NETAJI SUBHAS UNIVERSITY OF TECHNOLOGY

COCSC17

Machine Learning Project Report

Jyotshna Jha	Garima	Tarun Mittal
2021UCS1592	2021UCS1605	2021UCS1608

Problem Statement :

Banks play an integral role in the financial system of any country which directly affects its economic status and growth. The major roles of banks include accepting deposits from its customers, using those deposits to lend money to the borrowers in return for some interest, granting credits, discounting on bills etc. But the main source of profit for the banks is the interest it receives from lending money to the borrowers. A major problem faced by these banks is the failure of timely loan repayment by the borrowers. So, to tackle this problem, banks nowadays use some models to predict the possibility of loan repayment from the borrower. Factors like annual income, employment status, home ownership, current debt etc are taken into consideration to categorize the loan request as bad loan or not. So, our model basically aims to develop a similar model.

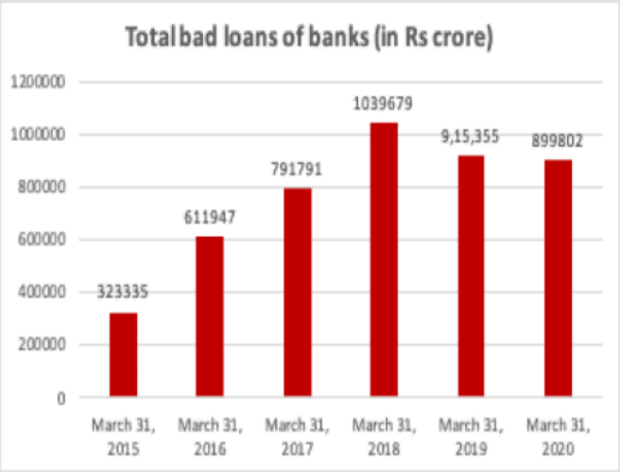


Fig. 1: Increase in the amount of bad loans

About The Project :

We are going to use streamlit library as a front end and python as a back end in this project and we well predict the loan amount on the basis of different factors like :

- Education
- Number of Dependents
- Income
- Loan Amount
- Credit History

and others. Also we are using different python libraries like:

- Pandas : Pandas Functions to get Deeper Insights into the Dataset i.e. isnull().
- Numpy : Used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform, and matrices.

- Sklearn / Scikit-Learn : It is a python library to implement machine learning models and statistical modeling.

- Seaborn : It is a library for making statistical graphics in Python.

- Streamlit : It is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science.

What are we trying to bring to focus with this model?

- Gender Biasness:
  In olden times, Banks and Loan providers used to prefer some specific genders when approving loan requests. Through automating the entire loan prediction model, based on previous trends, this discrimination can be reduced. Through this model, people can have proof that their loan request has a certain probability of being approved and through this prediction, the customer has less chances of being discriminated against.

- Greater Time management:
  Through the online availability of this loan prediction model, customers in need of a loan can avoid wasting their time on loan providers/Banks that are less likely to approve their loan based on past records. Customers save time and effort by only applying for a loan to those banks that have high and fair chances of approving their loan requests.

Overview :

- Getting Dataset: Choosing a real-life loan application dataset with balanced data is really important, because if the target variable has uneven distribution, then the model trained on that dataset would not be accurate.

- Exploratory Data Analysis: This module aims at analyzing the dataset, identifying the main attributes, their correlations etc. using visualization.

