PREDICTING HOUSE PRICE USING MACHINE LEARNING

PHASE 5 SUBMISSION DOCUMENT

Project Title : PREDICTING HOUSE PRICE USING MACHINE LEARNING

Phase 4: Development Part III

ABSTRACT

The trend of the sudden drop or constant rising of housing prices has attracted interest from the researcher as well as many other interested people. There have been various research works that use different methods and techniques to address the question of the changing of house prices. This work considers the issue of changing house price as a classification problem and discuss machine learning techniques to predict whether house prices will rise or fall using available data. This work applies various feature selection techniques such as variance influence factor, Information value, principle component analysis, and data petransformation techniques such as outlier and missing value treatment as well as different transformation techniques. The performance of the machine learning techniques is measured by the four parameters of accuracy, precision, specificity, and sensitivity. The work considers two discrete values 0 and 1 as respective classes. If the value of the class is 0 then we consider that the price of the house has decreased and if the value of the class is 1 then we consider that the price of the house has increased.

INTRODUCTION

Development of civilization is the foundation of the increase in demand for houses dayby da y. Accurate prediction of house prices has been always a fascination for buyers, sellers, and bankers also. Many researchers have already worked to unravel the mysteries of the predicti on of house prices. Many theories have been given birth as a consequence of the research wo rk contributed by various researchers all over the world. Some of these theories believe that t he geographical location and culture of a particular area determine how the home prices will increase or decrease whereas other schools of thought emphasize the socio-economic conditi ons that largely play behind these house price rises.We all know that a house price is a numb er from some defined assortment, so obviously prediction of prices of houses is a regression task. To forecast house prices one person usually tries to locate similar properties in his or he r neighborhood and based on collected data that person will try to predict the house price.

All these indicate that house price prediction is an emerging research area of regression that requires the knowledge of machine learning. This has motivated me to work in this domain. Reale state appraisal is an integral part of the property buying process. Traditionally, the app raisal is performed by professional appraisers specially trained for real estate valuation.

Tools required for predicting house price using machine learning:

What is GitHub?

GitHub is a web-based interface that uses Git, the open source version control software that lets multiple people make separate changes to web pages at the same time. As Carpenter notes, because it allows for real-time collaboration, GitHub encourages teams to work together to build and edit their site content.

How can GitHub help my team and me?

GitHub allows multiple developers to work on a single project at the same time, reduces the risk of duplicative or conflicting work, and can help decrease production time. With GitHub, developers can build code, track changes, and innovate solutions to problems that might arise during the site development process simultaneously. Non-developers can also use it to create, edit, and update website content, which Carpenter demonstrates in her tutorial.

How do I speak GitHub?

There are some common terms teams will need to understand when using GitHub. They are:

- **Git** — a tool that allows developers and others to use version control
- **GitHub** — one of many web interfaces for using Git
- **Organization (org)** — a grouping mechanism allowing teams to collaborate across many projects at once
- **Repository (repo)** — a folder in which all files and their version histories are stored
- **Branch** — a version of the repo that allows work without affecting other branches. Repos may have many branches for different possible changes being tested or considered, along with a default branch that serves as the source of truth.
- **Fork** — a new repository that inherits from a parent "upstream" repo. It is used to suggest changes to an "upstream" public repo by someone who doesn't have access to edit in the repo's home org.
- **Markdown (.md)** — a way to write content that converts plain text to formatted text.
- **Commit Changes** — a saved record of a change made to a file within the repo.
- **Pull Request (PR)** — a request for changes made to a branch to be pulled into another branch. Allows multiple users to see, discuss and review work being suggested.
- **Merge** — after a pull request is approved, the commit will be pulled in (or merged) from one branch to another and then, deployed on the live site
- **Issues** — allow users to report issues or bugs and track progress of assigning the fix for the issues.
- **Federalist** — a platform that securely deploys a website from a GitHub repository in minutes and lets users preview proposed and published changes.
- **Projects** — allows you to use GitHub for project management and tracking a set of issues, either for a specific repo or an entire org
- **Wiki** — a section of a repo made for hosting documentation. Documentation may be in the repo's README files instead.

Becoming fluent in GitHub terminology might seem intimidating at first, but the more team members engage with the platform, the easier it is to understand the ins and outs of GitHub.

How do I use GitHub?

