Ajay Iyer

DAT 402

Project 1

The dataset I chose for this project is Car Price Analysis dataset from Kaggle by Adeel Sherazi. After reading in the csv dataset through Python, I knew that I would want to find the model that can predict the price of the car so I thought Linear Regression would be perfect for this problem. The first thing I did was remove the first 3 columns of the data because I did not feel it was necessary to have those has some of my predictors for the price. After that I found how many NA values were in each column so I could remove those rows entirely to make the model as precise as possible. Next, I renamed some of the columns so manipulating them would have been easier in Python and R. Then, I used the label encoder to change some of the values that were categories to numbers so I could use them in the model. This was done for about 9 columns or predictors. Next, I double checked to make sure that all the types of data were int or float so that I would run into problems like these later. Then, I saved the updated dataset so I could do subset selection in R. After reading the dataset into R, I wanted to find the best model using the exhaustive selection method. I wanted to find the best model with 10 features/predictors for the price. The ones I got were aspiration, drive wheels, number of doors, location of the engine, width, height, number of cylinders, size of the engine, bore, stroke, and max rpm. After that, I did a summary of the model which found the coefficients and some statistics of the model. The multiple r-squared is 0.8843 while the adjusted r-squared is 0.878. After figuring out which 10 features I needed, I went back to Python and removed some columns so that only those 10 features show and the price variable. I performed the train test split with the test size being 20% so I could see if the data was a good fit for the model and I got an accuracy score of 0.891044 and I thought that was pretty good, but it can def be better. I printed the coefficients and intercept once again and I was getting something close to what the equation was when it was done in R. Then I used the x test data to predict what the output would be, and it was close to what the actual output. For the most part, the predictions were coming to be a little higher than what the data showed. Then, I took that updated dataset and read in R so I could figure what to do for bias-variance tradeoff. After going back to R, I used the KNN algorithm to plot the bias-variance tradeoff and that made 2 plots which were looking at k and Complexity of k vs outRMSE and inRMSE. I found that as the complexity increases, the graph was also increasing meaning there was high variance which is not what we desired for it to be in the model. After that I came back to Python to calculate the variance and bias, and the result was that the variance was around 62,000,000 which is ridiculously high, and the bias was 2,000,000. With such high variance, we can say that it does fit the data does fit but it might be overfitting to what we wanted.

<https://www.kaggle.com/datasets/adeelsherazi/car-price-analysis?resource=download>