Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model

Article *in* Transportation Research Record Journal of the Transportation Research Board · May 2018

DOI: 10.1177/0361198118773556

CITATIONS
119

READS
1,391

2 authors:

Fangru Wang
Georgia Institute of Technology
9 PUBLICATIONS **238** CITATIONS

SEE PROFILE

Catherine L. Ross
Georgia Institute of Technology
116 PUBLICATIONS **1,541** CITATIONS

SEE PROFILE

Predicting Travel Mode Choices in the Delaware Valley Region with Multinomial Logit Model and Extreme Gradient Boosting Model

Fangru Wang (Corresponding author)
Ph.D. Candidate
School of City and Regional Planning
Georgia Institute of Technology
760 Spring Street, Suite 213
Atlanta, Georgia, 30308
Phone: 646-453-9906
Email: fangru@gatech.edu

Catherine L. Ross, Ph.D.
Harry West Professor of City and Regional Planning / Civil and Environmental Engineering
Director, Center for Quality Growth and Regional Development
Georgia Power Professor of Excellence
760 Spring Street, Suite 213
Georgia Institute of Technology
Atlanta, Georgia, 30308
Phone: 404-385-5130 Fax: 404-385-5127
Email: catherine.ross@design.gatech.edu

Submission Date: February 22nd, 2018
Word Count: 5,508 Words + 2 Figures + 2 Large Tables + 2 Small Tables = 7,432

ABSTRACT

The multinomial logit (MNL) model and its variations have been dominating the travel mode choice modeling field for decades. Advantages of the MNL model include its elegant closed-form mathematical structure and its interpretable model estimation results based on random utility theory, while its main limitation is the strict statistical assumptions. Recent computational advancement has allowed easier application of machine learning models to travel behavior analysis, though research in this field is not thorough or conclusive. In this paper, we explore the application of the extreme gradient boosting (XGB) model to travel mode choice modeling and compare its result with an MNL model, using the Delaware Valley 2012 regional household travel survey data. The XGB model is an ensemble method based on the decision-tree algorithm and it has recently received a great deal of attention and use because of its high machine learning performance. The modeling and predicting results of the XGB model and the MNL model are compared by examining their multi-class predictive errors. We found that the XGB model has overall higher prediction accuracy than the MNL model especially when the dataset is not extremely unbalanced. The MNL model has great explanatory power and it also displays strong consistency between training and testing errors. Multiple trip characteristics, socio-demographic traits, and built environment variables are found to be significantly associated with people's mode choices in the region, but mode-specific travel time is found to be the most determinant factor for mode choice.

Key words: travel mode choice modeling; multinomial logit regression; machine learning; extreme gradient boosting model; Delaware Valley region

INTRODUCTION

Travel mode choice modeling is one of the most studied areas in travel behavior research and is a critical step in the travel demand forecasting process. Travel mode choices may vary by many different factors, including the level of service of certain travel modes, travelers' socio-demographic traits, and the built environment characteristics associated with each trip. In recent times we have seen the emergence of new travel modes like ride-sourcing and autonomous vehicles. These and other emerging transportation technologies and the availability of new travel modes will bring about changes in travel behavior including mode choice. Moreover, the emerging transport technologies may come with new data and information sources that may result in changes in the methods we use to analyze travel behavior and forecast travel mode choices.

The multinomial logit (MNL) model and its variations, such as the nested logit model, have been dominating the travel mode choice modeling field for decades. Advantages of the MNL model include its elegant closed-form mathematical structure and its interpretable model estimation result based on random utility theory. Its main limitation is the strict statistical assumptions, such as the independence of irrelevant alternatives (IIA), which often require very careful model specification and well-formatted data structures. Another limitation of the MNL model is that when the dataset is very unbalanced, meaning that different classes are represented very unequally, the estimation of the MNL model may be biased resulting in particularly higher prediction error for classes with smaller shares (*1*). Recent exploration of applying machine learning models to travel mode choice modeling has shown its merits and most of the exiting studies that compare machine learning models with statistical models suggest machine learning models could achieve similar or even higher prediction accuracy. The main strength of most machine learning models is that they do not have strict statistical assumptions and can therefore be applied to different data structures with great flexibility. Some machine learning techniques, such as ensemble methods, are found to be able to improve prediction accuracy of weak learners, such as low-depth decision tree models. In this exploratory research, we employ an extreme gradient boosting model and an MNL model to predict travel mode choices in the Delaware Valley (DV) region, and compare the results of the two models by comparing their average multi-class predicting errors. The extreme gradient boosting (XGB) model is a tree-based ensemble method that has recently gained great reputation by consistently wining machine learning

competitions on Kaggle (https://www.kaggle.com/). In this paper, we discuss the predictive results of the two models and compare and document the model implementation processes and implications for future travel mode choice modeling.

The rest of the paper is divided into four sections. Literature Review summarizes what factors have been identified as associated with travel behavior and reviews past studies that applied machine learning models to travel mode choice modeling. The Data and Methodology section explains how the two models are implemented and describes data sources. Results compare the prediction accuracy of the two models and provide interpretation of model estimation results. The Conclusions section evaluates the two models, summarizes the findings and points to potential efforts to improve travel mode choice modeling.

LITERATURE REVIEW

Extensive studies have been conducted to explore the factors associated with people's travel mode choices and what modeling structures should be applied to mode choice modeling. This literature review provides only a glimpse of the numerous studies by summarizing their findings.

Travel Mode Choice Modeling Techniques

Modeling travel mode choices is never an easy task as people's choice of travel modes intertwines with many different factors. Travel mode choice has been extensively modeled with random utility maximization theory for decades. One of the most commonly applied model is the MNL model. Stopher (*2*) and McFadden (*3*) are the earliest contribution to applying the MNL model to travel behavior modeling. The detailed model structure and applications were thoroughly discussed in the book by Ben-Akiva & Lerman (*4*). The MNL model captures the underlying mode choice process with utility maximization assumptions that travelers are rational decision makers who are fully informed and are able to choose the mode that has the largest utility for them. Though there are limitations in using utility maximization to represent the mode choice process (*5*), MNL models are effective in quantifying the effects of trip characteristics on people's mode choice. MNL model has a closed form mathematical estimation that was computable even decades ago. Another merit of the MNL model is that it can always replicate the shares of different classes.

The recent advancement in computation power has brought interest in applying machine learning techniques to modeling travel mode choices. Karlaftis & Vlahogianni (*6*) comprehensively review the differences and similarities of using statistical methods versus neural network models in transportation research. They concluded the merit of the neural network model is its flexibility in dealing with complex datasets and its great predictive power, while the most significant challenge is the lack of explanatory power compared to conventional statistical models. Similar to the merits of neural network models, most machine learning models do not have strict statistical assumptions behind the model estimation and are thus more flexible. Some machine learning techniques, such as bagging and ensemble methods, are often found to perform well with unbalanced datasets meaning that different classes account for very unequal shares.

