

# Modeling Stated Preference for Mobility-on-Demand Transit: A Comparison of Machine Learning and Logit Models

Xilei Zhao<sup>a</sup>, Xiang Yan<sup>b,\*</sup>, Alan Yu<sup>c</sup>, Pascal Van Hentenryck<sup>a</sup>

<sup>a</sup>*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology*

<sup>b</sup>*Taubman College of Architecture and Urban Planning, University of Michigan*

<sup>c</sup>*Department of Electrical Engineering and Computer Science, University of Michigan*

---

## Abstract

Logit models are usually applied when studying individual travel behavior, i.e., to predict travel mode choice and to gain behavioral insights on traveler preferences. Recently, some studies have applied machine learning to model travel mode choice and reported higher out-of-sample prediction accuracy than conventional logit models (e.g., multinomial logit). However, there has not been a comprehensive comparison between logit models and machine learning that covers both prediction and behavioral analysis. This paper aims at addressing this gap by examining the key differences in model development, evaluation, and behavioral interpretation between logit and machine-learning models for travel-mode choice modeling. To complement the theoretical discussions, we also empirically evaluated the two approaches on stated-preference survey data for a new type of transit system integrating high-frequency fixed routes and micro-transit. The results show that machine learning can produce significantly higher predictive accuracy than logit models and are better at capturing the nonlinear relationships between trip attributes and mode-choice outcomes. On the other hand, compared to the multinomial logit model, the best-performing machine-learning model, the random forest model, produces less reasonable behavioral outputs (i.e. marginal effects and elasticities) when they were computed from a standard approach. By introducing some behavioral constraints into the computation of behavioral outputs from a random forest model, however, we obtained better results that are somewhat comparable with the multinomial logit model. We believe that there is great potential in merging ideas from machine learning and conventional statistical methods to develop refined models for travel-behavior research and suggest some possible research directions.

**Keywords:** machine learning, mixed logit, mobility-on-demand, stated preference survey, travel mode choice, public transit

---



---

\*Corresponding Author.

*Email addresses:* xilei.zhao@isye.gatech.edu (Xilei Zhao), jacobyan@umich.edu (Xiang Yan), alayu@umich.edu (Alan Yu), pvh@isye.gatech.edu (Pascal Van Hentenryck)

## 1. Introduction

Emerging shared mobility services, such as car sharing, bike sharing, ridesourcing, and micro-transit, have rapidly gained popularity across cities and are gradually changing how people move around. Predicting individual preferences for these services and the induced changes in travel behavior is critical for transportation planning. Traditionally, travel behavior research has been primarily supported by discrete choice models, most notably the logit family such as the multinomial logit (MNL), the nested logit model and the mixed logit model. In recent years, as machine learning has become pervasive in many fields, there has been a growing interest in its application to mode individual choice behavior. A number of recent publications compared the results of machine-learning methods and logit models in modeling travel mode choices, with a particular emphasis on their respective out-of-sample predictive accuracy. These studies have often found that machine-learning models such as neural networks (NN) and support vector machines (SVM) perform better than logit models (e.g. Xie et al., 2003; Zhang and Xie, 2008).

However, the existing literature comparing logit models and machine learning for modeling travel mode choice has a number of important gaps. First, the comparisons were usually made between the MNL model, the simplest logit model, and machine-learning algorithms of different complexity. In cases where the assumption of independence of irrelevant alternatives (IIA) is violated, such as when panel data (i.e., data containing multiple mode choices made by the same individuals) are examined, more advanced logit models such as the mixed logit model should be considered. Second, previous mode-choice studies rarely applied machine-learning models for behavioral analysis (e.g., examining variable importance, elasticities, and marginal effects) and compared the behavioral findings with those obtained by logit models. In mode-choice modeling applications, the behavioral interpretation of the results is as important as the prediction problem, since it offers valuable insights for transit planners and policymakers in order to prioritize the design of service attributes. Third, existing studies rarely discuss the fundamental differences in the application of machine-learning methods and logit models to travel-mode choice modeling. The notable differences between the two approaches in the input data structure and data needs, the modeling of alternative-specific attributes, and the form of predicted outputs carry significant implications for model comparison. These differences and their implications, although touched on by some researchers such as Omrani et al. (2013), have not been thoroughly examined.

This paper tries to bridge these gaps: It provides a comprehensive comparison of logit models and machine learning in modeling travel mode choices and also an empirical evaluation of the two approaches based on stated-preference (SP) survey data on a proposed mobility-on-demand (MOD) transit system, i.e., an integrated transit system that runs high-frequency buses along major corridors and operates on-demand shuttles in the surrounding areas (Mahéo et al., 2017). The paper first discusses the fundamental differences in the practical applications of the two types of methods, with a particular focus on the implications of the predictive performance of each framework and their capabilities to facilitate behavioral interpretations. Then, we compare the performance of two logit models (MNL and mixed logit) and seven machine-learning classifiers, including Naïve Bayes (NB), classification and

regression trees (CART), boosting trees (BOOST), bagging trees (BAG), random forest (RF), SVM, and NN, in predicting individual choices of four travel modes and their respective market shares. Moreover, we provide behavioral interpretations of the best-performing models for each approach and contrasts the findings. We find that machine learning can produce higher out-of-sample prediction accuracy than logit models. Moreover, machine learning is better at revealing nonlinear relationships between trip attributes and the choice output but may produce unreasonable behavioral outputs if the computation of marginal effects and elasticities follows a standard procedure. To tackle this problem we propose to incorporate certain behavioral constraints into the procedure of calculating marginal effects and elasticities for machine learning models, and the results are improved as expected.

The rest of the paper is organized as follows. The next section provides a brief review of the literature in modeling mode choices using logit and machine-learning models. Section 3 explains the fundamentals of the logit and machine-learning models, including model formulation and input data structures, model development and evaluation, and model interpretation and application. Section 4 introduces the data used for empirical evaluation and Section 5 describes the logit and machine-learning models examined and their specifications. Section 6 evaluates these models in terms of predictive capability and interpretability. Lastly, Section 7 concludes by summarizing the findings, identifying the limitations of the paper, and suggesting future research directions. Table 1 presents the list of abbreviations and acronyms used in this paper.

Table 1: List of Abbreviations and Acronyms.

MNL	Multinomial logit
NB	Naive Bayes
CART	Classification and regression trees
RF	Random forest
BOOST	Boosting trees
BAG	Bagging trees
SVM	Support vector machines
NN	Neural networks
AIC	Akaike information criterion
BIC	Bayesian information criterion
Min	Minimum
Max	Maximum
SD	Standard deviation
SP	Stated-preference
RP	Revealed-preference
IIA	Independence of irrelevant alternatives
PT	Public transit

## 2. Literature Review

The logit family is a class of econometric models based on random utility maximization (Ben-Akiva et al., 1985). Due to their statistical foundations and their capability to represent individual choice behavior realistically, the MNL model and its extensions have dominated travel behavior research ever since its formulation in the 1970s (McFadden, 1973). The MNL model is frequently challenged for its major assumption, the IIA property, and its inability to account for taste variations among different individuals. To address these limitations, modelers have developed important extensions to the MNL model such as the nested logit model and more recently the mixed logit model. The mixed logit model, in particular, has received much interest in recent years: Unlike the MNL model, it does not require the IIA assumption, can accommodate preference heterogeneity, and may significantly improve the MNL behavioral realism in representing consumer choices (Hensher and Greene, 2003).

Mode-choice modeling can also be viewed as a classification problem, providing an alternative to logit models. A number of recent publications have suggested that machine-learning classifiers such as CART, NN, and SVM are effective in modeling individual travel behavior (Golshani et al., 2018; Hagenauer and Helbich, 2017; Omrani, 2015; Omrani et al., 2013; Wang and Ross, 2018; Xie et al., 2003; Zhang and Xie, 2008). These studies generally found that machine-learning classifiers outperform traditional logit models in predicting travel-mode choices. For example, Xie et al. (2003) applied CART and NN to model mode choices for commuting trips taken by residents in the San Francisco Bay area. These machine-learning methods exhibited slightly better performance than the MNL model. Based on data collected in the same area, Zhang and Xie (2008) reported that SVM can predict commuter travel mode choice more accurately than NN and MNL.

It is not surprising that machine-learning classifiers can perform better than logit models in predictive tasks. Unlike logit models that make strong statistical assumptions (i.e. constraining the model structure and assuming a certain distribution in the error term *a priori*), machine learning allows for more flexible model structures, which can reduce the model’s incompatibility with the empirical data (Christopher, 2016; Xie et al., 2003). More fundamentally, the development of machine learning prioritizes predictive power, whereas advances in logit models are mostly driven by refining model assumptions, improving model fit, and enhancing the behavioral interpretation of the model results (Brownstone and Train, 1998; Hensher and Greene, 2003). In other words, the development of logit models prioritizes parameter estimation (i.e. obtaining better model parameter estimates that underline the relationship between the input variables and the output variable) and pay less attention to validating the model’s out-of-sample predictive capability (Mullainathan and Spiess, 2017). In fact, recent studies have shown that the mixed logit model, despite resulting in substantial improvements in overall model fit, often results in poorer prediction accuracy compared to the simpler and more restrictive MNL model (Cherchi and Cirillo, 2010).

While recognizing the superior predictive power of machine-learning models, researchers often think that they have weak explanatory power (Mullainathan and Spiess, 2017). In other words, machine-learning models are often regarded as “not interpretable.” Machine-learning studies rarely apply model outputs to facilitate behavioral interpretations, i.e., to

test the response of the output variable or to changes in the input variables in order to generate findings on individual travel behavioral and preferences (Karlaftis and Vlahogianni, 2011). The model outputs of many machine-learning models are indeed not directly interpretable as one may need hundreds of parameters to describe a deep NN or hundreds of decision trees to understand a RF model. Nonetheless, many of the behavioral analyses applied in logit-model studies, such as the evaluation of variable importance, marginal effects, and elasticities, can be similarly implemented in machine-learning models via techniques such as partial dependence plots and sensitivity analysis (Golshani et al., 2018). Examining these behavioral outputs from machine learning models could shed light on what factors are driving prediction decisions and also the fundamental question of whether machine learning is appropriate for behavioral analysis.

Prediction and behavioral analysis are equally important in travel behavior studies. While the primary goal of some applications is to accurately predict mode choices (and investigators are usually more concerned about the prediction of aggregate market share for each mode than about the prediction of individual choices), other studies may be more interested in quantifying the impact of different trip attributes on travel mode choices. To our knowledge, mode-choice applications that focus on behavioral outputs such as elasticity, marginal effect, value of time, and willingness-to-pay measures have received even more attention than those that focus on predicting individual mode choice or aggregate market shares in the literature. This paper thus extends current literature by comparing the behavioral findings from logit models and machine-learning methods, beyond the existing studies that primarily focus on their predictive accuracy.

Finally, this paper points out other differences in the practical applications of these two approaches that have bearings on model outputs and performance, including their input data structure and data needs, the treatment of alternative-specific attributes, and the forms of the predicted outputs. Discussions of these differences are largely absent from the current literature that compares the application of logit models and machine-learning algorithms in travel behavior research.

