HOUSE PRICE PREDICTION USING MACHINE LEARNING

TABLE OF CONTENT

AIM & OBJECTIVE

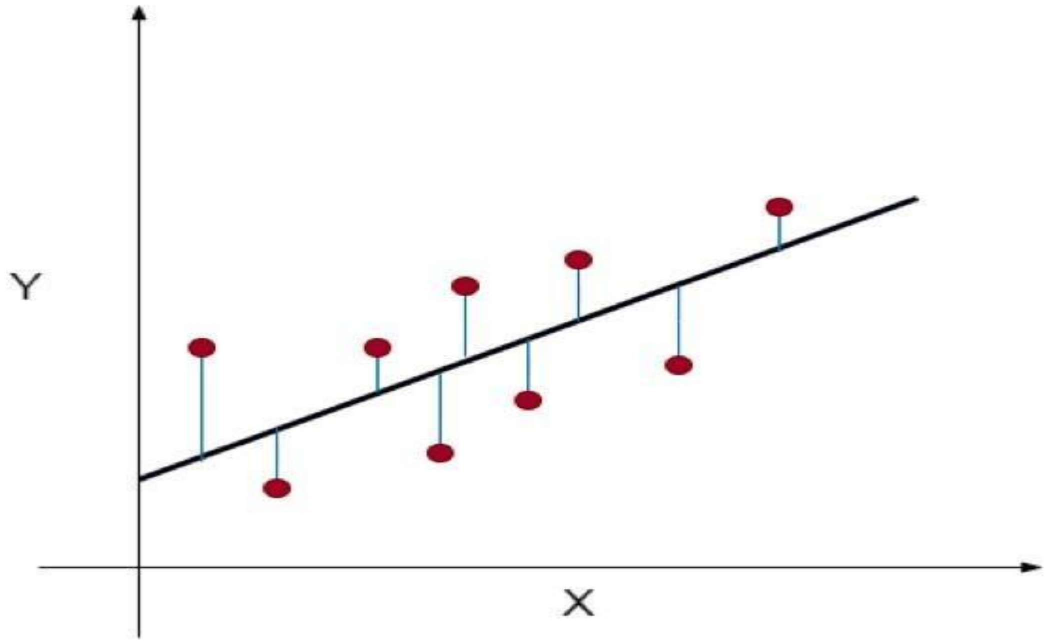
- People looking to buy a new home tend to be more conservative with their budgets and market strategies.
- This project aims to analyse various parameters like average income, average area etc. andpredict the house price accordingly.
- This application will help customers to invest in an estate without approaching an agent
- To provide a better and fast way of performing operations.
- To provide proper house price to the customers.
- To eliminate need of real estate agent to gain information regarding house prices.
- To provide best price to user without getting cheated.
- To enable user to search home as per the budget.
- The aim is to predict the efficient house pricing for real estate customers with respect totheir budgets and priorities. By analyzing previous market trends and price ranges, and alsoupcoming developments future prices will be predicted.
- House prices increase every year, so there is a need for a system to predict house prices in the future.
- 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.
- We use linear regression algorithm in machine learning for predicting the house price trends

PROPOSED SYSTEM

- Linear Regression is a supervised machine learning model that attempts to model a linear relationship between dependent variables (Y) and independent variables (X). Every evaluated observation witha model, the target (Y)'s actual value is compared to the target (Y)'s predicted value, and the major differences in these values are called residuals. The Linear Regression model aims to minimize the sum of all squared residuals. Here is the mathematical representation of thelinear regression:

$$Y= a_0+a_1X+ \varepsilon$$

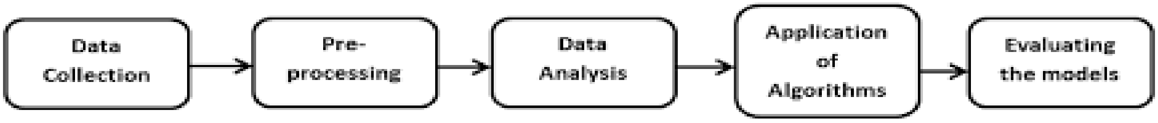
The values of X and Y variables are training datasets for the model representation of linearregression. When a user implements a linear regression, algorithms start to find the best fit line using **a_0** and **a_1**. In such a way, it becomes more accurate to actual data points; since we recognize the value of **a_0** and **a_1,** we can use a model for predicting the response.

