PVTA: Machine Learning for Transit Ridership Prediction

Anushka Basu, **Ian Birle**, **Disha Singh**
Centre for Data Science
University of Massachusetts Amherst

1. Introduction

The PVTA is a Regional Transit Authority that provides fixed route bus service, Dial-A-Ride and free paratransit service (transportation for seniors and people with disabilities who cannot use the regular, fixed route transit service), in 24 member communities in the Pioneer Valley.

PVTA is trying to better understand how micro-transit solutions could more effectively serve areas where riders are challenged with limited-service options. While in some areas, there are geographical or infrastructure restraints, there could be various reasons why certain areas, often more rural areas are underserved (having service as low as once per hour). In order to understand if these underserviced areas are potential regions for micro-transit, there is a need for thorough demographic data analysis and closely probe the communities here that can be catered through micro-transit.

In this work, we identify service areas with significant gaps in service. Additionally, we build a predictive model to find regions where ridership has the potential to grow given service enhancements, consequently, finding demographic features most influential in determining those ridership numbers. This understanding will help cater to areas with a more informed, people-driven approach. Essentially, we use QGIS and Tableau to show the correlation between the features and the ridership density.

2. Our Work

We perform our study on block group level wherein the ridership density for each block group was calculated by adding the ridership values for all stops that are within 3/4th mile distance of that block group. We used the PVTA provided stops shapefiles to calculate block group level ridership density.

2.1 Data Collection: We used PVTA provided shapefiles for stops consisting of stop level ridership information. We downloaded Block Group shapefiles and all other census data from the public website https://www.census.gov. We used a public API which allows easy download of user provided columns referenced by their Column name from any table in the census database.

2.2 Finding Underserved areas: We visualized the ridership density of PVTA serviced area over the geographical map of the Pioneer valley in QGIS and overlayed it with routes that have frequency of once per hour or more than an hour. We then manually select only those low ridership areas that are mainly serviced by low frequency buses reach. See Fig. 1

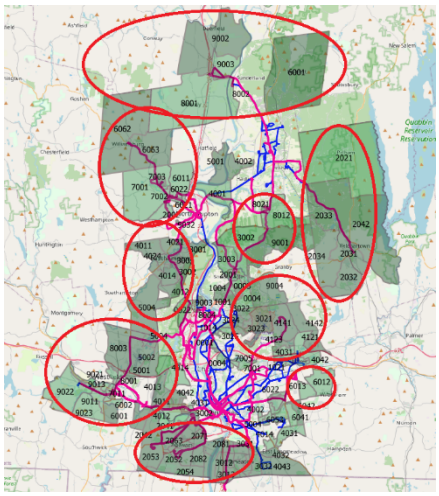

Figure 1: Green indicates low ridership. Pink routes indicate low frequency routes. Blue routes are all PVTA service routes, overlapped by the pink ones. The red circles are the manually selected potentially underserved areas with low ridership density

2.3 Ridership Prediction: We use 85 features from the Census data and feature engineer them into 54 most relevant features for our use case. We use Random Forest as our predictor in order to retain the interpretability of our results. We remove the '% of people using public transportation' column from our data as we are trying to predict the ridership for areas when the service is hypothetically high, hence we will never have the values for that column for our prediction samples. Instead, we take a generic column 'well served' which equals 0 for all

underserved areas and 1 for all well served areas. We toggle this column to 1 while predicting the ridership for all underserved areas. The model gives us the most relevant features influencing riderships in the form of feature importances. See Table 1.

2.4 Data Visualization: We draw gradient based maps in QGIS for all these features to inspect what feature dominates which block groups and the correlations it might have with true/predicted riderships. Finally, we consolidate all analysis in a Tableau Dashboard.

3. Conclusion & Future Work

PVTA considers this work as a foundation to what lies as a long road ahead. The dashboard can be extended to all features and understand which social indicator has the potential to govern ridership for a block group. Moreover, using transit data available at PVTA, it can be deduced which towns have high outgoing populations and to which town. From there, one can predict ridership for surrounding rural regions, the ridership that wants to be connected to a fixed route, but due to reasons (which can be derived from the dashboard) they are unable to do so. This strategy can be used by PVTA to strategize and extend micro-transit facilities in rural areas around ridership hubs.

Feature	**Importance**
Total:!!Car, truck, or van:!!Drove alone	0.244
Residential - single family_land_use_%	0.181
poverty_families	0.144
Commercial_land_use_%	0.120
Median age --!!Male	0.058
Population per SQFT	0.057
Total:!!Black or African American alone	0.050
Total:!!Bicycle	0.049

Table 1: Random Forest Feature Importances