

Application of Machine Learning in Predicting the Number of Bike Share Riders

Bongs Lainjo

Cybermatic Internation, Canada
bsuiru@icloud.com

Abstract

This study aims to investigate the factors influencing the demand for bike sharing to identify the variables that significantly predict the need for shared bicycles. The study aims to create more in-depth knowledge about bike-sharing models to enhance the ideas on designing, developing, implementing and utilization of bike-sharing models. A multiple regression model was used to model the demand for 703 Capital Bikeshare's shared bikes in the USA. The variations in the need for the bikes were assessed based on certain variables, such as the total number of registered and unregistered bikers and renters. Linear regression analyses were conducted to determine the factors that statistically and significantly predict the number of bikes rented. Machine learning classifiers: Random Forest, Decision Trees, Nearest Neighbor and XGBoost were used to determine the most important predictors of bike demand and the data analyzed in SPSS V25, R and PYTHON. Increase in demand for shared bikes was attributed to factors such as increase in temperature ($p = 0.000$), days of the week except the first day of the week ($p < 0.05$), the month of September ($p = 0.036$), spring season ($p = 0.000$), fall seasons ($p = 0.001$) and humidity ($p = 0.000$). A significant decrease in the demand for shared bikes was observed on the first day of the week ($p = 0.218$) and days with strong winds ($p = 0.113$). More people are likely to rent shared bikes on hot days, in September, during spring and fall seasons, on humid days, and all days of the week except on the first day of the week.

Keywords: Bike-Sharing Programs, User-Type, Monitoring Systems, Analytics, Predictive Models, Historical Data

JEL Codes:

How to Cite:

Lainjo, Bongs. (2022). Application of Machine Learning in Predicting the Number of Bike Share Riders. *International Journal Business, Management, and Economics*, 3(4). 339 - 370. DOI: <https://doi.org/10.47747/ijbme.v3i4.865>

1. Introduction

The objective of this paper is to highlight specific arguments along with ideas about designing, developing, and implementing bike-sharing models. The article further emphasizes understanding the aspects studied in the field, the weaknesses, and gaps or areas that need further study. Thus, the review focuses on demonstrating to the reader the reasons for conducting the research, including usefulness, necessity, importance, and validity. Bicycles are offered for rent across a city through public bicycle share programs (PBSP). In 2014, around 855 public bicycle-sharing programs were active in cities worldwide (Oppermann et al., 2018).

PBSPs are used for several purposes, including increasing cycling levels, facilitating the first and final kilometres of public transit trips, and reducing traffic congestion, to name a few (Ricci, 2015). These programs are used for both transit and leisure journeys, although they are primarily designed to make short trips (under 30 minutes) easier. Users of PBSPs are frequently charged an extra price for every journey that exceeds the prescribed time limit (usually 30 minutes). Increased bicycle availability for individuals who do not possess one, increased convenience of bicycling, and normalization of biking as a mode of transportation are all potential avenues for these initiatives to lead to increased population-level riding (Goodman et al., 2014). Bike-sharing systems (BSS) improve multimodality in transportation, urban accessibility, and mobility sustainability, more cities worldwide are implementing them to address increased air pollution, urban mobility, and changes in urban mobility patterns and behaviour, all of which have been aggravated by the recent pandemic crisis (Albuquerque et al., 2021). Shared bikes are gaining traction as an alternate mode of transportation that can address environmental issues, traffic congestion, and poor quality of life caused by car-oriented transportation systems while ensuring practical consequences (Gu et al., 2019).

Furthermore, shared bikes offer various advantages when constructing an integrated public transportation system with public buses and urban rail systems. To begin, a bike trip's typical mileage is roughly 3 to 5 km. However, it is feasible to travel more considerable distances using public transit. Second, the public transit catchment area may be expanded from an existing pedestrian area to a place that can be covered by bicycles. Third, highly requested shared bikes can accommodate leisure travel on existing transit at non-adjacent hours (save during rush hour, when traffic demand is concentrated during the day), potentially increasing the total utilization rate of public transportation. Bikes and walking or public transit are the most excellent way to achieve an integrated public transportation system (Böcker et al., 2020). Combining public transportation services, such as buses or metros, into high-capacity corridors where economies of scale can be built and strengthening cooperation with passages or feeders using bikes, an integrated public transportation system can increase the availability of public transportation (Gu et al., 2019). As a result, forecasting demand for shared bikes is critical for successfully implementing an integrated public transportation system that incorporates shared bikes.

In the data-driven business climate of the mobility sector, precisely predicting client demand is a critical component of success (Sohrabi et al., 2020). It is straightforward to identify global organizations that supply their varied services to clients based on demand forecast findings in the

data technology era. Companies have had short-term success because they have accurately forecasted demand based on internal and external factors. Most sectors throughout the globe are undergoing digital transformations based on machine learning and deep learning algorithms (Wang et al., 2019). Fast-growing start-ups, in particular, frequently rely on crucial business choices on data and algorithms rather than managerial expertise or intuition. The bike is an environmentally beneficial method of transportation that benefits individuals optimizing its exercise effects during the COVID-19 pandemic in a modern world where environmental concerns are becoming increasingly important. Since it was changed into a "shared" item a few years ago, the bike has received international notice. It is not only for recreational reasons but also for transportation (Gu et al., 2019).

When creating sustainable transportation systems, shared bike programs are seen as a way to respond to climate change and energy challenges in many places worldwide. A shared bike system is an essential and required tool for encouraging people to use bikes and advancing the implementation of an efficient urban transportation system. Despite its numerous benefits, many communities and organizations are hesitant to implement a shared bike program because of the drawbacks of high fixed expenses such as installation and operation. As a result, reliable demand forecasting is essential to keep a bike-sharing program running. The bike-sharing system was one of the first shared economy models in the transportation business. Capital Bikeshare, which started renting bikes in the Washington, D.C. region, held a data science competition on the Kaggle Competition platform to estimate consumer bike share demand (Fanaee-T & Gama, 2014). As a result, data scientists and analysts worldwide have attempted to forecast demand using various data mining approaches.

Much prior research has emphasized the relevance of model parameters, such as the distance between the rental and return locations (Younes et al., 2020; Ma et al., 2020). These works primarily focused on engineering feature approaches and statistical modelling within provided datasets. As a result, past research had model constraints since they only employed specific datasets. On the other hand, this research focuses on investigating a new characteristic from a live data source. Existing research has used statistical approaches to estimate bike demand, but few have used machine learning methodologies, which have recently come to the forefront. This research aims to offer a machine learning prediction model incorporating factors influencing shared bike demand. Furthermore, a predictive model that contains these factors is provided from the perspective of an integrated public transportation system.

The research paper focuses on discussing, understanding, and identifying the association between bike share (BS) riders and compelling parameters using a live data set. Through this study, more in-depth knowledge can be achieved on bike-sharing models, thereby enhancing future applications' design, implementation and utilization. Over the years, the concept of BS programs has increased globally. Thus, conducting this research paper has helped in understanding the arguments and theories which have been provided regarding the relationship between BS riders and compelling parameters. The research has also assisted in enhancing the developments made over the years on the selected topic and gaining insights regarding the gaps. Thus, understanding all these aspects will contribute to making a significant input in BS programming.

2. Literature Review

The literature review will cover the bike share programs worldwide, bike share models, methodologies used in previous literature and even gaps. The research strategy that yielded the information for this study was discussed in the first section of this review. A search of University student databases caused a flurry of results, including multiple articles about the Bike Share program. Following initial investigations through Academic Search Complete at the library, I discovered further particular subtopics, and the addition of healthcare-related databases revealed a wide range of relevant research. To find peer-reviewed online resources, the researcher will use government reports, Google Scholar, a literature matrix (created by the researcher to allow quick comparisons among publications to determine scope), and EBSCOhost. The search began with phrases and keywords such as Bike share, regression models, Nearest Neighbour, XGboost, Random Forest and Decision Trees. Using these keywords and *phrases* ensures a complete investigation of all aspects of Bike Share.

2.1 Bike Share Programs

According to Fishman (2015), BS has gained high and rapid popularity in the last decade. The concept of BS was introduced in the 1960s. However, the number of cities, which have been offering the service, has increased since the late 1990s. In this context, “Contemporary bike-share programs (BSPs) refer to the provision of bikes, which can be picked up and dropped off at self-serving docking stations. Typically, trips are of short duration (less than 30 min). The bicycles usually contain technologies allowing program operators to monitor activities in their respective docking station locations. Some are equipped with a global positioning system (GPS)” (Fishman, 2015, 2). Based on the study findings of Fishman, Washington, Haworth, and Watson (2015), the BS programs' key benefits include flexible mobility and physical activity benefits, emission reductions, reduced fuel use & congestion, financial savings, and multimodal transport connection support.

Additionally, one of the main benefits of BS is that bikes are perceived as a replacement car. On the other hand, the Institute for Transportation and Development (2018) found that in 2016, BS systems rider trips in the US accounted for over 28 million. Ridership, specifically in North America, has increased significantly since 2012 due to new BS systems every year. Since 2016, most new bike share systems operating in North America have used intelligent bikes. Furthermore, as BS continues to evolve, new operating systems across the US have emerged using dockless and stationless bikes. Additionally, in 2017, many dockless operators installed bikes in the US, Great Britain, China, Singapore, Australia, and Italy, among others. Correspondingly, ‘pedal assist e-bike share fleets’ have been launched in many North American cities since 2017.

