Project 2: Predicting the Sale Price of Houses in Ames

Jervin Seow, Karina Kong, Tan Si Hao

Problem Statement

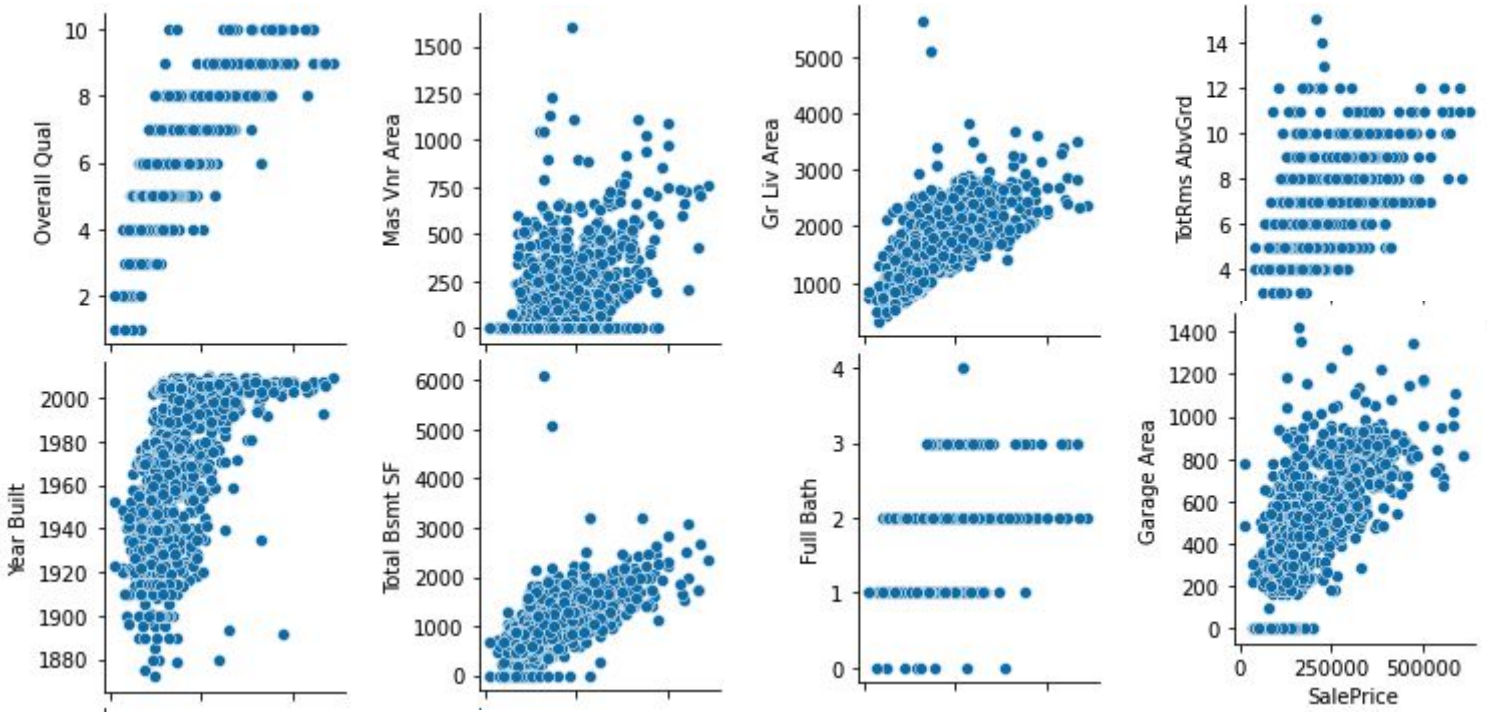
We are employees of a real estate agency in Ames, Iowa who have been tasked with doing market research for the company to find out the features of a house that are the **strongest predictors** of the sale price of the house and the **magnitude** to which these features affect the sale price of the house.

This information can help our company's agents determine a fair price range for each house and highlight the strongest features of each house they sell in order to maximise selling price for our customers.

