A Study on the Prediction of Bike Availability Using Machine Learning

Seung Yong Lee[1], Youngkeun Choi[2], Yangmi Lim[3*]

[1]College of Liberal Arts, Namseoul University, 91 Daehak-ro, Seonghwan-eup, Seobuk-gu, Cheonan-si, Chungcheongnam-do, Republic of Korea, leesky@nsu.ac.kr
[2]Division of Business Administration, College of Business, Sangmyung University Seoul, Republic of Korea, penking1@smu.ac.kr
[3]IT media department, Duksung Woman's University, 33 Samyang-ro 144-gil, Dobong-gu, Seoul, Republic of Korea, yosimi@duksung.ac.kr

The bike-sharing service has offered significant conveniences to urban residents, effectively augmenting the public transportation infrastructure. To achieve this, both operators and users need to forecast the precise number of available Bikes at docking stations. This engineering white paper is primarily centered on short-term predictions about docking station usage in London, England. The primary objective of this paper is to formulate a dependable Bike count prediction model utilizing machine learning methodologies. To accomplish this, various weather and day-related features serve as predictors, and a range of techniques spanning from linear regression to root mean square error evaluation were applied to construct the prediction model. The findings indicate that, concerning weather conditions, higher temperatures stimulate increased Bike usage, whereas heightened humidity and wind speed tend to decrease Bike usage. Regarding day types, Bike usage experiences a reduction on holidays and weekends.

Keywords: Sharing economy, Bike sharing, Machine learning.

1. Introduction

The swift advancement in mobile and Internet technologies presents numerous opportunities within the realm of the sharing economy. Bike sharing has emerged as a prominent facet of the sharing economy, gaining substantial popularity in recent times [1]. Bike-sharing services have witnessed widespread adoption across numerous major global urban centers. Bikes, serving as a vital mode of transportation, have offered considerable advantages for short-distance travel, contributing to the reduction of greenhouse gas emissions and promoting healthy exercise practices among cyclists. Shared mobility services are typically proffered through two modalities: docked bike sharing and undocked bike sharing. While the latter, the dockless bike-sharing approach, has engendered several challenges, including issues related to Bike abandonment, pedestrian path obstructions, and inadequate management, the former, docked bike-sharing programs, have witnessed a surge in popularity.

Efficient management of these Bikes poses a non-trivial challenge. The capacity constraints of docking stations impose limitations on the availability of Bikes for rental and return, leading to instances of station depletion or saturation. In response, manual repositioning of Bikes via dedicated trucks occurs according to a predetermined schedule (e.g., 8 a.m. and 2 p.m., as observed in cities like Suzhou, China). This approach aims to mitigate the consequences of an uneven distribution of Bikes within docking stations, thereby optimizing resource utilization. Recognizing the imperative for enhanced Bike management, this white paper seeks to harness advanced predictive models for the enhancement of bike-sharing system management. Short-term traffic forecasting represents a crucial facet of Intelligent Transportation Systems (ITS) research, offering the potential to predict traffic metrics such as flow rates, delays, speeds, and travel times. Methodologies employed in this domain encompass statistical techniques, nonlinear theories, and machine learning methodologies [2]. Wang conducted a comparative analysis of various machine learning models for local Bike rental demand forecasting, revealing that neural network-based and tree-based models exhibited superior predictive accuracy [3].

In recent years, a surge in research activity has underscored the pivotal role of machine learning in delivering state-of-the-art outcomes for short-term forecasting challenges, with a heightened focus on its capacity to adapt to data convergence issues [4]. Over the preceding decades, a multitude of investigations has sought to advance Bike-related predictive modeling employing diverse methodologies. For instance, studies such as Singhvi have introduced a logarithmic regression model to anticipate the Bike usage patterns during morning peak hours in New York City, incorporating variables like taxi utilization, weather conditions, and spatial factors to enhance predictive accuracy [5].

Hence, this investigation endeavors to delineate the key determinants for forecasting bike counts at individual stations through the application of machine learning techniques. While prior research predominantly gravitates towards macroscopic predictions within the domain of bike-sharing systems, there exists a compelling rationale to delve into station-level forecasting. The study scrutinizes a dataset comprising 17,414 instances sourced from Transport for London. Linear regression analysis, coupled with split validation facilitated by RapidMiner, is employed to explore the influential factors affecting bike counts in relation to weather conditions and day types. The revelations derived from this inquiry hold significant implications for the refinement of bike-sharing operational systems.

