Using Machine Learning methodologies on Mode Choice problems

T. Kesselring

2025.MME.9085

Delft University of Technology

Using Machine Learning methodologies on Mode Choice problems

by

T. Kesselring

Master Thesis

in partial fulfilment of the requirements for the degree of

Master of Science
in Mechanical Engineering

at the Department Maritime and Transport Technology
within the Faculty of Mechanical Engineering
of Delft University of Technology

To be defended publicly on Friday August 29, 2025 at 14:00.

Student number:	4955838
MSc track:	Multi-Machine Engineering
Report number:	2025.MME.9085

Thesis committee:

Dr. B. Atasoy	Chair of the committee & Daily Supervisor Delft University of Technology Faculty of Mechanical Engineering
Prof. Dr. Ir. E.B.H.J. van Hassel	Supervisor University of Antwerp
Ir. C. Yang	External Examiner Delft University of Technology Faculty of Mechanical Engineering

Date:	22-08-2025

TU Delft

Contents

List of Abbreviations

ALE	Accumulated Local Effects
ANN	Artificial Neural Network
BA	Balanced Accuracy
CM	Choice Modeling
CNL	Cross Nested Logit model
CNN	Convolutional Neural Network
DCM	Discrete Choice Model
DNN	Deep Neural Network
DT	Decision Tree
FCN	Fully Connected Neural network
FN	False Negatie
FP	False Positive
GB	Gradient Boosting
IIA	Independence of Irrelevant Alternatives
IWT	Inland Waterways Transport
LCCM	Latent Class Choice Model
LightGBDT	Light Gradient Boosting Decision Tree
LR	Logistic Regression
LRT	Light Rail Transit
MaaS	Mobility as a Service
MAE	Mean Absolute Error
MC	Mode Choice
ML	Machine Learning
MLP	Multilayer Perceptron
MNL	Multinomial Logit model
MNP	Multinomial Probit model
MXL	Mixed Logit model
NL	Nested Logit model
OD	Origin-Destination
RF	Random Forest
RFDT	Random Forest Decision Tree
RMSE	Root Mean Squared Error
RUM	Random Utility Maximization
SHAP	SHapely Additive exPlanation
SMOTE	Synthetic Minority Over-sampling TEchnique
SVC	Support Vector Classifier

TN True Negatie
TP True Positive

WLM Weighted Logit Model

XGB eXtreme Gradient Boosting

Glossary

Latent Class Model Latent class models are statistical models used to identify un-observable sub-groups, or latent classes, within a population based on observed variables. These models assume that the observed data are generated by a finite number of latent classes, each characterized by a unique pattern of responses on the observed variables. Through estimation techniques, latent class models uncover these underlying groups, allowing for insights into population heterogeneity and facilitating classification of individuals into meaningful categories.

Loss function A loss function measures how well the predictions of the model fit with the actual target values.

One-vs-One A One-vs-One model trains a separate classifier for each class pairing.

One-vs-Rest A One-vs-Rest model trains a classifier on each separate class independent of the other classes.

Random Residual The random Residual is the part of the utility equation that contains all unobserved attributes and other factors that influence the utility value of an alternative for an agent. As the attributes and other factors are not known this value is treated as a random value picked from a random distribution, the distribution type depends upon the model used.

Revealed Preference A Revealed Preference is an observed preference, if for example an assumed rational agent has the option between 2 products (product a and b) and it picks product a, then product a is a revealed preference. The key here is that revealed preference is based on actual observations

Shapely values A shepely value is a concept from cooperative game theory. This concept

SHapely Additive exPlanation An algorithm that helps visualize the effect of certain trained parameters on the overall outcome of the model for Machine Learning methodologies.

Stated Preference When a rational agents is asked whether they prefer product a over product b or the other way around, the result is then a Stated Preference. For scientific researches this is usually executed in the form of a survey

Supervised learning A Machine Learning approach where the data is labeled and each input set is matched with a distinct correct output label. The goal is that a trained model can predict output labels on unseen input data

Trip Chain A Trip Chain is a sequence of events that happen during a trip, for example a trip chain could be as simple as driving from home to the supermarket and back. But multiple activities can be done within one trip chain as well, such as driving from home to the supermarket to the library and then back to the home again. The entire trip is called a Trip Chain.

Unsupervised learning A Machine Learning approach where data is not labeled en there are no defined outputs. The goal is to find patterns and relationships within the data.

Weak learner A weak learner is a Machine Learning model that performs only slightly better compared to making random guesses.

Abstract

This thesis explores the use of Machine Learning (ML) techniques to model and predict transportation mode choice behavior, a critical component of urban mobility planning. Traditional mode choice modeling relies on Random Utility Maximization (RUM) theory, with models such as Multinomial Logit model and Mixed Logit model (MXL). While these offer interpretability, they often struggle with complex feature relationships and heterogeneity in large datasets. ML methods, by contrast, offer greater predictive power and flexibility, albeit with interpretability challenges. This study evaluates the performance of various ML models like Gradient Boosting and Random Forest Decision Tree on two datasets, with a primary focus on a detailed case study using Swiss travel data. The models are assessed under various configurations, including feature limitation, latent variable extraction, and SMOTE-based resampling. Comparative results demonstrate that ML models consistently outperform traditional Logistic Regression models in terms of F_1 and Balanced Accuracy metrics. Additionally, tools like SHapely Additive exPlanation (SHAP) are employed to enhance the interpretability of ML outcomes. The findings highlight the potential of ML to improve Mode Choice modeling, particularly when combined with theory-informed structures and advanced data balancing techniques.

1

Introduction

This chapter serves as a general introduction to the context of this thesis in the area of mode choice problems and the investigation of the potential of the machine learning methods for such problems. In Section 1.1 the relevance and general methodology of conducting Mode Choice (MC) research is explained and the nature of MC problems are discussed. In Section 1.2 the opportunities of applying Machine Learning (ML) to Mode Choice problems are discussed and the basis to the goal of this research is made. This basis is further developed into research questions in Section 1.3. In Section 1.4 the methodology used for each research question. Finally the structure of the research is explained in Section 1.5.

Mode choice problems refer to the decision-making process individuals or groups face when selecting a mode of transportation for a particular trip. These choices can include various options such as driving a car, taking public transit, cycling, walking, or using shared mobility services like ride-hailing. The decision-making process is influenced by multiple factors, including convenience, cost, travel time, safety, environmental impact, and personal preferences.
Understanding mode choice is crucial for transportation planning, urban development, and policy-making. By analyzing how people choose their modes of transport, planners and researchers can better predict travel patterns, optimize transportation networks, and identify areas where infrastructure improvements are needed. The insights gained from studying mode choice problems are essential for promoting sustainable transportation, reducing traffic congestion, and enhancing accessibility in cities.

Figure 1.1: *Two examples of wider bicycle paths in Manhattan (DOT, 2025b)*

Overall, mode choice problems play an important role in shaping transportation systems that are not only efficient but also equitable and sustainable, which are essential qualities for modern urban environments. One example of where a MC study has delivered results is in New York City (*NYC DOT - NYC Streets Plan*, 2021), like many large cities, New York faces large quantities of car traffic and as a result a lot of congestion within the city. To combat this the city wanted to rebalance how public streets prioritize different modes of transport in order to reduce the number of private motor vehicles, which take up a majority of street space despite not being the majority mode choice already. In combination with other studies related to the traffic and use of public space within NYC, the city was able to invest in relevant projects that would improve the attractiveness of other mode alternatives, such as cycling, walking and public transit, in order to reduce the number of people in private motor vehicles.

Follow up studies show that areas that have had infrastructural improvements, such as the examples shown in Figure 1.1, enjoy higher usage of alternative modes and lower pedestrian fatalities due to increased safety (DOT, 2025a; Rong & Freeman, 2024)

1.1. Mode Choice Problems

As mentioned MC is a key aspect of transportation planning and decision-making, focusing on how individuals or groups select a particular mode of transportation (car, bus, bike, walking, train, etc.) for their trips. These problems are central to understanding and predicting travel behavior, especially in the context of urban planning, public transportation design, and transportation policy. Moreover, these problems are becoming increasingly important for freight logistics as well in order to influence the chosen modes for addressing sustainability challenges. However predicting the behavior of entire groups can be a difficult task.

1.1.1. How to predict mode choice behavior

To make accurate predictions of a given choice it is important to acquire information about the system. For example, when predicting how much ice cream will be sold you would look at the predicted weather (will it be warm or cold?), what day of the week or year it is (is there a holiday, are people probably at work?), how expensive the ice cream is, etc. For mode choice systems this works the same, an analyst would observe important attributes of the system as a whole and also of each available alternative and based on historical data could predict an individuals mode choice behavior.
However it is fair to say that no two individuals are alike, where one individual would happily travel from A to B using a bicycle, another could choose to use a car provided the same circumstances in both scenario's. To compensate, the characteristics of individuals are typically also accounted for, e.g. someone's socio-economic background contains a lot of information that could explain the difference in what mode of transport they would use. Attributes such as age, occupation, living situation, etc. can thus all be used to help make better predictions. Unfortunately a new problem arises now in that the model quickly can become extremely complex with an increasing amount of variables to consider.
To do this, an analyst typically uses a so called Random Utility Maximization (RUM) based model. This theory was developed by McFadden (2001) and works on the basis that each individual has a 'utility' that they always want to maximize. Each alternative that can be chosen will provide an individual with a specific utility value and the alternative that has the highest utility will end up as the chosen alternative. To account for unobserved attributes and heterogeneity within the population there is a random distribution included in the model. This allows for the model to be relatively simple and still somewhat accurate.

1.1.2. Relevance of mode choice studies

As mentioned before, the goal of doing a mode choice study can vary but often focuses on transportation planning, decision making and policy shaping. A typical mode choice study thus also tends to focus on how impactful each observed attribute is on the system rather than looking at how well the model actually predicts. In a RUM based model each alternative will have a weight connected to it. This weight is determined during the training of the model and thus based on known historical data. By looking at the value and statistical outputs of each weight the analyst will be able to tell how impactful the corresponding attribute is on the system as a whole. This can in turn be used to possibly alter the system to show a more preferred behavior. For example, in the interest of combating climate change, it is favorable to entice more people to use public transport or walk/bike compared to using a private motor vehicle. Similarly, for freight transportation, it is

desirable to shift to waterways in order to reduce externalities in our urban and suburban areas. A mode choice study can be conducted on a specific area or between specific locations and the results will show which specific attributes are more important to the population as a whole in their decision to take a specific mode of transport and thus where the system can be improved to achieve the more favorable behavior in the future. Similarly it can thus also be used for private companies such as public transportation providers or cargo shippers to identify how they can change to make their alternatives more attractive and which technologies and policies they can adopt for improving the services they provide.
While it is rare for the models to be used to predict future travel behavior, it is still important for the model to be accurate in doing so. If the model is good at predicting it can in turn be concluded that the weights and variables that eventually will be used for evaluation are better representations of the system.

1.2. Machine Learning opportunities

While there are already good methods that can be used for MC problems, they also have their limitations. Complex relations between different attributes and heterogeneity within the population can be difficult to accurately capture and model, even with more advanced RUM based models. On top of that do RUM based models also require various pre-made assumptions regarding the used random distribution which can influence the accuracy of the results. With a ML model making such assumptions is not always needed, due to their 'data driven' nature. All context will be extracted out of the provided data rather than out of existing knowledge as would be the case with RUM based models which are 'theory driven'. This could possibly allow for an easier set up of a mode choice analysis. However it also negatively influences the ability to extract the required information out of the trained model, which can make it more difficult to draw accurate interpretations and conclusions.
Furthermore, there is also a global change in how data is collected and how widely available it is. With automated data gathering by sensors and computers, it is possible to get large amounts of data to train a model, but due to constraints such as privacy concerns this data is often not as in depth as data gathered using specifically designed surveys. This lack of available attributes will make it more difficult to get an accurate RUM based model. ML models however tend to do well with large datasets and could potentially still capture the more complex relations of the system despite having relatively less attributes as input.

1.3. Research questions

The main goal of this research is to investigate how ML can be used to better understand the MC process of a system, which is done by answering the following main research question:

What is the potential of machine learning models to accurately evaluate Mode Choice behavior for transportation systems?

To help answer this question, the following sub-research questions are formulated:

1. What Machine Learning methods can be used for MC problems?
2. What is the potential of using synthetic data generation and over/under sampling for imbalanced datasets?
3. How can integrated theory based knowledge improve ML models to make the most of both paradigms?

4. How can interpretable insights, such as the importance of features influencing mode choice, be extracted from trained ML models?

1.4. Research methodology

Each of the mentioned research questions in Section 1.3 will be answered using the following methodology respectively:

1. A literature review will asses various different ML techniques.
2. A combination of literature review and a direct comparison of several created ML models, with and without synthetic data generation techniques, will be used to asses the viability.
3. A direct comparison between several created ML models will be used to find out if integrating theory based knowledge improves the model results. Additionally a Logistic Regression model is included as it is well suited to represent a traditional logit based Random Utility Maximization model and can thus be used as a benchmark. The comparison will be done by evaluating the model accuracy with F_1 and Balanced Accuracy scores.
4. The created ML models will be evaluated using different techniques discussed in the literature review. The evaluations will be compared with already existing studies.

1.5. Report structure

To answer the research questions from Section 1.3 the report is structured as follows: In Chapter 2 various existing studies are being reviewed regarding combining ML and MC to identify problem areas and possible tools that can be used as solutions. After that in Chapter 3 this knowledge is applied to create the methodology for applying ML to MC. The created models are used on two case studies in Chapter 4:

- **Rhine-Alpine corridor Dataset**: A dataset containing shipping data (cargo transportation) between different location along the Rhine-Alpine corridor.
- **Switzerland Dataset**: A dataset containing passenger commuter data throughout Switzerland. Both of these datasets were chosen because of the unique challenges each of them provide. This gives the ML models the opportunity to be tested with both aggregate and disaggregate datasets as well as with passenger and cargo transportation systems. This chapter serves to answer the main research question by evaluating which model type has the best predictive accuracy and in turn can be best used to understand what the most important or influential variables are for choosing mode choice alternatives.

Finally in Chapter 5 a conclusion is drawn and the posed research questions are answered. Additionally some future research opportunities are discussed as well.

2

Literature

This chapter provides an overview of relevant research on the various Random Utility Maximization based models that are used on MC studies. First in Section 2.1 a more high level overview of MC literature is provided. After that in Section 2.2 the most prominent Random Utility Maximization based models are discussed and various existing literature papers are reviewed to identify the strengths and weaknesses of these models and how they are applied. Additionally in Section 2.3 some research that applies latent class models to MC problems is reviewed. In Section 2.4 the focus is shifted to research related to using Machine Learning on MC problems. Finally in Section 2.5 the most important differences between Random Utility Maximization based models and Machine Learning based models are highlighted and a research gap is identified.

2.1. Mode Choice literature

There are various reasons to perform a MC study and what modes are looked at during. Before diving deeper into the used methodology it is important to review how mode choice studies are performed on a higher level. Wu et al. (2019) performed an extensive review on a large collection of MC related papers. With this review they were able to identify the important research topics and most used keywords. Specifically the keywords can paint a good picture of what most studies have as research goals. In Figure 2.1 the results of this review study are displayed, from this it is clear that there is a large focus on methodology with Logit and Discrete choice models dominating the field. Additionally this review also gives a good insight in the purpose of the research and what modes are considered. With environmental protection and urbanization making up over half the keywords relating to purpose it thus also makes sense that the modes looked at the most are related to mass transport and sustainable alternatives to private vehicles.

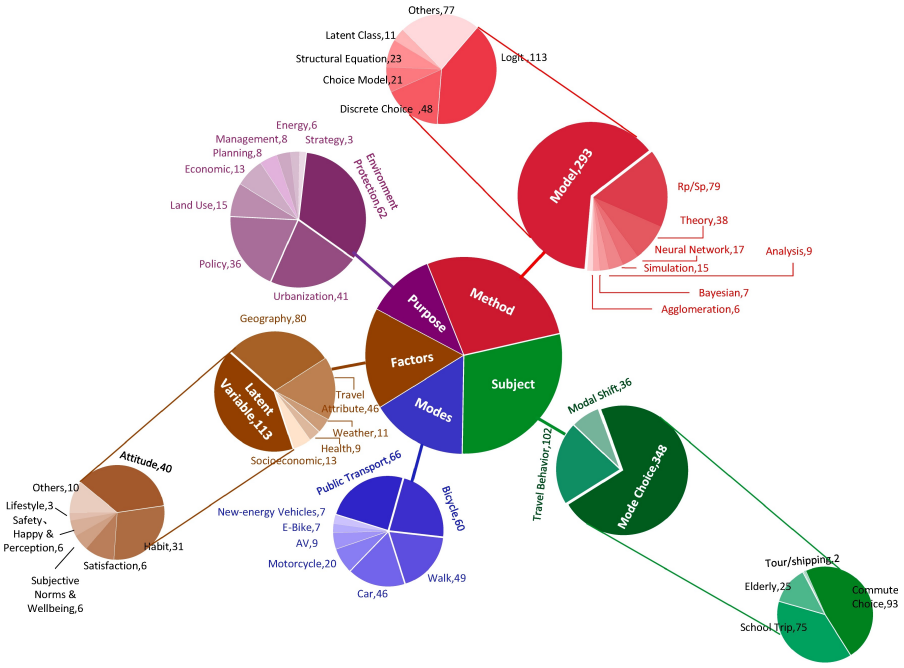


Figure 2.1: *The classification of keywords used in various MC related research papers (Wu et al., 2019).*

2.2. RUM based models

Based on the research review from Wu et al. (2019) it is clear that RUM based models are very popular as a use for MC studies. The most used ones, Multinomial Logit model, Nested Logit model and Mixed Logit model are reviewed in this section in Subsection 2.2.1, Subsection 2.2.2 and Subsection 2.2.3 respectively.

2.2.1. Multinomial Logit Model

The Multinomial Logit model (MNL), is a widely used model when faced with a MC problem, it benefits from the fact that it can be used for problems with 3 or more possible alternatives which a MC problem generally has. The upside of the MNL is that it is relatively simple to set up and use, this does however come with the downside that many complex characteristics of a MC problem may not be captured.

Masoumi et al. (2020) used MNL to investigate what factors are important in the mode choice a child (between the ages of 9 and 12 years old) takes to school. The data used for the research contains survey entries from 9 different cities in 7 European countries, this resulted in the researchers being able to make more comparisons because of the different infrastructure and general economic status between the countries.
Ding and Ning (2016) also applied MNL to identify key factors in MC decision making, however this research combined Revealed Preference and Stated Preference to also compare the current situation to a hypothetical situation where different policies were implemented. In this research this concerned travel behaviour in the Central Business District (CBD) of Nanjing, China. Ding and Ning (2016) investigated how modal shift from cars to mass-transit would occur when applying policies such as introducing a managed (dedicated) bus lane, decreasing ticket prices for public transit and increasing parking fees.
Eluru et al. (2012) also used MNL to predict the impact of policy changes, just like Ding and Ning (2016) the main goal is investigating what type of policies would be effective at getting more people to use other modes of transport over the car. The results of this research show that mainly the number of transits and transport time when using public transits impacts the decision maker when choosing between car and public transport.
The research of K. Wang et al. (2022) uses the MNL as a comparison tool to a different model type, namely the Multinomial Probit model (MNP). Nonetheless this research does serve as a good example of applying MNL to a MC problem. K. Wang et al. (2022) uses the data from a web-based travel survey issued in the Shanghai area as input to their model resulting in a data set of 1743 individual trips. A total of six alternatives / mode choices were considered in this research and the various resulting coefficients are well discussed. However the answer to the main research question (comparing the MNP to various other models including the MNL) actually shows that the MNP outperforms the other models.
Moreover Sekhar et al. (2016) used MNL as a comparison tool to a Random Forest Decision Tree (RFDT). A total of eight different modes were considered in this research and to answer its research question the total of correct predictions on a validation set of the two model types were compared with each other. This resulted in MNL scoring 78.01% while RFDT achieved an accuracy of 81.65%.
The research of Ton et al. (2020) uses multiple model types as well, however here MNL is not used as a model to compare others to but rather just one of the possible alternatives that can best fit the data. In contrast to the researches of K. Wang et al. (2022) and Sekhar et al. (2016) the other model types used in this research are the more common ones, namely the Nested Logit model (NL), Cross Nested Logit model (CNL) and the Mixed Logit model (MXL). Surprisingly and against the expectations of the researchers the MNL resulted in the best data fit even though this model type is the most simplified out of all 4 options. Ton et al. (2020) claim that this is a result of how

they had set the utility function: "because we include a relationship in the utility function between alternatives that contain the same modes via the estimation of mode-specific parameters "
MC does not only apply to passenger travel, also cargo operators have a wide variety of transport modes to choose from when transporting their goods from point A to B. The research of Pochan and Wichitphongsa (2020) investigates what some entrepreneurs would prefer in a hypothetical situation where infrastructure for high speed rail and normal double track rail existed to transport their goods between Bangkok and Chiangmai (Thailand). An interesting insight from this research however is that the mode of truck transport still covered more than half of the market share. The researchers stated the following about this: "... entrepreneur are still familiar with road transport and mainly used handling time more than transit time". This highlights an important factor in these types of hypothetical scenarios: 'resistance to change' (Dent & Goldberg, 1999). However, generally after some time this resistance fades away resulting in different real life outcomes compared to the predicted ones using MC models.

