# Ames Housing Project

•••

By: Edward Tjoe

# Problem Statement

Our client, Iowa Appraisals, is experiencing a loss of client and a dwindling market share in Ames, Iowa, due to its unreliable property appraisals.

They have approached us to create a model which provides meaningful estimates to prospective property sellers and buyers in the area and regain client satisfaction and market share from its competitors.

Our task is to create a predictive model for property appraisals in Ames, Iowa, based on historical data.

# Background

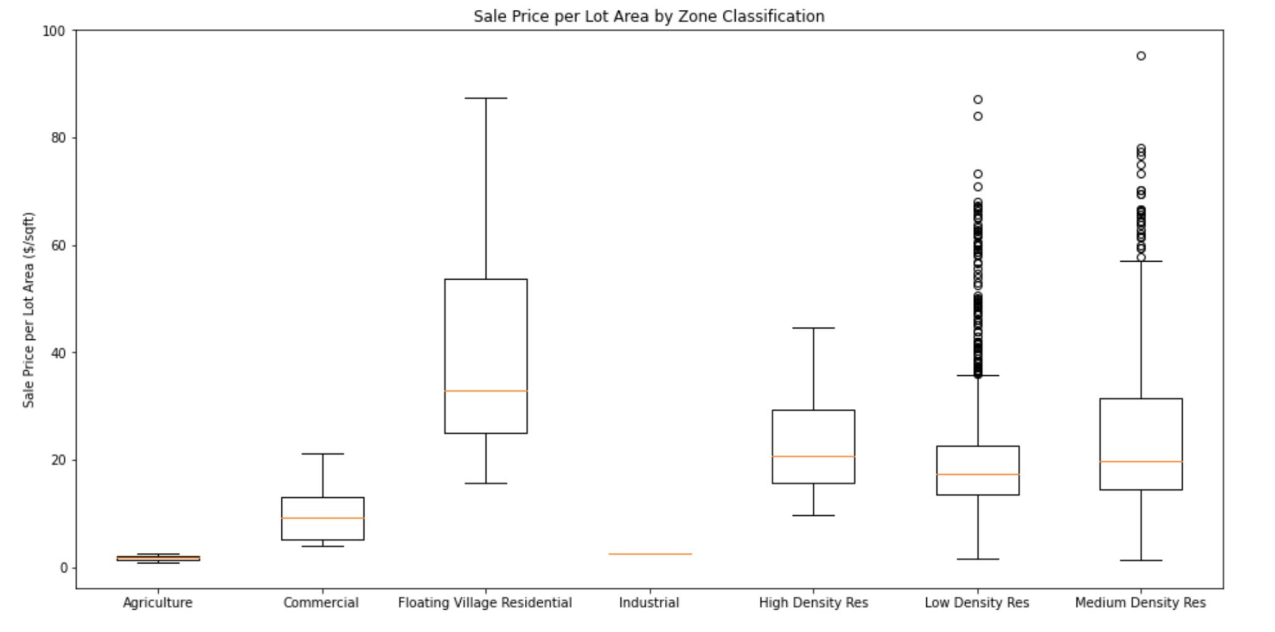
Real estate appraisals are a key component of the whole process of buying or selling a property. Appraisal agents such as Iowa Appraisals compete against one another to provide the greatest value to their clients.

One of the key metrics of client satisfaction is the based on attaining the fair market value of their homes. To fulfill the client's needs, the usage of data science is vital.

Conducted incorrectly, this leads to unsatisfied clients, lowering the competitiveness against other appraisal firms.

