**HOUSE PRICE PREDICTION USING MACHINE LEARNING**

**SB3001-NAAN MUDHALVAN**

|  |  |
| --- | --- |
| Date | 01 NOVEMBER 2023 |
| Name | M.ABDUL BASITH |
| Team ID | PROJ\_212169\_TEAM\_1 |
| Project Name | HOUSE PRICE PREDICTION USING ML |
| Maximum Marks |  |

|  |  |  |
| --- | --- | --- |
| **S.No** | **Content** | **PageNo** |
| **1** | **Index** | **2** |
| **2** | **Introduction** | **3** |
| **3** | **Abstract** | **3** |
| **4** | **Literature Survey** | **4-5** |
| **5** | **Problem Definition,designing&Objective** | **6-7** |
| **6** | **Import Dataset Data Cleaning&Analysis** | **7-10** |
| **7** | **Data Vizuvalization** | **11-13** |
| **8** | **Model Development & Evaluvation** | **14-17** |
| **9** | **Code Sample** | **18-21** |
| **10** | **Output Screenshots** | **23-24** |
| **11** | **Conclusion,futurework&reference** | **24-25** |

**2.Introduction :**

Abstract

House price forecasting is an important topic of real estate. The literature attempts to

derive useful knowledge from historical data of property markets. Machine learning

techniques are applied to analyze historical property transactions in India to discover

useful models for house buyers and sellers. Revealed is the high discrepancy between

house prices in the most expensive and most affordable suburbs in the city of Mumbai.

Moreover, experiments demonstrate that the Multiple Linear Regression that is based on

mean squared error measurement is a competitive approach.

Abstract

House price forecasting is an important topic of real estate. The literature attempts to

derive useful knowledge from historical data of property markets. Machine learning

techniques are applied to analyze historical property transactions in India to discover

useful models for house buyers and sellers. Revealed is the high discrepancy between

house prices in the most expensive and most affordable suburbs in the city of Mumbai.

Moreover, experiments demonstrate that the Multiple Linear Regression that is based on

mean squared error measurement is a competitive approach.

* Sure, I can help you with that. Here is a brief introduction to house price prediction using Python:
* House price prediction is a popular application of machine learning. It involves training a model on a dataset of house prices and their features, such as the number of bedrooms, the square footage, and the location. Once trained, the model can be used to predict the price of a new house based on its features.
* Python is a popular programming language for machine learning and data science. There are many libraries available in Python that make it easy to build machine learning models for house price prediction. Some of these libraries include NumPy, pandas, scikit-learn, and TensorFlow.

Abstract

House price forecasting is an important topic of real estate. The literature attempts to

derive useful knowledge from historical data of property markets. Machine learning

techniques are applied to analyze historical property transactions in India to discover

useful models for house buyers and sellers. Revealed is the high discrepancy between

house prices in the most expensive and most affordable suburbs in the city of Mumbai.

Moreover, experiments demonstrate that the Multiple Linear Regression that is based on

mean squared error measurement is a competitive approach.

Abstract

House price forecasting is an important topic of real estate. The literature attempts to

derive useful knowledge from historical data of property markets. Machine learning

techniques are applied to analyze historical property transactions in India to discover

useful models for house buyers and sellers. Revealed is the high discrepancy between

house prices in the most expensive and most affordable suburbs in the city of Mumbai.

Moreover, experiments demonstrate that the Multiple Linear Regression that is based on

mean squared error measurement is a competitive approach.

**3.Abstract:**

House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in India to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Mumbai. Moreover, experiments demonstrate that the Multiple Linear Regression that is based on mean squared error measurement is acompetitiveapproach.

4.Literature Survey

1. Introduction

House is one of human life's most essential needs, along with other fundamental needs such as food, water, andmuch more. Demand for houses grew rapidly over the years as people's living standards improved. While there arepeople who make their house as an investment and property, yet most people around the world are buying a house as their shelter or as their livelihood.

2. Attributes

* House price prediction can be divided into two categories, first by focusing on house characteristics, and secondly by focusing on the model used in house price prediction. Many researchers have produced a house price prediction model, including [1, 3, 6–8].
* A research undertaken by [9] analyses the existing housing price in Jakarta, Indonesia using the conceptual model and questionnaires. Based on the results, the attributes or factors affecting the house price differ for each house construction in Jakarta, therefore accepting the validity of this analysis as the main purpose of this research is to classify the factor or attributes affecting the house price. Various considerations influence the price of a house. According to [10], the factors influencing house prices can be classified into three categories: location, structural and neighborhood condition.

A. Locational

Location is considered to be the most significant feature of house price determination [6, 9–11]. [12] in his studyalso observed the significant of location attributes in deciding house price. The location of the property was classified ina fixed locational attribute. All of these studies point to the close association between locational attributes such asdistance from the closest shopping center, or position offering views of hills or shore, and house price variations.

B. Structural

Another significant feature influencing the house price is structural structure or some research has listed it asphysical attributes [10, 13]. Structural characteristic is a feature that people may identify, whether number of bedroomsand bathrooms, or floor space, or garage and patio. These structural attributes, often offered by house builders ordevelopers to attract potential buyers, therefore meet the potential buyers' wishes. In [14] in his earlier study, structuralattributes would be the key consideration for house hunters in determining what to purchase as such attributes representtheir market value. In their earlier study, [15] stated that all these attributes have a positive relationship to rising house prices [16].

C. Neighbourhood

Neighborhood qualities can be included in deciding house price. According to [13], efficiency of public education, community social status and proximity to shopping malls typically improve the worth of a property. There is a substantial rise in house prices from the fifth-class suburban community to affluent neighborhood as predicted [16]. Nonetheless, [13] study found that these qualities tend to be cultural based, as they are not similarly relevant in all cultures

3. Machine Learning Model

According to [20], the paradigm of evaluating the house demand can be classified into two classes which are the traditional method and the advanced valuation method. The traditional valuation scheme, including multiple regression method and stepwise regression process, whilst hedonic pricing tool, artificial neural network (ANN) and spatial analysis framework are advances valuation method. The model selection to be used to predict house price is quite critical as varieties of models are available. One of the most commonly utilized models in this research field is Regression Analysis which is used in many studies, including [3, 10, 21]. Another common model for house price predictions is the Support Vector Regression (SVR) [7, 22, 23].

