International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 09 | Sep -2017 www.irjet.net p-ISSN: 2395-0072

Predictive Modelling Analytics through Data Mining

Lakshay Swani[1], Prakita Tyagi[2]

[1] *Software Engineer, Globallogic Pvt. Ltd, Sector-144, Noida, UP, INDIA*
[2] *Software Engineer, Infosys, Bengaluru, INDIA*

Abstract – With this paper, we try to explore Predictive Analytics through a combination of various Data Mining techniques over Big Data. The survey provides an indication of acceleration in the area of Predictive Analytics for the enhancement of businesses and researchers by applying business intelligence for providing forecasting ability to proceed through the development of business. Also, the paper portrays the incorporation of the characteristics of Big Data as the base of Data Mining through the application of Apache Hadoop in achieving the above. Predictive analytics uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future. It is considered as the next frontier for innovation.

KEYWORDS: Predictive Analytics, Data Mining, Big Data, Hadoop, Machine Learning

1. INTRODUCTION

With the ever increasing online content generated at an accelerated rate, petabytes of data is produced each day. The explosion in the user generated data from the social media and businesses and organizations, there has always been a search for effective management and storage of the enormous data created. Examples of the above fact includes an approximation of over 7 terabytes of data generated each day by twitter, over 25 petabytes of data being handled by Google and over 500 terabytes of data being produced each day at Facebook [6]. Such extensive data might mean nothing when unorganized but when organized into recognizable structure, it could be extremely useful in analysis and future prediction of growth strategies for an organization [28]. This gave rise to the prediction domain through data mining [29]. What is required to obtain the information from the raw data, is highly computational systems that could process the raw data to convert it to meaningful data that could be recognized and analyzed. As the enormous data that is obtained online is unstructured and unorganized, the need arise to store, structure and process this available raw data in a cost effective and efficient way. Traditionally, the IT industry dint had the capabilities to handle such large sum of data. But with drastic excellence shown in the IT industry, various ways and tools have been devised for effective and efficient storage and processing of the data available [30][31]. Apache Hadoop is one such tool that has been emerging as one of the most successful in handing and managing big unorganized data [4].

1.1 Data Mining

Data mining is the process of discovering patterns and anomalies form large sets of data through the application of statistics, machine learning and database systems [17]. The objective of data mining is to extract the useful data from data set and process them to obtain information that has an understandable structure. It includes provision for automated and semi-automated analysis of large quantities of data set to extract recognizable patterns, dependencies between the data, groups or clusters of records and the relation between these records as well as the anomalies occurring within the system. Data mining could be applied to identify multiple groups and clusters of data and forward them for further analysis through machine learning or predictive analytics.

Data mining commonly involves the following 6 tasks

1. Anomaly Detection - In data mining, anomaly detection is the detection of events, items or observations that do not conform to a pattern that is expected or to an established normal behavior
2. Association Rule Learning – This involves discovering and describing the relationship between the various variables present within the data set. It is intended to identify and discover extensive rules within the data in the databases. This information could be useful in undertaking decisions before forth going any dilemma that could be an integral part of future perspective of the business of the organization [16].
3. Clustering – Clustering involves identifying similarities between the various objects of the data set and abstracting them into classes of similar objects with similar properties. It involves discovering structures and groups in the data have similar properties or are similar to each other in distinction, without using the available structures present in the data set.
4. Classification – It includes the process of creating generic or generalized known structures and properties and apply them to the data that has been clustered.
5. Regression – It involves development of a function that is responsible for predicting various properties and factors using regression techniques.

6. Summarization – This task involves and revolves around development of a compact representation of data set, including generation of report and visualization.

1.2 Need for Data Analytics

In all areas of business, a well-structured actionable data has endless benefits. With the high volume, variety and velocity of data that is being input to an organization, analysing the data could prove to be extremely useful for the aspect of the business. We provide the following key reasons highlighting the need of Big Data and strategic analysis within an organization referring to how it could help with the business growth.

