PREDICTING HOUSE PRICES USING MACHINE LEARNING

CONTENT

AIM & OBJECTIVE

People looking to buy a new home tend to be more conservative with their budgets and marketstrategies. 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).

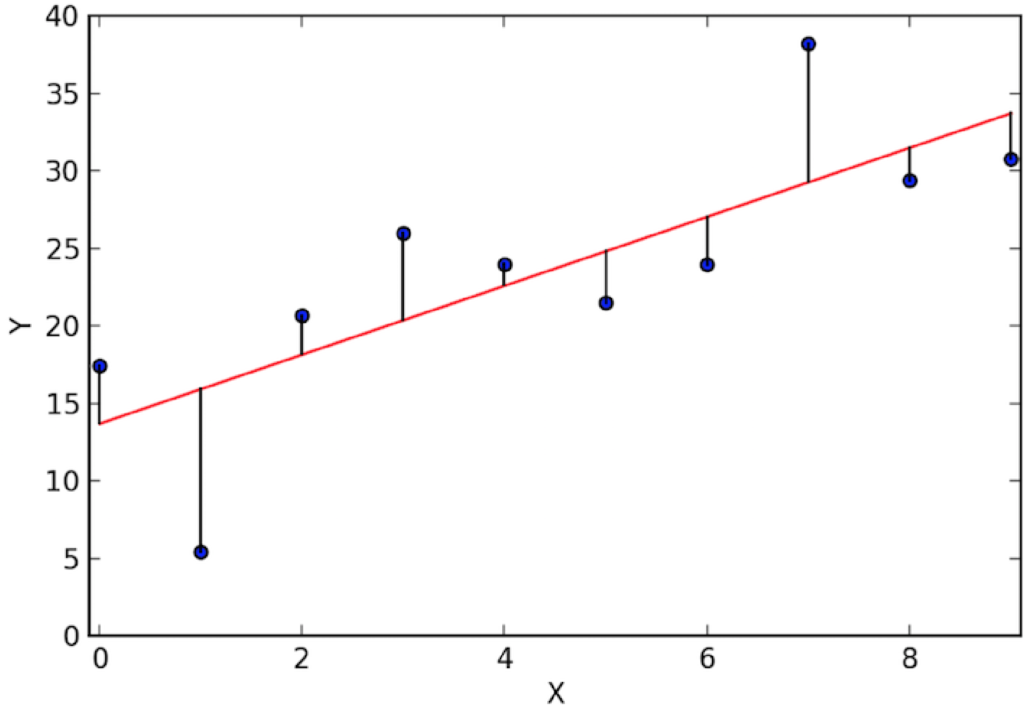
Every evaluated observation witha model, the target (Y)'s actual value is compared to the target (Y)'s predicted value, and the major differences in these values are called residuals. The Linear Regression model aims to minimize the sum of all squared residuals. Here is the mathematical representation of thelinear regression:

$$Y = a0 + a1X + \varepsilon$$

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 a0 and a1. In such a way,

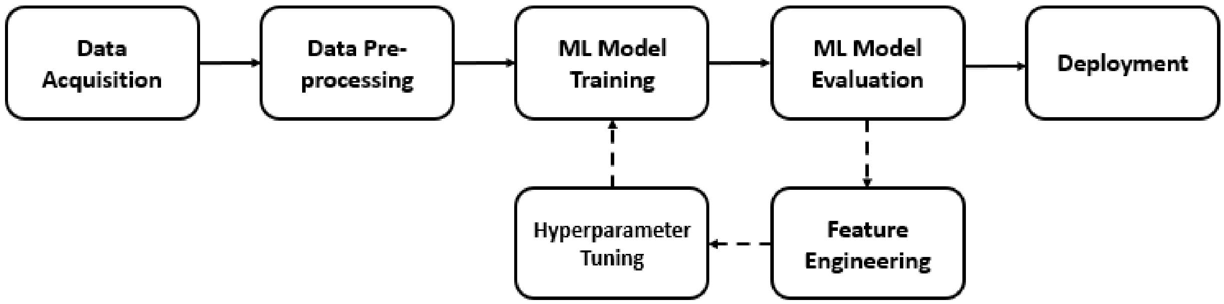
it becomes more accurate to actual data

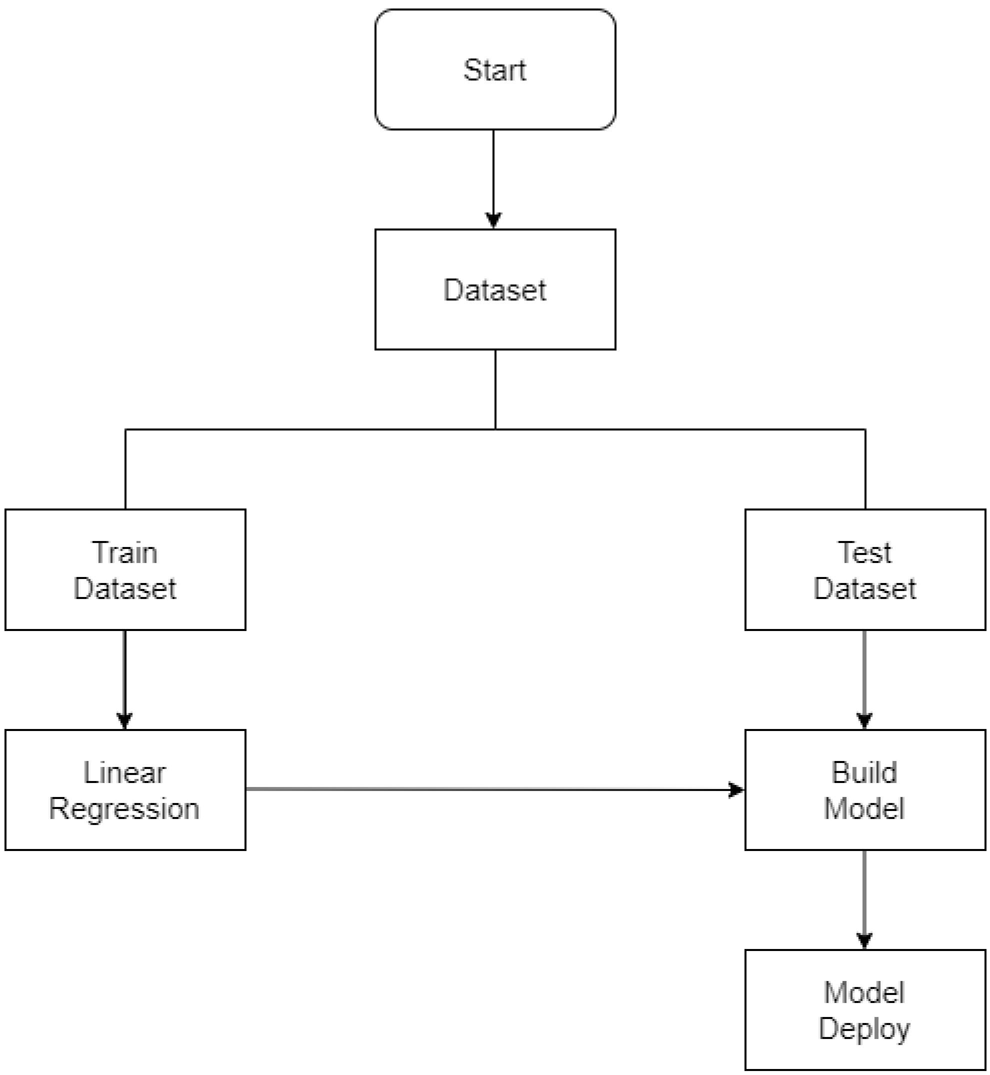
points; since we recognize the value of a0 and a1, we can use a model for predicting the response.



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.

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 members of XGBoost models, each plays an important role.

