Ames House Hunting

Machine Learning Project 2020

Anjali Pathak, Brandon Ryu,
Isabel Alvarez de Lugo, Stephen Kita

Data Overview

The data used featured about 2500 house sale records from **Ames, Iowa** between 2006-2010.

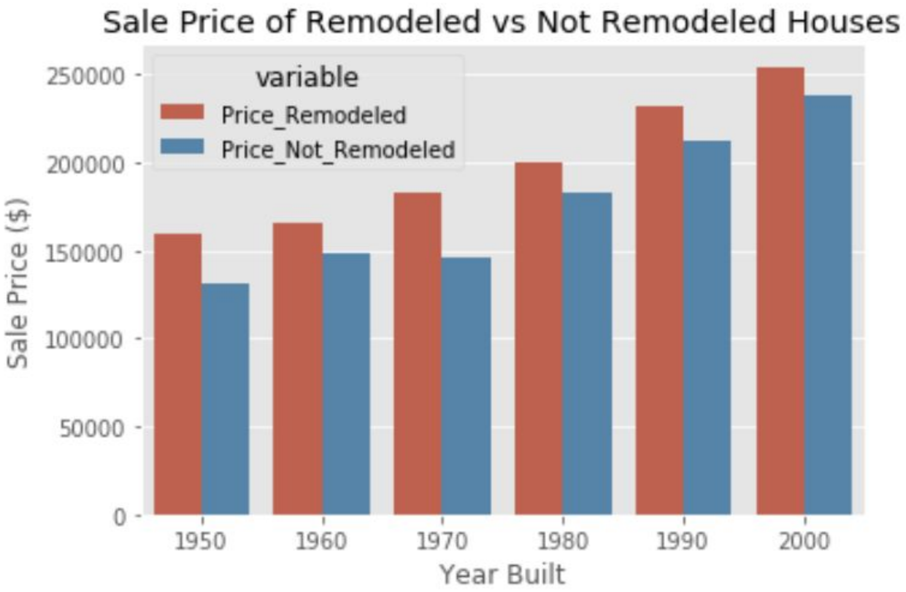
There are two datasets used: `Ames_Housing_Price_Data.csv` and `Ames_Real_Estate_Data.csv`.

The `Ames_Housing_Price_Data.csv` set contains 81 data columns, including the key feature SalePrice which will be used as the target of the predictive/descriptive modeling. 2580 observations (properties)

The `Ames_Real_Estate_Data.csv` set contains 90 data columns, including the key feature Prop_Addr which will be used to find the long-lat coordinates of the houses.

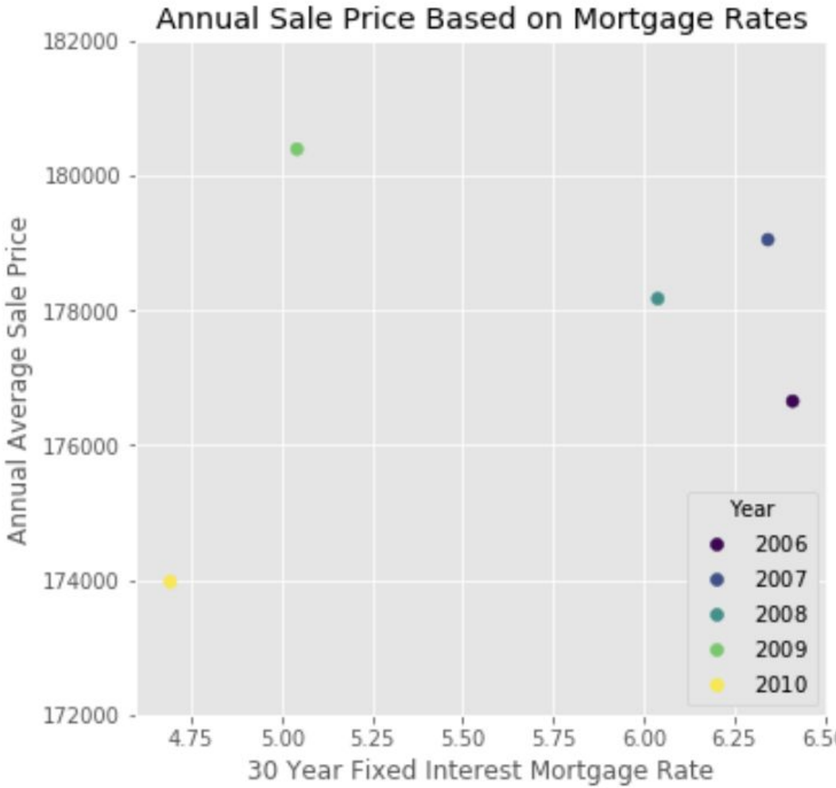
EDA: Housing Analysis

Does the Month Sold affect the Sale Price? Does Remodeling a home increase the Sale Price?



EDA: Housing Analysis

Can you see the effects of the Great Recession (2007-2009) in our data?



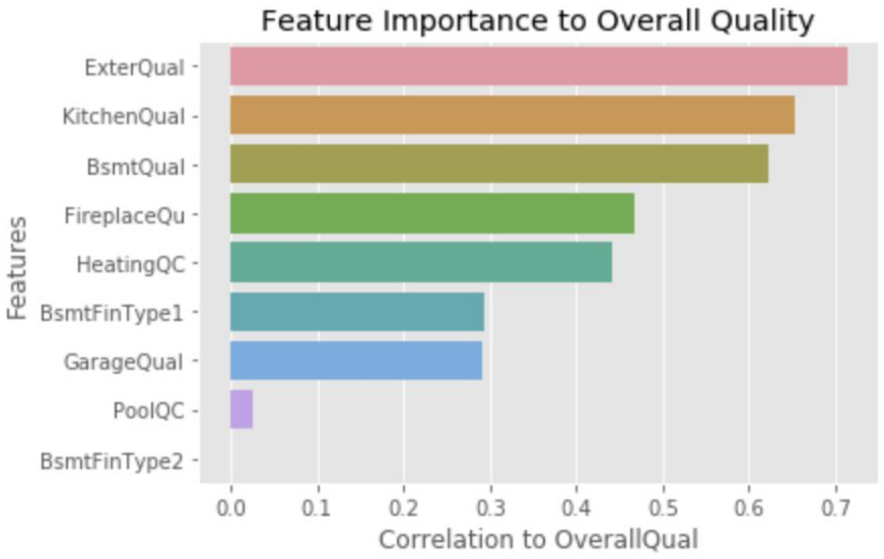
EDA: Housing Analysis

Does having certain optional additions affect Sale Price?