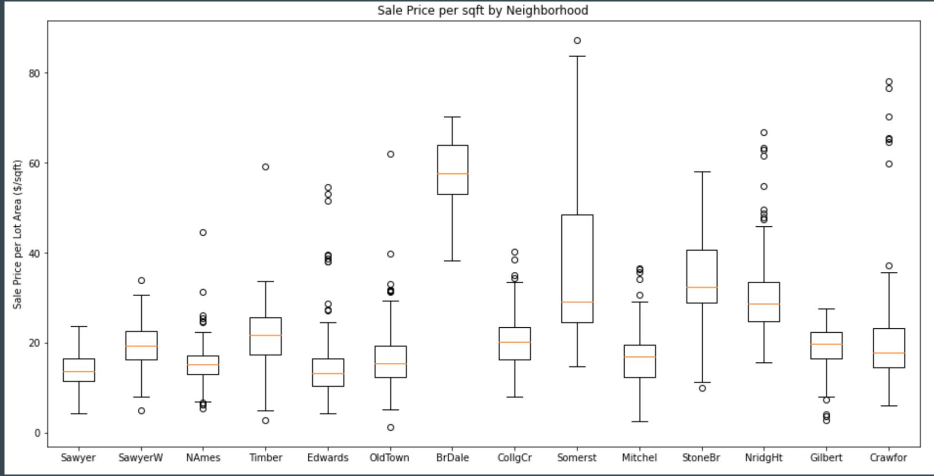
At the same time, agents who provide appropriate property values become a reliable source of appraisals.

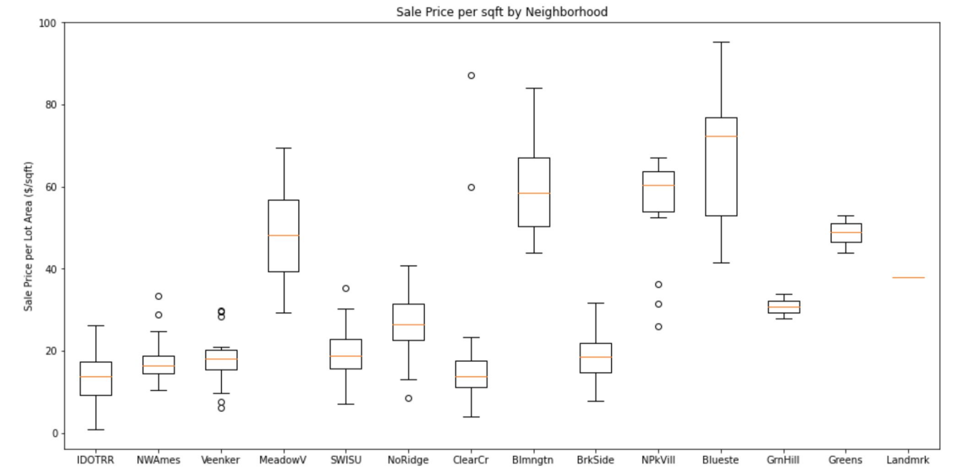
# Observation 1: Zoning Matters!



Property Zones are based on city plans, based on the initial planned usage of such land. As such, land appropriated for housing purposes are valued higher. This is an important component of our model.

