Deciphering House Price Dynamics Using Predictive Modeling on Housing Data

Allen Zou, Ting-An Lu, Zachary Chao, and Casey Hild

May 12, 2024

Abstract

Understanding the multifaceted dynamics of house price prediction in real estate markets is important for stakeholders involved in decision-making. This research focuses on unraveling these complexities through predictive modeling using the Ames Housing dataset, aiming to identify the factors influencing house prices and develop a robust predictive model to estimate them accurately. Employing a systematic methodology, including decision tree modeling, forward feature selection, k-fold cross-validation, and grid search hyperparameter tuning, the research investigates the intricate interplay of various property characteristics. The findings reveal predictors of house prices, such as the quality of rooms (e.g., basement, kitchen, etc.), number of amenities (bathrooms, utilities, etc.), and the area of certain rooms (basement, kitchen, etc.). While emphasizing insights for stakeholders, such as buyers, sellers, and investors, the study underscores the value of predictive modeling in strategic decision-making within the real estate domain.

1 Introduction

The real estate market is a complex ecosystem influenced by various factors, ranging from property characteristics to broader economic trends. Understanding the dynamics of house prices is essential for stakeholders involved in buying, selling, or investing in real estate. This research seeks to unravel the intricacies of house price dynamics through predictive modeling, offering insights that can inform strategic decision-making in the housing market.

In this study, we aim to address a research question: What factors influence house prices, and can we construct a predictive model to estimate house prices based on these factors? By delving into this question, we seek to unravel the complex interplay of factors shaping house price dynamics and explore the feasibility of developing a robust predictive model.

Based on the variables in the Ames Housing dataset, we hypothesize that certain factors such as location (i.e., the neighborhood variable, which describes the physical locations within Ames city limits), size of rooms (i.e., garage, basement, living area, etc. in square feet), and overall quality (i.e., categorical ratings for the aforementioned rooms such ”excellent,” ”good,” ”typical,” etc.), will have a significant impact on house prices. Specifically, they will have a positive relationship with house prices.

2 Related Work

Many previous papers have tackled the challenge of predicting housing prices in various locations. For the Ames Housing dataset, while some studies aim to generate the most accurate predictions using complex models such as neural networks, the primary focus has generally been on feature engineering. Clemmer performed extensive exploratory data analysis to identify the most important features for predicting housing prices [2]. Similarly, Nagarajan et al. tested various subsets of features to implement an ordinary least squares regression model [9]. These papers found that certain features, such as overall quality and total square footage, are some of the main factors in determining housing prices in Ames, Iowa. Another article from Truong et al. looks into housing price prediction using similar machine learning models as what we use and finds that factors such as locaiton, area, and population have the greatest effect on housing prices. [11]

3 Dataset

The Ames Housing dataset used in this study was compiled by Dean De Cock and serves as a modernized and expanded version of the often-cited Boston Housing dataset [6]. It consists of two comma-separated values (CSV) files: train.csv and test.csv. The train.csv file comprises a tabular format with 1,460 rows and 81 columns, while the test.csv file contains 1,461 rows and 81 columns. Each row in both datasets represents a distinct property sale transaction, serving as a unique sample in the dataset.

The first column of each dataset is labeled ”Id” and contains a unique identifier for each property sale. The target variable, ”SalePrice,” is included as the last column, representing the sale price of each property. The features of each property sale transaction are captured in the columns between the first and last columns of the dataset.

The dataset encompasses a wide array of features that describe various aspects of residential properties. These features include categorical variables such as MSSubClass, which identifies the type of dwelling involved in the sale, and MSZoning, which indicates the general zoning classification of the sale. Additionally, numerical variables such as LotFrontage (linear feet of street connected to property) and LotArea (lot size in square feet) provide quantitative measurements of property characteristics. Other essential features cover aspects such as property configuration (LotShape, LotConfig), neighborhood information (Neighborhood), and conditions related to the property (Condition1, Condition2). Variables like OverallQual and OverallCond rate the overall material/finish and condition of the house, respectively, while variables like YearBuilt and YearRemodAdd provide information about the original construction and remodel dates.

