Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

EFFECTIVE MODELLING FOR PREDICTIVE ANALYTICS IN DATA SCIENCE

P. Wazurkar, R. S. Bhadoria*

Department of Computer Science and Engineering, Indian Institute of Information Technology (IIIT) Nagpur, Seminary Hills, Nagpur 440006, Maharashtra, India

Received: 03 March 2018 / Accepted: 26 April 2018 / Published online: 01 May 2018

ABSTRACT

Predictive analytics includes many statistical and other empirical methods that create various data predictions as well as different methods for assessing predictive power. Predictive analytics not only helps in creating practically useful models but also plays an important role in building new theory for further study and research. Today, the use of available data to extract inferences and predictions by using predictive analytics has grown in the industry from being a small department in large companies to being an active component in most mid to large sized organizations. This paper addresses to reduce a particularly large gap of, the near-absence of empirical or factual predictive analytics in the mainstream research going on in this field by analyzing the issues faced in predictive modelling by the empirical determination of data with its experimental facts for latency pattern.

Keywords: Predictive Analytics, Big Data, Business Intelligence, Project Planning.

Author Correspondence, e-mail: robin19@ieee.org

doi: http://dx.doi.org/10.4314/jfas.v10i2.14

1. INTRODUCTION

The usage of advanced analytics is not very new, traditionally referred to as data mining or computational statistics was mainly used by the companies with deep pockets. Predictive analytics includes many statistical and other empirical methods that creates various data predictions as well as different methods for assessing predictive power. In the last decade, the field of big data has made great progress in predictive analytics sector by employing more advanced statistical modeling techniques and various different algorithms to support ongoing research. It has nowadays become common in the industry to see researchers and analysts to use the available big data in extracting different insights and developing better models by using various different algorithms so as to better facilitate the concerned problem for better prediction. Today, the use of predictive analytics has grown from being a small group in large companies to being an instrumental component of most mid- to large-sized organizations. These analytics mostly begins with business intelligence (BI) and later moves into the predictive analytics (PA) as the available data grows and the pressure to produce greater benefits from this data increases. Even small organizations, for profit and nonprofit, have been benefited from predictive analytics now, by using open source software so as to drive decisions on a small scale. Even after such development, many different opportunities for the further improvement in this domain remain. A lot of challenges are faced during the modelling process, which are required to be addressed so as to better facilitate the user.

This paper addresses to reduce a particularly large gap of, the near-absence of empirical or factual predictive analytics in the mainstream research going on in this field. As reducing this particular big gap tries to present a very important opportunity, to solve a large number of intricate problems like, customer segmentation and/or community detection in the social sphere, for credit scoring or predicting the next outcome of many time-driven events, generating new theory on the basis of preexisting data and developing new measures to solve various problems. We try to manifest that predictive analytics helps in developing and examining theoretical models by bringing a different perspective to the problem being solved.

Remaining part of the paper is organized into 6 Sections. Section-2 reviews the available literature and presents the related and background works done while section-3 discusses the challenges and issues faced during predictive modelling. In Section-4, we visit the methodology for predictive modelling while Section-5 about the result and analysis and finally Section-6 concludes the paper.

2. RELATED & BACKGROUND WORKS

Over the years the field of predictive analytics has evolved but the advent of increasing data in day to day lives demands for improvement in the ways these data is handled. Müller in 2016 stated that in the natural sciences, the evolution of the scientific method is often portrayed into four eras (Bell et al, 2009; Hey et al, 2009). During the earlier times, research was based on empirical observations, which was later followed by an era of theoretical science, in which building of some causal models was cultivated. As problems became complex the models started to evolve, and an era of computational research using simulations emerged (Müller 2016). Predictive analytics, has been mostly dealt with structured data, but with the increase in the amount of unstructured data, which nowadays constitutes 95% of big data requires new analytical methods to be developed (A Gandomi 2015).

