**PROJECT REPORT ON:**

**“HOUSING: PRICE PREDICTION”**

**SUBMITTED BY**

**Ashish Kumar Rathor**

**ACKNOWLEDGEMENT**

I would like to express my special gratitude to my SME as well as the “Flip Robo Technology” team for letting me work on the “Housing Price Prediction” project also huge thanks to my academic team “Data trained”. Their suggestions and directions have helped me in doing lots of research wherein I came to know about so many new things.

**Contents:**

1. Introduction

* Business Problem Framing:
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken

1. Analytical Problem Framing

* Mathematical/ Analytical Modelling of the Problem
* Data Sources and their formats
* Data Pre-processing Done
* Data Inputs-Logic-Output Relationships
* Hardware and Software Requirements and Tools Used

1. Data Analysis and Visualization

* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Key Metrics for success in solving the problem under consideration
* Visualization
* Run and Evaluate selected models
* Interpretation of the Results

1. Conclusion

* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

1. Reference
2. **INTRODUCTION**

**Business Problem Framing:**

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors to the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help companies increase their overall revenue, and profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modeling, Market mix modeling, and recommendation systems are some of the machine learning techniques used for achieving the business goals of housing companies. Our problem is related to one such housing company.

House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improving Real Estate efficiency. The aim is to predict efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, as also upcoming developments future prices will be predicted. ... the cost of the property depends on the number of attributes considered.

Now as data scientists our work is to analyze the dataset and apply our skills toward predicting house prices.

**Conceptual Background of the Domain Problem**

The real estate market is one of the most competitive in terms of pricing and the same tends to vary significantly based on numerous factors; forecasting property price is an important module in decision-making for both the buyers and investors in supporting budget allocation, finding property-finding stratagems, and determining suitable policies.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. This company wants to know:

• Which variables are important to predict the price of a variable?

• How do these variables describe the price of the house

1. Customer retention survives when the companies can fulﬁll customer expectations and
2. additionally maintain it in long-term relationships to ensure long-term buying decisions
3. [13–15]. The topic of customer retention is argued in business economics commonly
4. within the perspective of relationship marketing, which considers customer relation-
5. ships as one of the primary concerns with the long-term objective of developing and
6. maintaining them [16–18]. Many previous studies indicated that companies should
7. always manage customer satisfaction to achieve the retention stage. According to [19]
8. “satisfaction is an overall customer attitude towards a service provider”.In[20],
9. authors added that satisfaction is an emotional reaction regarding what customers
10. expect and what they receive, including the fulﬁllment of needs and goals. Customer
11. retention states a desired outcome in the future to satisfaction, so long-term of rela-
12. tionship is demonstrated by satisfaction. Although customer satisfaction does not
13. guarantee repurchase, it still plays a vital role in ensuring customer retention. While
14. many studies on customer retention had long focused on customer satisfaction, addi-
15. tional factors are stated as an inﬂuence in customer retention, such as trust and com-
16. mitment. [21], in “The Commitment-Trust Theory of Relationship Marketing,”which
17. is the most inﬂuential Relationship Marketing, suggests that the center of successful
18. relationship marketing is the relationship of commitment and trust. They urged the
19. importance of commitment and trust that leads to build a positive correlation between
20. company and customers and encourage efﬁciency, productivity, and effectiveness. The
21. degree of trust between service provider and customer is signiﬁcantly inﬂuenced by the
22. quality of the service, which results in an effective commitment to the provider, and
23. Customer retention survives when the companies can fulﬁll customer expectations and
24. additionally maintain it in long-term relationships to ensure long-term buying decisions
25. [13–15]. The topic of customer retention is argued in business economics commonly
26. within the perspective of relationship marketing, which considers customer relation-
27. ships as one of the primary concerns with the long-term objective of developing and
28. maintaining them [16–18]. Many previous studies indicated that companies should
29. always manage customer satisfaction to achieve the retention stage. According to [19]
30. “satisfaction is an overall customer attitude towards a service provider”.In[20],
31. authors added that satisfaction is an emotional reaction regarding what customers
32. expect and what they receive, including the fulﬁllment of needs and goals. Customer
33. retention states a desired outcome in the future to satisfaction, so long-term of rela-
34. tionship is demonstrated by satisfaction. Although customer satisfaction does not
35. guarantee repurchase, it still plays a vital role in ensuring customer retention. While
36. many studies on customer retention had long focused on customer satisfaction, addi-
37. tional factors are stated as an inﬂuence in customer retention, such as trust and com-
38. mitment. [21], in “The Commitment-Trust Theory of Relationship Marketing,”which
39. is the most inﬂuential Relationship Marketing, suggests that the center of successful
40. relationship marketing is the relationship of commitment and trust. They urged the
41. importance of commitment and trust that leads to build a positive correlation between
42. company and customers and encourage efﬁciency, productivity, and effectiveness. The
43. degree of trust between service provider and customer is signiﬁcantly inﬂuenced by the
44. quality of the service, which results in an effective commitment to the provider, and

