Estimating Housing Prices using Machine Learning Techniques

Yadhu Soppin and Jip Flinterman

Abstract

This research investigates the use of machine learning regression models to estimate housing prices. We used a Kaggle competition dataset which contained 79 variables, including numerical and categorical variables. Models such as Lasso Regression, Random Forest, and XGBoost were evaluated for their prediction accuracy and generalization. The results showed that Lasso Regression outperformed others in terms of R^2 and RMSE, while Random Forest provided the highest accuracy within 10% of the true sale price. Despite some overfitting, these models demonstrated the potential of machine learning in accurately predicting house prices. The study suggests further work to enhance model performance through improved feature interactions and dataset expansion.

Introduction

In this project, we aim to investigate the following question: *Can machine learning regression models accurately predict housing prices based on a comprehensive set of property features?* Our goal was to generate precise value estimates for homes, as small pricing errors can lead to significant financial consequences in real-world property valuation. Because we are predicting a continuous numerical variable (SalePrice), we approached this as a regression task rather than classification.

We hypothesized that: "Machine learning regression models trained on a comprehensive set of property features can accurately predict house prices with minimal error margins." To test this hypothesis, we used the "House Prices – Advanced Regression Techniques" dataset from a Kaggle competition, which is based on the Ames Housing dataset developed by Dean De Cock (2011). This dataset includes 79 explanatory variables describing residential properties in Ames, Iowa.

The dataset contains both numerical and categorical features, such as LotArea, OverallQual, YearBuilt, GarageCars, and Neighborhood. The target variable, SalePrice, represents the final selling price of each home. The distribution of SalePrice is right-skewed, with most homes clustered at lower price ranges and a smaller number of luxury homes extending the distribution tail. Many variables include missing values,

mixed data types, and differing scales, which required extensive preprocessing. These characteristics make the dataset ideal for testing multiple regression models such as Lasso, ElasticNet, and XGBoost, allowing us to explore their ability to handle complex, high-dimensional housing data.

Methods

To address our research question, we implemented a range of machine learning regression models to estimate housing prices as accurately as possible. Specifically, we trained and evaluated **Linear Regression, Ridge Regression, Lasso, ElasticNet, Random Forest Regressor**, and **XGBoost Regressor**. This diverse selection allowed us to compare both linear and non-linear approaches, as well as assess the impact of regularization and ensemble techniques on predictive performance.

We selected these models due to the dataset's high dimensionality and mixed data types. The traditional Linear regression served as a benchmark model. Other linear models such as Ridge and Lasso are well-suited for this task because they apply regularization to reduce overfitting, with Lasso additionally performing feature selection by shrinking some coefficients to zero. ElasticNet combines the strengths of both Ridge and Lasso, making it effective in datasets with correlated predictors. Tree-based ensemble models like Random Forest and XGBoost were chosen for their ability to capture complex non-linear relationships and interactions between features without extensive feature engineering.

Prior to model training, we applied a range of pre-processing techniques. Missing values were imputed: categorical features using the most frequent value, and numerical features using the mean. We separated the dataset into categorical and numerical features to allow appropriate transformations. Numerical features exhibiting high skewness (absolute skew ¿ 1) were log-transformed using np.log1p() to normalize their distributions and reduce variance instability. Categorical features were one-hot encoded after imputation (using the pd.get_dummies() function).

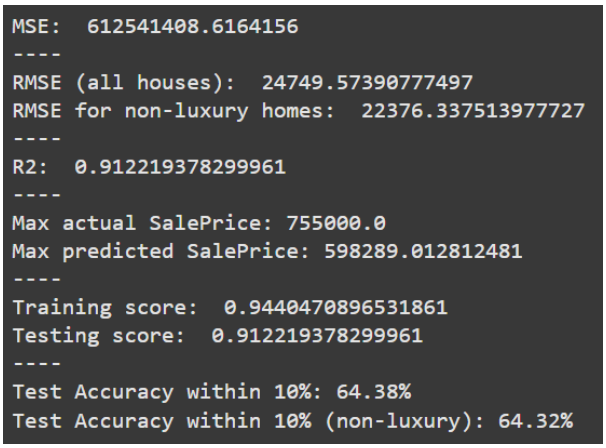
To identify key traits of high-end homes, we performed exploratory analysis on properties in the top 5% of various features and 2% of SalePrice. Based on these patterns, we engineered a binary feature called is_luxury to capture homes with luxury characteristics. In the test, a house was classified as luxury, if its price was in the 98th percentile and above. However, for the test, a house was considered luxury if it had a combination of various features in the 95th percentile, considering SalePrice data was not available in the test dataset. This feature was subsequently included in our models and consistently received one of the highest coefficients, indicating that it played a significant role in price prediction.

