House Price Prediction [Draft White Paper]

Business Problem

House Prices vary frequently, due to socio-economic factor changes. In is white paper I am trying to fit a model that will predict the price of the house depending on different factors.

Background/History

I was looking for various sources of data to predict the house price and initially I chose data from Kaggle for this project, the data contained everything, but no time related information was there. So, the ML model that will be coming out of the data will be extremely hard to use in future. So, I changed the dataset to another as mentioned below, that has the information of when the house was build, so it will act as an age of the house, which has major impact on the house price.

Data Explanation (Data Prep/Data Dictionary/etc)

The Data I used from Kaggle have several field about the house, out of which, some of the important fields are as follows.

<u>Numerical Variables:</u>

SalesPrice: This variable represents the sales price of the house
LotArea: This variable represents the size of a lot in square feet
OverallQual: This variable represents rates of the overall material and finish of the house
OverallCond: This variable represents rates the overall conditions of the house
1stFirSF: This variable represents the first floor in square feet
2ndFirSF: This variable represents the second floor in square feet
BedroomAbvGr: This variable represents bedrooms above grade (does not include basement bedrooms)
YearBuilt: This variable represents the original construction date.

<u>categorical variables:</u>

MSZoning: This variable identifies the general zoning classification of the sale
LotShape: This variable represents the general shape of the property
Neighborhood: This variable represents physical locations within Ames city limits
CentralAir: This variable represents central air conditioning
SaleCondition: This variable represents condition of sale

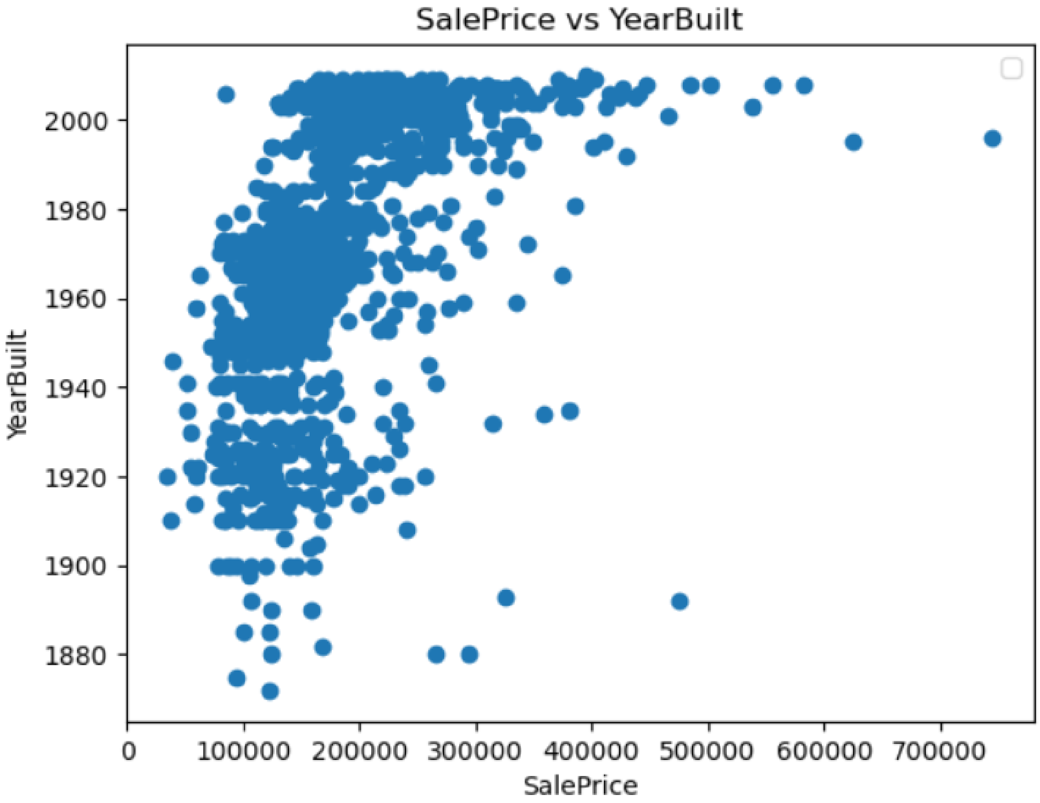
MoSold:	This variable represents month sold (MM)
YrSold:	This variable represents year sold (YYYY)