2. Related Study

Regression count modeling represents a prevalent technique for forecasting the real-time Bike inventory. In Austria, Rudloff and Lackner proposed a demand model for Bikes and return boxes, employing Poisson, negative binomial (NB), and hurdles models to predict the Bike count within a specified timeframe [6]. These models incorporated weather data, with particular emphasis on temperature and precipitation, alongside neighboring station information as regression variables. The findings indicated that the hurdles model outperformed the other two approaches. Wang et al. employed logarithmic linear and NB regression models, considering 13 independent variables encompassing socioeconomic, demographic, and geographic factors to forecast bike availability [7]. The results underscored the significance of all 13 highly fitting variables in both models. In a related study, Rixey utilized multivariate linear regression to pinpoint pivotal elements influencing bike-sharing ridership and subsequently estimated system ridership [8]. The research identified demographics, environmental conditions, and access to an extensive station network as key determinants within the multivariate linear regression model.

Given the vast scale and intricacies inherent in Bike Sharing System (BSS) data, the discourse within the engineering domain has prominently revolved around clustering analysis and visualization technologies. Numerous researchers have diligently employed these approaches to gain insightful perspectives by illuminating trends via visualization techniques [9]. To illustrate, Froehlich et al. harnessed Barcelona's 13th-week Bike station usage data to scrutinize BSS patterns. Their inquiry delved into human behavior, geographic considerations, and temporal relationships, subsequently endeavoring to prognosticate forthcoming Bike station utilization. Notably, temporal and spatial patterns were meticulously examined, revealing a degree of interdependence between stations. Docking stations were clustered utilizing the available Bike data, revealing that proximate stations exhibited close associations and grouping tendencies.

Kaltenbrunner et al. similarly embarked on an enhancement quest for Barcelona's BSS, leveraging docking station data [10]. Their investigation entailed discerning patterns of temporal and geographical mobility, facilitating the identification of imbalances within the BSS. Furthermore, they adeptly employed time series analysis techniques to forecast Bike counts at specific stations and times.

Vogel et al. ventured into the realm of bike-sharing data analysis with localized datasets, extracting patterns of Bike activity [11]. Employing cluster analysis techniques, Bike stations were categorized based on pickup and return activities, employing algorithms such as k-means, expectation-maximization, and sequential information bottleneck. The temporal activity of stations was instrumental in clustering the stations into five distinct groups, providing hourly average pickup and return statistics for each cluster. The linkage between these clusters and geographic information unveiled a propensity for adjacent stations to fall within the same cluster.

Feng and Hillston et al. introduced a novel moment-based predictive model reliant on time-based speed, drawing upon the Publication Continuous Time Markov Chain (PCTMC) to estimate Bike availability [12].

Gast and Massonnet et al. employed queuing theory-based time-mixed BSS models to predict Ronal probabilities [13]. Furthermore, novel metrics were proposed for model evaluation, diverging from the conventional root-mean-square error (RMSE). Fricker and Gast investigated the impact of station capacity on peer BSS performance, deploying probabilistic models and flow approximation techniques [14]. This model affords insights into optimizing station sizes to minimize imbalance issues.

Several recent studies within the engineering domain have embraced time series methodologies for the prediction of Bike counts at stations [15]. While these methodologies have exhibited remarkable performance in elucidating historical trends and forecasting future counts, they are not without limitations. For instance, Kaltenbrunner employed the Automatic Regression Transfer Average (ARMA) model to predict Bike availability at stations [10]. However, the ARMA model assumes a static mean and variance for observations over time, a premise incongruent with the dynamic nature of Bike station data.

Froehlich et al. introduced four models encompassing the last value, historical mean, historical trend, and Bayesian network [16]. They demonstrated that the Bayesian network model yielded the least predictive errors. However, the Bayesian network model did not directly provide the precise Bike count; instead, it categorized Bike availability into discrete intervals (e.g., 25%, 50%, 75%, and 100%). The algorithm selected one of these categories to describe bike availability.

Yoon et al. introduced a spatiotemporal prediction system employing the Automatic Regression Movement Integration Average (ARIMA) model to address the shortcomings inherent in ARMA models, particularly their inability to handle non-stationary data [15]. This novel approach incorporates seasonal trends and ambient information into the modeling process. The model's evaluation was conducted using a modest three-week dataset, revealing marginal enhancements in predictive performance compared to ARMA (with errors of 3.47 Bikes / Station for ARIMA vs. 3.50 Bikes / Station for ARMA). Nevertheless, it's important to note that ARIMA is inherently a static model, characterized by fixed coefficients and predictions confined to constant intervals. Additionally, ARIMA is regarded as a complex and challenging model to interpret.

