Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 21(7): 6471–6492.
DOI: 10.3934/mbe.2024282
Received: 08 January 2024
Revised: 02 March 2024
Accepted: 14 May 2024
Published: 02 July 2024

Research article

Divide-and-train: A new approach to improve the predictive tasks of bike-sharing systems

Ahmed Ali[1,2,*], **Ahmad Salah**[3,4], **Mahmoud Bekhit**[5,6] **and Ahmed Fathalla**[7]

[1] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

[2] Higher Future Institute for Specialized Technological Studies, Cairo 3044, Egypt

[3] College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Ad-Dhahirah, Sultanate of Oman

[4] Zagazig University Department of Computer Science, Faculty of Computers and Informatics, Zagazig, Sharkeya, Egypt

[5] Peter Faber Business School, Australian Catholic University (ACU), Sydney, Australia

[6] Kaplan Business School, Sydney, Australia

[7] Department of Mathematics, Faculty of Science, Suez Canal University, Ismailia, Egypt

* **Correspondence:** Email: a.abdalrahman@psau.edu.sa.

Abstract: Bike-sharing systems (BSSs) have become commonplace in most cities worldwide as an important part of many smart cities. These systems generate a continuous amount of large data volumes. The effectiveness of these BSS systems depends on making decisions at the proper time. Thus, there is a vital need to build predictive models on the BSS data for the sake of improving the process of decision-making. The overwhelming majority of BSS users register before utilizing the service. Thus, several BSSs have prior knowledge of the user's data, such as age, gender, and other relevant details. Several machine learning and deep learning models, for instance, are used to predict urban flows, trip duration, and other factors. The standard practice for these models is to train on the entire dataset to build a predictive model, whereas the biking patterns of various users are intuitively distinct. For instance, the user's age influences the duration of a trip. This endeavor was motivated by the existence of distinct user patterns. In this work, we proposed *divide-and-train*, a new method for training predictive models on station-based BSS datasets by dividing the original datasets on the values of a given dataset attribute. Then, the proposed method was validated on different machine learning and deep learning models. All employed models were trained on both the complete and split datasets. The enhancements made to the evaluation metric were then reported. Results demonstrated that the proposed method outperformed the conventional training approach. Specifically, the root mean squared error (RMSE) and mean absolute error (MAE) metrics have shown improvements in both trip duration and distance

prediction, with an average accuracy of 85% across the divided sub-datasets for the best performing model, i.e., random forest.

Keywords: bike-sharing system; *divide-and-train*; ensemble learning; machine learning; prediction; trip duration

1. Introduction

In a smart city, many tasks are fully monitored and recorded in datasets such as online park reservation [1], crowd flow prediction [2], and bike rides [3]. Bike riding in smart cities is a key source of information. In smart cities, it is expected that most trips within a city will be made by bikes. This is because there is a need for carbon-neutral cities; thus, there has been a great demand for bike-sharing systems [4]. About one billion users are projected to be part of the bike-sharing market in the future [5, 6]. The growing popularity of carbon-neutral bike sharing has made disparities in urban mobility worse. The rapid transformation of urban environments is tangible. Bike riding is one of these aspects that has witnessed a significant transformation. In smart cities, any bike trip is monitored by the BSS, and data is collected for every single trip. In addition, the BSS aims to lessen traffic, encourage exercise, and fight climate change.

There are several datasets of bike-sharing systems that are generated by the companies and cities dataset [7–9]. For instance, the New York City bike share dataset [10] is accessible to the public. It enables bicycle travel across all of New York City. This dataset includes data from hundreds of thousands of bike trips, gathered between January 2015 and June 2017. The dataset stores data about the users and the bike trips. Another bike-sharing dataset was collected in Los Angeles City between July 2016 and September 2021. This dataset includes information for over one million bike trips. These two examples outline the volume of the BSSs datasets. However, considering the entire dataset for training the ML and DL models has a key limitation which is the improper separation of different data patterns. For instance, while some datasets have two data patterns, it is easier to train the predictive model on a dataset with one data pattern. Splitting the dataset enables the effective determination of various patterns, such as user's age or gender, which aids in highlighting the relevant features within these datasets.

Machine learning and deep learning models are used heavily to predict data about BSS trips, such as trip duration, trip distance, and the number of trips in a given geographical area [11–13]. Predicting these data can help in deciding the proper number of bikes to be rented at a given bike station [4, 14, 15]. Several machine learning methods were utilized to predict the BSS data, such as random forest (RF), decision trees, and logistic regression. On the other hand, several deep learning models are used for the same purpose, such as a convolutional neural network (CNN), a gated recurrent unit (GRU), and long short-term (LSTM) [16].

The purpose of this study is to address two research concerns for station-based bike sharing systems. First, what is the performance gap between predictive models trained on the full BSS dataset and models trained on BSS sub-datasets divided by user attributes, namely, start station, age, and gender? Second, which machine learning or deep learning model's performance will the proposed method improve the most? This work is motivated by answering these two questions.

The proposed model commences with the *divide-and-train* training method which is inspired by the well-known *divide-and-conquer* method. The novelty of this research is that it is the first study to explore whether it is more advantageous to train on the entire BSS dataset as a single entity or to segment the dataset into smaller sub-datasets based on the values of a particular feature and then train on each sub-dataset independently. In the proposed method, the BSS dataset is split based on selective attributes. For instance, splitting the entire dataset on the gender attribute, there will be two sub-datasets obtained; one for male users and the other for female users. Afterward, the results of the predictive models trained on the complete dataset for both genders will be compared against the results of the same model trained on male users' data and similarly for the female users' data. The intuition of this proposed method is to reduce data complexity and thoroughly ease the predictive problem. Since the BSS dataset is characterized by its huge size, dividing it will still result in big sub-datasets which will be sufficient for training purposes. The list of the contributions of this work is as follows:

1) A new method for training the predictive models on the BSS dataset is proposed; it is called *divide-and-train*.

2) Using a real-world dataset, the proposed method is compared against the standard method of training. The obtained results reveal that the proposed method outperformed the standard training model by a huge performance gap.

3) Two different model types, namely, machine learning and deep learning, are compared to figure out which type will benefit the most from the proposed method, *divide-and-train*.

The rest of the paper is organized as follows. In Section 2, some of the existing methods are discussed. Section 3 exposed the proposed methods. The results are exposed and discussed in Section 4. Finally, the paper is concluded in Section 5.

2. Background and related work

Accurately forecasting the bike movement in an urban context is crucial for evaluating the behavior of people and the resource utilization within a city. The BSS usually operates in two modes of bike sharing, namely station-based and dockless systems. The idea behind BSS can be briefly described as a network of publicly accessible bikes distributed across a city available for renting by different users at affordable rates [17]. Station-based systems consist of a specific set of stations associated with bike slots that can be retrieved and returned to the same station from where they were first obtained. However, this leads to uneven demands since some stations may be completely crowded while others remain underutilized [18]. To address this problem, the dockless bike sharing system was developed, which enables users to have quick access to nearby bikes without the need for docked stations. This system saves users' time by locating nearby bikes, eliminates the expenses of constructing stations, and offers bike-sharing services at a low cost. Yet, dockless systems require accurate forecasting and estimation in order to prevent insufficient bike availability in heavily populated regions [19–21]. In this research, the main focus is on the station-based system.

This section presents an overview of the literature pertaining to the machine learning and deep learning models utilized in this study, specifically focusing on the GRU model, prediction, and clustering models.

