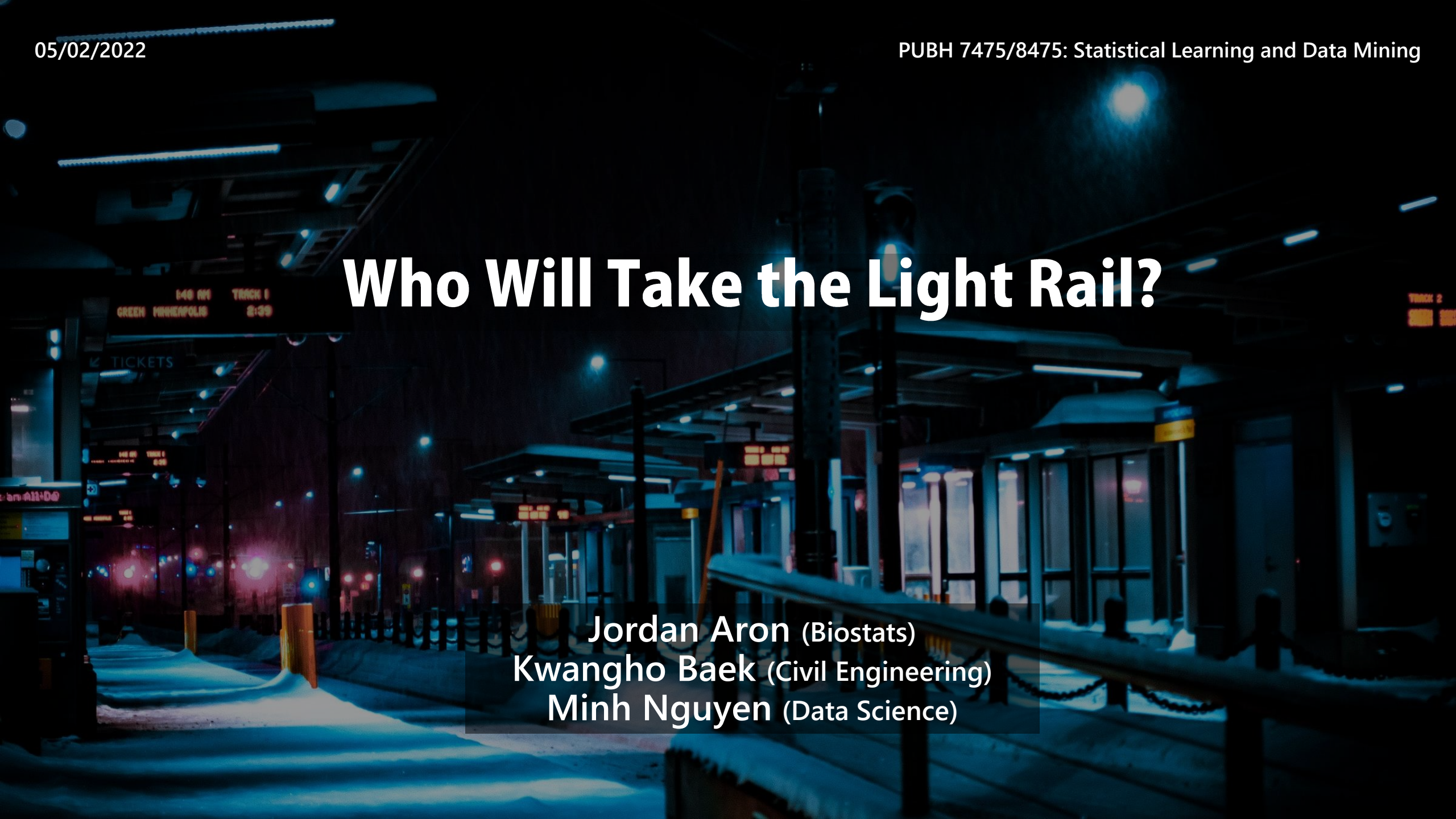


Who Will Take the Light Rail?

Jordan Aron (Biostats)
Kwangho Baek (Civil Engineering)
Minh Nguyen (Data Science)

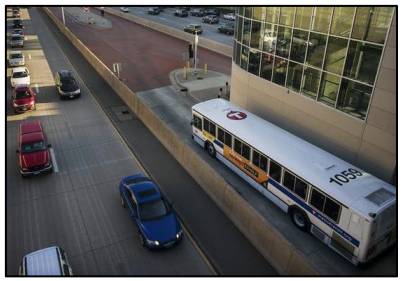
Structure

1. Introduction
2. Data Overview
3. Methodology
4. Conclusion and Discussion

Who Will Take the Light Rail?

Introduction

Travel Mode Choice Theory



(Metro Transit, Star Tribune)

Which Travel Mode Would you Take?

A function of

(1) socio-demographic attributes

(2) travel purpose (work/school/…)

(3) origin/destination location

(4) path attributes (cost/transit supply)

The most important thing is how a traveler perceives each transport mode for a given circumstances-> latent

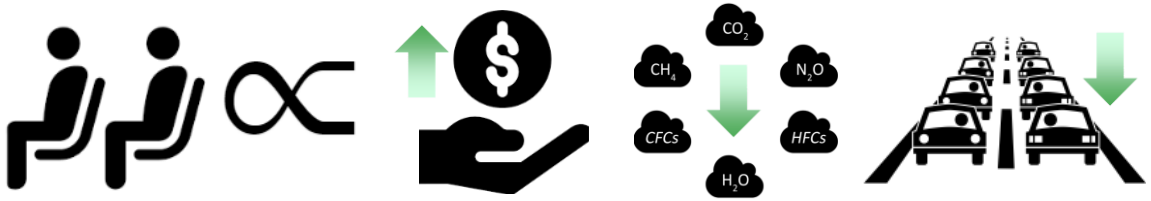
LRT Ridership Matters: Machine Learning Approach?

$715.3 M
Blue Line

$957.0 M
Green Line

$$\frac{Benefit}{Cost} > 1?$$

Inaccurate ridership prediction: tremendous social cost



Develop a mode prediction model, unlike the conventional one, which can handle large number of demographic variables that predicts the number and attributes of potential LRT riders.

Who Will Take the Light Rail?

Data

Data Source

2016 Transit On-Board Survey (OBS 2016) from Metro Transit, MN

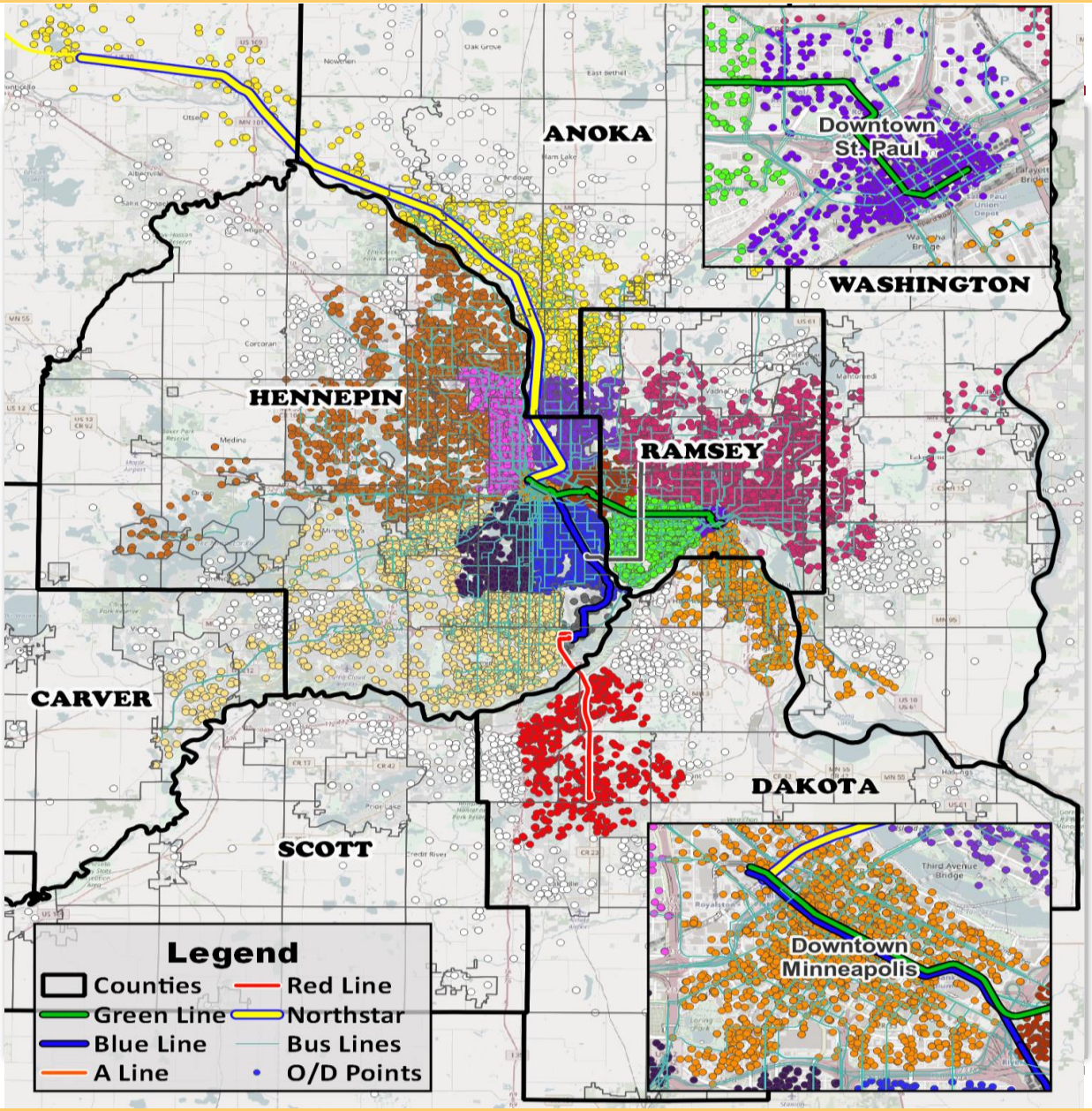
30,491
Observations (Passenger Count)

174
Demographic & Travel-related Attributes

- Travel Start Time
- Gender
- Age Group
- Access Mode
- Egress Mode
- Payment Method
- Transfers
- Income
- Employment or studentship
- Disability
- Ethnicity
- Home Language
- Fare Subsidized?
- Car Availability
- OD Coordinates
- OD Place Type
- …

Weights
Available for adjusting sampling error

OD: Origin and Destination



Who Will Take the Light Rail?

