# ECE20875 Final Project

Team Member Name(s): Hanyu Zhu, Siyue Shen

Purdue Username(s): zhu741, shen344

GitHub Username(s): zhu741, LuS6

GitHub Team Name: ssy (https://github.com/ECEDataScience/project-s20-ssy)

Team Project: Path 1: Bike Traffic

## Dataset:

The data set we are working with is “New York City - East River Bicycle Crossing” which contains daily bicycle counts for major bridges in NYC. The data points within the data sets includes date, day, high temperature in Fahrenheit, low temperature in Fahrenheit, the precipitation of that day, the bicyclist counts for each major bridge in NYC including Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge, and finally the total bicyclists count on all bridges.

We load all columns as lists beside the columns of date and day. Then, we convert them from string type into float type for further calculations. Note that the Precipitation column includes characters ’S’ (as snow) and ‘T’ (as Trace Amount), which needs extra modification to extract the useful data. For data points containing ‘S’, the modification we performed is to disregard the string ‘(S)’ and only use the float at front. For data points containing ‘T’, we regard it as 0 since it means trace amount which is negligible.

For question 3, in order to make our Precipitation data useful for the Multinomial Naive Bayes model, we modify Precipitation into a binary data set, which states 1 when it's raining and 0 when it's not raining.

## Methods:

## Question 1:

First, we normalize all the data. For bicycle counts of all four bridges and total count, we normalize them by minusing each data point in each dataset with the mean of the dataset and then dividing it with the standard deviation of the dataset. After normalization, all datasets are of the same scale, and thus, we could perform further calculations.

We need to choose 3 bridges out of 4 to install sensors, and there are four different combinations that we need to analyze with.

Combination 1: Brooklyn Bridge, Manhattan Bridge, and Williamsburg Bridge (BMW)

Combination 2: Brooklyn Bridge, Manhattan Bridge, and Queensboro Bridge (BMQ)

Combination 3: Brooklyn Bridge, Williamsburg Bridge, and Queensboro Bridge (BWQ)

Combination 4: Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge(MWQ)

Then, for each combination, we take the average of the 3 bridges' count for every single day in the dataset, and then, use the calculated average as the bicycle count for the specific combination. After making the bicycle count for the 4 combinations mentioned above, we then

calculate the mean square error (MSE) between each combination and the normalized total count. Finally, we would choose the combination with the lowest MSE to install sensors.

## Question 2:

The method we use in order to predict the number of bicyclists by using the next day's weather forecast is to implement a Ridge Linear Regression model on the data set. First we separate the data to 75% training data and 25% testing data. Then we trained the model with 75% of the normalized data set and tested it with 101 logarithmic lambda values. By finding the set where we have the lowest mean square error, we found the best lambda with the best model to predict.

The way we evaluate our model is by calculating our found model's coefficient of determination $r^2$ . If our $r^2$ is within the acceptable range ( $r^2 > 0.6$) to predict the number of cyclists on bridges then it proves that we can predict the number of cyclists by using the weather forecast. If our $r^2$ is not within the acceptable range ( $r^2 < 0.6$) to predict the number of cyclists, then it proves that we can not predict the number of cyclists.

## Question 3:

In order to determine whether we can use our data set to predict if it is raining based on the number of bicyclists on the bridges that day, we decide to use the multinomial Naive Bayes model on our data set since the input variable is bicycle count, which is a discrete variable. First we separate the data to 60% training data and 40% testing data. Then, we can classify our dependent (y in code) data set into two classes whether it's raining or not (raining [1] and not

raining [0]) , and calculate the density of each class within all the bicyclists counts. By importing metrics methods from the sklearn library, we are able to use accuracy_score to evaluate the accuracy of our predicted results.

## Results:

## Question 1:

The following data is the result we got from our algorithm.

*The mean square error between BMW and total 3.222660084541005.*

*The mean square error between BMQ and total 4.027232857799984.*

*The mean square error between BWQ and total 6.3326281318026565.*

*The mean square error between MWQ and total 2.925215520287073.*

The MWQ combination scores the lowest MSE score. Hence, we would install the sensors on **Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge** to get the best prediction of overall traffic.

## Question 2:

The following statement is the result we got from our program after running our algorithm mentioned in method for question2.