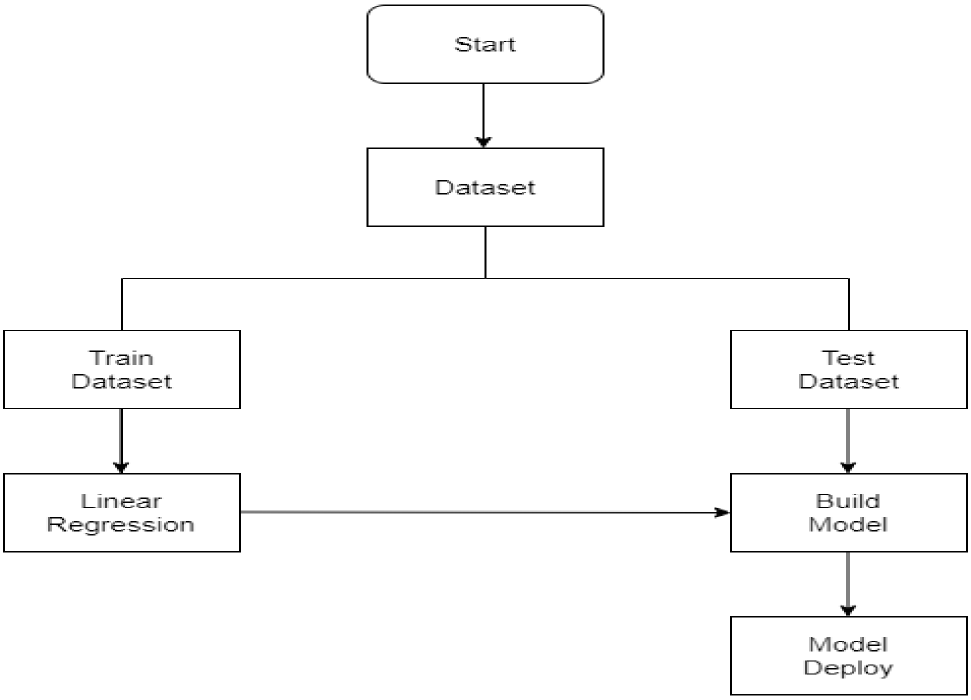


- As you can see in the above diagram, the red dots are observed values for both X and Y.
- The black line, which is called a line of best fit, minimizes a sum of a squared error.
- The blue lines represent the errors; it is a distance between the line of best fit and observed values.

- The value of the **a_1**is the slope of the black line.

BLOCK DIAGRAM





PROPOSED SYSTEM PHASES

Phase 1: Collection of data

Data processing techniques and processes are numerous. We collecteddata for USA/Mumbai real estate properties from various real estate websites. The data would be having attributes such as Location, carpet area, built-up area, age of the property, zip code, price, no of bedroomsetc. We must collect the quantitative data which is structured and categorized. Data collection is needed before any kind of machine learning research is carried out. Dataset validity is a must otherwise there is no point in analyzing the data.

Phase 2: Data preprocessing

Data preprocessing is the process of cleaning our data set. There mightbe missing values or outliers in the dataset. These can be handled by data cleaning. If there are many missing values in a variable we will drop those values or substitute it with the average value.

Phase 3: Training the model

Since the data is broken down into two modules: a Training set and Testset, we must initially train the model. The training set includes the target variable. The decision tree regressor algorithm is applied to the training data set. The Decision tree builds a regression model in the form of a tree structure.

Phase 4: Testing and Integrating with UI

The trained model is applied to test dataset and house prices are predicted. The trained model is then integrated with the front end using Flask in python

ALTERNATIVE REGRESSOR (XG BOOST REGRESSOR)

The results of the regression problems are continuous or real values. Some commonly used regression algorithms are Linear Regression and Decision Trees. There are several metrics involved in regression like root-mean-squared error (RMSE) and mean-squared-error (MAE). These are some key membersof XGBoost models, each plays an important role.

- **RMSE:** It is the square root of mean squared error (MSE).
- **MAE:** It is an absolute sum of actual and predicted differences, but it lacks mathematically, that's why it is rarely used, as compared to other metrics.

