House Price Prediction

By - Ganesh Walimbe

Problem statement

The goal is to understand the relationship between house features and how these variables affect the house price. Using more than one model, predict the price of the house using the given dataset. Please compare the accuracy of the models along with the drawbacks of each technique's assumptions before recommending the final prediction model.

Agenda

- Introduction to data
- Data preprocessing
- Performing exploratory data analysis
- Building an prediction model to predict house prices using supervised learning models.
- Conclusions

Introduction to data

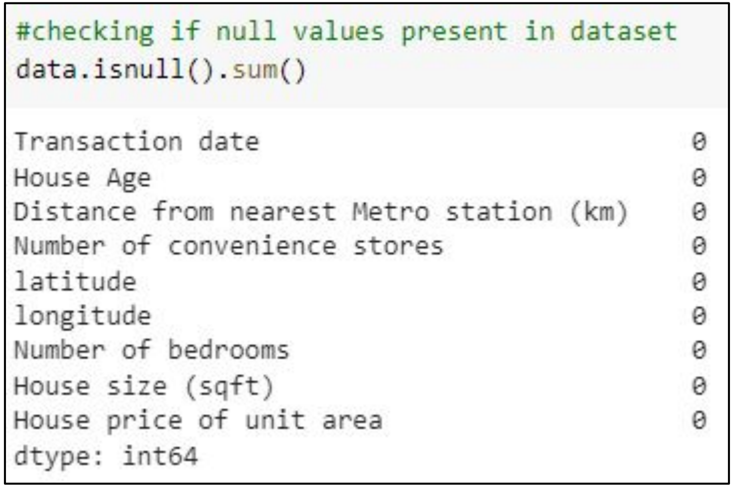
Data Description :

The data is containing 9 variables with 414 rows and those variables are as follows

- Transaction date
- House Age
- Distance from nearest Metro station (km)
- Number of convenience stores
- latitude
- longitude
- Number of bedrooms
- House size (sqft)
- House price of unit area

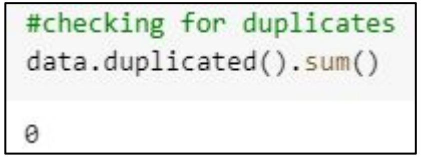
Data preprocessing

- There were no missing values present in dataset.
- Also there were no duplicates values



```
#checking if null values present in dataset
data.isnull().sum()
```

```
Transaction date                            0
House Age                                   0
Distance from nearest Metro station (km)    0
Number of convenience stores                0
latitude                                    0
longitude                                   0
Number of bedrooms                          0
House size (sqft)                           0
House price of unit area                    0
dtype: int64
```



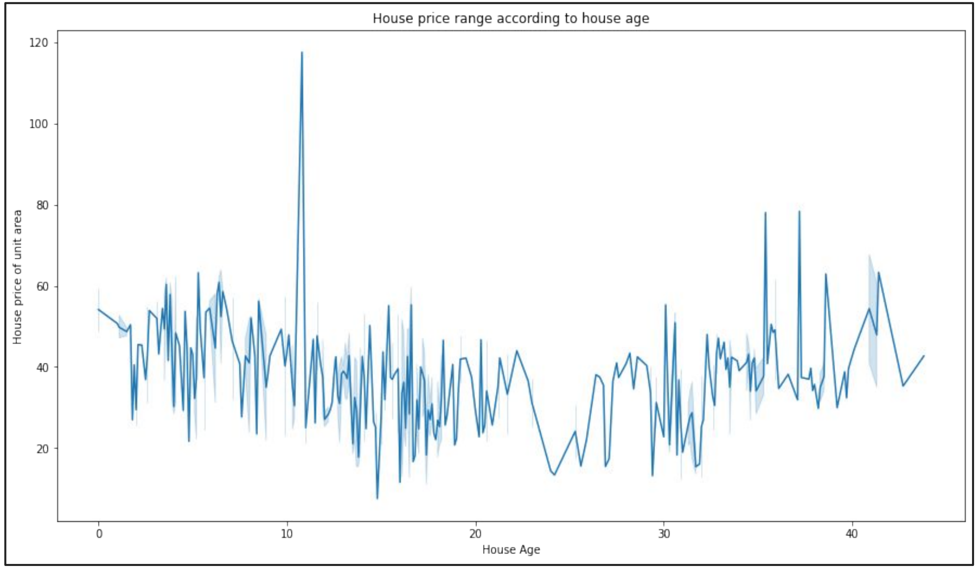
```
#checking for duplicates
data.duplicated().sum()
```

```
0
```

Exploratory data analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations

House price v/s house age
From House price v/s house age lineplot we can see that house price is maximum at 10 year old house

