New York Bike Traffic Analysis with Data Science

Evan Zhang

Objectives

With data on bike traffic across several bridges in New York City, along with other supplementary data on weather, day, etc., on a few core questions. The data is formatted with day-to-day data in rows, while the columns have data on the weekday, the day-month date, the high and low temperatures of the day, the precipitation (rainfall in height measured by inches), bicycle traffic count on four different bridges, and the total traffic on all four of those bridges. With this data, the primary method of prediction will be through regression and feature engineering, as the primary objective is to determine how reasonably well certain parameters could explain the following questions.

I. Which Three Bridges To Install Sensors?

In this question, the idea is to capture which of the bridges will give the greatest usage due to budget constraints in installing sensors to track overall traffic. This primarily means, we want to see which set of three bridges can we most reliably count on to be able to predict the remaining bridge's traffic, as the total traffic is already captured by the set of three.

III. Do Weather Conditions and Bicycle Traffic Correlate?

In this question, the idea is to predict under which weather predictions can we reasonably say we should allocate more police resources, as it could potentially be an explanatory factor in existing traffic on a given day. Essentially, can we use rain height (in inches), the high temperature, and the low temperature of a given day to accurately assess the traffic numbers on such a day?

III. Can We Predict The Weekday From Bicycle Traffic?

In this question, the idea is to ask whether one could reasonably figure out the day of the week based on the traffic data available. With this question, we are asking if there can be any conclusions drawn by seeing the total sum of bicycle traffic on the bridges on a given day with only this knowledge, be able to determine the day of the week.

I. Installation of Bridge Sensors

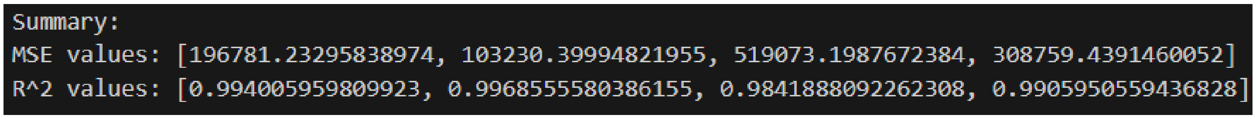
1.1. Procedure and Justification

The procedure begins by obtaining the pertinent information to the question, which essentially asks which three bridges we could reasonably say best predict the true bicycle traffic numbers on a given day. To do this, we will compose features comprising the Brooklyn Bridge, Manhattan Bridge, Queensboro Bridge, and Williamsburg Bridge. These features will target the total bicycle traffic on a given day. To make a comparison of three-bridge combinations, we will take the four unique combinations of the bridge traffic data Each feature matrix will contain data from three bridges, and a column of ones to add an intercept value. Afterward, we will create four different ridge regression analyses that automatically select the best alpha using the Scikit-learn RidgeCV, which will then cross-compare the groups' r^2 and MSE values against each other to make the best statement on which three bridges will make the most sense to place sensors on. The MSE and r^2 values will be stored in vectors for each three-combination of bridges. The feature groups matrices and respective metrics in the vectors follow this order:

1. Brooklyn, Manhattan, Queensboro
2. Brooklyn, Manhattan, Williamsburg
3. Brooklyn, Queensboro, Williamsburg
4. Manhattan, Queensboro, Williamsburg

We can reasonably expect this approach to work, because essentially we are taking regression models that find the best regularization parameter, and then compare which set of bridges are best able to predict the total bicycle traffic. It follows that if a set of three bridges can best account for the variability of the true total traffic, it would make sense to use the bridges defined in that set as the three bridges to install sensors onto. It also makes sense that the models will return a very high r^2 value, as the bridges are already accounting for 3/4ths of the total sum of bicycle traffic, so we will end up comparing very slightly marginally different r^2 values and picking the best out of those.

1.2. Results and Analysis



```
Summary:
MSE values: [196781.23295838974, 103230.39994821955, 519073.1987672384, 308759.4391460052]
R^2 values: [0.994005959809923, 0.9968555580386155, 0.9841888092262308, 0.9905950559436828]
```

