LINEAR REGRESSION PROJECT & CLASSIFICATION TREE

Team:

Ronit Naik
Nihal Mehta
Rutansh Trivedi
Masum Shah
Shubham Oberoi

CSC177

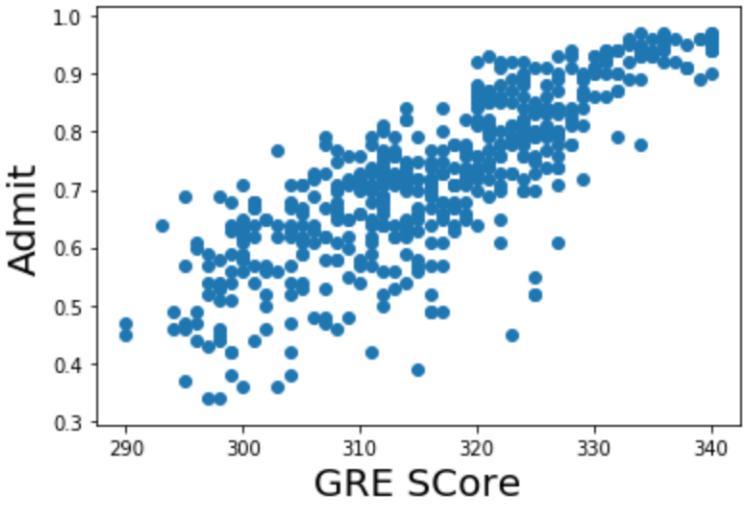
Introduction:

In simple linear regression a single independent variable is used to predict the value of a dependent variable. In multiple linear regression two or more independent variables are used to predict the value of a dependent variable.

1) Simple Linear Regression

Steps Performed:

1) The dataset we have used is about Admission prediction which has GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research and Chance of Admit as features.
2) We checked for rows with duplicate and Null values which is none. The dataset is clean.
3) For Simple linear Regression, we have selected GRE Score to predict chances of admission. The columns for simple linear regression are x as GRE Score and y as Chance of Admit.



4) Correlation between Admit and GRE Score can be seen in the above Scatterplot.
5) Normalized the value using z-score. We divided the training set into 80:20 ratio. After splitting the dataset, we converted x-train, x-test, y-train, y-test to NumPy array.
6) We trained the model using x-train, y-train. We get predicted y-value using x-test.
7) We calculated mean square error using y-test and y-predict.

2) Multiple Linear Regression:

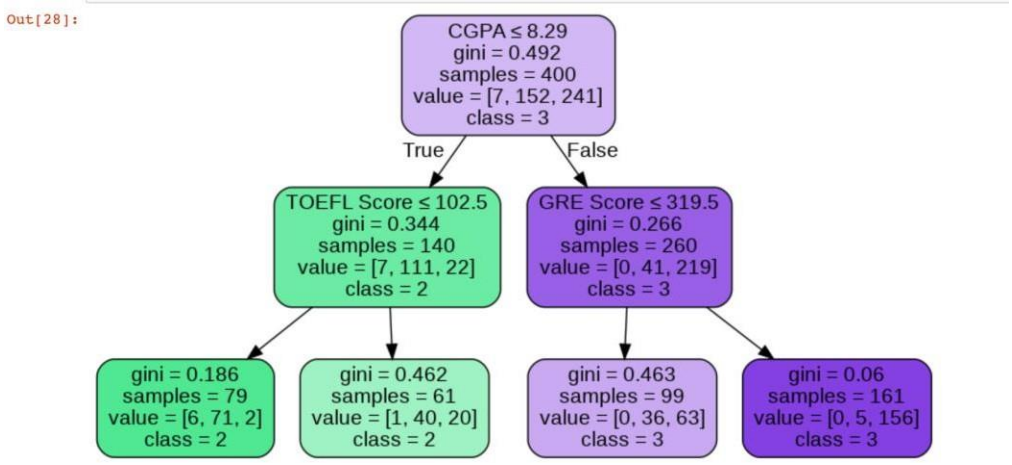
For multiple linear regression we used GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research (7 features) to predict the chances of admit.

3) Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

Steps Performed:

1) We used the same dataset of solving problem of chances of admit however now we have created categories out of 'Chances of admit' values.
2) To create the categories out of Chance of admit values we have used bins. We defined the bin range & bin names. We used bin attributes to discretize chance of admit values which gives us three classes because we have defined three bins.
3) We used GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research (7 features) as x and the above created chance of admit as y.
4) We spit the dataset 80:20 ratio. We created Decision Tree Classifier with Max depth of tree as 2.
5) We trained the above created classifier using x train y train. We performed predictions using x test.
6) Below is the visualization diagram that represents decision tree.



Insights:

In this Dataset, in simple linear regression we can see that GRE score is directly related to chances of getting admit from the university, while in the multiple linear regression we have taken the 7 features and checked those attributes relation with getting admit from the university. In classification we have created 3 categories to discretized chances of getting admits.

4) Housepricing Dataset (Old Dataset):

We use House pricing dataset which we had used in data-preprocessing project to predict sales price analysis using linear regression and multiple linear regression.

Steps performed:

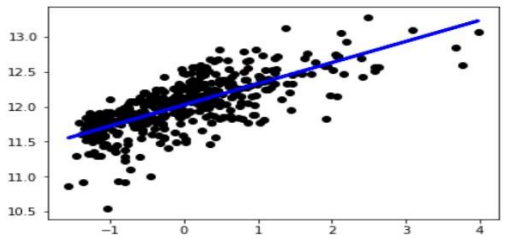
1) We performed data-preprocessing by removing duplicate data and filling null values with median values.
2) We performed feature importance analysis by using 'Random forest' for feature extraction. Out of 81 original features we selected top 30 features for framing our model.
3) We normalized numerical value using z-score. Detected and removed the outliers in the dataset.
4) We split the data into 70:30 ratio. For Simple linear regression we have predicted sales price using GrLivArea: Above grade (ground) living area square feet. For multiple linear regression we predicted sales price using GrLivArea, OverallQual: Overall material and finish quality and TotalBsmtSF: Total square feet of basement area.

Screenshots:-
Simple Linear Regression

```
In [132]: x= np.array(X_train['GrLivArea']).reshape(-1,1)
          model=LinearRegression()
          model.fit(x,y_train)
          print(model.score(x, y_train))
          x_new=np.array(X_test['GrLivArea']).reshape(-1,1)
          y_pred=model.predict(x_new)
          rmse=mean_squared_error(y_test, y_pred)
          print(rmse)
          plt.scatter(x_new, y_test,  color='black')
          plt.plot(x_new, y_pred, color='blue', linewidth=3)

          0.524421103018032
          0.06890071552592401

Out[132]: [<matplotlib.lines.Line2D at 0x7ff36b68c908>]
```



Insights:

In this Dataset, we have implemented simple linear regression and multiple linear regression on our old dataset. From the above diagram we can see that in simple linear regression attribute GrLivArea attribute is directly connected with the house's sale price.