Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?

Miguel Paredes
CSAIL & DUSP
Massachusetts Institute of Technology
Cambridge, Massachuetts 02139
Email: mparedes@mit.edu

Erik Hemberg
CSAIL
Massachusetts Institute of Technology
Cambridge, Massachuetts 02139
Email:hembergerik@csail.mit.edu

Una-May O'Reilly
CSAIL
Massachusetts Institute of Technology
Cambridge, Massachuetts 02139
Email: unamay@csail.mit.edu

Chris Zegras
DUSP
Massachusetts Institute of Technology
Cambridge, Massachuetts 02139
Email: czegras@mit.edu

***Abstract*—Discrete choice models are widely used to explain transportation behaviors, including a household's decision to own a car. They show how some distinct choice of human behavior or preference influences a decision. They are also used to project future demand estimates to support policy exploration. This latter use for prediction is indirectly aligned with and conditional to the model's estimation which aims to fit the observed data. In contrast, machine learning models are derived to maximize prediction accuracy through mechanisms such as out-of-sample validation, non-linear structure, and automated covariate selection, albeit at the expense of interpretability and sound behavioral theory. We investigate how machine learning models can outperform discrete choice models for prediction of car ownership using transportation household survey data from Singapore. We compare our household car ownership model (multinomial logit model) against various machine learning models (e.g. Random Forest, Support Vector Machines) by using 2008 data to derive, i.e. estimate models that we then use to predict 2012 ownership. The machine learning models are inferior to the discrete choice model when using discrete choice features. However, after engineering features more appropriate for machine learning they are superior. These results highlight both the cost of applying machine learning models in econometric contexts and an opportunity for improved prediction and better urban policy making through machine learning models with appropriate features.**

I. Introduction

Urban development is greatly influenced by the travel behavior of a city's residents [3]. For this reason, understanding and predicting the demand for travel and the demand for car ownership across the population is of key interest to city officials around the world. Transportation and urban planners use transportation demand forecasts to inform their policies and investments, thus placing high value on data and methods that provide demand projections.

Discrete choice models are a set of econometric tools that are widely used to explain transportation behaviors [7], including a household's decision to own a car [8], [2]. The typical goal of this type of modeling is to obtain unbiased estimators that provide behavioral and economic insights. An accurate discrete choice model fits the observed data with minimum error, typically assuming a linear model structure. Discrete choice models are selected because they provide interpretability, i.e. they elucidate how a categorical explanatory variable describing some aspect of human behavior or preference influences decision making. Once a discrete choice model is estimated, it can be used for prediction and to explore possible impacts of policy. However, while a discrete choice model can be used for these purposes, it is not explicitly estimated for them. In contrast, machine learning models are generally derived to explicitly maximize predictive accuracy, that is their performance on unseen data, rather than observed data which is used for "training". This is accomplished through rounds of modeling with different folds of training data and preliminary validation on held out folds. A model's performance is reported on data not in the training set. Machine learning models, e.g. decision trees or support vector machines (SVMs) are frequently non-linear, have different bias-variance tradeoffs and have regularization mechanisms that control model complexity (and indirectly generalization). Each uses a different technique for performance objective optimization and sometimes covariates are simultaneous selected with model derivation. Consequently, machine learning models generally outperform traditional econometric models on prediction [6]. Many ML modeling aspects that contribute to this advantage come however at the expense of interpretability. Unlike discrete choice models, less emphasis is on using features that are easily mapped to human behavior and well understood causal relationships and correlations.

Our research question is whether machine learning models can outperform discrete choice models on the prediction of household car ownership as 0, 1 or 2 or more cars. We hypothesize that because machine learning models are derived to reduce the prediction error instead of the estimation error, machine learning models should outperform discrete choice models on this class prediction task (at the expense of model and parameter interpretability, desirable parameter properties,

and behavioral theory soundness). In more detail, our first question is whether machine learning models will be superior when they use exactly the same variables we prepare for our discrete choice model. Because discrete choice variables are indicators of human preferences or characteristics that need to be represented in ways that have economic intepretability and significance (thus many times discretizing variable ranges into dummy variables to explore marginal effects), we then ask whether using the original variables and variables not consistent with a sound behavioral thoery better exploits machine learning for prediction.

