

A STATISTICAL APPROACH OF TIME SERIES

Submitted in partial fulfillment of the requirements for the
award of Master of Science degree in Mathematics

By

M. HARINI (40630002)



**DEPARTMENT OF MATHEMATICS
SCHOOL OF SCIENCE AND HUMANITIES**

SATHYABAMA

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

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MAY - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF MATHEMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **M. HARINI (40630002)** who carried out the project entitled "**A STATISTICAL APPROACH OF TIME SERIES**" under our supervision from **November 2021 to May 2022.**

Internal Guide

Dr. M. MALLIKA

External Guide

Head of the Department

Dr. M. NIRMALA, M.Sc., M.Phil., Ph.D.,
Head of the Department

Dr. M. NIRMALA of Mathematics

Sathyabama Institute of Science and Technology
(Deemed to be University), Chennai-600 119

Submitted for Viva voce Examination held on 12.05.2022

R. Narasimhan

Internal Examiner

(Dr. R. Narasimhan)

External Examiner

(S. Ganesan)

DECLARATION

I **M. HARINI (40630002)** hereby declare that the Project Report entitled "**A STATISTICAL APPROACH OF TIME SERIES**" done by me under the guidance of **Dr. M. MALLIKA** (Internal) and _____(External) at **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** is submitted in partial fulfillment of the requirements for the award of Master of Science degree in Mathematics.

DATE: 12.05.2022

PLACE: CHENNAI

Harini.M

SIGNATURE OF THE CANDIDATE

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. M. NIRMALA**, Head of the Department, Department of Mathematics 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. M. MALLIKA**, Assistant Professor, Department of Mathematics for her 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 Mathematics who were helpful in many ways for the completion of the project.

ABSTRACT

The house is a basic need of people to live a secure and hygienic life. The housing shortage among low-income groups is large, despite increases in the stock of quality housing in recent years, as house prices are high relative to incomes and access to credit is often difficult. In this project, the house price dataset has been collected from the famous website Kaggle. The missing values in the dataset are filled and the factors influencing price of the house are found, thus the dependent and independent variables are declared. The dataset was splitted into training and test sets where the train set is used to build the model and the test set is used to evaluate the model. The multiple linear regression model to predict house price was developed successfully using Scikit-learn in the tool python and evaluated by the error measures like MAPE, RMSE and R^2 score, which resulted the accuracy of the model is good and more effective to predict house price. Both the seller and buyer can make use of the house price prediction model. The house price prediction model is useful to both the seller and the buyer.

KEYWORDS:

House price prediction, Multiple linear regression (MLR), Mean absolute percentage error (MAPE), Root mean square error (RMSE), R^2 Score

TABLE OF CONTENTS

Abstract	i
List of figures	iii
List of tables	iii
List of abbreviations	iii
 CHAPTER – I	 1
1.1 Introduction	2
1.2 Time series analysis	3
1.3 A statistical approach of time series	3
 CHAPTER – II	 4
2.1 Literature survey	5
 CHAPTER – III	 10
3.1 Aim & Scope	11
 CHAPTER – IV	 12
4.1 Methodology	13
4.1.1 Multiple linear regression	13
4.2 Software tool used	14
4.3 Dataset	14
4.3.1 Handling missing data	15
4.4 Defining explanatory and response variables	15
4.4.1 Classification of variables	15
4.4.2 Dummy variables	16
4.4.3 Defining X and Y variables	17
4.5 Splitting dataset into train and test sets	18

CHAPTER – V	19
5.1 Results and Performance Analysis	20
5.1.1 Multiple linear regression model	20
5.1.2 MAPE	22
5.1.3 RMSE	22
5.1.4 R ² score	22
CHAPTER – VI	24
6.1 Summary and Conclusion	25
References	26

LIST OF FIGURES

Fig: 5.1 Regression plot	21
--------------------------	----

LIST OF TABLES

Table 4.1 The categorical variables of the dataset	17
Table 4.2 The independent variables of the dataset	18
Table 5.1 The intercept value	20
Table 5.2 Values of the regression coefficients	20
Table 5.3 Values of the error measures	23

LIST OF ABBREVIATIONS

MLR	Multiple linear regression
MAPE	Mean absolute percentage error
RMSE	Root mean square error
R ²	R squared

CHAPTER – I

1.1 INTRODUCTION

Housing is a key part of well-being and contributes to spatial and social mobility. Though every household's dream is to buy their own house, the forced lockdown has led to an increasing demand for buying houses to sustain better social distancing standards. According to the report from Knight Frank, India's real estate sector bounced back from the pandemic-induced slowdown with a record rise in sales. A new era of online home buying and selling has taken off. The rapid rise in population and incomes over the past decades has resulted in an increase in demand for housing. However, despite several public housing programmes that boosted supply, the housing shortage remains.

