GA DSI Project 2
Ames Housing SalePrice Prediction

Team 3:

Maybelle,
Anand
Tiek Leong

Contents

1. Problem Statement
2. Data used
3. Feature Selection ⇒ Linear Regression (20 Features)
4. MultiCollinearity ⇒ Ridge Regression (16 Features)
5. Group Feature Selection, Polynomial feature engineering
6. ElasticNet (8 Features)
7. Demo
8. Limitations
9. Recommendations

Problem Statement

We are a start-up aiming to set up a Sale Price recommendation tool where users can use to get a recommended Sale Price with just a few details about their property for sale.

This webpage would have specific fields for users to fill property details and submit.

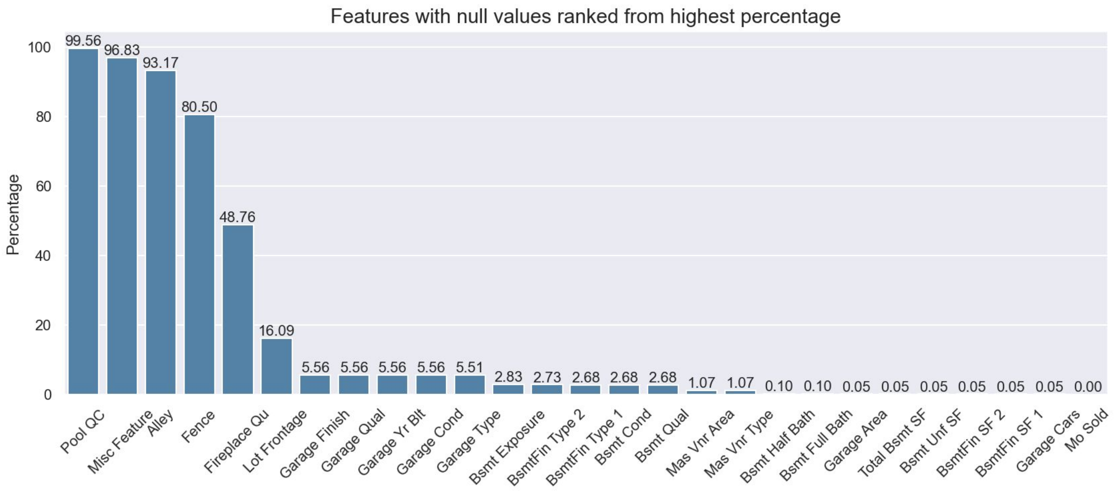
Using our proprietary machine learning algorithm, the web application generates a quote estimate for their property Sale Price.

Data used

- Ames Iowa Housing Data 2006 - 2010
- 81 Features, 2051 data points
- Target: Sale Price

