Features Selection for Ames Housing Price

Evan
Yu Fung
Su Ying
Leon

PropSci
Price Guaranteed

Problem Statement

Can’t get a good price for your house during this recession?

3 **cheap** fixes to increase your property value

Goals

- Pay less attention to irrelevant features
- Evaluate which features have positive or negative impact on sale price

Dataset

- Records from home/building sales in Ames, IA from 2006 - 2010
- 80 pieces of building details including:
    - Years of construction, sale, and remodel
    - Neighborhood, proximity to transportation/parks & recreation
    - Building type and municipal subclass
    - Building materials for exterior, roofing, masonry
    - Number of rooms, area in sq. ft.
    - Lot details such as size, shape, incline
    - Quality and condition ratings

Challenges

- Handling quantitative and qualitative values
- Interpretation of null values
- Sparse data with too many zeroes

Methodology

EDA

Data Cleaning

Exploratory Visualizations

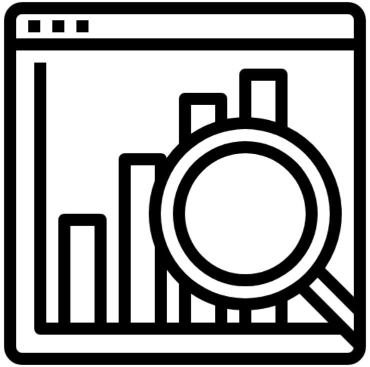
Modelling Methods

Business Recommendations

Exploratory Data Analysis

- Look at data for completeness
  - Any missing data?
  - Found an anomaly in year built

- Identified 2 outliers

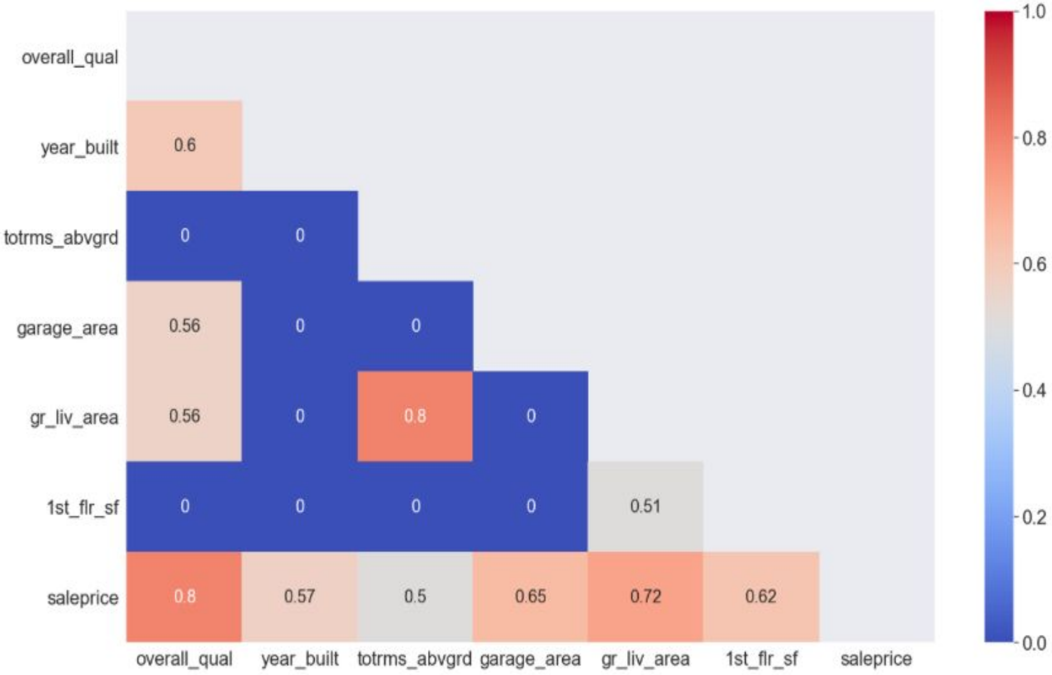


Data Cleaning

- Features are removed if they contain more than 50% null values
- Median values are given to missing quantitative values

