**Project Title**

**Loan Defaulters Model**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Team 4**

**Members:**

**Ambili CV**

**Arathi R Nair**

**Arravind S R**

**Easwer B I**

**Gautham Mohan R**

**Renuka Korisseril Chandran**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA, Apr 2023**

**List of Figures**

**Chapter 4**

Figure 4.1 Variables by type 15

Figure 4.2 Checking missing values 16

Figure 4.3 Defaulter Vs Non Defaulter 17

Figure 4.4 Univariate Analysis 18

Figure 4.5 Univariate Analysis Observations 19

Figure 4.6 Bivariate Analysis- Term 21

Figure 4.7 Bivariate Analysis- Grade 21

Figure 4.8 Bivariate Analysis - Home Ownership 21

Figure 4.9 Bivariate Analysis - Revolving Balance 21

Figure 4.10 Bivariate Analysis 22

Figure 4.11 Bivariate Analysis 23

Figure 4.12 Multivariate Analysis 24

**Chapter 5**

Figure 5.1 Label encoding Data 25

Figure 5.2 Handling Outliers 27

**Chapter 6**

Figure 6.1 Feature Scaling 28

Figure 6.2 Feature Reduction 29

**Chapter 7**

Figure 7.1 Model Building and Model Selection 30

Figure 7.2 Oversampling 31

Figure 7.3 Confusion Matrix 31

**CHAPTER 8**

Figure 8.1 Randomized Search CV 32

Figure 8.2 Grid Search CV 33

Figure 8.3 Pickling 34

**CHAPTER 9**

Figure 9.1 Result 35

**List of Abbreviations**

| np | NUMPY |
| --- | --- |
| pd | PANDAS |
| sns | SEABORN |
| le | LABEL ENCODER |
| Vif | VARIANCE INFLATION FACTOR |
| Skf | STRATIFIED K FOLD |
| rf | RANDOM FOREST |
| ML | MACHINE LEARNING |
| svm | SUPPORT VECTOR MACHINE |

**Table of Contents**

[**CHAPTER 1 7**](#_5ebb56cc96dy)

[**PROBLEM DEFINITION**](#_ccwd9j46y3bx)

[**CHAPTER**](#_5ebb56cc96dy) **2** [**7**](#_ccwd9j46y3bx)

[**INTRODUCTION**](#_iqop852p9rxc) **9**

[**CHAPTER 3 10**](#_k8om28d2z10b)

[**LITERATURE SURVEY**](#_z2of96csfw4x) **10**

[3.1 Prediction and Analysis of Financial Default Loan Behaviour Based on Machine Learning Model, Published by Herui Chen (2022)](#_14ms5vb7a3io) **10**

[3.2 Machine Learning Approaches to Predict Loan Default, Published by Wanjun Wu Bytedance Data Analysis Group, Beijing, China in Scientific Research Publisher (2022)](#_pv3alr4kd34d) **11**

[3. 3 A Naive Bayes approach to fraud prediction in loan default, Published by I O Eweoya & A A Adebiyi, Covenant University, Nigeria (2019)](#_s066qmm6onbg) **11**

[3.4 Loan default prediction using decision trees and random forest: A comparative study by Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath, Volume 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India 12](#_xfg9oca0dwt1)

[3.5 Predictions of Loan Defaulter - A Data Science Perspective by P. Maheswari, CH. V. Narayana,Published in: 2020 5th International Conference on Computing, Communication and Security (ICCCS) 12](#_r25poo6sdd8q)

[3.6 A study on predicting loan default based on the random forest algorithm by Lin Zhu,Dafeng Qiu,Daji Ergu(2019), Southwest Minzu university 13](#_n0q6p75xs3cy)

[**CHAPTER 4 13**](#_la3p831wet2h)

[**Dataset And Exploratory Data Analysis 14**](#_bg76q7hi6q4j)

[4.2 Exploratory Data Analysis 14](#_p7q56dfbv57d)

[4.2.1 Univariate Analysis 17](#_twbe4kaq1aj8)

[**CHAPTER 5 24**](#_n4w0ocxgcyhi)

[**Feature Engineering 24**](#_ebfnllfsndfn)

[5.1.2 One-Hot Encoding 24](#_9nfjy5wf8bij)

[**CHAPTER 6 27**](#_i00srcci9v89)

[**Feature Scaling and Feature Reduction 27**](#_abjocrvbgzlb)

[6.1 Feature scaling 27](#_3fin7761go56)

[6.2 Feature Reduction 28](#_re2egrr63v7)

**CHAPTER 7 30**

7.1 Oversampling 31

7.2 Confusion Matrix 32

**CHAPTER 8 33**

8.1 Randomized Search CV 33

8.2 Grid Search CV 34

8.3 Pickling 34

**9**[**. Result**](#_x4pyp9jil9p6) **35**

**10**[**. Conclusion 3**](#_r3z9zj6hsez4)**7**

[**References 3**](#_1cfd5ym60pys)**8**

**Abstract**

The main income earning assets for a bank are loans. At the end of 2014, loans accounted for 52.64% of total assets that banks held in the US. A bank’s profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-Performing Assets.

This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. The Logistic regression model is used.

Banks run into losses when customers don't pay their loans on time. Because of this, every year, banks have losses in crores, and this also impacts the country's economic growth to a large extent. In this project, we look at various attributes such as funded amount, location, loan, balance, etc., to predict whether a person will be a loan defaulter.Loan status column is used to predict whether a user is loan defaulter. The dataset comprises around 67,463 rows and 35 columns.

We are planning to implement various Algorithms like Logistic,SVM models, KNN,Random Forest,Decision Tree etc.to conclude the best one. We evaluate the performance of the model using metrics such as precision,recall,and F1 score.Using the results from the above analysis,we can easily predict whether a user is a loan defaulter or not.

# 

# CHAPTER 1

# PROBLEM DEFINITION

**1.1 Overview**

The goal of this project is to build a machine learning model that can predict if a person will default on the loan based on the loan and personal information provided. The model is intended to be used as a reference for the financial institution to help make decisions on loan defaulters, so that the risk can be lowered, and the profit can be maximized

Our Loan\_Defaulters dataset consists of approximately 67463 samples of loans granted by banks, with the full set of information about the borrower, the history of payments and the outcome of the loan.

The dataset is quite clean and the figures can be considered as ground truth, but lots of columns are either irrelevant, very sparse or non informative.It also does not contain any null or duplicated values. Some columns with corrupted information such as Employment Duration, Accounts Delinquent are removed.

The dataset is quite clean and the figures can be considered as ground truth, but lots of columns are either irrelevant, very sparse or non informative.It also does not contain any null or duplicated values. Some columns with corrupted information such as Employment Duration, Accounts Delinquent are removed.

Fields like ID and Payment plan are having unique values so we are removing this from our analysis. Public Record is not giving relevant information hence excluded. For Deliquency\_in the last two years and Inquiries\_Last 6 months column, It is observed from our analysis that more defaulters are appearing under 0 delinquency bucket and 0 inquiry status values ,so this variable is not useful for analysis. Revolving utility is not having much importance in analysis because no difference is seen between defaulters and non defaulters. Since total current balance is high for defaulters it is not useful. Collection 12 months Medical is having a single value so we can remove this. Total Revolving Credit Limit is having no impact on target variable. Batch Enrolled, Interest Rate are personal information and regarding loan respectively so no impact on target.

Moreover, the dataset is very unbalanced, with approximately 10% of loans considered as defaulted.The Loan status column consists of values 0 and 1 which means 0 as non-defaulter and 1 as defaulter.

Correlation is a technique for investigating the relationship between two quantitative, continuous variables in order to represent their inter-dependencies. Its correlation coefficient scales from -1 to 1, where 1 represents the strongest positive correlation, -1 represents the strongest negative correlation and 0 represents no correlation. The correlation coefficients between each pair of the dataset are calculated and plotted as a heatmap. From the heatmap, it is easy to locate the highly correlated features with the help of color coding.

Since the objective is to predict the outcome from the information gathered detailing the loan details, post considering the data concerning the history of payments or the current situation of a loan, we had excluded features for which the information is incomplete, or uninformative. Finally we decided around 9 significant features like Home Ownership, Term, Open Account, Grade, Total Accounts, Last week Pay, Total Collection Amount, Revolving Balance, Total Received Late Fee, etc.

**1.2 Problem Statement**

Banks use credit scoring models to evaluate potential loan default risks.These models generate a score that translates the likelihood of defaulting on a loan, making lending decisions easier. Developing a credit scoring model is time consuming. These models are also fixed and do not easily evolve with changing customer behavior to predict default more accurately. Machine Learning approaches can help enhance the accuracy of loan default prediction. The loan default prediction is a problem of binary classification.The objective of our project is to predict whether a loan will default or not based on a given dataset .

**CHAPTER 2**

# INTRODUCTION

Loan lending plays an important role in our everyday life and powerfully promotes the growth of consumption and the economy. Taking a loan has been inevitable for people since individuals around the world depend on loans to overcome financial constraints to achieve their personal goals, and organizations rely on loans to expand their production. In most of the cases, lending a loan is beneficial to both the borrowers and the lenders. However, loan default is still unavoidable, which carries a great risk and may even end up in a financial crisis. Therefore, it is particularly important to identify whether a candidate is a defaulter or not.

