Individual Project Report
Caroline Sklaver
DATS6103

1. Introduction

In this project, our goal was to determine the major factors contributing to London bike share use in 2015. To achieve this goal we began with a dataset from kaggle.com that we then cleaned, performed EDA, and machine learning algorithms. The cleaning did not require much, just renaming of columns, and creating dummy variables for one categorical feature, as the dataset did not have any missing values. The EDA included making graphs for visuals and calculating correlations amongst the independent variables. Next, machine-learning algorithms included linear regression and multiple linear regression analysis. The linear regressions were run between continuous predictor variables and total bike share count to determine relationships. Next, we used multiple linear regressions to create a model that predicts bike share use. These elements were made into a GUI application for ease of use and presentation demonstration.

2. Background

After choosing the London bike share data set from kaggle.com, it was clear that linear regression was the most appropriate model to run. With the dependent variable being a continuous count of bike share use, we ran simple and multiple linear regressions to determine the relationship of each column with bike use.

3. Personal Contributions

My portion of work in this project includes preprocessing, creating plots/figures, coding the model, creating the GUI demo, and explaining the mathematics behind linear regression algorithm in the final report. I put together and wrote all of the code and wrote the Introduction, Explanation of Data-Mining Algorithms, and Experimental Set-Up sections of the final report.

4. Results

Figure 1, the correlation matrix shows the high correlation between “temp” and “tempf”, which later results in our exclusion of “tempf” from the multiple linear regression in order to avoid multicollinearity.

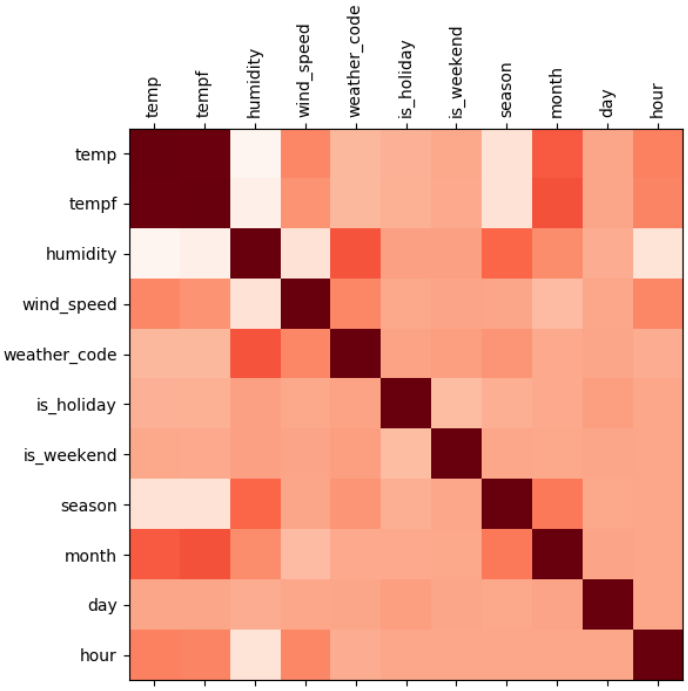

Figure 1: Correlation Heat map

Next, we used the GUI to look at the relationship between our continuous variables, “temp”, “humidity”, and “wind_speed” against bike count with a linear regression line (Figure 2, Figure 3, Figure 4).