Providing detailed descriptions for each variable directly in the paper might lead to clutter. The Ames Housing dataset we utilized in this study includes a comprehensive data description file (data_description.txt), which contains detailed descriptions of all variables. Readers interested in understanding the specific definitions of features can refer to this file for more information. The dataset is publicly available on Kaggle, and interested readers can access it via the link in the bibliography section.

Variable	Description
MSSubClass	Type of dwelling
MSZoning	Zoning classification
LotFrontage	Street connection length
LotArea	Lot size (sqare feet)
Street	Road access type
Alley	Alley access type
LotShape	Property shape
LandContour	Property flatness
Utilities	Available utilities
LotConfig	Lot configuration

Table 1: Examples of variables and their descriptions from the Ames Housing Dataset. The table presents a subset of variables from the dataset, providing details on details such as the type of dwelling, zoning classification, lot size, alley access types, available utilities, and lot configuration.

3.1 Data Pre-processing

The dataset initially comprised separate files for training and testing, with an unusual split of 1460:1461. To rectify this, the data were combined and shuffled, followed by an 80:20 train-test split. Due to the presence of numerous missing entries in at least half the rows and columns, preprocessing was necessary to ensure compatibility with scikit-learn, which does not support NaN values [10].

To address missing values, we first pruned columns with a substantial proportion of NaN values (approximately 20% or more). This step removed entries such as LotFrontage, Alley, FireplaceQu, PoolQC, Fence, and MiscFeature, resulting in a total of 73 out of 79 remaining columns (excluding SalePrice). One reason we chose to prune columns first was that no other column had anything close to 20% of its entries with NaN values. However, we chose not to prune rows because over half the entries contained missing entries. Given that all entries had a unique ID and there are 2,335 entries, this would remove a substantial amount of our data.

For the remaining dataset, we used the IterativeImputer from scikit-learn to handle missing values in numerical columns. The IterativeImputer models each feature with missing values as a function of other features, then uses that estimate for imputation. This approach allows for more sophisticated and accurate filling of missing data compared to simpler methods like mean or median imputation. Categorical variables were imputed using the SimpleImputer with the strategy set to most_frequent, which replaces missing values with the most frequent value (mode) of each column.

Following imputation, the dataset was split back into training and testing sets. This preprocessing approach allowed us to retain 80% of the original dataset, providing a robust foundation for subsequent modeling tasks involving linear regression, decision trees, and support vector machines.

4 Methodology

In this section, we provide an overview of the methodologies used for model selection and evaluation, as well as the steps involved in developing reliable predictive models. More detail about how we applied these methodologies to select our model will be provided in Section 4 (Model Selection). It should be noted that all the strategies employed here involve the use of scikit-learn's packages [10]. We employ a systematic approach, beginning with an explanation of why we chose the decision tree model, followed by feature selection, k-fold cross-validation, and grid search hyperparameter tuning. Each subsection outlines a aspect of our methodology, covering forward feature selection, hyperparameter optimization, cross-validation, and model evaluation. However, before delving into these methodologies, we will discuss why decision trees were chosen as our primary modeling technique.

4.1 Decision Trees

Decision trees offer strengths that are beneficial to modeling our the Ames Housing dataset. One of their advantages is their capability to capture nonlinear relationships between predictors and the target variable, providing flexibility in modeling complex data [1]. Moreover, decision trees inherently rank predictors based on their importance in predicting the target variable, facilitating feature selection and offering insights into variable importance. Their high interpretability is another significant advantage, as decision trees provide intuitive visualizations of decision paths, making it easy to understand the model's decision-making process.

However, it's important to note that decision trees also come with notable weaknesses and how we address them. They are prone to overfitting, especially with deep trees that may capture noise in the data rather than true patterns. Additionally, decision trees can exhibit instability, as small variations in the data can lead to significantly different decision trees, impacting model robustness. Furthermore, decision trees produce piecewise constant predictions, which may not capture gradual changes in the target variable across predictor space. Despite these challenges, employing strategies such as grid search for hyperparameter tuning can effectively mitigate issues of overfitting and instability, ensuring the development of more reliable decision tree models.

