House Value Prediction for Strategic Real Estate Decision-Making

Anna Dominic
amd9200@nyu.edu

Karthikeyan Shanmugam
ks6964@nyu.edu

Manisha Goyal
mg7609@nyu.edu

Siri Desiraju
scd4156@nyu.edu

Contents

1 Business Understanding

1.1 Objective

The real estate market is a critical component of the economy, with property values affecting a wide range of stakeholders from individual homeowners to large real estate developers. The primary objective of this project is to develop a predictive model that accurately estimates residential property prices in Ames, Iowa, based on a diverse set of features. By leveraging machine learning techniques and advanced statistical methods, this project aims to support various stakeholders in the real estate industry in data-driven decision-making.

1.2 Impact

This predictive model empowers stakeholders to make well-informed decisions, optimizing their investment strategies and enhancing service quality. The insights it provides shed light on the various factors that influence property values, fostering more strategic and calculated business actions. The model serves as a crucial tool for a range of applications, including investment decisions, renovation planning, and market analysis, offering valuable guidance to individuals and organizations alike.

1.3 Target Stakeholders

Real Estate Agents: For agents, the model provides a competitive edge by enabling them to predict the value of properties accurately and hence, set prices that are both attractive to buyers and satisfactory to sellers. This helps in optimizing listing prices and improving the quality of services offered.

Real Estate Developers: Developers can use the insights from the model to plan and execute projects that are more likely to attract buyers. By focusing on features that add value, developers can invest in profitable ventures and plan developments that meet market demands.

Homeowners: The model assists homeowners in making decisions about potential remodeling and upgrades. By understanding how different factors affect house values, homeowners can invest in renovations that maximize their property's market value.

Home Buyers: For buyers, the model acts as a tool to assess fair market prices and evaluate property investments, enabling well-informed purchasing decisions. It facilitates effective negotiations and helps in identifying opportunities for profitable investments while mitigating risks associated with overvalued assets.

2 Data Understanding

2.1 Data Source

The dataset for this project, sourced from Kaggle, comprises detailed information on over 1,400 properties in Ames, Iowa, spanning approximately four years. It includes 79 explanatory variables that describe almost every aspect of residential homes, from physical attributes like the lot size and building type to aesthetic

features and utilities. This comprehensive dataset allows for an extensive analysis of factors affecting house prices in this region.

2.2 Data Features

The dataset features a mix of approximately 35 numeric and 46 categorical variables. Some of these variables are summarized in Table 1 below:

Table 1: Features Description

Feature	Description
Property Type (*MSSubClass*)	Defines type of dwelling involved in the sale, such as 1-story or 2-story homes, split-levels, and duplexes
Lot Size (*LotFrontage, LotArea*)	Measures property's frontage in linear feet and total lot size in square feet
Building Quality (*OverallQual*)	Ranks overall material and finish of the house on a scale from 1 to 10
Year Built (*YearBuilt*)	Indicates original construction date

These variables, among others, provide a rich, multilayered perspective on the real estate market, offering critical insights into property valuation.

2.3 Potential Bias

Although the comprehensive detail provided helps minimize collection bias, some variables, such as those assessing the quality of materials or finishes, inherently contain a level of subjectivity. Additionally, potential biases may arise from non-random sampling or data reporting standards that differ over time or across categories. For the scope of this project, we recognize these possible biases but assume the data's credibility based on its comprehensive nature.

2.4 Preliminary Observations

An initial review of the dataset reveals a diverse range of properties, from modest single-family homes to larger, more complex structures. The variety in age, size, and style of the properties suggests that the Ames housing market caters to a wide demographic, which is essential for modeling realistic house price predictions.

3 Exploratory Data Analysis (EDA)

To better understand the underlying patterns and insights within the dataset, we performed some EDA. This analysis helped in identifying the distribution of various features, their relationship to house prices, and any potential correlations that could influence predictive modeling. By scrutinizing these aspects, we

gain a clearer picture of the factors that significantly impact house valuations, guiding the subsequent data preparation and feature engineering steps.

3.1 Distribution of Sale Price

The distribution of sale prices within the dataset provides insights into the range of property values. The histogram of sale prices (see Figure 1) illustrates a right-skewed distribution, where the majority of homes are clustered around the lower to mid-range price segments, with fewer homes extending into the higher price brackets. This skewness indicates that while most homes are affordable, there are a few premium properties that significantly differ in price due to superior features or locations.

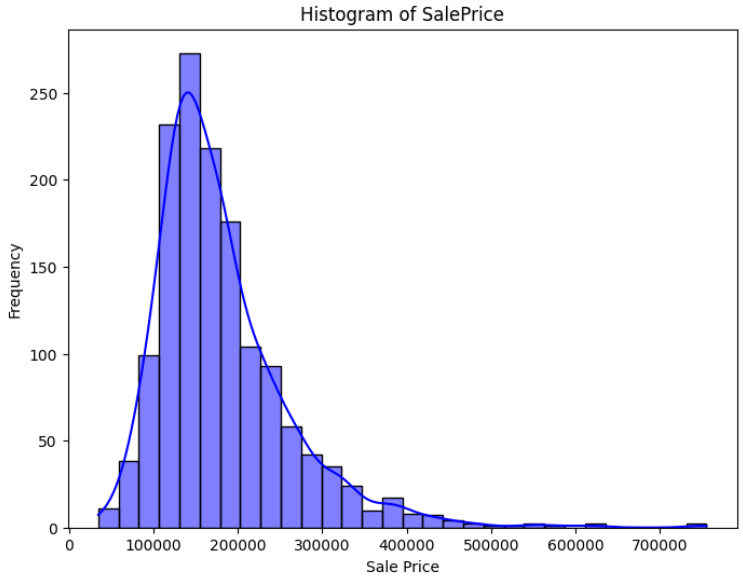

Figure 1: Sale Price Distribution

3.2 Correlations and Relationships

Analyzing the relationships and correlations between various house features and the sale price is essential for understanding what drives property values. The correlation matrix (see Figure 2) highlights strong associations between sale price and several house features. Key findings include:

- **Overall Quality (*OverallQual*)** shows a strong positive correlation with sale price, reinforcing that the quality of materials and finish significantly influences property values.
- **Living Area (*GrLivArea*)** and **Garage Cars (*GarageCars*)** also exhibit strong positive correlations with sale price, suggesting that larger living spaces and more garage capacity are highly valued by buyers.
- **Exterior Quality (*ExterQual*)** and **Kitchen Quality (*KitchenQual*)** are positively correlated with sale price as well, indicating that quality of the material on the exterior and well-appointed kitchens enhance a home's appeal and value.

