ECE - 20875 - Mini Project Report Very nicee

Haley Huntington
huntingh@purdue.edu
0031549183
github:huntingh
003-Chris Brinton

Rahul Maheswaran
rahulm@purdue.edu
0029004787
github:rahulmaheswaran
001-Chris Brinton

04 December 2020

Research Questions

1. You want to install sensors on the bridges to estimate overall traffic across all the bridges. But you only have enough budget to install sensors on three of the four bridges. Which bridges should you install the sensors on to get the best prediction of overall traffic?
2. The city administration is cracking down on helmet laws, and wants to deploy police officers on days with high traffic to hand out citations. Can they use the next day's weather forecast to predict the number of bicyclists that day?
3. Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges?

Data Cleaning

Before we looked at answering the questions, we examined the data set to make sure we understand what we are working with. The first thing we did was to use pandas, to parse the data as a dataframe. Then we checked the data type of each column and noted they were almost all of type object.

The dataset was the following size: `[214 rows x 10 columns]`
We had to use regex and as.type() functions to remove symbols such as (S) and (T) from the precipitation column, and ' , ' from all numbers before converting it to type floats and int respectively. Then the date column was converted to a **time series**, so we could use more applicable and known techniques to explore the data.

Column Name	Before	Cleaned
Date	object	datetime64[ns]
Day	object	object
High Temp (°F)	float6	float64
Low Temp (°F)	float64	float64
Precipitation	object	float64
Brooklyn Bridge	object	int64
Manhattan Bridge	object	int64
Williamsburg Bridge	object	int64
Queensboro Bridge	object	int64
Total	object	int64
dtype:	object	object

Question 1

To address question one, we need to be specific about what we want to achieve. To get the best prediction of **overall** traffic, we need to know more about the data. What are the average number of cyclists over each bridge, what sort of trends and variations do we see in the overall population over the recorded time?

To do this, we used seaborn and matplotlib to see how the dataset is distributed. The first step was just to see how many cyclists use the bridge over the recorded time period.

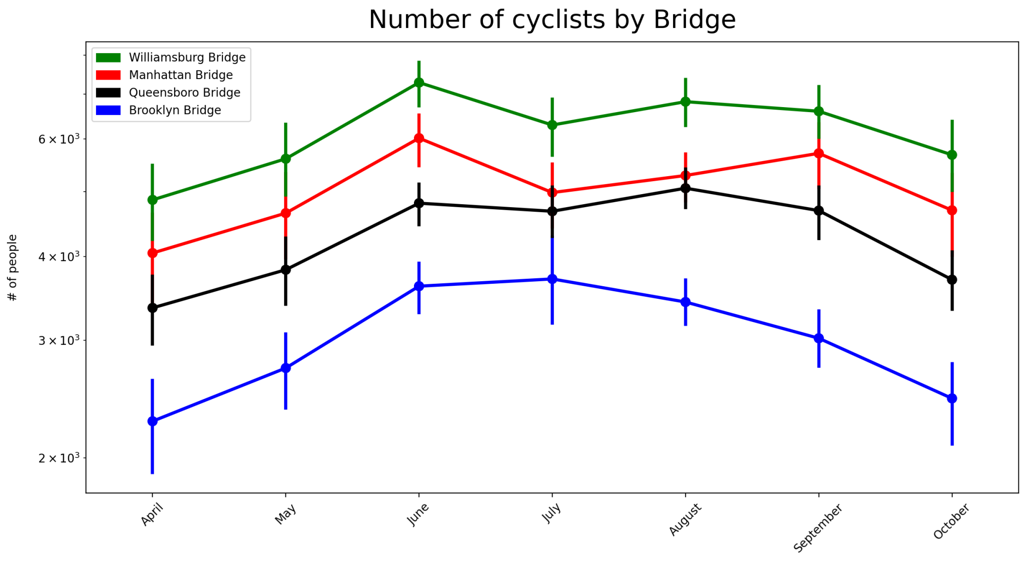

Figure 1: Plot showing number of cyclists using each bridge

This plot immediately reveals a few things. Williamsburg Bridge seems to have the highest cyclist traffic throughout the observed time period and Brooklyn Bridge has the lowest. The overall traffic for each bridge increases during spring and summer and drops as winter approaches.

