Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 10, October -2018

PREDICTIVE ANALYSIS USING BIG DATA IN HEALTHCARE SECTOR

[1]B.Lalithadevi, [2]Anirudh Bansal, [3]Sayon Bhattacharyya, [4]Param Saluja

[1] SRM IST Chennai Tamil Nadu 600089
[2] SRM IST Chennai Tamil Nadu
[3] SRM IST Chennai Tamil Nadu
[4] SRM IST Chennai Tamil Nadu

Abstract - *Health care analysis is a term used to describe the activity of analysing healthcare data from a variety of sources. These data contain information about cost, claim, pharmaceutical, clinical and patient behaviour. Electronic Medical Records contains appropriate data to perform analysis and highlight on different aspect of the data .The health care industry is a growing industry in the world .The healthcare data is growing with the advancement of time. So, we need to utilize these data in the most effective way. Hugh volumes of healthcare data remains under utilized and wasted .A definite procedure must be followed to capture , store and analyse the data set .We can achieve this idea only if we design an architecture which allows the storage of large data set and perform analysis on them . Hadoop is the best architecture to work on when large data sets are involved .Apache Hadoop is an open source ecosystem for handling large volumes of data using distributed processing.*

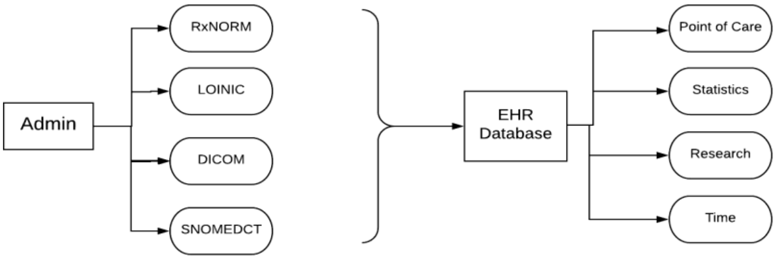
1. INTRODUCTION

Big Data Analytics is one of the hyped topics in Healthcare sector. Analysis is done in the area of medicine, effects, diseases etc. Based on the analysis, predictive analysis can be performed which can improve patient care, chronic disease management and hospital management. Healthcare analysis provides detailed information about the past and the present. This makes it convenient to perform analysis on a particular characteristic of data and predict the future scope.

To start with, a Data Warehouse is to be implemented. A Data Warehouse is a warehouse which will store volumes of healthcare data. All essential data related to healthcare should be available in the Warehouse. The integration of new data sources into the existing data warehouse system will further empower corporations and deeper analytics and insights. Using Hadoop as a warehouse will optimize performance, scalability and is cost effective

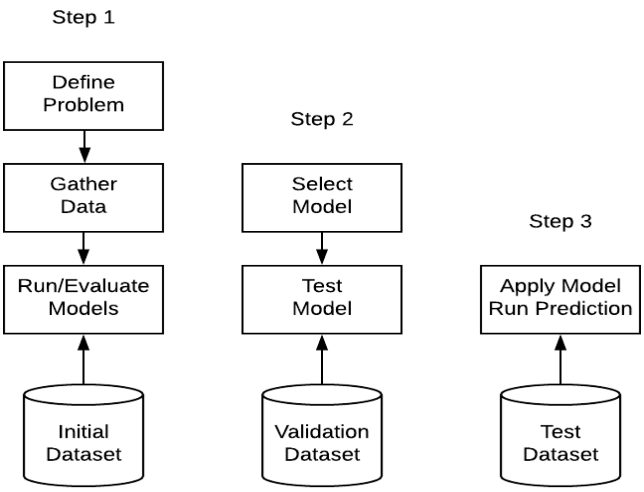
Big Data Analytics is one of the hyped topics in Healthcare sector. Analysis is done in the area of medicine, effects, diseases etc. Based on the analysis, predictive analysis can be performed which can improve patient care, chronic disease management and hospital management. Healthcare analysis provides detailed information about the past and the present. This makes it convenient to perform analysis on a particular characteristic of data and predict the future scope. To start with, a Data Warehouse is to be implemented .A Data Ware house is a warehouse which will store volumes of healthcare data. All essential data related to healthcare should be available in the Warehouse. The integration of new data sources into the existing data warehouse system will further empower corporations and deeper analytics and insights. Using Hadoop as a warehouse will optimize performance, scalability and is cost effective.

Figure -1: Data Flow



For predictive analysis to be effective, the data needs to be understood. First, we need to define the problem then gather the necessary data and evaluate different algorithm approaches. Second, we improve the process by selecting the best performing models and testing it with different data set to validate the working of the model. The final step is to run the model in the real world. Prescriptive analytics is a more specific term which includes evidence, recommendations and actions for each predicted outcome. Prediction should link to clinical priorities and measurable events such as cost effectiveness, clinical protocols or patient outcomes. These predictor-intervention sets are best evaluated within that same data warehouse environment. Many options exist when it comes to developing predictive algorithms. This gives a challenge to health care. Healthcare providers need to find a way to find a way to team up with expertise to develop appropriate prediction models.

