Predicting Home Sale Prices in Ames, Iowa

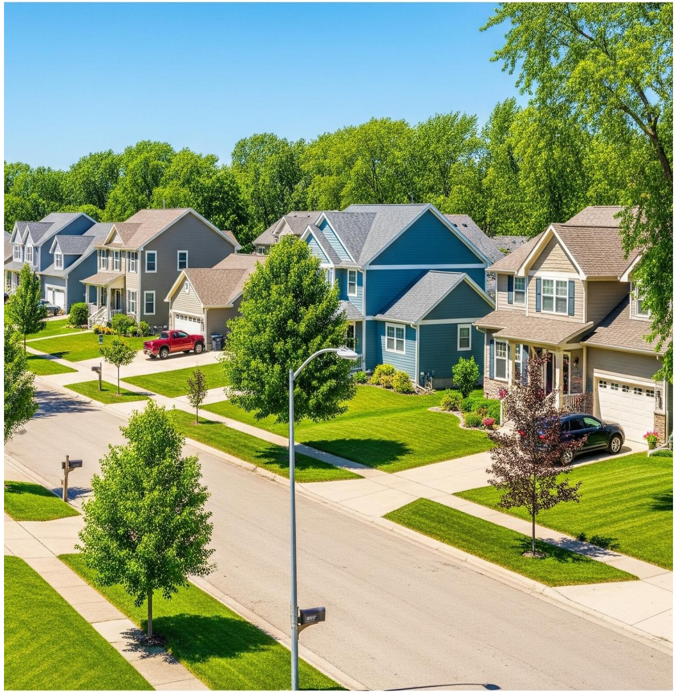
Gabriela Armenta • 05.30.2025

Outline

- **Objective & Background**
- Data Description
- Preprocessing & Exploratory Analysis
- Model Selection & Hyperparameter Tuning
- Results

Background and Objective

- Analysis is a ML project under NYC Data Science Academy.
- Objective: predict real estate sale prices and reveal their key drivers by
    - Performing feature engineering and exploratory data analysis.
    - Building predictive models to maximize R^2.
    - Interpreting model outcomes and uncover key predictors using SHAP.



Outline

- Objective & Background
- **Data Description**
- Preprocessing & Exploratory Analysis
- Model Selection & Hyperparameter Tuning
- Results

Dataset Overview

Dataset contains 2,580 house sale records from 2006 to 2010

Ames Iowa is a mid-size city with stable housing market

Dataset includes property details such as size and building type

Complied by Prof. Dean De Cock from Truman State University

Includes 79 predictors

Target variable: Sale Price

Continuous variable

US dollars

From 2006 to 2010

Feature Overview

- 43 objects and 38 numeric variables.
- Property location, zoning , utilities, and adjacent conditions.
- Building type, condition, and exterior finishes.
- Foundation, basement, living area layout and quality.
- Heating, cooling, and electrical systems characteristics.
- Garage and parking details.
- Outdoor features and amenities.
- Sale transaction information.

Outline

- Objective & Background
- Data Description
- **Preprocessing & Exploratory Analysis**
- Model Selection & Hyperparameter Tuning
- Results

Preprocessing and Exploratory Analysis

Steps

1. Initial exploration of dataset structure and content.
2. Examine pattern of missing data and impute missing values where replacements are known.
3. Encode ordered categorical features as numeric.
4. Visualize feature relationships with Sale Price.

Assessment and Imputation of Missing Data

- For 23 variables missing values signified true absence (e.g., no basement, no pool) and were recorded as 0 or None.
- Eight variables had only one or two missing entries (<0.01% of observations) and were dropped because imputation added negligible value.
- Masonry veneer type (10+ missing), masonry veneer area (20+ missing), and lot frontage area (460+ missing), were imputed using mean or median values as appropriate.

Ordinal Encoding and Data Pruning

Ordinal mapping

- Mapped values of 21 ordinal categorical variables to numeric codes where higher value meant better (e.g., heating quality scored from 0=poor to 4=excellent).

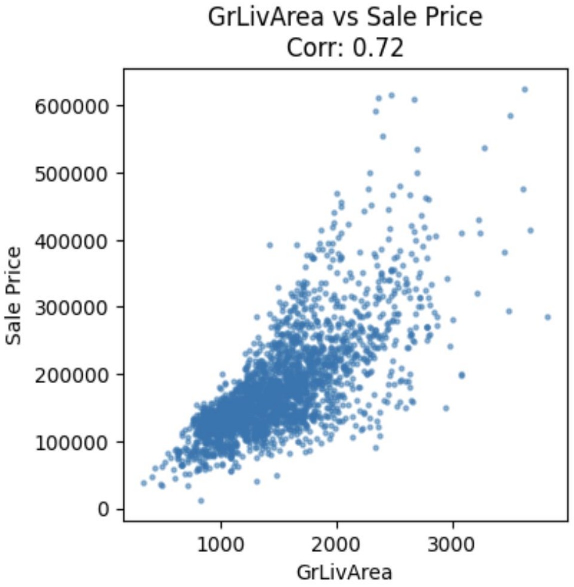
Dropped Ground Living Area > 4000

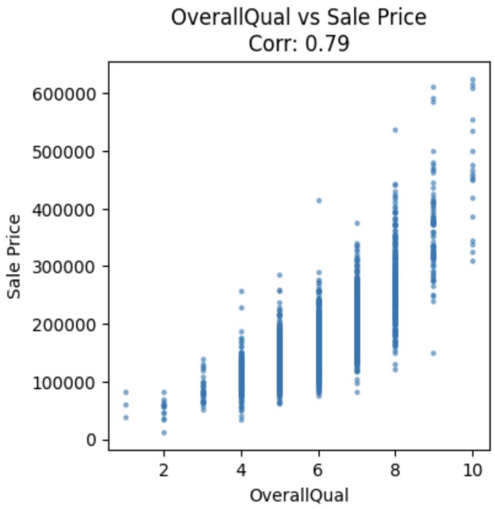
- Dropped outliers and unusual sales per dataset creator, Dean De Cock’s recommendation.

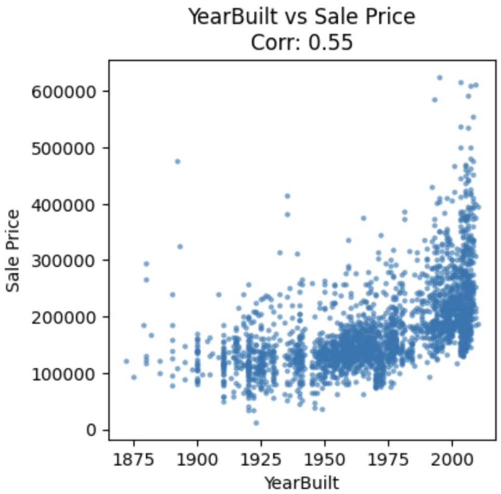
Drop Features with close to zero variance

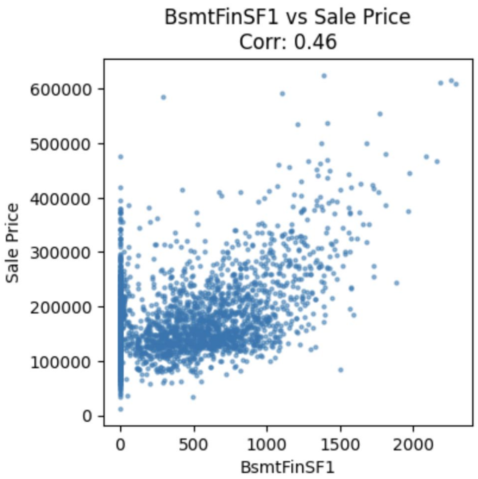
- Removed near-constant features (98% or more identical values) because their lack of variation provides negligible signal for learning relationships.

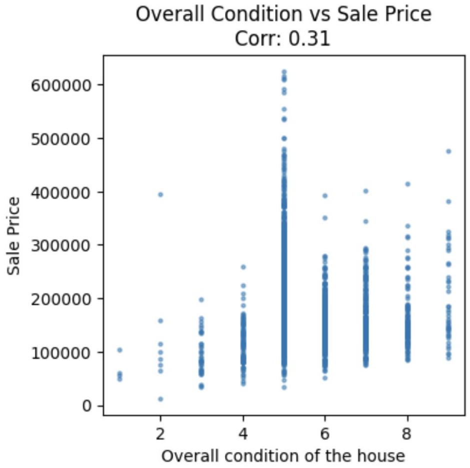
Exploratory Data Analysis of Numeric and Ordinal Features


GrLivArea vs Sale Price
Corr: 0.72

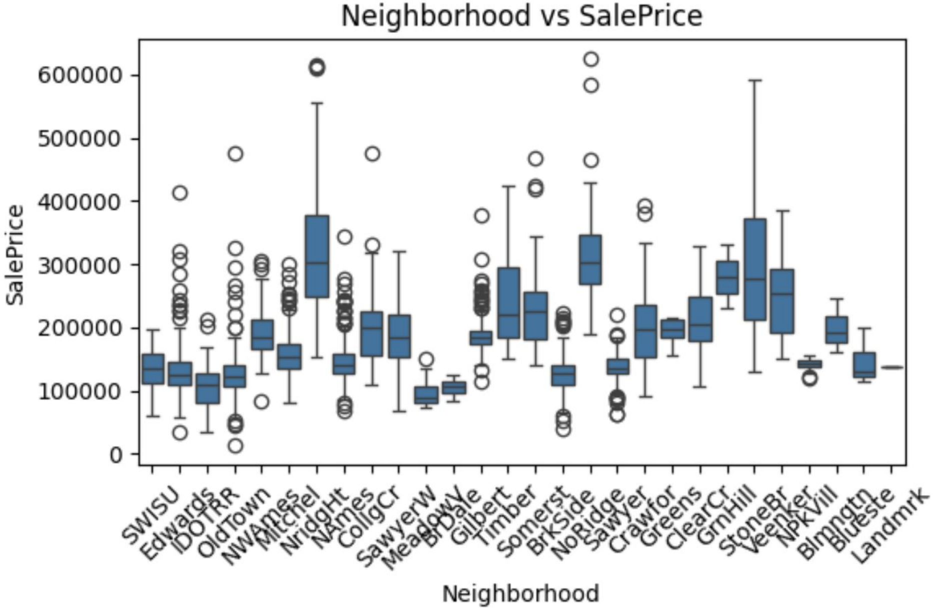

OverallQual vs Sale Price
Corr: 0.79

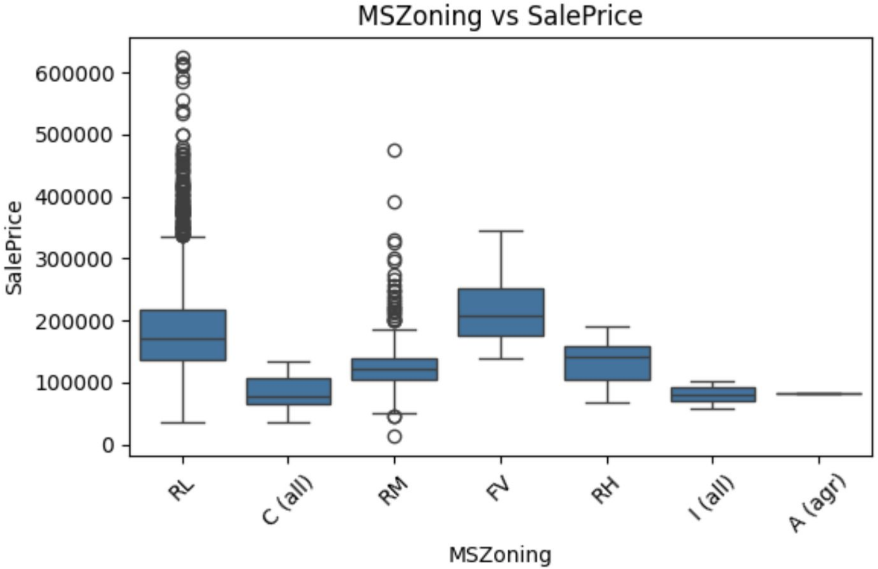

YearBuilt vs Sale Price
Corr: 0.55


BsmtFinSF1 vs Sale Price
Corr: 0.46

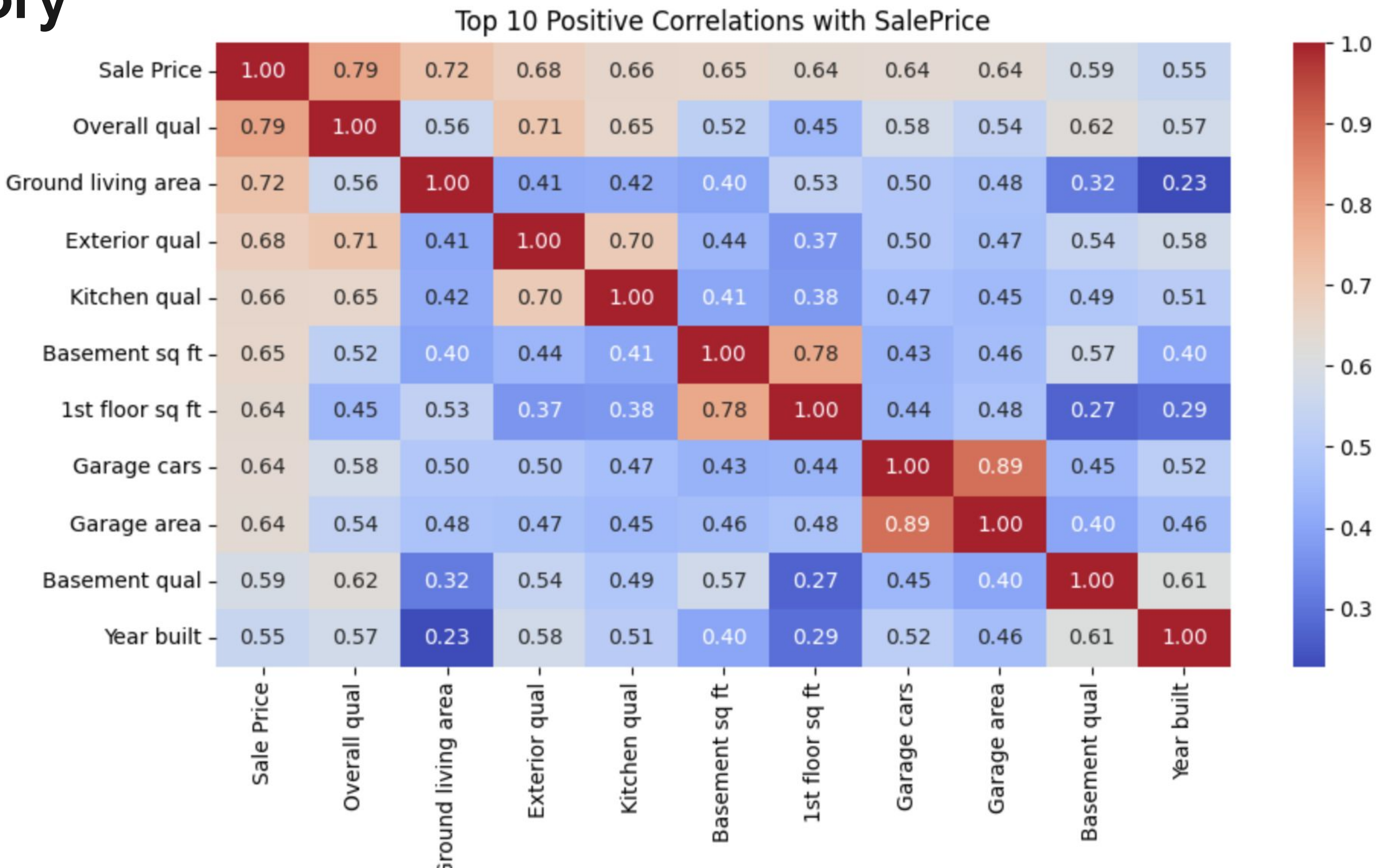

Overall Condition vs Sale Price
Corr: 0.31

Exploratory Data Analysis of Categorical Nominal Features





Exploratory Data Analysis: Heatmap



Outline

- Objective & Background
- Data Description
- Preprocessing & Exploratory Analysis
- **Model Selection & Hyperparameter Tuning**
- Results

Pipelines and Hyperparameter Tuning

Preprocessing pipelines

1. Identified numeric, categorical, and dummy columns.
2. Numeric processing:
   - Skewed: imputed median and MinMax scaling (0 and 1).
   - Symmetric: imputed mean and Standard scaling.
   - Dummies: none were present in the dataset.
3. Categorical processing:
   - Nominal: imputed mode and converted to dummies using one-hot encoding.
   - Ordinal: mapped to numeric codes. After that applied numeric processing.

Pipelines and Hyperparameter Tuning

Preprocessing pipeline

- Combined all feature-specific pipelines into a ColumnTransformer

Model Pipelines

- Created pipelines for random forest, XGBRegressor, and Elastic Net.

Hyperparameter Tuning

- Used Optuna to optimize each model’s parameters

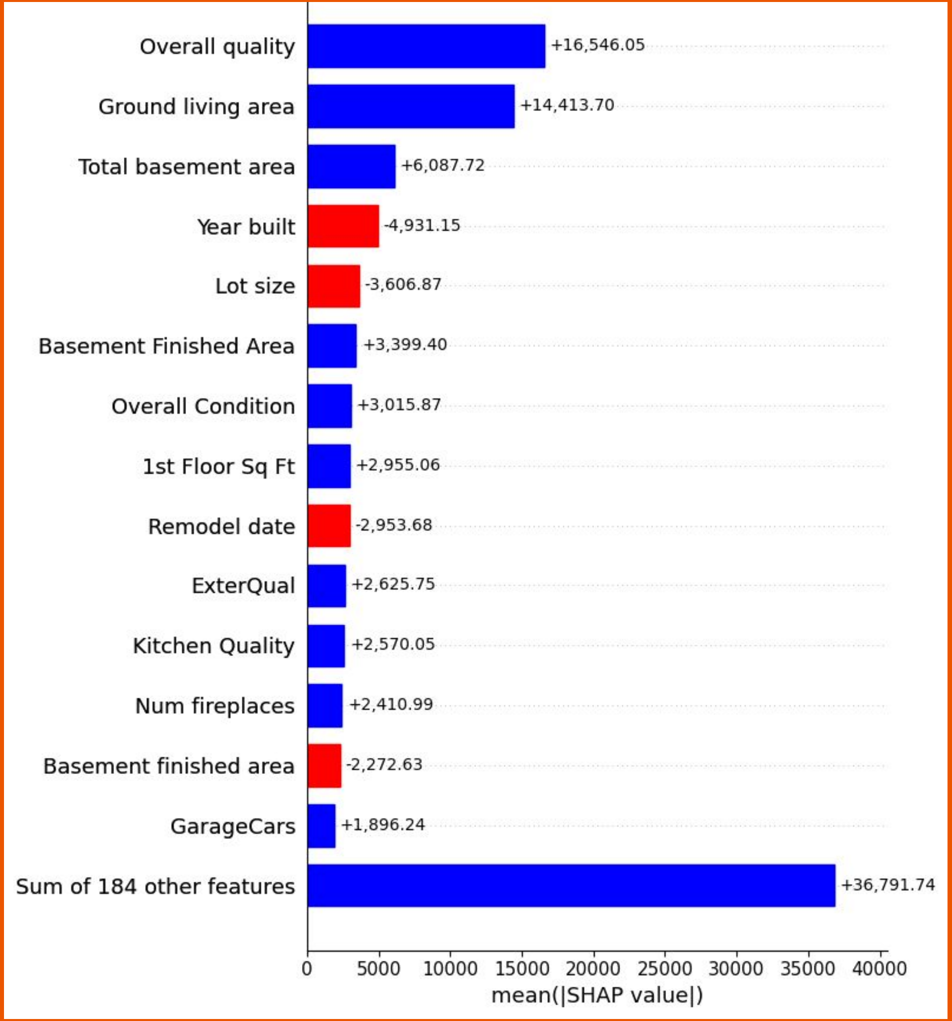
Outline

- Objective & Background
- Data Description
- Preprocessing & Exploratory Analysis
- Model Selection & Hyperparameter Tuning
- **Results**
- Conclusions and Next Steps

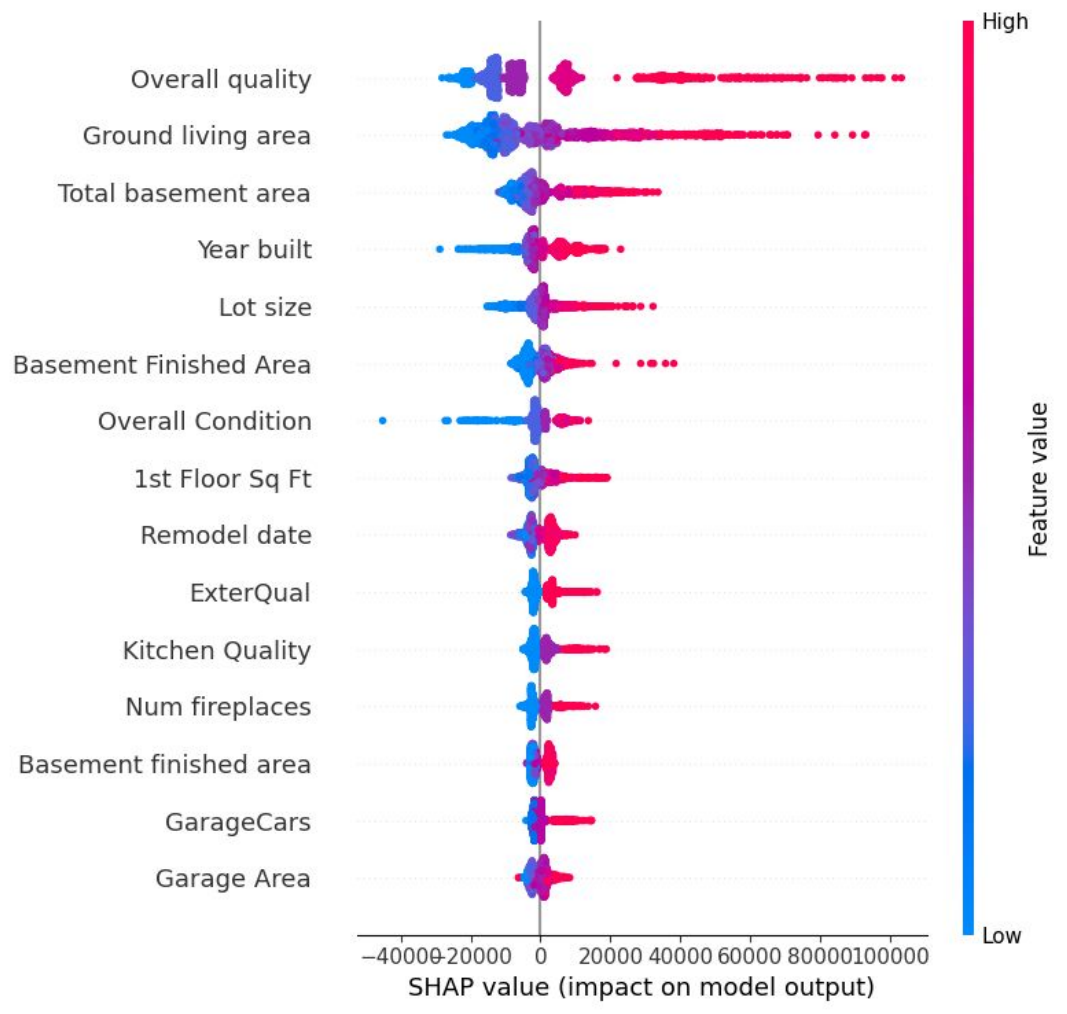
Results: Metrics

Random Forest	Elastic Net	XGBoost Regressor
MSE: 569281441.24	MSE: 496402737.18	MSE: 358492032.00
RMSE: 23859.62	RMSE: 22280.10	RMSE: 18933.89
MAE: 14474.70	MAE: 14937.00	MAE: 12216.03
R^2: 0.8968	R^2: 0.9100	R^2: 0.9350

SHAP Values


Overall quality
+16,546.05
Ground living area
+14,413.70
Total basement area
+6,087.72
Year built
-4,931.15
Lot size
-3,606.87
Basement Finished Area
+3,399.40
Overall Condition
+3,015.87
1st Floor Sq Ft
+2,955.06
Remodel date
-2,953.68
ExterQual
+2,625.75
Kitchen Quality
+2,570.05
Num fireplaces
+2,410.99
Basement finished area
-2,272.63
GarageCars
+1,896.24
Sum of 184 other features
+36,791.74
0
5000
10000
15000
20000
25000
30000
35000
40000
mean(|SHAP value|)

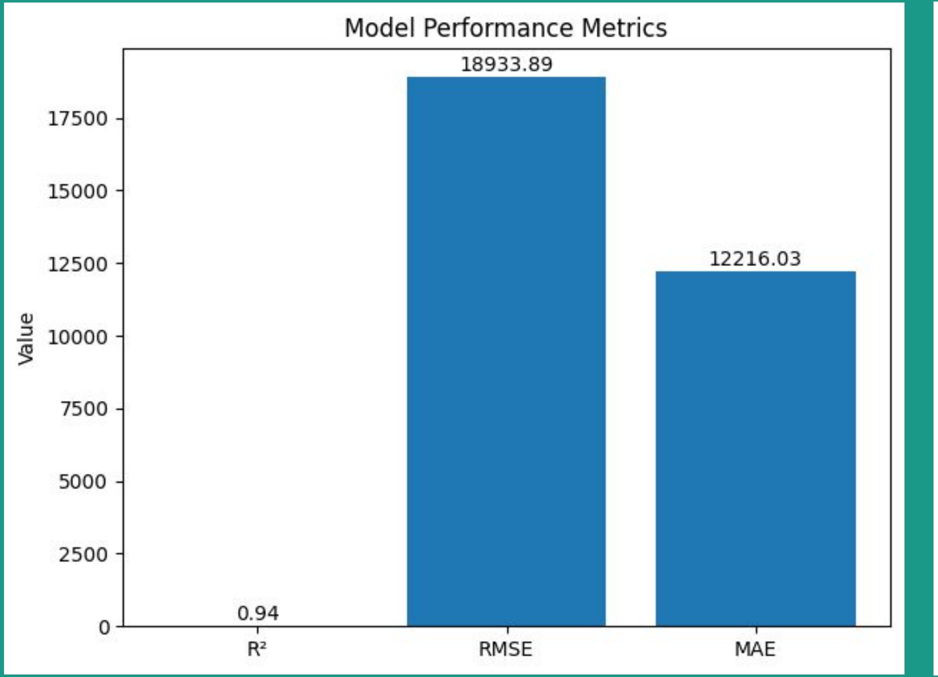
SHAP Values



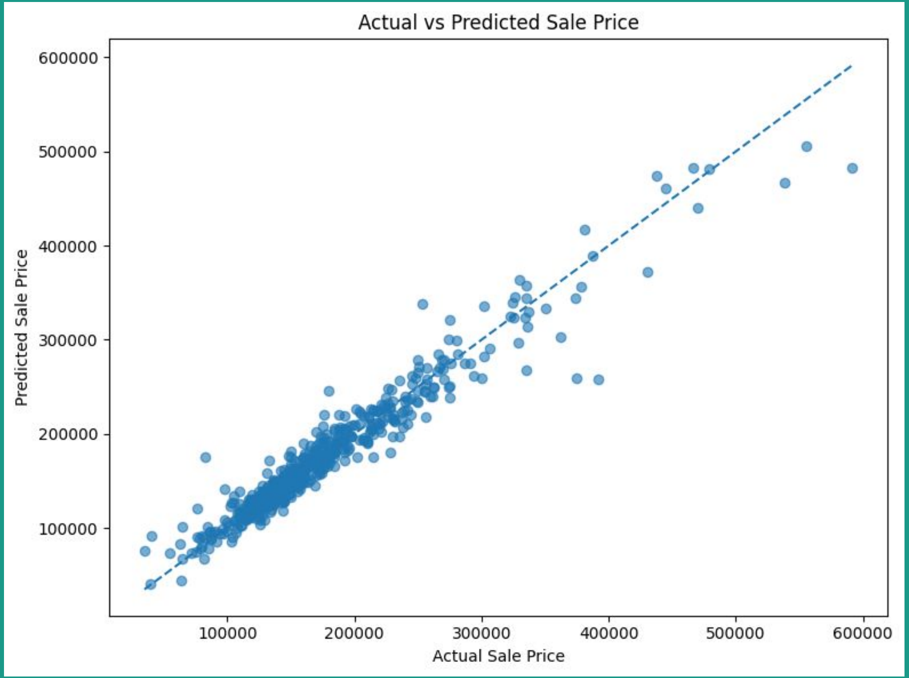
Overall quality
Ground living area
Total basement area
Year built
Lot size
Basement Finished Area
Overall Condition
1st Floor Sq Ft
Remodel date
ExterQual
Kitchen Quality
Num fireplaces
Basement finished area
GarageCars
Garage Area
−40000
−20000
0
20000
40000
60000
80000
100000
SHAP value (impact on model output)
High
Low
Feature value

Model Performance Plots

Performance Metrics



Actual vs. Predicted Sales Price



Outline

- Objective & Background
- Data Description
- Preprocessing & Exploratory Analysis
- Model Selection & Hyperparameter Tuning
- Results
- **Conclusions and Next Steps**

Conclusions

Overall quality was the top contributor to predictions of Real Estate Sale Price

XGBoost Regressor was the best performing algorithm

- Model explained 94% of the variability in the data.

Top 15 contributors to predictions focused on quality, size, age, and special features:

- **Quality and condition indicators** (overall , exterior, kitchen, condition) which indicate how “nice” the home is built or maintained.
- **Size and area metrics** (living area, basement, garage, lot) which describe how much usable space there is, both indoors and outdoors.
- **Age and updates** (year built, remodel date) give a sense of how old the home is and when it last received a major upgrade.
- **Features and amenities** (fireplaces, garage capacity) point to specific selling points that can drive buyer interest.

Conclusions: How home buyers and sellers could leverage these insights

- **Set Pricing Strategically:** Base list prices on measurable attributes (e.g., overall quality, living area, basement) instead of relying mostly on neighborhood comparables.
- **Highlight Key Features in Marketing:** Emphasize high-impact elements like kitchen quality, finished basement, fireplaces, and garage capacity in listings and ads.
- **Advise Sellers on Cost-Effective Upgrades:** Recommend targeted improvements (e.g., refreshing exterior finishes or updating the kitchen) that boost value rather than broad renovations.
- **Guide Staging and Presentation:** Showcase square footage (open living areas, functional basements) and quality details (finishes, fixtures) since these drive perceived value.

Conclusions: How home buyers and sellers could leverage these insights

- **Tailor Buyer Matching:** Filter listings by attributes buyers care about (e.g., move-in readiness, ample garage space) to connect them with the right homes.
- **Negotiate with Data:** Use objective measures (e.g., year built vs. remodel date) to support price adjustments, credits, or repair requests.
- **Allocate Time and Resources Efficiently:** Focus showings on properties whose high-impact features align with buyer priorities instead of low-predictive locations.

Next steps

- Use a larger dataset to build models for the US and not only for Iowa.

- Build an AI platform that can have interactive dashboards and an agent guiding decisions.