Nikita Ravi (ravi30)
Andrea Renee Levenson (alevens)
Project: Path 1

ECE 20875 Final Project: Bike Traffic

Describing the Data Set

This dataset is from the New York City Department of Transportation. Daily data was collected regarding the number of bicyclists going over the Brooklyn, Manhattan, Williamsburg, and Queensboro bridges. The date, highest temperature, lowest temperature, precipitation, and total bicyclists were also recorded.

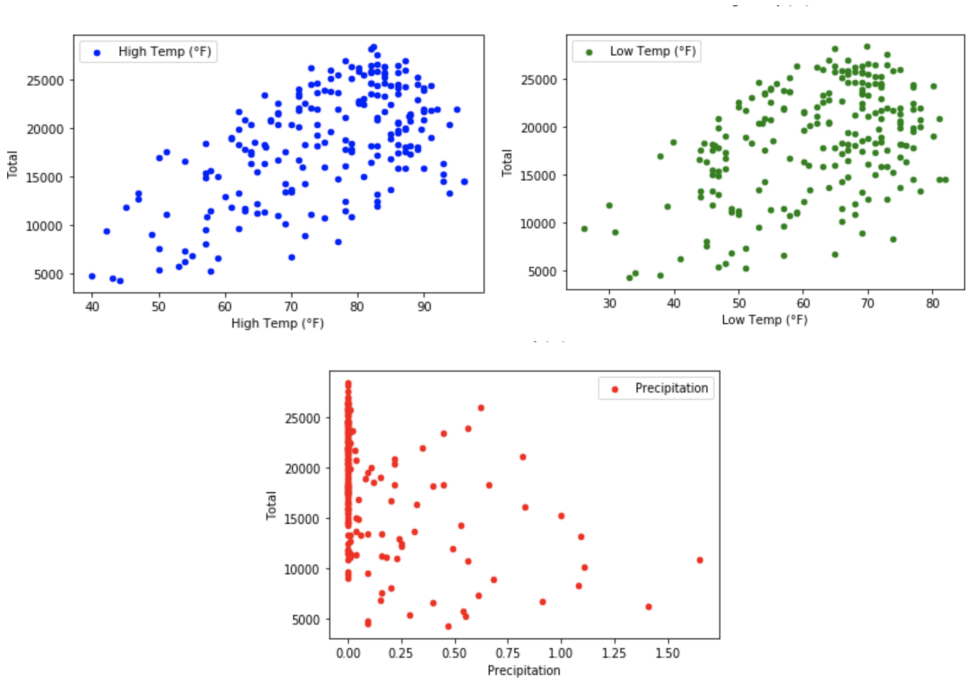
Analysis Chosen

Question 1

Question 1 asked us to install sensors on the bridges to estimate overall traffic, but the sensors can be installed on only three of the four bridges. To choose the three bridges, we decided to look at the average traffic of each bridge on weekdays, weekends, and overall. We could have simply taken the maximum values of each bridge and compared them, but that wouldn't be an accurate representation of overall traffic. Traffic changes daily because of various factors, like precipitation, temperature, time of the week, and many more. This is why we also decided to look at the weekday and weekend averages to see if the average traffic changes. If the average traffic is higher on the weekend, it may skew the overall average traffic, not allowing us to see that we want sensors on the bridges with the highest weekday traffic.

Question 2

Question 2 asked us to predict the number of bicyclists on a particular day using the weather forecast. Since we had labelled data (the number of bicyclists), we knew that our algorithm must use supervised machine learning, but we had to decide which type of supervised machine learning we wanted to use: regression or classification. In order to answer this question, we first plotted a graph of each independent variable (High Temperature, Low Temperature, and Precipitation) against the total number of bicyclists, as shown below:



From the graphs above we can clearly see that as both High Temperature and Low Temperature increases, the total number of bicyclists increases. We can also see that as precipitation increases, the total number of bicyclists decreases. Taking this into account, we believe that the best approach to solving this problem is by using regression.

