Pamukkale Univ Muh Bilim Derg, 29(8), 824-832, 2023
(41. Yöneylem Araştırması ve Endüstri Mühendisliği Ulusal Kongresi Özel Sayısı-YA/EM'2022)

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Prediction bike-sharing demand with gradient boosting methods

Gradyan artırma yöntemleriyle bisiklet paylaşım talebini tahminleme

Zeliha ERGUL AYDIN[1], *Banu ICMEN ERDEM*[1*], *Zeynep Idil ERZURUM CICEK*[1]

[1]Department of Industrial Engineering, Faculty of Engineering, Eskisehir Technical University, Eskisehir, Turkey.
zergul@eskisehir.edu.tr, bicmen@eskisehir.edu.tr, zierzurum@eskisehir.edu.tr

Received/Geliş Tarihi: 19.01.2023 Revision/Düzeltme Tarihi: 24.07.2023 doi: 10.5505/pajes.2023.39959
Accepted/Kabul Tarihi: 24.08.2023 Special Issue Article/Özel Sayı Makalesi

Abstract

The popularity of bike-sharing programs has increased the need for precise demand prediction techniques. In this work, the use of gradient-boosting techniques to forecast demand for bike-sharing systems is studied. The gradient boosting algorithms XGBoost, LightGBM, and CatBoost are used in this study to suggest an approach for predicting bike-sharing demand. Two real-world data sets were analyzed in this study, one for Konya and the other for Washington, D.C. Both datasets provide details about the day's particular characteristics and the weather. By using previous data to train a gradient-boosting model, we are able to make extremely precise predictions of future bike-sharing demand. CatBoost outperforms XGboost and LightGBM when all gradient boosting models are trained with the best hyperparameter sets.

Keywords: Bike-Sharing demand, Gradient boosting, Prediction, Machine learning.

Öz

Bisiklet paylaşım sistemlerinin artan popülaritesi, talebi doğru tahmin etme ihtiyacını artırmıştır. Bu çalışma, bisiklet paylaşım sistemlerinde talebi tahmin etmek için gradyan artırma yöntemlerinin kullanımını araştırmaktadır. Bu amaçla, XGBoost, LightGBM ve CatBoost gradyan artırma algoritmalarını kullanarak bisiklet paylaşım talebini tahmin etmek için bir yöntem önerilmektedir. Önerilen yöntem Konya ve Washington, D.C. olmak üzere iki gerçek dünya veri setine uygulanmıştır. Her iki veri setinde de hava koşulları ve günün belirli özellikleri gibi bilgiler yer almaktadır Geçmiş veriler üzerinde bir gradyan artırma modeli eğiterek, gelecekteki bisiklet paylaşımı talebine ilişkin son derece doğru tahminler yapılabilmektedir. Tüm gradyan artırma modelleri en iyi hiperparametre kümeleriyle eğitildiğinde; CatBoost, XGboost ve LightGBM'den daha iyi performans göstermiştir.

Anahtar kelimeler: Bisiklet paylaşım talebi, Gradian artırma, Tahminleme, Makine öğrenmesi.

1 Introduction

A service known as bike-sharing enables users to hire bicycles for brief periods of time, usually for a few hours or a day. Typically, bike-sharing programs have a network of stations where riders can hire and return their bikes. Users can rent a bike from any station in the network and return it to any other station. Shared bicycles are a practical and sustainable mode of transportation that can be used for errands, tourism, or fun. As it makes it simple for people to travel short distances or visit locations that are difficult to reach by public transportation, it is frequently utilized as a supplement to public transportation.

Several factors may affect the demand for rental bikes. These factors are weather, price, availability, public transit options, population density, bicycle infrastructure, cultural factors, and marketing efforts. The demand for bike-sharing may be higher on sunny, dry days and lower on wet or extremely hot days since people are more likely to utilize it when the weather is nice. Also, demand may be influenced by the price of bike rentals. People may be less likely to use the service if the price is too expensive. Another factor is availability. Lack of sufficient bikes at convenient locations could prevent individuals from using the service and related to that, people may be less likely to use bike-sharing if there are effective public transportation options nearby. In densely populated urban regions where individuals are more likely to utilize bicycles as a mode of transportation, there may be a stronger demand for bike-sharing. The fact that there are more bike rental companies in big and crowded cities confirms this situation. Another factor is ease of use. More individuals may use bike-sharing if there are designated bike lanes and other biking-friendly infrastructure in place. Biking may also be more widespread because it is more culturally accepted in some places while being less so in others.

Another aspect of bike-sharing that has benefits other than companies and users is its environmental benefit. Bike-sharing concepts help to achieve more sustainable mobility and reduce the congestion and environmental pollution brought on by motorized transportation [1]. Since bikes do not emit any emissions, sharing bikes can help cut down on air pollution, especially in cities where the large number of cars on the road can lead to poor air quality. Additionally, employing bike-sharing can help minimize the carbon impact of transportation because bikes don't emit greenhouse gases. By lowering the quantity of greenhouse gases like carbon dioxide that are emitted into the atmosphere, this can help ameliorate climate change. Utilizing bike-sharing can assist minimize traffic congestion, which can then help reduce air pollution and enhance people's quality of life overall. Bikes take up less space on the road than cars. Utilizing bike-sharing can help minimize noise pollution because bikes are much quieter than cars, especially in urban areas where there are often many cars on the road and noise levels can be very high. By promoting increased bike use as a means of transportation, bike-sharing

*Corresponding author/Yazışılan yazar

may also improve physical fitness [2] and lower the chance of developing chronic illnesses.