Predicting the outcome of a loan is a recurrent, crucial and difficult issue in banking and financial sectors. In the past, the evaluation primarily depended on manual review, which was time-consuming and labor-intensive. Recently, banks have opted for ma­chine learning approaches to automatically predict loan default since it can highly enhance the accuracy and the efficiency of the prediction. Machine learning models are rapidly evolving and have successful applications in various fields, motivating the bank industry to use them to predict loan default.

Here we use Loan\_Defaulters dataset to predict whether a customer is a defaulter or not. We perform analysis by the help of Machine learning Model developed in python language. Here we consider this as a classification problem.The Loan status column consists of values 0 and 1 which means 0 as non-defaulter and 1 as defaulter.

We are planning to implement various Algorithms like Logistic,SVM models, KNN,Random Forest,Decision Tree etc.to conclude the best one. We evaluate the performance of the model using metrics such as precision, recall and F1 score.

# 

# CHAPTER 3

# LITERATURE SURVEY

As an important core of the modern economy, the quality and management level of bank loans directly affect the economic system. Since loans are risky transactions of financial institutions, creditors or investors will undergo different levels of risk assessment before each loan transaction to assess whether creditors will repay the loans on time. In the past, commercial banks usually used 5C classification to make subjective decisions when evaluating the credit risk of credit users. They will assess and evaluate borrowers from five aspects: personal nature, credit constraints, solvency and market conditions, and credit services provided to customers. To a great extent, the judgment result depends on the subjective evaluation of risk appraisers in the evaluation process. Due to the rapid change of market economy, this assessment method can no longer meet the needs of borrowers or the risk management requirements of commercial banks. We need to establish a scientific and effective explanation model to evaluate the reputation of credit customers, so as to minimize the default risk and maximize profits. If the global economy is to develop steadily and healthily, commercial banks must implement a scientific credit risk management system at the core of the economy. This literature survey aims to explore the different ML techniques used in Loan Defaulters prediction.

## 3.1 Prediction and Analysis of Financial Default Loan Behaviour Based on Machine Learning Model, Published by Herui Chen (2022)

The paper forecasts and judges the credit degree and decision-making loan quota of customers in banks. The proposed improved logistic regression and neural network method can achieve better prediction results in financial risk prediction. The second part of the paper describes the financial loan risk prediction algorithm, focusing on the non-balanced data and penalty regression method. The third part is the empirical analysis of commercial bank loan data. By comparing with other algorithms, the method proposed in this paper has a better effect. This paper constructs four different models of logistic regression, decision tree, conventional AdaBoost, and the proposed prediction model in the pre-processing data, and compares the prediction effect and generalization performance of the four models horizontally.

## 3.2 Machine Learning Approaches to Predict Loan Default, Published by [Wanjun Wu](https://www.scirp.org/journal/articles.aspx?searchcode=Wanjun++Wu&searchfield=authors&page=1) Bytedance Data Analysis Group, Beijing, China in Scientific Research Publisher (2022)

In this paper, observed Random Forest and XGBoost algorithms to train the prediction model and compare their performance in prediction accuracy. In the feature engineering part, I used the variance threshold method and Variance Inflation Factor method to filter out unimportant features, and then input those selected features into Random Forest and XGBoost models. It turns out that Random Forest and XGBoost show **little** difference in the accuracy of their predictions since both get high accuracy of around 0.9 in the loan default cases.

## 3. 3 A Naive Bayes approach to fraud prediction in loan default, Published by I O Eweoya & A A Adebiyi, Covenant University, Nigeria (2019)

Credit or loan defaults have led to bank insolvency and nations entering recession, making life unbearable for people. An approach using Naïve Bayes yielded 78% accuracy. Using cross validation and features extraction based on the principal component analysis, the training and testing was done and 25% of the dataset used for testing. The accuracy of the model is good and the extent of false alarm as evident in the false positive rate is minimal.

## 3.4 Loan default prediction using decision trees and random forest: A comparative study by Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath, [Volume 1022](https://iopscience.iop.org/volume/1757-899X/1022), [1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India](https://iopscience.iop.org/issue/1757-899X/1022/1)

With the improving banking sector in recent times and the increasing trend of taking loans, a large population applies for bank loans. But one of the major problem banking sectors face in this ever-changing economy is the increasing rate of loan defaults, and the banking authorities are finding it more difficult to correctly assess loan requests and tackle the risks of people defaulting on loans. The two most critical questions in the banking industry are (i) How risky is the borrower? and (ii) Given the borrower’s risk, should we lend him/her? In light of the given problems, this paper proposes two machine learning models to predict whether an individual should be given a loan by assessing certain attributes and therefore help the banking authorities by easing their process of selecting suitable people from a given list of candidates who applied for a loan. This paper does a comprehensive and comparative analysis between two algorithms (i) Random Forest, and (ii) Decision Trees. Both the algorithms have been used on the same dataset and the conclusions have been made with results showing that the Random Forest algorithm outperformed the Decision Tree algorithm with much higher accuracy.

## 3.5 Predictions of Loan Defaulter - A Data Science Perspective by [P. Maheswari](https://ieeexplore.ieee.org/author/38548733600), [CH. V. Narayana](https://ieeexplore.ieee.org/author/37088577119),Published in: [2020 5th International Conference on Computing, Communication and Security (ICCCS)](https://ieeexplore.ieee.org/xpl/conhome/9276407/proceeding)

With the progress of technology and implementation of Data Science in banking, changes the face of the banking industry. Most of the banking, financial sectors and social lending platforms are actively investing in lending. But financial institutions might face huge capital loss if they approved the loan without having any prior assessment of default risk. Financial institutions always need a more accurate predictive system for various purposes. Predicting loan defaulters is a crucial task for the banking industry. Banks have an immensely large amount of data like customer's data, transaction behavior, etc. Data Science is a promising area to process the data and extract the hidden patterns using machine learning techniques. This paper uses statistical measures to preprocess the data and build an effective model that will predict the loan defaulter accurately.

## 3.6 A study on predicting loan default based on the random forest algorithm by Lin Zhu,Dafeng Qiu,Daji Ergu(2019), Southwest Minzu university

Recently, with the advance of electronic commerce and big data technology, P2P online lending platforms have brought opportunities to businessmen, but at the same time, they are also faced with the risk of user loan default, which is related to the sustainable and healthy development of platforms. Therefore, based on the Random Forest algorithm, this paper builds a loan default prediction model in view of the real-world user loan data on Lending Club. The SMOTE method is adopted to cope with the problem of imbalance class in the dataset, and then a series of operations such as data cleaning and dimensionality reduction are carried out. The experimental results show that: Random Forest algorithm out performs than logistic regression, decision tree and other machine learning algorithms in predicting default samples.

## 

# 

# CHAPTER 4

# Dataset And Exploratory Data Analysis

## 4.2 Exploratory Data Analysis

Exploratory Data Analysis is very important in Machine learning. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data. It also helps you prepare your dataset for analysis and Allows a machine learning model to predict our dataset better by giving you more accurate results.

**About this file**

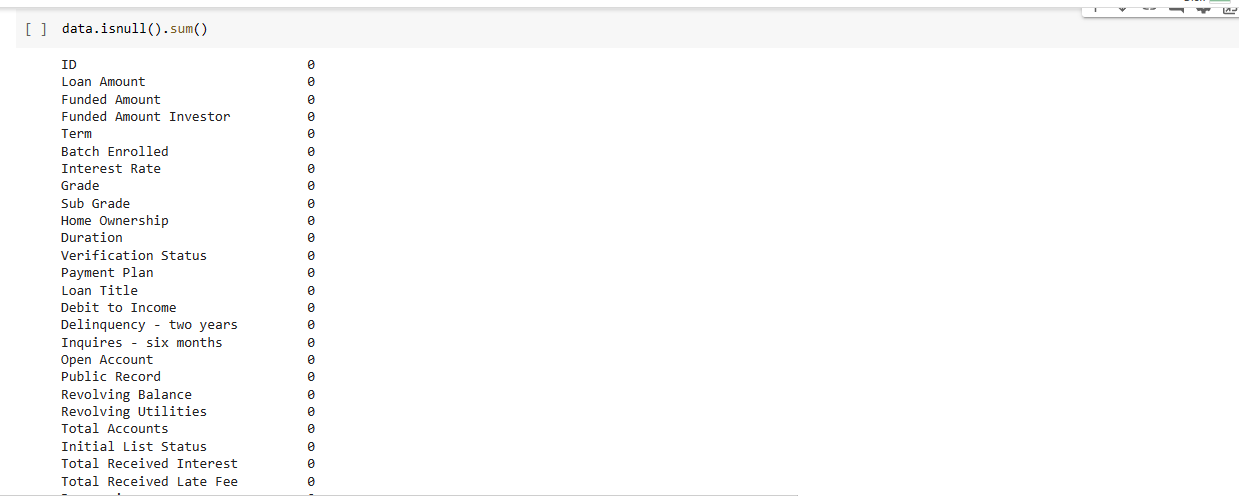
* ID: unique ID of representative
* Loan Amount: loan amount applied
* Funded Amount:loan amount funded
* Funded Amount Investor: loan amount approved by the investors
* Term: term of loan (in months)
* Batch Enrolled: batch numbers to representatives
* Interest Rate: interest rate (%) on loan
* Grade: grade by the bank
* Sub Grade: sub-grade by the bank
* Debt to Income: ratio of representative's total monthly debt repayment divided by self reported monthly income excluding mortgage
* Delinquency - two years: number of 30+ days delinquency in past 2 - years
* Inquires - six months: total number of inquiries in last 6 months
* Open Account: number of open credit line in representative's - credit line
* Public Record: number of derogatory public records
* Revolving Balance: total credit revolving balance
* Revolving Utilities: amount of credit a representative is using - relative to revolving\_balance
* Total Accounts: total number of credit lines available in - representatives credit line
* Initial List Status: unique listing status of the loan - - W(Waiting), F(Forwarded)
* Total Received Interest: total interest received till date
* Total Received Late Fee: total late fee received till date
* Recoveries: post charge off gross recovery
* Collection Recovery Fee: post charge off collection fee
* Collection 12 months Medical: total collections in last 12 months - excluding medical collections
* Application Type: indicates when the representative is an individual or joint
* Last week Pay: indicates how long (in weeks) a representative has paid EMI after batch enrolled
* Accounts Delinquent: number of accounts on which the representative is delinquent
* Total Collection Amount: total collection amount ever owed
* Total Current Balance: total current balance from all accounts
* Total Revolving Credit Limit: total revolving credit limit
* Loan Status: 1 = Defaulter, 0 = Non Defaulters
* Counting **Variables by type:**

