Project ECE 20875: Python For Data Science
Spring 2022

Project Team Information:

Mini-Project Spring 2022
ECE 20875-003
Joao Taff-Freire - Frucks - jtafffre@purdue.edu
Mateusz Romaniuk - romanczug - mromaniu@purdue.edu
Path 1: Bike Traffic

Descriptive Statistics:

We decided to use the first dataset. The data was gathered from four bridges in New York City (Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge and Queensboro Bridge). The data was gathered every day from April 1st to October 31st, and contains what day it was, the highest and lowest temperatures, the precipitation and how many cyclists passed through each bridge.

Approach:

1) **There are only enough sensors for three bridges. Which bridges should we install the sensors on to get the best prediction of overall traffic?**

The first step we took to determine this was to graph the frequency of the amount of cyclists in specified ranges and observe each bridge's distribution. A bridge that has a distribution that does not look like the others could be considered an outlier and be discarded. We also looked at the average daily traffic for each bridge, and the total daily average. The bridge with the average that is furthest away from the total average can also be a good indication of a bridge that does not accurately model the traffic.

2) **Can we use the weather forecast to predict the total number of cyclists?**

In order to answer this question, we decided to utilize linear Ridge Regression with the High Temperature, Low Temperature and Precipitation values as the independent variables. The Mean Squared Error and coefficient of correlation obtained in the Ridge Regression allows us to determine how effective our model can predict

traffic. We expect our model to be successful, as low temperatures and high precipitation seem like a good deterrent for cyclists.

3) **Can we use our data to predict whether it is raining based on the number of cyclists?**

The idea we came up with to answer this question was to plot a gaussian distribution with a model obtained using a Naive Bayes classifier. To do this we separated our data in two lists, one for days with precipitation and one for days without. For each list, we selected 75% of the data for testing and 25% for training, a 3:1 ratio. We used the training data to create the Gaussian distribution and compare it to the testing data. Similarly to question 2, we expect this prediction to be successful.

Analysis:

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?***

As a first step to answer this question, we plotted the distribution of cyclists as shown in the figure below. The X axis is the number of cyclists divided in bins with a determined range. The Y axis is the frequency of those ranges.

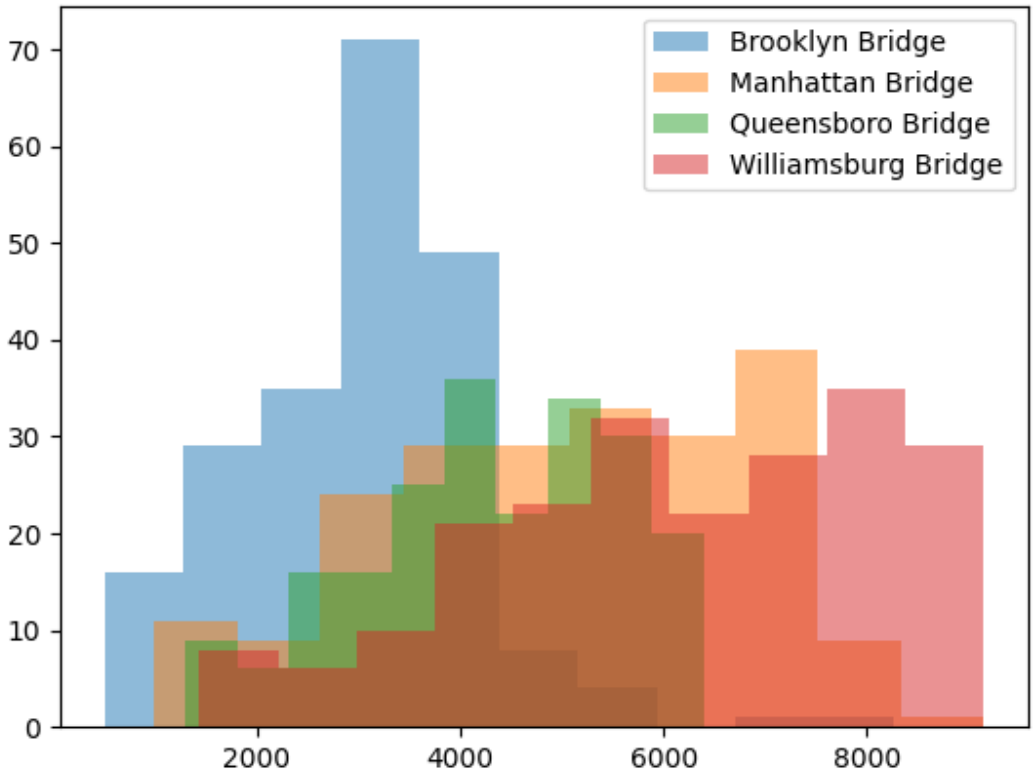

Figure 1. Distribution of cyclists in each bridge.