4. Related Work

A total of 14 articles were reviewed and evaluated to capture all attributes that influences the price of house. [3] in his article stated that square footage of unit of a house is the most importance variable in predicting price of a house, followed by number of bathrooms and number of bedrooms. Apart from that, the study suggests that the worth of the house increases by 2.6% if the floor space of the house is raised by 100 square feet.

5.Conclusions

This paper examined and analyzed the current research on the significant attributes of house price and analyzed the data mining techniques used to predict house price. Technically, houses with a strategic location such as the accessibility to shopping mall or other facilities tend to be more expensive than houses in rural areas with limited numbers of facilities.

The accurate prediction model would allow investors or house buyers to determine the realistic price of a house as well as the house developers to decide the affordable house price. This paper addressed the attributes used by previous researchers to forecast a house price using various prediction models. Taken together, the results of the survey have shown the potential of SVR, ANN and XGBoost in predicting house prices. These models were developed based on several input attributes and they work significantly positive with house price. In conclusion, the impact of this research was intended to help and assist other researchers in developing a real model which can easily and accurately predict house prices. Further work on a real model needs to be done with the utilization of our findings to confirm them.

5.Problem Statement

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price.

•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. This study is an exploratory attempt to use three machine learning algorithms in estimating housing prices, and then compare their results.

•In this study, our models are trained with 18-year of housing property data utilising stochastic gradient descent (SGD) based support vector regression (SVM), random forest (RF) and gradient boosting machine (GBM). We have demonstrated that advanced machine learning algorithms can achieve very accurate prediction of property prices, as evaluated by the performance metrics. Given our dataset used in this paper, our main conclusion is that RF and GBM are able to generate comparably accurate price estimations with lower prediction errors, compared with the SVM results.

•First, our study has shown that advanced machine learning algorithms like SVM, RF and GBM, are promising tools for property researchers to use in housing price predictions. However, we must be cautious that these machine learning tools also have their own limitations. There are often many potential features for researchers to choose and include in the models so that a very careful feature selection is essential.

•Second, many conventional estimation methods produce reasonably good estimates of the coefficients that unveil the relationship between output variable and predictor variables. These methods are intended to explain the real-world phenomena and to make predictions, respectively. They are used for developing and testing theories to perform causal explanation, prediction, and description (Shmuell, Citation2010). Based on these estimates, investigators can interpret the results and make policy recommendations. However, machine learning algorithms are often not developed to achieve these purposes. Although machine learning can produce model predictions with tremendously low errors, the estimated coefficients (or weights, in machine learning terminology) derived by the models may sometimes make it hard for interpretation.

•Third, the computation of machine learning algorithms often takes much longer time than conventional methods such as hedonic pricing model. The choice of algorithm depends on consideration of a number of factors such as the size of the data set, computing power of the equipment, and the availability of waiting time for the results. We recommend property valuers and researchers to use SVM for making forecasts if speed is a primary concern. When predictive accuracy is a key objective, RF and GBM should be considered instead.

•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.

**Designing**

**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. There are three factors that influence the price of a house which include physical conditions, concept and location.**

**Objective**

* Predict the house price
* Using two different models in terms of minimizing the difference between predicted and actual rating

**6. Data Analysis (EDA)**

First, Let’s import the data and have a look to see what kind of data we are dealing with:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline

# To clean the dataset we need to find a missing values and remove null values , find the outliers fill the correct values by using fillna() method

#We need to import the pandas library to work with the spreadsheet-like data enabling fast loading , aligning, manipulating, and merging , in addition to other key functions.

#We need to import the numpy library to working with numerical values as it makes it easy to apply mathematical functions.

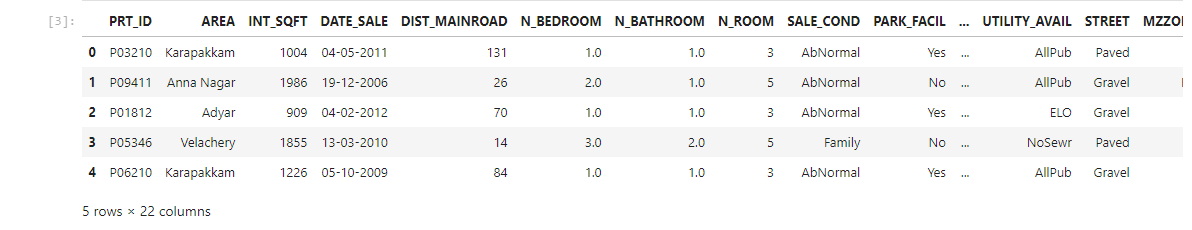
#We need to import the matplotlib.pyplot to use the pyplot functions. Matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc

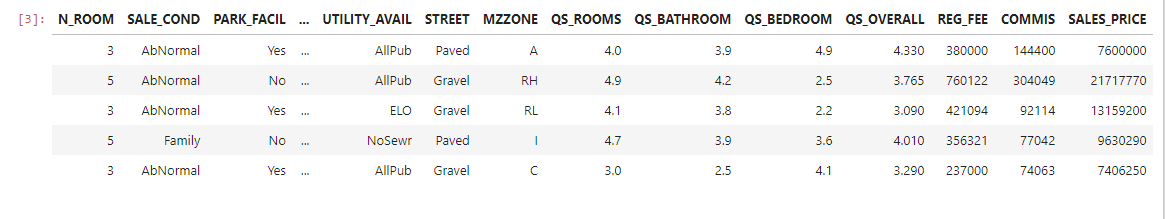
#import dataset

df=pd.read\_csv('train-chennai-sale.csv')

#The .read\_csv() function takes a path to a CSV file and reads the data into a Pandas DataFrame object.

df.head(5)





Let’s plot couple of features to get a better feel of the data

plt.figure(figsize=(20,20))

plt.subplot(321)

plt.boxplot(df["INT\_SQFT"])

plt.title("INT\_SQFT BOX PLOT")

plt.subplot(322)

plt.boxplot(df["DIST\_MAINROAD"])

plt.title("DIST\_MAINROAD BOXPLOT")

plt.subplot(323)

plt.boxplot(df["AGE\_OF\_BUILDING"])

