LOAN ELIGIBILITY PREDICTION

A PROJECT REPORT

Submitted by

BADIGI SATHWIKA [Reg No: RA2112704010032]

Under the Guidance of

Dr. Kalpana A V
(Assistant Professor, Department of Data Science and Business Systems)

In partial fulfillment of the Requirements for the Degree of

MASTERS OF TECHNOLOGY (INTEGRATED)
COMPUTER SCIENCE AND BUSINESS SYSTEMS

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

NOVEMBER 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this project report titled **“LOAN ELIGIBILITY PREDICTION”** is the bonafide work of **“BADIGI SATHWIKA [Reg No: RA2112704010032]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

Dr. Kalpana A V	Dr.M.Lakshmi
GUIDE	**HEAD OFTHEDEPARTMENT**
Assistant Professor	Dept. of DSBS
Dept. of DSBS	

Signature of Internal Examiner Signature of External Examiner

ABSTRACT

Banks are making major part of profits through loans. Though lot of people are applying for loans. It’s hard to select the genuine applicant, who will repay the loan. While doing the process manually, lot of misconception may happen to select the genuine applicant. Therefore we are developing loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicant. The time period for the sanction of loan will be drastically reduced. In this paper we are predicting the loan data by using some machine learning algorithms that is Decision Tree.

ACKNOWLEDGEMENTS

We express our humble gratitude to **Dr C. Muthamizhchelvan**, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr T.V.Gopal**, for his invaluable support.

We wish to thank **Dr Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department**, Dr M. Lakshmi** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We want to convey our thanks to our program coordinators **Dr.E.Sasikala,** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her inputs during the project reviews and support.

We register our immeasurable thanks to our Faculty Advisor**, Dr.K.Shantha Kumari,** Assistant Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Dr. Kalpana A V,** Assistant Professor, Department of Data Science and Business Systems, for providing me with an opportunity to pursue my project under his mentorship. He provided us with the freedom and support to explore the research topics of our interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Data Science and Business Systems staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

Badigi Sathwika

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
IOT	Internet Of Things
ML	Machine learning

LIST OF SYMBOLS

^	Conjunction

CHAPTER 1

INTRODUCTION

1.1 GENERAL

A loan is the core business part of banks. The main portion the bank's profit is directly come from the profit earned from the loans. Though bank approves loan after a regress process of verification and testimonial but still there's no surety whether the chosen hopeful is the right hopeful or not. This process takes fresh time while doing it manually. We can prophesy whether that particular hopeful is safe or not and the whole process of testimonial is automated by machine literacy style. Loan Prognostic is really helpful for retainer of banks as well as for the hopeful also.

Bank employees check the details of applicant manually and give the loan to eligible applicant. Checking the details of all applicants takes lot of time. The artificial neural network model for predict the credit risk of a bank. The Feed- forward back propagation neural network is used to forecast the credit default. The method in which two or more classifiers are combined together to produce a ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique. The process of classifiers is to improve the performance of the data and it gives better efficiency. In this work, the authors describe various ensemble techniques for binary classification and also for multi class classification. The new technique that is described by the authors for ensemble is COB which gives effective performance of classification but it also compromised with noise and outlier data of classification. Finally they concluded that the ensemble based algorithm improves the results for training data set.

1.2 MOTIVATION

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval.

1.3 PROBLEM STATEMENT

To deal with the problem, we developed automatic loan prediction using machine learning techniques. We will train the machine with previous dataset. so machine can analyse and understand the process . Then machine will check for eligible applicant and give us result.

Advantages

- Time period for loan sanctioning will be reduced.
- Whole process will be automated , so human error will be avoided
- Eligible applicant will be sanctioned loan without any delay.

CHAPTER 2

LITERATURE REVIEW

PAPERS REFERRED:

TITLE 1: Improving Information Quality in Loan Approval Processes for Fair Lending and Fair Pricing

AUTHOR: M. Cary Collins

YEAR: 2013

DESCRIPTION: Bank data management on loan approval processes has great room for improvements of information quality and data problems prevention especially with regards to fair lending and fair pricing practices. They first reviewed briefly typical data collection protocols deployed at many financial institutions for loan approval and loan pricing. Federal regulations mandate portions of these data protocols. While discussing the data capture and analysis for fair lending, they illustrated some initial key steps currently needed for improving information quality to all parties involved.

TITLE 2: Loan Credibility Prediction System Based on Decision Tree Algorithm

AUTHOR: Sivasree M S, Rekha Sunny T

YEAR: 2015

DESCRIPTION: Data mining techniques are becoming very popular nowadays because of the wide availability of huge quantity of data and the need for transforming such data into knowledge. Data mining techniques are implemented in various domains such as retail industry, biological data analysis, intrusion detection, telecommunication industry and other scientific applications. Techniques of data mining are also be used in the banking industry which help them compete in the market well equipped. In this paper, they introduced a prediction model for the bankers that will help them predict the credible customers who have applied for a loan. Decision Tree Algorithm is being applied to predict the attributes relevant for credibility. A prototype of the model has been described in this paper which can be used by the organizations for making the right decisions to approve or reject the loan request from the customers.

TITLE 3: Loan Approval Prediction based on Machine Learning Approach

AUTHOR: Kumar Arun, Garg Ishan, Kaur Sanmeet

YEAR: 2016

DESCRIPTION: With the enhancement in the banking sector, lots of people apply for bank loans but the bank has its limited assets which it grants to only limited people , so finding out to whom the loan can be granted is a typical process for the banks. So, in this paper , they tried to reduce this risk by selecting the safe person so as to save lots of bank efforts and assets. It was done by mining the previous records of the people to whom the loan was granted before and on the basis of these records the machine was trained using the machine learning model which gave the most accurate result. The main goal of this paper is to predict if loan assignment to a specific person will be safe or not. This paper has into four sections (i) Collection of data (ii) Comparing the machine learning models on collected data (iii) Training the system on most promising model (iv) Testing the system.

CHAPTER 3

OBJECTIVES

We will look into some of the current challenges faced to determine loan eligibility and how AI / machine learning can be used to address those challenges. The following are some of the current challenges:

- **Credit Scoring:** One of the most important factors in loan eligibility is credit score. Credit scoring is a statistical method of assessing the credit risk of a potential or existing customer. A borrower’s credit score is a numerical representation of their creditworthiness. In order to get a accurate picture of the borrower’s creditworthiness, lenders will often look at multiple factors such as payment history, credit utilization, outstanding debt, length of credit history, credit mix, new credit inquiries, employment history, current financial status, etc. Payment history is the most heavily weighted factor in most credit scoring models, followed by credit utilization. Other factors such as length of credit history, credit mix, and new credit inquiries may also be considered in some models. However, manually assessing all of these factors can be time-consuming and expensive. Additionally, human lenders are often biased, which can lead to inaccurate decision-making. Machine learning can be used to automate the credit scoring process. Credit scoring models are mathematical algorithms that lenders use to forecasting an individual’s credit risk. Credit risk is the probability of suffering a loss due to a borrower’s non-payment of a loan. Credit scoring models are generally based on machine learning algorithms that analyze different features as listed before in order to predict their future behavior. The accuracy of these predictions is important in order to make sound lending decisions. There are many different types of credit scoring models, but the most common ones used by lenders are the FICO score and the VantageScore. Credit scoring models are typically developed using historical data from a large population of consumers. Feature selection and engineering are important steps in the development of a credit scoring model. Once a model is developed, it can be used to score new consumers in real-time using machine learning. This is not only more efficient than manual credit scoring, but it also leads to more accurate decisions. Credit scoring models are constantly being updated and refined as more data becomes available.

