Predicting House Prices in Ames, IA

Kavan Pandya
March 27, 2020

Problem Statement

This project comes up with a tool for realtors and homeowners that helps:

1. Predict the price of a new house that has been listed on the market in Ames, IA.
2. Provides recommendations for house features that might affect sale price the most.

Understanding the problem

The DATA

The data set given contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA

Understanding Features

There are 81 unique features for 2051 houses for data collected between 2006 to 2010.

E.g. - Lot Area (Continuous):
Heating (Nominal):
HeatingQC (Ordinal):

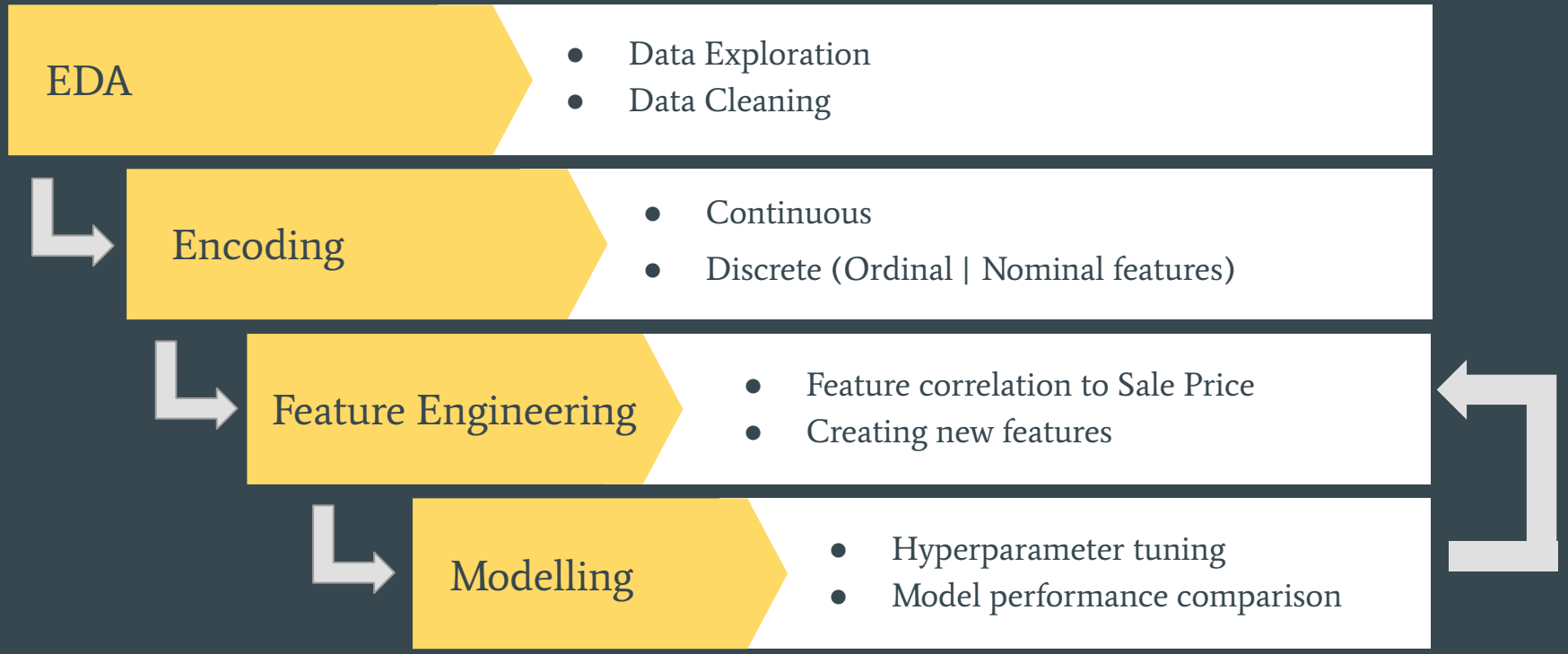
Choosing a model

Restricting ourselves to linear regression models, we pick a model from Linear Regression, Lasso, Ridge, and Elastic Net.

