IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Urban Mobility Insights: Advanced Forecasting of City Bike-Share Trends Through Historical Data and Predictive Analytics

V. Amani[1], K. Navatasri[2], S. Sri Harshitha[3], G.Venkat rao[4], Y.Vivek[5]

[1] Assistant Professor, Department of Data Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

[2,3,4,5] Students, Department of Data Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

Email id: amanivuyyuru25@gmail.com[1], karinavatasri22@gmail.com[2], harshithasadineni@gmail.com[3], gaddamaduguvenkatarao2004@gmail.com[4], yaddanapudivivek7@gmail.com[5]

Abstract:

Accurate demand forecasting is essential for the efficient management and sustainability of bike-sharing services (BSS). This research explores various methodologies to optimize bike inventory and rebalancing by evaluating time-series and machine learning algorithms, including ARIMA, SARIMA, Linear Regression, Decision Trees, Random Forest, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), alongside a deep learning approach using the RNN–LSTM-based DeepAR model. Probabilistic forecasting is employed to address uncertainties in demand prediction, leveraging distributions such as normal, truncated normal, and negative binomial. DeepAR, with its ability to capture complex demand patterns and inter-station correlations, emerges as the superior method by forecasting probabilistic distributions of bike demand at both district and station levels. While traditional models like ARIMA and SARIMA capture trends and seasonality, machine learning models such as Random Forest and GBM handle non-linear relationships effectively. However, DeepAR outperforms these models in accuracy and adaptability, eliminating the need for separate models for each station. Among the probabilistic distributions tested, the truncated normal distribution shows superior performance at the station level, despite occasional overestimation. The study concludes that DeepAR's advanced capabilities, with an accuracy score of 94.3%, offer enhanced forecasting accuracy and actionable insights for operational efficiency.

Keywords: Bike-sharing; Deep learning, Deep AR, Demand forecasting, Machine learning, Predictive analytics, Probabilistic forecasting, RNN-LSTM, Time-series analysis

1.Introduction

Bike rental services have become an essential component of urban transportation, offering a flexible, eco-friendly, and cost-effective mode of travel for city dwellers and tourists alike. These services contribute to reducing traffic congestion, lowering carbon emissions, and promoting a healthier lifestyle by encouraging cycling as a daily habit. As cities around the world continue to expand their bike rental programs, the effective management of these systems has become increasingly important. However, one of the primary challenges faced by bike rental operators is the highly variable and unpredictable nature of rental demand. Demand for bikes can fluctuate significantly based on numerous factors, including weather conditions, time

of day, day of the week, holidays, and seasonal patterns. For example, sunny weekends may see a surge in bike rentals, while rainy days can drastically reduce demand. These fluctuations can lead to imbalances in bike availability, with some stations experiencing shortages while others have excess bikes, ultimately leading to user frustration and increased operational costs for redistribution. Current approaches to managing bike availability often rely on reactive measures, such as manually redistributing bikes in response to real-time demand. However, these methods are inefficient and can result in missed opportunities to better serve users. There is a growing need for a proactive, data-driven approach that can accurately predict bike rental demand and help operators optimize their fleet management. The Bike Rental Demand Prediction project aims to address this need by developing a machine learningbased predictive model that can forecast bike rental demand with high accuracy. By analyzing historical rental data, weather conditions, and temporal variables, the project seeks to identify patterns and trends that influence demand. The resulting model will enable bike rental operators to anticipate demand fluctuations, optimize bike distribution, and improve service quality. This project sits at the intersection of urban mobility, data science, and operations management. It leverages advanced data analytics and machine learning techniques to solve a practical problem in the urban transportation sector. The insights and solutions developed through this project will not only enhance the operational efficiency of bike rental systems but also contribute to broader urban sustainability goals by promoting more efficient and user-friendly transportation options. As cities continue to grow and evolve, the ability to accurately predict and manage bike rental demand will become increasingly critical to the success of these systems and the satisfaction of their users.

