Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models

Rounaq Basu [a,b,*] and Joseph Ferreira [a,b]

[a] *Massachusetts Institute of Technology, Cambridge, MA 02139, USA.*
[b] *Singapore-MIT Alliance for Research and Technology, 1 Create Way, 138602, Singapore.*

Abstract

Rising vehicle ownership trends have led to significant increases in negative externalities associated with transportation such as pollution and congestion. While empirical studies have typically used only econometric frameworks, we must ask the question: *Can machine learning models outperform traditional econometric approaches?* Using a socio-demographic dataset from Singapore, 22 feature vectors were constructed using appropriate transformations and imputing missing data to predict a 6-class categorical ordinal variable. In our comparison of six different supervised learning algorithms with the multinomial logit (MNL) model, we found that the neural network (NN) model was the most robust and performed the best while generalizing to the test dataset with a predictive accuracy score almost 10% better than the MNL. Consequently, we used an ordinal logit classification approach with neural network binary classifiers (OLC_NN) to address the imbalanced classification problem. This model was seen to perform the best in terms of all performance metrics, even in comparison with the ordinal logit (OL) model.

We also used the econometric models to obtain insights into the household vehicle ownership decision-making process. Singapore's public transport system and strict regulatory practices influenced not owning any vehicle to be the most preferred alternative. A gender effect was also revealed, along with a strong indirect income effect through housing type and job sector. Additionally, the direct income effect was statistically significant and strongly positive in magnitude. An attitudinal aspect was noticed in households with young professionals, wherein they are strongly disinclined to own a car. Proximity to transit stations and taxi ownership were also found to be significant factors in influencing vehicle ownership negatively. This research paves the way for an integrated framework that incorporates both the econometric and supervised learning approaches to better predict the influence of disruptive changes.

Keywords: Econometrics; Machine learning; Vehicle ownership; Supervised learning; Autonomous vehicles

* Corresponding author. *E-mail address:* rounaq@mit.edu

1. Introduction

Car ownership and use is expanding throughout the world. Historically, economic development has been strongly associated with an increase in the demand for transportation and particularly in the number of road vehicles (Dargary et al., 2007). Vehicle ownership may promote work if employment opportunities and job searches are enhanced by reliable transportation (Baum, 2009). For example, vehicles may serve to reduce potential physical isolation from employment opportunities. Since the growth in vehicle ownership is continuing hand-in-hand with rapid urbanization, the strains are particularly severe in cities (Button et al., 1993). Rising vehicle ownership trends have led to significant increases in negative externalities associated with transportation such as pollution and congestion. This has motivated policy-makers and researchers to examine vehicle ownership trends more closely over the past couple of decades. The primary reason for this is understandable, since having access to a vehicle increases an individual's (or their household's) travel options, leading to greater mobility. Secondary reasons for this scrutiny include the need to predict future transport investment in road infrastructure and the commercial demand for new vehicles.

Transportation literature is quite rich in this field with several different approaches having been applied to examine the underlying factors influencing vehicle ownership. Researchers agree that such an exercise is highly contextual and is quite difficult to generalize at a large geographical scale. Most studies focus on a particular study region because of this very reason. A few examples can be found in England & Wales (Clark, 2009), Mexico City (Guerra, 2015), Montreal (Anowar et al., 2016), Honolulu (Ryan & Han, 1999), Bangkok (Dissanayake & Morikawa, 2010), the United States (Liu et al., 2014), China (Zhang et al., 2017) and in the Netherlands (Oakil et al., 2016). Although far and few in between, comparisons have been made across rural and urban regions by Dargary (2002) and Choudhary & Vasudevan (2017), across time by Wu et al. (2016), and across countries by Law et al. (2015).

While it is widely agreed that socio-demographics play a major role in determining vehicle ownership and use, other influencing factors have also been discovered. The built environment, including both neighborhood design characteristics and relative location, has potentially causal influences on vehicle ownership decisions, as was found by Zegras (2010) in Santiago de Chile and Macfarlane et al. (2015) in Atlanta. Wu et al. (1999) explored the idea of car ownership having a symbolic utility, which refers to the psychological satisfaction derived from owning and using a vehicle. Neighborhood-level population density is also considered to affect the amount of household automobile travel (Schimek, 1996). The effects of environmental knowledge and attitudinal perceptions on the numbers and types of vehicles owned per household were also explored by Flamm (2009). Therefore, it is important to understand the causal mechanics of vehicle ownership as it is a key determinant of choice of mode for travel decisions.

Economists generally model vehicle ownership as a function of per-capita income using econometric estimation techniques and obtain projections of the growth in car and total vehicle stock (Dargay & Gately, 1999). On the other hand, empirical studies on household car ownership have used two types of discrete choice modeling structures: *ordered* and *unordered* (Potoglou & Susilo, 2008). In ordered response structures, such as the ordered logit and ordered probit models, the choice of the number of household vehicles arises from a unidimensional latent variable that reflects the propensity of a household to own vehicles. Unordered response structures are based on the random utility maximization principle, which assumes a household associates a utility value across different car ownership levels and chooses the one with the maximum utility. The most common unordered response models are the multinomial logit and probit models, but only the multinomial logit has been used in practical applications because of its simple structure and low computational requirements. A slightly more sophisticated approach incorporating psychological and sociological factors as explanatory variables is the latent class model.

With access to big data, machine learning approaches are being widely used for various applications across the world. However, we see that there is no literature pertaining to the use of machine learning models for predicting and forecasting vehicle ownership. Can machine learning models outperform traditional econometric approaches such as the multinomial logit model? Therefore, *our first objective is to use different supervised learning algorithms to perform classification for a vehicle ownership target variable on a dataset from Singapore, and compare results with* *traditional econometric models*. Our empirical findings can provide insights into selecting better predictive algorithms and obtaining forecasts that are more accurate.

Although not without critique, econometric approaches use an underlying economic framework to explain behavioral decisions. This is particularly useful while trying to understand potentially disruptive phenomenon such as the introduction of electric vehicles or autonomous mobility-on-demand (AMoD) in the market. Supervised learning techniques are not applicable to such scenarios because such options are absent from current existing data. In order to evaluate the current and future prospects of electric vehicles, a few studies have tried to explore factors influencing ownership (Kim et al., 2015), penetration rates (Javid & Nejat, 2017), and vehicle usage (Danielis et al., 2018). While Basu et al. (2018) has explored the effect of AMoD on urban mass transit, a systematic analysis of the effect on private vehicle ownership is yet to be conducted. Therefore, *our second objective is to obtain insights into the underlying socio-demographic factors influencing household vehicle ownership decisions in Singapore using traditional econometric models.*

We chose Singapore as our study region for two reasons. First, Singapore is experiencing strong technological innovations in urban mobility and is scheduling pilots with AMoD and commercially available autonomous vehicles (AVs) in the near future. This enhances the importance of understanding current trends in vehicle ownership in Singapore (which is absent from literature) for better predictability while incorporating such disruptive services in choice sets. Second, we are engaged in the SimMobility project in Singapore that aims to build a city-scale agent-based microsimulation of demographic evolution, vehicle ownership changes, residential relocation, and urban transportation movements. Therefore, it is pertinent to compare machine learning techniques with econometric approaches before incorporating them into our simulation framework. A noteworthy point about Singapore is that the government uses a scheme called the Off-Peak Car (OPC) to reduce usage of cars in return for reduced car registration related fees and road taxes. We consider this as one of our mobility alternatives as it represents the spirit of low-sample disruptive technologies such as AVs.

In this paper, we first compare six different supervised learning algorithms with the multinomial logit model, which is the most popular approach for modeling the level of household car ownership as shown in transportation literature (Potoglou & Susilo, 2008). We construct the choice set drawing on the concept of Mobility-as-a-Service (MaaS) such that they can be considered as an ordered arrangement with incremental utility. This allows us to also consider the ordinal logit model, which is then compared with a classification algorithm that uses the best classifier obtained from the previous comparison and implements the ordinal logistic regression methodology. We estimate our models on a dataset obtained from a household survey and use the calibrated model to predict household vehicle ownership for a synthetic population of Singapore consisting of 1.77 million households in 2012, which is projected to grow to over 1.9 million by 2030. The need for this comparison stems from the observation that traditional econometric approaches do not have a high accuracy rate in prediction, leading to a cascading error effect in the predictions for the synthetic population that would introduce errors in all our urban simulations using SimMobility. Therefore, we aim to understand the factors influencing household vehicle ownership through econometrics while examining machine learning approaches for obtaining better predictive accuracy.

2. Background & Methodology

We provide the mathematical formulations and methodological details about our implementation of the econometric and machine learning models in the following sub-sections.

