ACKNOWLEDGEMENT

HOUSE PRICE PREDICTION IN AMES, IOWA: Comparing Lasso, Ridge, Random Forest, Gradient Boosting, and XGBoost Models

Prepared by: Group 22

Name	Student ID	Deliverables Assigned	Contribution
Quoc Huy Vu	7848298	Data Preprocessing	25%
Thanh Dat Nguyen	8699057	Model Implementation and Evalutaion	25%
Nguyet Tuyen Hoang	7719188	Introduction, Background Theory, Strengths and Limitations	25%
Zobair Yousuf Abdullah	7830361	Solutions, Discussions, Conclusion	25%

Subject Coordinator:

- Professor Lei Wang
- Dr Hui Luo

Resources and Tools: We acknowledge the use of the following resources and tools in completing this assignment:

1. Dataset: Ames Housing dataset provided for this assignment
2. Software and Libraries: Python, scikit-learn, pandas, matplotlib, seaborn
3. Development Environment: Jupyter Notebook, PyCharm
4. Generative AI: ChatGPT was used for refining report content, word usage, and structure.
5. Academic Resources: All academic papers and resources referenced in this report are properly cited in the References section using Harvard UOW citation style

Date of Submission: 30 May 2025

TABLE OF CONTENT

1. INTRODUCTION

1.1 Significance of House Price Prediction

House price prediction has been researched as it has significant effects on various stakeholders in the real estate industry. Accurate prediction models can help buyers in making smart decisions, sellers in estimating prices, and real estate agents in providing accurate valuations (Gao et al., 2019). As property markets keep changing and prices in different areas keep going up and down, it has become more and more crucial to be able to accurately estimate house prices.

1.2 Economic Impact of Real Estate Market

The real estate market is one of the biggest parts of the economy, and millions of people are affected by property transactions every year. In the past 10 years in the United States alone, house sales have gone up by 34%, and they reached a record high of 5.51 million last year (Gao et al., 2019). This significant market size shows how important it is to develop accurate prediction models that are about to account for many elements affecting house prices.

1.3 Challenges in House Price Prediction

Predicting house prices is hard because every property is different and many things affect them, such as their physical characteristics (size, age, condition), their location (neighbourhood quality, proximity to amenities), and the state of the market (interest rates, economic indicators). In the past, people have used basic regression models that don't show how these variables are related in a complicated way (Park and Bae, 2015). Traditional methods strongly relied on simple regression models that can’t capture the complex relationships between features (Park and Bae, 2015).

1.4 Dataset Overview

The Ames Housing dataset, used in this study, contains information on residential properties in Ames, Iowa. The dataset includes 2,930 observations and 82 features (79 common features, 2 identifiers and the tartget variable SalePrice) covering various aspects of residential homes, including:

- Property characteristics (lot size, square footage, number of rooms)
- Quality and condition ratings
- Location information
- Sale conditions and prices

Initial exploration of the dataset revealed several key insights:

1. The sale prices range from $34,900 to $755,000, with a mean of approximately $180,921
2. Several features show strong correlation with sale price, particularly overall quality, living area, and garage size
3. The dataset contains both numerical and categorical variables, with some missing values requiring appropriate handling

1.5 Report Overview

This report explores advanced machine learning approaches for house price prediction using the Ames Housing dataset. We analyze various modeling techniques, evaluate their performance, and identify the most influential factors in determining house prices. The report is structured as follows: Section 2 provides theoretical background on house price prediction methods; Section 3 details our methodology including data preprocessing, model implementation, evaluation and discussions.

2. BACKGROUND THEORY

2.1 Traditional House Price Prediction Methods

House price prediction has traditionally relied on hedonic price models, which decompose a property's price into the sum of contributions from its various attributes (Bourassa et al., 2010). The models usually implement multiple linear regression to estimate the relationships between features and prices from the housing dataset. Hedonic models are easy to use and understand, but they typically don't show how variables interact in complicated ways or how they relate to each other in a non-linear way. The hedonic price model, which comes from economics perspective, is the most common model used in traditional methods for predicting house prices and has been the subject of a lot of research (Gao et al., 2019).

2.2 Machine Learning Implementation

Machine learning has been used to estimate housing prices in the last few years. These methods are better than older ones in a number of ways. For example, they can model non-linear relationships, work with high-dimensional data, and uncover interactions between traits on their own (Pow et al., 2014).

