NYU

MTA Ridership Predictor

Jason Chan, Reggie Gomez
August 8, 2019

NYU

Introduction

Abstract, Motivation, Goodness and Actuation

Abstract

- New York City's MTA system has been strained with overcrowding and delays. If we can predict ridership, the **MTA will have a better tool to manage their limited resources.**
- We developed an application that **ingest turnstile, fare and weather data** from the between beginning 2015 to end 2018.
- We then to provide insight on how **weather and seasonal patterns affect ridership with a predictive model.**
- Lastly, we built a Tableau dashboard to visualize the predictions and computed insights.

*

Motivation

User	MTA Management
Beneficiary	MTA Management, City Planners and Riders
Importance	This application can help MTA and NYC make better resource management decisions.

Goodness

What steps were taken to assess the ‘goodness’ of the analytic itself?

- Input validation: correct data types and missing values
- Completeness of data: ratio of unique turnstiles reporting
- Machine learning scoring metrics: R^2 value, mean squared error.

Actuation/Remediation

What actuation or remediation actions are/could be performed by this application?

- The application itself does not provide actuation but, instead if provides information for the MTA to act on.
  - Optimize the utilization of resources and staff.
  - Less costly to taxpayers than capital investments.

NYU

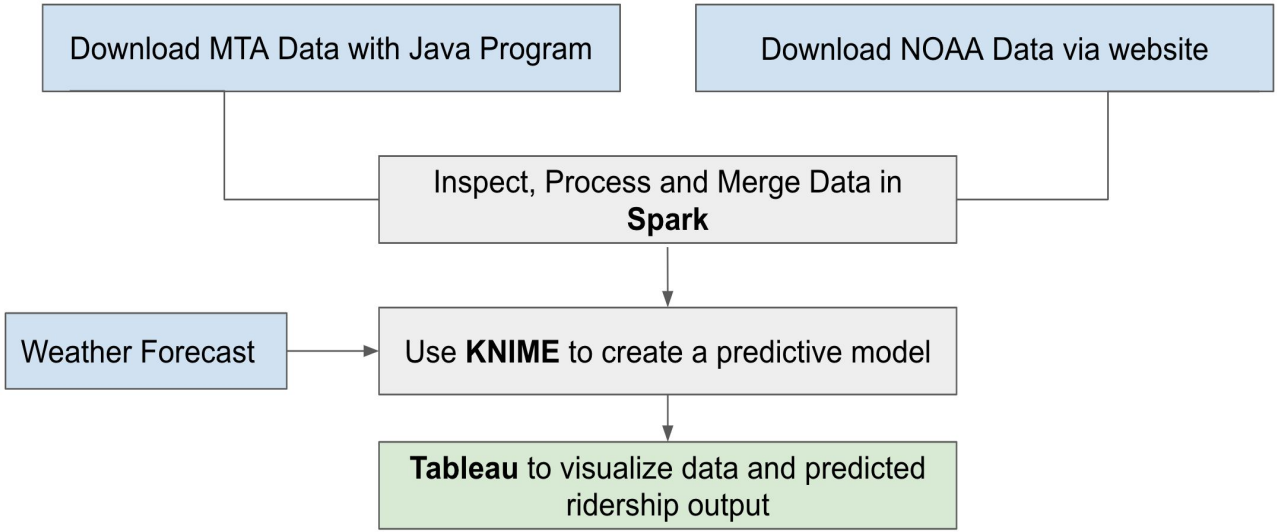
Methodology

Design Diagram and Data

Data Sources

- MTA Turnstile Data
  - Weekly reports of cumulative turnstile data for every station. Reported in 4 hour intervals.
  - size: ~9.5 GB
- MTA Fare Data
  - Weekly reports include a breakdown of all the fare types collected for the week.
  - size: ~400 MB
- NOAA Weather Data Summaries
  - Daily weather summary for a collection New York City region weather stations.
  - size: ~12 MB

Design Diagram



Platforms:

Java
Scala
Spark
KNIME
Tableau

Code Walkthrough

Calculating Cumulative Deltas

```
/*
determine what was wrong with turnstile by calculating cumulative deltas
*/

val list: List[(Long)] = reducedDateCountRDD.map(t=>t._2).collect().toList
val dateList: List[(String)] = reducedDateCountRDD.map(t=>t._1).collect().toLis

val daily = (list zip list.drop(1)).map({ case (a, b) => a - b })
val merged = dateList zip daily
val mergedRDD = sc.parallelize(merged)
val dailyAndCumu = mergedRDD.join(reducedDateCountRDD)
val outputBadDailys = dailyAndCumu.map(t => t._1 +',' +t._2._1+','+t._2._2 )
outputBadDailys.coalesce(1,true).saveAsTextFile("dailyandcumu")
```

Code Walkthrough

Calculating Ratio of Reporting Turnstiles by date

Schema: turnstile id, date, cumulative count

```
//check if all the turn stiles are reporting
val mappedStationTurnstiles = eltTurnstileSplit.map(line => line(0)).distinct
val uniqueTurnstile = mappedStationTurnstiles.count()

val turnStileReport = eltTurnstileSplit.map(v => (v(1), 1))
val turnStileReportDailyTotal = turnStileReport.map(t =>(t._1, t._2.toDouble)).reduceByKey(_+_).sortByKey(true)
val outputTurnStileReport = turnStileReportDailyTotal.map(t => t._1 + ',' + t._2+','+t._2/uniqueTurnstile)
```

NYU

Results

Insights and Obstacles

Insights

1. Unable to predict ridership reliably with weather data.
   - Multi-variable regression and gradient boosted trees R^2 : 0.25 to 0.38.
   - Low R^2: model explains very little of the variability of the response data around its mean
   - Correlation between fares and temp or fares and rain were below 0.19.

2. Our computed daily ridership did not remotely match MTA's own report
   - 2018 MTA reports daily avg of 4.7M. We computed 2.9M from fares.

3. A lot of turnstiles misreporting or offline
   - avg of ~72% of all unique turnstiles are being reported daily.

Obstacles

1. Turnstile data only showed **cumulative count** every 4 hours

2. Turnstiles sometimes go offline. We calculated a ratio to determine how many turnstiles are being reported per day.

3. We concluded turnstile data was really bad. Thus we used Fare data for ridership count. Turnstile data revealed daily turns ranged from negative to 1B+. Observation: Average turns per day for date range Jan 1, 2015 to Dec 31, 2018 was ~63 MILLION. This is a lot more than the 4.7M daily average

NYU

Conclusion

Insights and Obstacles

Summary

- Weather data may not be a strong predictor of MTA ridership.
- Past research shows reduced ridership during adverse weather, however, this is not what our data and predictive model showed us.
- Previous research also came from cities with residents that may not need to rely on public transit for commutes.
- Our research also showed us MTA data is of poor quality or incomplete.
- Demo: Link to Tableau Dashboards

Acknowledgements:
- New York University High Performance Computing
- Prof McIntosh

References

- 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.
- S. A. Changnon, “Effects of summer precipitation on urban transportation,” Climatic Change, vol. 32, no. 4, pp. 481–494, 1996.
- 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.
- 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.
- “Turnstile Data,” mta.info. [Online]. Available: http://web.mta.info/developers/turnstile.html.
- 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.
- “Fare Data,” mta.info. [Online]. Available: http://web.mta.info/developers/fare.html.