Subway Ridership Prediction Model

By Andrew Henkel

Abstract

I trained a multilinear regression model in PyTorch to determine the number of people riding the New York City subway given a set of weather conditions. The resulting model is composed of five input features (precipitation levels, snowfall, snow depth, average air temperature, and average wind speed), six hidden layers, and one output. It was trained for a thousand epochs (with a batch size of ten) using the Adam optimizer and the Mean Squared Error loss function. After training, I used my model to predict the ridership for ten sets of weather conditions. In the process, it yielded several important findings, including an apparent direct relationship between temperature and ridership, as well as an indirect relationship between precipitation and ridership. Both these findings disprove my initial hypothesis that ridership would actually increase when the weather is more severe. (However, there are some cases that actually fit my initial hypothesis, such as the direct relationship between wind speed and ridership.) I provide several possibilities as to why my initial assumptions were wrong, including the transition to working from home that was spurred on by the pandemic as well as the tendency for tourists to visit and travel in the city when the weather is better. I also describe several reasons why the predictions currently yielded by the model could be inaccurate, including the high MSE during training and the very large ridership numbers that exceed those recorded by the MTA. I then describe what might be done to remedy these issues.

Full Paper

The goal of my midterm project was to determine if the weather conditions in New York City affect the daily ridership of the subway. My initial assumption was that New Yorkers would generally prefer to walk to their destinations in nicer weather. Because of this, my hypothesis was that the subways would

be less crowded on days with decent/warm weather and more crowded on days with worse/colder weather.

To test this hypothesis, I used two datasets. The first was the daily ridership from the MTA starting from 2020, from which I used both the “Date” column and the “Subways: Total Estimated Ridership” (*MTA Daily Ridership Data: Beginning 2020 | State of New York*). The second was the daily summaries of climate information from the NCEI's Central Park station starting from March 1st, 2020 and ending at February 2nd, 2024 (*Daily Summaries Station Details: NY CITY CENTRAL PARK, NY US, GHCND:USW00094728 | Climate Data Online (CDO) | National Climatic Data Center (NCDC)*). I used the following columns from the NCEI dataset: “DATE” (for the date), “PRCP” (precipitation in inches), “SNOW” (snowfall in inches), “SNWD” (snow depth in inches), “TAVG” (average air temperature in Fahrenheit), and “AWND” (average wind speed in miles per hour). After downloading the datasets as CSV files and importing them as Pandas dataframes, I removed all rows containing dates before January 1st, 2022 to prevent the results from being affected by the low ridership from the pandemic. I also ensured all of the rows in the MTA dataset had corresponding rows in the NCEI dataset. Finally, I removed the date columns from both datasets and then treated the NCEI dataset as the “X” dataset and the MTA dataset as the “y” dataset.

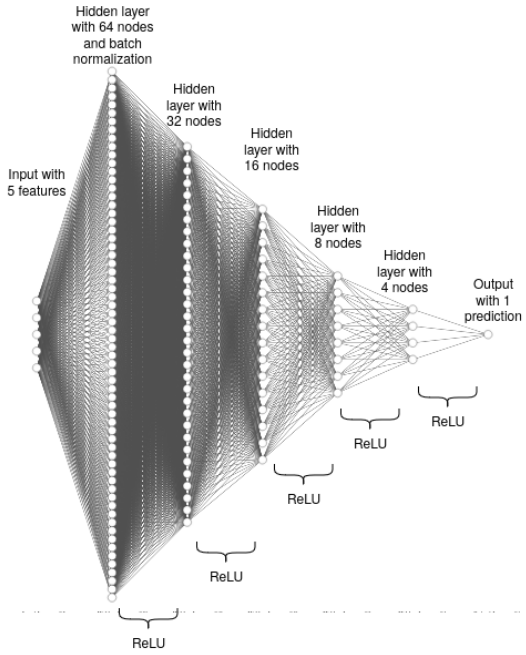

Figure 1: Diagram showing the layers of the multilinear regression model. Created using NN-SVG (by Alex Lenail) and diagrams.net.