### 3. Fundamentals of the Logit and Machine-Learning Models

This section discusses the fundamentals of the logit and machine-learning models. Table 2 presents the list of symbols and notations used in the paper and Table 3 summarizes the comparison between logit and machine-learning models from various angles. The rest of this section describes this comparison in detail.

#### 3.1. Model Development

Logit models and machine-learning models approach the mode choice prediction problem from different perspectives. Logit models view the mode choice problem as individuals selecting a mode from a set of travel options in order to maximize their utility. Under the random utility maximization framework, the model assumes that each mode provides a certain level of (dis)utility to a traveler, and specifies, for each mode, a utility function with two parts: A component to represent the effects of observed variables and a random error

Table 2: List of Symbols and Notations Used in the Paper.

Symbols	Description
$K$	Total number of alternatives
$N$	Total number of observations
$P$	Total number of features
$\mathbf{X}$	Input data for logit models containing $P$ features with $N$ observations for $K$ alternatives
$\mathbf{X}_{k,p}$	Feature $p$ for alternative $k, k = 1, \dots, K$ of $\mathbf{X}$
$\mathbf{X}_{k,-p}$	All the features except $p$ for alternative $k, k = 1, \dots, K$ of $\mathbf{X}$
$\mathbf{X}_{ik}$	A row-vector for the $i$ th observation for alternative $k, k = 1, \dots, K$
$\mathbf{X}_k$	Input data for alternative $k$ , $\mathbf{X}_k = [\mathbf{X}_{.k1}; \dots; \mathbf{X}_{.kP}]$ where $\mathbf{X}_{.kp} = [X_{1kp}, \dots, X_{Nkp}]$
$\mathbf{X}_i$	The $i$ th observation of $\mathbf{X}$ , $\mathbf{X}_i = [\mathbf{X}_{i.1}, \dots, \mathbf{X}_{i.P}]$ where $\mathbf{X}_{i.p} = [X_{i1p}; \dots; X_{iKp}]$
$\mathbf{X}_p$	The feature $p$ of $\mathbf{X}$ , $\mathbf{X}_p = [\mathbf{X}_{1,p}, \dots, \mathbf{X}_{N,p}]$ where $\mathbf{X}_{i,p} = [X_{i1p}; \dots; X_{iKp}]$
$\mathbf{Z}$	Input data for machine-learning models containing $P$ features and $N$ observations
$\mathbf{Z}_p$	Feature $p$ of $\mathbf{Z}$
$\mathbf{Z}_{-p}$	All the features except $p$ of $\mathbf{Z}$
$\mathbf{Z}_i$	$i$ th observation of $\mathbf{Z}$ , $\mathbf{Z}_i = [Z_{i1}, \dots, Z_{iP}]$
$U_k(\mathbf{X}_k \beta_k)$	Utility function for mode $k$
$\beta_k$	Parameter vector for alternative $k$ of MNL model
$\beta$	Parameter matrix of MNL model, $\beta = [\beta_1, \dots, \beta_K]$
$\hat{\beta}$	Estimated parameter matrix of MNL model
$\varepsilon_k$	Random error for alternative $k$ of MNL model
$\mathbf{Y}$	Observed mode choice data
$\hat{Y}_i$	Estimated mode choice for observation $i$
$\theta$	Parameter or hyperparameter vector for machine-learning models
$\hat{\theta}$	Estimated parameter or hyperparameter vector
$f(\mathbf{Z} \theta)$	Trained machine-learning models using $\mathbf{Z}$ and $\theta$
$p_{ik}$	Probability of choosing alternative $k$ of observation $i$
$\hat{p}_{ik}$	Probability prediction for choosing alternative $k$ of observation $i$
$I_k(\hat{Y}_i)$	Indicator function that equals to 1 if $\hat{Y}_i = k$
$P_k(\mathbf{X} \hat{\beta})$	Aggregate level prediction for mode $k$ based on $\mathbf{X}$ and $\hat{\beta}$ for logit models
$Q_k(\mathbf{Z} \hat{\theta})$	Aggregate level prediction for mode $k$ based on $\mathbf{Z}$ and $\hat{\theta}$ for machine-learning models
$E(\cdot)$	Arc elasticity
$M(\cdot)$	Marginal effect
$\Delta$	Constant

term to represent the effects of unobserved factors (Ben-Akiva et al., 1985). For example, the utility of choosing mode  $k$  under the MNL model can be defined as

$$U_k(\mathbf{X}_k|\beta_k) = \beta_k^T \mathbf{X}_k + \varepsilon_k, \quad (1)$$

where  $\beta_k$  are the coefficients to be estimated and  $\varepsilon_k$  is the unobserved random error for choosing mode  $k$ . Different logit models are formed by specifying different types of error terms and different choices of coefficients on the observed variables. For instance, assuming a Gumbel distributed error term and fixed model coefficients (i.e., coefficients that are the same for all individuals) produces the MNL model (Ben-Akiva et al., 1985). In the MNL,

Table 3: Comparison Between Logit and Machine-Learning Models

	Logit Models	Machine-Learning Models
Model formulation	$U_k(\mathbf{X}_k \beta_k) = \beta_k^T \mathbf{X}_k + \varepsilon_k$ $p_{ik} = \frac{\exp(\beta_k^T \mathbf{X}_{ik})}{\sum_{p=1}^K \exp(\beta_p^T \mathbf{X}_{ik})}, k \in \{1, \dots, K\}$	$\mathbf{Y} = f(\mathbf{Z} \theta), \mathbf{Y} \in \{1, \dots, K\}$
Commonly used model type	MNL, Mixed logit, nested MNL, generalized MNL	CART, BAG, BOOST, RF, NB, SVM, NN
Prediction type	Class probability: $p_{i1}, \dots, p_{iK}$	Classification: $k, k \in \{1, \dots, K\}$
Data input	$\mathbf{X}$	$\mathbf{Z}$
Model topology	Layer structure	Layer structure, tree structure, case-based reasoning, etc.
Optimization method	Maximum likelihood estimation, simulated maximum likelihood	Back propagation, gradient descent, recursive partitioning, structural risk minimization, maximum likelihood, etc.
Evaluation criteria	(Adjusted) McFadden's pseudo $R^2$ , AIC, BIC	Resampling-based measures, e.g., cross validation
Individual-level mode prediction	$\arg \max_k (\hat{p}_{i1}, \dots, \hat{p}_{iK})$	$\hat{Y}_i$
Aggregate-level mode share prediction	$P_k(\mathbf{X}_k \hat{\beta}_k) = \sum_i \hat{p}_{ik}/N$	$Q_k(\mathbf{Z} \hat{\theta}) = \sum_i I_k(\hat{Y}_i)/N$
Variable importance	Standardized beta coefficients	Variable importance, computed by using Gini index, out-of-bag error, and many others
Variable effects	Sign and magnitude of beta coefficients	Partial dependence plots
Arc elasticity for feature $p$	$E(\mathbf{X}_{k,p}) = \frac{[P_k(\mathbf{X}_{k,-p}, \mathbf{X}_{k,p} \cdot (1+\Delta) \hat{\beta}_k) - P_k(\mathbf{X}_k \hat{\beta}_k)]/P_k(\mathbf{X}_k \hat{\beta}_k)}{\Delta}, k \in \{1, \dots, K\}$	$E(\mathbf{Z}_p) = \frac{[Q_k(\mathbf{Z}_{-p}, \mathbf{Z}_p \cdot (1+\Delta) \hat{\theta}_k) - Q_k(\mathbf{Z} \hat{\theta}_k)]/Q_k(\mathbf{Z} \hat{\theta}_k)}{\Delta}, k \in \{1, \dots, K\}$
Marginal effects for feature $p$	$M(\mathbf{X}_{k,p}) = \frac{P_k(\mathbf{X}_{k,-p}, \mathbf{X}_{k,p} + \Delta \hat{\beta}_k) - P_k(\mathbf{X}_k \hat{\beta}_k)}{\Delta}, k \in \{1, \dots, K\}$	$M(\mathbf{Z}_p) = \frac{Q_k(\mathbf{Z}_{-p}, \mathbf{Z}_p + \Delta \hat{\theta}_k) - Q_k(\mathbf{Z} \hat{\theta}_k)}{\Delta}, k \in \{1, \dots, K\}$

the probability of choosing alternative  $k$  for individual  $i$  is

$$p_{ik} = \frac{\exp(\beta_k^T \mathbf{X}_{ik})}{\sum_{p=1}^K \exp(\beta_p^T \mathbf{X}_{ik})}, \quad (2)$$

Given the beta coefficient, the MNL can be associated with the likelihood function

$$\mathbf{L}(\beta) = \prod_{i=1}^N \prod_{k=1}^K \left[ \frac{\exp(\beta_k^T \mathbf{X}_{ik})}{\sum_{p=1}^K \exp(\beta_p^T \mathbf{X}_{ik})} \right]. \quad (3)$$

Maximum likelihood estimation can then be applied to obtain the "best" utility coefficients  $\hat{\beta} = \arg \max_{\beta} \mathbf{L}(\beta)$ . By plugging  $\hat{\beta}$  into Eqn. (2), the *choice probabilities* for each mode can be obtained. More complex logit models, such as the mixed logit and nested logit, can be derived similarly from different assumptions about the coefficients and error distribution. However, these models are more difficult to fit: They generally do not have closed-form solutions for the likelihood function and require the simulation of maximum likelihood for various parameter estimations. Observe also that logit models have a layer structure, which maps the input layer  $\mathbf{X}_i$  to the output layer,  $[p_{i1}, \dots, p_{iK}]^T$ .

Machine-learning models, in contrast, view mode choice prediction as a *classification* problem: Given a set of input variables, predict which travel mode will be chosen. More precisely, the goal is to learn a target function  $f$  which maps input variables  $\mathbf{Z}$  to the output

target  $\mathbf{Y}$  ( $\mathbf{Y} \in \{1, \dots, K\}$ ) as

$$\mathbf{Y} = f(\mathbf{Z}|\boldsymbol{\theta}), \quad (4)$$

where  $\boldsymbol{\theta}$  represents the unknown parameter vector for parametric models like NB and the hyper-parameter vector for non-parametric models such as SVM, CART, and RF. Unlike logit models that predetermine a (usually) linear model structure and make specific assumptions for parameters and error distributions, many machine-learning models are nonlinear and/or non-parametric, which allows for more flexible model structures to be directly learned from the data. In addition, compared to logit models that maximize likelihood to estimate parameters, machine-learning models often apply different optimization techniques, such as back propagation and gradient descent for NN, recursive partitioning for CART, structural risk minimization for SVM. Moreover, while logit models have a layer structure, machine-learning models have different model topologies for different models. For example, tree-based models (CART, BAG, BOOST, and RF) all have a tree structure, whereas NN has a layer structure.

