Bike Sharing in Boston, MA: Is there a relationship between age, user type, start station location and trip duration?

Adam M. Lang

DAT-520-X4545 Decision Methods and Modeling 19TW4

Professor: Shilpa Balan, Ph.D.

28 July 2019

**Abstract**

Bike Sharing (BLUEbikes) is a common method of transportation in the city of Boston despite numerous transportation options including Uber, Lyft, Subway and Bus. The city of Boston studies the patterns of bike usage to help with their goal of quadrupling bike use by 2030. They have looked at who uses the bikes, how long the bikes are being used for, and where the bikes are being used to determine city planning for future transportation infrastructure, money saving options, and environmental conservation. This study looked at the use of the BLUEbikes in the coldest month (January) of 2019 examining the aforementioned factors using Decision Tree Analysis in Rattle and R. A refined analysis was performed with a condensed data set, focusing on the top 5 most popular bike stations rather than all 225, which may have led to some limitations in the decision tree probabilities. However, meaningful results were obtained. It was found that customers are more likely to take longer trips, subscribers take shorter length trips. The average age was younger for customers and more evenly distributed for subscribers. People age 23 and younger were more likely to take trips from the Boylston St. at Jersey St. Station. Decision tree analysis was also validated using Tableau Analytical software.

**Table of Contents**

I

|  |  |
| --- | --- |
| Introduction | Pg. 3-4 |
| Research Question | Pg. 4 |
| Data Appraisal | Pg. 5-6 |
| Techniques/Methods | Pg. 6-9 |
| Evaluation | Pg. 9-11 |
| Decision Tree Model | Pg. 11-13 |
| Results | Pg. 13-15 |
| Limitations | Pg. 15-16 |
| Conclusion | Pg. 16-17 |
| References | Pg. 18-20 |
| Appendix A | Pg. 21-22 |
| Appendix B | Pg. 23-24 |

**Introduction**

Bike share is a program used by many cities for “on the go” transportation and exercise. It is known that 50% of members world wide use bike share systems less than once a month, men use them more than women, and commuting is the most common reason for use (Fishman, 2016). Even with the advent of “ridesharing” services such as Uber and Lyft, as well as regular subway, bus, and taxi services, many bike share programs are still thriving in cities across the United States (Graehler et al, 2018).

I currently live in the greater Boston area where bike share is a thriving entity despite the aforementioned transportation alternatives. BLUEbikes is the bike share program used across the greater Boston area. In 2018 the company’s data reports revealed the most popular station was “MIT at Mass Ave Amherst St.”, there were 1.7 million trips taken, 81% of all trips were taken by subscription members, the most popular stations used were MIT at Mass Ave. Amherst St, South Station -Boston, Coolidge Corner -Brookline, Davis Square - Somerville, and the most popular expansion station was Boylston St at Jersey St. (BLUEbikes, 2018).

BLUEbikes continues to publish their monthly user data for open public view. Although the company publishes yearly data analytics there isn’t much detailed analytical data looking at the coldest month of the year in Boston, which is January (Current Results, 2019). With the numerous other travel options available (Uber, Lyft, T-subway trains, taxi, car) and the city of Boston now known as having the #1 worst traffic in the United States (Gorzelany, 2019), it is worth taking a closer look at the data on the BLUEbikes program. The company continues to add bike docks in and around the city to foster more usage of the bike system instead of the other transportation options.

Recent studies discuss bike share dock locations being near bus, subway and taxi stations and how this may influence someone to take a shorter ride and use the bike to complete their journey. Or it may influence them to use the bike instead of other transportation means. It is noted that “in general, a one-unit increase in the frequency of public transit usage is significantly associated with a 4.0% increase in the probability of bike sharing usage and a 1.4% increase in the frequency of bike sharing usage (Zhang, 2018).” In addition, it is noted that bike share docks located near busy subway stations and bicycle infrastructure see overall greater utilization, and that greater population and employment generally predict greater bike usage. It is also seen that residential populations are associated with more trips by subscribers and on both weekdays and non-working days (Noland, 2016).

