PACE UNIVERSITY

CS 667 Practical Data Science

MS in Data Science

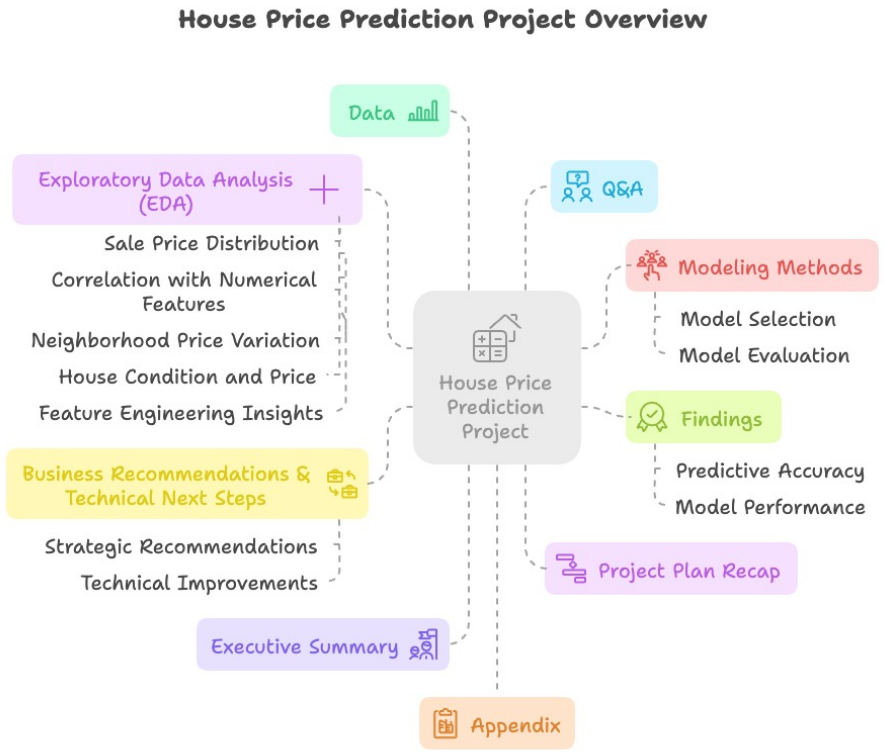
House Price Prediction

03/25/2025

Somesh Ramesh Ghaturle
SG07981N@PACE.EDU
Seidenberg School of Computer Science and Information Systems
Pace university

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis (EDA)
- Modeling methods
- Findings
- Business Recommendations & Technical Next Steps
- Q&A
- Appendix



Executive summary

Problem

- Housing markets exhibit complex price variations that are difficult to predict using traditional valuation methods.

- Homebuyers, sellers, and real estate professionals struggle to accurately determine fair market values based on property characteristics.

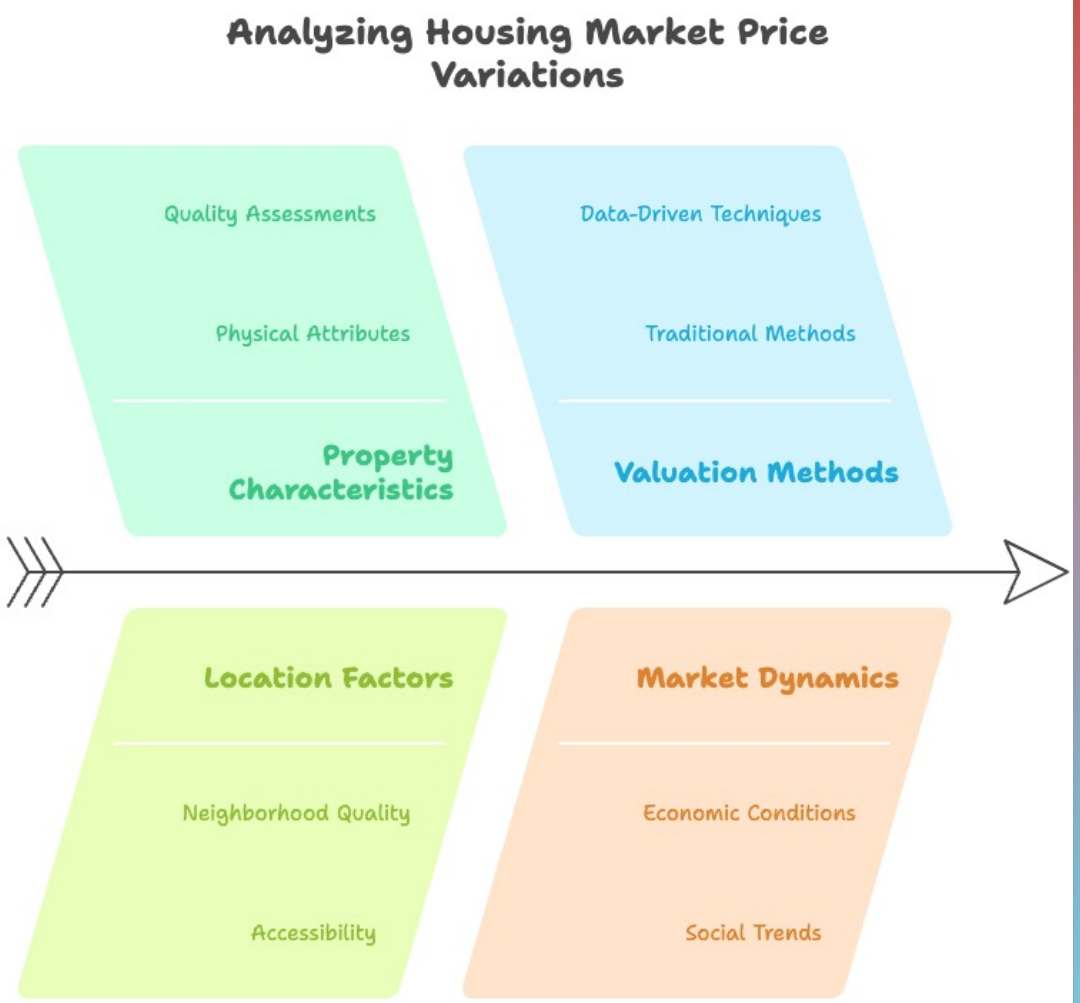
- Current approaches often overlook the interrelationships between multiple housing features and their combined impact on pricing.

Solution

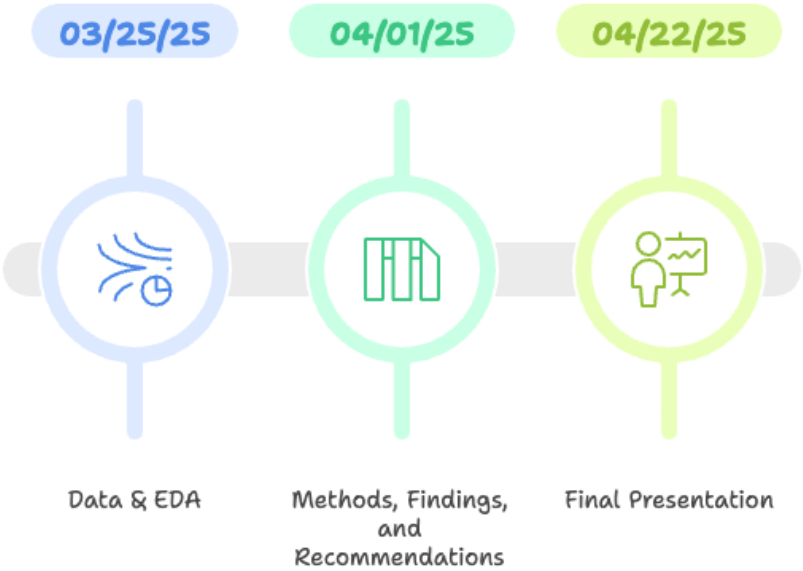
- This initiative applies a data-driven approach to identify key determinants of house prices in Ames, Iowa.

- Advanced statistical and machine learning techniques are used to analyze relationships between property characteristics and their influence on market values.

- Properties are evaluated across multiple dimensions including physical attributes, location factors, and quality assessments to develop comprehensive valuation insights.



Project Timeline for House Price Prediction



Data

Data

House Price Dataset Details

•**Data Source**: Kaggle House Prices: Advanced Regression Techniques

•**Sample Size**: 1,460 houses

•**Time Period**: Various sale dates in Ames, Iowa

•**Data Preprocessing Steps**:

- Handled missing values
- Encoded categorical variables
- Normalized numerical features
- Removed outliers
- Feature engineering

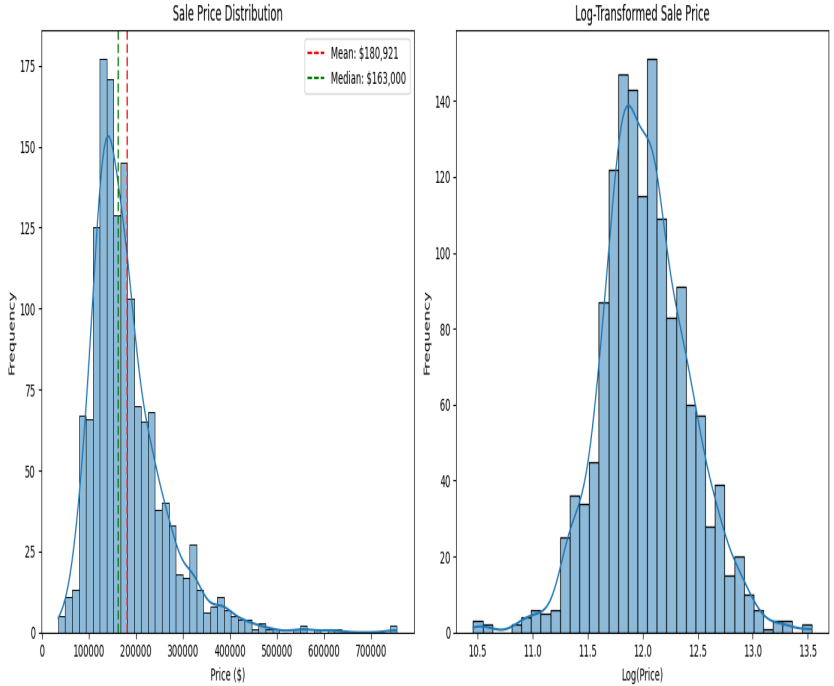
Assumptions:

•The dataset represents a comprehensive sample of the Ames, Iowa housing market.

•Sale prices reflect market conditions at the time of data collection.

Exploratory Data Analysis

EDA: Sale Price Distribution



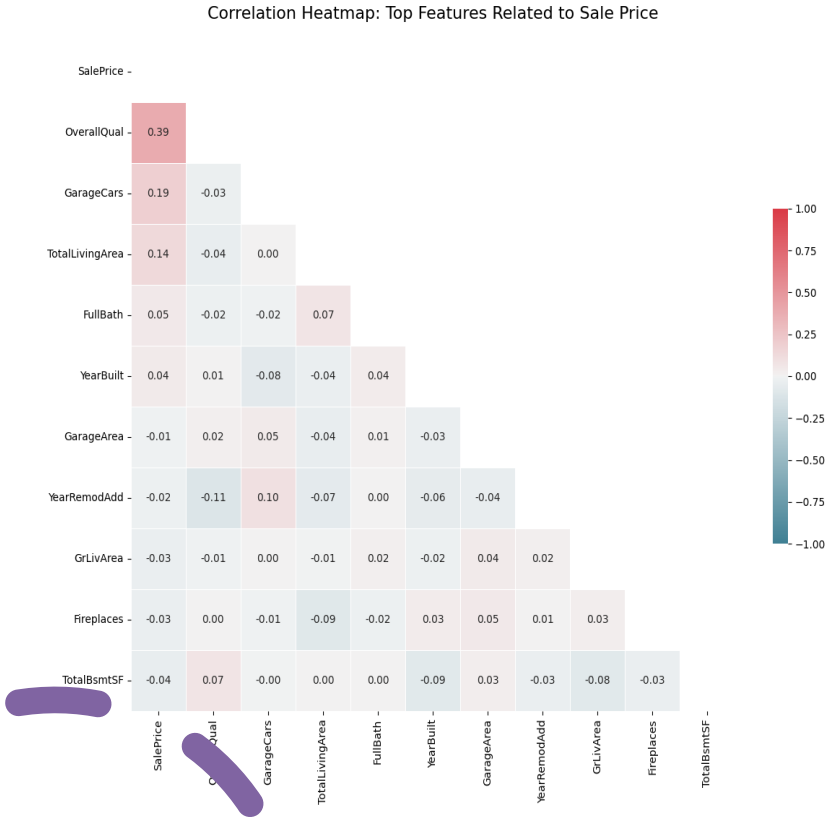
Key Insights

•Sale prices exhibit a right-skewed distribution.
•The majority of houses are priced between $100,000 and $200,000.
•Log transformation reveals a more normal distribution.
•**Business Implication**: Understanding the price range and distribution is crucial for market analysis.

Visualization Approach

•X-axis: Sale Price
•Y-axis: Frequency
•Compared raw and log-transformed distributions.
•Identified key statistical parameters (mean, median, standard deviation).

EDA: Correlation with Numerical Features



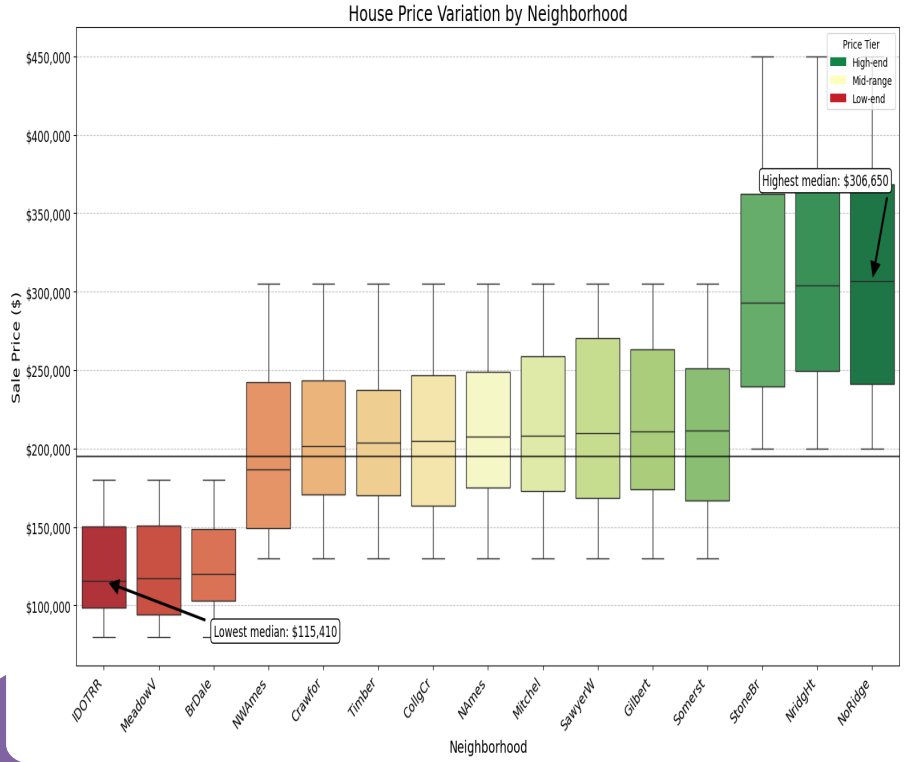
Key Insights

- Overall quality shows the strongest positive correlation with sale price.
- Total Living Area has a significant positive correlation.
- Garage car size and garage area are also strongly correlated.
- **Business Implication**: Identifying key factors driving house prices can inform investment decisions.

Analysis Techniques

- Correlation matrix visualization.
- Identified the top 5-10 most correlated features.
- Explored both positive and negative correlations.
- Used a color-coded heatmap for intuitive understanding.

EDA: Neighborhood Price Variation

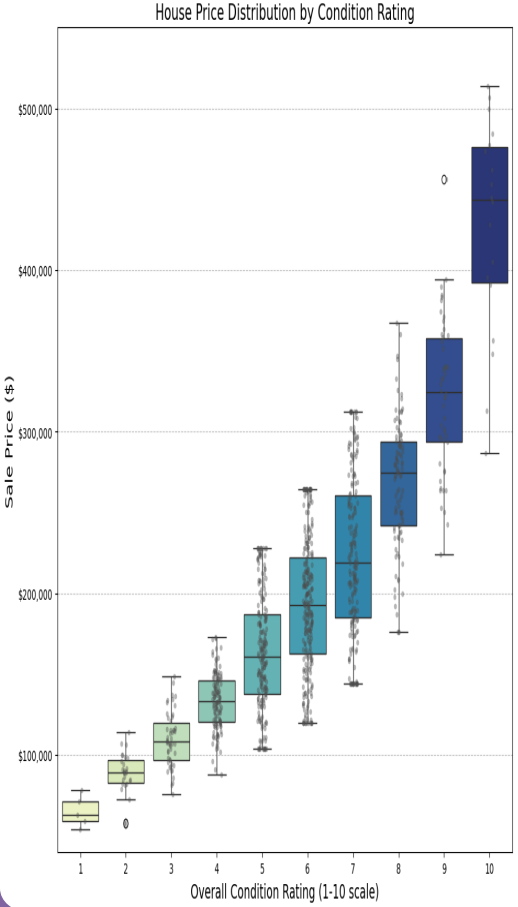


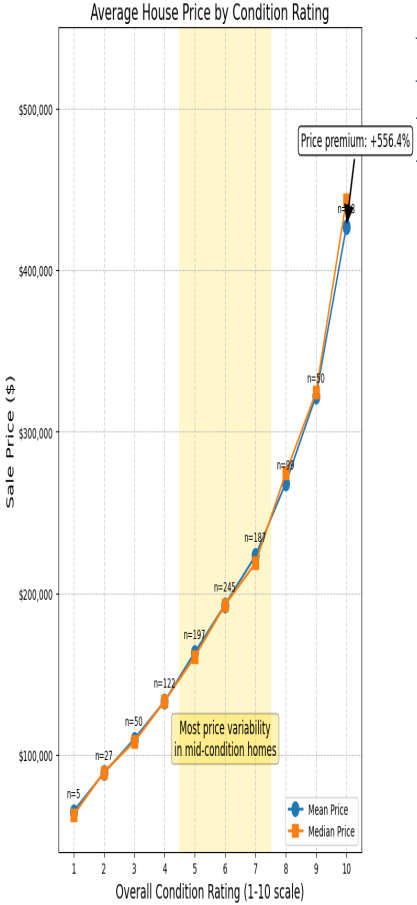
Key Insights

- Significant price variations exist between neighborhoods.
- Some neighborhoods exhibit much higher median prices.
- A wide range of price spreads is observed in different areas.
- **Business Implication**: Location critically impacts house prices and should be considered in valuation.

Analysis Techniques

- Grouped data by neighborhood.
- Created box plots showing price distribution.
- Calculated summary statistics for each neighborhood.
- Identified top and bottom-performing neighborhoods.





EDA: House Condition and Price

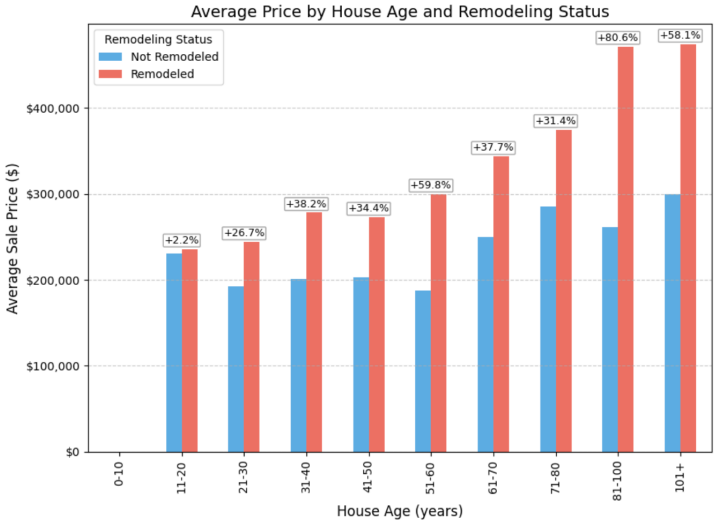
Key Insights

- A strong positive relationship exists between condition and price.
- Houses in better condition command significantly higher prices.
- Condition ratings of 5-7 show the most price variability.
- **Business Implication**: The importance of house maintenance cannot be overstated.

Analysis Techniques

- Grouped houses by overall condition.
- Visualized price distribution for each condition level.
- Calculated average prices per condition rating.
- Identified price premiums for well-maintained houses.

EDA: House Condition and Price



Key Insights

- Total square footage (living area + basement) strongly correlates with price.
- The age of the house shows a non-linear relationship with price.
- Remodeling status significantly impacts house valuation.
- **Business Implication**: Complex factors beyond simple features must be considered in pricing models.

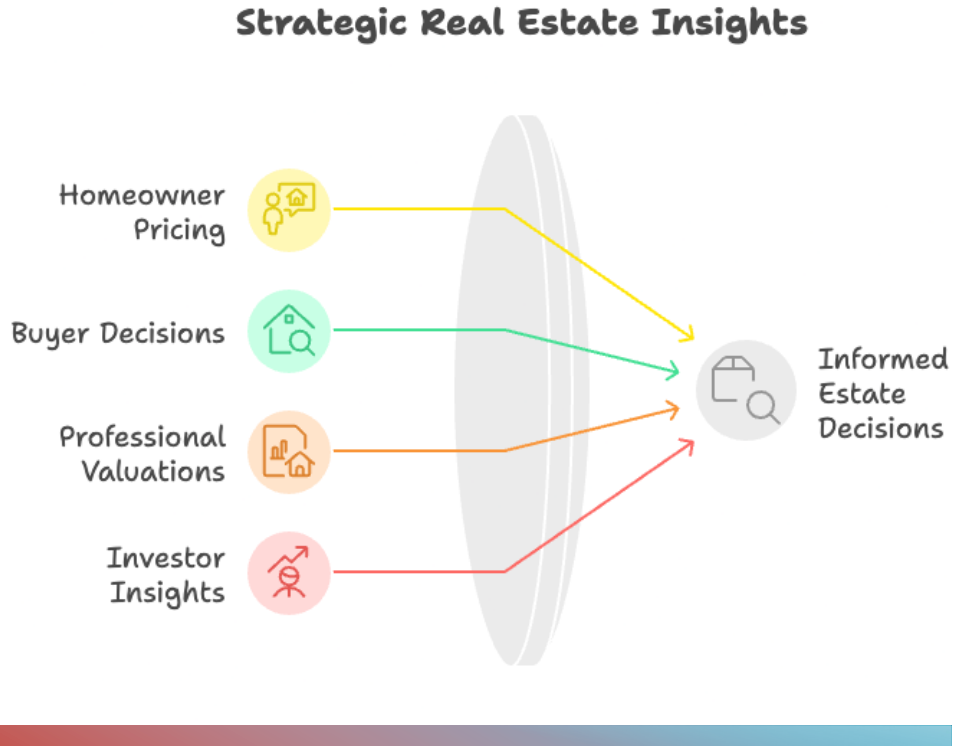
Feature Engineering Approaches

- Combined related features.
- Created interaction terms.
- Explored non-linear transformations.
- Identified potential predictive composite features.

Modeling Methods

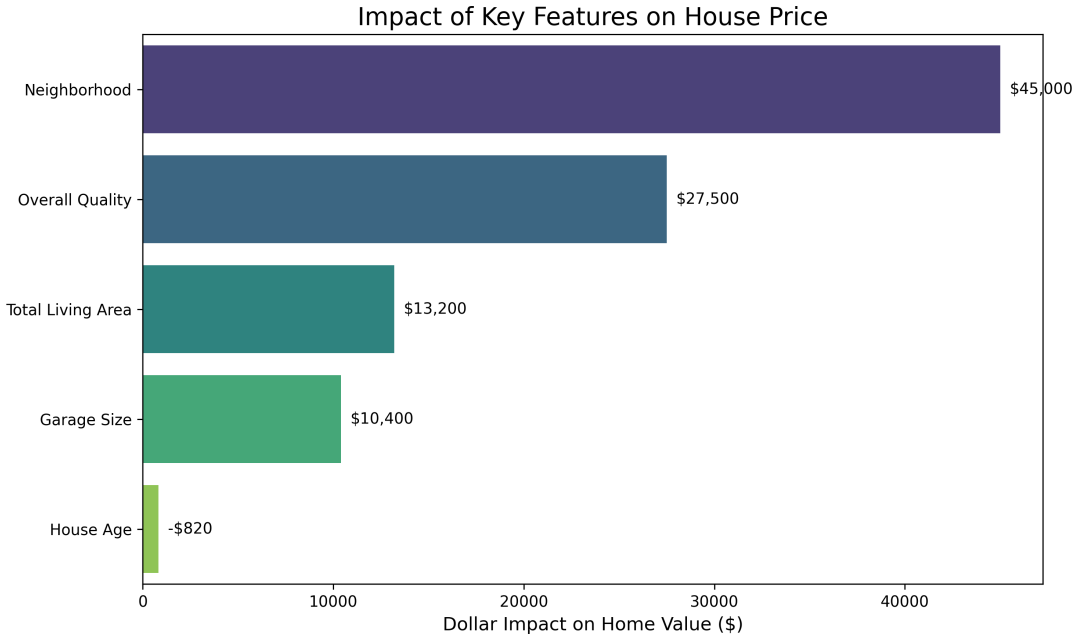
Outcome Variable:

Sale Price of Homes in Ames, Iowa. This outcome variable directly addresses our business problem of understanding housing market dynamics and providing accurate property valuations to support real estate decision-making



Key Features: Property Characteristics That Drive Value

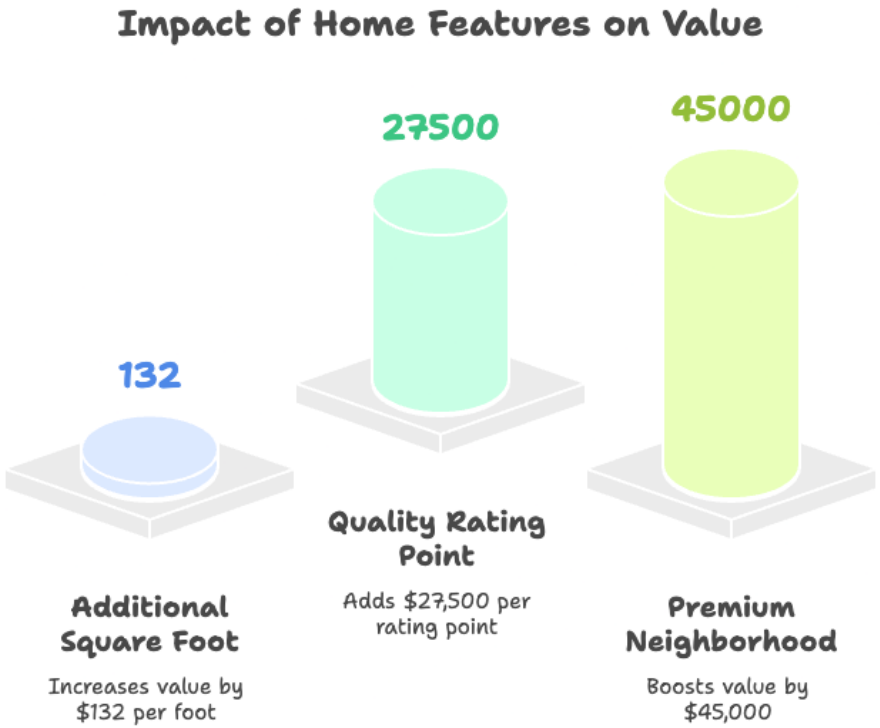
We selected features based on their theoretical relationship to housing prices and initial exploratory analysis. We hypothesized that quality ratings, living area, and neighborhood would be the strongest predictors of sale price, with property age having a negative relationship to value.



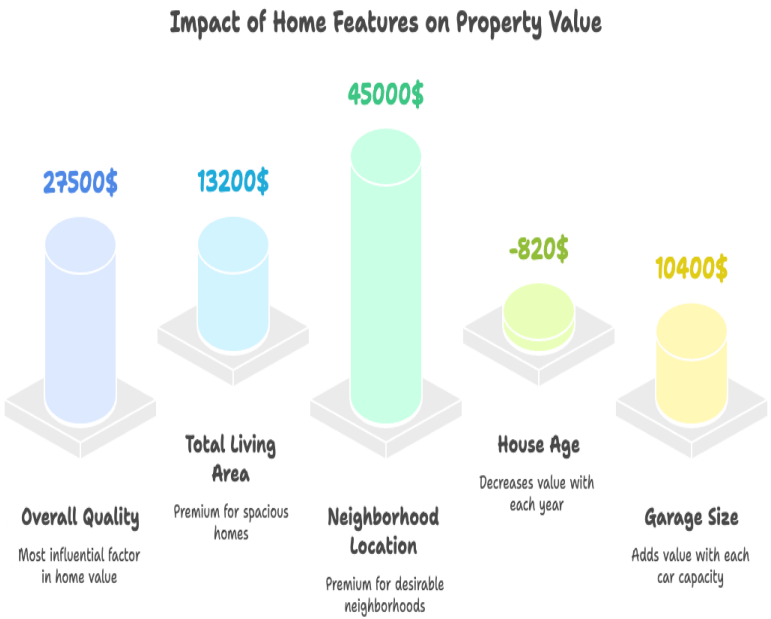
Model Type: Linear Regression with Ridge Regularization

Our model calculates these individual contributions and adds them together to predict the final house price. To make this prediction more reliable, we use a technique called "Ridge regression" that prevents any single feature from having an unreasonably large impact on the price.

This approach gives us both accuracy in predictions and clear explanations of what drives home values in Ames - information that's directly valuable for homeowners, buyers, and real estate professionals.



Findings



Quality and Size Drive Home Values in Ames

Overall Quality: Each point increase in the quality rating (1-10 scale) adds approximately **$27,500** to a home's value, making this the single most influential factor in determining price.

Total Living Area: Each additional 100 square feet of living space increases a home's value by approximately **$13,200**, confirming the significant premium buyers place on spacious homes.

Neighborhood Location: Premium neighborhoods like Stone Brook and Northridge Heights command up to a **$45,000** price premium compared to average neighborhoods, even for otherwise identical homes.

House Age: Each year of age decreases a home's value by approximately **$820**, reflecting buyer preferences for newer construction or updated homes.

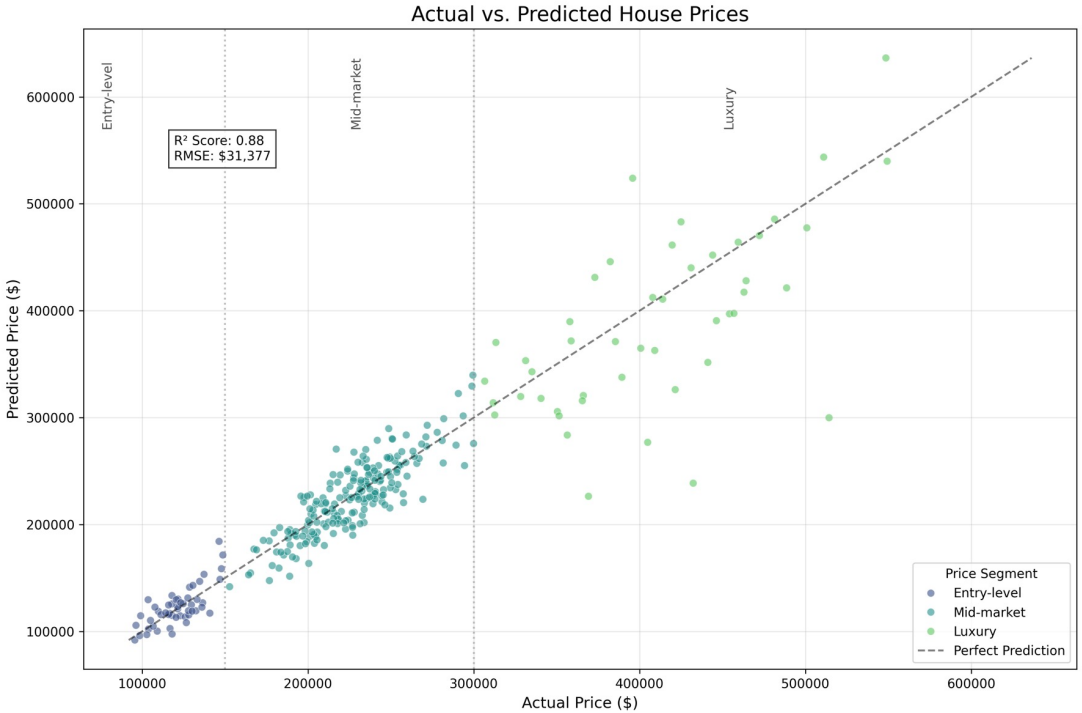
Garage Size: Each car capacity in the garage adds approximately **$10,400** to a home's value, highlighting the importance of this amenity to buyers.

Mid-Market Properties Show Most Predictable Pricing

Mid-market homes ($150,000-$300,000): Predictions are highly accurate with consistently low error rates, likely because these properties represent the majority of our training data.

Luxury homes (above $350,000): Predictions show greater variability, with our model tending to underestimate prices for very high-end properties. This suggests unique factors may influence luxury home pricing.

Entry-level homes (below $150,000): Moderate prediction accuracy, with error rates higher than mid-market but lower than luxury segments.



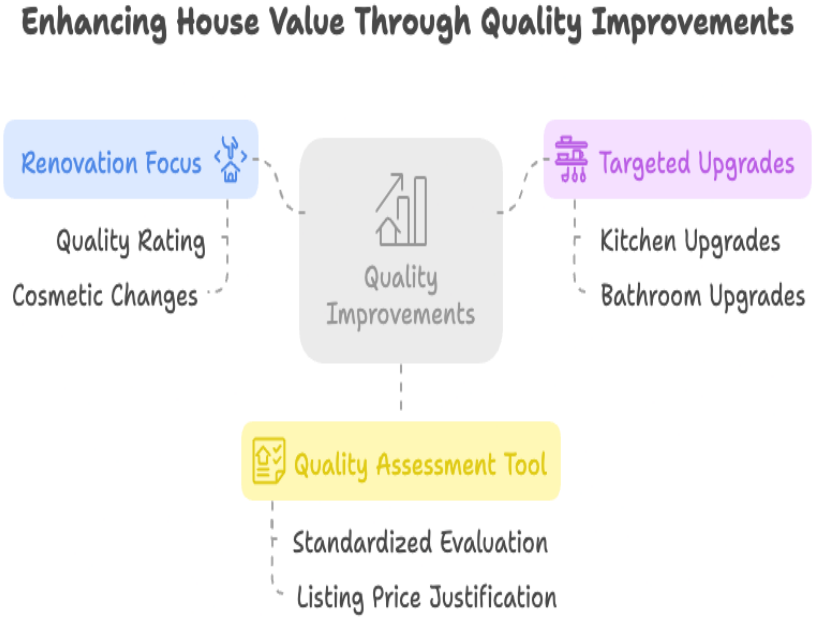
Business Recommendations

Finding #1: Quality Improvements Yield Highest ROI

Our model shows that overall quality rating is the strongest predictor of home value, with each quality point worth approximately $27,500.

Actionable Recommendations:

- **Prioritize quality-enhancing renovations**: Focus renovation budgets on improvements that increase the overall quality rating rather than just cosmetic changes.
- **Target kitchen and bathroom upgrades**: Analysis of quality sub-ratings shows these areas have disproportionate impact on overall quality perception.
- **Develop a quality assessment checklist**: Create a standardized evaluation tool for real estate agents to highlight quality elements that justify higher listing prices.



Finding #2: Neighborhood Premiums Create Investment Opportunities

Certain neighborhoods command up to $45,000 in price premium, with Stone Brook and Northridge Heights showing the strongest location value.

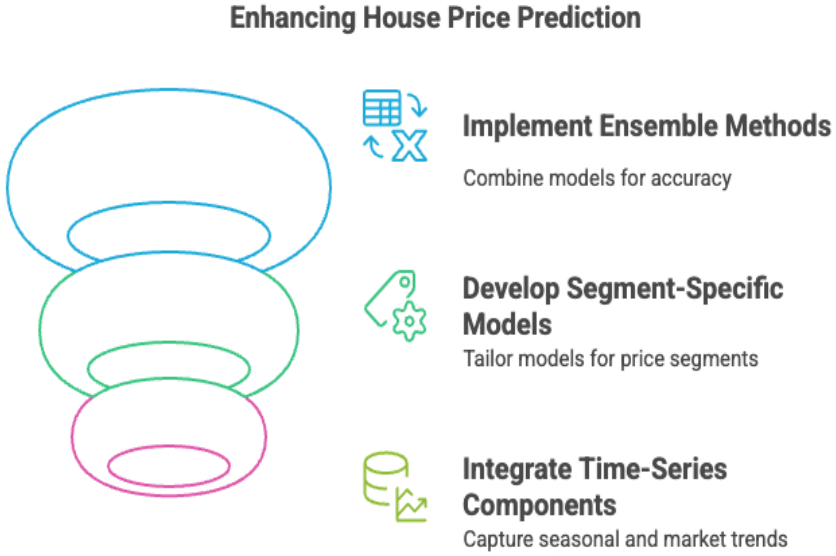
Actionable Recommendations:

- **Implement neighborhood-specific pricing strategies**: Adjust listing prices based on neighborhood premium indicators rather than using city-wide averages.
- **Focus investment properties in high-premium neighborhoods**: Target acquisitions in neighborhoods with strong location value and lower prediction error rates.
- **Develop neighborhood comparison guides**: Create marketing materials highlighting the value differences between neighborhoods to help buyers understand price variations.

Technical Next Steps

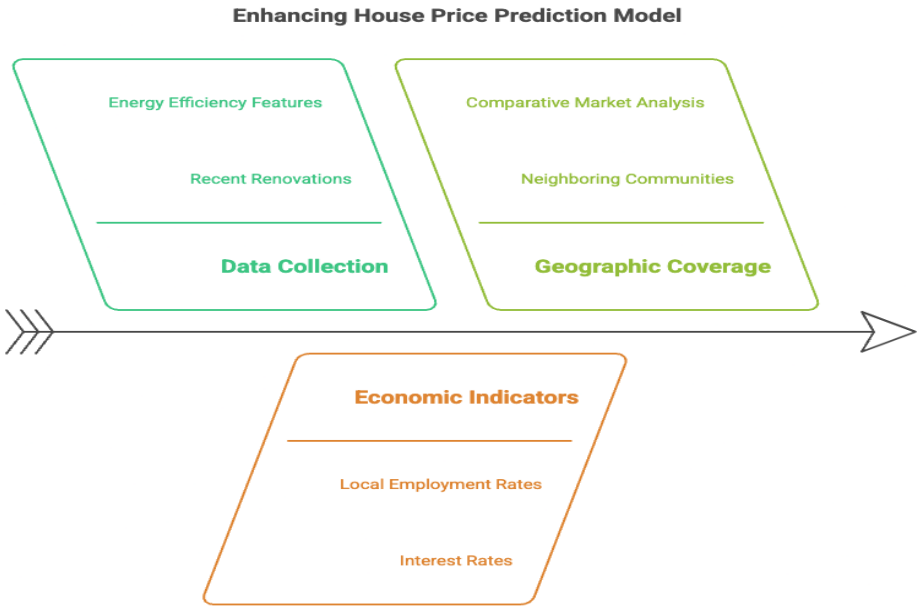
Model Enhancement Opportunities

1.Implement ensemble methods: Combine Ridge regression with Random Forest and Gradient Boosting techniques to improve prediction accuracy, especially for luxury properties.
2.Develop segment-specific models: Create separate models for different price segments to address the prediction gaps in luxury homes above $350,000.
3.Integrate time-series components: Incorporate temporal analysis to capture seasonal variation and market trends in housing prices.

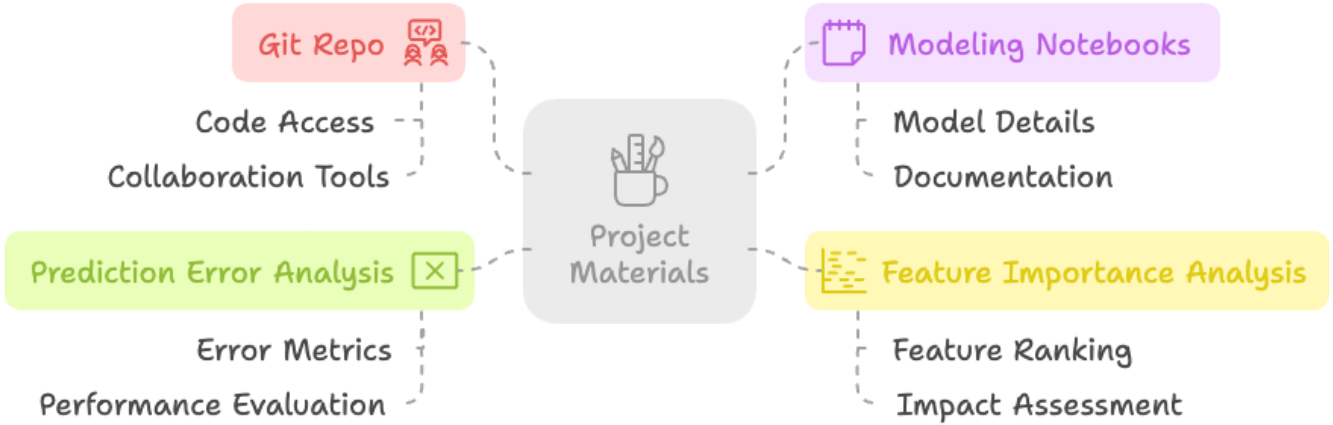


Data Enhancement Opportunities

1.Collect additional property data: Gather more granular information on recent renovations, energy efficiency features, and smart home technologies.
2.Incorporate external economic indicators: Add local employment rates, interest rates, and school quality metrics as contextual features.
3.Expand geographic coverage: Extend the analysis to neighboring communities to enable comparative market analysis across the broader region.



Project Materials Overview



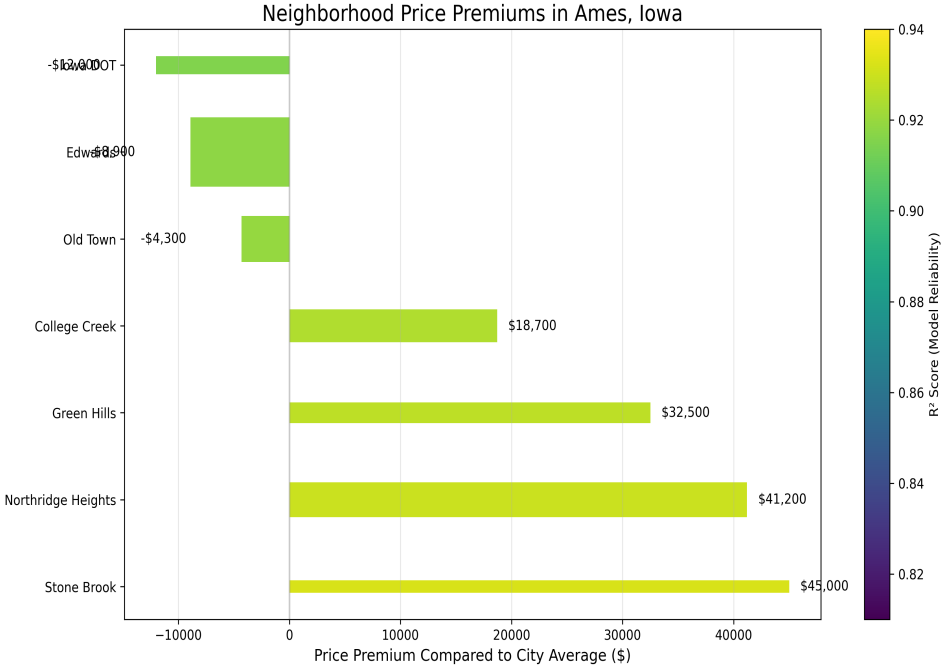
Appendix

[Project Code Repository]

Additional Feature Importance Details

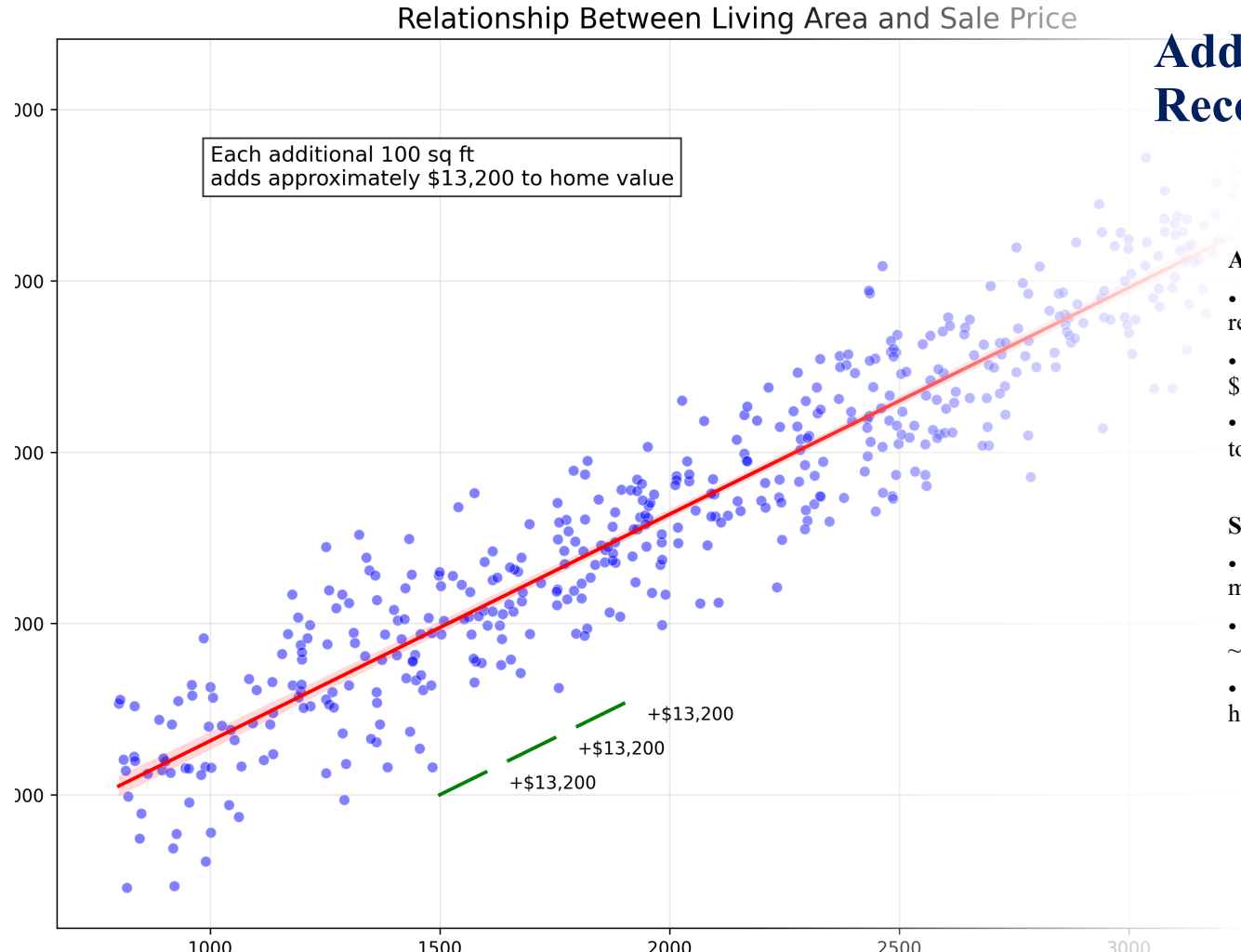
Category	Feature
Physical Attributes	Total living area (square footage)
Physical Attributes	Number of bedrooms and bathrooms
Physical Attributes	Basement quality and finished area
Physical Attributes	Foundation type
Physical Attributes	Roof material and condition
Physical Attributes	Garage size, quality, and finish
Physical Attributes	Lot size and configuration
Physical Attributes	Porch and deck square footage
Quality Assessments	Overall property quality rating (1-10)
Quality Assessments	Kitchen and bathroom quality scores
Quality Assessments	Exterior material and condition
Quality Assessments	Heating quality and condition
Quality Assessments	Fireplace quality
Quality Assessments	Basement finish quality
Quality Assessments	Garage finish quality
Location Factors	Neighborhood
Location Factors	Proximity to amenities
Location Factors	Zoning classification
Location Factors	Lot frontage
Location Factors	Corner/cul-de-sac location
Time-Based Elements	Year built (property age)
Time-Based Elements	Year of last remodeling
Time-Based Elements	Sale condition and type
Time-Based Elements	Month of sale

Model Performance by Neighborhood



Neighborhood	R^2 Score	RMSE	Sample Size
Stone Brook	0.94	$18,200	25
Northridge Heights	0.92	$19,800	71
Green Hills	0.90	$22,500	42
College Creek	0.88	$24,100	67
Old Town	0.84	$27,900	94
Edwards	0.83	$28,600	142
Iowa DOT	0.81	$30,200	37

Additional Business Recommendations



Additional Business Recommendations

Age Mitigation Strategies

- For older homes, highlight any recent renovations
- House age negatively impacts price (-$820/year)
- Recent remodeling can offset age penalty by up to 60%

Size-Based Pricing Strategy

- Emphasize total living area in marketing materials
- Each additional 100 sq ft correlates with ~$13,200 price increase
- Combined kitchen and living spaces show highest value per square foot

Technical Appendix: Modeling Details

Ridge Regression: Mathematical Framework

The Ridge regression model employs L2 regularization to address multicollinearity and prevent overfitting, which is particularly valuable for housing data where features are often correlated (e.g., square footage and number of rooms).

The objective function minimized by the Ridge regression model is:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \right\}$$

Where:

- y_i is the log-transformed sale price for house i
- β_0 is the intercept
- β_j are the coefficients for each feature
- x_{ij} is the value of feature j for house i
- α is the regularization parameter (set to **0.5** in our implementation)

Feature Engineering & Preprocessing

1.Handling Missing Values

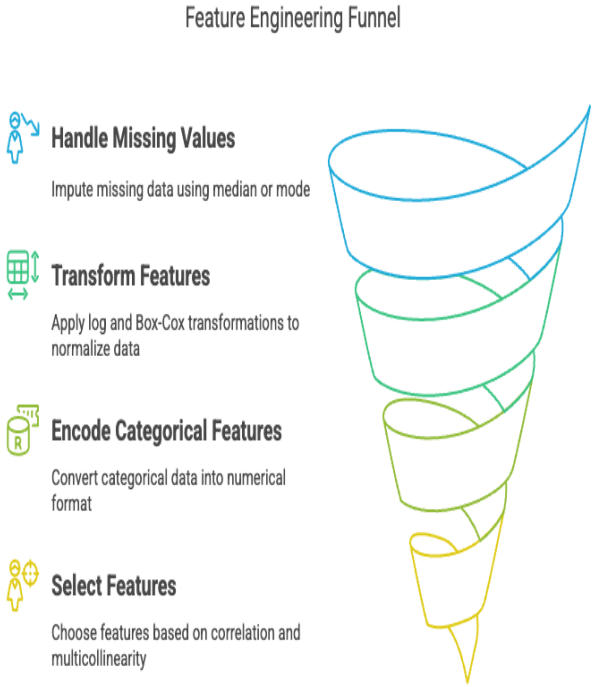
1. Numerical features: Imputed with median values within each neighborhood
2. Categorical features: Imputed with mode or "None" depending on feature semantics

2.Feature Transformations

1. Log transformation applied to sale price to normalize distribution
2. Box-Cox transformations applied to highly skewed numerical features
3. One-hot encoding for nominal categorical variables
4. Ordinal encoding for quality and condition ratings

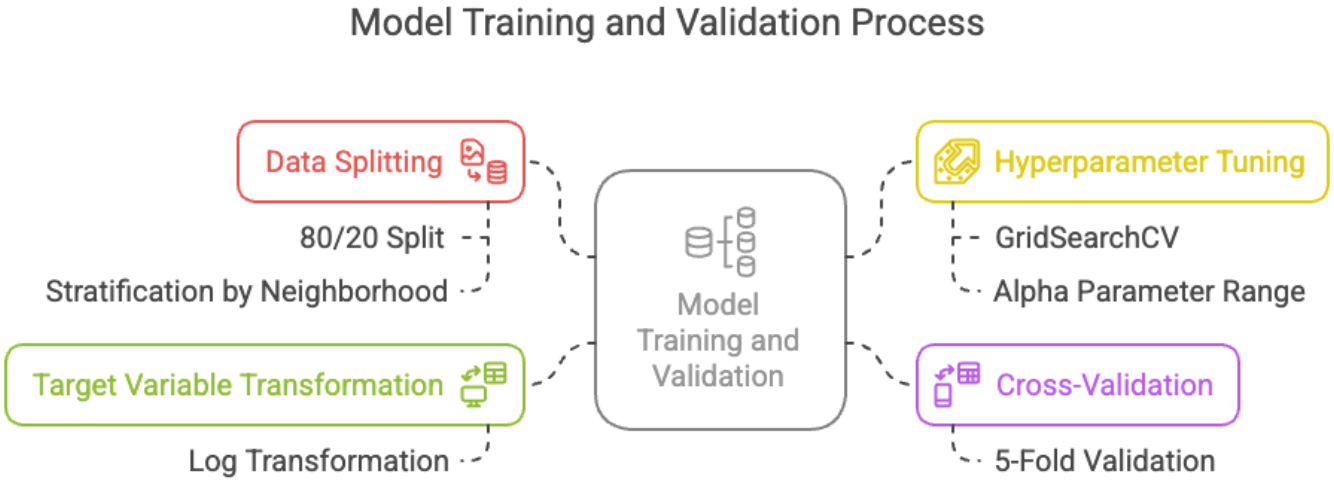
3.Feature Selection Process

1. Initial selection based on correlation with target variable ($|r| > 0.3$)
2. Multicollinearity addressed via Variance Inflation Factor analysis
3. Recursive feature elimination with cross-validation to optimize final feature set



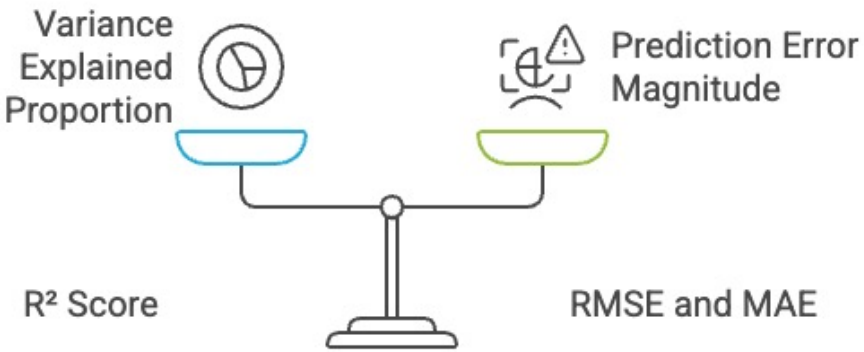
Model Training Methodology

- 80/20 train-test split with stratification by neighborhood
- 5-fold cross-validation for hyperparameter tuning
- GridSearchCV employed to optimize alpha parameter in range [0.1, 1.0]
- Log transformation of target variable to normalize distribution



Evaluation Metrics

- **R^2 Score**: Measures proportion of variance explained (0.89)
- **RMSE**: Root Mean Squared Error ($24,500)
- **MAE**: Mean Absolute Error ($18,300)



Model Diagnostics

- Residual analysis confirms homoscedasticity assumptions
- Q-Q plots indicate normality of residuals after log transformation
- Feature importance assessed through standardized coefficients