MA515 - Foundation of Data Science

Project Report: Credit

Submitted to: Dr. Arun Kumar
Submitted by: Apurva Pragya (2020MCB1229)
Submitted on : 30th November 2022

Link: google colab

INTRODUCTION

Dataset:

The data is stored in a CSV file named Credit.csv

Credit.csv contains 400 rows and 11 columns. The columns contain the following information of the user:

Income	Income of the user
Limit	Credit Limit
Rating	Credit Rating
Cards	No. of cards owned by the user
Age	Age of the user
Education	No. of years for which the user received education
Gender	Gender of the user
Student	Whether the user is a student or not
Married	Marital status of the user
Ethnicity	Ethnicity of the user
Balance	Balance of the user

First, The Categorical data in the columns 'Gender', 'Student', 'Married', and 'Ethnicity' was changed to numerical data. Since the data had no outliers and no missing values, no additional filtering was used.

Further, relationship analysis was found out on the variables. A histogram of each feature and a plot of relation between each variable and the target variable was made to analyze the distribution of data for each variable and study the relationship of each predictor with the target variable. It was noticed that 'Gender', 'Marital status', 'Ethnicity' and 'Student' is not correlated to credit rating. A heatmap for the correlation matrix was also created to evaluate the relationship between the variables. It was created the correlation matrix shown below:

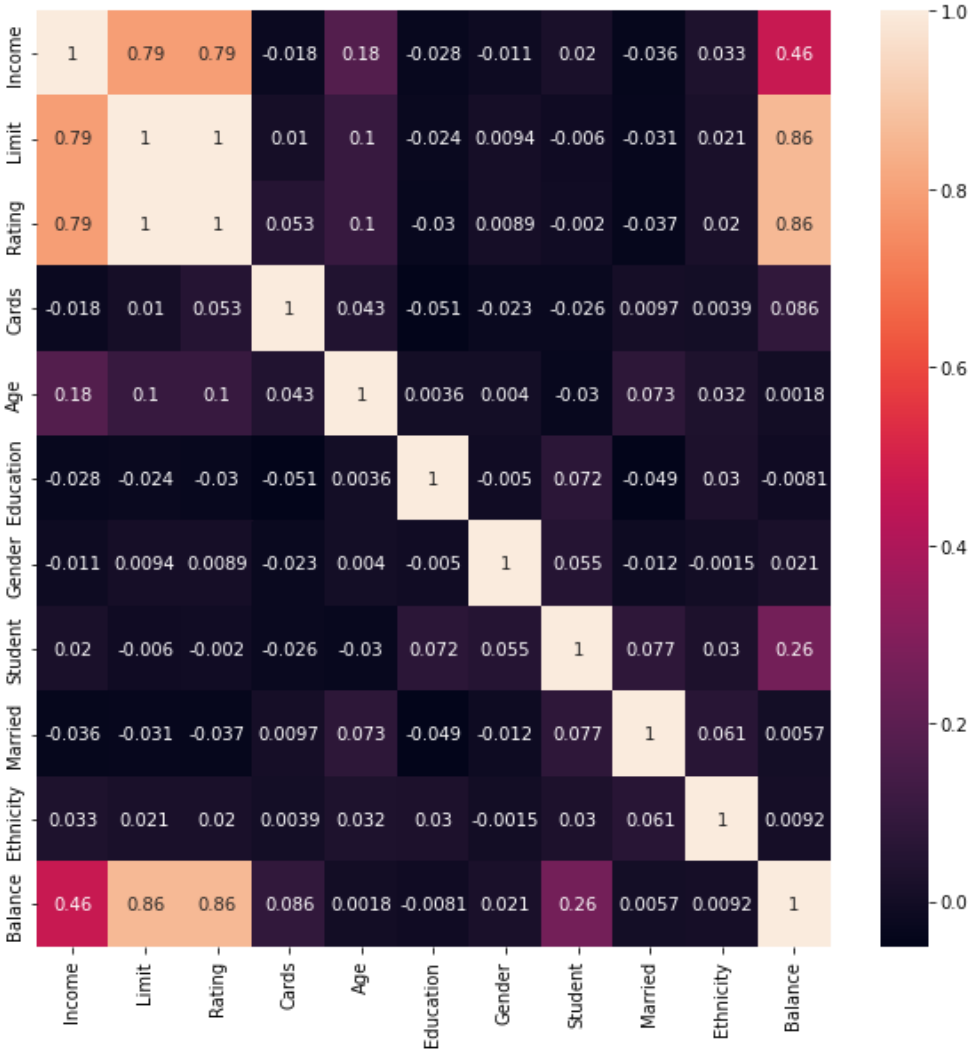

Fig 1 : Correlation matrix of the dataset

The above correlation matrix suggests a high correlation between Rating and Limit.

After finding the correlation and histograms for each feature, Regression analysis was done on the dataset. Since there are multiple predictor variables and single target variables, I have used multilinear regression to predict the credit rating.

The following steps were taken to make the regression model:

Multilinear regression:

I did not do feature scaling on this set of data since I am using linear regression, which uses matrices rather than gradient descent.

sklearn.linear_model.LinearRegression was used to create a multilinear regression model. The following coefficients were found after training this linear regression model on the training data:

a1=0.09642409

a2=0.06440052

a3=4.63981872

a4=0.01336632

a5=-0.2916367

a6=0.56793586

a7=0.52993203

a8=-2.59196404

a9=0.05806422

a10=0.00729516

The intercept is a0=31.29277471275327.

Hence the equation for multilinear regression is

$Y = a0 + a1*x1 + a2*x2 + a3*x3 + a4*x4 + a5*x5 + a6*x6 + a7*x7 + a8*x8 + a9*x9 + a10*x10$

Where,

x1 ='Income', x2='Limit', x3='Cards', x4= 'Age', x5='Education', x6= 'Gender', x7='Student', x8= 'Married', x9='Ethnicity', x10= 'Balance'

Using the above relationship, the model predicts the credit rating for training as well as for testing data, and attained values of R2 are 0.99566 and 0.99596 for training and testing data respectively.

Lasso:

For lasso regression, feature selection was performed.

Three features, namely "Ethnicity," "Married," and "Students," were shown to have the least significance in the prediction of credit rating.

After removing the non-significant predictors from the earlier data, new training and testing data were constructed. This brand-new data was then used to a Lasso regression model.

The lasso regression model's coefficients were as follows:

a1=0.10223557

a2=0.06431463

a3=4.54547386

a4=0.00635932

a5=-0.26884384

a6=0.16886591

a7=0.00760127

And the value of alpha was chosen as 0.1

Hence the equation for lasso regression model become:

$Y = a1*x1 + a2*x2 + a3*x3 + a4*x4 + a5*x5 + a6*x6 + a7*x7 + \alpha*\sum|a_i| \ , \ i = \{1,2,3,4,5,6,7\}$

Where,

x1='Income', x2='Limit', x3='Cards', x4= 'Age', x5='Education', x6='Gender', x7= 'Balance'

Using this relation, the model predicts the credit rating for the training and testing dataset and acquires 0.9956, 0.99598 values of R2 for the training and testing data respectively.

Findings:

The credit score is primarily influenced by the user's income, credit limit, number of cards held, level of education, gender, and account balance.

Conclusion:

- According to EDA, ethnicity, marital status, and student status do not significantly influence the prediction of credit ratings.
- It is clear that the feature selection boosted the R2 score.
- The model's theoretically predicted increase in accuracy and efficiency matched our outcomes.
- The model is now more efficient because the amount of space the data occupied has decreased and the model can now predict outcomes rapidly.

A key method for improving the efficiency of machine learning algorithms is feature selection. Additionally, it is very important to look for predictor collinearity. Utilizing dependent or irrelevant predictors will make the model less effective and add to its will increase the time complexity. So, they ought to be removed before fitting the model.