EXAMINING TRAVELERS WHO PAY TO DRIVE SLOWER IN THE KATY MANAGED LANES

A Thesis

by

FARINOUSH SHARIFI

Submitted to the Office of Graduate and Professional Studies of

Texas A&M University

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Mark Burris
Committee Members,	Yunlong Zhang
	Katie Turnbull
Head of Department,	Robin Autenrieth

August 2017

Major Subject: Civil Engineering

ABSTRACT

Many people believe that paying a toll to use a managed (tolled) lane will result in a shorter travel time than using the toll-free general-purpose lanes. However, there are times users pay to travel on the toll lane but go slower than the toll-free lanes. This research examined these “uneconomical trips” on managed lanes to discover potential reasons for these trips and help understand the lane choice behavior. Some potential factors considered were toll rate, traffic flow, and past trip experience.

Random forest and logistic regression methods were implemented to examine the impact and importance of variables on the probability of a user making an uneconomical managed lane trip. This thesis showed toll rate, traffic flow, travel time variability, and trip route are key factors in predicting uneconomical managed lane trips. One challenge of this study was the fact that a small percentage of trips were uneconomical trips, which leads the model to have some bias to the major class of trips. Therefore, resampling approaches including undersampling and synthetic minority oversampling technique (SMOTE) were implemented to balance the data. This study indicated undersampling technique and random forest lead to the model with the highest accuracy.

This study can help to better understand uneconomical managed lane trips and the main factors that cause these trips. Therefore, this study provides a better understanding of travel on managed lanes and general-purpose lanes.

DEDICATION

To my parents for believing in me and helping me live the best life!

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my academic advisor Dr. Mark Burris for all his contributions of time, ideas, and experiences. This thesis would not have been possible without his continuous guidance and encouragement. I also thank him for his excellent patience and caring throughout the whole thesis. I would also like to extend my gratitude to my other committee members, Dr. Katie Turnbull and Dr. Yunlong Zhang for their kind encouragement and valuable time. Special thanks to Dr. Gene Hawkins, who was willing to participate in my final defense committee at the last moment.

I would also like to appreciate Texas Department of Transportation (TxDOT), Harris County Toll Road Authority (HCTRA), and Houston TranStar for sharing the data for this project.

Finally, thanks to my family: Mom, Dad, Omid, and Farinaz for their continuous encouragement and support over the course of my education, and especially thanks to Mom and Dad for having their full support during the final stages of this research. None of my accomplishments would have ever been possible without them. Thank you!

CONTRIBUTORS AND FUNDING SOURCES

This study was supervised by a thesis committee including Dr. Mark Burris and Dr. Yunlong Zhang of the Department of Civil Engineering, and Dr. Katie Turnbull of the Department of Landscape Architecture and Urban Planning.

The data extraction and processing in Section 3 was conducted by Dr. Sunghoon Lee and AKM Abir of the Department of Civil Engineering on a study supported by Southwest Region University Transportation Center (SWUTC), Texas Department of Transportation (TxDOT), and Policy Research Center at Texas A&M Transportation Institute (TTI). All other work was completed by the student under advisement of Dr. Mark Burris of the Department of Civil Engineering.

There was no outside funding contributed to other parts of this research.

NOMENCLATURE

AUC	Area Under ROC Curve
AVI	Automated Vehicle Identification
BLR	Binary Logistic Regression
BRF	Binary Random Forest
E-ML	Economical Managed Lane
FHWA	Federal Highway Administration
GPL	General-Purpose Lanes
HCTRA	Harris County Toll Road Authority
HOT	High-Occupancy Toll
HOV	High-Occupancy Vehicles
ML	Managed Lanes
MRF	Multiclass Random Forest
NOAA	National Oceanic and Atmospheric Administration
OOB	Out-Of-Bag
ROC	Receiver Operating Characteristic
RTTD	Relative Travel Time Difference
SMOTE	Synthetic Minority Oversampling Technique
SOV	Single Occupancy Vehicles
TTD	Travel Time Difference
TTS	Travel Time Saving

TxDOT	Texas Department of Transportation
U-ML	Uneconomical Managed Lane

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. Overview

Katy Freeway is a 12-mile section of Interstate 10 (I-10) connecting the City of Katy to Downtown Houston. It consists of up to six general-purpose lanes (GPLs) and two managed lanes (MLs) in each direction. Some drivers on the MLs are required to pay a toll, depending on the time of day and number of passengers. During Monday to Friday, 5 am to 11 am, and 2 pm to 8 pm, high-occupancy vehicles (HOVs) with two or more occupants and motorcycles can use MLs for free. However, HOVs during all other times and single occupancy vehicles (SOVs) have to pay the toll that varies by time of day. The tollrate schedule is available on Harris County Toll Road Authority (HCTRA) website (https://www.hctra.org/KatyManagedLanes) and is shown in Table 1.

Table 1 Katy Managed Lanes Toll Rate Schedule

Dates	Direction	Time of Day	Toll at Eldridge (See Figure 1)	Toll at both Wilcrest and Wirt (See Figure 1)
Opening day (April 2009) to Sept 7, 2012	Westbound	Peak: 5-7pm weekdays	$1.60	$1.20
		Shoulder: 4-5 & 7-8 pm weekdays	$0.80	$0.60
		Off-peak: all other times	$0.40	$0.30
	Eastbound	Peak: 7-9am weekdays	$1.60	$1.20

Table 1 Continued

Dates	Direction	Time of Day	Toll at Eldridge (See Figure 1)	Toll at both Wilcrest and Wirt (See Figure 1)
		Shoulder: 6-7 & 9-10 am weekdays	$0.80	$0.60
		Off-peak: all other times	$0.40	$0.30
Sept 8, 2012 - Sept 7, 2013	Westbound	Peak: 4-6 pm weekdays	$2.20	$1.40
		Shoulder: 3-4 & 6-7 pm weekdays	$1.10	$0.70
		Off-peak: all other times	$0.40	$0.30
	Eastbound	Peak: 7-9 am weekdays	$2.20	$1.40
		Shoulder: 6-7 & 9-10 am weekdays	$1.10	$0.70
		Off-peak: all other times	$0.40	$0.30
Sept 7, 2013, to today	Westbound	Peak: 4-6 pm weekdays	$3.20	$1.90
		Shoulder: 3-4 & 6-7 pm weekdays	$2.10	$1.20
		Off-peak: all other times	$0.40	$0.30
	Eastbound	High Peak: 7-8 am weekdays	$3.20	$1.90
		Low Peak: 8-9 am weekdays	$2.60	$1.70

Table 1 Continued

Dates	Direction	Time of Day	Toll at Eldridge (See Figure 1)	Toll at both Wilcrest and Wirt (See Figure 1)
		High Shoulder: 6-7 am weekdays	$2.10	$1.20
		Low Shoulder: 9-10 am weekdays	$1.50	$1.00
		Off-peak: all other times	$0.40	$0.30

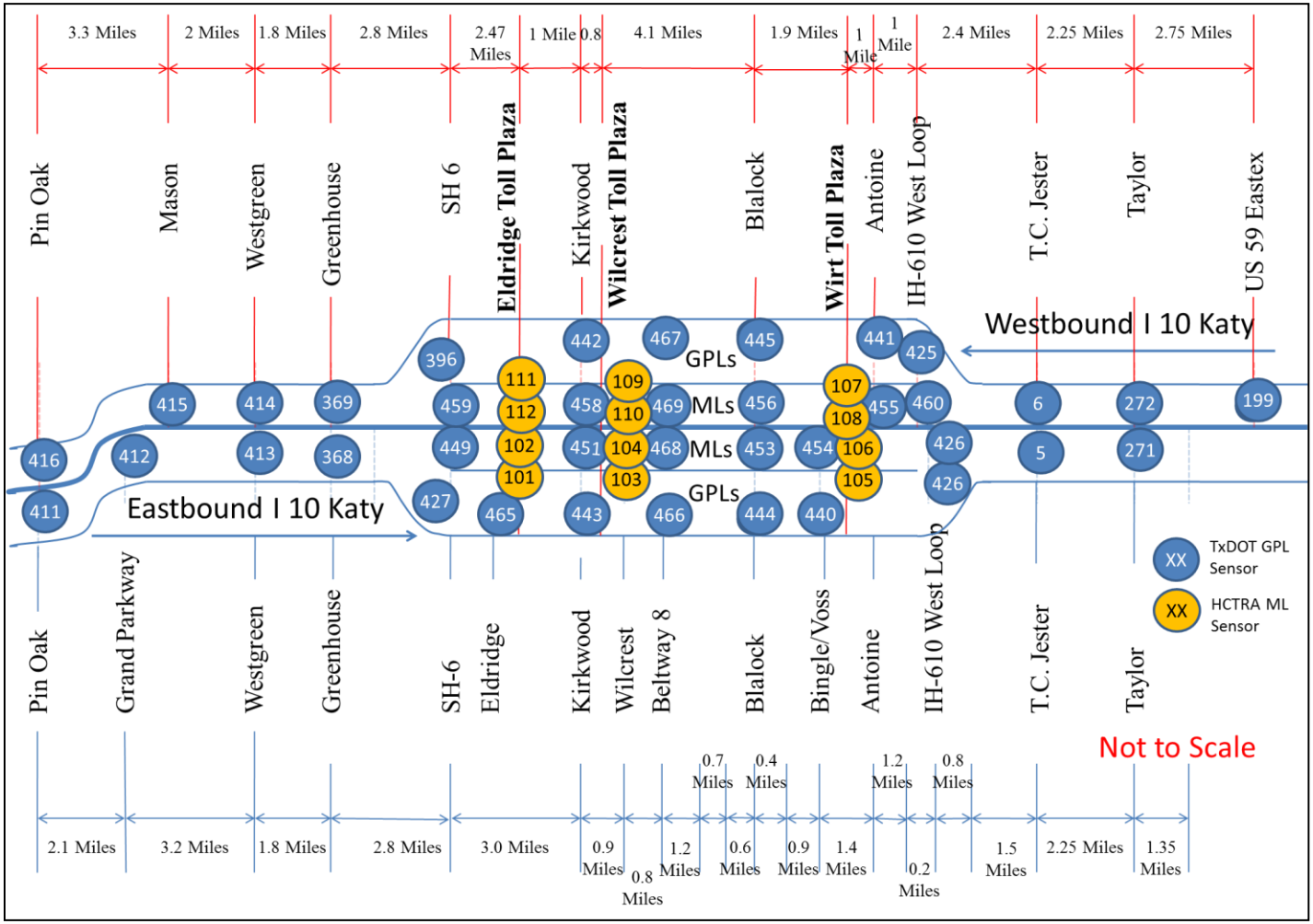

Figure 1 Automated Vehicle Identification (AVI) Sensors along Katy Freeway (Burris et al., 2016)

Generally, the MLs require less travel time than the GPLs, and it is advantageous to travel on these lanes. However, it is not always the least travel time route or the most economical route choice. In fact, approximately 11% of paid trips on the Katy MLs are 'uneconomical', meaning some drivers pay but experience a longer travel time (Burris et al., 2016). The objective of this research is to determine factors may be common among drivers who make these uneconomical trips. Is it intentional or unintentional? What factors may impact the decision the most? How do these trips affect future managed lane trips? This information should be beneficial in predicting ML travel.

Previous studies showed some possible variables that might have an impact on the travelers' lane choice decision in different conditions (Burris et al., 2016). These variables are time of day, trip length, travel time variability, trip history, and ML trip frequency. In this study, additional factors are included to improve the previous analysis. These factors include trip route, crashes, rain, and traffic flow. To determine if these variables are related to U-ML trips, pattern recognition methods are implemented. Random forest and logistic regression are applied to find possible patterns between dependent (U-ML trip) and independent variables.

1.2. Problem Statement

Katy Freeway has two MLs and at least four GPLs in each direction. The MLs generally have lower travel times than GPLs and usually save time for the traveler. Therefore, many drivers pay to use these MLs to save travel time. However, not all of these trips save travel time on the MLs. There are some trips on MLs that have a higher travel time than on GPLs despite paying a toll. Studies on almost three years of Katy

Freeway data have shown these trips account for almost 11% of total trips on the MLs (Burris et al., 2016).

For this study, nearly three years of Katy Freeway data was obtained from TxDOT automated vehicle identification (AVI) sensors and HCTRA sensors. The research will investigate paid ML trips with higher travel time than corresponding GPL trips; namely uneconomical managed lane trips (U-ML trips). The main focus of this study is to look into the U-ML trips, search for commonalities among these trips to establish some insight into this travel choice, and find the most relevant factors using pattern recognition methods.

1.3. Research Objectives

The main goal is to understand the U-ML trips better, leading to improved transportation planning models. To reach this goal, this study will:

1. explore ML trips, and specifically U-ML trips, and their characteristics
2. identify the most important variables affecting U-ML trips
3. investigate into the way these variables impact U-ML trips
4. estimate a model to predict U-ML trips.

1.4. Research Benefits

The main focus of this research is identifying common factors associated with U-ML trips. Pattern recognition methods help to recognize the factors and their ranking to determine the ones with the highest impact on this travel decision. As an example, the most important factor might be ML traffic flow. Possibly the higher the traffic flow on the MLs, the higher the probability of having a U-ML trip. Also, travel in the east

direction might show a high chance of having an uneconomical trip. Therefore, an eastbound ML trip taken when there is high traffic flow on the MLs would lead to an increased chance of a U-ML trip.

This study can help to explain the U-ML trips better and identify them based on the corresponding key elements, which leads to a better ML travel prediction and travel behavior understanding.

2. LITERATURE REVIEW

In this part, the existing literature on ML travel, machine learning techniques, and imbalanced data resampling will be provided to help with identifying the U-ML trips and associated parameters as well as establishing the best approach to model the U-ML trips.

2.1. Managed Lanes

As defined by the Federal Highway Administration (FHWA) publication (2012), managed lanes (MLs) are "designated lanes or roadways within highway rights-of-way where the flow of traffic is managed by restricting vehicle eligibility, limiting facility access, or and in some cases collecting variably priced tolls". One type of ML is HOT (High-Occupancy/Toll) lanes, which enables HOVs (High-occupancy Vehicles) to use the managed lane for free or lower toll. Other vehicles have to pay a higher toll to use MLs. Electronic toll collection and informative traffic-related message signs are typical features of HOT lanes.

One of the primary traffic management goals for priced MLs is congestion reduction, enabling the vehicles to travel at higher speeds and save travel time. That is, one of the main benefits of MLs is travel time saving (TTS) and people pay for this time saving. Also, MLs generally offer a more reliable travel time and help the environment by reducing the vehicle emissions and noise (FHWA, 2012).

Nevertheless, MLs may not always have a shorter travel time than the GPLs. In other words, drivers may pay to use the MLs, but their travel time on MLs is longer than

on the GPLs. These trips with higher travel time on MLs were previously studied by Burris et al. (2016). They focused on the Katy MLs, which was converted to a ML facility in 2009. Burris et al. (2016) examined travelers' lane choice behavior and the influencing factors. Also, they indicated Katy ML's TTS for paid ML trips ranged from -3.3 to over 20 minutes with an average of 2.6 minutes.

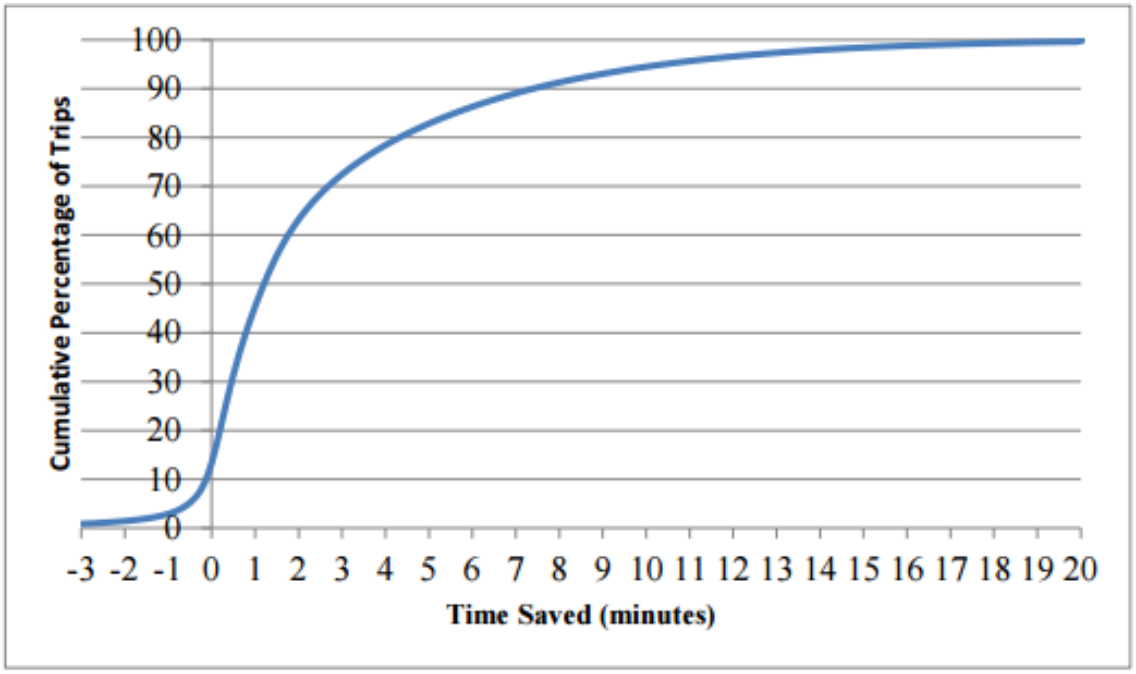

Figure 2 Travel Time Saved on Katy MLs (Burris et al., 2016)

As illustrated in Figure 2, approximately 11% (11.3% in 2012, 11.5% in 2013, and 10.8% in 2014) of the paid ML trips had a negative travel time saving or were slower on the MLs than on the GPLs (termed uneconomical trips). Burris et al. (2016) also observed that travelers were not willing to change their lane because of a bad trip

experience (a trip with speed much slower than all other vehicles' average speeds). In fact, travelers may not change their lane because of an uneconomical trip experience. Most of the ML studies center on how much these lanes would save travel time. Brownstone et al. (2003) studied the willingness to pay to reduce travel time on I-15. They concluded drivers are willing to pay up to $30 to reduce one hour of travel time. Also, Burris et al. (2007) conducted a survey of travelers to find their potential use of MLs. They concluded travel time saving and travel time reliability are the most important factors in choosing a toll lane. Sullivan (2000) also examined the perceived travel time saving along SR-91. He found the 48% (in the AM peak hour) and 23% (in the PM peak hour) believe in having a travel time saving of less than 15 minutes. Also, he noted approximately all of the respondents except a small minority in the AM peak hour overestimated their travel time saving. Buckeye (2012) evaluated the performance of I-394 in Minnesota. He used the speed as a measurement of effectiveness, and found the average speed for the ML is higher than the GPLs. However, he did not further study the days with a lower average speed for the ML. Burris et al. (2012) found a small difference between the ML speed and the GPL speed on I-394, and 35% of travelers paid for a travel time saving of less than one minute. He concluded that the small travel time saving obtained from this small speed increase cannot be the only reason of choosing ML over GPLs, and it might be as the result of avoiding a bottleneck or the higher reliability of ML. Kwon and Varaiya (2008) included negative travel time saving in their study on the effectiveness of the HOV system in California. The average travel time

savings between a random 10-mile HOV lane and the adjacent GPL is 1.7 minutes, and they noted some negative travel time savings for the random HOV facility.

However, no studies have been undertaken to confirm travel time loss or negative travel time savings on toll-paid MLs. In addition, most studies find that travel time savings is the most important factor in choosing to use MLs. In other words, the models assume travelers do not take MLs if they do not save travel time. Devarasetty et al. (2012) found the value of travel time and travel time reliability as the main incentives in choosing MLs. Later, Devarasetty et al. (2014) also noted that the travel time saving is recognized as the most influential factor in selecting a ML over GPLs by many studies, and the value of time and willingness to pay is calculated based on the fact that travelers pay to drive faster. Lam and Small (2001) computed the value of time to be $22.78 per hour for SR-91 in Orange County. They acknowledged the small travel time saving. However, they did not conduct any further study on travelers paying the toll and going slower. Gardner et al. (2013) examined the probability of a user choosing a HOT lane based on the cost to travel time savings ratio. However, they did not study the possibility of having a HOT lane trip with both monetary and time cost. This thesis focuses on the travel time loss of paid managed lanes users and implements various techniques to diagnose the pattern or relationship between the key variables and U-ML trip probability on the Katy Freeway.

2.2. Machine Learning

As stated by Bishop (2006), pattern recognition and machine learning are two interpretations of the same concept with separate fields of application. Pattern

recognition is grounded on engineering, while machine learning originated from computer science. They are both the effort of discovering regularities among the data or classifying the data by using computer algorithms. Pattern recognition and machine learning methods are characterized based on how they generate the output. The output can be either a data structure pattern or a set of variables. It is called supervised learning when there is a set of input data and output data to be trained. On the other hand, unsupervised learning attempts to find similarities among the data and classify them.

As mentioned by Bishop (2006), to evaluate the predictive power of the model, the dataset is usually divided into two groups, the training set and the test set. The training set is used to learn and fit the model. To assess the prediction of the model, an unseen dataset should be tested. This unseen dataset is called the test set. Hastie et al. (2001) remarked that it is problematic to find a general rule for percentages of training and test split. Dobbin (2011) also conducted a comprehensive review of the optimal split of the dataset, and stated that the training set should be 40% to 80% of the total data size. Consequently, this study will take 80% of the dataset as the training set to also make up for the loss of data in the resampling step.

A good model is one that predicts the test set outcome accurately. Hastie et al. (2001) noted that the confusion matrix analysis could evaluate the model's prediction ability. Confusion matrices represent the predicted amounts of each class versus the actual amounts of each class. Table 2 shows a typical confusion matrix.

Table 2 Confusion Matrix

		Prediction Condition	
		Positive Prediction	Negative Prediction
Actual Condition	Positive Actual	True Positive (TP)	False Negative (FN)
	Negative Actual	False Positive (FP)	True Negative (TN)

The first parameter is accuracy, which is defined as the number of true predictions divided by the total population as formulated in Equation 1. It is also a representation of the prediction error rate.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} = 1 - Error = 1 - \frac{FN+FP}{TP+FN+FP+TN} \quad (1)$$

Chawla (2005) argued that the accuracy is not a satisfactory parameter by itself, and defined other parameters to assess the model. False positive (FP) or Type I error is the probability of rejecting a null hypothesis when it is actually true. False negative (FN) or Type II error is the probability of not rejecting a null hypothesis when the alternative hypothesis is actually true. The FN quantity is usually more critical than the FP quantity because of the higher risk of not predicting a positive value. However, both of them are equally vital in this study. Other evaluation parameters are as follows:

$$Positive\ Predictive\ Value\ (PPV, Precision) = \frac{TP}{TP+FP} \quad (2)$$

$$True\ Positive\ Rate\ (TPR, Recall, Sensitivity) = \frac{TP}{TP+FN} \quad (3)$$

$$True\ Negative\ Rate\ (TNR, Specificity) = \frac{TN}{TN+FP} \quad (4)$$

$$False\ Positive\ Rate\ (FPR, Fall\ out) = \frac{FP}{TN+FP} \quad (5)$$

$$False\ Negative\ Rate\ (FNR) = \frac{FN}{TP+FN} \quad (6)$$

Chawla (2005) suggested that a good approach to model a classifier is to maximize the true rates and minimize the false rates. This is the foundation for the main technique of the model assessment in this study. Receiver Operating Characteristic (ROC) curve is a graph which displays the errors of TPR versus FPR. Actually, ROC curves express a tradeoff between the TPR (benefit) and the FPR (cost) of a classifying model (Fawcett, 2005).

The area under the ROC curve (AUC) is the main model evaluation parameter in this study. The AUC ranges from 0 to 1, and it indicates that each randomly chosen positive observation has this amount of probability to be classified positive truly rather than negative (Fawcett, 2005). The higher the AUC is, the better the classifier is.