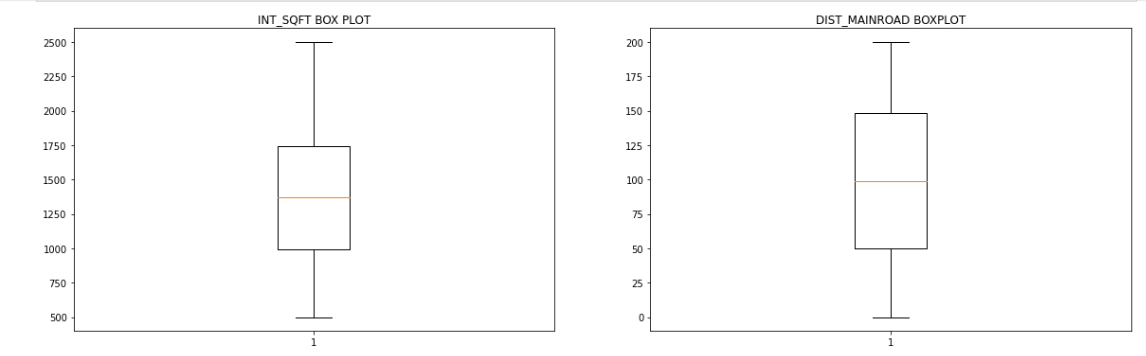
plt.title("AGE\_OF\_BUILDING BOXPLOT")

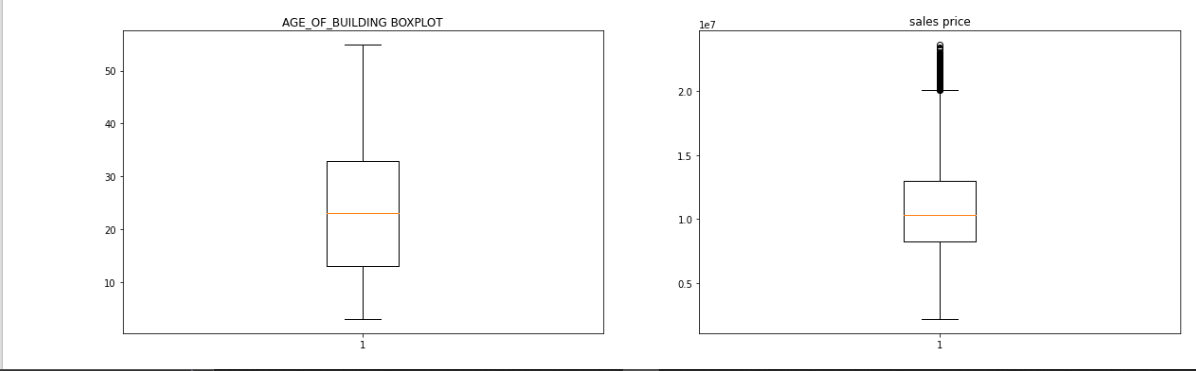
plt.subplot(324)

plt.boxplot(df["SALES\_PRICE"])

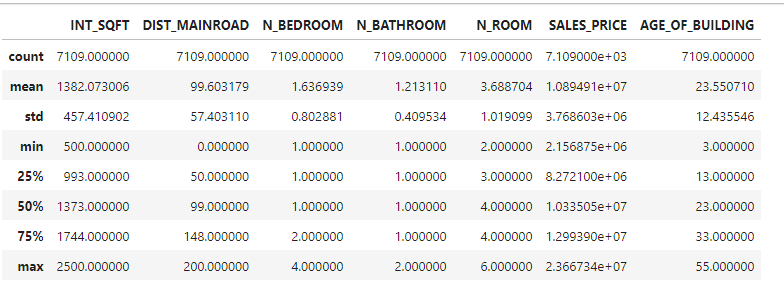
plt.title("sales price")

plt.show()





df.describe()



df["SALES\_PRICE"].max()

