International Journal of Research in Engineering, Science and Management
Volume-2, Issue-10, October-2019
www.ijresm.com | ISSN (Online): 2581-5792

Predictive and Descriptive Analytics using Big Data

Suvetha Suresh[1], V. Sanjana[2]
[1,2]*Student, Dept. of Computer Science and Engg., SRM Institute of Science and Technology, Chennai, India*

***Abstract*: There is a tremendous development in business processes which requires complex supporting system. Using advanced analytics and booming technological methods such as business intelligence systems, business activity monitoring, predictive analytics, behavioral pattern recognition, and "type simulations" can help business users continuously improve their processes. However, visualizing and interpreting the data to support the business decision making is still a challenging task and there is scope for advancement. The aim of the research is to find out a suitable technique from the field of machine learning and big data for internal cost prediction in consulting businesses and to provide interactive decision support. Improving the business decision is mostly based on how we analyze and make use of the data. Predictive analytics and data analytics play a major role in the era of big data. This paper examines the changes in customer behavior by interpreting and visualizing the data. On the other hand, predict an internal cost regarding the business using big data and machine learning.**

***Keywords*: Predictive analytics, Descriptive analytics, Apache spark, Knime, K nearest neighbor, Linear Discriminant Analysis, Logistic Regression, Gradient Design Neural Network, Optimal price**

1. Introduction

The aim of the research is to find out a suitable technique from the field of machine learning and big data for internal cost prediction in consulting businesses and to provide interactive decision support. Improving the business decision is mostly based on how we analyse and make use of the data. Predictive analytics and data analytics play a major role in the era of big data. This research examines the changes in customer behavior by interpreting and visualizing the data. On the other hand, predict an internal cost regarding the business using big data and machine learning. Data once obtained from a source can be a definitive resource to improves business model. Here the important scenarios taken into account is the pricing of products. Pricing of products are centralized on various categories. To find an optimal prince using the raw form of data obtained and converting them into considerable dataset is huge task. This can be obtained by using neural networks where gradient descent plays an important role. Major data nodes are shifted using big data and machine learning.

2. Related Works

This paper deals with examining the customer behaviour and interpret the internal cost regarding the business using big data and machine learning. Big data analytics provides these businesses the power to gather client knowledge, apply analytics and forthwith determine potential issues before it's too late. Machine Learning solutions to analyze, measure, and refine business processes. A business process model which can represent sequences of activities in order to service customers is essential to achieve business goals .By application of both big data and machine learning, we can automate analytical model building using the data collected which is in turn used to predict a customer's behavior and help us to interpret the internal cost. The systems already available examines the changes in consumer behavior and opinions due to the transition from a public to a commercial broadcaster in the context of broadcasting international media events. Another one scenarios is where a novel modeling framework which consists of a conceptual modeling language, a process and a tool for effective business processes reengineering using big data analytics and a goal oriented approach. These use process mining and process analytics techniques which utilizes process event log data and modeled process and real process confirm. Using this technologies, analyzers can find bottlenecks or structural problems of business processes and help enhance business processes. The most important difference is that they do not consider business goals and alternatives, i.e., goal-orientation. Thus, it is hard to diagnose how business processes aligned with business goals. Moreover, they also deal with large volume of data, but they do not use a distributed parallel processing. For analyzing problems, Fishbone diagram , Fault Tree Analysis, Pareto, Abuse Frames are used to analyze problems, but they do not support both certain and uncertain relationships between elements, and goal-orientation approaches. Additionally, although anti-goal and PIG (Problem Interdependency Graph) considers problem in a goal-orientation approach, they don't consider business processes. But in our model we use Gaussian distribution which helps in forming a bell shaped curve from which the change in aspects of raw data can be analysed over a period of time. The aim of our framework is to help reengineer business process by finding business problems in current (as-is) business models which are against business goals and transforming the as-is toward the next (to-be) business process models which are aligned with business goals, both of which are supported by the evidences

from big data.