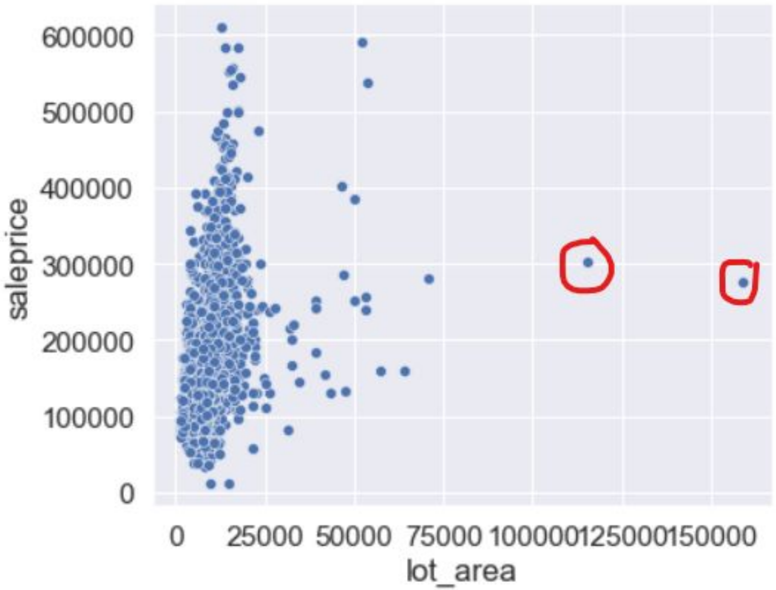
Missing Values

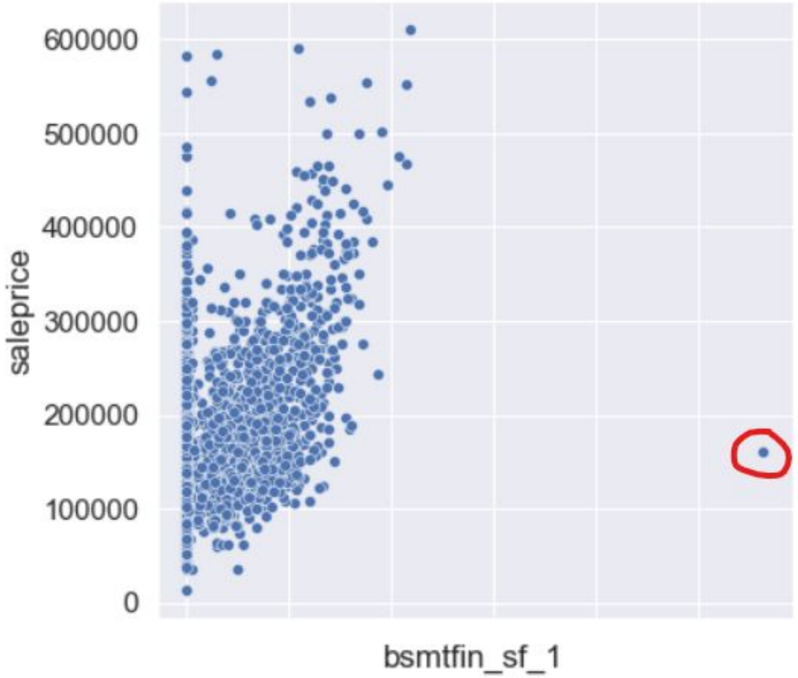
- Noted that features with high percentage of null values, were meaningful null values
- Garage Yr Blt feature dropped
- Lot Frontage missing values will be imputed after train-test-split



Outliers

Examples of Outlier Removal on certain features





We can see there are outliers here that we can remove

Label Encoding

- We used **ordinal encoding** to reduce the number of features required for One-Hot encoding. E.g Lot Shape Reg=4, IR1=3, IR2=2, IR3=1
- Superior attribute gets a higher score, to be consistent with Overall Qual and Overall Cond source encoding
- **New features** to represent presence and absence of some features e.g. pool, fireplace, garage, alley

