Predicting Housing Prices: The Dominance of Numerical Features, Effects of Data Transformation, and Challenges with High-Value Properties.

Ofek Kirshenboim

September 2024

Abstract

This study investigates the determinants of housing prices, focusing on the predictive power of numerical and categorical features, the effect of data transformations, and the accuracy of models for high-priced properties. First, we analyzed the significance of numerical features in predicting *SalePrice* using a random forest model. The feature importance analysis revealed that numerical features overwhelmingly dominate in predicting housing prices, with *OverallQual* and *GrLivArea* being the most influential. This demonstrates the crucial role of numerical characteristics such as quality, space, and structural attributes in determining property values.

Next, we applied a Box-Cox transformation to the *SalePrice* to address the skewness in its distribution. The transformed *SalePrice* exhibited a more symmetric distribution, which was expected to improve model performance. Visual comparisons of predicted versus actual values indicated that the Box-Cox transformation improved the linearity of predictions, reducing variance and thereby enhancing the model's predictive power.

Finally, we examined the model's performance across different price ranges and observed that higher-priced homes exhibited larger residuals, indicating lower predictive accuracy in this segment. This suggests that the model struggles to generalize well for expensive homes, potentially due to increased variance and more complex factors affecting high-value properties. The GitHub repository for this study is available at `https://github.com/Ofekirsh/statistic_theory.git`.

1 Introduction

Accurate prediction of house prices is crucial for stakeholders across the real estate ecosystem, including buyers, sellers, investors, and policymakers. House prices are influenced by a variety of features that can broadly be categorized into numerical and categorical variables, such as property size, age, quality, and

location [1]. Traditional methods for predicting house prices have relied heavily on linear regression models, which leverage basic attributes such as square footage, number of rooms, and neighborhood to provide a foundational level of accuracy [1]. These models, however, struggle to fully capture the complex interplay between different housing features and tend to underperform when considering high-value properties. Recent advancements in machine learning have introduced more sophisticated methods for price prediction, including decision trees, random forests, and gradient boosting, which are capable of modeling non-linear relationships between features [2, 4]. These models significantly outperform linear regression when it comes to incorporating a greater variety of housing attributes and capturing more nuanced interactions. Feature engineering and data preprocessing play a critical role in improving model performance, particularly when dealing with skewed data distributions. Techniques like the Box-Cox transformation have been employed to normalize skewed target variables such as *SalePrice*, leading to improved model robustness [3]. While machine learning models continue to evolve, one of the persistent challenges is understanding the contribution of different feature types, specifically numerical versus categorical features. In our study, we examine the relative importance of these features in predicting *SalePrice* and demonstrate that numerical features overwhelmingly contribute to the model's predictive performance. Attributes such as overall quality (*OverallQual*), above-ground living area (*GrLivArea*), and basement area (*TotalBsmtSF*) were identified as the most influential, suggesting that property value is highly driven by quantifiable factors. Another important aspect of our study involves transforming the target variable to enhance model performance. We applied a Box-Cox transformation to *SalePrice* to address the skewness present in the original distribution. This transformation resulted in a more normal distribution, which helped improve the predictive capability of the model [3]. The effectiveness of this approach was confirmed by analyzing scatter plots of predicted versus actual values, which showed a more consistent and linear relationship post-transformation. Lastly, we investigate the model's performance across different property price ranges and observe that predictive accuracy declines for higher-priced homes. This observation is highlighted by larger residuals in the high-price segment, indicating greater variability and potential biases in the model's predictions. Such challenges call for a more tailored approach to model high-value properties, possibly through the use of specialized data or more granular feature engineering. In this study, we present a comprehensive approach to house price prediction, combining advanced feature importance analysis, statistical transformations, and detailed performance evaluations. Our findings underscore the significance of numerical features, the benefits of transforming skewed data, and the challenges of predicting prices for high-value homes.