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners.

ALTERNATIVE REGRESSOR (RANDOM FOREST REGRESSOR)

Random Forest:

Random Forest is an ensemble learning method used for classification and regression tasks. It constructs multiple decision trees during training and merges them to get a more accurate and stable prediction. Key features of Random Forest include:

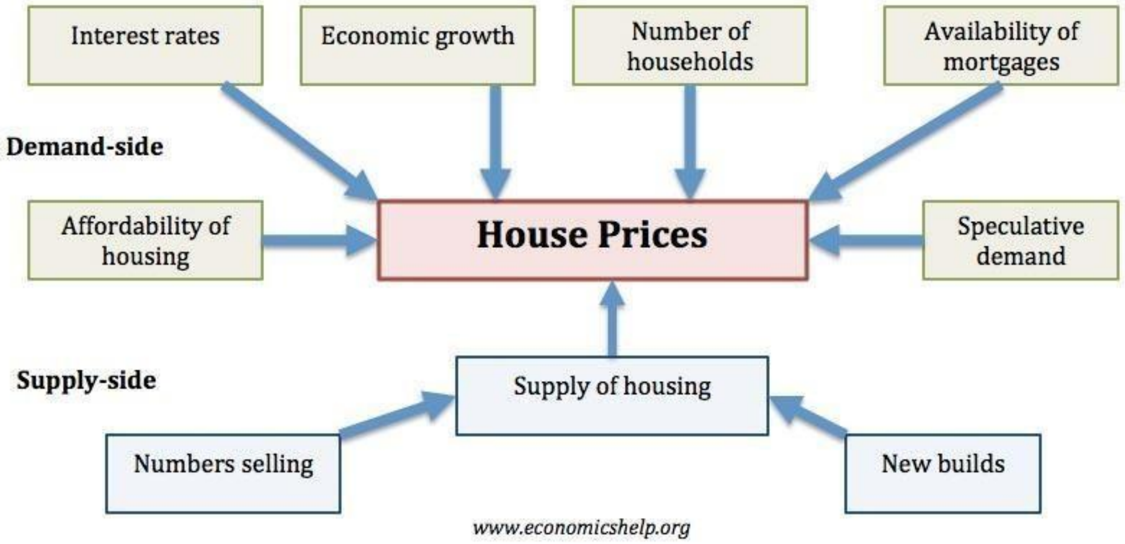
Ensemble Method: It builds multiple decision trees and merges their outputs to improve accuracy and prevent overfitting.

Random Sampling: Random Forest randomly selects subsets of the data and features to train each decision tree, reducing the risk of high variance and overfitting.

Voting or Averaging: For regression tasks, the final prediction is often the average of the predictions made by individual trees. This ensemble approach generally yields better predictive performance.

FACTORS THAT AFFECT HOUSE PRICING

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



www.economicshelp.org

- **Economic growth.** Demand for housing is dependent upon income. With higher economic 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. A period 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.

- **Geographical factors.** Many housing markets are highly geographical. For example, national house prices may be falling, but some areas (e.g. London, Oxford) may still see rising prices. Desirable areas can buck market trends as demand is high, and supply limited. For example, houses near goodschools or a good rail link may have a significant premium to other areas. This graph shows that first time buyers in London face much more expensive house prices – over 9.0 times earnings compared to the north, where house prices are only 3.3 times earnings.