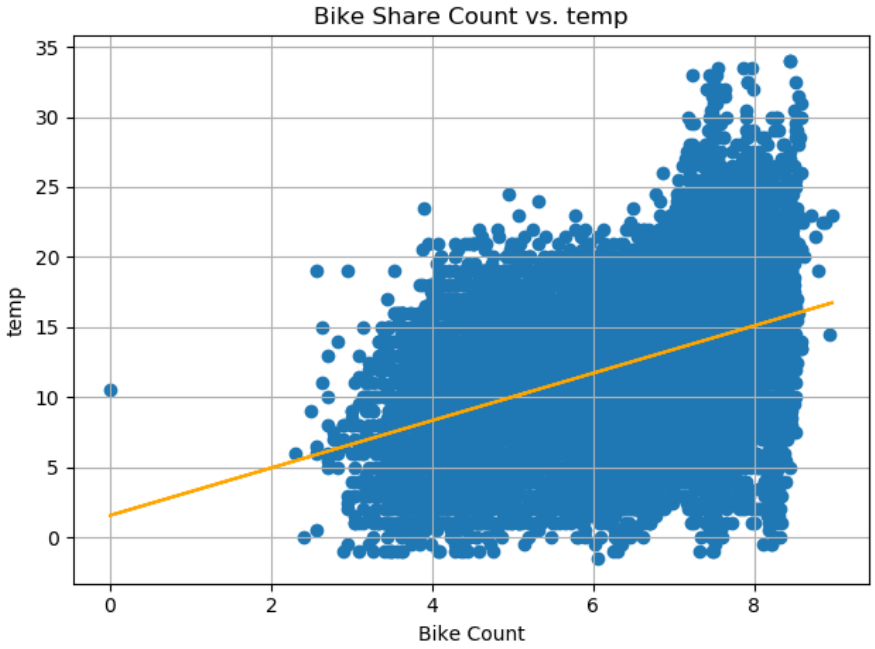

Figure 2

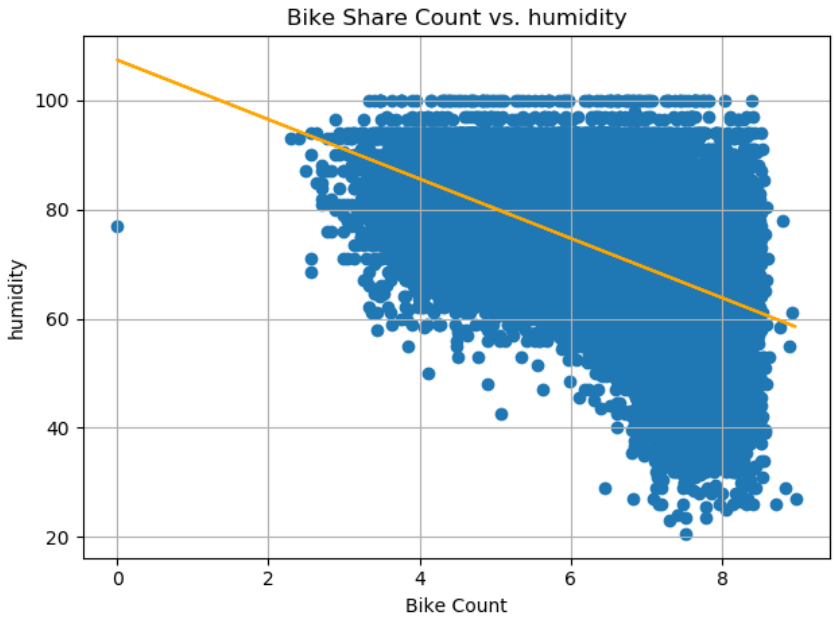

Figure 3

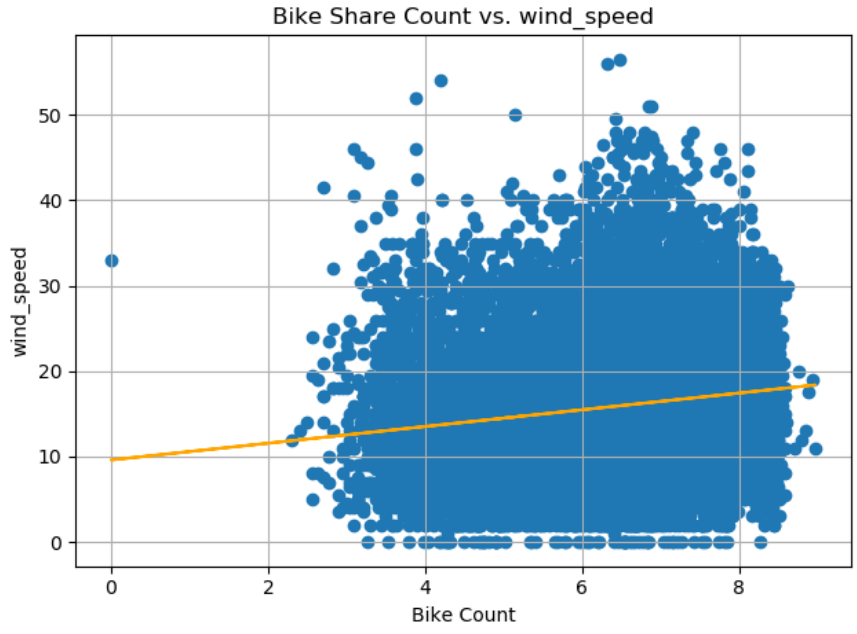

Figure 4

From these linear regression plots, we can conclude that temperature and wind speed have a positive correlation with bike count, while humidity has a negative correlation.

