Prediction of Bike Share Demand by Machine Learning: Role of Vehicle Accident as the New Feature

Tae You Kim, Catholic University of Korea, South Korea

Min Jae Park, Ajou University, South Korea

Jiho Shin, Business School Lausanne, South Korea

Sungwon Oh, Business School Lausanne, South Korea

ABSTRACT

In the fourth industrial revolution period, multinational companies, and start-ups have applied a sharing economy concept to their business and have attempted to better serve customer demand by integrating demand prediction results into their business operations. For survival amongst today's fierce competition, companies need to upgrade their prediction model to better predict customer demand in a more accurate manner. This study explores a new feature for bike share demand prediction models that resulted in an improved RMSLE score. By applying this new feature, the number of daily vehicle accidents reported in the Washington, DC area to the random forest, XGBoost, and LightGBM models, the RMSLE score results improved. Many previous studies have primarily focused on feature engineering and regression techniques within a given dataset. However, this study is meaningful because it focuses more on finding a new feature from an external data source.

KEYWORDS

Bike Share Demand Prediction, Daily Crashes Reported in Washington, DC, Machine Learning, New Feature Exploration, Sharing Economy

1. INTRODUCTION

Predicting customer demand accurately in the data-driven business environment of the mobility industry is a key factor of success (Sohrabi et al., 2020; Wessel, 2020). In the data technology period, it is very easy to find multinational companies that provide their various services to customers based on demand prediction results. In the short term, companies have achieved success because they have sufficiently predicted demand based on internal and external features. Digital transformation is occurring in most industries around the world based on machine learning and deep learning algorithms (Veres & Moussa, 2019; Wang et al., 2019). In particular, rapidly growing start-ups are often making their important business decisions based on data and algorithms, not on management experience or intuition.

DOI: 10.4018/IJBAN.288513

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

In a modern society where environmental problems are increasingly serious, the bike is an eco-friendly means of transportation that benefits from individuals maximizing its exercise effects during the COVID-19 epidemic. The bike has attracted worldwide attention since a few years ago as it has been transformed into a “shared” good. It is not just for leisure purposes, but for the transportation itself (Böcker et al., 2020; Gu et al., 2019). Shared bike programs in many cities around the world are considered a means of responding to climate change and energy crises when implementing sustainable transportation systems. A shared bike system is an important and necessary means to promote the use of bikes and further implement an efficient urban transportation system. However, despite its many advantages, the number of cities and businesses that are reluctant to introduce a shared bike program is increasing due to the disadvantages of high fixed costs such as installation cost or operating cost. Thus, an accurate demand prediction is required for the continuous operation of a bike sharing program. One of the earliest shared economy models in the mobility industry was the bike sharing system. In particular, Capital Bikeshare, which launched the bike rental business in the Washington, D.C. area, hosted a data science competition in the Kaggle Competition platform to predict customer bike share demand (Fanaee-T & Gama, 2014). In response, a number of efforts have been made by data scientists and data analysts around the world to predict demand by using various data mining techniques.

Shared bikes are emerging as an alternative means of transportation that can comprehensively improve environmental problems, traffic jams, poor quality of life caused by car-oriented transportation systems, and can guarantee practical effects (de Chardon, 2019; Gu et al., 2019). In addition, shared bikes have several advantages when establishing an integrated public transportation system in connection with public buses and urban rail systems. First, the average distance of a bike trip is around 3-5 km (last mile), but it is possible to travel longer distances in connection with public transportation. Secondly, the catchment area of public transportation can be extended from an existing pedestrian area to an area coverable by bike. Third, during non-adjacent hours (except rush hour when traffic demand is concentrated during the day), highly demanded shared bikes can accommodate leisure traffic on existing transit, which can increase the overall utilization rate of public transportation. Bikes are the best means to realize an integrated public transportation system by supporting traffic in combination with walking or public transportation (Böcker et al., 2020; Gu et al., 2019). An integrated public transportation system can increase the availability of public transportation by both merging public transportation services, such as buses or metros, into high capacity corridors where economies of scale are built and strengthening cooperation with corridors or feeders using bikes (Gu et al., 2019). Therefore, it is imperative to predict the demand for shared bikes to efficiently implement an integrated public transportation system that includes shared bikes.

In previous studies, many focused on the importance of features within the models, such as the distance between rental and return location (Ma et al., 2020; Younes et al., 2020). These studies primarily focused on engineering feature techniques and statistical modeling within given datasets. Therefore, there were model limitations from previous studies since they only used given datasets. This study, however, focuses on exploring a new feature from an external data source. Existing studies have predicted demand for bikes through statistical methodologies, but few have applied machine learning methodologies, which have recently been in the spotlight. The purpose of this study is to present a machine learning prediction model that uses variables that affect the demand for shared bikes. Furthermore, from the perspective of an integrated public transportation system, a predictive model that includes these variables is presented with the belief that traffic accident occurrence affects the demand for shared bikes as an alternative transportation system. The analysis results are expected to provide policy implications for efficient operation of a city’s integrated public transportation system.

The sequence of study is as follows. Section 2 provides a review of literature and reviews the history of bike share programs across the globe and in the U.S. and the prediction of bike share demand by various data mining and statistical approaches. Section 3 covers methodology and discusses data set overview, exploratory data analysis, and data preprocessing. In addition, the new feature is introduced

and applied to three different machine learning algorithms. Section 4 is the model and results section and compares each model's prediction performance with their RMSLE scores. Section 5 is the final section and concludes how the prediction of bike share demand could contribute to its operation and business. It also explores the limitations of this study and proposes further study topics.

