Spring 3-27-2012

A Business Intelligence Framework to Provide Performance Management through a Holistic Data Mining View

Masoud Pesaran Behbahani
Azad University (IR) in Oxford, Masoud@Kingston.ac.uk

Islam Choudhury
Kingston University London, i.choudhury@kingston.ac.uk

Souheil Khaddaj
Kingston University London, s.Khaddaj@kingston.ac.uk

A Business Intelligence Framework to Provide Performance Management through a Holistic Data Mining View

Masoud Pesaran Behbahani
Email: Masoud@Kingston.ac.uk
Azad University (IR) in Oxford, Oxford, UK
School of Computing and Information Systems, Kingston University London, UK
Islamic Azad University, Khorasgan Branch
Dr. Islam Choudhury
Email: I.Choudhury@Kingston.ac.uk
School of Computing and Information Systems, Kingston University London, UK
Dr. Souheil Khaddaj
Email: S.Khaddaj@Kingston.ac.uk
School of Computing and Information Systems, Kingston University London, UK

Abstract

Traditional views of business intelligence have mainly focused on the physical and human aspects of the organization. This paper tries to show that a new information view of business activities can make a platform for developing business intelligence and support performance management. To do that, the paper proposes a new framework that can be used to provide high level of business intelligence for performance management usage. The framework introduces a hierarchy of performance influencers and a new methodology for managing them. The new methodology introduces a holistic view towards data mining concepts. The framework can be served as a blueprint for the companies which use any of ecommerce business models.

Keywords: Business Intelligence, Data Mining, Performance Management

1.0 Introduction

In this paper, a new enhanced E-Business Analytical Framework (EBAF) is introduced that employs a new proactive data mining approach. This framework helps to improve performance and serves as a practice blueprint for any Small Medium Enterprise (SME) wishing to enhance its Customer Relationship Management (CRM) and lead generation, and will give competitive advantages to SMEs that utilize this framework. EBAF presents a conversion model as a template for wide range of business models to create layers of required mining structures. EBAF is highly scalable and can be extended to be even more useful in big enterprises. The approach is more business-driven, rather than current software-driven ones (Fernandez, 2011).

Business data is the heart of the methodology. Business transactions are producing huge amount of business data on a daily manner. Every website hit, every webpage visit, and every credit card payment, are updating values and creating new rows in Online Transactional Processing (OLTP) databases. Data streams and high-speed

generated instances of data have attracted attention of the data mining community over the last decade to develop new techniques to improve organizational performance of commercial enterprises (Gaber, 2012). Whereas companies are constantly searching for strategic methods in order to stay competitive, the huge amount of the business data seems to be a good resource for achieving this goal. Companies that conduct similar business activities collaborate with each other by sharing databases to gain the mutual benefit (Divanas, 2009; Dai, 2010; Rahbarinia, 2010). Various techniques of data mining have been presented by a variety of tools to effectively analyze these huge amounts of stored or streaming data (Wang, 2008). Data mining is a step further than data warehousing and even knowledge discovery in database process (Shearer, 2000). Whereas data warehousing is simply a method for organizing the data, data mining is a database application that can take advantage of data to find hidden patterns and unknown relationships (Bertino, 2005).

The paper introduces and explains the new multilayer data mining approach. A multilayer data mining methodology needs multilayer mining structure. To identify the components of these mining structure layers, the organization should be analysed. In fact traditional views of business activities, like that of Kotler and Kelly (2006) have mainly focused on the physical and human aspects of the organization. The information view of them started getting conceptualized with contributions from Holland and Naude (2004), Jayachandran et al. (2005) and Kumar Kar et al. (2010) by emphasizing on marketing activities. An analysis based on this point of view, resulted in a new multilayer mining structure concept. This structure can be a platform for a new multilayer data mining model. The multilayer data mining strategy can be generalized and utilized in any kind of organization to provide high levels of optimization.

2.0 Holistic Data Mining View

The proposed methodology for enhancing organizational performance includes a procedure based on a new data mining concept. For a deep perception of the procedure, mining structure and mining model terminologies must be clearly defined. Mining structure generally specifies the number and type of attributes and optionally partitioning the source data into training and testing sets. Data mining structures can

even contain nested tables to provide additional detail. In EBAF mining structures, EBAF conversion model components are being used as data source to prepare the structure for the middle layers of optimization.

Whereas mining structure stores information about the data source, mining model stores information derived from statistical processing of the data. Multiple mining models can be derived from a single mining structure. Each mining model includes bindings, metadata and patterns. The bindings that are stored in the model point back to the data cached in the mining structure. If the data has been cached in the structure and has not been cleared after processing, these bindings enable to drill through from the results to the cases that support the results. The metadata includes a list of the attributes from the mining structure that is used to build the model, the description of optional filters that are applied during process, and the algorithm that is used to analyse the data. It also includes the name of the model and the server where it is stored in. The main content of mining model i.e. patterns, can be in quite a few forms such as if-then rules that describe how objects are grouped together in a transaction, decision trees that can segment objects into groups, mathematical models with equations that describes patterns and can be used to forecast the future, and a set of clusters that define the characteristics of objects in the dataset.

In the proposed model which we name it multilayer data mining, a business is modelled into a multilayer data structure in which every layer can include several mining models involving various mining algorithms. The outcomes of each layer of mining models will be included in the mining structures of the next layer. Source data used in first layer data structures depend on the business model of the company. In a purchase business model, first layer of mining structures can be generated from the leaf level data such as company's products and services data, customers' demographic data, geographic data, behavioural data, recency/ frequency/ monetary transactional data, preferences/ interests/ hobbies data, psychographic data, propensity model, and media interaction data. In a bottom-up process, first layer of mining models are being produced. Based on these models, first layer of business conversion rate influencers which are introduced as EBAF conversion model components can be individually optimized. At the second layer of mining structures, these components are used as datasets to optimize the whole performance of the organization. The point in EBAF optimization is that conversion model components such as social media marketing, traditional media advertising, email marketing, Search Engine Optimization (SEO),

Pay Per Click (PPC) and other influencers on business efficiency which are individually optimized through previous data mining process are the inputs of next layer of data mining algorithms.

