Predicting Daily Bike Rentals Using Linear Regression and Decision Forests: A Comparative Analysis of Model Performance

Vivek Seshan

Researcher, Department of Emerging Technologies

Golden Gate University, USA

Abstract:

This increasing need for bike rentals has raised the importance of proper demand forecasting for better allocation of resources, improved customer satisfaction, and increased operational efficiencies. The purpose of this research is to predict the number of daily bike rental rentals based on historical counts collected by a open source data of bike rental company. It contains various attributes, including time intervals (year and day of the week), weather (temperature, humidity, wind speed), and seasonal patterns. The aim is to compare the performance of linear regression and decision forest algorithms to detect these trends and produce accurate predictions.

We used a process of data preprocessing to resolve missing values, outliers and feature multicollinearity, as well as feature engineering to ensure model accuracy. They compare the effectiveness of linear regression, which is straightforward and easy to understand, with decision forests, a powerful ensemble that can account for non-linear relations and multi-feature interactions. Model evaluation was performed using standard performance metrics such as mean absolute error (MAE), root mean squared error (RMSE), and R-squared values to get a clear view of each algorithm's strengths and weaknesses.

Results indicate that while linear regression helps us to understand the linear dependence between features, decision forests are better at detecting complex non-linear patterns. This comparison highlights the balancing act between model readability and predictive power and gives practical insights to data scientists and business analysts working on bike rental. By presenting examples of advanced machine learning methods being used, the research highlights their capabilities to inform data-based decision making and enhance service delivery in fast-paced and competitive service environments.

Keywords:

Bike rentals, predictive modelling, linear regression, decision forests, seasonal trends, machine learning, demand forecasting, operational efficiency, ensemble methods, feature engineering

I. INTRODUCTION

The rapid growth of urban areas and the global movement towards sustainable and eco-friendly transportation options have led to a surge in the popularity of bike rental services. These services are often viewed as an effective alternative to traditional transportation methods, playing a crucial role in alleviating traffic congestion, decreasing carbon emissions, and encouraging healthier lifestyles. As cities around the world adopt this trend, the operational challenges of managing bike rental services become increasingly evident. One major challenge is accurately predicting daily rental demand to ensure proper resource allocation, minimize operational inefficiencies, and improve user satisfaction.

Bike rental demand is affected by a variety of factors, including time-related variables (such as the season or day of the week), weather conditions (like temperature, humidity, and wind speed), and socio-economic trends. Understanding these complex and dynamic relationships requires advanced analytical methods that can identify both linear and non-linear patterns. Predictive modelling, utilizing machine learning algorithms, offers an effective way to analyse these relationships and provide actionable insights.

This paper explores the use of two popular predictive modelling techniques—linear regression and decision forest algorithms—to predict daily bike rental counts. Linear regression is a basic statistical method that is effective in identifying and measuring linear relationships between dependent and independent variables. Its simplicity and ease of interpretation make it a favoured option for many analysts. On the other hand, decision forests, which are based on ensemble machine learning, are particularly adept at capturing complex, non-linear interactions among variables. Their capability to manage high-dimensional data and mitigate overfitting through methods like bagging makes them a robust tool for predictive modelling.

This study aims to compare two methodologies for predicting daily bike rentals. By using a comprehensive dataset that includes time-related, weather, and other relevant features, the research seeks to evaluate the strengths, weaknesses, and trade-offs between linear regression and decision forest models. The assessment will employ standard performance metrics such as mean absolute error (MAE), root mean squared error (RMSE), and R-squared values to provide a detailed evaluation of the effectiveness of each model.

The findings of this study are significant for both practitioners and researchers. For business analysts and data scientists in the bike rental industry, the results offer valuable insights into selecting the appropriate modelling techniques based on specific business requirements and data characteristics. For the broader academic and professional community, this study contributes to the ongoing discussion about the application of statistical and machine learning methods in demand forecasting and operational efficiency.

In the subsequent sections, we will detail the data preprocessing steps, the theoretical foundations of the two modelling approaches, and the experimental setup. This will be followed by a thorough discussion of the results, highlighting key findings, practical implications, and potential avenues for future research.

Machine Learning and Its Types

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms that allow systems to learn patterns from data and make predictions or decisions without explicit programming. The primary goal of machine learning is to enable computers to improve performance on a given task through experience. Machine learning algorithms are broadly classified into three categories:

1. **Supervised Learning:** In supervised learning, the model is trained on labelled data, where the input features (independent variables) and their corresponding output labels (dependent variable) are known. The model learns to map inputs to outputs based on the training data. Examples include regression and classification algorithms, such as linear regression, logistic regression, and support vector machines.