The approach discussed in Delen 2013 presents the analytics in the cloud that will be creating great opportunities in the field of analytical development. The results discussed here show that theoretical predictive analytics tools would be more apparent to have larger impact. It also combines the theoretical insights available within large amounts of data. The advent of increase in the role of big data and predictive analytics in a retailing context has risen in importance due to increase in newer sources of data and improvement in techniques of theory, domain specific knowledge, and application of available statistical tools which is likely to continue unhampered (Bradlow 2017). The application of predictive analytics into online marketing, such as predicting the performance of online reviews by using a sentiment mining approach has shown that the sentiments negatively affect the performance of online reviewing by raising two exceptions, viz. positive sentiments in the title of review and neutral sentiments in the text of review (Salhan 2015). Predictive analytics is also influenced by the extraction of information from extremely large data sets and also from a large variety of data structures. While these models are more data-driven than the conventional illustrative statistical models, in the sense that former combines knowledge from pre-existing theoretical models in a less traditional way than the later (Shmueli 2011). The applications of predictive algorithms are not only limited to the online world. Health care industries are also transiting towards better utilizing it to provide quality services to humanity. The predictive models based on the data of individual health costs and outcome provides a “risk score” which improves costs and quality of health care (Tomar 2016).

3. CHALLENGES AND ISSUES IN PREDICTIVE ANALYTICS

Predictive analytics can be used to generate some significant improvements in efficiency of decision-making and thus improving on the return on investment. But predictive analytics isn’t always successful as they face a lot of challenges. In this section we discuss some issues which can hinder the process of creation of a better model for predictive analysis. Some of the most common reasons predictive models don’t succeed can be grouped into four categories:

- Issues with available resource management. Improper management of available resources like time, data, and number of analysts may lead to problems.
- Issues with data. The mismanagement in the data for model creation as well as model testing may create issues.
- Issues with modeling. Improper selection of different number of variables for modelling may induce unexpected issues.
- Issues in model deployment. After the model has been created the selection of improper audiences may lead to wrong results.

For models that try to estimate or predict specific values, an important step in the Business Understanding stage is identifying one or more target variables to predict. A target variable is actually a column in the available modeling data which contains the values to be estimated or predicted as predecided in the business objectives. The selected target variables can be numeric or categorical which depends on the type of model that will be built. Defining and selecting the target variable is a critically important task in a predictive modeling project as it is the only information the modeling algorithms use to derive what the modeler desires from the predictions. Algorithms do not possess a common sense to bring context to the problem in the way the modeler wishes. Thus, the target variable definition must be described or quantified as much as possible in the business objectives itself so that the issue of inappropriate results does not arise. To determine and rate a model as a good model depends on the specific interests of the organization and is specified as the business success criterion. After the successful determination of business success criterion it needs to be converted to a predictive modeling criterion so that the modeler can use it in a better way for selecting models. If the purpose of the model is providing highly accurate predictions to facilitate decisions to be used in the business different measures of accuracy will be used and thus this criterions should be taken care of.

4. METHODOLOGY FOR PREDICTIVE MODELLING

In this section, we briefly try to present an overview of the steps to be considered in the process of structuring and examining a predictive model and deduce a methodology to effectively model the available data for better predictions. Predictive analytics is the method for discovering the interesting and meaningful patterns in available set of data. It also includes the study of machine learning, pattern recognition and data mining. It is different from other analytics in many ways as it focuses on finding patterns in the data and follow a more data-driven approach, i.e. the algorithms derives the key characteristics for the models from the data itself and relies less on the hypothesis made by the analysts. Such data-driven algorithms deduce the different models from the data that includes the identification of variables for various parameters or coefficients.

In the context of predictive modeling, this paper uses the term predictive analytics to refer to the process of building and assessing of a model aimed at making factual predictions. It thus includes two components:

- Deploying various statistical models and other methods such as data mining algorithms specifically designed for predicting new/future observations or scenarios.
- Methods for evaluating the created model and assessing its predictive power.

