Rahnasto and Hollestelle
European Transport Research Review (2024) 16:43
https://doi.org/10.1186/s12544-024-00667-9

European Transport Research Review

ORIGINAL PAPER

Open Access

Comparing discrete choice and machine learning models in predicting destination choice

Ilona Rahnasto[1*] and Martijn Hollestelle[1]

Abstract

Destination choice modeling has long been dominated by theory-based discrete choice models. Simultaneously, machine learning has demonstrated improved predictive performance to other fields of discrete choice modeling. The objective of this research was to compare machine learning models and a multinomial logit model in predicting destination choice. The models were assessed on their predictive performance using metrics for both binary classification and probabilistic classification. The results indicate that machine learning models, especially a random forest model, could bring improvements in prediction accuracy. The more data was used in training the models, the better the machine learning models tended to perform compared to the multinomial logit model. With less data, the multinomial logit model performed comparatively well. The findings are relevant for the field of destination choice modeling, where evidence on the use of machine learning models is very limited. In addition, the unbalanced choice sets of destination choice models with multiple non-chosen alternatives increases the need for further research in model fit and parameter tuning.

Keywords Discrete choice modeling, Travel demand, Destination choice modeling, Machine learning, Predictive modeling, Random forest

1 Introduction

Travel behavior modeling has been dominated by theory-based discrete choice models, like multinomial logit models, for decades (e.g. [24, 26, 29]). While these models have their advantages in interpretability and straightforward estimation, they also come with strict statistical assumptions, which may, especially with wrong specifications or small amount of data, lead to inaccurate predictions. To add, Wang & Ross [26] argue that theory-based models might be too stiff to respond to changes resulting from emerging transport technologies and new streams of data and information.

Simultaneously, research in machine learning has been making advancement (e.g. [7]). One of the key advantages of these alternative modeling approaches is that, unlike theory-based models, machine learning models have been precisely developed to maximize predictive performance [7]. Compared to traditional discrete choice models, machine learning models don't rely on prior beliefs or behavioral theories but rather learn exclusively from the data itself [24, 29].

Recent literature reviews (see [10, 24]) state that novel empirical evidence of using machine learning in modeling travel behaviour is limited but promising. So far, these models have proven to bring significant improvements in predictive power and detecting underlying data patterns in travel demand modeling as well as other fields of discrete choice modeling [29]. However, van Cranenburgh et al. [24] mention that research on using machine learning specifically in modeling travel demand has

*Correspondence:
Ilona Rahnasto
ilona.rahnasto@ramboll.fi
[1] Digital Mobility Lab, Ramboll Finland Oy, Itsehallintokuja 3, Espoo 02601, Finland

mainly focused on mode and vehicle choice, and research on the application of machine learning in destination choice modeling seems very limited.

This paper contributes to the theoretical framework of comparing discrete choice and machine learning models and provide empirical insights on the predictive power of machine learning models in destination choice modeling. Given the strong reliance in practice on discrete choice model, one may assume that transitioning to a different paradigm is perceived with caution. Travel demand models are important tools to inform infrastructure investment decision making, and have therefore strict requirements in terms of accuracy and transparency. This research contributes to the practice by showing how machine learning can improve travel demand models and what the considerations and possible trade-offs are.

The objective was to investigate whether machine learning models could increase predictive accuracy in destination choice modeling. Specifically, the area of interest was predictive destination choice models that were evaluated in the in the context of an activity- and tour-based travel model simulation. The aim was to build compatible binary and probabilistic classification machine learning models and compare them to the multinomial logit model based on predictive power. This paper is based on one of the authors' master's thesis written in Aalto University in 2022 [15].

2 Background and context

2.1 Brutus travel demand model

The research was conducted in the context of an agent-, activity- and tour-based travel demand simulation, that aims to represent a full day's travel within the transport system of a daily urban system realistically. Brutus uses multinomial logit models in predicting travel-related choices. The model consists of both destination and mode choice models that are linked together in a tour-based approach. In this paper, we do not present Brutus in detail, and the reader may refer to e.g. Rijksen et al. [16] or Salomaa [19] for further description of the model.

The destination choice models are distributed over different segments, which is done for the tour-based destination selection, so that each destination within a tour can be modeled individually. The segments are formed of data points that share activity type in (i) *a*, previous endpoint, (ii) *b*, next endpoint and (iii) *i*, the destination of the trip, as well as (iv) modes of trip *A* and (v) *B*. These trips and endpoints are illustrated in Fig. 1. Each segment thus contains all alternatives hypothetically considered by individuals that made similar trips based on the above attributes and the actual realized trips within those attributes. The model estimation is then done separately for each segment.

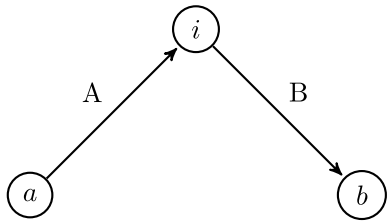

Fig. 1 Representation of trips *A* and *B* and endpoints *a*, *i* and *b*

The models in Brutus follow the theoretical framework of Discrete Choice Modeling as described by Ben-Akiva & Lerman [1]. The individual is assumed to behave based on utility maximization, i.e. choosing the alternative that gives them the most utility. The utility for the individual n for choosing the destination alternative i can be given with the utility function as

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

where V_{in} is the deterministic term and represents a linear combination of the measurable attributes of alternative i and individual n, and ϵ_{in} is the stochastic part representing uncertainty related to individual choice. The deterministic part can be further written as

$$V_{in} = \beta x_{in}. \tag{2}$$

In the above, β is the coefficient estimated from the choice data and x_{in} are the variables that characterize the utility of each alternative. As it is assumed that, given the same amount of land-use, longer trips are less attractive than shorter trips, β is ensured to receive a negative value in model estimation. In Brutus, the deterministic part of the utility function is specified as

$$V_{in} = \beta t_{tot} + \log m_i$$

where t_{tot} is total travel impedance for trips to and from destination i and m_i is the relative importance of land use in destination i in respect to the activity in destination i. The selection probability for destination i within a choice situation with a choice set C_n is given as

$$P(i \mid C_n) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}}. \tag{3}$$

For this research setting it is relevant to understand that in the destination choice model estimation of Brutus, (1) for each trip at most a hundred hypothetical destination alternatives are generated, (2) the travel survey observations are divided into smaller segments based on trip attributes (described above), and (3) the model estimation is done separately for each segment.