4.2 Forward Selection Overview

Feature selection is aimed at identifying the subset of relevant features that contribute most to the model's predictive performance while reducing dimensionality and computational complexity [4]. In our methodology, we employed a forward feature selection approach to iteratively select features based on their impact on model performance.

The forward feature selection process begins with an empty set of selected features and iteratively adds one feature at a time, evaluating the model's performance at each step. At each iteration, the algorithm identifies the feature that results in the greatest improvement in model performance, as measured by a chosen evaluation metric. This metric, in our case, was the mean squared error (MSE), a common measure of regression model accuracy. For future reference, the units of MSE (and MAE) values are in dollars.

To evaluate the performance of each feature subset, we utilized k-fold cross-validation, a robust technique for estimating the model's performance on unseen data. In our implementation, we employed a 5-fold cross-validation strategy, partitioning the dataset into 5 equally sized folds, training the model on 4 folds, and evaluating it on the remaining fold [7]. Specifically, with 2,335 data entries in our dataset, this change results in each fold containing approximately 467 data entries for validation and 1,868 data entries for training. This process was repeated 5 times, with each fold serving as the validation set exactly once.

During each iteration of the forward feature selection process, we trained a decision tree regression model on the preprocessed training dataset using the selected features plus the feature under consideration. We then used grid search with cross-validation to tune the hyperparameters of the decision tree.

Normally, the feature selection process continued until adding additional features no longer resulted in a significant improvement in model performance or began to degrade performance, but for the purpose of visualization, we continued to run forward selection until it depletes all features.

Upon completion of the feature selection process, we stored the selected features, their corresponding mean squared errors, and the hyperparameters of the best-performing models each iteration for further analysis and model evaluation.

4.3 One-Hot Encoding

Before training the decision tree models, categorical variables were encoded using one-hot encoding [3]. This preprocessing step converts categorical variables into binary vectors, with each category represented as a binary feature. One-hot encoding is necessary for decision tree models because scikit-learn's decision trees don't natively accept strings from categorical variables. By converting categorical variables into a numerical format through one-hot encoding, decision trees can effectively process categorical features and capture their relationships with the target variable.

4.4 K-Fold Cross-Validation

K-fold cross-validation is a widely used technique to assess the performance and generalization ability of a predictive model. The main idea behind k-fold cross-validation is to partition the dataset into k subsets, or "folds," of approximately equal size [7]. The model is then trained k times, each time using k-1 folds as the training set and the remaining fold as the validation set. This process allows each data point to be used for validation exactly once. The performance metric, such as mean squared error (MSE) for regression tasks or accuracy for classification tasks, is computed for each fold, and the average performance across all folds is reported as the overall performance estimate of the model. K-fold cross-validation helps mitigate the risk of overfitting by providing a more reliable estimate of a model's performance on unseen data compared to a single train-test split. Common choices for the value of k include 5 or 10, although the choice may vary depending on the size and nature of the dataset.

Our decision to employ 5-fold cross-validation was driven by the characteristics of our preprocessed dataset, which comprises 2,335 rows and 73 columns. With a moderate-sized dataset and a relatively large number of features, 5-fold cross-validation strikes a balance between bias and variance in model evaluation. By partitioning the dataset into 5 equally sized folds, each fold contains approximately 467 samples, ensuring that the training sets remain sufficiently large for model fitting while still allowing for comprehensive evaluation across multiple iterations. This method further reduces the risk of overfitting by exposing the model to diverse subsets of the data.

4.5 Optimizing Hyperparameters with Grid Search

In each iteration of the feature selection process, hyperparameters are optimized using scikit-learn's grid search to fine-tune the decision tree model [10]. Grid search is a systematic method for tuning hyperparameters by exhaustively searching through a predefined grid of parameter values and selecting the combination that yields the best model performance [3]. For the decision tree model, hyperparameters such as maximum depth, minimum samples split, minimum samples leaf, maximum features, and minimum impurity decrease are considered.

