CAPITAL BIKES ANALYSIS

A Comprehensive Analysis of Bike Usage and Trends

By:
Bezel Ndhlovu
Johannes Machinya
Genius Mpala

University of Notre Dame

MENDOZA COLLEGE OF BUSINESS

December 2024

Abstract

The rapid growth of urban bike-sharing systems highlights their potential as sustainable transportation solutions, yet operational challenges persist. This project analyzes data from Capital Bikes, a leading bike-sharing program in Washington, D.C., to uncover patterns and optimize operations. Leveraging over 16 million individual ride records alongside weather metrics, station details, and usage frequencies, our analysis explores temporal, spatial, and environmental factors influencing hourly bike rental demand per station.

Key findings include a 30.8% dominance of weekday rides by members and a significant weekend contribution by casual riders (25.8%). Peak demand aligns with commuting hours (5 PM), emphasizing the need for dynamic fleet allocation. Predictive modeling using Linear Regression, Random Forest, and XGBoost revealed Random Forest as the most effective, achieving an RMSE of 2.4317 while capturing complex non-linear relationships.

The study provides actionable insights to enhance fleet management, weather-responsive operations, and targeted marketing campaigns, contributing to a more sustainable and efficient bike-sharing ecosystem. Future directions include incorporating traffic and demographic data, real-time predictions, and user behavior studies to further optimize bike-sharing systems in urban areas.

This project bridges descriptive analytics and predictive modeling, offering a scalable framework for improving bike-sharing programs across the United States.

Table of Equations

Table of Figures

1. Introduction

The expansion of urban transportation systems has highlighted the need for sustainable mobility solutions, with bike-sharing programs emerging as a key approach to reducing traffic congestion, lowering carbon emissions,

and encouraging healthier lifestyles. Capital Bikes, a leading bike-sharing program in Washington, D.C, has amassed extensive data on daily rentals, weather conditions, station usage, and user behavior. This project seeks to harness this data to uncover insights into factors influencing bike-sharing behavior and to develop strategies for improving operational efficiency and user engagement.

Through the application of data analytics and machine learning techniques, the project aims to identify usage trends, predict demand patterns, and analyze the influence of external factors such as weather and seasonality on bike rentals. Using datasets provided by Capital Bikes, encompassing daily rentals, weather metrics, station details, and usage frequencies, the outputs will include descriptive analyses, predictive models, and actionable visualizations to support data-driven decision-making.

This work has significant implications for optimizing fleet management, minimizing idle bike inventory, and enhancing customer satisfaction, while contributing to sustainable transportation planning and promoting community well-being. By providing a framework for informed operational decisions, this project will help Capital Bikes improve system reliability and expand its reach.

2. Related Work

Research on bike-sharing programs has explored usage patterns, demand prediction, and operational optimization. Shaheen et al. (2010) provided a foundational review of global bike-sharing systems, identifying key adoption factors such as infrastructure, pricing, and weather, and underscoring the role of data-driven insights in enhancing program efficiency and sustainability. Faghih-Imani et al. (2014) used GPS data to analyze spatial and temporal dynamics, revealing how station location, weather, and time of day influence bike usage, and offering strategies for station placement and fleet rebalancing.

Advancements in predictive analytics have added depth to this field. For instance, Li et al. (2015) combined weather, temporal, and station data in machine learning models to forecast bike demand but encountered challenges with data imbalance and interpretability. Similarly, Zhao et al. (2019) applied deep learning to predict hourly bike demand, achieving higher accuracy at the cost of increased computational complexity and data granularity requirements.

While these studies have advanced the understanding of bike-sharing systems, they often focus on either descriptive insights or predictive modeling, with limited integration. Additionally, the complexity of some models can hinder their interpretability for key stakeholders. This project addresses these gaps by combining descriptive analytics and machine learning techniques to provide interpretable and actionable insights for Capital Bikes. By addressing challenges such as missing data and class imbalance, this work enhances model robustness and reliability. It contributes to the field by bridging the gap between exploratory insights and predictive analytics, offering a scalable framework for other bike-sharing programs.

3. Data Description

Our analysis focuses on a comprehensive dataset capturing various aspects of bike-sharing trends in Washington, DC. Below is a detailed overview of the data and preprocessing steps:

3.1 Data Overview

The project leverages a comprehensive collection of datasets to analyze and predict bike-sharing patterns. The Bike Rental Details dataset consists of over 16 million records, capturing information on individual rides, including timestamps, station details, user types (member or casual), and rideable types. These variables provide a granular view of rental behavior and user demographics. Complementing this is the Station List, which catalogs 916 stations with unique identifiers and names, facilitating spatial analysis of bike-sharing activity.