Figure -2: Steps Involved



2. OUTCOMES

Analyzing healthcare data is an important task in healthcare industry. With proper unbiased analyzation, correct decisions can be taken. Usually the outcomes can be categorized into Economic Cost, Survival Prediction and Diagnostic .Economic Cost refers to the cost involved in medicines and treatments. Cost of medicines and treatments are increasing day by day. People are finding it hard to spend so much in healthcare. It is witnessed that a lot of money gets wasted in the healthcare industry due to underutilization of resources. Proper utilization of resources will take the industry to a new level. The wastage of money can be analyzed by understanding the economic cost involved in it. Economic cost analysis will give us a goal to reduce money wastage and to produce cost effective quality product in the future. The cost can be predicted using a Linear Regression model. A linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. It is called simple linear regression model if there is only one independent variable or else it called multiple linear regression. The goal of the model is to predict, forecast and reduce error. By fitting the model to a set of observed data, additional values can be collected without the knowledge of independent variables. The fitted model can be used to make a prediction of the response. Variation in response variable can also be explained by quantifying the strength between response and explanatory variables.

It can be represented in a simpler form in which y is a dependent variable and X is a independent variable.

Diagnostic is the process of diagnosing a patient concerned with illness or any other health related problems. Diagnostic involves presence of a disease and name of the disease.

Figure -3: Formula

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf T} \boldsymbol\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

$$\mathbf{y} = X\boldsymbol\beta + \boldsymbol\varepsilon,$$

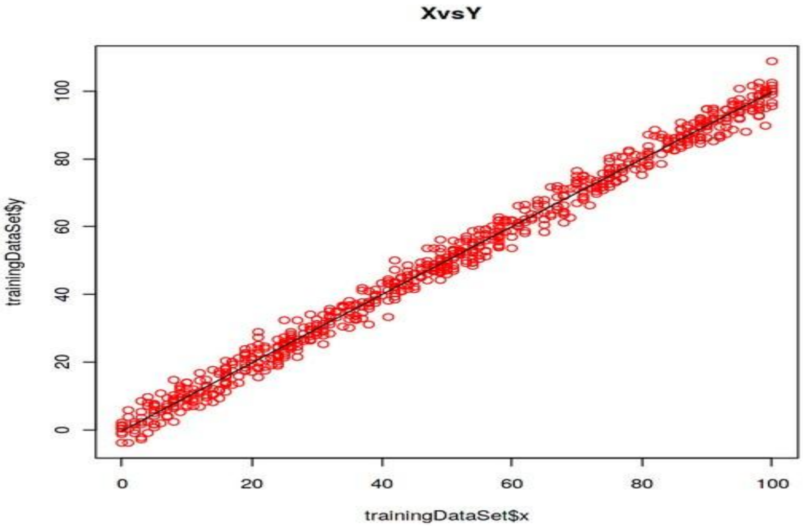
where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf T} \\ \mathbf{x}_2^{\mathsf T} \\ \vdots \\ \mathbf{x}_n^{\mathsf T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

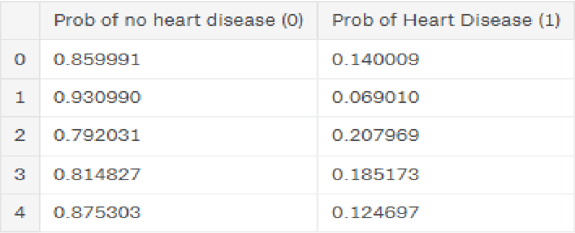
$$\boldsymbol\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Figure -4: Linear Regression



Diagnosis of a sick patients shows the presence of a disease. If not treated properly, the patient's condition may deteriorate. Sometimes, the disease may spread from one person to another if the disease is not treated well. Lack of treatment can result in worsening of health of the patient. It is advised to develop a good diagnostic model which detects the presence of disease and suggests proper medication and treatment. The presence of a disease is a binary outcome involving yes or no whereas the suggestion of treatments and medication involves categorical outcomes. Each category of disease has a set of precaution measures and treatments. Binary outcomes of diseases can be performed by using Logistic Regression Model. Logistic Regression Model is a statistical model which is used to predict binary outcome when binary dependent variables are applied on it. Logistic model is the model in which the probability is in the linear combination of independent variables. The two independent variables are 0 and 1 which represents outcomes like dead/alive, win/lose, healthy/sick. The goal of the model is to predict the presence of disease. P is defined as the probability of the event. The range of p lies between 0 and 1. P value closer to 1 state True nature whereas p value close to 0 denotes its false nature.