Feature	Average change in worth ($1000)	P value	Correlation with Sale Price
Pool	78	0.001801	0.061
Fireplace	72	0.000000	0.480
Finished Garage	71	0.000000	0.404
Been Remodeled	70	0.020231	-0.053
Porch	47	0.000000	0.291
Deck	46	0.000000	0.309
Finished Basement	29	0.000000	0.176

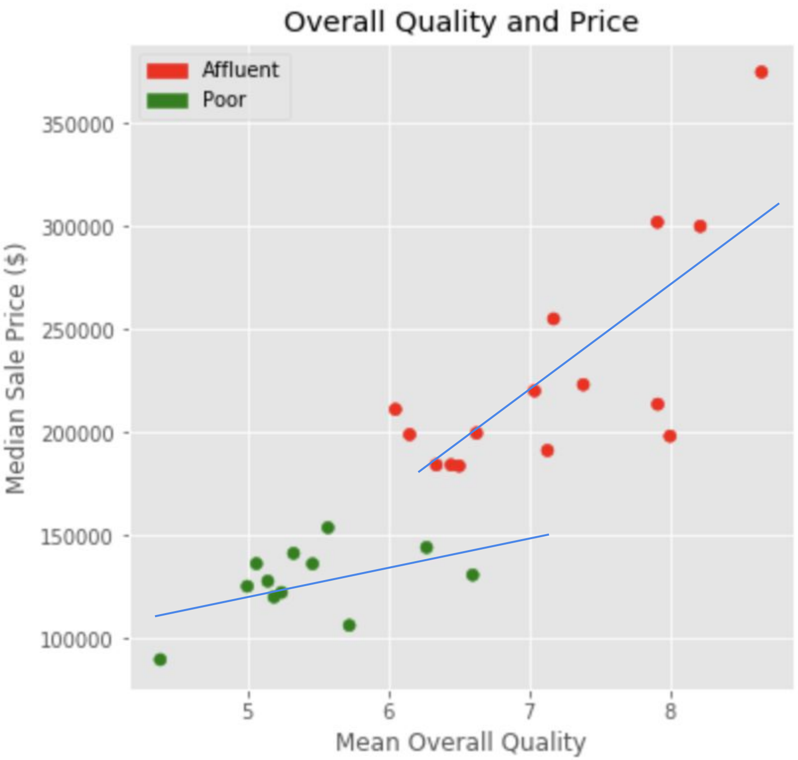
EDA: Housing Analysis

How does the Overall Quality affect Sale Price? What are the key features driving Overall Quality?



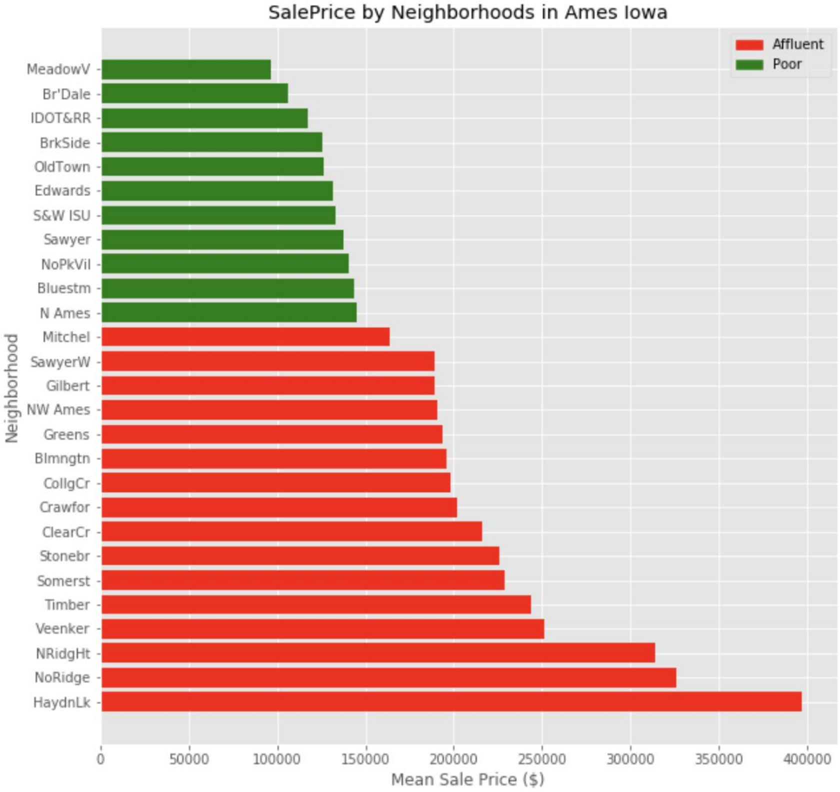
EDA: Neighborhood Analysis

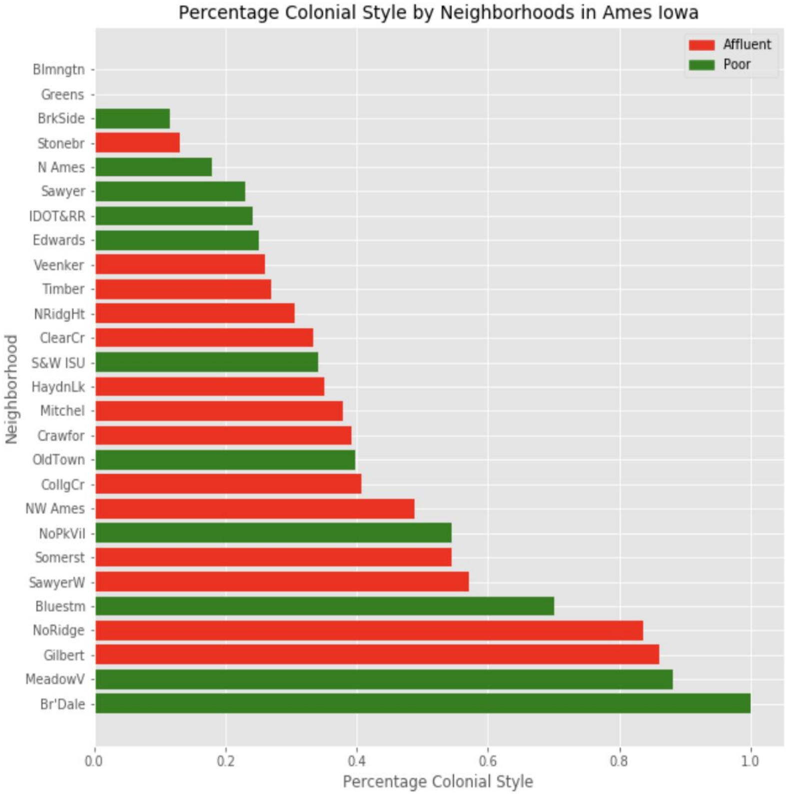
Does the price sensitivity on quality depend on the neighborhood?



EDA: Neighborhood Analysis

Which are the more expensive neighborhoods? What types of homes are popular in Ames?



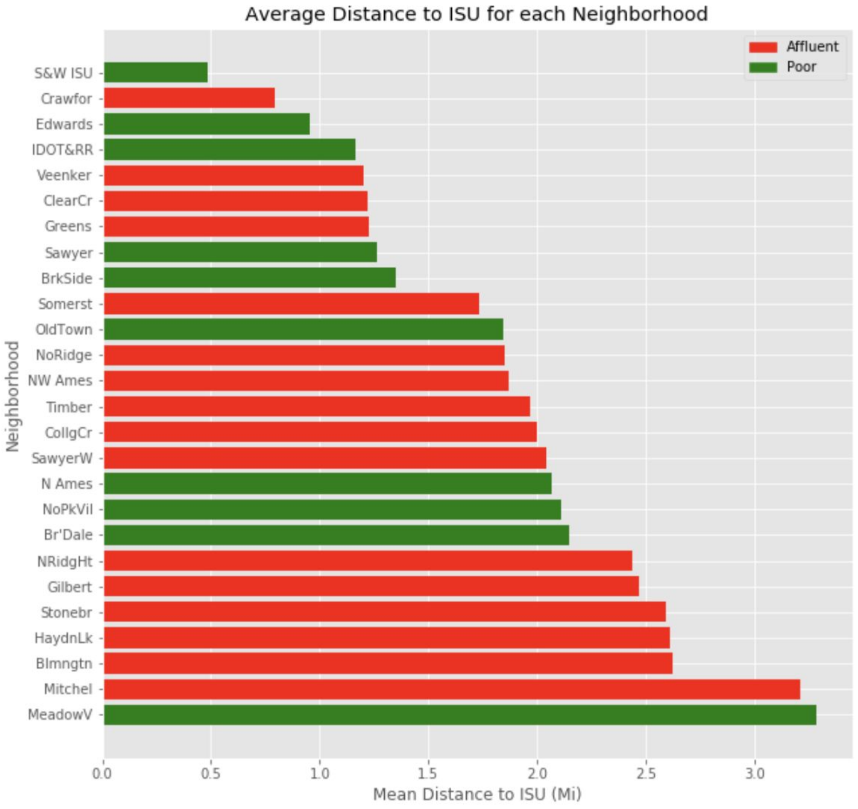


EDA: Neighborhood Analysis

Which Neighborhoods are closest to ISU?