2 Methods

2.1 Data

The dataset used in this study is the Ames housing dataset, a comprehensive collection of housing records from Ames, Iowa. This dataset includes 2,930 properties, each described by 79 features, both numerical and categorical, which provide information about the physical characteristics of the properties as well as their sales transactions. **Data Sources:** The Ames housing dataset is sourced from *Kaggle's House Prices: Advanced Regression Techniques* competition. The dataset was originally compiled by Dean De Cock to provide a rich alternative to the Boston Housing dataset for academic use in regression tasks. **Preprocessing and Cleaning:** Prior to analysis, several preprocessing steps were taken to prepare the data for modeling:

- **Handling Missing Values:** A significant portion of the dataset contained missing values, which were handled based on the nature of each feature. Features with more than 70% missing values, such as *Alley*, *PoolQC*, and *Fence*, were dropped from the dataset. For other features, missing values were imputed using median values for numerical features and the most frequent category for categorical features.
- **Encoding Categorical Features:** Categorical features were converted into numerical representations using one-hot encoding to enable their use in machine learning models.
- **Outlier Detection and Removal:** Numerical features were inspected for outliers using boxplots, and records that contained extreme values, particularly in the *GrLivArea* and *SalePrice* features, were removed to reduce the influence of skewed data on the model.

2.2 Machine Learning Models: Linear Regression and Random Forest

To predict the target variable, *SalePrice*, we employed two machine learning models: Linear Regression and Random Forest. These models were chosen to provide a comparison between a simple linear approach and a more complex, non-linear ensemble technique. **Linear Regression:** Linear regression aims to model the relationship between the independent variables (features) and the target variable (*SalePrice*) by fitting a linear equation of the form:

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon$$

where $\hat{y}$ represents the predicted *SalePrice*, β_0 is the intercept, β_i are the coefficients corresponding to each feature x_i, p is the number of features, and ϵ is the error term. The coefficients β_i are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared residuals:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i represents the actual *SalePrice*, $\hat{y}_i$ is the predicted value, and n is the number of observations. **Random Forest:** Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Each decision tree in the forest is trained on a random subset of the training data using bagging (bootstrap aggregation). The Random Forest algorithm's prediction is obtained by averaging the predictions of all individual trees:

$$\hat{y}_{\text{RF}} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

where T is the total number of decision trees in the forest, and $f_t(x)$ is the prediction from the t-th tree. Random Forest introduces randomness by selecting a random subset of features for each split in a decision tree, thereby decorating the trees and reducing variance. For feature selection, Random Forest uses the Gini impurity or mean squared error (MSE) to determine the optimal split at each node. The importance of each feature is computed based on its contribution to reducing impurity or MSE across all trees, providing insights into which features are most influential in predicting *SalePrice*. Both models were trained using the preprocessed dataset, with hyperparameters tuned using cross-validation to optimize performance. Linear regression served as a baseline to evaluate the performance gains from using a non-linear model such as Random Forest.

2.3 Statistical Tests

To validate the assumptions underlying the data and to assess the relationships between different variables, several statistical tests were performed: the Kolmogorov-Smirnov test, Chi-square test, and Pearson and Spearman correlation tests. **Kolmogorov-Smirnov Test:** The Kolmogorov-Smirnov (KS) test is a non-parametric test used to determine whether two samples are drawn from the same distribution. Specifically, it compares the empirical cumulative distribution functions (ECDFs) of the train and test datasets for a given feature. The KS statistic is given by:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

where $F_n(x)$ and $G_m(x)$ are the ECDFs of the two samples of sizes n and m, respectively. The null hypothesis of the KS test is that both samples are drawn from the same underlying distribution. A high KS statistic, along with a low p-value, indicates a significant difference between the two distributions, suggesting potential discrepancies between the train and test datasets. **Chi-square Test of Independence:** The Chi-square (χ^2) test of independence was

used to determine whether there is a significant association between categorical variables. To perform this test, a contingency table is created to summarize the frequency distribution of the variables. The test statistic is computed as:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} represents the observed frequency in cell (i, j), and E_{ij} represents the expected frequency under the assumption of independence, calculated as:

$$E_{ij} = \frac{(\text{row total})_i \times (\text{column total})_j}{\text{grand total}}$$

