Comparing Multiple Linear Regression (MLR) and Random Forest Regression (RFR) in Predicting Bike Rental

Introduction

Bike rentals are increasingly popular in many big cities nowadays due to them being cheaper, their potential to reduce traffic congestion and lower pollution. There are over 500 bike-sharing programs worldwide with a new generation of bike sharing systems automate the whole processes of renting and returning bikes; bringing many benefits to residents in urban areas such as health improvement and social connections [1]. However, understanding and predicting bike usage patterns is crucial for optimizing system operations and improving user satisfaction. In this study, we aim to find the factors that boost bike rentals. The dataset is sourced from *UCI Machine Learning Repository* [2] about Capital Bikeshare in Washington DC. Multiple Linear Regression and Random Forest Regression are compared in predicting bike usage.

Table 1: Descriptive Statistics

	Mean	Min	Max	Std
Temperature	0.497	0.02	1	0.193
Humidity	0.627	0	1	0.193
Windspeed	0.190	0	0.851	0.122
Casual Rider	35.676	0	367	49.305
Registered Rider	153.787	0	886	151.35
Total of Riders	186.463	1	977	181.38

Exploratory Analysis

- The dataset is made up of 17389 instances and 13 features which include temperature which was normalized. The total of bike users per hour is given as well as how many are registered users or casual riders.
- Table 1 provides some useful descriptive statistics of the features. Minimum normalised temperature is 0.02 (-8°C) and maximum temperature is 39°C. The maximum of bike riders recorded in an hour is 977.
- In Figure 1, a positive trend can be observed between temperature(normalized) and number of bike rentals. The majority of rentals is observed under clear/partly cloudy weather condition. Warmer weather conditions increase ridership while adverse weather conditions significantly reduce bike usage.
- In Figure 2, season trends can be analysed. Summer has the highest average bike rentals, followed closely by Fall. There is more variability for bike rentals in weather condition in Spring.
- In Figure 3, the positively skewed distribution of bike rentals is observed. Log transformation or standardisation may be beneficial in modelling.
- In Figure 4, hourly bike rentals is analysed. Peaks can be observed at 8 a.m. and 5 p.m., suggesting more bike rentals during peak hours.
- In Figure 5, correlation heatmap, temperature is the highest correlated to count of riders. Humidity is negatively correlated to the count of bike rentals.

