AGILE: GIScience Series, 3, 20, 2022. https://doi.org/10.5194/agile-giss-3-20-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

A machine learning based approach for predicting usage efficiency of shared e-scooters using vehicle availability data

Pengxiang Zhao [1], Aoyong Li [2], Petter Pilesjö [1], and Ali Mansourian [1]

[1]GIS Center, Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden
[2]State Key Laboratory of Automotive Safety Energy, School of Vehicle and Mobility, Tsinghua University, Beijing, China

Correspondence: Pengxiang Zhao (pengxiang.zhao@nateko.lu.se)

Abstract.

Shared electric scooters (e-scooters) have been rapidly growing in popularity across Europe over the past three years, which can bring various environmental and socioeconomic benefits. However, how to further improve the usage efficiency of shared e-scooters is still a major concern for micro-mobility operators and city planners. This paper proposes a machine learning based approach to predict the usage efficiency of shared e-scooters using GPS-based vehicle availability data. First, the usage efficiency of shared e-scooters is measured with the indicator Time to Booking at the trip level. Second, ten exploratory variables in time and space are calculated as features for the prediction based on the e-scooter trips and other related data. Last, three typical machine learning methods, including logistical regression, artificial neural network and random forest are applied to predict the usage efficiency by inputting the features. Besides, the variable importance is evaluated by taking the random forest model as an example. The results show that the random forest model yields the best prediction performance (accuracy = 71.2%, F1 = 78.0%), and the variables like the hour of day and POI density present high variable importance. The findings of this study will be beneficial for micro-mobility operators and city planners to design policies and strategies for further improving the usage efficiency of e-scooter sharing services.

Keywords. Micro-mobility, E-scooter sharing, Usage efficiency, Spatiotemporal analysis, Machine learning, Vehicle availability data

1 Introduction

In the past few years, there has been a proliferation of shared micro-mobility systems, especially e-scooter sharing services all over the world (e.g. McKenzie, 2020; Heumann et al., 2021). National Association of City Transportation Officials reported that people took 136 million trips on shared micro-mobility services in the United States (US) in 2019, with a 60% increase from 2018 (https://nacto.org/shared-micromobility-2019/). Among them, 86 million trips were made with shared electric scooters (e-scooters) in 109 US cities. In Europe, more and more micro-mobility operators are purchasing fleets of e-scooters to be deployed in many cities. For example, the two launched European e-scooter startups, namely TIER in Berlin and Voi in Stockholm, have operated e-scooter fleets in 130 and more than 70 European cities (https://sifted.eu/articles/escooter-market-updates/). Especially, the convenient characteristics of e-scooter sharing services, such as public availability, free floating, mobile payment, accelerate the rapid dissemination of e-scooters. These established e-scooter sharing systems allow users to rent e-scooters via a smartphone app, which have been indicated as an sustainable transport mode to support urban mobility, mainly for short- and medium-distance travel (e.g. Jiao and Bai, 2020; Dias et al., 2021; Hosseinzadeh et al., 2021).

Although the introduction of e-scooter sharing services is capable of mitigating the transportation problems of cities and bring social and environmental benefits, the unbalanced usage of e-scooters in urban space has always been one of the obstacles of fleet management and e-scooter sharing services. In particular, one of the most important goals of micro-mobility operators is to occupy the market, which normally leads to an oversupply of vehicles in the market, thereby causing a waste of resources and inefficient micro-mobility services. Considering the nature of dockless e-scooter sharing service, its great flexibility is also accompanied by the challenge of unpredictable usage patterns, such as an imbalanced distribution of e-scooters across a city. The inefficient deployment of vehicle fleet not only occupies the public urban space, but also increases the operation and maintenance costs of service providers. Besides, due to the limited battery capacity,

some vehicles can rapidly become out-of-charge during the course of the day if they are overused in quick succession (Losapio et al., 2021). It may also lower the usage efficiency of e-scooters if they are not charged on time. Furthermore, as a new form of urban mobility, their sudden appearance and rapid expansion has challenged city administrators and micro-mobility operators in efficient fleet management. Hence, it is urgent and necessary to investigate how to improve usage efficiency of micro-mobility services.

In recent years, vehicle availability data that records the GPS locations of all available vehicles (i.e., e-scooters in this study) from service providers has attracted notable attention in micro-mobility studies (e.g. Li et al., 2021; Zhu et al., 2020; Ziedan et al., 2021), which opens an avenue for fleet management by exploring shared micro-mobility usage patterns (e.g. Bai and Jiao, 2020; Caspi et al., 2020; Almannaa et al., 2021; Guo and Zhang, 2021). However, the existing studies are focused on analyzing shared micro-mobility usage patterns and the related influencing factors, little attention has been paid to predicting usage efficiency of e-scooter sharing services using GPS-based vehicle availability data.

To bridge the gap, this study aims to predict the usage efficiency of shared e-scooters to support e-scooter sharing services with vehicle availability data by developing a machine learning (ML) based approach. First, the usage efficiency of e-scooter sharing services is measured at the trip level based on the vehicle availability data from service providers. Second, the exploratory variables that include the spatial and temporal characteristics of the trip and its contextual information are used as independent variables, whereas the usage efficiency is regarded as dependent variable in terms of the indicator Time to Booking (TtB). Third, the potential of three typical machine learning methods, namely logistical regression, artificial neural network and random forest, is explored to predict the usage efficiency of shared e-scooters. The experiment is conducted based on the vehicle availability data collected in Stockholm, Sweden, which will be introduced in section 3.1.

