ISSN:0975-9646

Nishchol Mishra et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012, 4434- 4438

Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges

Nishchol Mishra[1], Dr.Sanjay Silakari[2]

School of IT, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India[1]
Professor& Dean, Comp. Sc. & Engg., Rajiv Gandhi Proudyogiki Vishwavidyalaya , Bhopal, India[2]

***Abstract*—Predictive analysis is an advanced branch of data engineering which generally predicts some occurrence or probability based on data. Predictive analytics uses data-mining techniques in order to make predictions about future events, and make recommendations based on these predictions. The process involves an analysis of historic data and based on that analysis to predict the future occurrences or events. A model can be created to predict using Predictive Analytics modeling techniques. The form of these predictive models varies depending on the data they are using. Classification & Regression are the two main objectives of predictive analytics. Predictive Analytics is composed of various statistical & analytical techniques used to develop models that will predict future occurrence, events or probabilities. Predictive analytics is able to not only deal with continuous changes, but discontinuous changes as well. Classification, prediction, and to some extent, affinity analysis constitute the analytical methods employed in predictive analytics.**

Keywords- Predictive Analytics; Predictive Modeling; Data Mining; Prediction.

I. INTRODUCTION

Predictive analytics is composed of two words predict & analysis, but it works in reverse *viz.* first analyze then predict. It is human nature to want to know and predict what the future holds.Predictive analytics deals with the prediction of future events based on previously observed historical data by applying sophisticated methods like machine learning. The historical data is collected and transformed by using various techniques like filtering, correlating the data, and so on. Prediction process can be divided into four steps: (1) collect and pre-process raw data; (2) transform pre-processed data into a form that can be easily handled by the (selected) machine learning method; (3) create the learning model (training) using the transformed data; (4) report predictions to the user using the previously created learning model.

II. PREDICTIVE ANALYTICS AND DATA MINING

The future of data mining lies in predictive analytics. The terms *data mining* and *data extraction* are often confused with each other; but there is a significant difference [20]. Data extraction involves obtaining data from one data source and loading itinto a targeted database. Thus one may 'extract' data from a source or legacy system to put it into a standard database or data warehouse. Data Mining, on the other hand, is the extraction of obscure or *hidden predictive information* from large databases or data warehouses. Also known as *knowledge-discovery*, data mining is the practice of searching for patterns in stores of data. To this end, data mining uses computational techniques from statistics and pattern recognition. Looking for patterns in data thus defines the nature of data mining.

A predictive analytical model is built by data mining tools and techniques. The first step consists of extracting data by accessing massive databases. The data thus obtained is processed with the help of advanced algorithms to find hidden patterns and predictive information. Though there is an obvious connection between statistics and data mining, methodologies used in data mining have originated in fields other than statistics.

Data mining lies at the confluence of several streams of applied knowledge such as database management, data visualization, machine learning, artificial intelligence and pattern recognition. Most data mining techniques includes genetic algorithms, decision trees, artificial neural networks, rule induction and nearest neighbour method.

Predictive analytics is used to determine the probable future outcome of an event or the likelihood of a situation occurring. It is the branch of data mining concerned with the prediction of future probabilities and trends. Predictive analytics is used to automatically analyze large amounts of data with different variables; it includes clustering, decision trees, market basket analysis, regression modeling, neural nets, genetic algorithms, text mining, hypothesis testing, decision analytics, and more [16].

The core element of predictive analytics is the 'predictor', a variable that can be measured for an individual or entity to predict its future behavior. For example, a credit card company may consider age, income, and credit history as predictors to determine the risk factor in issuing a credit card to an applicant.

Multiple predictors can be combined into a predictive model, which is then used to forecast future probabilities with an acceptable level of reliability. In predictive modeling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data become available [16].

Predictive analytics combines business knowledge and statistical analytical techniques which, when applied to business data, produce insights. These insights help organizations understand how people behave as customers, buyers, sellers, distributors, and so on.

