Enhancing House Price Predictions: A Comparative Analysis of Feature Selection Algorithms and the Lasso Model

Pavlo Mysak and Alexandra Kassian

School of Business: State University of New York at New Paltz

BUS 461 Business Analytics Capstone

Dr. Tao Li

October 9, 2024

Abstract

Predicting house prices with accuracy is a major challenge in real estate, traditionally approached through hedonic pricing models and machine learning (ML) models that highlight variable importance. However, the effectiveness of these conventional methods has not been compared to feature selection algorithms, which offer a systematic way to identify impactful predictors. This study evaluates various feature selection techniques, including Select K Best, Variance Threshold, and Recursive Feature Elimination (RFE) with Lasso Regression, to determine their ability to optimize linear regression models. We use Root Mean Squared Error (RMSE) as our primary performance metric, applying 5-fold cross-validation for robustness. By creating scenarios where the number of features exceeds the number of observations (m > n) and vice versa (n > m), we assess the stability of these selection methods. Our results indicate that feature selection algorithms not only enhance predictive performance compared to traditional hedonic and ML-based approaches, but also yield a new way to derive important housing features and their relation to prices.

Keywords: Feature Selection Algorithms, Lasso Regression

Introduction

Hedonic pricing models have been applied to the housing market for decades. The earliest references of the application often refer back to the 1920s researchers known as G.C. Hass and H.A. Wallace. Many credit G.C. Hass, from the University of Minnesota, as the first researcher to use the hedonic model to express price as a function of product characteristics. Although there have been doubts about whether G.C. Hass made a true significant impact, other research shortly followed behind (Colwell & Dilmore, 1999). H.A. Wallace, a researcher from the University of Iowa, and A.T. Court, an automobile industry analyst, both contributed to the development of modern hedonic analysis (Colwell & Dilmore, 1999). Their work laid the foundation for more refined approaches in the field. More specifically, A.T. Court (1939) is credited for pioneering the hedonic price analysis as his paper addressed the problems of nonlinearity and changes in underlying goods (Goodman, 1998). Griliches (1961) is credited for popularizing the model for modern use with his hedonic function for the prices of automobiles (Triplett, 2007). As research advanced, S. Rosen published a paper in 1974, formalizing the concept of hedonic pricing and providing a theoretical framework to explain how product prices can be broken down into the value of individual characteristics (Rosen, 1974). The hedonic pricing model became widely used in areas such as the car market, personal computer market, and primarily, the housing market.

Today, the total value of the residential housing market in the United States is estimated at $51.94 trillion, according to Zillow.com (Orphe Divounguy, 2023). Given that homes are a basic necessity and a part of daily life, it's no surprise that the housing market is one of the largest globally. Homes are often regarded as "bundles of utility-bearing attributes" that provide value and satisfaction to consumers (Kim et al., 2015). Given this significance, identifying the most sought-after attributes is important for both investors and consumers alike. However, most research focused on predicting house prices relies on either manually selecting variables for modeling or using ML models to assess variable importance. Although this approach can yield valuable insights, manual variable selection can introduce bias and ML models may run the risk of overfitting (Lahmiri et al., 2023). These limitations have been recognized throughout research, with many turning to ensemble models for their variable importance. Models such as random forests and other tree-based methods have the ability to handle both categorical and numerical data well, but there are clear limitations. Despite their flexibility, these models face challenges in accurately determining reference parameters, depending on the quality and availability of input data, and requiring increased computational time (Tran et al., 2008). Additionally, it is recommended that the reliability of the ensemble be verified to understand any represented uncertainties (Kim et al., 2020).

To avoid the issues underlying ML and specifically ensemble methods, the Least Absolute Shrinkage and Selection Operator (Lasso) model is an alternative for hedonic pricing feature selection. Unlike traditional methods, the Lasso model is a type of linear regression that employs regularization to improve the performance and minimize the prediction error (Fonti & Belitser, 2017). Regularization helps avoid overfitting by imposing a constraint on the coefficients of the model and shrinking some coefficients to zero. This characteristic of Lasso allows for the automatic selection of features as there will still be non-zero coefficients remaining after the regularization (Fonti & Belitser, 2017). This regularization term, known as the L1 penalty, is equal to the absolute value of the coefficients and is added to the loss function to form the overall cost function. The λ parameter controls the trade-off between fitting the training data and simplifies the model by penalizing large coefficients. If the penalty is weak, the model behaves more like a linear regression. The model may fit the data well and most of the coefficients will remain non-zero, but there is a high likeliness of overfitting. If the penalty is strong, more coefficients are pushed toward zero which reduces the number of variables. A strong penalty can prevent overfitting and improve the generalization of unseen data. Choosing the correct λ is usually done through cross-validation, where different values of λ are tested to find the one that minimizes the prediction error on testing data. For the purposes of our study, we employed the Lasso regression model as our benchmark model.

Numerous studies have effectively utilized Lasso regression for predicting housing prices and, ultimately, conducting feature selection. For example, Mohd et al. (2020) explored various real estate modeling techniques and found that Lasso regression was particularly effective in reducing multicollinearity. Their study highlighted the model’s ability to narrow down key predictors. Similarly, Mathotaarachchi et al. (2024) applied Lasso alongside other advanced ML methods and noted its ability to perform efficient feature selection. They concluded that Lasso’s feature selection capabilities allowed for a more precise focus on impactful variables. Vishwakarma and Singhal (2020) implemented Lasso within a hybrid multi-regression model, where it had one of the best scores in comparison to other models. Their findings underscored Lasso’s effectiveness in improving both performance and speed of prediction models. Lastly, Mullainathan and Spiess (2017) incorporated the Lasso model in their analysis and found that Lasso outperformed the ordinary least squares model in out-of-sample prediction accuracy. The study highlighted the model feature selection characteristics particularly in high-dimensional settings where multicollinearity was present. Collectively, these studies emphasize the robustness and adaptability of Lasso regression as a valuable method for analyzing housing markets.

The Lasso regression model is a favorable choice for modeling the housing prices, but our study extends beyond the feature selection properties of this model. Choosing the right features in the development of a predictive model is a crucial process, as it can fundamentally alter the direction and outcomes of the research. Numerous studies have underscored the significance of effective feature selection algorithms in uncovering meaningful patterns while minimizing computational costs. In our study, we employ the Select K Best, Variance Threshold, Recursive Feature Elimination and Select From Model algorithms to pinpoint the key predictors of housing prices. The Select K Best method ranks features based on statistical scores derived from their relationship with the target variable, utilizing mutual information for evaluation. We also applied the Variance Threshold technique, which eliminates features with low variance, thereby retaining those that contribute meaningful variability. Additionally, we implemented Recursive Feature Elimination (RFE) with Lasso regression, which iteratively ranks features based on importance while leveraging Lasso's L1 regularization to encourage sparsity. SelectFromModel, which selects features based on importance scores from trained models, was implemented, similarly with Lasso regression, as well as decision trees. SelectFromModel with decision trees was ultimately excluded from our analysis due to its limited overlap with features identified by other methods. By exploring these feature selection algorithms, we aim to construct a robust framework for identifying determinants in the housing market.

