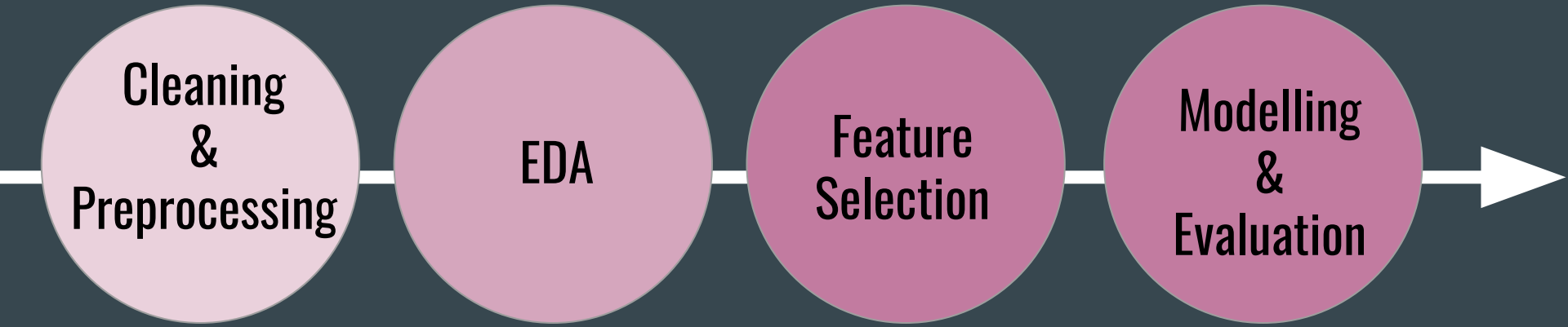
Ames Housing

Modeling and Analysis

Overview

1. Determine the best model for predicting Sale Price for houses in Ames. (R2 of at least 0.88, and should generalize well to new data within the Ames area)
2. Determine what the most important features are for said prediction
3. Less than 25 features to limit complexity

Workflow


Cleaning & Preprocessing
EDA
Feature Selection
Modelling & Evaluation

Data Cleanliness and Encoding

Outliers

Remove outliers that are skewing the data

“Missingness”

Fill with zero if NaN represents absence

Iterative Imputation/ Simple Imputation if truly missing

Dummying

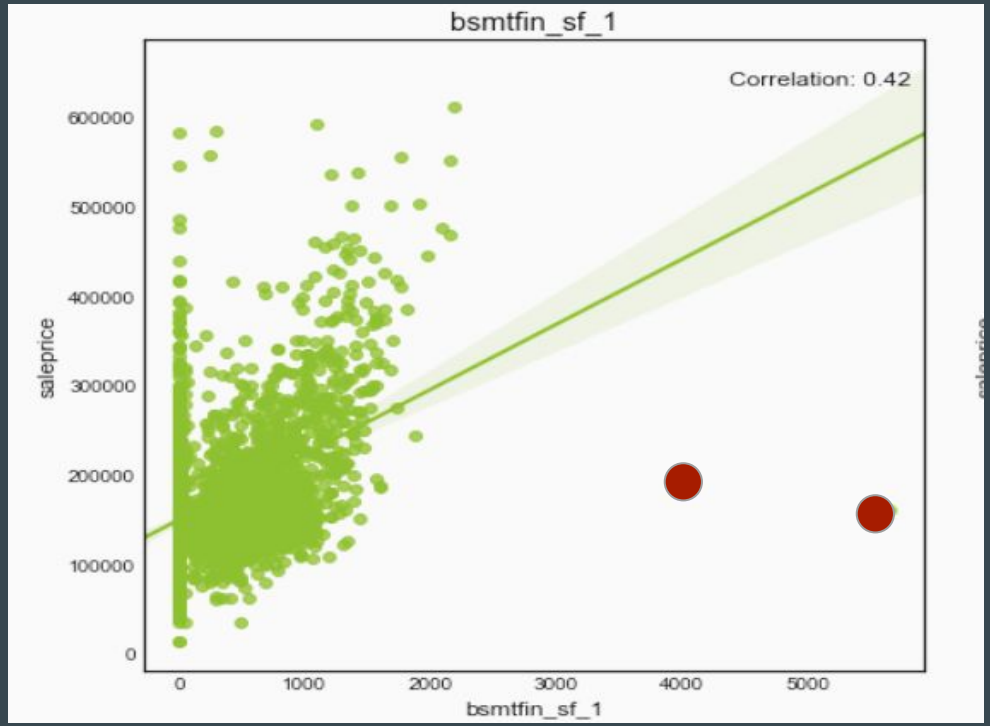
Categorical/nominal variables were one-hot encoded, or where applicable, binarized

