

Linear and Nonlinear Models for Ames Housing Prices

Yifan Yang

*Faculty of Arts & Science, University of Toronto, Toronto, Canada
1403651958@qq.com*

Abstract. Predicting residential sale prices requires models that can handle both additive patterns and more complex relationships among property characteristics. This study compares four approaches using the Ames Housing data: raw linear regression, log-transformed linear regression with smearing retransformation, a cross-validated regression tree, and a Random Forest. The 2,930 observations were divided into 70% training and 30% test samples, and all models were evaluated on the same held-out data using RMSE, MAE, and out-of-sample R^2 . The log transformation improved the linear benchmark, reducing RMSE from \$29,623.62 to \$27,731.26 and MAE from \$19,349.74 to \$18,347.86, while increasing R^2 from 0.8700 to 0.8861. Cross-validation improved the single tree, but its final test performance remained below both linear models. Random Forest achieved the strongest overall accuracy, with an RMSE of \$26,561.86, an MAE of \$17,864.80, and an R^2 of 0.8955. Permutation importance identified above-ground living area, overall quality, and neighborhood as the strongest predictors. The results indicate that nonlinear flexibility was most useful when many trees were combined, while the transformed linear model remained a competitive and more interpretable alternative.

Keywords: Housing price prediction, linear regression, decision tree, random forest, Ames Housing

1. Introduction

Residential property values are determined by a mixture of physical, quality-related, and location-specific characteristics. Floor area, construction age, neighborhood, and overall condition may all contribute to sale price, but their effects are unlikely to be identical across properties. This makes housing data a useful setting for comparing statistical models that impose different assumptions on the relationship between predictors and price. Hedonic approaches continue to be widely used in housing research, while machine-learning methods have become increasingly common in recent predictive applications [1]. Existing studies also show that both structural characteristics and location information can contribute meaningfully to residential price prediction [2].

Linear regression provides a transparent benchmark because it represents the response through a fixed combination of predictor effects. That structure is useful for interpretation, but it may be restrictive when the relationship between housing characteristics and price changes across observations. For example, additional living space may have a different value for houses of different quality, and some features may become important only beyond particular levels. Flexible predictive

methods can represent these types of threshold effects and interactions without requiring every nonlinear pattern to be written explicitly into the model [3]. Housing studies have reported cases in which machine-learning approaches outperform conventional hedonic specifications [3, 4], and ensemble methods may achieve further gains by combining information across multiple fitted models [5].

Tree-based methods are useful for examining this trade-off between interpretability and flexibility. A regression tree allows different regions of the predictor space to follow different prediction rules, which can capture abrupt changes and interactions that are difficult to represent with one global linear equation. A single tree, however, may be sensitive to the particular training sample. Random forest addresses this weakness by averaging predictions across many trees fitted from resampled observations and randomly selected predictor subsets. Previous housing applications have reported strong predictive performance for random forest models [4, 5]. Similar tree-based approaches have also been applied to larger residential datasets; for example, Tarasov and Dessoulavy-Sliwinski studied more than 10,000 housing transactions using random forest together with other machine-learning methods [6].

The empirical analysis in this study uses the Ames Housing dataset, which contains 2,930 residential property sales and detailed information on housing characteristics [7]. Four predictive specifications are examined: raw linear regression, log-transformed linear regression, a regression tree, and a random forest. All models are estimated using the same training sample and evaluated on the same held-out test set. Their predictive performance is compared using root mean squared error (RMSE), mean absolute error (MAE), and out-of-sample R^2 . The analysis focuses on whether the tree-based models provide better out-of-sample prediction than the two linear alternatives and on which housing characteristics contribute most strongly to prediction in the Random Forest model.

2. Method

2.1. Dataset preparation

The empirical dataset was obtained from the Ames Housing data introduced by De Cock [7] and accessed through the AmesHousing package in R. It contains 2,930 residential sales from Ames, Iowa. Sale_Price, recorded in U.S. dollars, served as the response variable. The modeling set included 11 predictors representing several aspects of each property, including living space, overall quality, neighborhood, age, basement area, and garage capacity.