1. Smarter Organizations – Through a well thought out analytics strategy, an organization work more efficiently and smartly in achieving its goals. For example, through thorough analysis of data patterns within a police department, it could be susceptible for them to identify the crime scenes and hotspots and in turn help the department to work efficiently in solving and preventing crimes from happening around the globe. In medical industry, proper analysis over the disease pattern over an area or group of people could help the doctors to effectively devise and predict the possibilities of a disease. In other case, weather forecasting industry has the analytics over the weather patterns as its base that helps the organization with proper and accurate weather predictions. Thus we see, ranging from criminal justice to real estate to health care to weather forecasting, Big Data analytics in being leveraged to provide effective and efficient outcomes.

2. Behavioural Marketing – The base of a successful business or organization is its reach to the targeted consumers. With effective and efficient marketing strategies, a business could bloom to its highest levels. With the marketing for the business reaching to its desired public, proper strategic analysis needs to be done so as to make the business reach to the target audience. This is where the behavioural marketing comes into play. Consumers are targeted on the basis of the websites they consume or searches for a commodity. This data is analysed for patterns to predict the audience to be notified of the organization. The data is collectively organized and analysed to gather the targeted audience and the marketing is done in the form of advertisements shown particularly to the category of audience for which the business being marketed could be useful. It provides the marketer with the ability to get in touch with the desire segment of the society hence leading to a well thought and effective marketing in turn making the business bloom.

3. Business Future Perspective – Big Data analysis can inevitably predict the future of the business keeping in mind the current scenarios. It helps the business in taking effective measurements that could possibly lead to the better future of the organization. With proper analysis, decision making abilities are provided to the organization and could lead to a perspective growth of its business. With the market trend changing rapidly, the trend of the changes that would be of impact to the business can be predicted efficiently and measures could be taken in handling those impacts. With strategic analysis, certain crucial decision points could be handled within an organization so as to achieve a foreseen goal of the organization.

2. Data Mining Techniques

1. **Association –** Association is one of the most familiar and most commonly known technique of data mining. In association, the similarities between different objects is identified for recurring patterns and a correlation is established between the objects. The objects are generally of the same type so as to define the correlation among them to identify the patterns. This could be exhibited through an example of a customer who always bought chocolates along with milk. The milk in this case gets associated with chocolates and thus a pattern is emerged that provides information about the sale of milk and chocolates and it suggests that the next time the person buys the chocolate, he'll be buying milk as well.
2. **Classification –** Classification, as the name suggests is the process of identifying an object by relating with it, multiple attributes that describe that object thoroughly. Classification is used to build up the reference or idea of an object by describing it using multiple attributes to define its particular class. Let us try to understand the same though an example of classification of cars. Given a car as an object, we can identify or classify it through multiple attributes such as shape, number of seats, transition type etc. Through proper comparison and analysis of the categories or attributes, one can identify the object to be classified into a similar kind of attribute.
3. **Clustering –** Once the attributes are applied to an object, through proper and considerate examination of the attributes or classes, one can identify it into groups of similar kind of objects. Such group of similar objects is referred to as a cluster. A cluster uses a single or a group of attributes or classes defined of an object as its base for segregation and

combine a set of results having correlating group of objects. Clustering works bi-directionally. It is useful in identifying a group of objects having a set of similar attributes or classes. Also it is useful in identifying the identification criteria for an existing cluster of objects.