Furthermore, since the outputs of logit models are individual choice probabilities, it is difficult to compare the prediction with the observed mode choices directly. Therefore, when evaluating the predictive accuracy of logit models at the individual level, a common practice in the literature is to assign an outcome probability to the alternative with the largest outcome probability, i.e.,

$$\arg \max_k (\hat{p}_{i1}, \dots, \hat{p}_{iK}). \quad (5)$$

This produces the same type of output (i.e., the travel mode choice) as the machine-learning models. Besides the prediction of individual choices, logit models and machine-learning methods are often evaluated based on their capability to reproduce the aggregate choice distribution for each mode, i.e., the market shares of each mode. For logit models, the predicted market share of mode  $k$  is

$$P_k(\mathbf{X}_k|\hat{\beta}_k) = \sum_i \hat{p}_{ik}/N, \quad (6)$$

and, for machine-learning methods, it is given by

$$Q_k(\mathbf{Z}|\hat{\boldsymbol{\theta}}) = \sum_i I_k(\hat{Y}_i)/N. \quad (7)$$

The calibration of the logit models is targeted at approximating aggregate market shares, as opposed to giving an absolute prediction on the individual choice (Ben-Akiva et al., 1985; Hensher et al., 2005). Thus, the predictive accuracy of the models may differ at the individual level and the aggregate level: Which of them should be prioritized depends on the project at hand.

Another important difference between the two approaches lies in the input data structures. The fitting of a logit model requires the data on all available alternatives. In other words, even if the attributes of non-chosen alternatives are not observed, their values need to



One observation for logit models

Person	Mode	Choice	Travel time	Gender	...
1	Car	0	8	Male	...
1	Walk	0	30	Male	...
1	Bike	1	20	Male	...
1	PT	0	15	Male	...

One observation for machine-learning models

Person	Mode	Travel time (Car)	Travel time (walk)	Travel time (bike)	Travel time (PT)	Gender	...
1	Bike	8	30	20	15	Male	...

Figure 1: Data Structure for Logit Models and Machine-Learning Models.

be modeled. In contrast, machine-learning algorithms require the observed (chosen) mode only and not necessarily information on the non-chosen alternatives. For example, many previous studies have only considered attribute values of the chosen mode (e.g., travel time of the chosen mode (Wang and Ross, 2018; Xie et al., 2003)) in their machine-learning models. We believe that it is better to also consider attribute values of the non-chosen modes since, if non-chosen mode travel times are not provided, the machine-learning model learns the mode choice based on the chosen mode travel time, which cannot be used to analyze mode changes or plan new transportation projects or services, like the mobility-on-transit system studied in the paper.

Figure 1 shows one observation that serves as the input to logit models and machine-learning models respectively. This difference has implications on the flexibility of these two types of models in modeling alternative-specific attributes (e.g., wait time is a transit-specific attribute). Due to this layered structure, logit models can easily accommodate these variables since each alternative has its own utility function and so alternative-specific attributes only enter the utility functions of the corresponding alternatives. While alternative-specific attributes can also be added into machine-learning models, the model does not explicitly capture that these attributes are associated only with certain alternatives.

### 3.2. Model Evaluation

When evaluating statistical and machine-learning models, the goal is to minimize the overall prediction error, which is a sum of three terms: the bias, the variance, and the irreducible error. The bias is the error due to incorrect assumptions of the model. The variance is the error arising from the model sensitivity to the small fluctuations in the dataset used for fitting the model. The irreducible error results from the noise in the problem itself. The relationship between bias and variance is often referred to as “bias-variance tradeoff,” which measures the tradeoff between the goodness-of-fit and model complexity. Goodness-of-fit captures how a statistical model can capture the discrepancy between the observed values and the values expected under the model. Better fitting models tend to

have more complexity, which may create overfitting issues and decrease the model predictive capabilities. On the other hand, simpler models tend to have a worse fit and a higher bias, causing the model to miss relevant relationships between input variables and outputs, which is also known as underfitting. Therefore, in order to balance the bias-variance tradeoff and obtain a model with low bias and low variance, one needs to consider multiple models at different complexity levels, and use an evaluation criterion to identify the model that minimizes the overall prediction error. The process is known as model selection. The evaluation criteria can be theoretical measures like adjusted  $R^2$ , AIC,  $C_p$ , and BIC, and/or resampling-based measures, such as cross validation and bootstrapping. Resampling-based measures are generally preferred over theoretical measures.

The selection of statistical models is usually based on theoretical measures. For example, when using logit models to predict individual mode choices, researchers usually calibrate the models on the entire dataset, examine the log-likelihood at convergence, and compare the resulting adjusted McFadden’s pseudo  $R^2$  (McFadden, 1973), AIC, and/or BIC in order to determine a best-fitting model. These three measures penalize the likelihood for including too many “useless” features. The adjusted McFadden’s pseudo  $R^2$  is most commonly reported for logit models, and a value between 0.2 to 0.3 is generally considered as indicating satisfactory model fit (McFadden, 1973). On the other hand, AIC and BIC are commonly used to compare models with different number of variables.

When applying machine-learning models, cross validation is usually conducted to evaluate a set of different models, with different variable selections, model types, and choices of hyper-parameters. The best model is thus identified as the one with the highest out-of-sample predictive power. A commonly-used cross validation method is the 10-fold cross validation, which applies the following procedure: 1) Randomly split the entire dataset into 10 disjoint equal-sized subsets, 2) choose one subset for validation, the rest for training; 3) train all the machine-learning models on one training set; 4) test all the trained models on the validation set and compute the corresponding predictive accuracy; 5) repeat Step 2) to 4) for 10 times, with each of the 10 subsets used exactly once as the validation data; and 6) the 10 validation results for each model are averaged to produce a mean estimate. Cross validation allows researchers to compare very different models together with the single goal of assessing their predictive accuracy. This paper compares the logit and machine-learning models using the 10-fold cross validation in order to evaluate their predictive capabilities at individual and aggregate levels.

Finally, when applying statistical models such as the logit models, researchers often take into account the underlying theoretical soundness and the behavioral realism of the model outputs to identify a final model (in addition to relying on the adjusted McFadden’s pseudo  $R^2$ , AIC and/or BIC). In other words, even though balancing the bias-variance tradeoff is very important, in statistical modeling, a “worse” model may be preferred due to reasons like theoretical soundness and behavioral realism. For example, since worsening the performance of a travel mode should decrease its attractiveness, the utility coefficients of the level-of-service attributes such as wait time for transit should always have a negative sign. Therefore, when a “better” model produces a positive sign for wait time, a “worse” model may be preferred. On the other hand, for machine-learning models, the predictive accuracy

is typically the sole criterion for deciding the best model, so machine-learning models may produce results that contradict the theoretical soundness or behavioral realism. This paper however shows that machine-learning models could also be selected based on behavioral realism through behavioral interpretation.

### 3.3. Model Interpretation and Application

The interpretation of outputs of logit models is intuitive and simple. Like any other statistical model, researchers can quickly learn how and why a logit model works by examining the sign, relative magnitude, and statistical significance of the model coefficients. Researchers may also apply these outputs to conduct further behavioral analysis on individual travel behavior, such as deriving marginal effect and elasticity estimates, comparing the utility differences in various types of travel times, calculating traveler willingness-to-pay for trip time and other service attributes. All of these applications can be validated by explicit mathematical formulations and derivations, which allows modelers to clearly explain what happens “behind the scene.”

In contrast, machine-learning models are often criticized for being “black box” and lacking interpretability (Klaiber and von Haefen, 2011). Some complex machine-learning models such as NN and RF may contain hundreds or even thousands of parameters and no human language can describe how exactly they work. In practice, it is often the case that more complex models have higher prediction accuracy, but increasing complexity inevitably reduces interpretability. Accordingly, machine-learning practitioners rarely attempt to directly interpret the model parameters. Instead they apply *model-agnostic* interpretability methods such as variable importance and partial dependence plots to extract explanations of their outputs (Molnar, 2018). On the one hand, variable importance measures in machine learning shows which variables have the most impact when predicting the response variable. Different machine-learning models have different ways to compute variable importance. For example, for tree-based models (such as CART and RF), the mean decrease in node impurity (measured by the Gini index) is commonly used to measure the variable importance. On the other hand, partial dependence plots measure the influence of a variable  $\mathbf{Z}_p$  on the log-odds or probability of choosing a mode  $k$  after accounting for the average effects of the other variables (Hastie et al., 2001). In recent years, as machine learning became increasingly popular in the study of societal systems, there has been a surge of research interest in the development of these model-agnostic methods to make machine-learning models and their decisions understandable (Vellido et al., 2012).

Arguably, the behavioral insights that one can obtain from the logit models (with parameter ratios, marginal effects, and elasticities) may also be generated by machine-learning models through the application of model-agnostic interpretability methods and sensitivity analysis. For example, for machine-learning models, the arc elasticity for feature  $p$  can be obtained by

$$E(\mathbf{Z}_p) = \frac{[Q_k(\mathbf{Z}_{-p}, \mathbf{Z}_p \cdot (1 + \Delta)|\hat{\boldsymbol{\theta}}) - Q_k(\mathbf{Z}|\hat{\boldsymbol{\theta}})]/Q_k(\mathbf{Z}|\hat{\boldsymbol{\theta}})}{\Delta}, \quad (8)$$

and the marginal effect for feature  $p$  can be computed as

$$M(\mathbf{Z}_p) = \frac{Q_k(\mathbf{Z}_{-p}, \mathbf{Z}_p + \Delta | \hat{\boldsymbol{\theta}}) - Q_k(\mathbf{Z} | \hat{\boldsymbol{\theta}})}{\Delta}. \quad (9)$$

In essence, all of these techniques, despite their obvious differences, measure how the output variable responds to changes in the input features. In the context of travel mode choices, they help researchers gain a better understanding of how individual choices of travel modes is impacted by a variety of different factors such as the socio-economic and demographic characteristics of travelers and the respective trip attributes for each travel mode. In the current literature, however, the behavioral findings gained from machine-learning models are rarely compared with those obtained from logit models. Since the goals of mode choice studies are often in extracting knowledge to shed light on individual travel preferences and travel behavior instead of merely predicting their mode choice, these comparisons are necessary to have a more thorough evaluation of the adequacy of machine learning. Machine-learning models that have excellent predictive power but generate unrealistic behavioral results are not useful in travel behavior studies.

#### 4. The Data for Empirical Evaluation

The data used for empirical evaluation came from a stated-preference (SP) survey completed by the faculty, staff, and students at the University of Michigan in the Ann Arbor campus. In the survey, participants were first asked to estimate the trip attributes (e.g., travel time, cost, and wait time) for their home-to-work travel for each of the following modes: Walking, biking, driving, and taking the bus. Then, the survey asked respondents to envision a change in the transit system, i.e., the situation where a new public transit (PT) system, named RITMO Transit (Jenkins, 2018), that fully integrates high-frequency fixed-route bus services and micro-transit services has replaced the existing bus system (see Figure 2). Text descriptions were coupled with graphical illustrations to facilitate the understanding of the new system. Each survey participant was then asked to make their commute-mode choice among *Car*, *Walk*, *Bike*, and *PT* in seven state-choice experiments, where the trip attributes for *Walk*, *Bike*, and *Car* were the same as their self-reported values and the trip attributes for *PT* were pivoted from those of driving and taking the bus. A more detailed descriptions of the survey can be found in Yan et al. (2018).