3. System Architecture

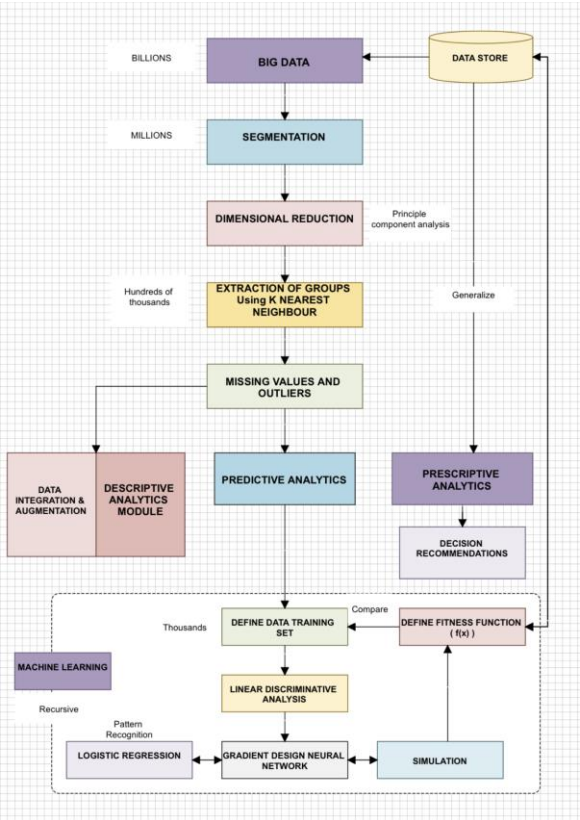

Fig. 1. System architecture

To predict optimal price for a product the system must deal with many variables and use cases because prediction of price is typically a complex structure. Let us consider a customer who buys a product from the supplier at supplier price. Revenue per product, gross profit, product details, product price and customer details are collected and stored in the data store (Big data source). This source is stored in cloud. This cloud storage is accessed by data bus for data extraction in database management system. Using big data tools like Apache spark and Knime, we process the available data with built-in module for faster computation. All these data are visualized and stored in cloud again.

Usually to predict optimal price for a product, we need millions and billions of data so that it can be accurate enough. All these billions of data in the big data are reduced to millions by segmentation approach. These millions of data are again dynamically reduced by principal component analysis. Hundreds of thousands of data are extracted and grouped using K Nearest Neighbor. Missing values and outliers are found. Basically, in the initial part we reduce the billions of data to use the necessary data in order to generate a module to predict the pricing of the product.

In descriptive analytics module, the data is augmented and integrated. Prescriptive analytics module is mainly for decision recommendation. In predictive analytics data training set is defined. Linear discriminant analysis algorithm consists of statistical properties of your data, calculated for each class. For a single input variable (x) the mean and the variance of the variable for each class is calculated. For multiple variables, the same properties calculated over the multivariate Gaussian, mainly the means and the covariance matrix. All These statistical properties are estimated from your data and plug into the Linear Discriminant Analysis equation to make predictions. All the given parameters are estimated in Linear Discriminant Analysis. Gradient Descent requires gradient vector from which the gradient descent training direction is computed and a suitable training rate is found. Pattern recognition is done by Logistic Regression. This gradient descent training rate is simulated and Fitness Function (f(x)) is defined. The defined Fitness Function (f(x)) is compared with the data training set. This machine learning process is done recursively and finally an optimal price for the product is predicted.

4. Proposed System

Changes happening in this world are reflected through data. The more things change, the more the changes are captured and recorded in the form of data. We can consider weather as example. Weather is not an entity that remains static, instead it differs from region to region and time to time. Therefore, it generates huge amount of data which can be stored and later used to analyze it for various purposes. Thus, Big data refers to a process that is used when data is mined using traditional methods and handling techniques which does not help in portraying the actual meaning of underlying data. Data that does not have structure or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data, which makes use of massive parallelism on readily-available hardware. The characteristics of big data can be based on the volume, variety, velocity, variability of collected data.

A. Apache Spark

Spark is a framework which provides a number of platforms, systems and standards which are interconnected with each other for various Big Data projects. Spark is open-source platform under the wing of the Apache Software Foundation. Spark has proven itself to be highly suitable for various Machine Learning applications. Spark does not have its own file system but it can be integrated with many file systems including Hadoop's HDFS, MongoDB and Amazon's S3 system.

Another important element of this framework is Spark Streaming, which helps in developing applications which perform analytics on streaming, real-time data such as automatically analyzing video or social media data, in real-time.

B. Knime

Knime is an analytics platform which is an open source

platforms used in data science to generalize and automate data. Knime has thousands of nodes which contains enough storage to store data in the node repository which allows you to drag and drop the nodes into the Knime workbench and help in analyzing them. It uses modular data pipelining concept. A collection of interrelated nodes creates a workflow which can be executed in a local manner as well as can be executed it in the Knime web portal after deploying the workflow into the Knime server. Knime helps to create the Guided Analytics process as a workflow by helping to automate the process. It provides graphical user interface which provides assembly of notes for data preprocessing.

