Predicting the Future Bike Shares: A Model Accuracy Analysis

Jan Skleničkа, Sebastian Černý, Vojtěch Chalupa, Andrea Matějková*

February 2, 2025

Abstract

Bike-sharing systems play an important role in sustainable urban mobility and represent a convenient and environmentally friendly transportation alternative. However, managing the availability of shared bikes requires accurate demand forecasting. This study evaluates the predictive performance of four regression models - XGBoost, Linear Elastic Network, Poisson regression, and Random Forest regression - using a London bike-sharing dataset from TfL Open Data covering the period from 2015 to 2017. The dataset contains hourly records of bicycle use along with weather and time variables. The models were evaluated based on the root mean square error (RMSE), with results showing that XGBoost outperforms the other methods and achieves the lowest RMSE. Optimized Random Forest regression also performed competitively, while Linear Elastic Net and Poisson regression showed significantly higher error rates, indicating their limited ability to capture complex, non-linear relationships. These findings help to determine which model is best for forecasting the demand for bike shares. The study underscores the importance of advanced modeling techniques for improving future urban transportation planning and optimizing resource allocation in bike-sharing systems.

*The authors used ChatGPT in order to expand vocabulary and writing style and for assistance with LaTeX. After using this tool, the authors reviewed and edited the content as necessary and takes full responsibility for the content of the publication.

1 Introduction

In recent years, bike-sharing systems have become a key part of sustainable urban mobility, offering an environmentally friendly alternative to traditional means of transport. Whether they are used by city residents or tourists visiting the sights, shared bikes are a great way to get around the city quickly and without traffic jams.

London is one of the cities with the most developed bike-sharing system, thanks in particular to the Santander Cycles scheme. In 2010, the scheme launched with 6,000 bikes, but now has more than double that number, reflecting its significant expansion and growing importance within London's urban mobility, which provides a flexible and affordable way to get around (BBC News, 2024). However, the growth in popularity of these services brings challenges associated with effectively managing bicycle availability and optimizing infrastructure, which presents the scope for reliable forecasting models to help with future demand estimation.

As the popularity of these services grows, so does the variety of tools ensuring efficient predictions. There are a range of statistical models for such demand forecasting, from the simplest regressions to complex machine learning approaches, and it is therefore crucial to evaluate their performance in a real-world environment. By systematically comparing different methodologies, we aim to determine the most effective approach to forecasting bike-share usage, and thus provide valuable insights for future planning.

Therefore, the aim of this study is to analyze and compare the performance of four regression models - Linear Elastic Net Regression, Poisson Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). For the analysis, we use the London bike-sharing database provided by the TfL Open Data dataset covering the period from the beginning of 2015 to the beginning of 2017, which includes information on the number of borrowings, weather conditions, time of day and daytime working patterns. We use Root Mean Square Error (RMSE) to assess their accuracy and determine the best-fitting model to evaluate historical bike share data.

Our analysis shows that the XGBoost model performs the best, showing the lowest RMSE and strong predictive ability. Its capability to capture non-linear relationships and interactions between variables makes it particularly suitable for predicting demand for shared bikes. These results suggest that gradient boosting methods can significantly improve future prediction accuracy and provide valuable insights for planning bike shares.

2 Data description

2.1 Dataset

For the scope of this project, the London bike-sharing dataset provided by TfL Open Data[1] was utilized. This dataset contains hourly time-series data on the number of bike shares in London during a period from 4.1.2015 to 3.1.2017, with a total of 17414 observations. It

[1]Powered by TfL Open Data. Contains OS data © Crown copyright and database rights 2016 and Geomni UK Map data © and database rights [2019].

includes data on bike-sharing usage, with weather conditions and calendar-based indicators. The data is structured with a timestamp variable, representing the date and time, which allows for data grouping.

The main target variable, cnt, represents the number of new bike shares recorded at a given time. The rest of the variables consists of various weather and seasonality-related factors, which are considered to influence the frequency of bike rentals. Their list, together with an explanation of their meaning, is presented in Table 1.