Binary Target Class Label

The Label

N=30,491

:LRT (Light Rail Transit) :Bus → :Transfer

isLRT=1

Direct

11,440
37.5%

If a passenger utilized at least one LRT route at any time during their travel

isLRT=0

Direct

19,051
62.5%

If a passenger only utilized bus route(s) during their travel

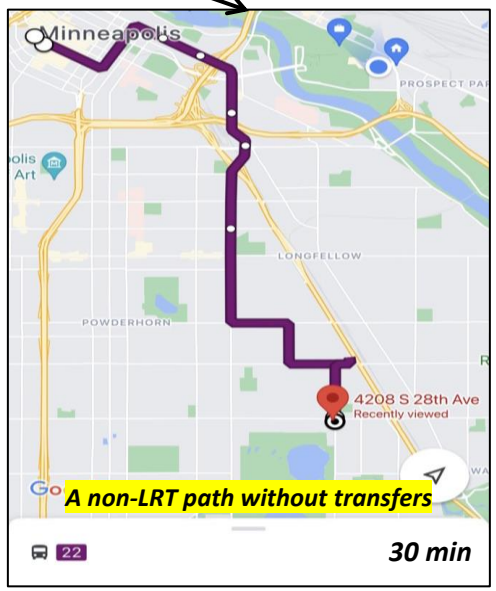
**The number and order of transfers of each path do not affect the classification*

Example

La Salle Plaza (DT MPLS) → Lake Hiawatha

LRT Path

Bus-only Path



isLRT=1

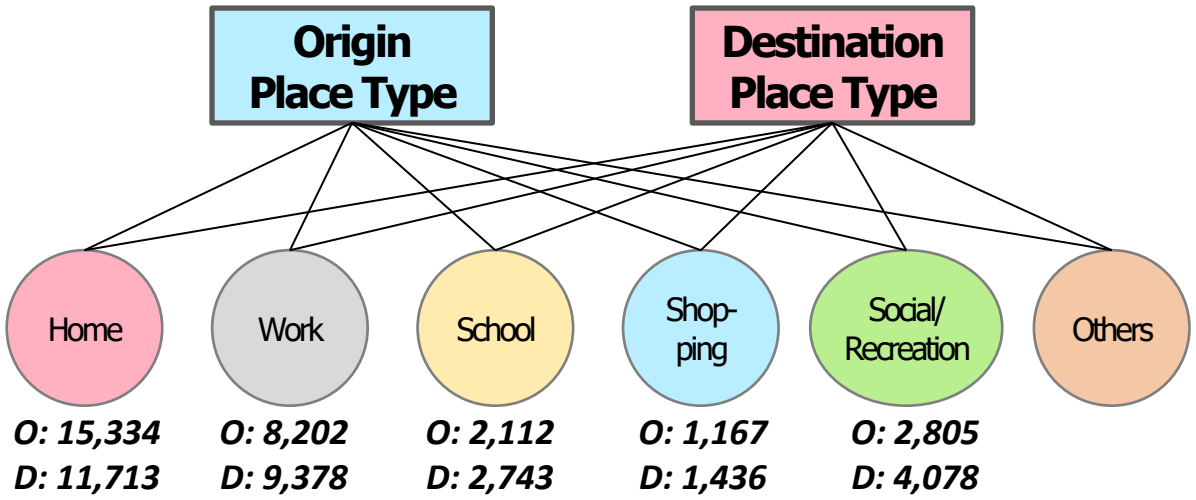
isLRT=0

Who Will Take the Light Rail?

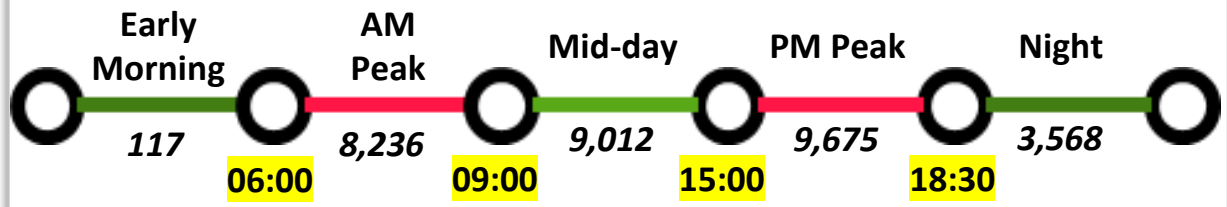
Predictors

N=30,491

Place Type (6 levels)

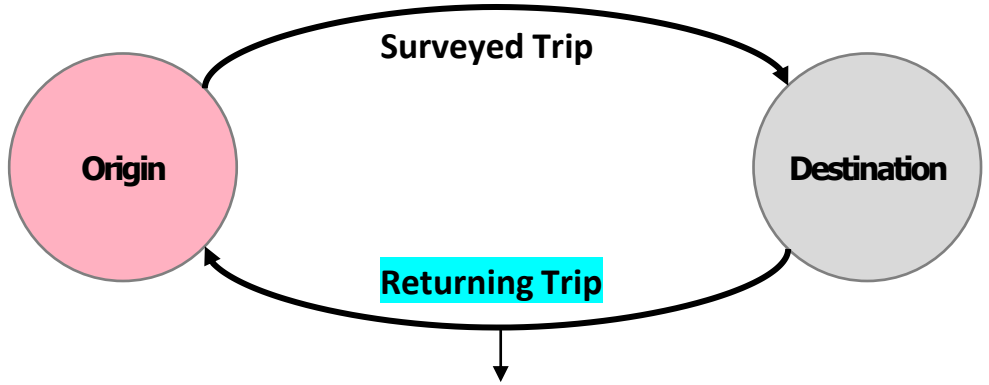


Time Period (5 levels)



N=30,491

Opposite trip planned? (binary)



If a respondent said this trip was planned: 1 (19,879), otherwise: 0 (10,612)

Vehicle Availability (binary)

Nobody could drop me off **x** No available vehicle in my household **x** I don't have a driver's license

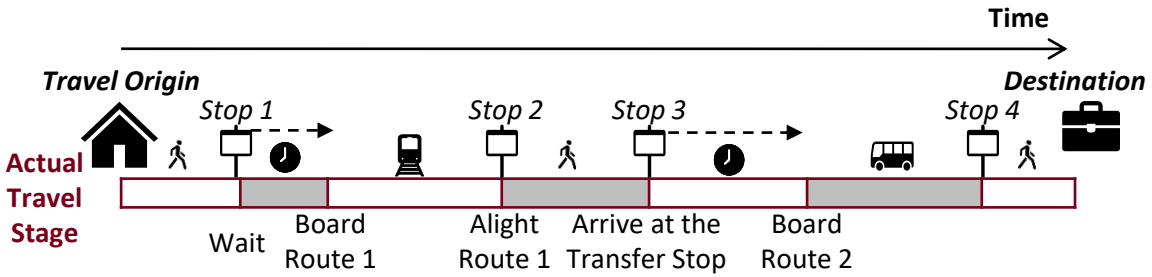
If the joint condition is true: 0 (22,014), otherwise: 1 (13,747)

Who Will Take the Light Rail?

Predictors

Motorized Access/Egress Mode and Transfer (binary)

N=30,491



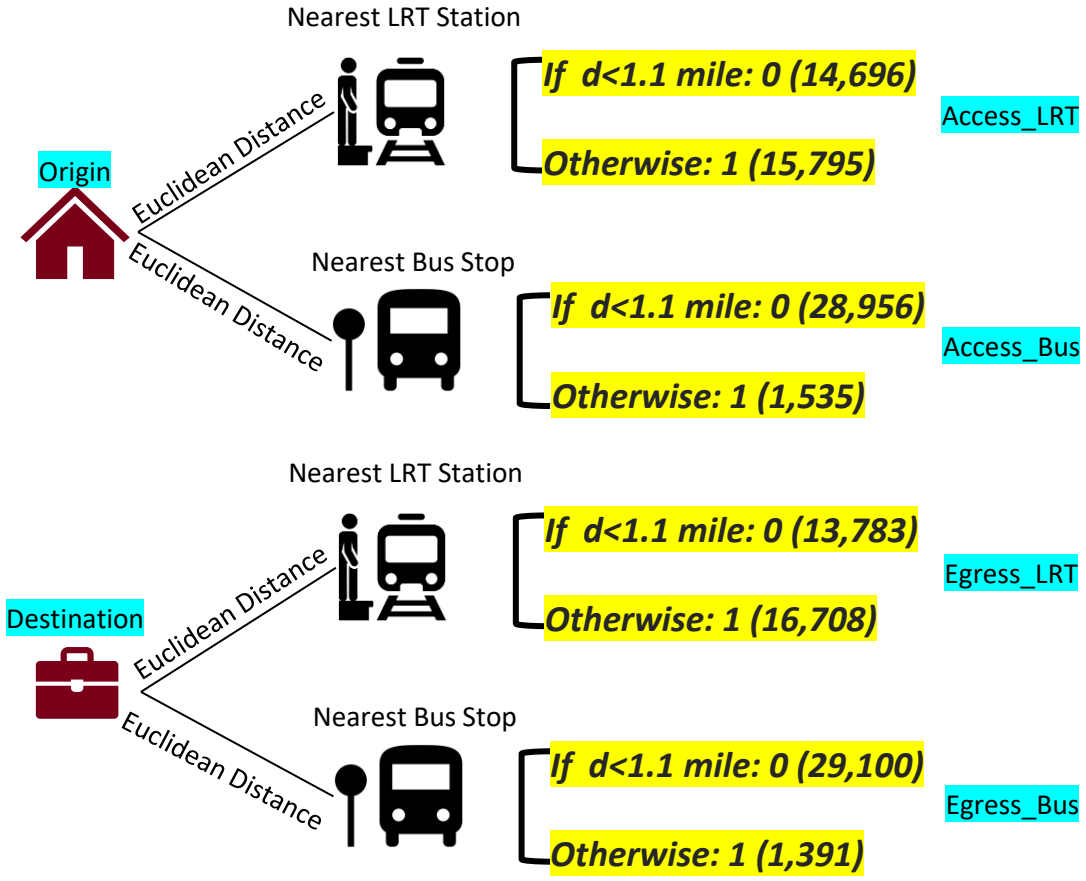
Terminology: *Access* | *Access Waiting* | *In-vehicle Time (IVT) 1* | *Transfer Walking* | *Transfer Waiting* | *IVT 2* | *Egress*