The remainder of this paper is structured as follows. Related work regarding usage efficiency of micro-mobility services and influencing factors of e-scooter usage are reviewed in section 2. Section 3 describes the data and software availability, and the used methods for predicting usage efficiency of shared e-scooters in this study. The experimental results are shown in section 4. We discuss and conclude this research in section 5.

2 Related work

2.1 Usage efficiency of micro-mobility services

The previous studies on analyzing usage efficiency of micro-mobility services are mainly concentrated on bike-sharing systems. For instance, Guo et al. (2017) used turnover rate to describe the bike-sharing usage, and identified the factors that have an influence on bike-sharing usage. It is found that the bike-sharing usage is affected by household bike ownership, travel time, and bike-sharing stations location, etc. Du et al. (2019) employed usage frequency to depict the usage of public bike-sharing systems, and developed a framework to explore the spatio-temporal usage patterns of free-floating shared bikes. It is reported that the factors like residential area, park and green area, and population size have a significant influence on the usage frequency of bike-sharing system. Gu et al. (2019) proposed a heuristic bike optimization algorithm to determine the optimal supply and distribution of bikes, thereby improving the usage efficiency of the free-floating bike sharing system. Wang et al. (2019) explored the bike usage in terms of rental duration and proposed an usage balancing design for bike-sharing systems towards efficient sharing. Li et al. (2020) measured the usage efficiency of bike-sharing service by calculating Time to Booking for each bike with GPS-based bike origin-destination data, and explored how it is influenced by the built environment and social-demographic characteristics with ordinary least squares (OLS) regression and geographically weighted regression (GWR) models. In recent studies, Li et al. (2022) further employed Time to Booking to measure the usage efficiency of e-scooter sharing services by conducting a comparison study in 30 European cities.

In summary, research into the usage efficiency of e-scooter sharing services is still limited. Nonetheless, several indicators, such as usage frequency, turnover rate, rental duration, time to booking, have been used to measure the usage efficiency of bike-sharing systems, which provide insights on evaluating the usage efficiency of e-scooter sharing services.

2.2 Influencing factors of e-scooter usage

Several studies have been conducted to explore the influencing factors of e-scooter sharing and examine the relationships between e-scooter usage and them. For instance, Caspi et al. (2020) explored the usage patterns of e-scooter sharing services using the trip data in Austin, Texas over about a six-month period. It is found that the usage of e-scooters is related with the areas with high employment rates and with bike infrastructure. Jiao and Bai (2020) examined the relationships between the e-scooter sharing usage and the surrounding environments using the shared e-scooter trips from April 2018 to February 2019 in Austin, TX. The results show that the factors, such as a shorter distance to the city center, the presence of transit stations, better street connectivity, are associated with the increased e-scooter usage. Huo et al. (2021) investigated the effects of the built environment on e-scooter sharing ridership using the multilevel negative binomial model based on the trip data in five cities of US. The results indicate that the e-scooter sharing usage is positively correlated with the

following factors, including population density, employment density, intersection density, land use mixed entropy, and bus stop density, etc. The study by Hosseinzadeh et al. (2021) identified the relationship between the e-scooter trip density and characteristics of sustainable urban development. By applying a Generalized Additive Modeling (GAM) approach to the e-scooter trip data in in Louisville, Kentucky, it is found that commercial land use percent, industrial land use percent, Walk Score and Bike Score have an influence on e-scooter trip density.

Overall, the existing studies have indicated that the urban built environment characteristics have a remarkable influence on e-scooter sharing usage, which can aid in evaluating the usage efficiency of e-scooter sharing services for fleet management. However, the research on predicting the usage efficiency of shared e-scooters is still scarce. This study aims to predict the use efficiency of shared e-scooters based on machine learning methods, which would support micro-mobility operator to further optimize e-scooter sharing services.

3 Method

3.1 Data and software availability

In this study, the vehicle availability data of e-scooter sharing services in Stockholm, Sweden was collected from one micro-mobility operator that has a high market share in Europe. The data spans from June 1^{st} to June 30^{th} 2021 to measure the usage efficiency of e-scooter sharing services and further predict it with machine learning. Each record in the data includes e-scooter id, timestamp, longitude, latitude, state of charge (SOC) for battery. The SOC is denoted by a number between 0 and 100 (%), where 100% represents the battery is fully charged.

The data processing is mainly concentrated on trip identification from vehicle availability data and trip outliers removal. First, the trips are identified from the collected vehicle availability data using the method in the study by Zhao et al. (2021), which has been demonstrated to be effective in trip identification. Next, the following criteria are adopted to remove the outliers of the e-scooter trips based on prior knowledge and existing studies (McKenzie, 2020), including: (1) trip duration is longer than one minute and less than 2 h, (2) trip distance is greater than 100 m and shorter than 15 km. Eventually, 708,974 valid trips are obtained. As shown in Figure 1, the extracted trip data is visualized in terms of trip ends, which are mainly concentrated on the inner city of Stockholm.

