Student thesis series INES nr 553

Applying machine learning approaches to model travel choice between micro-mobility services

Qilin Ren

2021
Department of
Physical Geography and Ecosystem Science
Lund University
Sölvegatan 12
S-223 62 Lund
Sweden

Qilin Ren (2021).

Applying machine learning approaches to model travel choice between micro-mobility services
Master degree thesis, 30 credits in *Geomatics*
Department of Physical Geography and Ecosystem Science, Lund University

Level: Master of Science (MSc)

Course duration: *January* 2021 until *June* 2021

Disclaimer

Applying machine learning approaches to model travel choice between micro-mobility services

---

Qilin Ren

Master thesis, 30 credits, in *Geomatics*

Supervisors

Pengxiang Zhao
Dep. of Physical Geography and Ecosystem Science, Lund University

Ali Mansourian
Dep. of Physical Geography and Ecosystem Science, Lund University

Exam committee:

Lars Harrie
Dep. of Physical Geography and Ecosystem Science, Lund University

Olive Niyomubyeyi
Dep. of Physical Geography and Ecosystem Science, Lund University

Abstract

Shared micro-mobility gradually becomes a crucial part of human daily transportation. To develop the shared micro-mobility, discovering the important influence factors of each travel mode is a key aspect. However, there are scarce studies that adopt machine learning methods to model travel choice between shared micro-mobility services and identify the crucial determinants for each mode. This study aims to apply four different machine learning methods (random forest, support vector machine, artificial neural network, and logistic regression) to simulate the decision schema in time and space and examine the factors that significantly influence citizens' travel mode choice in Zurich, Switzerland.

Collected data are used to build the choice set, which mainly includes trip attributes and external environment factors. Then, the four Machine Learning methods are used to examine the importance of each influence factor by permutation importance, and the performances of four ML models are compared. How the top 6 influence factors with higher importance affect the human choice of shared micro-mobility service is analyzed as follows.

The Random Forest model showed the best-predicted performance. With respect to the feature importance in the RF model, trip attributes, such as the duration and the length of the trip, are identified as the most important influence factor, followed by some POI type and density around the destination, like public facility, and education services. By contrast, weather affects people's choices slightly. It is noteworthy to see that dockless facilities always have priority when there is the same type of docked facilities without considering other variables, but docked services are still preferable under some situations. The results are valuable to policymakers and shared services provided by companies to adjust the shared micro-mobility system and contribute to better sustainable transportation.

Keywords: *Geography; Shared micro-mobility; machine learning; travel mode choices; transportation*

Acknowledgment

I would like to address my thanks to the Department of Physical Geography and Ecosystems Analysis, Lund University, Sweden, for offering me the opportunity to enhance my knowledge in GIS science by attending the Master’s Program in Geomatics. Although the pandemic influences the completeness of the study abroad experience, the eight months in Lund bring me super precious memories. I would also like to extend my thanks to my supervisors Pengxiang Zhao and Ali Mansourian, as well as examiner Lars Harrie. Their valuable feedback and inputs support me a lot. My opponent Emma Fogelström also offered great ideas for my work. At the same time, I would also like to thank my parents for the moral and financial support. Without their encouragement, the wonderful experience in Lund would not occur.

Table of contents

1. Introduction

1.1. Background

During the past decade, a newly appeared term “Micro-mobility” has been proposed. Micro-mobility represents traffic modes with small size, lightweight, lower speed, and widely used for short-trip becoming more and more popular (Yanocha 2019). The scope of this term includes both personal traffic tools, like the bike, and shared transportation, for example, shared bike or scooter. In recent years, shared transportation is spring up and starting to play a significant role in human’s daily life, especially the shared micro-mobility. It is the ideal complement to private and public types of short distances transport, which helps ease traffic congestion in urban area and ameliorates the attractiveness of public transport. Due to its characteristic of source-saving and flexibility, the micro-mobility service could become a significant substitute for petrol-based traffic tools and contributes to sustainable development and better connection between passengers and the public transportation system (Zhou et al. 2019). Thus, It is crucial to enhance the understanding of shared-micro-mobility as well as the factors influencing the likelihood of using shared-micro-mobility.

The term micro-mobility has an ambiguous definition that which kind of mobility mode is included (McKenzie 2020). In this project, only the main shared micro-mobile services will be analyzed due to data availability. Four representatives are explored recently: docked-bike, docked e-bike, dockless e-bike, and e-scooter. The first widespread shared-micro-mobility is docked bike. The docked facility requires users to bring a bike from a station to another station. Compared with the docked bike or docked e-bike, the dockless service equipped the lock on the vehicle itself, instead of a docking station. E-scooter is typically dockless as well.

BSS (bike-sharing system) is a newly appeared traffic choice which plays a more and more crucial role in individuals’ daily life. According to the global bike-sharing market research report, the market is forecast to arrive at 5 billion dollars by 2025, which is double the market in 2018 (Dublin 2020). The slogan of the biggest Chinese shared bike company is “Liberation of the last mile”, which reveals that except for short trips, people could also choose BSS combined with other public transportation. It implies that the BSS also helps to improve the utilization of other public transport like the metro and bus. As important shared and micro transportation, e-scooter as diminutive transportation is also becoming an upcoming trend in terms of mobility (Degele et al. 2018). Meanwhile, under the pandemic of Covid-19, micro-mobility as an individual traffic method also helps to avoid unnecessary touching in a hermetic space like buses. Therefore, researching on improving micro-mobility and other approaches that could replace motor vehicles, as well as make a thorough inquiry about what are the major factors that affect human beings choosing the transportation mode is critical.

There are plenty of empirical studies that have been conducted for examining the influencing factors of travel mode choice in previous years. The pertinent factors could be roughly classified as two main types: social-demographic factors, such as gender, age group, attitude, habits, trip purpose, and so on (Ding 2018; Klöckner and Matthies 2004; Sivasubramaniyam et al.2020; Tran et al. 2020; Zhao et al. 2020) as well as spatial-temporal context, for instance, topography, accessibility, weather, etc. (Bachand-Marleau st al. 2012; Bucher et al. 2020; Stojanovski 2019; Zhou et al. 2019). The methods of identifying the more crucial factors among the above factors are changing from statistical regression framework(e.g. linear regression model, multinomial logit model, etc.) (Bhat and Srinivasan 2005; Van Acker and Witlox 2011) to machine learning model (e.g. random forest, support vector machine, artificial neural network, etc.) (Cheng et al. 2019; Ermagun et al. 2015; Jahangiri and Rakha 2015; Shafique and Hato 2015). Based on the previous researches, only a few studies compared different machine learning methods for modeling travel mode choice on shared micro-mobility and analyzed the important influence factor in time and space. And therefore, analysis of travel mode choice is of great benefit for developing micro-mobility transportation. This paper tries to bridge these gaps.

The remainder of the paper is organized as follows. Section 2 presents a literature review on travel choice models and travel decision classification models between the four shared micro-mobility services, as well as shared micro-mobility studies based on these models. Section 3 presents the overview of the entire process in this study, the conceptual basis of proposed models, including their key assumptions, characteristics, and how to optimize the model. A fundamental state of the study area and introduction of the dataset in more detail are provided. The various preprocessing steps are covered as well. In the next, Section 4 compares the performances of different models and indicates the most virtual influence factors on shared micro-mobility choice according to the classification results. Section 5 analyses the results of this paper and compares them with other perspectives of similar results of other papers. The future research directions are also discussed in this part. Finally, Section 6 answers the research questions proposed at the beginning of this paper.

1.2. Project aim

This study intends to develop a deep understanding of the decision-making mechanism of commuters regarding the mode choice based on shared micro-mobility in a current urban transport system at a high spatiotemporal resolution. The machine learning methods including support vector machine, random forest, artificial neural network, and logistic regression are adopted to analyze multi-dimension impact factors and significance and compared these models on classification performance. This has the following three objectives:

- estimating the influencing factors that affect the choice among four shared micro-mobility transportation modes(docked bike, docked e-bike, dockless e-bike, e-scooter),
- compare the performance of different machine learning methods,
- evaluate the importance of each influencing factor.

In short, these three research questions are addressed:

1. What are the dominant factors that affect transportation mode choice between micro-mobility services (docked bike, docked e-bike, dockless e-bike, e-scooter)?
2. Which machine learning method gives the best accuracy?
3. According to the importance of each influence factor, what proposals of potential policy which may improve the utilization of each service could be put forward?

2. Related works

This section reviews related research on travel mode choice from diverse aspects of data resources, research subjects, and methods. Furthermore, the studies related to shared micro-mobility are reviewed. This literature review was used to identify the scope of previous studies of shared-micro-mobility and gave the basis of the entire methodology of this study.

2.1. Travel mode choice

Travel mode choice is roughly categorized as walk, bicycle, car, e-motorcycle, and public transport. Most studies are conducted among these mode choices. Private vehicle, walking, bicycle, and public transport (Marra and Corman 2020; Cui et al. 2020; Ton et al. 2020; Narayan et al. 2020; Wangsness et al. 2020) is the most frequently detected modes (Wu et al. 2020). During the previous decades, different researches have been conducted for simulating the travel choice mode as well as examining the associations between urban environment and transport mode choice among mentioned modes above. For instance, Holmgren and Ivehammar (2020) analyzed the factors that determine mode choice in situations where bicycles and walking are competitive options of private cars and public transportation (PT) in mid-size towns. A study in New Zealand explores the relative influence of these different types of motivational factors by investigating a wide range of different types of commuters, such as drivers, car passengers, bus users, cyclists, and pedestrians (Sivasubramaniyam et al. 2020).

However, with the development of society and techniques, the shared micro-mobility services are increasingly popular due to their convenience and environmental protection potential. The influx of micro-mobility services provides alternative ways of the motor vehicle in a shorter distance and leading to a substantive change in urban transportation in many cities(McKenzie 2020). Hence, it is crucial to examine the travel mode choice to optimize the shared micro-mobility services based on study results for better planning of city transportation system in the future.

The most used data source in previous studies is questionnaires and surveys in terms of transportation mode choice (Weis et al. 2021; Haas et al. 2018; Holmgren and P 2020; Scheiner and Holz-Rau 2007; Zhao et al. 2020) when respondents commute. In general, by questionnaires and surveys, more comprehensive information could be collected, including personal information, preference information, and so on. For example, an empirical evaluation was conducted based on a state-preference survey completed at the University of Michigan on the Ann Arbor campus (Zhao et al. 2020). The trip attributes for each travel mode (such as travel duration and travel distance), socio-demographic variables (e.g. gender and income), transportation-related residential preference variables (which is defined as the importance of transit availability of different travel modes when deciding where to live), and current/revealed travel mode choices are the main context of the survey. But the drawback of this kind of data is that the validation of data is hard to do, especially for the survey with heavy question volume (Mayo and Taboada 2020).

However, under this internet generation, trip information could be obtained through openly accessible company APIs (Application Programming Interfaces) (Reck et al. 2021). In this way, the accuracy of data is highly guaranteed compared with traditional survey methods. Moreover, compared with conventional data collection methods, due to its characteristic of automation collection, the GPS data could offer a tremendous amount of samples for analysis, which is more representative. However, GPS data could only offer the information of existed data, which means that it is impossible to learn of people who did not use the service. Why users do not choose the shared micro-mobility service is out of the study question. Except for crawling data by open APIs, the GPS data can be collected from smartphone sensors carried by users to infer what transportation mode the users have used (Jahangiri and Rakha 2015).

2.2. Influencing Factor

Influence Factors determinants of travel mode choice have been one of the major research directions so far. The determinants are usually classified into three categories: self-factors, external factors, and trip attributes. Self-factors normally includes subjective attitude and individual personal information; external factors are about environmental condition, like the weather; Trip attributes are duration, length and speed of the trip. To detect travel patterns, a variety of studies have paid attention to individual differences. For those works that take questionnaires as an information source, the question setting normally involves psychological (Ding et al. 2018) and social-demographic factors for individuals (Ding et al. 2018; Mayo and Taboada 2020) as a source of self-factors. Therefore, the collected answers from the survey are not only fact records but also some subjective awareness and preference. The advantage of this information source is that the attitude of informants is revealed, but data accuracy is sacrificed to some extent. For example, the social demographic characteristics (i.e.

gender, age, level of education, occupation, income, household type, etc.) are explored and identified how the different influence factors implicit different types of commuters' mode choices (Sivasubramaniyam 2021). The most frequently used psychological factors are the user's attitude to different traffic modes, such as levels of environmental friendly of each mode, the usefulness of each mode, and human's daily habit (Wu et al. 2020). For example, Tyrinopoulos's study indicated that the availability of parking space is the main factor which affects the preference of respondents towards car. Meanwhile, considering the trip purpose, using the car for work trips is indicated as the most dominant aim for the general population and female respondents, the following purpose are shopping and personal trips and finally leisure trips. An empirical model is created by comparing multinomial logit modal, nested logit modal, and error component mixed logit modal to research travel behaviors at the household level. How family formation affects each member's choice when they go out. In the same year, Mayo and Taboada ranked the crucial factors of transportation modes using the analytic hierarchy process. In their study, safety is ranked as the most important factor over travel cost, comfort, accessibility and attributes to the environment for the whole age group and regardless of income, travel distance, and gender of respondents (Mayo and Taboada 2020). However, due to the data limitation, it is difficult to collect the personal information and attributes and match with actual travel recorded by GPS data. Therefore, self-factor will not be considered in this study.

In addition, external factors also play important roles in travel mode choice studies. The contextual factors are divided into three aspects: policies, social norms, and geographical/climatic factors (Ding et al. 2018). The geographic and climate factors are frequently considered as potential influence factors for transportation choice, especially when considering travel mode choice from an objective aspect. For example, people may prefer transport mode that they could be covered like bus or car when it is raining day. Residential and firm density contributes to mobility patterns and transportation behaviors (Pouyanne 2004). The objective spatial conditions, as well as subjective location attitudes, influence mode choice are concluded (Scheiner and Holz-Rau 2007). The trip attributes normally include trip duration, trip distance, time of the trip, and so on (Jahangiri and Rakha 2015; Marra and Corman, 2020). Based on trip information, the other features could be collected.

