Loan Default

Less important features were dropped. To look at the highly correlated features, a correlation matrix was made as below. It was found that Secondary sanctioned amount is highly correlated with primary disbursed amount. 'PRI\_NO\_OF\_ACCTS\_new', 'PRI\_OVERDUE\_ACCTS\_new'are perfectly positively correlated and hence kept one. Similarly, 'SEC\_NO\_OF\_ACCTS', 'SEC\_ACTIVE\_ACCTS' are highly positively

correlated, hence keeping one also. Same way, 'SEC\_CURRENT\_BALANCE', 'SEC\_SANCTIONED\_AMOUNT', 'SEC\_DISBURSED\_AMOUNT' are highly positively correlated, hence keeping one. Feature which are not highly correlated with anyone: 'PRI\_ACTIVE\_ACCTS','PRI\_CURRENT\_BALANCE','PRI\_SANCTIONED\_AMOUNT','PRI\_DISBURSED\_AMOUNT','SEC\_OVERDUE\_ACCTS'.

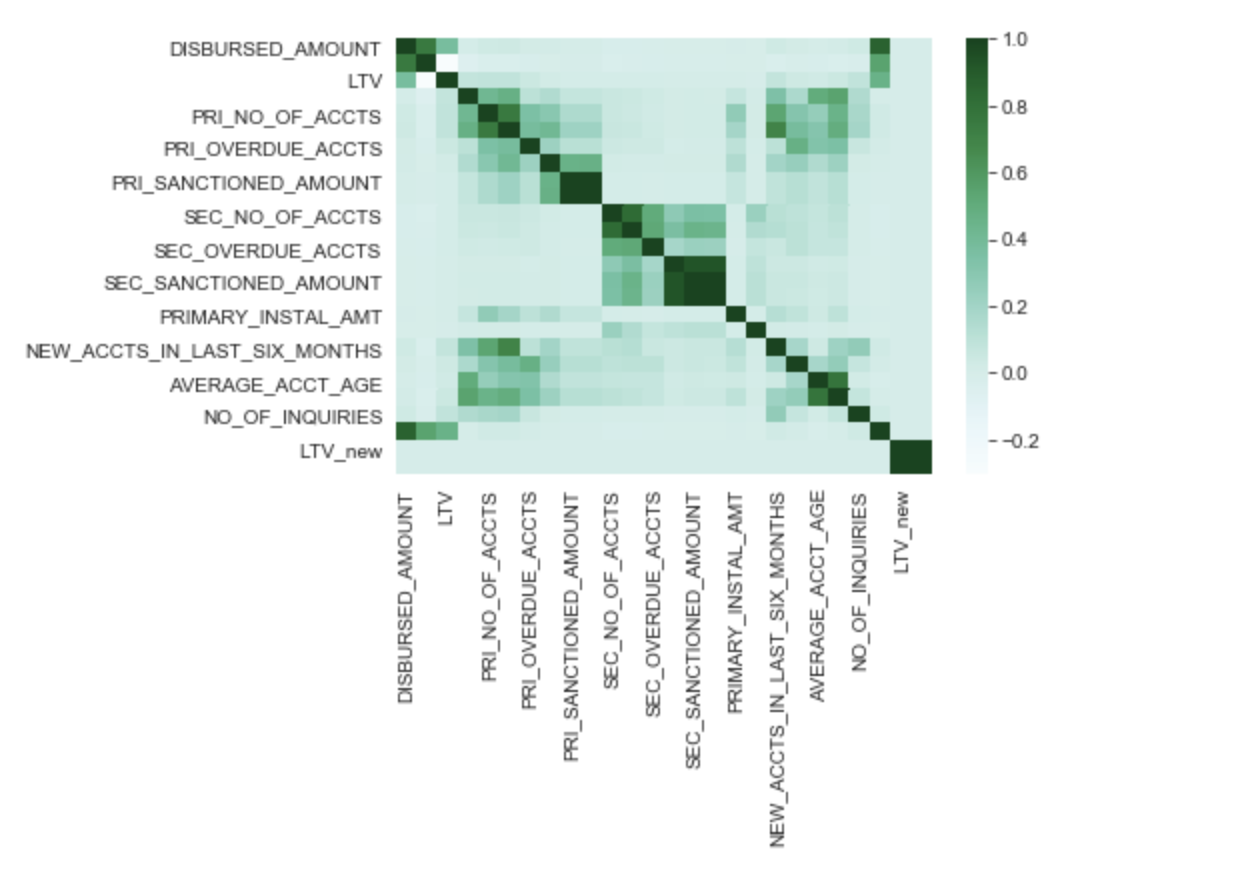
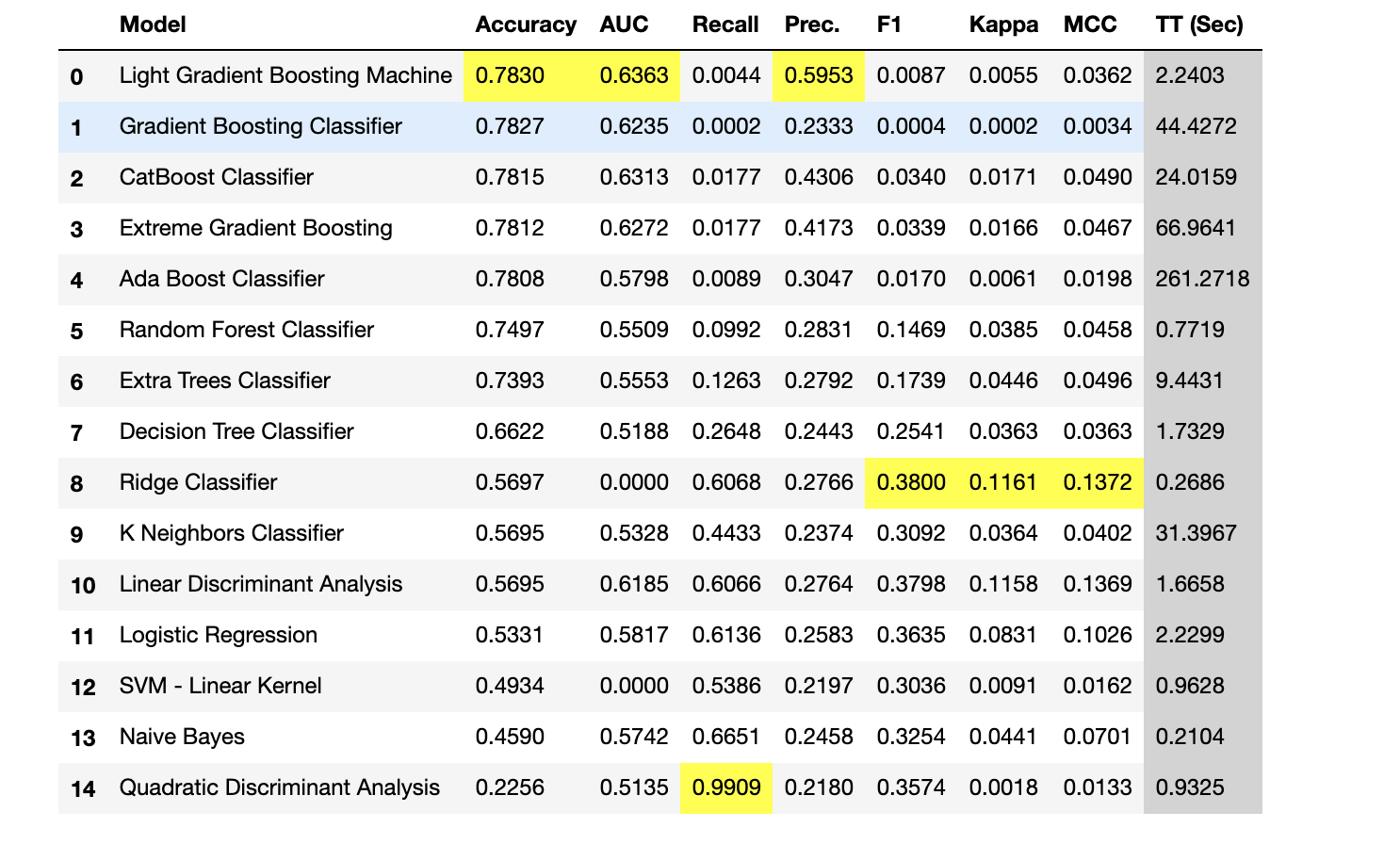


