House Value Prediction for Strategic Real Estate Decision-Making

...

Team G1
Anna Dominic
Karthikeyan Shanmugam
Manisha Goyal
Siri Desiraju

Business Problem

Objective:

Develop a model to predict house prices in Ames, Iowa.

Impact:

- Enables data-driven real estate decisions in investments, sales, and development
- Enhances understanding of factors influencing property values

Target Stakeholders

Who am I?	What is my goal?	How can the model help me?
Real Estate Agent	Set competitive prices to attract buyers & satisfy sellers	Use model predictions to enhance service quality and optimize listings
Real Estate Developer	Plan lucrative developments that captivate buyers	Use model insights to focus on profitable features and projects
Home Owner	Maximize home value with strategic remodeling	Use model for home value prediction and insights on remodel impact
Home Buyer	Find a home that meets personal needs at a fair price	Use model predictions to determine fair pricing and negotiate effectively

Understanding the Data

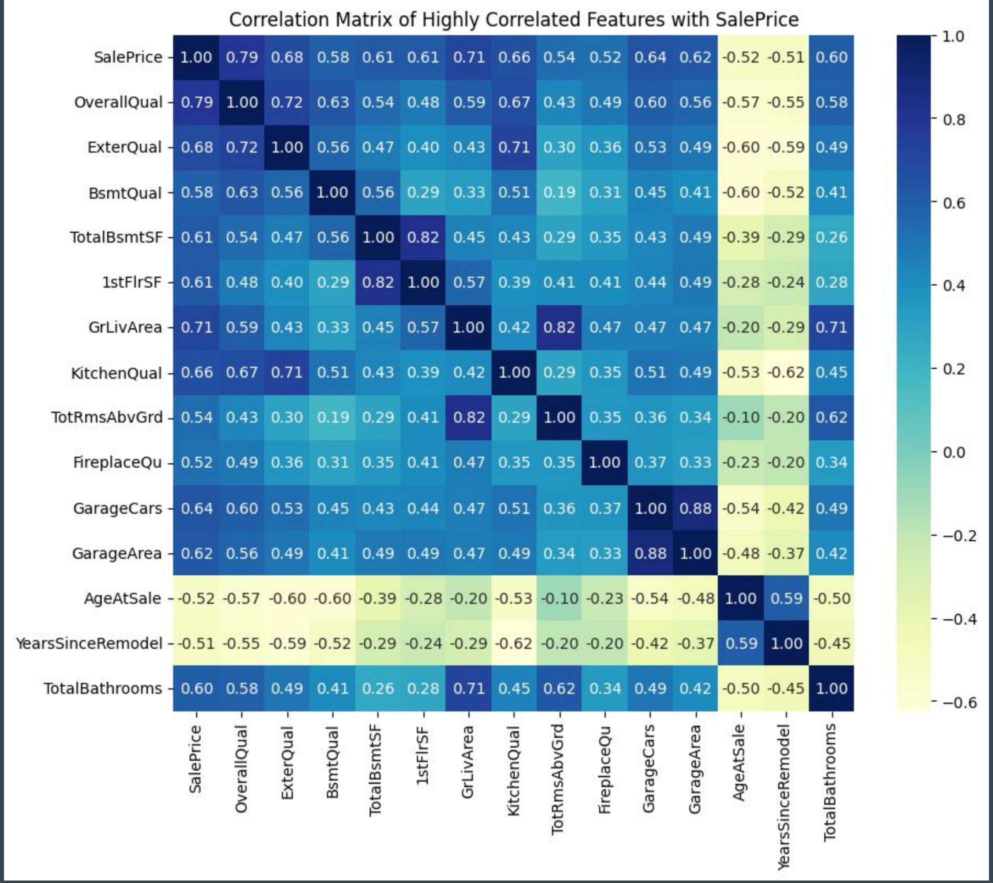
Dataset: House Prices of Ames, Iowa (from Kaggle)