**Research Question**

It is worth looking at whether bike share dock locations near major transportation centers (subway stations, bus stations, transportation hubs, colleges, businesses) see more bike use and for longer periods of time. All five of the most popular bike stations for BLUEbikes are located near bus and subway stations as well as colleges/universities and businesses/landmarks (BLUEbikes, 2019). The BLUEbike data states that 81% of trips in 2018 were taken by subscribers (BLUEbikes, 2018). It begs the question as to whether subscribers are the ones that are using the bikes for a longer period of time and from what stations. Since each of the stations are located near other forms of transportation, colleges and businesses, how do we know what age groups are using the bikes for shorter vs. longer trips from which stations? Is there a way to infer from the data the answers to these questions? For my research question I have decided to investigate the following:

**“During the coldest month of the year (January) in Boston, is there a relationship between age, user type (subscriber vs. customer), bike start station location and trip duration?”**

**Data Appraisal**

For this research question I will utilize the open data set published by BLUEbikes: <https://s3.amazonaws.com/hubway-data/index.html>. The dataset is a CSV file, entitled “201901-bluebikes-tripdata.zip, February 11, 2019 9:42:57 AM. This means it is the data from the first month of 2019 (January). The dataset is broken down into excel spreadsheet format: it has 69,873 total rows, which takes up 2.53 MB of data.

The BLUEbike data is published once a month by BLUEbikes and the cities where the bikes are located in Boston, Cambridge, Somerville, Brookline; these are the cities that own the bike sharing program. The cities study the data as it is part of their project “Go Boston 2030” where their goal is to increase bicycling in the city limits fourfold from what it is now. They study the data and how it relates to population and job growth, easing income inequality and making transportation more affordable, and Climate Resilience as more than a quarter of Boston’s greenhouse gas emissions are from transportation (Boston.gov, 2019).

The data set includes (BLUEbikes, 2019):

* Trip Duration (seconds)
* Start Time and Date
* Stop Time and Date
* Start Station Name & ID with latitude and longitude
* End Station Name & ID with latitude and longitude
* Bike ID
* User Type (Casual = Single Trip or Day Pass user; Member = Annual or Monthly Member)
* Birth Year
* Gender, self-reported by member

I am going to be using the utilities of data mining with the R program Rattle, and RStudio. It is well known that R/Rstudio can handle large data files up to 2GB of memory on a windows operating system so I should be able to analyze this data without having to resort to using something larger such as Hadoop (R documentation – memory limits). The variables I will be exploring include:

* Trip duration (in seconds): This will tell us the amount of use of each bike.
* Birth Year
* Bike station Names (start stations)
* User type (subscription or customer)

I hope to be able to classify if more subscribers are using the bikes, certain stations are seeing more usage of the bikes, if more customers use certain bike stations than subscribers and vice versa. Also, are the highly trafficked bike stations near public transportation hubs? Are they residential areas? Are subscribers using the bike hubs in public transportation or residential areas? Are these areas of high traffic?   The limitations of the dataset are the duration of each trip is in seconds not minutes or hours and the start station and end station names are categorical variables and can’t be ranked numerically. The longitude and latitude in RStudio can’t be linked to a geographic map unless you use the R ‘Geosphere Package’ (Hijmans, 2019) but I will instead be using Decision tree analysis provided by the data mining capabilities of Rattle.

**Techniques/Methods**

**a.** **Data Preparation**

Initially the dataset was loaded into Rattle and assessed, which provided a summary of the dataset using a partition of 70/15/15. Immediately noted were 69,872 observations with 14 input variables. Within those variables start and stop time both had well over 30,000 distinct inputs. “bike id” had over 2,000 and “trip duration” had over 3,670. Focusing on the variables I wanted to examine, “start station name” had 225 and “end station name” had 226. “Usertype” was only 2 and “birth year” was 66. Initially I ignored the variables with too many inputs (i.e. start and stop time, bike id, as well as the numeric variables longitude and latitude) as this would not produce a decision analysis on the data I was interested in.  I executed a decision tree analysis to test the data set using the target variable as usertype (this value must be less than 5 and I was not going to include gender), and variables: start station name, end station name, trip duration, birth year. Right away I could see that the size of the start station name and end station name being upwards of 225 variables caused over fitting of the model. The initial error of the root node was 0.069515. I experimented with this by leaving out start station name and then end station name, but regardless the 225 or more unique variables was causing over fitting of the data set training of the decision tree model.

