Travel Behaviour and Society 20 (2020) 22–35

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models

Xilei Zhao[a], Xiang Yan[b,*], Alan Yu[c], Pascal Van Hentenryck[d]

[a] Department of Civil and Coastal Engineering, University of Florida, USA
[b] Department of Urban and Regional Planning, University of Florida, USA
[c] Department of Electrical Engineering and Computer Science, University of Michigan, USA
[d] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, USA

ARTICLE INFO

Keywords:
Machine learning
Logit models
Mobility-on-demand
Stated-preference survey
Travel mode choice
Travel behavior

ABSTRACT

Some recent studies have shown that machine learning can achieve higher predictive accuracy than logit models. However, existing studies rarely examine behavioral outputs (e.g., marginal effects and elasticities) that can be derived from machine-learning models and compare the results with those obtained from logit models. In other words, there has not been a comprehensive comparison between logit models and machine learning that covers both prediction and behavioral analysis, two equally important subjects in travel-behavior study. This paper addresses this gap by examining the key differences in model development, evaluation, and behavioral interpretation between logit and machine-learning models for mode-choice modeling. We empirically evaluate the two approaches using stated-preference survey data. Consistent with the literature, this paper finds that the best-performing machine-learning model, random forest, has significantly higher predictive accuracy than multinomial logit and mixed logit models. The random forest model and the two logit models largely agree on several aspects of the behavioral outputs, including variable importance and the direction of association between independent variables and mode choice. However, we find that the random forest model produces behaviorally unreasonable arc elasticities and marginal effects when these behavioral outputs are computed from a standard approach. After the introduction of some modifications that overcome the limitations of tree-based models, the results are improved to some extent. There appears to be a tradeoff between predictive accuracy and behavioral soundness when choosing between machine learning and logit models in mode-choice modeling.

1. Introduction

Emerging shared mobility services, such as car sharing, bike sharing, ridesouring, 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 (a type of statistical models), most notably the logit family such as the multinomial logit model (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 modeling individual choice behavior.

Machine learning and conventional statistical models seek to understand the data structure based on different approaches. The logit models, like many other statistical models, are based on a theoretical foundation which is mathematically proven, but this requires the input data to satisfy assumptions such as the random utility maximization theory (Ben-Akiva et al., 1985). On the other hand, many popular machine-learning methods are non-parametric, such as decision trees, random forest, support vector machine, and neural networks, relying on computers to probe the data for its structure without a theory of what the underlying data structure should look like. In other words, while a logit model presupposes a certain type of structure of the data with its behavioral and statistical assumptions, machine learning “lets the data speak for itself” and hence allows forming more flexible modeling structures, which can often lead to higher predictive capability (e.g., higher out-of-sample predictive accuracy).

A number of recent empirical studies have verified that machine learning can outperform logit models in terms of predictive capability (e.g. Xie et al., 2003; Zhang and Xie, 2008; Hagenauer and Helbich, 2017; Wang and Ross, 2018; Lhéritier et al., 2018; Rasouli and

[*] Corresponding author at: 1480 Inner Rd, Gainesville, FL 32601, USA.
E-mail addresses: xilei.zhao@essie.ufl.edu (X. Zhao), xiangyan@ufl.edu (X. Yan), alanyu@umich.edu (A. Yu), pvh@isye.gatech.edu (P. Van Hentenryck).

https://doi.org/10.1016/j.tbs.2020.02.003
Received 22 August 2019; Received in revised form 25 January 2020; Accepted 5 February 2020
Available online 18 February 2020

Timmermans, 2014; Lindner et al., 2017; Cheng et al., 2019; Golshani et al., 2018). With a primary focus on prediction, however, these studies have paid less attention to the interpretation of machine-learning models for deriving behavioral insights. Notably, while some studies have analyzed the variable importance results of machine-learning models (e.g., Hagenauer and Helbich, 2017; Cheng et al., 2019; Wang and Ross, 2018; Lhéritier et al., 2018), other important aspects of behavioral outputs, such as direction of association, marginal effects, and elasticities, are rarely examined in the current literature. In many policy contexts, analysts not only need to know what factors are important determinants of mode choice, but also the direction and magnitude of their influence. For example, policymakers often apply the elasticities of auto use with respect to key policy variables (e.g., transit level of service) to inform transportation planning. Therefore, besides predictive performance, evaluating the credibility of machine learning for mode-choice modeling requires an investigation of its capability to generate sound behavioral outputs. This study presents such an analysis by comparing both the predictive accuracy and behavioral outputs of machine-learning models and logit models.

Existing studies that compare logit models and machine learning for modeling travel mode choice have two other major limitations. 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. However, to our knowledge, no existing study has compared the predictive performance and behavioral outputs of a mixed logit model with those of machine-learning models. Second, existing studies rarely discussed the fundamental differences in the practical 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 outcomes carry significant implications for model comparison. These differences and their implications, although touched on in some studies (e.g., Omrani et al., 2013), have not been thoroughly discussed.

This paper tries to bridge these gaps. It provides a comprehensive discussion of the fundamental differences between logit models and machine learning in modeling travel mode choices and also an empirical evaluation of the two approaches using stated-preference (SP) survey data on a proposed mobility-on-demand 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 similarities and differences in the practical applications of the two approaches, with a particular focus on the implications on the predictive performance of each approach and their capabilities to facilitate behavioral interpretation. The paper then compares the performance of two logit models (MNL and mixed logit) and seven machine-learning classifiers, including Naive Bayes (NB), classification and regression trees (CART), boosting trees (BOOST), bagging trees (BAG), random forest (RF), support vector machine (SVM), and neural networks (NN), in predicting individual choices of four travel modes and their respective market shares. Moreover, we interpret two logit models (MNL and mixed logit) and two machine-learning models (RF and NN) to extract behavioral insights and compare these findings.

The rest of the paper is organized as follows. The next section provides a brief review of the existing literature in modeling travel mode choices using machine-learning and logit models. Section “Fundamentals of machine-learning and logit models” compares the similarities and differences in the practical application of machine-learning and logit models for travel mode choice, which includes model development, evaluation, and interpretation. Section “The data for empirical evaluation” introduces the data used for empirical study and Section “Models examined and their specifications” describes the machine-learning and logit models examined and their specifications. Section “Comparison of empirical results” evaluates these models in terms of predictive capability and interpretability. Lastly, Section “Discussion and conclusion” 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 IIA assumption and its inability to account for taste variations among different individuals. To address these limitations, researchers 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 are effective in modeling individual travel behavior (Xie et al., 2003; Zhang and Xie, 2008; Omrani, 2015; Hagenauer and Helbich, 2017; Golshani et al., 2018; Wang and Ross, 2018; Wong et al., 2018; Lhéritier et al., 2018; Lindner et al., 2017; Rasouli and Timmermans, 2014). 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 travel mode choices for commuting trips taken by residents in the San Francisco Bay area. These machine-learning methods exhibited better performance than the MNL model in terms of prediction. 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. More recently, Lhéritier et al. (2018) found that the RF model outperforms the standard and the latent class MNL model in terms of accuracy and computation time, with less modeling effort.

It is not surprising that machine-learning classifiers can perform better than logit models in prediction tasks. Unlike logit models that predetermine a model structure (usually linear formulations), machine learning allows for more flexible model structures, which can reduce the model's incompatibility with the empirical data (Xie et al., 2003; Christopher, 2016). 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 (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 features and the output variable) and pay less attention to increasing the model's out-of-sample predictive accuracy (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 resulted 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 interpretation, 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 outputs of many machine-learning models are indeed not directly interpretable as one may need hundreds of parameters to describe a deep neural nets or hundreds of decision trees to understand a RF model. Nonetheless, with recent development in interpretable/explainable machine learning, a wide range of machine learning interpretation tools have been invented and applied to extract knowledge from the black-box models to facilitate decision-making (Doshi-Velez and Kim, 2017; Athey, 2017; Molnar, 2018). Variable importance, a widely-used machine-learning interpretation tool, has been applied in several recent studies that used machine learning to model travel behavior (e.g.., Hagenauer and Helbich, 2017; Golshani et al., 2018; Cheng et al., 2019). For example, Hagenauer and Helbich (2017) compared the variable importance results of machine-learning methods with MNL. Cheng et al. (2019) applied variable importance to evaluate the relative importance of explanatory variables of the RF model to help provide important insights into formulating transport policies. Examining these variable importance 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. In fact, traditional mode-choice applications have been more focused on behavioral outputs such as elasticity, marginal effect, value of time, and willingness-to-pay measures than on predicting individual travel mode choice. However, these behavioral outputs were rarely examined in machine-learning-based mode-choice analysis (Xie et al., 2003; Hagenauer and Helbich, 2017; Wang and Ross, 2018; Cheng et al., 2019). This paper thus extends the current literature by comparing various behavioral outputs from logit models and machine-learning models, including variable importance, direction of association between key variables and mode choice (based on examining partial dependence plots), elasticity, and marginal effects.