The resulting χ^2 statistic follows a Chi-square distribution with $(r-1)(c-1)$ degrees of freedom, where r and c are the numbers of rows and columns in the contingency table, respectively. A p-value less than the significance level indicates that the null hypothesis of independence can be rejected, suggesting that there is a significant relationship between the categorical variables. **Pearson and Spearman Correlation Tests:** Correlation tests were conducted to assess the relationships between numerical features. The Pearson correlation measures the linear relationship between two variables and is calculated as:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\text{Cov}(X, Y)$ represents the covariance between variables X and Y, and σ_X and σ_Y are the standard deviations of X and Y. Pearson's correlation coefficient ranges from -1 to 1, with values close to -1 or 1 indicating strong negative or positive linear correlation, respectively. In contrast, the Spearman correlation measures the monotonic relationship between two variables and is defined as the Pearson correlation of the ranked values of the variables. The Spearman rank correlation coefficient, ρ_s, is given by:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i represents the difference between the ranks of corresponding values and n is the number of observations. The Spearman correlation is beneficial for assessing non-linear relationships and is robust to outliers.

2.4 Box-Cox Transformation

The Box-Cox transformation is used to stabilize variance and make the distribution of data more normally distributed. For a given variable y, the Box-Cox transformation is defined as:

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

where λ is a parameter that determines the nature of the transformation, and $y > 0$. The optimal value of λ is typically chosen to maximize the likelihood of the transformed data being normally distributed. In our study, we applied the Box-Cox transformation to the target variable, *SalePrice*, to reduce skewness and improve model performance. Importantly, the transformation is reversible, allowing us to convert the transformed values back to the original scale using:

$$y = \begin{cases} (\lambda y' + 1)^{1/\lambda} & \text{if } \lambda \neq 0 \\ \exp(y') & \text{if } \lambda = 0 \end{cases}$$

This reversibility ensures that predictions made in the transformed space can be interpreted in terms of the original *SalePrice*.

3 Results

3.1 Numerical features are more significant predictors of SalePrice compared to categorical features

To evaluate the significance of numerical features in predicting *SalePrice*, we analyzed the feature importance from the trained random forest model[]. Figure 1 clearly shows that, out of the 15 most influential features in the model, 15 are numerical features. This overwhelming representation of numerical features indicates that they play a far more crucial role in determining property prices compared to categorical features. The most influential feature is *OverallQual*, which represents the overall quality of the house. It has the highest importance score, demonstrating that buyers tend to put significant emphasis on the quality of the property when deciding on its value. The second most influential feature is *GrLivArea*, which represents the above-ground living area of the house. Larger living spaces are strongly correlated with higher property values, making *GrLivArea* a critical factor. Similarly, *TotalBsmtSF*, which represents the total basement area, also ranks highly, showing that additional space in the form of basements contributes significantly to the price. Features like *2ndFlrSF*, *LotArea*, and *GarageArea* also rank among the top, further emphasizing the importance of space-related features in determining housing prices. These numerical features provide detailed, continuous information that allows the model to capture subtle variations in housing prices. The strong dominance of numerical features in the feature importance analysis demonstrates that numerical aspects such as quality, space, and structural attributes are critical in predicting property values. This finding suggests that accurate numerical data is essential for creating reliable predictive models for *SalePrice*.

