When Will the Bus Arrive?

An Interactive Qualify Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by:

Matthew J. Iandoli

Scott B. Gubrud

Date: 6 March 2020

Professor Jacob Whitehill, Advisor

Abstract

Public transportation decreases the amount of cars on the road and potentially provides a cleaner and more economic way of travel. With the many benefits of public transportation, some problems arise, mainly the predictability of bus arrival times. We collected bus arrival times through a real-time bus-tracker and weather predictions. Our main goal was to determine if we could predict when buses would arrive based on the weather. We then used the data to train a ridge regression model to predict lateness. Our models had some predictive power, but weather was not as useful a predictor as we initially thought it would be.

Contents

1 Introduction

As population and cities grow, the need for public transportation becomes more and more prevalent.

1.1 Benefits of public transportation

Publicly available transportation provides a mean of travel for those who do not have access to a car to get to work, get groceries and other community resources. Bus travel plays an important part in economic growth, better buses allow workers to take better jobs that they would have not been able to commute to, "Almost half the sample, and more than half of those who normally use the bus to commute to work, felt that a better bus service would give them access to a better job." [1] Public transport also allows those who do have a car an alternate way to travel, reducing traffic for other drivers. Benefiting by not having to drive, one can work, catch up on their latest TV show or just relax because they do not need to deal with the stresses of bad drivers and traffic. While it has benefits to the individual taking public transport also benefits the greater good, taking their carbon producing vehicle off the road to reduce their emissions on the planet. To give perspective, "If one in 10 Americans used public transportation regularly, the U.S. reliance on foreign oil could be cut by more than 40%, which translates to the amount of oil we import from Saudi Arabia each year" and "Communities that invest in public transit reduce the nation's carbon emissions by 37 million metric tons annually" [2][3]. This combined with the growing public concern about climate change makes public transit an attractive option for those who wish to reduce their carbon footprint while still getting around.

1.2 Flaws with public transportation

Public transportation has a lot of benefits such as people who can not afford or do not have a car, people who do not want to drive during their commute, and less car congestion. However, it does have its flaws and shortcomings. The most notable is its inconvenience, to take public transportation, one has to build their schedule around the schedule of the bus, train, or subway. Public transportation also can be unreliable which can put even more stress on people's schedule. The possibility of a bus being late may force a rider to take an earlier bus if they want ensure that they reach their designation on time. This problem is even greater for riders who need to ride on multiple buses to get to their destination. The routes in place by the transit authority may not be convenient to where a person lives or is trying to go which can make the commute longer due to changing routes or an increase amount of foot travel.

1.3 WRTA

The Worcester Regional Transit Authority (WRTA) supports 36 communities around central Massachusetts including the city of Worcester. Many WPI commuter students use the WRTA to get to and from campus everyday. The WRTA has a fleet of 52 buses (as of 2018) with 6 being electric buses, 17 being diesel-hybrid hybrid buses and the rest being clean diesel buses [4]. From a 2017 survey conducted by the WRTA to perform customer satisfaction, 42.7% of riders use the bus for transportation to work with 57.3% of riders using the bus 2 or more

times a day. Most importantly, 86.6% of riders come from a household with an annual income below $40,000 which exemplifies the benefit of public transportation to the community [4]. 86.6% can be compared to the median household income of $46,105 in Worcester, MA showing how important the WRTA is for providing inexpensive transportation for the less fortunate [5]. Serving over 3 million annual passenger trips and traveling over 2 million miles per year, the WRTA is integral to Worcester's economical infrastructure and growth.

1.4 Existing Technologies

1.4.1 Online Schedules

The WRTA was a website with general information, news and updates, maps and schedules. Users can visit the website and determine roughly when a bus will arrive at their desired stop. While the website publishes live updates, people can follow the WRTA's Twitter account or subscribe to their SMS text message service to receive real-time notifications about schedule changes. Unfortunately, the online schedules do not have all the stops' run-times available. The WRTA only publishes the start/end stops along with the larger intermediary stops forcing the user to interpolate the arrival time of their bus. An example of the online schedule[1] can be seen from Figure 1.

Route 1 - Inbound

Monday - Friday

Stop	Run #1	Run #2	Run #3	Run #4	Run #5	Run #6	Run #7
ROUTE 146 + WALMART 1145	6:15am	7:15am	8:15am	9:15am	10:15am	11:15am	**12:15pm**
Worcester Senior Ctr - Driveway 3631	6:20am	7:20am	8:20am	9:20am	10:20am	11:20am	**12:20pm**
GRANITE ST + ST ANNES 329	6:24am	7:24am	8:24am	9:24am	10:24am	11:24am	**12:24pm**
GRAFTON ST + RICE SQUARE 294	6:35am	7:35am	8:35am	9:35am	10:35am	11:35am	**12:35pm**
CENTRAL HUB AT UNION STATION 3130	6:45am	7:45am	8:45am	9:45am	10:45am	11:45am	**12:45pm**

Figure 1: An example of one of schedules for bus arrival times showing the schedule times for the inbound run of Route 1 on the weekdays. The schedule displays the stop's long name (usually the intersection of two roads or a landmark) with the stop-code underneath it. Each stop will have multiple runs throughout the day with some stops being skipped on certain runs depending on the route or time of day. The online schedule does not show all the stops and their times but just the"key" stops (the ones that are more popular). There are a handful of smaller stops in between each stop shown in the schedule.[1]

[1]https://www.therta.com/routes/1/

1.4.2 Bus Trackers

The WRTA provides an online service that allows people to track buses along their routes to within an hour of their arrival[2]. People can either track their bus from the website on their phone or activate text instructions for the buses next arrival.

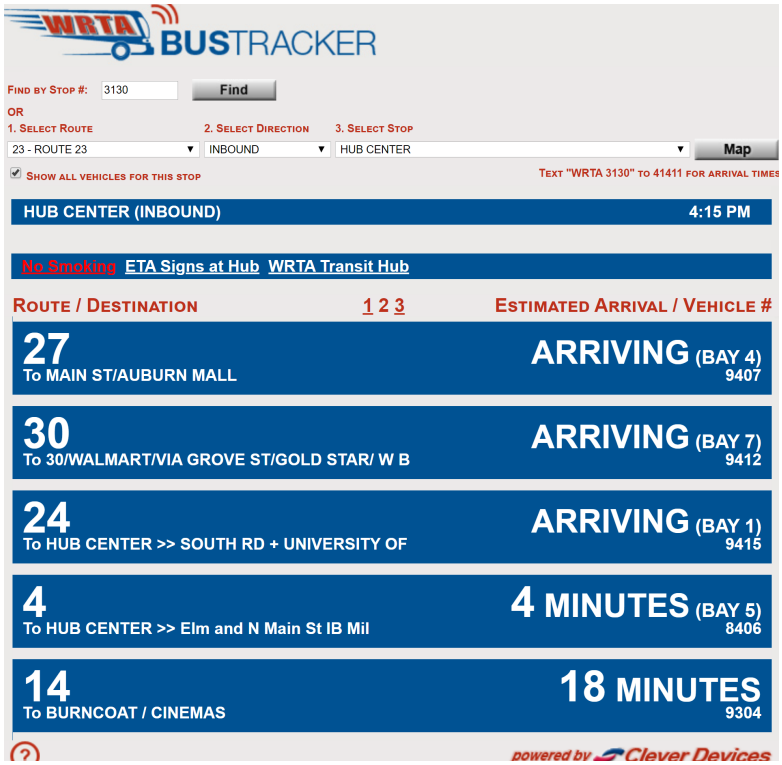

Figure 2: Bus tracker site showing the upcoming buses to a particular stop. The WRTA's real-time bus tracker that shows the upcoming buses to a particular stop. The site can look up stops by their stop-code or by searching with their route, direction, and name. The estimated arrival time is calculated by taking GPS location into account and their scheduled arrival time. When the bus has an estimated arrival time under 2 minutes, the site will display the bus as "arriving" and when the bus leaves the list, the bus has made its stop at the pick-up location.[2]

The users can either manually enter the stop code from the bus stop sign or scan the QR code directly from the sign. However, some bus stop signs are outdated and do not always have the stop code posted on the sign. This requires the users to manually search through the drop-down list on the website. Alternatively, they can find the stop code on Google Maps.