The grid search algorithm evaluates the model's performance using k-fold cross-validation as outlined in the previous subsection. By averaging the performance across all folds, grid search provides an estimate of the model's performance under various hyperparameter configurations. This process not only identifies the optimal hyperparameter values but also mitigates the risk of overfitting by ensuring that the chosen model parameters generalize well across different subsets of the data. The hyperparameter values that yield the lowest mean squared error (MSE) are selected as the optimal configuration for the decision tree model, ensuring that the model is both well-tuned and capable of generalizing to new data.

5 Model Selection

In the Model Selection section, we explore the application of methodologies described in the Methodology section to identify the optimal predictive model for our housing price prediction task. First, we detail the process of constructing a parameter grid tailored to our decision tree model. Subsequently, we elaborate on how we leveraged the outcomes derived from grid search cross-validation to select hyperparameters and a subset of features.

5.1 Grid Search Parameter Range

The param_grid dictionary for scikit-learn's grid search defines the range of values to explore during grid search. For each hyperparameter, such as max_depth, min_samples_split, min_samples_leaf, and min_impurity_decrease, we specified a list of potential values to test.

In particular, we defined a range of [0, 5, ..., 25, 30] for max_depth, [2, 3, ..., 6, 7] for min_samples_split, [2, 3, ..., 6, 7] for min_samples_leaf, and [0.0, 0.1, ..., 0.6, 0.7] for min_impurity_decrease, as these parameter values cover the range of values that GridSearchCV tended to search around. We did not alter the options for max_features—['auto', 'sqrt', 'log2']—as these represent constant strings.

5.2 Feature Subset and Hyperparameter Selection

This section illustrates how we visualize the model performance throughout the forward feature selection process using 5-fold cross-validation in each iteration of grid search and choose a model.

As described previously, the algorithm pinpoints the feature that yields the most substantial enhancement in model performance every iteration, gauged by MSE. Subsequently, based on the feature subset identified with the best-performing feature, we accumulate all the MSE values from the grid search on that particular model using the optimal feature subset. These aggregated MSE values form the basis for constructing a box and whisker plot, with each plot representing the MSE values from grid search on the best-performing feature subset for that iteratoin.

In Figure 1 (page 5), we present the box and whisker plots illustrating the MSE values of the model with the best-performing feature subset for all iterations of forward selection—comprising a total of 73 iterations, corresponding to the number of features in the preprocessed dataset. Notably, the initial iteration exhibits a markedly elevated box and whisker plot, indicating minimal improvement when incorporating a single feature. While the whiskers of nearly all other plots align approximately at the same lower bound, a low point emerges in the mean of the box and whisker plots at around iteration 16, representing a subset of 16 features.

Given the challenge of interpreting the box and whisker plots in Figure 1, we provide a more focused view in Figure 2, zooming into iterations 12 to 20. The box and whisker plot corresponding to iteration 16 seems to have a mean that's slightly lower than the other plots in the figure, with the upper end of its whisker aligning closely with adjacent plots. Another observation is that the lower end of the whiskers for the plot representing iteration 16 is one of the lower whiskers in the figure. All things considered, we opt for the model utilizing the subset of 16 features.

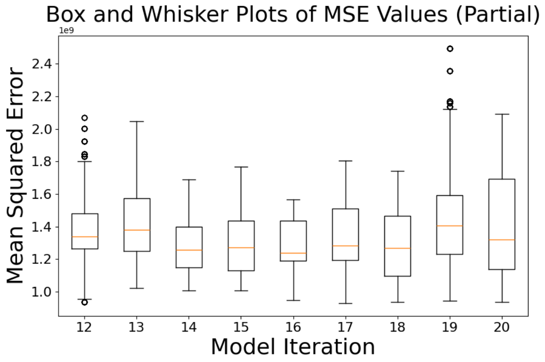

Figure 2: MSE values for iterations 12 to 20 of grid search during the forward feature selection process. This partial plot provides a focused view of model performance improvements during critical iterations. Iteration 16 shows a very slightly lower mean compared to adjacent iterations.