Singapore is one of the leading countries capitalizing on urban data and models to support decision making. The Singapore-MIT Alliance for Research and Technology (SMART) has been eveloping SimMobility, a state-of-the-art simulator of the transportation and land-use systems based on behavioral models. The objective of SimMobility is to predict the impact of different mobility interventions on travel demand and activities, both for passengers and freight, and on transportation networks and land-use [1]. Within SimMobility, discrete choice models are employed to represent the decision of urban agents, such as households choosing whether to have a car or not.

Using transportation survey data from Singapore [4], aggregated to the household unit, we compare machine learning models to discrete choice models for prediction of household car ownership. We use 2008 data to train our machine learning models and to estimate our discrete choice model and use these models to predict 2012 car ownership in households. We compare our household car ownership model (multinomial logit model (MNL)) against our 5 machine learning models (Random Forest, Support Vector Machines, Decision trees, Extreme Gradient Boosting, and an Ensemble of methods).

We proceed as follows; Section II describes the dataset used and provides summary statistics. Section III describes and compares discrete choice and machine learning methods. Section IV describes data processing and preparation for both the discrete choice model dataset and the machine learning model dataset, and lays out our experimental setup for both. Section V presents the results and discussion. Finally, we present conclusions and future work in Section VI.

II. Data

Singapore's Land Transport Authority (LTA) conducts the Household Interview Transportation Survey (HITS) every four to five years to better understand residents traveling behaviors. Approximately one percent of all the households in Singapore are surveyed with travel and sociodemographic questions. HITS is a key instrument for transportation and urban planners to better design their policies in response to Singapore's transportation needs.

We use Singapore's 2008 and 2012 HITS surveys, extracting the following variables for each respondent:

1) Type of residential property (private, landed, HDB, other)
2) Ethnicity (Chinese, Malay, Indian, other)
3) Motorcycle ownership
4) Geolocation of household
5) Employment Status (Full time worker, Part time worker, Self-employed, Student, etc.)
6) Type of Appointment (CEO, Executive, Professional, Assistant, Clerk, etc.)
7) Number of Children
8) Income Level
9) Taxi ownership

The HITS 2008 survey had 88,600 individual responses and the HITS 2012 had 85,880 individual responses. Both databases where aggregated to the household level, producing 8,901 households in 2008 and 6,310 households in 2012.

A. HITS 2012 Insights

Some interesting insights from a basic analysis of the 2012 HITS data are that, in general, landed households (HHs) have 2+ cars, private HHs have 1 car, and HHs in the "other" housing category usually do not own a car. Most 2+ car HHs are Chinese, further away from MRT stations, and more than 50% of their members are white collar professionals. In general, as HH car ownership increases, so does the HH size, the proportion of self-employed members, and the probability that there are white collar members. There is no association between HHs with no cars and HH size.

III. Methods

Our modeling proceeds in the same general manner regardless of whether we estimate a discrete choice model or derive machine learning model. First, the observational "raw" survey data is transformed into examples described by features, a.k.a. variables or predictors in a step we call feature engineering. After a model has been selected, some or all of the transformed examples, each paired with its outcome or response variable's value (a.k.a. label), are used to construct the model. Modeling requires parameter selection including specifying an objective. For the case of our discrete cohice model, we employ variables that are commonly used in this type of car ownership estimation. All models are constructed using 2008 examples. After derivation, we evaluate all models on the task of predicting which of three ownership classes, 0 cars, 1 car or 2+ cars, a HH in the 2012 dataset will fall into. We report the most common and simple prediction evaluation metric: accuracy, which is the ratio of correct predictions made to all predictions. We also examine confusion matrices which present predictions on the x-axis (in the columns) and observed outcomes on the y-axis (in the rows). The diagonal cells are the correctly classified cases, and the off-diagonal cells are the misclassified cases.