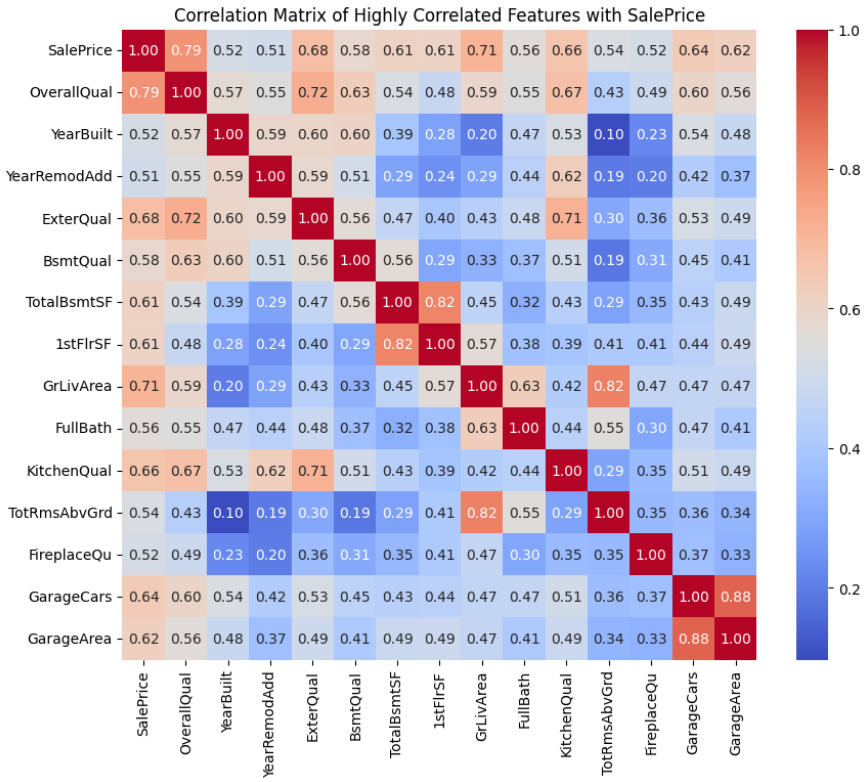

Figure 2: Correlation Matrix

3.3 Hypothesis Testing

We conducted certain hypothesis tests to validate some of the initial assumptions we had going into this project. Two of them along with their findings are presented below.

1. **Impact of Remodeling on Sale Price**

Hypothesis: Properties remodeled in the last 10 years sell at higher prices than those not remodeled, controlling for property age.

Methodology: Linear regression was used, incorporating whether a property was recently remodeled and its age as predictors of sale price.

Findings:

- **Model Performance**: The model's R-squared value is 0.273, indicating that approximately 27.3% of the variance in sale prices is explained by the model, while the adjusted R-squared is similarly 0.273, suggesting a minimal penalty for the number of predictors used.
- **Coefficients**: The coefficient for property age was -$1,375.37 per year, significant with a p-value less than 0.001, indicating that as properties age, their value decreases significantly. The analysis did not yield a significant coefficient for the remodeling variable, which might be due to multicollinearity or

other issues in the data.

- **F-Statistic**: The F-statistic is 548.7, with a corresponding p-value nearly zero (2.99e-103), strongly suggesting that the regression model as a whole is statistically significant.
- **Statistical Issues**: The lack of a significant coefficient for the remodeling effect might indicate strong multicollinearity, as suggested by the condition number and the presence of a warning regarding the potential singularity of the design matrix. Additional diagnostics or model adjustments may be necessary to isolate the effect of remodeling.

2. **Influence of Property Type and Style on Sale Price**

Hypothesis: Certain types of dwellings (*BldgType*) and architectural styles (*HouseStyle*) are more sought after and therefore sell at higher prices.

Methodology: ANOVA tests were conducted to assess the impact of these categorical variables on sale price.

Findings:

- **ANOVA Results**: For *BldgType*, the F-statistic is 13.01 with a p-value of 2.06×10^{-10}, indicating significant differences in sale prices across building types. For *HouseStyle*, the F-statistic is 19.60 with a p-value of 3.38×10^{-25}, showing significant differences in sale prices across house styles.
- **Tukey's HSD Test Results**: For *BldgType*, significant differences were found between several categories. Notably, single-family homes (*1Fam*) typically sell for more than duplexes, two-family conversions (*2fmCon*), and townhouse inside units (*Twnhs*), but not significantly different from townhouse end units (*TwnhsE*). For *HouseStyle*, significant price differences were observed. For instance, one-story (*1Story*) homes sell at higher prices compared to one and one-half story (*1.5Fin*) homes, and two-story (*2Story*) homes also command a premium over several other styles.
- **Post-hoc Analysis**: The Tukey tests detailed specific pairwise comparisons where significant price differences exist. For example, single-family homes sell for significantly more compared to duplexes and two-family conversions, with a difference ranging from approximately \$49,852 to \$57,331.

4 Data Preparation

4.1 Cleaning

The initial step in preparing our dataset involved a thorough cleaning process focused on ensuring data quality and integrity. The following are the key actions taken during the data cleaning phase:

Initial Identification of Missing Values: We began by identifying columns with missing values and assessing the extent of the missing data. This initial analysis categorized the missing values into three levels:

- **High Missing Values**: Features like *PoolQC* and *Alley* showed more than 50% missing data, indicative of the absence of certain property features (e.g. no pool or alley access). Given their informational

value in indicating the absence, we opted not to impute these values.