The dataset was then split into training and validation sets (70/30 split), and GridSearchCV was used for hyperparameter tuning where applicable — especially for regularized and ensemble models. This ensured that each model was optimally configured before evaluation. Models were assessed using Root Mean Squared Error

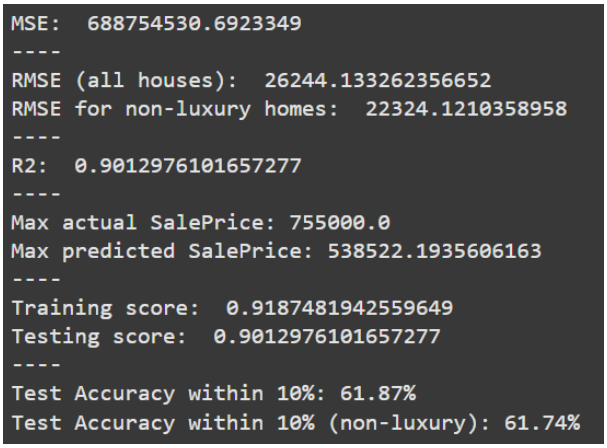
(RMSE) and R^2 score, which measures absolute prediction error and the proportion of variance explained, respectively.

Results

We evaluated multiple regression models to determine which approach could most accurately predict house prices. The models were assessed using key evaluation metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score, and test accuracy within 10% of the actual sale price for all houses and just non-luxury houses (more room for error for higher priced houses).Additionally, we measured performance on both the full dataset and the non-luxury segment (where is_luxury = 0) to assess consistency.

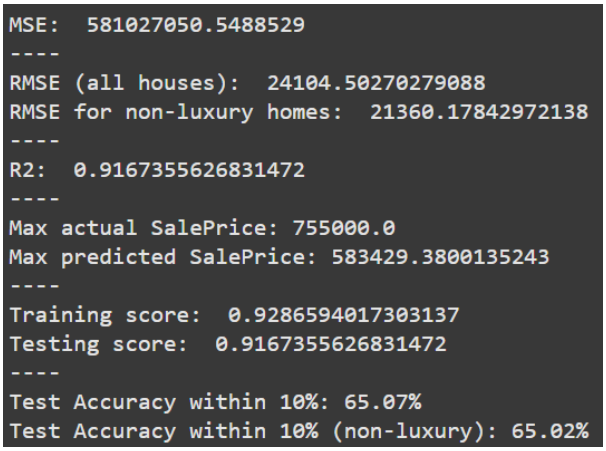


(a) Linear Regression results

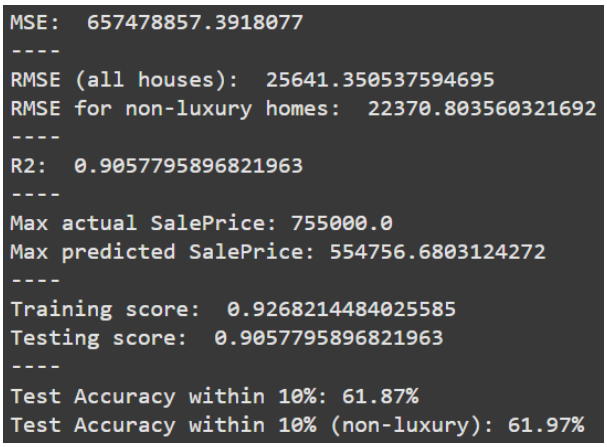


(b) Ridge Regression results

Figure 1: Linear and Ridge Regression Results



(a) Lasso Regression results



(b) ElasticNet Regression results

Figure 2: Lasso and ElasticNet Regression Results

Among the linear models, Lasso Regression performed best, achieving an R^2 of 0.917, an RMSE of 24,104,