Exploratory analysis was first used to examine the distribution of the response and several important predictor–price relationships. Sale_Price showed substantial right skewness, whereas its natural logarithm was much closer to a symmetric distribution. This pattern motivated the use of both the original response and a log-transformed response in the linear modeling stage. The exploratory plots also indicated that sale price generally increased with above-ground living area and tended to be higher for properties with stronger overall quality ratings.

Before fitting the models, categorical levels represented by fewer than 10 observations were grouped into an "Other" category to reduce instability from very sparse factor levels. After this preprocessing step, the working data contained one response variable and 11 predictors. A fixed random seed of 2026 was then used to divide the observations into 2,051 training cases and 879 test cases, corresponding to approximately 70% and 30% of the data, respectively. This single partition was retained for all four models so that their out-of-sample results were evaluated using the same test observations.

2.2. Linear regression models

Linear regression was used as the reference specification for the model comparison [2]. In the first model, Sale_Price was modeled directly as a function of the 11 selected predictors. For house i , the raw-price specification is written as equation (1).

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i \quad (1)$$

In equation (1), β_0 denotes the intercept, β_j is the coefficient associated with predictor j , and ϵ_i represents the model error for observation i . This specification provides the baseline against which the other three models are compared.

A second linear specification used the natural logarithm of Sale_Price as the response. The purpose of this transformation was to account for the pronounced right skewness observed in the original price distribution and to examine whether the linear fit improved on the transformed scale. The model is given in equation (2).

$$\log(Y_i) = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i \quad (2)$$

Standardized residual-versus-fitted plots and normal Q-Q plots were examined for both linear specifications. These diagnostics were used to compare changes in residual spread and departures from normality after the response transformation.

Predictions from the log-linear model were initially obtained on the log-price scale, so a retransformation step was required before they could be compared with predictions from the other models. Direct exponentiation can lead to retransformation bias. Therefore, Duan's smearing correction was applied using residuals from the fitted training model [8]. The resulting dollar-scale prediction is shown in equation (3).

$$\hat{Y}_i = \exp(\widehat{\log Y_i}) \left[\frac{1}{n} \sum_{k=1}^n \exp(\hat{\epsilon}_k) \right] \quad (3)$$

The bracketed term in equation (3) is the smearing factor estimated from the training residuals. Applying this correction placed the log-linear predictions back on the original sale-price scale, allowing RMSE, MAE, and out-of-sample R^2 to be calculated consistently across all four models.

2.3. Decision tree

The first nonlinear specification was a regression tree fitted with the training sample [1]. A relatively large candidate tree was initially constructed, allowing the data to determine a flexible set of predictor-based splits. Each terminal node produced a sale-price prediction equal to the mean response of the training observations assigned to that node. Model complexity was then selected through 10-fold cross-validation. Specifically, the complexity parameter cp was compared across candidate tree sizes, and the value corresponding to the minimum cross-validation error was retained for the final fitted tree.

2.4. Random Forest

The second nonlinear specification was a Random Forest fitted to the training data [1, 5]. The final forest contained 500 regression trees, with $mtry = 3$ predictors considered at each split. Individual

trees were built from bootstrap resamples of the training observations, while predictor selection at each node was restricted to a randomly chosen subset. The forest prediction was obtained by averaging the outputs from all trees, as expressed in equation (4).

$$\hat{Y}(x) = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (4)$$

In equation (4), $T_b(x)$ denotes the prediction from tree b for an observation with predictor vector x . Out-of-bag prediction error was recorded as additional trees were added to the ensemble, providing a way to assess whether the forest had stabilized before reaching 500 trees. Predictor importance was evaluated by permutation: each predictor was shuffled in turn, and the resulting increase in mean squared prediction error was used to measure how much the fitted forest relied on that variable.