The Usage Frequency dataset contains 873,318 daily records, tracking station-specific pickups and drop-offs. This data is critical for identifying demand trends and evaluating station performance. Additionally, the Weather Data includes 1,584 daily entries, covering a wide range of meteorological metrics such as temperature,

precipitation, wind, and solar radiation. These variables enable the study of environmental factors influencing bike rentals and help incorporate seasonality and weather conditions into predictive models.

Together, these datasets provide a rich foundation for exploring bike-sharing dynamics, uncovering trends, and developing strategies to optimize operations and enhance user engagement.

3.2 Preprocessing Steps

The preprocessing process involved multiple steps to prepare the data for analysis. Data cleaning was conducted to remove all missing values, reducing the dataset from 16 million records to a clean subset. This ensured consistency and integrity in the analysis. Subsequently, data integration combined weather, station, and time-based information with bike rentals, creating a consolidated view for analysis. The correlation between the dependent variable (total_rides) and independent variables was examined (see chart below), and variables with strong correlations and low multicollinearity were selected to inform feature selection and guide the linear regression model.

Feature engineering introduced new contextual variables, such as is_holiday, weekend, and seasonal indicators (e.g., season_Winter, season_Summer). The final dataset, consisting of 1,974,795 observations and 293 variables, incorporates weather metrics (e.g., temperature, humidity, precipitation), time-based attributes (e.g., hour, date, week, month, year), station identifiers, and contextual variables. These steps established a robust foundation for analysis and model development.