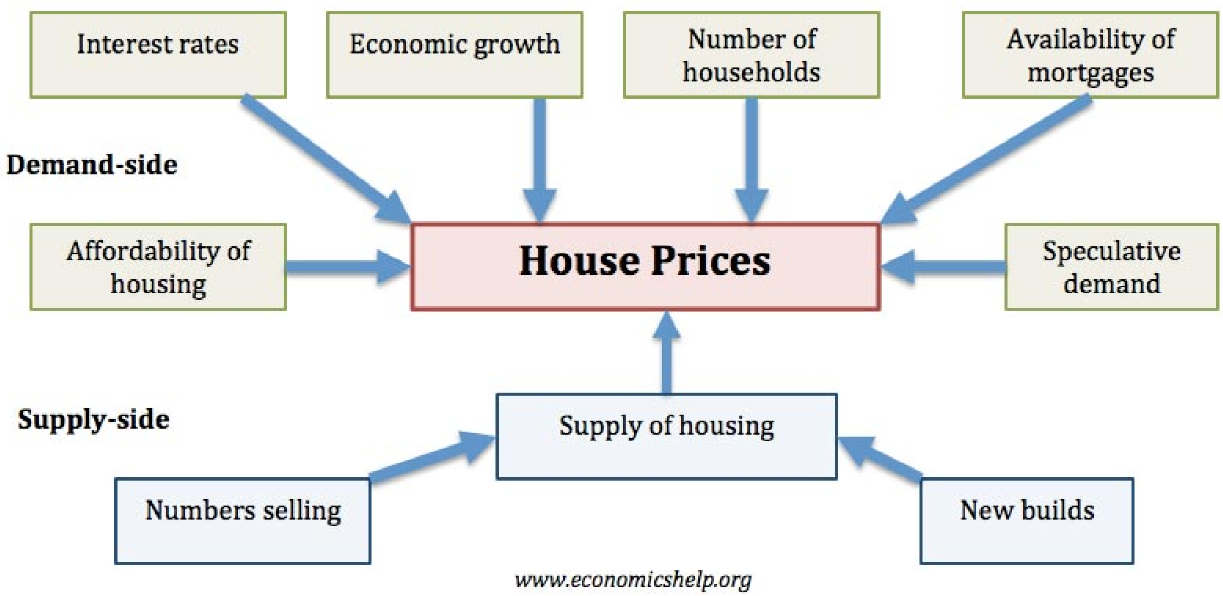
RMSE: It is the square root of mean squared error (MSE).

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.

