Using Travel Demand Survey Data to Predict Mode Shift by Machine Learning Techniques; Case Study: Bangkok

Watcharapong Wongkaew[1], Krittiphong Manachamni[2], Kongtup Wanichjaroenporn[3] Wachira Muanyoksakul

[1]Chulalongkorn University Transportation Institute, Bangkok, Thailand, watcharapong.w@chula.ac.th
[2]Chulalongkorn University, Bangkok, Thailand, Email: 6330011321@student.chula.ac.th
[3]University College of London, London, United Kingdom, kongtup.wanichjaroenporn.22@ucl.ac.uk
[4]Chulalongkorn University, Bangkok, Thailand, 6678002021@student.chula.ac.th

Keywords: Neural Network, Mode Shift, Transportation Hubs, Travel Demand Survey, Transportation Network Analysis

Extended Abstract

1. Introduction

In recent years, urbanization and population growth have resulted in escalating challenges for transportation systems in major cities worldwide. Among these cities, Southeast Asian Cities, produce most of the changes and give dynamic sense as a prime example for that. In major metropolitan areas like Jakarta, Manila, Bangkok, Kuala Lumpur, and Singapore, the strain on existing transportation infrastructure is evident. Insufficient public transportation options and inadequate road networks contribute to increased travel times, decreased productivity, and heightened environmental issues.

To address these challenges, countries in Southeast Asia are striving to develop efficient transportation solutions. Investment in public transportation projects, such as building and expanding rapid transit systems like MRTs, LRTs, and BRTs, is gaining momentum. These projects aim to shift travel mode that people use and for this. Mode shift prediction hold several significant importance in the context of transportation planning and urban mobility. By accurately predicting the mode shifts, we can identify areas where there is a higher potential for mode shift and focus on developing the new transportation hub.

Usually, mode choice and shift predictions are usually done by doing the travel demand survey and then using Revealed Preference (RP) and Stated Preference (SP) information (Idris et. Al, 2013). to create demand models or use machine learning detection method to better suit the bigger data of the survey. The most traditional way to do this is by using Travel Demand Survey (TDS) data and creating discrete mode choice analysis (Chalermphong, 2018). model to see whether demographics have been shifted or not.

These methodologies can be applied to many circumstances, although using Stated Preferences are biased toward hypothetical choices and analyzing hypothetical situations. While creating four step model (Cass et. Al, 2015) analysis is real choices based on the mode already implemented. These methodologies then were presented with many obstacles like obsolete data, difficulties in obtaining accurate costs, and many more.

Bangkok, the vibrant and dynamic capital of Thailand, stands as a prime example. To address these challenges, the Department of Rail Transport has devised and revised the Railway Master Map (M-Map 2) and recently rectified it to better suit the demand. Last month, Bangkok got the first monorail line (Yellow Line) in the outer ring of the city, and it presented a new mode of transportation that never seen before in the area.

While the city had many plans and models implemented on, with its ever-expanding population and increasing mobility demands, the need to develop efficient transportation solutions has become a pressing priority Bangkok had been presented with those challenges that have been attempted to address before, this prompted us challenges to create more complex model for solving one.

The Office of Transport and Traffic Policy and planning (OTP) aims to alleviate this challenge while planning the new transportation plan beyond the COVID-19 pandemic. It conducted Bangkok Travel Demand Survey in 2022, and with that the research team aims to use the data to help understand the transportation network better. Therefore, this is great opportunity to model the shifting mode choice in the area using the yellow monorail line.

This research aims to explore how machine learning can be applied to predict mode shifts in travel behaviour and identify potential transportation hubs in Bangkok. By analyzing comprehensive TDS data that includes vital information about travel patterns, preferences, and demographics, we can gain valuable insights into the city's transportation dynamics. The fusion of data-driven predictive models with cutting-edge machine learning algorithms (Law et. Al, 2023). offers a transformative opportunity to optimize urban mobility planning and adapt faster to the new digital world of mobility.

