Bank Loan Status Prediction

CBCPC22

PROJECT-REPORT

2023-24

Netaji Subhas University of Technology

Sector 3, Dwarka,

New Delhi-110078

Submitted By:

Vikas Kumar **2020UCB6012**

Gautam Kumar Singh **2020UCB6048**

Ankur Meena **2020UCB6054**

Under the guidance of: Prof. Manoj Kumar

Table Of Content

Bank Loan Status Prediction

Ankur Meena, Gautam Kumar Singh, Vikas Kumar

(Under the guidance of Prof. Manoj Kumar)

Netaji Subhash University Of Technology, Delhi

Abstract

The loan is quite possibly the main result of the financial foundations. Every one of the foundations is attempting to sort out viable business systems to convince more clients to apply for their loans. In any case, a few clients 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.

That's why we are making a Machine learning model to predict who can provide a loan or not based on property or attributes related to the finances of that particular person or user. This model is for the retailer who acquires a loan for purchasing inventory, expanding the store, marketing, side businesses, in an emergency, or covering other operational expenses

The dataset we used for model training is obtained from Kaggle. It includes 1,00,000 observations with many different features. However, due to some technical limitations, we have considered only 30,000 observations.

Logistic regression is a statistical method used for binary classification. However, when the relationship between variables is not linear this might not be a good approach. Hence, we are using many different models to train the dataset and obtain better results such as using **random forest, support vector machine (SVM), and voting classifier** with their prediction accuracy..

Introduction

In the dynamic landscape of the financial sector, the efficient and accurate evaluation of loan applications stands as a critical facet influencing the stability and growth of banking institutions. With an ever-increasing volume of loan applications, the need for robust, **data-driven decision**-making processes has become important. The project, "Bank Loan Status Prediction," represents a comprehensive exploration into the realm of machine learning (ML) to enhance and automate the loan approval system.

By leveraging historical loan data encompassing a diverse range of applicant information, financial metrics, and corresponding loan outcomes, this project seeks to **develop a predictive model capable of discerning patterns and making informed decisions.**

Furthermore, This project provides insights into the key factors influencing loan approval decisions. This transparency is essential for building trust with stakeholders, including financial institutions and loan applicants.

Motivation

The conventional approaches to evaluating loan applications are often time-consuming and may not fully leverage the wealth of information available in modern financial datasets.

The traditional credit scoring models often rely on a limited set of factors, potentially overlooking valuable information that can significantly contribute to the prediction of loan outcomes. By employing machine learning algorithms, we aspire **to develop a predictive model that considers a multitude of features, including historical financial data** and other relevant variables. This comprehensive approach seeks to provide a more holistic and nuanced understanding of an applicant's creditworthiness.

The outcomes of this project have far-reaching implications for both financial institutions and applicants. For banks, a more accurate loan prediction model can lead to better **risk management, reduced default rates, and increased profitability.**

Literature Survey

- Loan status prediction is a challenging task that has been the subject of much research in recent years. Machine learning algorithms offer a promising approach to this problem, as they can be trained on large datasets of historical loan data to learn the factors that are most predictive of loan status.

A variety of machine learning algorithms and techniques have been used for loan status prediction. Some of the most popular algorithms include:

- **Logistic regression:** Logistic regression is a simple but effective algorithm for **binary classification tasks**. It can be used to predict the probability that a loan will be approved or defaulted on.

- **Random forest:** A Random Forest is an **ensemble learning technique** used for both classification and regression tasks. It belongs to the family of ensemble models, which **combine the predictions of multiple individual models** to improve overall performance

- **Support Vector Machines:** Primarily, it is used for Classification problems. The **primary goal of SVM is to find a hyperplane in an N-dimensional space** (where N is the number of features) that best separates the data into different classes.

- **Voting Classifier:** A Voting Classifier is an ensemble learning method in machine learning that combines the predictions of multiple individual models to make a final prediction. It aggregates the results of multiple classifiers **and predicts the class label based on a majority vote (for classification tasks)** or an average (for regression tasks).

Related Work

- We followed a very similar work done by Chang Han in the same domain. However, the dataset used by him for the training purposes is completely different. The dataset he used had approximately 10,000 samples, while **our dataset had 30,000 samples**. Many of the features are also different in the two datasets.

