IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 3, March 2019

Selection of Best Algorithm for Regression Analysis of Various Algorithms in Data Science and Machine Learning

Supriya Patil

SCOE, Pune

Abstract: Regression Analysis one of the important techniques from data science, used for data analysis. In regression analysis the curve is designed so that the distance between curve and each point will be minimal. The association between or among dependent and independent variable is plotted with the help of this technique. Process is to observe the data, identify the relationship and trace the path for future so that the different values can use that path as a guideline. There are different types of regression based on number and nature of dependent and independent variable. This research article covers all mostly used Regression algorithm their properties usage; Based on the information about these algorithm, decision of selection of suitable algorithm can be made very easily. There different regression models are available, based on the applicability and frequency of use top 4 algorithms are considered for discussion. To support the research the mathematical equations and identity graphs are mentioned.

Keywords: Regression Models, Types of Regression, Classification and Regression, Selection of Appropriate Regression Model

I. INTRODUCTION

Classification and regression are two important technique used to categorize and analyses the data respectively. Once the data is classified by any one of the classification algorithm data possess similar properties, it means the data is of homogeneous in nature. Analysis of this data can be made easier with the help of regression technique. As regression requires training to select the path and testing to know the deviation of the target variable from path almost all algorithms are supervised algorithms. Basic structure of any algorithm requires two factors i) dependent variable ii) independent variable

Dependent Variable (Y):
Output or decision can be made with the help of value of this variable, it depends upon independent variable or variables.

Independent Variable (or variables) (X):
Input or decisive parameters decides value of Y. Values of these independent variable does not depend upon Y. For Example, how to decide the best car for me? There are various factors affecting the decision of buying the car. Throughout the paper same example is considered. For the sake of understanding the algorithm or approach correctly the situation or the number of factors will be altered accordingly. Here Dependent variable is decision to buy or not, Independent variables are budget, Mode of ignition, color, make, class, average, engine power etc. (while discussing the various types of the regression algorithms these variables will be selected as per the requirement)

Simple Linear Regression:
It is the most basic type of relapse. It is a method in which the reliant variable is persistent in nature. The connection between the needy variable and autonomous factors is thought to be straight in nature. We can see that the given plot speaks to a by one way or another straight connection between the mileage and dislodging of vehicles. The green focuses are the genuine perceptions while the dark line fitted is regression line.

As mentioned in this graph the observations are mentioned in the graph and the line which covers the maximum number of points with minimum possible distance from the regression line. If the example of buying the car is considered, then in case the cost is only parameter deciding the purchase of car then linear regression is best fit model to give the solution. Equation of the simple linear regression is

