| Who Will Take The Light Rails?    Spring 2022 PUBH 7475/8475 Project Report | system-wide-stacked.png |
| --- | --- |

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Jordan Aron

# Kwangho Baek

# Minh Nguyen

## Introduction

In the Twin Cities Metropolitan Area, two different transit services operate: buses and light rail transit (LRT) systems. It has been shown that the demographics and spatiotemporal ridership of LRT users and bus users significantly differ (Metro Transit, 2017). In terms of public transport mode choice, it is believed that taste heterogeneity matters. For example, a transit traveler concerned about in-vehicle crowding will avoid taking LRT in the morning peak if a similar but uncrowded bus alternative is available. Albeit not universally applicable, some heterogeneity can be inferred by a combination of demographic characteristics of passengers.

Therefore, predicting how many people will use the light rail transit (LRT) systems is a complex problem. It is also an important task for public transit planners, as the expected ridership is crucial in estimating the infrastructure’s social benefit. There have been many projects to predict the mode choice and ridership, but they have not included machine learning models.

Conventional transportation analysis frameworks that predict people’s mode choice have limits in considering taste heterogeneity effectively. Old models adopt only a few demographics as binary/ternary variables. This lack in complexity often leads to inaccurate prediction results (Train, 2008). More recent models exploit latent variables using the structural equation scheme, but are highly prone to bias by manually constructing such structures (Ben-Akiva et al., 2002). In recent years, machine learning has become a vital prediction approach due to the growth of real-time transportation data such as GPS tracking information. Additionally, machine learning can flexibly utilize demographic characteristics for predicting people’s mode choice (Koushik et al., 2020).

We fit four machine learning methods in order to accurately predict LRT use: logistic regression, decision trees, random forest, and neural networks. When deciding on which model to apply, a tradeoff must be made between a focus on predictive power or ease of interpretation. In this project we apply two models that have a much clearer interpretation (logistic regression and decision trees) and two models that favor prediction (random forest and neural network). When considering the use of such models, interpretation-focused models may be preferred unless there is a large difference in prediction accuracy. Afterwards, we compare and evaluate the results from our models to achieve the best approach to classify if a potential user used LRT as part of the public transportation use.

## Data

We used the 2016 Transit on-board survey data offered by Minnesota’s public transportation operator, Metro Transit. The dataset features 174 surveyed attributes (variables) of 30,491 transit passenger observations with full-scale travel and user characteristics, including travel origin, destination, trip purpose, number of transfers, used routes, and various socio-demographic factors. The majority of the demographic fields are categorical variables. For example, some data fields indicate which time window (on an hour-basis) they made the trip and how transit fare was paid.

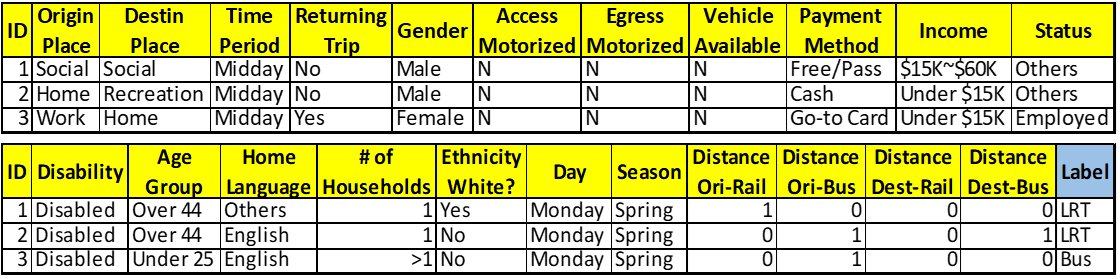
However, not all surveyed attributes can be used for predicting whether a new passenger will use LRT. First, about 25 columns are location information, namely each place’s latitude, longitude, zip code, and city. Second, some columns are duplicates as they represent the same information but coded differently, one as an integer and the other as natural language. Finally, we needed to reclassify some categorical variables as they are excessively classified that would comprise the model interpretability. Details for the variable selection will be provided in the next section.