Dataset

- The used dataset is obtained from the Kaggle. The following dataset has 30,000 samples with 18 features initially.
- After preprocessing and data analysis for the training set, we used 80% of the dataset for training and the other 20% is used for testing purposes.
- We came across some of the main features in the dataset: Credit Score, annual income, and current credit balance.
- Evaluation metrics that we will use: Accuracy, AUC Score, Precision, Recall and F1 Score

Problem Statement

- In the banking system, the most important issue is the approval of a loan to anyone and its return collection.
- Many of the banks face financial losses due to not recollecting the approved loan from individuals at the proper time.
- That’s why it has become very important to predict to which person the bank should approve the loan and to which person they should deny.
- Provide more accuracy than previous models of loan prediction.

Objective

- The primary objective of the project, "Bank Loan Status Prediction using Machine Learning," is to leverage advanced machine learning techniques such as logistic regression, random forest, and support vector machines to develop a robust and accurate predictive model for determining the approval or denial of loan applications.
- **This project aims to address the complexities and challenges associated with the traditional methods of assessing creditworthiness by harnessing the power of data-driven decision-making.**

Methodology

⇨ **Following steps representing project implementations**

- Initiation, from the dataset containing 30,000 rows and 18 columns.
- Preprocessing the data
    - Removing unrequired columns such as months since last delinquent, loan ID, and customers.
    - Removing rows containing null values for not mandatory columns
    - Filling mean values for mandatory columns such as credit score, annual income, etc.
    - Some of the null values are replaced with majority values of columns such as year in current job based on analysis
- To analyze the columns corresponding to output, we perform a seaborn plot.
- We also implement the Correlation of all the pairs of attributes to find the relation within columns. So, if two columns are too closely related then we will consider only one of them.
- we use the one-hot encoding technique on string-valued attributes to make it a good dataset. It will help us apply ML models to the dataset then will use the label encoding technique to make labels into 0 and 1 in our target attribute.
- After all work on the dataset, we move further partitioning data into two ratios of 80%-20% to be used as training and testing models.
- After splitting the dataset into a test and train set, we scaled all the train and test data for normalization.
- We have used various models for our project such as LogisticRegression, SVM, RandomForest, and VotingClassifier then based on their evaluation(precision, recall, f1-score,roc-auc-score) considering the model that provides better accuracy for our project.
- Then we also implement it on website where we fill the details and then based on training of these models we get the outputs.

Implementation, Results, and Discussion

- Initially, we have columns based on our datasets such as Loan ID, Customer ID, Loan Status, Current Loan Amount, Term Credit Score, Annual Income,Years in current job, Home Ownership, Purpose, Monthly Debt, Years of Credit History, Months since last delinquent, Number of Open Accounts, Number of Credit Problems, Current Credit Balance, Maximum Open Credit, Bankruptcies and Tax Liens. Considering a dataset of 30,000 observations.
- Handling null values, and performing various operations to get only necessary data according to our needs.

```
Loan ID                         0
Customer ID                     0
Loan Status                     0
Current Loan Amount             0
Term                            0
Credit Score                 5818
Annual Income                5818
Years in current job         1263
Home Ownership                  0
Purpose                         0
Monthly Debt                    0
Years of Credit History         0
Months since last delinquent    16080
Number of Open Accounts         0
Number of Credit Problems       0
Current Credit Balance          0
Maximum Open Credit             0
Bankruptcies                   61
Tax Liens                       6
```

- And for the column "Month since last delinquent" we have dropped this column as it has more than 50% null values in it.
- For other attributes like "credit score", "Annual Income" and "Bankruptcies" we have used mean value approaches for filing null values so that we get the most accurate output. 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.
- For the attribute "Years in the current job" (datatype String), we can not use the mean for filling null values so will use mode by plotting its count plot 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.

