New York City MTA Ridership Predictor

Jason Chan
Courant, New York University
New York City, United States

Reggie Gomez
Courant, New York University
New York City, United States

***Abstract*—**

New York City's Metropolitan Transportation Administration (MTA) system has been plagued with overcrowding. However, what are some of the external factors that affect ridership in the system? We developed an application that ingests transit and weather data from the between beginning 2015 to end 2018. Utilizing weather data, we built a predictive model that attempted to predict weekly ridership. However, we have found the model to be unreliable. We also discovered the transit data to be inconsistent. In the end, we concluded that weather might not be the best predictor of the ridership of the New York City MTA.

Keywords— predictive analytics, New York City, MTA, ridership, predictor, weather, machine learning, time series, turnstile, fare

I. Introduction

New York City's Metropolitan Transportation Administration (MTA) system has been plagued with overcrowding. However, what are some of the external factors that affect ridership in the system? For our research, we want to be able to predict New York City's MTA train based ridership with a 7-day weather forecast. Do more people ride the MTA when the weather is warmer? Alternatively, will more people ride the MTA when the weather is cooler? Thus, we want to be able to determine how ridership to the NYC MTA system changes based on weather conditions. We collected daily ridership data from the MTA from 2015 to 2018. We also collected historical daily weather data from NOAA. With those two datasets, we built an application that can use a weather forecast such as one from weather.com and predict the ridership for the next 7 days.

II. Motivation

The primary motivation for our research is to develop actionable insights for both the MTA's management and the riders of the MTA system.

One insight that can be derived from the data and application is the upcoming ridership amount of the system. This insight can be helpful to riders and the MTA. Another insight is that we can also observe seasonality or trends to the MTA readership over time.

These insights can result in decision making and resource optimization for the MTA. With this application, the MTA can benefit by knowing the upcoming forecast to whether or not to increase or decrease the number of trains that need to be online. MTA's management can shift staff to the popular stations on popular days. New York City Police Department can distribute their officers to stations more effectively if there is an anticipated ridership increase. With this information, the riders can change up their travel decisions such as leave earlier or later in the day to avoid the congestion. Riders may even consider alternate forms of transit such as cabs or bicycles. Riders may even consider alternate forms of transit such as cabs or bicycles.

III. Related Work

Research into how weather affects transportation patterns is not a new phenomenon. From commuters in cars to public transit, research into transit is very wide in its breadth.

Through survey research, Belgium researchers have concluded that weather affects people's decision on how they travel with automobiles [1]. The researchers found that people changed their decisions on how to commute on rainy days when driving by car. While the study was focused on driving behavior, the study mentions the people who participated in the study included individuals who were not drivers.

In Chicago, researchers looked at how weather affects automobile accidents, flight delays, and ridership of the CTA [2]. These researchers concluded a small decline of 3-5% in public transit ridership during rainy days. While a decline in ridership was included in this study, this study focused on motor vehicles and flight accidents related to adverse weather in the Chicago area.

A team of researchers in New York City looked at hourly ridership changes over the course of a year for the MTA for different weather patterns and station design [3]. The focus of this paper was on station design and how lawmakers can develop policies that can improve ridership and mitigate weather.

Lastly, researches from Washington state observed how weather and seasons affect ridership of the Pierce county bus system [4]. This paper looked at how wind, rain, snow, and temperature affected ridership numbers for different seasons. They have concluded that adverse weather affects system ridership negatively. They have also found some seasonal variation. This study is the closest to what we are trying to research into for the New York City MTA.

One of the major shortcomings of past research relates to the small datasets. Both researchers from NYC and Washington

admits the datasets have been small and did not include 24hr coverage. Our application addressed both as our datasets were large and included 24hr coverage.

IV. Datasets

To predict ridership, we needed both transit ridership data and weather data. The three primary data sources we chose were: MTA turnstile data, NOAA weather data, and MTA fare data. The entire dataset included data from 2010 to 2018. Original total dataset size was ~10gb. However, we only used data from the beginning of 2015 to the end of 2018 due to changes in reporting before mid-2014 by the MTA. Total size utilized for our dataset was 7gb.