In 2015, according to government estimates, about 40 million households faced housing shortages - which includes the homeless and households in sub-standard or congested dwellings or without basic amenities. The "Housing for all" (Prime Minister's Housing Scheme) by 2022, the project of the Indian Government aims to provide access to affordable housing which provides a better lifestyle for every household in India. Now it is extended to complete 80 lakh houses in 2022-2023. The consistent increase in raw materials prices is pushing up property prices in all eight prime residential markets.

House sales in India's eight prime housing markets increased 13% in 2021, when compared to the overall sales in 2020. Mid-income houses contributed 41% of the new houses from Rs.40 lakh to Rs.80 lakh and the affordable houses with the rate of Rs.40 lakhs contributed 21% in the 3rd Quarter of the year 2021 to the housing market. Chennai housing markets has an annual price hike of 5% in 2021. Certain projections state that the sales momentum is expected to increase in 2022 as prospective home buyers will continue to prefer bigger homes, better amenities and attractive pricing will keep them interested in sealing the deals. This project will be helpful to both the buyers and sellers to buy and sell the house with the house price prediction model built using multiple linear regression method.

1.2 TIME SERIES ANALYSIS

A time series is a set of observations taken at specified times, usually at equal intervals. Many sets of data appear as time series: daily rainfall amounts, hourly observations made on the yield of a chemical process, a weekly series of the number of road accidents, the number of new born babies in India in each year and so on. Examples of time series abound in so many fields, such as economics, business, engineering, the natural sciences and the social sciences. Time Series Analysis is a specific way of analyzing a sequence of data points collected over an interval of time. Time series are analyzed to discover past patterns of variability that can be used to forecast future values. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. The main objective of a Time Series Analysis is to develop a suitable model to describe the pattern or trend in data with more accuracy. These objectives may be classified as description, explanation, prediction and control.

1.3 A STATISTICAL APPROACH OF TIME SERIES

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. Statistics is applicable to a wide variety of academic disciplines, including natural and social sciences, government, and business. Machine learning models are statistical and probabilistic models that capture patterns in the data through the use of computational algorithms. A typical statistics course covers descriptive statistics, probability, binomial and normal distributions, test of hypotheses and confidence intervals, linear regression, and correlation. It is particularly applied in financial management, marketing management, and production, services and operations management. A typical "Business Statistics" course is intended for business majors, and covers descriptive statistics, it may include forecasting, time series, decision trees, multiple linear regression, and other topics from business analytics more generally. In this project, house price prediction model has been developed using multiple linear regression by the housing dataset.

CHAPTER – II

2.1 LITERATURE SURVEY

Abigail Bola Adetunji et al. 2022 explored the use of random forest machine learning techniques for house price prediction. The UCI machine learning repository Boston housing dataset with 506 entries and 14 features were used to evaluate the performance of the proposed prediction model. This study has further affirmed the prowess of random forest machine learning technique in predicting the prices of a house based on variables made available in the Boston housing dataset. A comparison of the predicted and actual prices revealed that the model achieved a prediction difference of ± 5 .

Adyan Nur Alfiyatin et al. 2017 have predicted house prices based on NJOP houses in Malang city with regression analysis and particle swarm optimization (PSO). PSO is used for selection of affecting variables and regression analysis is used to determine the optimal coefficient in prediction. The result from this research proved that a combination regression and PSO is suitable and get the minimum prediction error obtained, which is IDR 14.186.

Amey Thakur and Mega Satish 2021 implemented a house price prediction model of Bangalore, India and experimented with different machine learning algorithms to get the best model. When compared to all other algorithms, the decision tree algorithm achieved the lowest loss and the greatest R-squared. As housing prices fluctuate on a daily basis and sometimes exaggerated rather than based on worth, thus the major focus of this project is on predicting home prices using genuine factors.

