Uncovering taste heterogeneity and non-linearity for urban mode choice using SHAP

by

Thaddäus Christoph Weißhaar

to successfully complete the

HONOURS PROGRAMME MASTER IN "TRANSPORT, INFRASTRUCTURE AND LOGISTICS"

at the Delft University of Technology

Student number:	5611652	
Project duration:	February 1, 2022 – December 19, 2023	
HPM committee:	Dr. Ir. S. van Cranenburgh,	TU Delft, supervisor
	Dr. Ir. A.J. Pel,	TU Delft, supervisor

An electronic version of this report is available at http://repository.tudelft.nl/.

TU Delft

Uncovering taste heterogeneity and non-linearity for urban mode choice using SHAP

Thaddäus Christoph Weißhaar

***Abstract*—European cities are implementing diverse strategies to curtail car usage. Understanding the impact of these policies necessitates insights into mode choice behaviour. However, for conventional discrete choice models, utility specifications must be defined upfront, potentially leading to misleading policy recommendations. This problem is solved by Supervised machine learning (ML) models. However, they are challenging to interpret, which is crucial for evaluating transportation policies.**

We employ Shapley additive explanations (SHAP), a model-agnostic explainable artificial intelligence (XAI) tool to address this gap. The main advantages of SHAP are its foundation in game theory, the ability to highlight individual taste heterogeneity and non-linear effects. This paper aims to shed light on the potential of SHAP to improve current transportation mode choice models. Using a random forest (RF) model with the TreeSHAP estimation method, we compare SHAP insights with those derived from a traditional multinomial logit (MNL) model.

The results indicate that SHAP can detect the absolute importance of features. Substantial preference heterogeneity for car choice is perceived for features reducing car usage, as opposed to features increasing car usage. Non-linear effects, such as reciprocal functions and clustered patterns, are observed for certain features. MNL and RF models disagree on the importance and heterogeneity of features, and the MNL model fails to model highly nonlinear effects.

For policymakers, insights suggest that increasing parking fees and promoting car sharing may be feasible options. However, the efficiency of these measures may vary due to preference heterogeneity. The results underscore the need for further investigation into the reasons behind the different model results and different behaviour notions of SHAP, MNL and RF.

***Index Terms*—machine learning, random forest, Shapley values, SHAP, urban transportation, choice modelling, mode choice**

I. Introduction

European cities are adopting strategies to reduce car usage in their centres, including congestion pricing, pedestrian zones, and cycling infrastructure. They are also expanding public transport, introducing low-emission zones, promoting car-sharing and raising parking fees. These measures aim to improve sustainability, reduce congestion, and enhance urban living conditions. To assess the impact of such policy measures on citizens, insights into the factors determining mode choice are necessary. Previous studies (see Cornago, Dimitropoulos, and Oueslati [33], Hasan et al. [35] and Fan et al. [34]) highlighted the importance of various factors for choosing transportation modes in an urban setting, such as price, time, comfort, safety, socio-demographics and contextual factors like weather.

All of these studies used discrete choice models, which have the crucial limitation that the researcher has to define the utility specification upfront. Model specification is tedious since various theories must be tested to derive the best-fitting model. If some interactions between factors are neglected in the research, the models do not adequately represent reality, and conclusions might be misleading.

ML models like a RF can solve the interaction problem and fit data efficiently. They learn patterns to classify or regress the data. Supervised ML models can use flexible models with many parameters to model complex relations between input and output. Upfront, interactions between all factors are assumed. Further advantages include a higher predictive accuracy compared to discrete choice models [36] and the possibility of having non-textual data to increase the realism of choice experiments [17].

However, many ML algorithms, especially complex ones like RF, are so-called ”black-box” models and are challenging to interpret. Understanding why a model makes a particular prediction is crucial in applications like transportation policies, where decisions have significant consequences [36], and public representatives are held responsible for their decisions. The missing ability of parameters to be interpreted on individual and aggregated levels is the main reason why ML models have not been implemented widely in the transportation domain [29].

Recently, XAI methods have been proposed to interpret ML models better, mainly in other sectors like finance, healthcare and tech [32] [38]. Currently, there are at least 133 XAI methods available, however, just eight studies cover different XAI methods in the transportation domain (see Liu et al. [26], Deng [23], Parmar, Das, and Dave [27], [25], Saiyad, Srivastava, and Rathwa [31], [5], Alwosheel, Cranenburgh, and Chorus [21] and Huber et al. [24]).

SHAP by Lundberg and Lee [13] can visualise the complex relation between input and output variables for various models. It has become a famous XAI method in the machine learning community because of its theoretical foundation. SHAP estimates Shapley values, which have their origin in game theory and guarantee a fair distribution of importance among the factors for individual prediction contributions [39]. A SHAP value can be conceptualised in a mode choice setting as a percentage-point contribution of one feature to a probability difference. This difference is between the individual probability of choosing a mode and the average probability of this mode.

One key benefit of SHAP is highlighting individual taste heterogeneity via SHAP values. They can be computed for all factors disaggregated regarding factor levels and alternatives.

With current ML and random utility maximisation (RUM) models, it is not possible to account for it, despite various efforts (see Train [4], Keane and Wasi [8] and Dong and Koppelman [9]). Furthermore, SHAP can account for non-linear and adverse effects [30]. If included in models, potential non-intuitive demand reactions can be captured, leading to a higher effectiveness of policy measures.

This paper aims to shed light on the potential of SHAP methods to improve current transportation mode choice models regarding modelling preference heterogeneity and non-linear and interaction effects of features. Through this, transportation policies in the urban setting could be more targeted towards specific groups of individuals and more effective through the widespread incorporation of feature interaction effects. Furthermore, RF and RUM models will be compared to determine which model is superior to the underlying prediction model when using SHAP. First, the case study and the data set will be presented in section II. Then, in section III, the data analysis methods will be introduced theoretically with the used model specification for the case study. In section IV, the potential of SHAP with RF to model feature importance, feature signs, preference heterogeneity, non-linear effects and interactions will be highlighted and validated through SHAP in combination with a MNL model. Section V rounds off the paper with a discussion.

II. Case study and data set

To answer to which extent the SHAP method is suitable for highlighting heterogeneity, non-linearity and discovering interactions in urban mode choice modelling, a data set from Hillel, Elshafie, and Jin [14] is used. It investigates the mode choices of respondents on the multi-model transportation network of London, which includes the modes of walking, cycling, public transportation and driving by car. The data was collected between April 2012 and March 2015 and comprises 81,086 trips. 31,954 individuals made these in 17,616 households. For each trip, 32 variables were documented. A summary of all variables being used, their coded names, an explanation and their usage is provided in Table I.

For this research, two adjustments to the original data set were made. Firstly, trips with a travel distance shorter than 150m were eliminated. For these distances, mode choices are generally irrelevant since walking is the predominant mode. Secondly, fuel costs and congestion charges were summarised to car travel costs. This was done because respondents' sensitivity towards congestion charges and fuel costs is expected to be equal. Lastly, the access time for PT, the rail travel time, the bus travel time, and the transfer time were aggregated into PT travel time. This simplifies and homogenises the duration of PT, also leading to fewer factors.

After this adjustment, the data set contains 81,005 entries, the final choice of each respondent, and 12 factors. For the MNL model, the factors are divided into eight attributes and four covariates. To prevent overfitting of the RF model, the data has been split into training and test data, with 80% of the data set comprising the training data. Samples for the computation of SHAP values were drawn from the test data set.

III. Methods

As the introduction states, various XAI methods can explain black-box machine learning models. These can be divided into groups along two categories: global vs. local and surrogate vs. explanation generation methods. Whereas global methods try to explain the average outcome, local models explain specific data points [22]. Having the goal of improving the modelling of heterogeneity in mind, local models are more suitable because of their ability to assign the importance of various factors disaggregated towards individuals and modes.

Surrogate methods fit an inherently explainable model closely to the black box model. Explanation generation methods, on the other hand, use the output of the black-box ML model as input for the explanation function. It was not chosen for local explanation generation methods like ICE [7], counterfactual explanations [18], and LRP [10] because of the lower flexibility of these models regarding interpretations. Often, only one plot is generated to show the relation between factors. Secondly, their implementation in packages of programming languages like Python is less convenient than local surrogate models.

Regarding the two most popular local surrogate models, LIME by Ribeiro, Singh, and Guestrin [12], and SHAP by Lundberg and Lee [13], SHAP is preferred because of its theoretical foundation in game theory and with this a fair importance distribution of features. The second reason is the extensive Python package for SHAP, which covers insightful and visually appealing graphs. Two main SHAP estimation methods exist: the model-agnostic KernelSHAP and tree-based TreeSHAP. As shown later in this chapter, TreeSHAP is preferred because of its fast implementation. Therefore, it is possible to calculate exact Shapley values instead of estimations, and interaction values can also be computed. As a tree-based ML model, a RF [2] is used for model estimation II instead of XGBoost [11] because of its theoretical simplicity, its robustness towards overfitting [37] and computational speed.

Firstly, the feature importance table is introduced to provide an overall overview of the importance of features. Secondly, a summary and dependence plot will analyse feature signs, preference heterogeneity and non-linear effects. Lastly, a SHAP interaction table highlights interactions between features. To validate the overall feature importance, feature signs, heterogeneity and non-linear effects, it was chosen to use the inherently interpretable and widely used MNL model in combination with KernelSHAP. In the following, the theory and model specifications of the RF, MNL and SHAP methods will be shown. Particular focus is laid on SHAP plots for the visualisation of SHAP values.