I then used the X and y datasets to train a multilinear regression model consisting of five features, six hidden layers, and one output prediction. During the training process, the values in X where treated as the input features and the values in y as the outputted results. The model used batch normalization, ReLU activation, and the HE uniform initializer. During

the training process, the Mean Squared Error loss function was used along with the Adam optimizer (with a learning rate of 0.01). In addition, 70% of X and y were used for the actual training after being converted into PyTorch tensors, and 30% of these datasets were used for the testing. The model was trained for 1000 epochs in total with a batch size of 10. The training process was configured so that before the end of each epoch, the mean squared error will be calculated, and if it is lower than the lowest ("best") MSE calculated up until this point, then the weight calculated for the current epoch will be used for the next epoch. If not, then the weights associated with the best MSE will be restored. The final best MSE was 655619915776.0, and its root was 809703.6.

(Here are brief descriptions of the algorithms used: Batch normalization resets the previous layer so that the next layer can process its output more efficiently (Barabasi). The rectified linear activation function, or ReLU, reduces the chances of a vanishing gradient (Barabasi). The HE uniform initializer uses a uniform distribution to generate samples of values (Keras). The initial parameters of the model are then set to these values (Brownlee). The MSE is the average of the squares of the distances between the correct answers and the predicted answers (Barabasi). The Adam optimizer combines the ability to keep track of decaying gradients and decaying squared gradients to quickly converge on the right parameters in certain use cases (Barabasi).)

After the training process, I wrote up ten hypothetical weather conditions and used the model to predict the ridership on days with those conditions. During the process, I observed that the most significant factor affecting the ridership numbers was the average air temperature, with the ridership levels rising as the temperature rises. This generally occurs even if there is more precipitation or rainfall. When the temperature is constant, however, the next most significant factor that I observed is the amount of precipitation, which actually has an inverse relationship on the ridership levels. The third most influential factor is the average wind speed, which also has a direct relationship with the ridership numbers. The results are displayed below.

Table 1: Predicted ridership (people) based on weather conditions. It is sorted by temperature to convey the relationship between the temperature and the ridership numbers.

Precipitation	Snowfall	Snow depth	Avg. temp	Avg. wind speed	People
3	4	4	0	1	3307513
3	6	6	15	5	3463060
3	4	4	20	5	3684506
4	7	5	29	0	4733171
5	0	0	40	10	6320541
5	0	0	40	5	6139473
7	0	0	50	15.5	7802017
1	0	0	60	15	8531844
0	0	0	70	1	9508300
5	0	0	70	5	9316093

Overall, the air temperature appears to have a much larger impact on the ridership levels than the other factors, and since my initial hypothesis would have indicated that the ridership would increase as the temperature decreases, it has essentially been disproven by these findings. Additionally, my hypothesis would have indicated that the ridership would increase as the precipitation levels increase. The inverse appears to be true, again disproving my hypothesis. (In fact, the only noticeable trend that follows my hypothesis is that ridership increases as wind speed increases.) I believe that this can be explained to a certain extent. I had assumed that more people would prefer to take the subway when the weather conditions were worse, but I failed to take into account two aspects of the city. First is the fact that during the pandemic, many people have adjusted to working from home (Minkin). Because of this, some of the people who had commuted in worse weather in the past can now opt to stay home. Second

is that it can be assumed that tourists are more likely to travel in nicer weather, especially if they want to partake in some outdoor activity. In a circumstance like this, it doesn't matter how you get to your destination if the conditions are going to be abysmal when you get there. Because of this, more tourists would, in theory, commute when the weather is better.

In spite of these observations, there are still certain reasons to doubt the accuracy of the results. First is the fact that the numbers yielded by the model are absurdly large. In 2022, the MTA reported that the subway had a daily ridership of approximately 3.2 million ("Subway and Bus Ridership for 2022"). All of the numbers yielded by Table 1 surpass this amount, with the last entry in the table having a ridership of almost triple the number reported by the MTA. There could be several reasons why the numbers reported by the model are so large, and by extension, why the "best" mean squared error was so large as well. The first possibility is that either the learning rate of 0.01 or the duration of the training process (1000 epochs) isn't sufficient and therefore needs to be higher. Alternatively, the process of selecting the "best" MSE is flawed, and the model could yield better predictions when the model is simply trained to improve its current MSE, not discard it in favor of a better MSE after each training epoch. (I attempted this, but the MSE actually kept on increasing over time. Perhaps this trend would have changed if the learning rate or number of epochs were higher, but then the training process would have taken too long.) Another possibility is that the optimizer and the loss function are not suitable for this scenario. After all, New York City is full of people with different backgrounds, cultures, and agendas. Some people might have a need to travel on a specific day of the week, month, or year, regardless of the weather. (A typical example of this is a religious holiday.) For this reason, a loss function based on Mean Squared Error might be too sensitive to these