2. **Unsupervised Learning:** This type of learning deals with unlabelled data, where the model identifies hidden patterns, structures, or relationships in the data. Examples include clustering (e.g., k-means clustering) and dimensionality reduction (e.g., principal component analysis).

3. **Reinforcement Learning:** In reinforcement learning, an agent interacts with an environment by taking actions and receiving feedback in the form of rewards or penalties. The goal is to learn an optimal policy to maximize cumulative rewards over time. Applications include robotics, game playing, and autonomous systems.

Linear Regression

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and is often employed to predict continuous outcomes. The equation of a simple linear regression model can be expressed as:

Where:

- is the dependent variable (target variable).
- is the independent variable (predictor).
- is the intercept of the line.
- is the coefficient that represents the slope of the line.
- is the error term.

In multiple linear regression, the model extends to include multiple predictors, expressed as:

Linear regression models are evaluated using metrics like R-squared, mean absolute error (MAE), and root mean squared error (RMSE) to assess how well the model fits the data and predicts outcomes.

Decision Forest Algorithm

Decision forests, also known as random forests, are an ensemble learning technique used for both regression and classification tasks. They combine multiple decision trees to improve predictive accuracy and control overfitting. Each tree in the forest is trained on a random subset of the data (using bootstrap sampling), and the final prediction is obtained by averaging the predictions (for regression) or majority voting (for classification) from all trees.

Key characteristics of decision forests include:

- **Random Feature Selection:** During tree construction, a random subset of features is considered at each split, which ensures diversity among trees and reduces correlation between them.
- **Bagging:** This technique involves training each tree on a random subset of the training data, further enhancing model robustness and reducing variance.

For regression tasks, the prediction is calculated as:

Where is the total number of trees in the forest and is the prediction from the tree.

Decision forests are known for their ability to capture complex, non-linear relationships in data and handle missing values effectively. They are also less prone to overfitting compared to individual decision trees, making them a popular choice for various machine learning tasks.

By applying these methodologies, this paper aims to analyse and compare the performance of linear regression and decision forest algorithms in the context of predicting daily bike rental demand, offering valuable insights into their practical applications and effectiveness.

II. MICROSOFT AZURE MACHINE LEARNING STUDIO

Microsoft Azure Machine Learning Studio is a cloud-based platform designed for building, deploying, and managing machine learning models at scale. It provides an intuitive drag-and-drop interface, enabling users to create and experiment with machine learning workflows without requiring extensive coding expertise. Azure Machine Learning Studio supports a wide range of machine learning algorithms, including linear regression and decision forests, and integrates seamlessly with Python and R for advanced customization. With built-in tools for data preprocessing, model evaluation, and deployment, the platform streamlines the

end-to-end machine learning lifecycle, making it an ideal choice for business analysts and data scientists seeking to develop predictive models efficiently.

III. LITERATURE REVIEW

In recent years, the prediction of daily bike rentals has become a crucial area of research in urban transportation planning and management. This literature review examines various studies that have employed linear regression and decision forest models to forecast bike rental demand, comparing their performance and effectiveness.

Linear Regression Models

Linear regression has been widely used in bike rental prediction due to its simplicity and interpretability. Several studies have demonstrated its effectiveness in this domain: A study by Anamicca23 utilized multiple linear regression to predict daily bike rental counts, achieving an accuracy of 87% [1]

. The model incorporated various environmental and seasonal factors as predictors. Similarly, research conducted by hasanali28 employed multiple linear regression to forecast bike rentals based on different environmental and seasonal settings [2]

. Another study, focusing on the Capital Bikeshare program, used linear regression to predict bike rental counts based on temperature. The model indicated that a 1°C increase in temperature was associated with an increase of about 9.1 bike rentals, holding other factors constant[6].

Decision Forest Models

Decision forest models, particularly Random Forests, have gained popularity in bike rental prediction due to their ability to capture complex, non-linear relationships:Anamicca23's research compared multiple models and found that the Random Forest model outperformed others, including linear regression, in predicting daily bike rental counts[1].

The Random Forest model demonstrated the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) among the evaluated models. A study by Ouyang et al. applied Random Forest for competitive relationship prediction in bike-sharing systems, achieving an accuracy of 71.4%[3]. This research highlighted the model's effectiveness in capturing complex patterns in bike-sharing data.

