Problem 1 – Lasso regression (35%)

References:

- Ridge and Lasso Regression Explained
- How does Lasso regression(L1) encourage zero coefficients but not the L2? | by Mukul Ranjan | Medium
- What is LASSO Regression Definition, Examples and Techniques
- How to implement cross_val_score in sklearn

Using the dataset 'auto_mpg.csv' included in the assignment, perform lasso regression to predict car mileage as a function of the remaining variables in the dataset (excluding car name).

15% credit. Generate a plot of the MSE on validation data as a function of the regularization parameter via 10-fold cross-validation (see figure 6.12 in the ISL book).

After parsing through the data, doing necessary data type conversions, and splitting the data into features and target, a range of alpha values from 10^{-4} and 10^{4} were tested using 10 fold cross validation in order to observe and select the ideal tuning parameter.

The cleaned data was thus split into 10 partitions, with one partition being used for testing and the rest being used for testing. Thus, for 100 selected hyperparameter alphas, we tested the performance of the model 10 times using the above described testing/training strategy, and took the average MSE for all the models of a particular alpha tested and trained using the KFold cross validation strategy.

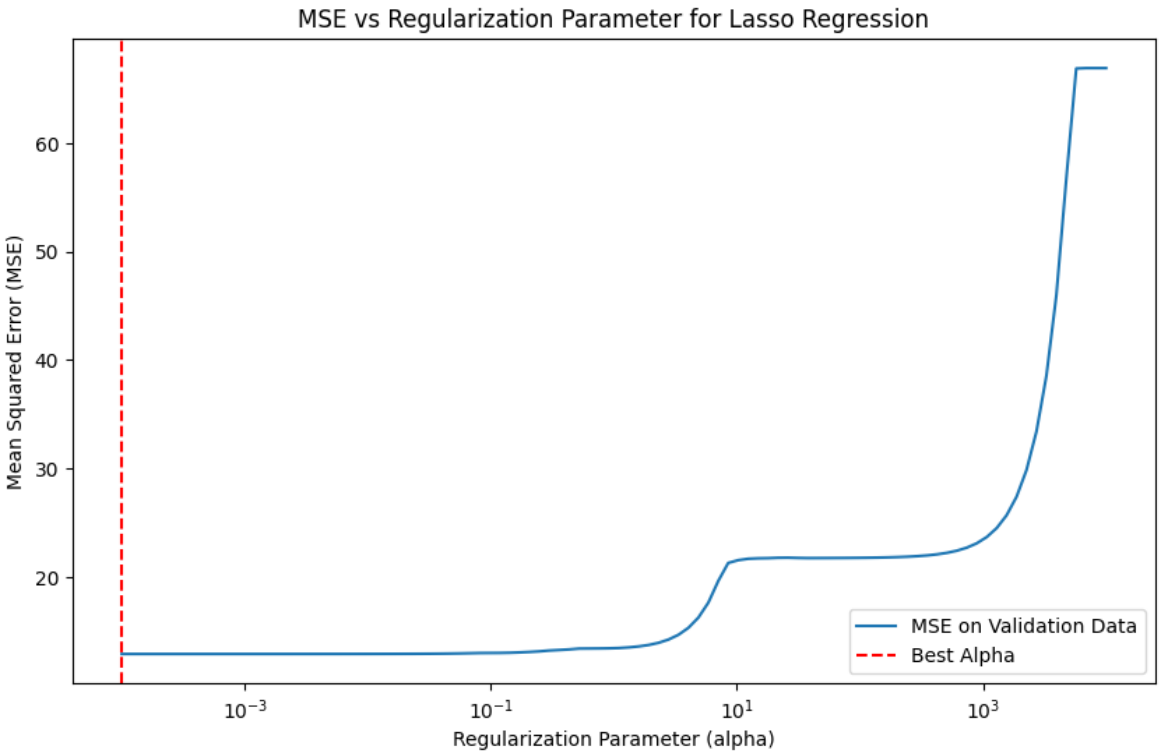

Figure 1: Average MSE vs Regularization Parameter for Lasso Regression

Discussion of findings:

Based on Figure 1, it appears that the ideal alpha, or penalty term, is the smallest term we tested in our range of alpha values. For this run, the ideal alpha was thus $\alpha = 0.0001$ with an MSE score of 12.9. Because our alpha value is relatively small, this tells us that while the coefficients are shrunk due to the penalty term, no particular coefficients are encouraged to be exactly 0. This tells us that an ideal model allows for more variables to have non-zero coefficients.

However, based on the graph above, there appears to be a growing trend of model performance decreasing (i.e. larger mean squared errors for models) as the Regularization Parameter increases. This is especially interesting, as in the figure above, the mean squared error appears to plateau at two different locations: for penalty terms between 10 and 100, and for penalty terms larger than around 1000. This tells us that penalty terms that lie within this region of the figure do not make meaningful contributions to the success of the model, as neither increasing or decreasing these alpha values incrementally gives us a model with increased or decreased performance (based on macroscopic view). As the Lasso Regression model attempts to regularize the linear regression model by using higher penalties, the model does not increase in performance.