[2]http://bustracker.therta.com/bustime/eta/eta.jsp

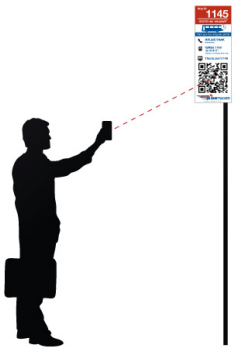

Figure 3: WRTA street sign displaying: stop code, route number, QR code, etc. Each stop is recognized by the bus sign attached to a post. Each sign will display what route(s) its on and some signs will display more detailed information like stop-code and QR-code to a link to the real-time bus arrival site for that particular stop.

1.5 Early Predictors

There are many different predictors that can be used to "guess" roughly when a bus will arrive. The most basic are when is the bus scheduled to come and what time it usually comes. For the most part, buses have a trend whether it be one specific bus is always early, always late, or etc. Another simple predictor is the day of the week it is. This is useful because per say Fridays could have more traffic in a particular location so the buses will tend to run late on Fridays. A more subtle predictor could be the weather, if its raining, snowing, the visibility. Poor weather usually (hopefully) causes drivers to use more caution while driving. Cautious driving can cause bus drivers to be a little late to make sure there passengers are safe. Another guess about weather is depending on how well the weather is that day, people are more or less likely to use public transportation that day. If its cold or rainy, people are less likely to walk or bike to where they need to go and will be more likely to take a bus which has temperature control and a roof. Depending on the bus routes, planned construction can have a big factor in whether or not a bus shows up on time either due to the traffic caused by it or any detours the bus has to make.

1.6 Causes of Lateness and Unpredictability

There are many causes of lateness that cannot be predicted, these include: car accidents, bus breakdowns, disorderly passengers and many others. Car accidents cannot be predicted and lead to an increase amount of car congestion and traffic. To an extreme, car accidents could make buses re-route and increase their lateness or even miss a stop. Another issue is although well designed, buses are not perfect and can malfunction or breakdown. A lot of malfunctions will not cause the bus to necessary be late, for example if the air conditioning on the bus breaks. However, engine problems and others issues can cause the bus to stop or even go out of commission. Malfunctions with the payment system can also cause the bus to stay at stops for prolonged periods of time. Another cause of concern can be disorderly passengers, whether it be break outs on the bus or at the stops themselves, while uncommon can cause the bus to run late.

1.7 Bus Stop Locations

Bus stop locations are attempted to be evenly distributed so that people have easy access to bus stops from their homes, places of work, etc. Transit authorities have this task and while evenly distributing these stops around their city, they also have to take into account exactly where on the street these stops are. It is so important because these locations can have a direct impact of the efficiency of the bus. An article written by Pace: Transit Support Guidelines discusses how its best for bus stops near intersections, are placed after the traffic light, also known as far-side bus stops, to maximize buses efficiency. The benefits of this come from the bus after stopping to integrate back into traffic easier through the gaps, avoiding blocking right hand turns, and the need to decelerate less [6].

1.8 Field Trip: A Trip to Worcester State with WRTA

To gain first hand experience in riding on one of the buses for the WRTA, our team scheduled a trip along a route that rides along the outside of the Worcester Polytechnic Institute (WPI)campus. The route including getting on a stop nearby WPI, continuing outbound to the Worcester State campus, stopping for a few minutes, and then travelling back inbound from Worcester State to WPI then back along its loop towards Union Station (see Figure 4).

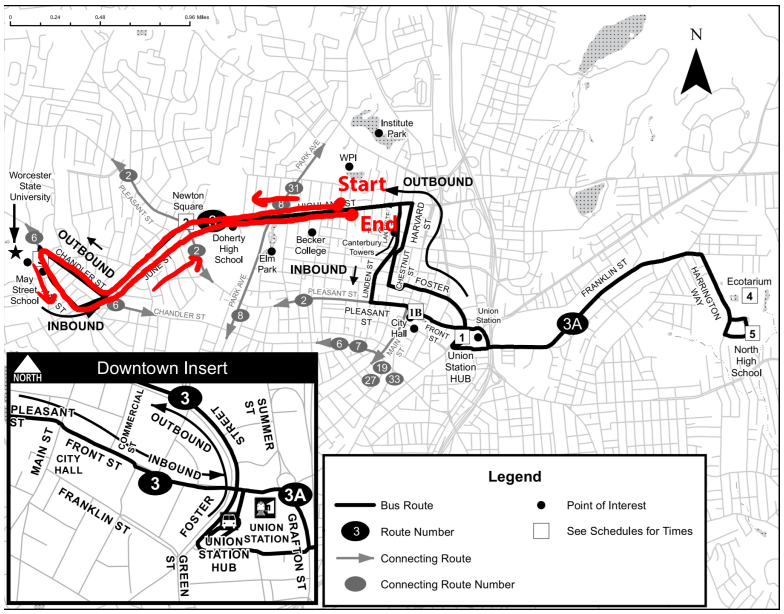

Figure 4: WRTA's map for route 3/3A with our planned route from WPI to Worcester State and back to WPI.

1.8.1 Attempt 1

To prepare for our trip, we wanted to simulate the experience of an average commuter of the WRTA as best as possible. So we started off by selecting a pick-up and drop-off spot and to find a time which would fit our team's schedule. For our pick-up location, we chose the Highland St. + Schussler Rd., a stop outbound on Route 3/3A. In conjunction, we chose a drop-off location that was across the street from our pick-up location for our general ease

which ended up being Highland St. + West. St., a stop inbound on Route 3/3A. We then chose a time most convenient for our team which ended up being around 12:15pm on October 31st.

Following WRTA's notice, "It is best to arrive at your bus stop at least five minutes early.", we started walking to our pick-up location shortly after noon from WPI's campus to arrive at least 5 minutes earlier than the scheduled time. When we arrived to our stop, we excitedly waited for our bus to arrive... but to our disappoint, no bus showed up. We tried to give the bus the benefit of the doubt but after 45 minutes or so of waiting we had to head back for other commitments. By the time we started to head back, we were able to see the next scheduled bus along Highland St. Hoping this was just a fluke, we decided to plan another trip.

1.8.2 Attempt 2

Two weeks later, on November 14th, 2019, we went for a second attempt to ride the bus. We took the exact same stops and schedule time as our first. To our prevail, the bus tracker was working this time and we were able to catch a bus that fit into our meeting time. As we got into the bus, it was fairly empty with only 2-4 passengers riding along with us at a time. This was understandable because it was 12:30pm on a Wednesday and not a busier time like 5 or 6pm. We rode the bus all the way to Worcester State University where the bus stopped for around 7 minutes before it started its route back up and brought us back to WPI's campus.

Figure 5: A picture of a the WRTA bus we rode on our second attempt.

