

Treleaf Technologies Nepal



Implement Different Machine Learning algorithm For Bank Loan Default Prediction

By

Manoj Nepali

Treleaf Technologies Nepal

1. Introduction

This report presents the analysis and results of a bank loan classification task, aiming to build a machine-learning model that predicts whether a customer will accept a personal loan based on provided information. The dataset consists of 15 columns, including customer demographics, financial details, and historical banking behavior, with a corresponding target column indicating loan acceptance.

2. Data Preprocessing and Exploratory Data Analysis (EDA):

I analyze the data and need to data preprocessing and exploratory data analysis. During data preprocessing i found that there are many missing values in the dataset. There are about 32% value are missing in Gender column and 23% values are missing in Home ownership. And also drop unnecessary column like Zip Code column and ID. After that i found the unique value of Home owner ship and Gender and handling the missing value like most frequent value(mode). And duing preprocessing in gender there are 4 unique value male(M),female(F),other(O),- and #. I drop all the rows except M and F values. I also found that there are some null values about 1% in Income and online column so i dropped all rows whose contain null values.

I displayed some bar chart, box plot to detect the outlier. From the box plot i found that there are outlier in income and age. Like in age column there are some value greater than 90 which is not good for our model so i dropped all the rows whose contain age greater than 90. Likewise i also handled income column aswell.

3. Observation

After data preprocessing and EDA I want to decide to make a model with different algorithm. First of all i split the dataset into dependent columns and target column. Before implement the spillited data into algorithm, i did feature scaling in the dataset. Like age and income there are highest value rest of other column value. After feature

scaling i implement the dataset into different model like logistic regression, decision tree, random forest, SVM, and boosting classifier. I found that i got about 93% in logistic regression model, in SVM i got 88%, Decision tree i got 98%, in Radnom forest i got 97% and in gradient boosting classifier i got 98%. Here see that gradient boosting and decision tree algorithm performs best. And i use confusion matrix to determine the accuracy, precision and recall value of each model. After that i saved the model using pickle.

4. Conclusion

In this bank loan classification task, we successfully developed and trained a machine-learning model to predict whether a customer will accept a personal loan based on various customer attributes and historical banking behavior. The Gradient Boosting algorithm emerged as the best-performing model, achieving an impressive accuracy of 98% on the testing data.

Best regards