CAPSTONE PROJECT -III

Home loan Prediction

U. V. Patel College of Engineering

Internal guide:
Prof. Menka Patel

Prepared By:
Vamsi Manikanta Mekala[20012531039]
Yash Rajesh Modi[21012532001]

B. Tech Semester-VII
(Computer Engineering- Artificial Intelligence)
Submitted to,
Department of Computer Engineering-Artificial Intelligence
U.V. Patel College of Engineering
Ganpat University, Kherva - 384012

U.V. PATEL COLLEGE OF ENGINEERING

09/12/2023

CERTIFICATE

TO WHOM SO EVER IT MAY CONCERN

This is to certify that Mr. MEKALA VAMSI MANIKANTA student of **B.Tech. Semester-VII (Computer Engineering - Artificial Intelligence)** has completed his full semester on site project work titled **“HOME LOAN PREDICTION”** satisfactorily in partial fulfilment of the requirement of Bachelor of Technology degree of Computer Engineering of Ganpat University, Kherva, Mehsana in the year 2023-2024 .

Prof. Menka Patel	**Dr. Paresh M. Solanki**
College Project Guide	**Head,ComputerEngineering**

U.V. PATEL COLLEGE OF ENGINEERING

09/12/2023

CERTIFICATE

TO WHOM SO EVER IT MAY CONCERN

This is to certify that Mr. **YASH RAJESH MODI** student of **B.Tech. Semester-VII (Computer Engineering – Artificial Intelligence)** has completed his full semester on site project work titled **“HOME LOAN PREDICTION”** satisfactorily in partial fulfilment of the requirement of Bachelor of Technology degree of Computer Engineering of Ganpat University, Kherva, Mehsana in the year 2023-2024 .

Prof. Menka Patel
College Project Guide

Dr. Paresh M. Solanki
Head,ComputerEngineering

ACKNOWLEDGEMENT

This satisfaction that successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation it made it possible, whose constant guidance and encouragement crown all efforts with success. I am grateful to our guide Prof. Menka Patel Mam for the guidance, inspiration and constructive suggestions that helpful us in the preparation of this project. I also thank my colleagues who have helped in successful completion of the project.

Table of Contents

Vision and Mission of Ganpat University	
Vision	It shall be the constant endeavour of Ganpat University to meet the educational needs of the youth in the areas of professional studies and provide state-of the art learning opportunities along with inculcation of values of commitment and uprightness.
Mission	Seek, search and offer programs that lead to symbiotic emergence of 'academic excellence' and 'industrial relevance' in education and research.

Vision and Mission of Computer Engineering Department	
Vision	Department aims to achieve its recognition as a leading contributor in the area of technical education of computer engineering by practicing latest principles, tools and technologies to cope with current and future challenges and hence contributing to global welfare.
Mission	1. To educate and inculcate strong fundamentals of science and computer engineering through best teaching learning practices. 2. To impart high quality education to acquire skills to conduct research and solve complex problems through modern tools, technologies and innovative practices. 3. Enabling youth for employability, social upliftment, following good moral practices and professional ethics. 4. Preparing youth for contributing in advancements of technology and society. 5. Encouraging students to be adaptive, courageous and life-long learners.

ABSTRACT

In India, the number of people or organization applying for loan gets increased every year. The bank employees have to put in a lot of work to analyse or predict whether the customer can pay back the loan amount or not (defaulter or non-defaulter) in the given time. The aim of this paper is to find the nature or background or credibility of the client that is applying for the loan. We use exploratory data analysis technique to deal with the problem of approving or rejecting the loan request or in short loan prediction. The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.

Loan prediction is a very common real-life problem that each retail bank faces at least once in its lifetime. If done correctly, it can save a lot of man hours at the end of a retail bank. Customers first apply for a home loan after that company validates the customer eligibility for the loan.

The loan eligibility process (real time) can be automated based on customer details I provided through for example filling an online application form. These details can be Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and many more. Hence the goal is to identify the customer segments that are eligible and in fact germane for loan amounts.

CHAPTER-1 Introduction

1.1 Purpose:

Home loan prediction refers to the use of predictive analytics and machine learning algorithms to assess the likelihood of a loan applicant being approved for a loan.

1.2 Problem Statement:

Dream 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.

1.3 Need of a solution:

This is a standard supervised classification task. A classification problem where we have to predict whether a loan would be approved or not. In a classification problem, we have to predict discrete values based on given set of independent variables.

CHAPTER-2 Literature survey

General

A literature review is a body of text that aims to review the critical points current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a reorganization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

Review of Literature Survey

Title : A benchmark of machine learning approaches for credit score prediction.

Author: Vincenzo Moscato, Antonio Picariello, Giancarlo Sperlí

Year : 2021

Credit risk assessment plays a key role for correctly supporting financial institutes in defining their bank policies and commercial strategies. Over the last decade, the emerging of social lending platforms has disrupted traditional services for credit risk assessment. Through these platforms, lenders and borrowers can easily interact among them without any involvement of financial institutes. In particular, they support borrowers in the fundraising process, enabling the participation of any number and size of lenders. However, the lack of lenders' experience and missing or uncertain information about borrower's credit history can increase risks in social lending platforms, requiring an accurate credit risk scoring. To overcome such issues, the credit risk assessment problem of financial operations is usually modelled as a binary problem on the basis of debt's repayment and proper machine learning techniques can be consequently exploited. In this paper, we propose a benchmarking study of some of the most used credit risk scoring models to predict if a loan will be repaid in a P2P platform. We deal with a class imbalance problem and leverage several classifiers among the most used in the literature, which are based on different sampling techniques. A real social lending platform (Lending Club) data-set, composed by 877,956 samples, has been used to perform the experimental analysis considering different evaluation metrics (i.e. AUC, Sensitivity, Specificity), also comparing the obtained outcomes with respect to the state-of-the-art approaches. Finally, the three best approaches have also been evaluated in terms of their explainability by means of different explainable Artificial Intelligence (XAI) tools.

Title : An Approach for Prediction of Loan approval using Machine Learning Algorithm.

Author: Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar

Year : 2020

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. 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. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model. The data is collected from the Kaggle for studying and prediction. Logistic Regression models have been performed and the different measures of performances are computed. The models are compared on the basis of the performance measures such as sensitivity and specificity. The final results have shown that the model produce different results.Model is marginally better because it includes variables (personal attributes of customer like age, purpose, credit history, credit amount, credit duration, etc.) other than checking account information (which shows wealth of a customer) that should be taken into account to calculate the probability of default on loan correctly. Therefore, by using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

Title : Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval.

Author: Amruta S. Aphale, Dr. Sandeep R. Shinde.

Year : 2020

In today's world, taking loans from financial institutions has become a very common phenomenon. Everyday a large number of people make application for loans, for a variety of purposes. But all these applicants are not reliable and everyone cannot be approved. Every year, we read about a number of cases where people do not repay bulk of the loan amount to the banks due to which they suffers huge losses. The risk associated with making a decision on loan approval is immense. So the idea of this project is to gather loan data from multiple data sources and use various machine learning algorithms on this data to extract important information. This model can be used by the organizations in making the right decision to approve or reject the loan request of the customers. In this paper, we examine a real bank credit data and conduct several machine learning algorithms on the data for that determine credit worthiness of customers in order to formulate bank risk automated system.

Title : Loan Approval Prediction Using Machine Learning

Author: Yash Divate, Prashant Rana, Pratik Chavan

Year : 2021

With the upgrade in the financial area loads of individuals are applying for bank advances however the bank has its restricted resources which it needs to allow to restricted individuals just, so discovering to whom the credit can be conceded which will be a more secure choice for the bank is a commonplace interaction. So in this task we attempt to decrease this danger factor behind choosing the protected individual in order to save bunches of bank endeavors and resources. This is finished by mining the Data of the past records of individuals to whom the advance was conceded previously and based on these records/encounters the machine was prepared utilizing the AI model which give the most precise outcome. The principle objective of this paper is to anticipate whether relegating the advance to specific individual will be protected or not. This paper is separated into four areas (i)Data Collection (ii) Comparison of AI models on gathered information (iii) Training of framework on most encouraging model (iv) Testing.

Title : Prediction for Loan Approval using Machine Learning Algorithm

Author: Ashwini S. Kadam, Shraddha R. Nikam, Ankita A. Aher, Gayatri V. Shelke,Amar S. Chandgude

Year : 2021

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. 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. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.

Title : Modern Approach for Loan Sanctioning in Banks Using Machine Learning

Author: Golak Bihari Rath, Debasish Das, BiswaRanjan Acharya

Year : 2021

Loan analysis is a process adopted by banks used to check the credibility of loan applicants who can pay back the sanction loan amount within regulations and loan amount term mentioned by the bank. Most banks use their common recommended procedure of credit scoring and background check techniques to analyze the loan application and to make decisions on loan approval. This is overall a risk-oriented and a time-consuming process. In some cases, people suffer through financial problems while some intentionally try to fraud. As a result, such delay and default in payment by the loan applicants can lead to loss of capital of the banks. Hence to overcome this, banks need to adopt a better procedure to find the trustworthy applicants for granting loan from the list of all applicants applied for the loan, who can pay can their loan amount in stipulated time. In the modern day age and advance of technology, we adopt a machine learning approach to reduce the risk factor and human errors in the loan sanction process and determine where an applicant is eligible for loan approval or not. Here, we examine various features such as applicant income, credit history, education from past records of loan applicants irrespective of their loan sanction, and the best features are determined and selected which have a direct impact on the outcome for loan approval.