There are not many existing studies that applied machine learning models to predicting travel mode choices. Among the limited number of such studies, all of them find that machine learning models can achieve at least similar prediction accuracy compared to conventional statistical models. Several studies find that the machine learning methods are substantially better (*7–11*). Machine learning models and statistic models are not comparable regarding their interpretation power, as most machine learning models do not allow meaningful interpretation. The commonly used machine learning methods in modeling travel mode choice includes decision trees (*7*, *12–15*), neural networks (*8*, *9*, *11*, *13*, *15*, *16*), support vector machines (*7*, *8*, *11*), and random forests (*10*). Some models' predictive power are improved by combining other machine learning concepts such as fuzzy sets and ensemble methods (*8*, *13*, *17*). Only three of the existing studies on applying machine learning methods to travel mode choice modeling use data from the U.S., as far as the authors can determine. The lack of understanding about the application of machine learning methods to travel mode choice modeling generally and specifically in the U.S. calls for more research and exploration.

Factors Associated with Travel Mode Choices

The factors associated with travel mode choices can be roughly categorized into three types: trip characteristics, personal/household factors, and neighborhood variables. Trip characteristics, often measured by level of service and trip-specific factors, can directly influence people's mode choices. Level of service is often measured by travel time and travel cost and they vary by travel modes (*18–21*). Frank et al. (*22*) found that travel time is the most important factor that influences people's choice of travel modes in the Puget Sound region, among many other built environment and socio-demographic variables. The departure time of a trip is also found to be related to people's mode choice (*23*) and it is correlated with the travel time of certain modes. Trip purpose is another important factor that directly affects people's mode choice. Most of the existing mode choice studies focus on commuting trips or home-based non-work trips and some others focus on special trip purposes or traveler groups, such as students' travel (*24–26*), trips to the airport (*27*), shopping trips (*28*) etc..

Personal and household variables reflect the socio-economic and demographic characteristics of travelers. Income, vehicle ownership, household size, number of workers, age, and gender, are found to be closely related to a traveler's mode choice in different studies. Income and vehicle ownership are directly determinant on people's choice of driving. Household size and number of workers also have effect on the mode choices of the household members and work-related trips. Special population groups, like the elderly, low-income, women, and disabled people, often yield special needs of travel options. The elderly are often found to have distinguishable activity and mode choice patterns (*29–32*). Similar to the elderly whose mobility is impaired by aging, low-income population's mobility is constrained by economic disadvantages and limited travel options in the area where they can afford to live (*33*, *34*). Female travelers are also found to present distinct travel behavior compared to men, though the findings are mixed. Cervero, (2002) (*35*) found that female travelers were more dependent on cars, probably because women are often undertaking both working and house work that require chain trips between work, shops, and child-care centers. Conversely, it is found that women have greater willingness to reduce car use and potential stronger preference for public transportation (*36*).

There are numerous studies of the relationship between built environment factors and travel behavior. Ewing & Cervero (37) developed a comprehensive review of studies of the relationship between travel behavior and the built environment and categorized the built environment factors into five "D" categories, including density, diversity, design, destination accessibility, and distance to transit. Policy and planning have been leveraged on these positive relationships between these built environment factors and people's mode choices to promote compact development. Land use is an indispensable component of the built environment and many studies have identified the association between land use measures and mode choice. Land use related metrics include the dominant land use type, land use mix, and the distance to certain land use types such as retail, etc. (22, 35, 38–40).

DATA AND METHODOLOGY

Data Preparation

Our major data source is the Delaware Valley Regional Planning Commission (DVRPC) household travel survey data collected in the Delaware Valley region in 2012. The Delaware Valley region consists of nine

counties in Pennsylvania and New Jersey (

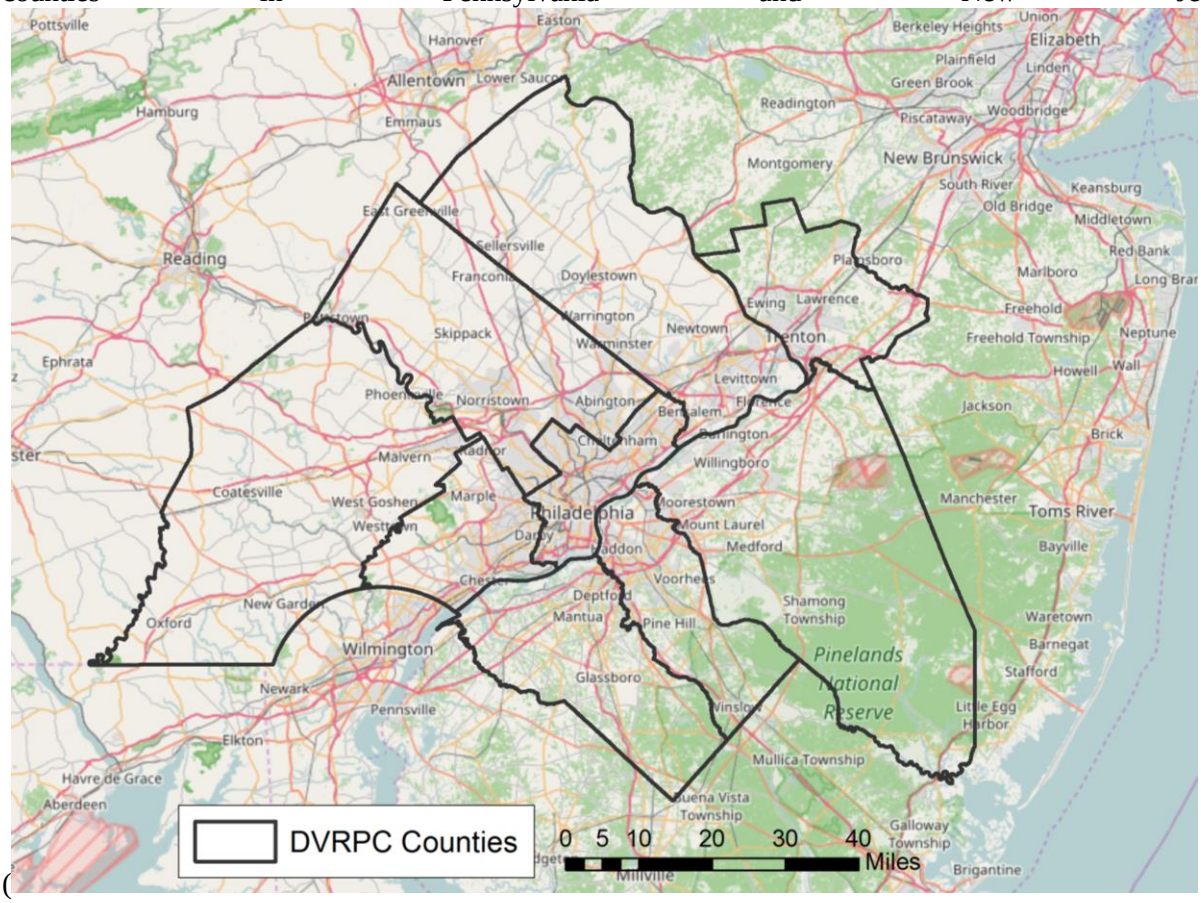

FIGURE 1). The dataset contains 81,940 unique trip records, but more than 20 thousand of them miss the information of travel mode and about 6 thousand miss the information of income levels. After the data cleaning process that removes the trips with missing values in the dependent and independent variables we want to model, there are 51,910 trips made by car (driver or passenger), biking, walking, or transit that we use for the mode choice analysis.