Although there is a sizable amount of literature on feature selection algorithms, the application of these algorithms toward housing price prediction is limited. Despite this, it is important to recognize the associated literature with these algorithms to support our analysis. In a study by Pudjihartono et al. (2022), the authors employed the Select K Best method to filter significant variables for predicting disease risk, showcasing its ability to streamline datasets and improve model accuracy. This research illustrates the method's effectiveness in identifying key predictors, which could similarly enhance housing price predictions. In the work by Patil et al. (2023), the authors combined correlation analysis with the Variance

Threshold method to enhance the diagnosis of myocardial infarction, demonstrating how this approach eliminates irrelevant features while retaining those with meaningful variance. The authors showcased that using variance-based approaches can reduce noise in data, which is a concept we can apply to our housing data. Furthermore, in the study by Darst et al. (2018), Recursive Feature Elimination (RFE) was employed alongside random forest models to identify key predictors in high-dimensional datasets, emphasizing its strength in managing correlated variables. Additionally, García-Magariño et al. (2019) explored the use of RFE in estimating missing real estate prices through agent-based simulations, reinforcing its role in optimizing predictive accuracy. These studies on RFE indicate the potential for the method to address gaps in housing price data and enhance overall predictive reliability. Lastly, Chanasit et al. (2021) developed a real estate valuation model that utilized boosted feature selection techniques, illustrating the importance of effective feature selection in enhancing model performance. Their results further validate the need for robust feature selection methods in real estate analysis, emphasizing its relevance for housing price modeling.

The purpose of this study is to assess whether specific feature selection algorithms are beneficial in reducing predictive error and to compare their performance against a baseline Lasso regression model. To compare the predictive error of the feature selection algorithms against the benchmark Lasso model, we employed the Root Mean Squared Error (RMSE) to be our metric for comparison. Our research will evaluate the effectiveness of these algorithms across two distinct scenarios: a traditional approach with high-dimensional data, and a subset-based data partition approach. We recognize that high-dimensional data is susceptible to overfitting due to feature variability and noise, so it is important to examine how different algorithms manage these challenges ("Feature Extraction," 2006; Kalina & Schlenker, 2015). To compare the features selected by the algorithms with those from the benchmark Lasso model, we will assess the overlap of selected features across algorithms and then analyze their feature overlap with the benchmark. In addition, our study will examine the significance and importance of features selected by these algorithms and compare them to the benchmark. Whether these algorithms produce similar feature importance rankings to Lasso would validate their utility in producing reliable and interpretable models without significantly increasing computational complexity.

Our research aims to provide valuable insights into the effectiveness of various feature selection algorithms, particularly in comparison to the benchmark Lasso model. By validation these algorithms, we can determine whether they produce results similar to those of the hedonic pricing model. These results can enhance the credibility and practical application of these algorithms. Our study supports the development of more interpretable predictive models for real estate and therefore improves pricing strategies and investment decisions. Additionally, the findings extend beyond real estate, offering methodologies that can be applied to other fields dealing with high-dimensional data and feature selection. Our work contributes to a broader understanding of how these algorithms perform across different conditions, and we hope to aid researchers and data scientists in making informed decisions. Moreover, our research addresses challenges in predictive modeling by managing multicollinearity and balancing model complexity with interpretability, which is particularly important in the era of big data and complex models.

Methodology

Lasso Regression Model

The Lasso (Least Absolute Shrinkage and Selection Operator) regression model is used as the benchmark model in this study due to its ability to perform both regularization and feature selection. By applying an L1 penalty to the regression coefficients, Lasso encourages sparsity as it reduces some coefficients to zero. The penalty process is effective as it selects a subset of the most relevant features. This helps prevent overfitting, especially in high-dimensional data scenarios, where the number of features exceeds the number of observations.

The model optimizes the following cost function:

$$min\left(\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \; + \; \lambda\sum_{j=i}^{p}|\beta_j|\right)$$

where y_i represents the observed values, $\widehat{y}_i$ represents the predicted values, λ represents the tuning parameter control the strength of regularization, and β_j represents the model coefficients.

A larger λ increases regularization, shrinking more coefficients to zero. This makes the Lasso an ideal benchmark model for evaluating the performance of other feature selection algorithms in this study.

Feature Selection

Feature selection is an important step in the development of predictive models, particularly when working with high-dimensional datasets where the number of features can exceed the number of observations. It involves selecting a subset of relevant variables (features) from the dataset, aiming to improve model interpretability, reduce overfitting, and enhance predictive accuracy. The process is especially important in regression and classification tasks where irrelevant or redundant features can obscure the underlying patterns and inflate computational costs. This section provides an overview of the feature selection algorithms employed in this study, their mechanisms, and a critical evaluation of their strengths and weaknesses. Additionally, we discuss the *SelectFromModel* method and why it was ultimately excluded from the final analysis.

The *Select K Best* method ranks features based on statistical scores associated with their relationship to the target variable. In this study, mutual information regression was used to evaluate and rank features, selecting the top K based on their individual performance in predicting house prices. Mutual information measures the dependency between two variables, capturing both linear and non-linear relationships. One advantage of *Select K Best* is its simplicity and efficiency. It directly ranks features based on their association with the target variable, making it an effective method when a quick assessment is needed. Additionally, its ability to capture non-linear dependencies makes it versatile for various data types. However, the algorithm evaluates each feature in isolation, ignoring potential interactions among features that could be significant in combination. Consequently, it may overlook important variables that become relevant only in the presence of others, limiting its capacity in scenarios where complex interactions drive the outcomes. Mutual information (MI) quantifies the amount of information obtained about one random variable through another. For continuous distributions, it is expressed as:

$$I(X;Y) \;=\; \int_{y}\int_{x} P_{(X,Y)}(X,Y) log(\frac{P_{(X,Y)}(X,Y)}{P_X(X)P_Y(Y)})dx\,dy$$

where $P_{(X,Y)}$ *(X,Y)* represents the joint probability density function (PDF) of *X* and *Y*, while P_X *(X)* and P_Y *(Y)* denote the marginal PDFs of *X* and *Y*, respectively. This integral calculates the sum of information for all possible combinations of values that *X* and *Y* may take.

The *Variance Threshold* method is an unsupervised feature selection approach that removes features with low variance, under the assumption that low-variance features are less likely to contain useful information. In this study, features with variance below a specified threshold were eliminated, ensuring that the remaining features contribute meaningful variability to the model. This method is straightforward and computationally efficient, as it requires only a simple variance calculation. It is particularly useful for eliminating constant or near-constant features that do not offer increased predictive power. The limitation of the *Variance Threshold* method is its insensitivity to the relationship between features and the target variable. Since it operates without considering how features relate to the outcome, it may retain irrelevant features if they exhibit sufficient variance or discard relevant ones with lower variance. Furthermore, it does not account for interactions among features, which may lead to suboptimal selection in complex datasets.