People and real estate agencies buy or sell houses, people buy to live in or as an investment and the agencies buy to run a business. Either way, we believe everyone should get exactly what they pay for. over-valuation/under-valuation in housing markets has always been an issue and there is a lack of proper detection measures. Broad measures, like house/Real-estate price-to-rent ratios, give a primary pass. However, to decide about this issue an in-depth analysis and judgment are necessary. Here’s where machine learning comes in, by training an ML model with hundreds and thousands of data a solution can be developed which will be powerful enough to predict prices accurately and can cater to everyone’s needs.Real Estate has become more than a necessity in this 21st century, it represents something much more nowadays. Not only for people looking into buying Real Estate but also the companies that sell these Estates. Real Estate Property is not only the basic need of a man but today it also represents the richness and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policymakers, and many. Investment in the real estate sector seems to be an attractive choice for investments. Thus, predicting the real estate value is an important economic inde

1. **Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem**

This particular problem has two datasets one is the training dataset and the other is the test dataset. I have built the model using the training dataset and predicted the Sale Price for the test dataset. By looking into the target column, I came to know that the entries of the Sale Price column were continuous and this was a Regression problem so I have to use all regression algorithms while building the model. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 80% null values and more than 85% zero values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. While checking the null values in the datasets I found many columns with nan values and I replaced those nan values with suitable entries like mean for numerical columns and mode for categorical columns. To get a better insight into the features I have used plotting like distribution plot, and bar plot. With this plotting, I was able to understand the relation between the features in a better manner. Also, I found outliers and skewness in the dataset so I removed outliers using the percentile method and I removed skewness using the yeo-Johnson (default). I used all the regression models while building the model then tuned the best model and saved the best model. At last, I predicted the sale price for the test dataset using the saved model of the training dataset.

**Data Sources and their formats**

The data was collected for my internship company – Flip Robo technologies in CSV (comma-separated values) format.

Also, I was having two datasets one is trained, and the other tests. I have built the model using the training dataset and predicted the Sale Price for the test dataset. My train dataset was having 1168 rows and 81 columns including the target, and my test dataset were having 292 rows and 80 columns excluding the target. In this particular dataset, I have an object, float, and integer types of data. I can merge these two datasets and perform my analysis, but I have not done that because of a data leakage issue. This is how my datasets look for me when I import those datasets into my Jupiter separately.

**Data Pre-processing Done**

* As a first step I imported the required libraries and I imported both datasets separately which were in CSV format.
* Then I did all the statistical analysis like checking shape, info about the dataset, etc.
* While checking the info of the datasets I found some columns with more than 80% null values, so these columns will create skewness in datasets so I decided to drop those columns.
* Then while looking into the dataset info I found some columns with more than 85% zero values this also creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 85% zero values.
* While checking for null values I found null values in most of the columns and I used **the imputation method** to replace those null values **(mode for categorical** column and **mean/median for numerical** columns).
* In the Id and utility columns the unique counts were 1168 and 1 respectively, which means all the entries in the Id column are unique and ID is the identification number given for a particular asset and all the entries in the Utilities column were the same so these two columns will not help us in model building. So, I decided to drop those columns.
* And all these steps were performed to both train and test datasets separately and simultaneously.