A total of 8,141 observations collected from 1,163 individuals were kept for analysis after a data-cleaning process. The variables that enter into the analysis include the trip attributes for each travel mode, several socio-demographic variables, transportation-related residential preference variables, and current/revealed travel mode choices. The travel attributes include travel time for all modes, wait time for *PT*, daily parking cost for driving, number of additional pickups for *PT*, and number of transfers for *PT*. The socio-economic and demographic variables include car access (car ownership for students and car per capita in the household for faculty and staff), economic status (living expenses for students and household income for faculty and staff), gender, and identity status (i.e., faculty vs staff vs student).

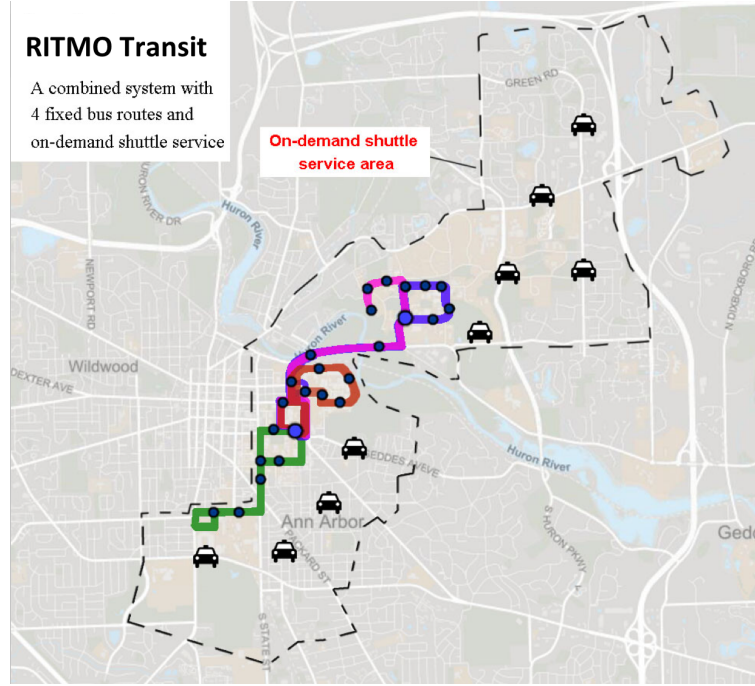


Figure 2: RITMO Transit: The New Transit System Featured with Four High-Frequency Bus Routes and On-Demand Shuttles Serving Approximately 2-Mile-Radius Area of the University of Michigan Campus.

The transportation-related residential preference variables are the importance of walkability/bikeability and transit availability when deciding where to live. Finally, current travel mode choices are also included as state-dependence effects (i.e., the tendency for individuals to abandon or stick with their current travel mode) are verified as important predictors of mode choice by many empirical studies. Table 4 summarizes the descriptive statistics on these variables, including a general description of each variable, category percentages for categorical variables, and min, max, mean, and standard deviation for continuous variables.

After extracting the data from the SP survey, we pre-processed the data and verified that all the independent variables have little multicollinearity (Farrar and Glauber, 1967). The existence of multicollinearity can inflate the variance and negatively impact the predictive power of the models. This study chose the variance inflation factor to determine which variables are highly correlated with other variables and found out that all variables had a variance inflation factor value of less than five, indicating that multicollinearity was not a concern.

## 5. Models Examined and Their Specifications

This section briefly introduces the logit and machine-learning models examined in this study. Since our dataset has a panel structure, usually a mixed logit model should be applied. However, we also fitted an MNL model as the benchmark for comparison, as previous studies generally compared machine-learning models with the MNL model only. Seven machine-learning models are examined, including simple ones like NB and CART,

Table 4: Statistics for Independent Variables and Dependent Variable (Travel Mode)

Variable	Description	Category	%	Min	Max	Mean	SD
<i>Dependent Variable</i>							
Mode Choice		Car	14.888				
		Walk	28.965				
		Bike	20.870				
		PT	35.278				
<i>Independent Variables</i>							
TT_Drive	Travel time of driving (min)			2.000	40.000	15.210	6.616
TT_Walk	Travel time of walking (min)			3.000	120.000	32.300	23.083
TT_Bike	Travel time of biking (min)			1.000	55.000	15.340	10.447
TT_PT	Travel time of using PT (min)			6.200	34.000	18.680	4.754
Parking_Cost	Parking cost (\$)			0.000	5.000	0.9837	1.678
Wait_Time	Wait time for PT (min)			3.000	8.000	5.000	2.070
Transfer	Number of transfers			0.000	2.000	0.328	0.646
Rideshare	Number of additional pickups			0.000	2.000	1.105	0.816
Income	Income level			1.000	6.000	1.929	1.342
Bike_Walkability	Importance of bike- and walk-ability			1.000	4.000	3.224	0.954
PT_Access	Importance of PT access			1.000	4.000	3.093	1.023
CarPerCap	Car per capita			0.000	3.000	0.529	0.476
Female	Female or Male	Female	56.320				
		Male	43.680				
Student	Students or faculty/staff	Student	73.517				
		Faculty or staff	26.483				
Current_Mode_Car	Current travel mode is Car or not	Car	16.681				
		Not Car	83.319				
Current_Mode.Walk	Current travel mode is Walk or not	Walk	40.413				
		Not Walk	59.587				
Current_Mode.Bike	Current travel mode is Bike or not	Bike	8.254				
		Not Bike	91.746				
Current_Mode.PT	Current travel mode is PT or not	PT	34.652				
		Not PT	65.348				

and more complex ones such as RF, BOOST, BAG, SVM, and NN. Most previous mode choice studies only examined a subset of these models (Chen et al., 2017; Omrani, 2015; Omrani et al., 2013; Wang and Ross, 2018; Xie et al., 2003).

### 5.1. Logit Models

We have already introduced the MNL model formulation in detail in Subsection 3.1, and so only the mixed logit model is presented here.

The mixed logit model is an extension of the MNL model, which addresses some of the MNL limitations (such as relaxing the IIA property assumption) and is more suitable for modeling panel choice datasets in which the observations are correlated (i.e., each individual is making multiple choices) (McFadden and Train, 2000). A mixed logit model specification usually treats the coefficients in the utility function as varying across individuals but being constant over choice situations for each person (Train, 2009). The utility function from alternative  $k$  in choice occasion  $t$  by individual  $i$  is

$$U_{ikt} = \beta_{ik}^T \mathbf{X}_{ikt} + \varepsilon_{ikt}, \quad (10)$$

where  $\varepsilon_{ikt}$  is the independent and identically distributed random error across people, alternatives, and time. Hence, conditioned on  $\beta$ , the probability of an individual making a

sequence of choices (i.e.,  $\mathbf{j} = \{j_1, j_2, \dots, j_\tau\}$ ) is

$$\mathbf{L}_{ij}(\boldsymbol{\beta}) = \prod_{t=1}^{\tau} \left[ \frac{\exp(\boldsymbol{\beta}_{ij_t}^T \mathbf{X}_{ij_{it}})}{\sum_k \exp(\boldsymbol{\beta}_{ik}^T \mathbf{X}_{ikt})} \right]. \quad (11)$$

Because the  $\varepsilon_{ikt}$ 's are independent over the choice sequence, the corresponding unconditional probability is

$$p_{ikj} = \int \mathbf{L}_{ij}(\boldsymbol{\beta}) g(\boldsymbol{\beta}) d\boldsymbol{\beta}, \quad (12)$$

where  $g(\boldsymbol{\beta})$  is the probability density function of  $\boldsymbol{\beta}$ . This integral does not have an analytical solution, so it can only be estimated using simulated maximum likelihood (e.g. Train, 2009).

In this study, the MNL models can be summarized as follows: 1) The utility function of *Car* includes mode-specific parameters for TT\_Drive, Parking\_Cost, Income, CarPer-Cap, and Current\_Mode\_Car; 2) the utility function of *Walk* includes mode-specific parameters for TT\_Walk, Student (sharing the same parameter for *Bike*), Female (sharing the same parameter with *Bike*), Bike\_Walkability (sharing the same parameter with *Bike*), and Current\_Mode\_Walk; 3) the utility function of *Bike* includes mode-specific parameters for TT\_Bike, Student (sharing the same parameter with *Walk*), Female (sharing the same parameter with *Walk*), Bike\_Walkability (sharing the same parameter with *Walk*), and Current\_Mode\_Bike; and 4) the utility function of *PT* includes mode-specific parameters for TT\_PT, Wait\_Time, Rideshare, Transfer, Student, PT\_Access, and Current\_Mode\_PT. We also specify three alternative specific constants for *Walk*, *Bike*, and *PT* respectively. The mixed logit model has the same model specifications except that travel time is generic across all modes. In addition, in order to accommodate individual preference heterogeneity (i.e. taste variations among different individuals), we also specify coefficients on all the level-of-service variables (i.e., travel time, Wait\_Time, Parking\_Cost, Transfer, and Rideshare) as random parameters. These random parameters are all assessed with a normal distribution. We use 1,000 halton draws to perform the integration. Both the MNL and mixed logit models are estimated using the NLOGIT software.

## 5.2. Machine-Learning Models

### 5.2.1. Naive Bayes

The NB model is a simple machine-learning classifier. The model is constructed using Bayes' Theorem with the naive assumption that all the features are independent (McCallum et al., 1998). NB models are useful because they are faster and easier to construct as compared to other complicated models. As a result, NB models work well as a baseline classifier for large datasets. In some cases, NB even outperforms more complicated models (Zhang, 2004). A limitation of the NB model is that, in real world situations, it is very unlikely for all the predictors to be completely independent from each other. Thus, the NB model is very sensitive when there are highly correlated predictors in the model. In this study, the NB model is constructed through the R package *e1071* (Meyer et al., 2017).

### 5.2.2. Tree-based Models

The CART model builds classification or regression trees to predict either a classification or a continuous dependent variable. In this paper, the CART model creates classification trees where each internal node of the tree recursively partitions the data based on the value of a single predictor. Leaf nodes represent the category (i.e., *Car*, *Bike*, *PT*, and *Walk*) predicted for that individual (Breiman, 2017). The decision tree is sensitive to noise and susceptible to overfit (Last et al., 2002; Quinlan, 2014). To control its complexity, it can be pruned. This study prunes the tree until the number of terminal nodes is 6. The CART model is obtained through the R package *tree* (Ripley, 2016).

To address the overfitting issues of CART models, tree-based ensemble techniques were proposed to form more robust, stable, and accurate models than single decision trees (Breiman, 1996; Hastie et al., 2001). One of these ensemble methods is BOOST. For a  $K$ -class problem, BOOST creates a sequence of decision trees, where each successive tree seeks to improve the incorrect classifications of the previous trees. Predictions in BOOST are based on a weighted voting among all the boosting trees. Although BOOST usually has a higher predictive accuracy than CART, it is more difficult to interpret. Another drawback is that BOOST is prone to overfitting when too many trees are used. This study applies the gradient boosting machine technique to create the BOOST model (Friedman, 2001). 500 trees are used, with shrinkage parameter set to 0.05 and the interaction depth to 10. The minimum number of observations in the trees terminal nodes is 10. The BOOST model is created with the R package *gbm* (Ridgeway, 2017).