**b. Data Manipulation**

I decided to experiment by manipulating the data set and cleaning it. I really wanted my research question to focus on what bike stations were influencing trip duration just as much as the variables for age and user type. However, it was clear that having 225 or more stations was not going to give any clear answers. I decided to focus the data set to start stations only since the bikes all have to start somewhere but not necessarily always end at a different station. I first tried narrowing the start station column to the most popular start station of 2018, “MIT at Mass Ave. Amherst St.”, then adding the second most popular start station “South Station”. I realized that although my root node error was less (0.05064), the data set was ignoring the start stations as a test variable in the modeling. I decided next to try expanding the start stations to the top 5 most popular start stations. This produced 5,470 observations and 14 input variables. This also cut the data set significantly from the initial 69, 872 observations, which made the data more workable. This seemed to train the data set without over fitting and gave me answers regarding birth year, start stations used by subscribers vs. customers and the average length of trip taken by certain age groups, subscribers, customers and from what stations.

**c. Model Optimization**

Initially I used the decision tree and its rules to evaluate model optimization. I then focused on the parameters, root node error and cross validation error. I started out each model with min split of 20, min bucket of 7, max depth of 3 and complexity of 0.0100. I would see what the input variables for the tree would be and the number of root nodes produced as well as the cross validation errors. I would then run a model with maximum complexity of 0.0000 and a minimum complexity of 1.0000. I then settled on working with just the min split, min bucket, and max depth keeping the complexity at standard of 0.0100. In the end I realized that min split of 20, min bucket of 7 and max depth of 3 with complexity of 0.0100 were appropriate producing cross validation errors of: 1.00, 0.57 and 0.59. The initial larger dataset with close to 70,000 observations produced cross validation errors of 1.00 and 0.95.

**d. Ethics and Legality**

The preparation and manipulation of this data set was done so under the current ethical and legal standards for open data sets released by the local, state and federal government (open data enterprise, 2016). Bike sharing data is part of the open transportation data sets the government releases for public use. Cities such as Boston are using this data to evaluate environmental issues, affordable transportation, health and fitness, vehicle accident data, and local economic growth (Boston.gov, 2019). The main concern with using this data has been the privacy risks. In 2014 a researcher was able to track down an individual bike share user in the London bike system, by way of his customer id and birth date, longitude and latitude coordinates of the bike station, time of day bike was used and link this to social media used at the same time (Mirani, 2014). This led to all open bike data sets removing customer id numbers (which this current data set does not have), which makes it harder to track someone. I still think that mining these open data sets does have its risks as you are using real user data. I however am not going to be using latitude and longitude, bike id, time of day and end station information so there will be no way to track a customer without these variables. I am also doing a similar general analysis of the data set that BLUEbikes publishes monthly and yearly so I am following the same standards they are as well as the city of Boston (BLUEbikes, 2018 and Boston.gov, 2019).

**Evaluation**

**a.** **Best Choices**

Decision trees are best used for classification problems with smaller data sets (Datla, 2015). I am going to institute a “bottom-up” decision tree which partitions the data into classes based on the results. This is best for research questions where we don’t know the outcome yet. “In a bottom-up approach we have a priori information on which group of instances belongs to a given node of the tree. It means we know the result of each node split before even generating the separating hyper plane (Barros et al, 2011).” A priori knowledge refers to being able to deduce information from the splits of the tree in a bottom up approach. We can’t use top-down approach because that requires us to already have the information and we prune the tree based on that. If done correctly a bottom up tree does not need pruning (Barros et al, 2011).

Initially the dataset for this study was quite large but with the manipulation to a smaller set this bike share data set should be excellent for tree analysis. Decision trees are known to have a very good combination of error rate and speed when compared to other supervised learning algorithms (Lim et al, 2000). Tree models are also a logic-based technique that supports a good understanding of the concepts underlying the data with pattern recognition and transparency of the results (Murth, 1998). Logic-based methods, such as decision trees, tend to perform better when dealing with discrete/categorical variables for which this data set has both (Kotsiantis, 2007). Discrete choice models such as decision trees are also widely used in transportation data evaluation to evaluate what factors people consider when choosing a mode of transportation such as bike, subway or taxi (Sekhar et al., 2016).

