Bike Sharing Demand Prediction - Capstone Project Report

Author: Ashik Abraham Baby

1. Introduction

Bike-sharing systems have emerged as a convenient and eco-friendly alternative for short-distance travel in urban cities.
Understanding and predicting rental demand helps operators optimize bike distribution, reduce imbalance, and enhance user satisfaction.
This project focuses on developing a machine learning model to accurately predict hourly bike rental demand using historical data combined with weather and temporal attributes.

2. Problem Statement

The goal of this project is to predict the total number of bikes rented in a given hour based on various factors such as time, weather, season, and holiday information. Accurate predictions can assist bike-sharing companies with planning, inventory management, and operational efficiency.

3. Dataset Description

The dataset consists of 17,379 rows and multiple features describing hourly bike rental activity. Major columns include:

- dteday – Date
- season – Spring, Summer, Fall, Winter
- year, month, day, hour
- holiday – Whether the day is a holiday
- workingday – Whether the day is a working day
- weather – Clear, Misty, Light Snow/Rain, Heavy Rain
- temp, atemp – Actual and perceived temperatures

• humidity, windspeed
• cnt – Target variable (total rentals)

The dataset is complete with no missing values, making preparation straightforward.

4. Data Preprocessing

Preprocessing steps included:

● Converting date strings into datetimes
● Extracting new time components (day, month, hour)
● Mapping categorical variables to meaningful labels
● Dropping irrelevant variables such as casual and registered
● Ensuring numeric types are consistent for modeling

Dataset was clean and required no missing-value treatment.

5. Exploratory Data Analysis (EDA)

EDA allowed us to understand demand patterns:

● Hourly analysis revealed morning and evening peaks (commute hours)
● Monthly and seasonal analysis showed highest demand in Summer and Fall
● Weather conditions greatly influenced demand (rain & snow reduced rentals)
● Higher temperatures correlated with higher usage
● Humidity had an inverse relationship with rentals
● Weekend vs weekday patterns differed significantly

6. Feature Engineering

To improve model performance, several derived features were introduced:

● Cyclical Encoding
  - hour_sin, hour_cos
  - month_sin, month_cos

● Peak Hour Indicators
  - is_morning_peak
  - is_evening_peak
  - is_peak

● Interaction Features
  - temp_humidity
  - temp_windspeed

● atemp_diff – difference between actual and apparent temperature

These features helped models better understand seasonality and human behavior patterns.

7. Model Development

Three models were trained:

1. Linear Regression
2. Random Forest Regressor
3. XGBoost Regressor

Train-test split was done with 80% training data and 20% testing data.

8. Model Evaluation

The models were evaluated using RMSE, MAE, and R^2 metrics.

Performance summary:

• Linear Regression: Moderate performance ($R^2 \sim 0.63$)
• Random Forest: Strong performance ($R^2 \sim 0.94$)
• XGBoost: Best performance ($R^2 \sim 0.957$)

XGBoost was selected as the final model due to superior predictive accuracy.

9. Feature Importance Analysis

XGBoost feature importance showed:

1. hour
2. temp
3. humidity
4. is_peak
5. weather
6. workingday
7. cyclic encodings (hour_sin, month_sin)

Time of day and environmental conditions are the biggest drivers of demand.

10. Residual Analysis

Residuals were examined to validate model reliability.

Findings:

- Residuals centered around zero → no systematic bias
- Higher errors during peak hours due to sudden surges
- Lowest errors in Winter due to stable usage patterns
- Summer had highest variability and thus higher error

This confirms the model generalizes well but faces challenges during high-variability periods.

11. Saving the Final Model

The final XGBoost model was exported as best_model.joblib for deployment, dashboard integration, or future inference.

12. Conclusion

The project successfully developed a machine learning-based prediction system for hourly bike rentals. XGBoost demonstrated superior performance,
capturing complex nonlinear relationships in the data. The insights obtained

are valuable for planning supply, improving availability, and enhancing user experience.

13. Future Work

Potential improvements include:

- Deploying a dashboard for visualizing our project.
- Hyperparameter optimization for further accuracy gains