1.9 Machine Learning

We both have some background knowledge on machine learning, having both taken Introduction into Artificial Intelligence (AI) which goes over a handful of AI topics including machine

learning. One of us has also taken the Machine Learning course which dives deeper into the math behind models and evaluation of model's accuracy. Some of the models including linear regression, naive Bayes classifier, quadratic discriminant analysis, random forest, and support vector machines (SVM's) as well as a brief overview of artificial and convolutional neural networks.

Given that bus schedules are know not to be 100% reliable and that existing predictive technologies clearly have room for improvement, we sought to generate our own predictions. Machine learning has been known to pick up small subtleties within data sets that normal statistic evaluation can miss. With this knowledge, we will try to apply machine learning to predict with greater accuracy when a given bus will arrive based on a set of predictors (weather conditions, day of week, etc.).

1.10 Related Work

Previous work has already been done on using machine learning and other modeling techniques to predict timings and delays for various parts of public transit networks [7].

One study used multivariate linear regression and random forests to predict bus travel times based on traffic, weather, and scheduled travel times. They worked with the Nashville Metropolitan Transit Authority which allowed them to collect real-time bus position data alongside weather and real-time traffic data which came from available APIs. The resulting predictions had an RMSE of 4 minutes and 50 seconds [8].

Another study tested a hybrid method that combines random forest with k-nearest neighbors against linear regression, k-nearest neighbors, SVM, and standard random forest. They used average bus dwell time, current bus speed, and the speed of buses ahead on the route to predict bus travel times for 2 routes in Shenyang, China. They obtained data by tracking the automotive vehicle location system for these buses. Their various models had RMSEs ranging from 26.37 seconds to 48.47 seconds with the random forest based on nearest neighbor having the lowest at 26.37 seconds [9].

A study in 2016 compared artificial neural networks (ANN's), SVM's and Bayes networks for predicting travel times. They used historical bus data and real-time bus speed and traffic data to predict bus travel times for the public transit system in Santiago, Chile. Their ANN was the most successful offering a 23% improvement over the existing system [10].

A study in 2002 used ANN's integrated with an algorithm to update prediction error in real-time in order to predict travel time. Using historical and real-time traffic and bus data, they built two models, one which predicted travel time for the entire run from one stop to another and a second model which made several predictions for the travel time to each intersection in between the two stops. The resulting models made predictions with root mean squared errors that ranged from around 60 seconds to around 390 seconds [11].

A study in 2007 used SVM's to predict bus travel time based on historical travel times and the travel time to get to the previous stop. They evaluated their model on transit route number 4 in Dalian economic and technological development zone in China and got root mean squared errors between 60 seconds and 130 seconds [12].

Another study from 2011 found SVM's to predict bus travel time was the more accurate than other models such as: ANN, k-nearest neighbor, and linear regression. This study took place in Hong Kong and was able to utilize new and historical arrival times to train their models. They also used surveys of passengers to test their trained models [13].

A different study that also used SVM's and ANN's combined them with Kalman-filtering and showed that the SVM-Kalman model was the most accurate. This study was based out of China and used bus arrivals from Shenzhen of two full weeks to evalute their proposed model [14].

2 Methodology

With the many benefits and flaws of public transportation in mind, we aim to provide better predictions of bus arrivals. In order to do this we needed to obtain data, define what it means to predict bus arrivals, build models to use the data to make predictions, and test how accurate those models are.

2.1 Pilot Data Investigations

The city of Worcester has the published data on the infrastructure of the public bus transportation, referred as General Transit Feed Specification (GTFS). This includes general information about the stops and routes such as geographical location, short/long names, scheduling, etc. However, this data does not include ridership or bus arrival time data. With the GTFS data-set, we were able to use this data to build up a database of relations between stops and routes. The GTFS does contain schedules for each route, however they are old and not up to date. We were able to write a script that pulls all the up-to-date schedules from the WRTA's website.

2.2 Collecting Data

We were able to obtain some PDFs of ridership data by request from the WRTA, however compression artifacts made any sort of automatic processing of this data prohibitively difficult. This meant our only option for data was to collect it ourselves. We were able to substitute bus arrival times with the WRTA's live bus tracker. We hypothesized we could continuously check the site and then record when each bus arrived. To confirm our hypothesis, we went around to a handful of stops nearby WPI's campus and checked to see if the online bus tracker was consistent with the actual bus. We were able to view that when the bus was removed from the live bus tracker, that was when it arrived at the specific stop. With this we would now be able to start collecting data from this point onward.

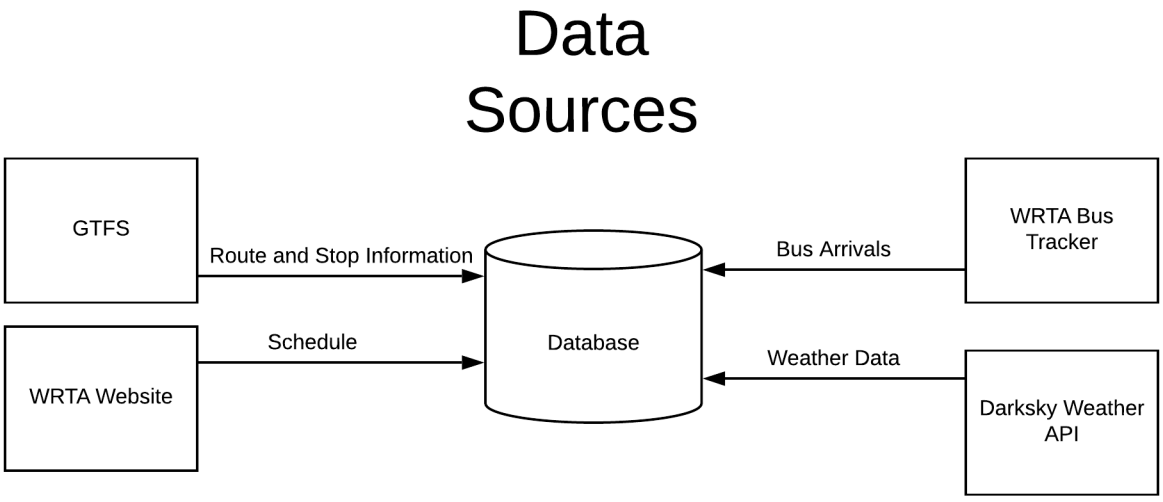

Figure 6: A visualization of the data sources that we use to populate our database. The GTFS data is used for relationships between stop-codes, long/short names, timetables, etc. while the other three data sources are used in our models.

Our intended analysis would require weather data to go along with our arrival time data. While historical weather data is available from several sources, we wanted our model to be based on weather predictions rather than actual weather data so that it could function better as a predictor. We could not find any historical source of weather predictions so we decided to collect them alongside our bus arrival data. We set up a system to collect weather predictions from the DarkSky weather API[3] for the following day each night at 9pm local time.

2.3 Data Alignment

Once we had data about when a bus arrived at a stop, we needed to figure out how to calculate how late (or early) a given bus arrival is. We considered several different methods for matching actual arrival times to scheduled arrival times.

2.3.1 Arrival Order

One of the simplest matching is for a given stop on a given day match the i^{th} actual bus arrival to the i^{th} scheduled bus arrival. There are a few potential problems with this method. Firstly this method assumes that there will be the same number of actual bus arrivals and scheduled bus arrivals which is not always the case for our data. Secondly this method does not produce lateness values that match the user experience. For example consider a bus stop with scheduled bus arrivals at 1:00, 2:00, and 3:00 for which buses actually arrive at 2:00, 3:00, and 4:00. This method will report that each bus arrived 1 hour late, however for a user that arrives at the bus stop at 2:00 would find a bus arrived exactly when it was scheduled to.

