Ames Housing Dataset

Group 6

Yong Tat
Elliot
Joseph
Jack

Problem Statement

We are freelance data analysts building a housing price prediction model.

Stakeholders
Our prime stakeholders are property agents/agencies based in Ames, who will use the model to manage their potential buyers' expectations.

Approach
Our approach of this project is to study the historical housing prices in Ames and the housing dataset features provided by Ames, Iowa Assessor's Office. We will build a linear regression model, and will be evaluating its performance using the Root Mean Squared Error (RMSE) metric.

Data source
The housing data set provides historical housing prices from 2006 to 2010, and 80 other features related to the property, locations and sales processes.

Methodology

Exploratory Data Analysis and Data cleaning

Visualizations
Box plots
Histograms
Scatter plots

Imputation of null values
26 features with null values

Removal of outliers
2 extremely large houses with low SalePrice

Feature Engineering

Feature combination
Combining features that have similar interest and create a new feature for the combined features

Encoding of ordinal categorical features
1, 2, 3, 4, 5...

One-hot encoding
Creation of dummy variables for categorical features

Feature Selection

Barplot
Check the distribution to see if any categories have majority of values

Boxplot
To observe the effect of categorical variable to the SalePrice

Scatterplot
Checking for linear correlation between numerical variables and SalePrice

Heatmap
To check the collinearity between the numerical variables

Model Building and Model Iteration

Regularization
Use of regularization models like Ridge, Lasso and ElasticNet for feature selection

RFE
Use Recursive Feature Elimination to select top features

Model Inference and Conclusion

Insights from our model
Magnitude and direction of coefficients

Model limitations
Linearity assumptions

Recommendations
How does this address our problem statement?

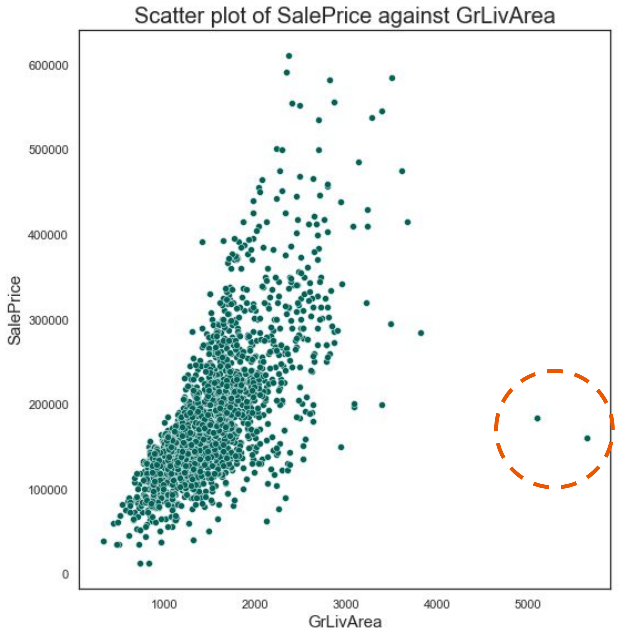
Data Cleaning

```
PoolQC          2042
MiscFeature     1986
Alley           1911
Fence           1651
FireplaceQu     1000
LotFrontage      330
GarageFinish     114
GarageCond       114
GarageQual       114
GarageYrBlt      114
GarageType       113
BsmtExposure      58
BsmtFinType2      56
BsmtFinType1      55
BsmtCond          55
BsmtQual          55
MasVnrType        22
MasVnrArea        22
BsmtHalfBath       2
BsmtFullBath       2
GarageCars         1
GarageArea         1
BsmtUnfSF          1
BsmtFinSF2         1
TotalBsmtSF        1
BsmtFinSF1         1
dtype: int64
```