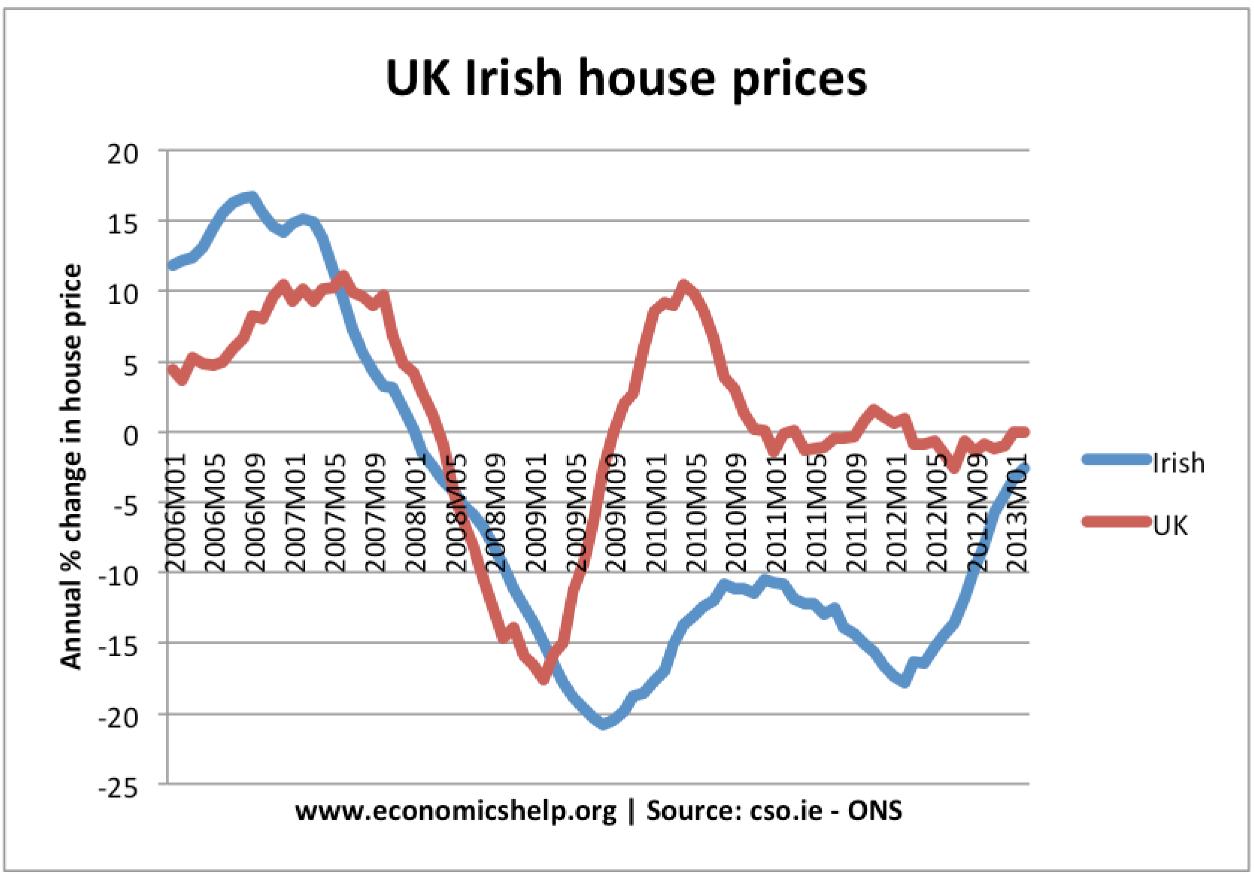
FACTORS THAT AFFECT HOUSE PRICING

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



• 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.

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’])

