

**Housing: Price Prediction**

Submitted by:

Abhilash Ashok

**ACKNOWLEDGMENT**

Foremost, I would like to express my sincere gratitude to my teacher Mr ShankarGouda Tegginmani for the sincere teaching of my Post Graduation Diploma in Data Science and Artificial Intelligence, for his patience, motivation, enthusiasm and immense knowledge.

Beside my teacher, I would like to thank the entire team of DataTrained, for their immense support for me round the clock, for their encouragement, insightful comments and all-time support.

My sincere thanks also go to Flip Robo Technologies for offering me the Internship opportunities in their groups and leading me working on diverse exciting projects. I would also like to thank Khusboo Garg for walking me through the internship.

I would also thank the DataTrained website for which I can access all the live classes and recorded classes for my reference when I am stuck anywhere.

I would like to thank W3schools.com and the YouTube channel as well to see some of the concepts that could be a add on.

Last but not the least I would like to thank my family and friends for supporting my journey entire out during the tough and at the ease times in my life.

**INTRODUCTION**

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 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

We are required to model the price of houses with the available independent variables.

This model will then be used by the management to understand how exactly the prices vary with the variables.

They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

Further, the model will be a good way for the management to understand the pricing dynamics of a new market

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

When we look at the dataset it is a vast dataset having all the information which is described with its equivalent. However on the overviewed it is found the data is not of uniform types, it is having all sort of categories like integers, float, strings etc.

Also, we see some of the features/label are being abbreviated with the values associated to the each, so need to identify and diagnose the same.

* Data Sources and their formats

When I look at the dataset, it is having all the variants along with some abbreviated features and labels. The data doesn’t seem of to be the standard type, we see all variants in the dataset.

I also see some of the features are not much necessary however it has just been added since its an import of the raw data.

I also see some of the features columns are having lot of null values which can be eliminated as it does not make anything effective.

* Data Preprocessing Done

Data cleaning is one of the most important step before we proceed anything further.

Firstly I would check for the information of the dataset where in which I get all the information like which types of data are present in the dataset.

Next, I would like to look into if at all any dataset is being missing, if so I would treat those missing dataset.

I would next choose which features are required and which has to be eliminated.

* Hardware and Software Requirements and Tools Used

1. Jupyter Notebook
2. Libraries : Numpy,Pandas,seaborn,matplotlib,etc

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The dataset we have received is of two different data, i.e. Train data and the test data

It means to say we need to perform the operation on the training data find the efficiency and when I treat with my test data it would show how well my data has gone through the model to execute

After my data is cleaned, I try to visualize the problem through graph for the better understanding or through numeric statistics how well my features and label are anodized inside the model.

I also would check for the multicollinearity problem that is existed, if so I would treat accordingly

* Testing of Identified Approaches (Algorithms)

Looking up on the Label, here we see it is SalePrice which is a continuous data.

I tried to train the data and then merge in the test data.

After the test data was being merged, I scaled in for the required algorithm, I did try

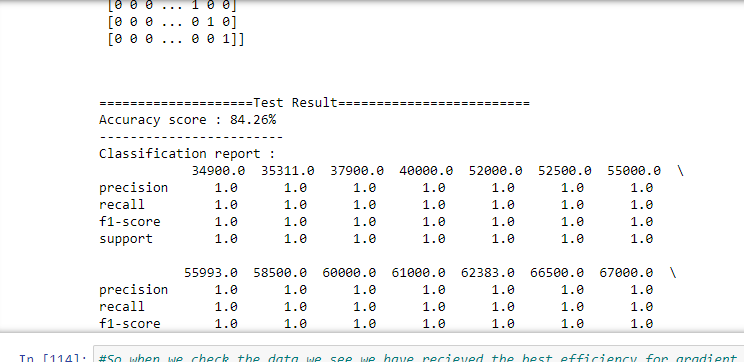
Listing down all the algorithms used for the training and testing.

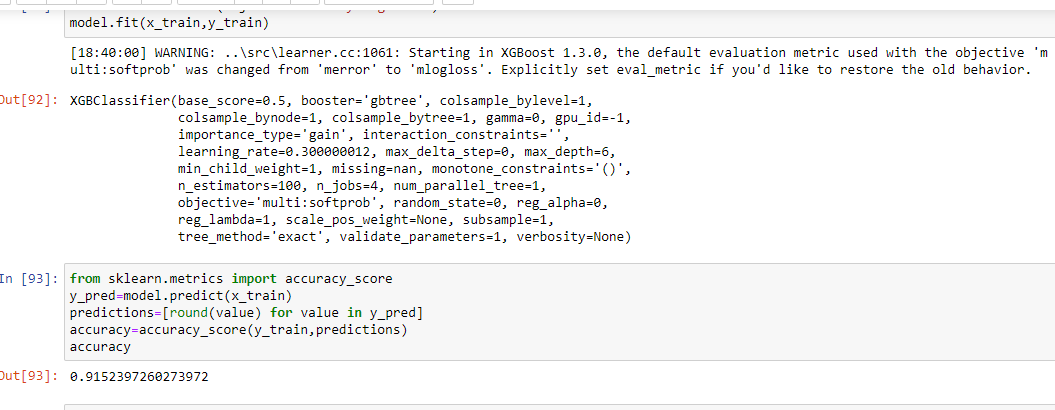
I performed different model on algorithms to check for which the one would give the best efficiency.

Run and Evaluate selected models

I have used XGB Regressor/Classifier,Random Forrest Classifier,Gradient Boosting Classifier,I check for the efficiency to see which one would give more,also I have used Logistic regression just to show, it cannot be much used on the continuous data.

However I see for the XGB regressor and Gradient Boosting Technique algorithm have given the best efficiency.





**Thank You**