Besides, the points of interest (POI) and road network data in Stockholm were collected from OpenStreetMap, which are used to calculate the exploratory variables for the prediction. According to the POI categories in the study by Zhao et al. (2017), the POIs in relation to human trips in e-scooters are extracted. The selected POIs are divided into seven categories, including work, shopping, dining, recreation, schooling, lodging, and medical facilities, as shown in Table 1. Three categories of roads that are suitable for e-scooter usage are chosen, including motorways, pedestrian ways and residential ways. The difference between the latter two types is that residential ways serve as an access to housing. Besides, bus stops data and administrative division data were also collected from TRAFIKLAB (https://www.trafiklab.se/) and Dataportalen Stockholm (https://dataportalen.stockholm.se/) respectively.

Table 1. POI categories and instances

Category	POI instances
Work	Office, Government, Company
Shopping	Shopping malls, Supermarket, Kiosk, Convenience store, etc.
Dining	Restaurant, Fast food
Recreation	Museum, Art gallery, Library, Theater, Bar, Attraction, etc.
Schooling	University, School
Lodging	Hotel, Guesthouse
Medical	Pharmacy, Doctor, Chemist, Dentist

The whole study is conducted on a computer with Intel(R) Core(TM) i7-4930K CPU 3.40GHz and 32.0 GB RAM, and the program is coded with Python language. The processed data and the code are open in GitHub (https://github.com/micromobility-research/usage_efficiency_prediction).

3.2 Measuring usage efficiency of e-scooters

In this study, the indicator Time to Booking is employed to measure the usage efficiency of e-scooter sharing services at the trip level, which quantifies each idle span of shared e-scooters. Time to booking is defined as the duration that a vehicle is booked again after the previous trip has ended, which has been used to evaluate the usage efficiency of bike-sharing services (Guidon et al., 2019; Li et al., 2020). To calculate the TtB values of e-scooters, all pairs of consecutive trips for each e-scooter are extracted first. Then, given a pair of consecutive trips for one e-scooter, the TtB can be calculated with Eq. (1).

$$TtB = ST_{Previous} - ET_{Current} \tag{1}$$

where $ST_{Previous}$ represents the start time of the current trip, and $ET_{Current}$ represents the end time of the previous trip.

From the definition of TtB, the inverse proportional relationship between usage efficiency and TtB is derived. Concretely, the higher the TtB is, the lower the usage efficiency is. It should be noted that it is impossible to calculate the TtB for the last trip of each e-scooter during the selected period since the related information that when the e-scooter is booked again is not available. In addition, considering that the vehicles with potential safety hazards are

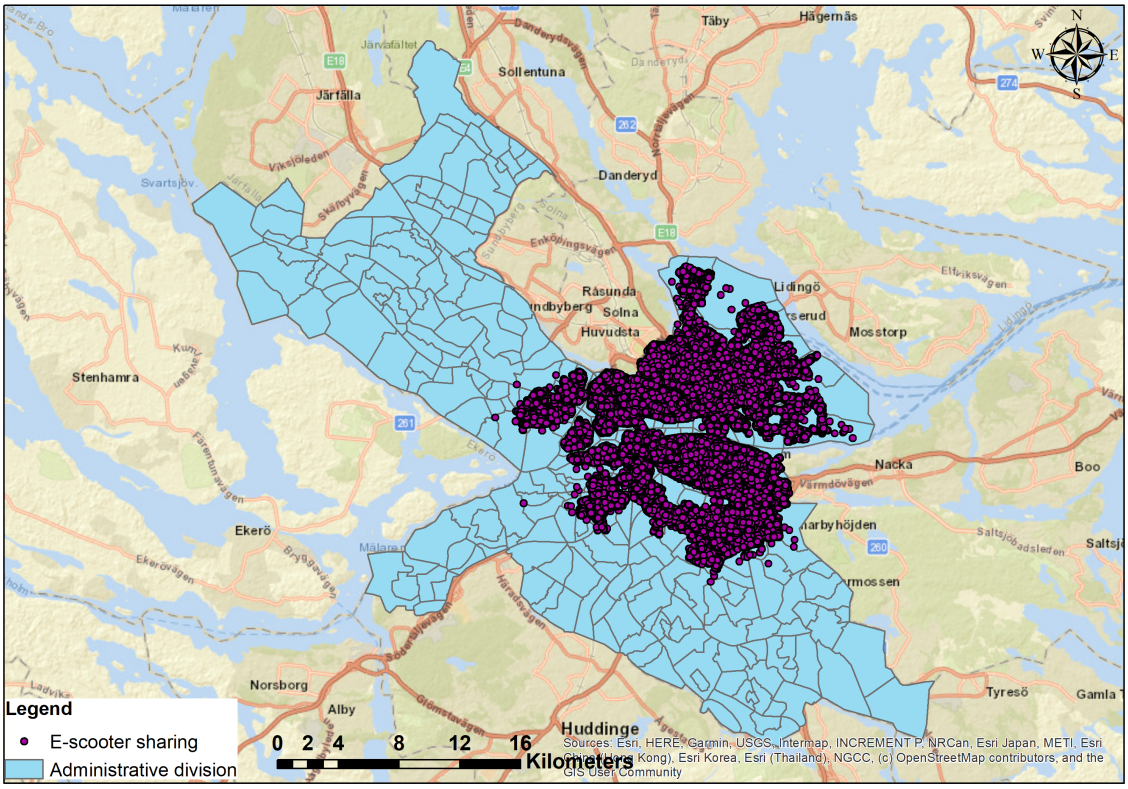

Figure 1. Study area and the extracted trip ends.

required to be maintained by micro-mobility operators before being reused, the TtB values that exceed three days are removed in this study. It is assumed that too long idle parking of e-scooters normally result from their unavailability in the market due to maintenance or some other reasons.