5.3 Final Hyperparameters and Features

Hyperparameter	Value
max_depth	10
max_features	'auto'
min_impurity_decrease	0.2
min_samples_leaf	3
min_samples_split	2

Table 2: Optimal hyperparameters for the decision tree model we chose, which was the one at iteration 16 of the forward selection process.

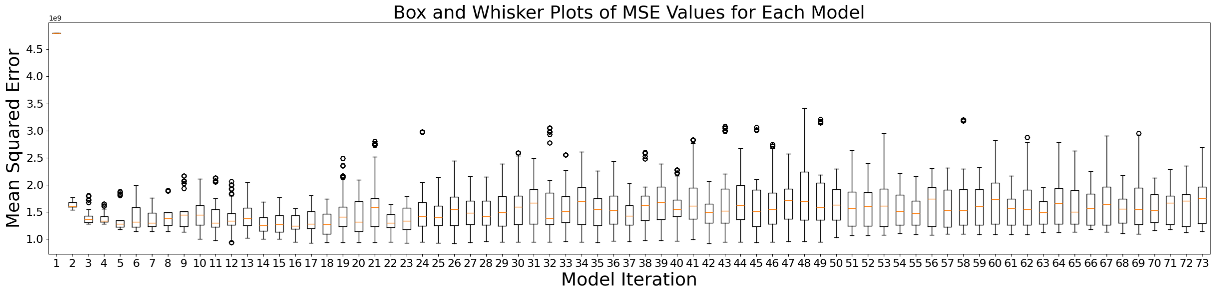

Figure 1: Each box and whisker plot corresponds to the MSE values from grid search on the best-performing feature subset for that iteration. The plot illustrates the MSE values obtained at each iteration of the grid search, representing model performance improvements as features are sequentially added.

The hyperparameters listed in Table 2 represent the configuration determined by grid search to be the most optimal for the decision tree model chosen.

Feature	Description
OverallQual	Overall material quality
GarageCars	Garage size
Utilities	Available utility types
Fireplaces	# of fireplaces
BsmtFullBath	# of basement bathrooms
KitchenQual	Kitchen quality
PavedDrive	Paved driveway
BldgType	Type of dwelling
GrLivArea	Above grade living area
GarageFinish	Garage interior finish
MSSubClass	Type of dwelling
1stFlrSF	First floor area (SF)
2ndFlrSF	Second floor area (SF)
KitchenAbvGr	Kitchens above grade
BsmtFinSF1	Type 1 basement area (SF)
BsmtCond	Condition of the basement

Table 3: List of selected features obtained for the model at iteration 16 of the forward selection algorithm. These features represent a subset of the original predictors and were chosen based on their contribution to improving the model's predictive performance.

The selected features, as shown in Table 5.3, were determined through the forward selection algorithm employed during the model selection process. These features represent a subset of the original set of predictors and were iteratively chosen based on their contribution to improving the model's predictive performance.

6 Results

6.1 Performance Metrics

The performance of each model was evaluated using two common metrics: Mean Absolute Error (MAE) and Mean Squared Error (MSE). MAE measures the average magnitude of errors in predictions without considering their direction [3]. Conversely, MSE calculates the average of the squares of the errors, giving higher weights to larger errors and penalizing outliers more heavily.

After retraining the decision tree model with the hyperparameters shown in Table 2, we evaluated its performance on the testing dataset. The resulting Mean Squared Error (MSE) was 823,650,565, while the Mean Absolute Error (MAE) was 20,194. While the MSE provides a measure of the average squared difference between predicted and true values, the MAE offers a more interpretable metric, representing the average difference between predicted and true values. Given that house prices in the Ames Housing dataset are typically in the hundreds of thousands of dollars [6], an MAE of 20,194 indicates that the model's predictions are somewhat adequate.