A. Random Forest

1) Theory: A RF classifier is suitable for predicting mode choice, assigning each response to one of the four modes.

Variable	Coding	Description	Usage	Values
CHOICE	travel_mode	chosen travel mode	choice	cycling driving public transportation walking
Pedestrian travel time	dur_walking	walking time	attribute	in hours
Cycling travel time	dur_cycling	cycling time	attribute	in hours
PT travel time	dur_pt_total	in-vehicle, access, egress and interchange time	attribute	in hours
Number of interchanges for PT	pt_n_interchanges	number of interchanges	attribute	#
PT travel cost	cost_transit	cost of public transport	attribute	in GBP
Car travel time	dur_driving	duration of car drive	attribute	in hours
Car travel cost	cost_driving_total	fuel and congestion charge cost	attribute	in GBP
Traffic variability	driving_traffic_percent	congestion on driving route	attribute	in % of usual travel time
Gender	female	gender of the respondent	covariate	1 if female, 0 otherwise
Age	age	age of respondent	covariate	in years
Car ownership	car_ownership	number of cars in a household	covariate	0: no cars 1: less than one car per adult in household 2: one or more cars per adult in household
Driving license	driving_license	whether the respondent possesses a driving licence	covariate	1 if the driver possesses a driving licence 0 otherwise

TABLE I: Used variables

Generally, the RF algorithm consists of five phases. First, data is sampled by randomly selecting subsets of the training data (with replacement) through bootstrapping. This creates diverse training sets for each tree. Then, for each decision tree in the forest, a random subset of features (in choice modelling attributes and covariates) is considered at each split, making the trees less correlated and increasing diversity. Each decision tree is constructed independently. The trees are built by recursively splitting data into subsets based on the selected features, aiming to maximise predictive accuracy. These splits continue until a stopping criterion, such as a maximum depth, is reached. When making predictions, each tree in the forest provides an output. For classifying transportation modes, the most frequent mode of the individual tree predictions is taken as the final prediction for one individual. This approach reduces overfitting [37]. The RF has been implemented using the scikit-learn library of Pedregosa et al. [6]. Several hyperparameters have been optimised to estimate the RF.

2) Hyperparameter settings: Several hyperparameters can be tuned to improve the model fit of the initial model. It was chosen to optimise regarding four hyperparameters, $n\_estimators$, $max\_features$, max_depth and $min\_samples\_split$ since these are the most critical hyperparameters for simple optimisation [40]. $min\_samples\_leaf$ has not been chosen because of its similarity with $min\_samples\_split$.

The first chosen hyperparameter is the number of trees in the ensemble, which is controlled by the $n\_estimators$ parameter. The $max\_features$ parameter in scikit-learn manages the number of features considered at each node. The max_depth parameter sets the maximum depth of each decision tree. Restricting the depth can help prevent overfitting but may also lead to underfitting. The $min\_samples\_split$ parameter determines the minimum number of samples to split an internal node. An increasing value can help prevent overfitting, mainly when dealing with smaller data sets.

For this multidimensional optimisation, a 2 stage process

Hyperparameter	Default value	Tested values
$n\_estimators$	100	30, 100, 300
$max\_features$	$\sqrt{n_{att}}$	1, $\sqrt{n_{att}}$, 7
max_depth	None	10, 30, None
$min\_samples\_split$	2	2, 10, 30

TABLE II: Initial hyperparameter search

was applied. First, the general parameter region is delimited through a grid search. Parameters are chosen to multiply mostly by a factor of 3, as some practitioners do. Default model values are added, and realistic $max\_features$ values have been used [20].

The max_depth range selected is similar to the number of features in the data set. An overview of the used hyperparameters is given in Table II. Each RF model is validated through k-fold cross-validation. The MNL model was estimated on the whole data set. It was decided to divide the training data set into three parts [15] without randomisation. Two-thirds of the training data set is used to train the data, and the third part is used for evaluation. This procedure is repeated three times for each hyperparameter specification. The cross-entropy loss was chosen as the decisive criterion since the negative cross-entropy loss (logistic loss or log loss) can be conceptualised as the log-likelihood value. Furthermore, the cross-entropy loss was averaged across the size of the different data sets (folds or test data set).

After determining the best-fitting model, a second grid search is conducted in the region with the preliminary best fit to find a local maximum. The test settings are displayed in Table III. Ultimately, both estimated RF models will be compared regarding their model fit on the test data, and the model that fits the data better will be chosen.

3) Model evaluation: The RF model is evaluated twofold. Firstly, the average cross-entropy loss is compared across the default, the intermediate and the final RF for all three folds and the test data. As described above, the average cross-entropy loss aligns closely with the log-likelihood, which is often used

Hyperparameter	1st	Tested	2nd
$n\_estimators$	300	250, 300, 350	300
$max\_features$	$\sqrt{n_{att}}$	2, $\sqrt{n_{att}}$, 4	$\sqrt{n_{att}}$
max_depth	None	25, 30, None	None
$min\_samples\_split$	10	8, 10, 12	8

TABLE III: Search for final parameters

for evaluating model performances of discrete choice models.

Secondly, a confusion matrix is used, a common tool to assess the model performance of ML models. Row and column totals have been added, with the column totals denoting the true values in the data set. The row totals depict the overall predictions for the various modes. Lastly, the percentage of true predictions row- and column-wise have been added. Row-wise prediction performance can be interpreted as the share of true positive predictions relative to false positive values. Out of all projections for a specific mode, this determines the share of true predictions. Column-wise prediction performance, however, depicts the share of true positives relative to false negatives. Out of all individuals who used a specific mode, this determines the share of accurate predictions.

B. MNL

As described earlier, to derive SHAP values from KernelSHAP, the MNL model will be used. As indicated in section II, the features in the data set have diverging usages. Some features will make up "attributes", and features labelled as "covariates" will comprise the alternative specific constants. They have been defined for each of the four modes except walking. The covariates gender, age, car ownership and driving licence possession are added.

The utilities for each alternative consist of the alternative-specific constants, attributes and interactions. To derive an optimal model, all covariates are assumed to influence the alternative specific constants. The attribute travel time is expected to affect all modes; costs are expected to impact PT and driving. Lastly, the number of interchanges is assumed to influence PT, and congestion is considered to affect car usage.

Certain interactions are defined to validate the RF based SHAP interaction values. Firstly, age might correlate with travel costs for PT and cars. This might be the case because of a hidden interaction of age with income, which is not provided in the data set. Furthermore, age might interact with owning a car and possessing a driving license. Lastly, car ownership and holding a driving license might correlate since driving is only allowed with a valid driving license.

Firstly, a model without interactions is estimated via the biogeme package from Bierlaire [3]. The least significant parameters are afterwards removed, and via the Likelihood ratio test, it is checked if models with fewer parameters are data-generating processes. In the second step, interactions are added to the model and the procedure of removing parameters and evaluating the model performance via the Likelihood ratio test is repeated.

The alternative specific constant specification and utility functions of the final MNL model can be found below.

$$
\begin{aligned}
asc\_cycling\_value &= asc\_cycling \\
&\quad + asc\_cycling\_shift\_female \cdot female \\
&\quad + asc\_cycling\_shift\_co \cdot car\_ownership \\
&\quad + asc\_cycling\_shift\_dl \cdot driving\_license \\
asc\_pt\_value &= asc\_pt \\
&\quad + asc\_pt\_shift\_age \cdot age \\
&\quad + asc\_pt\_shift\_co \cdot car\_ownership \\
&\quad + asc\_pt\_shift\_dl \cdot driving\_license \\
asc\_car\_value &= asc\_car \\
&\quad + asc\_car\_shift\_age \cdot age \\
&\quad + asc\_car\_shift\_co \cdot car\_ownership
\end{aligned}
$$

$$
\begin{aligned}
V\_Walking &= dur\_walking \cdot Beta\_TT \\
V\_Cycling &= asc\_cycling\_value \\
&\quad + dur\_cycling \cdot Beta\_TT \\
V\_PT &= asc\_pt\_value + dur\_pt\_total \cdot Beta\_TT \\
&\quad + cost\_transit \cdot Beta\_TC \\
&\quad + pt\_n\_interchanges \cdot Beta\_INTER \\
&\quad + Beta\_AGE\_TC \cdot age \cdot cost\_transit \\
V\_Car &= asc\_car\_value + dur\_driving \cdot Beta\_TT \\
&\quad + cost\_driving\_total \cdot Beta\_TC \\
&\quad + driving\_traffic\_percent \cdot Beta\_TRAF \\
&\quad + Beta\_AGE\_TC \cdot age \cdot cost\_driving\_total \\
&\quad + Beta\_AGE\_CO \cdot age \cdot car\_ownership \\
&\quad + Beta\_AGE\_DL \cdot age \cdot driving_license \\
&\quad + Beta\_CO\_DL \cdot car\_ownership \cdot driving\_license
\end{aligned}
$$

To evaluate the model performance, the same tools like for the RF model are used. Assessing MNL model performance via the average cross-entropy loss and confusion matrix enables comparability with the RF model. Despite the MNL model was estimated on the whole training data set, average cross-entropy losses have also been calculated for the various folds for comparability.

C. Shapley values

Shapley values by Shapley [1] originate from game theory and describe how a payout or win is distributed among players cooperating. To determine the "fair" share of one player, first, a payout is calculated for all possible sets of players (coalitions) collaborating, excluding the one player. Second, the average payout of these collaborations is subtracted from the average payout, including the player of interest. The result is the Shapley value for one specific player and one specific outcome.