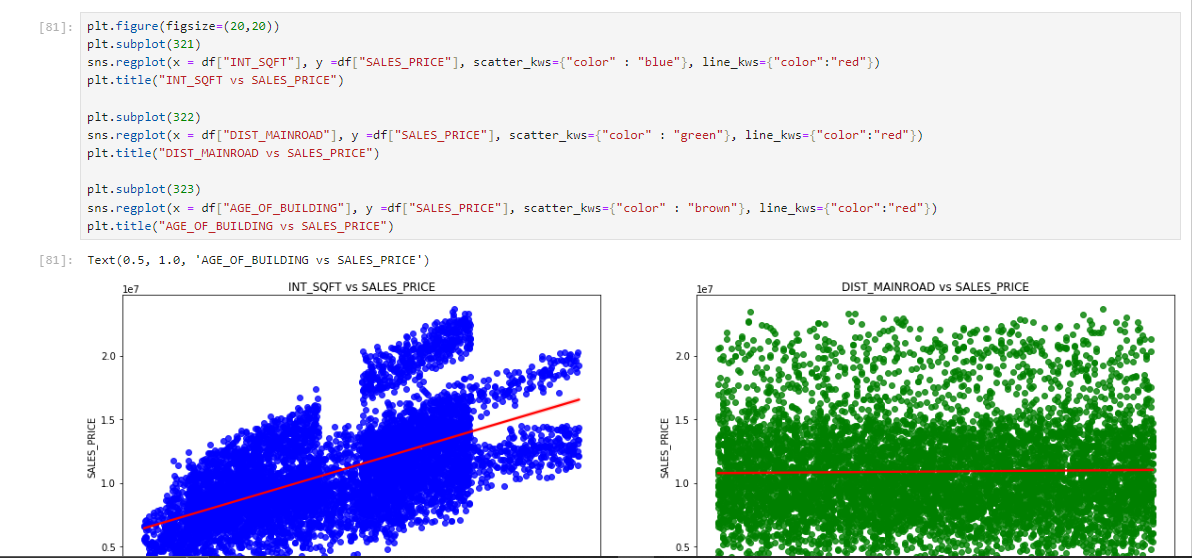
23667340

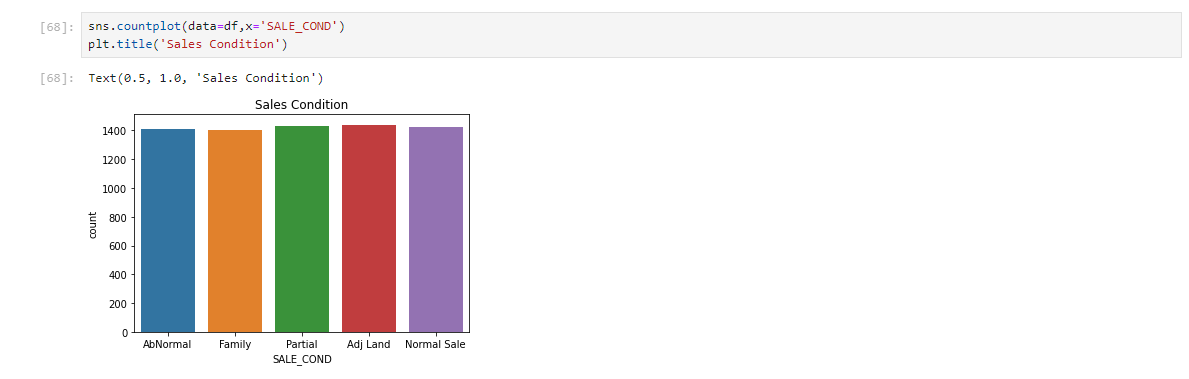
df.SALES\_PRICE.mean()

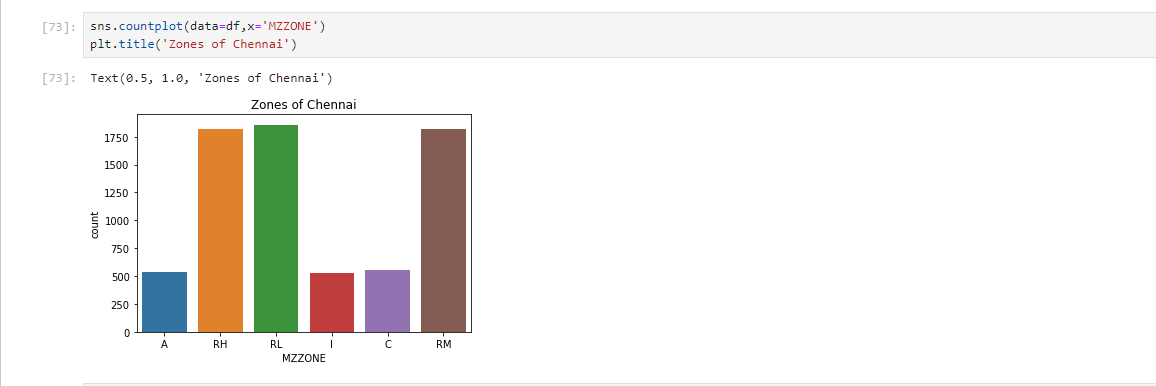
10894909.63918976

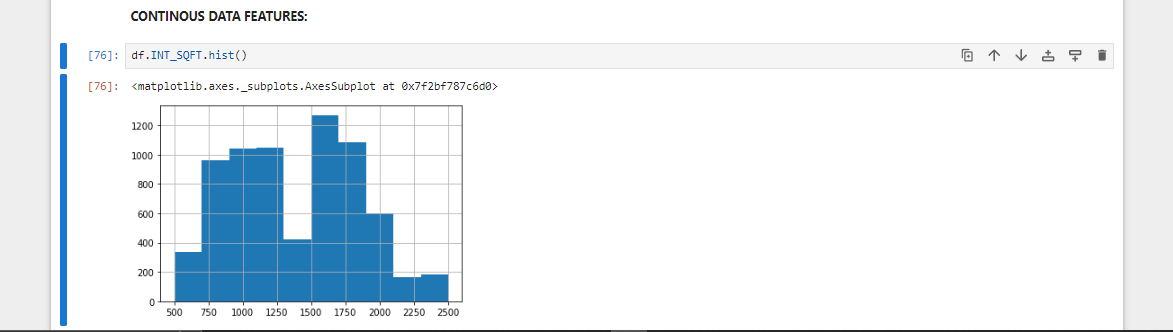
df.to\_csv('cleaned\_data')

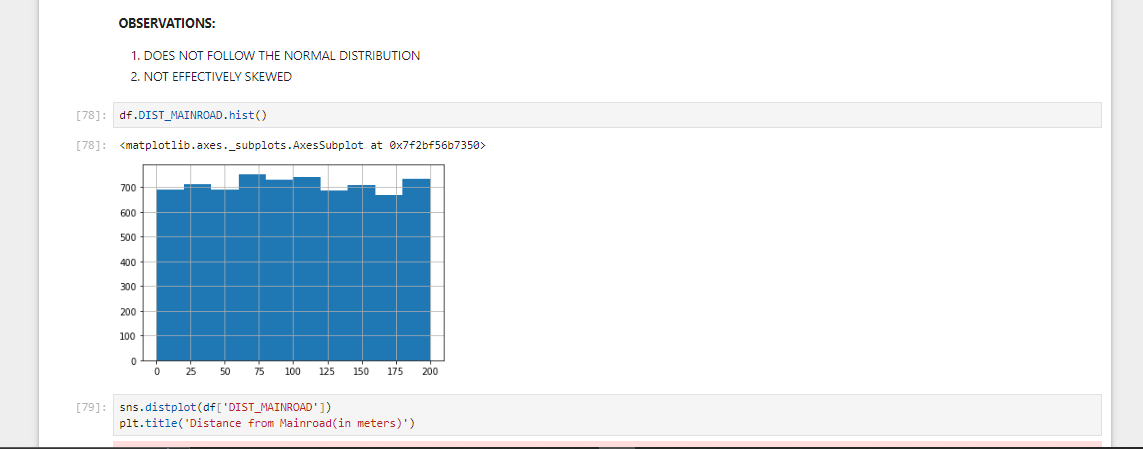
7.Data Vizuvalization:



