House Features and Sales Price Prediction in Ames

Jean
Dale
Clarence

Problem Statement

According to Millionacres' Home Buyer & Seller Survey, 52% of homeowners have concerns about selling their homes predominantly due to high uncertainty with regards to property valuation.

The CEO of TechProp Co., a technology real estate company, has requested for a model to be built to conduct higher accuracy valuation and optimise the price listing found on the company's real estate portal. He also requested for a highlight of the best features that brings the most value to houses to allow clients to make informed decisions.

This model will be trained on historical transactions in Ames property market.

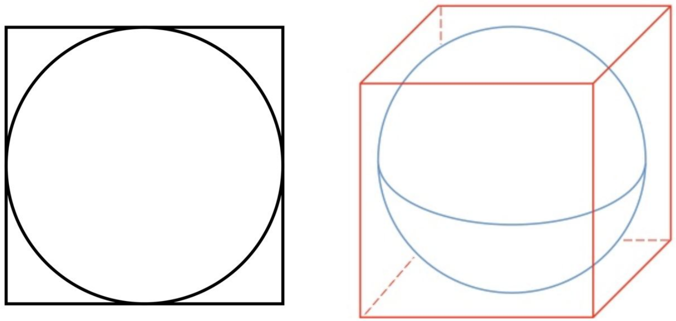
Methodology

Ockham's Law

- The simplest explanation is usually the right one. - William of Ockham

The Curse of Dimensionality

- A set of problems that arise with high-dimensional data.
- Number of features = number of dimensions
- Too many dimensions causes every observation to appear equidistant and no meaningful predictions can be formed.