[3]https://darksky.net/

2.3.2 Next Bus

Another method we considered was to match each scheduled arrival time to the first bus that arrives after it. This method was intended to more faithful to the experience of a user by having lateness be how long a user who arrives at the scheduled time would have to wait for a bus to arrive. This method while probably closer to the actual user experience of lateness still fails to match some parts of the user experience especially with early buses. To illustrate the problem consider our simple stop with arrivals scheduled at 1:00, 2:00, and 3:00 for which buses actually arrive at 1:00, 1:59, and 3:00. In this situation the bus that was scheduled to arrive at 2:00 will be marked as being 1 hour late but in actuality most users trying to catch the 2:00 bus would arrive in time to catch the bus that arrived at 1:59.

2.3.3 Closest Scheduled

The matching method the we ended up using for matching was to match every actual bus arrival to the closest scheduled bus arrival (allowing multiple arrivals to be mapped to the same scheduled time). This solves the problem of early buses but this can create a problem for extremely late buses. For our simple stop with scheduled arrivals at 1:00, 2:00, and 3:00 if buses actually arrive at 1:40, 2:40, and 3:40 this matching would mark the 1:40 and 2:40 arrivals as 20 minutes early rather than 40 minutes late but as most users probably do not arrive 20 minutes early for their bus this again does not really match the user experience perfectly. We decided to use this match despite its potential problem as the situation in which it differs from the user experience seemed to be the least likely.

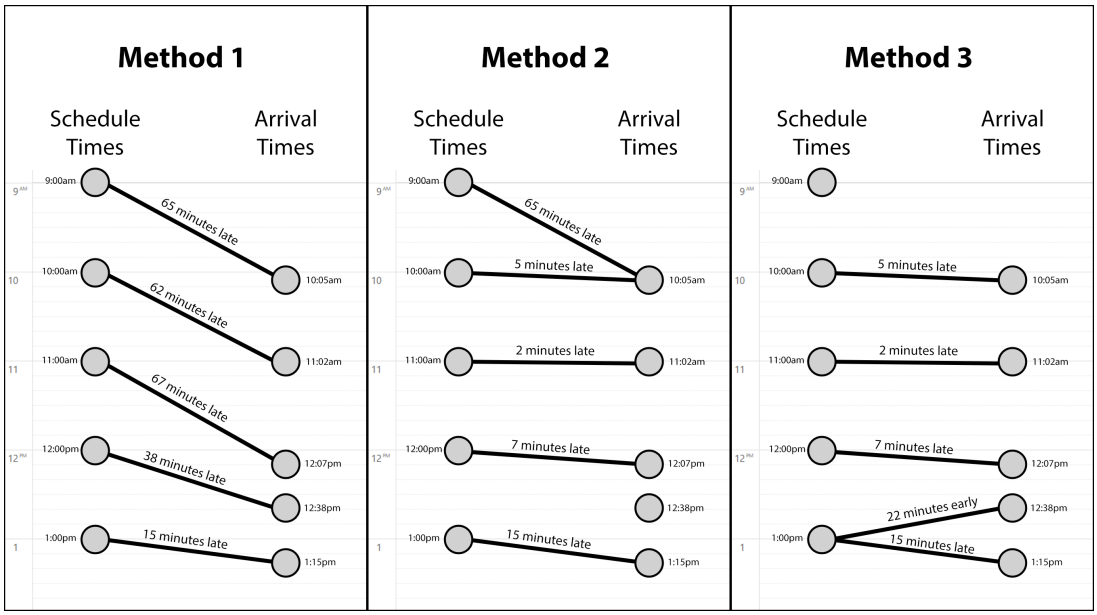

Figure 7: A side by side view of the 3 different matching methods. Each diagram shows how each method matches between scheduled and arrival times differently.

2.4 Defining the Prediction Problem

Once we had defined lateness, we needed to determine what precisely we were trying to predict and exactly what predictors we were going to base our prediction on. Over the course of our work, we ended up focusing on two different problems. We started our work on a simpler problem of predicting average absolute lateness of a stop over a day. After we had finished our work on the first problem, we moved on to trying to predict how late a bus would be for a specific scheduled stop.

2.4.1 Predictors

There are a variety of factors that impact when a bus will arrive (see 1.5 and 1.6), and the more of these factors that are included in a prediction the more accurate it can be. Unfortunately there is no good publicly available way to monitor many of these factors but fortunately weather, one of the factors we hypothesised would be very significant, is widely tracked. As we wanted to create something that would be more useful to a passenger in planning than the real time bus tracker, we decided to use weather predictions rather than real time data so that a prediction could be made a day in advance. The particular weather condition that we thought would be most impactful was precipitation so we used precipitation intensity and type (rain, snow, or sleet) as our weather predictors. It seemed reasonable to suspect that many of the factors that we could not directly track would have some geographical so we included the location of the stop as a predictor. We also included day of the week as a predictor, as stops can have different schedules for different days of the week and it seems likely that traffic and riders have weekly patterns.

2.4.2 The Average Lateness Problem

While we were working on designing our models and determining what we actually wanted to use as predictors, we worked to predict the average absolute lateness of all the buses that arrived at a stop over the course of a day. We used average absolute lateness rather than simply average lateness because we did not want early buses to cancel out late buses. This measure reflected how reliable a bus schedule was for a day but does not allow for direct prediction of when a bus will arrive.

2.4.3 The Scheduled Stop Problem

After finishing our work on the first problem, we moved on to the more useful problem of predicting how late a particular scheduled stop would be. Doing this meant that we had to include the expected arrival time as a predictor. This allowed us to make predictions about when a bus would actually arrive.

2.5 Solving the Prediction Problem

Now that we have defined our two prediction problems, we can implement a strategy to solve them. Since the average lateness problem is easier, we will tackle how to solve it first. Since the problem at hand deals with continuous data points, we will use a ridge regression as our prediction model. Ridge regression is a variant of a linear regression that tries to reduce the variance of the model and prevent over-fitting the sample data. With our predictors (X) and

our targets (y), we reduce the error (L2 loss) as well the model parameters (β) based on the regularization constant (λ). From this, we get the following loss function:

$$L(X, y; \beta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(X_i))^2 + \lambda \sum_j \beta_j^2 \tag{1}$$

With our predictors and our sample data, we can start to train our model. We divide our data into two sets, a training and a testing set (66.7% and 33.3% respectively) while keeping chronological order so there is no data leakage. Using scikit-learn's libraries, we are able to implement a ridge regression model and fit it with our data.

While similar, the scheduled stop problem is a bit more trickier. Instead of generating one model for all the stops and routes, we create a model for each stop, allowing us to get the predicted schedule time and not just the predicted average lateness for that particular day.

2.6 Evaluation

Once we had our models we need to evaluate them to see how accurate they were.. This would allow us to compare our models to one another and to other existing predictions. We chose to use Pearson's r in order to measure the accuracy of our models. Pearson's r is a measure of the correlation between two variables, in our case our predictions and our observations, that ranges from -1 to 1, with 1 being a perfect positive correlation, 0 being no correlation and -1 being a perfect negative correlation. In our context, a Pearson's r of 1 would mean our model is making perfect predictions, a Pearson's r of 0 would mean our models is no better than guessing randomly, and a Pearson's r of -1 would mean our model was perfectly predicting how off schedule a bus would be but would be wrong about whether the bus would be late or early.

2.6.1 Accuracy Over Time

In order to evaluate the accuracy of our models over time we separated out the most recent 33% of our data for testing and used the other 67% for training. We then trained the models on an expanding window of the training data that always started from the beginning of our data and moved the end of the window forward to add more data.