The multinomial logit models in Brutus are highly interpretable in terms of individual-related decision making. However, the tour-based segmentation sometimes leads to data scarcity and thus insignificant models and mispredictions. Due to the order of predictive modeling, especially the destination choice models are prone to be affected by the small amount of data in certain segments and were hence considered interesting for this research.

2.2 Machine learning compared to discrete choice models

As presented in Sect. 2.1, traditional discrete choice models are derived from econometric behavioural theories. In travel demand modeling, they have long been the established method, as they are relatively simple, have theoretical grounding in Random Utility Theory, and are flexible in including diverse attributes, making them analytically tractable for policy analysis.

These theories provide travel demand modellers an interpretable framework, where the relationship between the input and output variables is fairly easy to understand. However, van Cranenburgh et al. [24] explain that these theories impose structures and restrictions on the models that in practice may be violated and lead to misspecifications and erroneous predictions.

Machine learning models, on the other hand, are typically more flexible as they are less constrained by strict assumptions on pre-mapped functions or input variables [24, 29]. Indeed, with the less restricted form of machine learning models, one can overtake problems of model restrictions and may have the advantage of tabula rasa [24]. This data-driven approach of machine learning often leads to higher predictive performance [29]. Due to this quality, machine learning models often perform better with large streams of data and can be used with types of data that are incomprehensible for traditional theory-based discrete choice models [24].

However, with increasing predictive power the analyst often encounters a trade-off between accuracy and interpretability [9]. For a complex model it may be impossible to interpret how individual predictors are associated with the response [9]. As noted, machine learning models do not have a similar link to behavioral theory as the traditional models, which often means compromising in interpretability.

While machine learning models may bring challenges in interpretability, van Cranenburgh et al. [24] note that their flexibility often leads to higher goodness-of-fit. This, in turn, they claim, indicates accurate capturing of the underlying choice processes, and can further help understand behavioral patterns in travel demand [24]. This claim is supported by Zhao et al. [29], who found machine learning models to realistically reflect choice patterns and to serve as a tool to explore and identify nonlinear effects in travel demand utility functions.

While the potential of machine learning models is evident, according to van Cranenburgh et al. [24], the field of discrete choice modeling is still hesitant to make the shift. van Cranenburgh et al. [24] raise that this may stem from misconceptions towards machine learning or a lack of understanding about what machine learning and theory-based modeling have in common and what sets them apart. To encourage and guide the practitioners in enhancing discrete choice and travel demand modeling, more research and evidence on the use of machine learning models is needed, especially in applications where studies have not yet been conducted.

2.3 Novel results on machine learning in travel demand modeling

As the field of travel demand and transport modeling have traditionally been ruled by the multinomial logit model and its variations [26], research on machine learning applications in travel demand modeling is limited and, according to Wang & Ross [26], not rigorous enough. However, various scholars have applied different data-driven methods to certain travel demand applications. Below, we present some of the findings we considered most relevant related to this research.

The only study that the authors could find on machine learning models in the context of discrete destination choice was conducted by Thill & Wheele [21]. They studied a decision tree induction algorithm predicting destination choice, and they estimated the model twice differing the number of non-chosen alternatives per choice situation between the runs. They found that the prediction accuracy would increase when the training data included less non-chosen alternatives, which they argue to be due to overfitting and learning noise. On a higher level, they found the decision tree to perform satisfactorily and, compared to other discrete choice modeling results, very favorably. [21]

When comparing machine learning and logit models in mode choice modeling, Zhao et al. [29] found that a random forest model had much higher predictive accuracy compared to a multinomial logit model. They also found that random forest models could capture nonlinear relationships between input and output variables. This, they claim, can be an efficient tool in the specification of the utility functions of different alternatives [29]. The results of Hagenauer & Helbich [6] are as well in favor of a random forest model. They investigated classifiers similarly in a mode choice context, and found that in terms of prediction accuracy, random forest performs significantly better than the other classifiers, including the multinomial logit model [6].

Lee et al. [11] compared four types of artificial neural network models to a multinomial logit model in mode choice prediction, and found that the ANN models outperformed the multinomial logit model with prediction accuracies around 80% compared to 70% for multinomial logit model. They, similarly to Zhao et al. [29], also argue for the machine learning models' ability to capture non-linearity and biases in the data and, contrary to common beliefs, for their easy applicability [11]. Moreover, Wang & Ross [26] compared the application of an extreme gradient boosting model to a multinomial logit model in mode choice modeling, and found that the extreme gradient boosting model has a higher prediction accuracy than multinomial logit, especially when the dataset is balanced.

Zhang & Xie [27] studied a support vector machine in travel mode choice modeling and compared it to a traditional, yet very sophisticated multinomial logit model. Even with the relatively competitive performance of the multinomial logit, they still found the predictive power of support vector machines to have a slight edge [27].