Handling of null values

Categorical features were imputed with **'None'**

Numerical features were imputed with the **mean/median**

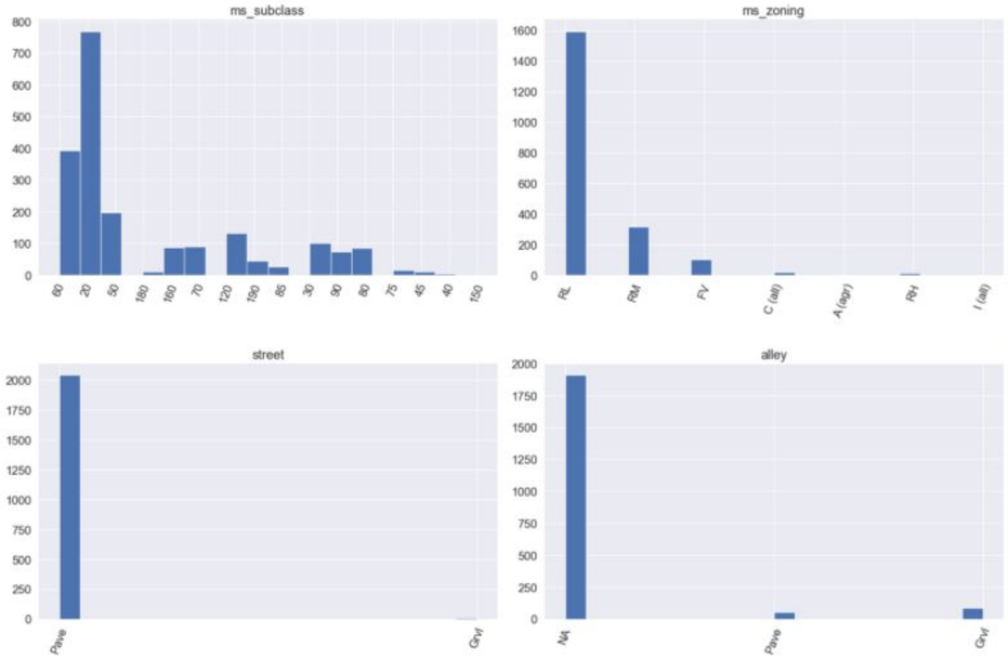


Dropping of outliers

A plot of SalePrice vs. GrLivArea reveals two transactions that had **GrLivArea > 5000 sq ft**

These were dropped to improve the linear fit of our model

Feature Selection (Bar Plot)

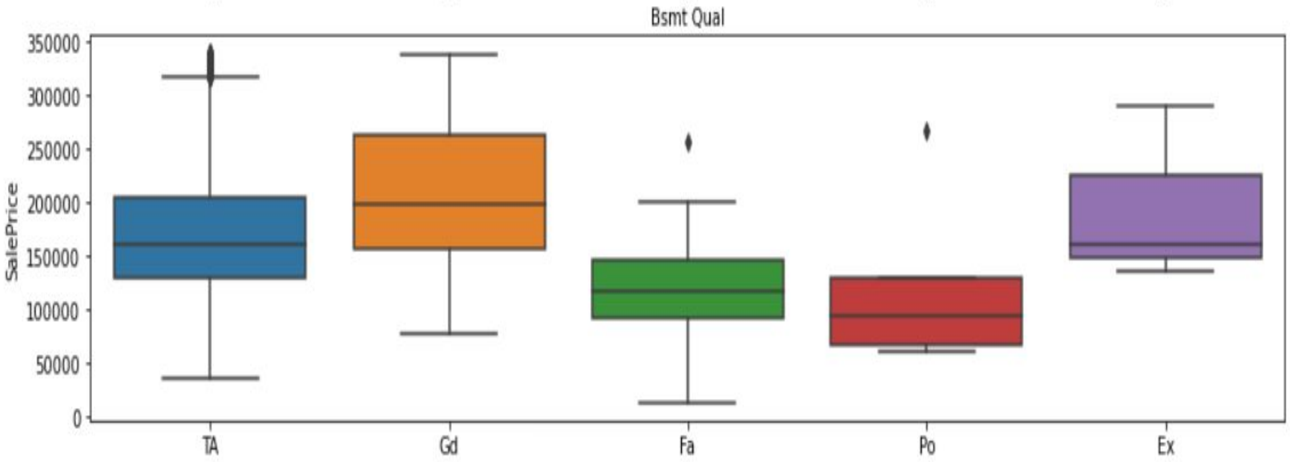



```
ms_subclass counts of unique rows in percentage:
20      37.53
60      19.18
50       9.66
120      6.44
30       4.93
70       4.39
160      4.29
80       4.20
90       3.66
190      2.24
85       1.37
75       0.78
180      0.54
45       0.54
40       0.20
150      0.05
Name: ms_subclass, dtype: float64
------------------------------------

ms_zoning counts of unique rows in percentage:
RL          77.89
RM          15.42
FV           4.93
C (all)      0.93
RH           0.68
A (agr)      0.10
I (all)      0.05
Name: ms_zoning, dtype: float64
------------------------------------

street counts of unique rows in percentage:
Pave    99.66
Grvl     0.34
Name: street, dtype: float64
------------------------------------

alley counts of unique rows in percentage:
NA      93.17
Grvl     4.15
Pave     2.68
Name: alley, dtype: float64
------------------------------------
```

We notice that there are plenty of features have one value is heavily over presented.

