

# **PROFESSIONAL TRAINING REPORT**

at

**Sathyabama Institute of Science and Technology  
(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of Bachelor of  
Technology Degree in Information Technology.

By

**S.Aakash**

**Reg no: 39120001**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

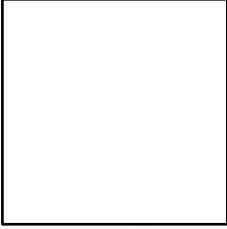
**SCHOOL OF COMPUTING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**

**CHENNAI – 600119, TAMILNADU**

**NOVEMBER 2021**



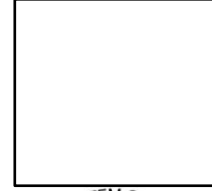
**SATHYABAMA  
INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE  
UNIVERSITY)**

**Accredited with Grade "A" by  
NAAC**

(Established under Section 3 of UGC Act,  
1956)

JEPPIAAR NAGAR, RAJIV GANDHI SALAI  
CHENNAI– 600119

[www.sathyabama.ac.in](http://www.sathyabama.ac.in)



**DEPARTMENT OF INFORMATION TECHNOLOGY**

## **BONAFIDE CERTIFICATE**

This is to certify that this project report is the Bonafide work of S.AAKASH (Reg no: 39120001) who carried out the project entitled “**Taiwan house price prediction Using Multiple Linear Regression**” under my supervision from June 2021 to November 2021.

Internal guide

Dr. R.Subhashini

Head of the department

Submitted for viva voce examination held on

Internal examiner

external examiner

**DECLARATION**

I, S.AAKASH hereby declare that the project report entitled **Taiwan house price prediction Using Multiple Linear Regression** done by me under the guidance of DR.R.Subhashini is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in Information Technology.

**DATE:**

**PLACE:**

**SIGNATURE OF THE CANDIDATE**

### **ACKNOWLEDGEMENT**

I am pleased to acknowledge my sincere thanks to the **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. T. Sasikala M.Ed., Ph.D., Dean**, school of computing, **Dr.R.Subhashini M.E., Ph.D, Head of the Department of Information Technology** 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.R.Subhashini**, for her valuable guidance, suggestions and constant encouragement paved the 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 Information Technology** who were helpful in many ways for the completion of the project.

## **TRAINING CERTIFICATE**



## ABSTRACT

Data analytics and visualization participate in a significant role in every single area of technology as per today's scenario. Even I can say every phase of our lives is surrounded by the implementation of new era technologies such as hospitality management, railway, transportation, health care, industry and so on. Data analytics

has been employed for many sectors since past decades like image processing, pattern recognition, medical diagnosis, predictive analysis, product recommendation. House price changes every year, so it is mandatory for a structure to foresee house prices in the future. House price prediction can help in fixing and thereby predicting house prices and customers can evaluate it. Our intention is to predict house prices using multiple linear regression techniques. The price of a particular location does depend on various factors like X1 transaction date, X2 house age, X3 distance to the nearest MRT station, X4 number of convenience stores, X5 latitude, X6 longitude.

Our intention behind proposing this paper is to employ data analytics techniques for predicting the price based on these metrics. The algorithm used in this analysis is multiple linear regression.

## TABLE OF CONTENTS

| CHAPTER<br>NO | TITLE | PAGE NO |
|---------------|-------|---------|
|---------------|-------|---------|

|   |                                             |   |
|---|---------------------------------------------|---|
|   | ABSTRACT                                    | i |
| 1 | INTRODUCTION                                | 1 |
| 2 | AIM AND SCOPE OF THE PRESENT INVESTIGATION  | 2 |
| 3 | EXPERIMENTAL / ALGORITHM USED               | 2 |
| 4 | RESULT AND DISCUSSION, PERFORMANCE ANALYSIS | 3 |
| 5 | SUMMARY AND CONCLUSION                      | 4 |
|   | REFERENCE                                   | 5 |
|   | APPENDIX                                    | 6 |
|   | A. SCREEN SHOT                              | 6 |
|   | B. SOURCE CODE                              | 8 |

## • INTRODUCTION

Real estate property can't be taken only as an important need of society but it also represents the status and reputation of an individual. Venture in real estate usually seems to be gainful because their possession values do not turn down rapidly. Transformation in the real estate price can control a mixture of conjugal sponsors, bankers, policy makers and so on. Venture in the real estate region seems to be an

eye-catching alternative for the investments. Investments is commerce that most people are interested in this globalization age. There are numerous often used for investment, for example, gold, stocks and property.