Moreover, Campbell, Cherry, Ryerson, and Yang (2016) stated that in the present scenario, China is the leading nation in the world in terms of the growth of public BS along with the private electric bike. The current projections confirm the feasibility of implementing large-scale shared e-bike systems nationwide. Public BS systems can be one of the fastest-growing public transportation modes in the world. The industry has been growing at an average rate of 37% since 2009. In this context, “The current trajectory of bike share adoption, the popularity of e-

bikes, and the presence of e-bike-share pilot projects in other countries all support a future of e-bike sharing in China” (Campbell et al., 2016, 400). Given the rapid evolution of transportation in China, it is not well understood how such a system will differ from standard bike-share and how both types of shared bikes (hereafter “shared bike” is used to refer to both bike-share and e-bikeshare) systems can best address the needs of urban China” (Campbell et al., 2016, 400). Based on the study findings of Guo, Zhou, and Li (2017), bike-sharing growth is receiving attention as societies are becoming highly aware of the significance of ‘active non-motorized traffic modes. Since BS is perceived as a good transport system, it increases the use of a bicycle, especially in the circumstances providing different pick-up, drop-off locations, self-service, etc., thereby making it convenient for users.

Furthermore, bike-sharing offers an efficient solution to the transport system and thus can be perceived as an alternative to other transit systems. Additionally, Kim, Ghimire, Pant, and Yamashita (2021) asserted that it is essential for riders to use a helmet while using a bike. A survey conducted in the year 2013 under ‘New York City’s bike share program’ found that about 85.3% of the riders did not wear helmets. This raises an important question – whose liability is it? Even though there are no statistics available on BS riders’ accidents, a well-designed and comprehensive program should include access to helmets. The additional challenge is motivating riders to use these helmets. If current rider safety behaviour and awareness are any indications, there is a strong likelihood that most BS riders will use them. This is a challenge that policymakers need to consider and prioritize in their respective BS programs.

2.2 Bike Sharing Models (NTBR and EXPANDED)

Different bike-sharing models can be used to predict flows in every station. Contextually, Tran, Ovtrachta, and d'Arcier (2015), robust linear regression models are one such model that helps in predicting flows. The developed environment variables used in the model are often identified within a buffer zone (300 meters) in every bike-sharing station. Thus, linear regression can be used during the busiest time of a weekday to predict the bike-sharing flow. Integrating a robust linear regression method can help improve rider needs and program optimization in general.

Furthermore, the arrival and exit flow at the hourly level can be integrated into the regression model at each station. On the other hand, Yang, Li, and Zhou (2019) highlighted another bike-sharing model system dynamics simulation. In this context, the simulation method helps in modeling factors along with operations, processes, and policies to be considered in the dockless bike-sharing programs operations. It further helps assess sustainable strategies that enhance the overall system performance. Maintaining an adequate balance between expenditure and revenue is essential, especially in the sustainable development of dockless bike-sharing programs in a specific area. Thus, both the income and expense of the dockless bike-sharing programs need to be fully considered in the respective system. Additionally, the system model performs different simulations and further evaluates the dynamic behavior of the individual system.

The study findings of Ottomanellia (2013, 204) further stated that the main objective of a dynamic simulation model is that it helps "to minimize the vehicles repositioning costs for bike-sharing operators, aiming at a high-level users satisfaction and if it increases with the probability to find an available bike or a free docking point in any station at any time. The proposed model

considers the dynamic variation of the demand". Based on the study findings of Soriguera, Casadoa, and Jiménez (2018), docking stations along with bikes are passive agents, while the amount and location are perceived as inputs to the simulation. Thus, it indicates that the simulation depends on the higher-level model for establishing the optimality in the respective information. Furthermore, "users and repositioning trucks are the active agents who take decisions, which result in an efficient flow of bikes between stations" (Soriguera, Casadoa, & Jiménez, 2018, 140). For instance, the case study of Barcelona's Bicing can be considered, wherein a 24-hour simulation was performed with the inputs from a real case study. Bicing is a bike-sharing system in Barcelona, which was selected as a benchmark.

Additionally, "the open data policy of the Barcelona council, including Bicing's data, was decisive in such selection. The Bicing open data portal includes the real-time occupancy of every station with a one-minute update frequency. These data allow assessing some aspects of the simulator's performance" (Soriguera, Casadoa, & Jiménez, 2018, 142).

3. Methodologies Used in Previous Literature (NTBR and EXPANDED)

In the study by Tran, Ovtracht, and d'Arcier (2015), the data were obtained from the bike-sharing trips from JC Decaux, the administrator of the Lyon bike-sharing system. This included data for every station during the year 2011. In this context, each trip helped provide information regarding the departure, arrival station, date, time of check-in & check out, and the subscriber type. Correspondingly, regarding the subscribers, Tran, Ovtracht, and d'Arcier (2015) focused on analyzing two bikes sharing users, comprising short-term subscribers having a 'one-day bike sharing subscription' and long-term subscribers with a yearly 'bike sharing subscription.' On the other hand, the research conducted by Kim, Ghimire, Pant, and Yamashita (2021) used the data gathered in June 2019. The authors selected a total of 25 sites located in urban Honolulu. In this study, stratified simple random sampling was integrated for selecting street segments with high, medium, and low traffic volumes that differed in lane configurations. This included multi-use paths, protected bikes, and shared and dedicated bike lanes. Visitors, shopping centers, attractions, activity generators, and commercial offices were also mapped and integrated into the sampling process.

During the fourth industrial revolution, multinational corporations and start-ups experimented with the sharing economy idea, attempting to better fulfill consumer demand by incorporating demand forecast findings into their operations. Companies must enhance their prediction model to estimate client demand better to survive in today's harsh competition. Kim et al. (2021) investigated a novel bike-sharing demand prediction model feature, which enhanced the RMSLE score. The RMSLE score results increased by adding this new feature to the number of daily car accidents recorded in the Washington, DC region to the random forest, XGBoost, and LightGBM models.

Machine learning techniques were used by Ashqar et al. (2017) to estimate the availability of bikes at San Francisco Bay Area Bike Share stations. As univariate regression methods, Random Forest (RF) and Least-Squares Boosting (LSBoost) were utilized, while as a multivariate regression technique, Partial Least-Squares Regression (PLSR) was used. The number of available bikes at each station was modeled using univariate models. PLSR decreased the

number of necessary prediction models and represented the network stations' geographical correlation. The results suggest that univariate models predict errors more accurately than multivariate models. The results of the multivariate model, on the other hand, are plausible for networks with a large number of spatially associated stations. According to the findings, station neighbors and the forecast horizon time are also essential factors. Fifteen minutes was the most efficient forecast horizon period for minimizing prediction error.

Citizens have benefited from the bike-sharing program, which has functioned as a valuable addition to public transportation. Each docking station has a defined space to keep bikes for docked bike-sharing service, and the station may be empty or saturated at various times. Bike-sharing companies often move bikes between stations by driving trucks based on their previous experiences, which might result in wasting human resources. Accessing this service is inefficient for operators and cumbersome for users. As a result, operators and riders benefit from accurately forecasting the quantity of available bike share in the stations. B. Wang and Kim (2018) focused primarily on short-term docking station utilization predictions in Suzhou, China. With one-month historical data, two new and highly efficient models, LSTM and GRU, are used to forecast the short-term available number of bikes in docking stations. As a comparison, Random Forest is utilized as a baseline. According to the results, RNNs (LSTM and GRU) and Random Forest may achieve good performance with tolerable error and comparative accuracies. In terms of training time, random forest is preferable, while LSTM with complicated structures can predict better in the long run. The highest discrepancy between actual data and anticipated value is just one or two motorcycles, indicating that the created models are suitable for implementation.

Because the number of bicycles is essential to the long-term success of dockless PBS, this study used OFO bike operation data in Shenzhen to test the implementation of a machine-learning method for quantity management. Zhou et al. (2020) employed two clustering methods to identify the bicycle gathering area. The available bike number and coefficient of available bike number variation were evaluated in each type of cycling gathering area. Second, using 25 impact variables, five classification algorithms were assessed on their accuracy in classifying the kind of bicycle collecting locations. Finally, the use of information gleaned from current dockless bike operating data to influence the numbering and administration of public bicycles was investigated. According to the findings, 492 OFO bicycle collecting places were classified as highly inefficient, normal inefficient, highly efficient, and average efficient. Around 110,000 bikes with minimal utilization were gathered in the highly weak and standard inadequate zones. The accuracy of the categorization algorithm will be impacted when more types of bicycle collecting areas are added. In five classification methods, the random forest classification had the most outstanding performance in detecting bicycle gathering area kinds, with an accuracy of more than 75%. In four different types of bicycle gathering spaces, there were notable variances in the features of 25 impact elements. It is possible to estimate area types using these criteria to maximize the number of bicycles available, save operating costs, and enhance usage efficiency. Using a machine learning technique, this research aids operators and the government in understanding the features of dockless PBS and contributes to the system's long-term sustainability.

In recent years, bike-sharing systems have seen remarkable expansion and scholarly interest. The key drivers to this growth were environmental awareness, technological advancements, and the desire for socially appropriate transportation choices. However, as these systems expand, businesses must constantly rebalance them to satisfy rising demand. As a result, running organizations continuously look for the best methods for predicting flow. Hamad (2020) investigated three machine learning techniques, focusing on the overlooked topic of multiple seasonality in time-series models. The study's goal was to look at the link between bike sharing, the weather, and the people who utilize it. The four strategies are then constructed and assessed to select the best-performing algorithm and recommend other research topics in traditional time series models.

