Ames Housing Data and Kaggle Challenge

Matthew Burrell

07/10/2022

Problem Statement

- Challenge is to create a regression model to predict sale price of a home
- What type of features are controllable that can increase the sale price of a home
- The model to develop is a OLS Ordinary Least squared
- OLS is a white box model meaning the coefficients gauge what effect sale price

Ames Data

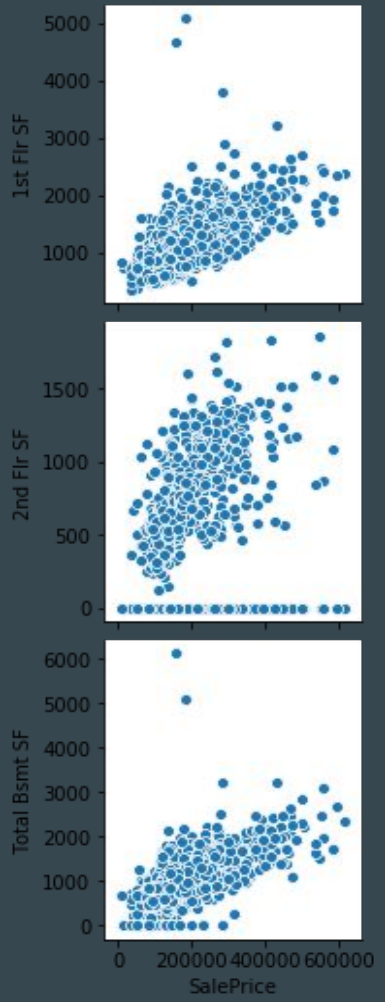
- The Dataset is exceptionally detailed
- It has 80 columns of data ranging from
  - Square feet of the first and second floor
  - To fireplace quality

EDA and Data Cleaning

- Missing variables mostly where from house without certain features
  - For example, an NaN in the fence column ment there was no fence
  - Replaced most missing futures with the missing feature
- One missing electrical entry
  - Replaced that with the most frequent electrical system: Standard Circuit Breakers
- All data types where correct
- Selected features
  - '1st Flr SF', '2nd Flr SF', 'Total Bsmt SF','Lot Area', 'Garage Area', 'Garage Type', 'Garage Cars' ,'Fireplaces', 'Fireplace Qu', 'Fence', 'Paved Drive', 'Street', 'Central Air'

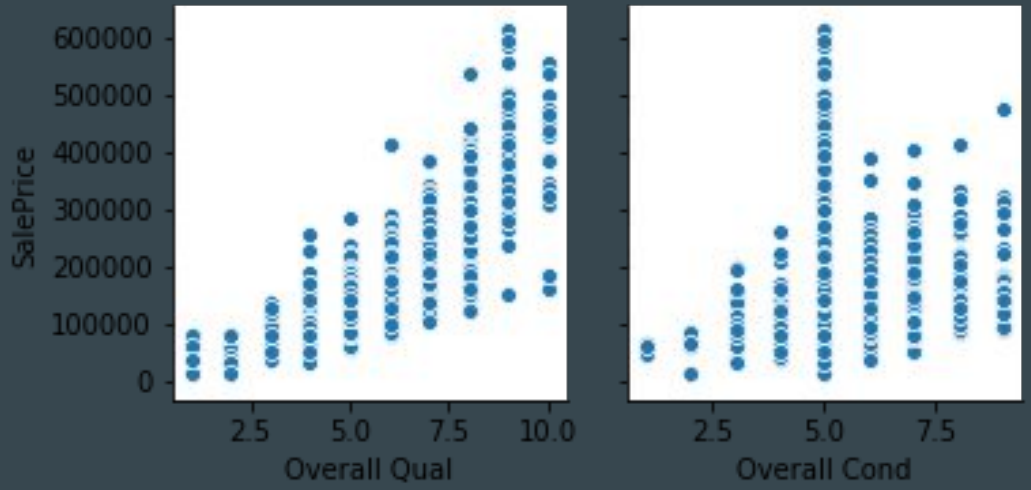
Feature Engineering

- Square footage
  - A lot of house did not have a 2nd floor
  - Some did not have a basement
- Combined all square footages into one total square footage variable
- Also, squared the total square footage