If access/egress modes are "on foot", "bike", "wheelchair" :0, otherwise 1

Access Motorized	0: 30,026 1: 5,735
Egress Motorized	0: 33,671 1: 2,090

Distances (four binary variables)

N=30,491



1.1 mile: 95% cutoff value of transit traveler's walking access distance

Who Will Take the Light Rail?

Predictors

Day/Season

N=30,491

Monday	*1,150*
Tuesday	*8,713*
Wednesday	*9,518*
Thursday	*10,721*
Friday	*389*

Jan~Mar	*Apr~Jun*	*Jul~Sep*	*Oct~Dec*
			
3,928	*9,229*	*6,663*	*10,676*

Others (Self-Explanatory)

N=30,491

Category	Level	Count
Gender	Female	17,088
	Male	18,673
Age Group	Age over 44	10,889
	Age 25-44	15,588
	Age under 25	9,284
Ethnicity	White	22,079
	Non-white	13,682
Annual Income (in 2016 USD)	Over $60K	16,814
	$15K~$60K	14,381
	Under $15K	4,566
Student/ Employment Status	Student	10,364
	Employed	22,402
	None of the above	2,995
Disability	Disabled	3,837
	Not Disabled	31,924

First 10 Rows of the Data

	ORIGIN_PLACE_TYPE	DESTIN_PLACE_TYPE	TIME_PERIOD	TRIP_IN_OPPOSITE_DIR
1	Social Visit / Community / Religious / Personal	Social Visit / Community / Religious / Personal	Midday	No
2	Your HOME	Recreation / Sightseeing / Restaurant	Midday	No
3	Work	Your HOME	Midday	Yes
4	Your HOME	Recreation / Sightseeing / Restaurant	PM Peak	Yes
5	Doctor / Clinic / Hospital (non-work)	Social Visit / Community / Religious / Personal	PM Peak	No
6	Work	Shopping	Midday	Yes
7	Your HOME	Work	Midday	Yes
8	Work	Your HOME	PM Peak	No
9	Social Visit / Community / Religious / Personal	Your HOME	PM Peak	No
10	Doctor / Clinic / Hospital (non-work)	Social Visit / Community / Religious / Personal	PM Peak	Yes

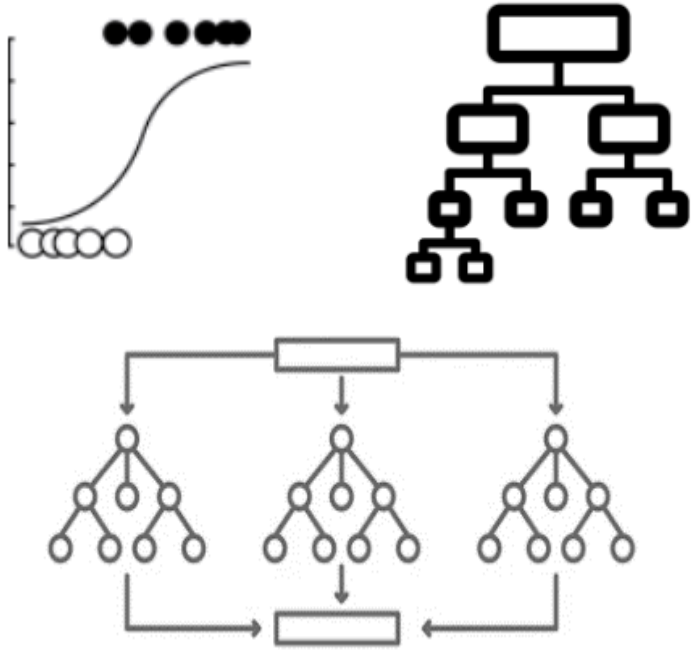
	GENDER	Access	Egress	Vehicle	pay	income	status	disable	ages
1	Male	Non-Motorized	Non-Motorized	Not Available	Free Ride or Pass Used	$15K~$60K	Others	Disabled	Over 44
2	Male	Non-Motorized	Non-Motorized	Not Available	Cash	Under $15K	Others	Disabled	Over 44
3	Female	Non-Motorized	Non-Motorized	Not Available	Go-to Card or Mobile	Under $15K	Employed	Disabled	Under 25
4	Male	Motorized	Non-Motorized	Available	Go-to Card or Mobile	$15K~$60K	Employed	Disabled	Over 44
5	Female	Non-Motorized	Non-Motorized	Available	Free Ride or Pass Used	$15K~$60K	Student	Disabled	Age 25-44
6	Female	Non-Motorized	Non-Motorized	Not Available	Cash	Under $15K	Student	Disabled	Age 25-44
7	Female	Non-Motorized	Non-Motorized	Available	Free Ride or Pass Used	$15K~$60K	Others	Disabled	Under 25
8	Female	Non-Motorized	Non-Motorized	Not Available	Free Ride or Pass Used	Under $15K	Employed	Disabled	Age 25-44
9	Male	Motorized	Non-Motorized	Available	Go-to Card or Mobile	Under $15K	Others	Disabled	Age 25-44
10	Female	Non-Motorized	Non-Motorized	Not Available	Go-to Card or Mobile	Under $15K	Others	Disabled	Over 44

	eng	HH	iswhite	Label	DAY	SEASON	Orig_Rail_Meter_Disc	Dest_Rail_Meter_Disc	Orig_Bus_Meter_Disc	Dest_Bus_Meter_Disc
1	Others	1	Yes	LRT	Monday	Spring	1	0	0	0
2	English	1	No	LRT	Monday	Spring	0	1	0	1
3	English	More than 2	No	Bus	Monday	Spring	0	1	0	0
4	English	1	Yes	LRT	Monday	Spring	1	0	1	0
5	English	1	Yes	LRT	Monday	Spring	0	0	0	0
6	English	2	Yes	LRT	Tuesday	Spring	0	0	0	0
7	Others	More than 2	No	LRT	Tuesday	Spring	0	0	0	0
8	English	1	Yes	LRT	Tuesday	Spring	1	0	0	0
9	English	1	No	LRT	Tuesday	Spring	0	0	0	0
10	English	1	Yes	LRT	Tuesday	Spring	0	1	0	0

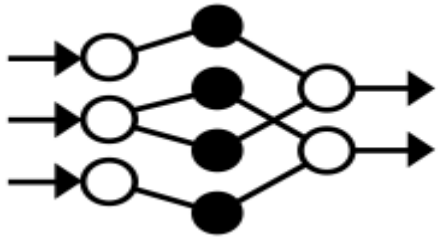
P=22 * N=30,491

Methods Overview

1. Implemented models:
    - Logistic regression
    - Decision trees
    - Random forest
    - Neural network

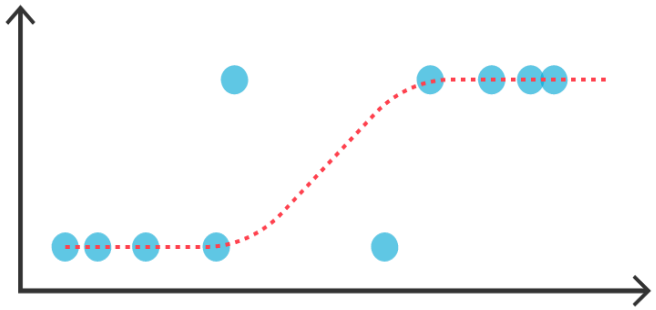


2. Training and Testing Process
    - Original dataset: 30,491 observations
    - 80:20 training testing data split (24,393 vs 6,098)



Logistic Regression

Logistic Regression Model



1. Used all (22) available predictors
2. Test error rate: 0.249
3. Simple to implement
4. Not doing so well in categorizing LRT users
5. Baseline model

Results

```
                Accuracy : 0.7506          Sensitivity : 0.5378
                  95% CI : (0.7395, 0.7614 Specificity : 0.8714
     No Information Rate : 0.6378          Pos Pred Value : 0.7038
     P-Value [Acc > NIR] : < 2.2e-16       Neg Pred Value : 0.7685
                                                Prevalence : 0.3622
                   Kappa : 0.4312          Detection Rate : 0.1948
                                      Detection Prevalence : 0.2768
  Mcnemar's Test P-Value : < 2.2e-16      Balanced Accuracy : 0.7046
```

6,098 Tests	Pred Bus	Pred LRT
True Bus	**3,389 (87%)**	500 (13%)
True LRT	1,021 (46%)	**1,188 (54%)**