coeff_df
```

14

```
from keras.layersimport 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)))
```

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

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, 13 colums containing data of the place, and the 14th column is the target column which contains the house prices. 18 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.

	0	1	2	3	4
id	7129300520	6414100192	5631500400	2487200875	1954400510
date	10/13/2014	12/9/2014	2/25/2015	12/9/2014	2/18/2015
price	221900	538000	180000	604000	510000
bedrooms	3	3	2	4	3
bathrooms	1	2.25	1	3	2
sqft_living	1180	2570	770	1960	1680
sqft_lot	5650	7242	10000	5000	8080
floors	1	2	1	1	1
waterfront	0	0	0	0	0
view	0	0	0	0	0
condition	3	3	3	5	3
grade	7	7	6	7	8
sqft_above	1180	2170	770	1050	1680
sqft_basement	0	400	0	910	0
yr_built	1955	1951	1933	1965	1987
yr_renovated	0	1991	0	0	0
zipcode	98178	98125	98028	98136	98074
lat	47.5112	47.721	47.7379	47.5208	47.6168
long	-122.257	-122.319	-122.233	-122.393	-122.045
sqft_living15	1340	1690	2720	1360	1800
sqft_lot15	5650	7639	8062	5000	7503

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 when one 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. 20 HEATMAP – for better understanding of which place is best suited for individual personal preference based on given dataset. This uses correlation concept

Here the varible x contains the value of the first 13 columns i.e the parameters that are required for calculating and predicting the house prices. The varible y contains the 14th column values which are the house prices.

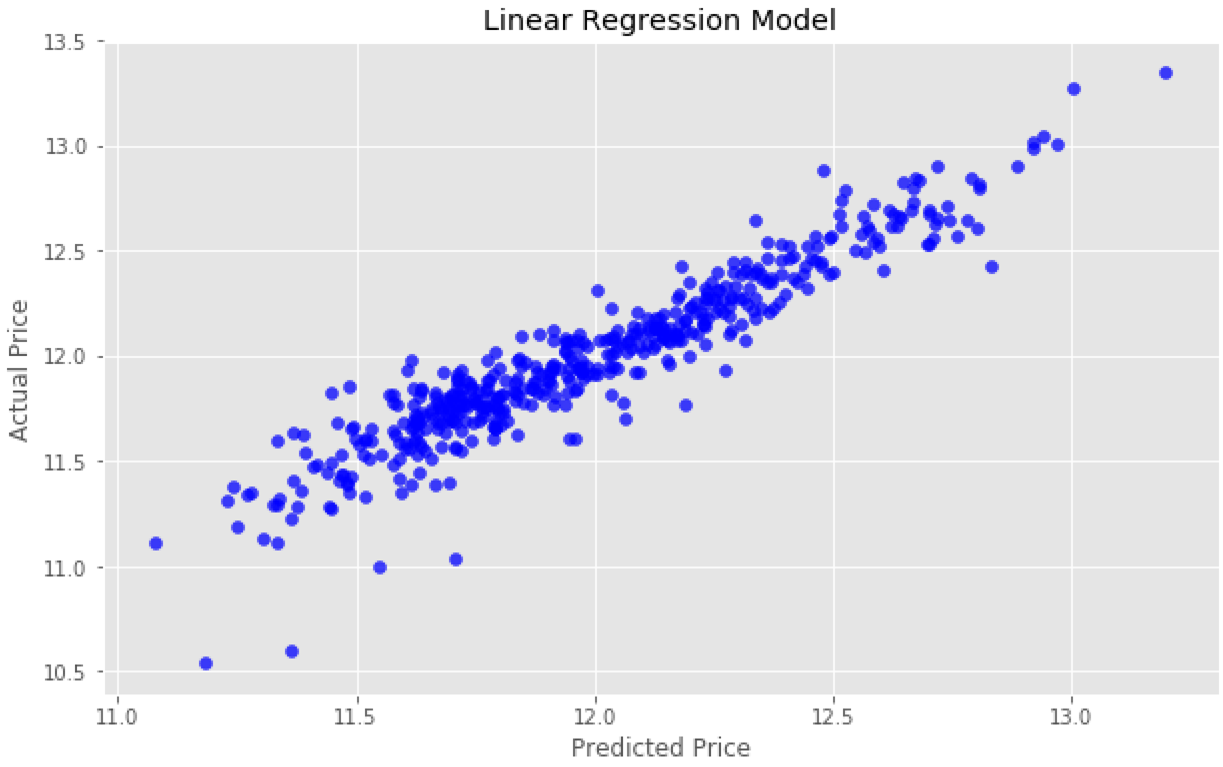
First we predict the values in y using the values in x . Then we compare the actual prices and predicted prices by using scatter plot. Then we find the r square error and mean square error between them . If the errors is less enough then we proceed for testing of the model since the training phase is over. If the error is large , thenwe use optimizers like adam, and repeat drop and fitting process for a set number of epochs to reduce the error.

The r square error or mean square error for good accuracy of the model in predicting the data is indicatednumerically also.

A model is good if these error values are less then 5.