2.2.2. Nested Logit Model

The Nested Logit model (NL) is a further development of the RUM theory, which allows for different alternatives to be grouped together (nested). This allows for alternatives within a nest to be correlated together, as a result the different alternatives within each nest will get more closely related Random Residual values. An example of how this can be useful can already be found in MC problem; When a decision maker faces the alternatives of traveling by driving, taxi/Uber, bike, bus, train/metro and walking it can be nested together in 3 categories namely: Car (driving and taxi/Uber), Public transportation (bus and train/metro) and Self powered (walking and biking). The alternatives within these categories have overlapping characteristics that can result in the decision maker viewing them similar and thus having a similar Random Residual value for them in their utility function. For example the group of public transport has the overlapping characteristic of not going directly from the origin of the decision maker to the desired destination and having to share the vehicle with others. It can be reasonably assumed that a decision maker has a certain like or dislike (utility or dis-utility) for these aspects and therefore it is useful and more accurate if these alternatives are correlated with each other.

The research from Shahikhaneh et al. (2019) applies NL to find the key parameters as to why motorcyclists in Mashhad, Iran choose the motorcycle over the relatively safer alternatives of Light Rail Transit (LRT) and Bus. Here the two alternatives where grouped together in a nest and 'main alternative' of the motorcycle is alone in a nest. The research states that the reasoning for this nesting structure is to better assess the actual research question: "How to get motorcyclists to use public transportation alternatives?". It thus makes sense to group the two public transportation alternatives together versus the motorcycle.
Qi et al. (2020) applied the NL in a more unique way such that rather than grouping different modes together in nests, this research only considered two modes and different Trip Chains. The research proposes two different nested models, one where the two modes of travel (private car and public transport) are the nests that each contain the same five Trip Chain sequences and a model that is the opposite (five nests with the different Trip Chains each containing the two modes). The goodness of fit tests that were done on the two models showed that the first model represented the data at a much higher level indicating that people more often choose the mode of travel based on the planned activities during the trip.
While the NL already introduces the ability of alternatives belonging to multiple groups/nests and thus having overlapping Random Residual values, the research of Wen et al. (2012) adds to this by also introducing a Latent Class Model. First the researchers identified the best nesting structure

by comparing goodness-of-fit data. Three different structures were looked at with an additional MNL to compare to, the three structures were: 1) A public transport nest while the other options were un-nested. 2) A car nest while the other options were un-nested. 3) A public transport nest and a car nest. Between these different structures the third one outperformed the others based on the likelihood ratio test. Later in the research the Latent Class Model is added to the MNL and best performing NL structure, both Latent Class Model versions performed better then their standard counterparts with significantly better goodness-of-fit. The Latent Class Model NL with four segments is the overall best according to the researchers: "it had the best goodness-of-fit and the ability to accommodate a flexible structure for the similarity among alternatives and variations in taste parameters.".
Hess et al. (2013) uses a different type of NL: Cross Nested Logit model (CNL). This model type is already mentioned in Subsection 2.2.1 in the research of Ton et al. (2020). CNL allows alternatives to be part of multiple nests resulting in even more correlation between similar alternatives. This particular research focuses on slightly more then just MC but rather looks at a trip consisting of three different main choices: Airport, Airline and Access Mode. In the research each alternative of those three choices is a nest, all final options will than belong to three nests based on the alternatives that are used in the option. This nesting structure is visualized in Figure 2.2.

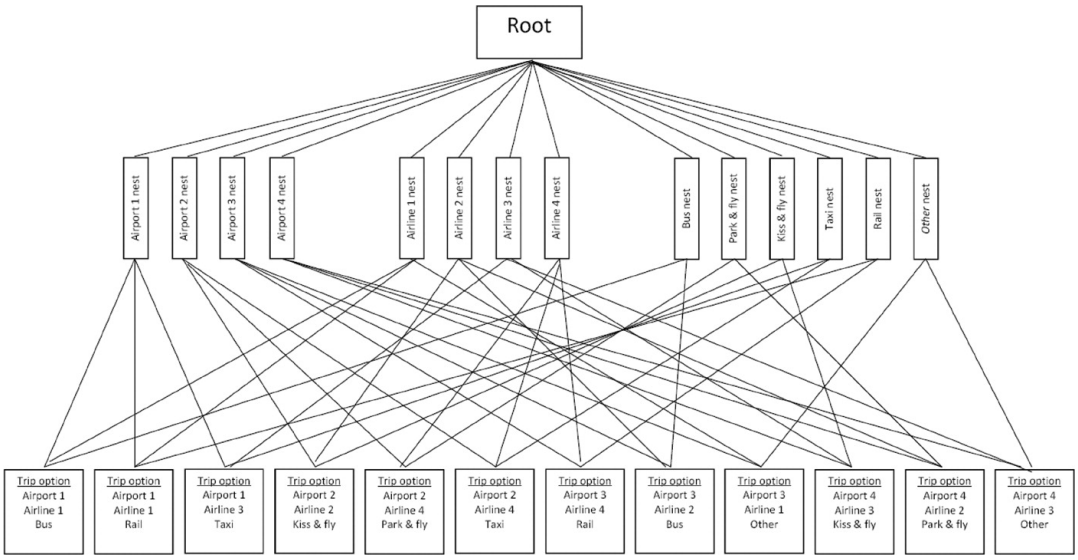


Figure 2.2: *The CNL structure used in the research of Hess et al. (2013) (Hess et al., 2013)*

Just as the research of Ton et al. (2020) multiple model types were also used to compare with and to check whether the CNL is significantly better then the simpler MNL and NL methods. Three different NL structures were used, each using nest of the alternatives from one of the main choices that make up a trip (NL1. Each airport is a nest NL2. Each airline is a nest NL3. Each access mode is a group). The NL1 structure shows a significantly better log-likelihood over the MNL structure while NL2 and NL3 are only slightly better and not statistically significant. This indicates that there is significant correlation between the different airport alternatives but not between the airlines and the access modes. However the CNL structure proves to have an even better log-likelihood than the MNL and NL1 structures. It also rejects all three NL structures on the basis of the χ^2 test. This concludes that unlike what the results of NL2 and NL3 show, there is actually significantly relevant correlation between these nests which only shows up within the CNL structure.

2.2.3. Mixed Logit Model

A Mixed Logit model (MXL), also known as a Logit Mixture Model, is an even further development of the RUM theory. MXL allows for more random distributions in order to capture correlations between different aspects such as the alternatives or observations. This can be a useful methodology as several situations imply some kind of correlation. With MNL there is no possibility to capture any correlation, this is a result of the Independence of Irrelevant Alternatives (IIA) assumption for the Random Residual. The NL partially solves this by allowing for correlation between alternatives that are placed within the same nest. MXL expands upon that by allowing even more flexible correlation structures through any type of random distribution.

Polydoropoulou et al. (2022) uses a MXL in order to capture the correlation between observations of the same individual, this is beneficial to this specific research as the observation data contains three separate observations for each individual which makes it important to be able to capture the correlation between these observations. The research itself focuses on sustainable last mile delivery where traditional package delivery is compared to more high tech delivery solutions with autonomous vehicles. The research is conducted in the form of a Stated Preference survey and eventually concludes that people preferred the traditional delivery methods.
Similarly the research of Ye et al. (2020) uses MXL because the "influencing factors" used in the study are not independent of each other. As a result the standard MNL can not be used due to the mentioned IIA assumption. Different to the other literature discussed, this research does not consider multiple mode choices as alternatives but rather how often the respondent would use bike-sharing as their mode of transport. This is combined with hypothetical travel scenarios that included: the distance to travel, weather, time of day and trip purpose; together with personal information the researchers determine what personal attributes as well as external attributes influence people's decisions to use bike-sharing. In order to achieve this goal the research also contains three different Mixed Logit models:

- The first model describes what impact individual attributes have on a person's travel characteristics (travel time and travel expense)
- The second model investigates how various factors such as age, travel distance, etc. influence an individual's bike-sharing usage
- The third model looks at an individual's willingness to shift to bike-sharing

Yang and Sung (2010) aim to analyze the effects of introducing a new mode to the market on passenger MC behaviour in Taiwan. Similar to Polydoropoulou et al. (2022) and Ye et al. (2020) this research also uses MXL with the intend to allow for heterogeneity among individuals. In order to analyze the effect that the introduction of a High-Speed Rail (HSR) line has on the transportation market in Taiwan the Stated Preference survey is split up into two sections. One section covers the situation before the introduction of HSR and one section with HSR. The MXL model is than applied to both data sets in order to visualize the differences before and after the introduction of HSR. Finally it was concluded that the introduction of HSR in the market did influence the relationship between the other alternatives to some degree.
Nicolet et al. (2022) use a weighted Mixed Logit model to estimate heterogeneous mode choice behaviour within a cargo route along the Rhine-Alpine corridor. The challenge with this case study is that the available data is aggregated and contains less context and details compared to most MC studies. To be able to extract useful results from the dataset they use a combination of a Weighted Logit Model (WLM) and a Mixture Logit model (or MXL). By assigning a weight to the model it is possible to correctly use the aggregated data, this is because this data set only shows the total flow for each Origin-Destination (OD) pair. The weighted factor within the model represents the total volume for each OD pair.

2.3. Latent Class models

A Latent Class Choice Model (LCCM) are considered an extension to Discrete Choice Model (DCM), similar to MXL a LCCM is able to capture unobserved heterogeneity in the population. LCCM does this by segmenting the population in different classes based on similarities between individuals. This has the benefit of allowing more interpretable values to represent why individuals within different classes make certain decisions. In contrast to having a singular random distribution describing the entire population, as would be the case with MXL (Hess et al., 2009).

Greene and Hensher (2003) looks closer into the differences between a MXL and LCCM. In the research they mention the trade-off between the two model types: The latent class structure is less vulnerable to false assumptions made by the analyst as there is no need to specify random distributions that describe the population, in turn this does mean that the overall model becomes less flexible. To compare the two models they use a Stated Preference choice experiment on what road type an individual would prefer to drive on by car. In this experiment a MNL was also included, with a log-likelihood test it was clear that this methodology did not perform nearly as well as MXL and LCCM. From this it can be concluded that unobserved heterogeneity is present within the data as this is what MNL is unable to capture. The two other methodologies performed nearly on the same level, Greene and Hensher (2003) concluded that neither model can be considered superior over the other and state that the usage of either methodology would be a result of the earlier mentioned trade-off.
Matyas and Kamargianni (2021) use LCCM to investigate heterogeneity is Mobility as a Service (MaaS) preferences. One of the most important parameters to figure out for a LCCM is the number of (latent) classes the model should have. To figure this out Matyas and Kamargianni (2021) ran six different models, each with one more class than the previous model, and ran various tests on the results to identify the best performing one. The tests chosen for this evaluation were the "rho-bar squared", Akaike Information Criterion (AIK) and Bayesian Information Criterion (BIC). Because the first ran model is a 1-class LCCM this is the same as running the model without any latent classes. This allowed the researchers to also identify if any heterogenous latent classes were present within the data. Although not all tests pointed to the same model being the best performing, combined together the 3-class LCCM was considered the best overall fit. The benefit of introducing latent classes is that researchers have the ability to directly see what type of groups are present within a population and what their specific implied importance is towards certain attributes and alternatives. In this case the 3-class model showed the following groups: MaaS avoiders, MaaS explorers and MaaS enthusiasts.
Lahoz et al. (2023) combine both LCCM and ML. This allows them to have the strong power ML methodologies provide for capturing unobserved behaviour and heterogeneity while maintaining a nicely interpretable model because of the LCCM. In the proposed framework a traditional LCCM is used with its two sub-models:

- A class membership model: This model is responsible for computing the probability an individual n belongs in a certain class k.
- A class-specific choice model: This model is responsible for assigning a probability an individual n chooses an alternative taking into account that this individual belongs in a certain class k.

An ANN is used to construct the latent variables for this model: "We propose a nonlinear relationship between the socio-characteristics of the individuals and the latent constructs by employing two densely connected layers, where one is a hidden layer and the other is the latent variables layer". Finally, the researchers conclude that the unique properties of both LCCM and a ML methodology

such as ANN indeed allow them to capture complicated nonlinear relations between various parameters. And while the created framework did not provide a substantial improvement in predictive performance, this captured nonlinear behaviour does allow for a better analysis in how socio-chatacteristics are related to some attitudinal indicators.

2.4. Machine Learning in Mode Choice problems

2.4.1. The benefits and challenges of Machine Learning in Mode Choice problems

Van Cranenburgh et al. (2022) identifies the potential that ML techniques can have on Choice Modeling (CM) by describing the similarities and differences between the two main modeling paradigms: Theory-driven and data-driven modeling. These paradigms mainly correspond to the 'classical' RUM based methodologies and ML respectively. The researchers conclude that both paradigms share the same underlying concepts and theory, albeit that both fields use different terminology for similar principles. The differences however are more interesting, where theory-driven methodologies can show a clear connection between the choices individuals make and the reasoning for them. These connections can directly be derived from the coefficients within for example the utility function. With data-driven methodologies this is however a lot harder or even impossible to do: "the underlying 'first principles' are unknown, or the systems under study are too complex to be mathematically described" (Ran & Hu, 2017). The focus and strong point of data-driven methodologies is the ability to make more accurate predictions based on out-of-sample data. Van Cranenburgh et al. (2022) also provide some recommendations on how ML can be directly applied to CM, for example by capturing random heterogeneity within the data. While the Mixed Logit model (MXL) is already designed to do just that with the usage of one or multiple random distributions for model parameters. However, these distributions still need to be defined by the analysts themselves which leaves room for error. On top of that there is also the possibility that a good fitting random distribution does not exist for the data. ML can be used here to help capture the random heterogeneity and in that way make the overall model more accurate.
Hillel et al. (2021) also review the usage of ML regarding CM and even specify more towards Mode Choice (MC). One of the main problems this research brings up is that there have not been many studies that sought out to "comprehensively compare ML techniques with each other and with Random Utility Maximization (RUM) models". As a result of this, most researches that apply ML to a MC problem do not have a good framework that supports which ML model can be used best for the problem. This is in contrast with how research that solely use RUM based methodologies where there is often clear reasoning why a specific model is used and sometimes results of multiple models are even compared within the research. One of the results of this problem is that many researches make some critical errors with their usage of datasets and validation in regards to ML methods.

2.4.2. Implementing Machine Learning to Mode Choice problems

As mentioned in Subsection 2.4.1 one of the challenges of using ML with MC problems is the lack of frameworks on to which researchers can choose the best ML methodology to use. Kashifi et al. (2022) sets out to make such a framework by using five different ML methods on a case study that uses three years worth of travel data in The Netherlands: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP) and Light Gradient Boosting Decision Tree (LightGBDT). The data set contains 230,608 trips in total recorded by 69,918 individuals. As can be expected the majority of the data (52.3%) showed mode preference towards the Car. This results in the dataset being imbalanced, to overcome this obstacle Kashifi et al. (2022) experiment with both over- and under sampling strategies. The research also attempts to solve another downside of using

ML, namely the lack of interpretability of the model. In order to solve this problem, the researchers implemented SHapely Additive exPlanation (SHAP), this allowed them to display how large each features influence is on what MC someone chooses.
Kim (2021) also recognizes the difficulty to appropriately interpret the relationship between input and output values when using ML methodologies, the so called "black-box" nature of ML and attempts to make a more interpretable model. In total, three model types were applied to the model: eXtreme Gradient Boosting (XGB), RF and Artificial Neural Network (ANN). For each model, the interpretability was enhanced by using various algorithms such as Accumulated Local Effects (ALE) in order to visualize the effect a input variable has on the alternatives: "the value of ALE measures the main effect of a variable at a specific value (or specific category) on the prediction.". This visualization allows the researcher to identify which factors are important in an individual's decision making process just as with RUM based models. The result of this ALE evaluation can even show a better representation due to its ability to display the non-linear relationship between variables and alternatives. Figure 2.3 shows an ALE evaluation on how the total travel time impacts an individuals choice to use certain alternatives. For example a low travel time has a negative impact on cars and a positive impact on transit and walking and thus people are more likely to use the latter two modes in situations where the total travel time is low. Meanwhile it shows that bike usage is nearly unaffected by the total travel time metric.

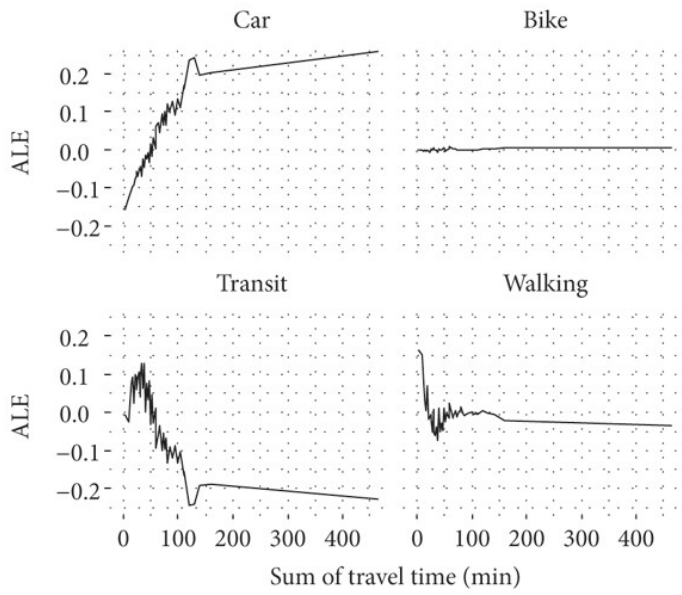


Figure 2.3: *An example of an ALE analysis for total travel time (Kim, 2021)*

F. Wang and Ross (2018) aim to truly test the performance difference between a ML and RUM based model. The research compares XGB to MNL on a household travel survey dataset. After filtering out invalid data points this resulted in a total usable dataset of 51,910 trips. However, F. Wang and Ross (2018) point out that the dataset is very unbalanced with car trips accounting for 83.20% of the dataset while biking has a mere 1.00%. Unlike Kashifi et al. (2022), there is no over- or under sampling tactics to counteract this unbalance. Instead F. Wang and Ross (2018) choose to train the model on two datasets, one with the complete dataset of 51,910 trips and another with the underrepresented bike alternative removed altogether. This method allows for some unique insights on how impactful an unbalanced dataset actually is for either model type. To compare the models, the dataset was split up randomly in 75% training data and 25% testing data 100 times for

each data set and subsequently the average prediction accuracy and robustness over these 100 runs was used. Overall, the XGB model performs better compared to the MNL for both datasets, while both models do better on the smaller dataset, the XGB model performs significantly better on it. This indicates that this model type is more sensitive to unbalanced datasets than the MNL model. Moreover, the research concludes that both models have their strengths and weaknesses in the training. For example, a mentioned strength of the XGB model is how easy it is to go through the fitting process where MNL requires careful monitoring to ensure that the model assumptions hold. Zhang et al. (2020) explore the potential of a so called Deep Neural Network (DNN), a neural network with multiple hidden layers. They claim that benefit of a DNN over a more regular Fully Connected Neural network (FCN) is its ability to deal well with large volumes of data and the versatile architecture. To test this, multiple model types are set up as comparison as well including also some RUM based models (MNL and NL) and a RF model. The DNN model used is one with two hidden layers, the first of these layers is designed such that it acts similar to the weights that are assigned in a RUM based model, this can be seen in Figure 2.4. The DNN model is then, along with the other models, trained on a large dataset. This dataset is a combination of multiple data sources and is eventually split into a 80-20% split for training and testing data respectively. Finally the accuracy of each of the 5 models is calculated, together with the so called Welch's t-test this shows that the DNN model performs significantly better. Closer compared to the FCN it is also shown that the DNN model saturates much later in the training process. Because of the earlier mentioned structure of the DNN model, it is also easier to extract data from the trained model on which attributes are influential for a decision maker similar to how this would be done with RUM based models.

