The Correlation of Housing Features and Sale Price in Ames, Iowa

By Alice Vadney, Athena Vo, Avery Donmoyer, and Naylor Stone

Abstract

This project investigates the factors influencing home sale prices in Ames, Iowa, using advanced machine learning techniques to predict future prices accurately. By analyzing a dataset containing diverse housing attributes, including square footage, lot size, overall quality, and the number of bedrooms and bathrooms, we sought to identify the features most strongly correlated with sale prices, with the eventual goal of predicting sales prices for other houses in the same neighborhood. The study combines data preprocessing, exploratory data analysis, and model building to provide actionable insights into real estate valuation.

Our approach began with cleaning and transforming the data, consolidating complex variables like bathroom types and creating new features such as total square footage and house age. Decision tree models were initially employed to evaluate the predictive power of individual features. The most effective tree model, determined by its R-squared (R^2) value, served as the basis for a random forest model, which enhanced predictive accuracy through aggregation and bootstrapping techniques. The final random forest model achieved an R^2 value of 0.87, demonstrating its robustness and reliability in predicting home prices.

The analysis revealed that square footage and overall quality are the strongest predictors of sale price, while variables like the number of bedrooms and the year sold had minimal impact. Visualizations, including scatter plots and residual analyses, confirmed the model's accuracy and highlighted its ability to minimize prediction errors. Despite its success, the model faced some challenges such as the exclusion of external economic factors that couldn't be accounted for in the dataset and the need to simplify complex data structures for inclusion in the model.

Introduction

Predicting home sale prices is a crucial task in real estate analytics, offering insights for buyers, sellers, and policymakers. This project explores the housing market in Ames, Iowa, using advanced predictive modeling techniques to identify the features most strongly correlated with home sale prices. By leveraging a dataset of housing listings, we aim to answer two central questions: Which housing features are most strongly correlated with the sale price of a home in Ames, Iowa? How accurately can future home sale prices be predicted based on these features?

Our dataset includes a wide range of housing attributes, such as the number of bedrooms, bathrooms, square footage, lot area, overall quality and condition, age of the house, and the year it was sold. Through careful data preprocessing and feature engineering, we refined these variables to create a streamlined dataset that effectively captures the key drivers of home value. This paper outlines the steps taken to analyze the data, build predictive models, and interpret the results, offering a comprehensive guide to understanding the factors that influence housing prices.

To approach this problem, we began with exploratory data analysis to uncover patterns and relationships within the data, followed by the creation of decision tree models to evaluate the predictive power of individual features. The most effective decision tree was then used as a foundation for a random forest model, which aggregates multiple decision trees to enhance accuracy and robustness.

We also had to think critically about the variables that were already present in the dataset and predict the ones that we thought would be most useful in our model. In order to get a start on this, we each took a few variables and ran our own exploratory data analysis and visualization to help inform us on the usefulness of each particular variable. When it came time to actually build our random forest model and run the regression, we had a good idea of which variables we should include, which variables not to, and which variables needed to be further investigated.

Key steps included data preprocessing: calculating new variables and combining related features for simplicity, model development: building and validating decision tree models to identify the best predictors, then bootstrapping these trees into a random forest model, and performance evaluation: using metrics such as R-squared values, scatterplots, and residual analysis to assess model accuracy.

Our random forest model achieved an R^2 value of 0.87, indicating strong predictive accuracy. This model reliably predicted home sale prices based on the training and testing datasets, as confirmed by visualizations such as kernel density plots and scatterplots of predicted versus actual prices. Residual analysis further validated the model, showing that prediction errors were minimal and normally distributed.

Square footage and overall quality were found to be the strongest predictors of sale price, underscoring the importance of house size and the house's overall status in determining a home's value, while variables such as the number of bedrooms and the year sold had the lowest predictive power of the group of variables, suggesting that they contribute less to overall price determination.

While the model performed well overall, several challenges emerged during the course of the project. In the beginning stages, many variables that should logically be one observation were split into subcategories in the initial dataset. For example, data on the number of bedrooms could be found in two different columns: the number of bedrooms above grade and the number below grade. To accurately assess the total number of bedrooms per house, that data had to be consolidated into one column for effective analysis. Additionally, the current economic condition of the housing market in which this data was collected (for example, various financial crises caused by economic events such as the COVID-19 pandemic, etc.) was not outwardly included in the dataset, even while these factors likely impacted housing prices. We're therefore unable to account for how any existing economic factor might have influenced the data used to train the model. Lastly, the dataset included too many variables for all of them to be effectively included in the model. Therefore, it's possible that some of the features excluded from the model could have been important predictive indicators, which could limit the model's overall comprehensiveness.

