HOUSE PRICE PREDICTION

A MINI PROJECT REPORT

18CSC305J - ARTIFICIAL INTELLIGENCE

Submitted by

SHASHANK NANDANWAR [RA2011003010779]
VANSH BHAVISHI [RA2011003010786]
SMARAJIT BAKSI [RA2011003010813]

Under the guidance of

Mrs. Renuka Devi

Assistant Professor, Department of Computer Science and Engineering

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

S.R.M. Nagar, Kattankulathur, Chengalpattu District

MAY 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that Mini project report titled **"HOUSE PRICE PREDICTION"** is the bona fide work of **SHASHANK NANDANWAR [RA2011003010779], VANSH BHAVISHI [RA2011003010786], SMARAJIT BAKSI [RA2011003010813]** who carried out the minor project under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE	**SIGNATURE**
Mrs. Renuka Devi	Dr. M. Pushpalatha
GUIDE	**HEAD OF THE DEPARTMENT**
Assistant Professor	Professor & Head
Department of Computing Technologies	Department of Computing Technologies

Abstract

House price prediction is a crucial task in the real estate industry, as it allows buyers and sellers to make informed decisions about property transactions. It involves using machine learning algorithms and data science techniques to estimate the value of a property based on various factors such as location, square footage, number of bedrooms and bathrooms, and other features. Accurate house price prediction can provide valuable insights into the housing market and help real estate professionals and consumers to understand the current market trends.

Linear regression is one of the most widely used algorithms for house price prediction, as it provides a straightforward and interpretable model for understanding how different features affect house prices. However, other machine learning techniques such as decision trees, random forests, and neural networks can also be employed to improve the accuracy and robustness of the predictions. With the increasing availability of data and advances in machine learning, house price prediction is becoming more accurate and reliable, making it an essential tool for anyone involved in the housing market. In conclusion, accurate house price prediction is essential for making informed decisions in the real estate industry, and machine learning techniques can provide valuable insights into the current and future market trends.

Contents

Problem Statement

The prices of houses vary from area to area in every city. Therefore, there is a need for some technology which will help in finding the idea house in their preferred locality and budget.

Traditionally brokers exist, who help in finding houses in the required location, but most of the times the brokers might not get you the ideal deal within your budget. Finding a house through is very time consuming and not cost effective. There is a need for a smart solution that can provide time saving and cost-effective solution for the above problem.

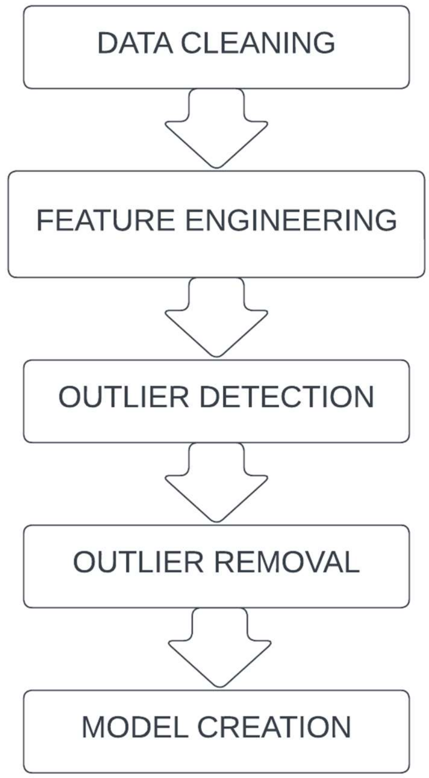
By addressing these challenges, the proposed project aims to contribute to an efficient AI model which will solve this problem by saving time, money and finding your ideal house.

Problem Explanation

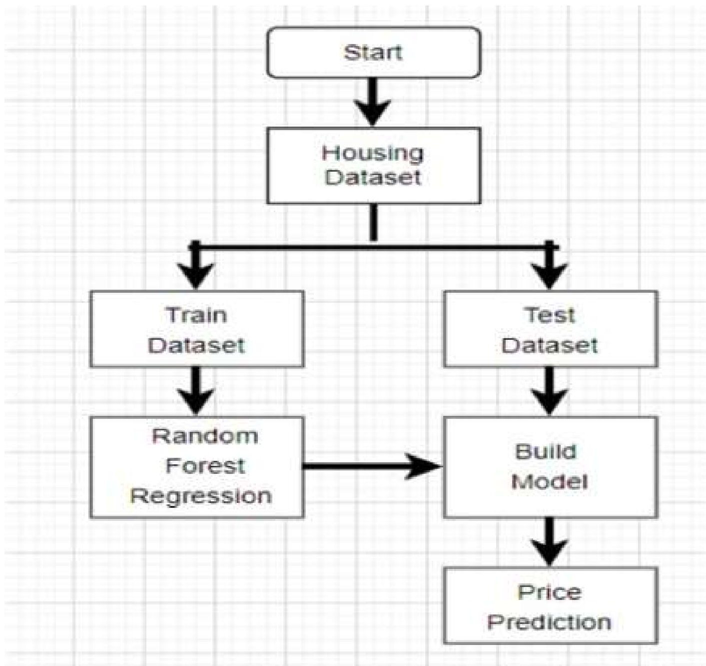
Real estate prices are indeed intricately linked with the economy, as they are influenced by a variety of factors such as interest rates, inflation, employment rates, and GDP growth. In addition to these macroeconomic indicators, local factors such as the availability of infrastructure, proximity to schools and healthcare facilities, and crime rates also impact the value of properties. However, despite the abundance of data available, accurately determining the fair market value of a property can be challenging. This is especially true in developing countries like India, where there is a lack of transparency in the real estate industry. The absence of reliable information can create an information asymmetry between buyers and sellers, which can result in either party being taken advantage of. Having accurate measures of house prices based on data analysis can bring more transparency and trust back to the real estate industry. For instance, a transparent pricing mechanism can reduce the scope for fraudulent activities such as underreporting the actual transaction value to evade taxes. Moreover, it can help buyers and sellers make informed decisions, reduce the time taken to close a transaction, and improve the liquidity of the real estate market.

