ECE 20875 Final Project Report

Team Member Name(s): Yuqing Fan and Yuxin Zhang

Purdue Username(s): fan230 and zhan2918

GitHub Username(s): vickyfyq and YuxinZh

GitHub Team Name: anorc

Project: Path 1: Bike Traffic

Dataset:

The features are 'Date', 'Day', 'High Temp (°F)', 'Low Temp (°F)', 'Precipitation', 'Brooklyn Bridge', 'Manhattan Bridge', 'Williamsburg Bridge', 'Queensboro Bridge', and 'Total'. Units for High Temp and Low Temp is Fahrenheit (°F). Unit for precipitation is millimeters (mm) per square meter. There are 214 samples in total. The range for High Temp (°F) is 56.2; The range for Low Temp (°F) is 55.9; The range for Precipitation is 1.65; The range for number of bikes on Brooklyn Bridge is 7760.0; The range for number of bikes on Manhattan Bridge 8155.0; The range for number of bikes on Williamsburg Bridge is 7708.0; The range for number of bikes on Queensboro Bridge is 5086.0; The range for Total bikes on bridges is 24102.0. The source of the dataset is the New York City Department of Transportation, and we downloaded the data from the webpage https://www.kaggle.com/new-york-city/nyc-east-river-bicycle-crossings/discussion/53852. This dataset is a daily record of the number of bicycles crossing into or out of Manhattan via one of the East River bridges.

Methods:

The first question asked us to estimate the overall traffic across all the bridges by installing sensors on three out of four bridges, due to the budget constraint. Since we can only install sensors on three bridges instead of four, we need to find out which three bridges contributed the most useful data for prediction of overall traffic. The method we used was polynomial regression data fit. There are three reasons why we choose polynomial regression: firstly, broad range of function can be fit under it; secondly, polynomial provides the best approximation of the relationship between dependent and independent variables, and thirdly, there are not too many outliers in the dataset so the estimation won't be affected too much. We normalized the data since the range of different features are different and it will increase the accuracy of our results. Before doing the polynomial regression, we separated the data into two sets - half of the data is used for training and the other half of the data is used for testing the accuracy. The training data are combined into four different groups, and each group has three out of four bridges. Then we used the function PolynomialFeatures and LinearRegression in library sklearn to

get the polynomial regressions of those groups, with the traffic on the three bridges as independent variable and total traffic as dependent variable and with a degree of 2, which is determined by the number of bridges in the group minus one. After we get the polynomial regressions, we test them with the testing data and calculate the MSE and r-square between estimated results and actual results. The group with the least MSE and largest r-square has the best estimation, and we are going to install sensors on the three bridges in the group to estimate overall traffic.

For the second question, the city administration wants to know if they can use the next day's weather forecast to predict the number of bicyclists that day. We decide to use ridge regression for this question. We chose this method because the data set has multicollinearity (there's correlations between predictor variables such as high temp and low temp). We chose three features, including high temp, low temp, and precipitation to represent the next-day weather. We divided the data into two data sets, one for training and one for testing. We also normalize the data since those features have different ranges and different units. Our target in this case is the total number of bicyclists on the corresponding date, and we also tried to predict the number of bicyclists on each bridge so that the city administration knows which bridge needs more police officers. Since there are three features, to prevent overfitting, we used regularization to penalize non-zero model coefficients. Then, we use Ridge and linear_model functions in library sklearn to find the ridge regression. To determine whether the weather can predict bicyclists on that day, we used MSE and mean absolute percentage error (MAPE). If the MAPE is under 25%, they can use the next day's weather forecast to predict the number of bicyclists that day; they can't if the MAPE is over 25%.

