Forecasting Boston MBTA Transit Dynamics: A Performance Benchmarking of Statistical and Machine Learning Models

Aiden Ruipeng Zhou*[1,4] Sai Siddharth Nalamalpu*[2,4] Kaining Yuan*[3,4], Patrick Bloniasz[4], Eugene Pinsky[4]

Western Academy of Beijing, 10 Laiguangying E Rd, Chao Yang Qu, Bei Jing Shi, China, 100102[1]; Sunset High School, 13840 NW Cornell Rd, Portland, OR 97229[2]; Woodbridge High School, 2 Meadowbrook, Irvine, CA 92604[3]; Boston University, 125 Bay State Road, Boston, MA 022154[4]

*Authors Contributed Equally

Introduction

Definitions:

- *Temporal Point Processes (TPPs)* are a random process where isolated events are scattered in time. In this case, we consider history-dependent point process models; these help consider factors such as the correlation of prior delays with future delays. These constitute the "self-exciting" elements of the model.
- MBTA: The Massachusetts Bay Transportation Authority; responsible for public transit through systems including trains, subways, and buses.

Inspiration: While at BU RISE, the MBTA system was a necessary tool to get around campus, be it to go to class, participate in social activities, or go on field trips. However, we sometimes experienced delays in the MBTA system, and so became curious about whether such behavior was predictable, and if so, how best to predict it.

Goals:

- Determine which factors influence delay and ridership prediction the most.
- Rank models on predictive accuracy.

Models:

- Random Forest (ML)
- Gradient Boost (ML)
- Multilayer Perceptron (ML)
- K-Nearest Neighbors (ML)
- Support Vector Machine (ML)
- Linear Regression (Statistical)
- Ridge Regression (Statistical)
- Lasso Regression (Statistical)
- Moving Average (Statistical)
- Poisson Regressor (Statistical)
- Point Process (only for delay data) (Statistical)

Methods

Target Metrics

We first compute RMSE values for varying models in predicting some "target metric":

- The total number of delays (named "Delay") in the next day
- The total number of gated station entries (named "GSE") on the next day

Input Data Splits

For each task, each model was provided with inputs including a set of metrics of the prior 5 days, which were shaped into a 1D tensor. Then, 100 cycles were run; in each cycle, a set of data points equivalent in length to the original data points was selected with replacement from the original data points.

In each test, the first 80% (ordered by date) were selected as a training set, and the remainder as a test set. This was:

- For delay counts: 1333 train data points, 333 test data points
- For gated station entries: 3355 train data points, 839 test data points

Input Data Blends

To evaluate performance changes given changes in provided data, the set of input data for each model test varied. The sets of metrics used were:

- Only the target metric
- The target metric and day of week
- The target metric and season
- The target metric and weather data
- The target metric, day of week, and season data
- The target metric, day of week, and weather data
- The target metric, season, and weather data
- The target metric, day of week, season, and weather data

Input Data Enhancements

To mitigate differing data requirements for models (e.g., multilayer perceptrons' superior performance given one-hot encodings), models were tested on each set of metrics up to 4 times. Tests included:

- Original metrics
- Scaled (via scikit-learn's StandardScaler) metrics
- One-hot encodings (if applicable) of metrics
- Scaled one-hot encodings (if applicable) of metrics

Test Summary

Overall, 27 tests were performed for each model in each test cycle. There were 10 models, amounting to 270 tests per cycle, 27000 per task, and 54000 tests in total. The minimum RMSE of all such tests was considered maximum model performance ("Any Data"), and the RMSE of the test given no additional data was considered raw model performance ("No Additional Data").

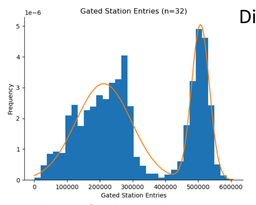
Homogeneous Point Process Model

Key Assumptions:

1. $P(A|B) = P(A)$
2. For any interval, $(t, t+\Delta t]$, $\Delta N_{(t,t+\Delta t]} \sim P(\mu)$ with $\mu = \lambda \Delta t$.
3. For any non-overlapping intervals, $(t_1, t_2]$ and $(t_3, t_4]$, $\Delta N_{(t_1,t_2]}$ and $\Delta N_{(t_3,t_4]}$ are independent.