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 3, March 2019

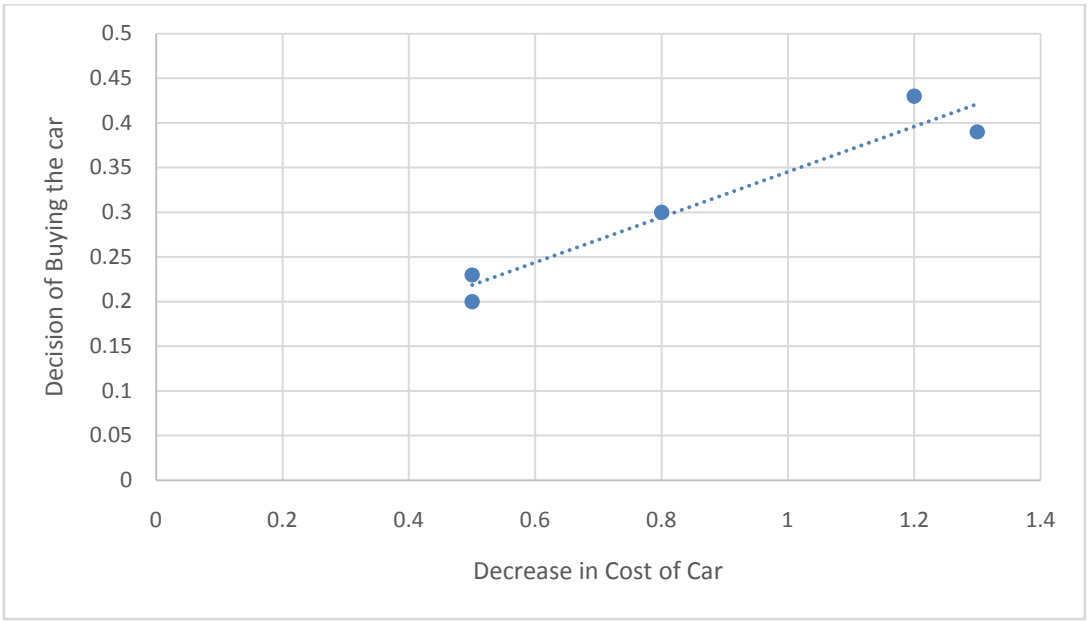

Figure 1: Simple Linear Regression

Y= a0+a1*x1+a2*x2+…+an*Xn

Where, a is coefficient, b is bias, y is dependent variable and x is independent variable, once the history or observations are collected Stochastic Gradient Descent help to decide values of weight used n regression. Epsilon or error can be also introduced to make the model more realistic. The range of values may cross the limit of 0 to 1 range.

Polygon Regression Model

In several situations data points cannot be covered with the help of straight line. That is why polynomial function is deployed to cover nonlinear points. Polynomial function is applied to make the decision to purchase the car. This graph reveals that as the increases in up to certain extent the decision of purchasing the car is resembling the straight line but from their onwards the decision making does not follow linear mode. As CNG+Petrol of Petrol+Electrical model drives the purchasing of a car.

Identity Equation of polynomial equation is

Y=a0+a1*x1+a2*X2^2+…+ a3*X3^3

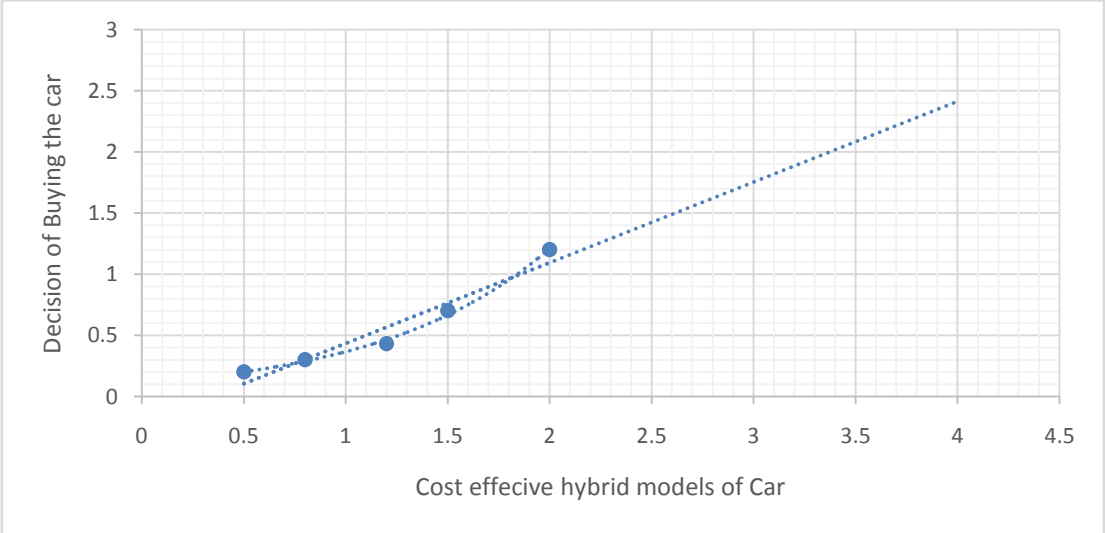

Figure2: Non – Linear (Polygon Regression Model)