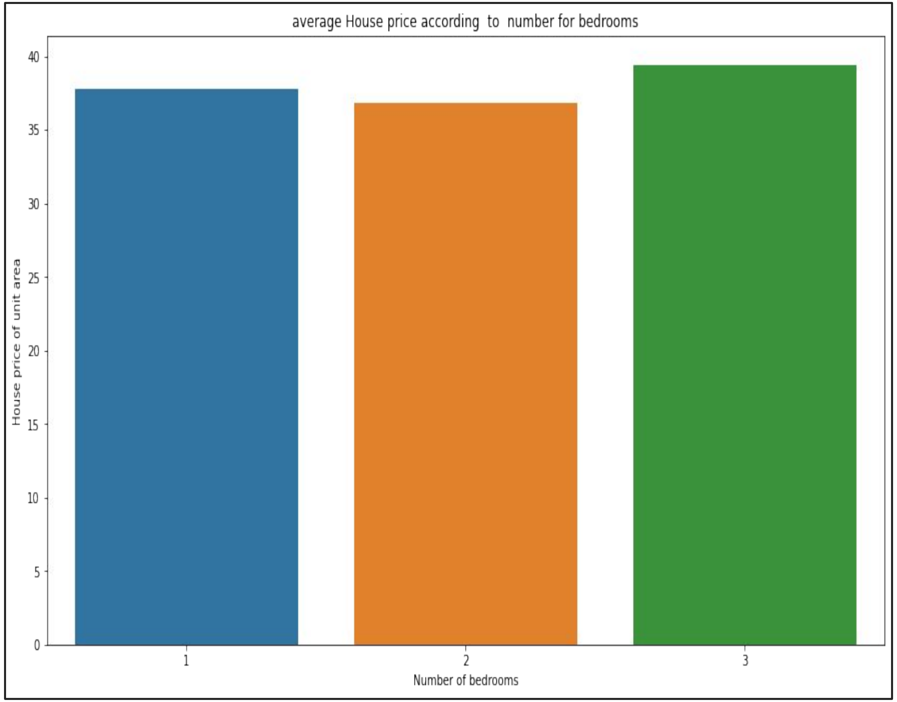


Exploratory data analysis

House price v/s number of bedrooms

From the barplot we can see that

1. Average price of houses with 1 bedroom is 37.74 per unit area
2. Average price of houses with 2 bedroom is 36.78 per unit area
3. Average price of houses with 3 bedroom is 39.43 per unit area

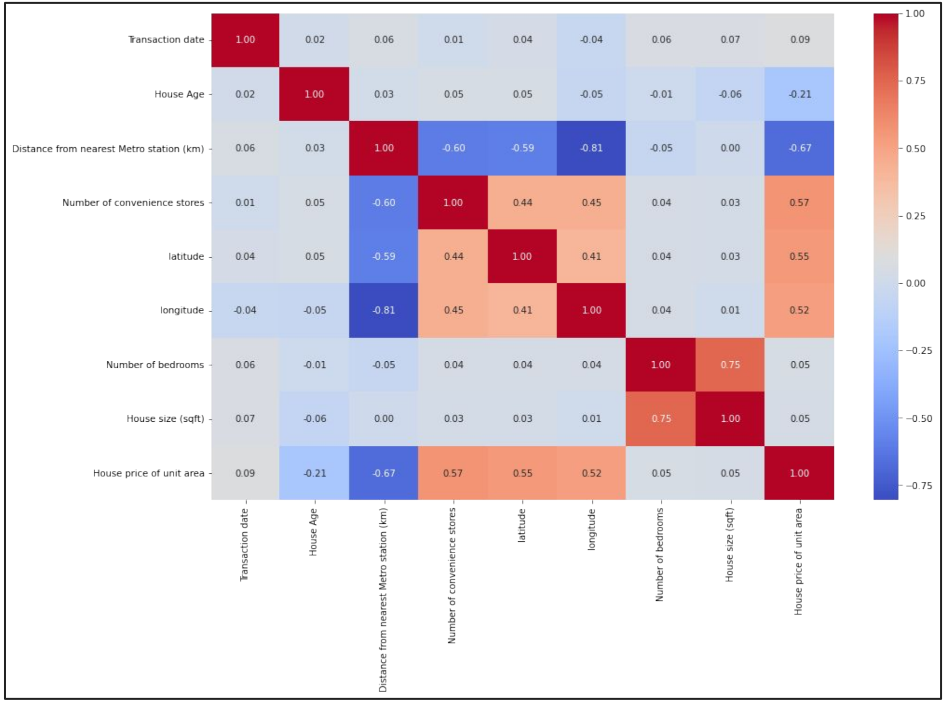


Exploratory data analysis

From correlationplot we can see that

1.There is highly positive correlationship between House price per unit area and Number of convenience stores,latitude ,longitude

2.There is highly negatively correlationship between distance from nearest metro station and House price per unit area