To choose the 3 bridges that will benefit from a sensor, we need to see which bridge has the most varying trend in usage. Even though we can install sensors in the 3 most highly used bridges, it does not predict the overall traffic pattern. *In fact, looking at this plot alone, we can observe that Williamsburg Bridge and Queensboro bridge have very similar trends* throughout the observed time period, where as even though Brooklyn Bridge sees lower traffic, it has a different pattern compared to the other three. This variation would be much more useful to account for when predicting overall traffic, so we need to observe how the bridges' traffic is distributed against each other.

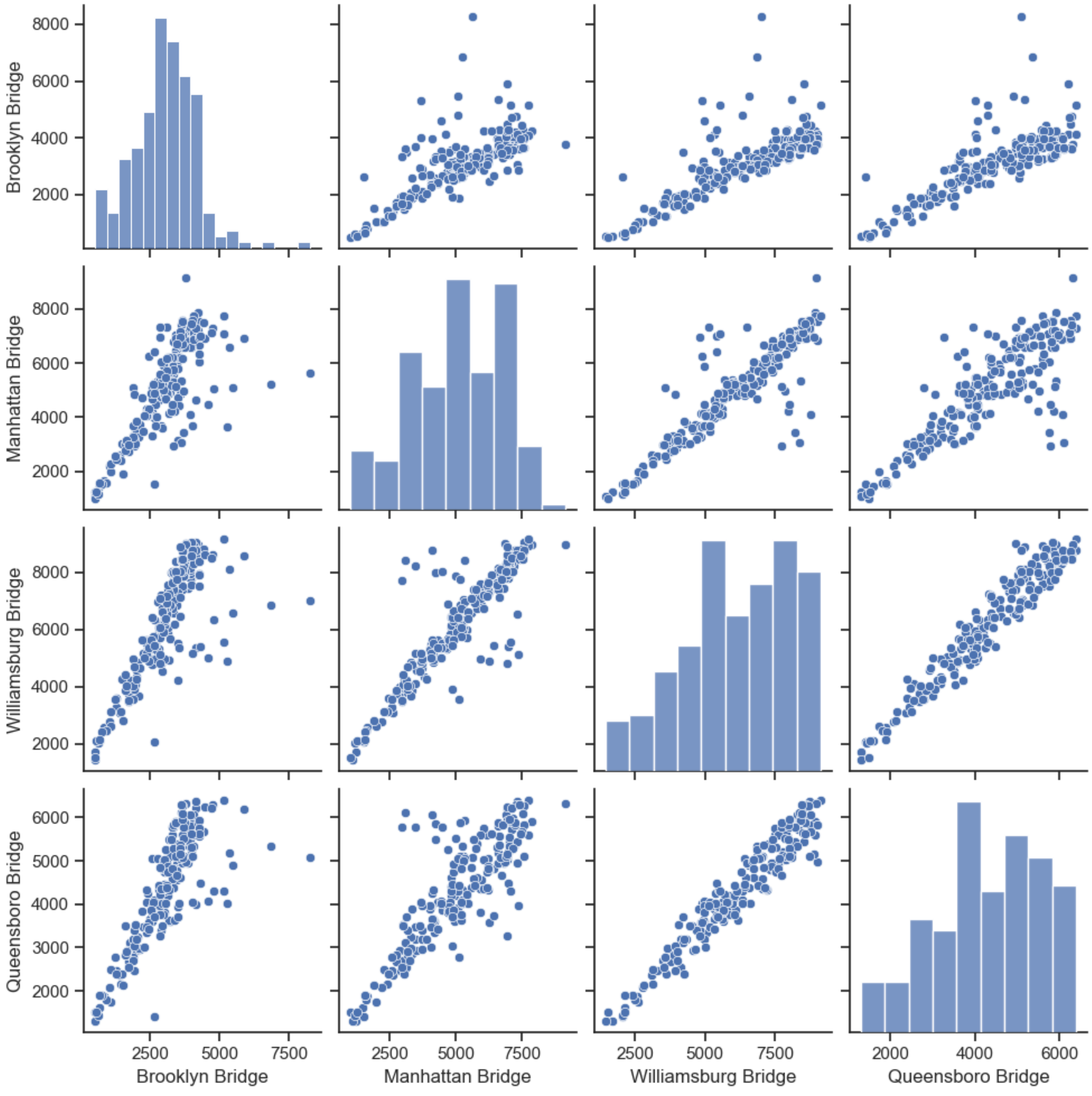

Figure 2: Distribution of data: Number of cyclists at each bridge

Looking at the histogram, it is clear that Brooklyn Bridge has a normal distribution, compared to a more one tailed distribution for Queensboro and Williamsburg. The scatter plot also shows a strong positive correlation between Williamsburg Brige and Queensboro Bridge. So it is clear that both these bridges have a very similar trend. To confirm this statistically, we used the Peason's Standard Correlation test for each bridge against each other (Eq. 0.1). We picked this statistical test because both variables are quantifiable, and we wanted to see if there is any linear relationship between them.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad [0.1]$$

