Pranav Sankar
Varun Ravi
April 23rd, 2022

# Mini Project - ECE 20875

**Project Team Information:**
Pranav Sankar - psankar193 - psankar@purdue.edu
Varun Ravi - ravi45 - ravi45@purdue.edu

We are doing Path 1.

**Descriptive Statistics:** For this project, we were asked to perform a series of statistical analyses given data about bikers on the largest four bridges in New York. The data tells us how many bikers were on the Brooklyn, Manhattan, Williamsburg, and Queensboro bridge along with weather factors such as temperature and precipitation from April 1st to October 31st.

**Approach:** For problem 1, we believed the best approach was linear regression. The problem was basically asking us to determine which three features (3 of the 4 bridges) can best predict the number of bikers on all four bridges per day and to determine this we sought to find the model that resulted in the smallest mean squared error.  For problem 2, we believed that the best method would be creating a model around the features and using the R-squared to determine how good of a model it was and whether or not it was possible to predict the number of bikes using the next day's weather. Finally, for problem 3 we went with a similar approach to the other two problems which is regression where the total number of bikers is the one feature and the probability of rain is the dependent variable. One major difference is that this problem contained a binary variable (whether it was raining) so we determined instead of using a linear regression we will proceed with a logistic regression because it will better fit the data. We then analyzed the logistic graph and confusion matrix obtained from the model created.

**Analysis:**

Problem 1 - To repeat, this problem is asking us to determine which three features can be best used to predict the total amount of bikers on the bridge. So to start, we created four feature matrices to represent all the combinations of bridges; these feature matrices paired with the total amount of bikers can be split to obtain our train and test X and y ( there will be a total of 16 parameters obtained because there are four feature matrices and we use the same y for each split). In addition, we decided it was best to randomly split the data with the standard 80-20 rule. We then normalized all parameters and obtained four different models representing each of the four cases. After testing the models with their respective test variables we calculated the MSE for all four cases and determined the sum of the Brooklyn, Manhattan, and Williamsburg were the optimal bridges to put the sensors on because their model had the lowest MSE compared to the other three.

Problem 2 - This problem assigns us the task of finding whether or not we can use temperature and precipitation to predict the number of bikes on the bridge on a given day. In order to do this, we created a feature matrix with high temp, low temp, and precipitation. We then split the matrix into a 80-20 (train-split) and then proceeded to train the model using ridge regression. We tested the accuracy of our model by using the R-squared. The R-squared for our model was surprisingly low. So we decided to split the 3 features into three separate feature matrices. We then made similar ridge regression models for the three features and found the R-squared to be underwhelming yet again. In order to get a better understanding and to check if we used the wrong method, we then used the matrix with all three features and created models for Linear, Ridge, and Lasso. However, after extensive testing, we conclude that there is too much variance by R-squared for the weather to determine the number of bikes on a particular day.

Problem 3 - To start this problem, we first manipulated the precipitation to binary (if there was rain append 1 else append 0). Next, we split the data and obtained the train and test X and y. Using the train parameters we constructed a model using logistic regression and plotted the predicted plot (Figure 1). Finally, we used the logistic prediction and test y to construct a confusion chart (Figure 2). From both of these visual representatives, it is clear that we can predict if it is raining given the total number of cycles: There are generally more cyclers when it is not raining than when it is. Looking at figure 1, when the function is at y = 1 (meaning is it raining) the amount of bikers is low (less than 15,000) and when the function is at y = 0 there are more (greater than 15,000) cyclers on the bridge. Additionally, when looking at the confusion chart we note when it is not raining around 78% of the selected days show an increased amount of cyclers. Vice versa when is it raining 80% of the selected days show a decreased amount of cyclers which further supports our claim.

