# Mode Choice Modelling with Machine Learning: A Sequential Tour-based Approach for Addressing Imbalanced Datasets

**Dimitrios Pappelis**
MaaSLab, Energy Institute, Bartlett School of Environment, Energy and Resources
University College London, United Kingdom, WC1H 0NN
Email: d.pappelis.19@ucl.ac.uk

**Emmanouil Chaniotakis**
MaaSLab, Energy Institute, Bartlett School of Environment, Energy and Resources
University College London, United Kingdom, WC1H 0NN
Email: m.chaniotakis@ucl.ac.uk

**Maria Kamargianni**
MaaSLab, Energy Institute, Bartlett School of Environment, Energy and Resources
University College London, United Kingdom, WC1H 0NN
Email: m.kamargianni@ucl.ac.uk

Word Count: 6,256 words + 3 tables (250 words per table) = 7,006 words

*Submitted [01/08/20]*

**ABSTRACT**

The continuous progress of machine learning has introduced numerous powerful classifiers that are examined as prominent alternatives to predict travellers' mode choices. However, most classifiers fail to capture the lower market share that characterizes the minority modes of transport. Although imbalanced choice datasets are common, this has been more apparent with the emergence of new modes and mobility services, which further fragment the mode choice composition. The problem is often magnified by biased sampling and measurement errors during the data collection process. The challenge of imbalanced classification in machine learning is subject of continuous multidisciplinary research, however its extensions in mode choice modelling, remain relatively unexplored. This paper provides empirical evidence of the effect that dataset imbalance might have on prediction measures and proposes a sequential tour-based framework for addressing skewed travel diary data. The framework is applied on a dataset from the city of Thessaloniki, Greece with a total of 5646 trips, using extreme gradient boosting (XGBoost). A set of performance metrics are used for the evaluation of the developed model and the output predictions are interpreted with partial dependence plots and state-of-the-art SHAP (SHapley Additive exPlanations) based on cooperative game theory. The results indicate that incorporating sequential effects can significantly improve the model's overall performance, especially with regards to recognition rates for the minority mode, without inducing bias within the trained classifier.

Key words: Mode Choice, Machine Learning, Classification, Imbalanced, Decision Trees, XGBoost, SHAP

## INTRODUCTION

Over the past decade, the flexible structure that characterizes advanced machine learning algorithms and their ability to process complex datasets, have motivated continuous research regarding their utilization in large-scale transport applications *(1)*. However, a recurring point of debate is the trade-off between their high predictive accuracy and lack of explainability or internal logic interpretation. This challenge is relevant to travel behavior modelling, as accurate demand forecasts only partially cover the question of interest. Understanding the causation motivating people's daily travel choices is critical to optimally design transport infrastructure (e.g. bus stops, bike lanes) and target social norms influencing behavioral trends. Furthermore, the design of deep 'black box' models entails the risk of misguided decision making, based on spurious correlations and artifacts in the training dataset *(2)*.

Within the context of mode choice modelling, this risk is magnified due to the imbalanced nature of transport data and information sources. The emerging mobility services have introduced new alternatives to be considered in everyday travel decisions, resulting in an increasingly fragmented market. As a result, it is a common paradigm in revealed preference data collection efforts, for some modes to receive much less observations than others or even be underrepresented. The presence of 'dominating' classes (such as car or public transport) versus the low market shares of minority modes (such as cycling or ridesharing) create an imbalance within the datasets used for evaluation. The inherent imbalance of the problem is often magnified by biased sampling and measurement errors during the data collection process. This creates a challenge for traditional machine learning algorithms that tend to provide biased predictions favouring the majority classes, as their design and evaluation is based on accuracy and its complement error rate *(3)*.

This paper aims to address these limitations by proposing a sequential tour-based approach for mode choice modelling with skewed travel diary data, to increase recognition rates for the minority mode and overall predictive accuracy. The extreme gradient boosting algorithm (XGBoost) is selected for the evaluation of the modelling concept. The proposed framework is applied on a revealed preference (RP) study from the city of Thessaloniki, Greece. Model performance is assessed with a set of prediction metrics, while sophisticated explanation methods are investigated to account for the opaque nature of the ensemble model. More specifically, this paper aims to contribute to existing literature on machine learning for mode choice in the following ways,

1. We provide empirical evidence on the effect that imbalanced datasets might have on prediction measures of classical machine learning approaches (e.g. decision trees) for mode choice.
2. We propose a sequential tour-based approach for increasing predictive accuracy and alleviating class imbalance in travel diary data without inducing bias in the classifier structure.
3. We apply and evaluate the proposed framework using extreme gradient boosting (XGBoost) on a case study from the city of Thessaloniki, Greece.
4. We interpret the 'black box' model predictions with partial dependence plots and state-of-the-art SHAP (Shapley Additive exPlanations) based on cooperative game theory.

We proceed as follows; Section 2 provides the background on imbalanced classification techniques and their applications within transportation modelling. Section 3 presents the dataset and provides summary statistics. Section 4 provides the design of the sequential tour-based modelling framework. Section 5 presents the results and the interpretation of the model output. Finally, we provide conclusions and future work in Section 6.

## BACKGROUND

The imbalanced classification problem is prevalent in numerous tasks such as rare disease diagnosis *(4)*, fraudulent transactions *(5)*, text recognition *(6)*, and many others across multidisciplinary fields of study. This ongoing field of research is relevant to transportation and particularly mode choice modelling, to evaluate the introduction of emerging mobility services which are currently underrepresented in everyday commuting. However, imbalanced classification with machine learning in transportation modelling is still relatively unexplored. Many solutions have been proposed to address imbalanced datasets in the predictive modelling problem of classification (assigning labels to a number of observations), with the four most prominent to be i) Enhanced data collection; ii) Data resampling; iii) Cost-sensitive Learning; and iv) Boosting.

In enhanced data collection, a larger dataset provides a balanced overview on the class frequency and is useful for the application of resampling techniques *(7)*. Addressing class imbalance at the data collection level is one of the solutions commonly found in previous transportation literature. The predictive ability of the disaggregate mode choice models by Wilson et al. was reduced due to the low representation of the bus and rail modes within the dataset *(8)*. To address this effect, they suggested the conducting of on-board surveys to enrich the estimation sample with observations on the less used modes; similarly, Nitsche et al. *(9)* enhanced underrepresented transport modes with further data collected to improve the accuracy of their models.