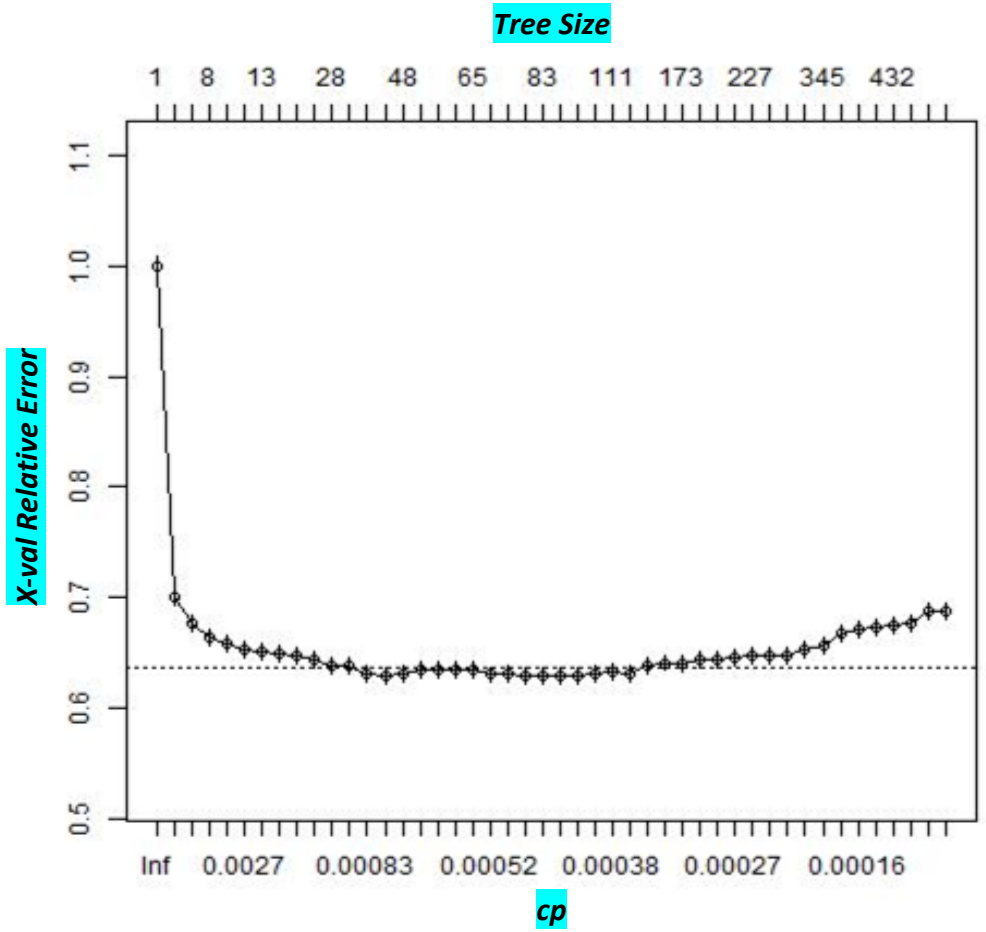
Decision Tree (Baseline Tree)

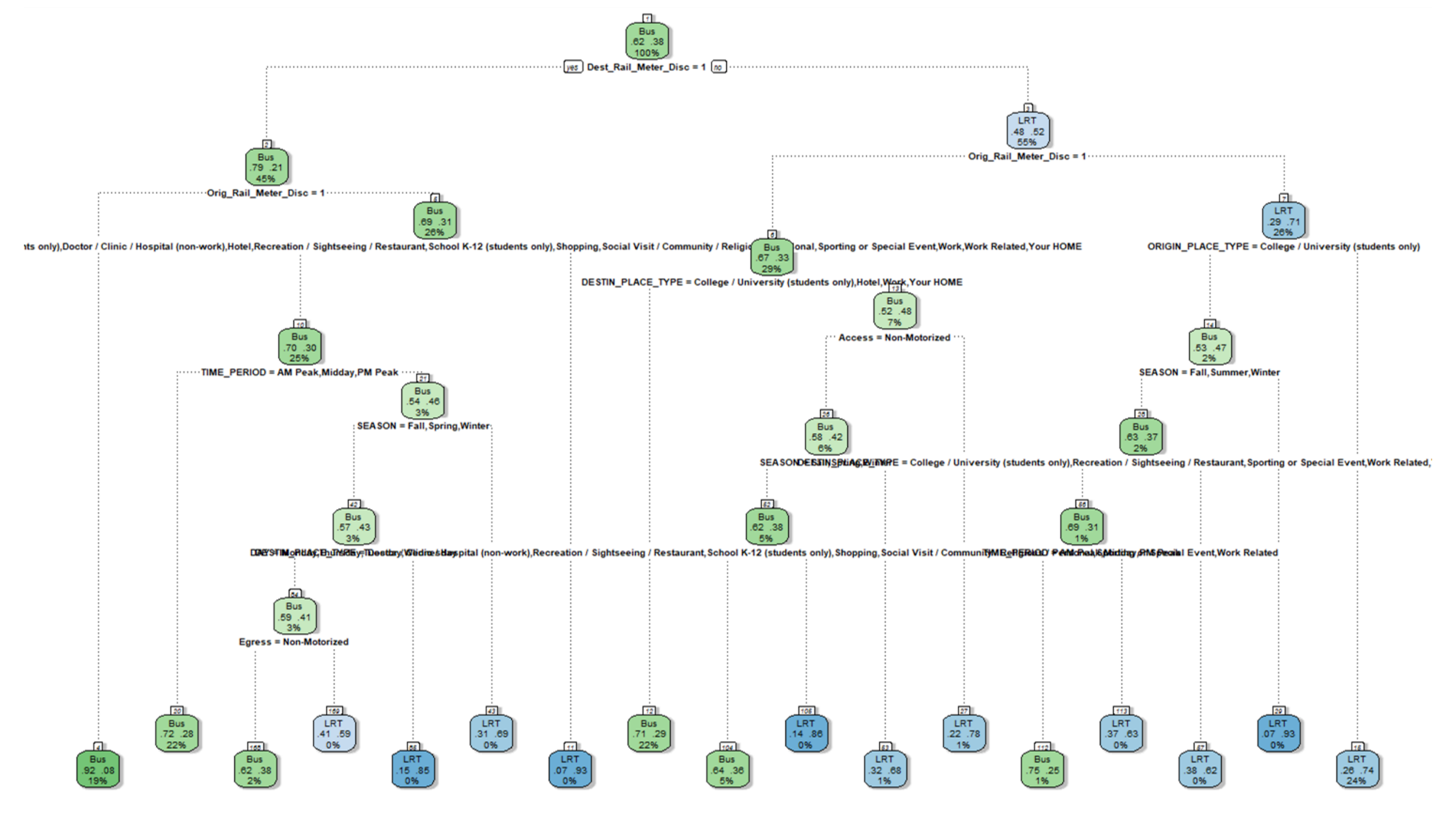
Baseline Decision Tree

1. Baseline tree was grown using all available predictors
2. 10 fold CV was implemented for tuning parameters
3. **Test error rate: 0.255**

6,098 Tests	Pred Bus	Pred LRT
True Bus	**3,190 (82%)**	699 (18%)
True LRT	850 (39%)	**1,350 (61%)**

Relative Error




Bus .62 .38 100%
yes Dest_Rail_Meter_Disc = 1 no
Bus .79 .21 45%
Orig_Rail_Meter_Disc = 1
Bus .69 .31 26%
LRT .48 .52 55%
Orig_Rail_Meter_Disc = 1
LRT .29 .71 26%
ORIGIN_PLACE_TYPE = College / University (students only)
Bus .67 .33 29%
DESTIN_PLACE_TYPE = College / University (students only),Hotel,Work,Your HOME
Bus .52 .48 7%
Access = Non-Motorized
Bus .70 .30 25%
TIME_PERIOD = AM Peak,Midday,PM Peak
Bus .54 .46 3%
SEASON = Fall,Spring,Winter
Bus .58 .42 6%
Bus .63 .37 2%
SEASON = Fall,Summer,Winter
Bus .57 .43 3%
Bus .59 .41 3%
Egress = Non-Motorized
Bus .92 .08 19%
Bus .72 .28 22%
Bus .62 .38 2%
LRT .41 .59 0%
LRT .15 .85 0%
LRT .31 .69 0%
LRT .07 .93 0%
Bus .71 .29 22%
Bus .62 .38 5%
Bus .64 .36 5%
LRT .14 .86 0%
LRT .32 .68 1%
Bus .69 .31 1%
LRT .22 .78 1%
Bus .75 .25 1%
LRT .37 .63 0%
LRT .38 .62 0%
LRT .07 .93 0%
LRT .26 .74 24%