There are numerous bike-sharing businesses worldwide, but a few of the most well-known ones are as follows: Ofo, Lime, Mobike, JCDecaux, Citi Bike, Santander Cycles, Vélib, Nextbike, Call a Bike and BikeMi, etc. More than 500,000 bikes are shared through more than 500 bike-sharing programs globally [3]. This makes it crucial for these businesses to correctly predict the demand for bicycles. Predicting demand for bike sharing is essential for a variety of reasons. It aids bike-sharing firms in maintaining their stock: A bike-sharing firm can guarantee that there are enough bikes available at each station to satisfy the needs of their users if they can predict the demand for bikes with any degree of accuracy. By doing this, it may be possible to avoid circumstances in which there are not enough bikes accessible, which may irritate users and result in missed sales. It helps bike-sharing businesses in improving their operations: Bike-sharing firms can better understand how their bikes are being used and where they are most required by predicting the demand for bikes. For instance, shifting bikes from low-demand to high-demand regions or changing their pricing to account for fluctuations in demand, this can help them optimize their operations. It also helps bike-sharing firms in making future plans: Bike-sharing firms may make well-informed decisions regarding things like expansion, new station placements, and new services with the support of accurate demand prediction. And finally, it helps bike-sharing firms in pricing optimization: Bike-sharing companies can adjust their pricing to more accurately represent demand for their services by understanding the demand for bikes at various times of the day. They can increase their sales and profitability as a result.

The major issue with bike-sharing systems is that it is impossible to predict how many rental bikes will be needed at any particular moment. The motivation for this study is to be able to estimate the number of bikes required for a bike-sharing systems to meet this issue. Two real-world data sets, one for Konya and the other for Washington, D.C. were used to predict the demand by using different Gradient Boosting Methods. There are several advantages of Gradient Boosting Methods:

- Often provide highly accurate prediction,
- Able to capture complex nonlinear relationships and feature interactions,
- Because of regularization, less prone to overfitting
- Can handle mixed types of data,
- No need to preprocess such as handling missing data, data scaling, and feature selection.

Of course, there are also drawbacks in addition to these advantages. Gradient Boosting Methods have a long training time, especially on large-scale datasets, because of the computational complexity of the ensemble strategy. Hyperparameter optimization is very essential for Gradient Boosting Methods, but this process is time-consuming due to the computational complexity just mentioned. Thanks to advantages, in many studies Gradient Boosting Methods are used to predict different outcomes successfully. Crop yield [4], effectiveness and accuracy of hydrogen gas storage [5], middle-aged and elderly depression [6], groundwater salinization [7], and success [8] are the some of the topics that were predicted by these methods.

Bike-sharing models have complex nonlinear relations between weather conditions, time of day, day of the week, holidays, and demand [9]. For this reason, gradient boosting methods are very fitting for bike-sharing models. Just at this point, although there are various prediction studies in bike sharing models, according to our knowledge there is no study using gradient boosting algorithms that have gained popularity in recent years. In addition, there is no study in Turkey that predict bike sharing. We believe that this study will be an important addition to the literature.

The remainder of the paper is structured as follows. The related works about our study are included in Section 2. The Gradient Boosting Methods we employed in this work and the experimental setup are provided in Section 3 together with the data sets and their characteristics. The findings achieved using the aforementioned strategies are contrasted in Section 4 using various evolution metrics. Finally, Section 5 summarizes the study's findings.

2 Related works

The modern bike-sharing system makes for an intriguing research topic since the widespread use of bike-sharing for urban transportation. Numerous related studies are conducted to learn the characteristics and advancements about it. Sathishkumar, Park, and Cho [10] focused on hourly bike demand prediction for Seoul bike-sharing system. This study used Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall, the number of bikes rented hourly, and date as a feature set. Before the prediction step, they used Boruta algorithm to reduce the size of the feature set. Linear Regression, Gradient Boosting Machine, Support Vector Machine (Radial Basis Function Kernel), Boosted Trees, and Extreme Gradient Boosting Trees were applied for bike demand prediction. Wang and Kim [11] considered the station-level short-term prediction model, unlike the other studies. They employed RF, LSTM, and GRU algorithms for rental bike demand at three stations in Suzhou Youon Public Bicycle Systems, China. The dataset includes station ID, the number of available bikes, and the timestamp. The predictions made by the three algorithms were accurate and near. They concluded that RF performs better than others since RF computing resources are less. Sathishkumar and Cho [12] also studied on Seoul bike-sharing system to predict bike-sharing demand. They compared five regression models: CUBIST, Regularized Random Forest, Classification and Regression Trees, K-Nearest Neighbour, and Conditional Inference Tree. They concluded that the regression model developed with Cubist was more successful than other models for bike-sharing demand prediction. Chang et al. [13] projected daily rental bike demand with an artificial immune system and artificial neural network using the Washington, DC dataset. They split the dataset 2011 years data as training data and 2012 years data as testing data. The eight popular forecasting algorithms used for comparison were Alternating Model Trees, Gradient Boosted Regression Trees, Bagging Regression Trees, Decision Table, Additive Regression, Support Vector Regression, Conjunctive Rule, and Locally Weighted Naive Bayes. Gao and Lee [14] present a moment-based approach to predicting demand for bike-sharing systems using a fuzzy C-means-based genetic algorithm with a back propagation network. To evaluate the proposed model, they ignored holiday days and considered a 2-h period to be a one-moment. Using a deep learning approach, Dastjerdi and Morency [15] focused on predicting of the demand for shared bikes in Montreal over the next 15 minutes. Prior to prediction, they used the Louvain algorithm to divide up clusters of closely knit bike-sharing stations in a community detection phase. They used a hybrid CNN-LSTM-based learning architecture to carry out short-term travel demand prediction at the community level as a result. Pan et al. [16] proposed a deep LSTM model to predict bike-sharing demand collected by Citi Bike System from the bike stations in New York City and Jersey City. The considered dataset includes each trip's start time, end time, start station, and end station. Jiang [17] prepared a comprehensive survey about bike-sharing and deep learning. With a classification for the prediction problems and models, the most recent studies on deep learning-based utilization prediction for bike-sharing are covered in this survey. Both inside and outside of bike share systems, several applications based on predicted bike usage are considered. Li and Zheng [18] propose a Hierarchical Consistency Prediction model to predict citywide bike usage in the next period. In order to create the model an Adaptive Transition Constraint clustering algorithm, a Similarity-based efficient Gaussian Process Regressor and a General Least Square formulation is proposed. Real-world data is used to make experiments and present the effectiveness of their model.