One way to achieve transparency in the pricing of properties is by leveraging technology and data analytics. Real estate analytics platforms can collect, process, and analyse vast amounts of data, including property sales history, comparable sales in the area, and property characteristics. Using sophisticated algorithms, these platforms can provide accurate estimates of a property's fair market value, taking into account both macro and local factors. In conclusion, accurate measures of house prices based on data analysis can bring more transparency and trust back to the real estate industry. This is particularly important for consumers in India, where a lack of transparency in the industry has often resulted in disputes and mistrust. Leveraging technology and data analytics can help create a more efficient and transparent real estate market, benefitting all stakeholders involved.

Flow Chart


DATA CLEANING
FEATURE ENGINEERING
OUTLIER DETECTION
OUTLIER REMOVAL
MODEL CREATION

Architecture



Algorithm used for the Problem

LINEAR REGRESSION

Linear regression is a popular algorithm used in house price prediction. It involves fitting a linear equation to the data, where the target variable (house price) is modelled as a linear combination of the input features (such as location, size, number of rooms, etc.). The goal is to find the best possible line that fits the data points, and use it to predict the price of a new house based on its features.

The linear regression algorithm works by finding the optimal values for the coefficients (slope and intercept) of the line that best fit the data. This is done by minimizing the sum of the squared differences between the predicted and actual values of the target variable. The resulting line represents the linear relationship between the input features and the target variable.

Once the linear regression model is trained on a dataset of historical house prices and features, it can be used to predict the price of a new house based on its features. To do this, the input features of the new house are fed into the model, and the algorithm uses the learned coefficients to calculate the predicted price.

It is important to note that linear regression assumes a linear relationship between the input features and the target variable, which may not always be the case in real-world scenarios. Additionally, linear regression may not be able to capture complex non-linear relationships between the features and the target variable. Therefore, it is important to use other algorithms as well and compare their performance to choose the best model for predicting house prices.

Implementation and Methodology

The implementation of linear regression in house price prediction typically involves the following steps:

1. Data preparation: The first step is to collect the data on historical house prices and their corresponding features, such as location, size, number of rooms, and other relevant information. The data must be pre-processed, which may involve cleaning the data, handling missing values, and removing outliers.

2. Feature selection and engineering: The next step is to select the most relevant features that will be used to predict the house price. Feature engineering may also be performed to create new features that may be more informative.

3. Splitting the data: The data is then split into training and testing sets. The training set is used to train the linear regression model, while the testing set is used to evaluate its performance.

4. Model training: The linear regression model is trained on the training set using an optimization algorithm such as gradient descent. The goal is to find the best values for the coefficients of the linear equation that will minimize the difference between the predicted and actual house prices.

5. Model evaluation: The trained model is evaluated on the testing set using various performance metrics such as mean squared error, root mean squared error, and R-squared. These metrics are used to determine the accuracy of the model and to compare it to other models.

6. Prediction: Once the model is trained and evaluated, it can be used to predict the price of a new house based on its features. The input features are fed into the model, and the model uses the learned coefficients to calculate the predicted price.

Overall, the implementation of linear regression in house price prediction involves collecting and pre-processing the data, selecting relevant features, splitting the data, training the model, evaluating its performance, and using it to make predictions.

Code

Open in Colab

```
In [ ]: HOUSE PRICE PREDICTION
```

```
In [ ]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np

        #data set
        from sklearn.datasets import fetch_california_housing
```

```
In [ ]: data=fetch_california_housing()
```

```
In [ ]: print(data.DESCR)
```

```
In [ ]: df=pd.DataFrame()
```

```
In [ ]: data.data
```

```
Out[ ]: array([[   8.3252    ,   41.        ,    6.98412698, ...,    2.55555556,
           37.88      , -122.23      ],
        [   8.3014    ,   21.        ,    6.23813708, ...,    2.10984183,
           37.86      , -122.22      ],
        [   7.2574    ,   52.        ,    8.28813559, ...,    2.80225989,
           37.85      , -122.24      ],
        ...,
        [   1.7       ,   17.        ,    5.20554273, ...,    2.3256351 ,
           39.43      , -121.22      ],
        [   1.8672    ,   18.        ,    5.32951289, ...,    2.12320917,
           39.43      , -121.32      ],
        [   2.3886    ,   16.        ,    5.25471698, ...,    2.61698113,
           39.37      , -121.24      ]])
```

```
In [ ]: data.data.shape
```

```
Out[ ]: (20640, 8)
```

```
In [ ]: #independent data
        df=pd.DataFrame(data=data.data,columns=data.feature_names)
        df.head()
```