Moreover, this paper discusses key differences in the practical applications of machine learning and logit models that have bearings on predictive performance and behavioral outputs, 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.

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
$\boldsymbol{X}$	Input data for logit models containing P features with N observations for K alternatives
$\boldsymbol{X_{k,p}}$	Feature p for alternative k, $k = 1, \ldots, K$ of $\boldsymbol{X}$
$\boldsymbol{X}_{k,-p}$	All the features except p for alternative k, $k = 1, \ldots, K$ of $\boldsymbol{X}$
$\boldsymbol{X}_{ik}$	A row-vector for the ith observation for alternative k, $k = 1, \ldots, K$
$\boldsymbol{X}_k$	Input data for alternative k, $\boldsymbol{X}_k = [\boldsymbol{X}_{.k1};\ldots;\boldsymbol{X}_{.kP}]$ where $\boldsymbol{X}_{.kp} = [X_{1kp}, \ldots, X_{Nkp}]$
$\boldsymbol{X}_i$	The ith observation of $\boldsymbol{X}$, $\boldsymbol{X}_i = [\boldsymbol{X}_{i.1}, \ldots, \boldsymbol{X}_{i.P}]$ where $\boldsymbol{X}_{i.p} = [X_{i1p};\ldots;X_{iKp}]$
$\boldsymbol{X}_p$	The feature p of $\boldsymbol{X}$, $\boldsymbol{X}_p = [\boldsymbol{X}_{1.p}, \ldots, \boldsymbol{X}_{N.p}]$ where $\boldsymbol{X}_{i.p} = [X_{i1p};\ldots;X_{iKp}]$
$\boldsymbol{Z}$	Input data for machine-learning models containing P features and N observations
$\boldsymbol{Z}_p$	Feature p of $\boldsymbol{Z}$
$\boldsymbol{Z}_{-p}$	All the features except p of $\boldsymbol{Z}$
$\boldsymbol{Z}_i$	ith observation of $\boldsymbol{Z}$, $\boldsymbol{Z}_i = [Z_{i1}, \ldots, Z_{iP}]$
$U_k(\boldsymbol{X}_k\|\boldsymbol{\beta}_k)$	Utility function for mode k
$\boldsymbol{\beta}_k$	Parameter vector for alternative k of MNL model
$\boldsymbol{\beta}$	Parameter matrix of MNL model, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$
$\widehat{\boldsymbol{\beta}}$	Estimated parameter matrix of MNL model
ε_k	Random error for alternative k of MNL model
$\boldsymbol{Y}$	Output mode choice data
$\widehat{Y_i}$	Estimated mode choice for observation i
θ	Parameter or hyperparameter vector for machine-learning models
$\widehat{\theta}$	Estimated parameter or hyperparameter vector
$f(\boldsymbol{Z}\|\theta)$	Machine-learning models as a function of $\boldsymbol{Z}$ given θ
p_{ik}	Probability of choosing alternative k of observation i
$\widehat{p}_{ik}$	Predicted probability for choosing alternative k of observation i
$I_k(\widehat{Y_i})$	Indicator function that equals to 1 if $\widehat{Y_i} = k$
$P_k(\boldsymbol{X}\|\widehat{\boldsymbol{\beta}})$	Aggregate-level prediction for mode k as a function of $\boldsymbol{X}$ given $\widehat{\boldsymbol{\beta}}$ for logit models
$Q_k(\boldsymbol{Z}\|\widehat{\theta})$	Aggregate-level prediction for mode k as a function of $\boldsymbol{Z}$ given $\widehat{\theta}$ for machine-learning models
$E_k(\cdot)$	Arc elasticity for alternative k
$M_k(\cdot)$	Marginal effect for alternative k
Δ	Constant

3. Fundamentals of machine-learning and logit models

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

3.1. Model development

Logit and machine-learning models approach the mode-choice prediction problem from different perspectives. Logit models are based on the assumption that individuals select a mode among a set of viable 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 utility to a traveler, and specifies, for each mode, a utility function with two components: A component that represents the effects of observed variables and a random error 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

Table 3
A comparison of machine-learning and logit models.

	Logit Models	Machine-Learning Models
Model formulations	$U_k(\boldsymbol{X}_k\|\boldsymbol{\beta}_k) = \boldsymbol{\beta}_k^T\boldsymbol{X}_k + \varepsilon_k$ $p_{ik} = \frac{\exp\left(\boldsymbol{\beta}_k^T\boldsymbol{X}_{ik}\right)}{\sum_{p=1}^{K}\exp\left(\boldsymbol{\beta}_p^T\boldsymbol{X}_{ik}\right)}, k \in \left\{1, \ldots, K\right\}$	$\boldsymbol{Y} = f(\boldsymbol{Z}\|\theta), \boldsymbol{Y} \in \{1, \ldots, K\}$
	MNL, mixed logit, nested MNL, generalized MNL	NB, CART, BAG, BOOST, RF, SVM, NN
Prediction types	Class probability: $p_{i1}, \ldots, p_{iK}$	Classification: k, $k \in \{1, \ldots, K\}$
Input data	$\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K]$	$\boldsymbol{Z}$
Model topologies	Layer structure	Layer structure, tree structure, case-based reasoning, etc.
Optimization methods	Maximum likelihood estimation, simulated maximum likelihood	Back propagation, gradient descent, recursive partitioning, structural risk minimization, maximum likelihood, etc.
Commonly-used evaluation criteria	(Adjusted) McFadden's pseudo R^2, AIC, BIC	Resampling-based measures, e.g., cross validation
Individual-level mode prediction	$\text{argmax}_k(\hat{p}_{i1}, \ldots, \hat{p}_{iK})$	$\hat{Y}_i$
Aggregate-level mode share prediction	$P_k(\boldsymbol{X}_k\|\hat{\boldsymbol{\beta}}_k) = \sum_i^N \hat{p}_{ik}/N$	$Q_k(\boldsymbol{Z}\|\hat{\theta}) = \sum_i^N \hat{p}_{ik}/N$
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 of feature p for alternative k	$E_k(\boldsymbol{X}_{k,p}) = \frac{[P_k(\boldsymbol{X}_{k,-p}, \boldsymbol{X}_{k,p}\cdot(1+\Delta) \mid \hat{\boldsymbol{\beta}}_k) - P_k(\boldsymbol{X}_k \mid \hat{\boldsymbol{\beta}}_k)] / P_k(\boldsymbol{X}_k \mid \hat{\boldsymbol{\beta}}_k)}{\|\Delta\|}, k \in \left\{1, \ldots, K\right\}$	$E_k(\boldsymbol{Z}_p) = \frac{[Q_k(\boldsymbol{Z}_{-p}, \boldsymbol{Z}_p\cdot(1+\Delta) \mid \hat{\theta}_k) - Q_k(\boldsymbol{Z} \mid \hat{\theta}_k)] / Q_k(\boldsymbol{Z} \mid \hat{\theta}_k)}{\|\Delta\|}, k \in \left\{1, \ldots, K\right\}$
Marginal effects of feature p for alternative k	$M_k(\boldsymbol{X}_{k,p}) = \frac{P_k(\boldsymbol{X}_{k,-p}, \boldsymbol{X}_{k,p} + \Delta \mid \hat{\boldsymbol{\beta}}_k) - P_k(\boldsymbol{X}_k \mid \hat{\boldsymbol{\beta}}_k)}{\|\Delta\|}, k \in \{1, \ldots, K\}$	$M_k(\boldsymbol{Z}_p) = \frac{Q_k(\boldsymbol{Z}_{-p}, \boldsymbol{Z}_p + \Delta \mid \hat{\theta}_k) - Q_k(\boldsymbol{Z} \mid \hat{\theta}_k)}{\|\Delta\|}, k \in \{1, \ldots, K\}$

$$U_k(\boldsymbol{X}_k|\boldsymbol{\beta}_k) = \boldsymbol{\beta}_k^T\boldsymbol{X}_k + \varepsilon_k, \tag{1}$$

where $\boldsymbol{\beta}_k$ contains the coefficients to be estimated and ε_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, the probability of choosing alternative *k* for individual *i* is

$$p_{ik} = \frac{\exp(\boldsymbol{\beta}_k^T\boldsymbol{X}_{ik})}{\sum_{p=1}^{K}\exp\left(\boldsymbol{\beta}_p^T\boldsymbol{X}_{ik}\right)}. \tag{2}$$

