HOUSE PRICES: ADVANCED REGRESSION TECHNIQUES

Author: Parth Shah, Prasad Sawant

ABSTRACT

Homebuyers all across the United States are confused about what price to pay for their dream home. Various factors are responsible for adjusting the sale price on a house. Houses vary in their location, area of living, quality of living, etc. In this project, we try to use the existing parameters to predict the selling price of a house.

1 THE PROBLEM

We are given a total of 79 parameters about a house, and using these parameters we have to predict the final price of the house. The data to be generated is numeric and hence various regression techniques would be useful in its accurate prediction. We have used random forest decision trees and logistic regression as our machine learning models.

2 DATA PREPROCESSING:

We visualized various columns in the data and realized a lot of them had missing data. These were affecting the prediction as the model was taking in '0' as the value for missing data which ultimately resulted in poor accuracy. We could fill in these missing values by enforcing a pattern in the data, filling it with mean, or any other statistical measure. But for data with more than 10% missing data, we would be introducing a lot of redundancy and/or false information. To overcome this problem, we decided to drop the columns with more than 10% of missing data. Luckily, most of the missing data was in the parameters which we believe did not have a substantial effect in determination of the sales price.

3 DATA EXPLORATION: ALL PARAMETERS

In order to gain more idea about the underlying data we decided to understand more of how various parameters affected each other and the target data. We plotted a heatmap and realized there were quite a few parameters which had a high degree of correlation with each other. The highly correlated variables can be used to replace/adjust the missing values as well.

For example, Garage cars (The number of cars in the garage) and garage year built had a high degree of correlation. Total number of rooms had a direct correlation with the total living area of the house. These relations, although obvious when read aloud, gave us an insight on how different parameters affect each other and more importantly, how they affected the sales price. In the heatmap the last column is the correlation of various parameters with the selling price of the house.

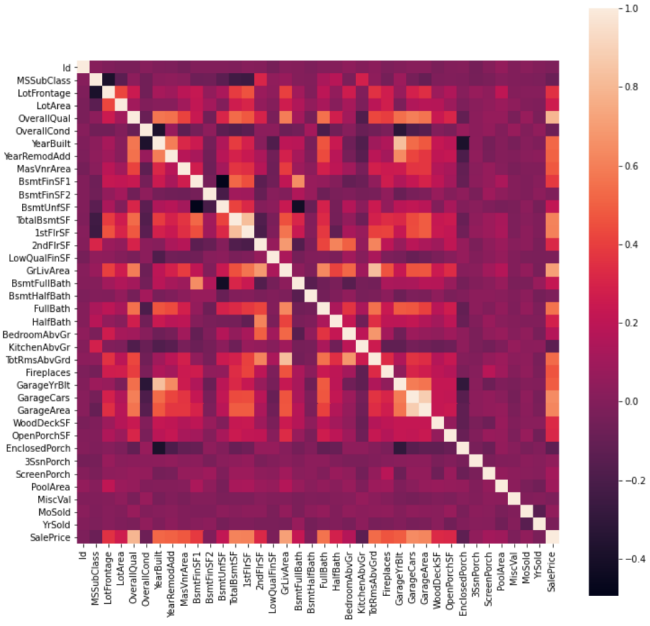

Figure 1: Heat-Map

4 Data Exploration: Specific parameters

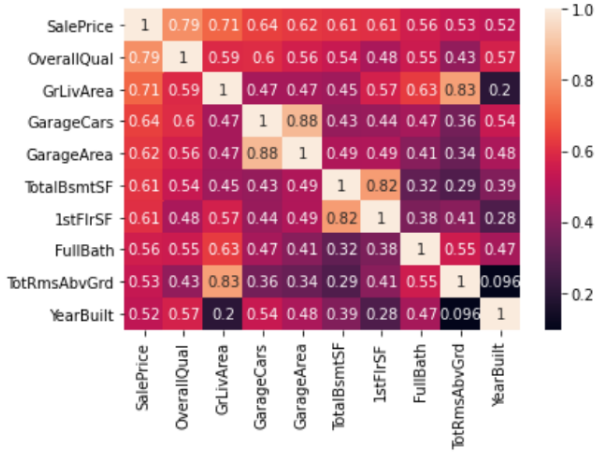

Figure 2: Correlation of various parameters

Parameters like overall quality, first floor square feet, living area, garage area etc. had the most say in determining the sale price of the house. We decided to explore these features further.