**Data Inputs- Logic- Output Relationships**

* I have used a bar plot for each pair of categorical features that shows the relation with the median sale price for all the subcategories in each categorical feature.
* And also, for continuous numerical variables I have used a scatter plot to show the relationship between the continuous numerical variable and the target variable.
* I found that there is a linear relationship between continuous numerical variables and Sale Price.

**Hardware and Software Requirements and Tools Used**

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

**Hardware required**: -

1. Processor — core i5 and above

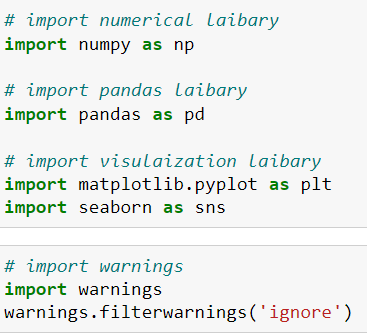
2. RAM — 8 GB or above

3. SSD — 512GB or above

**Software/s required**: -

1.Anaconda

**Libraries** **required :-**



* as follows:
* **import pandas as pd**: **pandas** is a popular Python-based data analysis toolkit that can be imported using import pandas as pd. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a NumPy matrix array. This makes pandas a trusted ally in data science and machine learning.
* **import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms basic linear algebra, basic statistical operations, random simulation and much more.
* **import seaborn as sns:** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in the exploration and understanding of data.
* **Import matplotlib. pyplot as plt:** matplotlib. pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some changes to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
* from sklearn.preprocessing import LabelEncoder
* from sklearn.preprocessing import StandardScaler
* from statsmodels.stats.outliers\_influence import variance\_inflation\_factor
* from sklearn.ensemble import RandomForestRegressor
* from sklearn.ensemble import AdaBoostRegressor
* from sklearn.ensemble import GradientBoostingRegressor
* from sklearn.tree import DecisionTreeRegressor
* from sklearn.svm import SVR
* from sklearn.neighbors import KNeighborsRegressor
* from sklearn.metrics import r2\_score,mean\_absolute\_error,mean\_squared\_error,root\_mean\_squared\_error
* from sklearn.model\_selection import cross\_val\_score
* from sklearn.model\_selection import KFold
* From sklearn.model\_selection import GridSearchCV

With these sufficient libraries, we can go ahead with our analytical skills.

**Data Analysis and Visualization**

**Identification of possible problem-solving approaches (methods)**

I have used the imputation method to replace null values. To remove outliers, I have used the percentile method. And to remove skewness I have used the yeo-Johnson method. To encode the categorical columns, I have to use Ordinal Encoding. Use of Pearson’s correlation coefficient to check the correlation between dependent and independent features. Also, I have used standardization. Then followed by model building with all regression algorithms.

**Testing of Identified Approaches (Algorithms)**

Since the Sale price was my target and it was a continuous column so this particular problem was a regression problem. And I have used all regression algorithms to build my model. By looking into the difference between the r2 score and the cross-validation score I found **RANDOM FOREST REGRESSOR** as the best model with the least difference. Also, to get the best model we have to run through multiple models and to avoid the confusion of overfitting we have to go through cross-validation. Below is the list of regression algorithms I have used in my project.

* RandomForestRegressor
* AdaBoostRegressor
* GradientBoostingRegressor
* DecisionTreeRegressor
* SupportVectorRegressor
* KNeighborsRegressor
* LineaRregression