Model performance depends on:

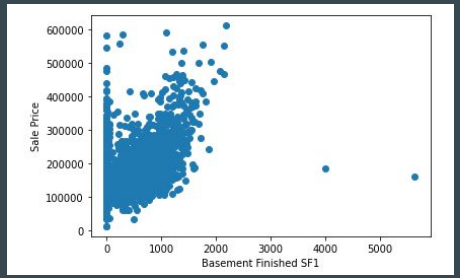
- Data
- Features
- Hyperparameters

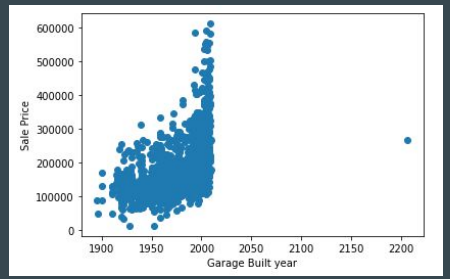
Process


EDA
Data Exploration
Data Cleaning
Encoding
Continuous
Discrete (Ordinal | Nominal features)
Feature Engineering
Feature correlation to Sale Price
Creating new features
Modelling
Hyperparameter tuning
Model performance comparison

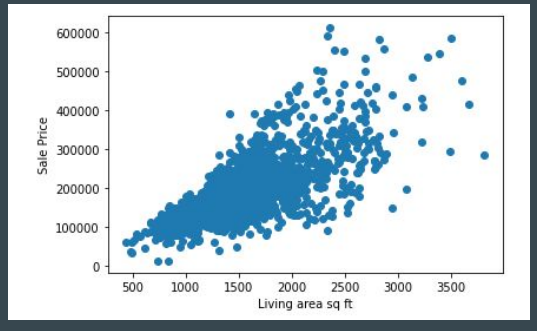
EDA

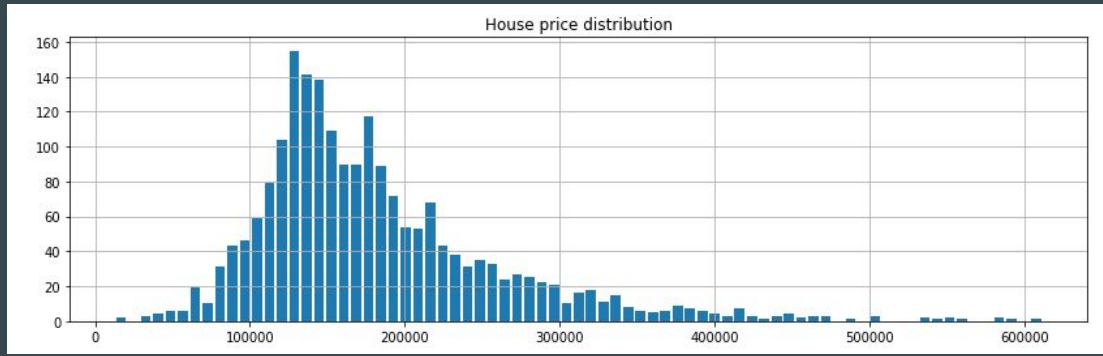
Identifying outliers and errors





Visualizing distributions





Encoding

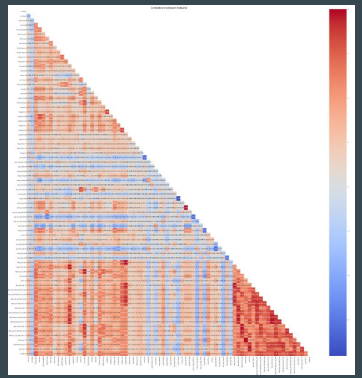
Discrete Variables:

- Ordinal - Ordinal Encoding
  - HeatingQC: Heating quality and condition
    i. Excellent - 4
    ii. Good - 3
    iii. Average/Typical - 2
    iv. Fair - 1
    v. Poor - 0

- Nominal - Dummy Encoding
  - Roof Style (Nominal): Type of roof
    i. Flat
    ii. Gable
    iii. Gambrel
    iv. Hip
    v. Mansard
    vi. Shed

