A STUDY ON PREDICTIVE ANALYTICS USING REGRESSION ALGORITHMS

S. Banumathi[1] , A. Aloysius[2]
1Assistant Professor, Department of Computer Application, Bishop Heber College, Trichy, Tamil Nadu, India. {banu15.sm@gmail.com}
2Assistant Professor, Department of Computer Science, St. Joseph's College, Trichy, Tamil Nadu, India. {aloysius1972@gmail.com}

*Abstract--**Big Data analytics is the technique of collecting, organizing and analyzing large sets of information to determine patterns and other uses in sequence. Big Data analytics can assist organizations to better understand the information contained in the data and will also help to identify the data that is most important to the trade and future dealing decisions. Predictive analytics does not inform what will happen in the future but it estimates what might happen in the future with an acceptable level of consistency, and includes imagine scenarios and risk consideration. This research article discussed types of regression algorithms: linear regression, logistic regression, and polynomial regression with statistical analysis process for estimating the relationships among variables.***

Keywords: Predictive Analytics, Big data, Regression, Prediction Models, Supervised Learning

I. INTRODUCTION

Big data is a collection of data sets so large and complex that it becomes complicated to process using readily available analysis tools. The challenges consist of capture, storage, search, sharing, transfer, analysis, and visualization [24]. Big data possibly will be found in three forms they are Structured, Unstructured and Semi-structured. Structured data refers to any data that reside in a preset field within a record or file. These include data contained within relational databases and spreadsheets. Structured data has the benefit of being easily entered, stored, queried and analyzed. Unstructured data is the contrary of structured data. Structured data usually resides in a relational database, and as a result, it is sometimes called relational data. This type of data can be plainly mapped into pre-designed fields. By dissimilarity, unstructured data is not relational and doesn't fit into these sorts of pre-defined data models. In addition to structured and unstructured data, there's also another category semi-structured data. Semi-structured data is information that doesn't exist in in a relational database but that does have some executive properties that make it easier to analyze. Big data can be described by the following characteristics: Volume, Variety- Variability, and Complexity. Big Data Analytics largely involves collecting data from different sources, manage it in a way that it becomes available to be consumed by analysts and finally deliver data products useful to the organization business. The method of converting fat amounts of unstructured raw data, retrieved from dissimilar sources to a data product useful for organizations forms the core of Big Data Analytics.

II. TYPES OF BIG DATA ANALYTICS

A. Prescriptive Analytics

The most valuable and most underused big data analytics technique, prescriptive analytics gives a laser-like focus to answer a specific question. This analytics is used to determine the best solution among a variety of choices, given the known parameters and suggests options for how to take advantage of a future opportunity or mitigate a future risk. Prescriptive analytics can also illustrate the implications of each decision to improve decision-making.

Table 1

Types of analytics

Prescriptive Analytics	**Diagnostic Analytics**	**Descriptive Analytics**	**Predictive Analytics**	**Outcome Analytics**
Forward-looking	Backward-looking	Backward-looking	Forward-looking	Backward-looking, Real-time, and Forward-looking
Focused on optimal decisions for future situations	Focused on causal relationships and sequences	Focused on descriptions and comparisons	Focused on non-discrete predictions of future states, relationship, and patterns	Focused on consumption patterns and associated business outcomes
Simple rules to complex models that are applied on an automated or programmatic basis	The relative ranking of dimensions/variable based on inferred explanatory power)	Pattern detection and descriptions	Description of prediction result set probability distributions and likelihoods	Description of usage thresholds
Discrete prediction of individual dataset members based on similarities and differences	Target/dependent variable with independent variables/dimensions	MECE (mutually exclusive and collectively exhaustive) categorization	Model application	Model application
Optimization and decision rules for future events	Includes both frequenters and Bayesian causal inferential analyses	Category development based on similarities and differences (segmentation)	Non-discrete forecasting (forecasts communicated in probability distributions)	Model application

B. Diagnostic Analytics

Data scientists turn to this technique when trying to determine why something happened. This analytics is useful when researching leading churn indicators and usage trends amongst most loyal customers. Examples of diagnostic analytics include churn reason analysis and customer health score analysis.