As noted earlier, machine learning methods fall into two categories of supervised and unsupervised learning. This study implements supervised learning techniques to predict U-ML trip likelihood. Logistic regressions and random forests are two methods of supervised machine learning for predicting the U-ML trips.

2.2.1. Logistic Regression

Logistic regression is a method of pattern recognition used in this study to predict the U-ML trips. As explained by Train (2003), logistic regression is a regression model for predicting discrete choice or categorical dependent variables. In other words, it can be employed for the data with binary or fail/win output, which is the focus of this study. The main form of the model is shown in Equation (7):

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{7}$$

Where:

$\beta_1 X$ = Regression coefficient multiplied by the independent variables

β_0 = Intercept of the linear equation

$p(X)$= Probability of X happening (between 0 and 1).

Logistic regression can help this study to examine the magnitude of each variable impact.

2.2.2. Random Forest

As defined by Brieman (2001), random forest is a technique of machine learning, which is an ensemble of decision trees. The main concept of random forest is to create a strong classifier by gathering all small decision trees together. Each tree in the random forest gets a set of input observations and produces a set of outputs or votes for the random forest output. The output of the random forest model is the mode or mean of all decision trees' outputs.

Random forest includes a large number of trees, say T trees. Each tree of $t \in \{1,...,T\}$ in the random forest is trained by using two-thirds of the main training set, and leaving one-third of the main training set out. The left-out part of the training set is termed "out-of-bag (OOB) data". This tree consists of multiple nodes for splitting different variables. To assign a splitting variable to each node of tree, a specific number of variables, say m, is selected from the tree input set. The number m is consistent and optimized for the whole random forest modeling procedure. The node finds the best variable among the selected m variables to split the data. This variable selection step

helps random forest to include minor effective variables in the model and reduce the variance significantly (James et al., 2013).

Random forests also do not need any type of model's validation. The OOB data will be tested by each tree, and the error of this prediction is named the overall OOB error. Overall OOB error decreases as the result of increase in the number of trees (James et al., 2013).

Random forests also rank the independent variables based on their importance in predicting the dependent variable. While testing the OOB data in each tree, the prediction error for the OOB data is recorded. Then, variables are permuted one by one, and the prediction error is computed for each permuted variable. The increase in normalized prediction error or decrease in accuracy is saved for each permuted variable and the associated tree. To rank the variables' importance, the decrease in accuracy for each variable is averaged over all trees in the random forest, and a new unique value, named "Mean Decrease in Accuracy", is created as an indication of the variable importance (Hastie et al., 2001)

Random forests are efficient and functional in dealing with large datasets, and they do not overfit. This flexibility and the variable importance ranking are the two main reasons to benefit from random forests in this study.

2.2.3. Applications in Transportation

This section will examine some previous transportation studies with application of logit and random forest techniques.

Sekhar et al. (2016) applied random forest and multinomial logit model to model the mode choice behavior of commuters in Delhi. 5000 stratified samples were collected by surveying households in Delhi. The final results showed the random forest had a higher accuracy (98.96%) than the logit model (77.31%).

Hagenaur and Helbich (2017) conducted a study to examine the mode choice classifiers. Their dataset was collected as Dutch National Travel Survey from 2010 to 2012. They also added environmental data to include weather conditions in their study. Seven different classifiers including random forest and multinomial logit model were suggested to model the mode choice. Random forest with accuracy of 0.961 was the most accurate model. Multinomial logit model, which was in use by the public services at the time of study, had a lowest accuracy (0.561).

Xiao et al. (2017) examined the transportation modes using Global Position System (GPS) data and modeling with tree-based models. They used AUC to evaluate the models, and found the ensemble models including random forest perform better than the traditional models. They also noted the ensemble models are like a black box, and the explanation of the final model may not be easy.

A great number of studies has studied lane choice behavior by using logistic regression models (Burris et al., 2007, Burris et al., 2016, Davarasetty et al., 2012, Davarsetty et al., 2014). This technique makes the variables of study and their impacts become more clear and easier to intuitively explain.

The current study will use both random forest and logistic regression model to evaluate their prediction along with finding the U-ML trip pattern.

2.3. Imbalanced Dataset

An imbalanced dataset is a dataset with approximately unequally sized classes (Chawla, 2005). In other words, there is a great gap between the proportions of each class in the dataset. It is usually valid for the datasets with a “rare” or “abnormal” event, which is the case of this study.

As explained by Chawla (2005), the challenge of an imbalanced dataset is that the models trained by this dataset will be biased to the major class. Models are generally developed in a way to increase accuracy. However, accuracy is not the best parameter for evaluation of a model.

Considering a case where 95% of the output is 0, and 5% of the output is 1 shows that if all models try to predict the zero value for all data and reach an accuracy of 0.95. The main interest of the study is predicting the rare event, and this model is not helpful in predicting any rare events. This fact is called the “paradox of accuracy”. This concept states that there may be some predictive models with a level of accuracy and higher prediction capability than other predictive models with higher accuracy (Zhu, 2007). Zhu (2007) suggested to use other parameters including sensitivity and specificity to evaluate the model. Chawla (2005) finds AUC the most useful parameter to evaluate the model’s goodness of fit.

To train a model, the imbalanced dataset should be resampled to develop an approximately equally sized classes dataset. Various techniques have been developed to balance the data as reviewd by Kotsiantis et al. (2006):

1. Undersampling: the first method aims to include all the minor class observations and randomly selects part of the major class observations in the dataset. In other words, this technique will downsize the major class. Using this approach may discard and lose part of the data.
2. Oversampling: this technique will include the whole major class and minor class observations plus adding some randomly sampled observations from minor class to balance the dataset. In other words, this technique will increase the minor class observations by sampling observations with replacement. This method may cause an overfitting issue in the final models.
3. Synthetic Minority Oversampling Technique (SMOTE): This approach increases the size of the minority class by creating synthetic examples rather than replicating them. In this method, a number of nearest neighbors is obtained by means of K-nearest neighbors. To create a new observation in this method, the difference of minor class observation and one of its nearest neighbors is computed and multiplied by a random number from zero to one. Later, this is added to the minor class observation to produce a new synthetic observation (Chawla, 2002).

While Batista et al. (2004) suggested that the oversampling techniques especially SMOTE resulted in more accurate models with higher AUCs, Blagus and Lusa (2013) argued that undersampling method led to a more accurate model for a high-dimensional imbalanced dataset. Current study will implement both undersampling technique and SMOTE to create a balanced dataset and find the most accurate technique.

3. DATA

In this study, a unique dataset obtained from Katy Freeway will be investigated. This chapter of the thesis will introduce Katy Freeway, dataset sources, and the main and sample dataset for this study.

3.1. Katy Freeway

Interstate 10 (I-10) is a major east-west interstate highway. Katy Freeway is a section of I-10 connecting the City of Katy to Downtown Houston. It is 12 miles with between four to six GPLs and two MLs in each direction (see Figure 3).

The Katy Freeway was converted to a ML facility in 2009. Most ML travelers are required to pay a toll. The HCTRA is responsible for toll rates and toll collection at three toll plazas. Tolls are electronically collected via EZ Tag or TxTag. Toll rates vary by time of the day, day of the week, and number of passengers. It turns to a high-occupancy toll lane (HOT) from Monday to Friday from 5 am to 11 am and 2 pm to 8 pm. During these hours, HOVs with two or more occupants and motorcycles can use MLs for free. However, HOVs during all other times and SOVs have to pay the toll that varies by time of the day. The toll rate schedule is available on HCTRA website (https://www.hctra.org/KatyManagedLanes) and is shown in Table 1. Katy MLs provide a free commute for the HOV drivers and a new commuting option for toll-paying SOV drivers.

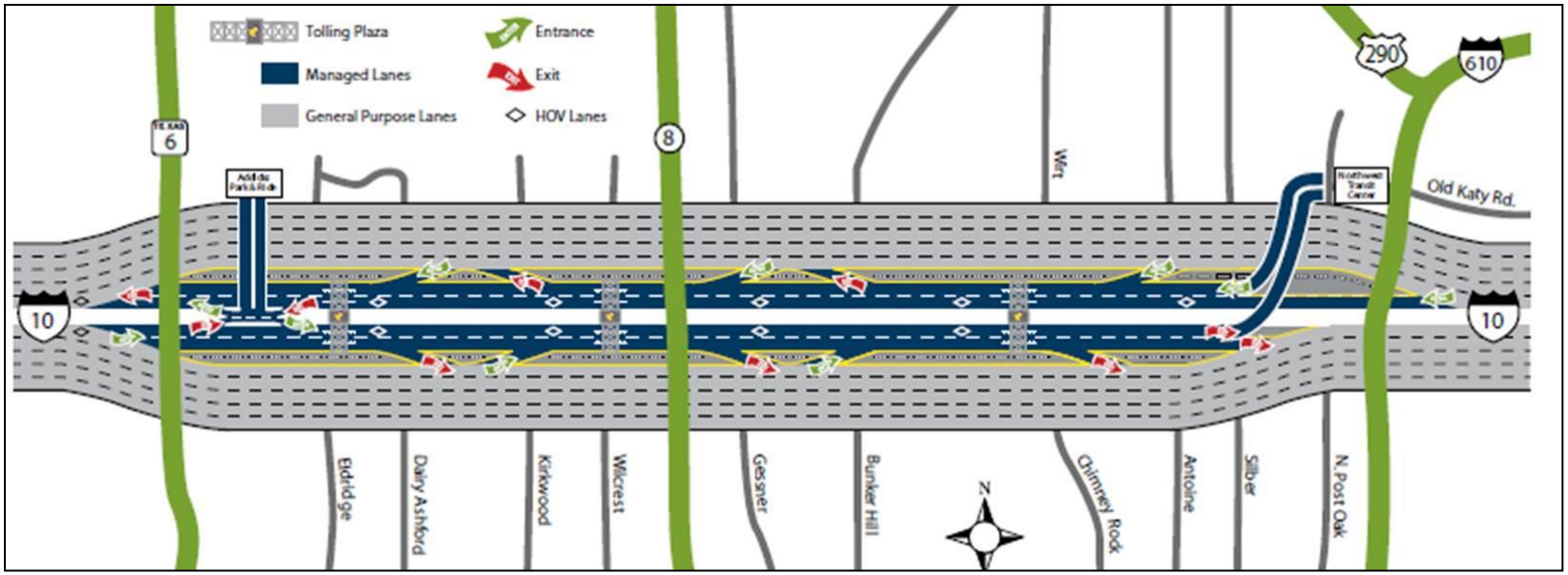

Figure 3 Katy Freeway (HCTRA, 2009)

3.2. Data Sources

The main dataset used in this study is from three main sources. The primary two data sources contain vehicles' trip information. These two datasets are combined to form a unique dataset including travel information from most of 2012, 2013, and 2014. The third dataset adds daily precipitation measurements to enrich the final dataset with environmental effects.

3.2.1. TxDOT AVI Data

The first part of data is acquired from automated vehicle identification (AVI) sensors operated by TxDOT. There are 38 AVI sensors with unique sensor numbers located along GPLs and MLs as illustrated in Figure 1. When they detect a vehicle, they record the vehicle's transponder ID, sensor ID, and detection time. All vehicles using MLs are required to have transponders. This data includes most of the trip records from 2012, 2013, and 2014 with transponder ID, AVI number, and detection time.

3.2.2. HCTRA Toll Data

The second part of the data is obtained from the HCTRA. They collect data from 12 AVI sensors at three toll plazas and use that to charge vehicles the appropriate toll rates. The AVI sensors are shown in Figure 1. This data also records vehicle's transponder ID, toll plaza ID, lane ID, and the detection time as the vehicle passes each sensor.

3.2.3. NOAA National Centers for Environmental Information

The third dataset is attained from NOAA (National Oceanic and Atmospheric Administration) to include daily precipitation effect in the study examination. NOAA

National Center for Environmental Information website (https://www.ncdc.noaa.gov/cdo-web/datatools/findstation) has a valuable source of daily weather summary for several stations in the Houston area. This study uses daily rain measurements from the closest station to Katy Freeway. This station is named “HOUSTON 11.8 WNW TX, US” with coordination (29.8066°, -95.5607°). The exact location of this station is mapped in Figure 4. To coordinate this data with two other datasets, the precipitation data is obtained from January 2012 to September 2014.

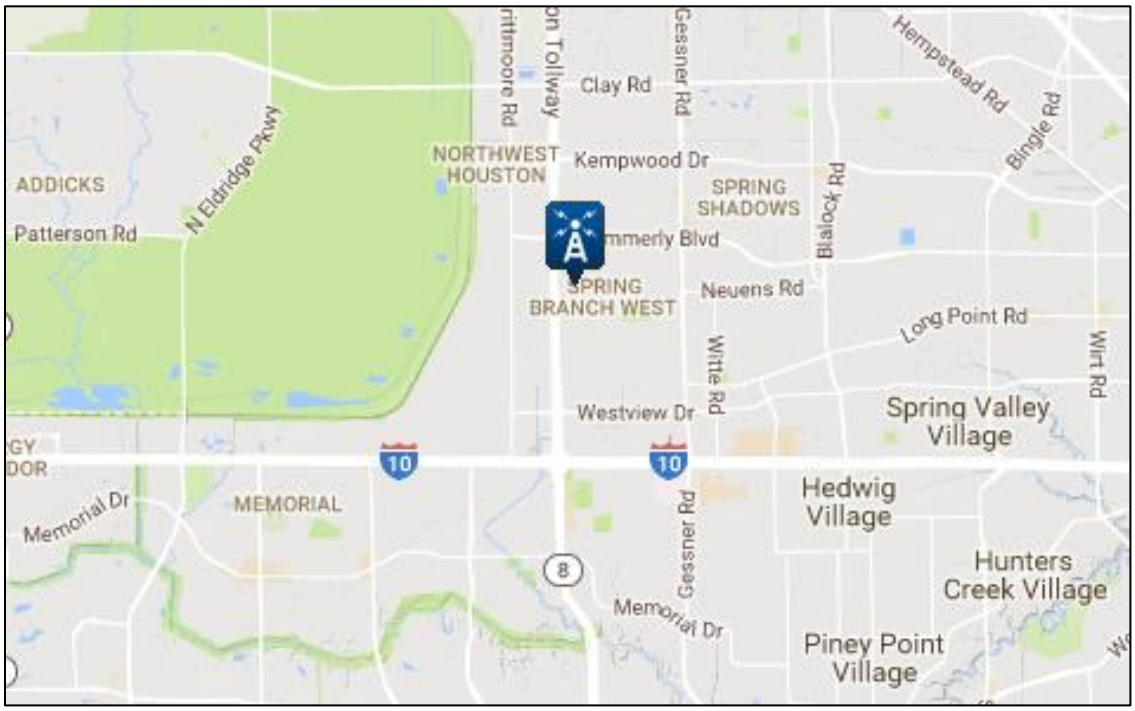

Figure 4 Rain Station Location

3.3. Main Dataset

TxDOT AVI sensors and HCTRA toll data are combined to form the vehicles' travel information data, which includes their trip route, trip time, and paid toll. Daily rain data is also combined to form the main dataset. The main dataset is starting from January 2012 to November 2012 and January 2013 to September 2014, and covering 7,013,587 ML trips. In this part, a series of cleaning and processing steps are implemented.

Firstly, the original transponder ID is changed to a random ID for each traveler to respect the anonymity of travelers. Next, those toll-free HOVs are excluded from the main dataset to focus on ML toll-paid users. Also, lane closure is the fourth dataset, which is derived from TxDOT for the three years of study and added to the main dataset to include the factor of lane closures.

3.3.1. Alternate GPL Trip

One key feature of this study is estimating an alternate GPL trip for each ML trip. This alternate GPL trip helps to compare ML trip travel time with a toll-free trip travel time to examine the economy of the trip. In other words, this alternate GPL trip travel time is the main factor in defining U-ML trips.

Also, some other GPL trip attributes like the number of vehicles with transponders on GPLs can be a clue for explaining the U-ML trips. Consequently, an alternate GPL trip is generated for each ML trip to both define U-ML trips and examine their related factors.

To characterize the alternate GPL trip, the start time for the alternate GPL trip is considered the same as the actual ML trip. Other attributes of the alternate GPL trip are estimated using three following scenarios:

1. Data from other vehicles on GPLs with the same start time are available to calculate the alternate GPL trip attributes.
2. Data from other vehicles on GPLs with the same start time are available for a part of the trip. This data will help to compute the partial alternate GPL trip attributes. Subsequently, alternate GPL trip attributes for the rest of the trip length will be assessed based on the average of other vehicles' attributes on the remainder of the length.
3. There are no data from other vehicles on any segment of GPLs at the start time of the actual trip. Therefore, the alternate GPL trip is generated using the average speed at that time of day on that segment length.

Based on this, the alternate GPL trip attributes are created. Among these attributes, the alternate GPL trip travel time is the main feature to detect U-ML trips.

3.3.2. Uneconomical Managed Lane Trip Identification

These two features indicate a U-ML trip:

1. Vehicles that pay a toll to use the MLs (not HOVs during the peak hours and motorcycles). This feature is already in the main dataset because all toll-free HOVs are excluded from the main dataset, and the main dataset is only focusing on paid ML trips.

2. Vehicles' actual ML trip travel time is longer than their alternate GPL trip travel time.

Therefore, the first classification of a U-ML trip can be placed as a binary parameter termed "$uneco_{binary}$", which is:

$$uneco_{binary} = \begin{cases} 1\ if\ (TT_{ML} \geq TT_{GPL})\ (Uneconomical\ ML\ trip) \\ 0\ if\ (TT_{ML} < TT_{GPL}) \quad (Economical\ ML\ trip) \end{cases} \tag{8}$$

Where:

TT_{ML}=Actual ML trip travel time

TT_{GPL}=Alternate GPL trip travel time

This classification only reflects the travel time saving or loss by the sign of the difference between actual ML trip and alternate GPL trip. The border between the two classes of this type of classification is the travel time difference of zero. In other words, this type of classification does not distinguish losing one minute in a two-minute trip from losing one minute in a ten-minute trip. Hence, another classification for U-ML trip identification is established to differentiate the economical and uneconomical ML trips with a wider margin. This new classification is defined with a variable termed "$Uneco_{multiclass}$".

$Uneco_{multiclass}$ is a variable dividing ML trips into three groups: economical ML (E-ML) trips, U-ML trips, and too close to decide or middle ML trips. The final class is developed for the ML trips with a small travel time saving or loss. To fairly adjust the interval for these ML trips, the relative travel time difference is defined as:

$$Relative\ Travel\ Time\ Difference\ (RTTD) = (TT_{ML} - TT_{GPL})/TT_{ML} \tag{9}$$

RTTD helps estimate the travel time loss more equitably. Referring to the previous example, RTTD is 0.5 for losing one minute in a two-minute trip. However, it is 0.1 for losing one minute in a ten-minute trip. RTTD makes the distinction clearer. To justify the classification of ML trips, $uneco_{multiclass}$ is defined as Equation (10).

$$uneco_{multi-class} = \begin{cases} 1 & if\ (RTTD > 0.05)\ (Uneconomical\ ML\ trip) \\ 0.5 & if\ (-0.05 \leq RTTD \leq 0.05)\ (Middle\ ML\ trip) \\ 0 & if\ (RTTD < -0.05)\ (Economical\ ML\ trip) \end{cases}$$

(10)

The second set includes ML trips with a small travel time loss or saving which is not significantly different from 0 (considering a p-value of 0.05 in each one-tail test).

$Uneco_{binary}$ and $Uneco_{multiclass}$ are the main response variables to be examined in this study.

3.3.3. Travel Behavior and Trip Frequency

To define the trip frequency in the ML trip examination, the number of previous month's trips is computed for each traveler. To evaluate the frequency, the number of previous trips in a same length of time should be considered for each trip. Previous month is selected because it is the closest period to the studied trip.

$$Total\ trip\ frequency = Past\ month\ total\ number\ of\ trips$$

$$ML\ trip\ frequency = Past\ month\ number\ of\ ML\ trips$$

Travel behavior is similarly a principal element to study ML trip classification. This element can be formulized as the percent ML trips (Equation 11):

$$Percent\ ML\ Trips = \frac{ML\ trip\ frequency}{Total\ trip\ frequency} \quad (11)$$

In the following step, variables of interest are selected among all available attributes and their type is assigned. Variables can be either numerical, measuring an observation's characteristic, or categorical, assigning a class to each observation. The dataset is reduced to the variables defined and classified in Table 3.

Also, an example of the dataset is shown in Table 4.

Table 3 Dataset Variables Definitions

Variable	Description	Variable Type	Class
$Uneco_{binary}$	1=U-ML trip 0=E-ML trip	Categorical	Output
$Uneco_{multiclass}$	1=U-ML trip 0.5=Middle ML trip 0=E-ML trip	Categorical	Output
TT_{ML}	Actual ML travel time	Numerical	Trip Length
TT_{GPL}	Alternate GPL travel time	Numerical	Trip Length
TTD	Travel time difference	Numerical	Trip Length
RTTD	Relative travel time difference	Numerical	Trip Length
Std	Standard deviation of ML travel time between the start and end sensors during the 10-minute interval at the time of travel over 20 weekdays prior to the trip	Numerical	Trip Length
Travel time variability	Coefficient of variation, travel time variability (Std/time)	Numerical	Trip Length
Weekday	1=Sunday, 2=Monday, 3= Tuesday, 4=Wednesday, 5=Thursday, 6=Friday, 7=Saturday	Categorical	Trip Time
Peak	1=1st hour of peak hour 2=2nd hour of peak hour Peak Hours: Weekdays, 7-9 am Eastbound, 4-6 pm Westbound	Categorical	Trip Time

Table 3 Continued

Variable	Description	Variable Type	Class
Shoulder	1=shoulder hour before peak hour 2=shoulder hour after peak hour Before Sep 8, 2012 Shoulder Hours: Weekdays, 6-7 am and 9-10 am Eastbound, 4-5 pm and 7-8 pm Westbound Off-peak Hours: All other times After Sep 8, 2012 Shoulder Hours: Weekdays, 6-7 am and 9-10 am Eastbound, 3-4 pm and 6-7pm Westbound Off-peak Hours: All other times	Categorical	Trip Time
Length	Travel Length of actual ML trip	Numerical	Trip Length
Direction	Direction of travel 0= Eastbound, 1=Westbound	Categorical	Geometry
Start sensor	Start sensor of the actual ML trip	Categorical	Geometry
End sensor	End sensor of the actual ML trip	Categorical	Geometry
Main lanes blockage	Number of main lanes blocked due to incidents	Numerical	Blockage
Frontage lanes blockage	Number of frontage lanes blocked due to incidents	Numerical	Blockage
HOV lanes blockage	Number of HOV lanes blocked due to incidents	Numerical	Blockage
Ramp lanes blockage	Number of ramp lanes blocked due to incidents	Numerical	Blockage
Shoulder lanes blockage	Number of shoulder lanes blocked due to incidents	Numerical	Blockage
Rain	Daily precipitation	Numerical	Rain
Total toll	Total toll paid	Numerical	Toll
Toll rate	Toll per length	Numerical	Toll

Table 3 Continued

Variable	Description	Variable Type	Class
ML traffic flow	Number of vehicles with transponders on ML that were traveled between the start and end sensors during the 10-minute interval at the time of travel	Numerical	Traffic
GPL traffic flow	Number of vehicles with transponders on GPL that were traveled between the start and end sensors during the 10-minute interval at the time of travel	Numerical	Traffic
ML trip frequency	Number of past month's paid ML trips	Numerical	Experience
Total trip frequency	Number of past month's total trips	Numerical	Experience
Percent ML trips	Rate of past month ML trips to total trips	Numerical	Experience

3.4. Sample Set

The Travel Survey Manual (Tierney et al., 1996) suggests a sample size of as small as 1000 for travel behavior studies. However, this study samples more data in order to consider all levels of variables and correlations among variables. Furthermore, the sample would be divided into two groups of training and test sets, and training set would be resampled. All these facts plus the software ability to analyze higher size of sample, the sample size is estimated as 1 trip each seven trips.