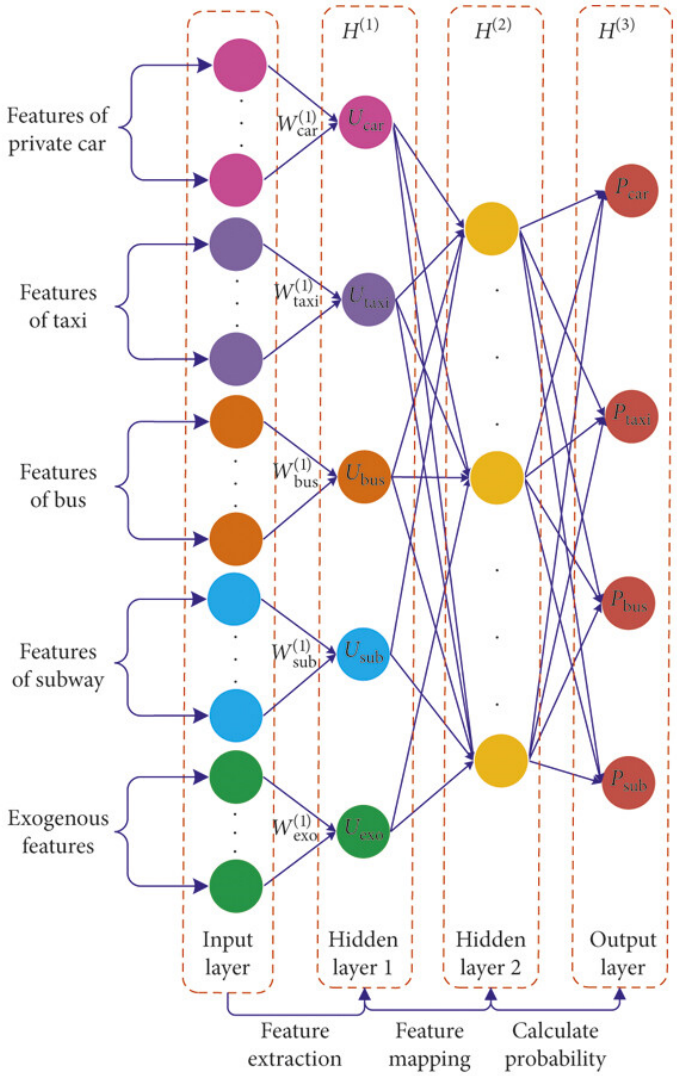


Figure 2.4: *A simplified view of the proposed DNN structure (Zhang et al., 2020)*

Salas et al. (2022) aim to make an accurate comparison between various ML and the MNL and MXL methodologies. The researchers want to measure how well each model performs by looking at how accurate they are on a test set of the data. The data is split up according to a stratified K-fold cross-validation scheme. This allows the researchers to train and test the model multiple times on different sets of the data (K-fold cross-validation) while ensuring that each training and test set remains representative to the entire data set (stratified). After evaluation they conclude that ML methodologies score significantly better than the MNL mainly due to unobserved heterogeneity in the datasets. However, the MXL can already easily account for this. Finally they do conclude that the Artificial Neural Network (ANN) does outperform all other evaluated methodologies, including the RUM based ones. This greater performance mainly shows in its ability to accurately predict choices within the test sets.

Li et al. (2024) have a very different approach compared to most other researches that aim to use ML on MC problems. Rather than using an extensive Revealed Preference dataset, they opted to use a questionnaire to gather Stated Preference data from individuals within Xi'an City in China. As a result the final dataset only consisted of 985 pieces of data, combining this with the 8:2 training:testing ratio this only leaves 788 data entries to train the XGB model with. This is even considered a low

number of data entries when using RUM based methodologies. Despite that, the researchers managed to get a 63% accuracy on their model.

2.5. Full Comparison

The literature reviewed in this chapter brings up a lot of different insights, challenges and benefits when it comes to the development of Mode Choice modeling. In this section the most important ones are highlighted and discussed more in depth.

2.5.1. Black-box

One of the most well known down sides of ML is its black-box nature. Most ML model types have a vast amount of variables that are set during the training phase. Often these variables do not have a clear indication on how exactly they affect the output data and it is thus extremely difficult to figure out which attributes are important. For many problems, this is not really an issue, for example when using ML to figure out whether a picture is a dog or a cat it is not really important which pixels lead to the conclusion and why. However, in the Choice Modeling field, determining which attributes are important and which aren't is generally actually one of the main research goals. It is thus important to be able to look into the black-box of a ML model. This problem is also highlighted by Van Cranenburgh et al. (2022) and the researches from Kashifi et al. (2022) and Kim (2021) set out to find good solutions to this problem. A variety of algorithms exist that can help uncover what attributes are more important than others such as SHAP and ALE. Most of these algorithms are able to give such a clear indication by changing specific variables within the trained model and identifying how that changes the results. While this method is a lot more time extensive than simply looking at the trained variables as would be sufficient with RUM based models, the results can actually be more detailed. This is a result of the non-linearity that ML models can have, Figure 2.3 is such an example of this.

2.5.2. Model types

As stated by Hillel et al. (2021), there does not really exist a good framework for using ML algorithms on MC problems. This lack of framework causes researchers to make critical mistakes with how they treat data and train/validate their models. This problem is exaggerated by the vast amount of different model types that exist within the ML sector. The research of Hillel et al. (2021) already mentions five types of supervised learning methodologies. Most of these methodologies also have different model types that fall inside, Kalimi (2023) even lists 50 different model types. Each model obviously has its own strengths and weaknesses but there seems to be little to no reasoning within the literature to pick one methodology over another one and why.

2.5.3. Unbalanced datasets

Another overlapping issue within the literature is the occurrence of unbalanced datasets. In order for the chosen ML algorithm to identify which attributes indicate what alternative is most likely the model requires a significant amount of data. As mentioned in Section 2.4.2 the theory-driven nature of RUM models allows them to get away with a much smaller sample size, providing that the granularity of the data is large enough. The data-driven nature of ML needs to make up for this lack of context in the form of more data. However a lot of MC problems tend to show a clear overrepresented alternative and also underrepresented ones, this can be seen in the datasets used by Kashifi et al. (2022) and F. Wang and Ross (2018). The former attempts to alleviate the problem by experimenting with over- and under-sampling from the dataset in order to make it more balanced. Both methods have their risks when it comes to how well the model will end up performing but it does allow the underrepresented alternatives to remain. The need to keep such alternatives in the

evaluation can be very important, for example when wanting to identify what attributes cause an alternative to be so rarely used compared to other more popular alternatives.

2.5.4. Research Goals

As shown in Section 2.5.2 and as discussed in the papers from Section 2.4.1 it is clear that there is no coherent structure as to which ML methods are to be used. As a result, a significant part of the existing literature either focuses on evaluating different methodologies or comparing them with others. Other researches that focus on extracting usable results out of datasets also struggle with finding the best working model for their data. For example, the research of Li et al. (2024) went through thirteen different models to eventually pick the most accurate one for further analysis.

2.5.5. Research gap

It is clear that using ML for MC studies already is a more common practice with plenty of researches focusing on finding different methodologies that work. However as mentioned in Subsection 2.5.4 the main focus is still on comparing different ML and logit based models with each other. There is little known on how these models could potentially be improved by applying theory based knowledge used in RUM based models. Applying methods such as the use of latent variables to help train the model could prove beneficial in ML just as they can be when applied to other methodologies.

3

Model Selection and Development

In this chapter the methodological approach is explained. In Section 3.1 the base models are selected and how each model works is briefly explained. Afterwards in Section 3.2 there is an in depth explanation on how the datasets are prepared for use in the ML models. Furthermore in Section 3.3 latent variables are introduced to explore whether applying expert knowledge to the dataset helps improve the results. Finally in Section 3.4 the technique to how the results of the trained models are analyzed is explained.

3.1. Machine Learning model types

Within the ML paradigm there are a vast amount of different methodologies with each having different strengths and weaknesses and thus different use cases. Most algorithms fall under one of two main branches within ML: Supervised learning and Unsupervised learning. The key differences of these two types of ML is that Supervised learning works with datasets that have input data that is paired with defined output labels. This type of dataset structure corresponds with how data in MC problems is represented. Historical data shows how a person or goods traveled between locations within an area of interest, this data represents the output label. All additional data collected describe potentially relevant factors and attributes for this trip, this represents the input data. A Supervised learning model can, when given enough data, learn how each input parameter influenced the agents decision to choose for a specific mode of transport.
Within the Supervised learning category there are a vast amount of different ML methodologies. In Subsection 2.5.2 this was listed as a key challenge when it comes to using ML on MC problems. As discussed there is no clear guideline on which model types work best under what circumstances. In order to ensure good results this research will therefore compare the results of the most commonly used methodologies as found in the reviewed literature in Subsection 2.4.2: Random Forest Decision Tree (RFDT), Gradient Boosting (GB), Support Vector Classifier (SVC) and Multilayer Perceptron (MLP). These model types will finally be bench-marked against a simple Logistic Regression (LR) model.

3.1.1. Gradient Boosting

Gradient Boosting consists of two different concepts combined: A boosting process and a gradient descent. The boosting part uses a so called Weak learner as the basis for the model. This Weak learner is evaluated based on the residuals also called the Loss function. A second Weak learner is now trained, however this model will no longer try and predict the expected target values. Instead this model will try to predict the residual of the first model based on the same input data. Equation 3.1 shows the mathematical approach to this. Here $F(x)$ represents the final model, $f_1(x)$ refers to the initial weak learner that tries to predict the target values based on the input x. All further $f_i(x)$ functions are new iterations that are added to the total function and represent the weak learners that try to predict the residual of the previous total function based on the input data x.

$$F(x) = f_1(x) + \sum_{i=2}^{m} f_i(x) \tag{3.1}$$

The gradient descent part of the model refers to how each new weak learner $f_i(x)$ is constructed. For each data point the model evaluates the gradient of the Loss function and based on whether this gradient is negative or positive, the model can now adjust the next iteration in the correct direction in order for the loss function to decrease.

3.1.2. Random Forest Decision Tree

A Random Forest Decision Tree model consists of multiple Decision Tree (DT) models where each tree is trained on a different subset of the total dataset. Sequentially each node on a tree will also take a random subset of features to consider for a split. With these two specifications each tree will be trained as any other DT where each split is based on trying to split the dataset in the most homogeneous way possible (when considering the target output labels or values). In the end the average result of all trees is taken as the final output of the model.

3.1.3. Multilayer Perceptron

A Multilayer Perceptron is a type of Artificial Neural Network (ANN). The architecture of a MLP consists of an input layer, one or more hidden layers and an output layer. The input layer has individual nodes for each input feature, each node is thus specifically assigned to one of the features in the dataset. The hidden layer(s) receive values from all nodes of the previous layer multiplied by some weight, this weight corresponds to how important the value is. Additionally the model has the opportunity to add a bias to a node which is independent from the input value. Lastly the final value is calculated for a node and it is passed through an activation function to introduce non-linearity into the model. The final layer is the output layer, here all values are combined and passed through another type of activation function (for example a Sigmoid function) which will determine the final output value of the model. The learning process of a MLP consist of changing the weights in the model, similar to GB this is done via gradient descent to identify which way certain values need to move to minimize the loss function.

3.1.4. Support Vector Classifier

A Support Vector Classifier model tries to create a boundary (or hyperplane) that will separate the data in its different classes. This boundary is created in order to maximize the distance between the hyperplane and the data points from the classes, the points closest to the hyperplane are also called 'support vectors' and they are the most important in forming the final hyperplane. However most datasets are impossible to perfectly separate due to outliers and noise present within the data, to account for this the model is allowed to draw the hyperplane in a way that some data is misclassified. The objective of the model is now to maximize the distance between the hyperplane and its support vector combined with minimizing the number of misclassifications.

3.1.5. Logistic Regression

To provide a baseline for evaluating the performance of ML models, a Logistic Regression (LR) model is trained alongside them. LR is particularly relevant in the context of mode choice modeling because it shares strong conceptual similarities with RUM based models such as the MNL model, which is widely used in transportation research. Both models are based on probabilistic decision-making, using the softmax function to estimate the likelihood of selecting a particular alternative from a finite set of discrete choices.

As a result, LR can serve as an effective benchmark to assess whether more complex ML models like GB or MLP offer substantial performance improvements over this well-understood baseline.

Using LR in this way allows for direct comparison with traditional discrete choice models and provides insight into whether the added complexity of ML methods is justified by significant gains in predictive accuracy or interpretability.

3.1.6. Single class classification

To help with understanding results and pinpointing strong and weak areas of the above mentioned models, a separate model will be trained for each alternative class within the dataset, also called

a One-vs-Rest. In Figure 3.1 the basic structure of this approach is visualized, first the dataset is split into an input and observation set. The input set contains all the features used for the mode choice evaluation and are thus also the variables that will eventually be evaluated and ranked on how important they are. The observations are the mode choices made corresponding to the input data. This data is split into a training and testing set where the training set is used to train the ML models and the testing set is used to evaluate how accurate the model is. Because the target will be the observations of a single mode, this structure will be applied separately for all mode choices considered in the data set. The 80/20 split training and testing sets will be identical for all mode choices however, this ensures consistent results for all modes and allows them to be compared with each other.

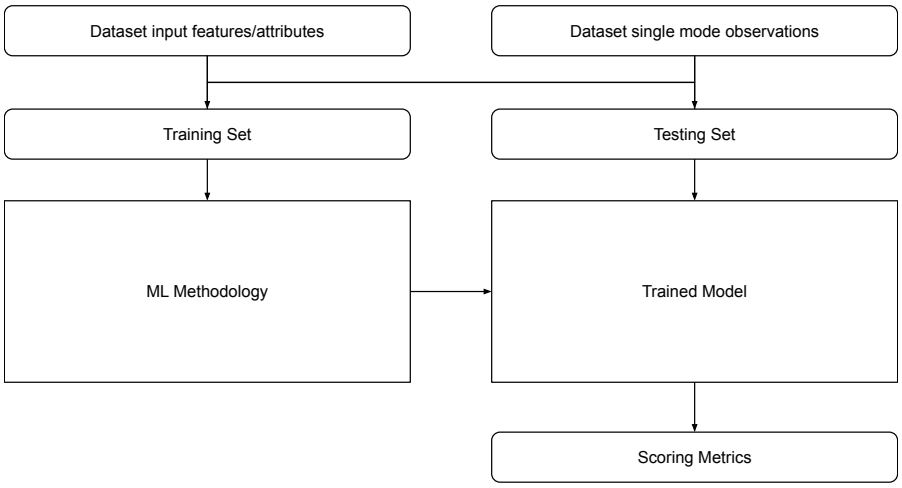


Figure 3.1: *Schematic of the base model setup.*

3.2. Dataset

The most important and time consuming part of using ML is the preparation of the to be used dataset. In this section some common problems with datasets will be discussed and possible solutions to these problems are explored.

3.2.1. Aggregate vs Disaggregate

Data concerning MC is collected in a variety of different ways. The two main techniques of acquiring data are Stated Preference and Revealed Preference. Stated Preference data is typically collected through surveys where individuals or agents are asked hypothetical scenario's in which they can state what alternative they would choose. This allows for the analyst to handcraft specific scenario's to analyze and conduct a targeted analysis. Revealed Preference is a more commonly used method when working with MC problems. Revealed Preference is data collected through observing individuals or agents making actual decisions in real world scenario's. This in turn allows for the analyst to look at specific areas of interest to uncover why these individuals choose for specific alternatives.
While Revealed Preference data can also be collected via surveys just like with Stated Preference, there are also other options to gather the required data. (Automatic) data gathering through companies or institutions such as railway operators can provide large datasets that cover the entire system. This however comes with the downside that the dataset is often aggregated. This often also changes how the data is represented, instead of having individual data points the information is aggregated and summarized. Not only does this result in specific individual data, such as socio-economic data, to be unavailable, it also results in the individual trips to be displayed as total traffic flows. For example, instead of having a separate data entry for everyone that traveled between point A and B there

will only be a couple of entries, each showing the total flow of traffic between these two points for a different mode of transport. This results in the dataset containing significantly less information compared to disaggregate datasets even though the amount of data represented within the dataset is often much higher.
This difference in how data is represented also warrants a different approach to analyzing the MC problem. With disaggregate level datasets it is logical to use classification supervised learning, it is known for each individual what mode of transport is chosen and it can thus be labeled with either a 0 or a 1. Using classification on aggregated datasets can however result in a fairly large loss of information for the model to learn. Combined with the already low amount of information that can be retrieved out of such a dataset can cause the results to be incorrect.
The solution would be to treat the problem as a regression problem. Instead of learning how attributes result in a specific alternative being the most chosen one, the model will learn how the attributes impact the different total traffic flows of each different alternative. Table 3.1 is an example of such an approach, the displayed values are the target output values that the model will try to predict using the specific attributes of each Origin-Destination (OD) pair.

Table 3.1: *An example of an aggregated MC problem dataset showing the total flow for each mode over different OD pairs.*

OD-Pair	Mode A	Mode B	Mode C
1	5500	2500	1000
2	150	700	200
3	8000	2500	750
4	4500	2000	1500

To further simplify the model, the dataset can be modified to where the output target is represented as shares of the total flow over the OD pair as shown in Table 3.2. This ensures that the model does not need to learn the specific absolute value of the traffic flows that can vary across the entire dataset.

Table 3.2: *The same example as shown in Table 3.1 but with its output targets converted into shares of the total flow along an OD pair.*

OD-Pair	Mode A	Mode B	Mode C
1	61.11%	27.78%	11.11%
2	14.29%	66.67%	19.05%
3	71.11%	22.22%	6.67%
4	56.25%	25.00%	18.75%

3.2.2. Imbalanced datasets

A common issue with datasets when using ML is dataset imbalance. For a ML model to learn the right relations between input variables and the desired output value there needs to be a good variety within the output in the training set. If this variety is not present the model wont learn how attribute values caused the target output. For example, if a dataset contains 100 entries of which alternative A is the correct output 99 times, the trained model would get a 99% accuracy rate when predicting that all outputs are alternative A. While this accuracy rate looks good on paper, the model shows that it is unable to accurately predict what situation leads to alternative A not being correct.
For MC problems imbalanced datasets are very common, as discussed in Subsection 2.5.3 one possible solution is oversampling.
An interesting technique for oversampling is the use of Synthetic Minority Over-sampling TEchnique (SMOTE). SMOTE will create new data points for the minority class rather than using exact duplicates for oversampling Chawla et al. (2002). To create these synthetic data points SMOTE will

look at a few closes neighbors of each existing data point in the dataset and create new ones in between them as shown in Figure 3.2.

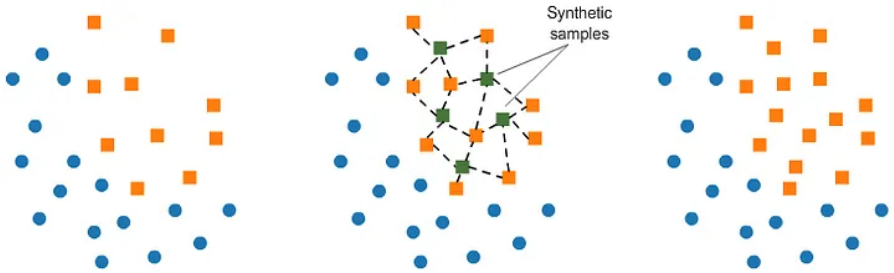


Figure 3.2: *An example of how SMOTE oversamples the minority (orange) class to make the dataset more balanced. Prudhvith, 2022*

The overall structure of the model will only change slightly as displayed in Figure 3.3, only the training set will be altered with the SMOTE technique while the testing set remains clean. This model will be used in addition to the model displayed in Figure 3.1, the results from both models can subsequently be compared to determine the influence SMOTE has.

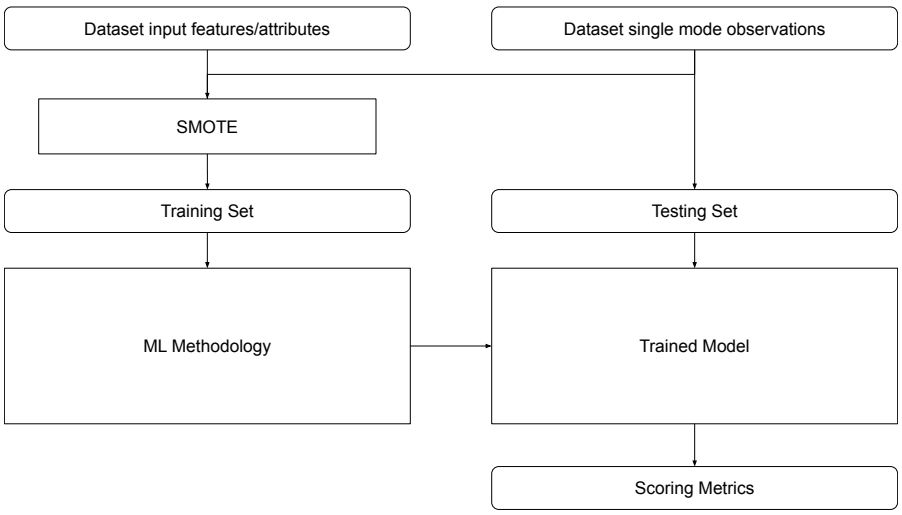


Figure 3.3: *Schematic of the model including SMOTE applied to the training set.*

3.2.3. Feature selection

An important step in any MC problem is selecting what features to include in the model. Generally speaking the more features added the better the model can perform as more relations between features can be created to distinguish alternatives from each other. However data collection is easier and cheaper with less features. Therefore it is important to find out if the developed models also have a satisfactory performance with a low number of features.
To evaluate the performance there will be several different so called 'experimental plans' where each experimental plan uses a different set of features. Most datasets used for MC problems are relatively similar in what features or attributes are included, however they will often also include relatively unique variables that are specific to, for example, the region of interest or chosen alternatives to be evaluated. The features can however be grouped together in different sets:

- **Attributes of the alternatives**: This group of attributes is specific to one or more alternatives and includes variables such as cost, distance, travel time, etc.