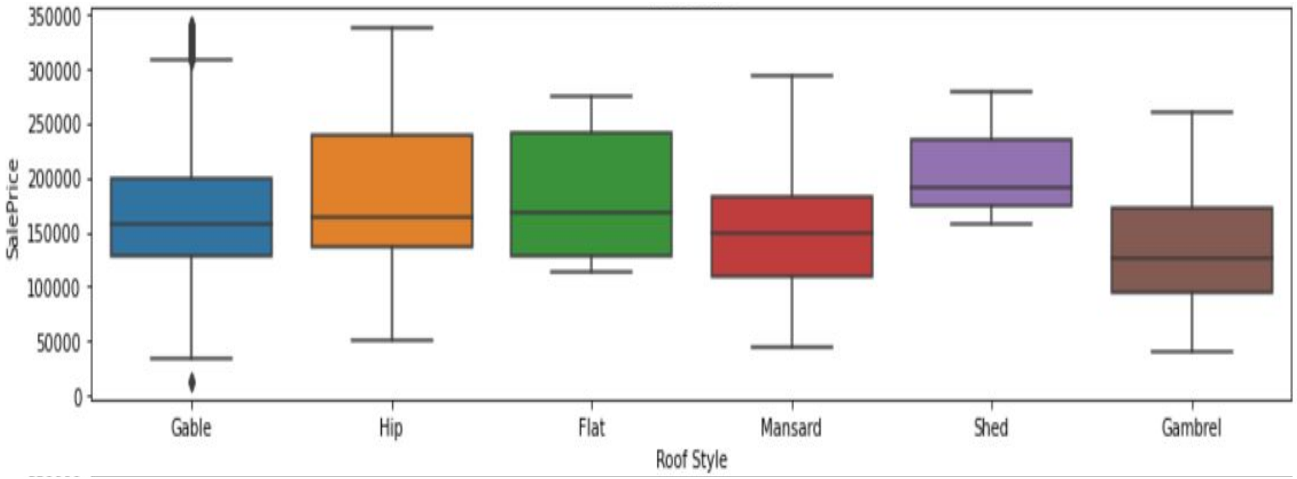
We will not use the features with frequency of one value more than 80% for my price prediction.

Feature Selection (Box Plot)



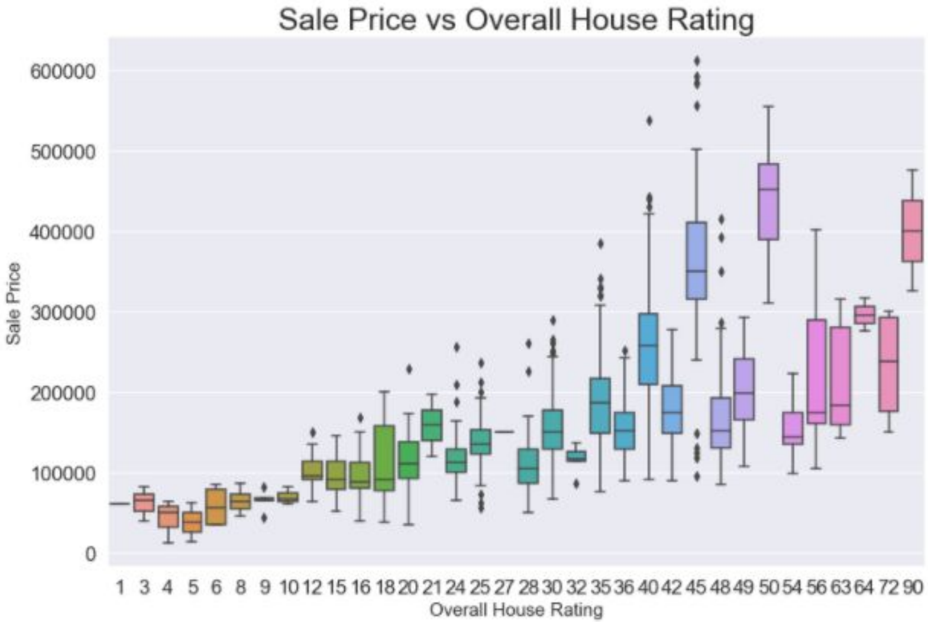
- If the categorical variable has an effect on SalePrice, the different categories will have different SalePrice.