3. Methodology

3.1 Dataset

We conducted an analysis of a sample comprising 17,414 data points, sourced from Transport for London. The data was gathered from three distinct sources: 1. Https://cycling.data.tfl.gov.uk/ 'Contains OS data © Crown copyright and database rights 2016' and Geomni UK Map data © and database rights [2019] 'Powered by TfL Open Data'. 2. Https://freemeteo.com - weather data 3. https://www.gov.uk/bank-holidays From 1/1/2015 to 31/12/2016. The cycling dataset was organized based on 'Start time,' representing the count of new bike shares grouped by hour, with long-duration shares excluded from the count. Drawing upon existing literature related to bike-sharing and considering three categories of bike count determinants, we examined the impact of five variables categorized into two groups: weather types and day types. These variables are enumerated and their definitions provided in Table 1.

Table 1: The variables in each category

Categories	Variables	Definitions
Bike counts	cnt	the count of a new bike shares
Weather counts	t1	real temperature in C
	hum	humidity in percentage
	windspeed	wind speed in km/h
Day types	isholiday	1 holiday / 0 non holiday
	isweekend	1 if the day is weekend

3.2 Analysis method

This study uses RapidMiner tool to conduct linear regression and machine learning analyses for the prediction of bike counts.

3.2.1 Linear regression

Linear regression analysis, a statistical method, is employed for quantitative prediction. This technique assesses the degree of correlation between variables. While classification methods are typically applied for predicting categorical labels, regression methodologies are tailored for predicting continuous values. In the context of linear regression analysis, the objective is to establish a linear association between a quantitative dependent variable, denoted as Y, and K independent variables. The fundamental model for multiple linear regression analysis is articulated as follows.

$$Y_i = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + ... + \beta_K \cdot X_K + \varepsilon_i \qquad (1)$$

Building upon the aforementioned discussion, we utilize a multivariable linear regression model to incorporate facility-related factors, proximity to the nearest landmark, and the popularity of said landmark as indicators of house prices. However, owing to inherent uncertainties in real-world scenarios, we must account for this uncertainty by integrating the variability, which can be simulated through the introduction of noise denoted as ΔP, during the price prediction process.

$$\mathrm{P} - \mathrm{A} = \begin{bmatrix} \mathrm{fi} \\ \mathrm{d} \\ \mathrm{q} \end{bmatrix} \mid \mathrm{distribution}(\Delta \mathrm{P}) \qquad (2)$$

In this model, we incorporate facility embedding represented by 'fi' denoting property types, 'd' signifying room types, and 'q' representing bed types.

The model comprises regression coefficients β0, β1, β2, ..., βK, with εi signifying the superordinate error inherent in measuring the dependent variable 'y.' Developing a regression model entails the determination of these regression coefficients, β0, β1, β2, ..., βK.

Typically, a sample drawn from the broader population is employed in the analysis. Given our lack of knowledge regarding the precise population regression coefficients (β0, β1, β2, ..., βK), we resort to the least squares method (OLS) to estimate them based on the provided sample data. This approach yields estimated coefficients denoted as β*0, β*1, β*2, ..., β*K, which minimize the sum of squared deviations between the model's predicted values, Y*, and the actual values, Y.

$$Y^*_i = \beta^*_0 + \beta^*_1 \cdot X_1 + \beta^*_2 \cdot X_2 + ... + \beta^*_K \cdot X_K + \varepsilon_i \qquad (3)$$

OLS can be employed when the dependent variable conforms to the following assumptions.

1. The dependent variable adheres to a standard normal distribution.

2. There exists a linear relationship between the independent and dependent variables.

3. Each observation maintains independence from others.

4. The variability of the dependent variable's values remains consistent across different values

of the independent variable, a condition referred to as homoscedasticity.

When these aforementioned assumptions hold true, the resultant predicted value is an unbiased estimate, and it yields the minimum mean square error when compared to other unbiased estimates. However, when the regression model is applied for prediction, and the first assumption is not met, even if the dependent variable follows an arbitrary distribution, the predictive estimates can still yield highly accurate results. This occurs due to the data mining approach's practice of using separate cases for training the model and validating its performance.

3.2.2 Data mining models

To thrive in an increasingly competitive marketplace, numerous companies are turning to data mining techniques for price prediction analysis. Achieving effective customer acquisition necessitates the development of a more precise and efficient rental price prediction model. Statistical and data mining techniques have been harnessed to construct these prediction models. Data mining techniques are instrumental in uncovering noteworthy patterns or relationships within data, enabling the prediction or classification of behavior through model fitting based on available data.

