IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 12, December 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

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

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112045 |

Regression Analysis in Data Science

Prathamesh Suryaji Mahadik , Swaraj Sayajirao Desai, Dr. Sampada Gulavani, Prof. Keerti Kadam

MCA Student, Bharati Vidyapeeth (Deemed to be University), Pune, Institute of Management, Kolhapur, India

MCA Student, Bharati Vidyapeeth (Deemed to be University), Pune, Institute of Management, Kolhapur, India

Department of MCA, Bharati Vidyapeeth (Deemed to be University), Pune, Institute of Management, Kolhapur, India

Department of MCA, Bharati Vidyapeeth (Deemed to be University), Pune, Institute of Management, Kolhapur, India

ABSTRACT: In the statistical literature predicting or learning numeric features is called regression. It is the subject of research in both Statistics learning and Machine learning. Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variable (known as independent variable). This paper reviews important techniques and algorithms for regression. Regression analysis is used to predict how far the change in the value of the dependent variable is due to the independent variable. Also the position of variables in regression analysis are not equal because one is the dependent variable and the other is the independent variable

KEYWORDS*:* Regression; Prediction; intercept; correlation; variables.

I. INTRODUCTION

Regression Analysis is a statistical process for estimating the relationships between the dependent variables or criterion variables and one or more independent variables or predictors. Regression analysis explains the changes in criteria in relation to changes in selected variables. For example, we use parents' height to predict the height of their children. In this case, the predictor variable is "Parents' height" and Dependent Variable is the "height of children". This is because its value is depending on the value of the Independent (Predictor) variable value . In such type examples, the regression analysis comes into life. It is used in numerous disciplines including Machine Learning and Artificial Intelligence. In machine learning, most research has been done for classification where the single predicted feature is discrete or nominal. Regression differs from classification in that the output or predicted features in regression problem is continuous. By considering importance of regression analysis, this paper consist of introduction of regression analysis, application of regression analysis is different fields and use of different regression techniques in data mining.

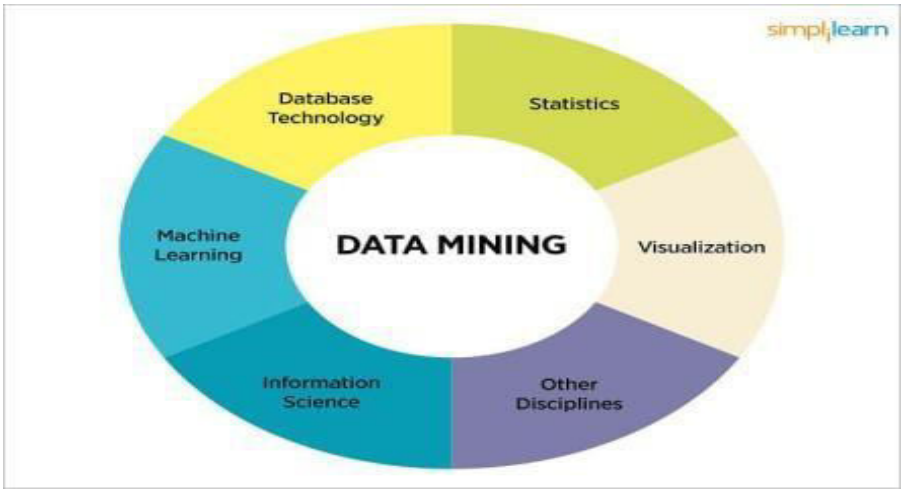

Fig 1 : Use of Data Mining in Different Discipline

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112045 |

1. What is Regression Analysis:

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them. When domain expert is not available, or the knowledge of expert is implicit then induction technique is developed in machine learning . These techniques are used to discover new rules, even in cases where formal knowledge is available. The most important advantage of induction techniques is that they enable us to extract knowledge automatically.The idea behind using induction for knowledge system is particularly accepted by a newly emerged discipline, Knowledge Discovery in Database (KDD) incorporates researches from various disciplines. The main sources of knowledge is a large databases, since they can store large amount of data belonging to many different domains. The use of automatic methods such as induction for knowledge discovery is viable, because it is difficult to find an expert for each different domain in database

2.Applications of Regression Analysis in Finance :