- **Moderate Missing Values**: Variables such as *LotFrontage* and *BsmtQual* displayed moderate levels of missing data (5-50%). These required tailored strategies for imputation or preservation.
- **Low Missing Values**: Features with less than 1% missing data, such as *MasVnrArea* and *Electrical*, suggested straightforward imputation opportunities without impacting the dataset significantly.

Strategies for Handling Specific Variables: For the variables identified with moderate to low missing values, we applied specific strategies, summarized in Table 2 below:

Table 2: Strategy for Handling Moderate to Low Missing Values

Feature	Strategy
LotFrontage	We imputed missing values using the median of the column
GarageYrBlt	For properties without a garage, we assigned a 0 to indicate the absence of a garage
MasVnrArea	In cases where *MasVnrType* was None, we set it to 0 to maintain consistency of the numerical variable
Electrical	Given the minimal number of missing entries, we imputed these with the category *Mixed*

Preservation of Meaningful NA Values: For some features, the NA values were meaningful and indicated the absence of a property feature, such as no basement or no garage. Therefore, these NA values were preserved to accurately reflect the property's characteristics.

Handling of Outliers: Outliers, such as properties with significantly high prices, were identified but retained in the dataset. These outliers are considered valuable for capturing the full spectrum of the real estate market, particularly for modeling luxury properties.

Final Adjustments: A small number of rows with unresolved missing values in *MasVnrArea* after attempts at imputation were removed from the dataset.

4.2 Feature Engineering

In the feature engineering phase, we enhanced the dataset by encoding existing categorical variables to numeric variables, developing new features, and normalizing the dataset.

Encoding Categorical Variables:

- **Ordinal Variables**: These include general quality and condition ratings (e.g. *ExterQual*, *BsmtQual*) and specific features like *LotShape* and *BsmtExposure*. We defined mappings for these variables, assigning integer values ranging typically from 0 (absent) to 5 (excellent), to convert these ordinal ratings into a numeric format.

- **Nominal Variables**: These include variables like *MiscFeature* and *GarageType*. We applied one-hot encoding to transform each category into a new binary column, where a 1 represents the presence of a category, and a 0 represents its absence. Special attention was paid to the variable *MSSubClass*, which, despite being numeric, categorizes types of dwellings and was treated as nominal by converting to string and then applying one-hot encoding.

Developed New Features: The newly developed features are summarized in Table 3 below:

Table 3: Description of Newly Created Features

Feature	Description
AgeAtSale	The property's age (difference in the year a property was sold and the year it was built)
YearsSinceRemodel	Time elapsed since the property was last remodeled
TotalBathrooms	Aggregated count of full and half bathrooms in the property
TotalBasementBathrooms	Aggregated count of full and half bathrooms in the basement
TotalPorchArea	Sum of all porch areas (including indoor, outdoor, and screened porches)

Normalization: The entire dataset was normalized using the min-max scaling technique, facilitating more effective model training by treating all features with equal importance during the learning process.

4.3 Feature Selection

After the data cleaning and feature engineering phases, our dataset expanded to include 213 features. To enhance model efficiency and interpretability, we applied feature selection techniques to identify the most influential variables. We evaluated three different methods: Lasso Regression, Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA), benchmarking their performance on a baseline linear regression model to determine their effectiveness.

1. **Lasso Regression**: This method applies a penalty to the size of the coefficients in regression models, effectively shrinking less important feature coefficients to zero, thus performing feature selection. Lasso Regression selected 88 features and achieved an R^2 score of 0.873, indicating strong predictive performance with a reduced set of features.

2. **Recursive Feature Elimination (RFE)**: RFE works by recursively removing the least important features based on the model weights. It drastically reduced the feature set to just 5 features, resulting in an R^2 score of 0.664, which was significantly lower than the other methods, suggesting that too much information may have been lost.

3. **Principal Component Analysis (PCA)**: PCA reduces dimensionality by transforming features into a set of linearly uncorrelated components. It retained 107 features and achieved an R^2 score of 0.872. While PCA provided a high degree of dimensionality reduction without a substantial loss in model

performance, the resulting components lack easy interpretability.

Given the importance of interpretability in our project, particularly for communicating key insights to stakeholders, Lasso Regression was chosen as our feature selection method. Its ability to maintain a balance between model simplicity and predictive accuracy, along with providing an interpretable set of features, made it the preferred choice.

5 Modeling

5.1 Model Development Strategy

The strategy for developing our predictive models was structured to balance interpretability with computational efficiency and predictive accuracy. The following an overview of our approach:

1. **Diverse Modeling Techniques**: We utilized both interpretable and complex models to harness their respective strengths. Models like Random Forest provide valuable insights due to their interpretability from a decision-tree structure, aiding in understanding the impact of various features. Simultaneously, we employed advanced models like XGBoost, known for their ability to handle complex, non-linear relationships and provide better predictive performance.

2. **Comprehensive Feature Analysis**: Our approach included a thorough analysis of all 213 available features and a selected subset of 88 features identified during feature selection (see section 4.3 Feature Selection). Exceptions include CatBoost and Ensemble methods, which were used only with the full feature set due to their specific characteristics that benefit from a broader data range. This comprehensive analysis was crucial in identifying the most influential features and assessing how the dimensionality and composition of the dataset influenced the effectiveness and accuracy of the predictive models.

3. **Parameter Optimization**: We utilized Grid Search combined with Cross-Validation to fine-tune our models, ensuring optimal performance.

4. **Evaluation Metrics**: To assess the performance of our models, we employed two key metrics: R^2 and Root Mean Square Error (RMSE). These metrics provided insights into the models' accuracy and error rates, helping us refine our approach and select the best-performing models.

5.2 Models

We evaluated various modeling techniques to predict house prices. The subsections below detail each model's performance, highlighting their effectiveness with various feature configurations and demonstrating the adaptability of different modeling techniques to the complexities of the dataset.