2.3. Travel mode choice models

Models have traditionally been estimating travel mode choice using a statistical regression framework. In 2004, a triangular model has been generated for presenting the complex interactivity among mobility patterns, urban forms, and economic or demographic characteristics (Pouyanne 2004). The Ordinary least square (OLS) regression and multinomial logit model are used to verify what elements may affect citizens' travel mode. OLS regression as one of the basic linear regression analyses, did not perform as well as the multinomial logit algorithm, which is one of the Discrete choice approaches (DCA). Various studies about transportation mode choice have been conducted previously by logit regression. In a logit model, the utilities (the output values of the utility specifications) are then passed through a logistic function to generate choice probabilities for each option in each considered choice situation. Other model structures, such as the Nested Logit (NL) and Cross-Nested Logit (CNL), are the derivative algorithms of logit model, which allow for these probabilities to be calculated given correlations among the options in the choice-set (Hillel 2021). In 2013, factors that affect travel choice have been investigated for public transportation in an urban area by probit modal (Tyrinopoulos and Antoniou 2013), which also proved potential solutions of increasing public transportation utilization by analysis of case study and questionnaire. Probit modal could be seen as an expansion of the logit algorithm, such research was conducted focusing on the shifting process to public transportation in developing countries. The used algorithm was still multinomial logit modal. Todor explored the relationship between travel choices and urban form in deep (Stojanovski 2019) for triggering city transformation and redesigning to better integrate sustainable travel alternatives in Sweden. This study illustrated the travel pattern and influence factors in visual, local, and regional scale individually, gives a more broad view of what affects travel choices.

Mode-choice modeling can also be treated as a classification problem. In recent years, more machine learning models are proposed to offer alternatives methods for modeling travel mode choices. The main merit of machine learning models is that ML methods learn to represent complex relationships in a data-driven manner (Cheng et al. 2019). The performances of ML models (i.e. Bayesian, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, etc.) has been compared with traditional regression methods in many studies (Jahangiri and Rakha 2015; Liu and Luo 2018; Zhao et al. 2020; Zhu et al. 2021). Zhao et al (2020) made a comparison between machine learning and logit models based on prediction and behavioral analysis of travel mode choice on a survey dataset with around 8 thousand records. The machine learning method with the best performance in this study is random forest. However, to a large extent, machine learning and logit models provide similar feature importance and if the influence that each independent variable has on the choice outcome is positive or negative. Based on the review study of Pineda-Jaramillo's study in 2019 and Hillel's study in 2021, the most frequently used machine learning models which always give the best accuracy when compared with other models are random forest and multi-layer perception. SVM also performs well sometimes. E-bike as an alternative way for a longer trip, which will encourage people to choose it was analyzed by Bucher in 2019, Environment friendly is an increasingly significant factor when people will travel around. One of the newest explorations paid attention to what factor would affect travel choice between an electric car and internal combustion engine cars (Bucher et al. 2020).

Overall, the logit modal is one of the popular algorithms which is widely used in terms of transportation mode choice topic in past projects. However, traditional regression-based models perform a bit poorly when dealing with massive attributes and nonlinear relationships (Zhu 2021). Meanwhile, more and more researchers start to apply various Machine Learning techniques to mode choice problems and get successful results (Hillel 2021).

2.4. Shared Micro-mobility

Shared micro-mobility has gradually become a hot topic. Various studies have investigated what factors contribute to the decision of people to use bike-sharing as means of transportation (Bachand-Marleau et al. 2012; Tran et al. 2015). A survey was conducted 2020 in Montreal, to determine what are the factors that encouraged individuals' willingness to use the shared bicycle system and the elements that influenced the frequency of use. The factor found to have the greatest effect on the likelihood for use of a shared bicycle system was the proximity of the home to docking stations (Bachand-Marleau et al. 2012). Factors contributing to the increased usage of bicycle-sharing in Montreal are identified further for providing recommendations of station size and location decisions. The early studies adopted linear models. With the innovation of shared transportation modes, researches on other shared micro-mobility was conducted. For instance, free-floating e-scooter ridesharing became a popular mobility trend. Because the journey was reserved only and it was digitally traceable, Degele analyzed the customer needs and motivations based on large datasets and got four customer clusters in 2018. One research investigated the factors affecting mode choice behavior with a focus on bike-sharing and explores the effectiveness of different policy options aiming at increasing bike-sharing ridership (Li and Maria 2018). Reck et al. applied a method to simulate the mode choice models among four different shared micro-mobility services by trip attributes (such as time of day, the elevation difference between origin and destination, trip distance, and so on) and new proposed elements: available vehicles surrounded with start point, which is recognized as important presentation of competition between different modes. In Zurich, a study focused on shared e-bike mode (offered by Smide company) with its key drivers of demand for free-floating e-bike-sharing, such as economic and social activities, quality of public transportation service, and availability of bike infrastructures (Guidon and Becker et al. 2019).

With the rise of shared micro-mobility, research shifted towards factors determining the use of shared micro-mobility services with other conventional traffic mode or between different micro-mobility modes. In a study of modeling the choices of travel mode between BSS and taxi, 13 models which are popular linear, nonlinear (not ensembles), and ensemble algorithms are used to model the preference, and the Random Forest model produced the best results compared with other methods. It also has a relatively high processing speed among these machine learning models (Zhou et al. 2019). The spatial-temporal factors that influence travel choices between these two alternative modes are explored. The difference in travel characteristics between docked and dockless bike-sharing is revealed, including riding distance, riding time, usage frequency, and spatial-temporal patterns by adopting geographically weighted regression (Ma et al. 2020). One of the newest studies compared the four frequently used shared micro-mobility using widely accessible vehicle location data, and then by using multinomial logit (MNL) model and normal error component logit-mixture model (NECLM), the joint effects of all attributes and competition effects between the different modes are revealed(Reck et al. 2021).

2.5. Research gap

Most of the aforementioned studies have separately examined different aspects of travel mode choices based on one different data source and analysis model. Although some previous studies have explored the factors determining travel mode choice between shared micro-mobility services, most of them have used surveys and questionnaires as data sources, and regression models as analysis methods (Ma et al. 2020; Reck et al. 2021; Zhou et al. 2019). However, none of the studies have adopted machine learning methods to explore the influence factors considering external impacts and trip attributes on mode choice among multiple shared micro-mobility services.

This study pioneers to address this gap by firstly exploring the GPS data crawled from open APIs of shared services providing companies to build the feature data set. This followed by applying four frequently used machine learning methods on the generated data set to explore the spatiotemporal influence factors of travel mode choice between four travel modes (docked bike, docked e-bike, dockless e-bike, dockless e-scooter) over the same spatial-temporal dimension. The performance of each ML method is compared as well.

3. Methodology

3.1. Overall Research Framework

The overall study follows a framework as presented in Figure 1. First, an attempt is made to collect GPS data on the four shared micro-mobility services as well as weather data, DEM data, and so on, which are potential factors influencing travel mode choice. These data are used to build the choice dataset to identify different travel modes. Second, split the dataset as a training dataset and a testing dataset by a certain ratio for model training and accuracy validation. Third, the important influence factors are found out for each method, and how they affect users' choices are analyzed. At the last, the potential measures that may improve the utility of each service are discussed.

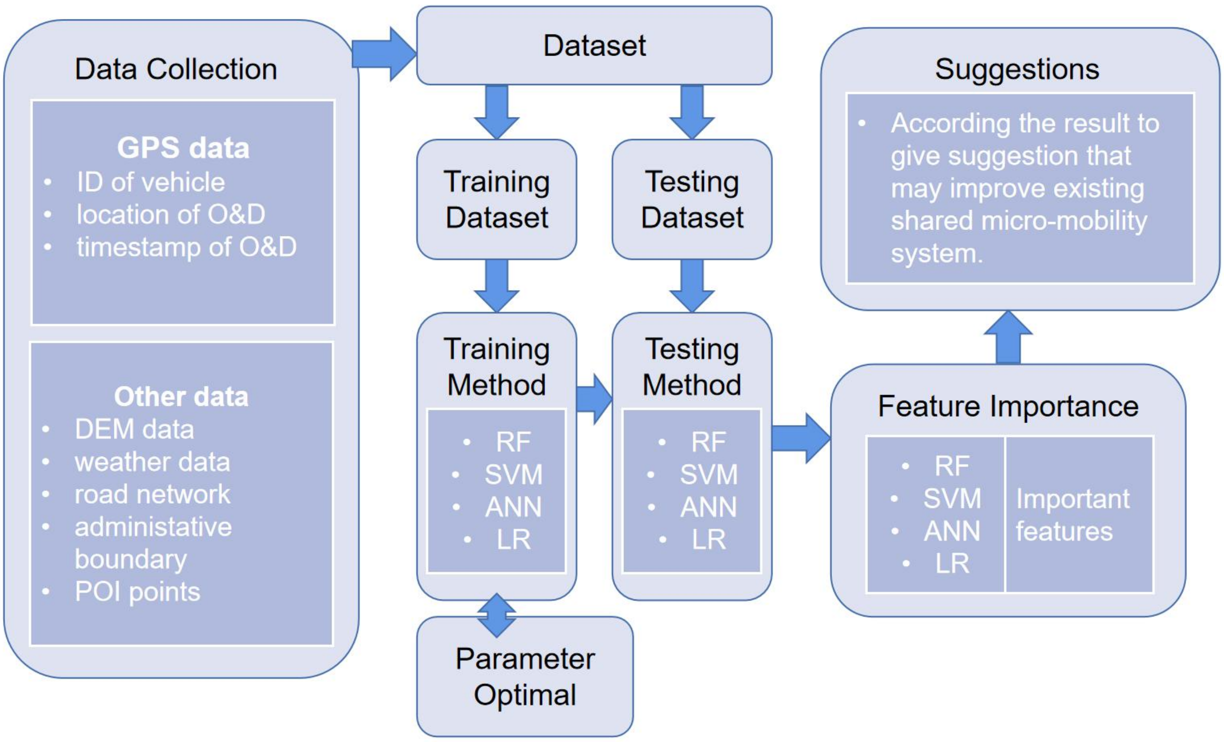


Figure 1 The overall framework of this study

3.2. Data Preparation

3.2.1. Data Collection

The data is collected in Zurich, which is the capital and the largest city of Switzerland, with 4.2 million residents in 2019 (Federal statistical office, 2021). The base map and location of the study area are presented in Figure 2. The main modes of transport in Zurich have been public transportation (46.9%) and private motorized transport (40.6%) in 2019 (Federal statistical office, 2021). There is 340 km of cycling lanes and tracks in Zurich, which provides shared micro-mobility a large development possibility. The main micro-mobility service providers in Zurich are listed in Table 1.

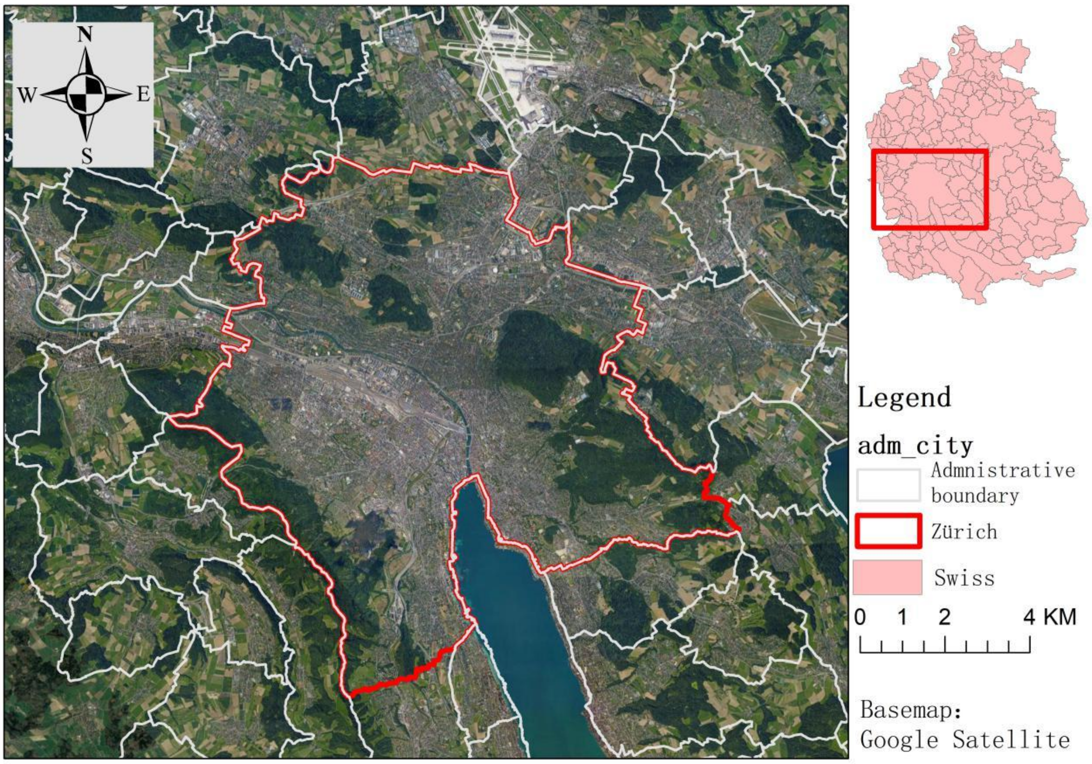


Figure 2 Overview map of study area. Based on data Swiss Admnidtative Boundary.