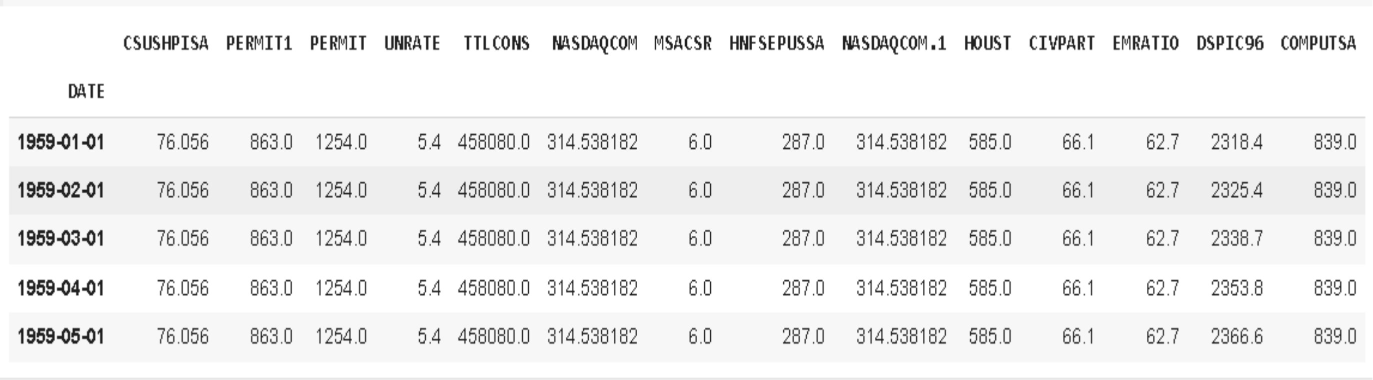
ADVANTAGE OF LSTM OVER OTHER MODELS

The LSTM model can be tuned for various parameters such as changing the number of LSTM layers, adding dropoutvalue 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. The output gate: Output Gate at LSTM selects the information to be shown as output.

EXPLANATION OF THE OUTPUT RESULTS AND THE DATASET

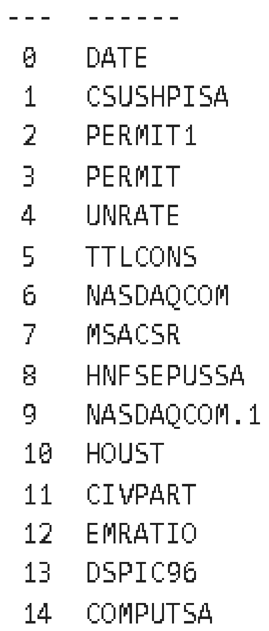


	CSUSHPISA	PERMIT1	PERMIT	UNRATE	TTLCONS	NASDAQCOM	MSACSR	HNFSEPUSSA	NASDAQCOM.1	HOUST	CIVPART	EMRATIO	DSPIC96	COMPUTSA
DATE														
1959-01-01	76.056	863.0	1254.0	5.4	458080.0	314.538182	6.0	287.0	314.538182	585.0	66.1	62.7	2318.4	839.0
1959-02-01	76.056	863.0	1254.0	5.4	458080.0	314.538182	6.0	287.0	314.538182	585.0	66.1	62.7	2325.4	839.0
1959-03-01	76.056	863.0	1254.0	5.4	458080.0	314.538182	6.0	287.0	314.538182	585.0	66.1	62.7	2338.7	839.0
1959-04-01	76.056	863.0	1254.0	5.4	458080.0	314.538182	6.0	287.0	314.538182	585.0	66.1	62.7	2353.8	839.0
1959-05-01	76.056	863.0	1254.0	5.4	458080.0	314.538182	6.0	287.0	314.538182	585.0	66.1	62.7	2366.6	839.0

First we import a sample data from sklearn library , you can get different types of sample data from Kaggle. The data taken here is the data of various parameters and the house prices in a given city called boston in the year between 1970 to 2020.

Here the data parameters are explained as follows:

Here for understanding purpose we have taken first 5 index/instance of data and printed them. In total there are 506 rows ofdata from the dataset , of which we have printed first 5 rows using head() function. There are 14 columns in total, i.e, 14 colums containing data of the place, and the 14th column is the target column which contains the house prices.