2.1. Econometrics

We select the multinomial logit model as a representative of the unordered response structure and the ordinal logit model as a representative of the ordered response structure. Note that our choice set of household vehicle ownership is constructed in an ordered fashion so that we can examine whether considering the incremental ordered response structure yields a better understanding of underlying preferences and consumer behavior. The mathematical framework of both these models are provided in the following sub-sections.

2.1.1. Multinomial logit (MNL)

Random utility theory is a structural component of behavioral theory (Manski, 1977) and is a well-explored concept in the field of economics. Expressed briefly, it states that individual n selects alternative i that has the highest utility U_{in} among those in the choice set C_n. Utility U_{in} is composed of a systematic utility component that can be expressed as a linear-in-parameters function of variables (V_{in}) and a random utility component (ϵ_{in}).

$$U_{in} = V_{in} + \epsilon_{in} = \beta' X_{in} + \epsilon_{in} \tag{1}$$

Therefore, the probability of individual n selecting alternative i from choice set C_n can be expressed as follows:

$$\begin{gathered} P(i|C_n) = P\big(U_{in} \geq U_{jn}, \forall j \in C_n\big) = P\big(U_{in} - U_{jn} \geq 0, \forall j \in C_n\big) \\ \Rightarrow \;\; P(i|C_n) = P\left(U_{in} = \max_j U_{jn}\,, \forall j \in C_n\right) \end{gathered} \tag{2}$$

Consider the case of binary choice as an example to obtain a tractable expression.

$$\begin{gathered} P_n(1) = P(U_{1n} \geq U_{2n}) = P(U_{1n} - U_{2n} \geq 0) = P(\epsilon_{2n} - \epsilon_{1n} \leq V_{1n} - V_{2n}) \\ \Rightarrow \;\; P_n(1) = F_{\epsilon_2 - \epsilon_1}(V_{1n} - V_{2n}) \end{gathered} \tag{3}$$

As can be seen from the expression above, this is the univariate CDF of $(\epsilon_2 - \epsilon_1)$. Similarly, an extension to three alternatives in the choice set would result in the bivariate CDF of $(\epsilon_2 - \epsilon_1)$ and $(\epsilon_3 - \epsilon_1)$. Different assumptions are made on the joint distribution of $\epsilon = \big(\epsilon_1, \epsilon_2, \dots, \epsilon_J\big)'$ leading to different models. For example, there are two major assumptions on the joint distribution of ϵ that lead to the logit model. Note that a logit model can be *binary*, i.e. there are only two alternatives in the choice set, or *multinomial*, i.e. there are multiple (greater than two) alternatives in the choice set.

First, ϵ_{jn} is independently and identically distributed (i.i.d.). This assumption of homoscedasticity leads to the following simplification:

$$f\big(\epsilon_1, \dots, \epsilon_J\big) = \prod_{j=1}^{J} f\big(\epsilon_j\big) \tag{4}$$

Second, ϵ_{jn} follows an extreme value distribution with parameters as zero and μ. Thus, we can express $\epsilon_{jn} \sim EV(0, \mu)\ \forall j$. The CDF and PDF expressions for such a distribution are as follows:

$$\begin{gathered} F(\epsilon) = \exp(-e^{-\mu\epsilon})\,, \qquad \mu > 0 \\ f(\epsilon) = \mu e^{-\mu\epsilon} \exp(-e^{-\mu\epsilon}) \end{gathered} \tag{5}$$

Based on these assumptions, we can arrive at a tractable expression for the choice probability of each alternative.

$$P(i|C_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}} \tag{6}$$

From the above expression, we can see that the logit model has an important property – *independence from irrelevant alternatives* (IIA). We can express the IIA property through the odds ratio as follows:

$$\frac{P(i|C_{1n})}{P(j|C_{1n})} = \frac{P(i|C_{2n})}{P(j|C_{2n})} \tag{7}$$

where $i, j \in C_{1n}\,; i, j \in C_{2n}\,; C_{1n} \subset C_n\,; C_{2n} \subset C_n$. However, this property is quite restrictive and, thus, the logit model is only appropriate when the alternatives are uncorrelated.

2.1.2. Ordinal logit (OL)

As opposed to the MNL which establishes a relationship between the covariates and the set of probabilities of the alternatives, the ordinal logit (OL) model is used to obtain expressions for the cumulative probabilities. An OL model for an ordinal response Y_i with K classes is defined by a set of $(K-1)$ equations as the last cumulative probability is necessarily equal to 1.

$$logit(p_{ki}) = \log\left(\frac{p_{ki}}{1-p_{ki}}\right) = \tau_k - \beta' X_{ki} \tag{8}$$

where $k = 1,2, \dots, K-1$ and $p_{ki} = \Pr(Y_i \leq y_k | X_i)$ is the cumulative probability. The parameters τ_k are called thresholds or cut-points, and are in increasing order $(\beta_1 < \beta_2 < \cdots < \beta_{K-1})$. An identification problem arises in the simultaneous estimation of the overall intercept β_0 (which is a part of the vector β') and all the $(K-1)$ thresholds, which can be solved by either omitting the overall constant from the linear predictor $(\beta_0 = 0)$ or fixing the first threshold to zero $(\tau_1 = 0)$. We use the former approach in our implementation for this paper.

It should be noted that the vector of slopes in the linear predictor β is not indexed by the class index k, which indicates that the effects of the covariates are constant across response categories. This is known as the parallel regression assumption, which yields $(K-1)$ parallel lines while plotting $logit(p_{ki})$ against a covariate. We purposely introduce the negative sign before β such that the interpretation is according to intuition. With this model specification, we can imply that increasing a covariate with a positive slope would be associated with a rise in the probabilities of the higher classes. We can now express the cumulative probability for class k in the following manner.

$$p_{ki} = \frac{\exp(\tau_k - \beta' X_{ki})}{1 + \exp(\tau_k - \beta' X_{ki})} = \frac{1}{1 + \exp(-\tau_k + \beta' X_{ki})} \tag{9}$$

The OL model is also known as the proportional odds model because the parallel regression assumption implies that the ratio of odds for two classes is constant across response categories. This can be expressed as the following:

$$\frac{odds_{ki}}{odds_{kj}} = \exp\left[\beta'\left(X_j - X_i\right)\right] \tag{10}$$

where odds for a class represent the proportionality of the odds of not exceeding that class, i.e. $odds_{ki} = p_{ki}/(1-p_{ki})$. Readers interested in further details about the modeling of ordinal outcomes in the setting of choice theory should refer to Greene and Hensher (2010).

2.2. Machine Learning

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. In machine learning, multiclass or multinomial classification is the problem of classifying instances into one of three or more classes. There are multiple algorithms, known as classifiers, which use a mathematical function to map input data to a category through a concrete implementation. We explore six different types of classifiers in this research. Note that we are interested in only classification, not clustering, algorithms as our objective of vehicle ownership prediction is a supervised learning problem (Kotsiantis et al., 2007).

2.2.1. Decision Tree (DT)

Decision Trees (DT) refer to a non-parametric supervised learning method that aims to predict the value of a target variable by learning a set of simple if-then-else decision rules inferred from the features. The deeper the tree, the more complex the decision rules and the better the model fit. DT models are simple to understand and interpret, and can be visualized easily. They require little data preparation, can handle both numerical and categorical data, and perform well even if the assumptions are violated by the true model from which the data is drawn. However, they tend to easily overfit by creating over-complex trees that do not generalize well. They can also be unstable as small variations in the data can result in a completely different tree being generated. Moreover, locally optimal decisions made at each node using greedy algorithms cannot guarantee to return the globally optimal model (Anyanwu & Shiva, 2009).

2.2.2. Random Forest (RF)

Many of the drawbacks associated with DT models can be overcome by using them within an ensemble learner like the random forest (RF) algorithm. The goal of an ensemble method is to combine the predictions of multiple base estimators built with a given learning algorithm in order to improve the generalizability and robustness over a single estimator. In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Note that, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Contrastingly, the split that is picked is the best split among a random subset of the features. Because of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree). However, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model. Our implementation combines classifiers by averaging their probabilistic predictions, instead of aggregating to a single class vote for each classifier as shown in Breiman (2001).