Figure 1.1. - MSE and r^2 vectors

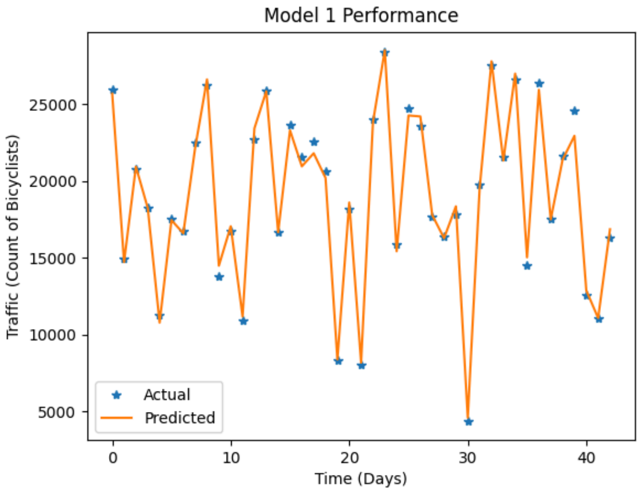


Figure 1.2. - Group 1 Regression Plot

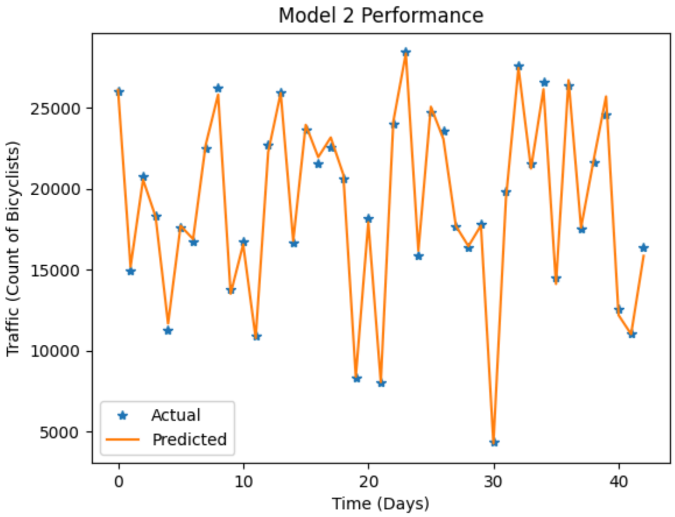


Figure 1.3. - Group 2 Regression Plot

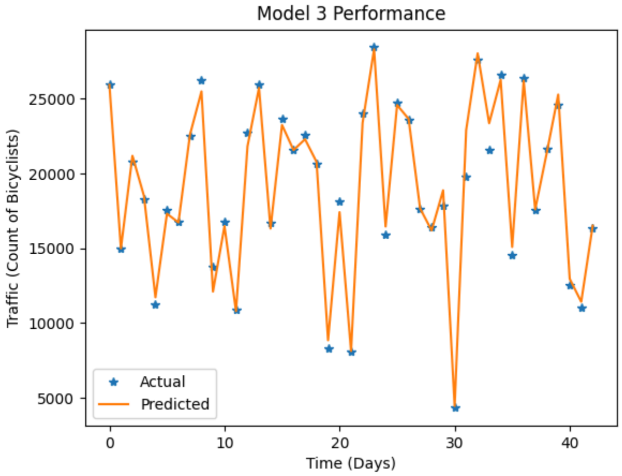


Figure 1.4. - Group 3 Regression Plot

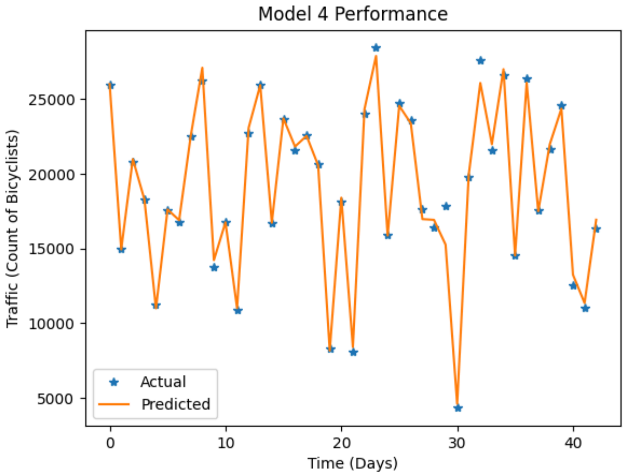


Figure 1.5. - Group 4 Regression Plot

From these figures that the regression models produced, we can deduce that the bridges are good explanatory features of the total traffic. This logically makes sense, as taking three bridges out of four will naturally be able to account for a lot of the values in the total traffic, as they are three-fourths of the total sum. However, we are looking for metrics that might suggest a group performs slightly better in the prediction of total traffic in comparison to other groups, which involves analyzing the r^2 and MSE values. From the vectors, we can see that Group 2 provides the lowest MSE at 103230.39994821955 and the highest r^2 at 0.9968555580386155. Because Group 2 provides metrics that suggest the best prediction of the true traffic count of a given day, we can reasonably say that the bridge traffic counts included in Group 2 (The Brooklyn, Manhattan, and Williamsburg bridges) are the most viable ones to measure to be able to predict the traffic of a given day. This suggests that we are best off installing sensors on the Brooklyn, Manhattan, and Williamsburg bridges for the best use of the sensors.