and a test accuracy of 65.07% (Fig. 2a). Lasso also demonstrated strong generalization, with nearly identical accuracy on non-luxury homes, and its regularization effect led to sparse, interpretable coefficients. Linear Regression and Ridge followed closely, though Ridge showed slightly higher error and lower accuracy (61.87%). ElasticNet, while balanced, did not outperform Lasso or Ridge, with a test R^2 of 0.906 and an accuracy of 61.87% (Fig. 1, 2b).

```
MSE:  598804075.468891
----
RMSE (all houses):  24470.4735440263
RMSE for non-luxury homes:  20585.202551296596
----
R2:  0.9141880152398114
----
Max actual SalePrice: 755000.0
Max predicted SalePrice: 559377.35
----
Training score:  0.9814038120157591
Testing score:  0.9141880152398114
----
Test Accuracy within 10%: 67.35%
Test Accuracy within 10% (non-luxury): 68.08%
```

(a) Random Forest results

```
MSE:  708840796.1158155
----
RMSE (all houses):  26624.064229861968
RMSE for non-luxury homes:  20937.84275363138
----
R2:  0.8984191355977997
----
Max actual SalePrice: 755000.0
Max predicted SalePrice: 557005.6
----
Training score:  0.9997522148253891
Testing score:  0.8984191355977997
----
Test Accuracy within 10%: 67.81%
Test Accuracy within 10% (non-luxury): 68.31%
```