While the advantages of shared bicycle use in terms of greater mobility, accessibility, and urban environmental quality are well established, the effects of increased bicycling on traffic safety need to be evaluated and managed further. Helmet use and behaviors of bike-share users and other cyclists are contrasted based on observational studies in one of the nation's most extensive and successful bike-share programs (Honolulu, Hawaii). In 25 different places throughout the city, 5431 bicycles, mopeds, motorbikes, and other two-wheeled vehicles were spotted. To examine the links between helmet wear, bicyclist characteristics, roadway, traffic violations, location, and environmental factors, K. Kim et al. (2021) employed two logistic regression models. Bikeshare users, visitors, ladies, and those carrying earbuds are less likely than other categories to wear helmets. Bicyclists who ride during rush hour and on weekdays and those who depend on conventional road lanes are likelier to wear helmets. Bikeshare riders are also more prone to break the law than other bicyclists. In addition to raising awareness of the traffic safety issues connected with the growing popularity of biking and bike share, the report includes recommendations for enforcement, education, engineering, and risk management. There is a pressing need to boost helmet wear and overall cyclist safety.

Citizens have benefited from the bike-sharing service, which has functioned as a valuable supplement to public transportation. Each docking station has a defined location to keep bikes for docked bike-sharing service, and the station may be empty or saturated at different times. Bike-sharing companies typically move bikes between stations by driving trucks based on previous experiences, which could waste human resources. Accessing this service is inefficient for operators and inconvenient for users. As a result, both operators and riders benefit from accurately forecasting the number of available bikes in the stations. Wang & Kim (2018) primarily focus on short-term docking station usage predictions in Suzhou, China. With one-month historical data, two new and highly efficient models, LSTM and GRU, are used to forecast the short-term available number of bikes in docking stations. As a comparison, Random Forest is utilized as a baseline. According to the results, RNNs (LSTM and GRU) and Random Forest may achieve good performance with tolerable error and comparative accuracies. In terms of training time, random forest is preferable, while LSTM with complicated structures can predict better in the long run. The highest difference between actual data and anticipated value is just one or two motorcycles, indicating that the created models are ready for implementation.

3.1 Gaps (NTBR)

There is one significant gap in the research, as adequate information could not be found regarding the streamlining association between bike share riders and compelling parameters. To enhance the overall understanding and knowledge of these models, improved methodologies can be integrated, such as conducting interviews with bike-sharing operators. This is because it can help obtain insights from the participant, thereby enabling the understanding of the implicit relationship among different factors. Such a strategy, in addition to a more preliminary and comprehensively exhaustive analysis – addressing covariate independence, homoscedasticity, outliers, collinearity, and others – is this research's central theme and focus. The importance of model factors such as the distance between the rental and return sites have been highlighted in several previous studies (Younes et al., 2020; Ma et al., 2020). These papers mainly focused on the datasets' engineering feature methods and statistical modeling. As a result, previous research had model restrictions since it only used specific datasets.

On the other hand, this study looks at a novel characteristic derived from a live data source. Existing studies have utilized statistical methods to estimate bike demand, but just a few have incorporated machine learning procedures, which have recently gained popularity. This study aims to develop a machine learning prediction model that considers elements that impact demand for shared bikes.

3.2 Contribution to the Research Field NTBR)

This study's importance or contribution will help better understand the concepts associated with bike-sharing in the present scenario globally. This study will further enhance the knowledge regarding the potential challenges and can act as a source of information in the future. Using a machine learning technique, our work assists operators and the government in better understanding the characteristics of bike sharing and contributes to the system's long-term sustainability.

3.3 The Study Settings

3.3.1 Capital share

This study will be based on a US bike-sharing provider, Capital Bikeshare Company (CBC). A mountain bike guided tour and rental facilities are part of the business. The program is jointly owned and sponsored by the District of Columbia and Arlington County, VA, and operated by Alta Bicycle Share, Inc. Its coverage includes both regions (see figure 1)

CBC provides rides for people of all skill levels. They will range from easy family rides to intense, fast-paced expert rides. These tours might take anywhere from a few hours to a week to complete. The company offers a package deal that includes bikes, transportation to and from the trailhead, lunches, and a personal tour guide to show them around the trails and provide information about the area. It also offers a complete bike rental fleet. The bikes range in price from low-cost city cruisers to full-fledged mountain bikes with full suspension. Customers are not required to be on tour to hire bicycles from the company. The company has 4500 bicycles and 500 stations.

Due to the ongoing Corona pandemic, the company's revenues have recently dropped significantly. In the current market environment, the company is struggling to stay afloat. As a result, it has decided to develop a thoughtful business plan to increase revenue as soon as the current lockdown ends and the economy returns to a healthy position. CBC hopes to understand better people's demand for shared bikes after the present Covid-19-related complications end across the country. They planned this to position themselves to meet people's needs whenever the situation improves, differentiate themselves from other service providers, and profit handsomely. They want to know what factors influence the demand for these shared bikes in the United States.

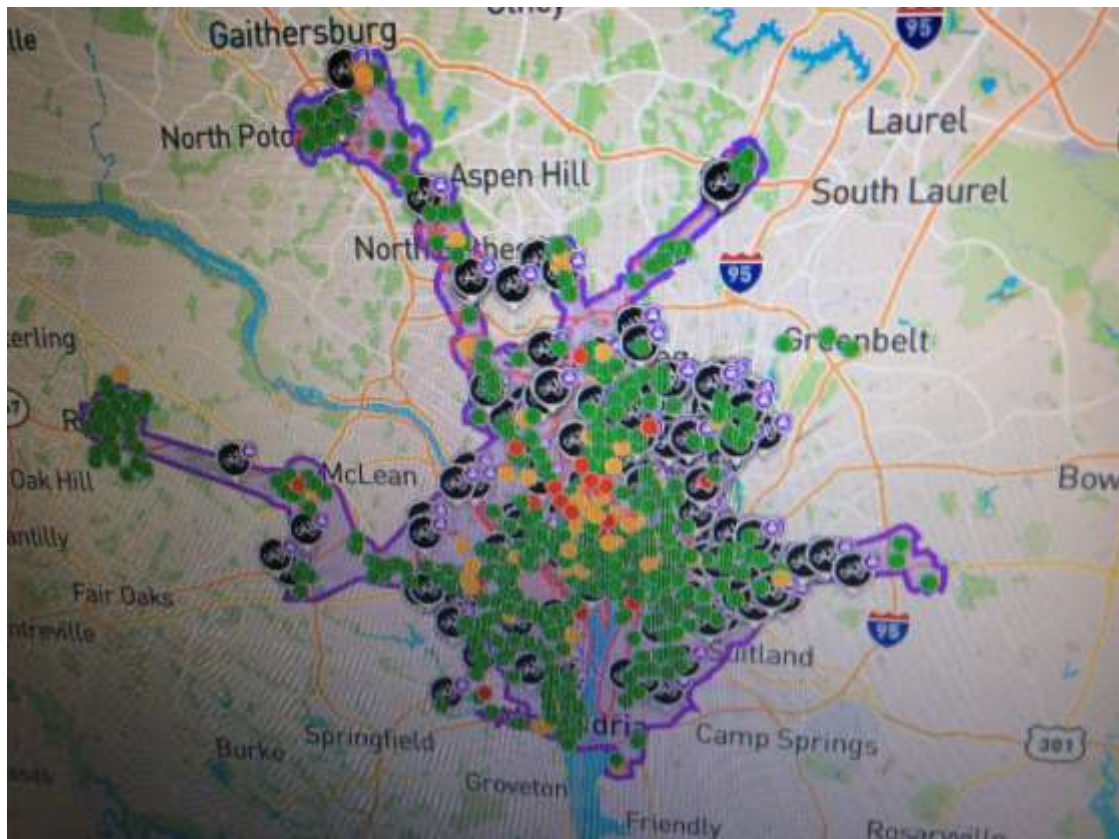


Figure 1: Capital Bikeshare (CB) Coverage: (Source: CB website)

3.3.2 Global distribution of bike sharing programs

The regional links are as follows:

[USA LINK](#) , [EUROPE LINK](#) , [ASIA LINK](#) , [CANADA LINK](#) , [S. AMERICA LINK](#), [AFRICA LINK](#)

3.4 Research limitations

In general, many scientific research studies are prone to different limitations. This one is no exception. The inexhaustive list of this study's shortcomings includes:

1. The study design is influenced by existing data sets.
2. Lack of a sampling protocol.
3. The data set is enhanced by a third party.
4. Inability to verify inconsistent data points – wrong data entry, recording errors, etc. – following on up outlier data.
5. Inadequate data cleaning
6. A limited number of cases, especially when divided into TRAINING (80 %) and TEST (20 %) Datasets
7. Model Replication only valid within the applied parameters

3.5 Data set

A bike-sharing system refers to a service that makes bikes accessible for shared use to individuals for a fee or free on a short-term basis. Such plans let users rent a bike from a "dock," which is frequently computer-controlled and where the user enters payment information, and the system unlocks the bike. After that, the bike can be returned to another dock in the same system. The original database has N=731 observations. These have been reduced to 703 cases after adjusting for deleted issues identified as outliers. Table 1 shows the data dictionary that details the data set attributes used in the study. The data used are free and publicly provided by BSC and Hadi Fanaee-T Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto: original data provider and data compiler, respectively. As indicated earlier, one database limitation has been the inability to include a comprehensive data-cleaning strategy during this process. Data collection for the study was conducted between 2011 and 2012 inclusive. In this dataset, there were no missing values.