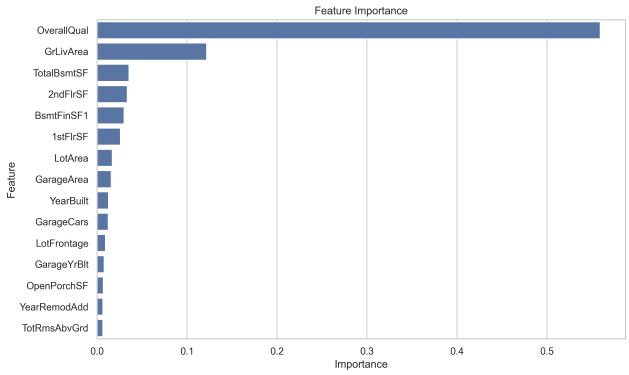

Figure 1: Feature importance plot showing the relative contribution of each feature in predicting *SalePrice*.

3.2 Applying a Box-Cox transformation on SalePrice will improve the performance of predictive models

To assess the impact of data transformation on the model's predictive performance, we applied a Box-Cox transformation to the *SalePrice* variable. This transformation aims to reduce skewness and improve the normality of the target variable. Figure 2 provides a visual comparison of *SalePrice* before and after the transformation. The original distribution is highly right-skewed, whereas the transformed distribution shows a more symmetric shape. This improvement in normality is expected to enhance the model's ability to capture the relationships between features and the target variable.

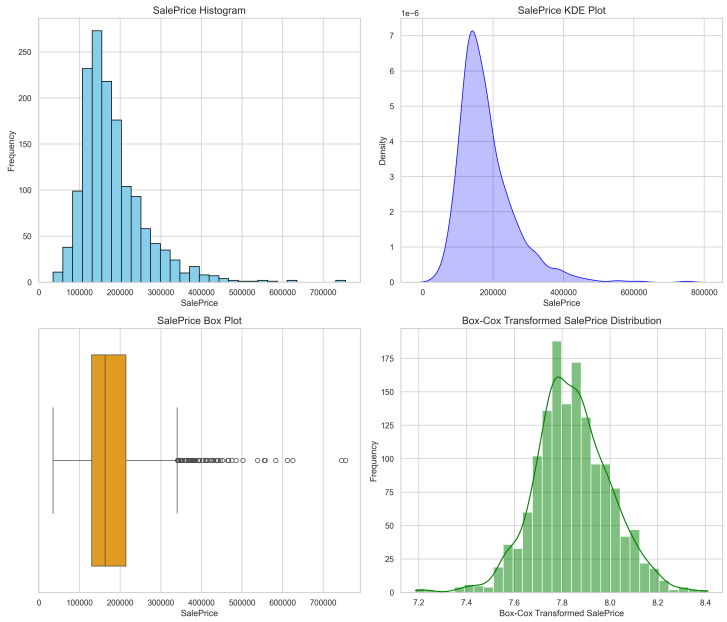

Figure 2: The four plots show different views of the *SalePrice* distribution. The top-left plot is a histogram of the original *SalePrice*, showing a right-skewed distribution with a peak around 130,000. The top-right plot is a kernel density estimate (KDE) of the original *SalePrice*, also highlighting the right-skewness with a sharp decline after the peak. The bottom-left plot is a boxplot of the original *SalePrice*, indicating the presence of multiple outliers on the higher end of the price range. The bottom-right plot is a histogram of the *SalePrice* after applying the Box-Cox transformation, showing a more symmetric distribution with reduced skewness.

The scatter plots in Figure 3 illustrate the predicted versus actual *SalePrice* values for both the original and Box-Cox transformed datasets. The predictions on the transformed *SalePrice* show a more linear relationship with reduced variance, indicating improved model performance after the transformation.