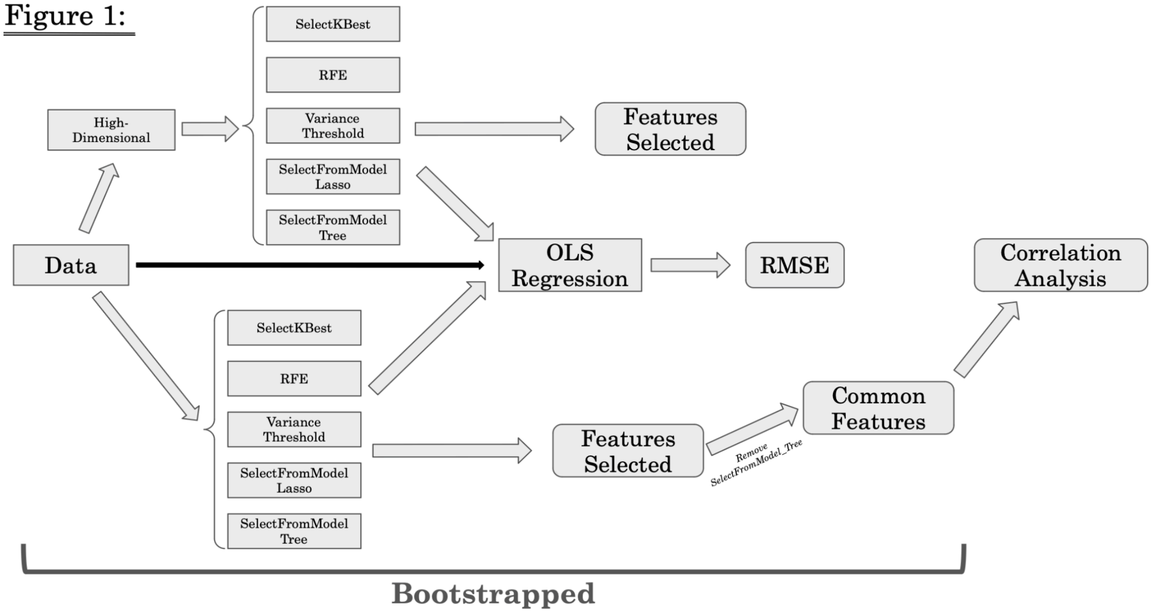
Recursive Feature Elimination (RFE) is a more iterative and sophisticated method for selecting features. It works by recursively fitting a model (in this case, Lasso Regression), ranking the features based on their importance, and then eliminating the least important ones. The process repeats until the desired number of features is selected. By pairing *RFE* with Lasso Regression, the algorithm benefits from the L1 regularization properties of Lasso, which penalizes less important coefficients, encouraging sparsity in the model. *RFE* with Lasso Regression is effective in handling high-dimensional data where there are more features than observations (m > n). The L1 penalty not only assists in reducing the feature set but also in improving model interpretability by retaining only the most significant features. This approach is particularly valuable in cases where it is necessary to identify a minimal set of predictors with strong explanatory power. A notable disadvantage of *RFE* is its computational intensity, especially in larger datasets. The iterative nature of the method means that it may require substantial computational resources as the number of features increases. Moreover, *RFE's* reliance on the choice of model (Lasso in this case) means that its effectiveness is tied to the performance of the underlying estimator. If the Lasso model struggles due to multicollinearity or other issues, *RFE* may yield suboptimal results.

SelectFromModel is an algorithm that selects features based on the coefficients or importance scores derived from a trained model. In this study, *SelectFromModel* was initially tested using two types of base estimators: a linear model (Lasso) and a non-linear model (Decision Tree Regressor). The algorithm selects features by retaining those with coefficients or importance scores above a specified threshold, while features with lower values are eliminated. This method leverages the interpretability of model coefficients (in the case of Lasso) or feature importance measures (in the case of Decision Trees) to determine which predictors are most relevant. When using Lasso, *SelectFromModel* aimed to capitalize on Lasso's sparsity-inducing properties, where features with the smallest coefficients (representing weak predictors) are set to zero. This straightforward mechanism provides a computationally efficient way to reduce the feature set while maintaining interpretable results. The Decision Tree Regressor, on the other hand, evaluates feature importance based on how much each feature reduces impurity across tree splits. This non-linear approach captures more complex interactions and relationships between features, making it particularly useful when linear assumptions are insufficient.

Despite these advantages, *SelectFromModel* using the Decision Tree Regressor—was removed from the final analysis. The features selected by this variation of *SelectFromModel* showed minimal overlap with those chosen by other algorithms, reducing the number of common variables identified. This limited the ability to confidently assert the importance of recurring features across methods, which is crucial for establishing reliable and generalizable predictors of house prices.

To provide an overview of the experimental design, Figure 1 presents a flow chart outlining each stage of the methodology, from feature selection through to model training and evaluation.

Figure 1:



Data Preprocessing and Transformation

The dataset used in this study comprises detailed information on housing attributes, including structural features (e.g., number of rooms, square footage, and age), lot characteristics (e.g., lot size and frontage), and neighborhood indicators (e.g., proximity to amenities). To prepare the data for analysis, several preprocessing steps are performed:

Missing data is addressed using context-specific imputation methods. For numerical variables, missing values are filled using group-based transformations or mean imputation, while categorical variables are replaced with a placeholder value indicating absence (e.g., "None").

Categorical features are converted into numerical format using one-hot encoding to make them compatible with machine learning algorithms. This transformation increases the dimensionality of the dataset, which is particularly relevant when analyzing scenarios where the number of features exceeds the number of observations.

The target variable (SalePrice) is log-transformed to mitigate skewness and stabilize variance, enhancing the linearity of the relationship between predictors and the target, and thus improving model performance.

Scenario Setup: Feature-to-Observation Ratio Variants

To explore the impact of different feature-to-observation ratios on the performance of feature selection algorithms, two distinct scenarios are established. These scenarios are designed specifically to evaluate the robustness and consistency of the feature selection algorithms themselves, independent of the model training process.

High-Dimensional Scenario (m > n):

In this setup, the dataset is manipulated to create a high-dimensional environment where the number of features (m) exceeds the number of observations (n). This is achieved by randomly selecting a subset of observation indices, ensuring that the number of observations (n) is smaller than the number of features (m). Specifically, the number of observations is set to $n = m - 50$, with consistency maintained across algorithms by using the same set of random indices for each evaluation iteration. This scenario is implemented solely in the feature selection phase to assess the robustness of the algorithms under high-dimensional conditions, mimicking real-world applications such as genomic studies or high-resolution spatial data where feature selection is crucial for reducing dimensionality, mitigating overfitting, and enhancing model interpretability.

It is important to note that this high-dimensional configuration is not applied to the model training process itself; the manipulation is limited to the feature selection phase to isolate the performance of the algorithms without influencing the subsequent modeling results. This approach ensures that the observed effects are attributable to the feature selection algorithms' capabilities in handling challenging, high-dimensional scenarios.

Traditional Scenario (n > m):

In the second scenario, the original dataset structure is preserved, where the number of observations (n) exceeds the number of features (m). This scenario reflects a conventional modeling environment commonly encountered in many regression analyses, providing a baseline against which the performance of feature selection algorithms can be evaluated. The entire dataset is used in this setup, allowing the feature selection algorithms to operate under typical conditions where they have sufficient data to accurately identify the most important predictors. This scenario serves as a control, demonstrating how algorithms perform when they are not constrained by high-dimensionality issues, thus allowing for a comparison of their stability and consistency across varying data conditions.

The study evaluates feature selection algorithms in both high-dimensional (m > n) and conventional (n > m) environments to assess their robustness and reliability across different data contexts. By testing these algorithms under varying scenarios, the analysis determines their sensitivity to changes in data structure. An algorithm that consistently identifies the same features in both high-dimensional and traditional setups demonstrates robustness, suggesting that it can be trusted in various practical situations where data configurations may not be ideal or may differ significantly from training to deployment.