2.2.3. Neural Network (NN)

We use the Multi-layer Perceptron (MLP) algorithm that learns a function f(.): $R^m \rightarrow R^o$ by training on a dataset, where m is the number of input dimensions and o is the number of output dimensions (Haykin, 2009). There can be one or more non-linear layers, called hidden layers, between the input and output layer. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation, followed by a non-linear activation function g(.): $R \rightarrow R$. The output layer receives the values from the last hidden layer and transforms them into the output values using the final activation function. The MLP classifier minimizes the cross-entropy loss function and trains using gradient descent where the gradients are calculated using backpropagation. While the MLP classifier performs well with non-linear models and on-line learning, it is sensitive to feature scaling and requires tuning several hyperparameters such as the number of hidden neurons, layers, and iterations. In our implementation, we used two hidden layers (100 neurons and 6 neurons respectively) in a feed-forward neural network architecture with rectified linear unit (ReLU) as the activation function for the hidden layers and the softmax activation function for the output layer.

2.2.4. Support Vector Machine (SVM)

A support vector machine (SVM) constructs a set of hyperplanes in a high-dimensional space that can be used for classification. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class, known as the margin, since a larger margin generally implies lower generalization error of the classifier. The objective function that the SVM tries to minimize can be expressed as follows (Hsu & Lin, 2002).

$$J(\theta,\theta_0)=\frac{1}{n}\sum_{i=1}^{n}L_h\left(y^{(i)}(\theta^T\phi(x)+\theta_0)\right)+\lambda\left|\left|\theta\right|\right|^2 \tag{11}$$

where $\phi(x)$ is the kernel or basis function of original feature vector x, y is the label vector, λ is the regularization parameter, $L_h(.)$ is the hinge loss defined as $L_h=\max(0,x)$, θ is the weight vector, and θ_0 is the bias. The first term of $J(\theta,\theta_0)$ measures the loss, which is a measure of the difference between true and predicted labels. The second term is the regularization term that prevents the classifier from overfitting the data and increases its generalizability. While other loss functions can be used, the most commonly used one is the hinge loss. Instead of using a linear kernel, we used a *radial polynomial basis* (rbf) kernel to extend generalizability and capture possible non-linear relationships in the data. We used the Pegasos algorithm to solve the optimization problem cast by the SVM, which has been found to extend well to non-linear kernels (Shalev-Shwartz et al., 2011).

2.2.5. Logistic Regression (LR)

Logistic Regression (LR) is a linear classification algorithm that models the probabilities describing the possible outcomes of a single trial using a logistic function. The objective function that the LR algorithm tries to minimize can be expressed as follows.

$$J(\theta,\theta_0)=-\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right)+\lambda\left|\left|\theta\right|\right|^2 \tag{12}$$

This is different from the SVM objective function in the sense that the loss function is the cross-entropy loss, also known as the log loss, instead of the hinge loss. These smooth functions make it easy to calculate the gradient and minimize loss. There are two techniques to implement this algorithm. The first is the more common one-vs-rest approach where separate binary classifiers are trained for all classes. The other is the true multinomial logistic regression approach that provides better calibrated probability estimates (Caruana & Niculescu-Mizil, 2006). We have implemented the latter for this research using the stochastic average gradient descent solver for faster convergence.

2.2.6. Stochastic Gradient Descent (SGD)

Stochastic gradient descent (SGD) is a simple yet very effective approach to fit linear models, which is particularly useful when the number of samples or the dimensionality of the problem is very large (Bottou, 2010). Similar to neural networks, SGD is also sensitive to feature scaling and requires tuning of hyperparameters. We implemented a multi-class SGD classifier through a combination of multiple binary classifiers in a one-vs-rest approach. For each of the N classes, a binary classifier is learned that discriminates between that class and all other $N-1$ classes. At testing time, we compute the confidence score (i.e. the signed distances to the hyperplane) for each classifier and choose the class with the highest confidence. Since we have already used a variant of SGD in our LR implementation and we are using a non-linear kernel in our SVM implementation, we selected the linear SVM as the objective function for implementing this algorithm.

2.2.7. Ordinal logit classification (OLC)

While the six classification algorithms described above will be compared with the MNL model, we need a separate approach for comparison with the OL model. To that effect, we use a modified rank classification algorithm that can be described through the following steps.

a) Create $(K-1)$ binary labels indicating association with cumulative classes such that label Y_1 indicates that the sample belongs to class 1, label Y_{12} indicates that the sample belongs to either class 1 or class 2, and so on.

b) Select the algorithm that performed the best in the comparison with MNL, and use it to train $(K-1)$ binary classifiers for the $(K-1)$ labels indicated above.

c) Use the trained classifiers to predict probabilities for each sample. Note that these probabilities are cumulative probabilities of belonging to one or more classes.

d) Individual class probabilities can be calculated from these cumulative probabilities such that the jth binary classifier trained on label $Y_{12\ldots j}$ predicts the probability of the sample belonging to a class up to the jth category, i.e. $Pr(Y_i \in \{1,2,\ldots,j\})$. The individual class probabilities can be expressed as follows.

   a. $P_i(1) = Pr(Y_i \in \{1\})$

   b. $P_i(j) = Pr(Y_i \in \{1,2,\ldots,j\}) - Pr(Y_i \in \{1,2,\ldots,j-1\}) = Pr(Y_i \in \{1,2,\ldots,j\}) - P_i(j-1)$

   c. $P_i(K) = 1 - Pr(Y_i \in \{1,\ldots,K-1\}) = 1 - P_i(K-1)$

e) The sample is attributed to the class having the highest probability for that sample.

2.2.8. Need for cross-validation

From the previous sub-sections, we can see that the SVC, LR and SGD algorithms involve a regularization parameter. Therefore, additional analysis is required to obtain the optimal value of the hyperparameter λ (which influences regularization) since it depends both on the training dataset and the learning algorithm. Therefore, 10-fold cross-validation was used to randomly create train and validation data sets from the overall training dataset (with replacement and shuffling) and specify a grid of values for λ. After extensive testing with different grid sizes, we decided to use $\lambda \in [0.01,10]$ with a step size of 0.01. Now, for each value of λ within this specified grid and each train dataset (90% of training samples), the learning algorithm was used to obtain a score/accuracy for the validation dataset (10% of training samples). Consequently, we obtained two 1000 x 10 matrices containing train scores and validation scores for each of these runs. By averaging out the score over the 10 cross-validation runs, a 1000 x 1 column vector of average scores was obtained. Since the validation score needs to be maximized, λ is said to be optimal for the highest score in the column vector of validation scores. Thus, we obtained an optimal hyperparameter λ^* for the training data set with which we used the learning algorithm on the entire training dataset to obtain optimal weights θ^*. Finally, we employed the optimal λ^* and θ^* to provide predictions for the test dataset.

3. Data

This research uses data gathered through a paper-based survey called the Household Information & Travel Survey (HITS) conducted in Singapore in 2012. The HITS survey is carried out once every four years and is used to collect data about households, individuals and their travel patterns for one observed working day. The survey contains three sections – household particulars, individual particulars and trip particulars. Socio-demographic characteristics about 9,584 households and 35,714 individuals were recorded in the first two sections. The final section contains data about each stage of a trip that each individual undertook with trip details such as point of origin/destination, travel time, mode, purpose, etc. This research focuses only on socio-demographic characteristics; the key variables of interest are summarized in Table 1.

Additionally, it is pertinent to mention that there are 116,624 residential postcodes in Singapore, so the residential location of each household is specified at the building level. Distance to transit is often considered an important determinant of long-term household decisions. It has been found that households living in proximity to transit stations are less likely to own a car (Kim and Kim, 2004). In light of this, we used locations of mass rail transit (MRT) stations in Singapore and postcode locations of each household to create a measure for distance to transit which captures the proximity of a household to the nearest MRT station (in meters).

We also noticed that the HITS survey sample underestimated taxi ownership in Singapore by a significant amount. This is a particularly critical flaw in the sampling technique because households with a taxi are unlikely to be owning an additional private car. After using sampling weights, the total taxi count in the weighted HITS population is around

18,500, while the actual taxi count in Singapore in 2012 was close to 25,500. Therefore, households that would be most likely to own a taxi based on employment, occupation and industry of individuals' jobs were identified. We then used an imputation method to randomly assign taxis to a subset of these selected households through a weighted iterative proportional fitting procedure such that the total taxi count reached close to 25,500.

The primary challenge of creating meaningful features is that most of the variables in the survey are at the individual level while the target variable for this project (vehicle ownership) is a household level variable. Therefore, the original variables shown in Table 1 had to be recoded and aggregated at the household level. The methodology for the feature engineering exercise is shown in Table 2. It should be noted that the individual level features were constructed such that they can be aggregated to the household level by taking a sum over all members of each household. Household income was scale-adjusted by dividing the sum of all individual incomes by 10,000.