- **Income Verification:** Another important factor in loan eligibility is income. It is a process that lenders use to confirm that an applicant has the financial means to repay a loan. Lenders need to verify the borrower's income in order to assess their ability to repay the loan. This process can be time-consuming, as it typically requires reviewing tax returns or pay stubs. Additionally, borrowers may falsify their income in order to qualify for a loan. Also, income verification can be a challenge, particularly for those who are self-employed or have multiple sources of income. The information related to income is typically provided in the form of tax returns, pay stubs, or bank statements. And, it is really cumbersome and time taking for humans to do a great job given the volume of request for loan eligibility. This is where machine learning / AI comes to the rescue. Machine learning can be used to automate the income verification process. By using data from previous loan applications, tax returns, bank statements, machine learning models can learn to identify patterns that are predictive of loan default. These patterns can then be used to automatically verify the income of new loan applicants. Classification models can be used to classify whether the income got verified or move the application to exception workflow which can then be processed by the humans. This is not only more efficient than manual income verification, but it also leads to more accurate decisions.
- **How much money to lend and at what terms:** Another challenge related to loan eligibility is to determine how much money to lend and at what terms. This decision is typically made by looking at the borrower's credit score and income. However, there are other factors that can be used to assess loan amount and terms. For example, the loan purpose (e.g., buying a car versus starting a business) can be used to determine loan amount. Additionally, the borrower's employment history can be used to assess loan terms. Machine learning can be used to automate the decision-making process for loan amount and terms. By using data from previous loan applications, machine learning models can learn to identify patterns that are predictive of loan default. These patterns can then be used to automatically determine loan amount and terms for new loan applicants.
- **Handling unstructured data:** Another challenge to determine loan eligibility is that data is often unstructured. This can make it difficult to extract the information that is

needed to assess loan eligibility. For example, a borrower’s credit history may be spread out across multiple sources, such as their credit report, public records, and tax returns. Machine learning can be used to automatically extract this information from unstructured data sources. By using data from previous loan applications, machine learning models can learn to identify patterns that are predictive of loan default. These patterns can then be used to automatically extract the information needed to assess the loan eligibility of new applicants.

CHAPTER 4

BLOCK DIAGRAM

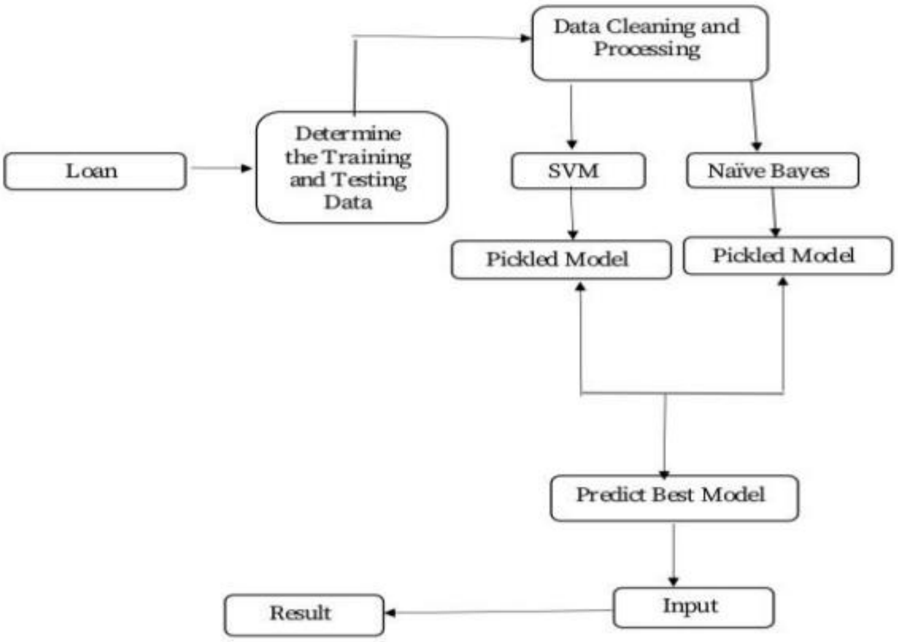
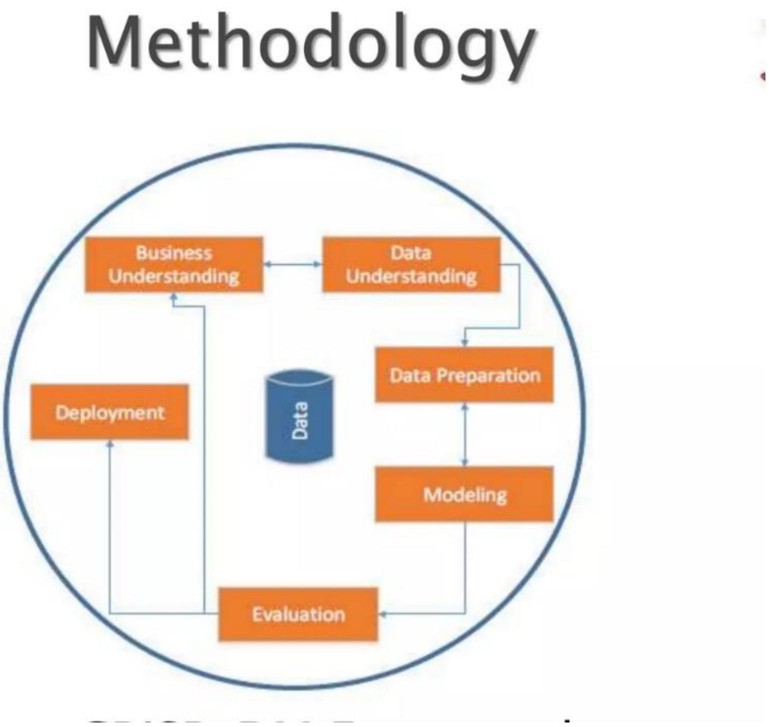


Fig -1: Loan Prediction Architecture



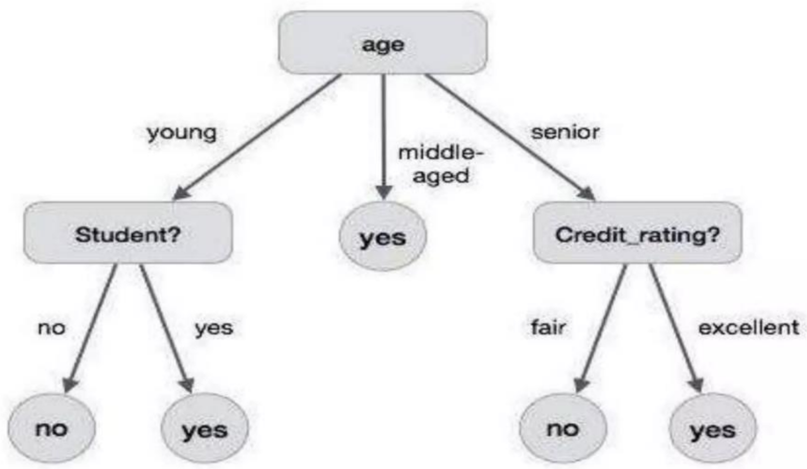
CHAPTER 5

WORKING MODEL

LOAN ELIGIBILITY PREDICTION TECHNIQUES

5.1 Decision trees

A supervised learning methodology, graphical representation of possible solutions to a choice based on certain situations and it is a tree-structured classifier. It starts with a root node where inside nodes represent the features of a dataset, branches symbolize the decision rules and each leaf node represents the result. In a decision tree and they have the purposes of deciding and communicating respectively. A decision tree plainly asks a question and then divides it into sub trees based on the answer. Although DT can solve classification and regression problems, it is most commonly used to solve classification problems. To find the dataset class, the algorithm searches at the top of the tree. It compares the root Trait with the record attribute and follows the offshoot on way to the next node, which it calculates depending on the relation.



Decision Tree Induction

Steps Working Of Decision Tree

In the first phase, start with S, which is the root node and includes the entire dataset. Second, discover the best Trait in the dataset using the Attribute Selection Measure. When the nodes cannot be categorized, in that time the final node is called a foliate node. Based on the labels, the root node is extra subdivided into the decision node and one leaf node. In the end, the node is divided into two leaves (accepted and declined offers).

5.2 Random Forest

Random Forest classifier finds decision trees in a subset of the data and then aggregates their information to that to get the full dataset's predictive power. Rather than relying on a single decision tree. The RF takes the predictions from each tree and forecasts the final output based on the majority votes of forecasts. Using a huge number of trees in the forest improves precision and eliminates the issue of over fitting. It predicts output with high precision, and it runs efficiently even with large datasets. It can also keep accuracy when a large proportion of data is lost. Random Forest can handle both classification and regression tasks. It can handle large datasets with high dimensionality. It improves the model's accuracy and avoids the over fitting problem. We use two-step training techniques in the process of tree-based Random Forest: First, we generate the random forest by mixing N trees together, and then we estimate for each of the trees we generate in the first phase .An ensemble algorithm employs the "random forest" artificial intelligence technique. Because it averts over-fitting by averaging the results, this approach outperforms single decision trees. Random Forest is an ensemble of diverse trees, similar to Gradient Boosted Trees, but unlike GBT, RF tree grow in parallel. Random Forests have a lot of uncorrelated trees. Because various trees are trained in parallel, the overall model diminishes a large number of variances. Random Forest treats each tree as a separate classifier that has been trained on resampled data.