Bridge [x]	**Compared Bridge [y]**	$R^2 score$
Brooklyn Bridge	Manhattan Bridge	0.7517130
	Williamsburg Bridge	0.7926036
	Queensboro Bridge	0.8132065
Manhattan Bridge	Brooklyn Bridge	0.7517130
	Williamsburg Bridge	0.8783772
	Queensboro Bridge	0.8389666
Williamsburg Bridge	Brooklyn Bridge	0.7926036
	Manhattan Bridge	0.8783772
	Queensboro Bridge	**0.9653991**
Queensboro Bridge	Brooklyn Bridge	0.8132065
	Manhattan Bridge	0.8389666
	Williamsburg Bridge	**0.9653991**

As expected Williamsburg Bridge and Queensboro Bridge traffic patterns are very similar with an R score of 0.965.

To validate this, we used the sklearn, linear regression model to create a model that **can predict the number of cyclists on Queensboro bridge, given the number of cyclists observed at Williamsburg Bridge**, **assuming** that all other factors will remain constant.

We divided the training data and test data at a 75:25 split, and found the best model to be described by the equation

$$\hat{y}(\boldsymbol{x}) = 0.63739706x + 369.7593588 \quad [0.2]$$

The model was tested with test data set, and we saw a model score of **0.945** which implies that 94.5 % of the values predicted were true. However, since the question does not ask about a model, we did not proceed to check the MSE. But with the above analysis, **we can conclude that the bridge that will not need a sensor is Queensboro Bridge, as it has a very high similarity in traffic with Williamsburg bridge. Therefore, having a sensor on the other three bridges, and predicting the number of cyclists on Queensboro bridge is the most suitable solution for this problem.**

Question 2

Question two attempts to relate the weather forecast and the number of bicyclists. To be able to answer this question, we had to select what features will affect the outcome of a model. For any given day, we can assume that 'weather' comprises of the Low and High temperature for that day, along with the Precipitation. We selected these to be our Features for this problem. To be able to decide what type of statistical test we wanted to use, we had to see the distribution and relationship between the features.

We decided to use a Ridge regression model for the following reasons. Looking at the joint plot between the otal number of cyclists and precipitation, there is a strong trend in the total