2.5. Evaluation metrics

Predictive accuracy was assessed on the held-out test sample using root mean squared error (RMSE), mean absolute error (MAE), and out-of-sample R^2 [9]. All calculations were performed after predictions had been expressed on the original sale-price scale, which allowed the four fitted models to be evaluated under the same criterion. RMSE places more emphasis on relatively large prediction errors because the residuals are squared, whereas MAE summarizes the typical absolute difference between predicted and observed prices. Out-of-sample R^2 measures how well the fitted predictions account for variation in sale prices in the test data. The three measures are defined in equations (5)–(7).

$$RMSE = \sqrt{\left(\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2\right)} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (7)$$

In equations (5)–(7), n denotes the size of the test sample, Y_i denotes the observed sale price for property i , $\hat{Y}_i$ denotes its predicted value, and $\bar{Y}$ denotes the mean observed price across the test observations. Models with smaller RMSE and MAE and larger out-of-sample R^2 were regarded as having better predictive performance.

3. Results and discussion

3.1. Linear regression results and diagnostics

Figure 1 shows a clear difference between the two response scales. Sale price on its original scale has a long upper tail, whereas the distribution becomes noticeably more balanced after taking logarithms. This transformation also produced better test-set predictions. The raw linear model obtained an RMSE of \$29,623.62, an MAE of \$19,349.74, and an out-of-sample R^2 of 0.8700. After modeling log sale price and applying the smearing correction, RMSE decreased to \$27,731.26 and

MAE to \$18,347.86, while R^2 increased to 0.8861. Thus, the transformed specification reduced both measures of prediction error and explained a larger share of the variation observed in the test sample.

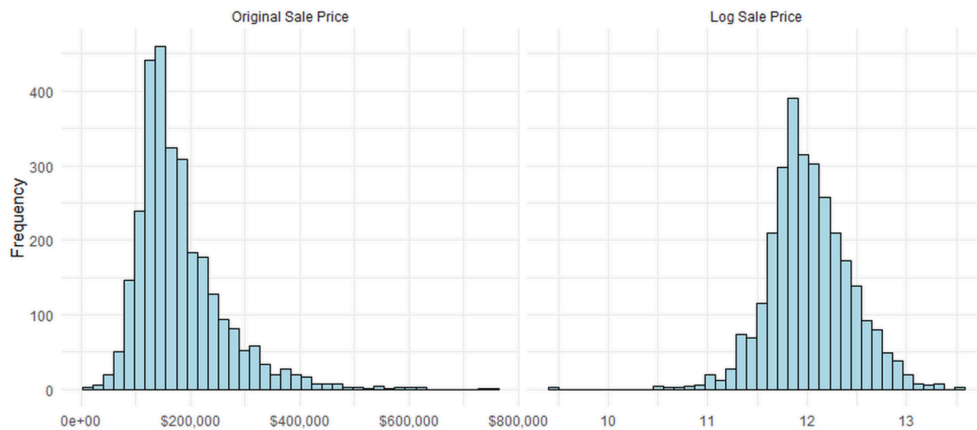


Figure 1. Sale-price distributions on the original and logarithmic scales (picture credit: original)

The residual diagnostics support the same comparison. In the raw-price model, residual dispersion became larger as fitted prices increased, particularly toward the upper end of the fitted range (Figure 2). This pattern is consistent with non-constant error variance. The log-linear model showed a more even residual spread across most fitted values, although several relatively large negative residuals were still present. Figure 3 leads to a similar conclusion from the normal Q-Q plots. Residuals from the log specification followed the reference line more closely through the middle of the distribution, while departures remained visible in the tails. The transformation therefore improved both predictive accuracy and the overall residual pattern, but it did not fully eliminate departures from the linear-model assumptions.