- **Socio-economic features**: This group of features are related to the individuals and describes their socio-economic background with variables such as age, gender, profession, private vehicle ownership, etc.
- **Cargo features**: This set of features describes the type of cargo being transported such as whether it is bulk, break bulk, container, etc. These features are of course only used when the MC problem concerns cargo shipment and replaces the socio-economic features that this type of problem does not have.
- **Trip specific attributes**: These attributes give information about the reason for a specific trip and the region it is held in.
- **Indicator variables**: This set of variables can show an individuals preference towards certain themes related to the different alternatives such as climate.
- **Latent features**: This is a set of variables not directly included in the dataset but rather created with the help of other features and attributes. Latent variables inherently don't represent anything specific but can still be useful for the models to create a more accurate result.

The different experimental plans will be created from a combination of above mentioned feature groups or different modeling strategies and are explained below, in Table 3.3 all experimental plans are displayed and for each of them the exact composition of feature sets used is shown.

Experimental plans:

1. **Full feature plan**: This experimental plan will use (almost) all available features and attributes from the dataset and the results of this plan are used to compare the other experimental plans with.
2. **Limited feature plan**: For this plan only the Attributes of the alternative and trip specific attributes will be used and only a limited number of socio-economic features. This is to simulate a situation where data collection is limited to automatically gathered information and thus where personal socio-economic features are not available due to, for example, privacy concerns.
3. **Latent variable plan**: In this experimental plan the Limited plan is expanded with latent variables. This experimental plan is used to test whether theory based knowledge can improve the results. The latent variables are further explained in Subsection 3.3.1.
4. **CNN plan**: Similar to the Latent variable plan this experimental plan is also an expansion on the Limited plan, however the addition now are variables created as a result of a Convolutional Neural Network (CNN) model with the help of the indicator variables. The results of this experimental plan can show whether combining data to form latent variables can be done with a CNN approach. This model is further explained in Subsection 3.3.2.
5. **Logistical regression model**: This model uses the exact same features and attributes as the Limited plan, however instead of using the ML methodologies shown in Section 3.1, a logistical regression model will be used. This will serve as an other way to compare results to.
6. **One-vs-One model**: Because of the required interpretability all above plans use the mentioned One-vs-Rest strategy as mentioned in Subsection 3.1.6. This model will use the same features and attributes as the Limited plan but the training process uses One-vs-One instead. This model can be used to compare how much performance is gained or sacrificed compared to the One-vs-Rest strategy and will thus also be run in tandem.

Table 3.3: *Overview of the experimental plans and feature sets used. (*: Partial use, **: Used to create the latent features, ***: Indirectly used to determine variables for the latent features)*

	Full Feature Plan	Limited Feature Plan	Latent Variable Plan	CNN Plan
Attributes of the Alternatives	x	x	x	x
Socio-Economic\Cargo Features	x	x*	x*	x*
Trip-specific Features	x	x	x	x
Indicator Variables	x		***	x**
Latent Features			x	x

3.3. Latent features

Latent features can provide useful data points for attributes that are inherently difficult or impossible to quantify on their own. In this research two sets of latent features will be used to study the impact they have on the overall results of the model.

3.3.1. Latent variables

The first latent feature set to be used is one developed by Atasoy et al. (2013). The latent variables are created based on a couple of Socio-economic variables with the help of some Indicator variables. With this technique it is possible to, once training is complete, get additional variables that tell how 'Pro-car' and 'Environmental friendly' an individual is without needing to know their responses to the indicator statements. In turn these latent variables can thus help the model in estimating what mode choice the individual used.
For this research these latent variables can be used as well to identify if they have the same positive impact on the ML models or if the model is already capable of making these connections during the training process without the help of additional input variables. The latent variable model will both be used in combination with the SMOTE technique mentioned in Subsection 3.2.2 and without. The model schematic is shown in Figure 3.4

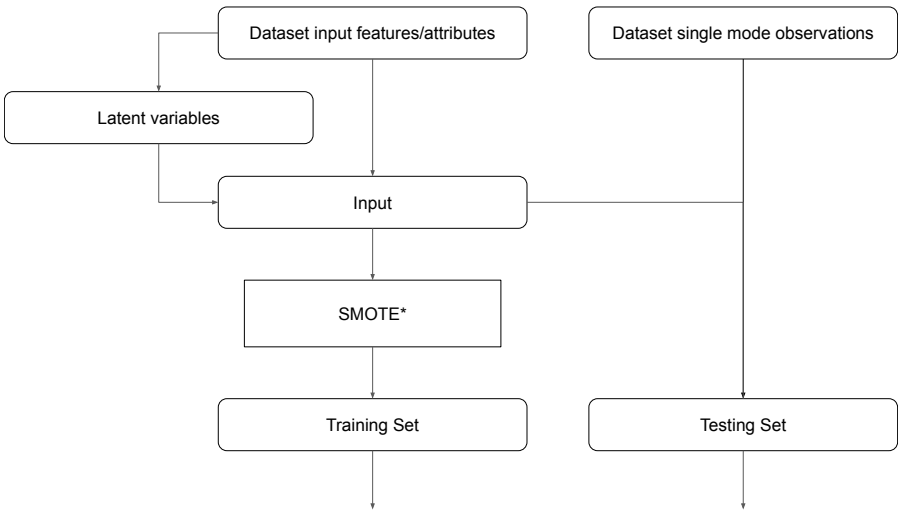
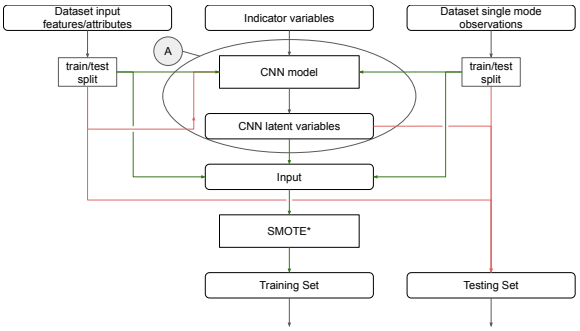


Figure 3.4: *The top part of the model schematic with latent variables. (*: The model will also be used without SMOTE)*

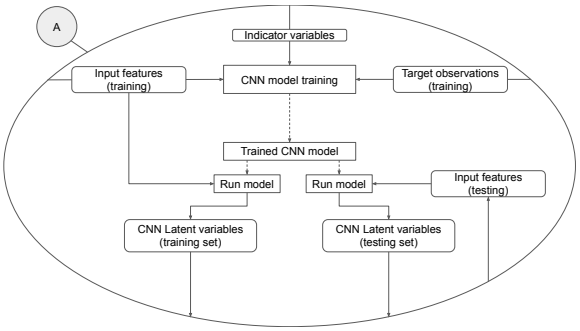
3.3.2. CNN model variables

In addition to the above mentioned latent variable model, a similar experiment will be done using a CNN model. The goal remains the same where latent variables are created and used to possibly

improve the ML models. As mentioned in Subsection 3.2.3 the CNN model will use the indicator variables from the dataset to help create the CNN latent variables. In Figure 3.5a the revised model architecture is shown where the CNN model also uses the indicator variables as an input in addition to the normal (limited) features and the single mode target observations. Important to note is that unlike with those 2 inputs, the CNN latent variable train/test split is not created using the same dataset. Instead the features are created by feeding the trained CNN model different inputs, namely the input training features and the input testing features as shown in Figure 3.5b.



(a) *The top part of the model schematic with latent variables created with a CNN model. Green lines indicate training data and red lines indicate testing data. (*: The model will also be used without SMOTE)*



(b) *Detailed view of CNN implementation. (note: input testing set flipped for ease of viewing)*

The CNN model itself is created with the goal to only use the indicator variables during training. In order to achieve this the model is initially split up in two separate input branches, one for the regular input features and another for the indicator features. Both branches go through various CNN layers before finally being merged and fed into the single output layer that contains the mode choice target values. In Figure 3.6 the full structure of the CNN model is displayed. As can be seen there is a special layer within the regular input branch (left) that extracts the latent variables, during training this layer is influenced indirectly by the indicator variables but after training these variables no longer influence the results in this layer anymore. Because of this the indicator variables are no longer needed after the model is trained, this is in line with how the latent variables are created as in Subsection 3.3.1.

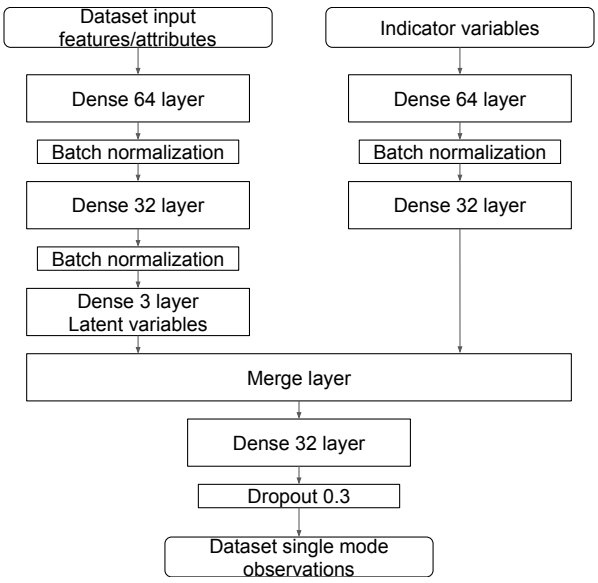


Figure 3.6: *The full structure of the CNN model used to create additional input variables for the main ML models.*

3.4. Model analysis

Analyzing a model consists of two important parts:

- Analyzing the accuracy and overall performance of the model
- Analyzing the results of the model

In Subsection 3.4.1 different methods for scoring the accuracy and overall performance are explained, in Subsection 3.4.3 the tools used to analyze the results of the model are shown.

3.4.1. Scoring Metrics

It is extremely import to verify how well the created model performs on a test dataset as this directly reflects how well conclusions can be made based upon the results of the model. After all a model that is not able to make accurate predictions has not learned any connections between the input attributes and the available alternatives. It is thus important to evaluate the prediction accuracy of the different models and compare those with each other to determine which model is 'better'.
There are various scoring methods available for both regression and classification models (Bajaj, 2025). Below the scoring metrics used to evaluate the model performance in this research are shown and explained.

Confusion Matrix

A confusion matrix is a way to visualize how well a model is able to predict binary targets. For a single binary target there are four possible classification states: True Positive (TP), True Negatie (TN), False Positive (FP) and False Negatie (FN). The prerequisites for each of these states are displayed in Table 3.4. A perfect model would have both the FP and FN values at 0, indicating that all predictions are correct. In reality this is not possible to achieve with complex datasets and thus the goal is to minimize both these values and therefore maximizing the TP and TN.

Table 3.4: *The classification names for the four possible prediction states.*

Classification Name	Predicted Target Class	Actual Target Class
True Negatie (TN)	0	0
False Positive (FP)	1	0
False Negatie (FN)	0	1
True Positive (TP)	1	1

However because the values for TN and TP are also dependent on how many total observations there are in the test set, it can be helpful for additional visual improvement to display the four values as percentages. This can be done by taking the sum over each row in a confusion matrix as shown in Table 3.5, and divide the values in that row with the sum. For example the TN value would be calculated as shown in Equation 3.2.

$$TN_{\%} = \frac{TN}{TN + FP} \tag{3.2}$$

Table 3.5: *The layout of a confusion matrix.*

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Scatter Plot

When dealing with a regression problem it is not possible to use a confusion matrix. To make it easy to still visualize the performance of the model a scatter plot can be used instead. In this plot the x-axis corresponds to the predicted values and the y-axis the actual values. Each observation in the test set will be plotted in this graph, in a perfect model all observations will be plotted exactly on the lin $y = x$ as this indicates that the prediction is exactly equal to the actual value. In reality it is thus easy to see how well a model performs by looking at how close the observations are to this $y = x$ line.

F_1-score

The F_1 score is a useful metric when dealing with binary classification models, specifically for models applied to unbalanced datasets. The scoring metric combines both the precision (Equation 3.3) and the recall (Equation 3.4) values into a single score by taking the harmonic mean of the two as can be seen in Equation 3.5 (Van Rijsbergen, 1979).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.4}$$

$$\begin{aligned} F_1 &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \end{aligned} \tag{3.5}$$

This harmonic mean will cause the F_1 score to only be high when both the precision and recall are performing well. With unbalanced datasets this is useful as a poorly performing minority class will actually impact this score negatively by a much greater amount compared to looking at just the precision score (Grandini et al., 2020).

Balanced Accuracy score

The Balanced Accuracy (BA) score is a balance between the recall for both the positive as the negative class in a model. Where the classical recall score only looks at TP (as shown in Equation 3.4), the BA also calculates the recall for the TN class. Both the recall scores are balanced together as shown in Equation 3.6.

$$\text{BA} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{3.6}$$

This scoring metric is perfect for the use on unbalanced classification datasets as it gives an equal weight to all classes and is therefore not skewed towards how the majority class performs (Grandini et al., 2020).

R^2 score

The R^2 score, also known as the coefficient of determination, is a common metric to evaluate regression models. The score can directly tell how well the model captures the variance in the data and is formulated as displayed in Equation 3.7. In this equation y_i represents the actual observed data point i, $\hat{y}_i$ is the predicted value corresponding to y_i and finally $\bar{y}$ is the mean of all the actual values y.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3.7}$$

A prefect model would have an R^2 score of 1 as there difference between the predicted and actual value would always be 0 and thus 100% of the variance is explained by the model. Furthermore a model that has an R^2 score of 0 has not captured any of the variance and is in performance similar to always predicting the mean value regardless of input data.

Root Mean Squared Error score

The Root Mean Squared Error (RMSE) score is a value that explains the average magnitude of the errors between predicted and actual values as formulated in Equation 3.8 where y_i and $\hat{y}_i$ represent the same values as in Equation 3.7.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.8}$$

The benefit of this scoring metric is that the RMSE score has the same unit as the target variable and can thus be easily related to the data itself. Another feature of the RMSE score is its sensitivity to large errors, on the one hand this can results in the model seemingly performing worse than it actually is due to a few large outliers. However this can also be a benefit when working with imbalanced datasets where it is important that the model also works well on a minority class or group. A relatively small group of large errors would have a great impact on the RMSE score.

Mean Absolute Error score

The Mean Absolute Error (MAE) score is very similar to the RMSE score. However in this approach all errors have a similar impact on the score and a small group of larger errors does not impact the score as heavily. The formulation for the MAE score is shown in Equation 3.9 where the values for y_i and $\hat{y}_i$ are again the same as in the previous equations.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{3.9}$$

3.4.2. Averaged results

To prevent the results from showing a unusual high or low score due to a favorable training and testing split, the models will be trained multiple times using different random splits. The eventual score will be the average over all trained models. To ensure that each model type and mode type can still be compared with each other fairly, each unique train-test is used on all of them before a new train-test split is generated.

3.4.3. Result analysis

SHapely Additive exPlanation

In Subsection 2.4.2 the use of SHapely Additive exPlanation (SHAP) is briefly mentioned as a tool to visualize the inner workings of a trained ML model. As the name suggests SHAP works on the principle of Shapely values a concept introduced in cooperative game theory Hart (1989). The idea behind this is that this value represents a fair way of how to distribute a reward among players based on their individual contribution to the game. When applying this to ML the players are the input features of the model and the value will represent how much each feature has contributed to the final output result. This property makes SHAP an extremely useful tool for MC problems, as with these studies the goal is often to find out how features influence an agents MC in order to better understand the system as a whole and possibly influence it.

4

Case Studies

To test how well the ML models formulated in Chapter 3 perform on MC problems, they will be applied to two existing cases. Firstly there is the Rhine-Alpine corridor dataset, explained in Section 4.1. This dataset is aggregated and can thus be used to show the potential of ML when used as a regression modeling tool. Secondly there is the Switzerland dataset, explained in Section 4.2. This is a more classic MC problem case study that is disaggregated and covers the MC behavior of passenger travel. This dataset is set up well to use ML on in the form of a classification model. After that inSection 4.3 some SHAP analyses are conducted to evaluate the models potential to actually be used as a tool for MC studies. Finally in Section 4.4 the results of the different models and experimental plans are compared and discussed.

4.1. Rhine-Alpine corridor dataset

The first dataset that will be used is the Rhine-Alpine corridor dataset (Nicolet et al., 2022). This dataset contains information on freight shipping along different regions along and close to the Rhine river. This corridor starts in Rotterdam and follows the river all the way up towards Basel. In this dataset three different modes of transport are considered: Road, Rail and Inland Waterways Transport (IWT). However only container shipping is considered in this case study and the use of dry bulk and other types of transportation are thus not included.

4.1.1. Dataset structure

The dataset shows the yearly container flow of two commodity types (Foodstuffs and Machinery) between 2011 and 2021 between set Origin-Destination locations. For each OD pair there are two entries, one for either of the commodities. The features and attributes recorded fall largely under the 'Attributes of the Alternatives' and 'Trip specific features' categories with most variables describing operating cost or travel time.
Because the data is collected in terms of total container flow rather than on a per container basis this dataset can be classified as an aggregate dataset. This means that the approach as described in Subsection 3.2.1 will be used in order to create the ML models. The dataset is made out of observations from 600 different OD pairs with two commodity types resulting in 1200 unique samples. Additionally there is eleven years of shipping data over these samples, however, there is only one variable available per OD pair for all included features and attributes. As a result the eleven years of data can not all be used as separate entries as there would not be any unique training input data and only one year can be picked as a target for the model, for this the year 2017 was chosen. In Table 4.1 the total shipping volume for the three transport modes are displayed, from this it can be concluded that there is a single majority class (Road) and two minority classes with specifically rail having only a tiny contribution to the total TEU flow. Additionally over the 1200 samples it is shown which mode holds the majority of volume shipped, these results indicate that if this problem were to be treated as a classification problem, the dataset imbalance would be even greater.

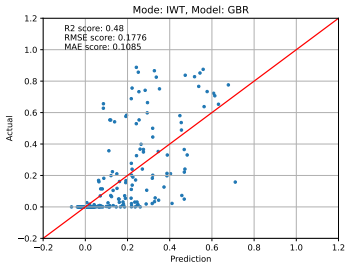
4.1.2. Base ML model

The initial full feature experimental plan results show poor performance on all three modes in Figure 4.1. None of the models is capable of reaching a R^2 score of 0.5 or higher indicating that the predictive power of the model is not sufficient. This is confirmed with the scatter plots where the two minority classes have most to all of the predictions in the lower half regardless of actual shares

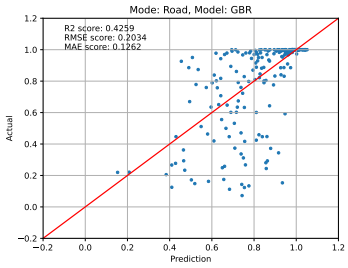
Table 4.1: *Shipping volume distribution over the three modes of transport during the year 2017 in the Rhine-Alpine corridor.*

Mode	Volume (TEU)	Share (%)	# of Max Share Cases	Share (%)
IWT	72,949,130	29.3%	148	12.3%
Road	169,685,385	68.2%	1037	86.4%
Rail	6,246,724	2.5%	15	1.3%
Total	248,881,239	100.0%	1,200	100.0%

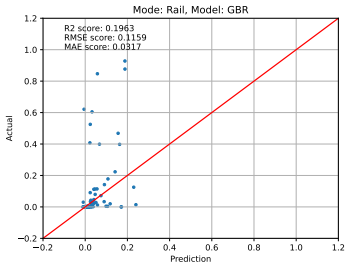
scoring higher on occasions. Meanwhile the majority class has the opposite problem where the model predicts very high even when the actual share is very low. From this it can be concluded that the class imbalance is too severe and the model only learns that the majority class always has a high share and the other two always have a low share and can not identify input variables as an indication to other results. As such for this research the usage of this dataset is discontinued.



(a) *Scatter plot showing the actual vs predicted shipping shares for IWT with a GB model.*



(b) *Scatter plot showing the actual vs predicted shipping shares for road transport with a GB model.*



(c) *Scatter plot showing the actual vs predicted shipping shares for rail transport with a GB model.*

Figure 4.1: *Scatter plots depicting the predicted vs actual values using a GB model with included the R^2, RMSE and MAE scores.*

4.2. Switzerland dataset

The second case study uses a Swiss dataset (Atasoy et al., 2013). This is a more traditional dataset in the sense that it covers the travel behavior of Swiss citizens in 2009 and 2010 across three different modes of transport: Private, Public and Soft (walking, cycling, etc.). This dataset has additional variables present in relation to the individuals and the alternatives compared to the Rhine-Alpine dataset from Section 4.1.