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

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\frac{\exp(\boldsymbol{\beta}_k^T\boldsymbol{X}_{ik})}{\sum_{p=1}^{K}\exp\left(\boldsymbol{\beta}_p^T\boldsymbol{X}_{ik}\right)}\right]. \tag{3}$$

Maximum likelihood estimation can then be applied to obtain the "best" utility coefficients $\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}}\boldsymbol{L}(\boldsymbol{\beta})$. By plugging $\hat{\boldsymbol{\beta}}$ into Eq. (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-term distribution. However, these models are more difficult to estimate. Notably, mixed logit models 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 $\boldsymbol{X}_i$ to the output layer, $[p_{ik}, \ldots, p_{iK}]^T$.

Machine-learning models, by contrast, approach 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 $\boldsymbol{Z}$ to the output target $\boldsymbol{Y}, \boldsymbol{Y}_i \in \{1, \ldots, K\}$, as

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

where θ represents the unknown parameter vector for parametric models like NB and the hyperparameter vector for non-parametric models such as SVM, CART, and RF. Unlike logit models that predetermine a model structure (usually linear formulations) and make specific assumptions for the parameters and error-term distributions, many machine-learning models are non-parametric, which allows for more flexible model structures to be learned from the data. In contrast 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, and 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.

Fitting a logit model necessarily requires information on both the chosen alternative and one or more of its competing alternatives. In other words, when the attributes of non-chosen alternatives are not observed, their values (at least for one of the non-chosen alternatives) need to be imputed in order to specify a logit model. By contrast, when modeling travel mode choices, machine-learning algorithms would still work even if the attributes of non-chosen alternatives are not modeled.[1]. However, not accounting for the information on the non-chosen alternatives is not ideal or even erroneous, since a given mode-choice outcome is a result of the differences in attribute values across alternatives (i.e., modal competition) rather than a result of the characteristics of the chosen alternative itself. Therefore, we believe that, like logit models, the non-chosen alternatives should also be considered in a machine learning model.

[1] Some previous studies have indeed only considered attribute values of the chosen mode in their machine learning models (e.g., Xie et al., 2003; Wang and Ross, 2018)

3.2. Model evaluation

3.2.1. Evaluation measures for model selection

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 data noise. 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.

The selection of logit models is usually based on theoretical measures. 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. All three measures penalize the likelihood for including too many "useless" features. The adjusted McFadden's pseudo R^2 is the most commonly used, 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 often used to compare models with different number of variables.

For machine learning, cross validation is usually conducted to evaluate models with different variable selections, model types, and choices of hyperparameters. The best model is 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 performance metrics; 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 on their predictive capabilities.

Note that both theoretical measures and resampling-based measures can be applied to evaluate machine-learning and logit models. Both types of measures have pros and cons. For example, AIC and BIC are often easier and faster to compute, but they may have poor performance when the assumed error-term distribution is not consistent with the realized residual distribution and/or the number of observations is not large enough compared to the number of attributes. By contrast, resampling-based measures such as cross validation do not rely on these assumptions and can be used to compare different classes of models.

It is worthwhile to point out that, 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, a "worse-fitting" model may be chosen due to greater 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 logically have a negative sign. Therefore, when a model produces a positive sign for wait time, it may be disregarded even if it has a better fit. Conventionally, predictive accuracy has been the sole criterion for deciding the best model in the machine learning. With a growing emphasis on machine-learning interpretation in recent years, however, some researchers have suggested that machine-learning models should be evaluated by both predictive accuracy and descriptive accuracy (Murdoch et al., 2019).

3.2.2. Individual-level and aggregate-level prediction

When evaluating the predictive accuracy of logit models at the individual level, a common practice in the literature is to assign an outcome to the alternative with the largest choice probability, i.e.,

$$\text{argmax}_k(\hat{p}_{i1},\ \ldots,\hat{p}_{iK}). \tag{5}$$

This approach produces a choice outcome as most off-the-shelf packages of machine-learning models does. Some machine-learning models (such as decision trees, RF, and NN) have a similar internal prediction mechanism, that is, predicting the choice probability for each alternative first and then outputting the alternative with the highest probability. Nevertheless, other machine-learning models such as SVM directly output the choice outcome without estimating the choice probabilities.

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

$$P_k\left(\boldsymbol{X}_k\middle|\hat{\boldsymbol{\beta}}_k\right) = \sum_i^N \hat{p}_{ik}/N, \tag{6}$$

and, for machine-learning methods, it is usually computed by

$$Q_k\left(\boldsymbol{Z}\middle|\hat{\theta}\right) = \sum_i^N I_k(\hat{Y}_i)/N. \tag{7}$$

However, using the proportion of predicted class labels to approximate the market share may not be ideal. Instead, as mentioned above, many machine-learning models can directly predict class probabilities at the individual level[2], so, for these models, we can estimate the aggregate-level market share by averaging the individual-level class probabilities:

$$Q_k\left(\boldsymbol{Z}\middle|\hat{\theta}\right) = \sum_i^N \hat{p}_{ik}/N. \tag{8}$$

By construction, machine-learning models focus on individual-level prediction. By contrast, the calibration of the logit models aims to replicate aggregate market shares rather than to predict individual choice (Ben-Akiva et al., 1985; Hensher et al., 2005). Thus, the predictive accuracy of these models may differ at the individual level and the aggregate level, and which of them to be prioritized depends on the context. While aggregate-level market share is usually given a greater weight in many policy applications, activity-based modeling requires precision on individual hit to avoid accumulated errors.

3.3. Model interpretation

Interpreting the outputs of logit models is straightforward and

[2] For models like SVM, Eq. (7) can be used to approximate the market shares directly, or some probability estimation methods (e.g., Wu et al., 2004) can be adopted to estimate the individual-level choice probability for class k.

intuitive. Like any other statistical model, researchers can readily examine the sign, magnitude, and statistical significance of the model coefficients. Researchers may conduct further behavioral analysis, 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 have explicit mathematical formulations and derivations, which allows modelers to clearly explain what happens "behind the scene".

By contrast, machine-learning models are often criticized for being "black-box" and lacking interpretability, which is a major barrier for machine learning in many real-world applications (Klaiber and von Haefen, 2011). To overcome this issue, a variety of machine-learning interpretation tools have been invented recently (e.g. Friedman, 2001; Goldstein et al., 2015; Molnar, 2018). The most commonly-used tools include variable importance and partial dependence plots (Molnar, 2018). Variable importance measures show the relative importance of each input variable in predicting the response variable. Partial dependence plots measure the influence of a variable $\boldsymbol{Z}_p$ on the log-odds or probability of choosing a mode k after accounting for the average effects of all other variables (Friedman, 2001). Notably, partial dependence plots may reveal causal relationships if the machine-learning model is accurate and the domain knowledge supports the underlying causal structure (Zhao and Hastie, 2019).