Feature Engineering

Correlation to Sale Price



New_Bsmt_Feat New_Ovr_Kit_Feat	0.902488
New_Ovr_Kit_Feat Exter Qual	0.890821
New_Gar_Feat New_Ovr_Kit_Feat	0.848087
New_Ovr_Kit_Feat 1st Flr SF	0.832614
New_Ovr_Kit_Feat	0.829210
New_Bsmt_Feat Exter Qual	0.824928
New_Gar_Feat New_Bsmt_Feat	0.815495
Overall Qual	0.799589
Exter Qual 1st Flr SF	0.797014
New_Ovr_Kit_Feat^2	0.795647
New_Gar_Feat Exter Qual	0.781692
New_Bsmt_Feat^2	0.766448
New_Gar_Feat 1st Flr SF	0.764487
New_Bsmt_Feat	0.756997
New_Bsmt_Feat 1st Flr SF	0.752954
Gr Liv Area	0.725804
Exter Qual^2	0.721485
Exter Qual	0.712082
Kitchen Qual	0.691005
Total Bsmt SF	0.661263

Creating new features

1. Garage Area*(Garage Qual + Garage Cond)
2. Basement SF*(Basement Cond + Basement Qual)
3. Living Area*(Overall Qual + Overall Cond + Kitchen Qual)

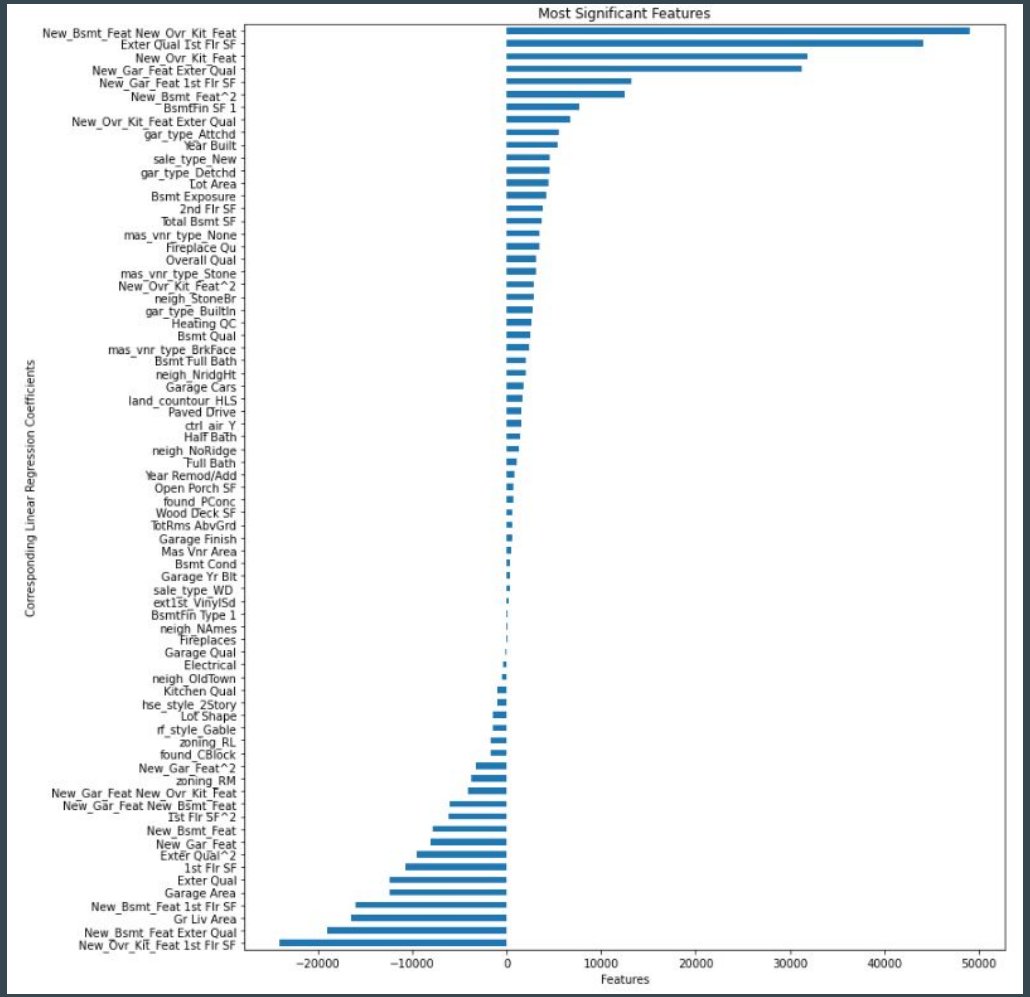
Polynomial Combination of these and:

1. Exter Qual
2. 1st Flr SF'

Modelling

Model Coefficients

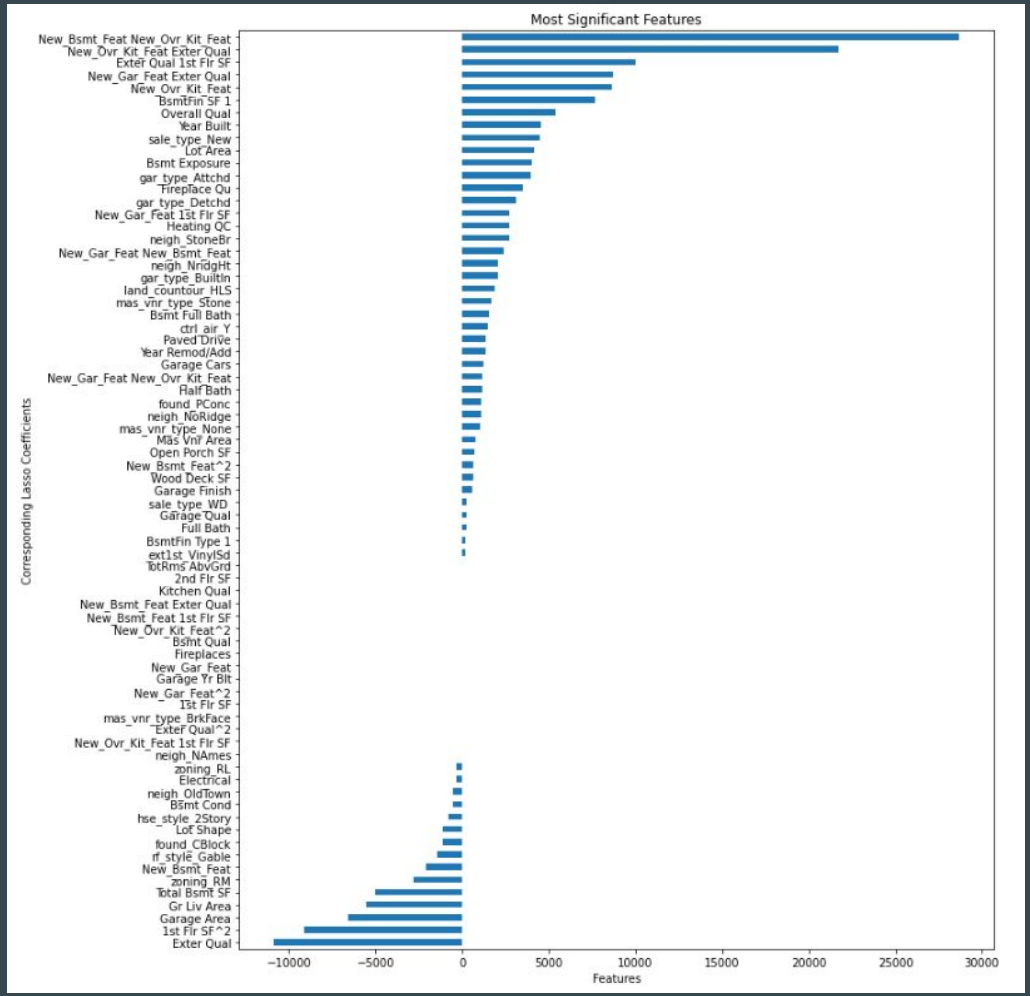
Linear Regression



Modelling

Model Coefficients

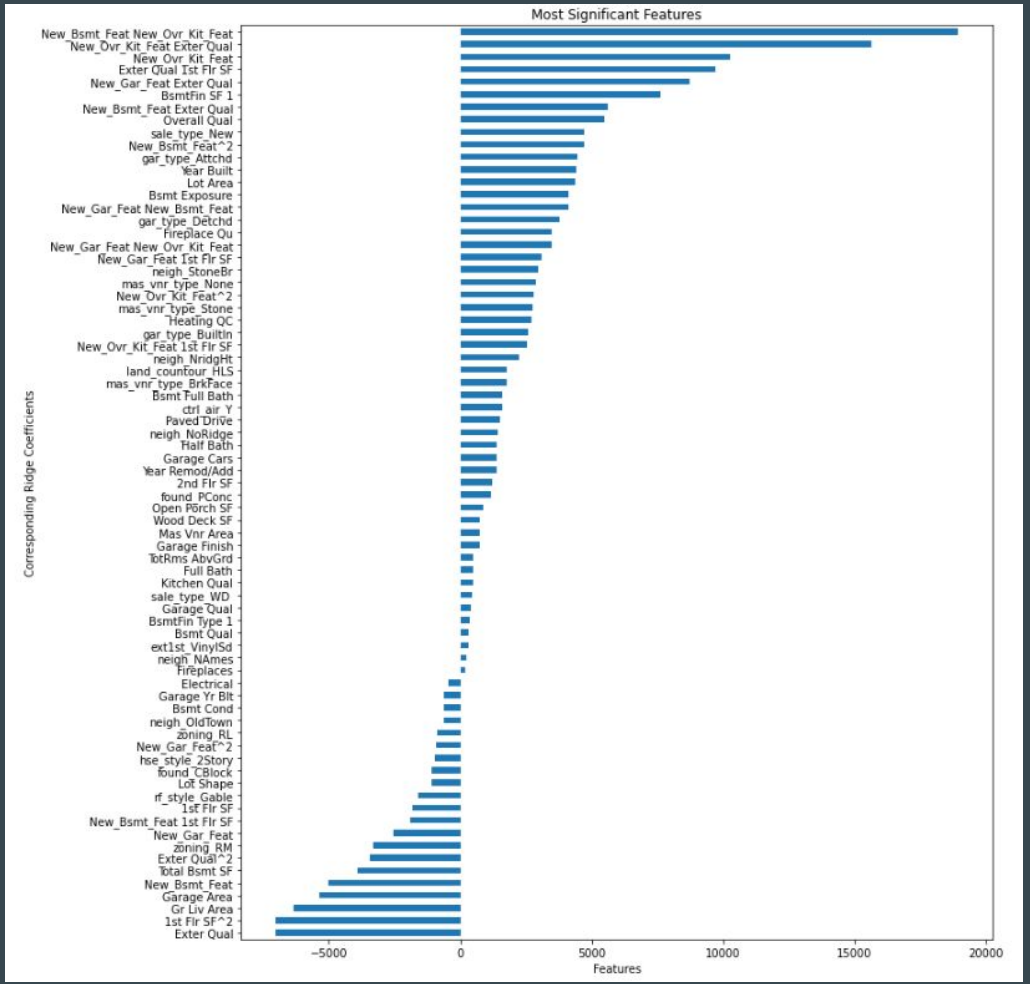
Lasso Regression



Modelling

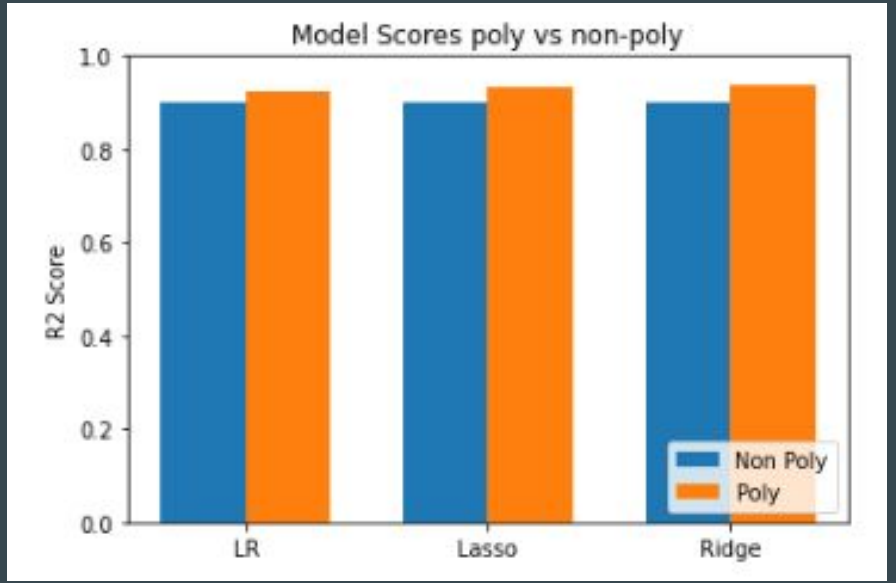
Model Coefficients

Ridge Regression



Modelling

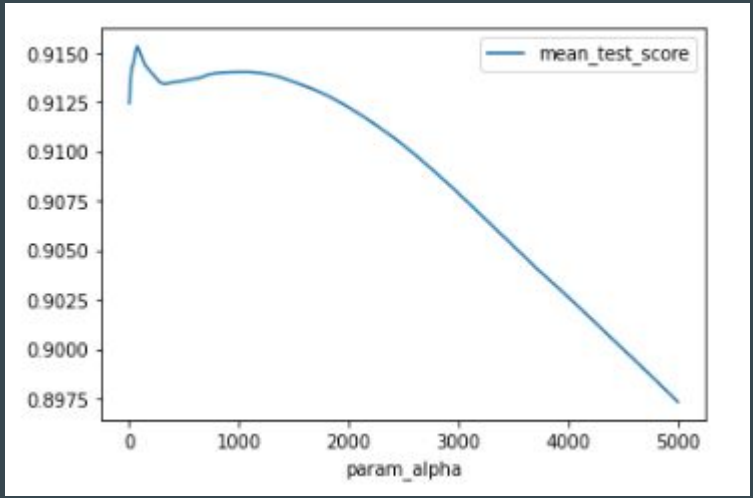
