Analysis of House Price Prediction using Linear Regression

Uttam Mahata (2022CSB104)

August 6, 2025

1 Introduction and Objective

The primary objective of this analysis was to develop a linear regression model capable of accurately predicting the final sale price of a house. The dataset used contains **81** features describing various aspects of residential homes.

2 Data Preprocessing

Before model development, the dataset underwent essential preprocessing steps to ensure data quality and suitability for linear regression.

- **Data Loading:** The `train.csv` dataset was loaded into a Pandas DataFrame.
- **Missing Value Imputation:** The `LotFrontage` column, identified as a numerical feature with missing values (read as 'NA' strings), was handled by replacing the missing entries with the **mean** of the column. This technique ensures that the statistical properties of the column are not significantly distorted.

3 Model Development and Evaluation

A systematic approach was taken to build and evaluate seven different regression models. The evaluation was performed using **5-fold cross-validation**, a robust technique that ensures the model's performance is generalizable and not dependent on a single random split of the data. The performance of each model was quantified using two key metrics:

- **Mean Squared Error (MSE):** Measures the average of the squares of the prediction errors. A lower MSE indicates a better fit.
- **R^2 Score (Coefficient of Determination):** Represents the proportion of the variance in the dependent variable (`SalePrice`) that is predictable from the independent variables (features). An R^2 score closer to 1 indicates a better fit.

The following models were constructed:

- **Simple Linear Regression:** SalePrice $\sim$ LotArea

- **Model 1 (Numerical):** SalePrice ∼ LotFrontage + LotArea
- **Model 2 (Numerical):** SalePrice ∼ LotFrontage + LotArea + OverallQual + OverallCond
- **Model 3 (Numerical):** SalePrice ∼ LotFrontage + LotArea + OverallQual + OverallCond + 1stFlrSF + GrLivArea
- **Model 4 (Mixed):** SalePrice ∼ LotArea + Street
- **Model 5 (Mixed):** SalePrice ∼ LotArea + OverallCond + Street + Neighborhood
- **Model 6 (Mixed):** SalePrice ∼ LotArea + OverallCond + Street + 1stFlrSF + Neighborhood + Year

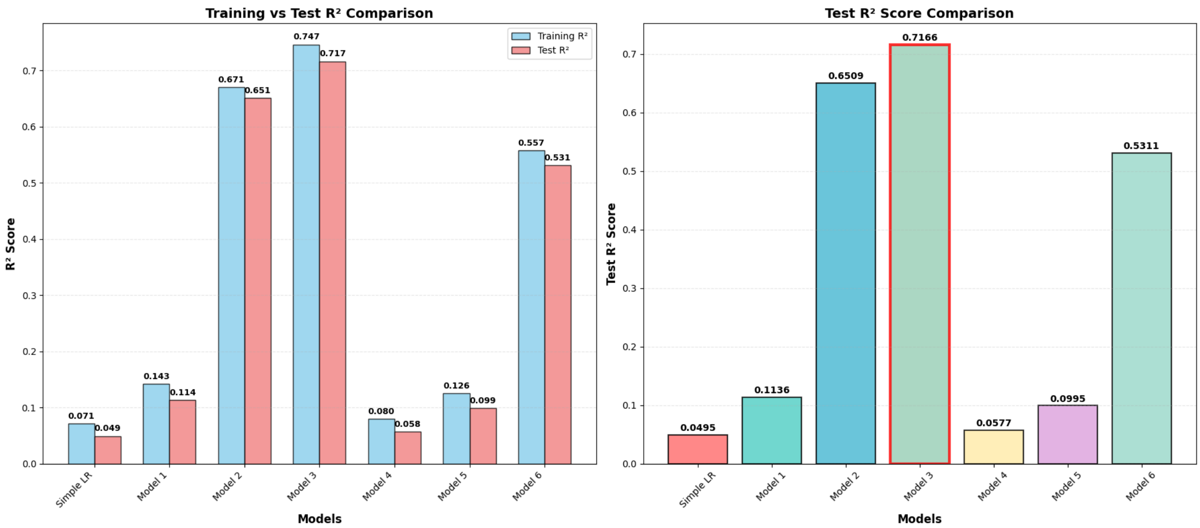
For the mixed models, categorical features were converted to a numerical format using **Label Encoding**.

4 Results and Discussion

The evaluation of the models yielded clear insights into the factors influencing house prices.

4.1 Model Performance Comparison

The performance of the models improved consistently as more relevant features were added. Model 3 emerged as the most accurate model based on the average test R^2 score.



