PREDICTING HOUSE PRICE USING MACHINE LEARNING

Abstract:

Data mining is now commonly applied in the real estate market. Data mining's ability to extract relevant

knowledge from raw data makes it very useful to predict house prices, key housing attributes, and many more. Research

has stated that the fluctuations in house prices are often a concern for house owners and the real estate market. A survey

of literature is carried out to analyze the relevant attributes and the most efficient models to forecast the house prices.

The findings of this analysis verified the use of the Artificial Neural Network, Support Vector Regression and XGBoost

as the most efficient models compared to others. Moreover, our findings also suggest that locational attributes and

structural attributes are prominent factors in predicting house prices. This study will be of tremendous benefit,

especially to housing developers and researchers, to ascertain the most significant attributes to determine house prices

and to acknowledge the best machine learning model to be used to conduct a study in this field.

Introduction:

Every single relationship in the current land business is working helpfully to achieve an upper hand over elective players. It is necessary to improve collaboration for a typical individual while achieving the greatest results. This research offers a system that uses backslide AI to forecast housing expenses. If you will sell a house, you want to see what retail cost to put on it. Moreover, a PC assessment can give you a precise measure. This backslide model is collected not only for expecting the expense of the house which is arranged accessible to be bought at this point also for houses that are a work in progress. The real estate market is a hero among the most preoccupied with valuing and continues to evolve.

Inversion is an AI device that urges you to do what is generally anticipated of you by taking - from the current quantifiable data - the association between the control boundary and the number of various boundaries. As indicated by this definition, the expense of a house relies upon the boundaries, for instance, the number of rooms, convenience, convenience, etc. In the impossible occasion that we utilize extortion that observes these cutoff points, we can compute lodging costs in a specific region of the world.

The objective part in The cost of the land property is included in this suggested model, as are the independent components: no. of rooms, no. of washrooms, cover locale, created area, the floor, age of the property, postal region, degree, and longitude of the property. Other than those of the referred to features, which stay generally expected at expecting the house costs, We've incorporated two different components: air quality and noise pollution. These features give a significant responsibility towards predicting property costs since the higher potential gains of these components will incite a reduction in house costs.

System design and architecture:

Step 1: Collection of data:

Information handling strategies and cycles are various. We gathered the information for Mumbai's land properties from different land sites. The information would have traits, for example, Location, cover region, developed region, age of the property, postal district, and so forth we should gather the quantitative information which is organized and ordered. Information assortment is required before any sort of AI research is completed. Dataset legitimacy is an unquestionable requirement in any case it is a waste of time to break down the information.

Step 2: Information preprocessing:

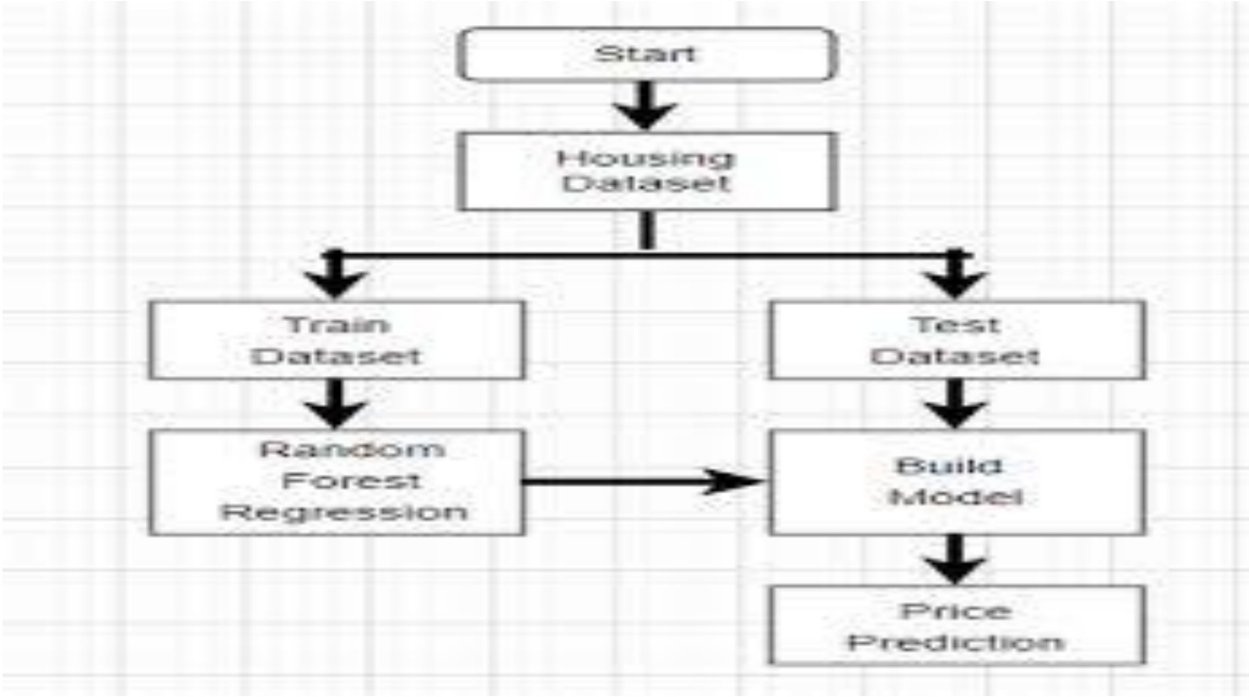
Data preprocessing is the most well-known approach to cleaning our instructive file. There might be missing characteristics or irregularities in the dataset. Data cleansing can help with these issues. Expecting a variable to have a large number of missing attributes we drop persons characteristics else substitute them with the typical worth.