- **ISU** is the largest employer of Ames, IA

- Neighborhoods with more convenient job commute.
    - Crawford
    - Edwards



EDA Conclusions

- As overall quality increased so did the Sale Price

- More affluent neighborhoods were more sensitive to changes in quality.

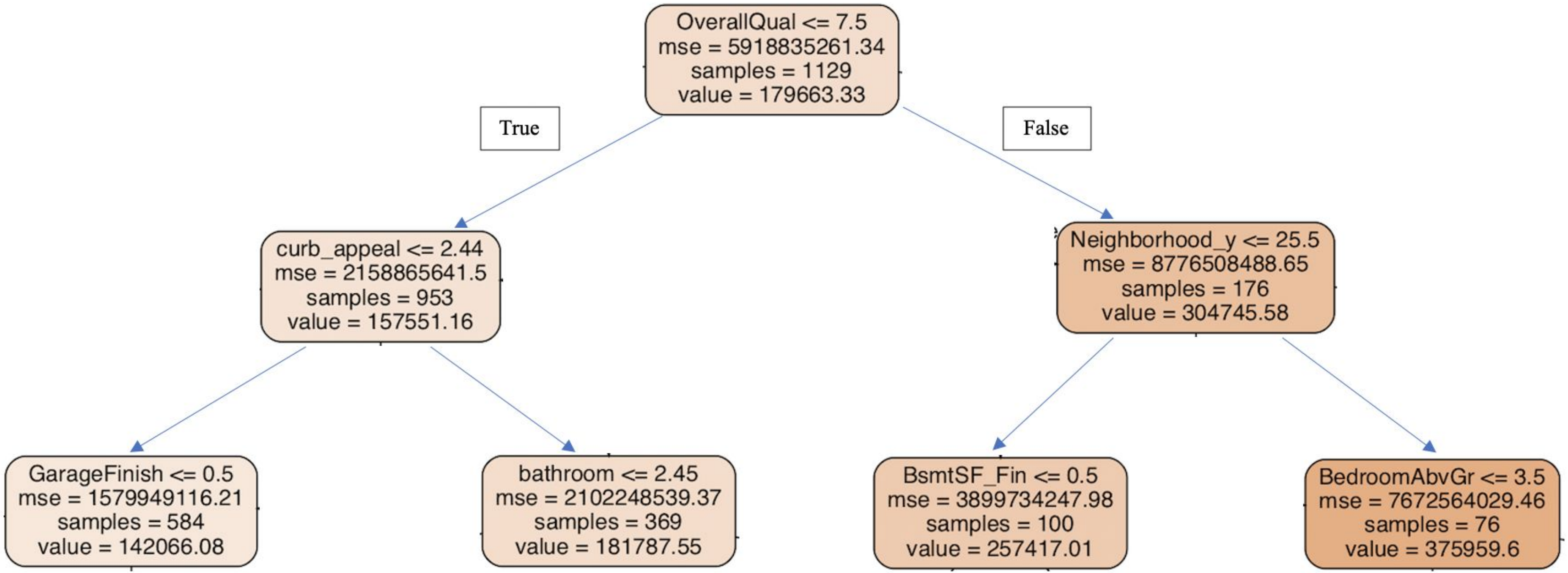
- There were moderate correlations with having a Fireplace and having a Finished garage and Sale Price

- 60% of the house in the dataset were Ranch style

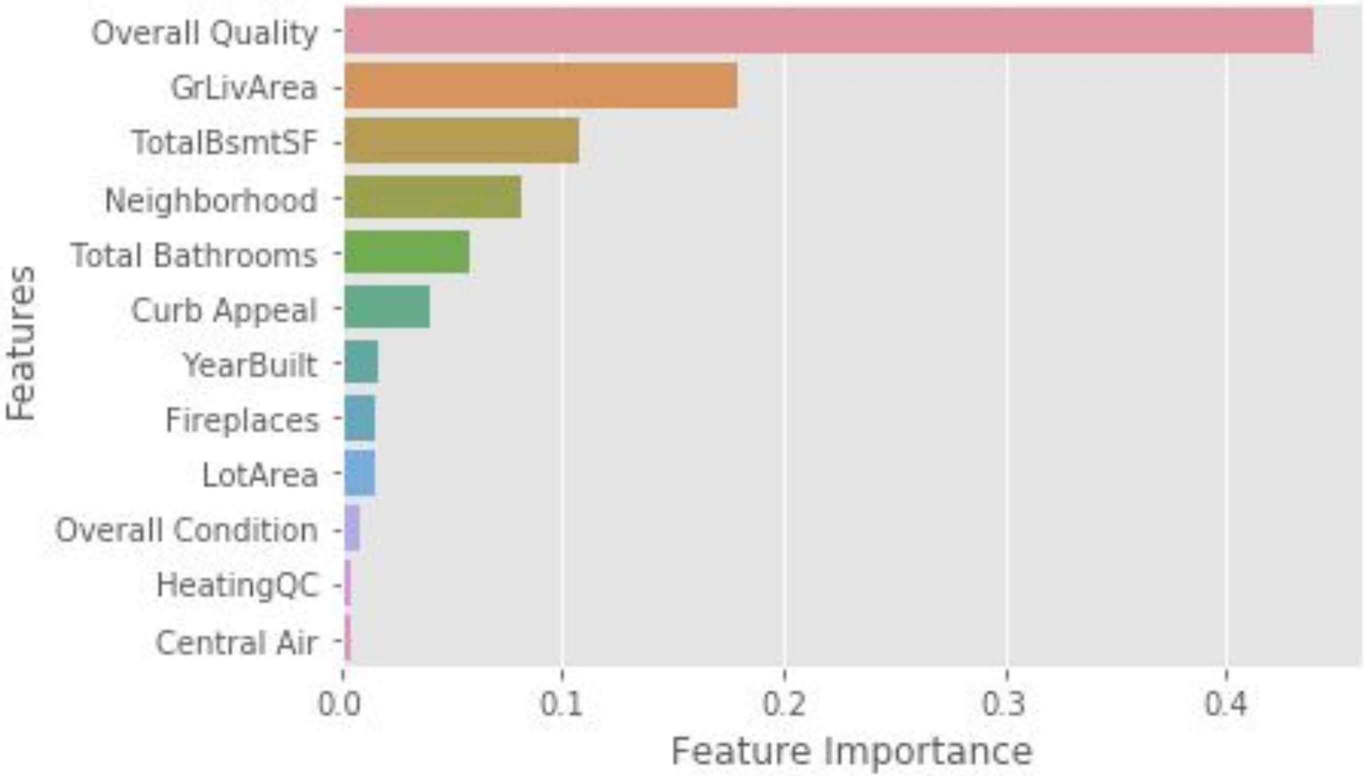
Features Engineered: Tree Based Models

- Total Bathrooms
  - Combined number of bathrooms into one feature
- Curb Appeal
  - Combined features related to curb appeal such as exterior quality, roof style, lot config, etc.
- Distance to College
- Distance to High School
- Total Porch Area
  - Combined all Porch sq. ft. features
- Basement
  - Transformed into binary - whether or not house has a basement