For this particular problem, we chose to model the data using linear, quadratic, cubic and ridge regression models because this will give us the ability to predict the number of bicyclists in the future at a range of high temperature, low temperature, and precipitation values.

We chose a linear regression model because, as mentioned before, it looks like there is a positive correlation between the high temperature and total number of bicyclists and between low temperature and total number of bicyclists, and a negative correlation between precipitation and total number of bicyclists.

We also chose quadratic and cubic regression because we wanted to account for a gradual increase or decrease in the number of bicyclists, given an increase or decrease in high temperature, low temperature, and precipitation. This gradual change can be prominently seen in the precipitation graph on the previous page.

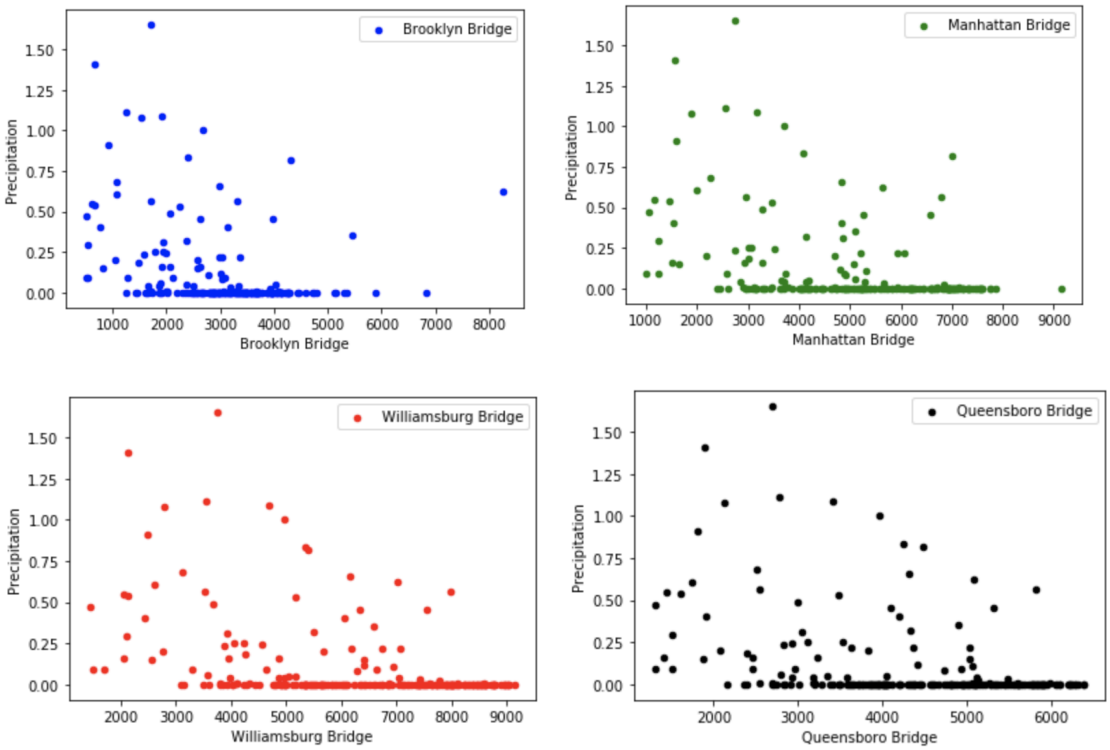
We chose ridge regression for linear, quadratic, and cubic because we wanted to know what the accuracy of our prediction is without overfitting the data. In other words, we wanted to avoid minimizing bias and maximizing variation, and attempt to get an accurate coefficient of determination.