**b. Agility**

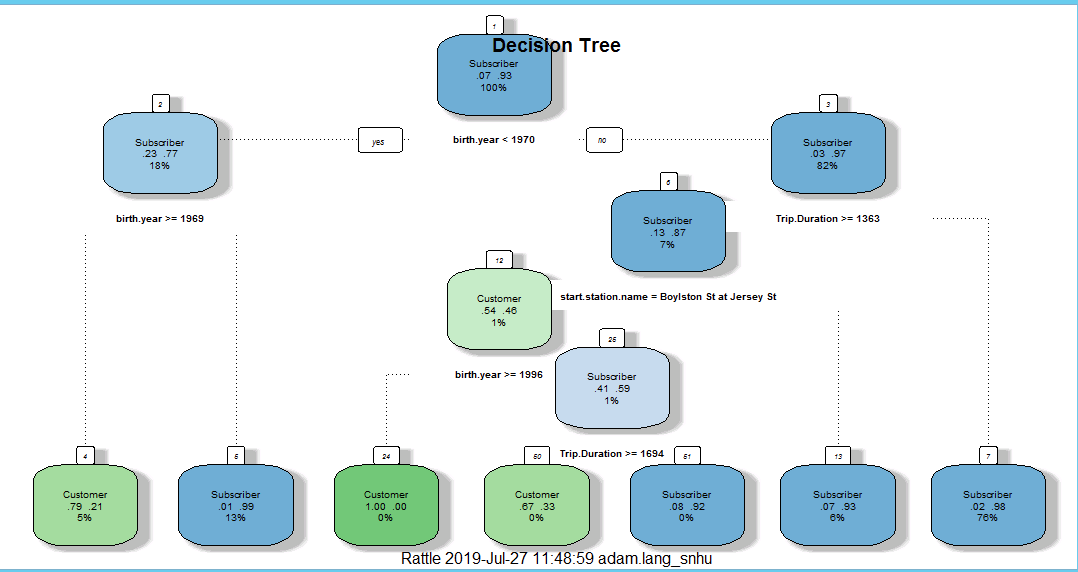
Being able to understand the discrete factors that cause certain groups of people (i.e. age, subscriber, customer, station location) to take longer or shorter bike trips will help organizations such as the city of Boston understand who is using the bike sharing program, and perhaps figure out places to add new bike stations to improve transportation options, as well as examine population and job growth, help with climate change, and even ease income inequality (Boston.gov, 2019). Knowing factors that consider longer vs. shorter bike trips may influence the addition of a new subway or bus station or a public park construction. City planning plays a huge role in having this data analyzed.

**c. Ethical Standards**

Since this dataset was obtained from a local government owned entity (BLUEbikes, Boston.gov), I know that using the data is considered ethical and legal. The city of Boston wants organizations of any type to have the data available to help with forecasting for city planning and other projects that affect urban development. I have also chosen to focus on variables in the data set that would not directly identify a person’s exact movement, which could be tracked via social media (Mirani, 2014). I am not using bike id, longitude/latitude, time of day, gender and do not intend to publish this data. I am using start station name and birth date, so while those are potential ways to track a subscriber or customer on their regular use of the bikes at certain stations, I do not intend to publish this data. Furthermore, the decision tree generalizes the splits and does not single out individual users from the data set, which makes this more ethical.

**Decision Tree Model**

Using the Model function in Rattle, a decision tree (Rpart) was created using “Usertype” as the target variable and the following as input variables: Birth Year, Trip Duration, Start Station Name. **(See Appendix A for model decision rules and error rates).**



**Fig. 1. Decision Tree Model**

The decision tree was constructed with Rattle and R. The parameters were set at min split 20, min bucket 7, max depth 3, complexity 0.0100. The root node error was 0.066876 (See Appendix A, Fig. 2.). Additional code was run to enlarge the font of the nodes: *FancyRPartPlot(crs$rpart, main = "Decision Tree", cex = 0.8, tweak = 1.7* (Williams, 2016). What does this decision tree tell us? Lets first remember that trip duration is in seconds and birth year is in actual year (i.e. 1970).

Starting at the root node if your birth year is less than 1970 this is a 7% chance you are a subscriber and 93% chance you are not a subscriber. Moving down the left side of the decision tree, if you are a subscriber whose birth year is <1970, there is a 23% chance you are a subscriber whose birth year is >=1969 but 77% chance your birth year is not >=1969. If your birth year is >=1969 and you are a subscriber there is then a 79% chance you are a customer and 21% chance you are not a customer. If your birth year is not >= 1969, there is a 1% chance you are a subscriber and 99% chance you are not. Out of those who have a birth year >=1969 or not, 5% are customers and 13% are subscribers. This end of the tree makes up 18% of the total.