Decision Tree from Random Forest



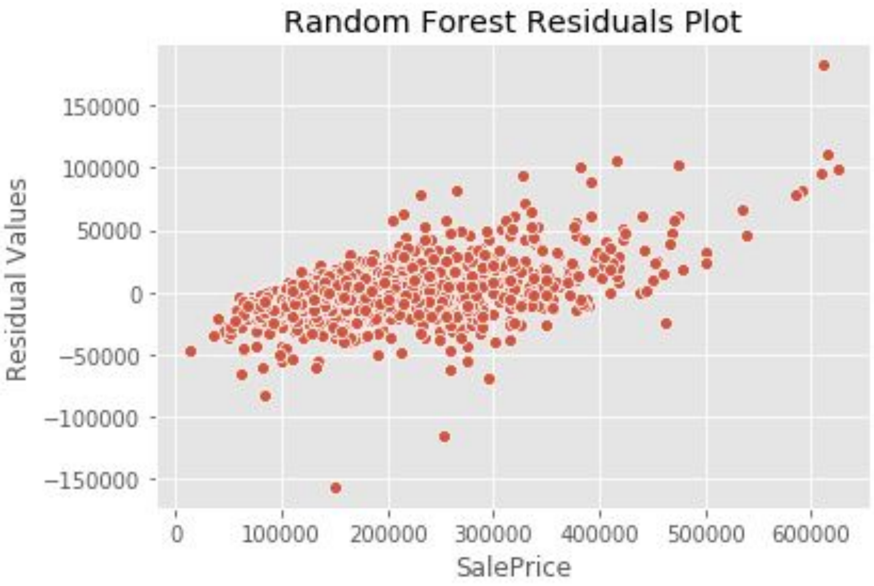
Feature Importance: Top 12 Features

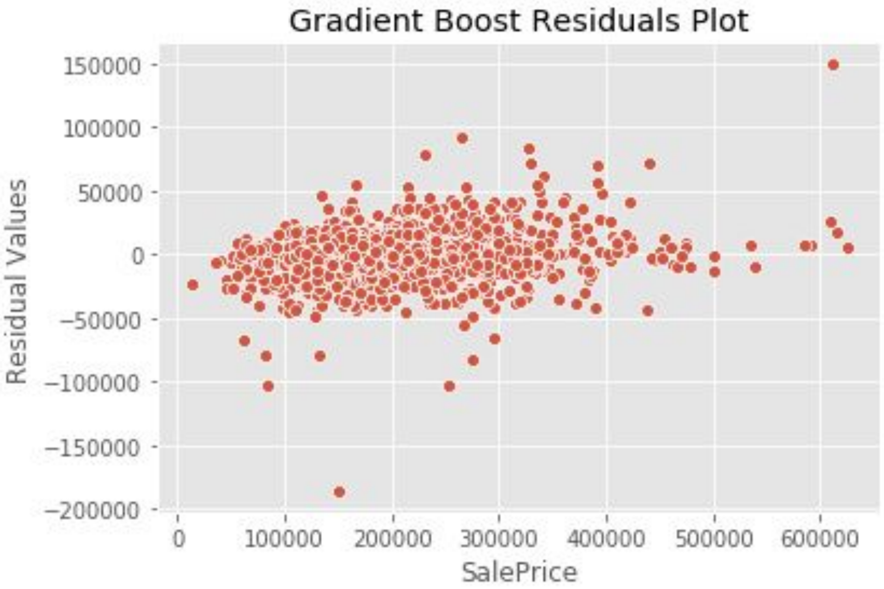


Tree Based Models: Scores

Model	R^2 Train	R^2 Test
Random Forest	0.967	0.895
Gradient Boost	0.971	0.911
XGBoost	0.963	0.901

Tree Models: Residuals Plots





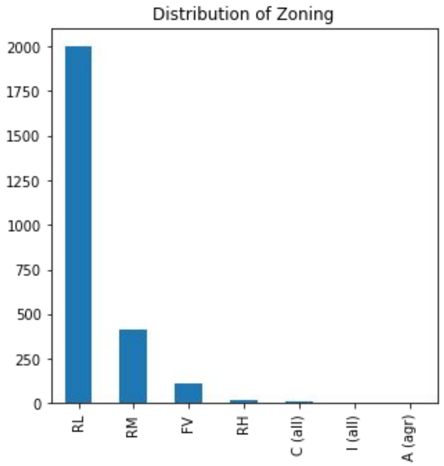
House Hunters - Ames, IA

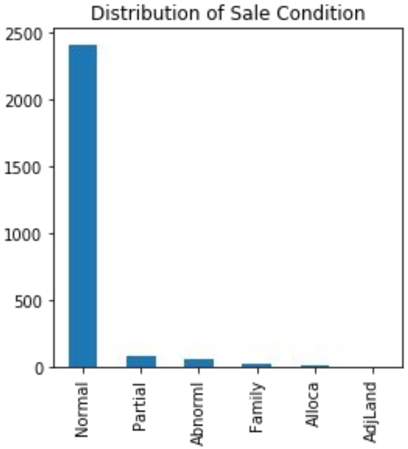
Objective: Create a house pricing model that is *interpretable* and can provide *recommendations* based on a buyer/seller's profile.

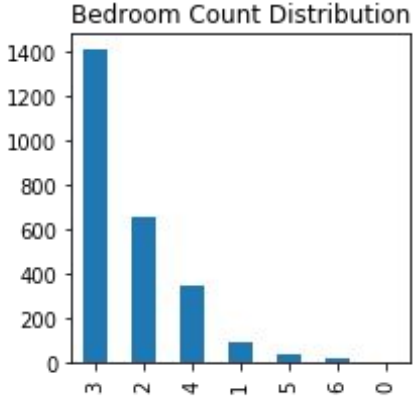
Solution: Multiple linear regression

Linear Model - Processing

- Remove outliers   (2580 => 2125 samples)

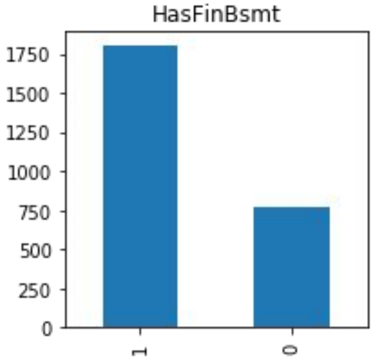


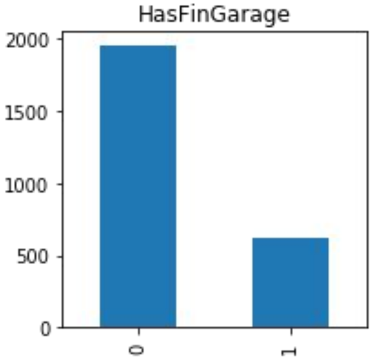


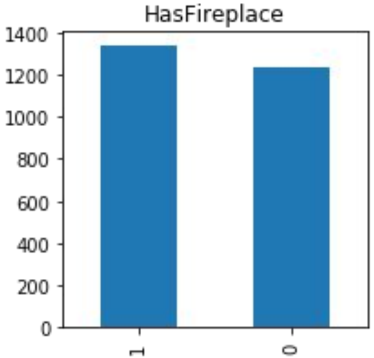


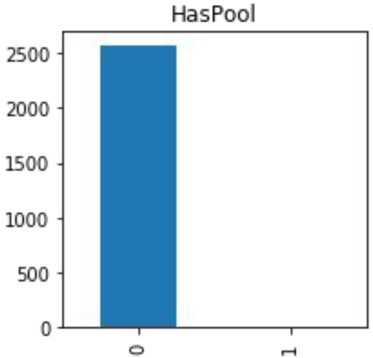
Linear Model - Transforming

- Create binary variables (15 new features)