4.2.1. Dataset structure

This dataset has separate entries for a single trip conducted with on of the mentioned modes of transport, as such it can be classified as a disaggregated dataset. For each entry there are 110 features, attributes and variables collected that can be used in the model training. Moreover, these variables can be split up in such a way that they populate each category mentioned in Subsection 3.2.3. The three modes are however not balanced well. After filtering there are 1906 separate entries of which most represent a trip taken with private transport as can be seen in Table 4.2.

4.2.2. Logistic Regression model

Similar to the Rhine-Alpine corridor dataset the initial model is a LR model that will provide a base line comparison to the various ML based models. As such the model is trained twice, once with the full feature plan and once with the limited feature plan.

Table 4.2: *Positive labels in the dataset for each mode.*

Mode	Positive Labels	Percentage
Public	536	28.1%
Private	1256	65.9%
Soft	114	6.0%

Full feature experimental plan

Initially the full feature experimental plan is used with both the One-vs-Rest and One-vs-One approaches. The F_1 and BA scores for these experiments are displayed in Table 4.3. For both Public and Private transport the model scores well (>0.8) on both metrics indicating that is capable of detecting both the TP and TN labels well. However for the Soft transport mode the metrics are significantly lower, especially concerning the F_1 score indicating that the model has trouble identifying the TP labels.

Table 4.3: *Average performance metrics (F_1 and BA) for the LR model using the full feature experimental plan.*

	Public	Private	Soft	One vs One
F_1	0.8300	0.8121	0.6797	0.6999
BA	0.8496	0.8213	0.7899	0.7542

These assumptions can be confirmed when looking at the averaged confusion matrices for the three modes in Table 4.4. As can be seen the negative cases are predicted correctly between 80% and 90% of the cases for all three modes. This is similar for the positive cases with the exception of the soft transportation mode where the TP score is only 68.5%.

Table 4.4: *Averaged confusion matrices for Public, Private, and Soft mode choices using a LR model with the full feature experimental plan.*

Public

	PN	PP
AN	87.0%	13.0%
AP	17.0%	83.0%

Private

	PN	PP
AN	80.1%	19.9%
AP	15.7%	84.3%

Soft

	PN	PP
AN	89.5%	10.5%
AP	31.5%	68.5%

The One-vs-One approach shows similar results albeit slightly worse on all aspects. While the F_1 and BA metrics are similar for One-vs-One compared to soft transport (see Table 4.3), it is clear from the confusion matrix displayed in Table 4.5 that all TP scores are lower compared to the One-vs-Rest counterparts.

Table 4.5: *Averaged confusion matrix for a LR model using the One-vs-One approach and full feature experimental plan.*

	Predicted		
	Public	**Private**	**Soft**
Public	80.0%	13.3%	6.7%
Private	9.9%	81.7%	8.4%
Soft	10.1%	25.0%	64.9%

Limited feature experimental plan

Using the limited feature experimental plan reduces the number of available features from 110 to 22. Unsurprisingly the performance of the LR model worsens as a result, as shown in Table 4.6 the scoring metrics are lower for every mode compared to the results from the full feature experimental plan shown in Table 4.3.

Table 4.6: *Average performance metrics (F_1 and BA) for the LR model using the limited feature experimental plan.*

	Public	Private	Soft	One vs One
F_1	0.7732	0.7736	0.5923	0.6205
BA	0.8093	0.7954	0.8152	0.7452

Notably when comparing the averaged confusion matrices there is actually a significant TP rate increase from 68.5% to 86.4% within the soft transport mode class. However this increase is at the cost of a higher FP rate as well and, considering that this is a minority class, that increase leads to an overall worse performing model.

Table 4.7: *Averaged confusion matrices for Public, Private, and Soft mode choices using a LR model with the limited feature experimental plan.*

Public

	PN	PP
AN	77.9%	22.1%
AP	16.2%	83.9%

Private

	PN	PP
AN	83.2%	16.8%
AP	24.1%	75.9%

Soft

	PN	PP
AN	77.6%	22.4%
AP	13.6%	86.4%

4.2.3. Base ML model

The base ML model uses the four methodologies mentioned in Section 3.1 and the full feature plan in order to get a basic understanding of what each ML model is capable of.

Table 4.8 shows that, similar as with the LR model, both the public and private transport modes perform better than the soft transport and One-vs-One approach. Comparing the different ML methodologies with each other reveals that, while not by much, GB performs the best for all modes, approaches and on both metrics.

Table 4.8: *Average performance metrics (F_1 and BA) using the four ML models with the full feature experimental plan.*

Model	Public	Private	Soft	One vs One
GB	0.8797	0.8653	0.7960	0.7872
RF	0.8546	0.8406	0.7379	0.7411
MLP	0.8571	0.8275	0.7297	0.7118
SV	0.8370	0.8178	0.6923	0.6895

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8728	0.8598	0.7686	0.7648
RF	0.8305	0.8289	0.6845	0.7002
MLP	0.8522	0.8257	0.6986	0.6976
SV	0.8565	0.8258	0.7647	0.7429

BA score

When evaluating the averaged confusion matrix for the GB model we can see in Table 4.9 that the results are biased towards what is provided in the dataset. When comparing the confusion matrices with Table 4.2 it is clear that the class with the most positive labels (private transport) also has the

best TP score but the worst TN score of the three modes. For the soft transport mode (the mode with the least positive labels) this is opposite.

Table 4.9: *Averaged confusion matrices for Public, Private, and Soft mode choices using a GB model with the full feature experimental plan.*

Public

	PN	PP
AN	94.7%	5.3%
AP	20.1%	79.9%

Private

	PN	PP
AN	79.4%	20.6%
AP	7.3%	92.7%

Soft

	PN	PP
AN	98.4%	1.6%
AP	45.6%	54.4%

Finally the One-vs-One approach shown in Table 4.10, shows slightly better results in terms of TP rates for both Public and Private transport compared to the One-vs-Rest approach. Interestingly both the F_1 and BA score of the One-vs-One approach are lower compared to the One-vs-Rest approach as can be seen in Table 4.8.

Table 4.10: *Averaged confusion matrix for the GB model using the One-vs-One approach and full feature experimental plan.*

	Predicted		
	Public	**Private**	**Soft**
Public	81.4%	17.9%	0.8%
Private	4.9%	93.4%	1.7%
Soft	11.7%	34.3%	54.0%

The GB model is the best performing, however the other models are not far behind, especially the RFDT model. In the subsequent sections the GB model is used as the main model to compare with, additionally better performing metrics per mode are also highlighted. The performances for all experimental plans and methodologies can however be viewed in Appendix B.

4.2.4. Limited features

To mimic more easily acquired datasets the limited features experimental plan uses fewer features and variables as the input of the model. This benefits model training time but has the potential of producing worse results compared to the full feature experimental plan. The goal is to evaluate whether the model is still capable enough of producing results that are useful for a mode choice study.

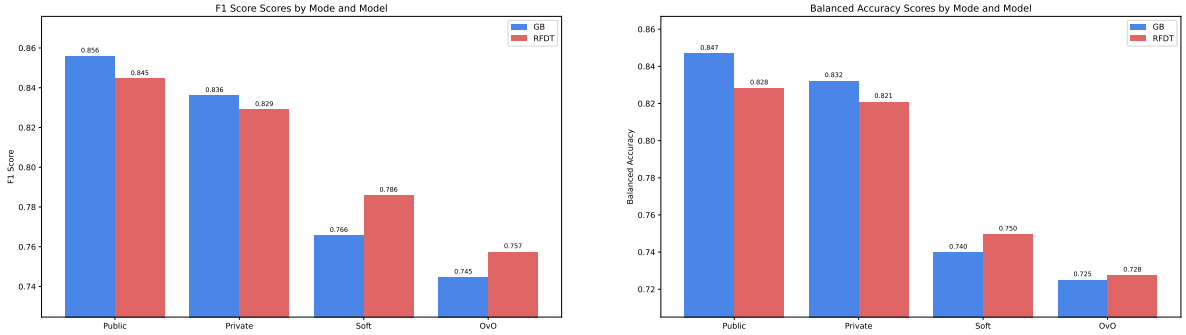
Similar as with the LR model, the performance metrics shown in Table 4.11 are lower compared to the full feature experimental plan (Table 4.8). However the decline is not as significant and the F_1 and BA scores for public and private transport are still greater than 0.8.

Table 4.11: *Average performance metrics (F_1 and BA) for the GB model using the limited feature experimental plan.*

	Public	**Private**	**Soft**	**One vs One**
F_1	0.8559	0.8360	0.7656	0.7446
BA	0.8469	0.8322	0.7398	0.725

With the limited feature dataset the RFDT model performs better for the soft transport mode and slightly when using the One-vs-One approach. The different values for both the GB and RFDT mod-

els is displayed in Figure 4.2.



(a) F_1 *scores of the GB and RFDT models.*

(b) *BA scores of the GB and RFDT models.*

Figure 4.2: *Differences in* F_1 *and BA scores between the best scoring models for the limited feature experimental plan.*

4.2.5. SMOTE implementation

Because minority classes are not an uncommon problem within the ML paradigm, it is possible to apply SMOTE to the dataset and balance out the classes by generating synthetic data. This can help the model with learning the positive labels associated with the minority class and thus provide better results.

Table 4.12 displays the performance metrics for the limited feature experimental plan over the three different modes and the One-vs-One approach. While the public and private transport modes see a minor decrease in performance as a result of SMOTE, the soft transport mode scores slightly better in terms of the F_1 score and makes a significant step up with BA. The BA score even rises above the one from the full feature plan. Similar as with the model without SMOTE, GB is the best performer

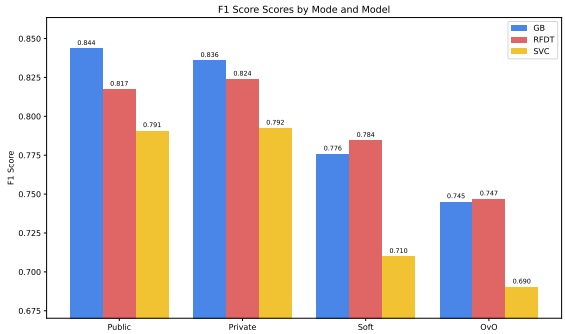
Table 4.12: *Average performance metrics (*F_1 *and BA) for the GB model using the limited feature experimental plan and SMOTE implementation.*

	Public	**Private**	**Soft**	**One vs One**
F_1	0.8435	0.8358	0.7757	0.7450
BA	0.8329	0.8389	0.7869	0.7523

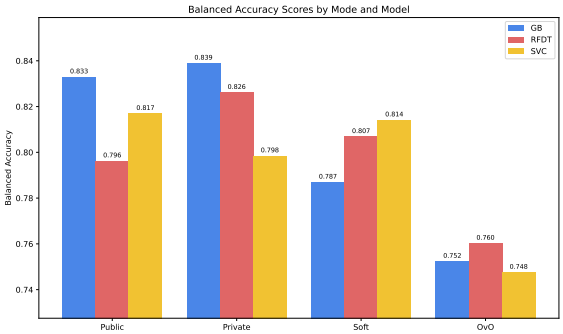
on both metrics for public and private transport but loses out again on the other two modes/strategies as shown in Figure 4.3. Again RFDT performs better however also SVC ranks highest but only for the BA score with the soft transport mode as shown in Figure 4.3b. However, Figure 4.3a shows that at the same time the model scores very poorly on the F_1 metric. It can thus be concluded that the TP rate has increased at the cost of a significant increase of FP rates as well.

4.2.6. Latent features

The latent variables were able to improve model performance for Atasoy et al. (2013) when applied to a logit model. In Table 4.13 the results are shown when the same variables were added as input features for the GB model. These results are without SMOTE modification to the training dataset as the results with SMOTE are similar but overall slightly more negative. Overall the metrics are near enough the same compared to only applying SMOTE as shown in Table 4.12. One slight difference is that the soft transport mode once again benefits. Furthermore when comparing this model with the other ML methodologies GB only loses out once. This is on the F_1 score for the soft transport mode, in turn GB still holds a higher BA score compared to RFDT as shown in Figure 4.4.

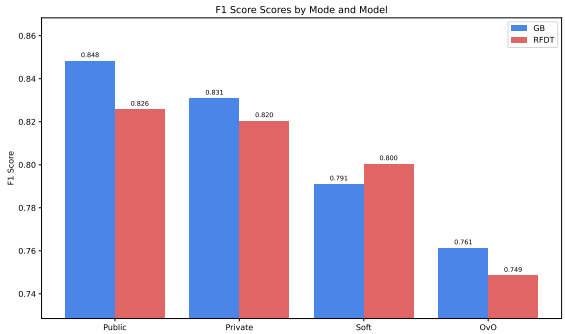


(a) F_1 *scores of the GB, RFDT and SVC models.*

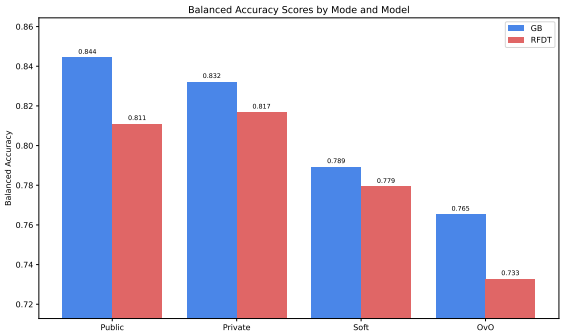


(b) *BA scores of the GB, RFDT and SVC models.*

Figure 4.3: *Differences in F_1 and BA scores between the best scoring models for the limited feature experimental plan using SMOTE.*



(a) F_1 *scores of the GB and RFDT models.*



(b) *BA scores of the GB and RFDT models.*

Figure 4.4: *Differences in F_1 and BA scores between the best scoring models for the limited feature experimental plan combined with the latent variables.*

Table 4.13: *Average performance metrics (F_1 and BA) for the GB model using the latent feature experimental plan without SMOTE implementation.*

	Public	Private	Soft	One vs One
F_1	0.8483	0.8309	0.7910	0.7613
BA	0.8444	0.8322	0.7893	0.7653

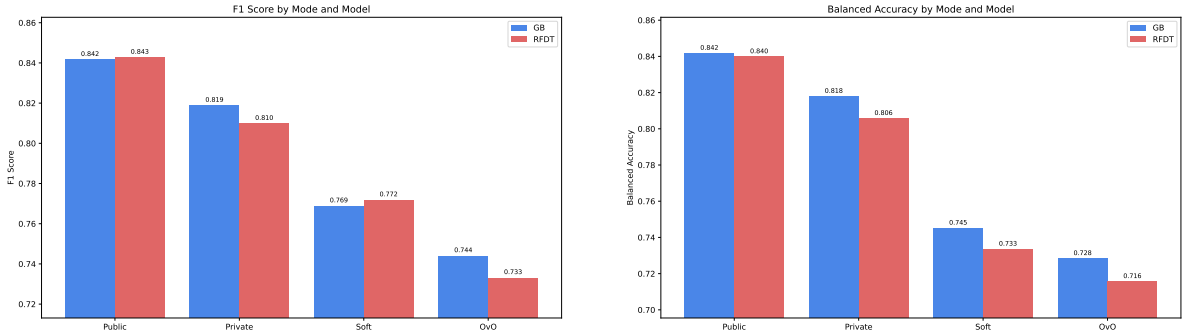
CNN features

Finally in this research there is also a more unique attempt at creating latent variables to potentially improve the model results, this is in the form of variables created using a CNN model. Similar as with the latent variable model, the performance was slightly better when not applying SMOTE on the training dataset. In Table 4.14 the performance metrics for the GB model are displayed. As can be seen the overall performance is worse compared to the latent variable model for all modes and the soft transport mode specifically.

Table 4.14: *Average performance metrics (F_1 and BA) for the GB model using the CNN feature experimental plan without SMOTE implementation.*

	Public	Private	Soft	One vs One
F_1	0.8416	0.8188	0.7688	0.7440
BA	0.8416	0.8178	0.745	0.7282

The other ML models do not perform much better with only the RFDT model having a better results with a higher F_1 score for the soft transport mode as shown in Figure 4.5a, however for the same mode the BA score is lower compared to the GB model as displayed in Figure 4.5b.



(a) F_1 *scores of the GB and RFDT models.*

(b) *BA scores of the GB and RFDT models.*

Figure 4.5: *Differences in F_1 and BA scores between the best scoring models for the limited feature experimental plan using CNN latent variables.*

To understand why the models with the CNN based latent variables performs poorly there are various options to inspect. The first possibility is that the CNN model did not train properly, to evaluate this an accuracy and loss function can be plotted. In Figure 4.6 it can be seen that the accuracy increased from the first training epoch to the last one as well as a decrease in loss. This indicates that the CNN was able to effectively learn the connections between the input and output variables. This leaves as options that the created latent variables are not of a good quality and as a result only confuse the model rather than help it.

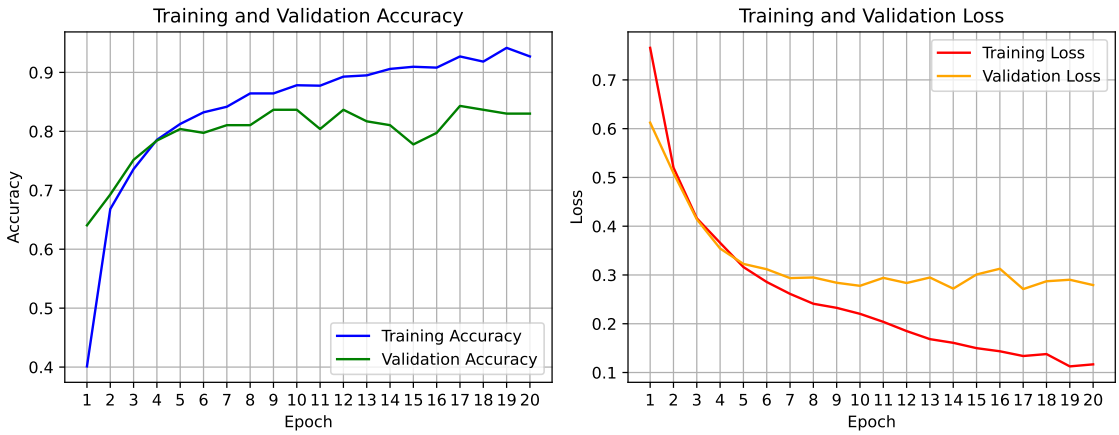


Figure 4.6: *Accuracy and Loss function for the CNN model training.*

4.2.7. Best performers

For each mode the best performing model can be picked based on the F_1 and BA score. However because these metrics are not tied together (i.e. F_1 can increase while BA decreases and vice versa) first a system needs to be introduced to identify what model is 'best'. For this research an equal balance between the two is chosen and thus the average score of the two metrics determines the best model. For this evaluation only models are considered that use the limited feature experimental plan or a derivative of it. The results of this evaluation are displayed in Table 4.15

Table 4.15: *Best performing models over all iterations (excluding the full feature experimental plan) for each different mode based on the average score between F_1 and BA.*

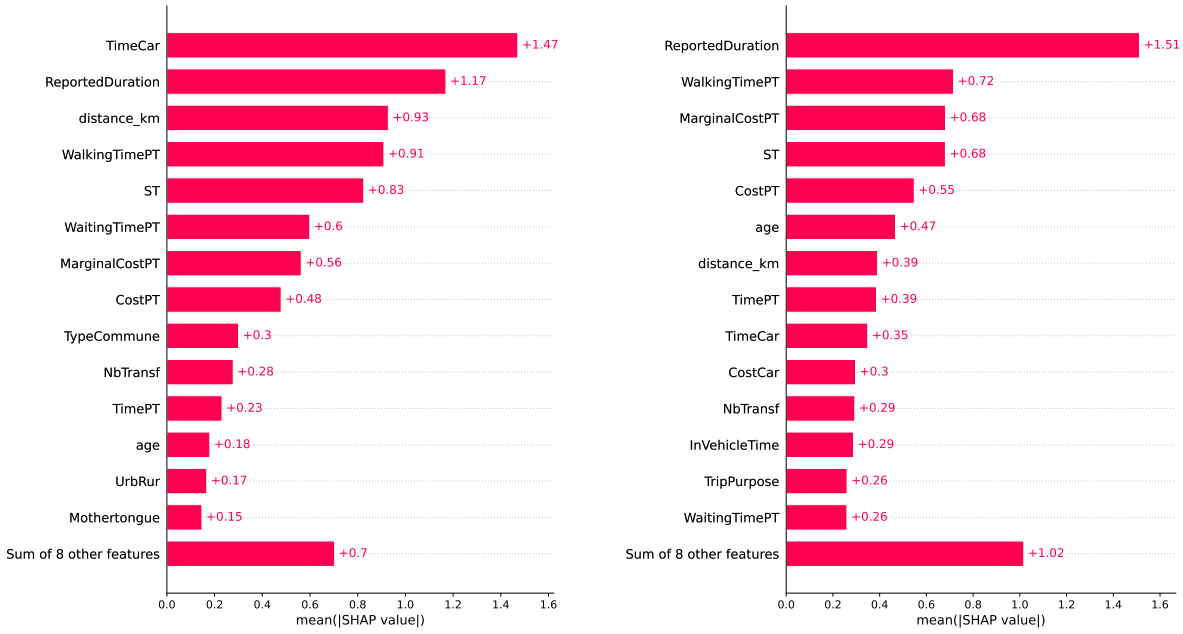
Metric	Public	Private	Soft	One vs One
Model	GB Limited	GB with SMOTE	RFDT with SMOTE	GB with latent variables
F1	0.8559	0.8358	0.7844	0.7613
BA	0.8469	0.8389	0.8068	0.7653
Average	0.8514	0.8374	0.7956	0.7633

This shows that overall GB is still the best performing model with it only being outperformed for the soft transport mode by RFDT.