Moving to the other side of the decision tree if we start at the root node again, and your birth year is not less than 1970 (so it is greater than 1970), you have a 3% chance of being a subscriber and 97% chance you are not. If you are a subscriber and your trip duration is >=1363 seconds, you have a 13% chance of being a subscriber and 87% chance you are not. If your trip duration is not >= 1363, therefore it is <1363, your birth year is >=1970 and you have a 2% chance of being a subscriber and 98% chance you are not. This is 76% total of the data set.

Moving back to the center of the tree, if your start station is Boylston St. at Jersey St. there is a 54% chance you are a customer and 46% chance you are not. This is overall probability of 1%. If your start station, is not Boylston St. at Jersey St., your start station is then going to be Coolidge Corner, Davis Square, MIT at Mass Ave, or South Station (i.e. the rest of the stations). If this is the case, your birth year is likely >=1970 and trip duration is >=1363. There is a 7% chance you are a subscriber and 93% chance you are not. The overall probability is only 6% of the data set.

Going back to if you are a customer whose start station is Boylston St. at Jersey St., if your birth year is >=1996, there’s a 100% chance you are a customer. Don’t forget that this also states that your trip duration is going to be >=1363 seconds. If your birth year is not >= 1996, so therefore it is <1996, there is 41% chance you are subscriber and 59% chance you are not. Moving forward from this, if your trip duration is then >=1694 seconds you are 67% chance customer and 33% chance you are not. The total probability is 0%. If you are a subscriber whose birth year is <1996 but trip duration is <1694 seconds, there is 8% chance you are subscriber and 92% chance you are not.

**Results**

Are the results of the decision tree reasonable? We have to make sense of all this information. To cross validate the information in this decision tree I ran some analytics in Tableau **(See Appendix B)**. The results showed that subscribers are of a broader age range at all 5 stations. Customers tend to be younger at all 5 stations especially South Station, Boylston St, and MIT at Mass Ave. The Tableau analytics also showed that Customers on average take longer trips from all 5 stations but the longer trips are from Coolidge Corner start station. Subscribers take on average shorter trips from all 5 stations.

If we look at the probabilities of this decision tree model is it telling us the same thing that Tableau is? Right away if we look at the right side of the tree it is telling us that you are more likely to be a customer than a subscriber and take a longer trip, but the overall probability is low (1%) that this trip is being taken from Boylston St. at Jersey St. Station. There is an almost 94% chance you are a customer taking a longer trip from the other 4 stations according to the decision tree so this does comply with our Tableau results. The longest trip on the decision tree mapped is >= 1694 seconds and there is a 67% chance of being a customer which again shows that customers are more probable to take longer trips.

What about age groups? If your birth year is >= 1969 you have a 79% chance of being a customer, and if your birth year is >=1996 you have basically a 100% probability of being a customer. This validates that customers tend to be younger overall (between age 50 and younger). Going back to the tree, if you are a customer whose start station is Boylston St. at Jersey St., if your birth year is >=1996, there’s a 100% chance you are a customer. Don’t forget that this also states that your trip duration is going to be >=1363 seconds. So we can infer that there is a large probability that people age 23 and younger will use the Boylston St. station and take a trip longer than 1363 seconds.

Looking at subscribers, if your birth year is <1970 you have an 18% chance of being a subscriber, but if your birth year is >1970 you have an 82% chance of being a subscriber. If your birth year is <1969 you have a 13% chance of being a subscriber (left side of the tree). It seems the probabilities are more evenly distributed for age groups of subscribers across the board. If we actually add the probabilities together for subscribers: birth year <1969 is .01, birth year <1970 is .07, birth year >1970 is .03, birth year is less than 1996 is .41 and .08. Total this would be about 60% probability which is more than half of the data set is distributed to subscribers age groups.

Accuracy of this model can be tested by looking at the error rates. The root node error is 0.66876. The cross validation errors for the nodes are: 1.00, 0.57 and 0.59 respectively. During development of the model these were actually higher at 0.6-0.9 so by tweaking the model with the parameters I was able to bring the cross validation errors down. Being that the decision tree chooses the splits based on the parameters, and also the probabilities of each variable occurring, it can only be as accurate as the data and the manual changes in the parameters that we make. However, I think that by validating the model with a broad analytical evaluation in Tableau helped prove that some of the tree makes sense based on what the data set is telling us.