In mode choice modelling, players can be conceptualised as features and the payout as the prediction of a probability of using a specific mode. In a mode choice setting, a Shapley value can thus be conceptualised as the contribution of one

variable towards the difference in probability between the individual and average predictions of all individuals.

Mathematically, the Shapley values $\phi_j(v)$ are calculated through Equation 1. v stands for the "value", "payout", or "prediction" the "players" or features (attributes/covariates) p are distributing. S denotes the number of features currently in a subset, excluding the feature of interest j. The formula states that the feature contributions towards the prediction are weighted and summed over all feature combinations [39]. The contribution of one feature can be considered "fair" if it satisfies the properties efficiency, symmetry, dummy and additivity [39].

$$\phi_j(v) = \sum_{S \subseteq \{1,\dots,p\} \setminus \{j\}} \frac{|\ S\ |!(p - |\ S\ | - 1)!}{p!} (v(S \cup \{j\}) - v(S)) \tag{1}$$

D. SHAP

1) SHAP: Since Shapley values are cumbersome to compute through Equation 1, various estimation methods have been developed to speed up computation times. One estimation family is SHAP by Lundberg and Lee [13], which computes Shapley values in multiple ways. KernelSHAP is a model-agnostic estimator, whereas TreeSHAP can only be used with tree-based ML models.

The general idea of SHAP is to utilise the concept of a coalition vector $z' \in \{0,1\}^p$, which describes which features are present in a coalition. The prediction values for each coalition can then be determined by a linear formula, with g being the explanation model and ϕ_j the SHAP values. SHAP values are defined as estimated Shapley values through KernelSHAP or TreeSHAP. ϕ_0 represents the average prediction of the model that is to be explained. Lundberg and Lee [13] describe the SHAP properties of local accuracy, missingness and consistency, which also satisfy the Shapley properties efficiency, symmetry, dummy and additivity as shown in the appendix of Lundberg and Lee [13] [39].

$$g(z') = \phi_0 + \sum_{j=1}^{p} \phi_j z'_j \tag{2}$$

2) KernelSHAP: Since it is computationally demanding to compute predictions for all possible features across all data points and output variables, these feature sets or coalitions are sampled for KernelSHAP. Features that are not members of a sampled coalition are randomised. However, some coalitions contain more information than others, and intuitively, the information density is the highest if features are analysed in isolation [39]. Therefore, these coalitions are the first to be added to the sampling "budget" K (see Lundberg and Lee [13]). In the second step, predictions for these coalitions are retrieved from the data, and thirdly, weights for the various coalitions are assigned according to their information density Lundberg and Lee [13]. Lastly, the weighted linear model of coalitions is fitted to the model, which is to be explained through KernelSHAP. This is done through the loss function in Equation 3, where the sum of squared errors is minimised.

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in K} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z') \tag{3}$$

$\hat{f}(h_x(z'))$ denotes the model prediction of the black box model for the instance x, $g(z')$ the SHAP model from Equation 2 and $\pi_x(z')$ the coalition weight highlighted in Equation 4 [39].

$$\pi_x(z') = \frac{M - 1}{\binom{M}{|z'|} |z'|(M - |z'|)} \tag{4}$$

3) TreeSHAP: TreeSHAP is developed explicitly for tree-based models to estimate Shapley values. From the root node, TreeSHAP keeps track of the number of coalitions following specific paths. All coalitions for one data point follow the trees at the same time, which reduces computation times.

At each leaf node, for each coalition, a conditional expectation $E_{X_j|X_{-j}}(\hat{f}(x)|x_j$ is computed, where x_j denotes the features included in the coalition and x_{-j} [39] indicates the features not included in the coalition of the instance x. The conditional expectation will be subtracted from the average prediction for this instance. This difference will then be split into marginal contributions of the various features.

This is done via backpropagation of all coalitions simultaneously for one instance. The marginal contribution mar_contri of the feature representing a node is the percentage of coalitions passing the node on the right side $\frac{r\_count}{tot\_count}$ minus the percentage of coalitions passing the node on the left side $\frac{l\_count}{tot\_count}$, as seen in Equation 5. The contributions are averaged through all the trees to derive SHAP values. Furthermore, the various subsets are weighted according to size, influencing SHAP values. The algorithm is in detail explained in [19].

$$mar\_contri = \frac{r\_count}{tot\_count} - \frac{l\_count}{tot\_count} \tag{5}$$

4) SHAP interaction values: Through the fast computational implementation of TreeSHAP, the exact computation of SHAP interaction values is also possible. The importance of a feature is split into main and interaction effects. Considered are only pair-wise correlations. For this, Shapley interaction indexes are used [19]. They are calculated through Equation 6, where the fraction is a weight and $\bigtriangledown_{ij}(S)$ the raw interaction value for a specific coalition S, an instance x and the features i, j. These features have to be different ($i \neq j$). Equation 7 computes the raw interaction value as the prediction value, including both features subtracted by the prediction values of coalitions, including only i or j, added by a coalition containing none of the features.

$$\phi_{i,j} = \sum_{S \subseteq \{1,\dots,p\} \setminus \{i,j\}} \frac{|S|!(p - |S| - 2)!}{2(p-1)!} \bigtriangledown_{ij}(S), \tag{6}$$

$$\nabla_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \tag{7}$$

SHAP interaction values are split equally, i.e. $\phi_{i,j} = \phi_{j,i}$, the total interaction effect is thus the summation of both values. The main effect is visualised in Equation 8.

$$\phi_{i,i} = \phi_i - \sum_{j \neq i} \phi_{i,j} \tag{8}$$

SHAP interaction values can be interpreted as the difference between the SHAP value for feature i when feature j is present and the SHAP value for feature i when feature j is absent.

5) Comparison TreeSHAP and KernelSHAP: In this paper, TreeSHAP was chosen instead of KernelSHAP for multiple reasons. Firstly, the computation of SHAP values is significantly faster compared to KernelSHAP. TreeSHAP can substantially reduce the computational effort from $O(TL2^M)$ to $O(TLND)$. T denotes the number of trees, L is the maximum number of leaves in any tree, D is the maximal depth of any tree, and M is the number of features. N is the number of background samples used, usually 200. Secondly, the significantly faster computation time allows SHAP values to represent Shapley values accurately. Therefore, no sampling is needed, and no measurement error prevails [19]. Lastly, KernelSHAP ignores correlations among features. It replaces feature values with random values and thus ignores possible correlations between these feature values. If this is the case, too much weight is put onto unlikely data points, which leads to inexact Shapley values [39]. TreeSHAP overcomes this problem by modelling conditionally expected predictions [16].

E. Visualisation of SHAP values

In this paper, it was decided to sample SHAP values because of their computationally demanding nature. A sample size of 250 was chosen to maximise insights while having adequate computation times. These 250 SHAP values generated from TreeSHAP with an RF model or KernelSHAP with an MNL model can be visualised through plots. Because of higher computational speed, the computation of exact values, and the possibility of computing interaction values, it was chosen to use TreeSHAP based on an RF model. The results will be validated through SHAP values derived from KernelSHAP based on the MNL model.

Firstly, this paper provides the overall importance of features through mean absolute SHAP values. They are visualised in the SHAP importance table, using a heatmap of the package seaborn [28]. Whereas the importance of features will be reported for all modes, more extended analysis steps are only conducted for car mode choice. The signs of the features will be analysed through a summary plot using the SHAP package [13], leading to insights into reasonable feature attributions. Taste heterogeneity, thus if feature values have the same or different effects across respondents, can also be visualised through a summary plot [30]. Through visualising SHAP distributions, policymakers gain insight into the effectiveness of measures directed at specific target groups. For example, insights are possible on how many respondents will shift mode if the price scheme is changed.

Furthermore, SHAP can detect non-linearity and adverse effects. Potential non-linear features are initially seen via the summary plot [30]. Then, all features assumed to affect car mode choice directly were confirmed regarding their non-linearity via dependence plots using seaborn. I.e. the attributes of car travel time, travel costs and traffic variability, as well as the covariates of car ownership, age and gender, are checked on non-linearity. SHAP values, however, do assume independence regarding other SHAP values. In the last analysis step, interactions will be tested through a seaborn heatmap with mean absolute SHAP interaction values for the mode car.

As a model-agnostic method, SHAP allows the values estimated through TreeSHAP to be validated with KernelSHAP and MNL. Using SHAP in combination with an MNL model would have the advantage of not missing inherent explainability while still getting more profound insight into heterogeneity. In this paper, all analysis steps besides SHAP interactions will be conducted for KernelSHAP with an RF model. Comparisons between the MNL and RF models regarding features' absolute importance, signs, heterogeneity, and non-linearity are thus possible. Lastly, comparisons of captured interactions in the MNL and RF model are also possible via comparing the SHAP interaction table with the magnitude and significance of interaction parameters derived from the MNL model.

In the following, the used plots, the SHAP importance table, the summary plot, the dependence plot and the SHAP interaction table will be explained in more detail regarding their interpretation.