4.3. SHAP analysis

For the SHAP analysis it is important to check which features are important for the outcome of the model. To compare how a ML model potentially improves the accuracy of a SHAP analysis the results for this are shown in Figure 4.7, Figure 4.8 and Figure 4.9. Here the SHAP analysis for the LR model is compared with the analysis for the best performing ML model as per Table 4.15 for all three modes of transport. From these figures it can be seen that the ML models show different importance levels for the evaluated features, important to note is that the values shown are not necessarily the same unit (scikit-learn developers, n.d.), thus comparing the displayed values between different model types can prove meaningless. However the ranking of importance, i.e. where a specific feature appears in the plot can still be compared as the overall meaning of the bars is still the same (highest mean absolute shap value means the biggest impact on the model). This is why for example the values in Figure 4.9b are significantly lower, but can still be compared with the plot in Figure 4.9a. These two plots also show something interesting, namely that with the RFDT model the distance is deemed the most impactful while in the LR model it does not appear in the four-

teen highest impact variables at all. Overall comparing the two models per mode shows that most features are approximately the same, however, the differences that do exist could potentially prove beneficial for mode choice studies. This is due to the F_1 and BA scores being higher for the ML models and those thus have more accurate results.



(a) *SHAP feature importance bar plot for the LR model.*

(b) *SHAP feature importance bar plot for the best performing ML model (GB).*

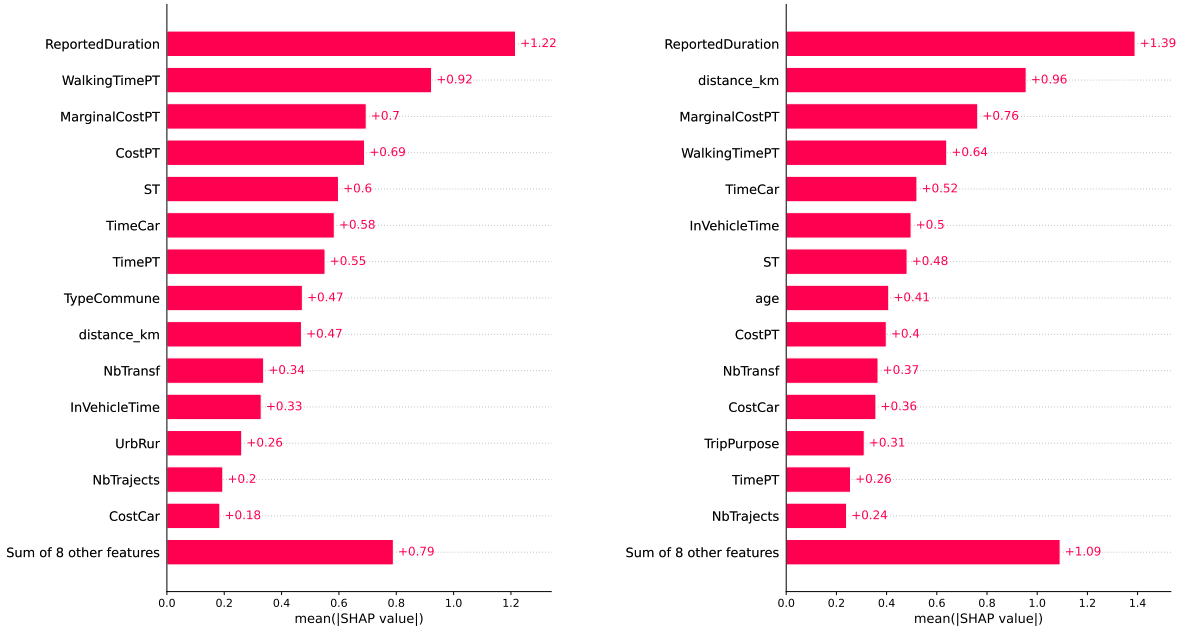
Figure 4.7: *SHAP analysis bar plots indicating the global importance of each feature based on the mean absolute value for that feature over all given samples for public transportation.*

4.3.1. SHAP results

For this specific case study the SHAP plots from figure 4.7, 4.8 and 4.9 can be used to determine which features are most impactful on the total system. Changes in the values for the highest ranked variables have the highest impact and enforcing changes in those can result in the respective mode of transport to be more or less attractive as an alternative to individuals. For example, with both private and public transportation it can be seen that the reported duration is ranked as the most important variable by quite some margin. This means that decreasing the travel duration for either mode will result in individuals to be more inclined to use that mode. Important to note is however that the shown SHAP plots do not actually provide the information on whether an increase or decrease in value will have a positive effect, in this case it is however reasonable to assume that a lower travel duration is seen as a positive change. There are different SHAP plots available that are capable of giving more in depth information like this, however for this research it is outside of the research scope and thus too time consuming to create.

4.4. Results analysis

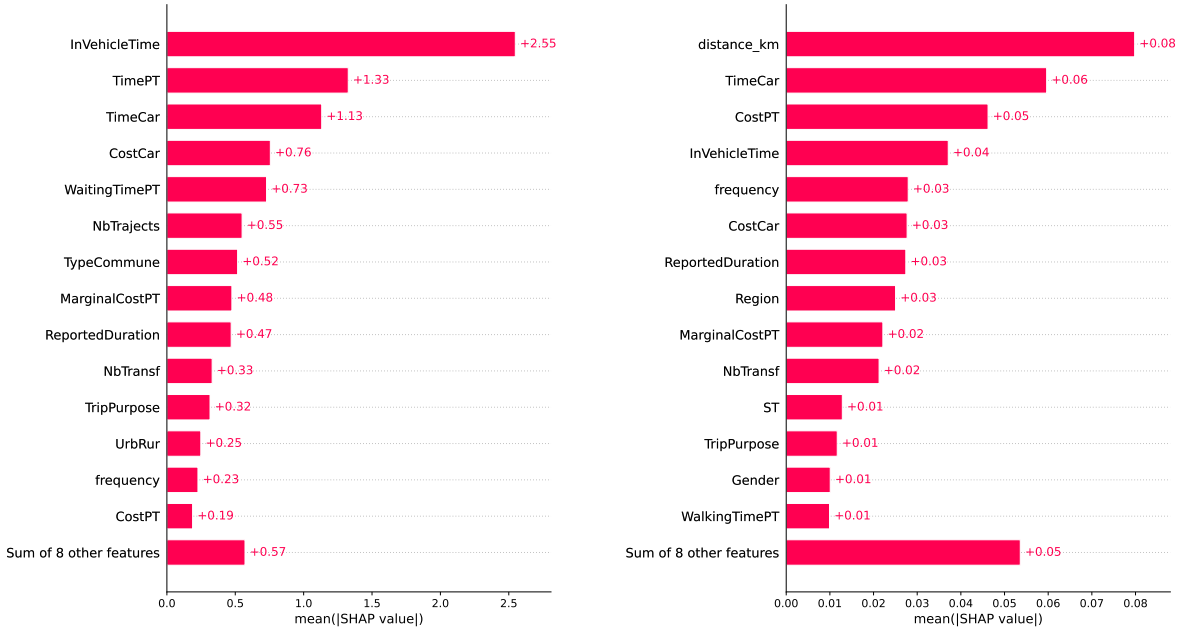
In this section the results are discussed and important findings are highlighted. Initially the comparison between the results of the LR model and the ML models are discussed in Subsection 4.4.1. After that in Subsection 4.4.2 the different ML model iterations are compared and model improvements and declines are discussed. Finally in Subsection 4.4.3 a more high level discussion is held.



(a) *SHAP feature importance bar plot for the LR model.*

(b) *SHAP feature importance bar plot for the best performing ML model (GB).*

Figure 4.8: *SHAP analysis bar plots indicating the global importance of each feature based on the mean absolute value for that feature over all given samples for private transportation.*



(a) *SHAP feature importance bar plot for the LR model.*

(b) *SHAP feature importance bar plot for the best performing ML model (RFDT).*

Figure 4.9: *SHAP analysis bar plots indicating the global importance of each feature based on the mean absolute value for that feature over all given samples for soft transportation.*

4.4.1. LR vs ML models

From the performance metrics it is clear that the LR model which represents the RUM based logit models is not capable of performing at the level that any of the ML model iterations is capable of. In Figure 4.10 it can be seen that the F_1 scores for all modes is significantly lower. When it comes to the BA score however, the soft transport mode actually scores really well. But because the F_1 score is extremely low this can be explained with a high TP rate at the cost of a high FP rate. This can also be confirmed when looking at the confusion matrix for soft transport in Table 4.7. The subsequent increase in the F_1 and decrease in BA score for the base ML model shows that the TP rate drops somewhat with the benefit of the FP rate dropping massively.

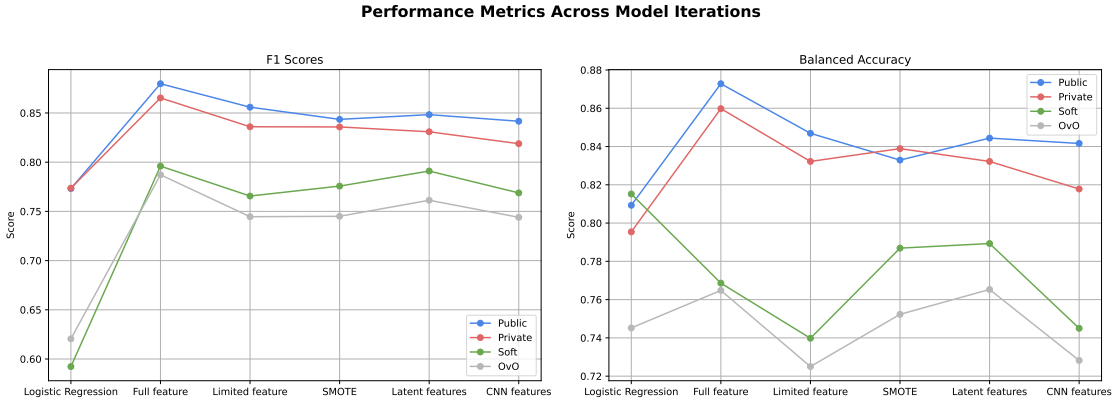


Figure 4.10: *F_1 and BA performance scores for each different model iteration on all modes with the GB model.*

4.4.2. ML model iterations

The initial full feature experimental plan shows a strong performance with mainly the F_1 scores for all modes ranking very high. However this model uses 110 different features, attributes and variables as inputs, this amount of data is usually not available when conducting a mode choice study. As such the limited feature run is used and as a result the input features drops to 22. With only a fifth of the features remaining it is harder for the model to make accurate predictions as a lot more information is no longer available. However, the limited feature experimental plan results show that the drop in performance is not severe. In fact for some models, such as the RFDT, the change can even cause an improvement as shown for the soft transportation class in Figure 4.11.

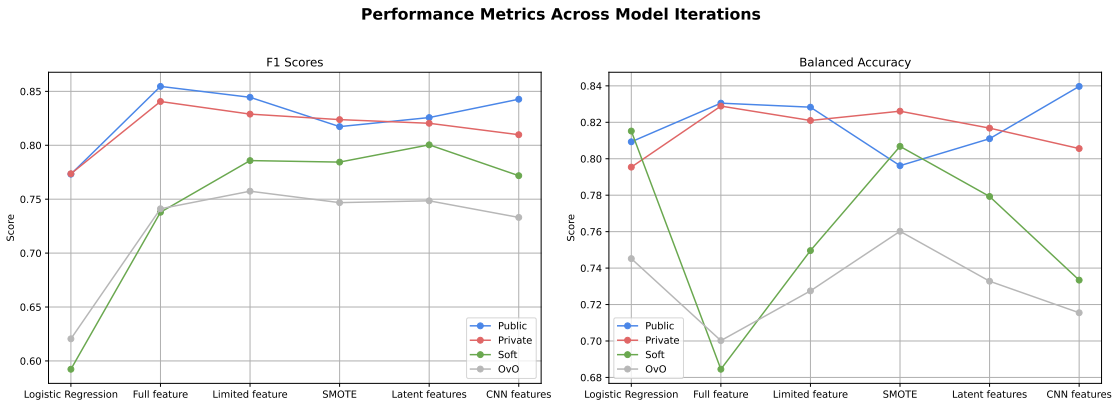


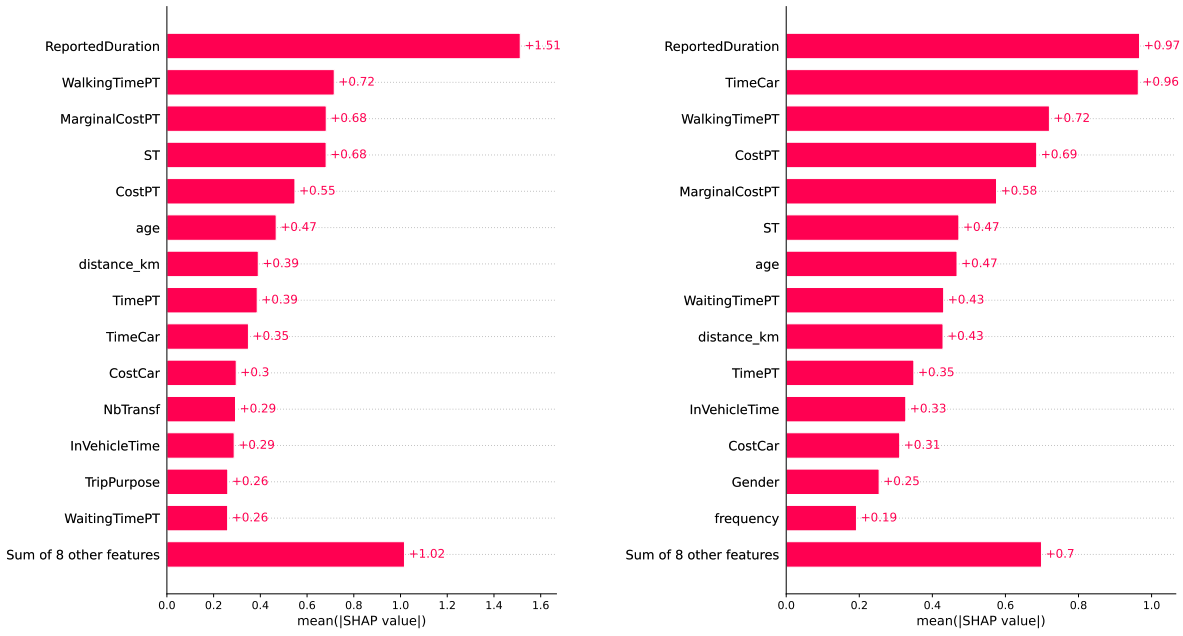
Figure 4.11: *F_1 and BA performance scores for each different model iteration on all modes with the RFDT model.*

When applying SMOTE to the training set an additional improvement can be seen for the soft trans-

port mode, especially for the BA score which, with the RFDT model even rises above the BA score for public transport. This however is also in part due to the SMOTE model having the opposite effect on this transport mode in both ML model types.
The latent variable model has mixed results, for the soft transport class the GB model has both the F_1 and BA score improve while with the RFDT this is only true for the F_1 score. Even though these results seem to overall improve the best soft transport results are created in the SMOTE only run as shown in Table 4.15.
Finally the CNN model proves to be a downgrade compared to most iterations, for the RFDT model the public transport scoring metrics both improve but the scores for the GB model on the base limited feature run still slightly outperform. It can thus be concluded that the CNN model is not capable of extracting meaningful latent variables with the help of indicator variables that the regular latent variable model was able to do. Considering the good performance of the CNN model itself as shown in Figure 4.6, the likely cause for this is that the created latent variables are actually not representative for the mode choices.

4.4.3. Overall results

Overall the performance changes are not of a significant magnitude, mainly for the public and private transport mode that seem to hover around the same score over all iterations. The only mode that actually has more significant changes is the soft transport mode, this is also a positive development as minority classes are generally of higher interest for mode choice studies as it is a class that needs to be improved to increase its usage. Even though most performance changes are small, the impact on the SHAP analysis can still be seen. In Figure 4.12 two GB models are compared on the public transport class. The first model (Figure 4.12a) does not have SMOTE applied and the second plot (Figure 4.12b) does. In Figure 4.10 it can be seen that the first model is only slightly superior to the second yet the SHAP analysis shows some differences.



(a) *SHAP feature importance bar plot for the GB model without SMOTE.*

(b) *SHAP feature importance bar plot for the GB model with SMOTE.*

Figure 4.12: *SHAP analysis bar plots between two GB models, one with and one without SMOTE applied.*

Furthermore it can be concluded that it is not ideal to use a singular ML model on all modes as the GB and RFDT models both are not superior on all of them. It is interesting to note that the

RFDT model seems to do specifically well on the minority class (soft transport). This result is in line with other studies where RFDT or similar variant are commonly used for unbalanced datasets (Leevy et al., 2018). This and the fact that the One-vs-One models constantly perform the worst on both scoring metrics (Figure 4.10 and Figure 4.11) also highlights how a one hot encoded binary classification is superior for the purpose of mode choice studies. It gives the model the chance to focus on a specific class and the SHAP plots can show the features that are important to that mode specifically because as shown in Section 4.3 those do not necessarily overlap.

5

Conclusion

The goal of this thesis was to investigate the potential of Machine Learning models to accurately capture Mode Choice behavior within transportation systems. The demand for more flexible, data-driven transportation modeling is growing as a result of modern data collection methods and the global increasing demand to limit private vehicle usage in favor for more sustainable alternatives. This research has explored the ability of various Machine Learning methodologies to address limitations observed in traditional Random Utility Maximization (RUM) models.
In doing so, this research addressed the following main research question:

> **What is the potential of machine learning models to accurately evaluate Mode Choice behavior of transportation systems?**

To answer this, the research was guided by four sub-questions:

1. **What Machine Learning methods can be used for Mode Choice problems?**
   The literature review identified a wide range of Machine Learning models applicable to Mode Choice problems, with supervised learning models like Random Forest Decision Tree (RFDT), Gradient Boosting (GB), Support Vector Classifier (SVC), and Multilayer Perceptron (MLP) stand out for their predictive accuracy and flexibility. However, no single dominant framework was evident, highlighting the need for structured model evaluation.

2. **What is the potential of synthetic data generation and over/under-sampling for imbalanced datasets?**
   The case studies showed that applying oversampling techniques such as Synthetic Minority Over-sampling TEchnique (SMOTE) can help alleviate performance issues on underrepresented classes (e.g., soft transport modes), though effects were modest. When the dataset imbalance is extreme SMOTE or comparable alternatives failed to achieve meaningful improvements. This confirms it's role as supportive tools that can give slightly more insights rather than a complete solution to dealing with minority classes.

3. **How can integrated theory-based knowledge improve Machine Learning models?**
   Incorporating domain-specific knowledge through the use of latent variables improved some Machine Learning models by adding contextual interpretability. However, gains in performance were not always consistent. Furthermore it is difficult to tell if theory-augmented Machine Learning models may be more beneficial for interpretability rather than raw accuracy.

4. **How can important information regarding mode alternatives be extracted from a trained Machine Learning model?**
   Through a feature importance analysis (using SHapely Additive exPlanation (SHAP)), it was demonstrated that Machine Learning models can reveal detailed and often nonlinear relationships between features and predicted choices. This offers a clear advantage over traditional models when explaining behavior, especially in disaggregated contexts, in addition it is also shown that results differ between a Logistic Regression and better performing Machine Learning model highlighting the relevance of using the better performing model to get a more accurate picture of which features are important and which are not.

Overall, this research found that ML models, particularly Gradient Boosting and Random Forest Decision Tree, can indeed outperform Logistic Regression models in predicting individual mode choices, especially when using One-vs-Rest classification with one hot encoded binary class targets. Yet, challenges persist in interpretability, framework consistency, and imbalanced datasets. While interpreting the final results with the help of feature importance analysis tools is no longer impossible or even difficult, they do not help with pinpointing ways to improve the model. When additional applied methods such as SMOTE and latent variables give mixed results, there is little to fall back on in order to figure out where the issue might be.