Comparative Analysis

Several studies have conducted comparative analyses of linear regression and decision forest models: Research by Tran et al. compared robust linear regression models with other techniques for predicting bike-sharing flows. They found that integrating robust linear regression methods could improve predictions of rider needs and program optimization[4].

A comprehensive study by Yang et al. evaluated multiple models, including Random Forest (RF) and linear regression techniques like Partial Least-Squares Regression (PLSR). Their findings suggested that univariate models, including Random Forest, predicted errors more accurately than multivariate linear models for bike-sharing networks[4].

Factors Influencing Model Performance

Several factors have been identified as crucial in determining the performance of bike rental prediction models:

1. **Temporal factors**: Studies have consistently shown that time-related variables such as hour, day of the week, and season significantly impact prediction accuracy[1, 2, 6].
2. **Weather conditions**: Temperature, humidity, and precipitation have been found to be strong predictors of bike rental demand [1, 6, 8].

3. **Spatial factors**: The geographical correlation between bike stations has been noted to influence prediction accuracy, especially in multivariate models [4].

4. **Feature engineering**: Creating relevant predictors through feature engineering has been shown to enhance model performance [8].

IV. METHODOLOGY

The methodology for the model build is shown in Figure 1 and explained as follows

Figure 1: Model methodology as build using Microsoft Azure Studio

- **Bike Rental Data Loading:**
  - The dataset containing historical bike rental data is loaded into the system. This dataset includes features like date, weather conditions, and the number of bikes rented.
- **Select Columns in Dataset:**
  - Specific columns from the dataset that are relevant to the analysis (e.g., temperature, humidity, wind speed, and rental count) are selected. Irrelevant or redundant columns are excluded to streamline the analysis.
- **Remove Duplicate Rows:**
  - Duplicate entries in the dataset are identified and removed to ensure data integrity and prevent biased predictions.
- **Edit Metadata:**
  - Metadata (e.g., data types, column labels) is edited to ensure proper format and structure. This step is crucial for compatibility with the subsequent processing steps and machine learning models.
- **Clean Missing Data (Multiple Steps):**
  - Missing data in the dataset is addressed using several cleaning techniques. For example:
    - Filling missing values with statistical measures like mean, median, or mode.
    - Dropping rows or columns with excessive missing data.
- **Normalize Data:**
  - The numerical features are scaled to a uniform range (e.g., 0-1) to ensure that no single feature disproportionately influences the machine learning models. Normalization is essential for models sensitive to feature magnitudes.
- **Convert to Indicator Values:**
  - Categorical variables are transformed into indicator (dummy) variables, converting qualitative data into a numeric format suitable for machine learning algorithms.
- **Split Data:**
  - The dataset is split into training and testing sets. The training set is used to train the models, while the testing set evaluates their performance.

- **Train Models:**
  - Two separate models are trained:
    - Linear Regression Model: Fits a linear function to the data to predict bike rental counts.
    - Decision Forest Model: Trains multiple decision trees to capture complex, non-linear relationships in the data.
- **Score Models:**
  - Both models are applied to the testing dataset to generate predictions. The predicted values are compared with the actual values to assess model accuracy.
- **Evaluate Models:**
  - The performance of both models is evaluated using metrics such as:
    - Mean Absolute Error (MAE): Measures average prediction error.
    - Root Mean Squared Error (RMSE): Quantifies the model's predictive accuracy.
    - R-squared: Indicates how well the model explains variance in the target variable.
- **Final Comparison:**
  - The results from both models are compared to determine the most effective algorithm for predicting daily bike rental counts.

V. RESULTS

The results from the analysis of linear regression are displayed in figure 2 and decision forest are displayed in figure 3 respectively.