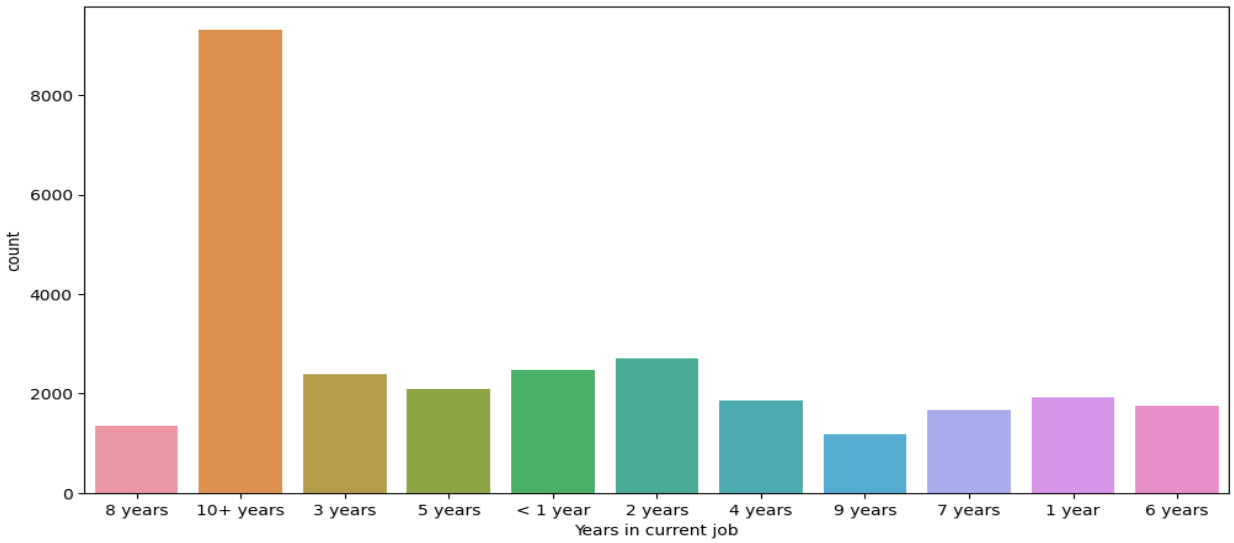


Fig 1 Count Plot of Years in Current Job

- we checked the correlation between different variables present by plotting the heatmap using the seaborn library as given in **Fig 2.**