I do still think there are confounding factors that are missing or present that could affect the outcome. The fact that only 5 start stations were used in the final data set (after cleaning) will affect the probabilities of each station being influenced by customer or subscriber and age groups. I used the top 5 most popular ride stations from 2018 and this data set is from January 2019, which a new trend could be starting. There could have been an error with choosing usertype as the target variable. However, the only other option is to choose “gender” as a target variable which did not work for some reason possibly because the values were 0, 1, 2 which are actually categorical and represent male, female or did not say. Rattle chooses the target variable based on values of less than 5 and the only two possibilities were usertype or gender. What if we had a different target variable would that have made the error rate change? I think it would. I also did not include the other variables in the data set such as longitude, latitude, end station name, bike id, start and stop times. Including those factors probably would have made the tree over grow to a random forest and make the probabilities even more error prone.

**Limitations**

A major limitation of this study is that it is a smaller data set that I cropped from the original larger data set. I decided to examine only the top 5 most popular bike stations (from 2018) rather than all 225 possible bike stations. The reasoning behind this was the original data set with 225 start and end stations made the data set over fit the training set and the decision tree was not able to be constructed in an understandable manner. I tried experimenting by using just one start station or two start stations but again this didn’t allow the bike stations variable to be included in the decision model and so I settled on using the top 5 stations. This is a huge limitation as it could have prevented me from finding probabilities for the other bike stations. That is a major limitation of the decision tree model in that it pre-selects the variables it will partition, but it can’t automatically partition variables. What I am getting at is I can’t train this decision tree to analyze each of the 225 bike stations to see which had the highest vs. lowest probabilities for customers and subscribers and birth year; the tree it would produce would be top down and fill an entire wall of a room. The major limitation of using decision tree analysis is it works very well for smaller data sets which I saw directly with this project by having to trim the original data set to make the decision analysis work.

I do wonder if perhaps there were more subscribers using bike stations that I did not include? Perhaps a certain age group was more likely to use a bike station I did not include? Another limitation was that I was not able to include longitude and latitude of the bike stations since the rattle algorithm for the CART model can’t understand the numeric values. I would have to use the R geocode program to actually evaluate these values. Another limitation was not including the end bike stations. Therefore I was not able to evaluate trip duration completely. If I was able to evaluate start and stop stations I could have perhaps seen different probability of taking longer or shorter trip based on the start vs. end station.

**Conclusion**

This study utilized decision tree analysis to study BLUEbikes the bike sharing program in Boston, MA. Decision tree analysis provided probabilities that showed customers were more likely to take longer trips than subscribers, but subscribers were more likely to take shorter trips. Although there are certain limitations to this data analysis due to reducing the original data set from 225 start bike stations to 5 of the most popular bike stations, there were still valuable conclusions identified. Customers were more likely to be of younger age, and subscribers had a broad age range. An interesting finding was that if you are a customer whose start station is Boylston St. at Jersey St., if your birth year is >=1996, there’s a 100% chance you are a customer and take a trip longer than 1363 seconds. So we can infer that there is a large probability that people age 23 and younger will use the Boylston St. station and take a trip longer than 1363 seconds. Being that Boylston St. Station is near the Fenway Park area, a major hub for transportation/night life/colleges & Universities/public parks, and this station was the #1 most popular new station added in 2018 (BLUEbikes, 2018), perhaps BLUEbikes and the city of Boston should consider adding more stations in this section of the city especially since there is a large probability younger people will use the bikes. And knowing from our tree that younger people are more likely to be customers and take longer trips, this could increase the number of bike trips in this section of the city, which is thriving with the need for on the go transportation. Overall further decision tree analysis should be performed using all 225 bike stations to more adequately assess if age, user type, station location and trip duration are related.

References

Barros et al. (2011) A bottom-up oblique decision tree induction algorithm. [2011 11th International Conference on Intelligent Systems Design and Applications](https://ieeexplore.ieee.org/xpl/conhome/6112291/proceeding). Date of Conference: 22-24 Nov. 2011

BLUEBikes (2018) 12 Days of Data. Retrieved from: <https://www.bluebikes.com/blog/12-days-of-data>

BLUEBikes System Open Data Sets. Retrieved from: <https://www.bluebikes.com/system-data>. And: <https://s3.amazonaws.com/hubway-data/index.html>

Boston.gov (2019) BIKE DATA. Measuring how many people ride bikes in the city. Retrieved from: <https://www.boston.gov/departments/boston-bikes/bike-data>

Campbell, K. B., and C. Brakewood. Sharing Riders: How Bikesharing Impacts Bus 22 Ridership in New York City. Transportation Research Part A: Policy and Practice, Vol. 23 100, 2017, pp. 264–282.