Practically, some behavioral outputs that one can extract from the logit models such as marginal effects and elasticities can also be obtained from machine-learning models by performing a sensitivity analysis. 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 lack the behavioral foundation required to obtain these measures.

In a machine-learning model, the arc elasticity for feature p of alternative k can be obtained by

$$E_k(\boldsymbol{Z}_p) = \frac{[Q_k(\boldsymbol{Z}_{-p}, \boldsymbol{Z}_p \cdot (1 + \Delta)|\widehat{\theta}) - Q_k(\boldsymbol{Z}|\widehat{\theta})]/Q_k(\boldsymbol{Z}|\widehat{\theta})}{|\Delta|}, \tag{9}$$

and the marginal effect for feature p of alternative k can be computed as

$$M_k(\boldsymbol{Z}_p) = \frac{Q_k(\boldsymbol{Z}_{-p}, \boldsymbol{Z}_p + \Delta|\hat{\theta}) - Q_k(\boldsymbol{Z}|\hat{\theta})}{|\Delta|}. \tag{10}$$

However, existing studies that compare machine-learning and logit models rarely compare these behavioral outputs. Since the goals of mode-choice studies are often in deriving generic behavioral insights instead of merely predicting individual mode choice, these comparisons are important to more thoroughly assess the applicability of machine learning in travel-behavior analysis.

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 on 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), fully integrating high-frequency fixed-route bus services and micro-transit services, has replaced the existing bus system (see Fig. 1). 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 stated-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. (2019).

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 (i.e., faculty vs staff vs student). 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 presents a descriptive profile of these variables.

After extracting the data from the SP survey, we processed the data and ruled out the concern for multicollinearity (Farrar and Glauber, 1967). The existence of multicollinearity can inflate the variance and negatively impact the predictive power of the models. We computed the variance inflation factor to determine which variables are highly correlated with other variables and found 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 machine-learning and logit models examined in this study. Since our dataset has a panel structure (repeated observations from the same individual), 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 were examined, including simple ones like NB and CART, and more complex ones such as RF, BOOST, BAG, SVM, and NN. Most previous mode choice studies only examined a subset of these models (Xie et al., 2003; Omrani, 2015; Wang and Ross, 2018; Chen et al., 2017).

5.1. Logit models

We have already introduced the MNL model formulation in SubSection 3.1, 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

$$U_{ikt} = \boldsymbol{\beta}_{ik}^T \boldsymbol{X}_{ikt} + \varepsilon_{ikt}, \tag{11}$$

where ε_{ikt} is the independent and identically distributed random error across people, alternatives, and time. Hence, conditioned on $\boldsymbol{\beta}$, the

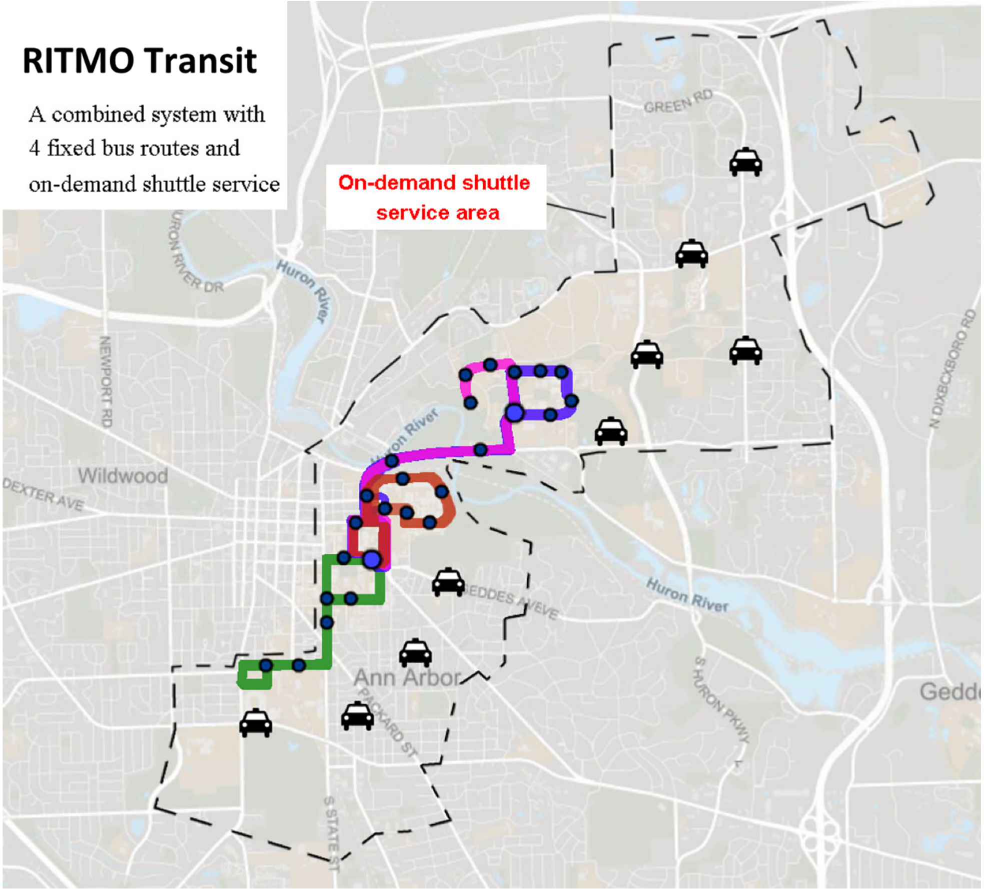


Fig. 1. RITMO Transit: The new transit system featured with four high-frequency bus routes and on-demand shuttles serving approximately 2-mile-radius area of the University of Michigan campus.

Table 4
A descriptive profile of the variables examined in this study.

Variable	Description	Category	%	Min	Max	Mean	SD
Dependent Variable							
Mode Choice		Car	14.888				
		Walk	28.965				
		Bike	20.870				
		PT	35.278				
Independent Variables							
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				
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[a]	Current travel mode is PT or not	PT	34.652				
		Not PT	65.348				

[a] Current_Mode_PT is not included for machine-learning models, since it can be represented by a linear combination of Current_Mode_Car, Current_Mode_Walk, and Current_Mode_Bike.

probability of an individual making a sequence of choices (i.e., $\boldsymbol{j} = \{j_1, j_2, \ldots, j_\tau\}$) is

$$\boldsymbol{L}_{i\boldsymbol{j}}(\boldsymbol{\beta}) = \prod_{t=1}^{\tau} \left[\frac{\exp\left(\boldsymbol{\beta}_{ij_t}^{T} \boldsymbol{X}_{ij_t t}\right)}{\sum_k \exp\left(\boldsymbol{\beta}_{ik}^{T} \boldsymbol{X}_{ikt}\right)} \right]. \tag{12}$$

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

$$p_{ik\boldsymbol{j}} = \int \boldsymbol{L}_{i\boldsymbol{j}}(\boldsymbol{\beta}) g(\boldsymbol{\beta}) d\boldsymbol{\beta}, \tag{13}$$

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).

The MNL model in this study is specified as follows: 1) The utility function of *Car* includes mode-specific parameters for TT_Drive, Parking_Cost, Income, CarPerCap, and Current_Mode_Car; 2) the utility function of *Walk* includes mode-specific parameters for TT_Walk, 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, 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, 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 specification. Moreover, in order to accommodate individual preference heterogeneity (i.e., taste variations among different individuals), coefficients on the selected level-of-service variables (i.e., TT_PT and Parking_Cost) are also specified as random parameters. The alternative-specific constant for *PT* is also assumed as a random parameter. These random parameters are all assessed with a normal distribution. We use 1,000 Halton draws to perform the numerical integration. Both the MNL and mixed logit models are estimated using the NLOGIT software.

5.2. Machine-learning models

We now briefly describe the machine-learning models examined in the paper. The hyperparameters of non-parametric models are tuned using grid search via the R package *caret* (Kuhn et al., 2008).

5.2.1. Naive Bayes

The NB model is constructed using Bayes' Theorem with the naive assumption that all the features are independent (McCallum et al., 1998). In practice, NB works well as a baseline classifier for large datasets. A limitation of the NB model is that it assumes all the predictors to be completely independent from each other, making it very sensitive when highly correlated predictors exist in the model. Here, the NB model is constructed through the R package *e1071* (Meyer et al., 2017).