Data

We worked with housing data from Kaggle for our project. The dataset included 79 explanatory variables used to predict the sale prices of homes in Ames, Iowa. These variables covered features within the house, the yard, and the surrounding neighborhood. From this dataset, we identified 10 key variables to focus on during our analysis: the number of bedrooms, the number of bathrooms, lot size, year built, year sold, garage capacity, floor plan square footage, type of utilities available, neighborhood, and house style.

This data was useful for determining the sale price of a house by offering a variety of factors that could influence homebuyers' willingness to pay a certain price. Square footage was included in the list of input variables because the size of a house is traditionally positively correlated with its price. Square footage is often viewed as a primary measure of a house's size and value (the larger the house, the higher the price). However, square footage was just one of many facets of a house's value.

We also analyzed the utilization of the house's space as a measure of value. To do so, we examined the number of bedrooms and bathrooms in the house, as well as the garage capacity. These variables served as additional indicators of the house's size, as larger houses tend to have more bedrooms, bathrooms, and garage space. Additionally, lot size was considered valuable when calculating the sale price. A larger lot traditionally increases a house's value, even if the square footage of the house itself does not increase.

We studied the year the house was last sold, which could indicate how modern the house's foundation, construction, and appliances were. Oftentimes, homeowners update appliances and other features before they sell in order to appeal more to buyers. Modern appliances are often highly sought after by homebuyers, so more recently sold houses might reflect greater desirability and, consequently, higher prices.

The style and features of the house also appeared to influence the final sale price. Homebuyers often have preferences for certain house styles (e.g., modern, craftsman, Victorian), especially in any given area. The style could thus drive the price of a home up or down depending on its marketability. Additionally, specific amenities may appeal more to some buyers, increasing a house's marketability.

Lastly, we considered neighborhood amenities, such as proximity to schools, pools, and other attractions. The neighborhood where a house is located can significantly affect its desirability and, therefore, its price.

One challenge we faced in using these variables was that many of them were split into multiple categories within the dataset. For instance, the number of bathrooms was represented as full baths above grade, basement full baths, half baths above grade, and basement half baths. We resolved this by combining these variables into a single metric, simplifying the data and making it easier to analyze meaningfully.

Methods

The primary goal of this project was to build a predictive model for housing sale prices in Ames, Iowa, using the 10 key variables mentioned previously. An observation in our study was one individual house sale record and all of its attached attributes, including house square footage, year built, number of bedrooms, number of bathrooms, neighborhood, and more. We approached the process of prediction with the plan of building a random forest model. Our data had already been split into training and test sets, and we used that built-in organization while building our model.

To begin, we built and trained a number of decision trees, using many different combinations of predictive variables to find the ones with the strongest predictive power. We then finalized one decision tree, built from the best combination of variables and bootstrapped that data into a random forest model to better capture data patterns and relationships. The final model was then trained on the training dataset.

We evaluated our final model's success through a number of indicators, including the value of R^2. As the value of a model's R^2 increases, becoming closer to 1.0, it is an indication that the model accurately explains a substantial proportion of the data variance, meaning its predictions are more accurate. For this project, we determined that a threshold of around 0.8 or higher would indicate strong predictive power, though the specific threshold will depend on what's typical in similar models built to predict housing prices. Additionally, we used a residual plot and a kernel density plot to evaluate the model visually.

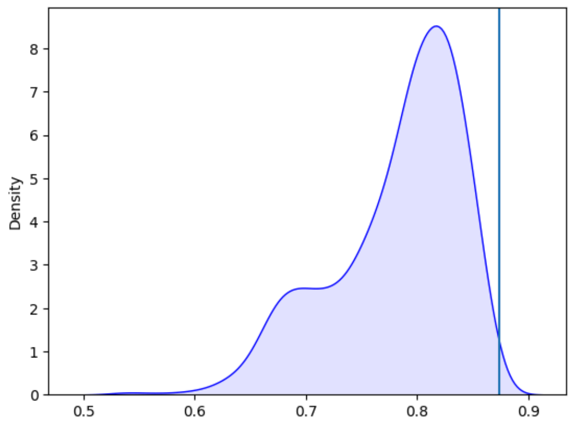
We also evaluated model performance across the individual decision trees to ensure that the bootstrapped data generated for the purpose of creating a random forest. If metrics such as R^2

vary widely across the individual decision trees, it may indicate an unreliable model from the bootstrapped data. Lastly, residual analysis can be an important factor in ensuring that there is no bias in the model, by over or underestimating certain price ranges, and that errors are uniformly distributed. If residuals show patterns, it may indicate missing variables or areas where the model could improve.