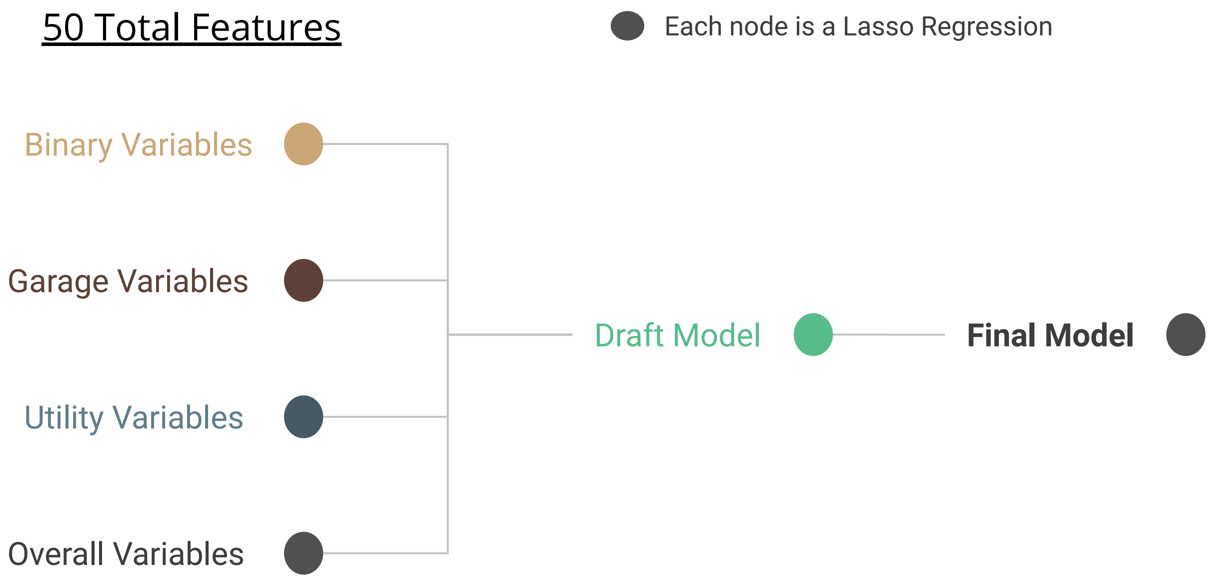








Linear Model - Feature Selection



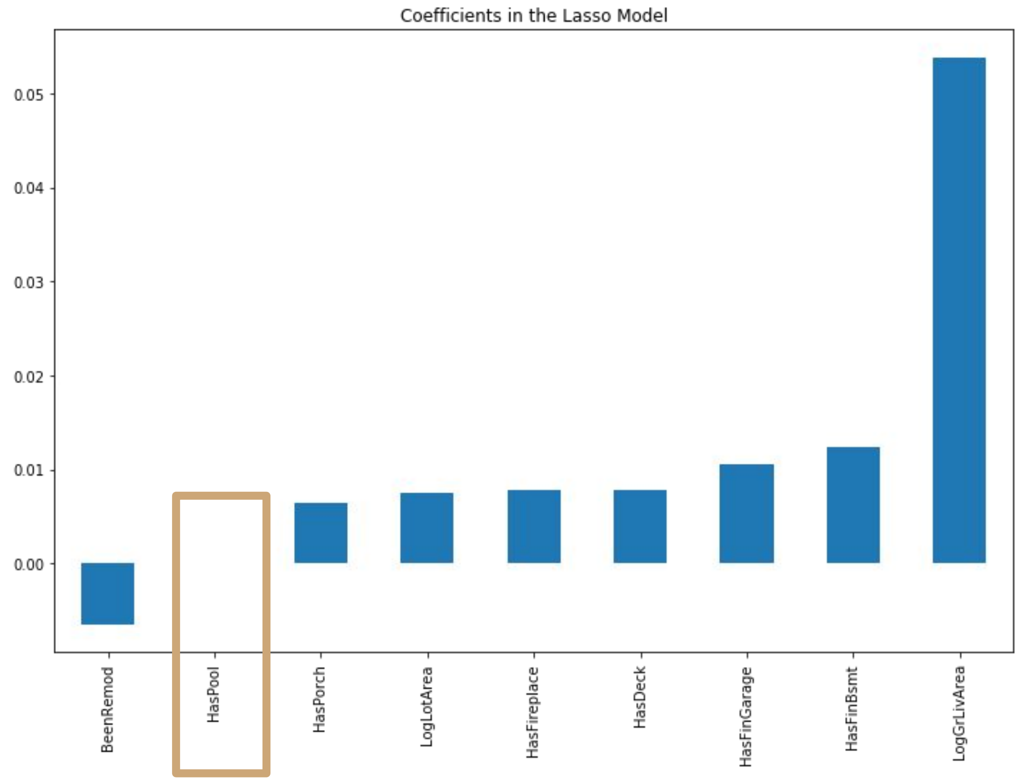
Feature Selection with Lasso

Dummify Variables	Lasso Regression	Watch Coefficients
Category: 1, 2, 3, 4 becomes Category_1, Category_2, Category_3, Category_4	Grid search for best penalization term (alpha)	Features with coeff = 0 can be dropped from the model. Categories with similar coeff can be grouped into binaries