1) SHAP Importance table: Despite being designed to highlight the individual importance of features, SHAP values can also provide insight into the overall importance of features. The most trivial to interpret is the SHAP Importance table, displaying SHAP importances for each mode separately and averaged. A SHAP importance for a specific mode and feature is calculated by taking absolute SHAP values for each individual and then averaging them overall. SHAP importances can be described as the average effect of a specific feature on the probability of choosing a particular mode. For example, in Figure 2a, car ownership significantly changes the probability of selecting the mode car for urban trips, namely by 15.12% on average. Gender, nonetheless, might only change the probability of choosing the mode car by 0.43%. SHAP values and SHAP importances tend to be scale-dependent, which means that if the probability of selecting a particular mode is lower in the data set, SHAP values and SHAP importances will also be lower.

2) Summary plot: In the summary plot, the importance distribution of one feature is visualised for one mode through SHAP values. For each feature, SHAP values for all individuals are plotted horizontally in a scatter plot. If some SHAP values are more frequent, they are stacked vertically above each other. Additionally, they are colour-coded depending on their respective feature values, where high feature values are

pink and low feature values are displayed in blue. A wide scattering of SHAP values suggests the importance of this feature for choosing modes varies significantly, and there is considerable taste heterogeneity. For example, in Figure 3a, extensive taste heterogeneity emerges for not owning a car. Low car ownership influences the probability of choosing the mode car differently. A concentrated plot, on the other hand, indicates homogeneity in tastes among individuals, like for owning a driving license in Figure 3a.

Additionally, non-linear effects can be detected by analysing the distribution of the SHAP value colouring of the plot [30]. If multiple colours are present in a cluster, it indicates that various feature values have the same effect. Clusters and preference heterogeneity can lead to non-linearity, which can be analysed in more detail in the dependence plot.

3) Dependence plot: A dependence plot is helpful to gain insight into the non-linear effects of the mode choice probability for specific features. On the x-axis, the feature values, and on the y-axis, SHAP values are drawn. SHAP values derived from an RF model are visualised in blue, whereas values derived from the MNL model are drawn in red. The relation between importance and attribute levels is linear in RUM-based models, whereas in ML models, the relation can be highly non-linear. For example, in Figure 4d, the RF model detects a significant rise in the probability of choosing the mode car for 0 to 10-year-olds due to their parents driving them. For individuals aged 10 to 20, however, the probability is significantly lower, which might be because parents do not want to drive their children anymore. For respondents aged 20 years old and older, age has almost no influence on the probability of choosing the mode car. The MNL model detects only a slightly higher probability of choosing the mode car for younger respondents.

4) SHAP interaction table: For evaluating interactions between features, SHAP interaction values [16] are adequate because of their theoretical background highlighted in III-D4. Each SHAP value is split into main and interaction effects. From 12 SHAP values per respondent, 88 interactions are computed (12 main and 66 interaction effects). Because of its high computational demand, it was chosen to sample 50 respondents from the already sampled 250 respondents. The main effects are on the main diagonal, and the rest of the matrix comprises the interaction effects, with the same interaction values above and below the main diagonal. This matrix is calculated over all respondents, which makes the calculation time-intensive. SHAP interaction values can be visualised on an individual level as well as on an aggregated level. For an individual, the summation of all SHAP interaction values and the mean prediction delivers the personal prediction for a respondent.

This paper focuses on an aggregated plot. Similar to the attribute importance plot, firstly, all SHAP interaction values are converted to absolute values, and then the mean of these absolute values is computed. Interaction effects are present above and below the main diagonal. SHAP interactions in Figure 5 can be interpreted like the following: On average, for choosing the mode car, the joint effect of pedestrian travel time and cycling travel time is 1.92%, and the main effects are 5.71% (pedestrian travel time) and 2.67% (cycling travel time). SHAP interaction values are implemented in the TreeSHAP algorithm and not in KernelSHAP. Therefore, they are just available for tree-based ML methods.

Estimated parameters	Value	Rob. std.err.	Rob. p-value
Beta_AGE_CO	-0,0136	0,0009	0,0000
Beta_AGE_DL	0,0160	0,0009	0,0000
Beta_AGE_TC	-0,0009	0,0003	0,0075
Beta_CO_DL	0,2459	0,0319	0,0000
Beta_INTER	0,8218	0,0226	0,0000
Beta_TC	-0,0941	0,0143	0,0001
Beta_TRAF	-2,6522	0,0671	0,0000
Beta_TT	-6,0300	0,0652	0,0000
asc_car	-2,7172	0,0576	0,0000
asc_car_shift_age	0,0097	0,0013	0,0000
asc_car_shift_co	1,4864	0,0381	0,0000
asc_cycling	-3,8490	0,0578	0,0000
asc_cycling_shift_co	-0,1326	0,0422	0,0017
asc_cycling_shift_dl	0,6232	0,0580	0,0000
asc_cycling_shift_female	-1,0289	0,0529	0,0000
asc_pt	-0,8188	0,0310	0,0000
asc_pt_shift_age	0,0088	0,0006	0,0000
asc_pt_shift_co	-0,2864	0,0202	0,0000
asc_pt_shift_dl	-0,4760	0,0271	0,0000

TABLE IV: Estimated Parameters of the MNL model

Model	Fold 1	Fold 2	Fold 3	Test data
MNL Model	0.7480	0.7550	0.7631	0.7503
Default RF	0.7220	0.7327	0.7147	0.6642
Intermediate RF	0.6364	0.6344	0.6297	0.6098
Final RF	0.6342	0.6294	0.6261	0.6063

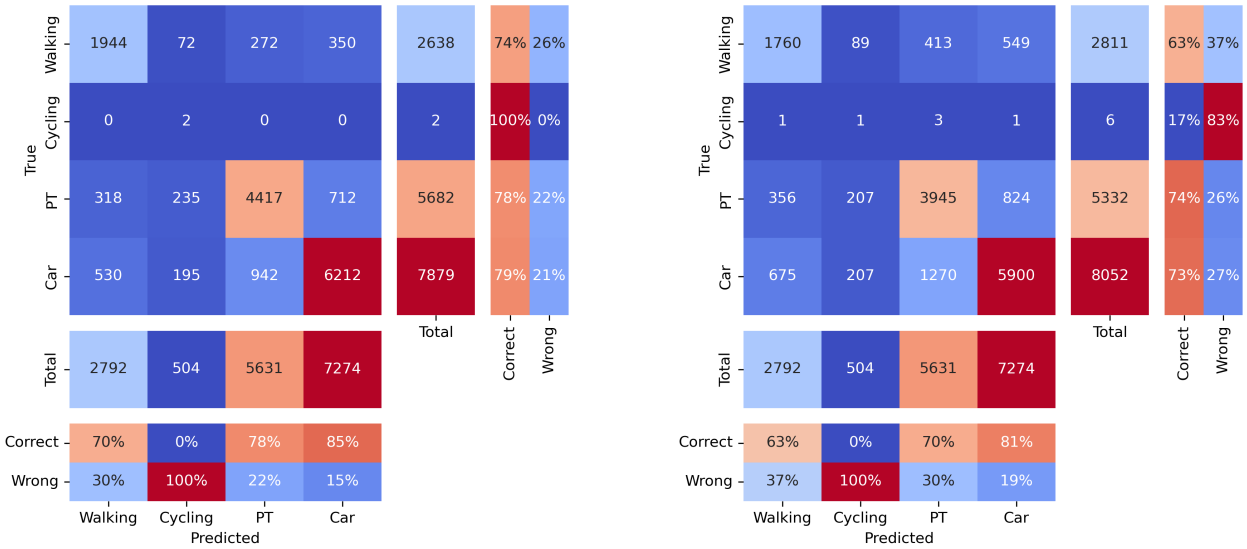
TABLE V: Average cross-entropy loss

IV. Results

A. RF and MNL model validation

1) MNL model estimation: An overview of the estimated MNL parameters is provided in Table IV. The robust p-value shows that all variables are significant on a 5% level. Most parameters have expected signs like $Beta\_TT$ and $Beta\_TC$. However, $Beta\_INTER$ leads to a higher preference for PT if there are more interchanges, which is opposite to general intuition.

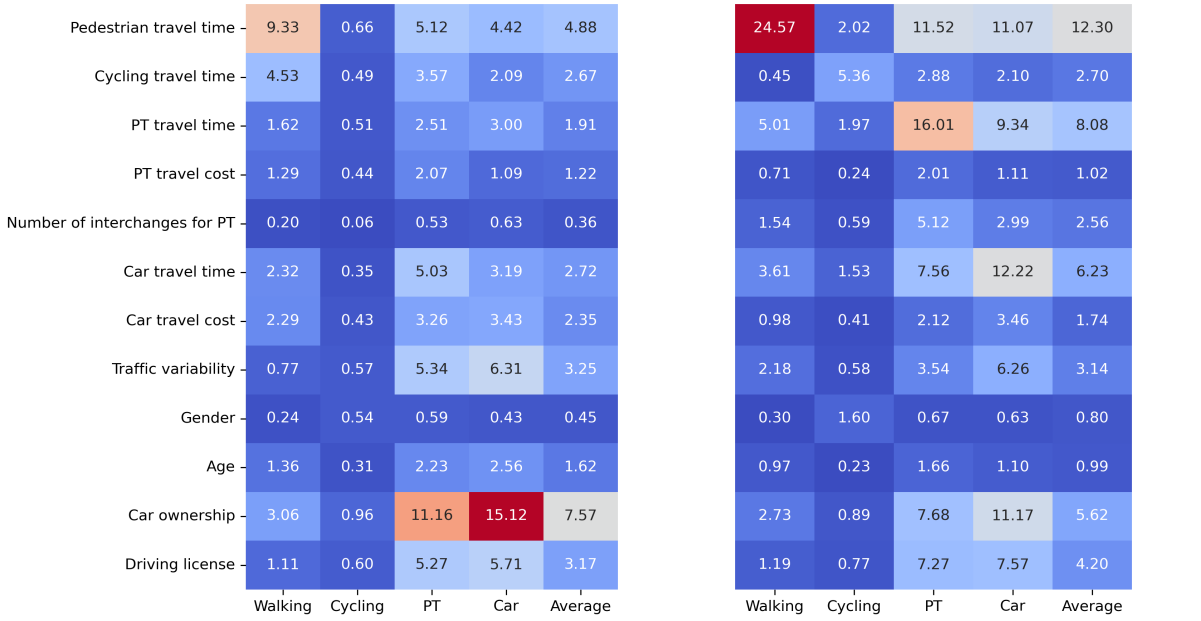
2) RF and MNL model evaluation: Apparent Table V is the significant difference between the MNL and the final RF model regarding the average cross-entropy value. The final RF model was chosen for further analyses because of its superior prediction performance for the test data set and all folds. As seen in Figure 1, both models have difficulties predicting the underrepresented mode cycling, with 0% of all cyclists expected to be cyclists. However, The other modes are reliably predicted; for example, 85% of all car users have been detected in the RF model and 81% in the MNL model. Most prediction errors occur in both models between car and PT. Generally, it can be concluded that both models predict the data set similarly well, with a better performance of the RF model.