Then during testing process we predict the future house prices using present and past data parameters ofhouses in an location. Then we plot this graphically as a house price over time graph.

For training the model , the error needs to be minimum for greater accuracy of model. The error between the actual and predicted price is plotted graphically using scatter plot. Here we can see that error is minimum sincethe data points of actual and predicted value are close to each other



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. 24

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).

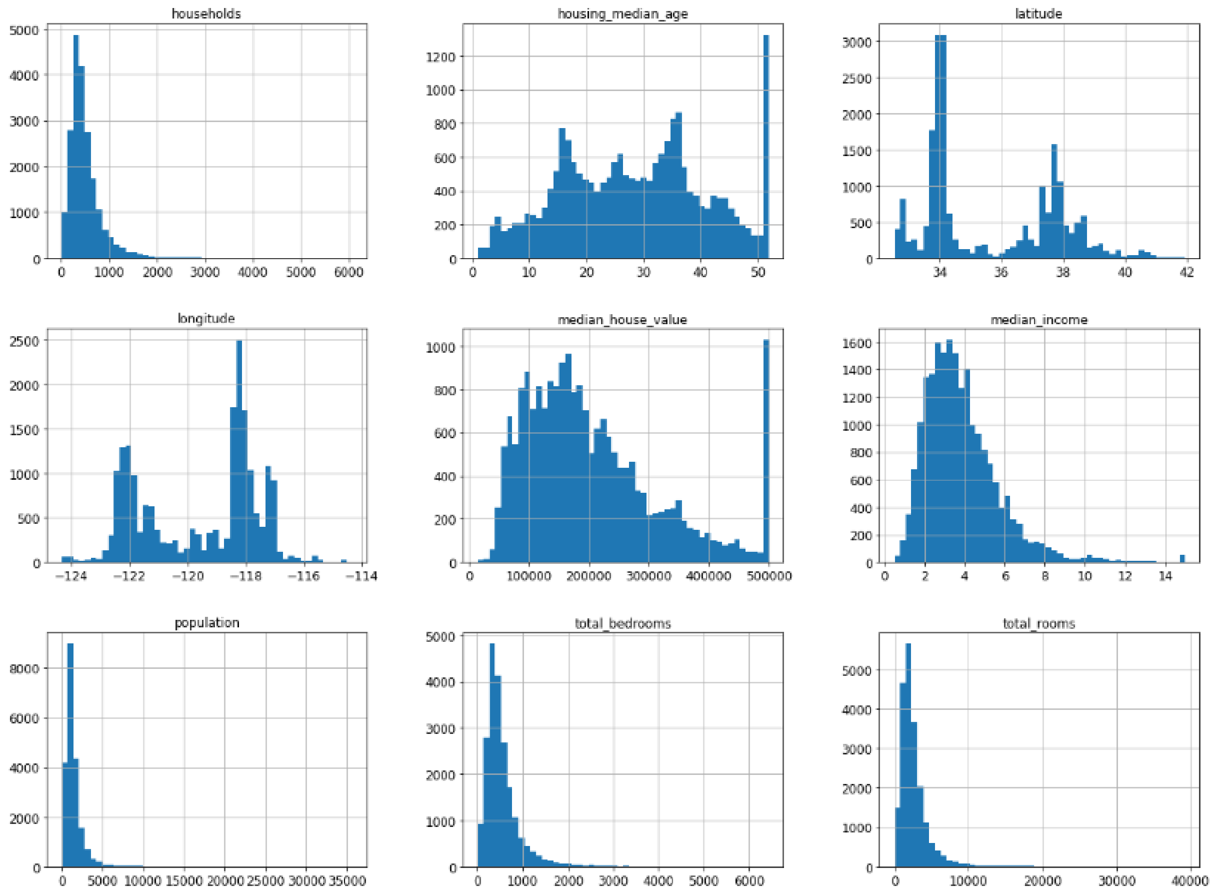
12. Then we can use matplotlib to plot a graph comparing the test andpredicted value to see the increase/decrease rate of values in each time of the year in a particular place. Based on this people will know when its best time to sell or buy a place in a given location.

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr.N.M.Nandhitha , Dean, School of electronics and Dr.T.Ravi, Head of the Department, Dept. of ECE for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Dr.S.Lalithakumari for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work. I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of ECE who were helpful in many ways for the completion of the project.


households
0 1000 2000 3000 4000 5000 6000
housing_median_age
0 10 20 30 40 50
latitude
34 36 38 40 42
longitude
−124 −122 −120 −118 −116 −114
median_house_value
0 100000 200000 300000 400000 500000
median_income
0 2 4 6 8 10 12 14
population
0 5000 10000 15000 20000 25000 30000 35000
total_bedrooms
0 1000 2000 3000 4000 5000 6000
total_rooms
0 10000 20000 30000 40000

CONCLUSION

Thus the machine learning model to predict the house price based ongiven dataset is executed successfully using xg regressor (a upgraded/slighted boosted form of regular linear regression, this gives lesser error). This model further helps people understand whether this place is more suited for them based on heatmap correlation. It also helps people looking to sell a house at best time for greater profit. Any house price in any location can be predicted with minimum errorby giving appropriate dataset.