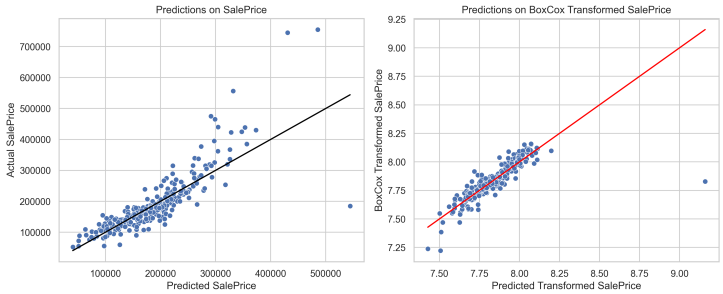

Figure 3: The scatter plots show the predicted versus actual values of *SalePrice*. The left plot presents the actual *SalePrice* on the y-axis and the predicted *SalePrice* on the x-axis, with a reference line representing perfect predictions. The right plot shows the predicted values of the Box-Cox transformed *SalePrice* on the x-axis and the transformed actual values on the y-axis, also with a reference line indicating perfect predictions.

3.3 Higher-priced homes tend to have lower predictive accuracy

To evaluate the model's performance across different price ranges, we analyzed the residuals of the predictions. Figure 4 presents the residual plot, highlighting that higher-priced homes tend to have larger residuals, suggesting lower predictive accuracy. This observation could be attributed to the higher variance in sale prices for expensive homes, which makes it more challenging for the model to generalize well.

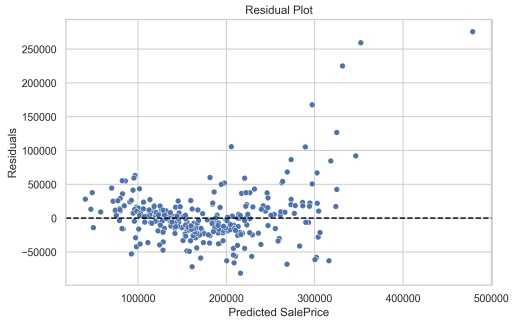

Figure 4: Residual plot showing the prediction errors for *SalePrice*. The x-axis represents the predicted *SalePrice*, while the y-axis shows the residuals, the difference between actual and predicted values.

4 Discussion

The findings from our analysis provide meaningful insights into the factors affecting housing prices, the effectiveness of data transformation, and the limitations of predictive models, particularly for high-priced homes. Our feature importance analysis shows a clear dominance of numerical features in predicting *SalePrice*. Of the top 15 most important features, 15 are numerical, with *OverallQual* and *GrLivArea* being the most influential. This highlights the critical role of continuous attributes, such as the quality of construction, living space, and overall size, in determining property value. Numerical data allows models to capture subtle variations that significantly affect the price, suggesting that accurate and detailed numerical data is vital for robust property valuation. To improve model performance, we applied a Box-Cox transformation to *SalePrice*, which successfully reduced skewness and made the distribution more symmetric. This transformation was effective in improving the model's predictive accuracy, as evidenced by a more linear relationship between the predicted and actual values for the transformed dataset. The enhanced linearity and reduced variance demonstrate that data transformation techniques can significantly improve the ability of predictive models to capture complex relationships. However, despite these improvements, the residual analysis revealed that the model struggled to accurately predict higher-priced homes, as evidenced by larger residuals in this price range. This suggests that high-value properties, which often exhibit greater variability and unique features, may require more complex modeling approaches or additional data to capture their nuances effectively. Future work should focus on incorporating more specific features that account for luxury elements and using advanced modeling techniques to better predict the value of high-priced properties.

References

[1] S. Rosen, *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition*, Journal of Political Economy, vol. 82, no. 1, pp. 34-55, 1974. Available: https://doi.org/10.1086/260169

[2] J. H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. Available: https://projecteuclid.org/euclid.aos/1013203451

[3] G. E. P. Box and D. R. Cox, *An Analysis of Transformations*, Journal of the Royal Statistical Society Series B (Methodological), vol. 26, no. 2, pp. 211-252, 1964. Available: https://www.jstor.org/stable/2984418

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2009. Available: https://web.stanford.edu/~hastie/ElemStatLearn/