Another well-known ensemble method is BAG, which trains multiple trees in parallel by bootstrapping data (i.e., sampling with replacement) (Breiman, 1996). The BAG model uses all independent variables to train the trees. For a  $K$ -class problem, after all the trees are trained, the BAG model makes the mode choice prediction by determining the majority votes among all the decision trees. By using bootstrapping, the BAG model is able to reduce the variance and overfitting problems of a single decision tree model. One potential drawback with the BAG model is that it assumes that all the features are independent. If the features are correlated, the variance would not be reduced with BAG. In this study, 600 classification trees are bagged, with each tree grown without pruning. The BAG model is produced with the R package *ipred* (Peters and Hothorn, 2017).

The RF model is also an ensemble method. Like BAG, RF trains multiple trees using bootstrapping (Ho, 1998). However, RF only uses a random subset of all the independent variables to train the classification trees. More precisely, the trees in RF use all the independent variables, but every node in each tree only uses a random subset of them (Breiman, 2001). By doing so, RF reduces variance between correlated trees and negates the drawback that BAG models may have with correlated variables. Similar to BAG, RF makes mode choice predictions by determining the majority votes among all the classification trees. Like other non-parametric models, RF is difficult to interpret. In this study, 600 trees are used and 10 randomly selected variables are considered for each split at the trees' nodes. The R package used for producing the RF model is *randomForest* (Liaw and Wiener, 2002).



### 5.2.3. Support Vector Machine

The SVM model is a binary classifier which, given labeled training data, finds the hyperplane maximizing the margin between two classes. This hyperplane is a linear or nonlinear (depending on the kernel) decision boundary that separates the two classes. Since a mode choice model typically involves multi-class classification, the one-against-one approach is used (Hsu and Lin, 2002). Specifically, for a  $K$ -class problem,  $K(K - 1)/2$  binary classifiers are trained to differentiate all possible pairs of  $K$  classes. The class receiving the most votes among all the binary classifiers is selected for prediction. SVM usually performs well with both nonlinear and linear boundaries depending on the specified kernel. However, the SVM model can be very sensitive to overfitting especially for nonlinear kernels (Cawley and Talbot, 2010). In this study, a SVM with a radial basis kernel is used. The cost constraint violation is set to 1.25, and the gamma parameter for the kernel is set to 0.4. The SVM model is produced with the R package *e1071* (Meyer et al., 2017).

### 5.2.4. Neural Network

A basic NN model has three layers of units/nodes where each node can either be turned active (on) or inactive (off), and each node connection between layers has a weight. The data is fed into the model at the input layer, goes through the weighted connections to the hidden layer, and lastly ends up at a node in the output layer which contains  $K$  units for an  $K$ -class problem. The hidden layer allows the NN to model nonlinear relationships between variables. Although NN have shown promising results in modeling travel mode choice in some studies (Omrani, 2015), NN models tend to be overfitting, and are difficult to interpret. In this paper, a NN with a single hidden layer of 10 units is used. The connection weights are trained by back propagation with a weight decay constant of 0.1. The R package *nnet* (Venables and Ripley, 2002) is used to create our NN model.

## 6. Comparison of Empirical Results

This section presents the empirical results of this study. Specifically, it compares the predictive accuracy of the logit models with that of the machine-learning algorithms. In addition, it compares the behavioral findings of the best-performing model (in terms of predictive accuracy) from each approach.

### 6.1. Predictive Accuracy

This study applied the 10-fold cross validation approach. As discussed above, cross validation requires sub-setting the sample data into training sample sets and validation sample sets. One open issue is how to partition the sample dataset when it is a panel dataset (i.e., individuals with multiple observations). One approach is to treat all observations as independent choices and randomly divide these observations. The other is to subset by individuals, each with their full set of observations. This study follows the first approach, which is commonly applied by previous studies (Hagenauer and Helbich, 2017; Wang and Ross, 2018; Xie et al., 2003).

As discussed in Subsection 3.1, the predictive power of the models may differ at the individual level (predicting the mode choice of a particular choice) and at the aggregate level (predicting the market shares for each travel mode). The calibration of logit models focuses on reproducing market shares whereas the development of machine-learning classifiers focuses on predicting individual choices. This study compares both the mean individual-level predictive accuracy and the mean aggregate-level predictive accuracy.

#### 6.1.1. Individual-Level Predictive Accuracy

The cross validation results for individual-level predictive accuracy is shown in Table 5. Note that, while the machine-learning methods predict a particular travel mode, logit models return probabilities for all available modes. The results assume that the travel mode with the highest predicted probability is selected as the predicted mode for the logit models. The best two models are RF and BAG, with a mean predictive accuracy equal to 0.856 and 0.843 respectively. However, the accuracy of the MNL and mixed logit model is only 0.640 and 0.592 respectively, much lower than the two best-performing machine-learning models.

The predictive accuracy of each model by travel mode is presented in Table 5. All models predict *Walk* most accurately. All machine-learning models have a mean predictive accuracy value between 0.795 and 0.929, whereas the MNL model has an accuracy of 0.860 and the mixed logit model 0.652. Both logit models and the two best-performing machine-learning models predict modes *PT* and *Bike* relatively better than mode *Car*. One possible explanation is that *Car*, with a market share of 14.888%, has fewer observations compared to other modes. Furthermore, many car owners always select mode *Car* regardless of changes in the PT profiles; such non-switching behavior creates a challenge to accurately predict mode *Car* (Hess et al., 2010).

Finally, it is somewhat surprising that the mixed logit model, a model that accounts for individual preference heterogeneity and has significantly better model fit (adjusted McFadden’s pseudo  $R^2$  is 0.58) than the MNL model (adjusted McFadden’s pseudo  $R^2$  is 0.36), underperformed the MNL model in terms of the out-of-sample predictive power. This finding is nonetheless consistent with the findings of Cherchi and Cirillo (2010). It suggests that the mixed logit model may have overfitted the data with the introduction of random parameters, and such overfitting resulted in greater out-of-sample prediction error.

It is also useful to compare the four models (two logit models and two best-performing machine-learning models) using a  $t$ -test to check if the mean accuracy of these models is significantly different from each other. The null hypothesis of the  $t$ -test is that the mean predictive accuracy of the four models is the same, while the alternative hypothesis is that the mean predictive accuracy of the four models is different. When the  $P$ -value is lower than the common thresholds, e.g., 0.1 or 0.05, the null hypothesis may be rejected. Since multiple comparisons for the four models must be conducted, the  $P$ -values obtained from these comparisons have to be adjusted otherwise, a null hypothesis could be incorrectly rejected by pure chance (Dunnett, 1955). Hence the  $P$ -values are adjusted by applying the Bonferroni correction (Rice, 2006), which requires  $P$ -values  $\leq \alpha/k$ , where  $\alpha$  represents the significance level (in this case,  $\alpha = 0.05$ ) and  $k$  is the number of individual significance tests (i.e.,  $k = 4$ ). The results of the adjusted  $t$ -tests are given in Table 6, where the numbers

Table 5: Mean Out-of-Sample Accuracy of Logit and Machine-Learning Models (Individual Level)

	Overall		Car		Walk		Bike		PT	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
<b>MNL</b>	0.640	0.016	0.393	0.028	0.860	0.017	0.414	0.032	0.698	0.027
<b>Mixed logit</b>	0.592	0.013	0.481	0.026	0.652	0.033	0.573	0.032	0.601	0.034
<b>NB</b>	0.602	0.013	0.612	0.046	0.852	0.018	0.368	0.038	0.529	0.030
<b>CART</b>	0.593	0.014	0.428	0.032	0.795	0.022	0.329	0.038	0.653	0.026
<b>BOOST</b>	0.836	0.010	0.775	0.028	0.916	0.014	0.798	0.031	0.817	0.028
<b>BAG</b>	0.843	0.016	0.774	0.018	0.916	0.020	0.834	0.023	0.817	0.029
<b>RF</b>	0.856	0.015	0.796	0.017	0.929	0.016	0.859	0.028	0.819	0.030
<b>SVM</b>	0.731	0.016	0.589	0.021	0.858	0.027	0.575	0.030	0.778	0.026
<b>NN</b>	0.643	0.014	0.445	0.046	0.863	0.026	0.420	0.040	0.678	0.028

Table 6: Adjusted  $t$ -tests Results for Testing Whether the Difference Between the Mean Predictive Accuracy of MNL, Mixed Logit, BAG, and RF Is Significant Using Bonferroni Correction ( $\alpha = 0.05$ ).

	MNL	Mixed Logit	BAG	RF
MNL		Significant	Significant	Significant
Mixed Logit	5.805e-5		Significant	Significant
BAG	1.273e-9	1.160e-10		Significant
RF	8.525e-11	1.656e-11	2.295e-3	

below the diagonal are  $P$ -values, and the numbers above the diagonal are testing conclusions (i.e., whether the mean difference is significant or not).

As a result, with all the adjusted  $P$ -values in the fourth row smaller than 0.05, the mean accuracy of RF (the best model) is significantly different from other models, which provides us with strong statistical evidence that RF has the best predictive performance at the individual level. With 85.6% predictive accuracy, it is advisable to apply machine learning to predict individual-level mode choices for the activity-based or agent-based transportation models. The results conclude with high confidence that the logit models are weaker than the best-performing machine-learning model in terms of prediction.

#### 6.1.2. Aggregate-Level Predictive Accuracy

We now turn to aggregate-level predictive accuracy. To quantify the sum of the absolute differences between the market share predictions and the real market shares from the validation data, we use the L1-norm, also known as the least absolute deviations. Taking machine-learning models as an example, let  $Q_k^*$  and  $\hat{Q}_k = Q_k(\mathbf{Z}|\hat{\theta})$  represent the true (observed) and predicted shares for mode  $k$ ,  $k = 1, 2, 3, 4$ . The L1-norm thus is defined as

$$\sum_{k=1}^4 |Q_k^* - \hat{Q}_k|. \quad (13)$$

The predictive accuracy results of the logit and machine-learning models at the aggregate level are depicted in Table 7. The results show that RF outperforms all the other models, with a prediction error of 0.043 and a standard deviation of 0.014. Notably, even though

Table 7: Mean L1-Norm Error for Mode Share Prediction

	<b>L1-Norm</b>	
	Mean	SD
<b>MNL</b>	0.048	0.020
<b>Mixed logit</b>	0.076	0.026
<b>NB</b>	0.297	0.043
<b>CART</b>	0.293	0.038
<b>BOOST</b>	0.067	0.022
<b>BAG</b>	0.048	0.019
<b>RF</b>	0.043	0.014
<b>SVM</b>	0.174	0.030
<b>NN</b>	0.221	0.037

logit models are expected to have good performance for market share predictions, RF has lower error compared to MNL (0.048) and mixed logit (0.076). Again, the MNL model resulted in a higher aggregate-level predictive accuracy than the mixed logit model.

In summary, the results show that RF is the best model (among those evaluated) for forecasting travel choice for a new transit system featuring very different parameters.

## 6.2. Model Interpretation

Recent advances in machine learning make models interpretable through techniques such as variable importance and partial dependence plots. Machine-learning results can be readily applied to compute behavioral outputs such as marginal effects and arc elasticities. However, other behavioral outputs such as the value of time, willingness-to-pay, and consumer welfare measures are hard to obtain from machine-learning models, because they are grounded on the random utility modeling framework and an assumption that individual utility can be kept constant when attributes of a product substitutes each other (e.g., paying a certain amount of money to reduce a unit of time). Machine-learning models lacks the behavioral foundation required to obtain these measures.