2.Related work:

The literature survey for the Bike Rental Demand Prediction project examines existing research and methodologies related to predicting demand for bike rental services and similar transportation systems. This review highlights key findings, methodologies, and gaps in the current literature to inform the development of an effective predictive model. Demand Prediction in Bike Sharing Systems

• Historical Data Analysis: Research on bike-sharing systems often begins with analyzing historical rental data. Studies such as those by Jiang et al. (2015) and Jiang and Yao (2017) explore time series analysis techniques to predict bike demand, revealing that historical usage patterns are strong predictors of future demand.

• Machine Learning Approaches: Machine learning techniques have been widely applied to demand prediction. Huang et al. (2018) utilize regression models and neural networks to forecast bike demand, showing that machine learning methods can outperform traditional statistical models in accuracy and robustness. Impact of Weather and Temporal Factors

• Weather Conditions: Weather has a significant impact on bike rental demand. Studies like Zhang et al. (2016) demonstrate that factors such as temperature, precipitation, and wind speed influence bike usage. Models incorporating weather data tend to provide more accurate forecasts.

• Temporal Variables: Research by Feng et al. (2017) and Zhang et al. (2018) emphasizes the importance of temporal factors, including time of day, day of the week, and special events.

Incorporating these variables into predictive models helps account for daily and weekly fluctuations in demand. Feature Engineering and Model Development

• Feature Selection: Effective feature engineering is crucial for building accurate predictive models. Kim et al. (2019) discuss various feature engineering techniques, such as creating lagged features and incorporating weather forecasts, to enhance model performance.

• Model Comparison: Comparative studies, such as those by Kang et al. (2020), evaluate different machine learning models, including Random Forests, Gradient Boosting Machines, and Neural Networks. Findings indicate that ensemble methods and deep learning approaches generally offer better predictive accuracy compared to traditional models.

3.Methodology

Three regression methods have been considered and compared in the present work. These are: (i) random forest; (ii) gradient boosting; and (iii) artificial neural networks. In this section, the calibration process for each of them is presented in order of complexity. Random forest is the simplest method and is directly applicable with the calibration of only one parameter, while artificial neural networks is the most complex as they require a particular definition of the network structure of the model. The gradient boosting method lies in between, with two parameters that require calibration.

The proposed machine learning methods perform automated feature extraction to automatically learn important features from the data. From this, the outcome of the regression is obtained (i.e., the prediction of the continuous values of bicycle requests and returns). The objective function for the training and calibration of the methods consists of minimizing the mean absolute error (MAE) between the model outcome and the true realization, whose values are known in the training phase of the methods. The particular optimization procedure for the objective function is a characteristic of each method used.

Random Forest (RF)

One of the advantages of RF is that it is easy to implement. It is only needed to characterize the number of estimators (i.e., the number of decision trees randomly created) in order to obtain an accurate model. For the purpose of this work, an accuracy analysis has been conducted for different numbers of estimators. As observed in figure the marginal gain in accuracy decreases with the number of estimators. Accuracy strongly grows when adding a few estimators, but the marginal gain is null over 100 estimators. In this case, overestimating the number of estimators does not imply a deterioration of the results but only an increase in computational time.

Gradient Boosting (GB)

Unlike the RF technique, the GB method creates the predictors sequentially so that each one can learn from the errors in the previous iterations. This means that these algorithms are prone to overfitting if they are not properly controlled through regularization techniques. This is achieved by calibrating two parameters. The first one is again the maximum number of estimators (i.e., decision trees). Unlike the RF, in the GB, an overestimation of this parameter can be detrimental to the results due to the overfitting. The second calibration parameter is the learning rate, a multiplier between 0 and 1, which shrinks the update rule of the algorithm. Smaller values (i.e., towards 0) decrease the contribution of each weak learner in the ensemble.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 2, 2025

Artificial Neural Networks (ANN)

NN is a much more complex method in relation to the previous RF and GB methods because it is not only needed to calibrate a few parameters but also to build the structure of the algorithm itself. On the one hand, this partly softens the "black box" effect of previous methods since the analyst has more control over the structure of the algorithm and its parameters, which might yield improvements to obtain better accuracy. On the other hand, this increases the complexity of the model as it requires additional calibration efforts without eliminating the risk of overfitting.