We observed that the distribution of the Brooklyn Bridge was extremely different from the rest, and that was the main reason we decided to put the sensors on the other three bridges. We also looked at the traffic daily average of each bridge. We obtained the following averages: Manhattan Bridge - 5052, Brooklyn Bridge - 3031, Queensboro Bridge - 4301, Williamsburg Bridge - 6161. When compared to the Total daily average (4636), Brooklyn Bridge's average was the furthest from it. That metric solidified our belief that the sensors should go to the Manhattan, Queensboro and Williamsburg bridges in order to best predict 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 (low/high temperature and precipitation) to predict the total number of bicyclists that day?***

After performing Ridge Regression, we obtained a Lambda value of 0.1 with a corresponding Mean Squared Error (MSE) of 13975003.011 and an R^2 (coefficient of

determination) value of 0.158. The model we obtained was the following: Total Traffic = -1938.759 * x_1 -1553.886 * x_2 + 4355.151 * x_3 + 18412.6. Shown below is a plot that shows how the Mean Squared Error changes according to the value of Lambda.

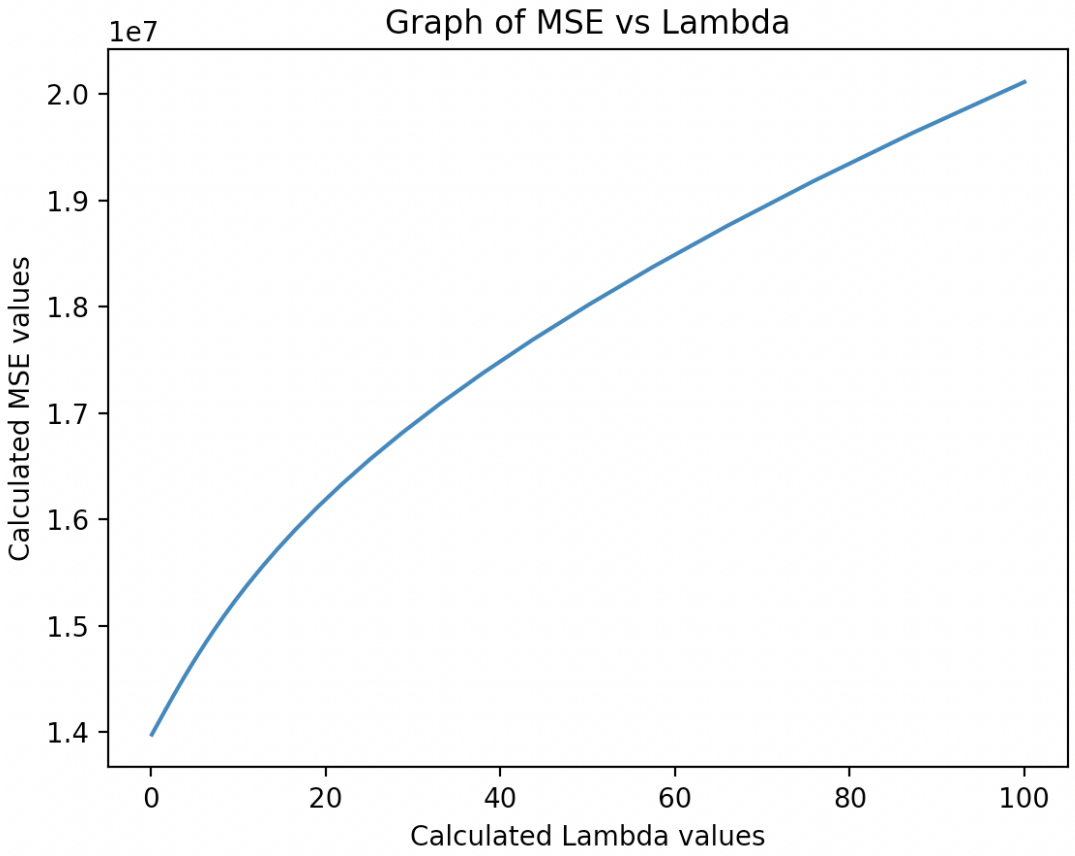

Figure 2. Plot of MSE vs Lambda.