A. Discrete Choice Models

Discrete choice models describe decision makers' choices among alternatives and examine situations in which the potential outcomes are discrete. Decision makers can be people, households, firms, or any other decision-making unit, and the alternatives might represent competing products, courses of action, or any other options or items over which choices must

be made. Discrete choice analysis consists of two interrelated tasks: specification of the behavioral model and estimation of the parameters of that model[9].

We estimate a multinomial logit model which is a type of a generalized linear model (GLM) that generalizes logistic regression to multi-class problems. It assumes that each original variable has been transformed to be case specific according to threshholds selected from the variable range. This engineering of variables (features) expands the dimensionality of the model while compressing the range of each variable. The output of MNL is a class likelihood and it can indicate whether a variable increases or decreases the probability of a given class by a specific percentage. It requires an assumption of independence of irrelevant alternatives, which is consistent with our problem, i.e. the odds of preferring one car ownership level over another do not depend on the presence or absence of other irrelevant alternatives. The coefficients in the MNL model are estimated in terms of maximum likelihood. We predict the outcome variable from the model's probabilities by selecting the outcome with the highest probability.

B. Machine Learning Models

Unlike many standard econometric techniques, machine learning (ML) methods were developed with the intention of maximizing prediction accuracy[5]. From a statistical perspective machine leanrning can be viewed as an automated exploratory analysis of large data sets. Feature engineering in machine learning does not demand the type of transformations used in discrete choice modeling. The models used in ML are typically non-linear, they use cross-validation and regularization (removing low impact model coefficients) with the goal of generalization to unseen observations for the task of prediction. Learning objectives vary and include e.g. Lk-norms. Optimization methods vary and data can be transformed into lower dimensional embeddings or higher dimensional embeddings with kernelization, e.g. in Support Vector Machines. For a more complete treatment see [5].

IV. Experiments

Using transportation household survey data from Singapore we examine whether machine learning models are more accurate at prediction that discrete choice models for prediction of household car ownership. We will look into the implications of the data pre-processing approach for discrete choice models and the impact on the performance of our machine learning models.

A. Data Preparation

For both 2008 and 2012 datasets, we pre-proccessed the original HITS surveys by aggregating the personal information to the household level. Then we created two sets of features: E_{DC} meeting discrete choice model assumptions and E_{ML} suited to machine learning.

1) E_{DC}*:* For E_{DC} we created dummy variables for some of the integer variables (HH size, number of certain types of professionals in the HH, number of students, number of children) or to indicate that the HH had at least a member with certain characteristics. For example, a binary variable `CEO` was created for whether one of the HH's members had a job description of CEO, or the present of two students in the household was represented through a binary variable (`STUDENT2`). Using geographical information system (GIS) tools we also created two distance variables (distance to the nearest subway station - `distMRT` - and distance to the central business district - `distCBD`). Our E_{DC} has a total of 37 variables. The engineered variables in E_{DC} are: *1)* Distance to the closest MRT station (`distMRT500, distMRT1000`) *2)* Distance to the Central Business District (CBD - `distCBD`) *3)* Number of fulltime workers in the HH (`FULLWORKER1, FULLWORKER2, FULLWORKER3p`) *4)* Number of CEOs in the HH (`CEO`) *5)* Number of white collar professionals in the HH (`WHITECOLLAR1, WHITECOLLAR2p`) *6)* Number of students in the HH (`STUDENT1, STUDENT2, STUDENT3p`) *7)* Number of children in the HH (`KIDS1, KIDS2p`) *8)* If the household has a member above 60 years (`ABOVE60`) *9)* Income dummy variables indicating HH income (`INC0, INC12` - we aggregated income1 and income2 categories because there were few of these, INC3, INC4, INC5, INC6) *10)* Housing type (HDB, `Private, Landed, Other Housing`, we aggregate the 3 types of HDB residences into one)

Missing values in the dataset were dealt with through list deletion, i.e. observations with an NA in one of the variables were elimnated from the dataset. The variable Income had the highest incidence of NA values (14.4% for 2008, and 31.9% for 2012). Future work will explore multiple imputation strategies, which were conducted for our discrete choice estimation model with no significant change in results.