After examining the scatterplot in Figure 6.1, which illustrates the relationship between predicted and true prices, a positive correlation is evident. The majority of points are clustered around a line, indicating reasonably precise predictions with minimal variance in the data distribution. Additionally, the calculated Pearson's correlation coefficient of 0.932 confirms a strong positive linear relationship between predicted and true prices.

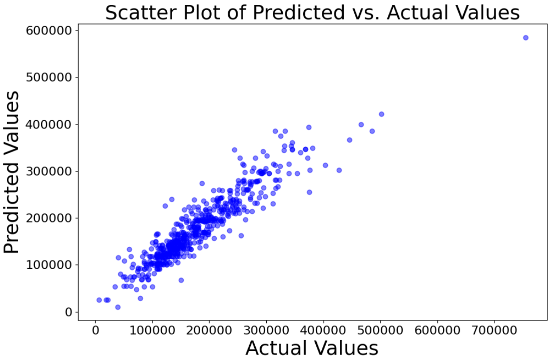

Figure 3: Scatterplot of predicted vs. true prices for the decision tree model. The x-coordinate represents the true price and the y-coordinate represents the predicted price. The strong positive correlation (Pearson's correlation coefficient = 0.932) between predicted and true prices suggests that the model's predictions are in line with the actual values.

7 Interpretation and Analysis

In this section, we interpret the decision tree model's variables. While interpreting decision trees can provide valuable insights into the underlying patterns within the data, in our analysis, we chose not to visualize the decision tree. This decision was influenced by a technical constraint related to the handling of categorical variables within the scikit-learn library. As mentioned in earlier sections, scikit-learn's models do not handle categorical variables natively; they require encoding them in some manner [10].

7.1 Visualization Problem

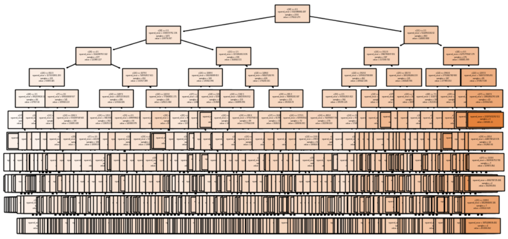

Figure 4: This figure shows a visualization of the decision tree. The extensive number of branches resulting from one-hot encoding of categorical variables complicates visualization and interpretation.

Because we utilized one-hot encoding to represent categorical variables, this approach results in the decision tree generating an extensive number of branches as shown in 4, rendering visualization challenging and interpretation cumbersome. Due to the sheer complexity and size of the resulting tree, attempting to visualize it would not provide clear or meaningful insights into the relationships between the predictor variables and the target variable. Therefore, instead of relying on visualizations, we examined the decision tree's feature importance to interpret it and extract actionable insights.

7.2 Feature Importances

Feature importance refers to the quantification of the relative contribution of each feature in a predictive model towards explaining the variability in the target variable [8]. It helps identify which features have the most significant influence on predictions, allowing for insights into the underlying relationships between input features and the target.

In this study, the importance value for each feature was computed using scikit-learn's decision tree implementation. The feature_importances attribute provides the importance of each feature in the decision tree model. This calculation is based on how much each feature contributes to reducing impurity (measured by Gini impurity) across all the nodes in the decision tree during the training process.

Impurity, in the context of decision trees, refers to the measure of uncertainty or disorder in a dataset. It quantifies how well a particular feature separates the data into classes. Gini impurity measures the probability of incorrectly classifying a randomly chosen element if it were randomly labeled according to the distribution of the classes in the node. It ranges from 0 to 0.5, where 0 represents perfect purity (all elements belong to the same class), and 0.5 represents maximum impurity (an equal distribution of classes). In decision tree algorithms, the goal is to minimize impurity when splitting nodes, meaning the feature that results in the greatest reduction in impurity (i.e., the feature that best separates the classes) is selected for splitting.

7.2.1 One-Hot Encoding Aggregation

When dealing with one-hot encoded features in the context of feature importance analysis, it's important to consider the aggregation of importance values back to the original categorical features [3]. One-hot encoding transforms categorical features into binary features, which can complicate the interpretation of feature importance.

