Tanner Young

Final

Applied Machine Learning

**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.**

A screenshot of a computer

Description automatically generated

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

MPG, displacement, and weight are skewed right. Acceleration has a normal shape distribution. Model year has a ragged plateau distribution.

**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.**

Based on the features provided I would choose Engine Displacement and Vehicle Weight. Generally, larger engine displacements tend to consume more fuel, leading to lower mpg. The weight of a vehicle significantly impacts its fuel efficiency. Heavier vehicles tend to consume more fuel, resulting in lower mpg.

**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.**

One potentially impactful feature that could enhance the prediction of MPG in the dataset could be Fuel Injection Type or System. The addition of information regarding the type of fuel injection system utilized in the vehicles could significantly influence the predictive power of the model for MPG. Different fuel injection systems (such as direct injection, port injection, or carbureted systems) have varying efficiencies in delivering fuel to the engine cylinders.

**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.**

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer program

Description automatically generated

**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.**

Displacement Coefficient (-0.0184): This coefficient implies that, on average, for every one-unit increase in engine displacement (cubic inches), the predicted mpg decreases by approximately 0.0184 units.

Weight Coefficient (-0.0057): For every one-unit increase in vehicle weight (presumably in pounds), the predicted mpg decreases by approximately 0.0057 units.

Considering the magnitudes of the coefficients, the displacement feature seems to have a relatively higher impact on the predicted mpg compared to the weight feature.

**A) Code for pipelines**

A screen shot of a computer

Description automatically generated

**B) Report the performance metrics for pipeline 1**

A screenshot of a computer program

Description automatically generated

Pipeline 1 Metrics:

Mean Squared Error: 14.991566354406825

R-squared Score: 0.7211724393728975

Intercept (Bias) - Pipeline 1: 23.60817610062893

Coefficients - Pipeline 1:

displacement: -1.8943995104230102

weight: -4.790095747900251

**C) Report the performance metrics for pipeline 2**

A screen shot of a computer

Description automatically generated

Pipeline 2 Metrics:

Mean Squared Error: 12.572387598221146

R-squared Score: 0.7661666511424955

Intercept (Bias) - Pipeline 2: 23.60817610062894

Coefficients - Pipeline 2:

displacement: 0.0

weight: -6.642864828588152

**D) Compare the performance of the three models.**

Pipeline 2 incorporates polynomial features (degree=3), scaling, and linear regression. It demonstrates better performance with an MSE of 12.5724 and a higher R-squared score of 0.7662.

Pipeline 2 shows the best performance among the three models, indicating that incorporating polynomial features helped capture a nonlinear relationship between features and mpg, resulting in improved predictive capability.

Pipeline 1 and Model 1 have similar performance metrics, but the coefficients differ due to preprocessing steps like scaling in Pipeline 1, which can affect the interpretation of the coefficients compared to the linear regression model without scaling.