For the third question, we are asked if we can use this data to predict whether it is raining based on the number of bicyclists on the bridges. We decided to use the naive bayes to predict if it is raining, because what we want in this problem is a simple probabilistic classifier with strong independence assumptions between the features. The feature total number of bicyclists on the bridges is used to build the model, and the target is if it's raining. Since what we need to predict is whether it's

raining, we can set raining as 1 and not raining as 0. When precipitation is larger than zero, we consider it's raining. If precipitation is zero, then it's not raining. Since the feature is a continuous number of bicyclists, we decide to use the gaussian naive bayes classifier to predict whether it's raining or not. We first separate data randomly by half to create a set of test data and a set of train data. We use the train data to train the model. Then we use the test data to accomplish the prediction and calculate the accuracy of prediction. The accuracy of prediction is determined by calculating (True Positives + True Negatives)/(True Positives + True Negatives + False Positives + False Negatives). If the accuracy is over 80%, we consider the model accurate and we consider the model not accurate otherwise.

Results:

For the first question, the four groups we have are : 1. 'Brooklyn', 'Manhattan','Queensboro', 2. 'Manhattan','Queensboro','Williamsburg', 3. 'Brooklyn', 'Queensboro','Williamsburg', 4. 'Brooklyn', 'Manhattan','Williamsburg'. The MSE we got from the four groups are 267447.47829, 440015.92205, 770408.50265, 92344.16867, respectively, and the r-squares are 0.99169, 0.98633, 0.97607 and 0.99713. The diagrams of the four groups are attached below in figure one, figure two, figure three, figure four below, respectively. From the results we found that group four had the best estimation of overall traffic with the lowest MSE of 92344.16867 and the highest r-square of 0.99713, so we are going to install the sensors on bridges in group 4, which are the Brooklyn Bridge, the Manhattan Bridge and the Williamsburg Bridge.

