International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

A Simple Linear Regression with Statistical Approach for Building Machine Learning Model and Prediction

Jayasudha.J [1]

Assistant Professor, Department of Computer Science, Kovai Kalaimagal College of Arts & Science,

Coimbatore, Tamilnadu, India[1]

ABSTRACT: Machine Learning is one of the fastest growing areas in the field of AI. Even it is introduced several years before, it requires lot of data to achieve better results. Now in this world everything is based on data, so plenty of data available everywhere for our research. Machine learning is implemented with several algorithms classified into Supervised Learning, Unsupervised Learning and Reinforcement Learning. Each category is having several algorithms for Regression, Classification, Clustering, Outlier Detection, Association, Dimension Reduction etc..Here I have discussed the concept of Regression with Simple Linear Regression using Simple Dataset and Python Coding with its Statistical and Graph outputs which shows the use case that where the algorithm could be implemented and where it can't be implemented. Based on the predicted results of dataset Simple Linear Regression is best suited for this model and there is no much deviation with original and predicted value, so Error value should be a minimum one.

KEYWORDS: Machine Learning, Supervised, Unsupervised, Reinforcement, Simple Linear Regression, Independent Variable, Dependent Variable.

I .INTRODUCTION

Machine Learning is a subset of AI which allows machine to learn from the past data. Machine Learning is a study of computer algorithms that improves its performance through its experience [2]. Learning is done through observations, instructions, data and better decisions. The main aim of Machine Learning is to learn automatically without human assistance. Due to that it saves money and time and it works with lots of real-life applications. Our traditional software is not having capability of predicting and justifying different objects based on learning, but machine learning is doing that with more number of records and predicts the same for real-time data. AI took birth in 1950's, but due to unavailability of data, AI was not a successful one. But now due to vast amount of data it is moving forward to stamp its success. [10] Different types of Machine Learning algorithms are available as shown in Fig 1.1.

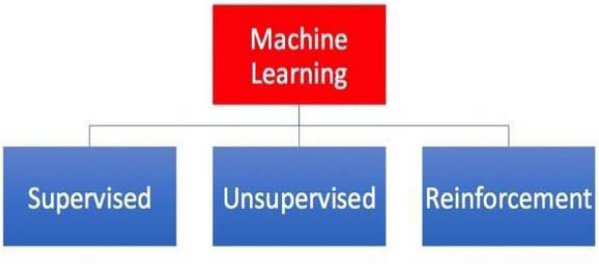

Fig.1.1. Machine Learning Algorithm Types

Supervised Learning: This type needs labeled data to train the model and based on that it predicts the new one. Here we are training the machine to learn the type of data through the input and then we are making the machine to learn for new inputs. Regression and Classification are the different types of Supervised Learning [2].
Unsupervised Learning: This type is not labeled into different categories. By using the implicit patterns or attributes it automatically identifies the results through similar features. Cluster Analysis is most widely used in Unsupervised Learning [2].

e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Reinforcement Learning: This is one of the hot areas of AI which reacts to the dynamic environment based on Rewards and Penalties. Here no training is provided and based on its results it reacts accordingly from its own experience [2].

II. LINEAR REGRESSION

Simple Linear Regression is a supervised machine Learning algorithm also called as Simple Linear Regression comes under the category of Regression, which is used for constructing efficient model for prediction, where the output is a continuous value and with a slope. It always works with Independent variable(X) and Dependent Variable(Y). This is a well-known algorithm for Machine Learning beginners. The concept behind this Linear Regression is every output is having some relationship in its input. So before selecting independent and dependent variables, we should make it very clear whether X affects the value of Y or not [1]. Linear Regression is classified into two types

- Simple Linear Regression (SLR).
- Multiple Linear Regression (MLR).

Normally Simple LR has one independent and one Dependent variable, but in case of Multiple LR it has many independent values are considered and with one Dependent variable. Mostly ML algorithms are designed based on Statistical model already defined. So both Simple and multiple LR is denoted using formulae.

SLR: $Y=\beta x+A$ (One Independent and Dependent Variable) [4], [7]
MLR: $Y=\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n+A$ (One dependent and More Independent Variables) [4], [7]
In the Second formula x1, x2…xn are all independent variables with one dependent variable (Y).
Y=Predicted Value
B=Slope/rate of Predicted
A=Y axis intercept=level of Y when X is 0. [8]

Independent Variable(X): An independent variable (X) is used to determine the value of dependent variable used for predicting values. This variable is also known as Explanatory variable.
Dependent Variable(Y): Here dependent variable depends on the independent variable and this value is the predicted one. This variable is known as Response Variable.