3 Methodology

The empirical part of the research was conducted as model estimation and evaluation in the context of destination choice in a travel demand simulation described in Sect. 2.1. The models were treated and evaluated as binary and probabilistic classification models, for which the aim was to predict the discrete choices and choice probabilities of destination alternatives as close to reality as possible. As the interest was the predictive power of the models, only the prediction outcomes were investigated. In this section, data, models, model estimation and validation as well as assessment metrics are described.

3.1 Data collection and processing

The travel survey data used in this research was collected within the Finnish National Travel Survey in 2016, for which the data descriptions and summaries are available online (see [23]). Home-to-home trips were removed as they do not contain an explicit destination and can therefore not be modelled with the existing destination choice model framework.

Land use data was geospatial grid data produced by Statistics Finland and the Finnish Environment Institute for urban structure observation [22]. Each grid cell represents a 250x250m^2 area for which the land use variables describe total aggregate quantities within that cell.

The transport network had been constructed for Brutus using two sources of data. The first of these was public transport and road network data produced by the Helsinki Regional Transport Authority [5]. The second source was OpenStreetMap [13], which was used to complement the public transport and road network with data on light transport. The networks had been combined to achieve a comprehensive network of motor traffic, walking and cycling. Then, connector links between the transport network and grid cells had been formed, which had been further used to calculate travel times between grid cells with different modes.

For each trip, a choice set with up to 100 non-chosen alternatives is generated. Attributes of the non-chosen alternatives are travel impedance data derived from the model network and destination characteristics from the land-use grid. These trips with their choice sets were segmented as described in Sect. 2.1. Segments in the survey with less than two trips were removed, as one trip i.e. choice situation isn't divisible into train and test sets. The data set obtained from the processing step consisted of 11,508 trips reported by 2784 unique individuals. The data that represented the generated alternatives and was used in destination model estimation consisted of 1,074,550 instances. Each trip had thus 93 alternatives on average.

3.2 Variables

The models were estimated twice for different variable sets. This was done to assess how models behave with varying amounts of input attributes and to see if the number of variables would affect the predictive power of the models. First, the models were estimated with all variables. Second, the models were estimated with reduced variables to reduce the risk of collinearity and to gain more predictive power. Reduction was done based on absolute Pearson's correlation coefficient being higher than 0.7, as suggested by Dormann et al. [4]. The correlation coefficients are presented in Appendix 1. The removed variables were travel time, travel impedance from previous location and travel impedance to next location. The complete list of variables used in both estimations are presented in Table 1.

Before model estimation, input variables were normalized. This was performed to obtain computational efficiency for all models that calculate Euclidean distances or use gradient descent. This is also supported by Iguyon & Elisseeff [8], who suggest that features should be comparable. Thus, numeric variables were normalized with interval scaling between [0, 1] (see e.g. [25]). This should be kept in mind if interpretation of effects of individual variables on predictions is performed. However, it does not affect the predictive capabilities of the models [14, 25] and it was thus considered appropriate for the scope of this study.

Table 1 Variables for model evaluation

Variable	Type	Round 1	Round 2
Age	Integer	x	x
Gender	Logical	x	x
Family size	Integer	x	x
Employed	Logical	x	x
Number of cars owned	Integer	x	x
Car user	Logical	x	x
Travel time	Numeric	x	
Travel impedance from previous location	Numeric	x	
Travel impedance to next location	Numeric	x	
Total travel impedance	Numeric	x	x
Travel cost	Numeric	x	x
Parking cost	Numeric	x	x
Amount of attraction	Numeric	x	x
Number of residents	Integer	x	x
Number of jobs	Integer	x	x
Number of service jobs	Integer	x	x
Number of accomodation and food service jobs	Integer	x	x
Healthcare in area	Logical	x	x
Grocery store floor area	Integer	x	x
Specialty store floor area	Integer	x	x
Assembly floor area	integer	x	x

3.3 Models

All estimated models are presented in Table 2. Models were estimated using R (version 4.1.0) and its open-source packages in Windows Subsystem for Linux 2 and Ubuntu (version 20.04.3).

The multinomial logit model is the model that is currently used in Brutus simulation model, and it was considered the baseline for the model comparison. The theoretical background for the model and its current implementation are described in Sect. 2.1.

To identify how the models performed compared to arbitrary choice, random selection was also used as a comparison model. All alternatives within a choice situation were assigned the same choice probability $p_i = 1/N$. For binary classification, the chosen alternative was picked arbitrarily.

Table 2 Summary of evaluated predictive models and their requirements

Model	R-package	Input variables require normalization	Choice probabilities require normalization
Multinomial logit model	mlogit (0.2.2)	No	No
Multinomial logit via neural networks	nnet (7.3.16)	Yes	Yes
Naïve Bayes	naivebayes (0.9.7)	No	Yes
Random forest	randomForest (4.6.14)	No	Yes
Support vector machine, linear kernel	e1071 (1.7.7)	Yes	Yes
Support vector machine, third degree polynomial kernel	e1071 (1.7.7)	Yes	Yes
Gradient boosted regression trees, logistic loss function	gbm (2.1.8)	Yes	Yes
Gradient boosted regression trees, AdaBoost loss function	gbm (2.1.8)	Yes	Yes

It should be noted that there is a delicate difference in how theory-based discrete choice models and machine learning models approach the choice prediction problem. This, as stated by Zhao et al. [29], leads to differences in how the models and their predictions should be treated and interpreted. In this paper, the modeling context mimicked the context of the theory-based probabilistic discrete choice modeling. Following the recommendations of Zhao et al. [29], the models were estimated based on the true (chosen) and false (not chosen) values, where the true values represented the reported trips, and the false values represented the generated, theoretically considered alternatives. The output was, depending on the assessment metric, either a logical value indicating choice (chosen/not-chosen) or a probability distribution indicating the probabilities of the alternatives being chosen.