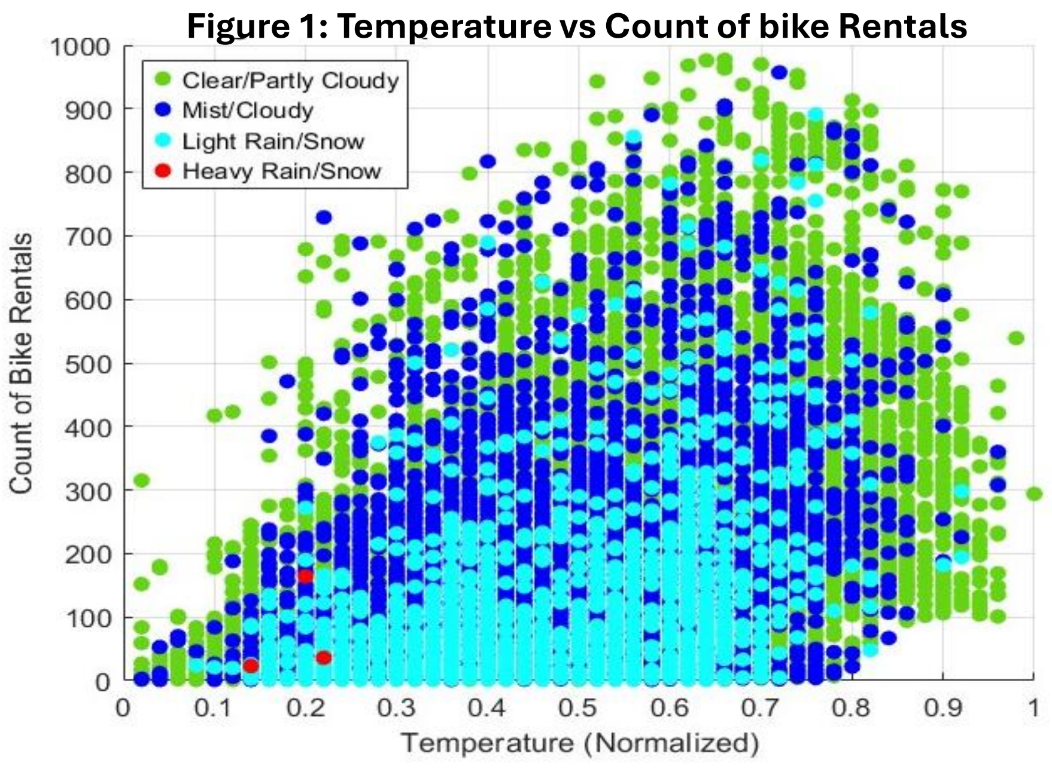


Figure 1: Temperature vs Count of bike Rentals

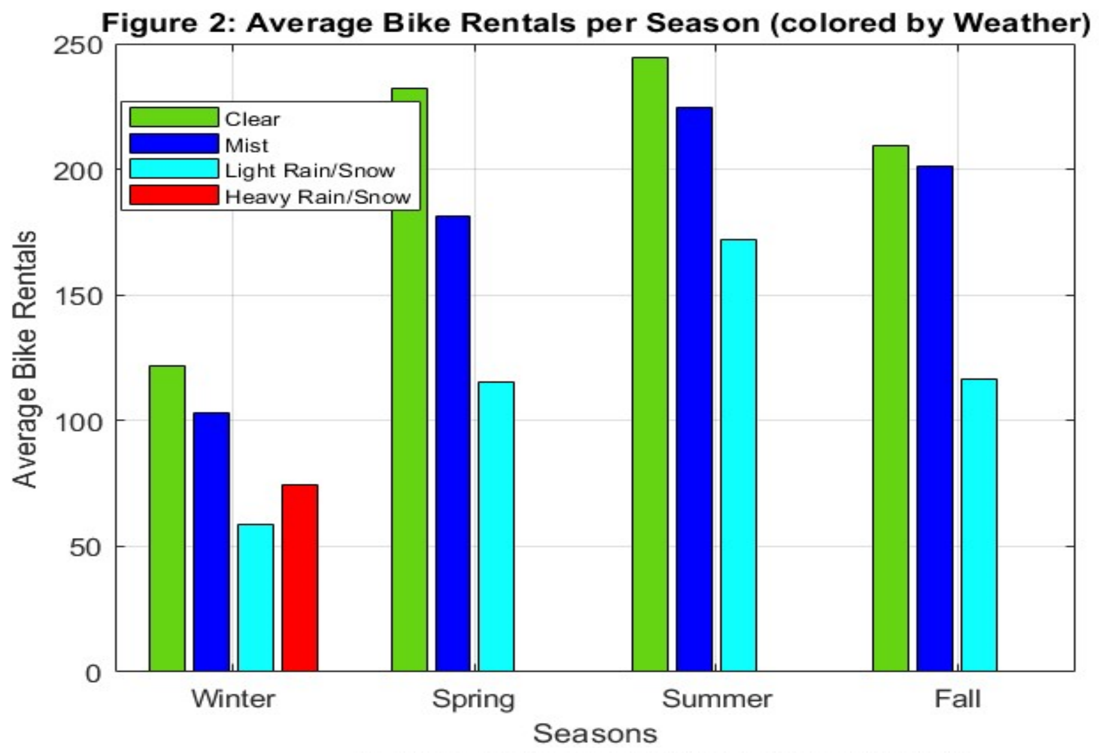


Figure 2: Average Bike Rentals per Season (colored by Weather)