4. **Prediction –** Prediction is referred to as a wide area of study that ranges from predicting about the failure of a machinery to identification of frauds and intrusions to even predicting the future aspect of the organization or business. When combined with the techniques of data mining, prediction involves various tasks such as analysis and creation of trends, classification, clustering, pattern discovery and relation. Through thorough examination and analysis over the past trends or events, one can make effective measurement regarding the future occurrence of that event.
5. **Sequential Patterns -** Sequential patterns are defined over a long period of time wherein trends and similar activities or events are identified to be occurring on a regular basis [23]. It is considered a very useful technique to identify the trends and similar events. Considering an example of a customer at a grocery store who buys a collection of objects regularly over the year or so. Through analytics over the historical data, patterns of grocery bought over the year can be used to provide the sequential pattern of what is to be added to the grocery list.
6. **Decision trees –** Decision trees are basically related to the above defined techniques. They can be either used for providing the criteria for selection or to provide support to the selection and use of specific data within the overall structure. A decision tree is started and created through a question that has more than a single result or option to be selected [14]. Each of the result in turn leads to another question that it can be categorized into a result set of more than one option. These subsequent questions leads to the categorization of the data set so as to facilitate the prediction based on the result set.
7. **Combinations –** In real world applications, several techniques are applied in combination. Exclusive use of a single technique is not valid. Clustering and combination are similar techniques and are used combined so as to achieve the desired effective result. Clustering can be used to identify the nearest neighbors which can be useful in refinement of the classification of the data set in use. In the same way, decision trees can act as base for the techniques of sequential patterns and is used to identify and build classification which when done over a longer period of time can be used to identify the patterns of similarity between the data set and could help in prediction.
8. **Long Term Processing –** Data analytics and predictive analytics is purely based on the data and information processed over a period of time. There is a need to record the data for a long term and then process the data for the patterns, classification, categorization and prediction. For example, for predictive learning and sequential patterns, there is a need to store and process the historical data and instances of information for building a pattern. As the time passes, new data and information is identified and processed along with the historical data and information and the analytics is applied on the whole set of data so as to cope with the additional data.

3. Predictive Analytics

Predictive analytics comprises of varied statistical trends and techniques ranging from machine learning and predictive modelling to data mining to efficiently analyze the historical data and information so as to process them to create predictions about the unknown future events [1][2]. As per the business aspect of predictive analytics, predictive analytics help in exploiting the patterns found in the historical business data to identify the risks and opportunities [3]. It captures the relationships between various factors to provide the assessment of risk or a potential threat and help guide the business through important decision making steps. Predictive analytics is sometimes described in reference to predictive modelling and forecasting. Predictive analytics is confined to the following three model that outlines the techniques for forecasting [26].

1. Predictive Models – Predictive models are the models that define the relationship between the various attributes or features of that unit. This model is used to assess the similarities between a groups of units providing assurance of the presence of similar attributes being exhibited by a group of similar units.
2. Descriptive Models – Descriptive models are the models that identify and quantify the relationships between the various attributes or features of the unit which is then used to classify them into groups. It is different from the predictive model in the ability to compare and predict on the basis of relationship between multiple behaviors of the units rather than a single behavior as is done in the predictive models.
3. Decision Models – Decision models are the models that identify and describe the relationship among all of the varied data elements present that includes the known data set upon which the model is to be defined, the decision structure that is defined for classification and categorization of the known data set as well as the forecasted or predicted result set on the application of decision tree on the known

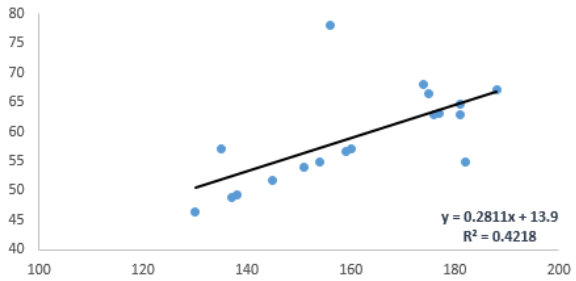
data set so as to identify and predict the results of the decisions based on multiple attributes or features of the data set.

4. Predictive Analytics Techniques

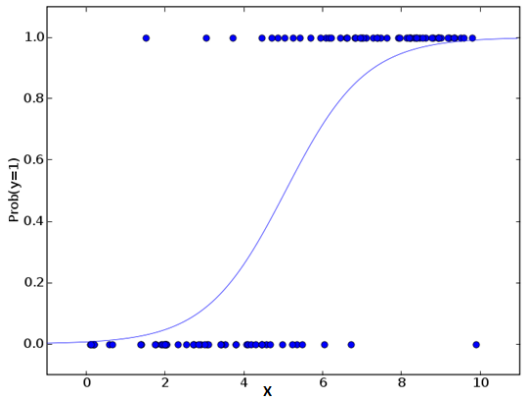
The techniques or approaches that can be used to conduct predictive analytics on a data set can be broadly defined and categorized as follows.