Table 1. Study Database Dictionary (Source: Author) The dataset was split into 'Train' and 'Test' datasets in the ratio of 80% to 20%.

4. Result and Discussion

The researcher is motivated by my interest in learning more about the factors influencing demand for these shared bikes. I am interested in knowing what factors influence demand for these shared bikes. I want to identify the variables that significantly predict the need for shared bicycles. A multiple regression model is required to model the demand for shared bikes with the supplied independent variables. I will use it to determine how needs vary depending on the attributes. According to Uyanık & Güler (2013), when we wish to forecast the value of a variable based on the importance of two or more other variables, we utilize multiple regression. The dependent variable, the total number of registered & unregistered bikers: and renters, is the variable we want to forecast. In this study, a linear regression model was used to determine the factors that statistically significantly predict the number of bikes rented. Machine learning classifiers, namely; Random Forest, Decision Trees, Nearest Neighbor, and XGBoost, were

employed to determine the most critical predictors; the most accurate classifier was considered the best. The data was analyzed using SPSS V25, R, and PYTHON.

4.1 Preliminary Analysis

4.1.1 Visual representations (Include HISTOGRAMS and BOXPLOTS with NARRATIVES here)

To begin, we looked at the distribution of the response variable Total Bike Rentals (cnt). According to the histogram in figure 1, the total number of rental bikes appears to follow a relatively normal distribution. The distribution's mean and variance are the same, and when the mean grows more extensive, the distribution approaches a normal distribution.

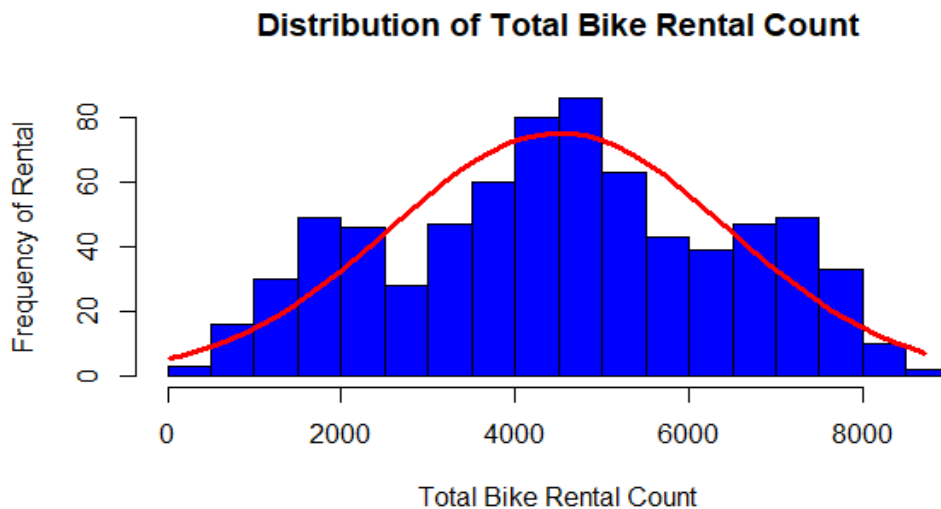


Figure 2: Normality test for bike rental count

To begin, we looked at the distribution of the response variable Total Bike Rentals (count). According to the histogram in figure 1, the total number of rental bikes appears to follow a relatively normal distribution. The distribution's mean and variance are the same, and when the mean grows more extensive, the distribution approaches a normal distribution.

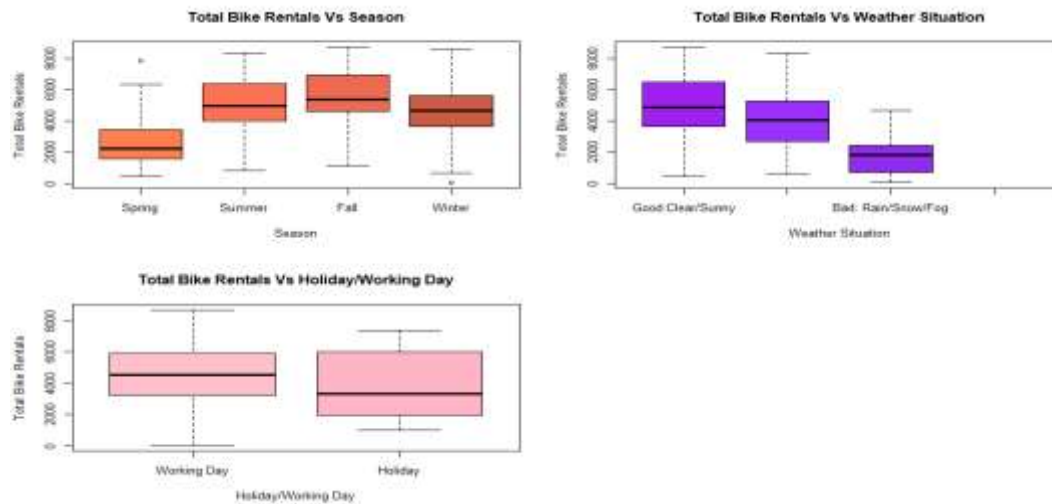


Figure 3: Bike rentals based on seasons

The graph depicts the association between the variable Total Bike Rentals(cnt) and the season. During the summer and fall, the average number of bike rentals is at its peak. The graph depicts the link between the variable Total Bike Rentals(cnt) and the holiday. We can see that the average number of bike rentals is more significant on weekdays than on weekends. The graph depicts the association between the variable Total Bike Rentals(cnt) and the weather. When the weather is terrible, bike rentals have a distinct downward tendency. The graph depicts the association between the variable Total Bike Rentals(cnt) and the year. We can see that the overall trend has risen during the two years. Moreover, there are many bike rentals each year during the summer and fall seasons.

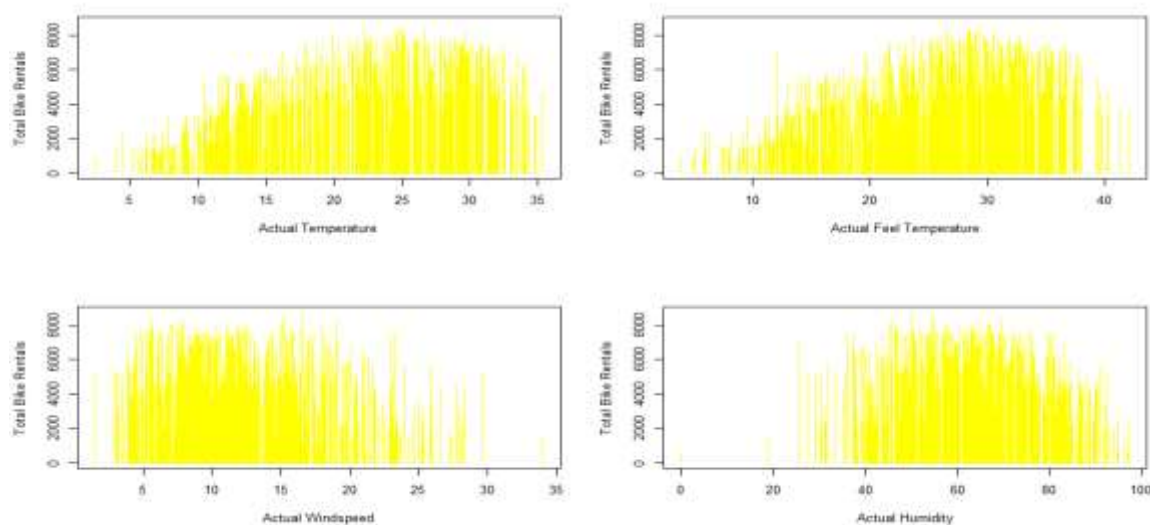


Figure 4: Descriptive statistics on bike rentals

These numerical variables appear to be distributed quite organically.

4.1.2 Descriptive Statistics

Table 1 shows the descriptive statistics: Mean, Median, Minimum, Maximum and Percentiles.

4.2 Verifying Regression Assumptions (NTBR accordingly)

When a researcher decides to use multiple regression to analyze data, one step in the process is to ensure that the data the researcher intends to analyze are compatible with various regression assumptions (Ernst & Albers, 2017). This is necessary because multiple regression can only be used if the data "passes" – the study's outcome - the assumptions required to get a valid result. The first assumption is that the dependent variable has to have a continuous or quantitative level of measurement. In this study, the dependent variable, the Total number of registered & unregistered bikers: renters (cnt), meets this criterion. The second assumption is that there should be more than one independent variable, which can be either continuous or categorical; the independent variables in this study meet the criterion. The third assumption is that a linear relationship between the dependent variable and every independent variable should be assessed using a scatterplot matrix. Dummy variables were generated from the categorical variables.

Figure 5 shows that the dependent variable only has a linear relationship with variables "temp and atemp"—figure 3 shows the relationship between the Response variable and the Categorical variables.