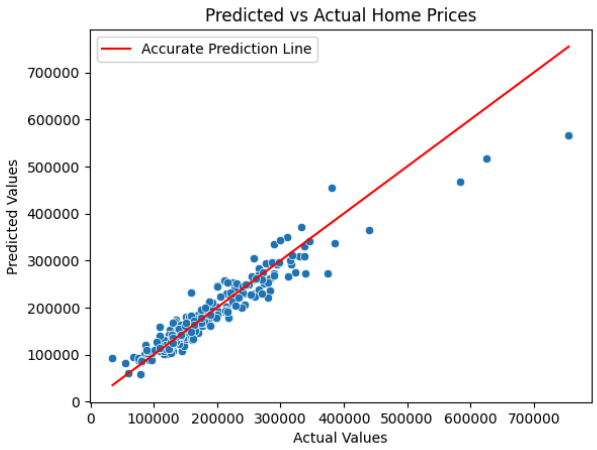
If our model turned out to be inaccurate, that could indicate that it was either underfit, meaning that there wasn't enough data for the model to make accurate predictions, or overfit, meaning the model had too much training data and produced over-generalized results. Failure might also indicate that our dataset was missing critical variables that significantly influence house prices, such as economic conditions, neighborhood-specific trends, or unrecorded features like recent renovations. It could also indicate that the housing market is more complex, influenced by non-linear interactions between features, than we had previously thought.

Results

To begin analyzing the data, we developed a series of decision tree models, testing various combinations of variables to understand their individual and collective impact on housing sale prices. These preliminary models allowed us to identify the features most indicative of price variations. After evaluating the decision trees based on their R^2 values, we selected the model with the highest R^2 value and bootstrapped that model to create a random forest, the results of which are discussed here.

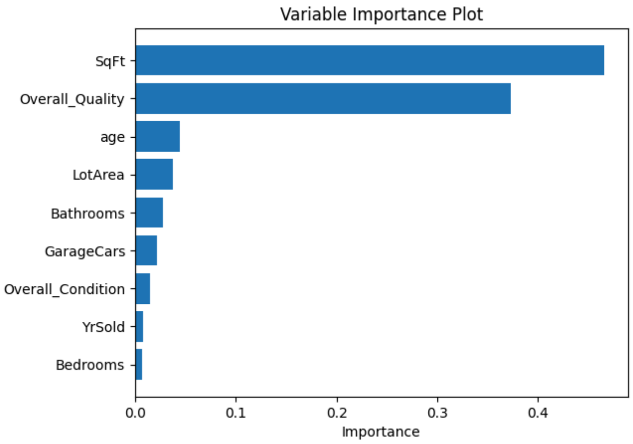


Kernel density plot of R^2 values, with the random forest model’s R^2 value indicated by the vertical blue line.



Scatterplot of predicted sale prices versus actual sale prices for the test dataset.

Our random forest model had an R^2 value of 0.87, indicating that it is reliably accurate at predicting the sales price of houses in Ames, Iowa based on the train and test datasets. To further assess the model's performance, we calculated a kernel density plot to situate the model's R^2 value within the greater context of the data. To further evaluate the model's performance, we plotted a kernel density plot to contextualize the R^2 value within the broader dataset. Additionally, a scatterplot of predicted vs. actual sale prices was generated, whose slope was close to one, further validating the model's accuracy. The residuals were centered around zero and exhibited a relatively normal distribution, suggesting minimal bias in the predictions.



Bar chart of importance variables in the random forest model.

In further analysis, we found that **overall house quality** and **house square footage** were the most significant predictors of sales price within the model. This was determined by calculating the mean decrease impurity, a metric that quantifies the contribution of each variable to the model's predictive power. Conversely, the **number of bedrooms** and the **year sold** were found to have the least predictive value.

Though we did not include all potential predictor variables from the dataset in this model, we explored the variables in our exploratory data analysis and determined that the variables in the model were the ones that were most likely to be correlated with home price.