2. LITERATURE REVIEW

2.1 Bike Share Program and Trend

Over the past 45 years, bike share programs have existed in three generational forms (DeMaio, 2003; DeMaio & Gifford, 2004). The first generation of a bike sharing program was launched by Witt Fiesten in Amsterdam in the Netherlands on July 28, 1965. A typical bike type, painted white, was provided for public rental use. Anyone could use the program and, if it was placed near a destination for the next user, it could offer continuous use, but it did not function properly and ended within a few days (DeMaio, 2009). The second generation of the bike share program was implemented in Denmark in 1991 and 1993 (Nielsen, 1993). It was a small program consisting of twenty-six bike rental service types at four stations and later hosted a massive program in Copenhagen called City Bike in 1995. Copenhagen Bike was specifically built for practical purposes using hard rubber tires and wheels with advertising plates. It could even be returned with the payment and receipt of coin deposits at certain locations throughout the city. The program was formalized when it was managed by a station company and a non-profit organization but it still suffered from theft, which resulted in the creation of a next-generation bike-sharing program with improved tracking capabilities (DeMaio 2009). The third generation of the bike-sharing program was the Bike about program at the University of Portsmouth in England, which was implemented in 1996. Since University of Portsmouth students used electronic cards to rent bikes (Black et al., 1998), the bike share system became smart through various technological improvements, including electronic locking and communication systems, smart card and mobile phone access, and on-board computers (Demaio, 2009).

The bike share program has grown slowly in the third generation, with one to two new bike share programs starting each year, and 1,500 bikes were implemented by JCDecaus in Lyon, France (Optimising Bike Sharing in Europe Cities, 2009). This was the largest of the third-generation bike-sharing programs. 15,000 members used bikes 6.5 times a day. Paris, the capital of France, also used the occasion to provide the program attention (Henley, 2005). Two years later, Paris launched its own bike-sharing program and grew in the city and suburban areas from about 7,000 to 23,600 bikes. This large-scale implementation has dramatically changed the flow of bike-sharing programs around the world. It has also been implemented outside Europe since 2008, with large numbers of programs in Brazil, Chile, China, New Zealand, South Korea, Taiwan, and the United States. As a result, there were about 60 bike-sharing programs in each country by the end of 2007 and there were about 60 third-generation bike programs worldwide (DeMaio, 2007). At the end of 2008, there were 92 Bike sharing programs (DeMaio, 2008). Bike-sharing programs have become more popular worldwide. Most bike-sharing programs are fourth-generation systems, which add demand response and multi-modal systems to third-generation systems (Parkes et al., 2013).

The cumulative demand for bike share programs has been increasing since 2007. There were an estimated 639 bike-sharing schemes operating in 53 countries located in nearly every region of the world in 2017, boasting a total of about 643,000 bikes. In the U.S., the total number of shared bike rides has exceeded 88 million since 2010. In 2016, the number of annual rides exceeded 28 million, which is equal to the annual number of rides on the Amtrak system. The five biggest bike share programs in the U.S. (Citi Bike in New York, Capital Bike share in greater Washington, D.C., Citi Bike in Miami, Divvy in Chicago, and Hub way in Greater Boston) contributed to 85% of all bike trips according to a report published by the National Association of City Transportation Officials (NACTO, 2017). This study used the capital bike share's business case in the Washington, D.C. area as it was the first mover in bike share systems in the U.S. There are many different reasons of joining

bike share programs in the Washington, D.C. area. Getting around quicker and more easily was the primary reason for joining the bike share program. It also indicates that people consider the bike share program an alternative travel options that costs less (LDA Consulting, 2017).

In terms of traffic congestion, Washington, D.C. was second among fifteen of the most congested cities in the U.S. where hours lost in congestion each year per capita was 155. The cost of congestion each year per driver was \$2,161, and the total cost of congestion each year for the city was \$4.6 billion. As a result of the extreme traffic congestion in Washington, D.C., there were many reported vehicle accidents. The total collisions have been increasing continuously since 2008 (Department of Transportation, 2018).

This study above indicates that bike share demand has been continuously increasing globally and that the main purpose of joining a bike share program is to get around faster and easier, which means that alternative transportation needs existed due to external events or circumstances such as traffic congestion or daily vehicle accidents. This study discusses predicting bike share demand not only with provided data but also with external data for better prediction results for bike share demand.

2.2 Bike Share Demand Prediction

The bike share demand forecasting study has different key influencing factors depending on the characteristics of a program's operating region. Also, the application of predictive models varies depending on researchers. In this section, we summarize and present the existing bike share demand forecasting study.

Alhusseini (2014) took two methodologies to predict bike share demand. The first approach used a support vector machine (SVM) algorithm to predict the demand, considering bike demand as a numerical attribute. Alhusseini's second approach used softmax regression and SVM algorithms, considering bike share demand a categorical attribute with five class labels. Du et al. (2014) applied time-by-hour data from a bike sharing demand dataset to predict the time-to-time bike share demand. Du estimated demand by stacking the generalized boosted model and random forest. In addition, the generalized linear models with elucidated net regulation (GLMNet), principal component revision (PCR), support vector revision (SVR), and conditional reference tree (Ctree) algorithms were used to forecast the demand. Lee et al. (2014) modified feature sets such that every categorical attribute was binary (had two possible values). The remainder of the attributes with two possible values, such as numeric and categorical, were used as is. For categorical attributes with multiple possible values, a new feature was created for each possible label such that each label has a separate column. They used a Poisson regression model, a neural network, and a Markov model to predict the number of rides.