Variable	**Description**
timestamp	Timestamp field representing the date and time for grouping the data.
cnt	The count of new bike shares recorded at a given timestamp.
t1	Real temperature in °C.
t2	"Feels like" temperature in °C.
hum	Humidity in percentage.
wind_speed	Wind speed in km/h.
weather_code	Coded representation of weather conditions (details below).
is_holiday	Dummy denoting holidays, 1 = Holiday, 0 = Non-holiday.
is_weekend	Dummy denoting weekend days, 1 = Weekend, 0 = Weekday.
season	Meteorological season: 0 = Spring, 1 = Summer, 2 = Fall, 3 = Winter.

Table 1. Variable Descriptions

The *weather_code* categories provide a classification of different weather conditions observed at the time of data collection. A code of 1 represents clear or mostly clear weather, although it may include minor occurrences of haze, fog, or patches of fog in the vicinity. 2 corresponds to scattered or few clouds, 3 indicates broken clouds, while fully cloudy conditions are coded under 4. Precipitation events are captured by several categories: 7 signifies rain, light rain showers, or drizzle, whereas 10 represents rain accompanied by a thunderstorm. Finally, 26 classifies snowfall, while extreme low temperature conditions leading to freezing fog are indicated by 94.

The entire dataset provides a well-structured basis for analyzing the relationship between bike-sharing demand and weather and time-related factors. This makes the data suitable for predictive modeling and comparative analysis across different forecasting approaches. Plots illustrating distributions of the original variables are provided in Figure 1.

2.2 Data preprocessing

As a first adjustment to the data, the *humidity* was divided by 100 in order to convert it to the 0-1 scale. Additionally, the *hour* variable was extracted from the *timestamp* to account for the effect of time of day. After this, to prepare the dataset for modeling, a structured preprocessing pipeline was implemented. This pipeline focused on handling our categorical