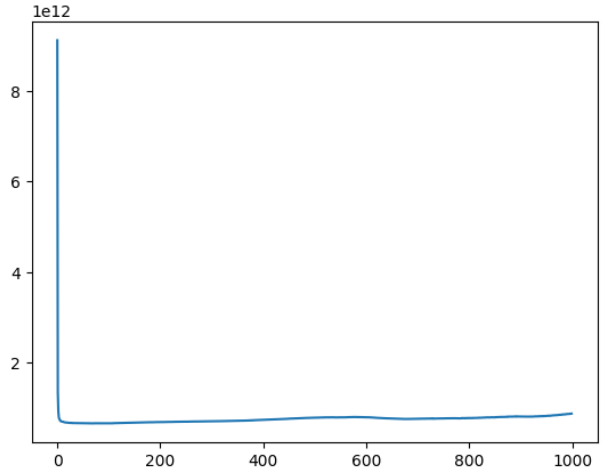

Figure 2: Graph showing MSE over the duration of the training process.

sorts of outliers. If I had more time, I would train a new model using a different loss function that might not be as sensitive to outliers, such as HuberLoss (*HuberLoss — PyTorch 2.2 Documentation*). However, even with these changes, it might still be very difficult to find a definitive relationship between the weather and the ridership numbers without focusing on a specific group of people whose travel habits aren't affected by these sorts of events.

To conclude, this project has yielded two important pieces of information. First, it provided some trends between the weather and the ridership numbers. These trends generally disprove my hypothesis and actually show that ridership increases when there's better weather instead of decreasing (with a few exceptions, such as the average wind speed). Second, it has introduced several possibilities why the training process might not be suitable for this kind of dataset (such as the loss function being too sensitive to seemingly random fluctuations in ridership). Therefore, if the predictions yielded by the model are actually accurate, then this can be of great benefit to the MTA in the hypothetical event that they actually use it. On the other hand, if they are not accurate, then there are several different ways I can attempt to remedy this situation.

References and Publications Used

Barabasi, Istvan. *Deep Learning Fundamentals and Intro to DL Frameworks Chapter-2*.

---. *Intro to Deep Learning Chapter-1*.

Brownlee, Jason. *Weight Initialization for Deep Learning Neural Networks - MachineLearningMastery.Com*. https://machinelearningmastery.com/weight-initialization-for-deep-learning-neural-networks/.

Daily Summaries Station Details: NY CITY CENTRAL PARK, NY US, GHCND:USW00094728 | Climate Data Online (CDO) | National Climatic Data Center (NCDC). https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail. Accessed 26 Mar. 2024.

HuberLoss — PyTorch 2.2 Documentation.

https://pytorch.org/docs/stable/generated/torch.nn.HuberLoss.html. Accessed 1 Apr. 2024.

Keras. *Keras Documentation: Layer Weight Initializers*. https://keras.io/api/layers/initializers/. Accessed 4 Apr. 2024.

Minkin, Kim Parker, Juliana Menasce Horowitz and Rachel. “COVID-19 Pandemic Continues To Reshape Work in America.” *Pew Research Center’s Social & Demographic Trends Project*, 16 Feb. 2022, https://www.pewresearch.org/social-trends/2022/02/16/covid-19-pandemic-continues-to-reshape-work-in-america/.

MTA Daily Ridership Data: Beginning 2020 | State of New York.

https://data.ny.gov/Transportation/MTA-Daily-Ridership-Data-Beginning-2020/vxuj-8kew/about_data. Accessed 26 Mar. 2024.

“Subway and Bus Ridership for 2022.” *MTA*,

https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2022. Accessed 26 Mar. 2024.