Fall 2019 Individual Report

Sayra J Moore

The George Washington University

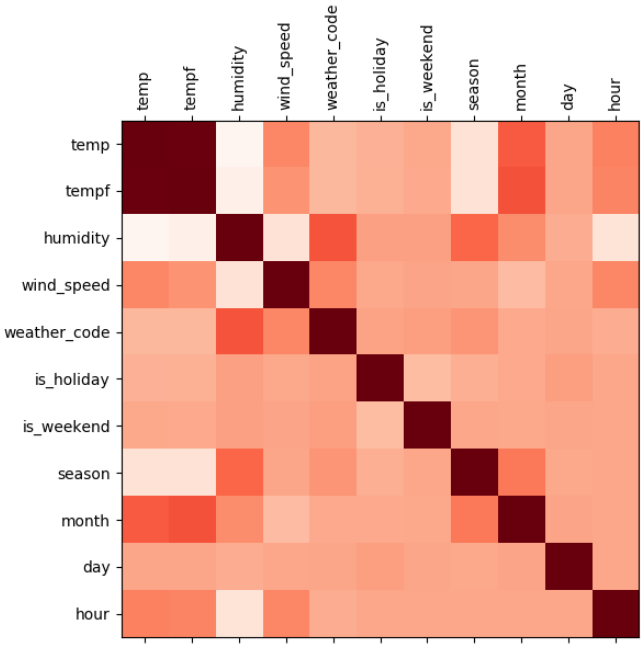
Fall 2019 Individual Report

Introduction

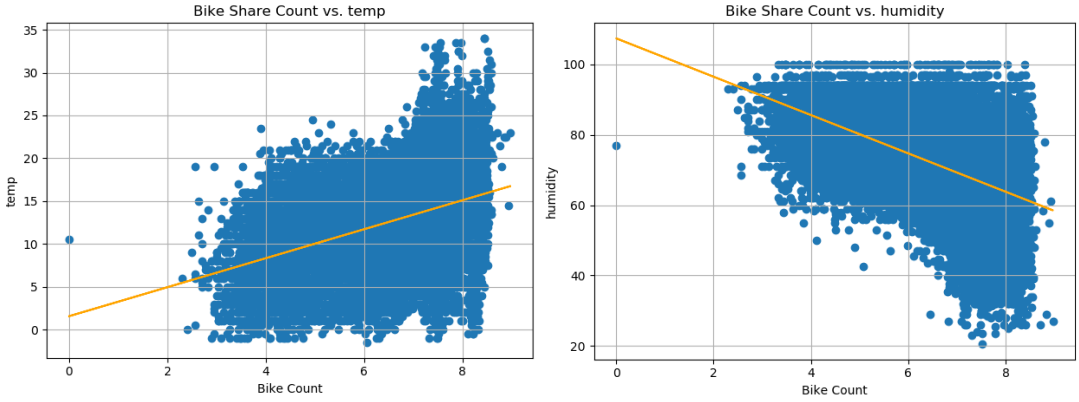
In this project, we are looking to build a model that will help us predict the number of bike shared at an instance with the given variables. The dataset was originally came from the UK government site 'Transport For London', and was merged with weather data from freemeteo.com/i-weather.com by timestamp. This merged dataset was then made available on kaggle.com. The 10 columns of this data set include: Timestamp, cnt (number of new bike shares), t1 (real temperature in Celsius), t2 ("feels like"), hum (%), wind_speed (km/h), weather_code (1:clear, 2:scattered clouds, 3:broken clouds, 4:cloudy, 5:rainy, 10:thunderstorms, 26:snowfall, and 94:freezing fog), is_holiday (1:holiday, 0:non-holiday), is_weekend (1:weekend, 0:weekday), and season (0:spring, 1:summer, 2:fall, and 3:winter)

Since This group is only two individuals, myself and Caroline, we decided to divvy up the work in half. Because we only ran one model, she took the coding part and I took the presentation and paper. So although we did not split the coding in half, the project itself was, and we worked every minute of it together side by side.