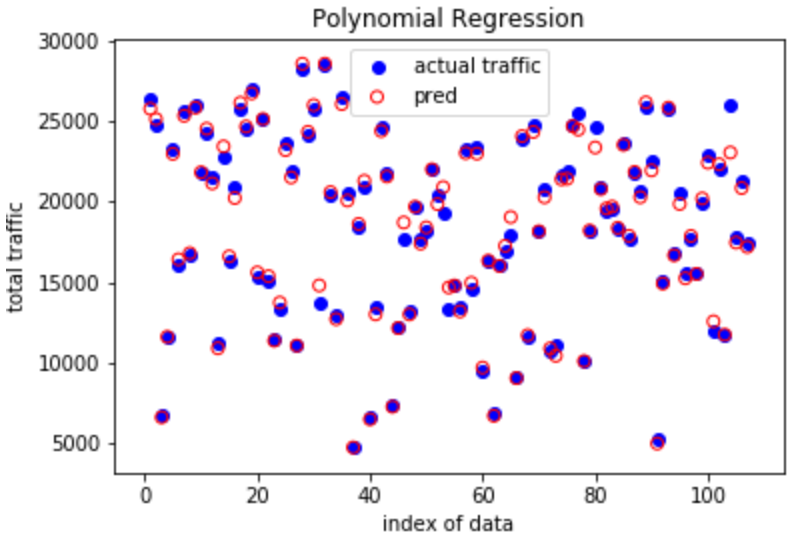

Figure 1: Estimation and Actual Overall Traffic of Group 1

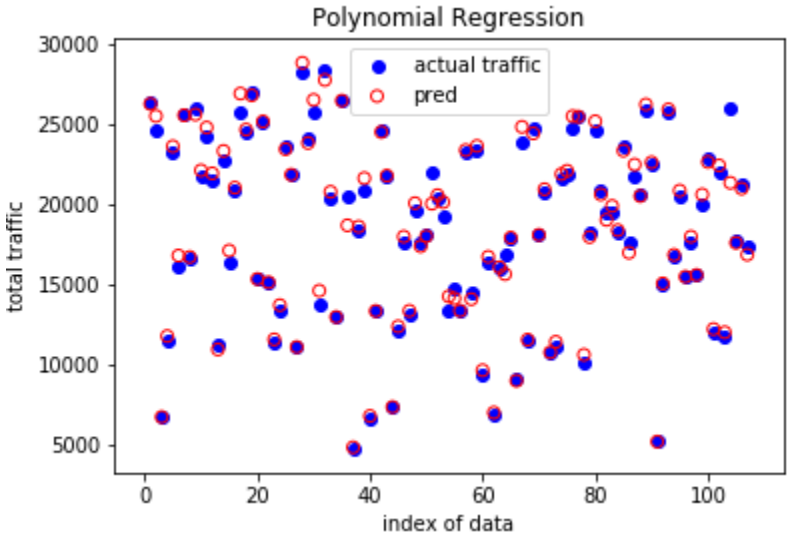

Figure 2: Estimation and Actual Overall Traffic of Group 2

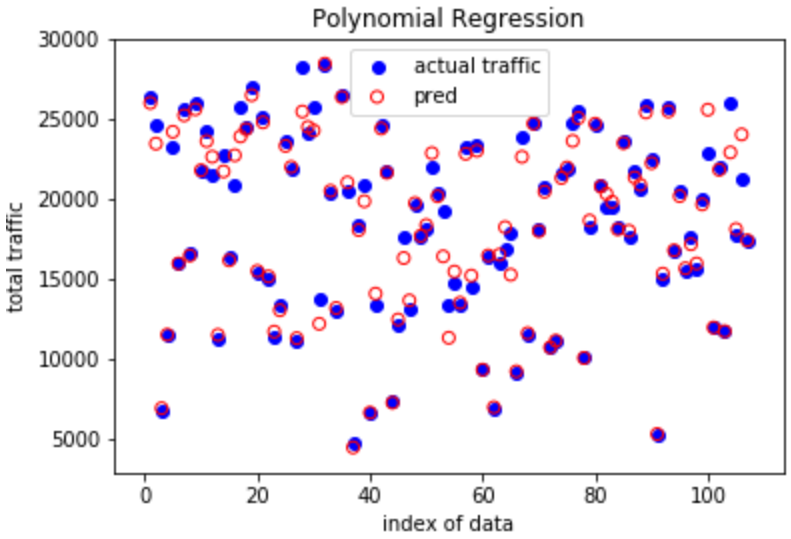

Figure 3: Estimation and Actual Overall Traffic of Group 3

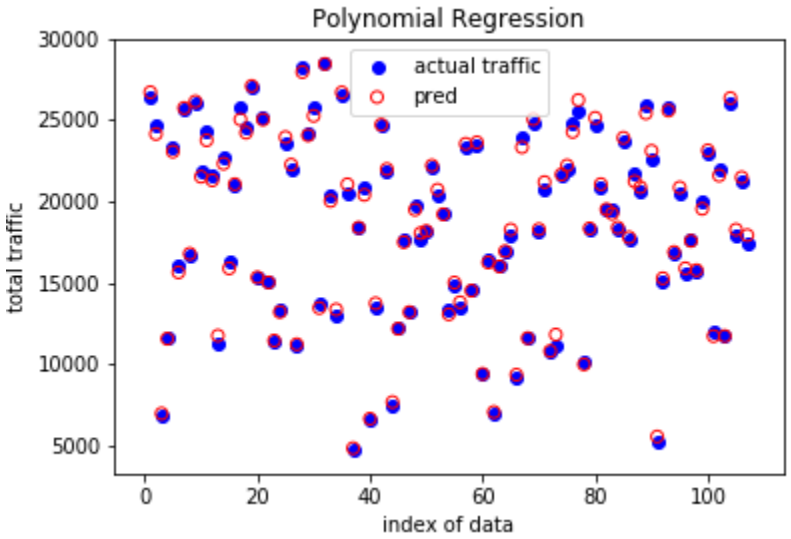

Figure 4: Estimation and Actual Overall Traffic of Group 4

For question 2, the best lambda tested for total traffic is 1.1220184543019636, which yields an MSE of 15305214.521989958. The diagram is attached below in

figure 5. The MAPE of the estimation is 18.560177%. The best lambda tested for Williamsburg Bridge is 0.26302679918953814, which yields an MSE of 1880618.0622138155, and the MAPE is 19.373186% (figure 6). The best lambda tested for Brooklyn Bridge is 0.8511380382023763, which yields an MSE of 689850.4924843025, and the MAPE of the estimation is 19.524537% (figure 9). The best lambda tested for Manhattan Bridge is 3.3884415613920256, which yields an MSE of 1803832.7924603703 (figure 8). The MAPE of the estimation is 23.T451737%. The best lambda tested for Queensboro Bridge is 0.1, which yields an MSE of 730044.0680356538 (figure 7). The MAPE of the estimation is 17.246927%. Since the MAPE of each bridge and the overall traffic are all less than 25%, we can say the city administration can use the next day's weather forecast to predict the number of bicyclists that day.