Another important note is that, differently from the multinomial logit model, the other models do not necessarily yield normalized probabilities for choice situations. This means that the sum of the choice probabilities within a choice set in the multinomial logit model is always 1, while for other models the probabilities need to be normalized to achieve comparability. Without normalization we would encounter practical contradictions: a machine learning model may not always predict a true choice for a choice situation, when in reality, the agent did make the trip to one of those destination. Hence, for compatibility and comparability, the output probabilitites were scaled with simple feature scaling so that the final probability were given as

$$p_i' = \frac{p_i}{\sum p}.$$

The authors remark that the above technique is merely a "quick fix", and that this contradiction in different modeling paradigms should lead to broader discussion on the use of differently derived models in the context of discrete choice modeling.

3.4 Model estimation and validation

Models were estimated in data segments that were divided by trip types. The segmentation follows the segmentation logic of Brutus described in Sect. 2.1. The models were run for in total 883 segments separately following a k-fold cross-validation technique (see e.g [18], p. 708). The goal of cross-validation is to see how well the models generalize on unseen data. The need for validation on unseen data was also observed by Thill & Wheeler [21] in their early study of using decision trees in predicting destination choice, where they did not use such a technique and hypothesised on encountering overfitting of the model. The k-fold technique has the advantage that it uses all the available data and gets thus accurate estimates [18].

The k-fold cross-validation was performed with $k = 5$ in each segment. The estimation data was divided into k parts with random sampling, however, so that data within a certain choice set C_n was in the same sample. Models were run for each $k \in [1, 2, 3, 4, 5]$ so that the k'th sample was set out as test data and the remaining $k - 1$ samples are set out as training data. The trained models would then predict a probability distribution for the data points in the k'th part of the data in the segment.

The predictions were assessed in segment groups that were formed based on number of observations (i.e. trips) in the segments. This was done to be able to assess the results in more depth and to understand how the amount of data affects model performance. The segments were divided into small, medium and large groups, where the intervals for segment division based on trip numbers were, respectively, $[2, 5]$, $[6, 10]$ and $[11, \infty]$.

The maximum number of trips in a segment was 174. The first interval consisted of 132 segments, 448 trips and 182,481 choice alternatives, the second of 504 segments, 4150 trips and 660,481 choice alternatives and the third of 87 segments, 4975 trips and 112,639 choice alternatives. The segment division could have been made in multiple various ways, and this approach was chosen due to its easy applicability and direct relation to amount of input data, which was one of the underlying reasons to assess novel models. The models were assessed with binary and probabilistic classification metrics, which are presented in Sect. 3.5.

3.5 Model assessment

The models were evaluated using metrics developed for (1) binary classification models and (2) probabilistic classification models. The two methods were selected to assess how the models performed in replicating reality in a wider sense, as choice prediction is rarely performed in isolation but rather as part of a larger model system. A model could for example predict the chosen alternative correctly, but the distribution of the choice probabilities related to other alternatives would not reflect reality. Hence, using only a simple binary assessment was thought to be too limiting, but considered useful for quantifying the results and understanding the differences of model performance in practice.

As the focus was firstly on applying machine learning in a destination choice context and secondly assessing their predictive power, methods for evaluating the models' fitting and interpretability were left out. Similarly, detailed parameter tuning for the models was not performed.

It is to be noted that selection criteria should be evaluated individually by use case by the analyst. In situations where the user is only interested in the output and getting correct predictions, focusing on metrics of predictive power is usually enough. However, if deeper analysis on the drivers and variables affecting the output is needed, interpretability should also be considered.

3.5.1 Binary classification evaluation

The modelled results were first evaluated with binary classification scores on an aggregate level to assess the models' ability to predict the correct chosen alternative as a baseline. A common practice in literature for individual level binary choice classification is to label, for each choice set C_n, the alternative with the highest choice probability as chosen [29], i.e.

$$\arg\max_i(p_1, ..., p_j).$$

The rest is labeled as not chosen. If multiple alternatives have the same choice probability, the true label is given randomly to one of these alternatives.

To assess the binary classification power, the positive predictive value (PPV), also called precision, was used. It indicates how often the predictions of the positive class are correct. In this context, it describes the share of a model's predictions for chosen destinations that were in fact chosen. Compared to the often used balanced accuracy metric, PPV does not consider each correctly predicted outcome as equally valuable. That is, it does not give value to the close to one hundred correctly predicted not-chosen alternatives, but rather evaluates the true choice of each choice situation. This is a suitable metric for predicting discrete choice with a large choice set, as for all choice situations we always predict exactly one chosen alternative. Positive predictive value is given as

$$PPV = \frac{TP}{TP + FP} = \frac{Number\ of\ true\ positives}{Number\ of\ positive\ calls} \quad (4)$$

where TP is number of true positive calls and FP is number of false positive calls. For best predictive power, positive predictive value should be maximized.

One should note, that as in each choice situation exactly one outcome is chosen and exactly one prediction for the chosen class is made, one will end up with the same score by using recall. That is, the number of false positives and false negatives is equal within a choice situation. Hence, instead of PPV, one could choose to use recall or F1-score as well and the overall performance rank of the models would be the same.

3.5.2 Probabilistic classification evaluation

The assessment of the probabilistic models was done with proper scoring rules, which is a common practice when evaluating probability distributions as predictions. Proper scoring rules are functions that compute scores based on the predicted probabilities and the events that actually occur. That is, they are used to determine the predictive performance when the estimated output is not a set of labels but a set of probabilities.

The three most commonly adopted strictly proper scoring rules were adopted as metrics, and they are presented below. For all scores, p denotes the predicted probability, x is the actual outcome of the event, where $x = 1$ is given for chosen alternatives and $x = 0$ otherwise, and N is the number of forecasting instances to be assessed. Note that for overall evaluation N is equal to the number of all predicted values, whereas for segment-based evaluation N is the size of the given segment.