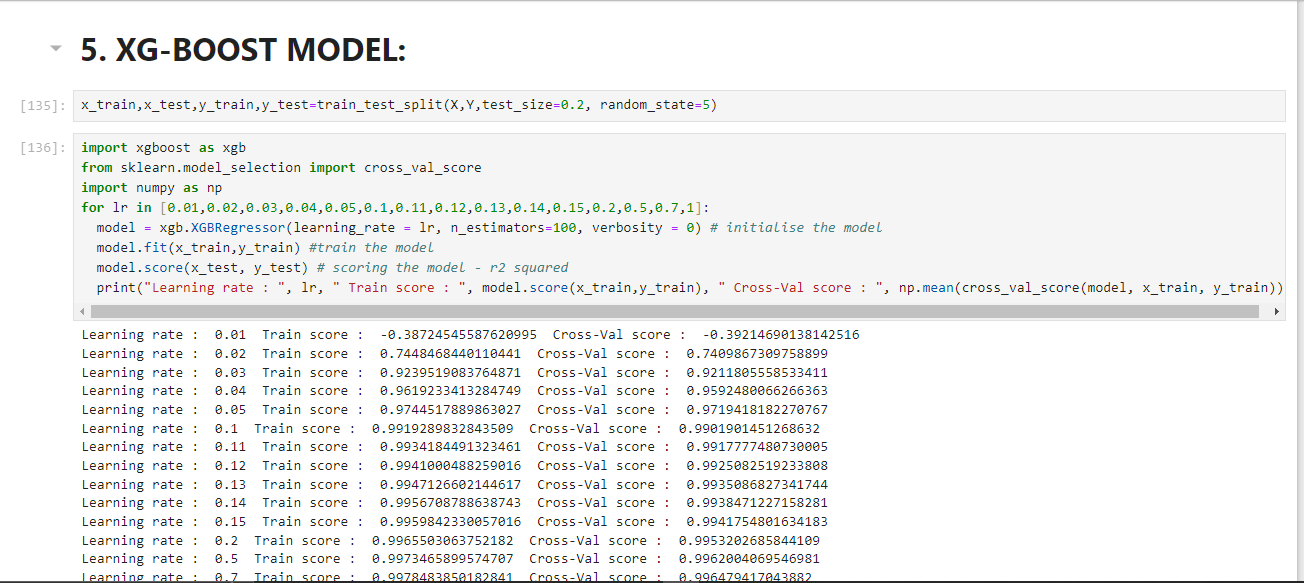




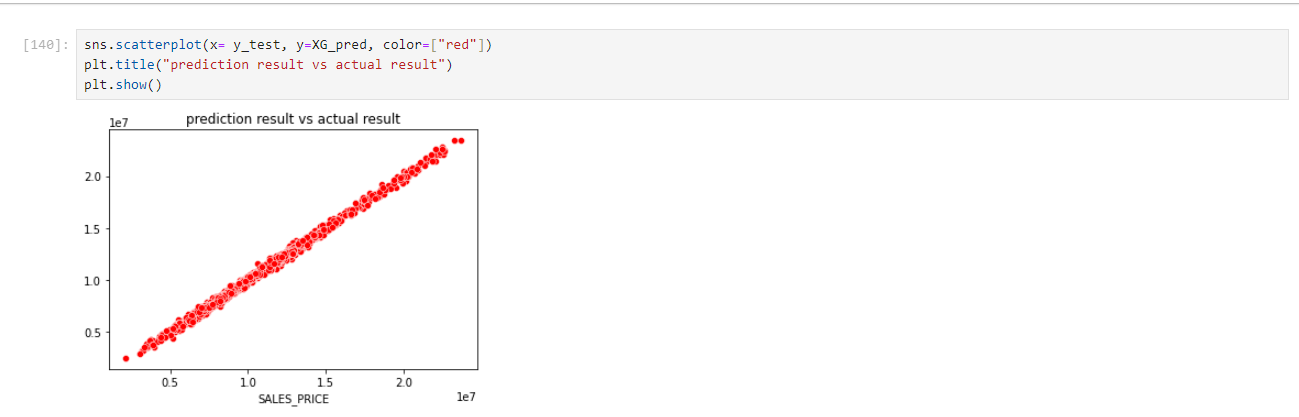












8.Model Development & Evaluvation:

1.LINEAR REGRESSION:

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

2. K-NEAREST NEIGHBOUR MODEL:

The K-Nearest Neighbor (KNN) algorithm is a popular machine learning t technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values. During the training phase, the KNN algorithm stores the entire training dataset as a reference.

3. DECISION TREE MODEL:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

4. RANDOM FOREST MODEL:

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

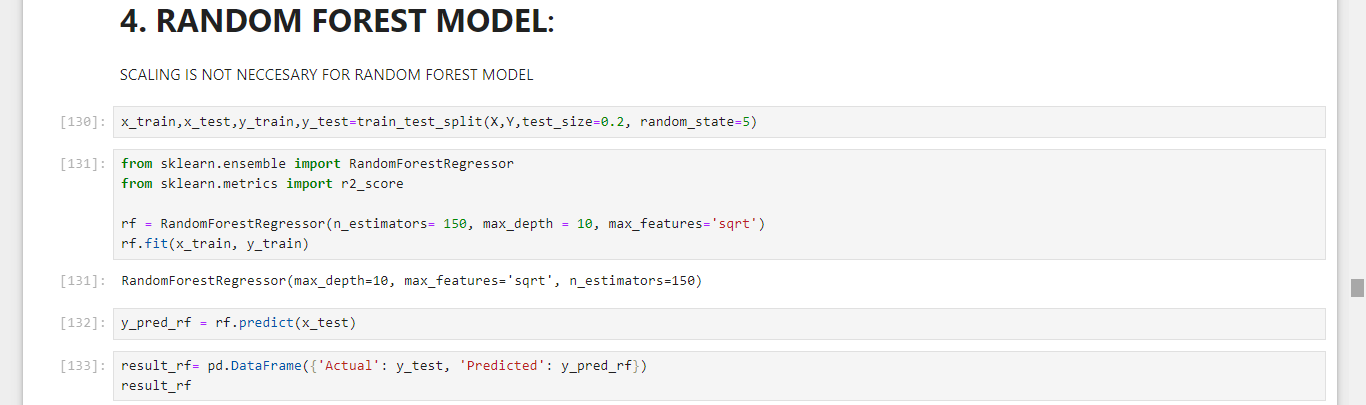
5. **XGBoost**:

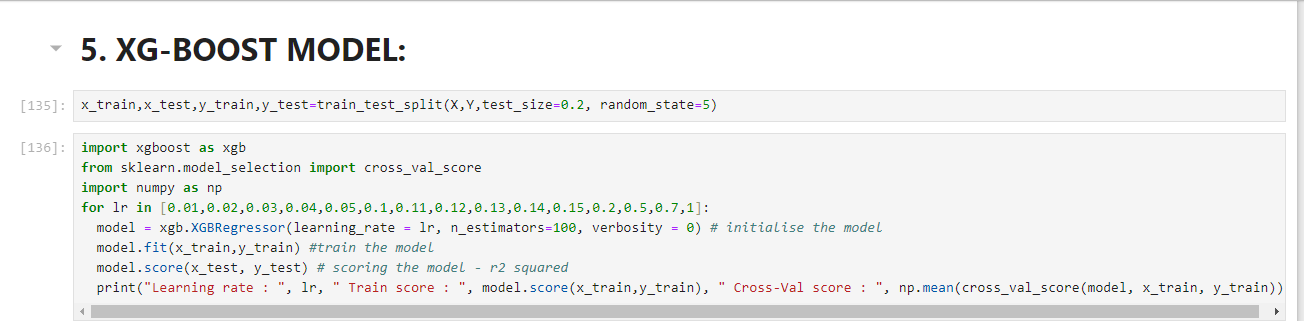
XGBoostis an optimized distributed gradient boosting library designed to be highly efficient**,**flexible and portable**.** It implements machine learning algorithms under the Gradient Boosting framework.

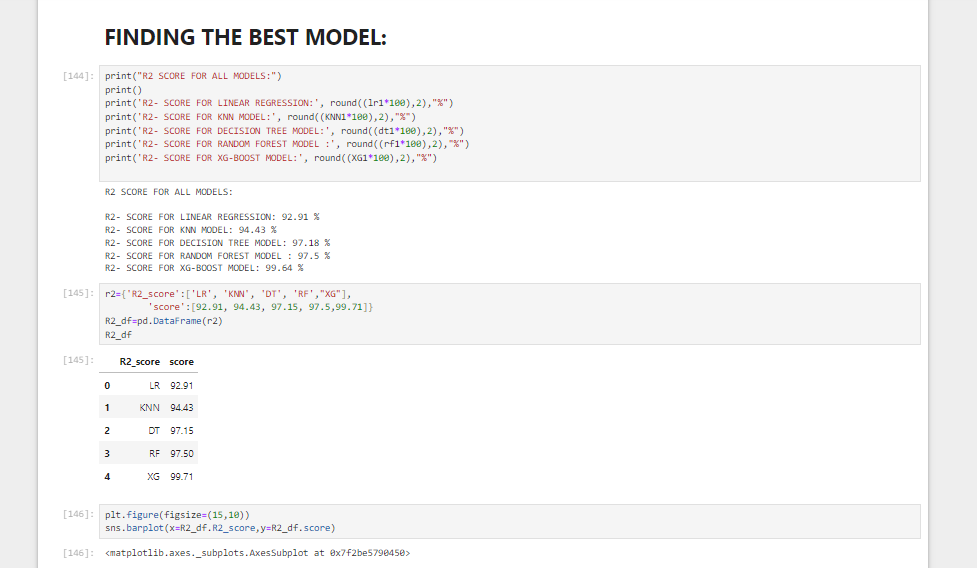


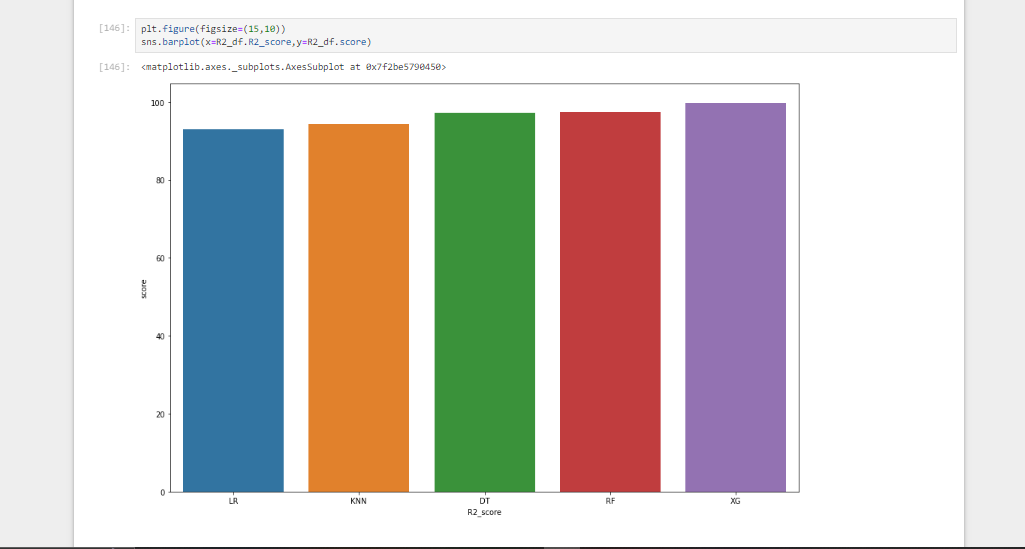












9.CODE SAMPLE:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express as px

%matplotlib inline

#import dataset

df=pd.read\_csv('train-chennai-sale.csv')

**ENCODING:**

**ORDINAL LABEL ENCODING:**

For AREA, SALE COND,UTILITY\_AVAIL,STREET AND MZ ZONE, we apply ordinal label encoding.

df1=df

df

df1["AREA"] = df1["AREA"].replace({"Karapakkam": 0,"Adyar": 1, "T Nagar": 6, "Chrompet": 2, "Velachery": 3, "Anna Nagar" : 5,"KK Nagar" : 4})

df1["SALE\_COND"] = df1["SALE\_COND"].replace({"Partial": 0,"Family": 1,"AbNormal": 2, "Normal Sale": 3,"Adj Land" : 4})

df1["UTILITY\_AVAIL"] = df1["UTILITY\_AVAIL"].replace({"ELO": 0, "NoSeWa": 1, "NoSewr" : 2, "AllPub": 3})

df1["STREET"] = df1["STREET"].replace({"No Access": 0,"Paved": 1, "Gravel": 2})

df1["MZZONE"] = df1["MZZONE"].replace({"A": 0,"C": 1, "I": 2, "RH": 3, "RL": 4, "RM" : 5})

df1.info()