The most applied solution to an imbalanced classification problem is to modify the composition of the training dataset (data resampling). It is an attempt to balance the class frequencies at the dataset level. There are two standard sampling methods that can be used: a) oversampling, replicating minority class examples and b) undersampling, discarding majority class examples. For the latter, although training time is decreased, the main drawback is the loss of information that comes with deleting examples from the training data *(10)*. On the other hand, when oversampling, no information is lost as the resampled training set contains all instances from the original dataset. Oversampling can be performed either by including duplicate or adding new minority class examples. The main drawback when duplicating examples, is the higher risk of overfitting *(11)*. Typically, sampling is only performed on the training dataset and not on the holdout set, to evaluate the resulting model on representative data of the target problem domain. Past research in this area includes random oversampling or under sampling, synthetic sampling with data generation, cluster-based sampling methods etc. *(12)*. The SMOTE oversampling algorithm was applied by Chang et al. for vehicle classification on an imbalanced dataset from a single magnetic sensor *(13)*.

In cost-sensitive learning, a learner is modified at the algorithmic level to incorporate varying penalty for each of the classes under consideration *(14)*. This solution addresses the assumption made by most machine learning classifiers, that the misclassification costs are equal among classes. In most real-world applications however, this assumption is not valid *(15)*. The cost of classifying an example incorrectly is typically greater than the cost of labelling it correctly. For instance, it is rational to reject a suspicious credit transaction, even if it is highly likely to be legitimate *(16)*. Thus, the implementation of cost-sensitive learning shifts the problem scope from accuracy optimization to the minimization of the total misclassification cost *(12)*. In previous transportation research, Tang et al. proposed a method for mode-switching decision tree induction that incorporates loss matrix selection, aiming to mitigate the classifier's difficulty in identifying the minority class *(17)*.

Finally, the concept of boosting is based on the observation that identifying many underlying rules is more feasible than a single accurate prediction rule *(18)*. In each iteration, a 'weak' learner (e.g. decision tree) runs over a different distribution (or weighting) of the training examples. The contribution of all sequentially built weak classifiers represents the model's predictions. As the underrepresented class instances are more likely to be misclassified, this is addressed in subsequent iterations towards the minimization of past errors, making this technique appropriate for alleviating class imbalance. In practical applications, the ensemble learning approach of tree boosting was applied by Chen et al. in a study on ridesplitting behavior for on-demand ride services *(19)*.

XGBoost is a state-of-the-art, scalable, open-source machine learning system based on the concept of tree boosting. Each decision tree 'learns' from the previous within the sequence, building towards an overall strong learner *(20)*. It has received wide acknowledgement for a series of winning performances in Kaggle competitions for machine learning applications *(21)*. One of its main advantages is the parallelizable nature of the core algorithm, granting both speed and scalability for training on large datasets. Within transportation, XGBoost is gradually rising in research interest and was selected for the scope of this paper. Wang and Ross concluded on a higher predictive accuracy of the XGBoost model compared to the Multinomial Logit Model (MNL) for mode choice, even though both models underperformed in unbalanced datasets *(22)*. Parsa et al. also applied the XGBoost algorithm to detect the occurrence of highway accidents using real time data *(23)*.

## DATA ANALYSIS

The primary data source for our models is a survey from Thessaloniki (Greece) conducted in 2014, based on individual travel diaries. The participants were asked to state their modes of transport and activities. The resulting dataset, after data cleaning and removal of missing values, consists of 2,610 individuals and a total of 5,646 trips. The trips were enriched with variables from Google Maps on distance and historical travel times for each alternative mode. The resulting dataset consists of 28 variables including both individual, household characteristics and trip-related attributes (Table 1).

Descriptive statistics graphs are provided in Figure 1. The mode choice set includes: 1.Car, 2.Bus, 3.MrC (motorcycle) and 4.Walk. As illustrated, there are significantly dominating modes and motorcycle (MrC) is present in only a few of the population classes. Hence, it is apparent that travelling using motorcycle is an underrepresented mode, accounting for only 171 trips in the whole dataset. This minority mode is not efficiently captured using basic machine learning techniques. For the experiments, the dataset was split into a training set (80% of the total instances) and a testing set (20% remaining instances) to evaluate the performance of the machine learning algorithms. A representative distribution of all classes was accounted for in the stratification. The models were implemented in Python 3.6 with the Scikit-learn machine learning module for medium-scale supervised and unsupervised problems *(24)*.

A key consideration in the selection of the XGBoost algorithm was the observed multicollinearity between variables. In fact, the historical travel times and distance variables from the Google APIs (Table 1) are naturally correlated. Therefore, the selected algorithm needs to account for such relationships between variables, to obtain valid predictions on feature importance. Decision trees address multi-collinearity to a great extent, as each split is determined on only one of the correlated features. In addition, within tree ensemble methods such as XGBoost, once a dominant feature has been learnt, the algorithm will minimize the complement error rate for future iterations, thus assigning a higher importance to only one of the correlated features. This is an important advantage of ensemble algorithms and particularly those that utilize decision trees.

