Transit Ridership Forecasting

Liam Dubay and Sebastian Lopez | Fall 2024

Overview

Many cities across the U.S. are increasing investment in public transportation in order to reduce greenhouse gas emissions and improve transportation safety. Public transit is getting increasingly expensive, and large transit projects can cost many billions of dollars. To build a more effective transportation system with limited funding, it is important to understand the factors that have the largest effect on ridership. We develop two models to predict public transit ridership per capita for 182 cities based on a number of factors such as vehicle statistics and funding allocation. The first is a bespoke model trained on 28 years of data for each city, while the second is trained on most city data at once.

Data Cleaning and Preprocessing

Data for most U.S. transit agencies from as far back as 1991 are available from the U.S. Department of Transportation's National Transit Database (NTD). We use the time-series summary data from 2023, the most recent available, and select three datasets: Total Funding Time-Series, Operating and Capital Funding Time-Series, and Service Data and Operating Expenses Time-Series. We apply the following cuts to eliminate rural and inactive transit agencies:

- Last Report Year == 2023
- Agency Status == Active
- Reporter Type == Full Reporter
- Reporting Module == Urban

NTD data are available for each transit agency, but for simplicity we combine all available data for each urbanized area (UZA). As the COVID pandemic severely impacted ridership for nearly every city, we exclude the years 2020 - 2023 from our data. This produces a dataset for 182 cities over 29 years (1991 - 2019).

We adjust all monetary values to 2019 dollars. Additionally, we control for population growth by linearly interpolating population estimates from the 2000, 2010, and 2020 Census for each urbanized area. All monetary and service data are calculated per capita, with the exception of fractional data (e.g., fraction of total funding for operations).

Ridership is measured by three different features: Unlinked Passenger Trips (UPT), the number of passengers who board public transportation vehicles; Passenger Miles Traveled (PMT), the cumulative sum of the distance ridden by each passenger;

and total fare revenues. We select UPT per capita as our predictive variable, as PMT will be affected by commute lengths and fare pricing differs between cities, agencies, and modes of transport.

Modeling Strategy

City-by-City Model

For each city we train a model using data from 1991 - 2018 and leaving 2019 as the testing set. Due to a lack of trend with time we create a baseline model based on the Native Forecast Gaussian Walk model. For the model we implement Lasso regression to select the features that influence UPT the most. We specifically use LassoCV that performs a 5-fold cross-validation to find the best model and tune the alpha hyperparameter. Our key performance indicator is simply how many times the difference between the model and the actual value is lower than the difference of the model and the baseline.

All-City Model

We train a second model on all cities at once, after randomly selecting 20% of cities to serve as our testing set. We include data from 1991 - 2018 in our training set, holding aside 2019 data for all cities for further validation.

We perform a 5-fold cross-validation test between five different regression models:

1. A baseline model which simply predicts the mean of the ridership data;
2. Single linear regression on total funding;
3. Multiple linear regression (MLR) on all features;
4. MLR with cross-validated lasso feature selection (LassoCV); and
5. XGBoost regression tree

The root mean squared error (RMSE) is our key performance indicator. Of the five models, XGBoost has the lowest RMSE and standard deviation. The MLR and MLR+LassoCV models are identical because LassoCV did not eliminate any features.

We use a grid search to tune the XGBoost hyperparameters. We use a learning rate of 0.1, a maximum depth of 3, and 50 estimators.

