

Predicting Departure Delay Durations

Denver RTD Light Rail



Jason Summer
2020

Agenda

01

Executive Summary

02

Problem Statement and Proposed Solution

03

Data Wrangling Effort

04

Data Exploration Analysis and Inferential Statistical Analysis

05

Predictive Modeling Evaluation

06

Opportunities for Improvement



Objective

Improve train efficiency and ridership through more transparent and widely available train departure times

A year's worth of RTD light rail departures were studied, alongside of surrounding weather and local sporting events. Correlating factors were identified, but more importantly 2 predictive models were pursued.

One model with a goal of supporting real-time predictions for immediate riders showed promising results. This model was able to estimate whether train departures would be 5-10, 10-15, or 15-20 minutes delayed with roughly 80% accuracy and 30-60 minutes delayed with 89% accuracy.

While areas of improvement are highlighted following the subsequent analysis, this real time model would be an effective addition to any real-time rider alert system.

Agenda

01

Executive Summary

02

Problem Statement and Proposed Solution

03

Data Wrangling Effort

04

Data Exploration Analysis and Inferential Statistical Analysis

05

Predictive Modeling Evaluation

06

Opportunities for Improvement

51% of train departures are over a minute late with the average train 1 minute and 42 seconds delayed.

THE DENVER POST

News ▾ Sports ▾ Business ▾ Entertainment ▾ Lifestyle ▾ Opinion ▾ Politics ▾ Classifieds ▾

Search



NEWS ▸ TRANSPORTATION

RTD cancels about 100 light rail trips Monday, days ahead of discussion on driver operator shortage

RTD says it's working to recruit more people, address service issues

NEWS ▸ TRANSPORTATION

Frozen switch leads to 15 to 20 minute delays for light rail passengers

NEWS ▸ TRANSPORTATION • News

More than 500 Denver International Airport flights canceled or delayed as snow rolls in

RTD bus and light rail service delays reported

NEWS ▸ TRANSPORTATION • News

Icy conditions slow Denver's bus, light rail and commuter rail systems

Traffic crashes on icy highways close major highway lanes during commute

NEWS ▸ TRANSPORTATION

Light rail service restored after power issues disrupted Wednesday rush hour service

Train riders should plan for additional travel time between affected stations

Light Rail Challenges

RTD faced increasing adversity in 2019 impacting train performance, ridership, and subsequently the brand's reliability.

Evolving Challenges



Driver Shortage

In November 2019, RTD surveyed riders to determine if service cuts or delays and cancellations are preferred in wake of a driver shortage



Weather Conditions

Riders sought refuge in public transportation during harsh weather conditions but experienced unexpected delays without notice



Outages

While rare, outages have left unknowing riders stranded, as seen on September 11th, 2019.



Real Time Signage

RTD's mobile ticketing app and real-time location website are limited in their delay notifications. Station signage reports schedule times. RTD has recommended enrolling in email alerts or following their twitter account.

Proposal

Provide RTD and riders alike with model outputs that can provide estimated delays in future train departures

Input Data Sources



Past Performance

Approximately 30% of light rail vehicles' and all commuter rail vehicles' station departures are tracked with some days removed by RTD due to extreme weather. The data was sourced directly from RTD.



Weather

Hourly weather was sourced from visualcrossing.com for 13 cities in Colorado. Data included information about ice, snow, rain, wind, heat index, temperature, humidity, etc.



Local Events

The starting day and time of games hosted locally by the Denver Nuggets, Colorado Avalanche, Denver Broncos, and Colorado Rockies were sourced from the Denver Post, sports-reference.com, and Altitude TV. Some schedules spanned multiple seasons.



Prior Delays

From the RTD data itself, the departure delays at prior stations were derived and included as lag feature consideration.

Currently, RTD's mobile ticketing **app does not alert riders of disruptions** and it cannot send texts. The **real-time location website is limited** and the electronic signs at **stations show scheduled times** instead of updated times.

RTD has indicated **investment in improved real-time alerting**, which was estimated to be a couple months out as of September 2019. Being able to **predict both immediate delays and high-risk delays further in the future** could fill the void of real-time and accurate delay updates or improve any current efforts to do so.

Agenda

01

Executive Summary

02

Problem Statement and Proposed Solution

03

Data Wrangling Effort

04

Data Exploration Analysis and Inferential Statistical Analysis

05

Predictive Modeling Evaluation

06

Opportunities for Improvement

Key RTD Data Wrangling

Pandas, numpy, matplotlib, seaborn, datetime, and dateutil packages were used for manipulation of sourced data.

✓ Station Identifier

A location mapping file was provided, however, not all station codes were accounted for from the rail data. A one source of truth of readable station names were coalesced from the mapping file's names and rail data's native cross street information.

✓ Arrival and Departure Times

Dates of scheduled station departures were merged with timestamps of actual arrivals and departures. The date stamps were adjusted accordingly in instances that the arrival/departure straddled midnight. New wait times and departure delays were calculated.