3.3 Exploratory variables

In this study, ten exploratory variables in time and space are selected as features to predict the usage efficiency of shared e-scooters. The temporal variables describe the occurrence time and date of idle parking, which can extracted from the identified e-scooter trips. The spatial variables depict the spatial relationships between parking location of e-scooter and its surrounding built environment, which are calculated based on the POI data and road network data in the study area. The variables are described as follows:

1. Start time of idle parking (Hour_day);
2. Idle parking occurred on weekday or weekend (Day_week);
3. Battery power of trip end (End_battery);
4. Distance to the nearest motorway (Distance_motorway);
5. Distance to the nearest pedestrian way (Distance_pedestrian);
6. Distance to the nearest residential way (Distance_residential);
7. Distance to the nearest bus stop (Distance_stop);
8. Distance to the nearest POI (Distance_POI);
9. Density of POIs in the corresponding administrative division unit (Density_POI);
10. Density of bus stops in the corresponding administrative division unit (Density_stop);

Table 2 displays the description of the features. The first three categorical variables are calculated based on the trips. The five distance-based variables are calculated based on the collected road network, bus stops and POIs dataset. The two density-based variables are obtained the density of POIs and bus stops in the corresponding administrative division unit by overlaying the trips ends with the administrative division map.

3.4 Machine learning method

In this study, we select three typical machine learning methods to implement the prediction of e-scooter sharing usage efficiency, including logistic regression (LR), artificial neural network (ANN) and random forest (RF). These ML methods have been widely used to GIS and urban transportation studies (e.g. Aditian et al., 2018; Pun et al., 2019; Bucher et al., 2020).

Table 2. Type and value of each feature.

Feature	Type	Value
Hour_day	Categorical	[0, 1, ..., 23]
Day_week	Categorical	[1, 2, ..., 7]
End_battery	Categorical	[1%, 2%, ..., 100%]
Distance_motorway	Continuous	[0, ∞]
Distance_pedestrian	Continuous	[0, ∞]
Distance_residential	Continuous	[0, ∞]
Distance_stop	Continuous	[0, ∞]
Distance_POI	Continuous	[0, ∞]
Density_POI	Continuous	[0, ∞]
Density_stop	Continuous	[0, ∞]

3.4.1 Logistic regression

Logistic regression transforms the input variables into the probability of an output variable using the logistic sigmoid function, which is a simple and efficient method for classification problems, especially binary classification (Menard, 2002). The output value is interpreted as the probability of an instance belonging to a particular class. In binary classification, logistic regression applies a logistic function to model a binary output variable. Unlike linear regression, the output range of logistic regression is bounded between 0 and 1 due to the introduction of sigmoid function. Additionally, as opposed to linear regression, a linear relationship between inputs and output variables is not required for logistic regression. The Python library Scikit-learn provides an API to implement the logistic regression.

3.4.2 Artificial neural network

Inspired by the information processing of biological neural networks in human brain, artificial neural network is comprised of a densely interconnected set of artificial neurons, where each neuron takes a number of real-valued inputs and produces real-valued output (Mitchell, 1997). The neuron (also called node) is the basic unit of computation in a neural network, which receives input from other neurons, or from an external data source and produces an output. Multiple neurons are capable of forming different networks, among which the feedforward neural network is the simplest type of artificial neural network. A feedforward neural network normally contains three types of layers, namely an input layer, a hidden layer and an output layer. Neurons from adjacent layers have weighted connections between them.

A typical example of a feedforward network with one or multiple hidden layers is called multilayer perceptron (MLP). Given a set of features X and an array of labels, the MLP is capable of learning the relationship between the features and the labels for classification. The important hyperparameters of the MLP are the number of hidden layers and the number of neurons in each hidden layer, which can be represented as a tuple ($hidden\_layer$, sizes), the penalty parameter for regularization $alpha$ as well as the learning rate.

3.4.3 Random forest

Random forest (RF) is a tree-based ensemble classifier, which constructs multiple decision trees to make predictions and performs voting for the predicted results (Breiman, 2001). The trees are generated by drawing a set of training samples with a bagging approach. Each tree depends on an independent random sample. For each decision tree in the forest, each node is split using a predetermined number of features randomly selected, which is different from traditional decision trees where the 'best' feature is used. Ultimately, the forest is created by growing the decision trees up to a user-defined number. In the case of classification, each tree votes and the membership class with the most votes is chosen as the final prediction. Hence, it is actually an extension over bagging.

As mentioned above, two hyperparameters normally require to be optimized in order to increase the predictive power of the random forest classifier, namely the number of decision trees (n_trees) and the number of features to be selected for the best split ($max\_features$). In addition, two other hyperparameters, the minimum number of samples at a leaf node ($min\_samples\_leaf$) and the minimum number of samples used to split an internal node ($min\_samples\_split$), will also be tuned.

3.5 Performance evaluation metrics

In this paper, four metrics are employed to evaluate the performance of the classifiers, including accuracy, F1 score, precision and recall. The four measures are calculated based on the following terms, namely True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN), which are commonly used in binary classification problem (Robert, 2014).

