Bike Sharing Demand Forecasting Report

Methods and Approach

The objective of this project was to forecast bike rental demand in Washington, D.C. based on historical usage data and corresponding weather information. To approach this problem, I had several hypotheses were regarding factors likely to influence the bike rentals, such as temperature, time of day, working days versus weekends, and varying weather conditions.

The training and test datasets were loaded and combined to create consistent exploration across all the data. During preprocessing, new features such as "hour," "day," and "month" were created from the "datetime" column to better capture time-dependent patterns. Irrelevant columns like "casual" and "registered" were dropped, as they were not available in the test set and the prediction target was the total "count."

Exploratory Data Analysis (EDA) was performed using visualizations such as histograms, boxplots, and count plots, taht revealed clear patterns. For example, bike rentals peaked during commute hours on working days and declined during unfavorable weather conditions. Numerical features were scaled using StandardScaler to optimize model performance, especially for linear models. Two models were trained: a Linear Regression model to establish a baseline and a Random Forest Regressor to capture more complex, non-linear patterns present in the data.

Results and Analysis

Model performance was evaluated using Root Mean Squared Log Error (RMSLE.

- **Linear Regression** provided a reasonable baseline but struggled to capture the complex interactions in the dataset, therefore it had a higher RMSLE.
- **Random Forest Regressor** had better performance with a lower RMSLE, but still modeling the non-linear relationships influencing rental behavior.

Feature importance analysis from the Random Forest model indicated that "hour," "temperature," "feels like temperature" (atemp), and "humidity" were the most critical predictors. This aligns with logical expectations, as the rental counts are sensitive to the time of day and the weather conditions.

Conclusion

- Random Forest outperformed Linear Regression with a lower RMSLE.
- Important features included 'hour', 'temp', 'atemp', and 'humidity'.

- Linear Regression is a **parametric** model (that assumes linearity), Random Forest is **non-parametric** (no assumptions).
- Bike rental count is highly dependent on time of day, temperature, and working days

I experimented with two models: Linear Regression and Random Forest Regressor. Linear Regression was chosen because it is simple, interpretable, and works well when there is a clear linear relationship between input features and the target variable. Random Forest was chosen for its ability to capture non-linear patterns and its robustness to outliers. After training and evaluation, Random Forest outperformed Linear Regression by achieving a lower RMSLE score, showing that it is a better predictive performance on this dataset. Linear Regression is a **parametric** model as it assumes a specific form for the data, whereas Random Forest is a **non-parametric** model that does not make strong assumptions about the data. Based on feature importance from Random Forest, the influencing features are 'hour', 'temp', 'atemp', and 'humidity'. Overall, this project hihglighted the importance of feature engineering and selecting flexible models when dealing with complex datasets.