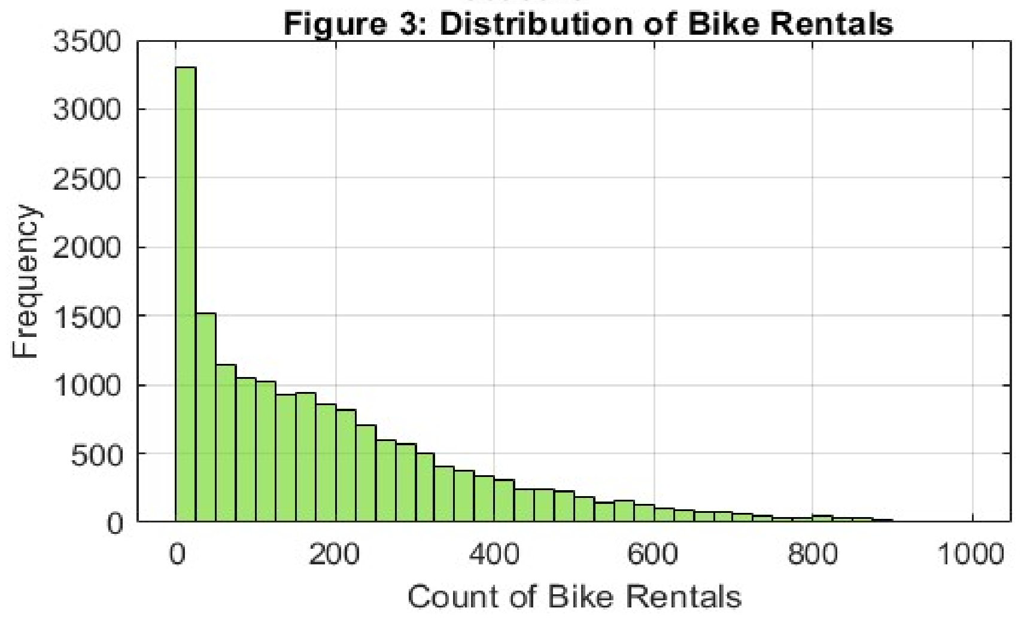


Figure 3: Distribution of Bike Rentals

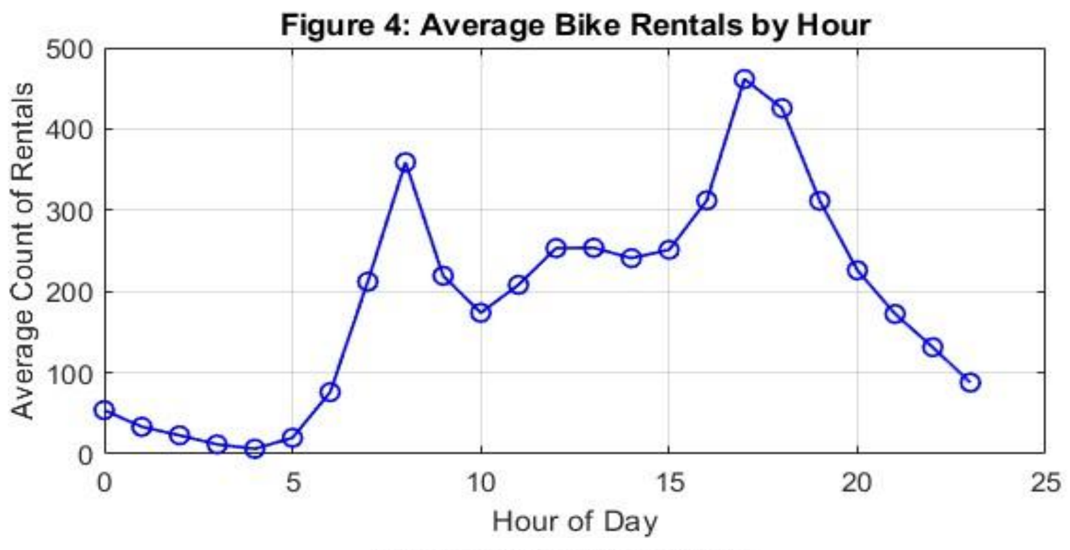


Figure 4: Average Bike Rentals by Hour

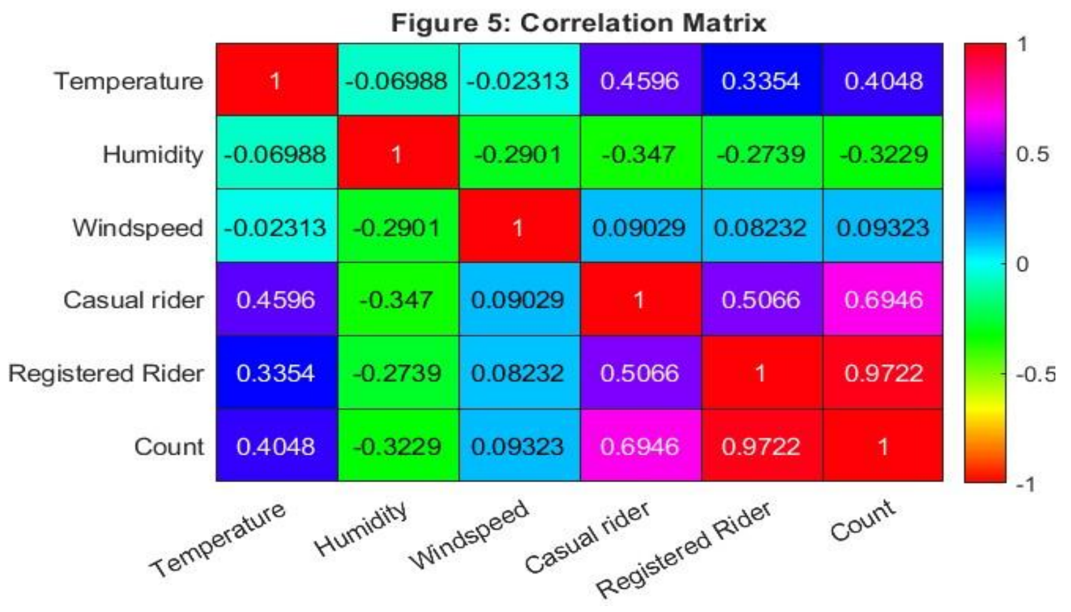


Figure 5: Correlation Matrix

MULTIPLE LINEAR REGRESSION (MLR)

Multiple Linear Regression is a statistical model to identify the relationship between a single dependent variable and multiple independent variables simultaneously. This method will provide a straightforward and interpretable way to analyse how changes in features affect bike rentals.

PROS

- Multiple Linear Regression is straightforward to implement, and the results can be easily interpreted.
- Multiple Linear Regression efficiently handles and analyses large datasets.
- The model can have a high predictive accuracy when the right features are implemented.
- We can assess the significance of each predictor and by how much the target variable is affected.

CONS

- Risk of overfitting if too many variables are included in the model.
- The key assumptions of the model such homoscedasticity and normality of residuals should be met to avoid inaccurate results. The parameters should be tuned carefully to obtain unbiased results.
- Multicollinearity issues if independent variables are not selected carefully and if the proper checks are not done.

RANDOM FOREST REGRESSION (RFR)

Random Forest Regression is an ensemble and supervised machine learning model to predict continuous numerical outcomes by building many decision trees and outputting the average prediction of the individual trees. This model implements bootstrap aggregation which reduces variance and improve accuracy. [3]

PROS

- The randomness in data sampling and feature selection helps prevent overfitting.
- Individual trees may be sensitive to noise, but averaging their predictions reduces this sensitivity.
- This model is versatile and can be used with both numerical and categorical features.
- The ensemble approach yields better results than a singular decision tree.

CONS

- The model is often considered a "black box," making it difficult to interpret individual predictions compared to linear regression models.
- A dataset with a lot of noise may lead to the model fitting the noise despite the ensemble approach.
- This model does not incorporate known relationships between attributes and targets
- Deep decision trees may be preferred but this can be resource intensive.

