# House Price Prediction

# Project Report

# INTRODUCTION

House price prediction main is to make an effective price prediction of the houses and check the accuracy of the model. It should identify the important home price attributes which are needed to the model’s prediction. The forecasting of the prices of houses is the main topic in the real estate industry, using machine learning techniques such as Linear Regression, Random Forest etc.

# METHODOLOGY

Python is the programming language used to run the code. Random Forest and Linear Regression models are used for training and evaluation of the dataset.

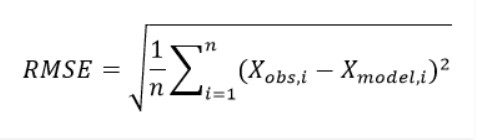
Boston house price dataset was used for the model prediction. In this model it has different phases such as data preprocessing, model building, testing the accuracy, evaluating and analyzing the model using some formulae such as mean square error (MSE), root mean squared error (RSME), mean absolute error (MAE).

# Mean Square Error:

Mean Square Error assess the average squared difference of the original and predicted values. It mainly measures the amount of error in the statistical models.

# Root Mean Squared Error:

Root mean squared error measures the difference between the predicted value by model and the actual values.



# Absolute Mean Error:

It is the measure of errors between paired observations expressing the same phenomenon.



# Preprocessing

The given dataset was for Boston house price list. Firstly, the libraries imported are:

* Pandas- this library is used for handling the data structure.
* NumPy- this library is for the linear algebra and the mathematics used.
* Seaborn- this library is used for data visualization.

Graphical user interface

Description automatically generated with medium confidence

* Now we have use data cleaning as the given dataset is not clean such as deleting duplicates, categorizing the data, decode the variables etc.

Calendar

Description automatically generated

# Data Visualization

The graphical representation of the data is known as data visualization. It is the way to see the trends, outliers and the patterns of the provided data. Data visualization is the technique used to analyze massive data and make data driven decisions.

Chart, bar chart

Description automatically generated

Fig: Data Visualization

Chart, histogram

Description automatically generated

Fig: Histogram

# Model building

It is important to know about the relationships in the existing data, constructing a correlation matrix gives that data and this is essential guide to model building. Below we can see the correlation matrix which gives brief overview of the dataset.

A red and white checkered flag

Description automatically generated with medium confidence

Correlation Matrix

# Training and validation

* + **Linear Regression**

Linear regression is the machine learning algorithm which is based on supervised learning. It is used to predict the values based on the independent variables given in the data set and mostly used to find the relationship between variables and forecasting.

**Testing the Data:**

* As shown in the figure we can see the training accuracy and the testing accuracy of the model using Linear Regression.

Text

Description automatically generated

# Evaluation

Checking the differences between actual and predicted prices:

Chart, scatter chart

Description automatically generated

* + Checking residual prices vs actual prices:

Chart, scatter chart

Description automatically generated

* + Checking Normality of errors:

Text

Description automatically generated

Chart, histogram

Description automatically generated

# Random Forest

Random Forest is a supervised machine learning algorithm that grows and combines multiple decision trees to form a "forest." It can be used in R and Python for classification and regression problems.

**Testing the data:**

As shown in the figure we can see the training accuracy and the testing accuracy of the model using Random Forest.

Text

Description automatically generated

**Evaluation**

* Checking the differences between actual and predicted price.

Chart, line chart

Description automatically generated

# CONCLUSION

I have used the 2 regression models i.e., Linear Regression and Random Forest for testing and evaluation of the given data set which is Boston house price. We can see that we get different values while using different regression models, also we can select anyone best among the above models comparing their values.

# REFERENCES

1. [Predicting House Prices with Machine Learning | by John Ade-Ojo | Towards Data Science](https://towardsdatascience.com/predicting-house-prices-with-machine-learning-62d5bcd0d68f)
2. [(PDF) House Price Prediction (researchgate.net)](https://www.researchgate.net/publication/349477129_House_Price_Prediction)
3. ht[tps://www.ibm.com/topics/logistic-regression](http://www.ibm.com/topics/logistic-regression)
4. [What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R](https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared) [Squared? - Studytonight](https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared)
5. [Mean absolute error - Wikipedia](https://en.wikipedia.org/wiki/Mean_absolute_error)
6. [Mean Square Error-Definition and Formula (byjus.com)](https://byjus.com/maths/mean-squared-error/)

By:

Vedvikas Reddy Magam