With our multiple linear regression model, our main results from the GUI are represented in Figure 5.

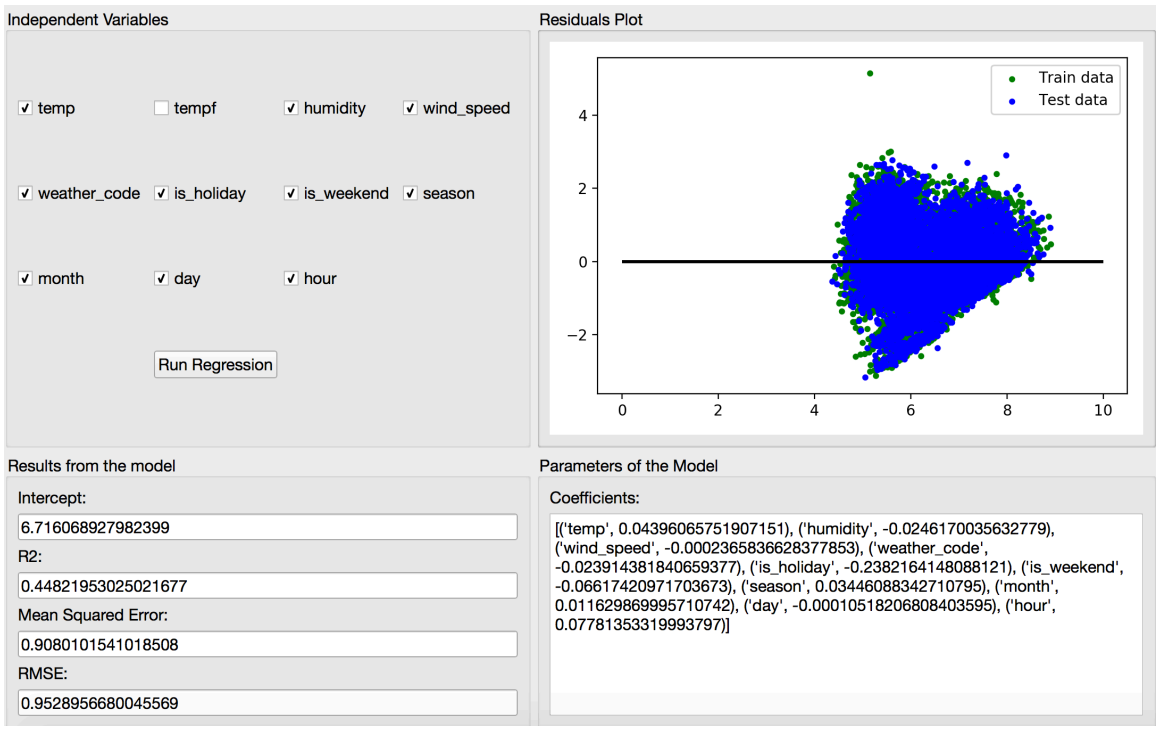

Figure 5

According to our R2 value, the features selected predict only 44.82% of the variation in bike count. This is not necessarily a bad model, however, there is much variation to still be explained. The coefficients printed show us the sign and magnitude of each predictor variables' relationship with bike count. Also, according to our MSE and RMSE, this is not a great model.

5. Summary and conclusions

In all, our final model is not great at predicting the number of bike-share uses in London. However, we were able to see the relationships between each variable and bike count, which is interesting information perhaps for bike share businesses to use.

Throughout this process I have learned how to create a GUI, which is an accomplishment and major take-away. I have also learned the importance of evaluating your model. Additionally, I learned which models are best based on the dataset at hand. In the future, I would like to implement other machine learning algorithms to other datasets, maybe one with a binary dependent variable. As far as this dataset goes, future research could be done on other features that may be predicting bike share usage in London.

6. Calculate the percentage of the code that you found or copied from the internet.

Including the demo found in DATS6103 git repository, I assume about 50% of the code in this project was copied from the internet.