Anand et al. 2021 has presented to predict accurate price of house without any loss based on various features on given data by using machine learning algorithms like linear regression, decision tree regression, K-means regression and random forest regression. From the analysis, the value of RMSE is 2.9131. The result of this research provides that the random forest regression gives maximum accuracy, it helps people to buy house on budget and reduces loss of money.

Atharva Chouthai et al. 2019 used machine learning algorithms to build the prediction model for houses. Housing sales prices are determined by numerous factors such as

area of the property, location of the house, material used for construction, age of the property, number of bedrooms and garages and so on. They used a data set of 100 houses with a number of parameters and used 50 percent of the data set to train the machine and 50 percent to test the machine. Here, machine learning algorithms such as logistic regression and support vector regression, lasso regression technique and decision tree are employed to build a predictive model and the results are truly accurate.

Fashoto et al. 2021 applied machine learning (ML) to predict maize yields for a single season in Eswatini. A ML model was trained and tested using open-source data and local data. This is done by using three different data splits. Predicting maize crop yields helps policymakers to make timely import and export decisions to strengthen national food security. However, none of the published work has been done to predict maize crop yields using machine learning in Eswatini, Africa.

Kodamagulla Kausthub 2021 addressed the usage of multiple linear regression technique in the case of predicting sales related to commercials, which are displayed in mainly three forms of media, namely TV, Radio & Newspaper. Using this analysis, many brands also improve the commercial content and relevance based on the customer's interest. Accuracy of the multiple linear regression algorithm for this model is calculated and was found to be 91.7%.

Nor Hafizah Hussin, Rahaini Mohd Said Aditya and Nadiah Ishak 2021 have presented a mathematical modelling using multiple linear regression (MLR) to predict the energy consumption for energy savings in Technology Campus, UTeM. Using the stepwise selection, the linear model obtained consist of the most influenced factors which is the number of lightings used and surrounding temperature in the building where the equation is $\hat{y} = 149.3 - 3.95b_1 + 0.0877b_2 + e$. Comparing the actual and predicted energy consumption at Technology Campus, UTeM showed that the MLR model obtained can be used to predict energy consumption and accounted for around 81% of the variance.

Oghoyafedo et al. 2017 has developed the Accident prediction model at Unsignalized Intersection using multiple linear regression method and the standard error of estimate was computed to show how close the observed value is to the regression line. As the issue of road accidents is an increasing problem in developing countries due to increasing road traffic/vehicle occupancy, geometric characteristics and road way condition, this study will enhance the production of updated design standards to enable practitioners to design unsignalized intersections for optimal safety and reduce the number of accidents at unsignalized intersections.

Olsen et al. 2020 this article has been tailored for individuals who are interested in learning more about multiple linear regression as a data analysis tool and its potential utility in pharmacy education research. Key terms needed to understand and interpret regressions are also discussed. Regression analyses could be helpful in advancing pharmacy educational scholarship by enabling scholars to better understand variables that may predict specific outcomes such as student achievement or program retention.

Quang Truong et al. 2020 explored various impacts of features on prediction methods. This paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation of regression and provide an optimistic result for housing price prediction.

Rath et al. 2020 have analysed the daily statistics of people affected by the disease are taken into account to predict the next day's trend in the active cases in Odisha as well as India. A valid global data set is collected from the WHO daily statistics and correlation among the total confirmed, active, deceased, positive cases are stated in this paper. Regression model such as linear and multiple linear regression techniques are applied to the data set to visualize the trend of the affected cases. Here a comparison of linear regression and multiple linear regression model is performed where the score of the model R^2 tends to be 0.99 and 1.0 which indicates a strong prediction model to forecast the next active cases. These models acquired remarkable accuracy in COVID-19 recognition.

Robart Konwar et al. 2021 in this paper, machine learning algorithms are explained in various combinations and the result for each algorithm is given based on the accuracy percentage to predict the exact prices of houses without bias to help both the buyers and sellers make their decisions and concluded that random forest algorithm provides more accuracy than support vector machines (SVM), linear regression algorithm and XG boost algorithm.

Shruti Shakhla 2018 studied and documented the performance of APPLE INC.'s stock price using multiple linear regression and gauged its performance using root mean squared error. Machine learning algorithms have proved to fetch benevolent results in predicting stock prices. The results are promising but can be improved by taking into consideration more parameters.