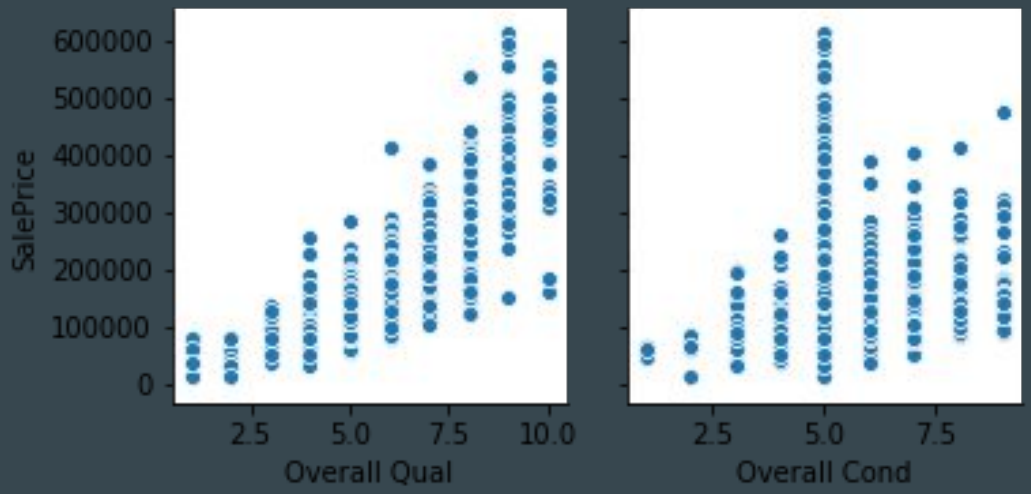
Feature Engineering cont

- Scale overall quality and condition
  - 10 very excellent
  - 9 excellent
  - 8 very good
  - 7 good
  - 6 above average
  - 5 average
  - 4 below average
  - 3 fair
  - 2 poor
  - 1 very poor



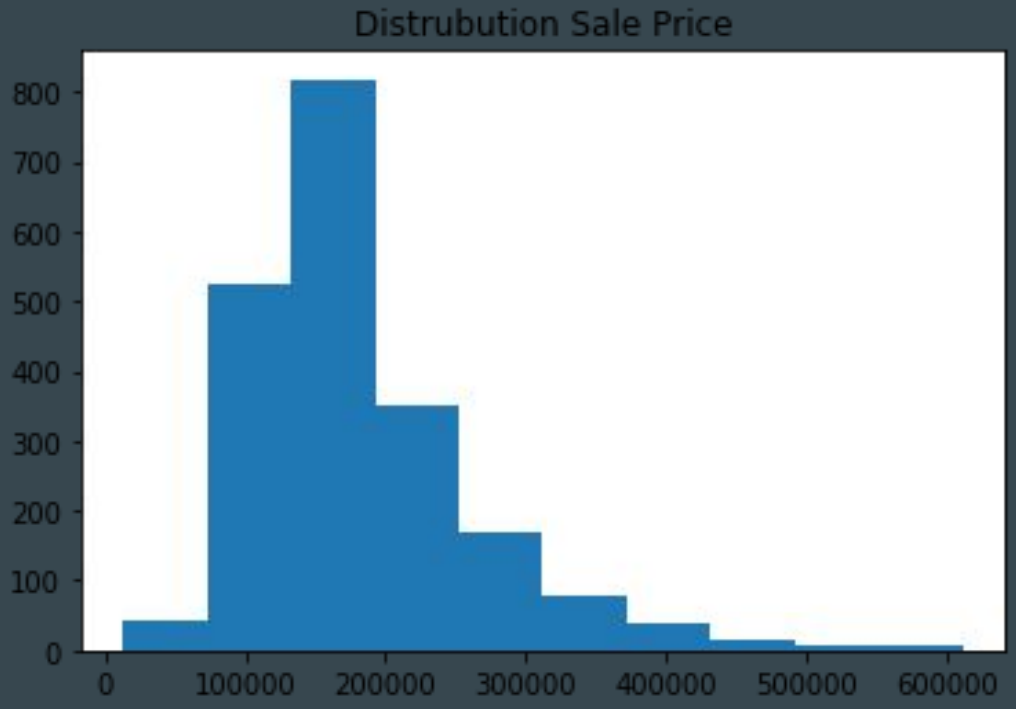
Feature Engineering cont

- Recategorized both quality and condition
- For quality
  - Above 9 excellent
  - 8 and 7 good
  - 6 and 5 is average
  - 5 to 3 fair
  - Below 3 poor
- For condition
  - Above 9 excellent
  - 8 to 6 good
  - 5 is average
  - 5 to 3 fair
  - Below 3 poor