*Best lambda tested is 0.1, which yields an MSE of 0.43183604633574585.*

Based on the printed statement above, the best lambda we get is 0.1. It is reasonable to be treated as 0 since our lambda scale is in logspace which could only  be infinitely close to zero but could never reach zero. Thus, by treating lambda as zero, we are able to perform linear regression on our dataset. The MSE of our predicted model is about 0.43. The relationship between MSE and Lambda is shown in the figure 2.1 below, which we could also observe that the MSE minimizes around lambda is zero.

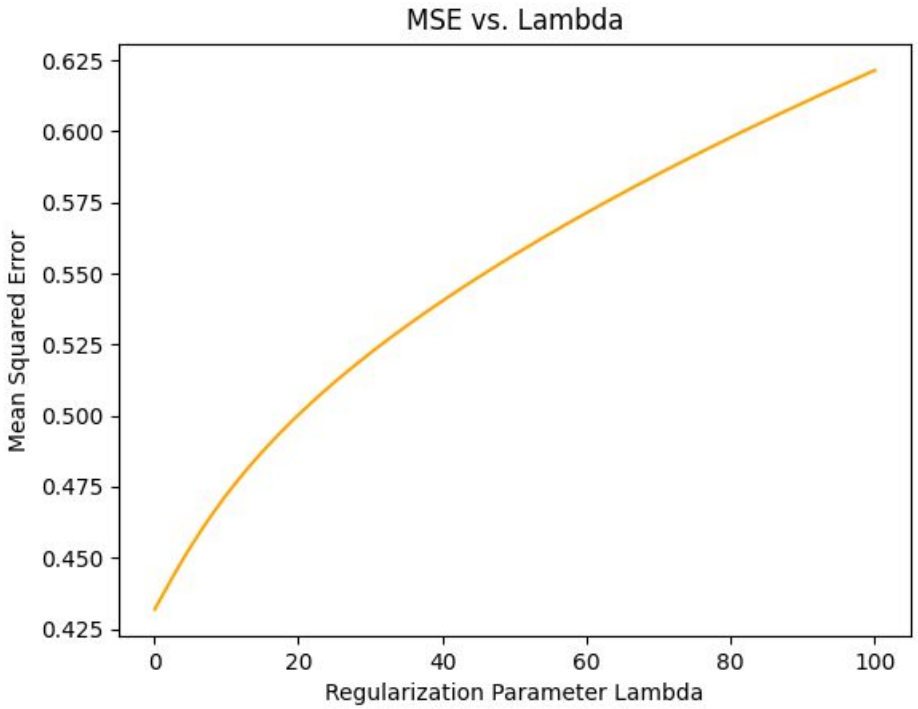


**Figure 2.1 Relationship between MSE and Lambda**

The linear model we get after training the dataset with linear regression algorithm is

$$Bicycle\ Count\ =\ 0.77T_H - 0.27T_L - 0.34P - 0.02$$

where, $T_H = High\ Temperature\ of\ the\ day$,

$T_L = Low\ Temperature\ of\ the\ day$,

$P = Precipitation\ of\ the\ day$

Based on our linear model, we use the reserved testing dataset to calculate the Coefficient of Determination ( $r^2$ ) of the predicted data and actual data.

*coefficient of determination: 0.6211262071677414*

The $r^2$ calculated for our linear model is 0.62, which shows moderate correlation between the weather and number of bicyclists on the bridges. **Thus, we conclude that it is reasonable to use next day's weather forecast to predict the number of bicyclists that day.**

## Question 3:

We used 40% of the data set as testing data and 60% of the data set as training data. The following statement is the result we got from our program after running our algorithm mentioned in method for question3.

*Multinomial Naive Bayes model accuracy (in %): 75.5813953488372*

The accuracy score means 75.6% predictions made by our trained Multinomial Naive Bayes model are correct when applying on the test dataset. 75.6% is a reasonably high accuracy for making predictions. **Thus, we conclude that we can use this data to predict whether it is raining based on the number of bicyclists on the bridges.**

## Reference:


I. *Homework 8: Linear Regression, Problem 2: Regularized regression* :

https://github.com/ECEDataScience/homework-8-s20-zhu741/blob/master/regularize-cv.py

II. *Class example on Naive Bayes from 4/17:*

http://www.cbrinton.net/ECE20875-2020-Spring/W14/naive-bayes.pdf