Step 3: The model's education:

We should train the model first since the data is separated into two modules: a Training set and a Test set. The objective variable is joined by the readiness set. The layout of educational assortment is computed using a decision tree regressor. A backslide model is collected as a tree structure by the Decision tree.

Step 4: Testing and Integrating with UI:

The test dataset is fed into the pre-programmed model, and home expenses are predicted. The front end, which includes Flask in Python, is then designed using the pre arranged model



Methodology:

A. Studied Algorithms:

During the time spent encouraging this model, different backslide computations were thought about. Straight backslide, Multiple straight backslide, Decision Tree Regressor, and KNN are all examples of machine learning techniques. were attempted upon the planning dataset. In any case, the decision tree regressor gave the most raised accuracy to the extent that expecting the house costs. The decision to pick the computation particularly depends on the angles and the type of data in the data that was used For our dataset, the decision tree computation is the best option.

B. Regressor for decision tree:

The decision tree regressor recognises quality components and trains a model like a tree to forecast data in the future to provide a massive result. The highest significance and minimum significance of a chart are gained by the decision tree regressor, which then separates the data as demonstrated by the system.

Network Search CV is a strategy for overseeing limit tuning that will beneficially create and study a model for each mix of calculation limits exhibited in a cross-section. System Lookup In this calculation, CV is utilised to determine the optimal impetus for max-significance, which is then used to construct the decision tree.

Implementation:

A. Data processing :

For their missing traits, age and floor restrictions were addressed. In addition, the preparation dataset's goal attribute is removed.

This is why the Pandas library is used. The objective characteristic's min, max, standard deviation, and mean were identified for the

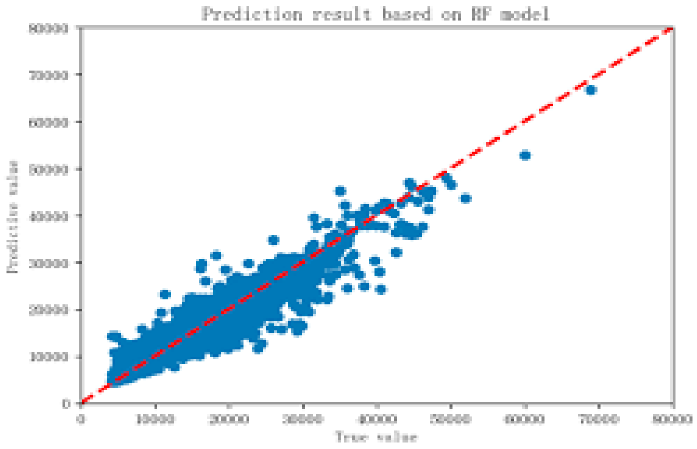
factual representation of the dataset. We divided the dataset into two parts: a preparation set (80%) and a test set (20%). (20 percent).

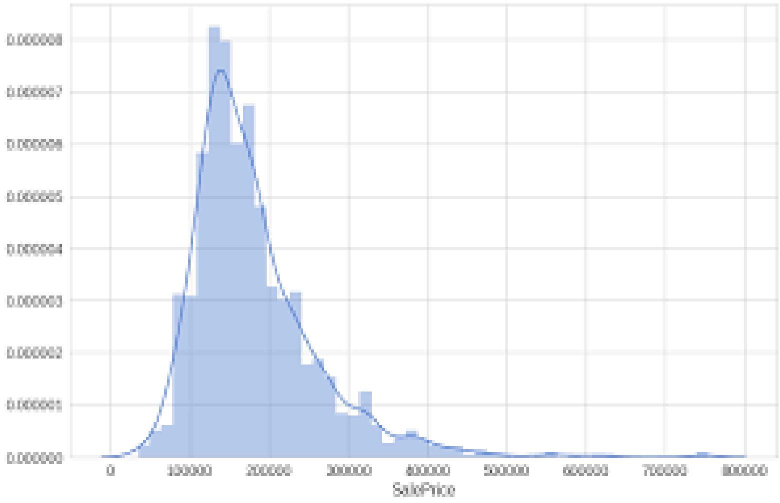
B. Max-profundity:

As previously said, system scan cv assists in determining the bush's maximum relevance. To visualise the various max-profundities

and unpredictable execution, we utilised Matplotlib.

Graph:





Future Scope:

The utilized pre-handling strategies truly do help in the forecast exactness. Nonetheless, exploring different avenues regarding various blends of pre-handling strategies to accomplish better expectation exactness.

- Utilize the accessible elements and assuming they could be joined as binning highlights has shown that the information got moved along.

- Preparing the datasets with various relapse strategies, for example, Elastic net relapse that consolidates both L1 and L2 standards. To grow the examination and check the execution.

The connection has shown the relationship in the neighborhood information. In this way, endeavoring to improve theneighborhood information is expected to cause rich with highlights that fluctuate and can give a solid connection relationship.

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. Location:

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

B. Structural:

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

C. Neighborhood:

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.

Program:

```
 from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

import pandas as pdd

# Loading the dataset

data_h = pdd.read_csv('kc_house_data.csv')

# Selecting the features and target variable

Features1 = ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors',

'zipcode']

target = 'price'

X1 = data_h[features1]

y1 = data_h[target]

# We will perform the data splitting into training and testing sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X1, y1, test_size=0.2,
random_state=42)
# instance of the Linear Regression model creation
model = LinearRegression()
# Training the model
model.fit(X_train, y_train)
# Making predictions on the test set
y_pred = model.predict(X_test)
# Evaluating the model
score = model.score(X_test, y_test)
print("Model R^2 Score:", score)
# Predicting the price of a new house
new_house = pdd.DataFrame({'bedrooms': [2], 'bathrooms': [2.5],
'sqft_living': [600], 'sqft_lot': [600], 'floors': [2], 'zipcode': [98008]})
predicted_price = model.predict(new_house)
print("Predicted Price:", predicted_price[0])
```

Output:

```
C:\Users\Tutorialspoint>python image.py
  Model R^2 Score: 0.5152176902631012
  Predicted Price: 121215.61449578404
```

Example program:

```
from sklearn import svm
from sklearn.svm import SVC
from sklearn.metrics import mean_absolute_percentage_error

model_SVR = svm.SVR()
model_SVR.fit(X_train,Y_train)
Y_pred = model_SVR.predict(X_valid)

print(mean_absolute_percentage_error(Y_valid, Y_pred))
```

Out put:

```
        0.18705129
```

Data Cleaning And integration:

The data obtained from the repository is in the form text file I have connected the text with the excel and the data is being extracted from the text file and moved into the excel file and it has been saved as the comma-separated file. Data cleaning is an iterative process, the first iterate is on detecting and correcting bad records the data taken from the repository have many inconsistencies and null values before loading into the machine learning models the data should be corrected in order to get the high accuracy of prediction as I am using the different tools for prediction, the cleaning process differs from one other but the ultimate goal is to gain more accuracy. The real estate data have some missing information they dont have the states name only latitude and longitude were given by using the R program I have identified the states they all seem to be the states present in the united states of America and the null values are removed to reduce the inconsistency.

Tools :

S.no	Name of the tool	Algorithms used
1	R studio	Random forset,Multiple ReGression
2	Rapid Miner	Support Vector Machine,Gradient Boosted Trees
3	Weka	Neural networks,Bagging

Conclusion:

The main goal of this project is to determine the prediction for prices which we have successfully done using different machine learning algorithms like a Random forest, multiple regression, Support vector machine, gradient boosted trees, neural networks, and bagging, so it's clear that the random forest have more accuracy in prediction when compared to the others and also my research provides to find the attributes contribution in prediction. So I would believe this research will be helpful for both the peoples and governments and the future works are stated below

Every system and new software technology can help in the future to predict the prices. price prediction this can be improved by adding many attributes like surroundings, marketplaces and many other related variables to the houses. The predicted data can be stored in the databases and an app can be created for the people so they would have a brief idea and they would invest the money in a safer way. If there is a possibility of real-time data the data can be connected to the H2O and the machine learning algorithms can be directly connected with the interlink and the application environment can be created.

THANK YOU