Exploratory Data Analysis of Ames Housing Prices

Management/Research Question:

Our primary question was: “Given a set of home features, how accurately can we predict the sale price of a home in Ames, Iowa?” Accurate price prediction can inform buyers and sellers (setting expectations), real estate agents and developers (pricing and investment strategy), lenders (risk assessment and loan-to-value decisions), and local governments (property tax assessment). By exploring the dataset and building predictive models, we aimed to understand which home features drive price differences and how a statistical/machine-learning model can improve on simple heuristics or averages.

Exploratory Data Analysis Findings:

Ames Housing dataset contains 1,460 observations across 81 variables describing each home’s characteristics. We set our target variable as SalePrice, the sale price of each home, and examined how various home features relate to price, identify potential outliers, and engineer new features to better capture non-linear effects. Our aim was to look at housing characteristics for the area and use them to forecast the value of similar homes. From our exploratory data analysis, we found that SalePrice values range from $34,900 to $755,000 with a median of roughly $163,000. The histogram and box-plot of SalePrice show a right-skewed distribution, which becomes approximately normal after log1p transformation, improving model performance and stabilizing variance.

Feature Engineering & Key Predictors:

To improve predictive performance, we expanded the dataset with new features designed to capture important housing characteristics and non-linear effects. Specifically, we created: **TotalSF** (Total square footage of the home), **TotalBaths** (Combined count of full and half bathroom), **HouseAge** (Difference between the year sold and the year built), **SinceRemodel** (Number of years since the last remodel or update), **IsRemodeled** (Indicator variable flagging homes that have been remodeled), **Piecewise GrLivArea** (Split Above-Ground Living Area into two segments at 2,000 sq ft), **OverallQual_sq** (Squared term for the overall quality score to model non-linear effects of quality), **HasGarage** (Indicator variable for whether a property includes a garage)

We combined these variables with key original features to form a comprehensive predictor set.

Model Building and Cross-Validation:

We modeled SalePrice using two different regression approaches to satisfy the assignment’s requirement for at least two models:

1. **Elastic Net Regression** – A linear model with L1/L2 regularization. Balances between Lasso and Ridge regression to handle correlated predictors
2. **Histogram**-Based Gradient Boosting Regressor (HGB) – A tree-based ensemble model capable of capturing complex non-linear relationships without explicit feature transformations.
3. **Extreme**-Gradient Boosting Regressor (XGB) – A gradient boosting decision tree model that utilizes weak prediction models sequentially. Each new tree corrects errors made by previous ones, optimizing a specified loss function.

All modeling was performed through a scikit-learn Pipeline that included data preprocessing and feature engineering. We all assessed model performance using 5-fold cross-validation, reducing the risk of overfitting

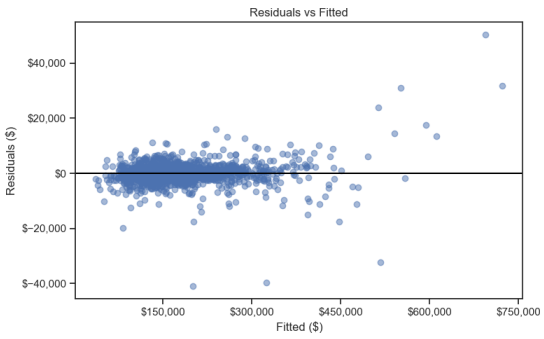
Results:

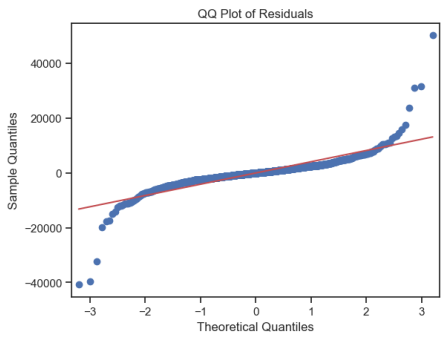
```
elastic
  RMSE: 34882 ± 12224, R²: 0.771 ± 0.200, RMSLE: 0.1506 ± 0.0207
hgb
  RMSE: 32559 ± 7142, R²: 0.815 ± 0.107, RMSLE:0.1481 ± 0.0164
```

The HGB model achieved lower RMSE and higher R^2, which indicated to us that it would provide better predictive performance on the training/validation splits. Thus, we selected the HGB Regressor as our final model for predictions on the test set.