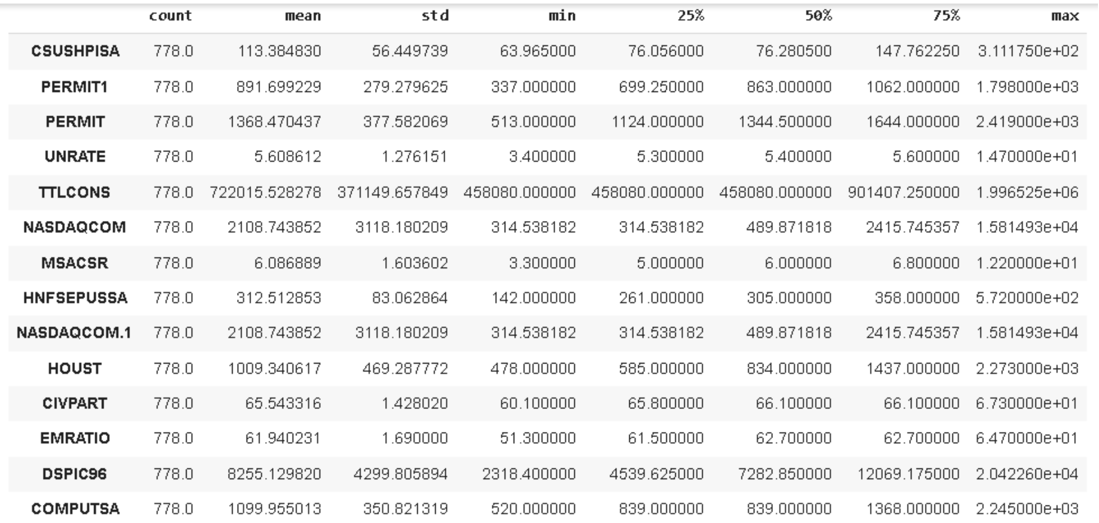


```
---  ------
 0   DATE
 1   CSUSHPISA
 2   PERMIT1
 3   PERMIT
 4   UNRATE
 5   TTLCONS
 6   NASDAQCOM
 7   MSACSR
 8   HNFSEPUSSA
 9   NASDAQCOM.1
 10  HOUST
 11  CIVPART
 12  EMRATIO
 13  DSPIC96
 14  COMPUTSA
```

Then we check if our data has some null values i.e missing values. Since if the data is incomplete , then there will be error during processing state which may lead to loss of accuracy in predicting model. Here in our given data , there is nomissing value as we can see.

Since our data contains no missing value, the program will skip the dropping phase in data processing, where data isdropped to increase accuracy and fit missing values in a way so that it is suitable for modelling.

Next we try to describe the data in such a way so that both people and machine find it easy to understand the given data . In order to do thiswe use the describe() function.



	count	mean	std	min	25%	50%	75%	max
CSUSHPISA	778.0	113.384830	56.449739	63.965000	76.056000	76.280500	147.762250	3.111750e+02
PERMIT1	778.0	891.699229	279.279625	337.000000	699.250000	863.000000	1062.000000	1.798000e+03
PERMIT	778.0	1368.470437	377.582069	513.000000	1124.000000	1344.500000	1644.000000	2.419000e+03
UNRATE	778.0	5.608612	1.276151	3.400000	5.300000	5.400000	5.600000	1.470000e+01
TTLCONS	778.0	722015.528278	371149.657849	458080.000000	458080.000000	458080.000000	901407.250000	1.996525e+06
NASDAQCOM	778.0	2108.743852	3118.180209	314.538182	314.538182	489.871818	2415.745357	1.581493e+04
MSACSR	778.0	6.086889	1.603602	3.300000	5.000000	6.000000	6.800000	1.220000e+01
HNFSEPUSSA	778.0	312.512853	83.062864	142.000000	261.000000	305.000000	358.000000	5.720000e+02
NASDAQCOM.1	778.0	2108.743852	3118.180209	314.538182	314.538182	489.871818	2415.745357	1.581493e+04
HOUST	778.0	1009.340617	469.287772	478.000000	585.000000	834.000000	1437.000000	2.273000e+03
CIVPART	778.0	65.543316	1.428020	60.100000	65.800000	66.100000	66.100000	6.730000e+01
EMRATIO	778.0	61.940231	1.690000	51.300000	61.500000	62.700000	62.700000	6.470000e+01
DSPIC96	778.0	8255.129820	4299.805894	2318.400000	4539.625000	7282.850000	12069.175000	2.042260e+04
COMPUTSA	778.0	1099.955013	350.821319	520.000000	839.000000	839.000000	1368.000000	2.245000e+03

Counts refers to the number of instances of data in each column i.e 506 since there are 506 rows of data for each columnMean refers to mean value of data in given colum.

Std means the standard value i.e the most common value in given set of data for a particular column.

Min refers the least data value in each column.

Max refers to the maximum data value in each column.

25% refers that 25 percentile of the data in that column is equal to or below that value.

Next we try to understand the correlation between the different values, in order to do that, the best way is by using heat map. Heat map is a representation of data in the form of a map or diagram in which data values are represented as colours.

