20875 Mini Project (Path 2)

Team

Adam Keith - keithaj
Ashwin Thampi - athampi

Dataset

The dataset is a csv file that contains information on bike traffic over four bridges over the course of seven months (April - October). The four bridges are the Brooklyn, Manhattan, Williamsburg and Queensboro bridges. The dataset contains the traffic over the respective bridges in addition to the total across all four bridges. Additionally, the high and low temperature - in degrees Fahrenheit - is given along with the precipitation - in inches.

Methods

Question 1: Find the three bridges that most accurately predict the total traffic.

For question one, the dates are assigned to x-values and the total traffic is assigned to the y-values. Next, an array of three bridge combinations is initialized. Then for each combination of three bridges, a model is trained and tested using linear regression and compared with the total data. An r squared value is calculated for each combination, with the highest being the bridge combination that best predicts overall traffic.

We chose to employ linear regression as it allows for a very simple comparison between the total traffic data and the predicted total traffic data using a combination of three bridges as the training data. This allows us to easily compare the results of each combination with that of the actual traffic data to determine the best predictor.

Question 2: Can next day's weather forecast be used to predict high traffic days?

The high temp, low temp and precipitation data are concatenated and used as the x-training data. This is then fitted to a model and compared with the total number of cyclists on the respective days using linear regression. The r-squared value across the y-values and predicted y-values is calculated and evaluated to determine if there is a correlation between the weather and the number of cyclists.

The reasoning for choosing linear regression is because we are simply comparing the relationship between two subsets of data. This allows us to visualize the similarities between the predicted traffic and actual traffic using the high temp, low temp and precipitation as parameters. This allows us to then use the r-squared value to identify a possible correlation in the data as stated above.

Question 3: Can you predict what day it is based on the number of cyclists?

The dataset is constructed by concatenating the number of cyclists per bridge. This data is then scaled using the standard scaler function. Then using k means clustering with a k value of seven - one for each day of the week - the predicted cluster is created with the k means and the scaled data and then mapped onto the corresponding day of the week.

The reasoning for selecting k-means clustering is that it can accurately group or ‘cluster’ the means so that patterns betweens the data of the same days of the weeks can be used to determine the current day of the week with reasonable accuracy.

Results

Question 1

Our findings showed that the Brooklyn, Manhattan and Williamsburg Bridge combination was the most indicative of the total traffic for any given day, as the r-squared score was closest to one compared to the other three combinations.

Bridge Combinations			**R-Squared**
Brooklyn	Manhattan	Williamsburg	0.9975
Brooklyn	Manhattan	Queensboro	0.9961
Brooklyn	Williamsburg	Queensboro	0.9856
Manhattan	Williamsburg	Queensboro	0.9830

Figure 1: R-Squared Per Bridge Combination

As shown in the table above, all four combinations of bridges are fairly accurate predictors of the total traffic, as all four r-squared values are within 0.2 of 1. However, since the Brooklyn, Manhattan, Williamsburg combinations yields the r-squared value closest to 1, it is the best predictor of the total traffic on any given day.

Question 2

The weather data can be used to predict the number of cyclists on a given day. When using the linear regression method illustrated in the methods section above, the test yields an r-squared value of 0.5770. An r-squared value between 0.5 and 1 illustrates a relationship between datasets. Therefore, we can say that the weather data is an accurate predictor of the overall traffic for any given day.

Question 3

We employed the k-means clustering illustrated in the above methods section. We generated the following k-means values for each day of the week.

Day	**K-Means**			
	Brooklyn	**Manhattan**	**Queensboro**	**Williamsburg**
Monday	3236.79	5915.36	4836.79	6988.00
Tuesday	1842.84	3333.32	2948.35	4140.90
Wednesday	4037.94	7099.98	5665.23	8338.45
Thursday	3527.10	4196.0	5818.2	8110.1
Friday	1013.05	1756.63	1863.74	2471.53
Saturday	2841.55	4714.05	3899.2	5551.56
Sunday	5770.29	5176.86	4573.57	6027.57

Figure 2: K-Means for bridge traffic by day of the week

Using the k-means values, the best match for the most recent data point is Saturday. Therefore, we can say that the day is Saturday.