Methods

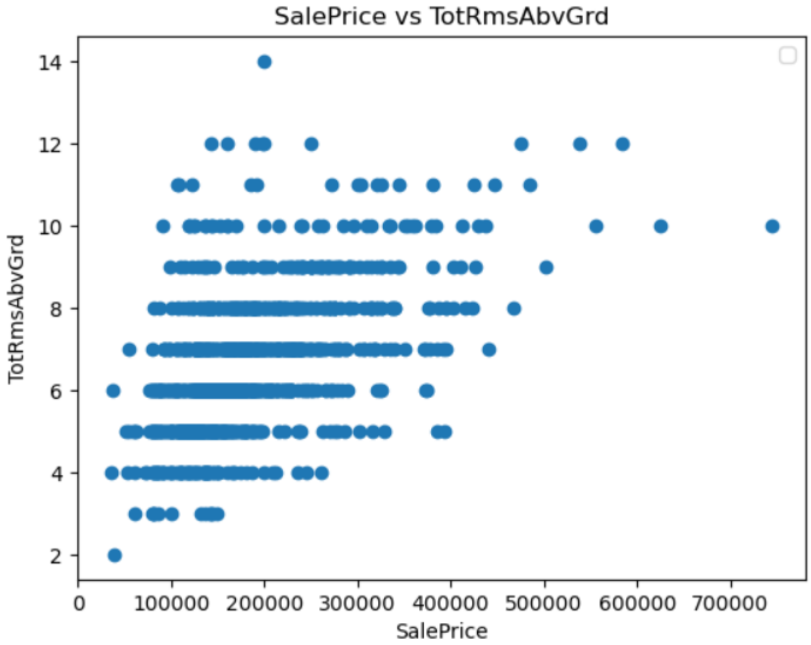
The prediction of house price from the data, take some analysis, I have used xgboost regressor model to predict the house price. Initially I used the train test split method to split the data into train and test, Then I have used the train data to fit the xgboost model. After the fitting done, I applied the model in test data, to check the accuracy of the model. R Squared represents the proportion of variance for a dependent variable explained by an independent variable in a regression model.

