Data Analysis for Bike Traffic in New York City

Zachary Au (zau)

Yash Agnihotri (yagnihot)

Path 2

https://github.com/zacharyau/Bikes-And-Bridges.git

The Dataset

The dataset for Path 2 is a csv file that describes bike traffic on four bridges in New York City: the Manhattan Bridge, the Queensboro Bridge, the Brooklyn Bridge, and the Williamsburg Bridge. It lists the number of bikers on each bridge per day as well as the total number of bikers on all of the bridges per day. The dates range from April 1st to October 31st. The day of the week is also listed. Finally, different weather conditions for each day are listed (high temperature, low temperature, and precipitation). There are 214 datapoints. We used an 80/20 split for testing and training data; meaning, there are 171 training datapoints and 43 testing datapoints.

Problem 1: Data Analysis

"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?"

We want to install sensors on the bridges with the most bikers, since we will be able to collect the most datapoints and get the most out of our investment. So, we decided to total the number of bikers across all 214 days per bridge and compare the results. The bridge with the least number of total bikers will not have sensors installed, since it will not provide as much data as the rest of the bridges.

We extracted the columns containing the number of bikers per day for a certain bridge using the Pandas library. Then, we converted the number of bikers into NumPy arrays. After that, we summed all of the elements in each array. Next, we created a dictionary with the name of bridges as keys and the total number of bikers as elements. We then used Python's sorting function to sort the dictionary. Finally, we printed the names of the three bridges with the three highest total number of bikers. These bridges will be the bridges selected for sensor installation. This method has the same effect as finding the average number of bikers on each bridge per day, since to find the average we would just divide each total by 214 (the number of datapoints).

Problem 1: Results

Our Python script printed the following in the terminal:

The sensors should be installed on:

- the Williamsburg Bridge (1318427 bikers)

- the Manhattan Bridge (1081178 bikers)

- the Queensboro Bridge (920355 bikers)

As shown above, our algorithm recommended that sensors be installed on the Williamsburg Bridge, the Manhattan Bridge, and the Queensboro Bridge. It also listed the total number of bikers on each bridge over the 214 day period. To visualize our results, we used Matplotlib to create a bar graph of our data.

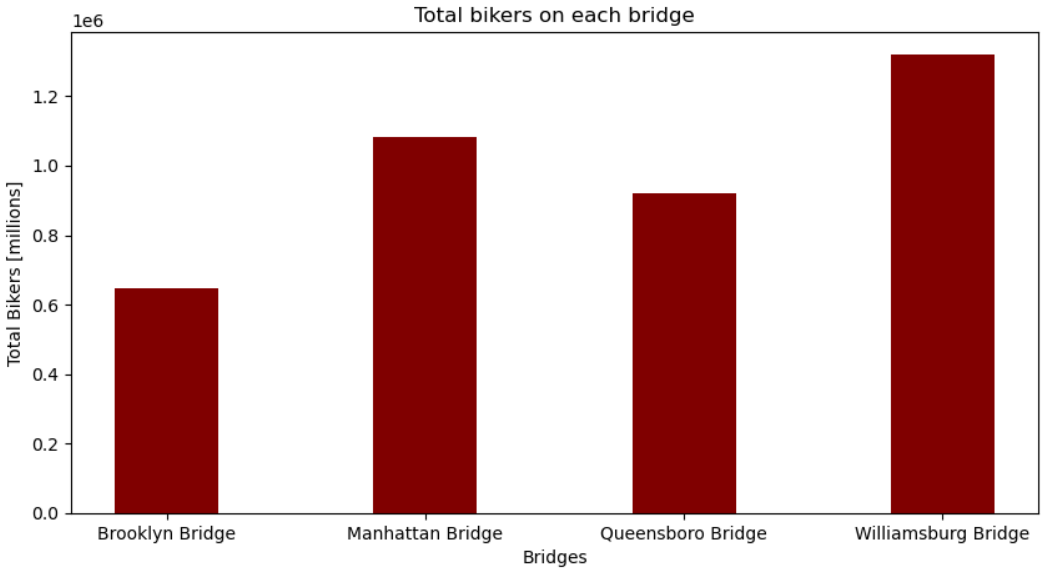


Figure 1: Bar graph showing the total bikers on each bridge over the 214 day time period

Figure 1 shows the bar graph generated by Matplotlib. It clearly shows that the Brooklyn Bridge received the least amount of bikers over the 214 day time period. The sensors should be installed on the Manhattan Bridge, Queensboro Bridge, and Williamsburg Bridge, since they received the biggest number of bikers and provide the most data about bike traffic.

Problem 2: Data Analysis

“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(low/high temperature and precipitation) to predict the total number of bicyclists that day?”