The Brier score [2], also known as an affine transformation of the quadratic score, was originally developed as a tool for weather forecast verification. It has thus been widely used in probabilistic modeling evaluation. It is given by

$$BS = \frac{1}{N}\sum_{j=1}^{N}(p_j - x_j)^2 \in [0, 1] \quad (5)$$

where j indicates the forecasting instance. For maximal predictive performance, the score is aimed to be minimized.

The spherical scoring rule was introduced by Roby [17] for psychological testing. It is expressed as

$$SS = \frac{1}{N}\sum\frac{p_i}{||\mathbf{p}||} = \frac{1}{N}\sum_{j=1}^{N}\frac{x_j p_j + (1 - x_j)(1 - p_j)}{\sqrt{p_j^2 + (1 - p_j)^2}} \in [0, 1] \quad (6)$$

and for maximal predictive performance, the score should be maximized.

The logarithmic score, which was the first characterization of any scoring rule, was introduced by Shuford et al. [20] and is given as

$$LS = \sum_{j=1}^{N}(x_j \ln p_j + (1 - x_j)\ln(1 - p_j)) \in]\infty, 0]. \quad (7)$$

For maximal predictive performance, the logarithmic score should be maximized. It should be noted, that the logarithmic score is not bounded from below. This means that the rule can in practice yield scores equal to minus

infinity. This is possible e.g. when a model gives a choice probability of zero to a true event.

4 Results

For binary classification, aggregate positive predictive values were calculated for both estimation rounds and they are presented in Table 3. For probabilistic classification, scores were calculated similarly for both estimation rounds. They were calculated aggregately for all predictions but also separately in segment groups. The segment groups were determined by amount of data as described in Sect. 3.4. Aggregate scores for predictions made with all variables are presented in Table 4 and with reduced variables in Table 5. The rest of the results are included in Appendix 2.

Table 3 Positive predictive value for all predictions

	Positive predictive value, all variables	Positive predictive value, reduced variables
Random forest	0.554580	0.520987
Multinomial logit model	0.397260	0.396073
Gradient boosted regression, AdaBoost loss	0.381995	0.368961
Gradient boosted regression, logistic loss	0.369135	0.356187
Multinomial logit via neural networks	0.340454	0.336434
Support vector machine, linear kernel	0.281109	0.260601
Support vector machine, polynomial kernel	0.269986	0.232534
Naive Bayes	0.266201	0.220467
Random choice	0.015381	0.015554

Table 4 Probabilistic scores for all predictions with all variables

	Brier	Logarithmic	Spherical
Random forest	0.006688	– Inf	0.992905
Multinomial logit model	0.008410	– Inf	0.991095
Gradient boosted regression, logistic loss	0.008544	– 39742.822580	0.990947
Gradient boosted regression, AdaBoost loss	0.008795	– 44971.704607	0.990651
Multinomial logit via neural networks	0.009190	– Inf	0.990321
Support Vector Machine, linear kernel	0.009777	– 51550.456456	0.989945
Support Vector Machine, polynomial kernel	0.009944	– 55850.431381	0.989832
Random choice	0.010622	– 51927.060075	0.989302
Naïve Bayes	0.018853	– Inf	0.980384

Table 5 Probabilistic scores for all predictions with reduced variables

	Brier	Logarithmic	Spherical
Random forest	0.006996	– Inf	0.992600
Multinomial logit model	0.008428	– Inf	0.991084
Gradient boosted regression, logistic	0.008522	– 39739.968200	0.990977
Gradient boosted regression, AdaBoost	0.008675	– 43422.048846	0.990776
Multinomial logit via neural networks	0.009063	– Inf	0.990442
Support vector machines, linear	0.009816	– 52371.694019	0.989915
Support vector machines, polynomial	0.010102	– 57663.467891	0.989704
Random choice	0.010541	– 62792.570686	0.989386
Naïve Bayes	0.010651	– Inf	0.988975

Looking at the aggregate scores in both binary and probabilistic classification contexts, Tables 3, 4 and 5, the random forest model outperformed other models, including the multinomial logit model. In segments with more data, as presented in Tables 9 and 12, the random forest but also other machine learning models, including gradient boosted regression trees and neural networks, performed better than the multinomial logit model. However, in segments with less data (see Tables 7 and 10), the multinomial logit model still performed relatively well.

Interestingly, the absolute scores of the multinomial logit model were in general worse in large segments compared to scores in small or medium segments. Some of the scores for other models, including gradient boosted regression with logistic loss, support vector machines with linear kernel and naïve bayes, also did not improve directly with more data.

Comparing the two rounds in probabilistic scores, the relative performance of the multinomial logit model was improved when reducing variables, while other models performed better with more variables. In small segments, however, all models performed better with less variables. Interestingly, when comparing the two rounds in binary scores in Table 3, all models, including the multinomial logit model, performed better with more variables.

As mentioned, the positive predictive value can be interpreted as the percentage of getting actual true predictions right out of all true predictions. Interpreting from Table 3, this means that random forest gets 55% of the true predictions right, while naive bayes gets 27%. For multinomial logit model, the true predictions are right in 40% of the choice situations. The increase from multinomial logit to random forest is 28 percentage points and 40%. On average, if the destination model predicts ten thousand trips, this would mean an increase of in 1500 correct predictions.

5 Discussion

The empirical results of this research shed light on how machine learning models perform in a destination choice context. So far, research on using machine learning in travel demand modeling has focused on predicting mode and vehicle choice.

