Predictive Data Mining Models for Marketing Performance Monitoring in Global Markets

Valerio Veglio[1] (University of Milano-Bicocca, Italy)
Fabio Del Bo (University of Pavia, Italy)

Abstract

Global companies have realized that the customer knowledge contains in marketing databases available to businesses is the key to predict accurate and punctual marketing performance in today's global business. But, much of this useful knowledge is hidden and untapped. Appropriate data mining tools, which are good at extracting and identifying valuable information and knowledge from huge databases, is one of the best supporting approach to make different marketing decision. In fact, analysing and understanding in advance the customer behaviour can represent the main corporation's strength to plan adequate marketing forecasting.

A data mining methodology is used to accomplish the research goals. This research demonstrates as General Linear Models (GLMs) such as Logistic Regression Models are one of the main tools able to set accurate marketing performance. Special focus is paid on the estimation of the probability of customer conversion in order to minimize the risk of churn within global companies.

Key words: predict data mining models, logistic regression, churn management, marketing performance.

Introduction

The main challenge for global companies is to identify the best models and methods able to forecast precise marketing performance in order to increase companies' profitability in today competitive landscape. An essential part of managing any organization is planning for the future. Nowadays, the long-run success of global firms is closely related to how well management is able to anticipate the future and draw up appropriate marketing strategies. Good judgment, intuition, and an awareness of the state of the economy may give a manager a rough idea or "feeling" of what is likely to happen in the future.

According to Daniel Kahneman (2002) it is obvious as qualitative approaches are not enough for predicting accurate and punctual business performance. In addition, Sato (2000) observes that the data mining analysis differs from the statistical data analysis.

Statisticians use sample observations to study the population parameters by estimation, testing and predictions with the main risk to estimate inaccurate business performance. As consequence, business intelligence tools are one of the main priorities for the Chief Information Office Director in every industry.

Data mining analysis has become an astonishing approach is so far the meaningful knowledge is often hidden in enormous databases and most traditional statistical methods could fail to uncover such knowledge. But, due to, a limited knowledge in this field, global corporations in order to forecast business performance fall into huge traps obtaining catastrophic results. Businesses often consider mathematical and statistical models too complex, inaccurate, expensive and very hard to elaborate the final results. On the contrary some predictive data mining models are simple to deduce and really accurate to plan future trends.

Bueren, Schierholz, Kolbe and Brenner (2004) define the knowledge as a strategic intangible resource at the base of competitive advantages. Besides, the most important type of knowledge would be appearing to be customer knowledge. In fact, an efficient utilization of customer knowledge determines the development of global corporations. This is particularly true in marketing area because of the proliferation of e-customer data

[1] Corresponding Author. Valerio Veglio is a PhD Student in Marketing and Management at Milano-Bicocca University.

collected in huge databases. Global companies have acknowledged that their marketing strategies should focus on identifying those customers who are likely to churn (Hadden, Tiwari, Roi and Ruta 2005).

In literature, churn management is defined as a set of techniques that enable firms to keep their profitable customers and it aims at increasing customer loyalty (Lejeune 2001). Additionally, enterprises can be identifying two groups of churners: voluntary or attrition and non-voluntary or forced churn. Non-voluntary churners are easier to detect because are the customers who have had their service or product withdrawn by the company. Instead, voluntary is more difficult to determinate because this type of churn occurs when customers make a conscious decision to terminate their service with the company (Hadden, Tiwary, Roy and Ruta 2005). This research proposes an extension of the churn definition focusing the attention on the concept of voluntary churners in digital contests. In this case a churner is a potential customer that does not buy the product or the service.

Nowadays, one of the main tools able to help marketers in the mentioned approach is data mining (Khak Abi and Glolamain 2010). Prediction of behaviour, customer value, customer satisfaction and customer loyalty are example of some of the information that can be extracted from the data that should already be stored within a company's database. However, to perform such a complex analysis of the information is necessary to either purchase commercial software or implement a solution based on one of the many data mining techniques that have been developed for this purpose.

