ECE 20875 Mini-Project Report

Team Members: Roshan Sundar, Nick (Ping-Hung) Ko

Team Usernames: rmsundar, ko109

Path 1: Bike traffic in NY city

Dataset:

The dataset that the city has provided overall is meant to represent 2016 bike traffic across various bridges in NY city. The first column is the 'Date' column which is the date of each day the data is collected - which spans a seven month timespan - as well as the 'Day' column which contains the corresponding day name (i.e. 'Monday') to the date. The 'High Temp', 'Low Temp', and 'Precipitation' columns contain the high temperature (F), low temperature (F), and raindrop (in) respectively for each day. These columns are meant to characterize the weather for each day.

The next columns are the 'Bridge' columns for Brooklyn, Manhattan, Williamsburg, and Queensboro bridges. They contain the number of bike riders across it each day. Below is a visual representation of the bike traffic:

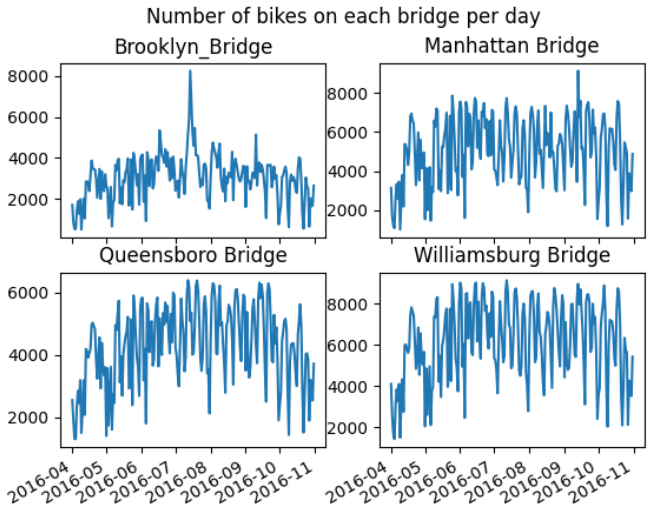

Figure 1: four graphs of the traffic across each bridge over time

Finally there is the 'Total' column, which shows the total (sum of) bike traffic across all four bridges every day.

Analysis:

Question 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?

Methodology:

To facilitate estimation of overall NYC bike traffic across all bridges with a limited budget, our team decided to first determine how each bridge impacts the overall traffic by performing a linear regression to predict the overall NYC traffic situation with traffic across each bridge as inputs, then examine the magnitude of the coefficients in front of each weight to determine the relative importance of each bridge. All training data are normalized before regression such that the magnitude of the coefficient in front of each bridge obtained from the training model is a direct indicator of impactfulness of each bridge.

We seek to compare the coefficients of each bridge from the most accurate model. Therefore, before we train our model, we randomly split our dataset in a 20% test and 80% training data then perform a 5 fold cross-validation to find a model with highest r-square value. Next, we compare the magnitudes of the coefficients to determine which bridge has the least contribution/impact on the overall traffic. Image below shows our result.

```
PS C:\Users\Nick\github\miniproject-f22-Ping-Hung> py .\problem1.py
equation:
0.1984975656821017Brooklyn + 0.306894096882792Manhattan + 0.3352100860380353Williamsburg + 0.2207202578682459Queensboro + 2.3781823990943957e-16

intercept of the model is 2.3781823990943957e-16
r^2 value = 0.9999981023368872
```


Result:

From a model that has r-square value of 0.999, we deduced that the coefficient in front of Brooklyn bridge is the smallest (0.1984) and thereby concluded that it has the least significance in impacting the overall traffic. As a result we deduced that we can simply exclude Brooklyn bridge and put sensors on the rest of three bridges.

Question 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?

methodology

As the officers are interested in studying whether there is a relationship between weather data and the total traffic, the best way to tackle this problem is to actually build a model. Thus, a linear regression with weather information being explanatory variables and total traffic being the subject variable is performed. Like problem one, we split our data randomly into 20% testing and 80% training, then we perform a 5-fold cross validation in search of the model that has the largest r-square value from 5 random 80% - 20% splits of data. The result is shown in the picture below:

```
equation:
0.9144051366721373High Temp + -0.3793737293528356Low Temp + -0.3534067166746457Precipitation + 4.472472208704233e-16

intercept of the model is 4.472472208704233e-16
r^2 value = 0.5758767564057958
```

Result:

As shown in the picture, the r-square value of the best model is about 0.578. This indicates that the regression itself only accounts for about 57.8% of variation in the subject variable (total bike traffic), and that the explanatory variables given (high temp, low temp, and precipitation) does not have a high correlation with the subject variable total traffic

Question 3: Can this data predict what *day* (Monday to Sunday) is today based on the number of bicyclists on the bridges?

Methodology:

The analysis model chosen was a k-nearest-neighbors algorithm. The justification is that the problem is one of classification, since there are seven discrete outcomes. In addition, given the distribution of the bike data, it is not trivial to make assumptions about the data and match it with an existing distribution. Therefore, a kNN is useful since it makes no prior assumptions about the data.

The training data for the kNN would consist of the bike traffic data across each bridge every day, as visually represented in Figure 1. Note, the total number of riders across all four bridges was used as input but was found to produce worse results, and was thus discarded. The output would be the day name, with 0 representing Monday, 1 representing Tuesday, 7 representing Sunday, and so on. The training and testing data would be split into 20% testing and 80% training randomly. The model would predict - from k=1 to k=6 - and compare to the expected output in order to validate performance. The metrics collected would be an accuracy from 0 to 1 and an AUROC score from 0 to 1.

Result:

Due to the randomness of the sampling, the kNN was run 5 times and the best result is the following: k = 5, acc = 0.279070, AUROC = 0.568254. The highest accuracy achieved by the model was 0.279. This shows that there is not a conclusive relation between the number of cyclists riding over the bridges in a given day and the day name. In addition, the AUROC score is very close to 50%, which is a bad result since it means that the model is essentially acting as a random classifier, which adds further weight to the idea that there is not a correlation between the number of cyclists and the day. A possible reason for this behavior could be that other environmental factors such as weather might play a role. It could also be human factors, such as the availability of other forms of transportation on a given day.