ACKNOWLEDGMENT

I take this opportunity to express our sincere gratitude to Ms. Sarishma for her valuable guidance in this project report without which it would not have been completed. I am very much grateful to her for her untiring assistance in this report and she has been encouraging us in eliminating the errors. The report has been developed as a result of her valuable advice.

I am thankful to my classmates and friends for their support and guidance. I also like to thank researchers and scholars whose papers and thesis have been utilized in this report.

Graphic Era Deemed to be University
Dehradun, Uttarakhand

A Mini Project Report
On

House Price Prediction using Multiple Regression(ML).

Name : Deepanshu

University Roll No. : 2016721

Course : B. Tech (CSE)

Semester : 3rd

Guided By : Ms. Sarishma

Department of Computer Science and Engineering

PROJECT-REPORT

1. INTRODUCTION:

1.1 Goal – The goal of the project is to predict the price of the Houses using different Regression Techniques taking into account various "features" of the dataset on which housing prices depend.

1.2 Motivation – Being extremely interested in everything having a relation with the Machine Learning and Data Science, the independent project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities.

1.3 System Requirements –

1.3.1 Operating System – Windows 10 / Ubuntu / Mac OS
1.3.2 System Architecture – 64-bit System Required
1.3.3 Software Required – Anaconda (Jupyter Notebook)
1.3.4 Additional Software Required – Web Browser

Or

2.3.1 Additionally this Notebook can also be accessed by uploading it on Google Colaboratory.

2. METHODOLOGY FOLLOWED-

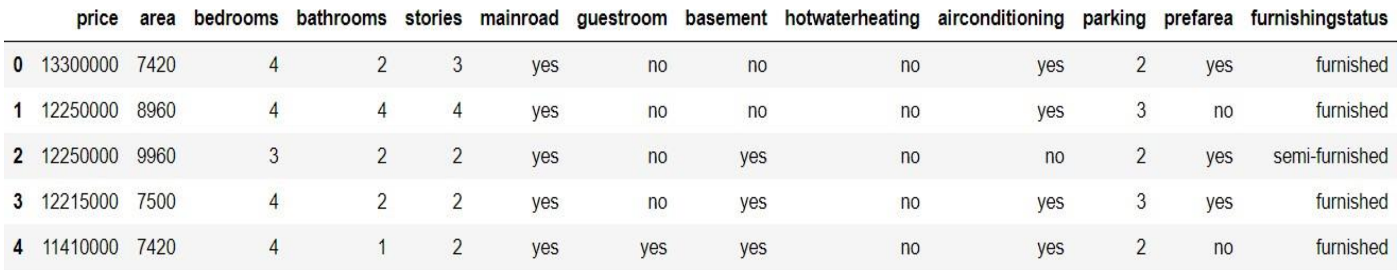
3.1 Libraries Used-

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Plotly

6. Sklearn

3.2 Data –

The crucial element in machine learning task for which a particular attention should be clearly taken is the data. Indeed, the results will be highly influenced by the data based on where did we find them, how are they formatted, are they consistent, is there any outlier and so on. At this step, many questions should be answered in order to guarantee that the learning algorithm will be efficient and accurate.



	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

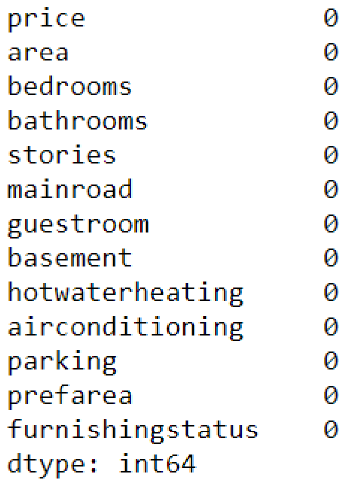
This is how the first 5 rows of our dataset looks like.

Our Dataset has 545 Rows and 13 Columns in total.