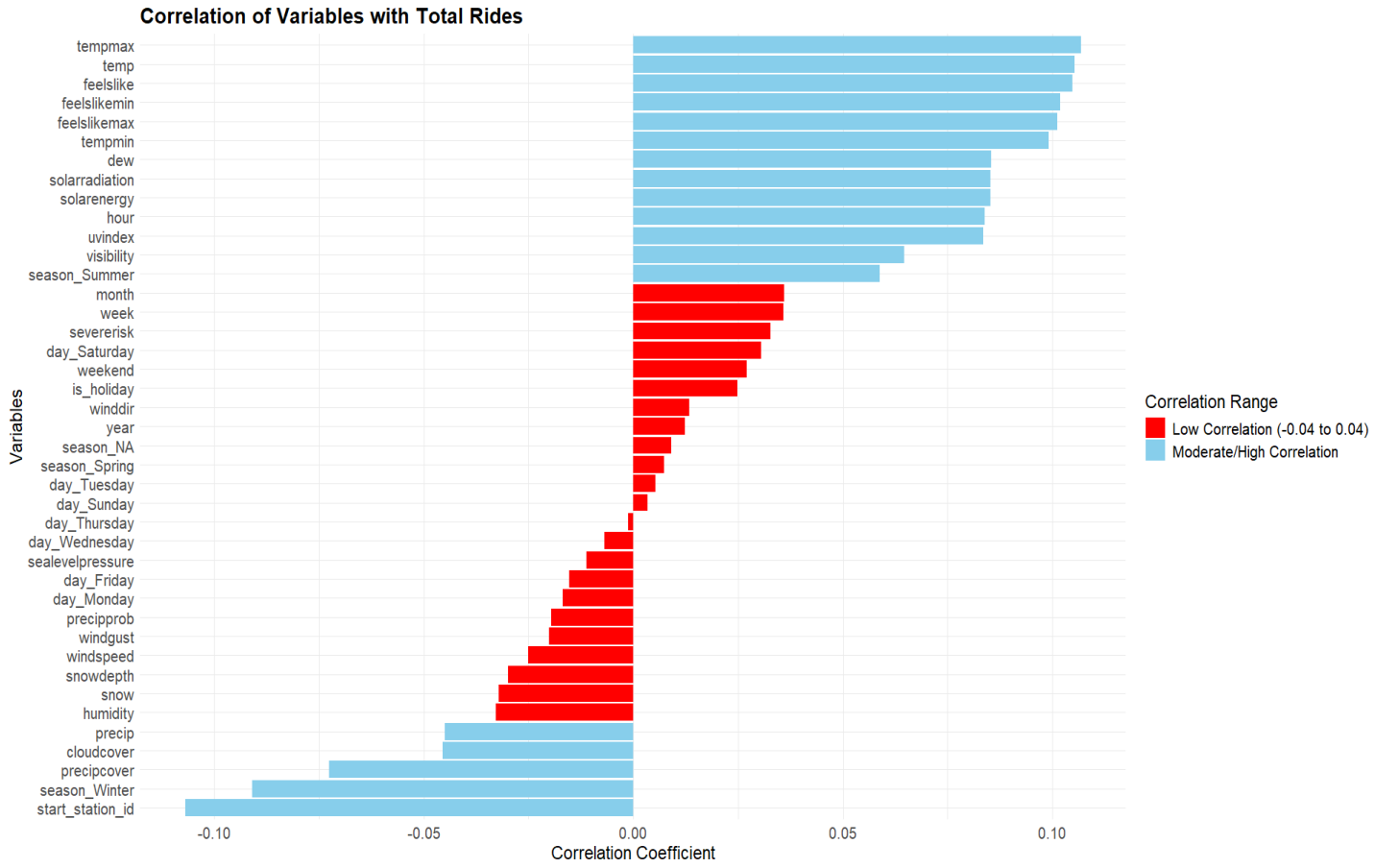


Figure 1Correlation of Variables with Rides

3.3 Data Splitting

I. **Training Set**: 80% of the dataset for model training and development.
II. **Test Set**: 20% reserved for final evaluation to ensure unbiased performance metrics.

3.4 Summary Graphs

The radial plot shows the hourly distribution of bike rentals, revealing peak usage at **5 PM**, likely aligned with evening commute times.

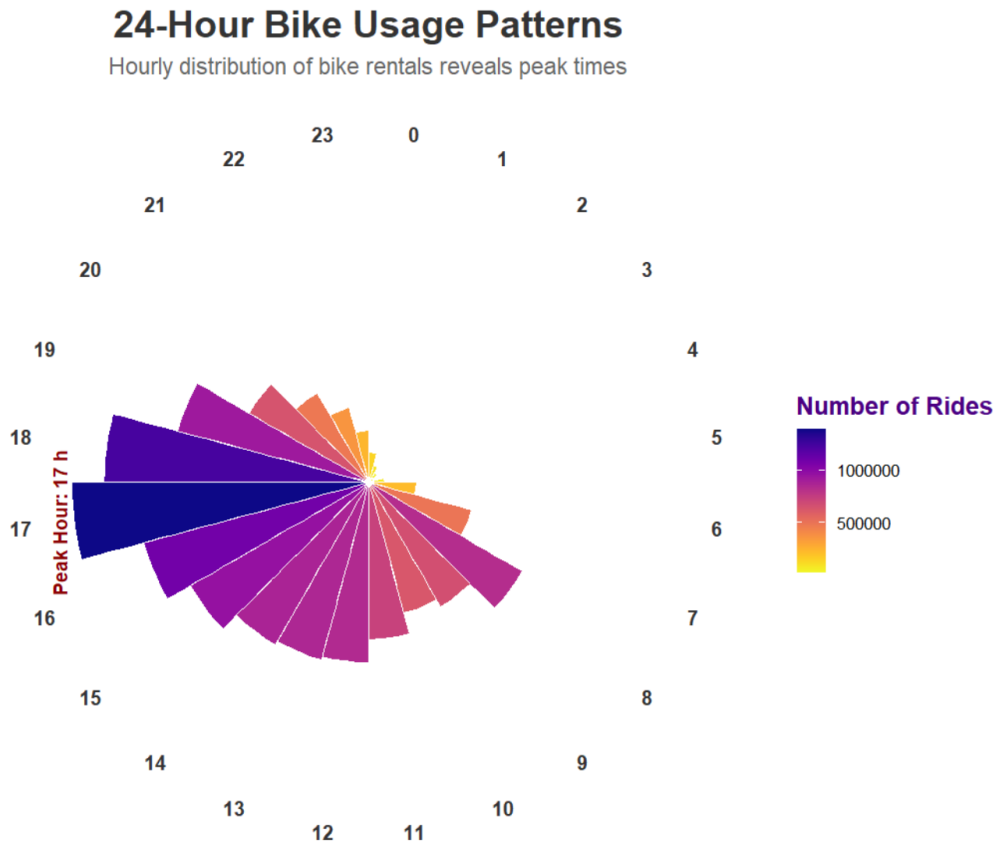


Figure 2: 24 Hour Bike Usage Patterns

The bar chart demonstrates average daily rides, segmented by user type (casual or member) and day type (weekday or weekend). **Members dominate weekday rides (30.8%)**, while casual riders contribute significantly on weekends (25.8%).