Also, Kim et al. (2018) proposed a study on predicting bike share demand using a graph convolutional neural network. Wang (2016) applied multiple linear regression, neural networks, decision trees, and random forests to predict the New York's CitiBike's hourly demand. Wang's study focused on the random forest ensemble method to improve accuracy. He created multiple trees by adding new independent attributes, improving the quality of each independent attribute, and transforming the dependent variable. Unlike other studies, Wang (2016) prepared the dataset for himself. His data preparation involved collecting information from three sources: one for historical weather data, one for individual ride information, and one for official holidays in New York. Godavarthy et al. (2017) studied the operational aspect, travel behavior, and travel mode shift using Great Rides Bike Share program data in Fargo, North Dakota, and found that the introduction of the bike share program in Fargo increased the bicycling behavior of North Dakota State University (NDSU) students and Fargo residents. And Liu et al. (2019) predicted bike share demand through standard long short-term memory (LSTM) models and multi-time step models. Pan et al. (2018) predicted bike share demand by using a recursive neural net (RNN) model and showed performance improvement. Li (2019) collected bike share data in Chicago and applied a Gaussian mixture model (GMM) to predict customer demand. Zeng et al. (2016) proposed a station-centric model which accounts for global features such as weather, user activity, and season.

Kaltenbrunner et al. (2010) used an auto-regressive moving-average (ARMA) model to forecast the number of bikes and docks for each station. Yoon et al. (2012) used a modified autoregressive integrated moving average (ARIMA) model to predict the available resources at each station. In their studies, spatial interaction and temporal factors were considered. Lim and Chung (2019) developed a model to predict the demand of Seoul bikes through a time series analysis in a statistical manner. In particular, the Holt-Winters method, which was used to measure the demand of electricity usage, was modified and used to predict demand. Sensitivity analysis was also performed on how the variation of parameters affected actual demand prediction. Chen et al. (2016) predicted excess demand for public bike rentals by implementing a weighted correlation network model. Li et al. (2015) proposed a hierarchical prediction model to predict the check-in/out of each station cluster. They first clustered stations according to a bipartite clustering algorithm based on the geographical locations of stations and transition patterns. They then predicted the traffic in the entire city by way of a gradient boosting regression tree model. The rental proportion across clusters was predicted by a multi-similarity-based inference model.

Furthermore, Min et al. (2017) proposed an analysis methodology that could identify the meaning of the data using visualization. Locational characteristics of stops were found according to utilization rates, and bike use patterns were studied that vary according to time, day, and month. Meanwhile, the use pattern between stops was found through an analysis of travel route, and the purpose of use was identified through an analysis of the destination ratio for each stop. Based on these data, the development direction of public rental bike systems in Daejeon was presented. Kim et al. (2012) studied the effects of the weather on public bike demand based on public bike data from Goyang-si, Korea with temperature variables showing a decrease in bike use if temperatures fell significantly or rose above 23°C. Precipitation has a negative effect on bike use, falling about sixty percent for every 10cm increase in rainfall. In addition, the number of clouds in the sky also has a negative impact on the use of bikes. These results show a similar trend, although there are differences in values compared to overseas studies. Lee et al. (2011) built a bike demand estimation model with features composed of the number of students, excluding elementary school students, and the number of passenger cars.

Vogel et al. (2014) conducted a case study on strategy establishment of bike sharing systems using data mining and explored activity patterns. Lee & Son (2019) proposed an optimal rebalancing strategy for public bike sharing systems in Seoul by using an integer programming method to estimate the number of bikes. Liu et al. (2016) developed a Meteorology Similarity Weighted K-Nearest-Neighbor (MSWK) regressor to predict station pick-up demand based on large-scale historic trip records to resolve the bike share demand rebalancing problem. Lu (2016) presented mathematical programming models that generate optimal daily allocation of bikes to the stations of a bike-sharing system. A time-space network was constructed to describe time-dependent bike flows within the system and a bike fleet allocation model that considered average historical demand and fixed fleet size was then established based on the time-space network. Lu's model targeted correcting demand asymmetry in bike-sharing systems where flow from one station to another is seldom equal to the flow in the opposing direction. The above mentioned bike share prediction studies are summarized in Table 1 below:

There were several previous studies about features and variables regarding bike share prediction. These previous studies only focused on the attributes of each variable, such as categorical or numerical. In contrast, this study focuses on a new feature exploration from an external data source and it demonstrates the importance of each feature per algorithm.