A. MTA Turnstile Data

The first data source we chose was the MTA turnstile data. For this dataset, a turn of the turnstile is a representation of a rider entering or exiting the MTA system. Thus, a daily total of the turns should represent the number of people who used the MTA system. MTA's turnstile data report is produced on a weekly basis for each turnstile for each station. The MTA collects turnstile data every single day at 4-hour intervals. The actual value of the turnstile is the cumulative turns count. The schema for the turnstile dataset in this report: MTA Unit Info, Turnstile Identification Info, Station, Lines, Date, Time, Entries, and Exits. Each weekly report is available on the MTA website as a CSV/TXT file[5]. The information that we are interested in is the Turnstile ID, Station, Date, Time, and Entries.

B. Weather Data

For our weather data, we used daily weather summaries from the National Oceanic and Atmospheric Administration (NOAA). The data from NOAA included a large amount of data types, but the ones we were interested in the most were: average wind, precipitation in inches, snow in inches, and average daily temperatures. NOAA's daily summaries also included weather readings from 63 weather stations around the New York City area. Weather summaries generated on a daily interval. Daily weather summaries can be requested via the NOAA website [6]. The schema for the NOAA dataset are: STATION, NAME, DATE, AWND, DAPR, DASF, MDPR, MDSF, PGTM, PRCP, SNOW, SNWD, TAVG, TMAX, TMIN, TOBS, TSUN, WDF2, WDF5, WSF2, WSF5, WT01, WT02, WT03, WT04, WT05, WT06, WT08, WT09, WT11. The values that were relevant to us were: Station, Name, Date, Average Wind Speed (AWND), total precipitation in inches(PRCP), total snow in inches (SNOW) and average daily temperature (TAVG).

C. MTA Fare Data

The third dataset we collected was the MTA Fare data. Fare data was similar to the turnstile data. However, unlike the turnstile report that reported cumulative turns of a turnstile, the fare report showed total types of fares per week that were swiped through the turnstiles or readers. Both are indicators of the number of riders that have gone through the system. Data included in the fare reports are station id, station and the twenty-six types of fares collected by the MTA. The Fare report is generated on a weekly basis. Each weekly report is available on the MTA website as a CSV file [7]. The schema for the Fare dataset: Remote ID, Station name, FF, SEN/DIS, 7-D AFAS UNL, 30-D AFAS/RMF UNL, JOINT RR TKT, 7-D UNL 30-D UNL, 7D-XBUS PASS, TCMC, RF 2 TRIP, RR UNL NO TRADE, TCMC ANNUAL MC, MR EZPAY EXP, MR EZPAY, UNL PATH 2-T, AIRTRAIN FF, AIRTRAIN 30-D, AIRTRAIN 10-T, AIRTRAIN MTHLY, STUDENTS, NICE 2-T, CUNY-120, CUNY-60,FF VALUE,FF 7-DAY,FF 30-DAY. Besides the Remote ID and Station name, all the other columns in the Fare data set represents all the different fare types the New York MTA collects.

D. Final Dataset

After processing and joining our datasets to create the final dataset. Our final dataset included the following 7 features/columns: week number, year, total riders for the week, average wind speed for the week, average precipitation for week, average snow for a week and average temperature for the week. Each record or row represented a week in a year.

V. Description of Analytic

The analytic that powers our application is a machine learning regression model. To ensure the reliability of this analytic, we compared two separate regression models. The models we used were the multilinear regression model, and the gradient boosted tree model. What these models did was take in multiple inputs and produce a predicted output. In our case, we built this model by training it on our data set that includes ridership and weather pattern data between 2015 to 2018. To generate output for this analytic, we just need to supply the weather forecast for the next week.