Bootstrapping Design

Bootstrapping is a central component of our analysis, serving as a robust resampling technique to quantify variability and provide reliable estimates of performance metrics in different scenarios. The primary purpose is twofold: (1) to calculate the Root Mean Squared Error (RMSE) and its confidence intervals, and (2) to assess the variability in selected features and its impact on RMSE, particularly in the high-dimensional scenario where feature variability, and thus performance variation, is expected to be more pronounced. To ensure the robustness of RMSE values and to account for random variability across different data samples, we employ bootstrapping with a high number of resamples (100 iterations). The bootstrapping process involves repeatedly sampling with replacement from the original dataset, creating multiple subsets on which the models are trained and evaluated. This approach allows us to obtain a distribution of RMSE values rather than a single point estimate. This distribution of RMSE values gives us a more reliable estimate of model performance compared to using a single dataset split. It also enables the calculation of confidence intervals for RMSE, offering insights into the uncertainty and variability in model performance. By capturing the 2.5th and 97.5th percentiles, we derive 95% confidence intervals that allow us to understand the range within which the true RMSE is likely to fall. This is crucial for assessing the stability of the models, particularly under different feature selection techniques.

In addition to deriving a distribution of evaluation values, we can similarly derive a distribution of percentages of common features in the high-dimensional scenario and the traditional scenario. The methodology involves applying each feature selection algorithm to multiple bootstrap samples and recording the selected features. We then calculate the percentage of features that overlap with those selected from the full dataset in each bootstrap iteration. By averaging these percentages and constructing confidence intervals, we further assess the stability and consistency of feature selection algorithms. This analysis provides valuable insights into which algorithms are more reliable in consistently selecting relevant features, even when the data structure varies.

Evaluation Metric

Root Mean Squared Error (RMSE) is selected as the primary evaluation metric. RMSE provides a direct measure of model accuracy, penalizing larger errors more significantly. This choice is appropriate given its widespread use in regression model evaluation, making it easy to compare results with other studies. RMSE's straightforward interpretation—representing the average magnitude of errors in the model's predictions—allows for clear insight into model performance and consistency across different feature selection techniques. By calculating RMSE for each fold in the cross-validation process and averaging these results, the study obtains a confidently representative measure of the model's performance. To convert our RMSE calculation to be in the same unit as Prices (dollar amounts), we use the following formulation to exponentiate the predictions before calculating our evaluation metric:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e^{y} - e^{\hat{y}})^{2}}$$

Consistency and Common Feature Analysis

For each algorithm, the percentage of common features identified across the high-dimensional (m > n) and traditional (n > m) scenarios is calculated. This consistency analysis assesses whether the same variables are selected despite variations in the feature-to-observation ratio. We utilize bootstrapping to obtain confidence intervals for this measure. A high percentage of overlap indicates that the algorithm is stable and reliable, capable of consistently identifying important predictors regardless of the data configuration. This is crucial in real-world applications where datasets can vary significantly in terms of dimensionality, and a reliable feature selection method must maintain its effectiveness across these variations. The overlap of selected features among different algorithms in the traditional scenario (n > m) is also evaluated. By examining the common features across these methods, the study aims to identify the most reliable and significant predictors that hold up under various selection processes. Variables that are consistently selected by multiple methods are likely to be strong predictors of house prices, as they show stability and relevance across diverse feature selection criteria and reinforce their importance beyond the constraints of any single algorithm. The degree of overlap among selected features provides a basis for validating the effectiveness of the feature selection algorithms themselves. If algorithms converge on a core set of predictors, it suggests that they are effectively capturing the essential information in the data, enhancing their credibility as reliable tools for feature selection in real estate modeling.

Correlation Analysis

Correlation analysis forms a critical part of our study, aimed at understanding the relationships between the selected features and the dependent variable— log sale price. After applying various feature selection algorithms and identifying features that consistently emerge as important across bootstrap samples, we proceed to evaluate their correlation with the dependent variable. This analysis is not only useful for validating the relevance of the selected features but also for interpreting the nature and strength of their relationships with the outcome. To quantify the direction and strength of the relationship between the selected features and the sale price, we calculate the Pearson correlation coefficient. This coefficient is a standardized measure, ranging from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values around 0 suggest no linear relationship. By computing these correlation coefficients for each feature identified as common across bootstrapped samples, we gain insight into whether these variables generally have a positive or negative effect on the sale price.

In addition to calculating the correlation coefficients, it is essential to determine whether these correlations are statistically significant. To achieve this, we conduct t-tests for each correlation coefficient. The t-test evaluates the null hypothesis that the correlation coefficient is zero (i.e., there is no significant linear relationship between the feature and the dependent variable). For each feature, we calculate the t-statistic using the correlation coefficient and the sample size, and we obtain the corresponding p-value. In this study, a p-value less than 0.05 indicates that we can reject the null hypothesis, suggesting that the observed correlation is statistically significant. This step is crucial because it ensures that the relationships we observe are not due to random chance but reflect genuine associations present in the data.

Results

Differences in Algorithm Performance

Table 1: RMSE ($) and 95% Confidence Intervals (Full Data Partition)

Algorithm	Full	95% CI Lower	95% CI Upper
SelectKBest	69188.66	62578.46	86038.75
RFE	104999.16	75148.29	135119.18
VarianceThreshold	81651.01	71343.49	98444.60
SelectFromModel_Lasso	115979.56	104204.91	139147.82
SelectFromModel_Tree	87719.76	77875.43	111380.03
Control Lasso	143403.25	129558.52	171619.61

Table 2: RMSE ($) and 95% Confidence Intervals (Subset Data Partition)

Algorithm	Subset	95% CI Lower	95% CI Upper
SelectKBest	77790.04	64781.70	98692.34
RFE	201238.44	82041.78	273614.68
VarianceThreshold	99298.79	73082.34	142694.55
SelectFromModel_Lasso	117431.58	85915.87	142282.83
SelectFromModel_Tree	81580.90	53069.82	115741.67
Control Lasso	—	—	—

The results presented in Table 1 demonstrate the performance of feature selection algorithms when granted access to the complete dataset, otherwise known as the traditional scenario (where $n > m$). *SelectKBest* yielded the lowest RMSE ($69,188.66), with a CI range of [62,578.46, 86,038.75], indicating its relatively high predictive accuracy and stability. *VarianceThreshold* achieved an RMSE of $81,651.01 with a CI range of [71,343.49, 98,444.60]. *RFE* followed, producing an RMSE of $104,999.16 with a wider CI range of [75,148.29, 135,119.18], suggesting more variability and lower accuracy in its prediction. The *SelectFromModel* algorithms, utilizing Lasso and Decision Trees, had mixed performances. *SelectFromModel_Lasso* exhibited the highest RMSE of the selection algorithms ($115,979.56) with a wide CI [104,204.91, 139,147.82], indicating lower predictive accuracy and high variability. On the other hand, *SelectFromModel_Tree* showed moderate performance with an RMSE of $87,719.76 and a CI of [77,875.43, 111,380.03], suggesting a balance between accuracy and stability. The benchmark Lasso recorded the highest RMSE ($143,403.25), with the widest CI range [129,558.52, 171,619.61], confirming the need for feature selection algorithms for improving model performance.

Table 2 illustrates the RMSE results under the subset data partition where the data demonstrates a highly-dimensional structure (where $m > n$). Recall that the high-dimensional configuration is applied solely during the feature selection phase and not during the model training process. This approach allows us to isolate and evaluate the performance of the feature selection algorithms without affecting the modeling results. By doing so, we ensure that any observed effects are directly related to the algorithms' ability to manage complex, high-dimensional scenarios. Notably, *SelectKBest* maintained its robust performance, yielding an RMSE of $77,790.04 with a CI range of [64,781.70, 98,692.34], slightly higher and wider than the full data partition (where $n > m$) but still relatively stable. *VarianceThreshold*'s performance was similar in both partitions, recording an RMSE of $99,298.79 with a CI of [73,082.34, 142,694.55], demonstrating consistency but also indicating moderate predictive power. The *RFE* algorithm's RMSE increased significantly to $201,238.44, and its CI range widened to [82,041.78, 273,614.68], highlighting its reduced effectiveness in the high-dimensional scenario. *SelectFromModel_Lasso* continued to exhibit high RMSE values ($117,431.58) and variability, as seen in its CI of [85,915.87, 142,282.83]. *SelectFromModel_Tree* had an RMSE of $81,580.90 with a CI of [53,069.82, 115,741.67], showing improved performance compared to other models, but still with wider variability compared to the full data partition. The RMSE for the benchmark Lasso in the subset partition was not reported, as it is methodologically infeasible to isolate the feature selection properties of Lasso when trained on a separate data partition. Specifically, feature importance cannot be derived from the subset data and subsequently used to train and evaluate the model on the full dataset, as was done with the other feature selection algorithms.