5.2.1 Lasso Regression

As mentioned in section 4.3 Feature Selection, Lasso Regression was utilized to harness its feature reduction capabilities via regularization. The optimal regularization parameter (α) was fine-tuned through cross-validation, settling at 100.0. This parameter effectively controls the sparsity of the model's coefficients, thereby enhancing the model's generalizability and avoiding overfitting. The model achieved an R^2 score of 0.873 and an RMSE of 27,854.35, indicating a relatively high degree of accuracy in predictions. Additionally, the model provides insightful interpretations of the data, highlighting the important features that influence house values.

5.2.2 Random Forest

For both feature sets, we utilized GridSearchCV combined with cross-validation to optimize the model parameters, and the results are as follows:

All Features: The optimal parameters were determined to be {*max_depth: 15, max_features: 0.5, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 300*}. This model configuration achieved an R^2 score of 0.903 and an RMSE of 24398.40, indicating an effective predictive performance with a substantial ability to capture the variance in house prices.

Selected Features: The same parameters as above were found to be optimal for this feature set. The model produced an R^2 score of 0.890 and an RMSE of 25930.91, showcasing a strong but slightly lesser performance compared to the full feature set.

The analysis reveals that the Random Forest model performs better when utilizing the full feature set, achieving higher accuracy and lower prediction error. This outcome suggests that the additional features, despite increasing model complexity, contribute significantly to capturing the dynamics of house pricing in the dataset. The results underscore the capability of Random Forest to handle a large number of features effectively, benefiting from the rich information to improve prediction accuracy.

5.2.3 Support Vector Machine (SVM)

For both feature sets, we utilized GridSearchCV combined with cross-validation to optimize the model parameters, and the results are as follows:

All Features: The best parameters found were {*C: 100, gamma: 'scale', kernel: 'linear'*}. With these parameters, the SVM model achieved an R^2 score of 0.550 and an RMSE of 52399.20. This performance indicates relatively low accuracy and high error.

Selected Features: The same parameters as above were found to be optimal for this feature set. This resulted in an R^2 score of 0.481 and an RMSE of 56310.40. The decrease in performance suggests that the model may rely on the comprehensive data provided by the full set of features to capture more complex patterns in the data.

The analysis of the SVM's performance highlights its sensitivity to feature selection, with a notable decline in performance when fewer features are used. Despite its robustness in handling linear separations, the SVM's performance falls short in comparison to other models.

5.2.4 Neural Network

Neural Networks were employed to explore the effectiveness of deep learning techniques in predicting house prices. To ensure efficient training dynamics, both the input features and the target variable were scaled to similar scales using the MinMaxScaler. The architecture for the neural network was designed to gradually reduce the dimensionality from the input layer to the output layer:

- **Input Layer**: Configured to match the number of features in the dataset.
- **Hidden Layers**: Comprising three layers with 128, 64, and 32 neurons respectively, each employing ReLU activation to introduce non-linearity and enhance learning.
- **Output Layer**: A single neuron output layer for regression, reflecting the continuous nature of the target.

The model was trained over 100 epochs using the Adam optimizer, with 'mean_squared_error' as the loss function.

All Features: Utilizing all available features, the model consisted of 37,761 trainable parameters and demonstrated relatively good predictive performance, achieving an R^2 score of 0.880 and an RMSE of 27,123.56.

Selected Features: In this case, the model included 21,761 trainable parameters and achieved an R^2 score of 0.868 and an RMSE of 28418.71.

The full-feature model slightly outperforms the latter demonstrating the neural network's ability to extract more nuanced patterns from a larger dataset, which can be crucial for complex regression tasks like house price prediction.

5.2.5 XGBoost

For both feature sets, we utilized GridSearchCV combined with cross-validation to optimize the model parameters, and the results are as follows:

All Features: Optimal parameters were identified as $\{$*colsample_bytree: 0.8, learning_rate: 0.05, max_depth: 3, n_estimators: 300, subsample: 1*$\}$. This setup yielded an R^2 score of 0.918 and an RMSE of 22,326.72, demonstrating the model's excellent predictive accuracy and its ability to explain a significant proportion of the variance in house prices.

Selected Features: The optimal parameters for the selected features were $\{$*colsample_bytree: 1, learning_rate: 0.1, max_depth: 3, n_estimators: 300, subsample: 0.8*$\}$. These settings resulted in an R^2 score of

0.911 and an RMSE of 23,276.64, indicating robust performance, albeit slightly lower than that achieved with the full set of features.

The comparative analysis indicates that XGBoost performs slightly better with the full set of features, benefiting from the broader range of data to enhance its predictive accuracy. This underscores XGBoost's capacity to effectively manage and utilize extensive datasets, optimizing the extraction of relevant patterns essential for accurate house price predictions.

5.2.6 CatBoost

Given the extensive dataset, we utilized CatBoost, which excels at processing raw datasets containing both numeric and categorical data, including those with missing values and unnormalized entries. To align with our comprehensive modeling strategy, we enhanced the raw dataset with the engineered features (see Table 3) such as *TotalBathrooms* and *AgeAtSale*.

We employed GridSearchCV and cross-validation and determined the optimal parameters to be {*depth: 6, iterations: 1000, l2_leaf_reg: 1, learning_rate: 0.1*}. The model achieved an R^2 score of 0.826 and an RMSE of 25,937.01. The integration of specific engineered features, along with CatBoost's adept handling of the raw data, showcases its utility in scenarios that require minimal data preprocessing but where strategic feature enhancement is feasible.

5.2.7 Ensemble Methods

Ensemble methods were employed to enhance the predictive accuracy by leveraging the combined strengths of multiple models. The base models used were Random Forest, XGBoost, and SVM, consisting of a good mix of high and low performance. Using all features, we explored three approaches to ensemble learning:

Averaging Predictions: This method aimed to reduce the variance and potential overfitting by averaging the predictive outcomes of the three base models. The R^2 score for this method was 0.863 with an RMSE of 28,927.59, providing a moderate baseline performance. The parameters used for each base model are summarized in Table 4 below.

