Australasian Transport Research Forum 2019 Proceedings
30 September - 2 October, Canberra, Australia
Publication Website: http://www.atrf.info

# Modelling mode choice with machine learning algorithms

Matthew John Richards[1,2], Jan Christoph Zill[2]

[1] School of Mathematics and Physics, The University of Queensland, Brisbane 4072, Australia

[2] Veitch Lister Consulting, 200 Mary Street, Brisbane 4000, Australia

Email for correspondence: jan.zill@veitchlister.com.au

## Abstract

This paper examines several machine learning methods to model mode choice decisions in the greater Melbourne area, based on the Victorian Integrated Survey of Travel and Activity (VISTA), a revealed-preference household travel survey. We compare the results to a well-calibrated nested logit model traditionally used in strategic transport models. We perform this comparison in two different ways. First, we compare all models trained on the same set of input features, as determined by constructing the discrete choice model. Second, we also compare the same nested logit model as before to machine learning models trained on the entire set of available features in the VISTA. We find that certain machine learning models consistently outperform the discrete choice model, but prediction accuracy is considerably better for models trained on all features. We therefore also investigate the interpretability of the best-performing machine learning models and investigate their sensitivity to selected features.

## 1 Introduction

The problem of mode choice determination is an important component of transportation modelling and forecasting (Dios Ortuzar and Willumsen 2011). Traditionally, mode choice models are econometric discrete choice models. These are based on the behavioural principle of random utility maximisation and they are therefore readily interpretable (Ben-Akiva, Lerman, and Lerman 1985; Hensher, Rose, and Greene 2005; Train 2009). However, discrete choice models usually require extensive work to specify and estimate. Often, a travel market is segmented by some explanatory variables, e.g. travel purpose, and then for each of those segments, a different functional form of the observed part of the utility is specified, depending on the fit to the data. The unobserved part of the utility has a fixed statistical distribution imposed (Ben-Akiva, Lerman, and Lerman 1985; Hensher, Rose, and Greene 2005; Train 2009). Additionally, random utility models are estimated on a per-decision maker case, meaning careful consideration of biases in the data are necessary (Dios Ortuzar and Willumsen 2011).

Machine learning (ML) techniques are fundamentally different in that they generally do not make assumptions regarding the structure of the data, but rather develop a notion of underlying structure through the fitting procedure. As such, they do not require the input data to fulfil such strong statistical assumptions. Their capacity to deal with binomial input features allows a single model to be trained, accounting for categories which would be typically segmented over for logit models, avoiding the time consuming iterative procedure of parameter estimation. While many machine learning models do

have tunable hyper-parameters, these can be optimised over in an automated fashion using a grid or random search. Consequently, ML techniques have started to appear in mode choice modelling research in recent years (C. Xie, Lu, and Parkany 2003; Zhang and Y. Xie 2008; Omrani et al. 2013; Omrani 2015; Hagenauer and Helbich 2017; Ma, Chow, and Xu 2017; Lee, Derrible, and Pereira 2018; Wang and Ross 2018; Cheng et al. 2019; Zhao et al. 2018). These works have generally found that ML outperforms multi-nomial logit (MNL) models in terms of out-of-sample prediction accuracy. However, mode choice models with detailed mode break-downs often display violation of the independence of irrelevant alternatives due to alternatives with shared unobserved attributes (Ben-Akiva, Lerman, and Lerman 1985). Therefore, nested logit (NL) models are generally better suited to mode choice modelling (Ben-Akiva, Lerman, and Lerman 1985; Dios Ortuzar and Willumsen 2011). Subsequently we compare several ML classifiers with a well-calibrated NL model (Veitch Lister Consulting 2014). Informed by the existing literature, we look to investigate logistic regression, the random forest, gradient boosting machines and neural networks.

The comparison of ML models to the NL model is performed in two different ways, motivated by the fact that one of the most important use-cases of strategic transport models is forecasting future travel demand. As such, variables that enter the model have to be available in future years and it is very important that the model is interpretable. In light of this, we compare the performance of ML algorithms with our NL model with similar input variables. We then compare the performance of the ML models using additional variables drawn from the full data set. We also investigate the interpretability of the best performing ML model, a gradient-boosted decision tree, in more detail and provide results for variable importance and partial dependence plots.

The rest of this paper is structured as follows. We first succinctly summarize the ML classifiers used in this work. We then give more detail on the discrete choice model used for comparison, before examining the data used for training and validation of our models. Next, we compare the results of the ML models to the NL model for identical input features. We contrast this to the performance of the ML models trained on all available input features, before inspecting the interpretability of the models. Finally, we present our conclusions.

# 2 Methodology

## 2.1 Machine learning classifiers

We present a brief but necessary discussion of the construction of the various machine learning methods presented. The mathematical specifics of the following classifiers are well documented and further details can readily be sought. All of the machine learning models are developed using the python package scikit-learn (Pedregosa et al. 2011) with the exception of the neural network, which uses the keras package (Chollet et al. 2015) via the tensorflow (Martín Abadi et al. 2015) implementation. The hyperparameters included below for each model were determined using cross validated grid searches over the range of feasible parameter inputs.