Feature Selection (Box Plot), cont'd



- If the categorical variable has no effect on SalePrice, the different categories will have around the same SalePrice.

Feature Engineering



For the new `overall` feature, we applied the polynomial features method to combine the `overall_qual` and `overall_cond`. These 2 features are the overall rating of the house in term of materials used, finishing and condition of the house. The sub-features are based on the existing pointing system.

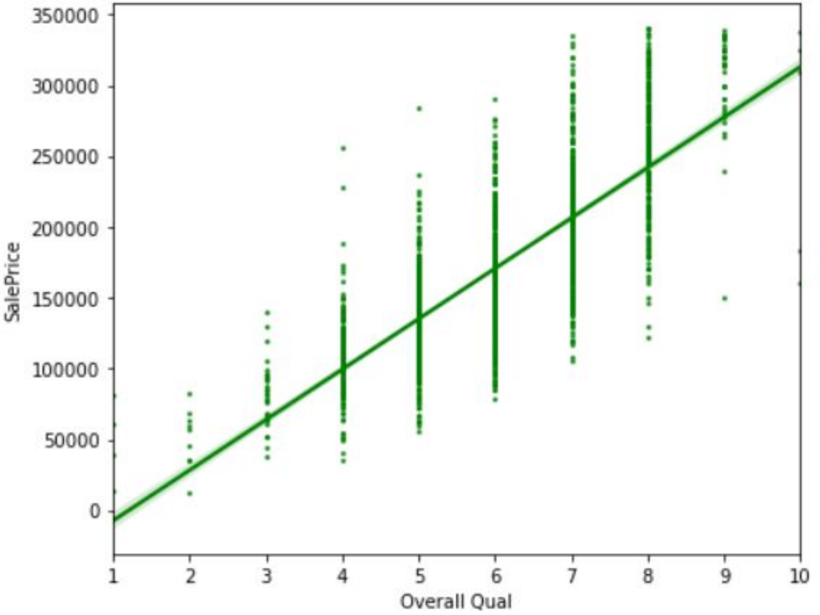
From the boxplot, you can see that increasing of the `overall` value will increase the `saleprice` which is good.

Feature Engineering

For the new `overall_age` feature, I have applied the polynomial features method to combine the `house_age` and `remod_age` features. All these features are related to age of the house with/ or without remodeling.

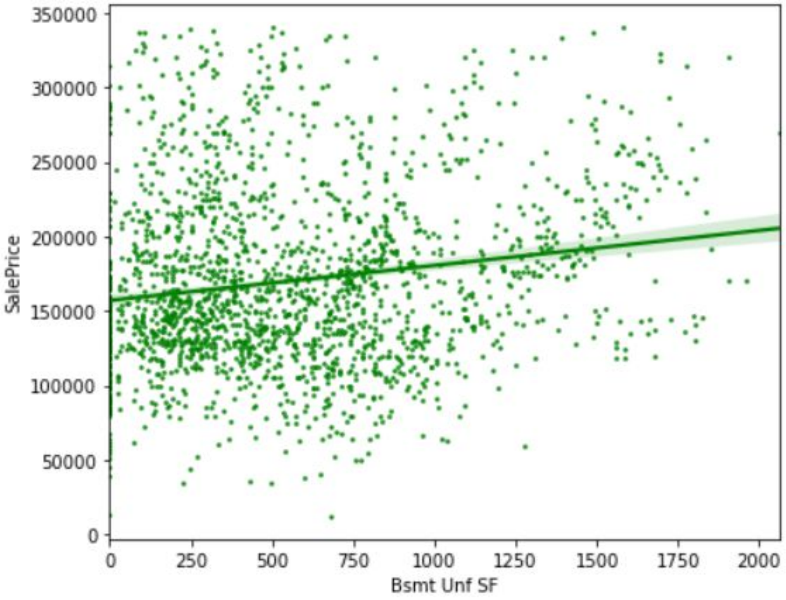
From the scatterplot, you can see that increasing of the `overall_age` value will decrease the `saleprice` which is expected.