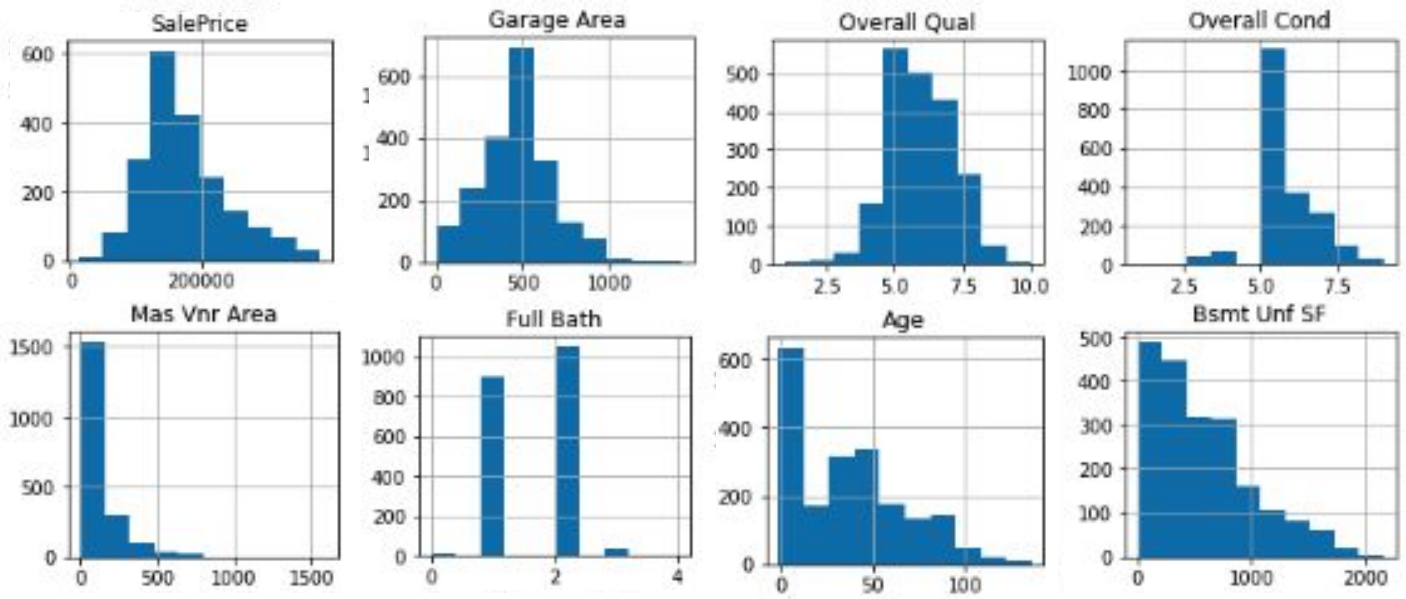
Understanding the predictors of pricing can lead to better margins

- Without time/resources/expertise to understand everything about a house, the buyer is unable to ascertain his own valuation perfectly.
- Because a buyer's resources are limited, our agents assist the price discovery process such that the buyer is able to arrive at a satisfactory valuation.
- If we are able to **highlight features that increase the perceived value and "downplay" features that decrease perceived value** - then the price discovered by the buyer will be higher than if he had total information about the house (which is impossible in reality).
- This means that we are able to **extract additional value from the buyer** by tilting the price discovery process in our favor.