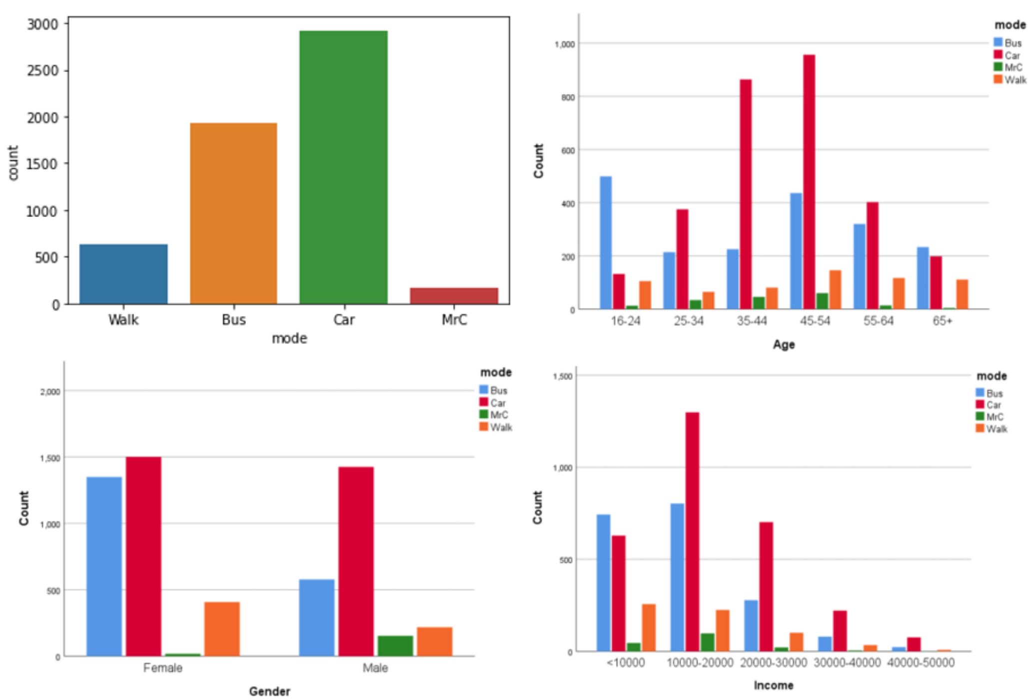


**Figure 1 Descriptive Statistics for Thessaloniki Dataset**

**TABLE 1 Independent variables for Thessaloniki dataset**

| Name | Type | Description |
|---|---|---|
| **Individual** | | |
| Gender | Boolean | Declared gender |
| Age | Categorical - 6 classes | Declared age |
| Income | Categorical - 5 classes | Declared income |
| Occupation | Categorical - 6 classes | Current occupation |
| Education | Categorical - 7 classes | Level of education |
| Driver’s License | Boolean | Ownership of driver license |
| **Household** | | |
| Household Size | Numerical – discrete | Total household size |
| Household over 16 | Numerical | Members over the age of 16 |
| Car Availability | Boolean | Availability of car for use |
| **Trip/Activity** | | |
| Start Time | Numerical - continuous | Declared trip start time |
| Duration | Numerical | Declared trip duration |
| Trip Day | Categorical - 7 classes | Day of the week |
| Start Activity | Categorical - 16 classes | Activity at origin location |
| End Activity | Categorical - 16 classes | Activity at destination location |
| Origin Location | Categorical - 11 classes | Municipality of trip origin |
| Destination Location | Categorical - 11 classes | Municipality of destination |
| **Google APIs** | | |
| Car Travel Time | Numerical - continuous | Historical travel time by car (seconds) |
| Car Distance | Numerical | Shortest travel distance by car (meters) |
| Bus Travel Time | Numerical | Historical travel time by bus |
| Bus Distance | Numerical | Shortest travel distance by bus |
| Bus Access Walk Time | Numerical | Access from origin walk time |
| Bus Access Walk Distance | Numerical | Access from origin walk distance |
| Bus Egress Walk Time | Numerical | Egress to destination walk time |
| Bus Egress Walk Distance | Numerical | Egress to destination distance |
| MrC Travel Time | Numerical | Historical travel time by MrC |
| MrC Distance | Numerical | Shortest travel distance by MrC |
| Walk Travel Time | Numerical | Historical travel time walking |
| Walk Distance | Numerical | Shortest travel distance walking |

## MODEL DEVELOPMENT

### Base case: Trip-based approach

For the base case, we applied a single decision tree, a common algorithmic approach that identifies rules to split a dataset based on different conditions. Decision Tree algorithms have been extensively investigated in literature *(25)*, characterized from their explainable structure. In the base-case scenario the individual’s sociodemographics, household and trip-specific variables are included, in combination with the historical distance and travel time values extracted from the Google APIs. Therefore, this is a solely trip-based approach, as the model is processing each travel diary instance separately, without assuming any dependence or correlation with previous trips. As expected, we received ‘naive’ results due to the strong imbalance within the dataset. The MrC instances were not captured by the classifier, providing biased predictions in favor of Car as the majority class. The initial training of the Extreme Gradient Boosting algorithm (XGBoost) improved the performance for two main modes (Car, Walk) but was also not able to capture the minority mode MrC. Figure 2 presents the confusion matrix for the base-case decision tree and XGBoost classifier on the validation dataset.

A classical approach to address the classification problem at this stage would be cost-sensitive learning. Nonetheless, this method entails certain limitations and was thus not considered for the scope of this study. Firstly, there is no insight available on the cost matrix during classifier training. This is important as the successful application of cost-sensitive learning relies on the accurate estimation of the supplied cost matrix *(26)*. A common, though heuristic, choice of assigning misclassification cost could be based on the inverse class distribution. In addition, the scope of the model under development is not an increase of the Area Under Curve (AUC) for a specific minority class of interest (MrC), but rather the design of a model that will accurately predict the probabilities for all modes, regardless of their occurrence. Therefore, applying cost-sensitive classification would bias the predictions of the classifier, increasing recall rates for the minority class at the cost of predictive accuracy for the majority modes. As a result, a more sophisticated approach is required to efficiently capture the mode choice decisions without inducing classification bias in the model output towards the minority mode.

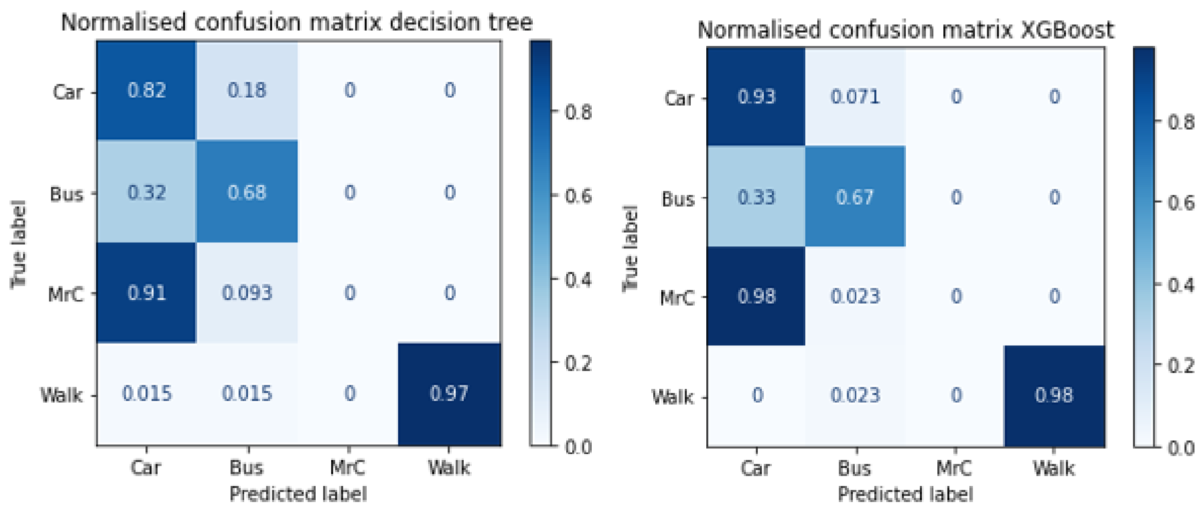


**Figure 2 Normalized confusion matrix for base-case models on imbalanced dataset**

**Extension: Sequential Tour-based Approach**

Travel diaries offer a variety of useful information for designing comprehensive travel behavior models. The value of this information has led to recent research on alternative ways to collect travel diaries (smartphones, tracking location devices etc.), as response rates for the classical data collection methods have decreased significantly *(27)*. In previous data driven approaches, Chang *(28)* used travel diary data to evaluate a fusion model using machine learning methods for mode choice. A main advantage of the travel diary structure is the ability to depict the time dependent ordering of the trips undertaken by the participants within a tracked day. In this study, we utilize this feature to design a sequential tour-based approach for addressing imbalanced mode classification.

