BUS RIDERSHIP PREDICTION: A MACHINE LEARNING APPROACH

By

BRANDON BULLARD

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE IN APPLIED ECONOMICS

WASHINGTON STATE UNIVERSITY
School of Economic Sciences

DECEMBER 2021

To the Faculty of Washington State University:

The members of the Committee appointed to examine the thesis of BRANDON BULLARD find it satisfactory and recommend that it be accepted.

Mark Gibson, Ph.D., Chair

Jake Wagner, Ph.D.

Alan Love, Ph.D.

ACKNOWLEDGMENT

I would like to extend my sincere thanks to Dr. Wagner for countless hours of mentorship throughout the development of this thesis. If I had questions or wavering confidence, Jake was always there to provide guidance and encouragement. Without his help, this thesis would not have been possible.

I would like to thank Dr. Gibson for being an excellent professor throughout my undergraduate and graduate studies at Washington State University. I truly appreciate him putting his trust in me and acting as committee chair when I decided to pursue the graduate program.

I would like to thank Dr. Love for being an all-star professor and overall gem of the human race. His expertise and compassion are evident in how he taught econometrics in Python alongside an entirely remote class.

To Dr. Iles, now Senior Lecturer at the Federation Business School in Australia, thank you for allowing me to serve as your research assistant my junior year. As a first-generation college student, I would not have known what graduate study entailed if not for our conversations that summer.

BUS RIDERSHIP PREDICITON: A MACHINE LEARNING APPROACH

Abstract

by Brandon Bullard, M.S.
Washington State University
December 2021

Chair: Mark Gibson

Bus ridership is a key component of transit systems nationwide and increasing the share of bus ridership is an important part of reducing externalities like congestion, pollution, and traffic accidents. However, bus ridership has been on the decline in recent years as personal automobiles remain the most popular mode of transport. In order to equip transit authorities with the information they need to make routing, location, and service decisions to induce demand, a predictive ridership demand model is developed. This model will serve as the foundation for a transit planning decision support tool. With the advent of automated passenger count (APC) systems and growing data availability in the transportation sector, machine learning methods are increasingly viable for prediction problems. This paper explores how best to develop a machine learning model for predicting bus ridership within a rural transit network.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Transit providers face many choices in the planning and operation of their routes and are often forced to make tradeoffs when facing competing interests. Transit users want quick, reliable, and cost-effective service, while transit providers want to maximize revenues by providing service to the greatest number of riders in the most efficient manner. In the literature this dichotomy is presented as the Transit Networking Problem (TNP) (Desaulniers and Hickman 2007). The TNP is multi-faceted and considers all variables within a transit provider's control including placement of stops, driver scheduling, route design, and route frequency, among others.

For a transit authority to make decisions within the constraints of the TNP, a decision-support tool based on a predictive ridership demand model can prove useful. Such a model will allow transit planners to understand the predicted changes in ridership caused by both internal and external factors. Internal factors relate to characteristics within the control of the transit authority like service frequency and placement of stops. External factors capture characteristics of an area such as population, income and walk quality among many others (Frei and Mahmassani 2013). Considering the growing amount of data available to researchers, machine learning methods are becoming a popular approach when faced with prediction problems. This thesis explores how to develop a framework that utilizes an optimal machine learning algorithm to predict bus ridership at the stop level by incorporating internal and external factors at existing bus stops.

Nationally, in place of public transit people are increasingly reliant on personal automobiles, and this trend is greater in rural areas (Rural Transit Fact Book, 2021; Analysis of Recent Public Transit Ridership Trends, 2020). Given the externalities associated with growing automobile use like pollution, congestion, and traffic accidents, it is important to grow the share of public transit users.

This research examines Pullman Transit (PT), the leading rural transit system in Washington. PT provides over 1.4 million rides annually with most of the ridership being students due to the nearby Washington State University (WSU). Like most transit networks, Pullman Transit is faced with challenging questions in their effort to meet service demands and increase ridership in a financially sustainable way. Where should bus stops be located? How frequently should each stop be serviced? Which routes should be driven to ensure all stops are serviced while minimizing rider travel time? What should the rider fare be? The answers to these questions are vital to the smooth and efficient operation of any transit network. To help answer these questions a machine learning ridership demand prediction model is developed using over 700,000 records from PT's Automated Passenger Count (APC) system.

CHAPTER TWO: LITERATURE REVIEW

Predicting transit demand has proven challenging, as ridership and service levels work in concert (Taylor 2008, Dill 2013, Beberri 2021, Boisjoly 2018, Chen 2015). The supply of transit induces demand, just as peak commuting times induce greater transit availability. Beberri et al (2021) uses a Poisson fixed-effects model to estimate the elasticity of ridership demand with respect to frequency. Frequency is measured as the number of stops on a route, and ridership demand is given by the sum of boardings and alightings (at each stop/on each route?). Using local stop-level data they find increased service frequency results in increased ridership, but that there are diminishing returns where a route is already popular.

There are a wealth of papers examining which additional variables are most important in predicting ridership. Taylor et al. (2008) uses two-stage simultaneous equation regression models of data from hundreds of urbanized areas throughout the U.S. The researchers investigate the effects of transit supply on demand as well as which variables had the most influence. They examined geographic, economic, population, and auto system characteristics and found that population, household income, percent college students, recent immigrants, and carless households to be important in explaining levels of ridership.

Chakrabarti (2015) focuses on how transit reliability affects ridership and finds that routes with greater adherence to an established schedule is associated with greater ridership. This effect is more pronounced on routes with greater headways presumably because of the higher consequence of missing a route.

An advantage to using disaggregated stop-level data is the ability to explore how the built environment around a stop influences ridership. Chakour and Eluru (2016) examine the city of Montreal to determine how both stop level infrastructure and the built environment influence bus ridership. While they also find that transit service characteristics like frequency and accessibility have the greatest impact, enhancements to the land like parks have a small but positive impact and inhibitors like major roads have a negative impact. With respect to spatial measurement of built environment variables, Pulugurtha and Agurla (2012) utilized spatial modeling methods to capture several attributes surrounding bus stops. They found that a quarter-mile buffer distance yielded the most meaningful estimates on ridership, and many papers that have followed use the same heuristic when gathering spatial data (Dill 2013, Chakrabarti 2015, Li 2020). With the advent of Automated Passenger Count and Geographical Positioning Systems (GPS) there is much greater data availability at the stop and route level. These systems are primarily used by transit authorities to evaluate changes in performance, but researchers can also use this technology to estimate demand at a much lower level of aggregation than previously available. Given this recent change, some of the earlier literature aggregates over the course of a day or entire route. This level of measurement also characterizes the environment with census tract levels, which averages about 4,000 inhabitants. At lower levels of aggregation, researchers have found smaller elasticities with respect to transit service characteristics like frequency and headway. Frei and Mahmassani (2013) for example estimate ridership using stop-level transit data using data from Chicago, Illinois transit system. Their paper finds much lower transit service elasticities with respect to ridership when comparing their results to similar studies at larger levels of aggregation.

Machine learning methods have been used in ridership prediction problems. Kawatani et al (2021) utilize gradient boosted decision trees for predicting bus travel time. Google is also using machine learning methods to predict bus delays (Fabrikant 2019). With respect to ridership prediction, Fontes et al. (2020) uses a neural network to predict bus ridership in European metropolitan areas based on weather conditions, finding improved model results when weather conditions are included. Li (2020) utilizes machine learning models in predicting ridership for hypothetical bus stops in the state of Delaware. In their analysis they find that jobs, the percent of people below the poverty line, and carless household features to have greater feature importance than built environment characteristics.

Contribution

The contribution of this paper comes from a few different areas. First, many papers utilizing stop-level ridership use data from metro areas with large ridership, whereas PT is in Pullman, WA, which has a population of about 34,000 (U.S. Census Bureau 2019). From this population most of the residents of the city are also students; total enrollment in 2019 at WSU's Pullman campus was 21,000. Many metropolitan areas also have other means of public transport, and changes to service characteristics in these other modes don't appear to be controlled for when analyzing bus ridership. Second, the characteristics of riders in metropolitan areas are different than this sample. Pullman's population is comprised primarily of students and people employed by the university, whereas other cities have a larger variety of commuters. While it is hard to derive riders' objectives for using public transit, metropolitan areas likely service people going to and from work, whereas PT services a large student population that does not directly pay for each ride. From a methodological approach, this thesis uses a bagging approach in contrast to the boosting methods in Kawatani et al and Li (2021;2020). In summary, this paper uses data from a small college town with a single bus network in contrast to the larger areas of study examined in other papers. In addition to having a smaller population, the characteristics of this population are different than other small towns and metro areas due to WSU's presence.

CHAPTER THREE: DATA

Tables 1-1.2 at the end of this chapter provide detailed descriptions of the data below. Table 5 in the appendix provides the accompanying summary statistics.