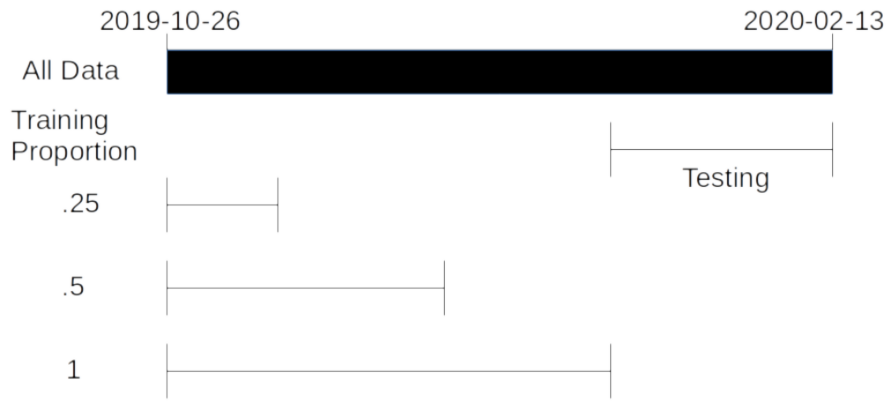

Figure 8: A visualization of our breakup of training and testing data. The training data always starts from the beginning of the data collection period and the testing data is always the last portion of the data. This prevents any data leakage from the training data into the testing data.

2.6.2 Regularization

In order to test for different values of λ, we separated out the most recent 33% of our data for testing then trained models using several different values of λ on the remaining data. Each of these models when then evaluated against the testing data.

2.6.3 Feature Representation

We tried different methods of encoding our features to see how that would effect the accuracy of our model. Our initial model includes a 1291 length one hot vector for the stop code in order to denote the location of the stop. We tried changing replacing this with the GPS location encoded by a vector containing the latitude and longitude of the stop as a predictor.

Our initial model has weather represented as a one hot vector of type (rain, snow, or sleet) and the value of the one hot is the chance of it occurring times the expected intensity. We tried separating the weather into a one hot of type and having the expected intensity times the chance as its own feature and we tried having the one hot of type with intensity and chance each being their own feature which increased our length of our predictor vector by one and two respectively.

2.6.4 Features Included

In order to determine which of our features was actually useful in making predictions, we trained a model while excluding each one of our predictors. This meant we trained four different models. The first used the stop, the route, and the weather as predictors. The second used the day of week, the route, and the weather as predictors. The third used the

day of week, the stop, and the weather as predictors. The fourth used the day of week, the route, and the stop as predictors.

3 Results

After applying our methodology, we now have the final results of the data and their models.

3.1 Accuracy Over Time

The accuracy of the aggregated model over time can be seen in figure 10. The accuracy of a few randomly selected stops over time can be seen in figure 11. These graphs show that the aggregated model and many of the individual stop models reach their peak accuracy quite quickly which indicates that training on more recent data and larger data sets does not increase the accuracy of the model.

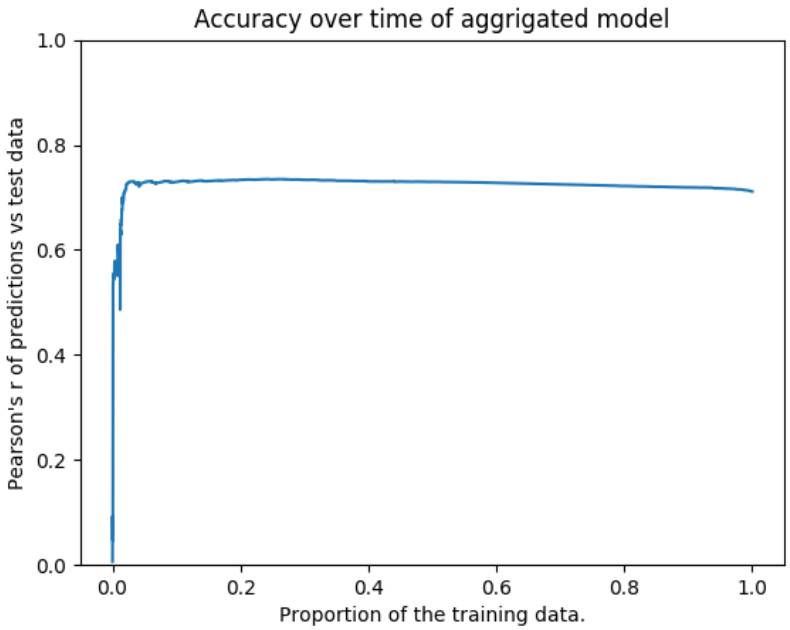

Figure 9: Accuracy over time of the aggregated model.

Figure 10: Accuracy over time of the aggregated model. Very early on we see a sharp spike in accuracy then a quick leveling off in accuracy that stays relatively consistent as more data is added.

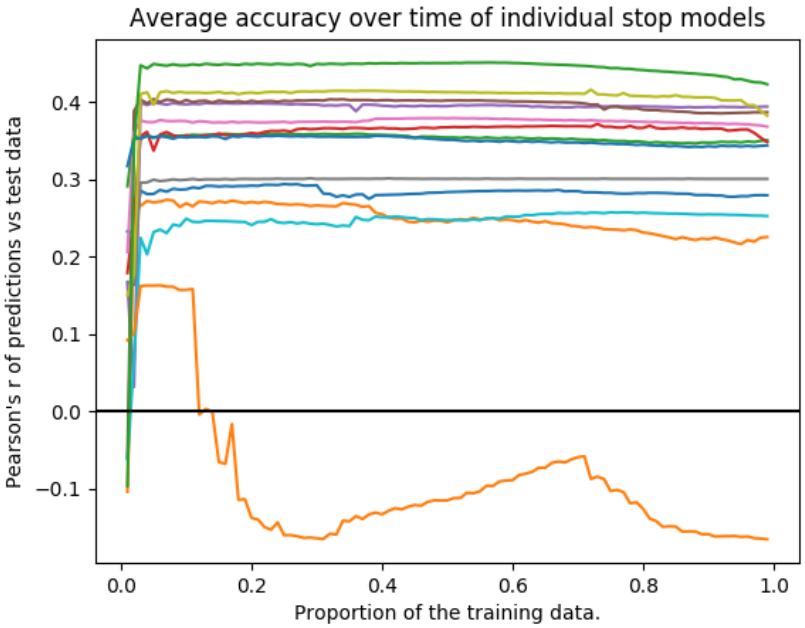

Figure 11: Accuracy over time of 20 randomly selected individual stop models. For most curves, it can be seen that the Pearson's r spikes up initially and levels out as the proportion of training data increases. However, one of the curves behaves different and drops significantly suggesting their is a substantial difference between the training and testing data.

3.2 Regularization

The accuracy of our aggregated model trained with different λ's can be seen in Table 1. As can be seen, different values for λ had almost no impact on the accuracy of the resulting model. Our other models were similarly unaffected by different values for λ.

Accuracy for Different values of λ

λ	Pearson's r
10.0	0.7391341179380303
01.0	0.7402577073825998
00.1	0.7402115666553709
00.01	0.7401825190717123
00.001	0.7401785558333898

Table 1: Accuracy of aggregate models with different values for λ. As can be seen, changes is to the value of λ have very little effect on the Pearson's r.

3.3 Feature representation

The model which had location encoded using GPS data resulted in an Pearson's r of .4238 which is significantly worse than the Pearson's r of .7402 that we got with our stop code based model. The models which had separated precipitation intensity and chance of precipitation

from the one hot of type both produced Pearson's r's of .7403 which did not significantly differ from the original model's Pearson's r of .7402.