The feature engineering technique proposed enables us to capture various nonlinear interactions and daily choice dependencies within the input dataset. For instance, in case the first trip of the day is undertaken by car, there is a high probability for subsequent trips to be affected by this initial choice. In the final trip of the day, the car is most likely to be returned home from the same individual. Furthermore, if an individual has selected a specific mode for a cyclical route (e.g. Bus), this might reveal a strong preference of Bus for the return trip to the original destination. Such a preference could be affected by various factors such as access and egress walking times, service frequency etc. To account for the majority of these tour-based effects and dynamic dependencies, we can restructure the travel diary in a way so that a subset or all variables from previous trips on a given day, are transferred in the remaining trips of the individual's schedule. Ultimately, our goal is to improve predicting performance on the 'naïve' base case scenario and efficiently capture the underrepresented mode, without including resampling or cost-sensitive bias in the model, as it will be 'informed' on MrC ownership and usage by specific individuals from their previous daily choices. Rashidi et al. *(29)* applied a similar sequential concept using the random forest method, to capture the serial correlation between trips towards a disaggregate travel demand modelling structure. A limitation of our proposed feature engineering technique on the training dataset is the increase of dimensionality for our problem. After a series of evaluation attempts, we decided upon the sequential inclusion of the last two trips for every individual. Therefore, this modification can be viewed as a dynamic 3-step memory horizon addition to the model. The modelling framework accounting for tour-based effects is depicted in Figure 3.

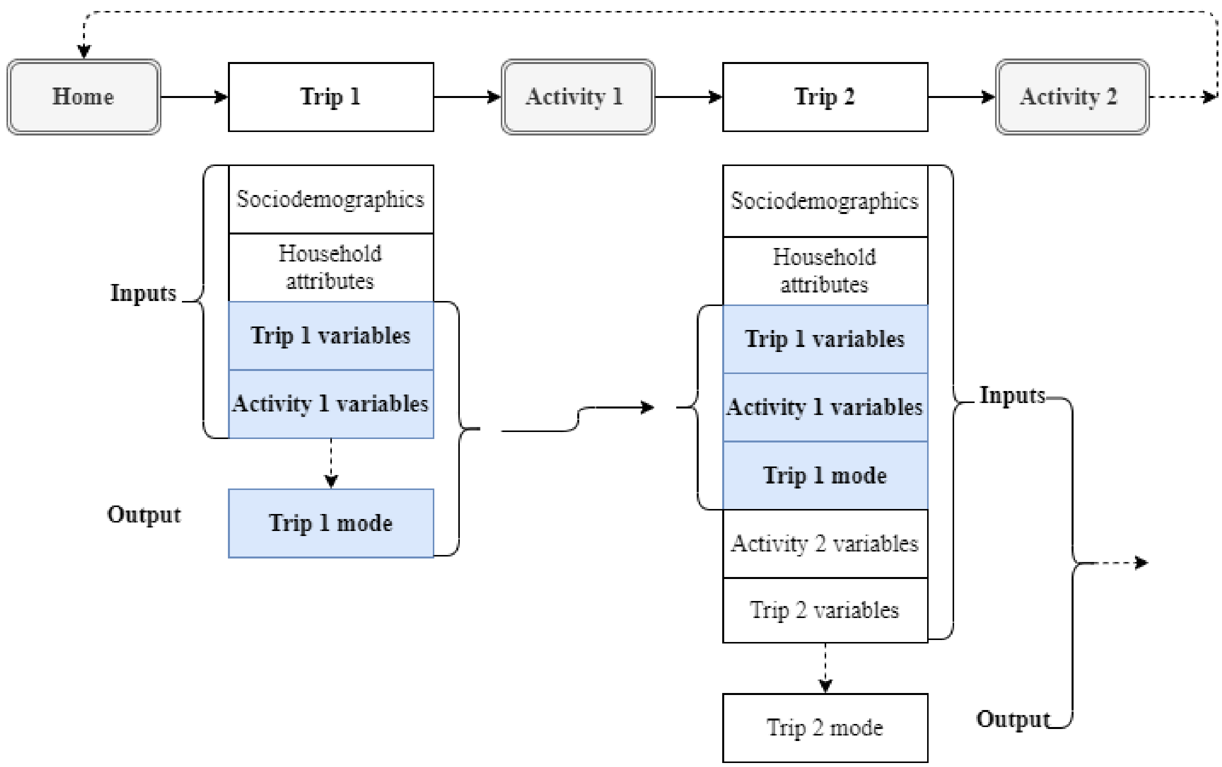


**Figure 3 Sequential modelling framework for tour-based effects**

The first step towards the evaluation of the sequential, tour-based approach was to apply it on an explainable machine learning algorithm, a single 3-level decision tree. The structure of the tree is depicted in Figure 4. The average shortest distance is a key factor in the tree structure, with a threshold of <800m for the classification of walking instances. In the second level, the tour-based effects become apparent, as the split criterion is based on the previous mode choice, further classifying the longer trip modes (Car, Bus, MrC). Finally, the ownership of a driver's license is the main factor determining the choice of Car over Bus usage. It is important to note that this modification allowed for the minority MrC to be captured to a basic extent. More information on the decision tree performance can be found in the Results section.

As the decision tree identified basic trip dependencies and minority mode instances at a satisfactory level, the next step is to apply this framework using the advanced extreme gradient boosting algorithm (XGBoost). Compared to single decision trees, the ensemble model is expected to offer higher predictive accuracy and overall performance. The advantage of using trees, however, is that their structure easily explains how each specific prediction is made (Figure 4). As the gradient boosting machine (GBM) fits numerous shallow trees in a stage-wise fashion, we are optimizing the residuals, thus lacking the intuitive interpretability provided by a single tree. The performance of the developed models is discussed in the following section.

## RESULTS

With regards to comparing the results, it should be noted that especially in imbalanced classification modelling, accuracy is not a proper evaluation measure, often referred to as the ‘accuracy paradox’, as it may lead to erroneous conclusions *(30)*. Therefore, we selected the following performance measures that give more insight into the performance of the model, based on the confusion matrix,

i) Recall or sensitivity, the ability of a classification model to identify all the relevant data points within the dataset,
ii) Precision, the ability of a classification model to identify only the relevant cases,
iii) F-measure, the weighted harmonic mean of the test's precision and recall, ranging between 0 and 1;
iv) Balanced accuracy, the average of sensitivity and specificity as computed for each class and averaged over the total number of classes *(31)*,
v) Cohen's Kappa, a measure of interrater reliability (or interobserver agreement). Values <0 indicate no agreement, 0.01-0.20 none to slight, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and 0.81-1.00 an almost perfect agreement *(32)*.