(a) Confusion Matrix for RF model

(b) Confusion Matrix for MNL model

Fig. 1: Confusion Matrix comparison



(a) Feature importance in % for the RF model

(b) Feature importance in % for the MNL model

Fig. 2: Comparison of the importance of features

B. Feature importance

Figure 2a visualises mean absolute SHAP values as a proxy for the overall importance of features. Car ownership, pedestrian travel time and traffic variability are the most important determinants on average, with mean average SHAP values of 7.57%, 4.88% and 3.25%, respectively. For each mode, different variables are important. For pedestrians, for example, the walking and cycling times are the most important factor, indicating strong substitution effects between the two modes. The most critical determinant for all modes besides walking is car ownership, implying that these modes compete more intensely with each other. Driving by car and using PT are closely intertwined, with car-related attributes, car ownership, owning a driving license, and traffic variability being the most important features for both modes.

Apparent is that modes' travel times influence not only the probability of their own mode but also of competing modes. This is intuitive since, for example, short pedestrian travel times might deter people from using the mode car. On average, pedestrian travel times influence the probability of using the car by 4.42%.

To summarise, travel times are the most relevant features for active modes. However, socio-demographics like car ownership and a driving licence are the most decisive for choosing the modes PT and car. This starkly contrasts travel costs being relatively unimportant for urban travellers in London. In the following subsections, reasons for selecting the mode car will be analysed in more detail. Firstly, the feature signs will be investigated. In a second step, light is shed on preference heterogeneity. Then, non-linearity is analysed and finally, feature interactions.

C. Feature signs

Most variable signs in Figure 3a are aligned with expectations according to the estimated MNL model. Car ownership emerges as a crucial factor, with low car ownership significantly associated with a lower probability of car choice. Further factors where the expected feature signs emerged are, among others, traffic variability, possessing a driving licence and high car travel costs. Notably, gender does not influence the probability of choosing a car.

However, some surprising results emerge. Only low pedestrian travel times negatively impact the probability of choosing the car. In contrast, higher pedestrian travel times have only a marginal effect. However, this effect is also present for the cycling travel time to a smaller extent. One explanation might be that both modes are only competitive with the mode car on short distances. High PT travel costs, which are anticipated to influence car probability positively, display no effect. Initially expected to have a minimal positive influence, age deviates from the linear pattern in the MNL model. High age almost does not influence the probability of car choice, while low age either significantly increases or reduces the probability.

Condensed, most features influence car choice as expected. However, some surprising outcomes can be observed, like only short pedestrian travel times affecting car choice or young individuals either preferring the mode car or not.

D. Preference heterogeneity

Preference heterogeneity manifests in various ways, highlighted in Figure 3a. High heterogeneity is observed in factors that consistently reduce the probability of choosing the car. Individuals experiencing low pedestrian and car travel times, high car travel costs and high traffic variability exhibit diverse preferences. The same goes for individuals not possessing a car or a driving license.

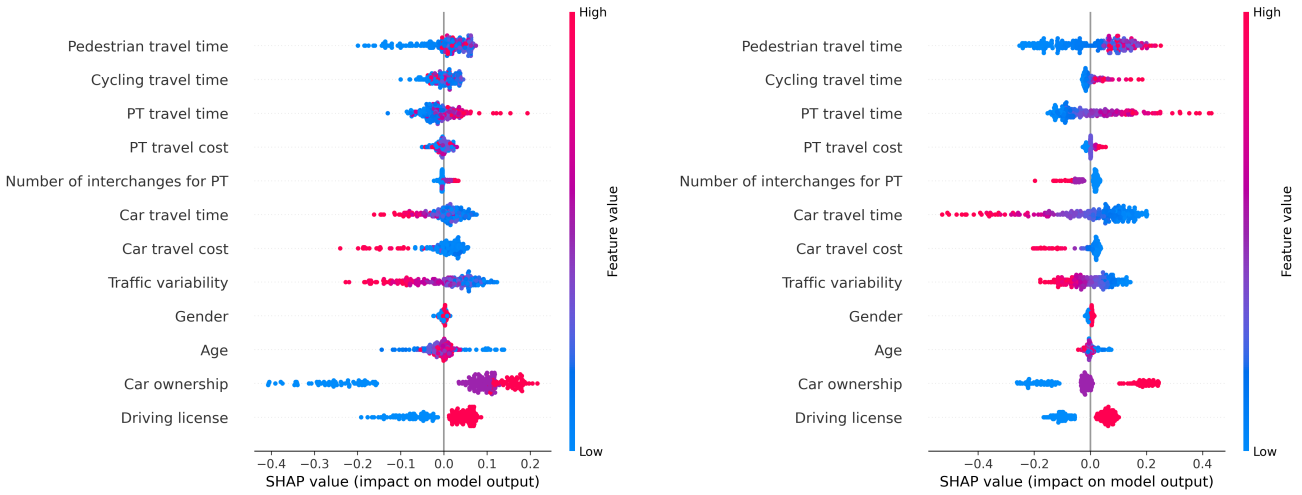
Age is highly heterogeneous, with some young individuals choosing the mode car often and others less. Furthermore, high heterogeneity is associated with high public transportation (PT) travel times, increasing the probability of opting for the mode car. Conversely, low heterogeneity mainly displays either positive effects or no effects on the probability of choosing the car. High car ownership, low congestion levels, and possessing a driver's license consistently affect car mode choice. In contrast, factors like gender, PT interchanges, and PT travel costs do not influence choosing the mode car.

The results suggest that features increasing car mode choice have mainly a homogeneous impact, whereas features reducing the probability of choosing the car are perceived heterogeneously. The only exception is age, with highly non-linear behaviour.

E. Non-linear effects

The examination of summary plots in Figure 3a also reveals insights into non-linear effects within variables, highlighting pedestrian travel times and age as potentially non-linear. In the following, features that influence the probability of choosing the mode car directly - as specified in the MNL model - will be analysed in more detail. As seen in Figure 4, car travel time and traffic variability exhibit linearity for the RF model, with potential deviations at high congestion levels. Conversely, reciprocal functions could be fitted to the importance of car ownership and driving license. These covariates exhibit in the RF model heterogeneity of importance for low feature values while demonstrating higher homogeneity values for higher feature levels.

Two clusters emerge for car travel costs in Figure 4b for the RF model. In the first cluster, an inverted quadratic function suggests a global maximum of importance for car travel costs. At this point, the mode car is most competitive with other modes. The second cluster is dominated by high travel costs, where heterogeneity is observed. An adverse structure explained by other covariates is observed in the case of age. While age does not influence car mode choice for most age groups, being below ten years notably increases the probability of using the mode car in the RF model. One plausible explanation is that parents often drive young children to kindergarten or other destinations. The mode car is chosen significantly less for individuals between 10 and 20 years old. This shift aligns with the assumption that their parents drive them less. More often, they now have to reach their destination



(a) Summary plot for the RF model

(b) Summary plot for the MNL model

Fig. 3: Comparison of Summary plots for the mode car

using other modes of transportation since they are not eligible for a driving licence yet.

In summary, significant non-linearity exists for the mode choice preferences of individuals in London. Age is highly non-linear for young ages. For car travel costs, clusters can be detected and owning a car and driving license can be described through reciprocal functions.

F. Interaction effects

Features do not influence car mode choice in isolation but interact with others. In SHAP, the main and interaction effects are intertwined. SHAP interaction values disentangle these effects and are highlighted in Figure 5 for the mode car. As mentioned in III, the results distilled from this analysis will be validated through interaction parameters in the MNL model.

As indicated by the SHAP feature importance table, the most influential variables were car ownership, traffic variability, and possession of a driving license. This can be confirmed by Figure 5, which indicates that most SHAP values provide a robust and accurate data representation. Generally, only limited interactions exist, with two exceptions. Specifically, interactions exist between owning a driving license and age and between owning a car and a driving license. For instance, owning a driving license changes the probability of choosing the mode car on average by 7.34%. If age had not been measured in the analysis, the importance of possessing a driving license changed by 2.64% points, and vice versa. Similarly, in the absence of car ownership, the importance of a driving license changes by 2.08% points.