1. True Positive (TP) is data sample classified as positive by the model that actually is positive.
2. False Positive (FP) is data sample the model recognizes as positive that actually is negative.
3. True Negative (TN) is data sample classified as negative by the model that actually is negative.
4. False Negative (FN) is data sample the model recognizes as negative that actually is positive.

Accuracy is the ratio of number of correct predictions conducted by the model over number of all kinds of predictions in classification problems, which is defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Precision refers to the number of TPs divided by the sum of the number of TPs and the number of FPs, which be denoted by:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall (also called true positive rate or sensitivity) is the number of TPs divided by the sum of the number of TPs and the number of FNs, which can be expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1 score is a way of combining the precision and recall of ML model, which is defined as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

3.6 Feature importance

Explainable artificial intelligence has been attracting more attention in recent years, which can provide more transparency to the ML algorithms and help people understand the cause of decision behind a model (Miller, 2019). Feature importance technique is capable of quantifying how useful the input features are at predicting a target variable by assigning a score to them, which is very important in feature engineering and interpretability of machine learning models. In the previous studies, different approaches have been proposed to evaluate the feature importance for different models (e.g. Olden et al., 2004; Strobl et al., 2007; Gregorutti et al., 2017; Williamson et al., 2021).

In this study, permutation-based variable accuracy importance (PVAI), as a commonly used method, is employed to assess the importance of input feature for the machine learning models. The rationale of the method is that the importance of a feature is measured by the increase in the prediction error of the ML model after the values of the feature are randomly permuted. If the prediction performance (e.g., accuracy) decreases more under the permutation of a feature, it implies that the feature has a higher importance (Breiman, 2001).

4 Experimental results

4.1 Usage efficiency analysis

In this section, the usage efficiency patterns of e-scooter sharing services are explored in time and space. According to Eq. (1), given a pair of consecutive trips, the TtB can be calculated. Since the TtB value for the last trip of each e-scooter during the selected period is not able to calculate, those trips are removed after calculating TtB values. Eventually, 702,138 trips with TtB are preserved for the experiment.

First, the statistical analysis is conducted by examining the probability distribution and cumulative probability distribution of TtB based on the trips. As shown in Figure 2, it can be observed that the idle parking less than one hour and greater than 8 hour occupies 45% and 20% respectively. It implies that there is a certain potential to further improve the usage efficiency of the shared e-scooters. In addition, we further calculate the mean and median of TtB, which are 326 min and 70 min respectively.

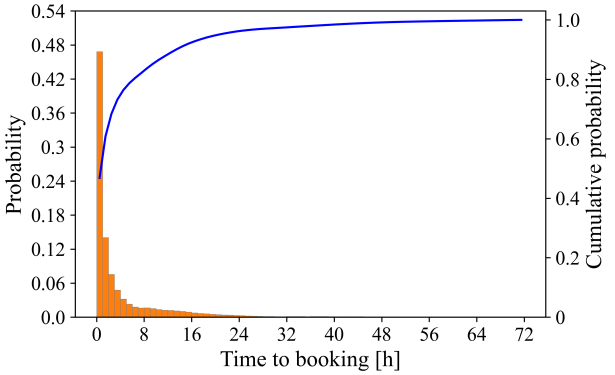

Figure 2. Probability distribution and cumulative probability distribution of Time to Booking.

Second, the temporal analysis is conducted to explore how the TtB varies over time. In this study, the entire idle parking is aggregated on hourly basis to describe the variations of TtB on weekday and weekend, as shown in Figure 3. It can be observed that the usage efficiency of the shared e-scooters presents the similar patterns on weekday and weekend. For example, the shared e-scooters display high usage efficiency (i.e., low TtB) during daytime (from 8:00-17:00) irrespective of weekday or weekend. However, the fluctuations of the usage efficiency are more pronounced in the nighttime, which is consistent with human travel behavior patterns.

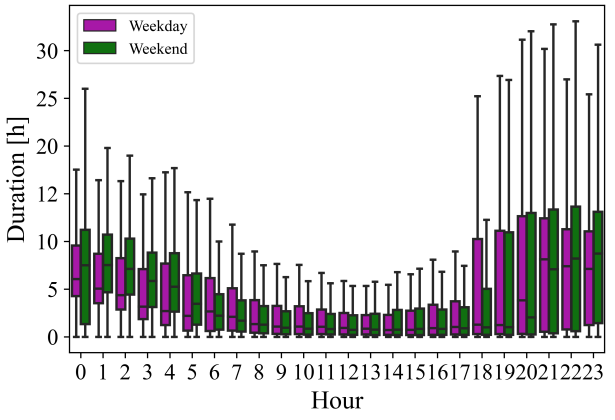

Figure 3. Temporal variations of Time to Booking on hourly basis on weekday and weekend.

Next, we further analyze how the usage efficiency of the shared e-scooters varies in space on weekday and week-

end. Figure 4 visualizes the spatial distributions on the usage efficiency in terms of TtB on weekday and weekend. Here, the median of TtB in each administrative division unit is calculated and employed to quantify the usage efficiency of e-scooter sharing services in different areas. The green color represents lower TtB (i.e., higher usage efficiency) while blue color represents higher TtB (i.e., lower usage efficiency). Grey color indicates that there is no data (or trips by e-scooter sharing service). By comparing the spatial variations of TtB on weekday and weekend, we can conclude that the areas with high usage efficiency of e-scooter sharing services are mainly concentrated on city center of Stockholm on both weekday and weekend. The areas with low usage efficiency are distributed in the periphery. One interesting finding is that the medians of TtB in some downtown areas exceed two hours. This is because the main city of Stockholm is connected by 14 islands and a peninsula through more than 70 bridges. The e-scooter sharing services do not display high usage efficiency in several islands of city center.

