Exploring the effectiveness of different modeling approaches for predicting travel mode choice: An analysis of Machine Learning versus Econometric methods

Mohammadjavad Javadinasr [1], Sina Asgharpour [2], Motahare Mohammadi [3], Abolfazl (kouros) Mohammadian [4], Joshua Auld [5]

1. Ph.D. Candidate, Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W. Taylor St, Chicago, IL 60607. Email: *Mjavad2@uic.edu*
2. Ph.D. Candidate, Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W. Taylor St, Chicago, IL 60607. Email: *sasgha3@uic.edu*
3. Ph.D. Candidate, Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W. Taylor St, Chicago, IL 60607. Email: mmoham73@uic.edu
4. Professor, Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 W. Taylor St, Chicago, IL 60607. Email: *Kouros@uic.edu*
5. Ph.D. Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439. Email: *jauld@anl.gov*

Abstract

The selection of transportation modes has been a popular topic in transportation planning, and it has been studied for several decades using random utility optimization. However, recent developments in machine learning techniques have opened up new opportunities for accurate prediction. In this study, we used data from the Chicago including activity and travel records for 12,000 households in a 24-hour period. Our objective was to develop travel mode choice models for commute trips through estimating a multinomial logit model to identify the factors that influence people's choices of commute travel mode. Notably, our study is the first in Chicago to consider seven travel modes, including walking, biking, walking to transit, driving to transit, auto driver, auto passenger, and TNC. Additionally, we employed a machine learning classifier to model the mode choice problem and compared its performance with the econometric model.

Introduction

Investigating the contributing factors in forming people’s decision to select a travel mode and ultimately finding accurate and reliable predictions for the share of each mode in the network is critical to transportation planning and development (Ortúzar and Willumsen 2011). Traditionally, the most widely employed approach to address the mode choice problem has been based on the concept of Random Utility Maximization (RUM) which is usually formulated through the multinomial logit models (MNL) (McFadden 1973). The simple parameter estimation procedure offered by the closed-form mathematical structure of the MNL models has resulted in widespread adoption of them in travel behavior studies (Hagenauer and Helbich 2017). Moreover, the MNL models can shed light on the contribution of each explanatory variable on the dependent variable (i.e., the outcome choice) by providing the marginal effect values which are very useful, especially for policy implications and analysis. More recently and especially in the past decade, the emergence of big data and the computational power and insights brought by machine learning (ML) techniques have led to incorporating data-driven approaches into travel demand modeling. ML models act as powerful classifiers with high accuracy power. Instead of using a prespecified utility function, ML techniques rely on the relationships between the explanatory variables and the outcome class (i.e., the mode choice) and are able to capture complex and nonlinear relationships (Hillel et al. 2021). However, ML techniques are mostly considered black box tools, meaning they provide their predictions without giving enough information on how they have reached such results.

In this study, we develop travel mode choice models for commute trips in Chicago by employing a dataset from the Chicago Metropolitan Agency for Planning (CMAP) travel tracker survey conducted in 2018-19. The motivation for this study was to estimate a mode choice module for POLARIS, which is an agent-based modeling framework for integrated travel demand and network simulations in Chicago metropolitan area (Auld et al. 2016). One of our contributions is to estimate a mode choice model through the inclusion of emerging modes such as TNC by using the most updated travel survey available in Chicago. To do so, we use a multinomial logit model as well as a machine learning classifier to develop models that represent the choice between several modal options, including Walk, Bike, Auto driver, Auto passenger, Walk to public transit, Drive to public transit (i.e., park & ride) and TNCs (e.g., Uber, and Lyft). Moreover, using the MNL model, we calculate the value of time, which is a critical policy variable for transportation-related decisions for auto, transit, and TNC modes. Finally, we will compare the results of our econometric and machine-learning mode choice models.

- **Methodology**

In this study, we utilized two approaches to address the mode choice problem. Firstly, we employed the well-established multinomial logit (MNL) model (i.e., econometric analysis) to characterize the underlying factors that affect the decision of people in selecting commute travel modes in Chicago. Secondly, we utilized Gradient Boosting as a classification tools.

