Application of Transformer Based Models for Discrete Choice Analysis

A thesis presented in part fulfilment of the requirements of the Degree of Master of Science in Transportation Systems at the Department of Mobility Systems Engineering, School of Engineering and Design, Technical University of Munich.

Supervisor	Dr. Santhanakrishnan Narayanan
	Univ.-Prof. Dr. Constantinos Antoniou
	Chair of Transportation Systems Engineering
Submitted by	Soma Shekhar Reddy
	0375129
Submitted on	Munich, 08th May 2025

Abstract

This thesis explores tabular Transformer models for travel mode choice prediction. Traditional Discrete Choice Models (DCMs) like Multinomial Logit (MNL) offer good interpretability but have limitations with complex patterns. Machine learning models like Random Forest and XGBoost often predict better but are harder to understand. This study tests whether Transformer models (TabTransformer and FT-Transformer) can provide reasonable predictions and interpretability. The study uses two datasets: a small West Midlands (WM) dataset and an extensive London Passenger Mode Choice (LPMC) dataset. The study compares Transformers against traditional models using accuracy, balanced accuracy, and other metrics. The study also analyzes interpretability through SHAP values and attention mechanisms. The results show that on the large LPMC dataset, FT-Transformer achieves 74.44% accuracy, performing almost as well as the best model (XGBoost at 74.72%) and better than MNL (72.54%). However, Transformers perform poorly on the small WM dataset, showing they need more data to work well. SHAP analysis reveals that Transformers capture expected travel behavior patterns. Attention analysis shows that with enough data, Transformers can discover meaningful relationships between travel features. This study presents benchmark results for Transformer models in transportation choice modeling, analyzes their interpretability through SHAP values, and examines ways to handle imbalanced transportation data.

Acknowledgment

I would like to express my heartfelt thanks to everyone who helped me complete this thesis. I am deeply grateful to my supervisors, Dr. Santhanakrishnan Narayanan and Prof. Dr. Constantinos Antoniou, for their constant guidance, support, and inspiration throughout this journey. Their expertise in modeling was invaluable to my research. I want to especially thank Santhanakrishnan Narayanan for agreeing to supervise my work, for his endless patience, and for sharing his valuable ideas and suggestions. This research could not have happened without both supervisors' interest in my topic and their willingness to guide me at every step. I also want to thank my family and friends who stood by me with their unconditional love and encouragement. Their belief in me gave me the strength to complete this work.

Contents

List of Figures

List of Tables

Glossary

A

AUROC - Area Under the Receiver Operating Characteristic Curve 26, 32, 33

C

CART - Classification And Regression Trees 23
CV - cross-validation 34, 35

D

DCMs - Discrete choice models 10, 12–17, 19
DNN - Deep Neural Networks 10, 18, 19, 36
DT - Decision Trees 10, 11, 20, 23, 33

G

GMPCA - Geometric Mean Probability of Correct Assignment 18, 35, 36

I

IIA - Irrelevant Alternatives 10, 13

L

LPMC - London Passenger Mode Choice 11, 12, 18, 19, 22, 24, 26, 27, 35, 46, 47

M

ML - Machine Learning 10, 12, 14–16, 18–20
MNL - Multinomial Logit 10–15, 18–20, 22, 23, 35, 36

N

NL - Nested Logit 13
NN - Neural Networks 10, 11, 14, 15, 18, 19, 36, 47

P

PHOEBE - Predictive Approaches for Safer Urban Environment 22

R

RF - Random Forest 10, 11, 15, 16, 18, 20, 23, 33

S

SHAP - SHapley Addiive exPlanations 10–12, 17–20, 27, 31, 35, 46, 47
SMOTE - Synthetic Minority Oversampling Technique 11, 12, 17, 19, 25, 31, 34, 35, 49
SVM - Support Vector Macines 10, 11, 14, 18, 19, 36, 47

V

VOT - Value of Time 13, 14

W

X

1. Introduction

Understanding how people choose between cars, public transit, or walking modes is crucial for transportation planning, infrastructure development, and mobility policies [2]. For decades, transport planners have relied on DCMs (Discrete choice models), especially the MNL (Multinomial Logit) model and its extensions, which are based on the Random Utility Maximization theory [3, 4].

These models provide significant advantages due to their robust theoretical foundation and interpretable utility coefficients. These coefficients can explain how various factors influence an individual's choices, which is crucial for policy analysis. However, DCMs also have certain limitations. The MNL model is based on strict assumptions, such as the Independence of IIA (Irrelevant Alternatives), and typically employs simplified linear utility specifications that may not adequately capture the complexities of human decision-making [5].

Over the past decade, ML (Machine Learning) techniques have increasingly been adopted for travel behavior modeling. Models such as DT (Decision Trees), RF (Random Forest), Gradient Boosting Machines (including XGB (Extreme Gradient Boosting)), SVM (Support Vector Macines), and NN (Neural Networks) are capable of learning complex, non-linear patterns directly from data, without requiring stringent assumptions about behavior [6, 7]. Numerous studies have demonstrated that these ML models frequently outperform traditional DCMs regarding predictive accuracy for mode choice tasks [2, 8].

Despite this improved predictive performance, ML models are frequently criticized as "black boxes," making it difficult to understand the underlying factors driving predictions [9, 10]. This presents a significant challenge for transportation modelers.

This trade-off between predictive accuracy and behavioral interpretability is not merely theoretical. A recent comprehensive study by Martín-Baos et al. [1] systematically evaluated a wide range of ML classifiers against the MNL model using both synthetic and real-world datasets. Their findings highlight that tree ensemble methods like XGB and RF demonstrated the highest predictive accuracy. However, behavioral analysis revealed significant inconsistencies – these models produced poor estimates of behavioral indicators like WTP (Willingness To Pay) and market shares. Additionally, their interpretations via SHAP (SHapley Addiive exPlanations) were found to be problematic from a behavioral perspective [1].

By contrast, NN, DNN (Deep Neural Networks), and SVM showed more balanced performance. While achieving slightly lower predictive accuracy than the tree-based models, they exhibited much better behavioral consistency, yielding plausible WTP estimates and SHAP values that aligned well with economic theory and MNL results.

The MNL model itself, though demonstrating lower predictive accuracy than the machine learning approaches, served as the robust behavioral baseline due to its strong theoretical foundation and direct economic interpretability.

This benchmark shows that selecting an ML model based purely on predictive metrics like accuracy can be misleading for policy analysis if behavioral consistency is compromised. Similar trade-offs have been observed in other studies [7, 10], which will be explored in depth in the literature review chapter.

1.1. Tabular Transformers

Recent advances in deep learning have led to adopting the Transformer architecture designed initially for language processing for tabular data. Models like TabTransformer [11] and FT-Transformer [12] use self-attention mechanisms to process tabular data in new ways.

TabTransformer focuses on learning contextual embeddings for categorical features through attention, while FT-Transformer processes numerical and categorical features through stacked attention layers. The self-attention mechanism allows these models to dynamically assess the importance of different features and capture complex, context-dependent interactions. This potentially offers more powerful representation capabilities than standard Neural Networks or the static splits of decision trees [12].

This raises a research question: Can these attention-based Transformer architectures achieve the predictive accuracy of models like XGB while maintaining behavioral consistency and interpretability?

1.2. Research Objectives

This thesis aims to address these gaps through the following objectives:

1. Benchmark the predictive performance (Accuracy, Balanced Accuracy, Macro F1) of DT, RF, XGB, TabTransformer, and FT-Transformer on the WM (West Midlands).
2. Compare TabTransformer and FT-Transformer LPMC (London Passenger Mode Choice) performance directly against the results reported by [1].
3. Evaluate SMOTE (Synthetic Minority Oversampling Technique) impact on generalization for the implemented models with the WM dataset.
4. Assess the behavioral consistency of Transformer SHAP values compared to MNL and Martín-Baos et al.'s NN and SVM findings.
5. Analyze insights derived from Transformer Attention analysis versus SHAP analysis.

1.3. Thesis Contributions

This thesis aims to make the following contributions:

- Provide the benchmark results for TabTransformer and FT-Transformer against the models evaluated by Martín-Baos et al. [1] on the LPMC and WM dataset.
- Contribute evidence regarding the behavioral consistency of SHAP values derived from Transformer models.
- Offer analysis of SMOTE effectiveness and limitations for these models in the mode choice context.

1.4. Thesis Outline

The remainder of this thesis is organized as follows: Chapter 2 provides a literature review covering DCMs, ML applications in choice prediction, transformer architectures for tabular data, and modern approaches to model interpretability. Chapter 3 details the methodology, including data description, preprocessing techniques, MNL specification, ML model configurations, evaluation metrics, and the approach to interpretability analysis. Chapter 4 presents the results, including MNL model estimates, comparative predictive performance of all models, SHAP analysis findings, attention visualizations, and interpretability comparisons. Chapter 5 discusses the implications of these results, addressing the research questions, examining trade-offs between different modeling approaches, and acknowledging limitations. Chapter 6 concludes with a summary of key findings, answers to the research questions, and suggestions for future research directions.

2. Literature Review

2.1. DCM in Transportation

DCMs have been widely used to understand travel behavior for decades. These models are based on the idea that people make rational choices to maximize utility [4]. DCMs are practical tools in transportation planning because they are easy to understand and have a strong theoretical foundation. Aboutaleb et al. [9] state that clear models are best for understanding, as they let you fully understand mechanisms and assumptions with parameters that give intuitive explanations. DCMs provide a framework for measuring traveler preferences and tradeoffs, which helps calculate economic measures like the VOT (Value of Time) and willingness to pay for service improvements [13, 14]. This makes DCMs helpful for policy analysis, where understanding causes is often as important as prediction accuracy.

The MNL, developed by McFadden[3], became the basic model for choice modeling because it is simple to calculate and use. In this model, the probability of choosing a particular mode depends on its utility compared to all other options:

$$P(i|x) = \frac{\exp(V_i)}{\sum_j \exp(V_j)} \tag{2.1}$$

Where $P(i|x)$ is the probability of choosing mode i, V_i is the utility of mode i, and the sum in the denominator adds up the utilities of all available modes.

However, the MNL models have limitations where choice alternatives are considered independent. The IIA comes from assuming error terms are independent. If a choice is added or removed, the odds between two choices stay the same, which does not match how people behave, causing problems in real use [7]. The classic example is the "red bus/blue bus problem." Suppose identical red and blue buses are introduced to compete with cars. In that case, the MNL predicts that both buses would take twice as much share from cars as a single bus would, which is unrealistic since travelers would likely see them as one option (public transport).

In order to address the challenges related to the violation of the IIA assumption, researchers have developed more advanced models, including the NL (Nested Logit) model and mixed logit. These approaches offer effective alternatives for analysis when the traditional IIA assumption is not applicable. The NL model puts similar choices into groups called "nests," allowing connection between options in the same nest [15]. Mixed logit models allow random taste variations and free substitution patterns [5]. These models are more flexible but harder to calculate, especially with large datasets or complex specifications.

Traditional DCMs present several limitations. Their performance depends on accurate specification, which requires significant expert knowledge. Additionally, these models often struggle with complex non-linear relationships unless the modeler explicitly programs them. As travel behavior becomes more complex, manually identifying relevant interactions becomes increasingly challenging [16]. As transportation systems evolve with shared mobility, autonomous vehicles, and mobility-as-a-service platforms, traditional DCMs face growing challenges in capturing modern travel decisions of a multidimensional nature [17, 18]. These emerging mobility patterns often involve decision processes that may not fit neatly into the traditional utility maximization framework [19]. DCMs also perform poorly with imbalanced datasets. When certain choices like car trips dominate the data, models may struggle to accurately predict less common choices like cycling, even when these are important for policy [2].

These limitations have motivated research into alternative approaches, including machine learning techniques for travel behavior modeling and hybrid methods that combine the interpretability of DCMs with the flexibility and predictive power of machine learning [7, 9, 20].

2.2. ML Applications in Mode Choice

Transportation researchers began exploring ML for travel behavior analysis in the 1990s, with early studies showing NN outperforming traditional MNL models in predicting travel mode choices [21].

Xie et al. [22] found that decision trees and neural networks achieved better prediction than MNL for commuting mode choices in San Francisco. Building on this, Zhang and Xie [23] demonstrated that SVM could predict commuter travel mode choice with 84.8% accuracy compared to MNL 79.7%. Hagenauer et al. [6] comprehensively compared seven ML algorithms against MNL, finding that random forest performed best with 91.4% accuracy, while MNL reached only 56.1%. Wang et al [2] showed that XGB achieved 84.5% overall accuracy compared to MNL 80.7%, with XGB performing notably better for predicting car use (91.4% vs. 83.0%) and walking (71.3% vs. 41.5%), though worse for transit use (19.3% vs. 33.3%). This study highlighted that ML advantages vary across transportation modes, with larger improvements typically for dominant modes.

Zhou et al. [24] compared ML techniques for modeling bike-sharing versus taxi choices in Chicago, finding that tree-based ensemble methods significantly outperformed models like logistic regression, highlighting the importance of appropriate metrics when dealing with imbalanced transportation datasets [24]. Wang et al[25] developed a deep neural network with alternative-specific utility functions (ASU-DNN) that achieved 87.3% accuracy, outperforming both standard MNL (71.9%) and conventional NN (84.5%) while producing reasonable VOT estimates.