SPLITTING THE DATASET:

**from sklearn.model\_selection import train\_test\_split**

**X=df1.drop(['SALES\_PRICE'],axis=1)**

**Y=df1['SALES\_PRICE']**

**X.shape**

**x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.2, random\_state=5)**

**x\_train.shape**

**x\_test.shape**

**5. XG-BOOST MODEL:**

**import xgboost as xgb**

**from sklearn.model\_selection import cross\_val\_score**

**import numpy as np**

**for lr in [0.01,0.02,0.03,0.04,0.05,0.1,0.11,0.12,0.13,0.14,0.15,0.2,0.5,0.7,1]:**

**model = xgb.XGBRegressor(learning\_rate = lr, n\_estimators=100, verbosity = 0) # initialise the model**

**model.fit(x\_train,y\_train) #train the model**

**model.score(x\_test, y\_test) # scoring the model - r2 squared**

**print("Learning rate : ", lr, " Train score : ", model.score(x\_train,y\_train), " Cross-Val score : ", np.mean(cross\_val\_score(model, x\_train, y\_train)))**

**xg\_model = xgb.XGBRegressor(learning\_rate = 0.7, n\_estimators=100, verbosity = 0)**

**xg\_model.fit(x\_train,y\_train) #train the model**

**xg\_model.score(x\_test, y\_test) # scoring the model - r2 squared**

**XG\_pred= model.predict(x\_test)**

**result\_XG= pd.DataFrame({"Actual": y\_test, "Predicted": XG\_pred})**

**result\_XG**

**sns.scatterplot(x= y\_test, y=XG\_pred, color=["red"])**

**plt.title("prediction result vs actual result")**

**plt.show()**

**plt.bar(range(len(model.feature\_importances\_)), model.feature\_importances\_)**

**plt.show()**

**print("R2 SCORE FOR ALL MODELS:")**

**print()**

**print('R2- SCORE FOR LINEAR REGRESSION:', round((lr1\*100),2),"%")**

**print('R2- SCORE FOR KNN MODEL:', round((KNN1\*100),2),"%")**

**print('R2- SCORE FOR DECISION TREE MODEL:', round((dt1\*100),2),"%")**

**print('R2- SCORE FOR RANDOM FOREST MODEL :', round((rf1\*100),2),"%")**

**print('R2- SCORE FOR XG-BOOST MODEL:', round((XG1\*100),2),"%")**

**FEATURE IMPORTANCE**

**from xgboost import plot\_importance**

**# plot feature importance**

**plot\_importance(model)**

**plt.show()**

**plt.figure(figsize=(8,8))**

**sns.barplot(x=feature\_scores,y=feature\_scores.index)**

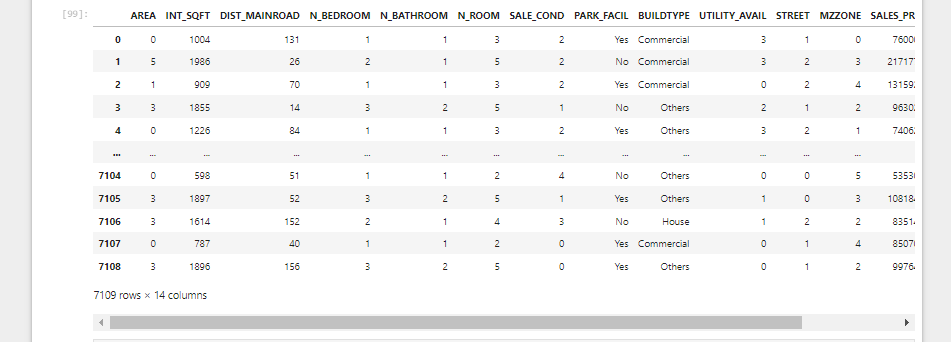
**plt.figure(figsize=(15,15))**

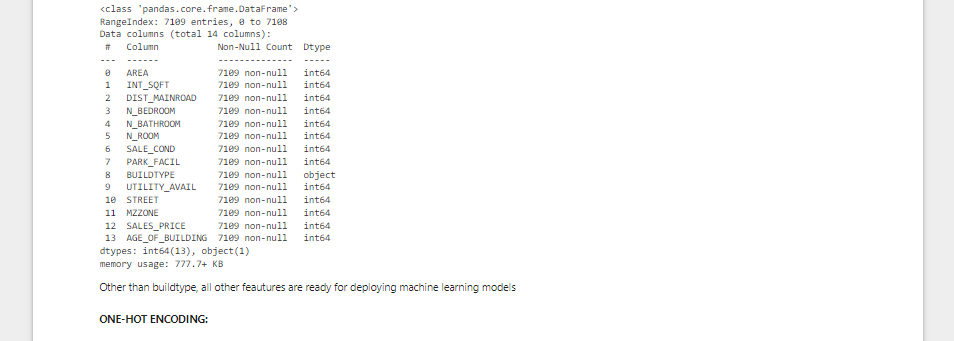
**colors = sns.color\_palette('bright')**

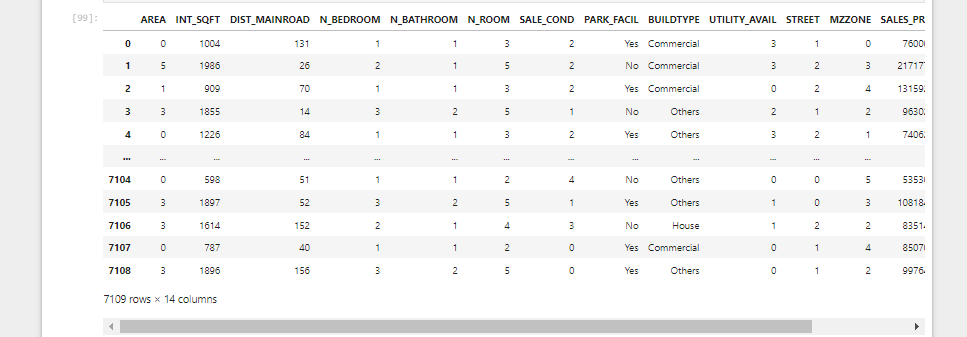
**plt.pie(score, labels=features,colors = colors, autopct = '%0.0f%%')**

