# Bank Loan Status Prediction

*Ankit Kumar, Arbaaz Khan, Mayank Gupta*
*Indraprastha Institute of Information Technology, Delhi*

## Introduction

The loan is quite possibly the main results of the financial foundations. Every one of the foundations are attempting to sort out viable business systems to convince more clients to apply their loans. In any case, there are a few clients who can't take care of the loan after their applications are endorsed.Different Financial establishments consider a few factors when affirming a loan.

Determining whether a given borrower will fully pay off the loan or cause it to be charged off (not fully pay off the loan) is difficult.

The dataset used for training the model is taken from Kaggle. It includes 1,00,000 observations with many different features.

The logistic regression is widely used to solve the classification problem. However, when the relationship between variables is not linear this might not be a very good approach. Hence, we are using many different models to train the dataset and obtain better results from them.

## Related Work

We followed a very similar work done by Chang Han in the same domai. However, the dataset used by his for the training purposes is completely different. The dataset he used had approximately 10,000 samples only, while our dataset has 1,00,000 samples. Many of the features are also different in the two datasets.

## Dataset and Evaluation

The used dataset is obtained from the Kaggle. The following dataset has 1,00,000 samples with 18 features initially but the dataset contains many null values so we take the mean of that attribute for filling that null values.

## Progress

Initially, our dataset contained many null values, so we first took care of those null values by dropping some of them and filling some of them using their mean and mode of that data on that attribute.

After it we are checking the Correlation of all the pairs of attributes. We found that some values of the Correlation of pairs are close to 1, which means we don't need these attributes in our dataset because these attributes will not affect our predicted values. Hence, we remove these attributes from our dataset.

After removing unnecessary attributes, we use the one-hot encoding technique on string-valued attributes for making it a good dataset. It will help us apply ML models to the dataset then will use label encoding technique to make labels into 0 and 1 in our target attribute.

After doing one-hot encoding, our dataset attribute increased to 45.

After preprocessing and data analysis for the training set, we used 80% of the dataset and the other 20% was used for testing purposes. The training set is further divided into validation using 20% of the training dataset.

After splitting the dataset into a test and train set, we scaled all the train and test data for normalization. Then we use the LogisticRegression on a 5-fold cross-validation classifier to predict the result.

## Results and Analysis

For handling Null values in case of attribute Years in current job (datatype String) we can not use mean of for filling null values so will use mode by plotting its countplot for checking the frequency of each unique data of that column.

As we can see in **Fig 1** “10+ years” has far more frequency count than others that’s why we have used this to fill null values in that column years in our current job.

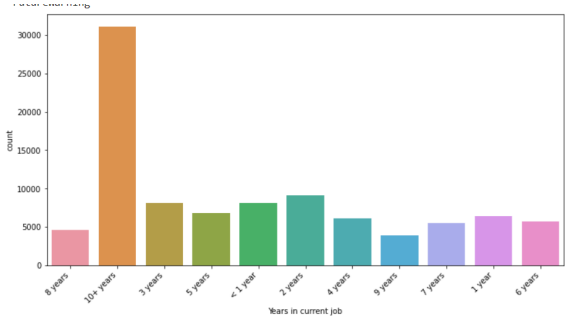


**Fig 1 Count Plot of Years in current Job**

For others attributes like “credit score”, “Annual Income” and “Bankruptcies” we have used mean value approaches for filing null values. And for some

attributes like “Maximum open credit” and “Tax liens” we have dropped their null values rows as they were less than 1% of the dataset.

And for the column “Month since last delinquent” we have dropped this column as it has more than 50% null values in it.

After handling Null values we have checked correlation between different variables present by plotting heatmap using seaborn library as given in **Fig 2.**

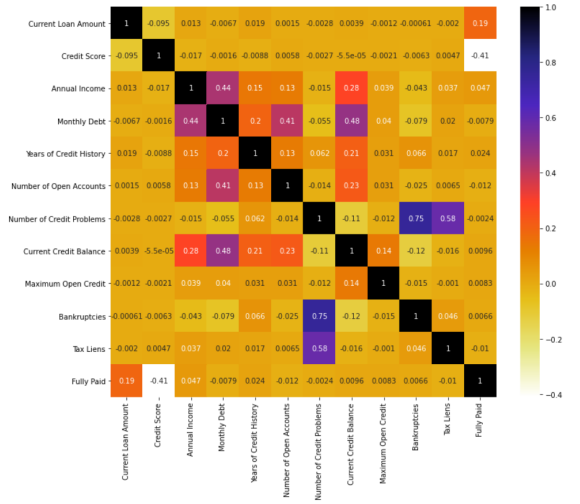


**Fig 2 Correlation Heatmap**

As we can see in the Fig 2 heatmap Number of credit problems has high correlation with both Bankruptcies and Tax liens so dropping that column (Number of credit problems) does not gonna affect our predicted result after training.

