# Predicting Bike Demand Using Linear Regression

[1]Md. Mobeen, [2]K. Sri Harshitha

*[1] Assistant Professor, CSE Department, NEC., Gudur*

*[2]CSE, NEC., Gudur*

***Abstract:*** *This project focuses on accurately predicting bike demand in urban bike-sharing systems using linear regression, aiding in optimal inventory management and customer satisfaction. Historical bike rental data is collected and pre-processed, including factors like temperature, humidity, wind speed, day of the week, and time of day. A linear regression model is trained using scikit-learn in Python, with performance evaluated through Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The results highlight the model's effectiveness in forecasting bike demand, providing valuable insights for operators to optimize inventory and enhance service quality. This study supports informed decision-making to improve the efficiency and reliability of bike-sharing services, benefiting urban commuters.*

***Keywords:*** *Bike-sharing systems, Linear regression, Forecasting, Historical data, Data preprocessing*

## I. INTRODUCTION

With the rise of urbanization and the focus on sustainable transportation, accurately predicting bike demand is crucial for urban planners, bike-sharing programs, and bicycle manufacturers. Linear regression can model the relationship between independent variables (e.g., time, weather, socio-economic factors) and bike demand. Historical data on bike rentals, including date, time, weather, and special events, were collected and pre-processed, addressing missing values, coding categorical variables, and measuring numerical features. Key factors influencing bike demand, such as time of day, day of the week, weather conditions, public holidays, and special events, were identified. A linear regression model was trained with this data, and its performance evaluated using metrics like mean squared error (MSE), root mean squared error (RMSE), and R-squared value. The model provides insights into factors affecting bike demand, aiding urban planners, bike-sharing programs, and manufacturers in resource allocation, infrastructure planning, and meeting commuter needs. This research supports sustainable transportation development and addresses urbanization challenges and the demand for environmentally friendly transport options.

## II. RELATED WORK

Bike-sharing systems have become a crucial part of urban transportation, offering convenience and sustainability. Research by Fishman, Washington, and Haworth (2013) provides insights into their operational, economic, and social aspects, highlighting their evolution and impact [1]. Leveraging big data, researchers have examined cycling patterns [2] and activity patterns within bike-sharing systems [4], as well as the influence of weather and events on usage [5]. Studies by Corcoran et al. (2014) [5] and Zhang and Mi (2018) [7] showcase environmental benefits, including reduced carbon emissions. Optimization strategies for bike availability are discussed by Raviv, Tzur, and Forma (2013) [3], while Tran, Ovtracht, and Faivre d'Arcier (2015) [9] explore the impact of the built environment on usage. Additionally, agent-based models for passenger transportation have been developed [10]. The importance of bike-sharing for environmental sustainability is further highlighted by big data analyses of their environmental benefits [8]. The WEKA workbench, as detailed by Frank, Hall, and Witten (2016) [6], has facilitated data mining in this field, enhancing understanding and system efficiency.

## III. METHODOLOGY

**Data Collection:**

For data collection, Historical bike rental data is collected from various sources, including Kaggle, containing daily demand, date, time, and weather conditions (temperature, humidity, wind speed, precipitation). The dataset includes variables such as "season" (1 to 4), "year" (0 and 1), "month" (1 to 12), "holiday" (0 and 1), "weekday" (0 to 6), "working day" (0 and 1), and "weather situation" (1 to 3).

**Data Preprocessing:**

In data preprocessing, we check for missing values; our dataset has none. For missing values, we use imputation techniques. We extract features like day of the week, month, and holidays from the date attribute. Categorical variables are converted into numerical values using one-hot or label encoding.

**Data Analysis:**

Exploratory Data Analysis (EDA) involves summarizing and visualizing data to understand its characteristics and relationships. Analysts use statistics and plots to examine feature distributions and correlations, guiding further analysis and modelling decisions.

**Data Splitting:**

To evaluate the linear regression model, we split the dataset into training and testing sets. Typically, 80% of the data is used for training and 20% for testing. Random splitting is essential to prevent bias and ensure the model generalizes well to unseen data.

**Model Selection:**

Linear regression models the relationship between a dependent variable and independent variables. It assumes a linear relationship and is widely used due to its simplicity,

interpretability, and computational efficiency, making it a common baseline in machine learning projects.

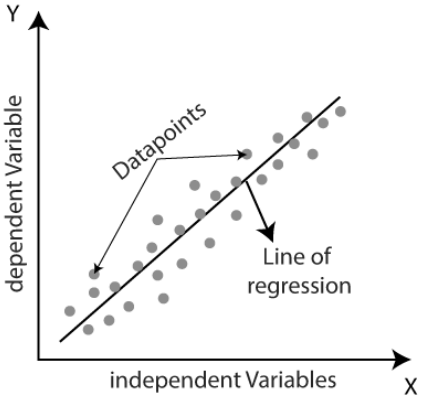


Fig. 1. Linear Regression

**Model Training:**

We train our linear regression model for bike-sharing demand forecasting using Sci-Kit Learn's Linear Regression function and Recursive Feature Elimination (RFE). This process fits the model to the training data, learning coefficients that minimize prediction errors. While effective, linear regression may not capture complex patterns as well as other techniques like decision trees or neural networks.

**Model Evaluation:**

We evaluate our linear regression model using MAE, MSE, RMSE, and R-squared ($R2$) score. Lower values for MAE, MSE, and RMSE, and a higher $R2$ score indicate better performance. These metrics help us understand and improve the model's accuracy in predicting bike demand.