✓ Time Delta Fields

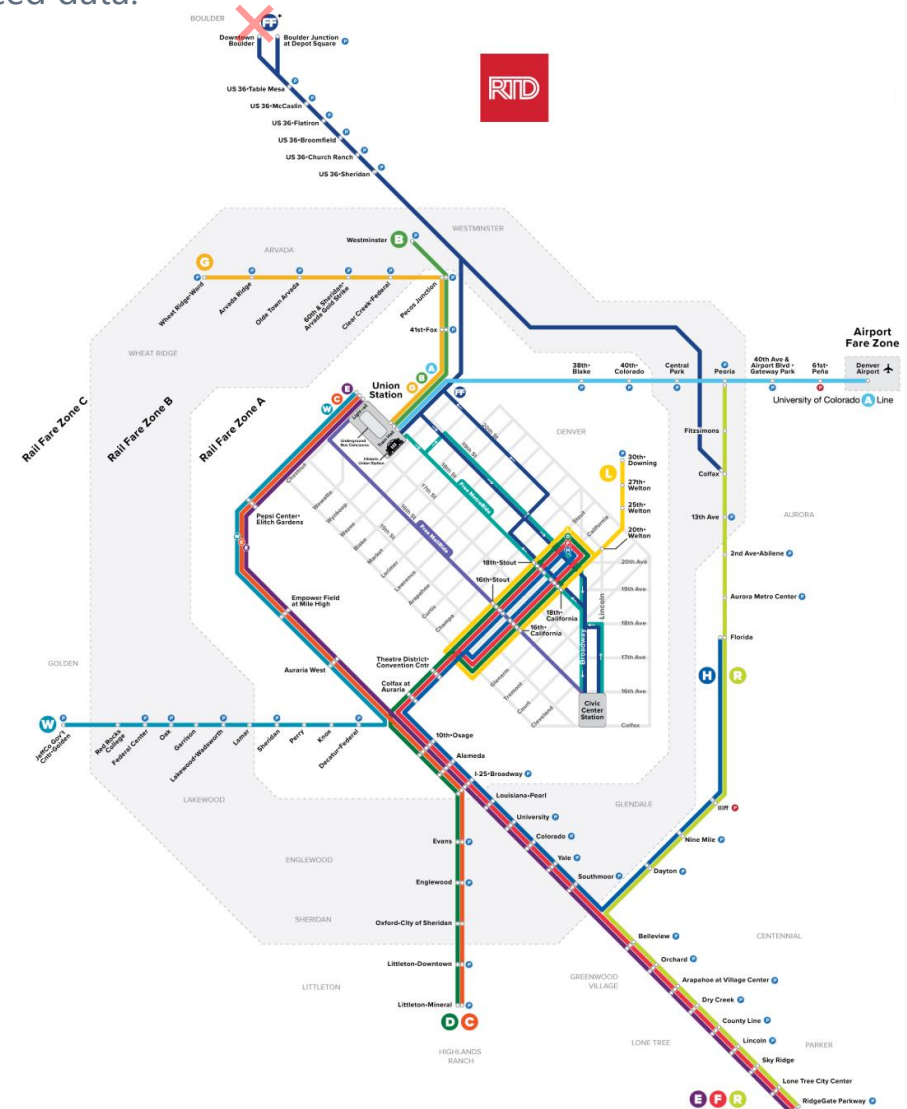
Time delta fields including station dwell times, delay durations, and delay flags were re-calculated based on new timestamps.

✓ End of Line Departures

Evaluation of the native sort order field indicated that some stations were missing, others out of order, and departure delays were captured for final stations. Thus, end of line stations were flagged to avoid skewing departure predictions

✓ Unnecessary Field Removal

RTD uses a similar data schemas to track non-rail vehicles. These extraneous fields and other unnecessary fields were removed for size reduction.



Key Weather Data Wrangling

Pandas, numpy, matplotlib, seaborn, datetime, and dateutil packages were used for manipulation of sourced data.

✓ NOAA Weather by County

Initial weather information was obtained from NOAA and sourced by county. Empty, duplicative, and inaccurate county values were researched and updated.

✓ Visual Crossing Weather by City

NOAA weather data was scarce and replaced by data sourced from visualcrossing.com by city. Again, city values had to be researched and updated and dummy indicator fields were created for 22 unique weather descriptors.

✓ Weather Data Merge

The visualcrossing.com weather data were merged into the train dataset by applying the closest hourly record of weather data to each scheduled train station departure time.

✓ Weather Descriptors

A combination of 22 unique weather descriptors were included in the weather data, such as 'thunderstorm' or 'hail'. 22 hot key indicator columns were initially created for each unique description. Subsequently for modeling preparation, ordinal weather fields were created for varying degrees of rain, ice, and snow.

Hourly Weather Data

- precipitation type and volume
- snow type and depth
- ice / hail type
- wind speed
- relative humidity
- temperature
- visibility

Key Sporting Event Data Wrangling

Pandas, numpy, matplotlib, seaborn, datetime, and dateutil packages were used for manipulation of sourced data.

✓ Rockies Schedule Sourcing

The 2019 Rockies game scores were obtained from sports-reference.com. However, start times were not included and were subsequently sourced separately from the Denver Post. Changes in the regular schedule were identified when these 2 sources were merged.

✓ Rockies Schedule Corrections

Research indicated 2 double-header days were played after prior game cancellations. Additional validation was performed by matching the opponent field of both the score and schedule files (mentioned above).

✓ Away Games

All away regular and postseason games were removed from the Rockies, Broncos, Avalanche, and Nuggets schedules.

✓ Consistent Datetimes

All game start times were re-formatted to a 24-hour clock. The Rockies data required additional correction for an inconsistent format and the 2nd game in a double-header did not have a published start time and was eventually removed. The Broncos schedule required conversion to local time.

✓ Hours Until Game Time

During modeling preparation, new fields were introduced to capture train departures occurring on the same day as local professional sporting games. The field captured the hours between scheduled departure and game start times.