Out[ ]:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25

```
In [ ]: #dependent data
        df['Target']=data.target
```

```
In [ ]: df.head()
```

Out[ ]:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

In [ ]:
```
def location(cord):
  Latitude= str(cord[0])
  Longitude= str(cord[1])

  location=geolocator.reverse(Latitude+","+Longitude).raw['address']

  # if the values are missing replace by empty string

  if location.get('road') is None:
    location['road']=None

  if location.get('county') is None:
    location['county']=None

  loc_update['County'].append(location['county'])
  loc_update['road'].append(location['road'])
```

In [ ]:
```
df=df.drop(labels=["Latitude","Longitude"],axis=1)
df.head()
```

Out[ ]:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	3.422

In [ ]:
```
from sklearn.linear_model import LinearRegression

model=LinearRegression()

model.fit(X_train,y_train)
```

Out[ ]: LinearRegression()

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [ ]:
```
y_pred=model.predict(X_test)
```

In [ ]:
```
from sklearn.metrics import r2_score

r2_score(y_test,y_pred)*100
```

Out[ ]: 100.0

In [ ]:
```
df.head()
```

Out[ ]:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467

Add data

In [ ]:
```
inp=np.array([8.3252,   41.0,   6.984127,      1.023810,      322.0,  2.555556])
```

In [ ]:
```
ip=inp.reshape((1,-1))
```

In [ ]:
```
model.predict(ip)
```

Out[ ]: array([2.555556])

Result

Our function, predict_price(location, sqft, bath, bhk), takes four input parameters: location, sqft, bath, and bhk. These parameters represent the features of a house that you want to use to predict its price. When we pass the values of these parameters into the function, it uses a trained machine learning model to predict the price of the house based on the input features. The machine learning model could be a linear regression model, a decision tree model, a neural network, or any other model that has been trained on historical data of house prices and their corresponding features. Once the machine learning model is trained, it can take in the input features and use them to predict the price of the house. The output of the function is the predicted price of the house, which is returned to the user.

Overall, our function provides a convenient way to predict the price of a house based on its features, which can be useful for real estate professionals, homebuyers, and sellers.

Conclusion

In conclusion, house price prediction is a complex and important task that involves using data science and machine learning techniques to predict the value of a house based on its features. With the vast amount of data available today, it is possible to build accurate models that can predict house prices with a high degree of accuracy. This can be very useful for real estate professionals, homebuyers, and sellers, as it can provide valuable insights into the housing market and help them make informed decisions. Linear regression is one of the most commonly used algorithms for house price prediction, but there are many other machine learning techniques that can be used depending on the nature of the data and the problem at hand.

Overall, house price prediction is a fascinating area of research that has the potential to revolutionize the real estate industry and bring more transparency and efficiency to the market.

Possible Future Enhancements

Here are some potential future enhancements for house price prediction:

1. Incorporating Additional Data Sources: In addition to traditional housing data, future models could include additional data sources such as crime rates, proximity to public transportation, and neighbourhood amenities. This would provide a more comprehensive view of the factors that influence house prices.

2. Spatial Analysis: Spatial analysis could be used to analyze the impact of geographical features such as terrain, climate, and natural disasters on housing prices.

3. Time Series Analysis: Time series analysis could be used to identify trends and seasonality in the housing market. This could help predict future housing prices and provide insights into market cycles.

4. Hybrid Models: Hybrid models that combine different AI techniques, such as combining deep learning and reinforcement learning, could provide more accurate and robust predictions.

5. Explainable AI: Explainable AI techniques could be used to provide transparent and interpretable models that help users understand how the model arrived at its predictions.

6. Real-Time Predictions: The ability to provide real-time predictions could be valuable for investors, buyers, and sellers in the housing market. This could be achieved by using streaming data and machine learning algorithms that can update predictions in real-time.

7. Transfer Learning: Transfer learning techniques could be used to train models on data from different geographic regions or housing markets. This could provide more accurate predictions for new markets without the need for significant amounts of local data.

Overall, future enhancements in house price prediction will likely involve the integration of multiple data sources, advanced AI techniques, and real-time predictions to provide accurate and actionable insights for stakeholders in the housing market.

References

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- https://www.kaggle.com/c/boston-housing
- https://towardsdatascience.com/house-price-prediction-using-linear-regression-technique-9ebf6e8c6d33
- https://towardsdatascience.com/predicting-house-prices-with-machine-learning-6f56c34f29dc
- https://medium.com/swlh/house-price-prediction-with-multiple-regression-and-feature-importance-ml-model-7d52a0206f68