5.2.2. Tree-based models

The CART model can build decision trees to predict a classification outcome (Breiman, 2017). However, decision trees are sensitive to noise and susceptible to overfit (Last et al., 2002; Quinlan, 2014). To control its complexity, we prune the tree until the number of terminal nodes is 6. The CART model is obtained via the R package *tree* (Ripley, 2016).

To address the overfitting issues of CART models, researchers have proposed the tree-based ensemble techniques to form more robust, stable, and accurate models than a single decision tree (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. In this study, we apply the gradient boosting machine technique to create the BOOST model (Friedman, 2001). We specify 400 trees, and we set the shrinkage parameter to 0.14 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). 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. RF is also a widely-applied ensemble method. Similar to BAG, RF trains multiple trees using bootstrapping (Ho, 1998). However, RF only uses a random subset of all the variables to train the classification trees, while BAG uses all the variables to train the trees (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 voting among all the classification trees. For the BAG model, 400 classification trees are bagged, with each tree grown without pruning. For the RF model, 500 trees are used and 12 randomly selected variables are considered for each split at the trees' nodes. The R package used for producing the BAG and RF models 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 hyper-plane maximizing the margin between two classes. For a multi-class classification problem, the one-against-one approach is used (Hsu and Lin, 2002). SVM usually performs well with both nonlinear and linear boundaries depending on the specified kernel. However, SVM 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 8, and the gamma parameter for the kernel is set to 0.15. 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. In this paper, a NN model with a single hidden layer of 18 units is used. The connection weights are trained by back propagation with a weight decay constant of 0.4. The R package *nnet* is used to create our NN model (Venables and Ripley, 2002).

6. Comparison of empirical results

This section presents the empirical results. Specifically, we compare the predictive accuracy of the logit models with that of the machine-learning algorithms. In addition, we compare the behavioral outputs of two machine-learning models (i.e., RF and NN) and two logit models (i.e., MNL and mixed logit).

6.1. Predictive accuracy

We applied the 10-fold cross validation approach. As discussed above, cross validation requires splitting 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

Table 5
Mean out-of-sample accuracy of machine-learning and logit models (individual level).

Model	All		Car		Walk		Bike		PT	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MNL	0.647	0.016	0.440	0.044	0.859	0.018	0.414	0.033	0.698	0.029
Mixed logit	0.631	0.008	0.513	0.031	0.797	0.014	0.413	0.038	0.673	0.027
NB	0.584	0.018	0.558	0.035	0.864	0.013	0.372	0.041	0.490	0.042
CART	0.593	0.014	0.428	0.032	0.795	0.022	0.329	0.038	0.653	0.026
BOOST	0.850	0.007	0.790	0.035	0.913	0.012	0.848	0.023	0.825	0.028
BAG	0.854	0.013	0.791	0.017	0.926	0.016	0.861	0.028	0.818	0.029
RF	0.856	0.012	0.797	0.022	0.928	0.016	0.859	0.021	0.820	0.027
SVM	0.772	0.012	0.701	0.027	0.878	0.026	0.681	0.033	0.770	0.026
NN	0.646	0.016	0.434	0.045	0.853	0.025	0.451	0.051	0.679	0.024

Table 6
Mean L1-Norm error for mode share prediction.

Model	Mean	SD
MNL	0.0399	0.0207
Mixed logit	0.0593	0.0268
NB	0.2771	0.0363
CART	0.0463	0.0280
BOOST	0.0291	0.0151
BAG	0.0253	0.0130
RF	0.0248	0.0128
SVM	0.0362	0.0218
NN	0.0493	0.0196

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 (Xie et al., 2003; Hagenauer and Helbich, 2017; Wang and Ross, 2018).

As discussed in SubSection 3.1, the predictive power of the models may differ at the individual level (predicting the mode choice for each observation) 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 aims at predicting individual choice. We compare the mean predictive accuracy at both the individual and aggregatelevels.

6.1.1. Individual-level predictive accuracy

The cross validation results for individual-level predictive accuracy are shown in Table 5. The best-performing model is RF, with a mean predictive accuracy equal to 0.856. However, the accuracy of the MNL and the mixed logit model is only 0.647 and 0.631 respectively, much lower than the RF model.

The predictive accuracy of each model by travel mode is also presented in Table 5. All models predict *Walk* most accurately. All machine-learning models have a mean predictive accuracy value between 0.795 and 0.928, whereas the MNL model has an accuracy of 0.859 and the mixed logit model 0.797. Both logit models and three ensemble machine-learning models (i.e., BOOST, BAG, and RF) 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. The notorious class imbalance problem causes machine-learning classifiers to have more difficulties in predicting the class with fewer observations.

Finally, it is somewhat surprising that the mixed logit model, a model that accounts for individual heterogeneity and has significantly better model fit (adjusted McFadden's pseudo R^2 is 0.536) than the MNL model (adjusted McFadden's pseudo R^2 is 0.365), 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.

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 $\widehat{Q}_k = Q_k(\boldsymbol{Z}|\widehat{\theta})$ represent the true (observed) and predicted market shares for mode k. The L1-norm thus is defined as

$$\sum_{k=1}^{4} \left| Q_k^* - \widehat{Q}_k \right|. \tag{14}$$

The predictive accuracy results of the machine-learning and logit models at the aggregate level are depicted in Table 6. The results show that RF outperforms all the other models, with a prediction error of 0.0248 and a standard deviation of 0.0128. Notably, even though logit models are expected to have good performance for market share predictions, RF has a lower prediction error compared to MNL (0.0399) and mixed logit (0.0593). Again, the MNL model resulted in a higher aggregate-level predictive accuracy than the mixed logit model.

In summary, the results show that RF has the best predictive performance and that logit models only outperform a minority of the machine learning models.

6.2. Model interpretation

This section interprets the results of two logit models (MNL and mixed logit) and two machine-learning models (RF and NN[3]). For the logit models, we interpret the coefficient estimates and calculate some behavioral measures, including marginal effects and arc elasticities. Also, we conduct comparable behavioral analysis on the RF and NN models by applying variable importance and partial dependence plots and by performing a sensitivity analysis.

It should be noted that the behavioral analysis conducted here is far from exhaustive, as mode choice model applications often go beyond what is covered here. In particular, recent advances in mode choice modeling, such as the development of mixed logit and latent class models, are mainly concerned about deriving insights on individual preference heterogeneity.

6.2.1. Variable importance and partial dependence plots

For traditional statistical models, standardized Beta coefficients

[3] The reasons for choosing these two machine-learning models are: 1) RF is the best-performing model among the seven machine-learning classifiers; and 2) NN is one of the most popular machine-learning classifier used for travel mode choice modeling.

Table 7
Outputs of the MNL and mixed logit models.

