**Literature Review**

Forecasting an individual's choice of transportation mode has garnered significant academic interest in recent years. The field of transportation and behavioural analysis has primarily relied on the extensive utilization of logit models in existing research. In general, logit models and logistic regression analysis have been employed, specifically to examine the connection between the likelihood of binary or ordinal responses and explanatory variables using the maximum likelihood estimation method (Trueck, 2009). According to Trueck (2009), the logistic function is given by the following expression:

A black and white math equation

Description automatically generated

The above can also be rewritten as:

A math equation with a white background

Description automatically generated

McFadden (1972, 1974) posed a significant challenge to the initial approach to comprehending travel behaviour. In his exploration of the analysis of transit behaviour, he advocated for the application of the multinomial logit model. Like binary logit models, the multinomial model adheres to the same principles and assumptions, with the notable difference being the presence of multiple alternative choices (Lee & Kim, 2023). Furthermore, because of the assumption of independence among the options, the total probabilities of all choices add up to 1 (Lee & Kim, 2023). Following the introduction of the logit model, subsequent research endeavours have sought to extend this model and investigate travel behaviour, thereby addressing certain constraints associated with the original logit model. Such examples are the nested logit model (Willis, 2014) and the mixed logit model proposed by McFadden & Train (2000). The nested logit model organizes choices into separate nests and permits varying correlations between these nests. As a result, correlations are consistent within each nest, but for options situated in different nests, the unobservable elements are uncorrelated and, in fact, entirely independent (Willis, 2014). The mixed logit model is defined as a standard multinomial model, also called “latent class model”, where the coefficients are chosen from a cumulative distribution, introducing an element of randomness (McFadden & Train, 2000).

**Factors influencing transport mode selection**

Since the appearance of logit models, numerous studies have surfaced with the objective of delving into the crucial factors that affect transportation mode selection, going beyond the conventional determinants like cost or travel distance. In a study conducted by Mayo and Taboada in 2020, they employed a hierarchy model to assess the factors influencing the choice of public transportation mode among respondents in the Philippines. Their survey results indicated that safety was the most significant consideration, followed by cost, comfort, and concerns about environmental sustainability (Mayo & Taboada, 2020). Another study by Donkor et al. in 2020 examined the role of emotions in transport mode selection, focusing on respondents in the city of Edinburgh. Their findings revealed that an individual's feelings and experiences related to public transportation, along with their socio-demographic characteristics, exerted a substantial influence on their transit behaviour (Donkor et al., 2020). Additionally, McCarthy et al. in 2017 conducted further research suggesting that the presence of young children in a family had an impact on transportation behaviour. Specifically, they proposed that families with children preferred car usage over other sustainable transit options, with psychosocial factors and household characteristics playing pivotal roles in the choice of transportation mode (McCarthy et al., 2017). The influence of various weather conditions in transport mode choice, have also been explored in the literature. Bocker et al. (2016) delved into weather-related factors and their connection to transit choices. Based on their analysis of travel diaries in the Netherlands, their results indicate that individuals who opt for walking or cycling modes are particularly affected by weather conditions (Bocker et al., 2016).

Additional research in the field of transportation focuses on mode selection during the COVID-19 pandemic and its impact on individual transit choices. To elaborate, Mussone & Changizi (2023) conducted a study that utilized a multinomial logistic regression model to investigate the factors influencing transportation mode choices prior, during, and post COVID-19 lockdowns. Their research, based on data from residents in Milan, Italy, indicated that socio-demographic factors, as well as individual preferences and concerns related to public transportation, played the most significant roles in predicting transport mode choices during the pandemic (Mussone & Changizi, 2023). It is worth noting that, despite mandatory contamination control measures, many residents in various countries expressed heightened concerns about the spread of the virus within public transportation during and after lockdown restrictions. Those concerns led residents to shift from relying on public transportation to utilizing private vehicles for their commuting requirements. The phenomenon is further investigated by Das et al. (2021), who employed a logistic regression model to examine travel behaviour and modal transitions. Their research findings indicate that demographic factors exert a considerable influence on preferences for switching transportation modes. Additionally, trip-related factors, including travel time and health conditions, demonstrate a robust association with the inclination to shift from public transportation to using cars (Das et al., 2021).

**Machine Learning essentials**

An alternative to the traditional logit models, comes with the introduction of machine learning.

