PROBLEM DEFINITION

A house is one of the most important necessities of life, as well as other basic necessities like food, water, and much more. Demand for housing has grown rapidly over the years as the standard of living of the people improves. Although there are people who make their homes as an investment and property, yet many people around the world buy a house as their living space or as a means of subsistence. According to , the housing market has a positive impact on the national currency, which is an important measure of the national economy.

Homeowners will buy furniture such as furniture and household items, and homeowners or contractors will buy raw materials to build houses to meet housing needs, which is an indication of the economic impact of the new housing market. Besides, buyers have the potential to invest heavily, and the construction industry is in good shape and can be seen by the country's high standard of housing. According to, many international organizations and human rights groups have emphasized the importance of housing.

The House focuses on the economic, financial, and political structure of each country. However, reported that housing price volatility has always been a problem for homeowners, real estate and real estate, apart from that he said that housing is not affordable as there has been significant inflation in several countries housing sector. The quality of life of the citizens and the economy of the country depends on the possible increase in housing prices. Ultimately, the issue will affect investors who make their home as an investment. The increase in housing demand occurs every year, resulting in an indirect increase in house prices every year.

The problem arises when there are many variations such as location and the need for a building that may affect the price of a house, so most participants including buyers and developers, real estate developers and the real estate industry would like to know the exact factors or factors that influence a home. price to help investors make decisions and help homeowners set home prices. Home price guessing can be done using multiple guessing models (Machine Learning Model) such as vector retrieval, neural artificial network, and more.

There are many benefits to real estate buyers, real estate investors, and real estate agents from the real estate model. This model will provide a wealth of information and information to real estate agents, real estate investors and real estate agents, such as estimating real estate prices in the current market, which will help them to determine house prices.

At the same time, this model can help potential buyers determine the features of the home they want in terms of their budget. Previous research has focused on analyzing house price characteristics. predicting house prices based on a separate machine learning model. However, this article combines both the predictable house price and the properties together. the literature review focuses on the pricing of home prices based on the machine learning model and analysis of the characteristics used primarily in previous research affecting house prices. This paper is organized as follows: the first phase summarizes the entire study. The second section describes the general characteristics used in real estate prices around the world.

DESIGN THINKING

There are few steps to solve how to predict the house prices in AI, they are

• Load Data and Packages.

• Analyzing the Test Variable (Sale Price)

• Multivariable Analysis.

• Impute Missing Data and Clean Data.

• Feature Transformation/ Engineering.

• Modeling and Predictions.

DATASET

<https://www.kagle.com/datasets/vedavyasv/usa-housing>

DATASET DESCRIPTION

The "USA\_Housing" dataset is a collection of data related to housing in various locations within the United States. It often includes information such as the price of houses, features or attributes of the properties, and other relevant data points.

Typical Columns:

Avg. Area Income: Averageincome of residents in the area.Avg. Area House Age: Average age of houses in the area.Avg. Area Number of Rooms: Average number of rooms in houses.Avg. Area Number of Bedrooms: Average number ofAvg. Area Number of Bedrooms: Average number of bedrooms in houses.Area Population: Population of the area.Price: The price at which a house was sold.Address: The address of the property.

This dataset is often used for tasks like housing price prediction, regression analysis, and understanding the factors that influence housing prices in different areas of the United States. You can perform data analysis and build predictive models to gain insights into housing trends and pricing factors.

REGRESSION ALGORITHM AND EVALUATION METRICS

1. Regression Algorithm Selection:

* Linear Regression: This is a good starting point as it's simple and interpretable. It assumes a linear relationship between the input features (like square footage, number of bedrooms, etc.) and the target variable (house price). If the relationship is roughly linear, linear regression can perform well.Decision Trees and Random Forest: These are flexible algorithms that can capture non-linear relationships. They can handle complex interactions between features, which might be present in real estate data.
* Gradient Boosting (e.g., XGBoost, LightGBM): These ensemble methods often yield high predictive accuracy. They're capable of modeling complex relationships, handling missing data, and feature selection.
* Decision Trees and Random Forest: These are flexible algorithms that can capture non-linear relationships. They can handle complex interactions between features, which might be present in real estate data.
* Gradient Boosting (e.g., XGBoost, LightGBM): These ensemble methods often yield high predictive accuracy. They're capable of modeling complex relationships, handling missing data, and feature selection.

1. Evaluation Metrics:

* Mean Absolute Error (MAE): This metric calculates the average absolute difference between predicted and actual prices. It's easy to understand as it represents the average dollar amount by which your predictions are off.
* Mean Squared Error (MSE): MSE squares the errors, which penalizes larger errors more heavily. It's often used but is sensitive to outliers. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, and it provides a measure of how much your predictions deviate, on average, from the true prices.
* R-squared (R2): R2 measures the proportion of the variance in the target variable that's explained by the model. A higher R2 indicates a better fit.
* Cross-Validation: It's important to use cross-validation techniques (e.g., k-fold cross-validation) to assess the model's performance on multiple subsets of the data. This helps ensure that the model generalizes well to unseen data.

LOAD THE HOUSE DATA SET AND PRE-PROCESS THE DATA

import numpy as np

import pandas as pd

import os

for dirname, \_, filenames in os.walk('/kaggle/input'):

    for filename in filenames:

        print(os.path.join(dirname, filename))

…/kaggle/input/usa-housing/USA\_Housing.csv

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

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

dataFrame= pd.read\_csv('/kaggle/input/usa-housing/USA\_Housing.csv')

X = dataFrame[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms', 'Avg. Area Number of Bedrooms', 'Area Population']]

y = dataFrame['Price']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

print("Mean Squared Error:{mse}")

…Mean Squared Error:{mse}

BUILDING THE PROJECT USING FEATURE SELECTION, MODEL TRAINING, EVALUATION

import numpy as np

import pandas as pd

import os

for dirname, \_, filenames in os.walk('/kaggle/input'):

    for filename in filenames:

        print(os.path.join(dirname, filename))

…/kaggle/input/usa-housing/USA\_Housing.csv

import pandas as pd

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

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

X = data[['Avg. Area Income', 'Avg. Area House Age', 'Area Population']]

y = data['Price']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.2,random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

* **LinearRegression()**

y\_pred = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, y\_pred)

rmse = mse \*\*05

print("Mean Squared Error:", mse)

print("Root Mean Squared Error:", rmse)

**…Mean Squared Error: 25217178477.03674**

**Root Mean Squared Error: 158799.1765628422**