2.1. Gated recurrent unit

A GRU model is built up from cells where a cell consists of an update $z_{(t)}$, reset gate $r_{(t)}$, and a hidden state vector $h_{(t)}$. The building block of a gate is one layer of neurons. The mathematical notation of the GRU cell mechanism is explained in Eqs (2.1)–(2.4). The inputs of a GRU cell are $h^{(t-1)})$ and $x^{(t)}$ for the previous cell's hidden state and the current input vector, respectively. On the other hand, a GRU cell produces an output hidden state which is denoted as $h^{(t)}$.

$$z^{(t)} = \sigma(W_z x^{(t)} + U_z h^{(t-1)} + b_z) \tag{2.1}$$

$$r^{(t)} = \sigma(W_r x^{(t)} + U_r h^{(t-1)} + b_r) \tag{2.2}$$

$$\tilde{h}^{(t)} = tanh(W_h x^{(t)} + U_h(h^{(t-1)} \odot r^{(t)}) + b_h) \tag{2.3}$$

$$h^{(t)} = z^{(t)} \odot h^{(t-1)} + (1 - z^{(t)}) \odot \tilde{h}^{(t)} \tag{2.4}$$

where $\odot$ is an element-wise multiplication, and $\sigma(.)$ and $tanh(.)$ are two activation functions of the neural network. As a feed-forward neural network consists of a set of weights, a GRU cell's feed-forward neural network weights are represented as W_z, W_r, and W_h. The recurrent neural network weights of a GRU cell are U_z, U_r, and U_h. Finally, the model biases are indicated by b_z, b_r, and b_h.

Deep learning algorithms, such as GRU, are recently utilized to anticipate station capacity and bike usage [22]. The rationale behind this lies in the advantages provided by deep learning in obtaining accurate predictions of outcomes in BSS. Researchers mainly focus on short-term demand forecasts and station-level flow. Yao and Feng [3] introduced a station-level based prediction method using a gated convolutional neural network (GCNN). Their objective was to demonstrate the diversity among the incoming and outgoing movements of bike stations. Li et al. [23] have enhanced the accuracy of short-term demand forecasts in urban BSS by using an irregular convolutional LSTM model. The main purpose was to recognize comparable temporal trends in bike utilization. This study was conducted using a dockless bike sharing system in Singapore and four docked systems in different cities, including Washington, D.C., Chicago, New York, and London. In a similar manner, the graphical convolutional neural networks (GCN) and GRU were utilized to identify how a BSS changes over time based on real historical data [24]. Other studies have specifically utilized deep learning techniques, particularly the GRU and LSTM, in order to obtain optimal predictions of users' demand and station capacities [25–28]. However, efficient prediction was lacking owing to many factors, including the restricted availability of data on real locations and the absence of seasonal aspects during the training of the data. Furthermore, it is necessary to include extra attributes such as gender, rental duration, and age of the users to get a precise forecast that aligns with the actual demand of the users.

2.2. Prediction methods

The existence of BSS in urban areas has been regarded as a green and sustainable mobile service that is increasingly gaining traction. The application of deep learning algorithms for forecasting BSS is crucial to understanding and anticipating their usage patterns. Historically, various methods have been employed to address the demand for these systems.

Li et al. [29] introduced an innovative technique called the spatial-temporal memory network (STMN) for the sake of predicting the utilization of short-term trips in BSSs. The proposed STMN model was utilized along with convolutional long short-term memory (LSTM) models to identify the spatial-temporal dependencies in the dataset. Their experiments were conducted on four distinct bike-sharing systems across different cities, including station-based systems in Chicago and New York and dockless systems in Singapore and New Taipei City. Similarly, Ma et al. [30] presented a groundbreaking method for estimating bike-sharing rental and return demand at individual station levels. The authors proposed a deep learning framework that utilizes multi-source data, such as weather information, user demographic data, land-use patterns, and historical bike-sharing trip data, to capture the spatiotemporal dynamics of bike-sharing usage. Implementing this model can enhance route selection, dynamic redistribution strategies, and the overall user experience.

Various methodologies have been suggested in the literature to address the complexities of predicting the movement of people and resources within a city. In [31], the authors provided an exhaustive analysis of the existing methods for predicting urban flow. As a case study, the authors compared various prediction techniques, including statistical, transfer learning, reinforcement learning, deep learning, and traditional machine learning. The hierarchical consistency prediction (HCP) model was proposed in [15] for predicting citywide bicycle usage. The HCP was comprised of three parts: the adaptive transition constraint (AdaTC) clustering algorithm, the similarity-based efficiency Gaussian process regressor (SGPR), and the general least square (GLS) formulation.

The main obstacle is to accurately select and extract the input features and use the temporal and spatial data to efficiently achieve customer satisfaction. Hence, deep learning models have emerged as one of the most effective approaches to addressing these difficulties. Liu et al. [32] proposed a novel approach to address these challenges by proposing a novel deep learning model for predicting bike sharing. The model, known as a convolutional autoencoder network (CAN), employed convolutional and autoencoder layers to capture both temporal and spatial data information. The authors evaluated the predictive accuracy of their model using real-world bike-sharing data from two cities and concluded that it outperformed existing models. In [16], the authors presented an alternative method for predicting the hourly rental demand for a bike-sharing system using a spatial-temporal attention mechanism. To extract spatial and temporal features, respectively, they used a CNN model and a GRU model.

For tracking the in-out flow prediction of bikes, the authors of [33] proposed a deep learning framework called the temporal attention graph convolutional network (TAGCN). This model takes spatial and temporal dependencies between stations into account, as well as the impact of hourly, daily, and weekly time scales on demand. In the study [34], the authors compared various techniques for short-term bike and slot availability forecasting. The focus of the study was on forecasts made at 15-, 30-, 45-, and 60-minute intervals, and the comparison included numerous characteristics and predictive models. Even with limited historical data, the results of the study indicated that deep learning models, particularly those employing bidirectional long short-term memory (BiLSTM), can accurately predict bike and slot availability.

The use of deep learning approaches proved to be beneficial in predicting the flow of bikes in urban environments. In addition, they successfully resolved issues related to data imbalance and the non-linearity of spatiotemporal data. However, there are also concerns that need to be investigated, including how to analyze seasonality data and weather effectively, select suitable input features, and determine the maintenance process of bikes in a dockless system. In general, the growing demand for

BSS has required the development of methods to obtain seamless and efficient services that should be delivered to BSS users.

2.3. Clustering methods

Current bike-sharing systems are typically split into two categories: station-based and dockless. Station-based bike sharing involves borrowing and returning bicycles to designated stations. The bike ride begins at a station and concludes with the bike being left on an empty dock. On the other hand, the dockless method imposes no limitations on where users can leave their bikes after use (i.e., bikes can be placed anywhere by the users). In [35], a rebalancing bike approach for station-based systems dependent on users' daily travel patterns was presented. The authors hypothesized that stations could become empty or crowded depending on user demand throughout the day. The main intent of this research is to prevent user discontent caused by missed demands during bike pick-up and drop-off. This process involved separating the extracted patterns into both positive and negative critical stations, which are then utilized to plan rebalancing operations.

To estimate the right number of bikes and free docks available to users, a polynomial-size clustering approach was devised [36]. This study examines both the needs at bike stations and the costs associated with rebalancing the inventory. Each cluster is comprised of a collection of self-sufficient stations. Hence, bike-balancing activities are handled within cluster stations. The stochastic demand for each station was calculated using mixed-integer programming techniques. This technique lacks information on the needs of other clusters, resulting in an uneven distribution of bikes based on the proportion of customer demand in each location.

In addition, authors in [37] present the unbalanced bike problem using clustering in station-based systems. The authors identified two groups of dissatisfied users: those who are unable to pick up bikes and those who cannot locate free docking stations. The authors built a real-time decision-support model for bike-sharing systems based on these classifications. This data-driven learning-based simulation method was presented to circumvent the time-consuming nature of simulations involving rebalancing issues. Another aspect that influences the re-balancing strategy is the problem of faulty bikes in dockless systems, where all shared bikes are considered inventory [38]. Unfortunately, the presence of defective bikes may lead to inadequate estimations of actual real-time price and quantity, posing rebalancing issues. In [39], an optimal framework model for resolving faulty bike-sharing concerns was proposed. K-means clustering was utilized to separate faulty bike sharing into multiple servicing stations to lower total recycling costs by optimizing the route. The authors considered an area in Beijing for study to verify the model's applicability. Authors in [40] recommended the placement of virtual stations be assigned according to user requirements. The clustering algorithm employed to produce a real-time answer was the K-means method. This research developed a mixed-integer linear programming model to enhance user demand, combined with the K-means method and CPLEX Optimizer to get the desired results.

3. Methodology

3.1. Overview

In the first stage of the proposed method, the BSS data are collected from a genuine BSS repository in order to ensure their quality and usability. The utilized dataset contains information about the user, including age and gender, as well as trip details. Subsequently, the proposed method divides the dataset

based on user characteristics or trip attributes, such as starting stations, to create clusters of observations with similar characteristics. This stage of data clustering permits the identification of distinct tendencies and patterns within each sub-dataset.