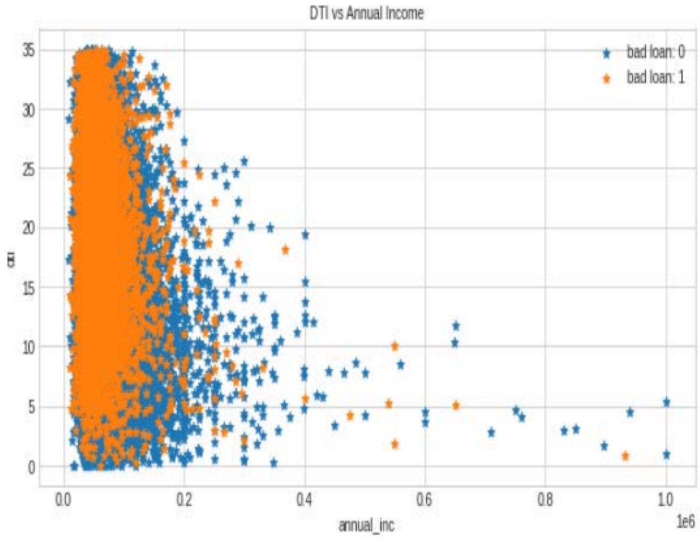


Fig.4: The scatter plot between annual income and DTI shows that with increase in annual income, DTI decreases and as a result the instances of bad_loan is higher for less annual income.

- Visualize the data using different graphics like histogram.
- Split the data into training and testing sets. The training set is used as an input by the model to understand the pattern and give predictions on unknown data. The test set is used to determine the efficiency of the trained model.
- Train the model on training data.
- Use streamlit to create a web application that allows users to input their own data and make predictions using the saved model.
- Include a button in the application to process the user input and make predictions.

Algorithms Used :

Logistic Regression -

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability of an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyzes the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision- making.

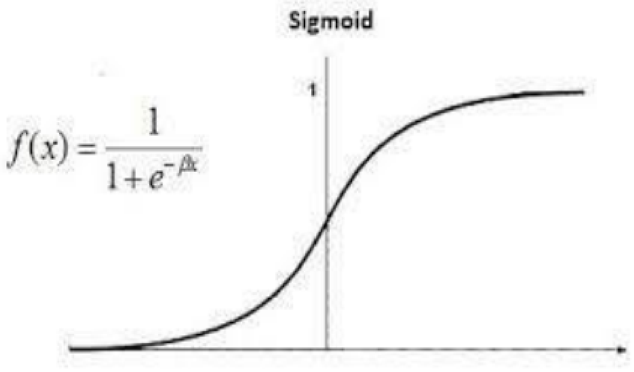
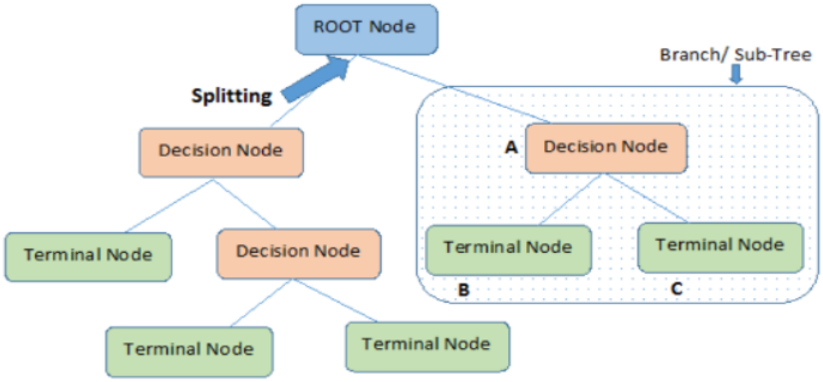


Fig.6: Sigmoid Function and Graph

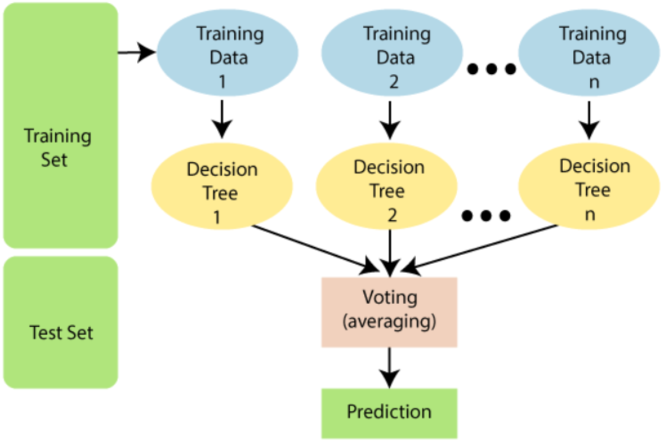
Decision Tree -

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving Classification problems. It is a tree- structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.