Question 3

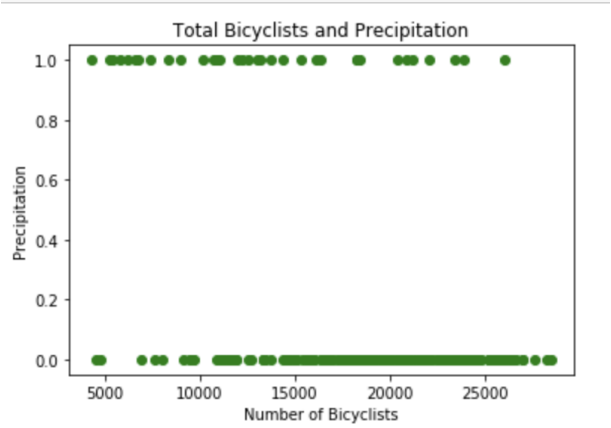
Question 3 asked us to use the data to predict whether or not it was raining. We make use of supervised machine learning because we have a labelled target value we are trying to predict.

But once again, we have to determine whether or not we want to use a classification model or a regression model. Unlike in question 2, in question 3 we have to determine if it will rain on a particular day, based on the number of people on each bridge.

In order to determine this, we once again plotted graphs for each independent variable against the precipitation values, as shown below:



From the graphs above, we can see that there are a smaller number of bicyclists on each bridge on days when there is a high precipitation, whereas there are a lot many bicyclists on days when there is a low precipitation value. As a result, we believe the best supervised machine learning algorithms to use is classification.



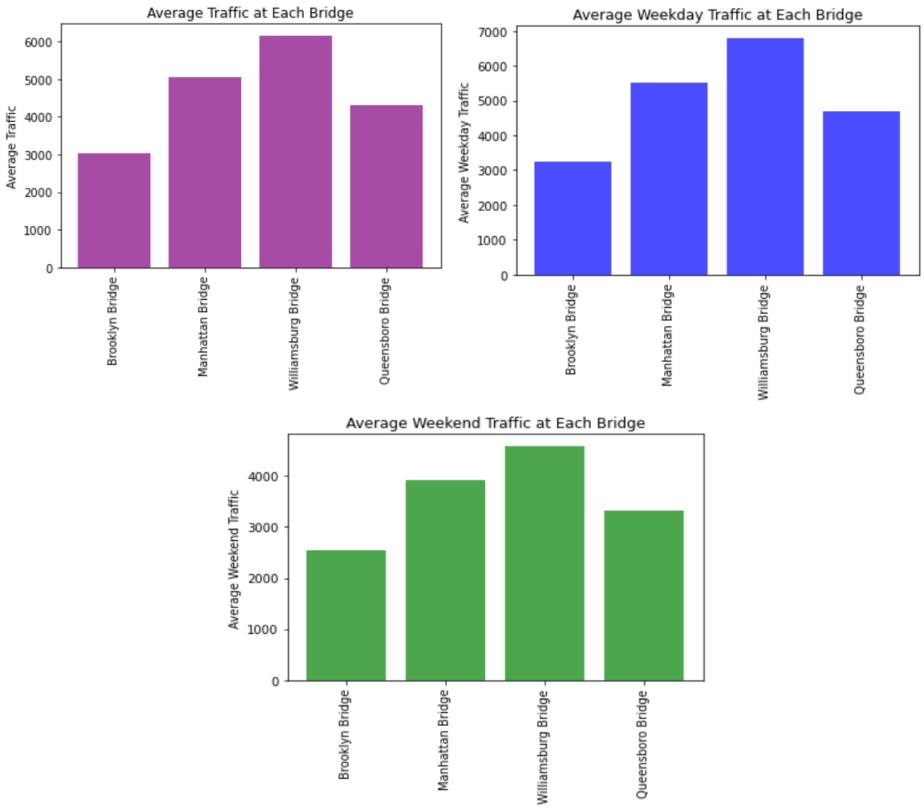
Looking at the graph on the left, you can clearly see that there is a divide between the values where there is no rain, and values where there will be rain. As a result of having this particular divide, we used logistic regression and support vector machine algorithms to predict the probability of it raining on a particular day based on the number of bicyclists since these two algorithms build this boundary between the two groups.