The Gradient Boosting method is an advanced version of decision-tree-based models, firstly developed by Friedman (2001). This method *boosts* the accuracy of a given learning algorithm by fitting a set of models. As a result, the ensemble of models outperforms the single model. The Gradient Boosting model, firstly, fits a simple decision tree to the train data, which usually has a poor fitting performance. Subsequently, it trains another decision tree to improve the error term (i.e., the difference between prediction and observation) in the previous tree. This process will continue to iterate until it gets to the minimized error term.

- **Results and Discussion**

- ***Data***

The main data source for this study was obtained from a comprehensive travel survey conducted by Chicago metropolitan Agency for Planning (CMAP). This activity-based survey was implemented between August 2018 and April 2019 and includes the activity and travel information of 12,391 households on an assigned travel day. Considering the significant impact of work trips on the transportation system and traffic flow, in this study, we are only focused on mandatory commute trips to model the mode choice of individuals. After the cleaning process, the final dataset for commute trips includes 11442 trips that are used for the modeling steps of this study. We considered seven travel modes including, walking, biking, transit, park and ride (i.e., drive to transit), auto driver, auto passenger, and TNC (e.g., Uber, and Lyft) in the choice set. With almost 36% share, the auto driver had the highest rank among all other modes in commute trips in Chicago.

- ***Multinomial Logit Model***

Table 1 presents the estimation results for the multinomial logit model for seven commute travel modes. In order to validate our econometric mode choice approach, we calculated the scalar measures of fit, i.e., McFadden's R2, for the estimated multinomial logit model which turned out to be 0.22. As can be seen in Table 2, the travel times corresponding to each mode are significantly negative in all utility functions which is consistent with the literature on mode choice modeling. People who possess a graduate degree are more likely to use active modes (i.e., walking or biking) for their commute trips. In the auto driver mode, the travel cost (e.g., the cost of gasoline) and the parking cost are two significant factors that negatively affect the utility of this mode. Moreover, people who own their homes and people with a higher number of vehicles in the household are more likely to use auto driver mode for commuting. This finding makes sense and can act as a proxy for the wealth of individuals.

Table 1. the results of the estimation of the MNL mode choice model for commute trips.

Variable	**Walk**	**Bike**	**Auto driver**	**Auto Passenger**	**Transit**	**Park & ride**	**TNC**
Constant	4.018***	REF[1]	3.229***	2.086***	4.174***	2.056***	-0.230
Walk Time	-0.140**						
Graduate Degree	2.433***	2.100***					
Flexible Work	1.045***						
Bike Time		-0.073*					
# Household Vehicle		-0.179**	0.385***				-0.503**
Employment		1.094***					2.624***
Auto Travel Time			-.0332**	-.0332***			-0.033**
Auto Cost			-.126***	-.126***			
Homeownership			.477***				
Parking Cost			-.00101*				
Transit Time					-0.0188*		
Transit Cost					-0.126**		

Transit Access Time					-0.076**		
Transit Egress time					-.077***		
Under 18					-2.586**		
Park & Ride Time						-0.018**	
Park & Ride Cost						-0.126**	
TNC Wait Time							-0.190**
TNC Cost							-0.101**

Note: 1) REF means the reference point.

* Significant at 90%, ** Significant at 95%, *** Significant at 99%.

Regarding the transit mode, the time and cost of the travel as well as the access/egress time revealed to be the significant determining factors. Moreover, travel cost and time are the only significant variables in the utility of the park & ride mode. One important finding is the significance of the TNC wait time in our MNL model suggesting in-vehicle travel time and travel cost are not the only alternative-specific features that influence the choice of TNCs. This finding can have policy implications, especially for companies such as Uber and Lyft to take measures that can lead to lower waiting times and increase the share of TNC mode in people's commute trips.

Another piece of information that can be extracted from table 2 is the value of times. The value of time (VOT) is the marginal substitution of the time associated with the cost of each mode (Abdel-Aal 2017). According to table 1, the VOT for auto modes (both auto driver and auto passenger) is $ 15.8 per hr., for transit modes (both transit and park & ride) is $ 8.9 per hr., and for TNC is $ 19.7 per hr. These findings are consistent with the literature (Lam and Small 2001) given that it is expected for TNC to have higher VOT compared to auto and transit modes, and for transit mode to have the lowest among the specified modes.