In the literature, in general, prediction models have been created on real case data as well as Seoul and Washington data sets. Machine learning algorithms and deep learning algorithmsare used in bike sharing prediction models, which are mostly created as a regression problem. However, none of these studies compared XGBoost, LightGBM and CatBoost gradient boosting algorithms. To fill this gap, in this study, the prediction performances of these three algorithms on the Washington dataset were compared. Also, when the literature is examined, it is seen that a bike sharing prediction model has not been created in Turkey yet. Based on this, in this study, we created a data set for the city of Konya by adding new attributes to the bicycle demands shared by the municipality. And again, we also compared the methods on this dataset.

3 Materials and methods

In this section, we go into detail on the datasets and methods used in the study.

3.1 Data description

Two real-world data sets were used in this study, one for Konya and the other for Washington, D.C. The 16 features in the Washington dataset are listed in Table 1 on an hourly basis. The Capital bike-sharing Company's Washington, D.C. bike-sharing dataset [19] includes rental bicycles from 2011 to 2012.

Figure 1 visualizes the Pearson Correlations between these features as a heat map. According to this figure, atemp is highly correlated with temp. When features are highly correlated with each other, it can cause problems with the regression models, known as collinearity. We removed the atemp feature to avoid collinearity.

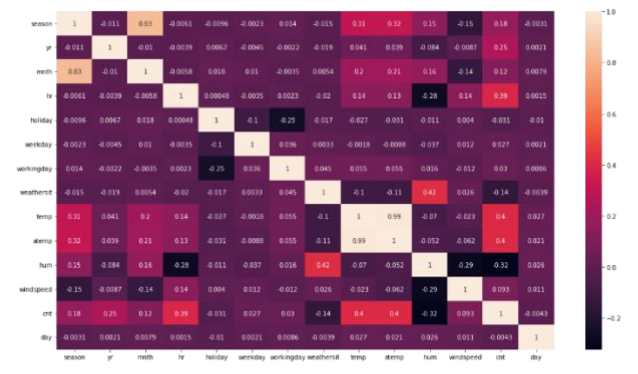

Figure 1. Correlation heatmap of features in the Washington dataset.

Table 1. Features and definitions in bike-sharing demand prediction.

Feature	Type	Description
Day	Categoric	Day of to month (1 to 31)
Season	Categoric	1. Springer, 2. Summer, 3. Fall, 4. Winter
Holiday	Categoric	0: non-holiday, 1: holiday
Weekday	Categoric	day of the week (0: Monday to 6: Sunday)
Workingday	Categoric	0: non-working day,1: workingday
Weathersit	Categoric	1. Clear, few clouds, partly cloudy or partly Cloudy; 2. Mist+cloudy, mist+broken clouds, mist+few clouds or Mist; 3. Light snow, light rain + thunderstorm+scattered clouds or Light rain+scattered clouds; 4. Heavy rain+ice pallets+Thunderstorm+mist, snow+fog
Temp	Numeric	Normalized temperature value in Celsius
Atemp	Numeric	Normalized feeling temperature in Celsius.
Hum	Numeric	Normalized humidity value
Windspeed	Numeric	Normalized wind speed value
Cnt	Numeric	count of total rental bikes
Yr	Categoric	0:2011, 1:2012
Mnth	Categoric	month (1 to 12)
Hr	Categoric	hour (1 to 12)

The hourly bicycle demands on a seasonal basis of Washington, D.C. dataset is given in Figure 2. According to this figure, there is a decrease in the demand for rental bikes during the winter months, and the demand for rental bikes increases during commuting hours.

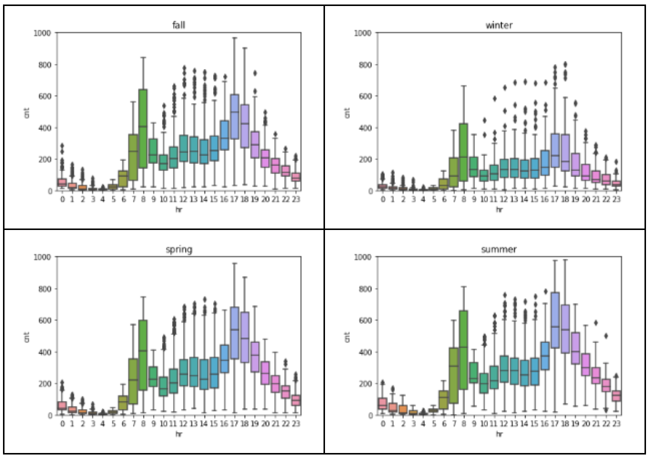

Figure 2. Hourly bicycle demands in Washington dataset.

The Konya data set was prepared by us using the data published by the Konya municipality [20]. Konya municipality has shared the information about the starting time of the rental, the end of the rental, the rental starting location, the rental ending location, the rental duration, the distance traveled, and the fair for each rental made in October, November, and December of 2021.