In one of the most comprehensive comparisons, Zhao et al. [7] tested multiple ML and

DCMs approaches, finding that RF had the highest prediction accuracy (85.6%), followed by Gradient Boosting (83.8%), with MNL achieving only 74.6%. ML models consistently show better predictive performance, especially for large datasets with complex patterns, with multiple studies documenting accuracy improvements of 5-20% compared to traditional DCMs [6, 7, 22, 23]. Wang and Ross [2] demonstrated ML ability to handle high-dimensional input by including over 30 variables in their model without performance degradation. While most studies confirm ML predictive advantages, some research shows that this can depend on context and metrics. Ali et al. [10] found that a well-specified MNL model with piecewise linear transformation of household income performed better than neural networks and gradient-boosting trees in terms of log-likelihood and mean absolute percentage error of market shares, suggesting specified traditional models remain competitive in specific contexts.

There appears to be a fundamental tradeoff between predictive accuracy and behavioral soundness when choosing between machine learning and logit models [7, 9]. While ML models typically predict better, their lack of theoretical foundation makes them less suitable for understanding behavior or policy analysis. For example, getting information about elasticities, the value of time, and other behavioral indicators need extra steps with ML models. Wang et al. [25] found that elasticities derived from NN showed higher variance and sometimes counterintuitive results, like positive values for cost elasticities. ML models also require larger datasets than traditional DCMs to perform well [26]. Paredes et al.[26] showed that with small samples (below 5,000 observations), MNL outperformed ML methods, with ML showing significant advantages only with datasets over 15,000 observations. ML models often demonstrate poor transferability to new contexts. Hillel et al. [20] found that RF models trained on London data achieved 83% accuracy locally but only 56% when applied to Manchester data, compared to MNL models, which showed a smaller performance drop (from 65% to 49%). ML models are sensitive to hyperparameter tuning, with minor configuration changes causing significant performance differences, requiring extensive cross-validation and careful optimization. Finally, ML models generally lack the theoretical foundation of economic choice theory that underpins DCMs, raising concerns about their ability to generalize beyond training data, particularly in forecasting scenarios where conditions may change substantially [9, 20].

2.3. Transformer Models for Tabular Data Analysis

Transformer models introduced by Vaswani et al. [27] use attention mechanisms to capture relationships between inputs and outputs. While initially developed for language tasks, they have expanded to handle other data types, including tabular data used in transportation modeling.

Huang et al. [11] introduced TabTransformer, which specifically adapts transformer architecture for tabular data. This model transforms categorical feature embeddings into robust contextual embeddings using self-attention transformer layers. Their extensive experiments demonstrated that TabTransformer outperformed other deep learning methods for tabular

data by at least 1.0% on mean and matched the performance of tree-based ensemble models in many cases. TabTransformer works by embedding categorical features, transforming them through transformer layers, and then combining them with numerical features. This approach captures complex interactions while maintaining interpretability.

In 2021, Gorishniy et al. [12] proposed FT-Transformer, which transforms categorical and numerical features into embeddings before processing. By treating each feature as a token, it models interactions more directly. Their tests across fifteen datasets showed that FT-Transformer outperformed other approaches on most tabular tasks, especially in its ability to work well across different datasets. More recently, Wang et al. [28] developed Transformer Choice Net specifically for choice prediction. It handles single, sequential, and multiple-choice scenarios, making it relevant for transportation modeling. It performed better than leading discrete-choice models without requiring custom tuning for each scenario.

Transformer models offer several advantages that make them promising for transportation choice modeling. They excel at capturing complex relationships between features without needing explicit specifications. Traditional DCMs struggle with complex non-linear relationships unless specifically defined. As transportation systems become more complex with shared mobility and autonomous vehicles, this ability to automatically discover patterns becomes increasingly valuable.

Transformers strike a good balance between predictive power and interpretability. While methods like RF provide limited explanations through feature importance, transformers offer rich interpretability through their attention mechanisms. Gorishniy et al. [12] showed that attention maps from FT-Transformer correlate highly with traditional methods for determining feature importance. This interpretability is crucial in transportation planning, where understanding why a prediction was made can be as important as the prediction itself. These models handle imbalanced datasets effectively - a significant advantage when modeling transportation choices, where car trips typically dominate datasets. This can improve the prediction of less common but policy-relevant modes like cycling or carpooling. Transformers process many features of different types simultaneously, reducing the need for manual feature selection. Traditional models become challenging to manage when there are many features and interactions. In contrast, transformers can effectively handle dozens or even hundreds of features simultaneously.

2.4. Handling Class Imbalance

Mode choice data is often imbalanced, with some modes (like car in many areas) dominating the dataset while others (like cycling) appear much less frequently. This imbalance creates challenges for ML models, which may focus on maximizing overall accuracy by predicting the majority class well while performing poorly on minority classes.

This is particularly problematic in transportation planning, where accurately modeling less common but policy-relevant modes (like cycling or carpooling) is often a priority. Standard accuracy metrics can be misleading in these situations, as a model that always predicts the majority class might achieve high accuracy while being useless for policy analysis. One approach to address the imbalance is to modify the training data. SMOTE is a popular algorithm that identifies and creates synthetic examples by interpolating between a minority class example and its nearest neighbors. These synthetic examples are then added to the training data to balance class distributions. SMOTE and its variants (like Borderline-SMOTE and ADASYN) aim to provide more balanced training data without simple duplication. However, these methods can introduce pitfalls. They may create unrealistic synthetic examples that don't match real-world distributions. They can also lead to overfitting if the synthetic examples don't generalize well. Furthermore, they change the underlying data distribution, which may be undesirable if the true distribution is important for subsequent policy analysis.

An alternative is to modify the learning algorithm rather than the data. Class Weighting assigns higher weights to minority class examples during training, often using inverse frequency or effective number of samples. Focal Loss modifies the loss function to focus more on difficult examples, using parameters gamma (which down-weights easy examples) and alpha (which balances classes). These methods keep the original data distribution intact while encouraging the model to pay more attention to minority classes, which may be preferable when preserving the real-world distribution is important for subsequent analysis. If the goal is to accurately simulate real-world market shares and choice probabilities for policy analysis, methods that change the training distribution (like SMOTE) or fundamentally alter the loss objective (like Focal Loss with high gamma) might produce models that don't reflect real-world behavior. On the other hand, if identifying and correctly classifying minority classes is the primary goal (e.g., for targeted policies to promote cycling), these methods might be beneficial despite their potential to distort overall probabilities.

2.5. Model Interpretability

While traditional DCMs offer inherent interpretability through their parameters, the black-box nature of many machine learning models poses significant challenges for transportation planning applications, where understanding the underlying decision mechanisms is crucial [9]. Transportation planners and policymakers not only require accurate predictions of mode choices but also clear explanations of how factors such as travel time, cost, and demographic characteristics influence these decisions in order to design effective policies and justify investments [7].

As machine learning models demonstrate increasingly superior predictive performance in mode choice analysis, the development of robust interpretability methods becomes critical to bridging the gap between prediction accuracy and behavioral understanding [20]. Among the various explainable AI techniques, SHAP has emerged as a comprehensive method for

interpreting model predictions. Grounded in cooperative game theory, SHAP values attribute the change in expected model predictions to each feature by conditioning on that feature. This method assigns an importance value to each input feature for a specific prediction while adhering to three essential properties: local accuracy, missingness, and consistency [29]. This makes SHAP especially valuable in transportation mode choice modeling, where understanding the relative importance of factors such as travel time, cost, and demographic characteristics is crucial.

Transformer models offer another avenue for interpretability through their attention mechanisms. Attention weights show which input tokens contribute most to generating each output token's representation at a specific layer/head. These weights can be visualized as heatmaps showing the pairwise relevance between features. Strong attention between two features might indicate an important interaction influencing the prediction. By analyzing these patterns, researchers can gain insights into how the model processes different features and which combinations it finds most relevant.

However, interpreting attention has limitations. Attention weights are internal model components, not direct attributions of output, making their interpretation less straightforward than SHAP values. Aggregation across multiple layers and heads is complex, as different heads may capture different types of relationships. Additionally, attention can be diffuse and inconsistent across different inputs, making it challenging to draw general conclusions. Despite these challenges, attention mechanisms provide a unique window into how transformer models process information and could complement SHAP analysis for understanding model behavior.

A recent comprehensive study by Martín-Baos et al. [1] systematically evaluated a wide range of ML classifiers against the MNL model using both synthetic and real-world datasets, including the LPMC dataset that will also be used in this thesis. Their methodology included testing multiple ML models (RF, XGB, SVM, NN, DNN) against MNL using proper validation methods to avoid data leakage. They evaluated both predictive performance (accuracy, GMPCA (Geometric Mean Probability of Correct Assignment)) and behavioral consistency (market shares, WTP) and analyzed model interpretability using SHAP. Their key findings revealed a clear tradeoff between prediction and behavioral consistency. Their study found that tree ensemble models (XGB, RF) showed the highest predictive accuracy but exhibited poor behavioral consistency. They produced unreliable WTP estimates and market share predictions, and their SHAP values often contradicted expected economic relationships. NN, DNN and SVM showed slightly lower accuracy but much better behavioral consistency. Their WTP estimates were reasonable, and their SHAP values aligned well with economic theory, making them more suitable for policy analysis. The MNL model, while having the lowest predictive accuracy, served as a robust behavioral baseline due to its strong theoretical foundation and directly interpretable parameters. The authors concluded that model selection should consider both predictive performance and behavioral consistency, with the choice de-

pending on the specific application. For policy analysis requiring behavioral understanding, models like NN, DNN and SVM offered the best compromise between accuracy and interpretability. This benchmark study provides a crucial reference point for our evaluation of transformer models, allowing us to determine where these newer architectures fall along the accuracy-interpretability spectrum.

2.6. Literature Gaps

Based on the literature review, several important research gaps have been identified that this thesis aims to address: Despite their promise, TabTransformer and FT-Transformer have not been systematically benchmarked against the full suite of ML models and DCMs evaluated by Martín-Baos et al. [1] on common travel behavior datasets like LPMC. While these transformer architectures have shown strong performance on general tabular data tasks, their effectiveness specifically for mode choice prediction remains unexplored.

While interpretability methods like SHAP can be applied to these models [9], their reliability and behavioral consistency when explaining tabular Transformers in the mode choice context remain unvalidated. Furthermore, the potential of the intrinsic attention mechanism as a tool for deriving behavioral insights and how it compares to SHAP in this domain is largely unexplored. There is limited understanding of whether transformer-derived SHAP values provide the same level of behavioral consistency as those from NN and SVM in the benchmark study.

Travel mode choice data is often highly imbalanced, with some modes being much more common than others. The effectiveness of standard techniques like SMOTE for improving Transformer generalization requires investigation, particularly given concerns about altering real-world distributions. No studies have systematically evaluated how imbalance-handling techniques affect transformer models in transportation mode choice.

The sensitivity of Transformer performance and interpretability to dataset size in the context of mode choice modeling is not well understood [26]. While transformers typically require substantial data for language tasks, their data requirements for tabular classification tasks in transportation are unknown. Understanding how these models perform with different dataset sizes would provide valuable guidance for practitioners.

This research aims to evaluate the predictive performance and interpretability of Transformer-based machine learning models (TabTransformer, FT-Transformer) compared to traditional machine learning methods and a benchmark MNL model for choice prediction, specifically investigating the utility of SHAP and attention analysis for extracting behavioral insights.

Specifically, this study seeks to answer the following key research questions:

How do TabTransformer and FT-Transformer compare against MNL and other ML models (RF, DT, XGB) in predicting choices on unseen data, particularly using balanced accuracy metrics?

To what extent do feature importance rankings and directional effects derived from SHAP analysis on transformer models align with or diverge from the significant variables and behavioral interpretations of the MNL model?

Can SHAP dependence plots or attention mechanisms reveal plausible non-linear relationships or feature interactions learned by transformers that are not explicitly captured by the linear MNL specification?

To address these questions, this research will develop an MNL model, train various ML models, including transformer-based architectures, apply SHAP and attention analysis to these models, and conduct comprehensive predictive performance and interpretability comparisons.

