Image Source: Unsplash (Jakub Zerdzicki)

# Predicting Housing Sale Prices

**03.03.2025**

—

Andrea Pena
Capstone 2: Project Report
Springboard

## Introduction

Predicting house prices is a crucial task in real estate, impacting homeowners, buyers, and investors. By identifying key factors influencing sale prices, this project provides valuable insights to guide better decision-making.

With a growing U.S. population—**346.6 million in 2024, up by 8.9 million in five years**—the demand for homeownership continues to rise. Understanding the drivers of property value can help homeowners maximize their sale prices, assist buyers in making informed purchases, and enable investors to identify profitable opportunities.

This project aims to **develop a predictive model for house prices by analyzing key property features** and evaluating their impact on sale price. By leveraging data-driven insights, we can refine real estate valuation strategies for the next five years.

## Data Overview

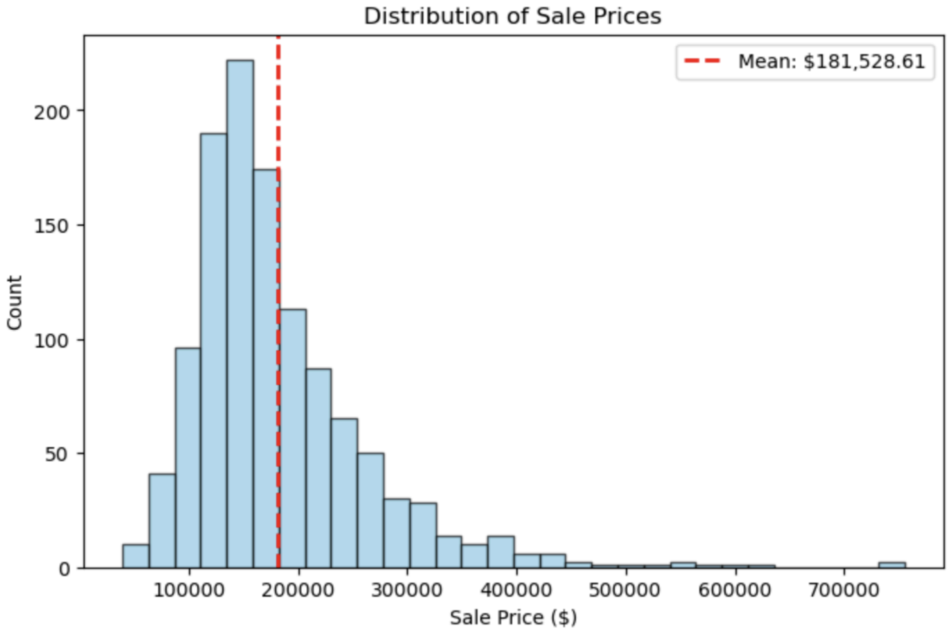
- **Dataset Source:** Kaggle (Gusthema's House Price Predictions using TensorFlow Decision Forests)
- **Rows/columns:** There are 1460 rows and 80 columns in the original dataset.
- **Data Cleaning:**
    - **Missing Data:** Five columns had over 50% missing values , while numerical and categorical missing values were imputed with the **median** and **mode**, respectively. Missing values were not dropped due to the small dataset.
    - **Duplicates:** None found.
- **Summary statistics:** Features such as Lot Area, Ground Living Area, Total Basement Size, and First Floor Size showed high variance, suggesting a strong potential as predictive features.

## Methodology

- **Exploratory Data Analysis:**
  - Created **correlation heatmaps, histograms, scatterplots, and barplots** to explore feature relationships.
  - Used **Pearson's correlation coefficient** to measure relationships between numeric variables and sale price.
    - Strong correlations were found with **Total Basement Size, First Floor Size, Ground Living Area, Full Bath, Total Rooms Above Garage, Garage Cars, and Garage Area**.

1. **Histograms**:
   - Features like **Total Rooms Above Garage, Garage Area, and Overall Quality** followed a roughly normal distribution.
   - **Ground Living Area, First Floor Size, and Total Basement Size** were right-skewed, suggesting the need for potential transformations.
   - **Figure 1**(below) has a right-skewed distribution, which indicates a small number of **high-value homes**, pulling the mean higher than the majority of sales.

**Figure 1: Sale Prices Histogram**



2. **Scatterplots**: Confirmed positive linear relationships for key numeric features.
3. **Barplots**: Categorical features showed minimal impact on Sale Price.

- **Preprocessing:**
  - **Categorical Encoding:** Applied one-hot encoding for categorical variables.
  - **Data Splitting:** Split into train (80%) and test (20%) sets.
  - **Feature Scaling:** Used StandardScaler to normalize numeric features.
- **Model selection:**
  - **Baseline Models:**
    - Started with **OLS, Lasso, and Ridge** regression for benchmark performance.
  - **Nonlinear Models:**
    - Tested **Random Forest, Gradient Boosting, and Support Vector Regression** to capture complex interactions.
- **Hyperparameter Tuning:**
  - Optimized **Lasso, Random Forest, and Gradient Boosting** using **GridSearchCV**, refining parameters to reduce overfitting and improve predictive accuracy.