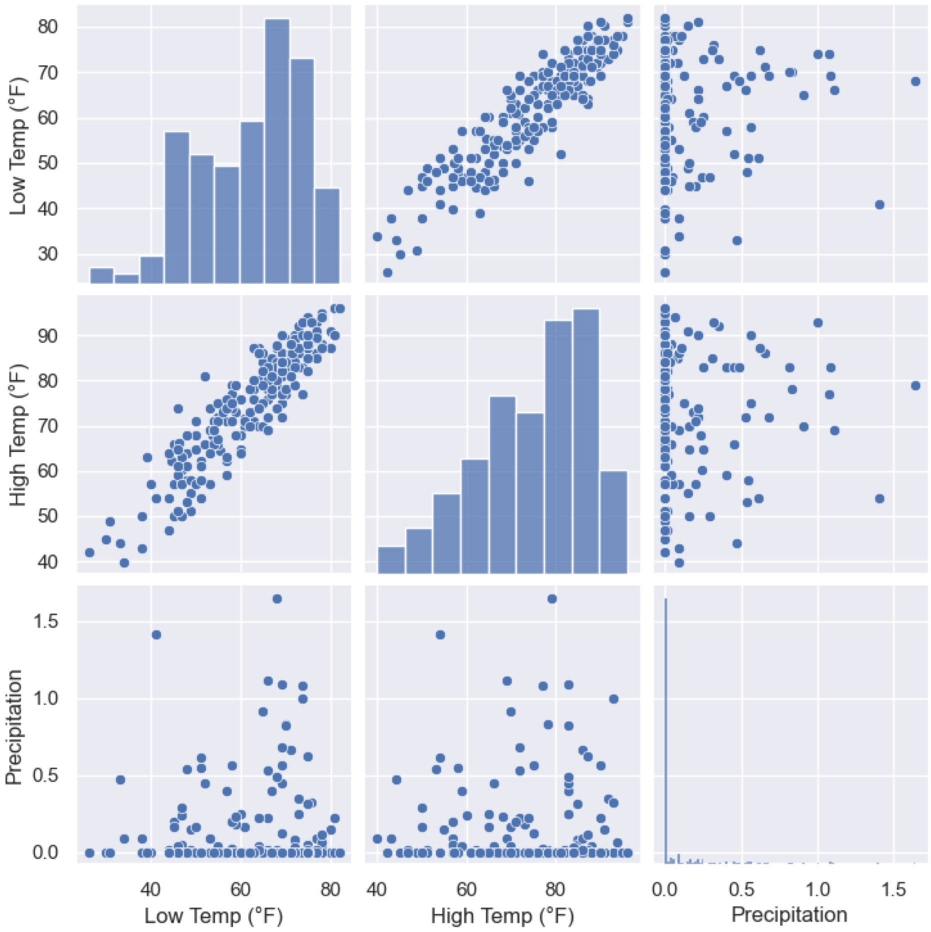

Figure 3: Distribution of features

number reducing as the precipitation increases. Given this strong dependency on one feature, using an unweighted linear regression model will most likely introduce a lot of bias in the model prediction process. Therefore, to prevent this, we used a Ridge Regression model, where we introduce regularization parameter lambda to account for this bias and preventing over fitting the model. First we take 75% of the data to train and 25% of the data to test, and run it through a loop that tries to find the best regularization parameter λ by evaluating the MSE error. Before we modeled the factors, normalized the values because they have different units. To normalize factors, we subtracted the mean from the values and then divided by the standard deviation of the factor. We found the best λ to be 2.75, which yields a MSE of 16759195.09 which is shown in figure 4 below.

Then we can use that λ to find the model that best represents the data. The equation is:

$$\hat{y}(\boldsymbol{x}) = -2126.5799498x_1 - 1403.72211222x_2 + 4306.60070065x_3 + 18469.2375 \qquad [0.3]$$

with an r^2 score of 0.5217946395017555. The values of the model have been normalized so the coefficients printed above can show us the importance of the different factors of the weather. First we will notice the largest coefficient is the third one, which corresponds to the High Temp (°F) column in our model. So we can gather from the model that the high temperature of the day is the most influential factor when bikers choose to ride. We can also see that because the value is positive, bikers generally chose to ride more when it is warmer. The second largest

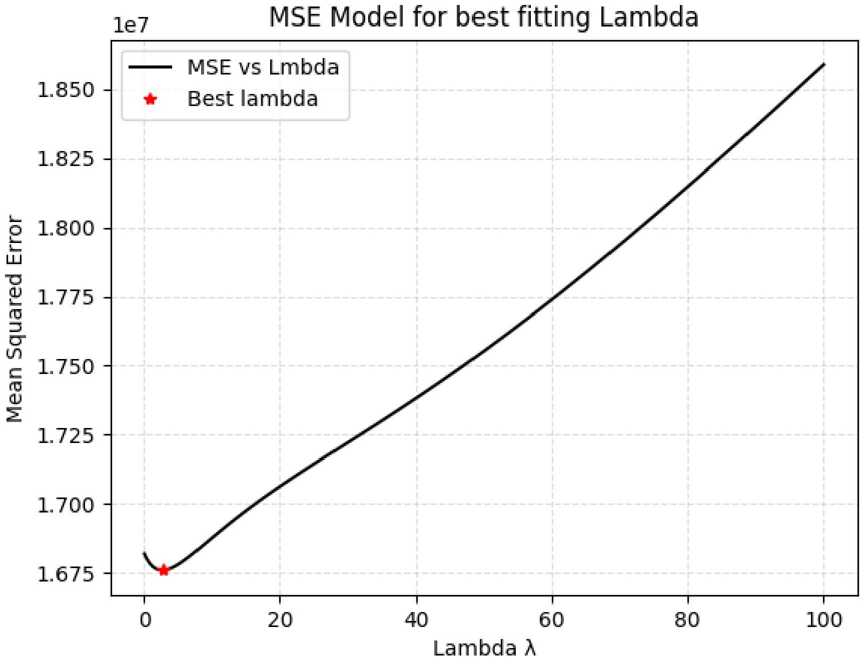

Figure 4: Finding the best Lambda Value

value is the first column which corresponds to precipitation. The negative sign in front of the precipitation coefficient tells us people choose not to ride when it rains. The other coefficient is the low temperature. The lowest temperatures tend to happen either very early or very late, or in other words, when it's dark. So it makes sense to see that the low temperature is not a very influential factor in our model. Through all of our calculations and analysis, we can conclude that you can predict the number of bicyclists by the conditions of the weather with our model with a 52 % accuracy. Increasing the number of features and getting more observations can help improve this model further.