Figure -5: Probabilities

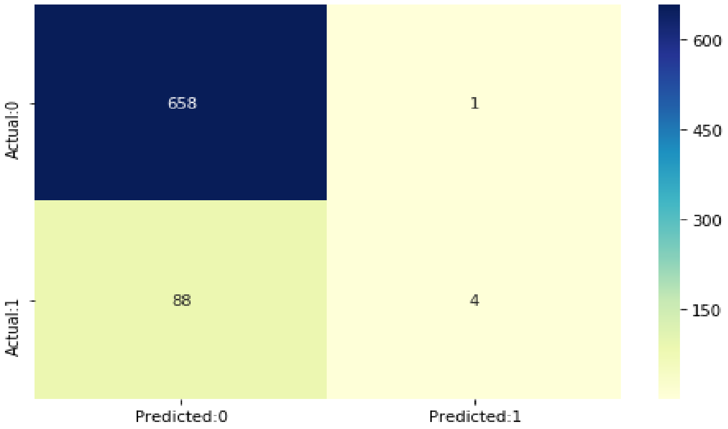


	Prob of no heart disease (0)	Prob of Heart Disease (1)
0	0.859991	0.140009
1	0.930990	0.069010
2	0.792031	0.207969
3	0.814827	0.185173
4	0.875303	0.124697

Figure -6: Formula

$$\operatorname{logit} p = \ln \frac{p}{1-p} \quad \text{for } 0 < p < 1.$$

Figure -7: Prediction



Presence of a disease is an important factor because the treatments will be based on the specified disease. Any error in detection of disease will cause a problem in analyzation of its prevention cures. The model should be fitted to a set of error free observed data. Then, the model should be tested with other datasets of the world to ensure its working. The presence of a disease was the first step, after the disease is detected preventive measures need to be considered. To ensure that, categorical outcome model is needed. Categorical model contains the disease name and its necessary treatments. Multinomial Logistic Regression is a classification method that is used to predict the probabilities of the different possible outcomes of a categorical dependent variable. It is similar to logistic regression model except the dependent variables are categorical rather than binary. In this model, there are K outcomes instead of 2 outcomes.

Figure -8: Formula

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_1 \cdot \mathbf{X}_i$$
$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_2 \cdot \mathbf{X}_i$$
$$\cdots\cdots$$
$$\ln \frac{\Pr(Y_i = K-1)}{\Pr(Y_i = K)} = \boldsymbol{\beta}_{K-1} \cdot \mathbf{X}_i$$

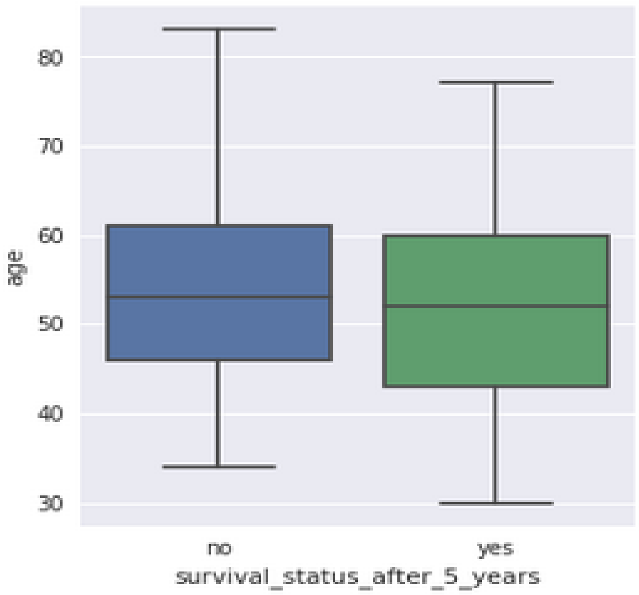
There are K outcomes and there is categorical dependent variable. Based on the health condition and the stage of the disease, the survival of a patient can be determined. The method of predicting the survival of patient is called survival prediction. This prediction can be done with the help of Survival Analysis model. Survival Analysis model focuses on analyzing the current event and predicting the duration of time until one or more events happen. Usually, survival data are not observed rather they are censored. In the survival analysis, primary focus is on the survival function.

Figure -9: Formula

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)\, du = 1 - F(t).$$

S(t) denotes the survival function.
Function f is event density. It is the rate of success or failure of a given event.
Function F is lifetime distribution function which denotes completion of survival function.

Figure -10: Survival Regression



3. CONCLUSIONS

We can conclude that big data analytics plays an important role in healthcare industry. It not only studies the data but predicts the future outcome. The prediction utilizes every aspect of the data and reduces waste of data. Predictive analysis promises to bring change towards a better system for the healthcare industry by minimizing wastage of data and maximizing the accuracy of analysis.

REFERENCES

[1] Nalini Priya. G, Kannan. An Empirical Study on Multivariate Data in Medical Decision Making Environment
[2] Chetna Kaushal, Deepika Koundal Big Data Application in Medical Domain
[3] M. G. Ruano, G. P. Almeida, F. Palma, J. F. Raposo, R. T. Ribeiro Reliability of Medical Databases for the use of Real Word Data and Data Mining Techniques for Cardiovascular Diseases Progression in Diabetic Patients
[4] Hang Zhao, Guijie Li, Wei Feng Research on Visualization and Application of Medical Big Data.
[5] Linear Regression Wikipedia https://en.wikipedia.org/wiki/Linear_regression Logistic Regression Wikipedia https://en.wikipedia.org/wiki/Logistic_regression.
[6] Multinomial Logistic Regression Wikipedia https://en.wikipedia.org/wiki/Multinomi.
[7] Survival Analysis Wikipedia https://en.wikipedia.org/wiki/Survival_analysis.