Ames Housing Dataset

Total Features

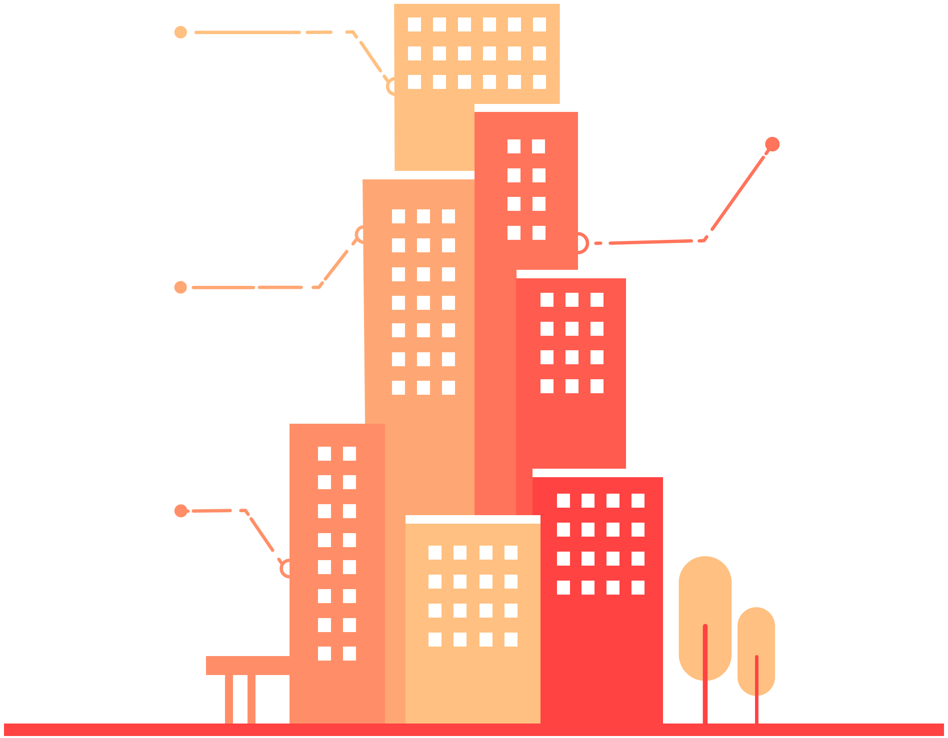
79 total features,
excluding ID

Numerical

36 features

Categorical

Ordinal: 23 features
Nominal: 20 features



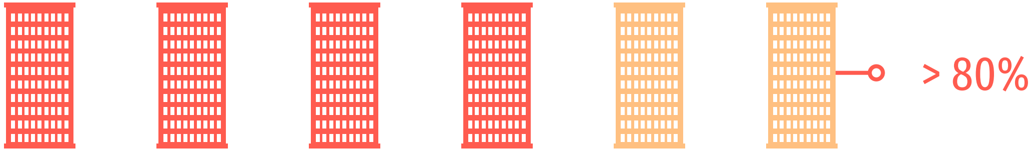
Target

Sale Price

Dataset Cleaning

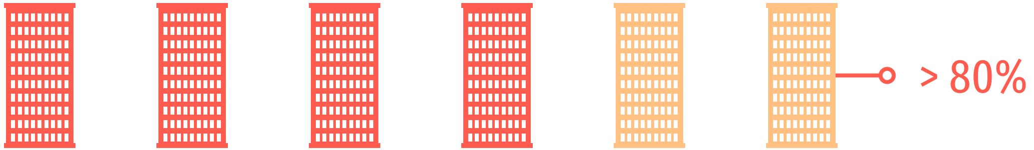
Features with missing data

Missing data percentage > 80%

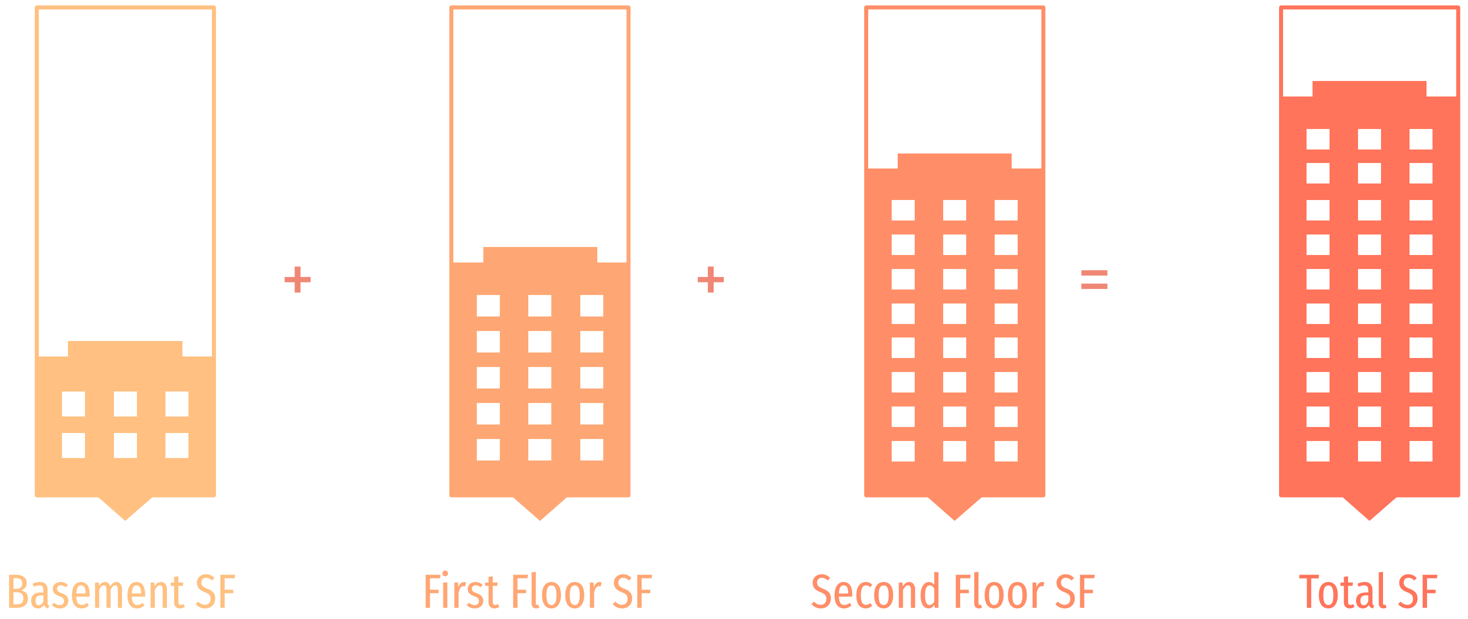


Features with high frequency of Mode Value

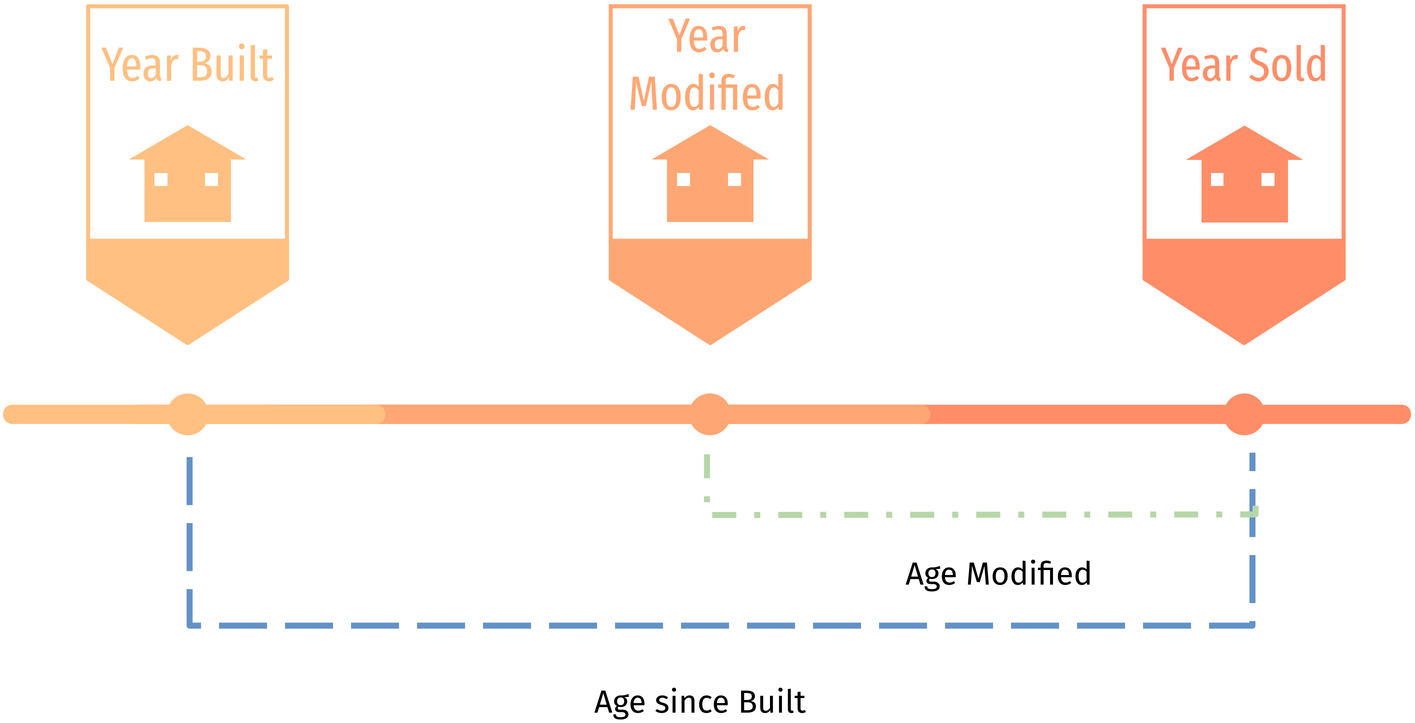
Mode data freq > 80%



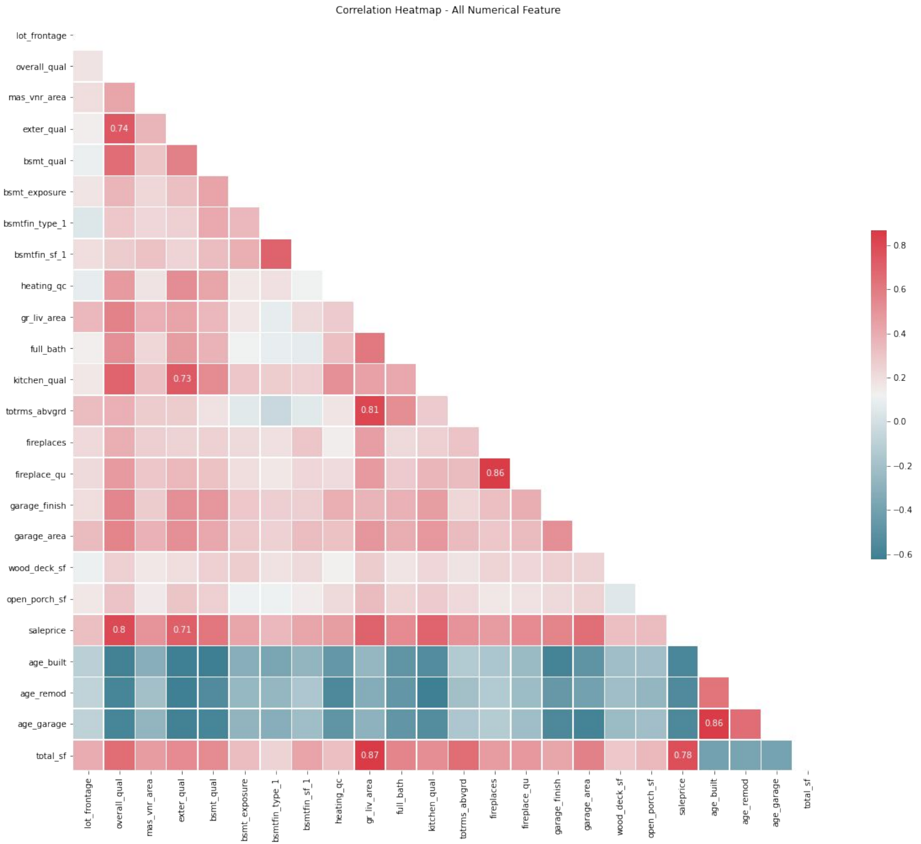
Feature Engineering



Feature Engineering



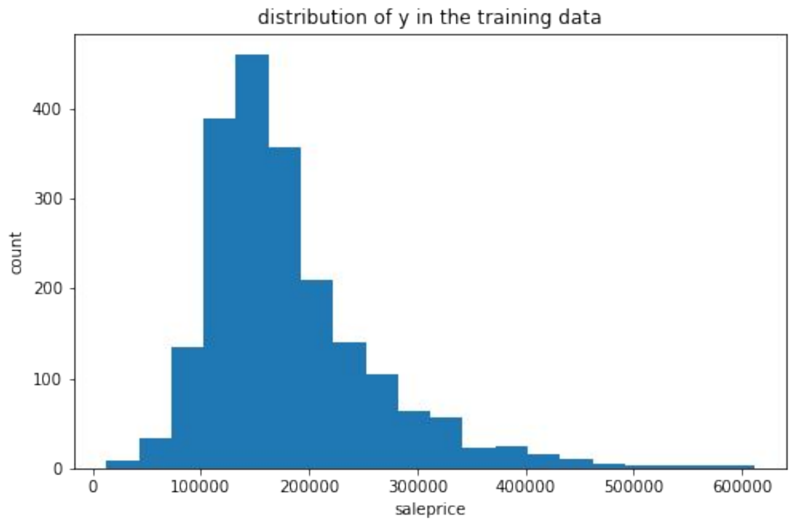
Exploratory Data Analysis

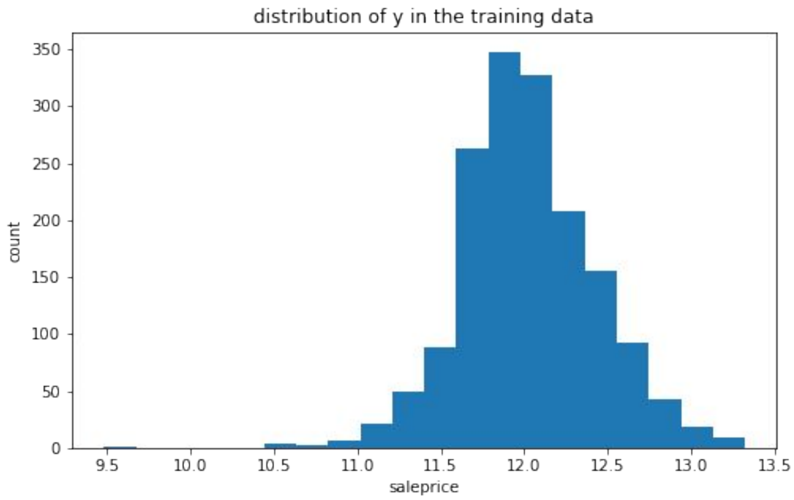