### *2.1.1 Logistic regression*

In its simplest form, logistic regression models a binary decision by partitioning the input space using a logistic curve, yielding a probability for each outcome. A multinomial

generalisation is used to extend to the multi-class outcome space whilst still retaining the output interpretation as probabilities. Logistic regression is fit using the procedure of maximum likelihood estimation, similar to discrete choice models, where in particular the SAGA optimisation algorithm (Defazio, Bach, and Lacoste-Julien 2014) is used. The models were constructed in the multi class sense rather than the one versus rest scheme and did not use class weightings. In order to improve convergence of the component iterative procedures, the input features are standardised with mean zero and variance one.

### *2.1.2 Neural network*

Neural networks consist of repeated layers of interconnected activation nodes, also referred to as neurons. The transformation of input features through hidden layers allows for complex interaction effects and structure to be effectively modelled. The training of weights in each layer is achieved through gradient descent using the back-propagation algorithm, or variations thereof (Yann, Bengio, and Hinton 2015). A neural network model using the ADAM optimisation algorithm (Kingma and Ba 2014) was trained with 4 hidden layers each consisting of 64 neurons. The binary cross entropy loss function and 100 training epochs were used. Finally, a batch size of 128 and 156 were used for the baseline and full feature sets respectively. In the same way as for logistic regression, the input features were standardised in order to improve results.

### *2.1.3 Random forest*

The random forest is an ensemble method which improves the performance of the constituent decision tree models. Decision trees recursively partition the input space into smaller regions using binary partitions of features in the input space. Eventually a terminal region is reached and the corresponding predictions are classified using a constant function for that particular segment or through sampling of the component observations. The shortcoming of such models is the difficulty of constructing a tree of the correct size such that it does not over-fit or under-fit the data (Hastie, Tibshirani, and Friedman 2001, pp 307-308).

The random forest addresses this by taking a collection of tree based classifiers, each trained on a random, independent sub-sample of input features. The prediction for any given input vector is the classification most voted for by the component estimators. These models are robust despite the potentially large number of underlying models within the ensemble (Breiman 2001). After hyper-parameter optimisation, the best performing model used 500 estimators with unconstrained maximum depth and terminal node size but requiring at least two observations to split an internal node.

### *2.1.4 Gradient boosting*

While also an ensemble tree method, gradient boosting instead uses a sequential rather than parallel approach to connect the component trees. Decision trees of relatively small size are fitted successively, where the next model attempts to minimise the loss function of the previously established classifier. A weighted average of all trained classifiers is used to determine the output prediction. Gradient boosting methods are found to perform particularly well in many contexts and produce interpretable results, although there is a time cost associated with this Hastie, Tibshirani, and Friedman 2001, p. 352. The trained models were determined to use the Friedman mean

squared error, learning rate of 0.1, logarithmic scaling of maximum features and 300 estimators.

## 2.2 Discrete choice model

Discrete choice models are heavily used in transportation modelling (Ben-Akiva, Lerman, and Lerman 1985; Hensher, Rose, and Greene 2005; Train 2009; Dios Ortuzar and Willumsen 2011). They can be derived from random utility theory and are therefore readily interpretable. However, they rely on the modeller specifying the functional form of the observed part of the utility and force the unobserved part of the utility to have a specific distribution. This means the specification and estimation is a rather manual process and usually involves several iterations.

The model we use here is Veitch Lister Consulting's Zenith model for Victoria (Veitch Lister Consulting 2014). It is a nested logit model with 38 segments (8 home-based trip purposes, each segmented by 0, 1, 2, 3+ cars per household, and six non-home based purposes), each with its own specification of the observed utility function. The utility functions are linear-in-parameters and contain costs for each mode (essentially travel time for car, bike, walk and a generalised cost consisting, among others, of travel time, wait time, and fare cost for public transport), and spatial constants for some aggregate regions. For example, the city centre has a spatial constant to account for parking costs for car. The considered modes for public transport include access and egress modes of walk, park, and kiss and ride. Parameter estimates were obtained by maximum-likelihood estimation with the biogeme package (Bierlaire 2016). For detailed specifications and validation see (Veitch Lister Consulting 2014).

To compare this model to the ML models, we micro-simulate the results of each trip in the VISTA survey, i.e. we draw a random error term for each observed trip and add it to the observed utility, then choose the alternative with maximum utility. The results are obtained in sample, i.e. full data was used for both training and testing. This very likely overestimates the accuracy of the discrete choice model. However, due to the afore-mentioned heavily restricted statistical structure and the very low number of variables, discrete choice models do not easily overfit data and therefore we consider the comparison appropriate.

## 2.3 Data

The VISTA dataset is comprised of trips recorded by randomly selected households who complete a travel diary recording all trips completed on a particular day (Victoria State Government Department of Transport 2018). The survey participants are drawn from the greater Melbourne region, including households in Ballarat, Bendigo and Geelong. An aggregation of the VISTA data from 2007-2013 is used, however trips occurring on weekends or during school holidays are excluded in order to model a standard week day. The resultant dataset comprises slightly under 150000 trips. Each entry contains details pertaining to the individual making the trip and information about their household which is summarised in Table 1. It must be noted that that these trips are not necessarily an unbiased representation of the true population and trips taken in the Greater Melbourne region; in practice, a re-weighting of predictions made informed by census population estimates is performed for the estimation of parameters in the nested logit model (Veitch Lister Consulting 2014). However, for the ML algo-

rithms considered here, this paper is only concerned with accuracy determined relative to this dataset, in line with common application.