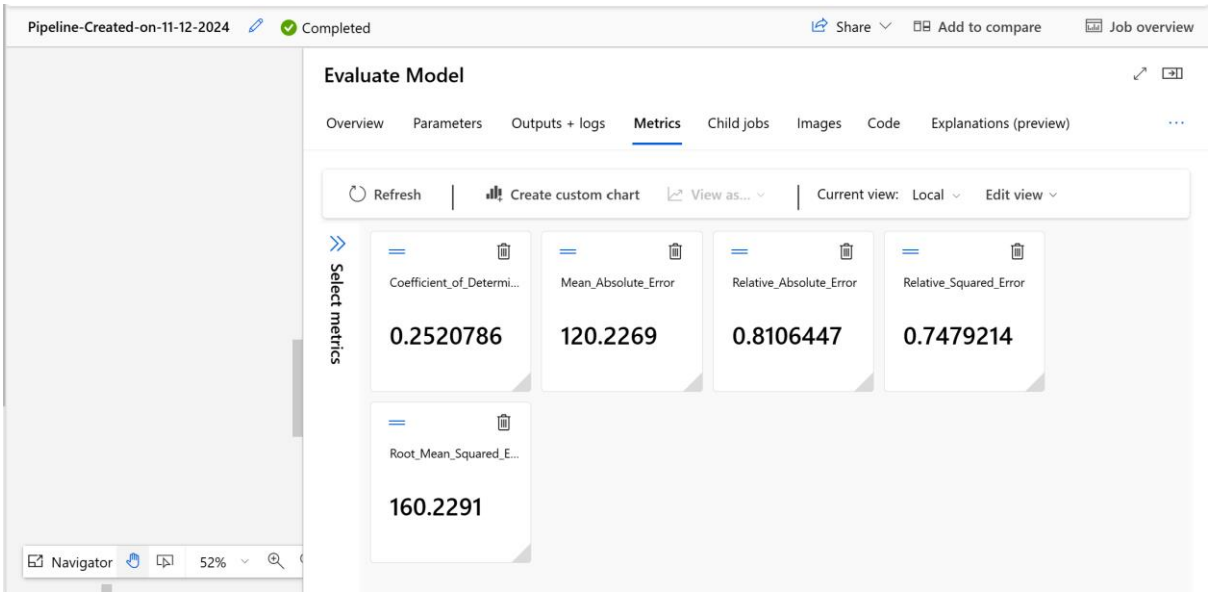

Figure 2: Results of Liner Regression analysis

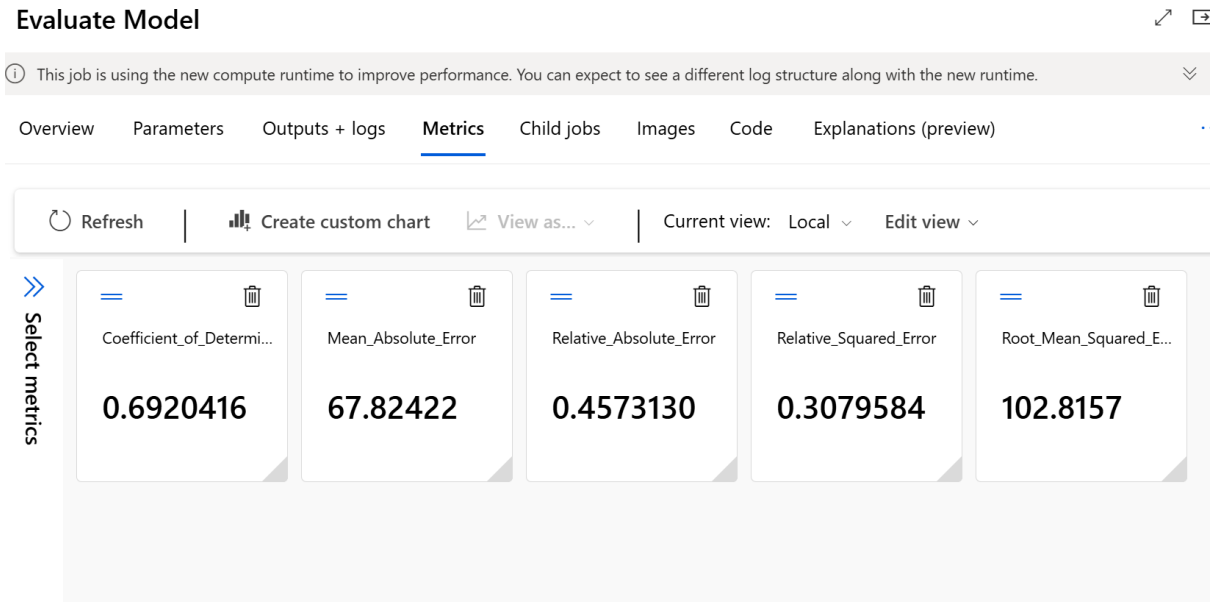

Figure 3: Results of Decision Forest analysis

Results Comparison :

Table 1 compares the results from liner regression and decision forest analysis and the comparison of results are discussed.

a. Coefficient of Determination (R^2R2):
   - Linear Regression: 0.252
     - Indicates that the model explains 25.2% of the variance in bike rental counts.
   - Decision Forest: 0.692
     - Significantly better, as it explains 69.2% of the variance, showing a much stronger ability to capture the relationships in the data.