Multiple related predictive models produce insights for making strategic company decisions such as exploring new markets, acquisitions, and retentions; finding up-selling and cross-selling opportunities; and discovering areas that can improve security and fraud detection. Predictive analytics indicates not only what to do, but also how and when to do it, and to explain 'what-if' scenarios [20].

The major objective of data mining is to build a model that can be used to predict the occurrence of an event. The model builders will extract knowledge from historic data and represent it in such a form that the resulting model can be applied to new situations. The process of analysing data sets extracts useful information on which to apply one or more data mining techniques in order to discover previously unknown patterns within the data, or find trends in the data which can then be used to predict future trends or behaviours. Data mining can be divided into two main categories: supervised (predictive) and unsupervised (descriptive).

In supervised learning, data is modelled from training data to find patterns within the data which can then be used to predict a label or value, given some set of parameters. Supervised learning is the process of creating predictive models using a set of historical data that contains the results we are trying to predict. The type of data determines whether this is done using a Regression or a Classification algorithm.

Regression is a statistical methodology that was developed by Sir Frances Galton (1822-1911), a mathematician who was also a cousin of Charles Darwin. Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent or response variable (which is continuous valued) [17]. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + c$) and determines the appropriate values for m and c to predict the value of y based on the input parameter, x. Advance techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as higher order polynomial equations [17]. Regression is a well-established and reliable statistical technique.

Classification is the set of data mining techniques used to fit discrete (categorical) data to a known structure in order to be able to form predictions for the class label of unlabelled data. Typically, classification algorithms are done in three phases, the first two phases, training and testing, use labelled data, that is, data which has known class labels. Training uses a portion of the data to fit a classifying model to the data. The testing phase then uses the models to try and predict the class labels, and validates the predictions using the actual values in order to determine how accurate the model is. The feedback from this determines how well the models work, and whether new models should be built. Once an acceptable model is built that passes the testing phase, the classifier is deployed on unlabelled data. This is called the deployment phase. Common classification algorithms include Bayesian classification, decision trees, back-propagation and neural networks, and genetic and evolutionary learners.

Unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Unlike supervised algorithms, unsupervised algorithms do not learn from historical data with known labels, hence, they perform without any supervision. Standard unsupervised techniques include clustering, characterization, association rule mining, and change and deviation detecting techniques.

Predictive Analytics consists of a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events [14]. Predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions. Predictive analytics use data-mining techniques in order to make predictions about future events, and make recommendations based on these predictions [19] [14]. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions [16].

Generally, the term predictive analytics is used to mean predictive modeling, "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to describe related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making but have different purposes and the statistical techniques underlying them vary. Predictive models analyze past performance to assess whereas Descriptive models quantify relationships in data [15] [16].

Decision models describe the relationship between all the elements of a decision — the known data (including results of predictive models), the decision and the forecast results of the decision — in order to predict the results of decisions involving many variables [16]. These models can be used in optimization, maximizing certain outcomes while minimizing others.

III. PREDICTIVE ANALYTICS TECHNIQUES

Predictive models analyze identify patterns in historical and transactional data to determine various risks and oppurtunities. Forecasting models capture relationships between many factors to allow assessment of the risks or potential associated with a particular set of conditions, guiding decision making for candidate transactions. Three basic techniques for Predictive analytics are Data profiling and Transformations, Sequential Pattern Analysis and Time Series Tracking [16]. Data profiling and transformations are functions that change the row and column attributes and analyses dependencies, data formats, merge fields, aggregate records, and make rows and columns [15]. Sequential pattern analysis identifies relationships between the rows of data. Sequential pattern analysis involves identifying frequently observed sequential occurrence of items across ordered transactions over time. Time Series Tracking is an ordered sequence of values at variable time intervals at the same distance [15]. Time series analysis gives the fact that the data points taken over time.

There are some advanced Predictive analytic techniques like Classification-Regression, Association analysis, Time series forecasting to name a few. Classification uses attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest [15]. Regression analysis is a statistical tool for the study of relations between variables. Association analysis describes significant associations between data elements. Time series analysis is employed for forecasting the future value of a measure based on past values [16].

