Loan Eligibility Prediction in machine learning

Tarun Kumar Arcot(1NT18CS174), Rakshith R (1NT19CS413)

Abstract

In today's world loan disbursement is one of the major business lines for the banks. However, Loan disbursement can be a risky business if the banks are unable to take an informed decision on whom to give the loan and whom not. Banks ask for various information and documents to ascertain the credibility of the client.

In this project an attempt has been made to use machine learning algorithms like Logistic regression, naive bayes, Decision tree and random forest to predict whether the next client who applies for the loan will repay the loan in time or not and thus whether the bank should disperse the loan. Instead of human taking the decision which may be subjective, machine learning algorithms can help take accurate decision based on multiple information provided by the customer. In today's world of AI & ML, decision of whom to give loan is decided by the bank based on various past data trends and by using mathematical algorithm. ML has reduced the errors considerably in disbursement of loans and thus improve loan recovery in time, hence higher profitability to the banks.

Introduction

Machine learning (ML) is **a type of artificial intelligence (AI)** that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. The proposal is to predict whether the specified person will be paying the loan or not. The machine learning methods are Decision tree and Naive Bayes algorithms to predict the loan eligibility criterion,

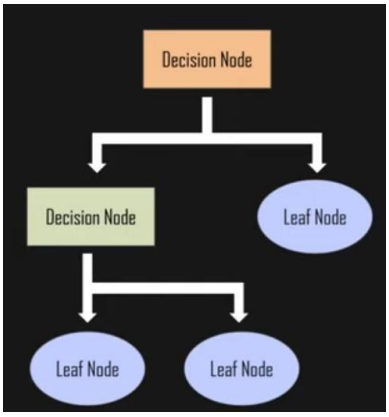
Data Set

There are two types of data sets used in the project 1)training data set and 2)testing data set. The data set includes:-
1) Loan_ID 2) Gender 3) Married 4) Dependents 5) Education 6) Self_Employed 7) ApplicantIncome 8) CoapplicantIncome 9) LoanAmount 10) Loan_Amount_Term 11) Credit history 12) Property_area 13) Loan_Status

Machine Learning Methods

1) Decision tree:-

A **Decision Tree** has many analogies in real life and turns out, it has influenced a wide area of **Machine Learning**, covering both **Classification** and **Regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. But in this we will use decision tree for classification.



2) 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. 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. **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**. Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

3) **Logistic Regression:-**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic Regression is much like the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. The equation for logistics equation: -