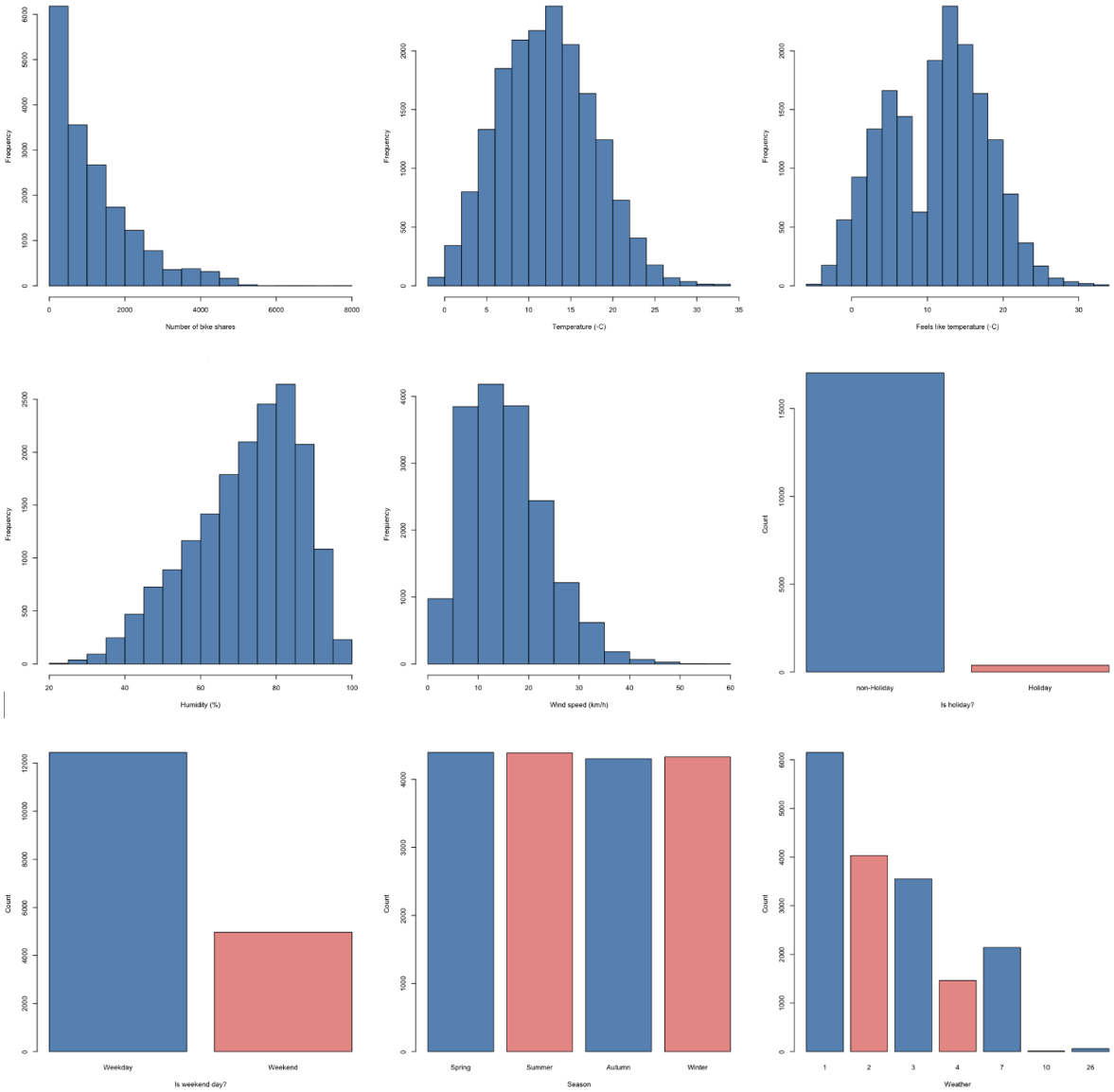

Figure 1. Plots representing distributions of the original variables

and numerical variables appropriately, applying transformations to optimize the data for machine learning models.

To facilitate their use in machine learning models, categorical variables, including *season* and *weather_code*, were converted into factors and one-hot encoded, creating separate binary columns for each category. Then, these were centered and scaled, ensuring a mean of 0 and variance of 1, making them suitable for modeling. As a result, the categorical variables no longer retained their original values, instead, they got transformed into continuous numerical values, which allowed for improved model performance while preserving the categorical distinctions.

All numerical variables, except *humidity*, i.e. temperature (*t1*, *t2*) and *wind_speed*, were

normalized, centered, and scaled to improve model stability by preventing any variable from dominating it. The *humidity* variable was left unnormalized, as it is already a percentage-based variable (ranging from 0 to 1 after rescaling), making additional transformation unnecessary. To capture non-linear patterns in bike demand, we transformed the *hour* variable using natural splines, allowing the model to better capture the fluctuations in usage throughout the day.

The pipeline was then trained and applied to the dataset, ensuring consistent transformations across all variables. This resulted in a well-structured dataset with standardized numerical features. The preprocessing approach ensured that our data is structured in a way that maximizes the effectiveness of predictive models. Lastly, to account for non-linear patterns, squared features of selected numerical variables were included, creating variables *t1_sq*, *t2_sq*, and *wind_speed_sq*.

The final dataset was divided into training and test sets, such that the part used for training the models contains 80% of all observations.

3 Methodology

This section describes the methodology used to analyze and compare models for predicting bike shares demand. We introduce the four selected methods and outline the steps taken to evaluate their predictive performance.

