

**HOUSING PRICE PREDICTION**

Submitted by:

**ABHISHEK JAIN**

**ACKNOWLEDGMENT**

I have taken the use of internet and some websites to study and analyze this dataset and done research on it.

* [Kaggle: Your Machine Learning and Data Science Community](https://www.kaggle.com/)
* [ResearchGate | Find and share research](https://www.researchgate.net/)

**INTRODUCTION**

* Business Problem Framing

Houses are one of the necessary need 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 in 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 the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

* Conceptual Background of the Domain Problem

Basic concepts of machine learning and python is required for solving this problem.

* Review of Literature

The two datasets are provided which are train.csv and test.csv. All the data is taken from these two datasets only.

* Motivation for the Problem Undertaken

House price prediction can help**determine the selling price of a house** and can help to arrange the right time to purchase a house. There are various factors that influence the price of a house which include physical conditions, concept and location.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

For doing analysis of this project , I have kept in mind many things about that how the problem can be visualized and represented in a simple way so I firstly I have imported all the necessary libraries required for data cleaning and viewing .

Secondly I have done analysis of each column separately so that I can find out what exactly each column of this dataset is pointing towards and what is the correlation between them.

Thirdly, I have checked that if any null value is present in the dataset or not as it can effect our model processing.

Finally, after all the data processing and EDA steps I have use various models for Machine Learning and selected the best model afterwards.

And in the last I saved the best predicted model to be available for further uses.

* Data Sources and their formats

The data sources which were provided to me were in form of csv files so it was easy for me to import it and then do further calculations on it.

The two csv files used in the project are train.csv and test.csv

* Data Pre-processing Done

There were many steps used in Data cleaning / pre-processing of data in this project by me.

1. Finding out null data in dataset
2. Dropping the columns which have large amount of null data
3. Identified the columns which are having only categorical data or numeric data and applied encoding methods to them.
4. Tried to visualize each and specific column using matplotlib library and tried to find a correlation between all.
5. Finding out the outliers and removing them which can effect the performance of our model.
6. Removing the skewness after working on outliers.
7. In the last using standardization techniques to further clean data and then saving it for machine learning processes.

* Data Inputs- Logic- Output Relationships

When I used the dataset train.csv and applied EDA methods on it and found out the correlation between each column , I saw that some columns are highly correlated to each other while some are not. Some columns have many null values because of which they need to be dropped while I have to changed some of the categorical columns into continuous columns so that EDA steps can be performed on it.

In the beginning we had 1168 rows and 81 columns.

After all the EDA steps –

removing null values, outlier removal, standardization of data

we get the final shape of dataset as 1168 rows and 195 columns.

Then ,splitting the dataset into two parts i.e independent(input) and dependent(output) variables we get new dimension of data as (1168,194) as x and (1168,1) as y to be used for **machine learning**.

So we have 1 column for output data and rest columns for input data to be used for ML.

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

No assumptions were made by me in this project for model prediction.

* Hardware and Software Requirements and Tools Used

Python 3.8 or above is used.

Jupyter notebook for coding work.

Libraries which are used in this project –

* **Pandas** for importing data.
* **Numpy** used in EDA steps.
* **Matplotlib** and **Seaborn** for visualization of data.
* **Sklearn** for standardization techniques, Machine learning.
* **Warning** for ignoring warnings in our code.

**Model/s Development and Evaluation**

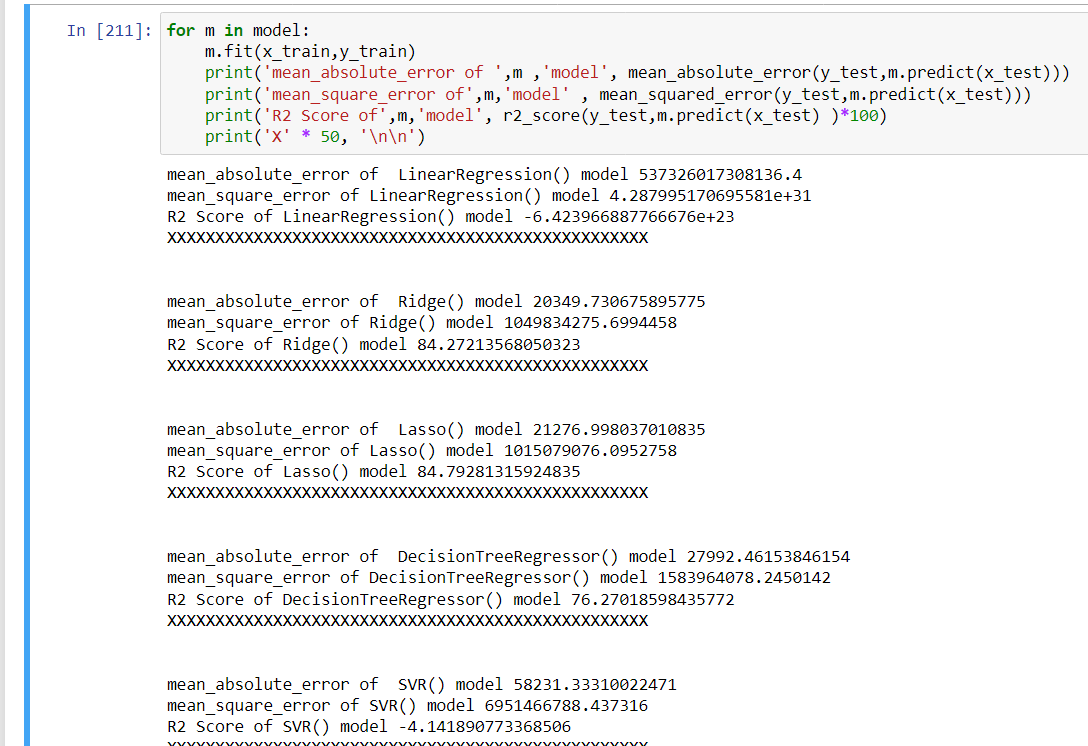
* Identification of possible problem-solving approaches (methods)