According to Zhou (2021), Machine learning is a method that enhances the performance of systems through computational learning from prior experiences. In the realm of computer systems, these experiences are embodied in the form of data. The central objective of machine learning is to create learning algorithms capable of constructing models based on this data. When the learning algorithm is supplied with experiential data, it yields a model capable of making predictions for new observations (Zhou, 2021). Based on the presence or absence of labelled training data, learning problems can be categorized into two groups: supervised and unsupervised learning. Supervised learning encompasses a training phase in which the algorithm is supplied with a dataset comprising pairs of input and corresponding output (referred to as labelled data). During this phase, the algorithm acquires the ability to make predictions or classifications by drawing insights from this labelled data (Zhou, 2021). In supervised learning, the primary tasks are categorized into regression and classification, depending on whether the prediction output is continuous or discrete. In the case of classification problems, when there are only two possible labels, it is referred to as a "binary classification problem" (Zhou, 2021). If there are multiple possible labels, it is termed a "multiclassification problem" (Zhou, 2021). Common algorithms that aim to solve regression and classification problems include Naïve Bayes Classifier, K-Nearest Neighbours, Decision Trees, Ensemble Learning, Boosting, Support Vector Machines and Neural Networks (Kubat, 2021). In contrast, unsupervised learning is a category of machine learning in which the algorithm is given input data but does not have access to predefined output labels. In this context, the algorithm's role is to autonomously identify patterns, structures, or relationships within the data, all without prior knowledge of the expected output (Zhou, 2021). The main task in unsupervised learning is clustering which involves the process of categorizing data points by identifying their similarities (Zhou, 2021). The most common technique for such problems is K-Means Clustering.

There is also a third form of learning called “Reinforcement learning” (Kubat, 2021). In this domain, the objective differs significantly. Here, the agent's role is not to induce knowledge from a pre-classified dataset but to engage in active experimentation with a system. The system, in turn, provides feedback in the form of rewards or penalties in response to the agent's actions. The agent's primary aim is to refine its behaviour by seeking to maximize rewards and minimize penalties as it interacts with the system (Kubat, 2021).

**Machine Learning applications in transport mode choice**

The adoption of machine learning in the modelling and prediction of transport mode choices has experienced a significant surge in recent times. In contrast to the prior logit models, the results indicate a notably improved ability to accurately predict transportation mode. This application is geared towards improving our understanding and predictive capabilities concerning individuals' decisions regarding the modes of transportation they choose. Numerous recent studies have sought to employ a range of machine learning models to predict transit behaviour. In 2015, Omrani conducted a study with the objective of forecasting the travel mode choices of individuals by applying machine learning techniques to national data from Luxembourg. His research outcomes revealed that artificial neural networks outperformed other alternative models in terms of predictive accuracy (Omrani, 2015). In their research conducted between 2010 and 2012 using data from the National Travel Survey in the Netherlands, Hagenauer and Helbich (2017) applied various machine learning classification models to predict travel mode choices effectively. Their results indicated that Random Forest model outperformed the others. Nonetheless, they noted that while trip distance emerged as the most critical predictor, the importance of variables varied among different classifiers and class labels (Hagenauer & Helbich, 2017). In a parallel study using the same data, Kashifi et al (2022) set out to also forecast transportation mode choices. They extended the previous research by incorporating additional machine learning techniques into their analysis. Their results revealed that boosting and LightGBDT exhibited superior predictive accuracy for the different classes, especially when utilizing both under and oversampling methods to address class imbalance. Furthermore, their analysis underscored that age, income, and distance were the most influential predictors in the context of transport mode prediction (Kashifi et al., 2022).

Other applications of machine learning in the prediction of transportation modes can be characterized as more tailored to specific scenarios, as they are designed for modeling particular situations. In a recent study, Bhuiya et al (2022) focused on modeling transport mode choices for individuals with limited mobility in Dhaka. Their research findings indicated that multi-nominal logistic regression and linear discriminant analysis models exhibited superior predictive accuracy, particularly considering a smaller dataset (Bhuiya et al., 2022). Additionally, Zhao et al (2020) investigate differences between machine learning and logit models based on trip diary recordings. Their study findings from staff and students within the University of Michigan suggest that when deciding between machine learning and logit models for transport modeling, it seems there is a trade-off between predictive accuracy and the alignment with behavioural principles (Zhao et al, 2020).

Recent research developments in this field have introduced more sophisticated approaches, such as the adoption of artificial neural networks and deep learning methods. For instance, Zhang et al. (2020) introduced a deep neural network model for classification using data from Beijing, and their results demonstrate that this network model outperforms the random forest model in predicting transportation modes (Zhang et al., 2020). Furthermore, in a study based on national travel data from the UK, Bei et al. (2023) introduce a deep neural network model that goes beyond mere travel mode prediction, also addressing the purpose of the trip. Their research indicates that the model they proposed surpasses the performance of basic multinomial logit models and single-task neural networks (Bei et al., 2023). Additionally, Wang, Mo, and Zhao (2021) present a "theory-based residual neural network" model that integrates discrete choice models with basic neural networks, using three separate survey datasets. Their results indicate that the model they propose not only achieves superior predictive accuracy but also exhibits greater resilience compared to straightforward neural networks or discrete models (Wang, Mo & Zhao, 2021).