3. Methodology

3.1. Methodology

This chapter outlines the comprehensive methodology employed to compare traditional discrete choice models with machine learning approaches for mode choice modeling. The methodology follows a structured approach encompassing data preparation, model implementation, training procedures, and comparative analysis

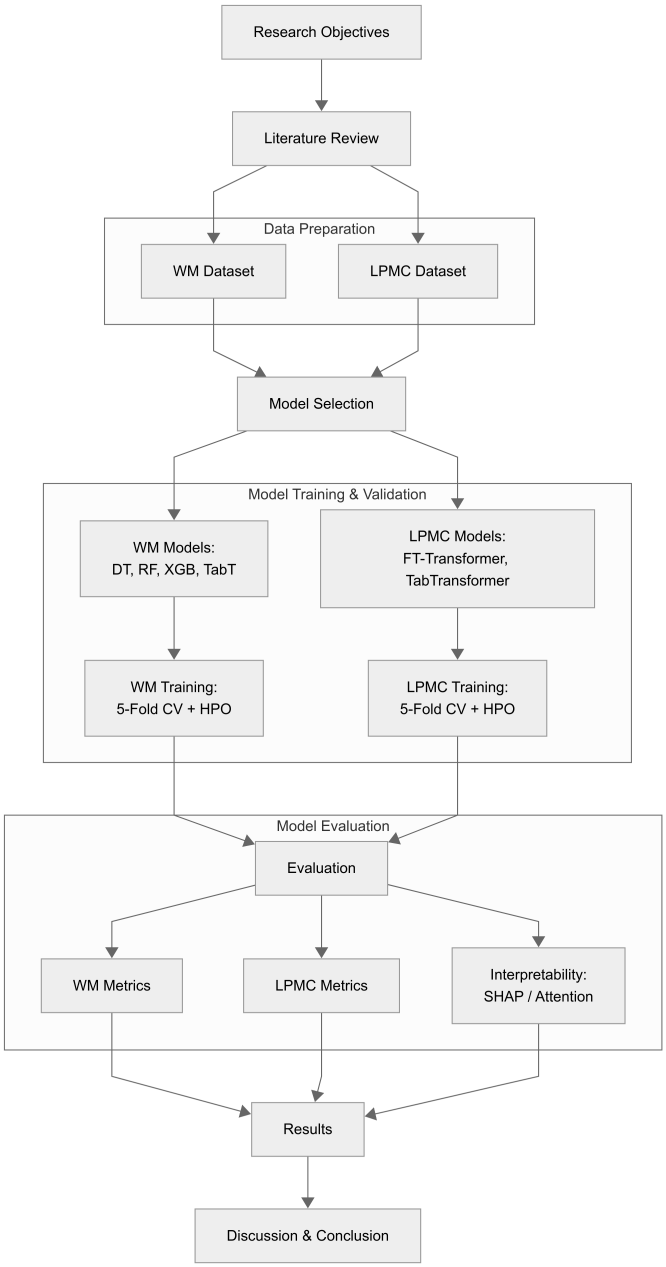


Figure 1 Methodology of the study

3.1.1. Datasets

The study utilizes two distinct datasets to ensure robust evaluation across different contexts and sample sizes. The WM dataset consists of Stated Preference survey data gathered from the WM region of the UK as part of the PHOEBE (Predictive Approaches for Safer Urban Environment) project. After preprocessing, the dataset comprises approximately 4,008 responses from 501 individuals. It includes socio-demographic factors such as age and gender, trip characteristics like purpose and time of day, and mode-specific attributes including travel times and costs. The target variable reflects the selected travel mode for each scenario, which consists of five choices: car (33.4%), public transport (29.8%), walking (23%), cycling (9.4%), and micro-mobility services (4.5%). The mode choices exhibit significant imbalance, with car being the dominant mode and micro-mobility services representing the smallest choices.

The LPMC dataset was collected by Hillel et al. [30] and consists of single-day travel diary data from the London Travel Demand Survey conducted between 2012 and 2015. This dataset was also used in the benchmark study by Martín-Baos et al. [1], enabling direct comparison of results. The dataset contains approximately 81,096 records, each of which corresponds to one trip made by one of the 17,616 participants in the survey, larger than the West Midlands dataset. From the original 31 variables, we use the same 20 variables that were used in the Martín-Baos et al. [1] study. The choice variable takes the following values: car (44.16%), public transport (35.28%), walking (17.56%), and cycling (2.98%). This dataset also exhibits considerable class imbalance for the cycling mode.

3.1.2. Data Preprocessing

For the WM dataset, the preprocessing phase involved several steps to ensure data quality. First, observations classified as "None" were removed from the analysis due to their low frequency and unclear nature. Analysis of response patterns revealed that approximately 139 of the respondents demonstrated lexicographic behavior, consistently choosing the same transportation modes across all scenarios. These patterns suggest the presence of response bias or strong habitual preferences, prompting their removal from the dataset. Features characterized by redundancy, correlation, and irrelevance were removed. The final dataset comprises 21 features, which were utilized for model training. Further details regarding the features can be found in the appendix. Unique individual identifiers were retained for grouped splitting but removed before model training. Missing values were imputed using K-Nearest Neighbors imputation with k=5, which preserves relationships between variables better than simple mean or mode imputation.

For the LPMC Data set, we use 20 features utilized in their Martín-Baos et al. [1] study, with no additional preprocessing to the dataset available on GitHub.

3.1.3. Models Implemented

The study implements and evaluates various machine learning methodologies while conducting interpretability assessments compared to the MNL model, which serves as the baseline.

MNL

The MNL model serves as the traditional discrete choice model baseline. This model operates on the principle that people select alternatives that provide them the most utility. The utility comprises a measurable part based on observable factors like travel time or cost, and a random part accounting for unobserved influences. This mathematical structure produces a simple formula where the probability of choosing an option equals the exponential of its utility divided by the sum of exponentials for all available options [3].

DT

DT creates a flowchart-like structure where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a prediction [31]. The model selects the best feature and threshold at each node based on measures such as Gini impurity, which evaluates how well a potential split separates the classes. The algorithm continues splitting the data into increasingly homogeneous subsets until reaching stopping criteria such as maximum depth or minimum samples per leaf [32]. The implementation used the standard CART (Classification And Regression Trees) algorithm via scikit-learn.

RF

RF enhances the DT method by forming an ensemble of trees, with each tree trained on a bootstrap sample of the original dataset [33]. This approach, called bagging, minimizes variance and mitigates overfitting risk. Furthermore, RF incorporates feature randomization at each split. The final prediction comes from aggregating all trees' predictions through majority voting. This ensemble approach significantly improves predictive performance compared to single decision trees while maintaining reasonable interpretability through feature importance measures [34]. The implementation used scikit-learn's RandomForestClassifier with tuned hyperparameters, including the number of trees, maximum depth, and maximum features.

XGB

XGB constructs an ensemble of weak predictive models sequentially [35]. Unlike RF, which develops trees independently, XGB creates each new tree to correct errors made by the existing ensemble. The model introduces several innovations compared to traditional gradient boosting, including L1 and L2 regularization terms, to mitigate overfitting and efficiently handle sparse data through sparsity-aware split finding. Additionally, XGB employs advanced tree pruning techniques and handles missing values by automatically determining the best direction for these gaps [36, 37]. The implementation used the xgboost library with tuned hyperparameters, including learning rate, maximum depth, subsample ratio, and regularization parameters.

TabTransformer

TabTransformer is an advanced deep-learning architecture specifically designed for tabular data [11]. The model processes categorical and numerical features differently. Categorical features pass through an embedding layer, converting them to dense vector representations. Each embedding includes a column identifier component and a value-specific component, helping the model differentiate between the feature identity and its specific category value. These embedded categorical features are then processed through transformer encoder layers, which include multi-head self-attention, feed-forward networks, layer normalization, and residual connections. The encoder layer is based on the attention mechanism established by Vaswani et al. [27]. This process transforms the initial embeddings into contextual embeddings, integrating information from other relevant features. Finally, these contextual embeddings are concatenated with the continuous features before processing through a multi-layer perceptron for the final prediction. The implementation used PyTorch-Tabular with a custom TabTransformerModel class, allowing attention weight extraction.

FT-Transformer

The FT-Transformer [12] adopts a unique approach by treating categorical and numerical features uniformly through a transformer-based architecture. Unlike TabTransformer, it processes all features using a unified tokenization method. Categorical features are embedded using standard embedding layers, while numerical features undergo linear projection followed by normalization. This design enables all features to coexist within the same embedding space, facilitating direct interactions within transformer layers. The model integrates learnable type embeddings into each feature representation, which helps distinguish among the various features. These embeddings are conceptually similar to the positional embeddings utilized in language transformers, as they share information about the types of features rather than their positions. The architecture is built upon multiple transformer encoder blocks that employ multi-head self-attention, feed-forward networks with GLU activation, a pre-norm configuration, and residual connections. The implementation leveraged PyTorch-Tabular with a custom FTTransformerModel class, which allows for the extraction of attention weights [38].

3.1.4. Experimental Design

The experimental design establishes a rigorous framework for model training, validation, and comparison.

Data Splitting

For the WM dataset, a 90/10 train/test split was performed using a stratified group k-fold due to the limited data. For the LPMC dataset, the last year of data (2014/15) was taken for testing, and the remaining years for training (2012/13–2013/14); that is, approximately 70% training and 30% test, similar to Martín-Baos et al. [1]. Following Hillel et al. [20], we avoided

bias in our model testing by ensuring that all trips from the same person appear only in the training set or only in the test set, never in both. This person-wise sampling approach gives us more accurate performance estimates by testing models on truly unseen data.

Cross-Validation

A 5-fold stratified group k-fold cross-validation approach was applied to the training set. This cross-validation ensures robust evaluation while respecting group structures to prevent data leakage between folds. The cross-validation maintained consistent class distribution across folds through stratification by transportation mode choice.

Hyperparameter Optimization

Hyperparameter optimization was implemented for all machine learning models to ensure optimal performance. The process employed a random search with 30 trials per model per dataset, efficiently searching the hyperparameter space. The PyTorch-Tabular model tuner interface to Optuna was used for transformer models, while scikit-learn's RandomizedSearchCV supported optimization for tree-based models. The primary metric optimized was balanced accuracy to prioritize performance across all classes, including minorities. Parameter grids covered each model's key architectural, regularization, and learning parameters. Tuning used a single train/validation split derived from the main 90% training data, with the best parameters then applied for the 5-fold cross-validation evaluation.

Class Imbalance Handling

The transportation choice datasets exhibited substantial class imbalance, with the car representing the dominant mode. For the WM dataset, Synthetic Minority Over-sampling Technique (SMOTE) and Focal loss were implemented to handle the imbalance. SMOTE was implemented using the imbalanced-learn library with the default strategy of sampling all minority classes to match the size of the majority class. SMOTE was applied only to the training portion of each cross-validation fold immediately before training the model for that fold. For evaluating the impact on generalization, models were also trained on the full SMOTE-sampled training set and evaluated on the original, unsampled 10% test set.

In addition to SMOTE, we implemented Focal Loss [39] as an algorithm-level approach for handling class imbalance. Unlike SMOTE, which modifies the training data, Focal Loss adapts the loss function to focus on hard-to-classify examples. The standard cross-entropy loss is modified by adding a factor $(1 - p_t)^\gamma$, where p_t is the probability of the true class and γ is a focusing parameter:

$$\text{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \tag{3.1}$$

This down-weights well-classified examples (high p_t), forcing the model to focus on difficult examples, which are often from minority classes. We specifically used Class-Balanced Focal Loss, which adds class weights to further address imbalance:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t) \tag{3.2}$$

where α_t is the weight for class t, set inversely proportional to class frequencies. We used $\gamma = 2.0$ based on recommendations from [39], balancing the focus on hard examples without overly distorting the model's probability estimates.

3.1.5. Evaluation Metrics

Multiple performance metrics were employed to comprehensively evaluate the model's predictive capabilities. For the West Midlands dataset, accuracy measures the overall percentage of correct predictions, while balanced accuracy calculates the average recall obtained in each class, making it crucial for imbalanced data. Macro F1-score provides the unweighted mean of F1-scores calculated for each class, treating all classes equally regardless of frequency. The weighted F1-score calculates the mean of F1-scores weighted by the support for each class, biasing toward majority classes. The evaluation also included precision and recall metrics in both macro and weighted variants. Confusion matrices are used to visualize misclassifications between classes, and they are presented as both raw counts and row-normalized values. AUROC (Area Under the Receiver Operating Characteristic Curve) measures overall discrimination ability, typically calculated using a One-vs-Rest approach for multi-class problems. The emphasis was placed on balanced accuracy and macro F1-score for evaluating success in handling class imbalance, as these metrics better reflect performance across all transportation modes regardless of their prevalence in the dataset. For the LPMC dataset, we used Accuracy and GMPCA.

The Geometric Mean of Predicted Probabilities of the Chosen Alternative (GMPCA) is a performance metric employed to assess how effectively a model predicts the actual choices observed in a dataset, while also accounting for the model's confidence in these predictions. It is particularly recommended for the behavioural analysis of choice models, as probability-based indices are considered essential for this purpose Hillel2019. The GMPCA is calculated using the expression:

$$\text{GMPCA} = \left(\prod_{n=1}^{N_{\text{test}}} P_{n,i_n} \right)^{\frac{1}{N_{\text{test}}}} \tag{3.3}$$

where N_{test} is the total number of observations in the test set, and P_{n,i_n} is the model's predicted probability for the alternative i_n that was actually chosen for the n-th observation. In

many applications, this index is subsequently expressed as a percentage by multiplying the result of Equation (3.3) by 100. A higher GMPCA value indicates that the model, on average, assigns higher probabilities to the correctly chosen alternatives.

3.1.6. Interpretability Analysis

An interpretability analysis was conducted to determine which models provide more actionable insights for transportation planning.

SHAP Analysis

SHAP values were computed using the Kernel Explainer from the SHAP library [29]. We chose this method because it works with any model type, allowing consistent comparisons across all our models, though it requires more computation time. A wrapper function providing model probabilities served as the prediction function for the explainer, with a logit link function employed to perform calculations in the log-odds space. The background data, crucial for approximating conditional expectations in SHAP, consisted of 10 centroids generated using k-means clustering applied to the respective training fold dataset. This specific approach to selecting the background data, along with other implementation details, followed the methodology found in previous benchmark studies in this domain [1]. SHAP values were computed for the entire test dataset for the WM dataset due to its smaller size. For the larger LPMC dataset, owing to computational constraints, we used 1000 randomly selected samples from the test dataset for SHAP value calculations. The outputs for interpretation included beeswarm plots and summary bar plots showing mean absolute SHAP values, which are standard visualizations associated with the SHAP framework.