Hours between Scheduled Departure and Sporting Events



Agenda

01

Executive Summary

02

Problem Statement and Proposed Solution

03

Data Wrangling Effort

04

Data Exploration Analysis and Inferential Statistical Analysis

05

Predictive Modeling Evaluation

06

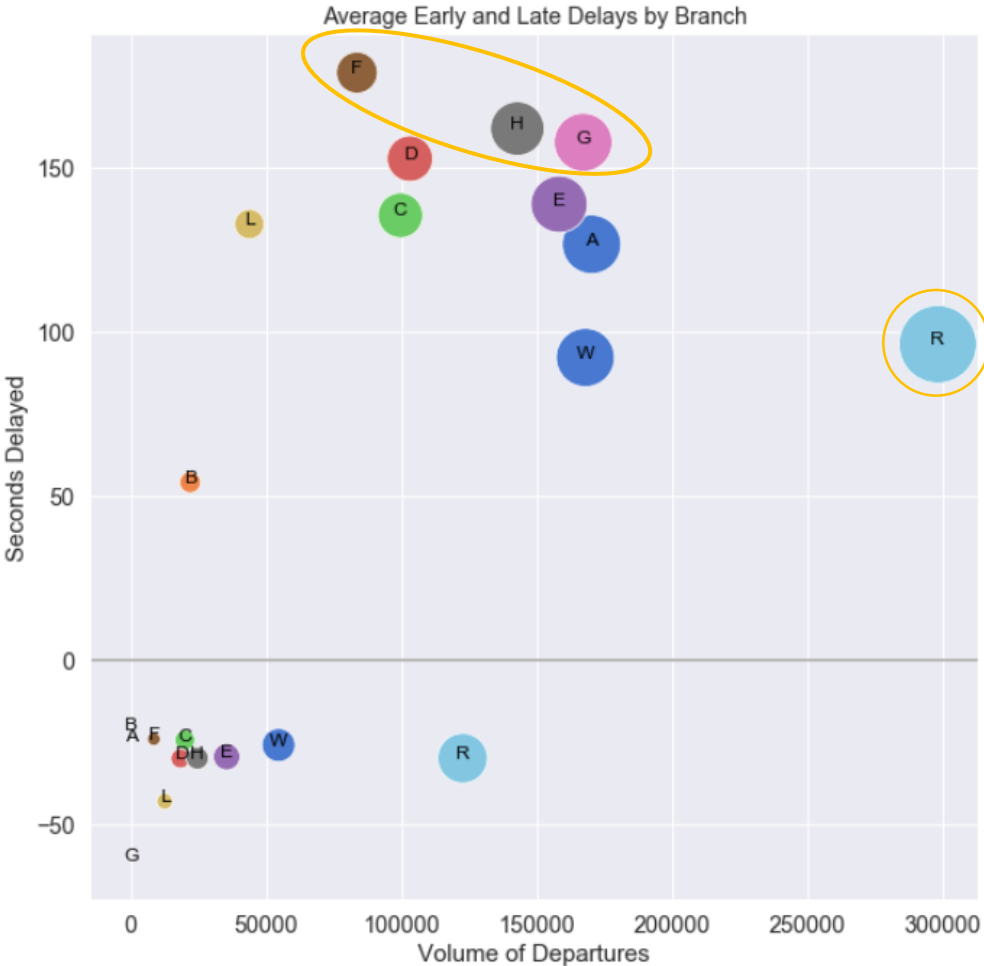
Opportunities for Improvement

Route Analysis

Certain routes show far more common delays while others reveal more extreme delays

Route Comparison

Route / Line	Overall Avg Seconds Delay
A	127
B	54
C	108
D	125
E	108
F	160
G	157
H	133
L	93
R	59
W	63



Nearly every route shows an average departure above 1 minute.

While lines F, G, and H show the highest average delays, respectively, F does not feature the highest volume of delays. On the contrary, G and H shows a high number of departures, thus extrapolating the impact of their delays.

Line R seems to feature the most departures with the average delayed departure around a minute and a half and the average on-time/early departure roughly 30 seconds before scheduled.

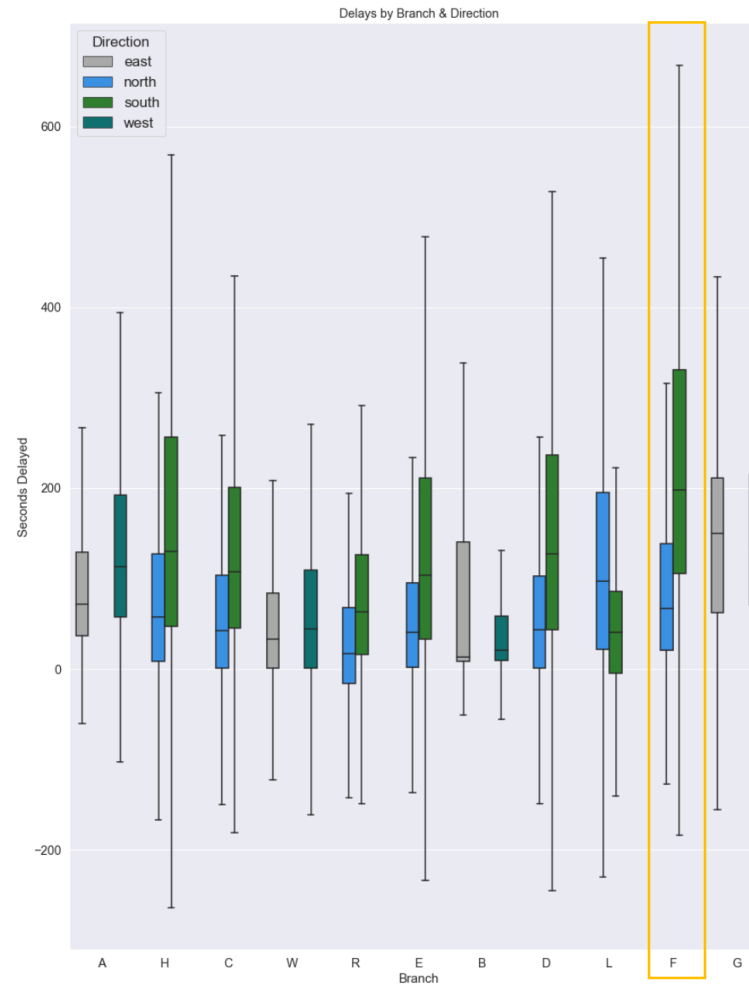
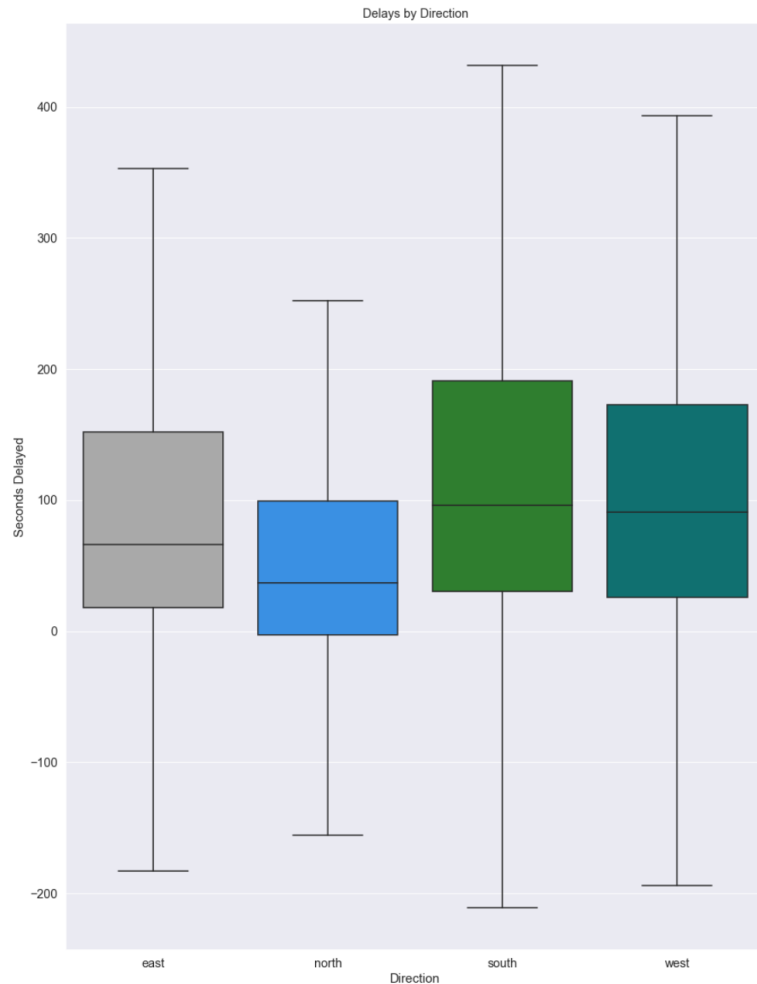
A one-way F test indicated a statistical significance (p-value ~ 0.00) in the difference in average departure delay among the different lines.