2. Mehodology

This study will be divided into 2 parts: the prediction model of transportation mode shift and experiment on transportation hub identification. Areas in Bangkok that will be used as the test areas are 1. All areas with urban railway line for prediction model part, and 2. Area of the new yellow monorail line for the experiment part. The model goal is mainly to accurately predict and determine relevant variables that affect mode shift of the population (Anantsuksomsri et. Al, 2020)., and the experiment will be conducted to examine the potential transportation hub in the future for yellow monorail line. (Lan et. Al, 2022).

With this application and method, we can diversify the travel demand survey data benefit. Therefore, this study would change how we use the travel demand survey data and tackle the problem of data availability and quality within the system.

2.1. Prediction of mode shift

In this section, we explored various machine learning techniques to try and predict the mode choice from the Travel Demand Survey data, collected before the opening of yellow monorail line. Then, we predict the mode shift and new potential transportation hub.

We explored and analysed comprehensive TDS data that includes information about travel patterns and demographics with both statistical methods and machine learning methods. We then create the model to predict the mode of travel of demographics first, then we used the models on the yellow line operated area and then compare the results to see potential shifts. To taking into account the impact of any occurrences related to the yellow monorail line. We hypothesize that this scenario assumes:

1. All types of transport are easily accessible and can reach the target area.
2. The best model accurately reflects consumer behaviour but is influenced by certain factors that may prevent them from choosing a particular travel style.

The TDS data which we use to determine mode choice are selected and filtered to 20 fields and described briefly below in **Table 1**.

Table 1 Travel Demand Survey data description

Information Type	**Data**	**Type of data**
Demographics	Household Type	Category
	Household Member	Numerical
	Household Age Group	Numerical
	Gender	Categorical
	Register	Categorical
Socioeconomics	Occupation	Categorical
	Income	Numerical
	Travel Ability	Categorical
	Has Personal Car	Categorical
	Has License for Personal Car	Categorical
	Has Motorcycle	Categorical
	Has License for Motorcycle	Categorical
Travel Behaviour	Total Trip	Numerical
	Amount of Trip before COVID	Numerical
	Amount of Trip after COVID	Numerical
	Mode of travel before COVID	Categorical
	Mode of travel after COVID	Categorical
	Trip sequence	Numerical
	Travel Objective	Categorical
	Travel Distance	Numerical

We have explored the statistics model that stems from discrete choice analysis like Binary Logit Model, Probit Model, Multinomial Logit Model and Nested Logit Model. We analyzed the theoretical part of those models, and then extract the logistic regression features of the model and then apply on to parts of the model. We then selected 5 candidates of standard and machine learning model to predict mode of travel: Multinomial Logistic Regression, Random Forest Classifier, Multi-Layer Perceptron (Akgöl et. Al, 2014), XG-Boost, and LGBM (Kashifi et. Al, 2022). which represent in **Table 2**.

Table 2 Comparison of candidate's model for mode-choice prediction

Model	Description
Multinomial Logistic Regression	Estimates the probability of multiple outcomes using logistic functions
Random Forest Classifier	Ensemble learning method that constructs multiple decision trees
Multi-Layer Perceptron	Artificial neural network composed of multiple layers of nodes
XGBoost	Gradient boosting decision tree algorithm

LightGBM	Gradient boosting decision tree algorithm

To optimize the model over the training period, we use loss functions and accuracy as a metric to better validate the model with data. The metrics we use are listed below in **(Table 3)** and use Adam (adaptive moment estimation) as optimizer in selected model. (Nam & Cho, 2020)

Table 3 Metrics for benchmark model

Metrics	**Description**
Accuracy	Percentage of all correctly classified observations.
F1-score	The evaluation metric that combines Precision and Recall.
Precision	Correct positive predictions relative to the total positive predictions.
Recall	Correct positive predictions relative to the total number of actual positive cases.

In this sub-part, we use the TDS data to determine mode choice by using algorithms and explored insights from the dataset. With the algorithm, we use that includes and transform information about travel patterns and demographics into interpretable results with statistics interpretation.

Since we will try to predict the mode shift of the area which yellow line operated, we create the buffer area with 5 km. travel distance from the transit line as for TOD (Transit Oriented Development) stated (Pongprasert & Kubota, 2019)., we then use all of other areas in Bangkok as training data, and then we divide dataset in to training set, test set with both external and internal dataset which explained below in **(Table 4)**.