The number of mining structure layers in the proposed methodology depends on the size of the business. For a small scale commercial company, we may organize the business influencers in fewer layers, but for a huge enterprise, probably more layers of influencers are needed to be constructed. Fig. 1 shows a sample of layers for a smaller business at the left side and for a bigger one at the right side. For the smaller business, the demographic data has been used in building first layer of mining structures to enhance social media marketing activities of the company. The resulted social media marketing in turn has been used in optimizing retention activities. But for a bigger company, as shown at the right side of the figure, we may use the demographic data as first layer of mining structures for increasing the effect of individual social media marketing components.

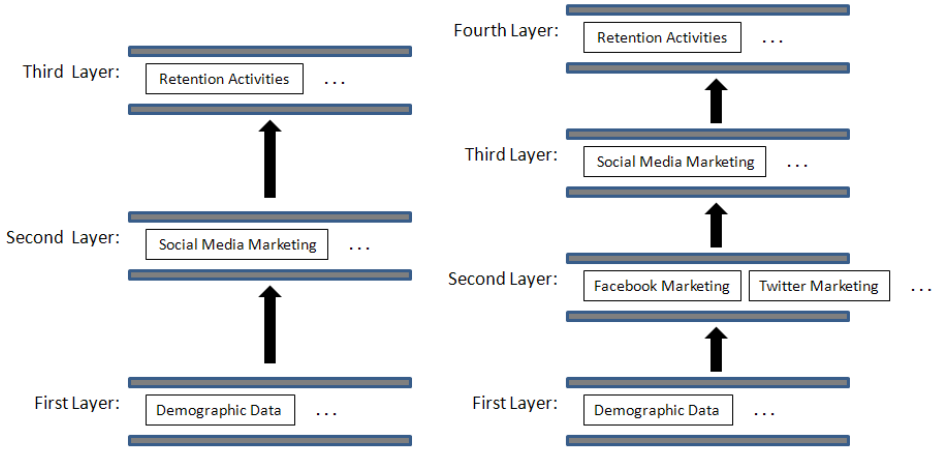


Figure 1. A large-scale company may need more layers of mining structures

Web 1.0 was the one-way interaction the user had with a webpage. People view a webpage and that's it, moving onto the next page. Web 2.0 is now the two-way interaction a user has with a website. People view a webpage but now they can place content onto the site and see what others are doing. In a modern Ecommerce, a successful business knows that how user generated content in social media channels and online communities can help to push forward the next generation of ecommerce. The new important feature of Ecommerce is its flexibility in communication between companies and customers. Unlike traditional ecommerce where the same message tends to be broadcast to everyone, emerge of mass customization lets the companies to

deliver customised content to groups of users. Facebook and Twitter marketing are shown in the figure as examples of these social media marketing influencers. By the term social media, we mean the web-based applications or channels for social interaction, compare to industrial or traditional media. Generally, social media marketing components which can participate in constructing a new layer of mining structures are blogs, social news services, social networking services such as Facebook marketing and Twitter marketing, community building services such as forums and wikis, social media sharing services such as YouTube, social classifications services and folksonomies like Del.icio.us as a social bookmarking service. A web service is a website that is designed to be used by programs rather than by people. The power of web services comes when disparate service providers are combined in mashup or different services of one provider combined in remixes. Both words originated in the music industry. We then may use the results of applying related mining models in constructing a new layer of mining structures to enhance social media marketing strategy of the company as a whole. This optimized social media marketing analytics then can be used in an endeavour to increase retention efficiency of the company.

The inspiration in multilayer data mining methodology originated during optimization process of a multivariable business environment, and then developed by an academic research project. Multilayer data mining model is functional in enterprises with any size, and obviously bigger enterprises might need more layers to reach the desired point.

3.0 Contextual Study

A five-stage conversion model including awareness, contact, engagement, conversion and retention phases is proposed to help identify mid-level mining structures in business domain. A simple overview of EBAF conversion model is presented in Fig. 2. We desire each phase to have the highest possible yield because they also act like funnels that each one feeds into the next. Inefficient awareness activities prevent the target audience know about the business services. Inefficient contact activities will limit the traffic to the site. An inefficient website with low conversion rate will restrict the number of customers. Inefficient retention follow-up activities will fail to extract additional value from the clients (Dimitriadez, 2006).

In addition to a conversion model for dealing with business activities, a model for assessing the success of the ebusiness is also needed. The left side of the figure shows how EBAF classifies the people to six main state groups, target audience, aware target audience, unique visitors, active unique visitors, actors, and finally the clients. A person in target audience may progress to a client state after traversing the required pathways. Some of these states groups may be divided to initial and repeat subgroups to provide more accurate assessments.