Columns like “basement type” were interpreted as Likert scales

Highlights of EDA

Outliers

Two significant outliers were removed, as they were significantly underpriced for their square footage.

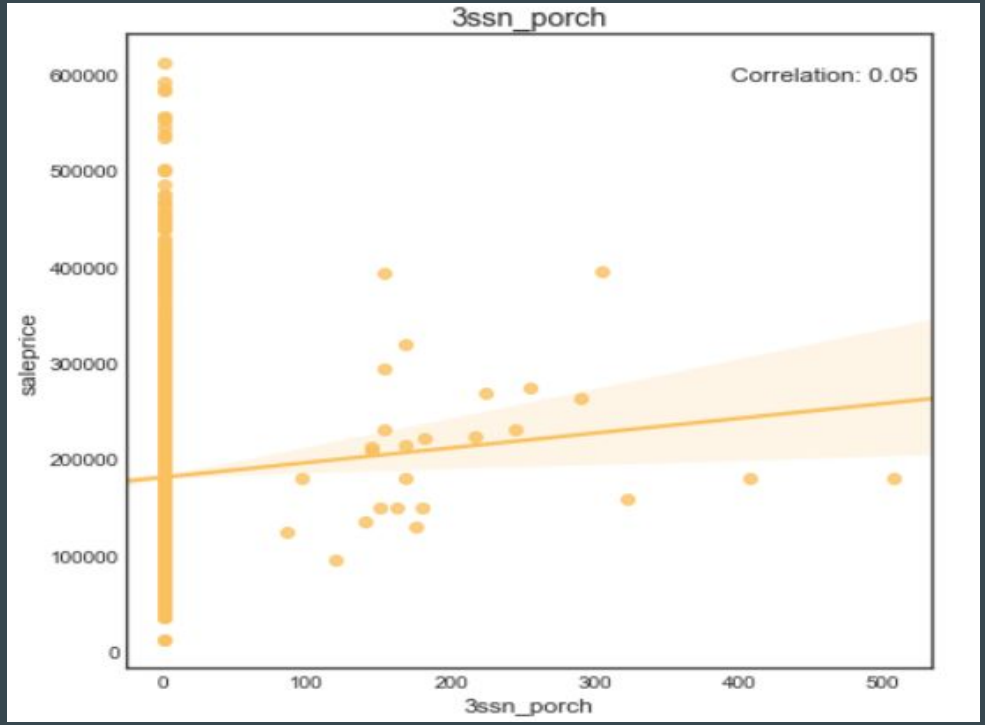


Highlights of EDA

Heavy skewedness

Due to the nature of encoding, some features were heavily skewed.

Such features can be encoded to binary instead (e.g. porch vs no porch)..