Problems with Data Multicollinearity

- Multicollinearity reduces the precision of our model
- Dependent variables should be *independent*.
- Features with high correlation weakens the statistical power of your regression model.
- Features with high correlation are therefore dropped.

Exploratory Data Analysis

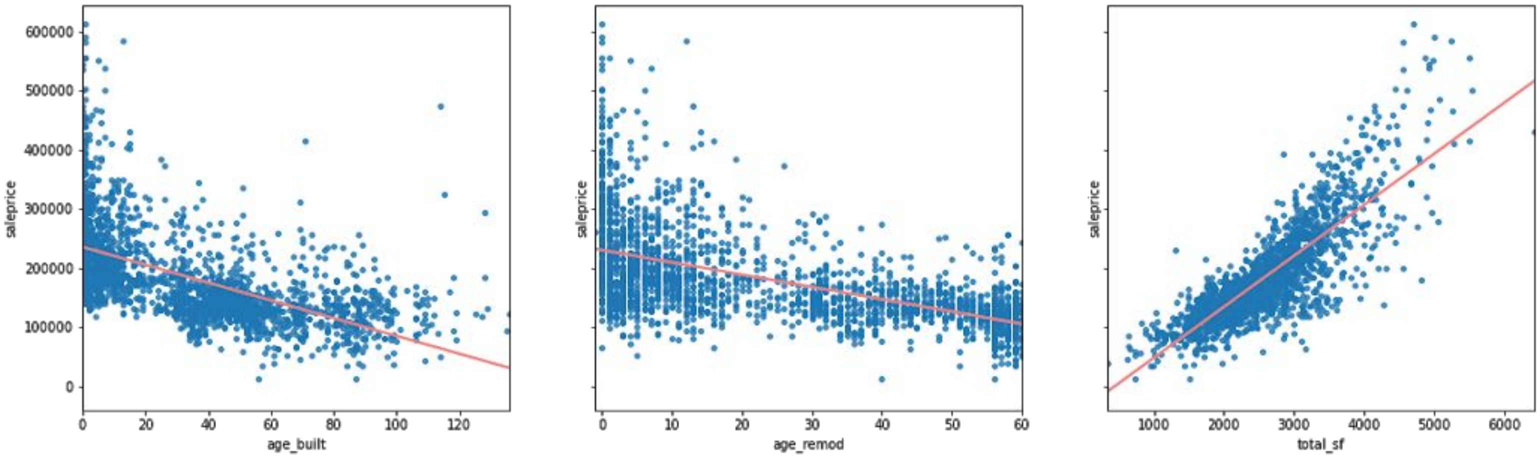




The distribution of the target was right-skewed

Hence, we did a log transformation to normalize it

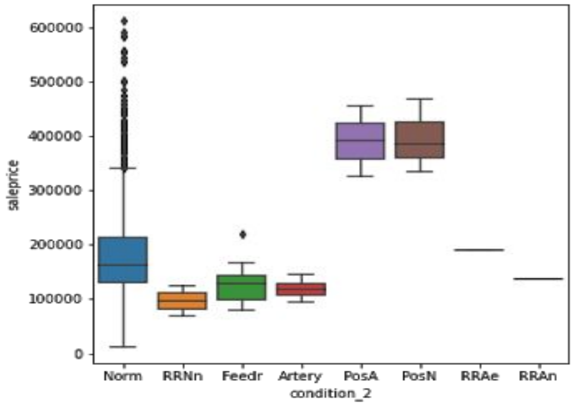
Exploratory Data Analysis

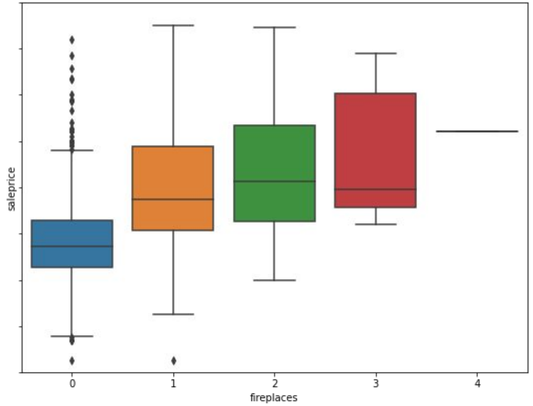


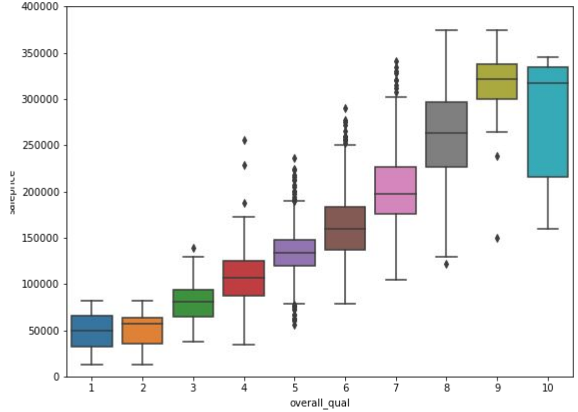
Age of the house have negative correlation to the saleprice. Same applies for modified house.