Importantly, these results echo the concerns highlighted in Chapter 1; Machine Learning offers a promising yet not fully mature alternative to theory-driven models. While it bypasses many of the assumptions RUM models require and have an easy initial set up, it introduces new challenges when it comes to data dependency, tuning complexity, and reduced transparency.

It can be concluded that Machine Learning has great potential, even as a plain model it already is capable of capturing the complex relations between variables which results in a strong predictive power. With the help of interpretability tools the models are capable of providing interested parties with targeted information on what features are critical.

5.1. Policy recommendations

The goal of this research was to uncover the potential of Machine Learning methodologies to be used for Mode Choice problems. This includes the ability to extract useful policy recommendations from the gathered results. While the used case studies were chosen based on data availability and not the actual need for a specific system change/improvement, for the sake of this research it is still worthwhile to translate the results into them. Based on the results from the Mode Choice study the most important variables for each transport mode can be identified and dependent on what changes are desired these can be improved. One of the most common goals for mode choice studies is to increase the use of more sustainable mode alternatives over the use of private cars. Based on figures 4.7, 4.8 and 4.9, a possible way to do such thing is to decrease the time it takes to travel by public transport while increasing the time it takes for private vehicles. This can be achieved by introducing dedicated transit roads or lanes and force private vehicles to take a longer route. This change can also help decrease the eventual costs of using public transportation as travel time and distance are down resulting in a higher capacity with the same equipment, furthermore can this also result in the driving distance when using private transport increasing due to more direct routes being converted to only allow public transport. These changes all positively impact the desired change as well and thus have the potential to snowball the increased usage of public over private transportation.

5.2. Future Research

While this research makes clear that Machine Learning is a viable alternative to use for Mode Choice modeling, the further attempted improvements to the model failed to provide overwhelming results. However, this should not mean that the explored avenues are a dead end and therefore several recommendations can be made for future research to potentially achieve more.

5.2.1. Datasets

One limitation that this research had was the data availability, while the Switzerland dataset was extremely detailed it still lacked the larger sample size that Machine Learning models generally need for training. Similarly the Rhine-Alpine dataset was not of a high enough quality to be used in a Machine Learning model. This is unfortunate considering there was shipping data for 600 Origin-Destination pairs on two commodity types collected over eleven years, this should be good for a

sample size of 13200. However the lack of available input features over this period left the dataset with only 1200 instead. This combined with a high class imbalance caused it to be unusable, however this does highlight the gap that an other dataset could potentially fill. It would be interesting to see how Machine Learning performs on a better structured aggregated dataset.
Generally research done on the topic of Mode Choice focuses on a specific region and a dataset related to it. However the usage of Machine Learning opens the door to applying transfer learning, this is a method already used to problems such as image recognition but can also potentially be used here. While Mode Choice studies generally need to be on a specific area this technique can help create a stronger final model as basic principles regarding Mode Choice are similar or the same regardless of the scope of the area. This could in turn also help with minority classes as in a small area there might be limited data available but when combined with a model that is already trained on similar data it is possible to get stronger results in return.

5.2.2. Model structure

Due to the limited knowledge of the author this research was limited to using existing libraries for the implementation of Machine Learning models. While this does mean that the used models are efficient and initially very easy to use, it has as a downside that they are not flexible and modular and thus incorporating theory driven knowledge was limited to the outside of the model and served more as a pre-processing step. Future research could focus on including theory driven structure into the core of Machine Learning models to potentially more efficiently get the best of both worlds similar to the work of Zhang et al. (2020). This also highlights the two different directions that can be taken in order to improve results: Better models or better data processing. Future research could be done to establish a well working framework for both direction on steps that can be taken to improve results.

Bibliography

[] Atasoy, B., Glerum, A., & Bierlaire, M. (2013). Attitudes towards mode choice in Switzerland. *disP - The Planning Review*, *49*(2), 101–117. https://doi.org/10.1080/02513625.2013.827518

[] Bajaj, A. (2025, May). Performance Metrics in Machine Learning [Complete Guide]. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

[] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

[] Dent, E. B., & Goldberg, S. G. (1999). Challenging "Resistance to change". *Journal of applied behavioral science/The Journal of applied behavioral science*, *35*(1), 25–41. https://doi.org/10.1177/0021886399351003

[] Ding, L., & Ning, Z. (2016). Estimating modal shift by introducing transit priority strategies under congested traffic using the multinomial logit model. *KSCE Journal of Civil Engineering*, *21*(6), 2384–2392. https://doi.org/10.1007/s12205-016-0640-0

[] DOT, N. (2025a, January). Vision Zero: New report shows street redesigns have brought largest declines in traffic deaths in predominantly Lower-Income neighborhoods and neighborhoods of color. https://www.nyc.gov/html//dot//html/pr2025/vision-zero-report-street-redesign.shtml?utm_source=chatgpt.com

[] DOT, N. (2025b, March). NYC DOT celebrates safer street designs, wider bike lanes across Manhattan avenues. https://www.nyc.gov/html/dot/html/pr2025/safer-across-manhattan-aves.shtml

[] Eluru, N., Chakour, V., & El-Geneidy, A. (2012). Travel mode choice and transit route choice behavior in Montreal: insights from McGill University members commute patterns. *Public transport*, *4*(2), 129–149. https://doi.org/10.1007/s12469-012-0056-2

[] Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2008.05756

[] Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation research. Part B: methodological/Transportation research. Part B, Methodological*, *37*(8), 681–698. https://doi.org/10.1016/s0191-2615(02)00046-2

[] Hart, S. (1989, January). *Shapley value*. https://doi.org/10.1007/978-1-349-20181-5\{_}25

[] Hess, S., Ben-Akiva, M. E., Gopinath, D., & Walker, J. L. (2009). Taste heterogeneity, correlation, and elasticities in latent class choice models. *Transportation Research Board 88th Annual Meeting*Transportation Research Board. https://trid.trb.org/view/881736

[] Hess, S., Ryley, T., Davison, L., & Adler, T. (2013). Improving the quality of demand forecasts through cross nested logit: a stated choice case study of airport, airline and access mode choice. *Transportmetrica. A, Transport science*, *9*(4), 358–384. https://doi.org/10.1080/18128602.2011.577758

[] Hillel, T., Bierlaire, M., Elshafie, M. Z., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of choice modelling*, *38*, 100221. https://doi.org/10.1016/j.jocm.2020.100221

[] Kalimi, A. (2023). List of machine learning models - Aamir Kalimi - Medium. https://medium.com/@codekalimi/list-of-machine-learning-models-61b51ad492f1

[] Kashifi, M. T., Jamal, A., Kashefi, M. S., Almoshaogeh, M., & Rahman, S. M. (2022). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel behaviour and society/Travel behaviour society*, *29*, 279–296. https://doi.org/10.1016/j.tbs.2022.07.003

[] Kim, E.-J. (2021). Analysis of travel mode choice in Seoul using an interpretable machine learning approach. *Journal of advanced transportation*, *2021*, 1–13. https://doi.org/10.1155/2021/6685004

[] Lahoz, L. T., Pereira, F. C., Sfeir, G., Arkoudi, I., Monteiro, M. M., & Azevedo, C. L. (2023). Attitudes and Latent Class Choice Models using Machine Learning. *Journal of choice modelling*, *49*, 100452. https://doi.org/10.1016/j.jocm.2023.100452

[] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal Of Big Data*, *5*(1). https://doi.org/10.1186/s40537-018-0151-6

[] Li, X., Shi, L., Shi, Y., Tang, J., Zhao, P., Wang, Y., & Chen, J. (2024). Exploring interactive and nonlinear effects of key factors on intercity travel mode choice using XGBoost. *Applied geography*, *166*, 103264. https://doi.org/10.1016/j.apgeog.2024.103264

[] Masoumi, H. E., Van Rooijen, M., & Sierpiński, G. (2020). Children's independent mobility to school in seven European countries: a multinomial Logit model. *International journal of environmental research and public health/International journal of environmental research and public health*, *17*(23), 9149. https://doi.org/10.3390/ijerph17239149

[] Matyas, M., & Kamargianni, M. (2021). Investigating heterogeneity in preferences for Mobility-as-a-Service plans through a latent class choice model. *Travel behaviour and society/Travel behaviour society*, *23*, 143–156. https://doi.org/10.1016/j.tbs.2020.12.002

[] McFadden, D. (2001). Economic choices. *The American Economic Review*, *91*(3), 351–378. https://doi.org/10.1257/aer.91.3.351

[] Nicolet, A., Negenborn, R. R., & Atasoy, B. (2022). A Logit mixture model estimating the heterogeneous mode choice preferences of shippers based on aggregate data. *IEEE open journal of intelligent transportation systems*, *3*, 650–661. https://doi.org/10.1109/ojits.2022.3208379

[] *NYC DOT - NYC Streets Plan* (tech. rep.). (2021, December). New York City Department of Transportation. https://www.nyc.gov/html/dot/html/about/nyc-streets-plan.shtml

[] Pochan, J., & Wichitphongsa, W. (2020). An Intercity Freight Mode Choice Model : A case study of High speed Rail Link Northern Line Thailand (Bangkok – Chiangmai). *MATEC web of conferences*, *308*, 04003. https://doi.org/10.1051/matecconf/202030804003

[] Polydoropoulou, A., Tsirimpa, A., Karakikes, I., Tsouros, I., & Pagoni, I. (2022). Mode choice modeling for sustainable Last-Mile Delivery: the Greek perspective. *Sustainability*, *14*(15), 8976. https://doi.org/10.3390/su14158976

[] Prudhvith, T. (2022). Why always using SMOTE is not recommended? - Tavva Prudhvith - Medium. https : / / medium . com / @prudhvithtavva / why - always - using - smote - is - not - recommended-ce21f981d31e

[] Qi, C., Zhu, Z., Guo, X., Lu, R., & Chen, J. (2020). Examining Interrelationships between Tourist Travel Mode and Trip Chain Choices Using the Nested Logit Model. *Sustainability*, *12*(18), 7535. https://doi.org/10.3390/su12187535

[] Ran, Z.-Y., & Hu, B.-G. (2017). Parameter Identifiability in Statistical Machine Learning: A review. *Neural computation*, *29*(5), 1151–1203. https://doi.org/10.1162/neco\{_}a\{_}00947

[] Rong, H. H., & Freeman, L. (2024). The impact of the built environment on human mobility patterns during Covid-19: A study of New York City's Open Streets Program. *Applied Geography*, *172*, 103429. https://doi.org/10.1016/j.apgeog.2024.103429

[] Salas, P., De La Fuente, R., Astroza, S., & Carrasco, J. A. (2022). A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. *Expert systems with applications*, *193*, 116253. https://doi.org/10.1016/j.eswa.2021.116253

[] scikit-learn developers. (n.d.). scikit-learn: machine learning in Python — scikit-learn 1.7.0 documentation. https://scikit-learn.org/stable/index.html

[] Sekhar, C. R., Minal, & Errampalli, M. (2016). Mode choice analysis using random forrest decision trees. *Transportation research procedia*, *17*, 644–652. https://doi.org/10.1016/j.trpro.2016.11.119

[] Shahikhaneh, A., Azari, K. A., & Aghayan, I. (2019). Modeling the transport mode choice behavior of motorcyclists. *Iranian journal of science and technology. Transactions of civil engineering/Civil engineering*, *44*(1), 175–184. https://doi.org/10.1007/s40996-019-00236-4

[] Ton, D., Bekhor, S., Cats, O., Duives, D. C., Hoogendoorn-Lanser, S., & Hoogendoorn, S. P. (2020). The experienced mode choice set and its determinants: Commuting trips in the Netherlands. *Transportation research. Part A, Policy and practice*, *132*, 744–758. https://doi.org/10.1016/j.tra.2019.12.027

[] Van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning - Discussion paper. *Journal of choice modelling*, *42*, 100340. https://doi.org/10.1016/j.jocm.2021.100340

[] Van Rijsbergen, C. J. (1979, January). *Information retrieval*. London ; Toronto : Butterworths.

[] Wang, F., & Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation research record*, *2672*(47), 35–45. https://doi.org/10.1177/0361198118773556

[] Wang, K., Bhat, C. R., & Ye, X. (2022). A multinomial probit analysis of shanghai commute mode choice. *Transportation*, *50*(4), 1471–1495. https://doi.org/10.1007/s11116-022-10284-x

[] Wen, C.-H., Wang, W.-C., & Fu, C. (2012). Latent class nested logit model for analyzing high-speed rail access mode choice. *Transportation research. Part E, Logistics and transportation review*, *48*(2), 545–554. https://doi.org/10.1016/j.tre.2011.09.002

[] Wu, L., Wang, W., Jing, P., Chen, Y., Zhan, F., Shi, Y., & Li, T. (2019). Travel mode choice and their impacts on environment—a literature review based on bibliometric and content analysis, 2000–2018. *Journal of Cleaner Production*, *249*, 119391. https://doi.org/10.1016/j.jclepro.2019.119391

[] Yang, C.-W., & Sung, Y.-C. (2010). Constructing a mixed-logit model with market positioning to analyze the effects of new mode introduction. *Journal of transport geography*, *18*(1), 175–182. https://doi.org/10.1016/j.jtrangeo.2009.01.005

[] Ye, M., Chen, Y., Yang, G., Wang, B., & Hu, Q. (2020). Mixed Logit models for travelers' mode shifting considering Bike-Sharing. *Sustainability*, *12*(5), 2081. https://doi.org/10.3390/su12052081

[] Zhang, Z., Ji, C., Wang, Y., & Yang, Y. (2020). A Customized Deep Neural Network Approach to Investigate Travel Mode Choice with Interpretable Utility Information. *Journal of advanced transportation*, *2020*, 1–11. https://doi.org/10.1155/2020/5364252

A

Research Paper

Using Machine Learning Techniques for Mode Choice Analysis: A Case Study on Swiss Travel Behavior

Thijs Kesselring, Edwin van Hassel, Bilge Atasoy
Department of Maritime and Transport Technology, Delft University of Technology, Delft, The Netherlands

***Abstract*—TODO**

I. Introduction

Understanding how individuals choose their mode of transport—such as driving, public transport, cycling, or walking—is vital for developing efficient and sustainable urban mobility. By analyzing how these individuals choose what mode of transport to use for specific trips it allows policy makers and urban planners to create an environment that encourages the usage of more favorable alternatives. This in turn can help prevent high amounts of traffic within cities which is a source of both noise and environmental pollution. An individuals decision making process depends on various factors, including their socioeconomic background, the trip characteristics, their own perceived environmental concerns, and overall policy constraints.

Traditionally, researchers have used Random Utility Maximization (RUM) models (Wu et al., 2019) to understand mode choice behavior. However, these models are theory driven and often require strong assumptions about distribution and linearity and may struggle with large, complex datasets as a result. The rise of Machine Learning (ML) in transportation modeling presents an opportunity to explore new, data-driven approaches that can overcome these challenges and make the overall process of modeling mode choice behavior easier and more accurate.

This study addresses the following research question: *What is the potential of machine learning models to accurately capture mode choice behavior in transportation systems?*
In order to answer this question several sub questions are created as well:

1) *What Machine Learning methods can be used for Mode Choice problems?*
2) What is the potential of synthetic data generation and over/under-sampling for imbalanced datasets?
3) How can integrated theory-based knowledge improve Machine Learning models?

To answer these questions, existing literature is reviewed on both the classic approach as on studies that already attempt to implement ML. Additionally several ML models are implemented and applied to an existing dataset regarding Swiss travel behavior.

II. Literature Review

A. Mode Choice Modeling Background

Mode choice modeling is a popular field in research. For a normal mode choice study the goal is to predict the mode of transportation an individual will choose given a set of alternatives and contextual factors, based of of that model it can then be determined what features and attributes have lead to that choice. However in the scientific community there is a strong drive to improve the modeling techniques. Wu et al. (2019) performed a literature review study in which they found that the majority of studies focuses on model improvements often aimed towards RUM-based models such as the Multinomial Logit Model (MNL), Nested Logit (NL), and Mixed Logit (MXL). These models have been widely used due to their solid theoretical foundations and interpretable outputs and most improvements focus on improving the capability of dealing with heterogeneity.

B. Random Utility Models

RUM models assume that individuals make choices to maximize their perceived utility. The utility of each alternative is typically modeled as a function of observed attributes and a random component to account for unobserved variation. While MNL is straightforward to implement, it suffers from the Independence of Irrelevant Alternatives (IIA) property. This however does not make it useless and for basic problems it is still a viable option. Masoumi et al. (2020) used MNL to what features and attributes are important in the mode choice behavior of a child's (between 9 and 12 years old) commute to and from school. Ding and Ning (2016) applied MNL to directly investigate the effects of specific policy changes. For this they created the model and later manually changed certain input variables to reflect changes in parking pricing and public transportation cost. The predicted results could then be compared to analyze how the population would likely respond. Too indicate the relevance MNL still has a research from Ton et al. (2020) compared the results of the model against that of a NL and MXL model. To their surprise MNL actually proved to be the model that was able to fit the data best. However this does not indicate that the latter two model types are not capable. NL and MXL models attempt

to address the limitation that MNL experiences with IIA by allowing for correlated alternatives or random taste variation, respectively. Shahikhaneh et al. (2019) used this property NL has by grouping certain public transportation alternatives together to help answer their main research question: "How to get motorcyclists to use public transportation alternatives?". The nested structure also allows researchers to get more creative with problems, for example Qi et al. (2020) use this structure to create two models, one where the same two modes are grouped together based on the trip chain that is used with them and another where the same five trip chain types are grouped together based on the transportation mode used. Based on what model performed best they were able to conclude that tourist tended to choose their mode of travel based on the type of trip chain they had planned (rather than the other way around). Finally MXL models provide even more flexibility, Polydoropoulou et al. (2022) to deal with a dataset that captures multiple samples from each individual and MXL allowed the researchers to capture the correlation between these samples. Ye et al. (2020) and Yang and Sung (2010) both used the MXL model specifically for the reason of capturing heterogeneity in the observed population indicating a strong need for the ability to model that.

C. Machine Learning Applications

Recent research has applied ML models such as Random Forests (RF), Gradient Boosting Classifiers (GBC), Multilayer Perceptrons (MLP), and Support Vector Classifiers (SVC) to mode choice problems. Studies have shown that these models can outperform traditional methods in terms of predictive accuracy, particularly in handling nonlinear interactions and high-dimensional data. For example Kashifi et al. (2022) compares several of these models with each other and manage to achieve impressive accuracy scores with the lightGBDT model. During their research they however did encounter a common problem for ML: Imbalanced datasets. To alleviate this problem Kashifi et al. (2022) experimented with under and oversampling strategies, with oversampling they were able to get good results even on the minority classes. Furthermore with the help of SHapely Additive exPlanation (SHAP) they were also capable of 'opening' the trained model and look in to the so called "black-box". In doing so the interpretability of the model raised and they were able to show how different features impacted the model. Wang and Ross (2018) try to compare a eXtreme Gradient Boosting (XGB) model with a basic MNL model. While the XGB model showed better performance one of the key takeaways from the research was that the MNL model required more attention in the set up and training in regards to the assumptions made for the random distribution. Zhang et al. (2020) attempt to make a hybrid model with the use of a Deep Neural Network (DNN) shaped like a RUM based model. When the model was tested on a large dataset it showed a higher accuracy than competing models which included logit models but also a RF model. While the results were still not through the roof the shape of the DNN model allowed the researchers to easily interpret the results and model itself without the need for additional tools and algorithms.

D. Literature review results

The conducted literature review highlighted a couple of areas that ML models struggle with when it comes to MC studies:

- **Black-box:** A well known downside of ML models is knowing how everything the model does between input and output works. While for some research fields this is not a big problem, with MC studies this connection between input and output is often the main goal.
- **Different model types:** Where logit models already have different variations, this is even more accurate for ML models. For example the research of Kalimi (2023) lists 50 different model types. This makes it extremely difficult to figure out what model best fits the available data.
- **Unbalanced datasets:** With MC studies it is very common that a specific mode is underrepresented within the data. Unfortunately the data driven nature of ML models can result in this class having poor performance after training because the model was not capable to learn the positive labels.

For these challenges this research needs to find some methods to resolve the issues. Fortunately the literature has also provided useful tools such as SHAP and Synthetic Minority Over-sampling TEchnique (SMOTE) to potentially use.

III. Methodology

A. Models

This study evaluates the predictive performance for five models of which one represents a logit model as a way to compare the other four with. These models are:

- Logistic Regression (LR)
- Gradient Boosting (GB)
- Random Forest (RF)
- Multilayer Perceptron (MLP)
- Support Vector (SV)

The LR model is used to represent a classic logit based model, this subsequently allows it to be used as a baseline model to compare with in this study. The other models are chosen based on the literature review as having the most potential.