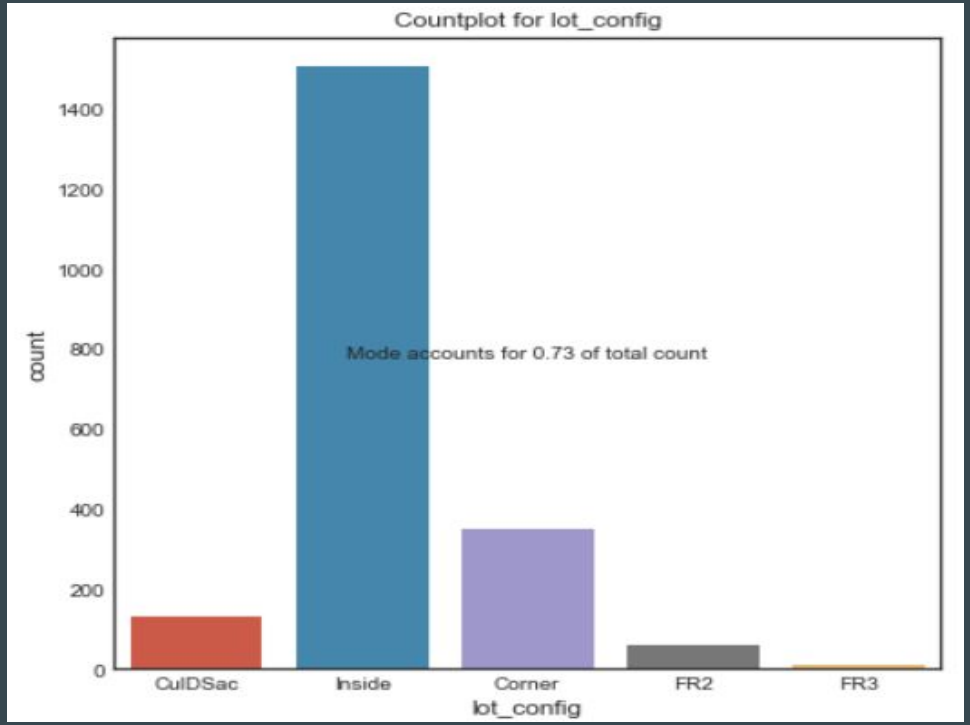


Highlights of EDA

Heavy skewedness 2

Similar problem with some categorical variables.

These can be binarized (e.g. “Inside” or “Not Inside” in this case)



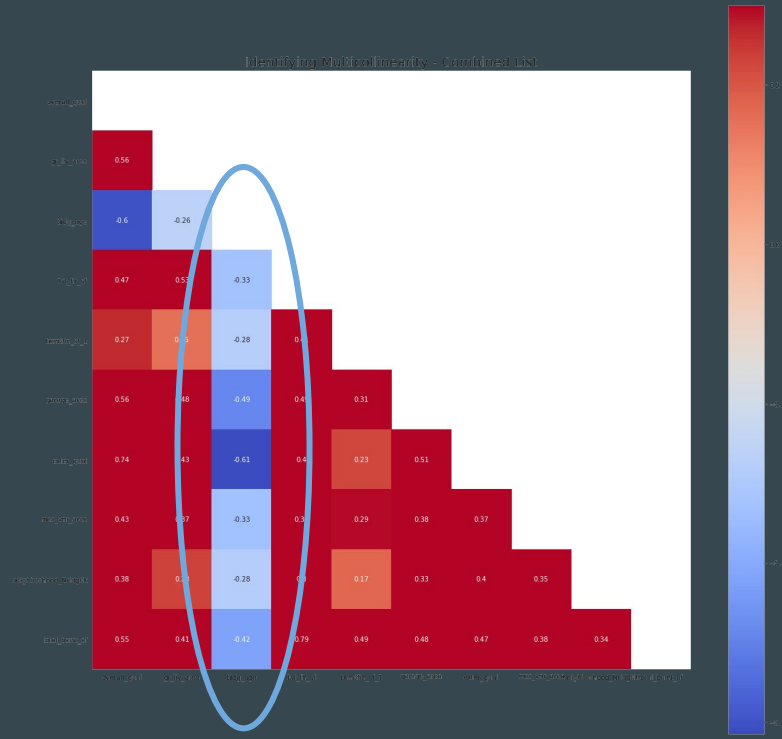
Multicollinearity

Dealing with Interaction Terms

The heatmap on the right is too small to interpret but the important thing to note is the blue line:

This represents building age - it is negatively correlated with all other features!