Ridership

The transit system data used in this analysis was provided by PT in location and service files. Location data includes the latitude, longitude, and the names of 223 bus stops, while the service data details boardings and alightings, time of day, stop name, and the corresponding bus name and route. After performing a merge between the location and service files, there were 50 stops that did not have exact coordinates. These coordinates are crucial for gathering additional data surrounding the bus stops, so Google's Places API was utilized to fill the missing values. From the complete dataset, boardings and alightings were summed at each stop by the hour, creating the ridership dependent variable *rid.* This sum is necessary to avoid an asymmetry problem in how the data is generated; some stops are used as drop off locations and as a result record zero boardings. An examination of the ridership data in figures 1-4 reveals there are spatial and temporal trends. In this analysis hourly aggregation is used to capture changes throughout the day that may impact ridership. Table 1 illustrates summary statistics of ridership at hourly and daily aggregations.

Figure 1: Hourly ridership 2019-2020

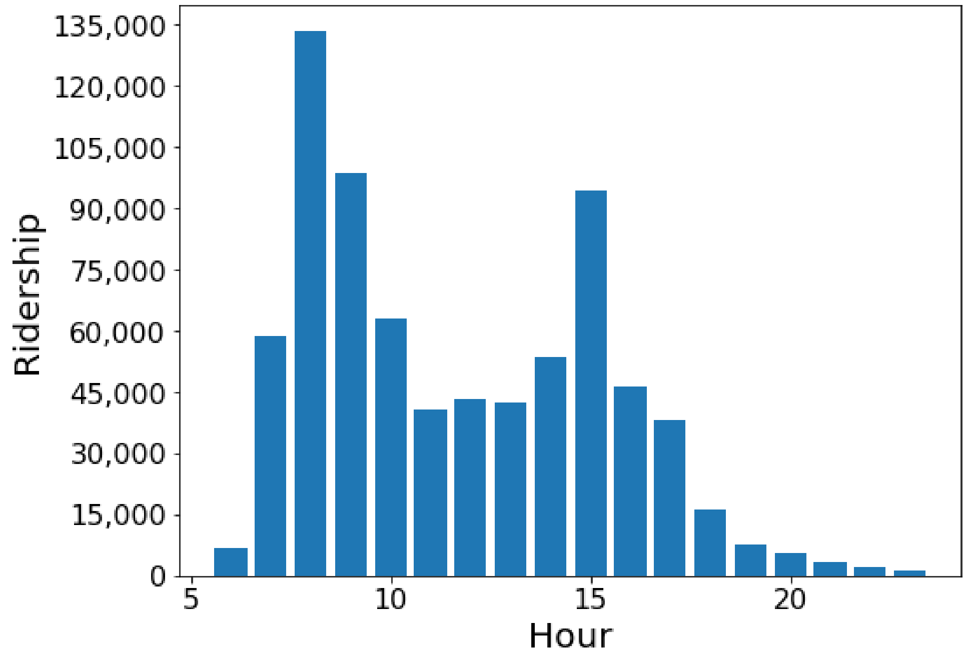


Figure 2: Weekly ridership 2019-2020

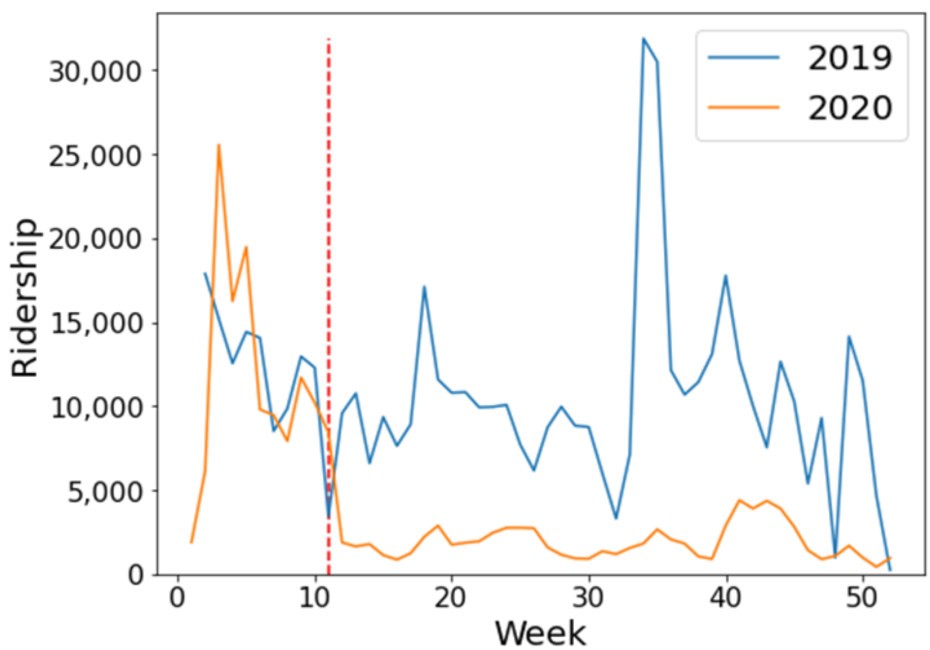


Table 1: Ridership summary statistics

Aggregation	Mean	Std	Median
Hour	2.68	8.36	0
Day	15.64	45.72	2

Figure 1 illustrates how ridership varies by the hour and confirms that ridership has two distinct peaks in the morning and afternoon. Both morning and afternoon peaks are a function of riders leaving to and from campus or a place of work. Figure 2 shows how ridership changes throughout the entire sample. First, the Covid-19 pandemic had an immense negative effect on ridership; the dashed line illustrates when WSU transitioned to online classes in response. This analysis will focus solely on 2019 ridership because it more closely resembles normal ridership levels. Before this negative shock, ridership appears to follow the flow of students according to WSU's 15-week semester system. Semesters begin in January, June, and August, with most students enrolling in the August semester. Ridership decays after its peak in August from student attrition, and the large negative troughs coincide with school holidays. Peaks throughout both semesters are hypothesized to coincide with examinations and the beginning of new semesters in the fall, spring, and summer. The cyclical nature of ridership is the primary motivation for controlling for time. A map of Pullman is below, illustrating how ridership is distributed spatially. The maps of boardings and alightings demonstrate the necessity for summing both; large points on one map occasionally don't correspond on the other.