	Transaction date	House Age	Distance from nearest Metro station (km)	Number of convenience stores	latitude	longitude	Number of bedrooms	House size (sqft)	House price of unit area
Transaction date	1.00	0.02	0.06	0.01	0.04	-0.04	0.06	0.07	0.09
House Age	0.02	1.00	0.03	0.05	0.05	-0.05	-0.01	-0.06	-0.21
Distance from nearest Metro station (km)	0.06	0.03	1.00	-0.60	-0.59	-0.81	-0.05	0.00	-0.67
Number of convenience stores	0.01	0.05	-0.60	1.00	0.44	0.45	0.04	0.03	0.57
latitude	0.04	0.05	-0.59	0.44	1.00	0.41	0.04	0.03	0.55
longitude	-0.04	-0.05	-0.81	0.45	0.41	1.00	0.04	0.01	0.52
Number of bedrooms	0.06	-0.01	-0.05	0.04	0.04	0.04	1.00	0.75	0.05
House size (sqft)	0.07	-0.06	0.00	0.03	0.03	0.01	0.75	1.00	0.05
House price of unit area	0.09	-0.21	-0.67	0.57	0.55	0.52	0.05	0.05	1.00

Building prediction models

To build the prediction we first have to perform training and test on the independent and dependent variables .I am going split data into 80% for training and 20% for testing ,after that i am going to apply various regression models on training and testing data i will find out which model is performing best on the dataset

.

The models i am going to use are as follows :

- Multiple linear regression
- Lasso regression
- Random forest regressor

Multiple linear regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables.

Assumptions:

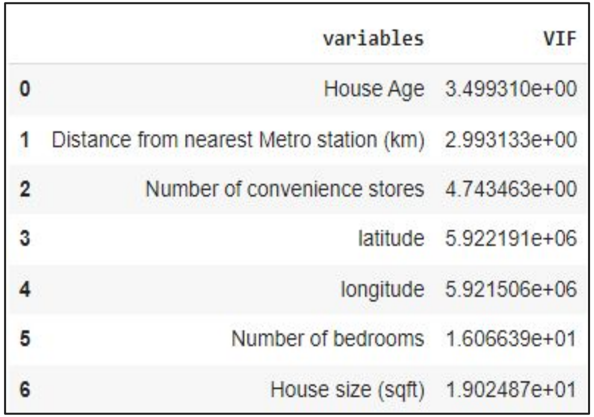
- The relation between the dependent and independent variables should be almost linear.
- Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of “best fit”.
- There should be homoscedasticity or equal variance in a regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X)
- There should not be multicollinearity in regression model. Multicollinearity generally occurs when there are high correlations between two or more independent variables.

Multiple linear regression

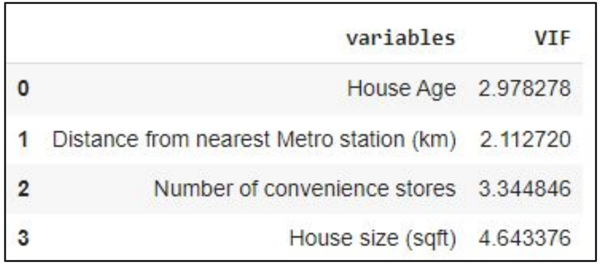
Checking assumptions of MLR

Checking for multicollinearity using VIF

- By using VIF we can see that there is high multicollinearity among the variables from
- We can remove variables with highest multicollinearity one by one by again checking multicollinearity after each removal to reduce multicollinearity.
- After removing highly correlated variables we can from dataframe that multicollinearity has been reduced so we can apply MLR on remaining variables



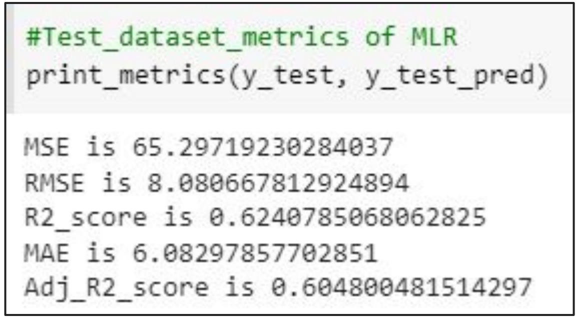
	variables	VIF
0	House Age	3.499310e+00
1	Distance from nearest Metro station (km)	2.993133e+00
2	Number of convenience stores	4.743463e+00
3	latitude	5.922191e+06
4	longitude	5.921506e+06
5	Number of bedrooms	1.606639e+01
6	House size (sqft)	1.902487e+01



	variables	VIF
0	House Age	2.978278
1	Distance from nearest Metro station (km)	2.112720
2	Number of convenience stores	3.344846
3	House size (sqft)	4.643376

Multiple linear regression

- After applying multiple linear regressions on training and testing data i got the r2 score of the model .
- The r2_score for the test set is 62% This means our linear regression model is performing well on the data