The VISTA data is additionally supplemented with mode dependent travel times produced by Veitch Lister's Zenith traffic model. For public transport, the generalised cost, incorporating travel time, wait time and fare costs among others, is used instead of travel time. Additionally, the zones are grouped into aggregate regions (CBD, Melbourne, Geelong and Regional) to include as geographically distinct categorical features, rather than using the arbitrary SA zoning regions.

The models are constructed to provide a single mode prediction for each trip from the aggregate categories of bicycle, car, public transport and walking. There are very few trips made by bike, comprising just under 1.5% of all trips, however this alternative is preserved to demonstrate the difficulty of working with extreme minority classes. The aggregation of bus, train and tram results in a public transport class comprising 7% of trips, which is still a relatively small segment. The vast majority of all trips are by car, making up 80%, with the remaining by foot. It should be noted that the category of car refers to any private vehicle, including motorcycles and utility vehicles. For any trips containing changes of mode, the predominant mode is included in the VISTA survey and this label is used.

**Table 1: Summary of features included in the VISTA dataset**

| Trip Details | Description | Person Details | Description |
|---|---|---|---|
| Departure/ arrival time | Times at which trip commences and finishes | Occupation | Occupation of person if employed |
| Orig/ dest. region | SA indicator of trip origin and destination | Study Type | Type of pri./sec./tertiary/ study completed by person if any |
| Orig/ dest. activity | Activity traveller undertook at trip origin and dest. | Full/ Part Time | Whether study/ occupation is full or part time |
| Purpose | Categorical purpose assigned to trip | Unemployed | |
| Distance | The distance travelled (km) | Gender | Male or female |
| Travel mode(s) | Mode or list of modes taken to complete trip, majority mode is specified | Car/ Motorbike Licence | Whether traveller has full/ learner/ probationary licence |
| | | Age | Age of traveller (yrs) |
| | | Relationship | Traveller is child, single, married, parent, grandparent |
| **Household Details** | **Description** | **Household Details** | **Description** |
| Dwelling type | Dwelling is house/ townhouse/ apartment | # Workers | (includes part time workers) |
| Dwelling ownership | House is owned/ rented/ mortgaged/ other | # Blue/ White Collar Workers | |
| Home location | SA region enclosing home | # Studying/ at school | |
| Household income quintile | Categorical encoding of household income | # Unemployed | |
| # Persons | Number of persons within household | # Cars | Number of cars available to household to use |
| # Dependants | | # Adult/ Child Bikes | Number of bikes available to household to use |

# 3 Results

To evaluate the machine learning models, we train two stages of models; a baseline set using equivalent features to the NL model, and secondly a dense feature set in-

cluding additional potential explanatory variables. For the baseline model, the included features are presented in Table 2. These are chosen to directly emulate the features used in the discrete choice model to provide a fair direct comparison. The second feature set makes use of all the features present in Table 1 with the exception of reported distance and arrival time to avoid any information leakage. This is intended to explore the extent to which rich feature sets benefit ML techniques.

**Table 2: Baseline features used to evaluate machine learning model performance**

| Feature | Description |
|---|---|
| Generalised cost | Alternative specific generalised cost (predominantly travel time, incorporates wait times and fares for public transport) |
| Trip purpose | Binary indicator variables corresponding to segments discussed in 2.2 |
| # Cars | Binary indicator for number of cars in household (0, 1, 2 or 3+) |
| Orig./ dest. parking region | Binary indicator variable for cost of parking zone at location; |

To evaluate performance, we note that there are two important scales on which to score the tested methods; individual and aggregate performance. Obviously these two are inherently linked, however it must be noted that the logit models are designed to emulate the overall mode share distribution (Dios Ortuzar and Willumsen 2011). In contrast, machine learning models optimise their accuracy with respect to predicting individual trips. Given this distinction, it is important to consider both metrics as both are practically important, and each methodology is intended to target one specific criterion. For individual level predictive power, the out-of-sample accuracy obtained via 10 fold cross validation is used, while the L1 norm of the predicted versus actual class counts is used for aggregate performance.

**Table 3: Mean out of sample accuracy for ML models trained on discrete choice equivalent input features**

| Model | Logistic Regression | | Neural Network | | Random Forest | | Gradient Boosting | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **Overall** | 0.8443 | 0.0045 | 0.8466 | 0.0027 | 0.8859 | 0.0035 | 0.8463 | 0.0574 |
| **Bike** | 0.0 | 0.0 | 0.0207 | 0.0067 | 0.1743 | 0.0052 | 0.1612 | 0.0190 |
| **Car** | 0.9576 | 0.0003 | 0.9516 | 0.0075 | 0.9566 | 0.0007 | 0.9283 | 0.0571 |
| **PT** | 0.4268 | 0.0012 | 0.5497 | 0.0240 | 0.5733 | 0.0029 | 0.5423 | 0.0528 |
| **Walk** | 0.4110 | 0.0017 | 0.4349 | 0.0401 | 0.6917 | 0.0032 | 0.5886 | 0.0767 |