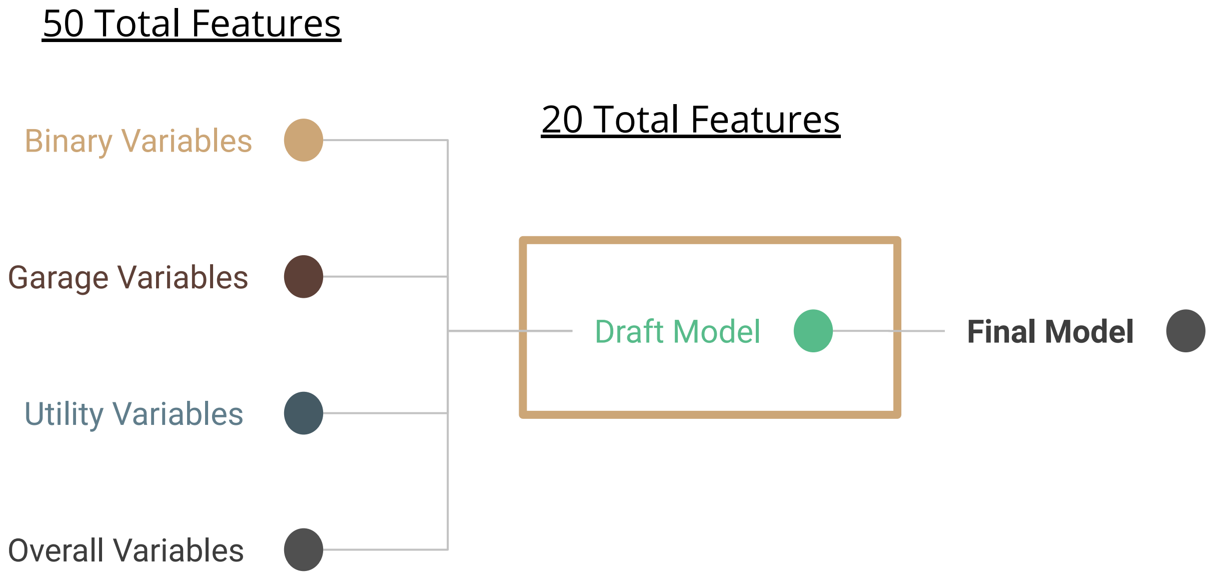
Binary Variables

Baseline
LogLotArea
LogGrLivArea

Test
BeenRemod
HasFinBsmt
HasFinGarage
HasPool
HasFireplace
HasPorch
HasDeck

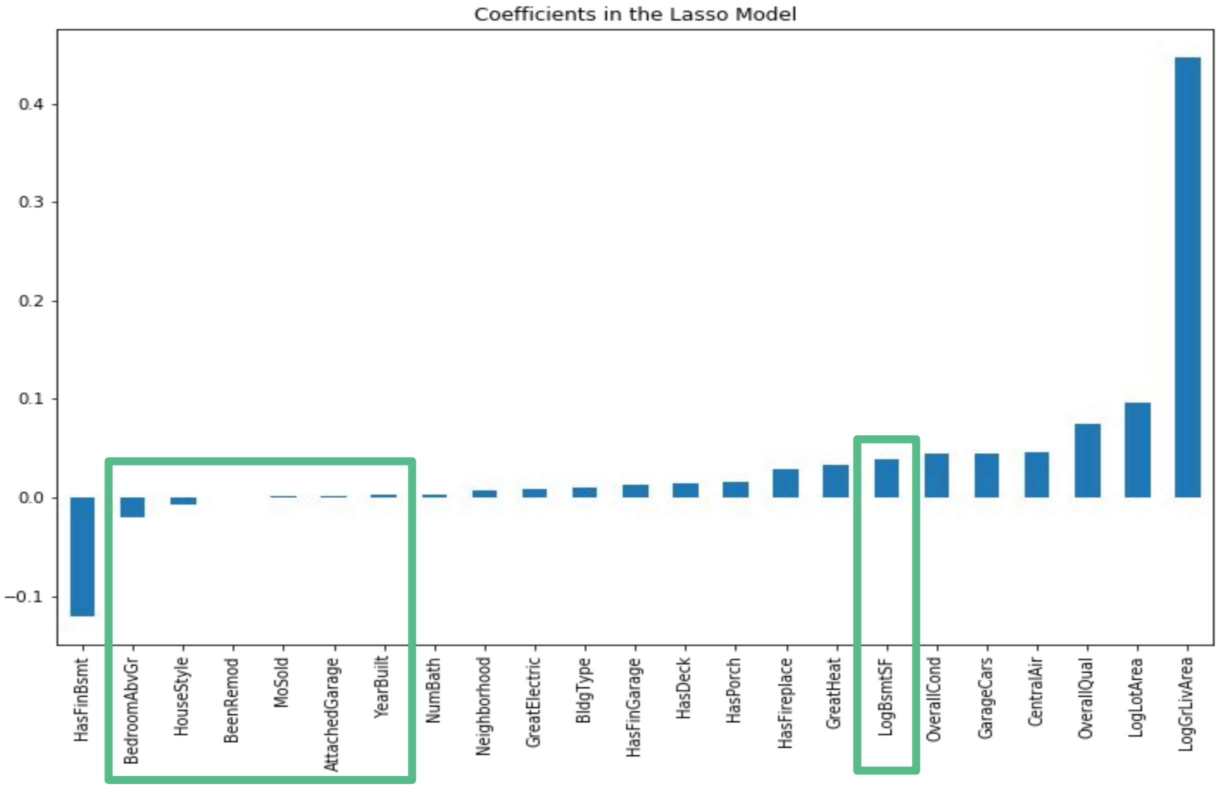


Linear Model - Feature Selection



Draft Model Rev 1

- <u>Remove</u>
  LogBsmtSF
  Bedroom
  BeenRemod
  HousStyle
  MoSold
  YearBuilt