Table 1 Companies which provide different type of shared micro-mobility

Service	Companies
Docked bike	Publibike
Docked e-bike	Publibike
Doecless e-bike	Bond
E-scooter	Lime, Bird, Tier, Voi and Circ

The records of shared micro-mobility use are crawling from open-accessible APIs of service provider companies in Switzerland between Feb 1 to Feb 29, 2020. The raw dataset includes the vehicle location data which are offered by shared micro-mobility provided companies in Zurich, Switzerland. Each company provides information on one type of shared service. After the preprocessing, the dataset contains 63132 records. Each record represents a trip, contains the information of id and type of the vehicle, the start/end location and time, the length, duration, and average speed of this trip. The other features will be generated based on the raw dataset. The sample size of each service for analysis is shown in Table 2:

Table 2 The sample size for each shared micro-mobility

Service	Sample size	Percentage
Docked bike	9286	14.71%
Docked e-bike	25808	40.88%
Dockless e-bike	5083	8.05%
E-scooter	22954	36.36%

According to the trip data, the docked e-bike has the most records which occupy almost 41 percent of all datasets, then trips of e-scooter have second large records in this dataset, while dockless e-bike has least trip data, with only 8.05 percent of all.

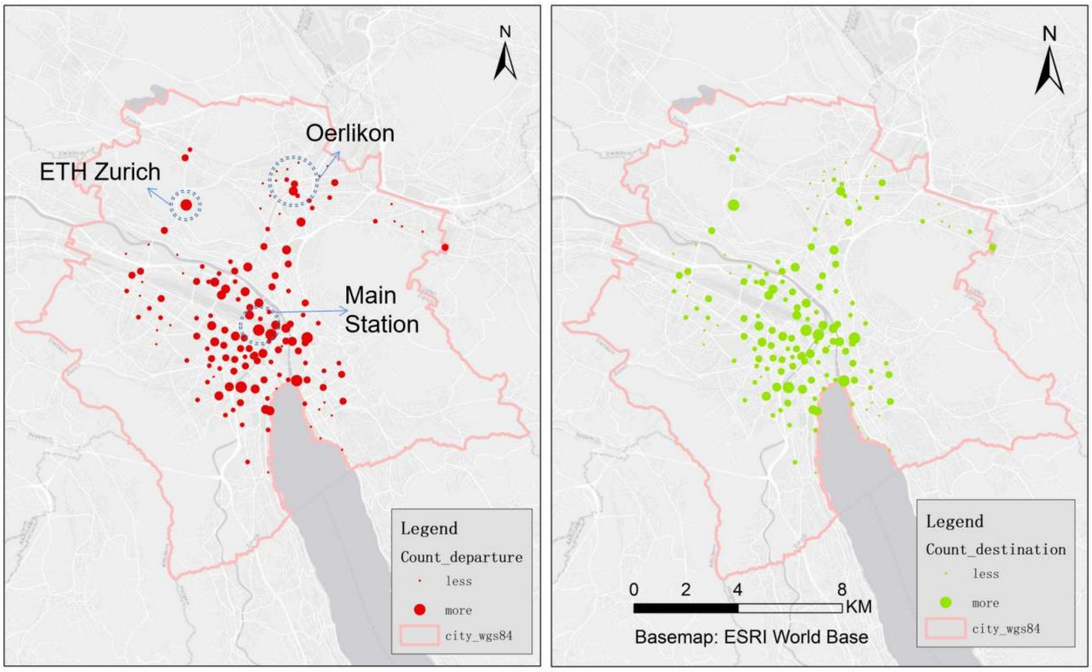
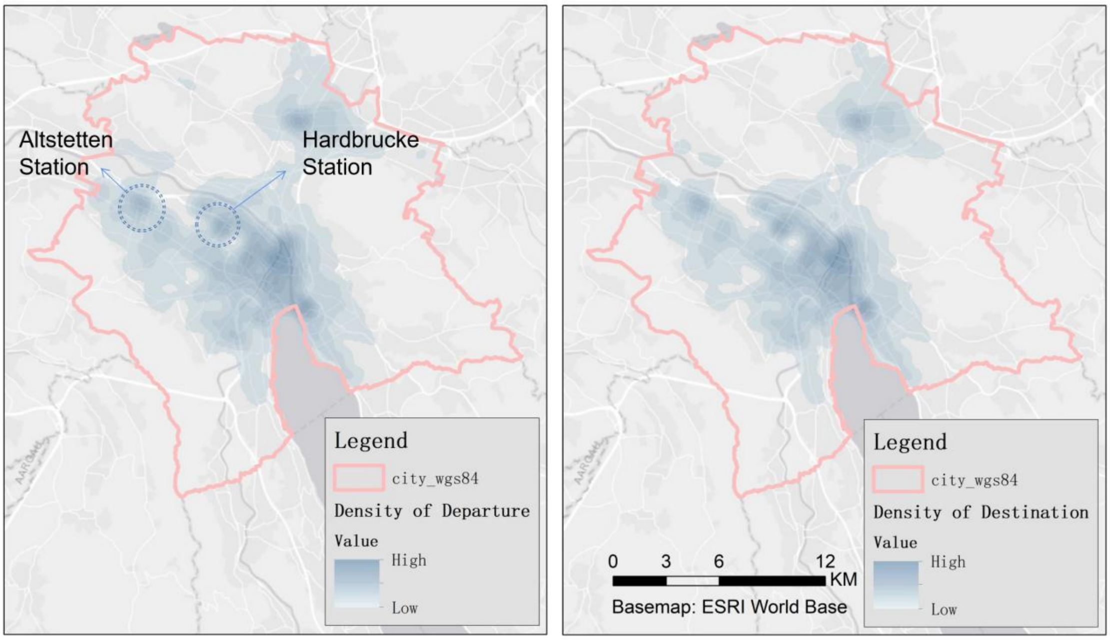


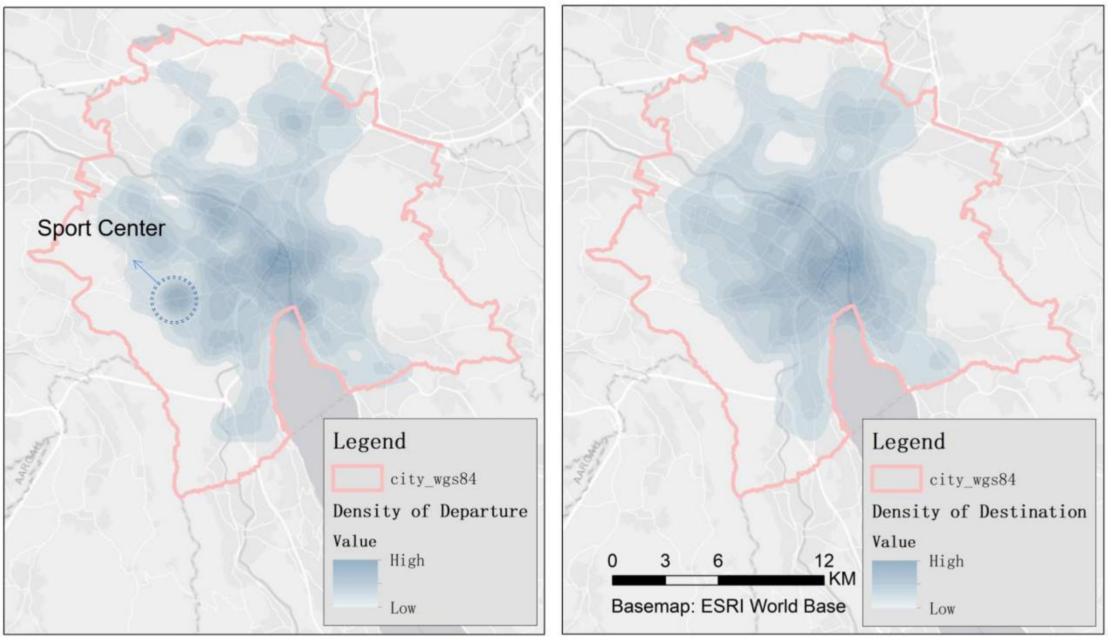
Figure 3 Docked sites and flows, presenting the flow size of each docked site. Bigger dots mean larger flows and verse visa. Based on data docked trip records.

The docked vehicles have permanent positions of start and endpoints, where users can rent a vehicle, and it is mandatory to bring back the vehicle to the same site or another site. The spatial pattern of origins and destinations of docked vehicles is visualized in the figure above (Figure 3), the size of symbols indicates the flux of each site. According to the figure, most docked stations are close to

Zurich main station, which usually have a higher flow. Oerlikon district and ETH Zurich are the other places with relatively concentrated docked vehicles flow.



A) E-scooter trips



B) Dockless e-bike trips

Figure 4 Heatmap of departure points and destination points of dockless services, based on data location of departure and destination points. Area with darker color is where the records of departure or destination points are denser.

From the map of the e-scooter (Figure 4.A), it is obvious that the e-scooter has a low utility rate around ETH Zurich University. Instead, except the downtown area around the main station, the e-scooter also shows two slightly dense areas around Altstetten station and Hardbrucke station. Moreover, the dockless e-bike trip heat map (Figure 4.B) presents a crowded gather of departure heatmap on the sports center.

The variables that enter into the analysis include the trip attributes, geographical and climatic factors, these factors are listed in Table 3. The trip attributes include trip duration, trip distance, the average speed of the trip, occurrence time of the trip (involving week/weekend and hour in a day). Geographic and climate data composed of DEM, road density, nearby POI (point of interest) types of destination, temperature, weather, humidity, wind speed. A detailed illustration of the final dataset with a description of each feature is shown in section 3.4.

Table 3 Analyzed influence factors

Geographic and climate factors	Weather	Temperature, weather, wind speed, humidity (based on the start time)
	Terrain	Elevation difference, road density difference, (between departure point and destination point)
Trip attributes	temporal	record day (weekday\weekend), Hour in a day, duration time of trip,
	spatial	distance of the trip, the POI type beside destination of each trip, residential density

In the next step, according to the hypothetical influence elements above, the relevant data used in this paper is collected from official Geoportals. The format and resources of used data are listed in Table 4.

Table 4 The used data in this study

Data	Format	Resource
Digital Elevation Model	Raster data	ALOS Data(12.5m) https://search.asf.alaska.edu/#/
POI points Apartment density Road network	Vector data	Zurich GIS Browser (0.01m) https://maps.zh.ch/
Bus stops	Vector data	https://opentransportdata.swiss/
Climate data	Table	

3.2.2. Data Processing

This section introduces each feature processing procedure and the supportive reason. The influence factors are generated from data in Table 4. The data are processed by ArcGIS and Python. According to the Table 3, there are in total 20 influence factors generated from vehicle trip data as the final dataset for travel mode choice classification.

Firstly, the road density is calculated by fishnet grid which is generated by the computer with 500m. Then the dissimilarity of DEM and road density between origin and destination is calculated and then join to the dataset as two features. They represent if the trip is flat or not and if the trip is towards the downtown district or remote region individually, to some extent. The number of apartment is counted by fishnet with 100m, which could prove whether the trip starts at residential area.

Then, for the POI-relevant factors, there are 56 sub-types (such as club, restaurant, crossing, school, university and so on) of POI data in the raw POI dataset, including more than 4 thousand points. These are summarized as 8 types in the end, as 8 influence factors (including food&beverage, public facility, medical service, tourist attraction, transportation service, road furniture, education service). Then, the number of each type of POI is summarized by a 100 meter buffer of end point for each trip as the value of influence factor. These factors might indicate the purpose of trip to some extent.

After processing all of the data, the dataset for machine learning is prepared, which contains a totally of 20 influence factors: duration (s), length(m), average speed (km/h), difference of road density (km/km2), DEM (m) at start and end points; temperature (℃), weather (weather includes 9 different types, which is represented by numbers, Table 5), wind speed (km/h), humidity (%), and surrounding POI types at endpoints; the occurring day (weekday is coded as 0, the weekend is coded as 1) and occurring time (in which hour of the day, maximum: 24).

Table 5 weather type and codes

Weather type	Code
Quite cool	0
Scattered clouds	1
Cool	2
Passing clouds	3
Partly sunny	4
Mild	5
Chilly	6
Refreshingly cool	7
Overcast	8

3.2.3. Feature Attributes

The entire dataset contains 21 variables for four travel modes with more than 63K records, which compose 20 independent variables and 1 dependent variable. The scope, unit, and the respective variable's type of the used features which are used for predictors in models are presented in Table 6. Nominal variables are encoded as dummy variables, such as POI types, services type, and weather types that described in section 4.2. When the training set and testing set are split, stratified sampling is used to make sure that the testing set has the same ratio of four services as the entire data set. The data set is separated as a training set and test set with the ratio 0.7:0.3.

Table 6 The type, coding, unit of each influence factor

Variables	Type	Coding/Scope
Duration of the trip (s)	Interval	[61,7199]
Length of the trip (m)	Interval	[101,8447]
Average speed of the trip (km/h)	Interval	[0.05,44.94]
Road Density Difference (m/m^2)	Interval	[-0.04, 0.04]
Elevation Difference (m)	Interval	[-203,211]
Temperature (℃)	Interval	[-3,18]
Weather type	Nominal	{0,1,2⋯7,8}
windSpeed (km/h)	Interval	[0,56]
Humidity (%)	Ratio	[0.31,0.98]
Week	Nominal	{0,1}
Hour	Ordinal	{0,1,2,3⋯22，23}
Apartment	Interval	[0,293]
Food&beverage	Interval	[0,26]
Public facility	Interval	[0,13]
Medical service	Interval	[0,4]
Accommodation service	Interval	[0,4]
Tourist attraction	Interval	[0,3]
Transportation service	Interval	[0,10]
Road furniture	Interval	[0,16]
Education service	Interval	[0,2]
Micro-mobility type	Nominal	[1,2,3,4]

3.3. Model

3.3.1. Model description

In this paper, identifying peoples' travel mode choice could be recognized as a multi-classification problem. To achieve the research aim, except the logit model (the most frequently used model in travel mode choice topic), this project is also suitable to use a supervised multi-classification algorithm to mimic individuals' choices under different situations. ML classification investigations consist of two main processes: model development and model evaluation. This section discusses four

ML methods (random forest, support vector machine, artificial neural network, and logistic regression) used for modeling travel mode choice. It includes an overview of the basic conceptual part and how the models are developed.

