Bike Sharing Demand Prediction

Paradesi Reshwanth

March 15, 2025

Contents

Chapter 1

Introduction

The objective of this project is to predict hourly bike-sharing demand using a combination of weather, calendar, and temporal features. This dataset (from UCI Bike Sharing) provides an opportunity to apply feature engineering, model selection, and advanced machine learning techniques. Our goal is to evaluate linear models (Ridge, Lasso) against an ensemble-based model (Histogram Gradient Boosting) and interpret the results.

Chapter 2

Methodology

Dataset: The dataset contains 17 original features and over 17,000 hourly records. **Steps followed:**

1. Data Cleaning and Feature Engineering (lag features, rolling averages, cyclical encodings).
2. Exploratory Data Analysis (EDA) to understand temporal and categorical patterns.
3. Model Training: Ridge, Lasso, Histogram Gradient Boosting with hyperparameter tuning.
4. Model Evaluation: RMSE, R^2, residual analysis, permutation importance.

Chapter 3

Exploratory Data Analysis (EDA)

3.1 Average Demand by Hour

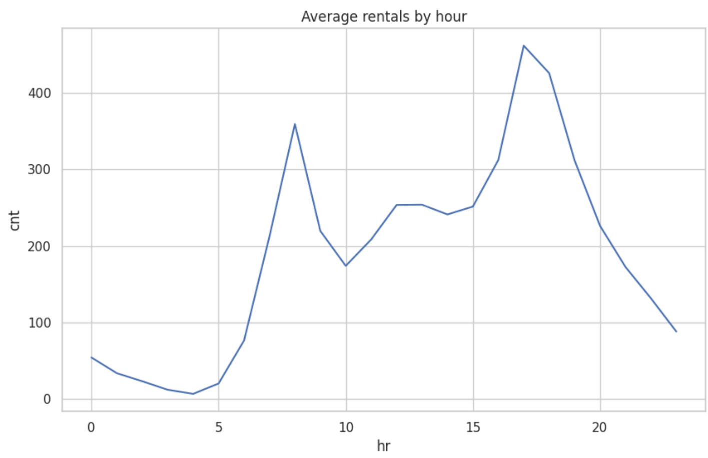

Figure 3.1: Average Bike Demand by Hour of Day

Observation: Demand is low during early morning hours, peaks sharply around 8–9 AM (commute), dips at noon, and rises again around 5–7 PM (evening commute). This confirms strong daily seasonality.

3.2 Average Demand by Season

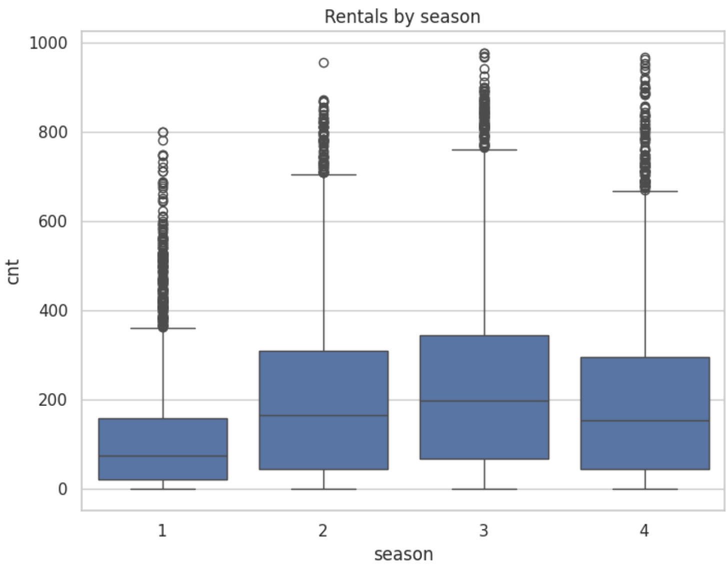

Figure 3.2: Bike Demand by Season

Observation: Demand is highest in summer and fall, and lowest in winter, showing clear dependency on weather/seasonal conditions.