Xiaojie Xu and Yun Zhang 2021 have demonstrated the usefulness of the machine learning approach to the house price forecasting problem in the Chinese market, by exploring neural networks for forecasting of house prices from one hundred major cities for the period of June 2010–May 2019, serving as the first study with such wide coverage for the emerging Chinese market through a machine learning technique. The results could be used on a standalone basis or combined with fundamental forecasting in forming perspectives of house price trends and conducting policy analysis.

Yanming Yang et al. 2018 introduced the mathematic model and calculation method of linear regression model and the parameter estimation and model test method of linear regression model is discussed. As it is indispensable to scientifically predict the consumption of aero-material spare parts and to make scientific decisions on aviation equipment maintenance resources and make full use of existing resources to improve maintenance capability, the application of linear regression model in forecasting the consumption of aero-material spare parts is analyzed by examples. According to the analysis results, the regression equation is modified as necessary to further improve the prediction accuracy. The results show that the linear regression model is feasible and effective for the prediction of aero-material spare parts consumption.

Yige Wang 2021 this paper, builds two forecasting models to predict house prices. It primarily takes a series of data processes, including missing value process, outlier removal, new variables creation, and correlation analysis. Finally, based on the results, this paper analyzes and compares the advantages and disadvantages of the two models. In terms of price prediction, the decision tree fitting effect is obviously more practical than OLS in the face of complex and nonlinear data. Therefore, in these two conditions, one is a large number of observations in this dataset, the other is complex and has noise samples and recommend to choose random forest.

Yin Libao et al. 2017 established the multiple linear regression model predicting CO₂ emission of coal-fired power plant, which provides guidance for the realization of online monitoring and to realize low-carbon development, and strive to complete the carbon emissions reduction targets of "13th Five-Year". The calculation of CO₂ emissions of Wanglong thermal power plant unit using multiple linear regression method is compared with both the results which is calculated by the methods provided in the 2006 IPCC Guidelines and the method of the coal consumption rate of power supply. Thus, the accuracy of this algorithm is verified.

Yun Hwan Kim et al. 2014 described the current trends in the prediction of crop pests using machine learning technology with the advent of data mining, the field of agriculture is also focused on it. This paper classifies and introduces SVM (support vector machine), multiple linear regression, neural network, and Bayesian network-based techniques, and describes some cases of their utilization.

Zihan Yan 2021 established a multiple linear regression model for predicting commercial housing prices, and the model is tested for multicollinearity, serial correlation and heteroscedasticity, and the optimal model is obtained. In addition, based on the grey correlation degree, this paper analyzes the impact of population trend and enterprise vitality on the economic vitality of Guangzhou, and studies the long-term and short-term impact of economic policy transformation on the provincial economy.

CHAPTER – III

3.1 AIM & SCOPE

Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series analysis can be applied to real - valued, continuous data and discrete numeric data. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering. In statistics, prediction is a part of statistical inference. The objective of this project is to build a multiple linear regression model to predict the house price effectively. To build the multiple linear regression model, the process includes finding the explanatory variables to predict the outcome of a response variable price, filling out the missing values in the dataset and splitting the house price dataset into training and test set. To validate the model's accuracy, error measures like MAPE, RMSE and R^2 score are used. House prices increase every year, so there is a need for a system to predict house prices in the future. The house price prediction model is very essential in filling the information gap and improving real estate efficiency. House price prediction can help the developer determine the selling price of a house and can help the customer to know the right time to purchase a house.

CHAPTER – IV

4.1 METHODOLOGY

4.1.1 MULTIPLE LINEAR REGRESSION

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

The Multiple linear regression (MLR) model can be represented as:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots \beta_k * X_k + \varepsilon$$

where,

Y = Dependent or Response variable.

X_i = Independent or Predictor variables $X_1, X_2, \dots, X_k$.

β_0 = Y - Intercept (constant term).

β_k = Regression (slope) coefficients for each predictor variable.

ε = The model's error term (also known as the residual).

Multiple regression attempts to predict a dependent or response variable y on the basis of an assumed linear relationship with several independent or predictor variables $X_1, X_2, \dots, X_k$.

For example, the yield of a crop may depend on the amount of nitrogen, potash, and phosphate fertilizers used. These variables are controlled by the experimenter, but the yield may also depend on uncontrollable variables such as those associated with weather.