**References**

Ababio-Donkor, A., Saleh, W. & Fonzone, A. 2020, "Understanding transport mode choice for commuting: the role of affect", *Transportation planning and technology,*vol. 43, no. 4, pp. 385-403.

Bei, H. et al. (2023) “Joint prediction of travel mode choice and purpose from travel surveys: A multitask deep learning approach,” Travel, behaviour & society, 33, p. 100625.

Bhuiya, M.M.R. et al. (2022) “Application of Machine Learning Classifiers for Mode Choice Modeling for Movement-Challenged Persons,” Future transportation, 2(2), pp. 328–346.

Böcker, L., Dijst, M. and Faber, J. (2016) “Weather, transport mode choices and emotional travel experiences,” Transportation research. Part A, Policy and practice, 94, pp. 360–373.

Das, S. *et al.* (2021) “Impact of COVID-19: A radical modal shift from public to private transport mode,” *Transport policy*, 109, pp. 1–11.

Hagenauer, J. and Helbich, M. (2017) “A comparative study of machine learning classifiers for modeling travel mode choice,” Expert systems with applications, 78, pp. 273–282.

Kubat, M. (2021) An Introduction to Machine Learning. 3rd edn. Cham: Springer International Publishing AG.

Kufel, J. et al. (2023) “What Is Machine Learning, Artificial Neural Networks and Deep Learning?-Examples of Practical Applications in Medicine,” Diagnostics (Basel), 13(15), p. 2582.

Lee, E. and Kim, S. (2023) *Geographic Information Systems for Intermodal Transportation*. Elsevier.

Mayo, F.L. & Taboada, E.B. 2020, "Ranking factors affecting public transport mode choice of commuters in an urban city of a developing country using analytic hierarchy process: The case of Metro Cebu, Philippines", *Transportation research interdisciplinary perspectives,*vol. 4, pp. 100078.

McCarthy, L., Delbosc, A., Currie, G. & Molloy, A. 2017, "Factors influencing travel mode choice among families with young children (aged 0-4): a review of the literature", *Transport reviews,*vol. 37, no. 6, pp. 767-781.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior.

McFadden, D. (1974) “The measurement of urban travel demand,” *Journal of public economics*, 3(4), pp. 303–328.

McFadden, D. and Train, K. (2000) “Mixed MNL models for discrete response,” *Journal of applied econometrics (Chichester, England)*, 15(5), pp. 447–470.

Mussone, L. & Changizi, F. 2023, "A study on the factors that influenced the choice of transport mode before, during, and after the first lockdown in Milan, Italy", *Cities,*vol. 136, pp. 104251-104251.

Omrani, H. (2015) ‘Predicting travel mode of individuals by machine learning’, *Transportation Research Procedia*, 10, pp. 840–849. doi:10.1016/j.trpro.2015.09.037.

Shanmugam, L. and Ramasamy, M. (2021) “Study on mode choice using nested logit models in travel towards Chennai metropolitan city,” Journal of ambient intelligence and humanized computing, pp. Journal of ambient intelligence and humanized computing, 2021.

Stefan Trueck, S.T.R. (2009) Rating Based Modeling of Credit Risk. 1st edn. San Diego: Elsevier Science.

Tamim Kashifi, M. et al. (2022) “Predicting the travel mode choice with interpretable machine learning techniques: A comparative study,” Travel, behaviour & society, 29, pp. 279–296.

Wang, S., Mo, B. and Zhao, J. (2021) “Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks,” *Transportation research. Part B: methodological*, 146, pp. 333–358.

Willis, K.G. (2014) “The Use of Stated Preference Methods to Value Cultural Heritage,” *Handbook of the Economics of Art and Culture*, 2, pp. 145–181.

Zhang, Z. et al. (2020) “A Customized Deep Neural Network Approach to Investigate Travel Mode Choice with Interpretable Utility Information,” Journal of advanced transportation, 2020, pp. 1–11.

Zhao, X. et al. (2020) “Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models,” Travel, behaviour & society, 20(C), pp. 22–35.

Zhou, ZH. (2021). Introduction. In: Machine Learning. Springer, Singapore. https://doi.org/10.1007/978-981-15-1967-3\_1