As a result of employing this this learn strategy and divide, the model's overall learning ability is increased .

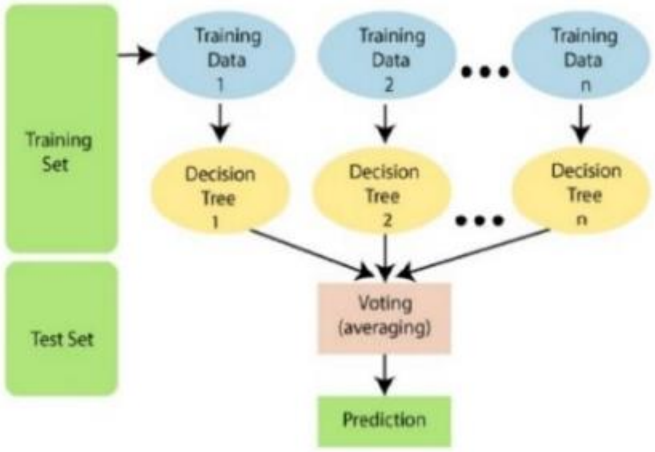


Fig. 3. General structure working of the RF

The Random Forest Working Steps

These steps illustrate Figure 3 above; in the first step, choose (K) as data points at random from the drill set. Second, construct the DT linked with the chosen data points (Subsets). Following that, select the digit (N) for the number of decision trees you wish to construct. Then, duplicate Steps 1 and 2. Finally, discover the predictions of eachdecision tree for new data points and assign the modern data points to the category that receives most votes. Clarify how RF works by using the following scenario: Assume you have a dataset with a variety of fruit images. As outcome, RF classifier will be given this dataset. Each decision tree is given a portion of the dataset to deal with. When a new data point occurs, the Random Forest classifier predicts the conclusion based on the majority of outcomes.

5.3 Logistic Regression

An algorithm that can be used for both regression and classification tasks, but it is most commonly used for classification.' _Logistic Regression is used to predict categorical variables using dependent variables. Consider two classes, and a new data point is to be checked to see which class it belongs to. The algorithms then compute probability values ranging between (0) and (1). Logistic Regression employs a more complex cost function, this cost function is known as the Sigmoid Function or the Logistic Function. LR also does not require independent variables to be linearly related, nor does it require equal variance within each group, making it a less stringent statistical analysis procedure. As a result, logistic regression was used to predict the likelihood of fraudulent credit cards .Clarify the working of LR through the following scenario: The default variable for determining whether a tumor is malignant or not is y=1 (tumor= malignant); the x variable could be a measurement of the tumor, such as

its size. The logistic function converts the x-values of the dataset's various instances into a range of 0 to 1. The tumor is classified as malignant if the probability exceeds 0.5. (As indicated by the horizontal line). As shown in the figure below:

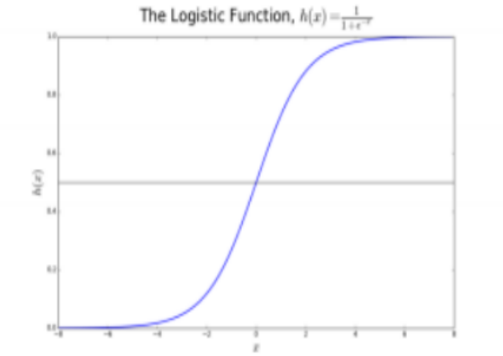


Fig. 4. Example of LR

Libraries for Data Analysis

The models are implemented using Python 3.7 with listed libraries:

Pandas

Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling.

Seaborn

Seaborn is a python library for building graphs to visualise data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data.

CHAPTER 6

PROJECT CODE

6.1 Algorithm

Step 1: Read the dataset.

Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.

Step 4: Feature selection are applied for the proposed models.

Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

Step 6: Then retrieve the best algorithm based on efficiency for the given dataset.

6.2 DATASET INFORMATION

# Dataset Information

A Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

This is a standard supervised classification task.A classification problem where we have to predict whether a loan would be approved or not. Below is the dataset attributes with description.

Variable	**Description**
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Credit history meets guidelines

Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

6.3 Program Code

```
#LOAD THE DATASET

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

train = pd.read_csv('train.csv')

#ANALYSE THE DATASET

train.head()

train.tail()

train.describe()

train.info()

train.isna().sum()

#FILL NULL VALUES

train['Gender'].fillna(
   value=train['Gender'].mode()[0]
)

train['LoanAmount'].fillna(
   value=train['LoanAmount'].median()
)
```

```
#WE WILL CREATE A COPY OF DATAFRAME AND WORK ON IT

df = train.copy()

df.isna().sum()

def fill_na_values(df: pd.DataFrame) -> pd.DataFrame:

    for feature in df:

        if df[feature].isna().sum() > 0:

            if df[feature].dtype == 'object':

                df[feature].fillna(
                    value=df[feature].mode()[0],
                    inplace=True
                )

            else:

                df[feature].fillna(
                    value=df[feature].median(),
                    inplace=True
                )

    return df

df = fill_na_values(df)

df.isna().sum()

import re

def remove_unwanted_characters(value: str) -> int:

    value = re.sub(r'[^0-9]', '', value)

    return int(value)

df['Dependents'] = df['Dependents'].apply(remove_unwanted_characters)
```

```
df['Dependents'] = df['Dependents'].astype(dtype='int64')

#VISUALIZE ALL THE FEATURES


import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

df.drop(
    labels=['Loan_ID'],
    inplace=True,
    axis=1
)

for feature in df:

    if df[feature].dtype == 'object':
        sns.countplot(
            data=df,
            x=feature
        )
        plt.title(f'Count plot of {feature}')
        plt.show()
    else:
        sns.displot(
            data=df,
            x=feature,
            kde=True
        )
        plt.title(f'Distribution plot of {feature}')
        plt.show()

#ENCODING ALL THE CATEGORICAL FEATURES

from sklearn.preprocessing import LabelEncoder


for feature in df:

    le = LabelEncoder()
    if df[feature].dtype == 'object':

        df[feature] = le.fit_transform(df[feature])


df.head()

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score

import numpy as np

X = df.drop(
    labels=['Loan_Status'],
    axis=1
).values

y = df['Loan_Status'].values

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=2022
)

def run_model(model) -> None:

    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)

    score = f1_score(y_test, y_pred)

    print(f'F1 Score - {score}')


lr = LogisticRegression(max_iter=500)
```

```
run_model(lr)


from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier()

run_model(rfc)


df.head()
```

CHAPTER 7

REQUIREMENTS

7.1Hardware requirements

Processor : Pentium i3 or higher.

RAM : 4 GB or higher.

Hard Disk Drive : 20 GB (free).

Peripheral Devices : Monitor, Mouse and Keyboard

7.2 Software Requirements

Operating system : Windows 8/11.