We labeled our target variable as 1 if a passenger utilized at least one LRT service at any time within their travel path. If no LRT service was used at any point within a passenger’s travel path then the target variable is 0. This coding strategy reflects the inherent purpose of LRT services that are expected to serve people’s travel demand quickly and massively from one transfer point to another (namely, LRT stations), fed by local bus services.

A total of 22 predictors are used in models by filtering, combining, and simplifying the raw data. Many of them are self-explanatory: binary biological sex, age group, annual income, disability, etc. Some reclassification resulted in the advent of self-explanatory variables, but some are not.

First, the surveyed place types for passenger’s trip origin and destinations had 9 attribute levels. We have simplified and integrated some types that share less than one percent to get six-leveled predictors. Second, We have combined three different variables to get a binary predictor called vehicle availability: having value 0 if a respondent said nobody could drop them off AND no available vehicles are in their household AND they do not have a driver’s license. Third, we can measure four distances for each passenger, which are distances from their trip origin to the nearest LRT station, nearest bus station, and those measured from their destination. We then discretize them into binary, depending on whether they exceed 1.1 miles (the 95 percentile transit walking distance) or not.

The figure below illustrates the first three rows of the used predictors and the target label.



## Methods

The dataset was processed as described in the previous section. The processed dataset with 22 predictors was then split into a training set and test set with a ratio of 80:20. The training set has 24,393 observations while the test set has 6,098. Models were evaluated based on their test error rate, which can be calculated by 1 - test accuracy, where:



## *Model 1. Logistic Regression*

Logistic regression is one of the linear models for classification that we studied in Chapter 4 and is widely used in binary classification problems as a baseline model, due to its simple implementation. In this project, we implement a logistic regression model by using the “glm” function for a generalized linear model from the “stats” package on CRAN. To fit the logistic regression model, we specified the option “family = binomial”. As this is the simplest model implemented, we achieve mediocre results with a training error rate of 0.25 and a test error rate of 0.24. Overall, Logistic regression is straightforward to implement, quick, and easy to interpret. However, its error rate is high, especially considering categorizing Light Rail Transit users, which is our main goal.

## *Model 2. Decision Tree*

Decision trees, also known as classification trees when predicting discrete classes, use conditional statements to predict the target variable. They are extremely useful methods when ease of interpretation is important. As noted above, we predict whether a person used the light rail as part of their public transportation usage. Decision trees resemble an inverted tree. At the top is the root node with no nodes above it. Below are internal nodes which are connected to the layers above by a single branch. Each internal node has two branches extending below. At the very bottom of the tree are terminal nodes which have one incoming branch from the level above and no outgoing nodes. These terminal nodes output the final prediction.