C. Descriptive Analytics

This technique is the most time-intensive and often produces the least value. However, Descriptive analytics is useful for uncovering patterns within a certain segment of customers. Descriptive analytics provide insight into what has happened historically and it will provide trends to dig into in more detail. Examples of descriptive analytics include summary statistics, clustering and association rules used in market basket analysis.

D. Predictive Analytics

Predictive analytics is most commonly used technique. Predictive analytics use models to forecast what might happen in specific scenarios. Examples of predictive analytics include next best offers, churn risk and renewal risk analysis [24].

E. Outcome Analytics

It refers to as consumption analytics, this technique provides insight into customer behavior that drives specific outcomes. This analysis is meant to help to understand customers better and learn how customers are interacting with products and services.

III. PREDICTIVE ANALYTICS

Predictive analytics extracts information from data sets in order to discover complex relationships, recognize unknown patterns, forecasting actual trends, find associations, etc. This allows us to anticipate the future and make the right decisions [10]. The applications of predictive analytics in business intelligence are uncountable. In business, as in life, the more you know about a likely outcome, the more confident you will be that the decision you are about to make is the right one. Predictive analytics can give you an idea of every possible probability so your team and your organization can assess the risks, the pursuant actions, and the potential for better managing results.

IV. REGRESSION

The statistical approach to forecasting vary in a dependent variable on the basis of change in one or more independent variables Regression analysis is a structure of predictive modeling method which investigates the association between a dependent (target) and independent variable (s) (predictor). There are various kinds of regression techniques available to construct predictions. These techniques are generally driven by three metrics (number of independent variables, type of dependent variables and shape of the regression line).

A. Linear Regression

It is one of the mainly well-known modeling techniques. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.In this method, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, with nature of regression line is linear. It is represented by an equation

$Y=a+b*X + e$ .....(Eq: 1)

, where a is an intercept, b is the slope of the line and e is error term. This equation can be used to expect the value of an objective variable based on given predictor variable(s).The difference between simple linear regression and multiple linear regression is that multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

B. Logistic regression

Logistic regression is a statistical method used to analyze a dataset in which there are one or extra independent variables that verify an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). Here the value of Y ranges from 0 to 1 and it can represent by the following equation.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrece…(Eq: 2)

ln (odds) = ln(p/(1-p))**.....**(Eq: 3)

logit (p) = ln(p/ (1-p))
=b0+b1X1+b2X2+b3X3....+bkXk**.....**(Eq: 4)

Logistic regression doesn't require the linear relationship between dependent and independent variables. It can hold various types of relations because it applies a non-linear log transformation to the predicted odds ratio.

C. Polynomial regression

Polynomial regression is a work of art of linear regression in which the correlation among the independent variable x and the dependent variable y is modeled as an nth order polynomial and the power of independent variable is more than 1. The equation below represents a polynomial equation:

y=a+b*x^2**.....**(Eq: 4)

In this regression method, the best fit line is not a straight line. It is rather a curve that fits into the data points.

V. LITERATURE SURVEY

Imran Naseem et al, proposed a new linear regression based on nearest subspace classification, this algorithm have compared and evaluated with standard database protocols. This algorithm consists of three approaches (PCA, LDA, ICA) that transform into low dimensional face space into two categories(reconstructive and discriminative). The main study of this approach have changed the non-frontal image into frontal image this can be estimated and produces a greater result with help of benchmark techniques. DEF approaches produced two major issues: the non-face partitions are rejected and overall recognition are combined in face segments. This study have identified the contiguous issues and related them to LRC approach with the help of DEF algorithm. LRC approach produces high recognition accuracies without any preprocessing action for face recognition and produced reliably for real scenarios [1].