Conclusion

This study sought to explore the factors most strongly correlated with home sale prices in Ames, Iowa, and to develop a predictive model capable of estimating these prices with accuracy. Through an extensive analysis of the dataset and the implementation of a random forest model, we identified key predictors of housing prices, such as overall house quality and square footage, while also addressing challenges in working with real-world data. Our findings provide insights into the housing market in Ames, Iowa, and set the stage for future work that can expand and refine these methods.

Our analysis revealed that overall house quality and square footage are the most significant predictors of sale price. These features consistently demonstrated the highest importance in the random forest model, as indicated by the variable importance plot. Conversely, variables such as the number of bedrooms and the year a house was sold showed relatively low predictive power, suggesting that they play a less critical role in determining home values.

The final random forest model achieved an R^2 value of 0.87, indicating a strong ability to predict home prices based on the selected features. Visualizations, including scatter plots and kernel density plots, confirmed the model's accuracy, with predicted values closely aligning with actual sale prices. Residual analysis further validated the model, showing a near-normal distribution of errors centered around zero.

While the results of this project are promising, there are opportunities for improving the accuracy of this work. One potential avenue is the collection of additional features about the houses that could provide additional predictive power. For example, data about the average household income by neighborhood, the availability of a range of public services, and rankings of local schools could all further strengthen a predictive model. However, this data wasn't included in our dataset and therefore couldn't factor into the model.

Another possible place for improvement would be to compare different types of predictive models. While the random forest model performed well, other machine learning approaches could be tested to determine whether they offer improved accuracy or computational efficiency. For example, we could have used a model like a neural network, which also works well with datasets containing many different variables, or we could have implemented hedonic pricing to get a better sense of each characteristic's individual effect on total price.

One weakness with our approach is the high dimensionality and multicollinearity within our dataset. The dataset contains a large number of features, some of which might be highly correlated with each other. This can increase the model's complexity, leading to overfitting, where the model performs well on the training data but poorly on unseen data. Random forests inherently provide feature importance scores. Thus, after training the model, we analyzed these scores to identify features that contribute minimally to the prediction.

One challenge that arose from the project was the geographic scope of the dataset. The reliance on a single dataset from Ames, Iowa, limits the generalizability of the findings to other housing markets. Real estate markets vary significantly across regions due to differences in economic conditions, buyer preferences, and other factors. Future research could apply the methods developed here to datasets from other locations to assess their broader applicability. It would be interesting to try applying our model with data from other areas and seeing how the strength of each variable changes spatially.

Another important note is that several features in the dataset contained missing values. Features like lot frontage, garage year built, or basement quality have many missing values. Additionally, outliers such as extremely high house prices or unusual feature values (e.g. unusually large lots) can distort the model’s performance. To mitigate this, we dropped features that have a high proportion of missing values to help reduce the complexity of the model.

Although random forests are less prone to overfitting than individual decision trees, the forest can still overfit the training data if its total number of trees is too large, or if those trees are too deep, capturing noise instead of the underlying pattern and resulting in poor generalization to unseen data. To mitigate this, we carefully tuned the hyperparameters such as the number of trees (*n_estimators*), maximum depth of each tree (*max_depth*), and minimum samples per leaf (*min_samples_leaf*). Random forests also provide an OOB error estimate, which we monitored to detect overfitting and help adjust the hyperparameters. Future iterations of the model could incorporate feature selection techniques such as principal component analysis (PCA) to reduce dimensionality and improve generalizability.

One of the limitations of the current analysis is that it focuses solely on housing-specific factors, without accounting for other external considerations that influence housing prices. While our dataset provides a rich set of 80 housing-related variables, these features alone may not fully explain correlations with sale prices. Economic conditions, census information, or regional amenities could play important roles. For example, unemployment rates or inflation during the data collection period could have shaped market behavior in ways our model cannot capture. Incorporating feature augmentation and multi-source integration could improve the model’s explanatory power. While housing data alone offers strong predictive capabilities, as evidenced by our model's R^2 value of 0.87, integrating external datasets could provide a more holistic understanding of the factors driving housing prices and support future studies in creating more accurate predictive frameworks.

Overall, this study provides a framework for analyzing and predicting home sale prices in a given region using machine learning techniques. Despite the challenges and limitations encountered, the findings show the importance of key features such as overall house quality and square footage in determining home values. By addressing the complications and extending the methods outlined in this project, future work can build on this foundation to develop even more accurate models for predicting the sales prices of homes.

Bibliography

Palermo, M. (2019). Ames Iowa Housing Data. Retrieved September 27, 2024 from https://www.kaggle.com/datasets/marcopale/housing.