Loan Default Forecasting using Data Mining

Bachelor in Technology
in
Information Technology

By,

College Id	Name	Email ID
171080015	Shubham Pakhare	sspakhare_b17@it.vjti.ac.in
171080071	Rutwik Chaudhari	rkchaudhari_b17@it.vjti.ac.in
171081053	Pranjal Shinde	pvshinde_b17@it.vjti.ac.in
171081064	Payal Bathija	prbathija_b17@it.vjti.ac.in

Under the guidance of:
Prof. S.G Bhirud

Department of Computer Engineering and Information Technology
Veermata Jijabai Technology Institute, Mumbai-400019

Contents

Chapter 1

Abstract

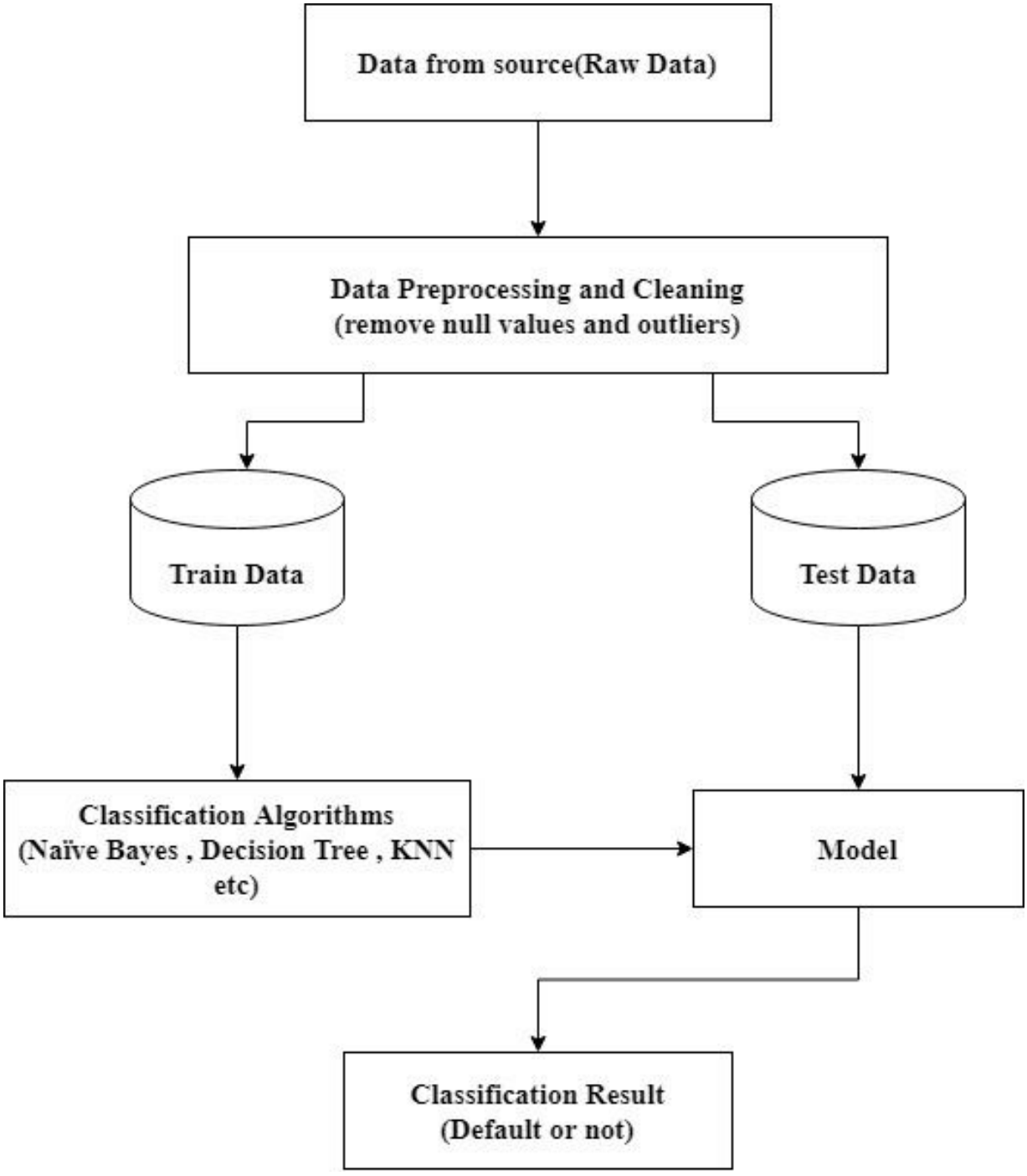
Estimation or assessment of default on a debt is a crucial process that should be carried out by banks to help them to assess if a loan applicant can be a defaulter at a later phase so that they process the application and decide whether to approve the loan or not. The conclusion derived from such assessments helps banks and other financial institutions to lessen their losses and eventually increase the number of credits. Hence, it becomes vital to construct a model that will take into account the different aspects of an applicant and derive a result regarding the concerned applicant. All available means to loan the money from their illicit activities are used for criminal activities in today's technology-based realm. The increasing number of bad debts resulting from commercial banks' loans reflects the growing problem of distraught banks within the economic system. We have used data mining algorithms to predict the likely defaulters from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future.