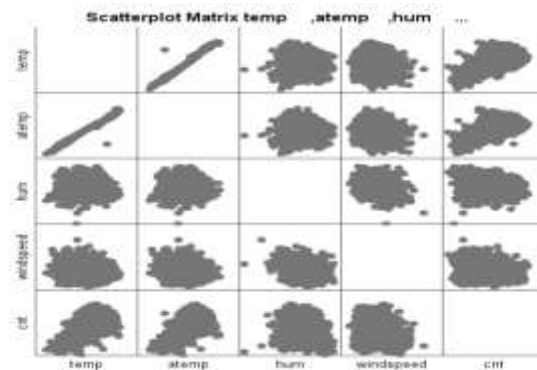


Figure 5: Scatter Plot with a linear correlation between (cnt) and two Explanatory (tempt & attempt) Variables

The fourth assumption is that there should be independent observations; this is checked using the Durbin-Watson statistic. The model shows the value of the Durbin-Watson statistic as 0.452. This implies a positive autocorrelation; this implies that the correlation between the data points is not modeled well enough. The fifth assumption is that the data should indicate homoscedasticity, which means that the variances remain similar as you move along the line of best fit. The scatterplots suggest some homoscedasticity; the points should be about the same Distance from the line. The sixth assumption is that there should not be multicollinearity; this occurs when there is a high correlation among independent variables - problem; this is assessed using the Variance

Inflation Factor (VIF) values. The acceptable values for VIF are those less than 10 (Hair et al., 2019). In this case, the independent variables with the highest values of VIF are; 'temp,' 'atemp,' and 'Winter.' The variables 'atemp' and 'Winter' will be excluded based on the high VIF values. The variable 'temp' will not be excluded despite having a high VIF value since, based on general knowledge, the temperature is likely to be an essential factor for bike rentals.

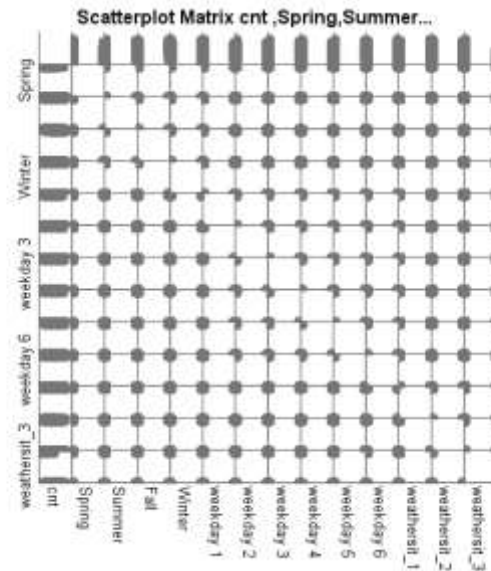


Figure 6: Linear relationship between the Response variable and the Categorical variables.

The seventh assumption is that there should not be outliers or points that are highly influential. Cook's Distance (see Table 1) was used to point out the influential data points; they were not included in the regression model because they have the propensity to reduce the predictive accuracy of the results and the statistical significance. Furthermore, David G. Keinbaum et al. also "measures the extent to which the estimates of the regression coefficients change when an observation is deleted from the analysis." The original data points were N=731; after removing the outliers, the number of observations included in the study reduced to N=703

Table 1: Outliers (Source: Author) OK

instant	Cook's distance
50	0.006915745
69	0.038708211
87	0.008525478
88	0.007635568
89	0.006170474

151	0.008282762
160	0.008569033
203	0.00941673
204	0.009301581
239	0.010159981
266	0.023678745
270	0.006406563
271	0.008344468
328	0.006543001
407	0.012217253
442	0.014635957
443	0.007062332
448	0.011428928
456	0.006164395
519	0.005926599
522	0.006122193
543	0.006580904
568	0.014896619
569	0.007468615
595	0.307200599
638	0.006708605
668	0.006884812
669	0.005721031

The eighth assumption is that the residuals should be approximately normally distributed (see figure 4).



Figure 7: Normal P-P plots that show that the residuals are approximately normally distributed.

Table 2 summarizes the statistics of all the variables used in the study. Table 3 shows the coefficients for the final regression model. The R Square shows how much of the total variation

in the dependent variable, the total number of registered & unregistered bikers: and renters (cnt), can be explained by the independent variables. In this case, 63.3 % can be explained. The F-statistic, $F(25,677) = 46.67$, $p < 0.05$, suggests that the regression model is a good fit for the model, that is, the regression model statistically significantly predicts the total number of registered & unregistered bikers: renters (cnt). Table 3 shows that the variables that significantly predict the total number of registered & unregistered bikers: renters (cnt) include Temperature, humidity, wind speed, the month of September, Spring, Fall, all weekdays except weekday one and weather situation 3 (Light Snow, Light Rain with Scattered Clouds), $p < 0.05$. Table 2 shows the descriptive statistics for both the dependent and the dependent variables.

Table 2: Descriptive statistics (Source: Author)

	N	Minimum	Maximum	Mean	Std. Deviation
Temp	703	.06	.86	.4936	.1831
Atemp	703	.08	.83	.4734	.1627
Hum	703	.25	.97	.6279	.1380
Windspeed	703	.02	.44	.1899	.0765
Cnt	703	431.00	8714.00	4516.4780	1902.2299
Jan	703	.00	1.00	NA (Not Appbl)	NA
Feb	703	.00	1.00	NA	NA
March	703	.00	1.00	NA	NA
April	703	.00	1.00	NA	NA
May	703	.00	1.00	NA	NA
June	703	.00	1.00	NA	NA
July	703	.00	1.00	NA	NA
Aug	703	.00	1.00	NA	NA
Sept	703	.00	1.00	NA	NA
Oct	703	.00	1.00	NA	NA
Nov	703	.00	1.00	NA	NA
Dec	703	.00	1.00	NA	NA
Spring	703	.00	1.00	NA	NA
Summer	703	.00	1.00	NA	NA
Fall	703	.00	1.00	NA	NA

Winter	703	.00	1.00	NA	NA
weekday 1	703	.00	1.00	NA	NA
weekday 2	703	.00	1.00	NA	NA
weekday 3	703	.00	1.00	NA	NA
weekday 4	703	.00	1.00	NA	NA
weekday 5	703	.00	1.00	NA	NA
weekday 6	703	.00	1.00	NA	NA
weathersit_1	703	.00	1.00	NA	NA
weathersit_2	703	.00	1.00	NA	NA
weathersit_3	703	.00	1.00	NA	NA

Table 3: Coefficients of Regression (Source: Author)

Model	Unstandardized Coefficients (Response Variable = cnt)			
	B	Std. Error	t	Sig.
1 (Constant)	4567.157	406.058	11.248	.000
temp	7031.969	648.074	10.851	.000
Hum	-3563.312	473.934	-7.519	.000
windspeed	-3471.414	640.650	-5.419	.000
Jan	235.395	279.960	.841	.401
Feb	210.804	283.372	.744	.457
March	352.045	288.488	1.220	.223
April	-11.964	387.999	-.031	.975
May	153.040	410.860	.372	.710
June	-470.427	421.167	-1.117	.264
July	-759.841	449.189	-1.692	.091
Aug	-164.138	429.240	-.382	.702

Sept	751.276	358.066	2.098	.036
Oct	286.497	253.176	1.132	.258
Nov	-179.087	237.306	-.755	.451
Spring	-1688.589	280.589	-6.018	.000
Summer	-651.577	344.391	-1.892	.059
Fall	-1089.325	319.478	-3.410	.001
weekday 1	202.147	164.052	1.232	.218
weekday 2	368.900	167.036	2.209	.028
weekday 3	445.015	165.848	2.683	.007
weekday 4	420.350	165.636	2.538	.011
weekday 5	420.175	166.046	2.530	.012
weekday 6	386.971	167.072	2.316	.021
weathersit_2	-192.036	121.059	-1.586	.113
weathersit_3	-1891.411	339.817	-5.566	.000

4.3. Machine Learning Results

4.3.1. Introduction

We claimed that no single learning algorithm could consistently outperform others across all data sets. As a result, we used an empirical approach to determine the accuracy of the candidate algorithms for the problem and then choose the one with the best accuracy in machine learning; a classifier is an algorithm that sorts or categorizes data into one or more of a set of "classes" automatically. A classifier is an algorithm - the principles that robots use to categorize data. On the other hand, the end product of your classifier's machine learning is a classification model. The classifier is used to train the model, and the model is then used to classify your data. Both supervised and unsupervised classifiers are available. Unsupervised machine learning classifiers are fed just unlabeled datasets, which they sort into categories based on data structures, pattern recognition, and anomalies. Training datasets are provided to supervised and semi-supervised classifiers, which teach them to categorize data into specified categories. Classification is a type of supervised learning in which the input data is also delivered to the objectives. Classification has numerous uses in various fields, including medical diagnosis, credit approval, and target marketing.

4.3.2. Types of Classifiers

The classifiers that were used in this study include; Random Forest, Decision Trees, Nearest Neighbor and XGBoost.

4.4. *Number of Clasifiers used*

4.4.1. Random Forest

A random forest is a machine-learning technique for solving classification and regression problems. It uses ensemble learning to solve complex problems by combining multiple classifiers. Many decision trees make up a random forest algorithm. Bagging or bootstrap aggregation is used to train the 'forest' formed by the random forest method. Bagging is a meta-algorithm that increases the accuracy of machine learning methods by grouping them. The random forest method determines the outcome based on decision tree forecasts. It forecasts by averaging or averaging the output of various trees. The precision of the result improves as the number of trees grows. A random forest method overcomes the drawbacks of a decision tree algorithm. It reduces dataset overfitting and improves precision.