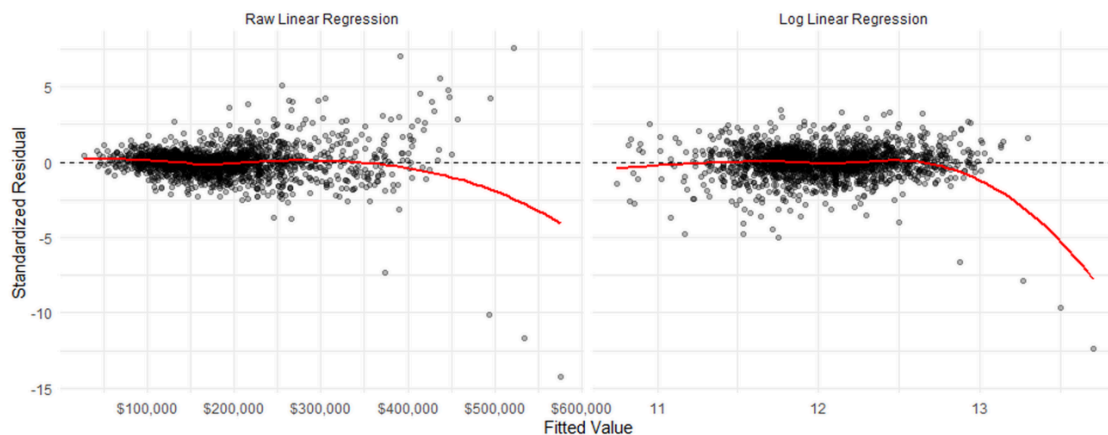


Figure 2. Residual-versus-fitted diagnostics for the two linear models (picture credit: original)

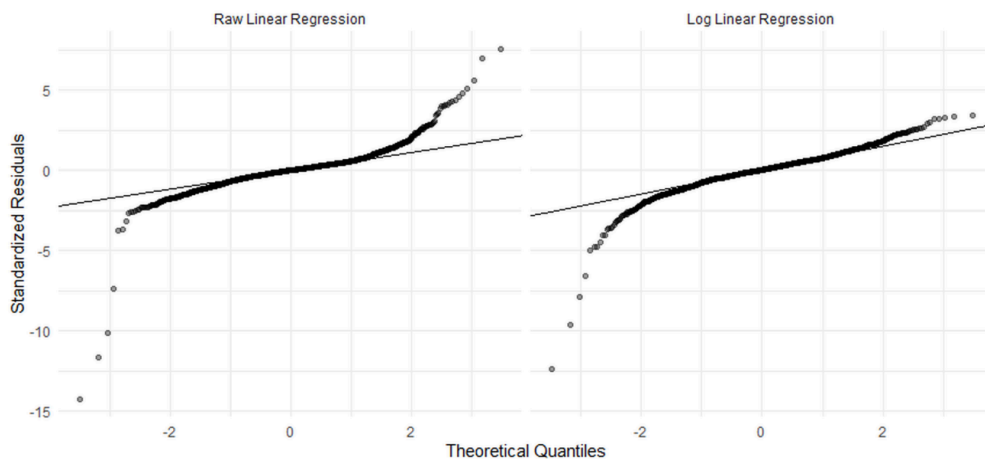


Figure 3. Normal Q-Q diagnostics for the two linear models (picture credit: original)

3.2. Decision tree results

Selecting tree complexity by cross-validation improved test-set performance. RMSE decreased from \$39,022.98 to \$34,532.54, MAE also declined, and test R^2 rose from 0.7745 to 0.8234 (Figure 4). Despite these gains, the tuned tree still underperformed both linear regression models. This indicates that a single recursive partition did not represent the housing-price structure as effectively as the stronger alternatives.

