Final Project

All of the code and graphs are on the Jupyter Notebook while the text answers required are on this word document.

**b) Based on the findings/observations in step a, decide whether the dataset requires any preprocessing (normalization or standardization). State the reason for your choice of preprocessing technique or if you decide not to do anything at this step**

For this project, I have decided to normalize the data. This is mainly because the data is not bell shaped and the data in different columns are on different scales. Because the data is skewed, standardization wouldn’t work well. Furthermore, the data for the different columns are all on different scales. This will cause the model to assign incorrect weights on columns.

**f) Justify your answer based on your understanding of how cross validation works and in the context of bias-variance trade off.**

After doing cross-validation on the different models, it can be seen that the linear regression model performs the best. I used negative mean squared error to evaluate the models, and the linear regression performed the best out of all the models.

The way cross-validation works is by splitting X and y into 5 equal “folds”. The model is then trained on 4 of the folds and evaluated on the other left-out fold. This process is repeated 5 times, with each of the 5 folds only used once for validation. The cross\_val\_score function then returns a list of the scores from each of the 5 evaluations. This gives an accurate assessment of how the model performs on unseen data.

Because the regularization is used to prevent overfitting, it works best on larger datasets and not the ones we are working with on this project. This set doesn’t have an issue with overfitting and the plain linear regression model performs better.

High-variance models are overly complex and overfit on the data. Ridge and lasso regressions are techniques that reduce overfitting and reduce the complexity of the model. Because our dataset does not struggle with this, ridge and lasso does not perform well.

**g) Explain if you find any discrepancy between the results in the cross-validation step and steps c and d**

After looking at the results from the Linear Regression model, OLS, Ridge Regression, and Lasso Regression, I do not find any discrepancies. The linear regression model has a r^2 value of 0.87 on the test set which is higher than any of the other models tested. It is not significantly greater, but is, on average, 1-2% greater than the other models. The OLS model performed the same as the Linear Regression Model.