We also chose to implement a Bernoulli distribution seeing as there are only two options: the probability that it will rain, and the probability that it will not rain on a certain day based on the number of bicyclists.

From research, we determined that any precipitation that is less than or equal to 0.2 will represent that it will not rain and anything greater than 0.2 will represent that it will rain on that particular day (Means, 2019).

Results of Analysis

Question 1



The charts above represent the average daily, weekday, and weekend traffic for all four bridges. The Manhattan Bridge, Williamsburg Bridge, and the Queensboro Bridge all have the highest daily, weekday, and weekend average traffic showing us the time of the week has no effect on average traffic. Since we decided to install the sensors on the three bridges with the highest average traffic, we would install the sensors on the Williamsburg Bridge, Manhattan Bridge, and the Queensboro Bridge.

Question 2

The following chart below shows the mean-squared-value (MSE), coefficient of regression (r^2) values, and the best regularization parameter (λ) value for each of the models built:

	Linear	Quadratic	Cubic	Ridge Linear	Ridge Quadratic	Ridge Cubic
MSE	17498379.32	17049458.16	103610650.74	17476450.30	17056507.13	15758031.68
r^2	0.58	0.59	-1.52	0.49	0.49	0.32
λ	-	-	-	1.23	0.10	14.17

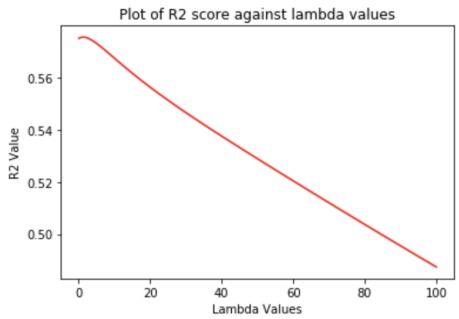
The linear and quadratic model had a coefficient of determination of 0.58 and 0.59 respectively which indicates that only 58% of the data and 59% of the data, respectively, fit the model. This is a bad model since there is a huge error between the actual data points and the regression line. In other words, the low coefficient of determination value indicates that there is a high variability in the data. Therefore, both the linear and quadratic models are bad models.

For the cubic model, the coefficient of determination is less than 0, it is -1.52. The coefficient of determination is only negative when the regression line is worse than a horizontal straight line, representing the null hypothesis. This occurs because the sum of squares of the regression line is greater than the sum of squares of the horizontal line ($1 - \frac{SS_{reg}}{SS_{tot}}$). Therefore, the cubic model was a bad model.

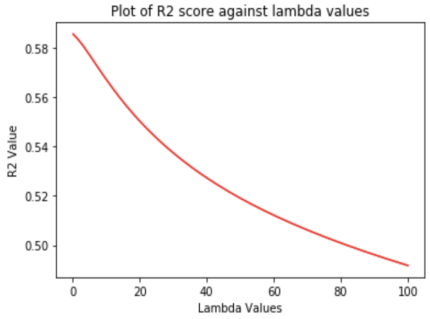
Considering that the linear, quadratic, and the cubic models are not good models to be used to predict the number of bicyclists based on weather forecasts, we decided to build a ridge model for linear, quadratic, and cubic and see if either minimizing the model parameters or minimizing the model error will somehow improve the coefficient of regression.

We determined the best regularization parameter by plotting a graph of a range of all the lambda values against their respective coefficient of determination as shown below:

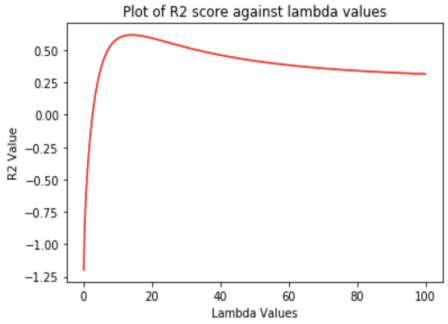
Linear Regression



Quadratic Regression



Cubic Regression



The best regularization parameter has the highest coefficient of determination value, which is shown on the table above.