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 3, March 2019

Ridge Regression Model:
Ridge regression model is used to solve the over fitting problem by considering the regularization or normalization process by introducing the bias or influencing factor. It is very difficult to clearly distinguish if there is very less variance among the parameter. To resolve this issue the influencing factor is introduced to change the structure of the existing graph so that different parameter can separately contribute in a decision-making process. In the example of car purchasing build quality is directly proportional to cost of the car, both curves may go hand in hand to clearly introduce decisive property by each Ridge regression model is used.

Identity equation is

$$Y=a+bx$$

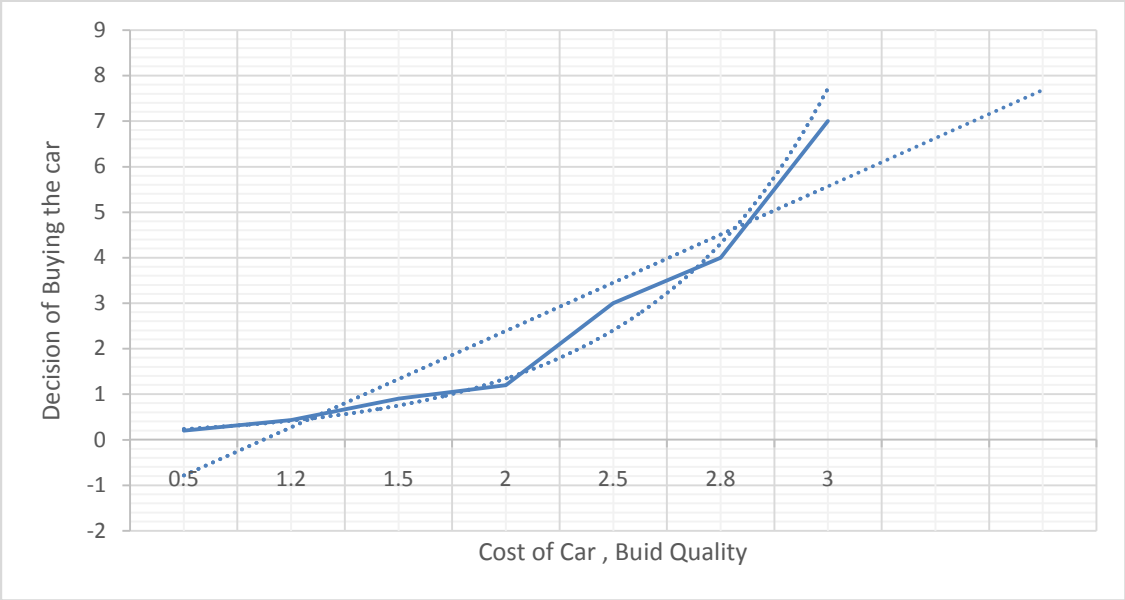

Figure3: Ridge Regression

Logistic Regression
Logistic regression is used when the response expected is binary in nature, that is dependent variable must be binary, independent variable might be of any form. Curve is "s" shaped curve used to deal with the problem of non-linearity. Identity equation of the logistic regression is

$$Y=(1/(1+e^{-xb}))+E$$

Where Y is output, x is an input, b is coefficient and E denotes error as it does not follow normal mode of distribution. In the example of cur purchasing the color of car may not have linear association with decision of buying a car. This is normalized by changing the curve by taking log of odds.