II. Weather Conditions as Predictors of Traffic

2.1. Procedure and Justification

The procedure begins by obtaining the pertinent information to the question, which asks under what weather conditions may cause higher traffic that could help understand optimal deployments of city traffic enforcement. The features comprising the feature matrix are the High Temperature, Low Temperature, and Precipiation which will target the total traffic on a given day. Then, we will fit it via the ridge regression model to get parameters to plot against the true bicycle traffic count data and find the mean square error and coefficient of determination values to determine the validity of the original premise of the question. Using cross-validation, we can find the best parameter of regularization (alpha) to use for the model, and we will then use that model's metrics to determine the validity of weather conditions as explanatory variables of a given day's bicycle traffic.

The method of approach here is straightforward, as we are looking to see if we can predict the total traffic of the bridges just by using the low temperature, high temperature, and precipitation conditions as valid explanatory values. Ridge regression is a straightforward way to be able to capture how fit an explanatory variable weather conditions are. With alpha being cross-validated, we can assume that we are getting the best ratio of regularization in the regression, such that we can say with reasonable certainty that our metrics of r^2 and the MSE values are the best available from ridge regression. As an aside, this approach makes sense if you think about whether or not a person would voluntarily choose to go into 'bad' weather more often than 'good' weather, which for more people would likely prefer the latter.

2.2. Results and Analysis

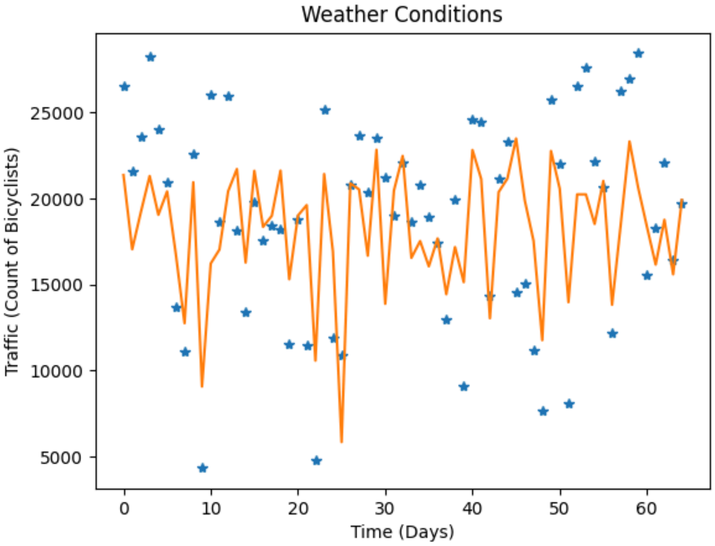


Figure 2.1. - Regression Plot Using Weather as Explanatory Variables

The **MSE** and **r^2** value we obtain from this regression model is 13809253.998073 and 0.600864 respectively at an optimal alpha level of **0.01**. With a ~60.01% variability in the target traffic values able to be explained by the model, then it is still able to make a decent enough guess about whether it can predict the true value of the bicycle traffic on a given day, so it makes sense to use it as a possible measure of how to spend police resources in NYC. However, the model metrics also suggest that purely relying on the weather conditions of high temperature, low temperature, and precipitation as the explanatory variables is not the best, as the model provided could not account for close to 40% of the variability of the true traffic values of a day. However, that doesn't mean it is completely unable to be used to predict traffic on a given day given the weather conditions. From a common knowledge standpoint, it is a valuable insight to see that generally if weather conditions are not as good, like high precipitation or extreme high/low temperatures, then people would generally choose to stay inside. That is not the full story of why a given person might choose a given day to ride their bicycle, as there are factors that might force a group of people to contribute to bicycle traffic counts. In that way, it does make some sense as to why the model cannot explain more variability of bicycle traffic on a given day. Overall, if the data could provide more weather conditions as explanatory variables, the model could be fine-tuned to a higher level, as the current condition parameters are not able to capture the full story of the bicycle traffic on a given day due to weather conditions.