Figure 3: Map of alightings 2019-2020

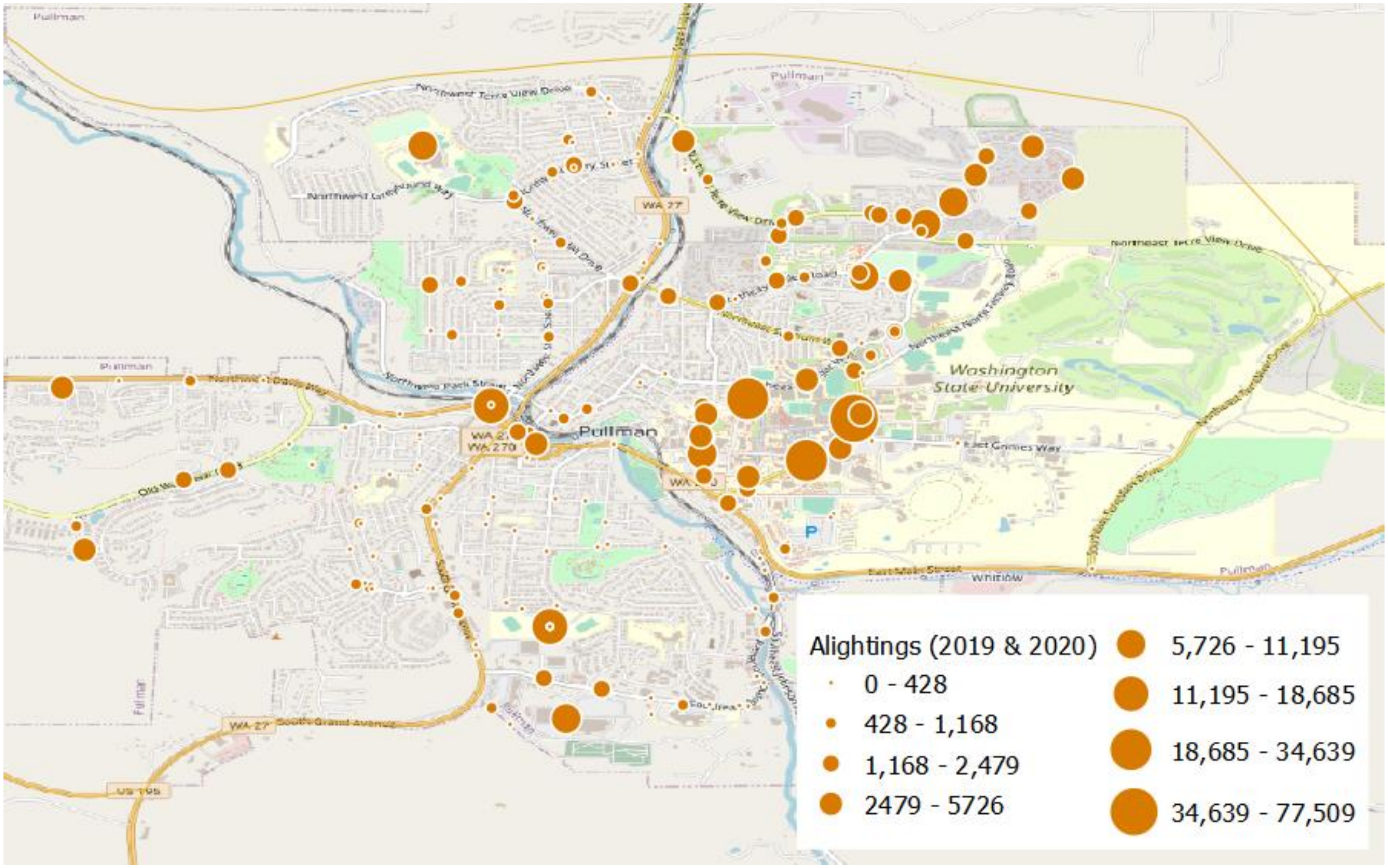
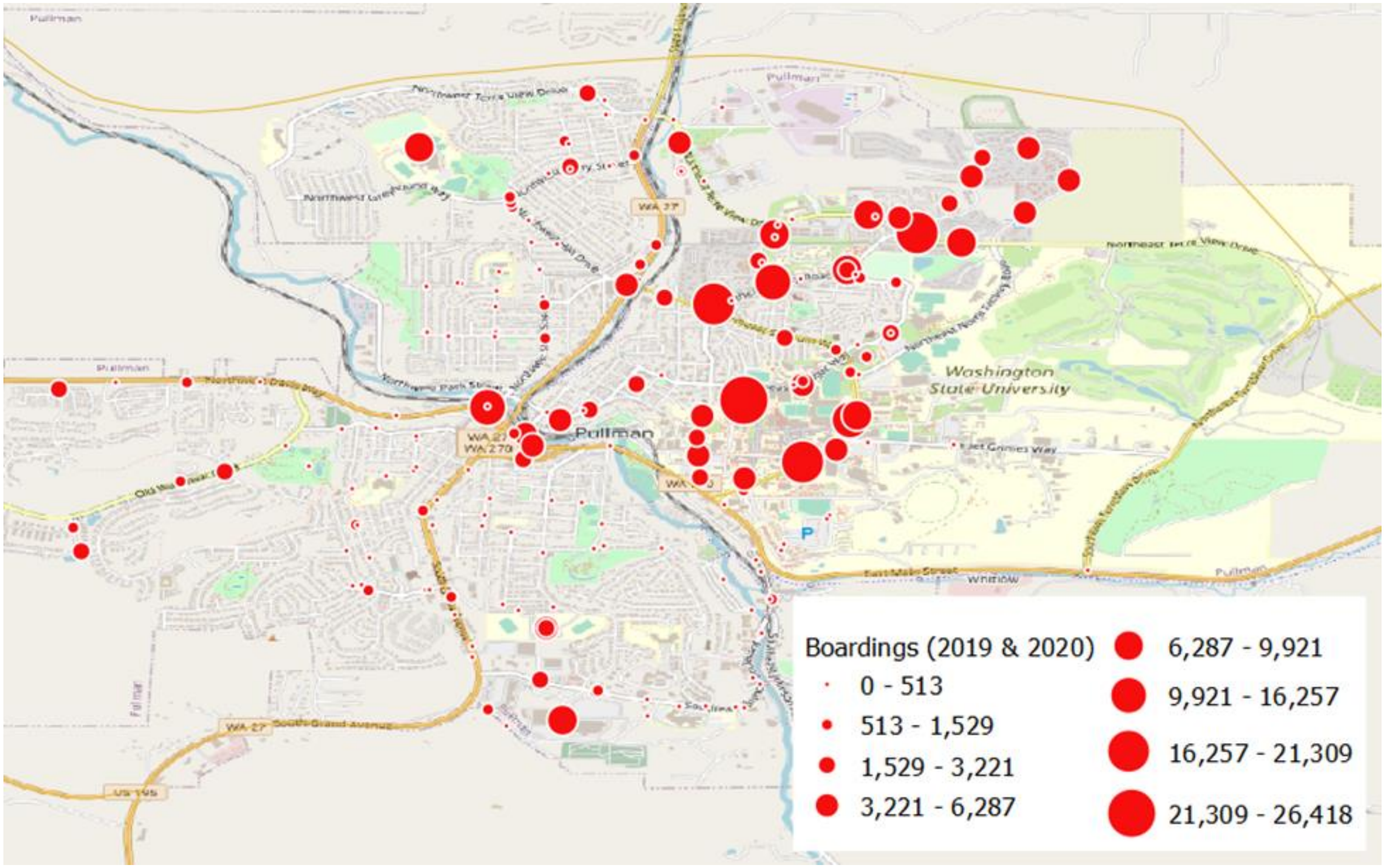


Figure 4: Map of boardings 2019-2020



Bus Network

PT operates in the city of Pullman Washington and includes 223 stops and 41 routes. These stops and routes have mixed purposes, transporting elementary school students, college students, and also serving the broader community. This analysis will focus on weekday ridership for the entire bus network since all busses are shared despite the variety of uses. Characteristics of the bus network are important since the transit authority can change them, and they also have the greatest direct influence on ridership (Berrebi 2021). Frequency is a variable that captures the number of times a bus services a stop aggregated by the hour. Each stop also has a travel time and distance for a bus or car to reach locations of interest. To identify locations of interest, a count of alightings were aggregated by stop. After identifying the most popular stops for alightings, the Google Maps API was used to calculate the time and distance necessary to reach this location of interest. Distance and time are two of the most important factors when choosing mode of transport, and these controls are designed to capture this effect. To better understand how they work, table 2 below provides an example.

Table 2: Travel time and distance controls

Prefixes	Description
busTime/driveTime	Travel time in seconds for bus or car
busDist/driveDist	Distance in meters for bus or car to travel
Morn/Aft	Morning or afternoon
Origin/Dest	Whether time/distance is to or from a location
Example	
busTimeAftOriginChinook	Time for bus to travel from the Chinook recreation center to this bus stop in the afternoon
Location	**Description**
Beasly	Beasly Coliseum
Sloan	University building
GrandMain	Intersection of streets Grand and Main
Spark	University building
Dissmores	Grocery store
TerreViewFairway	Intersection of streets TerreView and Fairway
Vogel	University building
Walmart	Grocery store
SRC	University building
ValleyStadium	Intersection of streets Valley and Stadium
CUB	University building
Safeway	Grocery store
Merman Valley	Intersection of streets Merman and Valley
SEL	Place of work, engineering firm
Highschool	Pullman High School

Socio-Demographic

Socio-demographic data is often used when analyzing bus ridership trends. For instance, population density near a stop is believed to have a direct relationship with ridership. Most socio-demographic variables used in this analysis come from the U.S. Census Bureau American Community Survey. There is a total of 12 types of census bureau data used in this analysis describing different characteristics of each respective block group. Rationale for the inclusion of these variables is included in the literature review, but they are broadly thought to have a relationship with ridership. From this selection other characteristics were created to explore each block group further. For example, the count of unemployed people divided by the labor force yields the unemployment rate for each block group. Employment characteristics are thought to be especially important for predicting ridership given that public transport serves as a means of commuting. Additional data that describes the number of jobs at the block group level comes from the U.S. Census Bureau at Longitudinal Employer Household Dynamics (LEHD) webpage. This data was enumerated by the 2010 census block.

Environmental

To control for the effects of the natural and built environment on ridership at each bus stop, three different data sources were utilized. Walkability and bike-ability are important factors when considering mode of transport. Those factors include the presence of sidewalks, bike lanes, and the distance and density of amenities or locations of interest nearby. The WalkScore index, developed by a private company of the same name, provides a number from 0-100 for any address summarizing these factors. For each bus stop, a WalkScore and BikeScore index are obtained through their API. Another feature at each stop is seating and shelter. These variables were provided by PT and are thought to be positively associated with ridership. Another type of environmental variables considered in this analysis is weather. This data was gathered from the National Oceanic and Atmospheric Administration, measuring hourly precipitation, wind speed, daily snow, and daily snow depth. The last set of environmental data considered are counts of types of places near bus stops. Google maps API was utilized to count the number of cafes, grocery stores, etc. within 1/8, 1/4, and 1/2 mile radiuses around each stop. Table 2.1 below this section details which places are used.