EDA: Scatterplots



EDA: Distribution of numeric features



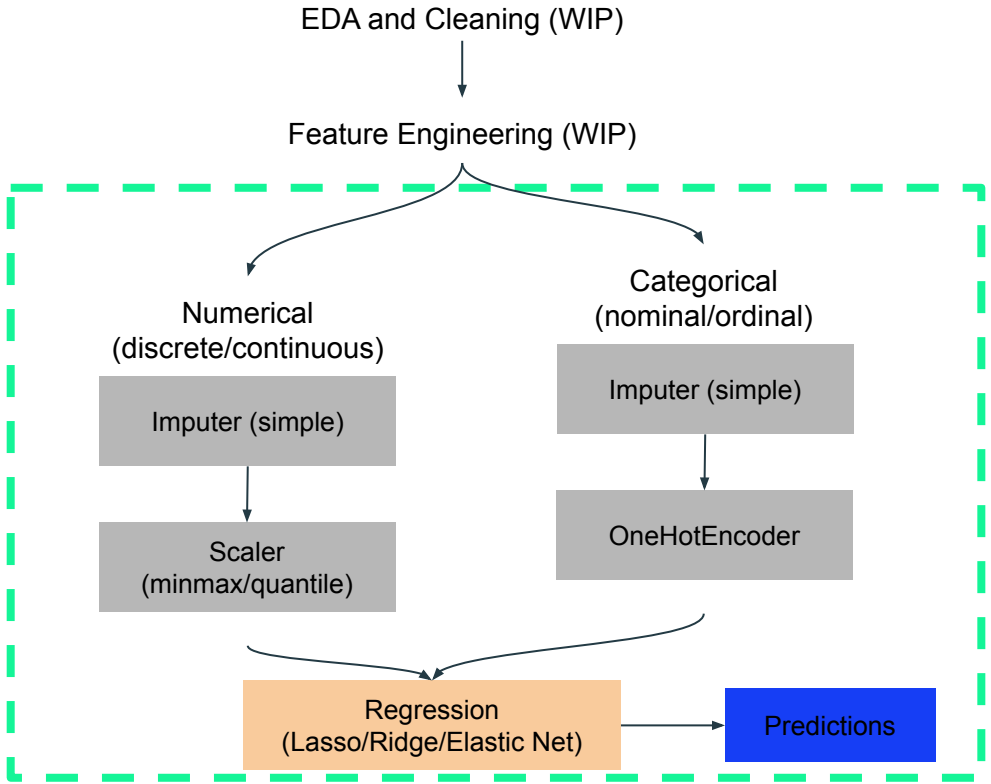
EDA: Boxplot

- Identify outliers for **SalePrice** that may affect the predictive ability of our model
- Address these outliers by
  - Removing them from the training dataset
  - Choosing a model that is robust to outliers

Model Preprocessing: Feature Selection & Engineering

1. Dropping of columns:
    a. PID (Parcel identification number)
    b. Features with very low to no correlation with sale price (i.e. < 10%)
2. Creating new features:
    a. 'age_at_sale' = 'yr_sold' - 'year_built'
    b. 'years_since_remodel' = 'yr_sold' - 'year_remod_add'
    c. Drop original features we utilised for engineering

Model Preprocessing: Imputing, Encoding and Scaling



Selecting the Final Model - Huber

5 regression models tested:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic Net Regression
5. Huber Regressor*

** A robust regularised regression that applies a l1 linear loss to samples that are classified as outliers, while inliers are subject to a less aggressive l2 loss.*

Evaluated across three criteria selected for this business problem:

1. Accuracy - Cross-validated r2 score and no overfitting
2. Interpretability of coefficients
3. Parsimony - Simplicity through less features

Model	Accuracy	Interpretability	Parsimony
Huber	**Good**	**Good**	Poor
ElasticNet	Average	Poor	**Good**
Lasso	Average	Poor	**Good**
Ridge	Average	**Good**	Poor
Linear	Poor	Poor	Poor

Model was tuned for prediction performance

Huber Regressor has two hyper parameters which were tuned via grid search:

1. Epsilon controls the number of samples that should be classified as outliers. The smaller the epsilon, the more robust it is to outliers.