3.3 Demand by Working Day

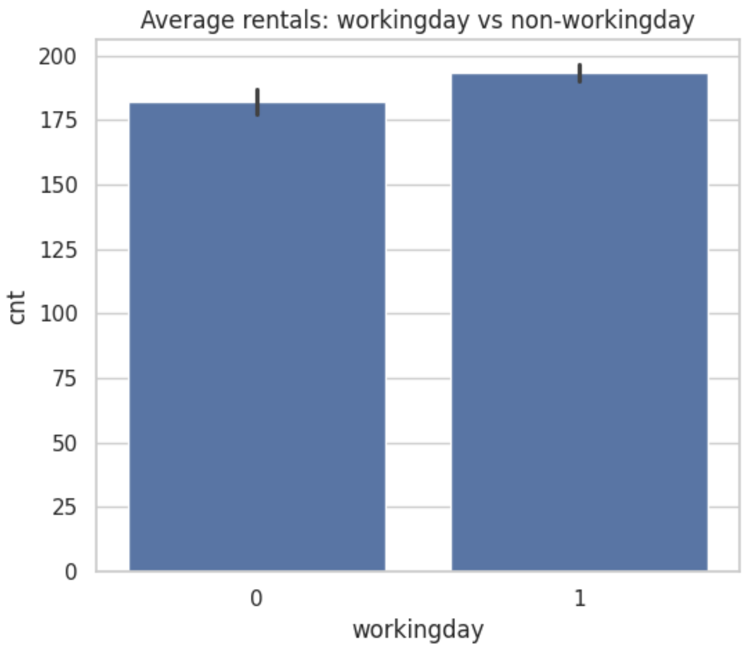

Figure 3.3: Working Day vs Holiday Demand

Observation: Working days exhibit strong commute-related peaks, while weekends/holidays show smoother, more evenly distributed usage.

3.4 Distribution of Rental Counts

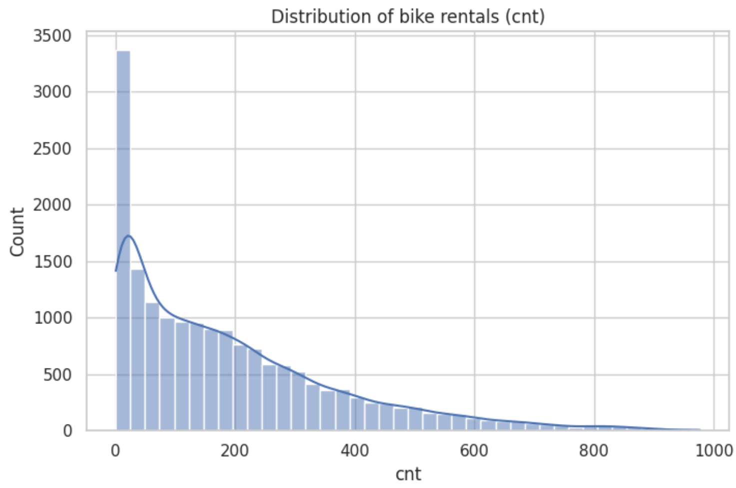

Figure 3.4: Distribution of Rental Counts

Observation: The distribution is right-skewed, with most hours having moderate demand and fewer hours experiencing very high usage.