The most important part of modeling a better predictive model for a particular project is the step in the very beginning when the predictive modeling project is actually defined. To set up a predictive model for a project is a very difficult task as the skills required to do it well are very broad, requiring knowledge of the business domain, databases, predictive modeling algorithms and techniques to mention a few. Very few individuals have all of these skill sets, and therefore setting up a predictive modeling project is inevitably a team effort. A step-by-step process provides a structure for analysis and facilitates the analyst of the steps that needs to be accomplished, also for the need of documentation and reporting throughout the process, which is valuable for the modelers. However, most of the practitioners are hesitant in describing the modeling process in linear, step-by-step terms as projects almost never proceed as planned due to the problems with data modeling.

On a broad note steps that should be followed for predictive analytical process are Understanding project, Understanding the available data, preparing the data, creation of model and its evaluation with deployment. These steps, and the sequence they appear, represent the most common sequence in a project.

Table 1. Describe different stage of Data Model

STAGE	DESCRIPTION
Understanding Project	Defining and describing the project.
Understanding the Data	Examining the data and identifying the problems.
Preparing the Data	Fix problems in data and deriving variables.
Creation of Model	Building predictive or descriptive models.
Evaluation of the Model	Assessing the model.
Deployment	Planning for usage of model.

A schematic for the steps involved in model building for explanatory and predictive modeling is shown in Figure 1. Although the main steps are the same, inside each step the predictive model dictates different operations and criteria.

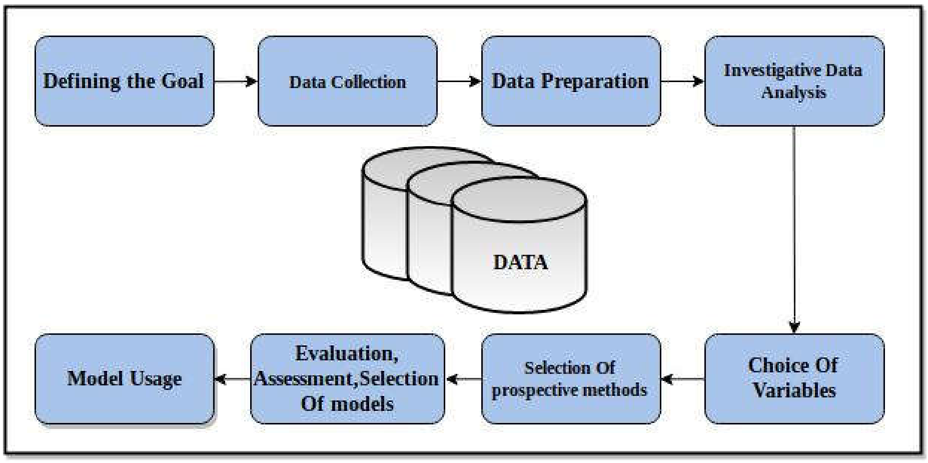
Defining the Goal
Data Collection
Data Preparation
Investigative Data Analysis
DATA
Model Usage
Evaluation, Assessment,Selection Of models
Selection Of prospective methods
Choice Of Variables

Fig.1. Steps involved in predictive modelling

4.1 Defining the Goal

Building a predictive model requires careful determination of what is specifically needed to be predicted and what are the goals to be fulfilled at the end of the project, as this impacts the type of models and methods used later on. The goals differ on the basis of the outcome expected like, the goal is known as prediction for a numerical outcome or classification for a categorical outcome. When the categorical outcomes are ranked according to their probability in the set of observations in a particular class is known as ranking.

4.1.1 Data Collection

In predictive analytics, closeness associated with the collected data to be used for modeling to the prediction context is a main consideration. In predictive analytics the sample size of data required is often large and thus results in lot of variables. The initial number of variables usually are numerous due to an effort to capture new sources of information and new relationships. Justification for every variable is based upon the combining theory, domain knowledge, and initial analysis.