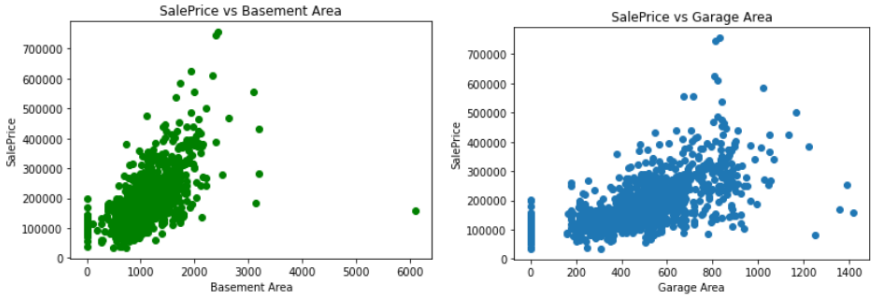

Figure 3: Sales price vs various parameters

When we see these graphs it's evident that there is a cluster of points which tend to form a pattern.. However, there are some points in the graph which have drastically different values than what represents a typical behavior. For example, in the 'Sales price vs Basement area' plot, there is a point for a very large basement area where we get a nominal sales price much lower than what it ideally should be. These points are outliers, which severely affected our model as these values are some exceptions and should not be included in our model prediction. We deleted these outliers and our model already made a huge progress in its accuracy.

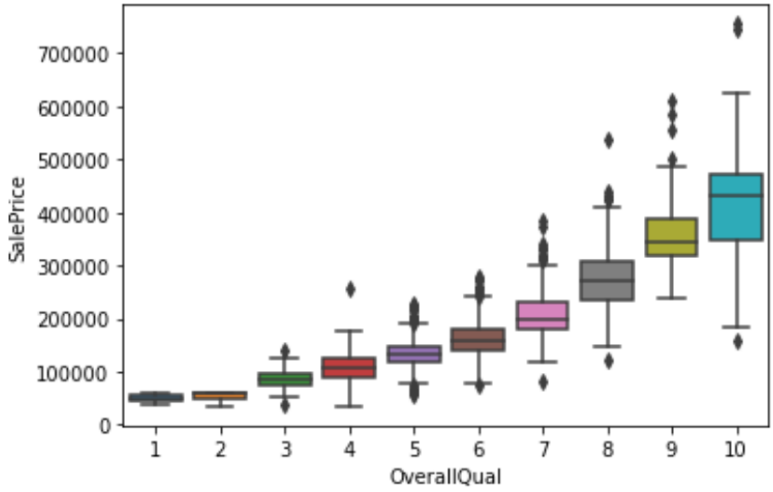

Figure 4: Sales price vs Overall Quality

In the plot above, the points above and below the rectangular boxes are outliers. However, there were a few parameters, like in the case of overall quality which had such a huge impact on determining the sales price that we decided to keep these outliers as deleting them would cost our model to degrade.

5 VALIDATION AND THE PROBLEM OF UNDER AND OVER-FITTING:

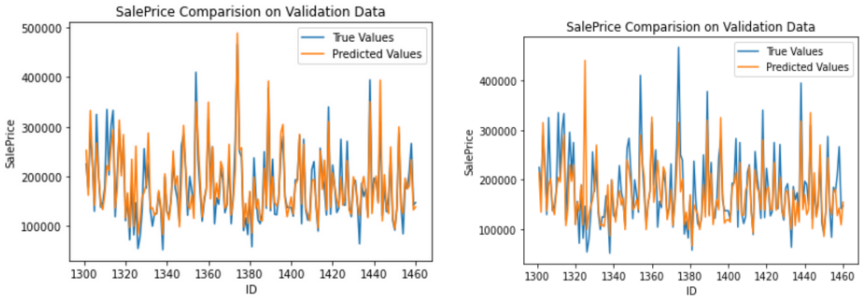

Figure 5: Left: Random forest, Right: Logistic Regression

We ran a random forest model on our data with '1000 estimators' as a tuned parameter. This gave a very low training error but huge validation error This was a clear case of overfitting and through trial and error we found that estimator = 100, was the best tuned parameter for our model when we ran it on the validation dataset. Comparing the accuracy of random forest with logistic regression, the accuracy of the former was much higher than logistic, leading us to conclude that random forest was a better model for this problem.

6 CONCLUSION

We went through various stages required in prediction of house process. From visualizing data, preprocessing data, choosing the right model, tuning the parameters and finally running our model on the test data. We conclude that our tuned random forest model was the best model for this problem.