Logo, company name

Description automatically generated

**Analysis of Factors affecting the House Prices and their Prediction**

**OUTLIERS [Shishir Tripathi, Ruchin Dhama, Gursimran Singh, Rajat]**

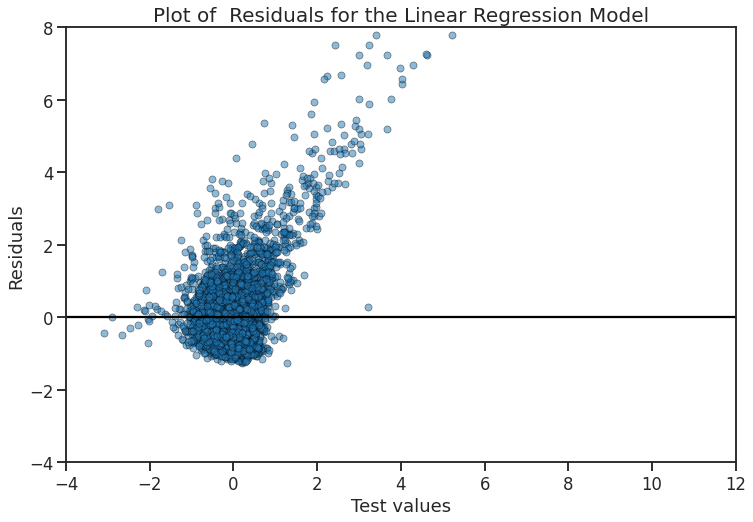
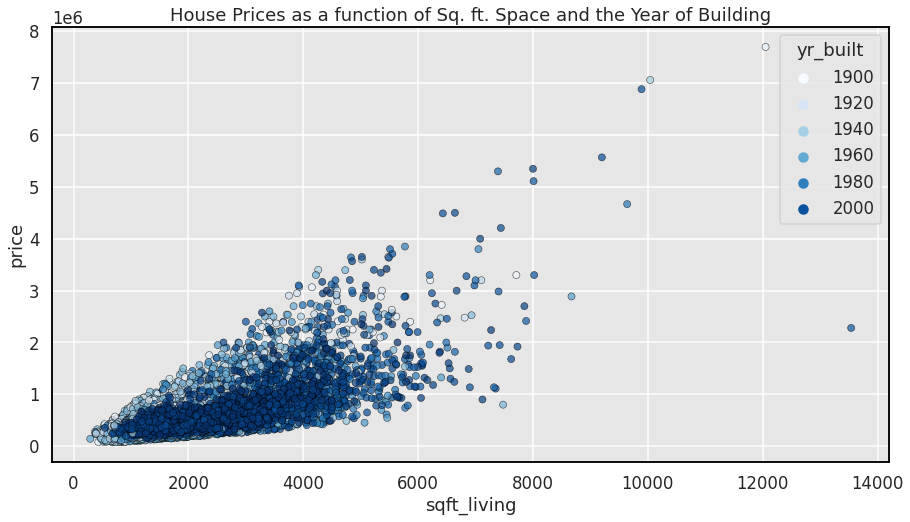
**Model Used**

Linear/Logistic Regression: Using this type of model we tried to figure out whether or not our data points have some linear/polynomial relationship.

KNN: Then we moved to KNN to make use of similarity amongst the more closer data points and based on that then predicting the price.

Tree based: Finally we applied tree based algorithm to figure out underlying relationship as well as how well the neighbours define a datapoint.

Ensemble Models: We also tried ensemble models like bagging and boosting, which outperformed all the basic models in results.

**Conclusion and Future Work**

In conclusion :

The more important features for a house price are not only the sqft area of that house but also the sqft area of nearby houses.

The grade of a house have a direct correlation with the house price.

The last renovation date is more important than the overall age of the house.

And finally to our surprise the number of bathrooms affects the house price much more as compared to the number of rooms in an house.

Future Work:

We can try to gather the data for more than one year so that we can even apply time series on the dataset and be able to predict the house prices in the future based on the features they have.

**Motivation and About the Project**

Buying anything in the market whose Price is not fixed gets a lot harder when the buyer doesn’t even know what the price is affected by, like in the case of buying a House.

Based on dataset and findings we want to:

* Explain what are the important factors of house(sqft area, floors condition etc.) which makes a house affordable.
* Show next time you go in market to buy house what you should pay more attention to.
* How to find a rough estimate of a house’s price.

**References**

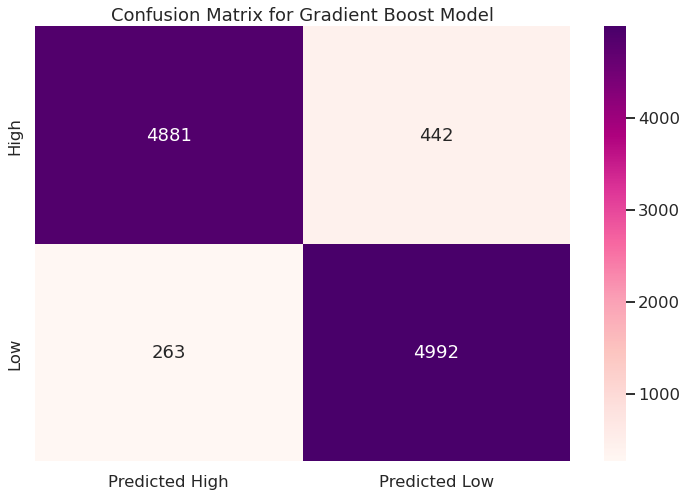
<https://numpy.org/doc/>

<https://pandas.pydata.org/docs/>

<https://scikit-learn.org/stable/>

**Results**

In Regression the best results we got from boosting algorithm with an R2 score of 0.90.

In classification as well we got our best results from boosting which gave us an accuracy of about 93% and recall and f1 score of 0.95 and 0.93 respectively for affordable houses class. 

**Data and Labels**

The dataset consists of historic data of houses sold between May 2014 to May 2015 house prices in King County an area in the US State of Washington, this data also covers Seattle. The dataset was obtained from Kaggle. This data was published/released under CC0: Public Domain. The dataset consisted of 21 variables and 21613 observations.

Initially the dataset was regression based but we converted it for classification and labels are “Affordable” and “Not Affordable”