Background :

In United States, residential properties vary in types. Due to their heterogenous nature, housing prices vary based on different features they have, such as square footage, number of bedrooms, number of bathrooms, age of the property etc.

Problem Statement

Develop a model using different modeling techniques to help a group of real estate investors to predict selling prices for homes by defining relationship between the sale price and various features of the property.

Requirements and Availability

To estimate the sale price for homes, we have Ames dataset which has detailed information on various features of each property spread over 84 columns in the dataset. Some columns are:

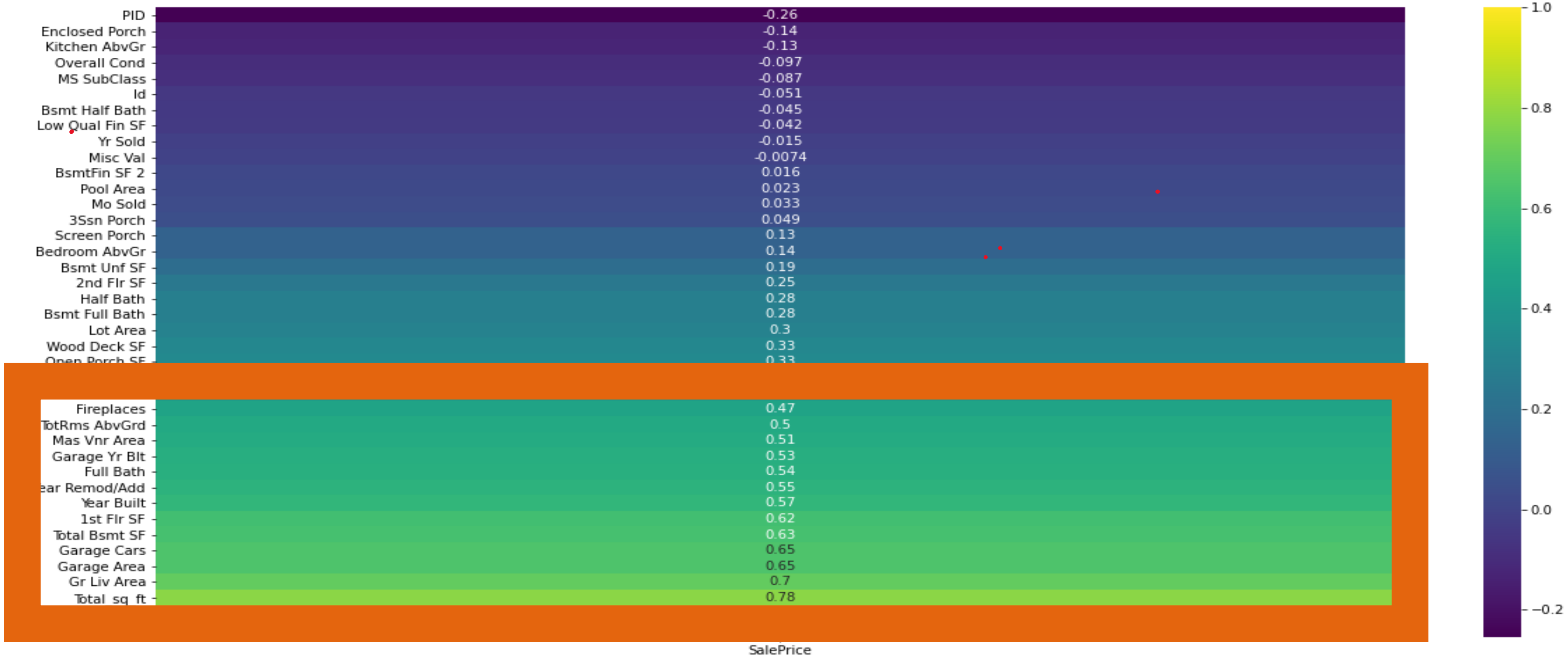
- Lot Area
- Overall Quality
- Basement details
- Bedroom details
- Bathroom details
- Any other additional features.

Process flow

- Import and clean the data
- Identify the most ideal and relevant features from the dataset
- Iterate with different models

Process flow continued…

- To select appropriate features , used below correlation matrix



Process flow continued…

As highlighted in the heatmap, below features had closest positive correlation with the Sale price and so were selected for modeling

- Lot Area
- 2nd Flr SF
- Total Bsmt SF
- Year Built
- Full Bath
- Bedroom AbvG
- Gr Liv Area
- Garage Area
- Garage Cars
- Wood Deck SF
- Total_sq_ft
- Year Remod/Add
- Overall Qual
- Bsmt Full Bath
- 1st Flr SF