Bo Zhang et al, proposed a new dynamic clustering algorithm for inter-signal correlation and clustering data for linear correlation. By the proposed idea, they invented energy efficient sampling method using joint linear regression and compressive sensing. Wireless sensors networks

are collections of continuous data such as climate, habitat and infrastructure monitoring. The sensor nodes follow the routine to collect readings and transmit to the sink usually made up of battery-powered and communicate using a radio transceiver. The approach of energy efficient sampling method using joint linear regression and compressive sensing produced as node works in periodically sampling mode and remaining modes works in compressed sampling mode with a very low rate. The processing capabilities are balanced in low devices complexity with low energy consumption in communication and signal.[2]

Kyung-Bin et al, presented the forecasting errors for the holidays are higher than weekdays. The prediction accuracy can be improved with help of deterministic, stochastic, Artificial Neural Net (ANN) and network-fuzzy methods. The study of reducing the error is based on the concept of fuzzy regression analysis. The data of previous years are loaded and the coefficients of the model are initiated by solving mixed linear programming to produces good accuracy. A prediction model with a fuzzy module in series is held on the neural net module such as programmed, hooked on account temperature, weekdays and seasonal variation. This study produces two benefits: the capability of approximating any non-linear function and other model determination through a learning process. Tanaka's approach held on to fuzzy arithmetic operations where both input and output data are fuzzy numbers.[3]

Duy-Huy et al, proposed an improved logarithmic maximum a posteriori (log-MAP) used for LTE (long term evolution) turbo decoding. This study discussed the approach of polynomial regression function to about compute the logarithmic term in jacobian logarithmic function. The main study of this approach is to replace the correction function with another function and reduce computational complexity. Turbo decoding with additive white Gaussian noise channel will offer a maximum performance than the max-log-MAP algorithm and high than a log-MAP algorithm.[4]

Suat Ozdemir et al, proposed the Wireless sensor networks data are secure using polynomial regression. This research work proposed novel approach based on data aggregation protocol which used to sensor the notes and represented the data in polynomial functions. Data aggregation approach is based on coefficients and base stations which enables to extract approximate network data from the aggregation. privacy preservation to data aggregation process while reducing the data communication.[5]

Hongran Li et al, proposed a model-free predictive control for nonlinear systems based on polynomial regression. The model-free predictive control does not use any mathematical models, it shows a satisfactory control performance when large datasets are available around the reference trajectory. The study preferable to the previously maintained datasets.[6]

Jackson Isaac et al, applied an analytical query model for data categorization in DBMS. A data set with 'n+1' attributes and' tuples, that analyzing the data within DBMS using SQL Query and UDF models is much more efficient than the current trend of exporting the data outside of

DBMS and analyzing. Optimized query model performed better in all the cases whereas UDF performed better. When the size of data grew larger than the main memory, when the logistic model is stored within the DBMS, new data can be classified in constant time using the Query processor.[7]

Jun Li et al, proposed a new semi-supervised segmentation algorithm, suited to high-dimensional data, of which remotely sensed hyper spectral image data sets are an example. The algorithm consists of two main steps they are semi supervised learning of the posterior class distributions followed by segmentation, which infers an image of class labels from a posterior distribution built on the learned class distributions and on a Markov random field. This study consists of training samples (selected by means of an active-selection strategy based on the entropy of the samples) which are used to improve the estimation of the class distributions and adopting a spatial multilevel logistic prior and computing the MAP segmentation with the α- expansion graph-cut-based algorithm, the segmentation accuracy achieved by method in the analysis of simulated and real hyper spectral scenes collected by the AVIRIS improves significantly with respect to the classification results proposed by the same algorithm using only the learned class distributions in spectral space.[8]