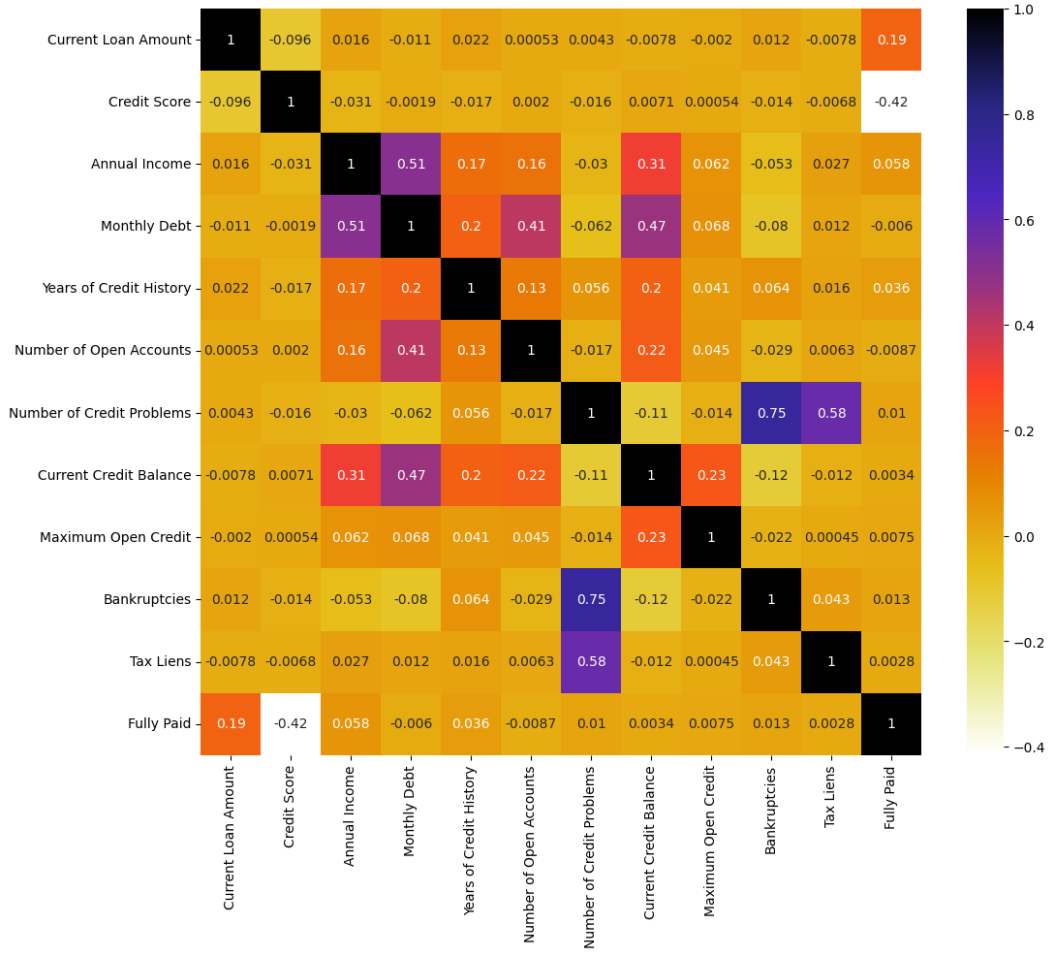
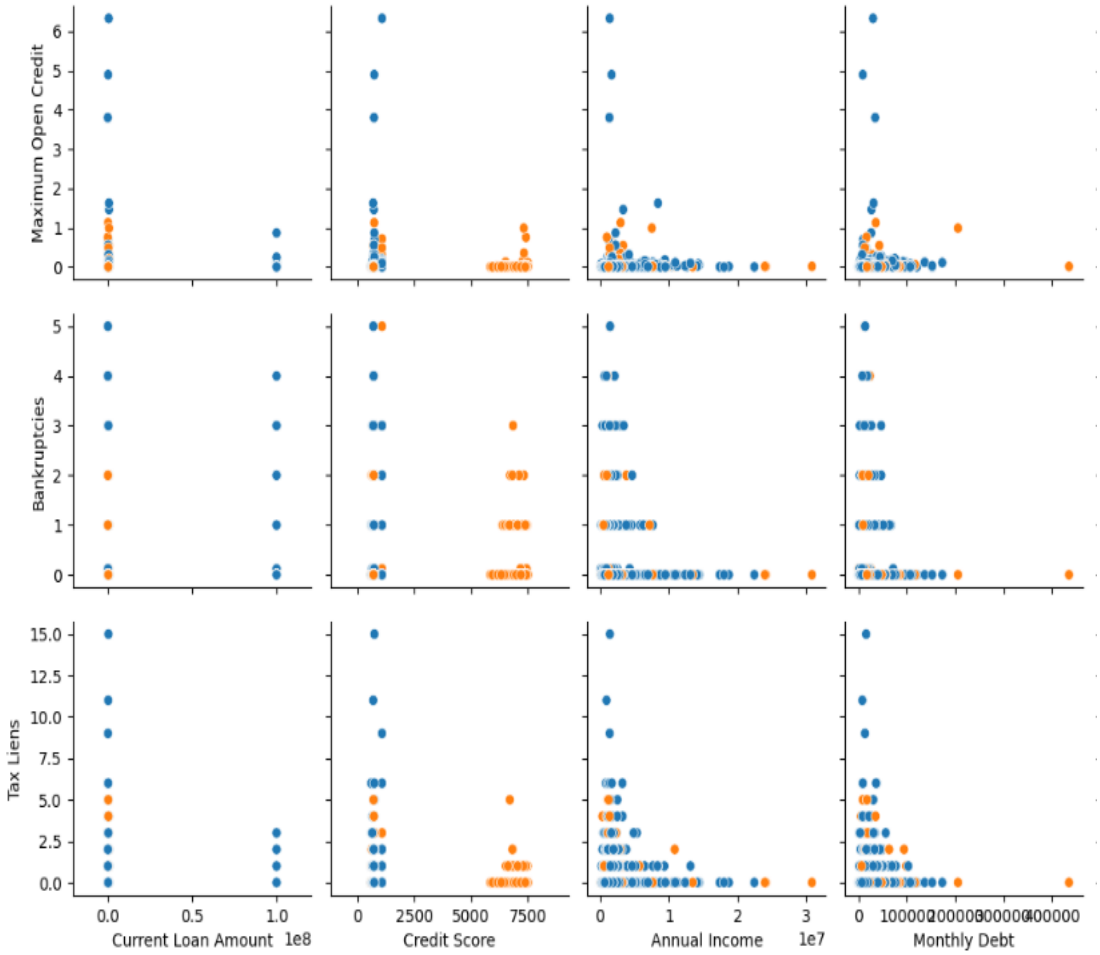


Fig 2 Correlation Heatmap

- As we can see in the Fig 2 heatmap Number of credit problems has a 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 a 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 a very important requirement for machine learning models to work accurately for improving the accuracy of the 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.
- Below the figure is the zoom view of the seaborn Pair-plot.