1. **Regression Analytics**
Regression techniques are focused on establishment of mathematical equations so as to model, represent and process the information from the available data set [25]. Some of the regression techniques being in use are described as follows.

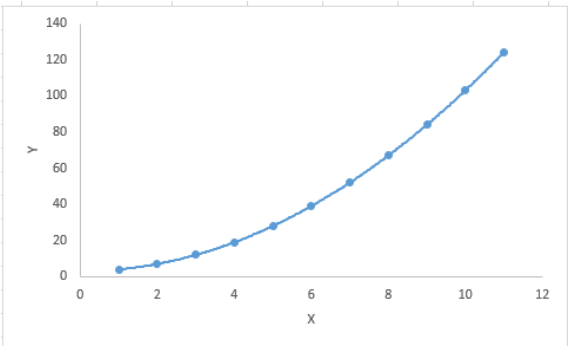
   a. **Linear Regression Model –** This technique establishes a linear relationship between the dependent variable y and multiple independent variables x. It is represented through the linear equation **y=a+bx+c**



   b. **Logistic Regression –** This technique is applied so as to find the probability regarding the success or failure of an event. This technique comes to use when the value of the dependent variable is binary.



   c. **Polynomial Regression –** In this technique, the prediction line is not a straight or linear one, but is a curve that fits the points of the data set being predicted upon.



   d. **Stepwise Regression –** This technique comes into play when there is a presence of multiple independent factors or variables. The best fit is predicted through stepwise incremental addition or removal of predictor variables as required for each of the step. This technique has the aim of achieving the maximum prediction power with the use of minimal number of predictor variables.

   e. **Ridge Regression –** Ridge regression technique is used where there is a multi-collinearity that is the data set has multiple independent variables with high extent or correlation. The ridge regression technique can be represented through the equation
   $y = a + b_1x_1 + b_2x_2 + b_3x_3 + ...+ e$

   f. **Lasso Regression –** Lasso regression is highly similar to ridge regression technique with less variance coefficients and high accuracy of the linear regression models. In this technique, variables having high correlation, only of the predictor variables is picked while all others are shirked to zero.

   g. **Elastic net Regression –** This technique is a combination of Ridge Regression technique and Lasso Regression technique. It enhances the accuracy of the best fit result and provides the advantage of no limitations over the number of variables selected and has the ability to suffer and withhold double shrinkage.

2. **Machine Learning Analytics**
Machine learning is a branch of AI (Artificial Intelligence) that was devised to provide the ability to computers to learn. These days, it is being used in various statistical models and methods for prediction of risks and opportunities and is found applicable in various fields such as banking fraud detection, medical diagnosis, natural language

processing and analysis over the stock market. Some of the methods commonly used for predictive analytics through machine learning are defined as follows.

a. **Neural Networks –** These are nonlinear modelling techniques wherein they learn the relationship between the inputs and the outputs through training. There are 3 types of trainings used by neural networks: supervised, unsupervised and reinforcement training. This technique can be applied for prediction, control and classification in various fields.
b. **Multilayer perceptron –** This technique consists of an output and an input layer with multiple hidden layers of nonlinear weights and is determined and defined through the weight factors by adjusting the weight of the network. The adjustment of the weights is done through a process called training the nets that contains the learning rules.
c. **Radial basis functions –** Radial basis function technique is built on the criteria of distance of data set with reference to the center. These functions are basically used for interpolation of data as well as smoothening of data.
d. **Support vector machines –** SVMs are designed and defined to detect and identify the complex patterns and sequences within the data set through clustering and classification of the data. They are also referred to as the learning machines.
e. **Naïve Bayes –** Naïve Bayes is deployed for the execution of classification of data through the application of Bayes Conditional Probability [8]. It is basically implemented and applied when the number of predictors is very high.
f. **k-nearest neighbours –** This techniques involves pattern recognition techniques of statistical prediction. It consists of a training set with both positive and negative values.
g. **Geospatial Predictive Modelling –** This modelling technique involves presence of occurrence of events over a spatial area with influence of special environmental factors. The occurrence of events is defined to be neither uniform not random but special in nature.