C. K-Nearest Neighbor

The process of learning begins with observations or data, such as examples, which we get from real world experience or from various other sources and use them to predict the upcoming possibilities. Thus machine learning helps in automating machines such as computer and other devices by making them learn everything by themselves and use if for solving problems. They use a lot of algorithms in order to provide solutions using machine learning.

K-Nearest Neighbors is one of the most basic and important classification algorithms in Machine Learning. It comes under supervised learning domain. Its major and widely used applications are in pattern recognition, data mining and intrusion detection. It is non parametric in nature which means, it does not make any underlying assumptions about how the data is distributed (where as in GMM algorithm it assumes a Gaussian distribution of the given data). Thus no prior assumptions are made from the data obtained. With the obtained data (also called training data), which classifies coordinates into groups identified by an attribute.

D. Linear Discriminant Analysis

Linear discriminant analysis is also called as normal discriminant analysis. It is a method which is used in various statistics problems, to recognize various patterns and in machine learning. Its main job is to find a linear combination of features that is used to characterize them or separate them into two or more classes of objects. The result produces by this classification is used as a linear classifier or as a dimensionality reduction which can be used for further classification. By estimating the probability of inputs LDA makes a prediction. The output class is decided by checking which class gets the highest probability. The theorem used to estimate these probabilities is Bayes theorem. Let's consider the output class to be (k) and input class to be (x), using this we can calculate the probability of each class to be:

P(Y=x|X=x) = (PIk * fk(x)) / sum(PIl * fl(x))

Where PIk refers to the base probability of each class (k) observed in your training data (e.g. 0.5 for a 50-50 split in a two class problem). In Bayes' Theorem this is called the prior probability.

PIk = nk/n

The f(x) above is the estimated probability of x belonging to the class. A Gaussian distribution function is used for f(x). Plugging the Gaussian into the above equation and simplifying we end up with the equation below. This is called a discriminate function and the class is calculated as having the largest value will be the output classification (y):

Dk(x) = x * (muk/siga^2) – (muk^2/(2*sigma^2)) + ln(PIk)

Dk(x) is the discriminate function for class k given input x, the muk, sigma^2 and PIk are all estimated from your data.

E. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are more than one independent variables that are used to determine an outcome. The outcome is measured with a dichotomous variable which means there are only two possible outcomes.

In logistic regression, can contain only dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Coefficients are generated by logistic regression and also its standard errors and significance levels of a formula to predict a *logit transformation* of the probability:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is called the probability of presence of the characteristic of interest. The logit transformation is also known by the term logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Rather than choosing parameters that helps in the minimization of the sum of squared error, estimation in logistic regression chooses parameters that maximize the availability of observing the sample values.

F. Gradient design neural network

It is commonly used in the training of deep neural networks. In general terms, back propagation is commonly used by the gradient descent optimization algorithm to adjust the weight of neurons by calculating the gradient of the loss function.

Backpropagation is also known by other terms such as "the backward propagation of errors", since an error is computed only at the output and distributed backwards throughout the all the network's layers. Changes in all weights are measured by a

gradient in error. Gradient can also be considered as the slope of a function. The higher the gradient, the steeper the slope and the faster a model can learn. But what if the slope is zero, the model stops learning. Mathematically, a gradient is defined as a partial derivative with respect to its inputs which are obtained. Gradient Descent can be thought as a minimization algorithm that minimizes a given function. There are three types of Gradient Descent, which differ in the amount of data they use.

5. Aspects of the system

The system architecture is divided into two main module, one is the big data part and the other one is machine learning part.

A. Big data approach

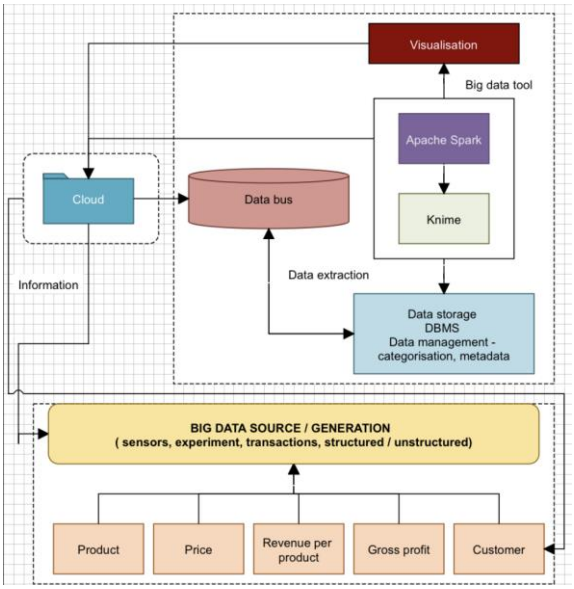

Fig. 2. Big data approach

In big data part, all the required details like Revenue per product, gross profit, product details, Product price and customer details are found in the big data source. To predict optimal price for the product we need billions and millions of data only then we can predict the price with higher accuracy rate. These sources are stored in cloud. Data bus has full authority to access the cloud for data extraction process in database management system. Big data tools like Apache spark and Knime are used to process the stored data with built-in module for faster computation. All these billions data are visualized and reduced to required amount of data to generate models.

