2021

# House Price Prediction

SATISH AGRAWAL DSC 680

BELLEVUE UNIVERSITY

## Executive Summary

Real state is one of the biggest industries and have millions of transactions every year. Real state includes a commercial complex, a multi-family building, a residential house, a plot etc. [James Gheen]. This project aims to look at only the price of a house. House price fluctuates a lot and there are variety of factors that impact the final price of a house [National Association of Realtors]. The size of the house/lot, amenities inside the house, upgrades, age and quality of materials used are few of the factors impact the price of a house. Other than what's included in the house, there are multiple external factors that impact the price of the house heavily, like market condition, job market, season and the time of the year to list a few.

There are various listing agencies helping list houses in the market and let user choose a property from a website. These companies have estimated price (for example zestimate from Zillow [3 Zillow]) and seller also chooses a listing price which may or may not be the same as estimated price. There is need of a mechanism to come up with a great estimated price given all (or most) of the important factors. This solution will provide a consistent way for seller, buyer and listing companies to see a fair price for the house and make their own decision.

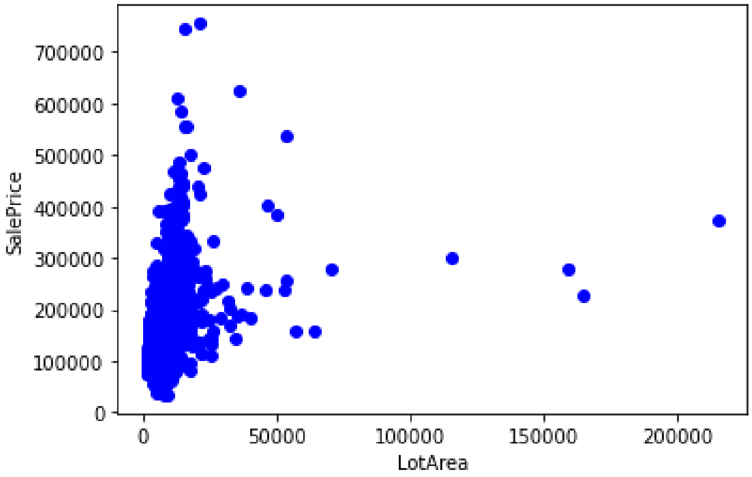
In this project I will build a machine learning model that takes important factors for consideration and trains a model which can accurately predict the price of a house.

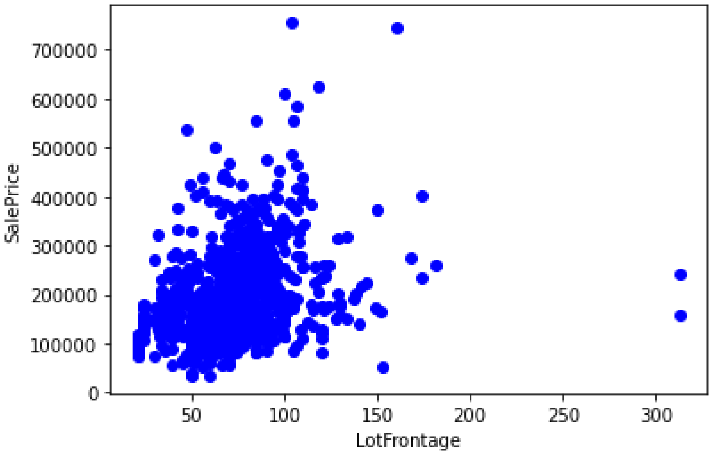
## Data Set and Missing values

The data set had quite a few columns with missing values. Machine learning algorithms cannot process attributes with missing values. Those missing values need to be handled before any modeling can start. There are variety of ways to impute missing values, for example, impute with most frequent value, impute with mean of all available values for the attributes, impute with min or max of all the values. These techniques work for specific cases and we need to evaluate which technique should be applied. There are other ways like predicting the values for missing values using the values available to train a imputation model. I normally avoid using imputing attributes with excessive missing values, unless that one attribute has strong correlation with the target variable. In this case I chose to drop attributes with more than 500 missing values. I imputed numerical attributes with the mean of available values and for categorical attributes, I used the most frequent technique to impute.

