INFO6105 40534 Data Sci Eng Methods SEC 16 Spring 2025 [BOS-2-TR]

Final Project Report

By: Sameeth Kumar Goud Talla

NU ID: 002810105

Title: Analysing How Weather Impacts Daily Bike Rental Demand

Video link :

https://drive.google.com/file/d/1kIbYJtS5lWICANJJj-P2kPCIt0vBgTmt/view?usp=sharing

Introduction

The rapid growth of urban bike-sharing programs necessitates precise demand projections and utilization analyses. Not standing with the fact that these systems provide a greener alternative to traditional modes of transportation and their successful implementation requires careful planning in order to meet investor demand. This project examines the effects of different weather and calendar-based variables on the daily total number of bike rentals. Data used in this study comprised daily rental data for two years in Washington, D.C., from the UCI Machine Learning Repository.

The project main goal is to use R's and data science tools to forecast how bike rental behavior will be change with the weather. In order to understand how temperature, humidity, windspeed, and other contextual elements influence rental patterns. we will construct several statistical models and conduct exploratory study. These results are useful for transportation authorities and urban planners looking for data - driven ways to maximize shared mobility services, in addition to being academically enlightening.

Dataset

Source - Machine Learning Repository at UCI

Used File - day.csv

Observations - 731 days

Target Variables:

- cnt (continuous) - Total no of bike rentals per day
- high_demand (binary) - Personalized classification label according to whether the number of rentals per day exceeds the median

Feature variables

Temp - normalized temperature

Humidity - normalized

Windspeed - The normalized wind speed

Seasons are classified as follows -

- 1 spring,
- 2 summer,
- 3 fall, and
- 4 winter.

weathersit -

- 1 means clear,
- 2 means mist, and

- 3 means rain or snow.

Federal holidays and working days are indicated by binary calendar characteristics.

Techniques Used

- **Linear Regression -** Used to provide coefficients for understanding the effects of the continuous target variable (cnt) and to model and predict it.
- **Logistic Regression -** Designed to use continuous and categorical information to distinguish between days with high and low demand.
- **Decision Tree Classification -** provides a rule-based paradigm for investigating predictor influence hierarchies and splits.
- **Cross-Validation (caret) -** 5-fold resampling is used to verify the robustness of the model.
- **Data Visualizations -** built to visualize relationships and ideas using ggplot2 and standard R charting tools.

R version 4.4.2 was used for all preprocessing, transformation, modelling, and graphing.

Results

Linear Regression - (cnt ~ multiple features):

The results of the multiple linear regression model showed that –

- Higher temperatures are associated with more bike usage, as evidenced by the strong positive association between temperature and rental demand ($p < 2e-16$).
- Windspeed and humidity were both statistically significant predictors with a negative correlation, indicating that users are impacted by physical resistance and discomfort.
- Contextual factors influence usage patterns, as evidenced by the statistically significant coefficients for season and weather.

Model fit metrics –

- Multiple R - squared - 0.522
- Adjusted R - squared - 0.517

The selected predictors can account for approximately 52% of the variation in daily bike rental counts, according to these values.

Logistic Regression - (high vs low demand):

Using predictors like temperature, humidity, wind speed, season, and weather, the logistic model categorized demand.

- On the test set, the model's accuracy was about 72.7%.
- Both the high and low classes performed almost equally, with sensitivity and specificity in balance.
- Preparation for days with high demand is made possible by this binary classification, which aids in demand forecasting decision-making.

Decision Tree :

The rpart approach was used to train a decision tree model –

- Temperature was the most significant variable (42), followed by season (28), hum (16), and windspeed (9), according to the variable importance score.
- Easy-to-understand decision guidelines were offered by the tree, such as the following: demand is high if temperature > 0.43 and humidity < 0.74.
- The model's explainability makes it perfect for stakeholders that want outputs that can be understood, even though it is marginally less accurate.