IDE Tool : Jupyter notebook

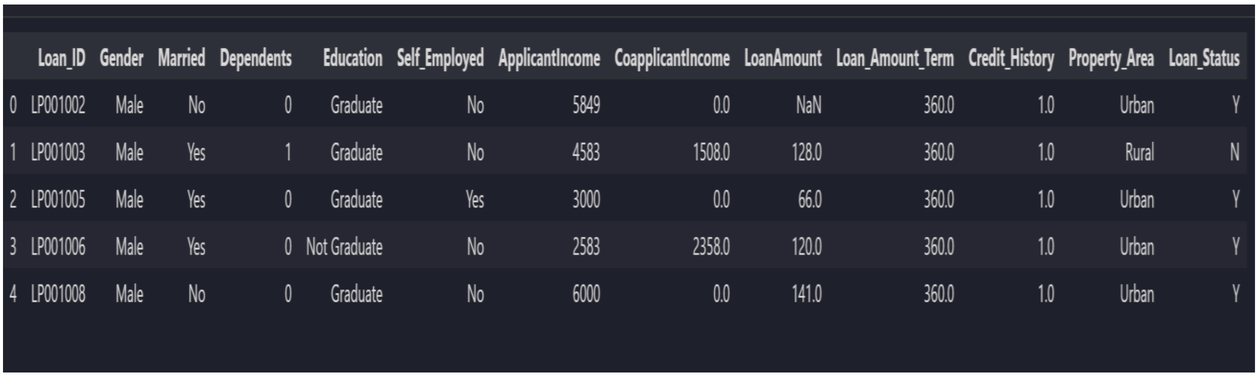
Coding Language : Python

APIs : Numpy, Pandas, Matplotlib

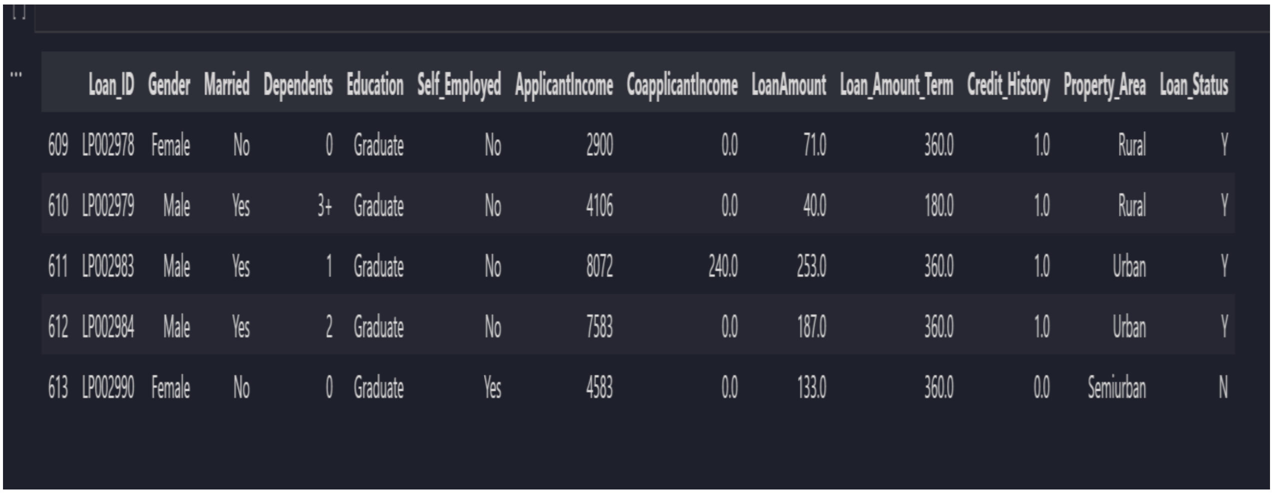
Dataset: Kaggle

CHAPTER 8

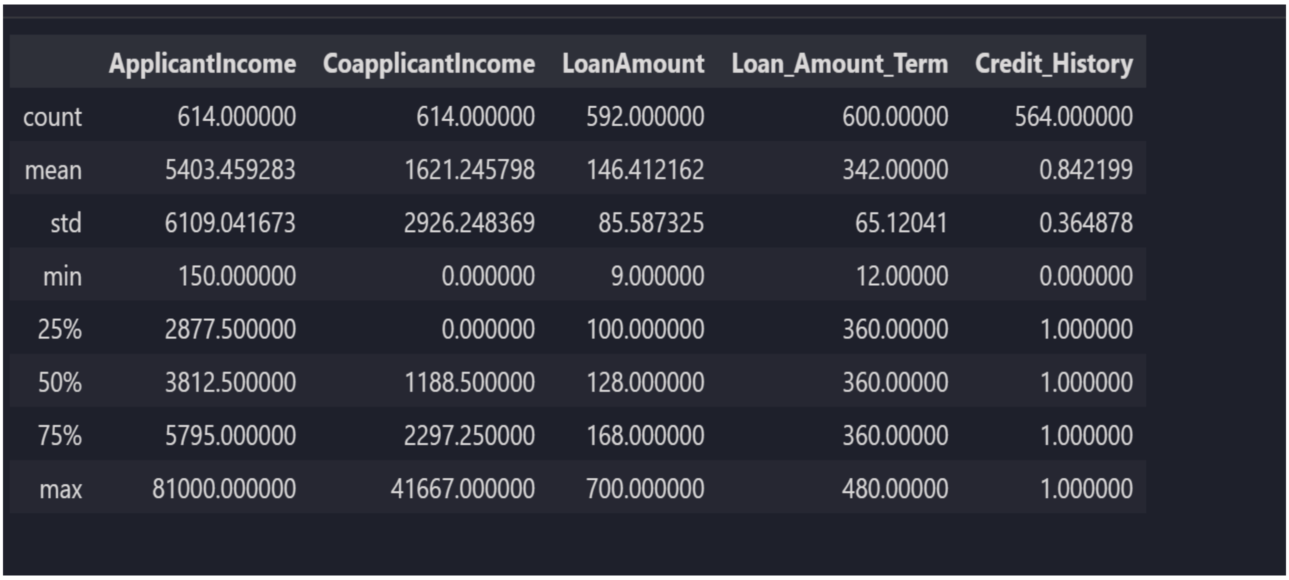
PROJECT FINDINGS



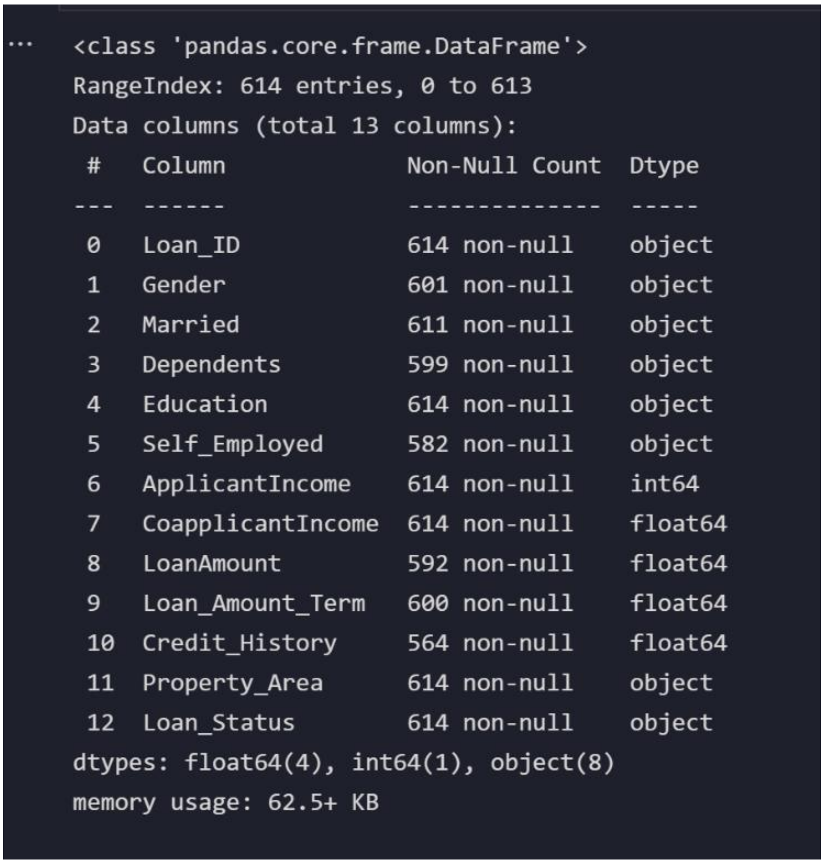
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y