Attention Analysis

To complement our SHAP analysis [29], we examined the attention mechanisms in our Transformer models after training. We collected attention weights from all model layers and heads during test set evaluation. These weights show how strongly different features connect to each other in the model's decision process. We averaged these weights across layers and heads, then created heatmap visualizations to see feature interaction patterns. This analysis allowed us to identify important features and feature relationships, compare patterns across models and datasets, and check alignment with SHAP results and transportation theory.

4. Data Analysis of West Midlands Travel Dataset

This section examines the West Midlands dataset after removing respondents who always chose the same travel mode (lexicographic behavior). The analysis covers 2,757 observations to understand factors influencing travel choices in this region.

4.1. Dataset Overview and Mode Distribution

The travel mode distribution in West Midlands exhibits a clear preference hierarchy (Figure 2). Private car represents the dominant choice, accounting for 33.4% of all trips. Public transport constitutes the second most popular option at 29.8%. Walking represents a substantial portion of journeys at 23.0%. Bicycle accounts for only 9.4% of trips, while micro-mobility options constitute the least utilized transportation mode at 4.5% of all journeys. This uneven distribution creates challenges for transport planners and presents the class imbalance problem for machine learning models.

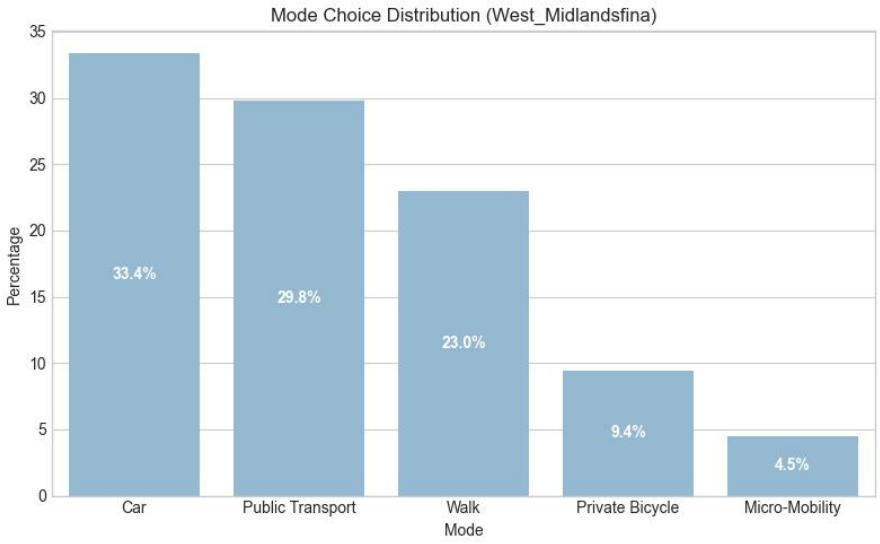


Figure 2 Mode Choice Distribution (West Midlands)

4.2. Factors Influencing Mode Choice

It should be noted that variables such as cost and risk are not stated by the respondents but rather represent values assigned during the survey design; hence, they are not of much significance for direct comparison across modes. Nevertheless, analyzing them based on the chosen choice (e.g., examining patterns when specific modes are selected) provides valuable insights into revealed travel behavior.

	Car			Public Transport			Micro-Mobility			Bicycle			Walking		
Variable	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Travel Cost (units)	4.99	2.87	6.42	2.54	1.40	3.21	2.96	1.46	4.78	0.18	0.08	0.36	-	-	-
Waiting Time (units)	5.00	3.50	6.50	11.43	7.00	16.90	4.81	3.50	6.50	2.91	2.10	3.90	-	-	-
Risk Score	2.55	1.80	3.40	-	-	-	25.87	19.70	36.70	29.37	22.90	42.50	9.32	6.90	12.80
Time Per Mile (min)	2.59	1.75	3.35	4.22	2.71	5.42	6.41	4.08	7.97	6.78	4.67	10.06	22.32	17.39	33.58

Table 1 Summary Statistics of Key Travel Variables by Mode

4.2.1. Cost and Time Relationships

The analysis reveals significant differences in travel costs across transportation modes (Table 1). Car travel exhibits the highest average cost at 4.99 units, approximately twice as expensive as public transport at 2.54 units. Micro-mobility travel averages 2.96 units in cost. Bicycle represents the most economical option at only 0.18 units on average.

Access, waiting, and exit times (AWET) demonstrate substantial variation across modes (Table 1). Public transport exhibits the longest non-travel time component at 11.43 units on average. This duration exceeds more than double the waiting time for cars at 5.00 units or micro-mobility options at 4.81 units. Bicycle demonstrates the shortest waiting time (2.91 units), which theoretically makes it attractive for time-sensitive travelers.

Travel speeds follow expected patterns (Table 1). Car travel achieves the fastest pace at 2.59 minutes per mile, followed by public transport at 4.22 minutes per mile. Micro-mobility and bicycle demonstrate similar speeds at 6.41 and 6.78 minutes per mile, respectively, while walking operates at a significantly slower pace of 22.32 minutes per mile.

4.2.2. Risk Perception Across Modes

Risk perception differs dramatically across transport modes and appears to strongly influence travel choices (Table 1). Bicycles are perceived as the most dangerous option, with a risk score of 29.37, followed closely by micro-mobility travel at 25.87. Walking demonstrates a moderate risk score of 9.32, while car travel is perceived as extremely safe with a score of only 2.55, approximately ten times safer than bicycle travel.

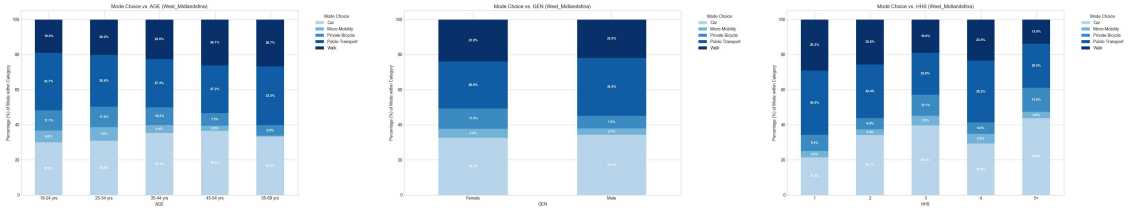
This substantial difference in safety perception likely explains why bicycles remain underutilized despite their advantages in cost and waiting time. The data suggests that many travelers are unwilling to trade perceived safety for cost savings. The risk hierarchy shown in Table 1 demonstrates clear patterns where active transport modes (bicycle and micro-mobility) are perceived as significantly more dangerous than motorized transport (cars), with walking occupying an intermediate position in the risk spectrum.

4.2.3. Demographic Influences

The dataset contains diverse demographic characteristics. The age distribution demonstrates reasonable balance across categories, with slightly higher representation in middle

age groups. The 25-34 years cohort represents 22.4% of respondents, followed by the 35-44 years group at 21.9%. Gender distribution is nearly balanced, with 49.0% female and 51.0% male participants. Household size exhibits variation, with 3-person households being most prevalent at 30.7%.

These demographic factors significantly influence travel mode choices, as illustrated in Figure 3. Distinct patterns emerge across age groups and gender categories.



(a) Mode Choice vs. Age (b) Mode Choice vs. Gender (c) Mode Choice vs. Household Size

Figure 3 Travel Mode Choice by Demographics

4.3. Summary: Patterns Supporting Model Analysis

The descriptive analysis reveals several empirical patterns from the West Midlands dataset that inform subsequent modeling approaches:

1. **Clear Mode Hierarchy with Class Imbalance**: Car travel dominates usage (33.4%), followed by public transport (29.8%), walking (23.0%), bicycle (9.4%), and micro-mobility (4.5%). This distribution creates the imbalanced dataset challenge for machine learning model implementation.

2. **Cost-Choice Relationships**: Travel costs exhibit significant variation across modes, with cars demonstrating the highest costs (4.99 units) and bicycles the lowest costs (0.18 units). Despite substantial cost differentials, expensive modes maintain popularity, suggesting factors beyond simple travel cost minimization influence choice behavior.

3. **Risk Perception as Mode Differentiator**: Risk scores vary dramatically across modes (bicycle: 29.37 vs car: 2.55), indicating that safety perceptions may override economic considerations in mode choice decisions.

These findings demonstrate that travelers face complex trade-offs where no single factor dominates choice patterns. Cars are expensive but fast and safe; bicycles are economical but perceived as risky; and public transport involves extended waiting times. These complex relationships suggest potential for advanced models to capture non-linear patterns, though the subsequent modeling results indicate significant challenges in learning these relationships, particularly with limited data.

5. Results

This chapter presents the findings from our experiments with minimal interpretation. The results are organized into four main sections: predictive performance, class imbalance handling using SMOTE, SHAP interpretability analysis, and attention mechanism visualization. Detailed interpretation follows in Chapter 5.

5.1. Predictive Performance

5.1.1. WM Dataset

Cross-Validation Results

Model	Accuracy (%)	Balanced Accuracy (%)	Macro / Weighted F1	Macro Prec. / Rec.
Random Forest	40.69 ± 1.38	25.28 ± 0.66	0.23 / 0.35	0.26 / 0.26
XGBoost	40.06 ± 1.04	26.85 ± 0.60	0.26 / 0.36	0.23 / 0.24
Decision Tree	38.73 ± 2.34	24.60 ± 0.83	0.23 / 0.35	0.32 / 0.27
TabTransformer	39.38 ± 2.18	23.65 ± 1.46	0.21 / 0.33	0.27 / 0.22
FT-Transformer	41.03 ± 2.13	23.35 ± 1.68	0.19 / 0.31	0.32 / 0.23

Table 2 Cross-validation performance metrics on the WM dataset

We evaluated five models on the small West Midlands dataset. Table 2 presents the cross-validation performance metrics. FT-Transformer achieved the highest accuracy at 41.03%, followed by Random Forest (40.69%) and XGBoost (40.06%). The accuracy figures may hide important performance problems.

Balanced accuracy, which gives equal weight to all travel modes regardless of their frequency, revealed substantially lower performance levels. All models scored below 27% on balanced accuracy, with XGB ranking highest at 26.85% while FT-Transformer ranked lowest at 23.35%. This large gap approximately 15-20 percentage points between regular accuracy and balanced accuracy indicates that models mainly succeeded at predicting the most common travel mode (Car) but failed with less common modes. The low Macro F1 scores, ranging from 0.19 to 0.26, further confirm this issue. Macro F1 measures both precision and recall across all classes, treating all modes equally regardless of their frequency. These scores demonstrate poor performance across all travel modes, particularly the less common ones that often hold significance for transportation planning.

Model	Accuracy (%)	Balanced Accuracy (%)	Macro / Weighted F1	Macro Prec. / Rec.
Random Forest	44.93	25.62	0.24 / 0.40	0.26 / 0.26
XGBoost	42.39	23.55	0.22 / 0.38	0.23 / 0.24
Decision Tree	45.29	27.47	0.28 / 0.42	0.32 / 0.27
TabTransformer	38.04	22.01	0.22 / 0.35	0.27 / 0.22
FT-Transformer	42.39	23.43	0.22 / 0.36	0.32 / 0.23

Table 3 Test set performance metrics on the WM dataset

he test set results revealed significant performance limitations for all models. With balanced accuracy remaining below 27% across all approaches, the models demonstrated fundamental difficulties in learning patterns for minority transportation modes. Decision Tree performed best with 45.29% accuracy, followed closely by Random Forest with 44.93%. TabTransformer performed worst with only 38.04% accuracy. This ranking change between cross-validation and test set results suggests model performance instability on this small dataset performance varied depending on the specific data split.

Balanced accuracy scores on the test set remained consistently low for all models, staying below 30%. Decision Tree achieved the highest balanced accuracy at 27.47%, while TabTransformer scored lowest at 22.01%. This pattern confirms that all models struggled to correctly predict less common travel modes.

Model	AUROC
Random Forest	0.63
XGBoost	0.64
Decision Tree	0.60
TabTransformer	0.65
FT-Transformer	0.65

Table 4 AUROC scores by mode on the WM dataset (Test)

AUROC scores in Table 4 provide different perspective on model discrimination ability. AUROC measures a model's ability to distinguish between classes, with values ranging from 0.5 (random guessing) to 1.0 (perfect prediction). Values above 0.7 are generally considered acceptable. TabTransformer and FT-Transformer tied for highest with scores of 0.65, followed by

XGB (0.64) and RF (0.63). DT scored lowest at 0.60. These AUROC values, all below 0.70, further confirm the models' limited ability to distinguish between different travel modes.