2) E_{ML}*:* To obtain E_{ML}, we go back to the original HITS datasets for 2008 and 2012 and aggregate them at the household level but do not add any dummy variables nor do we transform some of the variables from integer values to discretized binary variables representing each one of the units. Additionally, unlike discrete choice models where variable selection follows a sound behavioral theory and optimizes for explanatory power (by incorporating, deleting, or transforming variables in ways that maximize a criteria such as the adjusted rho-squared), machine learning models automate the feature selection process and can incorporate as many variables as needed (regardless of whether these variables are consistent with a sound behavioral theory). For this reason, we have a total of 38 features in E_{ML} out of which 13 features are not included in the discrete choice model. The E_{ML} include:

1) Distance to the closest MRT station (`distMRT`, we leave as an integer value instead of discretazing into 3 dummy variables as we did for E_{DC}) *2)* Distance to the Central Business District (`distCBD`) *3)* Number of fulltime workers in the HH (`FULLWORKER`, we leave as an integer value) *4)* Number of CEOs in the HH (`CEO`) *5)* Number of white

collar professionals in the HH (WHITECOLLAR, we leave as an integer value) *6)* Number of students in the HH (STUDENT, we leave as an integer value) *7)* Number of children in the HH (KIDS, we leave as an integer value) *8)* Income (INCOME, we leave as an integer value) *9)* Housing type (HDB123, HDB4, HDB5EF, Private, Landed, Other Housing, we do not aggregate HDB types) *10)* Number of people within an age category (BELOW4, AGE4TO9, AGE10TO14, AGE15TO19, ABOVE60, we do not aggregate all ages below 14 to obtain the number of kids, and we leave all categories with the total number of people in them as is) *11)* Number of people in a particular work condition (FULLTIME, PARTTIME, SELFEMPLOYED, RETIRED, UNEMPLOYED, STUDENT) *12)* Number of people in the HH that have a CEOs or Whitecollar job position (CEO, WHITECOLLAR) *13)* Number of people with a car or motorcycle licence (CAR_LICENCE, MC_LICENSE)

B. Setup

In the first experiment, we estimate a standard discrete choice model, MNL using E_{DC}. Likewise, we train four different ML models using E_{DC}. We train these models using cross-validation. We then use the 2012 HITS survey data as a test set to predict HH car ownership levels with each model. We compare the misclassification rates of all models.

In the second experiment we derive a second set of machine learning models with E_{ML}. Our algorithms use feature selection techniques and cross-validation. We then use 2012 HITS data again as a test set and predict the HH car ownership levels contrasting the new results to the MNL model.

We run all our experiments on Windows 10 using RStudio version 0.99.489, and R packages mlogit[1], randomForest[2], e1071[3], rpart[4], xgboost[5]. For all machine learning models we run the mentioned packages with their default parameters unless specified otherwise.

V. Results

A. Prediction from E_{DC}

First, we report the results of our discrete choice model that uses E_{DC}. We run a basic Multinomial Logit model with the following specification:

```
CAR_CHOICE = PRIVATE + LANDED + OTHER_HOUSING +
    MALAY + INDIAN + OTHER_RACE + HAS_MC + FULLWORKER1 +
    FULLWORKER2 +  FULLWORKER3p + WHITECOLLAR1 + WHITECOLLAR2p +
    STUDENT1 + STUDENT2 + STUDENT3p + KIDS1 + KIDS2p + INC12 +
    INC3 + INC4 + INC5 + INC6 + ABOVE60 + CEO + SELFEMPLOYED +
    HHSIZE3 + HHSIZE4 + HHSIZE5 + HHSIZE6plus + distMRT500 +
   distMRT1000 + distCBD + TAXI
```

The distribution of car classes is not balanced: 0 cars : 0.635, 1 car : 0.323, 2+ cars : 0.0413. The coefficients from MNL are shown in Table I.

The Log-Likelihood is -4999.9, the McFadden R^2 : 0.28429 and Likelihood ratio test: $\chi^2 = 3972.1$(p.value $\leq 2.22e-16$). These car ownership demand model results are consistent with other car ownership discrete choice models [2].