**Table 4: Mean out of sample accuracy for ML models trained on entire selection of input features**

| Model | Logistic Regression | | Neural Network | | Random Forest | | Gradient Boosting | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **Overall** | 0.8573 | 0.0036 | 0.8661 | 0.0034 | 0.8957 | 0.0025 | 0.9002 | 0.0030 |
| **Bike** | 0.0481 | 0.0020 | 0.3420 | 0.034 | 0.1132 | 0.0054 | 0.2818 | 0.0139 |
| **Car** | 0.9564 | 0.0004 | 0.9300 | 0.0089 | 0.9720 | 0.0004 | 0.9644 | 0.0006 |
| **PT** | 0.5042 | 0.0019 | 0.6057 | 0.0242 | 0.5979 | 0.0036 | 0.6488 | 0.0061 |
| **Walk** | 0.4939 | 0.0023 | 0.6815 | 0.0260 | 0.6717 | 0.0040 | 0.7113 | 0.0051 |

**Table 5: Evaluation of ML models aggregate level prediction using the L1 norm between the predicted and true class distributions for both input feature sets**

| Feature Set | NL Comparison | | Full | |
|---|---|---|---|---|
| **Model** | **Mean** | **SD** | **Mean** | **SD** |
| **Logistic Regression** | 0.1667 | 0.0008 | 0.1298 | 0.0008 |
| **Neural Network** | 0.1318 | 0.0236 | 0.0256 | 0.0128 |
| **Random Forest** | 0.0716 | 0.0018 | 0.1032 | 0.0014 |
| **Gradient Boosting** | 0.0971 | 0.0327 | 0.0647 | 0.0016 |

**Figure 1: Confusion matrix for random forest model using the NL comparison feature set. The number within each grid is the proportion of trips predicted as the particular column mode which are actually the mode of the correspondent row. Consequently, a perfect model would have have values of one on the main diagonal and zero otherwise. It can be seen that the main diagonal is the most commonly predicted for all classes excepting bike. Over-predicting the car class is the main source of classification error. Note that the figures should be treated as indicative rather than exact proportions as these results are stochastic due to the nature of the underlying models.**

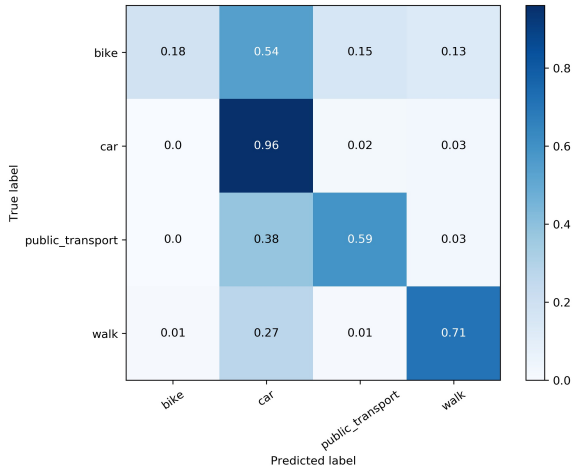



**Figure 2: Confusion matrix for the NL model evaluated using the respective input features. The behaviour is largely similar to the machine learning models; showing difficulty in dealing with the car class imbalance**

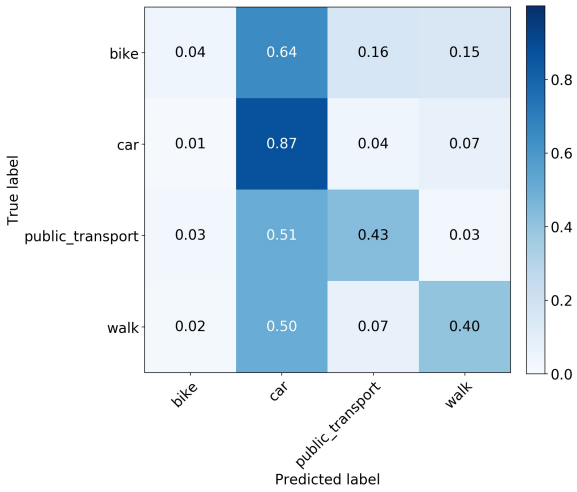
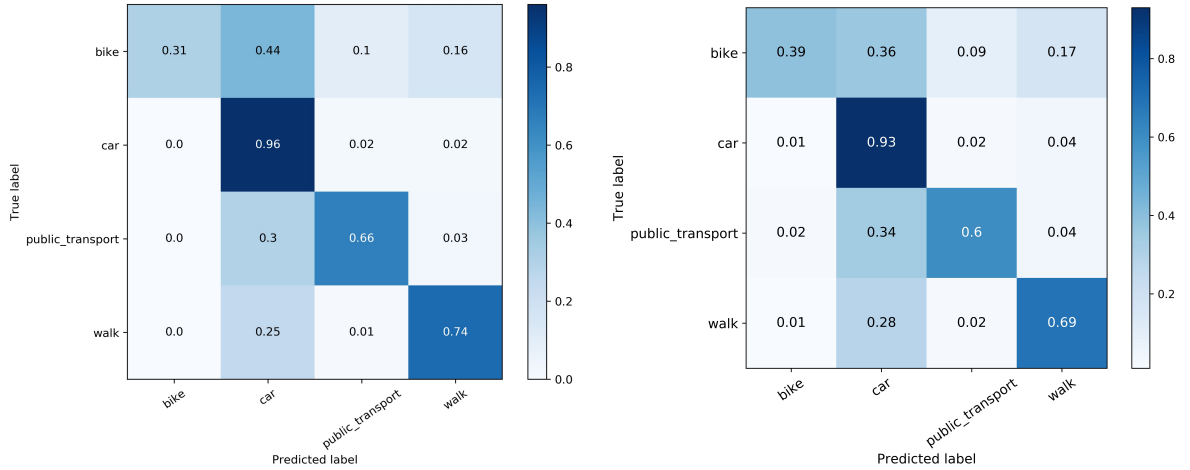