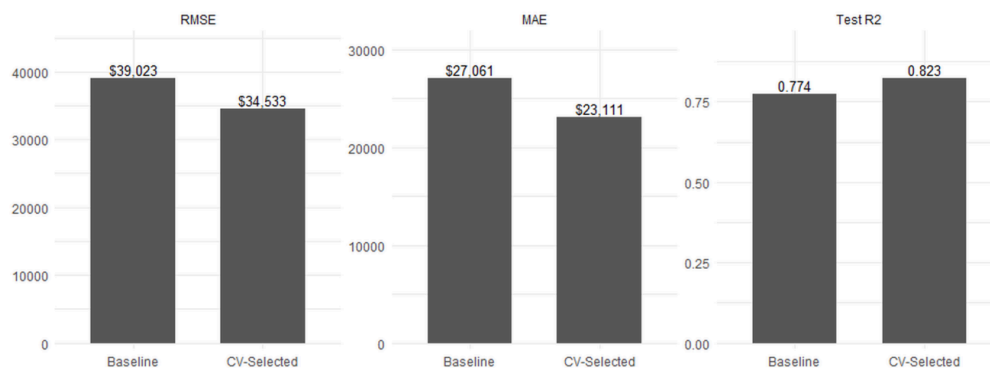


Figure 4. Test performance of the baseline and CV-selected decision tree (picture credit: original)

3.3. Random Forest results

The out-of-bag error declined rapidly as the first trees were added to the forest and then changed much more slowly as the ensemble became larger (Figure 5). Most of the improvement had already occurred by roughly 100–200 trees, after which the curve remained relatively stable. This pattern indicates that using 500 trees was sufficient for the fitted Random Forest and that further increases in forest size were unlikely to produce a meaningful reduction in prediction error.

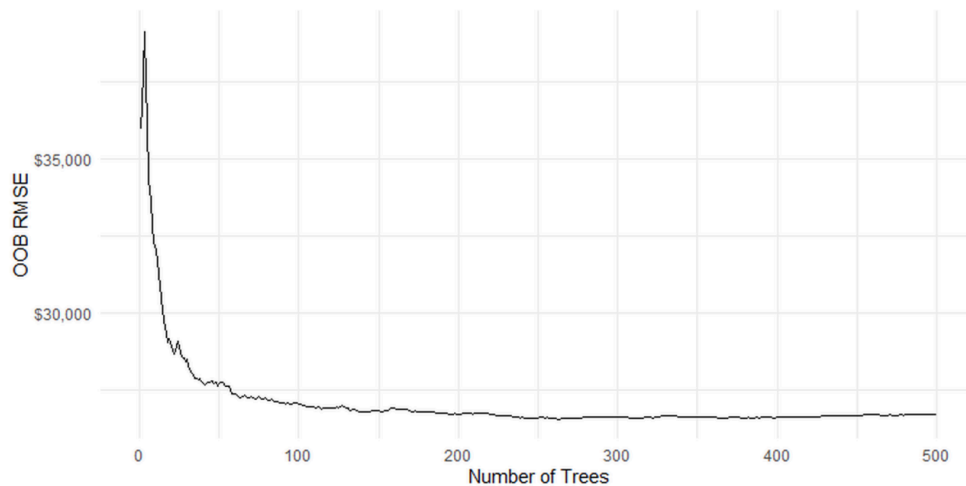


Figure 5. Out-of-bag RMSE across the 500-tree Random Forest (picture credit: original)

On the test set, Random Forest achieved the strongest overall performance among the four models. Its RMSE was \$26,561.86, its MAE was \$17,864.80, and its out-of-sample R^2 reached 0.8955. These values were better than those from the raw linear model, the log-linear model, and the single Decision Tree, although the log-linear specification remained relatively close.

Figure 6 reports permutation-based variable importance for the fitted forest. Above-ground living area produced the largest increase in prediction error when permuted, followed by overall quality and neighborhood. First-floor area, total basement area, and year built formed the next group of influential predictors. This ranking is broadly consistent with the exploratory patterns observed earlier in the analysis, where property size and quality were strongly related to sale price. These importance values describe predictive usefulness within the fitted Random Forest rather than causal influence.

