MA515 PROJECT REPORT

Foundations of Data Science

A Project Report Submitted by

Tata Charishma

2020MCB1252

Mathematics and Computing

Under the supervision of

Dr.Arun Kumar

Department of Mathematics

Indian Institute of Technology, Ropar

December 1,2022

Contents

Problem Statement :

Do exploratory data analysis on the data. Use logistic regression and KNN with different K to predict the credit score. Identify the most important features in the data. Compare the findings from different methods.

Data Preprocessing :

1. Taking Care of Missing Data :

   We will check for any NaN values and remove them if they are any. In given data, there are no NaN values.

2. Encoding Categorical Data :

   - Dependent Variable :

     We will take the Credit_Score in y and apply Label Encoder to change its categorical data to numerical data.

   - Independent Variable :

     Drop the name columns as it coincides with the ID column. Encode type_of_loan column using frequency encoding. Using pd.get_dummies and sending the columns that are required to encode we can change the remaining categorical data into numerical data.

3. Splitting dataset into training and testing :

   By taking test size as 0.2 divide the dataset into training and testing.

4. Feature Scaling :

If some column values are much larger than others we will scale those column values using Standard Scaler.

Exploratory Data Analysis :

- Find the summary of the data using data.describe() command. It gives count, mean, standard deviation, minimum and maximum values of each column.

- Plot the heat map to visualize the correlation between features. We can see that Annual_Income and Monthly_Inhand_Salary are highly correlated. So we can drop any one of the columns.

Logistic Regression :

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. fit() takes the independent and dependent values as parameters and fills the regression object with data that describes the relationship. By using that relationship we will predict x_test datapoints. After predicting we can see the confusion matrix as follows:

[[2329 87 1173]
 [508 3168 2150]
 [1444 1460 7681]]

From the Confusion Matrix, we can say that 13,140 values are predicted correctly.

Test accuracy from logistic regression is 0.6589.

Train accuracy from logistic regression is 0.6631.

KNN as Classifier :

The K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line. Therefore, a larger k value means smoother curves of separation resulting in less complex models. Whereas, smaller k values tend to overfit the data and result in complex models.

For K = 3
Model Accuracy = 0.6938
Confusion matrix = [[2291 146 1152]
 [515 3704 1607]
 [1290 1414 7881]]

For K = 50
Model Accuracy = 0.6804
Confusion matrix = [[2458   35 1096]
 [642 3259 1925]
 [1458 1236 7891]]

For K = 100
Model Accuracy = 0.6721
Confusion matrix = [[2461   27 1101]
 [654 3175 1997]
 [1489 1290 7806]]

For K = 200
Model Accuracy = 0.65815
Confusion matrix =[[2330   31 1228]
 [620 3129 2077]
 [1476 1405 7704]]

For K = 300
Model Accuracy = 0.65075
Confusion matrix =[[2229   40 1320]
 [571 3141 2114]
 [1420 1520 7645]]

For K = 500
Model Accuracy = 0.64085
Confusion matrix =[[1971 50 1568]
 [488 3134 2204]
 [1247 1626 7712]]
For K = 3, we got a maximum model accuracy of 0.6938.

Feature Importance :

- Fit the Random Forest Classifier.
- To get the feature importances from the Random Forest model use the feature_importances_ attribute.
- By the plot, we can see that Outstanding_Debt is the most important feature and next to that we have Interest_Rate, Credit_Mix, Credit_History_Age.

Comparison :

- Accuracy attained by Logistic Regression is 0.658.
- Highest Accuracy attained by KNN is 0.6938 at K =3.
- But this K value is very less.
- Accuracy attained by KNN at moderate K that is K = 200 is 0.658.
- By accuracy values we can see that both performances are the same. For some smaller values of K, we can observe that KNN has higher accuracy than Logistic Regression.