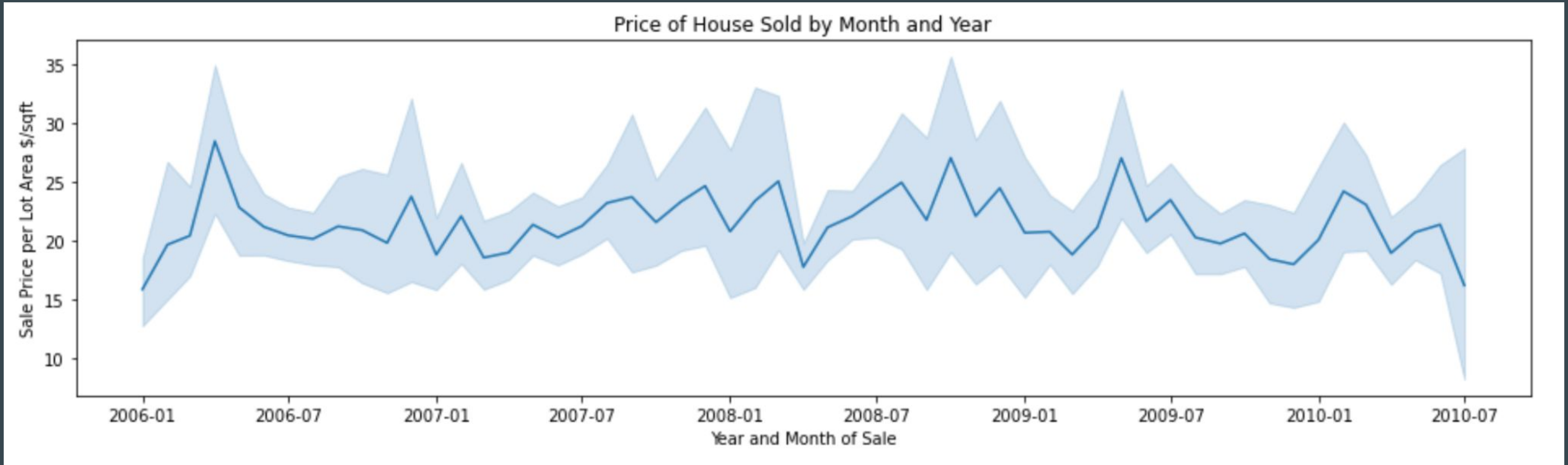
# Observation 2: Neighborhood Matters Too!





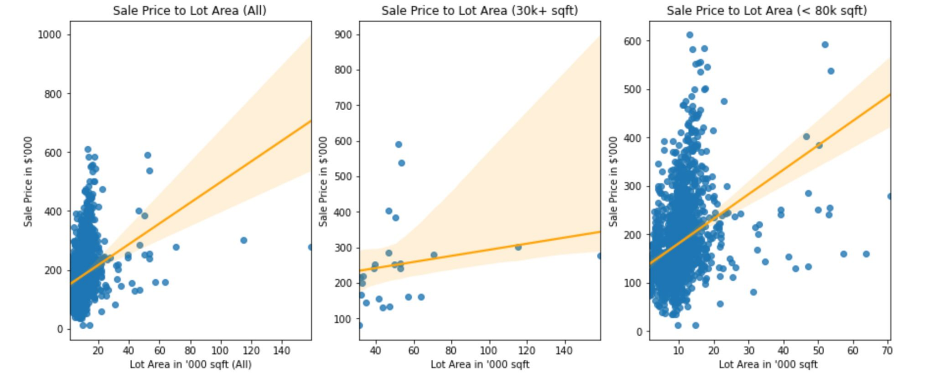
Comparing the boxplots of neighborhoods against property sale price per sqft, it seems that the neighborhood that a property belongs to plays a significant role in the sale price. This suggests that this variable should be included in our model.

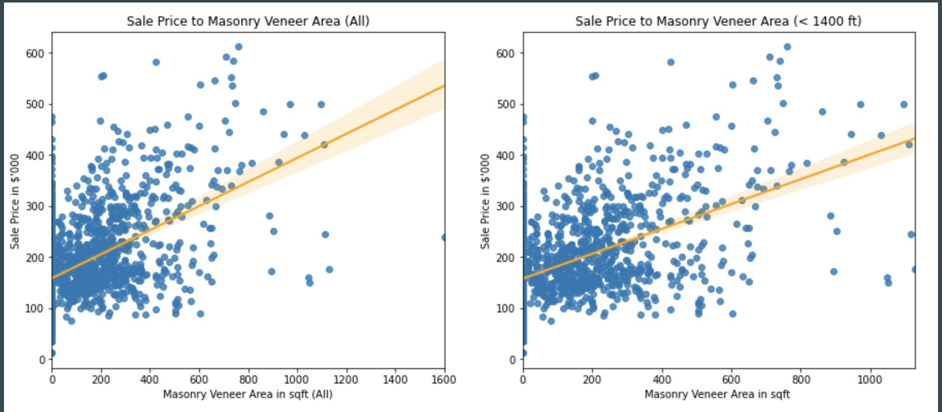
# Observation: Volatility in Sale Price



We observe volatility in sale price per lot area of properties throughout the months, but there is no reliable trend. Furthermore, this debunks the myth that ‘house prices always rise’. Seasonality of sale price per square footage may not be a reliable indicator and we therefore exclude this variable from our model.