Stacking Method: This technique involved using predictions from the base models as inputs to a final meta-model, a linear regressor, to compute the final prediction. The R^2 score improved to 0.916 and the RMSE decreased to 22,617.22, demonstrating enhanced predictive performance. The parameters used for each base model are summarized in Table 4 below.

Stacking Method with Hyperparameter Tuning: Extending the stacking approach, we applied hyperparameter tuning to both the base models and the meta-model to optimize performance. The fine-tuning involved adjusting parameters such as the learning rate for XGBoost and the regularization strength for SVM, among others. This calibration further refined the model's accuracy, leading to an R^2 score of 0.917 and an RMSE of 22,446.62, marking the highest performance among our ensemble techniques. The parameters used for each base model are summarized in Table 5 below.

Table 4: Parameters used in Averaging Predictions and Stacking Ensemble Methods

Model	Parameters
Random Forest	{*max_depth: 15, max_features: 0.5, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 300*}
XGBoost	{*colsample_bytree: 1, learning_rate: 0.05, max_depth: 3, n_estimators: 300, subsample: 0.8*}
SVM	{*C: 100, gamma: 'scale', kernel: 'linear'*}

Table 5: Parameters used in Stacking Ensemble Method with Hyperparameter Tuning

Model	Parameters
Random Forest	{*max_depth: 15, max_features: 0.5, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 300*}
XGBoost	{*colsample_bytree: 1, learning_rate: 0.01, max_depth: 3, n_estimators: 100, subsample: 0.7*}
SVM	{*C: 100, gamma: 'scale', kernel: 'linear'*}

These detailed ensemble strategies, particularly the stacking with hyperparameter tuning, illustrate the effectiveness of combining multiple modeling techniques. The incremental improvements with more sophisticated configurations highlight the potential of ensemble learning in tackling complex predictive tasks such as house price estimation.

5.3 Model Selection

The best performance metrics for each model, considering both all features and a subset of features, are summarized in Table 6 below. Lasso Regression, used as the baseline model, provided a benchmark for comparing the performance of more complex models.

Table 6: Summary of Best Results from Each Model

Model	Features Used	R^2	RMSE ($)
Lasso Regression	All	0.872	27854.35
Random Forest	All	0.891	25769.59
XGBoost	All	0.918	22326.72
SVM	All	0.550	52399.20
Neural Network	Lasso Selection	0.871	28105.04
CatBoost	All	0.826	25937.00
Ensemble Models (Stacking Method)	All	0.917	22446.62

Among the models evaluated, XGBoost demonstrated the highest R^2 score of 0.918 and an impressively low RMSE of 22,326.72. While the Ensemble Stacking Method also showed strong performance with an R^2 score close to that of XGBoost and a comparable RMSE, the relative simplicity and interpretability of XGBoost, coupled with its excellent predictive performance, made it the preferred choice. Additionally, XGBoost's ability to handle a wide range of data types and its robustness to overfitting solidify its suitability for providing reliable predictions and actionable insights.

6 Evaluation

6.1 Evaluating the Model Performance

The performance of the selected XGBoost model was rigorously evaluated to ensure its reliability and accuracy in predicting house prices. This evaluation utilized two primary methods:

1. **Actual vs. Predicted Values Plot**: This plot (see Figure 3), compares the actual house prices against the predictions made by the model. A closer alignment of the points along the diagonal line indicates better model accuracy. The plot shows a high concentration of points along the diagonal, suggesting that the model predictions are generally close to the actual values, thus confirming the model's effectiveness.

2. **Residuals Plot**: The residuals plot (see Figure 4), provides insights into the errors between the predicted and actual values. Ideally, residuals should be randomly distributed around the horizontal axis, indicating that the model does not systematically overestimate or underestimate the house prices. The plot displays a relatively even spread of residuals around zero, which supports the model's unbiased nature and its capability to handle various price ranges effectively. Additionally, the absence of patterns or systematic deviations in the residuals plot can be indicative of minimal overfitting, as it suggests that the model does not merely memorize the training data but generalizes well to new, unseen data.