Once the data has been clustered, three machine learning models, namely random forest, Catboost, and bagging, and two deep learning models (i.e., GRU and LSTM) are trained on each sub-dataset to predict the trip's distance and duration. This procedure entails the development of predictive models that are tailored to the particular characteristics of each sub-dataset. Following the model training phase, the performance of the proposed method is evaluated on each sub-dataset by testing the trained models to identify discrepancies and inconsistencies. Then, the most accurate model for each sub-dataset is reported, ensuring the most reliable and accurate outcomes possible. Then, the same machine learning models are trained on the complete dataset, and the trained models are evaluated as well. It is worth mentioning that, the reason behind choosing the aforementioned machine and deep learning models is that they achieved state-of-the-art performance in numerous applications [41].

Finally, the performance of the model trained on the whole dataset is compared to that of the same machine learning model trained separately on each sub-dataset. Comparing the two training approaches, if the model trained on the sub-dataset obtained higher prediction rates, then the proposed method enhanced the prediction task. The objective of this methodology is to enhance the predictive performance of machine learning models in bike-sharing systems. The proposed method contributes to more accurate and reliable predictions of journey distance and trip duration by taking into account the unique characteristics of different observation groups.

In the proposed work, the *divide-and-train* method utilized three key attributes to divide the bike-sharing system dataset into sub-datasets, enabling the development of more accurate and reliable predictive models. Age, gender, and start station were selected based on their potential impacts on cyclists' behavior.

Age is a significant factor in determining bicycling habits, as people at different stages of life have distinct physical capabilities and preferences. By dividing the dataset by age, we were able to identify distinct trends within different age groups, resulting in a more accurate prediction of the trip's duration and distance. This age-based data split allowed the machine learning model to account for the varying bicycling patterns of younger, middle-aged, and older users and to develop models tailored to each group's unique characteristics.

Gender is an additional significant factor that influences bicycle use. When developing predictive models, it is essential to account for the fact that male and female cyclists may have distinct riding habits, preferences, and physical capacities. By separating the dataset by gender, we were able to identify patterns that were unique to each gender group, thereby increasing the accuracy and dependability of our models. This gender-based segmentation ensured that the distinct biking behaviors of both male and female users were considered, resulting in a more nuanced understanding of bike-sharing usage.

The start station is an important determinant of user behavior as it reflects the geographic distribution of bike-sharing users and their travel patterns. By dividing the dataset based on start stations, we were able to create clusters of users who frequently began their trips from the same location. This allowed the predictive model to analyze the specific travel patterns and preferences associated with specific areas or neighborhoods, enhancing the accuracy of our models. This geographic segmentation not only helped us better understand the spatial dynamics of bike-sharing usage, but it also provided valuable insights into the factors that influence user selection of start stations.

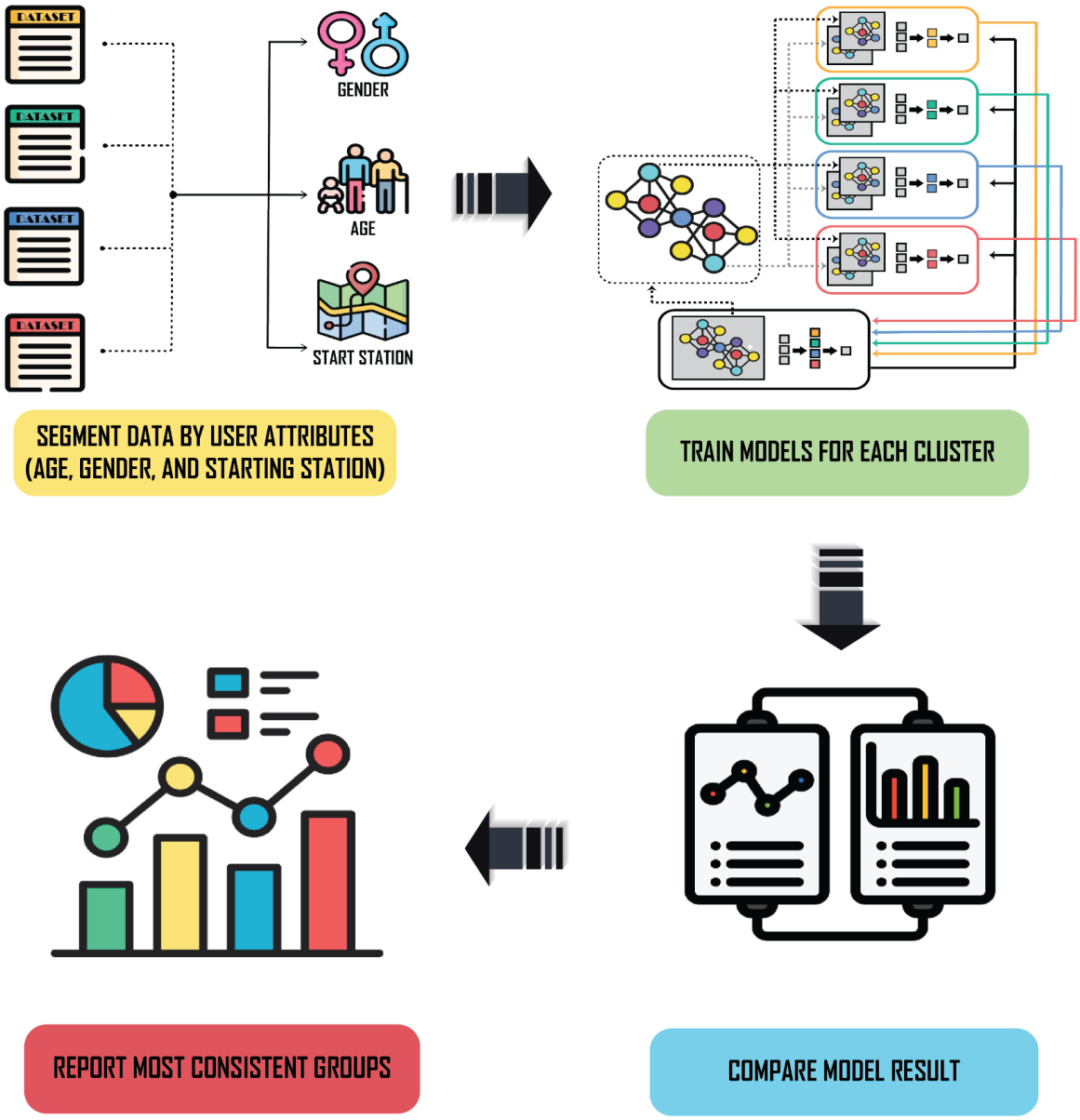


Figure 1. The predictive performance of the proposed methodology.

By dividing the dataset into sub-datasets based on these three attributes, the proposed predictive models are more accurate and trustworthy, tailored to the specific characteristics of each user group. This proposed technique is called *divide and train*. This is because the machine learning models and deep learning models are trained on the divided dataset instead of the complete dataset where the dataset is divided based on a given one attribute/feature.

In this study, we used three machine learning models, namely, random forest, Catboost, and bagging, and two deep learning models (i.e., GRU and LSTM) to predict trip distance and trip duration in bike-sharing systems. The random forest model is known for being reliable and able to handle many different kinds of data. It is made up of many decision trees that work together to make accurate predictions. Catboost is an innovative gradient-boosting model that excels at categorical features and generalization performance. The bagging model employs an ensemble of base learners trained on random subsets of the dataset in order to improve stability and reduce overfitting. By employing these cutting-edge machine learning models, our approach ensures accurate predictions tailored to the unique characteristics of each user group, thereby enhancing the overall performance and reliability of predictive models in the context of bike-sharing systems.

The default random forest model hyperparameters are the number of trees in the forest, the maximum tree depth, and the minimum sample size needed to split an internal node. Catboost's default hyperparameters are the learning rate, the depth of the trees, and the number of iterations of the gradient boosting process. The bagging model's default hyperparameters include the number of base estimators, the maximum number of samples for each base estimator, and the way the predictions of the base estimators are added together. Although set to their default values, these hyperparameters provided a solid foundation for the models and aided in achieving the desired level of predictive accuracy in the study.

