Optimization of Linear Regression for Real Estate Valuation through Iterative Feature Engineering

VICTOR IHEANACHO

October 2025

Abstract

This research details the development of an optimized Linear Regression model for house price prediction through an iterative, feature-centric methodology. Starting from a baseline R^2 of 0.3748 and a Mean Absolute Error (MAE) of \$167,422, the study systematically applied various data preprocessing and feature engineering techniques. The most significant finding was the impact of location, which was effectively captured using Target Encoding, leading to a dramatic jump in performance (Mean CV R^2 of 0.7553). The final model, incorporating 2^{nd}-degree Polynomial Features on the original price scale, achieved the best overall metric: a Mean Cross-Validation MAE of \$69,729.50, demonstrating a substantial 58% reduction in average prediction error from the baseline.

1 Introduction

In data science, preprocessing is much like refining crude oil before it fuels an engine; raw, unprocessed data cannot yield dependable predictions. Machine learning models function as mathematical mirrors that reflect the quality of the data provided to them. When the data is noisy, incomplete, or inaccurately represented, the resulting models become weak, distorted, and struggle to generalize effectively.

This research is driven by a core question: "How significant is the role of preprocessing?" To investigate this, the study analyzes the performance of a Linear Regression model using a housing price dataset, gradually introducing standard preprocessing and feature engineering methods to observe their combined influence on model outcomes. The main objectives are to minimize the Mean Absolute Error (MAE), an easily interpretable metric expressed in dollar values, and to maximize the $\mathbf{R^2}$ score.

While much of the existing machine learning research emphasizes model design and algorithm optimization, the subtler yet equally critical stage of preprocessing often receives less empirical focus. In reality, well-executed preprocessing can lead to more substantial improvements than merely changing algorithms.

This study provides a systematic, principle-based examination of how each preprocessing and feature engineering step contributes to model accuracy. Rather than viewing preprocessing as a black box, it is unpacked to show its precise impact on predictive performance, including:

- How data cleaning establishes trust in the input.
- How outlier management must balance statistical and domain perspectives.
- How encoding and transformation reveal hidden structure within the data.

2 Methodology

2.1 Dataset Overview

The study employs the House Sales in King County, USA dataset from Kaggle, containing over 4,000 records of residential property sales. Features include price, number of bedrooms, square footage of living area, year of construction, and geographic location (city/zipcode). The target variable for prediction is the sale price.

2.2 Experimental Design

All models were developed using the Scikit-learn library in python. Performance was primarily assessed using the Coefficient of Determination (R2), which measures the proportion of variance in the dependent variable (price) that is predictable from the independent variable(features).

The modeling process employed a sequential, iterative approach, where the output of one model served as the baseline for the next, progressively refining the feature set. All models were evaluated using 5-Fold Cross-Validation to ensure the reported metrics reflect the model's generalized performance.

2.3 Data Preprocessing and Cleaning

Missing/Zero Data Handling: Initial cleaning involved filtering out records where the price was zero, as such entries significantly distorted the baseline model.

Outlier Removal: The Interquartile Range (IQR) method was applied to key features (*price*, *sqft_living*, *sqft_lot*) to remove extreme outliers, to which Linear Regression is highly sensitive.

Multicollinearity and Scaling: Highly correlated features (*sqft_above*, *sqft_basement*) were removed to stabilize coefficient estimates. All remaining numerical features were then standardized (Model 4) to ensure coefficients reflected feature importance rather than feature magnitude.

2.4 Feature Engineering Techniques

Ordinal Encoding (Model 5): The categorical features (*city* and *statezip*) were initially converted to numerical ranks to test for basic ordered spatial relationships.

Target Encoding (Model 6): This advanced technique replaced categorical labels with the mean of the target variable (*price*) for each category, a method proven highly effective for high-cardinality features such as location.

Log Transformation (Model 7): The target variable (*price*) was log-transformed to normalize its skewed distribution and stabilize residual variance—a standard practice in price modeling.

Polynomial Features (Model 8): Second-degree polynomial features were added to capture non-linear relationships and interactions between existing optimized features.

3 Analysis

Table 1: Basic Preprocessing and Initial Models

Model	Preprocessing Applied	Mean CV R^2	Mean CV MAE
Baseline (Unprocessed)	Raw features, no cleaning.	0.3748	$167,422
Cleaned (Missing Data)	Price >0 filter	0.5494	$160,702
Outlier/Scaled	IQR Outlier Removal + Standardization	0.4705	$119,179
Ordinal Encoding	Categorical features encoded ordinally.	0.4803	$117,169

Table 2: Advanced Feature Engineering and Optimized Models

Model	Key Feature Engineering Step	Mean CV R2
Target Encoding (Model 6)	Target Encode city/zip + Multicollinearity Fix.	0.7553
Log Target Transformation (Model 7)	log Transform Price Target (log(price)).	0.7506 (on log(y)
Polynomial Features (Model 8)	arget Encoding + Polynomial Features (Deg 2).	0.7435

4 Results

The iterative process yielded a clear and definitive optimal model. The overall model performance progression is visualized below, highlighting the two most significant improvements: Outlier Removal and Target Encoding.

