Path: 1 – Bicycle Traffic

Group:

1) **Miguel Castilho Oliveira**
    a. **Purdue Username: mcastilh**
    b. **GitHub Username: MigCast9**
2) **Pedro Paulo Pinto**
    a. **Purdue Username: pintop**
    b. **GitHub Username: pepabrdp**

Dataset:

The dataset chosen gives information of bike traffic across four different bridges in New York City: Manhattan Bridge, Brooklyn Bridge, Queensboro Bridge, and Williamsburg Bridge. Additionally, it also provides information regarding the dates of when each data sample was collected, as well as weather characteristics, such as precipitation and temperature.

Analyses Methods:

1 – Sensor Data

The first analyses question inquired about the bridges which sensors should be installed on to get a better prediction of overall traffic. Considering there were enough sensors for only three of the four bridges, it was necessary to devise a method to disregard the bridge that least represented the data. To that end, the group decided that the bridge that presented the average daily bike traffic furthest from the average of all average daily bike traffic combined to be the one that would be eliminated.

For this, the group should calculate the average daily bike traffic for each bridge by summing their daily bike traffic and dividing it by the number of days. After calculating that number for each of the four bridges, it is necessary to calculate the total average daily traffic by summing up the numbers just found and dividing it by four. In sequence, all that is left is to check which average daily bike traffic is furthest from the total average daily bike traffic.

2 - Weather forecast to predict number of bicyclists

In this problem, it is necessary to consider the possibility of predicting the number of cyclists on the following day based on the weather forecast. To achieve that goal, what the group decided to do was a linear Ridge regression in which it will be necessary to use the values of Low Temperature, High Temperature, and Precipitation as the X parameters in order to try to predict the corresponding values for total traffic. In sequence, it is possible to compare the predicted total bike traffic to the actual values and calculate the Mean Squared Error to have some diagnostic of the model as well as the coefficient of correlation to see how effective the model is in predicting the correct values.

The team does not believe the weather forecast will be a good indicator of the number of bikers.

3 – Raining Prediction

The last problem has the objective to determine if it is possible to predict whether it is raining based on the number of bicyclists on bridges. In order to find out the solution to this task, the group decided to use a classification model, specifically the Naïve Bayes Model, plotted with a gaussian distribution. First, we separated the total number of cyclists in a day list in two different lists, one for days when there was precipitation and other for days without it. For each of these two we separated them between train data, 75% of total and test data, 25% remaining, resulting in a ratio of 3:1. The train data's served to create the Gaussian distribution, which we could compare the testing data to discover how accurate the classification model is.

Results:

1 – Sensor Data

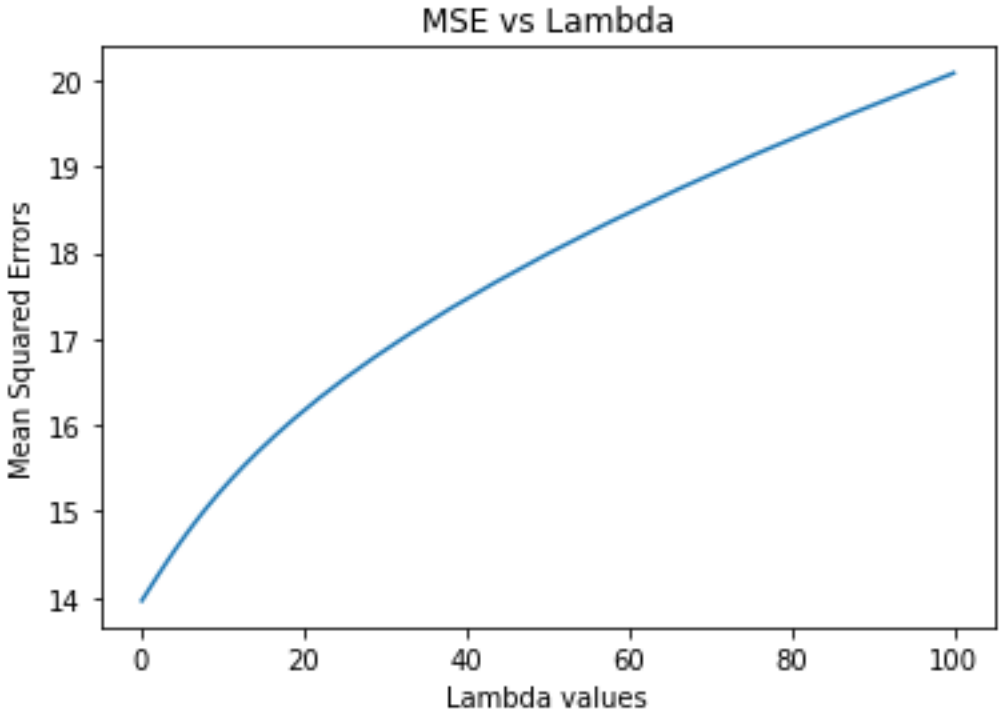
Through the implementation of the described method, the average daily bike traffic for the Manhattan Bridge was of 5052 bicycles, for the Brooklyn Bridge was of 3030 bicycles, for Queensboro Bridge was of 4283, and for the Williamsburg Bridge the average daily bike traffic was of 6160. After obtaining those values, the total average daily bike traffic obtained was of 4636 bikes, and the bridge that deviated the most from such value was the Brooklyn Bridge. Therefore, the bridges chosen to have the sensors installed are the Queensboro, Manhattan and Williamsburg Bridges to get the best sense of overall traffic.