Table 4 An Example of Dataset

Trip number	Randid	Direction	Week day	Peak	Shoulder	Main lanes blockage	Frontage lanes blockage	Ramp lanes blockage	HOV lanes blockage	Shoulder lanes blockage
1	934594339	1	2	0	0	0	0	0	0	0
2	934588670	0	7	0	0	0	0	0	0	0
3	934588670	0	3	0	0	0	0	0	0	0
4	934588670	0	5	0	0	0	0	0	0	0
5	934588670	0	6	0	0	0	0	0	0	0
6	934588670	0	7	0	0	0	0	0	0	0
7	934588670	1	7	0	0	0	0	0	0	0
8	934588670	0	4	1	0	0	0	0	0	0
9	934588670	0	7	0	0	0	0	0	0	0
10	934588670	0	7	0	0	0	0	0	0	0
11	934588670	0	4	0	0	0	0	0	0	0
12	934588377	0	4	0	0	0	0	0	0	0
13	934585941	1	2	0	0	0	0	0	0	0
14	934584384	1	4	0	0	0	0	0	0	0
15	934581638	1	6	0	0	0	0	0	0	0
16	934565133	1	7	0	0	0	0	0	0	0
17	934565133	0	3	0	1	0	0	0	0	0
18	934565133	0	2	0	1	0	0	0	0	0
19	934565133	0	3	0	1	0	0	0	0	0
20	934565133	0	4	0	1	0	0	0	0	0

Table 4 Continued

Trip number	Rain	TT_{ML}	std	Total toll	ML traffic flow	Length	Start sensor	End sensor	Travel time variability	TT_{GPL}
1	0	1.37	0.19	0.7	1	1.81	109	111	0.13	1.50
2	0	4.97	0.40	0.6	1	5.45	103	105	0.09	4.79
3	0	6.53	0.00	0.7	1	7.26	102	106	0.00	6.67
4	0	5.55	0.37	1	2	7.26	101	105	0.06	6.52
5	0	5.85	0.25	0.7	2	7.26	102	106	0.04	6.91
6	0	6.70	0.30	0.7	2	7.26	102	106	0.05	6.73
7	0	1.53	0.17	0.7	1	1.81	110	112	0.11	1.52
8	0	6.90	0.00	1.2	2	7.26	102	105	0.00	9.03
9	1.07	5.90	0.28	0.7	1	7.26	102	105	0.05	6.75
10	1.07	6.35	0.00	0.7	2	7.26	102	105	0.00	6.50
11	0	6.35	0.00	0.7	1	7.26	102	106	0.00	6.14
12	0	6.08	0.00	0.7	1	7.26	102	106	0.00	6.80
13	0	1.63	0.07	0.7	3	1.81	109	111	0.05	1.67
14	0	6.10	0.75	1	6	7.26	107	111	0.12	6.37
15	0	5.67	0.31	0.4	1	7.26	108	111	0.05	6.72
16	0	5.48	0.28	1	1	7.26	108	112	0.05	6.25
17	0	6.07	0.29	0.6	30	7.26	102	106	0.05	8.24
18	0.39	6.65	0.00	0.8	2	7.26	101	106	0.00	7.63
19	0	6.15	0.23	0.6	20	7.26	102	106	0.04	7.41
20	0	6.15	0.23	0.6	46	7.26	102	106	0.04	7.90

Table 4 Continued

Trip number	GPL traffic flow	Total trip frequency	ML trip frequency	Percent ML trips	Toll rate	TTD	$Uneco_{binary}$	RTTD	$Uneco_{multiclass}$
1	26	0	0	0	0.39	-0.14	0	-0.10	0
2	5	3	3	1	0.11	0.18	1	0.04	0.5
3	10	3	3	1	0.10	-0.14	0	-0.02	0.5
4	16	3	3	1	0.14	-0.97	0	-0.17	0
5	23	3	3	1	0.10	-1.06	0	-0.18	0
6	12	3	3	1	0.10	-0.03	0	0.00	0.5
7	38	3	3	1	0.39	0.02	1	0.01	0.5
8	27	3	3	1	0.17	-2.13	0	-0.31	0
9	12	3	3	1	0.10	-0.85	0	-0.14	0
10	12	3	3	1	0.10	-0.15	0	-0.02	0.5
11	0	3	3	1	0.10	0.21	1	0.03	0.5
12	8	1	1	1	0.10	-0.72	0	-0.12	0
13	67	0	0	0	0.39	-0.03	0	-0.02	0.5
14	5	0	0	0	0.14	-0.27	0	-0.04	0.5
15	12	2	2	1	0.06	-1.05	0	-0.19	0
16	5	4	4	1	0.14	-0.77	0	-0.14	0
17	39	4	4	1	0.08	-2.18	0	-0.36	0
18	25	4	4	1	0.11	-0.98	0	-0.15	0
19	43	4	4	1	0.08	-1.26	0	-0.21	0
20	31	4	4	1	0.08	-1.75	0	-0.29	0

4. METHODOLOGY

This study will use data acquired from four sources: AVI sensor data, toll data, lane blockage data, and rain data. The desired dataset is extracted from the combined datasets to focus on paid ML trips from January 2012 to November 2012 and January 2013 to September 2014. In this selected dataset, each ML trip is recorded as an observation, and all trip characteristics including start and end sensor, traffic flows, total toll, and precipitation measurements are documented. An alternate GPL trip and its attributes are defined for each ML trip. Also, two types of classification for U-ML trips are represented in equations (8) and (10).

The first classification technique for examining U-ML trips is using variable $uneco_{binary}$, which divides ML trips into two classes of economical and uneconomical trips. The main benefit of implementing this variable is simplifying the problem into a straightforward interpretable binary model. The second classification uses $uneco_{multiclass}$, which divides the dataset into three groups: E-ML trips, U-ML trips, and middle ML trips. This variable classifies trips where the travel time on the GPLs and MLs are almost identical into a new group. In this study, both of these classifications will be applied to the dataset to find the best classification.

This study focused on discovering the consistencies among ML trip attributes and ML trip classifications. The first step is to recognize the related attributes that may lead to having a greater chance of U-ML trips. These attributes include time of the trip, route of the trip, rain, lane blockages, toll, traffic flow, travel behavior and trip

frequency. Consequently, several hypotheses are investigated to select the related variables associated with ML trip classifications:

1. Day of the trip: The probability of having a U-ML trip may be diverse over the week. As an example, the possibility of U-ML trips may increase on the weekends because of the lower congestion on GPLs.
2. Time of the trip: Peak hours may decrease the likelihood of U-ML trips as a consequence of congestion on GPLs.
3. Route of the trip: Specific direction or start and end sensors may result in an increased probability of having a U-ML trip.
4. Length of the trip: The travel time for short distances are small, and accordingly their travel time difference between GPLs and MLs is also small. Their variation because of the congestion and other factors can be relatively large. This travel time variation on MLs and GPLs may lead to a U-ML trip. Therefore, length can have an impact on the likelihood of U-ML trips.
5. Safety variables, including crash and rain: Drivers may believe MLs are safer. Consequently, they would pay for that. Severe accidents on GPLs and weather conditions can cause drivers to pick safer lanes, assumedly MLs, no matter how much they have to pay, or how long it would take them.
6. Toll: A low toll paid per mile might be an incentive for drivers to select MLs over GPLs in particular circumstances. Thus, MLs may get congested while GPLs are faster.

7. Traffic flow: the traffic flows on the GPLs and MLs are the key features in examining TT_{ML} and TT_{GPL}. Higher traffic flow on MLs results in a greater possibility of having a U-ML trip. Similarly, lower traffic flow on GPLs leads to a higher likelihood of U-ML trips.
8. Travel behavior: As previously observed by Yang (1993) and Reddy (1995), drivers' experiences may have an impact on their future decisions, even greater than advanced trip information. They may be content and used to a route, and keep it the same. This fact can also be useful in the lane choice study, especially ML study. Therefore, examining the trip history and finding the travel behavior pattern may help to conclude that U-ML trips are the outcome of a daily routine.
9. Trip frequency: The number of times drivers use MLs or GPLs or Katy Freeway may add the concept of familiarity to the analysis. Part of U-ML trips may be as the consequence of unfamiliarity.

After defining the variables of interest, it is beneficial to create a variables' correlation table. It helps to discover how variables are correlated and if they are redundant or repetitive. At that time, a set of unique independent variables will be identified to be investigated by random forest and logistic regression methods.

Before studying the variables, sampling and resampling steps will be undertaken. First, a random sample of the main dataset will be selected and split into two groups of the training set (80%) and test set (20%). The training set will be our core dataset for estimating the models. This dataset will be resampled to create the balanced datasets

using undersampling technique and SMOTE. Nevertheless, the test set will always stay the same unseen set for testing the predictive power of the model.

Resampling is based on the dependent variables to be predicted. These variables are $uneco_{binary}$ in binary classification and $uneco_{multiclass}$ in multiclass classification. In other words, the size of classes is required to be resized for balancing. Also, both of the dependent variables show U-ML trips are a minor class with 5% to 11% of total trips. Therefore, they need to be resampled from the training set:

1. Analysis of binary classification: for this part of the analysis, three training sets will be prepared to be examined. The first set is the imbalanced training set, which is the same as the training split set. It is an imbalanced data because of the small proportion of U-ML trips (11% of the total trips). The models based on this sample are expected to be bias to the major class of E-ML trips. This is why two other training samples are developed. The second training set is obtained by undersampling the imbalanced training set. Hence, the percentage of minor and major classes are 50-50. The third approach is to use SMOTE on the imbalanced training set. This method makes the training set more balanced by applying oversampling technique using K-nearest neighbors practice to generate new observations. However, the final U-ML trip percentage of the total trips may not be exactly 50%.
2. Analysis of multiclass classification: the main training set is used as the imbalanced dataset in this analysis. The percentage of U-ML trips is 5% in this classification. Therefore, the dataset is extremely imbalanced. To resample the data, the undersampling method will be utilized to obtain three equally sized classes.

Random forest and logistic regression methods are developed for each training set in binary classification. The random forest can indicate the key classification factors, and logistic regression model shows each factor's impact on the U-ML trip likelihood. For multiclass classification, the random forest will be designed for both imbalanced and undersampled training sets. The final step is to apply all models on the test set to find the best classification, resampling, and pattern recognition techniques.

To evaluate the test models, they can be compared based on their accuracy, sensitivity, specificity, and ROC curve or AUC. Using this evidence, the final models will be constructed using the most significant independent factors and best resampling technique. The final models will include both random forest and logistic regression methods, and show the most efficient and easy to use models.

5. DATA ANALYSIS

In this chapter, the main analysis will be conducted in details. This began with a preliminary data analysis and continued to sampling and resampling procedures, random forest and logistic regression analyses, and lastly repeating the analysis with a new set of key variables.

5.1. Preliminary Data Analysis

This section is performing some exploratory analysis of the data to provide an introduction to the dataset.

The initial step is to investigate variables of study focusing on their averages and distributions. The total number of paid ML trips is 7,013,578 from January 2012 to November 2012 and January 2013 to September 2014. These trips are classified using either TTD or RTTD and as formulated in Equation (8) and (10).

5.1.1. Actual ML and Alternate GPL Travel Time

TT_{ML} and TT_{GPL} are the chief components of this study. These variables compute the travel time saving or loss on MLs and help to define U-ML trips.

TT_{ML} is the actual ML travel time, which is so diverse from less than 1 minute to over 35 minutes. The average TT_{ML} is 9.6 minutes. The cumulative TT_{ML} density plot is presented in Figure 5. TT_{GPL} is defined as the alternate GPL travel time, and ranges from 1 minute to over 45 minutes. The average value for TT_{GPL} is 12.2 minutes, which is greater than the average TT_{ML}. This means MLs have a lower travel time on average. The cumulative TT_{GPL} density plot is presented in Figure 6.

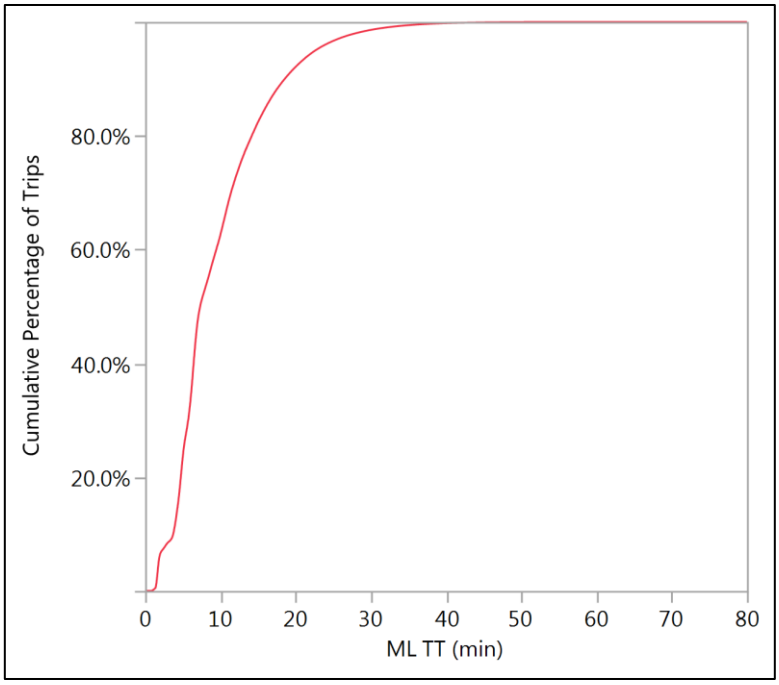

Figure 5 Actual ML Travel Time, TT_{ML} (min)

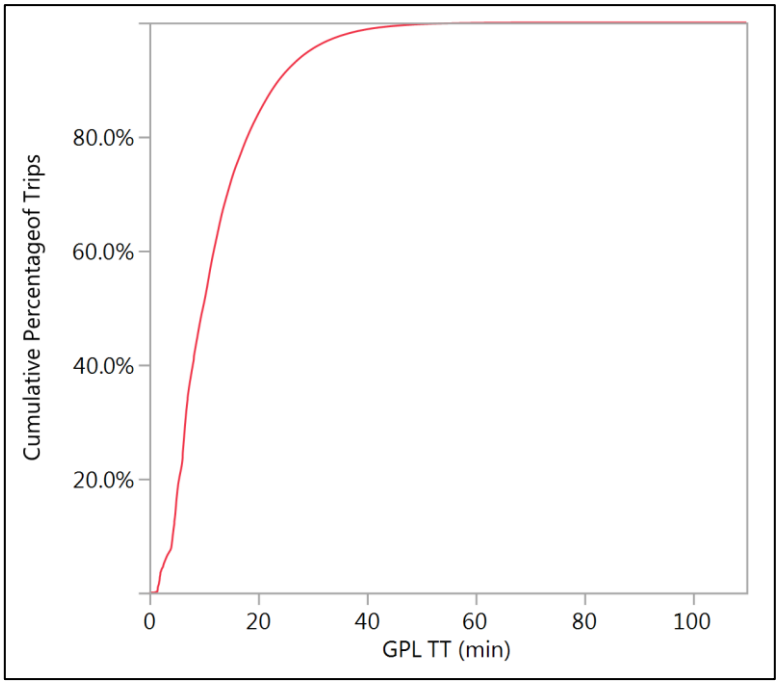

Figure 6 Alternate GPL Travel Time, TT_{GPL} (min)

5.1.2. TTD and RTTD

TTD is the travel time difference between actual ML and alternate GPL trips, which divides trips into two classes of negative and positive TTD. That is, positive TTD shows a U-ML trip. As shown in Figure 7, TTD ranges from less than -20 minutes to almost 3 minutes, and the average TTD is -2.6 minutes. Nearly 11% of paid ML trips have a positive TTD. In other words, using $uneco_{binary}$ variable to classify trips results in 786,448 U-ML trips (11% of total paid ML trips). Figure 9 shows ML trip divisions.

RTTD is the relative travel time difference between actual ML and alternate GPL trips classifying trips into three clusters of RTTD less than -0.05, RTTD more than 0.05, and RTTD in the middle. RTTD more than 0.05 shows a U-ML trip, however, RTTD between -0.05 and 0.05 is an average group which has a small travel time saving or loss. Consequently, it may not be fair to combine this group with the other two groups. As shown in Figure 8, RTTD ranges from less than -2 to near 2, and the mean RTTD is -0.3. $Uneco_{multiclass}$ variable is implemented to categorize trips based on RTTD intervals. This leads to 360,159 U-ML trips (5% of total paid ML trips). Figure 9 shows ML trip classification.

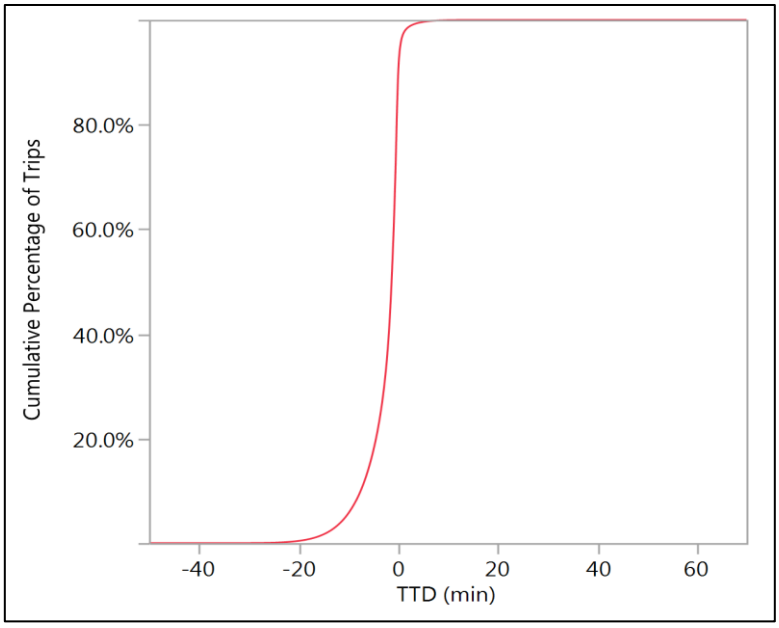

Figure 7 Actual and Alternate Travel Time Difference (min)

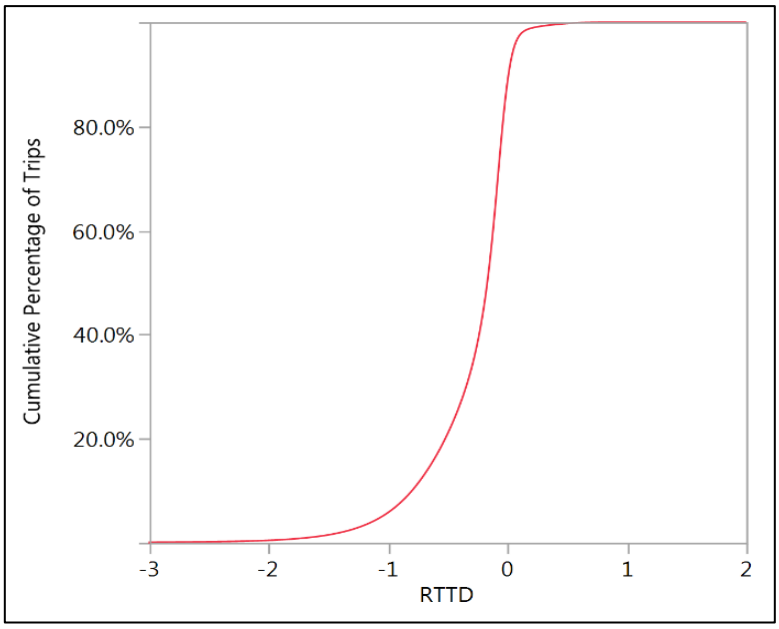

Figure 8 Relative Travel Time Difference between Actual and Alternate Trips

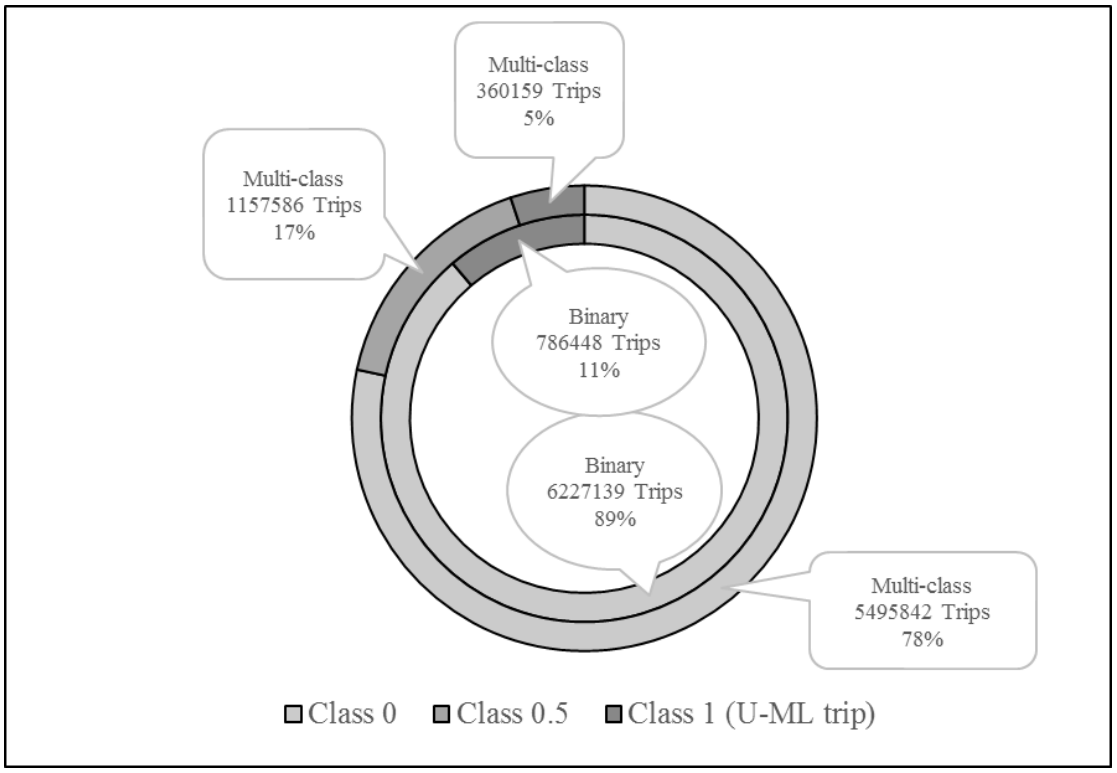

Figure 9 Binary and Multiclass Classes

As illustrated, 11% of paid ML trips are classified as U-ML trip in binary classification. This percentage is 5% in multiclass classification. In other words, 6% of paid ML trips have a small travel time saving (positive TTD) and can be categorized into a new middle group to study. Also, these small percentages show dataset is imbalanced with the major class of positive travel time saving (negative TTD, E-ML trips). Therefore, the imbalanced dataset techniques of resampling should be applied to design better models.

5.1.3. Time of Trip

Time of the trip can be measured using several variables. Three of the main elements of time of the trip can be marked as day of the week, peak hour, and shoulder hour.

One theory is that the U-ML trip rate may vary by the day of the week. Figures 10 and 11 show the distribution of binary and multiclass ML trip classes over the week. One key point in both types of classification is that the percentage of E-ML trips decreases over the weekends. This fact is more significant in multiclass ML trips with almost 8-14% drop of E-ML trips during the weekends. This fact is as the result of lower congestion during the weekend. As long as vehicles with transponders are a representative sample of the all vehicles using Katy Freeway, Figure 12 shows that the traffic flow is decreasing during the weekend.

The other variable conveying the time of the trip is the peak hour. The peak hour variable can be classified into three clusters of the non-peak hour, first peak hour, and second peak hour. Figures 13 and 14 exhibit the binary and multiclass ML trip distribution for three peak hour classes. It can be concluded that the percentage of U-ML trips increases during the non-peak hours. This factor is more outstanding in multiclass ML trip distribution with almost 13% drop in E-ML trips during the non-peak hours. This fact is also the consequence of lower traffic flow during non-peak hours, which may cause GPLs become faster. Nevertheless, some drivers may still choose MLs over GPLs and have a U-ML trip.