For the linear ridge regression and quadratic ridge regression has a regularization parameter of 1.23 and 0.1 respectively, which indicates that for this model, minimizing the model error was important. Both the linear ridge regression and quadratic ridge regression have a lower coefficient of determination than the actual linear and quadratic regression lines, which implies that the actual linear and quadratic models minimized bias and maximized variation, i.e. the model was overfitting.

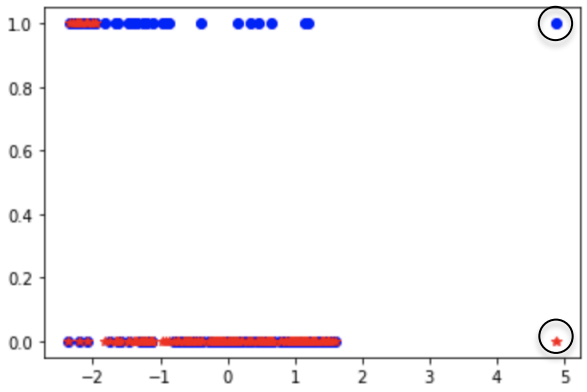
For the cubic ridge regression, the regularization parameter is 14.17 which indicates that for this model, minimizing the model parameters was important. This significantly increased the coefficient of regression from -1.52 from the cubic regression to 0.32 (cubic regression). However, this still is not a good enough model since only 32% of the data is represented by the regression line.

Overall, since all the models built have an extremely high MSE and extremely low coefficient of determination, we believe that the models are not a good representation to use to predict the number of bicyclists on a specific day based on weather conditions.

Question 3

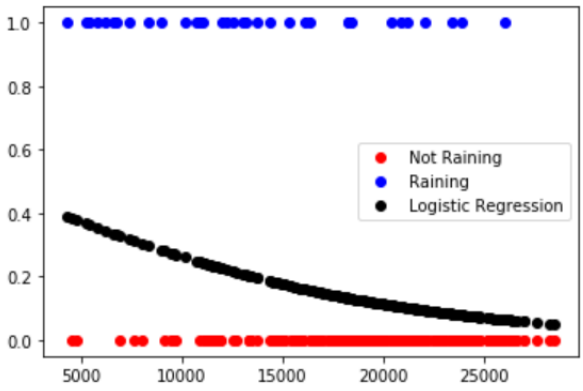
We created three different models to determine if we can predict if it is raining based on the number of bicyclists on the bridges. The accuracy of each model is shown in the table below:

	Logistic Regression	Naive Bayes	Support Vector Machine
Accuracy	0.860	0.744	0.698



The logistic regression model was a moderately good model in the sense that 86% of the data matched the model. The graph on the left shows the plot of the testing part of the X-values against the testing part of the Y-values (in blue) and a plot of the testing part of the X-values against the predicted Y-values (in red).

From this chart you can see that the majority of the predicted Y-values match the actual testing Y-values, but there are some minor differences such as the one circled in the graph. The only issue we found with this model is that the logistic regression curve does not match that of a sigmoid function, as shown on the right. However, since it still separates the two categories (i.e. raining and not raining), we believe that it is a moderately good model.



The last two models we created were eliminated because they both had low accuracies. The accuracy of the Naive Bayes model is 0.744, so only 74.4% of the data was represented by the model, and accuracy of the Support Vector Machine model is 0.698, so only 69.8% of the data was represented by the model.

Sources

Means, T. (2019, August 8). *What "Chance of Rain" Really Means*. Retrieved August 3, 2020, from https://www.thoughtco.com/chance-of-rain-3444366