*** Dependent data can be used with a Homogeneous Point Process, though Assumption 1 would be violated, by decreasing the window size to make the data independent.**

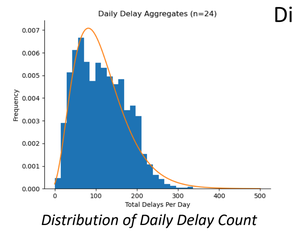
Data Distributions



Distribution of Daily Gated Station Entries

Distribution description:

1. Bimodal
2. Captured using two normally distributed curves



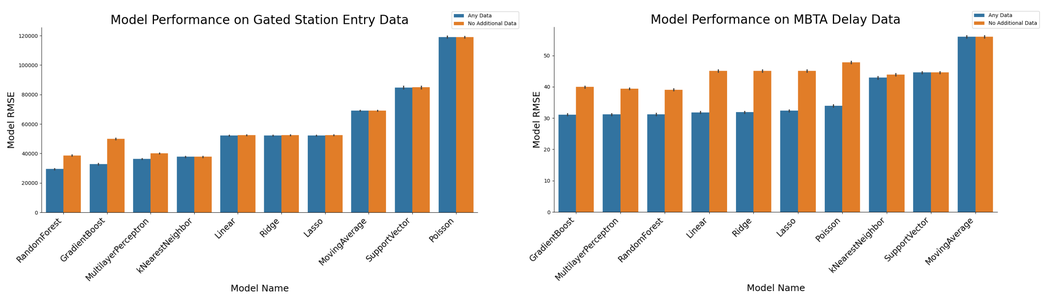
Distribution of Daily Delay Count

Distribution description:

1. Positive and right skewed
2. Captured by a Gamma distribution

Result 1

Random Forest Regression, Gradient Boost Regression, and Multilayer Perceptron Perform Best



Above are the results for the model comparison.
Models are ranked by their maximum performance (i.e., minimum RMSE) across all data blends. Each model has two attributes, displayed above:

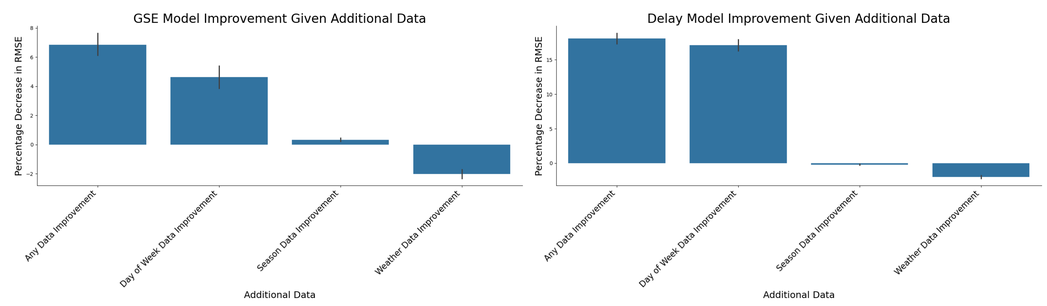
1. The aforementioned maximum performance value is labeled "Any Data."
2. RMSE value given the target metric only as input – labeled "No Additional Data" – is also provided.

The best 3 models are consistently Random Forest Regression, Gradient Boost Regression, and the Multilayer Perceptron. This suggests these models are the most effective for prediction in small-scale data (on the scale of thousands of data points).

Confidence Intervals Interpretation:
95% confidence intervals have been computed (black lines) and are fairly small, suggesting that model performance is relatively stable across bootstrapped datasets. Thus, we have high confidence that model performance rankings are not substantially affected by random variation.