- ***Machine learning model***

The entire dataset has 11,441 records with 68 explanatory variables and a target variable (i.e., mode choice class). Since some key variables contain a high percent of missing values (e.g., *transit time* with 44% missing), complete case analysis leads to loss of information. This issue is not the case for MNL model since it only needs features to be available in the corresponding utility functions. For instance, *access time to transit* is only available for trips that have transit as an available option in the choice set (please note that the available choice set is not the same for all observations). To tackle this issue in Machine Learning methods, we employed imputation by K-nearest Neighbors. Among different K values, K=10 (i.e., 10 neighbors) led to better imputation results in terms of bias reduction.

To train the models, we split the dataset into 80% train and 20% test datasets. Moreover, we employed 5-fold cross validation to calibrate the hyperparameters of the models.

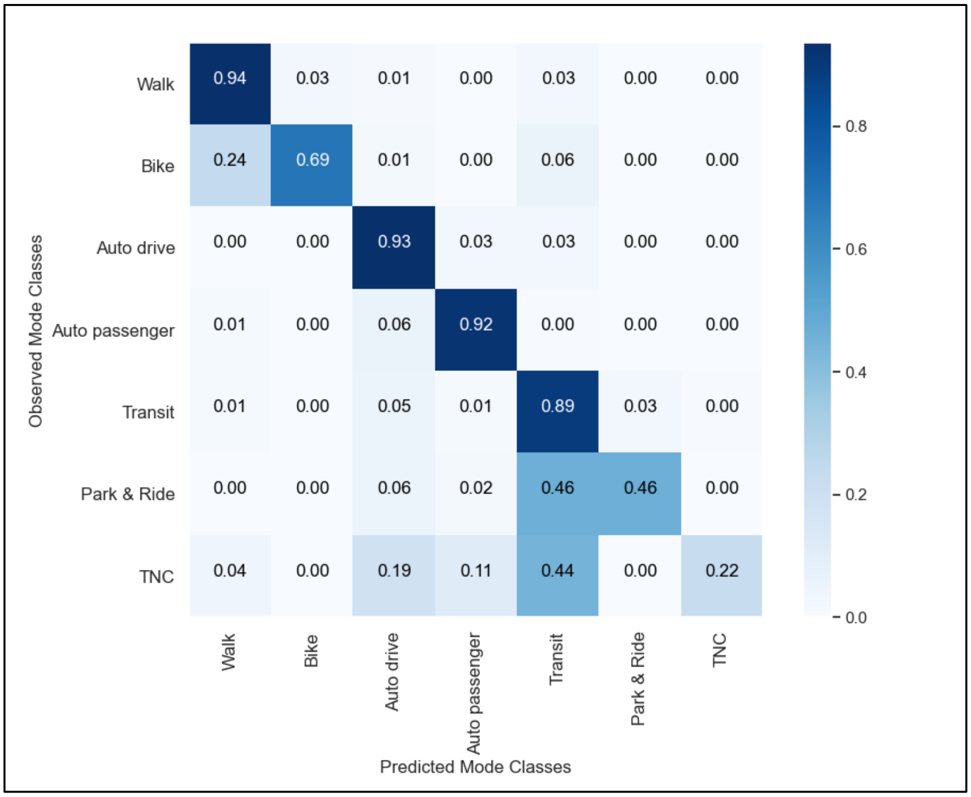
2 shows the accuracy of the individual and final models over 5-fold train, and test datasets, respectively. As shown, the fluctuation (i.e., coefficient of variation) of the accuracy of the Gradient Boosting model on train data is marginal implying a consistent performance. The model has a mean accuracy of 0.877 on train data. Although the training datasets of the MNL and Machine Learning model are different, the R^2 statistics can reflect the train accuracy of the MNL model. According to Table , the R^2 of the MNL is 0.22 implying the poor performance of the MNL compared to the Gradient Boosting model. This difference highlights the advantage of the machine learning models over econometric models in terms of performance.