Keywords—loan, credit, prediction, data mining

Problem Statement

Estimation or assessment of default on a debt is a crucial process that should be carried out by banks to help them to assess if a loan applicant can be a defaulter at a later phase so that they process the application and decide whether to approve the loan or not. The conclusion derived from such assessments helps banks and other financial institutions to lessen their losses and eventually increase the number of credits. Hence, it becomes vital to construct a model that will take into account the different aspects of an applicant and derive a result regarding the concerned applicant. All available means to loan the money from their illicit activities are used for criminal activities in today's technology-based realm. The increasing number of bad debts resulting from commercial banks' loans reflects the growing problem of distraught banks within the economic system. We have used data mining algorithms to predict the likely defaulters from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future.

Workflow


Data from source(Raw Data)
Data Preprocessing and Cleaning
(remove null values and outliers)
Train Data
Test Data
Classification Algorithms
(Naïve Bayes , Decision Tree , KNN
etc)
Model
Classification Result
(Default or not)

Chapter 2

Dataset Analysis

2.1 Dataset 1

2.1.1 Link:

https://www.kaggle.com/jannesklaas/model-trap

2.1.2 Data Source:

We obtained a home loan dataset from Kaggle. The dataset consists of various variables such as minority, sex, ZIP, rent, education, income, loan size, payment timing year, job stability and occupation.

2.1.3 Data Description:

The dataset has **160000** tuples and 15 attributes. 1 out of the 15 attributes is the target attribute viz. default.

The dataset is divided into : **80% training** data & **20% test** data

The dataset has 15 columns, namely :

1. unnamed: 0 (Categorical)
2. unnamed: 0.1 (Categorical)
3. minority (Binary)
4. sex (Binary)
5. zip (Numerical)
6. rent (Nominal)
7. education (Nominal)
8. age (Nominal)
9. income (Nominal)
10. loan_size (Nominal)
11. payment_timing (Nominal)
12. job_stability (Nominal)
13. year (Numerical)
14. default (Categorical)- Target
15. occupation (Nominal)

2.1.4 Data Preprocessing and cleaning

1. **Detection of missing values**

   No null values found in the dataset.

2. **Removing outliers**

   A few outliers are present in the dataset which have been detected using scatterplot. Further, those outliers were removed using Z-score, with a threshold value of 3.

```
<matplotlib.axes._subplots.AxesSubplot at 0xed364f0>
```

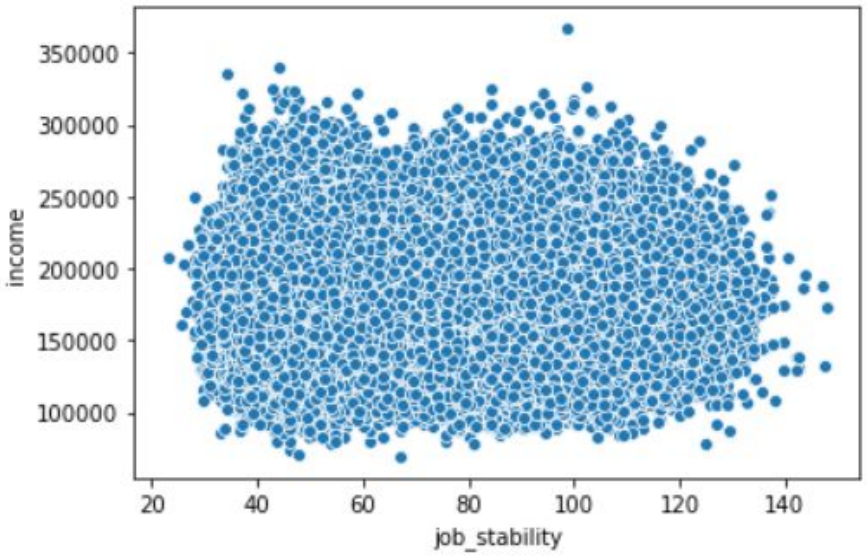


Figure 1 : Correlation between job_stablility and income

```
<matplotlib.axes._subplots.AxesSubplot at 0x11e879d0>
```