This section examines the behavioral findings from the best-performing logit model (MNL) and machine-learning model (RF). For the MNL model, we interpret the model results and calculate some behavioral outputs including marginal effects and elasticities. We then conduct comparable behavioral analysis on the RF model by applying variable importance and partial dependence plots and by performing sensitivity analysis. Finally, we compare and contrast the behavioral findings generated by the two models.

It should be noted that, while the mixed logit model is found to be inferior in terms of its predictive capacity (and hence its results are not discussed here), it can generate additional behavioral insights on individual travel that neither the MNL nor any machine-learning model can produce. Notably, the mixed logit model is very flexible in modeling (both observed and unobserved) preference heterogeneity, i.e., variations in traveler tastes for different attributes of the choice alternatives, among the study population. Since the MNL model does not recognize the panel data structure (i.e. repeated observations from the same individual), it has limited capacity to accommodate preference heterogeneity. For example,

one can only model *observed* taste variations using the market segmentation approach (Train, 2009).

### 6.2.1. Variable Importance and Effects

The outputs for the MNL model are presented in Table 8. The adjusted McFadden’s pseudo  $R^2$  for this model is 0.36, which indicates satisfactory model fit. All of the coefficient estimates are consistent with theoretical predictions. All level-of-service variables carry an intuitive negative sign, and all of them are statistically significant except for Parking\_Cost. Individual socio-demographic characteristics are associated with their travel mode choices. Unsurprisingly, higher-income travelers with better car access are more likely to drive than using alternative modes. Females are less likely to choose *Walk* and *Bike* than males, but there is no significant difference between the mode choice of students and faculty/staff. The model also shows that individual residential preferences and current travel mode choices are associated with their travel mode choices of *Car*, *Walk*, and *Bike*. However, people tend to have weak attachment to *PT* as shown by the small and insignificant beta coefficient. Individuals who value walking, biking, and transit access when choosing where to live are more likely to use these modes. Also, the model shows that travelers tend to stick to their current mode even when a new travel option is offered.

The last column of the table shows the  $X$ -standardized beta coefficients for the MNL model, which allow researchers to assess the relative importance of the independent variables, i.e., a coefficient of larger magnitude indicate a greater impact of the corresponding independent variable on the choice outcome (Menard, 2004). These results show that the most important variable in predicting mode choice is TT\_Bike, followed by the travel time variables for the other three modes, several revealed-preference (RP) variables (i.e., current travel modes), and some level-of-service attributes including Transfer, Rideshare, and Wait\_Time. These results are reasonable and generally consistent with findings in the existing literature.

There is growing research interest in developing techniques to interpret machine learning in order to help explain the decisions behind the complex models (Miller, 2017). This study applied widely used tools including variable importance measures and partial dependence plots to interpret the RF model and compare the behavioral findings obtained from the RF with those from the MNL models. Like ( $X$ -)standardized beta coefficients in a logit model, a variable importance measure can be used to indicate the impact of an input variable on predicting the response variable for machine-learning models. Unlike ( $X$ -)standardized beta coefficients that can show the direction of association between the input variable and the outcome variable with a positive or negative sign, however, variable importance measures provide no such information.

This study uses the Gini index to measure variable importance for RF. Figure 3 shows the relative variable importance of each input feature for the RF and MNL models relative to the maximum value. Note that the reason why Student has two values in the MNL model is that logit models have the flexibility of specifying alternative specific coefficients to account for different effects of a single feature on different alternatives. The ranking of the input features with respect to their relative importance in RF is generally consistent with

Table 8: Outputs of the MNL Model.

Variable	Alternatives	Unstandardized coefficients		X-standardized coefficients
		$\beta$	S.E.	$\beta\text{Std}X$
<i>Constants</i>				
Walk	Walk	2.882**	0.418	/
Bike	Bike	1.767**	0.412	/
PT	PT	3.260**	0.426	/
<i>Level-of-service variables</i>				
TT_Drive	Car	-0.075**	0.005	-1.138**
TT_Walk	Walk	-0.146**	0.004	-2.210**
TT_Bike	Bike	-0.163**	0.006	-2.461**
TT_PT	PT	-0.102**	0.009	-1.543**
Wait_Time	PT	-0.158**	0.018	-0.327**
Parking_Cost	Car	-0.134	0.096	-0.226
Rideshare	PT	-0.438**	0.042	-0.358**
Transfer	PT	-0.583**	0.050	-0.376**
<i>Socio-demographic variables</i>				
Income	Car	0.076*	0.030	0.103*
CarPerCap	Car	0.561**	0.083	0.267**
Student	Walk, Bike	-0.093	0.368	-0.041
	PT	-0.026	0.369	-0.011
Female	Walk, Bike	-0.174**	0.061	-0.086**
<i>Residential preference variables</i>				
Bike_Walkability	Walk, Bike	0.072*	0.033	0.068*
PT_Access	PT	0.112**	0.031	0.115**
<i>Current travel mode</i>				
Current_Mode_Car	Car	1.369**	0.094	0.510**
Current_Mode_Walk	Walk	1.291**	0.077	0.634**
Current_Mode_Bike	Bike	2.891**	0.120	0.796**
Current_Mode_PT	PT	0.090	0.073	0.043
<hr/>				
Sample size	1163			
Log likelihood at constant	-11285.82			
Log likelihood at convergence	-7160.62			
Adjusted McFadden's pseudo $R^2$	0.36			

\* significant at the 5% level \*\* significant at the 1% level.

that of the MNL model. The travel times of walking, driving, biking, and transit, and the revealed/current mode choice of biking have very high influence on their stated mode choice. Moreover, “Student” is neither important for the MNL model nor for the RF model. On the other hand, slight differences do exist. For example, PT\_Access and Bike\_Walkability have more importance for the RF model than they do for the MNL model. To conclude, the two model outputs on variable importance are very similar, which implies that both models

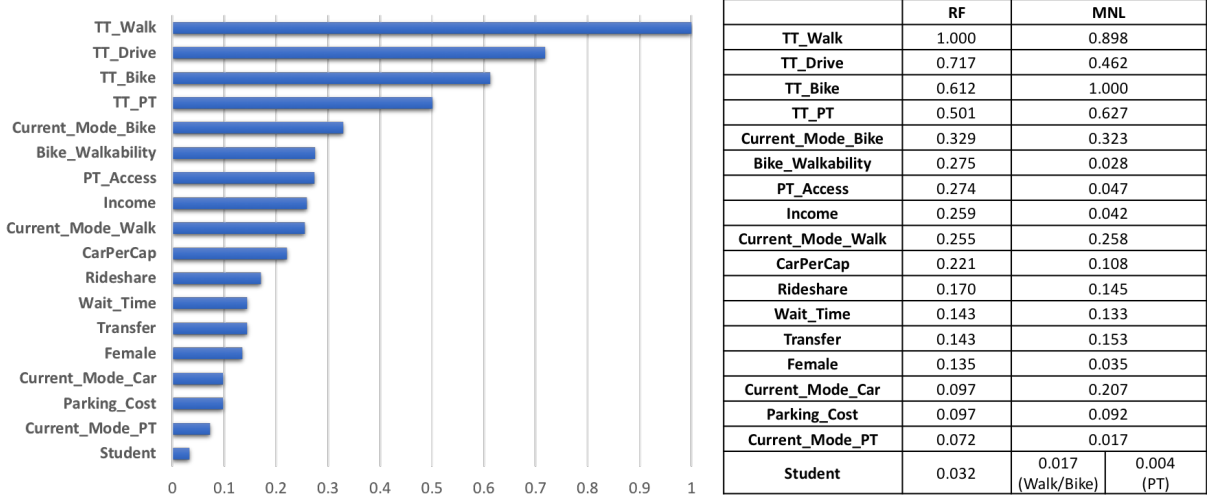
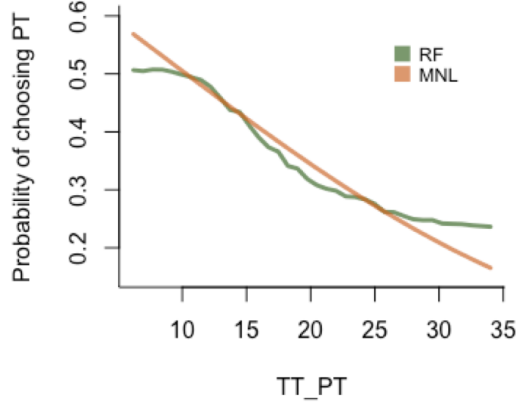


Figure 3: Relative Variable Importance Plot for RF and Comparison of MNL  $X$ -Standardized Coefficients Expressed as Relative to the Maximum.

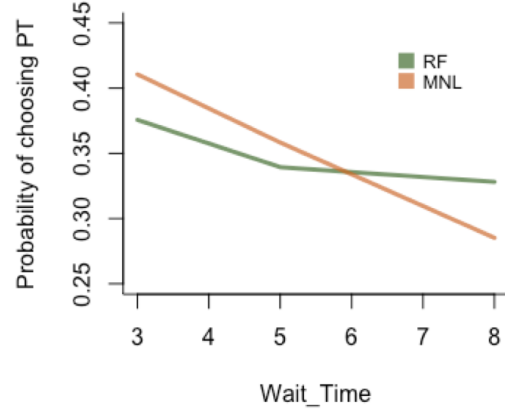
relied on similar information (variability of selected input features) contained in the sample data to predict the choice outcome.

Partial dependence plots are another important tool that helps interpret machine-learning models. Figure 4 presents how the probability of choosing  $PT$  changes as the value of the selected variable changes for RF and MNL. The shape of the curves sheds light on the direction and magnitude of the changes, which is similar to the beta coefficients estimated from the MNL model. However, the beta coefficients in logit models affect the utility of mode  $k$  (see Eqn. (1)) rather than the probability of choosing mode  $k$  (see Eqn. (2)). Accordingly, we translate utility estimates into probability estimates for the MNL model in order to compare it with RF directly.

