Milind Rathore
2020MCB1241

Exploratory Data Analysis Linear Regression and QDA Classification

Introduction

Data stored in OJ.CSV file with 18 Columns and 1070 Rows. By using the data set, we will find EDA on the dataset. And perform classification on Linear Regression & QDA.

Exploratory Data Analysis:

First, The Categorical data in the columns ‘Store7’ was changed to numerical data. Since the data had no outliers and no missing values, no additional filtering was used. Reading the data manually to get some insights into the format of the data. Finding the unique values in each column. Checking for missing values in the data. Storing the unique category of each categorical data.

Making the correlation matrix between the features. A heatmap for the correlation matrix was also created to evaluate the relationship between the variables. It was created the correlation matrix shown below: -

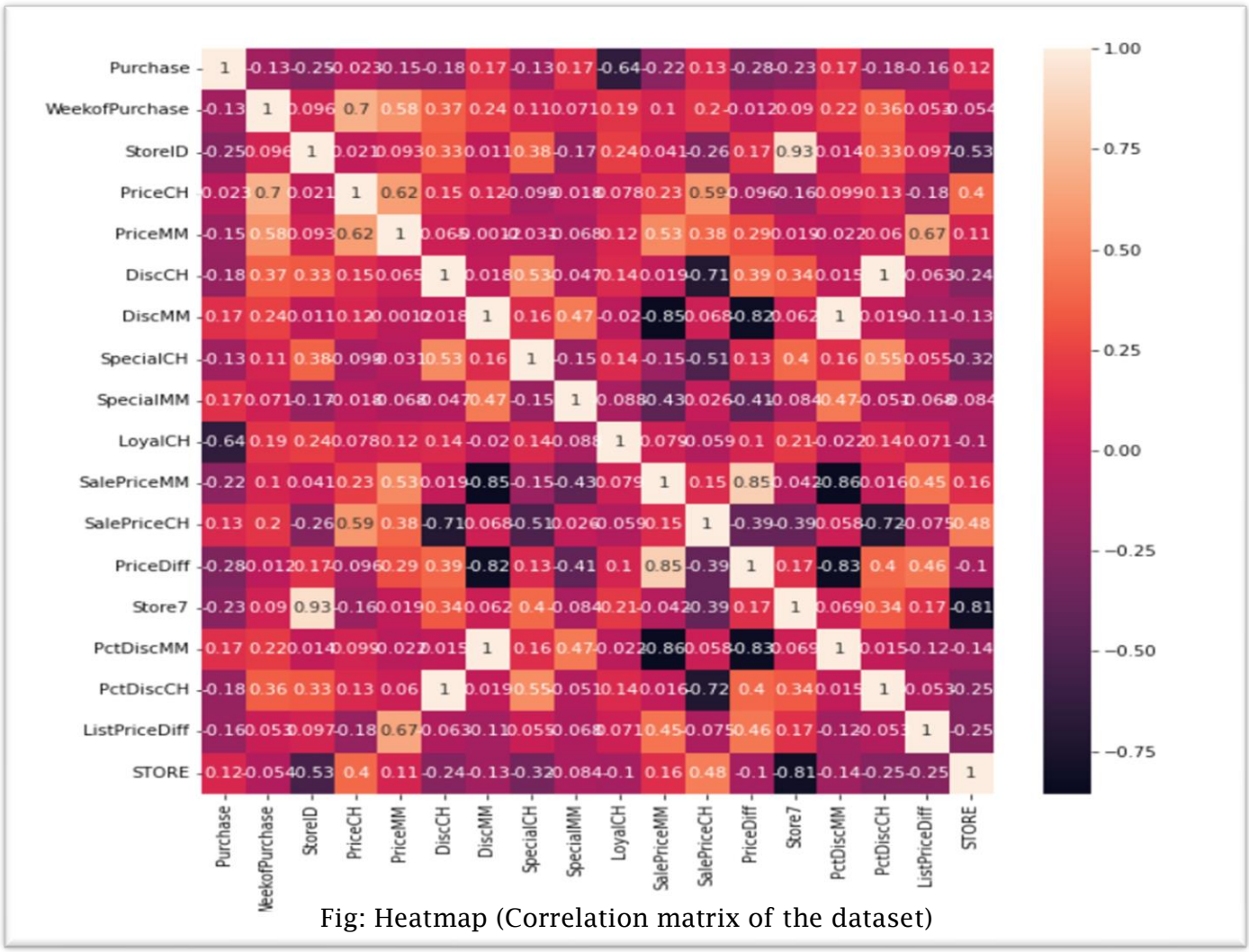

Fig: Heatmap (Correlation matrix of the dataset)

Shuffling the data and splitting it into training and test dataset

```
# shuffling the data and splitting it into training and test dataset
dataset.sample(frac=1)
#storing the predictor variable data in data_x
data_x= dataset.drop('Purchase',axis=1)
data_x=data_x.to_numpy()
#storing the target variable data in data_y
data_y= dataset['Purchase']
data_y=data_y.to_numpy()
#Splitting the data into testing and training data.

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(data_x,data_y,test_size=0.3)
```

Milind Rathore
2020MCB1241

Linear Regression Classification:

Train Data results before feature selection

```
Train Data Results Before Feature Selection:
              precision    recall  f1-score   support

          CH       0.87      0.88      0.87       464
          MM       0.80      0.79      0.79       285

    accuracy                           0.84       749
   macro avg       0.83      0.83      0.83       749
weighted avg       0.84      0.84      0.84       749

[[406  58]
 [ 59 226]]
```

Test Data results before feature selection

```
Test Data Results Before Feature Selection:
              precision    recall  f1-score   support

          CH       0.83      0.87      0.85       189
          MM       0.80      0.75      0.77       132

    accuracy                           0.82       321
   macro avg       0.82      0.81      0.81       321
weighted avg       0.82      0.82      0.82       321

[[164  25]
 [ 33  99]]
```

QDA:

Code;

```
# QDA
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
qda = QuadraticDiscriminantAnalysis()
qda.fit(x_train,y_train)

y_pred_qda_train = qda.predict(x_train)
y_pred_qda_test = qda.predict(x_test)

print('QDA on training Data :\n\n',classification_report(y_train,y_pred_qda_train))
print('\n\nQDA on testing Data :\n\n',classification_report(y_test,y_pred_qda_test))
```

QDA on Training data

```
QDA on training Data :

              precision    recall  f1-score   support

           0       0.68      0.83      0.75       464
           1       0.57      0.38      0.46       285

    accuracy                           0.66       749
   macro avg       0.63      0.60      0.60       749
weighted avg       0.64      0.66      0.64       749
```

QDA on Testing data

```
QDA on testing Data :

              precision    recall  f1-score   support

           0       0.70      0.86      0.77       189
           1       0.69      0.46      0.55       132

    accuracy                           0.69       321
   macro avg       0.69      0.66      0.66       321
weighted avg       0.69      0.69      0.68       321
```

Milind Rathore
2020MCB1241

Conclusion:

As we can observe that here Linear Regression is more considerably than Quadratic Discriminant Analysis (QDA) by checking their harmonic mean of precision and recall (i.e f1-score).