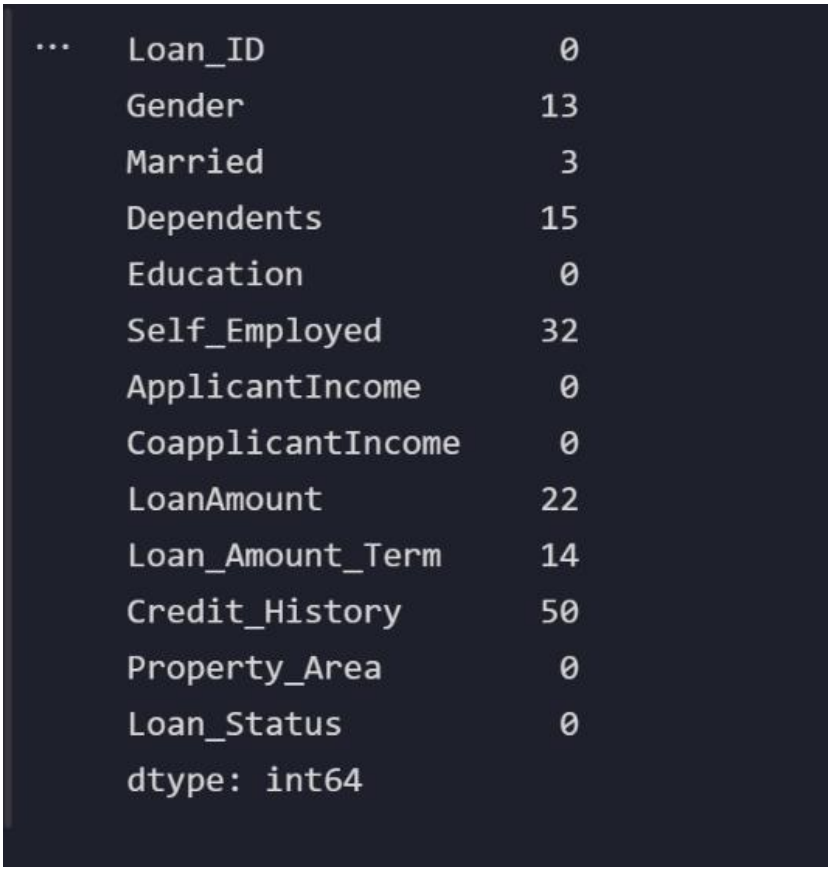
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
609	LP002978	Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0	Rural	Y
610	LP002979	Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0	Rural	Y
611	LP002983	Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0	Urban	Y
612	LP002984	Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0	Urban	Y
613	LP002990	Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0	Semiurban	N



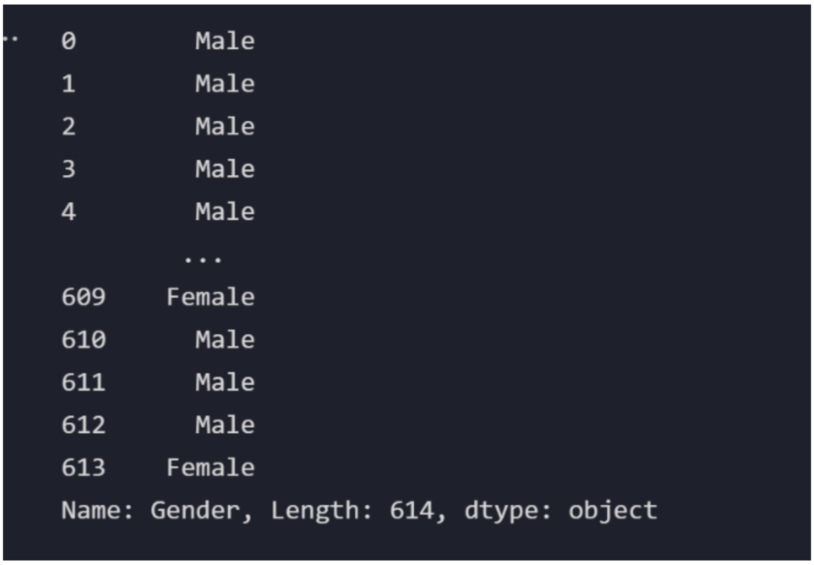
	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000



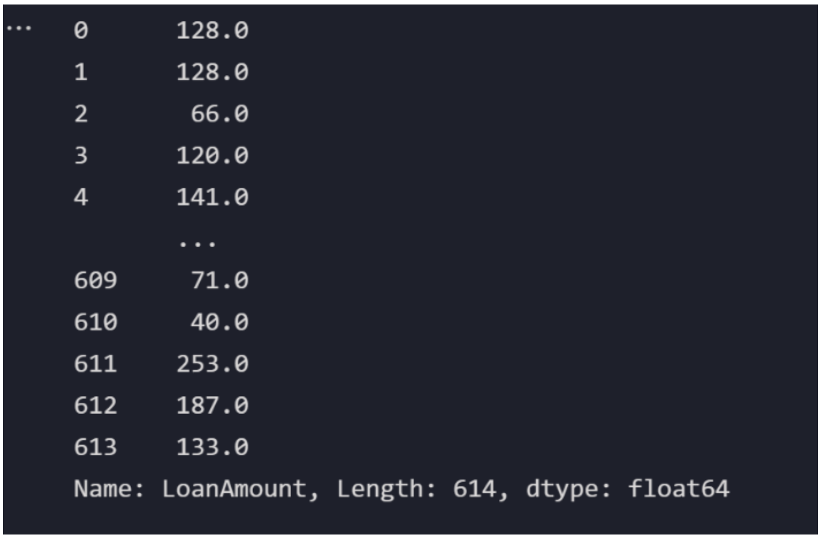
```
...  <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 614 entries, 0 to 613
     Data columns (total 13 columns):
      #   Column             Non-Null Count  Dtype
     ---  ------             --------------  -----
      0   Loan_ID            614 non-null    object
      1   Gender             601 non-null    object
      2   Married            611 non-null    object
      3   Dependents         599 non-null    object
      4   Education          614 non-null    object
      5   Self_Employed      582 non-null    object
      6   ApplicantIncome    614 non-null    int64
      7   CoapplicantIncome  614 non-null    float64
      8   LoanAmount         592 non-null    float64
      9   Loan_Amount_Term   600 non-null    float64
      10  Credit_History     564 non-null    float64
      11  Property_Area      614 non-null    object
      12  Loan_Status        614 non-null    object
     dtypes: float64(4), int64(1), object(8)
     memory usage: 62.5+ KB
```



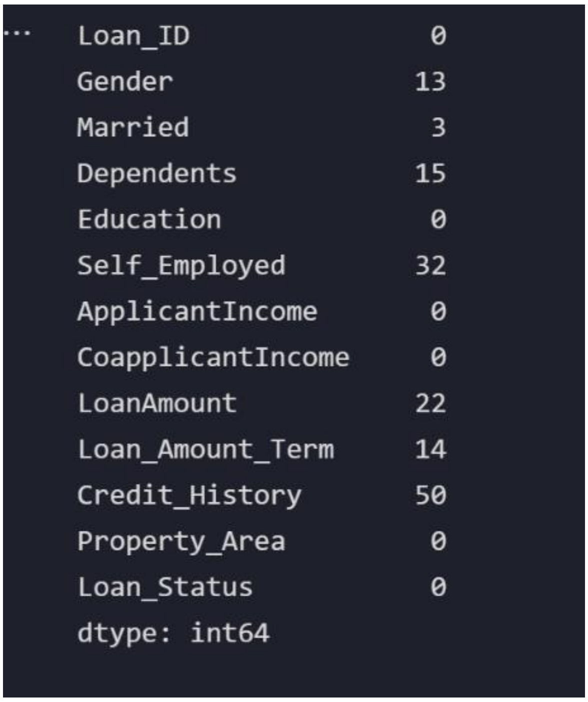
```
...  Loan_ID               0
     Gender               13
     Married               3
     Dependents           15
     Education             0
     Self_Employed        32
     ApplicantIncome       0
     CoapplicantIncome     0
     LoanAmount           22
     Loan_Amount_Term     14
     Credit_History       50
     Property_Area         0
     Loan_Status           0
     dtype: int64
```



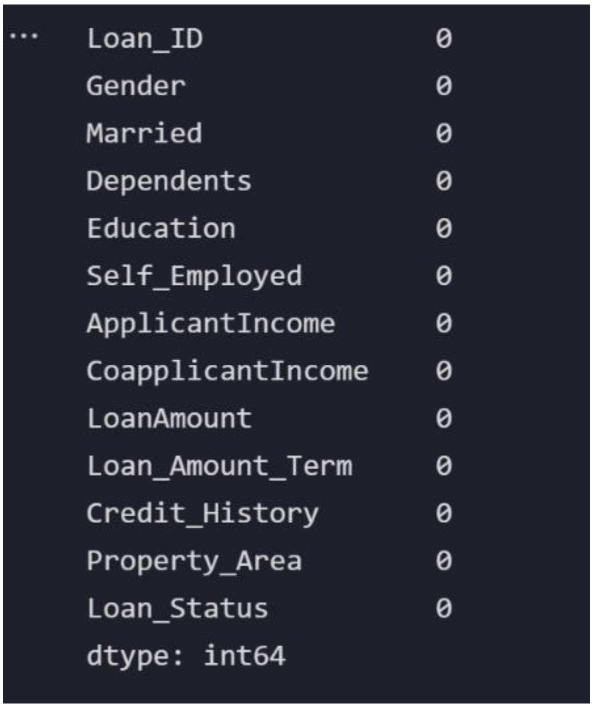
```
..  0         Male
    1         Male
    2         Male
    3         Male
    4         Male
             ...
    609     Female
    610       Male
    611       Male
    612       Male
    613     Female
    Name: Gender, Length: 614, dtype: object
```