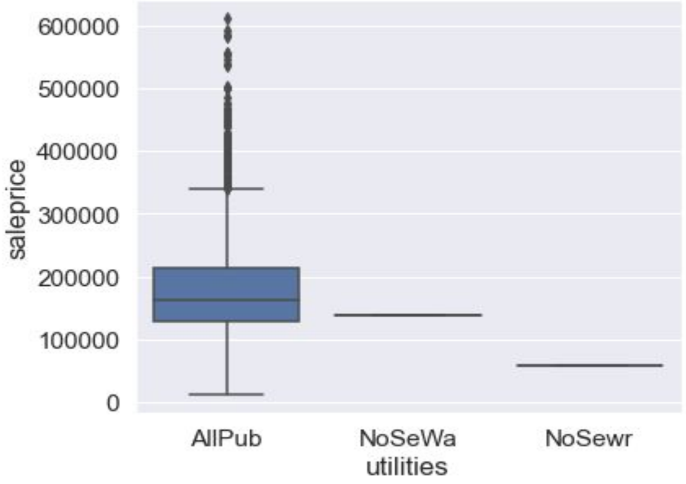
Value Count Grouping and Removal

Value Count Grouping

Misc Feature	
None	1986
Shed	56
Gar2	4
Othr	3
Elev	1
TenC	1

Misc Feature	
0	1986
1	65

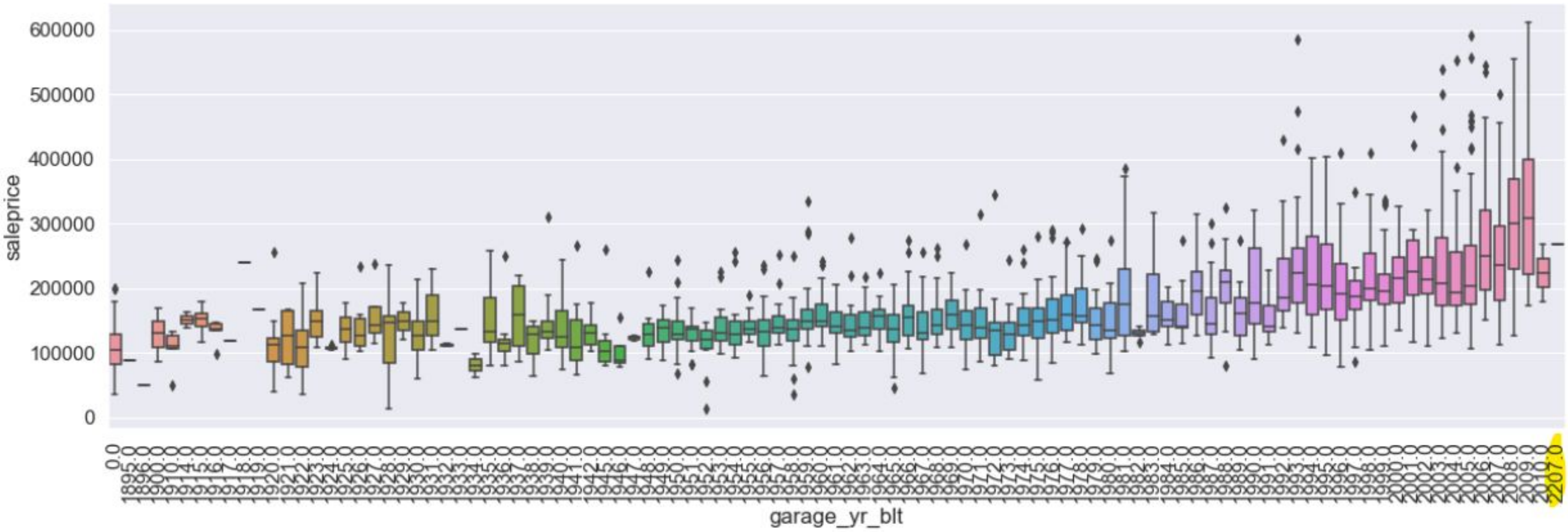
Feature Removal due to Value Count E.g. Utilities



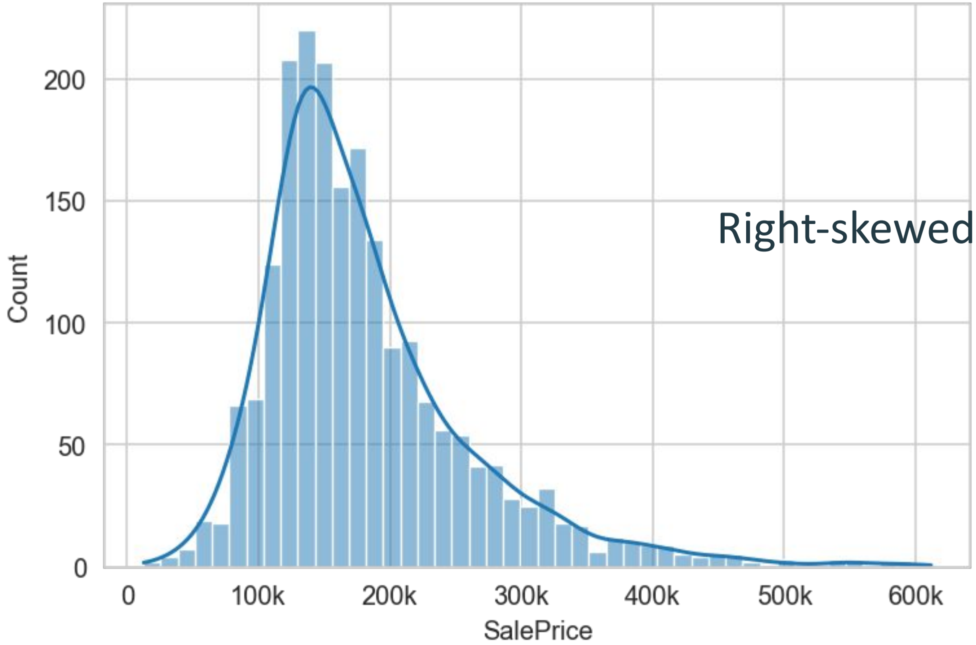

```
AllPub    2046
NoSeWa       1
NoSewr       1
Name: utilities,
```