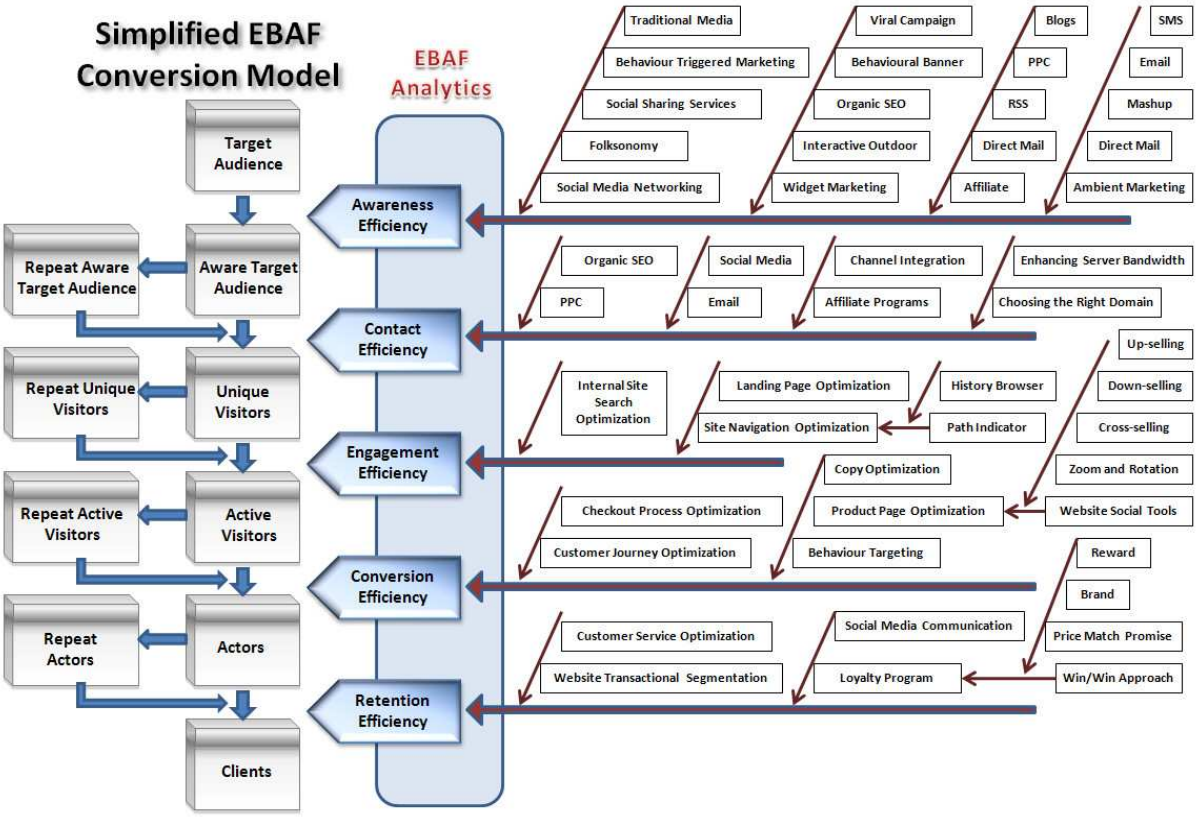


Figure 2. Simplified Proposed Conversion Model for Building Middle Layer Mining Structures

According to this classification, five efficiency coefficients can be defined: awareness efficiency, contact efficiency, engagement efficiency, conversion efficiency, and retention efficiency. Necessary efficiency variables defined in EBAF conversion model are shown in the table.

Table 1 - EBAF Conversion Model Variable Definition

Variable	Meaning	Description
Q_0	Target Audience	Number of potentially interested people with Web access at which the marketing message is aimed.

Q_1	Aware Target Audience	Number of potentially interested people with Web access who are aware of the site.
Q_{1R}	Repeat Aware Target Audience	Number of potentially interested people with Web access who has also been aware previously and are reminded again. This factor is important because people forget, and a successful marketing campaign should continuously revisit them.
Q_{1S}	Total Aware Target Audience	It's simply $Q_1 + Q_{1R}$
Q_2	Unique Visitors	Number of individual visitors. A visitor, who hits the site for several times in an interval, is considered as just one unique visitor.
Q_{2R}	Repeat Unique Visitors	Number of unique visitors who has made at least one (or a certain amount) previous visit(s) to a website.
Q_{2S}	Total Unique Visitors	It's simply $Q_2 + Q_{2R}$
Q_3	Active Visitors	Number of unique visitors who are engaged.
Q_{3R}	Repeat Active Visitors	Number of unique visitors who has made at least one (or a certain amount) previous engaged visit(s) to a website.
Q_{3S}	Total Active Visitors	It's simply $Q_3 + Q_{3R}$
Q_4	Actors	Number of visitors who have fulfilled the desired action.
Q_{4R}	Repeat Actors	Number of actors who has made at least one (or a certain amount) previous act(s).
Q_{4S}	Total Actors	It's simply $Q_4 + Q_{4R}$
Q_5	Clients	Number of loyal customers

This analytic model also highlights the key metrics of awareness, contact, conversion, engagement, conversion and retention efficiency.

3.1 Awareness Efficiency

The first stage of the model, awareness efficiency, represents the effectiveness of marketing endeavours to aware people of the ebusiness. Awareness efficiency is defined as total number of aware target audience divided by total number of target audience.

$$\text{Awareness Efficiency} = \frac{\text{number of total aware target audience}}{\text{number of target audience}} = \frac{Q_{1S}}{Q_0}$$

By the term awareness we mean marketing activities in middle layer mining structures that inform the target audience about the ebusiness products and services. These activities target the first stage of the EBAF analytics model which is illustrated at the left side of the fig. 2 and increase awareness efficiency. Damani and Damani (2007) illustrate and describe pros and cons of 21 different awareness channels including traditional TV commercials, on-demand/ IPTV commercials, radio, Internet radio, podcasts, traditional outdoor, interactive outdoor, newspapers, magazines, direct mail, public relation, ambient or guerrilla marketing and street graffiti, traditional online banner, behavioural media banner and rich media banner that can even allow full shopping within the banner, organic search engine optimization, pay per click, affiliates, email to Internet list, email to 3rd party list, SMS, RSS. Still there are additional channels including widget and gadget marketing, micro sites, and viral campaigns. These are good candidates to be optimized through data mining models and then being included in mid-level mining structures to provide more business intelligence.

3.2 Contact Efficiency

By the term contact, we mean marketing activities that are identified to be influencers to ease it for aware target audience to hit the website or generally contact the business. In an ecommerce case study, we define contact efficiency as total number of unique visitors divided by total number of aware target audience.

$$\text{Contact Efficiency} = \frac{\text{number of total unique visitsors}}{\text{number of total aware target audience}} = \frac{Q_{2S}}{Q_{1S}}$$

There are increasing numbers of channels that an interested consumer can contact the business. For example, usage of mobile phones has now extended from voice communications to the internet. An increase in extension of mobile internet technology and development of m-commerce applications has opened great opportunity for mobile service users. In Internet channel of contact, activities such as enhancing server speed and bandwidth, choosing suitable names for the site that can be easily guessed, using multiple names, affiliate programs and embedding hot links

in sponsored websites, banner ads on search engines, and organic search engine optimization are of most importance. An affiliate program is defined as a form of e-marketing that pays the affiliates for driving traffic to the advertiser or for subsequent transactions. By an affiliate program we usually mean a form of pure-commission selling. The affiliate website directs a visitor to a website or landing page.

Channel integration approach is essential for success of the business. That means from the customer's perspective, all contact channels are gates to the same place and a multi-channel retailing is just a single retail organisation that has multiple touch points, in the form of call-centre, physical store, mail, interactive TV, main website, web service, kiosk, and mobile. Although every channel has its own cost, but customers are willing to see similar prices across all channels. This enables them to research the product online and pick the items up in a local store. Some retailers have solved the problem by launching channel specific brands, so they can compete with increasing online competition without affecting retail store prices.