CHAPTER-3 AIM AND SCOPE OF THE PRESENT INVESTIGATION

3.1 PROJECT PROPOSAL:

The project proposal is the term of documents. A project can describe the project proposal. It is the set of all plans of a project. Like, how the software works, what are the steps to complete the entire projects and what are the software requirements and analysis for this project. In my project, I am doing all the steps and also risk and reward and other project dependencies in the project proposal.

3.1.1 Mission:

An online Web based machine learning application is very popular and well known to everyone. Now a day's everybody wants to get it and work with it. Loan prediction is mostly useful for bank employees in approving the loan application. This simple method gives fast and accurate results in approving the customer application.

3.1.2 Goal:

The goal is to develop a machine learning model for Loan Approval Prediction.

3.2 SCOPE OF THE PROJECT:

The scope of this paper is to implement and investigate how different supervised binary classification methods impact default prediction. The model evaluation techniques used in this project are limited to precision, sensitivity, F1-score.

3.3 OVERVIEW OF THE PROJECT:

The overview of the project is to provide a web-based machine learning application to the user. Therefore, the user can directly check the loan approval in our website over the internet. So, the user can easily check into this loan system prediction whether their loan will be approved or not.

3.4 EXISTING SYSTEM:

Anomaly detection relies on individuals' behaviour profiling and works by detecting any deviation from the norm. When used for online banking fraud detection, however, it mainly suffers from three disadvantages. First, for an individual, the historical behaviour data are often too limited to profile his/her behaviour pattern. Second, due to the heterogeneous nature of transaction data, there lacks a uniform treatment of different kinds of attribute values, which becomes a potential barrier for model development and further usage.

Third, the transaction data are highly skewed, and it becomes a challenge to utilize the label information effectively. Anomaly detection often suffers from poor generalization ability and a high false alarm rate. We argue that individuals' limited historical data for behaviour profiling and the highly skewed nature of fraud data could account for this defect. Since it is straightforward to use information from other similar individuals, measuring similarity itself becomes a great challenge due to heterogeneous attribute values. We propose to transform the anomaly detection problem into a pseudo-recommender system problem and solve it with an embedding based method. By doing so, the idea of collaborative filtering is implicitly used to utilize information from similar users, and the learned preference matrices and attribute embedding provide a concise way for further usage.

3.4.1 Disadvantages:

1. They had proposed a mathematical model and machine learning algorithms is not used.

2. Class Imbalance problem was not addressed and the proper measure were not taken

3.5 PREPARING THE DATASET:

This dataset contains 665 records of features extracted from Bank Loan data, which were then classified into 2 classes:

- Approve
- Reject

3.6 PROPOSED SYSTEM:

3.6.1 Exploratory Data Analysis of loan approval

Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

3.6.2 Data Wrangling

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

3.6.3 Data collection

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

3.6.4 Building the classification model

The predicting the loan approval, ML algorithm prediction model is effective because of the following reasons: It provides better results in classification problem.

It is strong in preprocessing outliers, irrelevant variables, and a mix of continuous, categorical and discrete variables.

It produces out of bag estimate error which has proven to be unbiased in many tests and it is relatively easy to tune with.

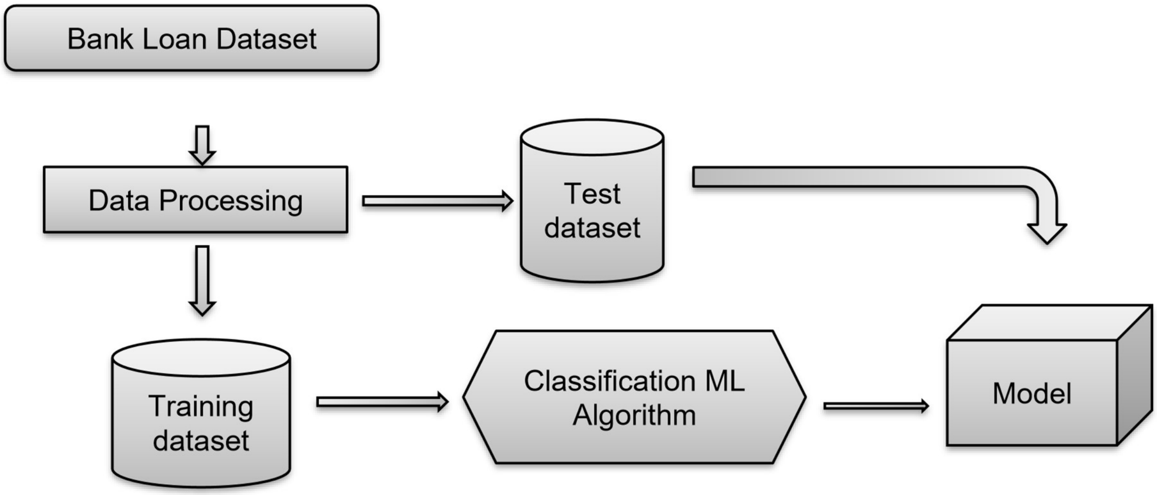


Fig.1: Architecture of Proposed model

3.6.5 Advantages:

1. Performance and accuracy of the algorithms can be calculated and compared.

2. Class imbalance can be dealt with machine learning approaches.

3.7 FLOW CHART:

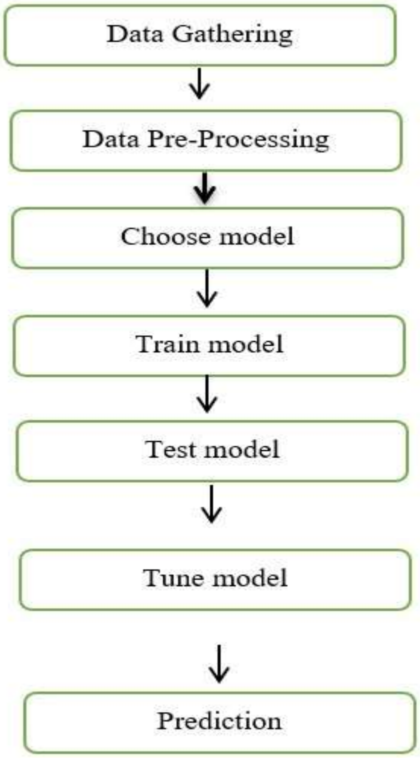


Fig.2: FLOW CHART

CHAPTER-4 SOFTWARE AND HARDWARE REQUIREMENT

. 4.1 SYSTEM STUDY:

To develop this model we use new modern technologies which are Machine Learning using Python for predicting and Flask is for user interface.

4.1.1 System requirement specifications:

a) Hardware requirements:

Processor	:	Intel
RAM	:	2GB
Hard Disk	:	80GB

b) Software requirements:

OS	:	Windows
Framework	:	Flask
Technology	:	Machine Learning using Python
Web Browser	:	Chrome, Microsoft Edge
Code editor	:	Visual Studio Code, Google Colab, Anaconda or Jupyter notebook.

CHAPTER-5 EXPERIMENTAL OR MATERIALS AND METHODS ALGORITHMS USED

5.1 SYSTEM SPECIFICATIONS:

5.1.1 Machine Learning Overview:

Machine learning is a field of study that looks at using computational algorithms to turn empirical data into usable models. The machine learning field grew out of traditional statistics and artificial intelligences communities. Through their business processes immense amounts of data have been and will be collected. This has provided an opportunity to re-invigorate the statistical and computational approaches to autogenerate useful models from data. Machine learning algorithms can be used to (a) gather understanding of the cyber phenomenon that produced the data under study, (b) abstract the understanding of underlying phenomena in the form of a model, (c) predict future values of a phenomena using the above-generated model, and (d) detect anomalous behaviour exhibited by a phenomenon under observation

5.1.2 Flask Overview:

Flask is an API of Python that allows us to build up web applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application.

5.1.3 STEPS TO DOWNLOAD & INSTALL PYTHON:

Download the Latest version of the **Python**executableinstaller (https://www.python.org/downloads/). Watch the PIP list where pip is the package installer for python. Now upgrade the pip and setuptools using the command.

```
Pip install --upgrade pip and Pip install --upgrade setuptools
```

5.1.3.1 IDE INSTALLATION FOR PYTHON

IDE stands for Integrated Development Environment. It is a GUI (Graphical User Interface) where programmers write their code and produce the final products**.** Best IDE is Pycharm. So download the pycharm new version and install the software (https://www.jetbrains.com/pycharm/download/)

5.1.3.2 PYTHON FILE CREATION

GO To FILE MENU > CREATE > NEW > PYTHON FILE >(Name Your Python File as “HOUSE PRICE PPREDICTION” > SAVE

5.2 PYTHON LIBRARIES NEEDED

There are many libraries in python. In those we only use few main libraries needed.

5.2.1 NUMPY LIBRARY

numPy is an open-source numerical Python library. NumPy contains a multi dimensional array and matrix data structures. It can be utilized to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines like mean, mode, standard deviation etc…,

Installation- (https://numpy.org/install/)

```
pip install NUMPY
```

Here we mainly use array, to find mean and standard deviation.

5.2.2 PANDAS LIBRARY

Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on Numpy package and its key data structure is called the DataFrame. DataFrames allow you store and manipulate tabular data in rows of observations and columns of variables. There are several ways to create a DataFrame.

Installation- (https://pandas.pydata.org/getting_started.html)

```
pip install PANDAS
```

Here we use pandas for reading the csv files, for grouping the data, for cleaning the data using some operations.