TABLE 1 summarizes the variables included in our model development process, their explanation, and data sources. The dependent variable is mode choice indicating whether the trip is made by car (driving or passenger), biking, walking, or transit. The independent variables include three types of factors suggested by the literature, including trip characteristics, personal/household features, and neighborhood-level variables. Most of the variables are constructed using publicly available datasets, as shown in the table, except that the travel time of alternative transportation modes are requested using the Google Maps Distance Matrix API. The DVRPC household travel survey data contains the reported travel time and modeled travel time for every trip, but it only has the travel time for the "chosen" travel modes. In order to implement discrete choice models, we also need to know the travel time by the "unchosen" modes to form a complete choice set for each individual. The Google Maps Distance Matrix API provides estimated travel time by driving, biking, walking, and transit at different times of day. We thus used it to request the estimated travel time of all the "unchosen" cases by specifying the departure time of each trip and the trip origin/destination (centroids of traffic analysis zones (TAZs) in the DVRPC region). For the chosen travel modes, we chose to use the modeled travel time, which are travel time between recorded departure TAZ and arrival TAZ modeled using DVRPC's travel demand model (TIM 2.2) based on mode choice. We chose to use the modeled travel time rather the reported travel time because reported travel time are often known as error-prone and also because both the "modeled" travel times and the travel times we acquired from Google API

are estimated using TAZ centroids, which make it more consistent between the chosen and unchosen modes. We include 51,910 unique trips from the DVRPC household travel survey data, but our final data panel included 207,640 observations as each unique trip has four entries for the four different travel modes.

Models and Implementation

We employ the XGB model and the MNL model to analyze travel mode choices in the DVRPC region and compare the predictive powers of the models and their implications. The XGB model is a tree-based ensemble method, so it allows understanding the importance of the independent variables in determining the dependent variable. Details of model implementation are documented as follows.

Multinomial Logit Model

The MNL model is based on random utility theory that assumes the utility of choosing a certain travel mode is a random variable that travelers always want to maximize. We implemented the model with maximum likelihood estimation using the "mnlogit" package in R. Except that the variable of travel time (by mode) is set as a generic variable, all other variables are included in the model as individual-specific variables as their values vary by individual travelers. We selected a set of independent variables when developing the MNL model using the whole dataset, mainly according to three criteria: first, a variable's sign needs to be consistent with existing theory; second, the variable is statistically significant; and third, groups of variables were selected by developing chi-squared tests to compare whether adding a group of variables improves the goodness of fit of the model. The MNL model prediction was implemented by conducting a Monte Carlo simulation process.

Extreme gradient boosting Model

The Extreme gradient boosting (XGB) Model is a tree-based ensemble method that can be used for both regression and classification. The XGB model was first proposed by Friedman (*41*). It is an ensemble method that is built upon iteratively growing low-depth decision trees based on the idea of additive training. To put it simple, in an XGB model, each low-depth decision tree is built to minimize a defined loss function, but each time the estimation puts more weights on the cases that are wrongly predicted by previously developed trees. The final model result is collectively determined by the results of all the developed trees. We implemented the XGB model using the "XGboost" package in R. In addition to estimating the model based on the independent variables we included, training the XGB model relies on tuning some hyperparameters for optimizing the model's performance. "Hyperparameter" is a term that can be used interchangeably with parameters, but in machine learning, it more often refers to the parameters that need to be set by the modeler or need to be tuned before the final model is learned based on independent and dependent variables. Hyperparameters directly influence the performance of a model and are often tuned using cross-validation techniques in machine learning. Seven hyperparameters need to be specified in the XGB model, including:

(1) "nrounds": the maximum number of iterations (similar to the number of trees to grow in our case);
(2) "eta": it controls the learning rate;
(3) "max_depth": it controls the depth of the tree;
(4) "min_child_weight": it controls the minimum value of sum of instance weight of a node;
(5) "subsample": it controls the number of samples supplied to a tree;
(6) "colsample_bytree": it controls the number of variables supplied to a tree;
(7) "max_delta_step": it controls regularization and is used to avoid overfitting.

The hyperparameters of "nrounds", "eta", and "max_depth" directly influence whether the model will overfit or underfit the data. To optimize the performance of the XGB model and also to avoid the overfitting problem. We first tuned the parameters of "nrounds" and "eta" together using grid-search and found that "eta" = 0.3 always produces the best model performance. The larger the value of "nrounds", the

more iteration the model will run, which will more likely result in the overfitting problem. We found that the hyperparameter of "nrounds" of 50 can achieve great model predicting performance yet does not generate big gaps between training and testing errors, a sign that the model is overfitting, so we set nrounds to 50. Then we used 10-fold cross-validation to tune all other parameters except "max_delta_step" in every run. The "max_delta_step" was set as 2 as it improves the performance of predicting rare case mode choices (e.g. trips made by biking). We set the maximum value that the hyperparameter "max_depth" can take to 10 to avoid overfitting, and specific values of "max_depth" were automatically set using cross-validation in every run. During the parameter tuning processes, we set the loss function as the multiclass classification error rate that is calculated as number of cases wrongly predicted divided by the number of all cases. In order to make the performance of the XGB model and the MNL model comparable, we included almost the same set of independent variables in both models, except that travel distance is included in the XGB model but not in the MNL model due to correlation between travel time and travel distance.

RESULTS

Descriptive Statistics

Four travel modes are included in the mode choice modeling analysis for the DVRPC region, including car (driving or passenger), biking, walking, and transit. The total number of unique trips included in our analysis after removing missing and outlier values is 51,910, among which 43,196 are made by car, 511 are made by biking, 5,467 are made by walking, and 2,736 are made by transit. One commonly encountered challenge of modeling travel mode choices is the issue of unbalanced datasets. In this case, the majority of trips are made by car while only about 1% trips are made by biking and only about 5% are made by transit. The dataset is very unbalanced and may cause the MNL and XGB models perform bad especially when predicting mode choices with smaller shares, such as biking and transit. TABLE 2 presents the descriptive statistics of all the variables used in the analysis.