A random forest algorithm's building component is decision trees. A decision tree is a decision-making tool with a tree-like structure. A basic understanding of decision trees will aid our understanding of random forest algorithms. There are three parts to a decision tree: decision nodes, leaf nodes, and root nodes. A decision tree method separates a training dataset into branches, each further divided into branches. This pattern repeats until a leaf node is reached. There is no way to separate the leaf node any farther. The nodes in the decision tree represent the attributes utilized to forecast the outcome. The leaves are connected to the decision nodes. The fundamental distinction between the decision tree and the random forest algorithms is that the latter randomly establishes root nodes and segregates nodes. The bagging method is used by the random forest to generate the required forecast. Rather than using a single sample of data, bagging includes using many samples (training data). These outputs will be ranked, and the one with the best score will be chosen as the final product. Random forest classification uses an ensemble methodology to achieve the desired result. Various decision trees are trained using the training data. This dataset contains observations and features that will be chosen at random when nodes are split (Sarica et al., 2017). Various decision trees are used in a rain forest system. There are three types of nodes in a decision tree: decision nodes, leaf nodes, and the root node. Each tree's leaf node represents the final output produced by that particular decision tree.

The other duty that a random forest algorithm does is regression. The principle of simple regression is followed by a random forest regression. In the random forest model, the values of dependent (features) and independent variables are passed. In terms of data extrapolation, random forest regression isn't ideal (Ren et al., 2017). Unlike linear regression, which uses present data points to estimate values outside of the observation range, nonlinear regression employs existing observations to estimate values outside of the observation range. This explains why the majority of random forest applications are related to classification. When the data is sparse, random forest does not generate good results. In this situation, the bootstrapped sample

and the subset of features will result in an invariant space. This will result in ineffective divides, which will have an impact on the outcome.

In our analysis, the variance explained is 85.15%, this implies that the out-of-bag predictions explain 85.15% of the target variance of the training set.

Table 4: Variable Importance

	IncNodePurity
yr	602435623.9
mnth	111016434.8
holiday	8734000.2
weekday	57549945.5
workingday	4866579
temp	724260733.4
atemp	290195256
hum	134594085.5
windspeed	57851177.8
season_2	3542049.4
season_3	788345.5
season_4	78673546.9
weathersit_2	11980229.5
weathersit_3	25934755.2
weathersit_4	0

According to the Random Forest model, the five most important variables are year, month, whether the day is a holiday or not, the day of the week and whether the day is a working day or not. The algorithm was then tested for negative values. The sum of the negative values is zero implying that the algorithm is not predicting any negative values

Table 5. Variable Importance

RMSE_Dectree1	MAD_De ctree1	MAPE_ Dectree1	RMSE_t rain_rf1	MAD_tr ain_rf1	MAPE_tr ain_rf1
854.6789	1566.626	19.99539	306.4635	1461.862	22.57648

The MAPE is 22.58, this implies that the Random Forest model has an accuracy of 77.43%.

4.5. Decision Trees

Decision Tree Analysis is a general-purpose predictive modeling tool with a wide range of applications. In general, decision trees are built using an algorithm that finds different ways to split a data set based on various conditions. It's one of the most popular and practical supervised learning methodologies (Patel & Prajapati, 2018). Decision Trees are a non-parametric supervised learning technique that can be used for regression and classification. The objective is to generate a model that predicts the values of targeted variables using simple decision rules inferred. A decision tree is a tree-like graph with nodes indicating the point at which we select an attribute, edges representing the responses to the query, and leaves represent the actual output. With a simple linear decision surface, they are employed in non-linear decision making (Zhang et al., 2020). The examples are classified using decision trees by sorting them along the tree from the root to a leaf node, with the leaf node supplying the classification to the example. Each node in the tree represents a test case for an attribute, with each edge descending from that node representing one of the test case's possible solutions. This is a cyclical procedure that occurs for each subtree rooted at the new nodes.

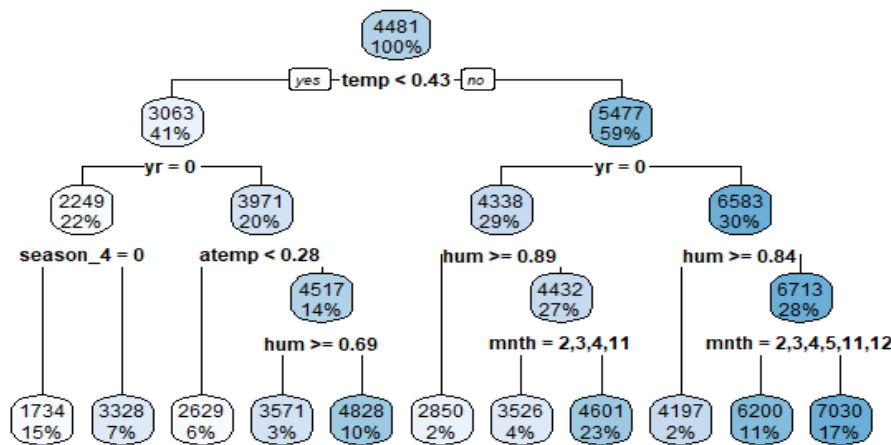


Figure 8: The Decision Tree

The Figure shows a decision tree that shows that the most important variables that predict the number of bikes rented are temperature, Year, Winter (season 4), humidity and the months of February, March, April and November. The algorithm was then tested for negative values. The sum of the negative values is zero implying that the algorithm is not predicting any negative values. The RMSE (Root Mean Square Error), MAD (Mean Absolute Deviation) and MAPE (Mean Absolute Percentage Error) are then calculated.

Table 6: Decision Tree Analysis of MAD, MAPE and RMSE

RMSE_Dectree1	MAD_Dectree1	MAPE_Dectree1
854.6789	1566.626	19.99539

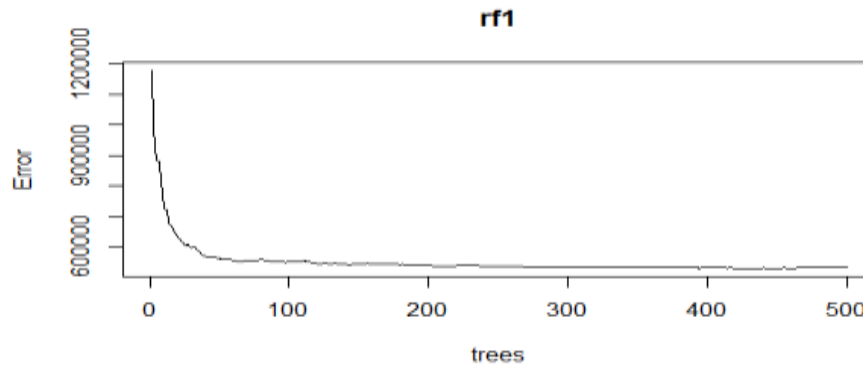


Figure 9: Accuracy Assessment of the Decision Tree

The plot shows that after 100 trees, the Error is stable, with no further drops in the error value. Let's fine-tune the model to determine the values of m_{try} . In the `tuneRF` function, set `ntreeTry = 100`.

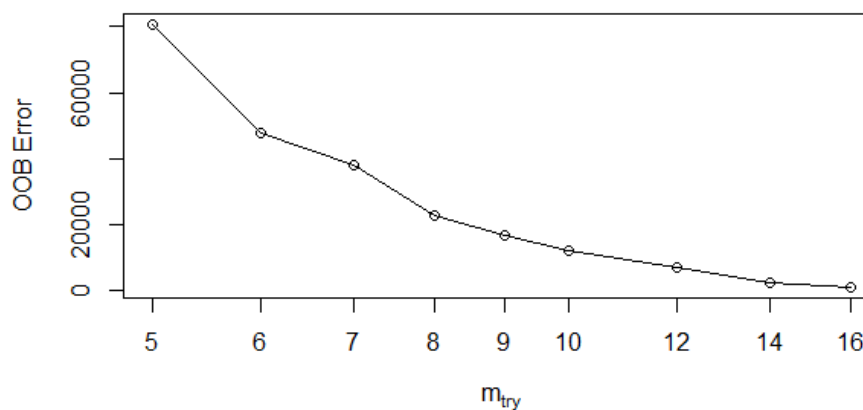


Figure 10: Accuracy Assessment of the Decision Tree

The suggested mtry value after tweaking these parameters is 12. Put these values in our model (mtry = 12 and ntreeTry = 100) and run it again. After fine-tuning, the variance explained is 85.12%, this implies that the out-of-bag predictions explain 85.12% of the target variance of the training set.

Table 8: Variable of Importance

Table 8: Variable of Importance

	%IncMSE	IncNodePurity
yr	1942636.02	592357292
mnth	267405.897	117574875
holiday	4272.415	7034547
weekday	42281.421	54917750
workingday	9532.101	6223112
temp	1921256.69	695179159
atemp	853648.574	339750731
hum	347429.819	126627480
windspeed	64558.113	60497177
season_2	8858.832	3215250
season_3	2331.159	1220109
season_4	174573.072	77609422
weathersit_2	39579.963	11856902
weathersit_3	58312.936	31275631
weathersit_4	0	0

After fine tuning the Random Forest model, the five most important variables are year, month, whether the day is a holiday or not, the day of the week and whether the day is a working day or not.