VI. Application Design

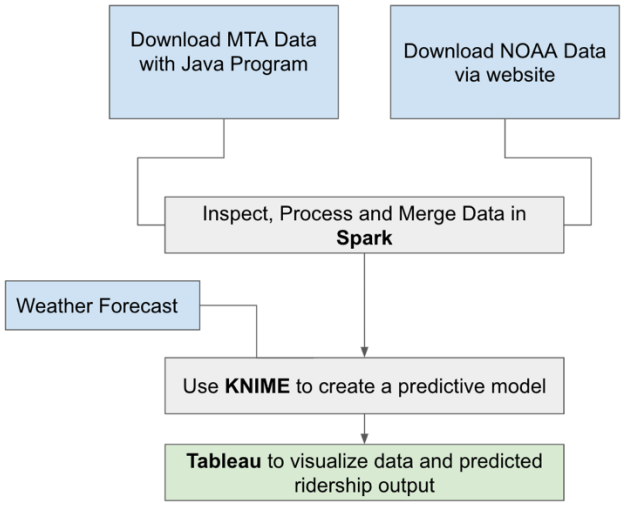

Fig. 1: Design Diagram

At a high level, our application had four distinct phases (see fig 1). The first phase is the data collection and ingestion phase. The second phase is the data processing phase. The third phase is predictive modeling phase. The last phase is the visualization phase.

A. Stage One: Data Collection

During the first phase, data collection and ingestion, we downloaded and moved our data to HDFS. To do this, we built two Java programs that read in the MTA's website source code. The programs would parse the source code to extract all the relevant URLs to the CSV files we needed. Then each program would download their respected CSV files into a local directory on our laptops. The Java program for the collecting the Fare data also inserted a date field into the CSV as the rows of data did not include the date. For the NOAA weather data, this data was downloaded from NOAA's website once NOAA fulfilled our requested date range. Once all of our data was downloaded, we moved it to our HDFS directory to prepare for the next phase: data processing.

B. Stage Two: Data Processing

The second phase of our application is data processing. The primary goal of data processing was to create the final dataset to be used in building our machine learning model during the predictive modeling phase. Data processing was the most computationally expensive phase. Thus all the work for the data processing phase was done in Apache Spark. Data processing comprised of three steps. The first step was to profile or inspect the data. We performed input validation, data type inspection, and checked the completeness of all our data in Spark. The next step, we performed some ETL work to shrink and clean up our data. We removed columns and reformatted data such as date and week values, so all that all three of our datasets were consistent. Lastly, we computed and merged our datasets into one final dataset. To do this, we grouped the weather data into weekly averages as that is the granularity we used due to the weekly reporting of the MTA data. We then took weekly summations of the MTA Fare/Turnstile data. Lastly, we joined the datasets and added a week number data field to the final dataset. As mentioned before, the fare and turnstile count represents the number of users in the system per week. This final data set is now the input data for the next phase: predictive modeling phase.

C. Stage Three: Predictive Modeling

During the predictive modeling phase, the goal was to build and evaluate our machine learning models on their abilities to predict ridership with weather data. To do this, we used KNIME as our tool of choice. In KNIME, we built workflows for both multilinear regression and gradient boosted tree regression. Before feeding our final dataset into our machine learning models, we performed an 80/20 partitioning on the data. 80% of the data would be used to train the model, while 20% of the data would be used to test the model. This partitioning also used a seed so multiple runs of the predictive model can be consistent. The test data would be used to generate the necessarily scoring metrics on how well the models performed. The metrics we are interested in were R^2 and the mean square error (MSE) values. For these models, one could also input data manually, such as next week's forecast. With the test data, the model predicts the ridership for the next week based on the weather data. The predicted value of the test data generated a new CSV file to be used in the next phase.

D. Stage Four: Visualization

The final phase of the application is to present the predicted values and intermediary datasets as a Tableau dashboard. The goal of this phase is to make sure the data is presented in a useful manner so that the MTA can easily use it and make decisions with it. Predicted ridership data were plotted on a bar graph to show how well the predictive model performed. The bar graph also included the original fare data right next to the predicted values. Thus, a user can easily compare visually how well the predictive model performed. Data from the intermediary datasets such as the turnstile count or weather information were also plotted on graphs against fare data to show how the weather patterns relate to the changes in fare intake for a given week. These graphs and charts are then assembled into online dashboards for viewing. Fig 2 shows one of our dashboards from Tableau.