The performance of the random forest model is similar to the results of Hagenauer & Helbich [6] and Zhao et al. [28], where random forests outperformed multinomial logit models in mode choice contexts. This result may be due to its capability to model variable interactions and nonlinear relationships, as argued by Lhéritier et al. [12], as well as the ability to handle large data sets and many variables [3]. Random forest is an aggregate model that yields the final predictions based on multiple boosted trees, which may explain its superiority in predictive capability.

Gradient boosted regression models' performance varied between groups of different amounts of input data. With more observations its performance was relatively better. As gradient boosted regression models are relatively complex and fit well on training data [26], they may suffer from overfitting and thus yield poor results especially when predicting choice with small shares of a very unbalanced dataset [26]. On the contrary, this quality may lead to better performance with more data and relatively less noise.

In this research, the larger choice set with multiple non-chosen alternatives may be the cause of differences in the performance of these models. The non-chosen alternatives in the choice set generate, some may even argue, irrelevant, variability in the data, to which complex models tend to respond by sometimes causing overfitting. This would be in line with the findings of Thill & Wheeler [21], where a decision tree algorithm predicted better with fewer non-chosen alternatives, which they suspected was due to having less noise in the learning data. However, to detect whether the model is in fact overfitting, one would have to look at both training accuracy and prediction accuracy (see e.g. [9]) which was left out of the scope of this study.

Interestingly, support vector machines with two different kernels did not perform better than the multinomial logit model in any of the comparisons. This is contrary to Zhang & Xie [27], who found that support vector machines provided the highest prediction accuracy when compared to multinomial logit and neural networks in predicting mode choice. Also, Hagenauer & Helbich [6] found support vector machines to perform significantly better than a multinomial logit model.

There might be multiple reasons behind this difference. Support vector machines with a third-degree polynomial kernel, i.e. a more complex model that is more prone to overfitting, performed better than support vector machines with a linear kernel only in the large segments with all variables. This would hint that, with less data, the more complex model overfits to the limited data points. This could be explained by the large choice set of "irrelevant" alternatives similarly as with gradient boosted models.

Another explanation to the relatively poor performance of support vector machines is that they might not be able to accurately separate the alternatives in the feature space in the training phase. Especially in the case of a

linear kernel and in contrast to the complex models, this could be the result of underfitting, which leads to higher model bias and poor generalization on new predictions. However, without assessing the model fit, it is hard to say whether the models are actually over- or underfitting. Further investigation on how the models are fitting and how to overcome issues related to it within a large choice set, be it e.g. alternative sampling or model tuning, should be conducted.

Similar to the results of Zhao et al. [28], naïve bayes performed worse than any other model and yielded even worse scores than random choice. Even though in the comparison of Hagenauer & Helbich [6], naïve bayes performed better than the multinomial logit model in prediction accuracy, it was still the worst of the other models, and only slightly better than the multinomial logit model. To assess the issue, as naïve bayes classifier assumes strictly the independence of the input variables, it could benefit from more aggressive variable rejection.

Overall, the more complex machine learning models like gradient boosted regression trees and neural networks seem to perform better with more data. As training data is increased, these models may generalize better to predictions on unseen data and capture underlying patterns in the data that less complex models might miss. However, counterintuitively, some models performed worse in large segments compared to medium segments. For the multinomial logit model, this was nearly consistent throughout the results. As also gradient boosted regression with logistic loss, support vector machines with linear kernel and naïve bayes peform better in the medium segment than the large segment, we suspect the data in larger segments may have some nonlinear structures that models relying on linear structures are not able to capture. To understand exactly why these models perform somewhat strangely, the input data should be investigated.

The comparatively good predictive performance of the multinomial logit model especially when estimated with reduced variables in segments with less data still proves its value even for purely predictive use. It may be that in situations where there is less data to learn from, the model benefits from having behaviorally derived assumptions as a starting point. In this research, the modeling context had been developed for the multinomial logit model following theoretical frameworks of discrete choice modeling, which may contribute to its performance. In addition to traditional models performing well with less data, their interpretability still makes them relevant. Machine learning models should not replace traditional models completely but bring additional flexibility and predictive accuracy.

6 Conclusion

The findings of this research are significant for the field of destination choice modeling, where evidence on the use of machine learning models is very limited. The introduction of machine learning models could bring major improvements in predictive accuracy already, but their application in this context needs more research.

The findings show that especially a random forest model can be more accurate than the multinomial logit model, even when models are trained with little data. With more data, also gradient boosted regression and neural network models could overperform the multinomial logit model in terms of predictive accuracy. Overall, when more data is available for model training, machine learning models tend to benefit from it more than the multinomial logit model. With less training data, a multinomial logistic model still performs well, as it has a starting point in behavioral assumptions without needing to learn as much from data from scratch.

We believe that large choice sets with many non-chosen alternatives may cause overfitting and poor predictions in some of models. Thus, models that perform well in predicting discrete choice with less alternatives should not automatically be considered good options for predicting within larger choice situations.

As the large number of non-chosen alternatives is very characteristic of destination choice modeling, we believe that further studies investigating model fit in unbalanced choice situations are vital. Potential models should also be further evaluated within choice sets of different sizes and weights of alternatives to understand how models could perform better even if data is unbalanced. Some models may benefit from model parameter tuning and variable selection, especially after knowing how they fit, and research focusing on specific model development could provide valuable insight as well.

6.1 Limitations

The variable selection and choice set design was optimized for the multinomial logit model, and it may not represent the true potential of some of the other models. The study was performed in a disaggregate activity-based model and data from Helsinki, Finland was used, and different contexts may yield different results. As already mentioned, both model fit and estimation error assessment and model parameter tuning were left out of the study, but could have an effect on the modeling results.

Appendix 1: Pearson's correlation coefficients

See Table 6.