3.5 Feature Correlation

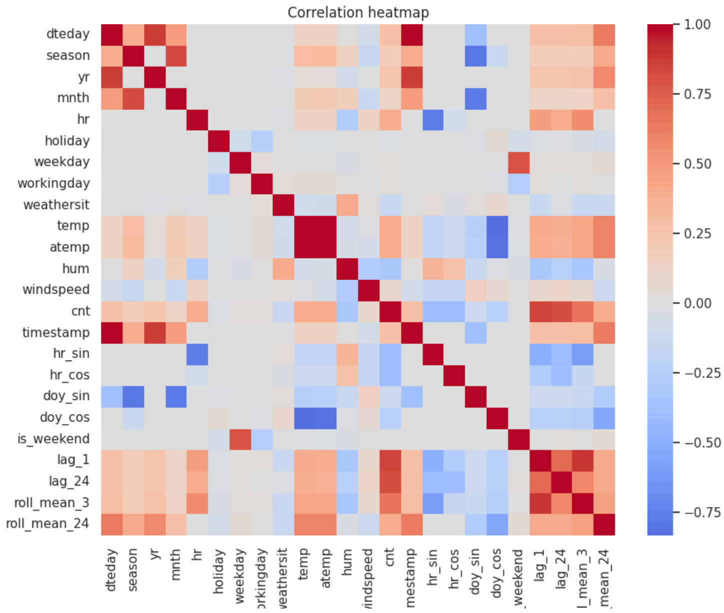

Figure 3.5: Feature Correlation Heatmap

Observation: Temperature and "feels-like temperature" are highly correlated, as expected. Humidity and weather situation show negative correlation with demand.

Chapter 4

Model Evaluation and Results

4.1 Ridge Regression

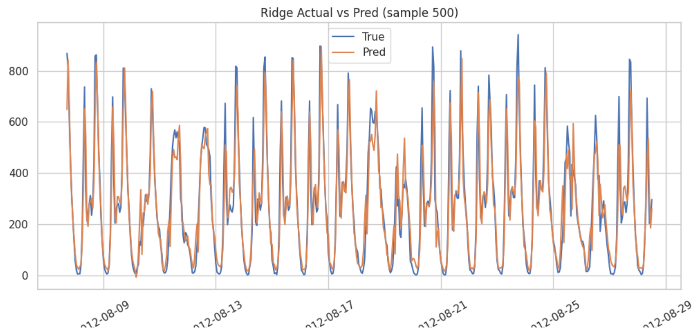

Figure 4.1: Ridge: Actual vs Predicted (Time Series)