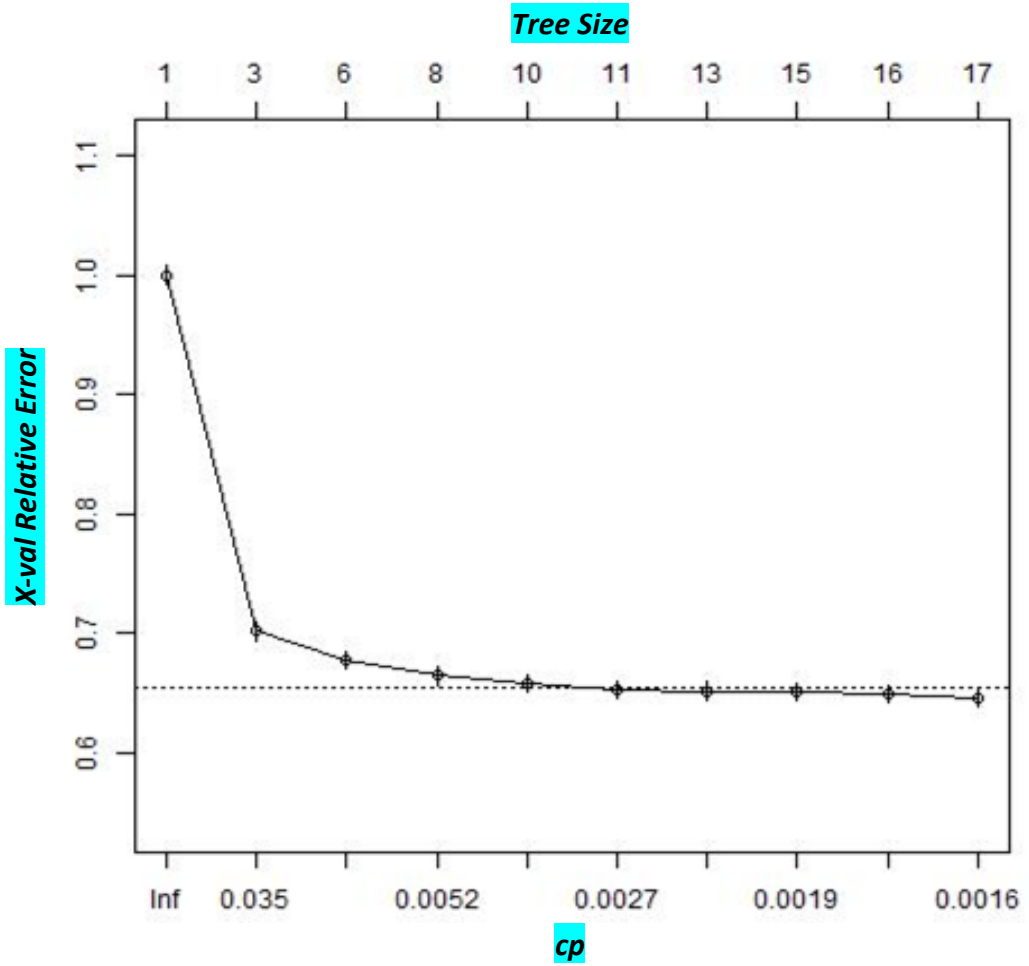
Pruned Tree

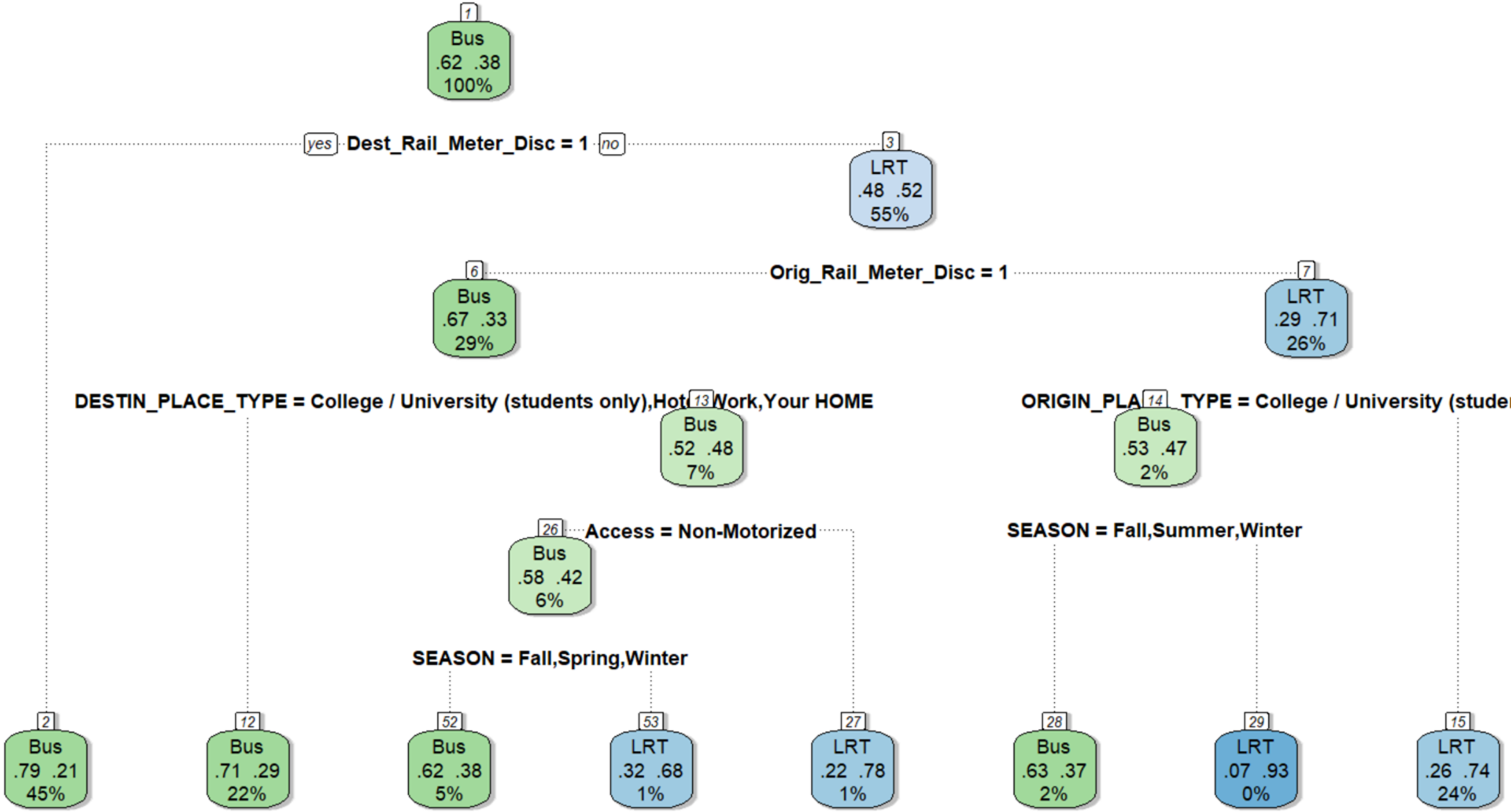
Pruning Result

1. Pruning resulted in 9 predictors to be utilized
2. 10 fold CV was implemented for tuning parameters
3. **Test error rate: 0.240**

6,098 Tests	Pred Bus	Pred LRT
True Bus	**3,430 (88%)**	459 (12%)
True LRT	1,005 (45%)	**1,204 (55%)**

Relative Error

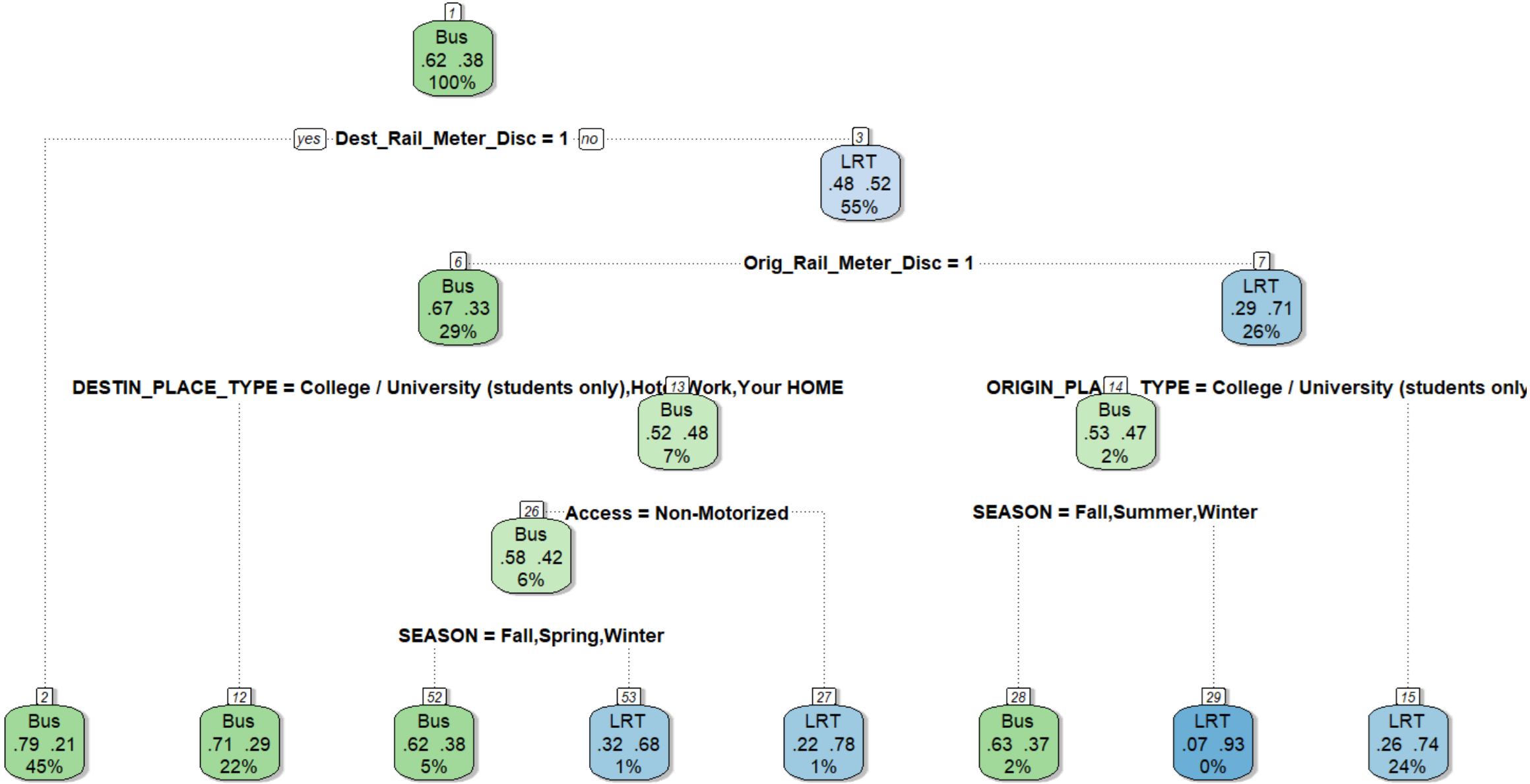



1
Bus
.62 .38
100%
yes
Dest_Rail_Meter_Disc = 1
no
3
LRT
.48 .52
55%
6
Bus
.67 .33
29%
Orig_Rail_Meter_Disc = 1
7
LRT
.29 .71
26%
DESTIN_PLACE_TYPE = College / University (students only),Hot
Work,Your HOME
13
Bus
.52 .48
7%
ORIGIN_PLA
14
TYPE = College / University (students only
Bus
.53 .47
2%
26
Bus
.58 .42
6%
Access = Non-Motorized
SEASON = Fall,Summer,Winter
SEASON = Fall,Spring,Winter
2
Bus
.79 .21
45%
12
Bus
.71 .29
22%
52
Bus
.62 .38
5%
53
LRT
.32 .68
1%
27
LRT
.22 .78
1%
28
Bus
.63 .37
2%
29
LRT
.07 .93
0%
15
LRT
.26 .74
24%