We summarized this dataset as hourly rental bike demand and added the weekday, working day, holiday, weather situation, temperature, wind speed, and humidity features.

To add weather-related features, we accessed the historical weather information of the relevant date of Konya with the Python Meteosat library [21]. We added the weekday, working day, and holiday attributes using the use google spreadsheet functions. Unlike the Washington data set, there is no season in the Konya data set, while the weather situation features take nine different values. This study apparently became the first to predict demand for bike-sharing models in Turkey by preparing and using the Konya dataset.

Correlation values for the Konya dataset are visualized in Figure 3. Even though there is a high correlation between the weekday and working day features, no action has been taken in order not to lose the information whether the relevant day is a weekend or not.

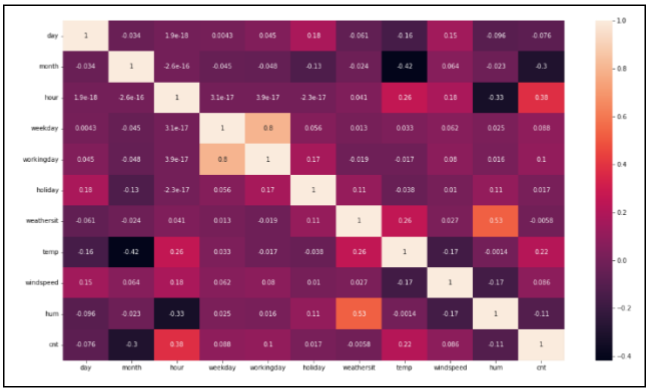

Figure 3. Correlation heatmap of features in the Konya dataset.

Figure 4 displays the hourly bicycle demand in the Konya dataset on a monthly basis. These numbers indicate that people in Konya rent bicycles between 10 am and 11 pm. Moreover, due to difficult weather conditions, bike demands declined in October and November. From the figure, it can be concluded that unlike the users in Washington, the users in Konya prefer to use the rental bikes during working hours, not for coming to work. This may be because the bike paths in Konya are less common than in Washington, and users don't prefer to rent a bicycle as a transportation vehicle in Konya.

3.2 Prediction methods

For regression and classification problems, gradient boosting is a machine learning technique that creates a prediction model in the form of a group of weak prediction models, often decision trees. Similar to earlier boosting methods, it builds the model incrementally. However, it generalizes such techniques by enabling the optimization of any differentiable loss function.

Natekin and Knoll [22] explained the main idea behind gradient boosting as to construct the new base learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The weak base learners in gradient boosting are typically decision trees. At each step, a new decision tree is trained to predict the residual error of the previous tree. The residual error is the difference between the predicted and actual values. The predictions of the new tree are then combined with those of the old tree to produce the final prediction.

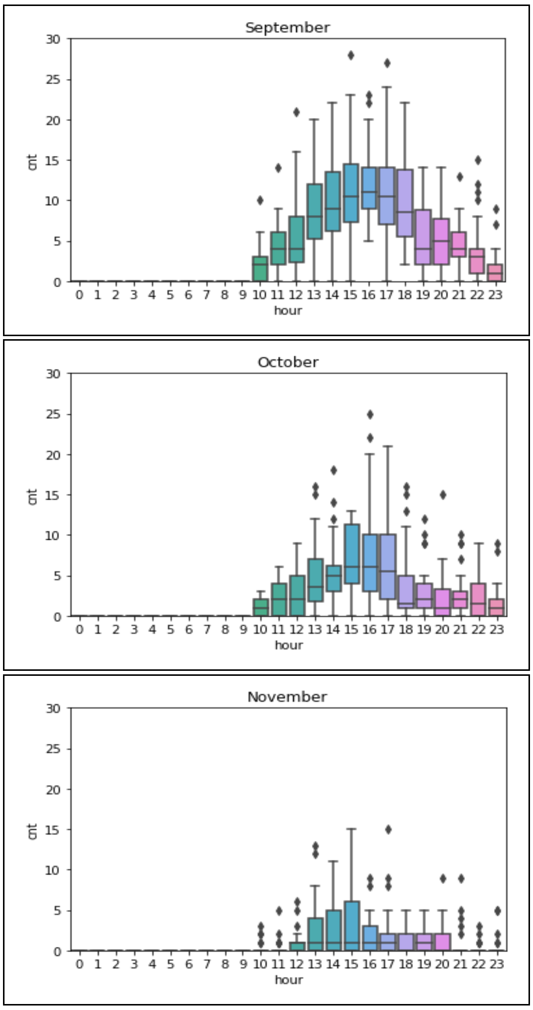

Figure 4. Hourly bicycle demands in the Konya dataset.

One of the main advantages of gradient boosting is that it can handle heterogeneous features, meaning that it can handle features that are on different scales and have different distributions. It is also resistant to overfitting and can handle high-dimensional data efficiently.

There are several gradient boosting algorithms, including Gradient Boosting, XGBoost, LightGBM, and CatBoost.

Let $(x, y)_{i=1}^{N}$ be a dataset where $x = (x_1, \ldots, x_d)$ denotes features and y corresponds to target. The objective of Gradient boosting is finding the functional dependence $x \to y$ with $\hat{F}(x)$ minimizing a specified loss function $\psi(y, f)$ [23].