A polynomial approach may be needed to deal with these interaction terms.



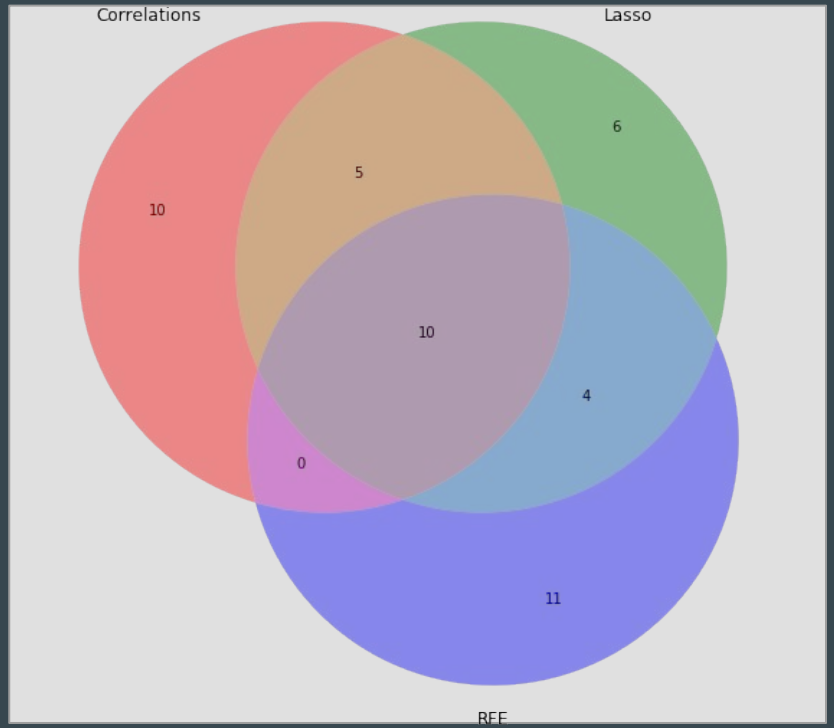
Feature Selection

Overlap of Feature Selection Methods

Three feature selection methods were used:

1) Filter (by correlation)
2) Wrapper (Recursive Feature Elimination)
3) Embedded (Lasso)

The features from all three methodologies were compared, and returned a list of 10 features that were shared.



Comparing Models:

1) 4 feature lists and 4 models were used
2) Poly + ElasticNet was the best performing model
3) The combined feature list (10 features) performed almost as well as the others (25 features)

