Final – Data and Regression

Background Information
For this final we are going to look at a data set with performance data about a number of automobiles. The original use of this data was as a testbed for graphical analysis packages at the 1983 American Statistical Association Exposition. After the exposition, it was added to the CMU Statistics Library. http://lib.stat.cmu.edu/datasets/

Quinlan cleaned the data a bit and used it in a paper he wrote in 1993. He added it to the UCI data sets at that point. (https://archive-beta.ics.uci.edu/ml/datasets/auto+mpg).

We will take the data set from Kaggle instead, because Kaggle has the data in the form of a CSV file. I did a quick scan to verify that the Kaggle version had the same data.

Here is a list of the features in the dataset https:
https://www.kaggle.com/uciml/autompg-dataset

1. *mpg* – Miles per gallon
2. *cylinders* – Number of cylinders in the engine
3. *displacement* – Displacement of the engine in cubic inches. (Roughly, it is the combined size of the cylinders in the engine)
4. *horsepower* – Amount of power developed by the engine.
5. *weight* – Base weight of the car in lbs.
6. *acceleration* – Time to go from 0 to 60 miles per hour in seconds.
7. *model year* – Year of manufacture a value from 70 to 82, indicating the years 1970 through 1982
8. *origin* – Manufacturer location (1-USA, 2-Europe, 3-Japan)
9. *car name* – Make and model

I have downloaded and created a link to the data set for convenience on the Canvas Final Assignment.

Question 1 (20 points):

A) Create a notebook and read in the ***autompg*** dataset and give a screen shot of the head of the data set.

B) Use kaggle to look at the distributions of the data values in the **autompg** dataset and point out any unevenness in the distributions.

C) Make an argument for which two input features you would want to use to best predict target mpg. If cars are not something you know much about, you can consult with someone more

knowledgeable (a friend or relative). You can also reach out to me. If you use outside help for this, document it in the submission.

Submission 1 of 3:

- (5) Screen shot of data set head.
- (5) Discuss distributions of features in autompg data set.
- (10) Discuss your two important features.
    - Outside reference if used.

Question 2 (20 points):

A) What ***new*** feature would you like in the ***autompg*** data set that could allow you to make a better prediction of the mpg (**target**)? Give a story for how it could affect mpg. You are allowed to consult.

B) In Scikit learn train a regression model on the ***autompg*** data set.
Steps to include:

a) ***Read Dataset***.
b) ***Select Features in DataSet for X and y*** *(Your 2 chosen input features and target)*
c) ***Linear Regression***
d) ***Train Model***
e) ***Evaluate Model.***

Take a screen shot of the code.

C) Report the standard metrics and the bias/coefficients. Take a screen shot.

D) Look at the coefficients and make an argument about the relative importance of the two input features. Remember to take into account any differences in the scale of the two features.

Submission 2 of 3:

- (5) Your desired feature and story.
- (5) Screen shot code.
- (5) Screen shot of metrics.
- (5) Discuss feature importance

Question 3 (20 points):
Create pipelined models on the autompg data set with the same two features from before and with mpg as the target.

The pipeline 1 will include

1) an Imputer with median strategy
2) a standard scalar
3) linear regression

The pipeline 2 will include

1) an Imputer with median strategy
2) polynomial features with degree=3
3) a standard scalar
4) linear regression

A) Code for pipelines
B) Report the performance metrics for pipeline 1
C) Report the performance metrics for pipeline 2
D) Compare the performance of the three models.

Note: A model with an R2 of 80% or better is a pretty good model. Over 90% is very good. As long as the model did not overfit, which would require sequestering a test set.

Note: You expect that the predicted value has an error bar of roughly plus/minus twice the MAE.

Submission 3 of 3:

- (5) Screen shot code for both pipelines
- (5) Screen shot of Metrics pipeline 1
- (5) Screen shot of Metrics pipeline 2.
- (5) Comparison of the models. (There should be three of them.)