Global and local modelling are the two main types of machine learning algorithms that can be used to anticipate property prices. Global models employ one prediction model on the complete dataset, while local models partition the data into smaller groups based on location or feature and make distinct models for each group (Gao et al., 2019).

The change from Single Task Learning (STL) to Multi-Task Learning (MTL) frameworks is a huge step forward in being able to guess how much a house will cost. In STL, each prediction task is distinct. In MTL, the tasks are connected in some way to improve overall performance. Gao et al. (2019) showed that MTL approaches work far better than traditional ones, especially when it comes to simulating how prices change based on where you are.

2.3 Justification of Chosen Models

In our project, we implemented 5 ML models that have been proven by previous research for house price prediction:

- **Ridge Regression:** Ridge regression is a popular linear model that helps prevent overfitting by adding a penalty to large coefficients (using L2 regularization). Instead of shrinking coefficients to zero like Lasso, Ridge spreads the penalty across all features, which is useful when you believe many features might each contribute a little to the prediction. In house price prediction, Ridge is a solid choice when you want a simple, stable model that can handle situations where several features are correlated, without dropping any features entirely.
- **Lasso Regression:** Lasso (Least Absolute Shrinkage and Selection Operator) regression performs both variable selection and regularization, making it ideal for datasets with many features. By applying L1 regularization, Lasso forces the coefficients of less important features to zero, effectively performing feature selection while training (Tibshirani, 1996). This characteristic is particularly valuable in house price prediction where numerous features may be present but only a subset significantly impacts the price, helping to create more interpretable and potentially more generalizable models.

- **Random Forest:** Decision trees divide the features into regions and assign each area a fixed value. To improve performance metrics like accuracy and model robustness, this ensemble method combines multiple decision trees to create a more powerful predictive model (Park and Bae, 2015). Random Forest can capture non-linear relationships, complex patterns, and feature interactions without requiring explicit specification.
- **Gradient Boosting:** This sequential ensemble method builds trees that correct errors made by previous trees. It has shown excellent performance in regression tasks by iteratively improving predictions through the addition of new models that focus on previously misclassified instances (Friedman, 2001). Gradient Boosting is particularly effective for house price prediction due to its ability to handle heterogeneous data and capture complex feature interactions.
- **XGBoost (XGB):** XGBoost is an advanced implementation of gradient boosting that uses a more regularized model formalization to control overfitting. It has demonstrated superior performance in various prediction tasks, including house price prediction, due to its ability to handle missing values, implement regularization, and process large datasets efficiently (Chen and Guestrin, 2016). XGBoost's tree pruning and parallel processing capabilities make it particularly suitable for the complex feature interactions present in housing data.

These models were selected were selected so that we can capture both linear and non-linear patterns. Together, they represent a comprehensive approach to house price prediction that balances interpretability, performance, and generalizability.

3. SOLUTIONS, IMPLEMENTATION, EVALUATION AND DISCUSSION

3.1 Design and Methodology

To tackle the Ames Housing price prediction task, we structured the project around a systematic workflow designed to ensure both data quality and model reliability:

- **Initial Data Exploration:** We began by loading the Ames Housing dataset and exploring its characteristics to identify patterns, unusual values, and data types.
- **Data Cleaning and Preprocessing:** Based on initial observations, we developed strategies for handling outliers and missing data.
- **Feature Engineering:** We transformed categorical variables and created new features to improve model performance.
- **Exploratory Data Analysis (EDA):** We conducted in-depth analysis to understand feature distributions and relationships with the target variable.
- **Model Development and Evaluation:** We implemented and compared multiple regression models using cross-validation and hold-out testing.

This structured approach ensured that all modeling decisions were evidence-based and grounded in thorough data understanding.

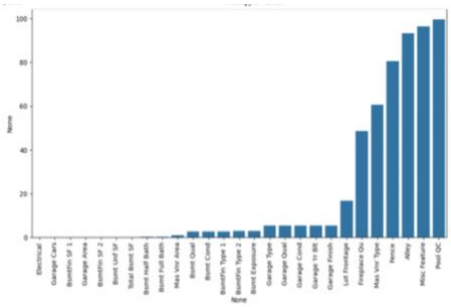
3.2 Data Preprocessing

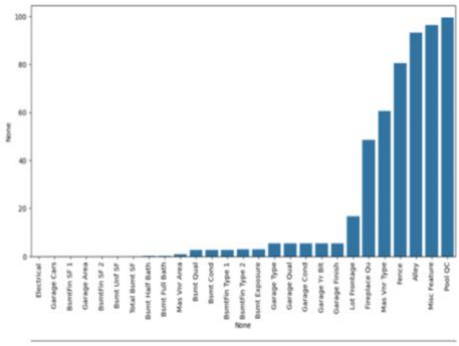
Our preprocessing workflow addressed several data quality challenges:

Outlier Detection and Removal: We identified outliers using scatterplots, particularly in the Gr Liv Area feature. Several extremely large houses with unexpectedly low sale prices were removed to prevent them from skewing our analysis.