The root mean squared error (RMSE) and mean absolute error (MAE) were employed as the error metric for evaluating the performance of the machine learning models. In regression problems, RMSE is a commonly employed metric for measuring the difference between the predicted and actual values. It computes the square root of the mean squared differences between the predicted and actual values, providing an interpretable and scale-sensitive performance metric for the model. Using RMSE, the study was able to compare the outcomes of the *divide-and-train* approach to those of the standard training approach, ultimately demonstrating the superior performance of the proposed method in predicting trip distance and trip duration for bike-sharing systems.

3.2. Implementation details

The data of each cluster/sub-dataset (as described in the previous section) is split into train and test sets with a splitting ratio of 80% and 20%, respectively. The training of the machine learning model is performed using default hyperparameter values. Thus, no hyperparameter tuning is applied in the training process. The reason behind that is that we want to study the clustering effect on machine learning models' performance. The utilized machine learning models in this work are random forest, bagging, and boosting due to the superior performance of these three models in the state-of-the-art methods.

For the utilized deep learning models, we designed a GRU model and an LSTM model. The choice of these two models stems from their proficiency in recognizing long-term dependencies, a crucial attribute for forecasting time series data characterized by cycles. This is particularly relevant for BSS

systems, which exhibit recurrent patterns in bike riding trips, including morning and evening peaks, as well as distinctions between working days and weekends. The GRU model architecture consists of a GRU layer of eight units. The GRU layer is followed by two fully connected layers of eight and four neurons, respectively. Finally, there is an output layer of a single neuron. The activation function of the fully connected layers is a rectified linear unit (ReLU), whereas the activation function of the output layer is linear. Furthermore, early stopping is utilized for stopping the training and avoiding overfitting when the generalization error increases. The Adam optimizer is used as an optimization algorithm to train the deep neural network. The proposed LSTM model architecture is similar to the GRU model, but the first layer is an LSTM layer instead of the GRU layer.

3.3. Evaluation metrics

The RMSE is the square root of the average of the squared differences between the predicted and actual values for a set of N observations, given by Eq (3.1). MAE measures the mean of the absolute differences between the actual values and the forecasted values, given by Eq (3.2).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2} \tag{3.1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \tag{3.2}$$

4. Results and discussion

4.1. Experimental setup

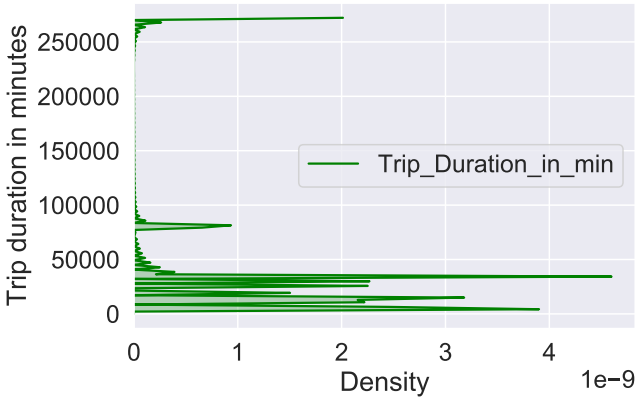
The experiments were conducted on a computer running 64-bit Windows-10 OS with two 2.6 GHz Intel 6-core processors. All of the utilized predictive models were implemented in the Python programming language version 3.9.16. Moreover, two Python packages are utilized to implement machine learning models, namely, the Scikit-learn [42] and Catboost [43] packages. Furthermore, the following libraries are utilized in our work: Pandas [44], NumPy [45], and Matplotlib [46]. The utilized dataset is publicly available alongside the code that was used to support the findings of this study as a GitHub repository*.

4.2. Dataset

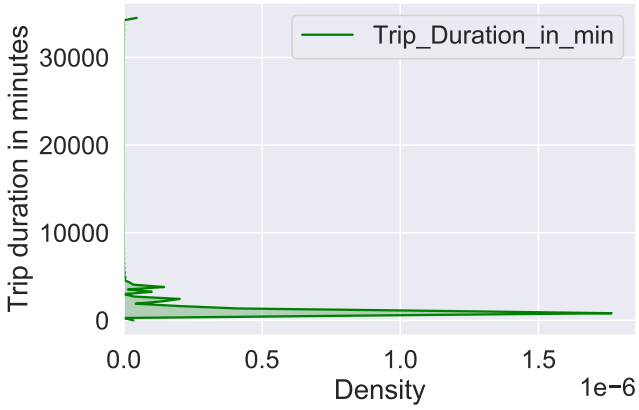
In this paper, the New York City bike share dataset is employed. It is a publicly available dataset that affords bike trips throughout the boroughs of New York City [10]. This dataset contains information on 735,502 anonymized trips collected between January 2015 and June 2017. The dataset has 17 features; the features represent information about the trip and the user. The primary features of the New York dataset are the trip duration in minutes, the start and end stations ID, the bike ID, the location of stations, the starting and ending points of the trip, and the start and end time of the trip. For the user's information, the dataset stores the user's gender and year of birth.

*https://github.com/stars-of-orion/BSS-Divide-and-Train

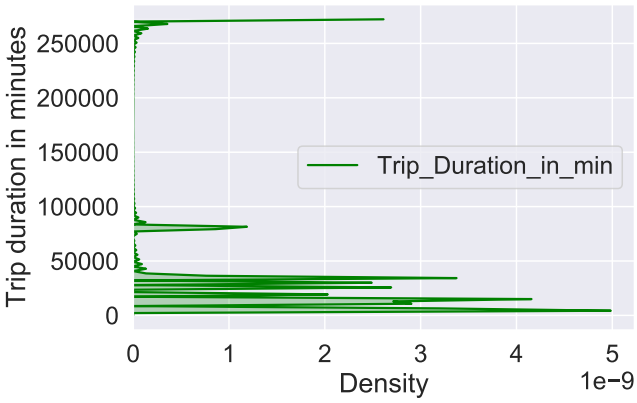
The dataset consists of 17 features. Four of these features (i.e., start time, end time, trip duration, and trip duration in minutes) were used to label the first predictive task, i.e., trip duration. For the trip distance label generation, another four features were utilized, namely, start station latitude, start station longitude, end station latitude, and end station longitude. Then, the dataset includes the station ID and station name features; we picked the station name to represent these two features in data splitting. The user type, gender, and birth year features were selected to split the data as well. The only unused feature in data splitting was the bike ID, as it can produce a huge number of groups.



(a) Data distribution for female and male users.



(b) Data distribution for female users.



(c) Data distribution for male users.

Figure 2. Data distribution for trip time duration in minutes based on the user's gender for the complete data and two split datasets.

One approach to show that two groups of data have a different patter is the data distribution. Thus, we proposed studying the distribution of the attribute before splitting and the distributions of the split groups. If there is a difference in the distribution between the complete dataset and the split dataset, then that means there are different patterns, as depicted in Figure 2. To illustrate the idea of different distributions of different value groups of the same attribute, we depicted the distribution of the trip duration attribute for the gender attribute including both female users' and male users' records in Figure 2a with the help of the kernel density estimation (KDE). Then, we depicted the record distribution using KDE for the trip duration of the female users' records and male users' records in Figure 2b,c, respectively. The three sub-figures of Figure 2 show different distributions based on the user's gender;