Left chart considers the average of both early and late departures
p-value of ~0.0 indicates the numpy minimum threshold of 1E-16 was exceeded

Direction Analysis

Train direction affects on-time departures, indicating the same route in a different direction can be less reliable

Direction Analysis



Simply comparing train direction shows significant overlap. However, northbound performance seems to show the lowest delays while southbound trains show the highest delays and most variability.

However, comparing train performance by direction for each line shows much higher separation between north and southbound trains. The greatest separation can be seen with line F.

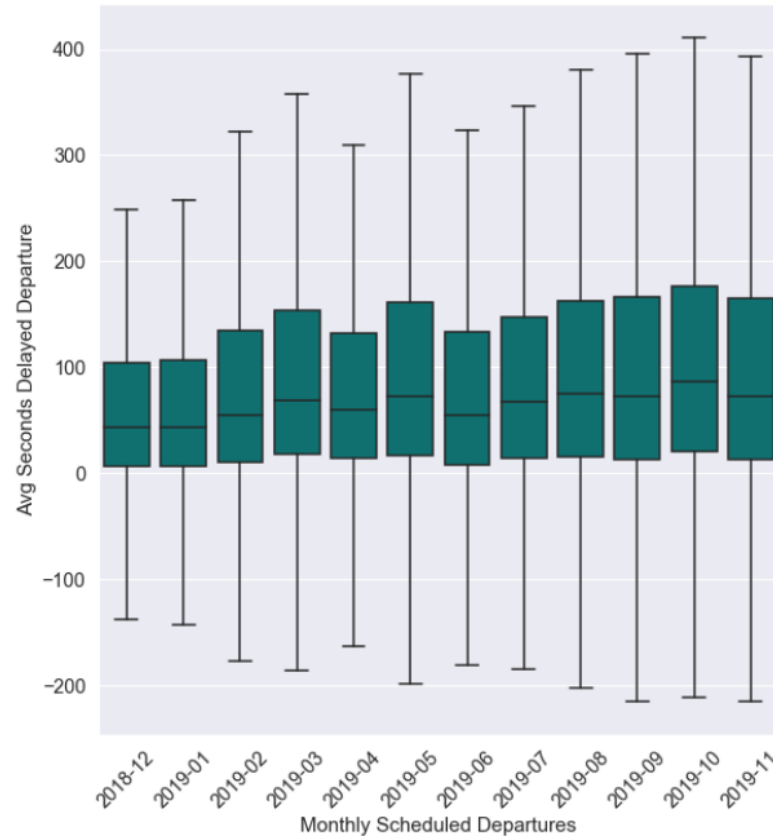
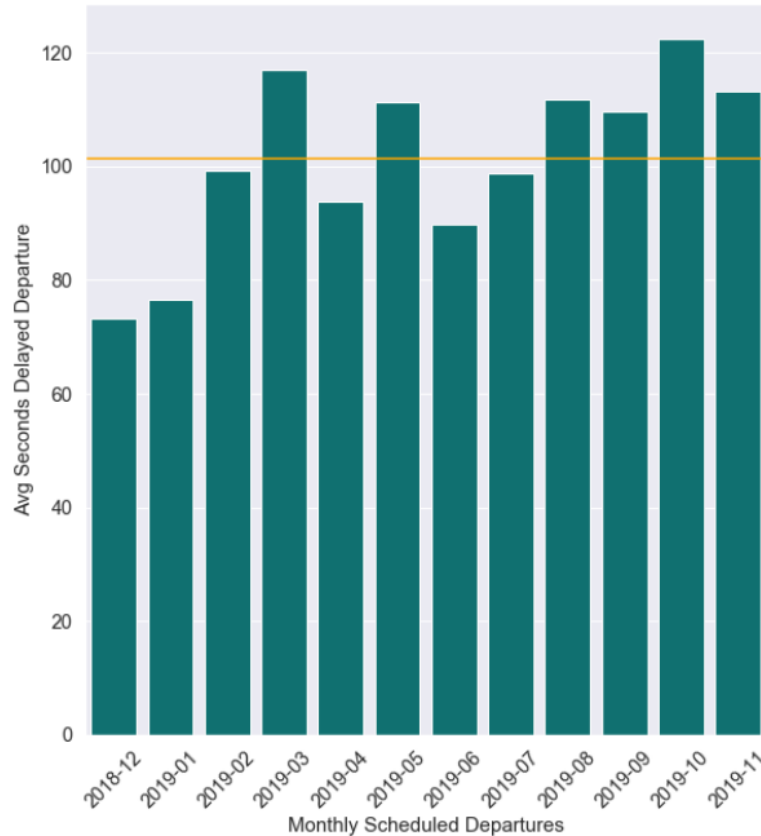
A permutation resampling technique used to compare the mean delay of southbound trains with all other directions resulted in a p-value of ~ 0.00 , indicating a significant difference between southbound trains and other directions.