One of the used models is Random Forest (RF), an algorithm based on multiple single DT (Decision Tree), which could be used for both classification problems and regression problems. This model is widely used in the data mining field. Compared with regression modeling, DT has no specific requirement of data format, which means both numerical and nominal types of variables could be explored (Yang and Zhou 2020). Random forest is an ensemble learning model with trees as the basic unit. Each tree will get a classification result and vote, and the forest finally chooses the category with the most votes as the final result. This model will take the dominant classified results for each classifier as the final class. It has the advantage of stability and only predicts incorrectly when half of the classifiers give wrong results (Cheng et al. 2019). In this model, the random feature is embodied in that the samples and features which are used to train a single DT are generated by random selection. Among them, the sample selection is the most random selection with replacement. The advantage of random extraction is that it greatly avoids the over-fitting problem caused by the high similarity between decision trees.

The second model is SVM (Support Vector Machine), a non-probabilistic binary linear classifier, which confirms the decision boundary between classes by finding the maximum margin hyperplane that is solved for the learning sample. This decision boundary could capture the general pattern in the training set, which has a high chance of performing well on the test set. The theoretical basis of the SVM method is non-linear mapping. The inner product kernel function are used to substitute the non-linear mapping to the high-dimensional space in order to obtain the best hyperplane. SVM was optimized to adapt multi-classification questions. One is pairwise classification, which constructs a binary SVM between every two classes. This method is called one-against-one (ovo) which is more suitable for practical application. Another is one-against-all(ovr), which uses binary SVM to classify one picked class with all of the rest classes as second class. Using SVM could avoid “dimension disaster” to some extent, which means the complexity of calculation is independent on the dimension of sample space but the number of support vectors.

The Artificial Neural Network (ANN) provides a universal and practical way to learn values from examples as real numbers, discrete values, or vector functions. An artificial neural network is composed of a series of simple connected units. In this network, each unit has a certain number of inputs and produces a single output. To solve more linearly inseparable data, in reality, a three-layer neural network should be built, including the input layer, hidden layer, and output layer. The number of neural nodes in the input layer (the first layer) is determined by the dimensionality of the input data. In this study, there are 20 influence factors, which means the data dimension is 20, and the input has 20 neuron nodes. In the same way, the number of nodes in the output layer (third layer) is determined by our classification number. There are four types of services, so the output layer has four neurons. According to the actual situation, the number of hidden layers and the number of nodes in each hidden layer could be adjusted. When there are more hidden layers and more hidden layer nodes, more complex data models can be processed, but with greater cost. The benefits of adding more levels are more in-depth representation features and stronger function simulation capabilities. Currently, there is no sound theory to guide the decision of hidden layers, generally set based on experience. ANN model usually uses the activation function, such as sigmoid function, tanh function, and so on. It is feasible to introduce nonlinear factors to the neuron so that the neural network can approach any nonlinear function arbitrarily. Therefore, the neural network can be used in many nonlinear models. The advantage of ANN is that it has a strong anti-noise ability and allows errors in training data. But this model usually needs a longer time to train. Neural networks can be used to solve classification problems as well as regression problems.

The last machine learning model, logistic regression (LR) is a classification algorithm, although it is called “regression”. For logistic regression, the two most prominent merits are the simplicity of the model and the strong interpretability of the model. This model can not only predict within the sample but also predict data outside the sample. At the same time, this simple model may easily lead to underfitting and inaccurate classification results compared with other complex models. Meanwhile, when the data set contains error points, the corresponding error of using regression will also be large. When using logistic regression to solve multi-category classification problems, similar to SVM, a classification algorithm called "one-vs-all" can be used. Finally, when making predictions, all the classifiers should be run, and then for each group of the input variable, the most likely output variable should be selected.

3.3.2. Method Development

In this study, the machine learning models are mainly achieved by Python, Scikit-learn package. The purpose of the model development process is to minimize the generalization error of each model.

Whether a model is applicable or how effective depends on the setting of hyperparameters to a large extent. By tuning the hyperparameters, the model can be optimized and its performance can be improved. Normally, the optimal value of model hyperparameters is found by the rule of thumb or through trial and error.

When the model performs poorly on unknown data (test set or out-of-bag data), it can be concerning that the generalization degree of the model is not good enough, and the generalization error is large. The generalization error is affected by the structure (complexity) of the model. The relationship between generalization error and model complexity could be accurately depicted in Figure 5. When the model structure is excessively complex, the model will overfit and the generalization ability is not enough, so the generalization error is large which indicates the worse prediction ability of model. When the model is too simple, the model will be under-fitting, and the fitting ability will be insufficient, so the error will be large as well. The attempt to find the best equilibrium could be seen as the process of model adjusting.

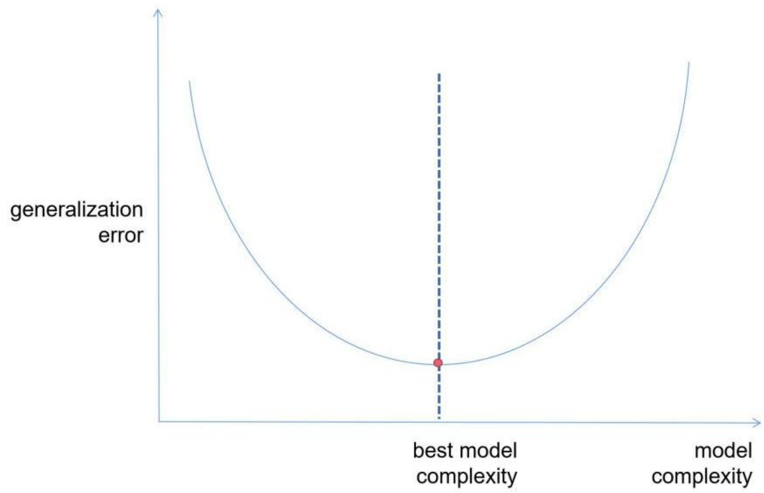


Figure 5 The relationship between model complexity and generalization, the best model complexity is where the generalization error is lowest.

There are four ML methods are selected, RF, SVM, ANN, and LR, which were used in similar researches and returned good results. Technically, the RF method is belonging to ensemble learning (Tin 1995), which combines multiple weak classifiers and gives a vote result in the end. Based on the theory, RF method may return a proper result of this study. But there will be differences in practice because of other elements, like data size, the number of variables and so on. Not that a certain method would be optimal for all application problems. The basic principle for choosing a classification algorithm is its accuracy. Therefore, it is necessary to compare several ML methods and find the proper one.

The random forest combines many trees, and it makes a prediction based on the most voting results. For the RF model, the number of the tree is represented by parameter: n_estimators. The number of trees in the forest versus the model performance (cross_val_score) is plotted in Figure 6. I found that once the tree number exceeded 50, the performance became stable by 10 trees as an interval. Although the best performance is generated by the most tree (200), the difference in error between 50 and 200 trees is extremely small. Therefore, considering the time-consuming, 50 trees are selected as one parameter in the final model.

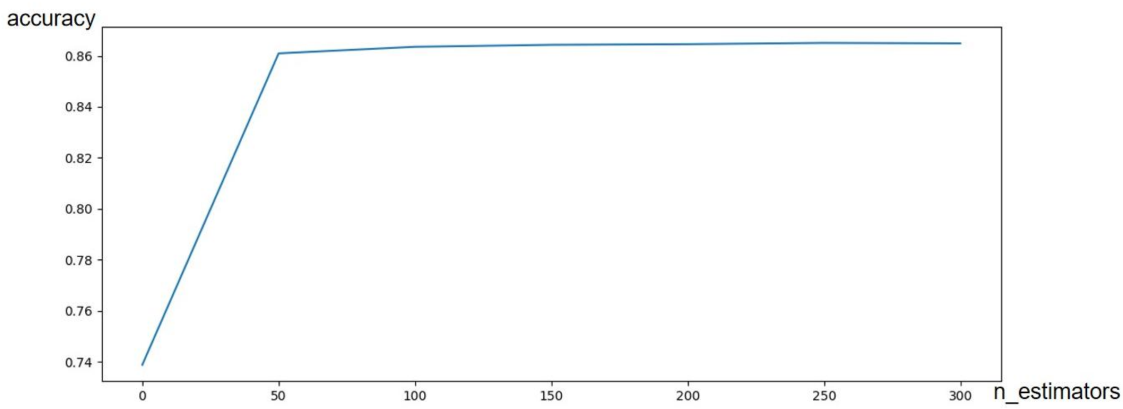


Figure 6 Change of accuracy score when number of estimator is enlarging for Random Forest method.

For the other parameters, the same approach is used to find the optimal value and build the final model. But the accuracy does not have an obvious increase for any attempt of parameters. This situation indicates that the model may almost achieve the best complexity after adjusting the n_estimators. GridSearchCV is also used to find the best combination of hyperparameters, which gives the similar results. After all of the parameters optimal, the accuracy of the model on the test set arrived at 0.72 and peaked up at 0.82 on the training set.

The GirdsearchCV is a method of hyperparameter optimal. First, a parameter space with possible candidate hyperparameter is set. Then each combination of the hyperparameters are scored by cross validation. In the end, the best combination with the highest accuracy is returned and recommended. GridSearchCV can guarantee to find the most accurate parameters within the specified parameter range, but this is also the flaw of grid search. It requires traversing all possible parameter combinations, which is very time-consuming in the face of large data sets and multiple parameters. Meanwhile, there is also probability that the best hyperparameter is not involved in parameter space, which could mislead modeller.

The classification result of the SVM model is highly relative to the selected kernel function and the penalty coefficient C. But how to choose the appropriate kernel function so far there is no more systematic method, only by constantly trying each kernel function. The most suitable kernel function can be found by comparing the SVM results of different kernel functions. The most commonly used kernel function contains linear, poly, RBF, sigmoid, and so on. In this study, poly, RBF, and sigmoid are used for finding the best function. The value of C is used to set the error scope in the model. A higher C value means fewer errors that are allowed to exist in the training dataset. Cross-validation is used to find the best C value. For different kernel functions, different hyperparameters affect model accuracy. For example, gamma is a significant parameter for RBF (Radial Basis Function) function. Increasing the gamma value will make the range of influence of each instance smaller and the decision boundary more irregular while decreasing the gamma value will make the influence range of each instance larger and the decision boundary flatter. The GridsearchCV is adopted, which gives the best combination of hyperparameters for the model. As a result, the final SVM model with an RBF kernel is generated, with “one-against-one” decision function, 0.05 as gamma value, and 5 as C value.

.

Multi-layers perception (MLP) is the used ANN model in this study. One of the most important parameters is the solver, which provides the neural network solution method to help it find the global minima by gradient descent for the model. There are three solution methods (such as “lbfgs”, “adam”, “sgd”) with different characteristics, which are used to optimize the weights. According to the real situation, the one with the best performance on classification accuracy. The adam solver performs better when the data size is relatively smaller, while adam’s method is more robust, and sgd will have the best performance when the parameters are adjusted better. Meanwhile, the number of hidden layers and the number of nodes in each layer also determinant the performance of a model to a large extent. A model with zero hidden layers can solve linearly separable problems. Thus, a non-linear problem usually uses at least one hidden layer. The GridSearchCV method is applied to choose the best combination of hyper-parameters to a given extent. Here, the adam solver is selected in the final MLP model, with a single hidden layer and 100 nodes in this layer.

The last method chosen for prediction is Logistic Regression (LR). The solver parameter determines the optimization method of LR loss function. There are four choices, liblinear, lbfgs, newton-cg, and sag, which are candidates in hyperparameter space. Because of the imbalanced dataset, the class_weight is set as “balanced”. Same with the SVM method, there are two classification methods, multinomial and one-against-rest. These perform same on binary classification problem. But since this study is a multi-classification problem, it is necessary to try both and find the one with better results. The GridSearchCV method is applied to choose the best combination of hyper-parameters to a given extent. In order to find a ideal combination of hyperparameters, GridSearchCV is again used here. As a result, the liblinear solver is selected in the final LR model, with a one against rest classification method.

4. Results

The results of this study are structured as follows: First, the model results are compared on their performance, their merits and drawbacks found in this study are summarized. This is followed by analyzing the impact of the important influence factors explained in section 3 on the intention to use the docked bike, docked e-bike, dockless e-bike, or e-scooter. This is done by running permutation importance for each model that is calculated based on a sub-set of the data (the test data set). Third, partial dependence is used to analyze the relationship between important variables and decisions.

4.1. Comparison of models

For RF, SVM, ANN, and LR models, this paper established these models by Python and mainly used scikit-learn package. During model training, in this paper, a 5-fold cross-validation method is adopted, the training data set is evenly split into equal 5 subsets, where any four subsets serve as training sets for the model, the remaining one is used as a test set to measure the prediction accuracy of the model for the training set. Then the accuracy of the testing set is calculated. By randomly sorting and extracting data from data sets, the scientific nature and consistency of the data used in models are ensured and the results are reasonable and explanatory.

In the model evaluation process, the modeler estimates the generalization error of the model, typically by testing on an out-of-sample test-set, which is in this study the testing dataset split randomly before training models. The model adjusting process is detailed described in section 3.6. Here is only the presentation of the final results.

To validate the classified results, the following accuracy parameters are logged: Overall Accuracy (OA), Recall, Precision and F1-score. These parameters are directly calculated from the classification confusion matrix and could be adapted to all models. OA means the probability of correct prediction for all samples. It is defined as the ratio of the total number of samples, correctly classified (on the diagonal of confusion matrix). However, this norm sometimes is not enough to generalize the quality of a model when the samples are not equal. When the records of each mode are not equal, the classification result is easily inclined towards the classes with more samples by classifying more to these classes, and sacrificing the accuracy of the class with small samples. Therefore, Recall, Precision and F1-score is attached as evaluation indicators to consummate understanding of each model. Compared with OA, Precision and Recall is needed to be computed for each class separately in stead of measuring the global sample prediction situation. The Precision is calculated by dividing the number of correctly classified samples by the number of all such predictions, while Recall is the number of the correctly classified samples divided by the number of actual class. F1-score is defined as a harmonic average of precision and recall. F1 could be seen as a comprehensive indicator which integrates Precision and Recall to evaluate the different ML methods on each class.