3. METHOD

3.1 Data

Bike sharing system is a tool for bike rentals that automates membership, leasing and return processes over the network, with kiosk locations throughout the city. The system allows people to rent bikes

Table 1. Summary of bike share demand prediction studies

Previous Studies	ML/DL Algorithms Studies	Statistical Analysis Studies	Feature/Variable Studies	Business Operations and Strategy Studies
Alhusseini(2014)	●	●		
Du et al.(2014)	●	●		
Lee et al.(2014)			●	
Kim et al.(2018)	●			
Wang(2016)	●	●		
Godavarthy et al.(2017)				●
Liu et al.(2019)	●			
Pan et al.(2019)	●			
Li(2019)	●	●		
Zeng et al.(2016)			●	●
Kaltenbrunner et al.(2010)		●		
Yoon et al.(2012)		●		
Lim et al.(2019)		●		
Chen at el.(2016)	●	●		
Li et al.(2015)	●	●		
Min et al.(2017)		●		●
Kim et al.(2012)			●	
Lee et al.(2011)		●		
Vogel et al.(2014)	●			●
Lee & Son(2019)		●		●
Liu et al.(2016)		●		
Lu(2016)		●		●

from one place and return them to another as needed. There are currently more than 500 bike-sharing programs around the world (Fanaee-T et al., 2013). The data generated by the system has been attractive to many researchers because the travel period, departure location, arrival location and elapsed time are explicitly recorded. In 2013, Capital Bike share corporation hosted a data science competition on Kaggle platform¹ and participants were asked to predict bike share demand in Washington D.C. area. The dataset consists of 10,866 rows. The datasets for its public bike share users in Washington, D.C. Metadata information is as shown in Table 2.

The data in this study are shared bike riding records and bike station status in the Washington DC area, as well as local weather information for that period. Washington DC data was chosen because unlike data in other cities where data was not disclosed or data was lost, the data was recorded without significant loss. The data on the operation of bikes contained information on the time of rental, return time, and place of rental. For the purpose of analyzing daily usage, the data were classified on a daily basis and the daily rental volume was set as a class. The data were then compiled by adding weekend and holiday information to the Washington DC region weather information (including highest and

Table 2. Bike share metadata

Fields	Description
datetime	hourly date + timestamp
season	1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday	whether the day is considered a holiday
workingday	whether the day is neither a weekend nor holiday
weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	temperature in Celsius
atemp	“feels like” temperature in Celsius
humidity	relative humidity
windspeed	wind speed
casual	number of non-registered user rentals initiated
registered	number of registered user rentals initiated
count	number of total rentals (casual + registered)

lowest temperature, humidity, wind volume, rainfall, etc.) and date information (Wessel, 2020). Thus, the processed data consisted of attributes and instances containing items converted to dummy variables. However, due to the nature of the open data, experiments related to daily usage have already been revealed on the Kaggle website. The purpose of this study is to expand the comparison group and improve the model using the already conducted study as baseline.

Exploratory data analysis (EDA) is a process for analyzing data, and refers to techniques for interpreting these processes, such as how data is collected and planned to perform easier and more accurate data analysis (Tukey, 1980). In statistics, exploratory data analysis is often an approach to analyzing data sets to summarize the nature of the data. Statistical models may or may not be used, but are primarily performed to determine what data can be communicated, apart from formal modeling or hypothesis verification work. The datetime field was divided into year, month, day, hour, minute and second after applying parsing function from pandas package. In addition, ‘dayofweek’ field was created additionally. After performing EDA and preprocessing, there were findings from each data field summarized below Figure 1.

For the date time-year variable, the demand of bike rental was higher in 2012 than in 2011. And, for the date time-month variable, from June to August, most customers rented more bikes and from December to February, less bikes. Even in the same winter season, December’s bike rental demand is higher than January’s. The day in the dataset only existed from 1 to 19 because of including only weekdays for date time-day variable and customers did not rent bikes at dawn, and that they borrow relatively much in the afternoon. Particular tendency for customers to rent bikes was from 7 a.m. to 9 a.m. and 4 p.m. to 7 p.m. according to date time-hour variable. For the number of total rentals count, log transformation was applied to make the target closer to the normal distribution.

The District of Columbia government in the U.S. shares hundreds of datasets online. According to the Open Data DC, D.C. believes that data’s greatest value comes from having it freely shared among agencies, and regional governments and with the public to the extent possible when considering safety, privacy, and security. As mentioned above, the one of the main reasons for joining bike share program was to get around easier and faster without any external events and circumstances. Thus, this study used daily vehicle accident reported data provided by Open Data DC. Crashes in D.C. dataset

Figure 1. Exploratory data analysis summary

which provides daily vehicle crash detail in Washington D.C. area derived from the Metropolitan Police Department's(MPD) crash data management system(COBALT) maintained by the District Department of Transportation(DDOT)'s attempt to summarize the most requested elements of the crash data. The dataset contains 221,107 rows with 56 fields. Data preprocessing was performed and the data was grouped by each day. Each row was grouped by report date as shown below and additionally created 'crash' feature as shown in Figure 2.

Several of the fields used in the analysis show missing and lack of data, and the field is organized and used for final analysis due to the redundant with target variable characteristics among the variables that correspond to the target variable candidates. Finally, selected features for the prediction are summarized in the Table 3.

3.2 Analytical Model

The ensemble technique is used to connect multiple machine learning models to create stronger prediction performance (Sagi & Rokach, 2018). In other words, just as a panel of experts is formed to gather and weigh various opinions to resolve a difficult problem, the objective of ensemble techniques is to obtain a more reliable predictive value than a single classifier by combining the predicted results of various models (Sagi & Rokach, 2018). Ensemble techniques, traditionally, have been divided into three types: voting, bagging, and boosting. Voting and bagging are the ways in which multiple classifiers decide the final forecast results through voting, but the difference between these two is