Finally, 22 explanatory features (21 socio-demographic features and taxi ownership) were obtained for 9,584 households with one additional ordered multi-class feature denoting vehicle ownership. This data set was separated into training and test data sets based on an 80-20 split. The training data set has 7,667 samples, while the test data set has 1,917 samples.

Table 1. Description of HITS survey questionnaire.

Section	Variable	Original Encoding	Examples
Household (HH)	Dwelling Type	14 categories	HDB 1-room, private flat, etc.
	Ethnicity	4 categories	Chinese, Malay, Indian, Others
	Household Size	Continuous	1, 2, etc.
	Available Vehicles	9 categories	Normal car, Taxi, LGV, etc.
	Vehicle Properties	5 categories	Individual registered, rental, etc.
	Sampling Weight	Continuous	50, 100, etc. (Can be fractional)
Individual (IND)	Age	17 categories	6-9 years old, 10-14 years old, etc.
	Resident Status	3 categories	Singapore citizen, Permanent Resident, Others
	Gender	2 categories	Male, Female
	Driving License	4 categories	Car, Motorcycle, Van/Lorry/Bus, None
	Employment Status	11 categories	Employed Full Time, Full Time Student, etc.
	Occupation	11 categories	Professional, service and sales worker, etc.
	Industry	12 categories	Manufacturing, construction, etc.
	Monthly Income	13 categories	No income, \$1-\$1000, \$1001-\$1499, Refused, etc.

4. Results & Discussion

First, we describe the results obtained from the econometric approaches in order to better understand the underlying effects driving vehicle ownership in Singapore. Second, we provide insights into the cross-validation approach by presenting results leading to selection of the optimal hyperparameters for the machine learning models. Finally, we compare the performance of all models and comment on the results.

4.1. MNL model estimation results

Since having no vehicle (*CAT_0*) had the largest share in the dataset, we considered that to be the base or reference alternative which implies that the utility for *CAT_0* is always specified to be zero for all households. It should also be noted that we used a weighted log-likelihood approach to estimate the MNL model since sample weights were specified for each household in the HITS sample. The estimation results from the MNL model are presented in Table 3. All the intercepts are seen to be negative, indicating that not owning any vehicle is the most preferred alternative. This might be due to Singapore's excellent public transport system, and strict regulatory practices and high taxes related to car purchases and driving licenses. We see that vehicle ownership, especially both car and motorcycle, is strongly influenced by the number of children in the household.

Table 2. Feature engineering of independent variables.

Level	Variable	Original Encoding	Feature Engineering
Household (HH)	Dwelling Type	14 categories	*One-hot encoding to obtain 3 binary features* • HDB_1: HDB with 1 room • HDB_2P: HDB with 2 or more rooms • PRIVATE: Privately owned
	Ethnicity	4 categories	*One-hot encoding to obtain 3 binary features* • MALAY: Malaysian • INDIAN: Indian • OTHER: Others
	Household Size	Continuous	*Dropped from consideration; handled through individual data*
	Available Vehicles	9 categories	*Combined to create 1 multi-class feature for private vehicle ownership* • ***CAT_0:*** No vehicle • ***CAT_1:*** 1+ motorcycle only • ***CAT_2:*** 1 off-peak car w/wo motorcycle • ***CAT_3:*** 1 normal car only • ***CAT_4:*** 1 normal car and 1+ motorcycle • ***CAT_5:*** 2+ normal cars w/wo motorcycle *1 binary feature created for taxi ownership* • TAXI_BIN: One or more taxi(s)
	Vehicle Properties	5 categories	
	Sampling Weight	Continuous	*Kept as is*
	Distance to nearest MRT station	-	*One-hot encoding to obtain 2 binary features* • DIST_MRT_500: Distance < 500 m • DIST_MRT_1000: Distance >= 500 m AND < 1000 m
Individual (IND)	Age	17 categories	*One-hot encoding to obtain 2 binary features* • CHILD: Age < 15 years • SENIOR: Age > 60 years
	Resident Status	3 categories	*1 binary feature to capture permanent residentship* • RESIDENT: Individual is a Singapore citizen (SC) or a permanent resident (PR)
	Gender	2 categories	*1 binary feature for male gender* • MALE: Individual is male
	Driving License	4 categories	*One-hot encoding to obtain 3 binary features* • CAR_LIC: Has a car license • MOTOR_LIC: Has a motorcycle license • NO_LIC: Does not have a license
	Employment Status	11 categories	*One-hot encoding to obtain 3 binary features* • WORKER: Employed full-time/ part-time/ self-employed • RETIRED: Retired • YOUNGPRO: Young professional (Worker and Age ∈ [25,34])
	Occupation	11 categories	*1 binary feature for white collar jobs* • WHITECOLLAR: Professional/ Associate professional & technician/ Legislator, senior official & manager
	Industry	12 categories	*1 binary feature for blue collar jobs* • BLUECOLLAR: Manufacturing/ Construction
	Monthly Income	13 categories	*A log-normal distribution was created for income and each individual's income was randomly sampled from this distribution from an interval based on the category mentioned in HITS. Missing/refused incomes were imputed based on other individual characteristics.*

Table 3. Estimation results from the multinomial logit (MNL) model on the training dataset [1, 2, 3].

Independent variable	***CAT_1***	***CAT_2***	***CAT_3***	***CAT_4***	***CAT_5***
Intercept	-4.696***	-5.792***	-1.996***	-8.606***	-7.012***
	(-14.82)	*(-15.28)*	*(-14.87)*	*(-15.15)*	*(-22.05)*
No. of children in HH	0.114	0.586***	0.304***	0.609***	0.457***
	(1.30)	*(6.12)*	*(7.45)*	*(4.37)*	*(5.48)*
No. of seniors in HH	0.044	0.147	-0.026	0.409**	0.295**
	(0.33)	*(0.61)*	*(-0.47)*	*(2.14)*	*(2.56)*
No. of citizens and permanent residents in HH	0.495***	0.460***	0.395***	0.936***	0.385***
	(3.83)	*(2.59)*	*(6.60)*	*(3.54)*	*(2.88)*
No. of males in HH	-0.129	-0.264***	-0.256***	-0.237	-0.235***
	(-1.20)	*(-2.14)*	*(-5.06)*	*(-1.35)*	*(-2.37)*
No. of HH members with car licenses	-0.613***	0.993***	1.259***	1.047***	2.184***
	(-4.42)	*(5.03)*	*(16.22)*	*(4.24)*	*(13.82)*
No. of HH members with motorcycle licenses	3.602***	1.340***	-0.152	2.693***	0.380
	(22.40)	*(7.14)*	*(-1.23)*	*(13.10)*	*(1.58)*
No. of HH members with no licenses	-0.430***	-0.462***	-0.233***	-0.776***	-0.072
	(-3.33)	*(-2.66)*	*(-3.65)*	*(-3.23)*	*(-0.49)*
No. of workers in HH	-0.037	0.141	-0.272***	0.109	-0.510***
	(-0.29)	*(1.08)*	*(-4.58)*	*(0.57)*	*(-4.03)*
No. of retirees in HH	0.041	-0.171	-0.003	-0.015	-0.507***
	(0.26)	*(-0.51)*	*(-0.02)*	*(-0.03)*	*(-2.95)*
No. of young professionals in HH	-0.078	0.034	-0.158***	-0.097	-0.054
	(-0.66)	*(0.20)*	*(-2.73)*	*(-0.56)*	*(-0.34)*
No. of white-collar employees in HH	-0.051	0.170	0.165***	-0.001	0.392***
	(-0.46)	*(1.26)*	*(3.07)*	*(-0.02)*	*(3.62)*
No. of blue-collar employees in HH	0.285**	0.139	0.087	0.044	0.265*
	(2.00)	*(0.69)*	*(1.20)*	*(0.17)*	*(1.84)*
HDB with one room	0.442	-	-2.339***	-	-
	(0.97)		*(4.00)*		
HDB with two or more rooms	0.224	-0.358*	-0.691***	-0.718**	-1.462***
	(1.19)	*(-1.99)*	*(-9.02)*	*(-2.54)*	*(-6.04)*
Privately owned house	0.342	-0.063	0.578***	0.752*	1.904***
	(0.95)	*(-0.26)*	*(5.09)*	*(1.80)*	*(9.81)*
HH of Malaysian ethnicity	0.514***	1.163***	-0.064	0.890***	-0.103
	(2.78)	*(5.39)*	*(-0.51)*	*(2.86)*	*(-0.26)*
HH of Indian ethnicity	0.309	0.263	-0.732***	-0.418	-1.762***
	(1.49)	*(1.04)*	*(-6.74)*	*(-0.92)*	*(-6.01)*
House within 500 meters of MRT station	0.043	-0.083	-0.253***	0.069	-0.611***
	(0.22)	*(-0.38)*	*(-2.92)*	*(0.22)*	*(-2.95)*
House within 500 to 1,000 meters of MRT station	0.227	0.042	-0.183***	-0.407	-0.172
	(1.29)	*(0.19)*	*(-2.33)*	*(-1.34)*	*(-1.00)*
HH income (divided by 10,000)	0.977***	0.631**	1.182***	1.395***	1.521***
	(3.90)	*(2.00)*	*(9.95)*	*(4.24)*	*(9.40)*
HH owns a taxi	-0.004	-0.783**	-1.454***	-1.205	-1.007
	(-0.09)	*(-2.07)*	*(-5.89)*	*(-1.29)*	*(-1.73)*

[1] Parameter is statistically significant at 90% (*), 95% (**), or 99% (***) confidence level.