# Crucial Removal of Outliers



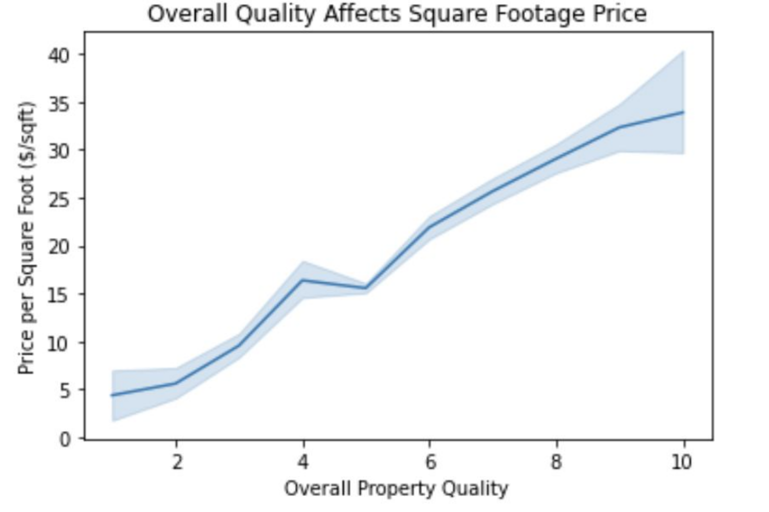


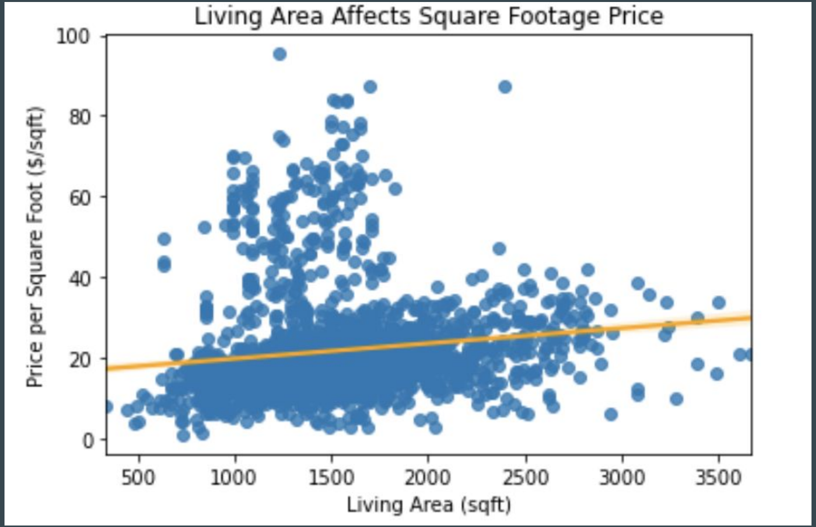
The figures to the left show the 95% confidence interval range before and after outliers are removed.

In removing the outliers, the confidence interval range drastically decreases.

Our process of removing outlier improve the predictive abilities of our models.

