

Ames Housing Sale Price Forecast

Team 3:

Jeffrey
Joel
Cathy

Content

- Intro & background information
- Problem Statement
- EDA & Exploratory Visualizations
- Pre-processing
- Feature Engineering
- Modeling
- Hyperparameter
- Results & Visualization
- Business Recommendations

Intro & background information

In July 2022, the median listing home price in Ames, IA was \$299K, trending up **10.8% year-over-year**. The median listing home price per square foot was \$200. The median home sold price was **\$304.5K**.

There are 41 neighborhoods in Ames. **Northridge** has a median listing home price of \$400K, making it the most expensive neighborhood. Hodge is the most affordable neighborhood, with a median listing home price of \$195.5K.

On average, homes in Ames, IA sell after **57 days** on the market.

Source: realtor.com

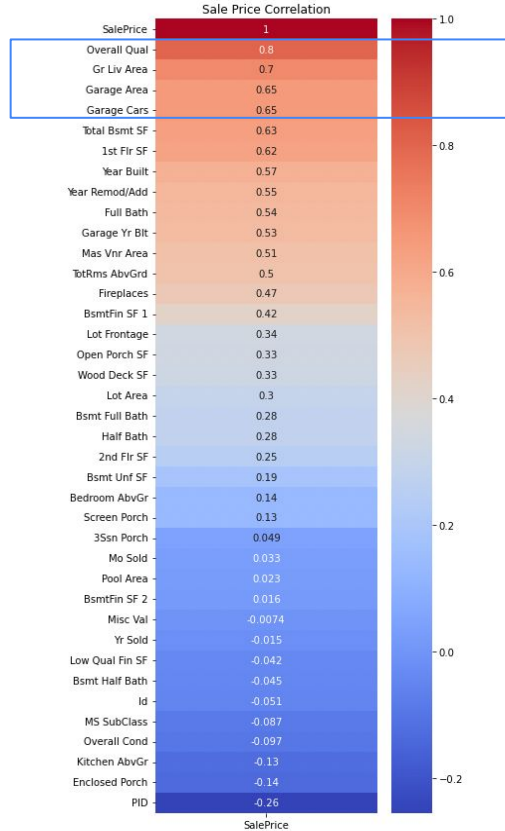
Median Listing Home Price vs. Median Home Sold Price



Problem Statement

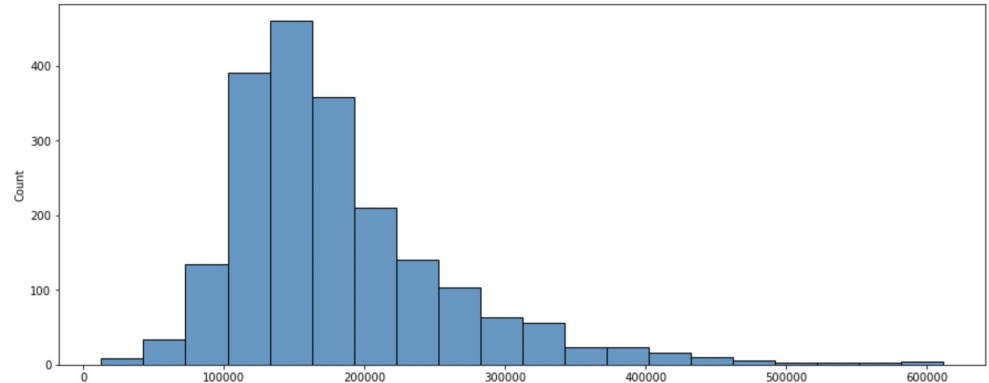
- Stakeholders: Realtor.com have requested the team to come up with a method to predict housing prices in Ames, Iowa.
- Problem Statement: We aim to provide a model that will allow investors to know which features are the most weighted in determining housing prices in Ames, Iowa.

EDA & Exploratory Visualizations

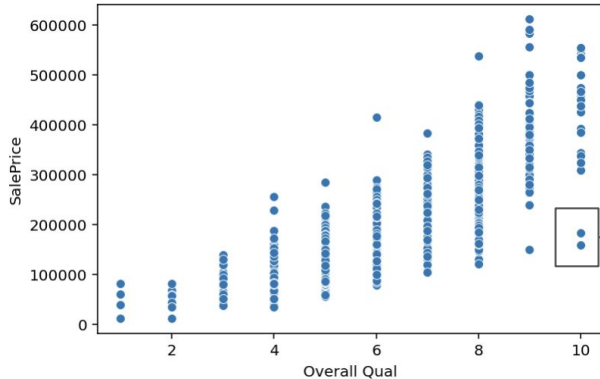
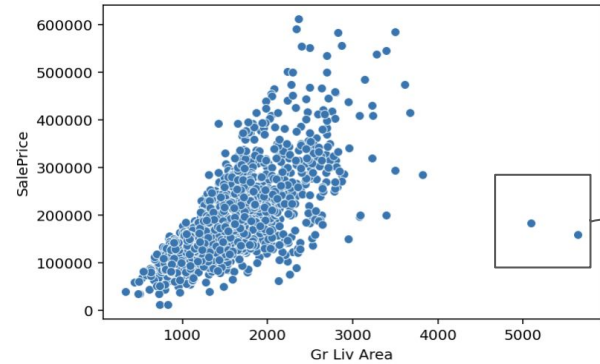


- Top correlated features (≥ 0.65) includes Quality, Gross Living Area & Garage

Distribution of Sale Price



EDA & Exploratory Visualizations



- Outliers are removed for better data representation.

Pre-processing

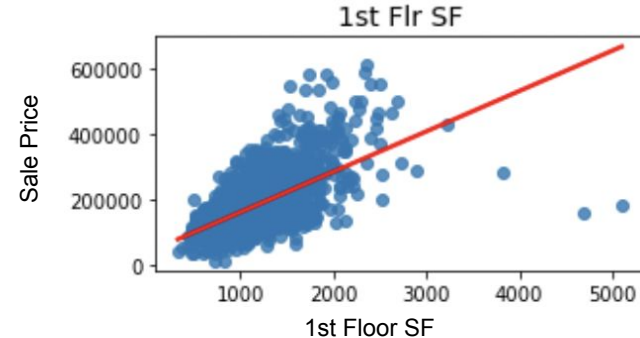
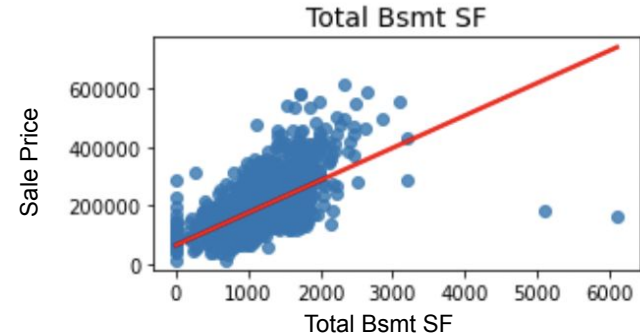
- Impute for missing data (26 Features)
- Dummify the categorical columns (> 40 categorical features)
- Train/test split the dataset
- Scaling the features (Standard Scaler)

	missing_count	%_missing_count
Pool QC	2042	99.561190
Misc Feature	1986	96.830814
Alley	1911	93.174061
Fence	1651	80.497318
Fireplace Qu	1000	48.756704
Lot Frontage	330	16.089712
Garage Finish	114	5.558264
Garage Qual	114	5.558264
Garage Yr Blt	114	5.558264
Garage Cond	114	5.558264
Garage Type	113	5.509508
Bsmt Exposure	58	2.827889
BsmtFin Type 2	56	2.730375
BsmtFin Type 1	55	2.681619
Bsmt Cond	55	2.681619
Bsmt Qual	55	2.681619
Mas Vnr Area	22	1.072647
Mas Vnr Type	22	1.072647
Bsmt Half Bath	2	0.097513
Bsmt Full Bath	2	0.097513
Garage Area	1	0.048757
Total Bsmt SF	1	0.048757
Bsmt Unf SF	1	0.048757
BsmtFin SF 2	1	0.048757
BsmtFin SF 1	1	0.048757
Garage Cars	1	0.048757
Mo Sold	0	0.000000

zero_values	
Pool Area	2042
3Ssn Porch	2025
Low Qual Fin SF	2018
Misc Val	1986
Bsmt Half Bath	1923
Screen Porch	1870
BsmtFin SF 2	1803
Enclosed Porch	1724
Half Bath	1308
Mas Vnr Area	1216
Bsmt Full Bath	1200
2nd Fir SF	1191
Wood Deck SF	1075
Fireplaces	1000
Open Porch SF	912
BsmtFin SF 1	657
Bsmt Unf SF	165
Garage Area	113
Garage Cars	113
Total Bsmt SF	54
Full Bath	8
Bedroom AbvGr	5
Kitchen AbvGr	2
Id	0

Feature Engineering

- Combined area features into 1 Total Area feature:
 - Total Bsmt SF
 - 1st Flr SF
 - 2nd Flr SF
 - Garage Area
 - Pool Area
 - 3Ssn Porch
 - Screen Porch
 - Wood Deck SF
 - Open Porch SF
 - Enclosed Porch
 - Mas Vnr Area
 - Low Qual Fin SF
- High correlation against Sale Price of 0.86



Modeling - baseline & production model

- OLS Model (Baseline)
- Ridge Model
- Lasso Model

R ² Score	OLS	Ridge	Lasso
Training	0.949	0.915	0.905
Testing	Close to 0 (too much noise)	0.774	0.884

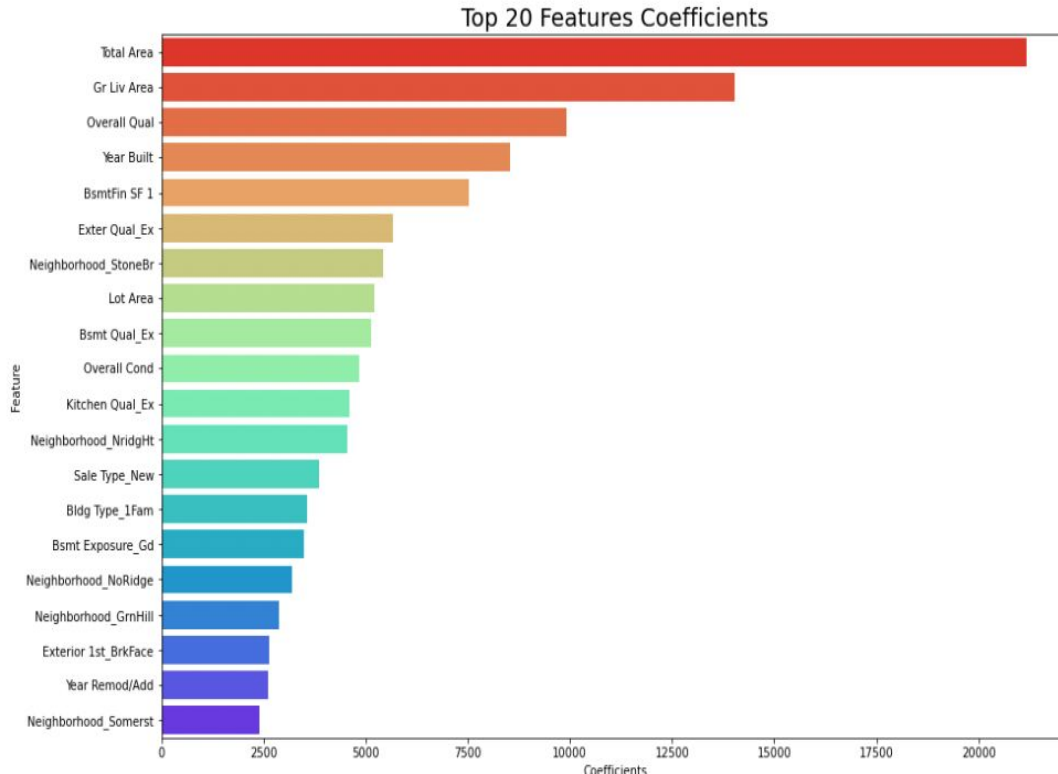
Hyperparameter tuning

Iteration	Range of alpha	CV	Optimal value of alpha	R ² (X_train, y_train)	R ² (X_test, y_test)
1	logspace (0,10,100)	3	849.75	0.9066	0.8822
		4	849.75	0.9066	0.8822
		5	849.75	0.9066	0.8822
2	logspace (2,3,100)	3	911.16	0.9048	0.8834
		4	849.75	0.9066	0.8822
		5	792.48	0.9082	0.8808
3	logspace (2.8,3,100))	3	919.68	0.9045	0.8835
		4	845.81	0.9067	0.8821
		5	799.89	0.9080	0.8809
4	linspace (915,925,1000)	3	920.97	0.9045	0.8835
		4	846.23	0.9067	0.8821
		5	803.22	0.9079	0.8810

*max_iter=50000 for all the experiments

- Parametric experiment on range of alphas and number of folds
- From logspace to linspace, to start with a larger range, and zoom in to fine tune alpha around the optimal obtained in previous iterations
- The R² values virtually stabilized since iteration 3 -> model saturated

Results & Visualizations



Areas and Quality has the highest coefficients in Lasso Model -> strongest features

Neighborhood Stone Brook and Northridge Heights - most valuable locations in real estate -> North of Ames.

Results & Visualizations



- Sale Price for top 3 neighborhood more than \$250K: NorthRidge, Northridge Heights, Stone Brook.

Business Recommendations

Strategy to maximize return in real estate investment

- By calculating the MAE and RMSE, our price estimator have a deviation of 8.6% off the actual Sale price
- Stakeholders should look to buy undervalued houses, that are listed more than 8.6% below the predicted price

Strategy to formulate optimum sale / rent price

- Houses in Northridge fetch the highest price.
- Location near Ada Hayden Heritage Park Lake or City center, Somerset

Important features that should influence decision making (Quantifiable ie. Area, Category types)

- General living area
- Year Built
- Lot Area
- Total Basement square feet
- Neighbourhood
- Basement Type 1 finished square feet