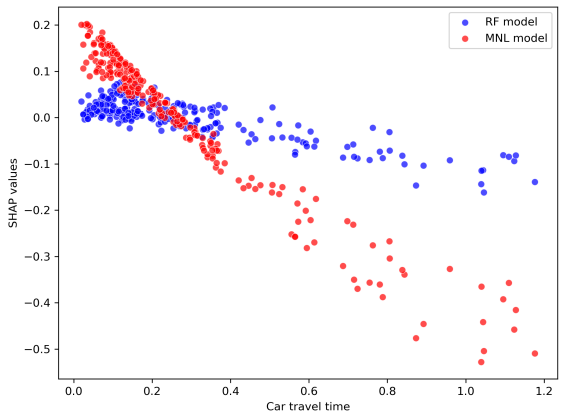
Furthermore, the highest relative interaction can be observed between cycling travel time and the travel times of other modes. On average, it changes the probability of choosing the mode car by 2.67%. When pedestrian travel time was not considered, the importance of cycling travel time changed by 1.92% points. Similarly, the absence of consideration for public transportation travel time resulted in a 1.24% change in the importance of cycling travel time.

G. SHAP validation

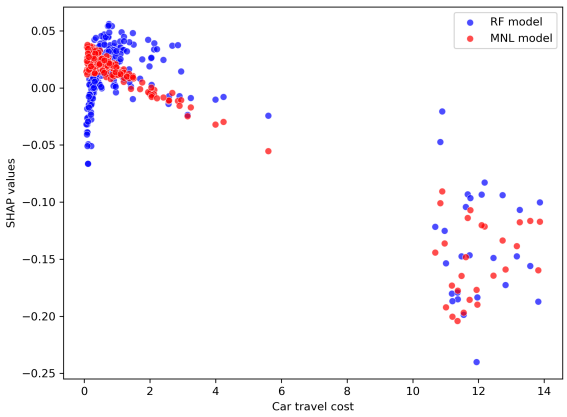
1) Feature Importance: The importance of features has turned out to be significantly different in the MNL model and a RF model, as it can be seen in Figure 2. Features diverge in their importance for overall values and particular modes. Across all modes, the RF model highlights car ownership and traffic variability as influential factors. In contrast, the MNL model places greater emphasis on pedestrian travel time, PT travel time, and car travel time. Travel times are the most important determinant for mode choice in the MNL model, whereas in the RF this is only valid for active modes. For example, PT travel times change the probability of using this mode by 16.01% in the MNL model compared to 2.51% in the RF model. Both models only agree on the unimportance of travel costs.

This underscores the importance of the underlying model for determining feature importance. Whereas the RF model predicts car ownership and traffic variability to be important, the MNL model puts more emphasis on travel times.

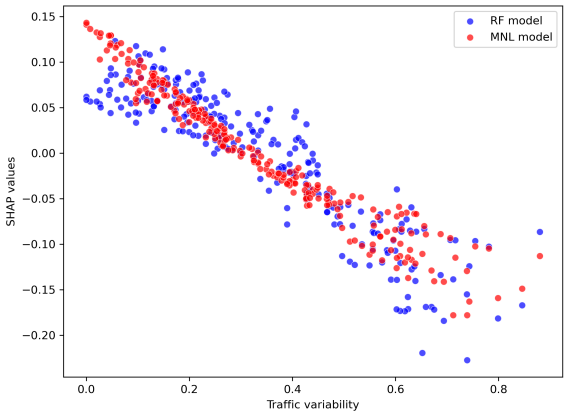
2) Feature signs: The MNL and RF models mostly agree on feature signs. Both models coincide with the effect of pedestrian travel time, car ownership and possessing a driving license as seen in Figure 3. Furthermore, they indicate that gender does not affect the probability of choosing the car mode. However, distinctions emerge for certain variables. Firstly, cycling travel time exhibits mixed effects in the RF model, while the MNL model indicates that longer cycling times lead to a higher probability of choosing the car mode. The MNL model sees thus greater competition between the mode cycling and car. Furthermore, the MNL model attaches a slight positive effect of age for lower age groups, whereas



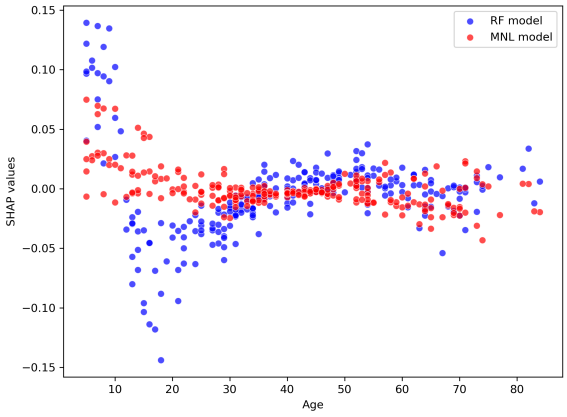
(a) Dependence plot for car travel time



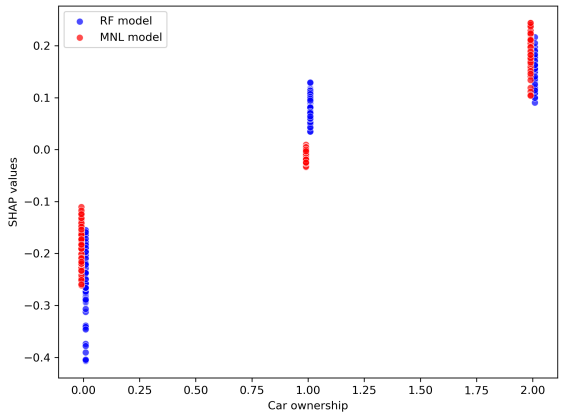
(b) Dependence plot for car travel cost



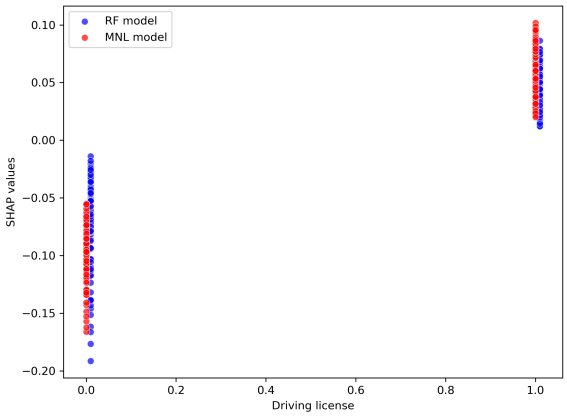
(c) Dependence plot for traffic variability



(d) Dependence plot for age



(c) Summary plot for car ownership



(d) Summary plot for driving license

Fig. 4: Dependence plots for the mode car

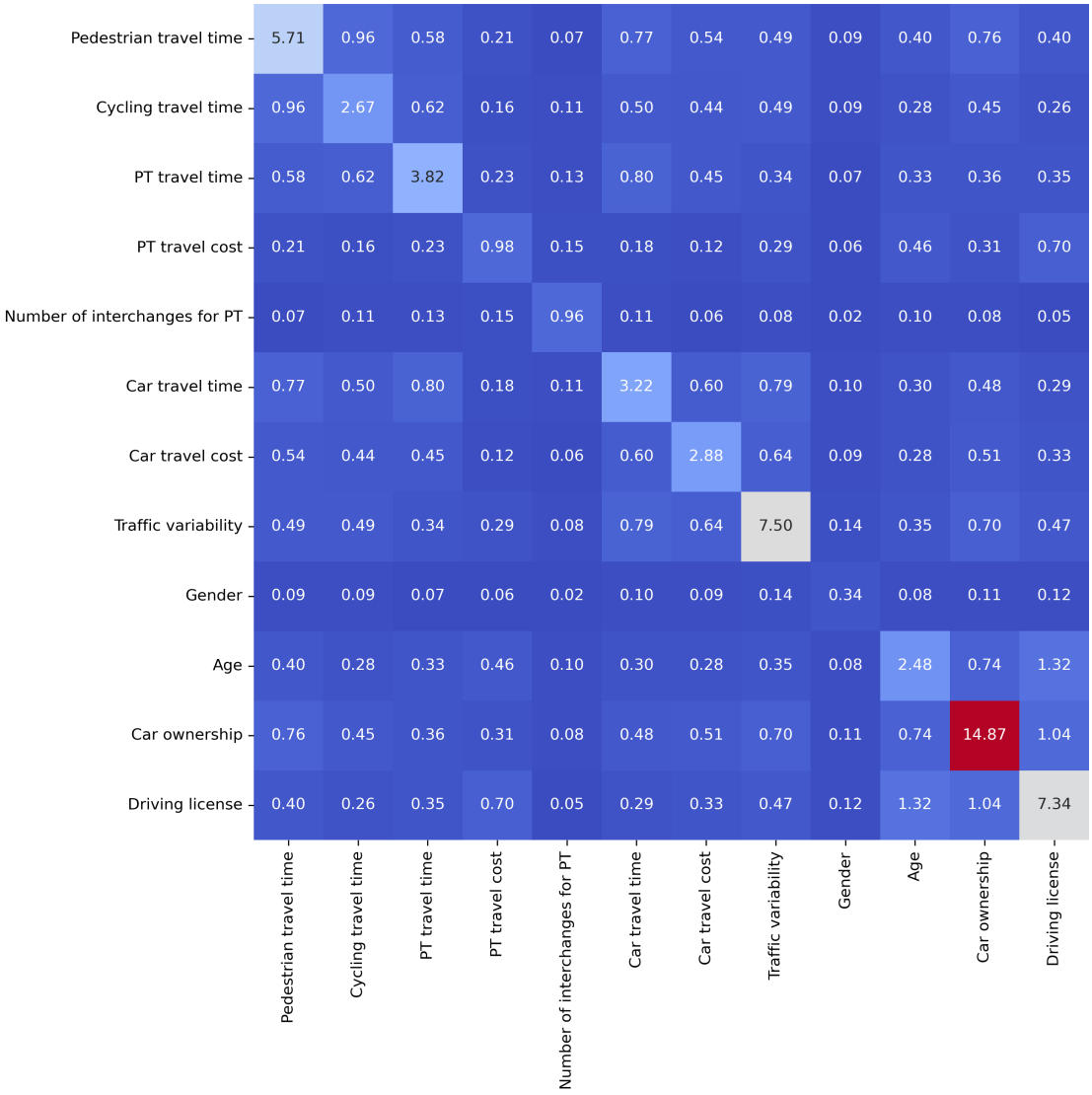


Fig. 5: Interaction plot for the mode car in %

the RF model detects positive and negative effects for young individuals.