Hasnat Khurshid et al, Segmentation and classification of multi temporal multi spectral SPOT 5 satellite images were presented using logistic regression and the registration process was conducted using commercial software and corresponding control points. The damage assessment and categorization was conducted within the BUA using MLR on a combination of different measures techniques. The proposed algorithm was evaluated using visual interpretation, statistical flood damage data produced by government agencies, and the damage reports of UNOSAT. The study was made with commercial and existing techniques of segmentation and damage assessment. The damage one produce was found consistent with ground facts and gives a lead for using SPOT 5 images for damage assessment.[9]

VI.FINDINGS

Prescriptive analytics seeks to determine the best solution or outcome among various choices, given the known parameters. Prescriptive analytics can also suggest decision options for how to take advantage of a future opportunity or mitigate a future risk, and illustrate the implications of each decision option [10]. In practice, prescriptive analytics can continually and automatically process new data to improve the accuracy of predictions and provide better decision options.

Prescriptive analytics consists of three types

1) Linear regression
2) Logistic regression
3) Polynomial regression

In linear regression, the single independent variable is used to predict the value of a dependent variable. LRC algorithm is compared and evaluated with standard database protocols. It consists of three approaches and it is divided into two categories. The three approaches are PCA(Principle Component Analysis), LDA(Linear

Discriminant analysis) and ICA (Independent Component Analysis). The two approaches are reconstructive and discriminative. Reconstructive approaches (such as PCA and ICA) are reported to be robust for the problem of contaminated pixels, whereas discriminative approaches (such as LDA) are known to yield better results in clean conditions. The LRC approach reveals a number of interesting outcomes than the Modular LRC approach for face identification. The LRC approach yields high recognition accuracies without requiring any preprocessing steps of face localization and/or normalization.

The fuzzy linear algorithm is based on Tanka's approach using Fuzzy linear regression where both input and output data are fuzzy numbers. In this regression, discussed forecasting for the holidays. Joint Linear Regression and Compressive Sensing (CS), discussed an energy efficient sampling method. Only one reference node works in periodically sampling mode and all other nodes in cluster work in compressed sampling mode with a very low sampling rate. After receiving data of all nodes, the sink node run prediction algorithm to roughly estimate the signal series of all nodes based on the signal series of reference node, then correct prediction error by using measurements of the nodes which work in compressed sampling mode.

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. This regression within DBMS, Optimized query performs is better in all cases,

Performance of UDF and R are almost similar when the data can be accommodated within the main memory and deteriorates when there is a large number of rows due to the use of summary tables for storing intermediate results. There is an exponential growth in the time taken by R when the data can no longer be stored in the main memory. When the logistic model is stored within the DBMS, new data can be classified in constant time using the Query processor.

In Semi-supervised Hyperspectral Image Segmentation, a comparison of the discussed method with other state-of-the-art classifiers in the considered(highly representative) hyperspectral scenes indicate that the discussed method is very competitive in terms of the (good) overall accuracies obtained and the (limited) number of training samples (both labeled and unlabeled) required to achieve such accuracies. In Segmentation and Classification in Remote Sensing Imagery, the resolution of SPOT 5 imagery is not sufficient for precise damage assessment and can only give a broad estimate of damage. The segmentation results were validated using Statistical measures like precision, recall, and dice coefficient on available ground truth. The results of change classification were compared and found consistent with the manual assessment report produced by UNO experts using Worldview 1 satellite imagery with sub-meter resolution.Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x. A log-map algorithm based on polynomial regression function, It can be easily

implemented in hardware involving shift registers, multiplications, comparators and addition operations. The simulation results show that the proposed algorithm outperforms the other Log-MAP-based algorithms, particularly are superior to the Max-Log-MAP algorithm with slightly increased complexity. Polynomial Regression Based Secure Data Aggregation for Wireless Sensor Networks, Data aggregation process is achieved using the polynomial coefficients and the base station is able to extract a good approximation of the network data from the aggregated data. The performance evaluation shows that the proposed scheme provides privacy preservation to data aggregation process while reducing the data communication.Model-Free Predictive Control for Nonlinear Systems in this regression, Model-free predictive control uses a set of short-length vectors that are selected from the input and output sequences of the controlled system. The vectors can be used to identify an autoregressive model, they are instead used to synthesize control input directly in order to obtain the desired output through locally weighted averaging.