	Predicted				
	Car	**PT**	**MM**	**Bike**	**Walk**
Car	90	14	0	2	17
PT	36	16	1	2	13
MM	3	0	0	0	1
Bike	17	4	0	3	2
Walk	30	8	0	1	16

Table 5 Decision Tree

	Predicted				
	Car	**PT**	**MM**	**Bike**	**Walk**
Car	92	10	0	1	20
PT	42	14	0	0	12
MM	3	0	0	0	1
Bike	22	3	0	0	1
Walk	32	4	0	1	18

Table 6 XGBoost

	Predicted				
	Car	**PT**	**MM**	**Bike**	**Walk**
Car	80	20	0	0	23
PT	42	13	2	1	10
MM	1	1	0	0	2
Bike	15	6	0	2	3
Walk	27	16	0	2	10

Table 7 TabTransformer

	Predicted				
	Car	**PT**	**MM**	**Bike**	**Walk**
Car	94	9	0	1	19
PT	49	9	0	0	10
MM	2	0	0	0	2
Bike	21	1	0	1	3
Walk	36	6	0	0	13

Table 8 FT-Transformer

The confusion matrices presented in Tables 5-8 clearly illustrate the specific failure patterns. In these matrices, rows show actual travel modes, columns show predicted modes, and the diagonal elements represent correct predictions. The confusion matrices revealed that models achieved most of their accuracy by defaulting to the majority class (Car). Most non-car trips were incorrectly classified as car trips, indicating the models failed to learn distinguishing patterns for alternative transportation modes.

For example, as shown in Table 6, XGB incorrectly labeled 42 'Public Transport' trips, 22 'Bike' trips, and 32 'Walk' trips as 'Car'. FT-Transformer displayed similar errors in Table 8, misclassifying 49 'Public Transport', 21 'Bike', and 36 'Walk' trips as 'Car'. The matrices also

revealed that most models completely failed to identify any 'Bike' and 'Micro Mobility' trips, often showing zero correct predictions for these categories. Correct predictions for 'Public Transport' and 'Walk' also remained very low, typically only 10-16 correct trips.

In summary, experiments on the West Midlands dataset demonstrated that while models achieved 40-45% overall accuracy, they performed poorly when predicting less common travel modes. The substantial gap between standard accuracy and balanced accuracy ranging from 15-20 percentage points indicates the models primarily learned to predict the majority class 'Car' while failing to correctly identify classes like 'Bike', 'Walk', and 'Public Transport'. The confusion matrices confirmed this pattern most non-car trips were incorrectly classified as 'Car'. Despite achieving moderate standard accuracy scores, no model provided reliable predictions across all travel modes when trained on this small, imbalanced West Midlands dataset.

5.1.2. Imbalance Handling

To address the class imbalance issue identified in the previous performance analysis, we applied two different techniques: SMOTE and Focal Loss.

SMOTE Results

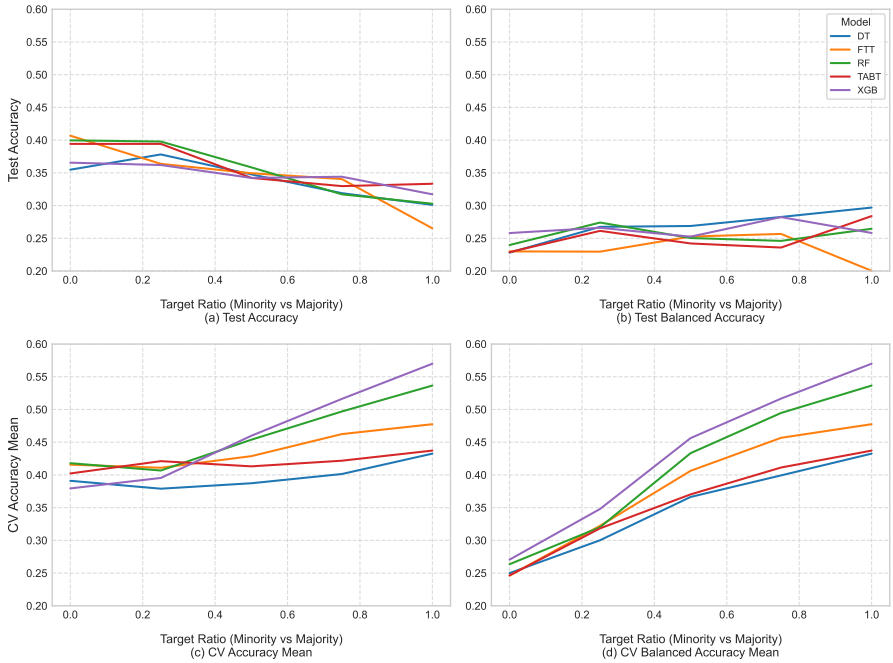


Figure 4 Impact of SMOTE Oversampling Ratio on Model Performance Metrics for the West Midlands Dataset. (a) CV Accuracy, (b) CV Balanced Accuracy, (c) Test Accuracy, and (d) Test Balanced Accuracy.

The impact of applying SMOTE oversampling to the WM training data was investigated by varying the target ratio of minority to majority classes from 0.00 (original data) to 1.00 (fully balanced). Figure 4 illustrates the trends for CV (cross-validation) and test set performance

across four metrics: CV accuracy, CV balanced accuracy, test accuracy, and test balanced accuracy.

During 5-fold CV, increasing the SMOTE ratio consistently improved the apparent performance for most models, boosting both mean CV accuracy and, more significantly, mean CV balanced accuracy. This suggests models learned effectively from the synthetically balanced data within the training/validation splits. When evaluating these SMOTE-trained models on the held-out test set, which represents the original distribution, revealed a negative effect on generalization. Test set accuracy decreased as the SMOTE ratio increased for most models, indicating overfitting to the synthetic training data. While test set balanced accuracy showed a slight improvement for several models at low SMOTE ratios (e.g., 0.25), performance typically declined significantly as the training data approached full balance (Ratio=1.00). This divergence between inflated CV metrics and degraded test set performance demonstrates that extensive SMOTE oversampling hindered the models' ability to generalize to the true data distribution on this dataset.

Focal Loss Results

As an alternative to data-level modifications, we also tested Focal Loss, which modifies the loss function to focus more on difficult-to-classify examples. Table 9 presents the results of applying Focal Loss to our models on the West Midlands dataset. Focal Loss showed some improvement in balanced accuracy for certain models. Most notably, XGB achieved 31.8% balanced accuracy with Focal Loss compared to 23.6% with standard training. TabTransformer similarly improved from 22.0% to 28.4%. However, an analysis of the SHAP values for models trained with Focal Loss revealed more scattered patterns with less clear relationships between features and predictions, suggesting a potential trade-off between balanced accuracy and behavioral interpretability.

Model	CV Acc. (%)	CV Bal. Acc. (%)	Test Acc. (%)	Test Bal. Acc. (%)
RF	44.5 ± 1.4	27.0 ± 1.2	43.2	26.5
DT	40.3 ± 1.9	26.2 ± 1.7	41.2	27.2
XGB	43.9 ± 0.9	28.7 ± 1.0	46.4	31.8
TABT	36.5 ± 2.1	24.6 ± 2.7	44.1	28.4
FTT	39.5 ± 0.6	26.9 ± 2.7	40.0	26.7

Note: CV metrics reported as Mean $\pm$ Standard Deviation. Accuracy metrics are percentages.

Table 9 Model Performance Metrics Summary

5.1.3. LPMC Dataset

This section presents the results of model evaluation on the LPMC dataset, with approximately 80,000 observations. We benchmarked our TabTransformer and FT-Transformer models against established machine learning algorithms and a MNL model, with comparative results sourced from Martín-Baos et al. [1]. Table 10 shows the test set accuracy and GMPCA

for all evaluated models.

Model	Test Accuracy (%)	Test GMPCA (%)
MNL	72.54	48.85
SVM	74.20	50.89
Random Forest	73.59	50.16
XGBoost	74.72	51.85
Neural Network (NN)	74.25	51.03
DNN	74.05	51.16
TabTransformer	72.80	48.93
FT-Transformer	74.44	51.40

Table 10 Test set Accuracy and GMPCA comparison on the LMPC dataset. Results for MNL, SVM, RF, XGB, NN, and DNN are sourced from Martín-Baos et al. [1]. TabTransformer and FT-Transformer results are from this study.

The FT-Transformer achieved 74.44% accuracy, a marginal improvement of 1.9 percentage points over the MNL baseline (72.54%) and only 0.28 percentage points below the best-performing XGBoost model. For the GMPCA metric, the FT-Transformer also ranks second at 51.40%, below only XGB at 51.85%. The TabTransformer model achived 72.80% accuracy and 48.93% GMPCA. These values position it seventh in accuracy and seventh in GMPCA, slightly above the MNL model, which achieved 72.54% accuracy and 48.85% GMPCA.

The difference between the highest-performing model XGB and the lowest-performing model MNL was 2.18 percentage points for accuracy and 3.00 percentage points for GMPCA. Among the five best-performing models (XGB, FT-Transformer, NN, SVM, and DNN), the accuracy values ranged from 74.05% to 74.72%, representing a difference of 0.67 percentage points.

5.2. SHAP Analysis Results

This section presents the results from applying SHAP analysis to our FT-Transformer models. SHAP analysis helps us understand how features influence model predictions. However, it should be noted that high SHAP importance does not necessarily reflect true behavioral causality or real-world feature importance. The patterns observed should be validated against domain knowledge and behavioral measures.

We applied this technique to models trained on the WM and LPMC datasets to identify which

features strongly influence transport mode predictions. In SHAP plots, each dot represents an observation, with red dots showing high feature values and blue dots showing low values. Positive SHAP values (right side) indicate that the feature increases the probability of choosing a particular mode, while negative values (left side) decrease this probability.

5.2.1. WM Dataset

For the WM dataset, we analyzed SHAP distributions for Car and Public Transport modes using the FT-Transformer model. For Car mode choice (Figure 5), household income (HHINC_Numeric) emerges as the most influential feature. Higher-income individuals (red dots) show positive SHAP values, indicating increased car choice probability, while lower-income individuals (blue dots) have negative SHAP values. This aligns with economic expectations as higher income enables car ownership.

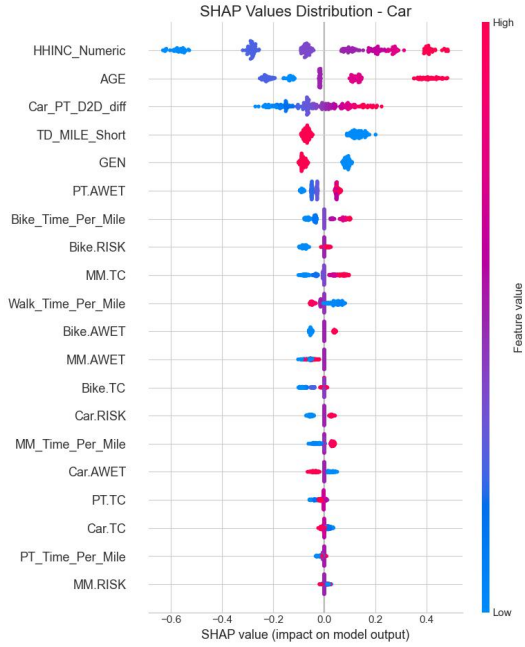


Figure 5 SHAP values for Car mode choice in WM dataset for FT-Transformer.

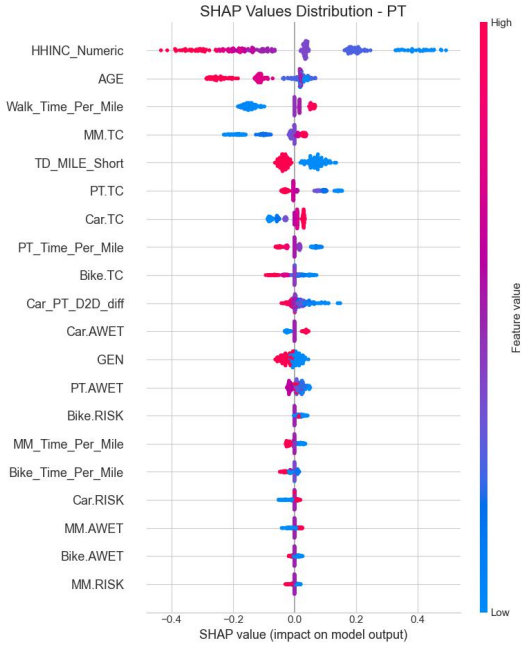


Figure 6 SHAP values for Public Transport mode choice in West Midlands dataset for FT-Transformer.

Age is the second most important feature, with older individuals (red dots) showing positive SHAP values and younger people (blue dots) showing negative values for car choice. The travel time difference between Public Transport and Car (Car_PT_D2D_diff, where positive values mean the Car is faster) ranks third in importance. When a Car offers a significant door-to-door time advantage over Public Transport (red dots, high positive feature values), it increases the probability of choosing a Car (positive SHAP values). Conversely, when Public Transport is faster or has a similar travel time to a Car (blue dots, low or negative feature values), the likelihood of choosing a Car decreases (negative SHAP values). TD_MILE_Short shows that high values (red dots) are associated with negative SHAP values, while low values (blue dots) show positive values. This suggests that shorter trips reduce car choice probability while longer trips increase it. Gender (GEN) displays a clear separation where females (blue dots) have positive SHAP values and males (red dots) have negative values.

Walk_Time_Per_Mile shows a counterintuitive pattern where high walk times (red dots) have negative SHAP values and low walk times (blue dots) have positive values. This counterintuitive behavior may indicate spurious correlations rather than meaningful relationships. Other features follow expected patterns: PT.AWET shows longer waiting times increase car attractiveness, cost variables (PT.TC, Car.TC) show expected economic relationships, and Bike.RISK shows high-risk values favoring car choice.