IV. PREDICTIVE MODELS

Although most experts agree that predictive analytics requires great skill and some go so far as to suggest that there is an artistic and highly creative side to creating models, generally predictive models need some basic steps of developing them [18]. These steps are:

Project Definition: Define the business objectives and desired outcomes for the project and translate them into predictive analytic objectives and tasks;

Exploration: Analyze source data to determine the most appropriate data and model building approach;

Data Preparation: Select, extract, and transform data upon which to create models;

Model Building: Create, test, and validate models, and evaluate whether they will meet project metrics and goals;

Deployment: Apply model results to business decisions or processes; and

Model Management: Manage models to improve performance (i.e., accuracy), control access, promote reuse, standardize toolsets, and minimize redundant activities.

Most experts hold the view that the data preparation phase of creating predictive models is the most time-consuming part of the process.

V. MODELING PROCESS:

There are various Modeling Process stages; some of them have been discussed over here as follows:

Purpose: This stage describes the objective of the project.

Obtain the data: Gathering data samples from various sources regarding the project.

Explore, clean and pre-process the data: Exploration can be performed by describing the variables, tokens and other terms which is used in project quite general. Sometimes these terms are in cryptic form or may be in short form, for which we have to tell the full explanation and the places where it can be used. We can also specify the conditions where these variables can be used.

Reduce the data and partition them into training, validation and test partitions: In this stage we try to reduce the variables or terms for the sake of simplicity. We can reduce number of variables by making the small group of similar purpose variables.

We will partition the data into a training set to build the model and a validation set to see how well the model does. This technique is a part of supervised learning process in classification and prediction problem. These problems can be used to develop other models and the value of outcome variables can be used in unknown places.

At this stage we can partition the data into training and validation. Training will build the model and partition will apply model on data to see how well the model does.

A Data mining endeavour involves testing multiple models, perhaps with multiple settings on each model. Starting from one model and test it one validation data might give us an idea about the performance of that model on such data. However when we choose the best performing model, the validation data no longer provide an unbiased estimate of how the models might do with more data. By playing the role in choosing the best model the validation data have become the part of the model itself.

Determining the data mining task: Data mining task in building the model is to find the objective.

Choosing the technique: The data which is divided into training and validation partitions can be used for creating the model by various techniques.

Use the algorithm to perform the task: In this stage we apply some of the algorithm to find fitted value (by applying algorithm on training data) and predicted value (by applying algorithm on validation data). Note that the predicted values would often be called the fitted values, since they are for the records to which the model was fit.

Prediction error can be measured in several ways.

1. Average error
2. Total sum of squared errors
3. RMS error(Root mean squared error)

Interpret the results: At this stage we try other prediction algorithms and see how they do error-wise. We might also try different settings on the various models. After choosing the best model (typically, the model with the lowest error on the validation data while also recognizing that ¨simpler is better¨), we use that model to predict the output variable in fresh data.

Deploy the model: After the best model is chosen, it is applied to new data.

VI. APPLICATIONS OF PREDICTIVE ANALYTICS

Predictive Analytics can be used in many applications. Here we cite some examples where it has made a positive impact [14].

*Medical decision support system:*Experts use predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions like diabetes, asthma, heart disease and other lifetime illnesses.

Fraud detection: Fraud is widely spread across industries. Cases of fraud appear in diverse fields such as credit card activations, invoices, tax returns, online activities, insurance claims and telecom call activities. All these industries are

interested in a) detecting frauds and bringing those responsible to book and b) preventing and monitoring fraud at reasonable costs [21]. Predictive modeling can help them achieve these objectives. The may also be used to detect financial statement fraud in companies.

Insurance: Similar to fraud, unexpectedly high and suspicious claims are the bane of insurance companies. They would like to avoid paying such claims. Though the objective is simple enough, predictive modeling has had only partial success in eliminating this source of high loss to companies. This is a promising area of further research. [21]

Health: While the systematic applications of predictive modeling in healthcare are relatively new, the fundamental applications are similar to those in the other areas. After all minimizing customer risk is the objective. In healthcare this is the risk of readmission, which can be reduced by identifying high risk patients and monitoring them. [21]

Financial prediction: Predictive analytics is useful in financial predictions.

Customer retention: By a frequent examination of a customer's past service usage, performance, spending and other behavior patterns, predictive models can determine the likelihood of a customer wanting to terminate a service sometime in the near future.

Analytical customer relationship management (CRM): Analytical Customer Relationship Management is a frequent commercial application of Predictive Analysis. CRM uses predictive analysis in applications for marketing campaigns, sales, and customer services to name a few.

VII. Related Work

R. Maciejewski et al. [5] proposed a model for spatiotemporal data, as analysts are searching for regions of space and time with unusually high incidences of events (hotspots), created a predictive visual analytics toolkit that provides analysts with linked spatiotemporal and statistical analytic views. The system models spatiotemporal events through the combination of kernel density estimation for event distribution and seasonal trend decomposition by loss smoothing for temporal predictions. *J. Yue et al.* [7] In this paper they specifically address predictive tasks that are concerned with predicting future trends, and proposed RESIN, an AI blackboard-based agent that leverages interactive visualization and mixed-initiative problem solving to enable analysts to explore and pre-process large amounts of data in order to perform predictive analytics. *R. M. Riensche et al*. [8] described a methodology and architecture to support the development of games in a predictive analytics context, designed to gather input knowledge, calculate results of complex predictive technical and social models, and explore those results in an engaging fashion. *Z. Huang et al.* [11] applied predictive analytics techniques to establish a decision support system for complex network operation management and help operators predict potential network failures and adapt the network in response to adverse situations. The resultant decision support system enables continuous monitoring of network performance and turns large amounts of data into actionable information. *Sanfilippo et al*.[9] Proposed New methods for anticipatory critical thinking have been developed that implement a multi-perspective approach to predictive modeling in support of Naturalistic Decision Making.*R. Banjade et al*. [2] this paper considers linear regression technique for analyzing large-scale dataset for the purpose of useful recommendations to e-commerce customers by offline calculations of model results. *V. H. Bhat et al.* [1] presents a novel pre-processing phase with missing value imputation for both numerical and categorical data. A hybrid combination of Classification and Regression Trees (CART) and Genetic Algorithms to impute missing continuous values and Self Organizing Feature Maps (SOFM) to impute categorical values is adapted in their work.*V. H. Bhat et al.* [10] proposed an efficient imputation method using a hybrid combination of CART and Genetic Algorithm, as a preprocessing step. The classical neural network model is used for prediction, on the pre-processed dataset. *N. Chinchor et al.* [4] this tutorial addresses combining multimedia analysis and visual analytics to deal with information from different sources, with different goals or objectives, and containing different media types and combinations of types. The resulting combination is multimedia analytics. *M. A. Razi et al.* [13] performed a three-way comparison of prediction accuracy involving nonlinear regression, NNs and CART models using a continuous dependent variable and a set of dichotomous and categorical predictor variables.

VIII. Oppurtunities & Challenges

It has been widely quoted that "information is the new oil". We've traveled from an industrial age, powered by hydrocarbons, to an information age driven by data. Predictive analytics, broadly defined, focuses on extracting features from data and building models that can predict future events:

Here we discuss some of the outstanding challenges in this field, with regard to: (i) privacy and ownership of data, (ii) analysis of user data, (iii) scaling of algorithms, and (iv) emerging data ecosystems & exchanges [22].

Privacy and Ownership of Data: Privacy and ownership of data is big issue. There is always conflict between producer and consumer of data, there are many organizations that believe that data should be open and that openness and interoperability provide them with a competitive advantage [22].

Analysis of User Data: The major focus of analysis of user data is in determining user's intent. This is certainly the focus of a lot of the predictive analytics used in online advertising, and the reason that search advertising is far more effective than display advertising [22].

Scaling of Algorithms: Having more data is always beneficial for data based system, due to Popularization of social media huge database repository has been created, we have to push

the limits in terms of scalability for systems . "The major problem associated with scaling algorithms is that communications and synchronization overheads go up and so a lot of efficiency can be lost, especially where the computation doesn't fit nicely into a map/reduce model"[22].