[2] Coefficient estimates are presented outside parentheses, while t-statistics are shown using italics inside parentheses.

[3] ***CAT_0:*** No vehicle; ***CAT_1:*** 1+ motorcycle only; ***CAT_2:*** 1 off-peak car w/wo motorcycle; ***CAT_3:*** 1 normal car only; ***CAT_4:*** 1 normal car and 1+ motorcycle; ***CAT_5:*** 2+ normal cars w/wo motorcycle

Interestingly, households with more senior citizens have multiple vehicles but prefer not to own only a car. This may be because such households have higher total incomes, or have diverse mobility needs in the case of large, joint-family households. Households with Singapore citizens or permanent residents prefer owning vehicles with the most preferred alternative being owning both a car and one or more motorcycles. An interesting observation is the effect of gender on vehicle ownership. Households with a higher number of males are less inclined to own any vehicle. Therefore, this might indicate that vehicle ownership in Singapore is strongly influenced by female members of the household. The effect of individuals owning licenses for cars or motorcycles is consistent with intuitive expectations, as seen from the signs of the coefficients. Households with a higher number of workers are strongly disinclined towards owning cars, which might again point towards Singapore's public transit system being preferred for commute trips. Similarly, a retiree-dominated household would not prefer to own cars, especially not a bundle of multiple vehicles.

Young professionals are strongly disinclined to own a car, but white-collar employees are more likely to own a car or multiple vehicles. This might be an attitudinal difference wherein younger people are more aware of sustainable transport trends and exhibit transit-friendly behavior. It may also be caused by an income effect since young professionals are not likely to be as wealthy as white-collar employees. On the other hand, blue-collar employees prefer to own motorcycles, thereby pointing to the class divide driven by income. Households living in government-provided housing (colloquially known as HDBs) are much less likely to own cars and more likely to own motorcycles. Contrastingly, households living in privately owned housing prefer car-dependent alternatives. This is another indication of the strong income effect evident in Singapore.

Households of Malaysian ethnicity are highly likely to own motorcycles and off-peak cars. This may be related to citizenship and the relative ease of obtaining vehicles and licenses. Interestingly, households of Indian ethnicity are strongly disinclined towards car-dependent alternatives. We also discover the effect of proximity to transit on vehicle ownership. Households located within 500 meters of an MRT station are less likely to own a car or have multiple vehicles compared to households located further away from the MRT. While the same can be said for households located between 500 to 1,000 meters of an MRT station, the magnitude of the proximity effect decreases. In line with our expectations, household income is a strong determinant of vehicle ownership wherein larger household incomes lead to multiple vehicle ownership. Taxi ownership has a negative effect on ownership of all other vehicles, with the strongest and most significant effect being on car ownership.

4.2. OL model estimation results

We present the estimation results of the OL model using the training dataset in Table 4. Recall from the OL model specification that the coefficients are actually log-odds and the odds can be represented by taking the exponents of the coefficients. There are two equivalent approaches of interpreting these results: (a) A negative log-odds value implies that the probability of the sample belonging to a higher class decreases with a unit increase in the independent variable, and vice-versa, and (b) A unit increase in the independent value will result in an increase in the probability of belonging to a higher class by a value equal to the odds. We will proceed with the latter interpretation, as it provides a direct magnitude of the change in probability. Note that odds being greater than one implies an increase in the probability which we term as positive influences on the mobility scale, while a decrease in the probability is associated with the odds being less than one which we term as negative influences on the mobility scale.

Several negative influences are easily noticeable, one of which is the number of household members with no licenses. Since our choice set contains private vehicle alternatives which require licenses, this is in line with expectations. As indicated by the MNL model, we find that households with higher number of workers or retirees or young professionals are less likely to be higher up on the mobility scale. We also notice strong indications of the indirect effect of income on vehicle ownership through housing type, where households living in public housing (HDBs) are more likely to be lower on the mobility scale. It is worth noting that living in an HDB with only one room has the strongest negative influence as it is arguably the "*poorest*" housing type. Proximity to transit stations also causes a negative impact on privately owned mobility, with the effect being less impactful as distance increases.

Finally, our hypothesis of taxi ownership having a strong negative influence on private vehicle ownership is also confirmed.

Table 4. Estimation results from the ordinal logit (OL) model on the training dataset [1, 2].

Parameters	Coefficient (Log-odds)	t-stat	Odds
Thresholds (β_k)			
CAT_0 | *CAT_1*	1.622***	169.12	-
CAT_1 | *CAT_2*	1.922***	199.33	-
CAT_2 | *CAT_3*	2.063***	213.39	-
CAT_3 | *CAT_4*	5.812***	470.34	-
CAT_4 | *CAT_5*	6.090***	481.73	-
Independent variables (β)			
No. of children in HH	0.273***	95.86	1.314
No. of seniors in HH	0.075***	16.73	1.078
No. of citizens and permanent residents in HH	0.405***	121.57	1.500
No. of males in HH	-0.160***	-44.86	0.852
No. of HH members with car licenses	0.999***	211.47	2.716
No. of HH members with motorcycle licenses	0.399***	66.29	1.491
No. of HH members with no licenses	-0.268***	-69.80	0.765
No. of workers in HH	-0.177***	-42.83	0.837
No. of retirees in HH	-0.127***	-19.31	0.881
No. of young professionals in HH	-0.122***	-31.27	0.885
No. of white collar employees in HH	0.163***	44.35	1.177
No. of blue collar employees in HH	0.095***	19.36	1.099
HDB with one room	-1.466***	-48.31	0.231
HDB with two or more rooms	-0.637***	-107.97	0.529
Privately owned house	0.856***	117.84	2.354
HH of Malaysian ethnicity	-0.040***	-4.64	0.961
HH of Indian ethnicity	-0.637***	-84.11	0.529
House within 500 meters of MRT station	-0.218***	-34.67	0.804
House within 500 to 1,000 meters of MRT station	-0.105***	-18.56	0.900
HH income (divided by 10,000)	0.806***	129.92	2.239
HH owns a taxi	-1.211***	-66.03	0.298

[1] Parameter is statistically significant at 90% (*), 95% (**), or 99% (***) confidence level.

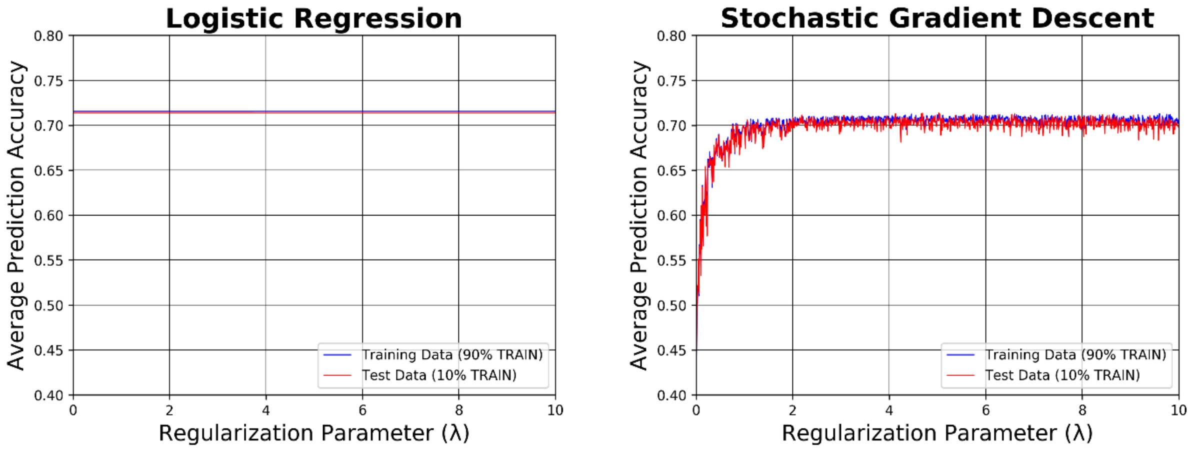
[2] ***CAT_0:*** No vehicle; ***CAT_1:*** 1+ motorcycle only; ***CAT_2:*** 1 off-peak car w/wo motorcycle; ***CAT_3:*** 1 normal car only; ***CAT_4:*** 1 normal car and 1+ motorcycle; ***CAT_5:*** 2+ normal cars w/wo motorcycle