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate)

There are two types of correlation, they are:

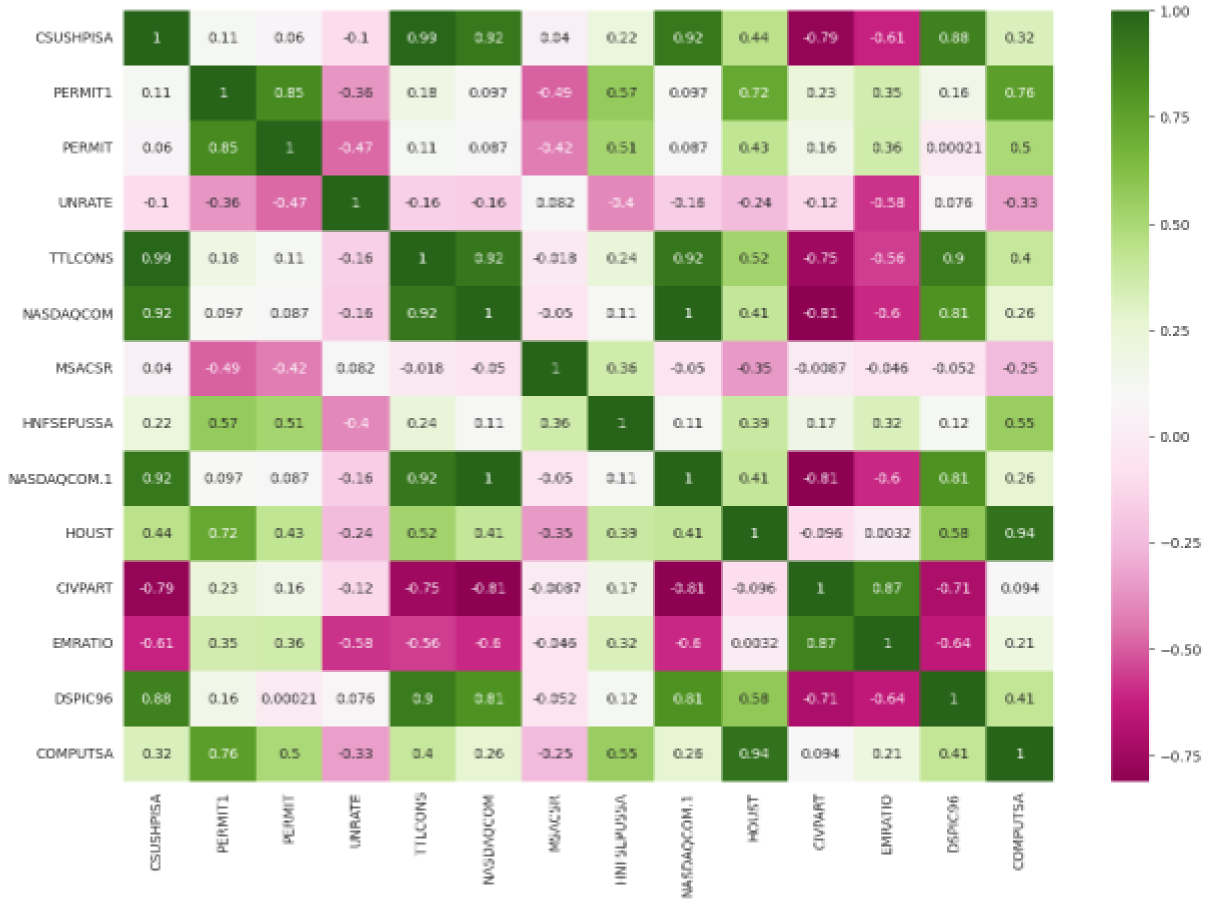
1. Positive correlation: A positive correlation is a relationship between two variables that move in tandem—that is, inthe same direction. A positive correlation exists whenone variable decreases as the other variable decreases, or onevariable increases while the other increases.

2. Negative correlation: Negative correlation is a relationship between two variables in which one variable increasesas the other decreases, and vice versa.

   In statistics, a perfect negative correlation is represented by the value -1.0, while a 0 indicates no correlation, and +1.0 indicates a perfect positive correlation. A perfect negative correlation means the relationship that existsbetween two variables is exactly opposite all of the time. These are two types of correlation are representednumerically and as well as by shade of colour in the heat map.

