

**HOUSING: PRICE PREDICTION**

Submitted by:

Ravinder Singh

**ACKNOWLEDGMENT**

1. data source, https://raw.githubusercontent.com/shughestr/PIMS\_2020\_Real\_Estate\_ data/master/sample\_clean.csv.

2. Kaggle competition, <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>.

3. Mckinsey report, https://www.mckinsey.com/industries/capital-projects-and-infrastructure/ our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate

4. Zillow zestimate, <https://www.zillow.com/blog/zestimate-updates-230614/>.

5. Great Learnings <https://www.mygreatlearning.com/blog/what-is-ridge-regression/>

**INTRODUCTION**

* Business Problem Framing

Which variables are important to predict the price of variable?

Key factors on which the price off a house depends in particular area eg. Total area of the house,Locality, Aminities ,Etc

How do these variables describe the price of the house?

By how much the Price varies according to the change in these key factors

* Conceptual Background of the Domain Problem

The Domain is Real Estate , If the Person preforming analysis is a localite or a person with theBackground then He can point out some of the Key features that might be Absent in the data set.

* Review of Literature

Correlation Matrix and pair plot Explain a lot abot the key factors on which the Sale Price depends

* Motivation for the Problem Undertaken

Obejctive is to develop a better understanding of the Multiple Linear Regression

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The problem is a Multiple Regresssion Problem so Model than can preform beter are Ridge as it helps us to reduce the error.

Also I have tried XGBoost Regression and Random Forest as they both works better for the Non – Linear relationship

* Data Sources and their formats

Two separate Data set were provided , both in Csv format named Train and test.

* Data Preprocessing Done

1. While reading the file some columns in Dataset had Na as the categorical value where as Pandas read it as missing values.So, I replaced those columns Na values to None.
2. MSSubClass, OverallQual, OverallCond needed to be converted to object type as they were read as Integer
3. LotFrontage and MasVnrArea needs to be converted to int or float type as they were read as Float.
4. Dropped the Garage Built year as mostly it was same the house built Year

* Data Inputs- Logic- Output Relationships

*GrLivArea :an increase of 1 square foot of house area above ground, the price will increase by 1.065185 times*

*1stFlrSF : if an increase in First Floor square feet area by 1 then price will increase by 1.052968 times*

*TotalBsmtSF :an increase of 1 square foot of basement area above ground, the price will increase by 1.046094 times*

*OverallQual\_9 if the overall material and finish of the house is Excellent, the price of house will increase by 1.039479 times*

*Neighborhood\_StoneBr: if Stone Brook is a nearby location, then the price of house will increase by 1.030926 times*

*GarageArea: an increase of 1 square foot of Garage Area ,the price will increase by 1.027725 times*

*GarageCars:an increase 1 car capacity of Garage Cars, the price will increase by 1.027691 times*

*CentralAir\_Y :if there is a Central air conditioning, the price will increase by 1.030009 times*

*LotArea : an increase of 1 square foot of Lot Area ,the price will increase by 1.033287 times*

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

We have assumed that there were Outliers(they can be the actual data points) in many columns so we capped them using Inter Quartile Range

Id in the Data set is not useful for the prediction of the house Price

* Hardware and Software Requirements and Tools Used

Ram : Minimum 4 GB

Processor : Intel I 3rd(8th generation)

Memory(HDD): 256 Gb+256gb(SSD)

(Alternative) Google Colab

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

In robust statistics, robust regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable

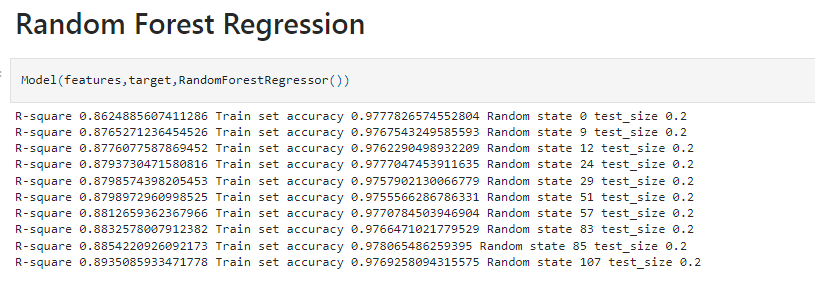
* Testing of Identified Approaches (Algorithms)

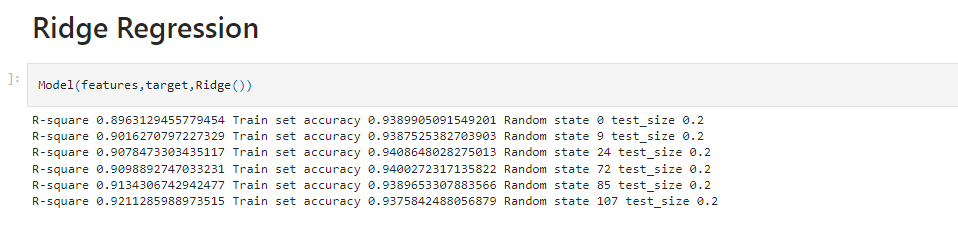
Ridge Regressor

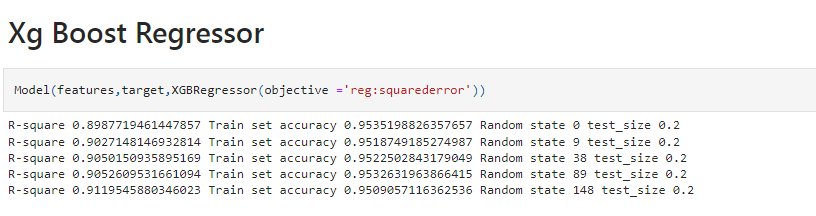
Random Forest Regressor

XgBoost Regressor

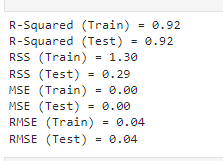
* Run and Evaluate selected models





The accuracy in the screen shot are actually R- Squared Score for the individual Models

* Key Metrics for success in solving problem under consideration



There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)

* Visualizations

Bar plot, Distance plot, Boxplot, line plot, pair plot

* Interpretation of the Results

Built Area of the House including Basement, First floor, Second Floor was one the most important area in the House Prediction in that area

**CONCLUSION**

* Key Findings and Conclusions of the Study

To make sure a model is robust and generalizable, we have to **take care it doesn't overfit**. This is because an overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data.

* Learning Outcomes of the Study in respect of Data Science

The objective of study of the multiple regression Problem is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

* Limitations of this work and Scope for Future Work

Limitations of the Solution is that the it can predict the price of a house which is in Over priced due to the Features that are been Installed by the owner of the House