Figure 2. Number of daily vehicle accident data preprocessing

X	Y	OBJECTID	CRIMEID	CCN	REPORTDATE	Grouping	Crash
-77.02589809	38.98315271	116904736	26610029	11001182	2011-01-03T05:00:00.000Z	2011-01-03	3
-77.02316436	38.92858957	116904732	26610024	11001005	2011-01-03T05:00:00.000Z		
-77.06432637	38.94432349	116904713	26609965	11001150	2011-01-03T05:00:00.000Z		
-76.97936706	38.89838024	116835528	26417537	11001081	2011-01-03T05:00:00.000Z	2011-01-04	3
-77.03810979	38.90724953	116905182	26610230	11001050	2011-01-04T05:00:00.000Z		
-77.03075962	38.90909866	116904905	26609897	11001250	2011-01-04T05:00:00.000Z		
-76.98830706	38.94233947	116904742	26610038	11001137	2011-01-04T05:00:00.000Z	2011-01-05	1
-77.08204347	38.9490809	116904976	26610154	11002018	2011-01-05T05:00:00.000Z		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3. Summary of feature selected

Fields	Selected	Excluded	The reason why the field was excluded
datetime-year	●		
datetime-month	●		
<i>datetime-day</i>		v	Missing data
datetime-hour	●		
<i>datetime-minute</i>		v	Lack of Data
<i>datetime-second</i>		v	Lack of Data
datetime-dayofweek	●		
season	●		
holiday	●		
workingday	●		
weather	●		
temp	●		
atemp	●		
humidity	●		
windspeed	●		
crush	●		
casual		v	Redundant with target variable (multicollinearity)
registered		v	Redundant with target variable (multicollinearity)
log_count	●		Target Variable

that voting generally combines models with different algorithms (González et al., 2020). In the case of bagging, however, each classifier is analyzed based on the same type of algorithm, learning involves taking data sampling differently. Boosting is learning and forecasting by giving weight to the following classifiers so that multiple classifiers can perform learning sequentially but correctly predict data that were not predicted earlier (González et al., 2020). The representative algorithm of bagging is random forest, and the representative algorithms of boosting are XGBoost (eXtra Gradient Boost) and LightGBM (Light Gradient Boost) (Weng et al., 2019).

Random forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Random forest is also an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. XGBoost is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by (Friedman, 2001). The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking (Chen & Guestrin, 2016). XGBoost initially started as a research project by Tianqi Chen as part of the Distributed Deep Machine Learning Community (DMLC) group. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. The same code runs on major distributed

environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples. LightGBM contains two novel techniques: Gradient-based One-Side Sampling and Exclusive Feature Bundling to deal with large number of data instances and large number of features respectively (Ke et al., 2017). LightGBM is a gradient boosting framework that uses tree based learning algorithms. We use bike sharing demand related information to generate vehicle accident information as the new feature. This study applies XGBoost and LightGBM as well as random forest, which had been shown to perform well in previous studies for demand forecasting.

In the model performance validation step, the performance of each prediction model is validated using the evaluation criteria. Generally, the following evaluation criteria were used to calculate the error between the predicted value and the performance value of the actual data as well as determine the accuracy of the predicted value quantitatively. The mean absolute error (MAE) represents the average of the absolute errors between the predicted values and the actual values. The mean absolute percentage error (MAPE) converts the difference between the predicted value and the actual value into a percentage. The root mean square error (RMSE) is the square root of the average of the squared residuals, and it is usually expressed as precision. Finally, the root mean squared logarithmic error (RMSLE) is the log value of the average of the residuals (Muhlestein et al., 2018). Prediction errors may occur in areas with large outliers as well as in areas with small outliers. RMSLE calculates an error by assigning penalties to items with underestimated outliers rather than to items with over-estimated outliers. RMSLE is robust to the outlier. RMSLE gives under estimation greater penalties than over estimation. In other words, a larger penalty is given when the predicted value is less than the actual value than when the predicted value is greater than the actual value. Among the performance measurements, in this study, it was determined that RMSLE was the most intuitive way to identify the effect of error on the data and can be applied well in the transportation demand forecasting performance including bike (Smibi and Menon, 2019). Therefore, RMSLE was considered in the model performance validation step. RMSLE was selected as a performance measure. RMSLE formula is illustrated below:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n \left(\log(p_i + 1) - \log(a_i + 1) \right)^2}$$

$p_i = \text{Predicted Value}$

$a_i = \text{Actual Value}$

4. RESULTS

We implemented a prediction model for 70% of the total dataset and applied 30% dataset to the prediction model to calculate the performance of the actual prediction accuracy of the prediction model via RMSLE. For data refining and analysis, the Python program was utilized in this research. Random Forest model showed 0.322 and 0.321 RMSLE score before and after adding the new feature respectively. XGBoost model showed 0.290 and 0.271 RMSLE score before and after adding the new feature respectively. LightGBM showed 0.304 and 0.294 RMSLE score before and after adding the new feature respectively. As a result, XGBoost showed the most improvement in predicting bike share demand and Random Forest showed the least. LightGBM also showed the improvement. Feature Importance is a base premise for feature selection technique and it is often used for simplification of models to make them easier to interpret by researcher, shorter training times, to avoid the curse of dimensionality and to reduce overfitting (reduction of variance). The results of prediction RMSLE score for each algorithm is shown in Table 4.

Table 4. Results summary

	Random Forest RMSLE	XGBoost RMSLE	LightGBM RMSLE
Before	0.322	0.290	0.304
After	0.321	0.271	0.294

Note: Before means prediction result without "crash" feature & After means prediction result with "crash" feature

Among the three analytical methodologies, XGBoost shows the best prediction performance. For the analysis by the Random Forest model, Date time-hour showed the highest and holiday feature showed the lowest feature importance score. Crash was ranked in 10th. And, for the analysis by the XGBoost, Date time-hour showed the highest and crash feature showed the lowest feature importance score. For the analysis by the LightGBM, Date time-hour also ranked in the highest feature importance and holiday showed the lowest. Crash feature was 3rd.