****

Figure 4.1 Variables by type

The dataset has 67463 observations and 35 variables including target,divided into 26 numeric and 9 categorical features.

* Checking for **missing values**:



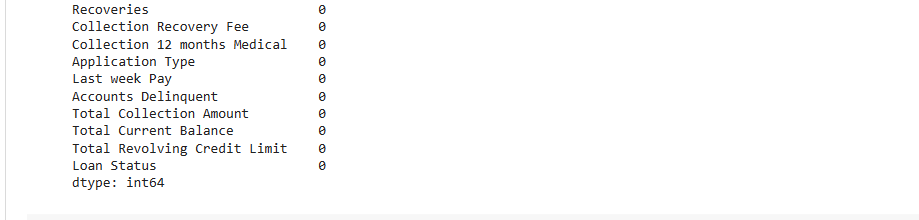


Figure 4.2 Checking missing values

We have no null values in our dataset.

* **Unbalanced data**: Target has 90.7% of non default results (value 0 ) against 9.3% of loan defaulters (value 1)

## 4.2.1 Univariate Analysis

Uni means one and this means that the data has only one kind of variable. The major reason for univariate analysis is to use the data to describe. The analysis will take data, summarize it, and then find some pattern in the data.

While plotting the values in Loan Status column,

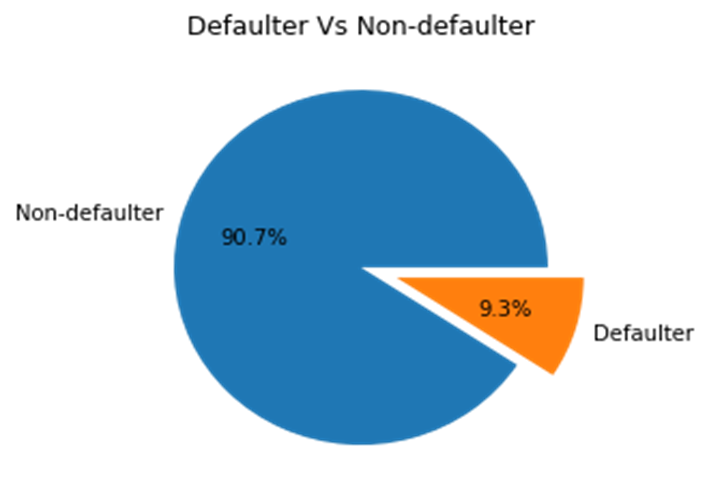
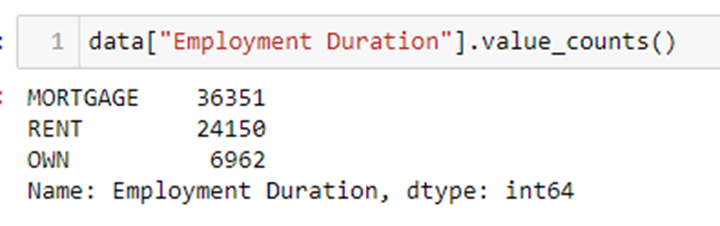
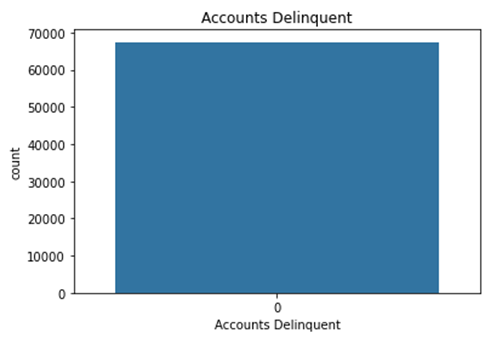
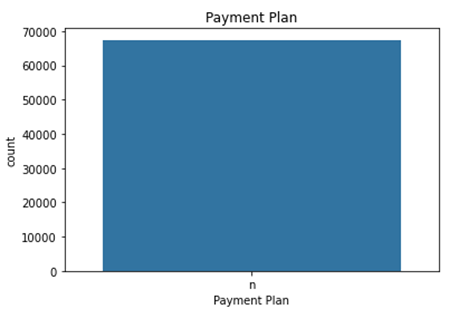


Figure 4.3 Defaulter Vs Non Defaulter

We can see the Non-defaulter observations are very high in our dataset. Defaulter observation is just 10% of Non-defaulter observation.It simply implies that our dataset is imbalanced.With this dataset, we are going to plot distplot and boxplot for numerical columns and bar plot for categorical column to find is there any unique characters within the columns.



While analyzing the dataset, ID is a unique column, so it’s a weak column.And column Employment Duration and Home Ownership didn’t hold relevant values.



Delinquent and Payment Plan columns contain a single unique value. So it is irrelevant.

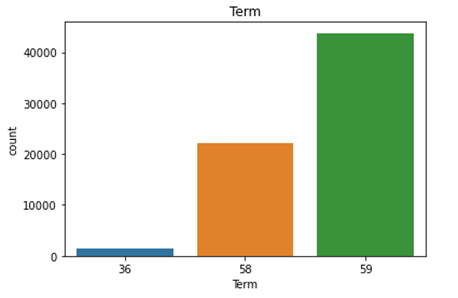
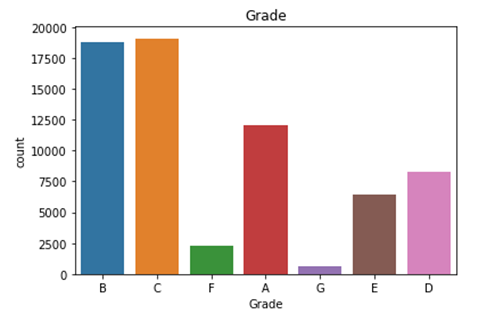


Figure 4.4 Univariate Analysis

The observations having Grade F and G were comparatively less than others and also having Term 36 months were less than others.



Figure 4.5 Univariate Analysis Observations

By analyzing box plot, the columns Funded Amount Investor, Interest Rate, Revolving Balance, Total Received Late Fee, Total Received Late Fee, Recoveries, Collection Recovery Fee, Collection Recovery Fee, Total Collection Amount, Total Current Balance, Total Revolving Credit Limit were contain outliers.

**4.2.2 Bivariate Analysis**

The term **bivariate analysis** refers to the analysis of two variables. The purpose of bivariate analysis is to understand the relationship between two variables.

The main Two types we will see here are:

1) Categorical v/s Numerical

2) Numerical v/s Numerical

* Around 9 columns hold a relevant relationship with our target column(Loan Status) like Home Ownership, Term, Open Account, Grade, Total Accounts, Last week Pay, Total Collection Amount, Revolving Balance, Total Received Late Fee, etc.