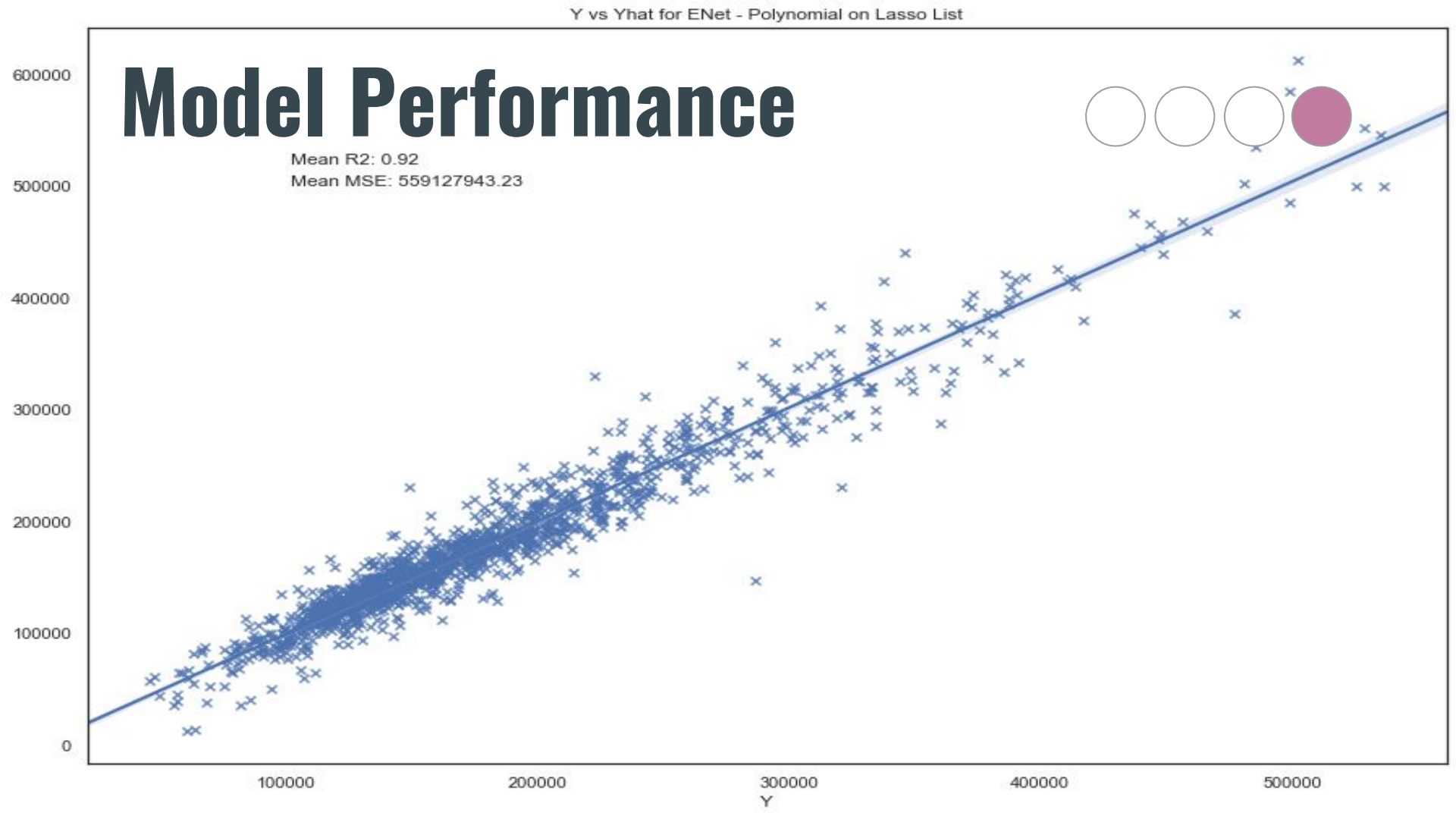
R 2

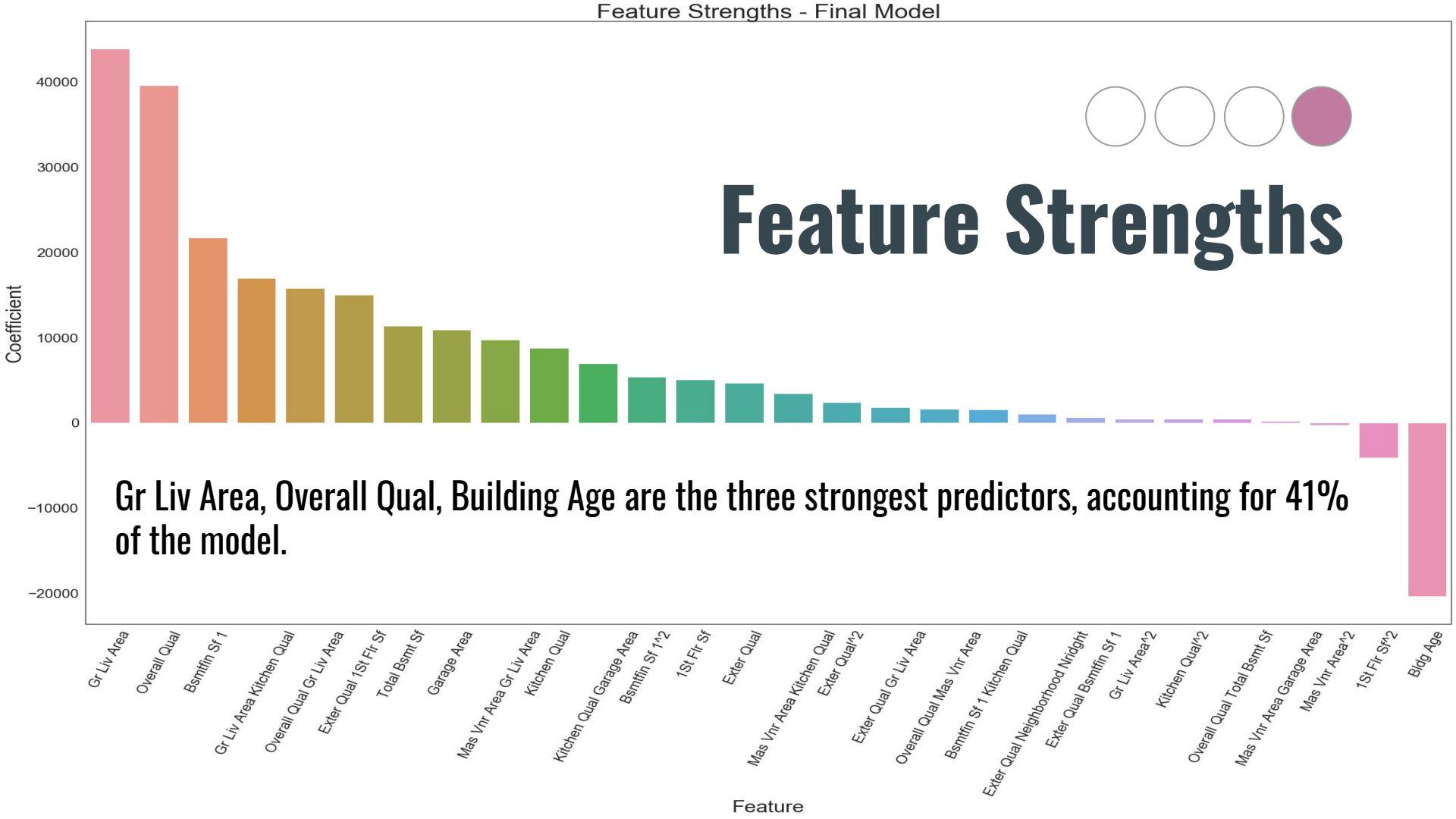
	Filter	Embedded	Wrapper	Combined
Ridge	0.87	0.89	0.89	0.87
Lasso	0.88	0.89	0.89	0.87
Enet	0.88	0.89	0.89	0.87
Poly Enet	-	0.91	-	0.90