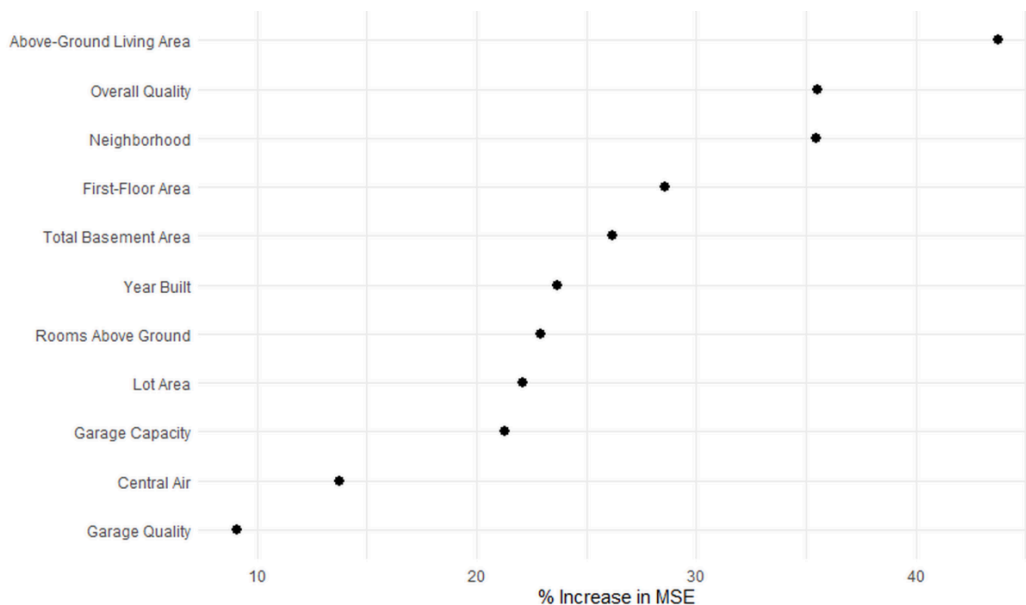


Figure 6. Permutation-based variable importance for the Random Forest (picture credit: original)

3.4. Overall model comparison

Table 1 and Figure 7 summarize the predictive ranking of the four fitted models on the test set. Random Forest produced the best values for all three evaluation measures, while the log-linear model ranked second. The raw linear model followed, and the single Decision Tree gave the weakest overall test performance. Compared with the log-linear specification, Random Forest reduced RMSE by \$1,169.40 and MAE by \$483.06, while increasing out-of-sample R^2 by 0.0094. The improvement was therefore present but not large.

The contrast between the two tree-based methods is particularly important. Although both can represent nonlinear relationships, the single Decision Tree performed substantially worse than either linear specification, whereas the Random Forest achieved the best prediction accuracy. In this analysis, additional flexibility was most effective when predictions were averaged across many trees rather than obtained from one recursive partition. The relatively strong performance of the log-linear model also shows that a simpler specification can remain competitive after an appropriate transformation of the response. Previous work using Ames Housing data has similarly shown that the best-performing method can depend on the models and specifications being compared [10].

Table 1. Test-set model performance

Model	RMSE (\$)	MAE (\$)	Test R^2
Raw Linear Regression	29,623.62	19,349.74	0.8700
Log Linear Regression	27,731.26	18,347.86	0.8861
Decision Tree	34,532.54	23,111.09	0.8234
Random Forest	26,561.86	17,864.80	0.8955