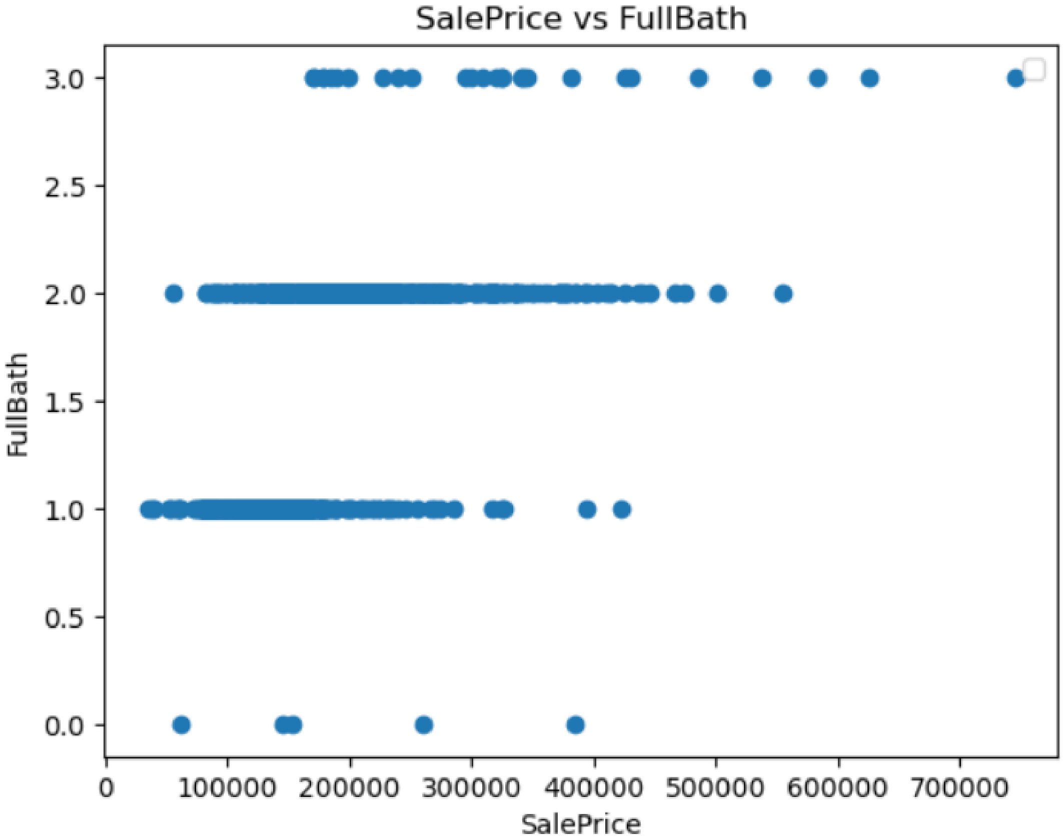
Analysis

I did some analysis on housing price, what are the reasons for house prices.
First is Year Build, looking at the image, it clearly says year build has some effect but not completely drive the price of of the house.

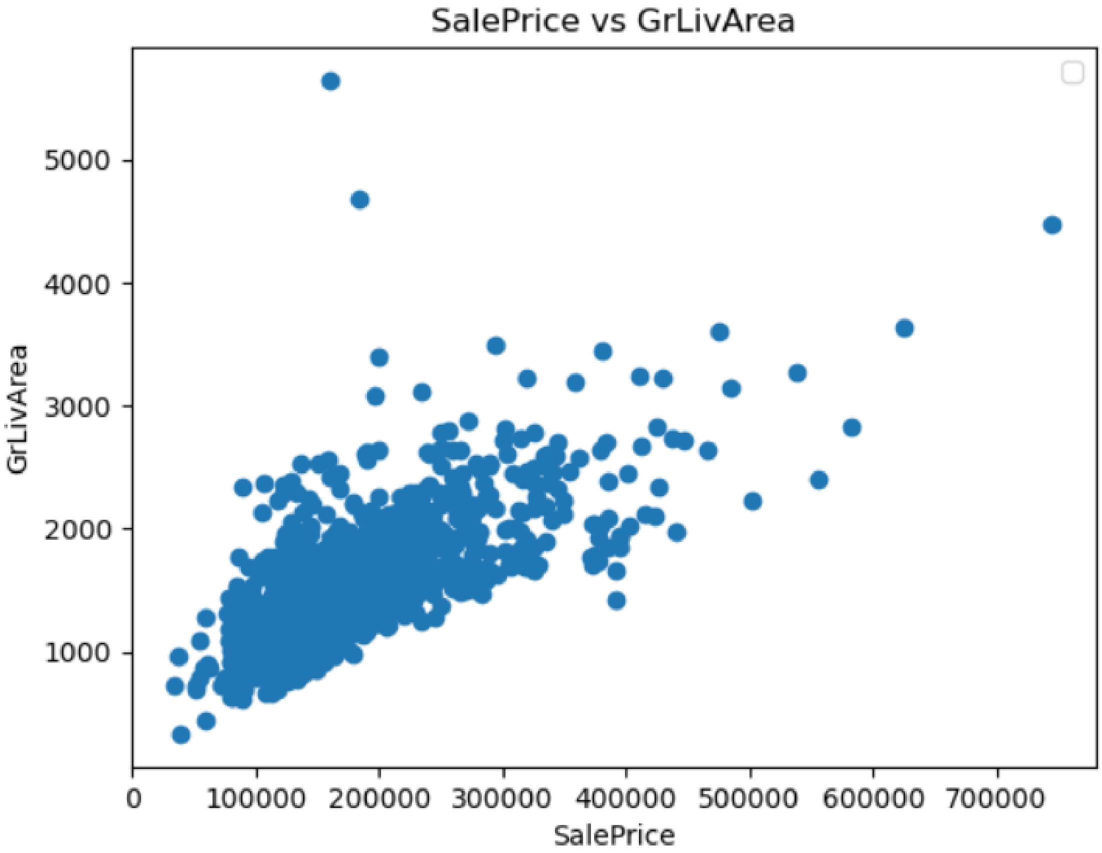


The number of rooms and number of full bathrooms have similar effects on house prices, if these increases, the house prices increase.