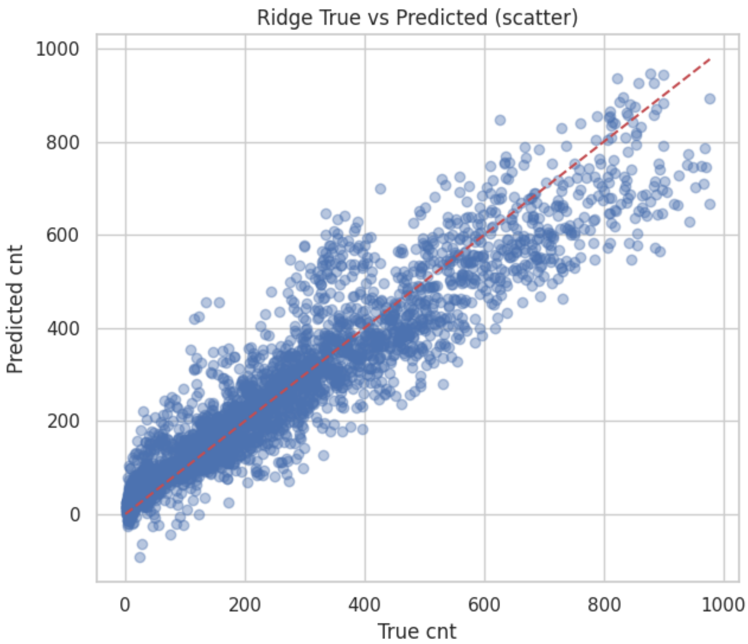

Figure 4.2: Ridge: True vs Predicted Scatter

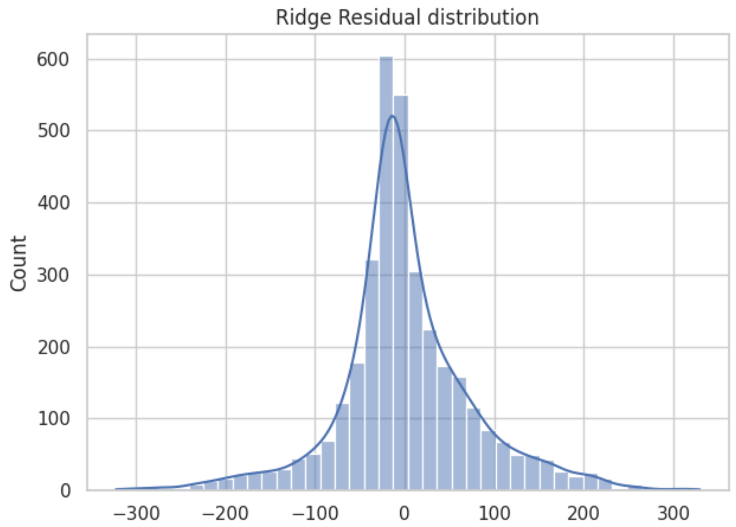

Figure 4.3: Ridge: Residual Distribution

Observation: Ridge regression captures overall seasonality but struggles with sharp peaks. Residuals are spread with moderate variance, confirming underfitting for complex patterns.

4.2 Lasso Regression

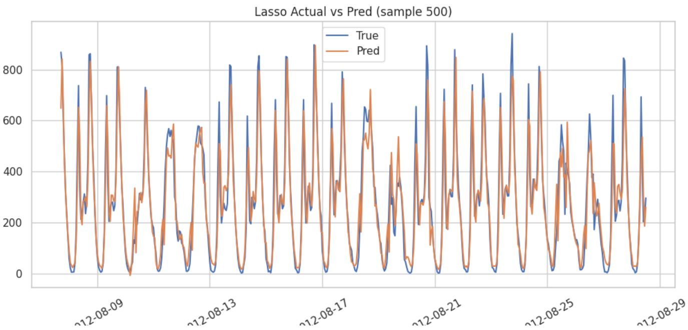

Figure 4.4: Lasso: Actual vs Predicted (Time Series)