Comparing Prediction Accuracy of the MNL Model and the XGB Model

Each of the two models were run 100 times to compare their average prediction accuracy and robustness to data changes. For each run, the dataset was randomly split into a training subset (75% of the data) and a testing subset (25% of the data) and the training errors and testing errors are averaged for the 100 times run and summarized in TABLE 3. In order to compare the predictive power of the MNL model and the XBG model, we examined the average total errors of the two models and also compare their prediction error for each of the four travel modes. The total error is calculated as the number of trips that are predicted to have the wrong mode choice out of the total number of trips. The mode-specific prediction error is calculated as the number of trips wrongly predicted out of the total number of trips made by that mode.

In addition to running the MNL model and the XGB model with the complete datasets, we also tested how the two models perform when biking, which has the smallest mode share, is removed from the data. Therefore, we also run each of the two models 100 times for a subset of the data that only contains the travel modes of car, walking, and transit. The results of the two models run with complete datasets and subsets of the data are shown in Table 3 respectively. The predicted mode shares of the two models are also presented in Table 3.

As TABLE 3 shows, for the complete choice set of four alternatives, both models show good overall prediction accuracy while the XBG model has lower errors than the MNL model. The XGB model has a total training error of 3.5% and a total testing error of 15.5%, while the MNL model has a total training error of 18.2% and a total testing error of 18.3%. This means that the XGB model has an overall prediction accuracy of 94.5% while the MNL model's overall prediction accuracy is only 92.7%. When the alternative of biking is removed from the dataset, both models perform even better, but the XGB model performs more significantly better compared to the result running with the complete dataset. When biking is removed, The

XGB model has a total training error of 3.7% and a total testing error of 10.5%, while the MNL model has a total training error of 17.8% and a total testing error of 17.9%.

Regarding mode-specific prediction accuracy, both models did similarly worst in predicting mode choices of biking, which is not surprising, as biking has much lower mode share, only about 1%, compare to other travel modes in the DVRPC region. Nevertheless, the XGB model still has a much better performance when predicting the choice of walking while it has lower performance of predicting the choice of transit, compared to the MNL model. The XGB model's walk-specific testing error is only 28.7% but the MNL model has a walk-specific testing error high as 58.6%. The XGB model's transit-specific testing error is 80.7% while the MNL model achieves a transit-specific testing error of 66.7%. Interestingly, when biking is removed from the model, the predicting errors of the XGB model regarding all the three alternatives significantly improve, which are all significantly lower than that of the MNL model.

The MNL model can always replicate the shares of all alternatives and it is why the predicted mode shares of the MNL model are very similar to the observed ones, shown in Table 3. The XGB model also performs great regarding predicted mode shares, and the difference between the observed mode shares and the predicted ones are all smaller than 2%. In summary, the XGB model can lower the overall predicting error compared to the MNL model. Both models do not perform well when dealing with extremely unbalanced data (when biking is included), but the XGB model performs significantly better regarding overall mode choices and all three specific modes when biking is removed.

Explanatory Results

The result of the MNL model estimated using the whole dataset is shown in TABLE 4 for evaluating the models' validity. Overall, the model has a great goodness of fit, as the adjusted Rho-squared (McFadden R-squared) of this model is about 0.71, indicating that about 71% of the information contained in the data is explained by this model. This model uses car as the base mode for estimation and the signs of all the variables included in this model are consistent with theory.

Several variables of trip characteristics are significantly associated with mode choices. Compared to the choice of car, "home-based other" trips are negatively associated with biking and transit, but are positively associated with walking. Biking trips are more likely to be in morning peak hours and less likely to occur in evening peak hours compared to trips by car. Walking trips are less likely to be in evening peak hours and late night compared to automobile trips, probably due to safety concerns. People prefer transit to car during morning peak hours, probably wanting to avoid roadway congestion. Biking trips are associated with longer activities at trip origins and transit trips are associated with shorter activities.

Multiple personal and household characteristics are found statistically significant in the model. Travelers from larger households are negatively associated with all three alternative travel modes, biking, walking, and transit, compared to car. The number of vehicles per capita and availability of a driver's license are also negatively associated with the three modes. Low-income travelers are more likely to choose walking and transit, while high-income travelers are less likely to use transit. Female travelers are less likely to choose the three alternative travel modes, which is consistent with the finding from Cervero (2002) (*35*) that female are more dependent on car. Children, elderly people, and travelers with disability are also found to be more dependent on cars according to the model result. Parking cost reduces people's likelihood to choose driving and transit subsidy encourages people to choosing alternative travel modes other than car.

Many of the built environment variables listed in TABLE 1 are highly correlated, such as the density measures and the entropy measures, so only variables of population density at both trip origin and destination are included in the final model. Population density at a trip's both ends are positively associated with biking, walking, and transit, though the direction of causation is unclear.

As illustrated previously, the XGB model is a tree-based ensemble method that allows measuring the "importance" a variable has in forming the final rule-based classifications. The importance measures used by the XGB model is calculated by integrating three importance-related measures, including "gain", "cover", and "frequency". "Gain" measures the improvement in accuracy brought by a feature to the

branches it is on; "cover" measures the relative quantity of observations concerned by a feature; and "frequency" counts the number of times a feature is used in all generated trees (*42*). FIGURE 2 plots the top 10 most important variables for the XGB model. Travel time is the most important variable influencing people's mode choices and it has an importance measure that is far larger than other independent variables. Trip distance is the second important variable and travel time and trip distance together account for more than 80% of the importance that all independent variables have. Vehicles per capita, number of household members in the trip, and multiple built environment variables are also associated with travel mode choices.

CONCLUSIONS

People's travel mode choices intertwine with many different factors and may have noticeable changes as technological advances, new travel modes and new data sources are available. Modeling travel mode choice is a critical step in travel demand forecasting and may also face challenges and opportunities as new travel modes and data sources become available. In this paper, we apply a popular machine learning model to modeling travel mode choices in the Delaware Valley region and compared the results with an MNL model by examining their multi-class prediction errors. The paper is among the limited number of studies that explore machine learning models' application in travel mode choice modeling and has included a relatively comprehensive list of independent variables that are ready for practical use.

Both the XGB model and the MNL models show high prediction accuracy for travel mode choices in the DVRPC region, but both models perform poorly for predicting the choice of biking that has an extremely small share of 1% in the dataset. Nevertheless, when biking is removed from the dataset, the performance of the XGB model exceeds that of the MNL model's substantially. Interestingly, after biking is removed from the dataset, the XGB model substantially exceeds the performance of the MNL model not only in predicting the choices of all three modes together, but also by significantly improving the accuracy for predicting every individual mode.

The main advantage of the MNL model is its great interpretability and it maintains high consistency between the training and testing errors. This may imply that the MNL model is able to effectively capture some generalizable relationships between the dependent and independent variables, so the developed model structure could fit more universally. The MNL model directly contributes to understanding the relationships between people's travel mode choices and other factors, and is useful for variable selection and deriving policy implications.