4.1.2 Data Preparation

The available data set is usually partitioned into two parts. The training set is used to fit the models. A holdout set is kept reserved for evaluating the final model on the basis of its predictive power and performance. A third data set i.e. validation set is commonly used for model selection. The final model, selected on the basis of validation set, is then evaluated for the holdout set.

4.1.3 Investigative Data Analysis

It consists of summarizing the data numerically and graphically, to reduce their dimension, and to handle the resulted outliers. As the number of predictors is often large, reducing their dimensions can help in reducing sampling variance and in turn increase predictive accuracy. The resultant variables can then be used as better predictors.

4.1.4 Choice of Variables

Predictive models are based on the association of predictors and the responses. Hence the variables are chosen on the basis of their perceptible properties. The response variable and its scale are chosen based upon the predictive goal, data availability, and precision of their measurement.

4.1.5 Selection of Prospective Methods

Predictive models often depend on nonparametric data mining algorithms (e.g., neural networks, classification trees, and k-nearest-neighbors). The flexibility provided by such methods enables in capturing complex relationships in the data without making restricting statistical assumptions.

4.1.6 Evaluation, Assessment and Selection Of models

To evaluate a model predictive accuracy is taken into account and is measured by applying the model to the defined holdout set by generating predictions. The models are assessed and selected on the basis of outcomes derived from the comparison of results drawn from the training and holdout sets.

4.1.7 Model Usage

Outcomes of predictive analytics depend on the predictive accuracy of the model. Performance measures (e.g., error rates) and plots (e.g., ROC curves and lift charts) are aimed towards conveying the predictive accuracy of the model. The assessed predictive power is again compared against naive and alternative predictive models available.

4.2 Algorithms for Predictive Analytics

Predictive modeling discusses the algorithms which are more of supervised learning approaches and tries to find the relationships among the inputs to one or more target variables. The target variable is the key.

Good predictive models requires to target Predictive modeling algorithms can be placed into a few categories:

- Local estimator algorithms.

- Global estimators that do not localize.
- Functional estimators that apply globally but can localize their predictions.

4.2.1 Logistic Regression Algorithm

The core of a logistic regression model is the odds ratio: the ratio of the outcome probabilities.

$$\textit{odds ratio} = \left\{\frac{P(1)}{1-P(1)}\right\} = \frac{p(1)}{p(0)}$$

This linear relationship is the statistical model logistic regression in computing:

$$\text{odds ratio} = \frac{P(1)}{1-P(1)} = w_0 + w_1 \times \ x_1 + \cdots + w_n \times x_n$$

Here, the values w_0 , w_1 , and so forth are the model coefficients or weights. The coefficient w_0 is the constant term in the model, sometimes called the bias term. The variables x_0 , x_1 , and so forth are the inputs to the model.

Some implementations of logistic regression highly depend on the variable selection options. The p value is used as a metric to decide whether to include or exclude a variable.

The best way to handle multi-class variable with N different levels is to divide them into N –1 dummy variable, and then build a single logistic regression model for each of these dummy variables. Each variable, as a binary variable, represents a variable for building a “one vs. all” classifier, meaning that a model built from each of these dummy variables will predict the likelihood. A record belongs to a particular level of the target variable in contrast to belonging to any of the other levels. While you could build N classifiers, modelers usually build N–1 classifiers and the prediction for the Nth level is merely one minus the sum of all the other probabilities:

$$p_r(focus_{=N}) = \sum_{i=1}^{i=N} P_r(focus_{=i})$$

Algorithm 1 Predict Model to Create and Collect

```
Input:  there are sufficient data to map like  N←All Given Data
        Define focus $\epsilon$ N and $C_N$ ←Create Collect
        Set Focus ∀ i∈ N, such that it covers all data in $C_N$
        Train set T ∀ $C_N$ to evaluate parameters $\emptyset_T$
            $\emptyset_T \leftarrow train\ (T, C_N)$
```

```
    for i = 1 to N do
        Check  Latency $L_i$ for each collect $C_i$
        // Perform the test on latency for each collect whether any data is remaining
    end for
return $\emptyset_T$
```