According to the OA of each ML method (Table 8), RF gives the best classification result both on testing set and training set, especially its performance on the training set with 82 percent correct rate, while the LR method gives the worst results with the lowest OA (0.56) on both training set and testing set. SVM and ANN methods built in this study performs similarly with around 68% and 67% overall accuracy on training set, as well as 65% on testing set.

Table 8 The assessment matrix of RF, SVM, ANN, LR model on classification of shared micro-mobility

		RF	SVM	ANN(MLP)	LR
Overall	Testing set	0.72	0.65	0.65	0.56
Accuracy	Training set	0.82	0.68	0.67	0.56
F1	Docked bike	0.06	0.24	0.15	0.12
	Docked e-bike	0.74	0.69	0.70	0.61
	Dockless e-bike	0.14	0.04	0.25	0.12
	E-scooter	0.84	0.77	0.77	0.67
Presicion	Docked bike	0.75	0.37	0.38	0.27
	Docked e-bike	0.62	0.57	0.62	0.60
	Dockless e-bike	0.79	0.13	0.51	0.16
	E-scooter	0.83	0.84	0.73	0.59
Recall	Docked bike	0.03	0.18	0.09	0.08
	Docked e-bike	0.93	0.85	0.82	0.63
	Dockless e-bike	0.08	0.02	0.16	0.09
	E-scooter	0.85	0.71	0.81	0.79

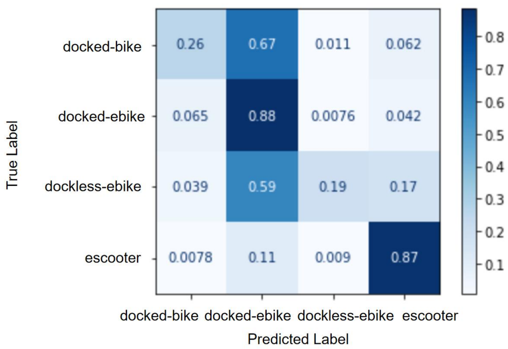
The RF model proves satisfactory. The result of classification between shared micro-mobility services is shown in Table 8. Compared with the results of previous studies (Bucher 2020; Liu and

Luo 2018; Zhou et al. 2019), the accuracy of classification on the training set is relatively high, and the model performs above-average level on the testing set.

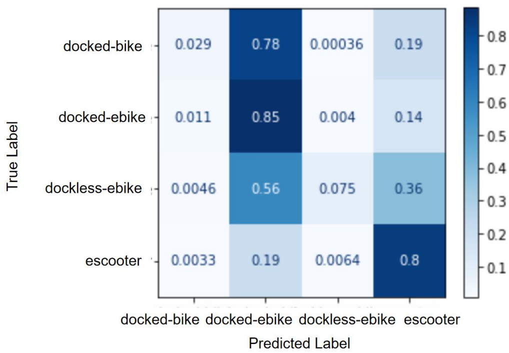
The above mentioned classification accuracy scores refer to the correct percentage of all categories. Classification accuracy is an easy measure of classifier, but it does not tell the potential distribution of response values, and the type of classifier that makes mistakes. Especially for the skewed dataset, accuracy could not be the primary performance indicator.

Therefore, Precision, Recall, and F1-score are crucial evidence of classification results considering the model performance on each class as well. The results show how the classification models perform on four modes individually. Precision calculates how many predicted samples are correctly predicted, while recall describes the ratio of samples that are correctly predicted for the original actual samples. F1-score could be regarded as a weighted average of model precision and recall.

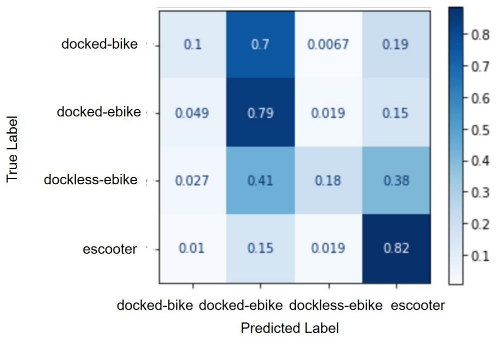
All models have the worst classification accuracy on identifying the dockless e-bike mode, which is the model with the least number of records, especially on the SVM model. It may result in an unequal proportion of each mode in the data set. Which modes are more difficult to be distinguished by different models could be discovered from confusion matric as well (Figure 7).



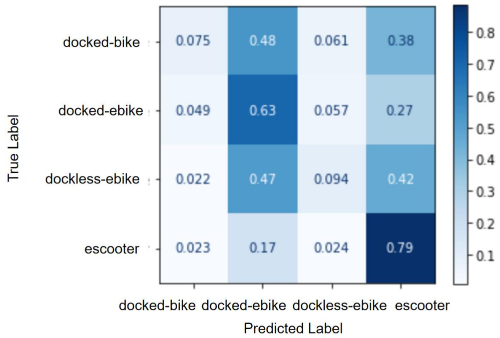
A) Random Forest



B) Support Vector Machine



C) Multi-layers Perception



D) Logistics Regression

Figure 7 Confusion matrix of travel model choices classification for the four machine learning methods

According to the confusion matrix of RF model, docked e-bike and e-scooter mode have the highest accuracy on prediction. The dockless e-bike has a significantly high possibility (59%) to be identified as docked e-bike. Meanwhile, 17% of dockless e-bike are classified as escooter, while only 19% actual samples are correctly classified. Moreover, 67% docked bike are predicted as docked e-bike incorrectly, with 26% identified correctly. Confusion matrix of these four methods perform similarly on a global view, since the identification of docked bike and dockless e-bike mode are quite poor. Docked bike and dockless e-bike mode are easily identified as docked bike mode by RF, whereas the classification of escooter by RF method is more reliable than the other methods. Moreover, MLP and SVM methods may confuse the part of other mode of samples as escooter. For the LR model, the prediction accuracy of modes with less samples are sacrificed for keep a high OA value, but it is still unsatisfactory. There still a lot percentage of error classified samples.

4.2. Importance of influence factors

Feature importance is a difficult concept to define in general, because the importance of each factor may be due to its (possibly complex) relationship with other factors. There are various approaches to calculate the importance of features for different models. For instance, the random forest algorithm could estimate the importance of a factor by looking at how much prediction error increases when (OOB) data for that factor is permuted while all others are left unchanged (Liaw and Wiener 2001). Importance of influence factors of logit regression could be shown by coefficients, it not only presents the significance of each feature but also indicates if the factor influences classification negatively or positively (Zhao et al. 2020).

However, this study intends to analyze the difference between methods, the important factor for each model are compared in the following steps. To keep the consistency of methodology, the Permutation Importance (PI) method (Permutation Importance 2021) is used to measure the feature importance of each model.

The process of calculating the feature importance is as follows: First, change the arrangement of the data in a column of the data table, keep the rest of the features intact, and then see how much influence it has on the prediction accuracy. If the results of the model change a lot, indicating that the feature is more important, and verse visa. The more important factors of each model in the top 7 are listed in Table 9.

Table 9 The Important factors of RF, SVM, ANN and LR model on shared micro-mobility modes classification

Rank	RF	SVM	ANN	LR
1	Duration(s)	Length(m)	Duration(s)	Length(m)
2	Length(m)	Public facility	Public facility	Duration(s)
3	Public facility	Speed(km/h)	Speed(km/h)	Speed(km/h)
4	DEM difference	Medical service	Transportation service	Public facility
5	Education service	Temperature(℃)	Road furniture	Education service
6	Speed(km/h)	Transportation service	Medical service	Transportation service
7	Food&beverage	Duration(s)	Education service	Accommodation service
…	…	…	…	…

To sum up, RF, SVM, ANN, and LR method have different ranks of important features. The situation of trip attributes have a strong impact on travel mode choice, although not all of trip attributes factors are ranked in first three. The surrounded POI points nearby orientation also have a great impact on model predictions. According to Table 9, Public facilities seem to be the most important type of POI that affects people’s choice, followed by Education services, Transportation services, and Food&beverage class. By contrast, weather conditions are less crucial for models.

4.3. Explanation of Model

Considering the relatively higher overall accuracy of the RF model, the relationship between influence factors with mode choice is presented based on the RF model. According to the rank of important features for the RF model, the most important factors are ranked as following: trip duration and distance, Public facilities surround endpoint, DEM difference (between departure point and destination), Education service surround endpoint, average speed of the trip, Food&beverage points surround endpoint, difference of road density (between departure and destination), Transportation points surround endpoint, and time of the day (hour). By contrast, the other weather condition and other POI types (Medical facility, Accommodation service, Tourist attraction and Road furniture) play a less crucial role in travel mode choice between shared micro-mobility services in the RF model.

Compared with feature importance which refers to the influence of a certain feature on model prediction, Partial Dependence Plots (PDPs) are used to reflect how these features affect the prediction. The working process of PDP is similar to permutation importance, Partial Dependence Plots are also performed on a well-trained (fitted) model. The value of a certain feature changes many times to produce a series of prediction results. Then the influence of each feature on the model results is produced. Following plots show the bivariate relationships between the prediction probability (i.e., the average likelihood of choosing a particular travel mode over others) for each mode, and each important factors.

The pdpbox package is mainly used for presentation of partial dependent relationship in this section. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter in the following PDP plots.

The duration of each trip is the most important feature in the RF model. Figure 8 presents how the trip duration change affects users' mode choice. Apparently, there is a conspicuous tendency that escooter (class 3) has obvious superiority on the trip with a short period at time, especially when the duration is shorter than approximately 400 seconds. With longer duration, the possibility of choosing escooter as transportation tool is decreasing rapidly with little fluctuates and keep negative when it is longer than 1,000 seconds. By contrast, the possibility of taking other modes for a trip is rapidly increasing when the trip is longer than around 400 seconds. Furthermore, for the trip with longer periods at time, docked e-bike (class 1) and dockless e-bike (class 2) are preferentially chosen compared with docked bike (class 0).

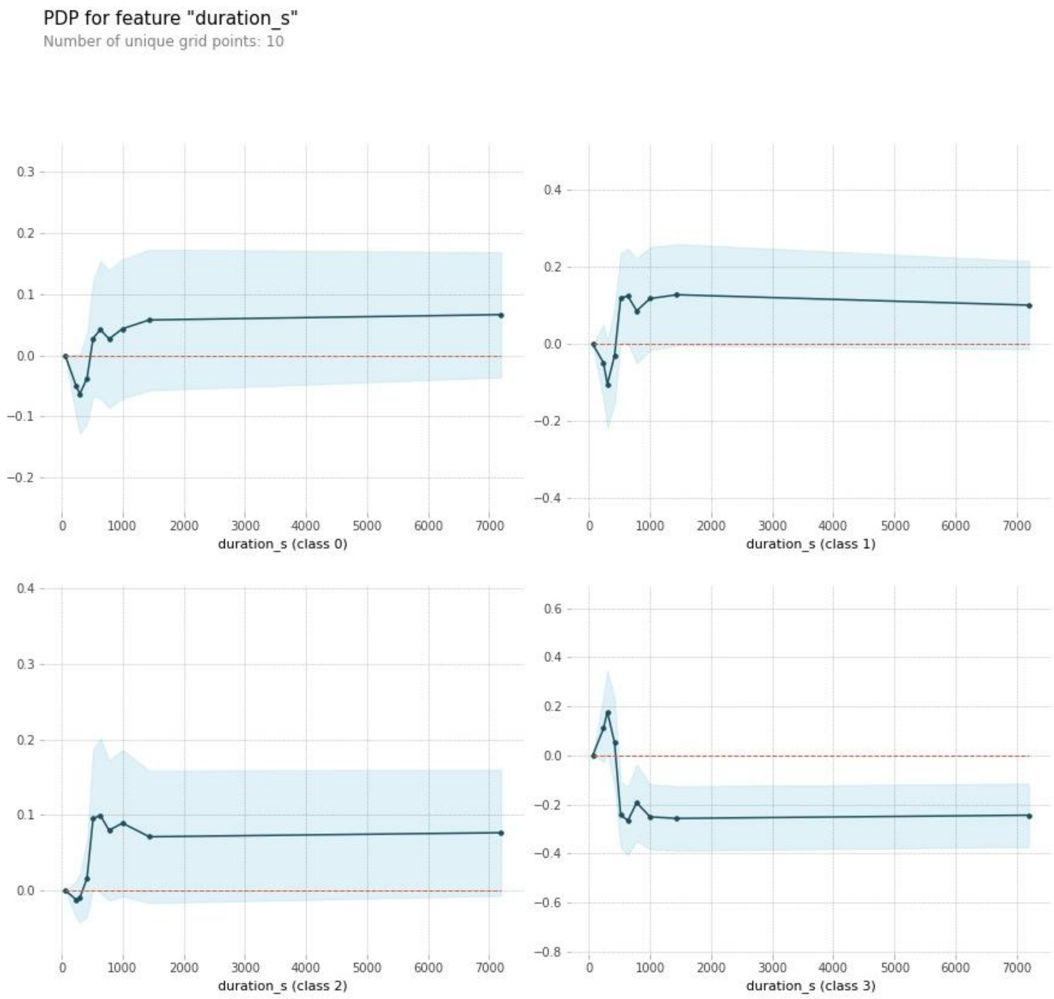


Figure 8 the partial dependent relationship between mode choice and trip duration based on RF method. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

The plot by trip distance (Figure 9) presents that with distance getting longer, docked services are more preferable for users, which docked e-bike (class 1) shows more tropesis. Until the distance longer than 3,000m, this trend does not rise further. By contrast, the trend of choosing an e-scooter (class 3) is declining until the distance longer than around 3,000m. When the trip distance is longer than 5,000m, the probability of choosing a dockless e-bike (class 2) is positive, and increasing slightly with a longer distance, while a shorter distance results in a negative impact on choosing a dockless e-bike between around 500m to 5,000m (class 2). This may indicate that a dockless e-bike (class 2) is more probably chosen when passengers are going to have a relatively long trip (more than 5,000m). There is an unexpected appearance that dockless e-bike has a positive probability to be chosen when the trip length is less than around 500m.