5. Applications of Predictive Analysis through Data Mining

1. **Customer Relationship Management (CRM) –** Analytical CRM is one of the most frequently used application of predictive analytics these days. The predictive analytics under this area is applied to the customer data to pursue and attain the CRM objectives defined for an organization. CRM makes use of these analysis in applications for increasing the sales targets, marketing and campaigns [2][5]. This not only impacts the business growth, but also makes the business customer eccentric through widening the base for customer satisfaction [7].
2. **Child Protection –** Child abuse is a serious offence and child protection is most sought after within any country [18]. Several child welfare organizations have applied the predictive analytics to flag high risk cases of child abuse [11]. The predictive models help in identifying from medical records, the cases that could fall under the child abuse criteria. This approach is termed as "innovative" by the Commission to Eliminate Child Abuse and Neglect Fatalities (CECANF) [19]. Using the predictive analytics, the child abuse related felonies have been identified at the earlier stage preventing from much further harm [20].
3. **Clinical decision support systems –** As defined, Clinical decision support (CDS) provides clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care [10][21]. It encompasses a variety of tools and interventions such as computerized alerts and reminders, clinical guidelines, order sets, patient data reports and dashboards, documentation templates, diagnostic support, and clinical workflow tools. [22] Experts have involved the predictive analytics to model the clinical data of patients so as determine the extent to which a patient might be exposed to a disease and predict the risk of development of certain conditions such as heart disease, asthma or diabetes. These approaches have been devised so as to predict both the state and level of the disease as well as the diagnosis and disease progression forecasting [12][13].
4. **Collection Analytics –** Many portfolios these days has a set of customer who doesn't make their payment within the defined time and the companies put up a lot of financial expenditure on collection of those payments [27]. Thus the companies have started applying the predictive analytics over their customers for effective analysis of the spending, usage and behavior of the customer who are unable

to make the payment and allocate the most effective legal agencies and strategies for each customer, thus increasing recovery significantly with lesser financial expenditure.

5. **Fraud Detection –** Fraud is one of the biggest challenges faced by businesses around the globe and can be of various kinds such as fraudulent online transactions, invalid credits [15], identity thefts and multiple false insurance claims. Predictive modelling can be applied to this area so as to model the data of the organization and detect such fraudulent activities [24]. These models have the capabilities to identify and predict the customers engaged in such activities. Many revenue systems too take in this consideration to mine out the non-tax payers and identify tax fraud [24].
6. **Project Risk Management –** Each company employs a risk management technique so as to increase their revenue. These risk management techniques involves the use of predictive analytics to predict the cost and benefit of a project within an organization and also helps organizing the work management so as to maximize the profit statement. These approaches can be applied ranging from projects to markets so as to maximize the return from the investment.

6. CONCLUSIONS

Predictive analytics is the future of Data Mining. This survey provided with the trends, techniques and applications of Predictive analytics through the application of data mining. Data Mining leading to predictive analytics is becoming key to every organization as it can be applied under various circumstances so as to highlight growth of the organization. Predictive Analytics aid not only in expansion of the business, but also prevents the degradation through analysis of the fraudulent activities.

REFERENCES

[1] Nyce, Charles (2007), Predictive Analytics White Paper(PDF), American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1

[2] Eckerson, Wayne (May 10, 2007), Extending the Value of Your Data Warehousing Investment, The Data Warehouse Institute

[3] Coker, Frank (2014). Pulse: Understanding the Vital Signs of Your Business (1st ed.). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42,more. ISBN 978-0-9893086-0-1.

[4] Hadoop over RDBMS: http://www.computerworlduk-.com/in-depth/applications/3329092/hadoop

[5] Fletcher, Heather (March 2, 2011), "The 7 Best Uses for Predictive Analytics in Multichannel Marketing", Target Marketing

[6] Bringing the Power of SAS to Hadoop .Combine SAS World-Class Analytic Strength with Hadoop's Low-Cost, High-Performance Data Storage and Processing to Get Better Answers, Faster.