Table 2.1: Data Descriptions

Variable	Measurement	Description
rid	Stop	Sum of boardings and alightings
income	Census block	Average income
population	Census block	Total population
incomePovertyRatio	Census block	Income to poverty ratio
degreePopOver25	Census block	Number of people with a degree
enrolledOver3	Census block	Number of people enrolled in school
ownerNoVehicle	Census block	Number of homeowners without a
renterNoVehicle	Census block	Number of renters without a vehicle
owner	Census block	Number of homeowners
renter	Census block	Number of renters
employed	Census block	Number of employed (place of
unemployed	Census block	Number of unemployed
labor_frc	Census block	Employed + Unemployed/Population
med_age	Census block	Median age
med_house_val	Census block	Median house value
n_jobs	Census block	Total number of employed people
n_jobs<29	Census block	Total number of employed people
n_jobs30_54	Census block	Total number of employed people
n_jobs>55	Census block	Total number of employed people
walkscore	Stop	Walkability
bikescore	Stop	Bike-ability
stopRouteVehicleFreq	Stop	Number of times bus services bus
stopRouteFreq	Stop	Number of time route services bus
stopFreq	Stop	Number of busses servicing bus stop
Shelter	Stop	1 if shelter, else 0
Simme Seat	Stop	1 if seat, else 0
gas_cpi	U.S. City avg	Gas Consumer Price Index
gas_pct_diff	U.S. City avg	Percent difference in gas_cpi from
daily_snowfall	City Weather	Daily snowfall (Inches)
daily_snowdepth	City Weather	Daily snow depth (Inches)
DailyDryBulbTemperature	City Weather	Daily temperature (Fahrenheit)
DailyPrecipitation	City Weather	Daily precipitation (Inches)
DailyWindSpeed	City Weather	Daily Windspeed (MPH)
Month	Time	Month of observation
Week	Time	Week of observation
DOY	Time	Day of year
DOW_num	Time	Day of week number (0 – M, 4 – F)
DOM	Time	Day of month
DOW_Sunday, DOW_Monday, DOW_Tuesday, DOW_Wednesday, DOW_Thursday, DOW_Friday	Time	Day of week controls

Table 2.2: Locations of interest count

Variables	Measurement	Description
cafeCount(100.5,402,804)	Stop	Number of cafés within 1/2, 1/4, 1/8 miles
churchCount(100.5,402,804)	Stop	Number of churches within 1/2, 1/4, 1/8 miles
restaurantCount(100.5,402,804)	Stop	Number of restaurants within 1/2, 1/4, 1/8 miles
transitstationCount(100.5,402,804)	Stop	Number of bus stops within 1/2, 1/4, 1/8 miles
universityCount(100.5,402,804)	Stop	Number of university buildings within 1/2, 1/4, 1/8 miles
supermarketCount(100.5,402,804)	Stop	Number of supermarkets within 1/2, 1/4, 1/8 miles
department_storeCount(100.5,402,804)	Stop	Number of department stores within 1/2, 1/4, 1/8 miles

CHAPTER 4: METHODS

The methods described below are used in a “Pipeline”; a series of steps that prepare the data for analysis by a model. First, the training data is split with 80% of the data used for training, and the remaining 20% for testing. In this split the stop names are separated so the model can be evaluated at unseen locations. Cleaning is the next step, and this simply involves removing the stop names, and datetime variables from the dataset, as well as imputing missing values. There are still controls for time like day, and day of week, but the datetime variable provided by PT does not work in the scikit learn package. For the machine learning models exclusively, the next step is feature generation. Here, each of the census variables are squared and interacted with one another to see if these transformations provide a better fit. Next is regularization, where LASSO is applied to reduce the number of features in the machine learning models. Finally, the estimator and hyperparameter spaces are selected based on bayes search cross validation and the performance of the model is observed. The regularization and hyperparameter tuning steps only apply to the machine learning models.

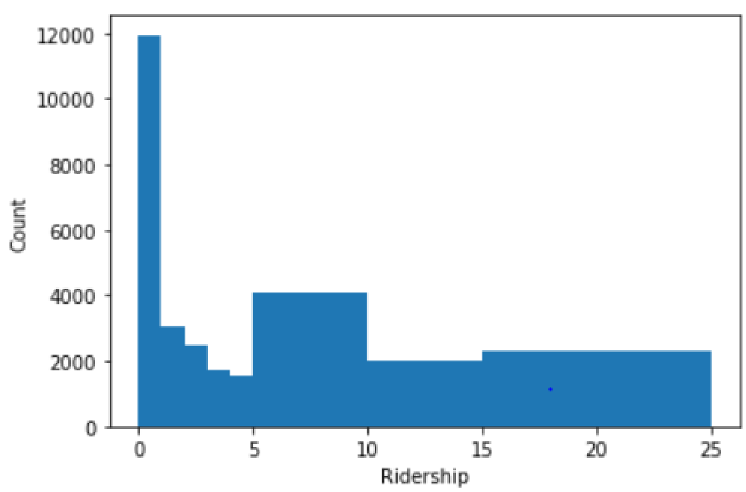
Data Processing

Before models can be formulated, transformations of the data are necessary such that each model will perform well. Stop names are also excluded from regression analysis, as the machine learning models will overfit if a stop is assigned to each observation. Weekends are removed because of the comparatively small sample, and all observations are aggregated by the hour.

Poisson Overview

Using the APC systems equipped on PT busses, all boardings and alightings are observed at the stop and route level. From this system the dependent variable, ridership (*rid*), is generated by summing both at each stop. However, since the ridership is measured as a count of the sum of boardings and alightings, it violates the ordinary least squares (OLS) assumption of normally distributed errors. Because it is impossible to observe a count less than zero, the variance will grow with the mean which presents heteroskedasticity. This can be remedied by taking a log transformation of ridership. However, since a majority of ridership observations include zeros, taking a log of ridership in this case is infeasible. This underscores the rationale for utilizing other models that can better handle a non-normally distributed dependent variable. Investigating the distribution of ridership in figure 5 below reveals that ridership is heavily skewed towards zero, more closely related to a Poisson distribution.

Figure 5: Ridership distribution



The Poisson model is better suited to nonnegative dependent variables, especially when the distribution is skewed towards zero. Many zeros are present because each observation is created when a bus passes by a stop regardless of whether it let passengers on or off. The Poisson regression is represented by the equation below:

$$E[rid_{it} | X_{it}] = e^{\beta x_{it} + \alpha_i + \mu_t + \epsilon_{it}}$$

Vectors x and β represent the explanatory variables and their coefficients, respectively. The *i* term signifies each stop, where *t* represents the time affecting all stops. The terms α_i represents the individual specific effects, μ_t is a linear time trend, and ϵ_{it} is the error.

Machine Learning Overview

Before developing the machine learning models, it’s necessary to understand the models used and their benefits. There are three major types of algorithms: Supervised Learning, Reinforcement Learning, and Unsupervised Learning. Supervised learning models are used where the variables are labeled and can be predicted in regression or classification problems given another set of variables. Unsupervised learning models are more useful where data is unlabeled, and the model can self-discover any naturally occurring patterns. Reinforcement learning methods assign positive values to the desired attributes to encourage the model, and negative values to undesired attributes (Ray 2017). Before choosing a model, one must consider the objective of the study, the nature of the data being used, and the desired accuracy of the model. The objective of this study is to predict ridership, so the appropriate analysis for this paper is a supervised regression algorithm because the target variable, ridership, is known.

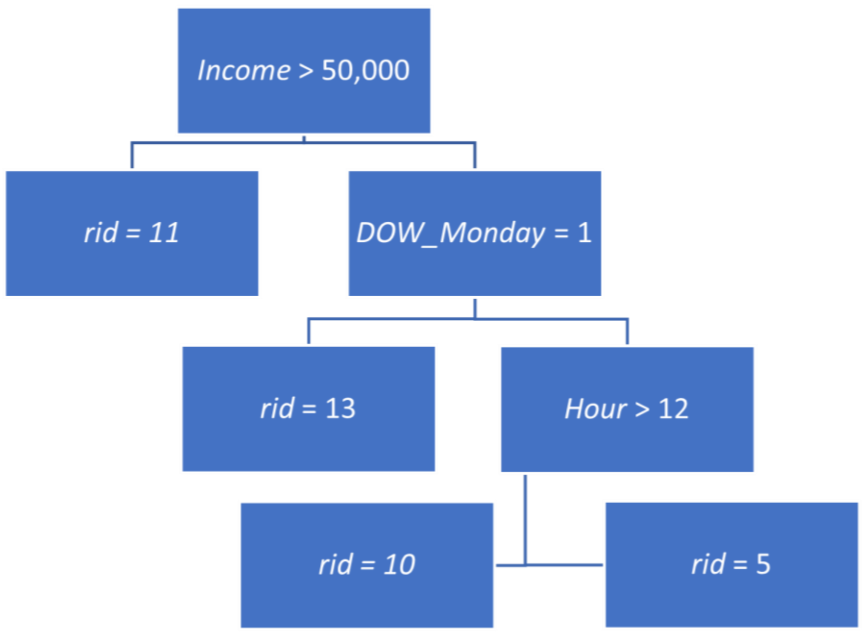
To test accuracy and avoid overfitting, data is split into testing and training sets. Finally, accuracy is important, but pursuing the highest in-sample accuracy score is not advised as this can lead to overfitting. This problem occurs where the model is trained heavily on the data of a training set, achieving excellent in-sample prediction while suffering in out of sample prediction. It is imperative that predictive models perform well both in- and out of sample.

To help improve prediction accuracy, many feature transformations of the census data are applied. These transformations result in a more complex and likely more predictive model, but they also increase the potential for overfitting. One method used to combat overfitting is called regularization. The regularization model used in this paper is called the "least absolute shrinkage and selection operator" or LASSO (Tibshirani 1996). LASSO, or L1 regularization, works by applying a penalty function to a model's loss term. This has the effect of reducing the coefficients of non-important variables to zero leading to a simpler model.