p-value of ~ 0.0 indicates the numpy minimum threshold of $1E-16$ was exceeded

Time of Year Exploration

While seasonality cannot be confidently determined with 12 months of data, there is a slight uptick in delays around Q4 2019

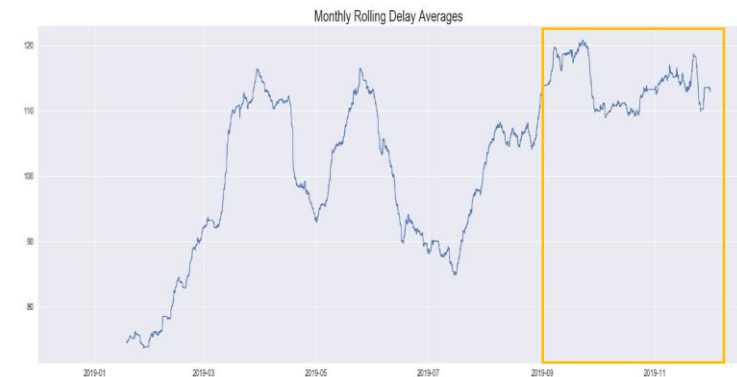
Calendar Analysis



There seems to be differences among the months. However, year-over-year is not possible for comparison.

Interestingly, September through November 2019 show higher than average delays. Media circulation of RTD operator shortages started around September and October 2019.

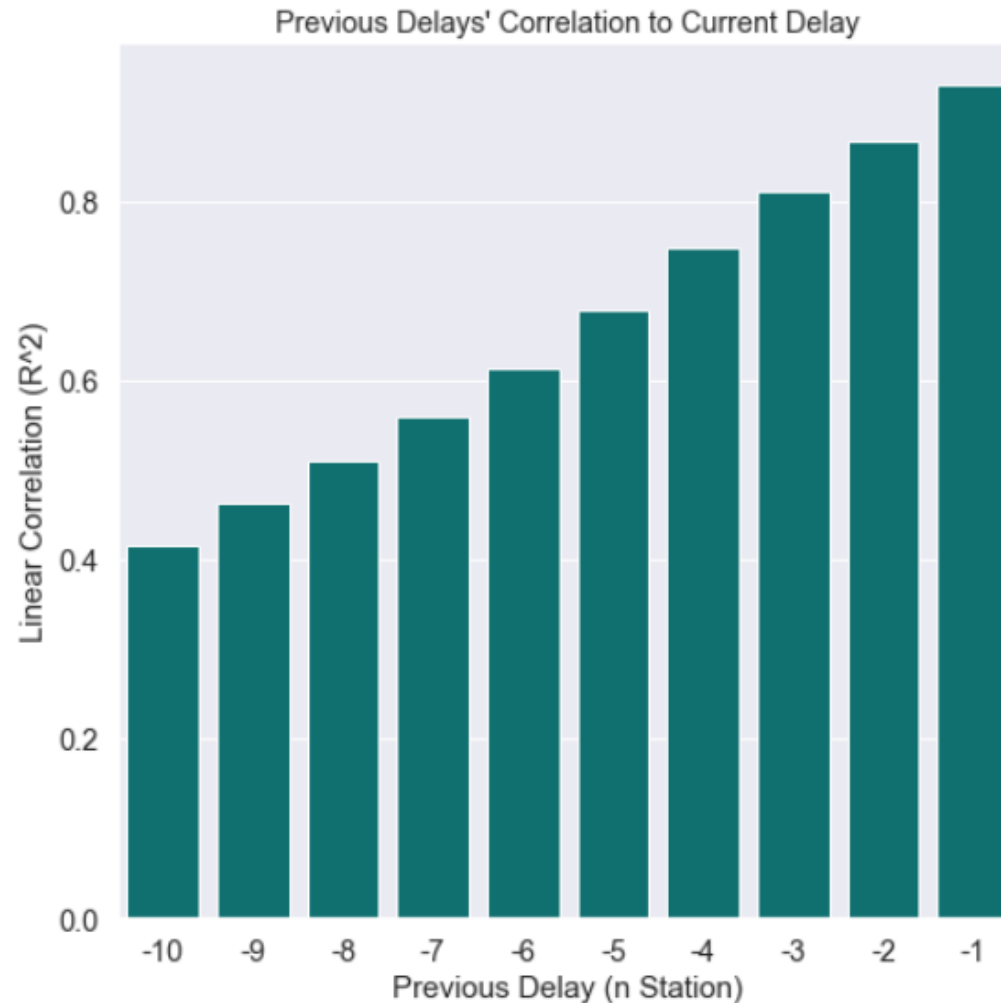
The below monthly rolling delay view shows this uptick more clearly.



Prior Delay Impact

As expected, immediately prior delays shows strong correlations to the current delay, but it is not constant across the route

Prior Delay Lag Indicators



The bar plot to the left shows the linear correlation between previous delays (up to the 10th prior station) on the current delay.