The city wants to know if we can predict the total number of bicyclists in a given day based off of the weather conditions. In other words, is the number of bicyclists on a given day correlated with that day's forecast? To study this relationship, we chose to use linear regression.

To perform linear regression in Python, we used the sklearn library. In Problem 2, we need the low temperature, high temperature, precipitation, and total bikers columns from the csv file. We used the same method described in Problem 1 to convert the csv columns into NumPy arrays. To create our feature matrix, we used NumPy's column_stack function to combine the high temperature, low temperature, and precipitation into one NumPy array, with each column representing x_1, x_2, and x_3. Next, we trained our linear regression model using the training data feature matrix. Then, we predicted datapoints using the model and compared the predicted datapoints to the test data. To have a quantitative answer to our question, we found the r^2 value.

$$E(a,b) = \frac{1}{N}\sum_{n=1}^{N}(y_n - (ax_n + b))^2$$

Equation 1: Mean Squared Error (MSE)

$$r^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \overline{y})^2} = 1 - \frac{MSE}{\sigma_Y^2}$$

Equation 1: Coefficient of Determination (r^2 value)

Problem 2: Results

Our Python script printed the following message in the terminal:

The r^2 score is: 0.595

This number has been rounded; the true r^2 value is 0.5954303929804975. This is not a good r^2 value. The r^2 value is the fraction of variance of data that is explained by the model. This means that our linear regression model does not fit the data well. Based off of only the r^2 value, it seems like we should choose another regression model or classifier and try again. This is why we created a visual to represent the situation.

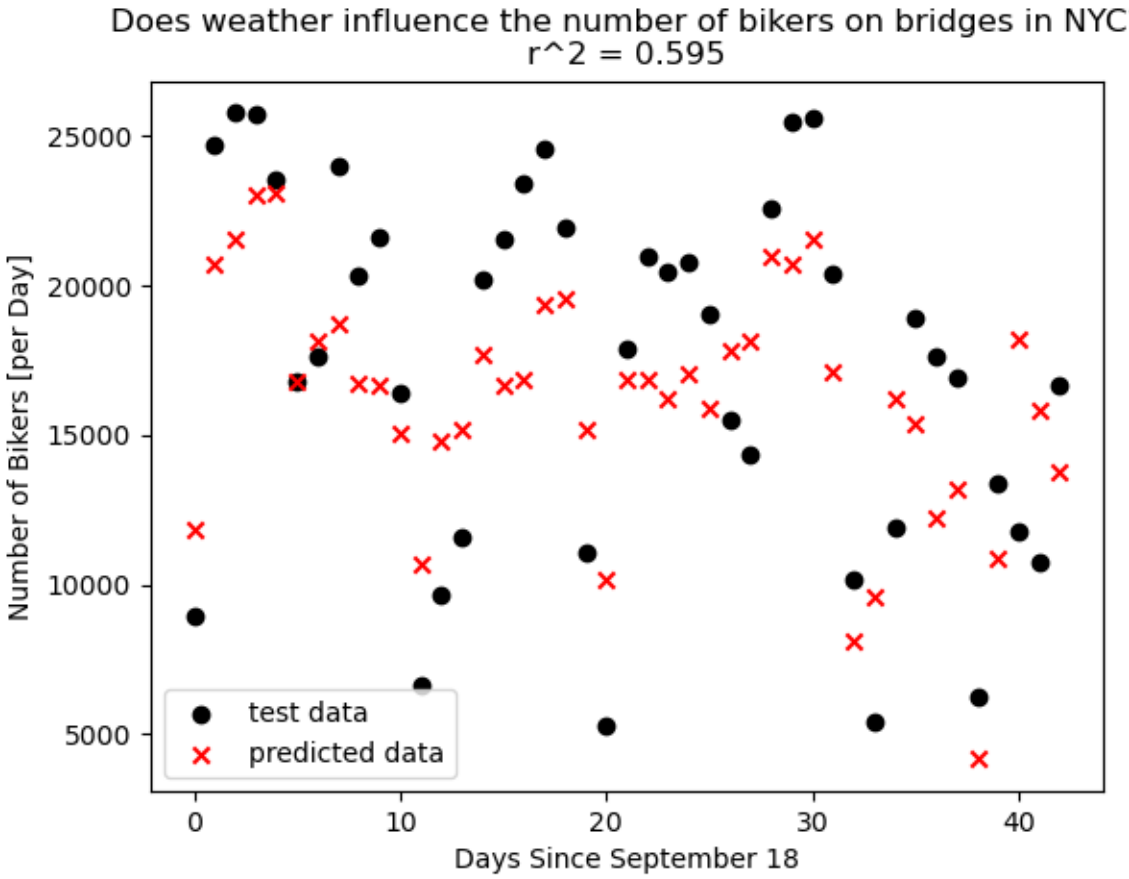


Figure 2: Scatter plot showing the predicted and actual data for the number of bikers on NYC bridges