5. RESULTS AND ANALYSIS

This section presents the experimental results associated with usage of above mentioned methodology by applying logistic regression algorithm and tries to find the relationships among the inputs as shown in Figure 2 and 3. Logistic regression is a linear classification technique for binary classification. Some implementations of logistic regression include variable selection options, usually forward selection or backward elimination of variables. The benefit of this approach is that the number of collect that justifies the need to classes which are associated to given input data. This classifiers scale linearly with the number of latency which found between two parameters during training as shown in Figure 2. The latency is the frequency of data appeared over period of time and this depicts the individual collect of the data among group of similar data. However, this approach groups all the data into a single value, nonetheless of the relations amongst variables, and therefore the model may suffer in accuracy as a result.

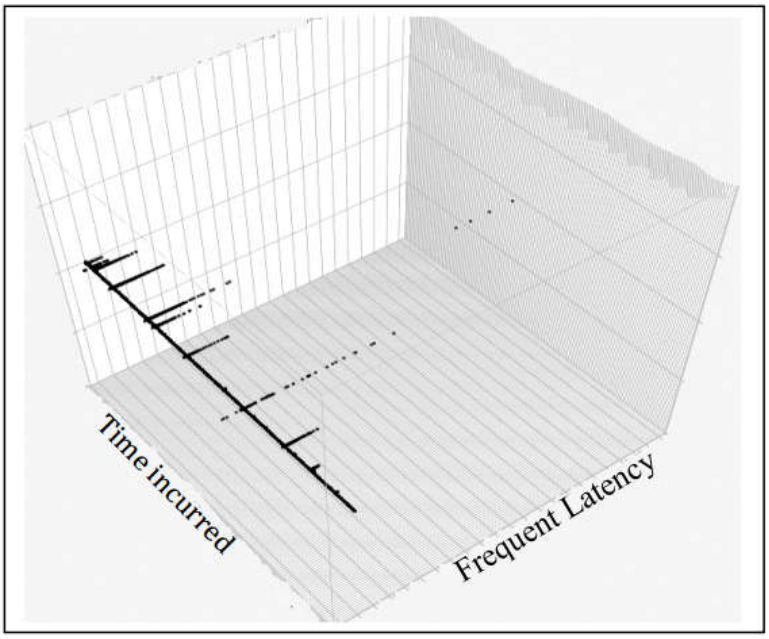

Fig.2. Shows the latency of data over time incurred

Further, with the motive of implementing predictive analytics algorithm is to segregate the dataset over frequent pattern found, that could be classified based on training. This scenario is tested for data coming from different geographical locations. All such data is mostly segregated at server situated in India and Kuwait as shown in Figure 3.