5.2.3 MATPLOTLIB LIBRARY

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Use interactive figures that can zoom, pan, update, visualize etc., Installation- (https://matplotlib.org/users/installing.html)

```
pip install Matplotlib
```

Here we use pyplot mainly for plotting graphs. **matplotlib**.**pyplot** is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

5.2.4 SCIKIT-LEARN LIBRARY

Scikit-learn is a free machine learning library for the Python. It features various algorithms like support vector machine, random forests, regression and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

```
Pip install Scikit-Learn
```

Installation-(https://scikit-learn.org/stable/install.html)

Here use scikit-learn’s regression methods for prediction purpose.

5.2.5 FLASK

Flask is an API of Python that allows us to build up web applications. It was developed by Armin Ronacher. Flask's framework is more explicit than Django's framework and is also easier to learn because it has less base code to implement a simple web-Application.

```
pip install flask
```

Here we use flask for the user-interface.

5.3 MODULES:

A modular design reduces complexity, facilities change (a critical aspect of software maintainability), and results in easier implementation by encouraging parallel development of different part of system. Software with effective modularity is easier to develop because function may be compartmentalized and interfaces are simplified. Software architecture embodies modularity that is software is divided into separately named and addressable components called modules that are integrated to satisfy problem requirements.

Modularity is the single attribute of software that allows a program to be intellectually manageable. The five important criteria that enable us to evaluate a design method with respect to its ability to define an effective modular design are: Modular decomposability, Modular Comps ability, Modular Understand ability, Modular continuity, Modular Protection.

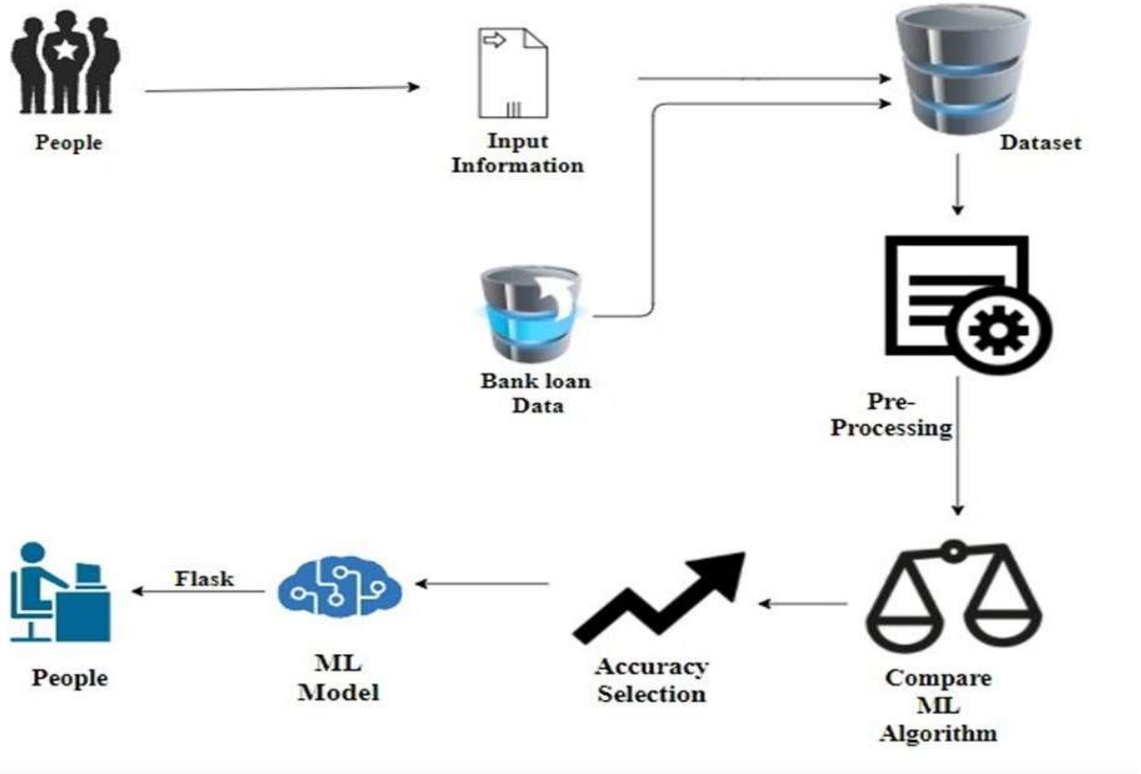


Fig.3: SYSTEM ARCHITECTURE

CHAPTER-6 PREDICTIVE MODELLING

6.1 Predictive Modelling

Predictive modelling is used to analyze the data and predict the outcome. Predictive modelling used to predict the unknown event which may occur in the future. In this process, we are going to create, test and validate the model. There are different methods in predictive modelling. They are learning, artificial intelligence and statistics. Once we create a model, we can use many times, to determine the probability of outcomes. So, predict model is reusable. Historical data is used to train an algorithm. The predictive modelling process is an iterative process and often involves training the model, using multiple models on the same dataset.

- **Creating the model:** To create a model to run one or more algorithms on the data set.
- **Testing a model:** The testing is done on past data to see how the best model predicts
- **Validating a model:** Using visualization tools to validate the model.
- **Evaluating model:** Evaluating the best fit model from the models used and choosing the model right fitted for the data.

6.2 Process Mode Used:

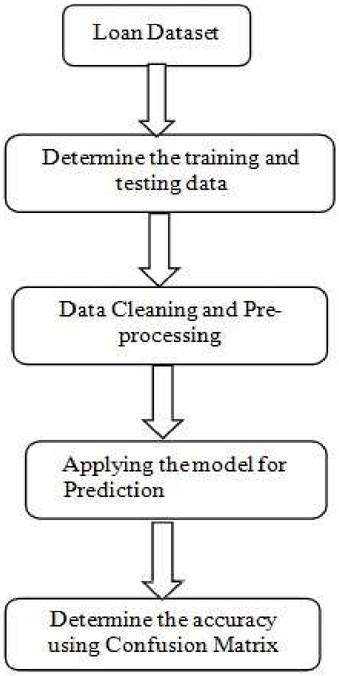

Loan Dataset
Determine the training and testing data
Data Cleaning and Pre-processing
Applying the model for Prediction
Determine the accuracy using Confusion Matrix

Fig.4 :Architecture of the proposed loan prediction model

6.3 What is hypothesis generation ?

Hypothesis generation is the process of listing out all the possible factors that can affect the final outcome.

This is a very important stage in a data science/machine learning pipeline. It involves understanding the problem in detail by brainstorming maximum possibilities that can impact the outcome. It is done by thoroughly understanding the problem statement before looking at the data.

Firstly, lets list out some of the most important factors which can affect the Loan Approval:

Salary: Applicants with high income should have more chances of loan approval.

Previous history: Applicants who have repayed their previous debts should have higher chances of loan approval.

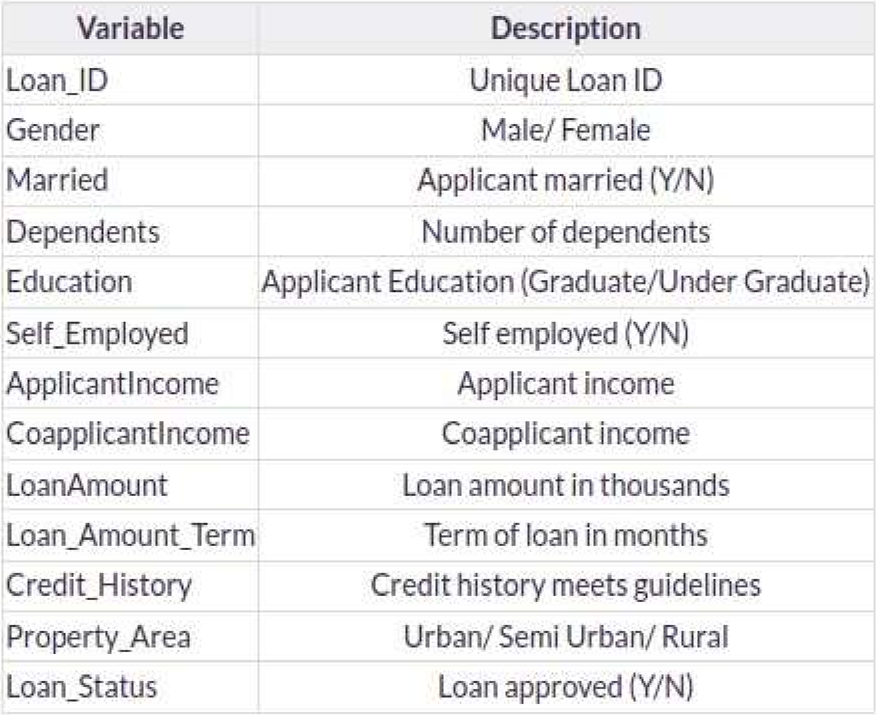
Loan amount: Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.

Loan term: Loan for less time period and less amount should have higher chances of approval.

EMI: Lesser the amount to be paid monthly to repay the loan, higher the chances of loan approval.

These are some of the factors which i think can affect the target variable, you can come up with many more factors.