2 - Weather forecast to predict number of bicyclists

After running through the described process of separating the data into training and test data in order to create a model and test it with the test data, it was possible to arrive at the follow model:

Total Traffic = -1942.78367 * x_1 -1543.50895 * x_2 + 4347.27326 * x_3 + 18412.6

As seen by the graph below, this model is obtained by the lambda value that creates the smallest Mean Squared Error, in which the Lambda value is 0.1 and the MSE is 13961.

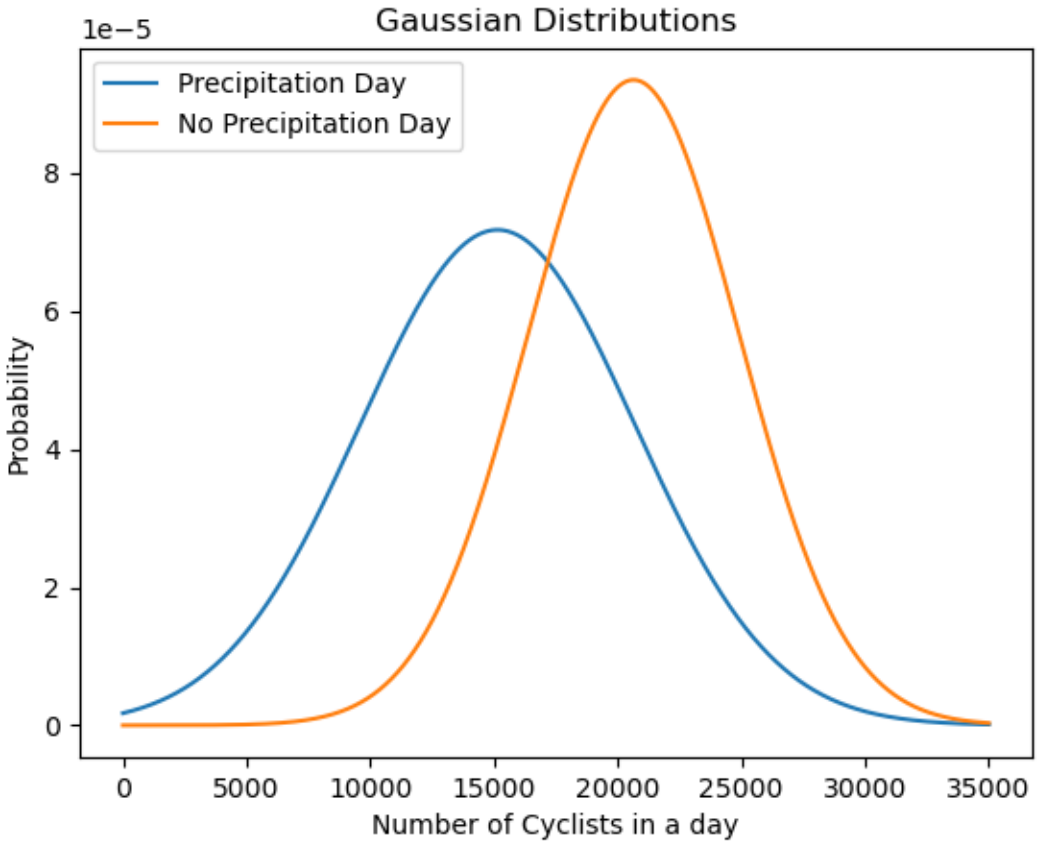


As observed, the MSE value obtained is extremely high, and moreover, the correlation coefficient yielded is at a low of 0.159, meaning that the weather cannot be used to accurately predict the number of bikers on a certain day.

3 – Raining Prediction

Out of the 214 datapoints available, we separated them between two categories, raining days and non-precipitation days, each respectively with 86 and 128 datapoints. To create the Naïve Bayes Model, we separated both categories in training and testing data, for each with a goal of separating on a 3:1 ration. The raining days got separated with 64 days for training and 22 for testing. As for the non-precipitation days datapoints, 96 were used for training and the remaining 26 were for

testing. After training the data, we built the two Gaussian Model for each training set, resulting in the following plot.



With the remaining test data, we could verify how accurate was the created models. For raining day tests, we got 54.54 % correct predictions, meaning 12 out of the 22 tests successfully predicted that it was raining day, based on the number of cyclists, which does not indicate a good relationship. Also, for no precipitation day test the accuracy was of 78.12 %, which indicates that 25 out of the 32 tests, correctly predicted that it was not a raining day based on the number of cyclists in a day, which is better for a prediction model, but not as high to be used as a predictor model. Finally, all the tests had a combined accuracy of 68.51 % to correctly predict if it is a raining day or a non-precipitation day, based on the number of cyclists on bridges in a day, indicating that only 37 cases passed out of 54. So, we can conclude that it is not reliable to use the dataset to predict whether it is raining or not, based on the amount of daily people crossing the bridges on their bikes.