2. Alpha is the regularization parameter.

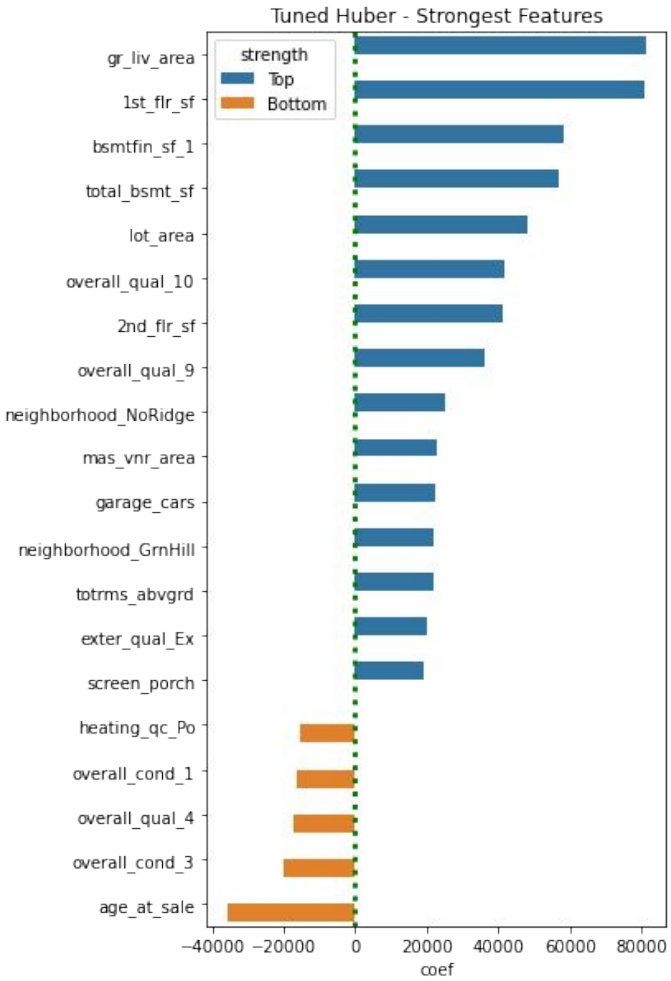
Delivers High Predictive Accuracy

0.87 Cross-validated r2

0.91 Train r2

0.93 Test r2

19752 Kaggle RMSE



Inference: Features that have the most effect on price

The regression coefficient is a measure of correlation between each feature and the response.

Positive

- Area - living area / 1st floor / basement / lot area
- Neighborhood - Northridge or Green Hills
- Garage capacity
- Total rooms
- Quality - overall quality / external equality

Negative

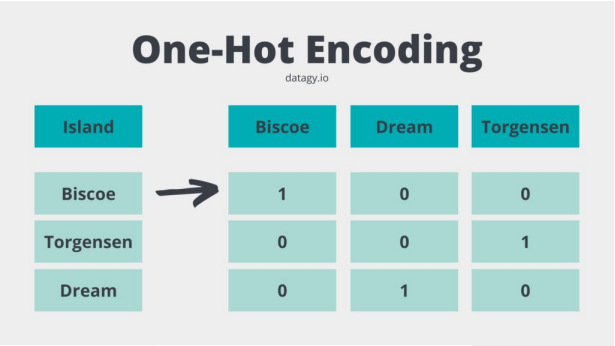
- Age at sale
- Poor overall condition
- Poor overall quality
- Poor heating quality

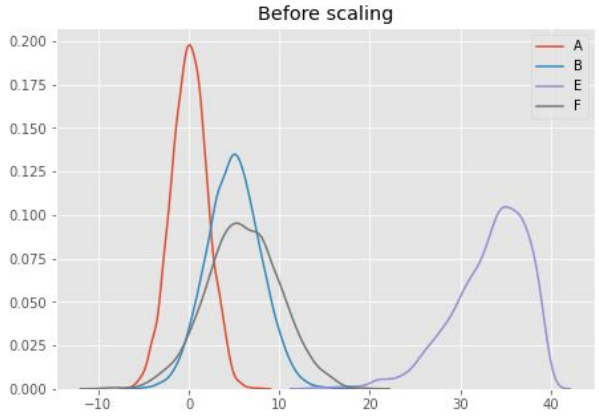
Recommendations

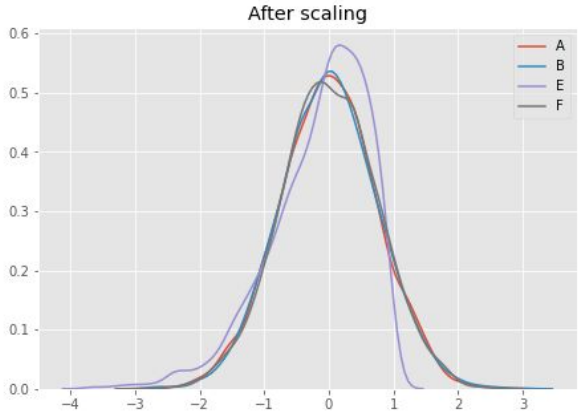
1. Features to highlight and downplay
    a. To highlight: House size, quality of the house and neighbourhood the house is located in
    b. To downplay: Age and poor quality of the house
2. Features to prioritise for refurbishment to maximise sale price
    a. The materials used to build the house have a stronger effect on Sale Price than the condition of the materials
    b. With a given budget for refurbishment, it is better to focus on material quality than condition
3. Neighborhoods to prospect for houses to sell
    a. Houses are more expensive in certain neighborhoods like Green Hills or Northridge
    b. Agents can focus on these neighborhoods to maximise commission
4. Agents can use this model to predict prices objectively prior to client negotiation

Model Preprocessing: Imputing, Encoding and Scaling

Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4







Methodology