3.4 Features included

Table 2 shows the accuracy of models trained while excluding different features. The significant drop in accuracy for the models that do not include stop or day of week indicate that those feature are much more useful predictors than the route or weather which see a very low drop in accuracy when excluding.

Accuracy for models trained while excluding certain predictors

Feature excluded	Pearson's r
None	0.74018
Day of week	0.64526
Stop	0.42385
Route	0.73436
Weather	0.73885

Table 2: Accuracy of the aggregate model with different predictors excluded. When excluding day of week or stop you can see a large drop in the accuracy of the resulting model. When excluding weather or route you see very little change in the accuracy of the resulting model.

4 Visualization

To get a better sense of the data, here are some general visualizations of different bus arrival times and stops.

4.1 Bus Arrivals

In order to see how consistent the buses are, we selected five random scheduled bus arrival times and graphed how late or early the actual bus was for those times over our entire data collection period. The results can be seen in Figure 12. We checked green stop (stop code 3500) to ensure that its extreme values were not a result of mismatching between arrivals and schedules but it was accurate and there simply were buses arriving around the midpoint between two arrivals scheduled an hour apart.

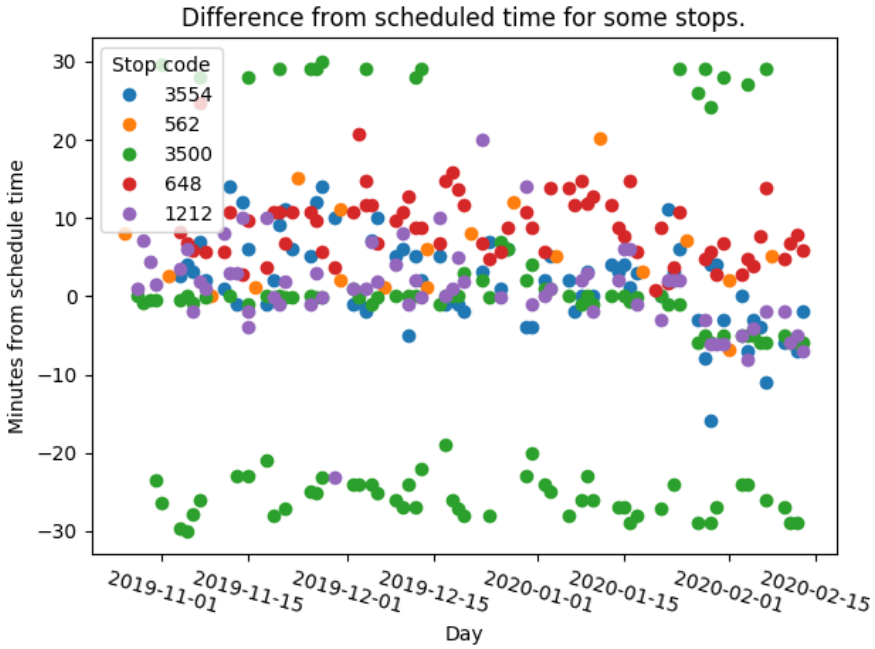

Figure 12: The lateness of five randomly selected scheduled times shown over our data collection period. This shows buses are late much more often then then early and that most of the time the buses are between zero and ten minutes late.

4.2 Bus Stop Lateness

In order to get a sense of what the ranges of values for stops were, we decided to look at the stops with the highest and lowest deviation from the schedule. We found that some of late stops have arrivals based on requests for pickup rather than a defined schedule. This means a rider has to call to be picked up which caused us to have many actual arrivals at that stop and only one scheduled arrival. Then because our matching method goes from actual arrivals to scheduled arrivals, all of those arrivals were matched to the one scheduled arrival time which inflates how off schedule that stop is. The stop with the lowest deviation is “Pleasant St. and Park Ave.”, which is right outside of WPI’s campus. The bus stop is also placed right after an intersection which relates back to bus stops being more efficient after the traffic light at an intersection [6]. Three out of the five bus stops with the lowest deviation are on Chandler St. which cuts right through the center of Worcester from Worcester State University which shows good efficiency of those particular stops and how stops on the same street will have similar lateness.

Highest deviation from schedule for stops in the system

Stop Code	Name	Deviation (min)
1202	PLANTATION TOWERS	88.3853
1146	SHOPPES + BLACKSTONE VALLEY	31.7349
1189	CINEMAS NORTH	30.9378
3087	SOUTHBRIDGE RD + WORCESTER RD	16.0913
6530	MAIN ST + OXFORD CENTER	15.4307

Table 3: The five stops with the highest deviation in our data set. Lateness is the average of the absolute value of the difference between the scheduled and actual arrival time over all arrivals in our data set.

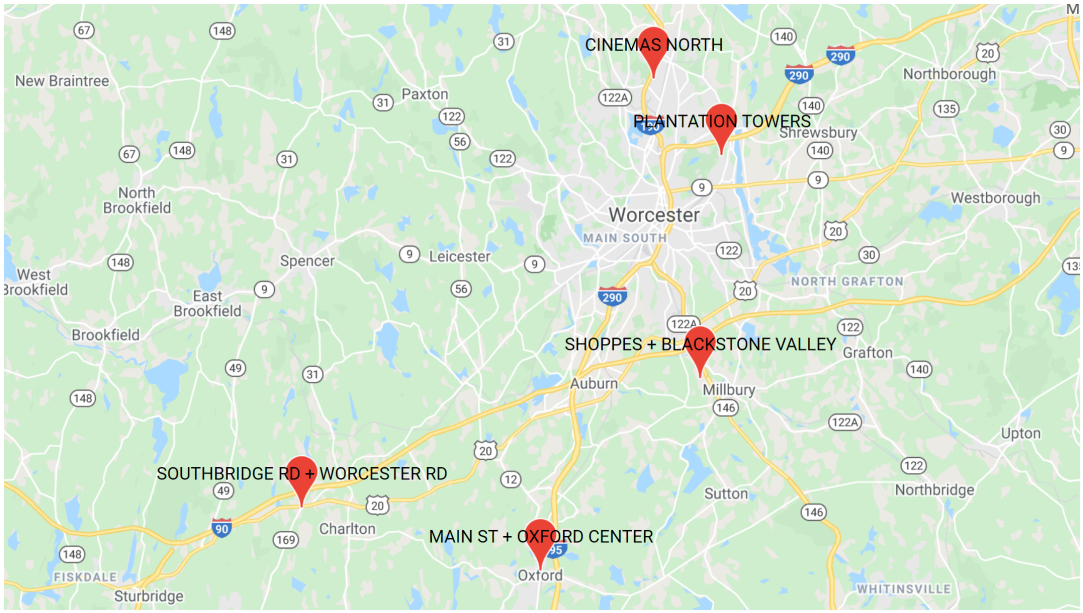

Figure 13: The five stops with the highest deviation plotted on Google Maps. All of the bus stops are outside of the center of Worcester which may be the cause of the increased in lateness.

Lowest deviation from the schedule for stops in the system

Stop Code	Name	Deviation (min)
869	PLEASANT ST + PARK AVE	3.6798
136	CHANDLER ST + BELLEVUE ST	4.0233
279	GRAFTON ST + SOUTH PLAZA	4.1309
130	CHANDLER ST + TATNUCK SQUARE	4.2345
161	CHANDLER ST + TATNUCK SQUARE	4.6566

Table 4: The five stops with the lowest deviation in our data set. Lateness is the average of the absolute value of the difference between the scheduled and actual arrival time over all arrivals in our data set.