III. Bicycle Traffic as a Predictor of Weekday

3.1. Procedure and Justification

This procedure begins with understanding a fundamental difference in data used in the question. We are using total bicycle traffic to see if we can accurately guess the day of the week. A quirk of the data is the discrete nature of a day of the week data, because it is categorical data, whether we express it as a numerical value or not. Because of the discrete nature of the day of the week, we are not able to use standard ridge regression as a potential model of the accuracy of prediction. The methodology used for this purpose involves a systematic approach integrating logistic regression and neural networks. In the initial steps, the dataset is preprocessed, with numerical values representing days of the week one-hot encoded to create binary vectors, acknowledging the categorical nature of the data. Logistic regression is then chosen as a baseline model due to its simplicity and efficiency in handling multi-class classification tasks. It is trained and evaluated on the dataset without the one-hot encoded labels, providing a straightforward interpretation of the relationship between input features and class probabilities. Subsequently, because there are likely holes in the explanation of the logistic regression, a neural network is constructed with an architecture designed for multi-class classification, now using one-hot encoding to represent the categorical days of the week. The neural network is compiled and trained using an appropriate loss function and optimizer, capturing non-linear relationships in the data. The use of one-hot encoding in both models ensures an accurate representation of the categorical nature of the days, which is particularly crucial for the softmax activation in the neural network's output layer.

The reasons for this approach lie in the strengths of both logistic regression by itself and a neural network. Multinomial logistic regression, as a baseline, provides simplicity and interpretability, while the neural network introduces complexity to capture intricate patterns. The choice of one-hot encoding aligns with the categorical nature of the days of the week, preserving the integrity of the data and enabling effective training of both models. By using both logistic regression and neural networks within a coherent methodology, this approach aims to compare the strengths of each model in a complementary manner to see if there is any ability for accurate predictions in the context of multi-class day-of-the-week classification.

3.2. Results and Analysis

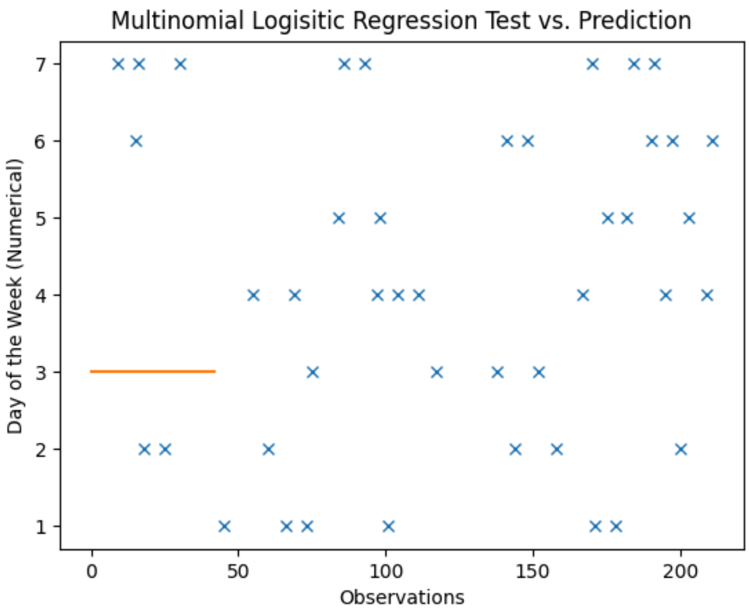
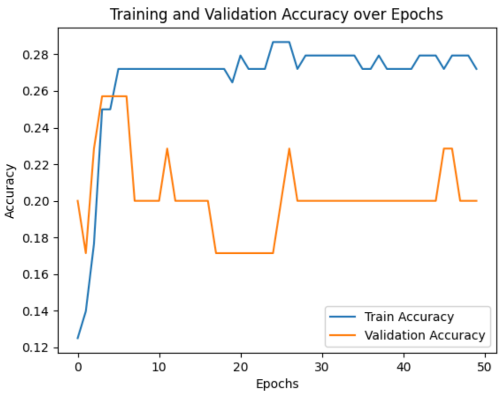
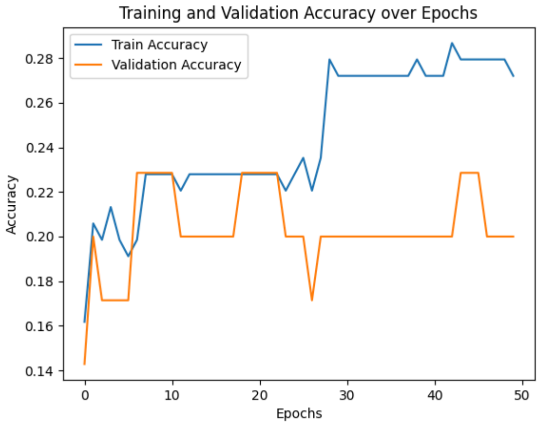
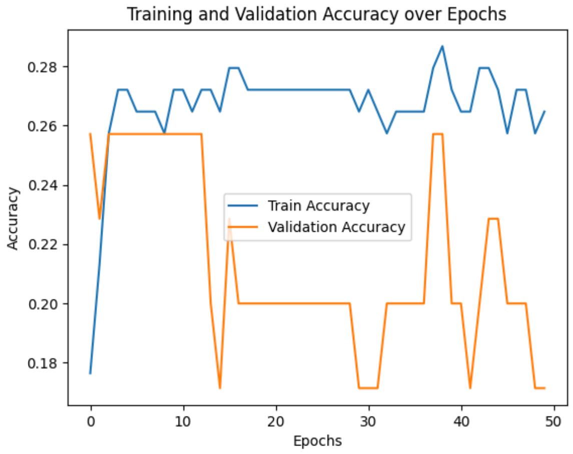
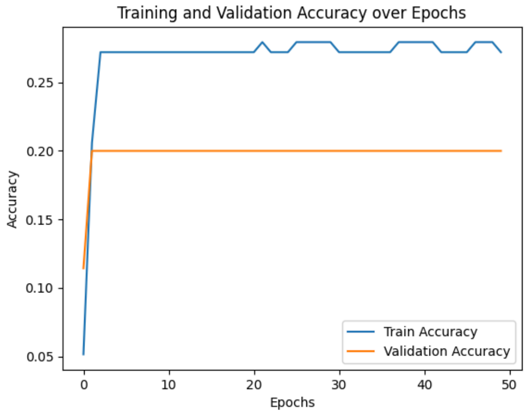