The average route has ~12 stations. However, showing the prior 10 delays shows an obvious pattern, in which the correlation increases as you approach the current delay. Similarly, prior stations further back are less linearly correlated to the current delay.

This pattern of increasing correlation indicates that delays in a line are not constant as one may expect. Instead, trains operators seem to make up for early delays as they progress in their route.

Some stations are marked as 'free running' indicating operators may leave earlier than scheduled. Providing notice of early departure can be just as important as late departure in many situations.

Local Sporting Event Impact

Identifying the departures that occur in the hours leading to the beginning of local sporting events and those following the sporting events may indicate an influx of riders and subsequent delay spikes.

Hours before Game Time



bars reflect the 25-75 percentile range, i.e. the middle 50% of departure delays

Delays tend to be rather variable throughout the entire dataset. However, showing delays before, during, and after local sporting events (occurring on the same calendar day), may allude to certain spikes in delays.

The 2 hours leading up to the beginning of Broncos and Avalanche games show obvious spikes in delays. A similar but less obvious pattern is seen with Rockies games. Interestingly, there is also a spike around where Broncos games would likely end, approximately 3-5 hours after kickoff.

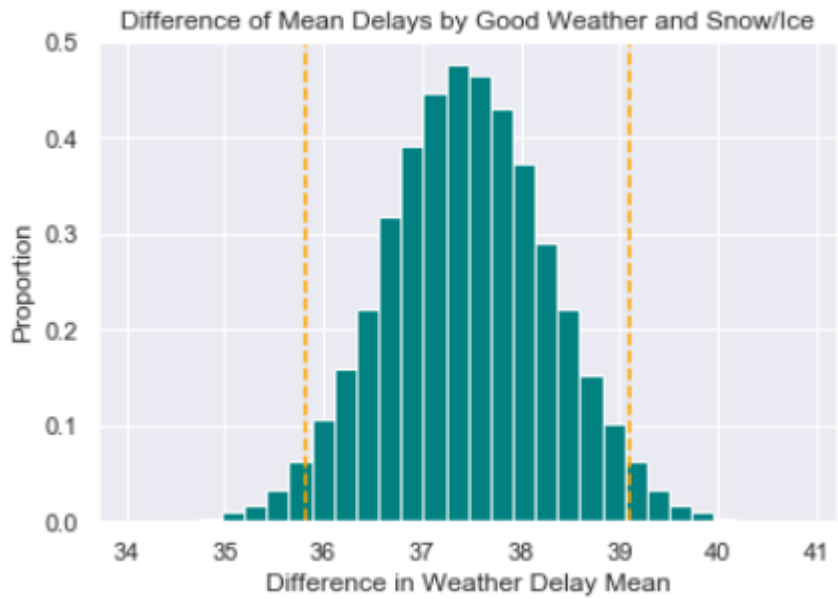
Weather Impact

A simple comparison of clear days with snow or rainy days indicates a statistically significant difference in average delays

Fine Tuning and Comparing Weather

Rain Description	Ordinal Rank	Ice Description	Ordinal Rank
Light Drizzle	1	Light Freezing Drizzle/Freezing Rain	1
Drizzle	2	Light Freezing Rain	2
Heavy Drizzle	3	Freezing Drizzle/Freezing Rain	3
Mist	4	Hail Showers	4
Light Rain	5		
Rain	6	Snow Description	Ordinal Rank
Heavy Rain	7	Light Snow	1
Rain Showers	8	Light Rain and Snow	2
Thunderstorm	9	Snow	3
		Blowing or Drifting Snow	4
		Heavy Snow	5

In addition to the quantitative fields for wind, temperature, etc., key weather descriptions were converted to ordinal indications of rain, ice, and snow conditions.



Using bootstrapped samples of 50,000, the mean delay of clear weather departures were compared to the mean delay of departures occurring around snow or rain conditions.

The 95% confidence interval of the difference of these means lies between the orange lines. The exclusion of 0 in this confidence interval indicates a statistically significant difference in the mean delays during clear and rain/snowy conditions.