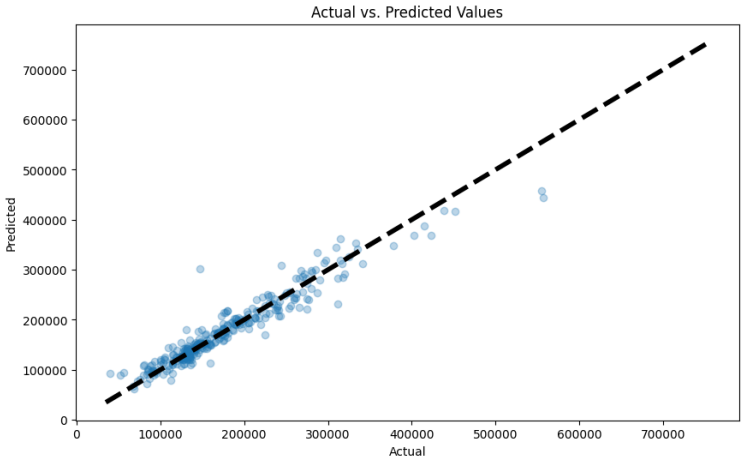

Figure 3: Actual vs. Predicted Values Plot

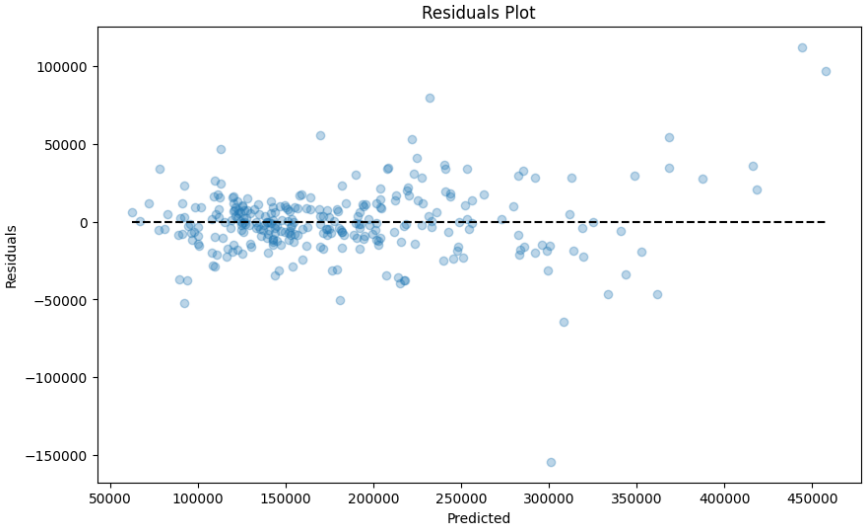

Figure 4: Residuals Plot

6.2 Identifying Key Features

We drew valuable insights into the factors most affecting house prices through XGBoost feature importance analysis and SHAP values analysis.

Feature Importance Plot: This chart (see Figure 5), ranks features based on their importance determined by the model. Notably, *GrLivArea*, *OverallQual*, and *TotalBsmtSF* are among the top features, indicating that the size of the living area, overall material and finish quality, and basement area size are critical determinants of house pricing.

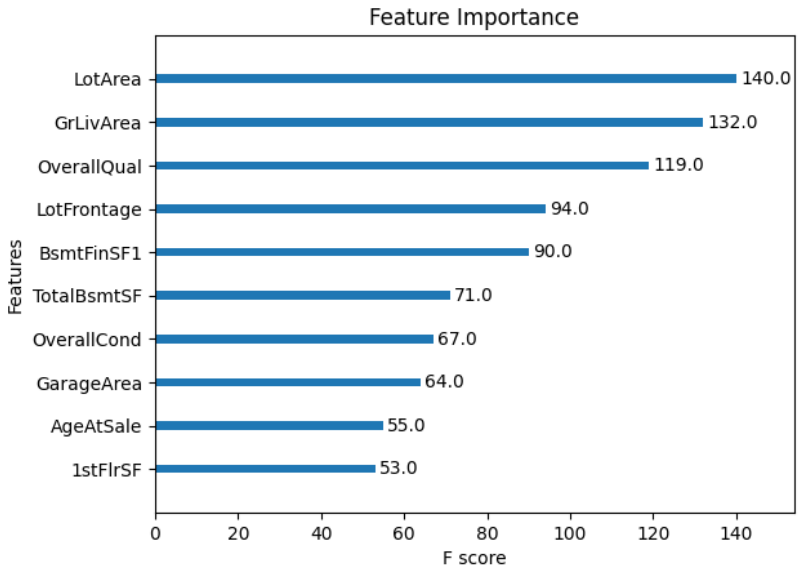

Figure 5: XGBoost Feature Importance Plot

SHAP Values Analysis:

SHAP Summary Plot: This plot (see Figure 6), uses SHAP values to show the impact of each feature on the model output. The color represents the feature value (red high, blue low). This analysis reveals that higher values of overall quality (*OverallQual*) and ground living area (*GrLivArea*) have a strong positive effect on the house prices, while features like basement quality (*BsmtQual*) and porch area (*TotalPorchArea*) also play significant roles but vary more in their impacts.

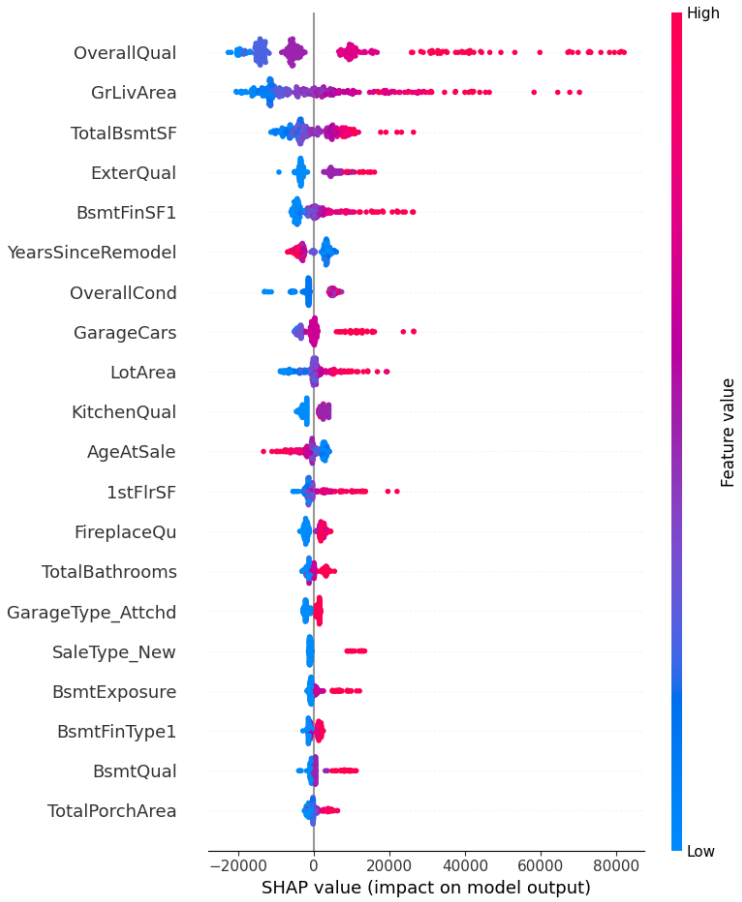

Figure 6: Shap Summary Plot

SHAP Explainer - Neighborhoods: The neighborhood analysis (see Figure 7), reveals the varying impact of location on house prices. Neighborhoods like Crawford and Brookside significantly increase house prices due to their desirable characteristics, while other areas like Stone Brook may detract from value, highlighting the critical role of location in real estate valuation.

SHAP Explainer - Dwelling Types: This analysis (see Figure 8), helps understand how different types of dwellings influence house prices. It indicates that newer single-story homes and larger, older two-story homes have significant impacts on pricing, reflecting market preferences for certain styles and historical qualities in homes.