In the demonstrations on this page, both presenters show how files are changed and merged in GitHub. This can be done by any member on the team, developers and non-developers, that has access to a GitHub repository. The following is a step-by-step method in which GitHub users can develop their websites:

- **Step 1** — Team members will open an issue via a project board.
- **Step 2** — Team members will create a new branch from the most recent version of the main branch in the repository where the entire team works to avoid conflicts.
- **Step 3** — Team members will add commits (edits or changes) to their respective branches.
- **Step 4** — Team members will open a pull request in which users can assign other team members to review content changes and internally discuss the details of the commits.
- **Step 5** — After waiting for the Federalist build to complete, team members can preview the change on a test version of the website and request reviewers to approve or comment on the change. Once the reviewers approve the pull request, the commits merge into the main branch and are published on the live site.

What else do I need to know about GitHub?

When starting a project using issues and project boards, write your content on external word processors or via Google Docs, and then, save these files to their respective project boards. These steps allow developers and content creators to have a master copy of the file(s), thus helping them track changes over the course of a project.

In addition, developers should consider downloading GitHub Desktop. GitHub Desktop allows users to do everything that could be done on GitHub's web interface, but locally on a user's machine.

GitHub is built to be a collaborative interface. By allowing multiple users to work on the same project simultaneously and requiring cross-team approval for pull requests, GitHub not only allows for, but encourages collaboration within design teams. This type of collaboration can help produce a higher level of quality control.

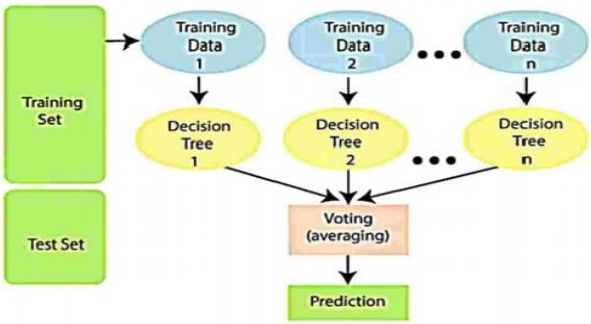
DESIGN FOR PREDICTION HOUSE PRICE USING MACHINE LEARNING

MODELS FOR HOUSE PRICE PREDICTION

Predicting house prices using machine learning involves building a regression model that can estimate the price of a house based on various input features. Here are some common machine learning models and techniques you can use for predicting house prices:

- **Linear Regression**: Linear regression is a simple and interpretable model that assumes a linear relationship between input features and the target variable (house price). You can use techniques like multiple linear regression for multiple features. Regularization techniques such as Lasso or Ridge regression can help prevent overfitting.

- **Decision Trees and Random Forests:** Decision trees and random forests can capture non-linear relationships between features and house prices. Random forests, in particular, are an ensemble of decision trees that can provide more robust predictions.



Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Use of Random Forest

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Working Of Random Forest algorithm

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

- **.Gradient Boosting**: Algorithms like XGBoost, LightGBM, and CatBoost are gradient boosting techniques that often perform well for regression tasks. They can handle complex feature interactions and provide good predictive accuracy.

- .**Support Vector Regression (SVR):** SVR is a regression technique that aims to find a hyperplane that best fits the data while allowing for a certain margin of error. It can be effective for house price prediction.

- **Neural Networks**: Deep learning techniques, such as feedforward neural networks or convolutional neural networks (CNNs), can be used for regression tasks. They are capable of modeling complex relationships but may require more data and computational resources.

- **K-Nearest Neighbors (KNN):** KNN is a non-parametric algorithm that predicts the house price based on the prices of the k-nearest neighboring houses in the feature space.

- **Bayesian Regression**: Bayesian regression models, such as Bayesian Ridge regression, can provide uncertainty estimates along with point predictions, which can be valuable for decision-making.

- **Feature Engineering**: Feature engineering is crucial in house price prediction. You may need to preprocess and engineer features, handle missing data, and transform categorical variables into numerical representations.

- **Regularization:** Regularization techniques like L1 (Lasso) and L2 (Ridge) can help prevent overfitting and improve model generalization.

- **Evaluation Metrics**: Use appropriate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or R-squared (R2) to assess the performance of your models.

- **Cross-Validation:** Employ techniques like k-fold cross-validation to assess the model's performance and prevent overfitting.

- **Hyperparameter Tuning**: Fine-tune hyperparameters of your models to optimize their performance. Techniques like grid search or random search can be used for this purpose.