3.3 Engagement Efficiency

Many of the website visitors may leave it immediately after viewing the landing page. According to a report by Palmer (2010), a website on average has less than 10 seconds to capture the visitor's interest. Engagement efficiency is defined as total number of active unique visitors divided by total number of unique visitors.

$$\text{Engagement Efficiency} = \frac{\text{number of total active unique visitsors}}{\text{number of total unique visitsors}} = \frac{Q_{3S}}{Q_{2S}}$$

By the term Engagement, we mean identified activities to engage the unique visitors and prevent them from getting back. We define these engaged page viewers as active visitors. Although the operational definition of an active visit is to some extent dependent on the business model, the distinctive feature of it is some interaction between the surfer and the webpage that could be as simple as viewing the offers or querying a database.

There are some activities that may be useful to increase engagement efficiency, including landing page optimization, and site navigation optimization. A landing page is the point at which an Internet visitor lands on the website. The landing page can be part of the main corporate website. It may be the home page or even might be several layers deep within the website. They also can be a part of a microsite which is specifically designed for a single audience or purpose. Landing page optimization

includes upgrading web servers to increase bandwidth, keyword follow-through, multiple browsers and screen resolution testing, creating trust by illustrating press recognition and awards. Site navigation optimization activities include keeping navigation consistent, keeping navigation clear by using path indicator and history browser, prioritizing navigation links, and offering a variety of navigation themes to visitors.

Internal site search optimization is another way to increase engagement efficiency. While most visitors impatiently go straight to site search, a study by Palmer suggests that internal site search users convert 3 times better than users who don't use search, assuming their query returns relevant results (2010).

3.4 Conversion Efficiency

Conversion efficiency is defined as total number of actors divided by total number of active visitors. An actor is a visitor who performs the desired action.

$$\text{Conversion Efficiency} = \frac{\text{number of total actors}}{\text{number of total active visitors}} = \frac{Q_{4s}}{Q_{3s}}$$

Conversion stage deals with strategies and activities that are identified to be effective to persuade the active visitor to take the desired action and can be used in related mining structure. Some of the most important strategies to improve this coefficient are customer journey optimization, copy optimization, stock management, behaviour targeting segmentation, space management, product page optimization, and checkout process optimization.

In an ecommerce website, products cannot be touched, tasted, or tested, but there are tactics and strategies that can be implemented on the product page to increase visitor engagement and help to convert them into actors. Many companies think about website segmentation as a complex, sluggish, and time-consuming project, but it doesn’t always need to be. At its simplest form, the active visitors can be split up to first time and repeat. This can be realized by cookies. If the visitor is a repeat, we can promote the product page of new or complementary offerings or special promotions targeted toward existing customers. In a product page optimization, up-selling, down-selling, and cross-selling can be considered. Up-selling means selling higher priced products or services to the customer who is considering a purchase, instead of the one he wanted to purchase (Jain, 2010). Down-selling is used when the customer, for some reason, decides to back down from the purchase. In this case we can offer him a

cheaper product, which has higher chances of being accepted. The goal here is to acquire a customer. Even if we will not profit as much as possible right away, eventually it will increase the overall profit margin. Cross-selling is a technique in which the salesperson recognizes what a customer may need and makes suggestions or recommendations. Cross-selling can be as simple as the waiter asking the customer if he wants a salad to go with his main course. Accurate cross-sell product recommendations based on past or current purchases will increase conversion. Although cross-selling is usually used to increase the profits, but it can also be used to solidify the relationship with the client and widen the customer's reliance on the company as an assured resource and decrease the likelihood of the customer switching to a competitor. The database of items you provide or sell should be organized with an affinity link that identifies them as possible cross-sell items for another product (Lau, 2004). As another attempt to optimize product page, many ecommerce websites have implemented technologies to allow zooming and changing angles in the product page to help the customer feel a better shopping experience.

Checkout process optimization is of highest importance in this phase. Surprisingly, some ecommerce websites permit cross-selling and up-selling activities in checkout page. These activities should be avoided in this stage. Emphasizing on security, auto-detecting credit card type, user friendly credit card errors, security code explanation, disabling finalize order button immediately after first click to avoid double billing, shipping time estimates, and the ability to bookmark receipt page, produce better results in checkout process optimization.

Companies should seriously consider periodic A/B testing, multivariable testing, user testing, consultancy and expert usability reviews, cart abandonment analysis, pinch-point analysis, customer feedback reviews and online surveys to scrutinize their ecommerce website and proactively prevent any plunge in conversion efficiency.

3.5 Retention Efficiency

The last, but probably the most important stage in the model is customer retention. Customers are the key asset of any organization and companies should plan and employ a clear strategy for retaining them (Zineldin, 2006; Almotairi, 2008). Retention efficiency is defined as total number of clients divided by total number of actors.

$$\text{Retention Efficiency} = \frac{\text{number of clients}}{\text{number of total actors}} = \frac{Q_5}{Q_{4s}}$$

In recent years, commercial companies are putting much more emphasis on Electronic CRM (ECRM) as a tool for managing customer relationship and to increase customer satisfaction and loyalty (Azila, 2011; Chang, 2005; Donio, 2006, Khalifa, 2005). A mining model based on transactional segmentation might be significantly effective in the field. Social media communication strategies are also proved to be effective.
Assuming $Q_{0S} = Q_0$ and $Q_{5S} = Q_5$, An overall average efficiency index η_{AV}, which can be thought of as a summary of the process, can be defined:

$$\text{Average Efficiency} = \eta_{AV} = \frac{1}{5}\sum_{i=1}^{5}\frac{Q_{is}}{Q_{is\text{-}1}}$$

This overall average efficiency factor can be weighted. This happens when some points of waterfall model have more importance in the business model. For example visits to the website may be considered as the most important criterion of its success. So a weighted average efficiency index is defined by applying coefficient:

$$\text{Weighted Average Efficiency} = \eta_{WAV} = \frac{1}{5}\sum_{i=1}^{5}\frac{\mu_i Q_{iS}}{Q_{iS\text{-}1}}$$

4.0 Empirical Study

Sample data mining models including logistic regression, time series regression, association rule, naive Bayes classification, and clustering segmentation, are selected (Yang, 2006; Kozielski, 2009) to clarify and support the idea. While different algorithms can be used to perform specific business tasks, the challenge is to choose the most appropriate one. Each algorithm produces a different result, and some algorithms can produce more than one type of result. Although in a real situation, each layer of mining models may utilize various data mining algorithms for achieving better optimization, in this paper for simplicity, in each layer only the results of one specific algorithm is mentioned.