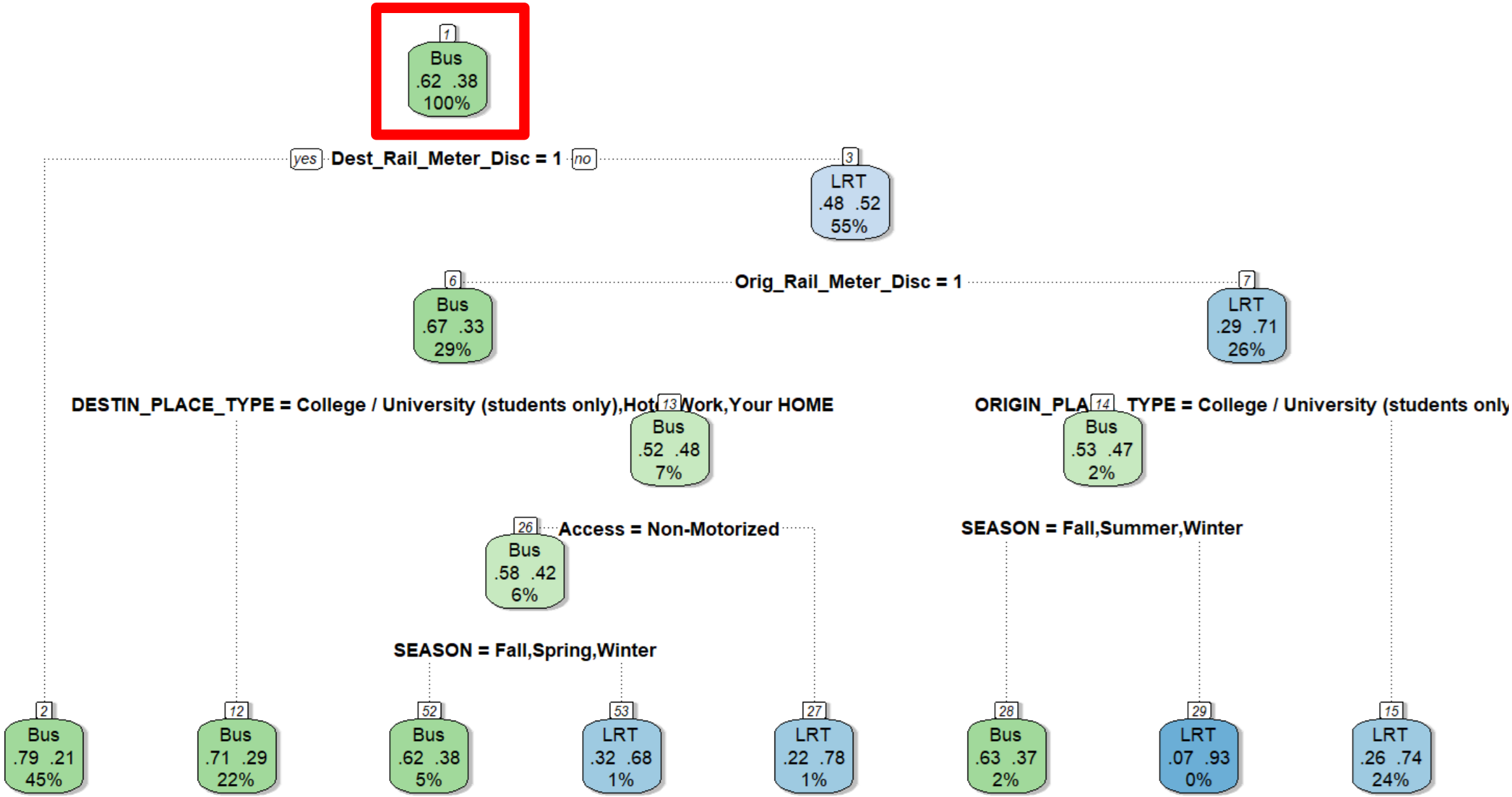
Most Pruned Tree

1. Pruning resulted in 6 predictors to be utilized
   - Discretized Distance from Origin to nearest LRT station
   - Discretized Distance from Destination to nearest LRT station
   - Origin Place Type
   - Destination Place Type
   - Access Mode (Motorized or not)
   - Season
2. Most straightforward interpretation
3. Test error rate: 0.249

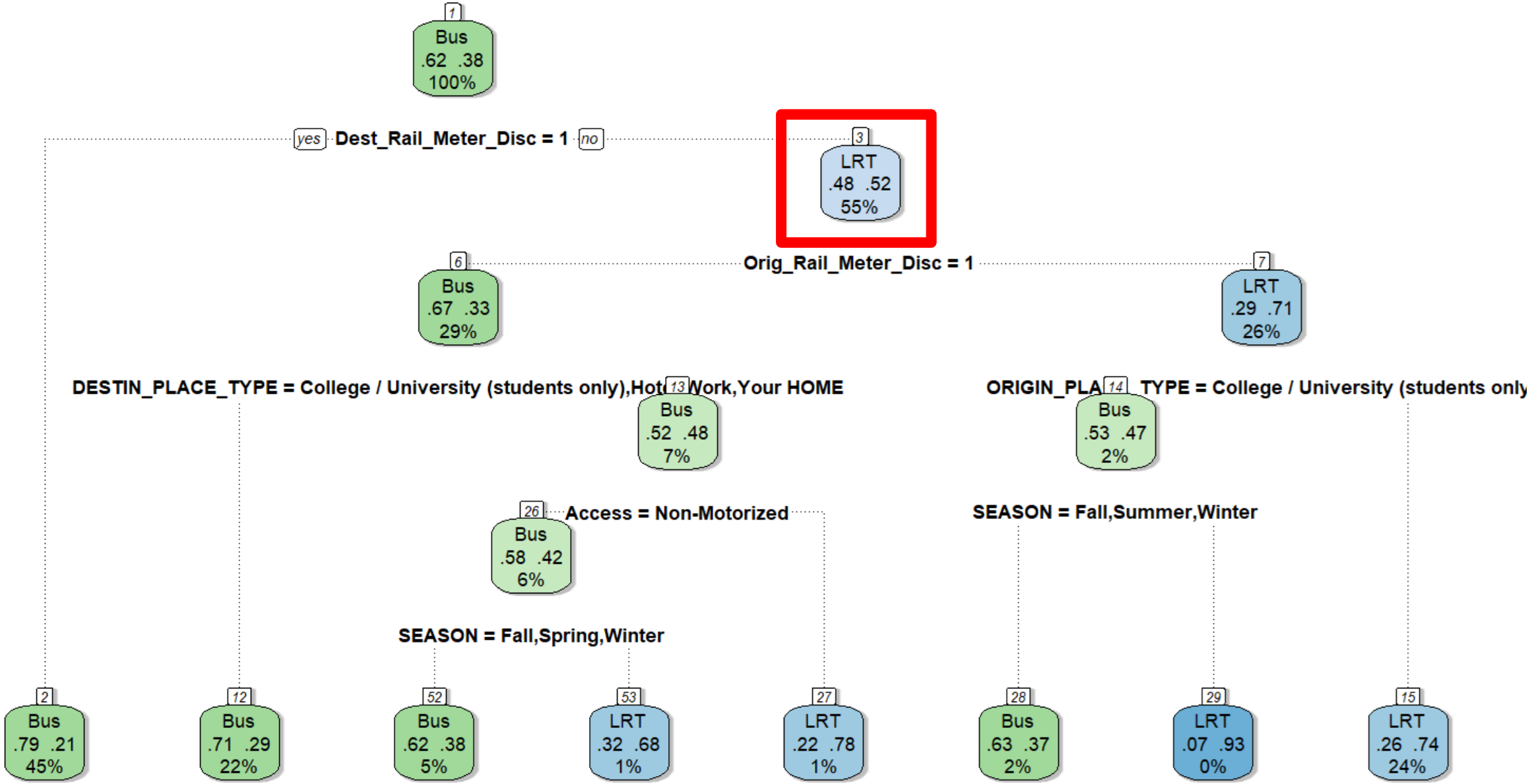
6,098 Tests	Pred Bus	Pred LRT
True Bus	**3,462 (89%)**	427 (11%)
True LRT	1,094 (50%)	**1,115 (50%)**