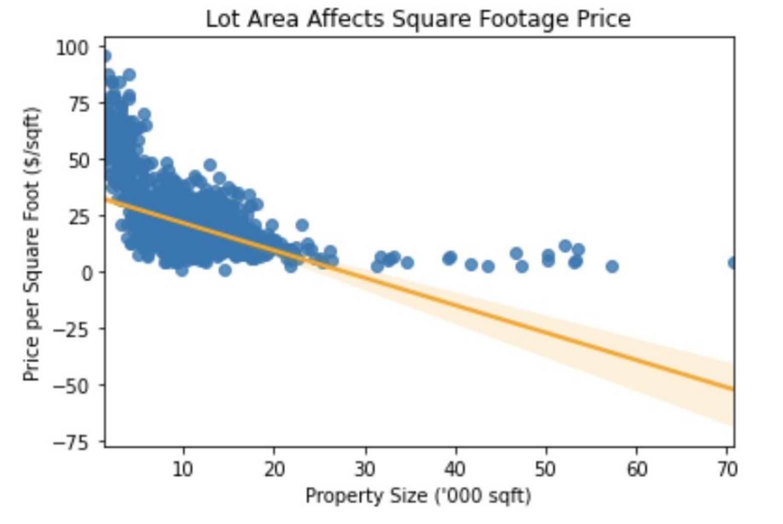
# Most Variables are Positively Correlated

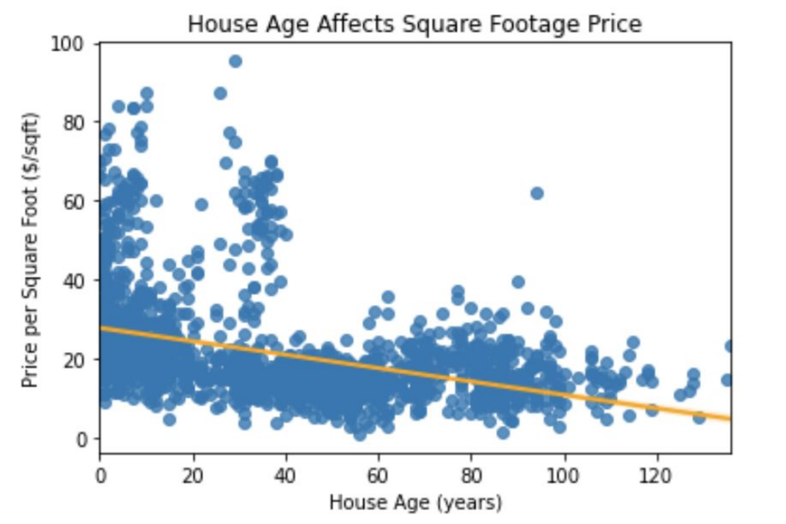




Most of our variables are positively correlated. For instance, the higher the quality or size of a property feature, the higher the property value per square foot. However, we note that some variables are more strongly correlated than others. As such, we should place greater focus on the high-impact variables.

# Some Variables are Negatively Correlated

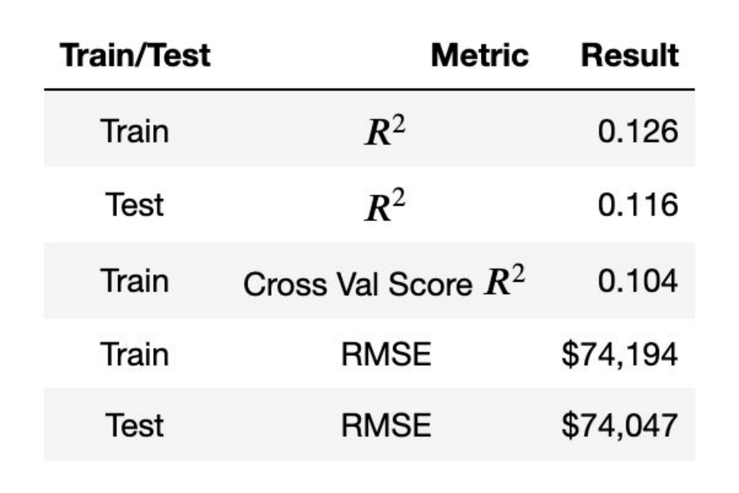




Negatively correlated variables are also present. There is an inherent premium to smaller units (shoebox units) due to the intense focus on the most crucial usage of space: housing/shelter.

As houses get older, a modest discount seems to be present, potentially due to the lower housing standards/quality of materials used.