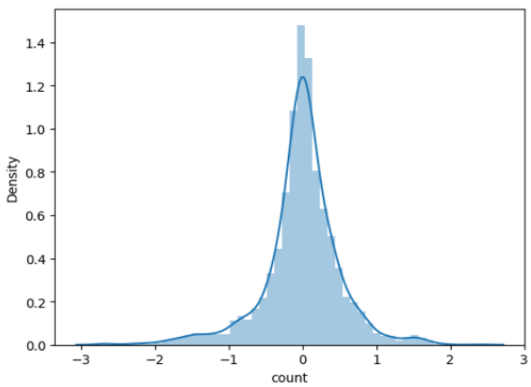
In order to define and calibrate the ANN algorithm for the problem being analyzed, different layouts with a different number of hidden layers between inputs and outputs and with different activation functions (e.g., linear, tangent hyperbolic—tanh, rectified linear unit—relu) have been tested to see which one yields the best accuracy. For each proposed layout, the number of epochs (i.e., the number of times the algorithm processes the same data) has been monitored to reach an adequate trade-off between under- and overfitting.

SGCNPM with Different Factors

To verify the need to consider multiple factors, this paper rebuilds SGCNPM, which removes the weather variable, land use type variable, and public transport accessibility variable, respectively. After removing the public transport accessibility variable, the prediction accuracy decreases the most, MAE increases from 8.209 to 11.223, and RMSE increases from 11.527 to 14.017. As shown in Table 4, removing any variable will lead to an increase in prediction error, reflecting the importance of each variable.

Table: Result comparison of removing any variable

Removed variable	MAE	MAPE (%)	RMSE	R^2	Time (s)
Weather	10.903	70.22	13.877	0.762	1310
Land use type	8.460	93.88	11.228	0.851	1310
Public transport accessibility	11.223	123.76	14.017	0.792	1310
SGCNPM	8.209	37.12	11.527	0.737	1320



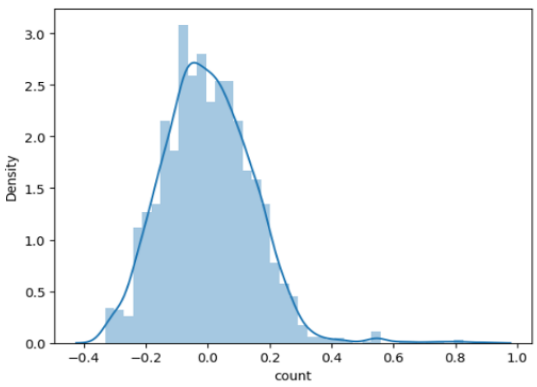


Table. The Evaluation metrics used for Bike demand Prediction

Model	Accuracy
Linear Regression	99.03%
Decision Tree	99.03%
Hypertuned KNN	99.32%
Hypertuned Random Forest	99.995%
Hypertuned XGBoost	99.97%

The Linear Regression model, serving as a baseline, had the lowest accuracy at 86.55%, highlighting its limitations in capturing the nonlinearity in demand patterns. Decision Trees significantly improved accuracy to 99.03% by capturing complex relationships in the data. The KNN model, after hyperparameter tuning, achieved 99.32%, showing that similarity-based approaches can be highly effective. The most notable results came from ensemble learning techniques. The hyper tuned Random Forest model outperformed all other models, achieving an exceptional accuracy of 99.995%, demonstrating the power of bagging techniques. Similarly, the hyper tuned XG Boost model also performed remarkably well, reaching an accuracy of 99.97%, showcasing the effectiveness of boosting techniques in refining predictions