We notice positive influences on private vehicle ownership caused by the number of children, the number of seniors, and the number of Singapore citizens (SCs) and permanent residents (PRs) in the household. Having more children requires more travel over a larger spread of destinations, which inherently requires increased mobility options for better accessibility. Seniors are more likely to have higher wealth (note the difference between wealth and income in this context) as well as diverse mobility needs, leading to a larger probability of being higher on the mobility scale. The intensively regulated atmosphere of Singapore makes it difficult for individuals who are not SCs or PRs to purchase a vehicle. In alignment with our previous discussion regarding licenses, we see that households with individuals having car and/or motorcycle licenses are more likely to be higher up on the mobility scale, which is an induced effect of our construction of the ordinal mobility scale. Regular workers, such as those employed in white-collar and blue-collar jobs, are more likely to own private vehicles. Finally, we notice very strong income effects on privately-owned mobility, both through the direct effect of household income and the indirect effect of privately-owned housing.

In summary, we see that our results from the MNL and OL models agree with each other, and the OL model results help quantify the jump in probabilities caused by an incremental change in the covariate over the ordinal mobility scale. We do not provide any discussion here regarding the threshold values as they require further elaboration on the latent variable formulation of the OL model. For brevity, they can be interpreted as the threshold values of utilities associated with each alternative. One important observation here is the utility associated with owning a private car (*CAT_3*) is almost three times that of the utility associated with the next lower alternative.

4.3. Cross-validation results

Figure 1 shows the average prediction accuracies (averaged by 10-fold cross-validation) for the LR and SGD models using these two learning algorithms as a function of the regularization parameter λ that can take values across the specified grid of [0.01,10]. 90% of the overall training dataset (denoted as TRAIN) was used as the training dataset and the remaining 10% of TRAIN was used as the test dataset in each of these 10 cross-validation iterations. The selection of these datasets was conducted randomly with replacement. We see that SGD converges slower than LR due to its inherent nature of selecting a sample at random for computing the gradient. LR, on the other hand, exhibits much more stable behavior. As mentioned earlier, the idea is to look at the test error and select the parameter that results in the lowest test error. Accordingly, we selected $\lambda^* = 2$ for LR and $\lambda^* = 4$ for SGD. It is also pertinent to mention that cross-validation time for SGD was 57 minutes as compared to 32 minutes for LR.



(a) **(b)**

Figure 1. Cross-validation results for (a) Logistic Regression, and (b) Stochastic Gradient Descent.

4.4. Model comparison results

While comparing the performance of different modeling approaches, it is necessary to consider a baseline model. We used the zero-information, also known as the zero-R, method to construct the baseline. This method assumes a prediction of the class with the maximum frequency (*CAT_0* in this application) to all data samples. Table 5 provides a comparison of model performance metrics between the baseline, the two econometric models and the seven machine learning models. Note that all performance metrics are calculated using sampling weights and the test dataset, except for the average training accuracy measure which used the weighted training dataset.

While accuracy is a good overall measure, it can be critiqued because it assumes equal costs for both types of errors (false positives and false negatives). Therefore, we should consider other measures such as precision, recall and F-measure. Recall is defined as the ratio of the total number of correctly classified positive examples and the total number of positive examples. High recall indicates the class is correctly recognized (small number of false negatives).

$$Recall = \frac{\#(True\ Positive)}{\#(True\ Positive) + \#(False\ Negative)} \tag{13}$$

Precision is the ratio of the total number of correctly classified positive examples and the total number of predicted positive examples. High precision indicates an example labeled as positive is indeed positive (small number of false positives).

$$Precision = \frac{\#(True\ Positive)}{\#(True\ Positive) + \#(False\ Positive)} \tag{14}$$

Since we now have two separate measures (precision and recall), it helps to have a measurement that represents both of them. We calculate an F-measure which uses harmonic mean in place of arithmetic mean as the harmonic mean punishes the extreme values more than the arithmetic mean.

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{15}$$

In addition to these measures, we also include a measure relevant to econometric models known as McFadden's pseudo R-squared. It compares a model with predictors (M_{Full}) to a model without predictors $\left(M_{Intercept}\right)$, i.e. using only the intercepts as explanatory variables) and uses the log-likelihood values of these two models. The ratio of the likelihoods suggests the level of improvement over the intercept model offered by the full model. Thus, a small ratio of log-likelihoods indicates that the full model is a far better fit than the intercept model. If comparing two models on the same data, McFadden's pseudo R-squared would be higher for the model with the greater likelihood.

$$\rho^2 = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}\left(M_{Intercept}\right)} \tag{16}$$

We can see that all the models tested in this application perform significantly better than the baseline in terms of prediction accuracy and robustness. MNL performs reasonably well and McFadden's R-squared indicates a good model fit (according to literature). Although having a slightly worse model fit, OL has a significantly improved accuracy score for both the train and test datasets. However, all the machine learning models outperform MNL with regard to all the performance metrics.

Table 5. Comparison of model performance metrics.

Metric	Baseline	Econometric Models		Machine Learning Models						
		MNL	OL	DT	RF	NN	SVM	LR	SGD	OLC_NN
Average train accuracy (%)	54.26	64.92	70.54	99.67	98.35	76.21	86.91	70.64	70.65	82.48
Average test accuracy (%)	55.47	63.34	70.03	64.37	70.74	73.12	66.89	71.94	72.28	70.78
Execution time (sec)	-	27.32	0.80	0.38	0.53	3.49	8.15	0.27	0.50	24.56
Average precision	0.31	0.64	0.66	0.64	0.68	0.70	0.69	0.68	0.70	0.69
Average recall	0.55	0.63	0.71	0.64	0.71	0.73	0.67	0.72	0.72	0.71
Average F-measure	0.40	0.64	0.69	0.64	0.69	0.71	0.68	0.69	0.69	0.69
McFadden's R-squared	-	0.36	0.28	-	-	-	-	-	-	-

DT has close to 100% accuracy on the training dataset because an over-complex model was created leading to overfitting and comparatively lower test accuracy. RF, being an extension of DT, has marginally lower training accuracy but performs significantly better on the test dataset, indicating much better generalizability. SVM, using non-linear kernels, performs quite well on the training dataset but fails to generalize as well. The performance of LR and

SGD are better than SVM but almost similar to each other. We notice that the best machine learning model, in terms of test accuracy, precision, recall and F-measure, is NN. Despite not having the highest training accuracy, it was the most robust and generalized best to the test dataset. Therefore, we used NNs in our OLC implementation, which we denote henceforth as OLC_NN. This model performs quite well on the training dataset, and marginally better than OL on the test dataset.

Table 6 shows the weighted aggregate shares of each class of the target variable using these models on the test dataset. We see that both MNL and DT are successful in creating a distribution close to the true distribution. While SVM performs decently, we notice that RF, NN, LR and SGD exhibit complications predicting classes with low sample sizes. Since the algorithms are trying to minimize overall prediction error as opposed to prediction error for each class, these machine learning models are treating low sample classes as outliers or noise and misclassifying them as high sample classes. This scenario is known as an imbalanced classification problem, and requires further investigation. While OL produces a biased prediction wherein it could not correctly predict even a single sample in the low-sample classes, OLC_NN is much better at maintaining a reasonable distribution. While it could not predict any sample in *CAT_4*, it should be noted that *CAT_4* has an extremely low sample size (0.89%) which is perhaps too low for training a predictive model.

We use two types of errors as performance metrics to compare the predicted market shares with the actual market shares. The first is the mean absolute error (MAE), which measures the average magnitude of the errors in a set of predictions without considering their direction.

$$MAE = \frac{\sum_{j=1}^{J}\left|y_j - \widehat{y_j}\right|}{J} \qquad (17)$$

The second type of error is the root mean squared error (RMSE), which is the square root of the average of squared differences between the predictions and actual observations. Note that the RMSE is particularly useful when large errors are particularly undesirable (such as skewed predicted market shares in our application) because large errors are given a relatively higher weight.