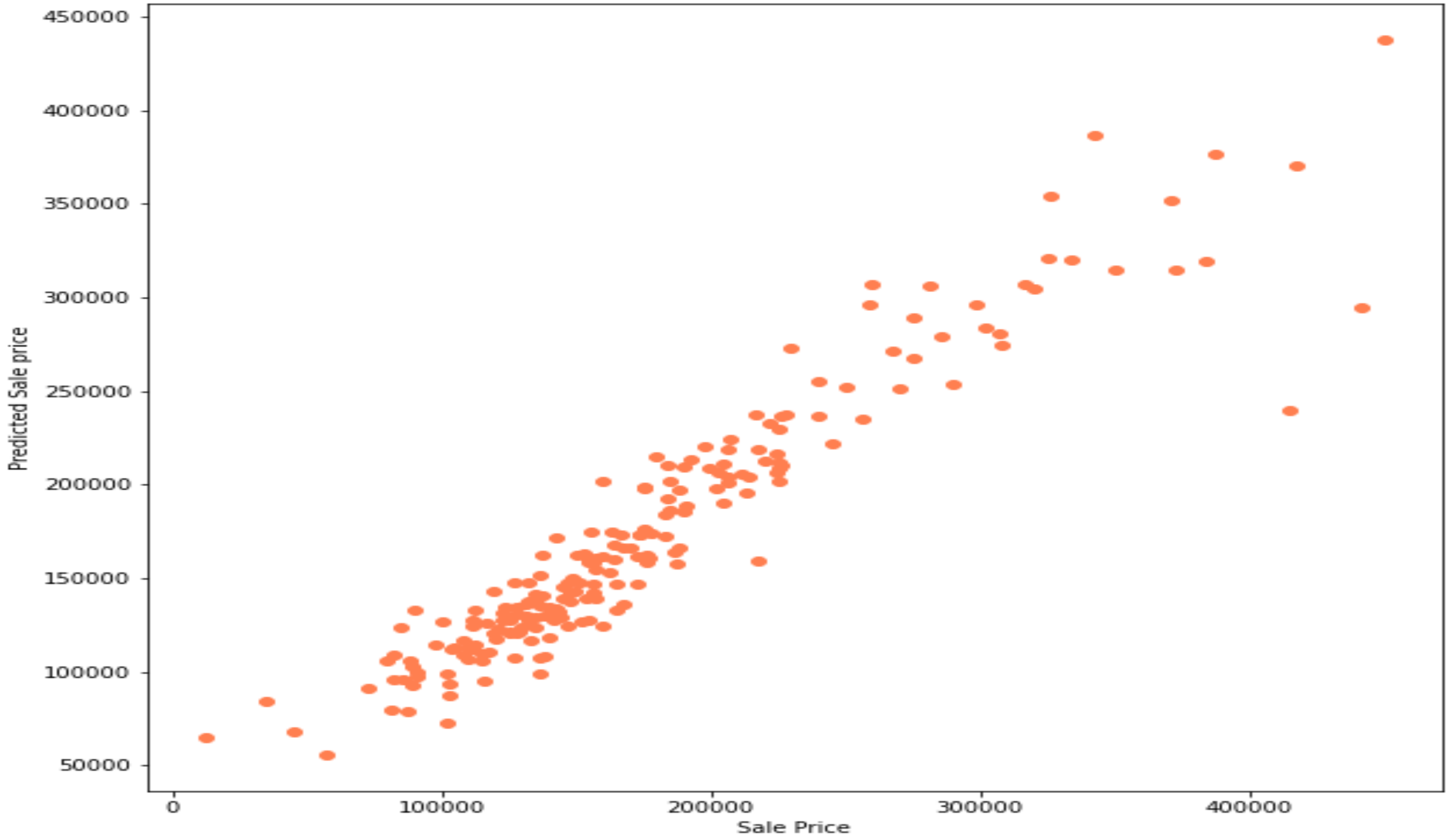
Process flow continued…

Model used : Multiple Linear Regression and generated
cross_val_score : 0.79

Feature Engineering : New Feature – Total Sq FT , Polynomial Features
Train / test scores After feature Engineering : 0.910, 0.89

Process flow continued…

Graph shows the scatter plot of Original Sale price Vs. Predicted Sale price. Values are very close to each other.



Process flow continued…

Coefficients generated from model for each feature used :

```
('Lot Area', -48.604815171176554),
 ('Overall Qual', 44705.1960148833),
 ('1st Flr SF', 3802.4872289198643),
 ('2nd Flr SF', 3741.810194591539),
 ('Total Bsmt SF', -3834.411309255544),
 ('Year Built', 20498.69676398044),
('Full Bath', -614290.7797889019),
 ('Bedroom AbvGr', 345736.7865048784),
 ('Gr Liv Area', -8302.308014638376),
 ('Garage Area', -657.6798609152339),
 ('Garage Cars', 179922.56397965524),
 ('Wood Deck SF', 237.54086716745678),
 ('Total_sq_ft', 3709.917234405921),
 ('Year Remod/Add', -52540.53383965394),
 ('Bsmt Full Bath', -285253.123285776)
```

Positive values for coefficients show positive impact on Sale Price , while negative coefficients show negative impact on the sale price.

Conclusion

Model successfully predicts the sale price using features selected.
Modifications or increase in below features will have direct positive impact on sale price.

- Overall Qual
- 1st Flr SF
- 2nd Flr SF
- Year Built
- Bedroom AbvGr
- Garage Cars
- Wood Deck SF
- Total_sq_ft