Machine learning aims at developing self-learning algorithms using dataset, such that machines can be enabled to project future activity based on the past data. It helps organizations in easily identifying and predicting the trends and patterns. It also facilitates the managers, executive and analysts to make use of its models in making decisions in an efficient manner. It helps the organization to adopt environmental without human intervention

Machine learning is used as an algorithm for building a model and thereby using that model for predicting a new dataset. The prominent difference of using conventional algorithms is that the result is oriented with the input data rather than focusing on a chain of different instruction sets. Supervised learning is based on building a model based on a labeled data set whereas unsupervised learning is totally oriented with an unlabeled data set. So far, the matter of house prediction is concern it is determined by following metrics: -

- X1 transaction date,
- X2 house age,
- X3 distance to the nearest MRT station,
- X4 number of conveniences store,
- X5 latitude,
- X6 longitude.

- **AIM AND SCOPE OF THE PRESENT INVESTIGATION**

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted

- **EXPERIMENTAL / ALGORITHM USED**

- MULTIPLE LINEAR REGRESSION

***Iteration***

Select one variable at a time from the variable importance table created using random forest

***Significance***

- Check the relationship of new variables along with existing variables by its t-value and probability of t- statistics.
- If R-square is improved, keep the variable, else drop it

***Model Accuracy***

- After adding the new variable, the model accuracy on train and test data using error rate and MAPE is checked.
- Drop the new variable if the model accuracy falls.

## **4.RESULT AND DISCUSSION, PERFORMANCE ANALYSIS**

### **4.1.MODEL PERFORMANCE:**

In this study dependent variables refer to output of the process, whereas independent variables refer to input of the process. Dependent variable is the price of the house per square meter in the current study. Independent variables are considered as house age, distance to the nearest MRT station, number of convenience stores in the living circle on the foot and geographic coordinate information (Latitude and Longitude). Prediction is based on multiple linear regression models. Therefore, the size of the testing set is accepted as 20% in this research.

### **4.2RELATIVE IMPORTANCE OF THE PROPERTY VARIABLES:**

This section provides information on the relative importance of the six variables used in the model development. Where the number of convenience stores in the living circle on the foot values specified that the variables have the least effect on prediction of real estate price, while distance to the nearest MRT station value highly significant effect of prediction real estate price. MRT stations are common in many countries around the world. This study safely suggests that real estate tenants, sellers and buyers initially consider whether there is an MRT station in the real estate before negotiating further. These are less important compared to the others in the process of prediction of real estate prices. Generally, this study provides preliminary information for real estate tenants, sellers and buyers about the importance of the property variables

## **5.SUMMARY AND CONCLUSION**

This study concentrates on data mining algorithms about prediction of real estate prices. While 80% of the dataset is used as a training set, the rest of it is used for the purpose of testing. To conclude, the gradient boosting algorithm provided the highest performance for the real estate price prediction according to results obtained. This study can be extended using a new dataset which consists of additional parameters.

## **6.REFERENCE**

- Fik, T. J., Ling, D. C. & Mulligan, G. F. (2003)."Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach," Real Estate Economics, Vol 31(4), 633-646
- Dubin, R., Pace, R. K. & Thibodeau, T. G. (1999)."Spatial Autoregression Techniques for Real Estate Data, "Journal of Real Estate Literature, 7(1), 79-5.