Preprocessing and Modeling

- Distribution of Sale price is skew right
- For OLS to work better, the target variable should be normally distributed
- Took the natural log to fix issue



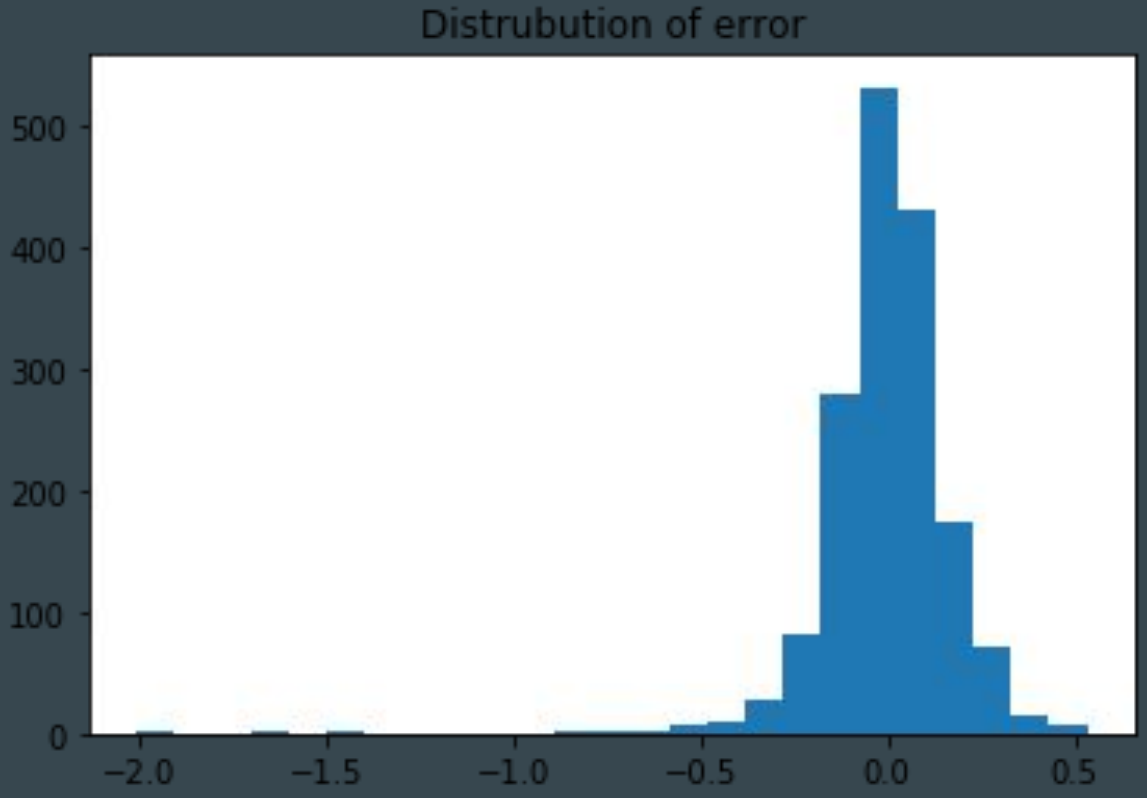
Pre and Mod cont.