For Public Transport choice (Figure 6), household income again emerges as the most influential feature. Lower-income individuals (blue dots) have positive SHAP values for PT choice, while higher-income individuals (red dots) have negative values, reflecting PT's role as a more affordable transportation option. Age ranks second, with younger individuals (blue dots) showing slightly positive SHAP values and older people (red dots) showing negative values for PT choice. Walk_Time_Per_Mile appears third, with longer walking times (red dots) encouraging PT use. Cost relationships follow expected patterns: Cheaper PT.TC is encouraging its use, while Car.TC shows expensive car travel, making PT more attractive. Travel time variables (PT_Time_Per_Mile, Car_PT_D2D_diff) display logical relationships where faster PT service and competitive door-to-door times favor PT choice.

Variable	**Coefficient**	**Std. Error**
PT	0.717***	(0.189)
TC	-0.049**	(0.022)
I(IVT/60)	-0.887***	(0.156)
I(AWET/60/TD)	-4.148***	(0.955)
RISK	-0.012**	(0.005)
I(TD $\times$ (Walk | PT))	-0.234***	(0.048)
I(AGEMP $\times$ PT)	-0.013***	(0.003)
I(AGEMP $\times$ MM)	-0.047***	(0.004)
I(AGEMP $\times$ Bike)	-0.029***	(0.003)
I(GEN_F $\times$ (MM | Bike))	-0.455***	(0.111)
Observations: 2,867		
Log-Likelihood: -3,867.898		
Significance codes: *** p<0.01, ** p<0.05, * p<0.1		

Table 11 Multinomial Logit Model Coefficients for West Midlands Dataset(PHOEBE project)

Table 11 presents coefficients from an MNL model estimated on the same WM dataset for reference. The SHAP analysis reveals consistencies and differences compared to the linear

MNL model (Table 11). The MNL model shows significant negative coefficients for travel time components, with waiting time (AWET/60/TD = -4.148) having a much larger impact than in-vehicle time (IVT/60 = -0.887). This is consistent with the SHAP results, where PT.AWET shows that more prolonged PT waiting times increase car choice probability. The MNL travel cost coefficient (TC = -0.049) is negative, indicating that higher costs reduce mode utility. The SHAP results show similar patterns to those of Car.TC has adverse effects on car choice and positive effects on PT choice.TC has adverse effects on PT choice when costs are high.

The MNL shows significant negative age interactions for PT (-0.013), MM (-0.047), and Bike (-0.029) modes. The SHAP results show that older people have negative SHAP values for PT choice, which aligns with the MNL negative age coefficient for PT. Older people show positive SHAP values for car choice, which is consistent if the Car is the reference category in the MNL model. However, the direct comparison is challenging because household income (HHINC_Numeric), the most important feature in both SHAP plots, is not included in the MNL model specification. This suggests that the FT-Transformer may be capturing important socioeconomic effects that the MNL model does not account for.

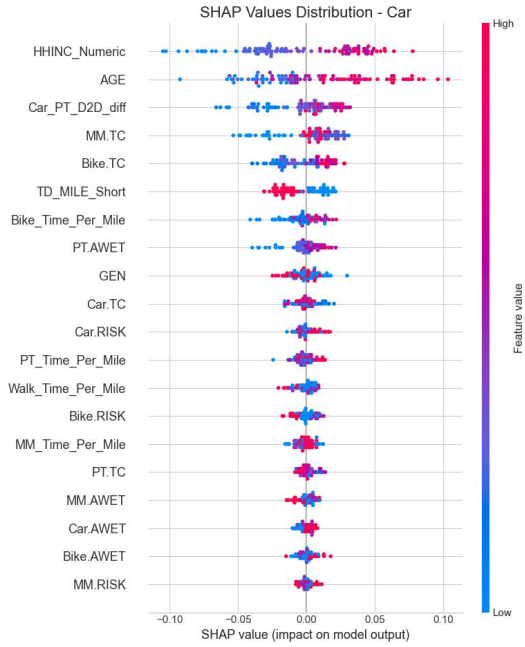


Figure 7 SHAP values for Car from RF model for West Midlands.

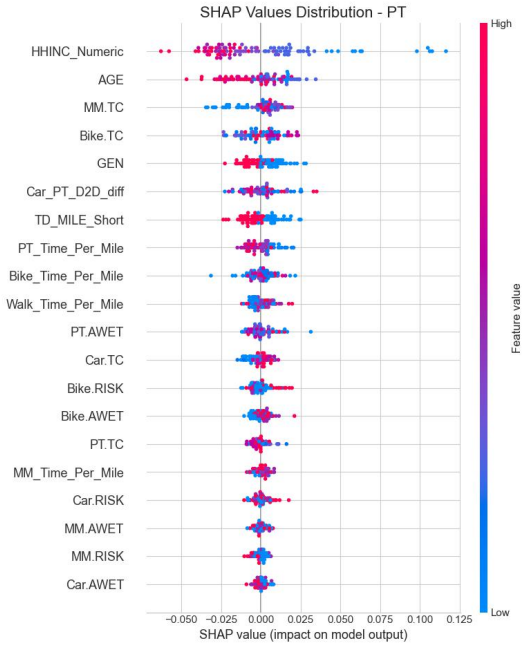


Figure 8 SHAP values for Public Transport from RF model for West Midlands.

Figures 7 and 8 show SHAP distributions from an RF model trained on the same dataset. The Random Forest model shows similar feature importance rankings compared to the FT-Transformer. HHINC_Numeric emerges as the most important feature for Car and PT choice, followed by age. However, the RF model displays more scattered SHAP value distributions, with points spread across both positive and negative ranges, making clear behavioral interpretation more challenging than the FT-Transformer patterns.

5.2.2. LPMC Dataset

Figure 9 and Figure 10 present the SHAP distributions for the FT-transformer for both the Car and Public Transport modes. For Car choice (Figure 9), driving duration (dur_driving) emerges as the most influential feature. Shorter driving times (blue dots) consistently appear with positive SHAP values, indicating an increased probability of car choice when driving faster. Longer times (red dots) show negative values, decreasing Car choice probability. Car ownership and driving license rank as the second and third most important features, with availability (red dots) strongly associated with positive SHAP values for Car choice, as would be expected.

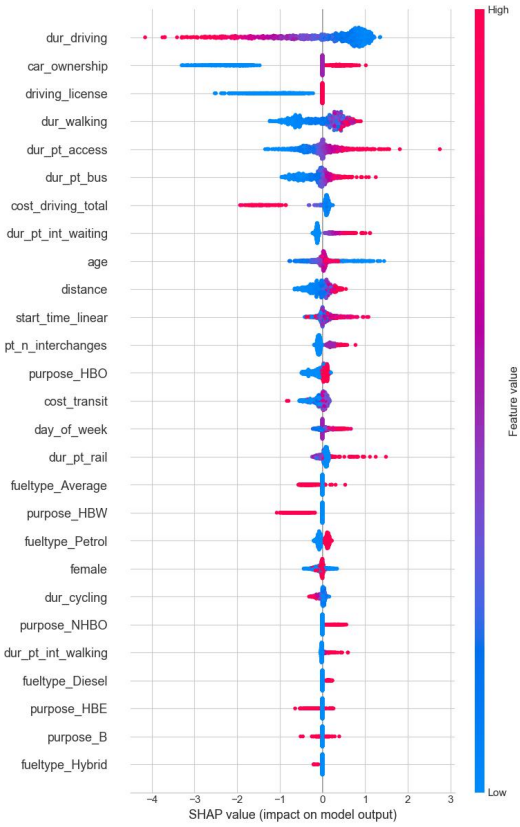


Figure 9 SHAP values for Car mode choice from FTtransformer model for LPMC.

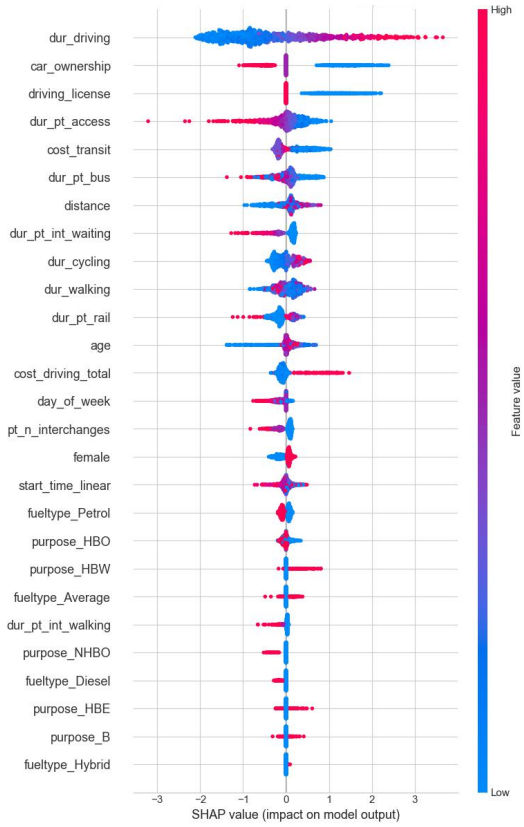


Figure 10 SHAP values for Public Transport mode choice from FTtransformer model for LPMC.

Public transport service quality variables appear in the middle ranks. Longer PT access times (dur_pt_access), bus times (dur_pt_bus), and waiting times generally align with positive SHAP values for Car choice, logically showing that poorer PT service increases Car use. Distance shows a complex distribution with points spread across the SHAP value range, suggesting non-linear effects. Demographic variables like age and gender are less strongly influenced in LPMC than in the WM dataset.

For Public Transport choice (Figure 10), driving duration again appears as the most important feature, but with inverse effects compared to Car choice. Longer driving times (red dots) align with positive SHAP values, indicating increased PT choice probability when driving takes

longer. Car ownership and driving license availability show substantial negative impacts on PT choice when present, ranking as the second and third most important features. Car cost (cost_driving_total) also ranks among the top features, with higher costs (red dots) corresponding to positive SHAP values for PT, showing the expected substitution effect.

PT service variables show expected relationships with some complexity. PT access time (dur_pt_access) displays negative SHAP values for longer times, as expected since longer access times discourage PT use. PT interchange waiting time (dur_pt_int_waiting) shows negative SHAP values for longer waits. However, PT bus duration (dur_pt_bus) shows a more complex distribution with positive and negative SHAP values across different journey times. Distance demonstrates a pattern where longer journeys generally align with higher PT choice probability, suggesting PT may be preferred for longer trips in LPMC.

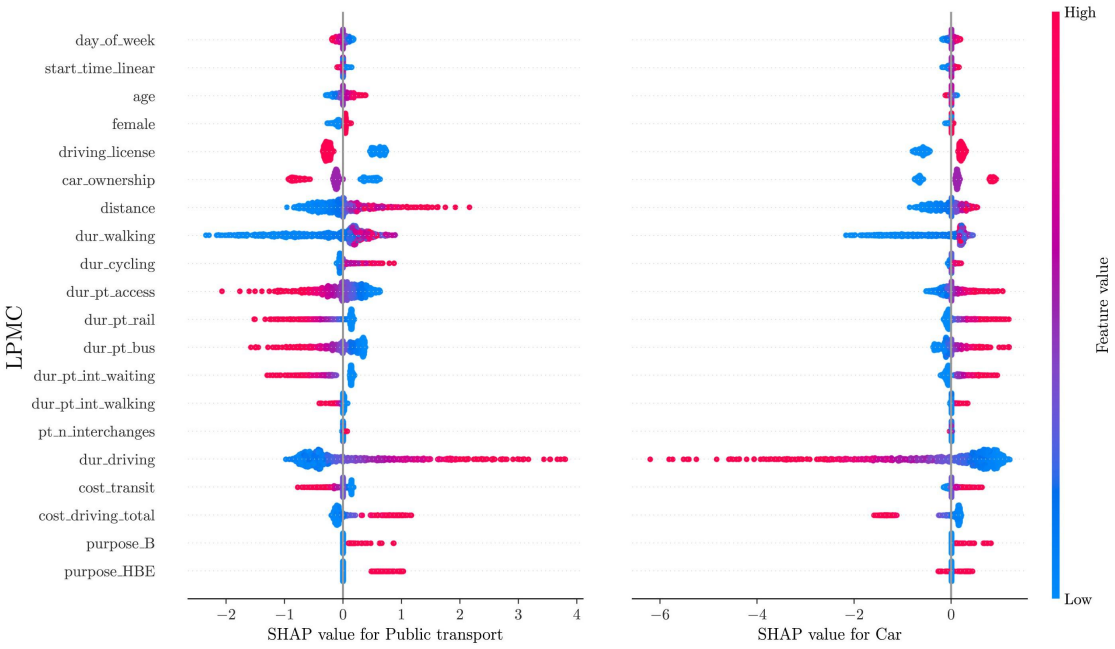


Figure 11 MNL shap Values for LPMC dataset [1]

Comparing our FT-Transformer SHAP results with the MNL SHAP values from Martín-Baos et al. (Figure 11) reveals similarities and important differences. Both models show driving duration (dur_driving) as having substantial influence, with wide SHAP value ranges indicating a high impact on predictions. The directional effects also align with behavioral expectations - shorter driving times favor car choice while longer times favor public transport.

However, notable differences emerge in the magnitude of the influence of different features. The FT-Transformer shows car ownership and driving license availability as highly influential features (ranking second and third), while in the MNL model, these variables display smaller SHAP value ranges, suggesting a more limited influence on predictions. This indicates that the FT-Transformer captures stronger interactions between car availability and travel deci-

sions compared to the linear utility specification of the MNL model.