3.1 Elastic Net Regression - Linear model

Elastic Net Regression is an approach valuable for its ability to balance feature selection and multicollinearity handling. In this first case, it represents a regularized linear model that combines Ridge regression (L2 penalty) and Lasso regression (L1 penalty)(Friedman *et al.*, 2010). Elastic Net Regression can be represented by the following equation:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right).$$

The Elastic Net model requires tuning of λ, expressing the regularization strength, and α, a mixing parameter that determines the balance between the L1 and L2 penalties. When $\alpha = 1$, the model reduces to Lasso. α equal to 0, reduces Elastic Net to ridge regression. These hyperparameters were optimized using cross-validation[2], selecting values that minimized the prediction error on the validation set. Only the "best" hyperparameters found are used in the subsequent prediction and evaluation of the model performance.

[2]In our analysis, we use k-Fold Cross-Validation (method = "cv", number = 5), in other words, splitting the dataset into 5 subsets (folds). The model is then trained on 4 (k-1) folds and validated on the remaining one. This is repeated 5 times (k-times) and averaged.

3.2 Poisson Regresion

Given the count-based nature of bike shares, another approach employed is Poisson regression, as it represents the most widely used model for count data (Pesta, n.d.). Similarly to the previous linear model, a regularized version of Poisson regression was chosen, that combines L1 and L2 penalties. Taking this into account, we get the following equation describing this method:

$$\hat{\beta} = \arg\min_{\beta} \left(-\sum_{i=1}^{n} [y_i \log \hat{y}_i - \hat{y}_i] + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right),$$

which means the ordinary least squares loss from the linear model gets substituted for the Poisson log-likelihood. Again, the hyperparameters λ and α were fine-tuned through cross-validation to optimize the balance and strength of penalties. This allowed us to find the "best" possible Poisson regression model.

3.3 Random Forest Regression

As next, we employed Random Forest regression, a powerful ensemble learning method based on decision trees. Unlike models such as Elastic Net or Poisson regression, Random Forest is a non-parametric model able to capture complex, non-linear relationships even without requiring explicit feature transformations. Random Forest works by generating multiple decision trees during training and aggregating their predictions to improve accuracy and robustness. The model is trained using bootstrap aggregation (bagging), where each tree is built on a randomly sampled subset of the training data. The final prediction is obtained by averaging the outputs of all individual trees in the forest, reducing variance, and preventing overfitting.

The best values for hyperparameters, such as the number of trees or minimum node size, were obtained using cross-validation to optimize the performance of the Random Forest Regression.

3.4 XGBoost model

Lastly, Extreme Gradient Boosting represents another type of ensemble supervised algorithm. It is a powerful machine learning algorithm that uses gradient boosted decision trees to optimize prediction performance (Chen & Guestrin, 2016). Unlike traditional decision tree-based models such as Random Forest, which builds trees independently, the construction of trees in XGBoost is parallelized, meaning that each other tree is trained to correct the residual errors of the previous trees.

XGBoost minimizes a differentiable loss function using gradient descent. This method also utilizes regularization, making it more robust against overfitting. The objective function thus consists of the loss function and regularization, and its general form could be represented by the following equation:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(m)}) + \sum_{m=1}^{M} \Omega(f_m),$$

where the first part includes the loss function and the second part stands for regularization, with M being the total number of boosting iterations. Similarly to the other approaches, the optimal parameters for the XGBoost model were found in the preceding grid search using cross-validation, in order to make it as efficient as possible.

After defining all the methods and finding all the desired optimal parameters for each of them, we trained the models on the training dataset and used them for prediction on the test set. Their prediction performance is then analyzed and compared between the individual approaches.

4 Results

Root mean square error (RMSE) metric was chosen to compare the performance of the individual models. It is a commonly used metric to evaluate model accuracy, specifically measuring how well the model‘s predictions align with the actual values. It quantifies the average magnitude of prediction errors, giving more weight to larger errors due to squaring.

The formula for calculating RMSE is thus defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

The Table 2 below reports the individual values of Root Mean Squared Error (RMSE) obtained for the four models used to assess the demand for bike sharing in London between 2015 and 2017.