Because the smallest tested alpha value was shown to gain the best MSE score, this tells us that the model could potentially be ideally fit by the Ordinary Least Squares Solution, as when alpha is equal to 0, Lasso Regression produces the OLS solution.

Overall, this tells us that the data provided can one, be ideally predicted using Linear Regression, and two, requires little to no regularization to prevent overfitting/underfitting. The relationship between the features and the car mileage is generally linear, and in general, this tells us that the simplicity of a linear regression model is sufficient enough to capture patterns in the data.

20% credit. Generate a plot of the regression coefficients as a function of the regularization parameter (see figure 6.12 in the ISL book).

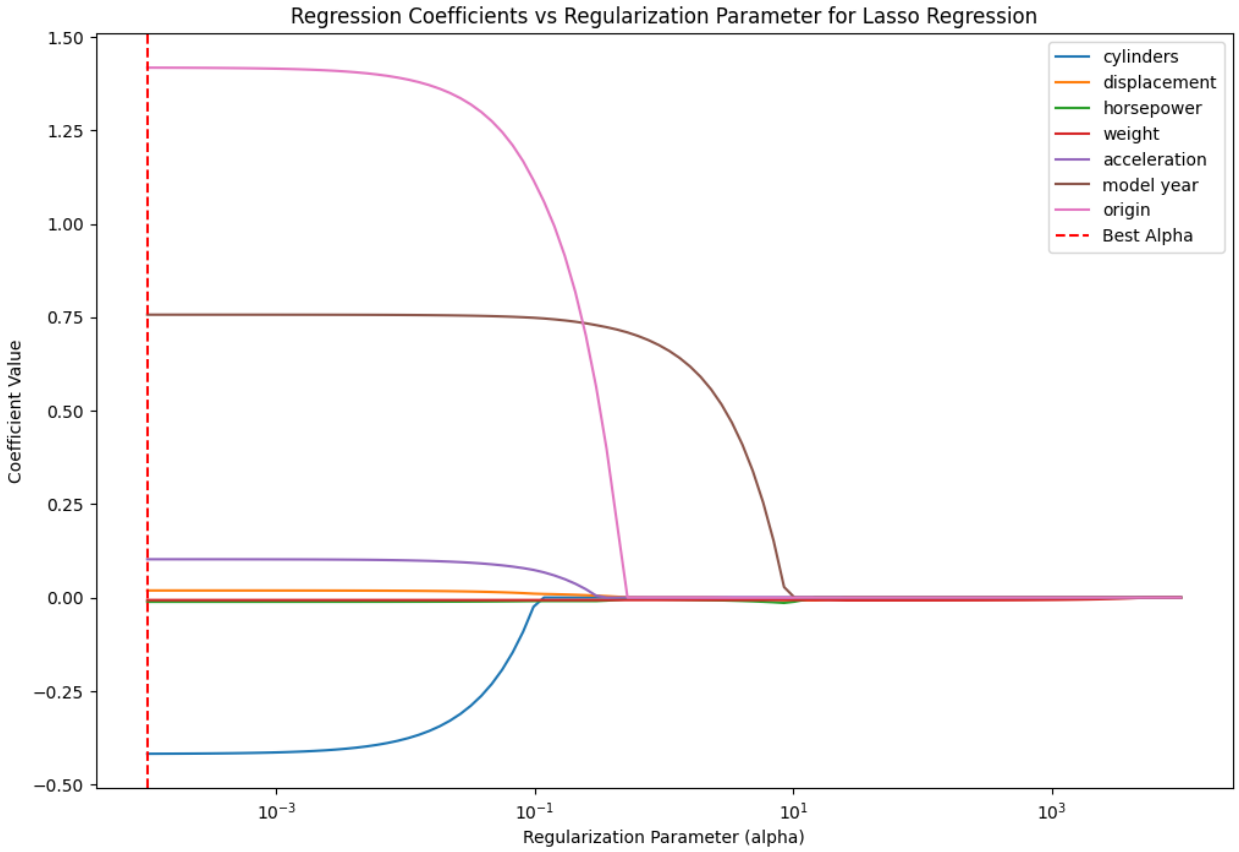

Figure 2. Regression Coefficients as a Function of the Regularization Parameter

Discussion of findings:

Based on the results of Figure 2 above, we see that as the regularization parameter increases, all coefficients approach zero. Initially, when the regularization parameter is negligibly small (i.e. so small that the behavior of the model is very similar to that of linear regression models), we see that some coefficients are close to, or already zero. This coincides with our understanding of Lasso Regression, as Lasso Regression works by adding a penalty term proportional to the sum of absolute values of coefficients. As the penalty term increases, or as the regularization parameter increases, coefficients are encouraged to approach or become zero. All coefficients approach zero when the penalty term is equal to 10, which tells us that all penalty terms greater than or equal to ten result in a model in which the regression coefficients are all zero.

Because the selected best penalty term is close to zero, this indicates that regularization is likely not needed for this dataset, as we do not need to encourage coefficients to approach zero.