>1400
properties

~4
years

~35
numeric

~46
categorical



Correlation Matrix of Highly Correlated Features with SalePrice

	SalePrice	OverallQual	ExterQual	BsmtQual	TotalBsmtSF	1stFlrSF	GrLivArea	KitchenQual	TotRmsAbvGrd	FireplaceQu	GarageCars	GarageArea	AgeAtSale	YearsSinceRemodel	TotalBathrooms
SalePrice	1.00	0.79	0.68	0.58	0.61	0.61	0.71	0.66	0.54	0.52	0.64	0.62	-0.52	-0.51	0.60
OverallQual	0.79	1.00	0.72	0.63	0.54	0.48	0.59	0.67	0.43	0.49	0.60	0.56	-0.57	-0.55	0.58
ExterQual	0.68	0.72	1.00	0.56	0.47	0.40	0.43	0.71	0.30	0.36	0.53	0.49	-0.60	-0.59	0.49
BsmtQual	0.58	0.63	0.56	1.00	0.56	0.29	0.33	0.51	0.19	0.31	0.45	0.41	-0.60	-0.52	0.41
TotalBsmtSF	0.61	0.54	0.47	0.56	1.00	0.82	0.45	0.43	0.29	0.35	0.43	0.49	-0.39	-0.29	0.26
1stFlrSF	0.61	0.48	0.40	0.29	0.82	1.00	0.57	0.39	0.41	0.41	0.44	0.49	-0.28	-0.24	0.28
GrLivArea	0.71	0.59	0.43	0.33	0.45	0.57	1.00	0.42	0.82	0.47	0.47	0.47	-0.20	-0.29	0.71
KitchenQual	0.66	0.67	0.71	0.51	0.43	0.39	0.42	1.00	0.29	0.35	0.51	0.49	-0.53	-0.62	0.45
TotRmsAbvGrd	0.54	0.43	0.30	0.19	0.29	0.41	0.82	0.29	1.00	0.35	0.36	0.34	-0.10	-0.20	0.62
FireplaceQu	0.52	0.49	0.36	0.31	0.35	0.41	0.47	0.35	0.35	1.00	0.37	0.33	-0.23	-0.20	0.34
GarageCars	0.64	0.60	0.53	0.45	0.43	0.44	0.47	0.51	0.36	0.37	1.00	0.88	-0.54	-0.42	0.49
GarageArea	0.62	0.56	0.49	0.41	0.49	0.49	0.47	0.49	0.34	0.33	0.88	1.00	-0.48	-0.37	0.42
AgeAtSale	-0.52	-0.57	-0.60	-0.60	-0.39	-0.28	-0.20	-0.53	-0.10	-0.23	-0.54	-0.48	1.00	0.59	-0.50
YearsSinceRemodel	-0.51	-0.55	-0.59	-0.52	-0.29	-0.24	-0.29	-0.62	-0.20	-0.20	-0.42	-0.37	0.59	1.00	-0.45
TotalBathrooms	0.60	0.58	0.49	0.41	0.26	0.28	0.71	0.45	0.62	0.34	0.49	0.42	-0.50	-0.45	1.00

Hypothesis Testing for Insights

Recently renovated properties sell at higher prices, controlling for the age of the property

OLS Regression Analysis

Market preferences reflect a higher valuation for certain dwelling types and styles that align with buyer demands

ANOVA & Tukey's HSD Post-hoc Analysis

Data Processing

Missing Values

Imputed moderate to low missing values and preserved meaningful 'NA's

High missing values evaluated for retention or exclusion

Encoding Categorical

Ordinal features mapped to numeric scales

Nominal variables transformed via one-hot encoding

Feature Engineering

Developed new features like 'AgeAtSale'

Normalized feature values using Min-Max scaling

After data preparation, there were **213 features**

Selecting Features for the Model

Feature Selection	Number of Features	R^2
Lasso Regression	88	0.873
Recursive Feature Elimination	5	0.664
Principal Component Analysis	107	0.872

Feature selection methods evaluated on baseline linear regression model

Lasso Regression has the highest R^2 value and is more interpretable than PCA

Model Development Strategy

Employed both interpretable (e.g., Random Forest) and complex (e.g., XGBoost) models to balance understandability with predictive power