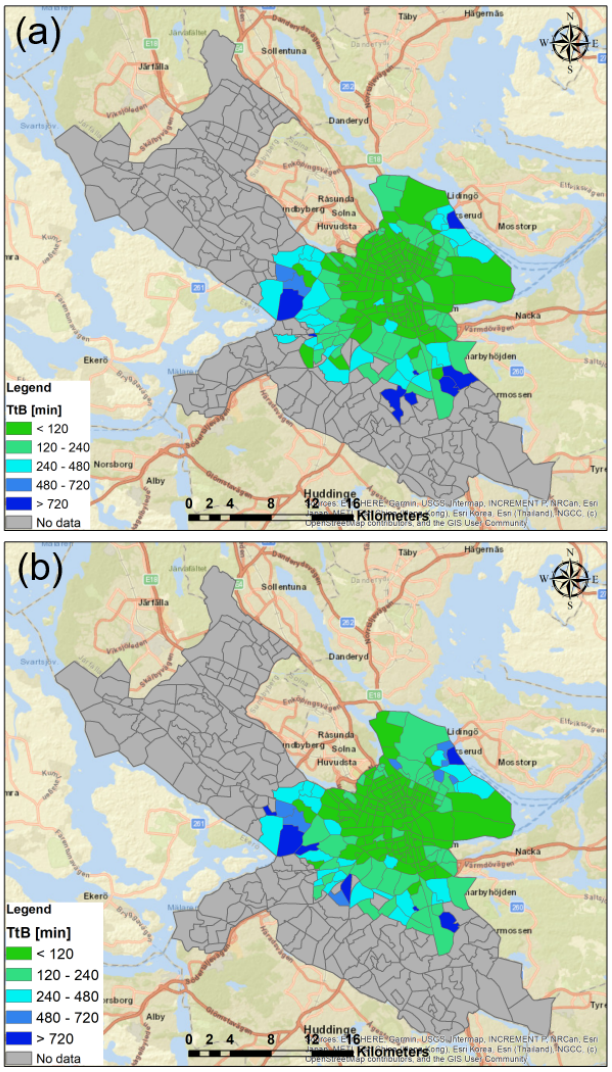

Figure 4. Spatial variations of Time to Booking on (a)weekday and (b) weekend.

4.2 Usage efficiency prediction

In this section, the above-mentioned three typical ML methods are used to predict the usage efficiency of shared e-scooters. According to the cumulative probability distribution of Time to booking in Figure 2, it is reported that the idle parking less than 2 hour occupies 60% of the entire idle parking of shared e-scooters. Here, we choose 2 hour as threshold to divide the whole idle parking of shared e-scooters into two parts, namely high usage efficiency and low usage efficiency. In this situation, the usage efficiency prediction is transformed into a binary classification problem. In this study, the idling parking with high usage efficiency belongs to class 1 (positive), and the idling parking with low usage efficiency belongs to class 0 (negative).

First, the whole dataset including high usage efficiency and low usage efficiency of idling parking is randomly split into 491,496 training data instances (70%) and 210,642 test data instances (30%). In the training and test data, the ratios of high usage efficiency instances to low usage efficiency instances are both 1:1.5. Second, grid search with 5-fold cross-validation is conducted to tune the hyperparameters of each classifier using training data. Moreover, the same training and test data are used in all the three models to make the performance evaluation results comparable. Figure 5 presents the prediction performance of the ML models in terms of the four evaluation metrics.

As shown in Figure 5, regarding the prediction performance in terms of accuracy and F1 score, the RF model yields the best performance (accuracy = 71.2%, F1 = 78.0%), which is closely followed by the ANN model (accuracy = 69.0%, F1 = 77.0%), and then followed by the LR model (accuracy = 63.2%, F1 = 74.5%). It is found that the selected ML models are capable of predicting the usage efficiency of shared e-scooters well. Concerning precision, the RF model also achieves the highest performance (precision = 72.8%), which is followed by the ANN model (precision = 69.8%) and then the LR model (precision = 64.1%). Taking the RF model as example, it implies that 27.2% of the predicted idling parking of high usage efficiency stems from the misclassification of the idling parking with low usage efficiency. It should be noted that, in terms of recall, the best prediction performance is achieved by the LR model (recall = 89.0%), followed by the ANN model (recall = 85.9%) and the RF model (recall = 83.8%). This is because the LR model obtains the lowest false-negative. In combination with the relatively low prediction accuracy of the LR model, it can be concluded that the LR model has a relatively poor prediction on the low usage efficiency instances compared with the RF model and the ANN model.