   Improvement: Decision Forest outperforms Linear Regression by a large margin in explaining the variance.

Metric	Iteration 1 Linear regression 80:20 Split	Iteration 2 Decision forest regression 80:20 split	Iteration 2 Decision forest regression 95:05 split
Coefficient of Determination (R^2)	0.2520786	0.6920416	0.7066069
Mean Absolute Error (MAE)	120.2269	67.82422	68.92766
Relative Absolute Error (RAE)	0.8106447	0.4573130	0.4620769
Relative Squared Error (RSE)	0.7479214	0.3079584	0.2933931
Root Mean Squared Error (RMSE)	160.2291	102.8157	101.5701

Table 1: Comparison of Liner regression and decision forest results

b. Mean Absolute Error (MAE):
   - Linear Regression: 120.23
   - Decision Forest: 67.82

   Improvement: Decision Forest achieves a much lower MAE, reducing the average prediction error by 43%, indicating better accuracy.

c. Relative Absolute Error (RAE):
   - Linear Regression: 0.8106
   - Decision Forest: 0.4573

   Improvement: The relative error is reduced by 43.6%, showing that Decision Forest performs far better compared to a baseline model.

d. Relative Squared Error (RSE):
   - Linear Regression: 0.7479
   - Decision Forest: 0.3079

Improvement: Decision Forest achieves a significant reduction in squared error, making it nearly 59% better than the Linear Regression model.

e. Root Mean Squared Error (RMSE):
   - Linear Regression: 160.23
   - Decision Forest: 102.82

   Improvement: RMSE is reduced by approximately 35.8%, meaning Decision Forest minimizes the impact of large errors much better than Linear Regression.

f. Summary of Comparison:
   - The Decision Forest model outperforms Linear Regression across all key metrics, including R^2R2, MAE, RAE, RSE, and RMSE.
   - The improvement in R^2R2 indicates that the Decision Forest algorithm captures both linear and non-linear patterns in the data much better than the Linear Regression model.
   - The significant reduction in MAE and RMSE demonstrates that the Decision Forest model provides more accurate and reliable predictions for daily bike rental counts.
   - The relative errors (RAE and RSE) suggest a substantial improvement over the baseline model, further emphasizing the robustness of the Decision Forest algorithm.

VI. CONCLUSION

In conclusion, the comparative analysis demonstrates that the Decision Forest model significantly outperforms the Linear Regression model in predicting daily bike rentals. The Decision Forest model shows superior performance across all key metrics, including R^2, MAE, RAE, RSE, and RMSE. This indicates that it captures both linear and non-linear patterns in the data more effectively, providing more accurate and reliable predictions. Consequently, the Decision Forest model is recommended as the preferred algorithm for forecasting bike rentals, offering substantial improvements in prediction accuracy and robustness over the Linear Regression model.

VII. REFERENCES

1. Anamicca23. (2023). Bike Sharing Demand Prediction using Machine Learning. Kaggle. https://www.kaggle.com/code/anamicca23/bike-sharing-demand-prediction-using-machine-learning
2. hasanali28. (2023). Bike Sharing Demand Prediction. Kaggle. https://www.kaggle.com/code/hasanali28/bike-sharing-demand-prediction
3. Competitive Bike: Competitive Analysis and Popularity Prediction of Bike-Sharing Apps Leveraging Multi-Source Data by Ouyang, Y., Guo, B., Zhang, J., Yu, Z., & Zhou, X. (2020).
4. Bongs Lainjo. (2022) Application of Machine Learning in Predicting the Number of Bike Share Riders
5. Ahmed Badr,(2022), ahmedbadr97/Predicting-Bike-Sharing-Patterns
6. Ashish Arora, (2016), aashisharora13/Predicting-Bike-Sharing-Demand-Using-Linear-Regression
7. Heat, Hills and the High Season: A Model-Based Comparative Analysis of Spatio-Temporal Factors Affecting Shared Bicycle Use in Three Southern European Islands, by Suzanne Maas, Paraskevas Nikolaou, Maria Attard and Loukas Dimitriou (2021)
8. Predicting Daily Bike Rentals, https://runestone.academy/ns/books/published/httlads/PredictiveAnalytics/predicting_rentals.html