Feature Selection (Scatter Plot)



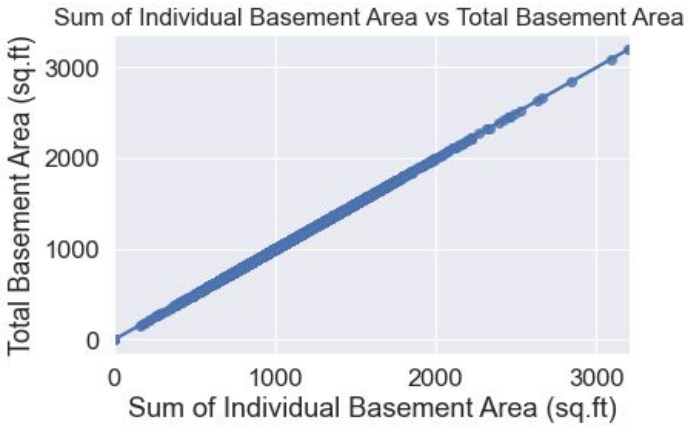
- If the variable has a linear correlation with SalePrice, it will follow the pattern like the diagram.

Feature Selection (Scatter Plot), cont'd



- If the variable has no linear correlation with SalePrice, it will follow the pattern like the diagram.

Feature Selection (Scatter Plot), cont'd



Comparison above shows that percentage of similarity between the sum of `bsmtfin_sf_1`, `bsmtfin_sf_2` and `bsmt_unf_sf` versus `total_bsmt_sf` is 100%.

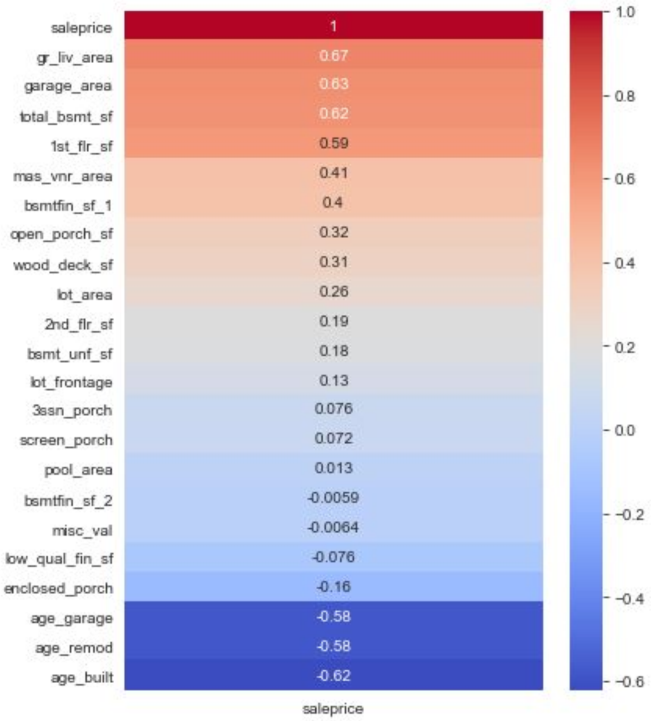
Scatter plot also shows that `total_bsmt_sf` is highly positive correlation with the sum of `bsmtfin_sf_1`, `bsmtfin_sf_2` and `bsmt_unf_sf`.

To avoid the multicollinearity, I will use **`total_bsmt_sf`** for my prediction instead of `bsmtfin_sf_1`, `bsmtfin_sf_2` and `bsmt_unf_sf` features.

For the continuous data, I will choose `gr_liv_area`, `garage_area` and `total_bsmt_sf` as my features.

- If the variables have co-linearity, it will follow the pattern like the diagram.
- The graph shows a near perfect co-linearity; total Basement Area is plotted against Sum of Individual Basement Area.
- Hence only 1 feature needs to be used.

Feature Selection (Heatmap)



- If the variable has low correlation with SalePrice, the feature can be dropped.