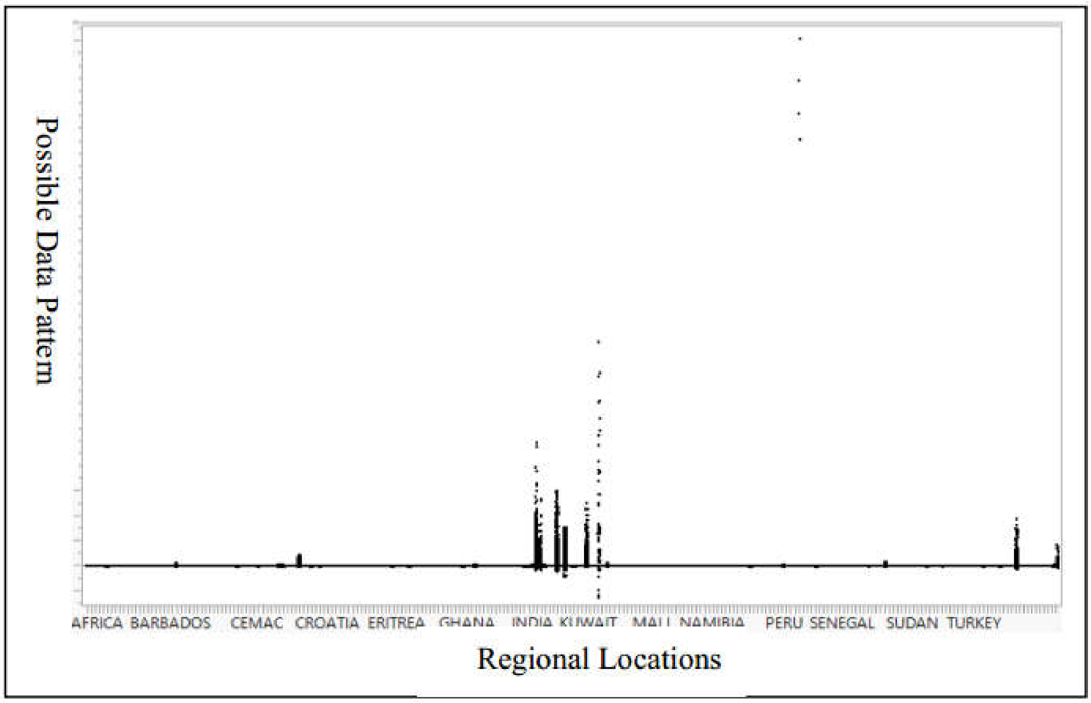

Fig.3. Shows the possible data pattern over different regional location

6. CONCLUSION

In this paper, we discussed need of predictive analytics in scientific research and also discussed the challenges and issues faced in the process of building and assessing such model. Later, we proposed an algorithm for create and segregate data based on different *collect*. This paper also discussed the latency to predictive the data based on different collects. This paper demonstrates the results for segregation of data based on different geographical locations. This paper convinces the main goal of the modeling and augmenting the different parameters which train to evaluate the substantial insight to get fruitful outcomes.

7. REFERENCES

[1] Müller, O., Junglas, I., Brocke, J. et al. Eur J Inf Syst (2016) 25: 289. https://doi.org/10.1057/ejis.2016.2

[2] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

[3] Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. Journal of Retailing, 93(1), 79-95.

[4] Delen, D., & Demirkan, H. (2013). Data, information and analytics as services.

[5] Mohammad Salehan, Dan J. Kim, Predicting the Performance of Online Consumer Reviews: A Sentiment Mining Approach to Big Data Analytics, Decision Support Systems (2015), doi: 10.1016/j.dss.2015.10.006.

[6] Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. Mis Quarterly, 553-572.

[7] Shmueli, G.” To Explain or To Predict?” Statistical Science 25.3 (2010): 289-310.

[8] Wazurkar, P.,Bhadoria, R. S., Bajpai D., (2017, November). Predictive Analytics in Data Science for Business Intelligence Solutions. In Communication Systems and Network Technologies (CSNT), 2017 Seventh International Conference on . IEEE.

[9] G.S. Tomar, N.S. Chaudhari, R.S. Bhadoria, and G.C. Deka, The Human Element of Big Data: Issues, Analytics, and Performance, FL:CRC Press, Sept. 2016.

[10] Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management, 36(5), 700-710.

[11] Wang, C-H., Cheng, H-Y., Deng, Y-T., Using Bayesian belief network and time-series model to conduct prescriptive and predictive analytics for computer industries, Computers & Industrial Engineering (2017), doi: https://doi.org/10.1016/j.cie.2017.12.003.

[12] Waller, Matthew & Fawcett, Stanley. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. Journal of Business Logistics. 34. . 10.1111/jbl.12010.

[13] Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

How to cite this article:
Wazurkar P, Bhadoria RS. Effective modelling for predictive analytics in data science. J. Fundam. Appl. Sci., 2018, *10(2), 202-212.*