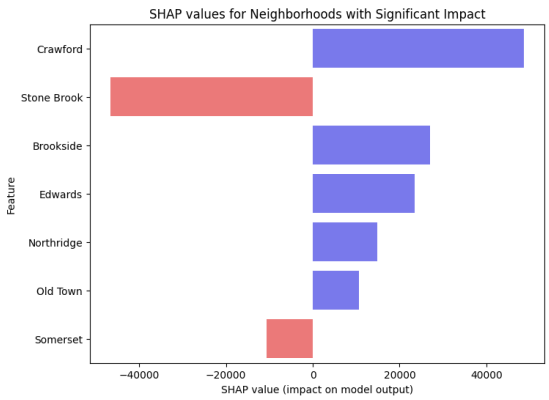

Figure 7: SHAP Explainer for Neighborhoods

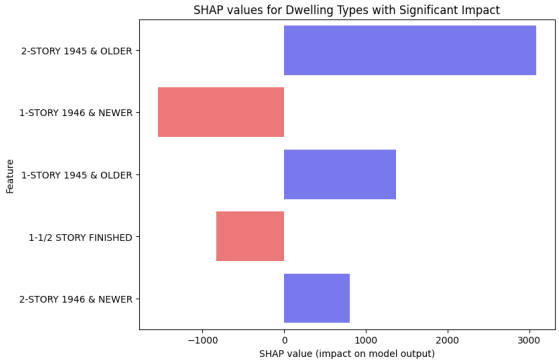

Figure 8: SHAP Explainer for Dwelling Types

Detailed SHAP Interaction Plots: These plots (see Figure 9 and Figure 10), further dissect the relationships between features and their SHAP values, illustrating how interactions between features like years since remodel (*YearsSinceRemodel*) and overall quality (*OverallQual*) or total number of bathrooms (*TotalBathrooms*) influence pricing predictions. Such insights are crucial for stakeholders focusing on renovations or real estate development as they highlight the specific attributes that could enhance property value.

In Figure 9, we deduce that factors such as the number of years since the most recent remodeling, number of bathrooms in the house and the total basmeent surface area have a positive effect on the perceived overall quality of the house, which in turn drives up the price of the house.

In Figure 10, we can see that factors such as overall condition (*OverallCond*), ground living area (*GrLivArea*) and total number of bathrooms (*TotalBathrooms*) are positively impacted by recent remodeling. The interplay of all these different features affect the price (*SalePrice*) of the house and are factors that different stakeholders are advised to keep in mind when making decisions.

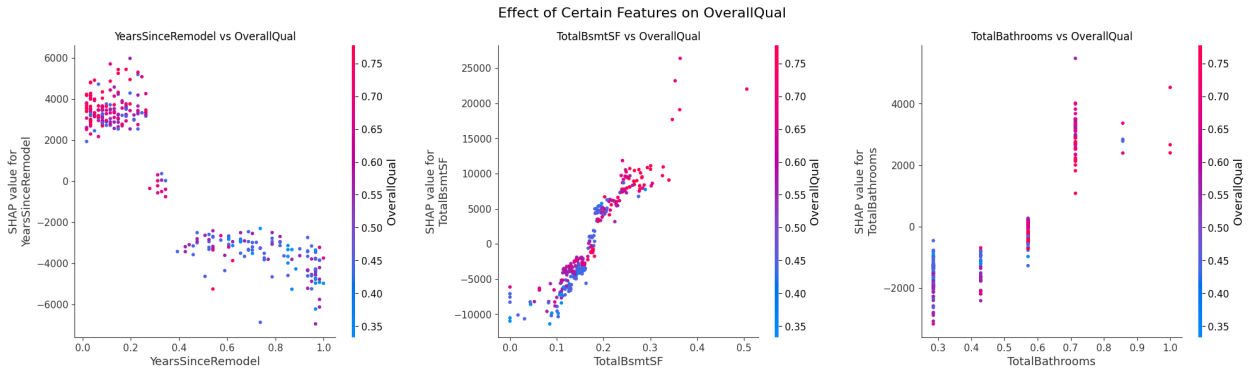

Figure 9: Shap Interaction Plots for OverallQual

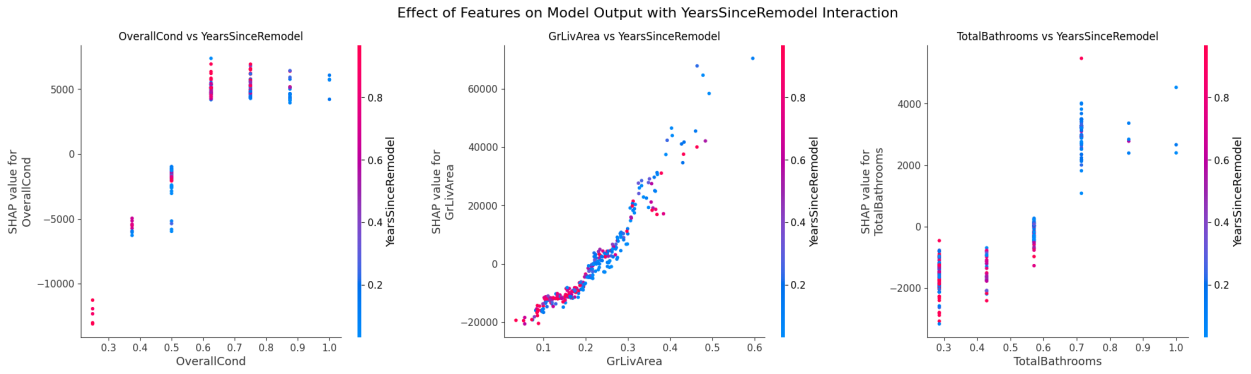

Figure 10: Shap Interaction Plots for YearsSinceRemodel

The detailed feature analysis provided by the XGBoost model not only aids in understanding the direct factors influencing house prices but also assists in strategic decision-making for property investments and enhancements. This level of insight supports the model's selection for deployment in real-world scenarios, where nuanced understanding of market dynamics is essential.

