Project 2 - Ames Housing Data and Kaggle Challenge

by: Matthew Edelmann

Problem Statement

We want to create a good model for determining the home price given a set of training and testing data.

Datasets

We use 2 data sets. One home training data set. The other is the home testing data set. We will train our data on the home train dataset. We will test it on the other data set.

* [`train.csv`](./data/train.csv): Home training data set

* [`test.csv`](./data/test.csv): Home testing data set

Cleaning the data

Null Values

First and formost we checked the list of of null values. We want to know how many null values are in each columns. We decided that columns with null values over 50. For columns that have null values between 1 and 50 we simply drop the null rows.

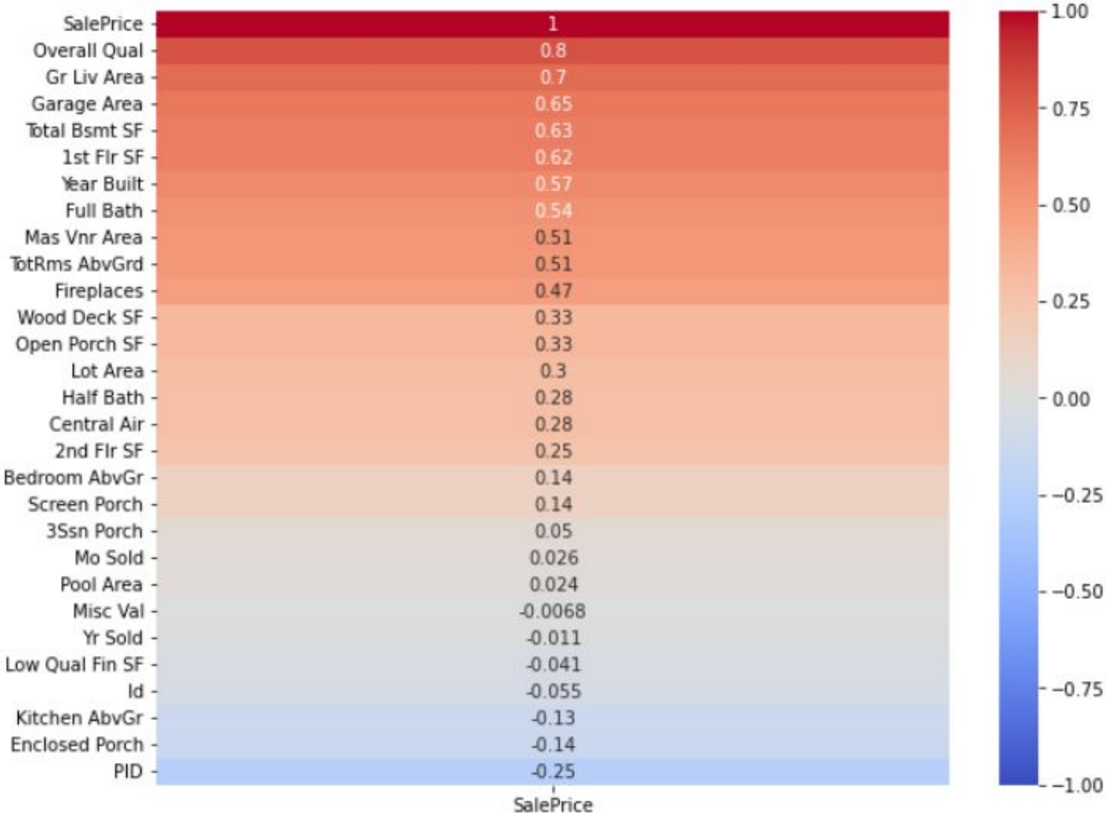
Removing duplicates

Looking at this list we see that there are many columns that are far too similar to each other. So we deleted near duplicate columns from each list.

Central Air

The last bit of cleaning that I did was to convert the 'Yes' and 'No' rows in the 'Central Air' column into 1's and 0's respectively. With that the data was cleaned.

Heat Map


SalePrice 1
Overall Qual 0.8
Gr Liv Area 0.7
Garage Area 0.65
Total Bsmt SF 0.63
1st Flr SF 0.62
Year Built 0.57
Full Bath 0.54
Mas Vnr Area 0.51
TotRms AbvGrd 0.51
Fireplaces 0.47
Wood Deck SF 0.33
Open Porch SF 0.33
Lot Area 0.3
Half Bath 0.28
Central Air 0.28
2nd Flr SF 0.25
Bedroom AbvGr 0.14
Screen Porch 0.14
3Ssn Porch 0.05
Mo Sold 0.026
Pool Area 0.024
Misc Val -0.0068
Yr Sold -0.011
Low Qual Fin SF -0.041
Id -0.055
Kitchen AbvGr -0.13
Enclosed Porch -0.14
PID -0.25
SalePrice
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Making dummies

Next I looked at the correlation between salesprice and the non object columns and only kept those which the absolute values of the correlation was at least 0.6. Then I made dummies of the object columns. Again, looking at this correlation, I only kept the columns which the absolute values of the correlation was at least 0.6.

The Columns

```
['Overall Qual',

 'Gr Liv Area',

 'Garage Area',

 'Total Bsmt SF',

 '1st Flr SF',

 'Exter Qual_TA']
```

Polynomials

We created a 2 degree polynomial of these columns. This was used for the model.

Conclusion

The model gave us These R2 scores:

Training R2: 0.8589933055335499

Testing R2: 0.879607496219049

And this cross validation score:

0.8219401785265875

That means the model was only slightly underfit and it explains about 82% of the data.

Further Research

Given more time I would like to make a model using OLS and Lasso.

QUESTIONS?