Figure 2 shows the test data and the data predicted by our model. As seen above, the distribution of bikers appears to be random. Therefore, it is impossible to predict the datapoints. Given the graph in Figure 2 and the r^2 value (0.595), we conclude that the police cannot use information about the weather to predict the number of bikers on the four bridges in New York City on a given day. Our linear regression results were inconclusive, indicating that there is no trend between weather and the number of bikers on NYC bridges.

Part 3: Data Analysis

"Can you use this data to predict what day (Monday to Sunday) is today based on the number of bicyclists on the bridges?"

We needed to reform the Monday through Sunday data so that a model can use it. Therefore, we represented the days of the week with numbers. Saturday was represented with six, Sunday was represented with zero, and so on. Next, we needed to use a "for loop" to add up all of the riders for each day of the week.

After transforming the data, we proceeded with choosing our neural network. With this data, we selected a neural network because, as the graph below illustrates, we specifically chose an MLP for this procedure due to its ability to capture complex, non-linear patterns in data, making them suitable for tasks where simpler models may fall short.

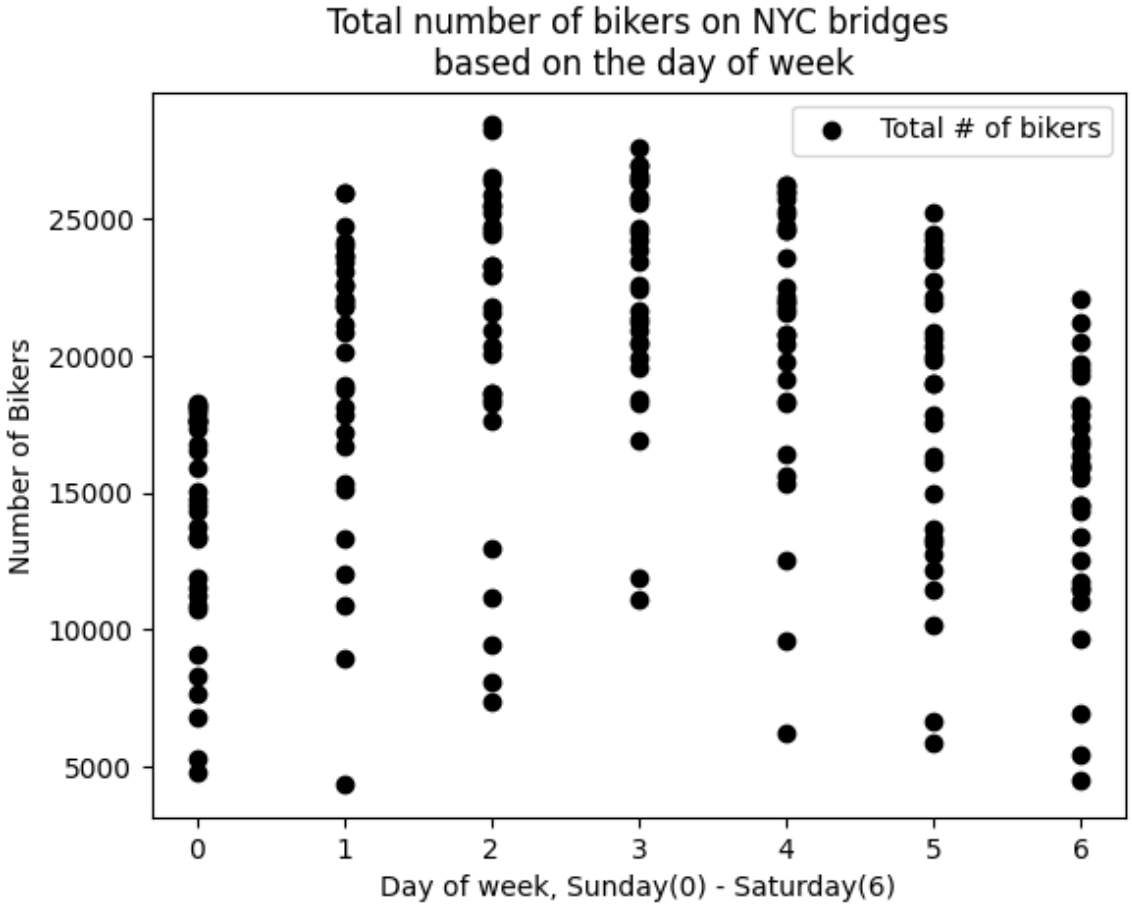