Removal of Data entry errors

Example: Garage Year Built Year 2207 removed

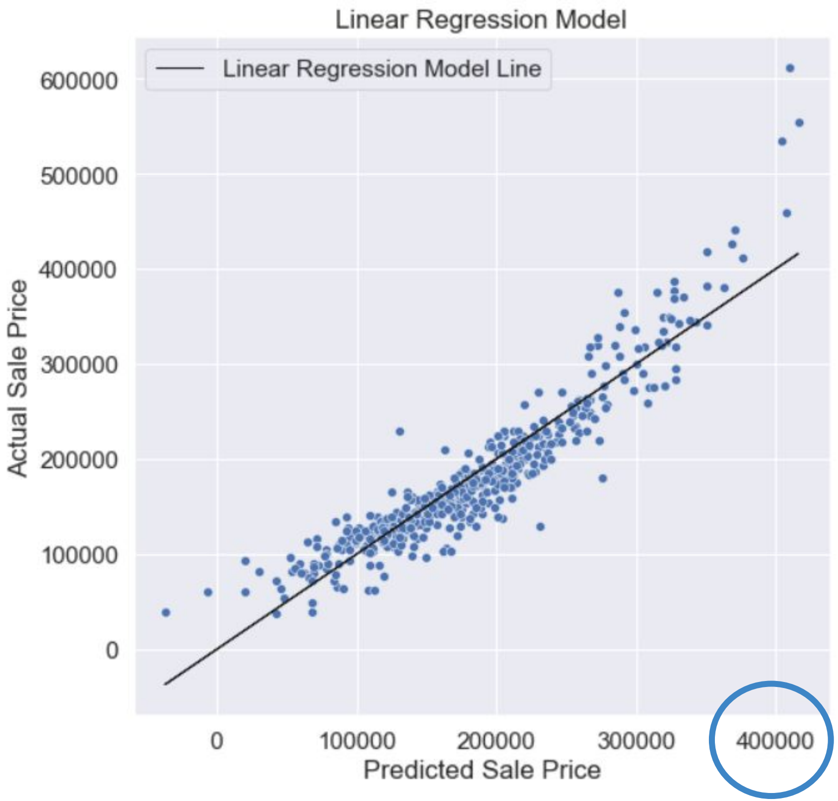


Target: SalePrice



Right-skewed distribution

Modelling - Linear Regression (20 Features)