III. WHY LINEAR REGRESSION?

It is because of simplicity and it can be used in various sectors. Regression is used in applications like housing, Investment, Sales Forecasting, BMI prediction, Risk Analysis etc., everything is based on our datasets and it creates a linear line as output with plotted points. The variables X, Y we are choosing should be linear and the variance between results should not be too much as it is homoscedastic. Based on these assumptions LR is mostly selected by developers for a simple problem having continuous values as input. Straight line should be a best fit line to make a best linear model.
The relationship between X and Y is denoted using a line known as Regression line The line may be in three ways ,Positive slope , Negative Slope and Zero slope as shown in Fig.3.1

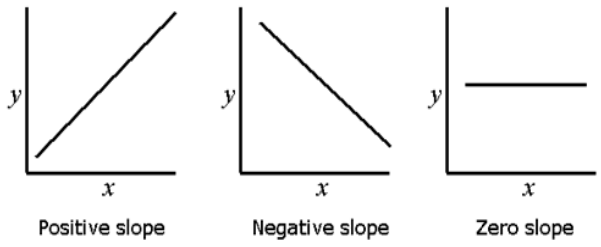

Fig.3.1. Different Regression Lines

The above Fig 3.1 indicates 3 different slopes that one may get as output. If the curve is Positive slope then our dataset is best suited for linear regression, if it is negative slope or Zero slope then it will not be considered as a best model.

e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Simple LR looks like this. Points are considered as our data in dataset and the line is known as Regression Line with a Positive slope [5] as shown in Fig.3.2.

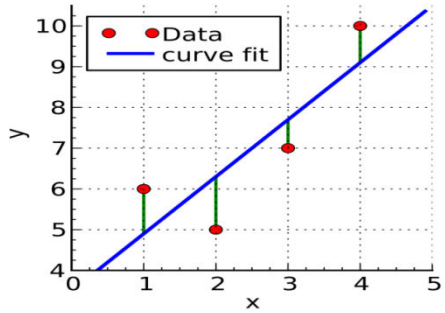

Fig.3.2 Regression Line with Positive Slope

IV. STATISTICAL CONCEPTS IN LINEAR REGRESSION

Our aim is to find the best fit line. Best fit line means the difference between the actual value Y and the predicted value Yp should be minimum [3]. This scenario is considered for all the values in the data set and it is summed up to find the error or deviation in it. In this way we may draw another line to find the error value. By comparing these two error values, select the line with minimum error ratio. This line is considered as a best fit line and Regression line and the predicted score or point for each value of X. The vertical points denote the errors of the prediction. Every point has small deviation and some with more deviation from original values.

area(X)	price(Y)
260	55000
300	56500
320	61000
360	68000
400	72500
420	75600
450	79000

Table.4.1 Area and Price Dataset

The above dataset in table 4.1 is used to construct a Linear Regression Model. Using this dataset new prices are predicted with different values for area (in Sq.ft).Before that a linear line is to be drawn for given values using the mean, standard deviation, correlation, slope and bias calculated using the formula.

M_X	M_Y	S_X	S_Y	r	β	A
358.571429	66728.57143	68.902968	9431.280275	0.990825	135.6218542	18098.4494

Table.4.2. Mean, Standard Deviation, Correlation, Slope and Bias Calculation

The slope (**β**) can be calculated as follows which is shown in Table.4.2.

$\beta = r. S_Y/S_X$ (0.990825 * 9431.280275/68.902968 = 135.6218542)

And the intercept (A) can be calculated as shown in Table.4.2.

$A = M_Y - \beta M_X$ (66728.57143-135.6218542 * 358.571429 = 18098.4494)

Where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line [6] in

e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

Fig. 3.2.is

$Y' = \beta x + A$
Y' = 135.6218542 (260) + 18098.4494. [8]
Y' = 53360.131459 (if the area=260)

In the same way other values are also predicted in the dataset in statistical method. Same is done in python using Jupyter Notebook with the help of libraries like numpy, Pandas, Matplotlib and Sklearn.