Gradient Boosting calculated $F^*(x)$ as the weighted sum of m functions [23]:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x). \tag{1}$$

Where ρ_m denotes is the weight of m^{th} function $h_m(x)$. These functions serve as the prediction model and are the additive approximation of $F^*(x)$ built up iteratively. The first approximation is constant as given in Equation (2):

$$F_0(x) = arg\ arg \sum_{i=1}^{N} L(y_i, \alpha)\ . \tag{2}$$

Consecutive models are minimizing Equation (3):

$$(\rho_m h_m(x)) = arg\ arg \sum_{i=1}^{N} L\big(y_i, F_{m-1}(x_i) + \rho h(x_i)\big). \tag{3}$$

3.2.1 XGBoost

XGBoost (eXtreme Gradient Boosting), first described by Chen and Guestrin [24], is an open-source software library that provides a gradient boosting framework for C++, Python, R, Java, Julia, and other languages. Due to its versatility in applications and capacity to handle large datasets, it has become widely used in the machine learning field. One of XGBoost's primary characteristics is its high level of efficiency while handling large datasets and missing information.

Gradient boosting, a machine learning technique that combines several weak models to create a strong model, is the fundamental idea behind XGBoost (eXtreme Gradient Boosting). Decision trees are employed as the foundation learners in XGBoost. XGBoost generates a collection of decision trees iteratively while training, with each tree seeking to correct the errors of the one before it. The gradient descent algorithm, which minimizes the loss function by iteratively updating the model parameters in the direction that reduces the loss, is used to train the trees.

The efficiency and effectiveness of XGBoost are improved over traditional gradient-boosting algorithms by a number of factors. For instance, it employs a split finding technique that takes into account sparsity and is quicker and more memory-efficient than conventional algorithms.

The objective of XGBoost is to minimize the sum of the loss term (l) and the regularization term (Ω) [25]:

$$L_{xgb} = \sum_{i=1}^{N} L\big(y_i, F(x_i)\big) + \sum_{m=1}^{M} \Omega(h_m), \tag{4}$$

where Ω is regularization term penalizes the model complexity to avoid overfitting. The regularization term Ω is given in Equation (5):

$$\Omega(h_m) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{5}$$

Where T and w denote the number of leaves of the tree and output scores of the leaves respectively.

3.2.2 LightGBM

LightGBM, firstly introduced by Ke et al. [26], is a variant of gradient-boosting decision trees with Gradient-based One-Side Sampling and Exclusive Feature Bundling techniques.

Gradient-based one-side sampling (GOSS) is a technique used in the LightGBM algorithm to improve training efficiency and reduce overfitting. Instead of random sampling, GOSS samples data points based on their gradient values. Data points with higher absolute gradient values are more likely to be included in the training dataset, while data points with lower absolute gradient values are more likely to be excluded. This results in a training dataset that is more diverse and representative of the overall distribution of the data, which can improve the model's performance.

The LightGBM algorithm employs the Exclusive Feature Bundling (EFB) technique to boost the effectiveness and precision of tree-based models. It works by classifying features that have a strong correlation and generating a distinct tree for each group of features. The idea behind EFB is that the model can better represent the underlying relationships in the data by building separate trees for each collection of linked attributes. This can result in increased model performance, particularly when the data contains a large number of correlated features.

3.2.3 CatBoost

CatBoost [27] is an open-source gradient boosting library developed by Yandex. The main idea behind CatBoost is to build gradient boosting models that are specifically designed to handle categorical data. One of the key features of CatBoost is that it can automatically handle missing values in the data, which is a common problem in real-world datasets. It does this by creating a separate tree for each missing value, which allows the model to make more accurate predictions even when there are missing values in the data. It is claimed that Catboost algorithm outperforms gradient boosted decision trees (GBDTs) XGBoost, LightGBM and H2O [28].

The goal of CatBoost is to lessen the prediction shift that occurs over training [8]. This change happens as a result of the fact that gradient boosting computes both the gradients and the models that reduce those gradients using the same instances during training. In the context of the CatBoost, a random permutation is a rearrangement of the rows of a dataset that is performed randomly during the training process. This is done to improve the generalization performance of the model by reducing the potential for overfitting the training data.

During training, CatBoost builds a series of decision trees to make predictions on the target variable. Each tree is trained on a subset of the data, and the rows of the data are randomly permuted before being used to train each tree. This helps to ensure that each tree sees a different combination of rows, which can help to reduce overfitting and improve the overall performance of the model.

4 Experiments

4.1 Experimental setup

We measured prediction performance with four evaluation metrics: Mean Absolute Error (MAE), R-Squared (R^2), Root Mean Squared Error (RMSE), and RSMLE (Root Mean Squared Log Error). Equations for computing these evaluation metrics are shown in Table 2.

Table 2. Evaluation metrics.

Metric	Formula
MAE	$\frac{\sum_{i=1}^{n}[y_i - \hat{y_i}]}{n}$
R^2	$1 - \frac{\sum_i (y_i - \hat{y_i})^2}{\sum_i (y_i - \bar{y})^2}$
MSE	$\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y} - y_i\right)^2}$
RMSLE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log\, log\,(\hat{y_i}+1) - log\, log\,(y_i+1)\,)^2}$

In this table, y_i is the actual bike demand, $\hat{y_i}$ is the bike demand prediction, $\underline{y}$ is the mean demand of the samples, and n is the sample size. Better prediction performance is shown by lower values of MAE, RMSE, and RMSLE as well as greater values of R^2. The closer R^2 is to one, the model a good fit for data; the closer R^2 is to zero, the less good fit the model is. MAE is the average absolute difference between the predicted and the actual demand. RMSE penalizes large errors more acutely than MAE. Often used to forecast sales and inventory demands, the RMSLE penalizes the underestimation more severely than the overestimation. All reported evaluation metrics were computed with 10-fold cross-validation to generalize the performance of prediction models. Withal we used optimized hyper-parameters with the Grid Search method for each algorithm.