**Key Metrics for success in solving the problem under consideration**

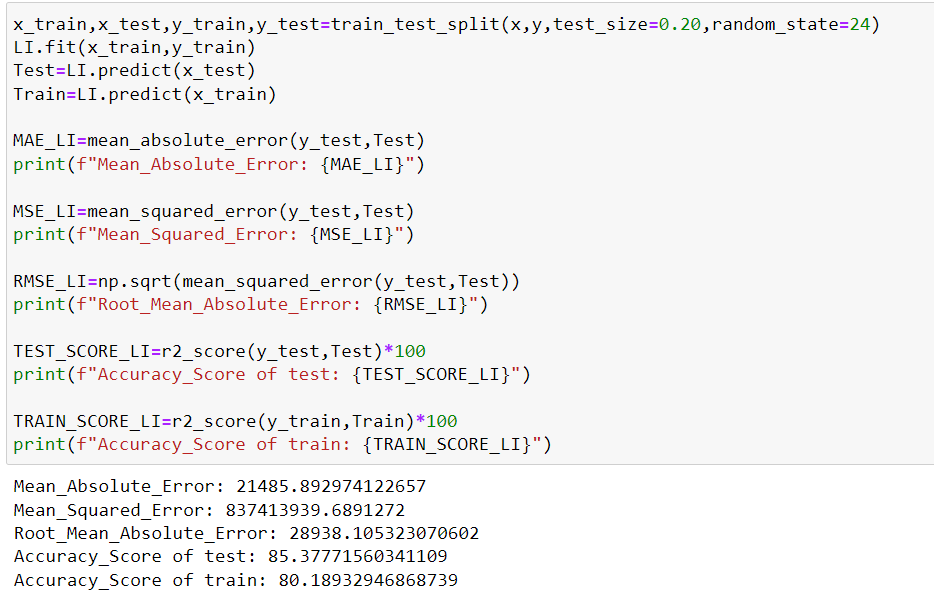
I have used the following metrics for evaluation:

* I have used mean absolute error which gives the magnitude of difference between the prediction of observation and the true value of that observation.
* I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.
* I have used the r2 score which tells us how accurate our model is.