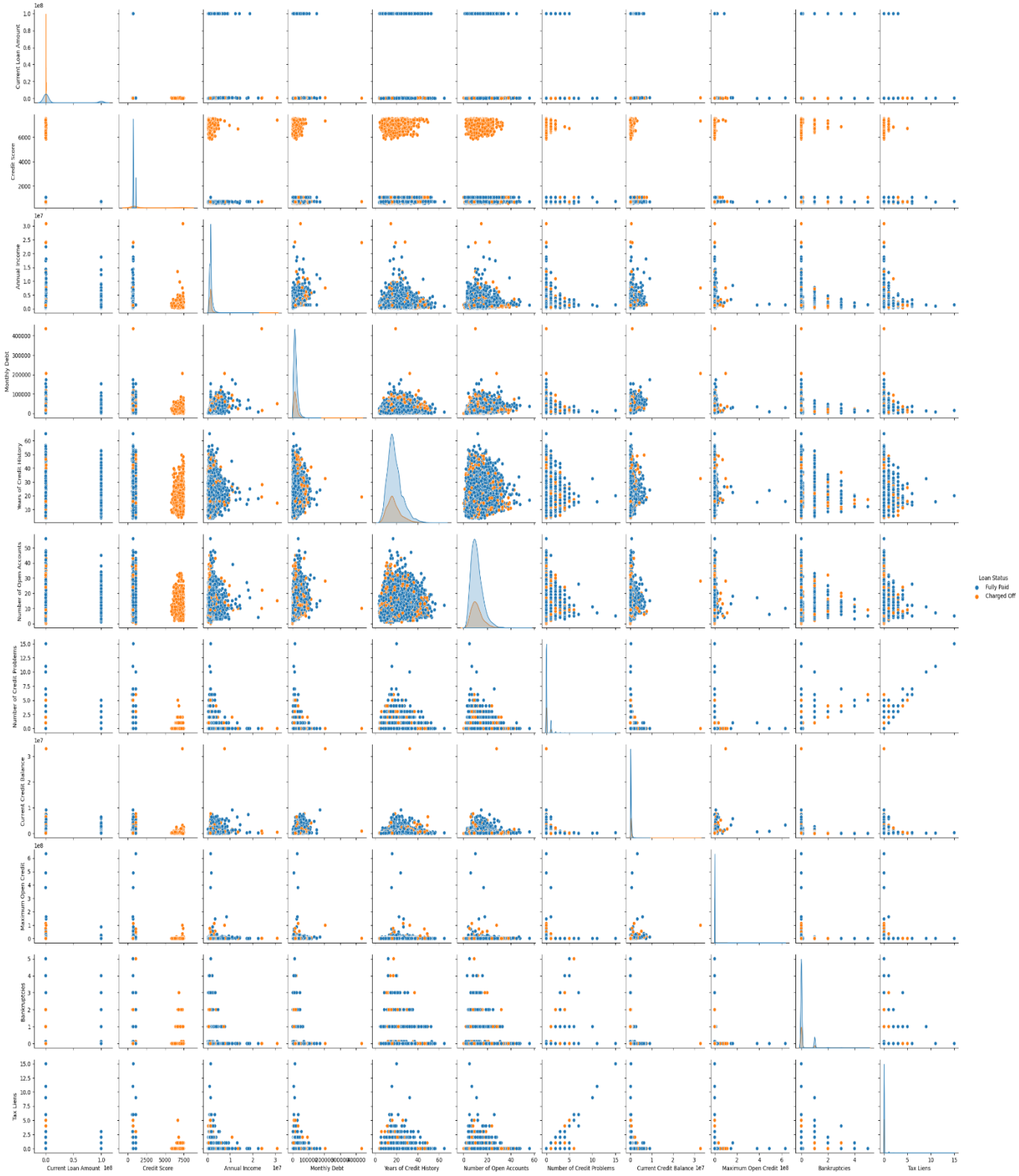


Fig 3 Pair-Plot between every attribute

- we use the one-hot encoding technique on string-valued attributes to make it a good dataset. It will help us apply ML models to the dataset then will use the label encoding technique to make labels into 0 and 1 in our target attribute. That includes breaking of attributes into multiple components. The final attributes are:
- ['Current Loan Amount', 'Credit Score', 'Annual Income', 'Monthly Debt', 'Years of Credit History', 'Number of Open Accounts', 'Current Credit Balance', 'Maximum Open Credit', 'Bankruptcies', 'Tax Liens', ' Term_Long Term', 'Term_Short Term', 'Years in current job_1 year', 'Years in current job_10+ years', 'Years in current job_2 years', 'Years in current job_3 years', 'Years in current job_4 years', 'Years in current job_5 years', 'Years in current job_6 years', 'Years in current job_7 years', 'Years in current job_8 years' 'Years in current job_9 years', 'Years in current job_< 1 year', 'Home Ownership_HaveMortgage', 'Home Ownership_Home Mortgage', 'Home Ownership_Own Home', 'Home Ownership_Rent', 'Purpose_Business Loan' 'Purpose_Buy House' 'Purpose_Buy a Car', 'Purpose_Debt Consolidation', 'Purpose_Educational Expenses', 'Purpose_Home Improvements', 'Purpose_Medical Bills', 'Purpose_Other', 'Purpose_Take a Trip', 'Purpose_major_purchase', 'Purpose_moving', 'Purpose_other' ,'Purpose_renewable_energy', 'Purpose_small_business', 'Purpose_vacation', 'Purpose_wedding']