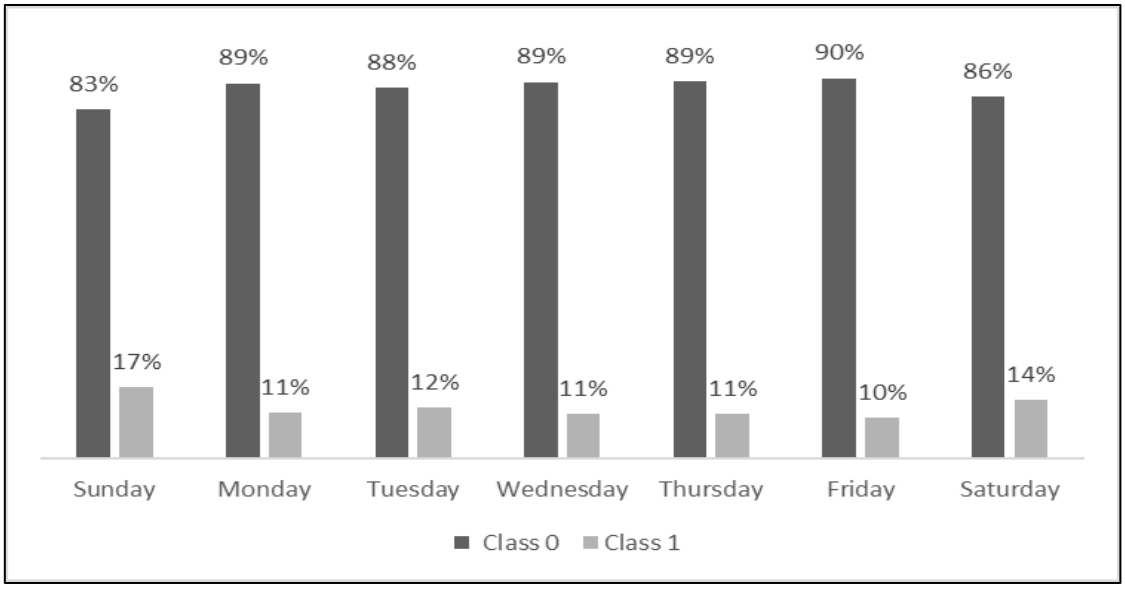

Figure 10 Binary ML Trip Classes Distribution for Weekdays

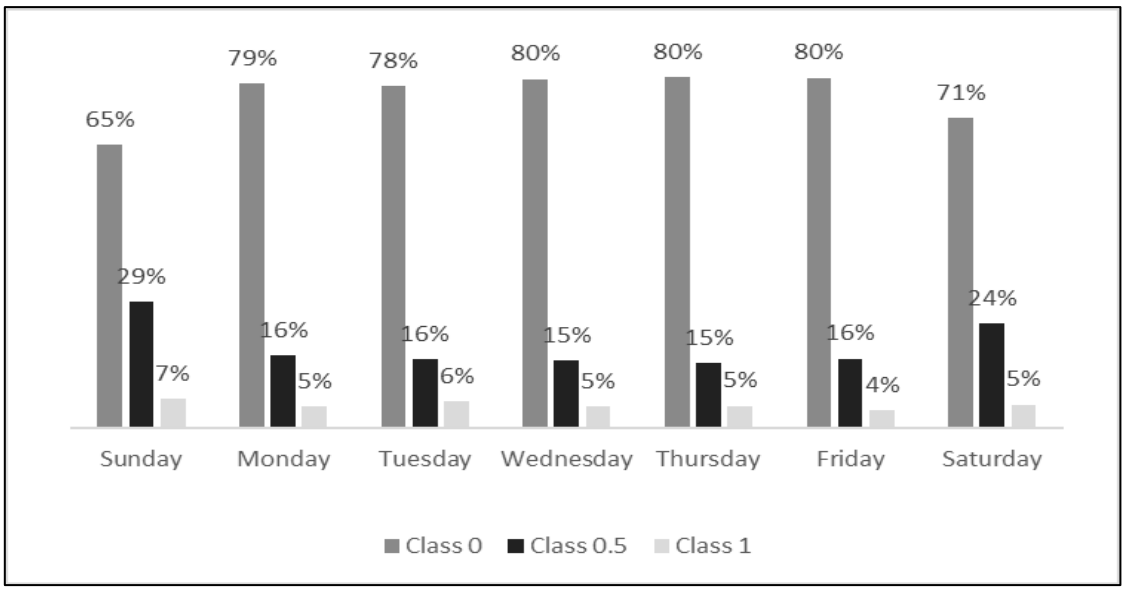

Figure 11 Multiclass ML Trip Distributions for Weekdays

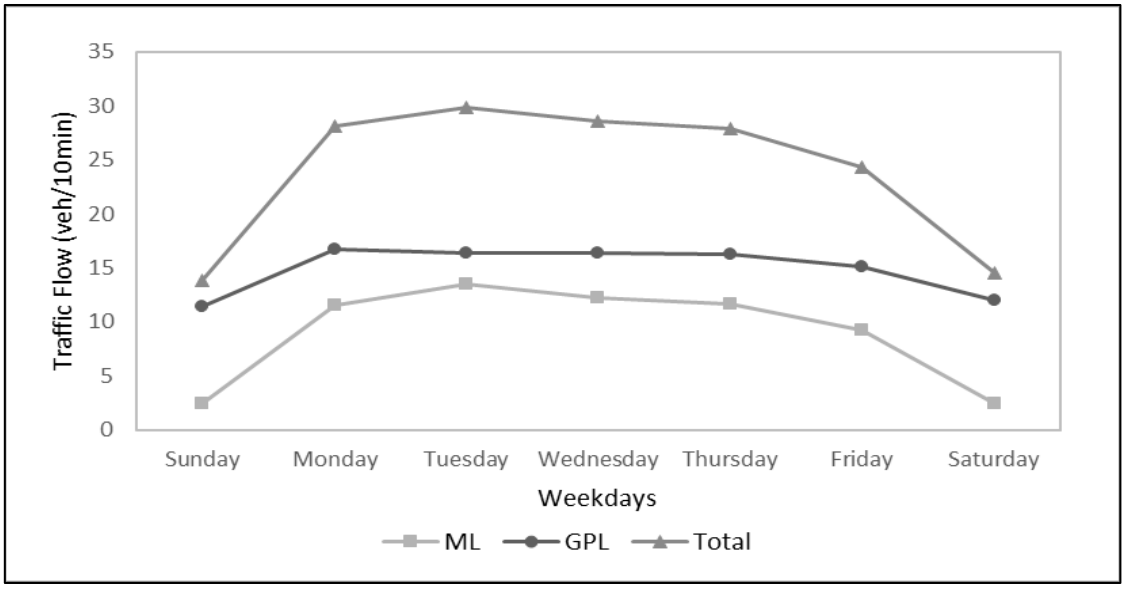

Figure 12 GPL and ML Traffic Flow over the Week

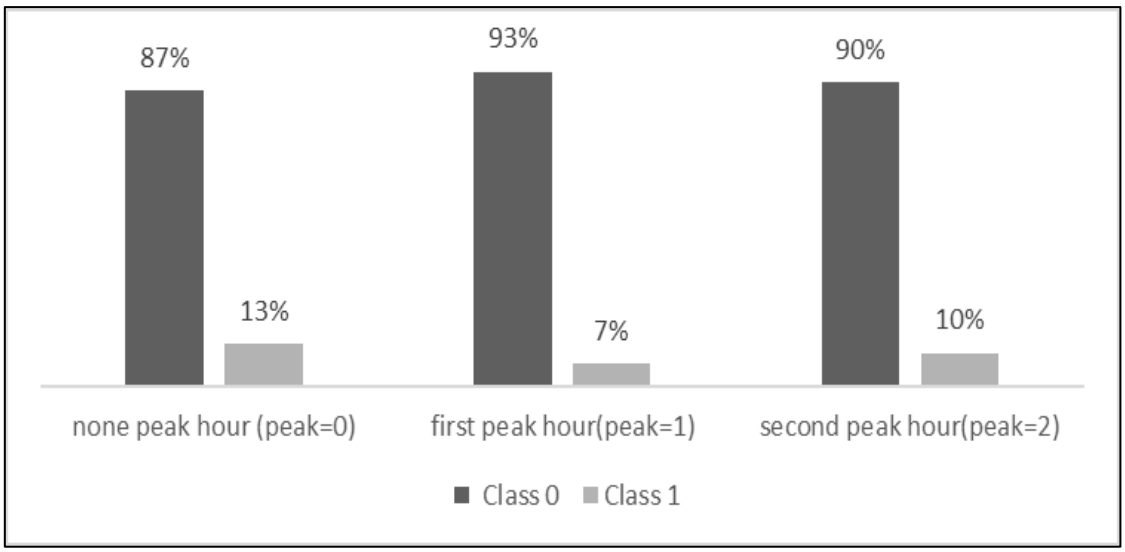

Figure 13 Binary ML Trip Distributions for Peak Hours

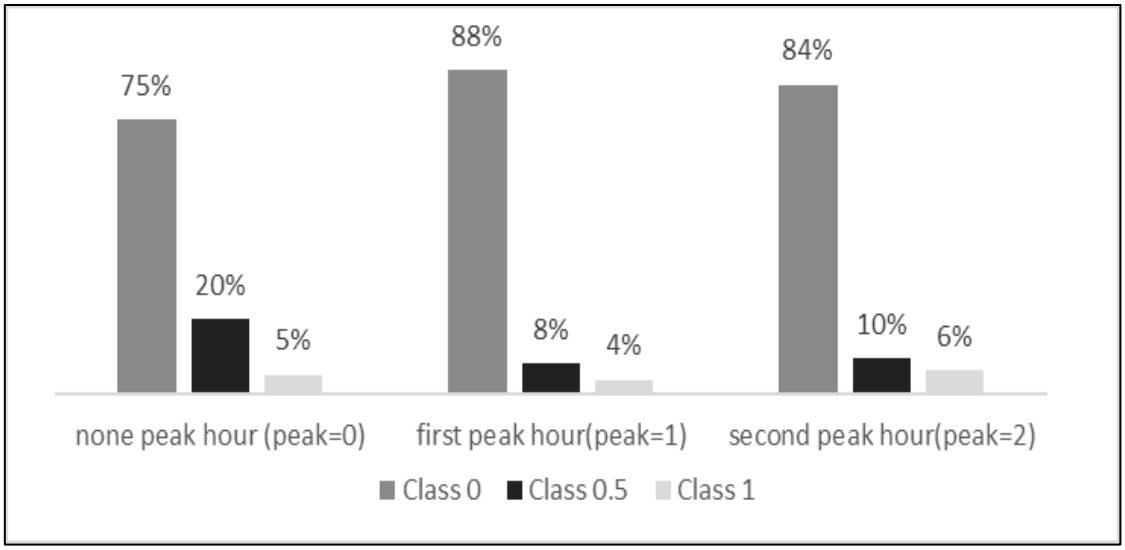

Figure 14 Multiclass ML Trip Distributions for Peak Hours

The third component of time of the trip is the shoulder variable. This factor classifies the non-peak hours into three classes of the non-shoulder hour, the shoulder hour before the peak hour, and the shoulder hour after the peak hour. Figures 15 and 16 exhibit the binary and multiclass ML trip distribution for different classes of shoulder hour factor. Similar to the peak hour, this factor also shows a drop in E-ML trips during the non-shoulder hours. This fact, which is more severe in binary classification, is again because of the lower traffic flow during the non-shoulder hours.

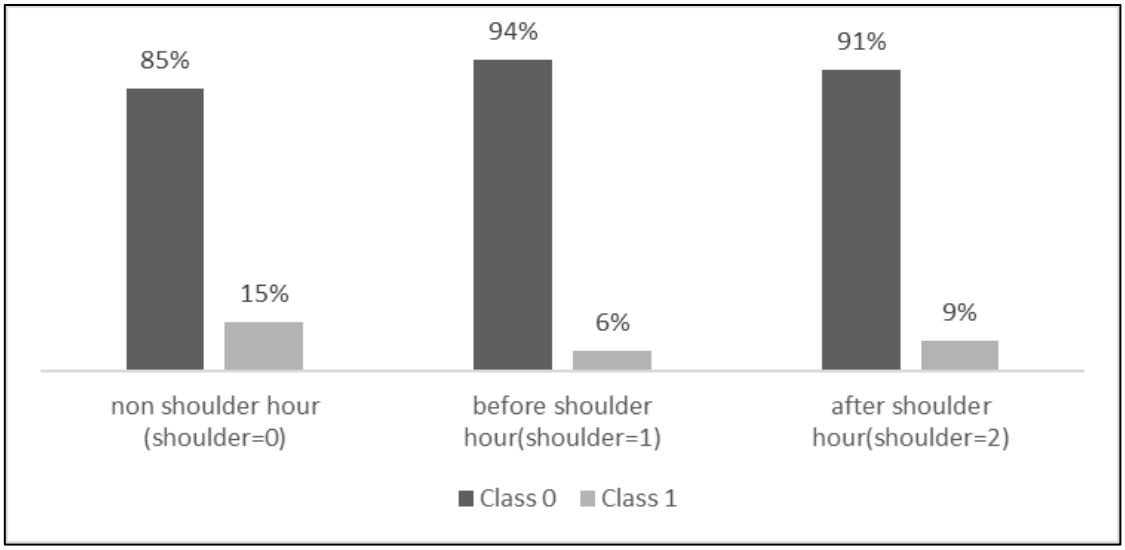

Figure 15 Binary ML Trip Distribution for Shoulder Hours

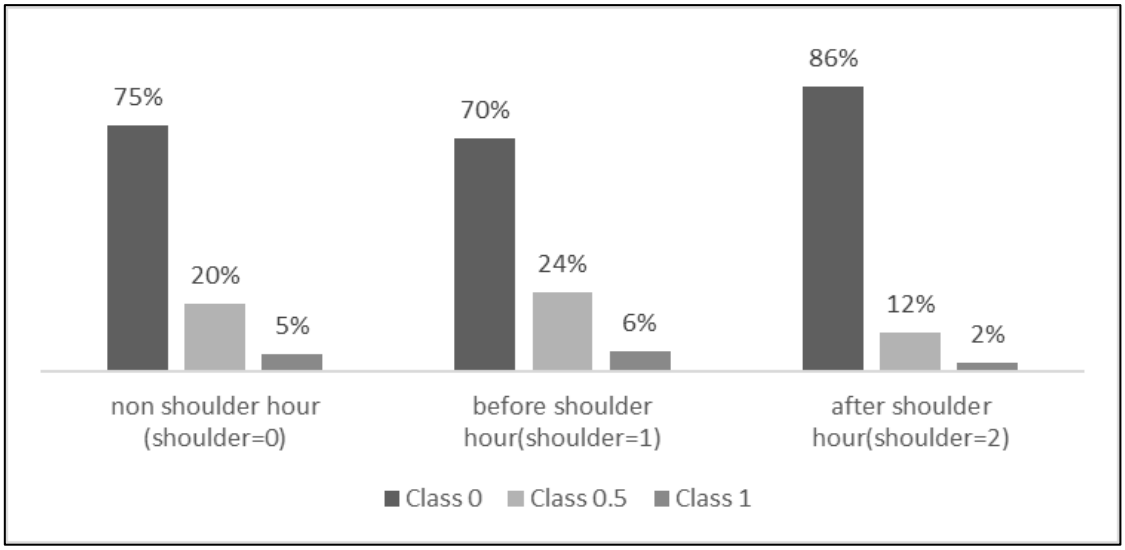

Figure 16 Multiclass ML Trip Distribution for Shoulder Hours

5.1.4. Trip Route

Route of the trip as well as the time of the trip can be characterized by several elements. Two key features are start sensor and end sensor which indicate the start and

end of the trip on Katy Freeway. As shown in Figures 17 and 18 for start sensors and Figures 19 and 20 for end sensors, the proportion of U-ML trips are wide-ranging among different start and end sensors. The location of these sensors and traffic flow at these sensors appears to impact U-ML trips.

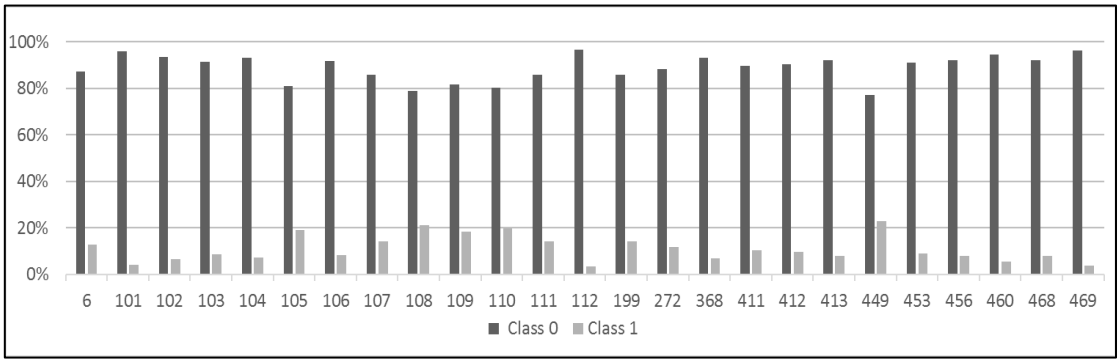

Figure 17 Binary ML Trip Distribution for Start Sensors

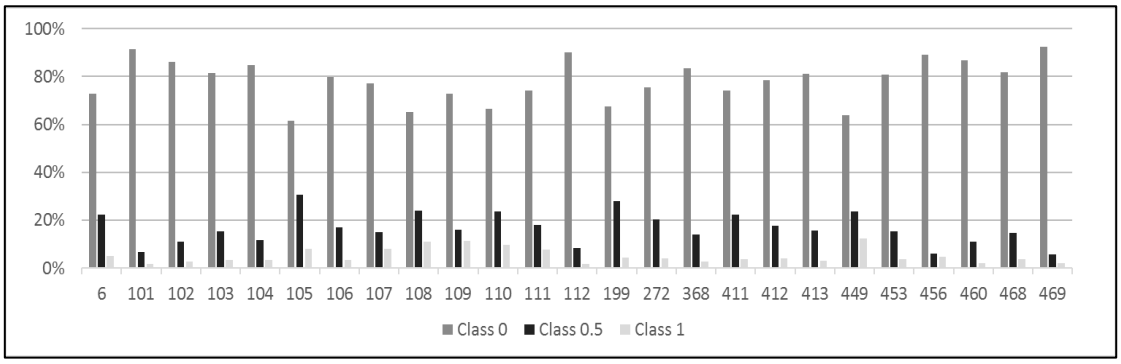

Figure 18 Multiclass ML Trip Distribution for Start Sensors

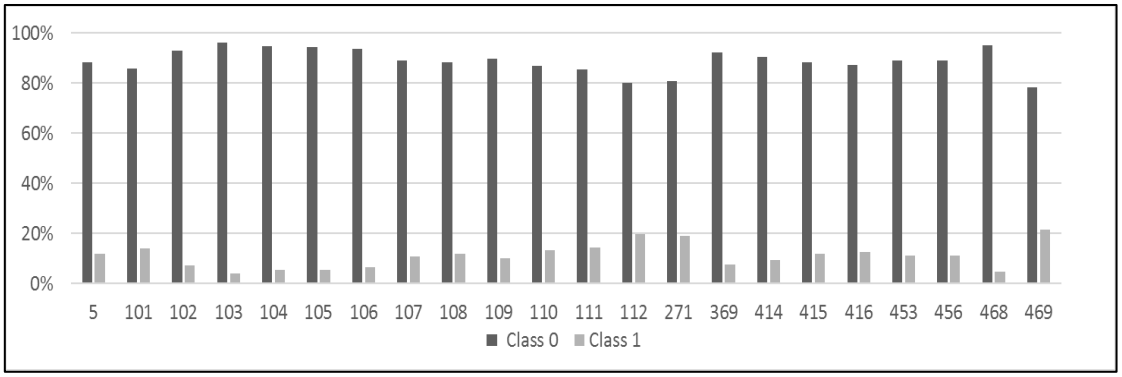

Figure 19 Binary ML Trip Distribution for End Sensors

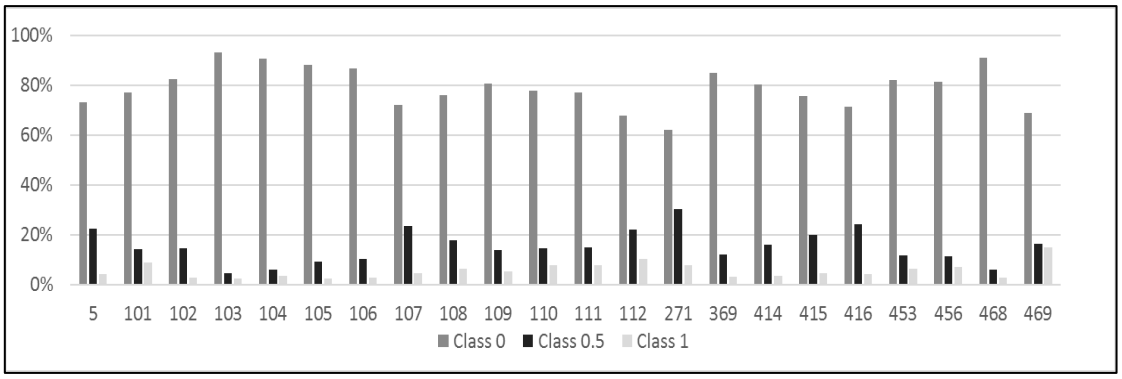

Figure 20 Multiclass ML Trip Distribution for End Sensors

The trip length is another feature affecting the U-ML trip ratio. Table 5 exhibits the average trip length for various classes of ML trips. Binary ML trip classification shows no significant variation between classes. However, the average trip length for multiclass ML trips does vary from 7.96 miles to 9.84 miles. It can be concluded that U-ML trips are more probable to occur over shorter distances compared to E-ML trips.

Table 5 Average Trip Length for Binary and Multiclass Classes

Class	Binary		Multiclass		
	E-ML	U-ML	E-ML	Middle ML	U-ML
Trip length (mi)	8.96	8.90	8.84	9.84	7.96

5.1.5. Rain and Blockages

Rain and blockages may cause unequal delays in travel time of MLs and GPLs. It may also result in some rerouting behavior, which causes drivers to change their routines. Table 6 shows the average rain measurement and blockages of each ML trip class.

Table 6 Average Rain and Blockages of Binary and Multiclass Classes

Class	Binary		Multiclass		
	E-ML	U-ML	E-ML	Middle ML	U-ML
Main lanes blockage	0.0021	0.0043	0.0017	0.0028	0.0055
Frontage lanes blockage	0.0003	0.0003	0.0003	0.0004	0.0003
Ramp lanes blockage	0.0002	0.0003	0.0002	0.0002	0.0003
HOV lanes blockage	0.0000	0.0001	0.0000	0.0001	0.0001
Shoulder lanes blockage	0.0012	0.0020	0.0010	0.0016	0.0023
Rain (in)	0.1134	0.1089	0.1139	0.1133	0.1061

As indicated, only main lanes blockage and shoulder lanes blockage are largely varying among the various ML trip classes; they may increase up to three times. U-ML trips have a greater average of main lanes blockage and shoulder lanes blockage than E-ML trips. This is as the result of rerouting behavior caused by the blockage. Drivers

reroute to MLs to pass a blockage. However, the GPL congestion is for a part of trip length, and it gets faster after the blockage. The average of precipitation is also changing between ML trip classes. U-ML trips have a lower average of precipitation than E-ML trips, which means U-ML trips are more likely during unrainy conditions.