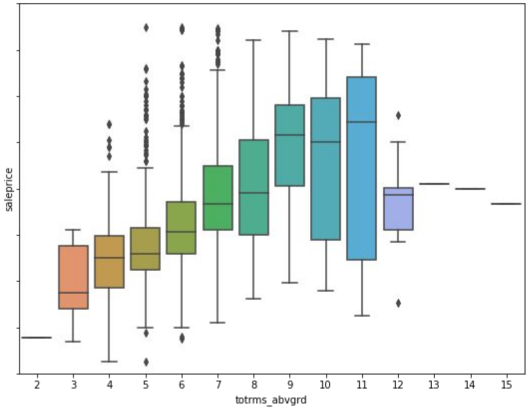
Total SF of the house have positive correlation to the saleprice.

Exploratory Data Analysis









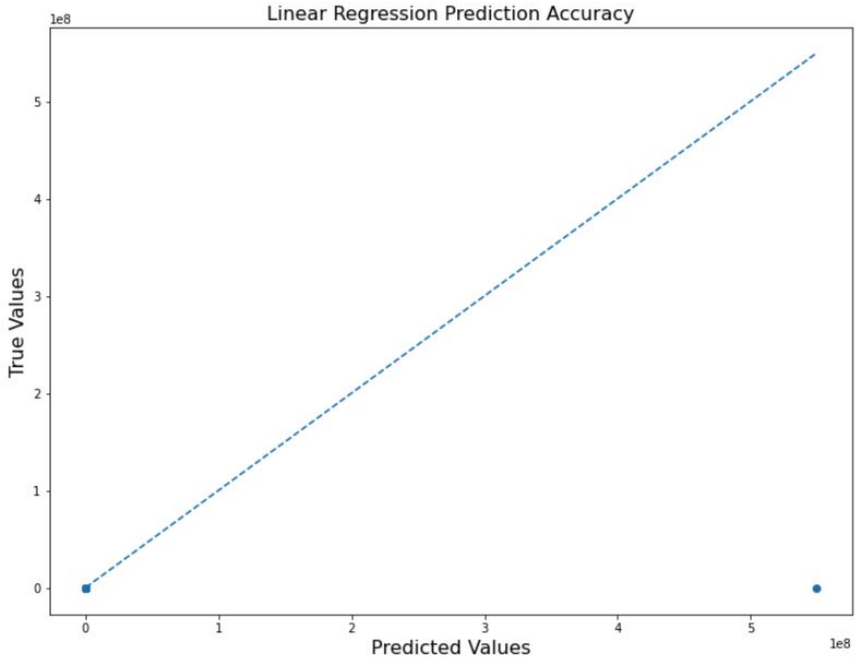
Categorical Features are evaluated with boxplots

1. Higher overall quality / number of fireplaces, drives a higher mean saleprice

2. Total rooms above ground up to 9 rooms increases the sale price. Saleprice is fluctuates between 100k to 300k with 9 - 11 rooms. Above 11 rooms, saleprice begins to dip.

3. Features with huge large gaps in mean are further evaluated and engineered prior to preprocessing

Modeling - Linear Regression



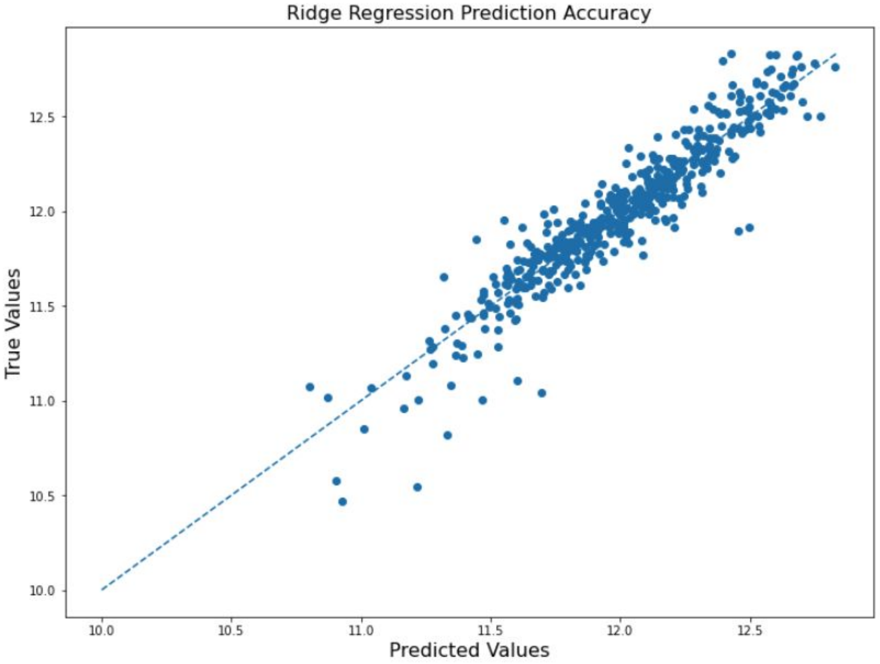

```
array([-8.63683981e-03,  6.04146279e-02,  5.28975341e-03,  9.55941047e-02,
        4.57179393e-02,  5.68946895e-02,  2.51351709e-02, -4.43803547e-03,
       -4.40816972e-03, -1.72751482e-02,  4.75685700e-02,  1.76068293e-02,
        5.29517081e-02,  2.72467587e-02,  1.68285659e-03,  2.90565651e-02,
        4.39680210e-03,  1.27446209e-02, -1.55015517e-02,  2.42610528e-02,
        3.06633018e-02,  4.11211341e-02,  6.21615976e-03, -3.74037965e-04,
        3.17672621e-03, -2.86257295e-03, -1.38458079e-02,  2.61580472e-02,
        5.49162862e+08,  1.09072990e-02,  4.19250491e-02,  3.79834920e-02,
       -4.70969033e+10, -2.96591388e+10, -2.10537768e+10, -8.05726095e+09,
       -5.52806634e+10, -2.50109999e-02, -1.58791759e-02, -3.61440798e-02,
       -6.49537125e-02, -3.04837940e-02, -8.85395350e-02, -3.90436712e-02,
       -1.07457569e-01, -8.17460774e-02, -5.88506241e-03, -6.99378844e-02,
       -4.07892644e-02, -6.79258896e-02, -1.17697566e-01, -2.73736794e-02,
       -6.88464428e-02, -3.37068075e-02, -3.92512277e-02, -1.01563561e-01,
       -3.47304062e-02, -8.02266973e-02, -7.14078151e-02, -5.57225892e-02,
       -1.71055761e-02, -4.05519700e-02, -2.66184545e-02, -5.95478879e+10,
       -1.36990013e+10, -9.30672159e+10, -8.40322341e+09, -1.45250150e+10,
       -8.38147596e+10, -2.83876994e+10, -4.09720752e+10, -6.26255993e-03,
        9.69533473e-02, -3.25291112e-03,  2.44299118e-03, -1.32895367e-02,
       -3.13918315e-02, -2.45044242e-02, -1.71247463e-02, -1.21670396e-02,
       -4.18850187e-02, -2.82931399e-02, -9.68646966e-03, -2.05463452e+10,
       -1.28471703e+11, -1.35433591e+11, -7.06444929e+10,  3.88799093e+09,
        6.03188649e+09,  6.04541229e+09,  1.54466039e+09,  6.34951017e+08,
        3.17800181e+08])
```