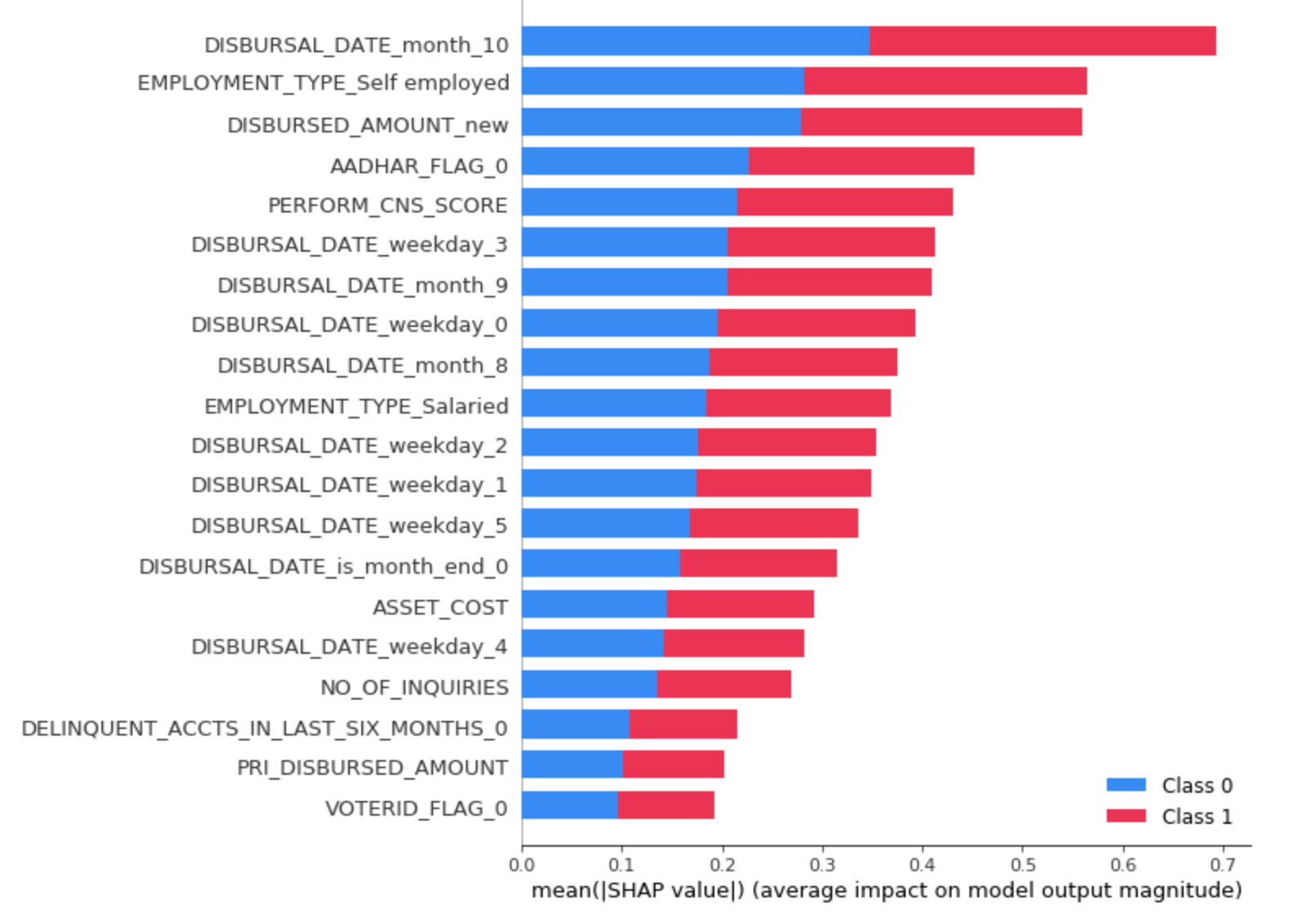
Figure 5: Correlation Matrix of selected features.