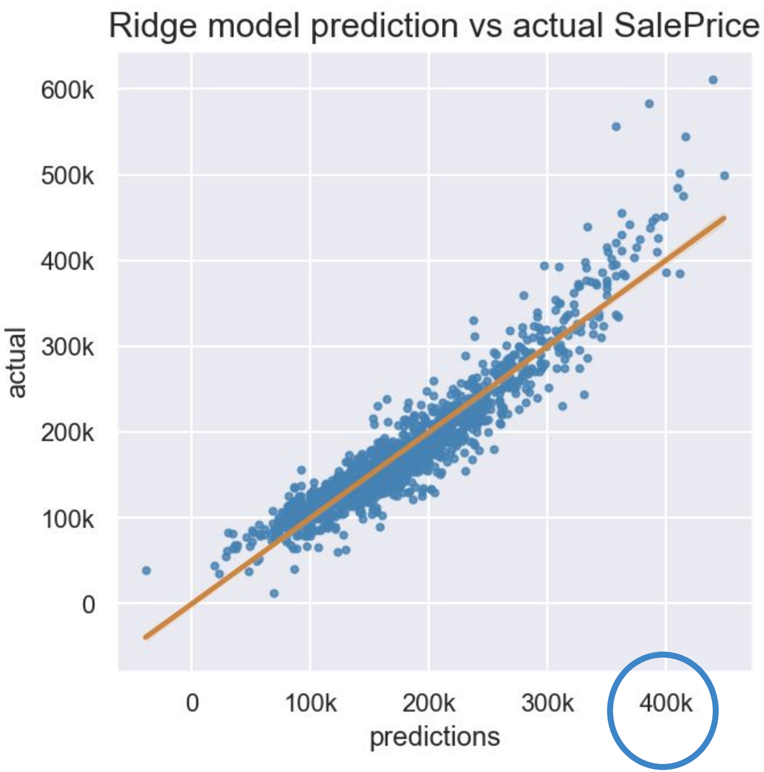
Features that had **correlation ≥ 0.3 (p-value < 0.05)** with Sales Price were retained

RMSE = root mean squared error

Baseline RMSE = 75639
Train RMSE = (25733)
Test RMSE = (25591)
% diff RMSE = 0.55%

Good generalisation. Performance better than baseline model.

Modelling - GridSearch CV Ridge (16 Features)



To reduce multicollinearity, 4 features were dropped

Baseline RMSE = 75639
Train RMSE = (26766)
Test RMSE = (26666)
% diff RMSE = 0.37%

Good generalisation. Performance compromised slightly

Group Feature Selection

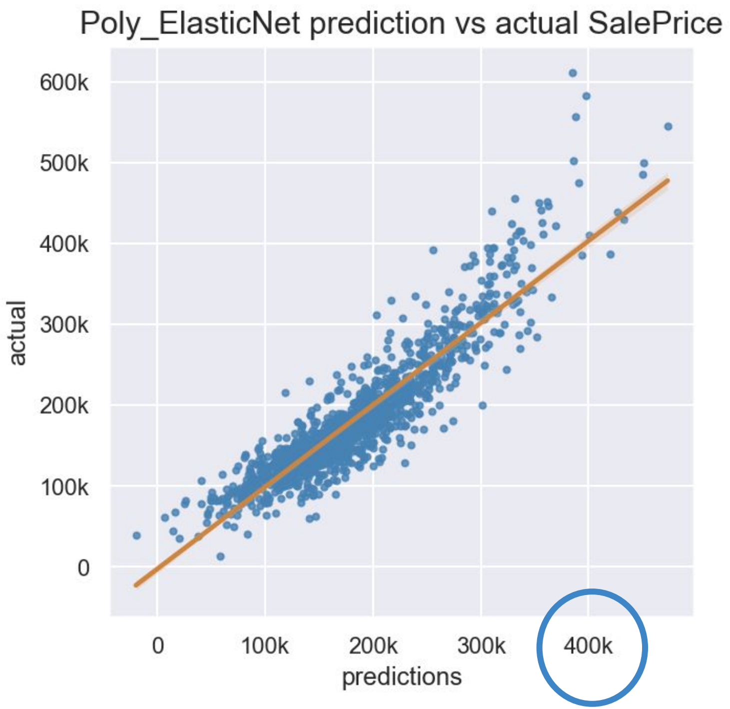
- Found common features in all 3 of group members' models, we further reduced the features used for model deployment
- **6 features** found:
  - **Overall Qual:** Rates the overall material and finish of the house
  - **Gr Liv Area:** Above grade (ground) living area square feet
  - **Total Bsmt SF:** Total square feet of basement area
  - **Garage Area:** Size of garage in square feet
  - **Year Built:** Original construction date
  - **Mas Vnr Area:** Masonry veneer area in square feet