We have made seaborn Pair-plot in **Fig 3** for checking the data distribution between different attributes so that we can decide whether we have to normalize the data by scaling or not because standardization of data is very important requirement for machine learning models to work accurately for improving the accuracy of model and as from Fig 3 we can see the data is highly concentrated at 0 or less than 1. So, before training we have to normalize it and we can’t use an overfitting method.

Fig 4 contains the confusion matrix we obtained for the test set using logistic regression approach. These values are used to calculate the Recall, precision, support and the F-1 score.

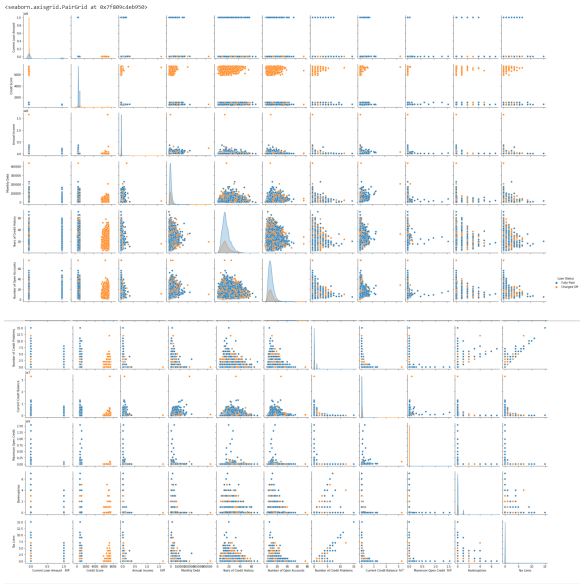

**Fig 3 Pair-Plot between every attributes**

These are the results that we generated using Logistic regression model (base model). In the future we plan to work on more advanced models to improve the accuracy of the system.

As we can see in **Fig 4** False negative value is very small, due to this we are getting a high recall of 1 because we know the formula of computing recall is **TP/(TP+FN)** so let's put the value in the formula.

Recall = 15413/(12+15413) which is equal to 0.9992.

The recall is very high which indicates that our model can predict 1 accurately.

for fully paid or label 1 in **Fig 5**.

As in **Fig 4** False negative are .06% which is negligible that’s why precision of label 0 comes nearly .99 (nearly 1).

**Precision= TN/TN+FN (for label = 0)**

919/919+12=0.99

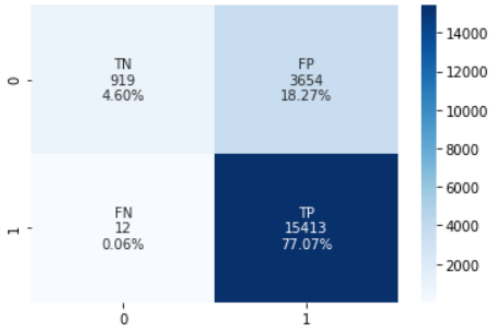


**Fig 4 Heatmap of Confusion matrix**

Classification Report For logistic Regression (Base model that we have used)
Accuracy using LR model: 0.8166816681668166
Mean Accuracy using 5-cross validation on LR model: 0.8197024628078509

| | precision | recall | f1-score | support | AUC_score |
|---|---|---|---|---|---|
| 0 | 0.99 | 0.20 | 0.33 | 4573 | 0.3999078943031485 |
| 1 | 0.81 | 1.00 | 0.89 | 15425 | 0.6000921056968515 |

**Fig 5 some evaluation metric**

And F1 scores can simply be calculated using formula.

**F1 Score = 2*(Recall * Precision) / (Recall + Precision)**

As from **Fig 5** AUC_SCORE for label 1 is higher than Label 0 it's mean Logistic regression model has high performance rate for predicting label 1 (Full paid)

## Future Work

Will use a 3- advanced ML model so that we can get some better results and then we will analyze the result and compare it with our previous one.

Models that we will use in future with gradient descent, decision tree etc approaches are Random Forest , Neural Network and gradient boosting classification.

Then will analyze those results and will do hyperparameter tuning for getting optimal parameters like loss function, n_estimators, max_depth depending on the attributes of the model.

Here we will implement random search with cross validation to select the optimal hyperparameters for models.

Working code(ipynb), dataset, ppt and report

**By Ankit**: Preprocessing, Base model working, Neural network implementation, Data analysis.

**By Arbaaz**: PPT ,Random forest model implementation, Data analysis.

**By Mayank**: Report, Preprocessing, gradient boosting classifier, analysis of different methods, Data analysis.

## Reference

**https://escholarship.org/content/qt9cc4t85b/qt9cc4t85b.pdf?t=pszej2&v=lg**