Question 3

In question three, we try to predict whether or not it is raining based on the number of bicyclists on the bridges. First thing we did was recognize the nature of units of "y". To categorise rain, we decided to consider any non zero precipitation as rain. Y in this case, is either true or false, 1 or 0, so we can apply a a logistic regression model.

We took the column of precipitation and made the values either 1 or 0 based on the precipitation. Next we plotted the logistic regression model shown in figure 5.

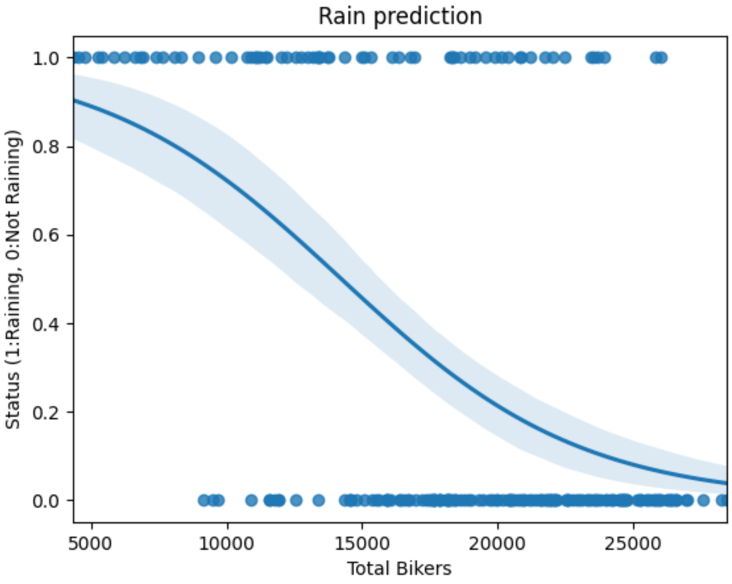

Figure 5: Logistic Regression Model.

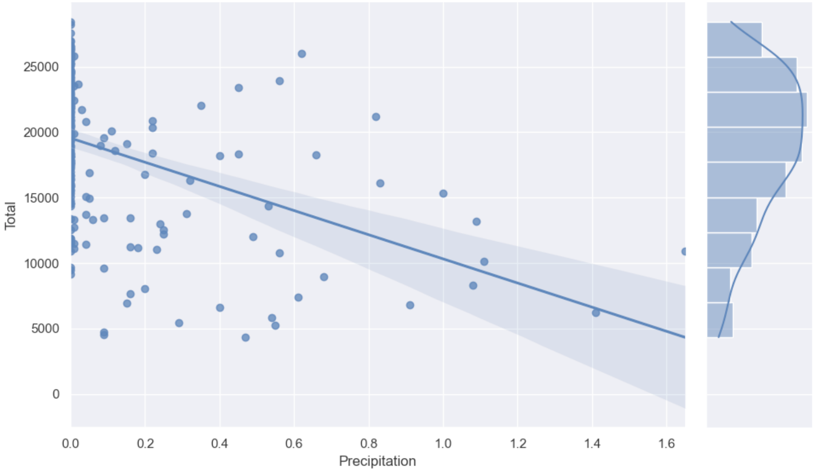

Figure 6: Joint plot of Precipitation and Total number of cyclists.

The model above supports our assumption that people would ride less when it's raining. On the left side of the line it's raining and there are less bikers, and on the right it is not raining and there are more bikers. In this case, we consider any value greater than or equal to 50% to predict rain. In other words, if there are 14,240 bikers or less, our model predicts it is raining. Once again, we recognize that this would be not be an accurate model, and therefore, in order to improve the accuracy of the model we recommend adding more features to it.

Improving the model

We could use a time series seasonal approach - including both the time/season of the year and along with the precipitation, high and low temperature. In addition, if we could get a larger data set, we would be able to increase the number of features and overall model score.

Since weather related predictions are hard to make based on one feature, we cannot accurately predict weather based on number of cyclists. However, with a larger dataset and more features, we could use multiple sklearn models and compare the overall model scores. Further more, the sampling could be done using the KFold method as we the data can be classified in groups. This sort of cross validation improves the reliability and accuracy of the model.

Code and References

All the python code for visualisations and calculations are available in our github repository: `https://github.com/ECEDataScience/project-fa20-very-nicee`

Equation links were taken from Wikipedia