HEATMAP – for better understanding of which place is best suited for individual personal preference based on given dataset. This uses correlation concept

```
<Axes: >
```



	CSUSHPISA	PERMIT1	PERMIT	UNRATE	TTLCONS	NASDAQCOM	MSACSR	HNFSEPUSSA	NASDAQCOM.1	HOUST	CIVPART	EMRATIO	DSPIC96	COMPUTSA
CSUSHPISA	1	0.11	0.06	-0.1	0.99	0.92	0.04	0.22	0.92	0.44	-0.79	-0.61	0.88	0.32
PERMIT1	0.11	1	0.85	-0.36	0.18	0.097	-0.49	0.57	0.097	0.72	0.23	0.35	0.16	0.76
PERMIT	0.06	0.85	1	-0.47	0.11	0.087	-0.42	0.51	0.087	0.43	0.16	0.36	0.00021	0.5
UNRATE	-0.1	-0.36	-0.47	1	-0.16	-0.16	0.082	-0.4	-0.16	-0.24	-0.12	-0.58	0.076	-0.33
TTLCONS	0.99	0.18	0.11	-0.16	1	0.92	-0.018	0.24	0.92	0.52	-0.75	-0.56	0.9	0.4
NASDAQCOM	0.92	0.097	0.087	-0.16	0.92	1	-0.05	0.11	1	0.41	-0.81	-0.6	0.81	0.26
MSACSR	0.04	-0.49	-0.42	0.082	-0.018	-0.05	1	0.36	-0.05	-0.35	-0.0087	-0.046	-0.052	-0.25
HNFSEPUSSA	0.22	0.57	0.51	-0.4	0.24	0.11	0.36	1	0.11	0.39	0.17	0.32	0.12	0.55
NASDAQCOM.1	0.92	0.097	0.087	-0.16	0.92	1	-0.05	0.11	1	0.41	-0.81	-0.6	0.81	0.26
HOUST	0.44	0.72	0.43	-0.24	0.52	0.41	-0.35	0.39	0.41	1	-0.096	0.0032	0.58	0.94
CIVPART	-0.79	0.23	0.16	-0.12	-0.75	-0.81	-0.0087	0.17	-0.81	-0.096	1	0.87	-0.71	0.094
EMRATIO	-0.61	0.35	0.36	-0.58	-0.56	-0.6	-0.046	0.32	-0.6	0.0032	0.87	1	-0.64	0.21
DSPIC96	0.88	0.16	0.00021	0.076	0.9	0.81	-0.052	0.12	0.81	0.58	-0.71	-0.64	1	0.41
COMPUTSA	0.32	0.76	0.5	-0.33	0.4	0.26	-0.25	0.55	0.26	0.94	0.094	0.21	0.41	1

ALGORITHM BRIEF OUTLINE

1. Import the python libraries that are required for house price prediction using linear regression. Example: numpy is used for convention of data to 2d or 3d array format which is required for linearregression model ,matplotlib for plotting the graph , pandas for readingthe data from source and manipulation that data, etc.

2. First Get the value from source and give it to a data frame and thenmanipulate this data to required form using head(),indexing, drop().

3. Next we have to train a model, its always best to spilt the data intotraining data and test data for modelling.
4. Its always good to use shape() to avoid null spaces which will cause error during modelling process.

5. Its good to normalize the value since the values are in very large quantity for house prices , for this we may use minmaxscaler to reducethe gap between prices so that its easy and less time consuming for comparing and values.range usually specified is between 0 to 1 using fittransform.

6. Then we have to make few imports from keras: like sequential for initializing the network,lstm to add lstm layer, dropout to prevent overfitting of lstm layers, dense to add a densely connected networklayer for output unit.

7. In lstm layer declaration its best to declare the unit, activiation,returnsequence.

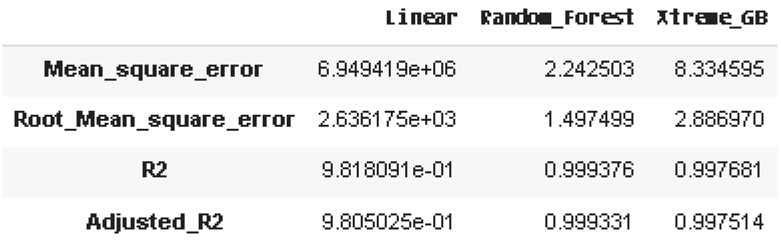
8. To compile this model its always best to use adam optimizer and set the loss as required for the specific data.

