New York City Subway 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 MTA system has been stressed with over crowding. But what are some of the external factors that affect the ridership in the system? We developed an application that ingest transit and weather data from the between beginning 2015 to end 2018. We then attempt to provide insight on how weather and seasonal patterns affect ridership with a predictive model.

Keywords—analytics, …

# Introduction

For our application, we want to be able to predict New York City’s MTA ridership with a 7 day weather forecast. Do more people ride the MTA when the weather is warmer? Or 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 historic daily weather data from NOAA. With those two datasets, our application can use a weather forecast such as one from weather.com and predict the ridership for the next 7 days.

# Motivation

One of 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 or congestion of the system. This 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.

Ultimately, these insights can result in action and decision making. With this application, the MTA can benefit by knowing the up coming forecast to whether or not to increase or decrease the number of trains that need to be online. 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.

# Related Work

Research into how weather affects transportation patterns is not a new phenomena. From commuters in cars to public transit, research into transit is very wide in it’s breathe. Through survey research, Belgium researchers have concluded that weather definitely affects people’s decision on how they travel between cars and public transit [1]. 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 ridership during rainy days. 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 policy makers can develop policy that can improve ridership and mitigate weather. Lastly, researches from Washington state looked at how weather and seasons affect ridership of the bus system [4]. They’ve concluded that adverse weather affects system ridership negatively.

One of the major short comings of the 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 will address both as our datasets will be large and have 24hr coverage.

# Datasets

To predict ridership, we needed both 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. However we only used data from beginning of 2015 to end of 2018 due to changes in reporting prior to mid 2014 by the MTA. Total size of our dataset was 10gb.

The first data source we chose was the MTA turnstile data. 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. The types of data in this reports include: 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].

For our weather data, we used daily weather summaries from 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 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 was 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 three 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].

After processing and transformation, our final dataset after merging include the following 7 features/columns: week number, year, total riders for week, avg wind speed for week, avg precipitation for week, average snow for week and average temperature for week. Each row will represent a week in a year.

# Description of Analytic

The analytic that powers our application is a regression model. The model we used was the multi-linear regression model and the decision tree model. What these models will do is 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 use this generate an output for this analytic, we will just need the forecast for the next week.

# Application Design

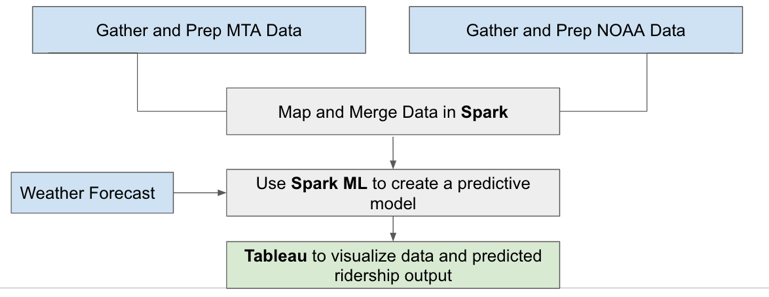


Fig. 1

Our application is built around the Spark distributed processing framework (see fig 1). First, our MTA and weather data is loaded on to the cluster. Spark will preprocess the data by removing unnecessary columns, performing calculations and transforming the data into the format mentioned in our dataset section. Then that data will be used to train our regression model using Spark’s MLLib. To use this application, we will feed it next weeks weather forecast to predict next week’s ridership count. With this ridership count, we can compare it to the previous week’s ridership and suggest increasing or decreasing the amount of trains. The final output information will be presented in Tableau.

Talk about algorithms use

Talk about using knime

# Actuation or Remediation

While our application can be automated to provide suggestions and forecast ridership, it is up the user weather 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.

# Analysis

(In this section, describe: Your experimental setup (tools, platforms), problems (with data, performance, tools, platforms, etc.). Describe what you learned. Discuss limitations of the application. Make recommendations for others, e.g. best practices.)

//process of each step 3-4 para

//bad data road block 2 para. How we found it

//how the quality of prediction perfomed 2 paragraphs

//ml metrics and weather or not we think its useful

//explaination of my investigation into the road block 2 paragraphs

//screenshot

# Conclusion

(One paragraph about the value, results, usefulness of your application.)

# Future Work

To improve prediction metrics, we will need to increase our dataset size and features. Another way of improving our prediction metrics is using higher quality data.

//additional research into the data

//additional feature selection

##### Acknowledgment

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

##### References

(Add references for all of the papers, texts, and data sources that you refer to in your paper. You may have websites to reference, the Spark book, the Hadoop book, etc. A reference is added below as an example.)

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.