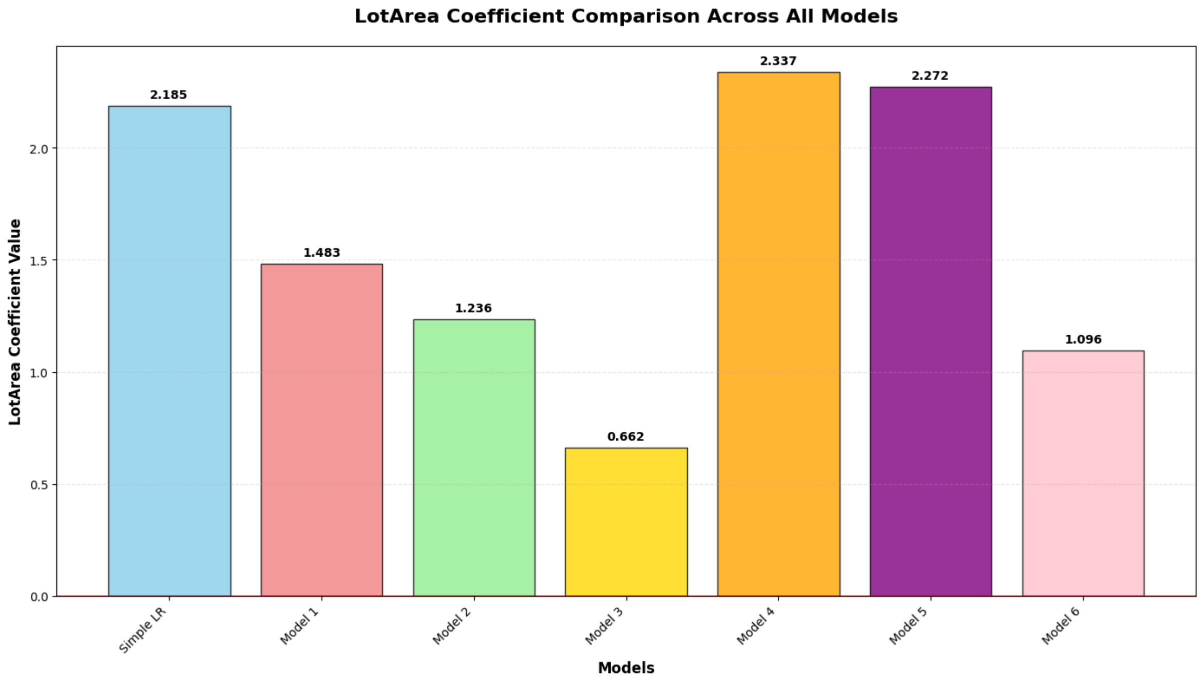
Model	Average Test MSE	Average Test R^2
Simple LR	5,995,826,265.93	0.0495
Model 1	5,555,904,889.54	0.1136
Model 2	2,169,949,347.89	0.6509
Model 3	**1,734,881,618.52**	**0.7166**
Model 4	5,937,730,454.03	0.0577
Model 5	5,668,347,710.49	0.0995
Model 6	2,912,636,507.51	0.5311

Table 1: Comparison of model performance across different linear regression models

4.2 Analysis of `LotArea` Coefficient

A key part of the analysis was to observe the change in the coefficient for the `LotArea` feature across all models. The coefficient represents the change in `SalePrice` for a one-unit increase in `LotArea`, holding other variables constant.

The analysis revealed that in the simple model, `LotArea` has a coefficient of **2.18**. However, as more powerful predictors like `OverallQual` (Overall Quality) and `GrLivArea` (Above Grade Living Area) were added in Models 2 and 3, the coefficient for `LotArea` dropped significantly to **0.66**.



LOTAREA COEFFICIENT COMPARISON

This demonstrates a critical concept in multiple regression: the importance of a single feature is relative to the other features in the model. The initial high coefficient of `LotArea` was likely capturing some of the effects of other, unincluded variables (e.g., larger lots often have larger, higher-quality houses). Once these more direct predictors were added, the unique contribution of `LotArea` itself was found to be smaller.

5 Conclusion

The analysis demonstrates that a multiple linear regression model can be an effective tool for predicting house sale prices. The key conclusions are:

1. **Feature Addition Improves Accuracy:** Model performance is drastically improved by adding relevant features. A simple, single-feature model is inadequate for this task.

2. **Model 3 is the Most Effective:** The model incorporating `LotFrontage`, `LotArea`, `OverallQual`, `OverallCond`, `1stFlrSF`, and `GrLivArea` provided the best balance of simplicity and predictive power, explaining nearly 72% of the price variance.

3. **Feature Importance is Contextual:** The predictive weight of any single feature, like `LotArea`, is highly dependent on the other features included in the model. Features like overall quality and living area are more dominant predictors of price than lot size alone.