## Outlier Detection and handling

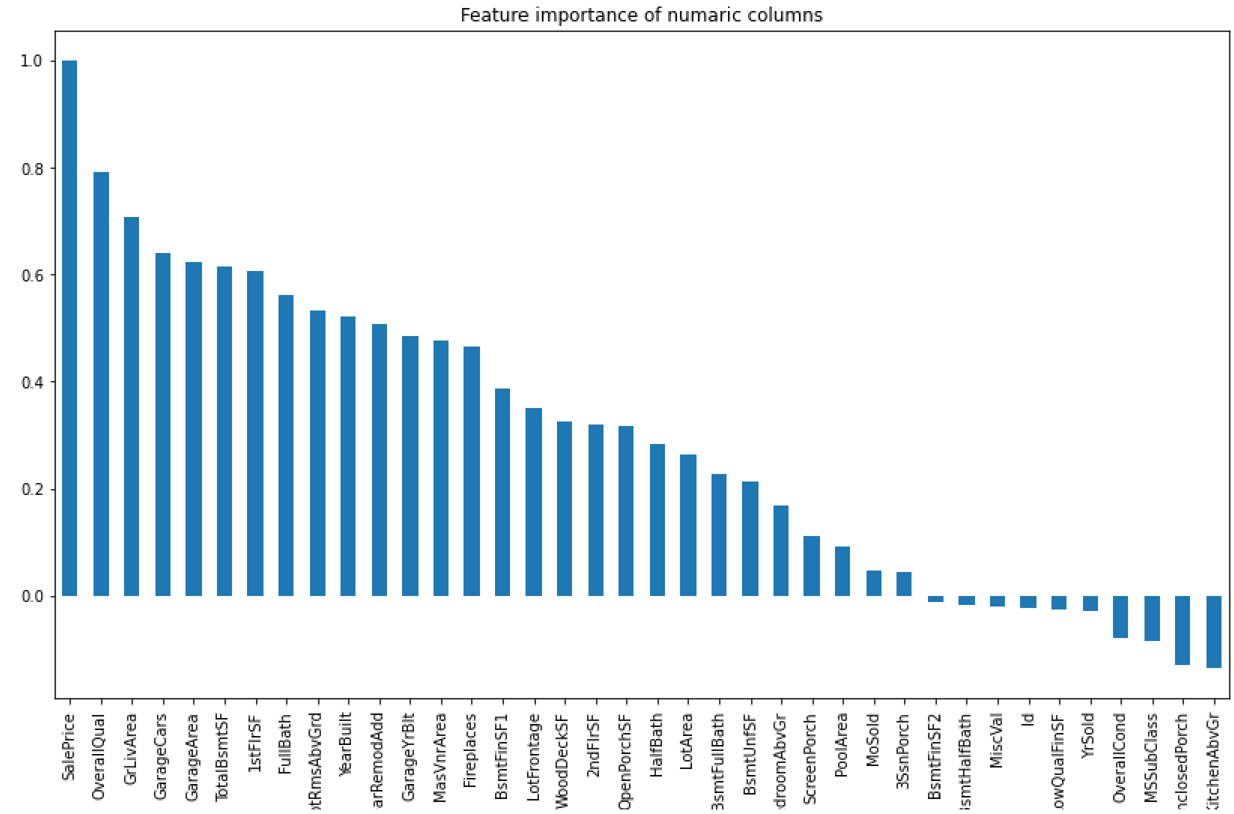
Outliers are the abnormal or extreme values for an attribute. For example, in a person database observation with the age of 200 years. Or in a school kid's dataset an observation with age of 81. These types of values exist due to variety of reasons including recording error. Outliers impact the model train and its predictability, and that is why it is one of the most important steps during the data analysis to handle outliers in any dataset. I looked at all the attributes in the dataset and identified any outliers. I use scatter plots to plot the data as dots in a two-dimensional plot against target variable which makes a bit easier to observe the outliers and come up with threshold to drop extreme values. In this data set there were a few attributes with outliers. For example, Gross living area had most of the observations between 500 to 3000 square feet and only a few more then 3000 square feet. I dropped all the observations with more than 4000 square feet. Similarly, I dropped observations with Lot size of more than 150000 sq feet with low sale price.