According to our discrete choice model, we see that higher income levels, increasing HH size, having a self-employed profession in the HH, having 2 or more children, having a HH member who is a CEO, living in a private or landed property, or being of Chinese descent increases the odds of a HH owning one or more cars. Conversely, living in an HDB, being of Malay, Indian or other race descent, having a motorcycle at home, having full time workers at home, having a HH member that is over 60 years old, living further away from an MRT station, and owning a Taxi are all associated with a lower odds of owning one or more cars. HHs with students show both associations with increased and decreased odds of having one or more cars.

We next use E_{DC} to train our ML models, all results are

TABLE I
Coefficients of MNL on E_{DC}. Signif. codes are 0:***, 0.001:**, 0.01:*, 0.05:., 0.1: , 1:.

Class	Feature	Estimate	Std. Error	t-value	$Pr(>\|t\|)$	Sig
1	(intercept)	-1.7690e+00	2.6041e-01	-6.7932	1.097e-11	***
2	(intercept)	-4.8123e+00	7.7258e-01	-6.2289	4.698e-10	***
1	PRIVATE	1.0831e+00	9.8878e-02	10.9541	< 2.2e-16	***
2	PRIVATE	1.5344e+00	1.8614e-01	8.2437	2.220e-16	***
1	LANDED	1.3999e+00	1.3571e-01	10.3149	< 2.2e-16	***
2	LANDED	3.0657e+00	1.9824e-01	15.4648	< 2.2e-16	***
1	OTHER_HOUSING	4.9332e-01	2.5438e-01	1.9393	0.0524642	.
2	OTHER_HOUSING	-1.6834e+01	4.0235e+03	-0.0042	0.9966617	
1	MALAY	-3.9385e-01	9.0294e-02	-4.3618	1.290e-05	***
2	MALAY	-3.1243e-01	2.7174e-01	-1.1497	0.2502518	
1	INDIAN	-1.2626e+00	9.3272e-02	-13.5362	< 2.2e-16	***
2	INDIAN	-2.3186e+00	3.1599e-01	-7.3378	2.172e-13	***
1	OTHER_RACE	-1.7221e+00	1.7274e-01	-9.9693	< 2.2e-16	***
2	OTHER_RACE	-2.2659e+00	3.9488e-01	-5.7383	9.564e-09	***
1	HAS_MC	-5.3935e-01	1.1661e-01	-4.6254	3.739e-06	***
2	HAS_MC	-1.9100e+00	5.0154e-01	-3.8082	0.0001400	***
1	FULLWORKER1	-4.0121e-01	1.2259e-01	-3.2728	0.0010648	**
2	FULLWORKER1	-6.7915e-01	2.5342e-01	-2.6800	0.0073632	**
1	FULLWORKER2	-7.0750e-01	1.4436e-01	-4.9010	9.535e-07	***
2	FULLWORKER2	-1.2388e+00	2.9750e-01	-4.1641	3.126e-05	***
1	FULLWORKER3p	-1.1820e+00	1.8500e-01	-6.3894	1.666e-10	***
2	FULLWORKER3p	-1.4398e+00	3.6629e-01	-3.9307	8.470e-05	***
1	WHITECOLLAR1	9.0963e-02	7.7261e-02	1.1774	0.2390545	
2	WHITECOLLAR1	1.2135e-01	1.7152e-01	0.7075	0.4792441	