- Despite selecting specific features to train our model, due to high dimensionality, the model could not interpret the high complexity resulting in a large variance in coefficients

RMSE Score:

- Testing set: 24783373
- Training set: 6.572824536029952e+22

Modeling - Ridge Regression

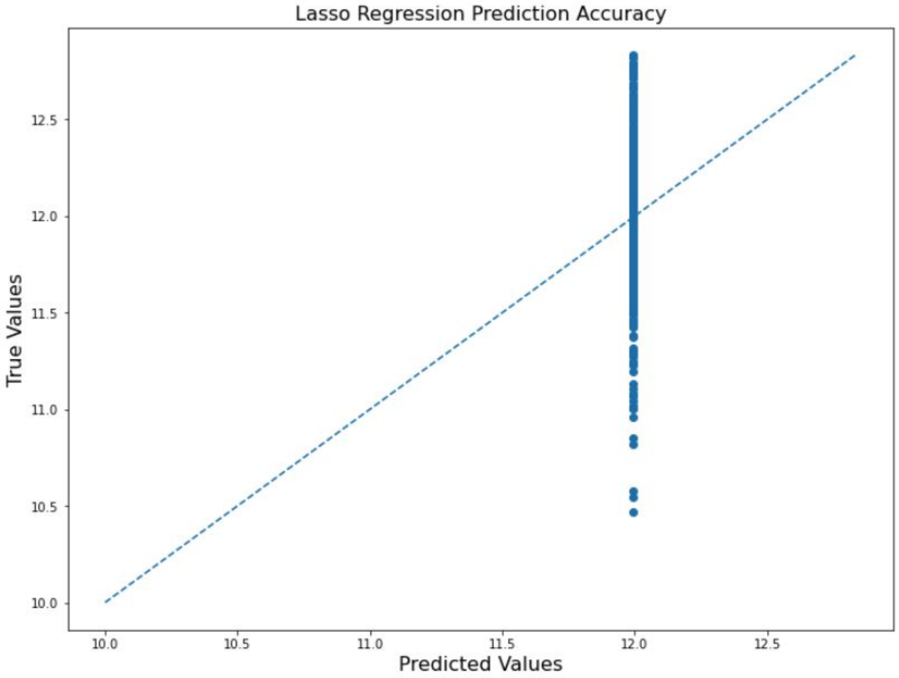



```
array([-7.32626355e-03,  5.98760046e-02,  4.28898614e-03,  9.55059078e-02,
        4.49524346e-02,  5.57260237e-02,  2.55278870e-02, -4.18236913e-03,
       -1.10567003e-03, -1.48776791e-02,  4.38350225e-02,  1.81395635e-02,
        4.86107534e-02,  2.66404453e-02,  1.40061899e-03,  3.02138116e-02,
        5.59160620e-03,  1.22556984e-02, -1.63062768e-02,  2.49239697e-02,
        3.06478668e-02,  4.08550862e-02,  5.98621859e-03, -4.93628092e-04,
        3.59893528e-03, -3.29825762e-03, -2.19281091e-02,  6.33298429e-03,
        0.00000000e+00,  5.48147062e-03,  4.53531857e-03,  6.59575126e-03,
       -9.36312604e-04,  7.19848595e-03, -5.98975872e-03, -5.66839576e-03,
        4.29683208e-05,  6.89636563e-04, -2.02742944e-03, -1.26127050e-02,
       -1.48515536e-02, -1.25478748e-03, -1.95569177e-02,  5.96247216e-03,
       -4.37704765e-02, -2.54162441e-02,  5.04296633e-03, -2.73817725e-02,
       -1.46206317e-02, -1.74556102e-02, -2.83516818e-02, -3.90913589e-03,
       -2.11113499e-02,  4.16421068e-03,  1.00654820e-02, -3.42561301e-02,
       -6.77092665e-03, -2.49101633e-02, -1.92703947e-02,  3.24892824e-03,
        9.49422730e-03, -3.52656402e-03, -4.50385316e-03,  5.16071903e-03,
        2.16993687e-03,  7.47570294e-03,  1.39669055e-03,  4.34964859e-03,
       -1.53040424e-02,  3.78659013e-03,  1.64790024e-03,  4.53460423e-04,
        0.00000000e+00, -1.08039422e-03,  1.25342136e-02, -3.57237407e-03,
       -1.22854441e-02, -5.18097890e-03, -2.03110608e-03, -6.26379388e-03,
       -1.60945230e-02, -9.29797613e-03, -1.55838181e-03, -3.69543047e-03,
        9.59050146e-03, -7.11225620e-03, -2.73116918e-03, -1.37545948e-02,
       -1.94739563e-03,  1.02576796e-02, -2.31637538e-03,  9.38700396e-03,
        2.61180499e-03])
```

- All variable coefficient had shrunk very close to zero, improving the model’s precision.