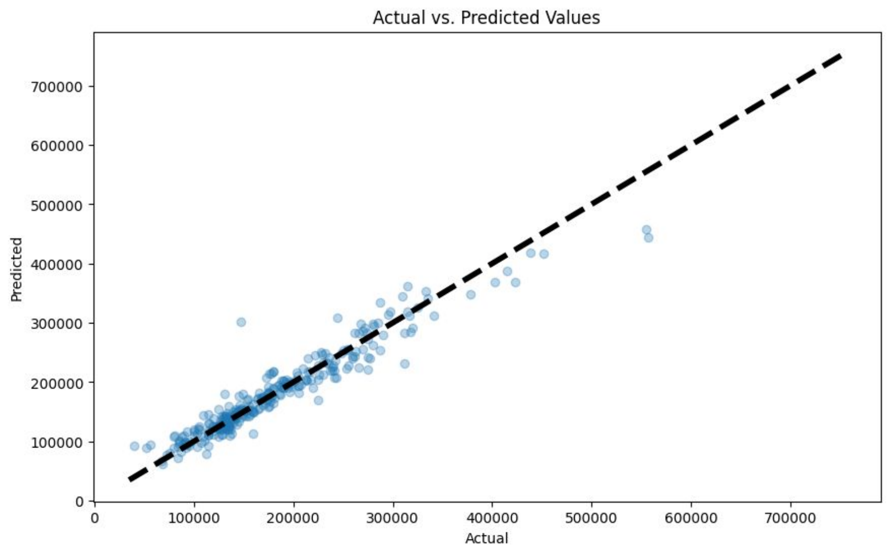
Conducted comprehensive feature analysis to determine model efficacy using both full and reduced feature sets

Utilized Grid Search with Cross-Validation to fine-tune model parameters, ensuring optimal performance

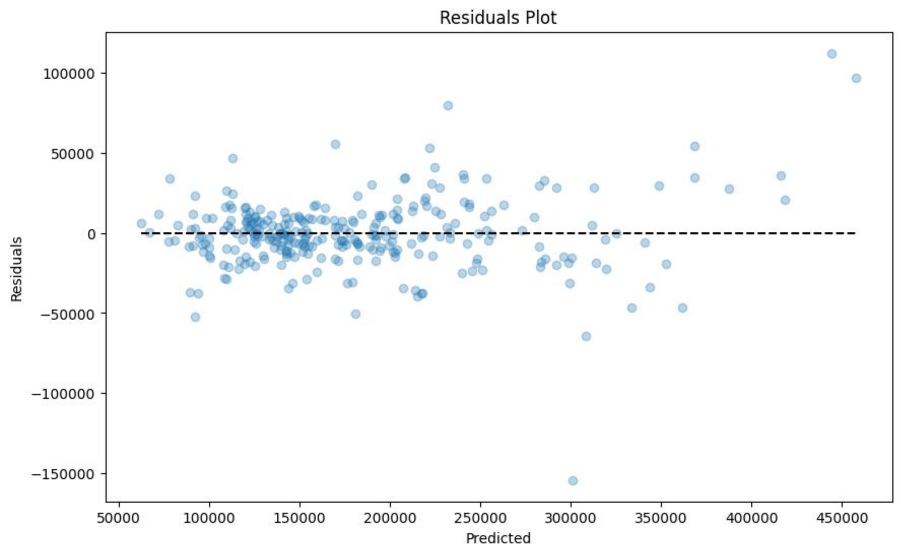
Model Results

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

Evaluating the Model

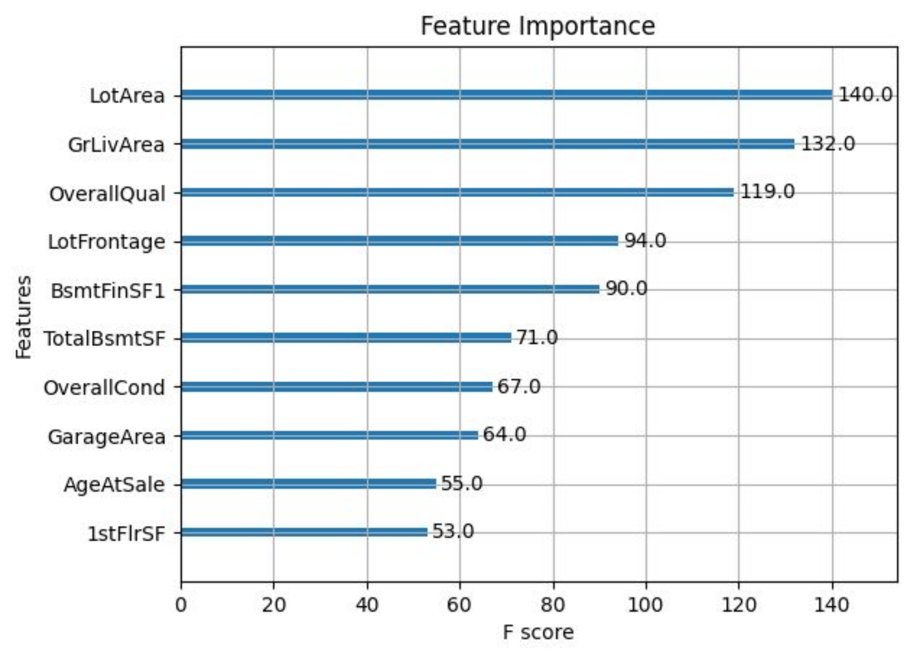


Actual vs Predicted Values

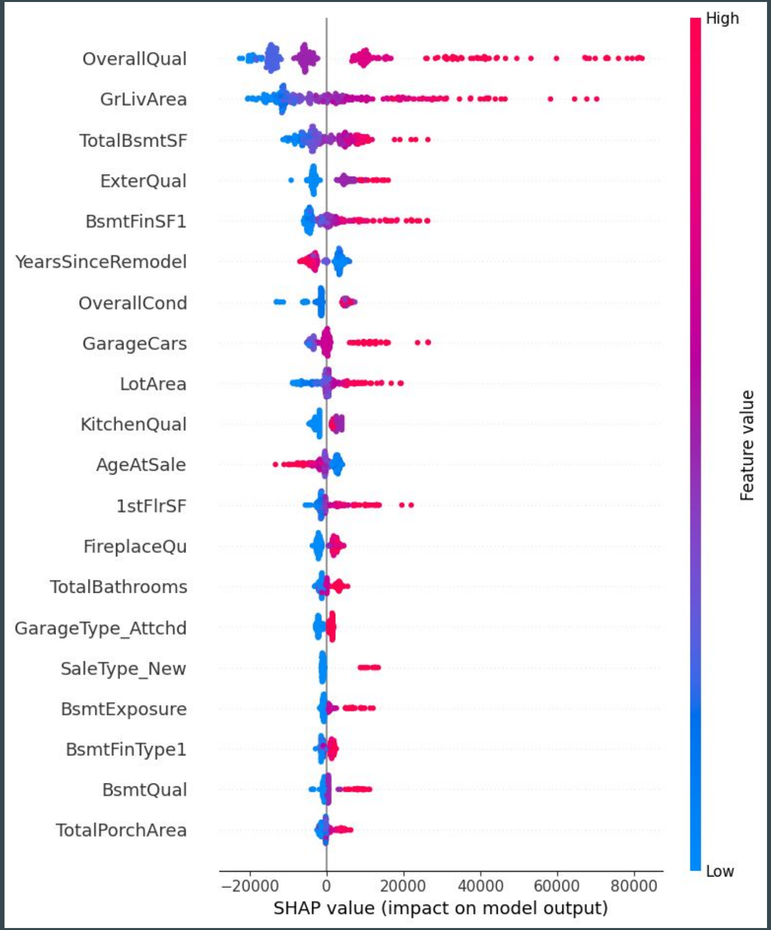


Residuals Plot

Identifying Key Features

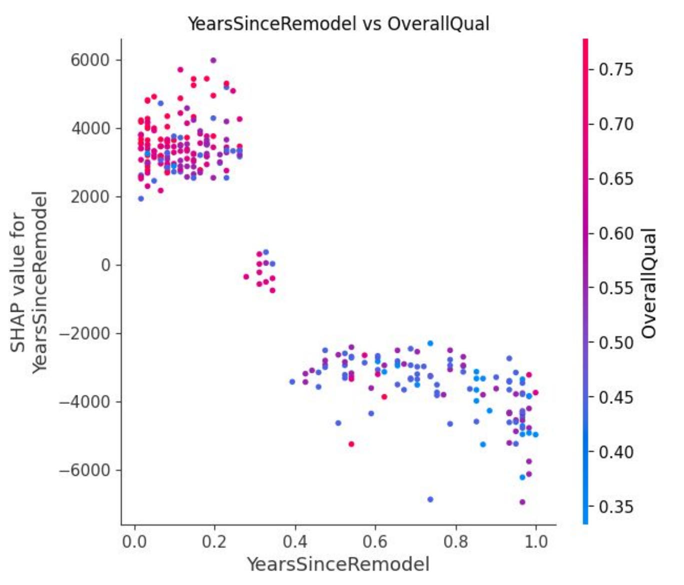


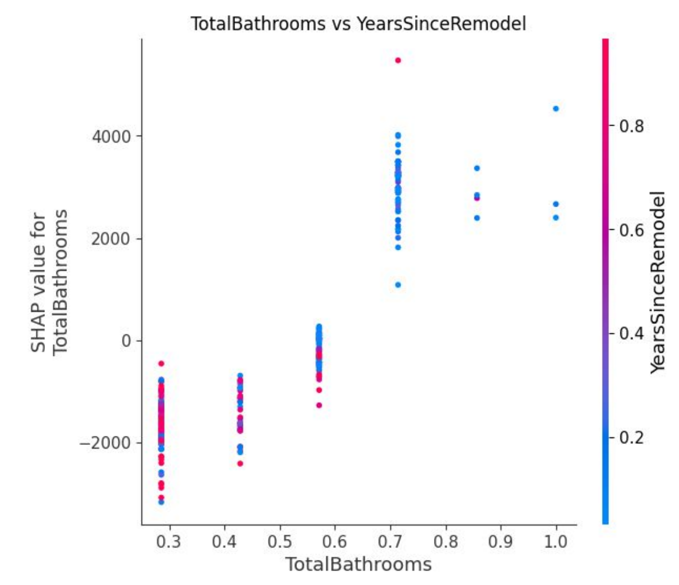
XGBoost Feature Importance



SHAP Explainer

Identifying Key Features

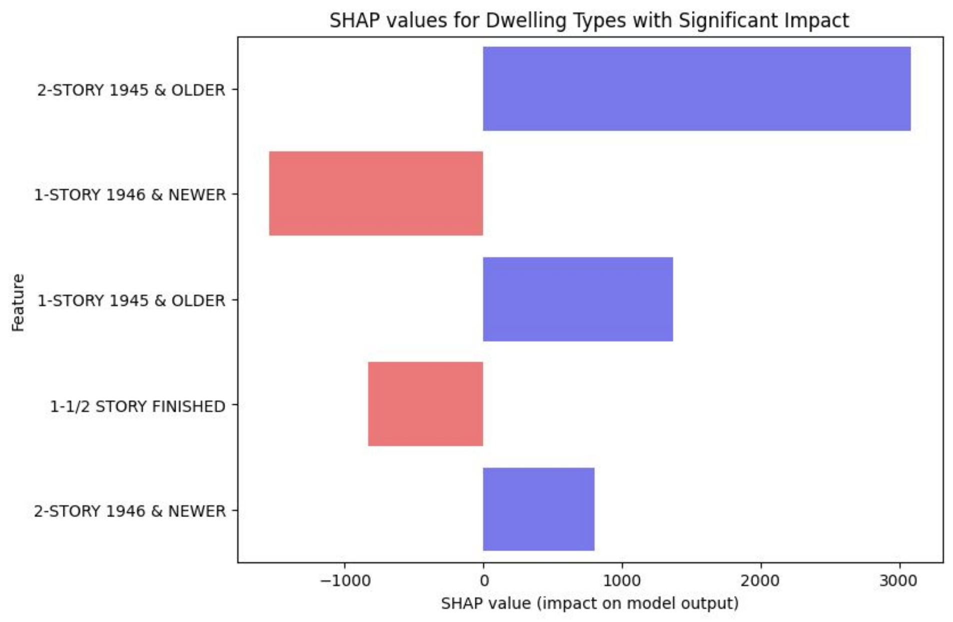




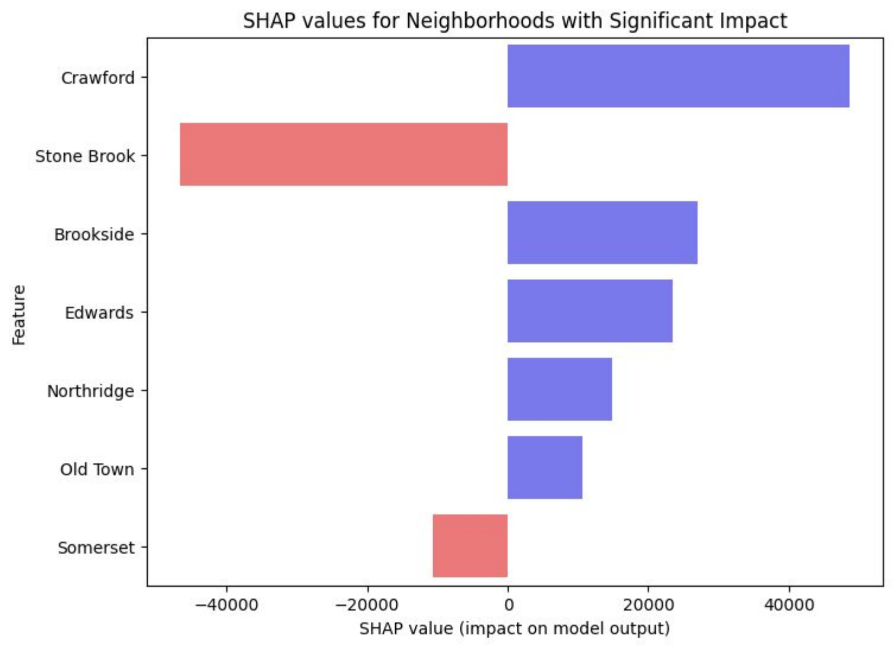
Factors like recent renovations drive up ‘Overall Quality’ of the house

Recent renovations have significant impact by improving ‘Number of Bathrooms’

Identifying Key Features



SHAP Explainer - Dwelling Types



SHAP Explainer - Neighborhoods

Recommendations to Stakeholders

Real Estate Agent

- Evaluate house prices with given features
- Guide premium buyers to Crawford neighborhood and budget buyers to Stone Brook neighborhood

Real Estate Developer

- Design dwelling types like ‘2-Story’
- Prioritize living space quality and garage size

Recommendations to Stakeholders

Home Owner

- Assess house value
- Evaluate targeted renovations to increase value:
    - Repaint and remodel interior and exterior finish
    - Ensure basement is finished
    - Remodel kitchen

Home Buyer

- Assess house value based on features and location
- Assess possible features and locations given a budget

Extending Scope to other Cities

Motivation: Increase business value and broaden applicability

Methodology: Leverage the best-performing model (XGBoost) to make predictions across diverse markets

City/State	R^2	RMSE	Important Features
New York	0.589	3218761.26	Property area, Zip code, Number of bathrooms
California	0.707	61945.62	Median household income, Distance to the ocean, Number of bedrooms
Seattle	0.577	344297.69	Apartment size, Zip code, Number of bathrooms

Future Work

Enhance Model Relevance and Accuracy

- **Cross-City Validation:** Explore the potential for better results with more comprehensive datasets
- **Local Economic Indicators:** Integrate metrics like employment rates and income levels to predict market shifts
- **Consumer Behavior:** Use surveys to capture buyer preferences and predict desirability trends
- **Real-Time Market Data:** Enhance predictions with live data on listings and sales