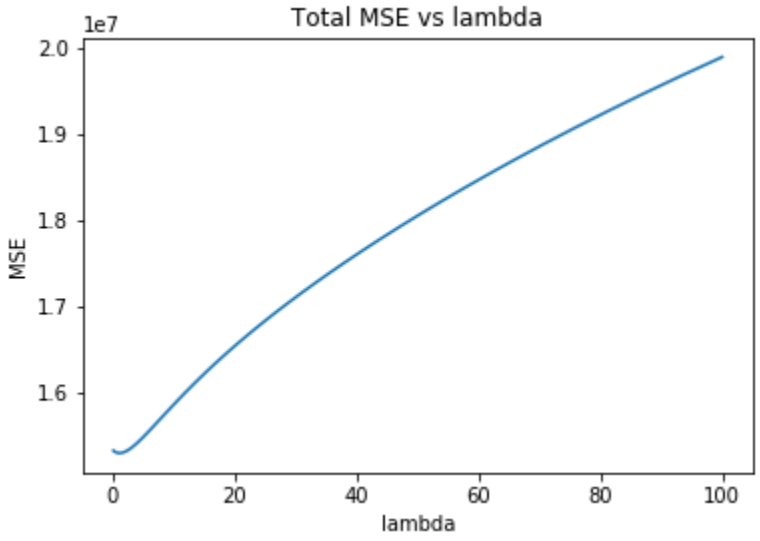

Figure 5: lambda vs MSE (Total Traffic)

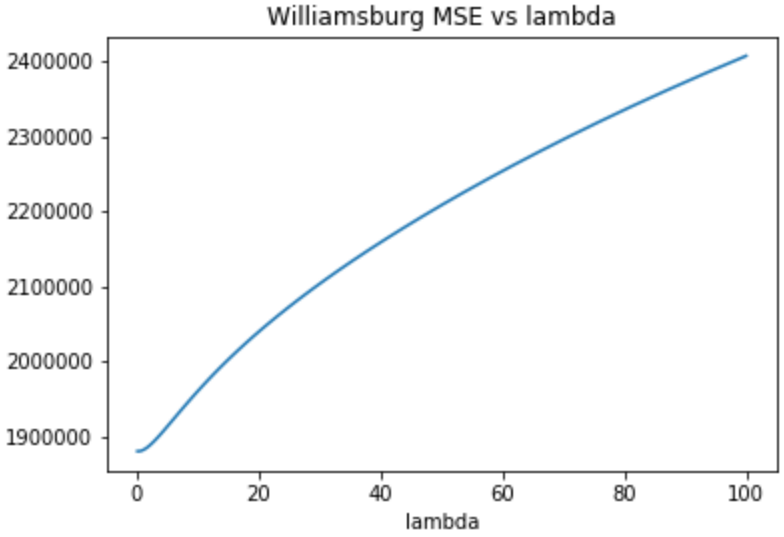

Figure 6: lambda vs MSE (Williamsburg Bridge)