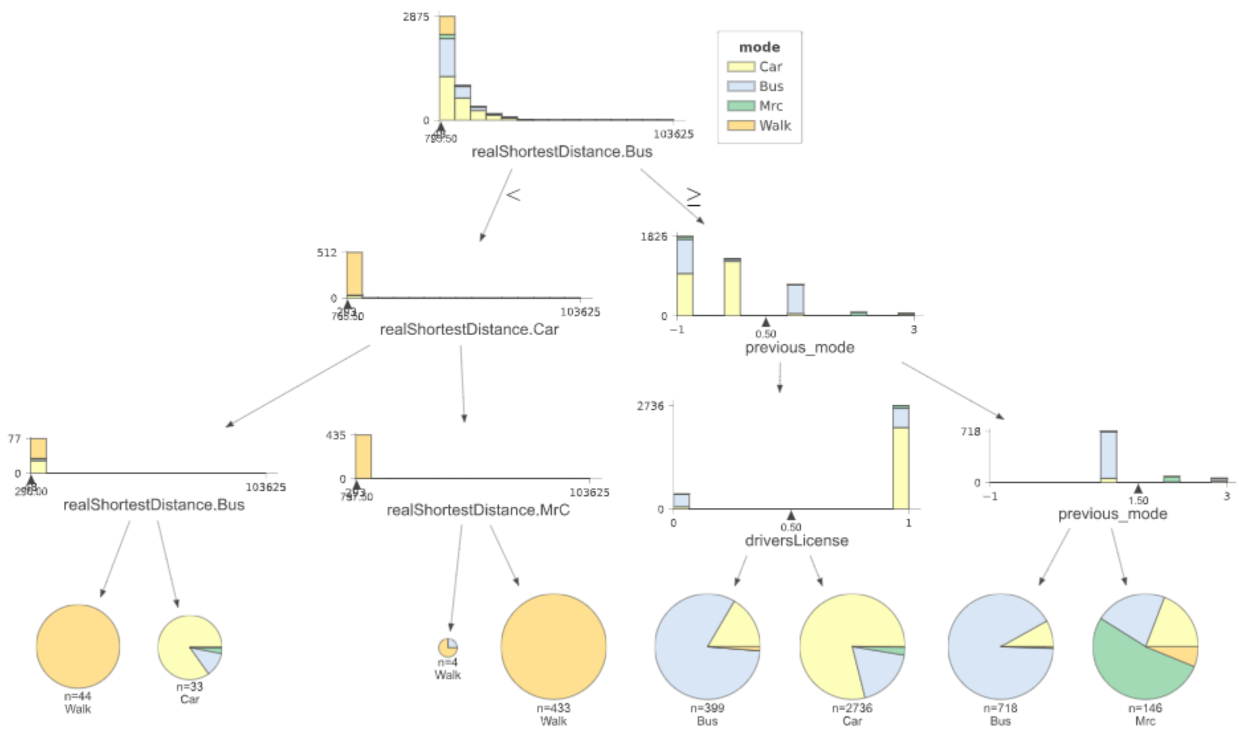


**Figure 4 Shallow Decision Tree with dynamic trip dependency**

The performance of the tour-based Decision Tree model on the validation set of 1,129 trips is summarized in Table 2. Compared to the base case scenario, there is a significant increase in performance, which is a good reference point considering the structure’s explainability. On the downside, the Decision Tree overestimates Car over Bus usage, which has a significant effect on the Precision and Recall rates of the two classes respectively. Nonetheless, the Recall rate of 70% for the minority mode is promising, considering that the decision tree was not able to capture any instance of this class in the base case scenario. Table 3 depicts the performance of the tour-based XGBoost model on the validation set. By incorporating past factors sequentially into the dataset, the model was able to capture nonlinear dependencies and relationships between the various features, thus greatly increasing prediction measures. All the majority class (Car, Bus, Walk) metrics performed at over 90%, with an overall Balanced Accuracy=0.924. Regarding the minority mode rates, Recall=0.70 and Precision=0.895, it is apparent that the class imbalance was alleviated to a significant level. The accurate identification of the minority mode did not affect the overall performance of the model on the majority modes, in contrast to the Decision Tree application. Moreover, the XGBoost matrix Cohen's Kappa=0.872 corresponds to an almost perfect agreement, in contrast to Kappa=0.711 for the decision tree, indicating substantial interobserver agreement. Therefore, it is apparent that including tour-based effects and dynamic factors significantly increased the recognition rate for the minority (MrC) and improved the overall predictive performance of the Gradient Boosting Machine (GBM).

**TABLE 2 Performance of tour-based Decision Tree model on validation set**

| Confusion Matrix | Car | Bus | **Predicted** MrC | Walk | Total |
|---|---|---|---|---|---|
| Car | 548 | 25 | 8 | 1 | 582 |
| Bus | 133 | 253 | 11 | 1 | 398 |
| MrC | 6 | 1 | 17 | 0 | 24 |
| Walk | 2 | 1 | 2 | 120 | 125 |
| Total | 689 | 280 | 38 | 122 | 1129 |
| **Performance Metrics** | **Recall** | **Precision** | **Specificity** | **F-Measure** | **Balanced Accuracy** |
| Car | 0.942 | 0.795 | 0.742 | 0.862 | 0.842 |
| Bus | 0.636 | 0.904 | 0.963 | 0.746 | 0.799 |
| MrC | 0.708 | 0.447 | 0.981 | 0.548 | 0.845 |
| Walk | 0.960 | 0.984 | 0.998 | 0.972 | 0.979 |
| Average | 0.811 | 0.782 | 0.921 | 0.782 | 0.866 |

**TABLE 3 Performance of tour-based XGBoost model on validation set**

| Confusion Matrix | Car | Bus | **Predicted** MrC | Walk | Total |
|---|---|---|---|---|---|
| Car | 543 | 36 | 1 | 2 | 582 |
| Bus | 37 | 360 | 0 | 1 | 398 |
| MrC | 5 | 2 | 17 | 0 | 24 |
| Walk | 0 | 1 | 1 | 123 | 125 |
| Total | 585 | 399 | 19 | 126 | 1129 |
| **Performance Metrics** | **Recall** | **Precision** | **Specificity** | **F-Measure** | **Balanced Accuracy** |
| Car | 0.933 | 0.928 | 0.923 | 0.931 | 0.928 |
| Bus | 0.905 | 0.9902 | 0.947 | 0.903 | 0.926 |
| MrC | 0.708 | 0.895 | 0.998 | 0.791 | 0.853 |
| Walk | 0.984 | 0.976 | 0.997 | 0.980 | 0.991 |
| Average | 0.882 | 0.925 | 0.966 | 0.901 | 0.924 |

**Model Interpretation**

The balance between interpretability and predictive accuracy in machine learning is subject to continuous research. Guidotti et al. *(33)* produced an extensive survey of methods for explaining black box models and classifying the different state-of-the-art approaches. For the scope of this study, we selected partial dependence plots and SHAP values to explain the output predictions of the XGBoost model.

*Partial Dependence Plots*

Partial dependence plots (PDPs) are used to illustrate the functional relationship between a small number of input variables and predictions. In one-way PDPs, the y-axis depicts the marginal effect of one feature on the outcome of the machine learning model. By visualizing mode choice dependency on the input variables of interest, we can extract useful information on the motivating factors that influence the choice behavior.