Table 2. Accuracy of Employed Methods on Train (5-fold Cross Validation) and Test Data

		Gradient Boosting
Train Data Accuracy (Cross Validation)	Fold 1	0.894
	Fold 2	0.898
	Fold 3	0.891
	Fold 4	0.898
	Fold 5	0.898
	Mean **(Coeff. of Variation)**	**0.896** **(0.0036)**
Test Data Accuracy	**Accuracy**	**0.897**

To understand the pattern of misclassification among classes, we used confusion matrix. Confusion matrix constitutes the frequency table showing the distribution of the predicted and observed (true) classes for the test data instances. Therefore, it provides a validation of a classification model with respect to ground-truth information. 1 illustrates the confusion matrices for the Gradient Boosting model. In this figure, the cell numbers represent the percentage of each observed class falling into a predicted class. Therefore, the summation of the cells in each row is 1. Besides, this percentage is also represented by color intensity. Based on 1, in general, our ML model has a good performance in class prediction.

In our employed dataset, *bike*, *park & ride*, and *TNC* classes are the minority classes with the associated frequencies between 1% to 3% from whole data while the other classes have a frequency over 17%. Predicting minority classes is one of the challenges of ML models. According to Figure 1, the Gradient Boosting model can correctly predict 69% of instances with *bike*, 46% of *park & ride*, and 22% of *TNCs*. This pattern pertains to the low accuracy of the machine learning models in predicting minority classes in unbalanced data. In these cases, the MNL models can be a good alternative to predict the minority classes, especially noting the benefits of the MNL model in interpretation and policy evaluation.



	Walk	Bike	Auto drive	Auto passenger	Transit	Park & Ride	TNC
Walk	0.94	0.03	0.01	0.00	0.03	0.00	0.00
Bike	0.24	0.69	0.01	0.00	0.06	0.00	0.00
Auto drive	0.00	0.00	0.93	0.03	0.03	0.00	0.00
Auto passenger	0.01	0.00	0.06	0.92	0.00	0.00	0.00
Transit	0.01	0.00	0.05	0.01	0.89	0.03	0.00
Park & Ride	0.00	0.00	0.06	0.02	0.46	0.46	0.00
TNC	0.04	0.00	0.19	0.11	0.44	0.00	0.22

Figure 1: Confusion Matrix of Gradient Boosting Model

As a remarkable advantage, Gradient Boosting provides the *feature importance* facilitating to understand the individual contribution of the explanatory variables. In figure 2, feature importance of top 16 variables is illustrated. As shown, among all variables, *Auto Travel Time* has the most significant effect on the target variable (i.e., mode choice) with the feature importance of 0.164. Furthermore, *Auto Availability*, *Walk Time*, *Number of Passengers*, and *Transit Cost* are the other significant variables affecting individuals' mode choice. As shown, mainly time and cost attributes constitute the important factors affecting individuals' mode choice in the Gradient Boosting model which is consistent with the results of the MNL. Most of the significant variables that appeared in the feature importance are also identified by the MNL model except for *Auto Availability*, *# of Auto Passengers*, *Destination in CBD*. This trend confirms that although MNL has a poor prediction performance, it can detect the most significant attributes affecting the target variable.