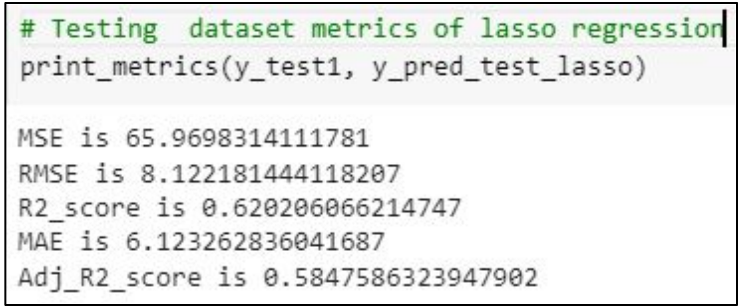
```
#Test_dataset_metrics of MLR
print_metrics(y_test, y_test_pred)
```

```
MSE is 65.29719230284037
RMSE is 8.080667812924894
R2_score is 0.6240785068062825
MAE is 6.08297857702851
Adj_R2_score is 0.604800481514297
```

Lasso regression

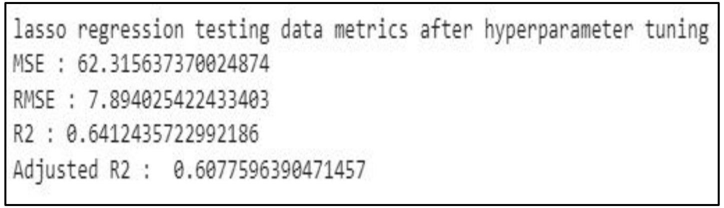
Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients

- After applying lasso regressions on training and testing data i got the r2 score of the model
- The r2 score for lasso regression for testing data is 62%.
- The r2 score for lasso regression after hyperparameter tuning for testing data is 64%.



```
# Testing  dataset metrics of lasso regression
print_metrics(y_test1, y_pred_test_lasso)

MSE is 65.9698314111781
RMSE is 8.122181444118207
R2_score is 0.620206066214747
MAE is 6.123262836041687
Adj_R2_score is 0.5847586323947902
```

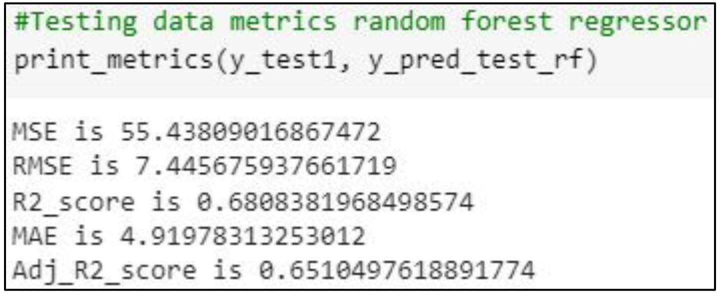


```
lasso regression testing data metrics after hyperparameter tuning
MSE : 62.315637370024874
RMSE : 7.894025422433403
R2 : 0.6412435722992186
Adjusted R2 :  0.6077596390471457
```

Random forest regressor

Random forests or **random decision forests** is an ensemble learning method for classification,regression and other tasks that operates by constructing a multitude of decision trees at training time.

- After applying random forest model on training and testing data i got the r2 score of the model .
- The r2 score for random forest model for training data is 68%.



```
#Testing data metrics random forest regressor
print_metrics(y_test1, y_pred_test_rf)

MSE is 55.43809016867472
RMSE is 7.445675937661719
R2_score is 0.6808381968498574
MAE is 4.91978313253012
Adj_R2_score is 0.6510497618891774
```

Conclusions

Conclusions from EDA :

From lineplot we conclude that house price is maximum at 10 year old house

- Average price of houses with 1 bedroom is 37.74 per unit area
- Average price of houses with 2 bedroom is 36.78 per unit area
- Average price of houses with 3 bedroom is 39.43 per unit area

From correlationplot we found out that

1. There is highly positive correlationship between House price per unit area and Number of convenience stores,latitude ,longitude.
2. There is highly negatively correlationship between distance from nearest metro station and House price per unit area because as the distance from the metro station increases the price of house decreases

Conclusion

Conclusions from Regression models :

1. The multiple linear regression is giving us R2 score of 62% for testing dataset but all the assumptions of Multiple linear regression are not satisfied.
2. Lasso regression is giving us R2 score of 62% and after hyperparameter tuning the R2 score has increased to 64% for testing dataset
3. Random forest regressor has given highest R2 score among all the model i.e. 68% for testing data and the important factors according to random forest model are Distance from nearest metro station ,House age,latitude,longitude and house size(sqft)

As the R2 score of random forest is highest among all the models so we can use random forest to predict the house prices.

Thank you