Python on the Google Colab Platform was used to implement the prediction algorithms. The CPU specification of the Google Colab is model 79, CPU Family 6, model name Intel(R) Xeon(R) CPU @ 2.20 GHz.

4.2 Results and discussions

The results of the prediction models are summarized in Table 3 for Washington city. In terms of the R^2 metric, which measures the proportion of the variance in the bike demand that is explained by the model, all three models show similar performance, with values ranging from 0.9547 for XGBoost to 0.9572 for CatBoost. The MSE, RMSE, and MAE metrics showed similar trends, with the CatBoost model having the lowest MSE (1405.7412), RMSE (37.4071), and MAE (23.3084) values among the three models. The MSE and RMSE values for the XGBoost model were higher than those for the CatBoost model but lower than those for the LightGBM model. The RMSLE value for the LightGBM model is 0.4510, which is lower than the RMSLE values for the XGBoost (0.5232) and CatBoost (0.5169) models. This indicates that the LightGBM model has a lower error in predicting bike-sharing demand in terms of underpredictions. It is worth noting that the difference in the RMSLE values between the three models was relatively small. So, we can conclude that the CatBoost model accurately predicted the bike demand with a relatively low error and the highest R^2 for Washington city.

Table 3. Average evaluation metrics of models for the Washington dataset.

Method	RMSLE	R^2	MSE	RMSE	MAE
XGBoost	0.5232	0.9547	1482.0372	38.4323	24.3303
LightGBM	**0.4510**	0.9505	1628.2098	40.3008	25.4060
CatBoost	0.5169	**0.9572**	**1405.7412**	**37.4071**	**23.3084**

A previous study by Xu et al. [29], which also utilized the Washington dataset, was used for comparing the performance of machine learning algorithms for bike sharing systems. In addition to the Random Forest, Gradient Boosting Decision Tree, and Artificial Neural Network models, additional hybrid models have been developed in this study. In the considered study, RMSLE values yielded 4.275, 11,432, and 13,057 for the RF, GBDT, and ANN models, respectively. It is seen that these values are considerably higher than the RMSLE values obtained for the gradient-boosting models proposed in our study. As a result, we can say that the gradient boosting models proposed for Washington City bike sharing demand prediction better capture the nonlinear relationships in the data.

The XGBoost model outperformed the other two gradient boosting models (LightGBM and CatBoost) in terms of the RMSLE metric, according to the data shown in Table 4 for Konya city, with a value of 0.2817. The RMSLE values for the LightGBM model are slightly higher but still within a similar range with XGBoost. The R^2 metrics of models indicate that the CatBoost and LightGBM models can better explain the variance in the bike demand compared to the XGBoost model. Besides, CatBoost had the highest R^2 value with 0.6564. The reason why the R^2 value obtained for Konya is lower than for Washington is that it contains only 3 months of data. As the number of days in the Konya data set increases, the forecast models will be more meaningful. To sum up, the CatBoost model had the lowest RMSE, MSE, and MAE values and the highest R^2 value for Konya city. Here, we can draw the conclusion that CatBoost is the best prediction model for rental bike demand prediction on both datasets.

Table 4. Average evaluation metrics of models for the Konya dataset.

Method	RMSLE	R^2	MSE	RMSE	MAE
XGBoost	**0.2817**	0.4983	9.2419	3.0266	1.5713
LightGBM	0.3118	0.6230	6.839	2.5927	1.4721
CatBoost	0.2889	**0.6564**	**6.3267**	**2.5109**	**1.3966**

To compare the methods in terms of execution time, the execution times with optimal parameters of each method are reported. The average training and prediction time of 10-fold for the examined datasets are displayed in Table 5. LightGBM was the fastest algorithm for both Konya and Washington applications according to training time. This result is in line with the findings of the [23],[30]. On the other hand, Catboost had the minimum prediction time of all prediction models for both cities. Training time is an important parameter in models that are retrained by periodically adding new data to the training set, while prediction time is critical in short-term and real-time prediction models. The prediction and training times of all three models are reasonable, given that hourly projections of bicycle demand are generated in this study. In this case, it will be more accurate to choose the best method according to the evaluation metrics. However, when the evaluation metrics of the models are the same, selection can be made according to the execution time.

Table 5. The average execution time of models.

	Training Time (in seconds)		
Dataset	XGBoost	LightGBM	CatBoost
Washington	0.2931	**0.0759**	1.7343
Konya	0.7855	**0.0501**	0.5056
	Prediction time (in seconds)		
Dataset	0.0018	0.0029	**0.0017**
Washington	0.0019	0.0021	**0.0012**
Konya	0.0018	0.0029	**0.0017**

A prediction interval is an estimation of the range within which, given the existing data, future observations are likely to fall. Prediction intervals are useful because they convey the forecasts' level of uncertainty. It is hard to assess the predictions' accuracy if we just provide point forecasts. However, it becomes obvious how much uncertainty is attached to each prediction if we additionally create prediction intervals. The predicted outputs are intervals that represent (with a given coverage probability) the most likely region defined by the upper and lower bounds of the interval to which the output of the uncertainties will belong [31].