Residual Analysis:

We plotted Residuals vs. Fitted Values to examine whether errors were randomly distributed. Most of the residuals for mid-priced homes cluster evenly around zero, suggesting the model captures the general trend well. However, as fitted values increase toward higher-priced homes, the residuals spread out and show larger deviations, indicating heteroscedasticity. This means **our model performs best on average-priced homes** but has greater uncertainty when predicting very high-priced properties.





Model Results and Recommendations:

Our analysis found that housing prices in Ames, Iowa, are strongly influenced by factors such as above-ground living area, basement size, overall quality, total square footage, number of bathrooms, and whether the home has been remodeled or has a garage. By adding new features like Total Square Footage, Total Bathrooms, and piecewise splits for living area, we were able to capture non-linear relationships that basic models would miss.

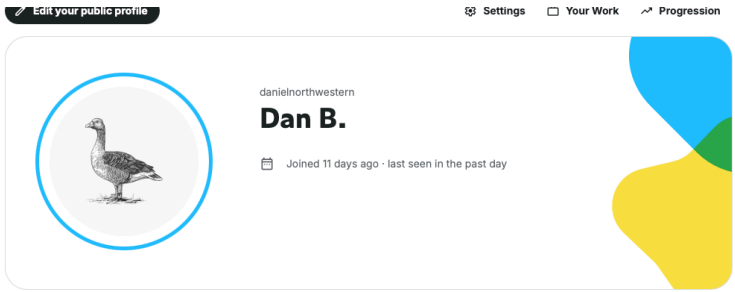
Our HGB model with a cross-validated RMSE of about $32,559Both models achieved R^2 values above 0.77, indicating that they explain a large portion of the variation in home prices.

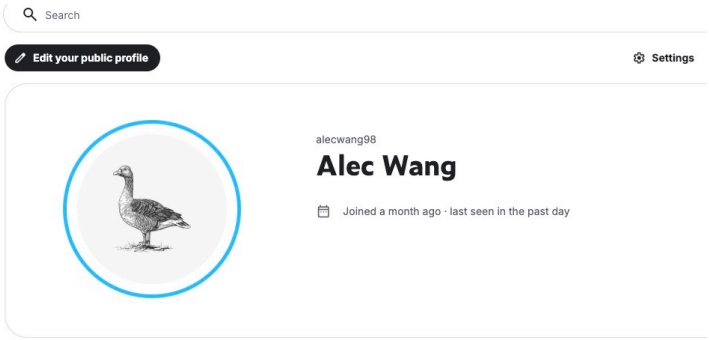
In summary, home sales prices in Ames is largely influenced by square footage, more bathrooms, higher quality finishes, and a remodeled interior. This analysis allows stakeholders to make data-driven decisions for purchases and investments.

Appendix

Kaggle Results:

User Name: Danielnorthwestern (Daniel Balette) - **0.14991**
User Name: alecwang98 (Alec Wang) - **0.13771**





kaggle

Search

Create

Home

Competitions

Datasets

Models

Benchmarks

Code

Discussions

Learn

More

Your Work

VIEWED

House Prices - Advan...

Titanic - Machine Lear...

EDITED

notebook134715ce78

KAGGLE · GETTING STARTED PREDICTION COMPETITION · ONGOING

Submit Prediction

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

All Successful Errors Recent

Submission and Description	Public Score
submission.csv Complete · now	0.14991

Submission Details

notebookcba7f1cfae - xgb test

Complete · 34s ago

Score: 0.13771

UPLOADED FILES

submission.csv (21 KiB)

DESCRIPTION

Notebook notebookcba7f1cfae | Version 20

40 / 500

RPATEL9877 · 13M AGO · PRIVATE

0

Module 2 Assignment Housing prices

Notebook Input Output Logs Comments (0) Settings

Competition Notebook	Public Score	Best Score
House Prices - Advanced Regression Te...	2.49223	2.49217 V4