To tune hyperparameters bayes search with cross-validation is used. In cross validation, several splits or "folds" are made on the data, the model is run on each fold, and then an average of the folds are taken to obtain an overall error estimate. Briefly, bayes search finds the minimum to an objective function in large problem-space. In this case, the objective is to arrive at the best model output given the variables included, so it randomly tries different combinations and returns the combination with the greatest validation score. The validation score used is mean absolute error, which is obtained by comparing predicted and actual estimates within the training set. Grouping is used to prevent the same set of stops being used in each of the folds, which might bias the estimates towards a particular set of stops. These steps make up the foundation for machine learning models to be fitted to the data.

Decision-Tree (DT) and Random Forest (RF) algorithms are used in this analysis. Tree-based methods involve segmenting the predictor space into a number of simple regions (Venables 1999). The motivation for using regression trees is that they are easily interpretable while vastly improving prediction accuracy. DT's work by taking each observation and partitioning an explanatory variable into different subsets (figure 6). In the example below, the decision tree first splits on income and predicts ridership will equal 11 for observations where income is less than $50,000. Then, another split is made where the day of the week is Monday. The prediction here finds observations on Monday in block groups that have > $50,000 income, and generates a prediction given these two conditions. This process goes on until a stopping criteria like minimum number of observations per leaf or maximum depth of the tree are met.

Figure 6: Example Decision Tree



Finally, the DT will stop being grown when a leaf or branch node has less than a minimum number of observations, or maximum depth has been reached. Setting the minimum number of observations required at a leaf node or setting the maximum depth of the tree are necessary to avoid overfitting the model. This process is called hyperparameter tuning, and in this model, it is performed by defining a set of values for the bayes search algorithm to search over. After running hundreds of times, the model will have tried many different combinations for parameters like max depth and outputs the best parameters to use when testing for out of sample prediction. A final model with exact hyperparameters will be selected when training and testing accuracy are roughly equivalent.

The RF model is a bagging method that utilizes the aggregation of several decision trees to make a final prediction (Breiman 2001). Bagging is short for “bootstrap aggregation”, and it works by randomly sampling from the training data with replacement, which further prevents overfitting. RF is a meta-estimator, meaning it simply uses the process of creating DT’s but aggregates the predictions of each one. However, an additional feature of the RF model that makes it distinct is that it limits the number of features that can be split at each node to some percentage of the total. This hyperparameter ensures that no one feature is relied on too heavily.

CHAPTER 5: RESULTS AND DISCUSSION

This section explores the rationale for choosing different models and their results. Each of these models were run in the software program Python, using the sklearn, and statsmodels packages (Pedregosa 2011, Seabold 2010).

Predictive Model Performance

In order to assess each model, it's necessary to evaluate the in-sample and out of sample predictive accuracy. Accuracy is measured in two ways, pseudo-R^2 and root mean squared error (RMSE). A pseudo R-squared is only useful when compared to another pseudo R-squared predicting the same outcome with the same data. A higher value for pseudo R-squared indicates better prediction of ridership. RMSE is another useful tool for examining predictive power. It is defined as the square root of the squared difference between observed and predicted values.

Poisson

The aim of the Poisson model is to deal with the skewedness that count data brings. First, when training the Poisson model, the likelihood function does not automatically converge due to the large number of variables included. To develop a good model, several control variables were left out. These include the 1/8 and 1/2 mile radiuses counting points of interest near stops and all time and distance measurements for stops to locations of interest. This also ensures that coefficients are estimated more precisely, and statistical significance is not affected by collinearity. The model is then estimated with a robust covariance matrix to prevent potential overdispersion and relax the assumption that variance must be equal to the mean. In-sample pseudo R-squared for this model is .52, so the model does a reasonable job predicting in-sample ridership. However, the out of sample RMSE is 82609.27, meaning that where ridership is predicted the model is off by an average of 82609.27 *rid*. Table 6 in the appendix details the full output of the Poisson model.

Decision Tree and Random Forest

Table 3: DT and RF In sample performance

Accuracy	Decision Tree	Random Forest
RMSE	23.12	22.39
Pseudo R^2	.72	.73

Table 3.1: DT and RF Out of sample performance

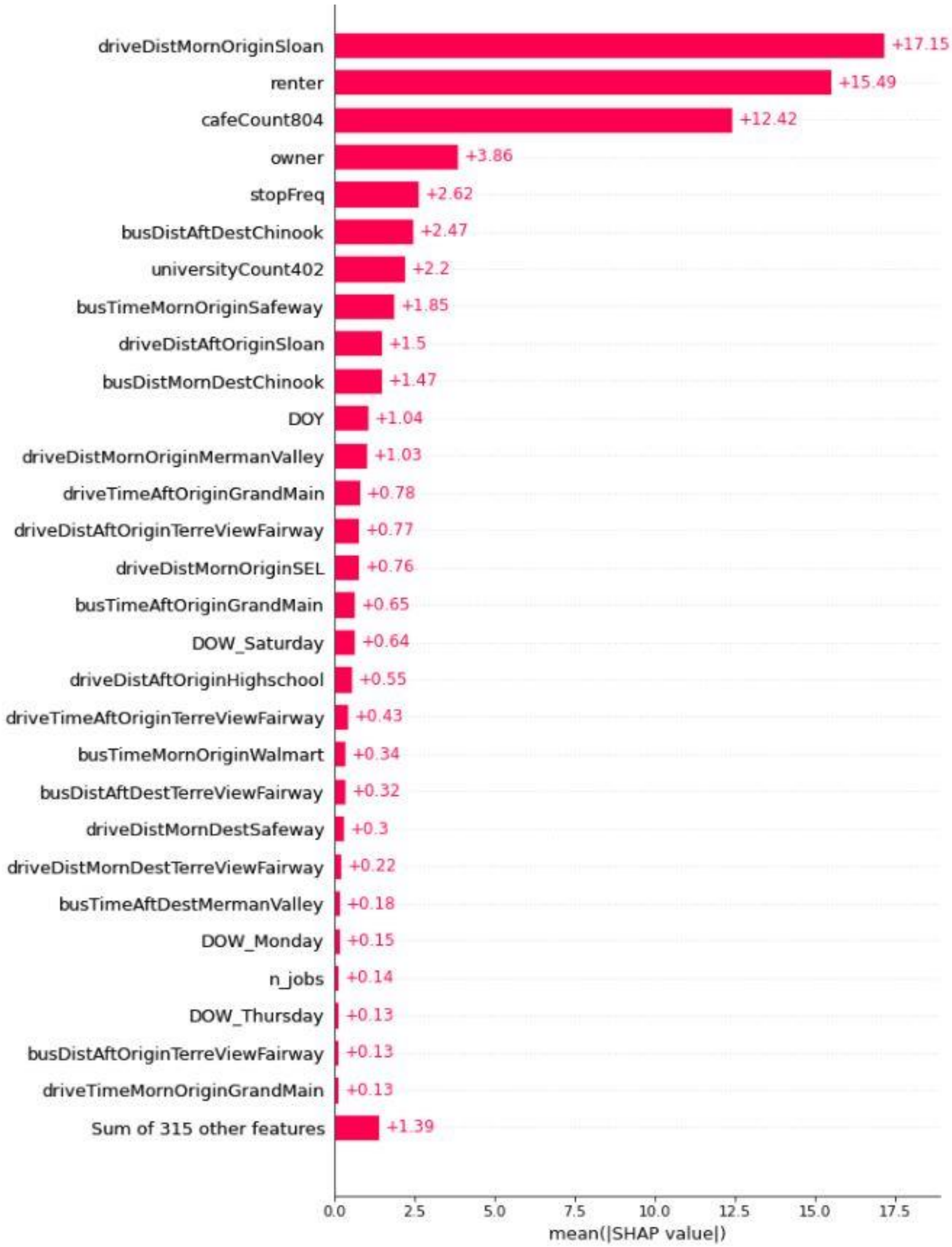
Accuracy	Decision Tree	Random Forest
RMSE	42.87	37.63
Pseudo R^2	.31	.47

After tuning the decision tree by allowing bayes search cross validation to search over hundreds of possible hyperparameters, its predictive performance is better than the Poisson model. However, to quote from Elements of Statistical Learning, "Trees have one aspect that prevents them from being the ideal tool for predictive learning, namely inaccuracy" (Hastie et al. 2009). In other words, the in-sample accuracy is much better, but the DT is shown to not be as flexible out of sample with a pseudo R-squared of .31, and RMSE of 42.87. This coefficient for RMSE means that where ridership is predicted, it is off on average by 42.87 *rid* at a stop.

Table 4: Decision Tree hyperparameters

LASSO α	Max Depth	Min Samples (Leaf)	Min Samples (Split)
1.33	304	24	103

Figure 7: DT SHAP values



SHAP values, an acronym from Shapley Additive Explanations, help break down a prediction to show the impact of each feature. For machine learning models like DT's and RF this is useful because the depth of a tree can make it hard to interpret which features are having the greatest impact on prediction. SHAP values interpret the impact of having a chosen value for a given feature in comparison to the prediction made if that feature was some baseline value. In figure 7 above, the features that influenced prediction the most were *DriveDistMornOriginSloan*, *renter*, *Cafecount804*, and *owner*. The first feature captures the distance to drive to Sloan Hall on the WSU campus from any of the bus stops. *Renter* captures the number of renters in a bus stops' block group. *Cafecount804* captures the number of cafes within half a mile, and *owner* is a count of the number of homeowners within a block group. Given the low value of alpha chosen by the bayes search algorithm, many features were not penalized and as a result were used in the final prediction.