Random Forest -

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.



$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

n = Number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

Support Vector Machine (SVM) -

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct Category in the future.

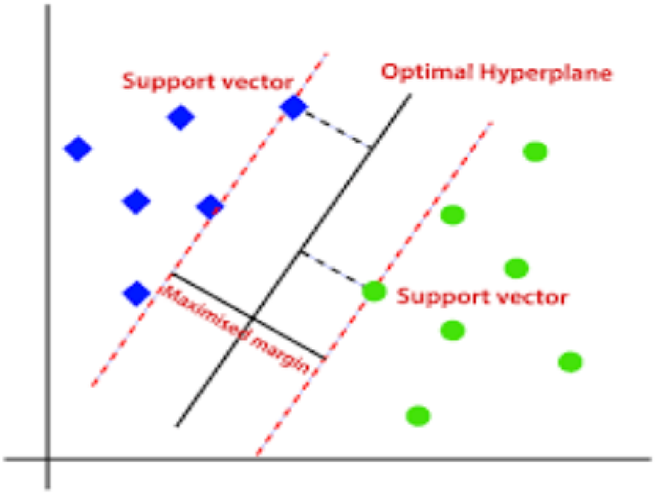


Fig.7: Optimal Hyperplane representation in Support Vector Classifier.

K-Nearest Neighbour (KNN) -

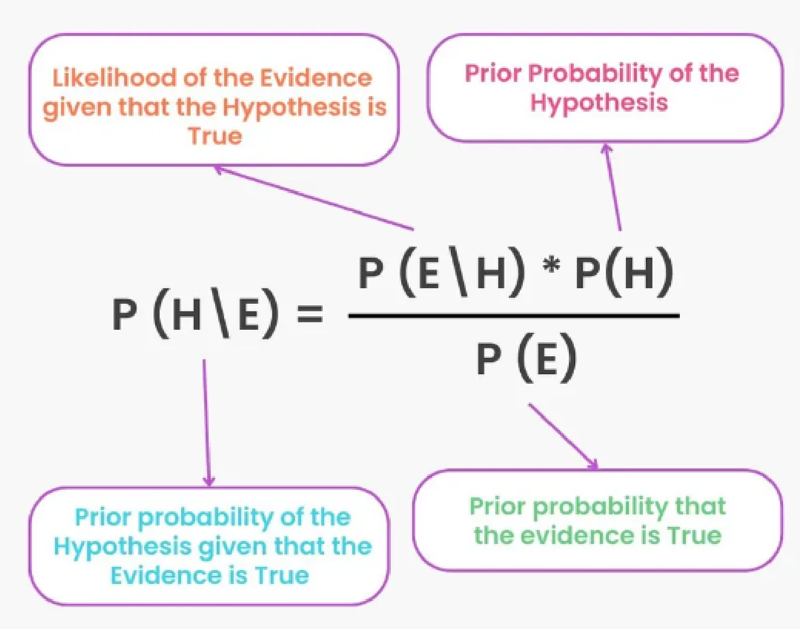
K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores 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-suited category by using K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}\left|x_i - y_i\right|$$

Naive Bayes -

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.It is mainly used in text classification that includes a high-dimensional training dataset. The Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.



$$P\,(H \backslash E) = \frac{P\,(E \backslash H) * P(H)}{P\,(E)}$$

Likelihood of the Evidence given that the Hypothesis is True

Prior Probability of the Hypothesis

Prior probability of the Hypothesis given that the Evidence is True

Prior probability that the evidence is True