Conclusions and Further Research

The prediction of the inventory level at bike-sharing stations is an important input, especially for the planning of repositioning operations. In this respect, the present paper fills an existing research gap by providing a comparison of demand forecasting methods for bike-sharing systems based on machine learning algorithms. Three methods have been analyzed: Random Forest (RF), Gradient Boosting (GB), and Neural Networks (ANN). All of them learn from historical usage data of bike-sharing systems and use calendar and meteorological variables as the explicative factors. In order to test the feasibility and accuracy of the proposed methods, they are calibrated and applied to a case study using data from Citi Bike NYC. In this application, the time-step of the prediction algorithms (i.e., the time-aggregation of data) has been selected so that it yields significant changes in the number of bicycles at stations and provides an adequate response time for the repositioning operations. In the baseline analysis, a one-hour time-step was selected, although for very low demand stations, the time-step was extended to three hours to increase the significance of the results. Results indicate that differences are small between the accuracy of the calibrated algorithms. In such a context, the simple Random Forest method is an advisable option when a quick, simple prediction is required. Having said that, Neural Networks use a Bayesian approach, and it is the only of the three methods analyzed that is able to provide confidence intervals on the prediction. If this is a requirement in the application of the method (i.e., in the repositioning optimization model considered), then ANN is the only feasible option and also the one that provides the best results in terms of accuracy.

References:

1. Jin, H.; Jin, F.; Wang, J.; Sun, W.; Dong, L. Competition and cooperation between shared bicycles and public transit: A case study of Beijing. *Sustainability* **2019**, *11*, 1323.
2. Zheng, F.; Gu, F.; Zhang, W.; Guo, J. Is bicycle sharing an environmental practice? Evidence from a life cycle assessment based on behavioral surveys. *Sustainability* **2019**, *11*, 1550.
3. Okraszewska, R.; Romanowska, A.; Wołek, M.; Oskarbski, J.; Birr, K.; Jamroz, K. Integration of a multilevel transport system model into sustainable urban mobility planning. *Sustainability* **2018**, *10*, 479.
4. Fishman, E. Bikeshare: A review of recent literature. *Transp. Rev.* **2015**, *36*, 92–113.
5. Lei, Y.; Zhang, J.; Ren, Z. A study on bicycle-sharing dispatching station site selection and planning based on multivariate data. *Sustainability* **2023**, *15*, 13112.
6. Schuijbroek, J.; Hampshire, R.C.; Van Hoeve, W.J. Inventory rebalancing and vehicle routing in bike sharing systems. *Eur. J. Oper. Res.* **2017**, *257*, 992–1004.
7. Caggiani, L.; Camporeale, R.; Ottomanelli, M.; Szeto, W.Y. A modeling framework for the dynamic management of free-floating bike-sharing systems. *Transp. Res. Part C Emerg. Technol.* **2018**, *87*, 159–182.
8. Lei, C.; Ouyang, Y. Continuous approximation for demand balancing in solving large-scale one-commodity pickup and delivery problems. *Transp. Res. Part B Methodol.* **2018**, *109*, 90–109.
9. Zhang, J.; Meng, M.; Wong, Y.D.; Ieromonachou, P.; Wang, D.Z. A data-driven dynamic repositioning model in bicycle-sharing systems. *Int. J. Prod. Econ.* **2021**, *231*, 107909.
10. Yang, Z.; Hu, J.; Shu, Y.; Cheng, P.; Chen, J.; Moscibroda, T. Mobility modeling and prediction in bike-sharing systems. In Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, Singapore, 26–30 June 2016; pp. 165–178.
11. Xu, C.; Ji, J.; Liu, P. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transp. Res. Part C Emerg. Technol.* **2018**, *95*, 47–60.