Report(Q1): Regression Analysis on Bike Sharing Demand

Dhyan Santoshbhai Patel (BT2024075)
Raj Jayeshkumar Gandhi (BT2024172)

December 8, 2025

1 Objective

The primary objective of this project is to predict the total count of bikes rented during a specific hour using the **Kaggle Bike Sharing Demand** dataset. The goal is to implement and compare various regression models—specifically Linear Regression, Polynomial Regression (degrees $d = 2, 3, 4$), and a Quadratic Model with interaction terms. A key focus of this iteration is advanced feature engineering, specifically Cyclical and One-Hot encoding, to better represent temporal and categorical data.

2 Methodology & Implementation

The solution was implemented in Python using `numpy` for matrix operations and `pandas` for data handling. The implementation adheres to the strict constraint of using the Normal Equation for optimization.

2.1 Advanced Feature Engineering

To improve model performance, we moved beyond raw numerical inputs and implemented specific encodings for time and categories:

2.1.1 Cyclical Encoding for Time

Time features like 'hour' (0-23) and 'month' (1-12) are cyclical. In a standard linear representation, Hour 23 and Hour 0 are numerically distant ($distance = 23$), but temporally adjacent. To capture this continuity, we mapped these features onto a circle using sine and cosine transformations:

$$x_{sin} = \sin\left(\frac{2\pi \cdot x}{P}\right)$$
$$x_{cos} = \cos\left(\frac{2\pi \cdot x}{P}\right)$$

where P is the period (24 for hours, 12 for months). This allows the model to interpret midnight (00:00) as continuous with the previous night (23:00).

2.1.2 One-Hot Encoding

Categorical variables such as ‘season‘ and ‘weather‘ do not have a strict ordinal relationship. We applied One-Hot Encoding to convert these into binary vectors. To prevent the "Dummy Variable Trap" (multicollinearity), one column was dropped from each category (e.g., $k-1$ columns for k categories).

2.2 Data Preprocessing

- **Leakage Prevention:** Columns `casual` and `registered` were dropped as their sum equals the target.
- **Standardization:** Z-score normalization was applied. Crucially, the mean (μ) and standard deviation (σ) were calculated **only on the training set** and then applied to the test set to ensure the test data remained unseen.
- **Train/Test Split:** A manual 80/20 split was performed using a fixed random seed (`123`) for reproducibility.

2.3 Model Formulation

The weights θ were calculated using the analytical **Normal Equation** with the pseudo-inverse (`pinv`) for numerical stability:

$$\theta = (X^T X)^{-1} X^T y$$

3 Results

The models were trained on the engineered features and evaluated on the held-out test set.

Model Name	**MSE**	R^2 **Score**
Interaction ($d = 2$)	**8869.83**	**0.7229**
Poly ($d = 4$)	11229.39	0.6491
Poly ($d = 3$)	11782.33	0.6318
Poly ($d = 2$)	12938.30	0.5957
Linear Regression	15485.58	0.5161

Table 1: Performance comparison of regression models on the test set.

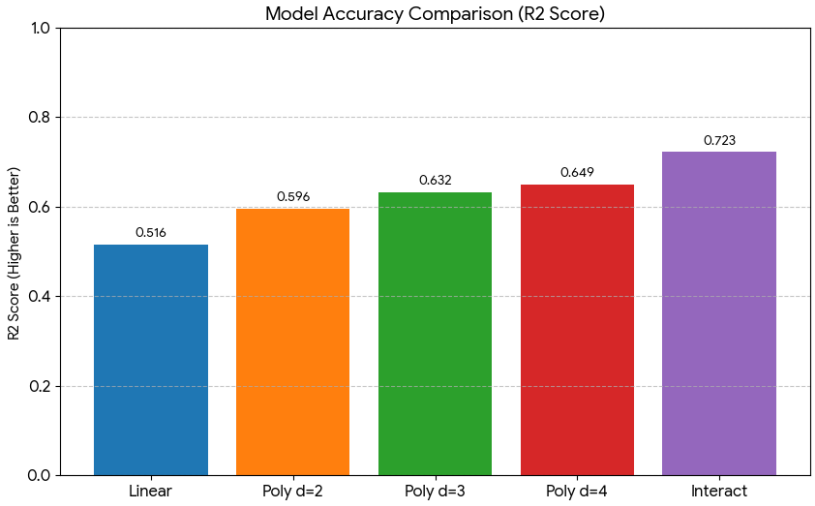

Figure 1: Optimization Trajectories

4 Analysis & Discussion

4.1 Feature Engineering Impact

The Linear Regression baseline improved significantly compared to previous iterations (from $R^2 \approx 0.38$ to 0.51). This validates the use of Cyclical Encoding: by representing time as coordinates on a circle, the linear model could effectively "connect" the late night and early morning hours, reducing the error at the day's boundaries.

4.2 Interaction vs. Polynomials

The **Quadratic Interaction Model** ($d = 2$) was the clear winner ($R^2 \approx 0.72$).

- **Synergy vs. Complexity:** While Polynomial models ($d = 3, 4$) try to fit the data by creating complex, "wiggly" curves, they treat features in isolation. The Interaction model, however, creates terms like $Temp \times Humidity$.
- **Physical Interpretation:** This captures conditional relationships. For example, high humidity might be bearable at low temperatures but drastically lowers bike demand at high temperatures. The Interaction model captures this "synergy," whereas pure polynomial models miss it.
- **Bias-Variance Trade-off:** The Interaction model adds necessary complexity (reducing Bias) without the excessive freedom of high-degree polynomials (avoiding High Variance/Overfitting).

5 Conclusion

The analysis confirms that the bike sharing demand dataset contains significant non-linear patterns and feature dependencies. The Linear Regression baseline is insufficient

due to high bias. The **Quadratic Model with Interactions** provides the most robust generalization by effectively modeling the curvature of the data and the synergistic effects between features, without suffering from the overfitting observed in higher-degree polynomials.