- We have used various model and doing their evaluation below.
- the confusion matrix, we obtained for the test set using the logistic regression, SVM, RandomForest, and VotingClassifier approaches. These values are used to calculate the Recall, precision, support, and the F-1 score.
- Explaining for logistic regression, confusion matrix, 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). The recall is very high which indicates that our model can predict 1 accurately.
  F1 scores can simply be calculated using a formula.
  F1 Score = 2*(Recall * Precision) / (Recall +Precision).
  AUC_SCORE for label 1 is higher than Label 0 which mean the Logistic regression model has a high performance rate for predicting label 1 (Fully paid).

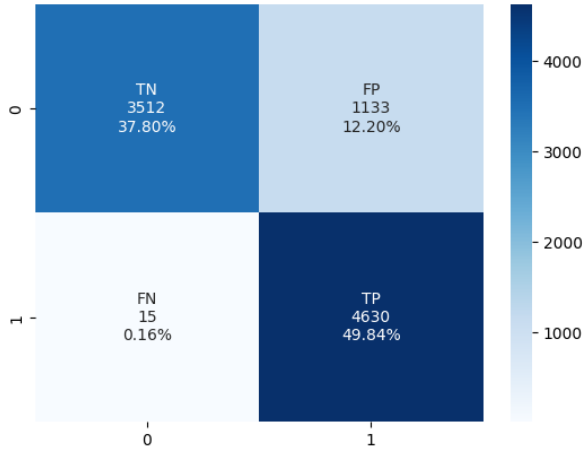
- **Results of logistic Regression**

score using logistic regression(base model) 0.8764262648008612

	precision	recall	f1-score	support
0	1.00	0.76	0.86	4645
1	0.80	1.00	0.89	4645
accuracy			0.88	9290
macro avg	0.90	0.88	0.87	9290
weighted avg	0.90	0.88	0.87	9290

ROC-AUC-Score 0.8764262648008612

some evaluation metric



Heatmap of Confusion matrix

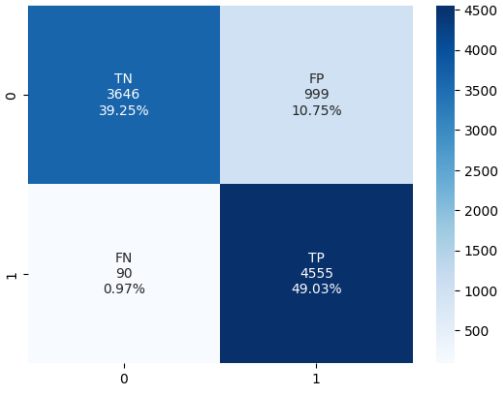
- **Result of Random Forest Classifier**

score for RandomForestClassifier 0.8827771797631863

	precision	recall	f1-score	support
0	0.98	0.78	0.87	4645
1	0.82	0.98	0.89	4645
accuracy			0.88	9290
macro avg	0.90	0.88	0.88	9290
weighted avg	0.90	0.88	0.88	9290