Most of the regression analysis is done to carry out processes in finances. By considering this, different applications of Regression Analysis in the field of finance and others relating to it is discussed below :

i)Forecasting : The most common use of regression analysis in business is for forecasting future opportunities and threats. For example, demand analysis, forecasts the amount of things a customer is likely to buy. When it comes to business, though demand is not the only dependent variable. Regressive analysis can anticipate significantly more than just direct income. Insurance firms
depend extensively on regression analysis to forecast policy holder credit worthiness and the amount of claims that might be filed in a particular time period.

ii)CAPM : The Capital Asset Pricing Model (CAPM), establishes the link between an asset's projected return and the related market risk premium, relies on the linear regression model. It is also frequently used by financial analysts to anticipate corporate returns and operational performance. The beta coefficient of a stock is calculated using regression analysis. Beta is a measure of return volatility in relation to the total market risk. Because it reflects the slope of the CAPM regression, we can rapidly calculate it in Excel using the SLOPE tool.

iii)Comparing with Competition : It can be used to compare a company's financial performance to that of a certain counter-part. It may also be used to determine the relationship between two firms' stock prices (this can be extended to find correlation between 2 competing companies, 2 companies operating in an unrelated industry etc). It can assist the firm in determining which aspects are influencing their sales in contrast the comparative firm. These techniques can assist small enterprises to achieve rapid success in a short amount of time.

iv)Identifying Problems : Regression is useful not just for providing factual evidence for management choices, but also for detecting judgement mistakes. A retail store manager, for example, may assume that extending shopping hours will significantly boost sales. However, RA might suggest that the increase in income isn't enough to cover the increase in operational cost as a result of longer working hours . As a result, this research may give quantitative backing for choices and help managers to avoid making mistakes based on their intuitions.

v)Reliable Source : Many businesses and their top executives are now adopting regression analysis to make better business decisions and to reduce guess work. Regression enables firms to take a scientific approach to management. Both small and large enterprises are frequently bombarded with an excessive amount of data. Managers may use regression analysis to filter through data and choose the relevant factors to make the best decisions possible.

3.Types of Regression Techniques : Each regression technique has some assumptions attached to it which we meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution. The techniques are discussed below :

i)Linear Regression : Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables . The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112045 |

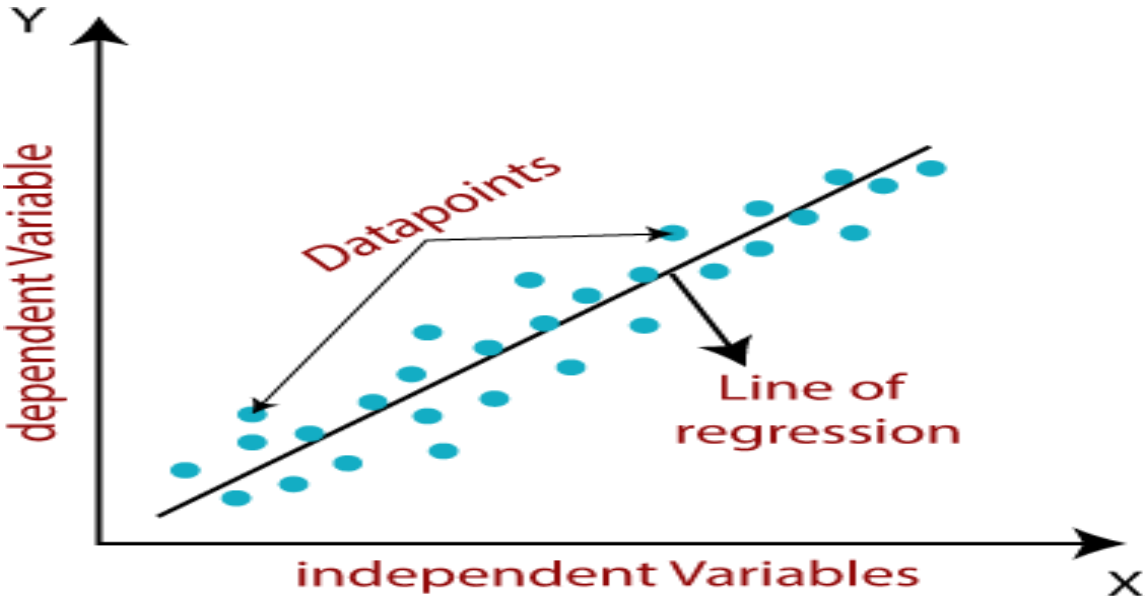

Fig.2 : Linear Regression Equation

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:

ii) Polynomial Regression : It is a technique to fit a non-linear equation by taking polynomial functions of independent variable. In the figure given below, you can see the red curve fits the data better than the green curve. Hence in the situations where the relation between the dependent and independent variable seems to be non-linear we can deploy Polynomial Regression models.