## Model Performance & Comparison

- Linear regression models were trained and tested, with **Lasso** achieving the best performance among them, yielding an RMSE of **34,262.05**. However, **Random Forest**, a nonlinear model, outperformed Lasso with a lower RMSE of **27,245.97**.
- The Random Forest model underwent hyperparameter tuning, which resulted in a lower test score (**27,173.45**), but a significant gap remained between the train (**11,054.71**) and test scores, indicating potential **overfitting**. After modifying the hyperparameters to reduce overfitting, the test score increased slightly to **28,653.73**, but the gap between the train (**26,085.00**) and test scores narrowed. This trade-off suggests improved generalization.
- To further improve performance, two additional nonlinear models, **Gradient Boosting** and **Support Vector Regressor**, were trained and tested. Gradient

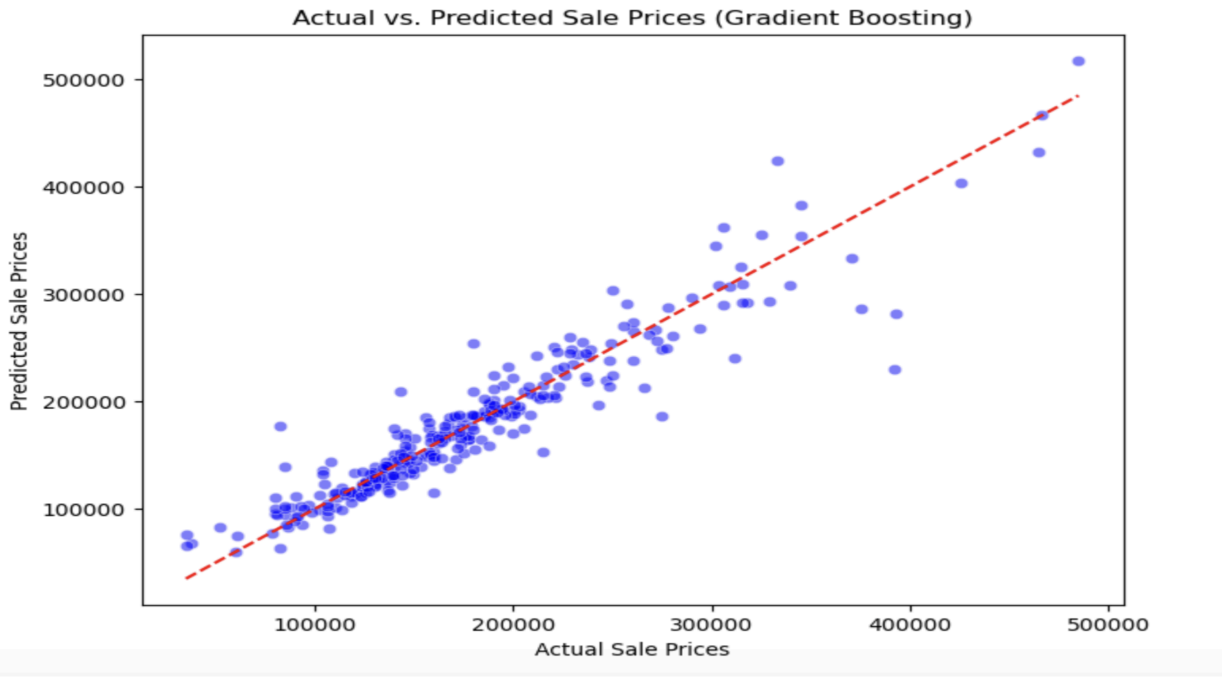
Boosting achieved the best performance among all models, with an RMSE of **25,844.64**.

- Following **hyperparameter tuning**, the Gradient Boosting model improved further, with train and test scores of **8,782.17**and **24,327.50**, respectively.
- Like RF, there was hyperparameter tuning for **GB**, two more times with **17,055.26** and **25, 405.89** then **20,307.08** and **25,492.86**, respectively. Each tuning shortened the gap with a slight increase in the test scores.
- The last evaluation had the **removal of outliers** to see if there was a change in test score, but it did increase(26,930.21) so the prior score is the best.

**Figure 2: Model Comparison Table**

| **Model** | **Train** | **Test** |
|---|---|---|
| Lasso Regression | -- | 34, 262.05 |
| Random Forest | -- | 27,245.97 |
| Tuned RF | 11,054.71 | 27,173.45 |
| Tuned RF | 26,085.00 | 28,653.73 |
| Gradient Boosting | 4,143.35 | 25,844.64 |
| Tuned GB | 8,782.17 | 24, 327.50 |
| Tuned GB | 17,055.26 | 25,405.89 |
| **Tuned GB** | **20,307.08** | **25,492.86** |
| GB(with outlier removal) | | 26,930.21 |

**Figure 3: Residual Plot**



## Conclusion

This project successfully identified key predictors of house prices, with **Overall Quality, Ground Living Area, Garage Cars, and Total Basement square footage** emerging as the most influential features. Among the models tested, **Gradient Boosting** provided the most accurate predictions, likely due to its ability to correct errors iteratively, capture complex relationships, and mitigate overfitting through controlled learning rates.