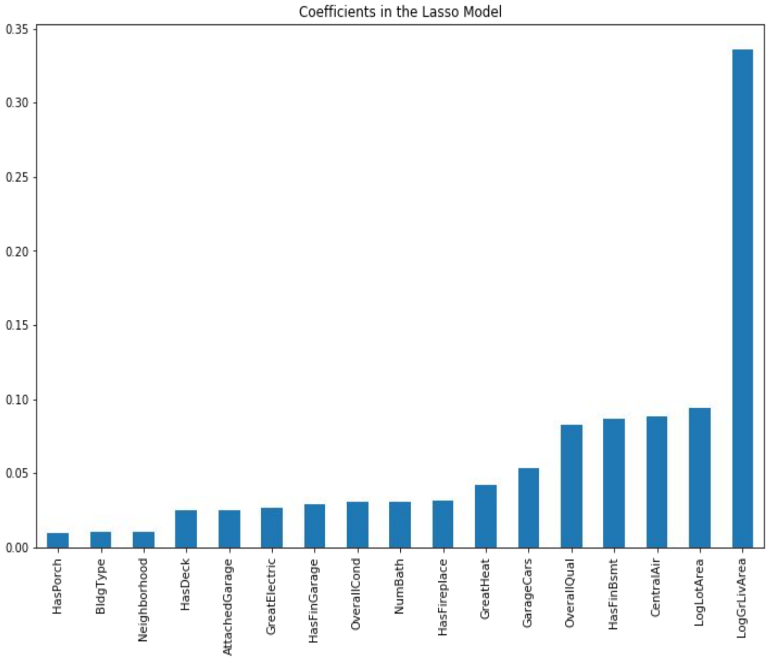


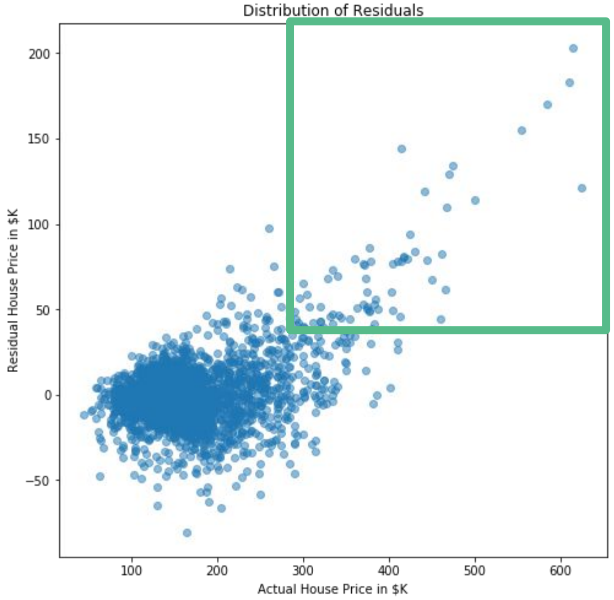
Draft Model Rev. 2

<u>Train Score</u>
0.908

<u>Test Score</u>
0.899

<u>Mean Error</u>
$22,620





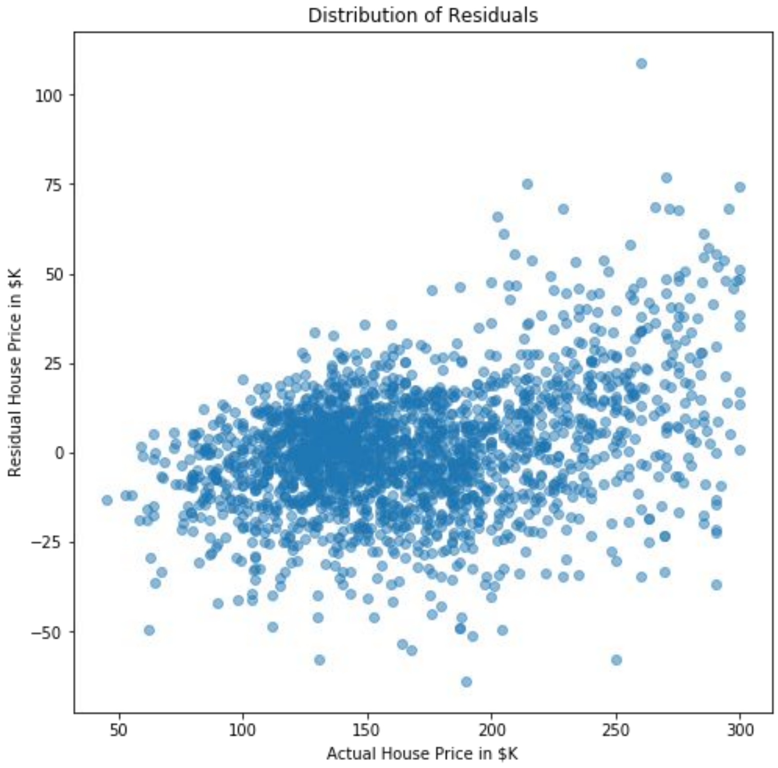
Final Model

Coeff	Features
0.313	**LogGrLivArea**
0.099	HasCentralAir
0.085	LogLotArea
0.079	HasFinBsmt
0.070	**OverallQual**
0.052	GarageCars
0.043	HasGreatHeat
0.037	NumBath
0.035	HasFireplace
0.033	HasAttchGarage
0.031	**OverallCond**
0.031	HasGreatElectric
0.023	HasFinGarage
0.023	HasDeck
0.013	HasPorch
0.011	BldgType
0.010	**Neighborhood**

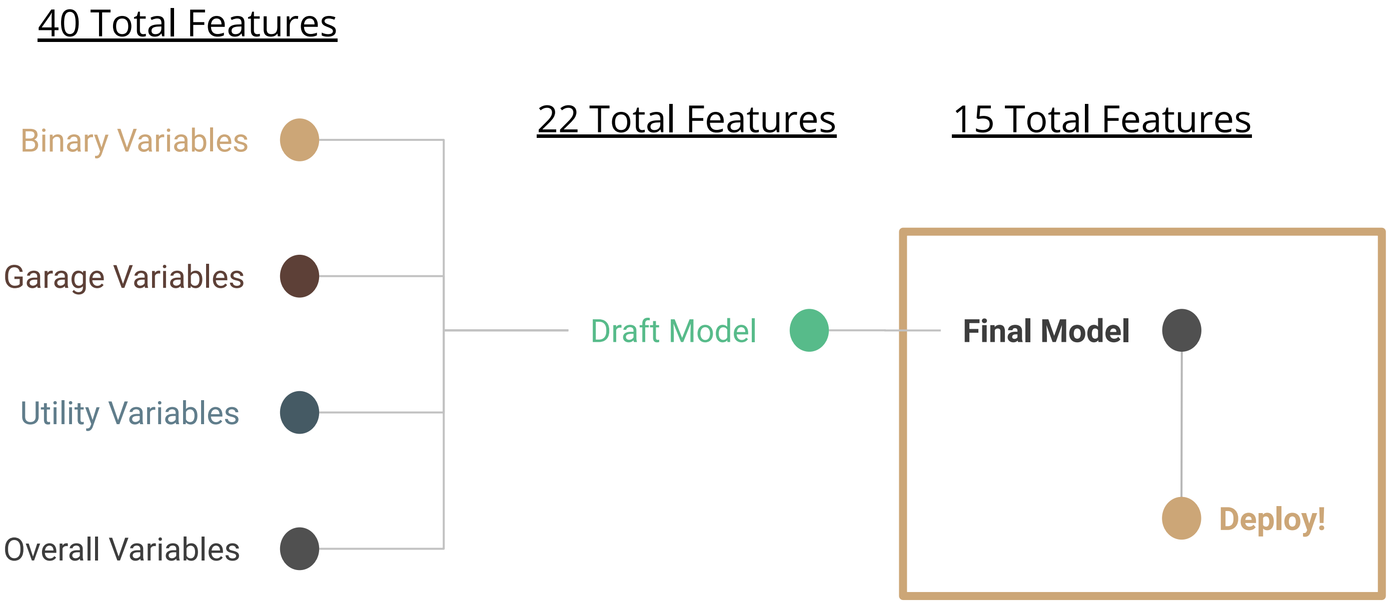
Train Score
0.890

Test Score
0.880

Mean Error
$17,213



Linear Model



Model Deployment

https://ames-housing-app.herokuapp.com/

Scenario 1: Over Budget Young Professional

Budget

100000

Please enter your budget in $

Gross Living Area (sqft)

700

Total above ground living area

Lot Area (sqft)

300

Total outside lot area

Overall quality

8

Overall material and finish of the house

Overall condition

8

Overall condition of the house

Neighborhood

Somerset

Building Type

1 Family

Number of Bathrooms

1

Number of cars in garage

1

Select features you would like in your home

- Finished Basement
- Finished Garage
- Fire Place
- Porch
- Deck
- Attached Garage
- Great Electric
- Great Heat
- Central Air

Predicted Price: $122,127

You are Over Budget.

Recommendations:

1. Reduce OverallQual to 6 to save $19,169. New predicted price: $102,958
2. Reduce Gross Living Area to 500sqft to save $12,696. New predicted price: $109,431
3. Reduce Lot Area to 100sqft to save $12,070. New predicted price: $110,057
4. Remove Central Air to save $11,401. New predicted price: $110,726.
5. Reduce OverallQual to 7 to save $9,994. New predicted price: $112,134
6. Reduce OverallCond to 6 to save $7,161. New predicted price: $114,967
7. Reduce Gross Living Area to 600sqft to save $5,990. New predicted price: $116,138
8. Reduce Lot Area to 200sqft to save $4,602. New predicted price: $117,526
9. Remove Great Heat to save $4,568. New predicted price: $117,560.
10. Reduce OverallCond to 7 to save $3,634. New predicted price: $118,493

Model Deployment

Scenario 2 - Under Budget College Professor

Budget

300000

Please enter your budget in $

Gross Living Area (sqft)

2700

Total above ground living area

Lot Area (sqft)

1300

Total outside lot area

Overall quality

8

Overall material and finish of the house

Overall condition

8

Overall condition of the house

Neighborhood

Stone Brook

Building Type

1 Family

Number of Bathrooms

3

Number of cars in garage

2

Select features you would like in your home

- Finished Basement
- Finished Garage
- Fire Place
- Porch
- Deck
- Attached Garage
- Great Electric
- Great Heat
- Central Air

Predicted Price: $276,533

You are Under Budget.

Recommendations:

1. Increase OverallQual to 9 to increase target by $24,645. New predicted price: $301,178.
2. Increase NumBath to 5 to increase target by $18,399. New predicted price: $294,931.
3. Increase GarageCars to 3 to increase target by $16,196. New predicted price: $292,729.
4. Increase GarageCars to 4 to increase target by $33,341. New predicted price: $309,874.
5. Increase NumBath to 4 to increase target by $9,051. New predicted price: $285,584.
6. Increase OverallCond to 9 to increase target by $8,482. New predicted price: $285,014.
7. Add Attached Garage to increase target by $7,481. New predicted price: $284,014.
8. Increase Gross Living Area to 2900sqft to increase target by $6,522. New predicted price: $283,055.
9. Change neighborhood to Northridge to increase target by $6,209.80. New predicted price: $282,742.45.
10. Increase Lot Area to 1500sqft to increase target by $3,774. New predicted price: $280,307.