4.2 SOFTWARE TOOL USED

To fit a multiple linear regression model from the House Price Dataset, it's not the easiest calculation to do by hand, so in most cases you'll need to use statistical software to see the results plotted on a graph for easier analysis.

Python is a high-level general-purpose programming language as a tool and it is a computer programming language often used to build websites and software, automate tasks and conduct data analysis. It can be used to create a variety of different programs and is not specialized for any specific problems.

As it's an open-source language, anyone can use it. To code, python is relatively easy to read and understand, as its syntax is more like English. Its straightforward layout means that you can work out what each line of code is doing.

Scikit – learn also known as sklearn is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN and is designed to interoperate with the python numerical and scientific libraries NumPy and SciPy.

Here, Scikit - learn is used in Python to build the Multiple linear regression model for the House Price Dataset.

4.3 DATASET

The house price dataset used in this project is taken from a website called Kaggle. In the house price dataset, the dependent (response) variable is price and the 7 independent (explanatory) variables are area, status, bhk, bathroom, year, location and builder where the price in lakhs, area in square feet, status in “ready to move” or “under construction”, number of bhk, number of bathrooms, number of years after finished building the house, location of the house and builder of the house.

4.3.1 HANDLING MISSING DATA

The missing data in the dataset is very common in real life datasets, where the missing data can be represented as a blank between the data or NAN (not a number). This shows that the information is not given in the dataset, which does not mean that there is no data. The prediction model could fail because of the missing data in the dataset.

In the house price dataset, there are many blanks where the number of years should be given. It is impossible to ignore this to take further steps to build multiple linear regression model. So, to handle the missing data there are many methods like

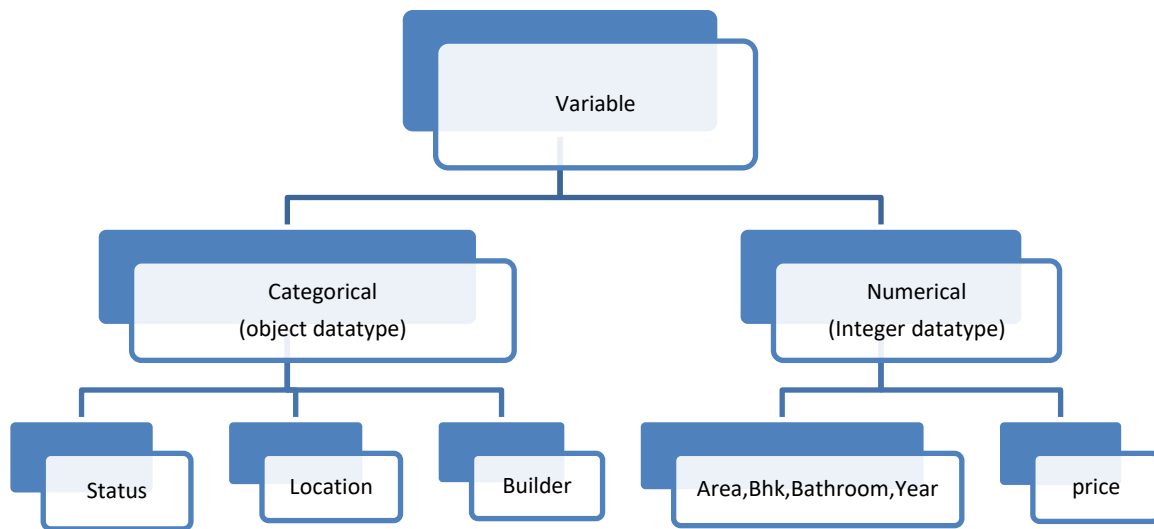
- Dropping the entire row
- Dropping the entire column
- Filling the missing values with mean, median, mode value with respect to that variable data
- Replacing with previous or next value
- Imputation of missing values using Scikit - learn library like KNN imputer

In this project, the missing values in the house price dataset are filled by the mean values of the present data in the variable year and also changed the data which are in the float datatype to integer datatype for all numeric variables.

4.4 DEFINING EXPLANATORY AND RESPONSE VARIABLES

4.4.1 CLASSIFICATION OF VARIABLES

The variables of dataset can be further classified as categorical and numerical variables, where both the dependent and independent variables can be distributed in the categorical and numerical variables.