Used **Regression model** as have both dependent and independent variables in our dataset.

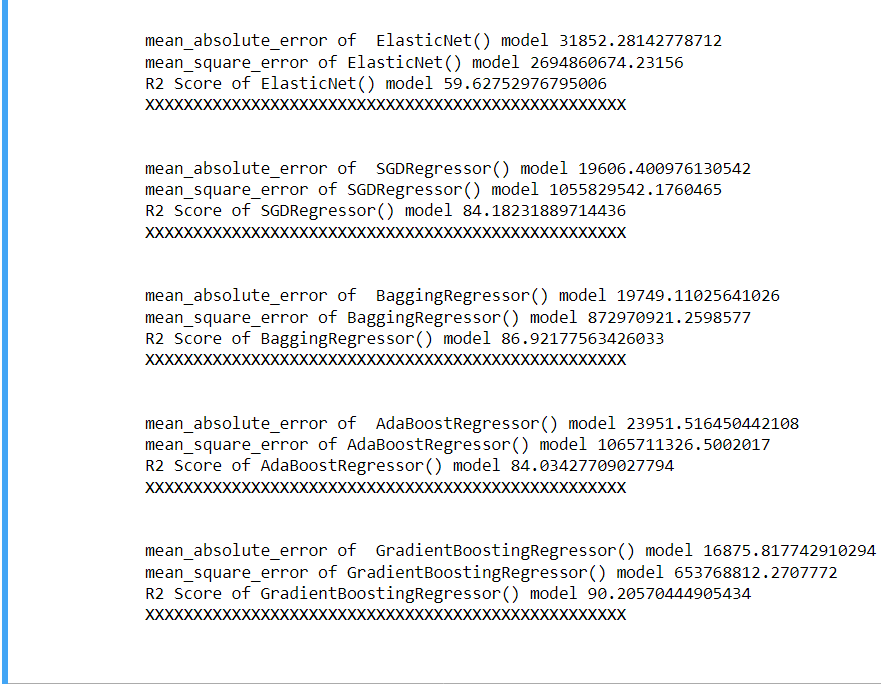
* Testing of Identified Approaches (Algorithms)
* Linear Regression
* Ridge Regression
* Lasso Regression
* SVR Regressor
* KNeighbors Regressor
* RandomForest Regressor
* AdaBoost Regressor
* GradientBoosting Regressor
* Run and Evaluate selected models