Result 2

Models Improve Given Day of Week (+4.64%, +17.1%)
Models Are Not Affected Given Season (+0.332%, -0.198%)
Models Worsen Given Weather (-2.01%, -1.99%)



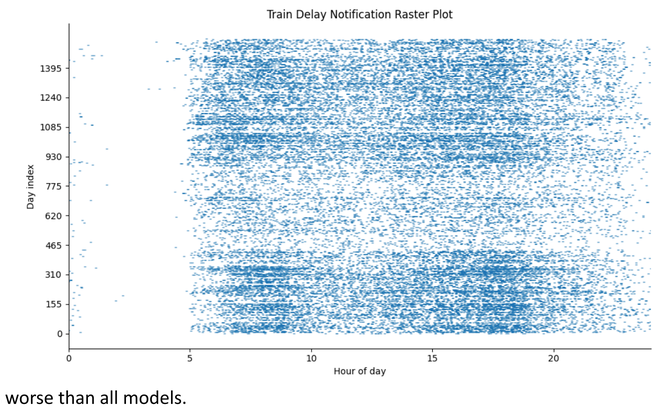
Above are results given different data blends, including 95% confidence intervals (black lines). Each category corresponds to some set of data blends:
"Any Data": Percentage improvement relative to RMSE given only the target metric of minimum RMSE on all tested data blends.
"Day of Week Data": Percentage improvement relative to RMSE given only the target metric of minimum RMSE on data with day of week and target metric.
"Season Data": Percentage improvement relative to RMSE given only the target metric of minimum RMSE on data with season and target metric.
"Weather Data": Percentage improvement relative to RMSE given only the target metric of minimum RMSE on data with weather and target metric.

Interpretation of Graphs/Results:

1. As seen, while improvement is more substantial in delay count data compared to gated station entry data, just including day of week data leads to improvement in both tasks. This improvement is close to the maximum improvement out of all data blends for both, demonstrating the relevance of day of week data to prediction.
2. Including season data has negligible effects on minimum RMSE, suggesting it is relatively irrelevant (likely mostly being discarded by models).
3. Including weather data led to worsening of predictions, suggesting that such data not only does not contain information relevant to predictions but also leads to a tendency of models to overfit.

Result 3

Homogeneous Point Process Model Does Not Perform Better (RMSE > 70)



- Performs worse than all models.
- In the future, on top of timestamp data, we could consider including weather and day of week information to determine if that would enhance model performance.

Interpretation of Results

- For delay count prediction, tree-based models and neural networks (Gradient Boost Regression, Random Forest, Multilayer Perceptron) achieved the lowest RMSE (RMSE <= 31.2), while the Point Process performed the worst (RMSE > 70).
- For gated station entry prediction, Random Forest, Gradient Boosting, and MLP again achieved the lowest RMSE (RMSE <= 36400), with Poisson regression performing the worst (RMSE = 119000).
- Providing all / some additional data in varying formats improved the gated station entry model RMSE by an average of 6.86%, and the delay count model RMSE by an average of 18.1%.
- Day-of-week features alone created a 4.64% average improvement in gated station entry prediction and a 17.1% average improvement in delay count prediction.

Conclusions

- Surprisingly, the models showed no performance boost with the addition of weather data (pressure, wind, precipitation, temperature, etc).
- As demonstrated, calendar data and seasonality dominate MBTA delay and gated station entries, while weather shows minimal effects at the daily aggregation level. Future work may want to explore more extreme events.
- Ensemble tree models (Gradient Boosting, Random Forest) and neural networks (MLP) consistently outperform other models and regressors, suggesting complex and non-linear patterns across the MBTA system.
- Through our research, riders can be alerted in advance of likely high-demand days. City planners and policymakers can gain quantitative evidence on which calendar factors (day-of-week, seasonality) drive ridership and delays, which can inform budget allocations and long-term investments.

Limitations/Future Steps

- Aggregating to daily counts may remove sub-daily patterns. Smaller intervals (15- or 30-minute) data could capture more complex patterns.
- We modeled system-wide totals rather than station- or line-level. Localized disruptions may be invisible when aggregated.
- In the future, we could incorporate holidays, special events in the Boston area, in order to explain the outliers in the data. The models could also be optimized using hyperparameter tuning through Bayesian optimization, for instance.

References

Acknowledgements

We would like to thank the RISE program at Boston University for the opportunity to conduct research. We are also grateful for the guidance provided by Patrick Bloniasz, Eugene Pinsky, Tharunya Katikireddy, and the other TAs. We would like to acknowledge that we are on the traditional land of the Massachusett and Wampanoag peoples, and we thank them for their continued stewardship of this land. Lastly, we thank our parents for supporting our participation in this program.