Data Ecosystems and Exchanges: **"**The emergence of data exchanges is clearly related to the problems with ownership of data. They allow data to be exchanged under a clear set of rules with ownership and conditions contractually determined. They also allow for a company to have a viable business model as a data provider and so provide useful data to the whole ecosystem without having to also compete on how the data is used"[22].

IX. CONCLUSION

The future of Data Mining lies in Predictive Analytics. This study mainly focuses on opportunities, applications, trends & challenges of Predictive Analytics in Knowledge discovery domain. Predictive Analytics is an area of interest to almost all communities and organizations. Predictive analytics is using business intelligence data for forecasting and modeling. Proper data mining algorithms and predictive modeling can refine search for targeted customers. Predictive Analytics can aid in choosing marketing methods, and marketing more efficiently. Predictive Analytics can be also helpful in Social Media Analytics.

REFERENCES

[1] V. H. Bhat, P. G. Rao, S. Krishna, and P. D. Shenoy, "An Efficient Framework for Prediction in Healthcare," *Most*, Springer-Verlag Berlin Heidelberg , pp. 522-532, 2011.

[2] R. Banjade and S. Maharjan, "Product Recommendations using Linear Predictive Modeling," 2011.

[3] Debahuti Mishra et al., "Predictive Data Mining: Promising Future and Applications", Int. J. of Computer and Communication Technology, Vol. 2, No. 1, pp. 20-28, 2010.

[4] N. Chinchor, J. Thomas, and P. Wong, "Multimedia Analysis+ Visual Analytics= Multimedia Analytics," IEEE *Computer Graphics*, 2010.

[5] R. Maciejewski et al., "Forecasting Hotspots - A Predictive Analytics Approach.," IEEE transactions on visualization and computer graphics, vol. 17, no. 4, pp. 440-453, May 2010.

[6] A. Guazzelli, K. Stathatos, and M. Zeller, "Efficient deployment of predictive analytics through open standards and cloud computing," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 32–38, 2009.

[7] J. Yue, A. Raja, D. Liu, X. Wang, and W. Ribarsky, "A blackboard-based approach towards predictive analytics," in *Proceedings AAAI Spring Symposium on Technosocial Predictive Analytics*, pp. 154–161, 2009.

[8] R. M. Riensche et al., "Serious Gaming for Predictive Analytics," in AAAI Spring Symposium on Technosocial Predictive Analytics. Association for the Advancement of Artificial Intelligence (AAAI), San Jose, CA, no. Zyda, pp. 108-113, 2009.

[9] Sanfilippo et al., "Technosocial Predictive Analytics in Support of Naturalistic Decision Making," *Symposium A Quarterly Journal In Modern Foreign Literatures*, no. June, pp. 144-151, 2009.

[10] V. H. Bhat, P. G. Rao, and P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques," *Architecture*, , Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009.

[11] Z. Huang et al., "Managing Complex Network Operation with Predictive Analytics,"Proceedings of the AAAI Spring Symposium on Technosocial Predictive Analytics, *Science*, pp. 59-65, 2009.

[12] C. Mccue, "Data mining and Predictive analytics in public safety and Security," *Analysis*, IEEE Computer Society, pp.12-18, August, 2006.

[13] M. A. Razi and K. Athappilly, "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models," *Expert Systems with Applications*, vol. 29, no. 1, pp. 65–74, 2005.

[14] http://en.wikipedia.org/wiki/Predictive_analytics.

[15] http://analyticsweb.com/predictive-analytics.

[16] http://www.articlesbase.com/strategic-planning-articles/predictive-analytics-1704860.html

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

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

[19] http://analyticsage.com.

[20] M Zaman, Predictive analytics; the future of business intelligencewww.mahmoudyoussef.com

[21] http://www.informationbuilders.com/blog/rado-kotorov/2276.

[22] http://www.quora.com/Predictive-Analytics/What-are-the-most-significant-challenges-and-opportunities-in-predictive-analytics.