HYPOTHESIS STATEMENT : Multiple Linear Regression and Random Forest Regression will both provide interesting results. Adjustments may be needed to certain variables to reduce randomness and noise to generate greater predictive performance in MLR. Random Forest Regression is expected to outperform Multiple Linear Regression since Random Forest is more flexible and robust to outliers and non-linear relationships. Multiple Linear Regression will produce more interpretable output, the significance of each parameter can be analysed and is expected to be less computationally taxing.

METHODOLOGY

The data is split into a 70:30 ratio for testing and training with 10-fold cross-validation to evaluate model robustness and mitigate overfitting. Ridge and Lasso Regressions are explored as regularization against multicollinearity. Polynomials and interaction terms are used to find any improvement or non-linear relationships in the MLR model. Stepwise regression is applied to identify significant relationships among predictors. The RFR model is tuned using grid search. Feature importance analysis is conducted to find significant variables in predicting bike rentals. To compare the two models, RMSE and R-squared are used.

CHOICE OF PARAMETERS AND EXPERIMENTAL RESULTS

Multiple Linear Regression: Enhanced MLR model with interaction terms outperformed other experimental MLR models and the baseline MLR model. The two most significant interaction terms are identified through stepwise regression. Stepwise regression is the procedure of selecting independent variables yielding the highest correlation with the dependent variable [4]. The two interaction terms that have been added to the MLR model are: 'temperature * workday and 'temperature * peak hour'. A comparison between actual and predicted values can be seen in figure 7.

Experimental Results

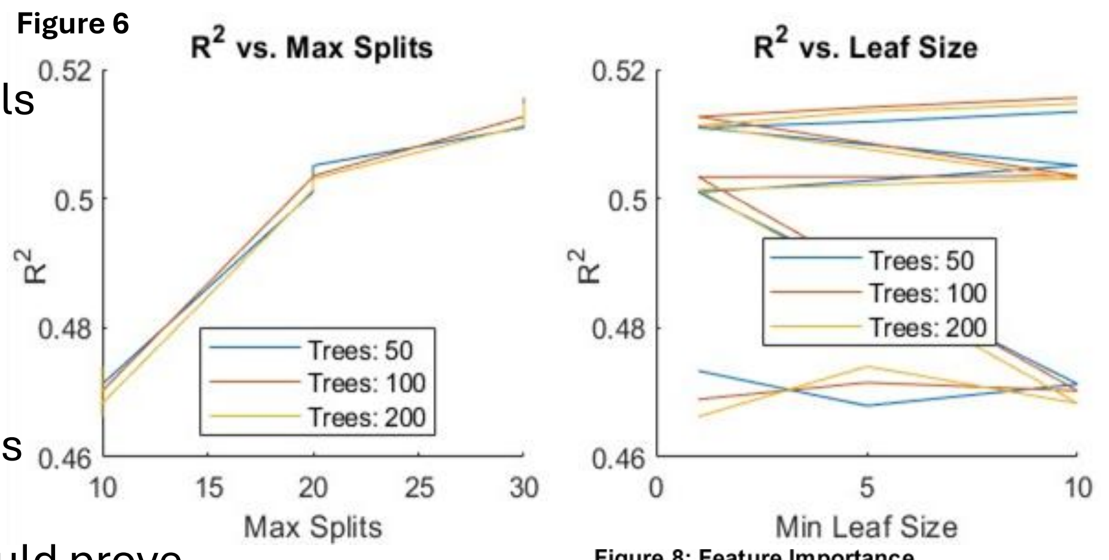
- Log-transformation has been applied to the dependent variable since a positive skewness was observed however this led to a lower r-squared compared to the baseline model.
- Ridge regression has not improved the model but further experiment with different regularization hyperparameter could prove otherwise, but ridge regression performed worse in cross validation since the dataset does not exhibit severe multicollinearity.
- Polynomial terms in the MLR model have led to rank deficiency and overfitting which lowers accuracy. No improvement in RMSE and R^2 were observed.