**Data Modeling:**

PyCaret classification module was used. This is a supervised machine learning module which is used for classifying the elements into a binary group based on various techniques and algorithms. The below table depicts various classification models and techniques. Using the below as a starting point I want to examine and tune some of the more effective models and find the most important features.

****

Initial impression is that the model is okay. The Light Gradient Boosting Machine models has better at maximizing the Accuracy and AUC. The Accuracy is 78.3% which is best among all the models tested. Similarly, AUC is also 63.2% which is highest among others. To evaluate the model, we also need to look at the other plots.



The above figure displays the SHAP values for the Light Gradient Boosting Machine model. The SHAP values for the various features show which features continue the most to the model positively or negatively. To read a SHAP value chart. Values to the right of 0 on the X-axis contribute positively to the target variable and values to the left contribute negatively to the target. Values that are pink have a higher impact on the target variable than values that are blue. So as loan default increases it positively affects default cases.

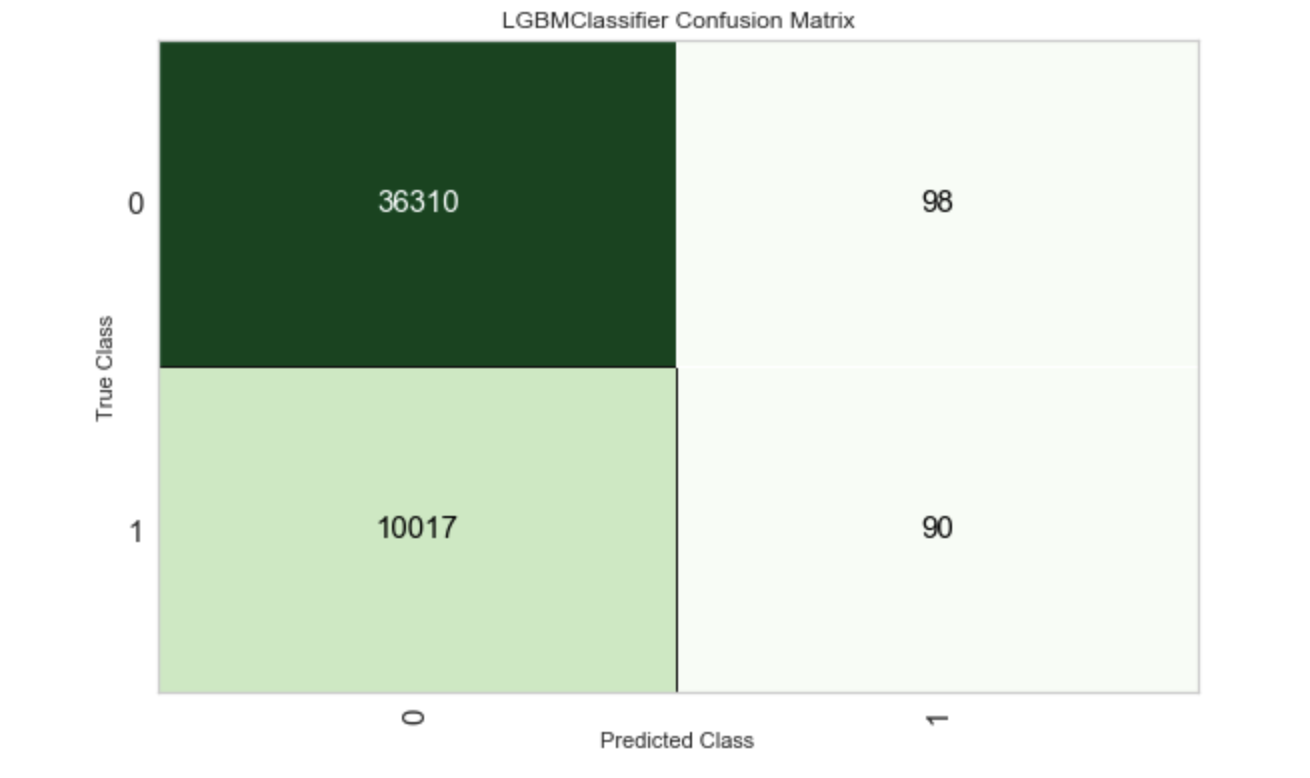


Figure 7: Confusion Matrix of Light Gradient Boosting Machine model.

