Loan Prediction Problem

This is a loan prediction problem from analyticsvidya.com website. Here we are given a training set and a test set. And, we are going to train a model using best suitable algorithm from machine learning and after that we will test the accuracy of our model using the test data set. Below are some of details about the loan prediction problem. We have taken the following details from analyticsvidya.com.

About Company

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.

Problem

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. Here they have provided a partial data set.

Data

|  |  |
| --- | --- |
| **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) |

Solution

Before discussing the solution, let’s discuss our working tools that we will be using in solving the problem. We have used the Rstudio as a working tool. Our dataset is stored in excel sheet format.

After analyzing the data set, we have concluded that the problem can be solved using the classification algorithm since the variable which we want to predict is categorical rather than numerical variable. After that, I have imported the dataset in to our Rstudio and started some data preprocessing. I have converted all the categorical variables into factors. Since our dataset has some missing data. So, I have replaced all the numerical missing data with the average of that column and the categorical missing data with the mode of that column. I have then applied feature scaling to all the numerical variables of our dataset.

After the data preprocessing step, I have trained the model using the glm function of the R programming language. After training the model, I applied k-fold cross validation to our model in order to check its accuracy. As a result, our accuracy was around 81.11%.

After training, I applied the testData to our model and predicted the Loan Status variables. I then uploaded the testData result to the analyticsvidya.com website and it gave me a test accuracy of about 78.47% which is an acceptable accuracy for someone who is a beginner in machine learning.

I have also uploaded the code of my model.

References

<https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/#ProblemStatement>