(b) XGBoost results

Figure 3: Random Forest and XGBoost Results

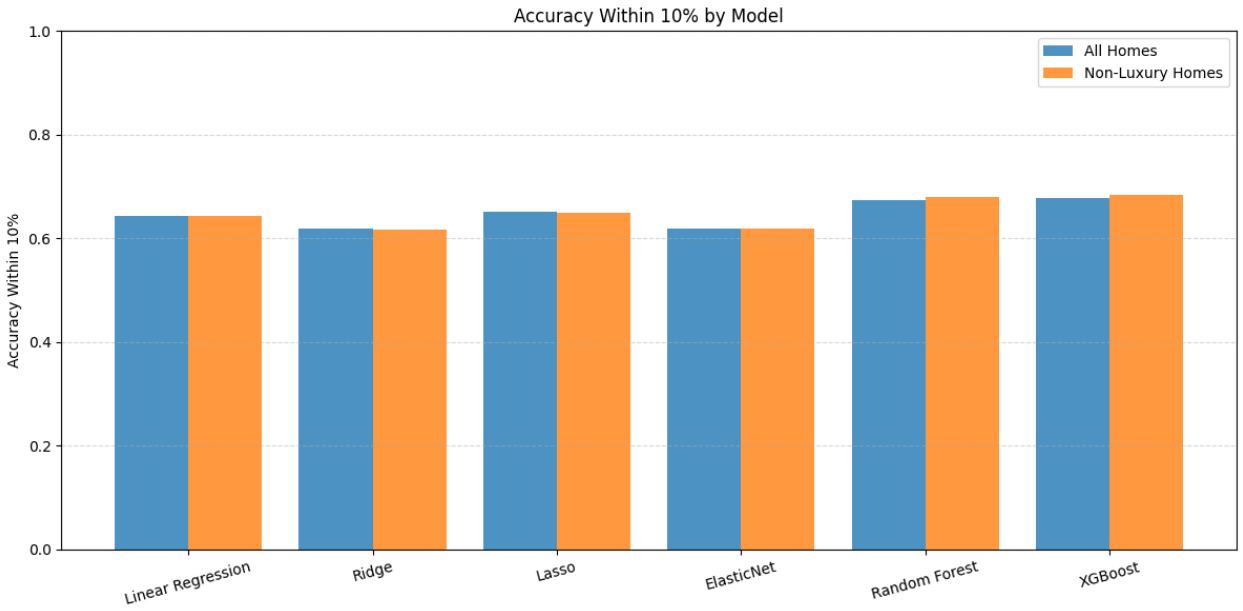


Figure 4: Accuracy Comparisons

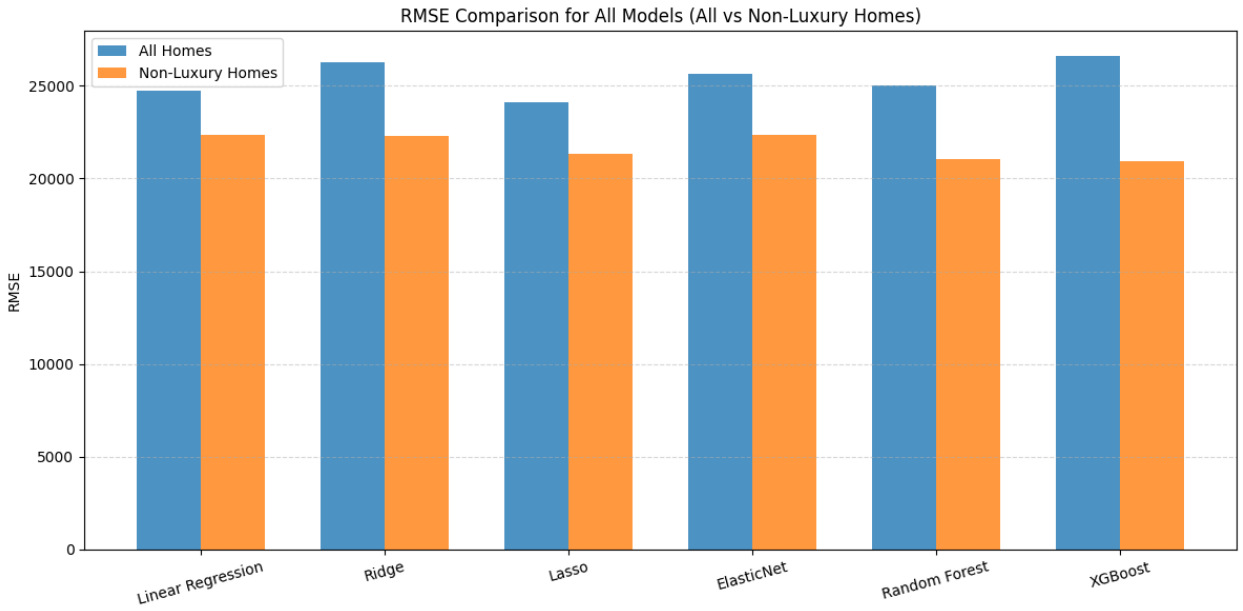


Figure 5: RMSE Comparisons

Ensemble models delivered superior results with regards to accuracy (Fig. 4). Random Forest achieved an R^2 of 0.914 with the lowest RMSE for non-luxury homes (20,585) and a test accuracy of 67.35%, outperforming all linear models (Fig. 3a). It was particularly strong at modelling typical home prices but showed mild overfitting (training $R^2 = 0.98$). XGBoost delivered the highest accuracy overall, with 67.81% of predictions within 10% of the actual SalePrice, and 68.31% for non-luxury homes. Despite near-perfect training performance, it maintained high generalization ($R^2 = 0.898$) (Fig. 3b).

Overall, while Random Forest and XGBoost achieved the highest accuracies and lowest RMSEs overall, the Lasso model demonstrated stronger generalization and a smaller RMSE gap between all homes and non-luxury homes. Given our aim to balance raw predictive power with stability across the price spectrum, Lasso proved to be the more suitable model (Fig. 5).

Discussion

Our research focused on accurately estimating housing prices using machine learning regression techniques on a complex dataset. The results affirmed our hypothesis: regression models can be tuned to generate precise price predictions, not just broad value ranges.

Among the linear models, Lasso Regression stood out by achieving the highest test R^2 score (0.917) and one of the lowest RMSE values (24,104). It also demonstrated strong generalization, with nearly identical performance on non-luxury homes, and flagged fewer outliers. This suggests that Lasso effectively balanced the bias-variance trade-off while maintaining interpretability.

Ensemble models offered better raw accuracy. XGBoost achieved the highest test accuracy within 10% at 67.81%, closely followed by Random Forest at 67.35%. However, both models showed signs of overfitting, with training R^2 scores near 1.0. This performance gap highlights a key limitation of tree-based models: while powerful, they can be overly sensitive to noise and easily overfit without careful regularization or pruning.