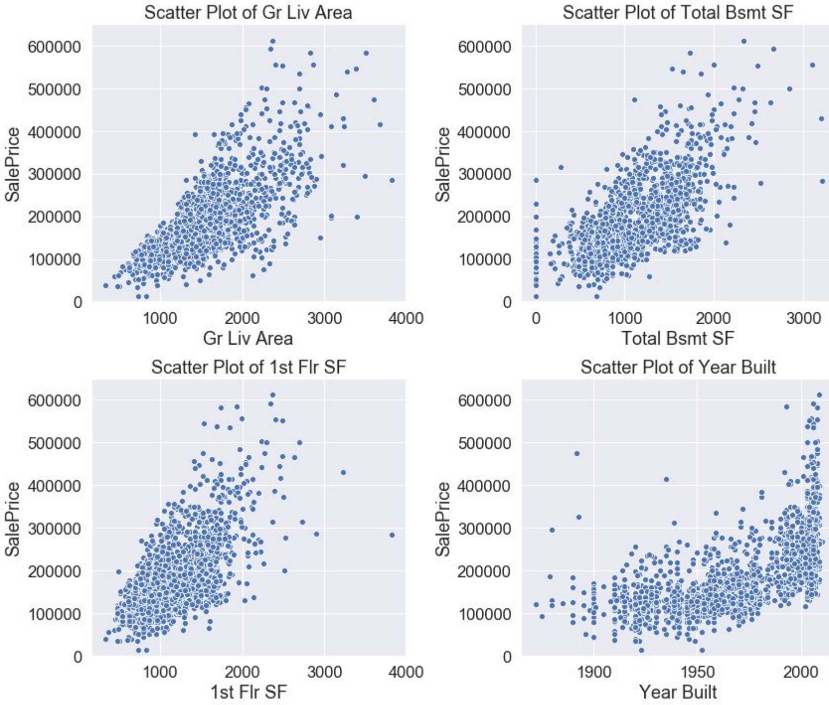
Feature Exploration

Heat map helps us visualize correlations between variables



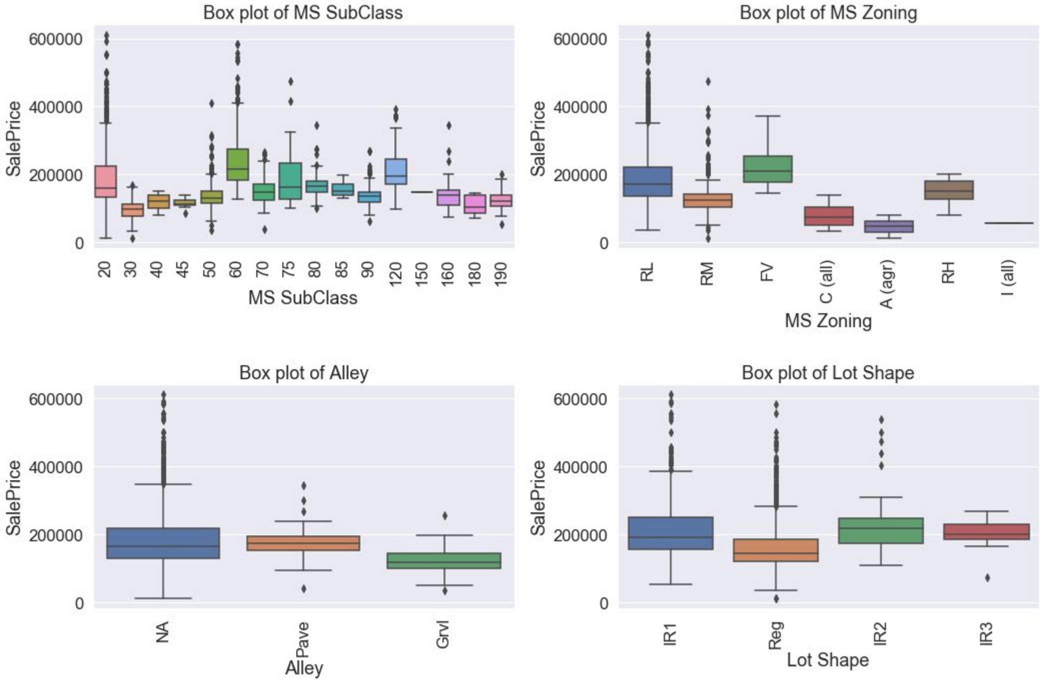
Feature Exploration

Scatter plots helps us visualize correlations between Sale Price and other numerical features.



Feature Exploration

Box plots helps us visualize relations between Sale Price and other categorical features.



Which *modeling approaches* get us the most accurate predictions?

Modelling techniques: Polynomial

- Increased Interaction terms for selected *numerical* variables
- [‘a’, ‘b’, ‘c’] -->

  [‘ab’, ‘bc’, ‘ac’, ‘a^2’, b^2’, c^2’, ‘a’, ‘b’, ‘c’]

Modelling techniques: Ordinal Values

- Assign numbers to ordered data
- [Ex, Gd, TA, Fa, Po] becomes [5, 4, 3, 2, 1]

Modelling techniques: Model Execution

- Feature scaling on numerical columns (standardize)
- Power transformation
- Train-Test-Split
- Hypertuning

Conclusion

The following features that have the most impact :

Lasso Regression model - best among other linear regression methods

Reliability: Error from model similar when applied on test data

Top features found to impact sale prices

Original 80 features are reduced to 12 features thus reducing complexity and overfitting.

Ground Living Area
Heating Quality
Fireplace Quality
Year Built
Overall Condition
Kitchen Quality
External Quality
Basement Quality
Neighborhood
Total Basement Area
Home functionality

Recommendation

- Improve kitchen quality
- Improve exterior quality
- Improve fireplace quality