B. Machine learning approach

In Linear Discriminant Analysis statistical property for the given data is calculated for each class. For a single input variable (x) the mean and the variance of the variable for each class is calculated. For multiple variables, the same properties calculated over the multivariate Gaussian, mainly the means and the covariance matrix. All These statistical properties are estimated from your data and plug into the Linear Discriminant Analysis equation to make predictions. All the given parameters are estimated in Linear Discriminant Analysis. Gradient Descent requires gradient vector from which the gradient descent training direction is computed and a suitable training rate is found. Pattern recognition is done by Logistic Regression. This gradient descent training rate is simulated and Fitness Function (f(x)) is defined. The defined Fitness Function (f(x)) is compared with the data training set. This machine learning process is done recursively and finally combining both the approach an optimal price for the product is predicted with higher accuracy rate provided that there are necessary required amount of data.

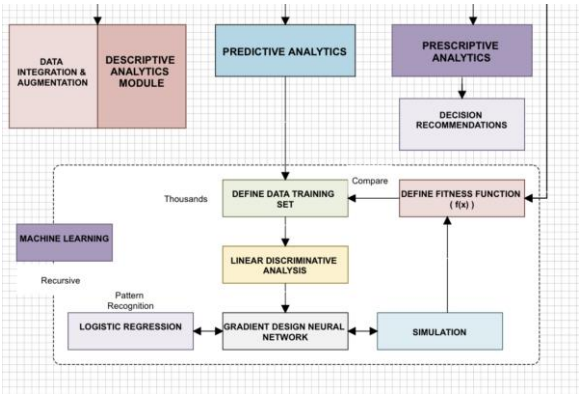

Fig. 3. Machine learning approach

6. Beneficiary of the system

- With the help of machine learning and big data, we can easily predict a products best price and demand also.
- High potential values that are difficult to process and analyze in reasonable time are analyzed easily in a statistical way.
- Valuable information from the chaos realty can be analyzed.
- Firm is benefitted by maximizing the profit and the quantity sold of the product.
- Constructs a mathematical optimization based on the given predictive formulas.
- The optimal price to maximize profit/revenue on the basis of predictive formulas produced by machine learning.
- Reliability cost investment is based on the forecasts made.
- Evaluating the relationship between the customer and the product bought by them, helps to generate a standard economic model that assumes the provider to maximize the product.
- Unexploitable ability to maximize the profit of the product.
- Maximum willingness to pay is reflected by the demand curve which optimal price is directly proportional to the demand.

7. Mathematical Modelling

Gaussian distribution which is also known as normal

distribution is a bell-shaped curve, and it is assumed that during any measurement values will follow a normal distribution with an equal number of measurements both above and below the mean value. The mean is the calculated average of all values, the median is the value at the center point also called as the mid-point of the distribution, while the mode is the value that was observed most frequently during the measurement. If a distribution is normal, then the values of the mean, median, and mode are the same. However, the value of the mean, median, and mode may be different if the distribution is skewed. Other characteristics of Gaussian distributions are as follows:

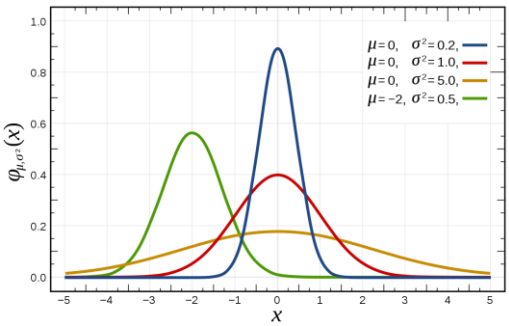

Fig. 4. Gaussian distribution graph

- Mean±1 SD contains 68.2% of all values.
- Mean±2 SD contains 95.5% of all values.
- Mean±3 SD contain 99.7% of all values

8. Future Scope

Predictive analytics involves the application of additional statistical methods and structural techniques to help develop the model. Data scientists often build multiple predictive analytics models and then select the best one based on its performance of the model.