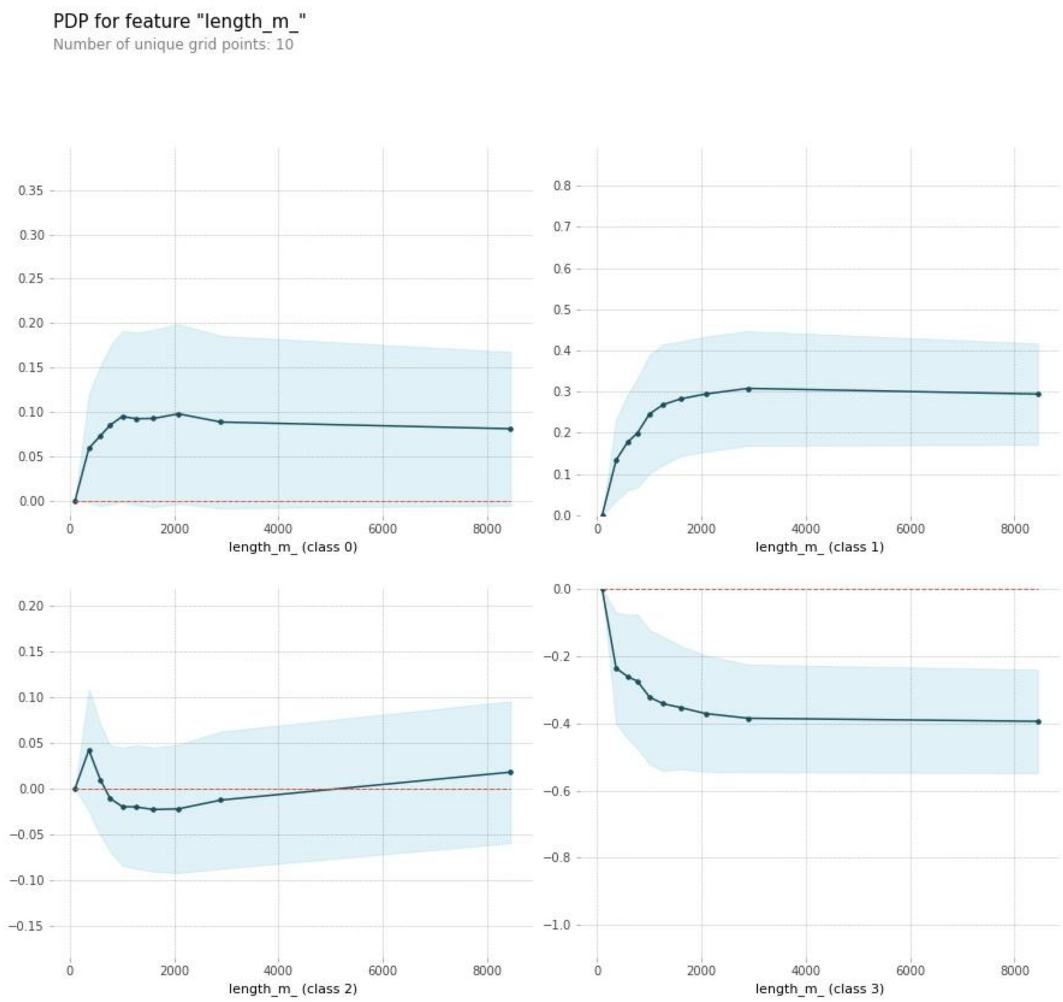


Figure 9 the partial dependent relationship between mode choice and trip length based on RF method. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

Travel mode choice could be highly related with location of destination. One of representative type of destination is public facility, which is comprised of cinema, museum, workshop, bank, fitness, and so on. According to Figure 10, public facilities surrounded endpoint of trip has positive effect on docked services (class 0 and class 1). This may result of docks arrangement that is always set nearby places with higher population flux. The public facilities are always typical buildings which are located in area with higher traffic flow. However, the possibility of choosing docked services will not keep growing with the number of surrounded public facilities of endpoints.

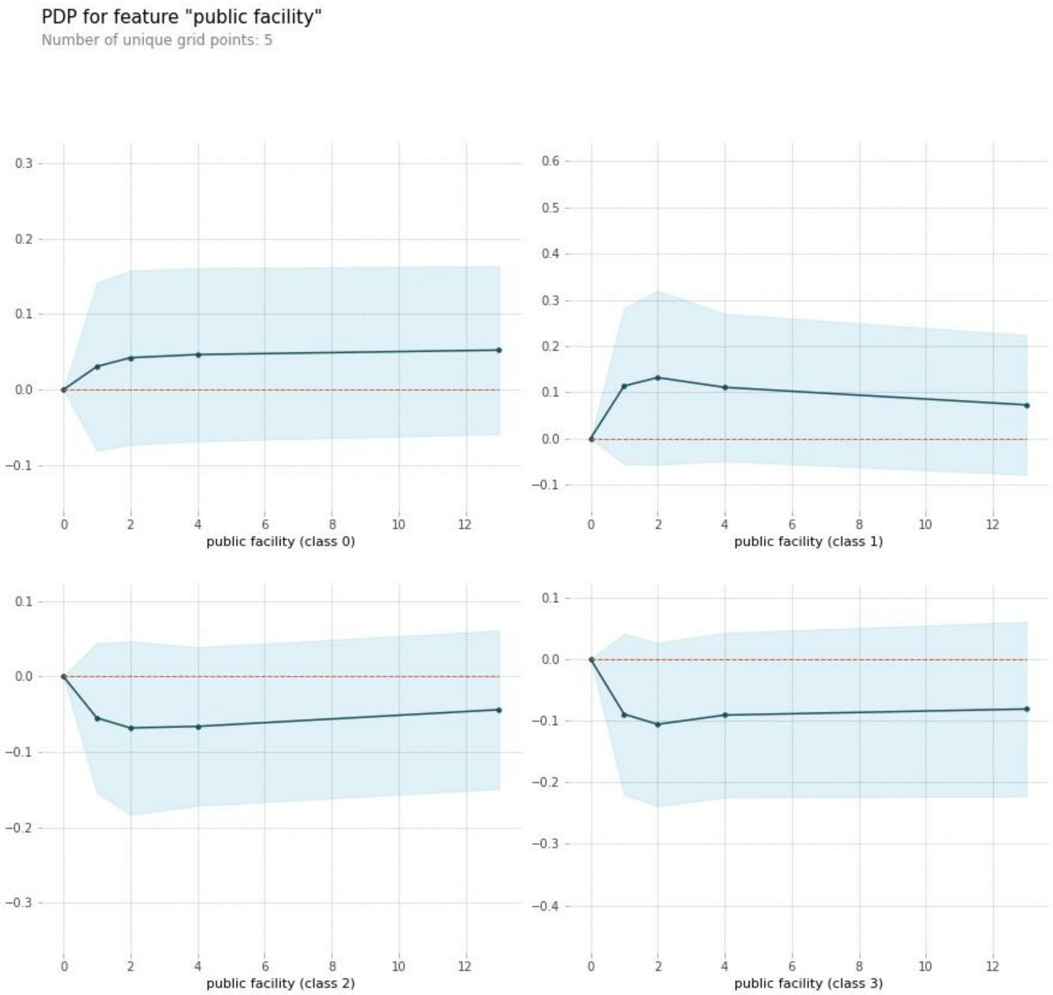


Figure 10 the partial dependent relationship between mode choice and public facility surround endpoint of trip based on RF method. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

To some extent, DEM difference between start point and endpoint represents the direction of inclination, if it is uphill or downhill. When the value of DEM difference is positive, this trip could be seen as a uphill trip, while negative value means downhill trip. Based on the partial dependence plots, users has a obviously preference at docked bike (class 0) when the value of DEM difference is between -20m and 20m, which indicates a gentle terrain. By comparison, although docked e-bike (class 1) also has higher possibility to be chosen when the terrain is flatter, there is no outstanding trend like docked bike. Moreover, escooter (class 3) is more likely to be used for a downhill trip instead of uphill trip.

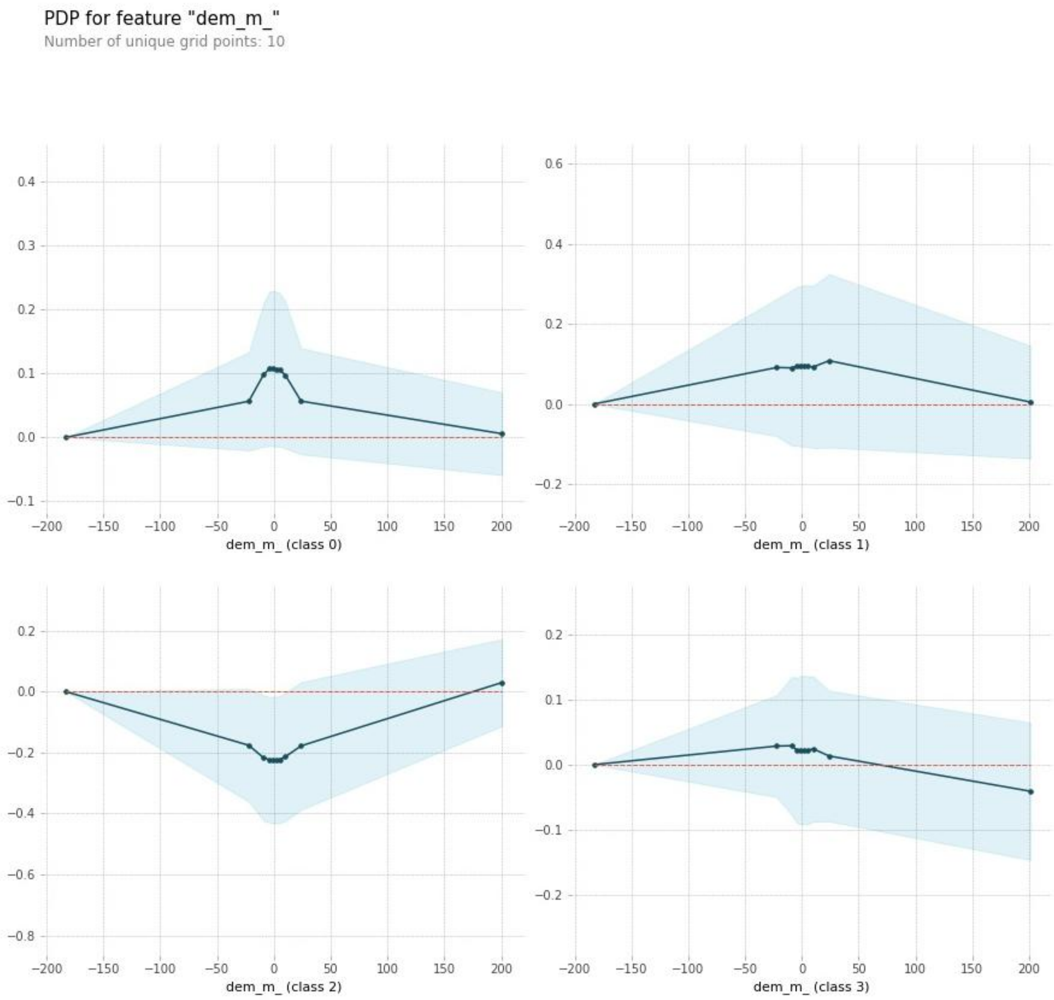


Figure 11 the partial dependent relationship between mode choice and DEM difference between started point and endpoint of trip based on RF method.The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

Another important POI type is Education services, including university, college ,school, and so on. Education service does not affect travel choice on docked e-bike (class 1) and dockless e-bike (class

2). By contrast, escooter (class 3) is more likely to be used to arrive education services instead of docked bike (class 0).

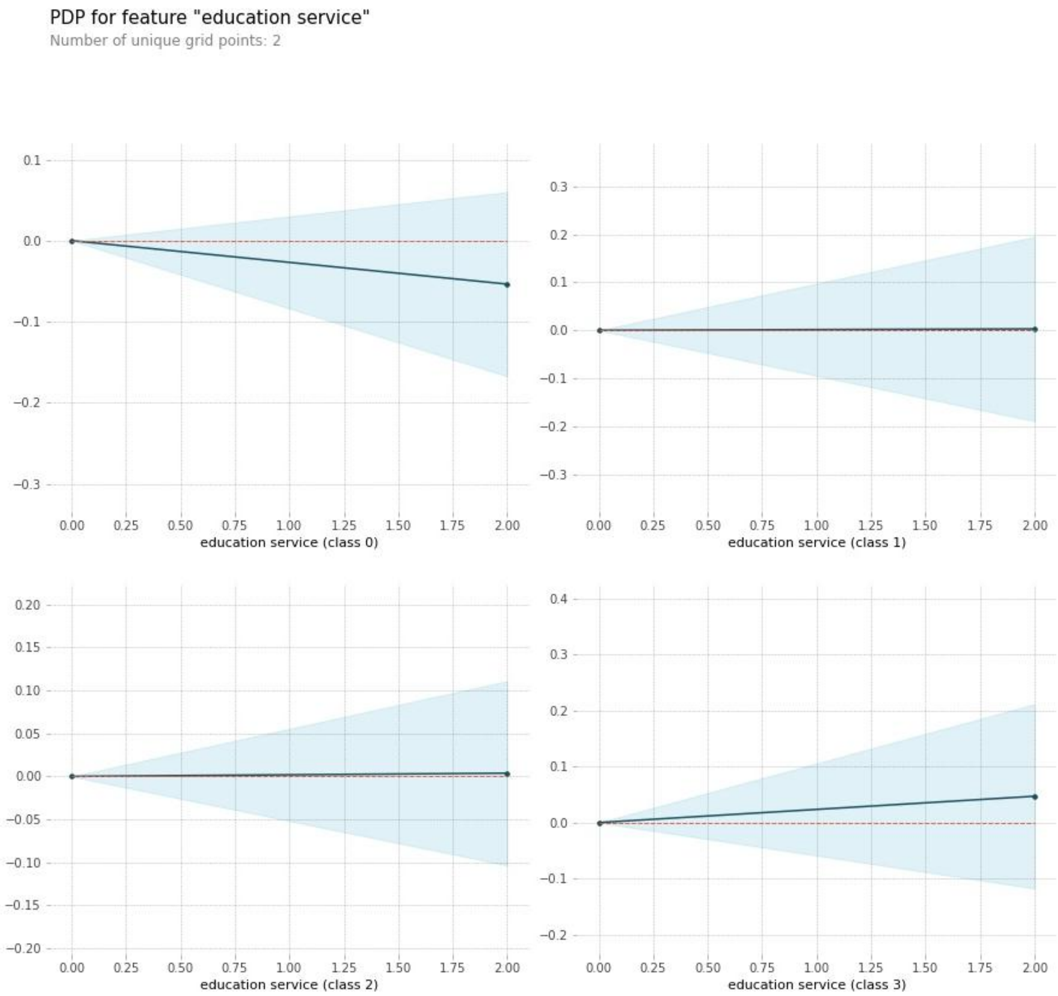


Figure 12 the partial dependent relationship between mode choice and number of education services surrounded endpoint of trip based on RF method. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

The next important factor is the average speed of the entire trip. Here, the average speed during the trip is calculated by the straight line distance between departure points and destination points. Considering the maximum speed of each mode is restrained by the hardware facility. The docked bike (class 0) could speed up to 25 kph (kilometer per hour) offered by Publibike, the same as e-scooter (class 3) (Füglister 2019), while the maximum speed of dockless e-bike (class 2) provided by

Bond company could arrive 45 kph. Here, the average speed during the trip is calculated by the straight line distance between departure points and destination points. According to Figure 13, although the e-scooter (class 3) has the same limitation of speed as dockless e-bike (class 2), when speed is faster than 10 kph, there is a gradually decreasing probability of utility. The increasing trends of docked e-bike (class 1) are fiercer than docked bike mode.