Figure 3.1. Multinomial Logistic Regression Baseline

Run	Accuracy Score	Training and Validation Accuracy Plot
1	0.27906976744186046	

2	0.2558139534883721	
3	0.23255813953488372	
4	0.27906976744186046	

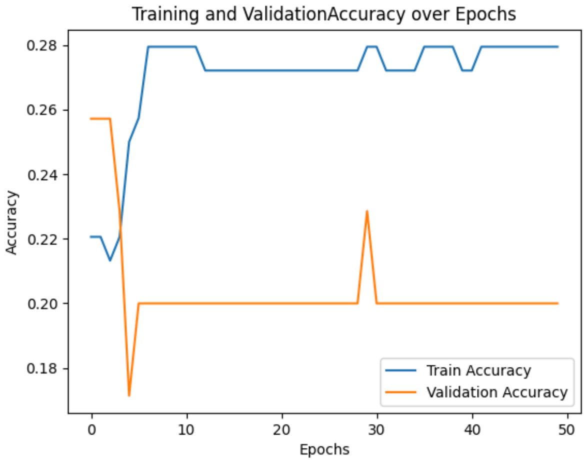
5	0.3023255813953488	

Figure 3.2. Five Trials of Neural Network Returns

The results of the predictive models, employing logistic regression and a neural network for day-of-the-week classification, are anticipated to provide insights into the effectiveness of these methodologies for the given task. The logistic regression model provided an accuracy score of **0.093023**, which means that logistic regression on one run is nowhere near enough to give accurate predictions of the weekday. Figure 3.1. Also makes apparent that the logistic regression has no oscillation beyond three, which means it is not able to capture anything meaningful enough to give more varied predictions. On the contrary, the neural network, with its capacity for non-linear learning, provided throughout a few runs accuracy scores of **0.2-0.3**, which is higher than that of the logistic regression model, but it still fails to give a good prediction with ranges like that. The evaluation metric of accuracy quantified the models' performances on the test set. With both models provided, it can be seen that total bicycle traffic has a difficult time predicting the day of the week. The reasons for this could be for many reasons, but a few could be in regards to how the design is fit and likely the data involved. To make predictions, the models would've needed to see that the data provided lots of patterns in traffic that could make a reasonable guess of the true day of the week, however, the data set provided ultimately is not that large, which could be masking any true patterns a reason for such low accuracy in predictions. Another reason could be that the models could do with more optimization in choices of parameters such as experimenting with different activators of the perceptrons, or need additional layers into the neural network to help tune the predictions. As a result, I can only cannot conclude that total traffic data can predict days of the week accurately and reliably, through the use of logistic regression and neural network models.