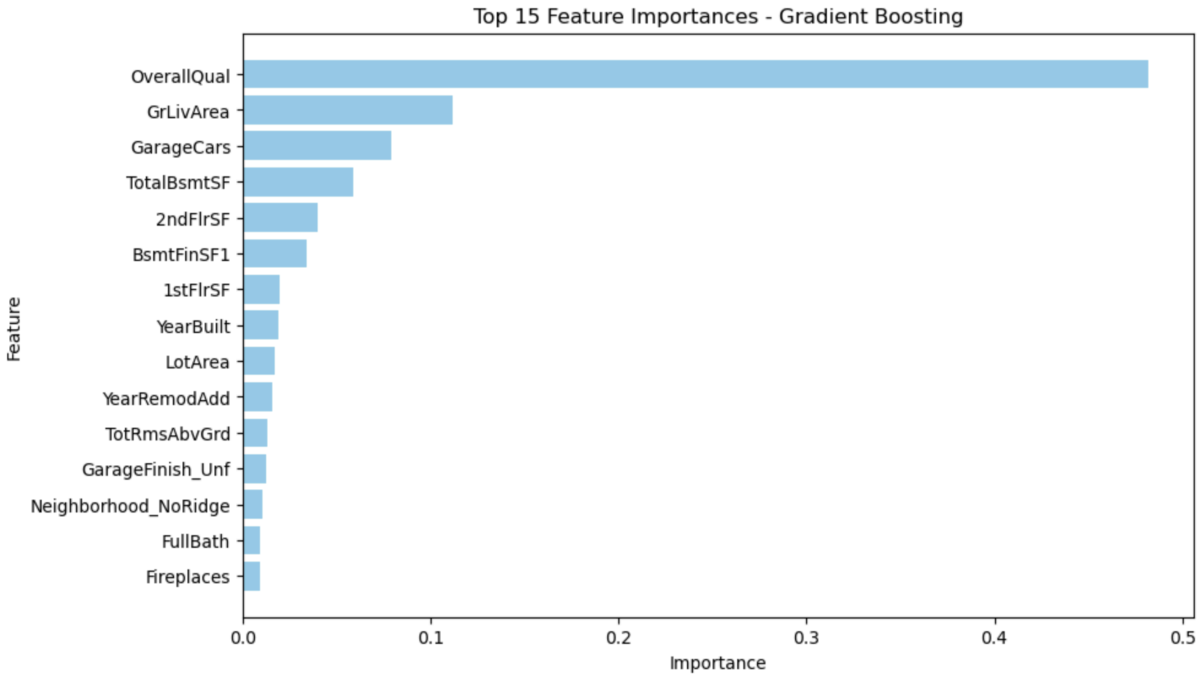
The results confirm the initial hypothesis that at least one independent variable has a **significant** impact on sale price. In fact, multiple factors influence home prices, emphasizing the importance of **property features** in real estate valuation

This model offers valuable insights for various stakeholders:

- **Homeowners** can use these insights to **prioritize renovations** that add the most value.

- **Home buyers** can make **data-driven purchase decisions** by understanding what features drive prices.
- **Real estate agents and investors** can refine **pricing strategies** and identify **high-value properties.**

**Figure 4: Feature Importance Bar Graph**



## Recommendations

- **Optimize Renovation Decisions for Homeowners**
  - Homeowners should **focus on improving Overall Quality, increasing living space, and expanding garages** to maximize sale price.
  - Renovations in less impactful areas, such as additional bathrooms or minor aesthetic changes, may yield **lower ROI**based on model
- **Guiding Pricing Strategies for Real Estate Agents**
  - Agents should emphasize **Overall Quality ratings, living area, and garage capacity** in listings and price negotiations
  - Homes with high **Total Basement square footage** can command premium pricing, especially in markets where finished basements add functional space.
- **Enhance Investment Decision-Making**

  - Real estate investors should **target properties with strong structural features** (e.g., large living areas, garage space) over superficial upgrades.
  - Focusing on **undervalued properties with potential for high-impact renovations** (e.g., increasing Overall Quality rating) can yield higher resale value.
- While the Gradient Boosting model performed best, further improvements are possible. Future work could involve fine-tuning hyperparameters, experimenting with different ensemble methods, or incorporating additional data sources (e.g., neighborhood amenities, economic trends)

## Data References

The dataset used in this project was obtained from **Kaggle** (Gusthema's *House Price Predictions using TensorFlow Decision Forests*). It contains historical housing data, including property characteristics and sale prices. Additionally, **population statistics** were sourced from Worldometer to provide contextual insights on market demand.