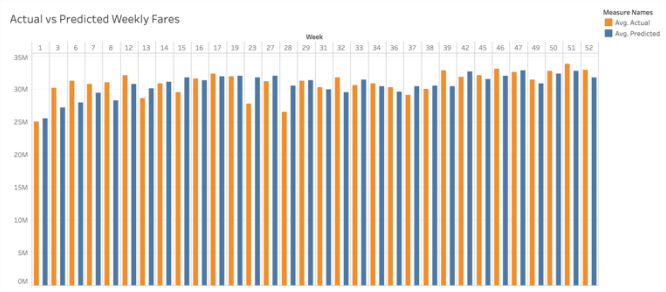

Fig 2. Tableau Dashboard

E. Performance Metrics

To measure the performance of our machine learning model, the performance metrics of R^2 and Mean Squared Error was used. While this is not a machine learning-focused paper, we will briefly discuss what R^2 and Mean Squared Error are and how their values were interpreted.

First, let's look at R^2. Fig 3 is the mathematical notation for the R^2 metric.

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Fig 3: R^2 formula.

SS_{res} stands for the sum of squares of residuals, and SS_{tot} stands for the total sum of squares. "R-Squared" is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)[8]. This value generally indicates how much of the model can explain the variability of the response data around its mean. Alternatively, in other words, if the R^2 value is closer to 1.0, the model can be trusted more while a value closer to 0 cannot be as trusted.

The Mean Squared Error (Fig 4) or MSE is the second metric we used to evaluate our predictive model.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Fig 4: Mean Squared Error.

The mean squared error is a risk function that measures the average squared difference between the estimated values and the actual value[9]. In other words, the further away the estimated value is from the actual value, the MSE will be higher. If the MSE is low, then our predictor is fairly good at predicting values that are close to the actual value.

With these two metrics, our goal is to have a high R^2 and a low MSE. That ensures our predictive model is to be trusted.

VII. Actuation or Remediation

While our application can be automated to provide suggestions and forecast ridership, it is up to the user whether or not to take such action. A rider of the MTA might see the ridership increase is predicted for the upcoming week and decide whether or not to ride the subway. MTA's management with the predicted ridership can decide to deploy more trains or more staff for the system.

VIII. Analysis

A. Development Environment

We carried out building this application both on our PCs and the New York University's (NYU) High-Performance Cluster (HPC).

For our personal computers, we used Apple MacBook Pros running MacOS 10.14. We primarily used our personal computers to develop the Java applications to scrape and download the MTA turnstile and MTA fare data. For Java, we used Java 8. Once all the necessary files were downloaded with our Java applications, we moved the data to New York University's High-Performance Cluster. Besides building our Java application, we used the desktop versions of KNIME and Tableau on our personal computers.

NYU's HPC, also known as DUMBO, is a 48 node Hadoop cluster that is operated by NYU's HPC department. The cluster's OS is CentOS 6.10. On the cluster, we used HDFS to manage our data between Spark. At the time of our research, the cluster ran Spark version 1.6.

B. Insights

After processing all the data and building the application, we've discovered three significant insights about the New York MTA system. The insights are: weather patterns may not be a good predictor of ridership, turnstiles were inconsistently reporting, and lastly our computed ridership numbers were not remotely close to the MTA's report.

a) First insight: Weather patterns may not be a good predictor of ridership. After feeding our data into our machine learning models, we evaluated the models with our performance metrics that we mentioned earlier. Both models scored poorly on both R^2 and MSE. R^2 ranged from 0.25 to 0.38. We also checked if there was a correlation between ridership with temperature. The correlation was only 0.19. With these performance numbers, we were unable to reach a high R^2 value and a low MSE. Thus, we determined weather patterns may not be a very good predictor of ridership.

b) Second insight: Turnstile data is unreliable. After processing the turnstile data, we found that we couldn't just calculate the difference of cumulative count of the turnstiles to determine the total daily riders for the MTA system. On some days based on the total turns, there were about ~7 million turns while on other days it was over 1 billion turns. If we were to interpret this to ridership, there were days over 1 billion+ riders. With further investigation, we noticed that turnstile would go offline and online without correctly being reported. We also noticed turnstiles would roll backward. Lastly, we observed that on average only ~71% of all unique turnstiles are being reported daily. With different turnstiles suddenly coming online or going offline, we saw inconsistent spikes or drops in cumulative count in the reports. Because of these issues, we've deemed the turnstile reporting as inconsistent, unreliable, and untrustworthy. For predictive modeling, we used the Fare data as an indicator of ridership instead of turnstile data.

c) Third insight: Our fare count does not match the MTA's reported count. The New York MTA reports a daily average of 4.6 million riders per day in 2018 [10]. Based on the fares collected in 2018, we only calculated 2.9 million daily riders from fares. While we understand that there are riders who do not pay, based on this calculation almost 40% of all the riders do not pay a fare to ride the system if this number is to be trusted. While we are not the MTA itself, this number seems relatively high in our opinion as the MTA reports that fare jumping cost the MTA $215 million[11]. The MTA claims fare jumping is at 4% [11].