Table 4 Proportion of Data splitting in the study

Data Split	**Train-test-split ratio**	**Area**
Training Set		
Training set	50%	Bangkok Metropolitan Area excluding Yellow Monorail Line Area
Validation set	17.5%	
Internal Testing set	17.5%	Urban Railway Line within Bangkok Metropolitan Area
Test Set		
External Testing set	15%	Yellow Monorail Line Area

With both statistical and machine learning models. We then created the model predict mode choices of the people within the internal dataset area, then we can train and calibrate our model within the internal area, validation with internal test set with parameters in **Table 1.**

2.2.Prediction of Mode Choice on the new transportation line

In this sub-part, we use models that trained and validated with internal data to predict on TDS data to determine mode choice of all of the external test set which is the area of the newly opened yellow monorail line, then analyze and interpret the error of the models and see the population that shifted to the yellow line.

After that, we compare prediction results from different models and datasets using the same evaluation metrics. The internal test set represents the same environment as the training and validation sets, so the model should perform well in F1-score due to the similar distribution of data. Conversely, the external test set represents trips within 5 km of the yellow line, which is an area of interest. The best model will perform well in both accuracy and F1-score, with minimal differences between the two.

The Random Forest Classifier showed the best performance in all metrics within the internal and external test sets. Although there was a drop in F1-score, precision, and recall, the performance was still the highest. A possible explanation for why the random forest model performs the best is that the ensemble method is a lot of voting decision tree and has a direct link to choice analysis.

3. Results

3.1. Mode Choice Prediction

To fine-tune all models, we focus on factors such as the complexity of the model, optimizer, learning rate, and loss function until we find the best setting for each model. Based on the performance evaluation in **(Table 5)**, we identified key features for each model in **(Figure 1A-E)**. Multinomial Logistic Regression helps us to choose the mode by providing insight into the weight of each mode and selecting the highest value. Other methods, especially Random Forest, identify descending feature importance shown in **(Figure 1C).** Moreover, examining Multilayer Perceptron model which also has SHAP values that visualize the impact of each parameter on the model as in **(Figure 1E).**