  RMSE Score:
- Testing set: 0.13114
- Training set: 0.12466

Modeling - Lasso Regression

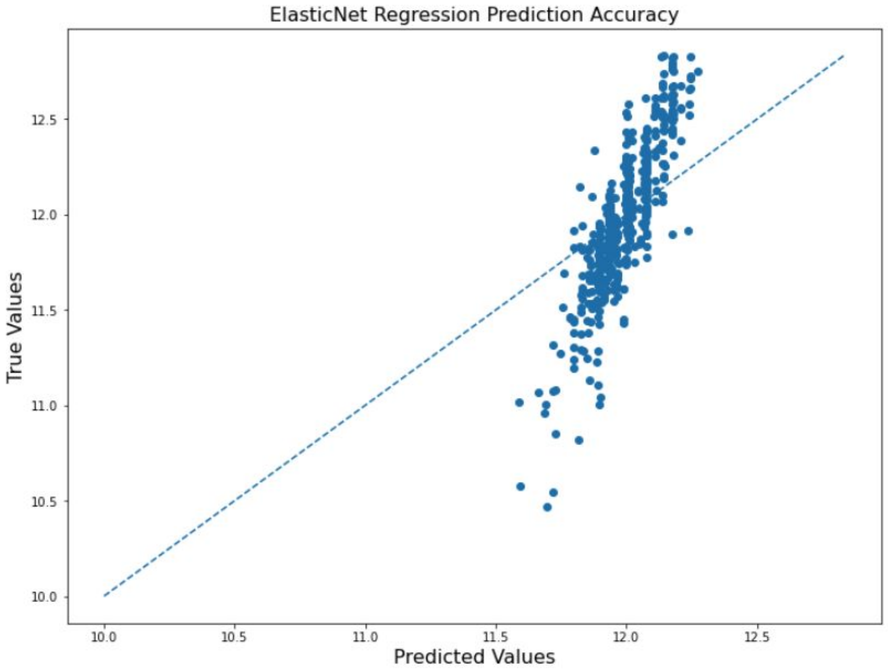



```
array([-0.,  0., -0.,  0., -0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
        0., -0.,  0.,  0.,  0., -0.,  0.,  0.,  0.,  0.,  0.,  0., -0.,
       -0.,  0.,  0., -0.,  0., -0.,  0.,  0., -0.,  0., -0.,  0., -0.,
       -0., -0.,  0.,  0.,  0., -0.,  0.,  0., -0., -0., -0., -0., -0.,
        0.,  0.,  0., -0., -0., -0.,  0.,  0.,  0.,  0.,  0., -0., -0.,
       -0.,  0.,  0.,  0., -0.,  0., -0.,  0.,  0., -0.,  0., -0., -0.,
        0., -0.,  0., -0., -0., -0.,  0., -0.,  0., -0., -0.,  0., -0.,
       -0.,  0.])
```

- The regularization method had force all the coefficients to be zero. This is therefore a poor model to use, although performed better than the linear model.

  RMSE Score:
- Testing set: 0.38969
- Training set: 0.36137

Modeling - ElasticNet Regression




```
array([-0.        ,  0.        , -0.        ,  0.08617862, -0.        ,
        0.00782843,  0.        ,  0.        ,  0.        ,  0.        ,
        0.        ,  0.        ,  0.        ,  0.        , -0.        ,
        0.        ,  0.        ,  0.        , -0.        ,  0.        ,
        0.        ,  0.02600591,  0.        ,  0.        ,  0.        ,
       -0.        , -0.        ,  0.        ,  0.        , -0.        ,
        0.        , -0.        ,  0.        ,  0.        , -0.        ,
        0.        , -0.        ,  0.        , -0.        , -0.        ,
       -0.        ,  0.        ,  0.        ,  0.        , -0.        ,
        0.        ,  0.        , -0.        , -0.        , -0.        ,
       -0.        , -0.        ,  0.        ,  0.        ,  0.        ,
       -0.        , -0.        , -0.        ,  0.        ,  0.        ,
        0.        ,  0.        ,  0.        , -0.        , -0.        ,
       -0.        ,  0.        ,  0.        ,  0.        , -0.        ,
        0.        , -0.        ,  0.        ,  0.        ,  0.        ,
        0.        , -0.        , -0.        ,  0.        , -0.        ,
        0.        , -0.        , -0.        , -0.        ,  0.        ,
       -0.        ,  0.        , -0.        , -0.        ,  0.        ,
       -0.        , -0.        ,  0.        ])
```

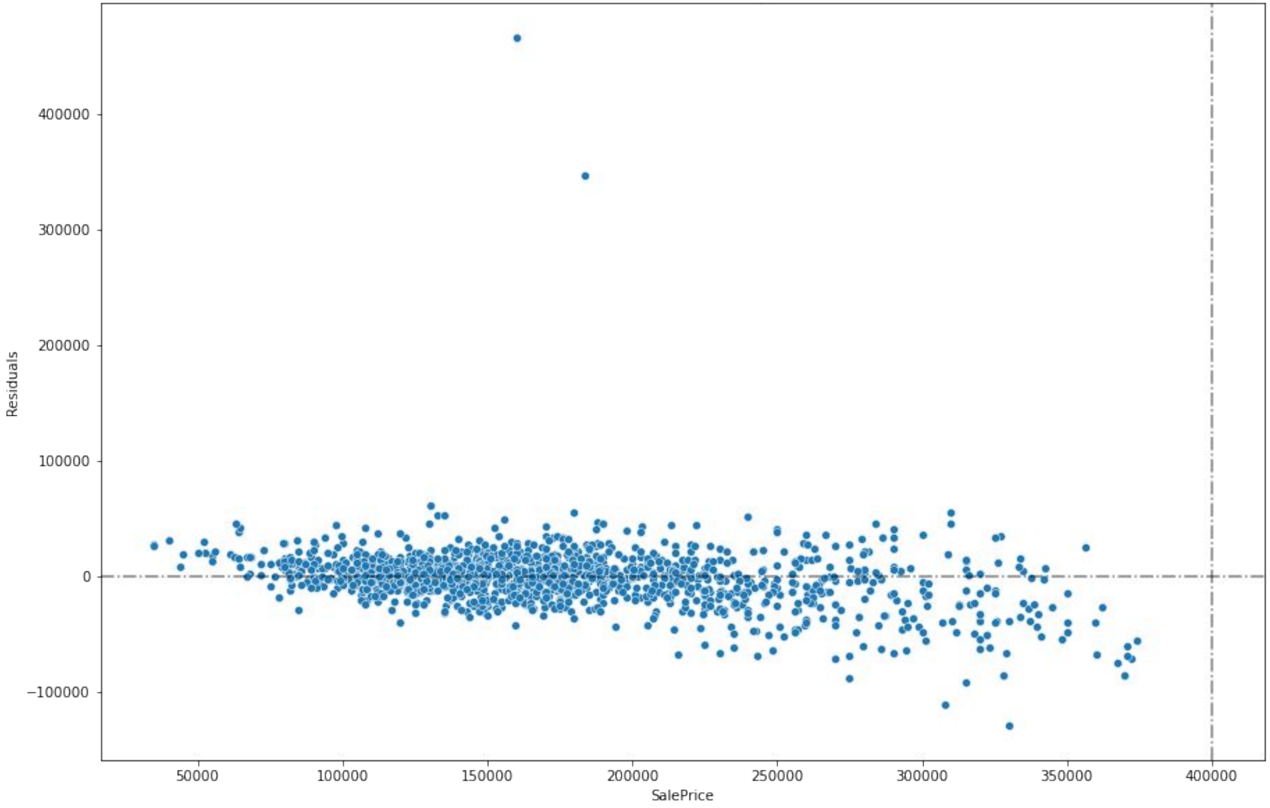
- With the combination of both Lasso and Ridge regularization, we see that the method had turned most coefficients zero and retained the strength of only 4 variables.