Missing Value Treatment: We analyzed missing data across all columns and applied context-appropriate strategies:

- Features with high proportions of missing values and low predictive importance (Fence, Alley, Misc Feature, Pool QC) were dropped
- For numerical features, missing values were imputed using zeros where appropriate (e.g., basement features) or median values
- For categorical variables, missing values were filled with 'None' or group-specific means (e.g., Lot Frontage by neighborhood)

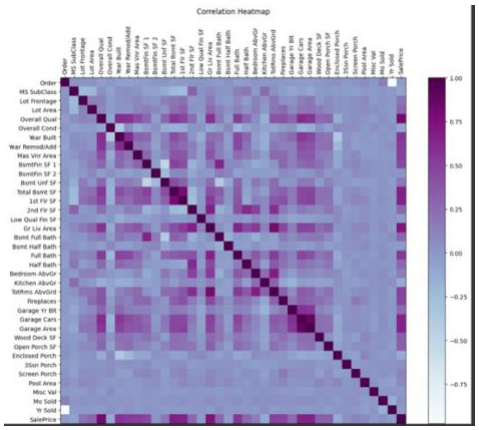




3.3 Exploratory Data Analysis (EDA)

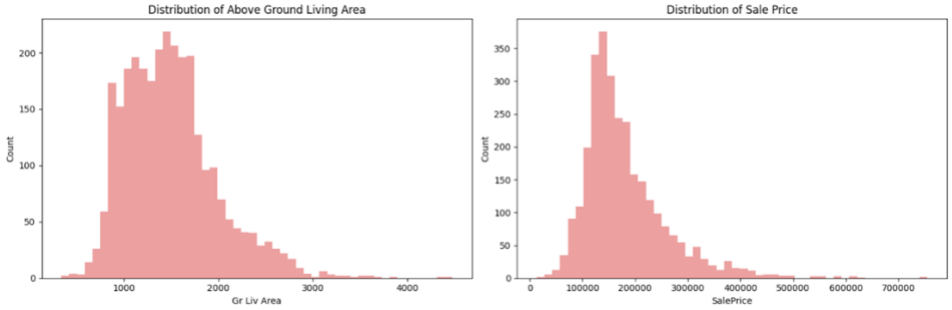
Our EDA process revealed critical insights about the dataset:

- **Correlation Analysis:** We started with a correlation heatmap covering all the numerical features, including SalePrice. This provided a quick overview of how the different features relate to each other. It was immediately clear that some variables such as Overall Quality, Gr Liv Area, Garage Area, and Total Basement SF have strong positive correlations with SalePrice, while others are only weakly linked.

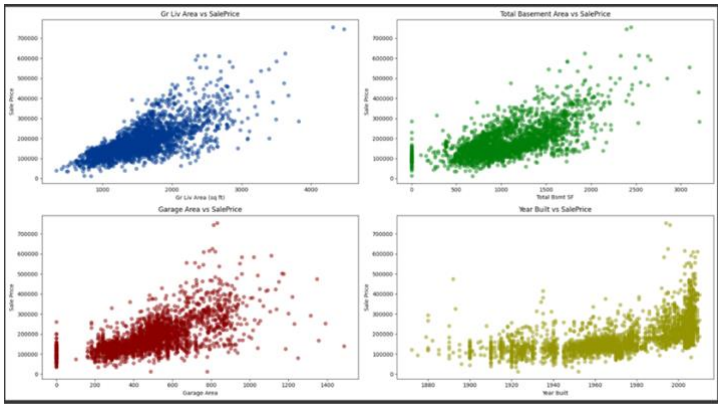


- **Distributions of Key Features**

Next, we checked out the distributions of some of the most important variables. Histograms for both above ground living area (Gr Liv Area) and SalePrice revealed a right-skewed pattern: most homes fall within a middle range, with just a handful of properties being very large or expensive. This pattern is pretty common in real estate datasets and helps explain why outlier handling is so crucial.