5.1.6. Toll

Toll can be either total toll or toll rate per mile. Table 7 exhibits the toll factors for ML trip classes. It shows that the average total toll and toll rate for U-ML trips are less than for E-ML trips in the binary classification. It can be deduced that when the toll rate is lower (during the non-peak hour), U-ML trips are more likely to happen. However, multiclass classification indicates E-ML and U-ML trips have the same average toll rate. The middle class has a much lower toll rate compared to the other two classes, meaning drivers paying the smallest toll rate have travel times very similar to GPLs.

Table 7 Average Toll Factors for ML Trip Classes

Class	Binary		Multiclass		
	E-ML	U-ML	E-ML	Middle ML	U-ML
Total toll	2.13	1.56	2.25	1.32	1.79
Toll rate	0.33	0.25	0.35	0.17	0.33

5.1.7. Traffic Flow

One principal element of this U-ML trip study is ML and GPL traffic flow. Fewer number of vehicles with transponders on the GPLs may lead to higher U-ML trips for drivers selecting MLs over GPLs. Conversely, congestion on MLs often results in a

U-ML trip for drivers on the MLs. In other words, U-ML trips may occur when a higher rate of vehicles with transponders using MLs comparing to E-ML trips. Table 8 shows the mean ML and GPL traffic flow for different ML trip classes.

Table 8 Average ML and GPL Traffic Flow for ML Trip Classes

Class	Binary		Multiclass		
	E-ML	U-ML	E-ML	Middle ML	U-ML
ML traffic flow (veh/10 min)	10.62	12.70	10.98	8.17	17.39
GPL traffic flow (veh/10 min)	16.20	12.14	16.48	13.67	11.31
Total Traffic flow (veh/10 min)	26.82	24.84	27.46	21.84	28.7
ML usage rate	0.40	0.51	0.40	0.37	0.61

5.1.8. Travel Behavior and Trip Frequency

Travel behavior implies that drivers may select MLs habitually. Table 9 shows the average ML trip frequency and the percent ML trips for the ML trip classes. The variation among classes is not significant in binary classification. Also, multiclass classification indicates a 4% increase in trip frequency from E-ML trip to U-ML trip. Therefore, travel behavior is not expected to be among the most important variables in predicting U-ML trips. However, it will still be considered in the initial stage of analysis.

Table 9 Average Travel Behavior for ML Trip Classes

Class	Binary		Multiclass		
	E-ML	U-ML	E-ML	Middle ML	U-ML
ML Trip Frequency	6.84	6.87	6.93	6.27	7.24
Percent ML Trips	0.42	0.40	0.42	0.39	0.41

5.2. Sampling and Data Partitioning

In this section, the first sample of the dataset will be derived from the main dataset. Later, the training and test sets will be formed for the further analysis.

To get the sample set, 1,001,941 trips (1 trip each seven trips) are randomly selected from all trips. This sample will be the main dataset to be examined in this study.

Then, data were split into two sets to train the model and test the predictive power of the model. The training set, which is used to train the model and find the best fit for the model, includes 801,554 trips (or 80% of sample trips). Figure 21 shows the training set classification. The number of U-ML trips in training set is 40,906 based on multiclass classification and increases to 89,747 trips in binary classification. Each class percentage in training set is the same as the main dataset, which shows the training set is a good representative of the main dataset.

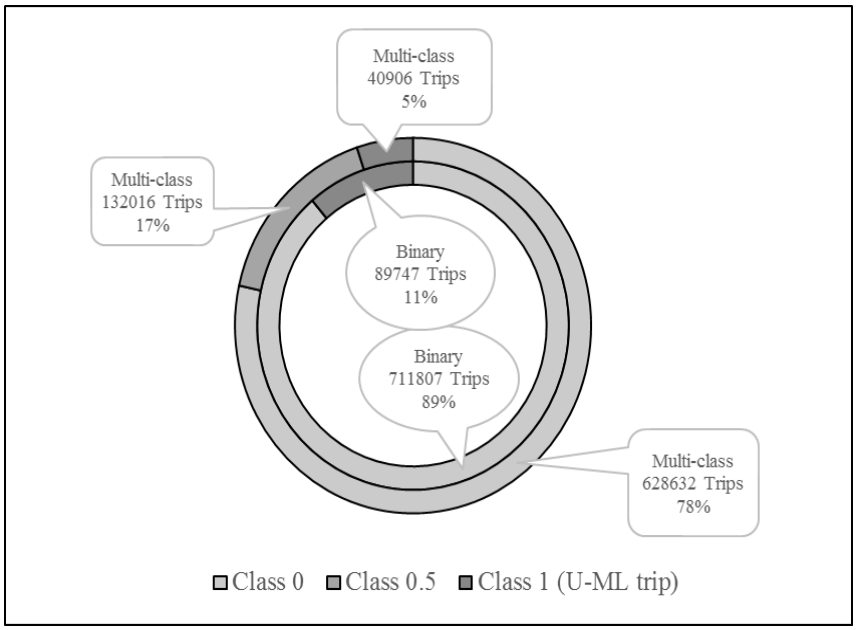

Figure 21 ML Trip Classifications

The second subset of data includes 200,387 trips or 20% of the sample trips. The test set is used to estimate the predictive powers and the accuracy of the model on a new dataset not used in model development.

5.3. Resampling

There is a small percentage (5%-11%) of U-ML trips in the training set (see Figure 21). Therefore, any models fitted to the data will be bias to the major class of trips, or E-ML trips. However, the focus of this study is to examine the causes of U-ML trips. Therefore, the data needs to be balanced and resampled.

Resampling of an imbalanced dataset is focused on the dependent variable to be predicted by the model. Therefore, the binary classification and multiclass classification are separated at this stage of the study, and the resampled datasets are generated for each of them separately.

5.3.1. Binary Classification

The techniques implemented to resample the training set for the binary classification analysis are undersampling and SMOTE. These two approaches generate new balanced datasets, named "undersampled" and "SMOTEd" datasets, using the variable $uneco_{binary}$. Table 10 shows the sizes of new developed training sets using each technique and the percentage of their balanced classes.

Table 10 Datasets Used in Binary Analysis

Class	U-ML	E-ML	Training set Size
Imbalanced Dataset	11.20%	88.80%	801,554
Undersampled Dataset	50.05%	49.95%	179,320
SMOTEd Dataset	42.86%	57.14%	628,229

5.3.2. Multiclass Classification

To resample the training set for multiclass analysis, undersampling method is implemented. This technique randomly decreases the size of E-ML trip and middle ML trip classes of variable $uneco_{multiclass}$ to be equal to U-ML trip class. The new training set size is named "undersampled dataset" and indicated in Table 11.

Table 11 Datasets Used in Multiclass Analysis

Class	U-ML	Middle ML	E-ML	Training set Size
Imbalanced Dataset	5.10%	16.47%	78.43%	801,554
Undersampled Dataset	33.33%	33.33%	33.33%	122,718

5.4. Variable Correlation

The variables to be included in the model should not be highly correlated. In other words, predicting variables ought to be independent unique variables. In this section, pearson's correlation coefficient is computed for all numerical variables within the dataset (see Table 12). Referring to this table, concerns regarding correlation are limited to the following five pairs:

1. Shoulder and main lane blockages: These two variables are related (ρ = 0.486) and it is logical. The shoulder lanes may get impassable as an effect of main lane blockage or accident.
2. Total toll and toll rate: Pearson's correlation is calculated 0.561 for these two factors since the toll rate is described as the total toll per mile of trip length. Therefore, only one of them can be included in the model formulation. In this study, the toll rate is the one to be considered in the model. It is worth noting that the toll rate and the trip length also have ρ = -0.360, which means they are moderately correlated. However, this correlation is not too significant to exclude one of them at the initial stage.
3. GPL traffic flow and trip length: If the trip length is short, and the GPLs are congested, drivers select MLs. However, they will choose MLs in spite of the low GPL traffic flow for longer trips. In other words, the GPL traffic flow for short ML trips may be higher than long ML trips. This association has formed the correlation factor of -0.457 between GPL traffic flow and trip length.
4. Toll rate and GPL traffic flow: Toll rate is increasing during the peak hours. Also, the number of vehicles passing all lanes are also increasing. Though, ML traffic flow

may increase to the capacity point, and after that other vehicles decide not to enter the MLs. Thus, GPL traffic flow is constantly increasing as the result of peak hour. This correlation between toll rate and GPL traffic flow will result in a factor of 0.437.

5. Percent ML trips and trip frequency: Travel behavior factors are correlated by ρ = 0.595. This correlation states that drivers who use Katy MLs frequently are more probable to select MLs over GPLs.

Despite these correlations, many of the variables were included in initial models to determine the superior variable. In subsequent models at least one, if not both, of the correlated variables were removed. Therefore, the binary and multiclass classification of ML trips is considered as the function of variables in Equation (12) at the initial stage.

Table 12 Variable Correlation (Pearson's Correlation Coefficient)

	Main lane blockage	Frontage lanes blockage	Ramp lanes blockage	HOV lanes blockage	Shoulder lanes blockage	Rain	Total toll
Main lanes blockage	1.000						
Frontage lanes blockage	0.000	1.000					
Ramp lanes blockage	0.086	0.115	1.000				
HOV lanes blockage	0.006	0.000	0.000	1.000			
Shoulder lanes blockage	0.486	0.000	0.037	0.005	1.000		
Rain	-0.002	0.005	0.011	-0.002	0.001	1.000	
Total toll	-0.009	-0.005	-0.006	-0.005	-0.005	0.000	1.000
ML traffic flow	-0.003	-0.004	-0.003	-0.003	-0.003	-0.021	0.260
Length	0.013	0.013	0.006	-0.003	0.011	-0.004	0.192
Travel time variability	0.025	0.003	0.005	0.001	0.015	-0.007	0.019
GPL traffic flow	-0.006	-0.003	-0.003	0.002	-0.003	0.000	-0.053
ML trip frequency	-0.005	-0.004	-0.004	0.001	-0.004	-0.003	0.134
Percent ML trip	-0.007	-0.005	-0.006	-0.001	-0.005	-0.007	0.124
Toll rate	-0.015	-0.007	-0.007	-0.003	-0.013	0.003	0.561

Table 12 Continued

	ML traffic flow	Length	Travel time variability	GPL traffic flow	ML trip frequency	Percent ML trips	Toll rate
Main lanes blockage							
Frontage lanes blockage							
Ramp lanes blockage							
HOV lanes blockage							
Shoulder lanes blockage							
Rain							
Total toll							
ML traffic flow	1.000						
Length	-0.212	1.000					
Travel time variability	-0.001	-0.006	1.000				
GPL traffic flow	-0.016	-0.457	0.091	1.000			
ML trip frequency	0.079	-0.027	-0.028	0.004	1.000		
Percent ML trip	0.075	-0.066	-0.037	0.034	0.595	1.000	
Toll rate	0.243	-0.360	0.061	0.437	0.083	0.107	1.000

$$uneco(binary, multiclass) = F(direction,\ weekday,\ peak,\ shoulder,\ main\ lanes\ blockage,\ frontage\ lanes\ blockage,\ ramp\ lanes\ blockage,\ hov\ lanes\ blockage,\ shoulder\ lanes\ blockage, rain,\ ML\ traffic\ flow,\ length,\ start\ sensor,\ end\ sensor, travel\ time\ variability,\ GPL\ traffic\ flow,\ ML\ trip\ frequnecy,\ Trip\ habit,\ Toll\ Rate) \quad (12)$$

5.5. Initial Binary Random Forest (BRF) Model

In this section, random forests are created for three training sets to discover the best training set and the best resampling technique. The number of trees developed in random forest method is selected to be 100 for the initial analysis.

5.5.1. Imbalanced Data

The training set, also known as the imbalanced dataset, is used to design a random forest for predicting the binary ML trip classification. This dataset consists of 801,554 observations with 11% of total paid ML trips as U-ML trips, which is shown in Table 10.

The random forest errors' plot shows the increase in number of trees have reduced the overall out-of-bag (OOB) error. However, the error of U-ML trip class increases when number of trees increases and reaches a constant number finally. This is the consequence of the imbalanced data. In other words, adding more trees to the random forest makes the models become more bias to the major (E-ML trip) class. All types of error reach a constant value after a certain number of trees, which indicates a certain number of trees are adequate for including all types of dataset variations in the analysis (see Figure 22).

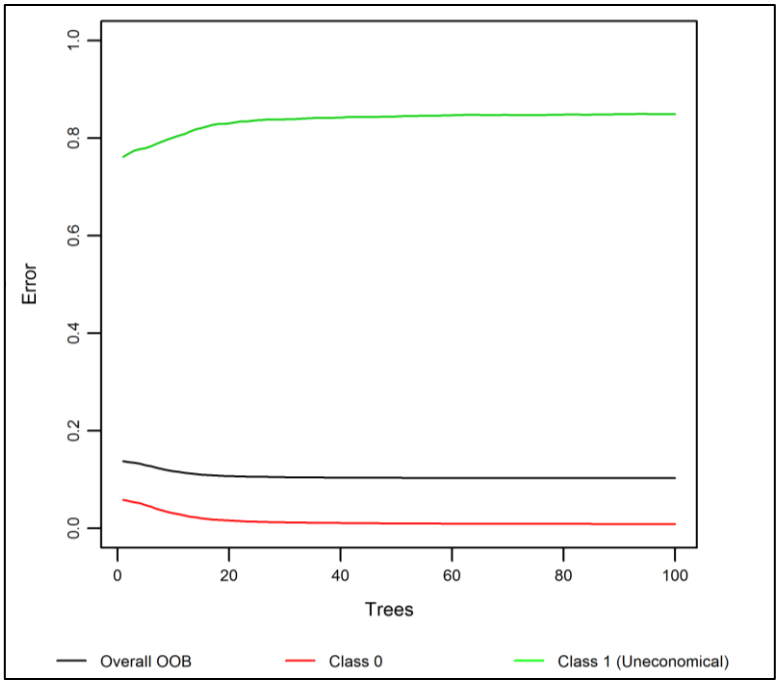

Figure 22 Errors Plot- Initial BRF for the Imbalanced Dataset

To test the model, the model is applied for predicting the test set's ML trip classification. To evaluate the model, a confusion matrix and the Receiver Operating Characteristic (ROC) curve were developed. As shown in Table 13, the accuracy of this prediction is 0.8975, which is the "paradox of accuracy" in this case. The model sensitivity for the test set is so high (0.992), despite the specificity is too low (0.150). In other words, the model predicts almost all of the U-ML trips as E-ML trips, causing the accuracy of the model to get close to the major class proportion (89%). This is why the model is not well-constructed and too bias to the major ML trip class. As indicated in Table 13, the AUC is a better goodness of fit measure than accuracy, and is close to 0.5, which shows the model has the probability of almost half to classify a positive value accurately.

Figure 23 also shows that the ROC curve is too close to the half line, which indicates the deficiency of this model. This model is not well-trained to examine its most important variables as the training set is imbalanced. Therefore, this training set is not good enough to build the final model.

Table 13 Model Specifications- Initial BRF for the Imbalanced Dataset

Accuracy	Sensitivity	Specificity	AUC
0.898	0.992	0.150	0.571

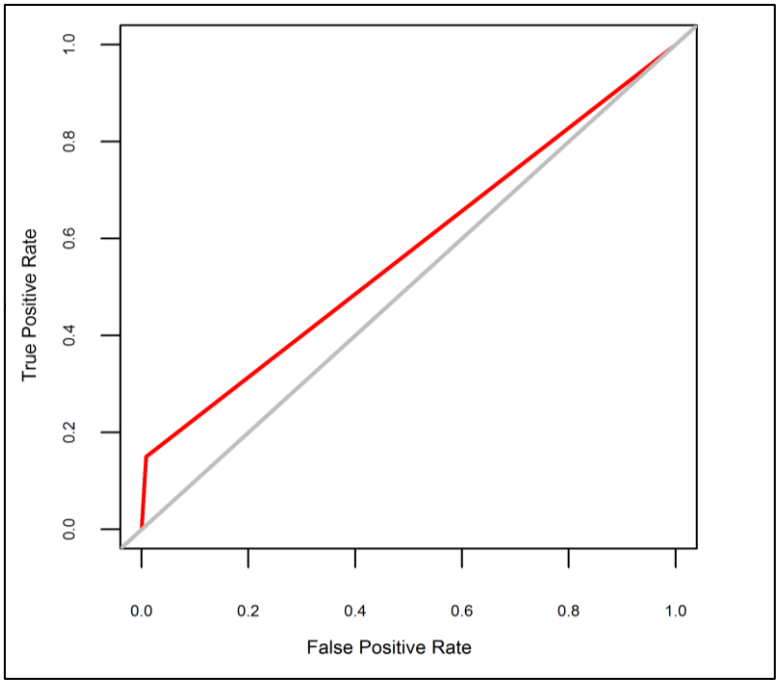

Figure 23 ROC Curve- Initial BRF for the Imbalanced Dataset

5.5.2. Undersampled Dataset

This dataset is generated by making the two classes of $uneco_{binary}$ variable equally sized. In other words, the major E-ML trip class is downsized to be equal to the number of U-ML trips in the training set, a total of 179,320 paid ML trips. The percentage of each ML trip class is almost 50 percent, and the dataset is well-balanced (see Table 10).

The random forest errors' plot displays that the all types of errors decrease by the increase in the number of trees, which demonstrates that the dataset is balanced and well-constructed. Also, the overall OOB error reaches 0.25, and the number of trees is adequate for the model (see Figure 24). The E-ML trip class has a higher error than U-ML trip class, which means the model does not predict economical class as well as uneconomical class. In other words, the test model's specificity is greater than its sensitivity. This is documented in Table 14.

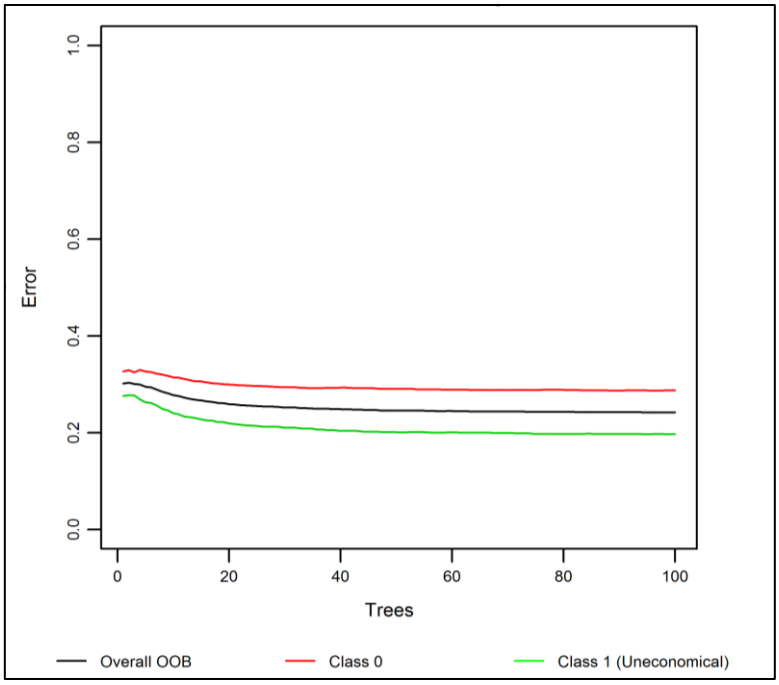

Figure 24 Errors Plot- Initial BRF for the Undersampled Dataset

Table 14 Model Specifications- Initial BRF for the Undersampled Dataset

Accuracy	Sensitivity	Specificity	AUC
0.725	0.714	0.807	0.761

As expressed in Table 14, the accuracy, specificity, and sensitivity of the test model are 0.72, 0.71, and 0.81 relatively. These three parameters show the model is well-fitted, and all of them are reasonably high. AUC, which is a better measure of goodness of fit, is 0.761. Correspondingly, Figure 25 displays how the ROC curve is covering a considerable more area than half of the square.

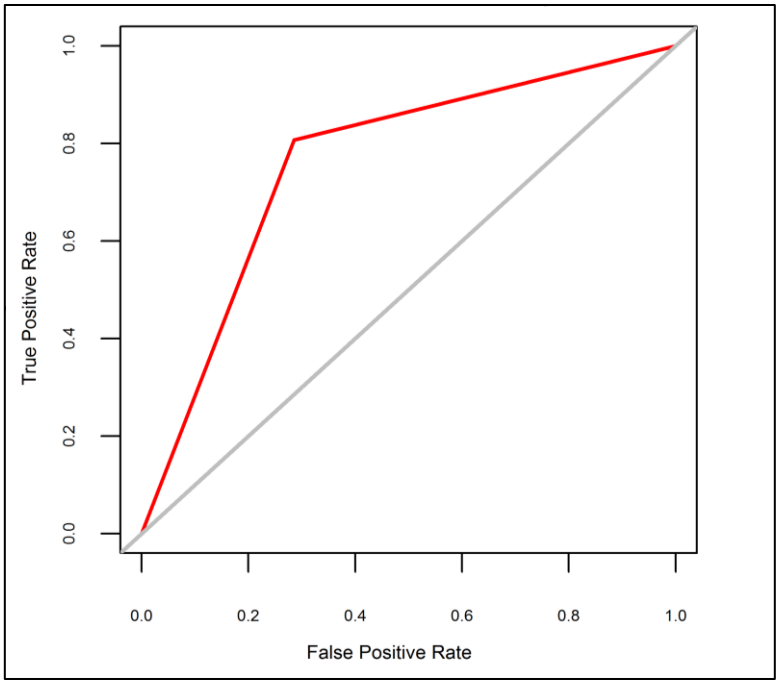

Figure 25 ROC Curve- Initial BRF for the Undersampled Dataset

Because this model is well-trained, the chief variables are also well-ranked by the mean decrease in accuracy factor. The Figure 26 shows their ranking. The GPL traffic flow is the most important variable. Next factors are travel time variability and ML traffic flow. Surprisingly, blockages, precipitation, and time of the trip are not effective factors. Even ramp blockage has a negative accuracy, which means it would be better to exclude this variable from the analysis to increase the accuracy.

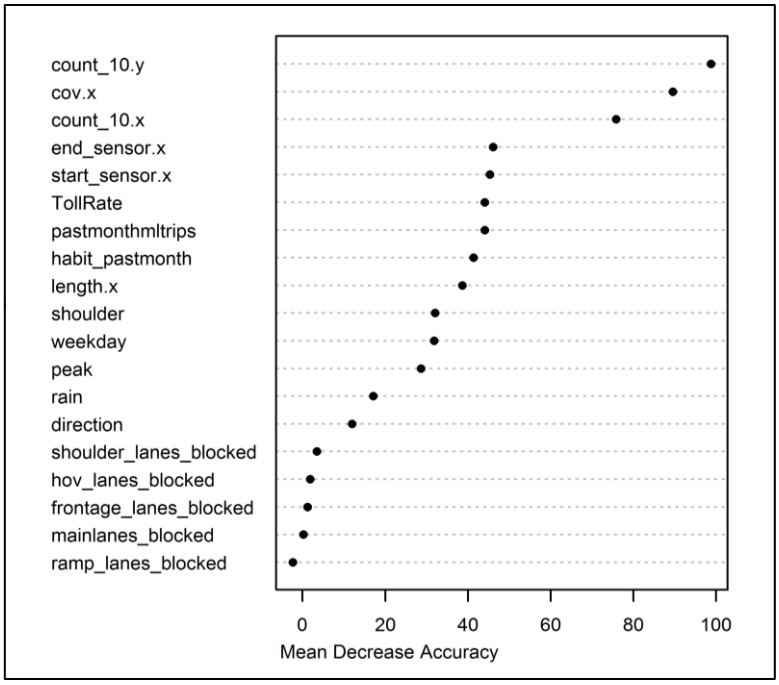

Figure 26 Variable Importance- Initial BRF for the Undersampled Dataset

5.5.3. SMOTEd Dataset

This dataset is established by applying SMOTE method on binary ML trip classification. The SMOTEd dataset is a balanced dataset with 42.86% of total paid ML trips as U-ML trips and the total size of 628,229 ML trips (see Table 10).