6.4 Variable Description : Given below the description for each variable



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)

Fig.5: variable description

CHAPTER-7 UML DIAGRAMS

7.1 USE CASE DIAGRAM

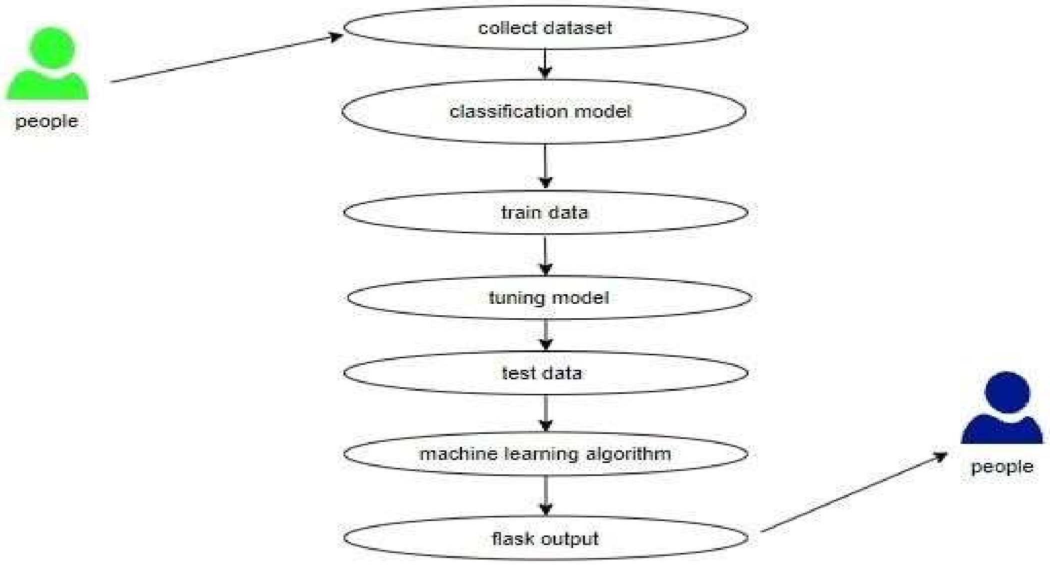


Fig.6 : USE CASE DIAGRAM

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner.

7.2 CLASS DIAGRAM

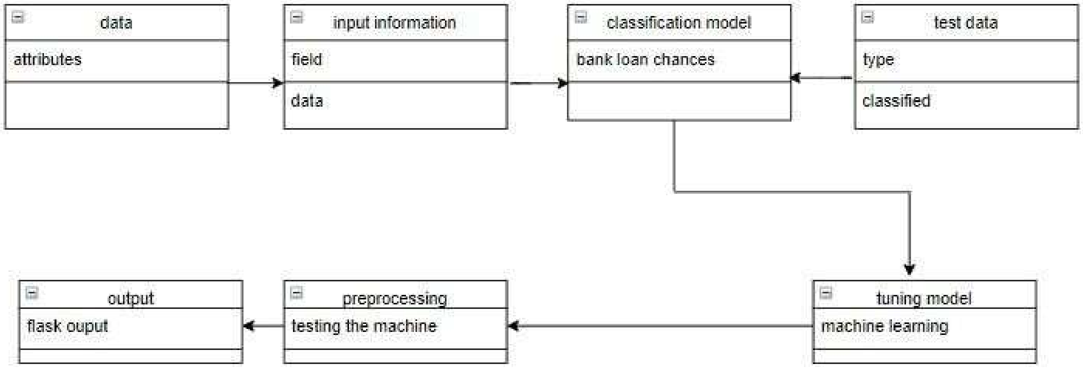


Fig.7 : CLASS DIAGRAM

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possible to make it correct.

7.3 ACTIVITY DIAGRAM

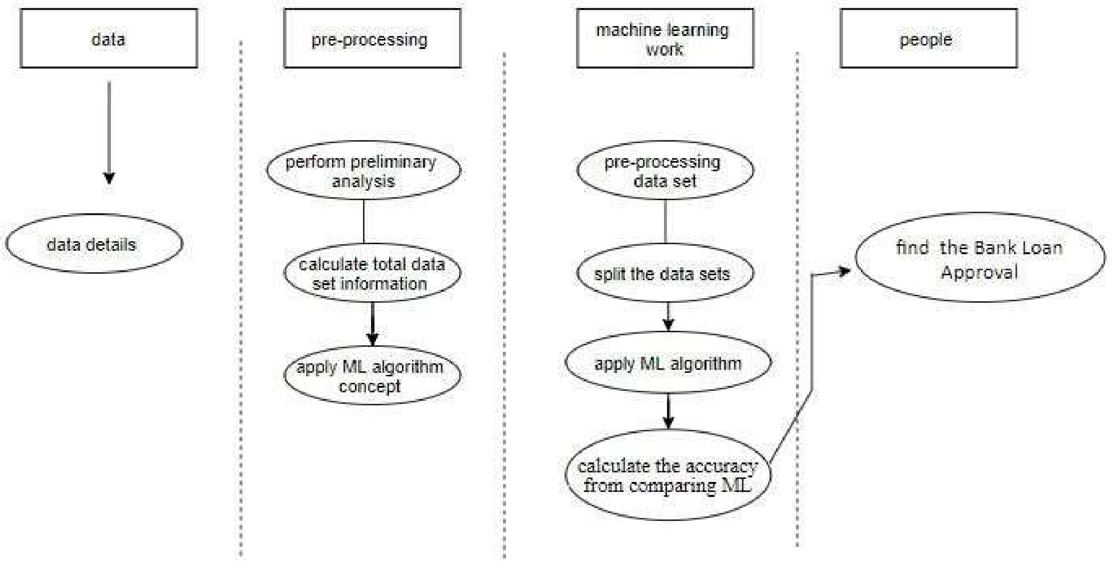

data
pre-processing
machine learning work
people
data details
perform preliminary analysis
calculate total data set information
apply ML algorithm concept
pre-processing data set
split the data sets
apply ML algorithm
calculate the accuracy from comparing ML
find the Bank Loan Approval

Fig.8: ACTIVITY DIAGRAM

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single.

7.4 SEQUENCE DIAGRAM

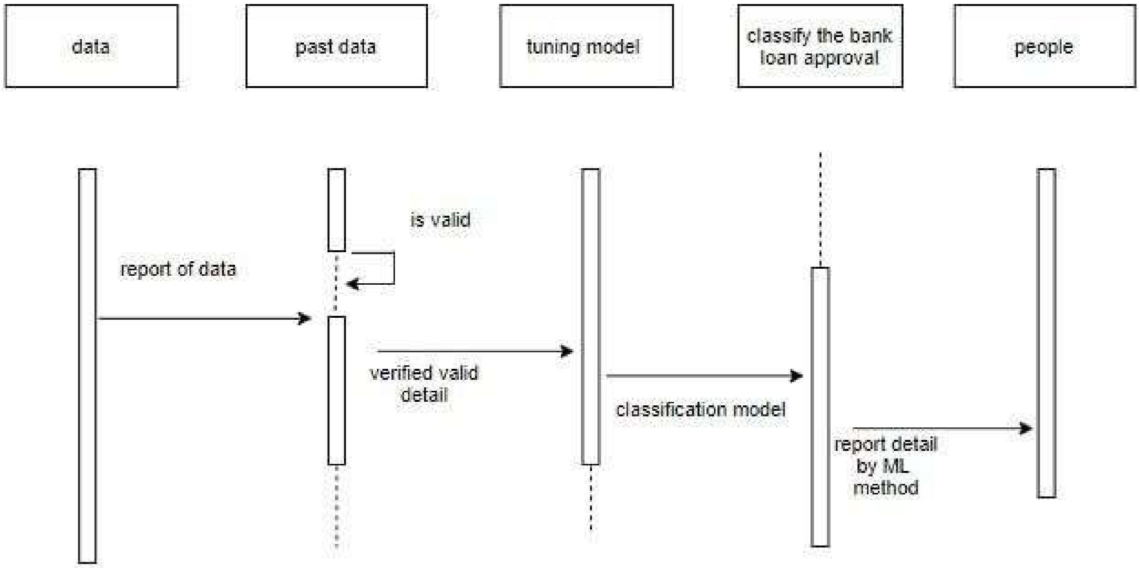


Fig.9 : SEQUENCE DIAGRAM

Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modelling, which focuses on identifying the behaviour within your system. Other dynamic modelling techniques include activity diagramming, communication diagramming, timing diagramming, and interaction overview diagramming. Sequence diagrams, along with class diagrams and physical data models are in my opinion the most important design-level models for modern business application development.

CHAPTER-8.MODULE DETAILS:

8.1 Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in realworld scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different data cleaning tasks using Python Pandas library and specifically, it focuses on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modelling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It’s important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it’s important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose & read the given dataset
- General Properties of Analyzing the given dataset

- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

8.2 Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process.

The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

MODULE DIAGRAM

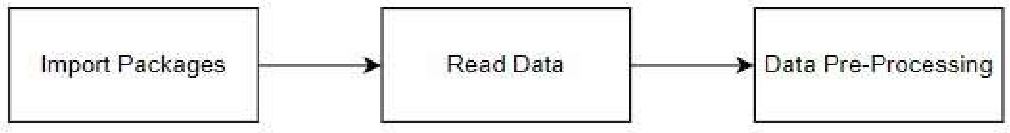


Fig.10

GIVEN INPUT EXPECT OUTPUT

input: data output: removing noisy data

8.3 Exploration data analysis of visualization

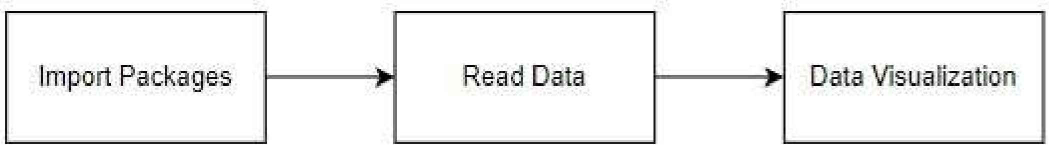
Data visualization is an important skill in applied statistics and machine learning.

Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

MODULE DIAGRAM



GIVEN INPUT EXPECT OUTPUT

input: data output:
visualized data

8.4 ALGORITHMS

8.4.1 Logistic Regression:

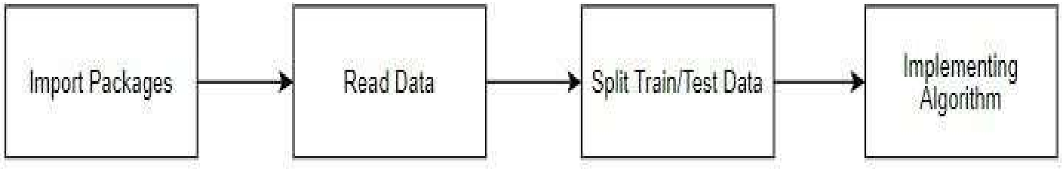
It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

MODULE DIAGRAM



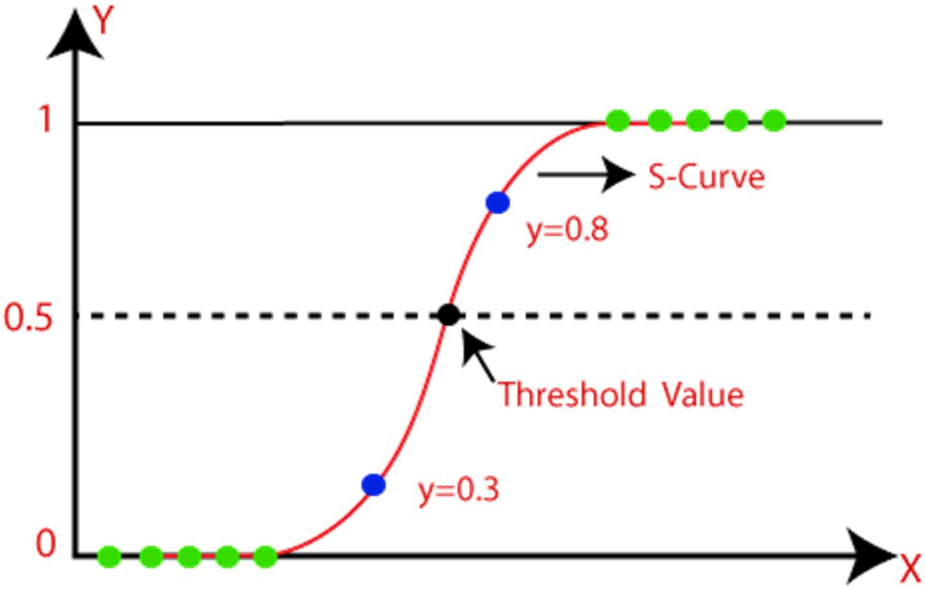


Fig.11: LOGISTIC REGRESSION

GIVEN INPUT EXPECT OUTPUT

input: data output: getting

accuracy

8.4.2 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees,* resulting in a *forest of trees,* hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

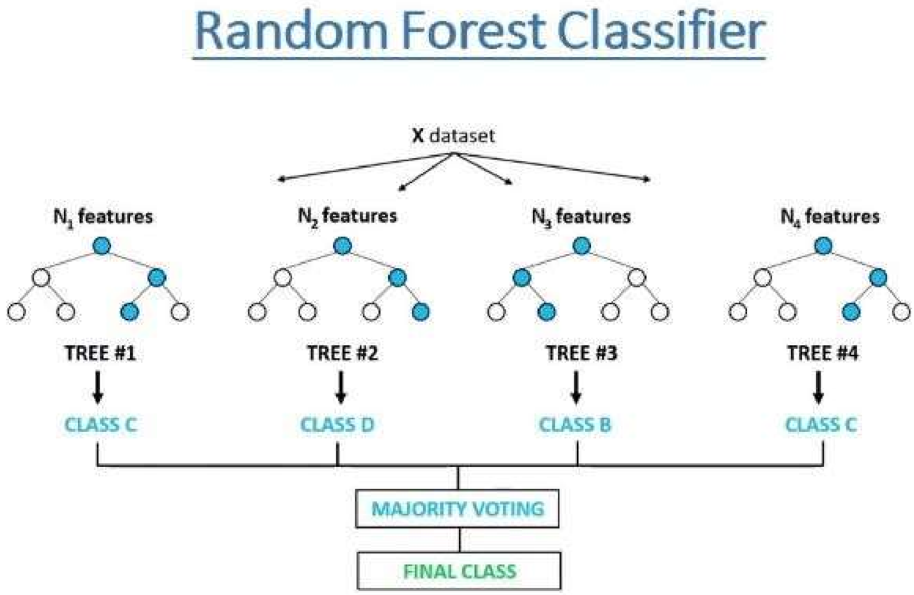


Fig.12: RANDOM FOREST CLASSIFIER

GIVEN INPUT EXPECT OUTPUT

input: data output: getting
accuracy

8.4.3 Decision Tree Classifier:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

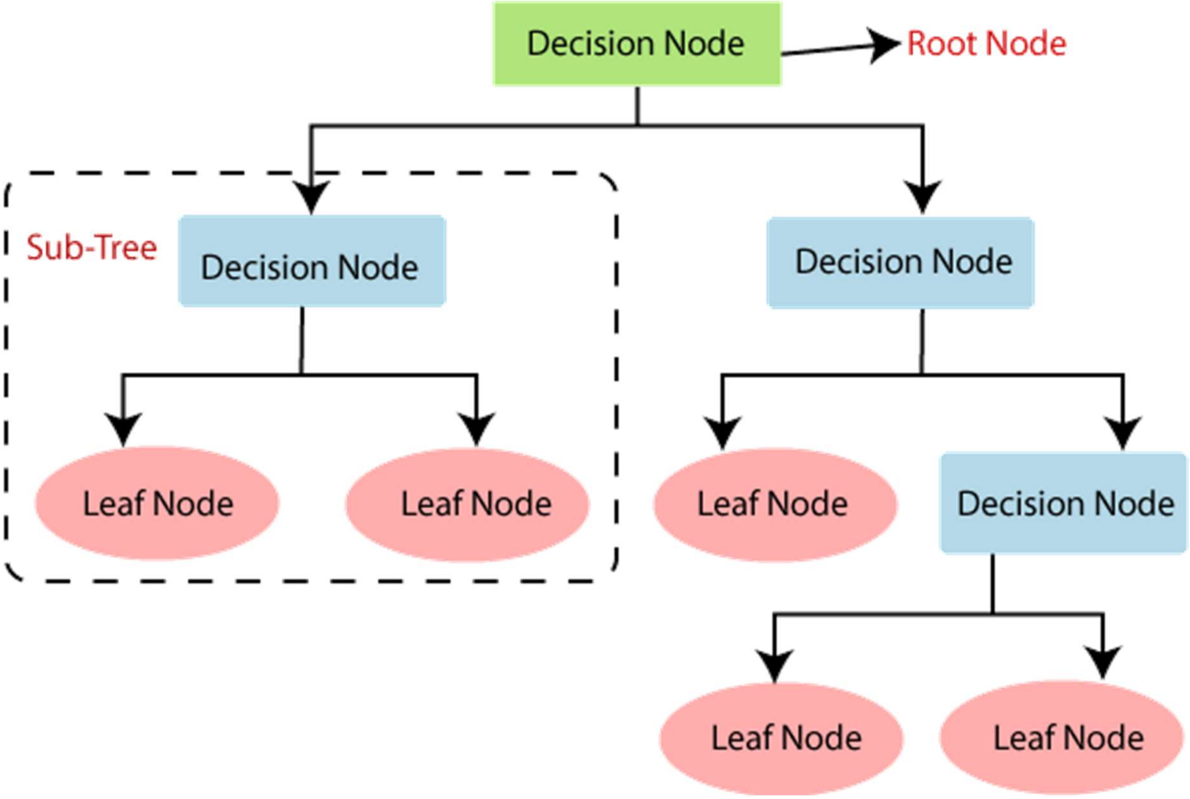


Fig.13: DECISION TREE CLASSIFIER

8.4.4 K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

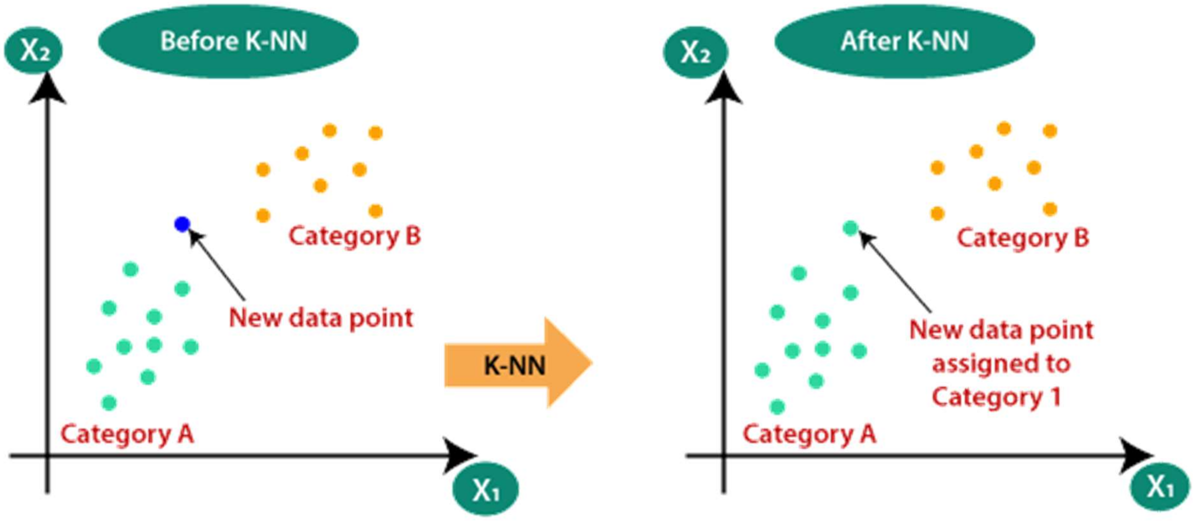


Fig.14: K-NEAREST NEIGHBOUR

8.4.5 Gradient Boost

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e., Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.

Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).

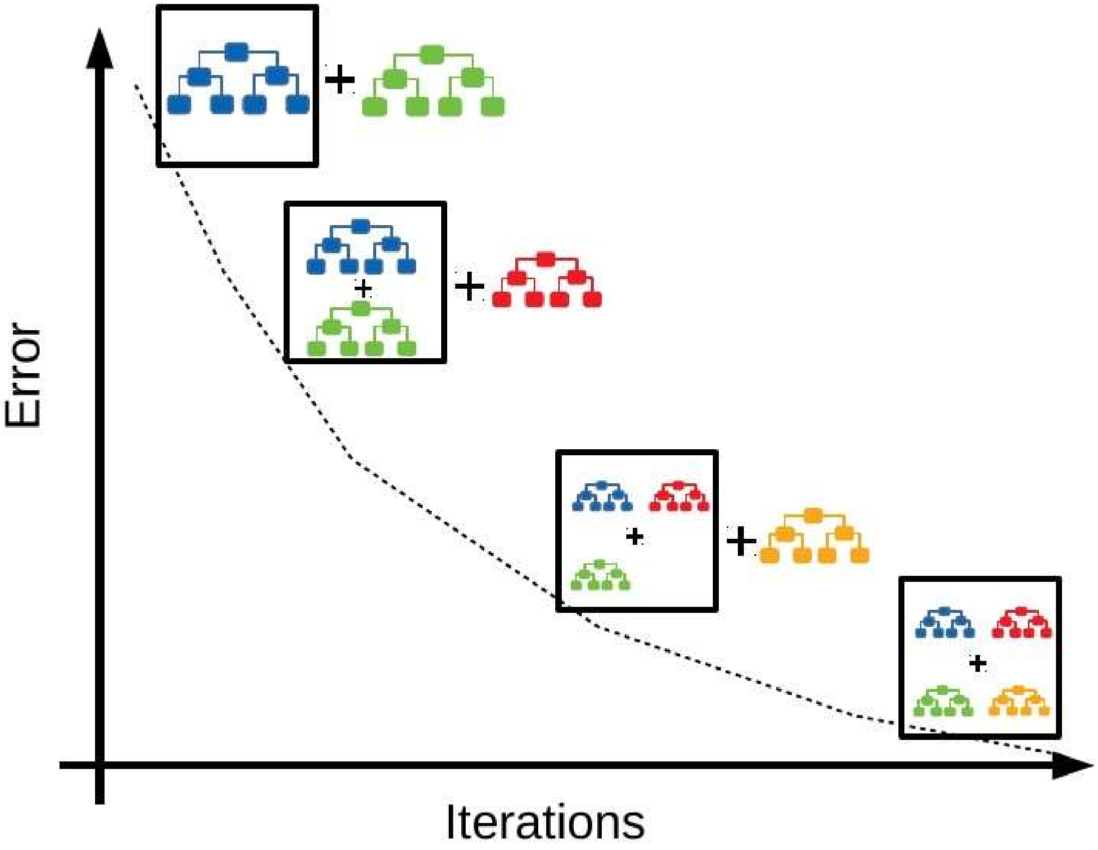


Fig.15: GRADIENT BOOST

8.4.6 Deployment Using Flask (Web Framework):

Flask is a micro web framework written in Python. It is classified as a micro-framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where preexisting third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself.

Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Flask was created by Armin Ronacher of Pocoo, an international group of Python enthusiasts formed in 2004. According to Ronacher, the idea was originally an April Fool's joke that was popular enough to make into a serious application. The name is a play on the earlier Bottle framework.

When Ronacher and Georg Brand created a bulletin board system written in Python, the Pocoo projects Werkzeug and Jinja were developed.

In April 2016, the Pocoo team was disbanded and development of Flask and related libraries passed to the newly formed Pallets project.

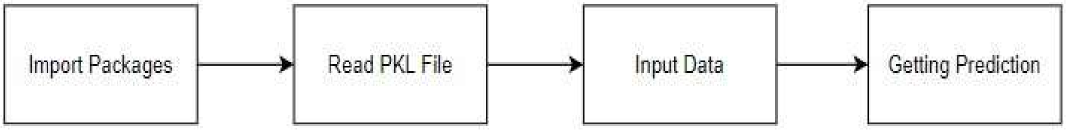
Flask has become popular among Python enthusiasts. As of October 2020, it has second most stars on GitHub among Python web-development frameworks, only slightly behind Django, and was voted the most popular web framework in the Python Developers Survey 2018.

The micro-framework Flask is part of the Pallets Projects, and based on several others of them.

Flask is based on Werkzeug, Jinja2 and inspired by Sinatra Ruby framework, available under BSD licence. It was developed at pocoo by Armin Ronacher. Although Flask is rather young compared to most Python frameworks, it holds a great promise and has already gained popularity among Python web developers.

Let's take a closer look into Flask, so-called "micro" framework for Python.

MODULE DIAGRAM



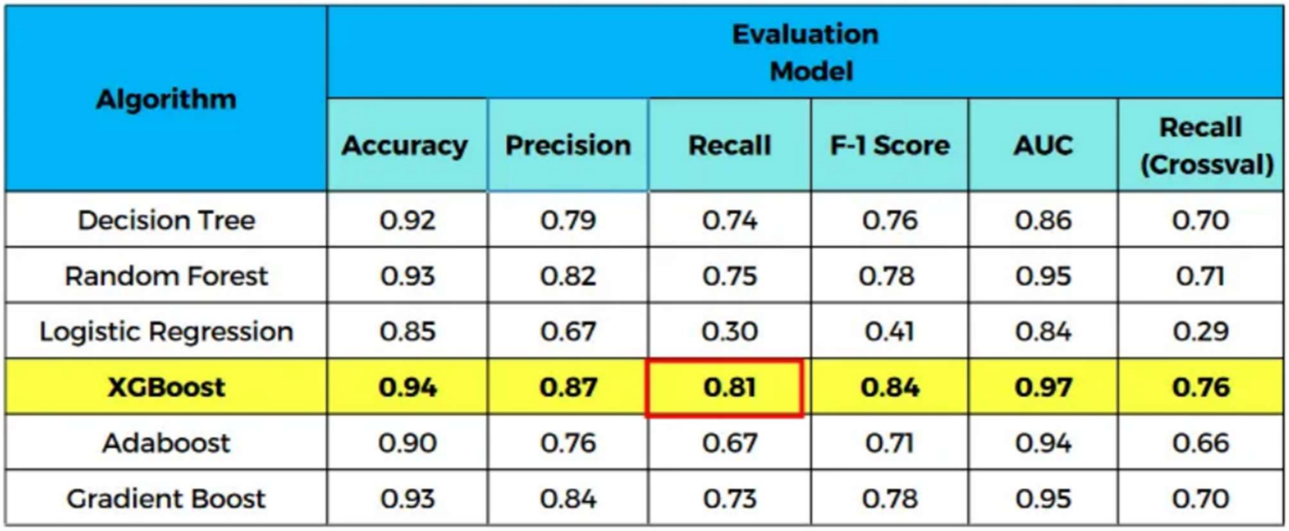
GIVEN INPUT EXPECTED OUTPUT

input : data values output :
predicting output

8.4.7 XG BOOST ALGORITHM

The evaluation metric used is Recall. XGBoost has the best algorithm with a very small gap between AUC train and test with cross results the highest validation of the others.

XGBoost stands for e**X**treme **G**radient **Boost**ing and it's an open-source implementation of the **gradient boosted trees** algorithm. It has been one of the most popular machine learning techniques in Kaggle competitions, due to its prediction power and ease of use. It is a **supervised learning** algorithm that can be used for **regression** or **classification** tasks.



Algorithm	Evaluation Model					
	Accuracy	Precision	Recall	F-1 Score	AUC	Recall (Crossval)
Decision Tree	0.92	0.79	0.74	0.76	0.86	0.70
Random Forest	0.93	0.82	0.75	0.78	0.95	0.71
Logistic Regression	0.85	0.67	0.30	0.41	0.84	0.29
XGBoost	**0.94**	**0.87**	**0.81**	**0.84**	**0.97**	**0.76**
Adaboost	0.90	0.76	0.67	0.71	0.94	0.66
Gradient Boost	0.93	0.84	0.73	0.78	0.95	0.70

Fig.16 Chart

Feature importance from XGBoost

Obtained recall 81%. Impact from model is the greater the loan amount approved, the guarantee provided increases and the greater the agreed loan amount, the offer to extend it payment period.