## **7.APPENDIX**

## • SCREEN SHOT

The first screenshot shows the initial data loading and inspection steps in the Jupyter Notebook:

```
In [3]: df = pd.read_excel("real estate valuation data set.xlsx")
df.head(10)
```

Out[3]:

|   | No | X1 transaction date | X2 house age | X3 distance to the nearest MRT station | X4 number of convenience stores | X5 latitude | X6 longitude | Y house price of unit area |
|---|----|---------------------|--------------|----------------------------------------|---------------------------------|-------------|--------------|----------------------------|
| 0 | 1  | 2012.910667         | 32.0         | 84.87882                               | 10                              | 24.98298    | 121.54024    | 37.9                       |
| 1 | 2  | 2012.910667         | 19.5         | 306.59470                              | 9                               | 24.98034    | 121.53951    | 42.2                       |
| 2 | 3  | 2013.583333         | 13.3         | 561.98450                              | 5                               | 24.98746    | 121.54391    | 47.3                       |
| 3 | 4  | 2013.500000         | 13.3         | 561.98450                              | 5                               | 24.98746    | 121.54391    | 54.8                       |
| 4 | 5  | 2012.833333         | 5.0          | 300.59840                              | 5                               | 24.97937    | 121.54245    | 43.1                       |
| 5 | 6  | 2012.666667         | 7.1          | 2175.03090                             | 3                               | 24.98005    | 121.51254    | 32.1                       |
| 6 | 7  | 2012.666667         | 34.5         | 623.47310                              | 7                               | 24.97033    | 121.53642    | 40.3                       |
| 7 | 8  | 2013.410667         | 20.3         | 287.60250                              | 6                               | 24.98042    | 121.54228    | 46.7                       |
| 8 | 9  | 2013.500000         | 31.7         | 5512.03890                             | 1                               | 24.98695    | 121.48458    | 18.8                       |
| 9 | 10 | 2013.410667         | 17.0         | 1733.18000                             | 3                               | 24.96731    | 121.51480    | 22.1                       |

```
In [23]: y = df["Y house price of unit area"]
x = df.drop(["No", "X1 transaction date", "Y house price of unit area"], axis=1)
```

```
In [24]: y.head()
```

Out[24]:

| 0 | 37.9 |
|---|------|
| 1 | 42.2 |
| 2 | 47.3 |
| 3 | 54.8 |
| 4 | 43.1 |

Name: Y house price of unit area, dtype: float64

```
In [25]: x.head()
```

The second screenshot shows the data splitting and model training steps:

```
In [24]: y.head()
```

Out[24]:

| 0 | 37.9 |
|---|------|
| 1 | 42.2 |
| 2 | 47.3 |
| 3 | 54.8 |
| 4 | 43.1 |

Name: Y house price of unit area, dtype: float64

```
In [25]: x.head()
```

Out[25]:

|   | X2 house age | X3 distance to the nearest MRT station | X4 number of convenience stores | X5 latitude | X6 longitude |
|---|--------------|----------------------------------------|---------------------------------|-------------|--------------|
| 0 | 32.0         | 84.87882                               | 10                              | 24.98298    | 121.54024    |
| 1 | 19.5         | 306.59470                              | 9                               | 24.98034    | 121.53951    |
| 2 | 13.3         | 561.98450                              | 5                               | 24.98746    | 121.54391    |
| 3 | 13.3         | 561.98450                              | 5                               | 24.98746    | 121.54391    |
| 4 | 5.0          | 300.59840                              | 5                               | 24.97937    | 121.54245    |