s **Run and Evaluate selected models**

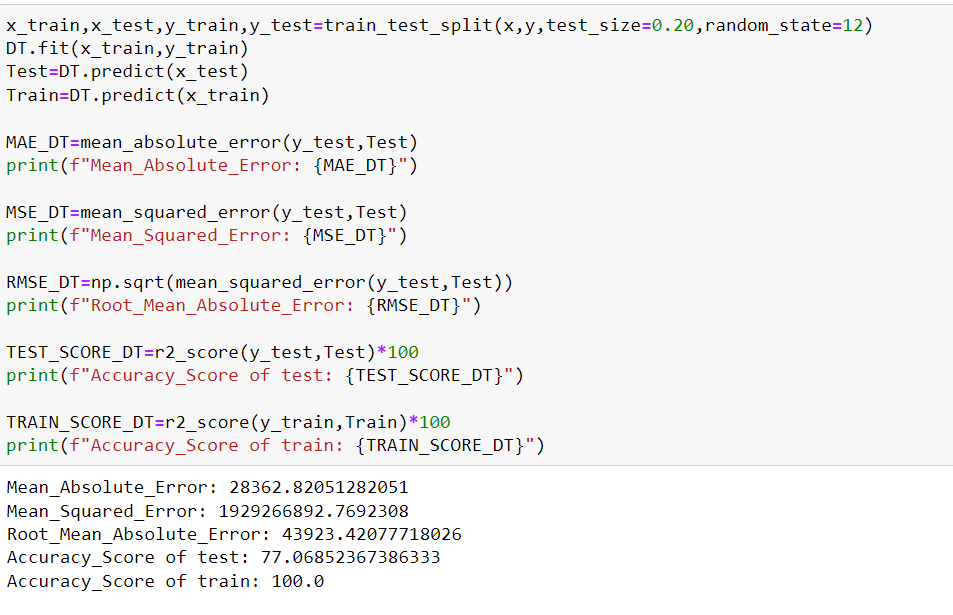
**1. Model Building:**

**A) LINEAR REGRESSION**



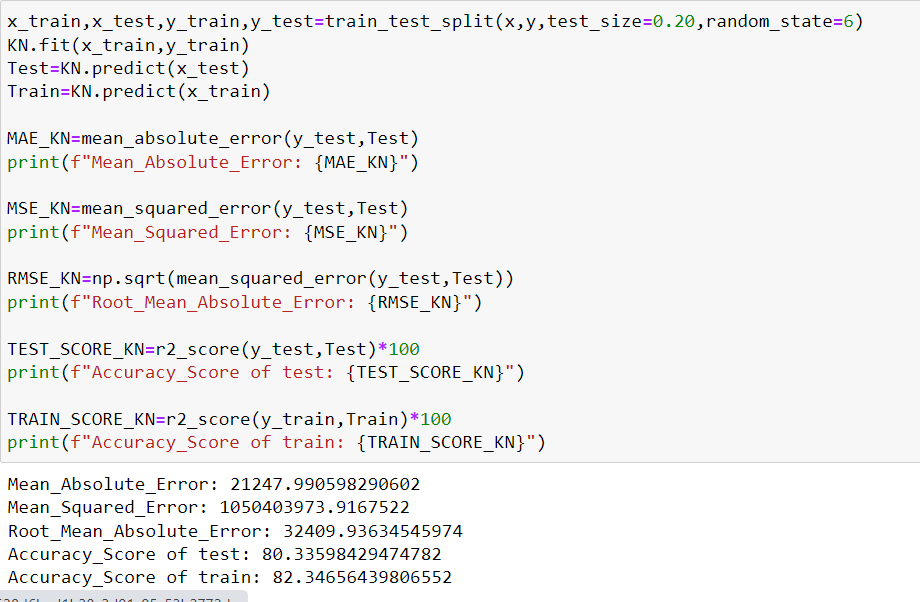
* Linear Regressor has given me 85.37% accuracy but still we have to look into multiple models.

**B) DECISION TREE REGRESSOR**



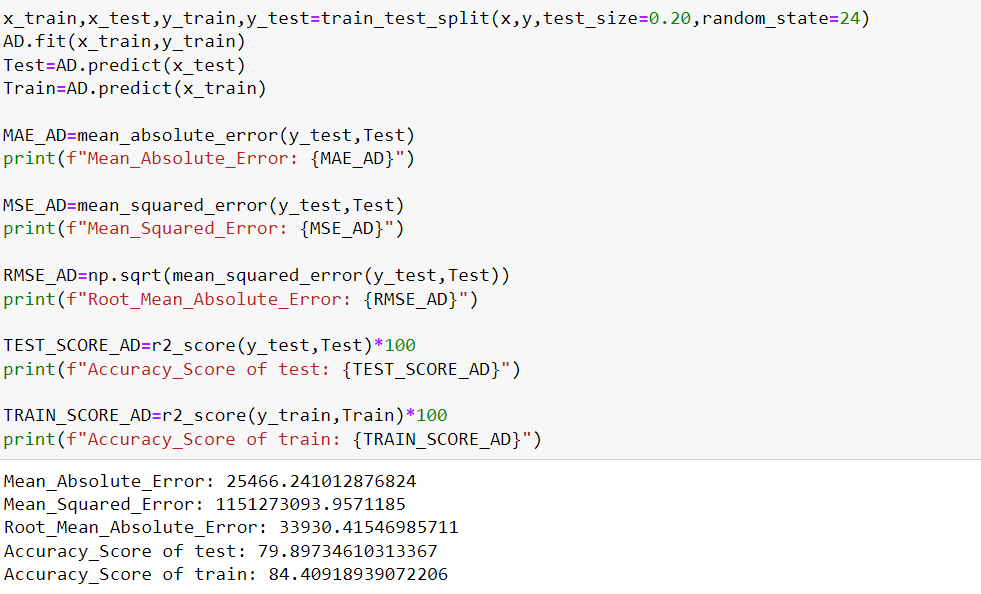
* DecionTreeRegressor is giving me 77.06% accuracy.So I go for other model for finding best model.