The numerical variables are also known as quantitative variables, where the measurement or number has a numerical meaning. For example, counts, people's height, total rainfall measured in inches etc.

The categorical variables are also known as qualitative variables. It is a variable that has two or more categories with no intrinsic ordering to the categories. For example, age group, gender, education level etc.

4.4.2 DUMMY VARIABLES

Here the house price dataset has the dependent variable price and some independent variables area, bhk, bathroom, year are numerical variables and the independent variables status, location and builder are categorical variables. To take further steps to build the MLR model, the categorical variables should be changed into numerical variables. The categorical variables are

Table 4.1: The categorical variables of the dataset

Status	Location	Builder
ready to move	location 1	builder 1
under construction	location 2	builder 2
	location 3	
	location 4	
	location 5	
	location 6	
	location 7	

Dummy variables are binary variables coded as 1 to indicate the presence of some attribute and as 0 to indicate the absence of some attribute. Typically, linear regression is used with numeric variables. The dummy variables are the numerical variables used in regression analysis to represent categorical data and can only take the values of zero and one.

The dummy independent variable which for some observations has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept.

Dummy variables are used frequently in time series analysis with regime switching, seasonal analysis and qualitative data applications. The dummy variables have some alternative names, such as indicator variables, binary variables and dichotomous variables. The dummy variable concept is applied to the house price dataset.

4.4.3 DEFINING X AND Y VARIABLES

The dependent variable price is Y and the independent variables ($X_1, X_2 \dots, X_{15}$) of the house price dataset are

Table 4.2: The independent variables of the dataset

Independent Variables	
area	X ₁
bhk	X ₂
bathroom	X ₃
year	X ₄
ready to move	X ₅
under construction	X ₆
location1	X ₇
location2	X ₈
location3	X ₉
location4	X ₁₀
location5	X ₁₁
location6	X ₁₂
location7	X ₁₃
builder1	X ₁₄
builder2	X ₁₅

4.5 SPLITTING DATASET INTO TRAIN AND TEST SETS

To build the multiple linear regression model, the training set of the data is used and for the prediction process the test set of the data is used.

To achieve this 80% of the dataset is splitted into training set and 20% of the dataset is splitted into test set.

CHAPTER – V

5.1 RESULTS AND PERFORMANCE ANALYSIS

5.1.1 MULTIPLE LINEAR REGRESSION MODEL

To develop the multiple linear regression model of the house price dataset, the intercept and the regression coefficients for each predictor variable are calculated using scikit – learn. Thus, the calculated values of intercept and the regression coefficients are represented in the tables 5.1 and 5.2 respectively.

Table 5.1: The intercept value

Intercept	Value
β_0	1.60874608

Table 5.2: Values of the regression coefficients

Coefficients	Values
β_1	0.05263027
β_2	10.08624066
β_3	-10.91739282
β_4	2.14948574
β_5	-3.5895784
β_6	3.5895784
β_7	-5.02276432
β_8	-2.5896122
β_9	-1.22242726
β_{10}	0.21765399
β_{11}	-7.27061063
β_{12}	-0.26717527

β_{13}	16.15493569
β_{14}	3.47659743
β_{15}	-3.47659743

$$\begin{aligned}
\text{Price} = & \beta_0 + (\beta_1 * \text{area}) + (\beta_2 * \text{bhk}) + (\beta_3 * \text{bathroom}) + (\beta_4 * \text{age}) \\
& + (\beta_5 * \text{ready to move}) + (\beta_6 * \text{under construction}) + (\beta_7 * \text{location1}) \\
& + (\beta_8 * \text{location2}) + (\beta_9 * \text{location3}) + (\beta_{10} * \text{location4}) + (\beta_{11} * \text{location5}) \\
& + (\beta_{12} * \text{location6}) + (\beta_{13} * \text{location7}) + (\beta_{14} * \text{builder1}) + (\beta_{15} * \text{builder2})
\end{aligned}$$

The developed multiple linear regression model is shown in the above equation.