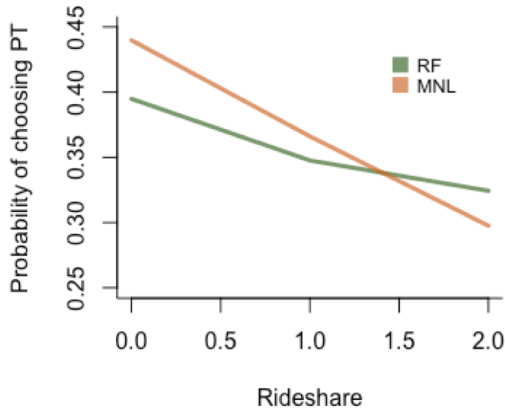
As shown in Figure 4(a), RF and MNL share a very similar trend for TT\_PT, with a very similar slope between 10 and 25 minutes. In addition, Figures 4(b)-4(d) show that RF and MNL have similar patterns for Wait\_Time, Rideshare, and Transfer, with RF having smaller slopes (the absolute values) compared to MNL. While MNL shows a nearly linear relationship between these features and the probability of choosing  $PT$ , RF presents some interesting findings for its nonlinear relationships: 1) For TT\_PT, RF has relative flat tails before 10 minutes and after 25 minutes, showing people tend to become insensitive to very short or very long transit times; 2) travelers are more sensitive to wait times less than 5 minutes; and 3) the choice probability of  $PT$  decreases more significantly from 0 to 1 ride share compared to from 1 to 2 ride shares. Therefore, unlike logit models that usually assume a linear relationship between the input variables and the utility functions, the partial dependence plots of machine-learning models can readily reveal the nonlinearities of mode choice responses to level-of-service attributes. As opposite the time-consuming hand-curating procedure required in logit models (often by introducing interactions terms) to reveal nonlinear relationships, machine learning algorithms search for nonlinearities automatically and thus can generate richer behavioral insights much more efficiently. Therefore, we believe that



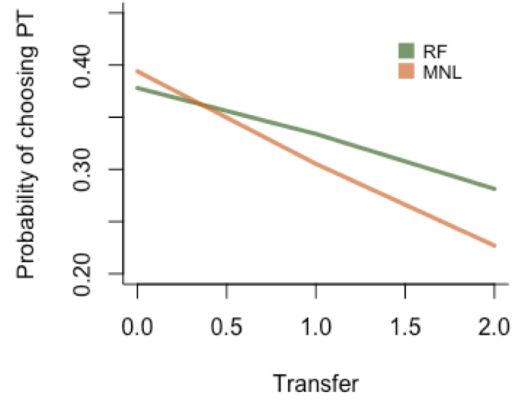
(a) Partial dependence on TT\_PT



(b) Partial dependence on Wait\_Time



(c) Partial dependence on Rideshare



(d) Partial dependence on Transfer

Figure 4: Partial dependence plots of variables for choosing  $PT$  as the travel mode

machine-learning models can serve as an exploratory analysis tool for identifying better specifications for the logit models in order to enhance the predictive power and explanatory capabilities of logit models.

The existence of nonlinearities, on the other hand, may prevent researchers from conveniently obtaining willingness-to-pay measures and valuation of time parameters like one can readily implement with a logit model. However, there is a caveat to the behavioral findings from the RF model. The partial dependence plots show that MNL agrees with RF well regarding the attribute TT\_PT but the impact of Wait\_Time, Rideshare, and Transfer on the choice probability of  $PT$  are smaller in the RF model than the MNL model. This is



discussed at length in the next section.

### 6.2.2. Arc Elasticity and Marginal Effects

Logit models are often applied to generate behavioral outputs such as marginal effects and elasticities to gain insights on individual travel behavior. Marginal effects (and elasticities) measure the changes of the choice probability of an alternative in response to a one unit (percent) change in an independent variable. This study calculates marginal effects and elasticities for the level-of-service variables associated with the proposed mobility-on-demand transit system, including TT\_PT, Wait\_Time, Rideshare, and Transfer. To facilitate the comparison of these behavioral outputs obtained from the MNL model with those generated by the RF model, arc elasticities were computed using Eqn. (8), since the RF model is not able to generate point elasticity estimates. Note that the data may present very nonlinear behavior and may not be sensitive to small changes, such as 1% and/or 1 unit change in a feature. Therefore, Table 9 presents the arc elasticity computed by applying 1%, 10%, 50%, and 100% increases for the selected feature. Similarly, the marginal effects are presented by applying 1, 2, and 5 units of increase for TT\_PT and Wait\_Time, as shown in Table 10.

The arc-elasticity and marginal-effect estimates are consistent with results shown in the partial dependence plots. The two models generate somewhat similar results regarding TT\_PT, but their outputs regarding Wait\_Time, Transfer, and Rideshare are drastically different. In general, the behavioral outputs of the MNL are quite reasonable. On the other hand, the RF model suggests that travelers are very insensitive to changes in Wait\_Time, Transfer, and Rideshare, which are inconsistent with findings in the existing literature (see Abrantes and Wardman (2011) for a meta-analysis of these behavioral outputs). For example, RF suggests that the impact of an additional transfer on the choice probability of public transit is smaller than that of an additional minute of travel time, which is very unlikely.

These results are puzzling for modelers: Why are the results on prediction and behavioral outputs inconsistent? Intuitively, a model is expected to capture traveler preferences (i.e., responses to different trip-attributes) reasonably in order to make accurate predictions of the choice outcome. However, our results show that RF has a significantly higher predictive quality even though it does not result in reasonable behavioral outputs for many trip attributes. Here we offer two potential reasons for this empirical contradiction. First, as discussed above, logit models define different utility functions for different modes and assume that alternative-specific attributes only affect the utility of their corresponding alternatives, i.e., Transfer, Wait\_Time, and Rideshare only affecting the utility of PT. In contrast, RF does not include such constraints and needs to “learn” them from the data by itself. When the data is not perfect, however, RF may stumble at representing travelers’ behavior realistically. For example, the results show that the RF model predicts that increasing the number of transfer associated with transit leads to a lower choice probability of choosing both *PT* and *Walk*, which implies a (unrealistic) negative cross-marginal-effect estimate of choosing *Walk* with respect to Transfer.

Second, there are specific limitations associated with using the RF model in estimating marginal effects and elasticities. To be specific, the prediction decisions of RF are based on splitting of input feature values at different nodes. These nodes can only take on discrete

Table 9: Arc Elasticity of *PT* Market Share with Respect to TT\_PT and Wait\_Time

		1% increase	10% increase	50% increase	100% increase
TT_PT	MNL	-0.86	-0.85	-0.77	-0.66
	RF	0.00	-0.99	-0.51	-0.30
Wait_Time	MNL	-0.34	-0.34	-0.33	-0.30
	RF	0.00	0.00	-0.08	-0.05

Table 10: Marginal Effects of *PT* Market Share with Respect to Transfer, Rideshare, TT\_PT, and Wait\_Time.

		Unit	1-unit increase	2-unit increase	5-unit increase
Transfer	MNL	Number	-10.93%	/	/
	RF	Number	-0.66%	/	/
Rideshare	MNL	Number	-8.23%	/	/
	RF	Number	-2.27%	/	/
TT_PT	MNL	Minute	-1.92%	-1.92%	-1.92%
	RF	Minute	-1.11%	-1.95%	-1.56%
Wait_Time	MNL	Minute	-2.96%	-2.96%	-2.96%
	RF	Minute	0.00%	-0.66%	-0.41%

value thresholds and thus the prediction decision becomes insensitive to values between and/or outside the thresholds. In the case study data, Wait\_Time, Transfer, and Rideshare only take three different values, since the data resulted from a SP survey where only three attribute levels were used to construct the state-choice experiments (see Yan et al. (2018) for more detail). For example, Wait\_Time has three values, i.e., 3 min, 5 min, and 8 min, and the decision trees inside RF split at 4 min, 5.5 min, and 6.5 min. As a result, RF has no ability to predict a different outcome for two observations with a Wait\_Time of 4.1 min and 5.4 min respectively or a Wait\_Time of 6.6 min and 10 min. In contrast, because the MNL model assumes a linear relationship between the utility functions and the independent variables, it is capable of distinguishing “between-thresholds” and “out-of-bound” observations. If Wait\_Time were to be observed for a larger variety of values like TT\_PT, it would have more reasonable behavioral outputs.

### 6.2.3. Revisiting Direct Arc Elasticities and Marginal Effects for Machine-Learning Models

It is possible to address these limitations through an alternative approach for calculating direct marginal effects and elasticities. Since the nature of the RF model structure makes it unsuitable to estimate these behavioral outputs in a standard fashion, the computations of elasticities and marginal effects need to include some behavioral constraints. In other words, while the logit models improve behavioral realism by constraining the model structure, machine-learning models may achieve similar results by constraining how the model results are applied to generate behavioral outputs. When degrading the value of an attribute associated with a given alternative, a behaviorally realistic response from a machine-learning model should be only changing the classification of those individuals currently using that

Table 11: Marginal Effects of *PT* Market Share with Respect to Transfer and Rideshare Using Proposed Approach for RF.

Values		0	1	2	Aggregate
Transfer	Marginal effects	-4.33%	-19.65%	-19.65%	-2.97%
	Market share	25.83%	6.38%	3.07%	
Rideshare	Marginal effects	-11.43%	-11.77%	-11.77%	-4.11%
	Market share	13.22%	10.12%	11.94%	

Table 12: Marginal Effects and Elasticity of *PT* Market Share with Respect to TT\_PT and Wait\_Time Using Proposed Approach for RF.

Values		<Max	Max	Aggregate
TT_PT	Marginal effects	-3.25%	-3.25%	-1.15%
	Elasticity	-1.08	-1.08	-1.08
	Market share	35.18%	0.10%	/
Wait_Time	Marginal effects	-3.10%	-3.10%	-1.09%
	Elasticity	-0.18	-0.18	-0.18
	Market share	26.94%	8.34%	/

alternative. For example, when a transfer is added to the PT alternative, it is expected that some individuals who are currently choosing PT may choose a different mode, whereas the classification for individuals who are currently choosing non-PT modes should not change. We thus propose to incorporate this behavioral constraint into the calculation of direct arc elasticity and marginal effect.

We illustrate the proposed approach by discussing the calculation of direct marginal effect of choosing PT with respect to Transfer. Since point estimates of marginal effects will be different for individuals with different attribute values due to nonlinear responses, marginal effects should be calculated at each data point first and then be projected back to the entire market based on their respective market shares. Denote the individuals choosing *PT* by  $\mathbf{Z}^{PT}$ , the marginal effect of choosing PT with respect to Transfer as by  $M(\mathbf{Z}_{\text{Transfer}}^{PT})$ , the marginal effect of choosing PT with respect to Transfer for those individuals with  $i$  transfers by  $M(\mathbf{Z}_{\text{Transfer}=i}^{PT})$ , and the market share of individuals currently choosing PT with  $i$  transfers by  $\phi_{PT, \text{Transfer}=i}$ . To allow RF to interpolate when the attribute values become out-of-bound after a change (e.g., Transfer becomes three, which is unobserved in the data, when one additional transfer is added to an individual who currently has two transfers), we assume that the marginal effects for out-of-bound observations are the same as those at the boundary (e.g.,  $M(\mathbf{Z}_{\text{Transfer}=2}^{PT}) = M(\mathbf{Z}_{\text{Transfer}=1}^{PT})$ ). Accordingly, the direct marginal effect of choosing *PT* with respect to Transfer can be expressed as:

$$M(\mathbf{Z}_{\text{Transfer}}) = \sum_{i=0}^2 M(\mathbf{Z}_{\text{Transfer}=i}^{PT}) \cdot \phi_{PT, \text{Transfer}=i}, \quad (14)$$

A similar approach can be applied to compute elasticity for continuous variables.

Table 11 presents the results of this approach for computing marginal effects of choosing *PT* with respect to Transfer and Rideshare, and Table 12 presents marginal effects and arc elasticities of choosing *PT* with respect to TT\_PT and Wait\_Time. Compared to results in Table 10, the results regarding Transfer, Rideshare, and Wait\_Time appear to improve significantly. In terms of marginal effects, these new results suggest that the impact of a transfer is approximately equal to three minutes of transit time and the impact of an additional pickup is approximately equal to four minutes of transit time. In addition, the RF model estimates that individuals value travel time slightly more than wait time. By contrast, the MNL results indicate that the effect of a transfer is roughly equal to five minutes of travel time and an additional pickup is roughly equivalent to four minutes and that individuals consider that wait time is 1.5 times as important as travel time by transit. Regarding the elasticity estimates, RF estimates a higher value for TT\_PT but a lower value for Wait\_Time compared to the MNL model. Overall, the behavioral results of the RF model and the MNL model are different but some comparable.