Aggregating the importance values of the one-hot encoded features back to their original categorical features allows for a clearer understanding of the importance of the original categorical variables in the model. This process involves summing up the importance values of all binary features corresponding to each original categorical feature. By doing so, we can prioritize the impact of the categorical variables themselves rather than individual binary representations.

7.2.2 Interpreting Feature Importances

The feature importance table (Table 4) displays the feature importance of each variable. Importance values are expressed as percentages. For reference, ”SF” stands for ”square feet.”

Table 4: Feature Importances from Decision Tree

Feature	Importance
BsmtCond	0.525%
BsmtFullBath	0.311%
BldgType	0.289%
KitchenAbvGr	0.155%
Fireplaces	0.092%
GarageCars	0.045%
MSSubClass	0.043%
BsmtFinSF1	0.041%
KitchenQual	0.021%
2ndFlrSF	0.008%
1stFlrSF	0.006%
GrLivArea	0.003%
OverallQual	0.000%
Utilities	0.000%
PavedDrive	0.000%
GarageFinish	0.000%

According to the model, if we examine variables by magnitude, ”BsmtCond” (basement condition) emerges as the most influential feature, contributing 0.525% to the predictive power. Following closely are ”BsmtFullBath” (number of basement full bathrooms) and ”BldgType” (type of dwelling), with importance values of 0.311% and 0.289% respectively. Other predictors include ”KitchenAbvGr” (kitchens above grade) and ”Fireplaces,” each contributing 0.155% and 0.092% respectively. Surprisingly, features like ”OverallQual” (overall quality) demonstrate minimal importance, indicating their limited impact on housing prices in this model. However, it’s noteworthy that the overall feature importances are strangely very low, suggesting potential limitations or biases in the dataset or model.

8 Discussion

To recap, we hypothesized that certain factors such as location, size of rooms, and overall quality, will have a significant positive impact on house prices. However, the analysis reveals that while variables like ”BsmtCond” and ”BsmtFullBath” play a part in predicting house prices, features like ”OverallQual” and ”Utilities” exhibit minimal importance. This suggests that our initial hypothesis may not fully align with the predictive model’s findings, indicating potential nuances or complexities in the determinants of house prices in the Ames Housing dataset. This is compounded by the fact that the feature importances have values below 1%. Further investigation into the dataset and model performance are required to provide deeper insights into these discrepancies.

9 Research Limitations and Future Directions

While our study provides valuable insights into predicting house prices using machine learning models, several limitations and avenues for future research should be considered.

Firstly, it’s important to recognize that Dean De Cock’s dataset is specific to residential homes in Ames, Iowa. While the findings offer insights into housing market dynamics in this locality, generalizing the results to other geographical areas should be done cautiously. The characteristics of the Ames housing market, such as demographics, economic conditions, and housing policies, may differ significantly from other regions, impacting the relevance and applicability of the findings. Additionally, housing markets vary in terms of supply-demand dynamics, cultural preferences, and regulatory environments, which can influence property prices and market trends differently. Therefore, while our analysis provides insights, it’s essential to exercise caution when extrapolating findings to other housing markets or datasets.

Some potential directions for validating the model with diverse datasets from different locations include external validation with datasets from other housing markets and exploring ways to adapt the model to different contexts. Additionally, if one wanted to study housing prices in another location, one would need to find another dataset that collected information about houses and their prices in other locations, let alone find one that has data that's compatible with the Ames Housing dataset.

Secondly, the peculiarities of the train-test split in our dataset may warrant reshuffling and redistributing the data or cross-validation for more robust model evaluation.

Thirdly, one notable limitation of our analysis is the unexpectedly low feature importance values observed in our models. The features overall exhibited seemingly negligible importance, despite the adequately accurate predictions of house prices, as indicated by the 20,194 MAE value. Future research should explore potential reasons for these discrepancies. Moreover, considering the possibility of imputation drastically affecting the results as there were a large number of missing entries, it might be necessary to collect additional data or refine feature engineering strategies to enhance model predictive power.