While both models share common ground in predicting feature signs for most features, the MNL model-based SHAP values do not detect adverse effects like those observed for young people in the RF model. Secondly, the MNL model aligns more with the expectation of mode competition between biking and driving by car.

3) Preference heterogeneity: In line with the different importance attribution of features, preference heterogeneity is significantly different for underlining RF and MNL models. The general observation from the RF model does not hold. It states that feature values reducing car choice are heterogeneously distributed, and those raising the probability are homogeneously distributed. Pedestrian travel time is the only similar feature regarding preference heterogeneity in Figure 3. Other factors have either higher or lower preference heterogeneity. Firstly, the MNL model reveals higher preference heterogeneity associated with longer travel times for the mode car and PT. This might relate to the fact that travel times are generally more critical for the MNL model.

Conversely, the MNL model portrays lower heterogeneity for owning a car and high congestion levels. This might be related to these features being considered less important in the MNL model. These differences underscore the importance of the assumed underlying model when using SHAP for highlighting preference heterogeneity.

4) Non-linear effects: The RF and MNL models also do not coincide regarding their notion of non-linearity for car mode choice. In the MNL model, only linear utilities have been defined. The MNL based SHAP values are thus predominantly linear, with some deviations at the tails in Figure 3, being in contrast with the RF based SHAP values.

Firstly, the MNL model cannot capture the reciprocal re-

lationship observed in the RF model for owning a car and a driving license. Furthermore, the inverted quadratic function for low car travel costs, followed by a clustering pattern for high costs, is inadequately represented in the MNL model. The relation is simplified to a linear relationship for low costs and a cluster for high fees. Regarding age, the RF model reveals strong positive and negative effects for specific age groups. In distinction, the MNL model only detects positive effects for very low ages and misses the negative probability for choosing the mode car for people aged between 10 and 20 years.

In conclusion, the MNL model's assumption of linearity restricts its ability to discern non-linear relationships in the data.

5) Interactions: Interactions detected in the RF model and visualised in Figure 5 are difficult to verify because of the different notion of RF based SHAP interaction values and interaction parameters in a MNL model. In the SHAP formulation, interaction values are described as the probability change of the importance of one feature when another is not present. In MNL however, interaction parameter describe the importance of correlation between two factors.

For covariates like owning a car, a driving license or age, interaction parameters in the MNL model can indicate if the found SHAP are correct. Adding interactions between attributes of different alternatives contradicts the assumption of MNL. It states that alternatives are independent of each other and that attributes contribute to the utility of only one alternative. One possible way to test the interaction between pedestrian and cyclist travel time would be to leave out one factor in the utility specification and measure the difference between the parameter values.

Despite the different notions of interactions, the interactions between socio-demographics can be confirmed by the MNL model. Interactions between attributes of various alternatives are not possible to confirm, and in the MNL model, an additional interaction turned out to be significant - the one between age and travel costs. The results suggest that interactions coincide between models, even though not all interactions could be tested, and some minor discrepancies exist.

V. Discussion

A. Main findings

The analysis of transportation mode choice using SHAP values from an RF model unveils that travel times are the most relevant for active modes. In contrast, socio-demographics like owning a car or a driving license are most decisive for choosing PT and driving by car. The signs of features mainly align with expectations, and surprising outcomes emerge, such as only short pedestrian travel times influencing car choice. The results suggest that features increasing car mode choice have mainly a homogeneous impact, whereas features reducing the probability of choosing to drive by car are perceived heterogeneously. Nonlinear effects, including reciprocal functions and clustered patterns, can be perceived for certain features. Limited overall attribute interactions are perceived, with notable exceptions for the relation between owning a car, a driving license and age.

Attempts to validate findings through MNL-based SHAP values have highlighted significant disparities, exempting similar feature signs and interactions for both models. The absolute importance of features varies significantly between both modes; for example, the MNL model assigns a higher importance to travel times across all modes. Furthermore, high heterogeneity is also observed in the MNL based SHAP values that increase the probability of car choice, as opposed to the RF model. The MNL model's general linearity aligns with features but misses highly nonlinear effects observed in the RF model, particularly regarding age. The results highlight the importance of the underlying prediction model when using SHAP.

B. Implications for researchers and policymakers

The main implication for researchers refers to the interpretation of the different outcomes. SHAP values derived from RF and MNL reveal significant differences regarding overall feature importance, preference heterogeneity and non-linearity. Therefore, researchers must decide which model describes the data-generating process. The advantage of the RF model is that relations are learned from the data. Therefore, the RF model does not influence feature relations.

Furthermore, a better prediction performance suggests that the RF model is better suited to describe the data. If the RF model is assumed to represent the underlying data process better, additional insights into non-linearity and feature interactions are possible. However, the MNL model has the advantage of being widely used in choice modelling. Furthermore, utility maximisation is defined as a theory of behaviour, and lastly, the estimated parameters can be used for economic appraisal. SHAP then adds them to an MNL model with importance values and highlights preference heterogeneity.

The following implications emerge for policymakers assuming that RF-based SHAP values are used. Since car ownership, pedestrian travel time and traffic variability turned out to be the most significant factors for car mode choice, increasing parking fees and promoting car sharing might be the most feasible options to reduce car usage. A further conceivable measure might be facilitating mixed neighbourhoods, such as lowering distances between inhabitants and their destinations. However, both measures targeted reduced car ownership - increased parking fees and promoting car sharing - might not be as efficient as suggested through the absolute importance values.

Since there is significant heterogeneity regarding low car ownership values, measures might lead to only groups reducing car usage as aimed for and others to a lesser extent. The reciprocal relationship of car ownership's importance infers that incentives for lower car ownership will lead to many households switching from one or more cars per adult to one or fewer vehicles per adult, not leading to significant impacts. Furthermore, the highly nonlinear relationship for age should

not be underestimated, especially policies for families with children, which might make sense to reduce car usage.

C. Limitations and research recommendations

The study is subject to several limitations. The findings apply to London, and generalising them to other cities may prove challenging. Moreover, the computational constraints of KernelSHAP dictated the sample size of 250 individuals. The same sample size was used for TreeSHAP to derive comparable outcomes, raising concerns about the results' robustness and generalisability. Additionally, a significant sampling limitation arises in calculating interaction values, where only 50 samples were used.

The absolute importance, heterogeneity and non-linearity of features, expressed through SHAP values, differs depending on the underlying prediction model. Therefore, conclusions and policy decisions based on SHAP also depend on the underlying model. Further insights are needed on how robust SHAP values are relying on other prediction models of the choice modelling domain (mixed logit, random regret) as well as the machine learning domain (gradient boosted trees, neural networks). Answering this question would help to determine, firstly, which underlying models can highlight specific structures in the data, like heterogeneity or non-linearity. Secondly, it would be possible to answer for which underlying models SHAP are delivering additional insight.

The last limitation is more of a theoretical nature. Underlying prediction models assume certain relations between input and output data. For example, a MNL model assumes rational decision makers that compare utilities of the different alternatives. The alternative with the highest utility is chosen. SHAP also assumes a relation between input and output data, assuming all features collaborate in distributing the output or prediction probability. If SHAP is used with MNL, it might be the case that according to the MNL model, one feature is an attribute of an alternative, whereas in the SHAP model, this feature is a player. The implications of different notions of behaviour of features in both models remain yet to be uncovered.

References

[1] L. S. Shapley. "17. A Value for n-Person Games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by Harold William Kuhn and Albert William Tucker. Princeton University Press, Dec. 1953, pp. 307–318. ISBN: 978-1-4008-8197-0. DOI: 10.1515/9781400881970-018. URL: https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html (visited on 10/31/2023).

[2] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. Aug. 1995, 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994. URL: https://ieeexplore.ieee.org/document/598994 (visited on 10/27/2023).

[3] Michel Bierlaire. "BIOGEME: A Free Package for the Estimation of Discrete Choice Models". In: *Proceedings of the 3rd Swiss Transportation Research Conference*. Ascona, Switzerland, Mar. 2003. URL: http://biogeme.epfl.ch.

[4] Kenneth E Train. "EM algorithms for nonparametric estimation of mixing distributions". In: *Journal of Choice Modelling* 1.1 (2008). Publisher: Elsevier, pp. 40–69.