# Baseline Model



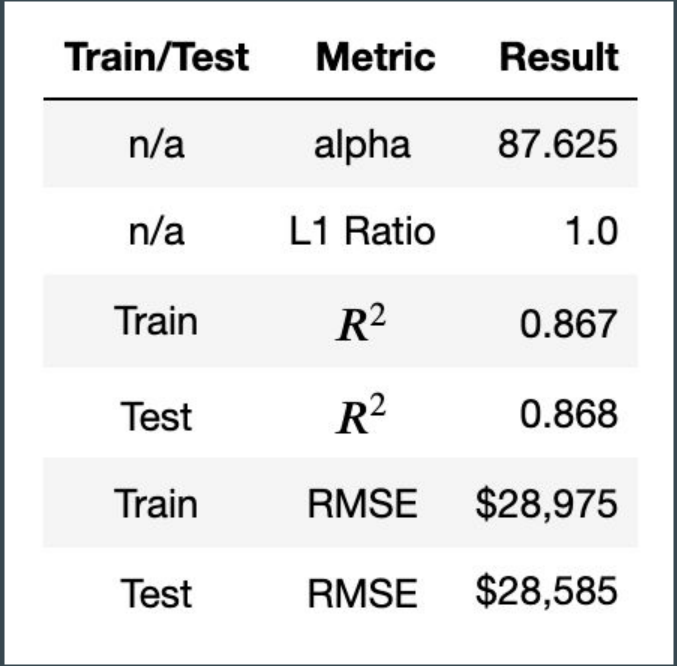
| Train/Test | Metric | Result |
|---|---|---|
| Train | $R^2$ | 0.126 |
| Test | $R^2$ | 0.116 |
| Train | Cross Val Score $R^2$ | 0.104 |
| Train | RMSE | $74,194 |
| Test | RMSE | $74,047 |

We create a baseline model based only on one variable: lot area.

This variable is used as it is the default variable to use in assessing a property.

Low $R^2$ and High Root Mean Squared Error suggests that this baseline model has a very limited predictive power.

# Elastic Net Model (Numeric Categories)



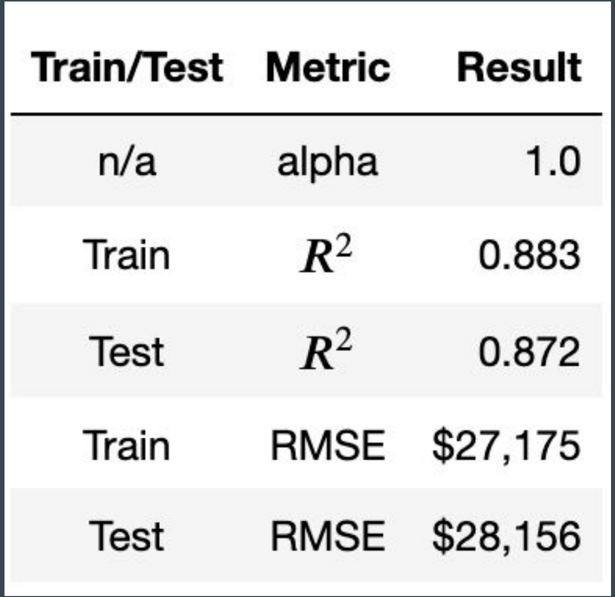
| Train/Test | Metric | Result |
|---|---|---|
| n/a | alpha | 87.625 |
| n/a | L1 Ratio | 1.0 |
| Train | $R^2$ | 0.867 |
| Test | $R^2$ | 0.868 |
| Train | RMSE | $28,975 |
| Test | RMSE | $28,585 |

We ran a elastic net model based only on numeric categories such as Garage Area and Lot Area.

The main purpose is to see whether there is a clear preference towards one type of penalization as opposed to the other.

Result: Enet model with large alpha range (1 to 100) suggests that a Lasso Model is more appropriate.