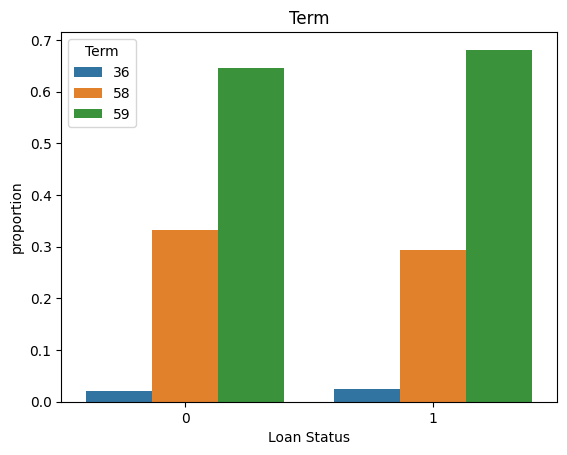
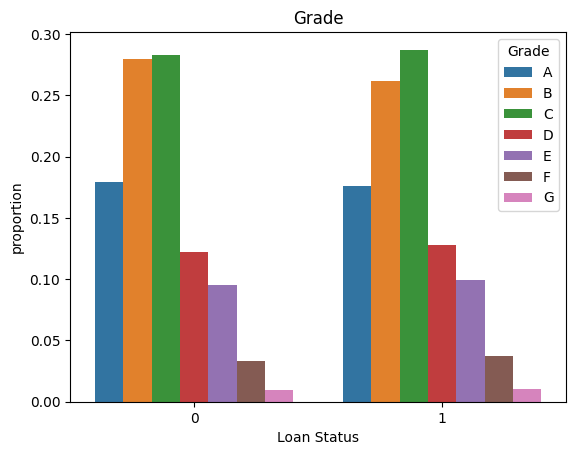
 

Figure 4.6 Bivariate Analysis- Term Figure 4.7 Bivariate Analysis- Grade

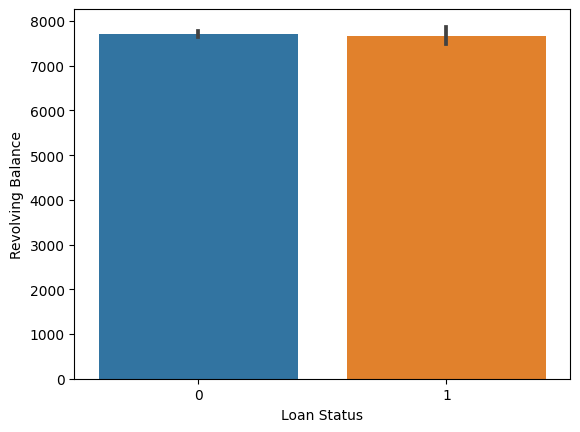
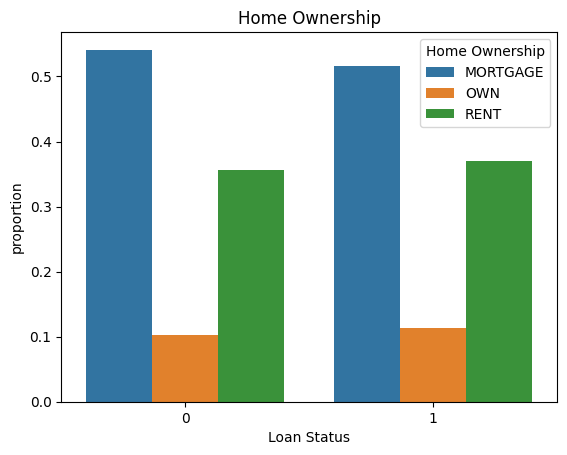
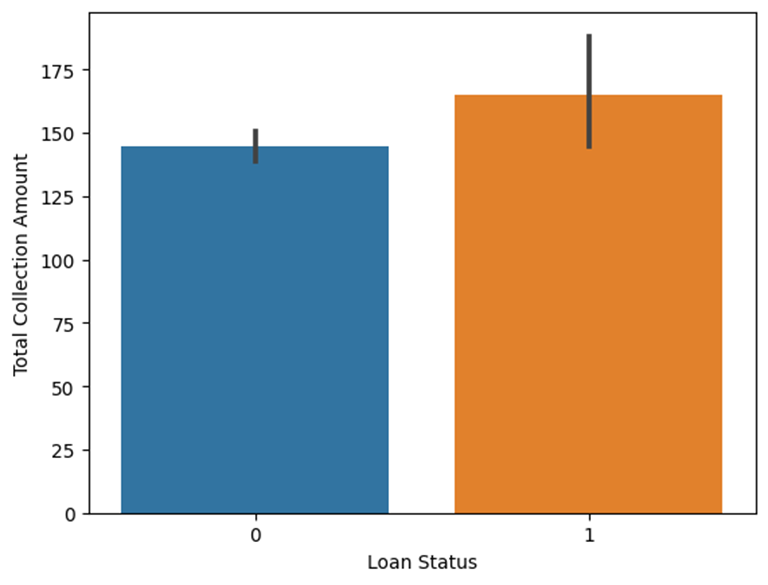
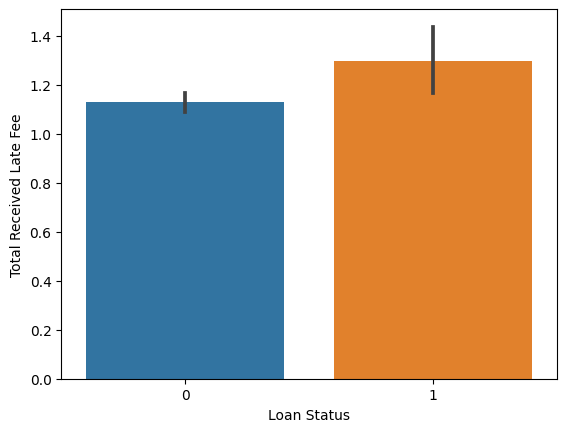


Figure 4.8 Bivariate Analysis - Home Ownership Figure 4.9 Bivariate Analysis - Revolving Balance



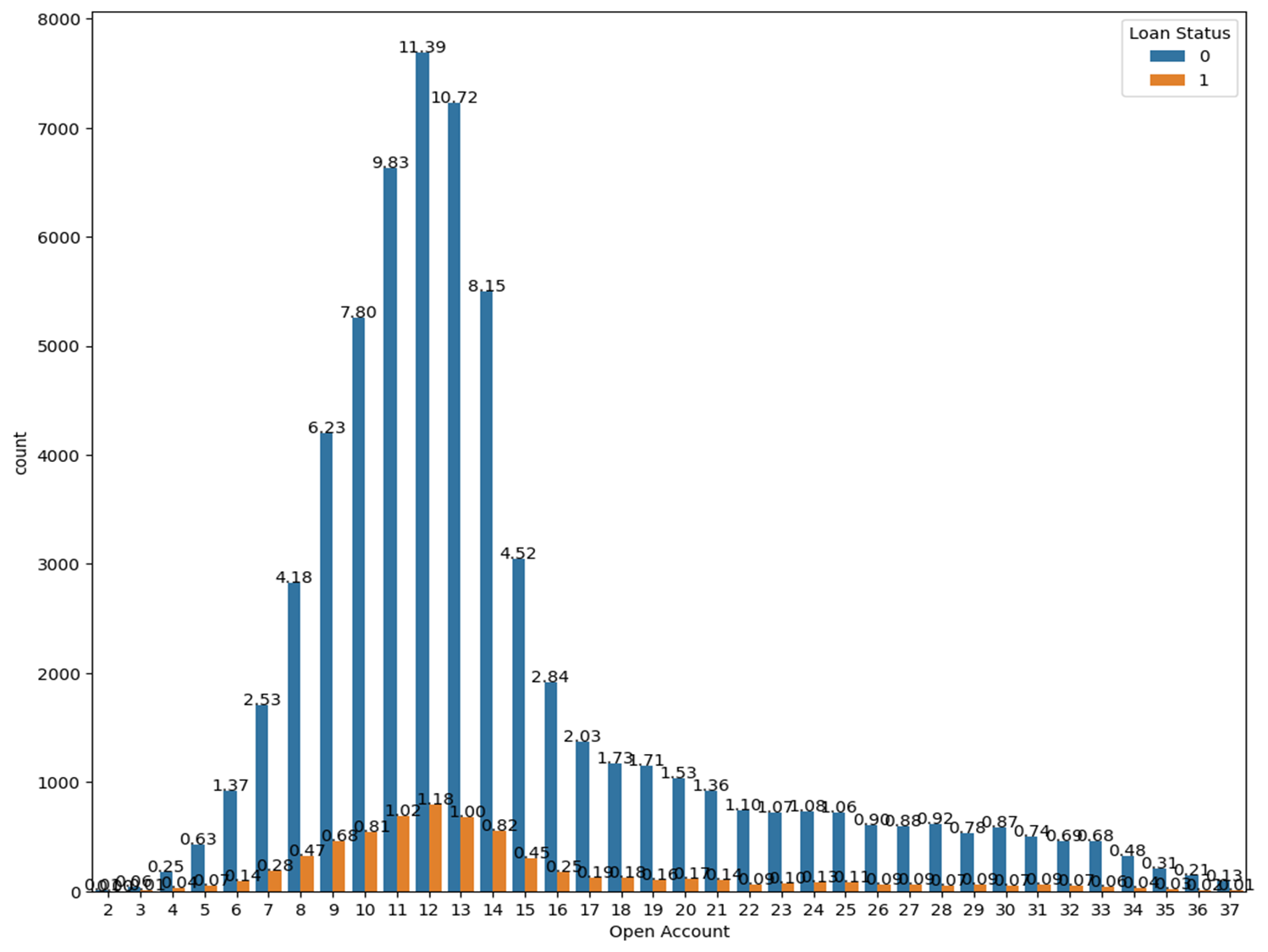
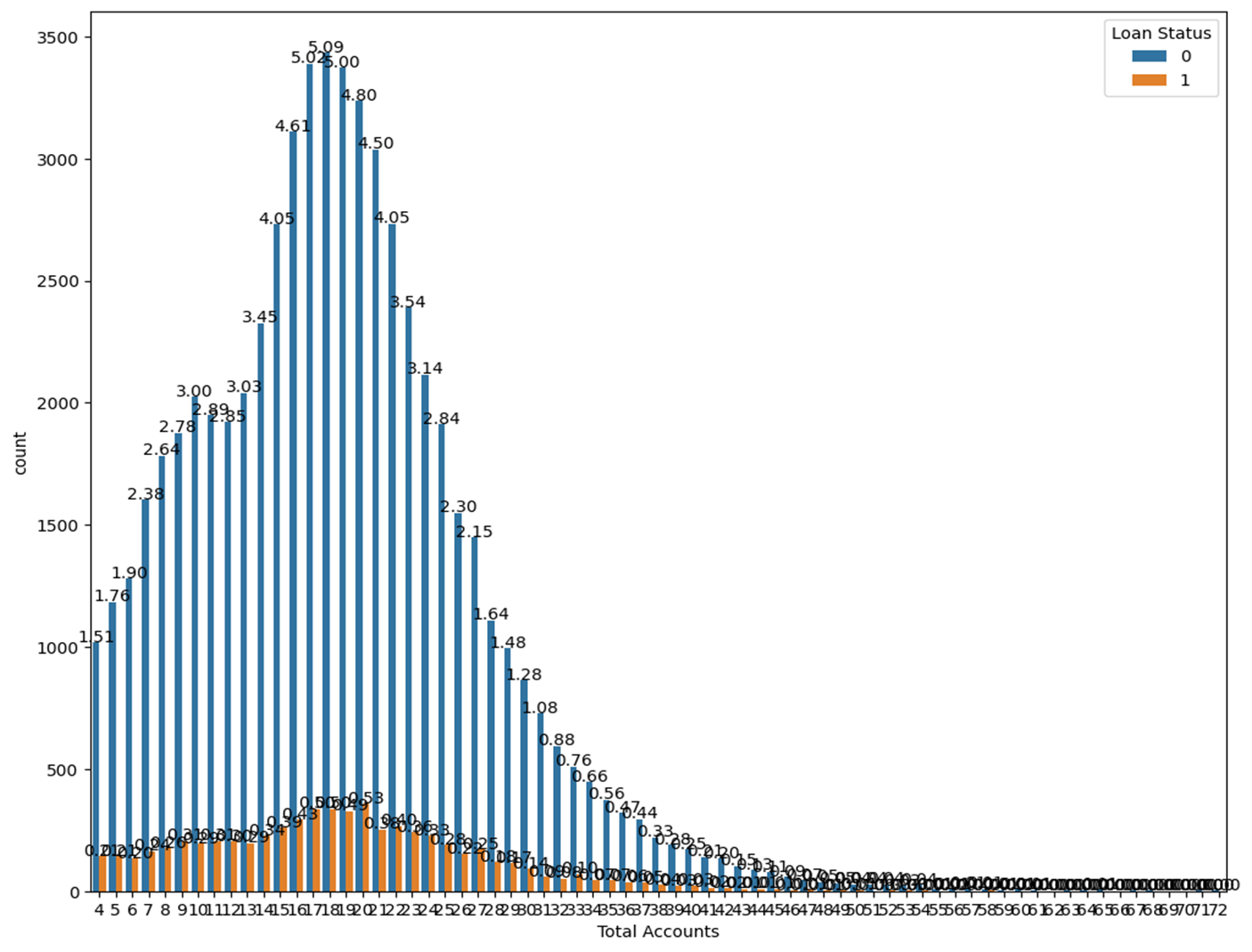