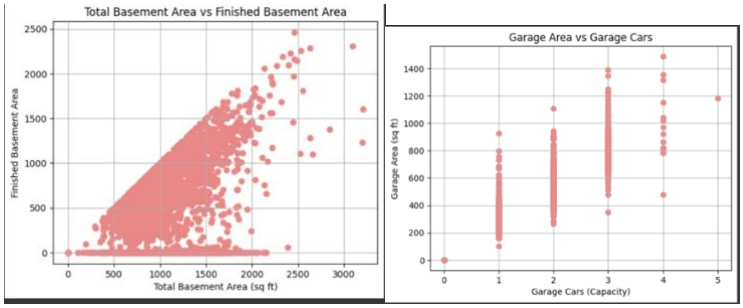


To dig deeper, we made scatterplots comparing SalePrice with several top features: Gr Liv Area, Total Basement SF, Garage Area, and Year Built. These plots confirmed what we suspected from the correlations. bigger houses, newer builds, and properties with larger garages or basements tend to sell for more. The scatterplot for Year Built was especially interesting, showing a jump in sale prices for newer homes.

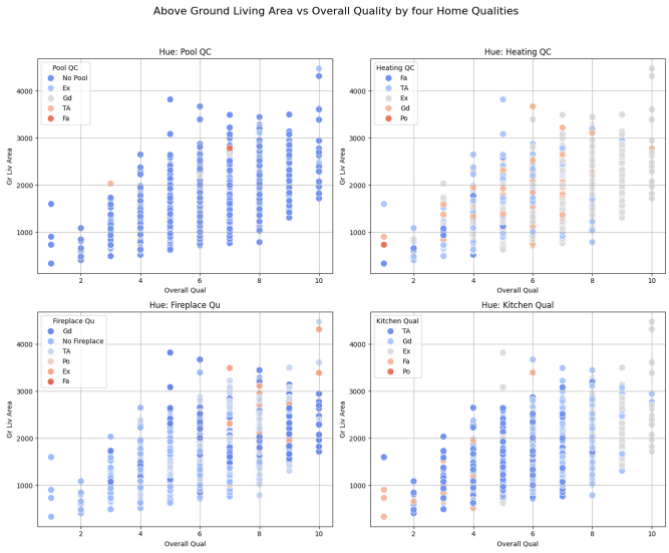


- **Feature Redundancy and Multicollinearity**

We also looked for features that might be telling us the same thing. For example, a scatterplot of Garage Area against Garage Cars showed distinct vertical groupings, which makes sense most garages are built to fit a certain number of cars. Similarly, comparing Total Basement Area to Finished Basement Area showed that a lot of homes have unfinished basements, which could be useful for engineering new features or grouping data**.**



Finally, we explored how different quality-related features (like Pool QC, Heating QC, Fireplace Qu, and Kitchen Qual) interact with core predictors such as Overall Quality and Gr Liv Area. By coloring scatterplots based on these categorical ratings, we could see patterns in how combinations of features impact house size and (indirectly) price. For instance, homes with higher overall quality often also have higher ratings for kitchens and fireplaces, but this isn't always the case.



These findings aligned with previous research (Gao et al., 2019; Park and Bae, 2015) emphasizing the importance of both physical characteristics and location-based factors in house price prediction.

3.4 Feature Engineering

We implemented several feature engineering techniques to prepare the data for modeling:

- **Categorical Variable Encoding:** Categorical variables were converted to string type and transformed through one-hot encoding to make them compatible with regression algorithms.
- **Feature Transformation:** The right-skewed SalePrice distribution was normalized using logarithmic transformation. Other numerical features with skewed distributions were similarly transformed using appropriate methods (log, square root, or Box-Cox).
- **Feature Selection:** We used correlation analysis and feature importance techniques to identify the most relevant predictors while addressing multicollinearity by removing redundant features.

3.5 Model Training and Evaluation

After preparing the dataset, we split the data into training and test sets. We then trained five different regression models: **Lasso Regression, Ridge Regression, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor**. We fine-tune Gradient Boosting and XGBoost for better performance. We evaluated model performance using widely used metrics:

- **Root Mean Squared Error (RMSE):** Measuring the square root of average squared prediction errors
- **Mean Absolute Error (MAE):** Measuring average absolute prediction errors
- **R-squared (R^2):** Indicating the proportion of variance explained by the model

Models	**RMSE**	**MAE**	**R^2**
XGBRegressor	20693.11	13427.79	0.9469
GradientBoostingRegressor	19622.18	13088.39	0.9522
RandomForestRegressor	24894.69	15387.27	0.9231
Lasso Regression	23236.01	15028.58	0.9330
Ridge Regression	22969.30	14889.27	0.9345