Variable	Alternatives	MNL			Mixed logit		
		Unstandardized coefficients		*X*-standardized coefficients	Unstandardized coefficients		*X*-standardized coefficients
		β	S.E.	βStd*X*	β	S.E.	βStd*X*
Constants							
Walk	Walk	2.792**	0.194	/	2.098**	0.286	/
Bike	Bike	1.678**	0.181	/	0.837**	0.273	/
PT	PT	3.224**	0.190	/	4.510**	0.560	/
Level-of-service variables							
TT_Drive	Car	− 0.075**	0.005	− 1.138**	− 0.110**	0.009	− 2.052**
TT_Walk	Walk	− 0.146**	0.004	− 2.203**	− 0.168**	0.005	− 3.158**
TT_Bike	Bike	− 0.162**	0.006	− 2.451**	− 0.205**	0.008	− 4.166**
TT_PT	PT	− 0.104**	0.009	− 1.563**	− 0.170**	0.034	− 2.931**
Wait_Time	PT	− 0.156**	0.018	− 0.323**	− 0.470**	0.046	− 1.005**
Parking_Cost	Car	− 0.148**	0.027	− 0.248**	− 0.440**	0.093	− 2.226**
Rideshare	PT	− 0.433**	0.042	− 0.354**	− 1.221**	0.103	− 1.064**
Transfer	PT	− 0.570**	0.047	− 0.368**	− 1.886**	0.174	− 1.298**
Socio-demographic variables							
Income	Car	0.075*	0.030	0.101*	0.026	0.059	0.318
CarPerCap	Car	0.560**	0.083	0.267**	0.505**	0.123	0.357*
Female	Walk, Bike	− 0.175**	0.061	− 0.087**	− 0.289**	0.111	− 0.043
Residential preference variables							
Bike_Walkability	Walk, Bike	0.069*	0.033	0.066*	0.320**	0.059	0.398**
PT_Access	PT	0.113**	0.031	0.115**	0.203	0.143	0.115
Current mode choice							
Current_Mode_Car	Car	1.366**	0.094	0.509**	1.418**	0.154	1.797**
Current_Mode_Walk	Walk	1.289**	0.077	0.633**	1.066**	0.088	0.478**
Current_Mode_Bike	Bike	2.899**	0.120	0.798**	3.628**	0.223	1.031**
Current_Mode_PT	PT	0.093	0.073	0.044	2.837**	0.362	1.313**
Random parameter standard deviations							
PT (Constant)	PT				3.766**	0.209	4.139**
TT_PT	PT				0.089**	0.014	4.664**
Parking_Cost	Car				0.907**	0.088	5.999**
Sample size		1163			1163		
Log likelihood at constant		− 11285.82			− 11285.82		
Log likelihood at convergence		− 7160.97			− 5234.94		
Adjusted McFadden's pseudo R^2		0.365			0.536		

Note: * significant at the 5% level, ** significant at the 1% level.

represent the strength of the effect of each independent variable on the outcome variable, and the variable with the largest standardized coefficient has the strongest influence. However, the utility of choosing a travel mode is a latent variable and thus unobservable, so it is not obvious how to standardize a latent variable in order to estimate the standardized Beta coefficients. If one is only interested in the rank order of the magnitude of the effects of the independent variables on the utility, however, standardizing the independent variables (i.e., *X*-standardization) is sufficient (Menard, 2004).

The outputs for the MNL and mixed logit are presented in Table 7. The coefficients presented here are *X*-standardized Beta coefficients. For both models, the results show that the most important variable in predicting the 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.

The adjusted McFadden's pseudo R^2 for MNL and mixed logit are 0.365 and 0.536, respectively, which indicates satisfactory model fit. All coefficient estimates are consistent with theoretical predictions. The level-of-service variables are all statistically significant and carry an intuitive negative sign. For both logit models, individual socio-demographic characteristics are strongly associated with mode choice. 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. Furthermore, the residential preferences and current travel mode choices of individuals are associated with their travel mode choices. Individuals who value walking, biking, and transit access when choosing where to live are more likely to use these modes. The model shows that travelers tend to stick to their current mode even when a new travel option is offered, but this inertia effect is weaker among transit riders. Furthermore, the random-parameter standard deviations in the mixed logit model are statistically significant, suggesting the existence of preference heterogeneity.

We now interpret the machine-learning models with variable importance measures and partial dependence plots. Like the *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. We use the Gini index to measure variable importance for RF. For NN, the variable importance is computed using the method proposed by Gevrey et al., 2003, which applies combinations of the absolute values of the weights. Unlike *X*-standardized Beta coefficients that have a positive or negative sign, however, variable importance measures provide no information on the direction of association between the independent variables and the outcome variable. Partial dependence plots are instrumental in this regard.

Table 8
Ranking of variable importance for RF, NN, MNL, and mixed logit.

Variable	RF	NN	MNL	Mixed logit
TT_Walk	1	16	2	2
TT_Drive	2	14	4	5
TT_Bike	3	13	1	1
TT_PT	4	11	3	3
Current_Mode_Bike	5	2	5	10
PT_Access	6	8	13	16
Bike_Walkability	7	6	16	13
Income	8	10	14	15
CarPerCap	9	7	11	14
Current_Mode_Walk	10	1	6	12
Rideshare	11	9	9	9
Transfer	12	5	8	8
Wait_Time	13	15	10	11
Female	14	3	15	17
Parking_Cost	15	12	12	4
Current_Mode_Car	16	4	7	6
Current_Mode_PT	/	/	17	7

Table 8 shows the ranking of variable importance for RF, NN, MNL, and mixed logit. Note that Current_Mode_PT is not included in the RF and NN models due to perfect multicollinearity with the other three current-mode-choice variables.[4] The ranking of the input features for RF is generally consistent with that of the two logit models, but NN has different variable importance results compared to RF, MNL, and mixed logit. For the RF model and the two logit models, the travel times of walking, driving, biking, and transit have very high influence on their stated mode choice; on the other hand, some differences do exist: For example, PT_Access, Bike_Walkability, Income, and CarPerCap are more important for RF compared to MNL and mixed logit.

Fig. 2 presents the partial dependence plots, which show how the probability of choosing *PT* changes as the value of the selected variable changes for RF and MNL. Like the Beta coefficients estimated from the MNL model, the shape of the curves sheds light on the direction and magnitude of the changes. However, the Beta coefficients in logit models affect the utility of mode k (see Eqs. (1) and (11)) rather than the probability of choosing mode k (see Eqs. (2) and (13)). Accordingly, we translate utility estimates into probability estimates for the MNL model to compare it with RF directly.

As shown in Fig. 2(a), RF, MNL, and mixed logit all show a similar decreasing trend for TT_PT, while NN presents a different pattern. As shown in Fig. 2(b)–(d), for Wait_Time, Rideshare, and Transfer, RF and NN differ from the two logit models. While MNL and mixed logit show a nearly linear relationship between these features and the probability of choosing *PT*, the two machine-learning models reveal some nonlinear relationships. For example, the following observations can be highlighted on the RF model: 1) For TT_PT, RF has relative flat tails before 10 min and after 25 min, 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 min; and 3) the choice probability of *PT* decreases more significantly from 0 to 1 rideshare compared to from 1 to 2 rideshares. Based on these observations, we specified piece-wise utility functions (i.e., specifying different coefficients for a variable in different data intervals) for the logit models (MNL and mixed logit). While not showing the model outputs here, we found that the model fit improved and that the coefficient estimates largely agreed with the nonlinearies revealed by the RF model.

In sum, partial dependence plots readily reveal the nonlinear association between level-of-service attributes and travel-mode choice. In contrast to the time-consuming hand-curating procedure required in logit models (often by introducing interactions terms) to reveal nonlinear relationships, machine-learning algorithms capture these nonlinearities automatically and thus can generate richer behavioral insights much more efficiently. Machine-learning models can thus 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.

6.2.2. Arc elasticity and marginal effects

We calculated marginal effects and arc elasticities for the level-of-service variables associated with the proposed mobility-on-demand transit system, including TT_PT, Wait_Time, Rideshare, and Transfer. Marginal effects (elasticities) measure the changes of the choice probability of an alternative in response to one unit (percent) change in an independent variable. Table 9 presents the results of these behavioral outputs.