Figure 4.10 Bivariate Analysis



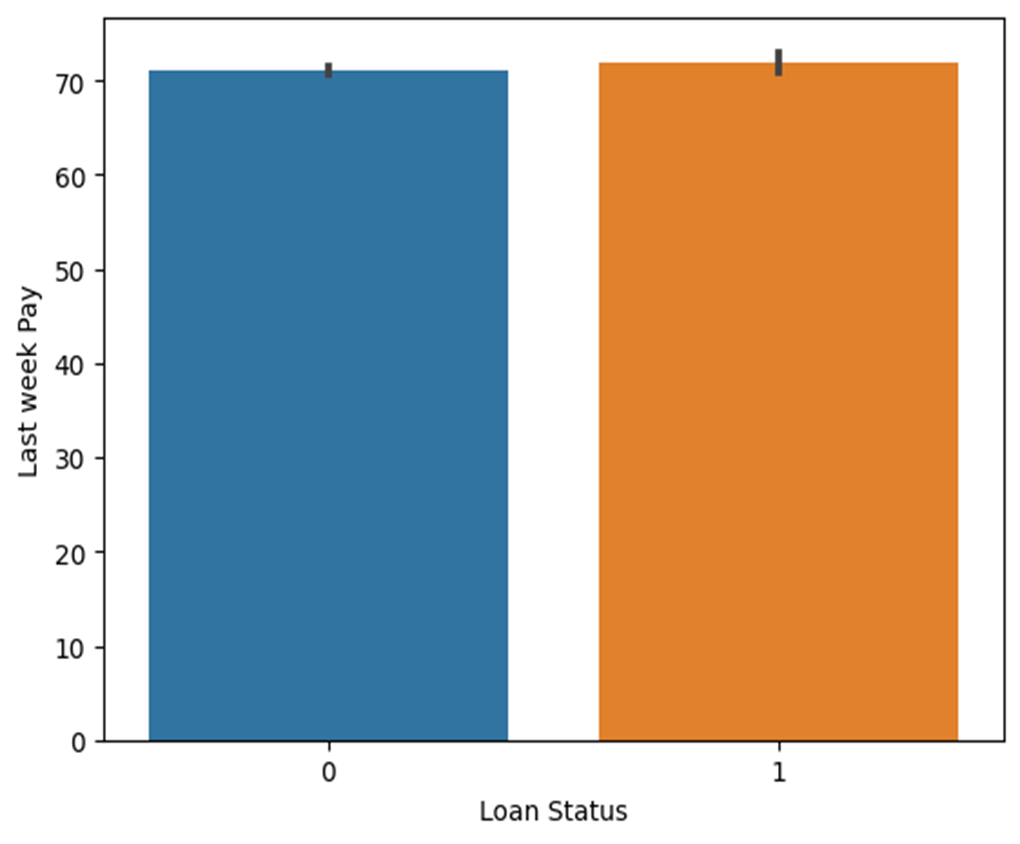
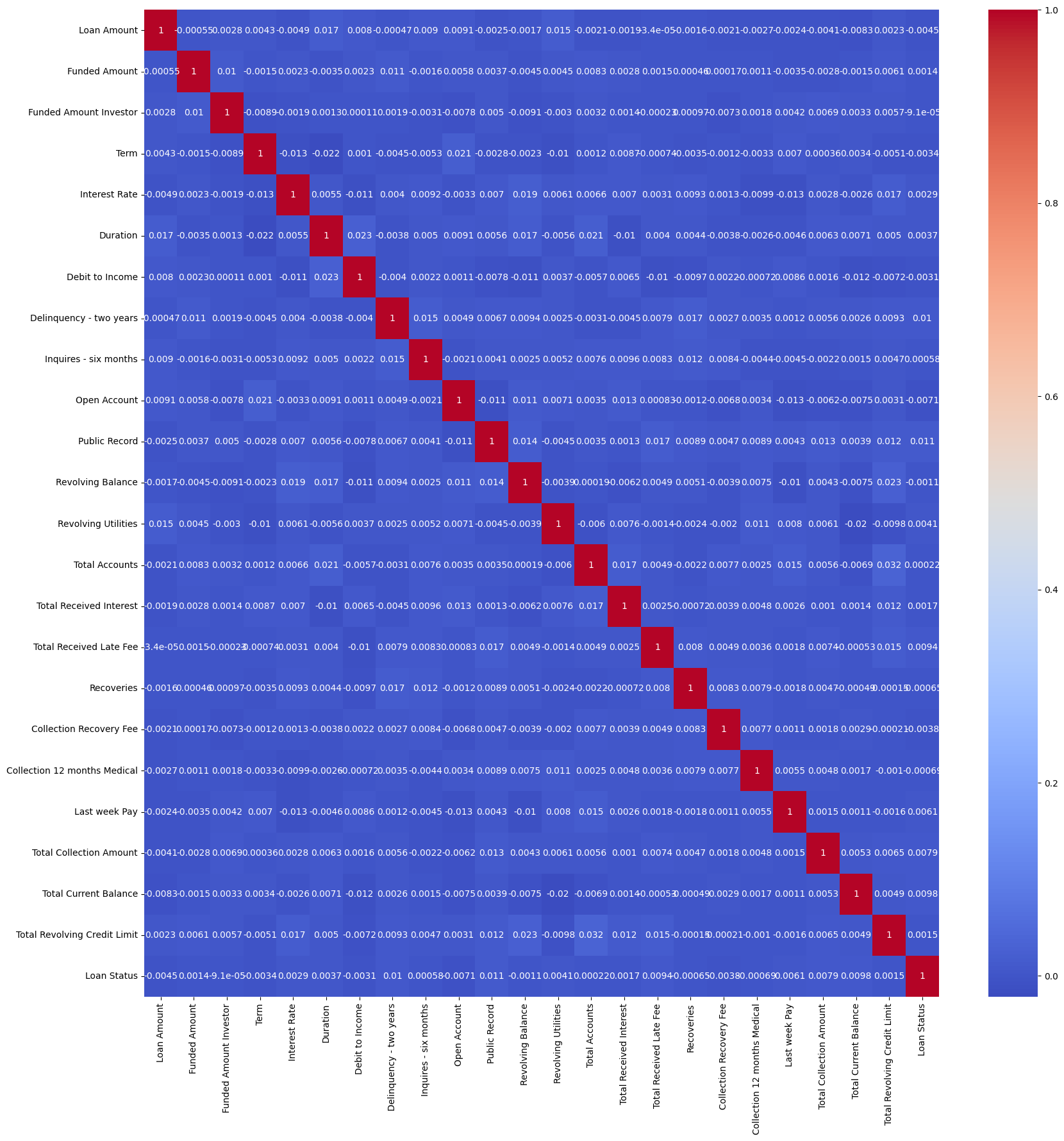


Figure 4.11 Bivariate Analysis

**4.2.3 Multivariate Analysis**

Correlation is a technique for investigating the relationship between two quantitative, continuous variables in order to represent their inter-dependencies. Its correlation coefficient scales from -1 to 1, where 1 represents the strongest positive correlation, -1 represents the strongest negative correlation and 0 represents no correlation. The correlation coefficients between each pair of the dataset are calculated and plotted as a heatmap. From the heatmap, it is easy to locate the highly correlated features with the help of color coding.