In scenarios where machine learning necessitates the separation of learning and testing datasets, the test dataset must adhere to the following criteria: First, it should be formatted consistently with the training dataset. Second, it must not overlap with the training dataset. Third, both datasets must exhibit data consistency. However, the creation of a test dataset that satisfies these requirements can be challenging. In the field of data mining, various validation frameworks leveraging a single dataset have been developed to address this challenge.

This study leverages the Split Validation operator provided by RapidMiner to facilitate this process. This operator partitions the input dataset into training and test datasets, enabling performance evaluation. In our investigation, we opt for relative segmentation among the operator's parameter options, utilizing 70% of the input data for the learning dataset.

3.2.3 Performance evaluation

Performance assessment leverages training data to evaluate the effectiveness of the generated model. Performance metrics encompass both technical performance measures and heuristic measures. The technical performance measures employed in this study elucidate model performance by generating models from training data, processing test data to construct models, and subsequently comparing the class labels of original verification cases with predicted class labels. The assessment of technical performance is stratified into supervised and unsupervised learning. The supervised learning methodologies employed in this study encompass classification and regression. All data utilized for learning and testing possess original class values. Performance evaluation is conducted by scrutinizing and analyzing the alignment between the original class values and the prediction outcomes. RapidMiner offers performance indicators for common classification problems, and among these, this study employs the root mean square error (RMSE) for evaluation.

Root Mean Square Error (RMSE) serves as a frequently employed metric when assessing the disparity between estimated values or those predicted by a model and real-world observations. It is well-suited for conveying precision. Each disparity is also referred to as a residual, and the root mean square deviation amalgamates these residuals into a singular measure. The root mean square deviation of the estimator concerning the estimate is defined as the square root of the mean square error.

4. Results

4.1 Linear regression

The outcomes of the linear regression analysis are as follows.

Table 2: The results of linear regression

Category	Variable	Coefficient	p-value
Weather types (3)	t1	44.191	0.000
	hum	-27.552	0.000
	windspeed	-2.802	0.003
Day types (2)	isholoday	-287.548	0.000
	isweekend	-212.722	0.000

The analysis unveiled the significance of all variables at the $p < 0.05$ level. Furthermore, all three room types exhibited significance at the $p < 0.05$ level. Notably, temperature demonstrated a positive correlation with bike counts. Conversely, humidity, wind speed, holidays, and weekends displayed negative associations, leading to a decrease in bike counts.

4.2 Performance evaluation

In numerous prediction scenarios, this study aims to impose greater penalties on predicted values that deviate further from the actual values, in contrast to those that exhibit proximity to the actual values. To achieve this objective, the study employs the mean of squared error values, referred to as the root mean squared error (RMSE). The RMSE formula is as follows:

$$\text{RMSE} = \text{root}\{(e_1\text{^}2 + e_2\text{^}2 + \dots + e_n\text{^}2) \ / \ n \ \} \quad (4)$$

Here, 'n' denotes the count of rows in the test set. While this formula may appear complex initially, it essentially encapsulates the following process:

- Taking the difference between each predicted value and the actual value (or error),
- Squaring this difference (square),
- Taking the mean of all the squared differences (mean), and
- Taking the square root of that mean (root).

Consequently, when reading from the bottom to the top, we arrive at the term 'root mean squared error.' When computing the RMSE value for the predictions generated by this study on the test set, the resulting RMSE is approximately 9.31. One advantageous feature of RMSE is its unit consistency with the predicted value, as it involves squaring and then taking the square root. This characteristic facilitates a straightforward interpretation of the error's magnitude in the same units as the prediction.

5. Conclusions

Bike sharing systems have witnessed global expansion and established themselves as a dependable mode of transportation. However, they grapple with logistical challenges, with some stations experiencing bike shortages while others become inundated. Addressing this issue necessitates the proactive prediction of bike demand, providing an opportunity for both cyclists and operating agencies to contribute to the solution. Cyclists can leverage predicted demand to plan and adjust their routes, while bike-sharing system (BSS) managers can strategically redistribute bikes using service trucks to balance station capacities. This research centers on short-term traffic forecasting, specifically predicting the availability of bikes at shared-bike stations, employing machine learning techniques. It employs a linear regression model to forecast bike counts at stations within London's bike-sharing systems. This study elucidates the determinants of bike counts, with a focus on mobility sharing. Concerning weather conditions, it is observed that higher temperatures correlate with increased Bike usage, whereas elevated humidity and wind speed are associated with reduced Bike usage. These findings underscore the influence of favorable weather conditions on heightened shared Bike usage. Additionally, for different day types, Bike usage decreases during holidays and weekends, underscoring the prevalence of shared Bikes as a weekday transportation choice for work or school-related commuting.