Table 6 Pearson's correlation coefficients

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
(1) age	− 0.07	− 0.46	− 0.02	− 0.15	0.29	− 0.08	− 0.01	− 0.01	− 0.01	0.08	0	− 0.17	0.03	0.01	0.01	0.02	0	0.07	0.02	− 0.03
(2) gender		− 0.02	− 0.01	0.04	0.15	− 0.03	− 0.01	− 0.01	− 0.01	0.05	− 0.02	0.01	− 0.03	− 0.01	− 0.01	− 0.01	− 0.02	− 0.02	− 0.01	− 0.01
(3) family size			− 0.09	0.45	− 0.02	− 0.1	− 0.02	− 0.02	− 0.02	0.08	− 0.04	0.12	− 0.08	− 0.04	− 0.03	− 0.04	− 0.01	− 0.06	− 0.03	0
(4) employed				0.04	0.28	0.01	0	0		0.06	0	− 0.14	− 0.02	− 0.01	− 0.01	− 0.02	− 0.04	− 0.04	− 0.01	− 0.02
(5) cars owned					0.46	− 0.25	− 0.03	− 0.02	− 0.02	0.3	− 0.08	0.02	− 0.11	− 0.07	− 0.05	− 0.07	− 0.04	− 0.05	− 0.04	− 0.04
(6) car user						− 0.27	− 0.03	− 0.02	− 0.02	0.33	− 0.04	− 0.13	− 0.05	− 0.03	− 0.02	− 0.03	− 0.03	0	− 0.01	− 0.04
(7) travel time							0.77	0.76	0.79	0.08	− 0.05	0	− 0.03	− 0.04	− 0.03	− 0.06	− 0.02	− 0.04	− 0.03	− 0.02
(8) A_imp								0.86	0.96	0.61	− 0.09	− 0.04	− 0.09	− 0.08	− 0.06	− 0.08	− 0.03	− 0.02	− 0.03	− 0.04
(9) B_imp									0.96	0.61	− 0.08	− 0.04	− 0.07	− 0.06	− 0.05	− 0.07	− 0.03	− 0.02	− 0.03	− 0.03
(10) imp										0.64	− 0.07	− 0.04	− 0.08	− 0.06	− 0.05	− 0.06	− 0.03	− 0.02	− 0.03	− 0.03
(11) cost											− 0.07	− 0.09	− 0.12	− 0.07	− 0.05	− 0.06	− 0.05	− 0.01	− 0.02	− 0.04
(12) park cost												0.2	0.37	0.64	0.36	0.66	0.03	0.22	0.33	0.42
(13) attr													0.12	0.29	0.17	0.32	0.09	0.45	0.25	0.2
(14) residents														0.16	0.14	0.35	0.2	0.17	0.04	0.08
(15) jobs															0.69	0.71	0	0.33	0.52	0.36
(16) service jobs																0.41	0.14	0.15	0.2	0.19
(17) acc. food jobs																	0.01	0.41	0.51	0.35
(18) healthcare in area																		− 0.02	− 0.04	0.02
(19) grocery store area																			0.44	0.12
(20) specialty store area																				0.17
(21) assembly area																				

Appendix 2: Probabilistic classification results

See Tables 7, 8, 9, 10, 11 and 12.

Table 7 Probabilistic scores, all variables, small segments

	Brier	Logarithmic	Spherical
Random forest	0.007742	– Inf	0.991767
Multinomial logit model	0.007865	– Inf	0.991705
Gradient boosted regression, logistic	0.008736	– 2529.700107	0.990750
Gradient boosted regression, AdaBoost	0.008946	– 3164.158663	0.990547
Multinomial logit via neural networks	0.009403	– Inf	0.990155
Support vector machines, linear	0.009842	– 3203.177694	0.989930
Support vector machines, polynomial	0.010002	– 3540.532401	0.989850
Random choice	0.010240	– 3752.135089	0.989700
Naïve Bayes	0.010285	– Inf	0.989393

Table 8 Probabilistic scores, all variables, medium segments

	Brier	Logarithmic	Spherical
Multinomial logit model	0.007770	– Inf	0.991737
Random forest	0.007774	– Inf	0.991740
Gradient boosted regression, logistic	0.008463	– 20347.566404	0.991047
Gradient boosted regression, AdaBoost	0.009040	– 25877.322791	0.990416
Multinomial logit via neural networks	0.009472	– Inf	0.990060
Support vector machines, linear	0.009619	– 25886.144363	0.990105
Naïve Bayes	0.009866	– Inf	0.989760
Support vector machines, polynomial	0.009958	– 28955.476820	0.989854
Random choice	0.010334	– 31467.797042	0.989601

Table 9 Probabilistic scores, all variables, large segments

	Brier	Logarithmic	Spherical
Random forest	0.005257	– Inf	0.994441
Gradient boosted regression, AdaBoost	0.008485	– 15930.223153	0.990942
Gradient boosted regression, logistic	0.008612	– 16865.556069	0.990857
Multinomial logit via neural networks	0.008846	– Inf	0.990635
Multinomial logit model	0.009674	– Inf	0.989813
Support vector machines, polynomial	0.009920	– 23354.422160	0.989804
Support vector machines, linear	0.009955	– 22461.134399	0.989759
Random choice	0.011290	– 16707.127944	0.988607
Naïve Bayes	0.029453	– Inf	0.969319

Table 10 Probabilistic scores, reduced variables, small segments

	Brier	Logarithmic	Spherical
Multinomial logit model	0.006530	– Inf	0.993010
Random forest	0.007280	– Inf	0.992277
Gradient boosted regression, logistic	0.008128	– 2599.199232	0.991390
Gradient boosted regression, AdaBoost	0.008612	– 3219.766363	0.990859
Multinomial logit via neural networks	0.008639	– Inf	0.990898
Support vector machines, linear	0.009096	– 3257.554065	0.990624
Support vector machines, polynomial	0.009708	– 3790.573007	0.990094
Random choice	0.010238	– 4134.021526	0.989701
Naïve Bayes	0.010337	– Inf	0.989281