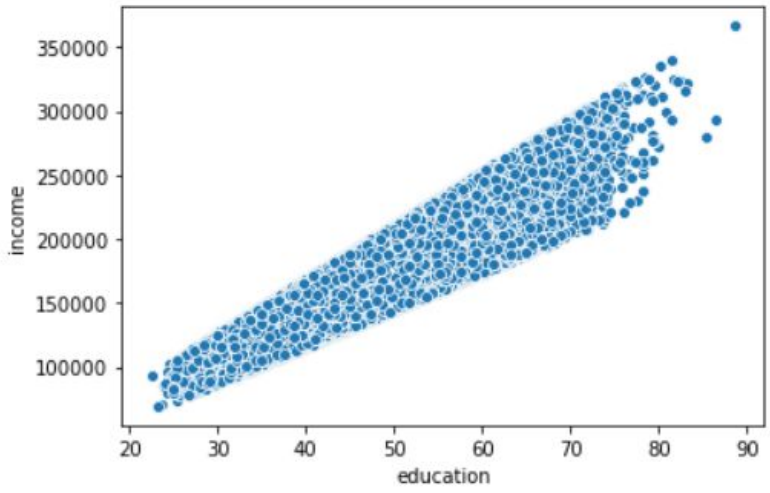


Figure 2 : Correlation between education and income

3. Correlation of attributes

In this section, we discussed the various attributes that affect the result, or in other words, the major variables that have an impact on the behaviour of the target attribute i.e. default. We analysed each of our attribute's behaviour and whether they have an impact on the target attribute. We also found the number of attributes on which the target attribute depends on. All of this was analysed using a heatmap. A **heatmap** is a visual representation of the correlation matrix.

The below plot helps to identify all the attributes that have a positive and a negative impact on the target attribute i.e. default.