This research proactively delves into the determinants of bike counts within the sharing economy domain, utilizing an official dataset sourced from Transport for London. The findings offer a comprehensive insight into the factors influencing bike counts within this emerging business paradigm. The primary objective of this paper is to formulate an optimal predictive model for bike counts, leveraging a restricted set of features encompassing weather and day types. Employing machine learning techniques, including linear regression and feature importance analyses, the study endeavors to attain superior predictive performance, as measured by RMSE. This methodological approach reveals latent patterns in bike counts. In addition to its methodological contributions, this study enriches the sharing economy literature by presenting a unified model that synthesizes the determinants of bike counts within this unconventional bike-sharing system. On a practical level, the research furnishes valuable insights for stakeholders, including bike-sharing providers, enabling them to assess their market positioning and enhance profitability.

Nonetheless, we acknowledge a notable limitation in this research. Our economic modeling approach is employed to scrutinize the dataset and discern the relationships between various factors and bike counts. However, we do not encompass any social or psychological variables that may influence bike counts. Consequently, it is imperative to undertake qualitative research to delve into the underlying motivations guiding users' decisions regarding bike usage. Future avenues of research in this study may encompass (i) investigating alternative feature selection methodologies, such as random forest feature importance, (ii) conducting further experimentation with neural network architectures, and (iii) procuring additional training data from other hospitality services, such as vrbo, to enhance the performance of the k-means clustering with ridge regression model.

References

1. Jaeseung R., (2022) Exploring the prospect of shared mobility solution focused on car sharing, (Journal of Digital Art Engineering & Multimedia, Vol.9, No.1, pp. 71-84 DOI : 10.29056/jdaem.2022.03.07, (Journal of Digital Art
2. Gang, C., Shouhui, W. & Xiaobo, X. (2016). Review of spatio-temporal models for short-term traffic forecasting. In 2016 IEEE International Conference on Intelligent Transportation Engineering (ICITE), (2016): 8-12.
3. [3] Wang, W. (2016). Forecasting Bike Rental Demand Using New York Citi Bike Data. Technological University Dublin.
4. Mohanned H. A., Mohanned, E., & Hesham, A. R. (2019). Dynamic linear models to predict bike availability in a bike sharing system. International Journal of Sustainable Transportation 14.3, (2019): 232-242.
5. Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., O'Mahony, E., Shmoys, D. B. & Woodard, D. B. (2015). Predicting Bike Usage for New York City's Bike Sharing System. In AAAI Workshop: Computational Sustainability. (2015): 1-5.
6. Rudloff, C. & Lackner, B. (2014). Modeling demand for bikesharing systems: neighboring stations as source for demand and reason for structural breaks. Transportation Research Record 2430.1, (2014): 1-11.
7. Wang, X., Lindsey, G., Schoner, J. E. & Harrison, A. (2015). Modeling bike share station activity: effects of nearby businesses and jobs on trips to and from stations. Journal of Urban Planning and Development 142.1, (2015): 1-21
8. Rixey, R. (2013). Station-level forecasting of bikesharing ridership: station network effects in three US systems. Transportation Research Record 2387.1, (2013): 46-55.
9. Froehlich, J., Neumann, J. & Oliver, N. (2009). Sensing and Predicting the Pulse of the City through Shared Bicycling. International Joint Conference on Artificial Intelligence 9, (2009): 1420-1426.
10. 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, (2010): 455-466.
11. Vogel, P., Greiser T. & Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. Procedia-Social and Behavioral Sciences 20, (2011): 514-523.
12. Feng, C., Hillston J. & Reijsbergen, D. (2017). Moment-based availability prediction for bike-sharing systems. Performance Evaluation 117, (2017): 58-74.
13. Gast, N., Massonnet, G. Reijsbergen D. & Tribastone, M. (2015). Probabilistic forecasts of bike-sharing systems for journey planning. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, (2015): 703-712.
14. Bacem Sameta,b,∗, Florent Couffina, Marc Zolghadria, Maher Barkallahb, Mohamed, Haddarb, Model reduction for studying a Bike Sharing System as a closed queuing network, 8th Swedish Production Symposium, SPS 2018, 16-18 May 2018, Stockholm, Sweden, DOI:10.1016/j.promfg.2018.06.055
15. Yoon, J. W., Pinelli, F. & Calabrese, F. (2012). Cityride: a predictive bike sharing journey advisor, In 2012 IEEE 13th international conference on mobile data management, (2012): 306-311.
16. Jon Froehlich, Joachim Neumann, Nuria Oliver, et al. Sensing and predicting the pulse of the city through shared bicycling. In IJCAI, volume 9, pages 1420–1426, 2009.