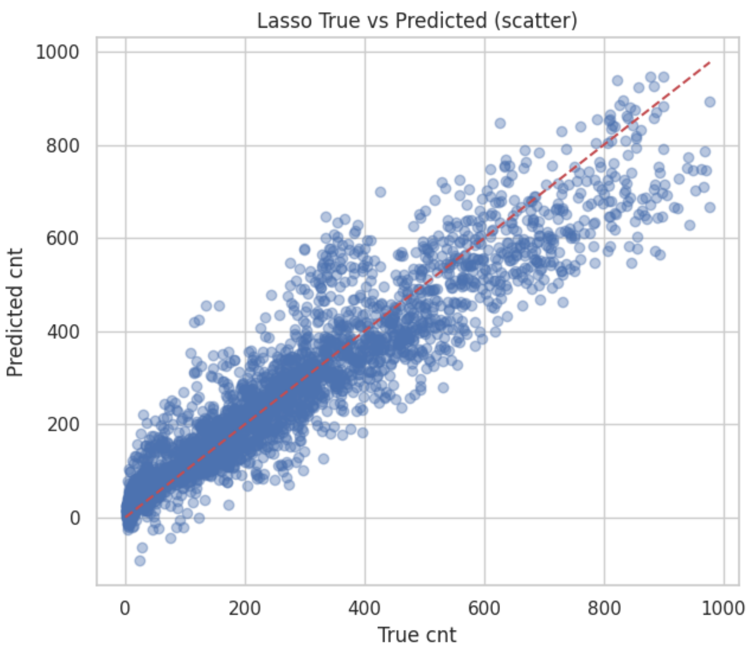

Figure 4.5: Lasso: True vs Predicted Scatter

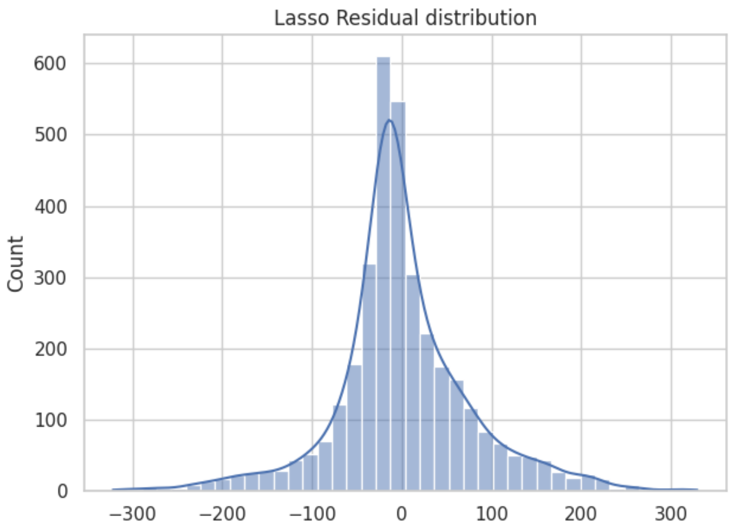

Figure 4.6: Lasso: Residual Distribution

Observation: Lasso shows very similar performance to Ridge, with nearly identical R^2 values. Regularization leads to slightly sparser models, but predictive strength remains comparable.

4.3 Histogram Gradient Boosting (Best Model)

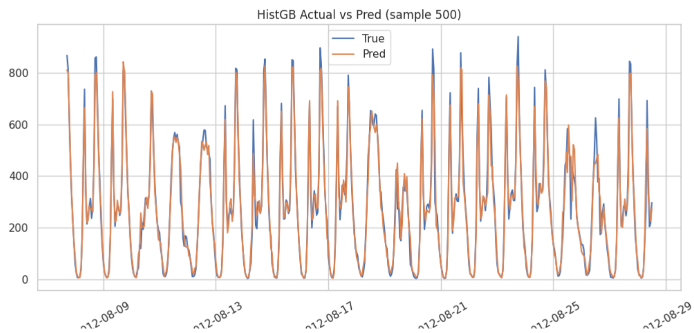

Figure 4.7: HistGB: Actual vs Predicted (Time Series)