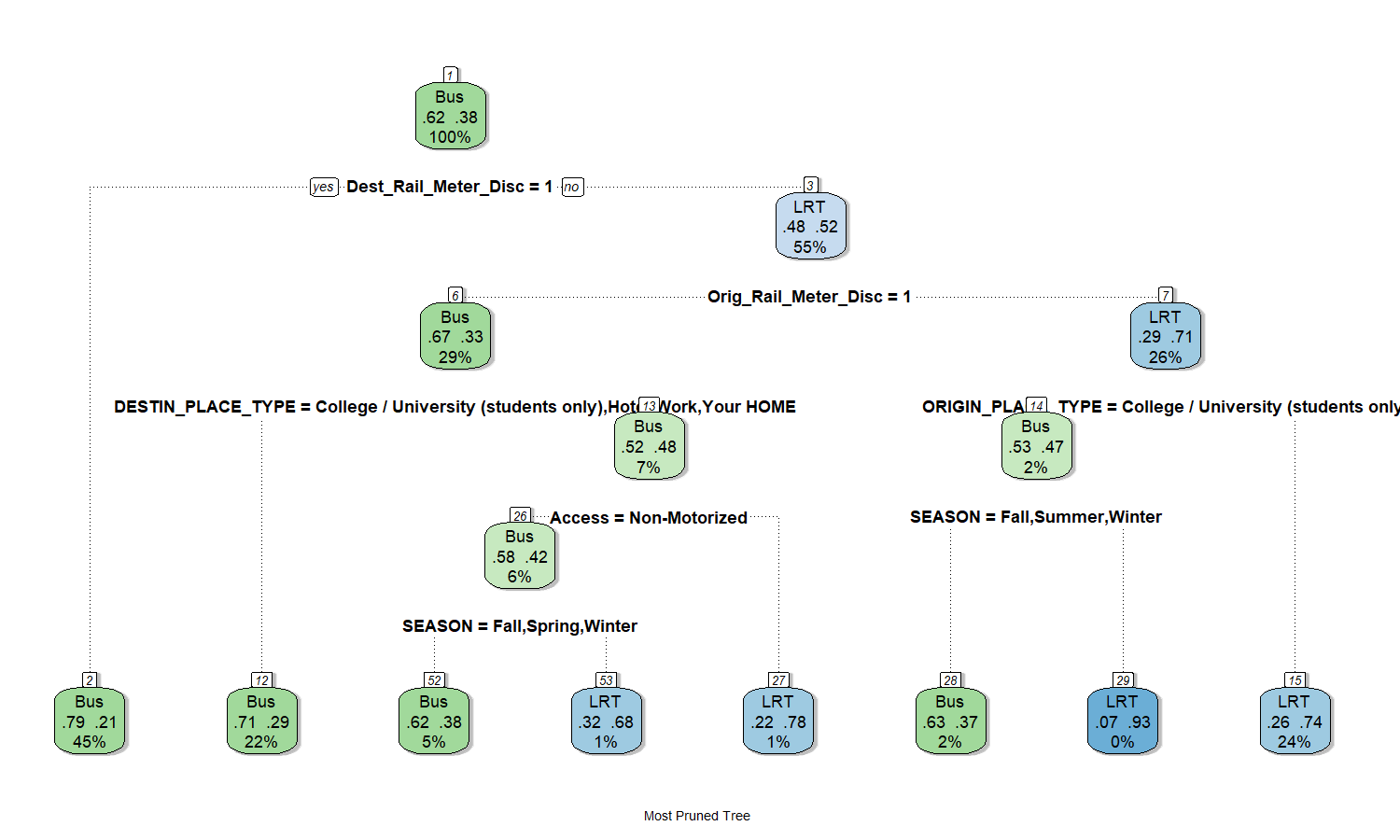
As all of the predictors are categorical, decisions are made depending on the exact value of the variable. For instance, if the root node is the discrete distance from the destination to an LRT station we will follow the left branch of the node if the distance is greater than 1.1 miles and follow the left branch if the distance is less than 1.1 miles.

Our initial classification tree was fitted using all of the available predictors. Although we realize that this is likely overfit, this model has the best positive predictive performance on LRT usage of 61% (which will be discussed later). We then prune this tree. To do this we calculated the 10-fold cross validation error on different shrinkage parameter values. We take the most parsimonious route, and choose the shrinkage parameter that is one degree larger than the shrinkage value within one standard deviation of the minimum. This tree will be labeled as the pruned tree.

The pruned tree performs well, using 9 unique variables and achieving a test error rate of 24%, which is the best of our classification tree methods (not including the ensemble random forest method). Although this pruned tree has the best overall classification tree accuracy, it does not have the best positive predictive accuracy of LRT usage. Looking at the compiled confusion matrix below for the classification trees, we see that the full tree accurately predicts 61% of the cases while the pruned tree only correctly predicts 55% of the cases. Finally, we pruned this tree once more using the similar method as above to achieve the most pruned tree.