Consistency of Feature Selection Across Scenarios

Table 3: Percentage of Common Features Across Data Partitions

Algorithm	Percentage (%)	95% CI Lower (%)	95% CI Upper (%)
SelectKBest	86.1	76.0	92.0
RFE	89.4	72.0	96.0
VarianceThreshold	96.9	92.0	100.0
SelectFromModel_Lasso	86.4	70.0	100.0
SelectFromModel_Tree	53.3	40.0	65.0

Table 3 presents the percentage of features selected by each algorithm that were consistent between the traditional scenario (where $n > m$) and high-dimensional scenario (where $m > n$). *VarianceThreshold* demonstrated the highest consistency, with 96.9% of features overlapping between partitions and a 95% CI range of [92.0%, 100.0%]. This suggests that the algorithm is robust in selecting stable features regardless of data dimensionality. This is intuitive, as the variance of independent variables will tend to stay consistent across randomly picked subsets. *RFE* also showed high consistency, with 89.4% of selected features common across partitions and a CI of [72.0%, 96.0%], indicating its reliability in feature selection despite its higher RMSE in the high-dimensional scenario. *SelectKBest* displayed a similar pattern, with 86.1% consistency and a CI of [76.0%, 92.0%], further corroborating its stability. In contrast, the *SelectFromModel* algorithms exhibited mixed results. *SelectFromModel_Lasso* achieved 86.4% consistency with a CI of [70.0%, 100.0%], while *SelectFromModel_Tree* had the lowest consistency, at 53.3% with a CI range of [40.0%, 65.0%]. These findings indicate that while *SelectFromModel_Lasso* maintains some level of stability, *SelectFromModel_Tree*'s feature selection is more sensitive to data partition changes, reflecting its variability in RMSE performance.

Common Feature Analysis

The common feature analysis, which assessed the overlap of features selected by different algorithms, demonstrated that *SelectKBest*, *RFE*, *VarianceThreshold*, and *SelectFromModel_Lasso* consistently identified a core set of features that correlated strongly with house prices. These core features included *GrLivArea, GarageArea, 1stFlrSF, TotalBsmtSF, Age* and *YearsSinceRemodeled*. The high level of agreement among these algorithms suggests that these features are robust and likely have a genuine impact on house prices. This reinforces the reliability of the features identified and provides a basis for developing predictive models that are not only accurate but also interpretable. The fact that these features remained consistent across algorithms indicates that they are essential drivers of property value, regardless of the specific method used.

Features Identified Across All Algorithms

Table 4: Correlation Coefficients and Significance Levels

Variable Name	**Coefficient**	**p-value**
GrLivArea	0.70	0.00
GarageArea	0.65	0.00
1stFlrSF	0.60	0.00
TotalBsmtSF	0.61	0.00
Age	-0.59	0.00
YearsSinceRemodeled	-0.57	0.00

Table 4 presents the correlation coefficients and significance levels for a selection of variables identified through the feature selection process. The table indicates significant positive and negative correlations between the variables and the target outcome. For instance, *GrLivArea*(0.70), *GarageArea* (0.65), and *TotalBsmtSF* (0.61), all of which have p-values of 0.00, indicating statistical significance at the 99% confidence level and Pearson's correlation coefficients over 0, indicating positive relationships with our target variable, Sale Price. *Age* (-0.59) and *YearsSinceRemodeled* (-0.57) show significant negative correlations with the target variable. These negative relationships suggest that as the age of the property increases or as more time passes since its last renovation, the target value decreases. This outcome aligns with the intuition that older properties or those not recently renovated may be valued lower or have diminished appeal compared to newer or recently renovated properties.

In addition to employing a combination of feature selection algorithms, we used Lasso regression as a robustness check to verify the consistency and validity of our selected features. Our combined feature selection approach identified six key predictors; all of these were also selected by the Lasso model. This high degree of overlap supports the reliability of our feature selection process, indicating that our algorithms are largely aligned with the Lasso regression results, thus reinforcing the robustness of our selected variables.

Table 5 shows the coefficients derived from the Lasso regression model, where the alpha parameter was optimized to include only the 15 most significant non-zero coefficients. The results indicate that *GrLivArea* and *TotalBsmtSF* exhibit substantial positive coefficients (0.0888 and 0.0267, respectively), aligning with their positive correlation values in the earlier analysis. Interestingly, Lasso identifies additional significant variables that were not as prominent in the correlation analysis. For example, *OverallQual* (0.1404), *GarageCars* (0.0504), *Fireplaces* (0.0111) and *BsmtFinSF1* (0.0063) are included in the model, suggesting that their influence may not have been fully captured through the combination of selection algorithms. Furthermore, categorical variables such as *MSZoning_RL* (0.0030) and *CentralAir_Y* (0.0000) are also retained, demonstrating Lasso's ability to capture both numerical and categorical relationships within the dataset. Negative coefficients for *Age* (-0.0234) and *YearsSinceRemodeled* (-0.0257) again emerge, consistent with the findings from the correlation analysis with the external feature selection components.

Table 5: Lasso Regression Coefficients

Feature	**Coefficient**
OverallQual	0.1404
GrLivArea	0.0888
GarageCars	0.0504
TotalBsmtSF	0.0267
Fireplaces	0.0111
BsmtFinSF1	0.0063
1stFlrSF	0.0034
MSZoning_RL	0.0030
GarageArea	0.0029
CentralAir_Y	0.0000
FireplaceQu_None	-0.0039
CentralAir_N	-0.0062
MSZoning_RM	-0.0115
Age	-0.0234
YearsSinceRemodeled	-0.0257

An interesting observation comes to light when comparing the specific features selected by Lasso with those from our combined feature selection algorithms. The Lasso model includes variables that are very similar in nature, such as *GarageCars* and *GarageArea* as well as *BsmtFinSF1* and *TotalBsmtSF*. In contrast, our combined feature selection algorithms chose only one variable from each of these pairs. This discrepancy suggests that Lasso might be capturing multicollinear relationships that our combined feature selection approach aims to minimize. Multicollinearity, the high correlation between predictor variables, can have a significant impact on feature selection. When two or more variables are highly correlated, they provide redundant information, and some algorithms, particularly those focusing on dimensionality reduction (such as principal component analysis or recursive feature elimination), tend to exclude these redundant variables to reduce multicollinearity. In our combined feature selection approach, this tendency results in the selection of variables that are distinct and measure different aspects of the target outcome. For instance, *GarageCars* and *GarageArea* are closely related metrics; both indicate the garage size or capacity, which logically influence the outcome in similar ways. Our combined feature selection algorithms select *GarageArea*, likely due to its more direct and comprehensive measurement of the garage space, while excluding *GarageCars* to avoid redundancy. Similarly, *BsmtFinSF1* and *TotalBsmtSF* capture similar dimensions of basement space, but our approach chose *TotalBsmtSF* as the more inclusive and therefore possibly more informative measure. Lasso regression operates differently than feature selection algorithms (unless they are powered by a Lasso component). While it penalizes the coefficients of correlated variables, it does not necessarily exclude them outright if they contribute significantly to the predictive power of the model. As a result, Lasso may retain multiple variables that are highly correlated as both offer predictive information that strengthens the overall fit of the model. This outcome explains why Lasso captures pairs of similar variables, whereas our combined algorithms select one representative from each pair to maintain a parsimonious and non-redundant model. The Lasso model's inclusion of correlated features may suggest that, when used as a robustness check, it serves to confirm that at least one variable from a correlated pair is important. For instance, if Lasso retains both *GarageCars* and *GarageArea*, it reinforces the validity of selecting one of these features in our combined approach, confirming that these types of variables are indeed influential.