# Lasso Model 1: 11 Variables



| Train/Test | Metric | Result |
|---|---|---|
| n/a | alpha | 1.0 |
| Train | $R^2$ | 0.883 |
| Test | $R^2$ | 0.872 |
| Train | RMSE | $27,175 |
| Test | RMSE | $28,156 |

Lasso model of both quantitative and qualitative features.

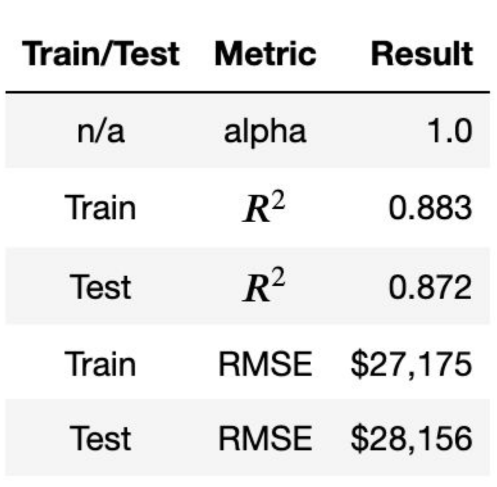
Observations:

1. Inclusion of qualitative features provide greater predictive power.
2. Feature selection allows us to have a better model even with lower number of variables.

Target RMSE: $32,044

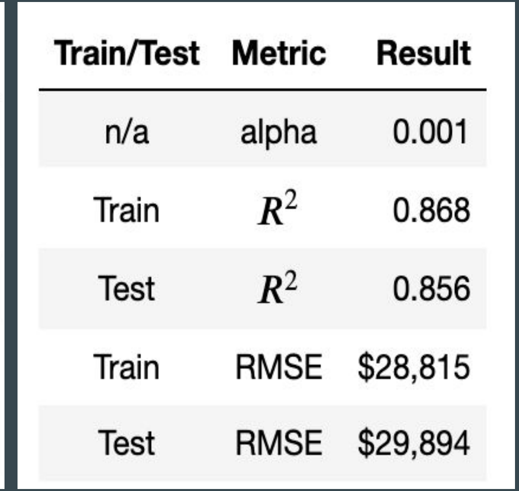
Variables Used: Neighborhood, Total Living Area, Overall Quality, Total Bsmt SF, House Age, Property SubClass, Garage Area, Lot Area, External Quality, 1st Floor Exterior, Heating Type

# Lasso Model 2: 7 Variables



| Train/Test | Metric | Result |
|---|---|---|
| n/a | alpha | 1.0 |
| Train | $R^2$ | 0.883 |
| Test | $R^2$ | 0.872 |
| Train | RMSE | $27,175 |
| Test | RMSE | $28,156 |

Lasso Model 1:
11 Variables



| Train/Test | Metric | Result |
|---|---|---|
| n/a | alpha | 0.001 |
| Train | $R^2$ | 0.868 |
| Test | $R^2$ | 0.856 |
| Train | RMSE | $28,815 |
| Test | RMSE | $29,894 |

Lasso Model 2:
7 Variables

Lasso model based on a subset of variables used in Lasso Model 1

Trade-off between bias and variance

Although we get a lower predictive value, the decrease of more than ⅓ of variables may be worth it.

Target RMSE: $31, 962

Variables Used: Neighborhood, Total Living Area, Overall Quality, Total Bsmt SF, House Age, Property SubClass, Garage Area, ~~Lot Area, External Quality, 1st Floor Exterior, Heating Type~~

# Conclusion and Recommendation

1. Only using Lot Area as a measure of Sale Price is insufficient.
2. Mix of quantitative and qualitative variables perform better due to the premiums associated with the qualitative variables not captured in quantitative variables.
3. Using less variables may be worth it: targets high-value variables and reduces noise in the predictive model
4. While this model is significantly better than the baseline, we can improve with more data. Some recommended data to obtain:
    a. Neighborhood school zone
    b. HOA fees
    c. Neighborhood crime rate
    d. Access to services (banks, supermarkets, etc)
    e. Access to highways and other public services