1	WHITECOLLAR2p	7.5711e-02	1.3231e-01	0.5722	0.5671564	
2	WHITECOLLAR2p	3.2997e-01	2.3120e-01	1.4272	0.1535183	
1	STUDENT1	8.5801e-02	8.4652e-02	1.0136	0.3107837	
2	STUDENT1	-2.9432e-01	1.8607e-01	-1.5818	0.1137044	
1	STUDENT2	2.7242e-01	1.0758e-01	2.5322	0.0113359	*
2	STUDENT2	-4.2835e-01	2.2892e-01	-1.8712	0.0613217	.
1	STUDENT3p	3.0469e-01	1.5781e-01	1.9307	0.0535189	.
2	STUDENT3p	-5.2509e-01	3.2649e-01	-1.6083	0.1077679	
1	KIDS1	1.1473e-01	9.4083e-02	1.2194	0.2226754	
2	KIDS1	-1.6763e-01	2.1547e-01	-0.7780	0.4365702	
1	KIDS2p	3.8690e-01	1.8508e-01	2.0905	0.0365732	*
2	KIDS2p	-3.1314e-02	3.8082e-01	-0.0822	0.9344656	
1	INC12	-3.2516e-01	1.5002e-01	-2.1674	0.0302029	*
2	INC12	-1.6454e-01	5.9777e-01	-0.2753	0.7831214	
1	INC3	8.0793e-01	1.6737e-01	4.8272	1.385e-06	***
2	INC3	8.4870e-01	6.1639e-01	1.3769	0.1685451	
1	INC4	1.6788e+00	1.7505e-01	9.5907	< 2.2e-16	***
2	INC4	2.5112e+00	5.5522e-01	4.5229	6.100e-06	***
1	INC5	2.1561e+00	1.9787e-01	10.8965	< 2.2e-16	***
2	INC5	3.5637e+00	5.7656e-01	6.1811	6.366e-10	***
1	INC6	2.8708e+00	2.0421e-01	14.0583	< 2.2e-16	***
2	INC6	4.7883e+00	5.7126e-01	8.3820	< 2.2e-16	***
1	ABOVE60	-2.0884e-01	7.2285e-02	-2.8891	0.0038631	**
2	ABOVE60	-1.7322e-01	1.5974e-01	-1.0844	0.2781762	
1	CEO	5.1387e-01	1.2483e-01	4.1166	3.845e-05	***
2	CEO	8.8771e-01	2.0008e-01	4.4367	9.134e-06	***
1	SELFEMPLOYED	1.8655e-01	1.0636e-01	1.7540	0.0794279	.
2	SELFEMPLOYED	5.2722e-01	2.0050e-01	2.6295	0.0085516	**
1	HHSIZE3	4.8378e-01	9.5250e-02	5.0791	3.793e-07	***
2	HHSIZE3	1.1843e+00	2.7747e-01	4.2681	1.971e-05	***
1	HHSIZE4	7.6383e-01	1.0804e-01	7.0700	1.550e-12	***
2	HHSIZE4	1.6895e+00	2.8486e-01	5.9308	3.014e-09	***
1	HHSIZE5	7.5341e-01	1.3432e-01	5.6091	2.033e-08	***
2	HHSIZE5	1.9222e+00	3.1753e-01	6.0536	1.416e-09	***
1	HHSIZE6plus	8.8358e-01	1.7253e-01	5.1213	3.034e-07	***
2	HHSIZE6plus	2.4499e+00	3.7507e-01	6.5318	6.498e-11	***
1	distMRT500	-1.2248e-01	7.3445e-02	-1.6676	0.0953921	.
2	distMRT500	-6.0583e-01	1.6948e-01	-3.5747	0.0003506	***
1	distMRT1000	-1.7331e-01	6.9273e-02	-2.5018	0.0123552	*
2	distMRT1000	-6.3918e-01	1.5390e-01	-4.1531	3.280e-05	***
1	distCBD	-1.3174e-03	4.8460e-03	-0.2719	0.7857288	
2	distCBD	-2.0160e-02	1.2541e-02	-1.6076	0.1079326	
1	TAXI	-1.3614e+00	2.3363e-01	-5.8270	5.644e-09	***
2	TAXI	-1.9170e+01	3.2646e+03	-0.0059	0.9953148	