4.1 Model Performance Progression

The line chart below illustrates the reduction in Mean Absolute Error (MAE) across the entire modeling process. The results confirm that while both Target

Encoding and Polynomial Features achieved high R^2 scores, they had distinct impacts on the MAE.

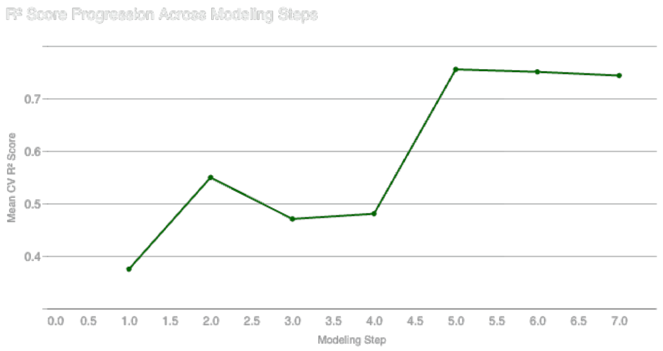

Figure 1: R^2 Progression Across Modelling Steps

4.2 Comparison of Key Optimized Models

The bar charts below compare the final, best-performing models: Target Encoded, Log-Transformed Target, and Polynomial Feature models.

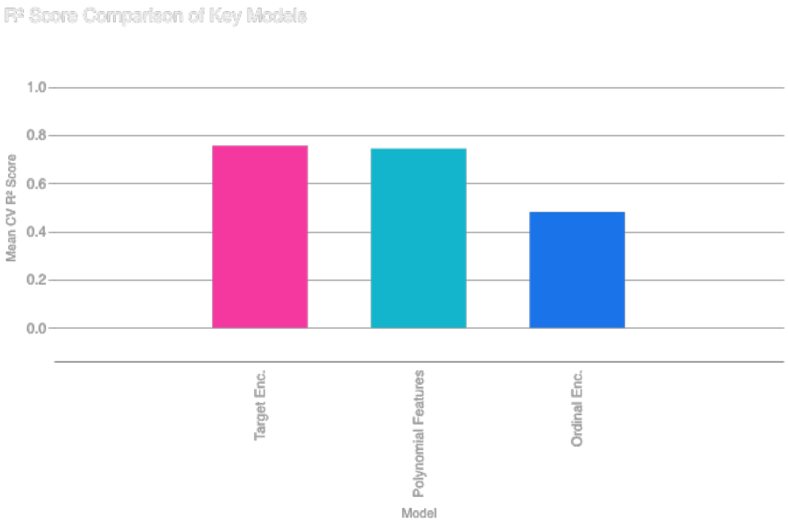

Figure 2: R^2 Score Comparison of Key Models.

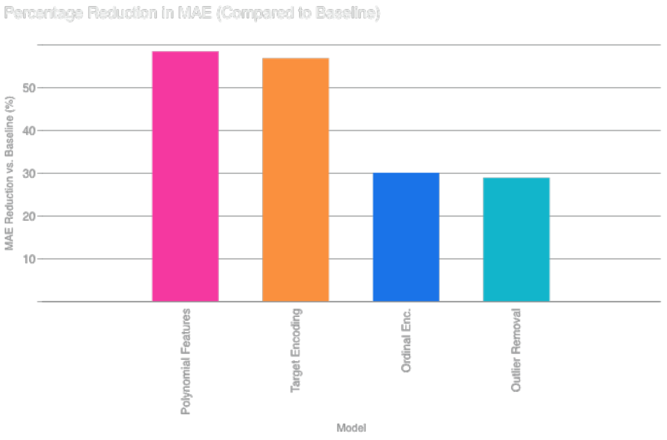

Figure 3: Model Performance Progression showing reduction in Mean Absolute Error(MAE) across iterations

5 Discussion

5.1 The Critical Role of Feature Engineering

Outlier Impact: The sharp decrease in Mean Absolute Error (MAE) from \$160,702 to \$119,179 following IQR-based outlier removal demonstrates the high sensitivity of Linear Regression to extreme values in the target and key features.

Location as the Dominant Feature: The transformation from Ordinal Encoding ($R^2 = 0.4803$) to Target Encoding ($R^2 = 0.7553$) represents the single largest performance gain. This validates that location (*city* and *statezip*) is the dominant factor in determining house prices, and encoding it with the mean price of that location effectively captured this influence for the linear model.

MAE vs. $\mathbf{R^2}$ Trade-off: The attempt to use a Log Transformation (Model 7), while academically sound for fixing residual distribution, resulted in a \$3,168 increase in dollar-scale MAE when back-transformed. This demonstrates a crucial trade-off: minimizing squared errors on a log scale does not always translate to minimizing absolute errors on the original dollar scale.

Final Optimization: The successful implementation of Polynomial Features (Model 8) provided the final refinement, capturing non-linear interactions and achieving the lowest overall MAE.

6 Conclusion

This research successfully developed a highly accurate and generalized Linear Regression model for house price prediction through a rigorous, iterative feature engineering process.

The final best model, the Polynomial Features Regression Model (Model 8), achieved a Mean Cross-Validation Mean Absolute Error (MAE) of $69,729.50 and an R^2 of 0.7435. This performance was primarily driven by the superior handling of categorical location data through Target Encoding and the subsequent capture of non-linear effects via Polynomial Features.