The most pruned tree has a test error rate of 24.9%, which places it between the full tree and the pruned tree. The main advantage of this tree however is its interpretability. By only using 6 unique variables this tree is by far the easiest to understand. The 6 variables used are: distance from origin to closest LRT station, distance from destination to closest LRT station, origin place type, destination place type, vehicle access and season. Below table summarizes confusion matrices of the three trees, followed by the visualization of the most pruned tree.

| Confusion Matrices | Full Tree | | Pruned Tree | | Most Pruned Tree | |
| --- | --- | --- | --- | --- | --- | --- |
| Pred Bus | Pred LRT | Pred Bus | Pred LRT | Pred Bus | Pred LRT |
| True Bus | 3,190 | 699 | 3,430 | 459 | 3,462 | 427 |
| True LRT | 859 | 1,350 | 1,005 | 1,204 | 1,094 | 1,115 |



Although it performs slightly worse, this tree may be used to illustrate the most important variables. We see that the root node measures whether the destination is within 1.1 miles to any LRT station. If it is not, the left branch is followed to node 2 where bus is predicted. If the destination is within 1.1 miles we take the right branch to node 3, which measures whether the origin is within 1.1 miles of a LRT station. If the origin is also within 1.1 miles of a LRT station we move right to node 7. After distance, this subtree finds that the origin place type and the season are important. If the origin place is not a college or university then this tree predicts LRT. If the origin is a college or university then we see that the weather is used. If it is currently spring then LRT is predicted, otherwise bus is predicted. The subtree starting with node 6 as its root is very similar except with one noted exception. The variable access is included. If access is not non-motorized (which is the same as if access is motorized) then LRT is predicted. If access is non-motorized, then the prediction depends on the season. As we can see, distance to and from the LRT station is the most important. Beyond that, one possible interpretation of this model is that when the weather is nicer (spring and summer) people are more likely to travel further distances to a LRT station as they don’t mind being outside. Additionally, with motorized access, people are more willing to travel to far-away yet convenient transit services.

## *Model 3. Random Forest*

Random Forest models are ensemble methods that combine the prediction of multiple decision trees to create a more accurate prediction. Random Forests have been shown to be an extremely powerful tool. In this project, we used the randomForest package provided by Breiman et al. on CRAN. We first set the ntree = 200, which indicates the number of trees to build. Although adding more trees to the random forest improves accuracy, it also increases the overall training time of the model. Moreover, we added additional parameters when we initialized the Random Forest model to further enhance the performance. For instance, we set mtry=2, nodesize=1, importance=TRUE, keep.inbag=TRUE. The result of this model on training data is a 19% error rate, and a 23.6% error rate on the test set. Although this is a lower error rate than the simple decision trees discussed above, the random forest model performed worse on accurately predicting LRT users. This is an easy method to implement, however it is also computationally more intensive. As we see a disparity between the test and training error rates, we can also conclude that this model is slightly overfit.

## *Model 4. Neural Network*

Neural networks are a collection of hidden layers and nodes containing a non-linear activation function that seeks to resemble how neurons work in the brain. As such, they are often seen as “black-box” models as it is not exactly clear how they arrive at their predictions. Although they may not be interpretable in the same sense that other statistical models are, they often compensate for this with higher predictive accuracy. Although we hypothesized that the neural network would have the smallest test error, this was not the case. The current general (albeit possibly false) consensus is that neural networks can be applied to every problem and outperform all other models, however our results suggest otherwise.

We fit a variety of neural networks ranging from 2 to 5 hidden layers, with each layer having between 32 and 256 nodes. We also used both sigmoid and rectified linear unit activation functions as well as dropout layers. The results were somewhat invariant to tuning and hyperparameters, with all neural nets having around 26% test error. One possible explanation for this is that tree methods may be better suited to this question. When choosing a method of public transportation, the thought process may be more similar to how a tree solves this. For instance, when considering what kind of public transportation to use people may discrete choices based on distance and the weather outside. This decision making process is adequately captured in tree methods and logistic regression.