Random Forest

Finally, the RF model ends up performing the best out of sample with a pseudo R-squared of .47 and a RMSE of 37.63. This captures how the combined estimations of many trees can enhance a model's out of sample prediction. Below in table 4.1 are the hyperparameters used for these estimates.

Table 4.1: Random Forest hyperparameters

LASSO α	**Max Depth**	**Min Samples (Leaf)**	**Min Samples (Split)**	**Num. Estimators**
1.46	102	20	113	200

Figure 8: RF SHAP values

From the RF SHAP values it is clear the RF and DT models emphasized different sets of variables for prediction. Given the modest increase in predictive performance, it is difficult to attribute which variables are truly the most useful with respect to prediction. The top features used for prediction in the RF model were *busTimeAftDestSafeway* and *renterNoVehicle*. The first variable captures the time for a bus to reach the local Safeway supermarket, while *renterNoVehicle* describes the number of renters in a block group that don't own a vehicle. The alpha value also indicates that most of the features were used to inform this prediction.

Limitations

The census data used in this analysis was enumerated during the 2010 census and as such does not reflect the exact makeup of Pullman in 2019. More timely data will be available in 2021 after the 2020 census is through processing. Additionally, block group level of measurement always contains at least 300 households. This can be problematic as some bus stops are in locations with low population density and cause the block group to capture a large area around a stop. With respect to both machine learning methods, the results indicate overfitting. There is a large discrepancy between in and out of sample predictions, and parity between the two is desired for the best performance. Further optimization of hyperparameters and pruning of the RF and DT could increase the out of sample predictive performance.

CHAPTER 6: CONCLUSION

Public transit plays a crucial role in reducing the externalities associated with automobile use like pollution, congestion, and traffic accidents. Encouraging bus ridership is especially important as other modes of public transport have large startup costs and require greater population density; both of which make small towns an infeasible location. In this thesis predictive models are developed to serve as the foundation of a decision support tool for local transit authorities to make better transit service decisions. From this analysis it is clear that machine learning is a viable approach for ridership prediction where complex datasets make estimation difficult. The random forest algorithm has been demonstrated to be most effective at out of sample predictive accuracy compared to the alternatives. However, these results do not support high enough predictive accuracy to be useful in the context of a decision support tool. Alternative levels of aggregation and other methods for prediction have proven to be more effective in terms of predictive accuracy (Li 2020; Dill et al 2013; Frei and Mahmassani 2013; Taylor et al 2009).

REFERENCES

Berrebi, Simon J, Sanskruti Joshi, and Kari E Watkins. "On Bus Ridership and Frequency." *Transportation Research. Part A, Policy and Practice* 148 (2021): 140–54. https://doi.org/10.1016/j.tra.2021.03.005.

Breiman, Leo. "Random Forests." *Machine Learning* 45, no. 1 (2001): 5–32. https://doi.org/10.1023/A:1010933404324.

Chakour, Vincent, and Naveen Eluru. "Examining the Influence of Stop Level Infrastructure and Built Environment on Bus Ridership in Montreal." *Journal of Transport Geography* 51 (2016): 205–17. https://doi.org/10.1016/j.jtrangeo.2016.01.007.

Chakrabarti, Sandip. "The Demand for Reliable Transit Service: New Evidence Using Stop Level Data from the Los Angeles Metro Bus System." *Journal of Transport Geography* 48 (2015): 154–64. https://doi.org/10.1016/j.jtrangeo.2015.09.006.

Desaulniers, Guy, and Mark D Hickman. "Chapter 2 Public Transit." *Handbooks in Operations Research and Management Science* 14 (2007): 69–127. https://doi.org/10.1016/S0927-0507(06)14002-5.

Dill, Jennifer, Marc Schlossberg, Liang Ma and Cody Meyer. "Predicting Transit Ridership at Stop Level: Role of Service and Urban Form." (2013).

Ding, Chuan, Donggen Wang, Xiaolei Ma, and Haiying Li. "Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees." *Sustainability (Basel, Switzerland)* 8, no. 11 (2016): 1100. https://doi.org/10.3390/su8111100.

Fabrikant, "Predicting bus delays with machine learning". Google AI Blog. (2019) https://ai.googleblog.com/2019/06/predicting-bus-delays-withmachine.htm

Fontes, Tânia, Ricardo Correia, Joel Ribeiro, and José Luís Borges. "A Deep Learning Approach for Predicting Bus Passenger Demand Based on Weather Conditions." *Transport and Telecommunication* 21, no. 4 (2020): 255–64. https://doi.org/10.2478/ttj-2020-0020.

Frei, Charlotte, and Hani S Mahmassani. "Riding More Frequently: Estimating Disaggregate Ridership Elasticity for a Large Urban Bus Transit Network." *Transportation Research Record* 2350, no. 2350 (2013): 65–71. https://doi.org/10.3141/2350-08.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning*. New York: Springer, 2009.

Kawatani, Takuya, Tsubasa Yamaguchi, Yuta Sato, Ryotaro Maita, and Tsunenori Mine. "Prediction of Bus Travel Time over Intervals Between Pairs of Adjacent Bus Stops Using City

Bus Probe Data." *International Journal of ITS Research* 19, no. 2 (2021): 456–67. https://doi.org/10.1007/s13177-021-00251-8.

National Academies of Sciences, Engineering, and Medicine. 2020. *Analysis of Recent Public Transit Ridership Trends*. Washington, DC: The National Academies Press.https://doi.org/10.17226/25

Pedregosa, Fabian, Gaeel Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (2011): 2825–30. https://doi.org/10.5555/1953048.2078195.

Pulugurtha, Srinivas S. and Mahesh Agurla, "Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods," Journal of Public Transportation, 2012, 15 (1), 33–52.

"Rural Transit Fact Book." *SURCOM - Rural Transit Fact Book*, North Dakota State University, 2021, https://www.ugpti.org/surcom/resources/transitfactbook/.

Seabold, Skipper, and Josef Perktold. "statsmodels: Econometric and statistical modeling with python." *Proceedings of the 9th Python in Science Conference*. 2010.

Stover, Victor W, and Edward D McCormack. "The Impact of Weather on Bus Ridership in Pierce County, Washington." *Journal of Public Transportation* 15, no. 1 (2012): 95–110. https://doi.org/10.5038/2375-0901.15.1.6.

Taylor, Brian D, Douglas Miller, Hiroyuki Iseki, and Camille Fink. "Nature And/or Nurture? Analyzing the Determinants of Transit Ridership Across US Urbanized Areas." *Transportation Research. Part A, Policy and Practice* 43, no. 1 (2009): 60–77. https://doi.org/10.1016/j.tra.2008.06.007

Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58, no. 1 (1996): 267–88. http://www.jstor.org/stable/2346178.

Venables, W. N., and Ripley, Brian D. *Modern Applied Statistics with S-PLUS*. 3rd ed. New York: Springer, 1999.

Wooldridge, Jeffrey M., 1960-. Introductory Econometrics: a Modern Approach. Mason, Ohio :South-Western Cengage Learning, 2012.

APPENDIX

Table 5: Summary Statistics

Variables	Mean	Std
doy	179.7	95.04
dow_num	2.36	1.66
month	6.41	3.12
week	26.31	13.58
dom	15.71	9.02
geoid	5.31E+11	1851.43
income	38634.31	23343.13
population	2359.34	911.48
incomepovertyratio	1982.96	1073.02
degreepopover25	641.17	437.38
enrolledover3	1376.31	874.8
ownernovehicle	14.8	16.92
renternovehicle	83.76	56.93
owner	259.69	271.22
renter	620.21	489.06
employed	1104.22	448.02
unemployed	112.17	113.98
med_age	27.32	7.25
med_house_val	286700.1	21351.43
walkscore	47.88	22.96
bikescore	38.53	14.11
stoproutevehiclefreq	47.08	36.67
stoproutefreq	46.77	37.11
stopfreq	51.47	62.88
shelter	0.24	0.4
simmeseat	0.1	0.28
bustimeaftoriginchinook	803.15	370.47
busdistaftoriginchinook	3311.65	2101.51
drivetimeaftoriginchinook	272.19	93.51
drivedistaftoriginchinook	2040.83	875.33
bustimeaftdestchinook	798.63	392.17
busdistaftdestchinook	3484.92	2664.81
drivetimeaftdestchinook	256.28	95.34
drivedistaftdestchinook	1963.73	911.37
bustimemornoriginchinook	771.53	389.42
busdistmornoriginchinook	3034.04	2024.91
drivetimemornoriginchinook	272.15	93.48
drivedistmornoriginchinook	2041.4	875.81
bustimemorndestchinook	799.49	374
busdistmorndestchinook	3554.62	2644.48
drivetimemorndestchinook	256.28	95.34