Nearest Neighbor and Deep Learning?

Because of its simplicity, ease of implementation, and efficacy, KNN (k-nearest neighbor) is a widely used classification technique. It is one of the top 10 data mining algorithms and has a wide range of applications. KNN has a few flaws that impair its categorization accuracy. It has a lot of memory requirements and a lot of time complexity (Taneja et al., 2014). The value of k must be chosen carefully in order for KNN to work. In real-world data sets, certain classes have more data points than others. In most circumstances, if k is a fixed, user-defined value, the result will be biased towards the majority class. Dynamic KNN is another good method for learning the best k value during training period (DKNN). It is based on the leave-one-out cross-validation method, which is a hybrid of eager and lazy learning (Gupta, 2012).

The similarity or difference between the training and test instances is measured by KNN using standard Euclidean distance. It takes into account the equal participation of all of the instance's

qualities, whether or not they are significant. As a result, when there are a high number of irrelevant qualities, the distance function's value becomes erroneous, which is referred to as the Curse of Dimensionality (GV, 2020). To solve this problem, assign varying degrees of priority to each attribute and weight each attribute differently when computing distance between two instances.

If $k=3$, the k nearest neighbor procedure is used to find the nearest neighbor of a new data item. If $k=3$, the three closest neighbors are verified, and the most common data item class is assigned to the new data item. This is a discussion of the KNN algorithm. So, how do we figure out the distance between k and the new data item? We can use Euclidean distance to calculate the distance between k and the new data point. For the KNN method, we may also use the hamming distance and Manhattan distance formulas to determine distance.

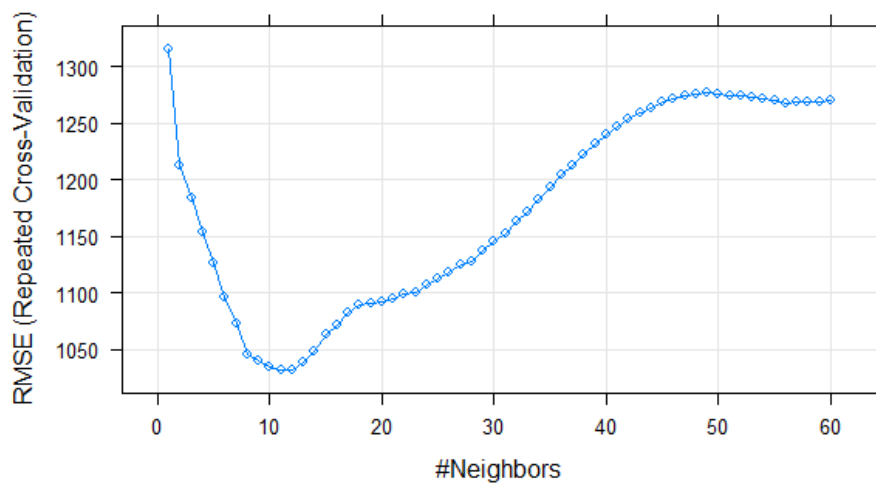


Figure 11: RMSE Analysis

RMSE was used to select the optimal model using the smallest value. The final value used for the model was $k=12$. The algorithm was then tested for negative values. The sum of the negative values is zero implying that the algorithm is not predicting any negative values.

Table 9: Varriables Analysis

Rsquared	Overall
atemp	100
temp	99.6821
yr	66.3144
season_3	24.6839
Hum	15.4853

Mnth	14.7678
Windspeed	14.6566
weathersit_3	13.4499
weathersit_2	4.5263
season_2	4.4449
Workingday	0.8906
Holiday	0.3038
Weekday	0.2115
season_4	0

According to the KNN algorithm, the five most important variables are temperature, year, fall, humidity and windspeed.

Table 10: Variabiles Analysis based on KNN Model

RMSE_train_knn	RMSE_test_kn	MAD_train_knn1	MAD_test_knn	MAPE_test_kn
1	n1		1	n1
955.9719	977.7355	1129.86	1241.963	26.94667

The MAPE is 31.84, this implies that the KNN model has an accuracy of 68.16%.

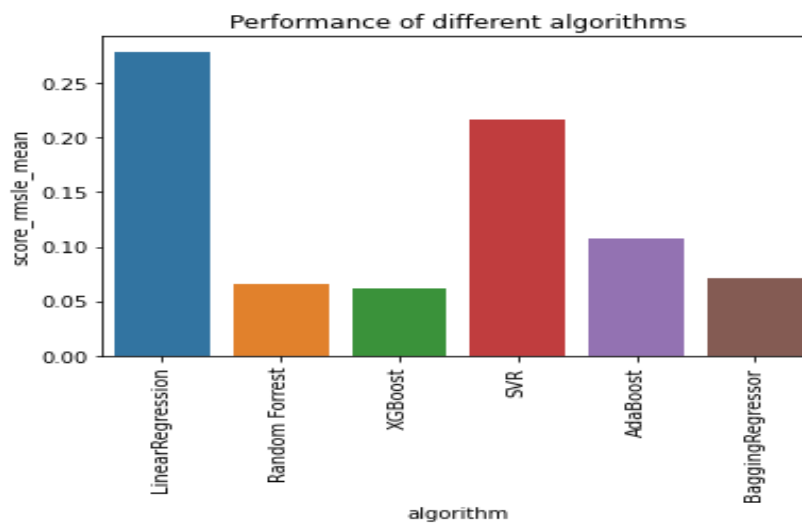
model comparison

A model comparison algorithm ran in Python generated the comparison below

Table 11: Python Model Comparison

	algorithm	score_rmsle_mean	score_std	time
0	LinearRegression	0.2787	0.1099	1.477
1	Random Forrest	0.0662	0.0407	2.716
2	XGBoost	0.0618	0.0248	1.747
3	SVR	0.2160	0.1017	0.210
4	AdaBoost	0.1076	0.0401	1.084
5	BaggingRegressor	0.0711	0.0482	0.414

Figure 12: Performance of various algorithms



Model comparison shows that XGBOOST is the best model based on rmsle

XGBoost

The XGBoost classifier is a machine learning technique that may be used to classify both structured and tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. XGBoost is a gradient boost technique with high gradients. As a result, it's a large Machine Learning algorithm with a lot of moving pieces. XGBoost is capable of handling huge, complex datasets. XGBoost is a strategy for ensemble modeling. XGBoost is a method of ensemble learning. It may not always be enough to rely on the outcomes of a single machine learning model. Ensemble learning is a method for combining the predictive abilities of numerous learners in a systematic way. The end result is a single model that combines the outputs of numerous models. The foundation learners, or models that make up the ensemble, could be from the same learning algorithm or from distinct learning algorithms. The most extensively used ensemble learning models are bagging, boosting, stack generalization, and expert mixtures. Bagging and boosting, on the other hand, are two highly appreciated ensemble learners. Though these two strategies can be applied to a variety of statistical models, decision trees have been the most popular. In this study, XGBoost was selected as the best model.

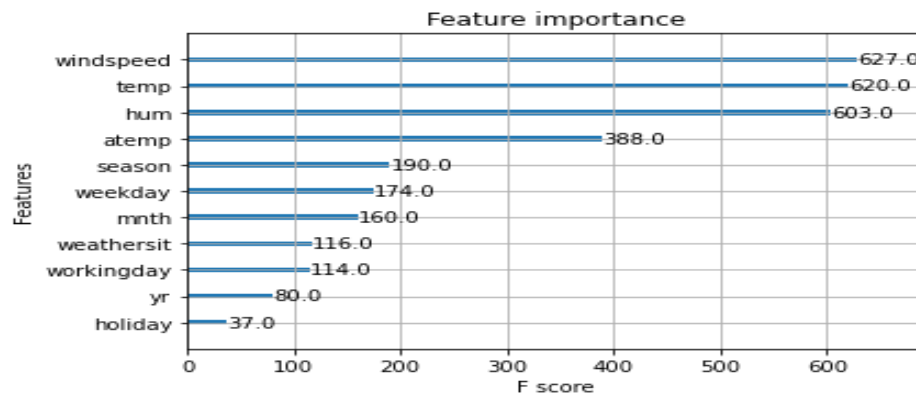


Figure 13: Model comparison of XGBOOST

The most important predictors according to the XGBoost algorithm are; Windspeed, Temperature, Humidity, Seasons, Whether it was a weekday or not, month of the year, weather situation, whether it was a working day or not, year and whether it is a holiday or not.

Discussion

The analysis revealed that an increase in temperature by 10 Celsius will lead to an increase in the number of renters by 7032; it also revealed that an increase in humidity by one unit leads to a decrease in the number of renters by 3564. The analysis also revealed that an increase in wind speed by one unit will decrease the number of renters by 3471; this is consistent with Nosal and Miranda-Moreno (2014). They used hourly data collected from induction loop counters on cycle lanes to evaluate the impact of weather on cycling in multiple North American cities. They established that temperature and humidity have significant impacts. The analysis also revealed that Spring and Fall seasons have a statistically significant effect on the number of bike renters; this is consistent with Lyu et al. (2021), who determined that factors like spring festival significantly affect biking. According to the researchers, these temporal characteristics suggest that the bike turnover rate could rise even higher if the efficiency of the bike supply is increased by proper bike relocation and location.