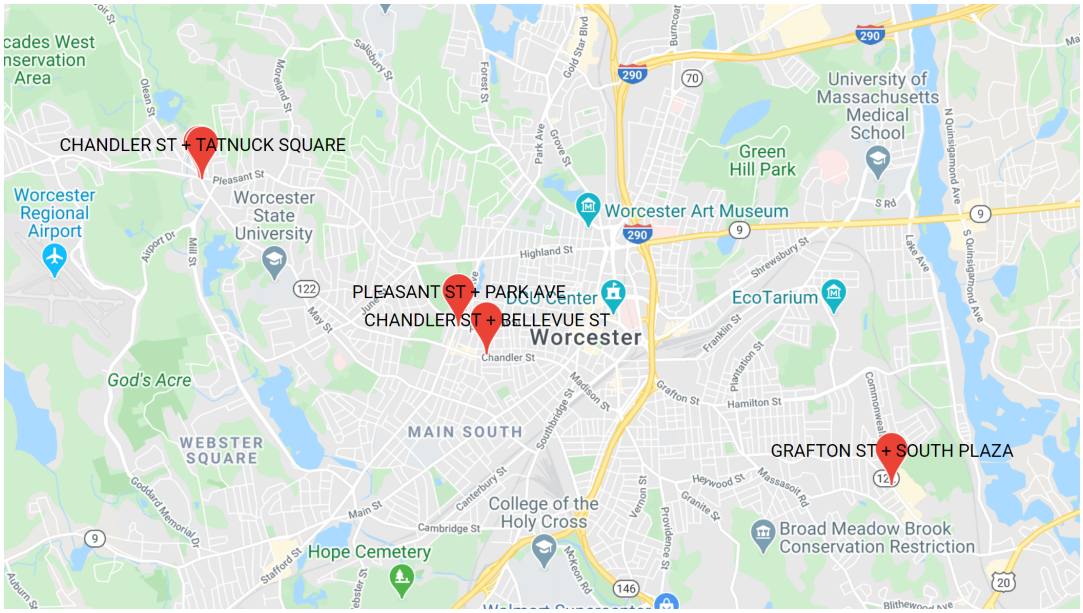

Figure 14: The five stops with the highest deviation plotted on Google Maps. Two out of the five bus stops are in the center of Worcester will the remaining three are along the main streets that cut through the center of the city.

4.3 Bus Stops

The WRTA serves the city of Worcester but also the surrounding cities that make up Worcester county. Figure 15 shows the distribution of bus stops that make up all the routes ran by the WRTA.

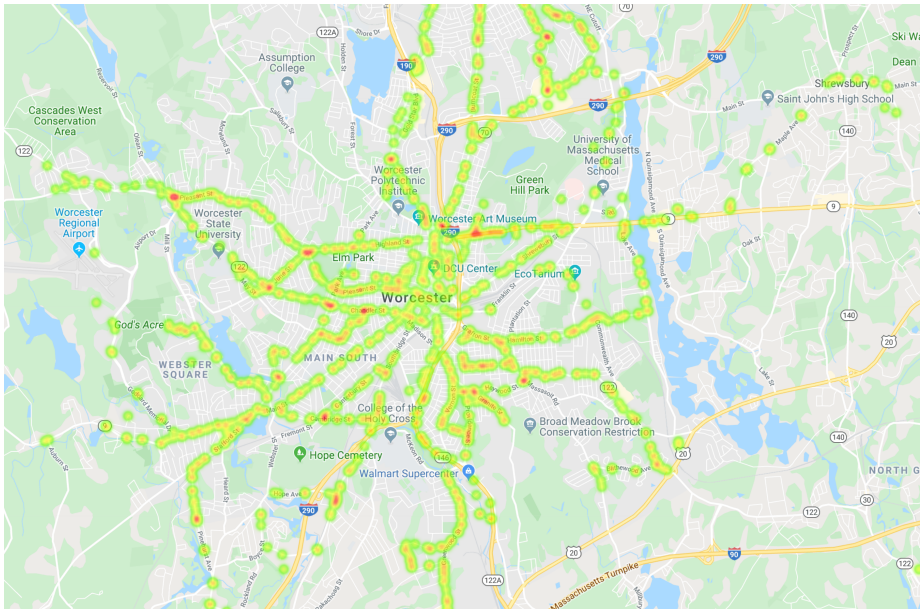

Figure 15: The distribution of the WRTA's stops around the Worcester county. Red representing the most dense and green representing the least dense. It can be seen that there are more stops near down-town Worcester, colleges, and intersections.

5 Conclusion

With the results in mind, we can start to draw some conclusions about our models.

5.1 Key Findings

In the end our final aggregated model had an accuracy of .74. We found that regularization had almost no impact on our results. At the beginning of our investigations we expected that weather would be at least a somewhat useful predictor of bus arrivals as in our experience precipitation tends to slow down all vehicles. However, our results indicate that weather is not a very significant predictor. However, it should be noted that the period during which we collected our data did not see a lot of snow or other extreme weather conditions which one would expect have more impact on road conditions. We found that instead stop and day of week we far better predictors of lateness. We also found that quantity of data was not very significant as our models reach peak accuracy quite quickly and simply hovered around that accuracy as they were trained on more data. Our individual stop models had accuracy's that ranged from -.15 to around .45.

5.2 Practical Value

The project could be taken a step further by being put into the users' hands to assist them with planning their use of public transportation. The majority of our predictive accuracy comes from static data(stop and day of week) so we could generate tables that predict how late a given stop would be on a given day of the week or how how consistent the bus will be with the schedule. This would allow users to make better plans and be more confident about using buses to get to places on time. An example of such a table is given in Table 5. A rider given this may be able to make better plans knowing that the bus will probably be later Tuesday and more on time Wednesday.

Predicted schedule offset(minutes) for stop 451

Scheduled Time	Monday	Tuesday	Wednesday	Thursday	Friday
05:30	5.81	8.03	2.83	4.33	4.56
06:10	5.71	7.93	2.73	4.23	4.47
06:30	5.66	7.88	2.68	4.18	4.42
07:10	5.56	7.78	2.58	4.08	4.32
07:40	5.49	7.71	2.51	4.01	4.24
08:30	5.36	7.58	2.39	3.89	4.12
08:40	5.34	7.56	2.36	3.86	4.10
09:30	5.22	7.44	2.24	3.74	3.97
11:00	5.00	7.22	2.02	3.52	3.75
11:50	4.87	7.09	1.89	3.40	3.63
12:40	4.75	6.97	1.77	3.27	3.51
13:30	4.63	6.85	1.65	3.15	3.38
13:40	4.60	6.82	1.62	3.13	3.36
14:20	4.50	6.73	1.53	3.03	3.26
15:20	4.36	6.58	1.38	2.88	3.11
15:50	4.28	6.50	1.30	2.81	3.04
17:25	4.05	6.27	1.07	2.57	2.81
18:15	3.93	6.15	0.95	2.45	2.68
18:25	3.90	6.12	0.92	2.43	2.66
18:58	3.82	6.04	0.84	2.34	2.58
19:58	3.67	5.89	0.70	2.20	2.43

Table 5: A table for a randomly selected stop that shows how late our model predicts the bus will be for a given scheduled time on a given day of the week. Each cell is colored based on how late the bus will be. The latest bus is in red and an on time would be white.

5.3 Reflection About the IQP

Applying machine learning to a real world problem was a lot different than our previous applications for class projects. The process of searching for, failing to find, and then collecting our own data was a much larger part of the project than we had originally anticipated. This coupled with the process of defining our problem was a very different experience from our work in the classroom where we would be given a data set and told explicitly what problem we were trying to solve.