Comparative Analysis
Comparing SHAP patterns across the WM and LPMC reveals similarities and context-specific differences. Both datasets show that longer travel times for a mode generally decrease its selection probability while increasing the alternatives. However, feature importance rankings differ between contexts. Household income ranks highest for Car choice in the West Midlands (Figure 5), while driving duration dominates in LPMC (Figure 9).

Demographic variables show a more substantial influence in the WM data, with household income and age being top-ranked features. In LPMC, mode-specific variables like driving duration and costs appear more dominant. Risk perception factors display consistent directional effects across both datasets, with higher perceived risk for a mode decreasing its selection probability.

For both datasets, the SHAP distributions reveal more complex patterns than simple linear relationships, particularly for variables like distance and age. However, some relationships appear counterintuitive (such as the walking time patterns in the WM dataset), highlighting the need for careful validation of SHAP-derived insights against transportation theory.

5.3. Attention Mechanism Analysis

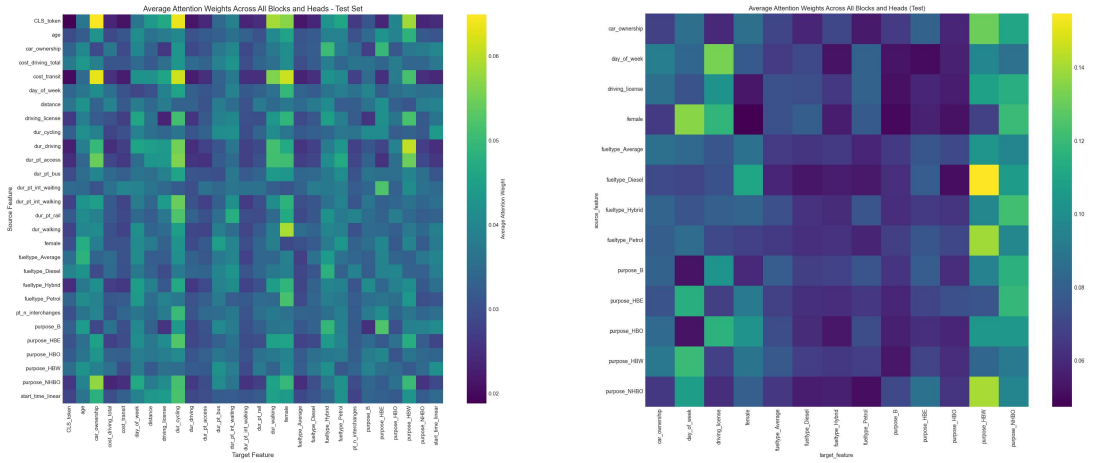
The attention mechanisms in transformer models provide insights into how these models process the data. However, attention weights should be interpreted cautiously as they represent internal model processing rather than direct feature importance [40].

Attention mechanisms show how transformer models connect different features, with higher attention values (brighter colors in heatmaps) indicating the model considers those features together when making predictions. The CLS token is a special element of FT-Transformer that collects information from all features for the final classification, with features receiving high attention from this token being particularly important for mode choice prediction.

5.3.1. LPMC Dataset Attention Patterns

For the FT-Transformer (Figure 12a), the attention weights generally range between 0.02-0.06, indicating a distributed pattern of feature interactions. The CLS_token row shows bright spots when attending to various features, suggesting which features are most important for classification decisions.

In the FT-Transformer attention matrix, we observe distinct clusters of related features. Transit-related features form interconnected groups with moderately high mutual attention, suggesting the model recognizes these features as conceptually related and processes them together. Similarly, driving-related features show higher mutual attention, indicating that the



(a) FT-Transformer

(b) TabTransformer

Figure 12 Average attention weights across all blocks and heads for LPMCdataset

model captures mode-specific feature relationships.

For the TabTransformer model (Figure 12b), the attention distribution shows stronger feature relationships with weights ranging between 0.06-0.14. Notable patterns include connections between demographic features (female, day_of_week) and trip purpose variables (purpose_NHBO, purpose_HBW), indicating that these categorical variables are processed as related groups.

Feature Importance from CLS Token Attention

Figure 13 presents the ranking of features based on their importance as determined by attention from the CLS token in the FT-Transformer. Car ownership emerges as the most important feature, followed by cycling duration (dur_cycling) and walking duration (dur_walking). This pattern makes sense in transportation modeling, as car availability strongly influences mode choice decisions.

The prominence of travel time features in the top rankings highlights how the model prioritizes time-related factors in transportation decisions. Distance appears in the middle of the ranking, suggesting it plays an important but not dominant role compared to car ownership and travel times. Demographic factors like female and driving_license rank highly, showing that the model recognizes that personal characteristics significantly influence travel mode choices. Trip purpose categories (purpose_HBW, purpose_HBO) also receive considerable attention, emphasizing how trip context affects transportation decisions. The relatively lower position of cost factors (cost_transit, cost_driving_total) than time-related factors suggests the model gives more weight to time considerations than monetary costs in mode choice prediction. This aligns with transportation theory, which states that time is often more critical in daily travel decisions than the cost for many travelers.

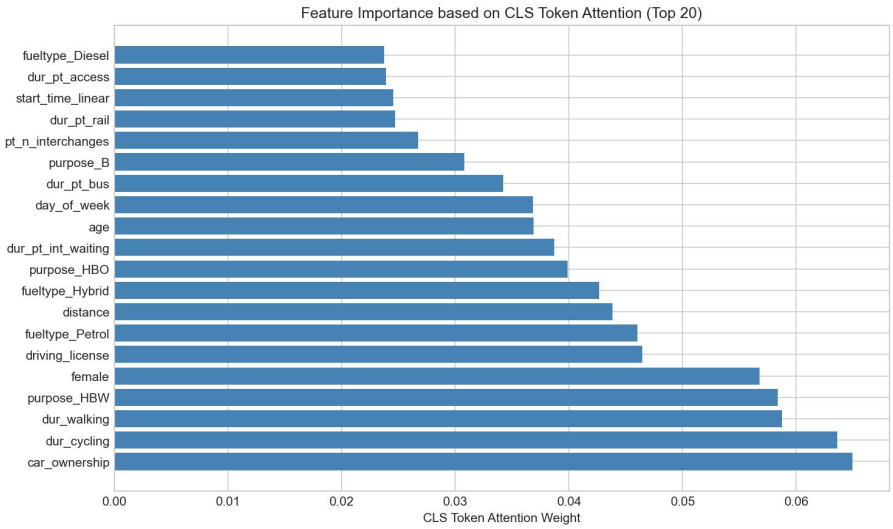
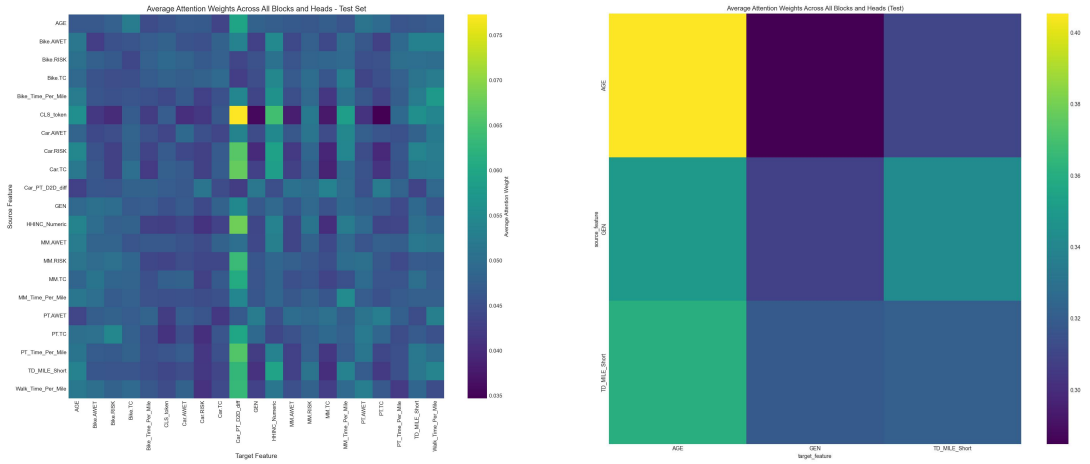


Figure 13 Feature importance based on CLS token attention for FT-Transformer (LPMC dataset)

5.3.2. West Midlands Dataset Attention Patterns

For the WM dataset, the attention patterns show clear differences between the two transformer models compared to the LPMC dataset.



(a) FT-Transformer

(b) TabTransformer

Figure 14 Average attention weights across all blocks and heads for West Midlands dataset

The TabTransformer attention pattern (Figure 14b) shows a simple structure with only three features as it processes only categorical features through transformer architecture: AGE, GEN (gender), and TD_MILE_Short. The attention weights range from approximately 0.28 to 0.40. The brightest areas indicate the strongest connections, with AGE showing strong self-attention and connections to other features.

The FT-Transformer attention matrix (Figure 14a) displays a much more comprehensive feature set as it processes all the variables through a transformer. The attention weights range from about 0.035 to 0.075, showing a much narrower and lower range than TabTransformer. The CLS_token row shows brighter spots when connecting to GEN and several other fea-

tures, but the overall pattern appears more distributed across the feature space. Most attention weights fall in the middle range, suggesting less confident feature relationships compared to the LPMC dataset.

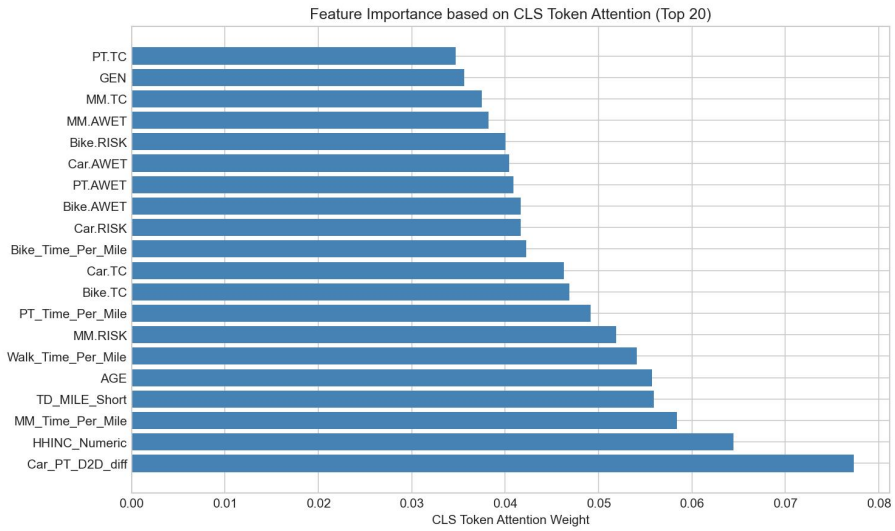


Figure 15 Feature importance based on CLS token attention for FT-Transformer (West Midlands dataset)

Figure 15 shows the feature importance derived from CLS token attention for the FT-Transformer on the West Midlands dataset. The ranking shows the top 20 features, with Car_PT_D2D_-diff receiving the highest attention weight, followed by HHINC_Numeric, MM_Time_Per_Mile, and TD_MILE_Short. Demographic features like AGE appear fifth in the ranking, and GEN appears in the lower portions, receiving less attention than transportation-specific variables. Mode-specific attributes such as Walk_Time_Per_Mile, MM.RISK, PT_Time_Per_Mile, and various cost and waiting time features appear throughout the ranking, though with less clear hierarchical organization than observed in the LPMC data.

Comparing feature type processing between datasets, we note that the LPMC models show balanced attention across transportation attributes (time, cost, mode-specific factors) and personal/trip attributes (purpose, demographics). In contrast, the West Midlands models show disproportionate attention to personal/trip attributes, particularly in the TabTransformer. The clustering of conceptually related features is much more evident in the LPMC dataset for both models. This logical grouping indicates that transformer models can autonomously discover natural feature relationships that align with transportation domain knowledge with sufficient data. This difference highlights the data requirements for transformer models to develop meaningful attention mechanisms. With sufficient data, as in the LPMC case, attention patterns reflect established transportation relationships. With limited data, the attention becomes less structured and potentially less reliable for behavioral interpretation.

6. Discussion

6.1. Discussions

This thesis evaluated tabular Transformer models (TabTransformer and FT-Transformer) for travel mode choice prediction by comparing these models with other machine learning methods and discrete choice models (DCMs). The study aimed to determine whether these deep learning models could combine high prediction accuracy with good behavioral interpretability. We used two datasets: WM and LPMC.

6.1.1. Predictive Performance and the Role of Data Size

The results show that dataset size strongly affects Transformer model performance. This matches findings from other transportation studies [26].

All models performed poorly on the WM dataset when predicting less common transportation modes, probably because of the smaller datasets. While FT-Transformer achieved reasonable overall accuracy, all models showed low balanced accuracy and Macro F1 scores. Our attempts to fix class imbalance had mixed results. SMOTE oversampling led to overfitting, with better cross-validation scores but worse test performance, as shown in Figure 4. This finding matches research by [41]. Focal Loss improved balanced accuracy for XGBoost and TabTransformer but made interpretations less clear. These results show the challenges of using complex models with smaller datasets.