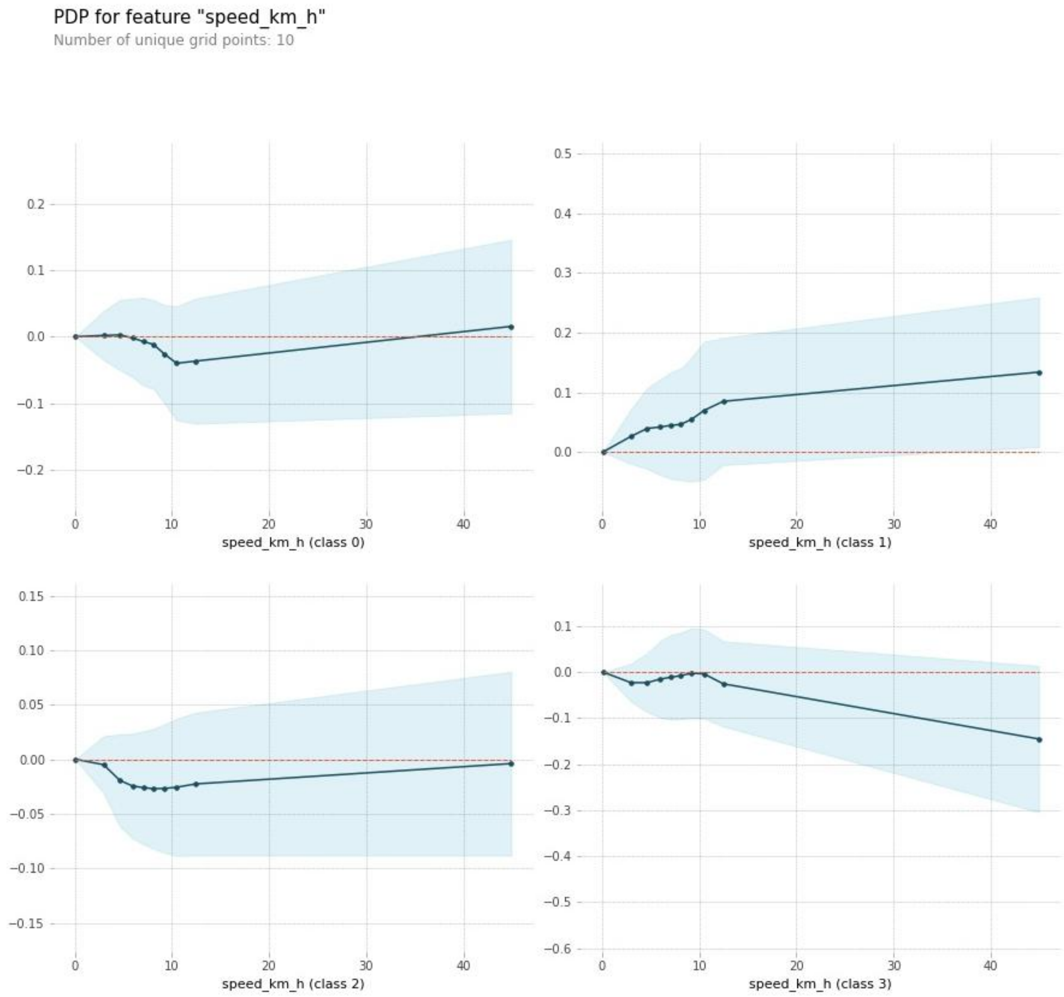


Figure 13 the dependent relationship between mode choice and average speed during the trip. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

If there are more Food & beverage POI around destination of trip, the possibility of using escooter (class 3) as travel mode is increasing, while docked facilities (class 0 and class 1) are less likely to be chosen. Moreover, dockless e-bike (class 2) shows slightly positive preference when the number of food&beverage POI is more than 4.

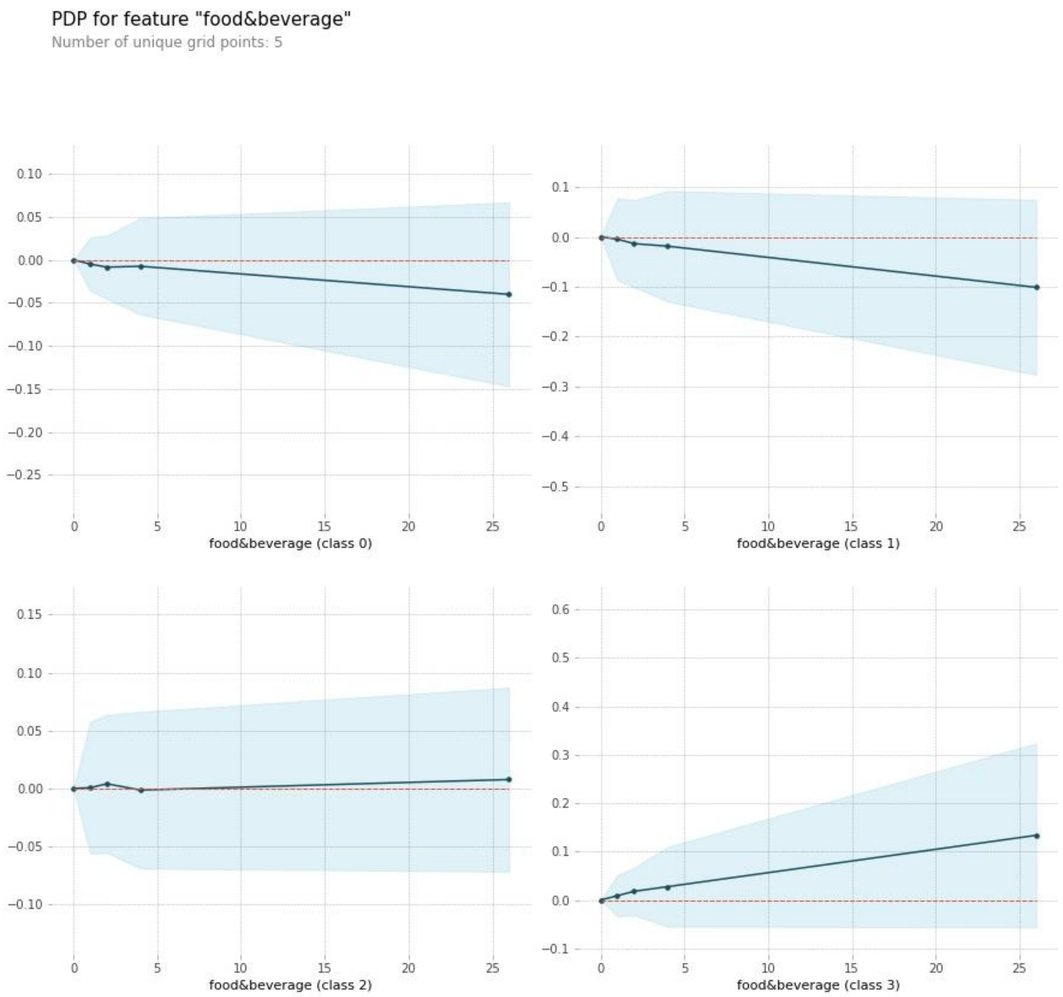


Figure 14 the dependent relationship between mode choice and the number of food&beverage POI around the destination of each trip. The blue shaded area indicates the confidence interval in PDP plots. The x-axis represents the value of an independent variable, and the y-axis is interpreted as change in the prediction from what it would be predicted at the base value or leftmost value. Class 0 to Class 3 are individually represented docked bike, docked e-bike, dockless e-bike, e-scooter

5. Discussion

5.1. Influence Factors

Overall, our analyses reveal that the mutual effects between four shared microservices, trip duration, trip length are the most important factors which may affect users' travel mode choice among docked bike, docked e-bike, dockless e-bike, and e-scooter. The preference to e-scooter is more negatively affected by longer distance and faster speed needed than the other modes. Instead, it affects preference to docked e-bike more positively. This is restricted by the hardware facility of various vehicles. To sum up, e-scooter normally undertake short trips in both spatial and temporal ways, while the others have superiority on longer distance or longer duration.

In general, the dockless e-bike has stronger competitive power compared with the docked e-bike, but less with the docked bike. It could be considered as that for the same type of vehicles (e-bike), dockless facilities always have priority when there are docked facilities at the same time. This finding confirms results from previous studies that dockless e-bike as the fifth generation of traffic is much more competitive compared with traditional docked (e-bike and bike) sharing system (Guidon et al. 2019). Reck has speculated that users are more likely to choose dockless services (e-bike and e-scooter) when there are available vehicles within a nice distance (Reck et al. 2021). Because of the fixed dock, docked services are limited significantly when people plan a trip. Therefore, improving the layout of dock sites or rise the number of dock sites in the area with higher flux could be useful approaches to increase the usage of docked micro-mobility (Bachand-Marleau et al. 2012; Rect et al. 2021; Li and Kamargianni. 2018; Wang et al 2015). How to find the most effective combination of location and number of dock infrastructures is a super valuable question for companies and governments to design and plan well.

Trip purpose has been discussed many times in previous studies (Wu et al. 2020), as a common conclusion, commuting purpose and has been recognized as the most frequent one when using shared systems (Faghih-Imani et al. 2015; Zhao et al. 2015). In this paper, this attempt is conducted by different POI points which are nearby the endpoints of trips. As a result, public facilities are the most frequent destination by using docked services (docked bike and docked e-bike). In contrast, going to education service is an important trip purpose for escooter users, while places offered food and beverage are also significant destination. However, for the other types of features generated by POI points, the purpose of the trip revealed by POI variables seems to have an scarcely effect on mode choice between the four shared micro-mobility modes. The reason that contributes to the result could be the inadequate quality of POI points. More meticulous POI data cleaning and the classified process could be a solution.

One more influence factors analyzed here is about urban environmental conditions, topography. Past study has confirmed that e-scooters are most likely on flat terrain (Rect et al. 2021). e-bikes (docked and dockless) are preferable at tortuous terrain (both uphill and downhill). These trends are approved again in this paper.

5.2. Model Analysis

Considering the models used in this study to predict travel mode choice, there are many differences between various models. By comparing the prediction accuracy of travel modes of four models, Random Forest performs the best on both training set and testing set, the accuracy of the training set is 82 percent, and 72 percent for testing set. In many previous studies, the RF model is confirmed as a great model with the highest accuracy compared with other models on travel mode choice classification researches (Cheng et al. 2019; Jahangiri and Rakha 2015; Zhao et al. 2020; Zhou et al. 2019). However, This model indeed performs unsatisfactorily on identifying docked bike and dockless e-bike modes, which need further modification in the future study. SVM has been explored as the model with ideal prediction accuracy sometimes (Cheng et al. 2019; Jahangiri and Rakha 2015; Liu and Luo 2018) when study travel mode choice topics. The truth that RF is more computationally efficient than SVM because of less execution time for the training model (Cheng et al. 2019) has been verified again in this paper. Meanwhile, the process of RF hyperparameters adjustment is more concise than SVM as well. ANN and SVM models also provided fairish prediction accuracy, which outperforms the LR model. Moreover, MLP as a basic ANN model is used in this paper, but there are many modification models based on MLP which could be the potential to offer better performance on travel mode choice simulation. It deserves to be further explored.

Although different models have distinct principles and performance, they all agree that the trip attribute of each shared micro mode is the most important influence factor. Only the SVM model gave a different result, but considering its lower prediction accuracy, it is not considered as reliable evidence. Based on the Partial Dependence Plots, non-linear associations between independent variables and dependent variables could be captured readily by machine learning models (Zhao et al. 2020). Therefore, machine-learning models could be served as an exploratory tool and applied to evaluate non-linear influences, although some ML models are kind of black-box models.

In short, the RF method usually has superior advantages in advancing travel mode choice classification. In particular, more and more advanced technologies enables different data sets of travel attributes data to be collected from various electronic devices. With more accessible travel information, the RF model could be seen as a robust model that can make good use of the big data. It has significant worth to develop the model or combine it with other types of model to capture more complicated associations between various datasets obtained from different sources. Due to the capability of the RF model in handling different types of independent variables and modeling complex non-linear relationships, it could combine with advanced data acquirement approaches as a promising method to contributes to travel mode choice analysis in the upcoming data era.