## Feature Selection

In this data set 81 attributes including the target variable sale price. As we include more features to the model training the process of training slows down. It is important to select which columns should be the part of the model training process. One way of selecting feature is to generate a correlation matrix and choose most correlated attributes to target variable. Also, if there are two attributes which are correlated in exactly same way to the target variable, they are redundant, and one should be dropped. I calculated the correlation matrix of all the attributes with the target variable and plotted them in a bar chart. Below chart shows the correlation of individual attributes with target variable. Values above 0.6 and below -0.6 (same strength of the correlation, but the reverse direction. Which means when value for the attribute increases, the value for target value decreases and vice versa) are considered the strongly correlated attributes.



## Methods

I chose a data set from Kaggle @ https://www.kaggle.com/bakar31/eda-house-price-prediction?select=train.csv. The method I implemented in this project was to, first split the data set into two datasets and set aside 20% of the observations for validation of the trained model. All the training is done only in the 80% of the entire dataset and in the end, I used the complete set to train a new model for production deployment.

After applying above stated data cleaning steps and a deep dive of exploratory data analysis, I started training models on the train set. This is a regression problem as target variable is a continuous variable. I chose Lasso, Ridge, Random Forest regressor, Gradient boost regressor, XGBoost regressor to train models using the train set. I prefer to train multiple models to checkout which one performs better, and it gives me a few options to choose from.

Each of these algorithms are trained using k-fold cross validation technique. I used the value of 10 for CV. This way it is easy to evaluate the performance of the model without using the test set, which can be only used once.

## Results

All the models performed pretty well and had a r2 score of more than 85% with XGBoost coming out as the best performing model with 90.7%.

## Discussion/conclusion

In this project, I decided to train multiple models. There are many advantages of training multiple models including, we get to compare them and choose the best performing model. One possible improvement in this project is to implement an exhaustive training and tuning the hyper parameters of individual algorithms. Finding the optimal hyperparameter values is crucial to have a better performing model.

House price is a very important data to predict as multiple stakeholders might have interest in finding the right sale price for a house. Irrespective of the market condition, economy and houses available in the market for sale, this model can be used to provide a guiding line.

## Additional Questions

Below are the questions I expect audience might be asking while presentation:

1. Other companies like Zillow already have models like this to predict the price of a house and come up with an estimate (z-estimate for Zillow). How is this model any different from those existing ones?
2. What are the top attributes of a house that you think affects most the price of a house?
3. Is this work extensible to other real estate property types, for example, land, commercial buildings, multi-family properties?
4. How do you plan to accommodate other real-world factors like demand and supply of the housing market, raw material supply situations, recession, etc.?
5. What do you propose to keep this model up to date and process new observations and keep learning to predict right price for a house?

6. Can this model be used to (or extended to) determine the price of a new construction?
7. There is a wide variety of builders from cookie cutters to extremely custom home builders who can customize everything in a house including the structures. Does this model account for at least the top-level known builders in the area?
8. Some houses have covenants and house owners' associations. Does this impact the price of a house positively or negatively?
9. Is there any correlation between the year the house was build and final price of the house?
10. How does a recently remodeled old house (more than 50 years) compare with new construction, if all other factors are unchanged?

## Acknowledgments

I have referred to Kaggle for the data set, data understanding and business understanding and some of the notebooks available in the Kaggle. I have also referred to multiple websites including data science articles from various authors, geeks for geeks, machine learning websites, mediam.com, and many others for the basic machine learning and statistics concepts and practical examples. I referred to Zillow website to understand some of the language used around house and estimated price of a house. I acknowledge Professor William's guidance and help in the success of this project.

## References

1. National Association of Realtors - https://www.nar.realtor/research-and-statistics/housing-statistics/existing-home-sales
2. Kaggle data set - https://www.kaggle.com/bakar31/eda-house-price-prediction?select=train.csv
3. Zillow https://www.zillow.com/homedetails/22803-Hansen-Ave-Elkhorn-NE-68022/91935264_zpid/

4. Redfin - https://www.redfin.com/NE/Omaha/16823-Pasadena-Ct-68130/home/63901488
5. Mark Chalon Smith - https://www.insurance.com/home-and-renters-insurance/home-insurance-basics/5-factors-that-affect-rates.html
6. Norada - https://www.noradarealestate.com/blog/housing-market-predictions/
7. James Gheen - https://www.biggerinvesting.com/6-types-of-real-property-infographic-real-estate-investing/