1. **KNEIGHBORS REGRESSOR:**

****

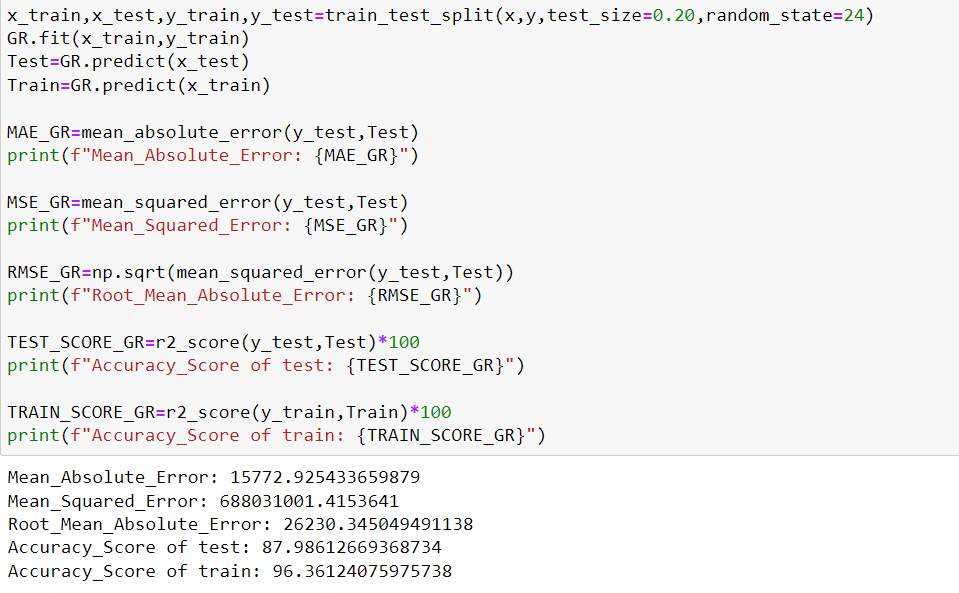
● KneighborsRegressor is giving me 80.33% accuracy

1. **ADA BOOST REGRESSOR**

****

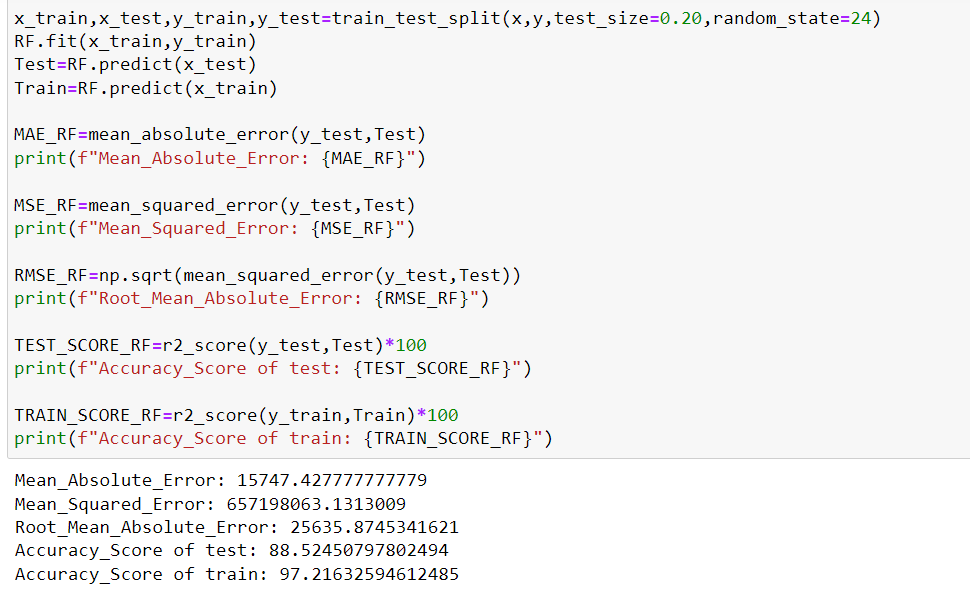
* AdaBoostRegressor is giving me 79.89% accuracy.