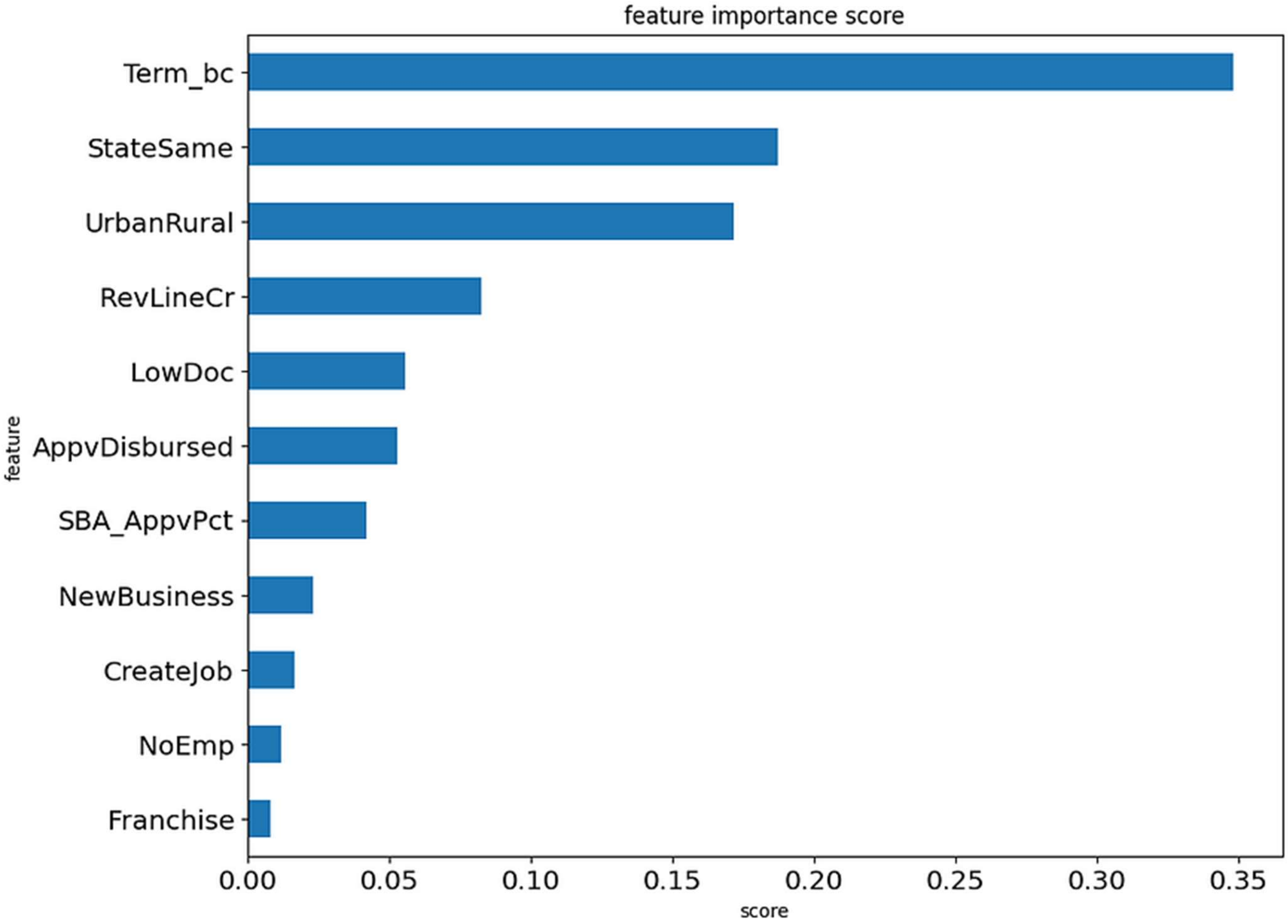

feature importance score
Term_bc
StateSame
UrbanRural
RevLineCr
LowDoc
AppvDisbursed
SBA_AppvPct
NewBusiness
CreateJob
NoEmp
Franchise
feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
score

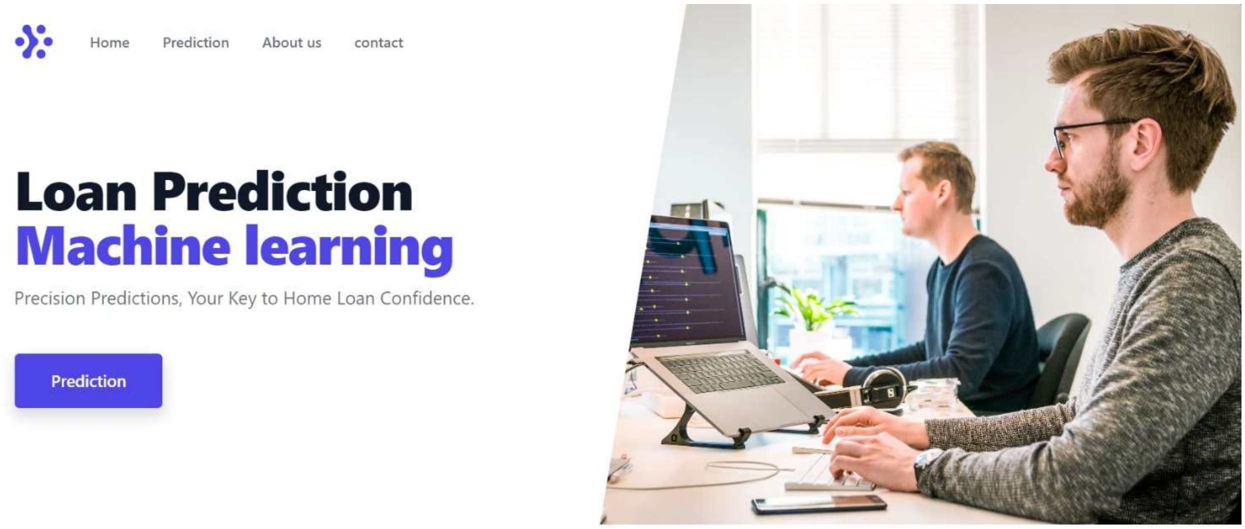
Fig.17

CHAPTER-9 Implentation

```
from flask import Flask, request, render_template
import pickle
import numpy as np

app = Flask(_name_)
model = pickle.load(open('model.pkl', 'rb'))

@app.route('/')
def home():
    return render_template("index.html")


@app.route('/predict', methods=['GET', 'POST'])
def predict():
    if request.method ==  'POST':
        gender = request.form['gender']
        married = request.form['married']
        dependents = request.form['dependents']
        education = request.form['education']
        employed = request.form['employed']
        credit = float(request.form['credit'])
        area = request.form['area']
        ApplicantIncome = float(request.form['ApplicantIncome'])
        CoapplicantIncome = float(request.form['CoapplicantIncome'])
        LoanAmount = float(request.form['LoanAmount'])
        Loan_Amount_Term = float(request.form['Loan_Amount_Term'])

        # gender
        if (gender == "Male"):
            male=1
        else:
            male=0

        # married
        if(married=="Yes"):
            married_yes = 1
        else:
```

```
    married_yes=0

# dependents
if(dependents=='1'):
    dependents_1 = 1
    dependents_2 = 0
    dependents_3 = 0
elif(dependents == '2'):
    dependents_1 = 0
    dependents_2 = 1
    dependents_3 = 0
elif(dependents=="3+"):
    dependents_1 = 0
    dependents_2 = 0
    dependents_3 = 1
else:
    dependents_1 = 0
    dependents_2 = 0
    dependents_3 = 0

# education
if (education=="Not Graduate"):
    not_graduate=1
else:
    not_graduate=0

# employed
if (employed == "Yes"):
    employed_yes=1
else:
    employed_yes=0

# property area

if(area=="Semiurban"):
    semiurban=1
    urban=0
elif(area=="Urban"):
    semiurban=0
```

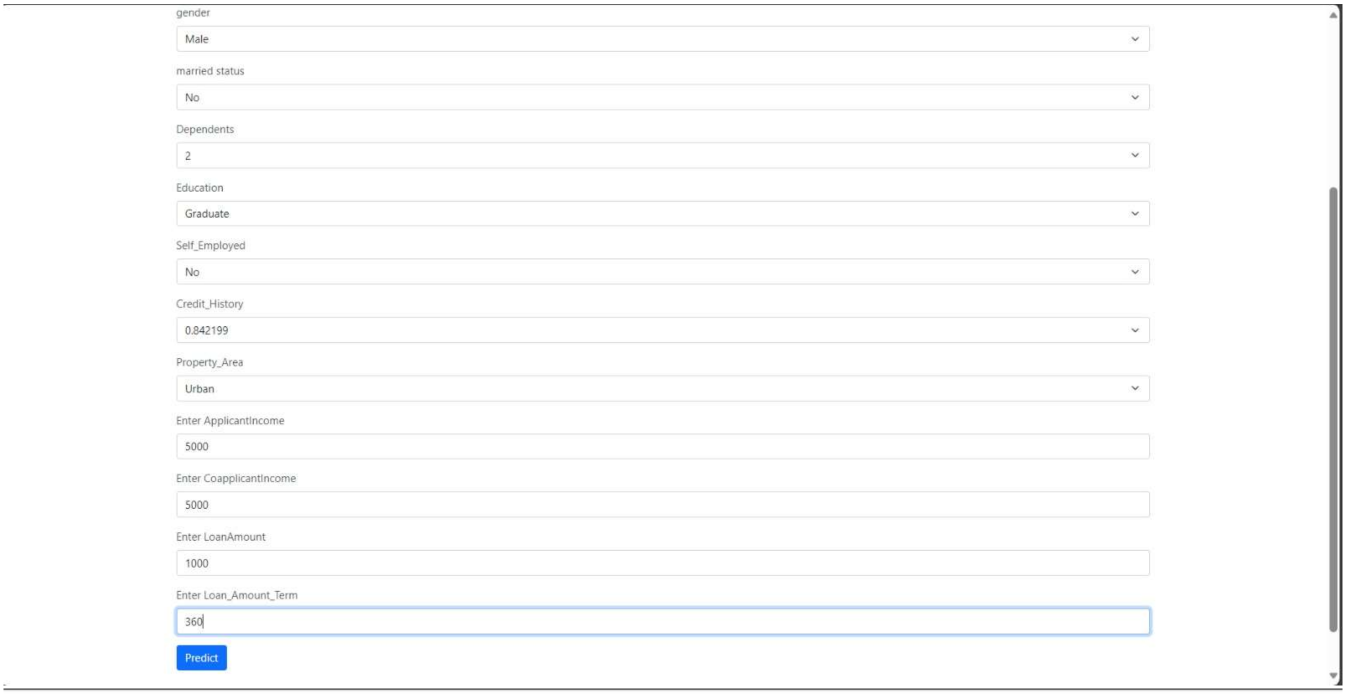
```
            urban=1
        else:
            semiurban=0
            urban=0


        ApplicantIncomelog = np.log(ApplicantIncome)
        totalincomelog = np.log(ApplicantIncome+CoapplicantIncome)
        LoanAmountlog = np.log(LoanAmount)
        Loan_Amount_Termlog = np.log(Loan_Amount_Term)

        prediction = model.predict([[credit, ApplicantIncomelog,LoanAmountlog,
Loan_Amount_Termlog, totalincomelog, male, married_yes, dependents_1, dependents_2,
dependents_3, not_graduate, employed_yes,semiurban, urban ]])

        # print(prediction)

        if(prediction=="N"):
            prediction="No"
        else:
            prediction="Yes"


        return render_template("prediction.html", prediction_text="loan status is
{}".format(prediction))




    else:
        return render_template("prediction.html")




if _name_ == "_main_":
    app.run(debug=True)
```