Regarding the effort of developing and implementing the models, the XGB model and the MNL model have different strengths and challenges. The advantage of the XGB model is it has very little limitation on the data structure and model specification. For example, travel distance is included in the XGB model and Figure 2 has shown that travel distance is an important independent variable in estimating mode choice, but it cannot be included in the MNL model due to the issue of correlation between travel time and distance. Also, though the XGB model requires effort of tuning some hyperparameters to optimize the model's performance, the whole model fitting process requires less attention and effort compared to the MNL model that requires very careful model specification and testing to examine whether the assumptions hold. In contrast, the MNL model can easily avoid the overfitting issue. Especially for a very unbalanced dataset, the XGB model may not suffer from overfitting issue for all choices combined, but may have overfitting issue when predicting the choice with small shares. For example, in this analysis, the hyperparameters of the XGB models are tuned by minimizing the multi-class predicting error and the overfitting issue is controlled at the whole dataset level. However, since biking only accounts for 1% of the data, the tuned hyperparameters will likely result in overfitting for predicting biking choices. Therefore, machine learning techniques that can handle choice-specific overfitting control will be very helpful in this situation.

Future mode choice modeling efforts should consider using machine learning techniques or integrating some machine learning techniques to conventional statistical modeling. The modeling results of

this study clearly demonstrates the advantages of the XGB model and the MNL model in different respects. The XGB model has higher prediction accuracy than the MNL model especially when the dataset is not extremely unbalanced. The MNL model allows intuitive interpretation that machine learning methods cannot surpass. An easy way to combine the advantages of the two types of models is to use MNL model as a before-hand variable selection and interpretation tool, while use machine learning models or machine learning techniques to perform or improve the mode choice forecasting. Unbalanced data is a notorious problem in machine learning and the analysis of this research may suggest that when the smallest share of a dataset is larger than 5% the XGB model can have significantly better performance compared to the MNL model. Of course, this may not hold true for other datasets, but avoiding extremely unbalanced datasets by collecting more samples for small share modes may be useful for improving models' performance in predicting travel mode choices. Machine learning models that perform well with unbalanced datasets are useful in this case.

The study has its limitations. The travel mode choices are modeled at the trip level without considering trip chaining effects or tour-level factors. How to incorporate both trip-level and tour-level considerations into travel mode choice analysis has been explored with statistical models, but has not been researched for machine learning models. This might be the next step for facilitating real-world applications of machine learning models to travel mode choice modeling. Another limitation is that the models are only applied to the Delaware Valley region, so the generalizability of the findings will need to be confirmed by future research. Nevertheless, this paper reveals the great potential of using machine learning method for improving travel mode choice prediction accuracy and provides an early contribution to documenting the implementation of the new machine learning model and techniques.

TABLES

List of All Tables

TABLE 1. Variables Included in the Two Models, and Their Names, Explanation, and Data Sources

Variable Name	Explanation	Data Source
Travel mode	Travel mode (dependent variable): 1= car; 2 = walk; 3 = bike; 4 = fixed route transit	DVRPC 2012 travel survey data
Trip Variables		
Travel time	Travel time by mode in minutes	DVRPC 2012 travel survey data and Google Maps Distance Matrix API data for the unchosen mode.
Morning peak	The trip is happening in the morning peak hour (7am - 9am)	DVRPC 2012 travel survey data
Evening peak	The trip is happening in the evening peak hour (4pm - 6pm)	
Late night	The trip is happening in late night (9pm - 4am)	
Activity duration	Activity duration in minutes	
Tour HBW	Dummy variable: the tour purpose is home-based work	
Tour HBO	Dummy variable: the tour purpose is home-based other	
Tour stops	Number of stops in the tour	
HH member	Number of household members in the trip	
Personal/Household Variables		
License	The traveler has a driver's license	DVRPC 2012 travel survey data
Low income	The traveler is low income (annual income < $30,000)	
High income	The traveler is high income (annual income >= $100,000)	
Young	The traveler is younger than 16	
Elderly	The traveler is elderly than 65	
Disability	The traveler has disability	
Female	The traveler is female	
Household size	The traveler's household size	
Employed	The traveler is employed	
Student	The traveler is a student	
Park must pay	Employee must pay for workplace parking out-of-pocket	
Transit subsidy	Employer offers to subsidize/pay for part of transit fare	
Number of vehicles per capita	Number of vehicles per capita in the household	
Neighborhood data (Calculated at census tract level for the DVRPC region)		
O/D: Population density	Population density (persons per sqml) at the origin or destination	ACS 5-year estimate 2011-2015
O/D: employment density	Employment density (jobs per sqml) at the origin or destination	LEHD 2012 data
O/D: employment entropy	Employment diversity (3-category entropy*) at the origin or destination	LEHD 2012 data
O/D: job-housing balance	Employment population balance (entropy metric) at the origin or destination	Calculated using the ACS and LEHD data above
O/D: bus stop density	Bus stop density (# bus stops per sqkm) at the origin or destination	Calculated using GTFS data
O/D: rail station density	Rail station density (# subway density per sqkm) at the origin or destination	Calculated using GTFS data
O/D: land use entropy	Land use entropy index (calculated as residential and commercial two-category entropy)	DVRPC land use GIS data

*Employment entropy is calculated as an entropy index considering three types of employment: retail, service, and finance.

TABLE 2. Mode Share and Descriptive Statistics

Travel Mode Share	**Car**	**Bike**	**Walk**	**Transit**	**Total**
Number of unique trips	43,196 (83.20%)	511 (1.00%)	5,467 (10.5%)	2,736 (5.30%)	51,910 (100.00%)
Continuous Variables					
Variable Name	**Mean**	**S.D.**	**Minimum**	**Maximum**	
Travel time	126.82	260.63	0	1772.22	
Tour stops	0.7	1.2	0	19	
Activity duration	124.8	173.2	1	1215	
HH member	0.4	0.7	0	6	
HH_WORK	1.4	1	0	6	
Vehicle per capita	0.8	0.4	0	8	
O/D: population density	23,760	66,285	0	506,607.20	
O/D: employment density	170,082	1,052,925	6.4	9,515,692	
O/D: bus stop density	184	693.1	0	4,790	
O/D: job-housing balance	0.7	0.2	0	1	
O/D: rail station density	12	64.6	0	501	
O/D: land use entropy	0.7	0.2	0	1	
Dummy Variables					
Variable Name	**Number of 1**		**Number of 0**		
Tour HBW	18,687	36%	33,223	64%	
Tour HBO	29,588	57%	22,322	43%	
Morning peak	10,901	21%	41,009	79%	
Evening peak	12,458	24%	39,452	76%	
Late night	2,076	4%	49,834	96%	
Household size	8,305	16%	43,605	84%	
Low income	6,748	13%	45,162	87%	
High income	20,244	39%	31,666	61%	
Female	28,550	55%	23,360	45%	
Young	6,229	12%	45,681	88%	
Elderly	11,939	23%	39,971	77%	
License	44,123	85%	7,787	15%	
Disability	1,557	3%	50,353	97%	
Park must pay	4,671	9%	47,239	91%	
Transit subsidy	2,076	4%	49,834	96%	
Student	8,824	17%	43,086	83%	

