

HOUSING PROJECT

Submitted by:

Pranjal Gandotra

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

This problem contains the prices of houses along with features of those houses, for example their price, if they have garage or not, number of rooms etc. In total there are around 80 unique features which we look at.

Our task is to create a machine learning model to predict prices of houses.

* Conceptual Background of the Domain Problem

A good understanding of financial regulations, related to housing properties is a must to understand this project.

* Review of Literature

These pieces of literature go in depth of how house prices vary.

<https://voxeu.org/article/what-drives-house-prices-some-lessons-literature>

<https://www.researchgate.net/publication/324417771_Selected_Review_of_the_Empirical_Literature_on_House_Price_Modelling_and_Forecasting_What_Does_the_Literature_Say>

* Motivation for the Problem Undertaken

To get a more accurate value of what a house can be sold for.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

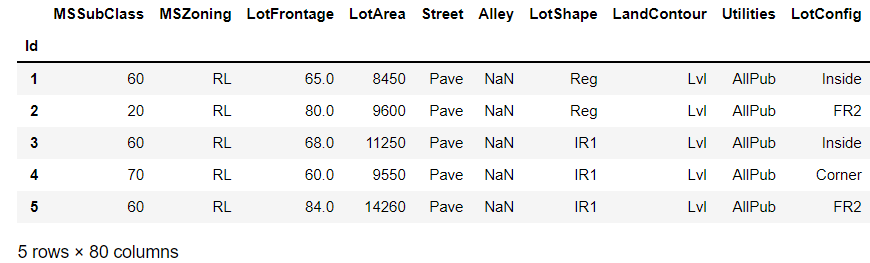
Inbuilt function such as standardising and log will be used in tackling this problem.

As the testing dataset doesn’t contain the prices. I will be testing the accuracy of my model by splitting the training data first.

The main metric will be Mean Absolute Error (MAE) for judging the accuracy of the trained models.

MAE = | Predicted price – Actual price |

* Data Sources and their formats

****

Snapshot of the data, this data was provided by FlipRobo.

Two files were provided, one containing a training data set and another file for testing our model. In total we have 80 columns worth of data and 1168 unique house prices in our training dataset.

* Data Pre-processing Done

Columns where more than 45% of the data was missing were dropped. As imputation with < 50% available data doesn’t work very well.

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

* State the set of assumptions (if any) related to the problem under consideration

The main assumption is that there is no selection bias in the data

which we have.

* Hardware and Software Requirements and Tools Used

Pandas, Seaborn, ploty and sickit libraries were used throughout the project.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