# Figure 4.12 Multivariate Analysis

# CHAPTER 5

# Feature Engineering

**5.1.1 Label Encoding data**

Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pe-processing step in a machine-learning project.

Label Encoder can be used to normalize labels. It can also be used to transform non-numerical labels (if they are hashable and comparable) to numerical labels.

Here the categorical columns in this dataset are **Home Ownership** and **Grade**.

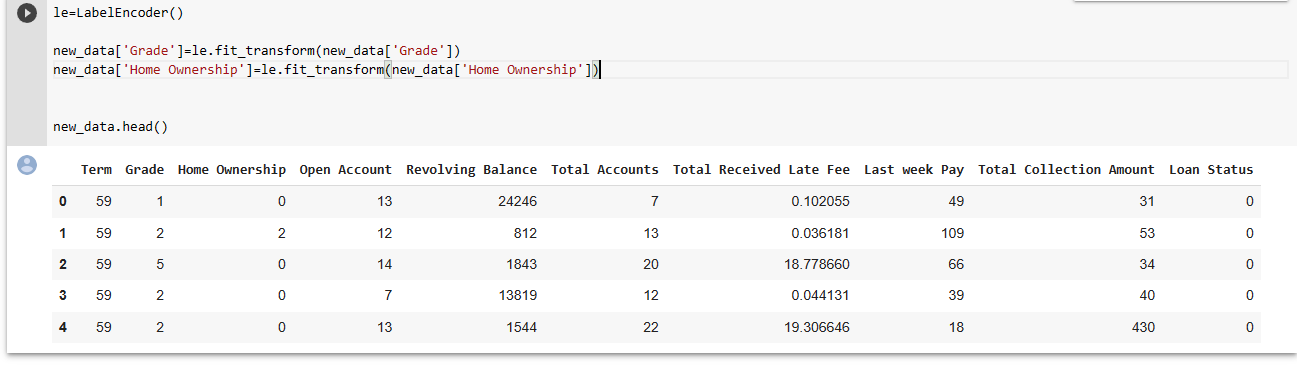


Figure 5.1 Label encoding Data

## 

## 5.1.2 One-Hot Encoding

One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model.

The advantages of using one hot encoding are: -

· It allows the use of categorical variables in models that require numerical input.

· It can improve model performance by providing more information to the model about the categorical variable.

· It can help to avoid the problem of ordinality, which can occur when a categorical variable has a natural ordering.

**5.2 Handling Outliers**

Handling outliers is an important step in data analysis and modeling. Outliers are data points that deviate significantly from the rest of the data, and they can have disproportionate impact on statistical measures and machine learning algorithms

Outlier handling is commonly used for continuous numerical columns in data analysis and machine learning tasks. Outliers are data points that deviate significantly from the overall pattern or distribution of the data. They can be the result of measurement errors, data entry mistakes

Instead of using traditional mean and standard deviation, consider using robust statistical measures such as median and Inter quartile range (IQR). These measures are less sensitive to outliers and provide a better.

We used Inter quartile Range IQR to detect outliers and we used the box plot to visualize the presence of outliers. The points that lie beyond the whiskers are detected as outliers .

In our data set there are total 9 columns in which we found 3 columns as continuous numerical column so we used IQR( Inter quartile Range ) method to detect outliers and used box plot to visualize the presence of outliers ,therefore we removed outliers in 2 column such as (Revolving Balance , Total Received Late Fee )

The box plot can be helpful in visually assessing the impact of these outlier handling techniques by comparing the box plots before and after outlier management. The updated box plot can provide insights into how the distribution has changed and whether outliers have been effectively managed.

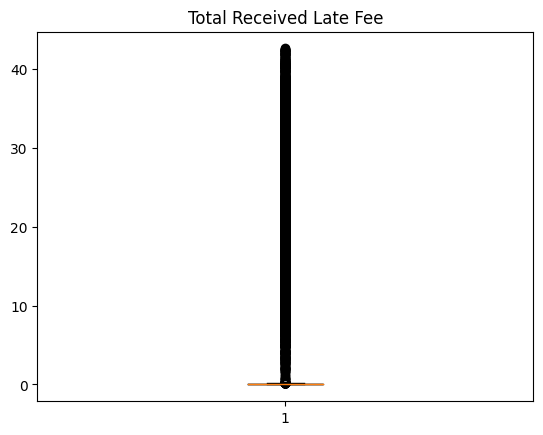
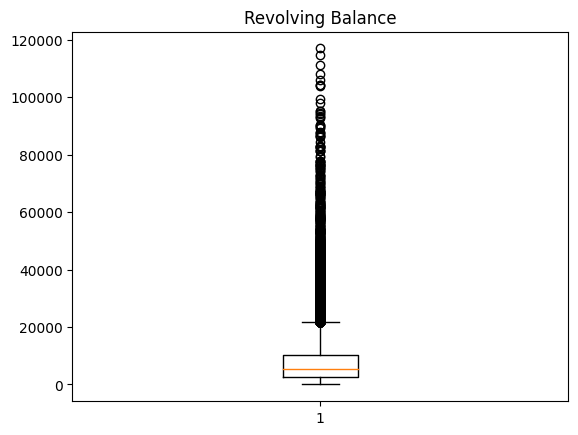


Figure 5.2 Handling Outliers

# CHAPTER 6

# Feature Scaling and Feature Reduction

## 

## 6.1 Feature scaling

Feature Scaling is a data preprocessing step for numerical features. Many machine learning algorithms like Gradient descent methods,KNN algorithm,linear and logistic regression,etc.requires data scaling to produce good results. Various scalers are used for this purpose. Here we use **MinMax Scaler.** This is a way of data scaling,where the minimum of feature is made equal to zero and the maximum of feature equal to one. MinMax Scaler shrinks the data within the given range,usually of 0 to 1.It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

In our dataset we have done scaling on continuous numerical columns like Total Received Late Fee,Total Collection Amount,Revolving Balance.

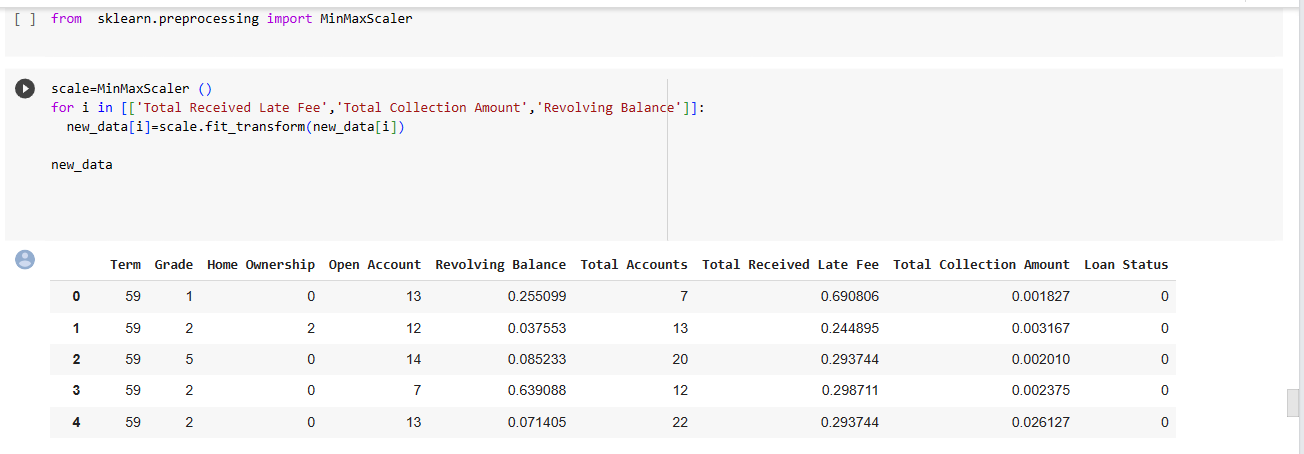


Figure 6.1 Feature Scaling

## 6.2 Feature Reduction

Feature reduction is also known as dimensionality reduction, is the process of reducing the number of features in a resource heavy computation without losing important information.A dataset contains a huge number of input features in various cases, which makes the predictive modeling task more complicated. Because it is very difficult to visualize or make predictions for the training dataset with a high number of features, for such cases, dimensionality reduction techniques are required to use.