Random Forest Regression: The baseline RFR model(100 trees and 10 splits), fitted by bagging, led to relatively high root mean squared error (RMSE). Grid Search Analysis has been performed to tune the following hyperparameters: Number of trees in the ensemble, Maximum Splits and Minimum Leaf Size. Grid search optimised RFR model, but at high computational costs.

Experimental Results

- The importance of each features can be identified in Figure 8, 'peak hour' has a high importance score followed by 'temperature', aligning with domain expectations.
- From Figure 6, it can be observed that increasing the number of splits improves R^2, suggesting a better model fit as the tree depth increases. R^2 fluctuates as leaf size changes, no clear trend can be identified.
- The best model in grid search, with R^2 being 0.515, used 100 trees, a maximum splits of 30 and a minimum leaf size of 10. In figure 9, it can be observed that around 30 splits, improvements in R^2 plateau. Higher values for splits and number of trees do not significantly improve the model after a point. The best performance occurs around 25-30 for max splits and 100-150 for number of trees.

Figure 6



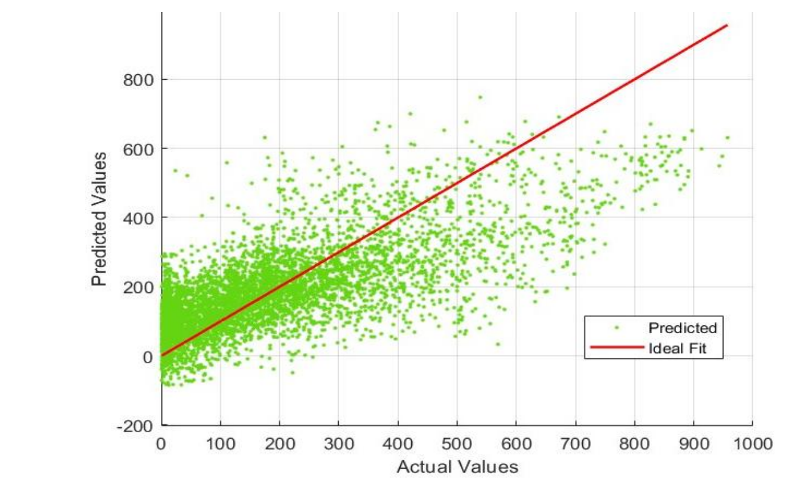


Figure 7: Actual vs Predicted values (MLR interaction terms test set)