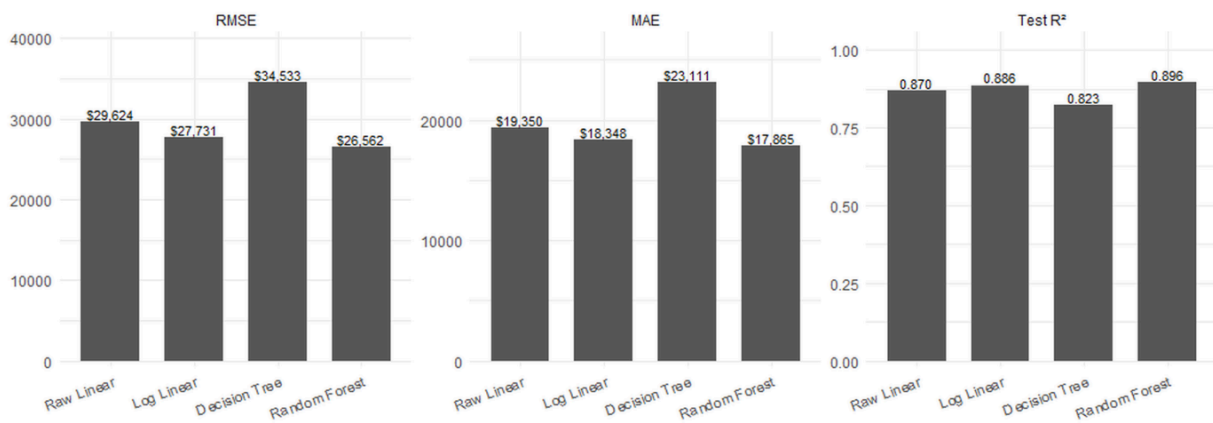


Figure 7. Test-set performance of the four fitted models (picture credit: original)

Figure 8 provides a visual comparison of predictions from the four models. Random Forest predictions lie closest to the 45-degree reference line for much of the test sample. In contrast, the Decision Tree produces several horizontal bands because houses assigned to the same terminal node receive the same predicted value. The largest prediction gaps occur among the most expensive properties. For several transactions above approximately \$500,000, predictions from all four models fall below the observed sale prices. Random Forest therefore remained the strongest model overall, but its advantage was less consistent for these high-price observations. The log-linear predictions

also remained relatively close to the reference line, indicating that the transformed linear specification achieved a useful balance between predictive accuracy and model simplicity.