drivedistmorndestchinook	1963.73	911.37
bustimeaftorigincub	906.07	376.52
busdistaftorigincub	2819.71	1912.28
drivetimeaftorigincub	260.17	94.43
drivedistaftorigincub	2237.59	949.05
bustimeaftdestcub	990.75	488.3
busdistaftdestcub	3890	3000.67
drivetimeaftdestcub	286.76	98.98
drivedistaftdestcub	2333.26	975.96
bustimemornorigincub	869.29	338.33
busdistmornorigincub	2572.54	1720.02
drivetimemornorigincub	259.87	94.27
drivedistmornorigincub	2236.74	949.45
bustimemorndestcub	862.27	407.33
busdistmorndestcub	3708.27	2612.85
drivetimemorndestcub	286.76	98.98
drivedistmorndestcub	2333.26	975.96
bustimeaftoriginspark	815.46	337.75
busdistaftoriginspark	3006.65	1936.37
drivetimeaftoriginspark	330.16	101.25
drivedistaftoriginspark	2500.2	1004.71
bustimeaftdestspark	1069.11	497.7
busdistaftdestspark	4067.37	2978.26
drivetimeaftdestspark	280.74	94.63
drivedistaftdestspark	2325.4	985.56
bustimemornoriginspark	810.77	311.72
busdistmornoriginspark	2857.24	1828.78
drivetimemornoriginspark	329.72	101.08
drivedistmornoriginspark	2489.71	1001.36
bustimemorndestspark	981.88	408.28
busdistmorndestspark	3660.5	2509.55
drivetimemorndestspark	280.74	94.63
drivedistmorndestspark	2325.4	985.56
bustimeaftoriginsloan	782.18	348.43
busdistaftoriginsloan	3032.1	2047.46
drivetimeaftoriginsloan	227.64	92.67
drivedistaftoriginsloan	1869.22	877.41
bustimeaftdestsloan	906.93	395.21
busdistaftdestsloan	3458.94	2636.32
drivetimeaftdestsloan	251.44	100.27
drivedistaftdestsloan	2048.35	946.63
bustimemornoriginsloan	816.96	384.81
busdistmornoriginsloan	3020.22	2057.65

drivetimemornoriginsloan	227.64	92.67
drivedistmornoriginsloan	1869.22	877.41
bustimemorndestsloan	905.84	406.84
busdistmorndestsloan	3564.33	2639.37
drivetimemorndestsloan	251.58	100.51
drivedistmorndestsloan	2040.77	933.92
bustimeaftoriginbeasley	798.14	487.02
busdistaftoriginbeasley	3102.61	2059.76
drivetimeaftoriginbeasley	240.86	114.46
drivedistaftoriginbeasley	2225.45	1135.87
bustimeaftdestbeasley	748.5	491.24
busdistaftdestbeasley	3680.29	2623.7
drivetimeaftdestbeasley	243.5	115.64
drivedistaftdestbeasley	2234.95	1122.98
bustimemornoriginbeasley	760.48	463.58
busdistmornoriginbeasley	2939.21	1917.77
drivetimemornoriginbeasley	240.86	114.46
drivedistmornoriginbeasley	2225.45	1135.87
bustimemorndestbeasley	715.98	457.5
busdistmorndestbeasley	3822.5	2829.44
drivetimemorndestbeasley	243.45	115.59
drivedistmorndestbeasley	2234.28	1122.41
bustimeaftoriginvogel	695.43	345.7
busdistaftoriginvogel	2650.6	1657.84
drivetimeaftoriginvogel	212.36	89.78
drivedistaftoriginvogel	1878.28	931.25
bustimeaftdestvogel	777.7	333.24
busdistaftdestvogel	2964.83	1915.31
drivetimeaftdestvogel	202.32	90.09
drivedistaftdestvogel	1936.17	984.84
bustimemornoriginvogel	700.16	294.21
busdistmornoriginvogel	2475.95	1616.69
drivetimemornoriginvogel	212.86	91.04
drivedistmornoriginvogel	1883.39	943.37
bustimemorndestvogel	644.91	285.65
busdistmorndestvogel	2533.85	1944.69
drivetimemorndestvogel	202.32	90.09
drivedistmorndestvogel	1936.17	984.84
bustimeaftoriginwalmart	1007.37	490.93
busdistaftoriginwalmart	5019.97	2291.87
drivetimeaftoriginwalmart	427.61	150.93
drivedistaftoriginwalmart	3519.02	1524.74
bustimeaftdestwalmart	1317.62	590.74

busdistaftdestwalmart	5746.75	3150.22
drivetimeaftdestwalmart	456.33	148.38
drivedistaftdestwalmart	3639.94	1553.89
bustimemornoriginwalmart	1050.28	454.22
busdistmornoriginwalmart	5132	2258.85
drivetimemornoriginwalmart	427.61	150.93
drivedistmornoriginwalmart	3519.02	1524.74
bustimemorndestwalmart	1308.71	571.31
busdistmorndestwalmart	5876.29	3191.01
drivetimemorndestwalmart	456.38	148.31
drivedistmorndestwalmart	3636.52	1559.23
bustimeaftoriginsel	1380.89	561.88
busdistaftoriginsel	5997.7	3670.02
drivetimeaftoriginsel	320.35	136.46
drivedistaftoriginsel	3101.56	1397.07
bustimeaftdestsel	1358.43	427.79
busdistaftdestsel	4625.25	2913.27
drivetimeaftdestsel	314.51	131.65
drivedistaftdestsel	3057.63	1386.7
bustimemornoriginsel	1384.57	549.21
busdistmornoriginsel	5883.42	3430.3
drivetimemornoriginsel	320.35	136.46
drivedistmornoriginsel	3101.56	1397.07
bustimemorndestsel	1354.05	442.22
busdistmorndestsel	4648.86	2783.95
drivetimemorndestsel	314.51	131.65
drivedistmorndestsel	3057.63	1386.7
bustimeaftoriginhighschool	1619.56	474.16
busdistaftoriginhighschool	4754.88	2554.73
drivetimeaftoriginhighschool	357.81	111.54
drivedistaftoriginhighschool	3186.75	1104.5
bustimeaftdesthighschool	1623.77	491.01
busdistaftdesthighschool	5519.88	3366.56
drivetimeaftdesthighschool	376.31	104.74
drivedistaftdesthighschool	3150.61	1099.97
bustimemornoriginhighschool	1590.95	432.85
busdistmornoriginhighschool	4908.09	2570.47
drivetimemornoriginhighschool	357.81	111.54
drivedistmornoriginhighschool	3186.75	1104.5
bustimemorndesthighschool	1608.19	499.66
busdistmorndesthighschool	5813.3	3432.1
drivetimemorndesthighschool	376.31	104.74
drivedistmorndesthighschool	3150.64	1100.05

bustimeaftorigingrandmain	764.19	382.32
busdistaftorigingrandmain	3070.19	2149.84
drivetimeaftorigingrandmain	211.76	90.72
drivedistaftorigingrandmain	1876.72	935.34
bustimeaftdestgrandmain	844.23	373.31
busdistaftdestgrandmain	3050.6	2012.47
drivetimeaftdestgrandmain	202.31	90.66
drivedistaftdestgrandmain	1936.91	986.87
bustimemornorigingrandmain	761.49	350.85
busdistmornorigingrandmain	2788.96	1964.7
drivetimemornorigingrandmain	212.22	91.87
drivedistmornorigingrandmain	1881.59	946.84
bustimemorndestgrandmain	740.71	305.57
busdistmorndestgrandmain	2983.24	2091.89
drivetimemorndestgrandmain	202.31	90.66
drivedistmorndestgrandmain	1936.91	986.87
bustimeaftorigindissmores	1004.68	330.46
busdistaftorigindissmores	2865.6	1961.34
drivetimeaftorigindissmores	256.47	86.8
drivedistaftorigindissmores	2028.9	886.97
bustimeaftdestdissmores	997.23	351.2
busdistaftdestdissmores	3931.59	2399.03
drivetimeaftdestdissmores	250.3	77.85
drivedistaftdestdissmores	2048.17	904.75
bustimemornorigindissmores	999.7	353.48
busdistmornorigindissmores	2651.2	1810.75
drivetimemornorigindissmores	256.47	86.8
drivedistmornorigindissmores	2011.97	882.87
bustimemorndestdissmores	1011.63	324.41
busdistmorndestdissmores	3551.64	2723.91
drivetimemorndestdissmores	250.3	77.85
drivedistmorndestdissmores	2048.17	904.75
bustimeaftoriginterreviewmerman	1265.15	667.05
busdistaftoriginterreviewmerman	4986.37	2970.75
drivetimeaftoriginterreviewmerma	275.9	143.07
drivedistaftoriginterreviewmerma	2630.66	1388.44
bustimeaftdestterreviewmerman	953.11	514.78
busdistaftdestterreviewmerman	3910.89	2433.7
drivetimeaftdestterreviewmerman	269.62	140.68
drivedistaftdestterreviewmerman	2614.5	1401.89
bustimemornoriginterreviewmerman	1242.62	642.9
busdistmornoriginterreviewmerman	4927.34	2762.02
drivetimemornoriginterreviewmerm	275.99	143.19