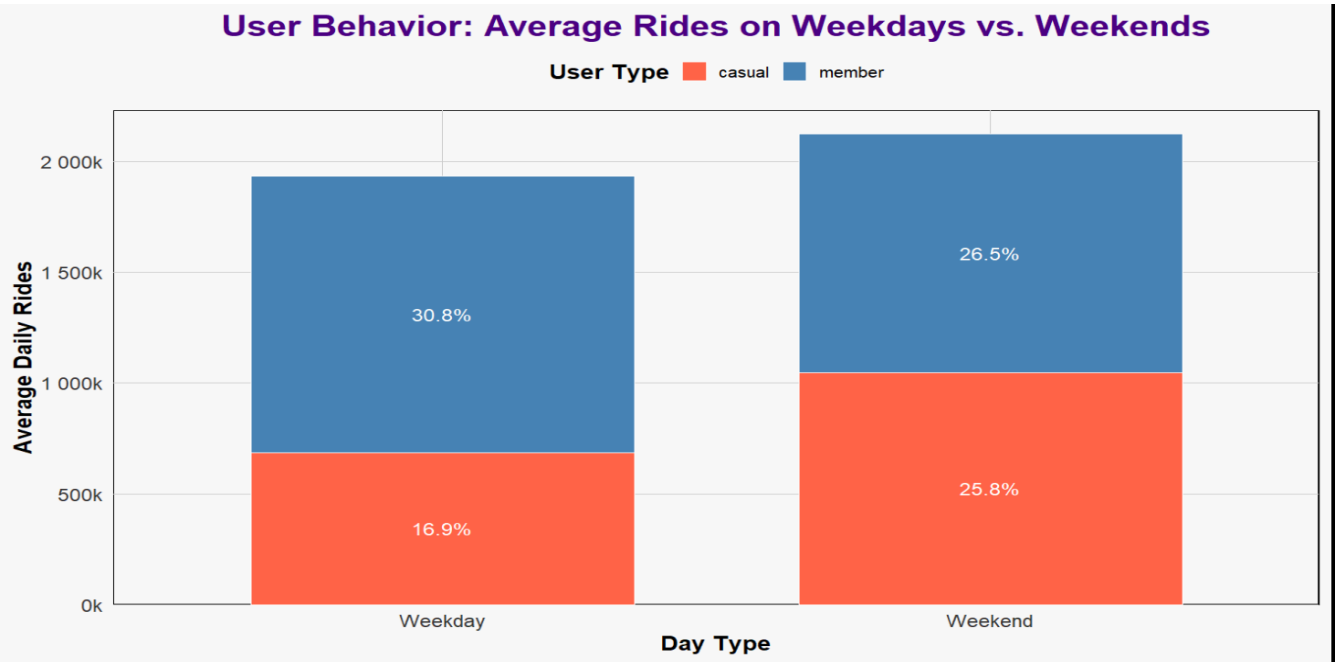


Figure 3: User Behavior: Average Rides on Weekdays vs Weeks

By preparing and analyzing this enriched dataset, we could extract meaningful insights about user behavior, station demand, and the influence of weather and temporal factors, laying the groundwork for predictive modeling and data-driven operational recommendations.

4. Methods

For this project, we applied three predictive modeling techniques: Linear Regression, Random Forest, and XGBoost. Each method was chosen to address specific challenges in predicting total_rides and to leverage their respective strengths for performance improvement.

Linear Regression was used as a baseline model due to its simplicity and interpretability. It assumes a linear relationship between the predictors and the response variable, making it suitable for understanding how individual features contribute to the prediction of total_rides. Linear regression's strengths lie in its computational efficiency and ease of interpretability, as it provides clear coefficients that quantify the effect of each predictor. However, its limitations include the inability to capture non-linear relationships and interactions among variables, which are likely present in our dataset, given the complex interplay between weather conditions, time variables, and ride demand.

Random Forest was employed to overcome the limitations of linear regression by modeling non-linear relationships and interactions. It is an ensemble learning method that builds multiple decision trees and aggregates their predictions, improving accuracy and reducing overfitting. Random Forest is particularly effective for handling large datasets with numerous features and complex patterns, as it does not require strong assumptions about the data distribution. Despite its strengths, the method has limitations, such as reduced interpretability and high computational cost, especially for large datasets.

XGBoost, a gradient boosting algorithm, was implemented to further enhance predictive accuracy. XGBoost builds decision trees iteratively, optimizing each tree to correct the errors of the previous ones. Its strengths include efficiency, the ability to handle missing data, and built-in regularization techniques to prevent overfitting. It is well-suited for this problem due to its capacity to capture subtle patterns and interactions in the data. However,

XGBoost requires careful hyperparameter tuning, and its computational demands can be significant, making it less interpretable than simpler models.

Together, these methods allowed us to explore both linear and non-linear relationships, evaluate their contributions to ride prediction, and develop a robust modeling framework tailored to the problem. Each method added unique insights and strengths, with random forest providing the best predictive performance among the three.

5. Results

To predict total_rides, we implemented and evaluated three advanced machine learning models, **Linear Regression**, **Random Forest** and **XGBoost**, along with detailed hyperparameter tuning to optimize performance. The models were trained on a dataset containing features such as weather conditions, temporal variables, station IDs and other predictors impacting ride demand. The performance of these methods was evaluated using Adjusted **R-squared** as the primary metric for the Linear Regression model, and **Root Mean Squared Error (RMSE)** as the primary metric for the Random Forest and XGBoost models.