4.1 Logistic Regression Model

The first regression model evaluated in experimental study is Logistic regression. Logistic regression is a regression technique that is optimized for binary models in

which dependent variable refers only to two variables. Examples of these yes-no models can be expressed as response to questions like: Is the customer loyal to the business? Is the customer a high value customer? Will the customer buy this product? In these cases, when the dependent variable refers to two values, standard multiple regression cannot be used. In this algorithm, a transformation of the dependent variable is going under prediction. This transformation is called the logit transformation (Gorunescu, 2011). We mention the transformation as logit (p) and define its formula as: Logit (p) = Ln (p/ (1-p)) = Ln (p) – Ln (1-p). In this formula, p is the proportion of objects with a certain characteristic e.g. the probability for a customer to remain loyal to a company or brand. The logistic transformation of any number of z which like probabilities, always takes on values between zero and one, is given by the inverse-logit: Logistic (z) = $Logit^{-1}$(z) = exp (z)/ (1+exp (z)) = 1/ (1+ exp(-z)). As fig. 2 shows, the value of the transform rapidly approaches to zero or one, making the transform suitable to be used in binary predictions.

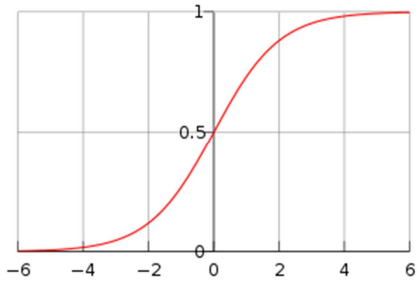
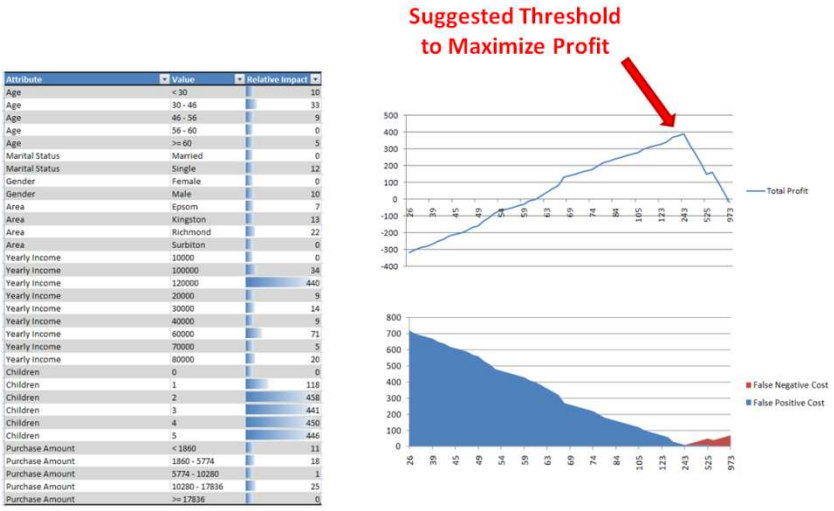


Figure 3. Logistic Transform

Now a Multiple Linear Logistic Regression can be defined as: Logit (p) = b0 + b1*X1 + b2*X2 + … + bn*Xn. The probability p of outcome variable can be derived by the equation: $p = \frac{1}{1+ e^{-(b0+b1*X1+b2*X2+\cdots+bn*Xn)}}$. As an example, the researcher evaluates the algorithm in optimizing a loyalty program. To minimize the cost, logistic regression algorithm is used to predict the outcome of the activity for each customer. Minimum requirement for determining threshold include false positive cost, false negative cost, true positive profit, and true negative profit. These values help the algorithm to find out the best threshold to maximize profit, as shown in fig. 4.



Attribute	Value	Relative Impact
Age	< 30	10
Age	30 - 46	33
Age	46 - 56	9
Age	56 - 60	0
Age	>= 60	5
Marital Status	Married	0
Marital Status	Single	12
Gender	Female	0
Gender	Male	10
Area	Epsom	7
Area	Kingston	13
Area	Richmond	22
Area	Surbiton	0
Yearly Income	10000	0
Yearly Income	100000	34
Yearly Income	120000	440
Yearly Income	20000	9
Yearly Income	30000	14
Yearly Income	40000	9
Yearly Income	60000	71
Yearly Income	70000	5
Yearly Income	80000	20
Children	0	0
Children	1	118
Children	2	458
Children	3	441
Children	4	450
Children	5	446
Purchase Amount	< 1860	11
Purchase Amount	1860 - 5774	18
Purchase Amount	5774 - 10280	1
Purchase Amount	10280 - 17836	25
Purchase Amount	>= 17836	0

Figure 4. Binary Prediction by Logistic Regression Algorithm

This threshold later will be used to evaluate next customers and predict about them. The figure also shows a table of relative impact of each influencer factor in prediction report. This table is just for getting a better understanding about the process. A small relative impact like having only one child shows that the related factors has only marginal effect on prediction. A zero relative impact like having no children shows that the factor does not affect the outcome. A factor with big relative impact like having more than one child is a strong indicator that the influencer is effective. The threshold obtained from profit diagram then can be used in companion with a calculator form for each object. Each factor in the form has a point assigned to it which has derived from the analysis by logistic regression algorithm. If the total score reaches the threshold, the marketing activity would be successful.

4.2 Time Series Regression Model

Integrating time dimension into other EBAF components helps to create a much more efficient business framework. In this experimental study, Demand Planning process is being considered using Time Series algorithm. This algorithm is another type of regression algorithms that is optimized for the forecasting of continuous variables such as sales, profits, temperatures, product values, stock prices and so on. A time series is a sequence of recorded values at regular intervals, such as yearly, monthly, weekly, daily, and hourly. Time series algorithms involves in complicated difference equation (Keogh, 2005; Chiu, 2003).

There are two main goals of the time series analysis. First goal is forecasting possible future values, starting from the observations already known data. Second goal of

analysis is identifying the nature of the phenomenon to obtain insights into the mechanism that generates the time series. A time series model can be decomposed into trend component T(t), cyclical component C(t), seasonal component S(t), and random error of irregular component E(t) (Gorunescu, 2011). The trend is a linear or non-linear component, and does not repeat within the time range. The Seasonal component repeats itself in systematic intervals over time. So an additive model of mentioned components can be expressed by this equation: Y(t)= T(t) * C(t) + S(t) + E(t). The next figure shows a mix of resulted reports in which a visual representation of the forecasted values (the dotted lines) is displayed, as well as the highlighted forecasted values that are appended back to the original data source.