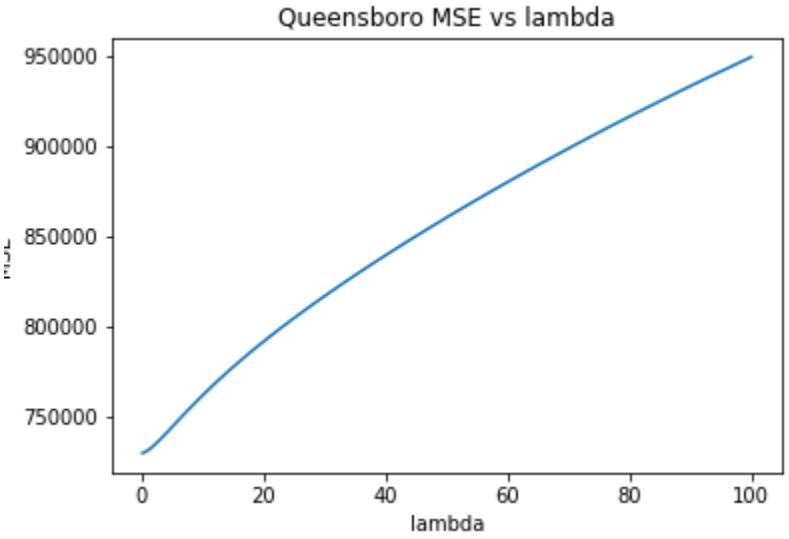

Figure 7: lambda vs MSE (Queensboro Bridge)

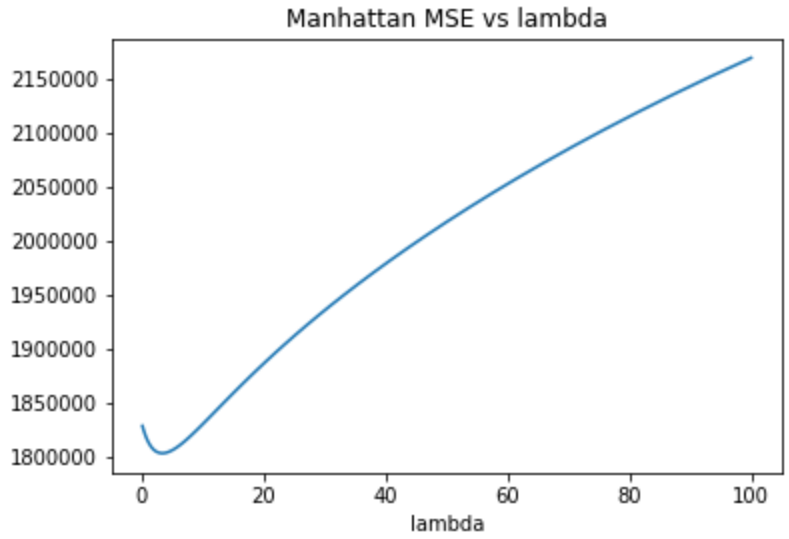

Figure 8: lambda vs MSE (Manhattan Bridge)

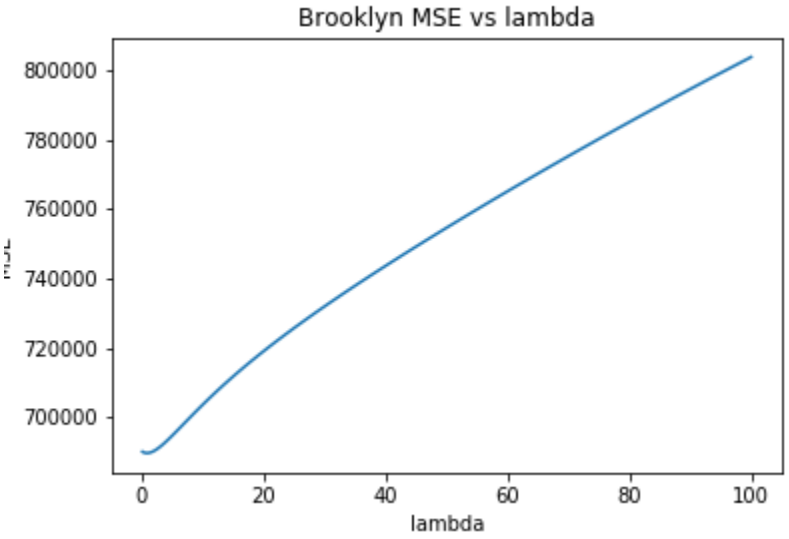

Figure 9: lambda vs MSE (Brooklyn Bridge)

For question 3, after we applied the method discussed in Methods part, we got that the accuracy of the model is 82.2429906542056%. Since this value is higher than 80%, we consider the method accurate enough, and we can use this data to predict whether it is raining based on the number of bicyclists on the bridges.