**All the screenshots are shown below in which we can see all the different types of models used for machine learning.**

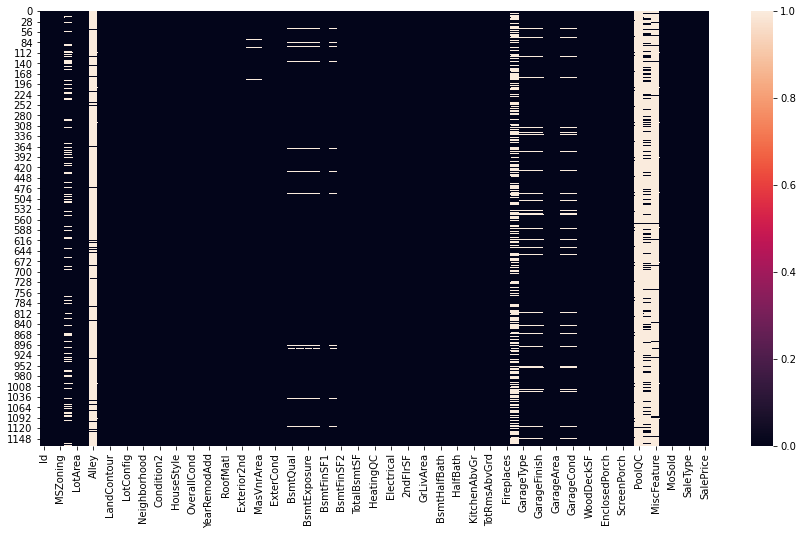




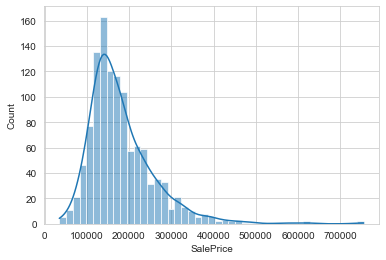




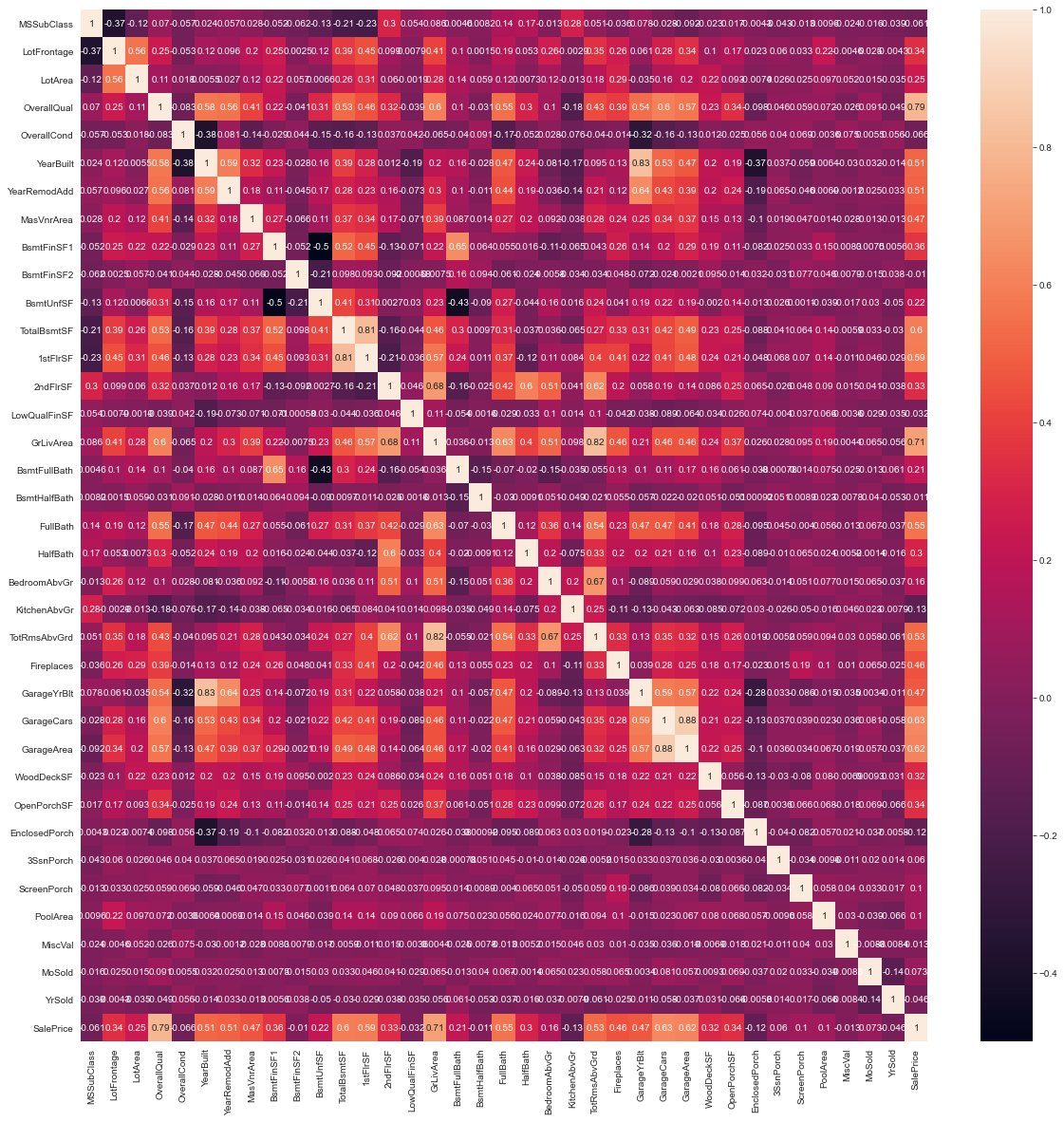
* Visualizations



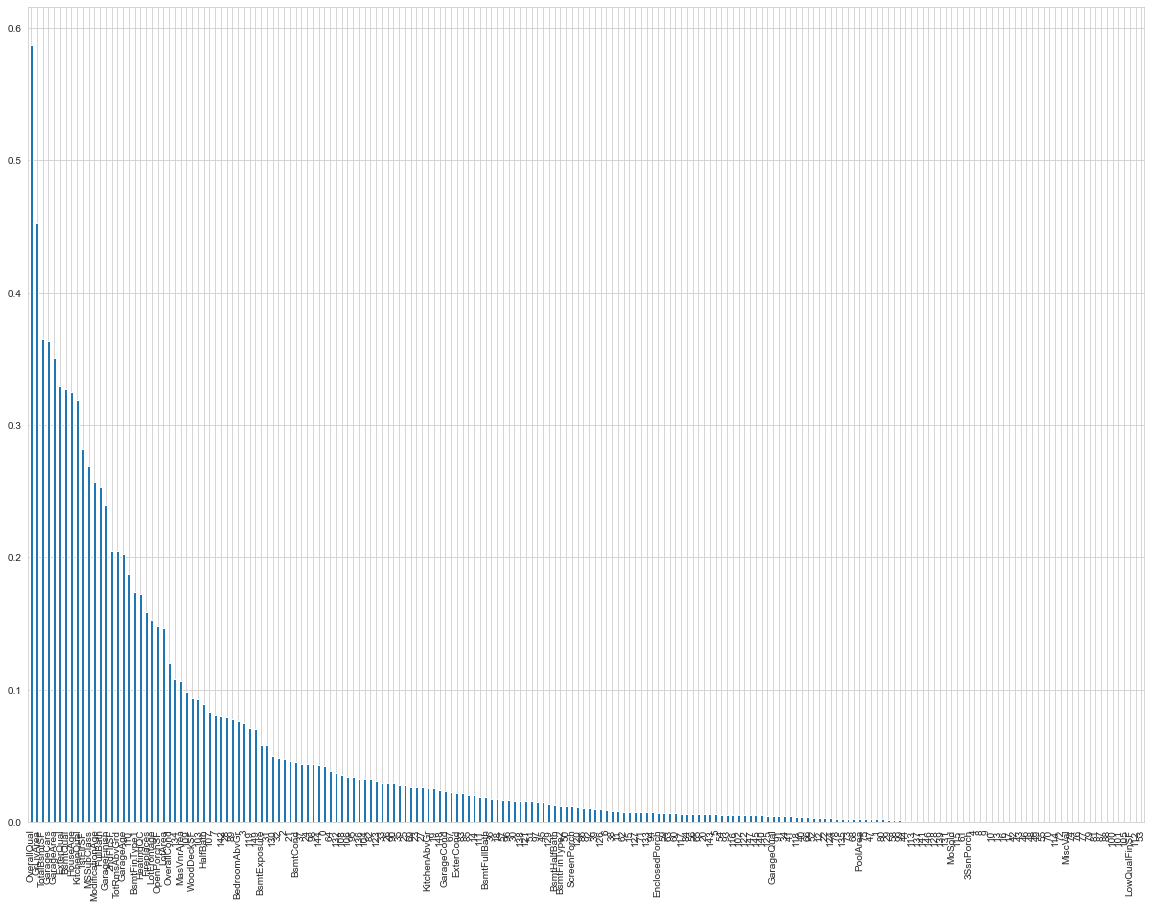
Null values Graph



Histoplot showing salesprice



Correlation graph



Feature Selection Graph showing all values

* Interpretation of the Results

After preprocessing data the final shape of dataset is

1168 rows

195 columns

Target variable is salesprice(y) – 1 column and 1168 rows

Other columns are input variables (x) – 194 columns and 1164 rows

**Data modelling**

Amongst all the models used Ridge model is the best . So we use it for hyperparameter tuning using GridSearchCV.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Using train.csv data the final shape of our dataset is 1168 rows and columns size is 195

But, when we used test.csv the rows size is 293 and columns size is reduced to 147 after feature selection.

We have to use regression model for machine learning as the output variable has continuous data. We have done hypertuning of our model also to improve accuracy ,but sometimes it does not work i.e hypertuning can reduce the accuracy of our model.



* Limitations of this work and Scope for Future Work
* The dataset provided to us is very small in size. l so our predictions can be wrong when working with a large dataset.
* There were large amount of null values in our dataset which was necessary to filter. If no null values were present, then we may have a chance to improve the accuracy of our model and there should be no necessity to drop many columns then.