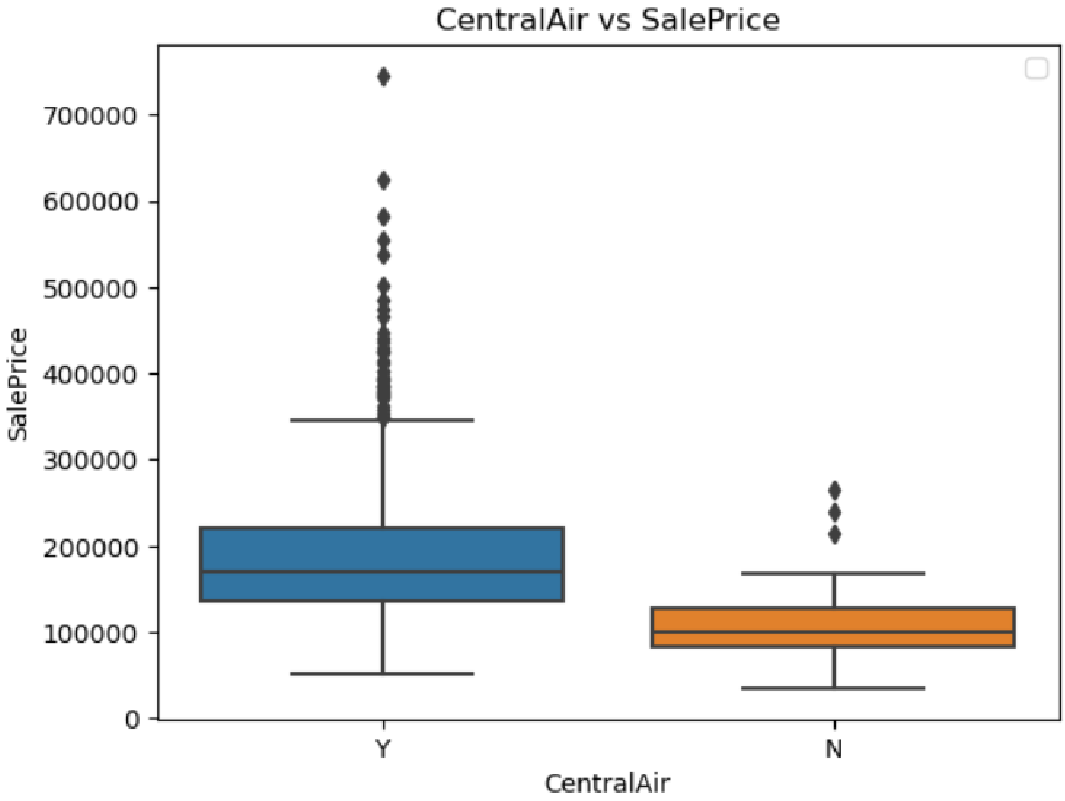




Living area in gound floor is important and has major impact on the house price.



It is obvious and can be seen in the plot that not having central air conditioning system decreases house price.



Conclusion

We have split the model into 80 and 20 for train and test, then fit the XGBoost model into it after selecting some categorical and continuous variable. Then I test the model on test data. The R-squared Score came as 0.8852383252668263. Which means 88% of variance for the dependent variable can be explained by the independent variable in the model.

Assumptions

The only assumption is that the encoded integer value for each variable should have an ordinal relation.