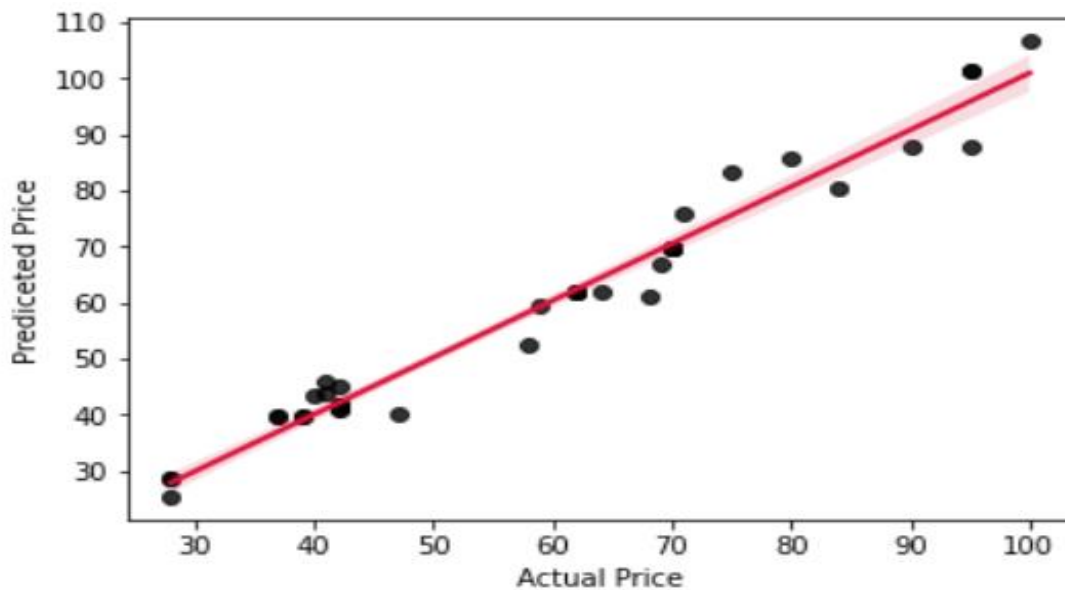


Fig 5.1: Regression plot

The Fig 5.1 shows the model was well fitted.

5.1.2 MAPE

The mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method in statistics. The formula to calculate MAPE is as follows:

$$\text{MAPE} = \frac{100}{n} * \sum_{i=1}^n \frac{|Actual\ value_i - Predicted\ value_i|}{|Actual\ value_i|}$$

mean absolute percentage error is commonly used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error. MAPE value of high numbers are bad model and low numbers are good model.

5.1.3 RMSE

The root mean squared error (RMSE) is the square root of the average of squared errors and it is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. The formula to calculate RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\text{predicted}_i - \text{Actual}_i)^2}{n}}$$

It can be said that RMSE values between 0.2 and 0.5 show that the model can relatively predict the data accurately.

5.1.4 R² SCORE

In statistics, the coefficient of determination is denoted by R² or r² (“r squared”) is the proportion of the variation in the dependent variable that is predictable from the independent variable.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

The better the linear regression fits the data in comparison to the simple average, the closer the value of R^2 is to 1.

Thus, the MLR model has been evaluated by the error measures MAPE, RMSE and R^2 score to confirm accuracy and performance.

Table 5.3: Values of the error measures

Error	Values
MAPE	0.0433
RMSE	3.5795
R^2 SCORE	0.9674

The Table 5.3 represents the values of the error measures of the MLR model.

CHAPTER – VI

6.1 SUMMARY AND CONCLUSION

In this project, the collected house price dataset has been analysed and prepared to build multiple linear regression model, thus the missing values are filled with mean values and categorical variables are well handled by introducing dummy variable method after the dataset got prepared the dataset had splitted into train set (80%) and test set (20%) the regression coefficients and intercept are calculated , then the model was fitted and evaluated by the error measures resulted in MAPE value of 0.0433, RMSE value of 3.5795 and with the R^2 of 0.9674 which shows it is an effective good model for house price prediction with 96% accuracy and plotted the fitted regression model successfully.