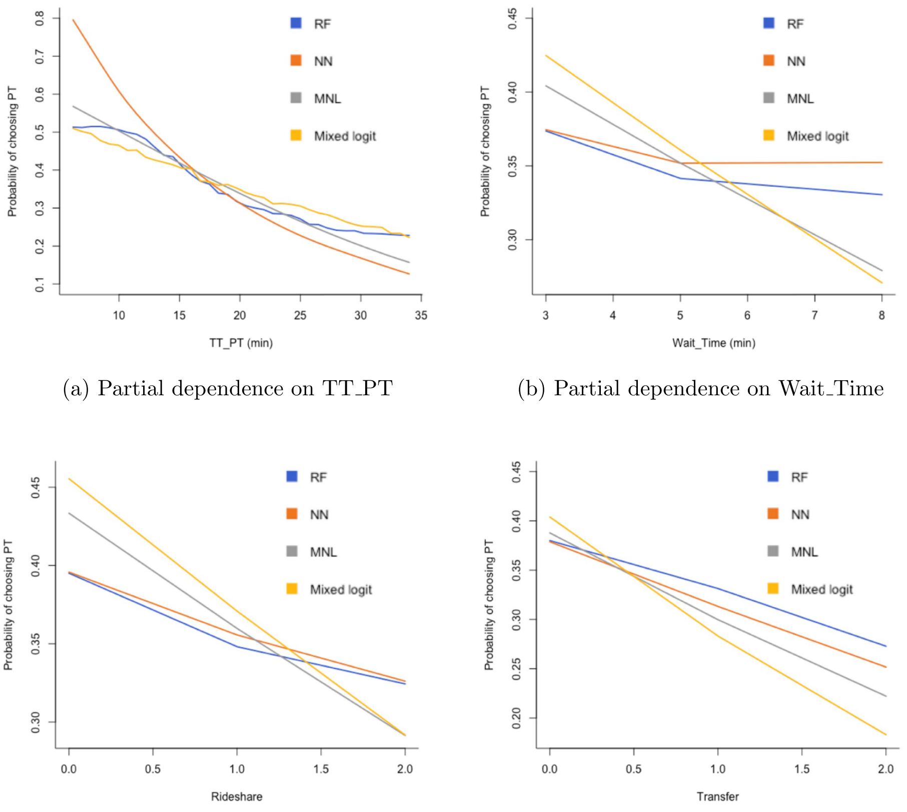
Note that we computed these behavioral outputs for the RF using both a standard approach and a modified approach. Computing arc elasticity and marginal effects for RF by using Eqs. (9) and (10) on the entire dataset (i.e., the standard approach) produces unrealistic results: 1) For Wait_Time, by adding 1 min, the marginal effect for RF is extremely small (− 0.01%); and 2) If converting the marginal effects of Transfer and Rideshare into relative value-of-time measures (by dividing their marginal effect estimates with that of TT_PT), we find that the penalty effect of one Transfer is equivalent to 2.8 min of TT_PT and the penalty of one Rideshare is equivalent to 1.3 min of TT_PT, which seem to be too small. This is possibly due to the nature of RF: RF consists of hundreds of decision trees that apply decision rules based on discrete values. Decision trees, which lack smoothness in prediction, become insensitive to values between the adjacent bounds learned from the original data, and are unable to make predictions outside the observed range (i.e., they cannot extrapolate).

To address these limitations, we propose a modified approach to compute these behavioral outputs for the RF model. First, We use $\Delta = 2$ min rather than $\Delta = 1$ min to compute the marginal effects of Wait_Time for RF. This helps solve the problem of the RF model being insensitive to very small changes in feature values (for a tree-based model, small changes of a feature value often cannot pass the thresholds for the splitting of trees and so the RF model may not react to such changes). Note that $\Delta = 2$ min is arbitrarily decided here, and future work should investigate how to decide an appropriate criteria for different features. In addition, we take a subset of the entire population to compute these two metrics by removing the "out-of-bound" instances, in order to counter the extrapolation issue of the tree-based models. The behavioral outputs obtained from this modified approach appear to be much more behaviorally sound compared to the results obtained from the standard approach. The marginal effects of Wait_Time, Transfer, and Rideshare increase significantly, whereas the marginal effect of TT_PT does not change since this variable is less vulnerable to the limitations of the RF model discussed above.

Table 9 shows that the arc-elasticity and marginal-effect estimates are all negative, indicating that when the level-of-service of transit gets worse, the travelers' preferences for transit will decrease. However, we find significant differences in the behavioral outputs across the four models. For Wait_Time, Transfer, and Rideshare, the marginal-effect estimates of logit models are larger than those of NN and RF. For TT_PT, the marginal effects and arc elasticity estimates for MNL, NN, and RF are similar in magnitude, whereas the estimates for the mixed logit model are much smaller.

Without the ground truth, it is difficult to judge the validity of these results. However, converting these marginal-effect estimates into relative value-of-time measures can help assess their relative soundness. We use TT_PT as the baseline to obtain relative value-of-time measures

[4] Removing Current_Mode_PT from estimation would not lead to any information loss in the RF and NN models, as these models can discern whether an individual has chosen public transit once their choices for the other three options are given. However, since logit models have a separate utility function for each alternative, all four current-mode-choice variables should be specified in the utility functions; removing Current_Mode_PT from the utility function of public transit would lead to information loss and thus is problematic.



(a) Partial dependence on TT_PT

(b) Partial dependence on Wait_Time

(c) Partial dependence on Rideshare

(d) Partial dependence on Transfer

Fig. 2. Partial dependence plots of variables for choosing *PT* as the travel mode.

Table 9
Marginal effects and arc elasticities.

Variable		Δ	MNL	Mixed logit	NN	RF	
						Standard Approach	Modified Approach
Wait_Time	Marginal effect	1 or 2 min	− 2.93%	− 2.96%	− 0.58%	− 0.01%	− 1.16% [a]
Transfer	Marginal effect	1 unit	− 10.69%	− 11.66%	− 6.27%	− 4.60%	− 5.10%
Rideshare	Marginal effect	1 unit	− 8.13%	− 7.74%	− 2.54%	− 2.08%	− 3.41%
TT_PT	Marginal effect	1 min	− 1.94%	− 0.87%	− 2.45%	− 1.63%	− 1.63%
	Arc elasticity	10%	− 0.89	− 0.49	− 1.28	− 1.07	− 1.08

Note: 2 min is used for the modified case of RF, while 1 min is used for other cases.

for other variables and obtain the following results. The penalty of a transfer is approximately equal to 5.5 min (MNL), 13.4 min (mixed logit), 2.6 min (NN), and 3.1 min (RF) of TT_PT. The penalty of a rideshare stop is equivalent to 4.2 min (MNL), 8.9 min (mixed logit), 1.0 min (NN), and 2.1 min (RF) of transit travel time. The value of one min wait time is equal to 1.5 min (MNL), 3.4 min (mixed logit), 0.2 min (NN), 0.7 min (RF) of transit travel time. The results of RF seem more realistic than those of NN since the latter implies extreme small penalty effects of a transfer, a ridesharing stop, and waiting time.

The existing literature generally finds that the penalty effects of a transfer is larger than 5 min of in-vehicle travel time (e.g., Iseki and Taylor, 2009; Garcia-Martinez et al., 2018), and the value of wait time is slightly larger than that of in-vehicle travel time (Abrantes and Wardmanbrantes, 2011). The behavioral outputs of the logit models appear to be more reasonable than those of RF, the better-performing machine-learning model. One possible reason is that without behavioral constraints put into the model specifications, the machine-learning models are completely data-driven and possibly learn some level of noise contained in the data, undermining its capability to generate reasonable behavioral outputs. If this is true, it creates a paradox for choice modeling: While machine-learning models have better predictive capability, they are inferior to traditional logit models when applied to derive generic behavioral insights.

There certainly can be another possibility: the results of RF are closer to ground truth. This possibility can be supported by two particularities in this study. First, TT_PT here consists of both in-vehicle travel time and out-of-vehicle travel time, and so using TT_PT—rather than the commonly used in-vehicle travel time—to construct the value-of-time measures naturally lead to smaller estimates than the literature. The other reason is that the proposed mobility-on-demand system here is described to respondents as an app-based system that not only provides accurate real-time information but also synchronizes the schedule of on-demand shuttles with that of the fixed-route buses. Therefore, passengers may perceive that the transfer will be much more convenient and that they can wait at more comfortable locations than traditional bus stops, which lead to smaller penalties for Transfer and

Wait_Time. It is beyond the scope of this paper to explore this issue, and we call for future machine-learning-based mode choice studies to switch from a mere focus on prediction to an emphasis on validating (and perhaps fixing) the behavioral outputs.

7. Discussion and conclusion

The increasing 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. On the other hand, many machine-learning models are built for prediction purposes but are often considered as difficult to interpret and are rarely used to extract behavioral insights from the model outputs. This study is one of the first attempts to validate the credibility of machine learning in producing sound behavioral findings.