## Discussion and Conclusions

In conclusion, we tried to predict whether a public transit user took the light rail as part of their public transportation trip, without using any path attributes. Often, other disciplines focus on travel purpose, origin/destination type, and path attributes, ignoring socio-demographic attributes. Although socio-demographic attributes were not found to be particularly important in any of our models, this in itself is an important finding as this corroborates the focus on other variables. Moreover, we discovered that more complicated methods, such as random forest and neural networks, did not outperform the simpler methods, logistic regression and decision trees, in this particular application.

The table below summarizes each model’s results. We can utilize test errors for evaluating model performance. However, as mentioned in the introduction, predicting LRT labels is particularly important since investing in rail transport requires a tremendous budget. Therefore, we include another important metric that can help quantify model performance, the LRT prediction accuracy rate. This metric represents the percent of correctly predicted LRT labels, and can be thought of as the model’s sensitivity.

| Model | Error Rate | | LRT Prediction Accuracy Rate |
| --- | --- | --- | --- |
| Train set | Test set |
| Logistic Regression | 0.250 | 0.249 | 54% |
| Decision Tree | 0.183 | 0.256 | 61% |
| Decision Tree (pruned) | 0.242 | 0.240 | 55% |
| Decision Tree (fully pruned) | 0.251 | 0.249 | 51% |
| Random Forest | 0.190 | 0.236 | 53% |
| Neural Network | 0.260 | 0.261 | 53% |

Although the lowest test error is achieved by using random forests, the gain is marginal and may not be found in repeated testing with other similar data sets. Furthermore, the fitted neural network has the worst performance out of all tested models. As both logistic regression and decision trees perform fairly well and offer an intuitive interpretation of the effects of the variables, we conclude that in this instance the added predictive power of more complex methods may not be worth the adverse effect they have on interpretability. Additionally, when LRT prediction is taken into account, a decision tree may be favored over a random forest model.

Overall, without path attributes, which are widely used in conventional methods, machine learning models still have potential. If path attributes were utilized for constructing advanced machine learning models, better accuracy and lower test error may be achieved. By doing that, we can incorporate information such as: how many minutes a passenger needs to access each LRT/non-LRT path option with a realistic access path, how long they have to wait, and how many transfers they would have to make. Important path attributes can be identified by taking passenger origin and destination location inputs (plus transit/passenger schedules, if applicable) or using the existing transit demand assignment algorithms.

In the future, potential extensions of this work include the application of other methods, such as SGDClassifier and LDA. Additionally, factors analysis should be considered as some variables might have an underlying relationship. Moreover, better feature engineering and variable selection can also enhance model accuracy. Finally, some more work on fine tuning models can be conducted, especially on the neural network which has a plethora of tuning parameter options.

## References

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Ben-Akiva, M., Mcfadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Munizaga, M. (2002). Hybrid Choice Models: Progress and Challenges. *Marketing Letters*, 13(3), 163-175.

Khani, A., Hickman, M., & Noh, H. (2014). Trip-Based Path Algorithms Using the Transit Network Hierarchy. *Networks and Spatial Economics*, 15(3), 635-653.

Koushik, A., Manoj, M., & Nezamuddin, N. (2020). Machine learning applications in activity-travel behaviour research: A review. *Transport Reviews*, 40(3), 288-311.

Metro Transit (2017). SURVEY INFORMATION FOR TRANSIT RIDERS. Website: https://metrocouncil.org/Transportation/Performance/Travel-Behavior-Inventory/Transit-On-Board-Survey/Survey-Information-for-Transit-Riders.aspx

Train, K. (2009). Discrete choice methods with simulation (2nd ed.). Cambridge University Press.