[5] R.S. Chalumuri et al. "Applications of Neural Networks in Mode Choice Modelling for Second Order Metropolitan Cities of India". In: *Proceedings of the Eastern Asia Society for Transportation Studies* 7 (2009). Type: Article, p. 134. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083693843&partnerID=40&md5=91d62df38af8ea351febae90cd6d0b71.

[6] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 06/11/2022).

[7] Alex Goldstein et al. "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation". In: (2013). Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1309.6392. URL: https://arxiv.org/abs/1309.6392 (visited on 09/28/2022).

[8] Michael Keane and Nada Wasi. "Comparing alternative models of heterogeneity in consumer choice behavior". In: *Journal of Applied Econometrics* 28.6 (2013). Publisher: Wiley Online Library, pp. 1018–1045.

[9] Xiaojing Dong and Frank S Koppelman. "Comparison of continuous and discrete representations of unobserved heterogeneity in logit models". In: *Journal of Marketing Analytics* 2 (2014). Publisher: Springer, pp. 43–58.

[10] Sebastian Bach et al. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". en. In: *PLOS ONE* 10.7 (July 2015). Ed. by Oscar Deniz Suarez, e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140. URL:

https://dx.plos.org/10.1371/journal.pone.0130140 (visited on 06/12/2022).

[11] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. event-place: San Francisco, California, USA. New York, NY, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: (2016). Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.1602.04938. URL: https://arxiv.org/abs/1602.04938 (visited on 06/12/2022).

[13] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[14] Tim Hillel, Mohammed Z E B Elshafie, and Ying Jin. "Recreating passenger mode choice-sets for transport simulation: A case study of London, UK". en. In: *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction* 171.1 (Mar. 2018), pp. 29–42. ISSN: 2397-8759. DOI: 10.1680/jsmic.17.00018. URL: https://www.icevirtuallibrary.com/doi/10.1680/jsmic.17.00018 (visited on 06/09/2022).

[15] Will Koehrsen. *Hyperparameter Tuning the Random Forest in Python*. en. Jan. 2018. URL: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74 (visited on 10/31/2023).

[16] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. "Consistent Individualized Feature Attribution for Tree Ensembles". In: (2018). Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.1802.03888. URL: https://arxiv.org/abs/1802.03888 (visited on 10/31/2023).

[17] Sander van Cranenburgh. "Blending computer vision into discrete choice models". en. In: (2020).

[18] Susanne Dandl et al. "Multi-Objective Counterfactual Explanations". en. In: *Parallel Problem Solving from Nature – PPSN XVI*. Ed. by Thomas Bäck et al. Vol. 12269. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 448–469. ISBN: 978-3-030-58111-4. DOI: 10.1007/978-3-030-58112-1_31. URL: http://link.springer.com/10.1007/978-3-030-58112-1_31 (visited on 09/28/2022).

[19] Scott M. Lundberg et al. "From local explanations to global understanding with explainable AI for trees". en. In: *Nature Machine Intelligence* 2.1 (Jan. 2020). Number: 1 Publisher: Nature Publishing Group, pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9. URL: https://www.nature.com/articles/s42256-019-0138-9 (visited on 10/28/2023).

[20] Sandeep Ram. *Mastering Random Forests: A comprehensive guide*. en. Oct. 2020. URL: https://towardsdatascience.com/mastering-random-forests-a-comprehensive-guide-51307c129cb1 (visited on 11/13/2023).

[21] Ahmad Alwosheel, Sander van Cranenburgh, and Caspar G. Chorus. "Why did you predict that? Towards explainable artificial neural networks for travel demand analysis". en. In: *Transportation Research Part C: Emerging Technologies* 128 (July 2021), p. 103143. ISSN: 0968-090X. DOI: 10.1016/j.trc.2021.103143. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001613 (visited on 06/09/2022).

[22] Nadia Burkart and Marco F. Huber. "A Survey on the Explainability of Supervised Machine Learning". In: *Journal of Artificial Intelligence Research* 70 (Jan. 2021), pp. 245–317. ISSN: 1076-9757. DOI: 10.1613/jair.1.12228. URL: https://jair.org/index.php/jair/article/view/12228 (visited on 10/04/2022).

[23] Yiling Deng. "Application of machine learning with a surrogate model to explore seniors' daily activity patterns". en. In: *Transportation Letters* (Aug. 2021), pp. 1–11. ISSN: 1942-7867, 1942-7875. DOI: 10.1080/19427867.2021.1969169. URL: https://www.tandfonline.com/doi/full/10.1080/19427867.2021.1969169 (visited on 10/05/2022).

[24] Tobias Huber et al. "Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps". en. In: *Artificial Intelligence* 301 (Dec. 2021), p. 103571. ISSN: 00043702. DOI: 10.1016/j.artint.2021.103571. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370221001223 (visited on 10/05/2022).

[25] Eui-Jin Kim. "Analysis of Travel Mode Choice in Seoul Using an Interpretable Machine Learning Approach". en. In: *Journal of Advanced Transportation* 2021 (Mar. 2021). Ed. by Inhi Kim, pp. 1–13. ISSN: 2042-3195, 0197-6729. DOI: 10.1155/2021/6685004. URL: https://www.hindawi.com/journals/jat/2021/6685004/ (visited on 10/05/2022).

[26] Yan Liu et al. "Dynamic activity chain pattern estimation under mobility demand changes during COVID-19". en. In: *Transportation Research Part C: Emerging Technologies* 131 (Oct. 2021), p. 103361. ISSN: 0968090X. DOI: 10.1016/j.trc.2021.103361. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X21003636 (visited on 10/05/2022).

[27] Janak Parmar, Pritikana Das, and Sanjaykumar M. Dave. "A machine learning approach for modelling parking duration in urban land-use". en. In: *Physica A: Statistical Mechanics and its Applications* 572 (June 2021), p. 125873. ISSN: 03784371. DOI: 10.1016/j.physa.2021.125873. URL: https://linkinghub.elsevier.

com / retrieve / pii / S037843712100145X (visited on 10/07/2022).

[28] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[29] Sander van Cranenburgh et al. "Choice modelling in the age of machine learning - Discussion paper". en. In: *Journal of Choice Modelling* 42 (Mar. 2022), p. 100340. ISSN: 1755-5345. DOI: 10.1016/j.jocm.2021.100340. URL: https://www.sciencedirect.com/science/article/pii/S1755534521000725 (visited on 06/09/2022).

[30] José Ignacio Hernandez. "Manuscript SHAP Covid". In: (2022).

[31] Gulnazbanu Saiyad, Minal Srivastava, and Dipak Rathwa. "Exploring determinants of feeder mode choice behavior using Artificial Neural Network: Evidences from Delhi metro". en. In: *Physica A: Statistical Mechanics and its Applications* 598 (July 2022), p. 127363. ISSN: 03784371. DOI: 10.1016/j.physa.2022.127363. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378437122002837 (visited on 06/11/2022).

[32] Yiming Zhang, Ying Weng, and Jonathan Lund. "Applications of Explainable Artificial Intelligence in Diagnosis and Surgery". en. In: *Diagnostics* 12.2 (Jan. 2022), p. 237. ISSN: 2075-4418. DOI: 10.3390/diagnostics12020237. URL: https://www.mdpi.com/2075-4418/12/2/237 (visited on 10/31/2023).

[33] Elisabetta Cornago, Alexandros Dimitropoulos, and Walid Oueslati. "The impact of urban road pricing on the use of bike sharing". en. In: *Journal of Environmental Economics and Management* 120 (July 2023), p. 102821. ISSN: 00950696. DOI: 10.1016/j.jeem.2023.102821. URL: https://linkinghub.elsevier.com/retrieve/pii/S0095069623000396 (visited on 10/31/2023).

[34] Dandan Fan et al. "Effects of congestion charging and subsidy policy on vehicle flow and revenue with user heterogeneity". In: *Mathematical Biosciences and Engineering* 20.7 (2023), pp. 12820–12842. ISSN: 1551-0018. DOI: 10.3934/mbe.2023572. URL: http://www.aimspress.com/article/doi/10.3934/mbe.2023572 (visited on 10/31/2023).

[35] Aisha Hasan et al. "Transit Behaviour and Sociodemographic Interrelation: Enhancing Urban Public-Transport Solutions". en. In: *Eng* 4.2 (Apr. 2023), pp. 1144–1155. ISSN: 2673-4117. DOI: 10.3390/eng4020066. URL: https://www.mdpi.com/2673-4117/4/2/66 (visited on 10/31/2023).

[36] Maarten Kroesen. *Modelling Paradigms*. Feb. 2023.

[37] Great Learning Team. *Random forest Algorithm in Machine learning: An Overview*. en-US. June 2023. URL: https://www.mygreatlearning.com/blog/random-forest-algorithm/ (visited on 10/27/2023).

[38] Kumar Dheenadayalan and Sumant Kulkarni. *Explainable Artificial Intelligence (XAI) and its real-world applications*. en. URL: https://www.zensar.com/insights/blogs/explainable-artificial-intelligence-xai-and-its-real-world-applications/ (visited on 10/31/2023).

[39] Christoph Molnar. *Interpretable Machine Learning*. URL: https://christophm.github.io/interpretable-ml-book/ (visited on 06/18/2022).

[40] Sharoon Saxena. *Random Forest Hyperparameter Tuning in Python — Machine learning*. URL: https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/# (visited on 10/31/2023).