The constructed random forest errors' plot shows the dataset is balanced, and all types of errors decrease by the increase in number of trees. The overall OOB error is 0.1, which is less than the overall OOB error of the undersampled trained random forest. Nevertheless, the U-ML trip class in this model has a greater error than E-ML trip class. In other words, the test model's specificity is lower than its sensitivity, and this model works better for identifying E-ML trips than U-ML trips. As shown in Table 15, this model has a high accuracy and sensitivity of 0.878 and 0.940. However, the specificity is

as low as 0.386. In other words, there is a great possibility for this model to fail in predicting some U-ML trips. AUC is 0.663 showing the model is not as good as the initial BRF trained by the undersampled dataset (see Figure 28). Since this resampling technique is inferior, further importance analysis is not conducted.

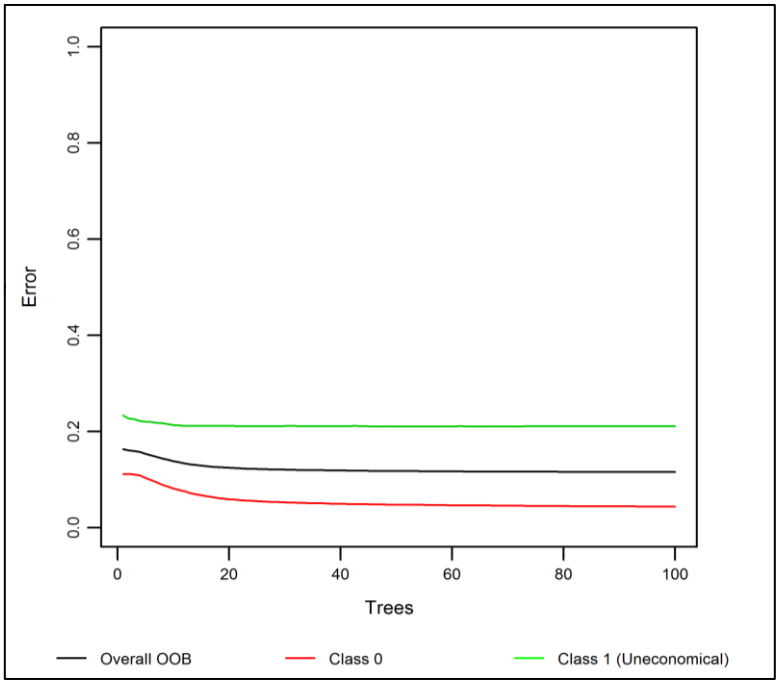

Figure 27 Errors Plot- Initial BRF for the SMOTEd Dataset

Table 15 Model Specifications- Initial BRF for the SMOTEd Dataset

Accuracy	Sensitivity	Specificity	AUC
0.878	0.940	0.386	0.663

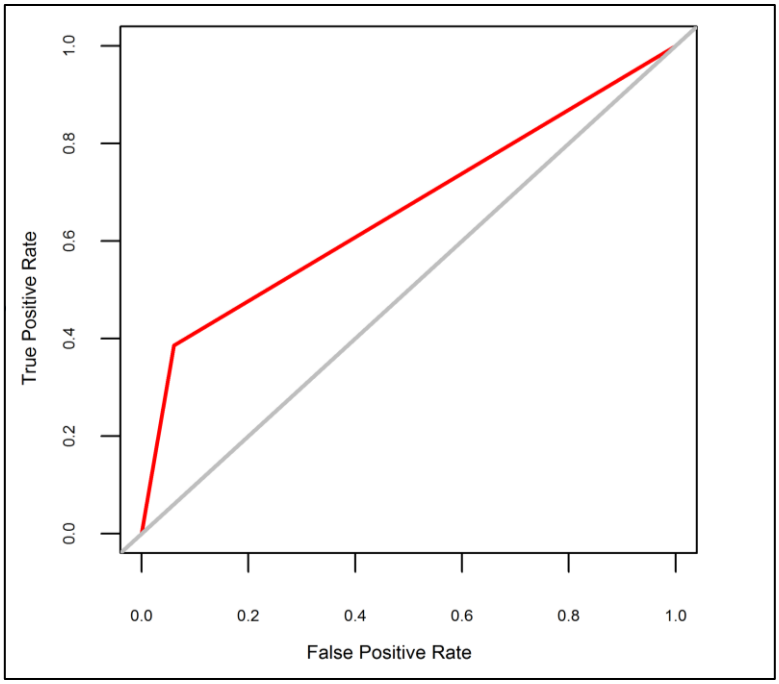

Figure 28 ROC Curve- Initial BRF for the SMOTEd Dataset

5.6. Initial Multiclass Random Forest (MRF) Model

5.6.1. Imbalanced Dataset

As specified in Table 11, this dataset consists of 801,554 paid ML trips with 5.1% of total paid ML trips as U-ML trips, 16.47% of total paid ML trips as middle ML trips, and 78.43% of total paid ML trips as E-ML trips. This dataset has one major class and two minor classes, which makes the dataset imbalanced. This issue is also significant in the random forest errors' plot (see Figure 29). The overall OOB error is decreased to 0.2, and the major class has a slight error. However, the minor classes' errors increase by the increase in number of trees. This is the imbalance effect causing the error become larger by adding more data to the model.

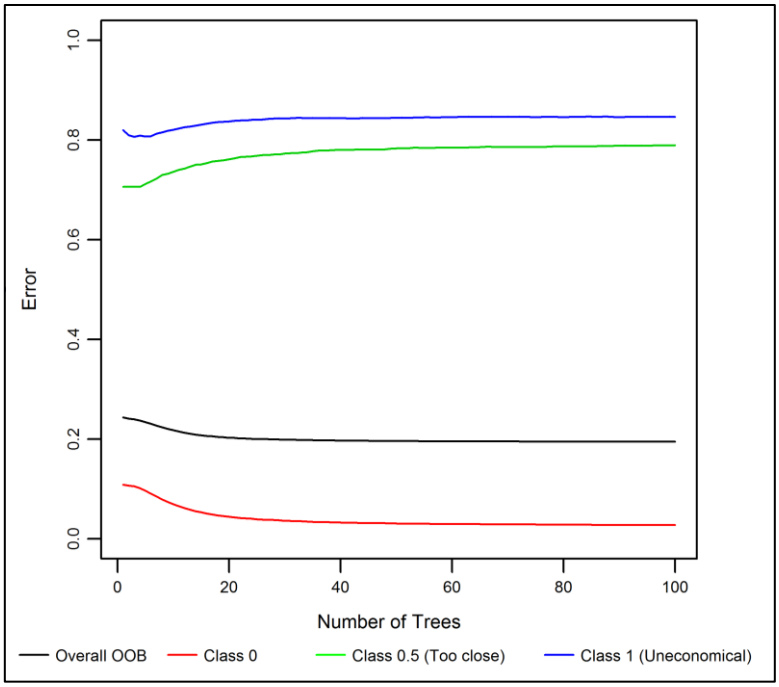

Figure 29 Errors Plot- Initial MRF for the Imbalanced Dataset

The accuracy and AUC for the test model is 0.806 and 0.603 relatively (see Figure 30). Despite it is not very deficient, the specificity and sensitivity for each class shows this model does not predict classes well (see Table 16). Either specificity or sensitivity is low for ML trip classes. In other words, the initial MRF model trained by imbalanced dataset fails in predicting all classes. Again, it shows imbalanced dataset is not a good dataset to train the final models.

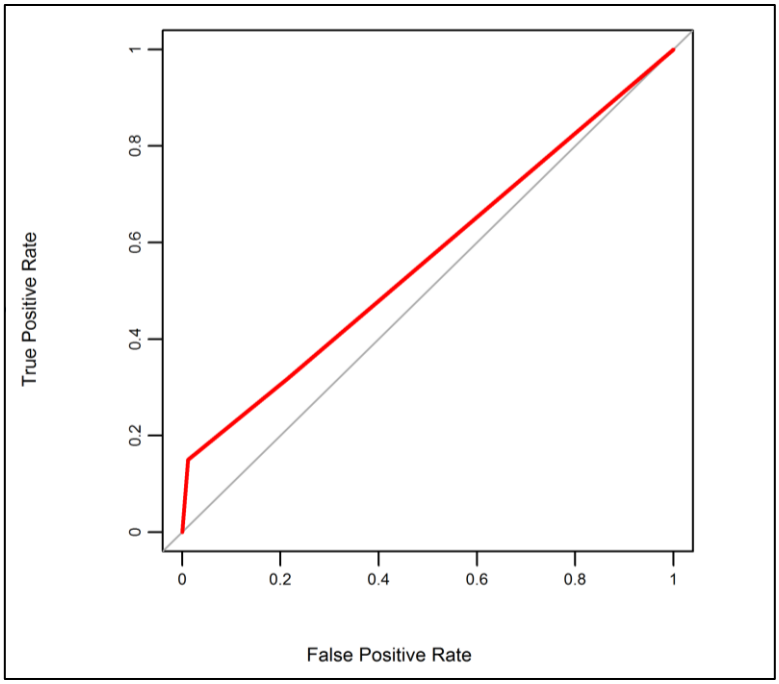

Figure 30 ROC Curve- Initial MRF for the Imbalanced Dataset

Table 16 Model Specifications- Initial MRF for the Imbalanced Dataset

Class	Sensitivity	Specificity
E-ML	0.975	0.239
Middle ML	0.203	0.968
U-ML	0.150	0.996

5.6.2. Undersampled Dataset

The size of undersampled subset used in multiclass analysis is 122,718 paid ML trips with 33.33% for each paid ML trip class. This data is balanced, and all paid ML trip classes' errors decline when the number of trees increases. The overall OOB error is 0.4, and the middle ML trip class error is greater than the other ML trip classes.

Subsequently, E-ML and U-ML trip classes will be better predicted by this model rather than the middle ML trip class (see Figure 31).

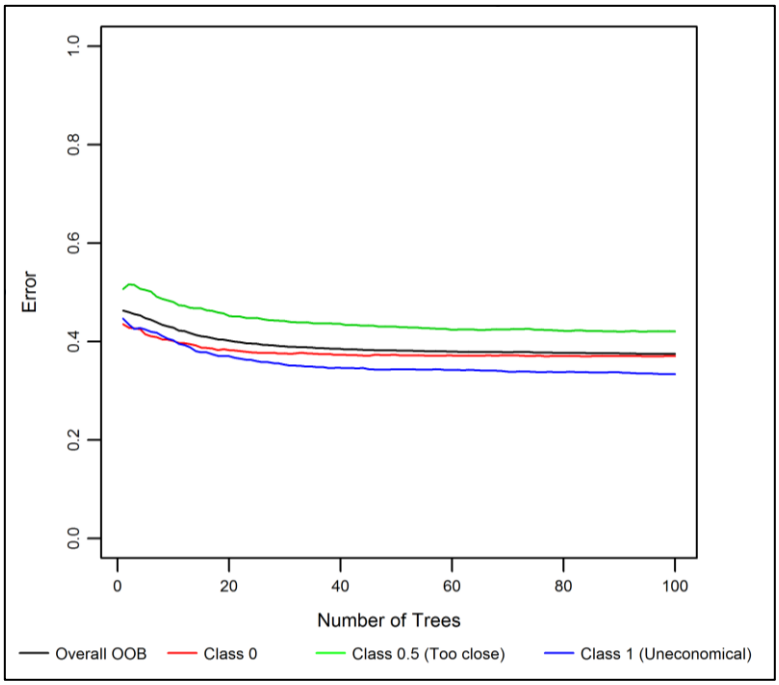

Figure 31 Errors Plot- Initial MRF for the Undersampled Dataset

The accuracy and AUC are 0.626 and 0.762 relatively (see Figure 32). Both these parameters show this model is well-designed for predicting multiclass ML trips. Also, specificity and sensitivity are equally adequate for all ML trip classes (see Table 17). Again, undersampling technique works well in training the model, and is a good choice for the final dataset resampling.

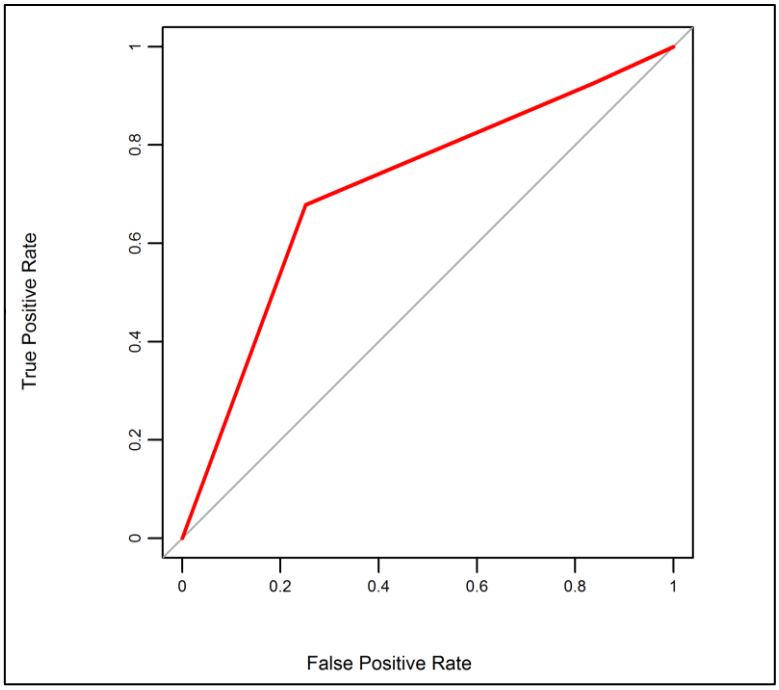

Figure 32 ROC Curve- Initial MRF for the Undersampled Dataset

Table 17 Model Specifications- Initial MRF for the Undersampled Dataset

Class	Sensitivity	Specificity
E-ML	0.631	0.858
Middle ML	0.586	0.757
U-ML	0.678	0.852

This model also presents the most important variables in multiclass analysis of ML trips by the means of the mean decrease in accuracy factor. Figure 33 shows the ranking of these variables. Travel time variability, GPL traffic flow, and ML traffic flow are the most important ones in multiclass. However, blockages and rain are not significant in this classification.

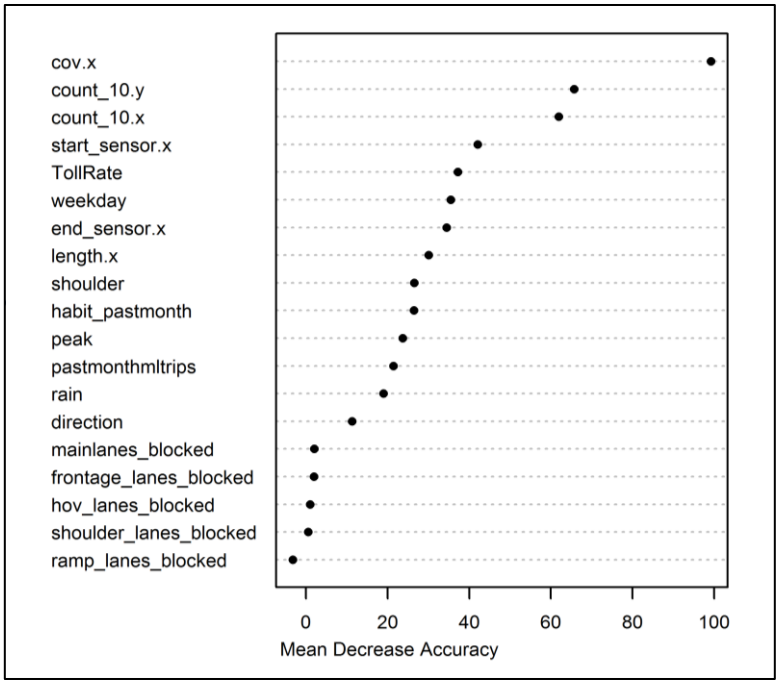

Figure 33 Variable Importance- Initial MRF for the Undersampled Dataset

5.7. Initial Binary Logistic Regression (BLR) Model

5.7.1. Imbalanced Dataset

The imbalanced training set is used to train a logistic regression model for predicting the binary ML trip classification. The size of this dataset is 801,554 paid ML trips with 11% of total paid ML trips as U-ML trips (see Table 10).

This dataset is significantly imbalanced as same as observed in section 5.5.1. In addition, the paradox of accuracy can be easily noticed. The accuracy and sensitivity of the test model in this analysis are 0.889 and 0.999 relatively, while the specificity is 0.021 (see Table 18). It shows that the model classifies almost all of the U-ML trips as the major E-ML trip class. Therefore, this model is not well-fitted. AUC is revealed as

0.510, which is insignificant and the model is unproductive for training the final models (see Figure 34).

Table 18 Model Specifications- Initial BLR for the Imbalanced Dataset

Accuracy	Sensitivity	Specificity	AUC
0.889	0.999	0.021	0.510

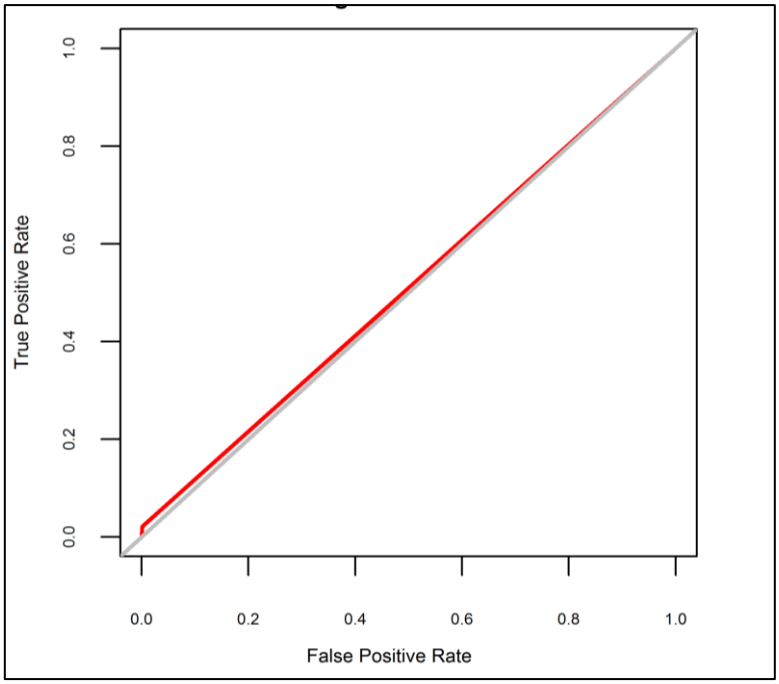

Figure 34 ROC Curve- Initial BLR for the Imbalanced Dataset

5.7.2. Undersampled Dataset

This dataset is created by random undersampling of the major E-ML trip class to have the same size as the U-ML trip class in the main training set. Therefore, the new undersampled dataset will have two equally sized classes with a total size of 179,320 paid ML trips. The proportion of each class is almost 0.5, and the dataset is balanced (see Table 10).

The model's accuracy, sensitivity, and specificity are 0.652, 0.650, and 0.671 relatively (see Table 19), which shows the model is well-fitted. The ROC curve is also plotted and shows that AUC is 0.660 (see Figure 35). This model confirms the undersampled set is a good set for training final models.

Table 19 Model Specifications- Initial BLR for the Undersampled Dataset

Accuracy	Sensitivity	Specificity	AUC
0.652	0.650	0.671	0.660

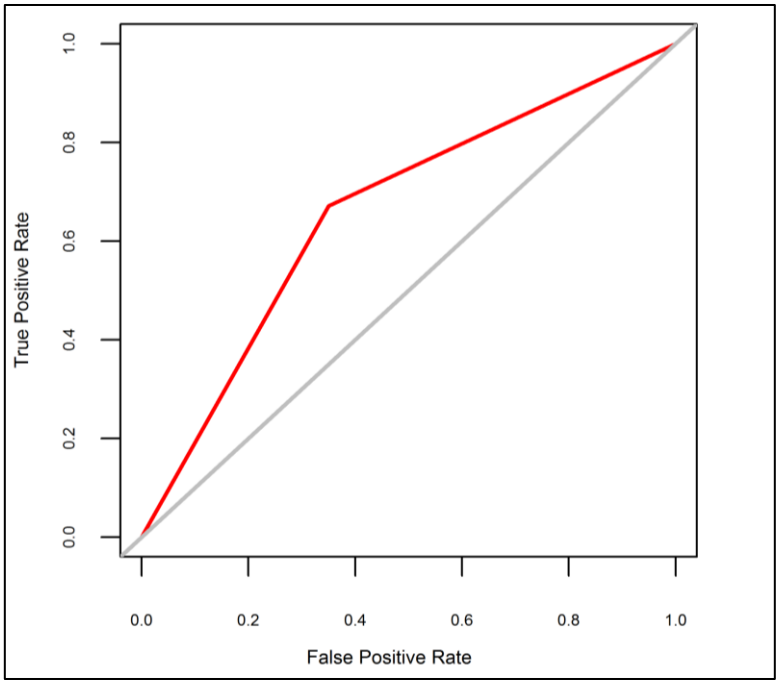

Figure 35 ROC Curve- Initial BLR for the Undersampled Dataset

5.7.3. SMOTEd Dataset

This dataset is developed by applying SMOTE on the binary ML trip classification. This dataset is a balanced dataset with 42.86% of total paid ML trips as U-ML trips and the total size of 628,229 paid ML trips (see Table 10).The model is well-balanced, but it is not functioning well as indicated in Table 20.

The model's accuracy is 0.736, but the specificity is 0.408. Also, the area under the ROC curve, AUC, is 0.593 (see Figure 36), which shows that the model is inefficient, and SMOTEd set is not functional for training the final models.

Table 20 Model Specifications- Initial BLR for the SMOTEd Dataset

Accuracy	Sensitivity	Specificity	AUC
0.736	0.777	0.408	0.593

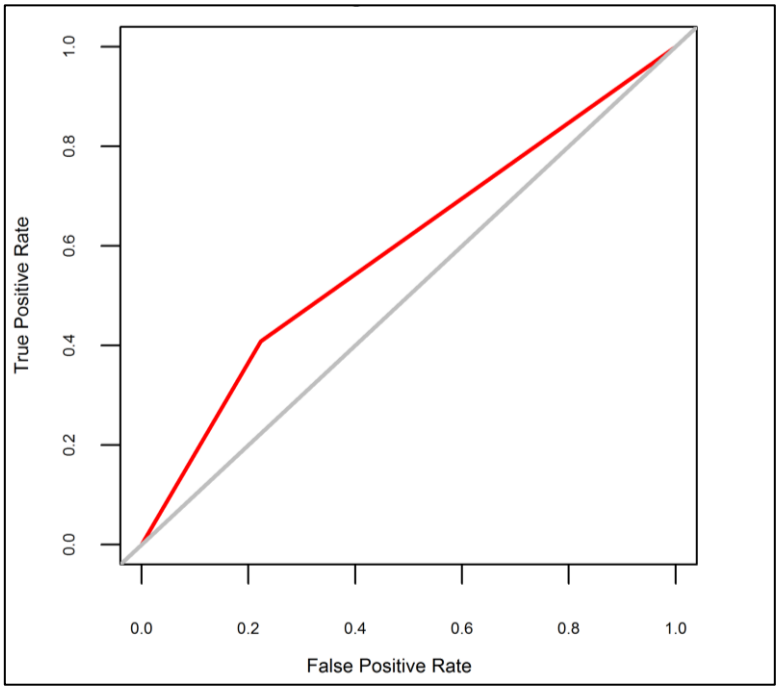

Figure 36 ROC Curve- Initial BLR for the SMOTEd Dataset

5.8. Discussion of Initial Models

Initial binary random forest (5.5), multiclass random forest (5.6), and binary logistic regression (5.7) models were trained by imbalanced, undersampled, and SMOTEd datasets. Then, a test set of data is analyzed by the models to test each model's accuracy. All cases are discussed regarding their accuracy, sensitivity, specificity, and AUC. In all cases the best resampling technique was undersampling. The most unfailing

goodness of fit is AUC. As illustrated in Table 21, the highest AUC belongs to the undersampled dataset in all pattern recognition methods.