Agenda

01

Executive Summary

02

Problem Statement and Proposed Solution

03

Data Wrangling Effort

04

Data Exploration Analysis and Inferential Statistical Analysis

05

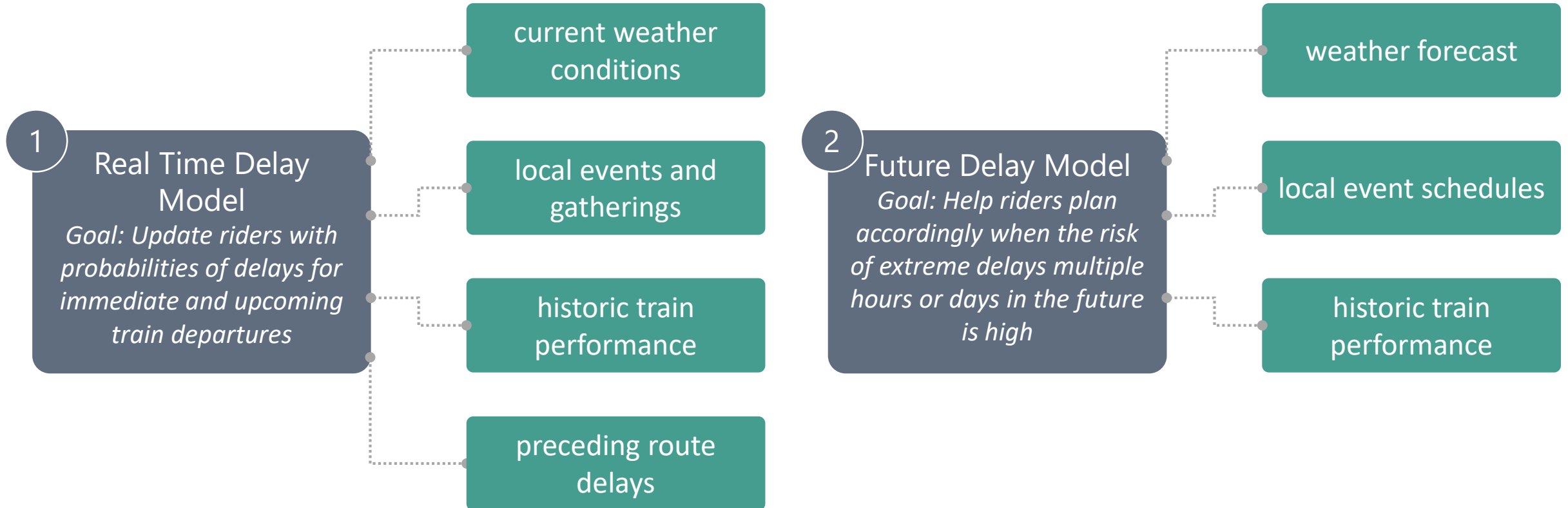
Predictive Modeling Evaluation

06

Opportunities for Improvement

2-Model Approach

Differentiated by feature selection, a real time prediction model and a future prediction model were pursued using random forest regressors

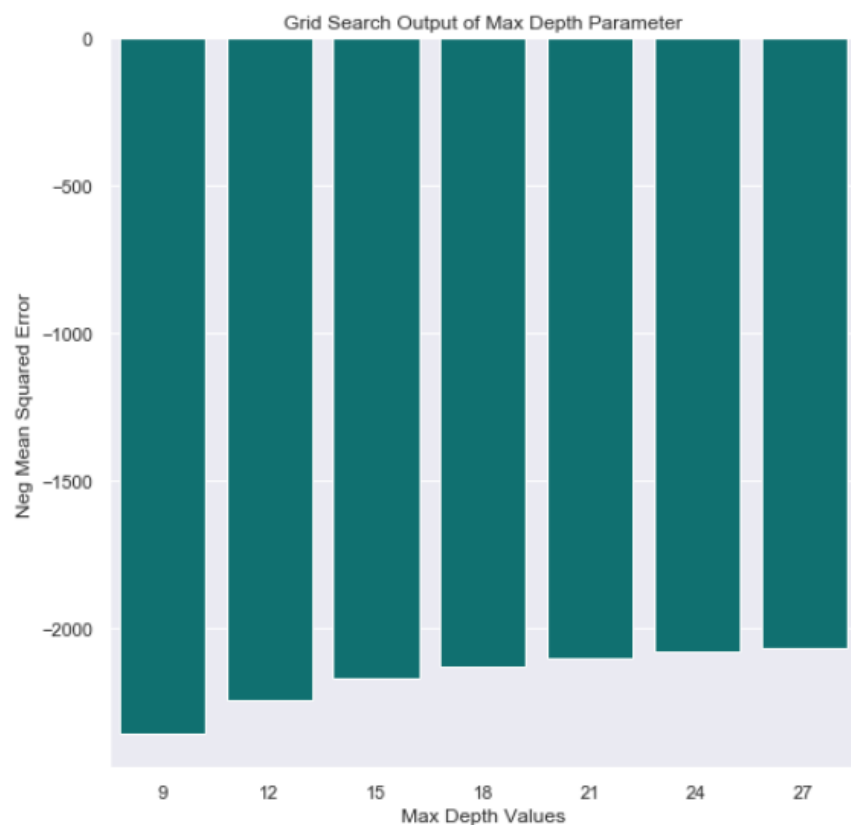


The approach for both models is presented in the following section. However, as will be noted, the real-time delay model provides quality predictions while the future delay model likely requires additional data for usable predictions.

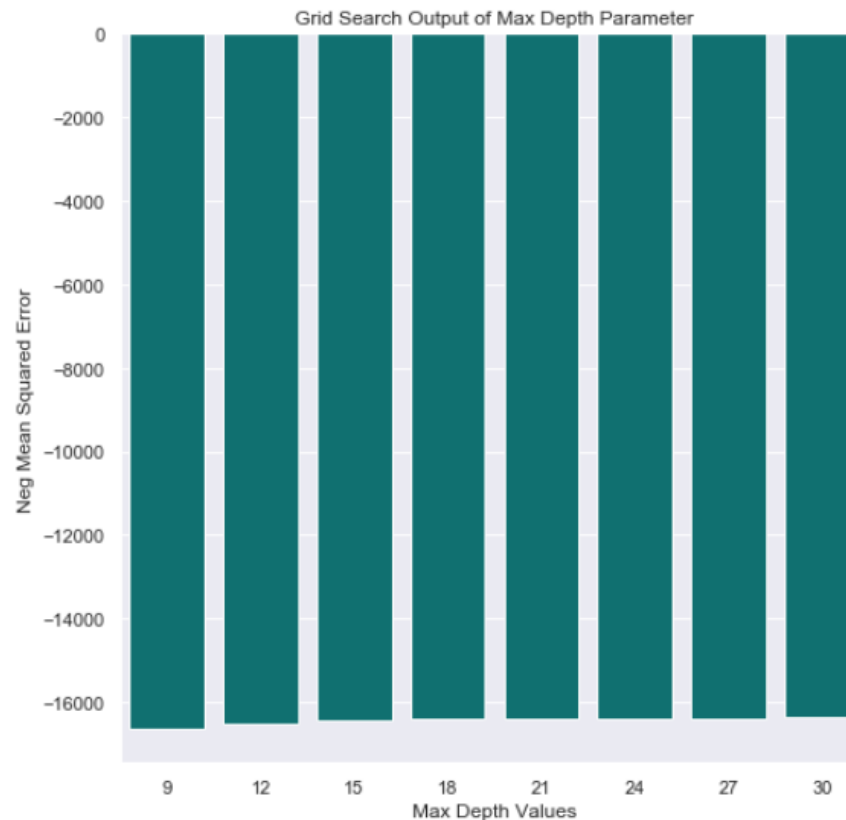
Hyperparameter Tuning

Due to the sheer variability of features, a random forest regressor was used and the max depth hyperparameter tuned

Real Time Delay Model



Future Delay Model



For each model, the dataset was split into training, validation, and testing (final holdout) sets.