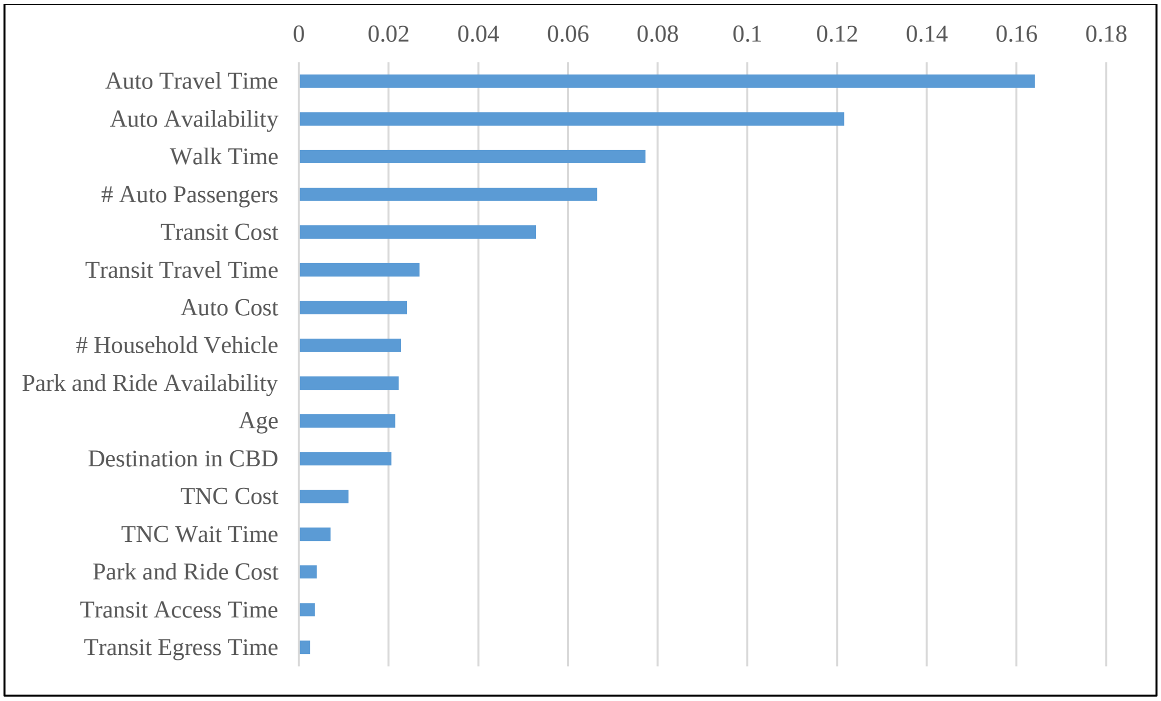


Figure 1: Feature Importance of Gradient Boosting Model

- ***Discussion***

In this study, we developed travel mode choice models for commute trips in Chicago by employing a dataset from CMAP travel survey conducted in 2018-19. We considered seven travel modes including walking, biking, walking to transit, driving to transit (i.e., park and ride), auto driver, auto passenger, and TNC in the choice set. We employed econometric (i.e., Multinomial Logit Model) and machine learning (i.e., Gradient Boosting) techniques to address the mode choice problem. According to our results, travel times and costs are the most contributing factors in forming individuals' decision to select a mode. In comparison, the MNL framework did not require imputing all the missing values due to the different utilities associated with the different choice sets defined for each observation. Moreover, consistent with the previous literature, we also corroborated the higher accuracy of ML techniques compared to the MNL model. Nonetheless, the accuracy power of the ML model was not equal in predicting all outcome classes. In general, in the presence of unbalanced data, the precision of our ML models to predict the minority classes decreased. The MNL model provided useful and interpretable alternative-specific variables related to the minority classes (such as the TNC wait time) that are very valuable for policy implications.

ACKNOWLEDGMENT

"The work described in this paper is sponsored by the U.S. Department of Energy (DOE) Vehicle Technologies Office (VTO) under the Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility Laboratory Consortium, an initiative of the Energy Efficient Mobility Systems (EEMS) Program. The submitted manuscript has been created by the UChicago Argonne, LLC, Operator of Argonne National Laboratory (Argonne). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357.

REFERENCES

Abdel-Aal, M. M. M. (2017). "Value of time determination for the city of Alexandria based on a disaggregate binary mode choice model." *Alexandria Engineering Journal*, Elsevier, 56(4), 567–578.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., and Zhang, K. (2016). "POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations." *Transportation Research Part C: Emerging Technologies*, Elsevier Ltd, 64, 101–116.

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, Institute of Mathematical Statistics, 29(5), 1189–1232.

Hagenauer, J., and Helbich, M. (2017). "A comparative study of machine learning classifiers for modeling travel mode choice." *Expert Systems with Applications*, Pergamon, 78, 273–282.

Hillel, T., Bierlaire, M., Elshafie, M. Z. E. B., and Jin, Y. (2021). "A systematic review of machine learning classification methodologies for modelling passenger mode choice." *Journal of Choice Modelling*, Elsevier, 38, 100221.

Lam, T. C., and Small, K. A. (2001). "The value of time and reliability: measurement from a value pricing experiment." *Transportation Research Part E: Logistics and Transportation Review*, Pergamon, 37(2–3), 231–251.

McFadden, D. (1973). "Conditional logit analysis of qualitative choice behavior."

Ortúzar, J. de D., and Willumsen, L. (2011). *Modelling transport*.