**Decision tree regression**

**Random forest regression**

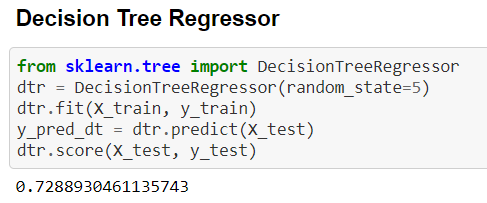
**Lasso regression**

**Support vector regression**

**Linear regression**

**XGB boost regression**

* Run and Evaluate selected models



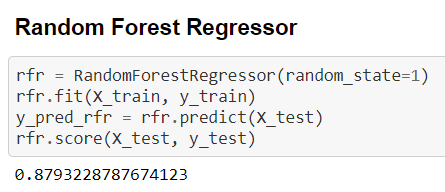
Mean absolute error = 28864.67

Mean squared error = 1779561053.36

Median absolute error = 19000.0

Explain variance score = 0.73

R2 score = 0.73



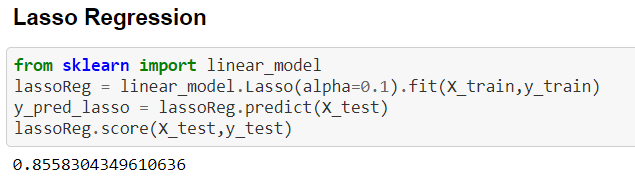
Mean absolute error = 19050.39

Mean squared error = 792131304.27

Median absolute error = 13370.5

Explain variance score = 0.88

R2 score = 0.88



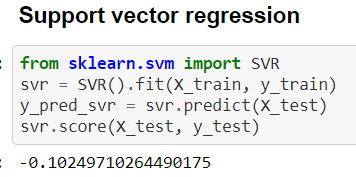
Mean absolute error = 21373.51

Mean squared error = 946337005.91

Median absolute error = 15125.05

Explain variance score = 0.86

R2 score = 0.8



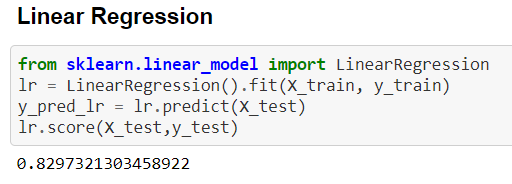
Mean absolute error = 62435.41

Mean squared error = 7236852014.26

Median absolute error = 45956.44

Explain variance score = 0.0

R2 score = -0.1



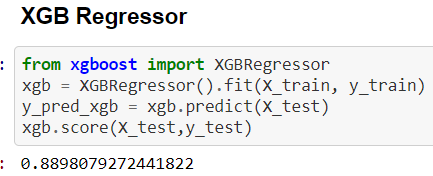
Mean absolute error = 22242.31

Mean squared error = 1117647722.17

Median absolute error = 15164.24

Explain variance score = 0.83

R2 score = 0.83



Mean absolute error = 18543.73

Mean squared error = 723306865.63

Median absolute error = 13697.06

Explain variance score = 0.89

R2 score = 0.89

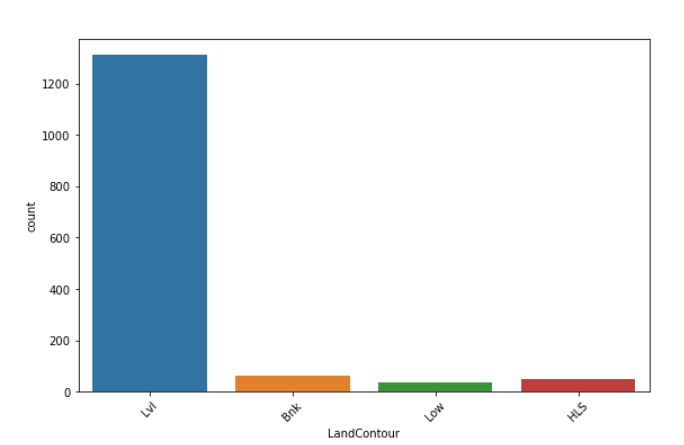
With this data we can observer that the model with XGB regression performed the best.

* Key Metrics for success in solving problem under consideration

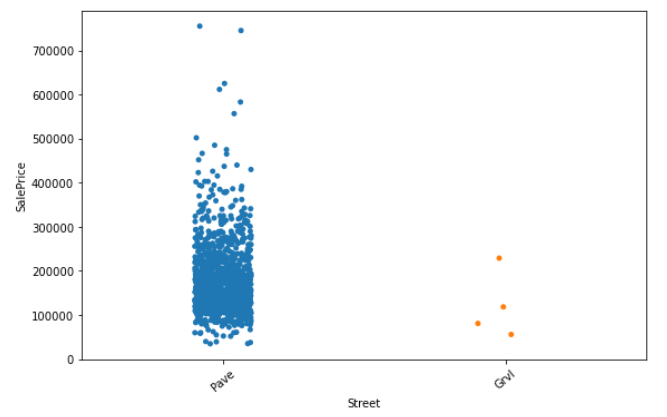
Mean Absolute error; Mean absolute error is the modulus of the difference between the predicted price and actual price.

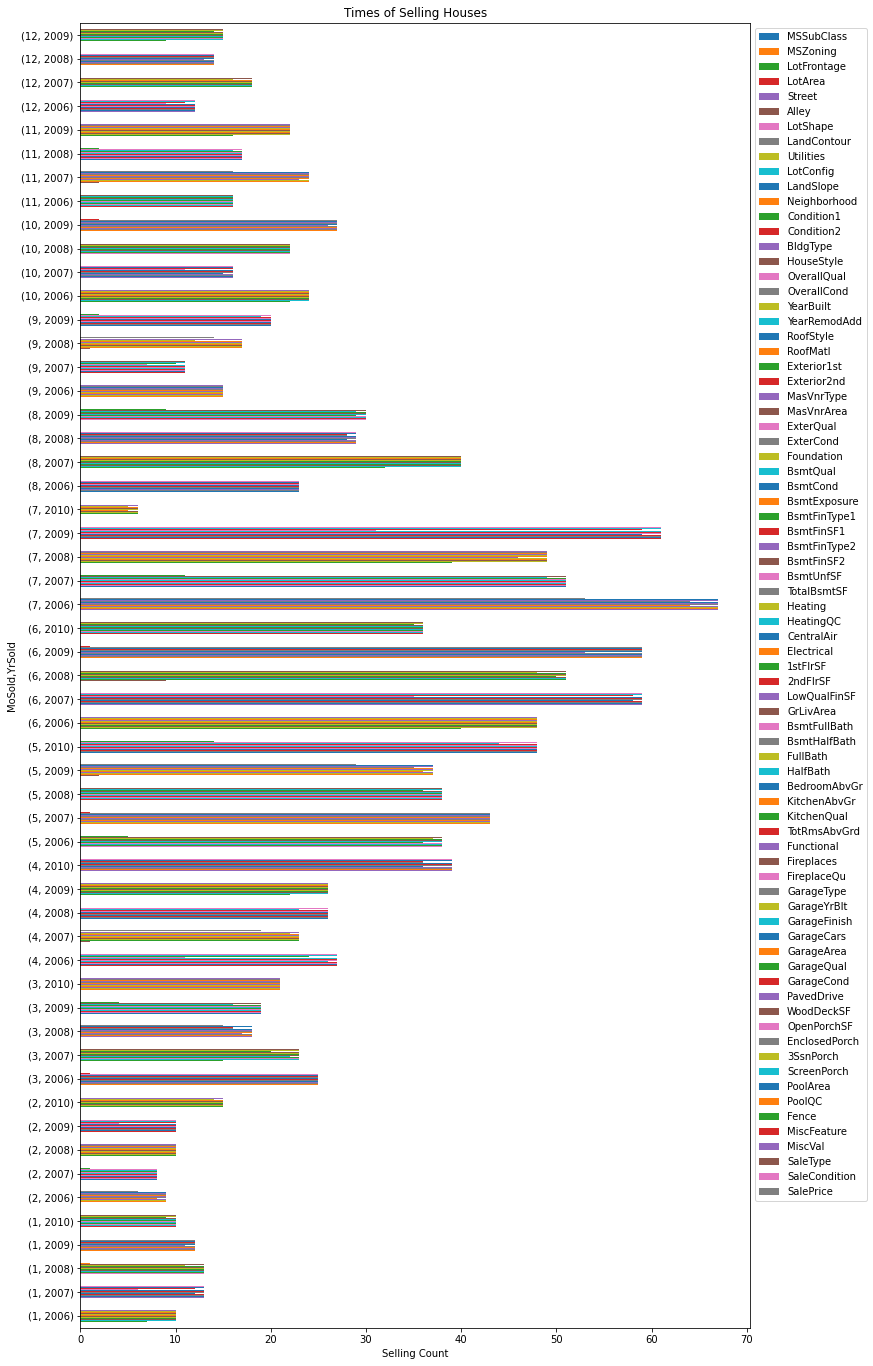
The lower the MAE the better is the model.