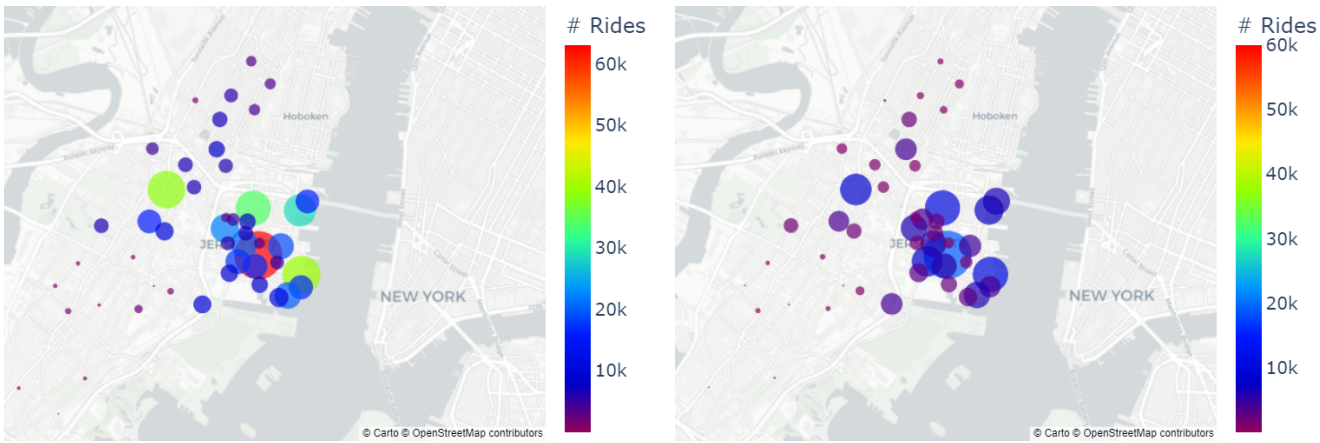
thus, there is a need to handle the male users' records in a different way than the female users' records. This emphasizes the idea of the proposed work to split the data based on the user's gender (i.e., female and male) and then train two models, each model on a sub-dataset.

Similar to the proposed work in [47], we performed a significant analysis for the two data groups of the female users' data and male users' data (i.e., trip duration and trip distance). The visual data distributions of these two data groups are not normal, as indicated in Figure 2b,c. The curves of these two figures are not bell-shaped. Thus, we examined whether there was a significant difference between these two groups using the Kruskal-Wallis and Mann-Whitney U tests. We assumed that the null hypothesis (H_0) is that there is no significant difference between these two groups. We analyzed the p-value, which is the probability under the assumption of no effect or no difference of data groups. If the p-value is less than 0.05, then there is a significant difference between the two groups. As the reported results of Table 1 show that the p-values are less than 0.05, then the null H_0 is rejected, and it can be concluded that there is a significant difference between these two groups of data. Thus, it is better to train on each data group/sub-dataset independently.

Table 1. Significant analysis for the trip duration and trip distance of the female and male users' data.

	Test	**p-value**
Male-Female trip duration groups	Kruskal-Wallis	< .001
Male-Female trip duration groups	Mann-Whitney U	< .001
Male-Female trip distance groups	Kruskal-Wallis	< .001
Male-Female trip distance groups	Mann-Whitney U	< .001

4.3. Results



(a) Density of rides for male users.

(b) Density of rides for female users.

Figure 3. Density for rides based on the user's gender: (a) male users and (b) female users.

The first point of comparison is a visual comparison between the density of rides between the female and male users which is depicted in Figure 3, similar to the work proposed in [48]. The circle size and color indicate the number of rides for each bike station. The circle color is the primary indicator of the number of rides followed by the circle size. For instance, if there is a large blue circle and smaller

green circle, then the smaller green circle indicates more number of rides relative to the larger blue circle, as the green color indicates higher density. The circle size can be used in comparison for circles with the same color. The two sub-figures of Figure 3 include different patterns of rides. Figure 3 proves the idea of different patterns based on the user's gender. In Figure 3a, there are green and red circles indicating a high number of rides while Figure 3b includes different color patterns indicating a lower number of rides. Thus, each gender's bike-sharing data should be handled as a separate dataset.

The second point of comparison is the improvement in the RMSE metric due to the dividing of the dataset into sub-datasets. The utilized equation for measuring the improvements or declines of the RMSE metric when the machine learning model is trained on the complete dataset is denoted in Eq (4.1). In Figure 4, the complete dataset RMSE improvements of trip duration prediction are depicted against all other age ranges for the Catboot model, which achieved a moderate performance between the machine learning models and the other two deep learning models. Other ML models produced similar improvement patterns; thus, we selected one model as representative of the machine learning models to reduce the number of figures. The greatest improvements were obtained for age ranges 60 to 65, 55 to 60, with more than 86% score improvement, and 45 to 50, with more than 87%. Similarly, the RMSE metric improvements in trip distance prediction are depicted in Figure 5. As shown in Figure 5, the average improvement is about 85%. Finally, the heat map is depicted in Figure 6 for the RMSE improvements for trip duration prediction when the data is divided based on the trip start station. Figure 6 shows a close RMSE score for predicting the divided sub-datasets with an average improvement of 85%.

$$\frac{(Full\_dataset\_score - subdataset\_score)}{Full\_dataset\_score} \times 100 \tag{4.1}$$

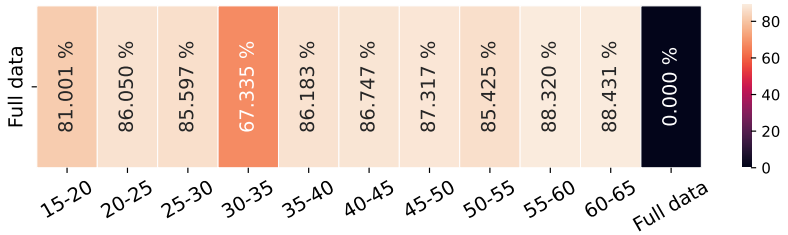


Figure 4. A heat map of RMSE metric for the trip distance prediction for different age classes using the ensemble Catboost model.

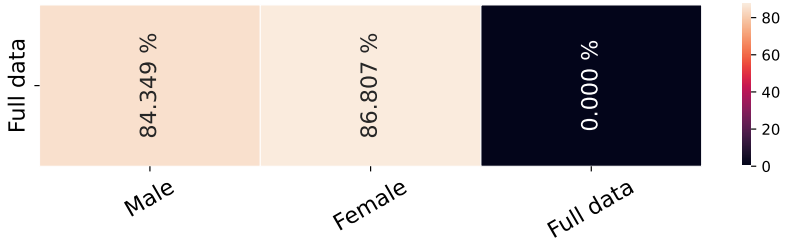


Figure 5. A heat map of RMSE metric for the trip distance prediction for different genders using the ensemble Catboost model.

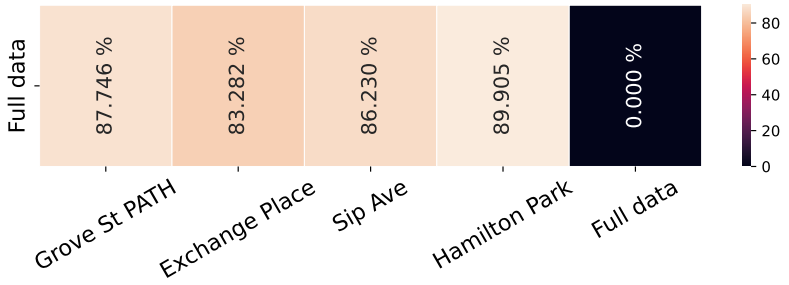


Figure 6. A heat map of RMSE metric for the trip distance prediction for different start stations using the ensemble Catboost model.

Table 2. MAE scores for the divided dataset by gender against the complete dataset for predicting trip duration.

Model	**Type**	**Male**		**Female**		**Full data**
			Imp. %		**Imp. %**	
Random Forest	ML	3.1	64.0%	4.0	53.5%	8.6
Catboost	ML	8.4	45.5%	8.7	43.5%	15.4
Bagging	ML	3.2	64.8%	4.4	51.6%	9.1
GRU	DL	7.8	34.5%	9.2	22.7%	11.9
LSTM	DL	8.0	45.2%	9.6	34.2%	14.6

Table 3. RMSE scores for the divided dataset by gender against the complete dataset for predicting trip duration.