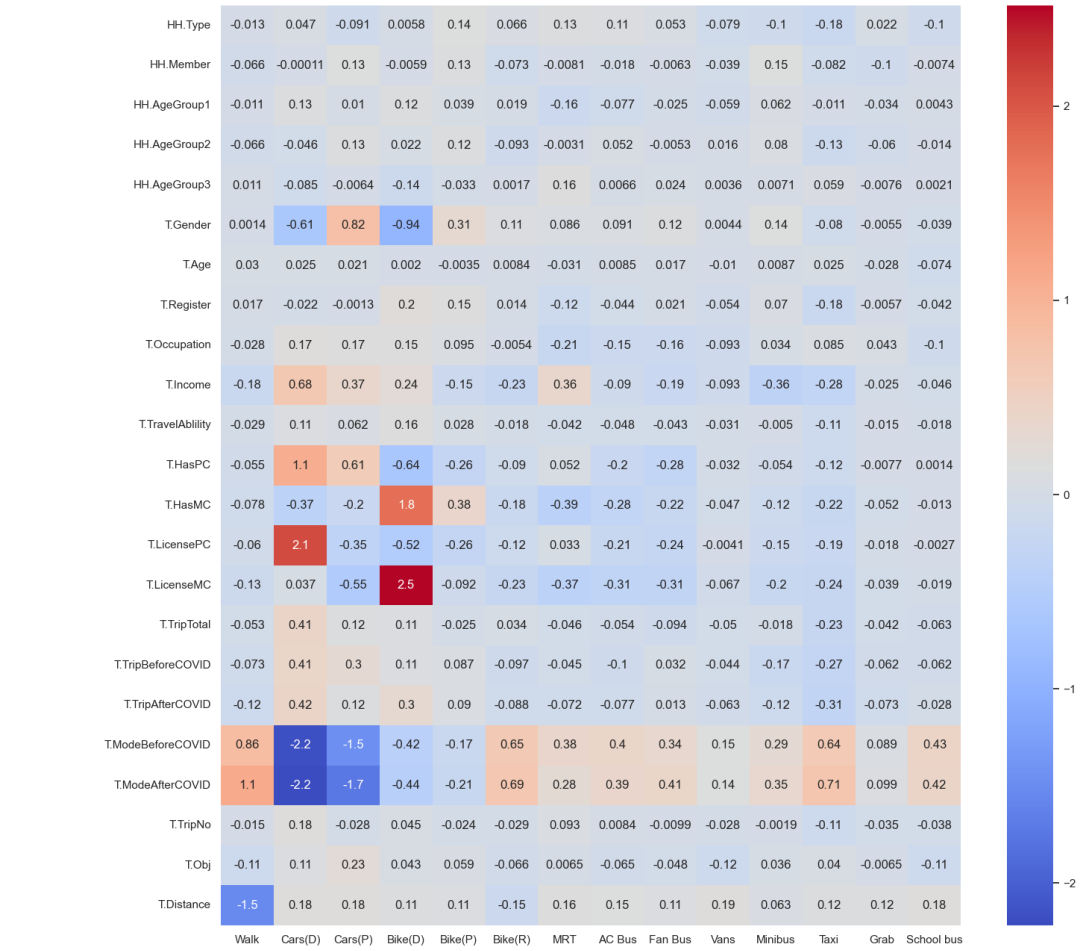

Figure 6 (A) Features Importance from Multinomial Logistic Regression's weights

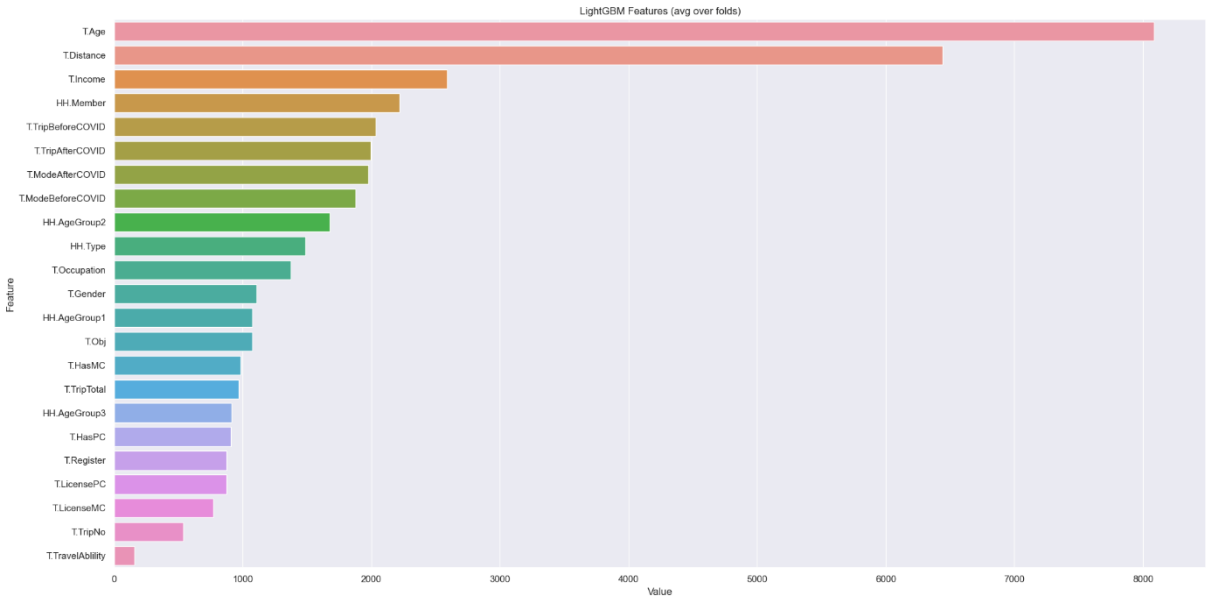

Figure 6 (B) Features Importance for LightGBM's features importance

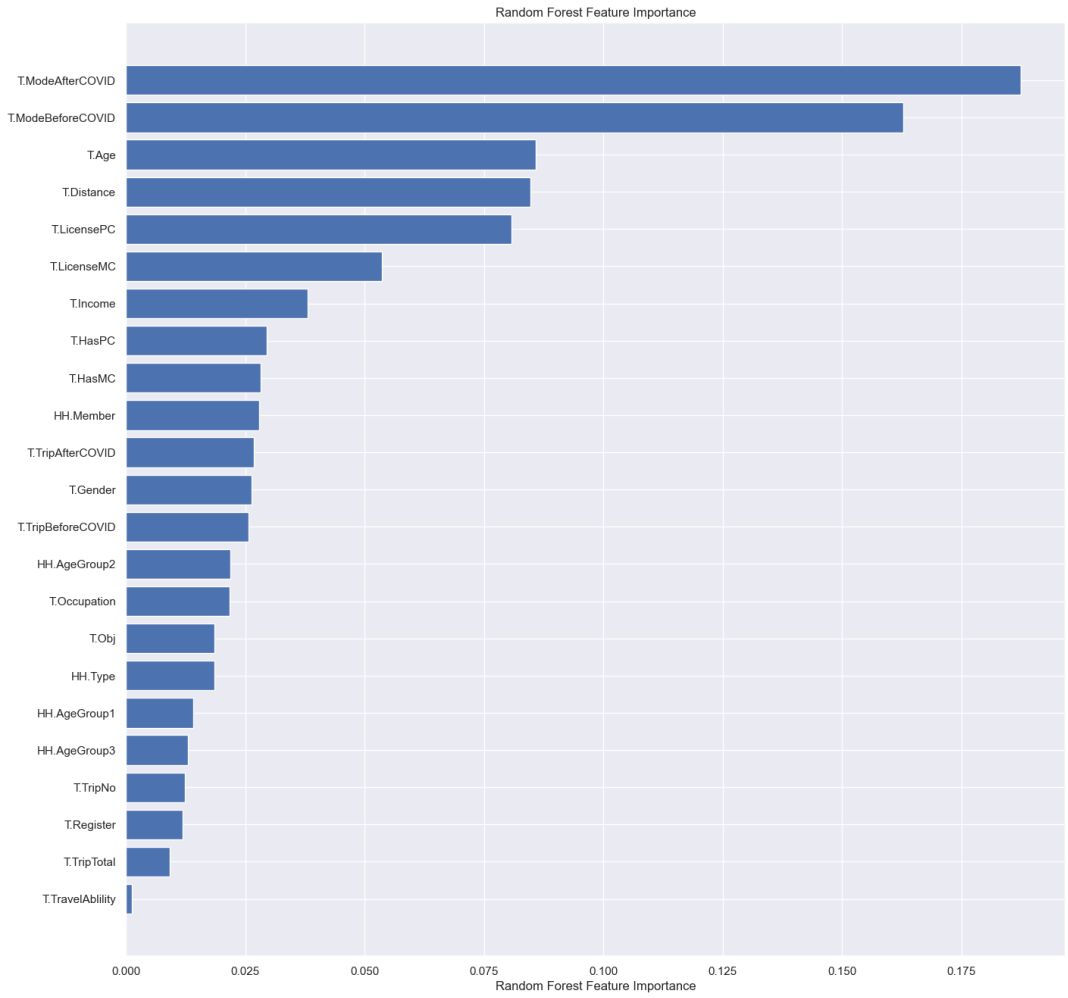

Figure 6 (C) Features Importance for LightGBM's features importance

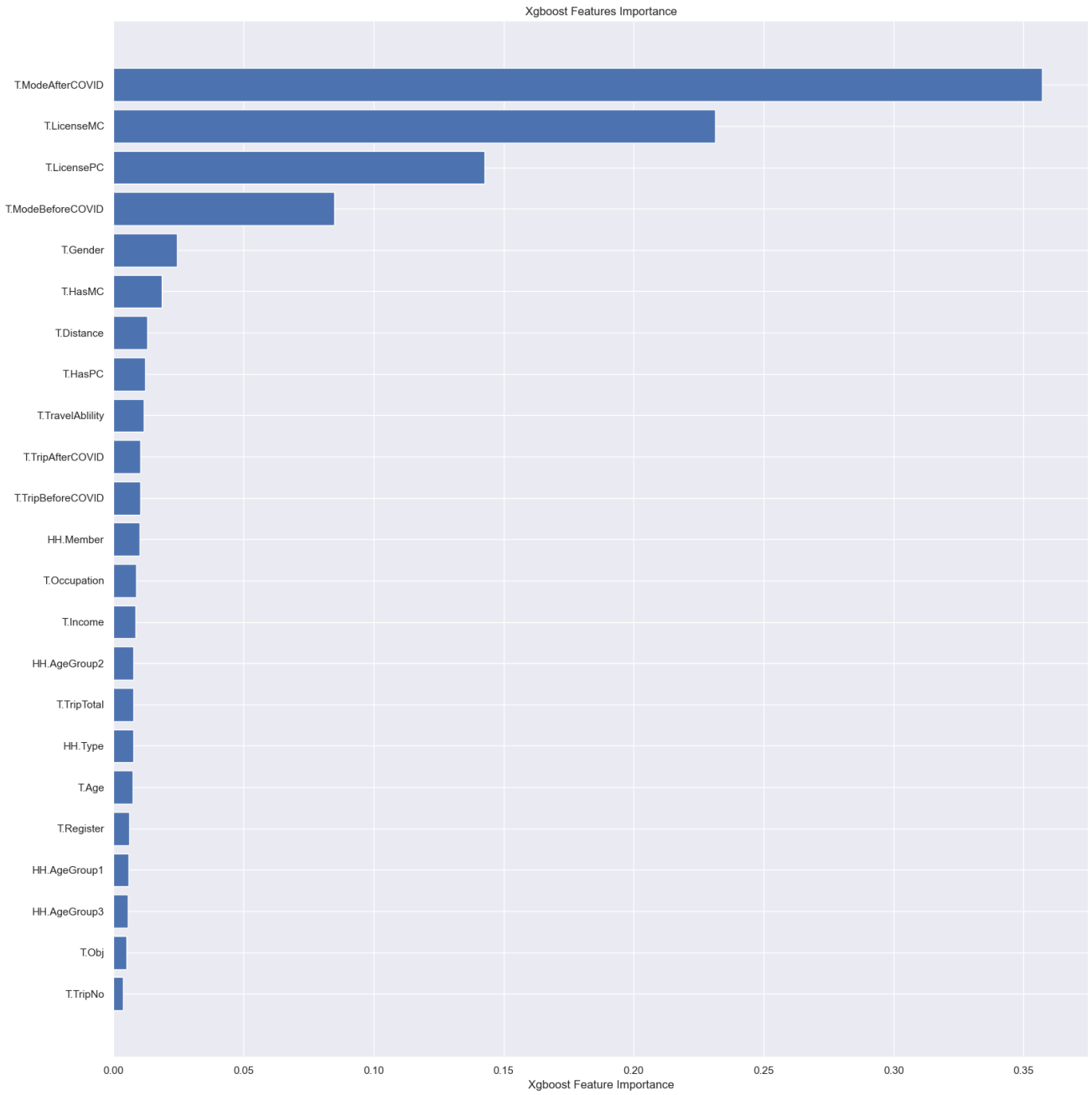

Figure 6 (D) Features Importance for XGBoost's features importance

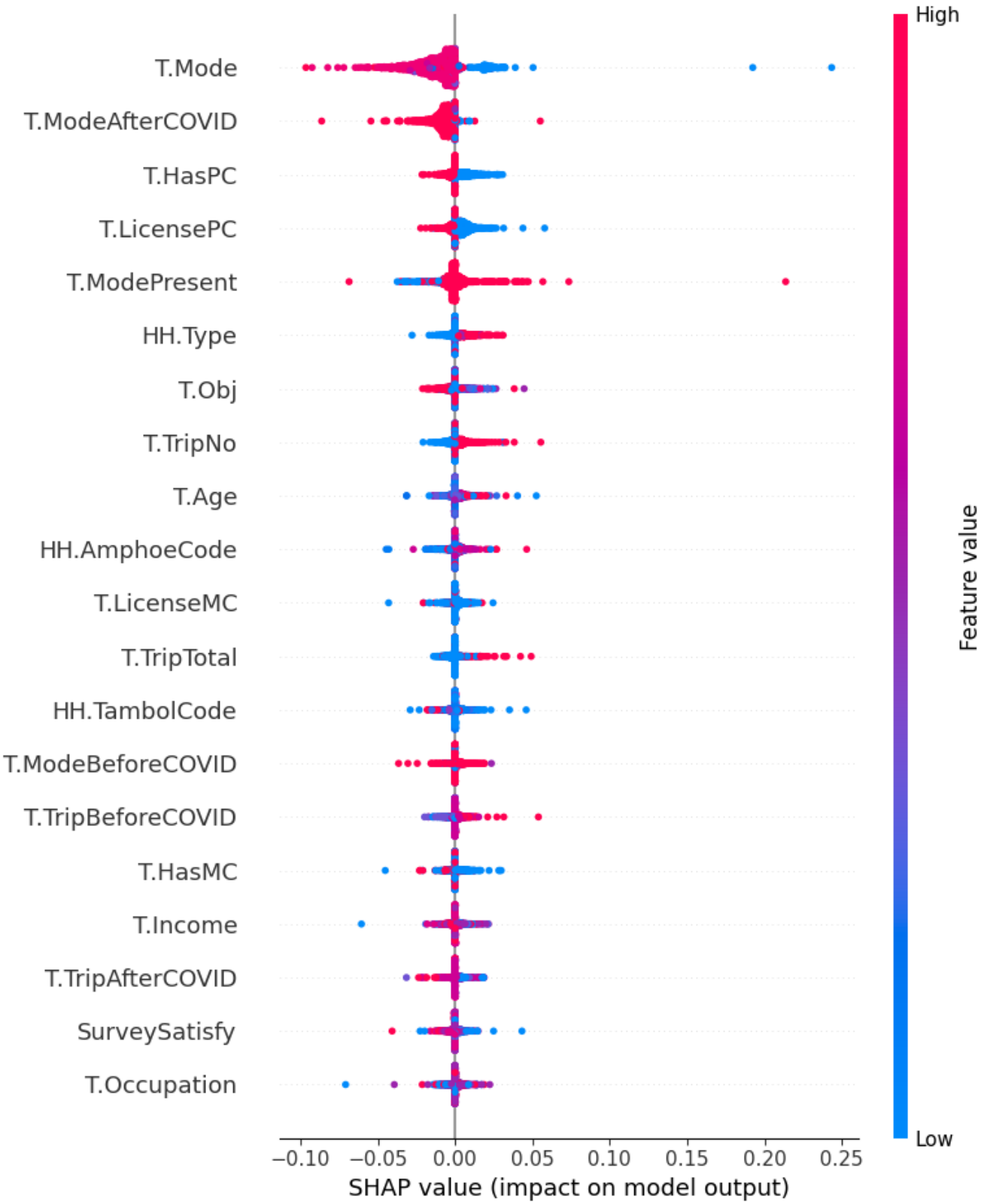

Figure 6 (E). Features Importance from MLP's SHAP value