**plt.show()**

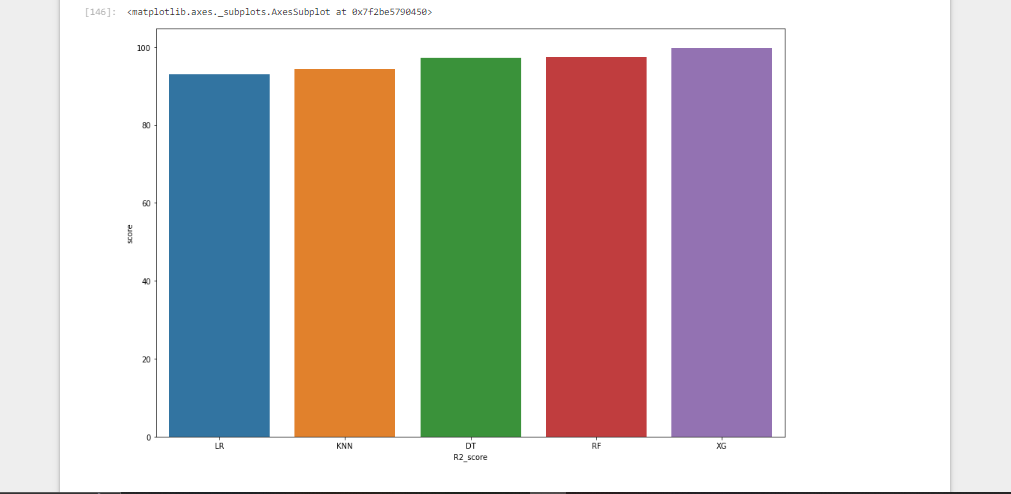
10.OUTPUT:

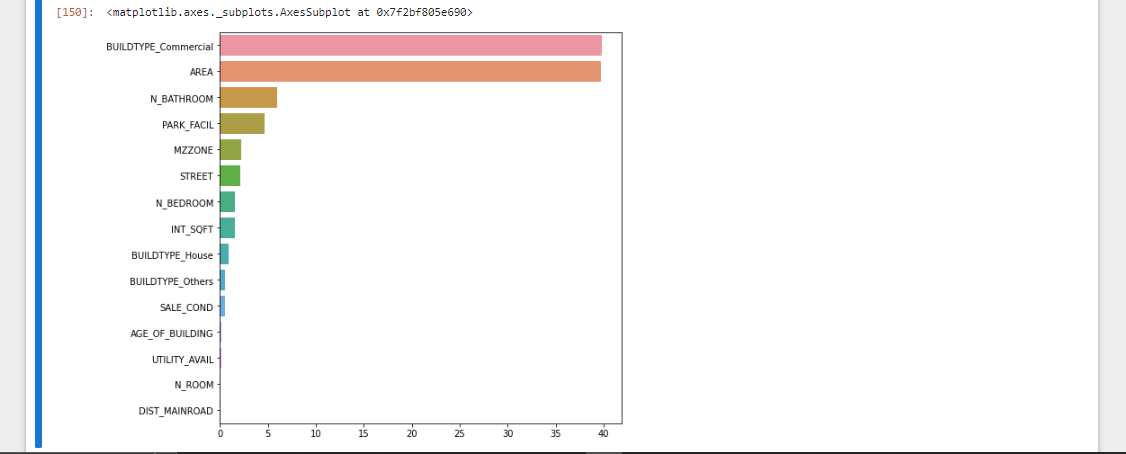


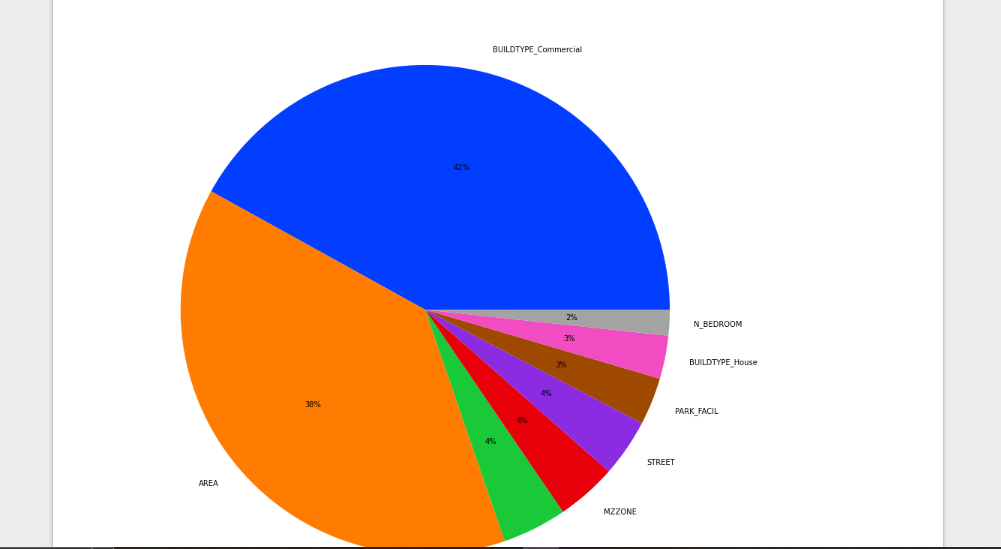












11.Conclusion:

The paper entitled “House Price Prediction Using Machine Learning” has presented to predict house price based on various features on given data. From our analysis we set value of RMSE as 2.9131889. In this model we have to add additional features like tax, air quality so it become different from other prediction system. It helps people to buy house in budget and reduce loss of money.

Future Work:

This paper is currently working on deployment using flask and automate the result file. Use another country housing data set for prediction. This paper is also in other sectors as well as other countries, is yet to be explored.

Reference:

The machine learning model is given the test data but without the price of the properties in order to predict the price for them given the various features for the properties. The predicted price is then compared to the actual price in the test data.