Modeling King County Bus Ridership

Maggie Stark, Hannah Murphy, Alexander Van Roijen, Jacob Warwick

MOTIVATION

Bus route planning and optimization is an integral part of city planning. However, no model or data sources exist to measure total ridership across the King County Metro system.

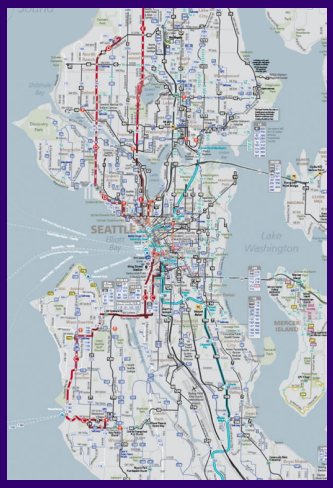
- ORCA payment card transactions are available, but only measure riders who pay with Orca
- Automatic person counter (APC) measures all riders but is only present for 60% of trips.

Goal: Predict total ridership (APC) across route, direction, and time of day using Orca transactions and route metadata.

DATA

- **21.5M Orca Transactions**
- **APC Ridership Counts**
- **Approximately 170 Routes**
- **Route Direction & Geographic Region**
- **Season**
- **Day of Week**

http://kingcounty.maps.arcgis.com/apps/webappviewer/index.html?id=3e239c9048604de8a1c73b72679bc82e



METHODOLOGY

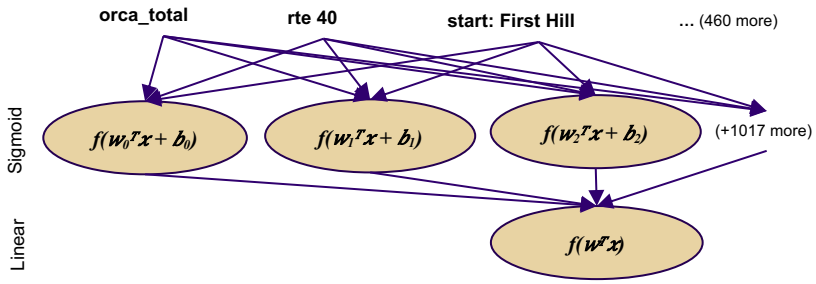
Data Source: Orca farecard transactions and automatic person counters (APC) present on about 60% of buses in the survey period.

Pipeline: filtered Orca and APC data to relevant dates, routes, and trip ids; merged the Orca and APC data sets; created features based on the type of Orca account, geographic region, season, and day of the week; and aggregated data into 15 minute, 30 minute, and 1 hour intervals for modeling.

- Snow days removed; data represent atypical ridership
- RapidRide data removed due to data collection issues

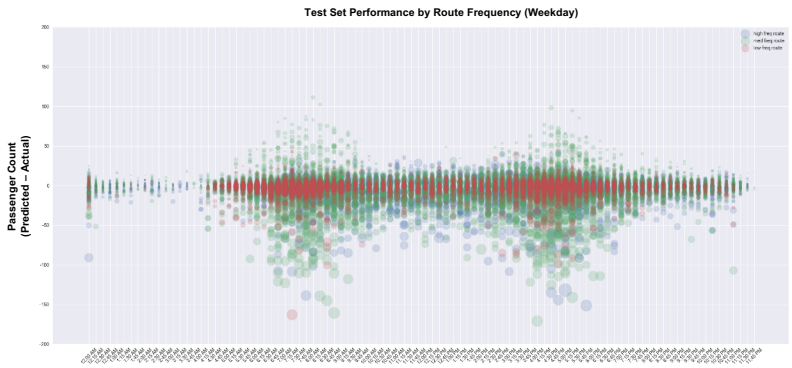
MODEL SELECTION

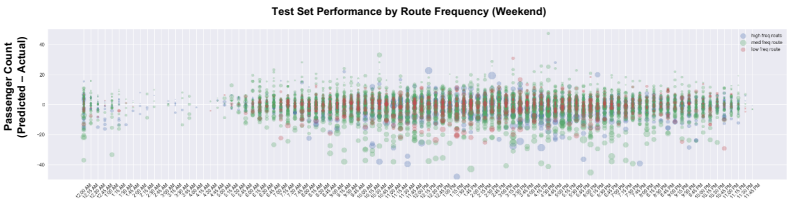
After evaluating networks, gradient-boosted trees, kernel SVM, and gaussian process regression, the best performing model was a shallow, wide neural network with sigmoid activation. This model was trained using SGD/Nesterov and 23.5% dropout for 13 epochs.



MODEL PERFORMANCE

Test set mean absolute error: **5.5077** passengers per route / direction / 15 minutes

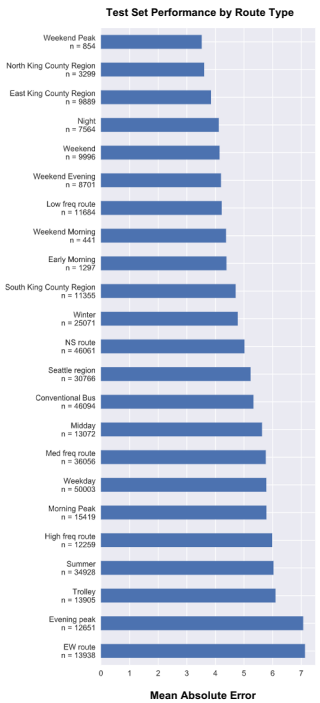




Left: Residuals for weekdays and weekends. Our model is less accurate during weekday peak usage times (morning and evening commutes).