Limitations

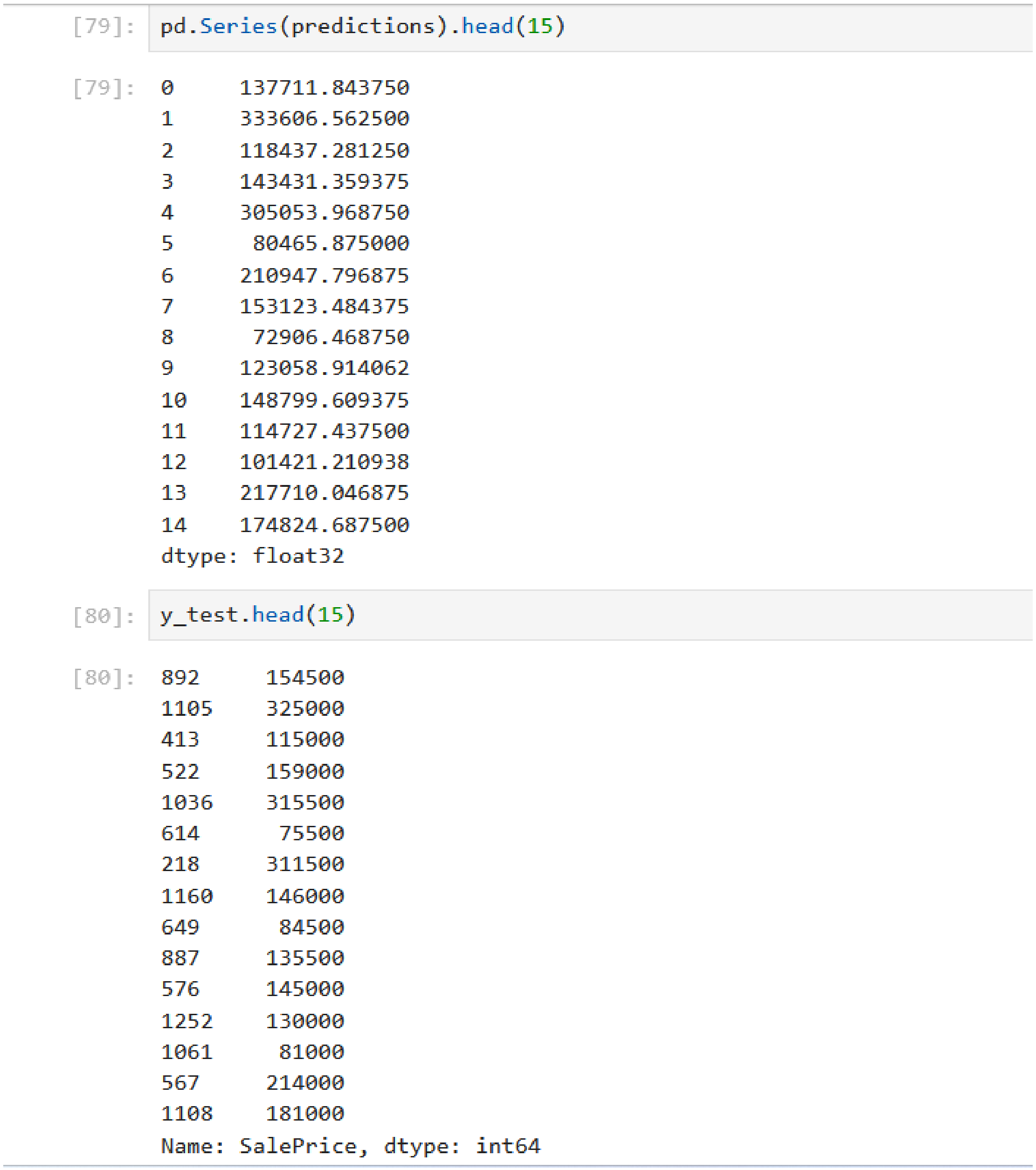
The house price prediction involves several attributes. There are a few attributes, those could have a major impact on house price, are not available in the dataset for this model fitting. Some of the attributes like the information on locality about crime rate and walking scores sometime have significant impact on House Pricing. Another principal factor is school district reading for Elementary, Middle and High School.

Challenges

There are two major challenges in this project development identifying attributes for feature selection: I used correlation techniques to identify the numerical attributes and did some analytical analysis for the categorical variables. Another challenge was to choose the model, here I have to predict the house price, which is a continuous variable, so I have choosen the XGBoost regression.