Overall, 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 machine-learning and logit models side by side using cross validation to examine their capabilities for prediction and behavioral analysis. The results show that the best-performing machine-learning model, i.e., RF, significantly outperforms the logit models in prediction both at the individual level and the aggregate level, probably due to its capability to capture variable interactions and model nonlinear relationships (Lhéritier et al., 2018; Cheng et al., 2019). In fact, most machine-learning models outperform the logit models. Somewhat surprisingly, the mixed logit model underperforms the MNL and most machine-learning models in terms of the out-of-sample predictive accuracy, which may result from overfitting the training set.

Moreover, building on previous work that applied variable importance to interpret machine-learning models (Cheng et al., 2019; Hagenauer and Helbich, 2017; Wang and Ross, 2018; Lhéritier et al., 2018; Golshani et al., 2018), we further examined partial dependence plots and computed several behavioral outputs (marginal effect and arc elasticities) from the machine-learning models. Comparing these results with the behavioral findings of the logit models generates valuable insights. We find that machine-learning and logit models largely agree on variable importance and the direction of influence that each independent variable has on the choice outcome. In addition, the partial dependence plots show that machine-learning models can readily capture nonlinear associations. Therefore, machine-learning models may serve as an exploratory tool used to identify nonlinear effects, which can inform logit-model analysts the specification of nonlinearities in the utility functions of different alternatives. This procedure is much more efficient than the hand-curating procedure typically done with statistical models.

However, there are significant differences in the behavioral outputs (marginal effects and arc elasticities) generated from the machine-learning and logit models. In particular, we show that, for tree-based models such as RF, computing marginal effects and elasticity using a standard approach may result in unreasonable behavioral results. To obtain more realistic behavioral outputs, one need to incorporate some modifications to the computation procedure in order to overcome the limitations of tree-based models. However, even with these modifications, the RF's behavioral results are somewhat inconsistent with the existing literature. Therefore, there appears to be a tradeoff between predictive accuracy and behavioral soundness when choosing between machine learning and logit models in mode-choice modeling. More research is needed to identify the nature of the problem. Future research should consider developing approaches to refine machine-learning models to fix any illogical behavioral results; for example, imposing behavioral constraints to the risk functions of machine-learning models may lead to more realistic behavioral outputs. In addition, one may examine which machine-learning models are more suitable than others for behavioral analysis. Future studies may also consider evaluating different travel mode choice datasets and examining more input features.

This study sheds light on many new research directions in applying machine learning for travel-behavior research. The first topic is concerned with 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. We have not addressed this important topic in this paper. The second topic is related to the reporting bias associated with the SP data. The SP data are generally considered as containing reporting bias due to the hypothetical nature of state choice experiments. It is unclear how how the SP reporting bias may affect the performance of logit models and machine-learning models, respectively. Moreover, logit models using joint RP and SP data have been proposed to correct for the SP reporting bias (Train, 2009), but theoretically sound machine-learning models that can facilitate such a joint estimation process are yet to be developed.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xilei Zhao: Conceptualization, Methodology, Formal analysis, Writing - original draft, Supervision, Funding acquisition. **Xiang Yan:** Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Writing - original draft. **Alan Yu:** Software. **Pascal Van Hentenryck:** Supervision, Writing - review & editing, Funding acquisition.

Acknowledgements

This research was partly funded by the Michigan Institute of Data Science (MIDAS), by the U.S. Department of Energy (Grant 7F-30154), and by the U.S. Department of Transportation through the Southeastern Transportation Research, Innovation, Development and Education (STRIDE) Region 4 University Transportation Center (Grant P0147836 - Project B3).

References

Abrantes, P.A., Wardman, M.R., 2011. Meta-analysis of UK values of travel time: an update. Transp. Res. Part A: Policy Practice 45 (1), 1–17.

Athey, S., 2017. Beyond prediction: using big data for policy problems. Science 355 (6324), 483–485.

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. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 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. J. 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. J. Mach. Learn. Res. 11 (Jul), 2079–2107.

Chen, X.M., Zahiri, M., Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: an ensemble learning approach. Transp. Res. Part C: Emerg. Technol. 76, 51–70.

Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.

Cherchi, E., Cirillo, C., 2010. Validation and forecasts in models estimated from multiday travel survey. Transp. Res. Rec.: J. Transp. Res. Board 2175, 57–64.

Christopher, M.B., 2016. Pattern Recognition and Machine Learning. Springer-Verlag, New York.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Farrar, D.E., Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. Rev. Econ. Stat. 92–107.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann.

Stat. 1189–1232.

Garcia-Martinez, A., Cascajo, R., Jara-Diaz, S.R., Chowdhury, S., Monzon, A., 2018. Transfer penalties in multimodal public transport networks. Transp. Res. Part A: Policy Practice 114, 52–66.

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model. 160 (3), 249–264.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graphical Stat. 24 (1), 44–65.

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 Behav. Soc. 10, 21–32.

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst. Appl. 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.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844.

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13 (2), 415–425.

Iseki, H., Taylor, B.D., 2009. Not all transfers are created equal: towards a framework relating transfer connectivity to travel behaviour. Transp. Rev. 29 (6), 777–800.

Jenkins, K., 2018. New app reinvents University bus system to be more like Uber. The Michigan Daily.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. Transp. Res. Part C: Emerg. Technol. 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. Environ. Resource Econ. 1–17.

Kuhn, M., et al., 2008. Building predictive models in r using the caret package. J. Stat. Software 28 (5), 1–26.

Last, M., Maimon, O., Minkov, E., 2002. Improving stability of decision trees. Int. J. Pattern Recogn. Artif. Intell. 16 (02), 145–159.

Lhéritier, A., Bocamazo, M., Delahaye, T., Acuna-Agost, R., 2018. Airline itinerary choice modeling using machine learning. J. Choice Model.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22. URL:http://CRAN.R-project.org/doc/Rnews/.

Lindner, A., Pitombo, C.S., Cunha, A.L., 2017. Estimating motorized travel mode choice using classifiers: An application for high-dimensional multicollinear data. Travel Behav. Soc. 6, 100–109.

Mahéo, A., Kilby, P., Van Hentenryck, P., 2017. Benders decomposition for the design of a hub and shuttle public transit system. Transp. Sci.

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. J. Appl. Econometrics 15 (5), 447–470.

Menard, S., 2004. Six approaches to calculating standardized logistic regression coefficients. Am. Stat. 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.

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. J. Econ. Perspectives 31 (2), 87–106.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Interpretable machine learning: definitions, methods, and applications. arXiv:1901.04592.

Omrani, H., 2015. Predicting travel mode of individuals by machine learning. Transp. Res. Proc. 10, 840–849.

Omrani, H., Charif, O., Gerber, P., Awasthi, A., Trigano, P., 2013. Prediction of individual travel mode with evidential neural network model. Transp. Res. Rec.: J. Transp. Res. Board 2399 (1), 1–8.

Quinlan, J.R., 2014. C4. 5: Programs for Machine Learning. Elsevier.

Rasouli, S., Timmermans, H.J., 2014. Using ensembles of decision trees to predict transport mode choice decisions: effects on predictive success and uncertainty estimates. Eur. J. Transp. Infrastruct. Res. 14 (4).

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.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. fourth ed. 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. Transp. Res. Rec.: J. Transp. Res. Board.

Wong, M., Farooq, B., Bilodeau, G.-A., 2018. Discriminative conditional restricted boltzmann machine for discrete choice and latent variable modelling. J. Choice Model. 29, 152–168.

Wu, T.-F., Lin, C.-J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res. 5, 975–1005.

Xie, C., Lu, J., Parkany, E., 2003. Work travel mode choice modeling with data mining: decision trees and neural networks. Transp. Res. Rec.: J. Transp. Res. Board 1854, 50–61.

Yan, X., Levine, J., Zhao, X., 2019. Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. Transp. Res. Part C: Emerg. Technol. 105, 683–696.

Zhang, Y., Xie, Y., 2008. Travel mode choice modeling with support vector machines. Transp. Res. Rec.: J. Transp. Res. Board 2076, 141–150.

Zhao, Q., Hastie, T., 2019. Causal interpretations of black-box models. J. Business Econ. Stat. 1–10.