Conclusion

Model	Pros	Cons
Multiple Linear	Easy to interpret Easy to deploy	Hard to meet assumptions Inaccurate > $300,000
Random Forest	Easy to meet assumptions Easy to tune hyperparameters	Not great for regression Inaccurate > $400,000
Gradient Boost	Easy to meet assumptions Accurate for all price ranges	Easy to overfit due to high variance Harder to interpret

A **combination model** that uses

- Multiple Regression: Sale Price < $300,000

- Gradient Boost: Sale Price > $300,000

Future Work

Data Analysis	**In Depth Neighborhood:** Grocery stores, bars, cafes, parks.
Feature Engineering	**Driving Distance:** Integrate with MapQuest to replace Ellipsoidal distance
Modelling	**Bundled Linear:** Do the coefficients change if trained on different neighborhoods?
Web App	**Intelligence:** Smarter recommendations (effect of changing multiple features)

Thank You

Extra Slides

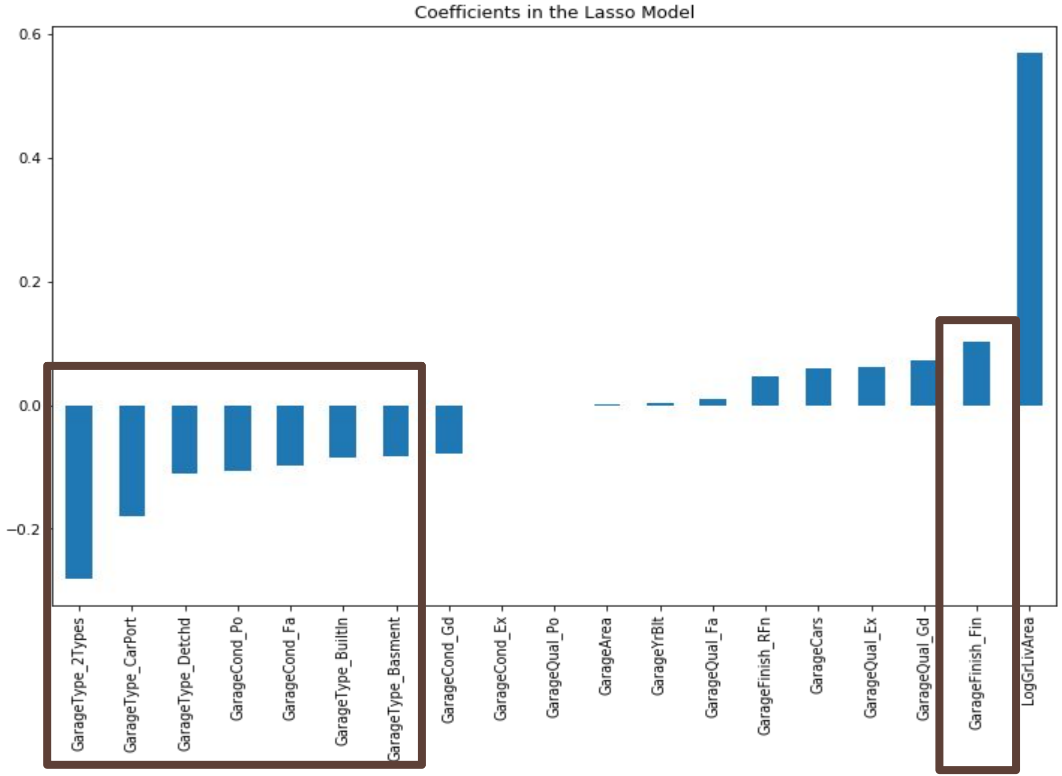
Garage Variables

Test
- *GarageType*

GarageYrBlt
- *GarageFinish*

GarageCars
GarageArea
GarageQual
GarageCond

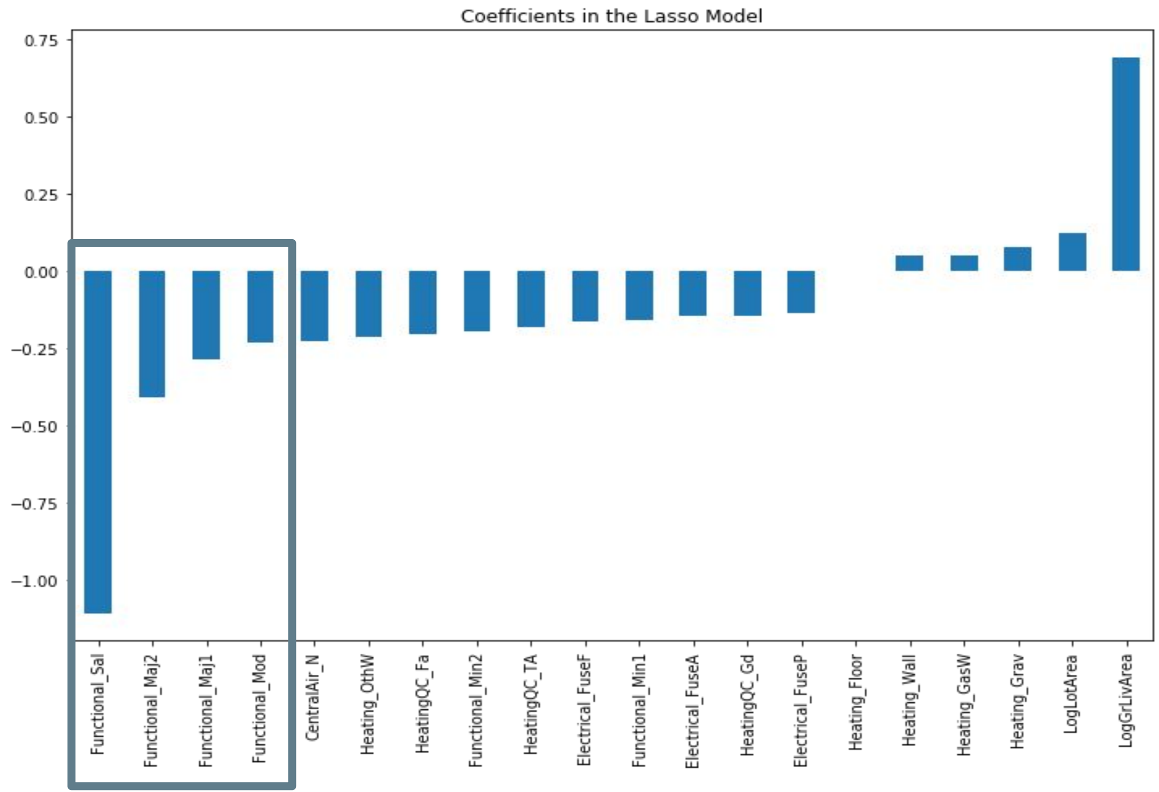
Baseline
LogLotArea
LogGrLivArea



Utility Variables

<u>Test</u>
Heating
HeatingQC
CentralAir
Electrical
Functional

<u>Baseline</u>
LogLotArea
LogGrLivArea



Overall Variables

<u>Test</u>
OverallQual
OverallCond

<u>Baseline</u>
LogLotArea
LogGrLivArea