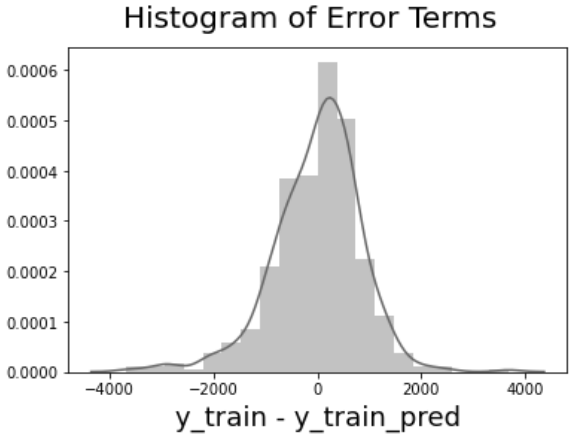


Fig. 2. Residual Analysis of Linear Regression

Normality of error terms ensures symmetric distribution around mean, indicating unbiased predictions. A mean of 0 ensures the model neither overestimates nor underestimates, enhancing reliability and interpretation of coefficients.

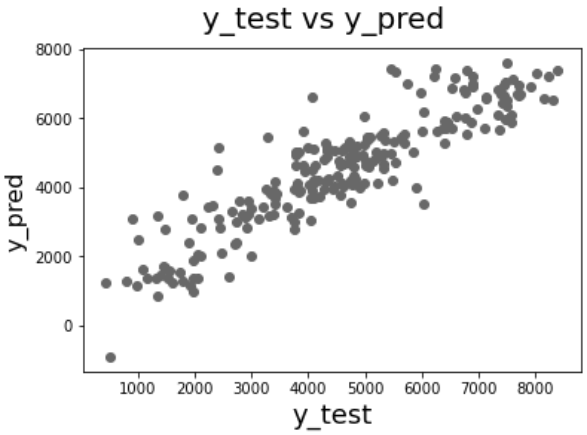


Fig. 3. Model Evaluation

The graph displays a strong correlation between predicted and actual y_test values, confirming the model's accuracy in predicting bike rental demand. Minimal dispersion around the line of perfect correlation further supports the model's reliability and predictive power.

**Predictions:**

To predict bike rental demand, we'll use the "predict" function in Python with scikit-learn. This generates estimates based on independent variables. By applying our model to new data, we can accurately forecast demand, helping companies optimize operations and improve customer satisfaction. This ensures our model generalizes well to unseen data, making it a valuable tool for future predictions.

## IV. RESULTS AND ANALYSIS

The linear regression model showed high R-squared, adjusted R-squared, high F-value, and low Prob(F) value, indicating a strong fit. **Season_spring** and **weathersit_Rainy** were significant predictors with high negative coefficients, while **mnth_Nov** had a relatively high p-value (0.096). The final model included **temp**, **season_spring**, and **mnth_Sep**, dropping less influential variables like **mnth_July**. The model equation clarified the impact on bike rental demand: biking counts increase in September with moderate temperatures, but decrease during rainy, misty days, or holidays.

## V. CONCLUSION

In conclusion, this study developed a predictive model for bike-sharing demand, considering factors like temperature, season, weekday, and weather conditions. Recommendations include

placing stations strategically, planning for peak hours, implementing dynamic pricing, developing infrastructure along popular routes, scheduling maintenance efficiently, and offering weekend discounts. These strategies aim to improve bike-sharing efficiency and user satisfaction, providing valuable insights for operators and urban planners.

## VI. REFERENCES

[1] Fishman, Elliot, Simon Washington, and Narelle Haworth. "Bike Share: A Synthesis of the Literature." *Transport Reviews* 33, no. 2 (2013)

[2] Romanillos, Gustavo, Martin Zaltz Austwick, Dick Ettema, and Joost De Kruijf. "Big data and cycling." *Transport Reviews* 36, no. 1 (2016)

[3] Raviv, Tal, Michal Tzur, and Iris A. Forma. "Static repositioning in a bike-sharing system: models and solution approaches." *EURO Journal on Transportation and Logistics* 2, no. 3 (2013)

[4] Vogel, Patrick, Torsten Greiser, and Dirk Christian Mattfeld. "Understanding bike-sharing systems using data mining: Exploring activity patterns." *Procedia-Social and Behavioral Sciences* 20 (2011)

[5] Corcoran, Jonathan, Tiebei Li, David Rohde, Elin Charles-Edwards, and Derlie Mateo-Babiano. "Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events." *Journal of Transport Geography* 41 (2014)

[6] Frank, Eibe, Mark A. Hall, and Ian H. Witten. *The WEKA workbench*. Morgan Kaufmann, (2016)

[7] Zhang, Yongping, and Zhifu Mi. "Environmental benefits of bike sharing: A big data-based analysis." *Applied energy* 220 (2018)

[8] Ricci, Miriam. "Research in Transportation Business & Management." (2015)

[9] Tran, Tien Dung, Nicolas Ovtracht, and Bruno Faivre d'Arcier. "Modeling bike sharing system using built environment factors." *Procedia Cirp* 30 (2015)

[10] Hajinasab, Banafsheh, Paul Davidsson, Jan A. Persson, and Johan Holmgren. "Towards an agent-based model of passenger transportation." In *Multi-Agent Based Simulation XVI: International Workshop, MABS 2015, Istanbul, Turkey, May 5, 2015, Revised Selected Papers 16*, pp. 132-145. Springer International Publishing, (2016)

[11] Fanaee-T, Hadi & Gama, João. Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence. (2014)