For future students starting their On-Campus IQP, we would recommend that they set up regularly scheduled meeting times. This helps build habit for working on a more independent style project while continually getting work done. For any project involving a code-base, we would recommend keeping the code documented and that they maintain one file giving a brief description of what script does and why it was created.

5.4 Machine Learning

Our previous experience with machine learning had focused on the theory and how machine learning actually actually functions. However, this project using machine learning did not

require a much knowledge about the math behind the machine learning. Instead a large part of the project was focused on collecting, managing, and selecting data on which to train the model.

References

[1] James Laird Peter Mackie and Daniel Johnson. *Buses and Economic Growth*. Tech. rep. University of Leeds, 2012. URL: https://www.its.leeds.ac.uk/fileadmin/user_upload/News/BusesEconomicGrowth_FINAL-REPORT.pdf.

[2] *Benefits of Riding the Bus*. URL: https://www.therta.com/services/benefits-of-riding-the-bus/ (visited on 02/29/2020).

[3] *Public Transportation Facts*. URL: https://www.apta.com/news-publications/public-transportation-facts/ (visited on 03/04/2020).

[4] WRTA. "2017 Annual Report: Worcester Regional Transit Authority". In: *Worcester Regional Transit Authority* (2017). URL: https://www.therta.com/wp-content/uploads/2018/04/Final-Annual-Report.pdf.

[5] *Economy in Worcester, Massachusetts*. 2016. URL: https://www.bestplaces.net/economy/city/massachusetts/worcester (visited on 02/29/2020).

[6] Pace. "Bus Stop Location & Roadway Design". In: (2015). URL: http://www.pacebus.com/guidelines/04c_stop_location_roadway_design.asp.

[7] R. Choudhary, A. Khamparia, and A. K. Gahier. "Real time prediction of bus arrival time: A review". In: *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. Oct. 2016, pp. 25–29. DOI: 10.1109/NGCT.2016.7877384.

[8] A. Oruganti, F. Sun, H. Baroud, et al. "DelayRadar: A multivariate predictive model for transit systems". In: *2016 IEEE International Conference on Big Data (Big Data)*. Dec. 2016, pp. 1799–1806. DOI: 10.1109/BigData.2016.7840797.

[9] Bin Yu, Huaizhu Wang, Wenxuan Shan, et al. "Prediction of Bus Travel Time Using Random Forests Based on Near Neighbors". In: *Computer-Aided Civil and Infrastructure Engineering* 33.4 (2018), pp. 333–350. DOI: 10.1111/mice.12315. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mice.12315. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12315.

[10] Nikolas Julio, Ricardo Giesen, and Pedro Lizana. "Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms". In: *Research in Transportation Economics* 59 (2016). Competition and Ownership in Land Passenger Transport (selected papers from the Thredbo 14 conference), pp. 250–257. ISSN: 0739-8859. DOI: https://doi.org/10.1016/j.retrec.2016.07.019. URL: http://www.sciencedirect.com/science/article/pii/S0739885915300895.

[11] Steven I-Jy Chien, Yuqing Ding, and Chienhung Wei. "Dynamic Bus Arrival Time Prediction with Artificial Neural Networks". In: *Journal of Transportation Engineering* 128.5 (2002), pp. 429–438. DOI: 10.1061/(ASCE)0733-947X(2002)128:5(429). eprint: https://ascelibrary.org/doi/pdf/10.1061/\%28ASCE\%290733-947X\%282002\%29128\%3A5\%28429\%29. URL: https://ascelibrary.org/doi/abs/10.1061/%5C%28ASCE%5C%290733-947X%5C%282002%5C%29128%5C%3A5%5C%28429%5C%29.

[12] Yu Bin, Yang Zhongzhen, and Yao Baozhen. "Bus Arrival Time Prediction Using Support Vector Machines". In: *Journal of Intelligent Transportation Systems* 10.4 (2006), pp. 151–158. DOI: 10.1080/15472450600981009. eprint: https://doi.org/10.1080/15472450600981009. URL: https://doi.org/10.1080/15472450600981009.

[13] Cong Bai, Zhong-Ren Peng, Qing-Chang Lu, et al. "Dynamic Bus Travel Time Prediction Models on Road with Multiple Bus Routes". In: *Intell. Neuroscience* 2015 (Jan. 2015). ISSN: 1687-5265. DOI: 10.1155/2015/432389. URL: https://doi.org/10.1155/2015/432389.

[14] S. Moridpour, T. Anwar, M. T. Sadat, et al. "A genetic algorithm-based support vector machine for bus travel time prediction". In: *2015 International Conference on Transportation Information and Safety (ICTIS)*. June 2015, pp. 264–270. DOI: 10.1109/ICTIS.2015.7232119.

A Data Samples

Sample of weather data

for time	precip intensity	precip probability	precip type
2019-10-26 00:00:00-04	0.0006	0.2	rain
2019-10-27 00:00:00-04	0.0388	0.94	rain
2019-10-28 00:00:00-04	0.0017	0.18	rain
2019-10-29 00:00:00-04	0.0065	0.45	rain
2019-10-30 00:00:00-04	0.0003	0.11	rain
2019-10-31 00:00:00-04	0.0192	0.97	rain
2019-11-01 00:00:00-04	0.0202	0.58	rain

Table 6: A sample of a weeks worth of weather data

Sample of bus arrivals

arrivaltime	stop code	routenumber	routenumber
2020-01-24 14:50:05.389335-05	1121	30	8405
2019-12-16 18:39:15.195627-05	3130	3	9365
2019-11-27 11:24:53.886827-05	3002	30	8402
2019-12-10 12:30:57.759942-05	726	7	2354
2020-01-18 15:02:09.861566-05	246	2	3362
2019-11-15 15:55:08.600261-05	404	16	2350
2019-11-04 11:44:24.066416-05	1133	8	9407

Table 7: A sample of seven randomly selected bus arrivals from our data set

Sample of stop data

stop id	stop code	stop lat	stop lon	stop name
822027	716	42.25441	-71.83224	MAYWOOD ST + LOVELL ST
821069	921	42.26451	-71.78805	SHREWSBURY ST + CROSS ST
821701	1120	42.31251	-71.79759	WEST BOYLSTON ST + ASSUMPTION AVE
821959	138	42.26136	-71.81971	CHANDLER ST + DEWEY ST
821335	4085	42.1915	-71.76052	MAIN ST + MILLBURY CENTER
821710	1129	42.30102	-71.80017	WEST BOYLSTON ST + ERICKSON ST
2383135	3574	42.260452	-71.800602	SALEM ST AT MAIN LIBRARY

Table 8: A sample of seven randomly selected stops from our data set

A sample of route data

route id	route short name
3663	11
3681	30
3674	24
3870	A
3682	31
3678	27
3680	3

Table 9: A sample of seven randomly selected routes from our data set

Sample of schedule

arrival timestamp	stop id	route id	direction id	stop code	route short name
2019-11-26 20:10:00-05	821001	3689	1	621	7
2020-01-21 13:05:00-05	821242	3690	0	580	8
2019-09-11 06:30:00-04	822166	3681	0	3130	30
2019-12-07 14:07:00-05	821806	3685	0	432	4
2019-11-05 05:55:00-05	821670	3680	0	448	3
2019-12-20 17:36:00-05	821190	3670	1	642	19
2020-01-04 18:25:00-05	821396	3687	0	666	5

Table 10: A sample of seven randomly selected scheduled arrivals from our data set