Here we use **Variance Inflation Factor(VIF)** to measure the amount of inflation in the variance of the regression coefficients due to multicollinearity. Specifically, VIF is calculated as the ratio of the variance of coefficient in a model with only one predictor. A VIF value of 1 indicates no multicollinearity, while values greater than 1 indicate increasing levels of multicollinearity.

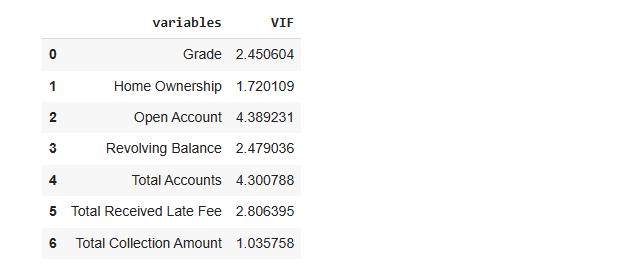


Figure 6.2 Feature Reduction

**CHAPTER 7**

**MODEL BUILDING AND MODEL SELECTION**

Machine learning models are powerful tools used to efficiently and effectively perform vital tasks and solve complex problems.Machine learning models are created from machine learning algorithms,which are trained using either labeled,unlabeled,or mixed data. Different machine learning algorithms are suited to different goals,such as classification or prediction modeling.We use Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, XGB Classifier, Naive Bayes Classifier etc.

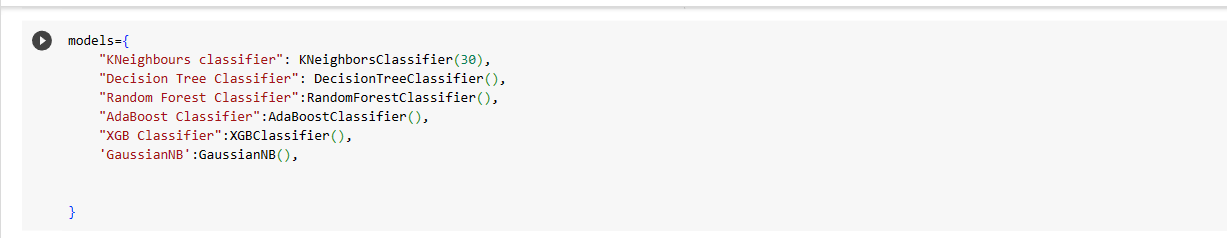
****

Figure 7.1 Model Building and Model Selection

**7.1 OVER SAMPLING OF DATA**

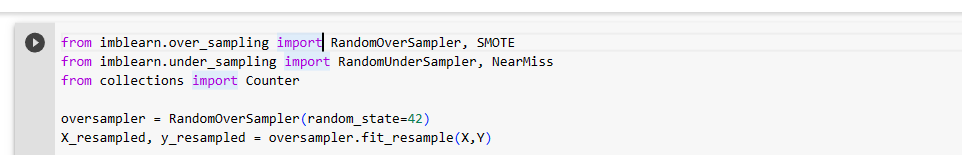
Oversampling can be defined as adding more copies to the minority class. Oversampling can be a good choice when you don't have a ton of data to work with.Here we use Random Oversampler for balancing the data. Random Oversampling involves randomly selecting examples from minority classes,with replacement,and adding them to the training dataset.  


Figure 7.2 Oversampling

**7.2 CONFUSION MATRIX**

The confusion matrix is a matrix used to determine the performance of classification models for a given set of data. It evaluates the performance of the classification models,when they make predictions on test data,and tells how good our classification model is. With the help of the confusion matrix, we can calculate the different parameters for the model,such as accuracy, precision,etc.

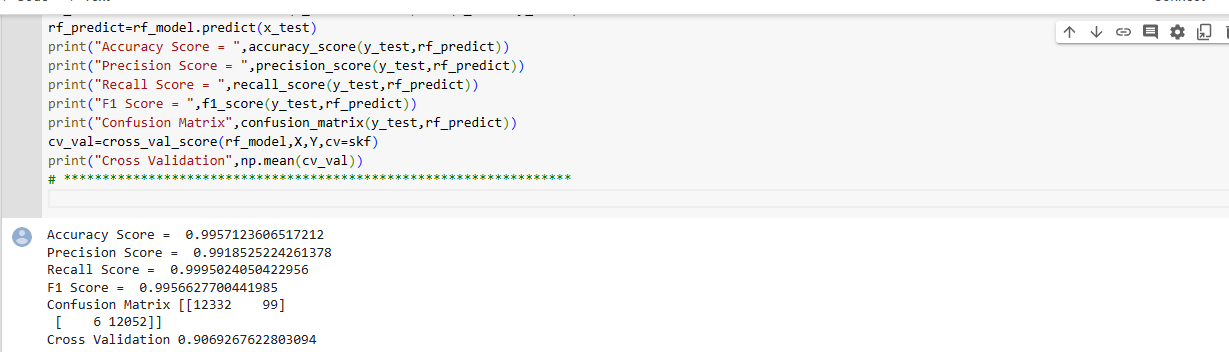


Figure 7.3 Confusion Matrix

**CHAPTER 8**

**HYPER PARAMETER TUNING**

Hyperparameter tuning refers to the process of selecting the best hyperparameters for a machine learning algorithm to optimize its performance. Hyperparameters are configuration settings that are not learned during the training process, but rather set by the user before training begins.These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

Here we used Randomized Search CV and Grid Search CV

**8.1 Randomized Search CV**

Randomized search is a valuable technique for hyperparameter tuning, allowing you to efficiently search through the hyperparameter space and find good configurations for your machine learning models.

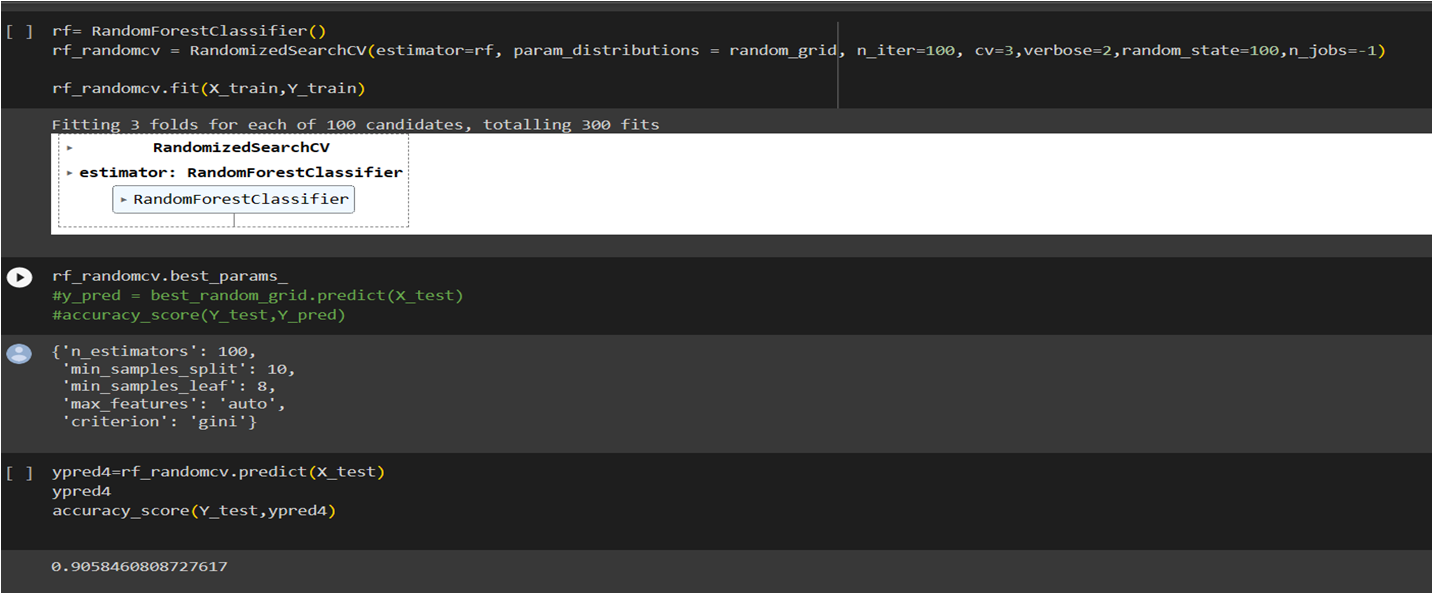
****

Figure 8.1 Randomized Search CV

**8.2 Grid Search CV**

Grid Search CV is a powerful tool for hyperparameter tuning, allowing you to systematically explore the hyperparameter space and identify the best settings for your machine learning model



Figure 8.2 Grid Search cv

**8.3 PICKLING**

Pickling is a powerful feature in Python that allows you to serialize and deserialize objects, providing a convenient way to store and retrieve data structures in a persistent and portable format.

The pickle file preserves the internal state and structure of the object, allowing you to retrieve it and work with it again without the need to re-create or retrain the object.