- **Data Preprocessing**: Data cleaning, normalization, and scaling are important steps in preparing your data for machine learning models.

- **Feature Selection**: Identify and select the most relevant features that have the most impact on predicting house prices.

Remember that the choice of model and techniques depends on the specific dataset and the problem you are trying to solve. It's often a good practice to try multiple models and compare their performance using cross-validation to determine the best approach for predicting house prices.

SAMPLE CODE

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

%matplotlib inline

HouseDF = pd.read_csv('USA_Housing.csv')
HouseDF.head()
HouseDF=HouseDF.reset_index()
HouseDF.head()
HouseDF.info()
HouseDF.describe()
HouseDF.columns
sns.pairplot(HouseDF)
sns.distplot(HouseDF['Price’])
sns.heatmap(HouseDF.corr(), annot=True)

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

y = HouseDF['Price’]

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=101)

from sklearn.linear_model import minmaxscaler

lm = minmaxscaler(feature_range=(0,1))

lm.fit_transform(X_train,y_train)

print(lm.intercept_)

coeff_df = pd.DataFrame(lm.coef_,X.columns,columns=['Coefficient’])
```

```
from keras.layers import Dense,Dropout,LSTM
        from keras.models import Sequential
        model = Sequential()
        model.add(LSTM(units = 50,activation = 'relu',return_sequences = True,input_shape =
        (x_train.shape[1], 1)))
        model.add(Dropout(0.2))

        model.add(LSTM(units = 60,activation = 'relu',return_sequences = True))
        model.add(Dropout(0.3))

        model.add(LSTM(units = 80,activation = 'relu',return_sequences = True))
        model.add(Dropout(0.4))

        model.add(LSTM(units = 120,activation = 'relu'))
        model.add(Dropout(0.5))

        model.add(Dense(units = 1))

        model.compile(optimizer='adam', loss = 'mean_squared_error’)
        model.fit(x_train, y_train,epochs=50)

        print(lm.intercept_)

        coeff_df = pd.DataFrame(lm.coef_,X.columns,columns=['Coefficient’])
        coeff_df

        predictions = lm.predict(X_test)

        scale_factor = 1/0.02099517
        y_predicted = y_predicted * scale_factory
        y_test = y_test * scale_factor

        plt.scatter(y_test,predictions)

        sns.distplot((y_test-predictions),bins=50);
```

```
plt.figure(figsize=(12,6))
plt.plot(y_test,'b',label = 'Original
Price') plt.plot(y_predicted,'r',label =
'Predicted Price') plt.xlabel('Time')
plt.ylabel
('Price')
plt.legend
()
plt.show(
)