REFERENCES

- [1] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala ^a, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique", The 8th International Conference on Information Technology and Quantitative Management, Procedia Computer Science 199, (2022), 806 - 813.
- [2] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017, 323 - 326.
- [3] Amanda A. Olsen, Jacqueline E. McLaughlin, Spencer E. Harpe, "Using Multiple Linear Regression in Pharmacy Education Scholarship", Currents in Pharmacy Teaching and Learning, 12 (2020), 1258 - 1268.
- [4] Amey Thakur, Mega Satish, "Bangalore House Price Prediction", International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 09 | Sep 2021, 193 - 196.
- [5] Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, Dr. Vinayk A. Bharadi, "House Price Prediction using Machine Learning", IRE Journals, Volume 4 Issue 11 | May 2021, 29 - 33.
- [6] Atharva Chouthai, Mohammed Athar Rangila, Sanved Amate, Prayag Adhikari, Vijay Kukre, "House Price Prediction using Machine Learning", International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03 | Mar 2019, 4403 - 4406.
- [7] Kodamagulla Kausthub, "Commercials Sales Prediction using Multiple Linear Regression", International Research Journal of Engineering and Technology (IRJET), Volume: 08, Mar 2021, 826 - 829.

- [8] Mohammed Azam Sayeed, Noor Ayesha, Mahadeva B K, Shanavaz H, "Estimation of Iron using Multiple Linear Regression Models", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 8 Issue II, Feb 2020, 733 - 747.
- [9] Nor Hafizah Hussin, Rahaini Mohd Said Aditya, Nadiah Ishak, "Modelling and predicting energy consumption in laboratory buildings using multiple linear regression", Jurnal Ilmiah Matematika, Vol. 8, No. 1, April 2021, pp. 33 - 39.
- [10] Oghoyafedo N.K, Ehiorobo J. O, Nwankwo E, "Accident Prediction Model at Unsignalized Intersection Using Multiple Linear Regression Method", Nigerian Journal of Environmental Sciences and Technology (NIJEST), Vol 1, No. 2 July 2017, pp 379 - 389.
- [11] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, "Housing Price Prediction via Improved Machine Learning Techniques", 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019), Procedia Computer Science 174 (2020), 433 - 442.
- [12] Robart Konwar, Angshuman Kakati, Bhagyashree Das, Divyansh Borah Shah, Dr. Monoj Kumar Muchahari, "House Price Prediction Using Machine Learning", International Journal of All Research Education and Scientific Methods (IJARESM), Volume 9, Issue 6, June -2021, 3308 - 3316.
- [13] Shruti Shakhla, Bhavya Shah, Niket Shah, Vyom Unadkat, Pratik Kanani, "Stock Price Trend Prediction Using Multiple Linear Regression", International Journal of Engineering Science Invention (IJESI), Volume 7, Issue 10 Ver II, Oct 2018, pp 29 - 33.
- [14] Smita Rath, Alakananda Tripathy, Alok Ranjan Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model", Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14, (2020),1467 - 1474.

- [15] Stephen Gbenga Fashoto, Elliot Mbunge, Gabriel Ogunleye and Johan Van den Burg¹, "Implementation of Machine Learning for Predicting Maize Crop Yields using Multiple Linear Regression and Backward Elimination", *Malaysian Journal of Computing*, 6(1): (2021), 679 - 697.
- [16] Xiaojie Xu, Yun Zhang, "House price forecasting with neural networks", *Intelligent Systems with Applications*, 12 (2021), 200052, 1 - 13.
- [17] Yanming Yang, Lulu Sun, Chaoran Guo, "Aero-Material Consumption Prediction Based on Linear Regression Model", 8th International Congress of Information and Communication Technology (ICICT-2018), *Procedia Computer Science* 131, (2018), 825 - 831.
- [18] Yige Wang, "House-price Prediction Based on OLS Linear Regression and Random Forest", 2021 2nd Asia Service Sciences and Software Engineering Conference (ASSE '21), February 2021, 24 - 26.
- [19] Yin Libao, Yao Tingting, Zhou Jieliang, Liu Guicai, Liao Yanfen, Ma Xiaoqian, "Prediction of CO₂ emissions based on multiple linear regression analysis", The 8th International Conference on Applied Energy – ICAE2016, *Energy Procedia* 105, (2017), 4222 - 4228.
- [20] Yun Hwan Kim, Seong Joon Yoo, Yeong Hyeon Gu, Jin Hee Lim, Dongil Han, Sung Wook Baik, "Crop Pests Prediction Method using Regression and Machine Learning Technology: Survey", 2013 International Conference on Future Software Engineering and Multimedia Engineering, *IERI Procedia* 6, (2014), 52 - 56.
- [21] Zihan Yan, "Real estate price model based on multiple linear regression and analysis of influencing factors of regional economic vitality development", *Academic Journal of Business & Management*, Vol. 3, Issue 4: (2021), 18 - 21.