Model	RMSE	St. Dev.
Baseline	23.4	1.4
SLR	13.0	1.0
MLR	10.4	0.46
MLR+LassoCV	10.4	0.46
XGBoost	5.82	0.39

Results

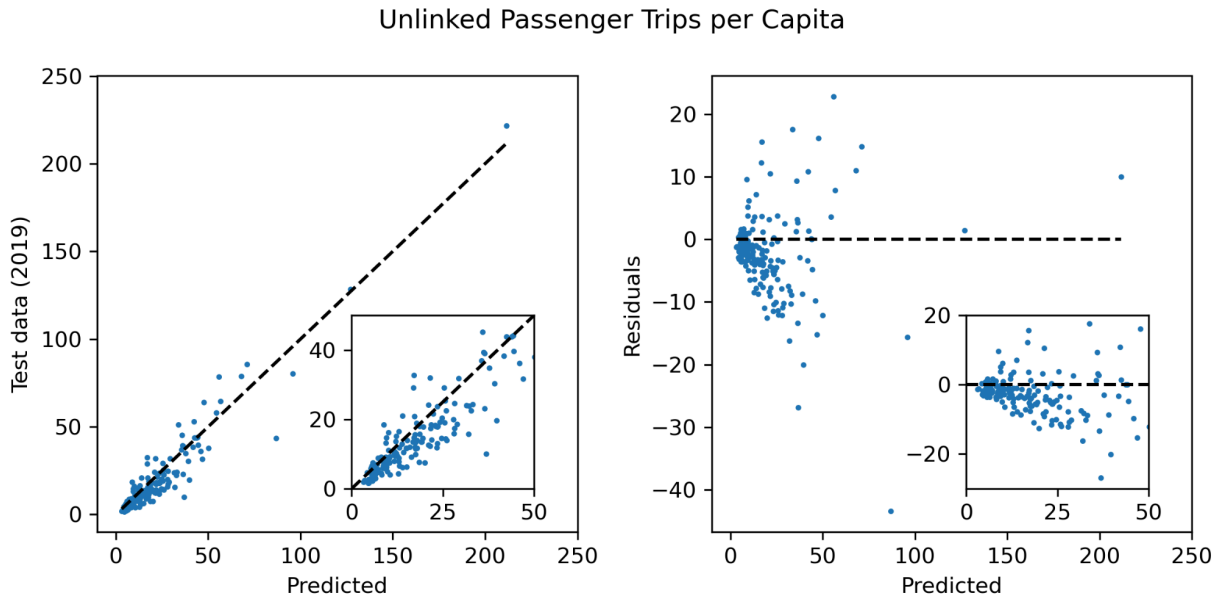
City-by-City Model

We test the LassoCV model against the 2019 values for each of the cities. We find that for 182 cities the model performs better than the baseline for 36 cities. For the remaining cities the baseline is closer to the actual values indicating our model does not accurately describe the 2019 UPT values.

All-City Model

We test the XGBoost ridership predictions against the remaining 20% of data from 1991 - 2018. The XGBoost model yields an RMSE of 9.23, compared to the baseline with an RMSE of 21.2. It tends to perform worse for cities with ridership >~50 UPT per capita, but that is a small minority of all cities.

We then extend our all-city model to predict 2019 ridership for all 182 cities. We achieve an RMSE of 7.44, but once again our model is out-performed by the naive baseline with an RMSE of 0.95. The plot below compares the XGBoost predictions to the test data and residuals. The lower-right inset shows some trend between the residuals and predicted ridership, indicating that the all-city model does not fully capture the 2019 data.



In summary, the all-city model performs better than the city-by-city model thanks to the much larger training set. However, its performance may be skewed by a handful of cities with very high ridership.

Limitations and Future Work

Our work has revealed the difficulty of forecasting public transit ridership from year to year. There are many factors at play, not all of which are captured by even our extensive dataset. The fact that our models were consistently out-performed by a naive baseline prediction shows that it is very hard to move the needle on transit ridership, even in cities with robust public transportation systems.

There are a number of ways our models could be extended in the future. Our dataset does not include crime and safety statistics, which could have an effect on ridership. We could also better account for delayed correlation between features - for example, transit projects can take many years to complete, so ridership may correlate more strongly with capital expenditures several years in the past. Cities of different sizes may need unique transit solutions and by extension unique models. Finally, data which are not included in the NTD, such as population changes and local developments, might be necessary to improve the model accuracy.