9. We can fit the model to run for a number of epochs. Epochs are the number of times the learning algorithm will work through the entire training set.

10. Then we convert the values back to normal form by using inverse minimal scale by scale factor.
11. Then we give a test data(present data)to the trained model to get the predicted value(future data).

12. Then we can use matplotlib to plot a graph comparing the test andpredicted value to see the increase/decrease rate of values in each time of the year in a particular place. Based on this people will know when its best time to sell or buy a place in a given location.

COMPARISM OF DIFFRENT MODEL



	Linear	Random_Forest	Xtreme_GB
Mean_square_error	6.949419e+06	2.242503	8.334595
Root_Mean_square_error	2.636175e+03	1.497499	2.886970
R2	9.818091e-01	0.999376	0.997681
Adjusted_R2	9.805025e-01	0.999331	0.997514

These values represent the performance of three different machine learning models in predicting the S&P/Case-Shiller Home Price Index. The models are:

- Linear Random Forest: A linear regression model that uses multiple decision trees to make predictions.
- Xtreme_GB: A gradient boosting machine model that uses multiple weak learners to make predictions.
- Xtreme_GB (tuned): A tuned version of the Xtreme_GB model that has been optimized for performance.

The following is a brief explanation of each column in the table:

- Model: The name of the machine learning model.
- Mean Square Error (MSE): A measure of how close the predicted values are to the actual values. A lower MSE indicates better performance.
- Root Mean Square Error (RMSE): The square root of the MSE. It is interpreted in the same way as the MSE.
- R-squared (R2): A measure of how well the model explains the variation in the data. A higher R2 indicates better performance.
- Adjusted R-squared: A variation of the R2 statistic that takes into account the number of parameters in the model. It is a more accurate measure of model performance when comparing models with different numbers of parameters.

Based on the values in the table, the Xtreme_GB (tuned) model has the best performance, with the lowest MSE and RMSE and the highest R2 and adjusted R-squared. This suggests that the tuned Xtreme_GB model is the best model for predicting the S&P/Case-Shiller Home Price Index.

CONCLUSION

- Upon analysing the data through three distinct processes, namely EDA (Exploratory Data Analysis), correlation matrix examination, and machine learning modelling, a set of common and highly important features emerge. These features are:

- Personal Income & Outlays: It directly reflects changes in home prices, which have a significant impact on the S&P Home Price Index (HPI).

- Employment rate, which often leads to increased demand for housing and higher home prices, influencing the S&P HPI.

- Total Construction Spending (TTLCONS): It signifies the level of construction activity, which affects housing supply and demand, consequently impacting the S&P HPI.

- New privately own housing Directly measures property prices, affecting the value of the S&P HPI.

- CPI-Adjusted Price: Reflects changes in housing costs, impacting home prices and the S&P HPI.

- NASDAQ Composite Index (NASDAQCOM): The performance of tech companies can influence economic growth, job creation, and housing demand, affecting the S&P HPI.

- Monthly Supply of New Houses (MSACSR): The supply of new houses relative to demand impacts home prices, which, in turn, influences the S&P HPI.

SOFTWARE TOOLS

- colob
- R Square
- Adjusted R Square
- MSE
- RMSE
- MAE

REFERENCES

- Real Estate Price Prediction with Regression and Classification, CS 229 Autumn2016 Project Final Report
- Gongzhu Hu, Jinping Wang, and Wenying Feng Multivariate Regression Modellingfor Home Value Estimates with Evaluation using Maximum Information Coefficient
- Byeonghwa Park , Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction , Volume 42, Pages 2928-2934 [4] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, 2015. Introduction to Linear Regression Analysis.
- Iain Pardoe, 2008, Modelling Home Prices Using Realtor Data
- Aaron Ng, 2015, Machine Learning for a London Housing Price Prediction Mobile Application
- Wang, X., Wen, J., Zhang, Y.Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. Optik-International Journal for Light and Electron Optics, 125(3), 14391443