Results

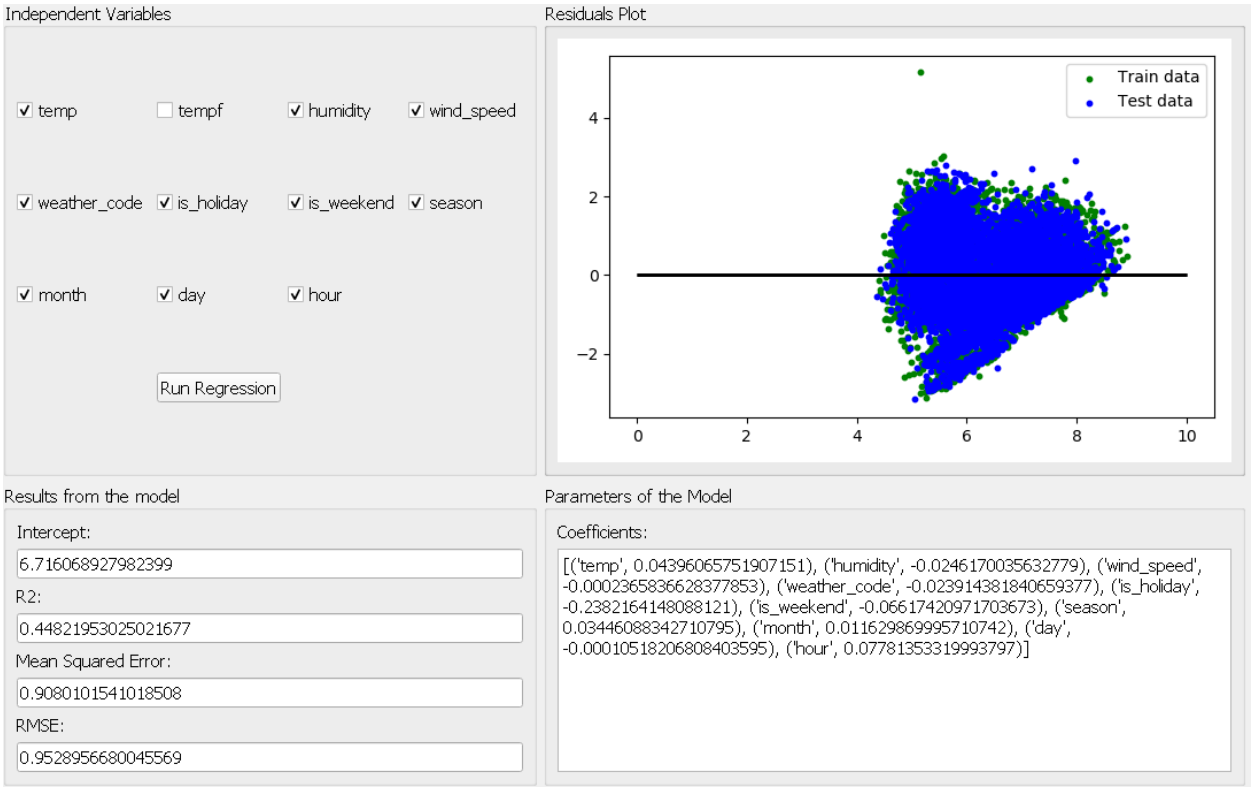


Before running the model, we did some EDA like a correlation plot to see how highly the variables were correlated to each other. Aside from variables being highly correlated with themselves, we saw that variables such as “weather_code” and “humidity” were highly correlated, and so was “month” to “temp” and “tempf”.



After doing that, we also ran plots to see the relationship between some of the variables and the target. First, we see that “temp” and Bike Count have a positive direct correlation, meaning as one increases, so does the other. But in the second figure, we see that “humidity” and

Bike Count have a negative indirect relationship. This means that as “humidity” increases the Bike Count decreases.



We used Multiple Linear Regression because our target value was not categorical nor were we trying to solve a classification problem, therefore it ruled out many of the model that we learned in class. The equation that comes from Multiple Linear Regression helps us apply the values of the variables, to then calculate the final number that is our target, in this case the number of bike shares purchased. The equation goes as follows: $\hat{Y}=b_0+b_1X_1+b_2X_2+...b_nX_n$. Then the values that we observe (ex. 85% humidity on a sunny day in June with a wind speed of 20) get plugged into the X in the equation to then find the $\hat{Y}$, or number of bike shares purchased.

Conclusion

After doing our EDA and building the model, we saw that we didn't have a great model and that the model only explained about 45% of the variation. If we were using the data for classification (if they did or did not purchase a bike share), then we would have potentially had a much better model and would have been able to fun many different models. Overall for this dataset, it was not great and would have much rather used the data for classification if our target would have allowed it. I believe that the variables we had for it would work well.

As for the amount of code that I personally wrote, it is too little to calculate. As mentioned earlier, it was just two of us so we split it between the code(gui) and doing the paper with interpretations and putting together the PowerPoint slide. We both worked together side by side every chance we had and did our absolute best to help each other out. I would absolutely do another Machine Learning study with my group mate Caroline!