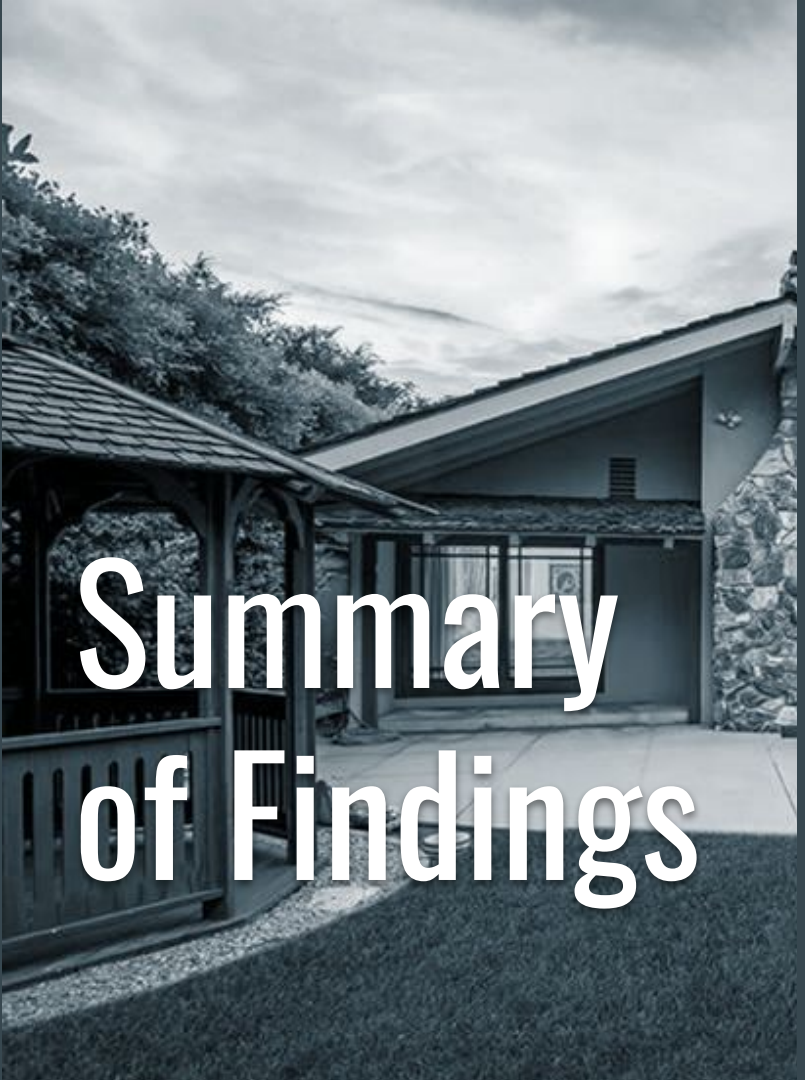
MSE

	Filter	Embedded	Wrapper	Combined
Ridge	7.44614e+08	6.523629e+08	6.45936e+08	7.885942e+08
Lasso	7.56769e+08	6.469722e+08	6.4887e+08	7.863289e+08
Enet	7.39538e+08	6.585488e+08	6.38745e+08	7.820008e+08
Poly Enet	-	4.529110e+08	-	5.815977e+08

Model Performance




Feature Strengths - Final Model
Feature Strengths
Gr Liv Area, Overall Qual, Building Age are the three strongest predictors, accounting for 41% of the model.
Coefficient
40000
30000
20000
10000
0
−10000
−20000
Gr Liv Area
Overall Qual
Bsmtfin Sf 1
Gr Liv Area Kitchen Qual
Overall Qual Gr Liv Area
Exter Qual 1St Flr Sf
Total Bsmt Sf
Garage Area
Mas Vnr Area Gr Liv Area
Kitchen Qual
Kitchen Qual Garage Area
Bsmtfin Sf 1^2
1St Flr Sf
Exter Qual
Mas Vnr Area Kitchen Qual
Exter Qual^2
Exter Qual Gr Liv Area
Overall Qual Mas Vnr Area
Bsmtfin Sf 1 Kitchen Qual
Exter Qual Neighborhood Nridght
Exter Qual Bsmtfin Sf 1
Gr Liv Area^2
Kitchen Qual^2
Overall Qual Total Bsmt Sf
Mas Vnr Area Garage Area
Mas Vnr Area^2
1St Flr Sf^2
Bldg Age
Feature



Summary of Findings

1) For feature selection, the intersection between the three feature selection methods yielded the features with the highest predictive power

2) Polynomials were needed to factor in correlations between the independent variables, and this led to a problem of over dimensionality. Using a smaller list yielded as much predictive power as using 25! features.

3) The ElasticNet model was the best performing in terms of both R2 and MSE.