Polynomial Feature - Ground Living Area

	Gr Liv Area	Gr Liv Area^2	Gr Liv Area^3	Gr Liv Area	Overall Qual	Mas Vnr Area	Year Built	Total Bsmt SF	Garage Area
0	1441.0	2076481.0	2.992209e+09	1441.0	8.0	456.0	1999.0	776.0	492.0
1	1604.0	2572816.0	4.126797e+09	1604.0	5.0	0.0	1958.0	1604.0	576.0
2	1150.0	1322500.0	1.520875e+09	1150.0	5.0	176.0	1955.0	1078.0	288.0

- Ground Living Area is the 2nd most correlated to SalePrice
- Overall Qual will be imputed for new data because it is a subjective measure for our use case
- Polynomial feature engineering more features resulted in overfitting

Modelling - GridSearch CV ElasticNet (8 Features)



Mean Absolute Error (MAE)

Baseline RMSE = 75639
Train MAE = 21742
Test MAE = 22177
Test RMSE = (29884)
% diff RMSE = 1.84%

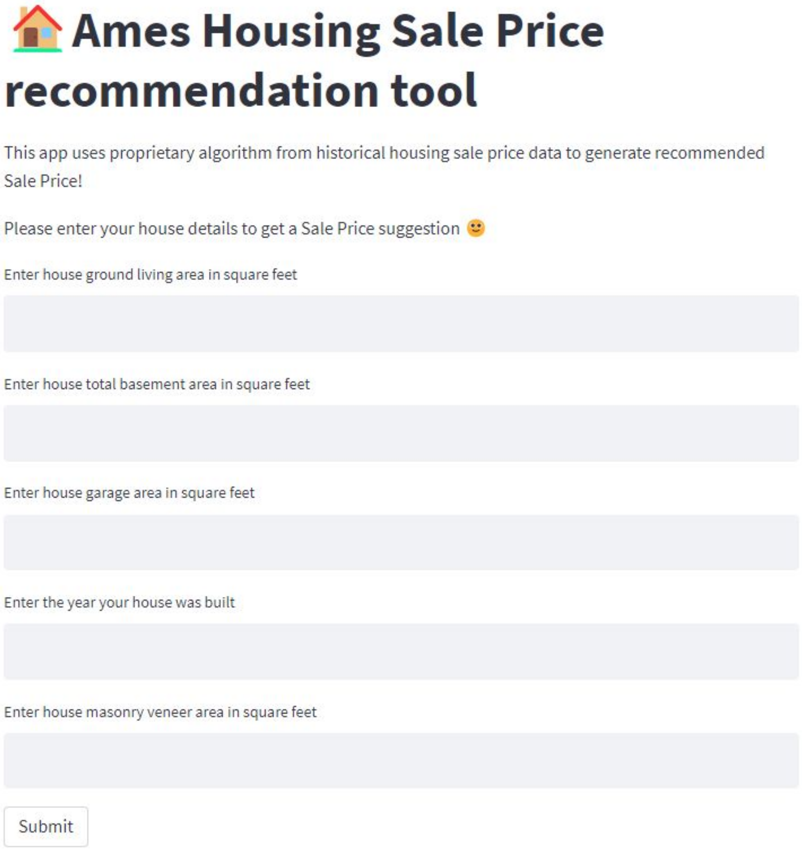
Generalisation is maintained.
Compromised on performance.

Demo

[Streamlit Cloud link]

Actual value: **164500**

Recommended: **170965**



Limitations

- With SalePrice right skewed distribution - model has limited performance for recommendations above $400K.
- Features with good predictive value for SalePrice tend to have high multicollinearity with each other.
- This model does not take into account any economic or external factors that may affect Sale Price: for example, housing loan interest rates, unemployment figures, inflation, natural disasters etc.