1. **GRADIENT BOOSTING REGRESSOR**

****

**GradientBoostingRegressor is giving me 87.98%**

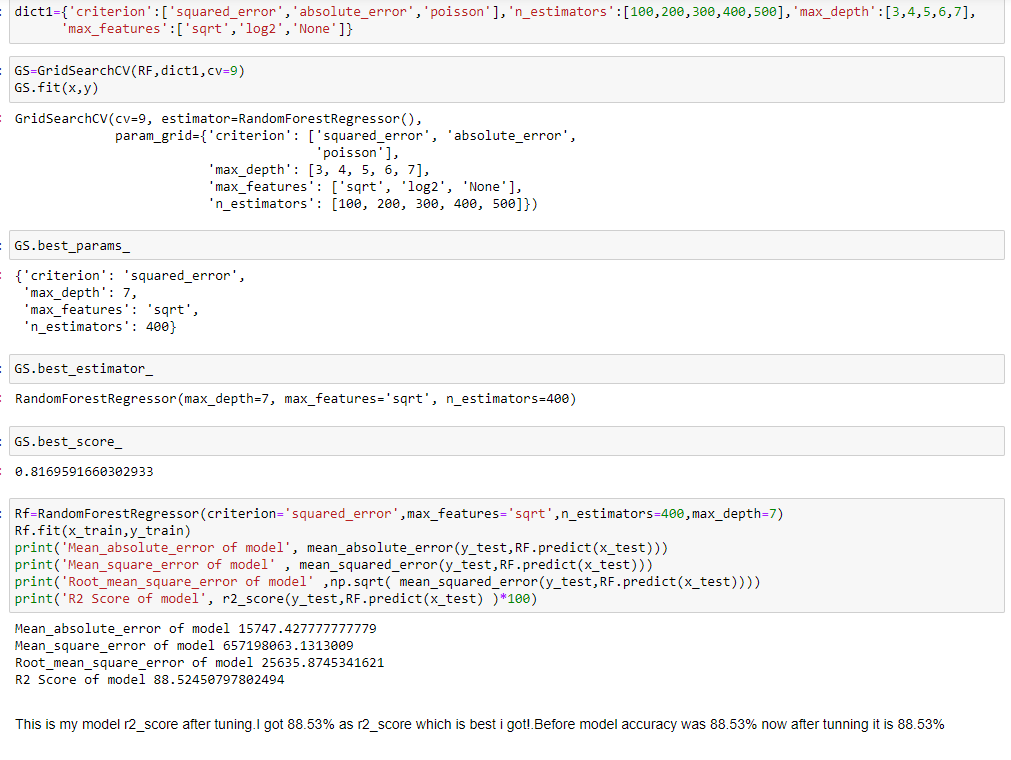
1. **RANDOM FOREST REGRESSOR**

****

RandomForestRegressor is giving me 88.52% accuracy.

* **By looking into the difference of model accuracy and cross validation score I found Random Forest Regressor as the best model.**

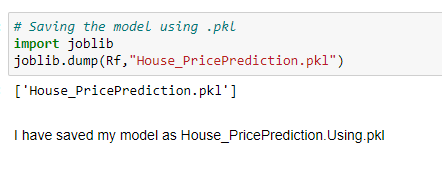
**2. Hyper Parameter Tunning:**

****

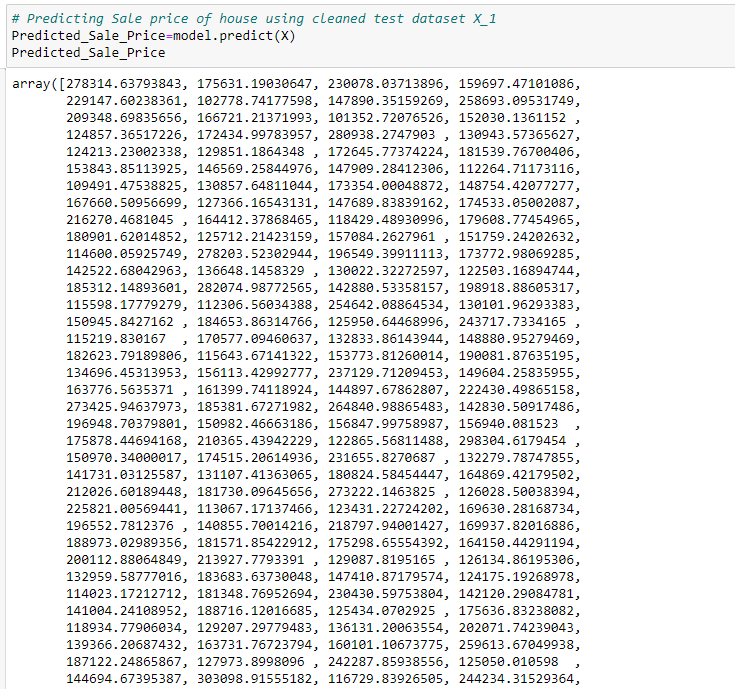
* **I have choosed all parameters of RandomForestRegressor, after tunning the model with best parameters I have increased my model accuracy from 88.52% to 88.54%. Also mse and rmse values has reduced which means error has reduced**