RMSE Score:
- Testing set: 0.93929
- Training set: 0.87868

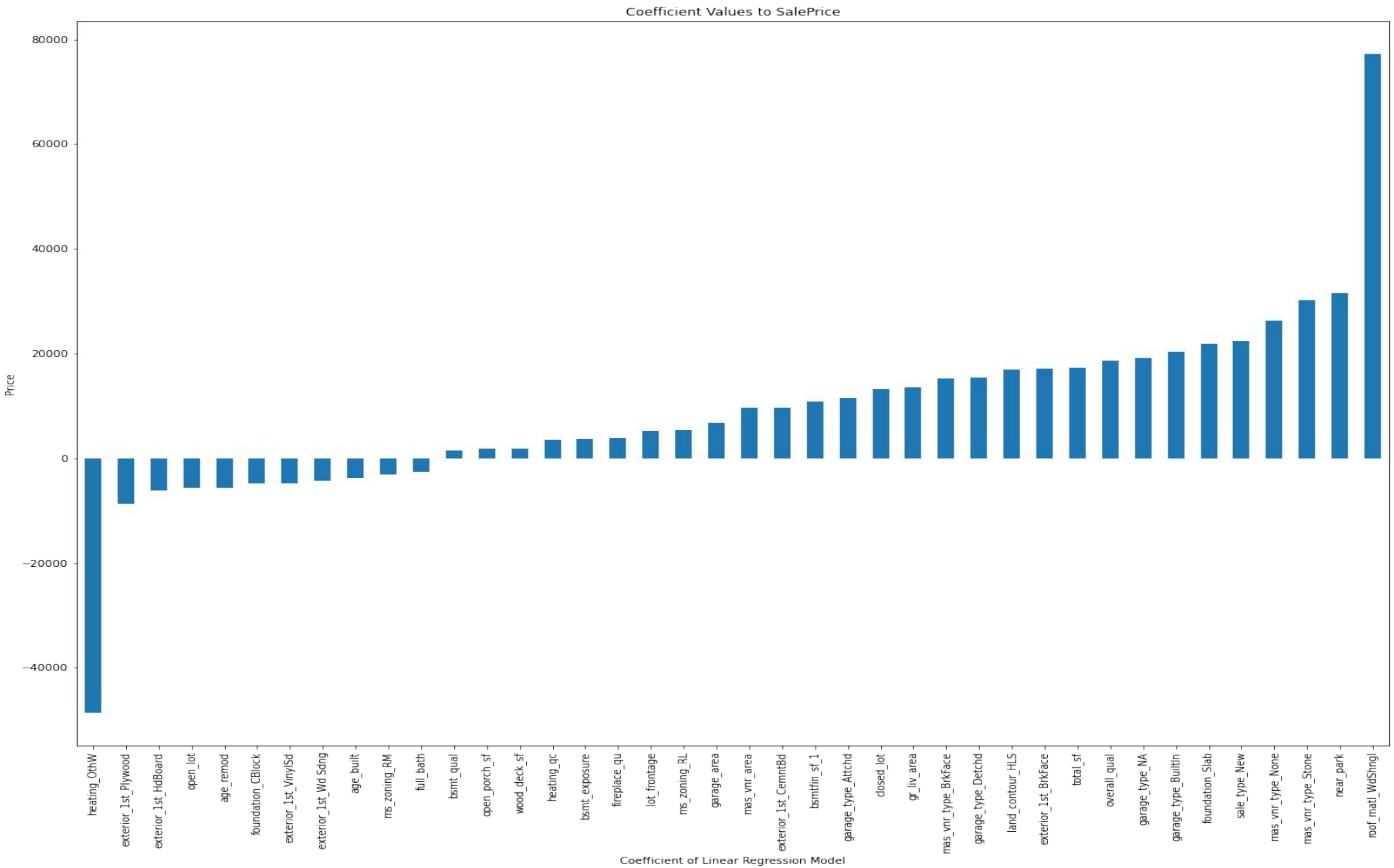
Model Selection

Regression Model	R Squared		RMSE	
	Train Set	Test Set	Train Set	Test Set
Linear	0.881892	-4.04e+15	6.57e+22	24783373
Ridge	**0.85286**	**0.856018**	**0.12466**	**0.13114**
Lasso	0	-7.16e-07	0.36137	0.38969
ElasticNet	0.40878	0.419018	0.87868	0.93929

Residual Error on Training Set



Model feature-saleprice coefficients



Hypothesis testing

***H*0: There is no correlation between the features and the saleprice of houses in the Ames Housing Dataset**

$H0$: $\varrho = 0$

***HA*: There is a correlation between the features and the saleprice of houses in the Ames Housing Dataset**

HA: $\varrho \neq 0$

Independent variables:
1. 'overall_qual'
2. 'total_sf'
3. 'gr_liv_area'
4. 'garage_area'
5. 'bsmt_qual'
6. 'exterior_1st_BrkFace'
7. 'wood_deck_sf'
8. 'exterior_1st_CemntBd'
9. 'exterior_1st_VinylSd'
10. 'garage_type_Builtin'
11. 'garage_type_Attched'

Dependent variables: 'saleprice'

Significance level: 0.45% (adjusted down for the Bonferroni correction)

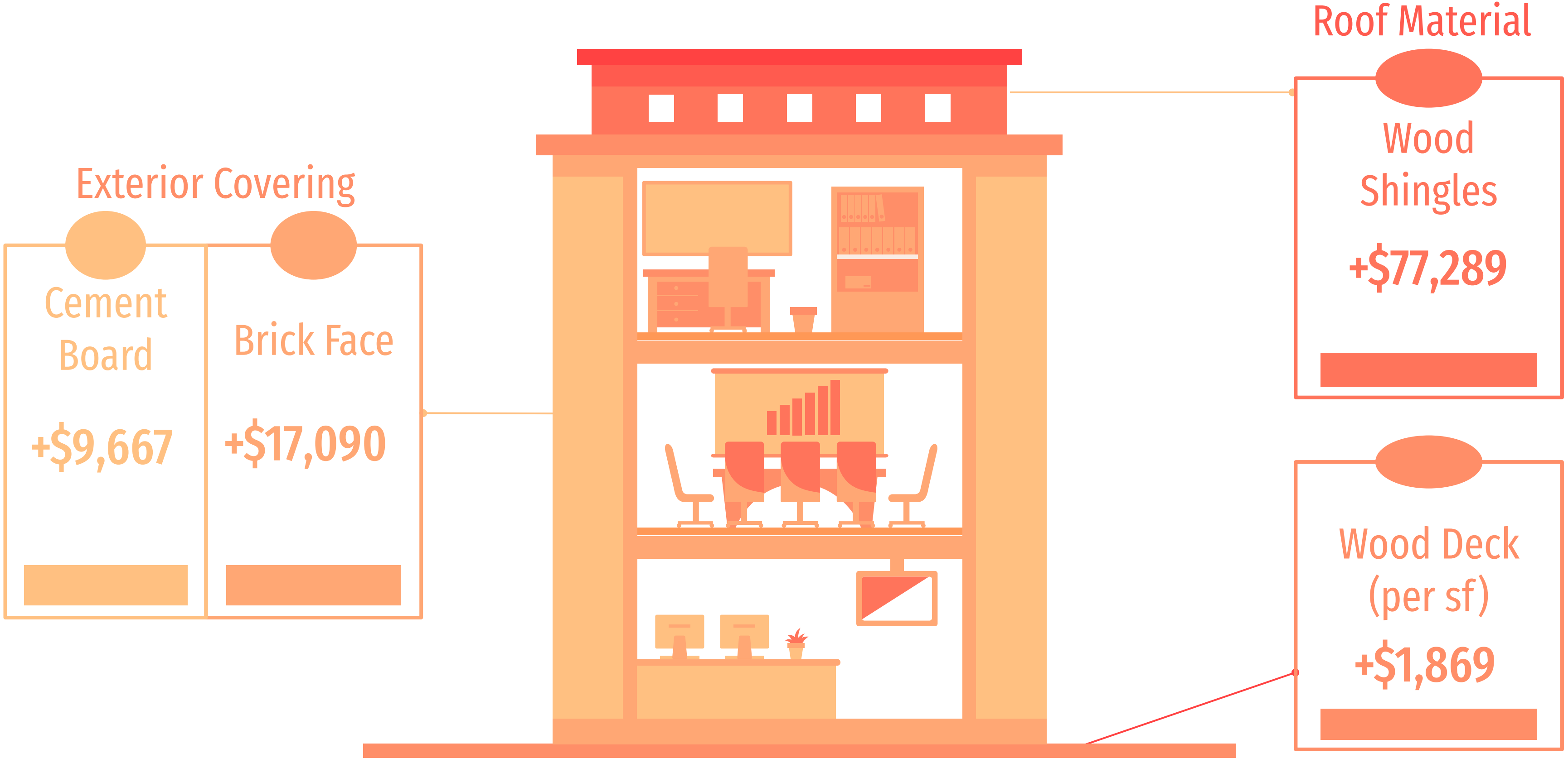
All p_values were less than 0.0001 hence the null hypothesis was rejected