Discussions and Limitations

The findings of this study offer insights into the performance and reliability of various feature selection algorithms in the context of real estate price prediction, particularly when confronted with different data structures. Our experimental design, which deliberately created scenarios of both traditional (where $n > m$) and high-dimensional (where $m > n$) data partitions, has yielded results that not only show the strengths and weaknesses of each algorithm but also provide broader implications for the field of predictive modeling and feature selection.

An analysis of Root Mean Square Error (RMSE) values across these scenarios indicates that applying subset constraints generally leads to elevated RMSE values, suggesting a decline in performance for most algorithms in highly dimensional data. For example, the RMSE for the *VarianceThreshold* algorithm increased from \$81,651.01 in the full data scenario to \$99,298.79 in the subset scenario, with confidence intervals showing similar trends of increasing variability across all algorithms. This variance supports the notion that constraining feature selection adds uncertainty to the selection process, affecting the model's overall performance. Even in the cases where an algorithm improved in average RMSE performance in the high-dimensional scenario (For example, *SelectFromModel* with Decision Trees), this still comes with drastic increases in the width of confidence intervals, indicating less stability in predictive power.

A key observation is the consistent performance of the *SelectKBest* algorithm across both data partitions. Achieving the lowest RMSE of $69,188.66 in the traditional scenario and $77,790.04 in the high-dimensional scenario, *SelectKBest* demonstrated stability and predictive accuracy. This performance suggests that the mutual information criterion used by *SelectKBest* effectively captures complex, potentially non-linear relationships in real estate data. Its robust evaluation across various data structures highlights its potential as a reliable tool for feature selection in diverse real-world applications where the relationships between features and dependent variables are not immediately apparent. In contrast, other algorithms, particularly *Recursive Feature Elimination (RFE)* and *SelectFromModel* variants, exhibited greater variability between the two scenarios. Notably, the RMSE for *RFE* rose sharply from $104,999.16 in the traditional scenario to $201,238.44 in the high-dimensional scenario. This substantial increase suggests that recursive methods may face challenges with high-dimensional data, where the risk of overfitting is significant. Such results underscore the importance of considering the dimensionality of the data when choosing feature selection methods, particularly in domains like genomics or high-resolution spatial analysis.

The Benchmark Lasso model is employed as a benchmark for comparison, as it does not implement any external feature selection process and utilizes the entire dataset as presented. In the full data partition scenario, the benchmark Lasso model records the highest RMSE of $143,459.03 among all methodologies, suggesting that the incorporation of feature selection is indeed beneficial for enhancing model performance.

The consistent feature analysis provided further insights into the stability of these algorithms by determining the percentage of variables in common per selection algorithm across the two scenarios (high-dimensional and traditional). *VarianceThreshold*, with a 96.9% overlap of selected features between the two scenarios, emerged as the most consistent algorithm. The high feature similarity of *VarianceThreshold* is largely attributable to its selection criterion being based solely on the variance of individual features, independent of their relationship with the target variable. This method's stability across different data partitions is logical: features with high variance in the full dataset are likely to maintain high variance in random subsets, leading to consistent selection. However, while this method ensures stable feature selection, it doesn't necessarily select the most predictive features. This is reflected in its RMSE values ($81,651.01 in the traditional scenario and $99,298.79 in the high-dimensional scenario), which, while relatively stable, are not the lowest among the algorithms tested. Stability in feature selection does not always translate to optimal predictive power.

SelectKBest was a balanced performer, maintaining both high feature consistency (86.1% overlap) and low RMSE values ($69,188.66 in the traditional scenario and $77,790.04 in the high-dimensional scenario). This further emphasizes the effectiveness of mutual information criteria in identifying features that are both consistently important and predictive across varying data structures in the real estate domain. The limited increase in RMSE when moving to the high-dimensional scenario indicates that the features selected by *SelectKBest* are not only informative in the original, lower-dimensional dataset but also retain their predictive relevance when the number of observations is reduced. This consistency suggests that Mutual Information Criteria is effective at prioritizing features that are genuinely influential rather than those that may appear important due to noise or spurious correlations. This property is especially important in real estate, where a variety of factors—ranging from the physical attributes of properties to locational and socioeconomic variables—interact in complex ways to determine housing prices.

In examining the performance of *Recursive Feature Elimination (RFE)* across traditional and high-dimensional scenarios, we encountered an intriguing and seemingly contradictory phenomenon. While *RFE* showed a substantial increase in RMSE when moving from the traditional to the high-dimensional scenario (from $104,999.16 to $201,238.44), it simultaneously demonstrated a high consistency rate, with an average of 89.4% of the same features selected across the two scenarios. Confidence intervals for this consistency ranged from 72% to 96%, suggesting that the algorithm maintained a relatively stable feature selection pattern. However, the notable increase in RMSE and the widened confidence intervals when applied to the high-dimensional scenario suggest that this consistency might not necessarily equate to predictive effectiveness. The core of this contradiction may lie in the tendency of *RFE* to overfit when operating in high-dimensional environments. High-dimensional settings introduce a larger number of potential feature interactions, leading to an increased risk of the algorithm selecting features that appear statistically significant within the specific context of the training subset but may not generalize well to the full dataset. *RFE*, by its iterative nature, may capture patterns specific to the high-dimensional partition but exclude critical features that are essential for the model's performance when all data is incorporated, as displayed in the subset RMSE values. High-dimensional data inherently carries more noise, and if *RFE* fails to select key predictive features while retaining others that may appear relevant only within the high-dimensional context, the RMSE increase is expected. The consistency rate of 89% between traditional and high-dimensional feature sets demonstrates that *RFE* can indeed identify many overlapping features; however, the discrepancy is significant enough that just 3 or 4 variables—if they are key predictive features—can markedly impact model performance. This phenomenon underscores the importance of key variables in complex datasets such as real estate, where only a small set of features may capture a substantial amount of the predictive information. In such cases, omitting these critical features or replacing them with less predictive alternatives could lead to dramatic changes in RMSE, even when the overall overlap of features between scenarios remains high. In our results overview, we noted that our Benchmark Lasso model tends to include variables that are similar in nature, such as *GarageCars* and *GarageArea*, or *BsmtFinSF1* and *TotalBsmtSF*. These variable pairs represent different dimensions of the same underlying feature—garage capacity and basement area, respectively. Now, let us consider the impact of reducing the number of observations, thereby simulating a high-dimensional scenario as done in the current study. In such scenarios, the model's capacity to distinguish between subtle differences in the predictive power of similar variables becomes compromised. Consequently, the alternative, less pronounced measure of these variable pairs—such as *GarageCars* instead of *GarageArea* or *BsmtFinSF1* instead of *TotalBsmtSF*—may appear more influential to the selection algorithm under these restricted conditions. This shift occurs because the reduced sample size increases the variability and instability of feature importance rankings, making it more likely for the algorithm to select features that seem important only within the context of the limited, high-dimensional subset of data. If this phenomenon repeats across several variable pairs, the cumulative effect is a selection of features that do not optimally represent the most predictive dimensions of the dataset when evaluated in a broader context. This misalignment between selected features and actual predictive relevance is a plausible explanation for the observed increase in RMSE and the expansion of confidence intervals when the model is applied to the high-dimensional scenario. By selecting less robust and potentially noisier variables due to the constraints of the reduced sample size, the model becomes more prone to overfitting, capturing spurious relationships that do not generalize well to the full dataset.

