Project: Bike Rental Prediction Using Random Forest

Introduction This document outlines the development and evaluation of a predictive model for daily bike rental counts. The project aimed to leverage environmental and seasonal factors to forecast bike rental demand.

Data Overview: The dataset comprised several variables, including:

* **Season**: Categorized into spring, summer, fall, and winter.
* **Year**: Indicated by 0 (2011) and 1 (2012).
* **Month**: Ranging from 1 to 12.
* **Holiday**: Binary indication of a holiday.
* **Weekday**: Numbered from 0 (Sunday) to 6 (Saturday).
* **Weather Situation**: Categorized from clear conditions to heavy rain.
* **Temperature**: Normalized readings in Celsius.
* **Feeling Temperature**: Normalized perceived temperature.
* **Humidity**: Normalized percentage.
* **Wind Speed**: Normalized values.
* **Count**: Total number of bikes rented.

Initial Approach: Initially a multiple linear regression model was constructed, which identified ‘casual’ and ‘registered’ as significant predictors. However, due to their high correlation with the target variable, they were excluded from subsequent models to avoid multicollinearity.

Challenges Encountered: The linear regression model encountered issues with negative predictions, which are not plausible in the context of bike rental counts. This led to the decision to explore alternative modelling techniques.

Model Development: The Random Forest algorithm was selected for its robustness and ability to handle non-linear relationships. The model focused solely on environmental and seasonal variables, excluding user type indicators.

Data Preparation and Analysis: The data was imported into R Studio and underwent preliminary analysis. It was then divided into training and test sets to facilitate model training and evaluation.

Model Training and Testing: A Random Forest model was trained using the training set. The model’s performance was assessed through predictions on the test set, followed by the creation of visual plots for further analysis.

Model Evaluation: The model achieved an R-squared value of 0.8860, indicating a strong fit to the data. Although the Mean Squared Error was 453616.4, this was deemed acceptable given the wide range of the ‘count’ variable. Visualizations confirmed a good correlation between actual and predicted counts, particularly for higher values. The residuals were predominantly centred around zero, suggesting accurate predictions.

Conclusion: The project successfully met its objective to predict daily bike rental counts based on environmental and seasonal factors. The Random Forest model demonstrated strong predictive capabilities, aligning with the project’s goals.