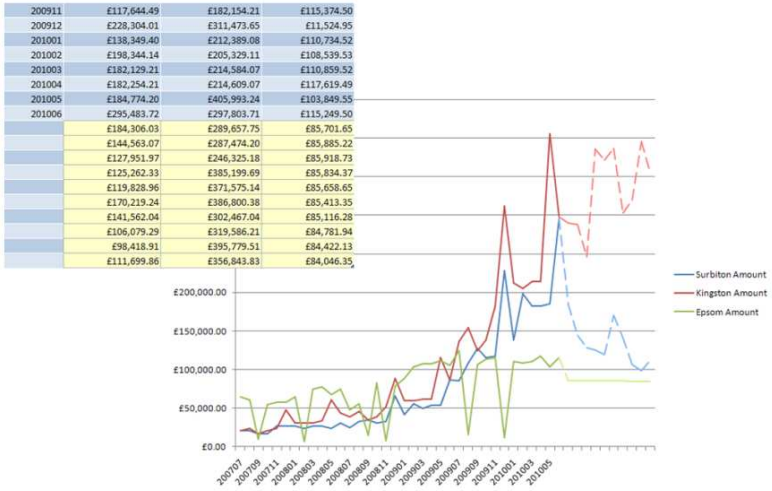


Figure 5. Forecasting Results using Time Series Regression Algorithm

4.3 Association Rule Model

As discussed before, different data mining models can be used in each mining structure layer to provide more enhancements. For example we used association rule algorithm in EBAF first layer in two important recommendation engine and market basket analysis subjects. Association rule algorithms can discover correlations between different attributes in a dataset and locate desired itemsets and rules (Wang, 2010). Itemset is a group of items in a case or in a transaction. An association model consists of a series of itemsets and the rules in the X→Y form that describe the mutual relationship and interdependence between items, within the cases.

4.4 Naive Bayes Classification Model

In the second layer of enhancement, the dataset includes items that have been previously optimized as targets. In a case study a classification model is used for this layer. Decision tree algorithm and Naive Bayes algorithm are most important

classification algorithms. Given an object with attributes {A1, A2... An}, we wish to classify it in class C. According to Naïve Base algorithm, the classification is correct when the conditional probability Pr(Ck|A1, A2... An) reaches its maximum among other classes. Based on Bayesian Theorem, we have: Pr(Cj|A1, A2... An) = Pr(A1, A2, ..., An | Cj) * Pr(Cj) / Pr(A1, A2, ..., An). Assuming mutual independence of attributes for a given class C, results: Pr(A1, A2, ..., An |Cj) = Pr(A1|Cj) * Pr(A2|Cj) * ... * Pr(An|Cj).

By estimating all the probabilities Pr (Ai|Cj) for all attributes Ai and classes Cj, a new object can be classified to class Ck if the probability associated to it is maximized among the other classes:

$$\mathbf{Pr(Ck)} * \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{P\,r(Ai|Ck)}$$

One of main advantage of this algorithm is that it can easily handle irrelevant input attributes. Other advantages include robustness to noise and missing values (Gorunescu, 2011).

The input attributes through the case study are components of a low level of conversion model components e.g. Facebook marketing as a social media marketing, Google Adwords as a kind of PPC, and organic SEO. Search engines are the most common tools that new visitors use to find a company's website. PPC advertising refers to a marketing activity in which a company pays to a website owner if his visitor actually clicks on the company's advertisement and is redirected to the ad link. Email marketing is divided to two third-part and in-house solutions. Considering these influencers, a dataset for a second layer mining structure in an experimental study in EBAF is shown in fig. 6:

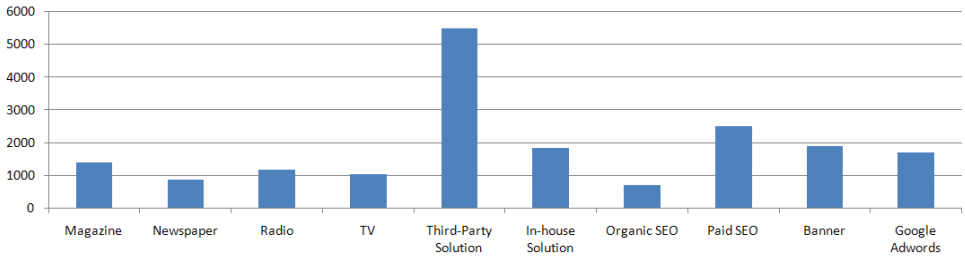
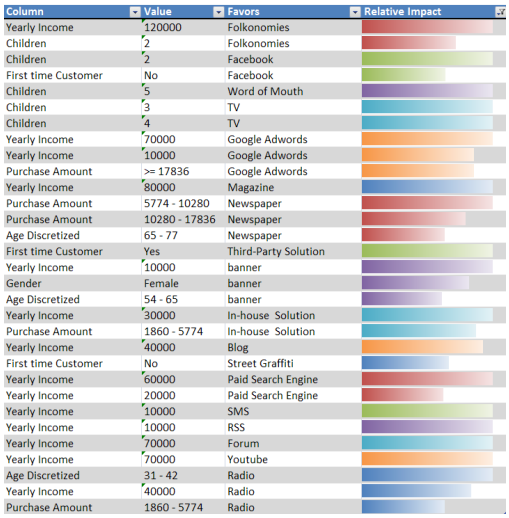


Figure 6. A Dataset for Second Layer of Mining Structures

When the algorithm analyse the data for identifying key influencers, it creates predictions that correlates each column of data with the specified outcome, and then uses the confidence score for the predictions to identify the factors that are the most

influential in producing the targeted outcome (Zhang, 2007). ID columns or other columns that have a lot of unique values are not considered in building the related mining structure, but as Fig. 7 shows, leaf level data like customer's demographic, geographic, and transactional data can be analysed.