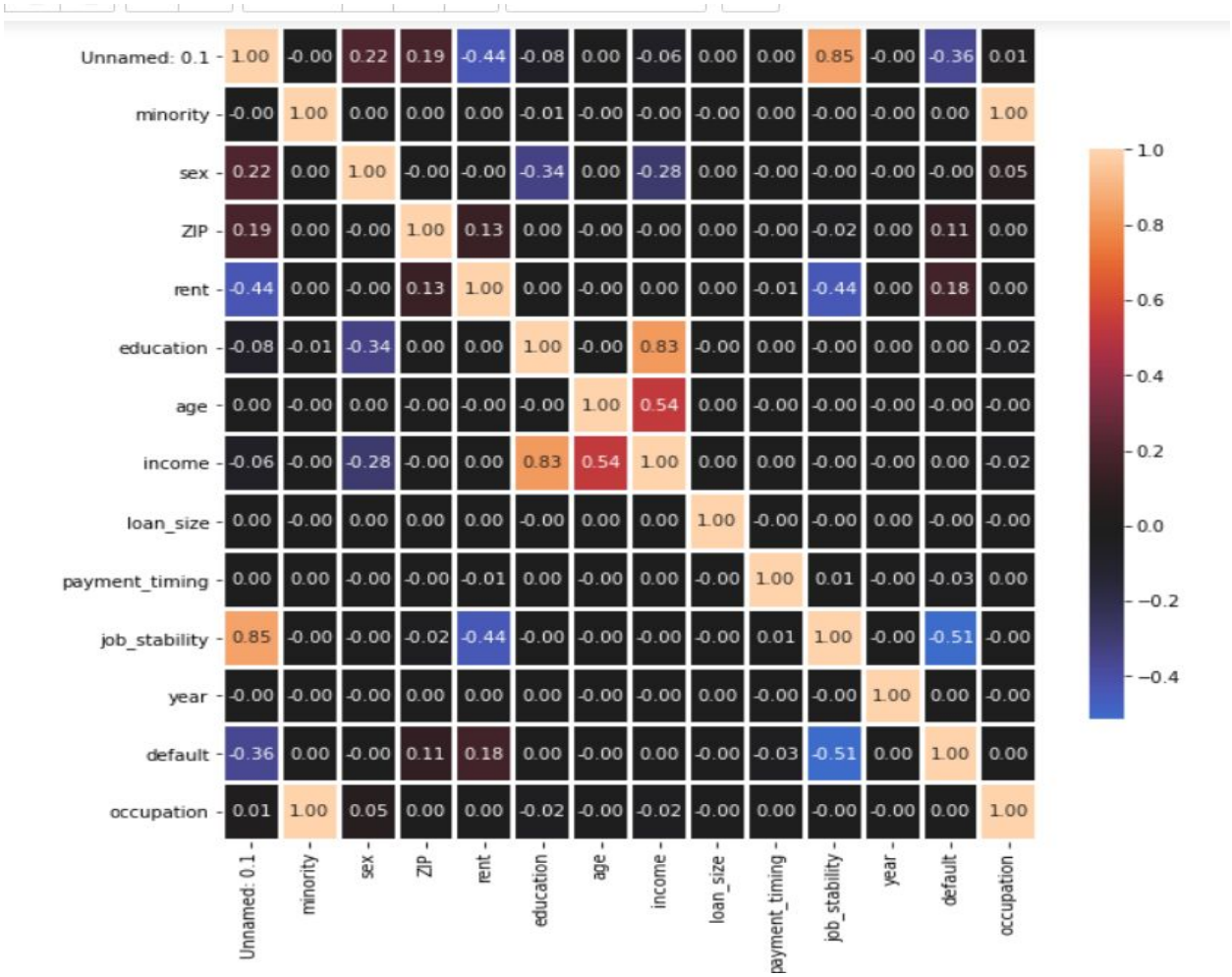


Figure 3 : Heatmap 1

MAJOR NEGATIVE ATTRIBUTES AFFECTING DEFAULT

More the job stability, less is the chance of being defaulter

Sr No.	Attribute	Description
1.	job_stability	Describes whether a person has a stable job or no

MAJOR POSITIVE ATTRIBUTES AFFECTING DEFAULT

More the rent, more the chance to be defaulter

Sr No.	Attribute	Description
1.	rent	States whether a person pays a rent or no

CORRELATED ATTRIBUTES

The attributes that are highly correlated to each other are as follows. These attributes are positively dependent on each other

Sr No.	Attribute 1	Attribute 2
1.	age	income
2.	education	income

4. **Feature Selection**

   From the above heatmap, we have selected the following important attributes:

   1. Unnamed 0.1
   2. Zip
   3. Rent
   4. Education
   5. Age
   6. Income
   7. Job_stability

2.1.5 Advantages of using the dataset:

1. Big dataset with many attributes
2. Balanced dataset with unique values for each attribute
3. No missing values

2.1.6 Disadvantages of using the dataset:

1. Many attributes have very less correlation with the target variable, as a result they need to be eliminated
2. High dimensionality
3. Some attributes are not self explanatory.

2.1.7 Performance Report:

Sr No	Model	Accuracy	Precision	Recall	F-1 Score
1	Logistic Regression	0.88686	0.87471	0.88686	0.87499
2	Random Forest	0.98171	0.98189	0.98171	0.98128
3	Decision Tree	0.98523	0.98594	0.98523	0.98542
4	Naive Bayes	0.93167	0.94958	0.93167	0.93612
5	KNN	0.81703	0.77778	0.81703	0.79334

3.1.8 Comparison of Different models Applied:

Above table represents the values obtained for the various metrics from the different models.

According to the table, The models **Random Forest and Decision Tree** have greater values (>0.98)

1. Since the **Precision and Accuracy** of 2 models are similar ~0.98, we will compare using F1-score.
2. According to **Recall & F1-score**, Decision Tree has greater value (0.99) than Random Forest.

3. Since the false negatives cost is the highest, the most optimal model will be the one with the minimum false negatives. In other words, a model with higher sensitivity/ Recall will fetch a higher net revenue compared to other models.
4. Since, Recall of Logistic Regression and KNN classifiers for predicting the defaulters is less than 0.90, we will eliminate these models.

   Therefore, we can infer that the **Decision Tree Classifier** is doing prediction well for our dataset.

2.2 Dataset 2

2.2.1 Link:

https://www.kaggle.com/burak3ergun/loan-data-set

2.2.2 Data Source:

We obtained a home loan dataset from Kaggle. The dataset consists of various variables such as gender, marital status, dependents, education, employment, applicant income, co applicant income, loan amount, loan amount for term, credit history, property area and loan status.