Considerations

- Years provided for data is from 2006 to 2010, which is over a decade old and may not reflect the current market figures and could result in stale data
- All this data was taken during the 2008 financial crisis which could have implications on data quality
- There could be other features not initially included in the raw training data that were not considered as we were limited to the data at hand e.g. Sale Condition

Recommendations

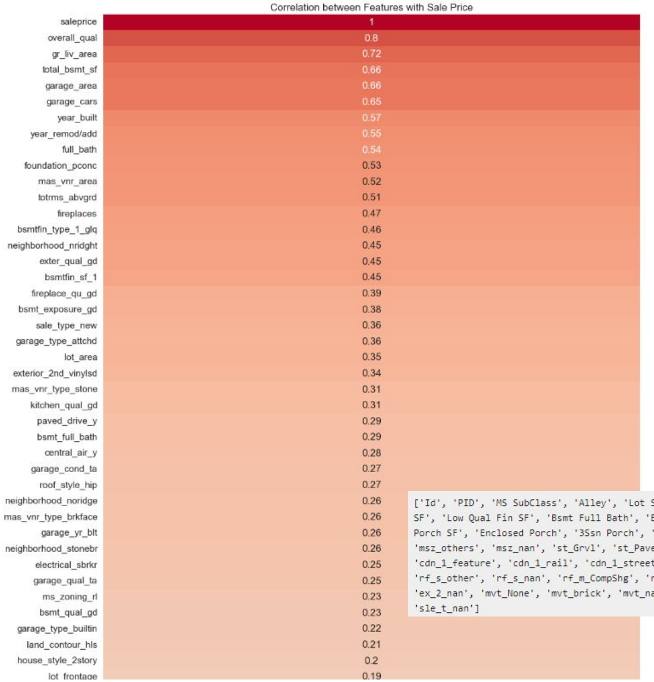
- Although GridSearchCV ElasticNet linear model was chosen for deployment, we recommend trying different models to achieve better prediction performance
- Collect more recent data (newer than 2010)
- Collect more data points with house saleprices above $400k
- Get access to Sale Condition feature
- Collect data on other features not in current data that may have influence on house saleprice

Double Bam!!!

Any Questions?

Correlation Heat Map between Features and Sale Price Pre - Feature Selection

Skipped



- After Data Cleaning and processing, we have 138 features (Correlation heatmap on the left)
- We conducted feature selection to reduce the number of features and to reduce multicollinearity

```
['Id', 'PID', 'MS SubClass', 'Alley', 'Lot Shape', 'Utilities', 'Land Slope', 'Overall Cond', 'Exter Cond', 'Bsmt Qual', 'Bsmt Cond', 'BsmtFin Type 2', 'BsmtFin SF 2', 'Bsmt Unf SF', 'Heating QC', 'Central Air', 'Electrical', '2nd Flr
SF', 'Low Qual Fin SF', 'Bsmt Full Bath', 'Bsmt Half Bath', 'Half Bath', 'Bedroom AbvGr', 'Kitchen AbvGr', 'TotRms AbvGrd', 'Functional', 'Garage Finish', 'Garage Cars', 'Garage Qual', 'Garage Cond', 'Paved Drive', 'Wood Deck SF', 'Open
Porch SF', 'Enclosed Porch', '3Ssn Porch', 'Screen Porch', 'Pool Area', 'Pool QC', 'Fence', 'Misc Feature', 'Misc Val', 'Mo Sold', 'Yr Sold', 'SalePrice', 'have_bsmt', 'have_garage', 'have_pool', 'have_fence', 'msz_RL', 'msz_R_others',
'msz_others', 'msz_nan', 'st_Grvl', 'st_Pave', 'st_nan', 'ld_c_Lvl', 'ld_c_unLvl', 'ld_c_nan', 'lt_c_Corner', 'lt_c_Inside', 'lt_c_others', 'lt_c_nan', 'nbrhd_pc_50010', 'nbrhd_pc_50011', 'nbrhd_pc_50014', 'nbrhd_nan', 'cdn_1_Norm',
'cdn_1_feature', 'cdn_1_rail', 'cdn_1_street', 'cdn_1_nan', 'cdn_2_Norm', 'cdn_2_other', 'cdn_2_nan', 'bdg_t_Duplex', 'bdg_t_fam', 'bdg_t_twnhs', 'bdg_t_nan', 'hse_s_1story', 'hse_s_2story', 'hse_s_split', 'hse_s_nan', 'rf_s_Gable',
'rf_s_other', 'rf_s_nan', 'rf_m_CompShg', 'rf_m_other', 'rf_m_nan', 'ex_1_MetalSd', 'ex_1_VinylSd', 'ex_1_cement', 'ex_1_other', 'ex_1_wood', 'ex_1_nan', 'ex_2_MetalSd', 'ex_2_VinylSd', 'ex_2_cement', 'ex_2_other', 'ex_2_wood',
'ex_2_nan', 'mvt_None', 'mvt_brick', 'mvt_nan', 'fdn_CBlock', 'fdn_PConc', 'fdn_other', 'fdn_nan', 'htg_GasA', 'htg_other', 'htg_nan', 'grge_t_Attchd', 'grge_t_Detchd', 'grge_t_other', 'grge_t_nan', 'sle_t_Con', 'sle_t_Oth', 'sle_t_WD',
'sle_t_nan']
```