The 86.4% consistency exhibited by SelectFromModel_Lasso suggests that linear models like Lasso can maintain a stable feature selection process, even when faced with dimensional changes in the data. This level of consistency, similar to that achieved by SelectKBest, indicates that Lasso's regularization properties effectively penalize and shrink coefficients, leading to a focused and consistent selection of features. However, despite this stability, the high RMSE values and wide confidence intervals (CIs) for SelectFromModel_Lasso in both traditional and high-dimensional scenarios reveal a critical limitation: while the model maintains consistent feature selection, it may not be adequately capturing the underlying complexities of the data. High RMSE values ($115,979.56 in the traditional scenario and $117,431.58 in the high-dimensional scenario) and wide CIs ([104,204.91, 139,147.82] and [85,915.87, 142,282.83], respectively) for this model suggest that when the real estate data includes non-linear relationships, Lasso’s linear approach struggles to fit the model accurately. This limitation may stem from the fact that Lasso operates under the assumption that the relationships between predictors and the target are predominantly linear. In a domain like real estate, where interactions between variables such as property size, neighborhood characteristics, and market dynamics are likely to be non-linear, relying solely on Lasso’s linear framework can lead to oversimplified models that fail to capture critical patterns. Consequently, even if the features selected by Lasso remain consistent, their predictive power may be compromised when the model cannot fully adapt to the underlying data structure (linear vs non-linear variable interactions).

On the other hand, SelectFromModel_Tree’s performance, with a significantly lower consistency rate of 53.3%, underscores the variability that can arise when using non-linear base models such as decision trees. Decision trees, being more flexible than linear models, can capture a wide range of complex interactions and non-linearities. However, this flexibility comes at the cost of stability. The lower overlap in feature selection between the traditional and high-dimensional scenarios suggests that decision trees may be more sensitive to changes in data structure and sample size, potentially leading to different sets of features being chosen when the number of observations varies. This sensitivity is likely because decision trees can split on different features depending on the structure and partitioning of the data, which can lead to variability in the selected feature set. Despite the variability, SelectFromModel_Tree exhibited more favorable RMSE values than Lasso in both scenarios. In the traditional scenario, its RMSE of $87,719.76 and CI [77,875.43, 111,380.03] indicate that while it is less consistent in feature selection, it achieves a balance between accuracy and variability that Lasso struggles to attain. In the high-dimensional scenario, SelectFromModel_Tree’s performance showed further improvement, with an RMSE of $81,580.90 and a CI of [53,069.82, 115,741.67]. The lower RMSE suggests that the non-linear nature of decision trees enables the model to adapt to the complex patterns within the real estate data more effectively, even as the number of observations decreases. However, the wider variability compared to the full data partition indicates that while decision trees are powerful in capturing intricate relationships, they require careful tuning and validation to ensure that the model remains reliable and not overly sensitive to data fluctuations (Bertsimas & Dunn, 2017; García Leiva et al., 2019).

These findings collectively indicate that the ideal feature selection method should balance consistency with predictive power, and that this balance may shift depending on the specific characteristics of the dataset and the analytical goals. In the context of real estate price prediction, algorithms like *SelectKBest*, which leverage mutual information, appear to offer a robust approach to feature selection across different data structures. However, the significant variations in performance across algorithms and scenarios highlight the need for careful consideration of the specific context and objectives when choosing a feature selection method.

Our common feature analysis revealed a core set of features consistently identified across multiple algorithms, including *TotalBsmtSF*, *GarageArea*, *GrLivArea*, and *OverallQual*. The recurrence of these features across different selection methods not only reinforces their importance in predicting house prices but also provides valuable insights into the key drivers of real estate valuation. This finding has significant implications for both theoretical understanding of housing markets and practical applications in real estate assessment and investment strategies.

The Benchmark Lasso model, previously used to compare predictive performance, was also employed to serve as a benchmark for evaluating the effectiveness of the feature selection algorithms. Notably, the Benchmark Lasso model retained all variables that were also selected by our feature selection algorithms, reinforcing the alignment between traditional regularization methods and our approach. Furthermore, the selected variables exhibited similar relationships with the target variable across both the Lasso and feature selection models. This consistency suggests that variable selection algorithms, particularly those that incorporate a regularization mechanism like Lasso, are adept at pinpointing important predictors in a way that aligns with traditional econometric models such as hedonic regressions.

Variable selection algorithms offer an additional advantage by often yielding more streamlined and interpretable models. One of the key benefits observed in this study is that the feature selection algorithms tend to prioritize parsimony by selecting one representative feature from distinct dimensions or clusters of highly correlated variables. For instance, *GarageArea* was selected over *GarageCars* in the combined feature selection algorithms. Lasso's retention of both these multicollinear variables (among others) highlights a potential trade-off: while including correlated variables may improve predictive performance in the short term, it might also lead to challenges in interpreting the model, as the influence of each variable could overlap. This indicates that feature selection algorithms, through their inherent dimensionality reduction processes, provide models that are not only more interpretable but also potentially less prone to multicollinearity, which can be a significant issue in traditional hedonic models.This behavior can be attributed to the nature of how the algorithms process the data. Feature selection methods such as *SelectFromModel* with Lasso and *Recursive Feature Elimination (RFE)*, which rely on model-driven importance ranking, tend to select a single feature to represent a distinct aspect of the dataset. In the case of our analysis, variables like *GrLivArea*, *1stFlrSF*, and *TotalBsmtSF* were chosen as they capture different dimensions of property characteristics, such as living spaces, and basement size, respectively. Each of these variables provides unique information that contributes to a multidimensional understanding of property value. The selection algorithms naturally avoid redundancy by discarding highly correlated features that do not add significant incremental value. In contrast, the Lasso model, while effective in reducing the number of predictors through penalization, tends to retain multiple correlated features when they offer incremental improvements in model fit, even if they represent the same underlying dimension (e.g., *GarageArea* and *GarageCars*).

The implications of this study are twofold. First, the consistent selection of core features across multiple algorithms underscores their importance in predicting outcomes, lending support to the idea that certain variables serve as fundamental drivers of variation in real-world datasets. Second, the ability of feature selection algorithms to eliminate redundant variables without compromising predictive performance offers a streamlined approach to modeling that balances efficiency with accuracy. This can be particularly useful in fields like real estate valuation, where interpretability is crucial for decision-making. By selecting only one variable from highly correlated pairs, feature selection algorithms provide clearer insights into the relationships between predictors and outcomes, making the models easier to interpret and more transparent for practitioners. Moreover, our findings suggest that these methods are particularly advantageous when combined with correlation analysis, which can provide insights into the relationship between selected variables and the dependent variable. By systematically choosing representative features from each group of multicollinear covariates, the feature selection process not only reduces dimensionality but also ensures that the retained variables measure distinct and complementary aspects of the phenomenon under study. This process helps minimize the potential distortions caused by multicollinearity, which can complicate the interpretation of traditional hedonic models.