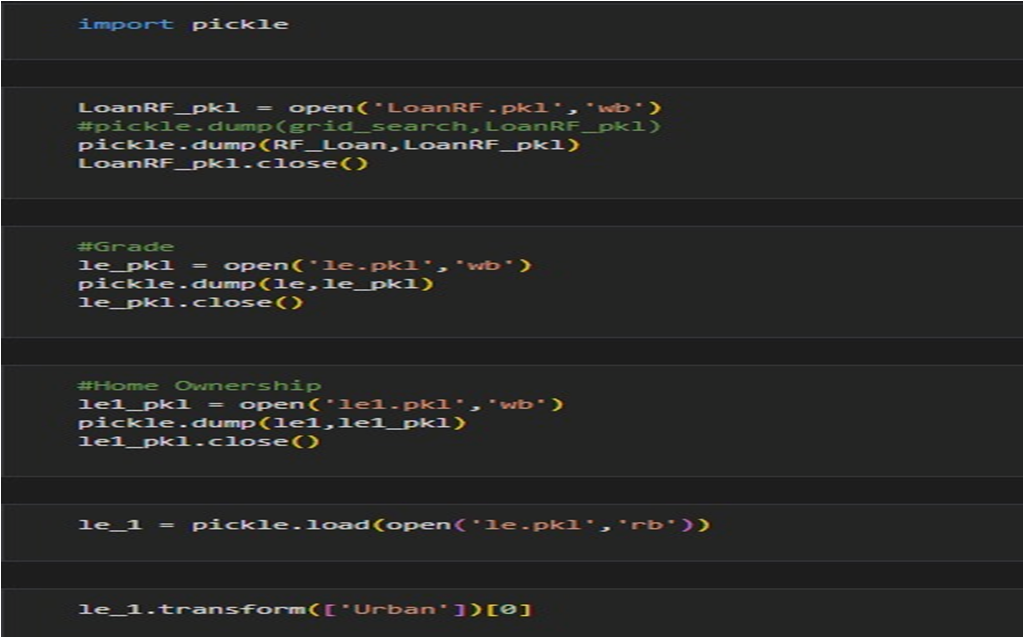


Figure 8.3 Pickling

# 9. Result

By pickling our model and label encoded and scaled data we created a web application using Visual studio.Thus our application is working fine and able to predict defaulters and analyze non defaulters.

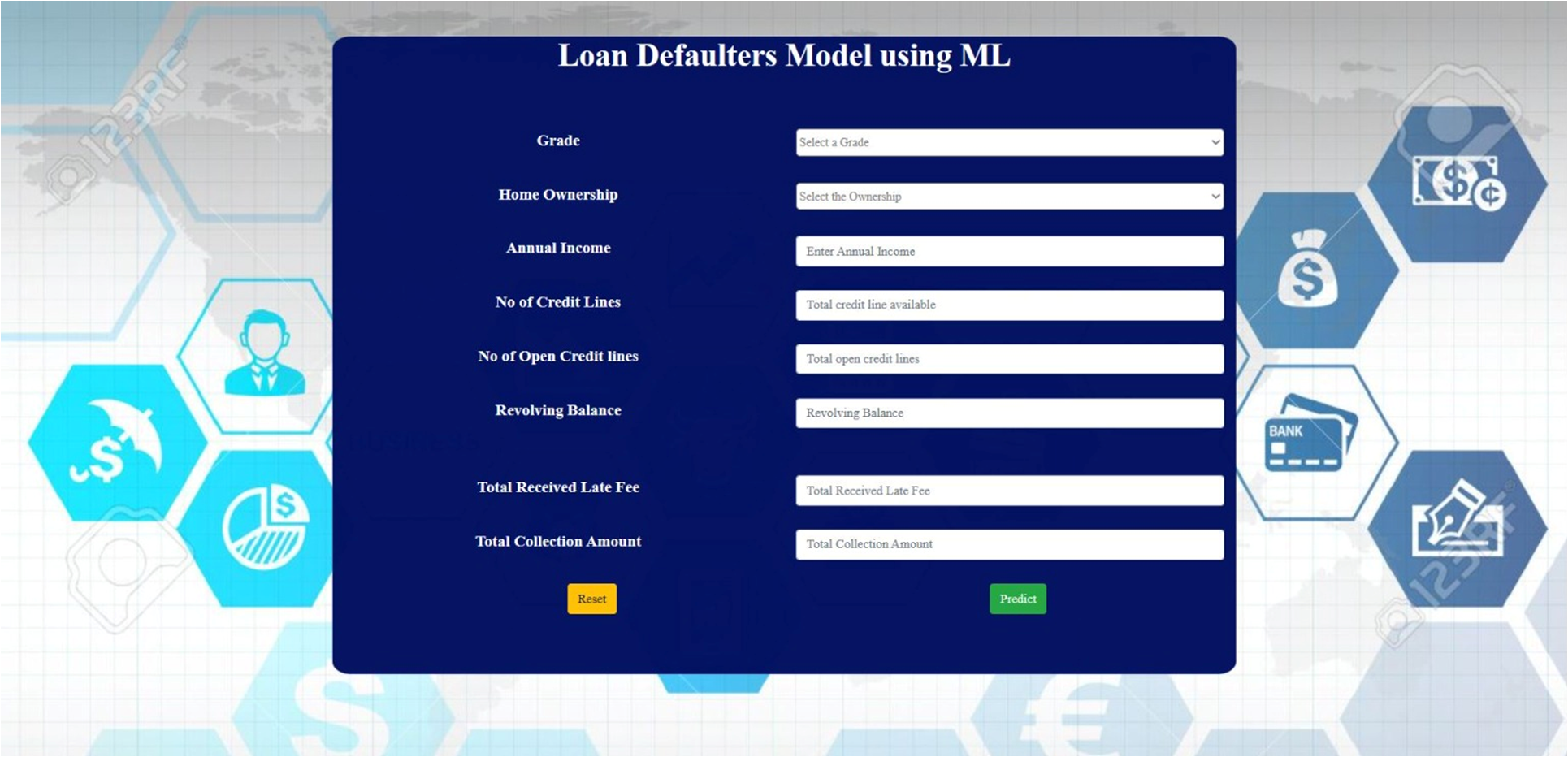


Figure 9.1 Result

# 8. Conclusion

Thus we implemented a machine learning model using the algorithm Random Forest classifier. It was showing the highest accuracy on our dataset so we selected this model. It is actually a collection of decision trees.It works fine when we pickled the model by selecting best parameters from random search cv and grid search cv. So our application is having high accuracy in classifying defaulters and non defaulters of loans.

# 

# 

# 

# 

# 9.References

[1] Loan default prediction using decision trees and random forest: A comparative study by Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath, [Volume 1022](https://iopscience.iop.org/volume/1757-899X/1022), [1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India](https://iopscience.iop.org/issue/1757-899X/1022/1)

[2]Predictions of Loan Defaulter - A Data Science Perspective by [P. Maheswari](https://ieeexplore.ieee.org/author/38548733600), [CH. V. Narayana](https://ieeexplore.ieee.org/author/37088577119),Published in: [2020 5th International Conference on Computing, Communication and Security (ICCCS)](https://ieeexplore.ieee.org/xpl/conhome/9276407/proceeding)

[3] Madaan, M., Kumar, A., Keshri, C., et al. (2021) Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study. IOP Conference Series: Materials Science and Engineering, 1022, 012042. <https://doi.org/10.1088/1757-899X/1022/1/012042>

[4]A study on predicting loan default based on the random forest algorithm by Lin Zhu,Dafeng Qiu,Daji Ergu(2019), Southwest Minzu university

[5] Ma, X., Sha, J., Wang, D., et al. (2018) Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. Electronic Commerce Research and Applications, 31, 24-39. <https://doi.org/10.1016/j.elerap.2018.08.002>

[6] Nalić, J. and Švraka, A. (2018) Using Data Mining Approaches to Build Credit Scor- W. J. Wu DOI: 10.4236/iim.2022.145011 164 Intelligent Information Management ing Model: Case Study—Implementation of Credit Scoring Model in Microfinance Institution. 2018 17th International Symposium Infoteh-Jahorina (INFOTEH), East Sarajevo, 21-23 March 2018, 1-5. https://doi.org/10.1109/INFOTEH.2018.8345543

[7] Baesens, B., Roesch, D. and Scheule, H. (2016) Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119449560

[8] Han, J.T., Choi, J.S., Kim, M.J., et al. (2018) Developing a Risk Group Predictive Model for Korean Students Falling into Bad Debt. Asian Economic Journal, 32, 3-14. <https://doi.org/10.1111/asej.12139>

[9] Marqués, A.I., García, V. and Sánchez, J.S. (2012) Exploring the Behaviour of Base Classifiers in Credit Scoring Ensembles. Expert Systems with Applications, 39, 10244-10250. <https://doi.org/10.1016/j.eswa.2012.02.092>

[10] Wang, Y. and Priestley, J.L. (2017) Binary Classification on Past Due of Service Accounts Using Logistic Regression and Decision Tree. Grey Literature from PhD Candidates. <http://digitalcommons.kennesaw.edu/dataphdgreylit/>

[11] Li, X.H. (2013) Using “Random Forest” for Classification and Regression. Chinese Journal of Applied Entomology, 50, 1190-1197. [12] Li, Z., Li, S., Li, Z., et al. (2021) Application of XGBoost in P2P Default Prediction. Journal of Physics: Conference Series, 1871, 012115. https://doi.org/10.1088/1742-6596/1871/1/012115