Interpretation of Regression Coefficients for Selected Best Regularization Parameter:

The above figure tells us that models with small regularization parameters indicate that displacement, acceleration, horsepower, weight, and potentially model year are not good predictors of car mileage. The coefficients of the best regularization parameter tell us that good predictors of car mileage are origin, model year, and cylinders, as for each unit of origin, the car mileage increases by approximately 1.37 mpg.

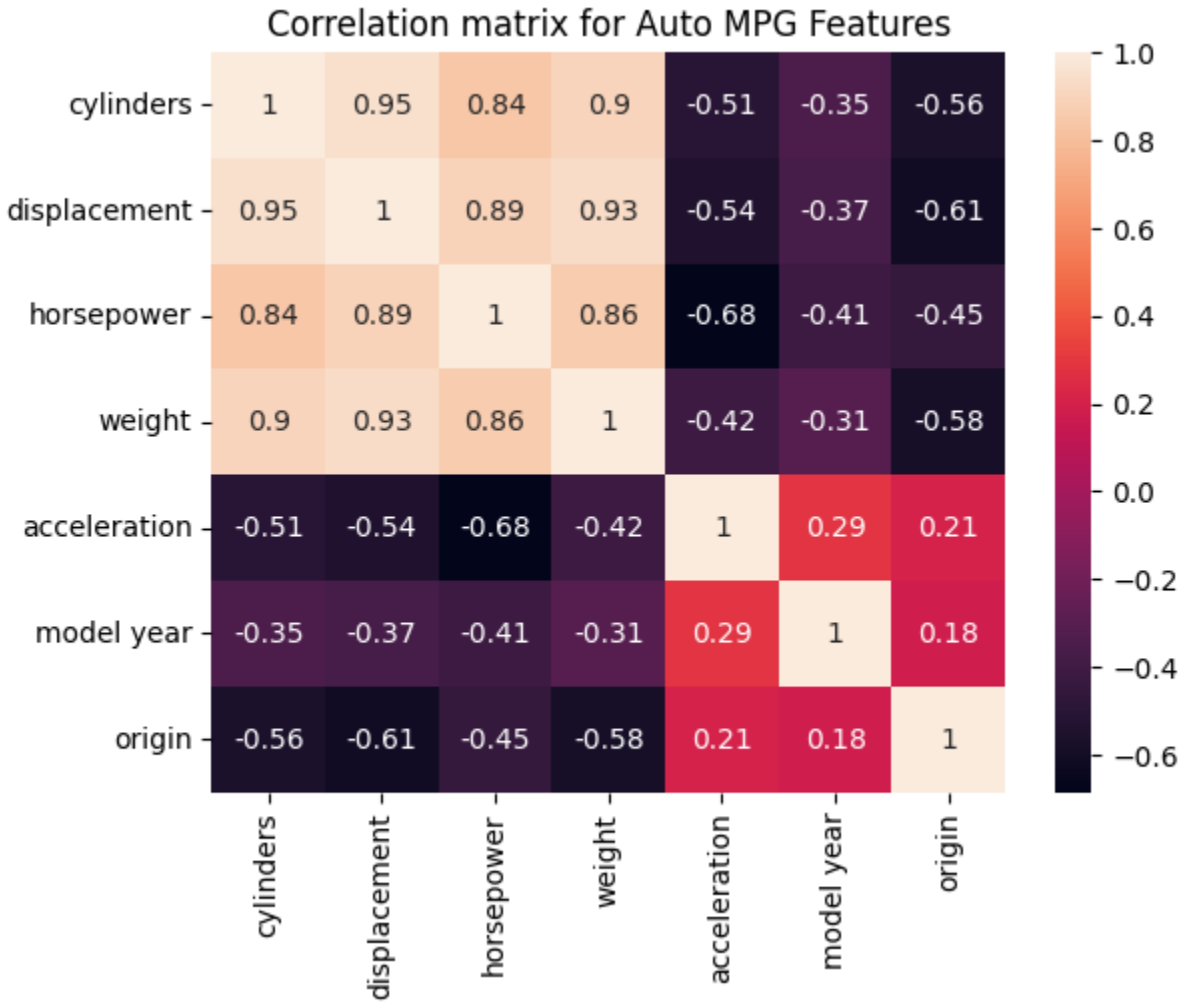

Figure 3. Correlation Matrix for Feature Dataset

Looking at Figure 3 helps to better understand the features in this dataset, and provide deeper insight into the weights assigned to each feature in Figure 2. In the above heatmap, we can see that there exists high correlation for certain features; this includes cylinders and displacement, displacement and weight, weight and cylinders, and so on.

It is interesting to note such features with high correlation coefficients, such as that of cylinders and weight. The coefficient for cylinders is notably much higher than that of weight, as for the ideal alpha value, the weight coefficient is close to, if not already 0. This somewhat makes

sense, as the more cylinders a car has, the heavier it is, as higher cylinder engines are much heavier. Thus, while the model above implies that origin, model year, acceleration, and cylinders are good predictors of car mileage, the presence of potential collinearity between features cannot be ignored, as we cannot say that features are significant based simply on their coefficient values. In order to make further interpretations of significant features, calculated p-values are needed for each coefficient to see which features are the best predictors of car mileage.