With the combination of inconsistent turnstiles and fare collection, our predictive model was not to be trusted. If anything, this poor data reporting from the MTA may uncover other problems for the MTA if further research and probing was performed.

C. Obstacles

We faced two significant obstacles while building our application. All the obstacles revolved around the quality of the MTA data. We've spent over 80% of our time inspecting and cleaning due to the poor quality of data.

The first significant obstacle with the MTA data was the fact that the data was missing and incomplete. Due to issues with the turnstile or misreporting, we found large portions of the MTA data missing. Since the turnstile data was cumulative, missing data caused any attempt to calculate daily totals to be fruitless.

The second major obstacle was that the data was inconsistent. We've found that the fare count is not even close to the rider count reported by the MTA without explanation. We

discovered only about ~71% of all the turnstiles were being reported. On top of that, we've also found that turnstiles were being resettled randomly. A better practice would be to reset all the counters of the turnstiles the same day every month or every week. That way, even with bad turnstiles going offline, we can know next week's data will not be dependent on last week's bad cumulative data.

IX. Conclusion

In conclusion, we built an application that predicts ridership based on past fare and weather data. However, that prediction is unreliable due to poor performance metrics. Previous research suggests links between ridership and weather. However, our research suggests differently. We do acknowledge some of the previous research also came from other cities with different demographics and transit patterns. That may play a role in how the results turned out. Lastly, we also found some major data problems with the MTA that may affect our ability to predict reliably, or that is close to MTA's own reported rider counts.

X. Future Work

We can revisit this project in the future by improving our input data. To improve prediction metrics, we will need to increase our dataset size and features. We can also consider other factors besides weather to improve our predictive model's performance. Another way of improving our prediction metrics is by using higher quality data or looking for data that links ridership besides fare and turnstile data.

Acknowledgment

We would like to thank New York University's High-Performance Computing for providing our service and support with their Hadoop and Spark cluster.

References

1. A. J. Khattak and A. D. Palma, "The impact of adverse weather conditions on the propensity to change travel decisions: A survey of Brussels commuters," Transportation Research Part A: Policy and Practice, vol. 31, no. 3, pp. 181–203, 1997.
2. S. A. Changnon, "Effects of summer precipitation on urban transportation," Climatic Change, vol. 32, no. 4, pp. 481–494, 1996.
3. A. Singhal, C. Kamga, and A. Yazici, "Impact of weather on urban transit ridership," Transportation Research Part A: Policy and Practice, vol. 69, pp. 379–391, 2014.
4. V. Stover and E. Mccormack, "The Impact of Weather on Bus Ridership in Pierce County, Washington," Journal of Public Transportation, vol. 15, no. 1, pp. 95–110, 2012.
5. "Turnstile Data," mta.info. [Online]. Available: http://web.mta.info/developers/turnstile.html.
6. National Centers for Environmental Information and Ncei, "Climate Data Online Search," Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC). [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/search.
7. "Fare Data," mta.info. [Online]. Available: http://web.mta.info/developers/fare.html.
8. "Coefficient of determination," Wikipedia, 26-Jul-2019. [Online]. Available: https://en.wikipedia.org/wiki/Coefficient_of_determination.
9. "Mean squared error," Wikipedia, 06-Aug-2019. [Online]. Available: https://en.wikipedia.org/wiki/Mean_squared_error.
10. "Annual Subway Ridership." Mta.info, web.mta.info/nyct/facts/ridership/ridership_sub_annual.htm.
11. Bloomberg.com. [Online]. Available: https://www.bloomberg.com/news/articles/2018-12-03/nyc-turnstile-jumpers-bus-fare-cheats-costing-mta-215-million.