Capital Bikeshare System

Jen Kelleman | Carnegie Mellon | Data Science Capstone | Feb 2025

Project Overview

This project analyzes the **Capital Bikeshare System** in Washington, D.C., using data from 2011 to 2012. The primary goal is to understand bike rental patterns and predict future bikeshare system usage.

Dataset Characteristics

The dataset includes 14 variables, across 17,380 rows. Data is collected in two-hour intervals over a two-year period. The dataset provides rich metrics on 3 main categories, including user demographic (registered vs. casual use), weather conditions, and time-based variables.

Research Questions and Goals

1. Environmental and Seasonal Factors: How do these factors influence hourly and daily bike rental counts?
2. Future Rental Bike Prediction: Can the number of bike rentals be predicted based on weather-based variables? Focusing on humidity, and windspeed.

Exploratory Data Analysis (EDA)

As part of the the EDA analysis, I've cleaned and transformed the data. After removing N/A and unnecessary rows, I've retained 14,877 observations. To further isolate the variable impact, I've performed feature engineering and data enrichment, adding 11 new variables to the existing dataset (total: 25)

- **Hourly Bike Rentals (Heatmap)**: Shows peak hours during morning and evening rush hours (i.e., 7am-9am and 5pm-7pm),
- **Weather Conditions & Bike Rentals (Scatter Plots)**: Indicates a clear relationship between favorable weather conditions and increased bike rentals.

Time of Year & Bike Rentals (Line Graph): Reveals seasonal trends with peaks during warmer months and unexpected high rentals in December.

- **Days of the Week & Bike Rentals (Bar Chart)**: Shows slightly higher rentals on weekends.

Prediction Modeling

The predictive variables I tested were Humidity, Wind Speed, and Temperature. The response variable was Bike Rental Count.

Linear Regression Model

Key metrics include:

- **R-squared**: 0.23, indicating that approximately 23% of the variability in bike rental counts can be explained by the predictors.
- **p-values**: 2.2e-16, indicating that the predictors are very significantly associated with the bike rental count.
- **Pearson correlation:** Humidity = -0.31, Temp = 0.39, Wind Speed = 0.89
- **Spearman correlation:** Humidity = -0.36, Temp = 0.41, Wind Speed = 0.12

Model Selection, Tuning, Outliers

Given the low r-square value of the linear regression model, I explored different machine learning algorithms, such as PCA, random forest, and gradient boosting. With hyperparameter tuning and PCA modeling, I found 100% variance in bike rental usage. Major success!

Principal Component Analysis (PCA) Model

Using a PCA helped to reduce the dimensionality and transform this large set of variables into a smaller subset of uncorrelated variables called principal components. Outliers were heavily skewing the data, so I tuned the hyperparameters. I removed outlier data points (i.e., days with bikes rentals over 800 bikes). Key metrics include:

- **PC3:** 100% variance of the data can be captured by PC3.
- **Standard deviation:** PC1 explains approximately 43% of the data. PC2 explains 34% of the data. PC3 explains 23% of the data.
- **p-values**: 2.2e-16, indicating that the predictors are very significantly associated with the bike rental count.
- **Coefficients:** PC1 = 47.68, PC2 = -67.53, PC3 = 19.06
- **Residuals**: RMSE = 153.805. This value indicates the average magnitude of the errors between the predicted and actual bike rental counts.

Main Conclusions and Takeaways

Summary of insights into predicting future bikeshare system usage. Knowing these insights can help to optimize bike availability and improve the service efficiency.

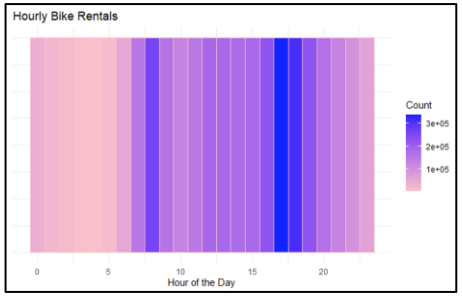
- **Time-based:** Bike rental usage peaks during commuting hours. Specifically, 7am-9am and 5pm-7pm.
- **Weather-based:** Higher temperatures, lower windspeed, and lower humidity all generally correspond to an increase in bike rentals.
- **Seasonality:** Bike rental usage peaks during the summer months (May, June, July, September).
- **Potential limitations:** No information on the exact location of the bikeshare systems, no user type segmentation (e.g., commuters vs. leisure users), and ambiguities in the data.
- **For future explorations,** I recommend during a peak hours analysis, demand forecasting, and operational adjustments.
- **Peak hours analysis:** Conduct a detailed analysis of peak hours to understand the factors driving high demand during these times. Look at commuter patterns, in particular. **Demand forecasting:** Use the hourly data to build predictive models for bike rental demand. **Operational adjustments:** Investigate the operational aspects, such as bike availability and maintenance schedules, to ensure that bikes are available at peak hours.