In summary, the selected three ML methods achieve the satisfying performance on the prediction of e-scooter sharing services at the trip level. Among them, random forest displays the best prediction performance

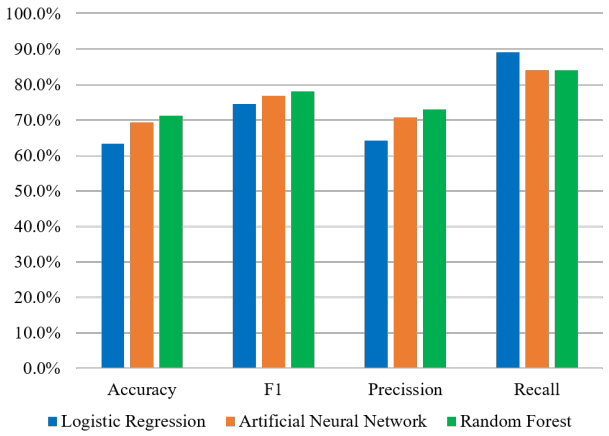

Figure 5. The prediction performance of the three employed ML methods in terms of the evaluation measures.

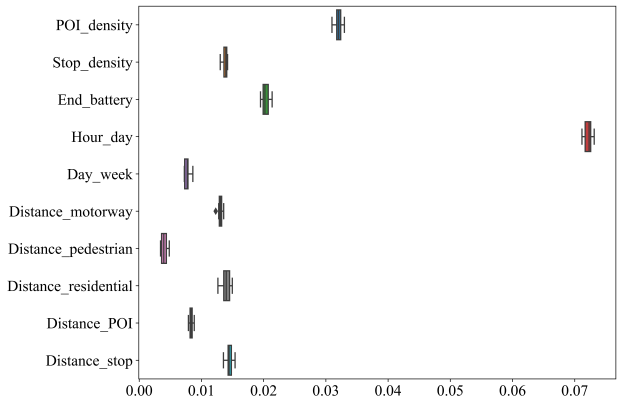

Figure 6. Feature importance in the Random forest model.

4.3 Variable importance analysis

In this work, the PVAI method introduced in section 3.6 is implemented to measure the importance of the ten variables/features. Specifically, each feature is permuted 10 times in the test data. The importance is measured by the reduction in accuracy. We here present the variable importance analysis result by taking the random forest model as an example.

Figure 6 presents the importance of each feature with boxplot. It is found that the hour of day has the highest feature importance in the RF model, which is supported by the findings in Figure 3. From Figure 3, it is found that the shared e-scooters display different usage efficiencies in terms of Time to Booking. Next, POI density is displayed as the second important features in the RF model. POI density describes the intensity of the spatial distribution of POIs in an area, which is positively associated with pedestrian activity intensity. Since the usage efficiency of shared e-scooters relies heavily on their surrounding pedestrian activities, it explains why POI density presents high feature importance. It should be noted that battery power of trip end is also displayed as an important feature in the RF model. Compared with the above-mentioned three features, other features have low importance values.

5 Conclusion and future work

The introduction of e-scooter sharing services to cities results in significant environmental and socioeconomic benefits, while some existing related issues still require to be solved. Due to the flexibility nature of shared e-scooters, how to further improve the usage efficiency of e-scooter sharing services has always been one of the main concerns of micro-mobility operators and transport planners. The emergence of GPS-based vehicle availability data provides the possibility to investigate the usage efficiency of shared e-scooters at a high spatial and temporal resolution. In addition, machine learning has been demonstrated its usefulness in the fields of GIS and transportation. Driven by the goal of urban sustainable urban mobility, we investigate the prediction on the usage efficiency of shared e-scooters based on the GPS-based vehicle availability data and machine learning models in this study. The main findings of this study are summarized as follows.

First, the vehicle availability data collected from a micro-mobility operator is employed to measure the usage efficiency of e-scooter sharing services in Stockholm in terms of Time to Booking. The temporal and spatial analysis results suggest that the usage efficiency of shared e-scooters has notable variable patterns over time and space. Second, three typical machine learning models, including logistic regression, artificial neural network and random forest, are used to predict the usage efficiency of shared e-scooters based on the extracted ten variables. With regards to the evaluation metrics, the random forest model achieves the best prediction performance. Last, the variable importance analysis is implemented based on the random forest model. The results show that three features, including hour of day, POI density, battery power of trip end, present high importance compared with other features. This study has important implications with respect to further optimize e-scooter sharing services by micro-mobility operators.

However, there are several limitations in the current study, which could be considered as directions for future work. First, the current features only consist of several basic variables in time and space. Some socioeconomic and weather variables require to be taken into account, which have a potential to further improve the prediction performance of the machine learning models. Second, only the feature importance analysis is implemented to improve the interpretability of machine learning models. It would be more meaningful to attempt some more interpretability analysis techniques of machine learning, such as partial dependence plots, feature interaction. Last but not least, this study divides the usage efficiency of shared e-scooters into high and low categories by setting a threshold two hour according to the statistical distribution of TtB. It would be

more realistic to take the prediction of TtB as a regression problem to meet the various requirements of micromobility operators.

References

Aditian, A., Kubota, T., and Shinohara, Y.: Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia, Geomorphology, 318, 101–111, 2018.

Almannaa, M. H., Ashqar, H. I., Elhenawy, M., Masoud, M., Rakotonirainy, A., and Rakha, H.: A comparative analysis of e-scooter and e-bike usage patterns: Findings from the City of Austin, TX, International Journal of Sustainable Transportation, 15, 571–579, 2021.

Bai, S. and Jiao, J.: Dockless E-scooter usage patterns and urban built Environments: A comparison study of Austin, TX, and Minneapolis, MN, Travel behaviour and society, 20, 264–272, 2020.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

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

Caspi, O., Smart, M. J., and Noland, R. B.: Spatial associations of dockless shared e-scooter usage, Transportation Research Part D: Transport and Environment, 86, 102 396, 2020.