Model	**Type**	**Male**		**Female**		**Full data**
			Imp. %		**Imp. %**	
Random Forest	ML	94.8	84.2%	75.5	87.4%	598.5
Catboost	ML	225.5	39.9%	162.5	56.7%	375.2
Bagging	ML	94.2	82.7%	108.9	80.1%	545.9
GRU	DL	266.4	31.5%	210.1	46.0%	389.1
LSTM	DL	266.4	31.5%	210.2	46.0%	389.3

The RMSE scores have been evaluated in Tables 3 and 5 . The results have been obtained by evaluating several machine learning models, namely random forest, Catboost, and bagging, along with the LSTM and GRU deep learning models. In Tables 3 and 5, the RMSE scores of the entire dataset (i.e., the complete dataset) are compared against those of the divided datasets based on the user gender in order to predict the trip distance and trip duration, respectively. The Catboost algorithm has superior performance when applied to the entire dataset. On the other side, the random forest algorithm outperforms the other three algorithms when applied to the split dataset. The RMSE metrics for all of the utilized models were improved for the proposed method, *divide-and-train*, as indicated in the *Imp.%*

column of Table 5. In addition, the proposed method has improved the RMSE for all of the three predictive models in the gender-based data split, as mentioned in Table 3.

Table 4. MAE scores for the divided dataset by gender against the complete dataset for predicting trip distance.

Model	Type	Male		Female		Full data
			Imp. %		Imp. %	
Random Forest	ML	95.6	4.8 %	87.6	12.7%	100.4
Catboost	ML	216.7	12.9%	185.6	25.4%	248.9
Bagging	ML	103.8	1.4%	96.1	8.7%	105.3
GRU	DL	296.4	31.6%	266.5	38.5%	433.2
LSTM	DL	360.3	18.8%	455.7	-2.7%	443.6

Table 5. RMSE scores for the divided dataset by gender against the complete dataset for predicting trip distance.

Model	Type	Male		Female		Full data
			Imp. %		Imp. %	
Random Forest	ML	220.3	84.0%	205.1	85.1%	1373.7
Catboost	ML	357.2	84.4%	301.1	86.6%	2282.6
Bagging	ML	243.0	-0.5%	225.6	6.7%	241.9
GRU	DL	486.5	97.8%	439.3	98.1%	22595.4
LSTM	DL	582.0	97.4%	639.2	97.2%	22608.8

The proposed method is evaluated on a different attribute split, e.g., trip start station. Tables 7 and 9 list the RMSE of the complete dataset against the top four trip start stations in the number of trips for predicting trip distance and duration, respectively. The behavior was similar to the results of Tables 3 and 5. The complete dataset prediction is in the last column of Table 5, and the random forest model achieved the lowest RMSE. In other words, the random forest model performs best for the split dataset for predicting both trip distance and duration. For instance, in Table 3, the improvements in the RMSE scores are almost six times greater for trip duration prediction. The complete dataset RMSE score is 598.5 while the male dataset and female dataset RMSE scores are 94.8 and 75.5, respectively. Besides, we can conclude that the predictive models can benefit from the proposed method of data split with different levels. The same experiments were repeated on the MAE metric where the results were reported in Tables 2, 4, 6, and 8.

The fourth point of the comparison is the proposed model's bias variance. We proposed using cross-validation analysis and the confidence interval (CI) statistical analysis for the best performing model, i.e., random forest, for predicting the trip distance of male users' data and female users' data. First, the cross-validation analysis was preformed at different folds, i.e., k for the RMSE metric. The k values are 2, 3, 4, and 5. Then, the CI for 95% was calculated to depict the confidence area using Eq (4.2). Finally, the results were depicted in Figure 7. In Figure 7, the curve represents the average RMSE value of the k-fold and the gray area around the curve shows the variance of the used k values. Figure 7a,b

show the variance of predicting the trip distance of the male users' and female users' data, respectively. Figure 7a shows a slighty higher bias and less model variance in comparison to Figure 7b. The results show a small margin of variance.

$$CI = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right) \tag{4.2}$$

where $z = 1.96$, σ represents the standard deviation of the k results of the k-fold cross-validation, and n represents folds, $n = k$.

4.4. Discussion

The reported results of the MAE and RMSE are similar to each other. The overall results show that the performance of the ML models in general was better than the deep learning models (i.e., GRU and LSTM). On average, the random forest model achieved the highest accuracy among all of the models for the split datasets. For the deep learning models, the GRU model slightly outperformed the proposed LSTM model.

To answer the first and second questions of this work in Section 1, the performance improvement should be investigated. For the first question, the performance gap between training the model on the complete dataset and the divided datasets varied from one model to another. There is no fixed improvement for all models. For the second question, the average performance gap for all of the reported results, Tables 2–9 show that the random forest model achieved the highest improvement in the performance on average after utilizing the proposed *divide-and-train* method. The highest reported improvement for one experiment was in Table 5 for the GRU model where the improvement hit 98.1%.

The proposed *divide-and-train* method can be utilized to improve the prediction of trip distance and trip duration. The accurate prediction of trip distance and duration can lead to an accurate estimation of the number of bikes in a station. As a result, the BSS company can make decisions in advance to face any expected shortage in the number of bikes per station through the proper logistic operation. The analysis findings provide significant insights into the trip characteristics of female and male users, as well as the influence of start station selection on trip duration and distance. These findings not only enhance the current knowledge base but also have multiple practical implications and indicate potential avenues for future research.

Our research contributes to the advancement of knowledge regarding gender-related differences in travel patterns, supported by significant statistical support ($p < .001$ for both the duration and distance of trips among different gender cohorts). Significant accuracy gains were observed in the prediction of trip duration and distance using random forest and bagging models when applied to gender-divided datasets. These improvements ranged from 43.5% to 87.4% when compared to the full dataset. Similarly, upon examining data categorized by start station names, we noticed enhanced accuracy in predictions, especially with random forest and GRU models. The significant enhancements in the accuracy of predicting trip duration and distance when datasets are partitioned based on gender or start station indicate that transportation systems can gain advantages from planning that take into account gender and implementing strategies specific to each station. Urban planners and policymakers can utilize these insights to create public transportation systems that are safer and more accessible, accommodating the varied needs of their users. Moreover, the notable differences in the performance of models, such as the random forest and LSTM models demonstrating significant enhancements in mean absolute

error (MAE) and root mean square error (RMSE) scores when applied to divided datasets, highlight the possibility of implementing machine learning models customized for particular demographic or geographic groups. This has the potential to improve the effectiveness and user satisfaction of transportation services. This demonstrates the potential of focused data analysis in improving predictive models. These findings emphasize the significance of taking demographic and geographic factors into account when analyzing data.

By examining the potential implications of our findings on comparable datasets and diverse domains, this research establishes a foundation for developing advanced, contextually relevant analysis methodologies in statistical analysis and machine learning. By means of these efforts, we can enhance our comprehension of the complex patterns of human behavior and enhance the development and provision of services across diverse sectors.

Table 6. MAE scores for the divided dataset by start station name against the complete dataset for predicting trip duration.

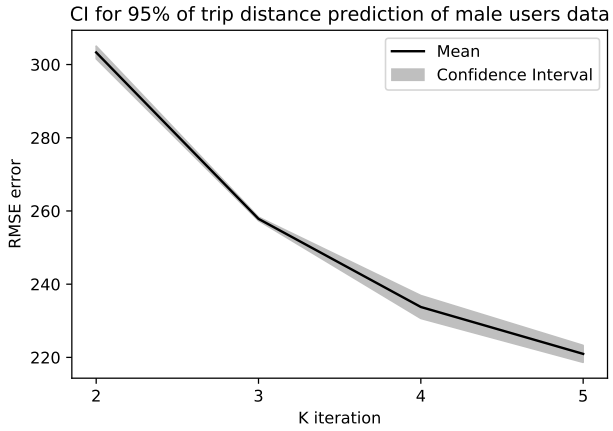
Model	Grove St PATH	Exchange Place	Sip Ave	Hamilton Park	Full data
Random Forest	3.1	3.7	3.2	3.0	8.6
Catboost	6.9	7.6	5.9	6.0	15.4
Bagging	3.2	3.8	3.3	3.2	9.1
GRU	7.0	9.8	6.5	5.2	11.9
LSTM	7.6	9.4	7.4	7.0	14.6

Table 7. RMSE scores for the divided dataset by start station name against the complete dataset for predicting trip duration.

