Ames Iowa Housing Price Prediction

BY NATI MARCUS

Problem Statement

- Given the Ames housing dataset, how well can I predict house prices for Ames on unseen data? To answer this question I will be conducting Linear Regression Analysis to calculate the root mean squared error for housing prices. Doing so will allow me to evaluate around how accurate in dollars the prediction is. I will be comparing this figure to the "baseline model" RMSE, i.e the average error if you were to predict the median housing price of the training data.

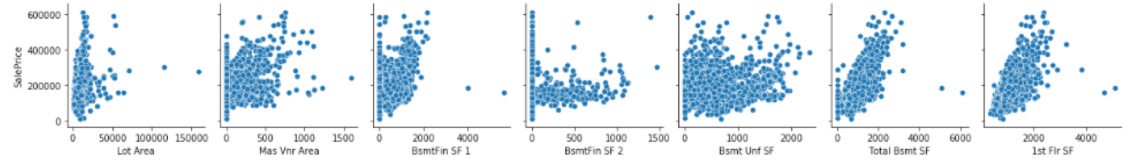
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Source: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e

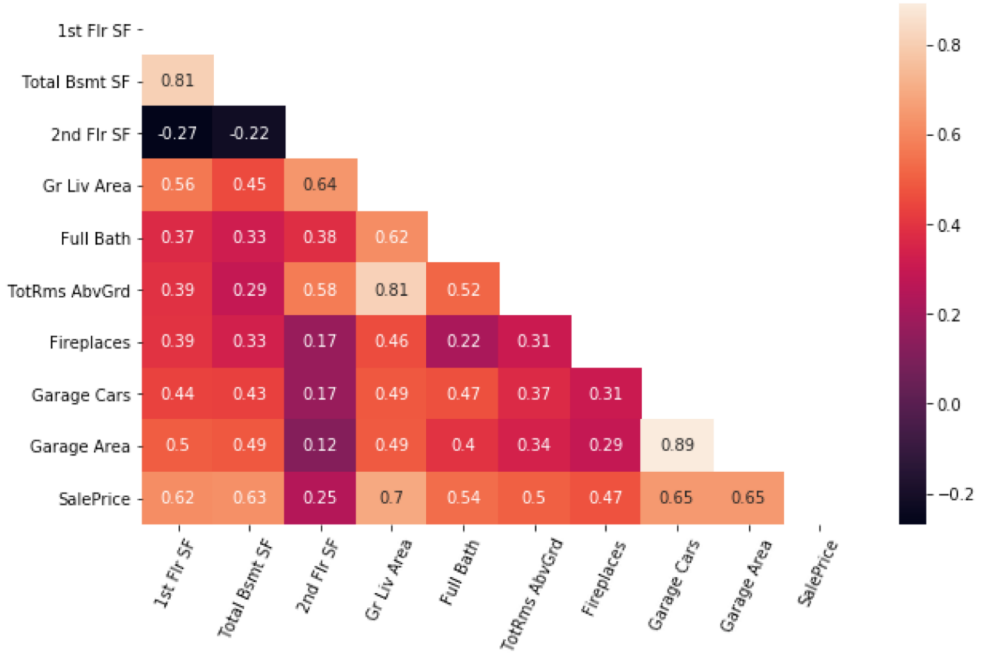
Data Overview

- Training Set
  - 2051 rows, 81 columns
  - ~43 quantitative features, ~37 qualitative features, 1 outcome variable
- Test Set
  - 878 rows, 80 columns
  - ~43 quantitative features, ~37 qualitative features, no outcome variable

```
sns.pairplot(data=df,x_vars = ['Lot Area','Mas Vnr Area','BsmtFin SF 1','BsmtFin SF 2','Bsmt
'Total Bsmt SF','1st Flr SF'],
             y_vars = ['SalePrice']);
```



Pairplot of some of the numerical features to check linear relationship with `SalePrice` value



Heatmap to check correlation of numerical features and `SalePrice` value

Methodology

- Begin by exploring data
  - Cleaning, feature selection, pair plot, heatmap
- Wrote main regression function that printed scores based on whether or not the data was scaled or had any form of regularization
- Conducted multiple regression analyses, increasing in complexity
  - First Model only Numerical Values
  - 2nd -4th Models Interaction Terms and Poly Features Added
  - 5th Model included all above plus dummy variables (best model)
  - 6th and Final dropped columns of 5th but yielded worse result
- Evaluate based on RMSE, baseline model

Source: https://commons.wikimedia.org/wiki/File:Scikit_learn_logo_small.svg

Model 5 Overview:

- Numerical columns included 1st floor sqft, total bsmt sqft, 2nd flr sqft, above ground living area, # of full baths, total rooms above ground, # of cars for garage, and garage area
  - Features were scaled
  - Also used polynomial features
- Categorical (dummy) columns were MS Zoning, Utilities, Neighborhood, Overall Qual, Overall Cond, Exter Qual, Exter Cond, Bsmt Cond, Heating, Heating QC, Central Air

Source: https://www.expert.ai/blog/machine-learning-definition/

Result of Best Model (Model 5)

- On test data, the RMSE was $27,977.02
  - Improved prediction from baseline model (RMSE of ($68,946.92)) by 2.5X
  - On average, model is off by around $27K when attempting to predict house prices on unseen data

17	Nati Marcus	27977.02...	2	1h

kaggle

Source: https://en.wikipedia.org/wiki/Kaggle

Conclusions and Recommendations

- The model I constructed yielded a significantly better RMSE result compared to the baseline model RMSE on both the training and testing data.
- However, the model is still not as accurate as it could be, so I would suggest using a more systematic approach to selecting features in order to optimize this model.
- In conclusion, the model is a good start, but more work needs to be done to obtain a more accurate house price prediction.