The analysis also revealed that the month of September had a significant effect on the number of bikers. It was also established that all the days of the week had a significant effect on the number of renters except the first day of the week; this is consistent with Lyu et al. (2021), who established that different days of the week attract different numbers of bikers, for instance in Ningbo, China, the weekends register between 20%-30% lower rental numbers than the weekdays. The analysis also revealed that weather situation 3 (Light Snow, Light Rain with Scattered Clouds) had a statistically significant effect on the bike rentals.

5. Conclusion (NTBR and EXPANDED accordingly)

Bike-sharing systems have become popular in recent years all around the world. Although this trend has resulted in many studies on public cycling systems, there have been few previous studies on the factors influencing public bicycle travel behaviour. A bike-sharing system is a

service in which individuals can borrow bikes for a fee or free for a limited period of time. Many bike share programs allow users to borrow a bike from a system, which is usually computer-controlled. The user enters payment information, and the system unlocks the bike. After that, the bike can be returned to a system-wide dock. The study's goal is to figure out how much demand there is for shared bikes across the country based on compelling parameter estimates. Rental firms arrange this to position themselves to meet people's requirements whenever the situation improves overall, allowing them to stand out from other service providers and earn handsomely. My focus is to discover which variables are essential in predicting shared bike demand. How well those variables accurately characterize the bike's requirements The service provider organization has amassed a vast dataset on daily bike requests across the market based on some parameters which can reliably be applied in predicting potential demand.

The regression analysis carried out determined that the variables that significantly predict the total number of registered & unregistered bikers: renters (cnt) include Temperature, humidity, wind speed, the month of September, Spring, Fall, all weekdays except weekday one and weather situation 3 (Light Snow, Light Rain with Scattered Clouds). Comprehending the temporal features of bike-sharing usage may aid service providers and policymakers in improving bike-sharing services. The study also employed machine learning algorithms to determine the most important predictors of the number of bikes rented, the XGBoost was identified as the best algorithm in terms of accuracy, the algorithm determined that the most important predictors include; Windspeed, Temperature, Humidity, Seasons, Whether it was a weekday or not, month of the year, weather situation, whether it was a working day or not, year and whether it is a holiday or not.

Bike sharing programs (BSP) continue to evolve and expand at a rapid pace. Many countries have implemented various BSP concepts and techniques since the 1960s. There are a variety of versions available, ranging from dockless to electronic real-time monitoring systems. Recreation, errands, work, and other activities may all be done with these BSP. And all signs point to the introduction of more complex and inventive rider-friendly technology in the future. The goal of this article is to examine current variables established by various operators and streamline them using analytics to discover the most appealing ones. There is a lack of standardization and a single criterion on what is required and what is not, given the contents of existing data sets. There appear to be two elements in common: user type and device type (registered and unregistered, and duration of each trip). This article is based on historical data provided by a single operator in the Washington, District of Columbia, United States. Several variables were tested, including categorical and continuous data types. Eight of the 18 were deemed acceptable and contributed significantly to the development of usable and reliable predictive model. Bike-sharing systems have grown in popularity around the world in recent years. Despite the fact that this trend has resulted in a slew of studies on public bicycle systems, there have been few studies on the factors that influence public bicycle travel behavior. Bike-sharing is a computer-controlled system that allows people to rent bikes for a price or for free for a limited time.

References

- Albuquerque, V., Sales Dias, M., & Bacao, F. (2021). Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review. *ISPRS International Journal of Geo-Information*, 10(2), 62. <https://doi.org/10.3390/ijgi10020062>
- Ashqar, H. I., Elhenawy, M., Almannaa, M. H., Ghanem, A., Rakha, H. A., & House, L. (2017). Modelling bike availability in a bike-sharing system using machine learning. 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). Published. <https://doi.org/10.1109/mtits.2017.8005700>.
- Caggiani, L. & Ottomanelli, M. (2013). A dynamic simulation-based model for optimal fleet repositioning in bike-sharing systems. *Procedia-Social and Behavioral Sciences*, 87, 203-210.
- Campbell, A. A., Cherry, C. R., Ryerson, M. S., & Yang, X. (2016). Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. *Transportation Research Part C*, 64 (2016), 399-414.
- Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions. *PeerJ*, 5. <https://doi.org/10.7717/peerj.3323>
- Fishman, E. (2015). Bikeshare: A review of recent literature. *Transport Reviews*, 36(1), 1-22.
- Fishman, E., Washington, S., Haworth, N., & Watson, A. (2015). Factors influencing bike share membership: An analysis of Melbourne and Brisbane. *Transportation research part A: policy and practice*, 71, 17-30.
- Goodman, A., Green, J., & Woodcock, J. (2014). The role of bicycle sharing systems in normalising the image of cycling: An observational study of London cyclists. *Journal of Transport & Health*, 1(1), 5–8. <https://doi.org/10.1016/j.jth.2013.07.001>
- Gu, T., Kim, I., & Currie, G. (2019). Measuring immediate impacts of a new mass transit system on an existing bike-share system in China. *Transportation Research Part A: Policy and Practice*, 124, 20–39. <https://doi.org/10.1016/j.tra.2019.03.003>
- Guo Y, Zhou J, Wu Y, Li Z (2017) Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China. *PLoS ONE*, 12(9), 1-19.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis*. Cengage Learning.
- Hamad, S. Y. Y. (2020). Analysis and Prediction of Bike Sharing Traffic Flow. Technical University of Munich Chair of Transportation Systems Engineering. Published.
- Institute for Transportation and Development. (2018). The bikeshare planning guide. *IDTP*, 5-110.

- Kim, K., Ghimire, J., Pant, P., & Yamashita, E. (2021). Bikeshare and safety: risk assessment and management. *Transportation Research Interdisciplinary Perspectives*, 9, 1-9.
- Kim, T. Y., Park, M. J., Shin, J., & Oh, S. (2021). Prediction of Bike Share Demand by Machine Learning. *International Journal of Business Analytics*, 9(1), 1–16. <https://doi.org/10.4018/ijban.288513>
- Kleinbaum D.G et al. (2014, 2008, 1998) *Applied Regression Analysis and other Multivariate Mehods*, Cengage Learning
- Lyu, C., Bie, J., & Wang, H. (2021). Discovery of temporal variations of public bike ridership using the internet of things for builda ing smart city. *Microprocessors and Microsystems*, 104320. <https://doi.org/10.1016/j.micpro.2021.104320>
- Maizlish, N., Woodcock, J., Co, S., Ostro, B., Fanai, A., & Fairley, D. (2013). Health Cobenefits and Transportation-Related Reductions in Greenhouse Gas Emissions in the San Francisco Bay Area. *American Journal of Public Health*, 103(4), 703–709. <https://doi.org/10.2105/ajph.2012.300939>
- Nosal, T., & Miranda-Moreno, L. F. (2014). The effect of weather on North American bicycle facilities: A multi-city analysis using automatic counts. *Transportation Research Part A: Policy and Practice*, 66, 213–225. <https://doi.org/10.1016/j.tra.2014.04.012>
- Oppermann, M., Möller, T., & Sedlmair, M. (2018). Bike Sharing Atlas: Visual Analysis of Bike-Sharing Networks. *International Journal of Transportation*, 6(1), 1–14. <https://doi.org/10.14257/ijt.2018.6.1.01>
- Ricci, M. (2015). Bike sharing: A review of evidence on impacts and processes of implementation and operation. *Research in Transportation Business & Management*, 15, 28–38. <https://doi.org/10.1016/j.rtbm.2015.03.003>
- Sohrabi, S., Paleti, R., Balan, L., & Cetin, M. (2020). Real-time prediction of public bike sharing system demand using generalized extreme value count model. *Transportation Research Part A: Policy and Practice*, 133, 325–336. <https://doi.org/10.1016/j.tra.2020.02.001>
- Soriguera, F., Casado, V., & Jiménez, E. (2018). A simulation model for public bike-sharing systems. *Transportation Research Procedia*, 33, 139-146.
- Tran, T. D., Ovtracht, N., & d’Arcier, B. F. (2015). Modelling bike sharing system using built environment factors. *Procedia Cirp*, 30, 293-298.
- Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, 106, 234–240. <https://doi.org/10.1016/j.sbspro.2013.12.027>
- Wang, B., & Kim, I. (2018). Short-term prediction for bike-sharing service using machine learning. *Transportation Research Procedia*, 34, 171–178. <https://doi.org/10.1016/j.trpro.2018.11.029>

- Wang, Y., Zhang, D., Liu, Y., Dai, B., & Lee, L. H. (2019). Enhancing transportation systems via deep learning: A survey. *Transportation Research Part C: Emerging Technologies*, 99, 144–163. <https://doi.org/10.1016/j.trc.2018.12.004>
- Yang, T., Li, Y., & Zhou, S. (2019). System dynamics modelling of dockless bike-sharing program operations: A case study of mobile in Beijing, China. *Sustainability*, 11(6), 1–20.
- Zhou, Q., Wong, C. J., & Su, X. (2020). Machine Learning Approach to Quantity Management for Long-Term Sustainable Development of Dockless Public Bike: Case of Shenzhen in China. *Journal of Advanced Transportation*, 2020, 1–13. <https://doi.org/10.1155/2020/8847752>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4>).