2.2.3 Data Description:

The data set consists of 13 Columns(Attributes) and 614 records, which are as follows:

1. Loan_ID (Nominal)
2. Gender (Categorical)
3. Married (Categorical)
4. Dependents (Categorical)
5. Education (Categorical)
6. Self_Employed (Categorical)
7. ApplicantIncome (Numerical)
8. CoapplicantIncome (Numerical)
9. LoanAmount (Numerical)
10. Loan_Amount_Term (Ordinal)
11. Credit_History (Binary)
12. Property_Area (Numerical)
13. Loan_Status (Binary)

2.2.4 Data Preprocessing and cleaning:

1. **Detection of missing values**
   - The dataset has missing values for various attributes, so we remove them by applying various preprocessing techniques.
   - There are missing values for Gender (13), Married(3),Dependents(15),Self_Employed(32),LoanAmount(22),Loan_Amount_Term(14) and Credit_History(50).
   - For the binary and categorical attributes, we replace the missing values with the most frequently occuring value of that particular attribute.
   - For Numerical Values, the missing values have been filled with the mean of all the values of that attribute as all the records are unique.

2. **Correlation of attributes:**
   In this section, we discussed the various attributes that affect the result, or in other words, the major variables that have an impact on the behaviour of the target attribute i.e. default. We analysed each of our attribute's behaviour and whether they have an impact on the target attribute. We also found the number of attributes on which the target attribute depends on. All of this was analysed using a heatmap. A **heatmap** is a visual representation of the correlation matrix.
   The below plot helps to identify all the attributes that have a positive and a negative impact on the target attribute.