Table 11 Probabilistic scores, reduced variables, medium segments

	Brier	Logarithmic	Spherical
Multinomial logit model	0.007918	– Inf	0.991601
Random forest	0.008201	– Inf	0.991383
Gradient boosted regression, logistic	0.008559	– 20931.525900	0.990969
Gradient boosted regression, AdaBoost	0.009055	– 25366.349494	0.990409
Multinomial logit via neural networks	0.009476	– Inf	0.990059
Support vector machines, linear	0.009824	– 27579.699485	0.989952
Support vector machines, polynomial	0.010120	– 30503.976477	0.989733
Random choice	0.010330	– 31886.204874	0.989606
Naïve Bayes	0.010433	– Inf	0.989246

Table 12 Probabilistic scores, reduced variables, large segments

	Brier	Logarithmic	Spherical
Random forest	0.005467	– Inf	0.994152
Gradient boosted regression, AdaBoost	0.008217	– 14835.932989	0.991214
Gradient boosted regression, logistic	0.008540	– 16209.243067	0.990920
Multinomial logit via neural networks	0.008633	– Inf	0.990831
Multinomial logit model	0.009867	– Inf	0.989625
Support vector machines, linear	0.009921	– 21534.440469	0.989755
Support vector machines, polynomial	0.010144	– 23368.918406	0.989606
Random choice	0.010849	– 26772.344286	0.989065
Naïve Bayes	0.010938	– Inf	0.988632

Acknowledgements
We would like to express our gratitude to Rainer Kujala and Jukka Luoma for their support and guidance.

Author contributions
All authors contributed to the design of the study. Data collection was performed by Martijn Hollestelle. Ilona Rahnasto designed and performed model estimation, validation and assessment. Ilona Rahnasto took the lead in writing the manuscript, and Martijn Hollestelle provided critical feedback and helped shape the final manuscript. All authors read and approved the final manuscript.

Funding
This article has been published open access with support of the TRA2024 project funded by the European Union. The work was partly financed by Ramboll Finland Oy.

Availability of data and materials
All data availability is described in Sect. 3.1. The travel survey data, HLT, should be requested from the Finnish Transport and Communications Agency, Traficom.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no conflict of interest.

Received: 24 October 2023 Accepted: 6 August 2024
Published online: 21 August 2024

References

1. Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. MIT Press.
2. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.
3. Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society, 14*, 1–10.
4. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46.
5. Elolähde, T., Räty, P., & West, J. (2019). *Helsingin seudun työssäkäyntialueen liikenne-ennustejärjestelmän kysyntämallit 2017*. Technical report, HSL Helsingin seudun liikenne.
6. Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications, 78*, 273–282.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning data mining, inference, and prediction (12th printing)*.
8. Iguyon, I., & Elisseeff, A. (2003). *An introduction to variable and feature selection*.
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning—With applications in R|Gareth James|Springer*.
10. Koushik, A. N., Manoj, M., & Nezamuddin, N. (2020). Machine learning applications in activity-travel behaviour research: A review. *Transport Reviews, 40*(3), 66.
11. 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: Journal of the Transportation Research Board, 2672*(49), 101–112.
12. Lhéritier, A., Bocamazo, M., Delahaye, T., & Acuna-Agost, R. (2019). Airline itinerary choice modeling using machine learning. *Journal of Choice Modelling, 31*, 198–209.
13. OpenStreetMap contributors. (2020). *OpenStreetMap*.
14. Patro, S. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. In *IARJSET*.
15. Rahnasto, I. (2022). *Comparing discrete choice and machine learning models in predicting destination choice* (MSc thesis).
16. Rijksen, H., Hollestelle, M., Koopal, R., Brederode, L., & Moilanen, P. (2019). Development of a hybrid travel demand model combining agent based microscopic and gravity based macroscopic approaches. In *European transport conference, Dublin*.
17. Roby, T. B. (1965). Belief states and the uses of evidence. *Behavioral Science, 10*(3), 66.
18. Russell, S., & Norvig, P. (2010). *Artificial intelligence a modern approach* (3rd edn.).
19. Salomaa, O. (2011). *An accessibility-based simulation model for destination and mode choice of trips* (Master's Thesis). Espoo.
20. Shuford, E. H., Albert, A., & Edward Massengill, H. (1966). Admissible probability measurement procedures. *Psychometrika, 31*(2), 66.
21. Thill, J. C., & Wheeler, A. (2000). Tree induction of spatial choice behavior. *Transportation Research Record, 1719*, 66.
22. Tilastokeskus. (2019). *Tilastokeskuksen tuottama ruutuaineisto yhdyskuntarakenteen seurantaan (YKR)*.
23. Traficom. (2016). *Valtakunnallinen henkilöliikennetutkimus*.
24. van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning—Discussion paper. *Journal of Choice Modelling, 42*, 100340.
25. Wan, X. (2019). Influence of feature scaling on convergence of gradient iterative algorithm. In *Journal of Physics: Conference Series* (vol. 1213).
26. Wang, F., & Ross, C. L. (2018). Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(47), 35–45.
27. Zhang, Y., & Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record: Journal of the Transportation Research Board, 2076*(1), 141–150.
28. Zhao, D., Shao, C., Li, J., Dong, C., & Liu, Y. (2010). Travel mode choice modeling based on improved probabilistic neural network. In *Proceedings of the conference on traffic and transportation studies, ICTTS* (vol. 383).
29. Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society, 20*, 22–35.

Publisher's Note