George Kalpakis
Path 2
gkalpaki
https://github.com/george6982/miniproject.git

Data Overview

Path two provides data pertaining to the amount of bicycle traffic across various bridges in New York City. Specifically, it provides the number of bikes which crossed a bridge on a certain day, for 214 different days, for four different bridges (Brooklyn, Manhattan, Williamsburg, and Queensboro). The data also provides other information about each day, such as what day of the week it fell on and weather information in order to glean the influence of these variables.

Task 1:

While the city would like to predict bicycle traffic on all four bridges, the project's budget allows for only three sensors. Task 1 asks us to determine which bridge should be left out while maintaining the best picture of overall traffic. In other words, we are to find which bridge can be most accurately predicted given data from the other three.

To solve this problem, four "folds" were considered. Each fold produced a model which was trained with the input containing three features, one for each of the three the bridges, and the output as the data collected for the remaining bridge.

In this analysis, all of the data was used for training, and none was left for validation. This is because the task asks us to find which bridge is optimal over the others, not to find the absolute performance of each model. Additionally, because the model is very simple (three feature linear regression) compared to the amount of data provided, it is very unlikely these models will overfit to the data they were trained with.

The performance of the three bridges to predict the data of the final bridge will indicate which bridge should lack a sensor. The RMSE (root mean squared error) of each model was chosen as the metric to measure performance. The resulting performance of the four models is provided in the table below.

Bridge left out	**RMSE**
Brooklyn	641
Manhattan	809
Queensboro	310
Williamsburg	435

As can be seen, traffic on the Queensboro bridge was most closely predicted by the other three, with a typical error of only a few hundred bikers. Therefore, the bridge which should lack a sensor is the Queensboro bridge.

Task 2:

The next task is to predict the total number of bikers provided with the weather prediction for that day. However, to use weather data alone would be a naive approach, as the day of the week has a large impact on bike traffic. With this in mind, instead of using weather data to directly predict total bike traffic, it was used to predict the *ratio* of total bike traffic to the *average* bike traffic for that day of the week. These average values were tabulated in the excel file which the data was originally presented in. There are three input features related to weather: the daily low temperature, daily high temperature, and precipitation level.

In order to test how well the model will perform on unseen data, 20% of the data was used to validate the models, while the remaining 80% was used for training. This test/train ratio is standard across many applications where the model complexity is small compared to the amount of data provided.

When training a model, we would first like to see if any of these features lack relevance. With a linear regression model, if the coefficient corresponding to a feature is very close to 0, it is likely the output is entirely independent of the feature, in which case it is not even worth including it in the model. A hypothesis test may be used to determine the likelihood that an output is independent of a certain feature. However, to ensure a fair comparison between features (raw temperature data is higher valued than precipitation), the features must first be normalized to a similar scale. Before fitting a model, each feature was normalized to have mean 0 and variance 1. A model was fitted which contained the following coefficients:

Feature	**Coefficient**
High temperature	0.2738
Low temperature	-0.1142
Precipitation level	-0.0996

These values may be thought of as the relative importances of the features. As can be seen, none are negligibly near 0 compared to the others. Therefore, it can be concluded that all three features are important to the prediction of traffic, and none will be thrown away.

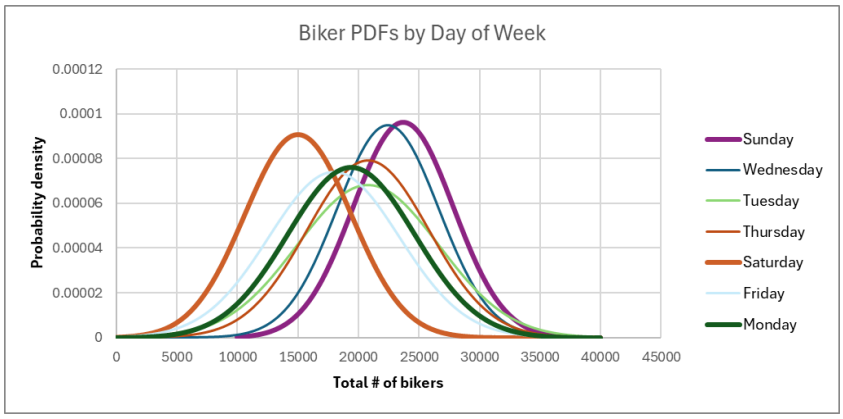
The RMSE of the model on the validation set was found to be 0.164. To map this to the RMSE of the total number of bikers, it depends on the day of the week (because of the previous decision to predict total/day of week average). However, as an average across all days of the week, this model performs with an RMSE of 3041 bikers per day. With an average of 18,500

bikers per day, this corresponds to a typical error of 16%. Indeed, weather forecasts may be used to predict bike traffic with an imperfect, yet quite useful accuracy.

Task 3:

Task 3 asks us to use traffic data to predict which day of the week the data was collected on. To solve this problem, a few methods are available.

First, an attempt was made at using a gaussian mixture model (GMM) based on the total number of bikers per week. The data for each day of the week was isolated, and the mean and variance were calculated for each. These values were used to produce a normal distribution of total bike traffic for each day, the graphs of which are presented below.



Using a GMM to predict day of the week involves taking the observed number of bikers and considering the likelihood that it is on a certain day based on the heights of the bell curves at that location. However, upon inspection of the figure above, it becomes clear that this approach will not be able to make predictions with any sort of confidence. The variance of these bell curves are simply too large, causing them to overlap each other and providing no x location where one bell curve overpowers the others significantly.

This justifies a new approach: k-means clustering, and this time using data from the four bridges individually, not just the total number. If we imagine each data point as a location in a 4D space, if this data happens to appear in seven distinct “clusters”, then it is very probable that each of these clusters corresponds to a day of the week. The algorithm to find these clusters will be a standard k-means algorithm.

Seven clusters were found that provided a local optimum. The following figure provides the percent of the day of the week that was most common for that cluster.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
% points of leading day	31% Thursday	25% Sunday`	27% Wednesday	19% Tuesday	37% Saturday	Unclear	38% Sunday

As can be seen, the results of this approach are again lackluster, as no cluster corresponded to a certain day of the week with even 50% confidence. To add to this, after multiple tests, each random initial position of cluster centers resulted in quite different final results. It is apparent that this method cannot be used to predict the day of the week with confidence.

From these results, it may be concluded that traffic data cannot be used to reliably predict the day of the week because the volume of bike traffic simply undergoes too great of random fluctuations from day to day.