Table 21 Initial AUC Summary

AUC	Initial BRF	Initial MRF	Initial BLR
Imbalanced	0.571	**0.603**	0.510
Undersampled	0.761	**0.762**	0.660
SMOTEd	**0.663**	--	0.593

The key independent variables are shown by their importance results in the random forest models using the best resampling technique (undersampling). These variables are ranked in sections 5.5.2 and 5.6.2.

The most important variables are computed by summing up the mean decrease in accuracy from initial BRF and MRF for undersampled datasets. The total decrease in accuracy for all variables is indicated in Figure 37. As illustrated in this graph, the first three factors are very important to the model. These factors are travel time variability, GPL traffic flow, and ML traffic flow. The next three variables are start sensor, toll rate, and end sensor. These six variables are selected for the final model noting that they have no significant correlations, and they are unique independent variables. Also, it is noteworthy that these six variables are a representative of time of the trip, route of the trip, cost of the trip, and traffic flow.

As indicated in sections 5.5.2 and 5.6.2, an efficient number of trees to have both a small OOB error and not large worthless number of trees is 50. The OOB error begins to stay constant from this point (see Figure 24 and Figure 31).

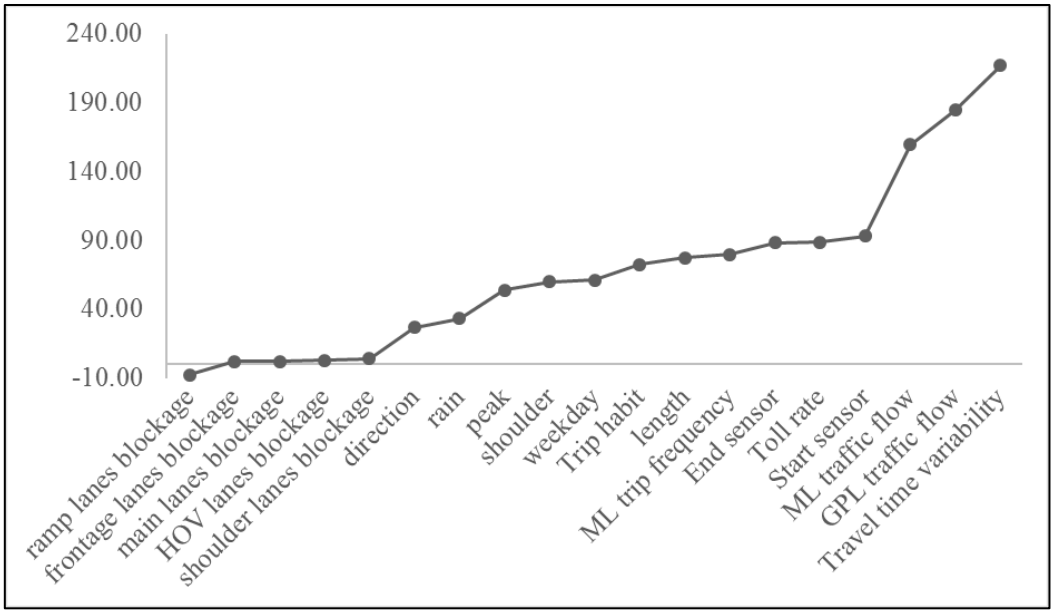

Figure 37 Total Decrease in Accuracy

5.9. Final Models

The final models are generated for the most important variables identified from the initial analysis. These factors are GPL traffic flow, ML traffic flow, toll rate, start sensor, end sensor, and travel time variability. The main formula used in this analysis as indicated in Equation (13).

$$uneco(binary, multi-class) \sim F(ML\ traffic\ flow,\ startsensor,\ endsensor,\ travel\ time\ variability, GPL\ traffic\ flow,\ TollRate) \quad (13)$$

Also, these models use the undersampling technique to generate balanced training sets. The undersampled datasets are the same undersampled datasets from the initial analyses. Table 22 shows the size and class proportions of the undersampled datasets.

Table 22 Undersampled Datasets

	U-ML trip	Middle ML	E-ML trip	Training set Size
Binary	50.05%	---	49.95%	179,320
Multiclass	33.33%	33.33%	33.33%	122,718

In this section, both random forest and logistic regression models are designed to discover variables importance ranking and their type of impacts.

5.9.1. Final Binary Random Forest (BRF) Model

The number of trees established in the random forest method was 50 as concluded from initial analysis.

A small fragment of a sample tree from this model is illustrated in Figure 38. The outcome from this tree is combined with the outcome from 49 other trees, and the mode of this combination will be the binary ML trip classification.

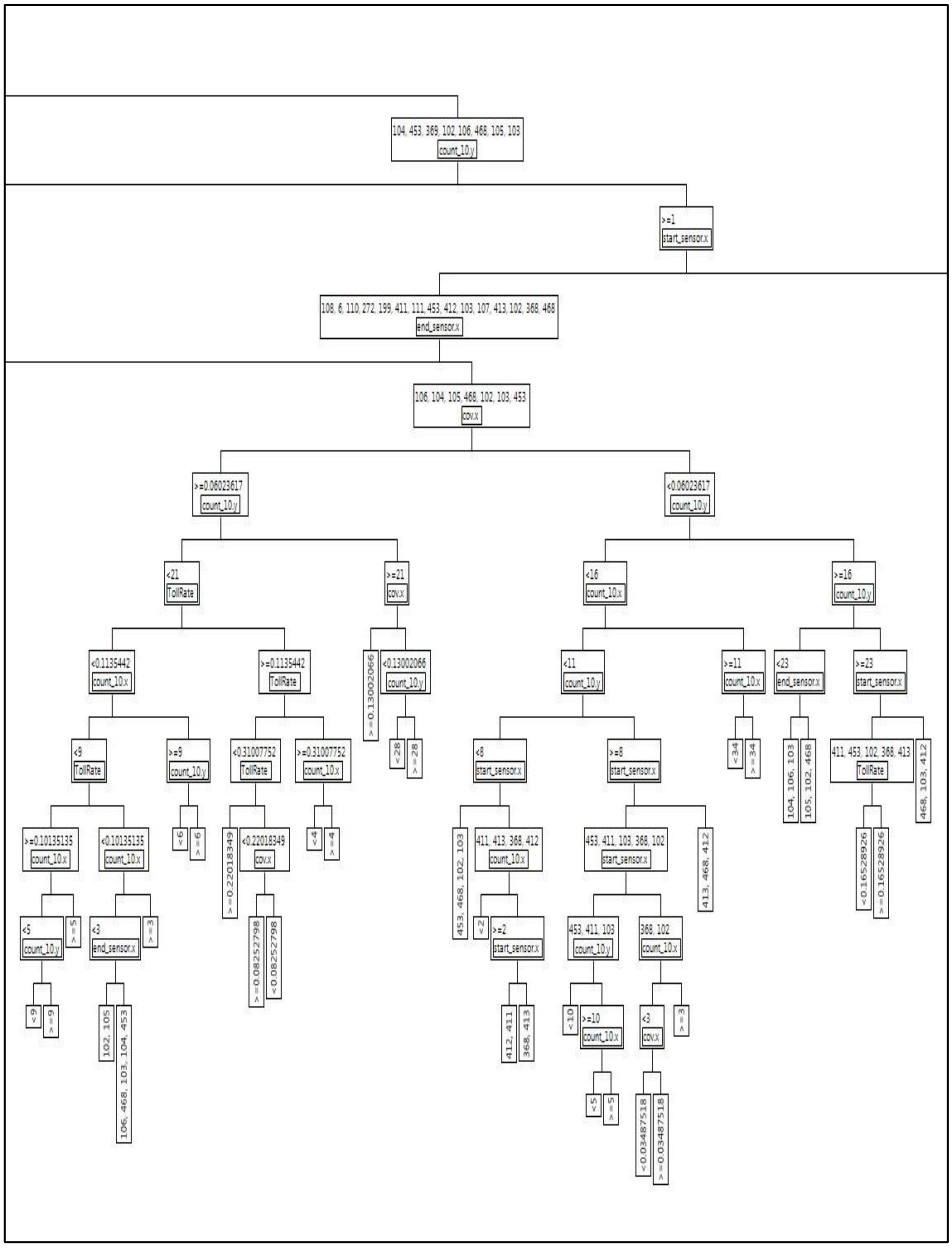

Figure 38 Part of a Sample Tree in the Final BRF

By adding more trees to the model, the overall OOB error and classes' errors decrease and reach 0.2 to 0.3 (see Figure 39), which has not significantly changed from initial BRF errors (see Figure 24). In other words, exclusion of some variables and limiting the model to only six variables did not increase the errors notably. It shows that these six variables can explain a great part of the U-ML trip pattern. The errors' plot also shows that the data is balanced, and the growing number of trees would lead to a better model with a lower error value.

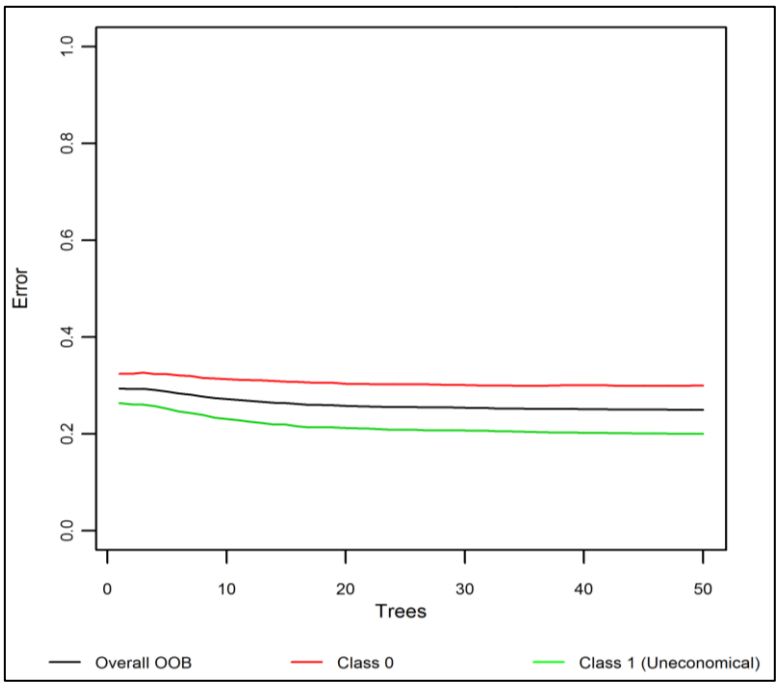

Figure 39 Errors Plot- Final BRF

In addition, the U-ML trip class has a lower error rate than E-ML trip class. Thus, U-ML trip class will be more accurate than other ML trip classes. In other words, the model specificity is higher than its sensitivity as indicated in Table 23. The accuracy,

sensitivity, and specificity are 0.717, 0.706, and 0.802. The equally high and adequate values of these three parameters show the model is well-constructed. Also, AUC is 0.754 (see Figure 40), which has dropped slightly from the initial BRF value, 0.761 (see Table 14). However, the drop is 0.8%. Therefore, the model does not decrease in accuracy much when excluding the many variables.

Table 23 Model Specifications- Final BRF

Accuracy	Sensitivity	Specificity	AUC
0.717	0.706	0.802	0.754

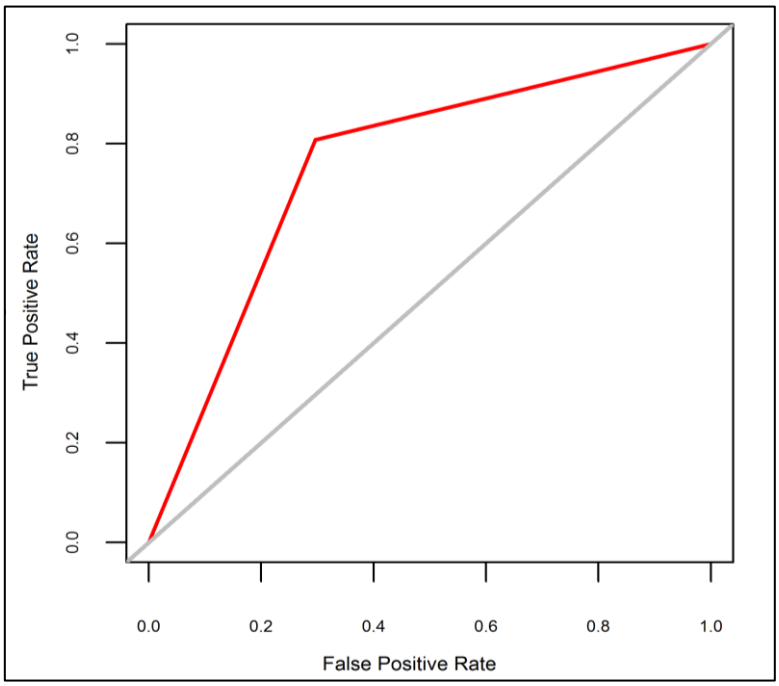

Figure 40 ROC Curve- Final BRF

Also, the new importance ranking of the variables changed from the initial BRF model. Figure 41 shows the variables' importance in terms of mean decrease in accuracy. As indicated in Figure 41, removing any of these variables would greatly decrease the accuracy of the model.

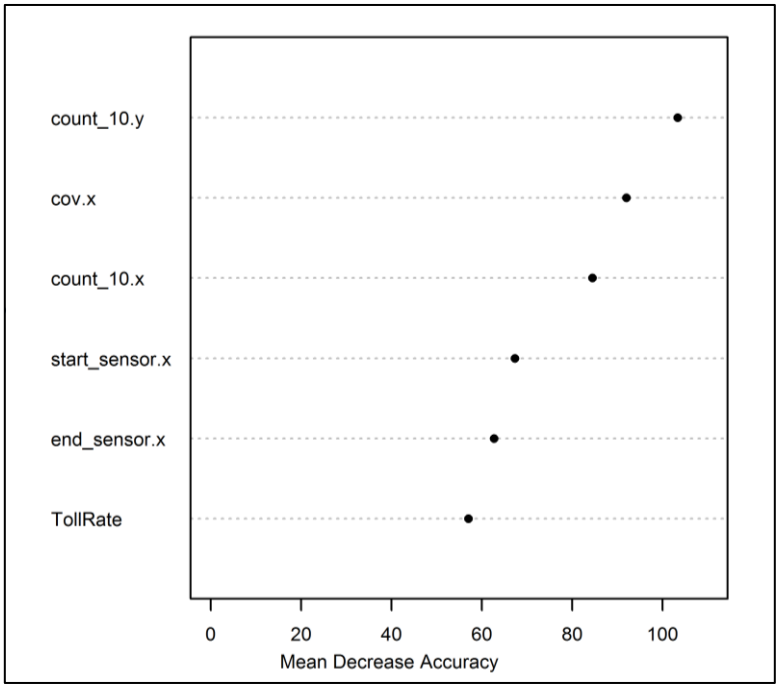

Figure 41 Variables Importance- Final BRF

The variable ranking order is the same as the initial BRF model ranking except for start sensor and end sensor variables. This fact shows other excluded variables have sort of correlation with these two factors, and omitting them made these two factors impact more distinct.

This model is a simple binary and well-constructed random forest model. It clearly shows the ranking of each variables' importance, and has a high accuracy in

predicting U-ML trips Nevertheless, the effect of each variable on the likelihood of U-ML trip cannot be easily interpreted from this model.

5.9.2. Final Multiclass Random Forest (MRF) Model

This model is obtained by generating 50 trees. Increasing the number of trees reduces the overall OOB error and all class errors (see Figure 42). Comparing the final MRF error plot with initial MRF error plot (see Figure 32), it can be concluded that the errors do not drop remarkably by excluding many variables. In addition, the model can identify the extreme classes (economical and uneconomical classes) with a smaller error rate than when there is also a middle class. The overall OOB error is higher than the final BRF model's overall OOB error.

The model's specifications are illustrated in Table 24. Sensitivity of the middle class is 0.557, but other ML trip classes' sensitivities and specificities are high. The accuracy of the model is 0.614 and AUC is 0.752 (see Figure 43). Comparing to an AUC of 0.762 derived from the initial MRF analysis, the model accuracy has not changed much.

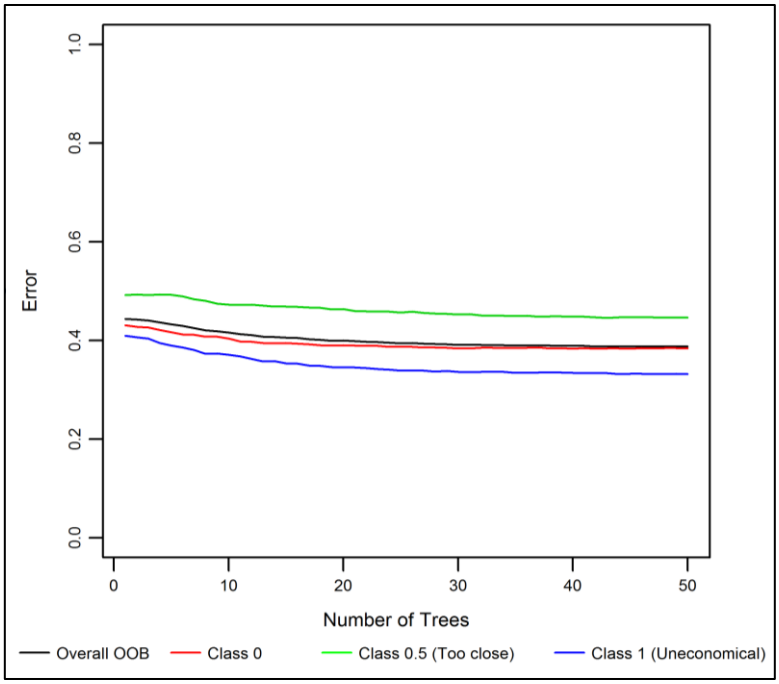

Figure 42 Errors Plot- Final MRF

Table 24 Model Specification- Final MRF

Class	Sensitivity	Specificity
E-ML	0.622	0.849
Middle ML	0.557	0.760
U-ML	0.676	0.839

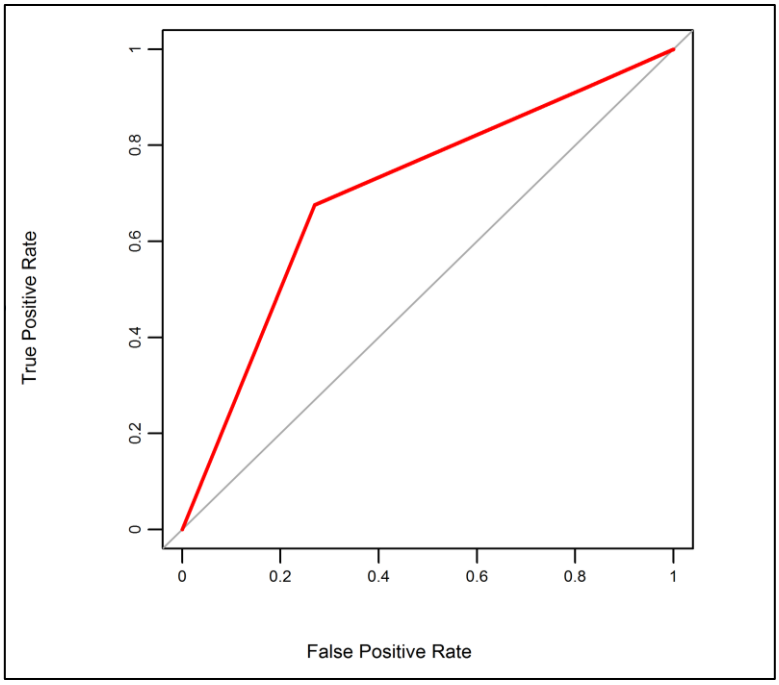

Figure 43 ROC Curve- Final MRF

The variable importance rankings (see Figure 44) show that the travel time variability, and GPL traffic flow, and ML traffic flow are the most important variables. This model is also well-constructed. However, it is difficult to measure the magnitude of each variable's impact.

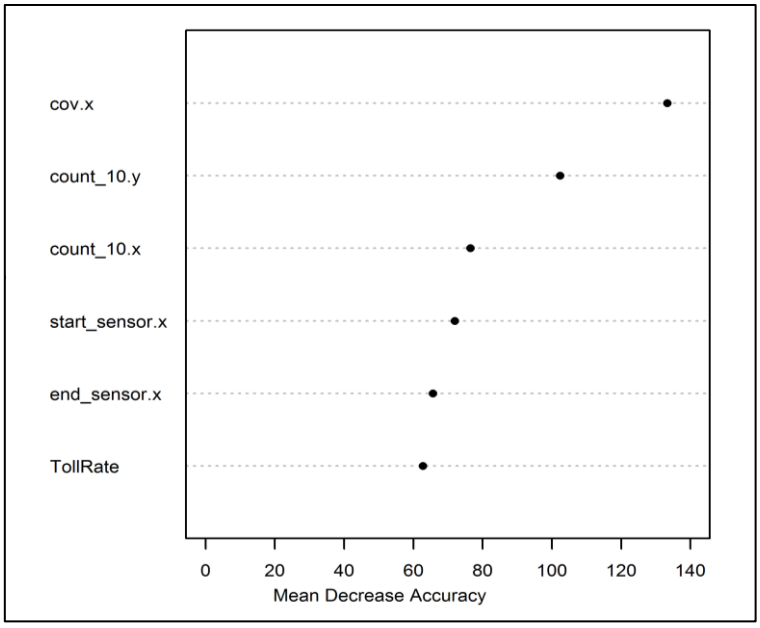

Figure 44 Variables Importance- Final MRF

5.9.3. Final Binary Logistic Regression (BLR) Model

To predict future U-ML trips and estimate their probability, the final BLR model may be the best model. This model's specifications are shown in Table 25. Adequate and equally high values of accuracy, sensitivity, and specificity indicate that the model is well-designed. Also, ROC curve is showing that the model is predicting accurately (see Figure 45). Accuracy and AUC may have dropped comparing to the initial BLR model, however, the model and variable coefficients are less complicated and easier to use for future ML trip estimation.

Table 25 Model Specifications- Final BLR

Accuracy	Sensitivity	Specificity	AUC
0.614	0.608	0.664	0.636

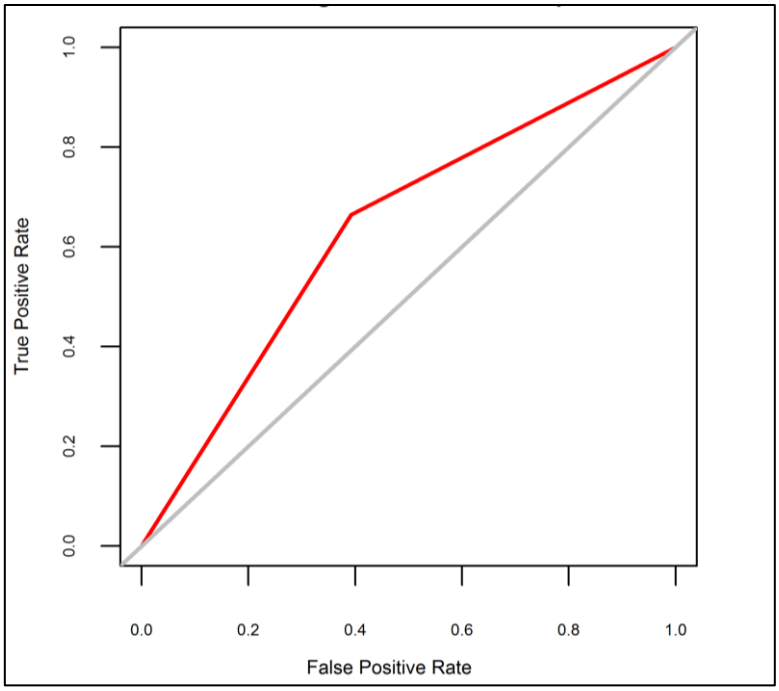

Figure 45 ROC Curve- Final BLR

The main advantage of the final BLR method is to easily predict future ML trip classifications. Table 26 illustrates how each factor affects the U-ML trip probability. In other words, a positive value indicates an increased likelihood of a U-ML trip by an increase in the associated variable. The lowest p-value suggests a strong association between the variable and U-ML trip likelihood.