According to **(Table 5)**, the top model achieves 85% accuracy, which is considered state-of-the-art in predicting mode choice. Even if people prefer other modes of transportation, they may face obstacles accessing the route or the route may not exist yet. Therefore, it is important to utilize available modes of transport for the trip.

DISCUSSION

Model results and interpretation

Table 5 shows that demographic data has been used to predict mode choice using different computational techniques. The model's accuracy is around 80%, with F1-scores ranging from 0.36-0.71. The recall is low, between 0.36-0.67 in the internal test set, indicating that the model struggles to detect positive cases.

In the internal test set, the random forest model has the lowest recall and F1-score for predicting the Minibus mode choice. For predicting the metro mode choice, the recall and F1-score are good, indicating excellent performance in detecting positive results for the metro mode. Therefore, the mode-choice model can be used to predict shifts, especially when using the metro.

XGBoost and LightGBM are among the top performers in our study, but random forest has a great perform in F1-scores, making it the superior model. It can be used for computing mode-shift and transportation hubs to minimize errors when other substitute models are unavailable.

In the external test set, the random forest model achieved a recall of 0.80 and an F1-score of 0.62. However, the model exhibited poor precision performance, emphasizing the need to shift to using the metro. The transportation hub change has resulted in an increase in the number of passengers benefiting from the opening of the yellow line, especially at Lat Phrao 71 and Hua Mak stations.