Figure 5 depicts the two-way partial dependence of car usage on joint values of income and trip distance. As expected, higher values of trip distance increase the probability of car usage, with the greater increase observed in the region of 0-10km. This can be interpreted as a threshold value that individuals with car availability consider using alternative modes because of the shortest trip distance. In addition, the increase of income positively correlates with car usage as the main mode, reaching a point of diminishing returns for values >30,000 euros. This was expected considering the annual costs of maintaining a car are generally higher than the explored alternative modes (e.g. Bus, MrC). Furthermore, the partial dependence of Bus mode choice on the age distribution indicates that a lower age is linked to higher probabilities of Bus usage- for the age group of 16-24 years in particular- as it is characterized mostly by students that, in majority, do not own a driver's license or have access to a Car. Although this insight is important in explaining the output of our GBM, it needs to be clarified that the causal interpretation provided by partial dependence plots are relevant only with regards to the validity of our developed model, and not necessarily to the actual real-world decision making *(33)*.

Partial dependence plots are a useful tool in model interpretation, but they entail an important limitation, in the form of the independence assumption. It is a rare occurrence that the features of interest are not correlated with any other feature of the model. For instance, computing the Bus PDP (Figure 5) for a specific age range (e.g. 16-24 years), we need to average over the marginal distribution of income, which also includes higher values (>50,000 euros). This observation can be considered unrealistic for such a young age. Furthermore, PDPs may not account for hidden heterogeneous interactions, as they are based on average marginal effects across all individuals *(33)*. Therefore, while we can gain some useful insight on the model output, it is important to explore more ways of interpretation for the model under development.

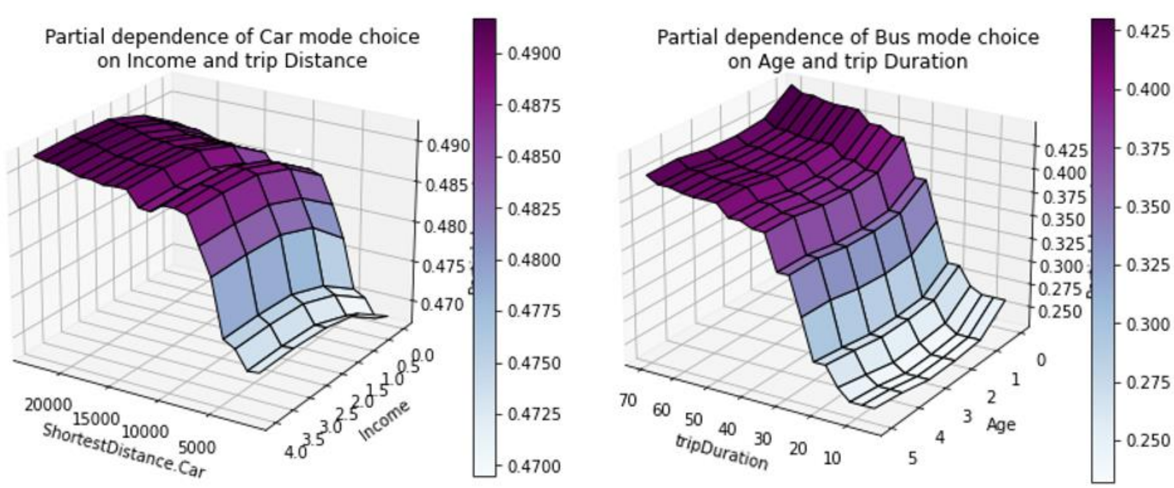


**Figure 5 Two-way partial dependence plots for Car and Bus modes**

*SHAP (SHapley Additive exPlanations)*

SHAP is a state-of-the-art machine learning interpretation approach, based on the work of L. Shapley in cooperative game theory *(34)*. The Shapley values attribute the total payoff from a cooperative game to the corresponding players. In 2017, Lundberg and Lee developed a package in Python that enables the estimation of SHAP for various techniques including XGBoost *(35)*. SHAP values are increasingly utilized by researchers within transportation for model interpretation. Mihaita et al. *(36)* employed SHAP to analyse the impact of different features on accident duration for traffic safety. Parsa et al. also used SHAP values to explain a GBM for the detection of highway traffic accidents *(23)*. The background of the Shapley values framework is presented below *(34)*.

Formally, a cooperative game is played by a set of players $N = \{1, \ldots, N\}$ termed the grand coalition. The game is characterized by a set function $u: 2^M \rightarrow R$ such that $u(S)$ is the payoff for any coalition of players $S \subseteq N$. Shapley values are built by examining the marginal contribution of a player to the existing coalition S. The Shapley value method satisfies a set of desirable axioms,

*i) Additivity Axiom*: For any two $u_1$ and $u_2$, $\psi_i(N, u_1 + u_2) = \psi_i(N, u_1) + \psi_i(\mathrm{N}, u_2)$ for each $i$, where the game $(N, u_1 + u_2)$ is defined by $(u_1 + u_2)(S) = u_1(S) + u_2(S)$ for every coalition $S$.

*ii) Symmetry Axiom:* For any $u$, if $i$ and $j$ are interchangeable then $\psi_i(N, u) = \psi_j(N, u)$.

*iii) Dummy Axiom* For any $u$, if $i$ is a dummy player then $\psi_i(N, u) = 0$.

*Theorem:* Given a coalition game $(N, u)$, there is a unique payoff division $x(u) = \varphi(N, u)$ that divides the full payoff of the grand coalition and that satisfies the Additivity, Symmetry and Dummy axioms, the Shapley value,

$$\varphi_j(N, u) = \frac{1}{N!} \sum_{S \subseteq \{x_1, \ldots, x_m\} \backslash x_j} |S|!(|N| - |S| - 1)![u(S \cup \{x_j\}) - u(S)]$$

Applying this framework for machine learning model interpretation, the Shapley value of a feature is its contribution to the payout, weighted and summed over all possible feature value combinations. As a result, in order to calculate the exact Shapley value, all possible sets of feature values have to be evaluated with and without the j-th feature *(33)*. Calculating the exact SHAP values for more than a few features is NP-hard. A computational effective approximation can be achieved with Monte-Carlo sampling, averaging the quantity within the expectation of a random sample. The above methodology was applied for the XGBoost model using the SHAP Python package *(37)*.