3.3 Description of all the Columns –

Column	Datatype	Unique Values	Description
Price	Integer	Continuous Variable	Sale Price of the House
Area	Integer	Continuous Variable	Area of the house
Bedrooms	Integer	1, 2, 3, 4, 5, 6	Number of Bedrooms
Bathrooms	Integer	1, 2, 3, 4	Number of Bathrooms
Stories	Integer	1, 2, 3, 4	Number of Floors
Main Road	String	yes, no	Is house on Main Road
Guestrooms	String	yes, no	Does it have a Guestroom
Basement	String	yes, no	Does it have a Basement
Hotwater Heating	String	yes, no	Is it Installed
Airconditioning	String	yes, no	Is AC Installed
Parking	Integer	0, 1, 2, 3	Number of Parking
Prefarea	String	yes, no	Is it buyer's preferred area
Furnishing Status	String	Un, Semi, Furnished	Furnishing status of house

3.4 NULL Values in Each Column –



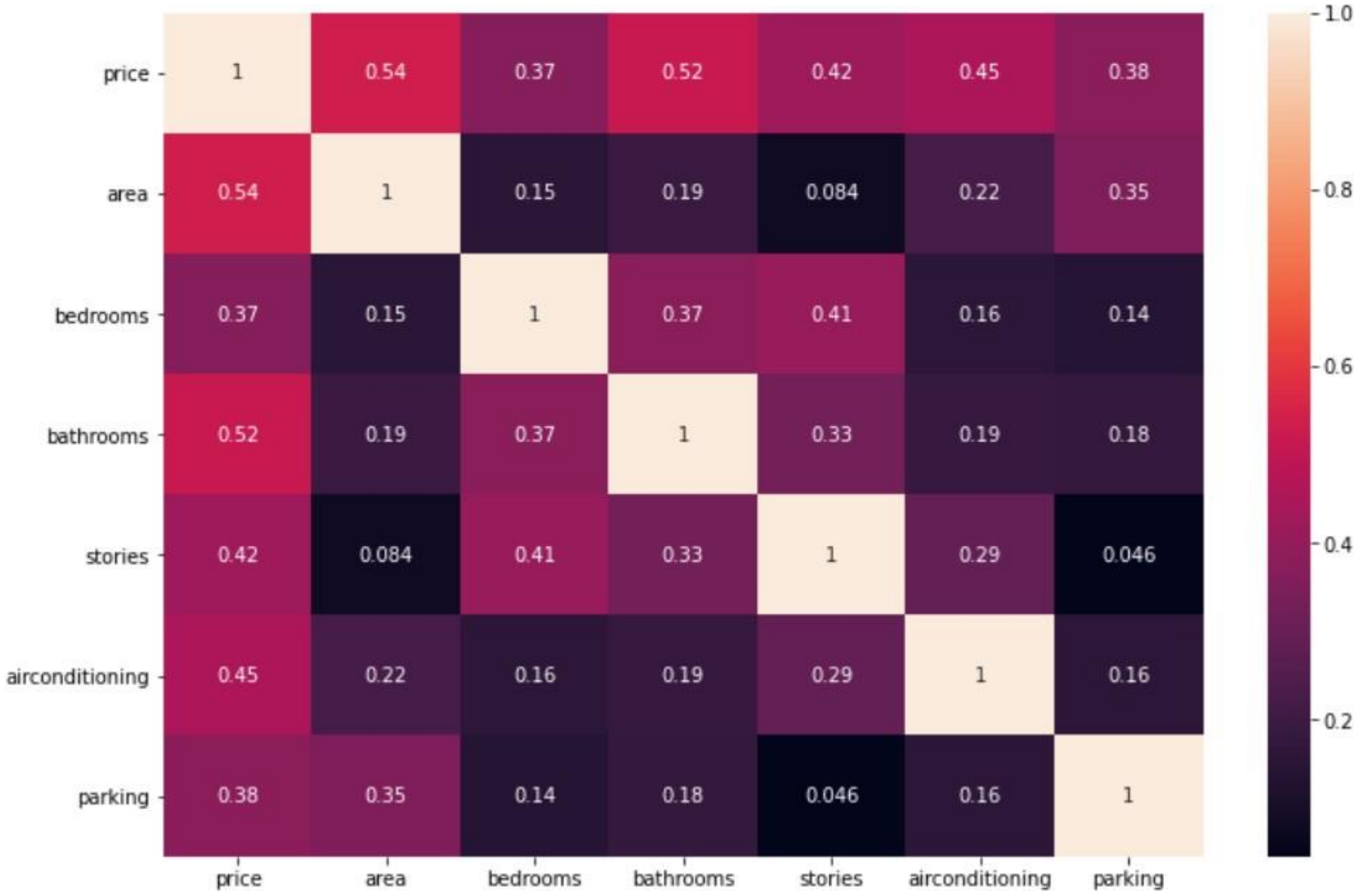
```
price               0
area                0
bedrooms            0
bathrooms           0
stories             0
mainroad            0
guestroom           0
basement            0
hotwaterheating     0
airconditioning     0
parking             0
prefarea            0
furnishingstatus    0
dtype: int64
```

As we can observe, there are no NULL values in any of the Columns so no Data Cleaning is required.