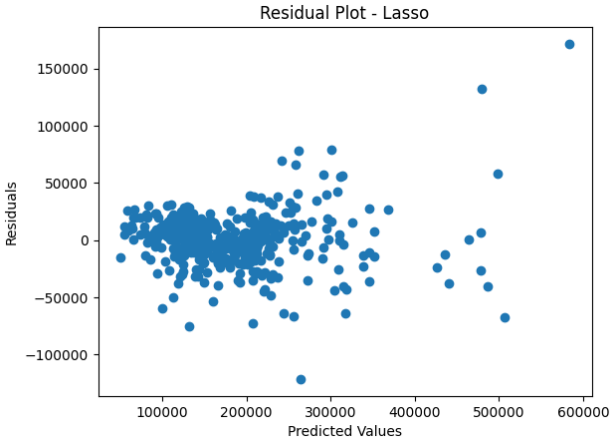


Figure 6: Residual Plot - Lasso

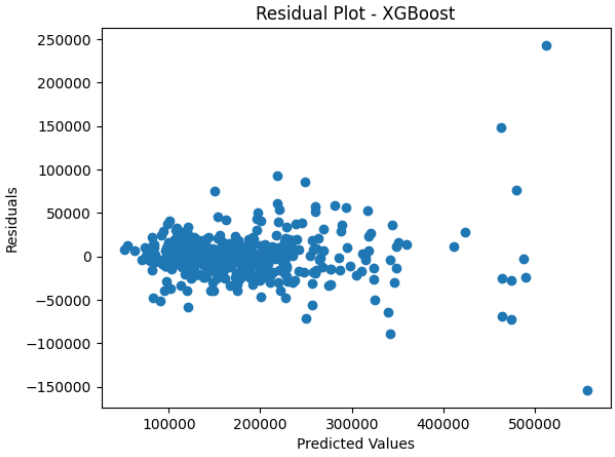


Figure 7: Residual Plot - XGBoost

Interestingly, simpler models like Lasso and Ridge performed nearly as well and were more stable. However, residual plots (Fig. 6 & 7) revealed a common trend across all models — heteroskedasticity. Residuals were tightly clustered for mid-range predictions but increasingly dispersed for higher-priced homes. This pattern indicates the models struggled with generalizing to luxury homes, likely due to their low representation in the dataset and more variable feature combinations.

However, a major strength of our approach was the thorough pre-processing pipeline, including imputation,

normalization of skewed features, and one-hot encoding of categorical variables. The engineered is_luxury feature proved useful across all models and consistently ranked as a top predictor.

That said, our work had limitations. We did not use an external test set beyond the validation split, which may have introduced some bias during hyperparameter tuning. Additionally, the dataset reflects a specific region and time frame (Ames, Iowa), limiting generalizability. Future work could address these issues by introducing cross-validation across multiple folds, enhancing feature interaction terms, and experimenting with alternative encoding strategies such as target encoding.

Surprisingly, despite its complexity, XGBoost did not outperform simpler models like Lasso in terms of RMSE or R^2. This suggests that model complexity does not guarantee better generalization, especially when the dataset size or diversity is limited.

Conclusion

In this research, we aimed to estimate housing prices accurately using machine learning regression models trained on a dataset with a wide range of property features. Our results confirmed that several models—including Lasso Regression, Random Forest, and XGBoost—were able to deliver strong predictive performance. Among these, Lasso Regression emerged as a top performer, achieving an R^2 of 0.917 and 65.07% accuracy within 10% of actual prices, while maintaining strong generalization. Random Forest and XGBoost achieved even higher prediction accuracy within 10%, though both exhibited signs of overfitting.

These findings support our hypothesis: machine learning regression models can predict house prices with relatively minimal error when properly tuned and preprocessed. However, we also observed consistent underestimation of high-end home prices across all models, suggesting challenges in generalizing to rare or extreme cases—likely due to limited representation in the dataset.

To improve model performance further, especially on luxury properties, future research should explore richer feature engineering, such as incorporating interaction terms between categorical and numerical variables. Additionally, expanding the dataset to include more high-priced homes could reduce bias and improve model generalizability across the full price spectrum. We also recommend implementing external cross-validation, rather than relying solely on a train/validation split, to obtain a more robust estimate of model performance. Finally, experimenting with alternative encoding strategies (e.g., target encoding) may enhance the learning process for complex categorical variables.

Word Count (total excluding titles): **1888**