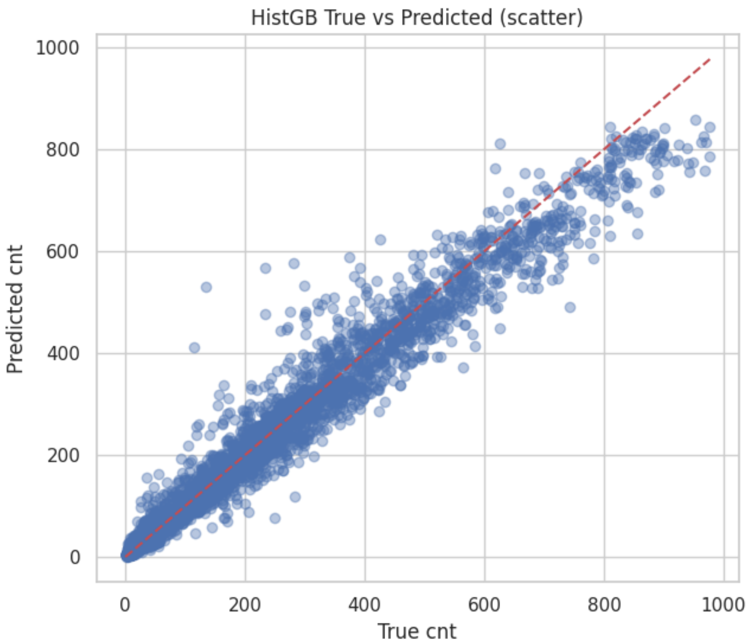

Figure 4.8: HistGB: True vs Predicted Scatter

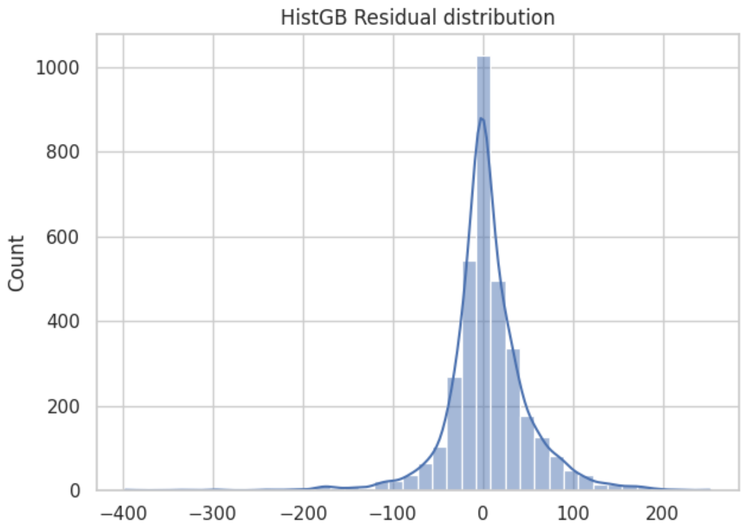

Figure 4.9: HistGB: Residual Distribution

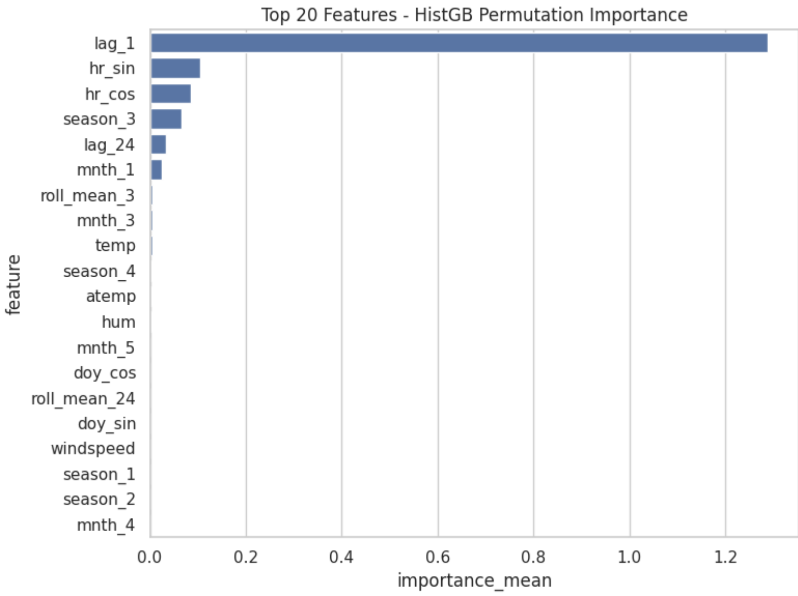

Figure 4.10: HistGB: Feature Importance (Permutation)

Observation: HistGB achieved the best performance with Test RMSE $\approx$ 45.6 and $R^2 = 0.957$. - Time-series alignment shows it tracks peaks and troughs very well. - Scatter plot is tightly aligned around the diagonal (ideal predictions). - Residuals are small and centered, with fewer large errors compared to Ridge/Lasso. - Feature importance highlights *hour of day*, *temperature*, and lag features as the most influential.

Chapter 5

Conclusion

The analysis shows that while linear models provide a baseline ($R^2 \approx 0.88$), Histogram Gradient Boosting significantly outperforms them ($R^2 = 0.957$). Temporal features (hour, lag, rolling averages) and weather-related variables were key in driving accurate predictions. Future extensions could include deep learning architectures (RNN, LSTM) for sequential modeling.