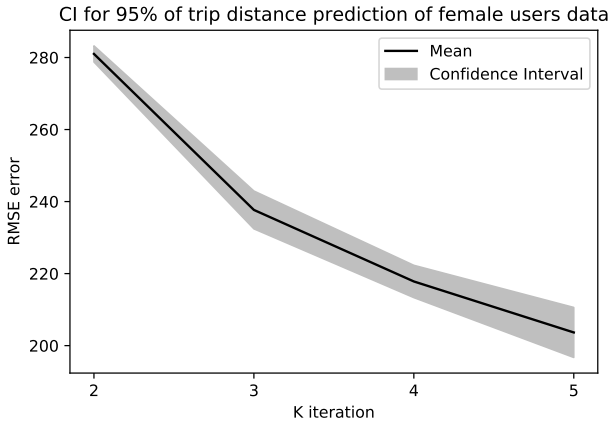
Model	Grove St PATH	Exchange Place	Sip Ave	Hamilton Park	Full data
Random Forest	25.4	36.9	25.8	25.7	598.5
Catboost	45.9	47.6	34.4	44.6	375.2
Bagging	25.8	41.3	28.5	27.5	545.9
GRU	68.5	48.4	61.8	45.7	389.1
LSTM	70.4	49.1	62.5	46.1	389.3

Table 8. MAE scores for the divided dataset by start station name against the complete dataset for predicting trip distance.

Model	Grove St PATH	Exchange Place	Sip Ave	Hamilton Park	Full data
Random Forest	80.4	113.7	92.0	69.2	100.4
Catboost	167.5	231.3	195.5	132.3	248.9
Bagging	85.5	123.4	101.3	74.3	105.3
GRU	230.0	430.3	293.1	250.3	433.2
LSTM	425.6	1028.2	1163.2	853.9	443.6



(a) CI for 95% confidence level for predicting trip distance of male users' bike riding dataset.



(b) CI for 95% confidence level for predicting trip distance of female users' bike riding dataset.

Figure 7. Random forest model's CI for 95% confidence level of bike rides based on the user's gender: (a) male users and (b) female users.

Table 9. RMSE scores for the divided dataset by start station name against the complete dataset for predicting trip distance.

Model	Grove St PATH	Exchange Place	Sip Ave	Hamilton Park	Full data
Random Forest	178.0	252.8	202.3	172.2	1373.7
Catboost	279.7	381.6	314.3	230.4	2282.6
Bagging	190.6	272.7	224.8	189.3	241.9
GRU	400.0	995.1	470.7	409.8	22595.4
LSTM	626.4	1313.8	1364.0	937.3	22608.8

5. Conclusions

The current work proposed an answer regarding the question of whether the predictive model can benefit from training on many BSS sub-datasets instead of training on the complete BSS dataset. The proposed work aimed to improve the BSS trip duration and distance by dividing the BSS dataset based on the user's attributes; the proposed method is called *divide and train*. For instance, the BSS was partitioned into two distinct sub-datasets, one for female users and another for male users. Next, a set of predictive models trained on a complete BSS dataset was compared against the predictive models trained on divided BSS sub-datasets. The proposed method was thoroughly tested on a real dataset for the BSS of New York City. The two metrics utilized to assess accuracy were the RMSE and MAE. The obtained results showed that the proposed method contributes to drastically improving the trip duration and distance. The proposed method succeeded in achieving 84.2% and 87.4% in predicting the trip duration using random forest for male and female partitions. In addition, the obtained results demonstrated that the proposed method significantly reduces both the trip duration and distance. Specifically, it enhances the accuracy of predictions, showcasing a seven-fold improvement in the RMSE metric for

predicting trip duration, and a six-fold improvement for trip distance predictions. Future work includes three directions as follows: 1) testing the proposed method on the free-floating bike-sharing systems, 2) investigating the proposed method in other applications, rather than the BSS, and 3) analyzing the new BSS usage patterns during the COVID-19 pandemic.

Future studies could investigate the suitability of the *divide − and − train* method in contexts where demographic or spatial factors, such as healthcare or retail, have an important effect on results. Furthermore, investigating the fundamental factors contributing to the observed differences based on gender and variations based on start station could provide a more profound understanding of user behavior and preferences.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors extend their appreciation to the Deanship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number (IF2/PSAU/2022/01/20968).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. X. Y. Ni, D. J. Sun, Q. C. Lu, Q. Chen, A proportional allocation model for parking reservation systems considering entrance capacity constraints, *IEEE Intell. Transp. Syst. Mag.*, **16** (2024), 162–173. https://doi.org/10.1109/MITS.2023.3316276
2. G. Xiao, L. Chen, X. Chen, C. Jiang, A. Ni, C. Zhang, et al., A hybrid visualization model for knowledge mapping: Scientometrics, SAOM, and SAO, *IEEE Trans. Intell. Transp. Syst.*, **25** (2024), 2208–2221. https://doi.org/10.1109/TITS.2023.3327266
3. X. Yao, J. Feng, An end to end two-stream framework for station-level bike-sharing flow prediction, *Expert Syst. Appl.*, **247** (2024), 123273. https://doi.org/10.1016/j.eswa.2024.123273
4. Y. Zhou, Q. Li, X. Yue, J. Nie, Q. Guo, A novel predict-then-optimize method for sustainable bike-sharing management: a data-driven study in china, *Ann. Oper. Res.*, **2022** (2022), 1–33. http://doi.org/10.1007/s10479-022-04965-0
5. I. Otero, M. Nieuwenhuijsen, D. Rojas-Rueda, Health impacts of bike sharing systems in europe, *Environ. Int.*, **115** (2018), 387–394. http://doi.org/10.1016/j.envint.2018.04.014
6. V. Albuquerque, M. S. Dias, F. Bacao, Machine learning approaches to bike-sharing systems: A systematic literature review, *ISPRS Int. J. Geo-Inf.*, **10** (2021), 62. http://doi.org/10.3390/ijgi10020062

7. L. Caggiani, R. Camporeale, Z. Hamidi, C. Zhao, Evaluating the efficiency of bike-sharing stations with data envelopment analysis, *Sustainability*, **13** (2021), 881. http://doi.org/10.3390/su13020881
8. M. A. Butt, S. Danjuma, M. S. B. Ilyas, U. M. Butt, M. Shahid, I. Tariq, Demand prediction on bike sharing data using regression analysis approach, *J. Innovative Comput. Emerging Technol.*, **3** (2023). https://doi.org/10.56536/jicet.v3i1.52
9. L. Cheng, J. Yang, X. Chen, M. Cao, H. Zhou, Y. Sun, How could the station-based bike sharing system and the free-floating bike sharing system be coordinated?, *J. Transp. Geogr.*, **89** (2020), 102896. http://doi.org/10.1016/j.jtrangeo.2020.102896
10. *New York City Bike Share Dataset*. Available from: https://www.kaggle.com/akkithetechie/new-york-city-bike-share-dataset.
11. C. Rudloff, B. Lackner, Modeling demand for bikesharing systems: neighboring stations as source for demand and reason for structural breaks, *Transp. Res. Rec.*, **2430** (2014), 1–11. http://doi.org/10.3141/2430-01
12. H. Yang, K. Xie, K. Ozbay, Y. Ma, Z. Wang, Use of deep learning to predict daily usage of bike sharing systems, *Transp. Res. Rec.*, **2672** (2018), 92–102. http://doi.org/10.1177/0361198118801354
13. W. Wang, *Forecasting Bike Rental Demand Using New York Citi Bike Data*, Master's thesis, Technological University Dublin, 2016.
14. B. Wang, I. Kim, Short-term prediction for bike-sharing service using machine learning, *Transp. Res. Procedia*, **34** (2018), 171–178. http://doi.org/10.1016/j.trpro.2018.11.029
15. Y. Li, Y. Zheng, Citywide bike usage prediction in a bike-sharing system, *IEEE Trans. Knowl. Data Eng.*, **32** (2019), 1079–1091. http://doi.org/10.1109/TKDE.2019.2898831
16. C. Wirtgen, M. Kowald, J. Luderschmidt, H. Hünemohr, Multivariate demand forecasting for rental bike systems based on an unobserved component model, *Electronics*, **11** (2022), 4146. http://doi.org/10.3390/electronics11244146
17. H. Lin, Y. He, S. Li, Y. Liu, Insights into travel pattern analysis and demand prediction: A data-driven approach in bike-sharing systems, *J. Transp. Eng. Part A. Syst.*, **150** (2024), 04023132. https://doi.org/10.1061/JTEPBS.TEENG-8137
18. C. M. Vallez, M. Castro, D. Contreras, Challenges and opportunities in dock-based bike-sharing rebalancing: a systematic review, *Sustainability*, **13** (2021), 1829. https://doi.org/10.3390/su13041829
19. X. Ma, S. Zhang, T. Wu, Y. Yang, J. Yu, Can dockless and docked bike-sharing substitute each other? Evidence from Nanjing, China, *Renewable Sustainable Energy Rev.*, **188** (2023), 113780. https://doi.org/10.1016/j.rser.2023.113780
20. Z. Chen, D. van Lierop, D. Ettema, Dockless bike-sharing systems: What are the implications?, *Transport Rev.*, **40** (2020), 333–353. https://doi.org/10.1080/01441647.2019.1710306
21. Y. Wang, Z. Zhan, Y. Mi, A. Sobhani, H. Zhou, Nonlinear effects of factors on dockless bike-sharing usage considering grid-based spatiotemporal heterogeneity, *Transp. Res. Part D Transp. Environ.*, **104** (2022), 103194. https://doi.org/10.1016/j.trd.2022.103194