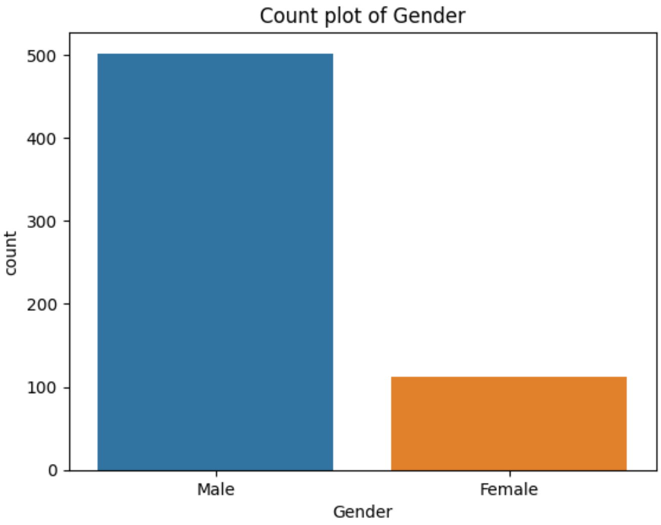
```
...  0      128.0
     1      128.0
     2       66.0
     3      120.0
     4      141.0
            ...
     609     71.0
     610     40.0
     611    253.0
     612    187.0
     613    133.0
     Name: LoanAmount, Length: 614, dtype: float64
```

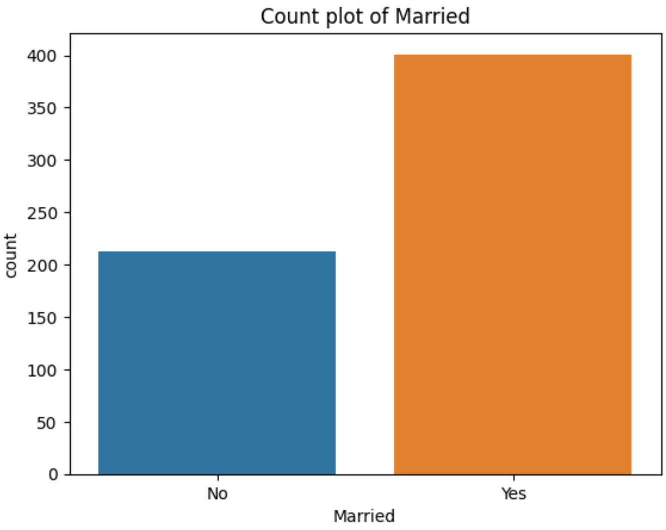


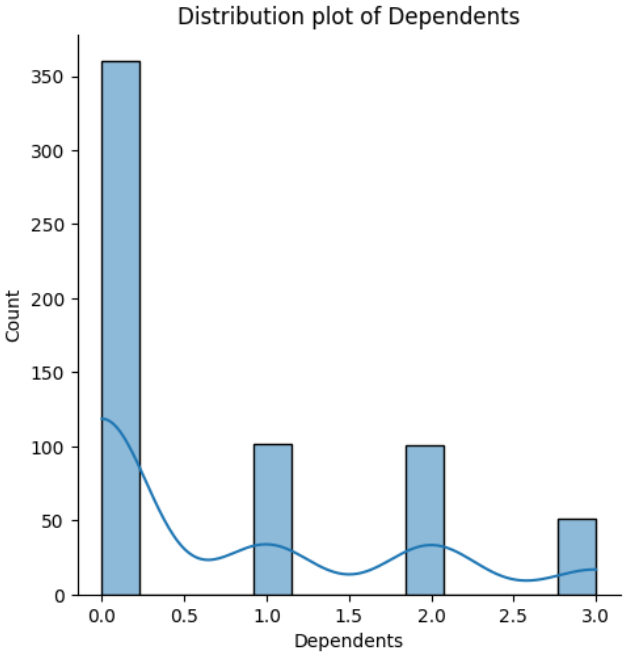
```
...  Loan_ID               0
     Gender               13
     Married               3
     Dependents           15
     Education             0
     Self_Employed        32
     ApplicantIncome       0
     CoapplicantIncome     0
     LoanAmount           22
     Loan_Amount_Term     14
     Credit_History       50
     Property_Area         0
     Loan_Status           0
     dtype: int64
```

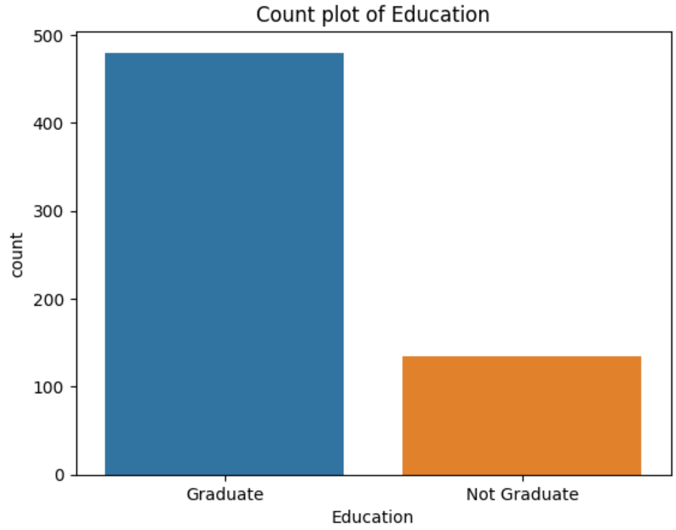


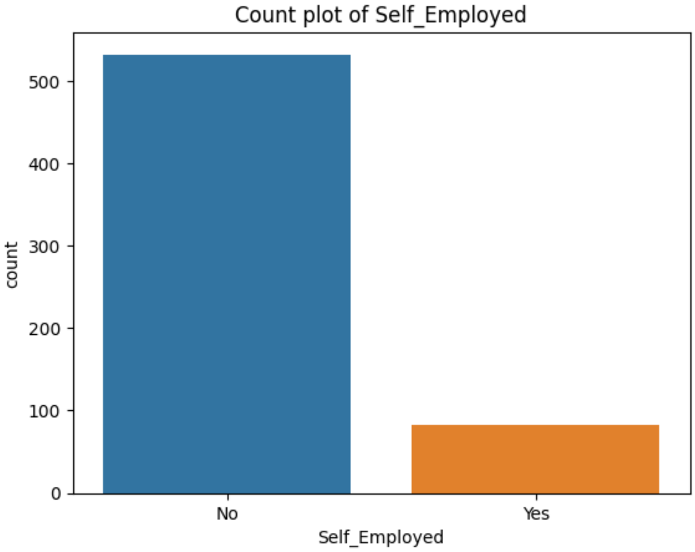
```
...  Loan_ID              0
     Gender               0
     Married              0
     Dependents           0
     Education            0
     Self_Employed        0
     ApplicantIncome      0
     CoapplicantIncome    0
     LoanAmount           0
     Loan_Amount_Term     0
     Credit_History       0
     Property_Area        0
     Loan_Status          0
     dtype: int64
```