On the larger LPMC dataset, FT-Transformer showed strong predictive performance (74.44% accuracy, 51.40% GMPCA), ranking second only to XGBoost and performing better than the MNL baseline model and TabTransformer, as shown in Table 10 suggests that with enough data, FT-Transformer can predict almost as well as the best tree models. TabTransformer consistently performed worse than FT-Transformer on both datasets. This suggests that FT-Transformer's approach to processing numerical and categorical features works better [12].

6.1.2. Behavioral Consistency and Interpretability: SHAP Analysis

The objective was to determine if tabular Transformers could maintain behavioral consistency while making accurate predictions. SHAP analysis was employed for this evaluation, with findings compared to MNL results and the benchmark study by [1].

On the LPMC dataset, FT-Transformer's SHAP values matched economic theory and have similar rankings to the MNL model from the benchmark study. For example, driving duration was very important for car choice (shorter times increasing probability), and car ownership/license availability were key positive factors, as shown in Figure 9. Higher driving costs logically increase the probability of public transport choice. This suggests that, unlike tree-

based models in the benchmark study, the FT-Transformer can learn reasonable relationships from large travel datasets. It potentially offers a good balance between accuracy and interpretability similar to NN/SVM [1].

The SHAP analysis on the smaller WM dataset as seen in Figures 5 and 6. Age and Gender were more important compared to the LPMC dataset, and cost features sometimes acted as distance indicators. These patterns were less clear than on LPMC. Comparing FT-Transformer and Random Forests SHAP values on WM (Figures 7 and 8) it can suggest that the FT-Transformer exhibits marginally clearer patterns even when applied to smaller datasets, although both models encountered challenges in their performance on the WM. The cross-dataset comparison showed that while basic relationships (like negative time/cost effects) were captured consistently by FT-Transformer, the importance of features varied significantly. This reflects differences between the WM and LPMC dataset characteristics.

6.1.3. Interpretability: Attention Mechanism Analysis

The attention mechanisms of transformers offer valuable insights that extend beyond what is provided by SHAP. While SHAP focuses on explaining changes in output, attention weights illustrate the model's internal focus [42]. There are differing perspectives on the use of attention for interpretability; some researchers question its explanatory power [40], while others recognize the significance of attention patterns in understanding model behavior [43]. This study analyzes average attention weights and the CLS token's attention for feature ranking, aligning with approaches from original transformer studies [11, 12]. Both transformer models on the large LPMC dataset showed clear attention patterns, as seen in Figure 12. FT-Transformer showed logical grouping of related features (e.g., transit attributes, driving attributes). The CLS token focused on important features like transit cost, age, and interchange waiting duration, as shown in Figure 13. TabTransformer showed stronger, more concentrated attention patterns. Attention patterns on the smaller WM dataset were less clear, especially for FT-Transformer, as seen in Figure 14. TabTransformer showed strong pairwise attention, perhaps focusing on specific correlations in the limited categorical data. This further shows that dataset size affects how we interpret attention patterns reliably.

SHAP and attention analysis identified similar important features (e.g., travel times/costs, demographics). SHAP helps understand how each feature impacts the final choice, while attention shows how the model processes information and reveals feature interactions.

6.1.4. Implications and Limitations

Our findings suggest that FT-Transformer, with large data, could be helpful in travel behavior modeling. It may help solve the trade-off between accuracy and interpretability highlighted by [1]. It predicts almost as well as the best models while maintaining behavioral consistency comparable to NN/SVM models. However, large datasets are needed for it to work well. TabTransformer seems less effective for this specific task.

Both SHAP and attention analysis help interpret tabular Transformers. SHAP helps assess behavioral plausibility, while attention shows the model's internal processing. Both methods work better with larger datasets.

However, we have several limitations. We used only two datasets and specific models. As [28] shows, model performance can vary significantly depending on various factors, and there is uncertainty in model comparisons.

The interpretability analysis presented here has several limitations. First, SHAP values indicate feature influence on model predictions rather than authentic causal relationships in transportation behavior. Second, the analysis relies on visual inspection of plots rather than quantitative validation against established behavioral measures such as willingness-to-pay estimates. Third, some counterintuitive patterns suggest the models have learned spurious correlations rather than meaningful transportation relationships. Future research should include quantitative validation of SHAP-derived insights against transportation economic theory and empirical behavioral measures.

7. Conclusions

This study investigated tabular Transformer models (FT-Transformer and TabTransformer) for travel mode choice modeling by comparing these models with traditional machine learning approaches (Decision Trees, Random Forest, XGBoost) and the Multinomial Logit model on two datasets. It used SHAP and attention analysis for interpretability.

The study found that FT-Transformer could balance predictive accuracy and interpretability well, but only with large datasets. On the large LPMC dataset, FT-Transformer predicted almost as well as XGBoost. At the same time, SHAP analysis showed good behavioral consistency, similar to NN/SVM models in the benchmark study [1]. When trained on enough data, it handles the accuracy-interpretability trade-off better than tree models.

Dataset size was critical for transformer performance and interpretability. All models performed poorly on the smaller WM dataset, and attention patterns were less clear. TabTransformer consistently performed worse than FT-Transformer on both datasets.

SHAP analysis and attention mechanism analysis provide complementary insights. SHAP helps check behavioral consistency against economic theory, while attention analysis shows how the model processes feature internally. Both methods work better with larger datasets.

Standard class imbalance techniques like SMOTE and Focal Loss had limited success on the small, imbalanced WM dataset. SMOTE led to overfitting [41], while Focal Loss slightly improved balanced accuracy but made SHAP interpretations counterintuitive.

This research contributes:

- Systematic benchmarks for tabular Transformers on mode choice datasets
- Assessment of their behavioral consistency using SHAP analysis
- Attention analysis to understand the model behavior
- Analysis of SMOTE effectiveness and limitations for these models in the mode choice context.

In summary, while not perfect for all situations, FT-Transformer has potential, showing that with enough data, high predictive accuracy, and meaningful behavioral interpretability can coexist in complex machine learning models for travel behavior analysis.

7.1. Future Research Directions

This work has several limitations, suggesting future research directions. We only used two datasets and specific model configurations. Future studies should test these findings with more travel behavior datasets and newer transformer models.

A possible direction would be to integrate MNL models with transformer architectures. This could combine MNL's theoretical foundation with transformers' flexible pattern learning. Could build on previous hybrid work by [25], who developed a neural network with utility functions that maintained both good predictions and reasonable behavioral outputs.

Evaluating behavioral metrics (e.g., willingness-to-pay, etc.) for transformer models would be necessary, as Shap analysis alone will not guarantee behavior dependency.

Rule-based transformer architectures incorporating transportation choice theory could address the smaller dataset limitations. By adding domain knowledge directly into the attention mechanisms, such models might need less data to learn behaviorally consistent patterns. Specialized architectures like Transformer Choice Net designed specifically for choice prediction could be relevant for transportation applications[28]. This model can handle single, sequential, and multiple-choice scenarios, making it promising for complex travel behavior modeling.

Transfer learning could help overcome data size limitations. Pre-training transformer models on large transportation datasets before fine-tuning them on smaller regional datasets might maintain accuracy and behavioral consistency with limited data. This would help planning agencies with limited data collection resources.

Further research into handling class imbalance for transformer architectures is needed, given the limitations of SMOTE and Focal Loss observed in this study. Multi-model approaches could address class imbalance problems. Instead of data augmentation techniques like SMOTE, developing specialized models for minority transportation modes and combining them could improve prediction for less common but important modes.

Bibliography

[1] J. Á. Martín-Baos, J. A. López-Gómez, L. Rodriguez-Benitez, T. Hillel, and R. García-Ródenas, "A prediction and behavioural analysis of machine learning methods for modelling travel mode choice," *Transportation Research Part C: Emerging Technologies*, vol. 156, p. 104318, 2023.

[2] F. Wang and C. L. Ross, "Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 47, pp. 35–45, 2018.

[3] D. McFadden, "Conditional Logit analysis of qualitative choice behavior," 11 1972.

[4] M. E. Ben-Akiva and S. R. Lerman, *Discrete choice analysis: theory and application to travel demand*, vol. 9. MIT press, 1985.

[5] K. E. Train, *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press, 2009.

[6] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, vol. 78, pp. 273–282, 2017.

[7] X. Zhao, X. Yan, A. Yu, and P. van Hentenryck, "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models," *Travel Behaviour and Society*, vol. 20, pp. 22–35, 2020.

[8] C. R. Sekhar, Minal, and E. Madhu, "Mode choice analysis using random forrest decision trees," *Transportation Research Procedia*, vol. 17, pp. 644–652, 2016.

[9] Y. M. Aboutaleb, M. Danaf, Y. Xie, and M. Ben-Akiva, "Discrete choice analysis with machine learning capabilities."

[10] A. Ali, A. Kalatian, and C. F. Choudhury, "Comparing and contrasting choice model and machine learning techniques in the context of vehicle ownership decisions," *Transportation Research Part A: Policy and Practice*, vol. 173, p. 103727, 2023.

[11] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings."

[12] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, “Revisiting deep learning models for tabular data.”

[13] D. A. Hensher, “Hypothetical bias, choice experiments and willingness to pay,” *Transportation Research Part B: Methodological*, vol. 44, no. 6, pp. 735–752, 2010.

[14] K. A. Small, C. Winston, and J. Yan, “Uncovering the distribution of motorists’ preferences for travel time and reliability,” *Econometrica*, vol. 73, pp. 1367–1382, 6 2005.

[15] D. McFadden, “Econometric models for probabilistic choice among products,” *The Journal of Business*, vol. 53, p. S13, 1 1980.

[16] J. L. Walker and M. Ben-Akiva, “Advances in Discrete Choice: Mixture Models,” in *A Handbook of Transport Economics* (A. de Palma, R. Lindsey, E. Quinet, and R. Vickerman, eds.), Chapters, ch. 8, Edward Elgar Publishing, 2011.

[17] A. Vij and J. L. Walker, “How, when and why integrated choice and latent variable models are latently useful,” *Transportation Research Part B: Methodological*, vol. 90, pp. 192–217, 2016.

[18] S. Shaheen and A. Cohen, “Shared ride services in North America: definitions, impacts, and the future of pooling,” *Transport Reviews*, vol. 39, pp. 427–442, 7 2018.

[19] D. A. Hensher, M. J. Beck, and E. Wei, “Working from home and its implications for strategic transport modelling based on the early days of the COVID-19 pandemic,” *Transportation Research Part A Policy and Practice*, vol. 148, pp. 64–78, 4 2021.

[20] T. Hillel, M. Bierlaire, M. Z. Elshafie, and Y. Jin, “A systematic review of machine learning classification methodologies for modelling passenger mode choice,” *Journal of Choice Modelling*, vol. 38, p. 100221, 2021.

[21] P. Nijkamp, A. Reggiani, and T. Tritapepe, “Modelling inter-urban transport flows in Italy: A comparison between neural network analysis and logit analysis,” *Transportation Research Part C Emerging Technologies*, vol. 4, pp. 323–338, 12 1996.

[22] C. Xie, J. Lu, and E. Parkany, “Work travel mode choice modeling with data mining: Decision trees and neural networks,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1854, no. 1, pp. 50–61, 2003.

[23] Y. Zhang and Y. Xie, “Travel mode choice modeling with support vector machines,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2076, no. 1, pp. 141–150, 2008.

[24] X. Zhou, M. Wang, and D. Li, “Bike-sharing or taxi? modeling the choices of travel mode in chicago using machine learning,” *Journal of Transport Geography*, vol. 79, p. 102479, 2019.

[25] S. Wang, B. Mo, and J. Zhao, “Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions,” *Transportation Research Part C: Emerging Technologies*, vol. 112, pp. 234–251, 2020.

[26] M. Paredes, E. Hemberg, U.-M. O’Reilly, and C. Zegras, “Machine learning or discrete choice models for car ownership demand estimation and prediction?,” in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pp. 93–98, IEEE, 2017.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need.”

[28] H. Wang, X. Li, and K. Talluri, “Transformer choice net: A transformer neural network for choice prediction.”

[29] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[30] T. Hillel, M. Z. E. B. Elshafie, and Y. Jin, “Recreating passenger mode choice-sets for transport simulation: A case study of london, uk,” *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction*, vol. 171, no. 1, pp. 29–42, 2018.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. 10 2017.

[32] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, pp. 81–106, 3 1986.

[33] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[34] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[35] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.

[36] D. Nielsen, “Tree boosting with xgboost: Why does xgboost win "every" machine learning competition?,” Master’s thesis, Norwegian University of Science and Technology, Trondheim, Norway, Dec. 2016. Supervisor: Håvard Rue.

[37] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 8 2020.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[40] S. Jain and B. C. Wallace, “Attention is not Explanation,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 3543–3556, Association for Computational Linguistics, June 2019.

[41] C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, *Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics*, vol. 334 of *International Series in Operations Research & Management Science*. Cham: Springer International Publishing and Imprint Springer, 1st ed. 2022 ed., 2022.

[42] Z. Niu, G. Zhong, and H. Yu, “A review on the attention mechanism of deep learning,” *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[43] S. Wiegreffe and Y. Pinter, “Attention is not not explanation,” p. 11 – 20, 2019. Cited by: 527.

Declaration

I hereby confirm that the presented thesis work has been done independently and using only the sources and resources as are listed. This thesis has not previously been submitted elsewhere for purposes of assessment.

Place, 08th May 2025, Signature