In this study, in addition to point estimates, interval predictions at a 95% confidence level were also generated for the Washington dataset. 80% of the dataset is allocated as training and 20% as testing. After applying parameter optimization via Grid Search for each model; for the test dataset, the average widths of the prediction intervals obtained for the XGBoost, LightGBM, and CatBoost models were calculated as 16.75, 100.29, and 96.36, respectively. The width of a prediction interval represents the range within which the future observation is likely to fall. A narrower width indicates a more precise prediction, while a wider width suggests a larger range of uncertainty. With an average width of 96.36, the prediction intervals generated by CatBoost are wider compared to XGBoost but narrower than LightGBM. This implies that CatBoost provides a moderate level of uncertainty in its predictions. The LightGBM model has a higher level of uncertainty and variability in its predictions compared to the other two models. XGBoost stands out with the smallest average width, implying that the prediction intervals generated by XGBoost are relatively narrow. This indicates a higher level of precision and confidence in its predictions compared to CatBoost and LightGBM. The actual test data rates included in the prediction intervals are 22.27%, 79.748%, and 72.18% for XGBoost, LightGBM, and CatBoost, respectively. Considering the average width, XGBoost gave a more precise estimation range, but the actual test data it contains is quite low at 22.27%. LightGBM, with the highest average width, covered the most test data. The CatBoost model, on the other hand, performed relatively well in terms of both the average width and the amount of actual test data it contained. This supports the results obtained for point predictions.

Figure 5 visualizes the prediction intervals obtained with XGBoost, LightGBM, and CatBoost for a part of the test dataset, respectively.

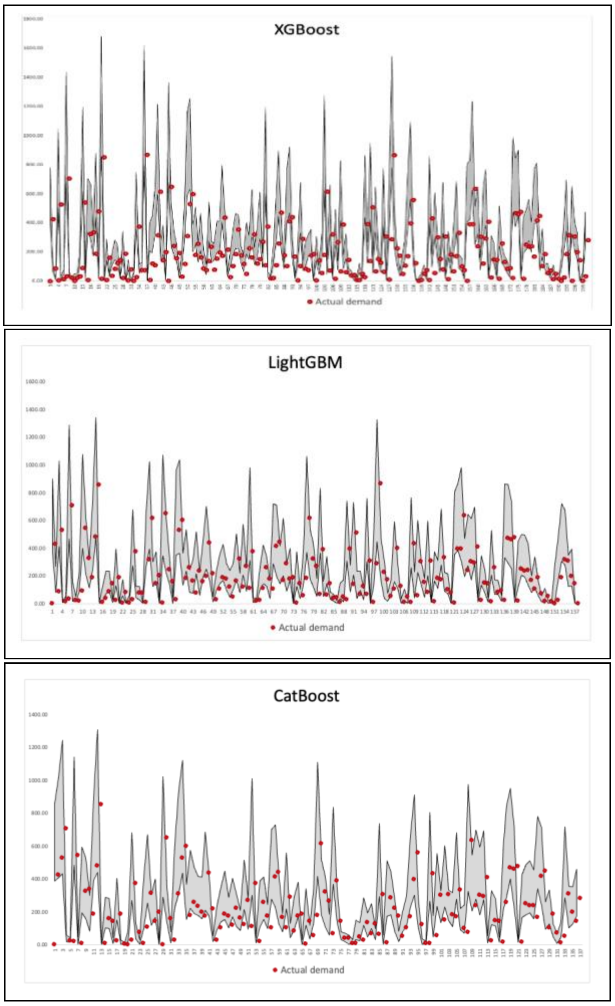

Figure 5. Prediction intervals for test dataset with XGBoost, LightGBM and CatBoost.

5 Conclusion and future works

In conclusion, this study applied gradient boosting methods, specifically XGBoost, LightGBM, and CatBoost, to predict bike-sharing demand. The results demonstrated that all three models achieved high performance. Comparing the three gradient boosting methods, we found that XGBoost and LightGBM showed similar results, while CatBoost had a slightly better performance. Besides, LightGBM and CatBoost had the minimum training and prediction times, respectively.

However, it should be noted that this study only considered a limited set of features and two specific bike-sharing systems. While the Konya dataset was compiled by us, the Washington dataset is openly accessible. The only public dataset we could access on bike-sharing demands in Turkey was for Konya. Nevertheless, Konya dataset only includes a period of 3 months, so one of the limitations of the study is scarcity of the Konya dataset. So, we carried out the first demand forecasting study for bicycle sharing in Turkey.

Further research could investigate the impact of additional features such as location of station, bike rental fee, bicycle infrastructure, cultural factors, and marketing efforts. Another topic for future studies is using a large-scale dataset for a comprehensive seasonality analysis in Turkey. Additionally, the model could be extended to other bike-sharing systems to evaluate its generalizability.

Overall, this study provides a promising approach for accurately forecasting bike-sharing demand using gradient boosting methods, which can be valuable for bike-sharing companies and city planners in making informed decisions. We have shown that using different gradient boosting methods can result in similar performance, and choosing the appropriate method is depend on the specific requirements and use case.

6 Acknowledgment

This study is supported by Eskisehir Technical University Scientific Research Projects Committee (ESTUBAP-22ADP274).

7 Author contribution statements

In the scope of this study, Zeliha ERGUL AYDIN made contributions to concept generation, literature review, data curation, drafting the article, and getting results in the context of this study. Banu ICMEN ERDEM contributed to the idea's development, the literature review, the drafting of the article, the evaluation of the results, and the spelling. Zeynep Idil ERZURUM CICEK participated in the idea's development, the literature review, the writing of the article, and getting results.

8 Ethics committee approval and conflict of interest statement

"There is no need to obtain permission from the ethics committee for the article prepared".

"There is no conflict of interest with any person / institution in the article prepared".

9 References

[1] Maggioni F, Cagnolari M, Bertazzi L, Wallace SW. "Stochastic optimization models for a bike-sharing problem with transshipment". *European Journal of Operational Research*, 276(1), 272-283, 2019.

[2] Tekouabou SCK. "Intelligent management of bike-sharing in smart cities using machine learning and Internet of Things". *Sustainable Cities and Society*, 67, 1-14, 2021.