Model Comparison



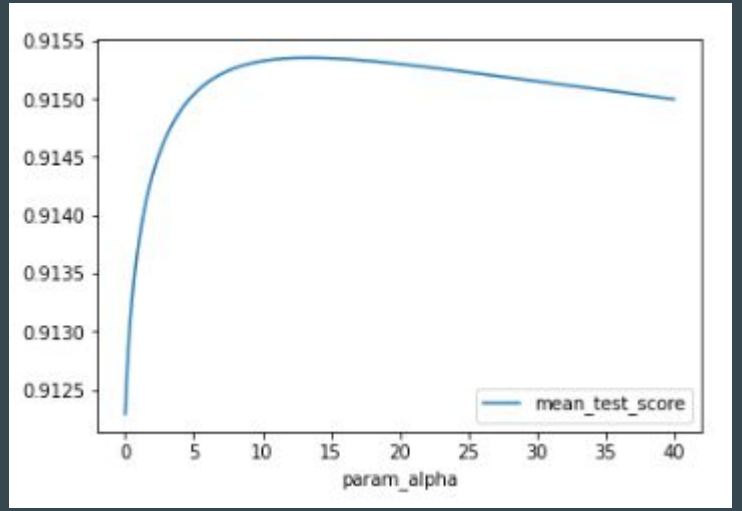
Modelling

Hyperparameter tuning

Lasso Regression



Ridge Regression



Interesting Features

Total Bsmt SF, Bsmt Qual, Bsmt Cond, Overall Qual, Overall Cond, Kitchen Qual, and Gr Liv Area

- Stone Brook neighbourhood seems to have the highest correlation.

- Also having a garage (attached or detached),
- Fireplace Quality,
- Heating Quality,
- Lot size

- Exterior Quality
- Garage Area
- Gr Living Area !!!

Next steps:

1. Discuss with Homeowner/Realtors what features are important for them and create combination features from them

2. Use a different machine learning model.

Questions?