* Visualizations

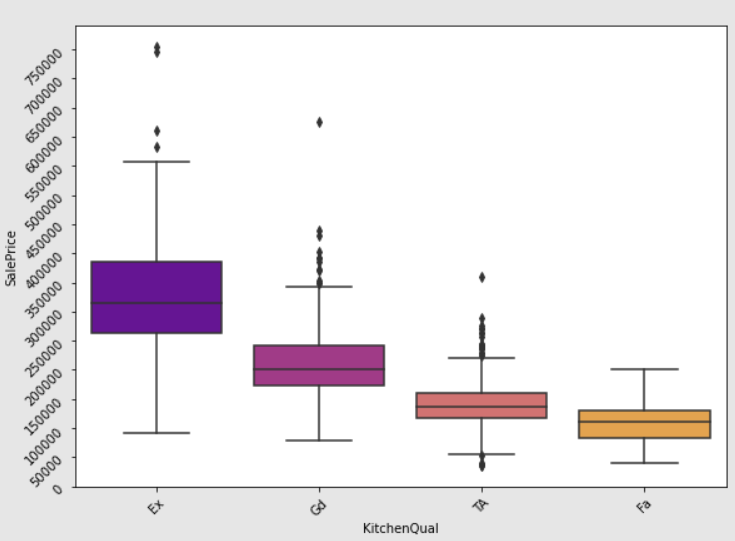
Count plot for all the categorical variables were created, here is one such example. This graph contain how the land is contoured around the house

An example of strip plot. This plot contains the price house in relation if the street around the house is paved or not.





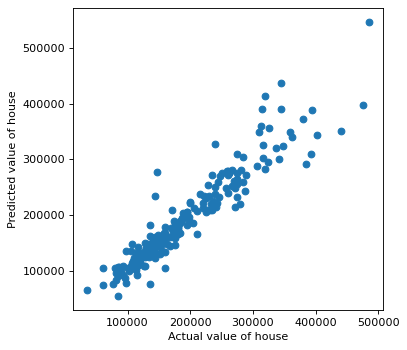
The above graph contains the data for each the number of houses sold, grouped by first the month of the year and then the year of the year. This gives us the trend of around which time the houses are sold.



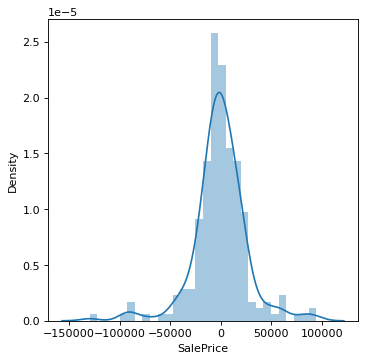
Here in the plot above we can see how kitchen quality affects the price of a house.

* Interpretation of the Results

Results were concluded from scatter plot of the predictions vs actual values and the mean absolute error between the two.



We can observe the straight line, which means we are getting accurate results.



This is the distribution plot of the MAE, most of the predictions have a low MAE which is good.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Predicting of the house prices is a complex task, here we have used complex machine learning algorithms to help us solve the problem.

We have used the MAE as the main metric.

* Learning Outcomes of the Study in respect of Data Science

XGB boost algorithm works best for this particular data set, hyper parameter tuning was performed and optimal parameters were found.

* Limitations of this work and Scope for Future Work

The dataset only contains the houses for one city. If we want to make a model which can predict the price for any house, we need a lot more data from all countries and cities across the world. Then we will be able to achieve a generalised model to predict house prices.