High quality data will be consistent in its format, reflecting at real world scenario it describes, and will enable reliable, reproducible research. With sophisticated AI systems and predictive models can be used to cede control and stick to what we know will continue to deliver growth which directly implies accuracy rate for the predictions will be increased. In future, Predictive analytics will become a self-fulfilling prophecy.

9. Conclusion

In this paper, we proposed a system to predict optimal price of the product using big data approach. In machine learning we used Linear discriminant analysis (LAD) in which the file system in distributed and mean, variance of the system is predicted in an effective way. An optimal price forecast should correct any persistent biases in the predictions. Overall aim of this project is to provide a quantitative tool to optimally predict a products price and maximize the profit.

References

[1] Salo, J. and Karjaluoto, H. (2006) "IT-enabled supply chain management", Contemporary Management Research, 2(1), 17-30.
[2] Sathi, A. (2012) "Big Data analytics: Disruptive Technologies for Changing the Game", Boise, ID: MC Press.
[3] Sen, J. (2015) "Architecture of Big Data Ecosystem for Business Applications", CBS Journal of Management Practices, 2(1), 21 – 40.
[4] Sharifi, H. & Zhang, Z. (2001) "Agile Manufacturing in Practice Application of a Methodology", International Journal of Operations & Production Management, 21(5-6), 772–779. Stevension, W.J. (2009) "Operations Management", Tata McGraw – Hill, 9th Ed.
[5] Stock, J.R. (2013) "Supply Chain Management: A Look Back, A Look Ahead", Supply Chain Quarterly, 2, 22–26.
[6] Thompson, J. (1967) "Organization in Action", McGraw-Hill, New York, NY.
[7] Trkman, P., McCormack, K., de Oliveira, M.P.V., & Ladeira, M.B. (2010) "The Impact of Business Analytics on Supply Chain Performance". Decision Support Systems, 49 (3), 318-327.
[8] Van Donk, D.P. & Van der Vaart, T. (2005) "A Case of Shared Resources, Uncertainty and Supply Chain Integration in the Process Industry", International Journal of Production Economics, 96(1), 97–108.
[9] Vickery, S., Jayaram, J., Droge, C. & Calantone, R. (2003) "The Effects of an Integrative Supply Chain Strategy on Customer Service and Financial Performance: An Analysis of Direct versus Indirect Relationships", Journal of Operations Management. 21, 523–539.
[10] Waller, M.A. & Fawcett, A.C. (2013) "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management", Journal of Business Logistics, 34(2), 77–84.
[11] Yahalom, R., Klein, B., & Beth, T. (1993). "Trust Relationships in Secure Systems – A Distributed Authentication Perspective". Proceedings of 1993 IEEE Computer Society Symposium on Research in Security and Privacy, pp. 150-164.
[12] Yan, J., Xin. S., Liu, Q., Xu, W., Yang, L., Fan, L., Chen, B. & Wang, Q. (2014) "Intelligent Supply Chain Integration and Management Based on Cloud of Things", International Journal of Distributed Sensor Networks, 1-15
[13] L. J. Zhang, "Editorial: Big Services Era: Global Trends of Cloud Computing and Big Data," IEEE Trans. Services Computing, vol. 5, no. 4, 2012, pp. 467–468. 9. W.M.P. van der Aalst, "A Decade of Business Process Management Conferences: Personal Reflections on a Developing Discipline," Business Process Management, A. Barros, A. Gal, and E. Kindler, eds. Springer, 2012, pp. 1–16.
[14] W.M.P. van der Aalst, "Process Mining," Comm. ACM, vol. 55, no. 8, 2012, pp. 76–83.
[15] M. zur Muehlen and R. Shapiro, Handbook on Business Process Analytics, Springer, vol. 2, 2010.
[16] S. Rizzi, "Collaborative Business Intelligence," Proc. First European Summer School (eBISS 11), Springer, 2011, pp. 186–205.
[17] C. Costello, "Incorporating Performance into Process Models to Support Business Activity Monitoring," doctoral dissertation, Dept. of Information Technology, National Univ. of Ireland, 2008. 14. Business Process Analytics Format Specification, Workflow Management Coalition (WfMC), Feb. 2012;
[18] www.wfmc.org/Download-document/Business-Process-Analytics-Format-R1.html.
[19] O. Molloy and C. Sheridan, "A Framework for the Use of Business Activity Monitoring in Process Improvement," E-Strategies for Resource Management Systems: Planning and Implementation, E. Alkhalifa, ed., IGI Global, 2010.