In summary, these results indicate that the new feature 'crash' does impact on the accuracy of each model. According to the feature importance above, though XGBoost showed improvement on the score, the new feature was the least important feature. On the contrary, in LightGBM model, it showed third important feature among other features. It implies that the number of daily vehicle accidents reported in the Washington D.C. area may be a significant feature to predict bike share demand in Washington D.C. Overall, date time-hour was the most important feature in all three models. In fact, date time-hour showed similar time-series pattern with crash reported as shown in Figure 3.

5. CONCLUSION

Today, bike sharing systems are in operation in many countries around the world. The increase in number of system equipment installations and their operation are expected to continue into the future as citizens' bike share demands increase. From an individual health perspective, a bike sharing system

Table 5. Feature importance summary

Importance Rank	Random Forest Feature	XGBoost Feature	Light GBM Feature
1	datetime-hour	datetime-hour	datetime-hour
2	atemp	datetime-year	humidity
3	temp	workingday	crash
4	humidity	datetime-dayofweek	windspeed
5	datetime-month	datetime-month	temp
6	datetime-year	atemp	atemp
7	season	season	datetime-dayofweek
8	datetime-dayofweek	temp	datetime-month
9	workingday	humidity	weather
10	crash	weather	season
11	weather	holiday	datetime-year
12	windspeed	windspeed	workingday
13	holiday	crash	holiday

Figure 3. Vehicle accidents and bike share demand comparison by hour

is essential to reduce air pollution such as fine dust particles caused by the use of internal combustion engine vehicles and to ultimately solve urban transportation problems such as traffic congestion and public transportation inadequacy. Recently, many governments have increased their use of bike sharing systems to protect themselves and their citizens from infectious diseases such as COVID-19. As the number of bike sharing systems increases, corresponding problems of various forms also arise. First, a demand imbalance arises at certain times and at certain rental locations where bikes are not available. Also, most bike-sharing systems use internal combustion engine vehicles such as trucks when moving and rearranging bikes' location. Thus, when demand for bikes is not accurately predicted, the problems of additional cost and time, as well as air pollution, also occur. Therefore, accurately predicting bike demand makes it possible to more effectively operate a transportation system. Furthermore, predicting demand for shared bikes is also crucial for a city's smooth and efficient operation of an integrated public transportation system.

With respect to integrated public transportation systems, this study presents a predictive model by adding traffic accident data to the main variables. A machine learning prediction model is presented using variables that affect demand for shared bikes. It has been confirmed that the existing methodology is more predictable, and it has also been confirmed that urban traffic accident data act as a major variable for increasing predictability. The findings are expected to provide policy implications for efficient operation of a city's integrated public transportation system. This study demonstrated a new feature exploration for better prediction performance of bike share demand through machine learning algorithms. Random Forest, XGBoost, and LightGBM were used in the experiment. The new feature, the number of daily vehicle accidents reported in the Washington, D.C. area, was added to the algorithms and each RMSLE score was compared.

As a result, all of the three algorithms' RMSLE scores were improved. Many previous studies focused on EDA, preprocessing, statistical estimation, and data mining techniques. This study, however, is meaningful as it focuses on exploring a new feature from an external data source from Open Data DC. This new feature exploration will help companies in the mobility industry resolve business operation problems such as bike supply rearrangement (so-called inventory re-balancing problems) or excess supply (cost) of bikes or bike station decks, which often arise from a failure to accurately predict customers' bike share demand.

Academically, this study presented new factors affecting the use of bicycle sharing systems. In particular, it makes an academic contribution to the presentation of a new predictive model of bicycle use, taking into account environmental variables. In addition, we present a new analytical methodology to present a suitable methodology for the prediction. The analysis results of this study show that when people use shared bicycle systems, they are heavily influenced by external factors due to the

system characteristics associated with other public transport systems. It is necessary to strategically respond to bicycle demand by developing a comprehensive model that reflects considerations of external variables such as weather and road conditions. This will help to establish an eco-friendly transportation system in the city center and reduce traffic congestion.

In the COVID 19 situation, the bike sharing system has become an important means of transportation to eliminate the risk of infectious diseases. Adaptation to transportation during the COVID-19 pandemic is essential to mitigate the spread of the virus and prepare for other health crises. While many people live farther away socially and have reduced overall mobility, efficient transportation and mobility in cities are still important to connect people with each other (Jobe & Griffin, 2021). The Bike sharing system provides citizens with the opportunity to use transportation in a safe and healthy way to meet comfort. As cities eventually reopen, bike-sharing policies and users' knowledge of behavior can affect their willingness to choose these shared mobility systems. Government agencies in each country will be informed of prevention guidelines to users of the Bikes haring system, which will be key to ensuring proper compliance. These restriction and recommendations allow bike-sharing operators to ensure the health and safety of their users and continue to provide bike-sharing as a healthy travel option during the COVID-19 pandemic (Bergantino et al, 2021).