Current Results (2019) Boston Temperatures: Averages by Month. Retrieved from: https://www.currentresults.com/Weather/Massachusetts/Places/boston-temperatures-by-month-average.php

Datla, MV (2015) "Bench-Marking of Classification Algorithms: Decision Trees and Random Forests - A Case Study using R," (in English), 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-Tact-15), Proceedings Paper p. 7, 2015.

Fishman E. (2016) Bikeshare: A Review of Recent Literature. *Transport Reviews.* Vol. 36(1):92-113.

Graehler et al. (2018) Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?  Paper Presented at the 98th Annual Meeting of the Transportation Research Board. January 13-17, 2019, Washington, D.C.

Gorzelany J. (Feb 11, 2019) Do You Live In The City With The Worst Traffic? Retrieved from: <https://www.forbes.com/sites/jimgorzelany/2019/02/11/here-are-the-u-s-cities-suffering-the-worst-traffic-congestion/#17f4f3376e36>

Hijmans R. (May 26, 2019) Package ‘geosphere’. Retrieved from: <https://cran.r-project.org/web/packages/geosphere/geosphere.pdf>

KOTSIANTIS, SB (2007) Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249-268.

Lim et al. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. Machine Learning, journal article vol. 40, (3): 203-228.

Mirani, L. (April 16, 2014) London Transport’s Bike-Share Privacy Slip Raises Concerns. Retrieved from: <http://skift.com/2014/04/16/london-transports-bike-shareprivacy-slip-raises-concerns>

Murthy, SK (1998) Automatic Construction of Decision Trees from Data: A Multidisciplinary Survey. Data Min. Knowl. Discov., vol. 2, (4):345-389.

Noland et al. (2016) Bikeshare trip generation in New York City. Transportation Research Part A: Policy and Practice. Vol. 94: 164-181.

Open Data Enterprise (2016) Briefing Paper on Open Data and Privacy. Retrived from: <http://reports.opendataenterprise.org/BriefingPaperonOpenDataandPrivacy.pdf>

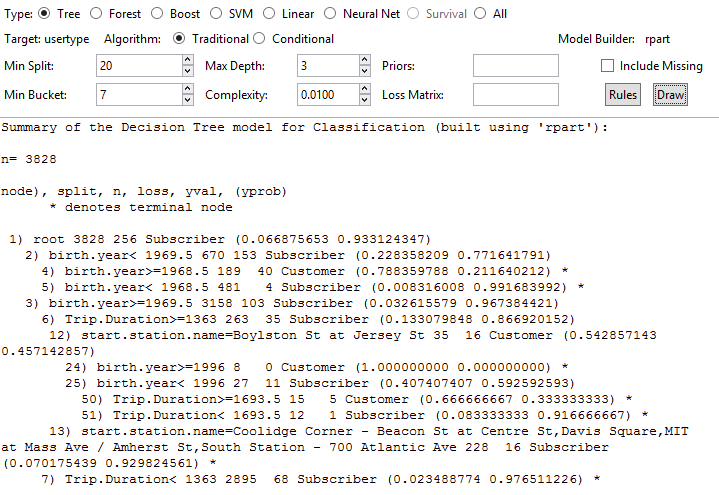
R Documentation – memory limits in R (n.d.) Retrieved from: <https://stat.ethz.ch/R-manual/R-devel/library/base/html/Memory-limits.html>

Sekhar et al. (2016) Mode Choice Analysis Using Random Forest Decision Trees. Transportation Research Procedia. 17:644-652.

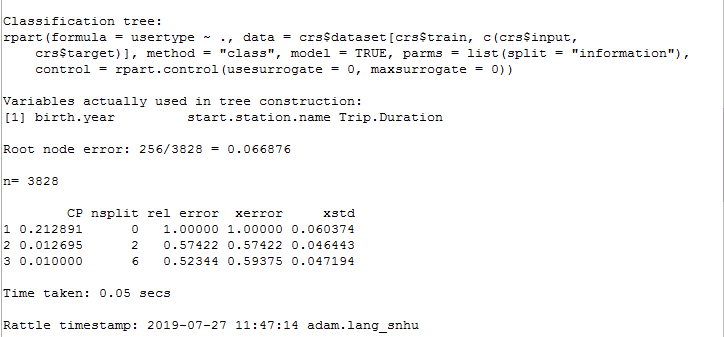
Williams G (2016) Decision Tree - node labels font size. Retrieved from: <https://groups.google.com/forum/#!topic/rattle-users/-jJmx-K0ydY>

Zhang et al. (2018) Associations between Public Transit Usage and Bike Sharing Behaviors in the United States. Sustainability. 10(1868):1-20.