3.6 Discussion

Model Evaluation and Analysis

Based on the model comparison metrics above, we can observe 6 key insights:

- **Gradient Boosting really shines:**
  Out of all the models we tried, Gradient Boosting Regressor delivered the best results by a noticeable margin. It had the lowest RMSE and MAE, and the highest R^2 score, meaning it was the most accurate at predicting house prices. The reason for this strong performance is likely because Gradient Boosting builds a series of trees that keep learning from the mistakes of the previous ones—so it gets really good at catching complex trends in the data.
- **All the models did pretty well:**
  All models achieved high R^2 values, above 0.92 (with Random Forest fluctuate a bit around that) That means each one could explain more than 92% of the variation in house prices—a pretty solid outcome.
- **Ensemble models have an edge:**
  The "ensemble" methods, like Gradient Boosting and XGBoost, generally did a better job than the plain linear models (Ridge and Lasso). This makes sense because house prices depend on lots of interacting factors, and ensemble models are great at capturing those kinds of complex, non-linear relationships.
- **Random Forest was good, but not the best:**
  Random Forest gave us solid predictions, but it didn't match up to the other models. This was to be expected, since we did not implement any regulations to it.
- **XGBoost vs. Gradient Boosting :**
  We expected XGBoost to come out on top, since it's usually known for its advanced features and strong regularization. But in this case, standard Gradient Boosting did better. This implies the fine-tune parameters we use for XGBoost hyperparameters weren't fully optimized.

- **How far off were our predictions, on average?**
  Across the board, the mean absolute errors (MAE) for our models were usually between $13,000 and $15,000. Considering how much homes can vary in price, that's a pretty reasonable margin of error. It means our models are generally making predictions that would be useful in practice.

Strengths and Limitations

The implemented approach demonstrates several strengths:

- The ensemble methods effectively capture non-linear relationships and feature interactions without requiring explicit specification
- The feature engineering process successfully addresses data quality issues and improves model performance
- The models provide reasonable accuracy in predicting house prices across different price ranges

However, certain limitations should be acknowledged:

- We did not implement cross-validation so it does not show how stable our models are.
- Some potentially relevant external factors (e.g., economic indicators, interest rates) are not included in the dataset

4. CONCLUSION

This study has demonstrated the effectiveness of machine learning approaches for house price prediction, with ensemble methods showing particularly strong performance. The results highlight the importance of both property characteristics and location-based factors in determining house prices, consistent with findings from previous research.

The superior performance of Gradient Boosting suggests that capturing complex non-linear relationships and feature interactions is crucial for accurate house price prediction. The feature importance analysis provides valuable insights for stakeholders in the real estate market, identifying key factors that drive property values.

Several limitations of the current approach present opportunities for future research. Incorporating temporal dynamics and external economic indicators could enhance prediction accuracy, particularly in volatile markets. Additionally, exploring deep learning approaches and geospatial modeling techniques may further improve performance, especially for location-centered prediction tasks.

In practical terms, the findings from this study can assist various stakeholders in the real estate market. Buyers can make more informed decisions by understanding the key factors influencing house prices, sellers can set more competitive prices based on accurate valuations, and real estate agents can provide more reliable guidance to their clients.

REFERENCES

Bourassa, SC, Cantoni, E & Hoesli, M 2010, 'Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods', *The Journal of real estate research*, vol. 32, no. 2, pp. 139–160.

Chen, J-H, Ong, CF, Zheng, L & Hsu, S-C 2017, 'Forecasting spatial dynamics of the housing market using Support Vector Machine', *International journal of strategic property management*, vol. 21, no. 3, pp. 273–283.

Gao, G, Bao, Z, Cao, J, Qin, AK & Sellis, T 2022, 'Location-Centered House Price Prediction: A Multi-Task Learning Approach', *ACM transactions on intelligent systems and technology*, vol. 13, no. 2, pp. 1–25.

Limsombunchai, V 2004, 'House Price Prediction: Hedonic Price Model vs. Artificial Neural Network', *IDEAS Working Paper Series from RePEc*.

Park, B & Bae, JK 2015, 'Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data', *Expert systems with applications*, vol. 42, no. 6, pp. 2928–2934.

Pow, N, Janulewicz, E & Liu, L 2014, 'Applied machine learning project 4 prediction of real estate property prices in Montreal', *Course Project Report*, McGill University.