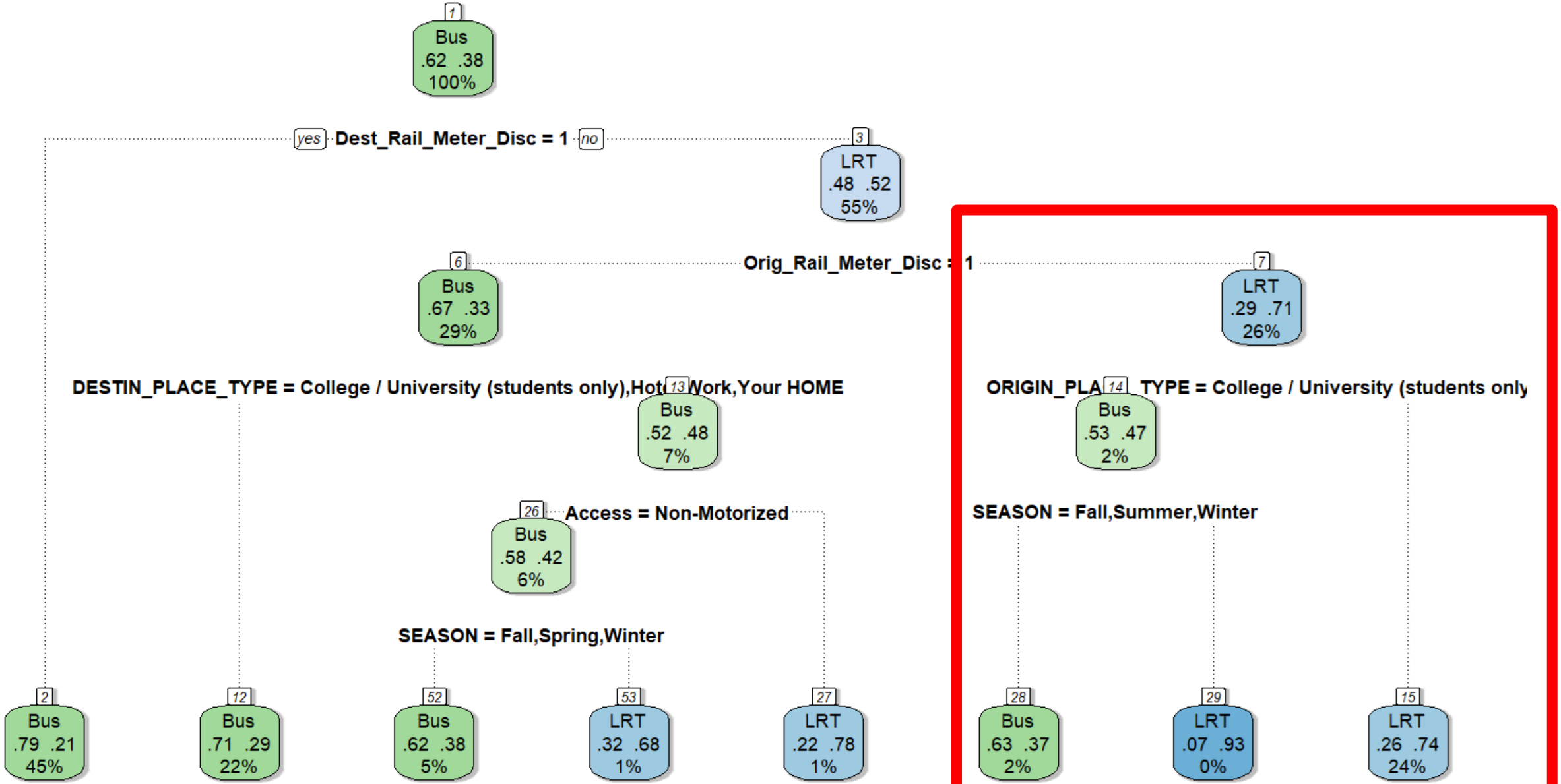
Most Pruned Tree



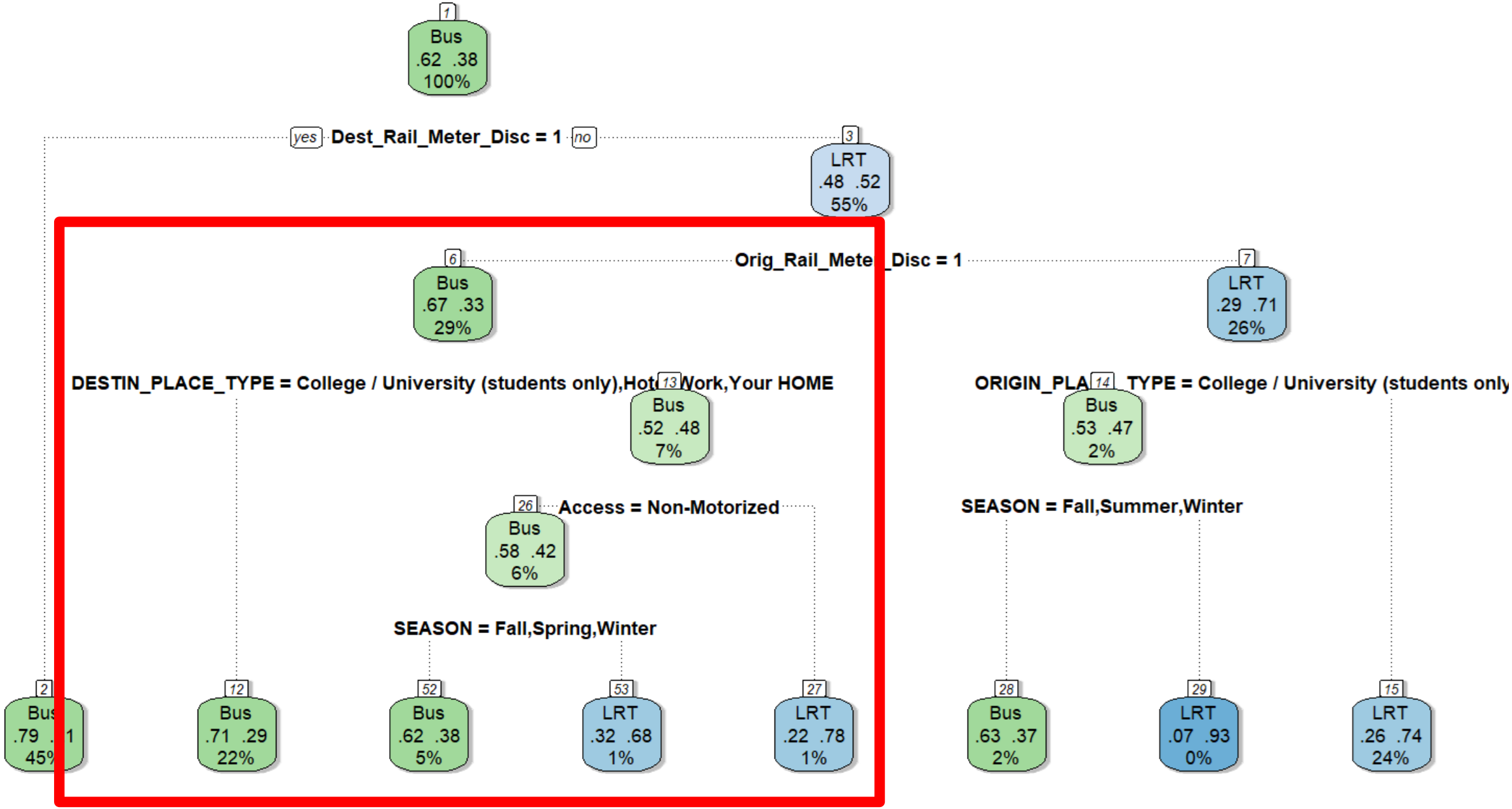
Most Pruned Tree



Most Pruned Tree



Most Pruned Tree



Most Pruned Tree

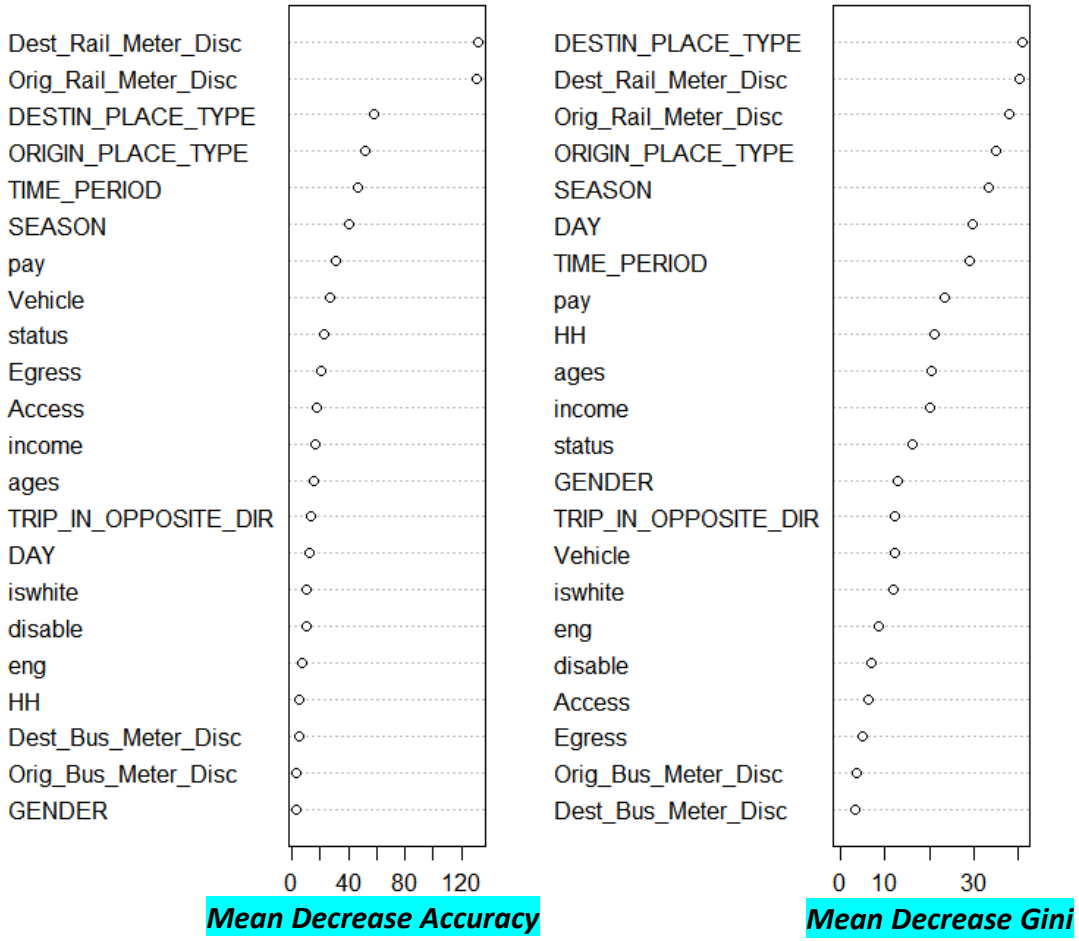
Random Forest

Model Fitting

1. All predictors were used
2. **Test error rate: 0.236**
3. Simple to implement, but higher time complexity
4. Slightly overfitting, but acceptable
5. Aligning features importance with DT
6. Does not doing so well with categorizing LRT

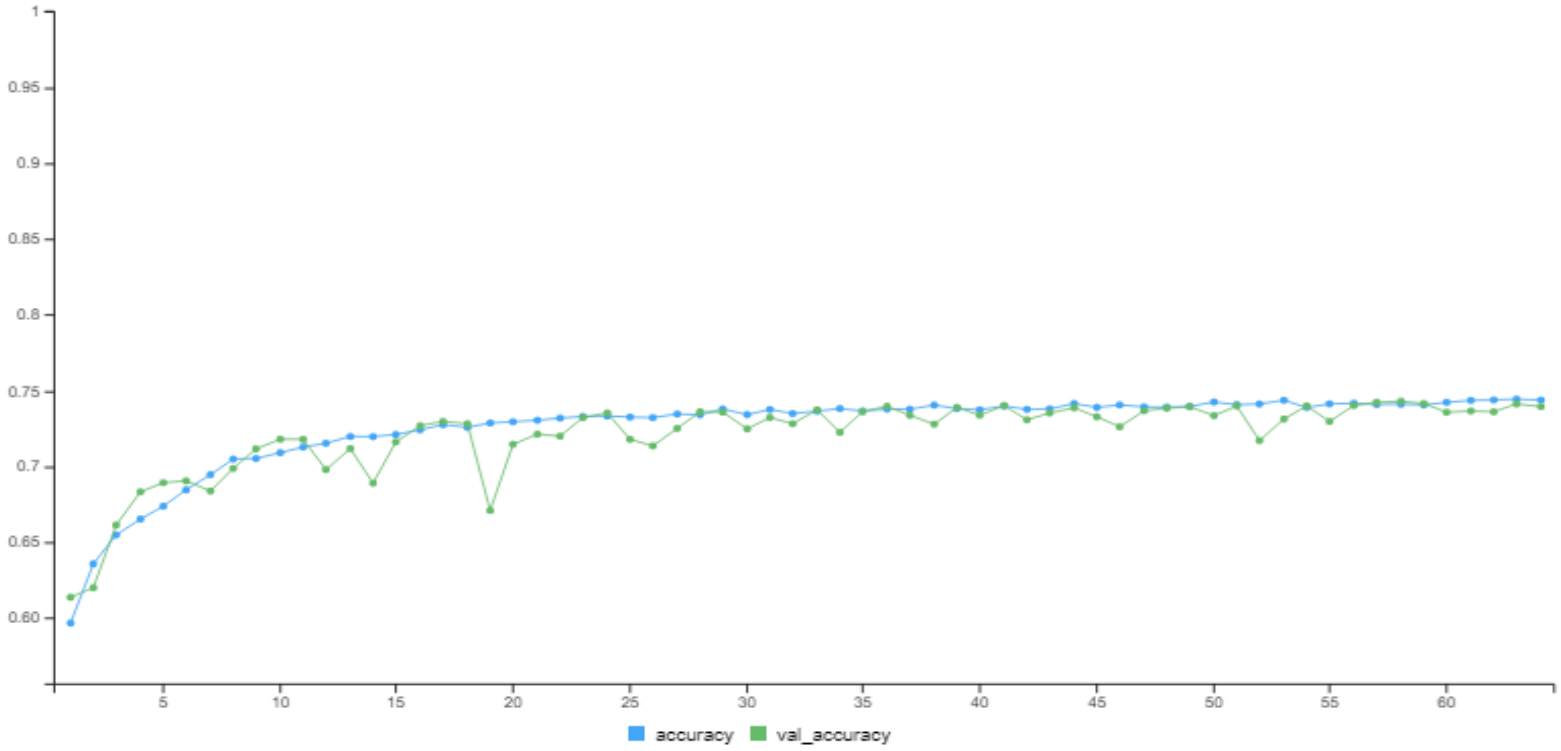
6,098 Tests	Pred Bus	Pred LRT
True Bus	**3,480 (89%)**	409 (11%)
True LRT	1,034 (47%)	**1,175 (53%)**

Importance Plots



Neural Network

1. FNN with between 1-5 layers and 32-256 nodes per layer and different activation functions
2. All results were similar, with the lowest test error of 26.0% with 2 layers of 32 nodes



6,098 Tests	Pred Bus	Pred LRT
True Bus	**3,375 (87%)**	514 (13%)
True LRT	1,050 (47%)	**1,159 (53%)**

Conclusion

Discussion

1. Smallest test error rate of 23.6% (Random Forest)
2. Behavioral interpretation with decision trees and RF's predictor importance plot are available for better interpretability