TABLE 3. Average Training Errors, Testing Errors, and Predicted Mode Shares of 100 Model Runs

	MNL Model - 4 Alternative				**XGB Model - 4 Alternative**			
	Training Error		Testing Error		Training Error		Testing Error	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Total	18.2%	0.000004	18.3%	0.000007	3.5%	0.001651	15.5%	0.000148
Car	8.8%	0.000003	8.8%	0.000009	1.8%	0.000548	8.6%	0.000056
Biking	94.6%	0.000139	94.9%	0.000291	19.4%	0.059651	95.2%	0.000636
Walking	58.1%	0.000054	58.6%	0.000163	5.5%	0.006854	28.7%	0.001171
Transit	66.5%	0.000102	66.7%	0.000370	1.8%	0.000548	80.7%	0.000650

	MNL Model - 3 Alternative				**XGB Model - 3 Alternative**			
	Training Error		Testing Error		Training Error		Testing Error	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Total	17.8%	0.000003	17.9%	0.000008	3.7%	0.000485	10.5%	0.000065
Car	8.3%	0.000002	8.3%	0.000007	0.6%	0.000047	3.5%	0.000005
Walking	58.5%	0.000063	58.5%	0.000153	8.5%	0.005302	27.8%	0.000495
Transit	67.2%	0.000092	67.4%	0.000302	0.6%	0.000047	65.9%	0.001284

Predicted Mode Shares	**MNL Model 4 Alternative**		**MNL Model 3 Alternative**		**XGB Model 4 Alternative**		**XGB Model 3 Alternative**	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Total	100.0%	100.0%	99.4%	95.3%	100.0%	100.0%	98.4%	95.6%
Car	83.3%	83.0%	82.9%	83.5%	84.0%	84.5%	84.2%	83.5%
Biking	1.0%	1.0%	1.0%	0.3%				
Walking	10.6%	10.5%	10.4%	9.0%	10.6%	10.7%	9.8%	9.1%
Transit	5.1%	5.5%	5.2%	2.5%	5.4%	4.8%	4.4%	3.0%

TABLE 4. MNL Model Result

Generic variable	Est.	S.E	Sig.						
Travel time	-0.0047	0.0001	***						
Individual-specific Variables	Biking			Walking			Transit		
	Est.	S.E	Sig.	Est.	S.E	Sig.	Est.	S.E	Sig.
(Intercept)	-0.74	0.25	**	1.30	0.10	***	1.39	0.12	***
Tour HBO	-0.26	0.11	*	0.07	0.04	.	-0.71	0.06	***
Morning peak	0.38	0.12	**	0.06	0.04		0.39	0.06	***
Evening peak	-0.02	0.12		-0.24	0.04	***	0.09	0.06	
Late night	0.11	0.23		-0.27	0.09	**	-0.12	0.13	
Activity duration	0.0007	0.0003	**	0.0001	0.0001		-0.0006	0.0001	***
Household size	-0.50	0.04	***	-0.42	0.02	***	-0.41	0.02	***
Vehicle per capita	-3.03	0.14	***	-1.96	0.06	***	-2.27	0.07	***
Low income	-0.23	0.15		0.30	0.05	***	0.68	0.06	***
High income	-0.19	0.11	.	0.03	0.04		-0.27	0.06	***
Female	-0.87	0.10	***	-0.18	0.03	***	-0.22	0.05	***
Young	0.16	0.23		-0.31	0.08	***	-1.05	0.11	***
Elderly	-0.89	0.16	***	-0.62	0.05	***	-0.39	0.07	***
License	-0.46	0.19	*	-1.25	0.07	***	-1.84	0.08	***
Disability	-0.96	0.36	**	-0.62	0.10	***	0.03	0.11	
Park must pay	1.17	0.11	***	0.78	0.05	***	1.10	0.06	***
Transit subsidy	1.28	0.13	***	0.80	0.07	***	0.96	0.08	***
O: population density	6.8E-06	4.9E-07	***	8.4E-06	2.7E-07	***	7.8E-06	2.9E-07	***
D: population density	8.2E-06	5.0E-07	***	9.8E-06	2.9E-07	***	9.9E-06	3.1E-07	***
Pseudo R-squared (McFadden R-squared) = 0.701									

Note: "O" and "D" stand for "origin" and "destination" of a trip respectively.

Sig. indicates significance codes: '***' = 0 – 0.001; '**' = 0.001 – 0.01; '*'= 0.01 – 0.05; '.'= 0.01 – 0.1; ' '= 0.1 – 1