Model	**Test RMSE**
XGBoost	232.1232
Linear Elastic Net Regression	726.5146
Poisson Regression	694.8062
Optimized Random Forest Regression	242.2423

Table 2. Root Mean Squared Error (RMSE) for different predictive models

The results show that the XGBoost model achieved the lowest RMSE (232.1252), indicating that it can be considered the most accurate method of forecasting the bike-sharing demand among the models evaluated. The optimized Random Forest regression comes close with an RMSE of 242.2423. This suggests that the optimization process has effectively improved the model's performance.

In contrast, both the Linear Elastic Net regression (RMSE = 726.5146) and the Poisson regression (RMSE = 694.8062) show significantly higher errors, suggesting that these Elastic Net regression models are less suitable for capturing complex relationships in the given data set. The significantly higher RMSE values for these models suggest that nonlinear approaches such as Extreme Gradient Boosting and Random Forest are more effective for this prediction task. Overall, the results highlight the superiority of ensemble-based methods over traditional

regression models for predicting bike shares, with XGBoost showing the best generalization ability.

5 Discussion and conclusion

As discussed in the Results section, our findings suggest that ensemble-based machine learning methods, in our case XGBoost and Random Forest, over-perform methods based on traditional regressions. These conclusions are consistent with previous research on the performance of predictive models. For example, both Moulaei *et al.* (2022) and Subudhi *et al.* (2021) came up with a similar result, i.e., they showed better performance of ensemble methods such as Random Forest, while predicting mortality in Covid-19. The accuracy of forecasting with Random Forest models was also confirmed in the comparative analysis of Osisanwo *et al.* (2017).

Although we find that the XGBoost model along with the Random Forest regression best predicts the demand for bike sharing, which seems to align with other literature on this topic, some limitations should be noted. For example, the dataset covers only two years (2015-2017), which may not fully capture long-term trends, changes in user behavior, or the transformation of the city. Extending our data by several years or comparing these results with conclusions from other investigations using an updated dataset could improve the robustness of the findings. Verification with more recent data could show the true validity of this analysis and confirm whether its conclusions can continue to hold over time. At the same time, additional machine learning methods could be included in the comparison to see if any of them will further outperform the four methods we have chosen.

To conclude, this paper focused on conducting an analysis comparing predictive performance of four selected models: Linear Elastic Net regression, Poisson regression, Random Forest, and Extreme Gradient Boosting. For this purpose, a London bike-sharing dataset from TfL Open Data was utilized covering the period from 2015 to 2017. Upon optimizing the performance of all the methods examined, the empirical findings demonstrate that ensemble machine learning methods such as XGBoost and Random Forest outperform the other approaches to forecasting the demand for bike shares.

References

BBC News (2024): "Santander cycles: London's bike hire scheme sees record growth." Accessed: 31 Jan. 2025.

Chen, T. & C. Guestrin (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.

Friedman, J., T. Hastie, & R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software* **33(1)**: p. 1.

Moulaei, K., M. Shanbehzadeh, Z. Mohammadi-Taghiabad, & H. Kazemi-Arpanahi (2022): "Comparing machine learning algorithms for predicting covid-19 mortality." *BMC medical informatics and decision making* **22(1)**: p. 2.

Osisanwo, F., J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi *et al.* (2017): "Supervised machine learning algorithms: classification and comparison." *International Journal of Computer Trends and Technology (IJCTT)* **48(3)**: pp. 128–138.

Pesta, M. (n.d.): "Poisson regression." Faculty of Mathematics and Physics, Charles University (MFF UK). Accessed: January 30, 2025.

Subudhi, S., A. Verma, A. B. Patel, C. C. Hardin, M. J. Khandekar, H. Lee, D. McEvoy, T. Stylianopoulos, L. L. Munn, S. Dutta *et al.* (2021): "Comparing machine learning algorithms for predicting icu admission and mortality in covid-19." *NPJ digital medicine* **4(1)**: p. 87.