B. Data Preprocessing

One of the most important steps with training ML models is how the training (and testing) data is processed. There are two main approaches to presenting the target data to the model: One vs One (OvO) and One vs Rest (OvR). The first approach is a multiclass method where the model will attempt to train each mode against each other and finally create a model that can predict which mode is chosen based on the input data. The other approach uses binary one hot encoded target data to train and test on. As a result a separate model is trained for each mode and the predicted output is either a 0 indicating this mode was not used, or a 1 indicating that the mode was used.

This research will include both approaches to determine which approach has the best results. To deal with the imbalanced data SMOTE will be applied to the dataset in order to generate synthetic data for the minority class. During model evaluation results are compared both with and without SMOTE training data.

C. Feature Sets

For the purpose of comparing potential improvements the model will be trained using different feature sets:

- **Full features:** All observed variables.
- **Limited features:** A subset of features selected based on domain relevance and perceived general availability.
- **Latent features:** Extracted using existing variables.
- **CNN features:** Latent features extracted by training a Convolutional Neural Network (CNN).

The full feature set contains almost all features that are available with the dataset and is thus expected to provide the best results. However, in order to simulate a realistic mode choice study a second feature subset (Limited features) is created that only contains features that are related to the alternatives, trips and only a very limited number of socio-economic features. The latent features set builds upon this subset by creating two latent variables. These variables are created using the results of Atasoy et al. (2013).

D. CNN model

Finally the CNN features set aims to create three latent variables by training a separate CNN model in a similar way as how the latent variables are created. This CNN model has two input branches of which one is strictly used during training and contains indicator variables. These variables consist of multiple opinion questions where respondents could quantify to what degree they agreed or disagreed with a statement.
The CNN structure itself is displayed in Figure 1. It contains two input branches, the right branch is only used during training of the CNN model and uses the indicator variables. In the left branch the other input variables from the experimental setup are provided. After some CNN layers the two branches are merged and trained on the mode choice observations. The goal is that the left branch of the trained model has been influenced by the indicator variables and thus the 'latent variables' are extracted before the merge layer so that the model does not require new indicator variables for unseen data in order to create new latent variables. When implementing the CNN model it is trained beforehand and the latent variables are extracted and added as training data for the mode choice ML models. During testing the trained model is used to create the latent variables as explained above and those latent variables are again used to test the mode choice ML models.

E. Model Training and Evaluation

The different models defined earlier are trained on the different feature sets according to the experimental setups as displayed in Table I. Each experimental setup is trained on all model types, to evaluate the trained models they are scored based on some metrics. The first metric is a normalized confusion matrix, this allows for an easy overview of how well the model does on accuracy. Because of the normalization the differences in representation within the dataset (unbalanced classes) can be ignored in reviewing the results as both the majority as the minority class have the same visual importance. The different trained models are also graded using the F_1 and the Balanced Accuracy (BA) score, shown in equation 1 and 2 respectively. Both scores are useful to use when dealing with unbalanced datasets as the resulting score will actually be impacted negatively if the minority class under performs.

$$F_1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{BA} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{2}$$

To ensure that the results are not a(n) (un)lucky occurance with how the dataset was split into a training and testing set, each model is trained 10 times with a different split (but equal for each model) and the results are averaged. This means that the results for each model can be compared with the other models as all of them used the same train-test split.

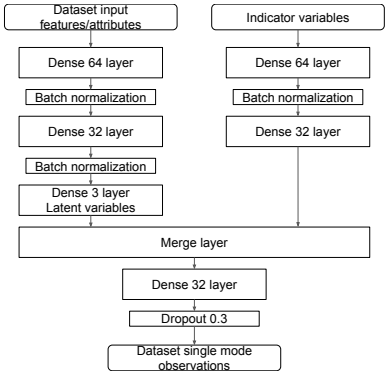


Fig. 1. The full structure of the CNN model used to create additional input variables for the main ML models.

TABLE I
OVERVIEW OF THE EXPERIMENTAL SETUPS AND THE FEATURE SETS USED FOR THEM. (*: PARTIAL USE, **: USED TO CREATE THE LATENT FEATURES, ***: INDIRECTLY USED TO DETERMINE VARIABLES FOR THE LATENT FEATURES)

	Full Feature Plan	Limited Feature Plan	Latent Variable Plan	CNN Plan
Attributes of the Alternatives	x	x	x	x
Socio-Economic or Cargo Features	x	x*	x*	x*
Trip-specific Features	x	x	x	x
Indicator Variables	x		***	x**
Latent Features			x	x

IV. CASE STUDY: SWITZERLAND

A. Dataset Description

The Swiss dataset includes detailed trip and individual-level data, such as age, income, household size, travel distance, and

selected mode of transport. Transport modes are categorized into Public (e.g., bus, train), Private (e.g., car), and Soft (e.g., walking, cycling). The data is inherently imbalanced, with the soft transport mode as a clear minority class as shown in Table II.

TABLE II
POSITIVE LABELS IN THE SWITZERLAND DATASET PER MODE.

Mode	Positive Labels	Percentage
Public	536	28.1%
Private	1256	65.9%
Soft	114	6.0%

B. Results

Initially the Full Feature and Limited Feature experimental setups were used to train the LR model. The resulting F_1 and BA scores are displayed in table III and IV respectively.

TABLE III
AVERAGE PERFORMANCE METRICS (F_1 AND BA) FOR THE LR MODEL USING THE FULL FEATURE EXPERIMENTAL SETUP.

	Public	Private	Soft	One vs One
F_1	0.8300	0.8121	0.6797	0.6999
BA	0.8496	0.8213	0.7899	0.7542

TABLE IV
AVERAGE PERFORMANCE METRICS (F_1 AND BA) FOR THE LR MODEL USING THE LIMITED FEATURE EXPERIMENTAL SETUP.

	Public	Private	Soft	One vs One
F_1	0.7732	0.7736	0.5923	0.6205
BA	0.8093	0.7954	0.8152	0.7452

From these results it can already be concluded that there is a significant drop in performance between the two setups. This however is to be expected considering the limited feature run contains a lot less information for the model to train on. Running the full feature setus with the four machine learning methodologies resulted in higher F_1 and BA scores compared to the LR model, this can be seen in table V and VI. From this it can also be observed that the Gradient Boosting method has the highest scores across all modes of transport, as a result the remainder of the research focuses mainly on this method over the others.

TABLE V
AVERAGE F_1 SCORES THE FOUR ML MODELS WITH THE FULL FEATURE EXPERIMENTAL PLAN.

Model	Public	Private	Soft	One vs One
GB	0.8797	0.8653	0.7960	0.7872
RF	0.8546	0.8406	0.7379	0.7411
MLP	0.8571	0.8275	0.7297	0.7118
SV	0.8370	0.8178	0.6923	0.6895

TABLE VI
AVERAGE BA SCORES THE FOUR ML MODELS WITH THE FULL FEATURE EXPERIMENTAL PLAN.

Model	Public	Private	Soft	One vs One
GB	0.8728	0.8598	0.7686	0.7648
RF	0.8305	0.8289	0.6845	0.7002
MLP	0.8522	0.8257	0.6986	0.6976
SV	0.8565	0.8258	0.7647	0.7429

When using the limited feature setup with the GB method a similar drop in performance can be observed as with the LR method. However as can be seen in Table VII the trained model still outperforms the LR model and even produces better or near equal results compared to the full feature setup used with the LR model.

TABLE VII
AVERAGE PERFORMANCE METRICS (F_1 AND BA) FOR THE GB MODEL USING THE LIMITED FEATURE EXPERIMENTAL PLAN.

	Public	Private	Soft	One vs One
F_1	0.8559	0.8360	0.7656	0.7446
BA	0.8469	0.8322	0.7398	0.725

Furthermore the other setups are also used with the implementation of the Latent variables and the CNN model. In table VIII and IX the F_1 and BA scores are shown for both the latent as the CNN variable setup respectively.

TABLE VIII
AVERAGE PERFORMANCE METRICS (F_1 AND BA) FOR THE GB MODEL USING THE LATENT FEATURE EXPERIMENTAL PLAN WITHOUT SMOTE IMPLEMENTATION.

	Public	Private	Soft	One vs One
F_1	0.8483	0.8309	0.7910	0.7613
BA	0.8444	0.8322	0.7893	0.7653

TABLE IX
AVERAGE PERFORMANCE METRICS (F_1 AND BA) FOR THE GB MODEL USING THE CNN FEATURE EXPERIMENTAL PLAN WITHOUT SMOTE IMPLEMENTATION.

	Public	Private	Soft	One vs One
F_1	0.8416	0.8188	0.7688	0.7440
BA	0.8416	0.8178	0.745	0.7282

It can be observed that the CNN variable setup is outperformed by the Latent variable setup for all transport modes. Additionally both setups are either worse or nearly the same compared to the limited feature setup results. This indicates that the additional variables do not help the Gradient Boosting model in the learning stage and in fact actually seem to confuse the model slightly. Another interesting development is that the results when using SMOTE as a pre-processing technique on the data actually also resulted in worse overall performance.

C. Best performers

Finally over all the iterations and models the best performing setup can be chosen. This is done based on the average score each model has over the F_1 and BA score. In Table X the results of this are shown. This confirms that the GB model is generally the best performing method. However when looking at the minority class the RF model has the best results, this is in line with other studies such as the research of Leevy et al. (2018) where RF model performs better on a minority class compared to other standard ML models. Additionally it can be seen that SMOTE does have a positive impact on the results and is in part responsible for the best results for both the minority classes.

TABLE X
BEST PERFORMING MODELS OVER ALL ITERATIONS (EXCLUDING THE FULL FEATURE EXPERIMENTAL PLAN) FOR EACH DIFFERENT MODE BASED ON THE AVERAGE SCORE BETWEEN F_1 AND BA.

Metric	**Public**	**Private**	**Soft**	**One vs One**
Model	GB Limited	GB with SMOTE	RF with SMOTE	GB with latent variables
F1	0.8559	0.8358	0.7844	0.7613
BA	0.8469	0.8389	0.8068	0.7653
Average	0.8514	0.8374	0.7956	0.7633

V. CONCLUSION AND FUTURE RESEARCH

This research aimed to answer the question as to "*What the potential of machine learning models to accurately capture mode choice behavior in transportation systems is?*". In order to answer this question several sub questions were posed and answered as followed:

1) **What Machine Learning methods can be used for Mode Choice problems?**:
   Supervised models such as RF GB, SVC, and MLP showed strong predictive performance as a result of a literature review, though no universal best model exists.
2) **What is the potential of synthetic data generation and over/under-sampling for imbalanced datasets?**:
   Oversampling techniques like SMOTE can slightly improve minority class results but are not a complete solution.
3) **How can integrated theory-based knowledge improve Machine Learning models?**:
   Latent variables can add interpretability but yield inconsistent performance gains.

Overall, Gradient Boosting and Random Forest outperformed Logistic Regression, especially in One-vs-Rest classification. However, interpretability, dataset imbalance, and framework consistency remain challenges. ML models provide strong predictive power and, with interpretability tools, can highlight critical features for decision-making. Yet, they also bring issues of data dependency and tuning complexity.

A. Future research

1) Datasets: The Switzerland dataset, while detailed, lacked sufficient size, and the Rhine-Alpine dataset suffered from limited features and imbalance. Better structured, larger datasets are crucial. Transfer learning could also be promising, allowing models trained in one region to improve performance in another and strengthening results for minority classes.

2) Model structure: This work used standard ML libraries, limiting integration of theory within models. Future research could focus on hybrid approaches that embed theory-driven knowledge directly into model architecture. Progress can be made either by improving datasets or developing more flexible models, ideally leading to a framework combining both directions.

ACKNOWLEDGEMENTS

This research was conducted as part of a Master's thesis in Mechanical Engineering at TU Delft under the supervision of Dr. B. Atasoy.

REFERENCES

Atasoy, B., Glerum, A., and Bierlaire, M. (2013). Attitudes towards mode choice in Switzerland. *disP - The Planning Review*, 49(2):101–117.

Ding, L. and Ning, Z. (2016). Estimating modal shift by introducing transit priority strategies under congested traffic using the multinomial logit model. *KSCE Journal of Civil Engineering*, 21(6):2384–2392.

Kalimi, A. (2023). List of machine learning models - Aamir Kalimi - Medium.

Kashifi, M. T., Jamal, A., Kashefi, M. S., Almoshaogeh, M., and Rahman, S. M. (2022). Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel behaviour and society/Travel behaviour society*, 29:279–296.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal Of Big Data*, 5(1).

Masoumi, H. E., Van Rooijen, M., and Sierpiński, G. (2020). Children's independent mobility to school in seven European countries: a multinomial Logit model. *International journal of environmental research and public health/International journal of environmental research and public health*, 17(23):9149.

Polydoropoulou, A., Tsirimpa, A., Karakikes, I., Tsouros, I., and Pagoni, I. (2022). Mode choice modeling for sustainable Last-Mile Delivery: the Greek perspective. *Sustainability*, 14(15):8976.

Qi, C., Zhu, Z., Guo, X., Lu, R., and Chen, J. (2020). Examining Interrelationships between Tourist Travel Mode and Trip Chain Choices Using the Nested Logit Model. *Sustainability*, 12(18):7535.

Shahikhaneh, A., Azari, K. A., and Aghayan, I. (2019). Modeling the transport mode choice behavior of motorcyclists. *Iranian journal of science and technology. Transactions of civil engineering/Civil engineering*, 44(1):175–184.

Ton, D., Bekhor, S., Cats, O., Duives, D. C., Hoogendoorn-Lanser, S., and Hoogendoorn, S. P. (2020). The experienced mode choice set and its determinants: Commuting trips in

the Netherlands. *Transportation research. Part A, Policy and practice*, 132:744–758.

Wang, F. and Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. *Transportation research record*, 2672(47):35–45.

Wu, L., Wang, W., Jing, P., Chen, Y., Zhan, F., Shi, Y., and Li, T. (2019). Travel mode choice and their impacts on environment—a literature review based on bibliometric and content analysis, 2000–2018. *Journal of Cleaner Production*, 249:119391.

Yang, C.-W. and Sung, Y.-C. (2010). Constructing a mixed-logit model with market positioning to analyze the effects of new mode introduction. *Journal of transport geography*, 18(1):175–182.

Ye, M., Chen, Y., Yang, G., Wang, B., and Hu, Q. (2020). Mixed Logit models for travelers' mode shifting considering Bike-Sharing. *Sustainability*, 12(5):2081.

Zhang, Z., Ji, C., Wang, Y., and Yang, Y. (2020). A Customized Deep Neural Network Approach to Investigate Travel Mode Choice with Interpretable Utility Information. *Journal of advanced transportation*, 2020:1–11.

B
Results

In this appendix all Switzerland dataset results from the different model iterations are shown for all four ML methodologies and for each mode.

B.1. Model F_1 and BA results

B.1.1. Full feature results

Results from the full feature experimental plan.

Base results

Table B.1 shows the F_1 and BA scores for the full feature experimental plan for all four ML methodologies and transport modes.

Table B.1: *Average performance metrics (F_1 and BA) using the four ML models with the full feature experimental plan.*

Model	Public	Private	Soft	One vs One
GB	0.8797	0.8653	0.7960	0.7872
RF	0.8546	0.8406	0.7379	0.7411
MLP	0.8571	0.8275	0.7297	0.7118
SV	0.8370	0.8178	0.6923	0.6895

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8728	0.8598	0.7686	0.7648
RF	0.8305	0.8289	0.6845	0.7002
MLP	0.8522	0.8257	0.6986	0.6976
SV	0.8565	0.8258	0.7647	0.7429

BA score

SMOTE results

Table B.2 shows the F_1 and BA scores for the full feature experimental plan with SMOTE applied to the training set for all four ML methodologies and transport modes.

Table B.2: *Average performance metrics (F_1 and BA) using the four ML models with the full feature experimental plan and SMOTE.*

Model	Public	Private	Soft	One vs One
GB	0.8802	0.8636	0.8129	0.7989
RF	0.8196	0.8368	0.7969	0.7732
MLP	0.8500	0.8395	0.7566	0.7154
SV	0.8263	0.8257	0.7123	0.7164

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8708	0.8647	0.8018	0.7951
RF	0.7936	0.8329	0.7814	0.7628
MLP	0.8439	0.8416	0.7631	0.7274
SV	0.8392	0.8207	0.7129	0.7167

BA score

B.1.2. Limited feature results

Results from the limited feature experimental plan.

Base results

Table B.3 shows the F_1 and BA scores for the limited feature experimental plan for all four ML methodologies and transport modes.

Table B.3: *Average performance metrics (F_1 and BA) using the four ML models with the limited feature experimental plan.*

Model	Public	Private	Soft	One vs One
GB	0.8559	0.8360	0.7656	0.7446
RF	0.8445	0.8289	0.7858	0.7574
MLP	0.8304	0.8066	0.7725	0.7330
SV	0.8246	0.8103	0.6750	0.6966

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8469	0.8322	0.7398	0.7250
RF	0.8283	0.8210	0.7496	0.7275
MLP	0.8266	0.8085	0.7669	0.7381
SV	0.8406	0.8226	0.8164	0.7770

BA score

SMOTE results

Table B.4 shows the F_1 and BA scores for the limited feature experimental plan with SMOTE applied to the training set for all four ML methodologies and transport modes.

Table B.4: *Average performance metrics (F_1 and BA) using the four ML models with the limited feature experimental plan and SMOTE.*

Model	Public	Private	Soft	One vs One
GB	0.8435	0.8358	0.7757	0.7450
RF	0.8173	0.8238	0.7844	0.7468
MLP	0.8222	0.7901	0.7441	0.7065
SV	0.7907	0.7923	0.7098	0.6902

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8329	0.8389	0.7869	0.7523
RF	0.7962	0.8261	0.8068	0.7602
MLP	0.8131	0.7922	0.7670	0.7263
SV	0.8169	0.7984	0.8141	0.7475

BA score

B.1.3. Latent variables results

Results from the latent variables experimental plan.

Base results

Table B.5 shows the F_1 and BA scores for the latent variables experimental plan for all four ML methodologies and transport modes.

Table B.5: *Average performance metrics (F_1 and BA) using the four ML models with the latent variables experimental plan.*

Model	Public	Private	Soft	One vs One
GB	0.8483	0.8309	0.7910	0.7613
RF	0.8257	0.8204	0.8004	0.7485
MLP	0.8320	0.8150	0.7718	0.7320
SV	0.8184	0.8021	0.6977	0.6852

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8444	0.8322	0.7893	0.7653
RF	0.8110	0.8168	0.7793	0.7328
MLP	0.8301	0.8164	0.7828	0.7401
SV	0.8411	0.8200	0.8602	0.7802

BA score

SMOTE results

Table B.6 shows the F_1 and BA scores for the latent variables experimental plan with SMOTE applied to the training set for all four ML methodologies and transport modes.

Table B.6: *Average performance metrics (F_1 and BA) using the four ML models with the latent variables experimental plan and SMOTE.*

Model	Public	Private	Soft	One vs One
GB	0.8315	0.8321	0.7878	0.7454
RF	0.8101	0.8101	0.7766	0.7308
MLP	0.8323	0.8040	0.7601	0.7103
SV	0.7992	0.7814	0.6980	0.6809

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8205	0.8323	0.8024	0.7514
RF	0.7877	0.8104	0.7918	0.7412
MLP	0.8232	0.8042	0.7917	0.7373
SV	0.8260	0.7855	0.8129	0.7425

BA score

B.1.4. CNN variables results

Results from the CNN variables experimental plan.

Base results

Table B.7 shows the F_1 and BA scores for the CNN variables experimental plan for all four ML methodologies and transport modes.

Table B.7: *Average performance metrics (F_1 and BA) using the four ML models with the CNN variables experimental plan.*

Model	Public	Private	Soft	One vs One
GB	0.8416	0.8188	0.7688	0.7440
RF	0.8427	0.8098	0.7718	0.7331
MLP	0.8320	0.7844	0.7551	0.7136
SV	0.8327	0.7915	0.6811	0.6574

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8416	0.8178	0.7450	0.7282
RF	0.8397	0.8056	0.7334	0.7155
MLP	0.8328	0.7862	0.7474	0.7191
SV	0.8504	0.8097	0.8236	0.7379

BA score

SMOTE results

Table B.8 shows the F_1 and BA scores for the CNN variables experimental plan with SMOTE applied to the training set for all four ML methodologies and transport modes.

Table B.8: *Average performance metrics (F_1 and BA) using the four ML models with the CNN variables experimental plan and SMOTE.*

Model	Public	Private	Soft	One vs One
GB	0.8286	0.8098	0.7592	0.7250
RF	0.8248	0.7893	0.7749	0.6977
MLP	0.8165	0.7895	0.7243	0.6864
SV	0.8160	0.7873	0.6975	0.5842

F1 score

Model	Public	Private	Soft	One vs One
GB	0.8197	0.8119	0.7713	0.7336
RF	0.8131	0.7884	0.8003	0.7056
MLP	0.8087	0.7927	0.7550	0.7236
SV	0.8326	0.7827	0.7950	0.6436

BA score