- Multicollinearity
  - Total square footage is highly correlated with total square footage squared of 0.92
    - Dropped square footage
  - Garage area was highly correlated with garage cars
    - Dropped Garage area

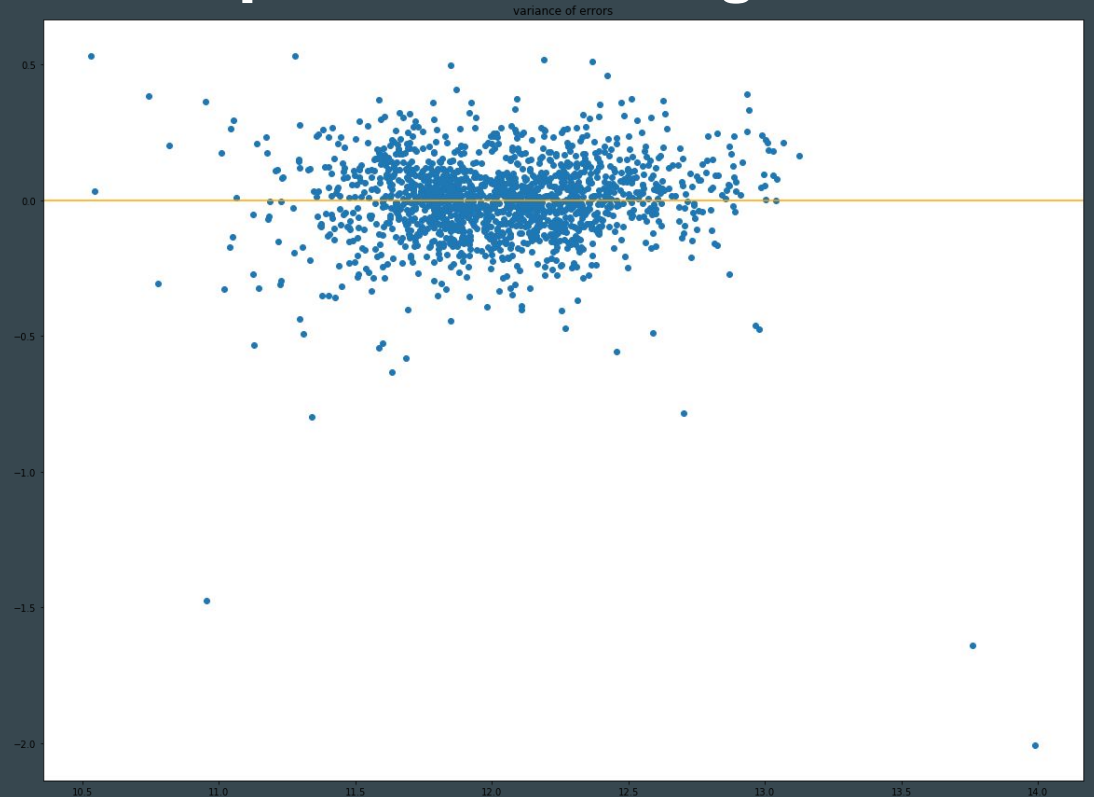
Preprocessing and Modeling

- Split train set
  - Used 20%
  - And a random state of 42
- Trained on a OLS model
- Model r2 results
  - train score 0.8442
  - test score 0.8487
  - cross val score 0.8222

Evaluation and Conceptual Understanding



Evaluation and Conceptual Understanding cont



Evaluation and Conceptual Understanding cont

- Mean absolute error 0.10917104843581744
- Mean Squared Error 0.02590429862342500 2
- Root Mean Squared Error 0.16094812401337583

Explaining some coefficient metrics

- For homes that are in excellent condition, there are 10.2% more valuable than a home with an average score for condition, while holding all else constant.
- For homes that are in good condition, there are 2.7% more valuable than a home with an average score for condition, while holding all else constant.
- For homes that are in fair condition, there are 15.7% less valuable then a home with an average score for condition, while holding all else constant.
- For homes that are in fair condition, there are 22% less valuable then a home with an average score for condition, while holding all else constant.

A funny coefficient

- Holding all else constant, a home that has an ben franklin Stove is 0.09% less valuable, then a home without a fireplace.
- Someone would rather have no fireplace then have this thing in their home

Conclusion

- The OLS model does a fair job with predicting sale prices of homes
  - train score 0.8442
  - test score 0.8487
  - cross val score 0.8222
- To increase the price of a home the best bet is to improve condition of the home
  - homes that are in excellent condition, there are 10.2% more valuable
- Next steps to improve the model
  - adding more features from ames data
  - Scale features
  - Using other machine learning models like lasso or ridge regression