ML traffic flow increase the likelihood of a U-ML trip. Also, high GPL traffic flow shows a high number of vehicles using GPLs, which decreases the U-ML trip likelihood. One of the most important factors is the travel time variability as shown by final MRF model. The higher the travel time variability, the higher the U-ML trip probability. In addition, high toll rates lead to a smaller number of U-ML trips. Sensor 469 is arbitrarily set as the base sensor, with a coefficient equal to 0. All other sensor

coefficients are evaluated versus this sensor. The highest and lowest coefficients for the start point are for sensors 109 and 101. The sensors with highest and lowest coefficients for the end point are 271 and 369. A thorough look into these coefficients shows that in general the longer the trip is, the lower the chance to have a U-ML trip (see Figure 1). The lowest p-value (the strongest association) is for variables: GPL traffic flow, toll rate, start sensor 109, and travel time variability.

Table 26 Parameters Estimates- Final BLR

Variable	Estimate	Std. Error	Z value	P-Value
(Intercept)	-0.07	0.18	-0.39	6.95E-01
MLs traffic flow	0.01	0.00	24.24	7.68E-130
start_sensor.x6	0.91	0.04	24.34	7.97E-131
start_sensor.x101	-1.07	0.22	-4.84	1.33E-06
start_sensor.x102	-0.75	0.22	-3.35	8.22E-04
start_sensor.x103	-0.69	0.22	-3.11	1.84E-03
start_sensor.x104	-0.89	0.26	-3.44	5.74E-04
start_sensor.x105	0.20	0.22	0.88	3.78E-01
start_sensor.x106	-0.72	0.24	-2.99	2.75E-03
start_sensor.x107	0.93	0.04	26.09	4.52E-150
start_sensor.x108	1.30	0.05	27.21	4.76E-163
start_sensor.x109	2.72	0.04	60.53	0.00E+00
start_sensor.x110	2.09	0.09	22.74	1.91E-114
start_sensor.x111	1.94	0.05	39.64	0.00E+00
start_sensor.x112	-0.08	0.30	-0.27	7.88E-01
start_sensor.x199	0.93	0.05	20.25	3.33E-91
start_sensor.x272	0.74	0.04	17.50	1.46E-68
start_sensor.x368	-0.91	0.22	-4.11	4.02E-05
start_sensor.x411	-0.62	0.22	-2.76	5.77E-03
start_sensor.x412	-0.86	0.22	-3.86	1.14E-04
start_sensor.x413	-0.92	0.22	-4.11	4.01E-05
start_sensor.x449	0.32	0.22	1.46	1.44E-01

Table 26 Continued

Variable	Estimate	Std. Error	Z value	P-Value
start_sensor.x453	0.10	0.23	0.42	6.78E-01
start_sensor.x456	0.54	0.17	3.24	1.19E-03
start_sensor.x460	0.22	0.05	4.82	1.42E-06
start_sensor.x468	-0.90	0.23	-3.84	1.25E-04
end_sensor.x5	0.49	0.12	4.00	6.24E-05
end_sensor.x101	0.91	0.12	7.37	1.77E-13
end_sensor.x102	0.17	0.16	1.06	2.87E-01
end_sensor.x103	0.65	0.13	5.12	3.01E-07
end_sensor.x104	0.64	0.19	3.37	7.57E-04
end_sensor.x105	0.23	0.12	1.85	6.44E-02
end_sensor.x106	0.33	0.13	2.59	9.53E-03
end_sensor.x107	-0.31	0.18	-1.68	9.34E-02
end_sensor.x108	-0.33	0.27	-1.23	2.19E-01
end_sensor.x109	-0.64	0.18	-3.49	4.74E-04
end_sensor.x110	-0.70	0.21	-3.33	8.76E-04
end_sensor.x111	-0.47	0.18	-2.59	9.63E-03
end_sensor.x112	-0.39	0.18	-2.11	3.48E-02
end_sensor.x271	0.93	0.12	7.61	2.71E-14
end_sensor.x369	-1.41	0.19	-7.55	4.32E-14
end_sensor.x414	-1.08	0.18	-5.91	3.40E-09
end_sensor.x415	-0.90	0.18	-4.95	7.33E-07
end_sensor.x416	-0.97	0.19	-5.26	1.43E-07
end_sensor.x453	0.44	0.45	0.99	3.20E-01
end_sensor.x456	-0.21	0.22	-0.93	3.52E-01
Travel time variability	4.91	0.11	43.71	0.00E+00
GPLs traffic flow	-0.02	0.00	-49.04	0.00E+00
Toll rate	-1.04	0.02	-48.72	0.00E+00

5.10. Discussion of Results from the Final Models

This study found that the undersampling technique was best for balancing the data. The undersampled data was used to train three models: final BRF, MRF, and BLR. The random forest methods provide the ranking of variables' impacts and the logistic regression model provides the magnitude of variables' impacts.

Table 27 compares AUCs from the initial analysis and the final models. It shows that the exclusion of many variables did not decrease the AUC of the final model very much. Instead the model is now easier and more practical for future use.

In both initial and final models, BRF and MRF models are almost equally accurate. However, it is easier to predict two classes instead of three classes. Therefore, BRF model is the focus of this discussion.

Table 27 Final AUC Summary

AUC	BRF	MRF	BLR
Initial Full Model	0.761	0.762	0.660
Final Model	0.754	0.752	0.636

Table 28 shows variable importance ranking and their significance in the final models. The analysis of variance (ANOVA) for BLR model is implemented. This analysis sequentially compares the smaller model with the next more complex model. This test is conducted for each variable by comparing the full model and the model without the variable. The Wald Chi-squared test evaluates this comparison by generating p-values, which shows the significance of the variable in the model. A large p-value in

BLR model shows that the model without the corresponding variable is essentially the same. As indicated, all p-values are small, and all variables have a significant impact on ML trip classification, but there might be slight differences in rankings.

As indicated by random forest models, the most important variables for ML trip classification are GPL traffic flow, travel time variability, ML traffic flow, start sensor, end sensor, and toll rate. However, their impact on the U-ML trip rate (for example, does the variable increase or decrease the likelihood of a U-ML trip) is unclear in random forest models. The BLR model provides more information on a variables' impact on U-ML trip rate based on the estimated coefficients.

Table 28 Final Variables' Impacts

Variables	Final BRF Mean Decrease Accuracy	Final MRF Mean Decrease Accuracy	Final BLR ChiSquare Test P-value
GPL traffic flow	103.29	102.35	< 0.0001
Travel time variability	91.97	133.24	< 0.0001
ML traffic flow	84.47	76.43	< 0.0001
Start sensor	67.26	71.90	< 0.0001
End sensor	62.71	65.61	< 0.0001
Toll rate	56.99	62.73	< 0.0001

GPL traffic flow is the most important variable in BRF model. Also, the BLR model shows that the GPL traffic flow is an influential variable in predicting the trips (see Table 26). The decrease in GPL traffic flow will result in an increase in U-ML trip likelihood as the result of a shorter travel time on GPLs.

The next important variable in BRF model is travel time variability. This variable is also significantly important in the BLR model (considering the associated p-value). The increase in travel time variability will cause an increase in U-ML trips.

ML traffic flow rate is the third most important variable in the BRF model and a significant variable of the BLR model. The increase in ML traffic flow likely indicates longer ML travel time does indicate a higher likelihood of a U-ML trip.

The next two variables of importance are start sensor and end sensor. Start and end sensors are two categorical variables, and each class of them has a specific coefficient in the BLR model. To verify the BLR model, an investigation of U-ML trip likelihood for each sensor pair (start and end sensor) is conducted. U-ML trip likelihood for each sensor pair with more than 5000 U-ML trips is computed using the main dataset. Table 29 shows the probability of most and least likely routes for U-ML trips. Also, Figure 46 maps the most and least likely routes for U-ML trips. It can be observed that sensor 271 is the most likely end point for U-ML trips. Also, sensor 101 is the least likely start point for U-ML trips. These two facts were noted in the BLR model. Figure 46 also shows the most and least number of U-ML trips are happening around a specific number of sensors.

The final variable in BRF model is toll rate. This variable is also significant in BLR model, and the decrease in toll rate (non-peak hours) will result in a higher likelihood of a U-ML trip.

Table 29 Most and Least Likely Routes for U-ML Trips

Sensor Pair	Total Trips	Economical Trips	Uneconomical Trips	U-ML Percentage	U-ML trip Likelihood
101-103	179012	173718	5294	3%	**Least likely**
101-105	594085	570433	23652	4%	
368-105	108321	103258	5063	5%	
101-5	159621	152328	7293	5%	
368-101	96886	91336	5550	6%	
449-271	22843	17683	5160	23%	**Most Likely**
449-5	30316	22300	8016	26%	
105-271	273065	197828	75237	28%	
108-112	43330	30681	12649	29%	
449-101	45320	24892	20428	45%	

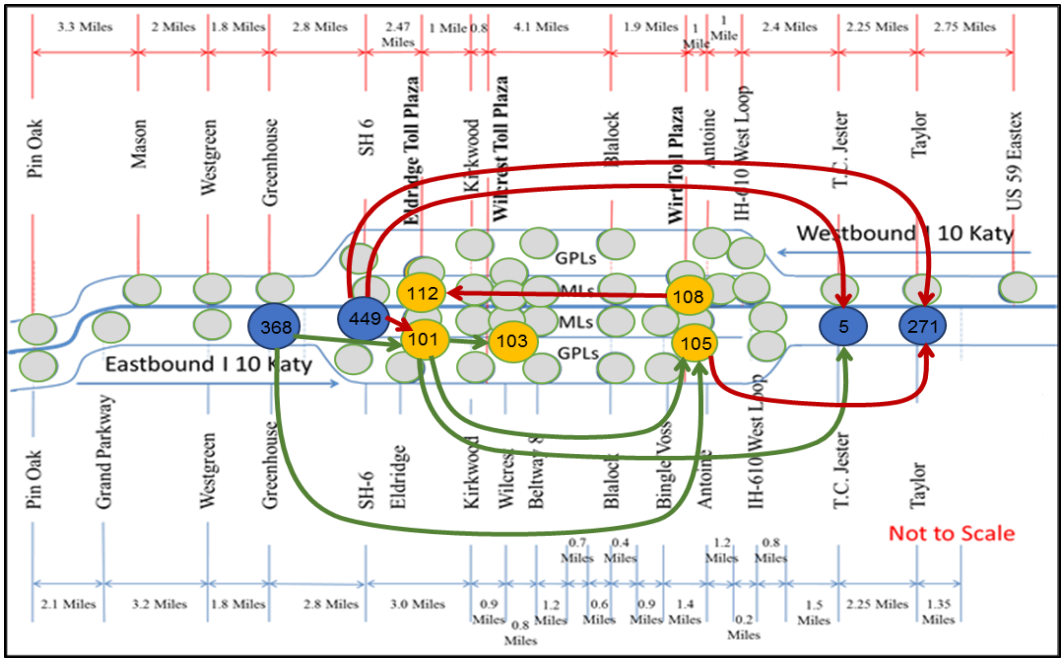

Figure 46 Most and Least Likely Routes for U-ML Trips (Red arrows show the most likely routes, and green arrows show the least likely routes)

6. SUMMARY AND CONCLUSIONS

The initial objective of this study was to identify the uneconomical managed lane (U-ML) trips and the factors associated with these trips. Managed Lane (ML) trips are expected to save travel time. However, Burris et al. (2016) showed 11% of total paid ML trips on the Katy Freeway had a negative travel time saving, and termed them "uneconomical managed lane (U-ML) trips".

To perform this study and examine U-ML trips and related factors, a unique dataset was obtained from different sources. The first set was the data collected by AVI sensors along the Katy Freeway and operated by TxDOT. This set includes the transponder ID, sensor ID, and the detection time for each vehicle passing each sensor. The next dataset was acquired from HCTRA. This dataset presents the collected tolls at toll plazas, and includes the transponder ID, toll plaza ID, lane ID, and detection time. The third dataset was obtained from NOAA for daily precipitation measurements in the Katy Freeway area. The first two datasets were combined to form a dataset containing vehicle ID, passed sensor IDs and detection times, and total tolls paid at different toll plazas. In other words, this combined set provides the route of the trip, time of the trip, and cost of the trip for each trip. The precipitation also added a significant measurement for weather conditions for each trip. This combined dataset was reduced to include only paid ML trips. Also, an alternate GPL trip was computed for each ML trip. The alternate GPL trip is an artificial trip starting at the same point and same time of the actual ML trip. Using GPL vehicle data, the alternate GPL trip attributes were computed.

To build the final dataset, ML trips were needed to be classified based on their travel time savings. Two forms of classification were suggested in this study. One is a binary classification based on TTD, and the other is multiclass classification based on RTTD. A preliminary analysis showed the ML travel time ranges from less than 1 minute to over 35 minutes with an average of 9.6 minutes. Similarly, the alternative GPL trip may take from less than 1 minute to over 45 minutes with a mean time of 12.2 minutes. The travel time difference (TTD) also ranges from -20 minutes to 3 minutes with an average of -2.6 minutes, showing most of the ML trips save travel time. 11% of ML trips are uneconomical. The relative travel time difference (RTTD) computes travel time loss on MLs relative to the ML travel time. RTTD ranges from -2 to 2 with an average of -0.3. RTTD helps to introduce a middle ML trip class with a negligible travel time saving or loss, which has different attributes from economical and U-ML trip classes.

As noted earlier, the main ML trip dataset is imbalanced since there is a small proportion of U-ML trips (see Figure 21). The main dataset consists of 7,013,587 ML trips from almost every month from January 2012 to September 2014. To get a sample set, 1,001,941 trips, or one-seventh of all trips, were randomly selected. This sample set was also divided into two groups: a training and a test set. The training set accounted for 80% of the sample size or 801,554 trips. The binary classification and multiclass classification of ML trips in the training set showed U-ML trips are 5-11% of total ML trips. Thus, new balanced training sets were generated using resampling techniques, undersampling or SMOTE. However, the resampling techniques are highly dependent on

the proportion and number of classes. Consequently, resampling was conducted for binary classification and multiclass classification separately (see Table 10 and 11).

The distribution of each variable among ML trip classes was explored to find the type and magnitude of their influence on the percentage of U-ML trips. First, the ML traffic flow has a higher average for U-ML trips than E-ML trips. Conversely, the GPL traffic flow has a lower mean for U-ML trips. These two facts state that the congestion on MLs and the lower GPL traffic flow will lead to a longer ML travel time and shorter GPL trip, which is a U-ML trip. The next factor examined was the time of the trip, including weekday, peak hour, and shoulder hour. An examination of U-ML trip rate for each day of the week indicates 8-14% drop of U-ML trip rate over the weekends. This impact is as the result of a large drop in ML traffic flow over the weekends. Likewise, U-ML trip rate decreases by 6-13% during the non-peak and non-shoulder hours as the result of smaller ML traffic flow during these times. The other element of ML trips is the route of the trip, which is characterized by a start sensor, an end sensor, and trip length. ML trip classes distribution over start and end sensors showed how U-ML trip rate might significantly change from a sensor to sensor. Also, U-ML trip's average length is 7.96, which is smaller than other ML trips, showing U-ML trips are more likely over short distances. Also, the average trip frequency for U-ML trips is 7.24, which is slightly higher than other ML trip classes. Trip frequency, percent ML trip, rain, and blockages did not indicate any significant variation among ML trip classes in this analysis.

Three types of initial analysis were performed to find the most important variables and the best resampling technique to design the final models. Initial analyses

include binary random forest, multiclass random forest, and binary logistic regression. Each of these models is trained by imbalanced, undersampled, and SMOTEd datasets. The undersampled dataset with equally-sized classes performed the best overall.

The preliminary models also provided information on the most important factors to include in the final models. These factors are ML traffic flow, GPL traffic flow, toll rate, travel time variability, start sensor, and end sensor. These variables are independent of each other (referring to the correlation table), and they are inclusive, adding different trip attributes to the final models. Therefore, the final models were built using these six variables and undersampled datasets.

The final models are BRF, MRF, and BLR. The best parameter for evaluation is AUC, which shows almost the same AUC for both BRF and MRF models. However, their variable importance rankings are different. BRF model is the best model because of the easy binary classification and higher AUC. It can better predict future ML trip classes based on GPL and ML traffic flow, vehicles' start and end sensors, toll rate, and travel time variability. However, if there is a need for studying the impact of each of these variables separately, the BLR model is superior in that the coefficient of each variable indicates its impact on the likelihood of U-ML trips.

BRF model showed the most important variable is the GPL traffic flow. The next variables are travel time variability, ML traffic flow, toll rate, start sensor, and end sensor. Its AUC is 0.756, which is a high value in random forest modeling. BLR model has a lower AUC of 0.636 compared to BRF model. However, it provides an insight into each variable's impact. As observed, high ML traffic flow will lead to congestion on

MLs and a higher likelihood of a U-ML trip. Also, high GPL traffic flow will lead to congestion on the GPLs and a lower chance of a U-ML trip. A lower toll rate increases the probability of U-ML trips. This is not surprising since lower toll rates occur during the less congested times. Therefore, toll rate acts like a peak hour variable. Also, fewer people may choose to pay the higher toll rates and enter the MLs during periods with high toll rates. Next, an increase in the travel time variability will increase the U-ML trip probability as the result of the higher variance in the expected travel time. Also, the start sensor and end sensor are among the most important factors causing U-ML trips. However, they are categorical variables and their impact is not relatively positive or negative. Besides, their combination in the model is more critical than their individual impacts.

There were some limitations in this study. The first limitation is that many demographic attributes likely influence U-ML trips. Drivers may select MLs regardless of its toll or travel time based on their wealth and income. There was no available dataset on drivers' demographic characteristics.

REFERENCES

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM Sigkdd Explorations Newsletter, 6*(1), 20-29. doi:https://doi.org/10.1145/1007730.1007735

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*: Springer.

Blagus, R., & Lusa, L. (2013). SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinformatics, 14*(1), 1-106. doi:https://doi.org/10.1186/1471-2105-14-106

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Brownstone, D., Ghosh, A., Golob, T. F., Kazimi, C., & Van Amelsfort, D. (2003). Drivers' Willingness-To-Pay to Reduce Travel Time: Evidence from the San Diego I-15 Congestion Pricing Project. *Transportation Research Part A: Policy and Practice, 37*(4), 373-387. doi:https://doi.org/10.1016/S0965-8564(02)00021-6

Buckeye, K. (2012). Performance Evaluation of I-394 MnPASS Express Lanes in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board*(2278), 153-162. doi:https://doi.org/10.3141/2278-17

Burris, M., Nelson, S., Kelly, P., Gupta, P., & Cho, Y. (2012). Willingness to Pay for High-Occupancy Toll Lanes: Empirical Analysis from I-15 and I-394. *Transportation Research Record: Journal of the Transportation Research Board*(2297), 47-55. doi:https://doi.org/10.3141/2297-06

Burris, M., Sadabadi, K., Mattingly, S., Mahlawat, M., Li, J., Rasmidatta, I., & Saroosh, A. (2007). Reaction to the Managed Lane Concept by Various Groups of Travelers. *Transportation Research Record: Journal of the Transportation Research Board*(1996), 74-82. doi:https://doi.org/10.3141/1996-10

Burris, M., Spiegelman, C., Abir, A., & Lee, S. (2016). *Travelers' Value of Time and Reliability as Measured on Katy Freeway* (PRC 15-37 F). Retrieved from College Station, TX: http://tti.tamu.edu/documents/PRC-15-37-F.pdf

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: an overview. *Data Mining and Knowledge Discovery Handbook, Springer*, 853-867.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling TEchnique. *Journal of Artificial Intelligence Research (JAIR), 16*, 321-357. doi:https://doi.org/10.1613/jair.953

Devarasetty, P., Burris, M., & Shaw, W. (2012). Do Travelers Pay for Managed-Lane Travel as They Claimed They Would? Before-and-After Study of Travelers on Katy Freeway, Houston, Texas. *Transportation Research Record: Journal of the Transportation Research Board*(2297), 56-65. doi:https://doi.org/10.3141/2297-07

Devarasetty, P. C., Burris, M., Arthur, W., McDonald, J., & Muñoz, G. J. (2014). Can Psychological Variables Help Predict the Use of Priced Managed Lanes? *Transportation Research Part F: Traffic Psychology and Behaviour, 22*, 25-38. doi:https://doi.org/10.1016/j.trf.2013.10.006

Dobbin, K. K., & Simon, R. M. (2011). Optimally Splitting Cases for Training and Testing High Dimensional Classifiers. *BMC medical genomics, 4*(1), 31. doi:https://doi.org/10.1186/1755-8794-4-31

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:https://doi.org/10.1016/j.patrec.2005.10.010

Gardner, L. M., Bar-Gera, H., & Boyles, S. D. (2013). Development and Comparison of Choice Models and Tolling Schemes for High-Occupancy/Toll (HOT) Facilities. *Transportation Research Part B: Methodological, 55*, 142-153. doi:https://doi.org/10.1016/j.trb.2013.06.006

Hagenauer, J., & Helbich, M. (2017). A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications, 78*, 273-282. doi:https://doi.org/10.1016/j.eswa.2017.01.057

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning. *Springer*.

HCTRA (2009). *Retrieved from* https://www.hctra.org

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 6): Springer.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling Imbalanced Datasets: A Review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25-36.

Kwon, J., & Varaiya, P. (2008). Effectiveness of California's High Occupancy Vehicle (HOV) System. *Transportation Research Part C: Emerging Technologies, 16*(1), 98-115. doi:https://doi.org/10.1016/j.trc.2007.06.008

Lam, T. C., & Small, K. A. (2001). The Value of Time and Reliability: Measurement from a Value Pricing Experiment. *Transportation Research Part E: Logistics and*

Transportation Review, 37(2), 231-251. doi:https://doi.org/10.1016/S1366-5545(00)00016-8

Perez, B. G., Fuhs, C., Gants, C., Giordano, R., & Ungemah, D. H. (2012). *Priced Managed Lane Guide*. (FHWA-HOP-13-007). Federal Highway Administration, US Department of Transportation Retrieved from https://ops.fhwa.dot.gov/publications/fhwahop13007/fhwahop13007.pdf.

Reddy, P. D., Vaughn, K. M., Abdel-Aty, M. A., Jovanis, P. P., & Kitamura, R. (1995, 1995). *Design of a Recurrent Neural Network for Analyzing Route-Choice Behavior in the Presence of an Information System.* Paper presented at the Intelligent Transportation: Serving the User Through Deployment. Proceedings of the 1995 Annual Meeting of ITS America.

Sekhara, C., & Madhu, E. (2016). Multimodal Choice Modeling Using Random Forest Decision Trees. *International Journal for Traffic & Transport Engineering*, *6*(3). doi:http://dx.doi.org/10.7708/ijtte.2016.6(3).10

Sullivan, E. (2000). *Continuation Study to Evaluate the Impacts of the SR 91 Value-Priced Express Lanes: Final Report*. California Polytechnic State University, San Luis Obispo.

Tierney, K., Decker, S., Proussaloglou, K., Rossi, T., Ruiter, E., & McGuckin, N. (1996). *Travel Survey Manual*. U.S. Environmental Protection Agency, Federal Highway Administration, US Department of Transportation Retrieved from http://ntl.bts.gov/lib/4000/4500/4529/1392.pdf.

Train, K. (2003). *Discrete Choice Methods with Simulation*: Cambridge University Press.

Yang, H., Kitamura, R., Jovanis, P. P., Vaughn, K. M., & Abdel-Aty, M. A. (1993). Exploration of Route Choice Behavior with Advanced Traveler Information Using Neural Network Concepts. *Transportation*, 20(2), 199-223.

Xiao, Z., Wang, Y., Fu, K., & Wu, F. (2017). Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. *ISPRS International Journal of Geo-Information*, *6*(2), 57. doi:https://doi.org/10.3390/ijgi6020057

Zhu, X. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*: Igi Global.