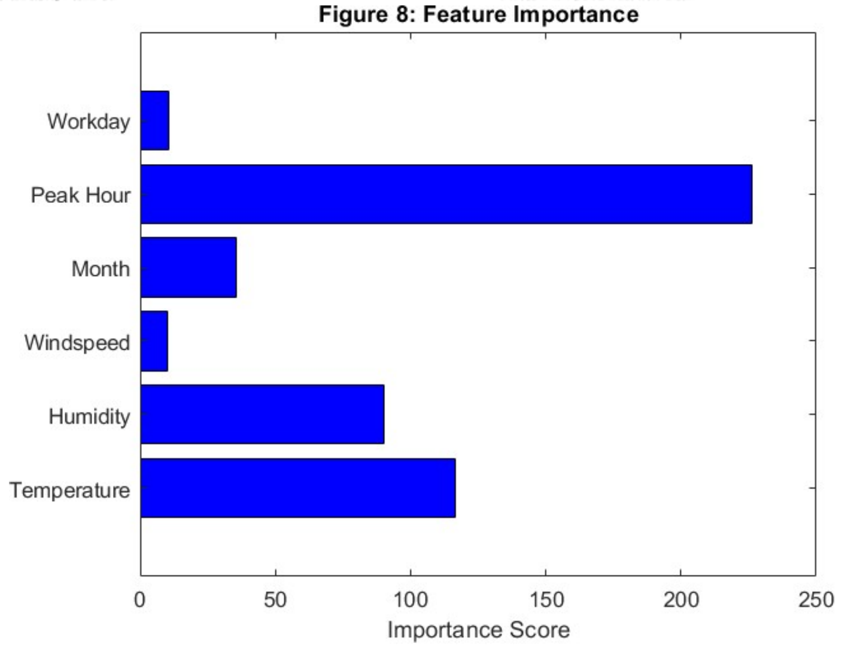


Figure 8: Feature Importance

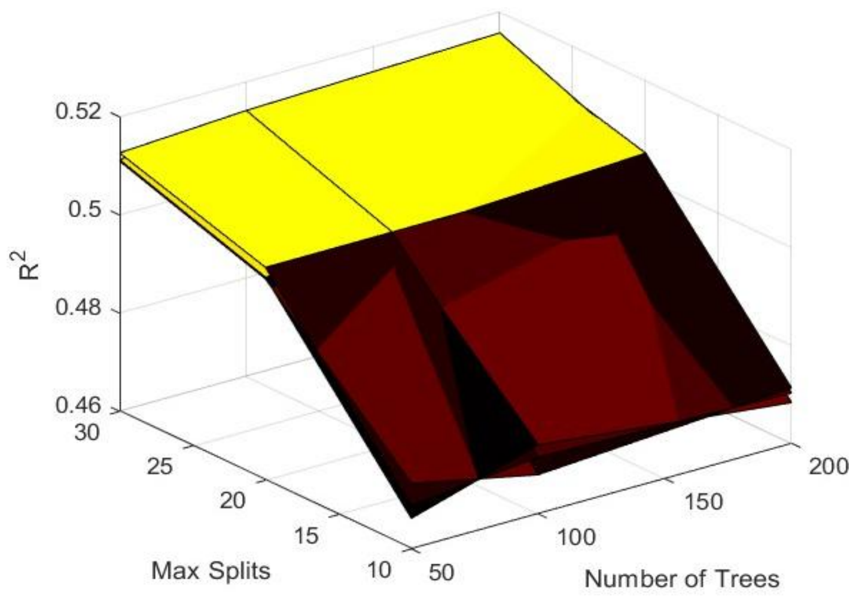


Figure 9: 3D Display of Hyperparameter Tuning

RESULTS ANALYSIS AND DISCUSSION

- During cross-validation, the R^2 of the MLR model was slightly higher compared to the RFR model, indicating that MLR explain marginally more variance than RFR in training data. On the other hand, RMSE were comparable, suggesting that both models had similar predictive accuracy.
- During testing, the MLR model with interaction terms again slightly outperformed the optimised RFR model (Table 2). MLR proves to be more advantageous for this dataset.
- MLR was significantly faster and tasks 95% faster than RFR.
- MLR provides interpretable coefficients and significant interaction between features were identified. RFR's robustness to non-linear relationships performed well on unseen data despite minimal preprocessing and showed us the most important features.
- In Figure 10, the REC curve indicates that prediction accuracy from both models are similar at different tolerance levels. RFR slightly outperforms MLR at higher tolerance levels, probably due to the robustness of RFR in handling variability and outliers. It has also been observed in the residual plots that MLR struggled to capture non-linear relationships while RFR's ensemble approach can solve this issue but at higher computational costs.
- The dataset was previously used in a study for Anchor Regression, which is an interpolation between OLS and TSLS (Two-Stage Least Squares). This method showed improvement in prediction stability and robustness compared to OLS [5]. This method is worthy further exploration for this dataset since it is particularly beneficial in scenarios where there are temporal factors.
- This study main focus was to compare machine learning techniques in prediction, however the results such as the main determinants of bike rentals could be implemented when designing such a system in similar urban cities.