from sklearn import metrics

print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test,
predictions)))
```

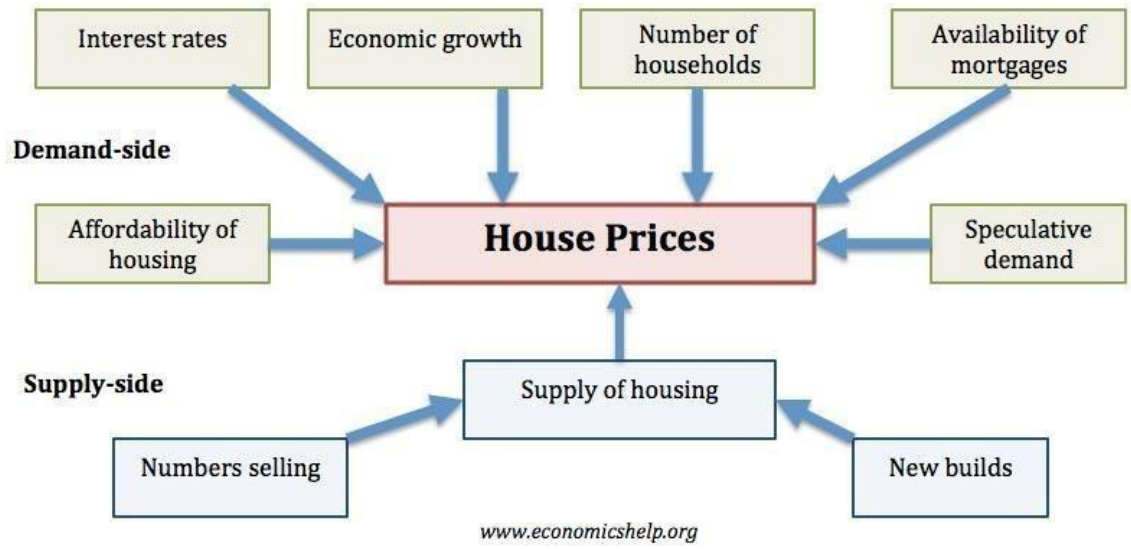
FEATURES OF HOUSE PRICE PREDICTION

House price prediction is a complex task that involves the use of various features to estimate the value of a property accurately. These features, also known as predictors or input variables, can vary depending on the specific machine learning model and dataset used, but here are some common features that are often considered when predicting house prices:

1. **Location**: The location of the property is a critical factor. Features related to location may include the neighborhood, proximity to schools, parks, public transportation, and the distance to amenities like shopping centers and hospitals.
2. **Property Size**: Features like the total area of the property (in square feet or square meters), the number of bedrooms, bathrooms, and the size of the yard can significantly impact the price.
3.
4. **Property Age and Condition**: The age of the property and its overall condition, including recent renovations or updates, can affect its value.
5. **Amenities and Features**: Special features such as a swimming pool, garage, fireplace, central heating, air conditioning, or smart home systems can influence the price.
6. **Historical Sales Data**: Past sales data of similar properties in the area can be a useful feature for predicting the current property's price.
7. **Market Trends**: Current real estate market conditions, such as supply and demand, interest rates, and economic factors, can play a role in price prediction.
8. **Crime Rate**: Safety is a concern for potential homebuyers, so the local crime rate can be a relevant feature.
9. **School Quality**: Proximity to good schools and the overall quality of the education system in the area can be a significant factor for families.
10. **Transportation**: Access to public transportation and commuting options can affect property values.
11. **Local Amenities**: Proximity to parks, shopping centers, restaurants, and other local amenities can influence the price

GRAPH STRUCTURE FOR HOUSE PRICING:

In order to predict house prices, first we have to understand the factors that affect housepricing.



- **Economic growth.** Demand for housing is dependent upon income. With highereconomic growth and rising incomes, people will be able to spend more on

  houses; this will increase demand and push up prices. In fact, demand for housing is often noted to be income elastic (luxury good); rising incomes leading to a bigger % of income being spent on houses. Similarly, in a recession, falling incomes will mean people can't afford to buy and those who lose their job may fall behind on their mortgage payments and end up with their home repossessed.

- **Unemployment.** Related to economic growth is unemployment. When unemployment is rising,fewer people will be able to afford a house. But, even the fear of unemployment may discouragepeople from entering the property market.

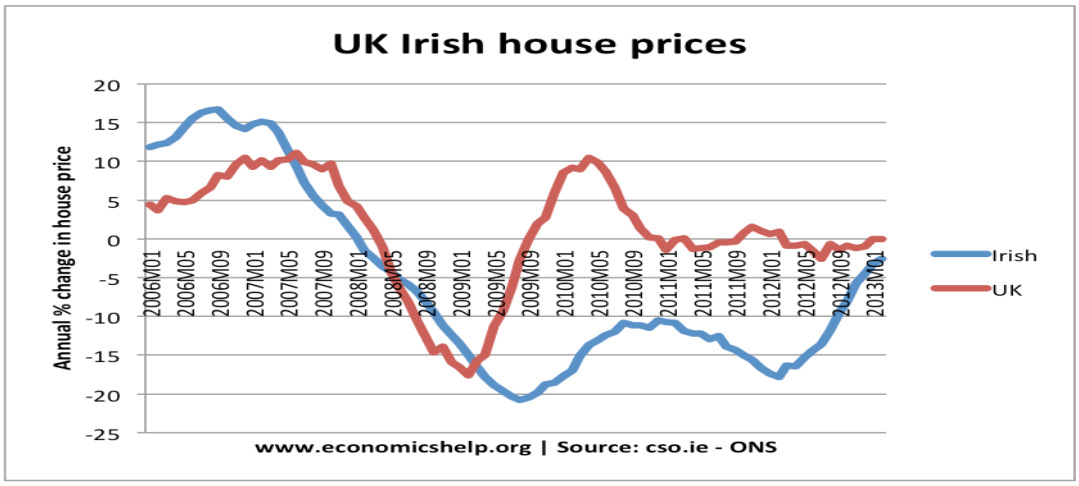
- **Interest rates.** Interest rates affect the cost of monthly mortgage payments. Aperiod of high- interest rates will increase cost of mortgage payments and will cause lower demand for buying a house. High-interest rates make renting relatively

more attractive compared to buying. Interest rates have a bigger effect if homeowners have large variable mortgages. For example, in 1990-92, the sharp rise in interest rates caused a very steep fall in UK house prices because many homeowners couldn't afford the rise in interest rates.