Due to the extremely low coefficient of determination (the closer to 1 the better, it determines how well the independent variable can predict the dependent variable) and the extremely high MSE (lower is better, it measures the difference between the estimated values and the actual values), we can confidently claim that the weather cannot accurately predict the total number of cyclists.

3) ***Can you use this data to predict whether it is raining based on the number of bicyclists on the bridges (hint: The variable raining or not raining is binary)?***

Using the training data we separated for our Naive Bayes classifier, we were able to plot the models for rainy days and non-rainy days. That plot is shown below.

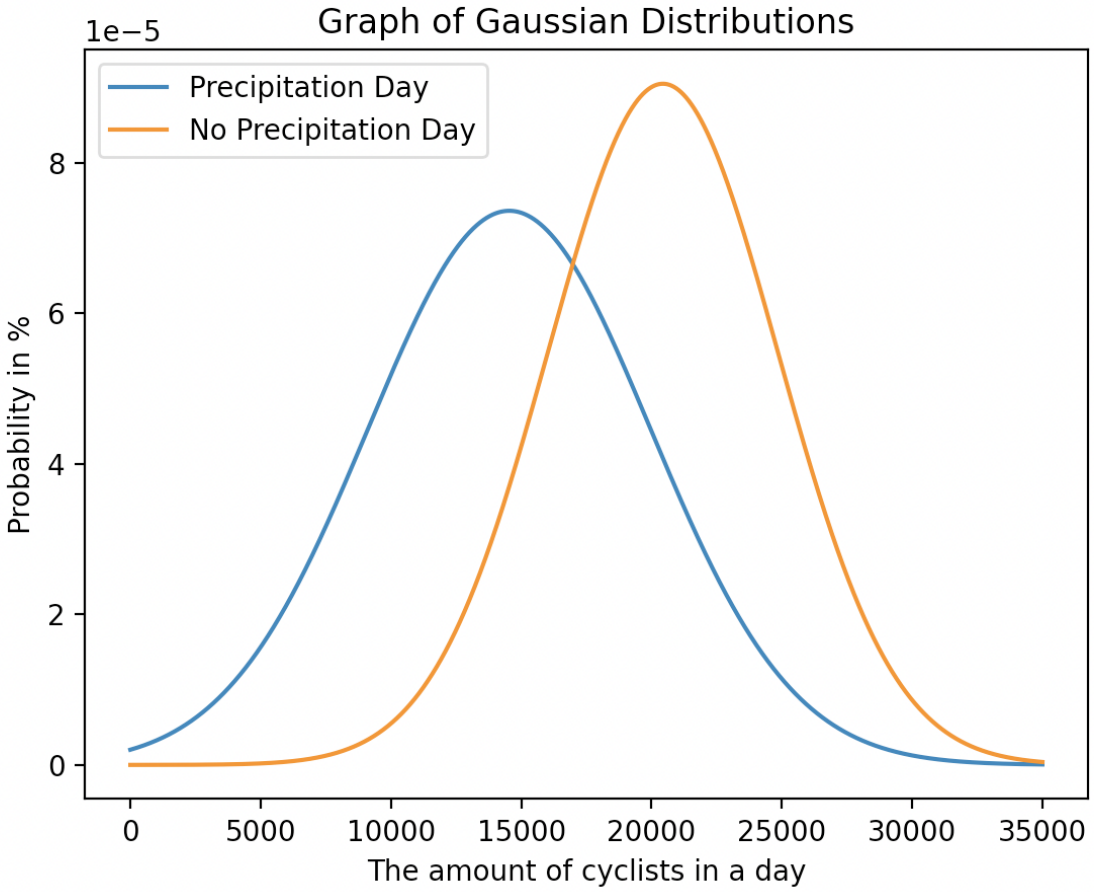

Figure 3. Plot of the gaussian distributions for days with and without rain (training data).

We used the data we separated for testing to determine the accuracy of our models. We obtained a test accuracy of 58.823% for days with precipitation, and 78.378% for days without precipitation. Our total accuracy was 72.222%. We believe this is not accurate enough to properly predict whether it is raining based on the number of bikers on the bridges, especially with a dataset as small as ours (17 tests for rainy days and 3 7for non-rainy days).