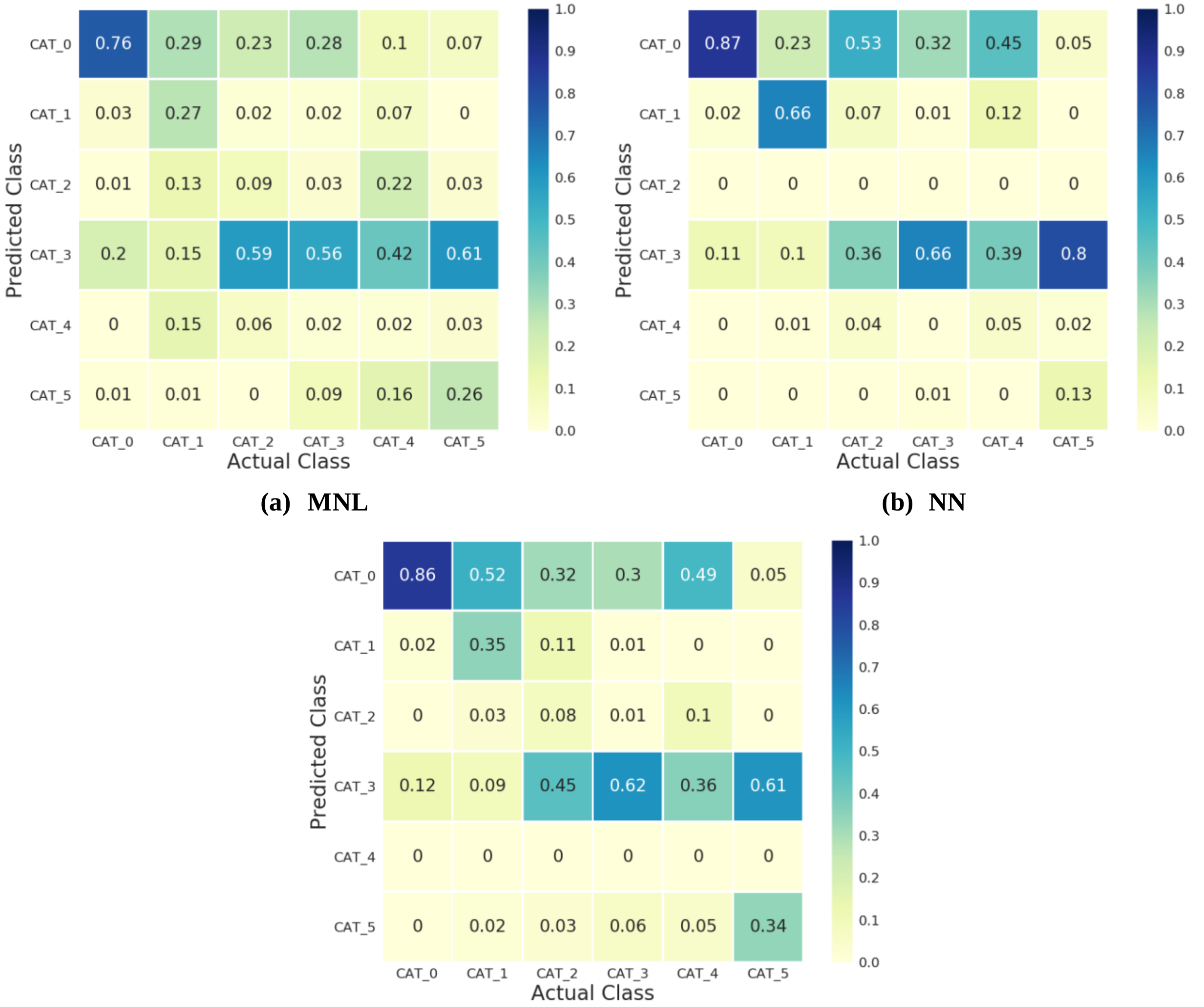
$$RMSE = \sqrt{\frac{\sum_{j=1}^{J}\left(y_j - \widehat{y_j}\right)^2}{J}} \qquad (18)$$

Table 6. Comparison of actual and predicted weighted aggregate shares (in %) of target variable classes in the test dataset [1].

Class	Actual	Baseline	Econometric Models		Machine Learning Models						
			MNL	OL	DT	RF	NN	SVM	LR	SGD	OLC_NN
CAT_0	55.47	100.00	53.68	60.45	57.07	62.85	61.38	48.78	58.31	62.50	60.80
CAT_1	4.45	-	3.96	-	3.72	3.44	4.64	5.90	1.28	1.72	3.08
CAT_2	2.20	-	2.39	-	1.96	0.42	-	1.84	-	-	1.04
CAT_3	32.64	-	34.34	37.48	31.96	31.65	32.61	36.20	38.32	35.37	31.17
CAT_4	0.89	-	1.09	-	0.94	0.33	0.38	0.81	-	0.10	-
CAT_5	4.33	-	4.53	2.07	4.35	1.31	0.99	6.46	2.09	0.30	3.91
Performance Metrics											
MAE	-	14.84	0.76	3.27	0.55	2.46	2.03	2.38	2.84	3.25	1.77
RMSE	-	22.70	1.04	3.62	0.78	3.39	2.92	3.27	3.19	3.79	2.41

[1] ***CAT_0:*** No vehicle; ***CAT_1:*** 1+ motorcycle only; ***CAT_2:*** 1 off-peak car w/wo motorcycle; ***CAT_3:*** 1 normal car only; ***CAT_4:*** 1 normal car and 1+ motorcycle; ***CAT_5:*** 2+ normal cars w/wo motorcycle

We now examine the confusion matrices for MNL, NN and OLC_NN to better understand the imbalanced classification phenomenon. The actual classes are placed on the x-axis, while the predicted classes are placed on the y-axis such that each cell is the proportion of samples belonging to the actual class classified as the predicted class and each column sum equals to one. Indeed, classes with large samples (*CAT_0* and *CAT_3*) are being predicted quite frequently even when the actual classes are different. On the other hand, low-sample classes (*CAT_2* and *CAT_4*, in particular) fail to maintain even a good self-prediction, as seen from the major diagonal values for those classes in Figure 2. This shows that our hypothesis of the imbalanced classification problem is true. While there are techniques to address this problem, they are beyond the scope of this research and we provide comments about them in the following section. Our examination of the three confusion matrices leads to the conclusion that OLC_NN performs the best in terms of individual class prediction accuracy, followed by NN and MNL respectively. Considering all performance metrics from Table 5, Table 6, and Figure 2, it is clear that OLC_NN is the best modeling approach among all the nine models tested in this application.



(a) MNL

	CAT_0	CAT_1	CAT_2	CAT_3	CAT_4	CAT_5
CAT_0	0.76	0.29	0.23	0.28	0.1	0.07
CAT_1	0.03	0.27	0.02	0.02	0.07	0
CAT_2	0.01	0.13	0.09	0.03	0.22	0.03
CAT_3	0.2	0.15	0.59	0.56	0.42	0.61
CAT_4	0	0.15	0.06	0.02	0.02	0.03
CAT_5	0.01	0.01	0	0.09	0.16	0.26

(b) NN

	CAT_0	CAT_1	CAT_2	CAT_3	CAT_4	CAT_5
CAT_0	0.87	0.23	0.53	0.32	0.45	0.05
CAT_1	0.02	0.66	0.07	0.01	0.12	0
CAT_2	0	0	0	0	0	0
CAT_3	0.11	0.1	0.36	0.66	0.39	0.8
CAT_4	0	0.01	0.04	0	0.05	0.02
CAT_5	0	0	0	0.01	0	0.13

(c) OLC_NN

	CAT_0	CAT_1	CAT_2	CAT_3	CAT_4	CAT_5
CAT_0	0.86	0.52	0.32	0.3	0.49	0.05
CAT_1	0.02	0.35	0.11	0.01	0	0
CAT_2	0	0.03	0.08	0.01	0.1	0
CAT_3	0.12	0.09	0.45	0.62	0.36	0.61
CAT_4	0	0	0	0	0	0
CAT_5	0	0.02	0.03	0.06	0.05	0.34

Figure 2. Confusion matrices for the (a) Multinomial logit, (b) Neural network, and (c) Ordinal logit classification with neural network model predictions on the test data set.