V. BUILDING LINEAR REGRESSION MODEL USING PYTHON

Several steps we have to follow before building a model, the following Fig.5.1. represents the steps in building a machine learning model [9].

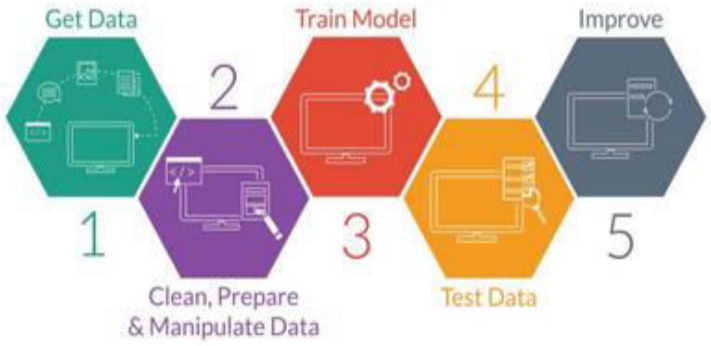

Fig.5.1. Steps to build a Machine Learning Model

Get Data: After deciding what we are going to predict, we have to collect the data and the values really affect the output based on the requirements. In our model I have chosen a simple dataset which is used to predict the price of sq.ft area. Only two column values with area and price are here. So the process of collecting data is done. Using python libraries import the dataset using the below coding [10].

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('house.csv')
X = dataset.iloc[:, 0:1]
y = dataset.iloc[:,1]
```

Clean, Prepare and Manipulate Data: After collecting data, we have to clean it, means data may be of any type and there may be null or missing values .So converting data type, filling the missed values using the mean values of that column should be done. Here we are not having such problem, so moving to training part.

Train Data: Usually based on our input ML algorithms learn and predict. So divide the dataset into training and testing set. Training is like making the machine to learn,Testing is like checking whether it observed the training to predict the new future values. But training and testing is done using our dataset only. Based on the number of records algorithms will give better result[10]. Here we are having very small data set. With that we are giving 80% for training and 20% for testing (test_size=0.2)

```
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

After training and testing we should check with the Linear Regression model in Sci-Kit using sklearn to import Linear Regression. Because in ML Sci-Kit learn is having many algorithms that we want to use,based on the model we've to import it. After that we have to fit our trained model into it [10].

```
# Fitting Linear Regression to the dataset
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(X, y)
```

After Training and Testing is done, we have to visualize the data to check whether our model is fit with Linear Regression [10]. The plot in Fig 5.2 with Regression Line implies this is a best fit line (Positive /slope) and the dataset fits on this model.

```
# visualizing the Linear Regression results
 def viz_linear():
    plt.scatter(X, y, color='red')
    plt.plot(X, lin_reg.predict(X), color='blue')
    plt.title('Price Prediction(Linear Regression)')
    plt.xlabel('Area(Sq.Ft)')
    plt.ylabel('Price(in Rs.)')
    plt.show()
    return
viz_linear()
```

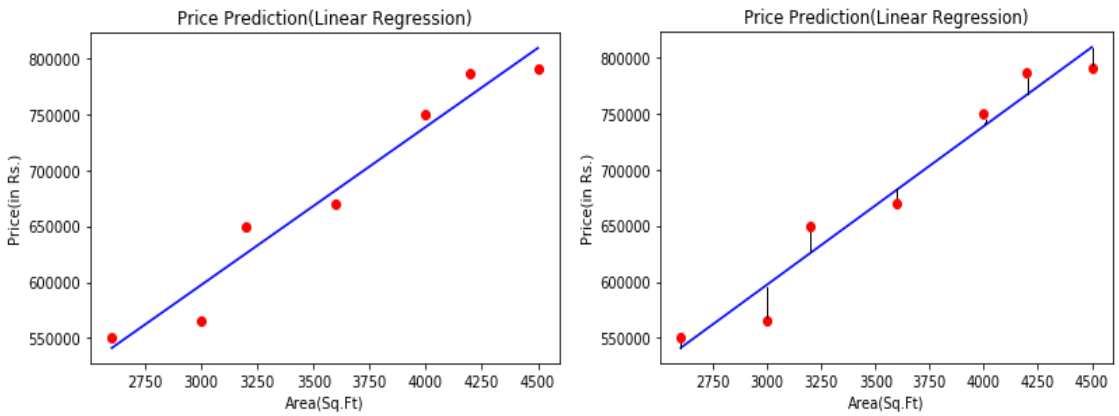

Fig.5.2. Regression Line for Dataset & Regression Line with deviations.