ROC-AUC-Score 0.8827771797631861

some evaluation metric



Heatmap of Confusion matrix

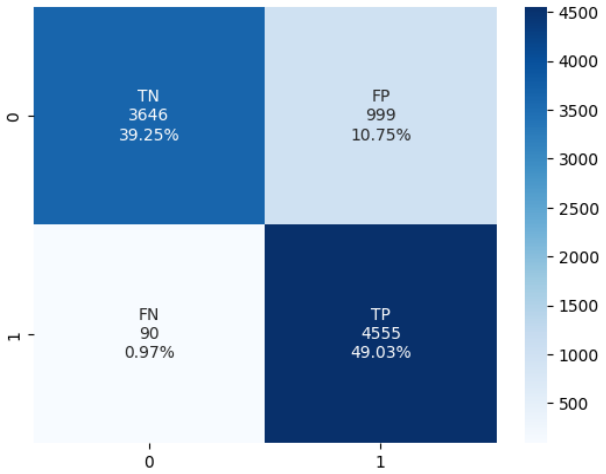
- **Result of SVM**

score for SVM classifier 0.8748116254036599

	precision	recall	f1-score	support
0	1.00	0.75	0.86	4645
1	0.80	1.00	0.89	4645
accuracy			0.87	9290
macro avg	0.90	0.87	0.87	9290
weighted avg	0.90	0.87	0.87	9290

ROC-AUC-Score 0.87481162540366

some evaluation metric



Heatmap of Confusion matrix

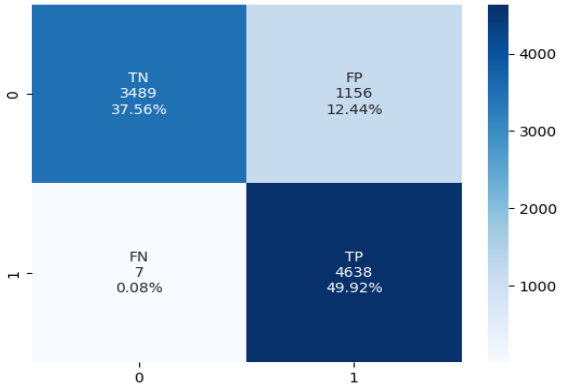
- **Result of VotingClassifier**

score for soft_VotingClassifier 0.8848116254036599

	precision	recall	f1-score	support
0	1.00	0.76	0.86	4645
1	0.80	1.00	0.89	4645
accuracy			0.88	9290
macro avg	0.90	0.88	0.88	9290
weighted avg	0.90	0.88	0.88	9290

ROC-AUC-Score 0.884811625403543

some evaluation metric



Heatmap of the Confusion matrix

⇨ We have also passed the dataset for testing all the models and get the result for each model whether the loan is approved(1) or rejected(0) based on observations.

Real Life implementation

Bank Loan Status Prediction

Current Loan Amount

445412.00 − +

Credit Score

709.00 − +

Annual Income

1167493.00 − +

Monthly Debt

5214.74 − +

Years of Credit History

17.12 − +

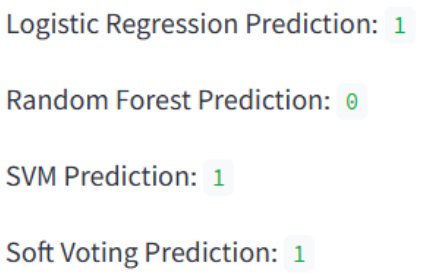
Number of Open Accounts

0.99 − +

Current Credit Balance

228189.94 − +

- In actual calculation loan for the above data is approved and the result of our models are shown below



These are the results that we generated using the pair plot. In the future, we will decide which model will give the best result by analyzing these pair-plot graphs.

Conclusion and Future Work

- we will analyze the result and compare it with our previous one with few more advance model.
- Models that we will use in the future with gradient descent, decision tree, etc approaches are Neural Network, and gradient boosting classification.

- Then will analyze those results and will do hyperparameter tuning to get optimal parameters like loss function, n_estimators, and max_depth depending on the attributes of the model.
- Here we will implement a random search with cross-validation to select the optimal hyperparameters for models.

Reference

- https://ijarsct.co.in/Paper1165.pdf
- Machine-Learning-Projects-Code/Loan Status Prediction Using Machine Learning.ipynb at main · DataThinkers/Machine-Learning-Projects-Code · GitHub
- Predicting Loan Approval Status (alexforrest.github.io)
- Loan Approval Prediction using Machine Learning – GeeksforGeeks
- Loan Status Prediction | Kaggle