6. LIMITATIONS AND FURTHER STUDY

This study had some limitations because there could always be outliers in practice and the model may fail to predict them. The new feature, the number of daily accidents reported, might be a regionally specific factor that can vary. In addition, there may be some countries or states that do not provide daily accident reports to the public. There may be better machine learning or deep learning models that can show better results with proper hyper parameter tuning. Future studies can use geographical information such as longitude and latitude data of bike rental decks, subway station locations, or locations of daily vehicle accidents reported. The distances between those location points can provide a meaningful approach to the data. Predicting mobility demand for other transportation services besides bike shares, such as car pools, vans, or electric motorcycles, can also be considered in the future. Since we analyzed this study with Washington DC data, it is difficult to generalize the findings to bicycle sharing systems around the world. Therefore, it is necessary to analyze data from more cities to draw comprehensive conclusions and present models that can reflect characteristics between cities. Also, a model presentation that reflects the changing nature of public transportation systems in urban areas after COVID 19 is also required.

REFERENCES

- Alhusseini, M. (2014). *Prediction of Bike Sharing Systems for Casual and Registered Users*. Academic Press.
- Bergantino, A. S., Intini, M., & Tangari, L. (2021). Influencing factors for potential bike-sharing users: An empirical analysis during the COVID-19 pandemic. *Research in Transportation Economics*, 86, 101028. doi:10.1016/j.retrec.2020.101028
- Black, C., Faber, O., & Potter, S. (1998). *Portsmouth Bikeabout: A Smart-Card Bike Club Scheme*. The Open University.
- Böcker, L., Anderson, E., Uteng, T. P., & Throndsen, T. (2020). Bike sharing use in conjunction to public transport: Exploring spatiotemporal, age and gender dimensions in Oslo, Norway. *Transportation Research Part A, Policy and Practice*, 138, 389–401. doi:10.1016/j.tra.2020.06.009
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324
- Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S., & Jakubowicz, J. et al. (2016). Dynamic cluster-based over-demand prediction in bike sharing systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 841–852). ACM. doi:10.1145/2971648.2971652
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM. doi:10.1145/2939672.2939785
- Consulting, L. D. A. (2017). 2016 Capital Bikeshare Member Survey Report. Capital Bikeshare.
- de Chardon, C. M. (2019). The contradictions of bike-share benefits, purposes and outcomes. *Transportation Research Part A, Policy and Practice*, 121, 401–419. doi:10.1016/j.tra.2019.01.031
- DeMaio, P. (2003). Smart bikes: Public transportation for the 21st century. *Transportation Quarterly*, 57(1), 9–11.
- DeMaio, P. (2007). *What a year for the bike-sharing*. Academic Press.
- DeMaio, P. (2008). *Bikesharing Wrapup for 2008*. The Bikesharing Blog.
- DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4), 41–56. doi:10.5038/2375-0901.12.4.3
- DeMaio, P., & Gifford, J. (2004). Will smart bikes succeed as public transportation in the United States? *Journal of Public Transportation*, 7(2), 1–15.
- Department of Transportation. (2018). *Traffic Safety Statistics Report for the District of Columbia (2015-2017)*. Author.
- Du, J., He, R., & Zhechev, Z. (2014). Forecasting Bike Rental Demand. In *CS 229 Machine Learning Project*. Stanford University.
- Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3), 113–127. doi:10.1007/s13748-013-0040-3
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.
- Godavarthy, R., Mattson, J., & Taleqani, A. R. (2017). *Evaluation Study of the Bike Share Program in Fargo, North Dakota* (No. SURLC 17-005). SURLC.
- González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64, 205–237. doi:10.1016/j.inffus.2020.07.007
- 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. doi:10.1016/j.tra.2019.03.003

- Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., & Banchs, R. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bike-based public transport system. *Pervasive and Mobile Computing*, 6(4), 455–466. doi:10.1016/j.pmcj.2010.07.002
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146-3154). Academic Press.
- Kim, Won-Kyung, & Son. (2018). Bike Sharing Demands Prediction based on GCN. In *Proceeding on 2018 KISS Conference* (pp. 832-834). Korea Information Science Society.
- Kim, D-J., Shin, H-C., Park, J-S., & Lim, H-J. (2012). The Impact of Weather on Bike Usage - Focus on Usage of Bike-sharing System in Goyang. *Journal of Transport Research*, 77-88.
- Lee, C., Wang, D., & Wong, A. (2014). *Forecasting utilization in city bike-share program* (Vol. 254). Technical report, CS 229 2014 Project.
- Lee, E.-T., & Son, B.-S. (2019). Optimal Rebalancing Strategy for Public Bike-Sharing System in Seoul. *Journal of the Korean Society for Transplantation*, 37(1), 27–38. doi:10.7470/jkst.2019.37.1.027
- Lee, Kim, & Choi. (2011). Development of Regression-based Bike Direct Demand Models. *KSCE Journal of Civil Engineering*, 31(4), 489-496.
- Li, Y., Zheng, Y., Zhang, H., & Chen, L. (2015). Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 33). ACM.
- Lim, H., & Chung, K. (2019). Development of Demand Forecasting Model for Seoul Shared Bike. *Journal of Korea Contents*, 19(1), 132–140.
- Liu, J., Sun, L., Chen, W., & Xiong, H. (2016). Rebalancing bike sharing systems: A multi-source data smart optimization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1005-1014). ACM. doi:10.1145/2939672.2939776
- Liu, X., Gherbi, A., Li, W., & Cheriet, M. (2019). Multi Features and Multi-time steps LSTM Based Methodology for Bike Sharing Availability Prediction. *Procedia Computer Science*, 155, 394–401. doi:10.1016/j.procs.2019.08.055
- Lu, C. C. (2016). Robust multi-period fleet allocation models for bike-sharing systems. *Networks and Spatial Economics*, 16(1), 61–82. doi:10.1007/s11067-013-9203-9
- Ma, X., Ji, Y., Yuan, Y., Van Oort, N., Jin, Y., & Hoogendoorn, S. (2020). A comparison in travel patterns and determinants of user demand between docked and dockless bike-sharing systems using multi-sourced data. *Transportation Research Part A, Policy and Practice*, 139, 148–173. doi:10.1016/j.tra.2020.06.022
- Min, Mun, & Lee. (2017). Demand Forecast for Public Bikes (“Tashu”) in Daejeon using Random Forest. In *Proceeding on 2018 KISS Conference* (pp. 969-971). Korea Information Science Society.
- Muhlestein, W. E., Akagi, D. S., McManus, A. R., & Chambless, L. B. (2018). Machine learning ensemble models predict total charges and drivers of cost for transsphenoidal surgery for pituitary tumor. *Journal of Neurosurgery*, 131(2), 507–516. doi:10.3171/2018.4.JNS18306 PMID:30239321
- NACTO. (2017). *Bike Share in the US: 2010-2016*. National Association of City Transportation Officials.
- Nielsen, B. H. (1993). *The bike in Denmark: present use and future potential*. Ministry of Transport Dänemark.
- Optimising Bike Sharing in European Cities. (2009). *France. FR_Rennes_ Old System tab*. http://www.obisproject.com/palio/html.run?_Instance=obis&_PageID=4&_LngID=21&_CatID=722&pic=4&_Checksum=1012982978/
- Pan, L., Cai, Q., Fang, Z., Tang, P., & Huang, L. (2018). *A Deep Reinforcement Learning Framework for Rebalancing Dockless Bike Sharing Systems*. arXiv 2018, arXiv:1802.04592.
- Parkes, S. D., Marsden, G., Shaheen, S. A., & Cohen, A. P. (2013). Understanding the diffusion of public bikesharing systems: Evidence from Europe and North America. *Journal of Transport Geography*, 31, 94–103. doi:10.1016/j.jtrangeo.2013.06.003

- Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, 8(4), e1249. doi:10.1002/widm.1249
- Smibi, M. J., & Menon, V. (2019). Modeling Compensation of Data Science Professionals in BRIC Nations. In *Emerging Technologies in Data Mining and Information Security* (pp. 631–638). Springer. doi:10.1007/978-981-13-1951-8_57
- 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. doi:10.1016/j.tra.2020.02.001
- Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34(1), 23–25.
- Veres, M., & Moussa, M. (2019). Deep learning for intelligent transportation systems: A survey of emerging trends. *IEEE Transactions on Intelligent Transportation Systems*.
- Vogel, P., Saavedra, B. A. N., & Mattfeld, D. C. (2014,). A hybrid metaheuristic to solve the resource allocation problem in bike sharing systems. In *International Workshop on Hybrid Metaheuristics* (pp. 16–29). Springer. doi:10.1007/978-3-319-07644-7_2
- Wang, W. (2016). *Forecasting bike rental demand using New York Citi Bike data* (MSc thesis). Technological University Dublin, School of Computing College of Science of Health.
- 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. doi:10.1016/j.trc.2018.12.004
- Weng, T., Liu, W., & Xiao, J. (2020). Supply chain sales forecasting based on lightGBM and LSTM combination model. *Industrial Management & Data Systems*, 120(2), 265–279. doi:10.1108/IMDS-03-2019-0170
- Wu, C., & Kim, I. (2020). Analyzing the structural properties of bike-sharing networks: Evidence from the United States, Canada, and China. *Transportation Research Part A, Policy and Practice*, 140, 52–71. doi:10.1016/j.tra.2020.07.018
- Yoon, J. W., Pinelli, F., & Calabrese, F. (2012, July). Cityride: a predictive bike sharing journey advisor. In *2012 IEEE 13th International Conference on Mobile Data Management* (pp. 306–311). IEEE. doi:10.1109/MDM.2012.16
- Younes, H., Zou, Z., Wu, J., & Baiocchi, G. (2020). Comparing the Temporal Determinants of Dockless Scooter-share and Station-based Bike-share in Washington, DC. *Transportation Research Part A, Policy and Practice*, 134, 308–320. doi:10.1016/j.tra.2020.02.021
- Zeng, M., Yu, T., Wang, X., Su, V., Nguyen, L. T., & Mengshoel, O. J. (2016). Improving Demand Prediction in Bike Sharing System by Learning Global Features. Machine Learning for Large Scale Transportation Systems (LSTS). *KDD: Proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining*, 16.

ENDNOTE

- ¹ Data: Kaggle <https://www.kaggle.com/c/bike-sharing-demand/data>

Min Jae Park is the Associate Professor of Dept. of e-business, Ajou University. He was Academic Director of Aalto Univ. Executive MBA in Seoul and an Assistant Professor in the Seoul School of Integrated Science and Technologies (aSSIST), Rep. of Korea and obtained a Ph.D. Degree in the School of Business and Technology Management at the Korea Advanced Institute of Science and Technology (KAIST). His research interests include R&D management, innovation by technology like IoT, and data mining. His research articles have been published in *Technological Forecasting and Social Changes*, *Public Management Review*, *International Journal of Information Management*, *Telecommunications Policy*, *Government Information Quarterly*, etc.