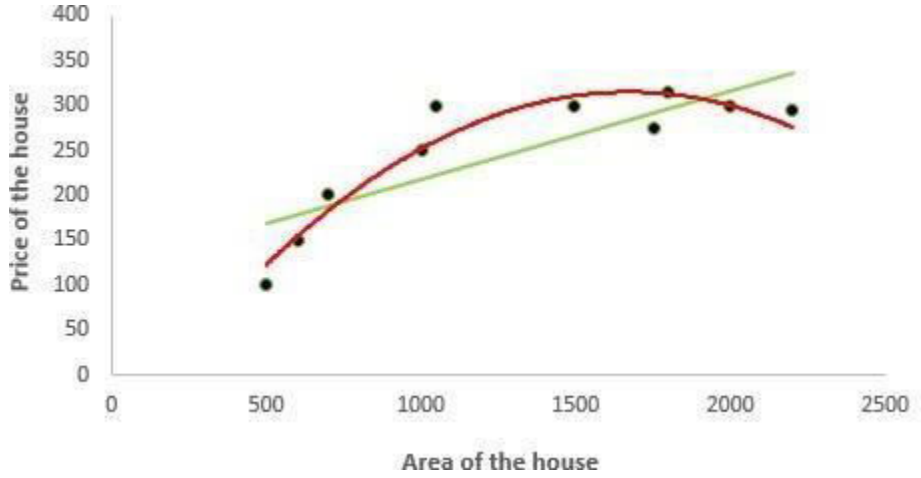

Fig. 2 : Polynomial Regression

Here we can create new features like

$X_1 = x, X_2 = x^2, \ldots, X_k = x^k$

and can fit linear regression in the similar manner.. In case of multiple variables say X1 and X2, we can create a third new feature (say X3) which is the product of X1 and X2 i.e.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112045 |

$X_3 = X_1 * X_2$

iii) Logistic Regression : In logistic regression, the dependent variable is binary in nature and independent variables can be continuous or binary. In multinomial logistic regression, you can have more than two categories in your dependent variable. Here my model is:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + --- + b_k X_k)}}$$

logistic regression equation

iv)Ridge Regression : It is a technique for analyzing multiple regression data. When multicollinearity occurs, least squares estimates are unbiased. A degree of bias is added to the regression estimates, and as a result, ridge regression reduces the standard errors.

v)Lasso Regression : It is a regression analysis method that performs both variable selection and regularization. Lasso regression uses soft thresholding and selects only a subset of the provided covariates for use in the final model.

vi) Elastic Net Regression : Elastic Net regression is preferred over both ridge and lasso regression when one is dealing with highly correlated independent variables.
It is a combination of both L1 and L2 regularization.
The objective function in case of Elastic Net Regression is:

$$\text{Min}\ (\sum \varepsilon^2 + \lambda_1 \sum \beta^2 + \lambda_2 \sum |\beta|) = \text{Min}\ \sum (y - (\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k))^2 \lambda_1 \sum \beta^2 + \lambda_2 \sum |\beta|$$

4.Conclusion : Regression techniques are the most sought after data analytical methodologies. These techniques are employed by data scientists in a wide range of fields such as banking, insurance, retail, pharmacy, e-commerce and caried other fields. Mostly the linear and logistic regressions are induced into the data inventory by those companies which indulge in generating a large amount of data. Even though much research has been done on regression, two important things need to be done like the fitness of the estimation or accuracy and the other is the interpretability of the constructed model. New research can be conducted in these traditional directions. Additionally, research need to be directed to increase the efficiency, both in computational complexity and storage. Also it requires to pay more attention to the construction systems that deal with outliers, noise, missing values and irrelevant features is very important, since databases today have a very large number of records and attributes.

REFERENCES

1.https://www.geeksforgeeks.org/types-of-regression-techniques/
2. https://www.analyticsvidhya.com/blog/2021/07/all-you-need-to-know-about-polynomial-regression/
3.www.researchget.com
4.www.wikipedia.com

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details