7 Recommendations and Impact

Based on the insights derived from our predictive model, we have formulated targeted recommendations for different stakeholders within the real estate market. These recommendations are aimed at leveraging the model's findings to optimize investment strategies, improve property values, and assist buyers and sellers in making informed decisions.

Real Estate Agents:

- **Evaluate House Prices with Given Features**: Agents can use the predictive model to accurately assess house prices based on specific property features, enhancing their ability to advise clients on fair pricing and investment returns.
- **Guide Premium and Budget Buyers**: Premium buyers should be directed towards the Crawford neighborhood due to its higher property values, while budget buyers might find more affordable options in the Stone Brook neighborhood, aligning purchase decisions with financial strategies and lifestyle preferences.

Real Estate Developers:

- **Design Two-Story Dwellings**: Developers should focus on constructing two-story dwellings, which are shown to be highly valued in the market, especially in premium neighborhoods.
- **Prioritize Living Space Quality and Garage Size**: Emphasize the quality of living spaces and the size of garages in new developments, as these features significantly influence house valuations.

Home Owners:

- **Assess House Value**: Owners should periodically assess their property's market value using the predictive model, considering current market trends and property conditions.
- **Evaluate Targeted Renovations to Increase Value**: Repaint and remodel interior and exterior finish as enhancements to the interior and exterior finish can significantly boost property appeal and value. Ensure basement is finished as completing or renovating the basement can add substantial usable space and increase the property's overall value. Remodel kitchen as modern and well-equipped kitchens are crucial for increasing home valuation, reflecting current buyer preferences for high-quality amenities.

Home Buyers:

- **Assess House Value Based on Features and Location**: Buyers should evaluate potential homes based on essential features and location to ensure investment in properties that meet their needs and financial considerations.
- **Assess Possible Features and Locations Given a Budget**: Understand what features and neighborhoods are feasible within a set budget to maximize the value and satisfaction of their investment. Buyers should explore neighborhoods like Stone Brook for affordability and consider features typically available within their budget, such as smaller lot sizes or older homes needing renovation. Premium neighborhoods like Crawford offer higher-end features like modern kitchens and spacious garages but at a higher cost.

These recommendations are designed to provide actionable advice tailored to the specific needs and roles of each stakeholder group, enabling them to make strategic decisions that align with market dynamics and personal or business goals. By applying the insights gained from the predictive model, stakeholders can enhance their competitive edge, achieve better financial outcomes, and realize optimal property values in the real estate market.

8 Extending Scope to Other Cities

To broaden the applicability and increase the business value of our predictive model, we utilized XGBoost, our best-performing model, to estimate house prices in diverse markets such as New York, California, and Seattle. This expansion aimed to test the model's adaptability and effectiveness across varying market conditions, thereby enhancing its relevance for stakeholders operating in different geographic regions. We obtained and cleaned relevant housing datasets for these cities and applied the XGBoost model to predict house prices. The summarized results of these efforts are in Table 7 below:

The results reveal that the model's performance varied significantly depending on the quality and scope of the available datasets. The R^2 scores and RMSE values indicate that the model struggled to capture market nuances as effectively as it did in Ames, Iowa. These performance differences underscore the importance of securing adequate and reliable input features that reflect the unique market conditions of each region.

Table 7: Performance Metrics of Predictive Models for Different Cities

City/State	R^2	RMSE ($)	Important Features Observed
New York	0.589	3,218,761.26	Property area, Zip code, Number of bathrooms
California	0.707	61,945.62	Median household income, Distance to ocean, Number of bedrooms
Seattle	0.577	344,297.69	Apartment size, Zip code, Number of bathrooms

9 Future Work

As the predictive model evolves, expanding its capabilities and enhancing its accuracy are pivotal for capturing real-world market dynamics more effectively. The application of the model across different cities could benefit significantly from the inclusion of more comprehensive datasets. By analyzing diverse markets with a variety of data inputs, the adaptability and predictive performance of the model could be refined to better suit the nuances of each location. Integrating local economic indicators such as employment rates and income levels could also enhance the model's ability to respond to economic shifts. These metrics can provide deeper insights into how market fluctuations influence house prices, offering a more robust tool for stakeholders in the real estate sector.

Consumer behavior is another area where deeper insights could improve model accuracy. Utilizing surveys and behavioral studies to capture shifts in buyer preferences and trends could help in anticipating future market demands and adjusting housing price predictions accordingly. Finally, incorporating real-time market data could significantly increase the model's responsiveness and accuracy. Access to up-to-date information on listings, sales, and market fluctuations could transform the model into a more dynamic and immediately applicable tool in real estate transactions. These enhancements could ensure that the predictive model remains a vital asset for real estate agents, developers, and consumers, keeping pace with market changes and consumer trends.

10 Conclusion

This project successfully developed a predictive model using XGBoost to estimate residential property prices in Ames, Iowa, demonstrating substantial accuracy through high R^2 scores and low RMSE values. Recommendations based on the model's insights provide actionable strategies for stakeholders in the real estate market, aiding them in their decision-making processes. Future enhancements focusing on integrating local economic data and real-time market information are anticipated to further improve the model's utility, making it an indispensable tool in real estate valuation. The project sets a foundation for advanced applications of machine learning in real estate, offering pathways for both continued research and practical implementation in market analysis.

11 Appendix

11.1 Group Contribution

We would like to emphasize that the success of this project is attributed to the equal and significant contributions of all group members. Each phase, including coding, data analysis, result generation, and the drafting of the presentation and final report, was collaboratively completed.

11.2 Data Source and Code Repository

Datasets:

- Ames: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data
- New York: https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market
- California: https://www.kaggle.com/datasets/shibumohapatra/house-price
- Seattle: https://www.kaggle.com/datasets/samuelcortinhas/house-price-prediction-seattle

Source code: https://github.com/manisha-goyal/housing-value-predictor.

Please get in touch should you run into issues with the datasets or source code.