5.1 Linear Regression Results

The linear regression model achieved an R-squared value of 0.1347, indicating that only about 13.47% of the variance in total bike rentals is explained by the predictors. Although most predictors were statistically significant, the low R-squared value highlights the inadequacy of a linear relationship in capturing the complexity of the data.

5.2 Random Forest Results

The Random Forest model was chosen for its ability to handle non-linear relationships and provide insights into feature importance. Key hyperparameters, such as the number of trees (ntree), were tuned during model training, which utilized bootstrapped sampling to generate multiple decision trees and averaged their predictions to reduce variance. The model achieved a test RMSE of 2.4317, demonstrating the strongest predictive capability and effective handling of feature interactions. However, Random Forest faced challenges, including computational intensity for larger datasets and limited interpretability. Additionally, hyperparameter tuning demanded substantial computational resources.

5.3 XGBoost Results

XGBoost, a gradient boosting algorithm, was selected for its advanced optimization techniques and ability to detect subtle patterns in the data. Extensive hyperparameter tuning was conducted, optimizing parameters such as max_depth (set to 3), min_child_weight (set to 1), eta (learning rate, set to 0.05), and nrounds (set to 100). Early stopping with a patience of 20 rounds was employed to prevent overfitting, and RMSE was used to evaluate each iteration. The final XGBoost model achieved a test RMSE of 2.478, underperforming the Random Forest model. Moreover, XGBoost required careful hyperparameter tuning, which was computationally expensive, and its interpretability was lower than simpler models like Random Forest and Linear Regression.

5.4 Evaluation Metric

The primary metrics used to evaluate model performance were R-squared (R^2) and Root Mean Squared Error (RMSE). R-squared (R^2) evaluates the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. While RMSE was used to evaluate non-linear models like XGBoost and Random Forest, R^2 was employed to assess the performance of the Linear Regression model and is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Equation 1: R Squared

RMSE measures the average magnitude of prediction errors in units of the dependent variable, making it particularly useful for regression problems as it penalizes larger errors more than smaller ones. Lower RMSE values indicate better model performance. Across Random Forest and XGBoost models, RMSE was computed as:

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

Equation 2 : Root Mean Squared Error

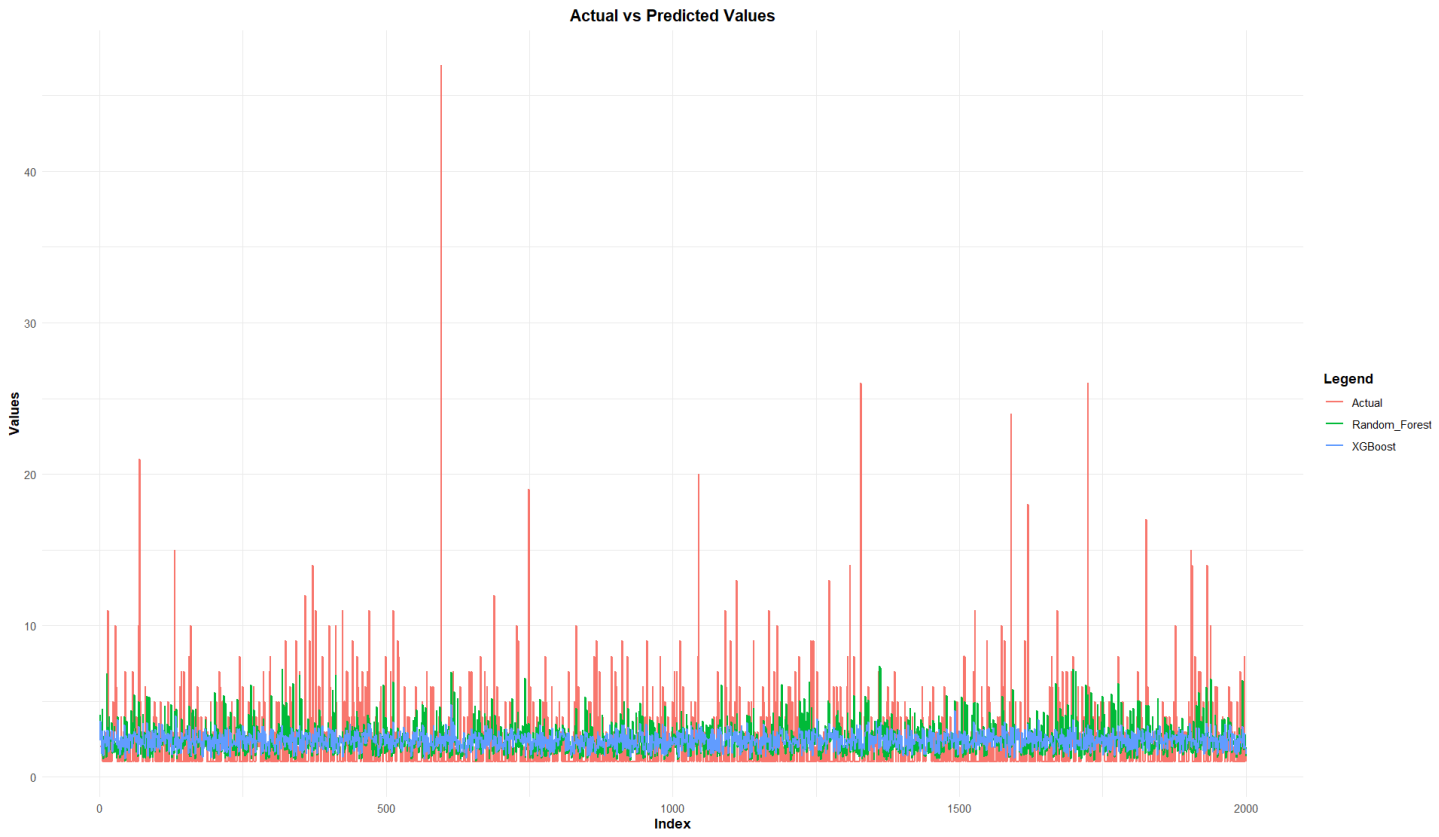


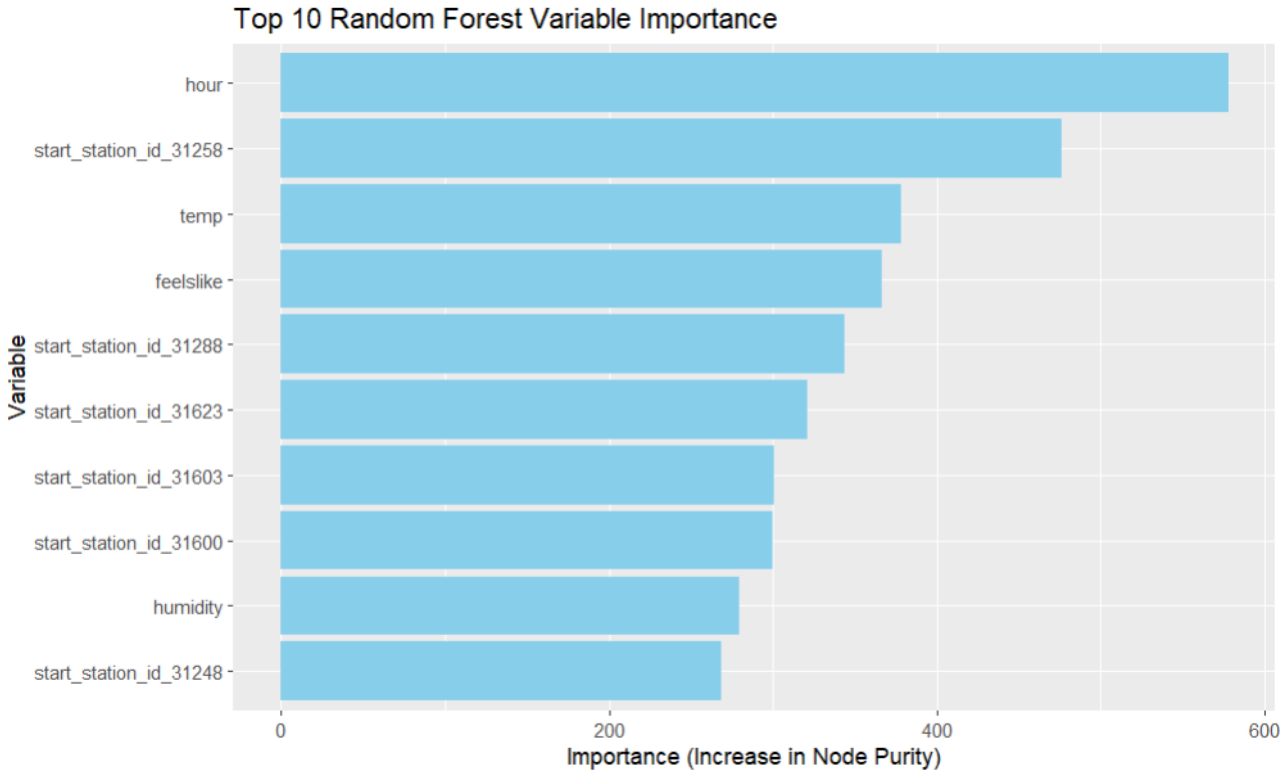
Figure 4: Actual vs Predicted

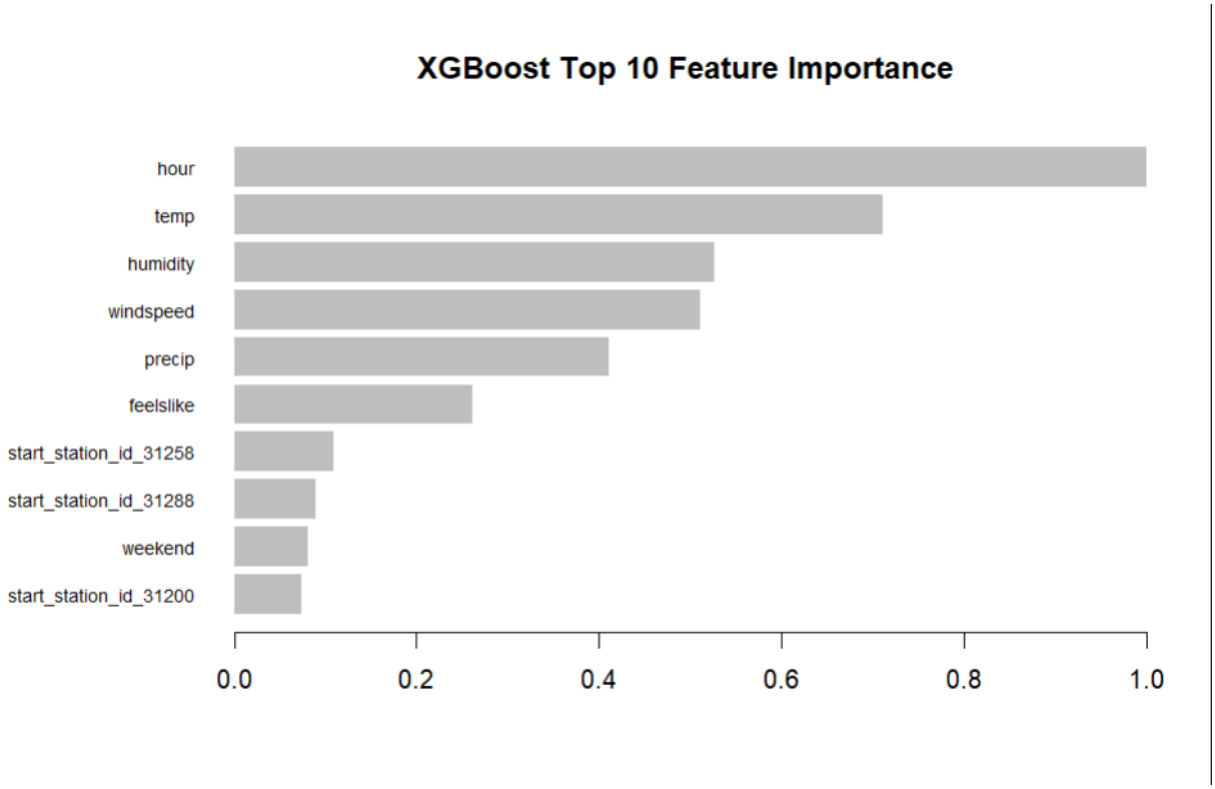
6. Challenges and Observations

The development process faced several challenges, including extensive hyperparameter tuning for the Random Forest and XGBoost models, which increased computational complexity and time requirements. The dataset's non-linear relationships and complex interactions necessitated advanced models like XGBoost to effectively capture these dynamics. Overfitting was another concern, addressed through techniques such as early stopping and cross-validation to ensure the models generalize well to unseen data. The final results highlight the importance of careful hyperparameter optimization and model selection for robust predictive performance. Among the three models, Random Forest emerged as the most effective, achieving the lowest RMSE and successfully capturing nuanced relationships in the data, as demonstrated by the 'Actual vs Predicted Values' chart.

7. Discussion: Results and Insights from the Bike Rental Prediction Models

The primary objective of this project was to predict total bike rentals using three predictive models: Linear Regression, Random Forest, and XGBoost. By analyzing patterns in the data and assessing the models' performance, we identified key insights that can inform operational decisions and strategic planning.





The analysis revealed the critical role of temporal factors, particularly the hour of the day, as the most significant driver of bike rental demand. Peak usage aligns with commuting hours, underscoring the importance of dynamically adjusting fleet availability to meet these surges. The inclusion of station-specific identifiers among the top predictors in both Random Forest and XGBoost models suggests that station characteristics significantly impact rental behavior, offering opportunities to refine fleet placement strategies. Weather variables, including temperature, humidity, and windspeed, remain vital predictors, with a non-linear relationship observed for temperature—demand rises in moderate weather but drops in extremes. Precipitation also has nuanced effects, highlighting the need for weather-responsive strategies such as dynamic pricing or providing protective gear.