Pranav Sankar
Varun Ravi
April 23rd, 2022

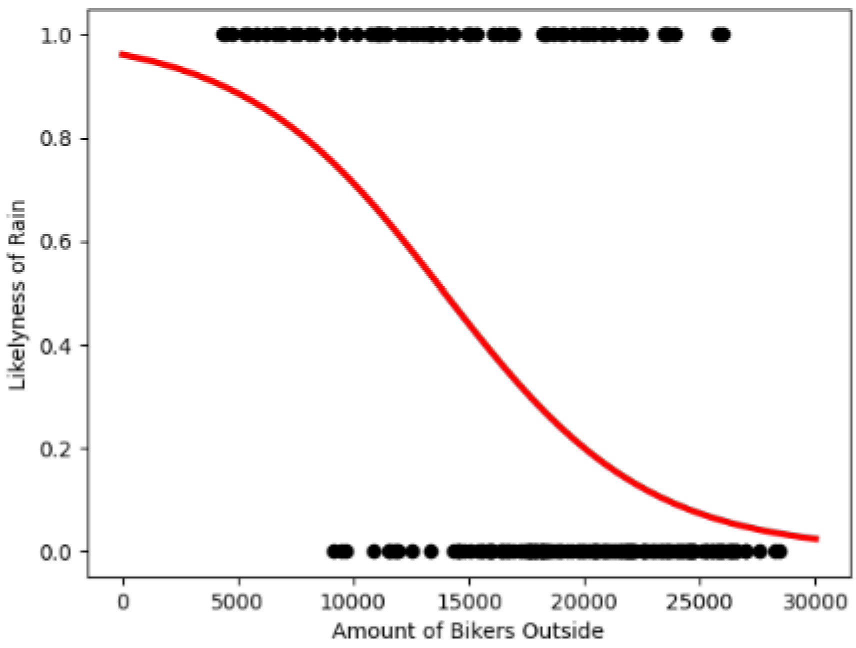

Figure 1

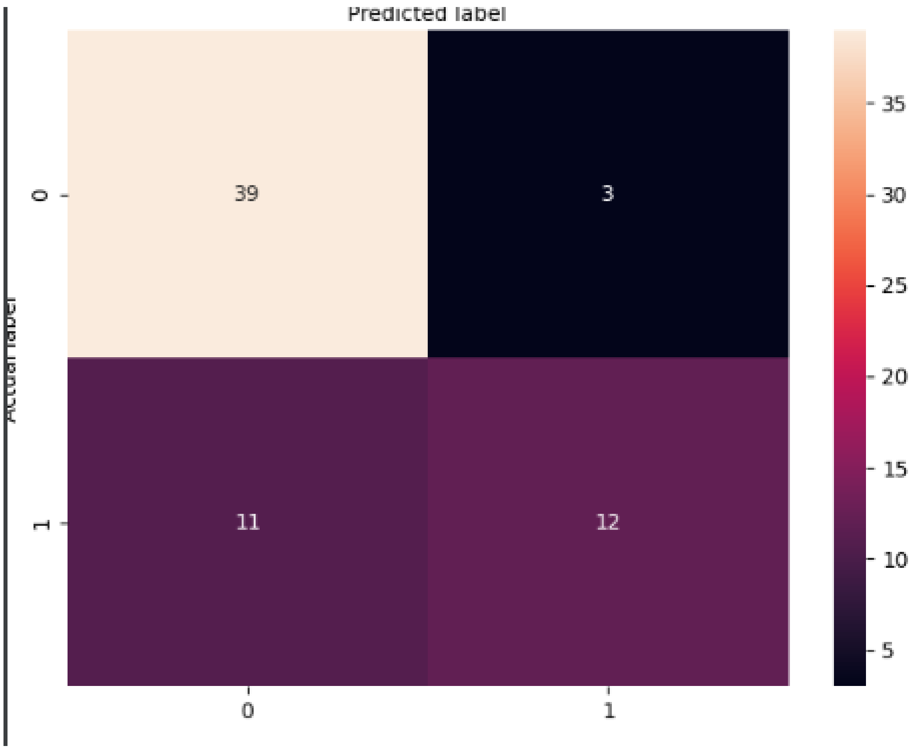

Figure 2