3. The main purpose of the models are predicting & analyzing LRT for given demographic/geographic input -> The accuracy for LRT could be another metric (False LRT is costlier than False Bus)
4. Baseline Decision Tree had the most accurate LRT prediction
5. Incorporating path attributes (in-vehicle time, # of transfers…) predictors derivable from shortest path passenger assignment algorithms would significantly increase the model accuracy

Results Summary

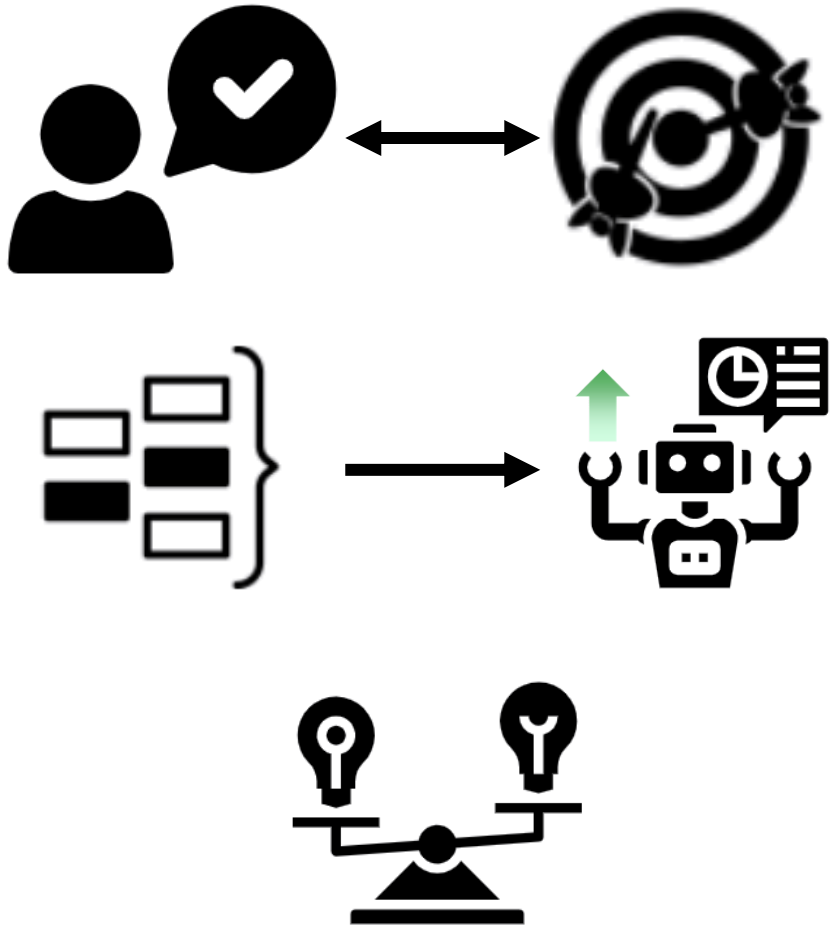
Method	Training Error Rate	Test Error Rate	Predicted LRT Given True LRT*
Logistic Regression	0.250	0.249	1,188 (-46%)
Baseline Decision Tree	0.183	0.256	1,350 (-39%)
Pruned Decision Tree	0.242	0.240	1,204 (-45%)
Most Pruned Tree	0.251	0.249	1,115 (-50%)
Random Forest	0.190	0.236	1,175 (-47%)
Neural Network	0.260	0.261	1,159 (-47%)

** Numbers in the parentheses: The difference proportion of the predicted LRT ridership compared to the actual ridership (2,209)*

Future Work

- **Potential approaches**

1. Implementing more models such as SGD Classifier or LDA
2. Applying Factors analysis since variables might have underlying relationships
3. Better feature engineering
4. Fine tuning Neural Network
5. Comparing model interpretability versus prediction accuracy



Thank you for your attention!

Questions and Answers