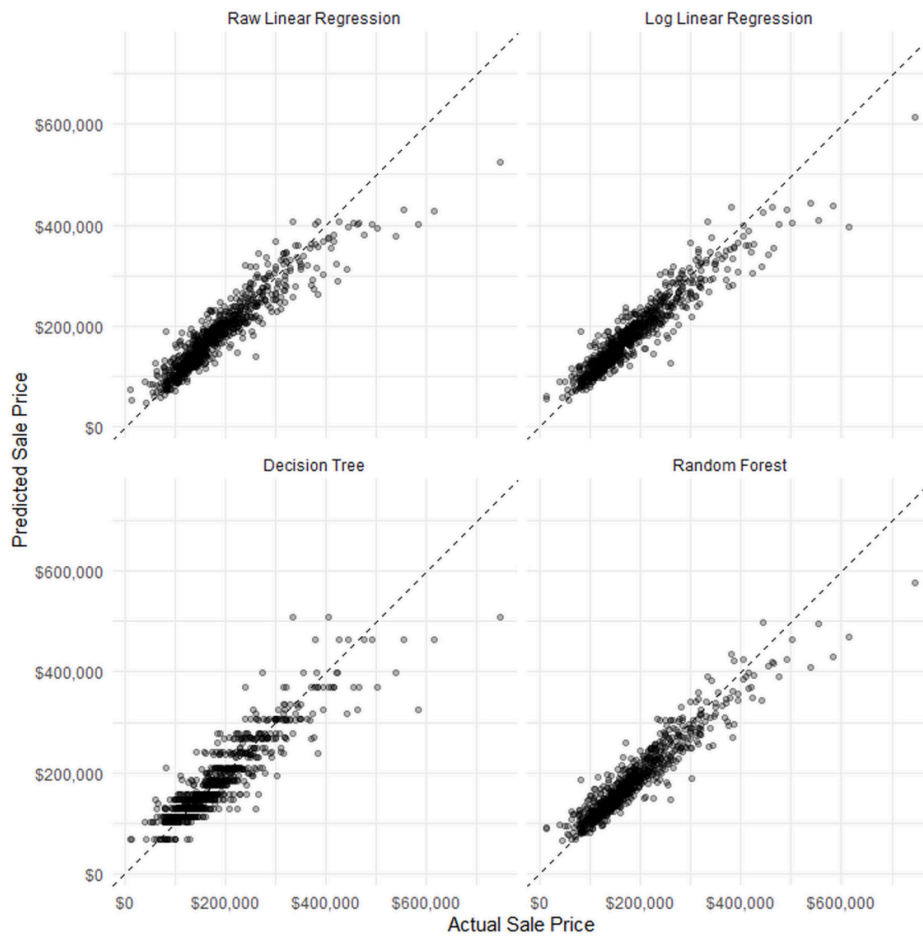


Figure 8. Actual and predicted sale prices on the common test set (picture credit: original)

4. Conclusion

This study compared two linear specifications and two tree-based models for predicting sale prices in the Ames Housing data. Using the same held-out test sample for all four methods, Random Forest produced the strongest predictive results, while the log-linear model ranked second and remained relatively close. The raw linear model performed less well, and the single Decision Tree produced the weakest test accuracy. These results show that allowing nonlinear relationships does not by itself guarantee better prediction. In this analysis, the largest gain came from combining many trees rather than relying on one recursive partition.

The log transformation improved both the predictive performance and the residual behavior of the linear model, making it a competitive alternative despite its simpler structure. For the Random Forest, above-ground living area, overall quality, and neighborhood contributed most strongly to prediction according to permutation importance. At the same time, Figure 8 showed that the most expensive homes remained difficult for every model, with several high-value transactions receiving predictions below their observed sale prices. This suggests that the available structural and

neighborhood variables did not fully capture the factors associated with the highest-priced properties.

Because the analysis is based only on transactions from Ames, Iowa, the conclusions should not automatically be generalized to other housing markets. A useful extension would be to repeat the comparison in other cities and examine whether additional geographic or economic variables improve prediction for expensive properties. More flexible ensemble methods, such as gradient boosting, could also be compared with the Random Forest and log-linear benchmarks used here.

References

- [1] Moreno-Foronda, I., Sánchez-Martínez, M. T., & Pareja-Eastaway, M. (2025). Comparative analysis of advanced models for predicting housing prices: A review. *Urban Science*, 9(2), 32.
- [2] Jaroszewicz, J., & Horynek, H. (2024). Aggregated housing price predictions with no information about structural attributes—Hedonic models: Linear regression and a machine learning approach. *Land*, 13(11), 1881.
- [3] Soltani, A., & Lee, C. L. (2024). The non-linear dynamics of South Australian regional housing markets: A machine learning approach. *Applied Geography*, 166, 103248.
- [4] An, S., Song, Y., Jang, H., & Ahn, K. (2025). Toward transparent and accurate housing price appraisal: Hedonic price models versus machine learning algorithms. *Financial Innovation*, 11, 141.
- [5] Deng, L., & Zhang, X. (2025). Boosting the accuracy of property valuation with ensemble learning and explainable artificial intelligence: The case of Hong Kong. *The Annals of Regional Science*, 74, 32.
- [6] Tarasov, S., & Dessoulavy-Śliwiński, B. (2025). Algorithm-driven hedonic real estate pricing—An explainable AI approach. *Real Estate Management and Valuation*, 33(1), 22–34.
- [7] De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).
- [8] Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383), 605–610. <https://people.stat.sc.edu/hoyen/pastteaching/stat704-2022/Notes/Smearing.pdf>
- [9] Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced machine learning techniques for predictive modeling of property prices. *Information*, 15(6), 295.
- [10] Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics*, 3(1), 30–45.