Model Building & Iteration

- Regularization was done with Ridge, Lasso and ElasticNet models
- ElasticNet performed the best, with lowest Adj R2 score of 0.8708 and RMSE of 21790.

```
Ridge(alpha=94.37878277775381) Performance for 126 features.
----------------------------------------------------
Estimate of Testing Adj. R2: 0.8914
Training Adj. R2: 0.9119
Test Adj. R2: 0.8706

Estimate of Testing RMSE: 19624
Training RMSE: 17791
Test RMSE: 21804
```

```
Lasso(alpha=448.700000000002) Performance for 126 features.
----------------------------------------------------
Estimate of Testing Adj. R2: 0.8916
Training Adj. R2: 0.9078
Test Adj. R2: 0.8696

Estimate of Testing RMSE: 19570
Training RMSE: 18207
Test RMSE: 21888
```

```
ElasticNet(alpha=0.09540000000000015, l1_ratio=0.30000000000000004)
----------------------------------------------------
Estimate of Testing Adj. R2: 0.8913
Training Adj. R2: 0.9121
Test Adj. R2: 0.8708

Estimate of Testing RMSE: 19634
Training RMSE: 17772
Test RMSE: 21790
```

RFE

- Using RFECV, 55 features was the optimal number given
- For a more business friendly and interpretable model, 25 features were used but this had increased RMSE

```
# Use enet model with RFECV
selector = RFECV(enet_model126, step=1, cv=5)
selector = selector.fit(X_train_scaled, y_train)
selector.n_features_
```

```
55
```

```
ElasticNet(alpha=0.058999999999999996, l1_ratio=0.2) Performance for 25 features.
----------------------------------------------------
Estimate of Testing Adj. R2: 0.8960
Training Adj. R2: 0.9035
Test Adj. R2: 0.8795

Estimate of Testing RMSE: 19940
Training RMSE: 19407
Test RMSE: 22999
```