5. Conclusion & Future Work

The growing usage of vehicles by households and the increasing diversity of vehicle options (due to disruptive technologies and services) have serious policy implications for traffic congestion and air pollution. Consequently, it is important to accurately predict the vehicle holdings of households as well as the vehicle miles of travel by vehicle type to project future traffic congestion and mobile source emission levels. Therefore, we attempted to compare supervised learning algorithms with traditional econometric approaches in order to find a model that performs the best in predicting household vehicle ownership. Using a socio-demographic dataset from Singapore, 22 feature vectors were constructed using appropriate transformations and imputing missing data to predict a 6-class categorical ordinal variable. The neural network model was found to be the most robust and performed the best while generalizing to the test dataset. However, the dataset had some dominant high-sample classes that resulted in an imbalanced classification problem wherein the low-sample classes had high prediction errors due to those instances being treated as outliers. The ordinal logit classification using neural network binary classifiers was able to address this issue, albeit to a limited extent, while maintaining an almost similar predictive accuracy score. Techniques to appropriately handle such problems (such as oversampling low sample classes, undersampling high sample classes, combining oversampling and undersampling, or using algorithms to create synthetic samples based on clustering) should be explored in order to better improve prediction accuracy and generalizability.

The econometric models provided interesting insights into the household vehicle ownership decision-making process. Singapore's excellent public transport system, and strict regulatory practices regarding purchasing a car and obtaining a driving license influenced not owning any vehicle to be the most preferred alternative. A gender effect was also revealed, wherein households with a higher number of males are less inclined to own any vehicle. Households with a higher number of workers are strongly disinclined towards owning cars. A strong indirect income effect through housing type (government-provided or private-owned) and job sector (white-collar or blue-collar) was also evident through ownership of motorcycles for low-income households and cars for high-income households. Additionally, the direct income effect was statistically significant and strongly positive in magnitude. An attitudinal aspect was noticed in households with young professionals, wherein they are strongly disinclined to own a car. Proximity to transit and taxi ownership were found to be significant factors in influencing vehicle ownership negatively.

We hypothesize that taxi ownership and vehicle ownership in Singapore are intricately linked and there is a conditional dependence between the two variables. A natural extension would be to try the nested logit model as an additional econometric approach. Since the HITS survey does not provide data that can differentiate between the two nests, using only socio-demographic data would not be enough for estimating such a model. This is another area where machine learning can come in handy, wherein we can employ multi-label classification algorithms since each instance has two labels (one for taxi ownership and one for vehicle ownership). This research paves the way for an integrated framework that incorporates both the econometric and supervised learning approaches in order to better predict the influence of disruptive changes.

Acknowledgements

This research was supported in part by the National Research Foundation (NRF) Singapore through the Singapore MIT Alliance for Research and Technology's Future Urban Mobility IRG research program. In addition, we appreciate the support of the Singapore Land Authority (SLA) and Singapore Land Transport Authority (LTA) in providing data and other helpful information.

References

Anyanwu, M. N., & Shiva, S. G. 2009. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security, 3(3), 230-240.

Basu, R., Araldo, A., Akkinepally, A. P., Nahmias Biran, B. H., Basak, K., Seshadri, R., ... & Ben-Akiva, M. 2018. Automated Mobility-on-Demand vs. Mass Transit: A Multi-Modal Activity-Driven Agent-Based Simulation Approach. Transportation Research Record.

Baum, C. L. 2009. The effects of vehicle ownership on employment. Journal of Urban Economics, 66(3), 151-163.

Bhat, C. R., & Sen, S. 2006. Household vehicle type holdings and usage: an application of the multiple discrete-continuous extreme value (MDCEV) model. Transportation Research Part B: Methodological, 40(1), 35-53.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT 2010 (pp. 177-186). Physica-Verlag HD.

Breiman, L. 2001. Random forests. Machine learning, 45(1), 5-32.

Button, K., Ngoe, N., & Hine, J. 1993. Modelling vehicle ownership and use in low income countries. Journal of Transport Economics and Policy, 51-67.

Caruana, R., & Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

Choudhary, R., & Vasudevan, V. 2017. Study of vehicle ownership for urban and rural households in India. Journal of Transport Geography, 58, 52-58.

Clark, S. D. 2009. The determinants of car ownership in England and Wales from anonymous 2001 census data. Transportation research part C: emerging technologies, 17(5), 526-540.

Danielis, R., Giansoldati, M., & Rotaris, L. 2018. A probabilistic total cost of ownership model to evaluate the current and future prospects of electric cars uptake in Italy. Energy Policy, 119, 268-281.

Dargay, J. M. 2002. Determinants of car ownership in rural and urban areas: a pseudo-panel analysis. Transportation Research Part E: Logistics and Transportation Review, 38(5), 351-366.

Dargay, J., & Gately, D. 1999. Income's effect on car and vehicle ownership, worldwide: 1960–2015. Transportation Research Part A: Policy and Practice, 33(2), 101-138.

Dargay, J., Gately, D., & Sommer, M. 2007. Vehicle ownership and income growth, worldwide: 1960-2030. The Energy Journal, 143-170.

Dissanayake, D., & Morikawa, T. 2010. Investigating household vehicle ownership, mode choice and trip sharing decisions using a combined revealed preference/stated preference Nested Logit model: case study in Bangkok Metropolitan Region. Journal of Transport Geography, 18(3), 402-410.

Flamm, B. 2009. The impacts of environmental knowledge and attitudes on vehicle ownership and use. Transportation research part D: transport and environment, 14(4), 272-279.

Greene, W. H., & Hensher, D. A. 2010. Modeling ordered choices: A primer. Cambridge University Press.

Guerra, E. 2015. The geography of car ownership in Mexico City: a joint model of households' residential location and car ownership decisions. Journal of Transport Geography, 43, 171-180.

Haykin, S. S. 2009. Neural networks and learning machines (Vol. 3). Upper Saddle River, NJ, USA. Pearson.

Hsu, C. W., & Lin, C. J. 2002. A comparison of methods for multiclass support vector machines. IEEE transactions on Neural Networks, 13(2), 415-425.

Javid, R. J., & Nejat, A. 2017. A comprehensive model of regional electric vehicle adoption and penetration. Transport Policy, 54, 30-42.

Kim, H. S., & Kim, E. 2004. Effects of public transit on automobile ownership and use in households of the USA. Review of Urban & Regional Development Studies, 16(3), 245-262.

Kim, D., Ko, J., & Park, Y. 2015. Factors affecting electric vehicle sharing program participants' attitudes about car ownership and program participation. Transportation Research Part D: Transport and Environment, 36, 96-106.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

Law, T. H., Hamid, H., & Goh, C. N. 2015. The motorcycle to passenger car ownership ratio and economic growth: a cross-country analysis. Journal of Transport Geography, 46, 122-128.

Liu, Y., Tremblay, J. M., & Cirillo, C. 2014. An integrated model for discrete and continuous decisions with application to vehicle ownership, type and usage choices. Transportation Research Part A: Policy and Practice, 69, 315-328.

Macfarlane, G. S., Garrow, L. A., & Mokhtarian, P. L. 2015. The influences of past and present residential locations on vehicle ownership decisions. Transportation research part A: policy and practice, 74, 186-200.

Manski, C. F. 1977. The structure of random utility models. Theory and Decision, 8(3), 229-254.

Oakil, A. T. M., Manting, D., & Nijland, H. 2016. Determinants of car ownership among young households in the Netherlands: The role of urbanisation and demographic and economic characteristics. Journal of Transport Geography, 51, 229-235.

Potoglou, D., & Susilo, Y. 2008. Comparison of vehicle-ownership models. Transportation Research Record: Journal of the Transportation Research Board, (2076), 97-105.

Ryan, J., & Han, G. 1999. Vehicle-ownership model using family structure and accessibility application to Honolulu, Hawaii. Transportation Research Record: Journal of the Transportation Research Board, (1676), 1-10.

Schimek, P. 1996. Household motor vehicle ownership and use: How much does residential density matter?. Transportation Research Record: Journal of the Transportation Research Board, (1552), 120-125.

Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. 2011. Pegasos: Primal estimated sub-gradient solver for svm. Mathematical programming, 127(1), 3-30.

Wu, G., Yamamoto, T., & Kitamura, R. 1999. Vehicle ownership model that incorporates the causal structure underlying attitudes toward vehicle ownership. Transportation Research Record: Journal of the Transportation Research Board, (1676), 61-67.

Wu, N., Zhao, S., & Zhang, Q. 2016. A study on the determinants of private car ownership in China: Findings from the panel data. Transportation Research Part A: Policy and Practice, 85, 186-195.

Zegras, C. 2010. The built environment and motor vehicle ownership and use: Evidence from Santiago de Chile. Urban Studies, 47(8), 1793-1817.

Zhang, Z., Jin, W., Jiang, H., Xie, Q., Shen, W., & Han, W. 2017. Modeling heterogeneous vehicle ownership in China: A case study based on the Chinese national survey. Transport Policy, 54, 11-20.