Applied Artificial Intelligence

CS451

Course Project Report

Khizar Hayat Khan 2016194

Introduction:

The problem that we are going to solve is binary classification problem. We have given a data set about a bank that has several properties about a person who has got a loan or who has rejected to get a loan from a bank. The variables of the data set are given below.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Not 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 | (Target) Loan approved (Y/N) |

The goal of this project is to build a model that will accurately predict whether to give a loan to a new customer or not based on several characteristics of the customer.

Tools:

We have used Jupyter Notebook and have used the following libraries of Python.

1. Pandas for getting and manipulating data
2. Scikit Learn for machine learning
3. Matplotlib and Seaborn for data visualization

Methodology:

We have followed the following methodology in our project

1. After importing data into our environment, we identified the categorical and numerical data in the data set.
2. We then found the number of missing values in each column of the data set.
3. We handled the missing values of categorical data by filling it with most frequent value.
4. Similarly, we handled the missing values of numerical data by filling it with the mean of the column.
5. After this, we converted the categorical data into 1 and 0 format.
6. Then, we split the dataset into train and test set.
7. We applied different algorithms using 5-fold cross validation and stored the accuracy
8. We then did some feature engineering and re applied different algorithms
9. Lastly, we tuned the hyperparameters of each algorithm

Analysis:

Now let’s analyze out result. Following is the accuracy table that we have gotten after applying each machine learning algorithm.

|  |  |
| --- | --- |
| **Algorithm Name** | **Accuracy (%)** |
| Logistic Regression | 68.56 |
| Support Vector Classifier (Linear) | 79.50 |
| Decision Tree Classifier | 80.13 |
| Random Forest Classifier | 78.83 |

One of the cons of Logistic Regression is it doesn’t handle large number of categorical features. This could be one of the reasons that it is not performing well on our dataset. Other reason could be the hyperparameter of the algorithm. We will compare the result after we reach to analysis after feature selection and performance tuning.

Support Vector using Linear kernel, Decision Tree and Random Forest Classifier has performed good in terms of accuracy. We can improve the performance of each algorithm by selecting the important features from the data set and removing the unimportant features. Some of the variables in our data set are continuous variables and can possibly have outliers which can affect the accuracy of the model. We remove those outliers as well. Another method for improving the accuracy of the model is to properly tune the hyperparameters of the model. A hyperparameter is a parameter whose value is used to control the learning process.

We did exactly as describe in the above paragraph and we have gotten the following accuracies for each model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Accuracy (%) before feature engineering etc.** | **Accuracy (%) after feature engineering etc.** | **Change in Accuracy (%)** |
| Logistic Regression | 68.56 | 81.65 | 13.09 |
| Support Vector Classifier | 79.50 | 81.83 | 2.33 |
| Decision Tree Classifier | 80.13 | 80 | -0.13 |
| Random Forest Classifier | 78.83 | 80.55 | 1.72 |

Here we can see that Logistic Regression has improved a lot as compare to other algorithms. We have dropped two categorical variables from our data set because we thought it is not a good feature. Also, we have changed the hyperparameter of Logistic Regression. We trained the algorithm on an array of hyperparameter and selected the one which had higher accuracy. These can be the reason due to which Logistic Regression has improved a lot.

Support Vector Classifier has improved 2.33% because the initial hyperparameter of the classifier was not an optimized one. After we have change its hyperparameter, we have gotten an improved accuracy.

The accuracy of Decision Tree Classifier has dropped a little bit. This could be due to dropping of two categorical variables that we taught are not good predictors.

Similarly, the performance of the Random Forest Classifier has improved because we have tried out this algorithm with different number of trees and selected the one with higher accuracy.

Conclusion:

When solving a machine learning problem, data pre-processing and feature selection are important steps to consider. Without doing these steps we will not be able to get good models. Also, it is recommended to tune the hyperparameters of algorithms. It helps to get good results as well.