Column	Value	Favors	Relative Impact
Yearly Income	120000	Folkonomies	
Children	2	Folkonomies	
Children	2	Facebook	
First time Customer	No	Facebook	
Children	5	Word of Mouth	
Children	3	TV	
Children	4	TV	
Yearly Income	70000	Google Adwords	
Yearly Income	10000	Google Adwords	
Purchase Amount	>= 17836	Google Adwords	
Yearly Income	80000	Magazine	
Purchase Amount	5774 - 10280	Newspaper	
Purchase Amount	10280 - 17836	Newspaper	
Age Discretized	65 - 77	Newspaper	
First time Customer	Yes	Third-Party Solution	
Yearly Income	10000	banner	
Gender	Female	banner	
Age Discretized	54 - 65	banner	
Yearly Income	30000	In-house Solution	
Purchase Amount	1860 - 5774	In-house Solution	
Yearly Income	40000	Blog	
First time Customer	No	Street Graffiti	
Yearly Income	60000	Paid Search Engine	
Yearly Income	20000	Paid Search Engine	
Yearly Income	10000	SMS	
Yearly Income	10000	RSS	
Yearly Income	70000	Forum	
Yearly Income	70000	Youtube	
Age Discretized	31 - 42	Radio	
Yearly Income	40000	Radio	
Purchase Amount	1860 - 5774	Radio	

Figure 7. Key Influencers Identification by Naive Bayes Algorithm

In a Naive Bayes model, all columns must be either discrete columns or be discretized during data preparing process. That is because the algorithm cannot use continuous columns as input and cannot predict continuous values. If the input column contains continuous numeric values, EBAF segments the numeric values into buckets. The number of buckets to be generated is calculated by using the following rule:

Number of buckets = sqrt (Number of distinct values of data in the column)

4.5 Clustering Segmentation Model

Third layer includes components that previously enhanced in second layer of enhancement. As an example the dataset can contains information about customers' main awareness and contact channels. We choose segmentation technique as the case study in this layer (Jin, 2006). During developing EBAF conversion model, the fact that segmentation is extremely valuable in all the phases was revealed. It was also discovered that any situation may need its special kind of segmentation strategy. Although segmentation, as a whole, is necessary for a successful awareness campaign, transactional segmentation, as an example, may not be helpful in that area but definitely is very useful in increasing retention efficiency of a website. Segmentation algorithms split data into groups, or clusters, of items that have similar properties.

Before applying the clustering algorithm as the main segmentation technique, it is useful to explore the dataset.

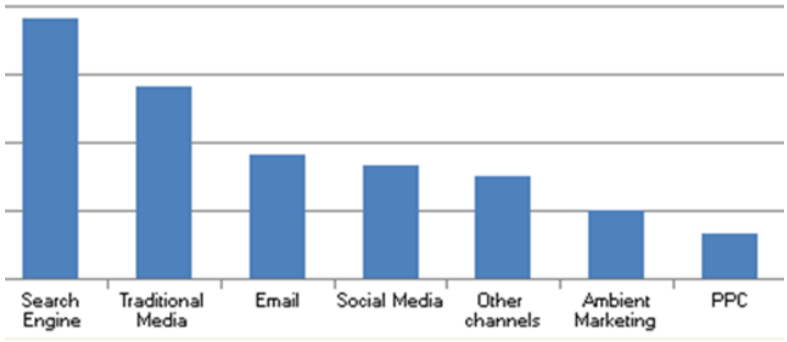


Figure 8. A Dataset for Third Layer of Mining Structures

Customers' demographic data, geographic data, behavioural data, transactional data, and other leaf level data can be considered in the clustering model, but attributes like ID, first name, last name, and email address should be excluded from mining structure because they have distinct values and do not have any effect on analysis. Initial results of applying algorithm shows the clusters and also shows that which object belong to which cluster. A category characteristics result reveals valuable details about the similarities in each category (Kogs, 2006). The results also shows the relative importance that indicates how important the attribute and value pair is as a distinguishing factor for the category. We would then explore the generated customer segments in further detail to better understand the overall attributes of customers who belong to this category (Chen, 2007; Gorunescu, 2011). For this ecommerce case study, Fig. 9 shows details about the cluster that represents most valuable customers. This cluster was selected and scrutinized with highest attention because finding out about characteristics of this cluster helps in improving marketing campaigns and to enhance the business.

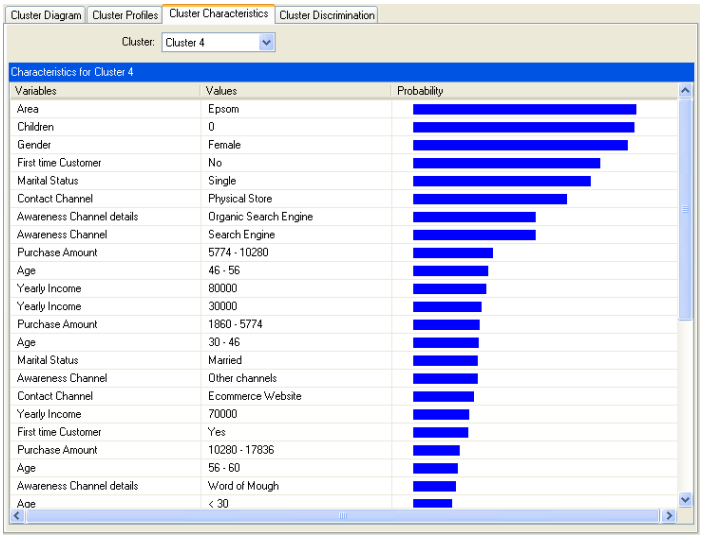


Figure 9. Cluster Characteristics for the cluster that represents most valuable customers

Segmentation algorithms, as well as other algorithms in this research, have been widely used in business applications to support decision-making. It is important to know that segments change over time as well as customers tend to change behaviours. Therefore, the algorithms should be recalculated on a more recent data source to check for correctness.

5.0 Conclusion

It appears from the preceding discussions and experimentations that the proposed multilayer data mining approach to an ebusiness framework may increase overall amount of business intelligence that an enterprise can gain. The concept contains a new methodology and its associated mining structures and mining models. The paper used this novel methodology and introduced an optimized framework called EBAF, to provide intelligence for SMEs and help them to gain competitive advantages. To support the theory, an experimental study consisting of various algorithms applying on different mining structure layers presented to provide a better understanding of the concept and to be a proof of usability of the new methodology. The next step of this research is planned to integrate the methodology into multidimensional data and cube structures and also deal with fast-generated data streams to support real time decision making.