Loan Prediction
Machine learning

Precision Predictions, Your Key to Home Loan Confidence.

Output:



gender

Male

married status

No

Dependents

2

Education

Graduate

Self_Employed

No

Credit_History

0.842199

Property_Area

Urban

Enter ApplicantIncome

5000

Enter CoapplicantIncome

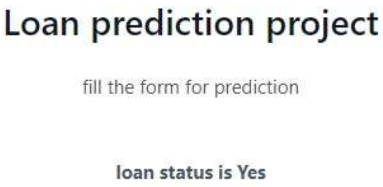
5000

Enter LoanAmount

1000

Enter Loan_Amount_Term

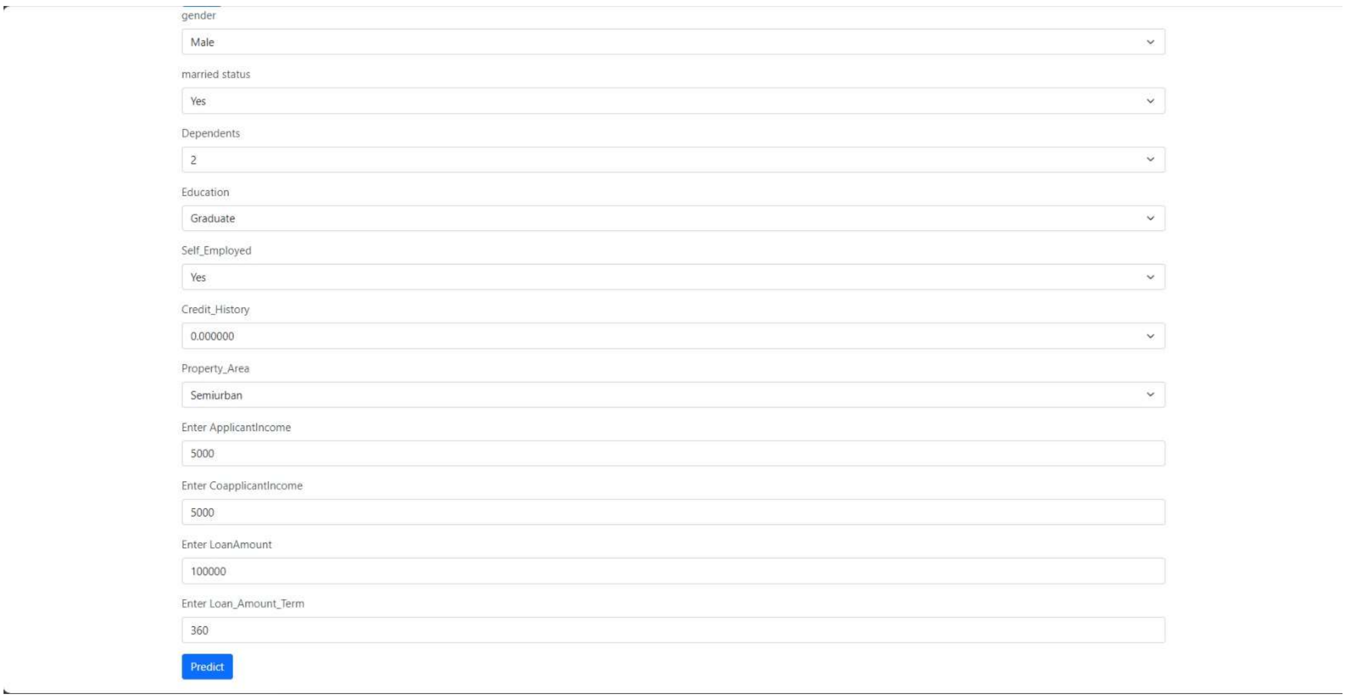
360



Loan prediction project

fill the form for prediction

loan status is Yes



gender

Male

married status

Yes

Dependents

2

Education

Graduate

Self_Employed

Yes

Credit_History

0.000000

Property_Area

Semiurban

Enter ApplicantIncome

5000

Enter CoapplicantIncome

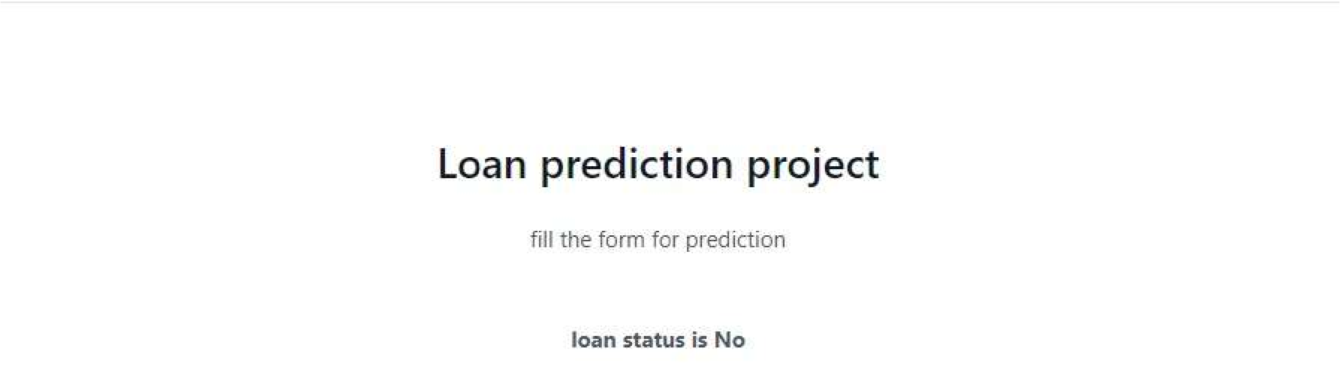
5000

Enter LoanAmount

100000

Enter Loan_Amount_Term

360



Loan prediction project

fill the form for prediction

loan status is No

Chapter-10 SUMMARY AND CONCLUSION

In conclusion, home loan prediction is a critical application of predictive analytics and machine learning in the financial industry. It serves as a valuable tool for lenders, borrowers, and regulatory authorities, addressing various needs and challenges within the home loan approval process.

For lenders, home loan prediction offers the ability to assess the creditworthiness of applicants accurately, optimize lending practices, and manage risk efficiently. This results in improved profitability, compliance with regulations, and enhanced portfolio management.

Borrowers benefit from home loan prediction by gaining insights into their likelihood of loan approval and the potential interest rates they might receive. This empowers them to make informed decisions and improve their financial standing when necessary.

Regulatory bodies benefit from the increased transparency and fairness that predictive models bring to lending practices. These models help prevent discriminatory practices and ensure that lending institutions adhere to established regulations.

Furthermore, home loan prediction supports efficiency, automation, and data-driven decision-making, contributing to a more streamlined and customer-friendly loan application process. It also aids in default prevention and the promotion of financial inclusion by evaluating creditworthiness based on objective criteria.

Overall, home loan prediction is a valuable solution that not only benefits individual stakeholders but also contributes to the stability and responsible growth of the housing and lending industries. It represents a significant step toward making the lending process more efficient, fair, and accessible for all.

Chapter-11 References

[1] Amruta S. Aphale , Dr. Sandeep R. Shinde, 2020, Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval, International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 08 (August 2020)

[2] Ashwini S. Kadam, Shraddha R. Nikam, Ankita A. Aher, Gayatri V. Shelke, Amar S.Chandgude, “Prediction for Loan Approval using Machine Learning Algorithm” (IRJET) Volume: 08 Issue: 04 | Apr 2021.

[3] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490494, doi: 10.1109/ICESC48915.2020.9155614.

[4] Rath, Golak & Das, Debasish & Acharya, Biswaranjan. (2021). Modern Approach for Loan Sanctioning in Banks Using Machine Learning. Pages={179-188} 10.1007/978-981-15-5243-4_15.

[5] Vincenzo Moscato, Antonio Picariello, Giancarlo Sperlí, A benchmark of machine learning approaches for credit score prediction, Expert Systems with Applications, Volume 165, 2021, 113986, ISSN 0957-4174.

[6] Yash Divate, Prashant Rana, Pratik Chavan, “Loan Approval Prediction Using Machine Learning” International Research Journal of Engineering and Technology (IRJET) Volume: 08 Issue: 05 | May 2021.