6. Conclusion

This study firstly explored the potential influence factors that affect passengers' mode choice between four different micro-mobility modes (docked bikes and docked e-bikes, dockless e-bike, dockless e-scooter) by Machine Learning methods. A widely accessible vehicle location dataset acquired from companies' open APIs is used to conduct analyses. Overall, according to these analyses, questions that are proposed in Section 1 could be answered.

Firstly, our results suggest that the dominant influence factors which affect transportation mode choice between four shared micro-mobility services are trip attributes (the distance of trip in both temporal and spatial way). Secondly, the performance of four Machine Learning methods is compared based on the generated dataset. Random Forest is verified as the model with the best performance on classification accuracy and implementation difficulties. Associations between independent variables and dependent variables based on RF are explained readily as well. By contrast, Linear Regression method gives the worst classification accuracy. Furthermore, based on the feature importance and partial dependence plots, Some suggestions are proposed for advancing the shared micro-mobility system. For instance, e-scooter is preference for trip aimed on education service compared with other micro-mobility travel modes, thus dropping more escooters around educational area could be a effective method to enhance development of escooter. Similarly, docked services (docked bike and docked e-bike) are more likely to be used when the destination of the trip is belonging to public facilities, therefore, docked facility should have a enough volume around public facilities to provide more convenience.

Generally, dockless e-bikes are more competitive compared with a docked e-bike. Will dockless e-bike take the place of the docked e-bike in the transportation system? Is it a more convenient and efficient mode? These could be valuable to study in future researches. Related governments department and shared micro mobility service providers can leverage these associations to develop more feasible shared micro-mobility regulations and optimize present facilities layout.

This study has some limitations which could be complemented in future works. First and foremost, our analysis only paid attention to the influences of external factors and trip attributes on mode choice analyses. To give more supportive evidence of mode choice analyses, personal information could be extended into features, such as sociodemographic information, micro-mobility service membership, personal preferences). The combination of the specific empirical data and survey data could be used to build a high-quality dataset. Meanwhile, other alternative modes (like public transport and walking which are recommended for relieving urban congestion and keeping sustainable development) could provide more valuable analysis. In addition, this study is constrained to one city, Zurich. Different city backgrounds related to shared micro-mobility facilities could contribute to a completely distinct situation. Thus, the results of this paper that are relative with influence factors apply to Zurich only, but the RF model could be the first recommended model for the travel mode topic. In the end, shared micro-mobility contributes to the transportation system in many aspects. This paper only discussed how different situations affect users' preference on mode choice.

7. Reference

Bachand-Marleau, J., B. H. Y. Lee, and A. M. El-Geneidy. 2012. Better Understanding of Factors Influencing Likelihood of Using Shared Bicycle Systems and Frequency of Use. *Transportation Research Record: Journal of the Transportation Research Board* 2314: 66–71. doi:10.3141/2314-09.

Bhat, C. R. 1998. Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. *Transportation Research Part A: Policy and Practice* 32: 495–507. doi:10.1016/S0965-8564(98)00011-1.

Bucher, D., R. Buffat, A. Froemelt, and M. Raubal. 2019. Energy and greenhouse gas emission reduction potentials resulting from different commuter electric bicycle adoption scenarios in Switzerland. *Renewable and Sustainable Energy Reviews* 114: 109298. doi:10.1016/j.rser.2019.109298.

Bucher, D., H. Martin, J. Hamper, A. Jaleh, H. Becker, P. Zhao, and M. Raubal. 2020. Exploring Factors that Influence Individuals' Choice Between Internal Combustion Engine Cars and Electric Vehicles. *AGILE: GIScience Series* 1: 1–23. doi:10.5194/agile-giss-1-2-2020.

Cheng, L., X. Chen, J. De Vos, X. Lai, and F. Witlox. 2019. Applying a random forest method approach to model travel mode choice behavior. *Travel Behaviour and Society* 14: 1–10. doi:10.1016/j.tbs.2018.09.002.

Cui, B., G. Boisjoly, L. Miranda-Moreno, and A. El-Geneidy. 2020. Accessibility matters: Exploring the determinants of public transport mode share across income groups in Canadian cities. *Transportation Research Part D: Transport and Environment* 80: 102276. doi:10.1016/j.trd.2020.102276.

Degele, J., A. Gorr, K. Haas, D. Kormann, S. Krauss, P. Lipinski, M. Tenbih, C. Koppenhoefer, et al. 2018. Identifying E-Scooter Sharing Customer Segments Using Clustering. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1–8. Stuttgart: IEEE. doi:10.1109/ICE.2018.8436288.

Ding, Z., X. Jiang, Z. Liu, R. Long, Z. Xu, and Q. Cao. 2018. Factors affecting low-carbon consumption behavior of urban residents: A comprehensive review. *Resources, Conservation and Recycling* 132: 3–15. doi:10.1016/j.resconrec.2018.01.013.

Dublin, J. 2020 .Global Bike Sharing Market Research Report 2019-2025.. Retrieved 2 January, from https://www.globenewswire.com/

Ermagun, A., Rashidi, T.H., Lari, Z.A., 2015. Mode choice for school trips: long-term planning and impact of modal specification on policy assessments. *Transport.* Res. Rec.2513, 97–105.

Guidon, S., H. Becker, H. Dediu, and K. W. Axhausen. 2019. Electric Bicycle-Sharing: A New Competitor in the Urban Transportation Market? An Empirical Analysis of Transaction Data. *Transportation Research Record: Journal of the Transportation Research Board* 2673: 15–26. doi:10.1177/0361198119836762.

Haas, M. C., R. G. Hoogendoorn, C. E. Scheepers, and S. Hoogendoorn-Lanser. 2018. Travel mode choice modeling from cross-sectional survey and panel data: the inclusion of initial nonresponse. *Transportation Research Procedia* 32. Transport Survey Methods in the Era of Big Data:Facing the Challenges: 268–278. doi:10.1016/j.trpro.2018.10.049.

Hillel, T., M. Bierlaire, M. Z. E. B. Elshafie, and Y. Jin. 2021. A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling* 38: 100221. doi:10.1016/j.jocm.2020.100221.

Holmgren, J., and P. Ivehammar. 2020. Mode choice in home-to-work travel in mid-size towns: The competitiveness of public transport when bicycling and walking are viable options. *Transportation Research Procedia* 48. Recent Advances and Emerging Issues in Transport Research – An Editorial Note for the Selected Proceedings of WCTR 2019 Mumbai: 1635–1643. doi:10.1016/j.trpro.2020.08.204.

Jahangiri, A., Rakha, H.A., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans. Intell. Transp. Syst. 16 (5), 2406–2417.

Klöckner, C. A., and E. Matthies. 2004. How habits interfere with norm-directed behaviour: A normative decision-making model for travel mode choice. *Journal of Environmental Psychology* 24: 319–327. doi:10.1016/j.jenvp.2004.08.004.

Li, W., and M. Kamargianni. 2018. Providing quantified evidence to policy makers for promoting bike-sharing in heavily air-polluted cities: A mode choice model and policy simulation for Taiyuan-China. *Transportation Research Part A: Policy and Practice* 111: 277–291. doi:10.1016/j.tra.2018.01.019.

Liaw, A., and M. Wiener. 2001. Classification and Regression by RandomForest. *Forest* 23.

Liu, Chunyu, Luo Xia. 2018. A Travel Mode Choice: a Comparision of Machine Learning and Multinomial Logit Model. *China Transportation Review* 1000-713X (2018)08-0057-07

Ma, X., Y. Ji, Y. Yuan, N. Van Oort, Y. Jin, and S. Hoogendoorn. 2020. A comparison in travel patterns and determinants of user demand between docked and dockless bike-sharing systems using multi-sourced data. *Transportation Research Part A: Policy and Practice* 139: 148–173. doi:10.1016/j.tra.2020.06.022.

Marra, A. D., and F. Corman. 2020. Determining an efficient and precise choice set for public transport based on tracking data. *Transportation Research Part A: Policy and Practice* 142: 168–186. doi:10.1016/j.tra.2020.10.013.

Mayo, F. L., and E. B. Taboada. 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* 4: 100078. doi:10.1016/j.trip.2019.100078.

McKenzie, G. 2020. Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services. *Computers, Environment and Urban Systems* 79: 101418. doi:10.1016/j.compenvurbsys.2019.101418.

Narayan, J., O. Cats, N. van Oort, and S. Hoogendoorn. 2020. Integrated route choice and assignment model for fixed and flexible public transport systems. *Transportation Research Part C: Emerging Technologies* 115: 102631. doi:10.1016/j.trc.2020.102631.

Permutation Importance. 2021. https://kaggle.com/dansbecker/permutation-importance. Accessed July 21.

Pineda-Jaramillo, J. D. 2019. A review of Machine Learning (ML) algorithms used for modeling travel mode choice. May.

Pouyanne, G. 2004. *The influence of urban form on travel patterns. An application to the metropolitan area of Bordeaux*. ersa04p244. *ERSA conference papers*. ERSA Conference Papers. European Regional Science Association.

Reck, D. J., H. Haitao, S. Guidon, and K. W. Axhausen. 2021. Explaining shared micromobility usage, competition and mode choice by modelling empirical data from Zurich, Switzerland. *Transportation Research Part C: Emerging Technologies* 124: 102947. doi:10.1016/j.trc.2020.102947.

Scheiner, J., and C. Holz-Rau. 2007. Travel mode choice: affected by objective or subjective determinants? *Transportation* 34: 487–511. doi:10.1007/s11116-007-9112-1.

Shafique, M.A., Hato, E. Use of acceleration data for transportation mode prediction.*Transportation* 42, 163–188 (2015). https://doi.org/10.1007/s11116-014-9541-6

Sivasubramaniyam, R. D., S. G. Charlton, and R. J. Sargisson. 2020. Mode choice and mode commitment in commuters. *Travel Behaviour and Society* 19: 20–32. doi:10.1016/j.tbs.2019.10.007.

Stojanovski, T. 2019. Urban Form and Mobility Choices: Informing about Sustainable Travel Alternatives, Carbon Emissions and Energy Use from Transportation in Swedish Neighbourhoods. *Sustainability* 11. Multidisciplinary Digital Publishing Institute: 548. doi:10.3390/su11020548.

Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

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

Tran, T. D., N. Ovtracht, and B. F. d'Arcier. 2015. Modeling Bike Sharing System using Built Environment Factors. *Procedia CIRP* 30: 293–298. doi:10.1016/j.procir.2015.02.156.

Tran, Y., T. Yamamoto, and H. Sato. 2020. The influences of environmentalism and attitude towards physical activity on mode choice: The new evidences. *Transportation Research Part A: Policy and Practice* 134: 211–226. doi:10.1016/j.tra.2020.02.012.

Tyrinopoulos, Y., and C. Antoniou. 2013. Factors affecting modal choice in urban mobility. *European Transport Research Review* 5. SpringerOpen: 27–39. doi:10.1007/s12544-012-0088-3.

Van Acker, V., Witlox, F. Commuting trips within tours: how is commuting related to land use?.*Transportation* 38, 465–486 (2011). https://doi.org/10.1007/s11116-010-9309-6

Wang, C.-H., G. Akar, and J.-M. Guldmann. 2015. Do your neighbors affect your bicycling choice? A spatial probit model for bicycling to The Ohio State University. *Journal of Transport Geography* 42: 122–130. doi:10.1016/j.jtrangeo.2014.12.003.

Wangsness, P. B., S. Proost, and K. L. Rødseth. 2020. Vehicle choices and urban transport externalities. Are Norwegian policy makers getting it right? *Transportation Research Part D: Transport and Environment* 86: 102384. doi:10.1016/j.trd.2020.102384.

Weis, C., M. Kowald, A. Danalet, B. Schmid, M. Vrtic, K. W. Axhausen, and N. Mathys. 2021. Surveying and analysing mode and route choices in Switzerland 2010–2015. *Travel Behaviour and Society* 22: 10–21. doi:10.1016/j.tbs.2020.08.001.

Wu, L., W. Wang, P. Jing, Y. Chen, F. Zhan, Y. Shi, and T. Li. 2020. Travel mode choice and their impacts on environment—a literature review based on bibliometric and content analysis, 2000–2018. *Journal of Cleaner Production* 249: 119391. doi:10.1016/j.jclepro.2019.119391.

Yang, W., and S. Zhou. 2020. Using decision tree analysis to identify the determinants of residents' CO2 emissions from different types of trips: A case study of Guangzhou, China. *Journal of Cleaner Production* 277: 124071. doi:10.1016/j.jclepro.2020.124071.

Yanocha, D., and A. Mackenzie. 2021. The Electric Assist: Leveraging E-bikes and E-scooters for More Livable Cities. *Institute for Transportation and Development Policy*. https://www.itdp.org/publication/electric-assist/. Accessed April 18.

Zhao, J., J. Wang, and W. Deng. 2015. Exploring bikesharing travel time and trip chain by gender and day of the week. *Transportation Research Part C: Emerging Technologies* 58: 251–264. doi:10.1016/j.trc.2015.01.030.

Zhao, X., X. Yan, A. Yu, and P. Van Hentenryck. 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society* 20: 22–35. doi:10.1016/j.tbs.2020.02.003.

Zhou, X., M. Wang, and D. Li. 2019. Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning. *Journal of Transport Geography* 79: 102479. doi:10.1016/j.jtrangeo.2019.102479.

Zhu, Z., L. Sun, X. Chen, and H. Yang. 2021. Integrating probabilistic tensor factorization with Bayesian supervised learning for dynamic ridesharing pattern analysis. *Transportation Research Part C: Emerging Technologies* 124: 102916. doi:10.1016/j.trc.2020.102916.

Zurich | Federal Statistical Office. 2021. https://www.bfs.admin.ch/bfs/en/home/statistics/regional-statistics/regional-portraits-key-figures/cantons/zurich.html. Accessed April 23.