References

Almotairi, M. (2008) "*CRM Success Factors Taxonomy*," in Proc. of European and Mediterranean Conference on Information Systems, pp. 29-35.

Azila, N. and M. Noor, (2011) "*Electronic customer relationship management performance: Its impact on loyalty from customers' perspectives*,"

International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 1, No. 2, June.

Bertino, E., Fovino, I. N., Provenza, L. P. (2005) "*A framework for evaluating privacy preserving data mining algorithms*," Data Mining Knowledge Discovery, vol. 11, pp. 121–154.

Chang, T.M., Liao, L.L., and Hsiao, W.F. (2005) "*An Empirical Study on the e-CRM Performance Influence Model for Service Sectors in Taiwan*," in Proc. IEEE International Conference on e-Technology, e-Commerce and e-Service, pp. 240-245.

Chen, J. R. (2007) "*Making clustering in delay-vector space meaningful*," Knowledge and Information Systems, vol. 11, no. 3, pp. 369-385.

Chiu, B., Keogh, E., Lonardi, S. (2003) "*Probabilistic discovery of time series motifs*," in Proc. of Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D.C., 493–498.

Dai, B. R. and Chiang, L. H. (2010) "*Hiding frequent patterns in the updated database*," in Proc. of International Conference on Information Science and Application, Seoul, South Korea;, pp. 1–8.

Damani R. and Damani, C. (2007) Ecommerce 2.0: The Evolution of Ecommerce, London, United Kingdom: Imanco plc.

Dimitriades, Z.S. (2006) "*Customer Satisfaction, Loyalty and Commitment in Service Oganizations: Some Evidence from Greece*," Management Research News, vol. 29, no. 12, pp. 782-800.

Divanis, A. G. and Verykios, V. S. (2009) "*Exact knowledge hiding through database extension*," IEEE Trans. Knowledge Data Eng., vol. 21, pp. 699–713.

Donio, J. P., Massari, and G. Passiante, (2006) "*Customer Satisfaction and Loyalty in a Digital Environment: An Empirical Test*," Journal of Consumer Marketing, vol. 23, no. 7, pp. 445-457.

Fernandez, M. T. (2011) "*Business strategy model*," International Journal of Innovation, Management and Technology, Vol. 2, No. 4, pp. 301-308.

Gaber, M. M. (2012) "*Advances in data stream mining*," WIREs Data Mining and Knowledge Discovery, vol. 2, pp. 79-85.

Gorunescu, F. (2011) "Data mining: Concepts, Models and Techniques," Berlin, Springer-Verlag LLC.

Holland P. C. and Naude, P. (2004) "*The metamorphosis of marketing into an information-handling problem*," The Journal of Business & Industrial Marketing, vol. 19, no. 3, pp. 167-178.

Jain, M. K., Dalela, A. K., and Tiwari, S. K. (2010) "*Fuzzy Mathematical Model for Up-Sell Solution*," International Journal of Innovation, Management and Technology, Vol. 1, No. 1.

Jayachandran, S., Sharma, S., Kaufman P., and Raman, P. (2005) "*The Role of Relational Information Processes and Technology Use in Customer Relationship Management*," Journal of Marketing, vol. 69, pp. 177-192.

Jin, R., Goswami, A., and Agrawal, G. (2006) "*Fast and exact out-of-core and distributed k-means clustering*," Knowledge and Information Systems, vol. 10, no. 1, pp. 17-40.

Kar, A. K., Pani, A. K., and Kumar S. (2010) "*A Study On Using Business Intelligence For Improving Marketing Efforts*," Business Intelligence Journal, vol. 3, no, 2, pp. 141-150.

Keogh, E., Lin, J., Fu, A. (2005) "*HOT SAX: efficiently finding the most unusual time series subsequence*," in Proc. of the 5th IEEE International Conference on Data Mining (ICDM 2005), Houston, TX, pp. 226– 233.

Khalifa, M. and Shen, N. (2005) "*Effects of Electronic Customer Relationship Management on Customer Satisfaction: A Temporal Model*," in Proc. of 38th Annual Hawaii International Conference on System Sciences, pp. 171-178.

Koga, H., Ishibashi, T., and Watanabe, T. (2006) "*Fast agglomerative hierarchical clustering algorithm using Locality-Sensitive Hashing*," Knowledge and Information Systems, vol. 12, no. 1, pp. 25-53.

Kotler P. and Keller K. L. (2006) Marketing Management, 12th ed. New York: Prentice hall.

Kozielski, S. And Wrembel, R. (2009) "New Trends in Data Warehousing and Data Analysis," Berlin, Springer-Verlag LLC.

Lau H. L. and K. Chow, (2004) "*A database approach to cross selling in the banking industry: Practices, strategies and challenges*," The Journal of Database Marketing, vol. 11, no. 3.

Palmer, J. (2010) Ecommerce Roadmap, Best Practices of Today's Successful Ecommerce Sites, New York: Palmer Web Marketing, LLC.

Rahbarinia, B., Pedram M. M., Arabnia H.R., and Alavi, Z. (2010) "*A multi-objective scheme to hide sequential patterns*," in Proc. of International Conference on Computer and Automation Engineering, Singapore, pp. 153–158.

Shearer, C. (2000) "*The CRISP-DM model: The new blueprint for data mining*," Journal of Data Warehousing, vol. 5, pp. 13–22.

Wang, J., Hu, X., Hollister, Zhu, K. D. (2008) "*A comparison and scenario analysis of leading data mining software*," International Journal of Knowledge Manage, vol. 4, pp. 17–34.

Wang, P. (2010) "*Research on privacy preserving association rule mining a survey*," in Proc. of IEEE International Conference on Information Management and Engineering, Chengdu, China, pp. 194–198.

Yang, Q., and Wu, X. (2006) "*10 challenging problems in data mining research*," International Journal of Information Technology and Decision Making, vol. 5, no. 4, pp. 597–604.

Zhang, J. D., Kang, K., and Silvescu, A. (2007) "*Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data*," Knowledge and Information Systems, vol. 9, no. 2, pp. 157-179.

Zineldin, M. (2006) "*The Royalty of Loyalty: CRM, Quality and Retention*," Journal of Consumer Marketing, vol. 23, no. 7, pp. 430-437.