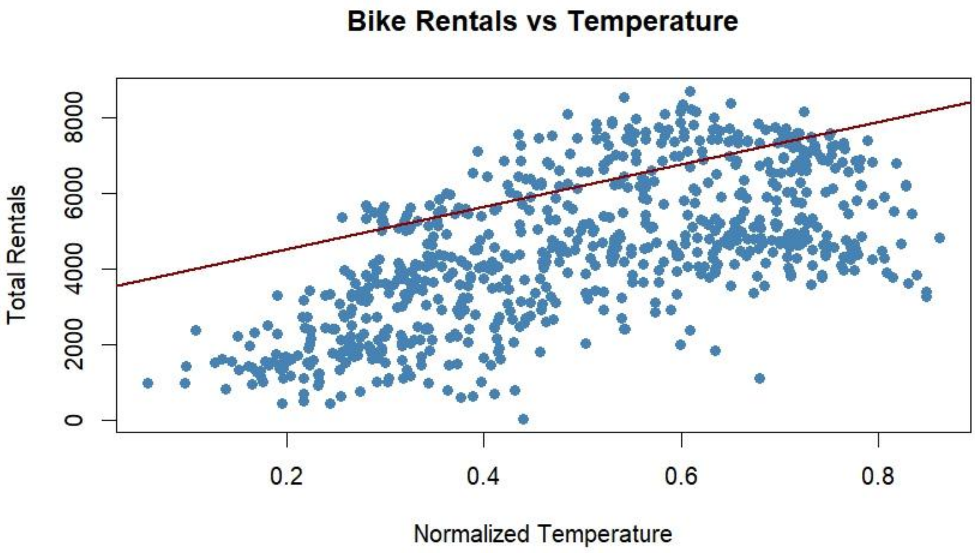
caret::train Logistic Model:

A logistic model was also assessed with 5-fold cross-validation and caret::train:

- Accuracy of cross-validation: ~77.9%
- Kappa value: around 0.56
- Verifies the categorization approach's dependability across various data subsets.

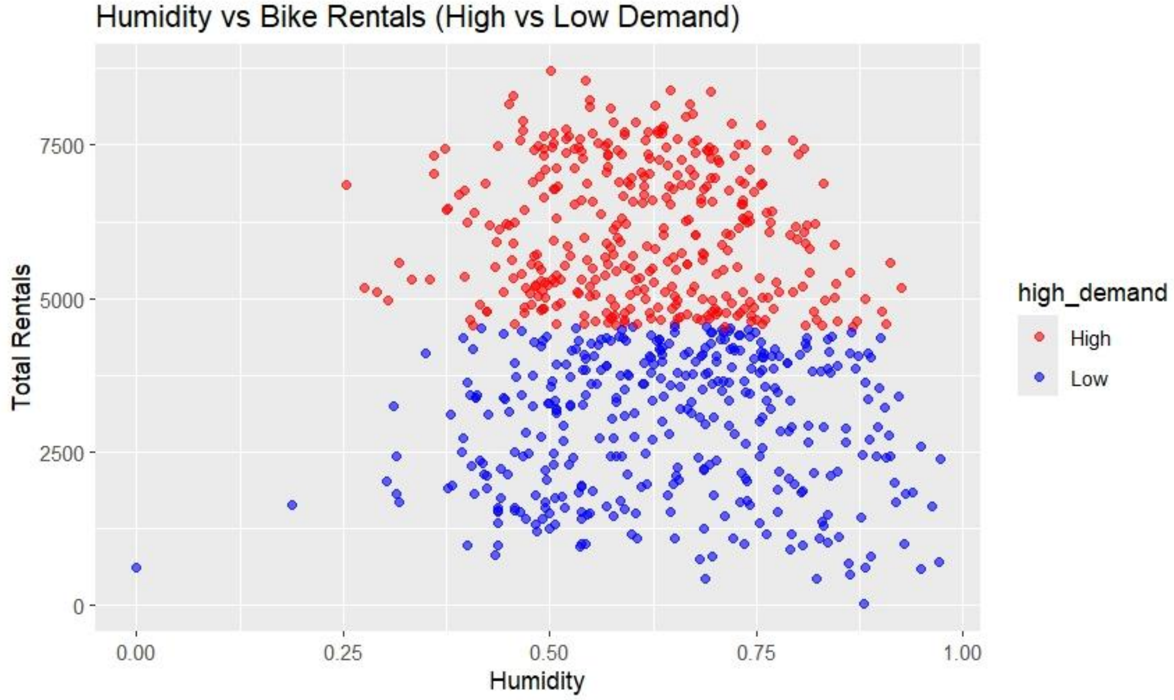
Visualizations

Figure 1 - Bike Rentals vs Temperature:



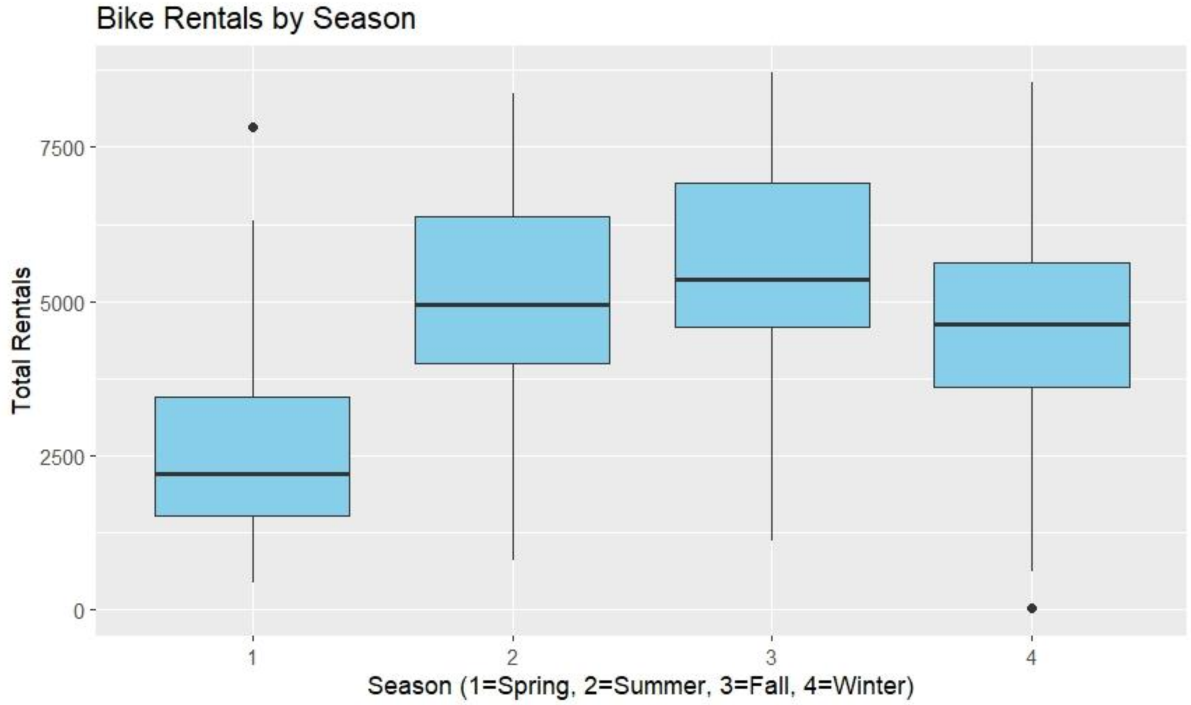
The scatterplot illustrates the positive linear correlation between total rentals and normalized temperature. There are usually more rides on warmer days.

Figure 2 - Humidity vs Bike Rentals (High/Low Demand):



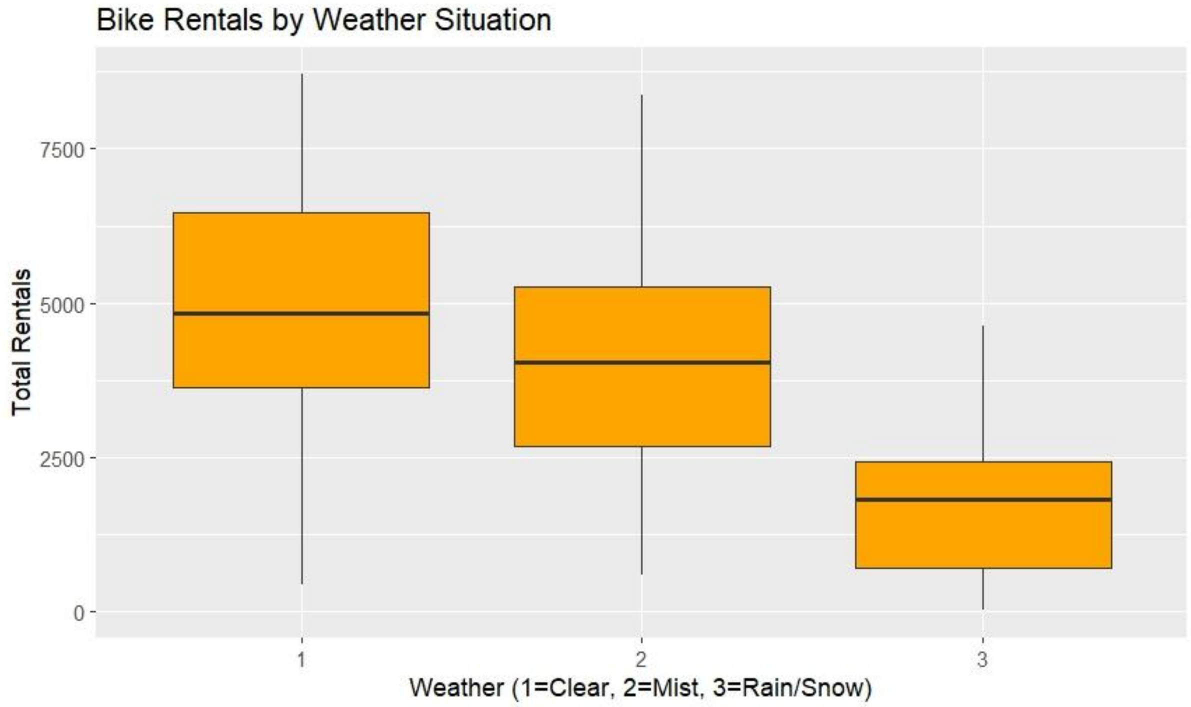
This coloured scatterplot compares humidity to demand by classifying days as either high or low. Days with high demand seem to be more closely associated with lower humidity.