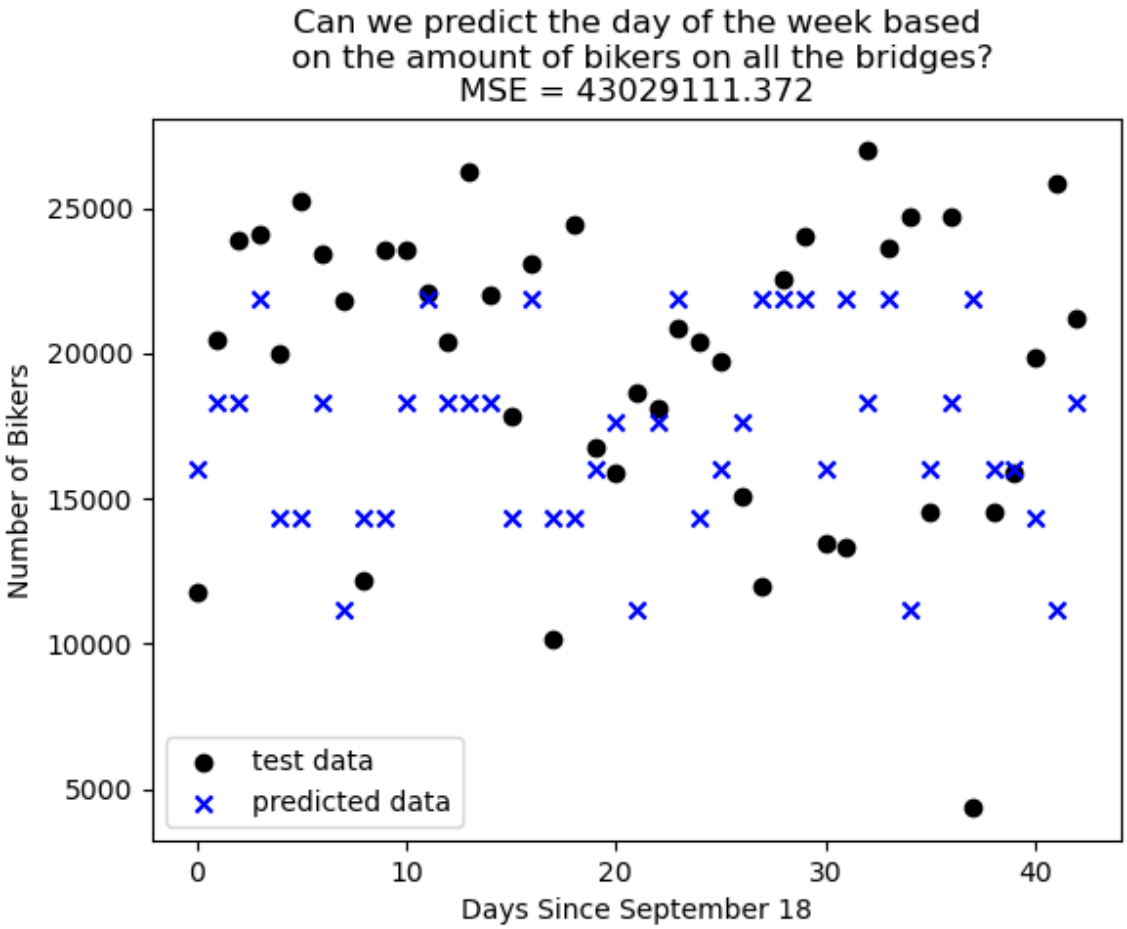


Figure 3: Dot plot showing the number of bikers on each day of the week

We can make use of the sklearn libaray's ability to not only train the model but also to separate training and test data. This allows us to partition the data so that the test data is not seen by the model prior to execution. We chose an 80-20 split (80% training data, 20% testing data) to make sure the MLP has enough data to work with. After that, we used the training dataset to train the model, and the outcomes are shown below.



After training the model on the training data, we used the sklearn mean_squared_error function to find the MSE of the predicted and actual data. The graph above illustrates how the MSE we discovered would vary from 2 million to 9 million. As a result, we can say that the model is unable to accurately predict the testing data.

The graph also makes it evident that the data is non-linear. To elaborate, the model's accuracy is dependent on variables such as dataset size, data quality, and parameter selection. There are likely more variables, such as bad weather, illness, events in New York City, holidays, etc., that have an impact on the number of bikers on NYC bridges. In conclusion, we cannot predict what day of the week it is based off of the number of bikers on the four NYC bridges.