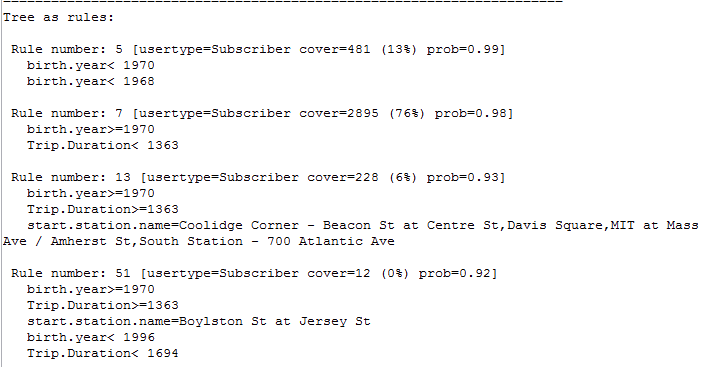
**Appendix A**

****

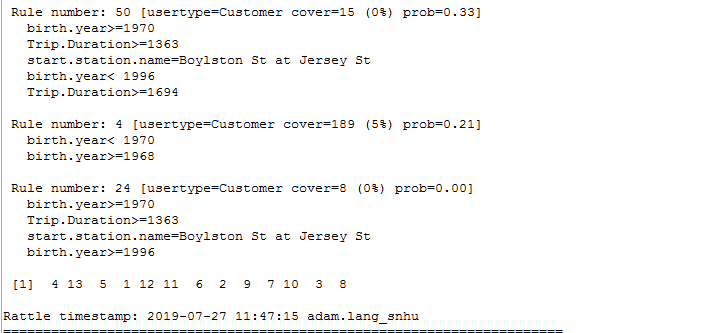
**Fig. 1. Summary of Decision Tree model for Classification**

****

**Fig. 2. Error Matrix**

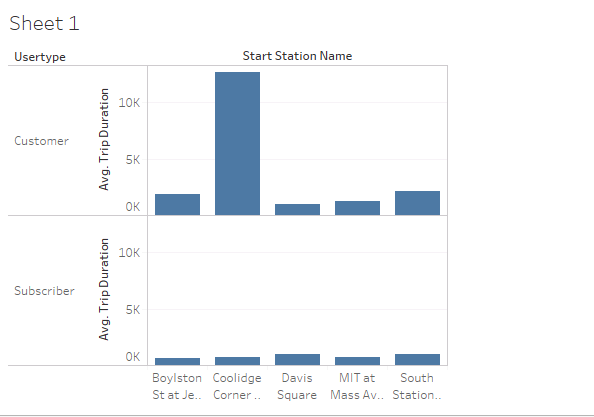
****

**Fig. 3a. Tree as Rules**

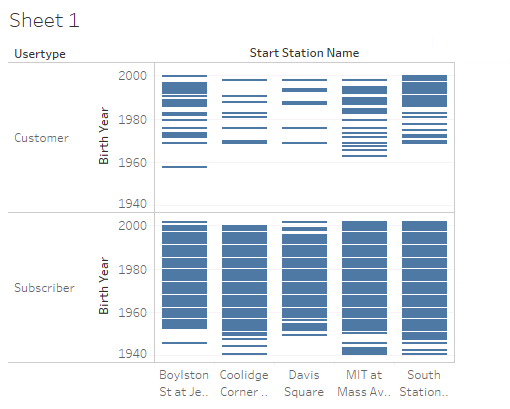
****

**Fig. 3b. Tree as Rules (continued)**

**Appendix B: Tableau Analytics**

****

**Fig.1. Tableau analytics: Customers on average take longer trips from all 5 stations but the longer trips are from Coolidge Corner start station. Subscribers take on average shorter trips from all 5 stations.**

****

**Fig. 2. Tableau Analytics: Subscribers are of a broader age range at all 5 stations. Customers tend to be younger at all 5 stations especially South Station, Boylston St, and MIT at Mass Ave.**