Figure 3 - Rentals by Season:



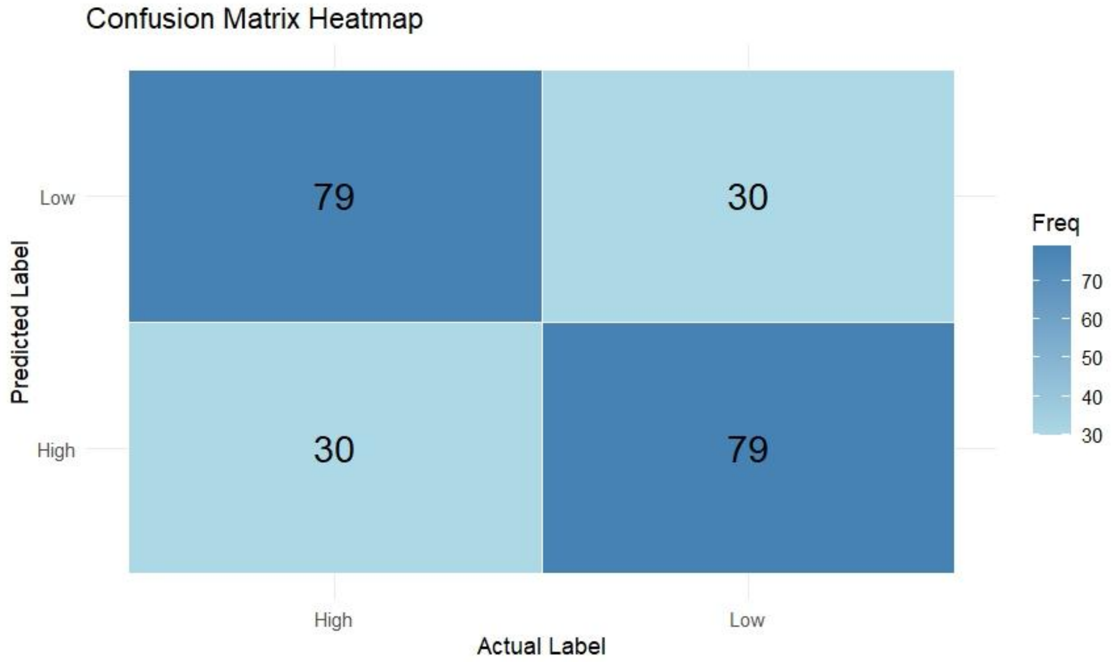
The distribution of bike rentals by season is shown in a boxplot. The median count is highest in the fall and lowest in the spring.

Figure 4 - Rentals by Weather Situation:



This boxplot compares total rentals across different weather types. Clear weather leads to the highest rental counts, while rain/snow significantly reduces demand.

Figure 5 - Confusion Matrix Heatmap:



A heatmap that shows the performance of a logistic regression model. Each of the remaining 79 accurate high/low classifications has strong predictive accuracy.

Discussion & Conclusion

The results unequivocally show that the demand for daily bike rentals is significantly influenced by temperature. Days with high temperatures tend to have more rentals, while wind and humidity discourage use because they are uncomfortable. User behaviour insights are further refined by season and weather, which demonstrate that activity is encouraged by clear, temperate conditions.

Both the logistic and decision tree models produced excellent classification results, with accuracies of more than 72%. For bike-sharing operations, these models can be utilized to create real-time notifications or staffing decisions. Key insights are easier for non-technical stakeholders to understand thanks to the visuals' efficient validation of the statistical results.

All things considered, this project demonstrates a useful application of statistical modelling in urban mobility systems. The method can be added to real-time demand forecasting dashboards and expanded to additional cities. Time-series forecasting and the incorporation of external data sources, such as public events or traffic disruptions, may be part of future research.

References:

- UCI Machine Learning Repository - Bike Sharing Dataset
- R Documentation - caret, ggplot2, rpart, reshape2
- INFO6105 Course Materials