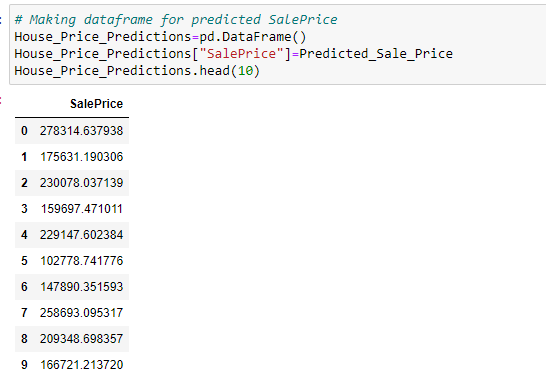
**3. Saving the model and Predicting SalePrice for test data:**

* I have saved my best model using .pkl as follows**.**

****

**Predicting SalePrice for test dataset:**

****

****

* **I have predicted the Sale Price for test dataset using saved model of train dataset, and the predictions look good. I have also saved my predictions for further analysis.**

**Interpretation of the Results**

* This dataset was very special as it had a separate train and test datasets. We have to work with both datasets simultaneously.
* Firstly, the datasets were having null values and zero entries in maximum columns so we have to be careful while going through the statistical analysis of the datasets.
* And proper plotting for the proper type of features will help us to get a better insight into the data. I found maximum numerical continuous columns were in a linear relationship with the target column.
* I notice a huge number of outliers and skewness in the data so we have to choose proper methods to deal with the outliers and skewness. If we ignore these outliers and skewness, we may end up with a bad model which has less accuracy.
* Then scaling both train and test dataset has a good impact like it will help the model not to get biased.
* We have to use multiple models while building model using train dataset as to get the best model out of it.
* And we have to use multiple metrics like Mae, Mse, Rmse and r2\_score which will help us to decide the best model.
* I found RandomForestRegressor as the best model with 88.54% r2\_score. Also, I have improved the accuracy of the best model by running hyper parameter tunning.
* At last, I have predicted the Sale Price for test dataset using saved model of train dataset. It was good!! that I was able to get the predictions near to actual values.

**4.CONCLUSION**

**Key Findings and Conclusions of the Study**

In this project report, we have used machine learning algorithms to predict the house prices. We have mentioned the step-by-step procedure to analyse the dataset and finding the correlation between the features. Thus, we can select the features which are not correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a csv file was generated consisting of predicted house prices. Hence, we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the DataFrame of predicted prices of test dataset.

**Learning Outcomes of the Study in respect of Data Science**

I found that the dataset was quite interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove missing value and to replace null value and zero values with there respective mean, median or mode. This study is an exploratory attempt to use five machine learning algorithms in estimating housing prices, and then compare their results.

To conclude, the application of machine learning in property research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices. Future direction of research may consider incorporating additional property transaction data from a larger geographical location with more features, or analysing other property types beyond housing development.

**Limitations of this work and Scope for Future Work**

* First draw back is the data leakage when we merge both train and test datasets.
* Followed by more number of outliers and skewness these two will reduce our model accuracy.
* Also, we have tried best to deal with outliers, skewness, null values and zero values. So it looks quite good that we have achieved a accuracy of 88.54% even after dealing all these drawbacks.
* Also, this study will not cover all regression algorithms instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones.
* **This model doesn't predict future prices** of the houses mentioned by the customer. Due to this, the risk in investment in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process.

**5.References**

1. <https://www.diva-portal.org/smash/get/diva2:1456610/FULLTEXT01.pdf>
2. <https://ieeexplore.ieee.org/document/8882834/references#references>
3. <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>
4. <https://www.kaggle.com/anmolkumar/house-price-prediction-challenge/tasks?taskId=2304>