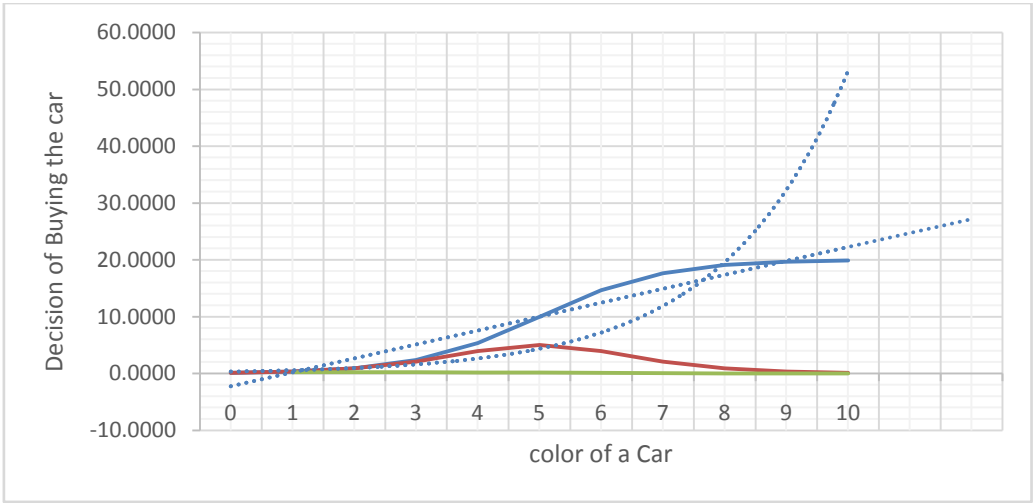

Figure 4: Logistic Regression

II. COMPARATIVE ANALYSIS

Regression Modeling requires thorough analysis of problem statement. To decide the suitable model of regression understanding of number of dependent parameters is necessary. Along with number the interdependent relation among these independent variables also plays the crucial role in the process of decision making. Behavior of dependent variable with respect to independent variable is required to be examined carefully. These careful observations help in deciding the nature of a curve. Once the curve type is decided further projection can be plotted easily. Linearity, Multicollinearity, Overfitting, underfitting of the curve, selection of sampling, selection of the distribution, visualization of error these are few more parameters which are supposed to be considered in order to make the regression correctly.

III. CONCLUSION AND FUTURE WORK

Nature of curve along with identity equation and consideration of one example through- out helped in understanding of regression and the process of applying it. The problem of decision making of car purchase based on parameter like cost, color, brand, built quality and type of engine is considered to analyze the suited regression model for a given problem and situation. Along with the regression model used in this research paper there are few more regression model available which may also extend the discussion of selection of suitable model further. Scope of this paper is restricted for four algorithms; rest of algorithm or approaches might be covered during the implementation of system. This is considered as a future scope to be completed.

REFERENCES

[1]. Philip Russom, "Big Data Analytics", TDWI Best Practices Report, 2011.
[2]. Alfredo Cuzzocrea, Il-Yeol Song, C. Davis Karen, "Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!", DOLAP'11, October 28, 2011.
[3]. James R. Evans, Carl H. Lindner, Business Analytics: The Next Frontier for Decision Sciences, Decision Science Institute, March 2012.
[4]. Surajit Chaudhuri, Umeshwar Dayal, Vivek Narasayya, "An Overview of Business Intelligence Technology", Communications of the ACM, vol. 54, no. 81.1, August 2011.
[5]. Shmueli Galit, Otto R. Koppius, "Predictive Analytics in Information Systems Research", Mis Quarterly, vol. 35, no. 3, pp. 553-572, September 2011.
[6]. RS Michalski, JG Carbonell, TM Mitchell, Machine learning: An artificial intelligence approach, Springer-Verlag, 2013.
[7]. S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", J. of Informatica, vol. 31, pp. 249-268, 2007.
[8]. Gaf Seber, AJ Lee, "Linear regression analysis", Wiley Series in Probability and Statistics, 2012.
[9]. DC Montgomery, EA Peck, GG Vining, "Introduction to linear regression analysis", Wiley Series in Probability and Statistics, 2015.
[10]. Zhang Xuegong, "Introduction to Statistical Learning Theory and Support Vector Machines", Acta Automatica Sinica, 2000.