The above confusion matrix shows that 90 of them are true positive (loan default), 36310 are true negative (not a loan default) cases. Only 98 cases were found False negative. But there is a very high false positive rate of this model. But overall accuracy of this model was 78.3% which is acceptable.

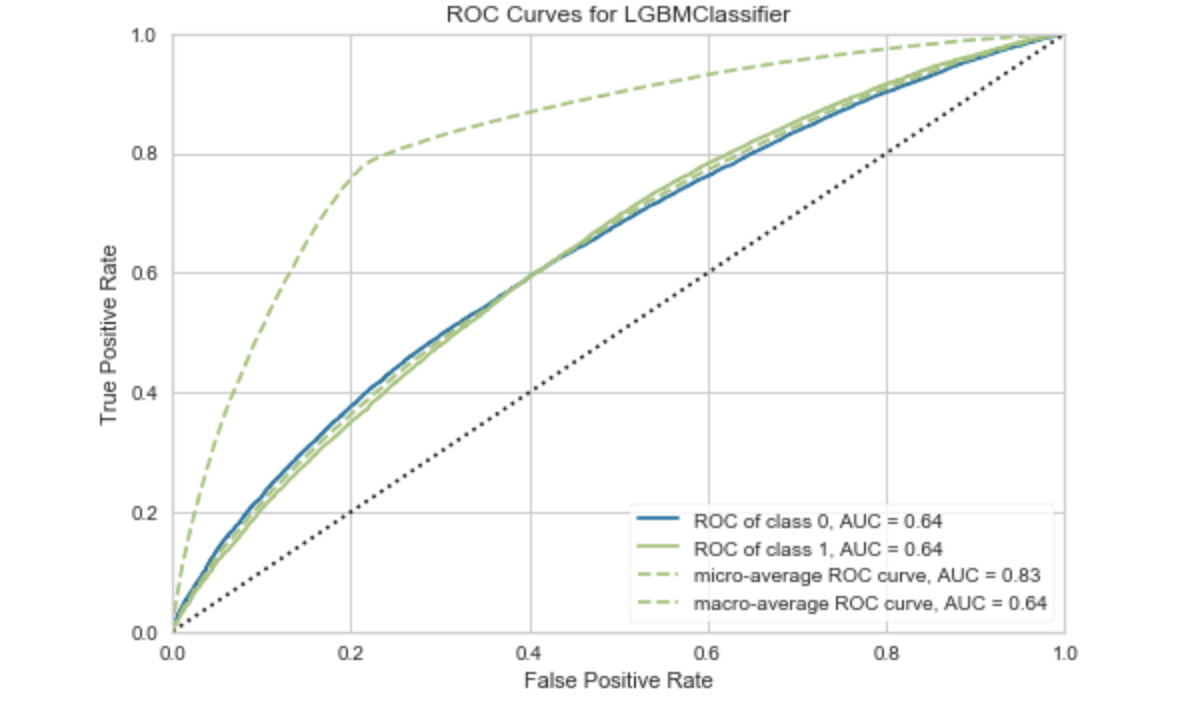


Figure 8: AUC curve of Light Gradient Boosting Machine model

The area under the curve for this model was 63.36% which looks okay for this type of dataset.