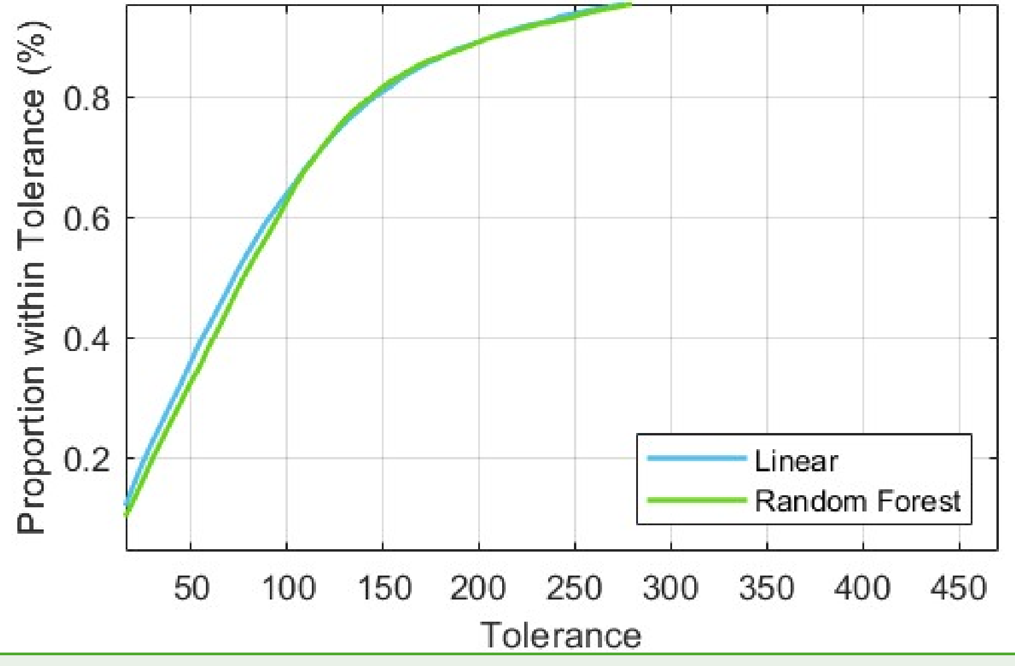


Figure 10: Regression Error Characteristic Curve

Table 2: Performance Metrics

	TRAINING		TESTING		
	R^2	RMSE	R^2	RMSE	Time taken
MLR	0.517	125.70	0.527	125.49	0.243s
RFR	0.512	126.44	0.514	127.15	5.901s

LESSONS LEARNED

- When implementing machine learning models, a selected number of features believed to explain the target variable should be initially used. New additional features should be added until no further improvement is observed in performance metrics. This is especially true for linear regression model which require adjustments to variables and many combination of linear equations.
- Dummy variables should be selected carefully as this can lead to sparsity; feature selection or regularization should then be used.
- Hyperparameters should be tuned with care to avoid overfitting.

FUTURE WORK

- Use other machine learning models such as SVM Regression or other versions like Anchor Regression which may be better fit for this dataset .
- Random forest could further be improved through parameter selection, eliminating features with less predictive power.
- Log-transformation across all variables could yield more accuracy and higher predictive power as some variables were skewed.
- Consider time-based cross validation if temporal dependencies have been observed and to avoid data leakage.

REFERENCES

[1] Fanaee, H. (2013) Bike sharing, UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset (Accessed: 01 November 2024).
[2] Fifteen (2023) From cars to bikes: The impact of bike-sharing on cities and their residents, Fifteen. Available at: https://fifteen.eu/en/resources/blog/from-cars-to-bikes-the-impact-of-bike-sharing-on-cities-and-their-residents (Accessed: 15 November 2024).
[3] Trehan, D., n.d. Why Choose Random Forest and Not Decision Trees. Towards AI. Available at: https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees?utm_source=chatgpt.com [Accessed 30 November 2024].
[4] Lewis-Beck, M.S., 1978. Stepwise regression: A caution. Political Methodology, 5(2), pp.213–240. Cambridge University Press on behalf of the Society for Political Methodology. Available at: https://www.jstor.org/stable/25791533
[5] Rothenhäusler, D., Meinshausen, N., Bühlmann, P. and Peters, J., 2024. Anchor regression: heterogeneous data meet causality. [online]

Sajjaad Jurawon – 2024