In the absence of ground truth, one cannot determine which model represents individual travel behavior more accurately than the other. The arguments can go either way. Some may argue that results obtained from the RF model should be more accurate given its superior predictive quality. Others, on the hand, may suggest that the behavioral outputs from any machine-learning model are unreliable because machine learning is not inherently built for the estimation task and lacks model selection consistency (Mullainathan and Spiess, 2017). We believe that this is largely an unresolved (untouched) issue in the literature and further theoretical and empirical studies are needed to shed light on the appropriateness of applying machine learning for parameter estimation and inference tasks. Our goal here is to start the conversation on these issues and to explore sound approaches to compute behavioral outputs from machine-learning algorithms in order to understand consumer choice behavior.

## 7. Discussion and Conclusion

The increasingly popularity of machine learning in transportation research raises questions regarding its advantages and disadvantages compared to conventional logit-family models used for travel behavioral analysis. The development of logit models typically focuses on parameter estimation and pays little attention to prediction (i.e., lack of a procedure to validate out-of-sample prediction accuracy). On the other hand, machine-learning models are built for prediction but are often considered as difficult to interpret and are rarely used to extract behavioral findings from the model outputs.

This paper aims at improving the understanding of the relative strengths and weaknesses of logit models and machine learning for modeling travel mode choices. It compared logit and machine-learning models side by side using cross validation to discover their predictive and interpretability capabilities. The results showed that the best-performing machine-learning model, the random forest model, significantly outperforms the logit models both at individual and aggregate levels. Somewhat surprisingly, the mixed logit model underperformed the multinomial logit model in terms of the out-of-sample predictive quality, which may result from overfitting. Moreover, to interpret the RF model, we applied three tech-

niques, including variable importance, partial dependence plots, and sensitivity analysis, to extract behavioral insights from the model outputs. Some of the results (e.g., on travel time by transit) were illuminating, revealing additional behavioral information compared to the MNL model due to the RF model’s ability to better capture the nonlinear effects of an independent variable on the choice output. This indicates that machine learning can, at minimum, serve as an exploratory analysis tool to reveal nonlinearities; researchers can then apply such information to specify logit models that can better represent behavioral preferences and have better predictive capabilities, which should be much more efficient than a hand-curating procedure typically done with statistical models.

However, a direct application of standard approaches to compute behavioral outputs from the RF model such as marginal effects and elasticities leads to unrealistic behavioral findings. This is because the machine-learning models studied here lack the behavioral assumptions applied in logit models (i.e., constraining alternative-specific attributes to only affect the utility of their corresponding alternatives). Moreover, the RF algorithm, a tree-based model, is not capable of distinguishing “between nodes” and “out-of-boundary” observations. To address these limitations, the paper proposed an alternative approach for estimating arc elasticities and marginal effects. This approach imposes behavioral constraints to the process of generating behavioral outputs from machine-learning model results, which leads to findings that are more realistic and somewhat comparable with the MNL’s outputs.

Overall, these results are encouraging and identify many new research directions in applying machine learning to model travel behavior and forecast travel demand. Prediction and interpretation are two major topics in modeling individual choice behavior. Traditionally, each approach has focused on one aspect and ignored the other. We have demonstrated that both approaches can be applied to make predictions and infer behavior. Nonetheless, there are several major topics in travel-behavior research that we have not examined in depth. First is regarding preference heterogeneity. The development of the mixed logit model has mostly been driven by its capability to capture both observed and unobserved preference heterogeneity among individuals. Machine learning models have limited power in accommodating observed preference heterogeneity through a market segmentation approach but they cannot account for unobserved heterogeneity because they do not recognize a panel data structure. Second is on mechanisms to correct the reporting bias associated with stated-preference data. The stated-preference data are generally considered as containing reporting bias due to their hypothetical nature. Joint revealed-preference and stated-preference models have been proposed to correct for this bias (Train, 2009), but to our knowledge no machine learning algorithms allow such a joint estimation process. Finally, the differences in the output formats of logit models and machine learning. As discussed in Section 6.1.1, the logit model outputs a choice probability for each alternative whereas machine learning outputs a class (i.e. predicted mode). To facilitate the comparison with machine learning, the common practice is to alter the outcome of logit models (o.e. assigning the alternative with the highest choice probabilities as the predictive class). It is not clear if this practice has major implications on the predictive accuracy results.

There is great potential in merging important ideas from machine learning and logit models to develop more refined models for travel-behavior research. Besides addressing

the limitations mentioned above, other possible research directions include: 1) examining which machine learning models are more suitable than others for behavioral analysis; 2) incorporating the behavioral assumption that assume alternative-specific, which are enabled by the “layered” data structure of logit models, into machine learning algorithms.

## Acknowledgements

This research was partly funded by the Michigan Institute of Data Science (MIDAS) and by Grant 7F-30154 from the Department of Energy.

## References

- Abrantes, P. A., Wardman, M. R., 2011. Meta-analysis of UK values of travel time: An update. *Transportation Research Part A: Policy and Practice* 45 (1), 1–17.
- Ben-Akiva, M. E., Lerman, S. R., Lerman, S. R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Vol. 9. MIT press.
- Breiman, L., 1996. Bagging predictors. *Machine Learning* 24 (2), 123–140.
- Breiman, L., 2001. Random forests. *Machine Learning* 45 (1), 5–32.
- Breiman, L., 2017. *Classification and Regression Trees*. Routledge.
- Brownstone, D., Train, K., 1998. Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics* 89 (1-2), 109–129.
- Cawley, G. C., Talbot, N. L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11 (Jul), 2079–2107.
- Chen, X. M., Zahiri, M., Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C: Emerging Technologies* 76, 51–70.
- Cherchi, E., Cirillo, C., 2010. Validation and forecasts in models estimated from multiday travel survey. *Transportation Research Record: Journal of the Transportation Research Board* (2175), 57–64.
- Christopher, M. B., 2016. *Pattern Recognition and Machine Learning*. Springer-Verlag New York.
- Dunnett, C. W., 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50 (272), 1096–1121.
- Farrar, D. E., Glauber, R. R., 1967. Multicollinearity in regression analysis: The problem revisited. *The Review of Economic and Statistics*, 92–107.
- Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.
- Golshani, N., Shabanpour, R., Mahmoudifard, S. M., Derrible, S., Mohammadian, A., 2018. Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. *Travel Behaviour and Society* 10, 21–32.
- Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications* 78, 273–282.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics New York, NY, USA:.
- Hensher, D. A., Greene, W. H., 2003. The mixed logit model: The state of practice. *Transportation* 30 (2), 133–176.
- Hensher, D. A., Rose, J. M., Greene, W. H., 2005. *Applied Choice Analysis: A Primer*. Cambridge University Press.
- Hess, S., Rose, J. M., Polak, J., 2010. Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research Part D: Transport and Environment* 15 (7), 405–417.
- Ho, T. K., 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8), 832–844.

- Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks* 13 (2), 415–425.
- Jenkins, K., Jan 2018. New app reinvents University bus system to be more like Uber. *The Michigan Daily*. URL <https://www.michigandaily.com/section/research/new-app-helps-turns-university-bus-system-demand-service>.
- Karlaftis, M. G., Vlahogianni, E. I., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies* 19 (3), 387–399.
- Klaiber, H. A., von Haefen, R. H., 2011. Do random coefficients and alternative specific constants improve policy analysis? An empirical investigation of model fit and prediction. *Environmental and Resource Economics*, 1–17.
- Last, M., Maimon, O., Minkov, E., 2002. Improving stability of decision trees. *International Journal of Pattern Recognition and Artificial Intelligence* 16 (02), 145–159.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News* 2 (3), 18–22. URL <http://CRAN.R-project.org/doc/Rnews/>
- Mahéo, A., Kilby, P., Van Hentenryck, P., 2017. Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science*.
- McCallum, A., Nigam, K., et al., 1998. A comparison of event models for naive bayes text classification. In: *AAAI-98 Workshop on Learning for Text Categorization*. Vol. 752. Citeseer, pp. 41–48.
- McFadden, D., 1973. Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (Ed.), *Frontiers in Econometrics*. Academic Press New York, New York, NY, pp. 105–142.
- McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. *Journal of applied Econometrics* 15 (5), 447–470.
- Menard, S., 2004. Six approaches to calculating standardized logistic regression coefficients. *The American Statistician* 58 (3), 218–223.
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. URL <https://CRAN.R-project.org/package=e1071>
- Miller, T., 2017. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*.
- Molnar, C., 2018. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.
- Mullainathan, S., Spiess, J., 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31 (2), 87–106.
- Omrani, H., 2015. Predicting travel mode of individuals by machine learning. *Transportation Research Procedia* 10, 840–849.
- Omrani, H., Charif, O., Gerber, P., Awasthi, A., Trigano, P., 2013. Prediction of individual travel mode with evidential neural network model. *Transportation Research Record: Journal of the Transportation Research Board* 2399 (1), 1–8.
- Peters, A., Hothorn, T., 2017. ipred: Improved Predictors. R package version 0.9-6. URL <https://CRAN.R-project.org/package=ipred>
- Quinlan, J. R., 2014. C4. 5: Programs for Machine Learning. Elsevier.
- Rice, J., 2006. *Mathematical Statistics and Data Analysis*. Nelson Education.
- Ridgeway, G., 2017. gbm: Generalized Boosted Regression Models. R package version 2.1.3. URL <https://CRAN.R-project.org/package=gbm>
- Ripley, B., 2016. tree: Classification and Regression Trees. R package version 1.0-37. URL <https://CRAN.R-project.org/package=tree>
- Train, K. E., 2009. *Discrete choice methods with simulation*. Cambridge University Press.
- Vellido, A., Martín-Guerrero, J. D., Lisboa, P. J., 2012. Making machine learning models interpretable. In: *ESANN*. Vol. 12. Citeseer, pp. 163–172.
- Venables, W. N., Ripley, B. D., 2002. *Modern Applied Statistics with S*, 4th Edition. Springer, New York,

iSBN 0-387-95457-0.

URL <http://www.stats.ox.ac.uk/pub/MASS4>

Wang, F., Ross, C. L., 2018. Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transportation Research Record: Journal of the Transportation Research Board*.

Xie, C., Lu, J., Parkany, E., 2003. Work travel mode choice modeling with data mining: decision trees and neural networks. *Transportation Research Record: Journal of the Transportation Research Board* (1854), 50–61.

Yan, X., Levine, J., Zhao, X., 2018. Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. *Transportation Research Part C: Emerging Technologies*.

URL <http://www.sciencedirect.com/science/article/pii/S0968090X18310398>

Zhang, H., 2004. The optimality of naive Bayes. *AA* 1 (2), 3.

Zhang, Y., Xie, Y., 2008. Travel mode choice modeling with support vector machines. *Transportation Research Record: Journal of the Transportation Research Board* (2076), 141–150.