**Figure 3: Confusion matrix comparison between gradient boosting and neural network on the full feature set. Very similar behaviour is exhibited for both models, over-predicting the car mode whilst otherwise making few incorrect classifications.**




(a) Gradient boosting

(b) Neural Network

## 3.1 Comparison with same input variables

Inspecting the results for the NL model comparison in Table 3, it is clear that predicting the minority mode bike is a difficult task for all models. This is unsurprising given that there is a severe class imbalance. Given that the models optimise accuracy, there is little consequence to predicting this particular class poorly. In the reverse case, the expectation that all models predict car most accurate is met, with the mean accuracy exceeding 0.90 for every model. The logistic regression model is the best predictor of the car class, but the random forest is only marginally behind here. The forest is also most accurate for all other classes and has the highest overall model accuracy.

Public transport appears difficult to predict with only 57% of these trips correctly identified in the best case; the random forest model. This however makes sense given that other than travel time, no mode specific information has been included. Indeed, examining the random forest confusion matrix in Figure 1, it is clear that the model had difficulty distinguishing car travel from other modes, with most classification error resulting from false positives predicting car. While this is expected to an extent given it is the majority class, it may also suggest the lack of distinguishing features between modes contributes to the inaccuracy observed. Over-prediction of the car class is most impactful for the bike mode where the majority of trips are classified incorrectly. Besides the random forest, the other models all perform similarly in terms of overall accuracy of 0.84. These totals however are attained differently, with logistic regression prioritising car at the expense of other modes (particularly bike which is never predicted), while gradient boosting has higher accuracies balanced across the classes.

The most significant result here however is the performance of random forest in Figure 1 relative to the reference NL model in Figure 2. For every class, the machine learning algorithm outperforms the traditional model by a considerable margin. Given that this feature set is designed to work well for NL models it is expected that using a richer feature set will extend this performance margin.

## 3.2 Comparison of ML models with all input features

Comparing Tables 3 and 4, it is clear that all models exhibit notable performance improvements when trained on the full feature set. Logistic regression and the random forest exhibit marginal shifts of about 1%, while the neural network moves by almost 2% to 0.866. The main change is in the case of the gradient boosting model however, which has an increase of mean overall accuracy of 5.5% to 0.9002 (standard deviation 0.0030); resulting in highest overall accuracy. Additionally, it should be noted that the standard deviation in the mean has decreased by an order of magnitude from the comparison features, indicating this is a more certain result. Furthermore, the individual class scores see a corresponding increase of around 10% for the minority classes and 4% for the car class. The random forest model which performed best on the comparison features actually exhibits a decline in accuracy for the walking and cycling classes. This is largely related to the fact that the car mode predictive accuracy has increased. Given the imbalanced nature of the data this trade-off is worthwhile in terms of optimising overall accuracy. This does raise the important question of whether accuracy is the most appropriate metric to optimise model performance against. Unfortunately, this is largely decided by the context of application. In certain settings, it may be desirable to optimise an accuracy weighted by the class sizes, resulting in the better prediction of minority classes at the expense of overall accuracy.

Evaluating the aggregate level performance, as seen in Table 5, the random forest model has the smallest L1 norm deviation of 0.0716 with trained on the comparison features. This indicates that it best matches the overall mode share distribution and makes sense given that this was also the most accurate model on the individual trip level for these inputs. More notably however, is that with the full feature set, the neural network has the best performance in regards to overall mode distribution. Although this model has a significantly larger standard deviation, even in the worst case, the neural network still marginally outperforms the gradient boosting model. It is not immediately evident why the neural network does better in this regard, given that it under-performs the gradient boosting model in all classes except bike. Examining the confusion matrix in Table 3, there is little difference other than the better performance of the gradient boosting model. The majority of classification error results from over-predicting car in both cases, so it is unclear why the predictions and misclassification balance closer to the true distribution for the neural network model. The decline of the random forest aggregate performance on the full feature set is logical given the model focuses on the car class for individual level accuracy, so would expectedly suffer in terms of the aggregate distribution.