- **Consumer confidence**. Confidence is important for determining whether people want to take therisk of taking out a mortgage. In particular expectations towards the housing market is important; if people fear house prices could fall, people will defer buying.

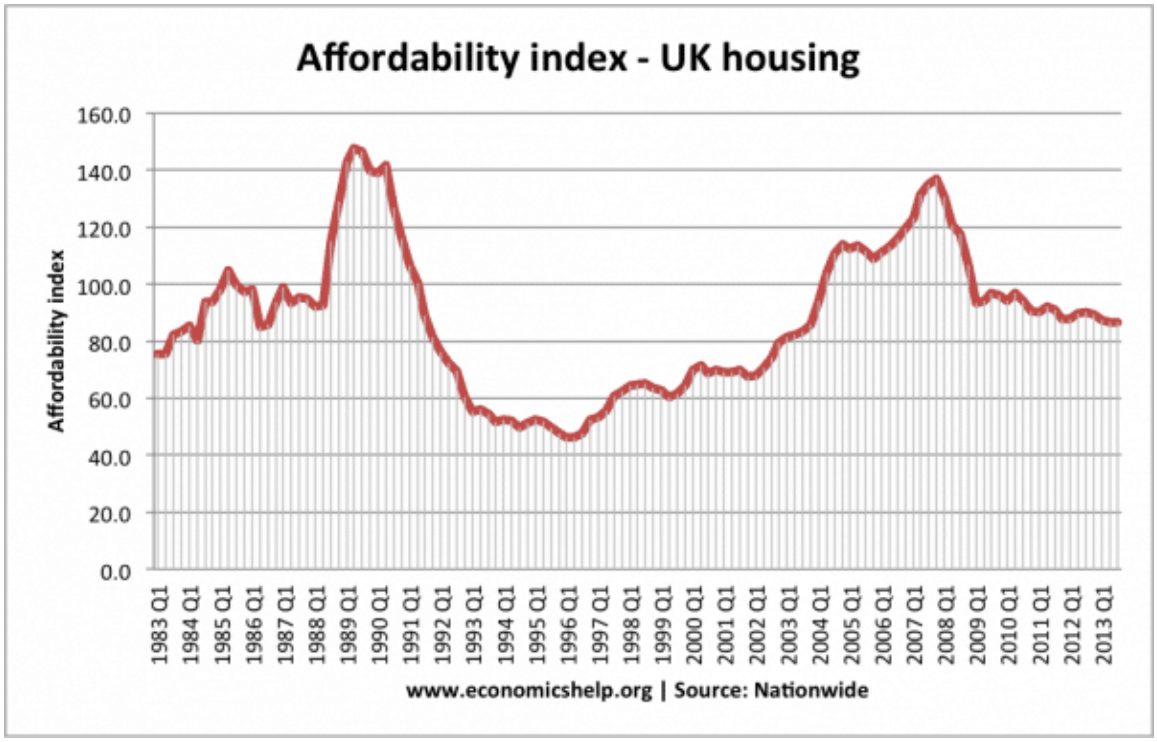
- **Mortgage availability**. In the boom years of 1996-2006, many banks were very keen to lend mortgages. They allowed people to borrow large income multiples (e.g. five times income). Also, banks required very low deposits (e.g. 100% mortgages). This ease of getting a mortgage meant that demand for housing increased as more people were now able to buy. However, since the credit crunch of 2007, banks and building societies struggled to raise funds for lending on the money markets. Therefore, they have tightened their lending criteria requiring a bigger deposit to buy a house. This has reduced the availability of mortgages and demand fell.

- **Supply**. A shortage of supply pushes up prices. Excess supply will cause prices to fall. For example, inthe Irish property boom of 1996-2006, an estimated 700,000 new houses were built. When the property market collapsed, the market was left with a fundamental oversupply. Vacancy rates reached 15%, and with supply greater than demand, prices fell.



By contrast, in the UK, housing supply fell behind demand. With a shortage, UK house prices didn't fall as much as in Ireland and soon recovered – despite the ongoing credit crunch. The supply of housing depends on existing stock and new house builds. Supply of housing tends to be quite inelastic because to get planning permission and build houses is a time-consuming process. Periods of rising house prices may not cause an equivalent rise in supply, especially in countries likethe UK, with limited land for home-building.

- **Affordability/house prices to earnings.** The ratio of house prices to earnings influences the demand. As house prices rise relative to income, you would expect fewer people to be able to afford. For example, in the 2007 boom, the ratio of house prices to income rose to 5. At this level, house prices were relatively expensive, and we saw a correction with house prices falling.



Another way of looking at the affordability of housing is to look at the percentage of take-home paythat is spent on mortgages. This takes into account both house prices, but mainly interest rates and the cost of monthly mortgage payments. In late 1989, we see housing become very unaffordable because of rising interest rates. This caused a sharp fall in prices in 1990-92.

BENEFITS FOR HOUSE PRICE PREDICTION:

Predicting house prices using machine learning can offer various benefits to both homebuyers and sellers, as well as real estate professionals and investors. Here are some of the key advantages:

- **Improved Accuracy**: Machine learning models can analyze large datasets, considering numerous features and variables to make more accurate price predictions compared to traditional methods

.

- **Faster Decision-**Making: Buyers and sellers can make quicker decisions with access to real-time or near-real-time price estimates, helping them respond to market changes promptly.

- **Data-Driven Insights**: Machine learning models can provide insights into the factors that influence property prices, helping buyers and sellers understand the market better.

- **Pricing Transparency**: Machine learning models can make the pricing process more transparent by considering various factors, such as location, property size, condition, and recent sales, which can lead to more equitable and informed pricing.

