Name of the project

[HOUSING PROJECT]

SUBMITTED BY:

SHIVAM SHARMA

INTRODUCTION

Problem Statement:

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. 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 company is looking at prospective properties to buy houses to enter the market. We 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. For this company wants to know:

Business Goal:

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.

Motivation for the Problem Undertaken

- **Which variables are important to predict the price of variable?**

- **How do these variables describe the price of the house?**

Data Sources and their format

- **Train.csv file downloaded from PMT**
- **Test.csv file downloaded from PMT**

Data Preprocessing and EDA

The steps followed for the cleaning and EDA of the data:-

1) First we used Drop function for delete some columns from the dataset which having some IDS and maximum Null values (ex- ID, Alley, PoolQC, Miscfeature, Fence)
2) Then we filled all the NANs in the datset columns which having Nans by the means of their mean(if any continuous data) and by the most occurring element in that column(in the categorical data)
3) Now seprate the data which having continous data put in a different dataframe i.e data_con
4) The data which having Object datatype put in a different dataframe i.e – data_obje
5) Now after filling all the NANs we will encode the Object Datatype from the data frame(data)
6) Now we will check the correlation between the independent variables and between the independent variables and dependent variables.
7) Plot a Heatmap to check clearly the correlations between all
8) We drop the columns which are highly correlated between the independents variables
9) Plot some regression plots between the Label and the Highly positively and negatively correlated variables
10) Plot some swarm and catplots for check the datapoints between the labels and the variables
11) Now we are going to check the skewness for the continuous data columns by the help of Distribution plots
12) Now we remove the skewness from the columns which are not in the range by the means of Log ,cbrt,sqrt transformation
13) Again see the distribution of the data by plotting the distribution plot after removing the skewness

14) As we removed the skewness from the continuous data so that we can now easily conact the dataframes again to make a new dataframe(datas)

15) Now our new dataframe is skewed free ,null free,standardise and ready for the model building

16) Seprate the output variable from the dataframe and train the data for the precition

Algorithm used in this project:-

1. Linear Regression
2. Lasso
3. Random forest Regressor
4. Adaboost Regressor
5. Gradient boosting Regressor

BEST Algo. From these And why?

Gradient boosting Regressor is the best algo from all of these algo which is used in this data to predict because the difference between the cross val score and the accuracy score is min for the GBDT algo AND IT ALSO GIVE BETTER ACCURACY(approx. 90%)that's why we USED THIS **Algo.**

- **Lets hypertune this algo with the help of Grid searchCV**
- **Save the model for later predictions**

 # Now my model is ready to predict the dataset which were provided in this project(df – Dataframe which going to be tested with this model)

CONCLUSIONS

THE DATAFRAME CONTAINS THE DATA WHICH RELATED TO THE PROPERTY SO THAT WE CAN PREDICT THE SALE PRICE OF THAT PROPERTY BU THE HELP OF THE INDEPENDENT VARIABLES

KEYFINDING:-

WE DON'T CHECK THE OUTLIERS IN THIS DATASET BECAUSE SOME POINTS MIGHT BE AN OUTLIERS BUT WE CANT SAY BECAUSE IT IS RELATED TO THE REAL TIME PROPERTY SO THAT EVERY PROPERTY MAY VARY THEIR AREAS AND TYPE OF MATERIALS AND ETC THAT'S Y WE DID NOT GO TO REMOVE THE OUTLIERS FROM THIS DATA