Bike Sharing Dataset Analysis and Prediction Report

Part 1: Exploratory Data Analysis and Prediction Model

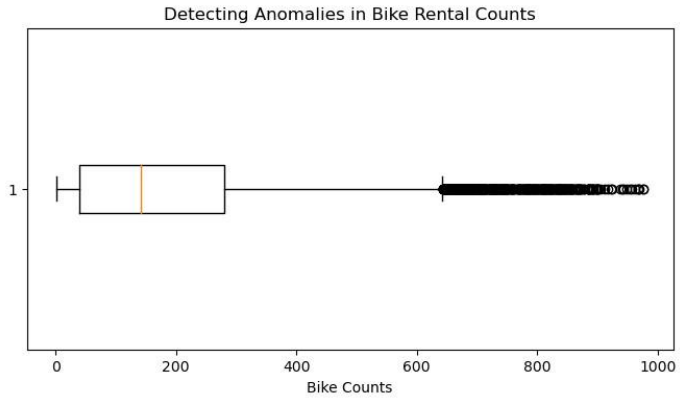
Introduction

This report presents an analysis of the Bike Sharing Dataset from the UCI Machine Learning Repository. The dataset contains information about bike rental counts, weather conditions, and various features related to bike sharing.

Data Analysis

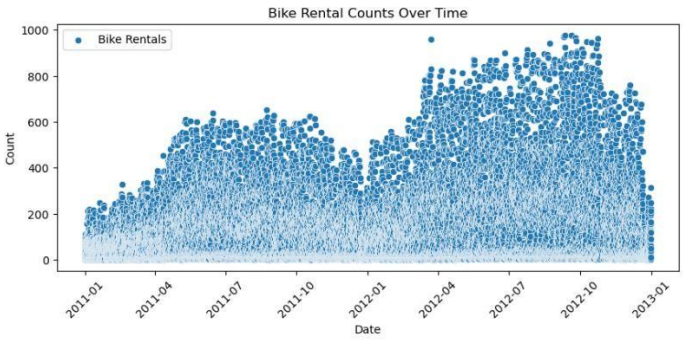
Anomalies in Bike Rental Counts

When plotting a boxplot, I observed anomalies in bike rental counts, with values ranging from approximately 630 to 977.



Yearly Trends

I noticed an increase in bike sharing counts in 2012 compared to 2011, indicating a growing trend in bike rentals.



Hourly Patterns

Hourly analysis revealed that bike sharing was less popular from 4 am to 5 am but surged during the evening from 5 pm to 6 pm, suggesting commuter patterns.

Monthly and Seasonal Patterns

Monthly analysis showed that bike sharing was less in January and peaked in September. Additionally, I found that bike sharing counts were lower in the spring season but significantly higher in the fall season.

Weather Impact

The weather situation had a significant impact on bike sharing. Clear and partly cloudy weather conditions resulted in higher bike sharing counts, while heavy rain and foggy conditions led to lower counts.

Model Selection

For our predictive model, I chose the Random Forest Regressor for several reasons:

- Ensemble Learning: The Random Forest model leverages ensemble learning, combining multiple decision trees to make predictions. This approach reduces overfitting and improves generalization.

- Feature Importance: Random Forest provides feature importance scores, aiding our understanding of factors affecting bike rental counts.

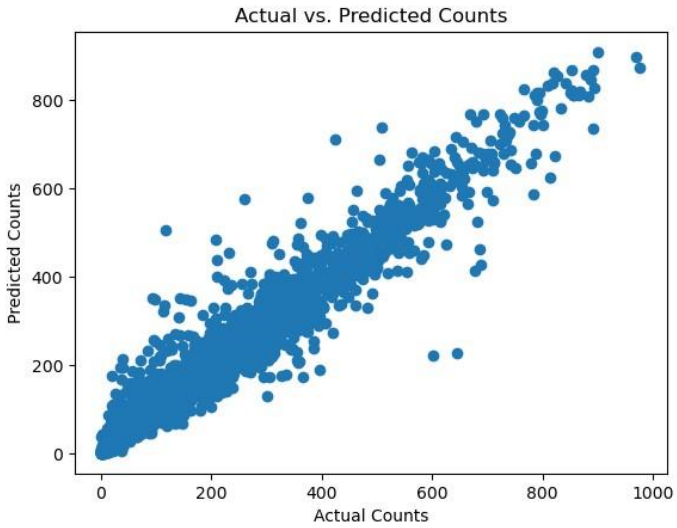
- Non-Linear Relationships: The model captures non-linear relationships between features and the target variable.

- Robustness: Random Forest is robust to outliers and versatile for diverse datasets.

I also considered the XG Boost model but ultimately selected Random Forest due to its faster training time and comparable results.

Model Evaluation

The Random Forest Regressor model performed well with a Mean Squared Error (MSE) of 1730.80 and an Rsquared (R2) value of 0.95, indicating a good fit to the data.



Future Considerations

When deploying the code for daily prediction services, consider regular maintenance, scalability, monitoring, documentation, and security measures to ensure the model's performance and data integrity.

This report provides a concise overview of the analysis and code development process for building a prediction model for bike rental counts, considering the characteristics and insights derived from the Bike Sharing Dataset. The Random Forest Regressor was chosen due to its compatibility with the dataset, and it delivered promising results.