Figure 6 illustrates the mean SHAP values of the features in order of significance, for the Car and Bus classes, respectively. It is apparent that the variable *mode1* -encoding the mode of the previous trip- is a significant factor in the predictions, justifying the improvement in the performance of the model with the inclusion of dynamic factors. With regards to the SHAP values calculated for Bus, the historical distance and travel time variables have a strong influence on its selection probability. Owning a driver's license has a negative effect, while higher values of egress walking distance appear to demotivate travelers in selecting Bus for their trips. Finally, age also seems to influence the Bus class prediction, with younger groups opting for it in a more usual basis than the aging part of the population. Therefore, feature importance indicates that accounting for tour effects and ordering dependencies is critical in the performance of the GBM as they play a key role in everyday activity planning. People tend to plan ahead and optimize their joint within-day schedule rather than individual trips.

For the explanation of individualized predictions, we proceed to randomly select a female from the validation set, travelling from work to home, and depict the feature contributions for each classification class in relation to the base value (Figure 7).

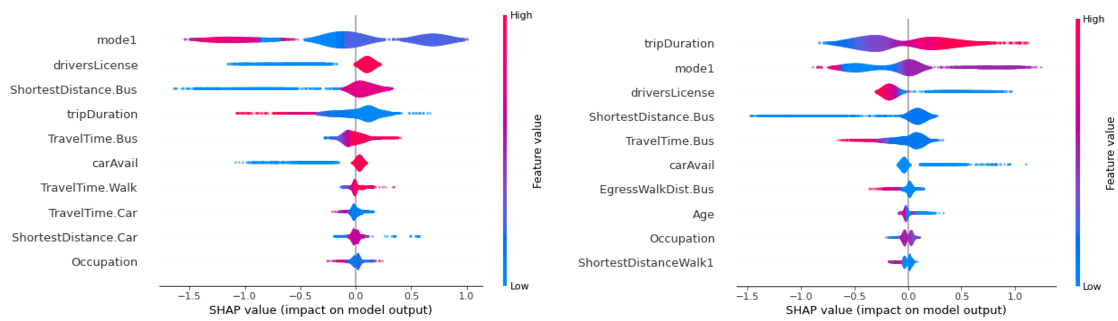


**Figure 6 Mean SHAP values of significant features for Car (left) and Bus (right)**

The GBM correctly classifies this individual as a commuting car driver with high confidence. The main positive factors include the past mode from the previous trip, in addition to the historical values on distance and trip duration. The choice of MrC was the second most probable for the given individual, with past mode choice and female gender contributing negatively to its probability. A possible explanation for this distinction might be the higher number of men that opt towards MrC ownership compared to women. Travelling on foot was disregarded by the classifier, mostly due to the higher values of travel time and distance, granting the specific trip impractical for travelling on foot.

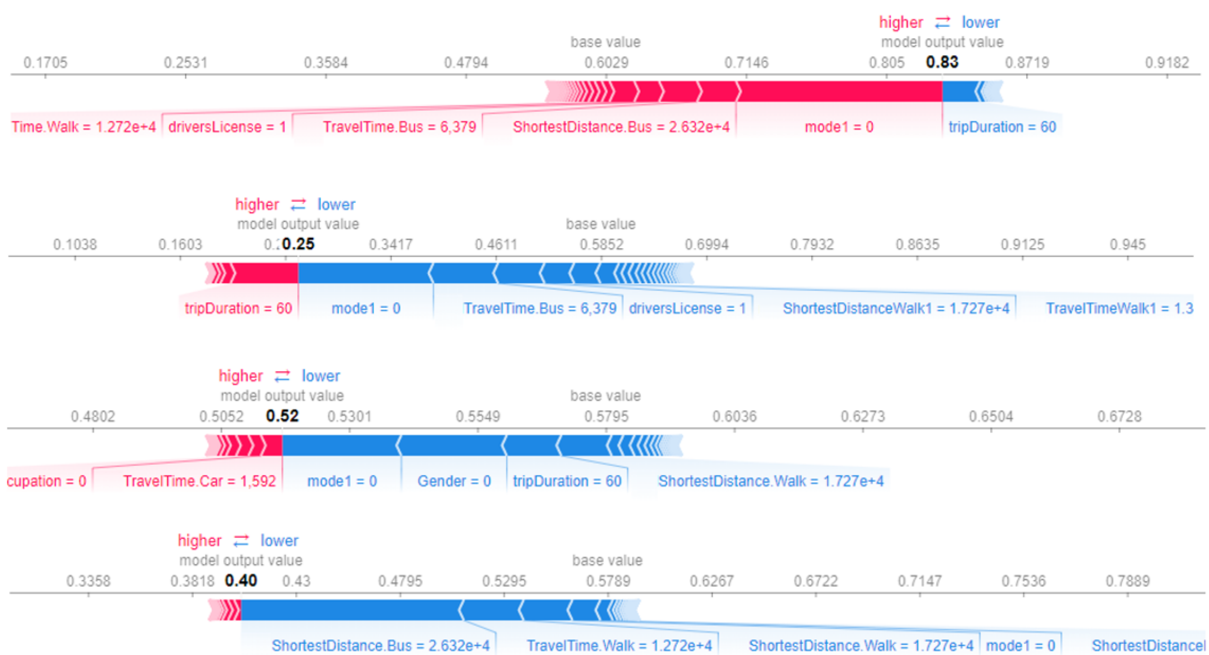


**Figure 7 SHAP feature contribution on individual predictions for Car, Bus, MrC, Walk**

Finally, the individualized predictions are explored further by generating synthetic data for a feature of interest, to identify the functional relationships and threshold values that would lead to a shift of contribution and potentially change of behavior for this specific person (Figure 8). It is apparent that the current trip value duration of 60 min is over this person's marginal value of positive contribution for the Bus mode, while the threshold of positive/negative contribution is predicted at 25min duration trips.

Advanced interpretation methods (e.g. SHAP) for complex 'black box' models can be of great value for transportation planning and policy applications. Judging from personalized and -to an extent- explainable model predictions, we can gain insight on urban infrastructure design, requirements (e.g. parking spots, EV charging stations) and travelers' needs.