Furthermore, some aspects of data exploration remain to be addressed to provide us with a deeper understanding of the dataset. Exploring the underlying structure of the data and identifying strong correlations between predictors could enhance the interpretability and robustness of our models. Visualizations such as mean values and standard deviations could provide valuable insights into the distribution and variability of the data, aiding in identifying potential outliers or anomalies. Future research should focus on conducting more comprehensive data exploration to better characterize the dataset and uncover hidden patterns or relationships that could further improve model performance and explain the discrepencies in the feature importances.

Fourthly, given that only a subset of variables had some level of feature importance, exploring dimensionality reduction techniques like Principal Component Analysis (PCA) could be beneficial. PCA can help identify latent variables and reduce the dimensionality of the feature space while preserving the most critical information, potentially improving model performance and interpretability [5].

Lastly, there were concerns about continuously updating and retraining the model in response to dynamic changes in society. As societal factors evolve over time, such as changes in demographics, economic conditions, and housing policies, the predictive model must adapt accordingly to maintain its effectiveness. Therefore, implementing a mechanism to periodically update the dataset and retrain the model with new data is important. This may involve regular data collection efforts to capture the latest market trends, as well as reevaluation of model performance and recalibration of predictive features.

10 Conclusion

In this study, we embarked on a journey to decipher the intricate dynamics of house prices by employing predictive modeling techniques on the Ames Housing dataset. Through preprocessing, feature selection, and model evaluation, we aimed to uncover the key factors influencing house prices and develop an accurate predictive model.

Preprocessing involved imputing missing data and ensuring compatibility with machine learning algorithms, while feature engineering encompassed selecting relevant predictors and transforming variables to enhance predictive power. We employed techniques such as forward feature selection, k-fold cross-validation, and grid search hyperparameter tuning to iteratively refine our model and optimize its performance. By systematically evaluating various models and configurations, we sought to identify the most influential predictors and construct a predictive framework capable of accurately estimating house prices.

We sought to identify the most influential predictors and construct a predictive framework capable of accurately estimating house prices. Despite our initial anticipation that traditional factors like such as location, size of rooms, and overall quality will have a positive relationship with house prices, they showed minimal importance in our model.

Additionally, several limitations and future research directions should be considered. The dataset, specific to Ames, Iowa, limits generalizability due to varying market dynamics in other regions. Validation with diverse datasets and adapting the model to different contexts are potential solutions. The train-test split's peculiarities should be examined in more depth. The low feature importance values, despite accurate predictions, indicate a need for further exploration of data structure and correlations. Dimensionality reduction techniques like PCA could enhance model performance and interpretability. Additionally, continuous model updates are necessary to adapt to societal changes, requiring periodic data collection and model recalibration.

In conclusion, our study contributes to advancing the understanding of housing market dynamics and provides valuable insights for stakeholders involved in real estate decision-making. By leveraging predictive modeling techniques and data-driven approaches, we can empower individuals and organizations to make informed decisions in the dynamic and complex realm of real estate.

References

[1] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Taylor Francis, 1984.

[2] John Clemmer. Exploratory data analysis of housing in ames, iowa. https://www.kaggle.com/code/leeclemmer/exploratory-data-analysis-of-housing-in-ames-iowa, 2017. Accessed: 2024-06-12.

[3] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[5] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[6] Kaggle. Ames Housing Dataset. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data, 2016. Accessed: May 8, 2024.

[7] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013.

[8] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2023.

[9] Aparna Nagarajan, Emmanuel Ogwal, Sirisha Yellajosyula, Yezhou Jiang, Sean Xu, and Steven Choi. Grp12 report - house prices in ames, iowa. *ResearchGate*, 2020.

[10] Scikit-learn Contributors. Scikit-learn documentation, n.d.

[11] Quang Truong, Minh Nguyen, Hy Dang, and Bo Mei. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 174:433–442, 2020. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.