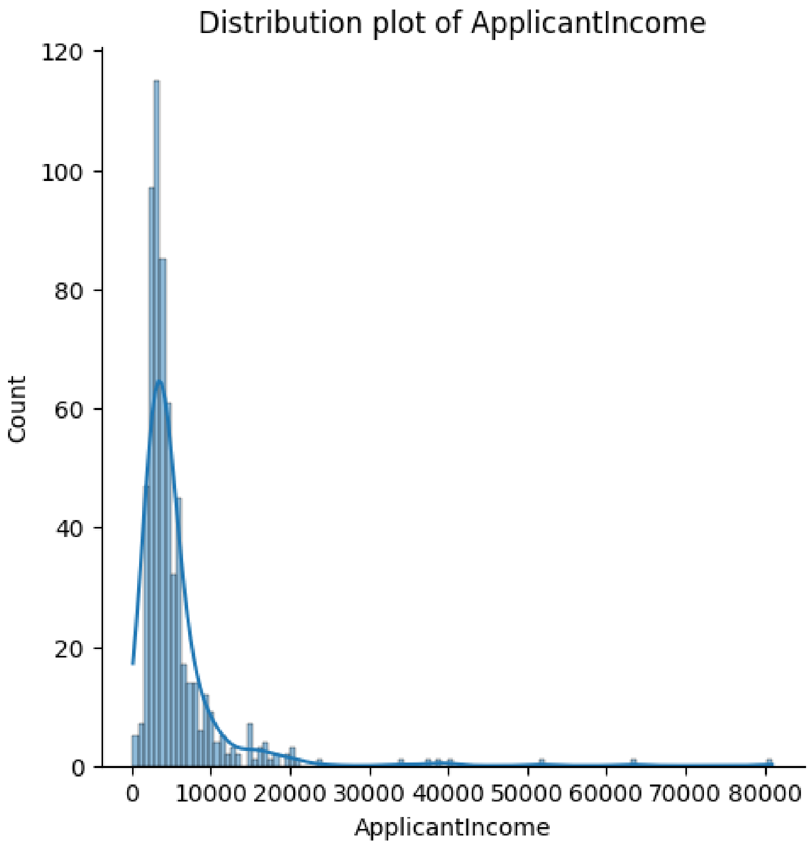


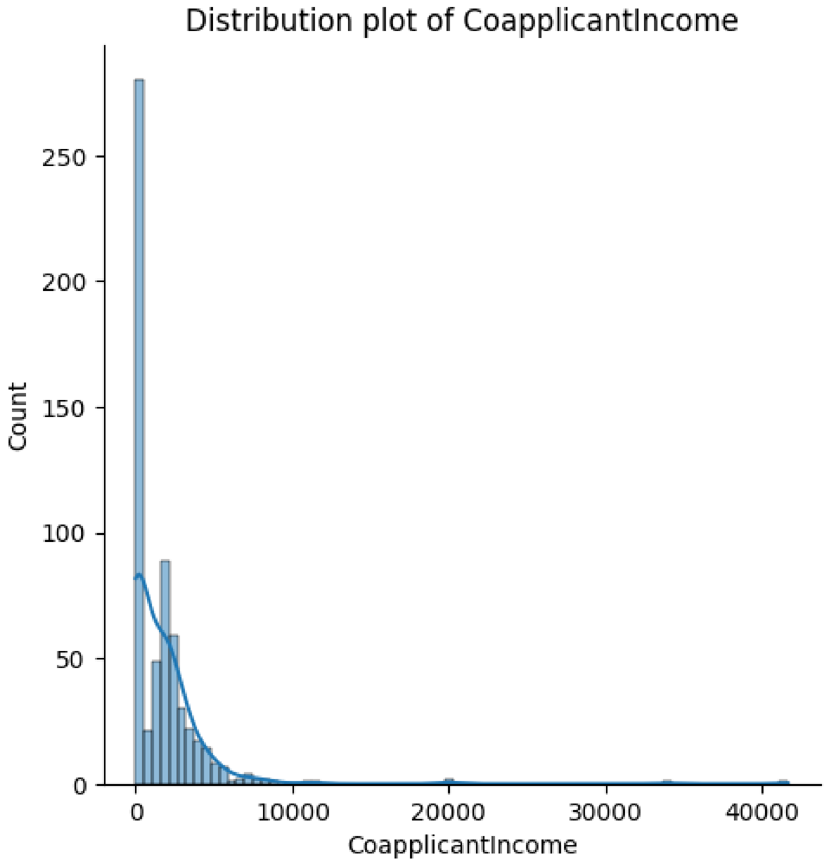


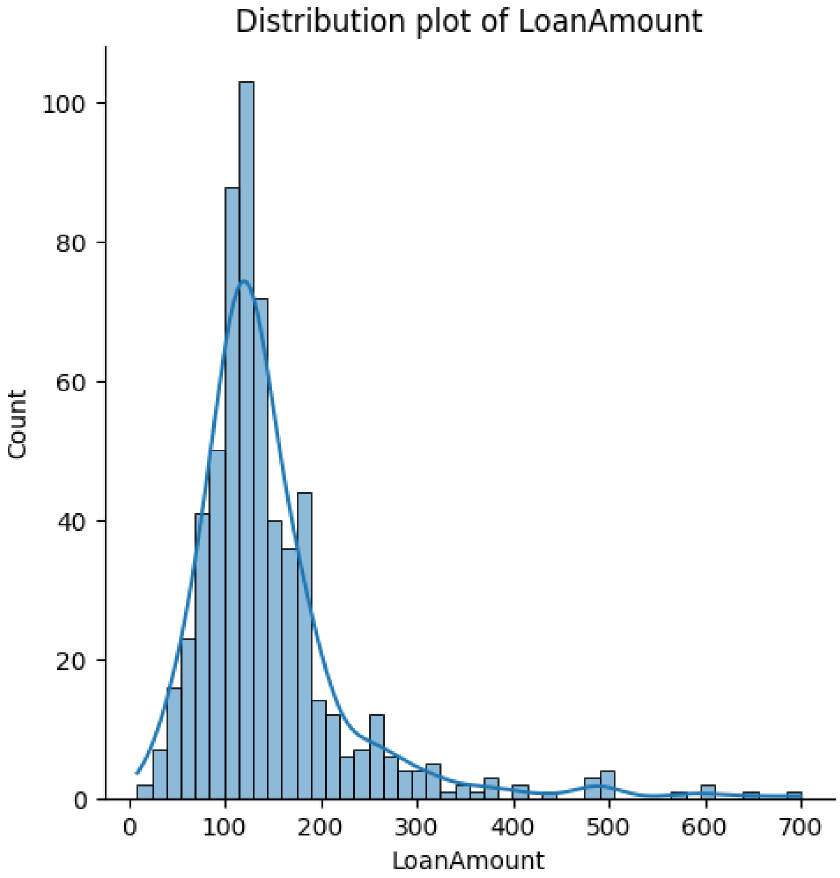

Distribution plot of Dependents
Count
350
300
250
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Dependents

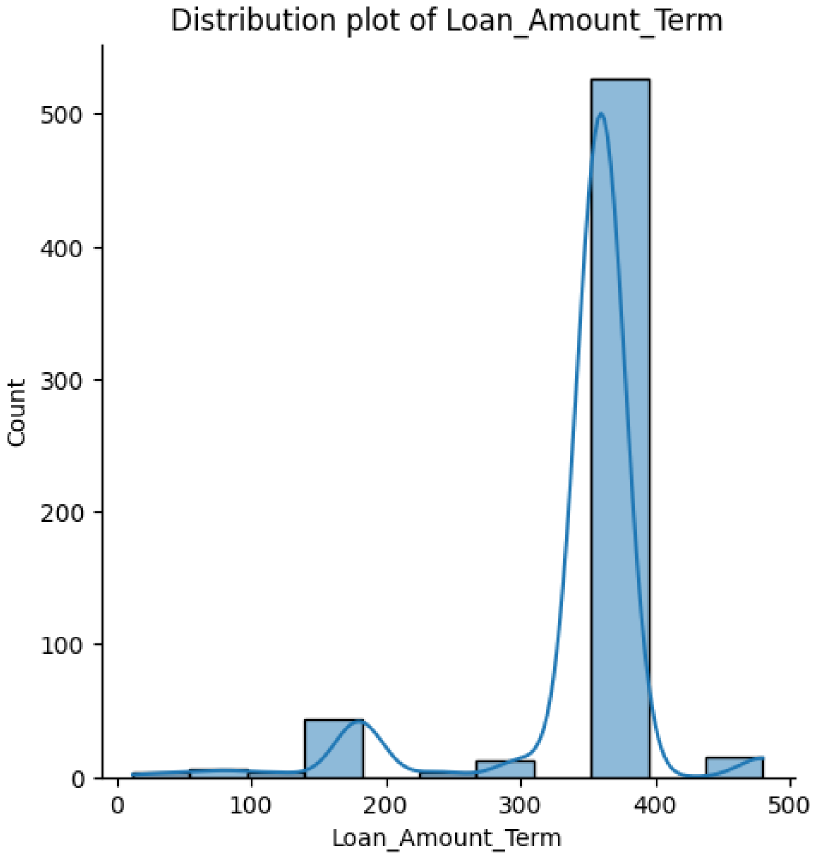

Count plot of Education
count
500
400
300
200
100
0
Graduate
Not Graduate
Education