Dias, G., Arsenio, E., and Ribeiro, P.: The role of shared E-Scooter systems in urban sustainability and resilience during the Covid-19 mobility restrictions, Sustainability, 13, 7084, 2021.

Du, Y., Deng, F., and Liao, F.: A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system, Transportation Research Part C: Emerging Technologies, 103, 39–55, 2019.

Gregorutti, B., Michel, B., and Saint-Pierre, P.: Correlation and variable importance in random forests, Statistics and Computing, 27, 659–678, 2017.

Gu, Z., Zhu, Y., Zhang, Y., Zhou, W., and Chen, Y.: Heuristic Bike Optimization Algorithm to Improve Usage Efficiency of the Station-Free Bike Sharing System in Shenzhen, China, ISPRS International Journal of Geo-Information, 8, 239, 2019.

Guidon, S., Becker, H., and Axhausen, K.: Avoiding stranded bicycles in free-floating bicycle-sharing systems: using survival analysis to derive operational rules for rebalancing, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 1703–1708, IEEE, 2019.

Guo, Y. and Zhang, Y.: Understanding factors influencing shared e-scooter usage and its impact on auto mode substitution, Transportation research part D: transport and environment, 99, 102 991, 2021.

Guo, Y., Zhou, J., Wu, Y., and Li, Z.: Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China, PloS one, 12, e0185 100, 2017.

Heumann, M., Kraschewski, T., Brauner, T., Tilch, L., and Breitner, M. H.: A Spatiotemporal Study and Location-Specific Trip Pattern Categorization of Shared E-Scooter Usage, Sustainability, 13, 12 527, 2021.

Hosseinzadeh, A., Algomaiah, M., Kluger, R., and Li, Z.: E-scooters and sustainability: Investigating the relationship between the density of E-scooter trips and characteristics of sustainable urban development, Sustainable cities and society, 66, 102 624, 2021.

Huo, J., Yang, H., Li, C., Zheng, R., Yang, L., and Wen, Y.: Influence of the built environment on E-scooter sharing ridership: A tale of five cities, Journal of Transport Geography, 93, 103 084, 2021.

Jiao, J. and Bai, S.: Understanding the shared e-scooter travels in Austin, TX, ISPRS International Journal of Geo-Information, 9, 135, 2020.

Li, A., Zhao, P., Huang, Y., Gao, K., and Axhausen, K. W.: An empirical analysis of dockless bike-sharing utilization and its explanatory factors: Case study from Shanghai, China, Journal of Transport Geography, 88, 102 828, 2020.

Li, A., Zhao, P., Haitao, H., Mansourian, A., and Axhausen, K. W.: How did micro-mobility change in response to COVID-19 pandemic? A case study based on spatial-temporal-semantic analytics, Computers, Environment and Urban Systems, 90, 101 703, 2021.

Li, A., Zhao, P., Liu, X., Mansourian, A., Axhausen, K. W., and Qu, X.: Comprehensive comparison of e-scooter sharing mobility: Evidence from 30 European cities, Transportation Research Part D: Transport and Environment, 105, 103 229, 2022.

Losapio, G., Minutoli, F., Mascardi, V., and Ferrando, A.: Smart Balancing of E-scooter Sharing Systems via Deep Reinforcement Learning, 2021.

McKenzie, G.: Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services, Computers, Environment and Urban Systems, 79, 101 418, 2020.

Menard, S.: Applied logistic regression analysis, vol. 106, Sage, 2002.

Miller, T.: Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence, 267, 1–38, 2019.

Mitchell, T. M.: Artificial neural networks, Machine learning, 45, 81–127, 1997.

Olden, J. D., Joy, M. K., and Death, R. G.: An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data, Ecological Modelling, 178, 389–397, 2004.

Pun, L., Zhao, P., and Liu, X.: A multiple regression approach for traffic flow estimation, IEEE Access, 7, 35 998–36 009, 2019.

Robert, C.: Machine learning, a probabilistic perspective, 2014.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC bioinformatics, 8, 25, 2007.

Wang, S., He, T., Zhang, D., Liu, Y., and H. Son, S.: Towards efficient sharing: A usage balancing mechanism for bike sharing systems, in: The World Wide Web Conference, pp. 2011–2021, 2019.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M.: A general framework for inference on algorithm-agnostic vari-

able importance, Journal of the American Statistical Association, pp. 1–38, 2021.

Zhao, P., Kwan, M.-P., and Qin, K.: Uncovering the spatiotemporal patterns of CO2 emissions by taxis based on Individuals' daily travel, Journal of Transport Geography, 62, 122–135, 2017.

Zhao, P., Haitao, H., Li, A., and Mansourian, A.: Impact of data processing on deriving micro-mobility patterns from vehicle availability data, Transportation Research Part D: Transport and Environment, 97, 102 913, 2021.

Zhu, R., Zhang, X., Kondor, D., Santi, P., and Ratti, C.: Understanding spatio-temporal heterogeneity of bike-sharing and scooter-sharing mobility, Computers, Environment and Urban Systems, 81, 101 483, 2020.

Ziedan, A., Shah, N. R., Wen, Y., Brakewood, C., Cherry, C. R., and Cole, J.: Complement or compete? The effects of shared electric scooters on bus ridership, Transportation Research Part D: Transport and Environment, 101, 103 098, 2021.