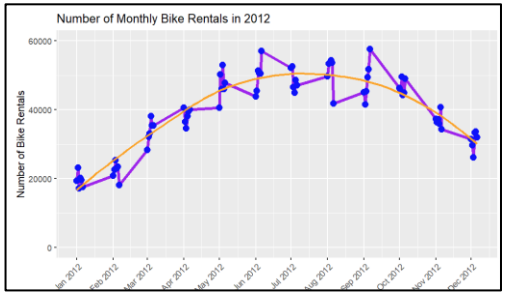
Capital Bikeshare System

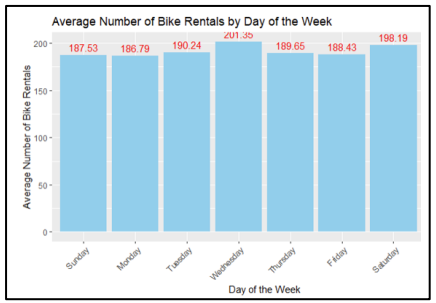
Jen Kelleman | Carnegie Mellon | Data Science Capstone | Feb 2025

Appendix

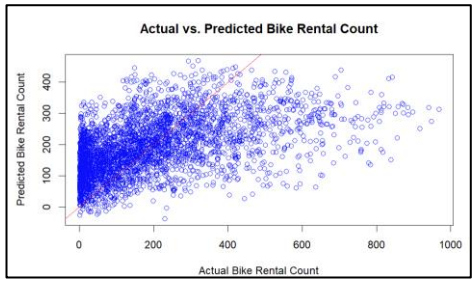
Exploratory Data Analysis

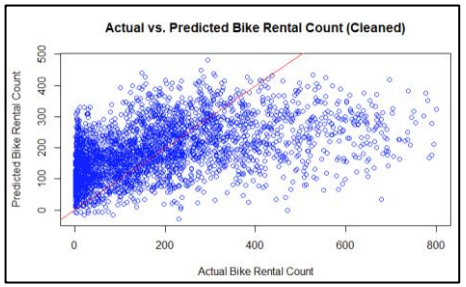


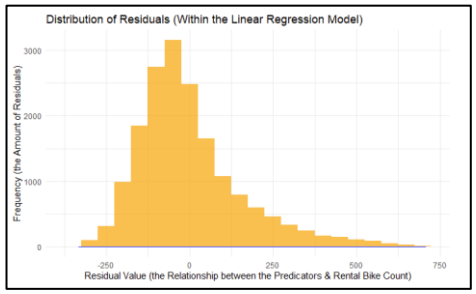


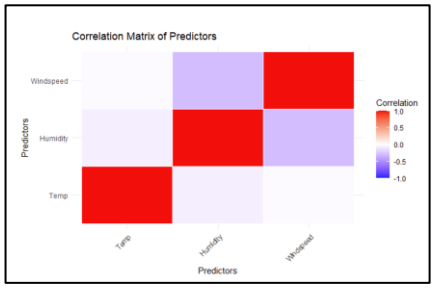


Linear Regression Model

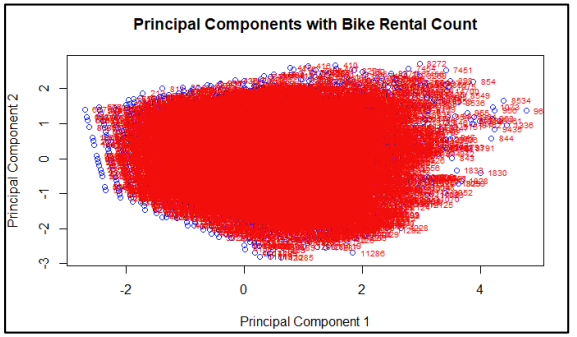








PCA Model



Cook's Distance of the residuals