Training	6% of data
Validation	24% of data
Test (Holdout)	70% of data

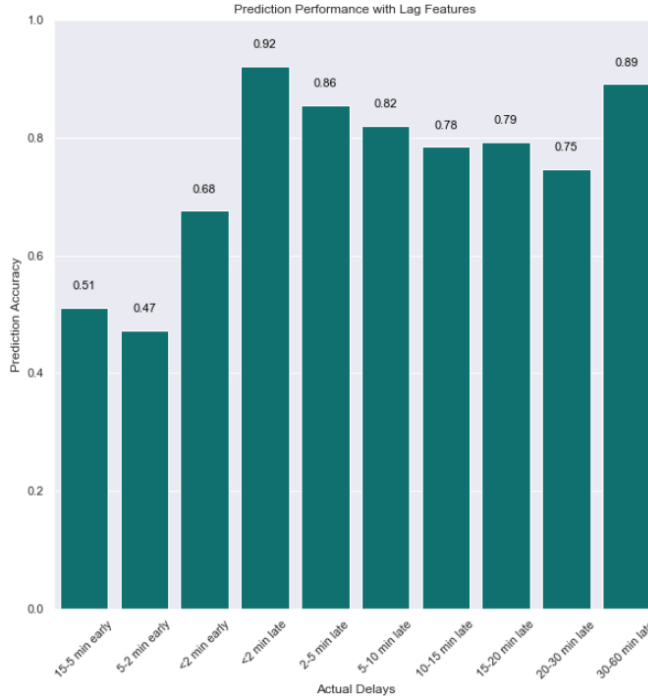
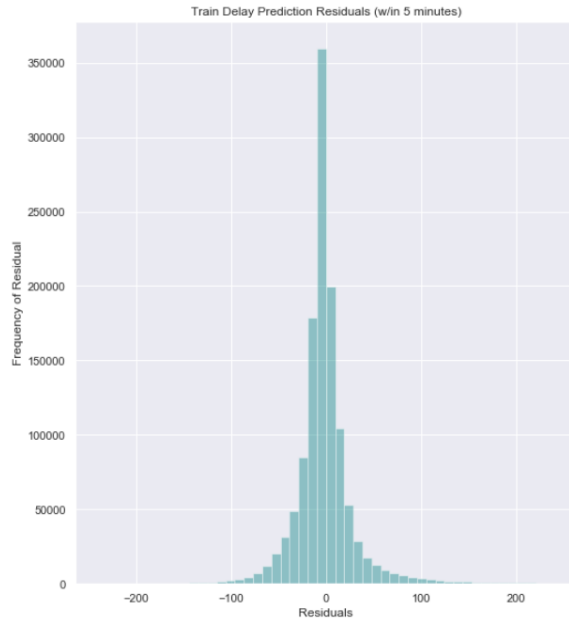
The training set was used to create the results to the left, indicating the optimal max depth value as produced by 3-fold cross validation grid searches.

The optimal max depth value was then re-trained and checked for over-fitting using the validation set.

Max depths of 27 and 18 were selected for final predictions based on reasonable amounts of over-fitting and best generalization abilities on the validation sets.

Model Performance

Categorizing departure delays into reasonable ranges shows prediction opportunity for delays between 2 minutes and 60 minutes using the real-time model



Cutting delays into more digestible time estimates, as shown in the bar chart, indicates that delays between 2 minutes and 60 minutes can be appropriately categorized into their respective range with around 80% accuracy.

As shown via the histogram, prediction residuals are normal, with the bulk of the residuals within a range of 100 seconds of the actual delay.

The scatterplot indicates increasing residual size and variance as the delay increases, which is expected as unforeseen circumstances, i.e. mechanical issues, not captured in the data may be responsible for large delays.

This model could thus be used to update riders' probability distributions of potential delays as trains near.

The future delay model, however, did not perform well without the lag features included in the real-time model. See Appendix for similar benchmarking of this model and the Opportunities for Improvement section for ways to improve both models.

Agenda

01

Executive Summary

02

Problem Statement and Proposed Solution

03

Data Wrangling Effort

04

Data Exploration Analysis and Inferential Statistical Analysis

05

Predictive Modeling Evaluation

06

Opportunities for Improvement

Opportunities for Improvement

Improved data quality, richness, and increased data considerations could lead to improved real-time performance and possibly future delay predictions beyond real-time

✓ Improved RTD Data

Severely delayed trains due to inclement weather, newer lines and stations, and unmonitored stations and trains were excluded in this dataset. Richer and more representative train records would likely improve performance

✓ Seasonality Studies

Multi-year data would help to understand prospective changes due to seasonality. Additionally, data spanning several years could also assist in estimating the true impact of the operator shortage on delays in late 2019 and onward.

✓ Ridership Information

Riders are not tracked in a congruent fashion as trains and rider data is not available to the public. However, having estimated ridership for each train departure would allow one to determine if capacity has an affect on delays. Furthermore, it could help to identify high impact areas based on rider volume.

✓ Additional Local Events

In addition to games hosted by the Broncos, Nuggets, Rockies, and Avalanche, widening the consideration of local events would further expand the model and its usage. For instance, large scale events, such as festivals, could advertise with more accurate transportation recommendations.

✓ Alternative Time-Based Features or Models

Instead of the 4 prior delay lag features considered in this model, alternative time-based features could have been added. For example, a time-based window feature. Alternatively, a regression analysis on the appropriate number of lag features or window length could have been conducted as well.

✓ Geospatial Data

Including a spatial analysis of routes between stations would help to estimate the impact of the distances and identify possible areas of high variance, such as streetlights or queuing spots.

Appendix

Future Delay Model Output