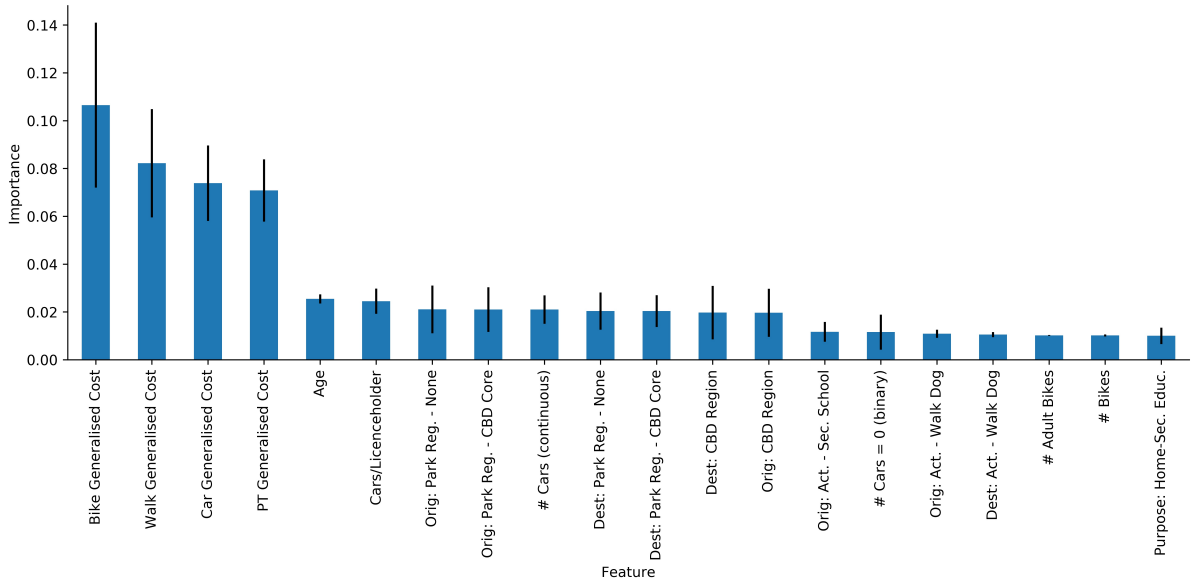
## 3.3 Predictive insights

In order to determine the value of features to the ML models, variable importance is examined. This is computed by considering permutations in an input feature and examining the effect this has on model predictions, with larger changes implying greater reliance of the model on that particular feature. The feature importance is first presented for the gradient boosting model as it was the most successful. Additionally, importance for a random forest is also included to give some indication of the consistency of this metric between models. For such ensemble methods, feature importance is simple to compute as the average of the importances obtained from the individual tree classifiers (Hastie, Tibshirani, and Friedman 2001, pp 368, 593–594).

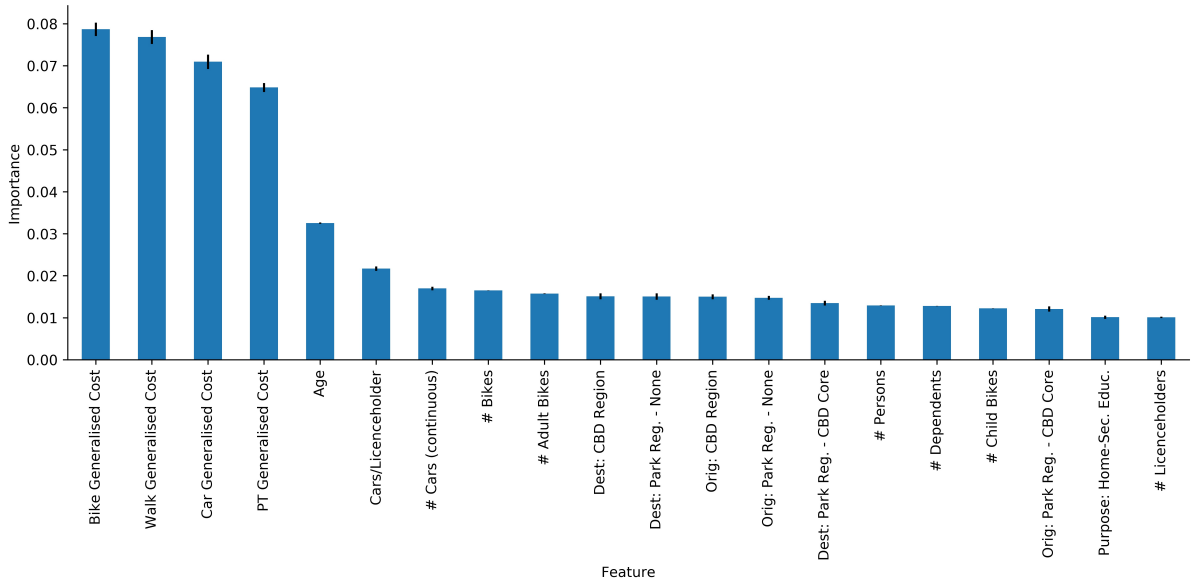
There is reasonable association between the two importance plots in Figure 4 with both models regarding the same first 5 features as most important. The prevailing generalised costs were the basis for the comparison model features and given the reasonable results it makes logical sense that they are valued highly by the model. Whilst it is expected age would impact mode choice somewhat, it is interesting to note the high degree of importance placed on this variable. Cars per licence-holder and number of cars encode information about the viability of choosing the car mode so it is logical these are valued by both models. For the CBD region and parking region indicators which follow, there is not a distinct importance ordering, as they share near identical mean and non trivial variance for the gradient boosting case. This does however still indicate that their collective importance reflects the potentially different mode dynamics in the central business district. In terms of remaining features, gradient boosting model makes use of some of the binary encodings of the categorical trip purpose and origin/destination activity features. In contrast, the random forest values some household informational figures; number of persons and dependants. Examining the significantly higher standard deviation in the importance weights in the gradient boosting model, this can likely be attributed to the sequential model construction, as opposed to the parallel construction of forests. Given that successive trees are designed to correct prior ones, this construction may suffer somewhat when particular input features are permuted.