VI. CONCULSION

The future of Data Mining, Big data lies in Predictive Analytics. This research mainly focuses on regression algorithms of Predictive Analytics domain. In this research various regression algorithms such as linear, logistic and polynomial regression technique are discussed, in which these algorithms helps to business analysts in order to build models to predict trends, make tradeoff decisions, and model the real world for decision making support. In this research, we discussed various application using this three regressions. So every regression predicts the various result. In a future, comparison of these algorithms to shows which is the best algorithm in predictive analysis would be done.

REFERNCES

[1] Imran Naseem, Roberto Togneri, Senior Member, IEEE, and Mohammed Bennamoun "Linear Regression for Face Recognition", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 11, November 2010.

[2] Bo Zhang, Yulin Liu, Jiwei He,and Zhaowu Zou "An Energy Efficient Sampling Method through joint Linear Regression and Compressive Sensing", 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP)June 9 – 11, 2013.

[3] Kyung-Bin Song, Member, IEEE, Young-Sik Baek, Dug Hun Hong, and Gilsoo Jang, Member, IEEE "Short-Term Load Forecasting for the Holidays UsingFuzzy Linear Regression Method", Ieee Transactions On Power Systems, Vol. 20, No. 1, February 2005.

[4] Duy-Huy Nguyen and Hang Nguyen," An improved Log-MAP algorithm based on polynomial regression function for LTE Turbo decoding" ,IEEE-IC-2015.

[5] Suat Ozdemir and Yang Xiao "Polynomial Regression Based Secure Data Aggregation for Wireless Sensor Networks", IEEE Globecom 2017.

[6] Hongran Liy and Shigeru Yamamoto "Polynomial Regression Based Model-Free Predictive Control for Nonlinear Systems" SICE

Annual Conference 2016 ,Tsukuba, Japan, September 20-23, 2016

[7] Jackson Isaac and Sandhya Harikumar, "Logistic Regression within DBMS". IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 32, No. 11, November 2019.

[8] Tomas Pranckevi□ius and Virginijus Marcinkevi□ius "Application of Logistic Regression withPart-Of-The-Speech Tagging For Multi-Class Text Classification"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210-227, Feb. 2009.

[9] Jun Li, José M. Bioucas-Dias, Member, IEEE, and Antonio Plaza, Senior Member, IEEE," Semisupervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression With Active Learning", IEEE Transactions On Geoscience And Remote Sensing, Vol. 48, No. 11, November 2010.

[10] Banumathi. S and Aloysius. A, "Big Data Prediction Using Evolutionary Techniques: A Survey", Journal of Emerging Technologies and Innovative Research, Vol. 3, Issue.9, Pg. 89-91, Sep 2016.

AUTHORS PROFILE

S. Banumathi doing her Doctoral Degree in Computer Science at Bharathidasan University, Tiruchirappalli, Tamilnadu, India. She has more than Seven years of teaching and research experience. She is currently working as Assistant Professor, Department of Computer Applications, Bishop Heber College, Tiruchirappalli, Tamilnadu, India. She had published several papers in international journal. She had presented various papers in national and international conferences. Presently she is doing her research on Big data analytics, Machine learning, Cloud computing, Wireless Adhoc Networks, IoT domains.

Dr. A. Aloysius is working as Assistant Professor in Department of Computer Science, St. Joseph's College, Trichy, Tamil Nadu, India. He has 17 years of experience in teaching and research. He has published many research articles in the National/ International conferences and journals. He has also presented research articles in the International Conferences on Computational Intelligence and Cognitive Informatics in Indonesia. He has acted as a chair person for many national and international conferences. His current area of research is Cognitive Aspects in Software Design, Big Data, and Cloud Computing.