Future Uses/Additional Applications

I have fitted a XGBoost model to predict the house price and tested the model in the testing sample. I have checked the 15 samples from the test data to see how the prediction came up, the estimated price is quite like the actual value.



```
[79]: pd.Series(predictions).head(15)
```

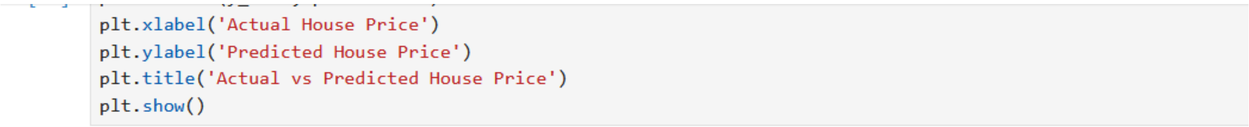
```
[79]: 0     137711.843750
      1     333606.562500
      2     118437.281250
      3     143431.359375
      4     305053.968750
      5      80465.875000
      6     210947.796875
      7     153123.484375
      8      72906.468750
      9     123058.914062
      10    148799.609375
      11    114727.437500
      12    101421.210938
      13    217710.046875
      14    174824.687500
      dtype: float32
```

```
[80]: y_test.head(15)
```

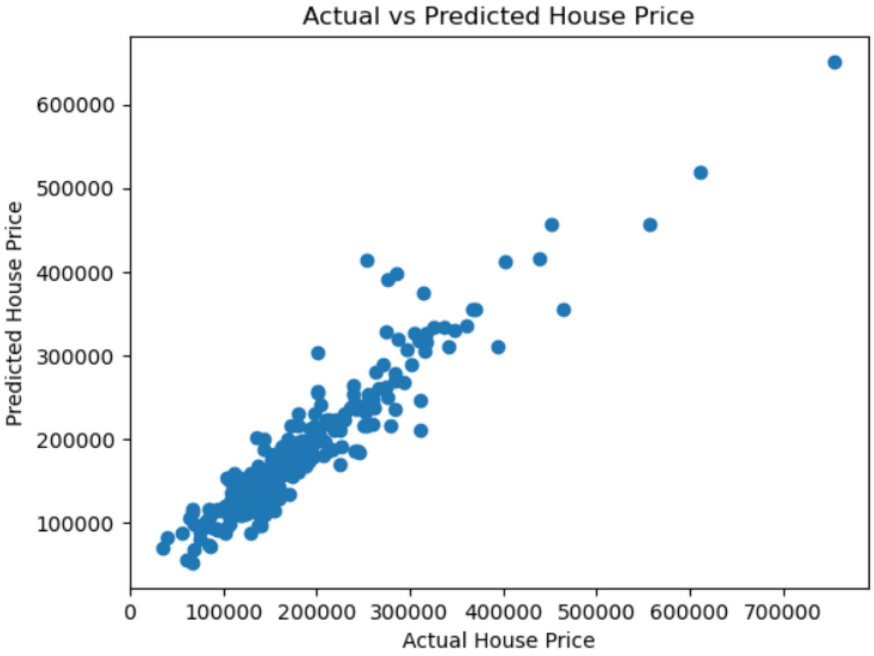
```
[80]: 892     154500
      1105    325000
      413     115000
      522     159000
      1036    315500
      614      75500
      218     311500
      1160    146000
      649      84500
      887     135500
      576     145000
      1252    130000
      1061     81000
      567     214000
      1108    181000
      Name: SalePrice, dtype: int64
```

Recommendations

I have checked the accuracy of the model, but fitting the Actual vs Predicted House Price. Looking at the graph it looks like the fitting is quite accurate, as the predicted and actual house prices are similar. I can recommend this model showing the accuracy chart. This model can be used to predict the House Price.



```
plt.xlabel('Actual House Price')
plt.ylabel('Predicted House Price')
plt.title('Actual vs Predicted House Price')
plt.show()
```



Implementation Plan

There are multiple ways to implement this model, one of the ways is to host a website and create an API. If we host a website with form, that will take the information about the house, and it will predict the house price.

Ethical Assessment

There may be several ethical implications for house price perdition, as house price is an economic information, we must be very sure before publishing this in the outside world, there are several decisions, that can be made based on the estimated price. As the consumer knows the house that they want the prediction, consumers will have the address and other information about the location, now if the prediction came wrong, it would create a wrong impression about the locality and about the house. The owner of the house can be impacted by that.