[7] Barkin, Eric (May 2011), "CRM + Predictive Analytics: Why It All Adds Up", Destination CRM

[8] Mrutyunjaya Panda, Manas Ranjan Patra. A Comparative Study of Data Mining Algorithms for Intrusion Detection.

[9] McDonald, Michèle (September 2, 2010), "New Technology Taps 'Predictive Analytics' to Target Travel Recommendations", Travel Market Report

[10] Moreira-Matias, Luís; Gama, João; Ferreira, Michel; Mendes-Moreira, João; Damas, Luis (2016-02-01). "Time-evolving O-D matrix estimation using high-speed GPS data streams". Expert Systems with Applications. 44: 275–288. doi:10.1016/j.eswa.2015.08.048.

[11] Stevenson, Erin (December 16, 2011), "Tech Beat: Can you pronounce health care predictive analytics?", Times-Standard

[12] Lindert, Bryan (October 2014). "Eckerd Rapid Safety Feedback Bringing Business Intelligence to Child Welfare" (PDF). Policy & Practice. Retrieved March 3, 2016.

[13] "Florida Leverages Predictive Analytics to Prevent Child Fatalities -- Other States Follow". The Huffington Post. Retrieved 2016-03-25.

[14] Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, Wei-Yang Lin (2009). Intrusion detection by Machine Learning : A Review

[15] Finlay, Steven (2014). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st ed.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 1137379278.

[16] Siegel, Eric (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (1st ed.). Wiley. ISBN 978-1-1183-5685-2.

[17] Ian H. Witten and Eibe Frank. Data Mining : Practical Machine Learning Tools and Techniques.

[18] "New Strategies Long Overdue on Measuring Child Welfare Risk - The Chronicle of Social Change". The Chronicle of Social Change. Retrieved 2016-04-04.

[19] "Eckerd Rapid Safety Feedback® Highlighted in National Report of Commission to Eliminate Child Abuse and Neglect Fatalities". Eckerd Kids. Retrieved 2016-04-04.

[20] "A National Strategy to Eliminate Child Abuse and Neglect Fatalities" (PDF). Commission to Eliminate Child Abuse and Neglect Fatalities. (2016). Retrieved April 4, 2016.

[21] "A Roadmap for National Action on Clinical Decision Support". JAMIA. Retrieved 2016-08-10.

[22] "Predictive Big Data Analytics: A Study of Parkinson's Disease using Large, Complex, Heterogeneous, Incongruent, Multi-source and Incomplete Observations". PLoS ONE. 11: e0157077. doi:10.1371/journal.pone.0157077.

[23] http://www.articlesbase.com/strategic-planning-articles/predictiveanalytics-1704860.html

[24] Schiff, Mike (March 6, 2012), BI Experts: Why Predictive Analytics Will Continue to Grow, The Data Warehouse Institute

[25] www.cs.uiuc.edu/~hanj, Jiawei Han and Micheline Kamber,2006.

[26] Wayne W. Eckerson, "Predictive Analytics : Extending the Value of YourData Warehousing Investment", www.tdwi.org, 2006.

[27] Dhar, Vasant; Chou, Dashin; Provost Foster (October 2000). "Discovering Interesting Patterns in Investment Decision Making with GLOWER – A Genetic Learning Algorithm Overlaid With Entropy Reduction". Data Mining and Knowledge Discovery. 4(4).

[28] http://www.hcltech.com/sites/default/files/key_to_monetizing_big_data_via_predictive_analytics.pdf

[29] "Predictive Analytics on Evolving Data Streams" (PDF).

[30] Ben-Gal I. Dana A.; Shkolnik N. and Singer (2014). "Efficient Construction of Decision Trees by the Dual Information Distance Method" (PDF). Quality Technology & Quantitative Management (QTQM), 11(1), 133-147.

[31] Ben-Gal I.; Shavitt Y.; Weinsberg E.; Weinsberg U. (2014). "Peer-to-peer information retrieval using shared-content clustering"(PDF). Knowl Inf Syst. 39: 383–408. doi:10.1007/s10115-013-0619-9.