	feature	corr_coef	p_val
0	overall_qual	0.805498	0.000000e+00
1	total_sf	0.835440	0.000000e+00
2	gr_liv_area	0.722026	3.258443e-263
3	garage_area	0.650506	4.209760e-197
4	bsmt_qual	0.615598	6.231406e-171
10	garage_type_Attchd	0.365005	1.238628e-52
8	exterior_1st_VinylSd	0.343485	1.971370e-46
6	wood_deck_sf	0.329478	1.195444e-42
9	garage_type_BuiltIn	0.210061	9.650031e-18
7	exterior_1st_CemntBd	0.191627	5.705991e-15

Ames City Housing Top Features

Basement Quality (0-5)	Garage Area (per sf)	Ground Living Area (per sf)	Total SF (per sf)	Overall Quality (1-10)
+$1,494	+$6,679	+$13,499	+$17,325	+$18,659

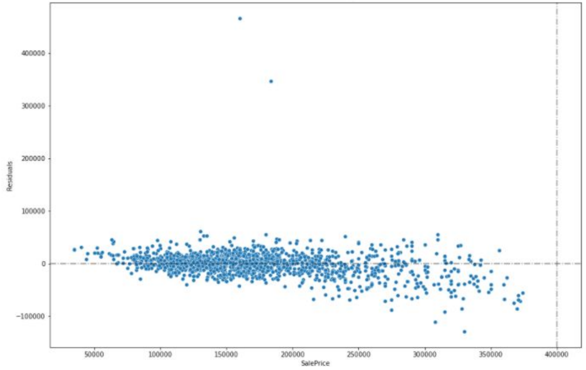
Recommendations to Increase Valuation

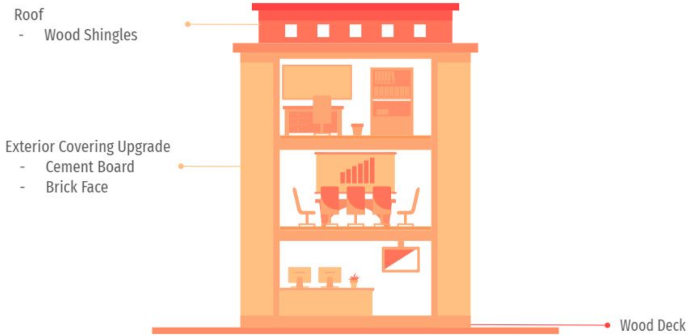


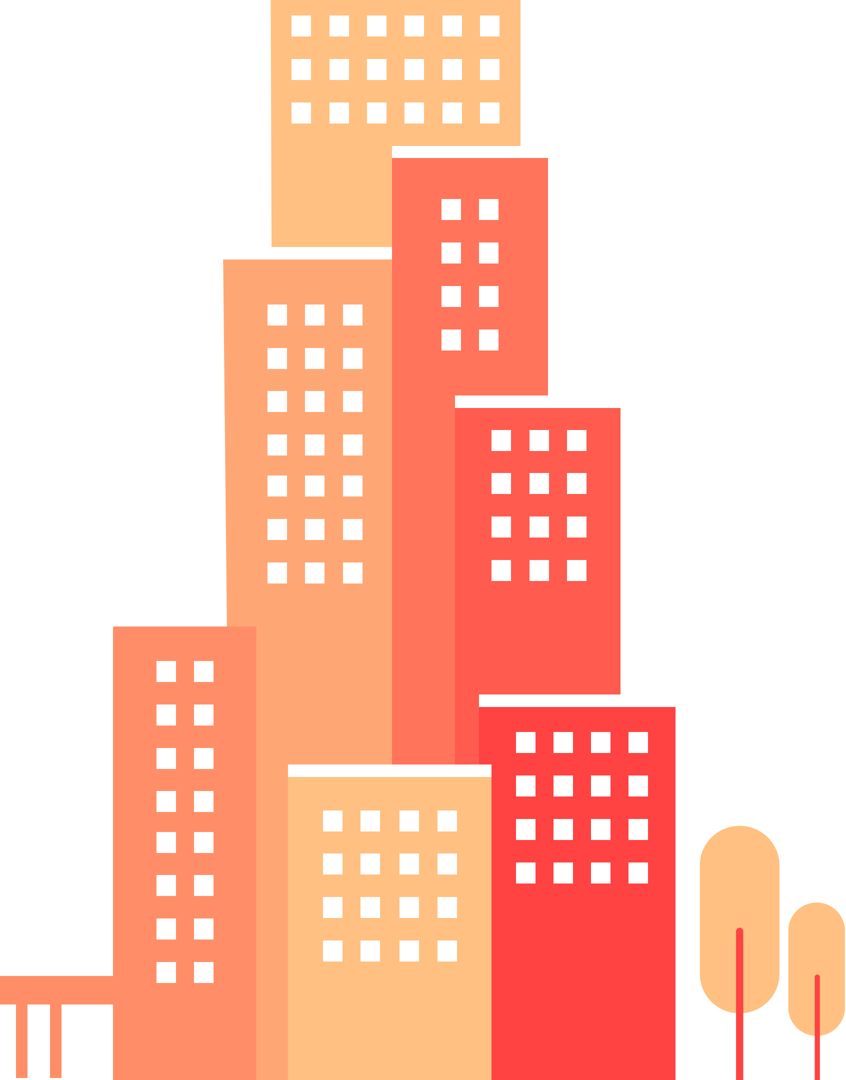
Conclusion

- Features used are correlated to the saleprice
- $250,000 and under: Model did well
- Above $250,000: higher variance
- Limitations: Less data above $250,000
- Top features are usually fixed and cannot be changed
- Recommended features for upgrading:
  - Wood Shingle roof
  - Cement Board or Brick Face exterior
  - Wood Deck

Residual Error on Training Set







Questions?