22. W. Jiang, Bike sharing usage prediction with deep learning: a survey, *Neural Comput. Appl.*, **34** (2022), 15369–15385. https://doi.org/10.1007/s00521-022-07380-5

23. X. Li, Y. Xu, X. Zhang, W. Shi, Y. Yue, Q. Li, Improving short-term bike sharing demand forecast through an irregular convolutional neural network, *Transp. Res. Part C Emerging Technol.*, **147** (2023), 103984. https://doi.org/10.1016/j.trc.2022.103984

24. C. Song, S. Zhou, W. Chang, Y. Xiao, Y. Fu, L. Yang, A short-term demand of bike-sharing forecasting model based on spatio-temporal graph data, in *2023 28th International Conference on Automation and Computing (ICAC)*, IEEE, (2023), 1–5. https://doi.org/10.1109/ICAC57885.2023.10275167

25. S. Zhou, C. Song, T. Wang, X. Pan, W. Chang, L. Yang, A short-term hybrid TCN-GRU prediction model of bike-sharing demand based on travel characteristics mining, *Entropy*, **24** (2022), 1193. https://doi.org/10.3390/e24091193

26. J. Y. Xu, Y. Qian, S. Zhang, C. C. Wu, Demand prediction of shared bicycles based on graph convolutional network-gated recurrent unit-attention mechanism, *Mathematics*, **11** (2023), 4994. https://doi.org/10.3390/math11244994

27. B. Pan, L. Tian, Y. Pei, The novel application of deep reinforcement to solve the rebalancing problem of bicycle sharing systems with spatiotemporal features, *Appl. Sci.*, **13** (2023), 9872. https://doi.org/10.3390/app13179872

28. X. Chang, J. Wu, H. Sun, X. Yan, A smart predict-then-optimize method for dynamic green bike relocation in the free-floating system, *Transp. Res. Part C Emerging Technol.*, **153** (2023), 104220. https://doi.org/10.1016/j.trc.2023.104220

29. X. Li, Y. Xu, Q. Chen, L. Wang, X. Zhang, W. Shi, Short-term forecast of bicycle usage in bike sharing systems: a spatial-temporal memory network, *IEEE Trans. Intell. Transp. Syst.*, **23** (2021), 10923–10934. http://doi.org/10.1109/TITS.2021.3097240

30. X. Ma, Y. Yin, Y. Jin, M. He, M. Zhu, Short-term prediction of bike-sharing demand using multi-source data: a spatial-temporal graph attentional LSTM approach, *Appl. Sci.*, **12** (2022), 1161. http://doi.org/10.3390/app12031161

31. P. Xie, T. Li, J. Liu, S. Du, X. Yang, J. Zhang, Urban flow prediction from spatiotemporal data using machine learning: A survey, *Inf. Fusion*, **59** (2020), 1–12. http://doi.org/10.1016/j.inffus.2020.01.002

32. B. Wang, H. L. Vu, I. Kim, C. Cai, Short-term traffic flow prediction in bike-sharing networks, *J. Intell. Transp. Syst.*, **26** (2022), 461–475 .http://doi.org/10.1080/15472450.2021.1904921.

33. W. Zi, W. Xiong, H. Chen, L. Chen, TAGCN: Station-level demand prediction for bike-sharing system via a temporal attention graph convolution network, *Information Sciences*, **561** (2021), 274–285. http://doi.org/10.1016/j.ins.2021.01.065

34. E. Collini, P. Nesi, G. Pantaleo, Deep learning for short-term prediction of available bikes on bike-sharing stations, *IEEE Access*, **9** (2021), 124337–124347. http://doi.org/10.1109/ACCESS.2021.3110794

35. M. Cipriano, L. Colomba, P. Garza, A data-driven based dynamic rebalancing methodology for bike sharing systems, *Appl. Sci.*, **11** (2021), 6967. http://doi.org/10.3390/app11156967

36. J. Schuijbroek, R. C. Hampshire, W. J. Van Hoeve, Inventory rebalancing and vehicle routing in bike sharing systems, *Eur. J. Oper. Res.*, **257** (2017), 992–1004. http://doi.org/10.1016/j.ejor.2016.08.029
37. A. Maleki, E. Nejati, A. Aghsami, F. Jolai, Developing a data-driven learning-based simulation method as a decision support tool for rebalancing problem in the bike-sharing systems, *Available at SSRN 4329723*. http://doi.org/10.2139/ssrn.4329723
38. M. Du, L. Cheng, X. Li, F. Tang, Static rebalancing optimization with considering the collection of malfunctioning bikes in free-floating bike sharing system, *Transp. Res. Part E Logist. Transp. Rev.*, **141** (2020), 102012. http://doi.org/10.1016/j.tre.2020.102012
39. S. Chang, R. Song, S. He, G. Qiu, Innovative bike-sharing in china: Solving faulty bike-sharing recycling problem, *J. Adv. Transp.*, **2018** (2018). http://doi.org/10.1155/2018/4941029
40. Z. Sun, Y. Li, Y. Zuo, Optimizing the location of virtual stations in free-floating bike-sharing systems with the user demand during morning and evening rush hours, *J. Adv. Transp.*, **2019** (2019). http://doi.org/10.1155/2019/4308509
41. A. Fathalla, A. Salah, M. A. Mohamed, N. I. Lestari, M. Bekhit, A novel dual prediction scheme for data communication reduction in IoT-based monitoring systems, in *International Conference on Internet of Things as a Service*, Springer, **421** (2021), 208–220. https://doi.org/10.1007/978-3-030-95987-6_15
42. A. Pajankar, A. Joshi, Introduction to machine learning with scikit-learn, in *Hands-on Machine Learning with Python: Implement Neural Network Solutions with Scikit-Learn and PyTorch*, Springer, (2022), 65–77. https://doi.org/10.1007/978-1-4842-7921-2_5
43. A. V. Dorogush, V. Ershov, A. Gulin, Catboost: gradient boosting with categorical features support, preprint, arXiv:1810.11363. http://doi.org/10.48550/arXiv.1810.11363
44. N. Bantilan, pandera: Statistical data validation of pandas dataframes, in *Proceedings of the Python in Science Conference (SciPy)*, (2020), 116–124.
45. J. Unpingco, Numpy, in *Python Programming for Data Analysis*, Springer, (2021), 103–126. https://doi.org/10.1007/978-3-030-68952-0_4
46. S. Cao, Y. Zeng, S. Yang, S. Cao, Research on python data visualization technology, in *J. Phys.: Conf. Ser.*, IOP Publishing, **1757** (2021), 012122. https://doi.org/10.1088/1742-6596/1757/1/012122
47. A. Sanmiguel-Rodríguez, Bike-sharing systems: Effects on physical activity in a spanish municipality, *Phys. Act. Rev.*, **10** (2022), 66–76. http://doi.org/10.16926/par.2022.10.22
48. Y. Chen, Y. Zhang, D. Coffman, Z. Mi, An environmental benefit analysis of bike sharing in New York city, *Cities*, **121** (2022), 103475. http://doi.org/10.1016/j.cities.2021.103475

AIMS Press