[1]https://cran.r-project.org/web/packages/mlogit/index.html
[2]https://cran.r-project.org/web/packages/randomForest/index.html
[3]https://cran.r-project.org/web/packages/e1071/index.html
[4]https://cran.r-project.org/web/packages/rpart/index.html
[5]https://cran.r-project.org/web/packages/xgboost/index.html

TABLE II
DISCRETE CHOICE VS. MACHINE LEARNING MODEL PREDICTION ACCURACY RESULTS ON E_{MNL}.

Prediction Acuracies on 2012 HITS data, E_{MNL}

Multinomial Logit (MNL)
Accuracy 0.743

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3712	458	0
Observed	1 Car	981	957	14
	2+ Cars	30	137	20

Extreme Gradient Boosting
Accuracy 0.742

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3713	456	1
Observed	1 Car	986	949	17
	2+ Cars	35	136	16

Decision Tree Classifier
Accuracy 0.718

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3763	407	0
Observed	1 Car	1283	669	0
	2+ Cars	46	141	0

SVM Linear Kernel
Accuracy 0.642

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3911	244	15
Observed	1 Car	1808	138	6
	2+ Cars	180	7	0

Random Forest
Accuracy 0.734

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3627	542	1
Observed	1 Car	947	992	13
	2+ Cars	33	140	14

in Table II. We present both the prediction accuracy of our models and the confusion matrix on 2012 HITS. All machine learning models where run using the default parameters in their corresponding R packages, except for the random forest model, for which we use 1500 trees.

Using prediction accuracy as a measurement metric, our Multinomial Logit (MNL) predicts 2012 HITS slightly better (0.743) than the best machine learning model (XGB, with 0.742). However, when we look at the confusion matrices we observe that the slight superiority of the MNL model is less obvious when it comes to types of prediction or classification errors. For example, while the MNL model outperforms or virtually ties the XGB model in correctly classifying car ownership levels in almost all cells, the Decision Tree Classifier (DTC) outperforms the MNL model in classifying the number of HHs with 0 car ownership, and has less misclassification error for the HHs it predicts will have 1 car. However, the DTC performs poorly in comparison to the MNL on the lower triangle part of the matrix. We also observe that the DTC does not predict that any HH will own 2+ cars. Similarly to

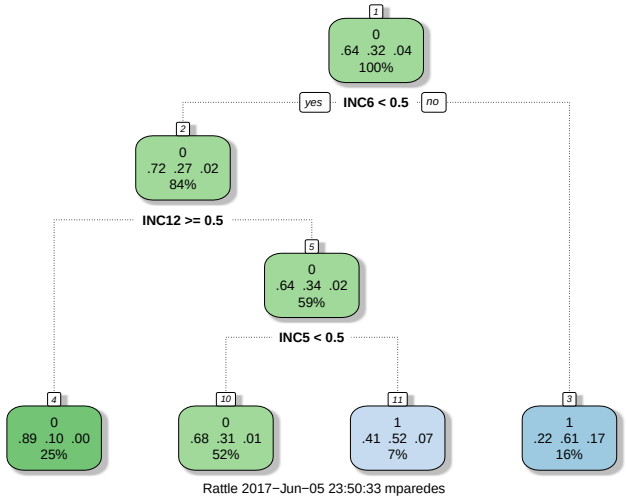

Fig. 1. Decision tree model

our discrete choice model results, our ML model shows an accuracy level that is consistent well performing models in similar classification tasks. The Decision Tree model is shown in Figure 1 and its most important features.

Income is the most important variable for the prediction of a HHs car ownership in our DCT model. Each node shows the classification, the probability of each category (0, 1 or 2+ cars) at that node, and the percentage of observations that fall in that node. The bottom row adds to 100%. For example, the first variable that the DTC classifies HHs on is whether their income is in the highest category (`INC6`). If a HH is in the highest income category (`INC6` > 0.5), then it directly classifies those households as having 1 car. If the HH is not in the highest income category (`INC6` < 0.5), then the DCT starts evaluating the HH on other income levels (`INC12` and `INC5`), and finally it classifies on whether the HH is of Indian ethnicity. This demostrates that for the DCT the income variable and the indian ethnicity variable are the most important features.

B. Prediction Using E_{ML}

Using E_{ML} and some ML parameter tuning, we obtain improved prediction accuracies. The results for the ML methods on the E_{ML} data set are shown in Table III. We run our random forest model with all default parameters except that we use 1500 trees. Our DCT model is run with all default parameters. Our extreme gradient boosting is run with a max depth of 6, an eta = 0.01, column sample by tree of 0.7, and all other parameters were the default. Finally, our ensemble between the random forest and the extree gradient boosting was produced by averaging the estimated probabilities. The results can be seen in Table III.