Results

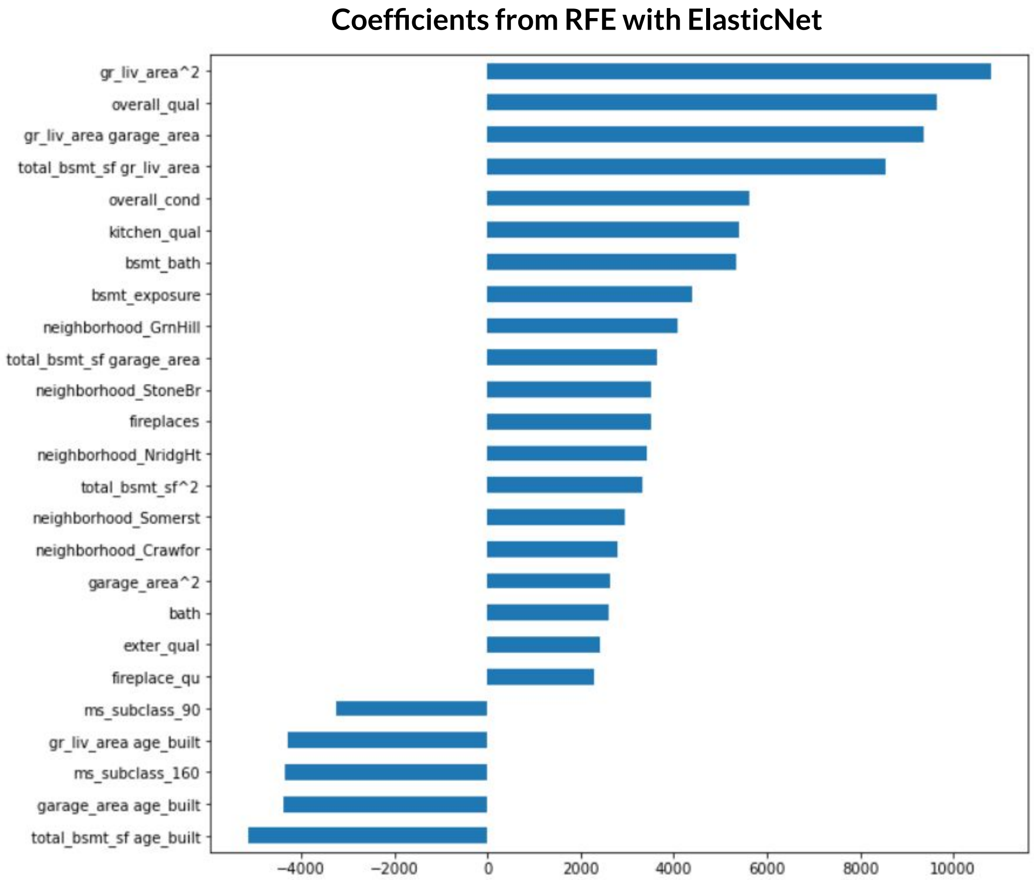
Adj R2: 0.8795
RMSE: 22999

Top positively correlated features

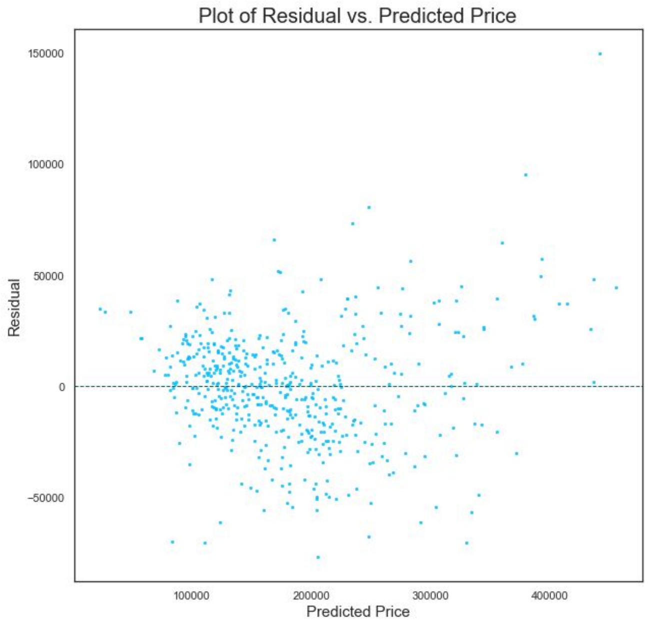
1. Above Ground Living Area
2. Overall Quality
3. Garage Area

Top negative correlated features

1. Bsmt Age
2. Garage Age
3. MS Subclass 160 (2-STORY PUD - 1946 & NEWER)



Results



- Majority of predicted prices roughly coincide with actual prices
- However, tends to undervalue houses > 350k
- Residual plot shows that **errors are not homoscedastic**
- Residuals are more **sparsely distributed** as price increases

Recommendations

Property agents/agencies can consider making use of the model to better understand the factors affecting SalePrice, so as to manage their potential buyers' expectations:

- **Gr Liv Area** (Above ground living area square feet) and **overall_qual** (overall material and finish of the house) adds the most value to a home.
- Combined effect of **Total Bsmt SF** (Total square feet of basement area) and **age_built** (Age of the property, calculated from year built) hurt the value of a home the most.
- As the size of one's house is typically already fixed, homeowners who hope to increase the value can work towards remodelling the kitchen, fireplace and basement area. Regular facade maintenance also goes a long way in bringing up the price of the house.
- The neighbourhoods of **Green Hills**, **Stone Brook**, **Northridge Heights** might be good investments.
- This model will not generalize well to other cities since it includes specific neighbourhoods by name. To make it more universal, neighbourhoods could be classified into different types instead e.g. urban, suburban.

Conclusion

- ElasticNet model with 25 features yielded an adjusted R2 score of 0.8795 and RMSE of 22999.
- Identified top few positively and negatively correlated features that can best predict housing prices in Ames
- Future Improvements:
    - Make use of polynomial features during feature engineering
    - Improving the model’s generalizability to other cities
    - Consider other aspects such as world economy crisis and changes in state government housing policies which may impact housing prices
    - Consider other machine learning algorithms

Thank You Everyone

Any Questions?

Overall

The housing prices are recorded from 2006 till 2010 whereby many other aspects are not considered such as world economy crisis, changes in state government housing policies, housing demand, land availability for development etc. All of these aspects will lead to fluctuations in housing prices.

From my perspective, we have to include all the factors capable of impacting housing prices. Doing so, we can have a better prediction of the housing price.

In addition, it is recommended that the model has to be revisited and updated with new information to ensure the model is improved.

Further research should be carried out on how to improve the feature engineering, further experiment with other models and start different polynomial features transform to improve the price prediction model.