Despite the valuable insights gained from this study on feature selection algorithms, several limitations should be acknowledged. While our primary objective was to evaluate the effectiveness of various feature selection techniques in reducing predictive error, our study did not encompass market dynamics or economic factors influencing pricing. This focus may limit the applicability of our findings in scenarios where those contextual factors play a crucial role. Additionally, the dataset utilized in our analysis is sourced from a Kaggle competition, which may not be representative of all housing markets. This dataset is constrained by specific regional characteristics that may not include those relative to other areas. Although we performed the necessary pre-processing steps in the cleaning of our data, the quality could still impact the precision of our models.

Although we were constrained by time and resources, we did not explore a comparative performance of feature selection algorithms that utilized log-transformed Sale Price and raw Sale Price. Future research could benefit from investigating these transformations and whether different features are selected with these algorithms. This can provide a better understanding of the implications on predictive accuracy and interpretability across algorithms. In addition, our study only focused on select algorithms due to limited research, potentially overlooking other methods that could yield different insights. In addition to exploring other methods, we did not choose to explore other performance metrics besides RMSE. Although RMSE is a standardized benchmark, this metric may not fully capture all dimensions of model performance in terms of interpretability and computational efficiency. These limitations underscore the need for future research to explore these results more comprehensively.

Conclusions

This study explored the efficiency of various feature selection algorithms in improving predictive accuracy and variable importance for housing price prediction models. By comparing these algorithms to our benchmark Lasso model, we were able to assess their performance in both high-dimensional ($m > n$) and traditional (n

> *m*) data environments. Our results showed that feature selection algorithms, particularly *SelectKBest* and *VarianceThreshold*, had notably reduced predictive error when compared to our benchmark Lasso model. These findings suggest that a more systematic approach to feature selection can improve model performance and common machine learning techniques in this domain.

The robustness of the feature selection algorithms was further demonstrated through their ability to consistently identify key housing attributes, such as *GrLivArea*, *TotalBsmtSF*, and *GarageArea*. Our benchmark Lasso model also identified these attributes with the coefficients being non-zero after regularization. This consistent selection across various methods not only highlights the importance of these attributes but also confirms that feature selection algorithms are capable of uncovering reliable predictors even in high-dimensional scenarios. Although some algorithms, like RFE, showed variability in performance across different data structures, overall, feature selection techniques offered a balanced approach by maintaining accuracy while enhancing model interpretability.

While this study focused primarily on advancements in feature selection for predictive modeling, it also underscores the broader potential of feature selection algorithms in fields beyond real estate. The ability to systematically reduce dimensionality while preserving predictive features offers significant advantages for researchers and data scientists. Future research can extend these findings by exploring additional algorithms, incorporating economic or contextual variables, and applying these methods to different datasets or industries. Ultimately, our findings contribute to a deeper understanding of predictive modeling and provide insights for improving real estate price prediction models and beyond.

References

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082. https://doi.org/10.1007/s10994-017-5633-9

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, *5*(2), 65–75. https://doi.org/10.1007/s13748-015-0080-y

Chanasit, K., Chuangsuwanich, E., Suchato, A., & Punyabukkana, P. (2021). A Real Estate Valuation Model Using Boosted Feature Selection. *IEEE Access*, *9*, 86938–86953. https://doi.org/10.1109/access.2021.3089198

Colwell, P. F., & Dilmore, G. (1999). Who Was First? An Examination of an Early Hedonic Study. *Land Economics*, *75*(4), 620. https://doi.org/10.2307/3147070

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*, *19*(S1). https://doi.org/10.1186/s12863-018-0633-8

Feature Extraction. (2006). In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35488-8

Fonti, V., & Belitser, E. (2017). *Feature Selection using LASSO*. https://vu-business-analytics.github.io/internship-office/papers/paper-fonti.pdf

García Leiva, R., Fernández Anta, A., Mancuso, V., & Casari, P. (2019). A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design. *IEEE Access*, *7*, 99978–99987. https://doi.org/10.1109/ACCESS.2019.2930235

García-Magariño, I., Medrano, C., & Delgado, J. (2019). Estimation of missing prices in real-estate market agent-based simulations with machine learning and dimensionality reduction methods. *Neural Computing and Applications*, *32*(7), 2665–2682. https://doi.org/10.1007/s00521-018-3938-7

Goodman, A. C. (1998). Andrew Court and the Invention of Hedonic Price Analysis. *Journal of Urban Economics*, *44*(2), 291–298. https://doi.org/10.1006/juec.1997.2071

Kalina, J., & Schlenker, A. (2015). A Robust Supervised Variable Selection for Noisy High-Dimensional Data. *BioMed Research International*, *2015*, 1–10. https://doi.org/10.1155/2015/320385

Kim, H., Park, S. W., Lee, S., & Xue, X. (2015). Determinants of house prices in Seoul: A quantile regression approach. *Pacific Rim Property Research Journal*, *21*(2), 91–113. https://doi.org/10.1080/14445921.2015.1058031

Kim, T., Shin, J., Kim, H., & Heo, J. (2020). Ensemble‐Based Neural Network Modeling for Hydrologic Forecasts: Addressing Uncertainty in the Model Structure and Input Variable Selection. *Water Resources Research*, *56*(6). https://doi.org/10.1029/2019wr026262

Lahmiri, S., Bekiros, S., & Avdoulas, C. (2023). A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization. *Decision Analytics Journal*, *6*, 100166. https://doi.org/10.1016/j.dajour.2023.100166

Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. *Information*, *15*(6), 295. https://doi.org/10.3390/info15060295

Mohd, T., Nur Izzah Jamil, Noraini Johari, Abdullah, L., & Suraya Masrom. (2020). An Overview of Real Estate Modelling Techniques for House Price Prediction. *Springer*, 321–338. https://doi.org/10.1007/978-981-15-3859-9_28

Mullainathan, S., & Spiess, J. (2017). Machine Learning: an Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Orphe Divounguy. (2023, September 26). *The Value of Residential Real Estate Broke a New Record $52 Trillion*. Zillow. https://www.zillow.com/research/total-market-value-2023-33031/#/

Patil, A., Bharate, P., & Junaid, M. (2023). Enhancing Myocardial Infarction Diagnosis with Efficient Machine Learning Techniques Through Combination of Correlation and Variance Threshold Feature Selection. *International Journal of Life Sciences*, *12*(4). https://www.ijlbpr.com/uploadfiles/31vol12issue4pp172-180.20231019100711.pdf

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, *2*(927312). https://doi.org/10.3389/fbinf.2022.927312

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. https://www.jstor.org/stable/1830899

Singh, D. A. A. G., Appavu, S., & Leavline, E. J. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. *International Journal of Computer Applications*, *136*(1), 9–17. https://doi.org/10.5120/ijca2016908317

Tran, L. M., Rizk, M. L., & Liao, J. C. (2008). Ensemble Modeling of Metabolic Networks. *Biophysical Journal*, *95*(12), 5606–5617. https://doi.org/10.1529/biophysj.108.135442

Triplett, J. (2007). *Title: Zvi Griliches' Contributions to Economic Measurement*. https://www.nber.org/system/files/chapters/c0890/c0890.pdf

Vishwakarma, S., & Singhal, S. (2020). House Price Forecasting Based on Hybrid Multi-regression Model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3601507