FIGURES

List of All Figures

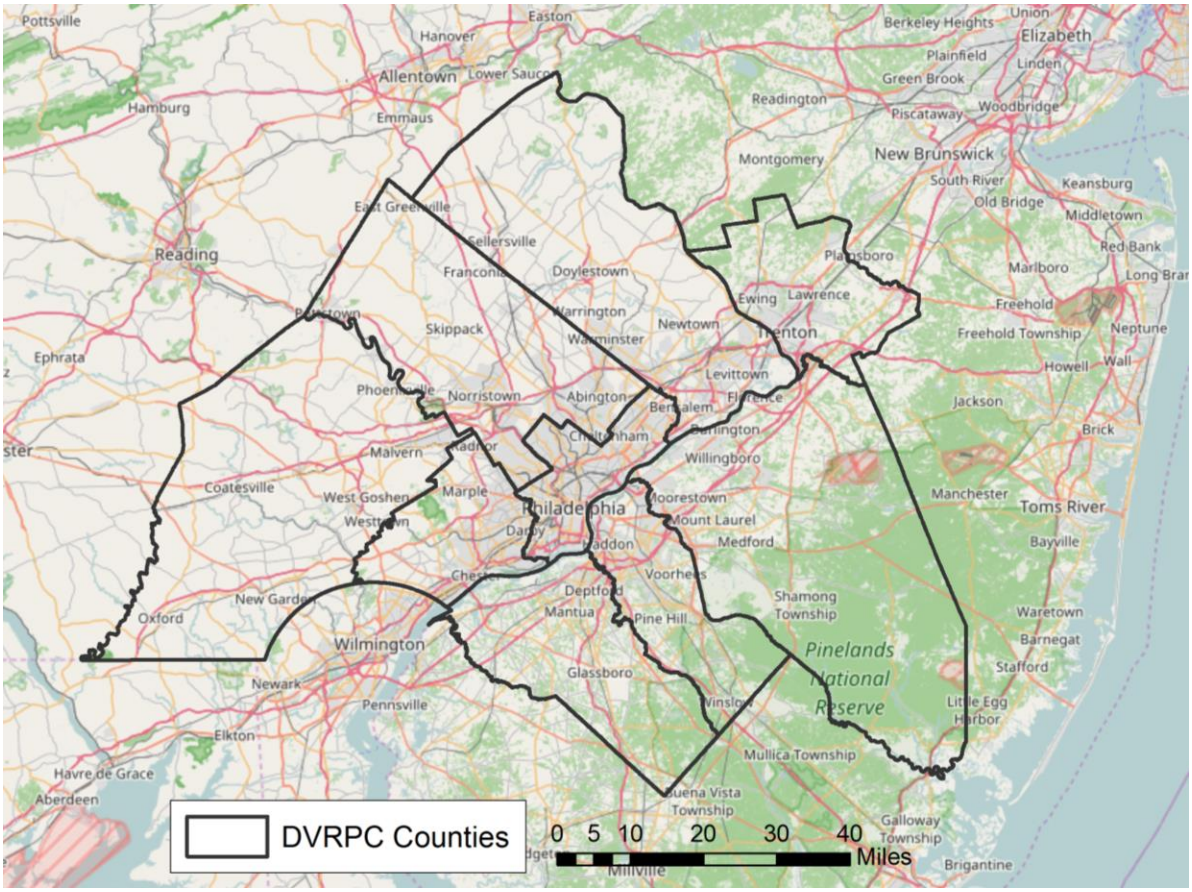

FIGURE 1. Study Area of the DVRPC Region

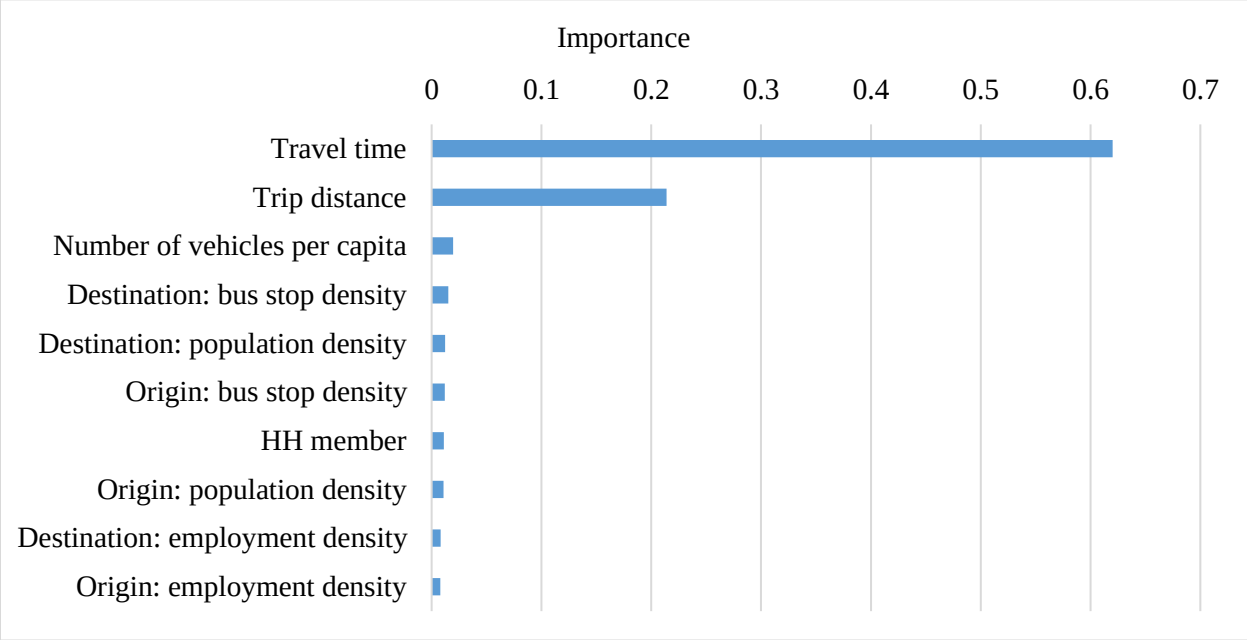

FIGURE 2. Top 10 Most Important Independent Variables of the XGB Model

The authors confirm contribution to the paper as follows: study conception and design: Fangru Wang and Catherine L. Ross; data collection: Fangru Wang; analysis and interpretation of results: Fangru Wang; draft manuscript preparation: Fangru Wang and Catherine L. Ross. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. King, G., and L. Zeng. Logistic Regression in Rare Events Data. *Political analysis*, Vol. 9, No. 2, 2001, pp. 137–163.
2. Stopher, P. A Multinomial Extension of the Binary Logit Model for Choice of Mode of Travel. *Northwestern University, unpublished*, 1969.
3. McFadden, D. The Measurement of Urban Travel Demand. *Journal of Public Economics*, Vol. 3, No. 4, 1974, pp. 303–328. https://doi.org/10.1016/0047-2727(74)90003-6.
4. Ben-Akiva, M., and S. R. Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT press, 1985.
5. Gärling, T. Behavioural Assumptions Overlooked in Travel Choice Modelling. *Travel Behaviour Research: Updating the state of play*, 1998, pp. 3–18.
6. Karlaftis, M. G., and E. I. Vlahogianni. Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights. *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 3, 2011, pp. 387–399.
7. Biagioni, J. P., P. M. Szczurek, P. C. Nelson, and A. Mohammadian. Tour-Based Mode Choice Modeling: Using an Ensemble of (Un-) Conditional Data-Mining Classifiers. 2008.
8. Omrani, H., O. Charif, P. Gerber, A. Awasthi, and P. Trigano. Prediction of Individual Travel Mode with Evidential Neural Network Model. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2399, 2013, pp. 1–8. https://doi.org/10.3141/2399-01.
9. Rao, P. S., P. K. Sikdar, K. K. Rao, and S. L. Dhingra. Another Insight into Artificial Neural Networks through Behavioural Analysis of Access Mode Choice. *Computers, Environment and Urban Systems*, Vol. 22, No. 5, 1998, pp. 485–496.
10. Sekhar, C. R., Minal, and E. Madhu. Mode Choice Analysis Using Random Forrest Decision Trees. *Transportation Research Procedia*, Vol. 17, 2016, pp. 644–652. https://doi.org/10.1016/j.trpro.2016.11.119.
11. Zhang, Y., and Y. Xie. Travel Mode Choice Modeling with Support Vector Machines. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2076, 2008, pp. 141–150.
12. Celikoglu, H. B. Application of Radial Basis Function and Generalized Regression Neural Networks in Non-Linear Utility Function Specification for Travel Mode Choice Modelling. *Mathematical and Computer Modelling*, Vol. 44, No. 7, 2006, pp. 640–658.
13. Shukla, N., J. Ma, R. Wickramasuriya, N. N. Huynh, and P. Perez. Tour-Based Travel Mode Choice Estimation Based on Machine learning and Fuzzy Techniques. 2015.
14. Wets, G., K. Vanhoof, T. Arentze, and H. Timmermans. Identifying Decision Structures Underlying Activity Patterns: An Exploration of Machine learning Algorithms. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1718, 2000, pp. 1–9.
15. Xie, C., J. Lu, and E. Parkany. Work Travel Mode Choice Modeling with Machine learning: Decision Trees and Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1854, 2003, pp. 50–61.
16. Hensher, D. A., and T. T. Ton. A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 36, No. 3, 2000, pp. 155–172.
17. Vythoulkas, P. C., and H. N. Koutsopoulos. Modeling Discrete Choice Behavior Using Concepts from Fuzzy Set Theory, Approximate Reasoning and Neural Networks. *Transportation Research Part C: Emerging Technologies*, Vol. 11, No. 1, 2003, pp. 51–73.
18. Bhat, C. R. Work Travel Mode Choice and Number of Non-Work Commute Stops. *Transportation Research Part B: Methodological*, Vol. 31, No. 1, 1997, pp. 41–54. https://doi.org/10.1016/S0191-2615(96)00016-1.