$$\log\left[\frac{y}{1-y}\right] = b0 + b1x1 + b2x2 + b3x3 + \cdots + bnxn$$

4) **Random forest:-**

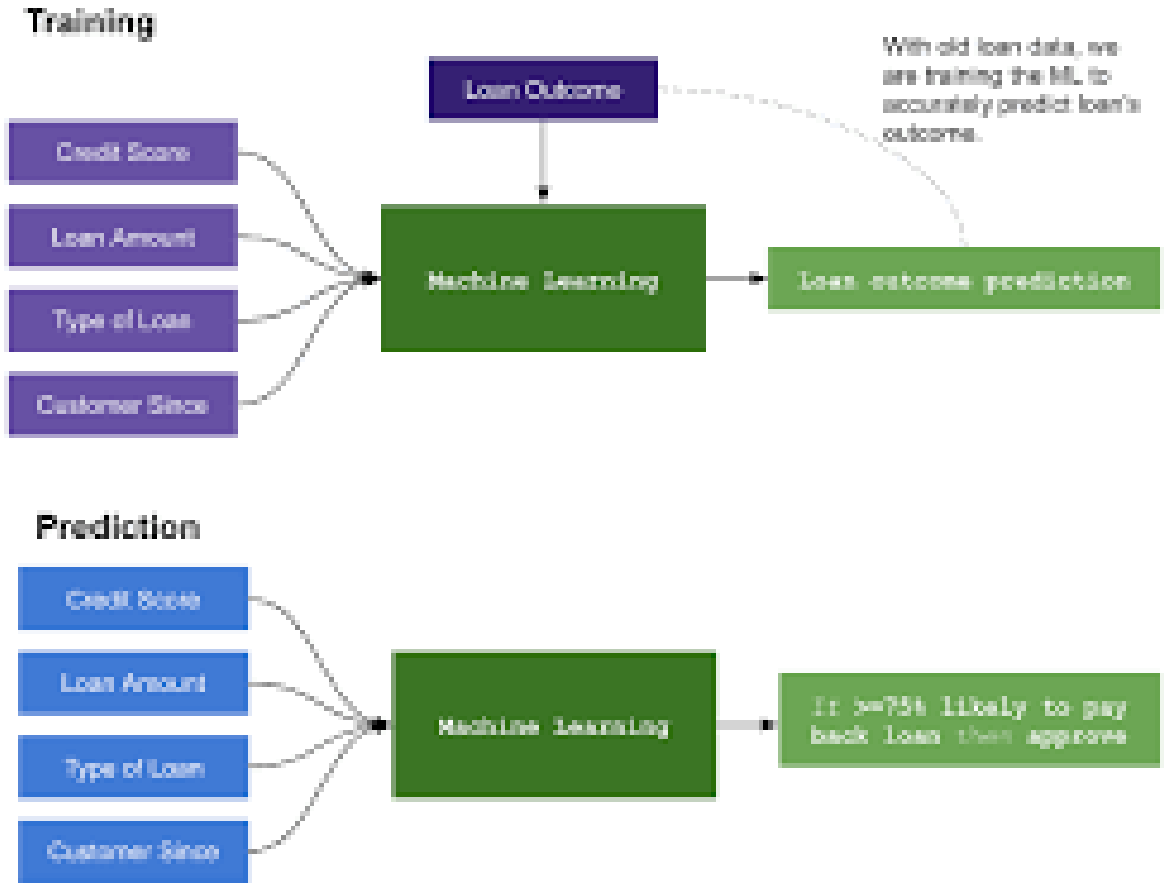
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

5) **Support Vector Machine:-**

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. This best decision boundary is called a hyperplane.

The graph is:



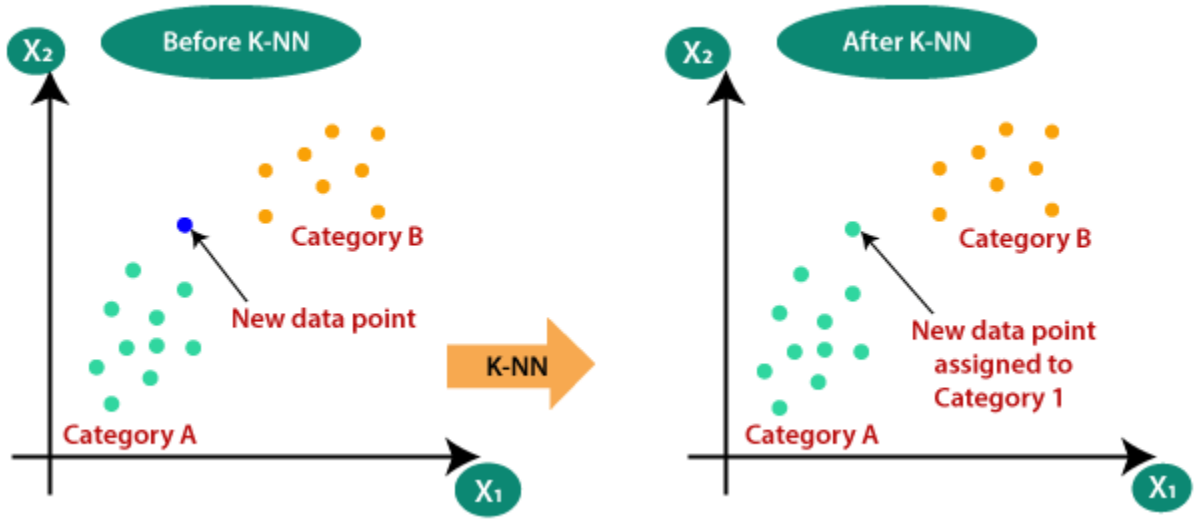
6) **KNN Algorithm:-**

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 suite 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.



Assessment:-

1) Accuracy
2)confusion matrix
(precision,recall,f1 - score, support)

Presentation and Visualization

Will use boxplot and histogram

Bibliography

[1] https://en.wikipedia.org/wiki/Decision_tree

[2] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[3] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset