Predicting Ames Iowa House Prices

GA DSI Project 2

Julia Kelman

How can we get the best predictions?

Problem Statement

Real estate, Rental & Leasing :

4 Industry contributing to Iowa Gross Domestic Product

We are a **property management company** looking to expand to the Iowa market.

The **Portfolio feature** for our App allows individuals who own real estate properties to **track the value of their homes**.

We need to predict house prices with the highest level of accuracy for the portfolio feature of our App.

We plan to **solve** this problem by using the **Ames Housing Dataset** to build a **regression model** able to predict house prices with the highest R^2 and lowest RMSE.

This model will **inform** which **features** our employees should record during their monthly **inspections**.

Data:
The origin

From: Ames Iowa Assessor's Office

Assessment of **individual residential properties** sold in Ames, IA from **2006** to **2010**

2930 Observations
81 Variables

Workflow

Data Cleaning

- Nulls
- Correcting data types

Feature Prep

- Reformating non-numeric features
- Feature engineering
- Dummying categorical features

EDA

- VIsualization to identify trends

Model Prep

- Determining X, y
- Log y
- Data scaling

Modeling

- Baseline model
- Linear Regression, Lasso, Ridge

Model Selection

- Calculating R^2, RMSE

Model Evaluation

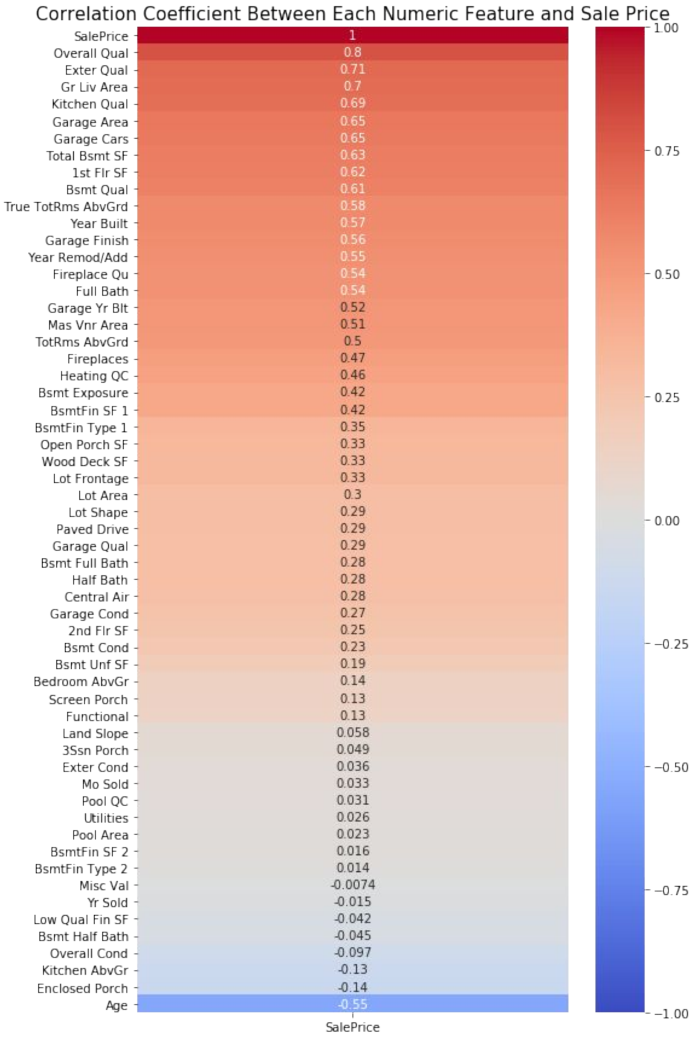
- Visualization

Recommendions

EDA part 1:

Top 10 Features with strongest correlation to Sale Price:

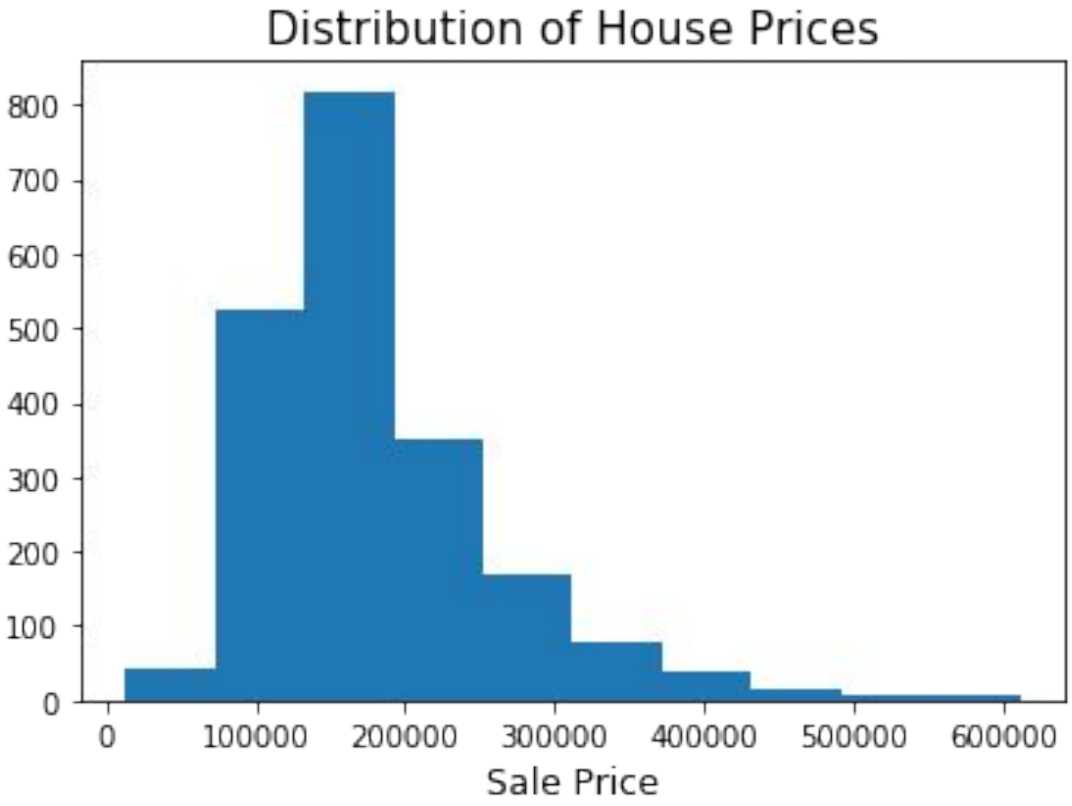
1. Overall Quality
2. External Quality
3. Above ground living area
4. Kitchen Quality
5. Garage Area
6. Garage Cars
7. Total Basement Sq Ft
8. 1st Floor Sq Ft
9. Basement Quality
10. True Total Rooms Above Ground



EDA part 2:

Investigating Sale Price Distribution

- Right Skew
- Take the log of sale price in our model



Modeling

X: Every feature except Id, PID, Sale Price, and features included in engineered feature

Y: log(Sale Price)

75 Features:

- **73** Original
- **2** Engineered:
  - Age (Yr sold - Yr remodeled)
  - True Total Rooms Above Ground (TotRms Abv Gr + Full Bath + Half Bath)

4 Models:

- Baseline
- Linear Regression
- Ridge
- Lasso

Model Selection

Based on R^2 and RMSE

	Baseline	Ridge	Lasso
R-squared	Train: 0.0 Test: -0.0136	Train: 0.9297 Test: 0.8960	Train: 0.9256 Test: 0.9141
RMSE	Train: 79,558 Test: 85,053	Train: 19,213 Test: 26,091	Train: 18,972 Test: 23,558

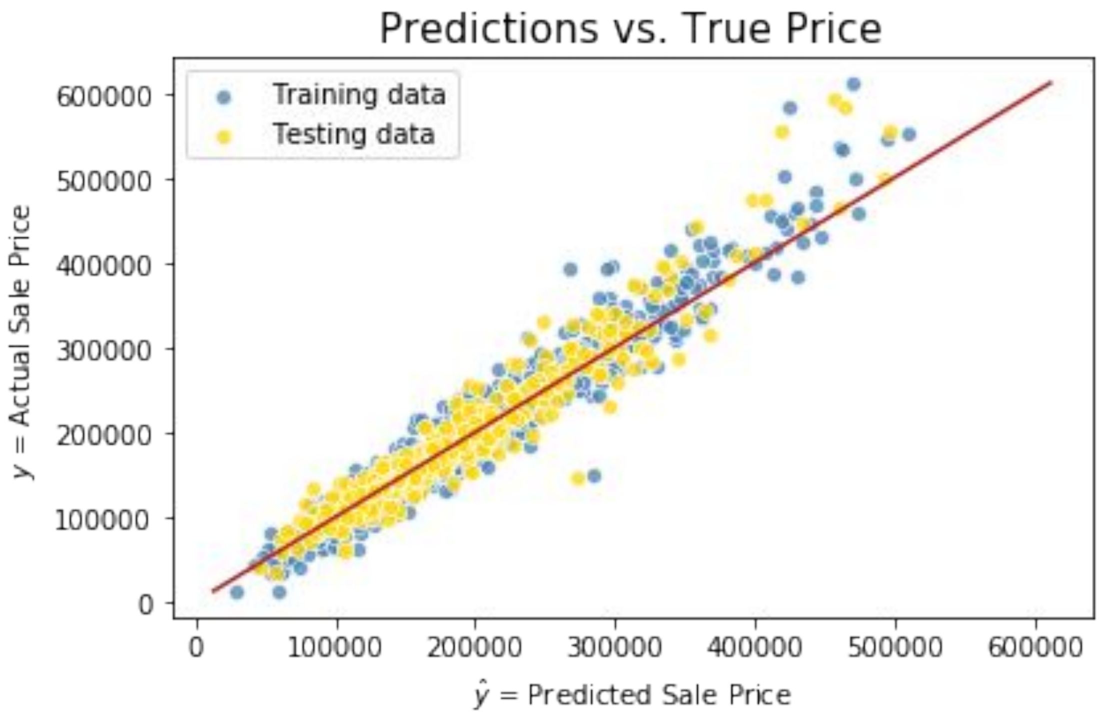
Model Selection

Based on R^2 and RMSE

	Baseline	Ridge	Lasso
R-squared	Train: 0.0 Test: -0.0136	Train: 0.9297 Test: 0.8960	Train: 0.9256 Test: 0.9141
RMSE	Train: 79,558 Test: 85,053	Train: 19,213 Test: 26,091	Train: 18,972 Test: 23,558

- 91% of the variation in sale price is explained by our model (relative to our baseline)
- True prices are approximately $23,500 from predicted value

Model Evaluation: Our Predictions



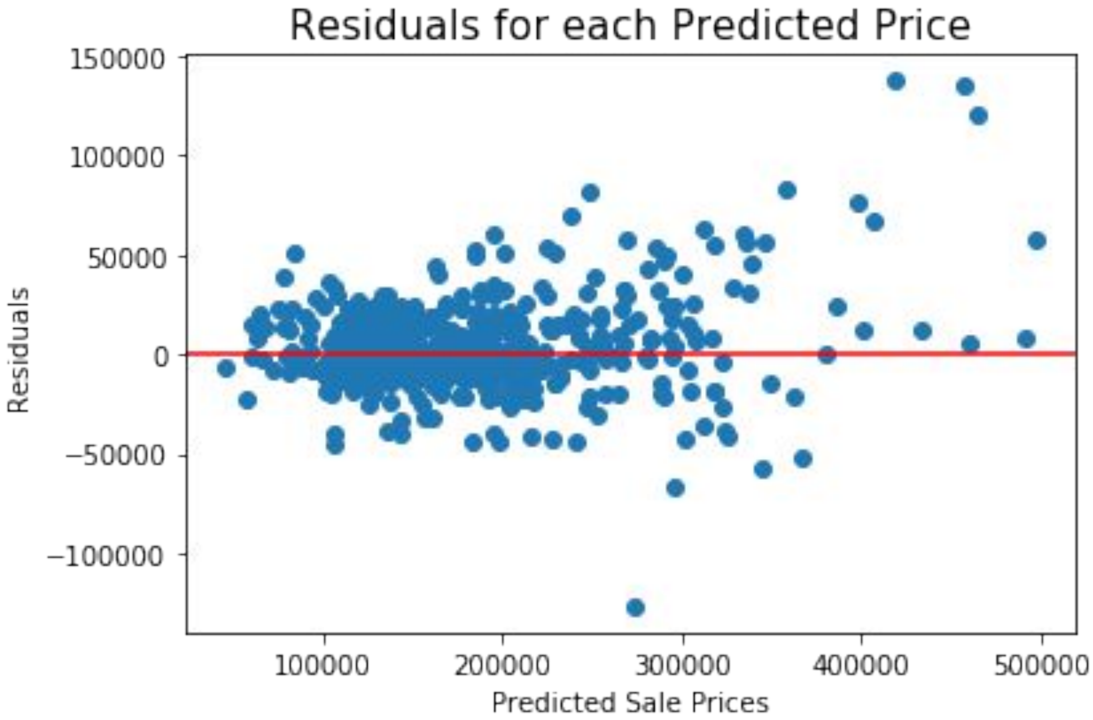
Observations:

Good predictions for both the **training** and **testing** data

Findings:

Potential for **improvement** for **higher price** range

Model Evaluation: Residuals



Observations:

Acceptable **residuals** for houses with prices **below $250,000**

Findings:

Potential for **improvement** for **higher price** range

Model Evaluation: Features

Top 5 Features
with strongest coefficients:

1. Above Ground Living Area
2. Overall Quality
3. Year Built
4. Overall Condition
5. Finished Basement Sq Ft

A 1 unit increase in those features leads to the biggest expected increase in sale price
(all else held constant)

Conclusion

Lasso Regression model with **all features** included gives us the **highest predictive power**.

Recommendation

Create **procedure** for employees to **collect data on those variables** during visits.
Only **17** features are **subject to change** after initial assessment.

References

1. https://www.iowadatacenter.org/quickfacts
2. http://jse.amstat.org/v19n3/decock/DataDocumentation.txt