Route Frequency Definition:
Low (red): <1 trip per hr
Med (green): between 1 trip per hr and 1 trip per 15 mins
High (blue): 1 or more trip per 15 mins

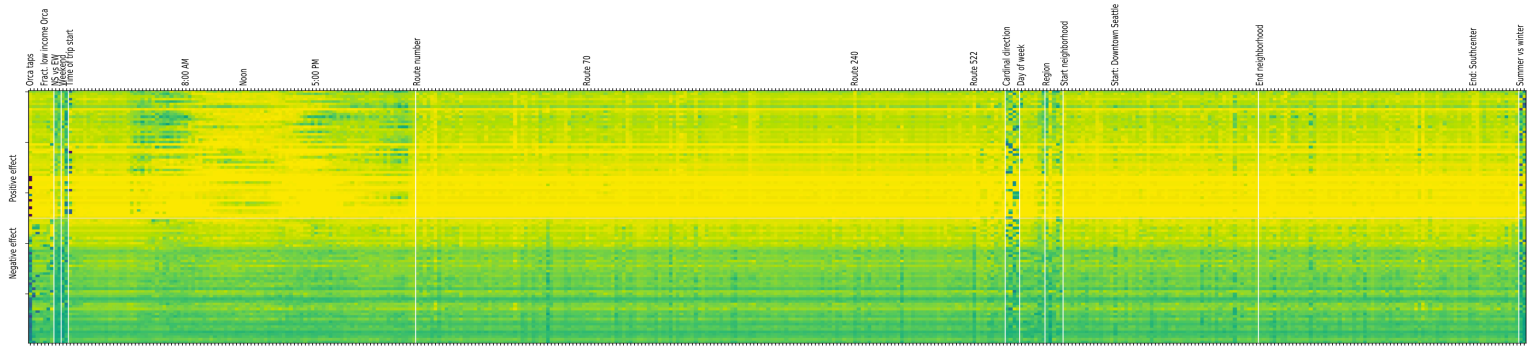
Right: Mean absolute error by route type. Routes traveling east-west and during the weekday evening peak were the worst performers, with a mean absolute error of 7.



MODEL ANALYSIS

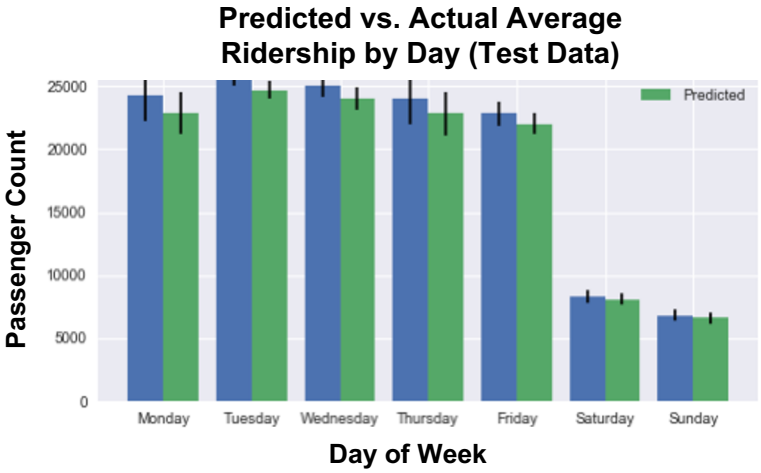
We subtracted the bias terms from the first layer input weights and re-applied the sigmoid function to assess per-feature prediction effects within each neuron.

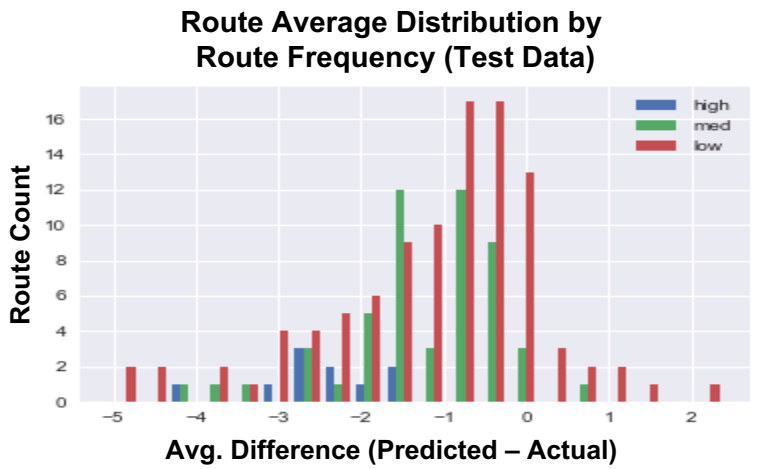
- The model learned to differentiate peak weekday/weekend times
- More subtle interactions exist between start and end neighborhoods, route number, direction, season, and other features.



Top 50 positively and negatively contributing neurons and the effect of each feature on the neuron's prediction.

TOTAL RIDERSHIP ESTIMATION





CONCLUSIONS & FUTURE WORK

We have created a model that predicts total ridership over the King County Metro bus system from transactional data with high accuracy. This will allow planners to create ridership estimates for individual routes, directions, regions, and times, to better predict system usage and plan for future expansion. We hope to see this model used to address issues of equity, access to transit, and system planning.

Future work could use this model as a basis to predict rider destination. Once reliable data from RapidRide are made available, we hope this model could be trained to predict those routes. As data becomes available, we propose that a similar model could be applied to transit systems in Pierce, Kitsap, and Snohomish counties.

ACKNOWLEDGMENTS

We would like to thank Mark Hallenbeck at TRAC for his time, patient explanations, and expertise, as well as Dmitri Zyuzin for his help gathering the data for this project. We would like to thank Megan Hazen for her guidance in this capstone project and for her expertise on neural networks.