Improve: After Visualization [6] we will come to a conclusion that our dataset best fits to this model. That's why before choosing a model we should plot a graph to know, under which category or model this dataset will give best results. Half of the dataset is giving minimum error ratio while remaining giving more variations. If we have chosen a large dataset, we might have reduced that error ratio too. Without understanding the points dispersal we are unable to choose an algorithm. Now everything is done and we have to check our model with a new value of area: 460 as shown below. It gives the predicted value of 80484.50351053. By visualization we conclude that this will be a best fit for our model [10].

```
# Predicting the price for new area(values).
lin_reg.predict ([[460]])
```

Output: array ([80484.50351053])

By analyzing previous values we do have an idea of what will be the next value approximately, if it is not like that or any change then we have to give more records as input and more training and testing should be done and error checking also should be done with Error checking functions Mean Square Error, Root Mean Square error etc..This gives us the

e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 7, July 2020 ||

level of error that is occurred in our model. Based on that we have to retune it.

Fig.5.2 the dots are original points and the small line connecting the points(Y) and regression line (Y') is the variation a nd the difference in variation is listed in the below Table 5.3. Here Y-Y´ is the error of prediction and both positive erro rs and negative errors should be considered so we are squaring the values.

area(X)	price(Y)	Y´	Y-Y´	(Y-Y´)²
260	55000	53360.13	1639.87	2689172.3
300	56500	58785.01	-2285.01	5221247.9
320	61000	61497.44	-497.44	247448.9
360	68000	66922.32	1077.68	1161400.8
400	72500	72347.19	152.81	23350.4
420	75600	75059.63	540.37	292000.9
450	79000	79128.28	-128.28	16457.0

Table.5.3. Predicted output and error values.

VI. CONCLUSION

Linear Regression is a simple model to work with dataset of continuous values. Datasets may change and we require accurate values in case of Healthcare data, so based on the data linearity we are able to choose linear regression. If the data points are in curve shape we may use polynomial regression for better results because in such situation linear regression will not give better results as the line is non-linear one. Then Data Collection, Cleaning data, Training and Testing should be done with valid data to get good results. Even this is the simple algorithm that every beginner is interested and easy to learn, several algorithms like Decision Tree, Random Forest, Support Vector Machine etc.. are available. Depends upon the nature of the problem we select, Values in Datasets, choice of algorithm may differ for better results.

REFERENCES

[1] Kumari, Khushbu, and Suniti Yadav. "Linear Regression Analysis Study." Journal of the practice of Cardiovascular Sciences 4, no.1 (2018): 33.
[2] Pavithra, D., and A. N. Jayanthi. "A Study on Machine Learning Algorithm in Medical Diagnosis.", International Journal of Advanced Research in Computer Science 9, no. 4 (2018).
[3] Stanton, Jeffrey M. "Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors". Journal of Statistics Education 9, no. 3 (2001).
[4] Yildiz, Baran, Jose I. Bilbao, and Alistair B. Sproul, "A Review and Analysis of Regression and Machine Learning Models on Commercial Building Electricity Load Forecasting.", Renewable and Sustainable Energy Reviews 73 (2017): 1104-1122.
[5] Freedman, David A., "Statistical Models: Theory and Practice". Cambridge university press, 2009.
[6] Chan, Y. H. "Biostatistics 201: Linear Regression Analysis." Age (years) 80 (2004): 140.
[7] Schneider, Astrid, Gerhard Hommel, and Maria Blettner. "Linear Regression Analysis: part 14 of a Series on Evaluation of Scientific Publications." Deutsches Ärzteblatt International 107, no. 44 (2010): 776.
[8] De Leege, Arjen, Marinus van Paassen, and Max Mulder. "A Machine Learning Approach to Trajectory Prediction." In AIAA Guidance, Navigation, and Control (GNC) Conference, p. 4782. 2013.
[9] Zou, Kelly H., Kemal Tuncali, and Stuart G. Silverman. "Correlation and Simple Linear Regression.", Radiology 227, no. 3 (2003): 617-628.
[10] Swamynathan, Manohar, "Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python", Apress, 2019.