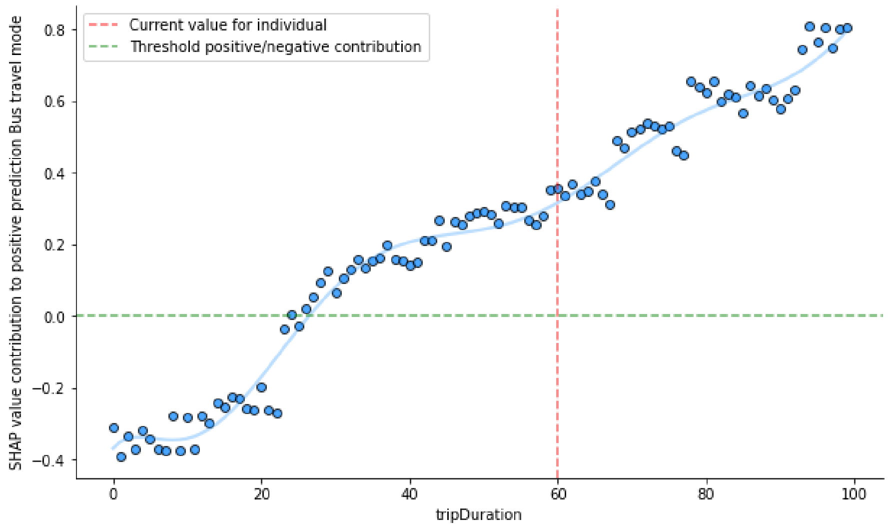


**Figure 8 SHAP value contribution of trip duration on Bus mode for individual**

## CONCLUSION

Creating effective classification models from imbalanced datasets is a challenge within many scientific domains. Typical machine learning algorithms tend to favor the 'dominating' classes and are thus inefficient in providing predictions for the minority class, which is often of great interest. This ongoing field of research is relevant to mode choice modelling, to evaluate the introduction of emerging mobility services which are currently underrepresented in everyday commuting. In this paper, we propose a tour-based modelling framework using extreme gradient boosting and apply it on imbalanced travel diary data from the city of the Thessaloniki. The results indicate that the XGBoost algorithm performed significantly better with the inclusion of tour-related effects and dynamicity factors within the training dataset, especially with regards to the identification of the minority mode. The output predictions were interpreted with partial dependence plots and the game theoretic SHAP approach, providing useful insight on feature importance and variable relationships. Future work includes working towards the implementation of these and other promising methods of machine learning modelling and interpretation for large-scale transport applications.

## ACKNOWLEDGEMENTS

The research reported in this paper has been supported by European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 815269, project HARMONY. The authors would like to thank CERTH-HIT and particularly Georgia Aifantopoulou for the kind provision of the data.

## STATEMENT OF CONTRIBUTIONS

The authors confirm contribution to the paper as follows: Study conception and design; all authors; Introduction: all authors; Background: Dimitrios Pappelis; Data Analysis: all authors; Model Development: Dimitrios Pappelis, Emmanouil Chaniotakis, Results and Model Interpretation: all authors; Conclusion: Dimitrios Pappelis; All authors reviewed the results and approved the final version of the manuscript.

## REFERENCES


1. Karlaftis MG, Vlahogianni EI. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies. 2011 Jun 1;19(3):387-99.

2. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018 Aug 22;51(5):1-42.

3. Branco P, Torgo L, Ribeiro R. A survey of predictive modelling under imbalanced distributions. arXiv preprint arXiv:1505.01658. 2015 May 7.

4. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks. 2008 Mar 1;21(2-3):427-36.

5. Wei W, Li J, Cao L, Ou Y, Chen J. Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web. 2013 Jul 1;16(4):449-75.

6. Li Y, Sun G, Zhu Y. Data imbalance problem in text classification. In2010 Third International Symposium on Information Processing 2010 Oct 15 (pp. 301-305). IEEE.

7. Brownlee J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery; 2020 Jan 14.

8. Wilson FR, Damodaran S, Innes JD. Disaggregate mode choice models for intercity passenger travel in Canada. Canadian Journal of Civil Engineering. 1990 Apr 1;17(2):184-91.

9. Nitsche P, Widhalm P, Breuss S, Maurer P. A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. Procedia-Social and Behavioral Sciences. 2012 Jan 1;48:1033-46.

10. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter. 2004 Jun 1;6(1):20-9.

11. Drummond C, Holte RC. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. InWorkshop on learning from imbalanced datasets II 2003 Aug 21 (Vol. 11, pp. 1-8). Washington DC: Citeseer.

12. He H, Ma Y, editors. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons; 2013 Jun 7.

13. Xu C, Wang Y, Bao X, Li F. Vehicle classification using an imbalanced dataset based on a single magnetic sensor. Sensors. 2018 Jun;18(6):1690.

14. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 2016 Nov;5(4):221-32.

15. Thai-Nghe N, Gantner Z, Schmidt-Thieme L. Cost-sensitive learning methods for imbalanced data. InThe 2010 International joint conference on neural networks (IJCNN) 2010 Jul 18 (pp. 1-8). IEEE.

16. Elkan C. The foundations of cost-sensitive learning. InInternational joint conference on artificial intelligence 2001 Aug 4 (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.

17. Tang L, Xiong C, Zhang L. Decision tree method for modeling travel mode switching in a dynamic behavioral process. Transportation Planning and Technology. 2015 Nov 17;38(8):833-50.

18. Schapire RE. The boosting approach to machine learning: An overview. InNonlinear estimation and classification 2003 (pp. 149-171). Springer, New York, NY.

19. Chen XM, Zahiri M, Zhang S. Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. Transportation Research Part C: Emerging Technologies. 2017 Mar 1;76:51-70.

20. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001 Oct 1:1189-232.

21. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

22. Wang F, Ross CL. Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. Transportation Research Record. 2018 Dec;2672(47):35-45.

23. Parsa AB, Movahedi A, Taghipour H, Derrible S, Mohammadian AK. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention. 2020 Mar 1;136:105405.

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

25. Brathwaite T, Vij A, Walker JL. Machine learning meets microeconomics: The case of decision trees and discrete choice. arXiv preprint arXiv:1711.04826. 2017 Nov 13.

26. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Berlin: Springer; 2018 Oct 22.

27. Prelipcean AC, Yamamoto T. Workshop Synthesis: New developments in travel diary collection systems based on smartphones and GPS receivers. Transportation Research Procedia. 2018 Jan 1;32:119-

28. Chang X, Wu J, Liu H, Yan X, Sun H, Qu Y. Travel mode choice: a data fusion model using machine learning methods and evidence from travel diary survey data. Transportmetrica A: Transport Science. 2019 Nov 29;15(2):1587-612.

29. Rashidi TH, Hasegawa H. An Innovative Simultaneous System of Disaggregate Models for Trip Generation, Mode, and Destination Choice. 2014.

30. Fernando A, Barrenechea E, Business H, Herrera F, Galar M. A Review on ensembles for the class Imbalance Problem. IEEE Transactions on Systems Man and Cybernetics: Part C: Applications and Reviews. 2012;42.

31. Urbanowicz RJ, Moore JH. ExSTraCS 2.0: description and evaluation of a scalable learning classifier system. Evolutionary intelligence. 2015 Sep;8(2):89-116.M. L. McHugh. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3):276{282, 2012.

32. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica. 2012 Oct 15;22(3):276-82.

33. Molnar C. Interpretable Machine Learning. Lulu. com; 2020 Feb 28.

34. Shapley LS. A value for n-person games. Contributions to the Theory of Games. 1953;2(28):307-17

35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. InAdvances in neural information processing systems 2017 (pp. 4765-4774).

36. Mihaita AS, Liu Z, Cai C, Rizoiu MA. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. arXiv preprint arXiv:1905.12254. 2019 May 29.

37. Lundberg SM, Erion GG, Lee SI. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888. 2018 Feb