[3] Otero I, Nieuwenhuijsen MJ, Rojas-Rueda D. "Health impacts of bike-sharing systems in Europe". *Environment international*, 115, 387-394, 2018.

[4] Huber F, Yushchenko A, Stratmann B, Steinhage V. "Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches". *Computers and Electronics in Agriculture,* 202, 1-11, 2022.

[5] Sripetdee T, Jitmitsumphan S, Chaimuengchuen T, Burana-amnuay M, Chinkanjanarot S, Jonglertjunya W, Ling TC, Phadungbut P. "Extreme gradient boosting machine for modeling hydrogen gas storage in carbon slit pores from molecular simulation data". *Energy Reports,* 8(16), 16-21, 2022.

[6] Zhang C, Chen X, Wang S, Hu J, Wang C, Liu X. "Using CatBoost Algorithm to Identify Middle-aged and Elderly Depression National Health and Nutrition Examination Survey 2011–2018". *Psychiatry Research,* 306(11), 1-8, 2021.

[7] Tran A, Tsujimura M, Thang H, Nguyen T, Binh D, Dang T, Doan QV, Bui D, Anh Ngoc T, Vo P, Thuc P, Pham TD. "Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta". *Vietnam. Ecological Indicators,* 127, 1-14, 2021.

[8] Jhaveri S, Khedkar I, Kantharia Y, Jaswal S. "Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns". *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 27-29 March 2019.

[9] Salaken SM, Hosen MA, Khosravi A, Nahavandi S. "Forecasting bike sharing demand using fuzzy ınference mechanism*". 22nd International Conference, ICONIP 2015,* İstanbul, Turkey, 9-12 November 2015.

[10] Sathishkumar VE, Park J, Cho Y. "Using data mining techniques for bike-sharing demand prediction in metropolitan city". *Computer Communications*, 153, 353-366, 2020.

[11] Wang B, Kim I. "Short-term prediction for bike-sharing service using machine learning". *Transportation Research Procedia*, 34, 171-178, 2018.

[12] Sathishkumar VE, Cho Y." A rule-based model for seoul bike-sharing demand prediction using weather data". *European Journal of Remote Sensing*, 53(1), 166-183, 2020.

[13] Chang PC, Wu J, Xu Y, Zhang M, Lu X. "Bike-sharing demand prediction using artificial immune system and artificial neural network". *Soft Computing,* 23, 613-626, 2019.

[14] Gao X, Lee G. "Moment-based Rental Prediction for Bicycle-sharing Transportation Systems Using a Hybrid Genetic Algorithm and Machine Learning". *Computers & Industrial Engineering*, 128, 60-69, 2018.

[15] Dastjerdi A, Morency C. "Bike-sharing demand prediction at community level under COVID-19 using deep learning". *Sensors*, 22(3), 1-18, 2022.

[16] Pan Y, Zheng R, Zhang J, Yao X. "Predicting bike-sharing demand using recurrent neural networks". *Procedia Computer Science*, 147, 562-566, 2019.

[17] Jiang W. "Bike-Sharing usage prediction with deep learning: a survey". *Neural Comput & Applic,* 34, 15369-15385, 2022.

[18] Li Y, Zheng Z. "Citywide bike usage prediction in a bike-sharing system". *IEEE Transactions on Knowledge and Data Engineering*, 32(6), 1079-1091, 2020.

[19] Fanaee-T H, Gama J. "Event labeling combining ensemble detectors and background knowledge". *Prog Artif Intell*, 2, 113-127, 2014.

[20] Konya Açık Veri Portalı. "Paylaşımlı Kiralık Bisiklet Kullanım Verileri". https://acikveri.konya.bel.tr (06.12.2022).

[21] Meteostat. "The Weather's Record Keeper". https://meteostat.net/en/ (06.12.2022).

[22] Natekin A, Knoll A. "Gradient boosting machines, a tutorial". *Frontiers in Neurorobotics*, 7, 1-21, 2013.

[23] Bentejac C, Csörgö A, Martinez-Munoz G. "A comparative analysis of gradient boosting algorithms". *Artificial Intelligence Review*, 54, 1937-1967, 2021.

[24] Chen T, Guestrin C. "XGBoost: A scalable tree boosting system". *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California, USA, 13-17 August 2016.

[25] Ergul Aydin Z, Kamisli Ozturk Z. "Performance analysis of XGBoost classifier with missing data". *Manchester Journal of Artificial Intelligence and Applied Sciences*, 2(2), 166-170, 2021.

[26] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. "LightGBM: A highly efficient gradient boosting decision tree". *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach California, USA, 4-9 December 2017.

[27] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. "Catboost: unbiased boosting with categorical features". *Advances in Neural İnformation Processing Systems*, 31, 6638–6648, 2018.

[28] Dorogush AV, Ershov V, Gulin A. "CatBoost: Gradient Boosting with Categorical Features Support". *arXiv preprint*, 2018. https://doi.org/10.48550/arXiv.1810.11363.

[29] Xu, T, Han, G, Qi, X, Du, J, Lin, C, Shu, L. "A hybrid machine learning model for demand prediction of edge-computing-based bike-sharing system under ınternet of things". *IEEE Internet of Things Journal,* 7(8), 7345-7356, 2020.

[30] Apaydın, M, Yumuş M, Değirmenci, A, Karal Ö. "Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* 28 (5), 737-747, 2022.

[31] Marín, LG, Cruz, N, Sáez, D, Núñez, A. "Prediction interval methodology based on fuzzy numbers and its extension to fuzzy systems and neural networks". *Expert Systems with Applications*, 119, 128-141, 2019.