But there are string values in our dataset and the Regression Algorithms and most in-built methods work only on numerical values. So, we have to encode the string values to some numerical values. We do this by encoding 'yes' as 1 similarly no as '0' in all columns with string datatype except furnishing status.

In Furnishing Status column, we apply One Hot Encoding.

3.5 Checking the Co-relation of Price with different columns-



There are more columns but we are mostly interested with columns which are at-least 35% correlated with the Price Column.

Now we perform Interactive Data Visualisation on the above features of our dataset which helps us to understand the data better and also explore the outliners in the data.

(Please refer to the project file to see the Interactive Plots)

3. PREPARING DATA AND PREDICTING THE HOUSE PRICES –

3.1 Removing the outliners of the data– We removed the data where number of bathrooms were 4 and number of bedrooms were 6 as they were very less in number and could result in wrong prediction of the prices.

3.2 Splitting the Dataset– We first separate our target variable (‘price’) from the rest of our variables, then we split the data in the ratio of 80:20 (80% for training our model and 20% for evaluating our model) keeping the ratio of ‘airconditioning’ variable same (Stratified Split) as it is highly correlated to our target variable.

3.3 Using Regression Techniques to predict the value of House Prices–

3.1 Regression Technique – Regression Analysis is a form of predictive modelling technique which investigates the relationship between dependent variables (X) and an independent variable (y).

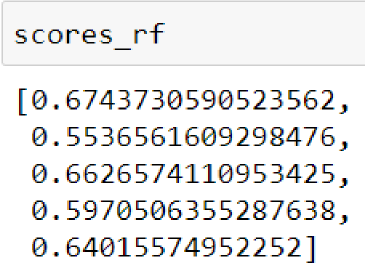
In the Project I have used **Multiple Linear Regression**, **Gradient Boosting Regressor** and **Random Forest Regressor** to predict values and evaluate our model.

3.4 Checking the goodness of our Prediction–

R-Squared Method – It is a statistical measure of how close the original data points are to the fitted regression line. It is also known as Coefficient of multiple determination.

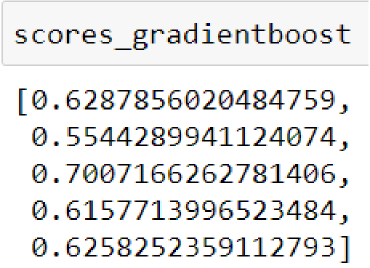
4. Final R-Squared Scores of our Models–

In the project, I have resampled the data 5 times and recorded the R-Square values of all three Regression algorithms which are as follows –



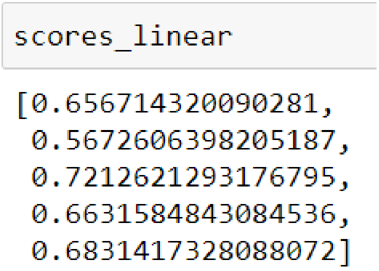
```
scores_rf
```

```
[0.6743730590523562,
 0.5536561609298476,
 0.6626574110953425,
 0.5970506355287638,
 0.64015574952252]
```



```
scores_gradientboost
```

```
[0.6287856020484759,
 0.5544289941124074,
 0.7007166262781406,
 0.6157713996523484,
 0.6258252359112793]
```



```
scores_linear
```

```
[0.656714320090281,
 0.5672606398205187,
 0.7212621293176795,
 0.6631584843084536,
 0.6831417328088072]
```

From this, we can conclude that **Multiple Linear Regression** is the best Regression Algorithm for our dataset as it has produced the maximum **R^2 score of 72%**.

5. References-

5.1 - https://pandas.pydata.org/

5.2 - https://numpy.org/

5.3 - https://matplotlib.org/

5.4 - https://seaborn.pydata.org/

5.5 - https://plotly.com/