### Creating Training and Testing Data

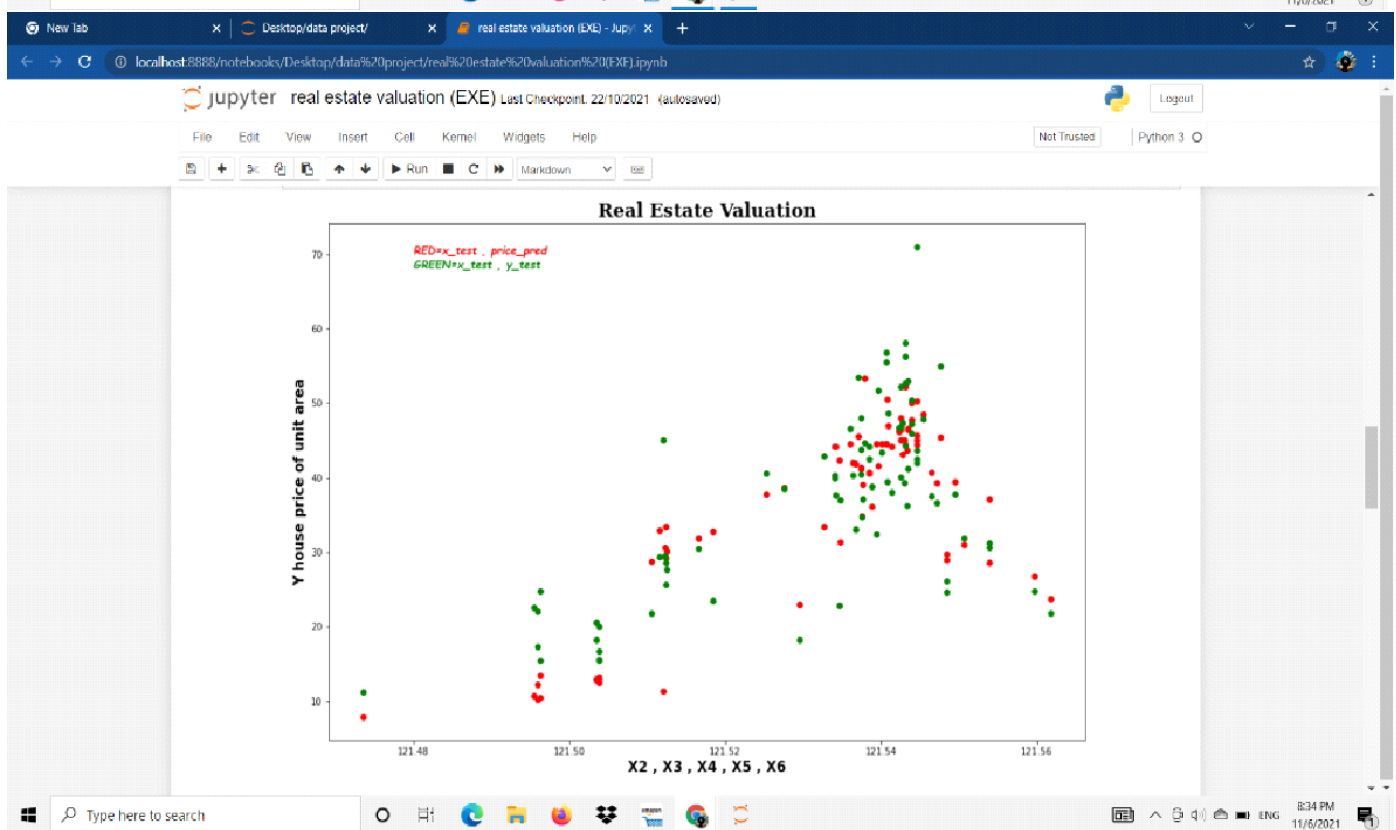
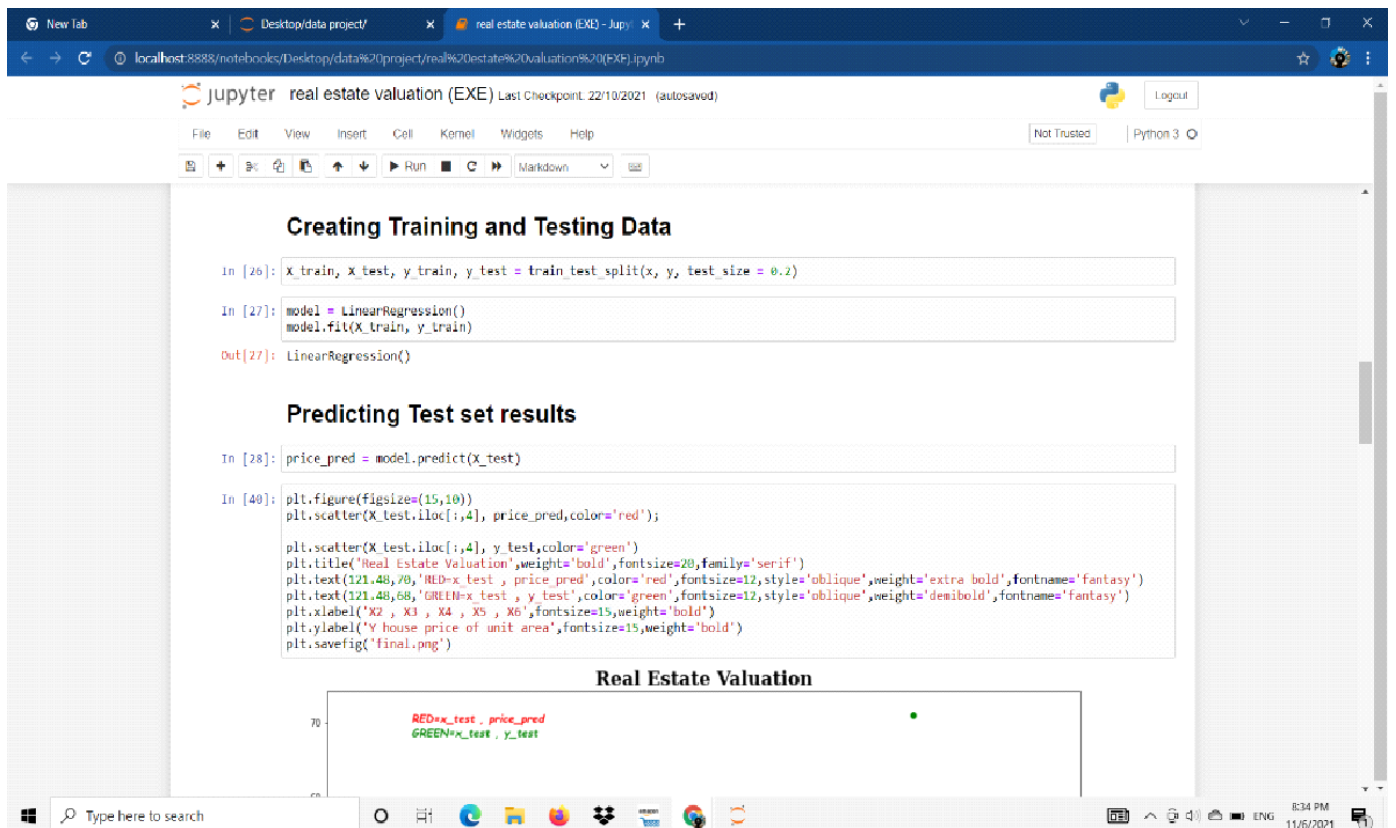
```
In [26]: X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)
```

```
In [27]: model = LinearRegression()
model.fit(X_train, y_train)
```