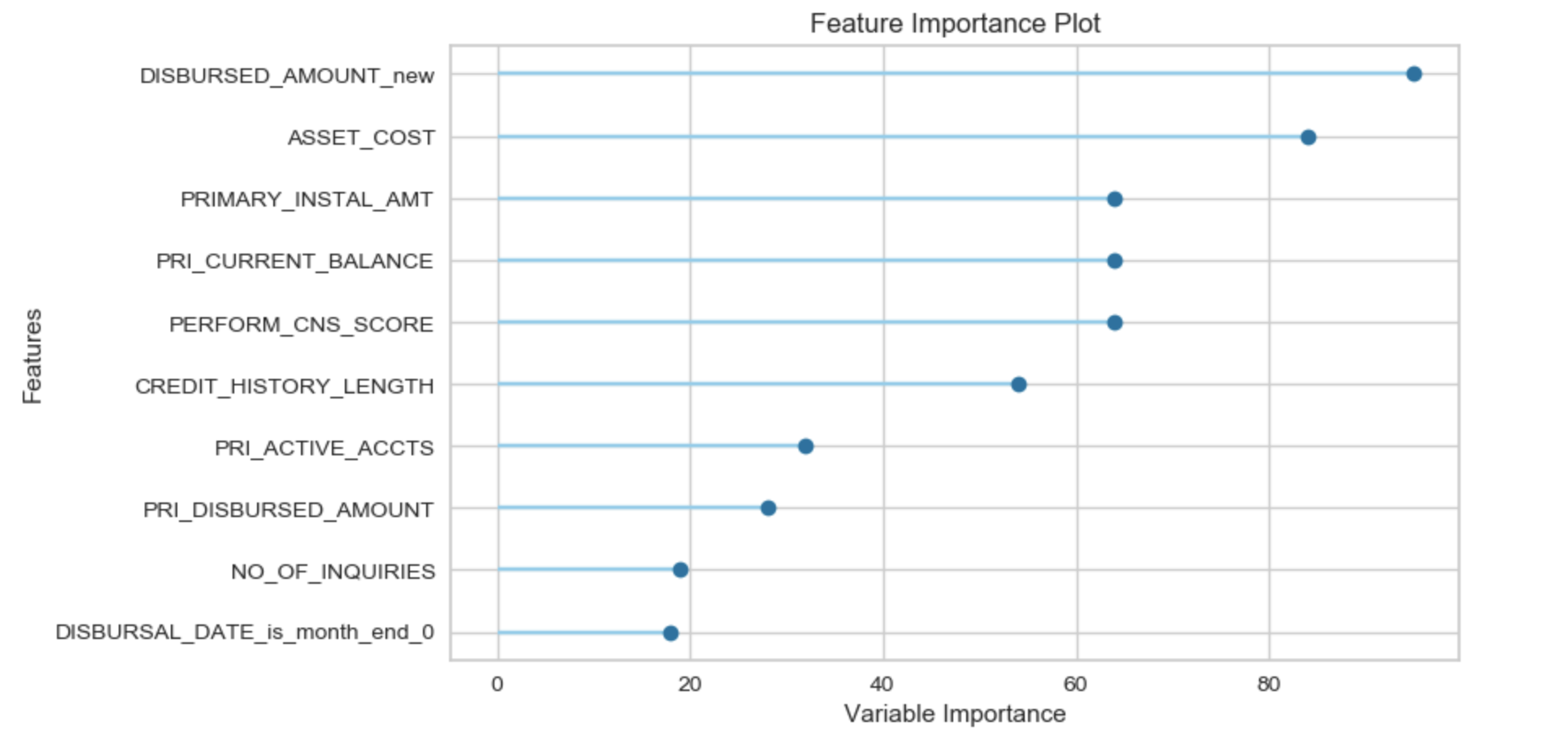


Figure 7: Feature importance of Light Gradient Boosting Machine model

Regardless of other features, the same features impact the model.

**Conclusions:**

With the Accuracy 78.7% and AUC 68% shows that this model is okay for prediction of loan default use cases. The most important features which impact the model are Disbursed amount and asset cost to the loanee.

**Future Research:**

I feel that the most important feature of the dataset is the imbalance in outcomes and how we can deal with that. I think it would be interesting to inspect the records that cause the most error. Different machine learning architecture such as neural networks could provide better results. Deep learning can be another way to discover and push the boundaries of what is possible in predicting the outcome of loans. We could as well ensemble models to make better predictions.

On a side note, relaxing the strict requirements could help capture the false negative market and that could lead to better profitability to the bank that finds a way. This is on the business side and not on the machine learning side, but nonetheless, business problems are a driver in the implementation and deployment of machine learning models. Being able to serve this under-served niche would open opportunities to those customers and potentially increase the profit margin.