Correlation between features

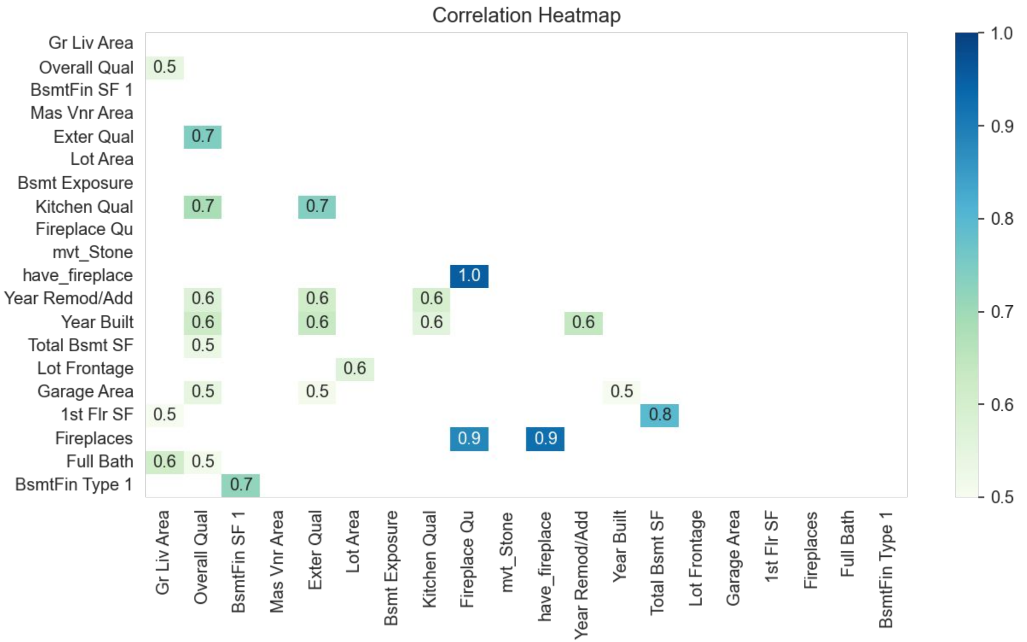
Skipped

feature	vif
Gr Liv Area	20486.869948
Overall Qual	17.869798
BsmtFin SF 1	12.615395
Mas Vnr Area	7.695555
Exter Qual	3.592153
Lot Area	3.457068
Bsmt Exposure	3.397382
Kitchen Qual	3.233269
Fireplace Qu	2.647413
mvt_Stone	2.642132
have_fireplace	2.585894
Year Remod/Add	2.581521
Year Built	2.362629
Total Bsmt SF	2.171844
Lot Frontage	2.124630
Garage Area	1.888327
1st Flr SF	1.682331
Fireplaces	1.648372
Full Bath	1.374341
BsmtFin Type 1	1.373617
const	1.231542

- We used correlation heatmaps and Variation inflation factor scores to assess for multicollinearity between features.
- We found VIF score being not very useful for feature selection for this dataset and target.
- When all features with high VIF scores were removed, it results in a significantly poorer performing model
- So we had to be selective with which feature to drop

Skipped

Correlation Heatmap



- have_fireplace, Fireplace Qu and Fireplaces have high correlation
- BsmtFin SF 1 has high VIF score and low correlation with **SalePrice**
- Total Bsmt SF and 1st Flr SF have high correlation with **SalePrice,** so they were not dropped