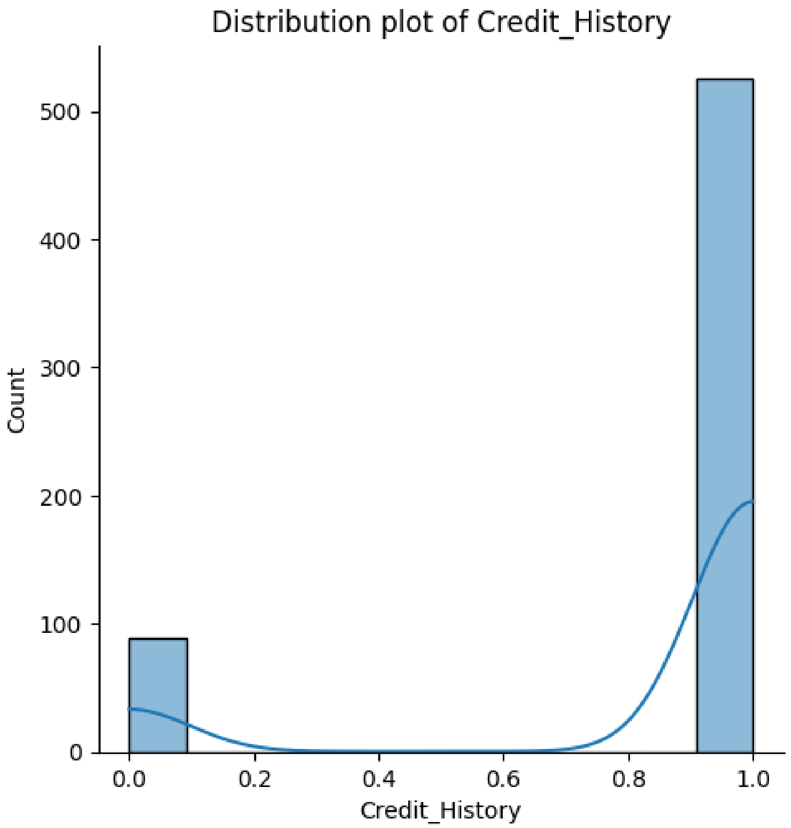

Count plot of Self_Employed
count
500
400
300
200
100
0
No
Yes
Self_Employed

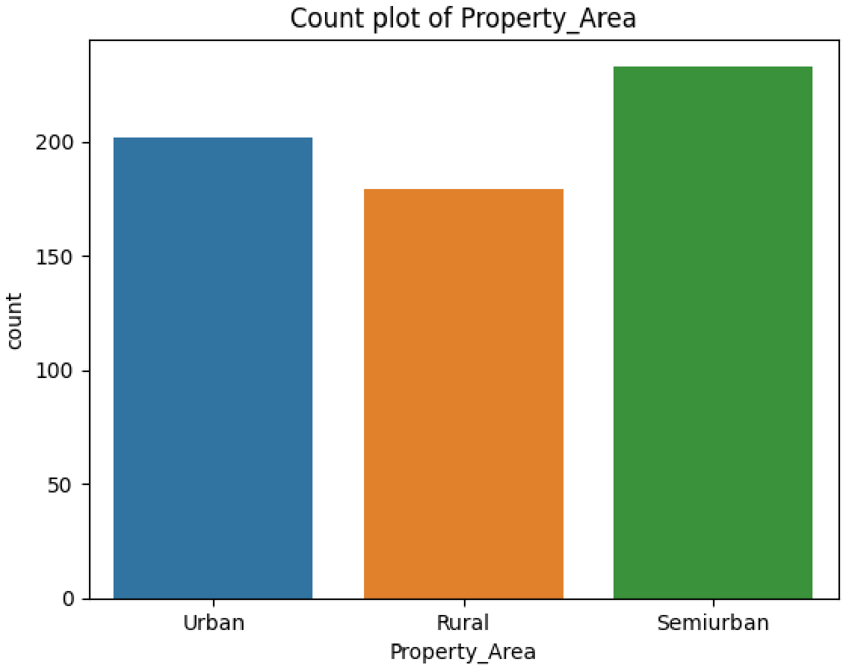

Distribution plot of ApplicantIncome
Count
0
20
40
60
80
100
120
0
10000
20000
30000
40000
50000
60000
70000
80000
ApplicantIncome

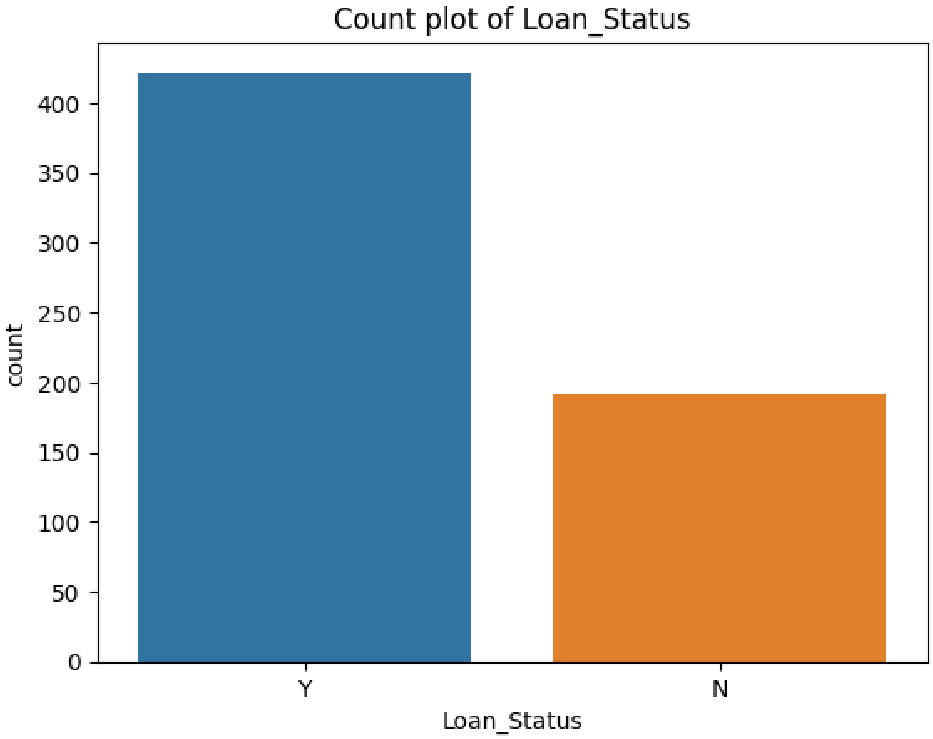

Distribution plot of CoapplicantIncome
Count
0
50
100
150
200
250
0
10000
20000
30000
40000
CoapplicantIncome


Distribution plot of LoanAmount
Count
0
20
40
60
80
100
0
100
200
300
400
500
600
700
LoanAmount


Distribution plot of Loan_Amount_Term
Count
0
100
200
300
400
500
0
100
200
300
400
500
Loan_Amount_Term


Distribution plot of Credit_History
Count
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Credit_History


Count plot of Property_Area
count
0
50
100
150
200
Urban
Rural
Semiurban
Property_Area


Count plot of Loan_Status
count
400
350
300
250
200
150
100
50
0
Y
N
Loan_Status

CHAPTER 9

CONCLUSION

In this work, customer’s data is analyzed using machine learning techniques. Python PYSPARK machine learning package is used to train the model and publicly available loan dataset is collected from Kaggle for implementation. Machine learning models decision tree, random forest, support vector machine, k-nearest neighbor, and decision tree with adaboost are implemented in this work. The model’s results are analyzed in terms of accuracy parameter, and performance is represented in graphical format.

CHAPTER 10

FUTURE ENHANCEMENTS

The analysis of implemented models reveals that machine learning models are suitable for customer loan elig ibility prediction and identified that the existing models are not giving 100% accuracy. Means there is a need to design an innovative model for customer loan eligibility prediction over banking sector. In this connection, we are proposing a deep learning model for customer loan eligibility prediction for future work.

PAPER PUBLICATION STATUS

Publication process not yet started