Out[27]: LinearRegression()

### Predicting Test set results

```
In [28]: y_pred = model.predict(X_test)
```



## B. SOURCE CODE

```
import pandas as pd
import matplotlib.pyplot as plt

import numpy as np
from sklearn.linear_model import LinearRegression

from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
df = pd.read_excel("Real estate valuation data set.xlsx")
df.head(10)
y = df['Y house price of unit area']
x = df.drop(['No','X1 transaction date','Y house price of unit area'],axis=1)
y.head()
x.head()
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.2)
model = LinearRegression()
model.fit(X_train, y_train)

price_pred = model.predict(X_test)
plt.figure(figsize=(15,10))
plt.scatter(X_test.iloc[:,4], price_pred,color='red');

plt.scatter(X_test.iloc[:,4], y_test,color='green')
plt.title('Real Estate Valuation',weight='bold',fontsize=20,family='serif')
```

```

plt.text(121.48,70,'RED=x_test ,
price_pred',color='red',fontsize=12,style='oblique',weight='extra
bold',fontname='fantasy')

plt.text(121.48,68,'GREEN=x_test ,
y_test',color='green',fontsize=12,style='oblique',weight='demibold',fontname='fantasy')

plt.xlabel('X2 , X3 , X4 , X5 , X6',fontsize=15,weight='bold')
plt.ylabel('Y house price of unit area',fontsize=15,weight='bold')
plt.savefig('final.png')
print('Coefficients: \n', model.coef_)

print('Intercept: \n', model.intercept_)

#MSE

print('Mean squared error: %.2f'
      % mean_squared_error(y_test, price_pred))

#cod

print('Coefficient of determination: %.2f'
      % r2_score(y_test, price_pred))
y_pred=model.predict(X_test)
print(y_pred)
pred_y_df=pd.DataFrame({'Actual Value':y_test,'Predicted Value':y_pred, 'Difference':
y_test-y_pred})
pred_y_df[0:30]

```