While Thailand has the eBUM model for predicting large-scale shifts using trip generation methods, this paper focuses on using behaviours as the primary factor for predicting mode shift and understanding people based on their persona and background. Additionally, the model's error is still mixed with the real shifted impact from the opening of the Yellow Line metro in Thailand.

CONCLUSIONS

In conclusion, this study delved into various machine learning techniques to predict mode choice and mode shift using TDS data, enabling us to forecast potential new transportation hubs along the proposed metro line. Among the methods explored, Random Forest emerged as the most accurate, achieving an 85% prediction accuracy. However, further optimization of the model could potentially yield even better results. Utilizing our model's predictions, we identified key characteristics of potential new transportation hubs, including their association with upcoming neighborhoods, pre-existing market and commercial districts, and established residential areas/flats. These insights would provide valuable guidance for policy planning and developing new transportation infrastructure in the region of adopting Transit-Oriented Development (TOD) strategies by aligning transportation hubs with urban development, to enhancing connectivity and accessibility.

Despite the valuable contributions of this study, there are certain research limitations that warrant acknowledgment as followed, unbalanced data between modes of transport hampering the potential of model. Moreover, encoded TDS data without real data proven very hard to distinguish and affect the results of the model. For future research, it is essential to explore additional variables and real-time data sources to enhance the model's predictive capabilities. Moreover, conducting in-depth case studies could further refine and validate the model's accuracy in various urban contexts. This study's findings try to offer new approaches and valuable insights for transportation planning and development, paving the way new policy planning and solutions in the future.

Acknowledgement

The authors would like to express sincere gratitude to Dr. Saksith Chalermphong, Professor of Transportation Engineering and Dr. Pathinan Thaithatkul of the Transportation Institute, Chulalongkorn University for advice throughout the entire research process. Lastly, we would like to express our appreciation to Office of Transport and Traffic Policy and Planning which had provided the data that contributed completeness of this research project.

References

Idris, A. O. Modal Shift Forecasting Models for Transit Service Planning. Graduate Department of Civil Engineering University of Toronto, 2013.

Chalermpong, S. Discrete Choice Analysis for Transportation Engineering. Chulalongkorn University Press, Bangkok, 2018.

Garber, N. J., and L. A. Hoel. Traffic & Highway Engineering, Fourth Edition. Cencage Learning, Toronto, 2009

Cass N., and J. Faulconbridge. Commuting practices: New insights into modal shift from theories of social practice. *Transport Policy*, 2015. http://dx.doi.org/10.1016/j.tranpol.2015.08.002

Lawson, C. T., E. Krans, E. G. Rentz, and J. Lynch. Emerging trends in household travel survey programs. *Social Sciences & Humanities Open*, 2023. https://doi.org/10.1016/j.ssaho.2023.100466

Anantsuksomsri, S., M. A. Turquist, and N. Tontisirin. Using Household Survey to Forecast Household Mode Choice and Trip Sharing. *Environment Asia*, 2020. https:// doi: 10.14456/built.2020.3

Lan, T., H. Cheng, Y. Wang, and B. Wen. Site Selection via Learning Graph Convolutional Neural Networks: A Case Study of Singapore. *Remote Sensing*, 2022. https://doi.org/10.3390/ rs14153579

Kashifi, M. T., A. Jamal, M. S. Kashefi, M. Almoshaogeh, and S. M. Rahman. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society*, 2022. https://doi.org/10.1016/j.tbs.2022.07.003

Akgöl, K., M. M. Aydin, Ö. Asilkan, and B. Günay. Prediction of Modal Shift Using Artificial Neural Networks. *TEM Journal*, vol. 3, no.3, pp. 223-229, 2014.

Nam D., and J. Cho. Deep Neural Network Design for Modeling Individual-Level Travel Mode Choice Behavior. *Sustainability*, 2020. https://doi:10.3390/su12187481

Pongprasert, P., and H. Kubota. TOD residents' attitudes toward walking to transit station: a case study of transit-oriented developments (TODs) in Bangkok, Thailand. *Journal of Modern Transport*, 2019. https://doi.org/10.1007/s40534-018-0170-1