We see that the E_{ML} based models outperform the E_{DC} models on the prediction task based on accuracy. The random forest model has the highest accuracy, but has a higher rate of false positives (those cases where the ML model predicts a

TABLE III
MACHINE LEARNING MODEL PREDICTION ACCURACY RESULTS ON E_{ML}

Prediction Acuracies on 2012 HITS data, E_{ML}

Random Forest
Accuracy 0.799

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3505	665	0
Observed	1 Car	423	1512	17
	2+ Cars	4	163	20

Decision Tree
Accuracy 0.763

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3263	907	0
Observed	1 Car	404	1548	0
	2+ Cars	10	177	0

Extreme Gradient Boosting
Accuracy 0.749

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3957	213	0
Observed	1 Car	1185	766	1
	2+ Cars	57	129	1

Ensemble Random Forest & Extreme Gradient Boosting
Accuracy 0.789

		Predicted		
		0 Cars	1 Car	2+ Cars
	0 Cars	3822	348	0
Observed	1 Car	797	1155	0
	2+ Cars	17	169	1

HHs will have no, one, or two or more cars and it does not), specially for the two or more car HHs. However, the random forest has the lowest false negative rate for the two or more car HH case (cases when the ML model predicts that a HH will not have two or more cars and it does). All ML models have no two or more car HH false positive rates. The ensemble model did not outperform one individual component.

VI. CONCLUSION & FUTURE WORK

We conclude that both estimation models and prediction models, such as discrete choice model and machine learning models are useful, and not interchangable without extra effort. The MNL anchors the results given an interpretable theory. While the ML models minimize the prediction error by taking all data and not constraining the specification to linear relationships in the parameters.

We use 2008 data to train our machine learning models and to estimate our discrete choice model and use these models to predict 2012 car ownership in households. We compare our household car ownership model (multinomial logit model) against our 5 machine learning models (Random Forest, Support Vector Machines - SVMs, Decision trees, Extreme Gradient Boosting, and an Ensemble of methods) and find, that the machine learning models underperform on the prediction task on the data set prepared for MNL(E_{MNL}). As a consequence we assemble a data set for the ML models E_{ML} and the predictive accuracy of the ML methods outperform the MNL modell on the prediction task by 10%. However, our machine learning models are not interpretable, except for the decision tree classifier, which uses rules to classify HHs into categories.

These results show that machine learning and econometric models, having different objectives and mechanics, require datasets pre-processed in different ways. By not pre-processing or not doing feature engineering to obtain a dataset typical of machine learning methods, but rather training our machine learning models on the discrete choice model data, we do not fully take advantage of machine learning models. These findings suggest that if demand prediction is what is needed in a transportation context, machine learning methods might be able to aide urban and transportation planners.

Future work will look further at additional feature engineering and feature selection methods for both discrete choice models and machine learning models, and more advanced discrete choice models such as those using latent variables.

ACKNOWLEDGMENT

The authors would like to thank the Singapore-MIT Alliance for Research and Technology (SMART) for funding and research support, as well as the Government of Singapore for the data.

REFERENCES

[1] Muhammad Adnan, Francisco C Pereira, Carlos Miguel Lima Azevedo, Kakali Basak, Milan Lovric, Sebastián Raveau, Yi Zhu, Joseph Ferreira, Christopher Zegras, and M Ben-Akiva. Simmobility: A multi-scale integrated agent-based simulation platform. In *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*, 2016.

[2] Chandra R Bhat and Jessica Y Guo. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological*, 41(5):506–526, 2007.

[3] Robert Cervero and Kara Kockelman. Travel demand and the 3ds: density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3):199–219, 1997.

[4] CHOI Chik Cheong and Raymond Toh. Household interview surveys from 1997 to 2008–a decade of changing travel behaviours. *Editorial Team*, 52, 2010.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[6] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *The American economic review*, 105(5):491–495, 2015.

[7] Daniel McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974.

[8] Kenneth Train. A structured logit model of auto ownership and mode choice. *The Review of Economic Studies*, 47(2):357–370, 1980.

[9] Kenneth Train. *Discrete choice methods with simulation*. Cambridge university press, 2003.