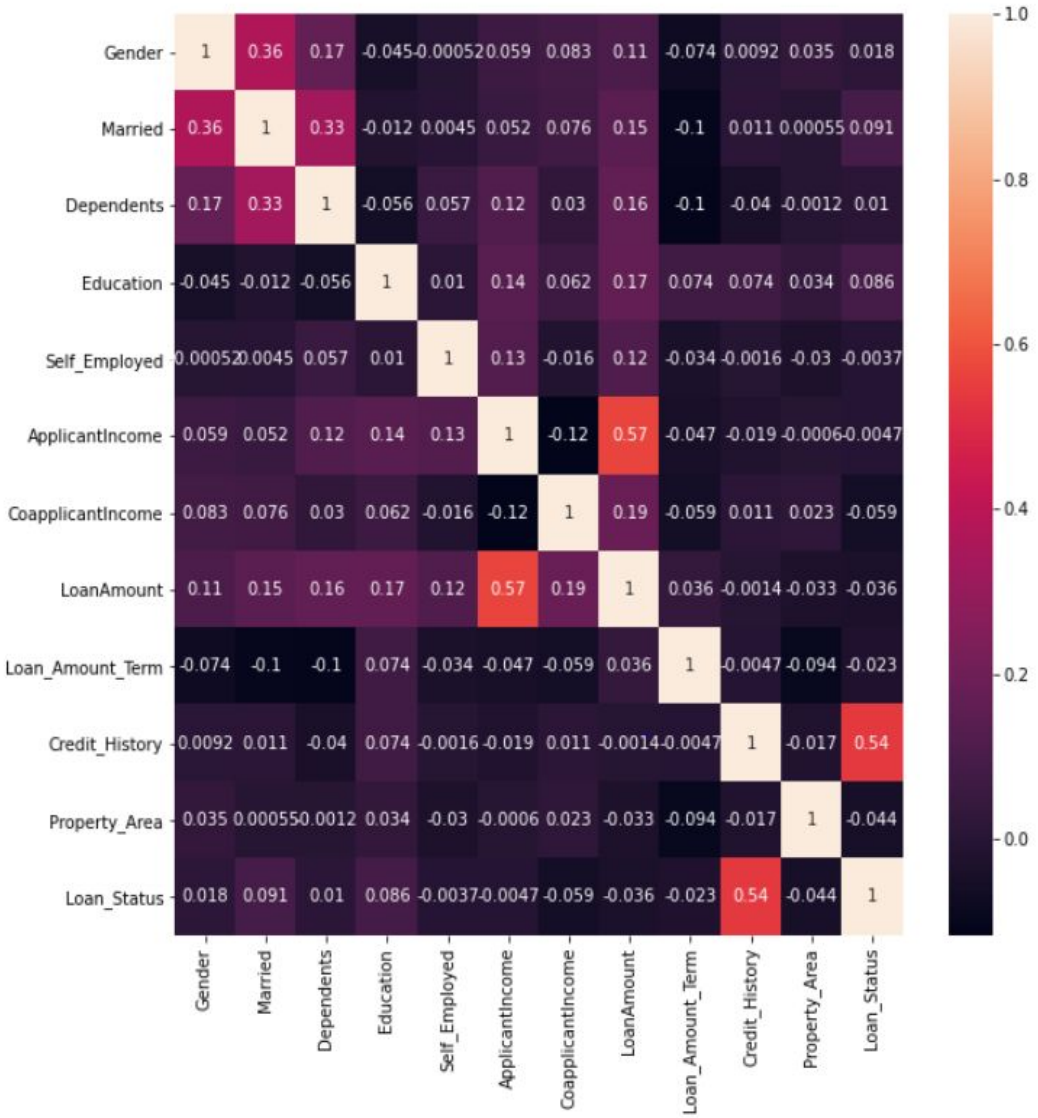


Figure 4 : HeatMap 2

MAJOR POSITIVE ATTRIBUTES AFFECTING Loan_status (target value):

Sr. No	**Attribute**	**Description**
1.	Credit_History	Describes record of borrower's responsible repayment of debts

CORRELATED ATTRIBUTES

The attributes that are highly correlated to each other are as follows. These attributes are positively dependent on each other

Sr No.	Attribute 1	Attribute 2	Description
1.	LoanAmount	ApplicantIncome	Describes the loan amount and income of the applicant respectively.
2.	Gender	Married	Describes the gender and marital status of the applicant respectively.

- From the heatmap we can see that certain attributes like Gender,Dependents,Self_Employed,CoapplicantIncome have relatively less correlation with the target attribute. So we drop these columns.
- We further split the data set into training data(80%) and testing data (20%).
- Then we apply feature scaling to the training data to further improve model performance.

Feature Selection:

From the above heatmap, we have selected the following important attributes:

1. Credit_History
2. Education
3. ApplicantIncome
4. LoanAmount
5. Loan_Amount_Term
6. Married
7. Property_Area

2.2.5 Advantages of the Data set:

1. Considers real life scenarios of Marriage and Dependency.
2. Also, it considers the previous Credit history of the person.

2.2.6 Disadvantages of the Data set:

1. The data set has relatively few values,as a result we can only achieve accuracy only upto a certain level.
2. There are only a few attributes which have strong correlation with the target attribute.

2.2.7 Performance Report:

Sr No	**Model**	**Accuracy**	**Precision**	**Recall**	**F-1 Score**
1	Logistic Regression	0.80487	0.83316	0.804878	0.777847
2	Random Forest	0.76422	0.824735	0.7642276	0.7118120
3	Decision Tree	0.788617	0.7989039	0.7886178	0.7634164
4	Naive Bayes	0.796747	0.815176	0.7967479	0.7705985
5	KNN	0.78048	0.7840871	0.780487	0.7562943

2.2.8 Comparison of Different models Applied:

1. Logistic regression performs better than the rest of the models.
2. Random forest doesn't give much accuracy with this dataset if we use n_eliminator=30
3. If n_eliminator = 75-100, Random Forest gives similar accuracy as Logistic Regression.
4. Thus, **Logistic Regression** seems to be performing better than other models as it gives higher accuracy and higher recall indicating less number of false negatives.

Chapter 3

References

1. **Dataset 1:**
   https://www.kaggle.com/jannesklaas/model-trap

2. **Dataset 2:**
   https://www.kaggle.com/burak3ergun/loan-data-set

3. **Reference Paper:**
   Loan Default Forecasting Using Data Mining
   2020 International Conference for Emerging Technology (INCET)
   Belgaum, India, 5-7 June 2020

4. **Machine learning model to predict loan default**
   https://medium.com/@pankajgurbani01/machine-learning-model-to-predict-loan-default-9c33f87e1a38

5. Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA)
   Article in Journal of Computational and Theoretical Nanoscience · August 2019

6. Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications
   New Zealand Journal of Computer-Human Interaction ZJCHI 2,2 (2017)
   Zakaria, Dmitriy

7. Loan Credibility Prediction System Based on Decision Tree Algorithm
   28-09-2015 SCMS School of Technology and Management Cochin, Kerala, India