The repetition of similar features regarding the number of bicycles within a household raises a known issue with feature importance, where the importance of correlated features is split amongst them (Molnar 2019, section 5.5.5). Consequently, there is a collective importance associated to the number of bicycles, but it is unclear precisely whether it is a particular aspect that is important, as correlation of these features means they are all regarded as valuable. For categorical variables, there is also difficulty in interpreting feature importance, given that for model prediction they are encoded as independent binary variables. For both models, the home-secondary education purpose segment is useful and origin and destination activities are also exploited by gradient boosting. Ideally the collective contribution of all the purpose or origin activity segments would be evaluated. Unfortunately, this is not the sum of the individual purposes as this does not account for interaction effects. Further development to analyse the collective contribution may help clarify the value of including such features.

**Figure 4: Comparison of variable importance for the two most successful models; random forest and gradient boosting. Note that this is a truncation of the 20 most important features of those the model was trained one. The collective sum of all the importance score of all features is one.**

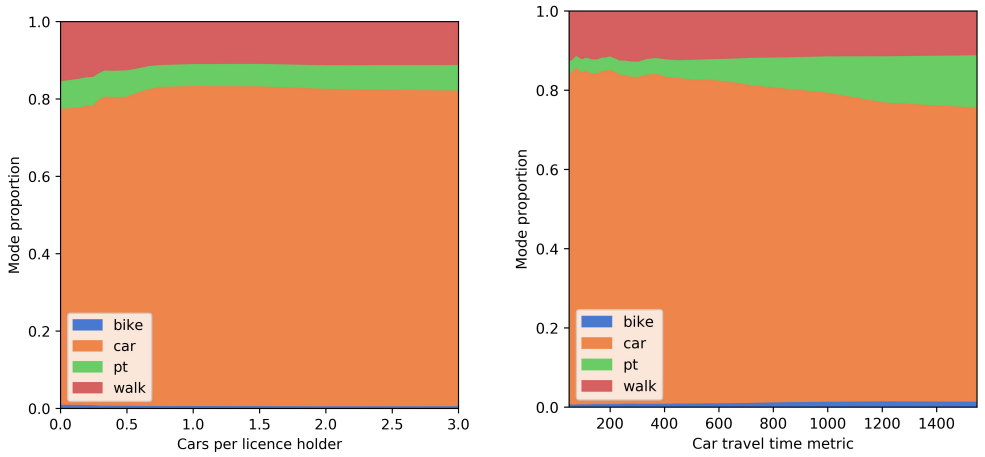



(a) Gradient boosting




(b) Random Forest

**Figure 5: Partial dependence plots for the features cars per licence holder and car travel time resulting from the gradient boosting model. Relatively little change in proportion occurs throughout the range of sample values for both plots. Note that the x axis in (b) is not a literal time in minutes or seconds, but rather a scaling proportional to the time**




(a) Cars per licence holder partial dependence

(b) Car travel time partial dependence

To supplement the feature importance plots, partial dependence plots are produced, which examine the effects of perturbing a single input feature in isolation. Whilst often a useful diagnostic tool for interpreting machine learning models, they unfortunately appear to be of limited usefulness for the mode choice problem. Examining Figure 5, very little change is exhibited in the mode distribution aside from at extreme values for both cars per licence holder and car travel time. Indeed, for both of these quantities, varying a single parameter in isolation results in potentially non physical scenarios. Adjusting cars per licence holder does not adjust the related number of cars and persons, or underlying indirect factors such as household income. Likewise, varying the travel time by car can produce infeasible scenarios as all travel times are correlated to an extent. Perturbing a single feature however does not account for this, and consequently partial dependence can produce trips with extreme travel times by car and marginal travel times by foot. While this is possible in special scenarios, it is certainly not standard behaviour. Furthermore, the high degree of correlation between these features potentially means that all mode travel times can be correlated in the same sense for a particular mode.

This highlights a potential extension for further research regarding machine learning and the mode choice problem. Machine learning techniques are constructed in a versatile manner, which potentially discards known problem specific information. In particular, for mode choice, a potential extension would be to restrict certain features from partitioning particular classes. For example, it is reasonable that public transport travel time and cost should not dictate a decision between travelling by car and walking, however this is perfectly legal for standard machine learning models. By placing such restrictions on the underlying decision trees, it may be possible to produce more effective and interpretable models for this particular problem.