drivedistmornoriginterreviewmerm	2629.07	1386.53
bustimemorndestterreviewmerman	889.71	430.39
busdistmorndestterreviewmerman	3779.38	2301.84
drivetimemorndestterreviewmerman	269.62	140.68
drivedistmorndestterreviewmerman	2614.5	1401.89
bustimeaftoriginvalleystadium	962.93	479.09
busdistaftoriginvalleystadium	3004.3	2225.59
drivetimeaftoriginvalleystadium	215.06	101.37
drivedistaftoriginvalleystadium	1927.51	1004.53
bustimeaftdestvalleystadium	953.98	428.18
busdistaftdestvalleystadium	3704.57	2494.34
drivetimeaftdestvalleystadium	202.97	101.22
drivedistaftdestvalleystadium	1917.72	1031.78
bustimemornoriginvalleystadium	951.07	475.91
busdistmornoriginvalleystadium	2824.78	2039.64
drivetimemornoriginvalleystadium	215.06	101.37
drivedistmornoriginvalleystadium	1927.51	1004.53
bustimemorndestvalleystadium	855.59	355.78
busdistmorndestvalleystadium	3100.1	2459.09
drivetimemorndestvalleystadium	202.97	101.22
drivedistmorndestvalleystadium	1917.72	1031.78
bustimeaftoriginsafeway	1009.03	483.02
busdistaftoriginsafeway	5162.76	2427.82
drivetimeaftoriginsafeway	325.54	147.59
drivedistaftoriginsafeway	3194.22	1478.23
bustimeaftdestsafeway	1269.76	699.41
busdistaftdestsafeway	5482.34	2948.54
drivetimeaftdestsafeway	328.72	148.01
drivedistaftdestsafeway	3238.94	1518.33
bustimemornoriginsafeway	1048.36	466.21
busdistmornoriginsafeway	5298.83	2478.76
drivetimemornoriginsafeway	325.54	147.59
drivedistmornoriginsafeway	3194.22	1478.23
bustimemorndestsafeway	1287.6	581.47
busdistmorndestsafeway	5404.32	3127.1
drivetimemorndestsafeway	328.72	148.01
drivedistmorndestsafeway	3238.94	1518.33
bustimeaftoriginmermanvalley	1187.36	597.81
busdistaftoriginmermanvalley	4172.42	2638.22
drivetimeaftoriginmermanvalley	251.25	124.61
drivedistaftoriginmermanvalley	2275.31	1206.27
bustimeaftdestmermanvalley	994.7	466.45
busdistaftdestmermanvalley	3356.36	2275.17

drivetimeaftdestmermanvalley	243.09	122.47
drivedistaftdestmermanvalley	2260.48	1218.47
bustimemornoriginmermanvalley	1203.07	611.47
busdistmornoriginmermanvalley	4120.48	2488.15
drivetimemornoriginmermanvalley	251.25	124.61
drivedistmornoriginmermanvalley	2275.31	1206.27
bustimemorndestmermanvalley	929.96	405.44
busdistmorndestmermanvalley	3549.6	2106.58
drivetimemorndestmermanvalley	243.14	122.54
drivedistmorndestmermanvalley	2261.36	1219.51
bustimeaftoriginterreviewfairway	1089.79	579.88
busdistaftoriginterreviewfairway	4079.18	2187.05
drivetimeaftoriginterreviewfairw	284.55	141.74
drivedistaftoriginterreviewfairw	2794.99	1379.8
bustimeaftdestterreviewfairway	931.83	502.08
busdistaftdestterreviewfairway	4455.1	2583.06
drivetimeaftdestterreviewfairway	288.05	142.19
drivedistaftdestterreviewfairway	2808.97	1401.94
bustimemornoriginterreviewfairwa	1080.65	571.99
busdistmornoriginterreviewfairwa	3894.71	1952.28
drivetimemornoriginterreviewfair	284.55	141.74
drivedistmornoriginterreviewfair	2794.99	1379.8
bustimemorndestterreviewfairway	922.25	468.01
busdistmorndestterreviewfairway	4483.44	2666.91
drivetimemorndestterreviewfairwa	288.4	142.54
drivedistmorndestterreviewfairwa	2812.39	1404.67
bustimeaftoriginsrc	1105.42	532.01
busdistaftoriginsrc	3489.38	1956.56
drivetimeaftoriginsrc	327.54	127.36
drivedistaftoriginsrc	2870.55	1233.42
bustimeaftdestsrc	975.26	514.46
busdistaftdestsrc	3966.29	2732.97
drivetimeaftdestsrc	321.32	126.65
drivedistaftdestsrc	2838.66	1215.56
bustimemornoriginsrc	1083.33	510.74
busdistmornoriginsrc	3404.14	1811.5
drivetimemornoriginsrc	327.54	127.36
drivedistmornoriginsrc	2870.55	1233.42
bustimemorndestsrc	902.1	386.13
busdistmorndestsrc	3725.8	2574.47
drivetimemorndestsrc	321.28	126.6
drivedistmorndestsrc	2837.9	1214.84
restaurantcount804	13.6	10.41

restaurantcount402	4.33	5.4
restaurantcount201	1.19	2.46
restaurantcount1005	0.29	0.84
transit_stationcount804	27.12	8.2
transit_stationcount402	8.31	3.35
transit_stationcount201	2.89	1.35
transit_stationcount1005	1.66	0.91
cafecount804	3.8	3.22
cafecount402	1.28	1.86
cafecount201	0.42	1.18
cafecount1005	0.11	0.41
churchcount804	4.98	3.94
churchcount402	1.53	1.84
churchcount201	0.38	0.66
churchcount1005	0.14	0.41
universitycount804	14.47	21.88
universitycount402	4.59	9.45
universitycount201	1.08	2.91
universitycount1005	0.2	0.57
supermarketcount804	0.13	0.33
supermarketcount402	0.03	0.16
supermarketcount201	0	0.03
supermarketcount1005	0	0
department_storecount804	0.08	0.27
department_storecount402	0.04	0.19
department_storecount201	0	0
department_storecount1005	0	0
gas_cpi	232.82	16.99
gas_pct_diff	0.89	5.96
daily_snowfall	0.11	0.54
daily_snowdepth	0.79	2.89
dailydrybulbtemperature	55.09	17.67
dailyprecipitation	0	0.01
dailywindspeed	8.41	5.44
dow_sunday	0.04	0.2
dow_monday	0.16	0.37
dow_tuesday	0.19	0.39
dow_wednesday	0.19	0.39
dow_thursday	0.18	0.38
dow_friday	0.19	0.39
dow_saturday	0.05	0.22
n_jobs	782.11	955.37
n_jobs29	287.71	262.52

n_jobs30_54	362.52	519.69
n_jobs55	131.88	218.95

Table 6: Poisson model results

Variables	**Ridership (*rid*)**
doy	0.0253
	(0.0232)
dow_num	-0.00996
	(0.00866)
month	-0.909
	(0.705)
week	0.0266***
	(0.00504)
dom	-0.0289
	(0.0231)
income	-2.43e-05***
	(2.50e-06)
population	0.00222***
	(0.000248)
degreepopover25	-0.00423***
	(0.000315)
enrolledover3	-0.000513***
	(0.000192)
ownernovehicle	-0.0197***
	(0.00275)
renternovehicle	-0.00846***
	(0.00129)
owner	0.00950***
	(0.00111)
renter	0.00425***
	(0.000570)
employed	-0.00379***
	(0.000332)
unemployed	-0.00824***
	(0.000727)
o.labor_frc	-
med_age	-0.144***
	(0.0151)
med_house_val	-4.93e-06***
	(6.63e-07)
walkscore	-0.0108***

	(0.00123)
bikescore	-0.00736***
	(0.00121)
stoproutevehiclefreq	0.0234***
	(0.00238)
stoproutefreq	-0.0314***
	(0.00239)
stopfreq	-0.00504***
	(0.000134)
shelter	0.661***
	(0.0356)
simmeseat	0.578***
	(0.0561)
restaurantcount402	0.0671***
	(0.00666)
transit_stationcount402	-0.0405***
	(0.00466)
cafecount402	0.212***
	(0.0195)
churchcount402	-0.338***
	(0.0118)
universitycount402	0.0421***
	(0.00308)
supermarketcount402	-1.418***
	(0.0716)
department_storecount402	-2.130***
	(0.0929)
gas_cpi	-0.00226*
	(0.00117)
gas_pct_diff	-0.0103***
	(0.00239)
daily_snowfall	0.0234
	(0.0200)
daily_snowdepth	-0.0104**
	(0.00437)
dailydrybulbtemperature	0.00422***
	(0.00101)
dailyprecipitation	0.571
	(1.588)
dailyavgwindspeed	-0.00188
	(0.00260)
n_jobs	0.00140
	(0.00140)
n_jobs29	0.000981

	(0.00118)
n_jobs30_54	-0.00218
	(0.00214)
o.n_jobs55	-
Constant	10.09***
	(0.811)
Observations	26,827
Robust standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	