- **Personalized Recommendations**: For buyers, machine learning models can recommend properties that align with their preferences and budget, improving the house-hunting experience.

- **Competitive Edge for Sellers**: Sellers can gain a competitive advantage by pricing their properties more accurately, potentially leading to faster sales and higher offers.

- **Risk Mitigation**: Investors and real estate professionals can use machine learning models to assess the risk associated with property investments, helping them make more informed decisions.

- **Forecasting Market Trends**: Machine learning can be used to forecast housing market trends and identify emerging opportunities or risks for both buyers and investors.

- Cost Savings: Automated valuation models (AVMs) powered by machine learning can save costs associated with traditional property appraisals.
- Scalability: Machine learning models can be easily scaled to accommodate a wide range of properties, making it feasible to predict prices for multiple properties simultaneously.
- Continuous Learning: Machine learning models can adapt and improve over time as more data becomes available, leading to better predictive accuracy.
- Accessibility: With the proliferation of online real estate platforms and apps, consumers have easy access to property price predictions and market insights.

ADVANTAGES :

- The LSTM model can be tuned for various parameters such as changing the number of LSTMlayers, adding dropout value or increasing the number of epochs.

- Long Short Term Memory (LSTM)
- LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:
- The input gate: The input gate adds information to the cell state

- The forget gate: It removes the information that is no longer required by the model. Theoutput gate: Output Gate at LSTM selects the information to be shown as output.

- 1.

DISADVANTAGES:

- The quality of data used for training the prediction model is crucial. Inaccurate, incomplete, or outdated data can lead to unreliable predictions. Additionally, obtaining comprehensive and up-to-date real estate data can be challenging in some regions.
- : Cleaning and preparing the data for modeling can be time-consuming. Handling missing values, outliers, and dealing with categorical variables can be complex and may require domain knowledge.
- Overfitting occurs when a model is too complex and fits the training data too closely, leading to poor generalization on unseen data. This can result in inaccurate predictions on new house prices.
- : Choosing the right model and its hyperparameters can be challenging. More complex models might capture intricate patterns but are more prone to overfitting, while simpler models may underfit the data.
- : Selecting the most relevant features or variables for predicting house prices can be challenging. Not all available features may be useful, and engineering new features often requires domain knowledge.

CONCLUSION:

In conclusion, house price prediction is a complex and vital aspect of the real estate industry. It involves the analysis of numerous variables, including location, property features, economic trends, and market conditions. Several methods and techniques, such as regression analysis, machine learning algorithms, and deep learning models, can be employed to forecast house prices.

The accuracy of house price predictions depends on the quality and quantity of data, as well as the appropriateness of the chosen model. It is essential to continually update and refine the prediction models as new data becomes available and mark

HOUSE PRICE PREDICTION

USING MACHINE LEARNING TECHNIQUES