# 4 Conclusion

The performance of machine learning methods were assessed on the mode choice problem, relative to a NL model. Overall, promising results were found with models performing well in regards to both individual level predictive accuracy and the aggregate mode share. Additionally, these models were found to be more accurate than the NL reference implementation. Extending ML to rich feature sets of detailed, individual specific level data, resulted in incrementally higher predictive accuracy. The best performing model was gradient boosting with a mean predictive accuracy of 0.9002 (standard deviation 0.0030). Interpretative evaluation was conducted, examining feature importance and partial dependence of models. Whilst insight gained was consistent with the econometric understanding of influential factors, further research is required to effectively investigate the collective importance of highly correlated features and categorical features. Finally, it was proposed that custom machine learning models taking advantage of the established relationship between mode specific input features and the choice set could produce more accurate and interpretable results.

# 5 Acknowledgements

The authors acknowledge the industry experience program of the School of Mathematics and Physics at the University of Queensland that resulted in the precursor to this work, Tim Veitch and Pedro Camargo for inspiring discussions, and Ben Cowley for proof-reading.

# 6 References

Abadi, M. et al., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, : USENIX Symposium on Operating Systems Design and Implementation, viewed Jun. 4 2019, https://www.tensorflow.org/

Ben-Akiva, M. E., Lerman, S. R. & Lerman, S. R., 1985. Discrete choice analysis: theory and application to travel demand. Cambridge, Massahusetts: MIT press

Bierlaire, M., 2016. PythonBiogeme: a short introduction. :Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Breiman, L., 2001. Random Forests. Machine Learning, 1 Oct., Volume 45, pp. 5-32, doi: 10.1023/A:1010933404324

Cheng, L. et al., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behaviour and Society, Volume 14, pp. 1-10, doi: 10.1016/j.tbs.2018.09.002

Chollet, F. & others, 2015. Keras. :Version 2.2.4, software, https://keras.io

Defazio, A., Bach, F. R. & Lacoste-Julien, S., 2014. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. arXiv: the Computing Research Repository, viewed May 16 2019, http://arxiv.org/abs/1407.0202

Dios Ortuzar, J. & Willumsen, L. G., 2011. Modelling transport. West Sussex, United Kingdom: John Wiley & Sons

Hagenauer, J. & Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications, 2.Volume 78, doi: 10.1016/j.eswa.2017.01.057

Hastie, T., Tibshirani, R. & Friedman, J., 2001. The Elements of Statistical Learning. 2 ed. New York: Springer Science+Business Media

Hensher, D. A., Rose, J. M. & Greene, W. H., 2005. Applied choice analysis: a primer. Cambridge: Cambridge University Press

Kingma, D. & Ba, J., 2014. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, viewed Jun. 2 2019 https://arxiv.org/pdf/1412.6980.pdf

Lee, D., Derrible, S. & Pereira, F. C., 2018. Comparison of Four Types of Artificial Neural Network and a Multinomial Logit Model for Travel Mode Choice Modeling. Transportation Research Record, Volume 2672, pp. 101-112, doi: 0.1177/0361198118796971

Ma, T.-Y., Chow, J. Y. J. & Xu, J., 2017. Causal structure learning for travel mode choice using structural restrictions and model averaging algorithm. Transportmetrica A: Transport Science, Volume 13, pp. 299-325, doi: 10.1080/23249935.2016.1265019

Molnar, C., 2019. Interpretable Machine Learning. : Christoph Molnar, viewed Apr. 23 2019, https://christophm.github.io/interpretable-ml-book/

Omrani, H., 2015. Predicting Travel Mode of Individuals by Machine Learning. Transportation Research Procedia, Volume 10, pp. 840-849, doi: 10.1016/j.trpro.2015.09.037

Omrani, H. et al., 2013. Prediction of Individual Travel Mode with Evidential Neural Network Model. Transportation Research Record, Volume 2399, pp. 1-8, doi: 10.3141/2399-01

Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, Volume 12, pp. pp. 2825--2830

Train, K. E., 2009. Discrete choice methods with simulation. Cambridge: Cambridge university press

Veitch Lister Consulting, 2014. Zenith model for Victoria validation report. :Veitch Lister Consulting, viewed Jun. 6 2019, https://veitchlister.com.au/wp-content/uploads/2018/08/ZML-VIC_ZenithVictoria_ModeChoice_RevB.pdf

Victoria State Government Department of Transport, 2018. Victorian Integrated Survey of Travel and Activity. :Victoria State Government, viewed May 14 2019, https://transport.vic.gov.au/about/data-and-research/vista

Wang, F. & Ross, C. L., 2018. Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. Transportation Research Record, Volume 2672, pp. 35-45, doi: 10.1177/0361198118773556

Xie, C., Lu, J. & Parkany, E., 2003. Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. Transportation Research Record, Volume 1854, pp. 50-61, doi: 10.3141/1854-06

Yann, L., Bengio, Y. & Hinton, G., 2015. Deep learning. Nature, 5, Volume 521, pp. 437-438, doi: 10.1038/nature14539

Zhang, Y. & Xie, Y., 2008. Travel Mode Choice Modeling with Support Vector Machines. Transportation Research Record, Volume 2076, pp. 141-150, doi: 10.3141/2076-16

Zhao, X., Yan, X., Yu, A. & Hentenryck, P. V., 2018. Modeling Stated Preference for Mobility-on-Demand Transit: A Comparison of Machine Learning and Logit Models. arXiv: the Computing Research Repository, viewed Jun. 1 2019, http://arxiv.org/abs/1811.01315