19. Bowman, J. L., and M. E. Ben-Akiva. Activity-Based Disaggregate Travel Demand Model System with Activity Schedules. *Transportation Research Part A: Policy and Practice*, Vol. 35, No. 1, 2001, pp. 1–28.
20. Cervero, R., and M. Duncan. Walking, Bicycling, and Urban Landscapes: Evidence from the San Francisco Bay Area. *American journal of public health*, Vol. 93, No. 9, 2003, pp. 1478–1483.
21. Dissanayake, D., and T. Morikawa. Investigating Household Vehicle Ownership, Mode Choice and Trip Sharing Decisions Using a Combined Revealed Preference/stated Preference Nested Logit Model: Case Study in Bangkok Metropolitan Region. *Journal of Transport Geography*, Vol. 18, No. 3, 2010, pp. 402–410. https://doi.org/10.1016/j.jtrangeo.2009.07.003.
22. Frank, L., M. Bradley, S. Kavage, J. Chapman, and T. K. Lawton. Urban Form, Travel Time, and Cost Relationships with Tour Complexity and Mode Choice. *Transportation*, Vol. 35, No. 1, 2008, pp. 37–54.
23. Ye, X., R. M. Pendyala, and G. Gottardi. An Exploration of the Relationship between Mode Choice and Complexity of Trip Chaining Patterns. *Transportation Research Part B: Methodological*, Vol. 41, No. 1, 2007, pp. 96–113.
24. Ewing, R., W. Schroeer, and W. Greene. School Location and Student Travel Analysis of Factors Affecting Mode Choice. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1895, 2004, pp. 55–63.
25. Klöckner, C. A., and T. Friedrichsmeier. A Multi-Level Approach to Travel Mode Choice – How Person Characteristics and Situation Specific Aspects Determine Car Use in a Student Sample. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 14, No. 4, 2011, pp. 261–277. https://doi.org/10.1016/j.trf.2011.01.006.
26. McDonald, N. C. Critical Factors for Active Transportation to School Among Low-Income and Minority Students: Evidence from the 2001 National Household Travel Survey. *American Journal of Preventive Medicine*, Vol. 34, No. 4, 2008, pp. 341–344. https://doi.org/10.1016/j.amepre.2008.01.004.
27. Gupta, S., P. Vovsha, and R. Donnelly. Air Passenger Preferences for Choice of Airport and Ground Access Mode in the New York City Metropolitan Region. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2042, 2008, pp. 3–11.
28. Bhat, C. R. Analysis of Travel Mode and Departure Time Choice for Urban Shopping Trips. *Transportation Research Part B: Methodological*, Vol. 32, No. 6, 1998, pp. 361–371.
29. Georggi, N. L., and R. M. Pendyala. An Analysis of Long-Distance Travel Behavior of the Elderly and the Low-Income. University of South Florida, 2000.
30. Golob, T. F. A Simultaneous Model of Household Activity Participation and Trip Chain Generation. *Transportation Research Part B: Methodological*, Vol. 34, No. 5, 2000, pp. 355–376. https://doi.org/10.1016/S0191-2615(99)00028-4.
31. Kim, S., and G. Ulfarsson. Travel Mode Choice of the Elderly: Effects of Personal, Household, Neighborhood, and Trip Characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1894, 2004, pp. 117–126.
32. Kuppam, A. R., and R. M. Pendyala. A Structural Equations Analysis of Commuters' Activity and Travel Patterns. *Transportation*, Vol. 28, No. 1, 2001, pp. 33–54.
33. Giuliano, G. Low Income, Public Transit, and Mobility. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1927, 2005, pp. 63–70.
34. Murakami, E., and J. Young. *Daily Travel by Persons with Low Income*. US Federal Highway Administration Washington, DC, 1997.
35. Cervero, R. Built Environments and Mode Choice: Toward a Normative Framework. *Transportation Research Part D: Transport and Environment*, Vol. 7, No. 4, 2002, pp. 265–284. https://doi.org/10.1016/S1361-9209(01)00024-4.

36. Matthies, E., S. Kuhn, and C. A. Klöckner. Travel Mode Choice of Women The Result of Limitation, Ecological Norm, or Weak Habit? *Environment and Behavior*, Vol. 34, No. 2, 2002, pp. 163–177. https://doi.org/10.1177/0013916502034002001.
37. Ewing, R., and R. Cervero. Travel and the Built Environment. *Journal of the American planning association*, Vol. 76, No. 3, 2010, pp. 265–294.
38. Pinjari, A. R., R. M. Pendyala, C. R. Bhat, and P. A. Waddell. Modeling the Choice Continuum: An Integrated Model of Residential Location, Auto Ownership, Bicycle Ownership, and Commute Tour Mode Choice Decisions. *Transportation*, Vol. 38, No. 6, 2011, p. 933.
39. Rajamani, J., C. Bhat, S. Handy, G. Knaap, and Y. Song. Assessing Impact of Urban Form Measures on Nonwork Trip Mode Choice after Controlling for Demographic and Level-of-Service Effects. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1831, 2003, pp. 158–165.
40. Zhang, M. The Role of Land Use in Travel Mode Choice: Evidence from Boston and Hong Kong. *Journal of the American Planning Association*, Vol. 70, No. 3, 2004, pp. 344–360. https://doi.org/10.1080/01944360408976383.
41. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics*, 2001, pp. 1189–1232.
42. Understand Your Dataset with XGBoost — Xgboost 0.6 Documentation. http://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html#feature-importance. Accessed Jul. 20, 2017.