Optimization Project Report

Regression on Bike Sharing Demand

1. Problem Statement & Objectives

The objective of this experiment was to predict the hourly bike rental demand (count) using environmental and time-based features from the Bike Sharing Demand dataset. The dataset consists of 10,886 data points, which we expanded into 12 distinct features after preprocessing.

We evaluated five specific model configurations using an 80-20 chronological train-test split (8,708 training samples and 2,178 test samples). The models included a Linear Regression baseline, Polynomial Regression models (degrees 2, 3, and 4) without interaction terms, and a Quadratic model that included interaction terms. Model selection and performance analysis were based on Root Mean Squared Error (RMSE) and R^2 scores on the unseen test set.

2. Methodology

2.1 Preprocessing & Leakage Prevention

- Feature Engineering: We extracted temporal features including hour, weekday, month, and year from the datetime timestamp to capture cyclic patterns in demand.
- Feature Pruning: To prevent target leakage, we explicitly dropped the casual and registered columns, as their sum equals the target variable count.
- Split & Scale: We utilized a time-based split (first 80% for training) rather than a random shuffle to respect the temporal nature of the data. A StandardScaler was fit exclusively on the training data and then applied to the test data to ensure no information leakage occurred regarding feature distributions.

2.2 Model Implementations

- Linear Regression (Baseline): A standard linear fit using the 12 scaled features.
- Polynomial Models (Degrees 2, 3, 4 - No Interactions): We implemented a custom feature generator to create powers of features (x^2, x^3, …) independently for each variable. This allowed us to capture non-linearities (like the curve of demand vs. temperature) without exploding the feature space with cross-products.
  - Features generated: Degree 2 (24 features), Degree 3 (36 features), Degree 4 (48 features).
- Quadratic with Interactions: We used PolynomialFeatures(degree=2) to generate a full set of interaction terms (e.g., temp x humidity), resulting in 90 total features.

×

3. Results

The following table summarizes the performance of all five models on the test set, sorted by predictive accuracy (R^2):

RANK	MODEL	DEGREE	INTERACTIONS	RMSE	R^2
1	Polynomial	3	No	158.42	0.4693
2	Polynomial	4	No	158.80	0.4674
3	Polynomial	2	No	167.75	0.4057
4	Linear Regression	1	No	184.04	0.2847
5	Quadratic	2	Yes	184.37	0.2821

Best Model: The Polynomial model with degree=3 (no interactions) was the top performer, achieving an R^2 of 0.4693. This represents a significant improvement over the linear baseline ($R^2 = 0.2847$) and the complex interaction model.

4. Analysis

4.1 Why Polynomial degree=3 Won

The cubic polynomial model struck the best balance for this dataset. By including terms up to x^3 for each feature independently, the model successfully captured the non-linear nature of key variables–most notably the hour of the day, which likely follows a complex "M" shape (morning and evening peaks) that a straight line cannot fit. With 36 features, it maintained enough complexity to model these curves without introducing the excessive noise found in the interaction model.

4.2 Bias-Variance Tradeoff

- Linear (High Bias): The low R^2 score (0.2847) confirms that a simple linear equation significantly underfits the data. It fails to account for the obvious non-linear relationships between time/weather and bike demand.
- Polynomial degree=4 (Variance creeping in): While close in performance to the cubic model, the degree 4 model saw a slight decrease in R^2 (0.4674) and increase in RMSE. This suggests the beginning of overfitting, where the model starts chasing noise in the training data rather than generalizing patterns.
- Quadratic with Interactions (Overfitting/Multicollinearity): Despite having the most features (90), this model performed the worst ($R^2 = 0.2821$). This counter-intuitive result is likely due to severe multicollinearity introduced by the interaction terms (e.g., temp and atemp are already highly correlated; multiplying them creates redundant noise), causing the model weights to become unstable.

4.3 The Failure of Interactions

Counter-intuitively, the Quadratic model with interactions performed the worst ($R^2=0.2821$), slightly below even the simple linear baseline. Despite having the most features (90), it failed to generalize. This is likely due to overfitting and multicollinearity. With so many interaction terms derived from a limited set of base features, the model likely "memorized" noise in the training set that did not exist in the future test data. This validates our decision to prune interaction terms in the other polynomial models.

5. Conclusion

The Polynomial degree=3 model (No Interactions) proved to be the optimal choice. It effectively captures the non-linear single-variable effects—such as the curve of demand throughout the day or across temperatures—while avoiding the pitfalls of overfitting seen in higher-complexity models.

The experiment highlighted that for this specific dataset, adding feature complexity via interactions was detrimental. A simpler approach focusing on the power-relationships of individual features yielded the most robust predictions, offering an RMSE of 158.52 bikes per hour.