Seasonal and day-of-week trends reveal consistently higher rentals on weekends and holidays, which could be leveraged through targeted promotions or discounts to enhance engagement. Conversely, demand drops sharply during winter, indicating a need for proactive strategies such as winter-specific marketing campaigns or incentivized pricing to sustain usage levels. Station-specific demand, as highlighted by the importance of station IDs, suggests opportunities for geospatial optimization, enabling Capital Bikes to address localized rental patterns and redistribute resources effectively.

While modeling was constrained to 10,000 rows due to computational power limitations, the models demonstrated robust predictive alignment with observed trends, particularly in periods of stable demand. However, discrepancies during sharp demand spikes suggest room for improvement by integrating contextual factors such as large-scale events or abrupt weather changes. These findings establish a strong foundation for operational improvements, enabling Capital Bikes to optimize fleet distribution, develop targeted strategies, and enhance customer satisfaction within the bike-sharing ecosystem.

7.1 Conclusion and Future Work

This report analyzed bike rental data in Washington, D.C., using predictive models to identify factors influencing demand and provide actionable insights for operational optimization. Key predictors included temporal factors like the hour of the day, weather conditions such as temperature and humidity, and day types like weekends and holidays. Advanced models such as Random Forest demonstrated superior accuracy (RMSE: 2.4317), outperforming XGBoost (RMSE: 2.478) and Linear Regression, the latter limited by its inability to capture the non-linear relationships in the data. Recommendations include dynamically adjusting bike resources to meet peak demand, implementing weather-responsive strategies, and launching targeted marketing campaigns to address demand fluctuations. Future enhancements could focus on expanding features with traffic and demographic data, developing real-time prediction systems, conducting geo-spatial analyses for station optimization, and performing user-level behavioral studies. These advancements would enable data-driven strategies to improve operational efficiency, customer satisfaction, and the overall sustainability of the bike rental system.

Contributions

I. **Data Collection and Cleaning:**
Genius Mpala, Bezel Ndhlovu, Johannes Machinya

II. **Data Analysis and Visualization:**
Genius Mpala, Bezel Ndhlovu, Johannes Machinya

III. **Model Development and Evaluation:**
Bezel Ndhlovu: Contributed to model building and testing, optimizing parameters for improved results.
Johannes Machinya: Focused on advanced tuning techniques and overall model evaluation. Both team members compared model performances to derive insights.

IV. **Report Writing and Presentation:**
Genius Mpala: Drafted sections on data preparation, analysis, and overall findings, while also organizing the final presentation.

Bibliography

1. Shaheen, S. A., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: Past, present, and future. Transportation Research Record, 2143(1), 159–167.

2. Faghih-Imani, A., Anowar, S., Miller, E. J., & Eluru, N. (2014). Heterogeneity in bike-sharing systems: A GPS-based analysis of temporal and spatial dynamics. Transportation Research Part C: Emerging Technologies, 47, 66–84.

3. Li, Y., Zheng, Y., Zhang, H., & Chen, L. (2015). Traffic prediction in a bike-sharing system. Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, 1–4.

4. Zhao, Z., Deng, X., Zhang, K., & Lu, F. (2019). Deep learning for bike-sharing demand prediction: Models, methods, and results. Neural Computing and Applications, 31(6), 1575–1591.

5. Zhu, M. (2022). Short-term prediction of bike-sharing demand using multi-source data: A spatial-temporal graph convolutional network approach. Applied Sciences, 12(3), 1161.

6. April, Y., & Parvatia, M. S. (2021). Regression Analysis Final Project: Predicting Bikes Rental Demand Using Weather and Holiday Season. RStudio Publications.

7. Capital Bikes Data. Provided datasets including bike rental details, weather metrics, station lists, and usage frequency data for the Washington, D.C., bike-sharing program.

8. RStudio Documentation. R Programming for Data Science. Available at: https://r4ds.had.co.nz
9. XGBoost Documentation. XGBoost: Scalable and Flexible Gradient Boosting. Available at: https://xgboost.readthedocs.io

10. Microsoft Excel and PowerPoint for dataset cleaning and visualizations.

11. IEEE Xplore. An Effective Data-Driven Approach to Predict Bike Rental Demand. https://ieeexplore.ieee.org

12. SpringerLink. Bike Sharing Usage Prediction with Deep Learning: A Survey. https://springerlink.com

13. ND.edu Academic Database Access for reference and background materials on transportation modeling.

Appendix

Refer to attached R scipt