The key goal for global firms is to discover knowledge in huge database and to make sense out of the data in order to improve the probability of customer conversion and decries the number of churner.

The purpose of this article is to demonstrate the strong link between data mining and development of accurate marketing strategies within global companies in order to maximize the probability of customer conversion and minimize the number of churners. Special attention is paid on GLMs in particular on the Logistic Regression Model because of it seems the best tool able to predict the main marketing and sale activities generated by prospects before to get a customer status. The Criteria based on the Loss Functions such as the Confusion Matrix[2] and the Receiver Operating Characteristic (ROC) curve[3] have been used to evaluate the goodness of fit of the predictive model developed.

This paper is further structured as follows. Firstly, despite the current lack in literature, a brief theoretical framework on the main difference between statistical data analysis and data mining analysis is presented in this paper. Particular focus is paid on the data mining process and competitive knowledge discover in database (Section 2). We then explain the main features of the database analysed and the research methodology developed in this research (Section 3). After, we show the main results of the data mining analysis developed (Section 4). Finally, conclusions and proposes suggestion for future research conclude the article (Section 5).

Theoretical Framework

According to Sato (2000) is possible to consider the main features that differentiate both statistical data analysis and data mining analysis. First, data mining analysis is governed by the need to uncover, in a timely manner, emerging trends, whereas statistical data analysis is related to historical fact and it is based on observed data. Second, statistical data analysis focused on finding and explaining the major source of variation in the data instead data mining analysis endeavours to discover, not the obvious source of variation, but rather the meaningful, although currently overlooked information. Third, statistical data analysis

[2] The Confusion Matrix contains the number of the elements that have been correctly or incorrectly classified for each class. The main diagonal shows the number of observations that have been correctly classified for each class; the off-diagonal elements indicate the number of observations that have been incorrectly classified. The disadvantage of this measure is that it is not very robust concerning the choice of the cut off value in the "a posteriori" probabilities (Baesens et al. 2002 and Giudici 2003).

[3] The ROC curve is obtaining by graphing, for any cut-off value, the false positive (1-specificity: proportion of non-events forecasted as event) on the horizontal axis and the sensitivity (proportion of events forecasted as such) on the vertical axis. Each point on the curve corresponds to a particular cut-off point, trading off sensitivity and specificity.

manages data related to a specific research questions while data mining analysis explores data collected for different purposes other than the aim of the research. Giudici (2003) observes that data mining is not just about analysing data; it is a much more complex process where data analysis is just one the aspect. Turban, Aronson, Liang and Sharda (2008, p. 305) define data mining as "the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge in databases". "To apply a data mining methodology means following and integrated methodological process that involves translating the business needs into a problem which has to be analysed, retrieving the databases needed to carry out the analysis, and applying a statistical technique implemented in a computer algorithm with the final aim of achieving important results for taking a strategic decision" (Giudici 2003, p. 6). The strategic decisions will itself create new measurement needs and consequently new business needs, setting off what has been called "the virtuous circle of knowledge" induced by data mining (Berry and Linoff 2011). Exploring databases for hidden competitive information through sophisticated data mining tools is the matter of concern for the business organizations. Effective use of information technology is the first step to achieve the data sufficiency that helps the companies in making time to time business decisions. Rapid advances in information and sensor technologies along with the availability of database management technologies combined with the breakthroughs in computing technologies, computational methods and processing speeds, have opened the floodgates to data dictated models and pattern matching (Fayyad and Uthurusamy 2002; Hand et al. 2001). Today, it is clear as the use of sophisticated and computationally intensive analytical methods are expected to become even more common place with recent research breakthroughs in computational methods and their commercialization by leading vendors (Grossman et al. 2002). "In the recent years computing environment search engines plays a crucial role in digging out the hidden facts that are processed through well define decision support system" (Murthy 2010, p. 2). These decision systems utilize available information and through data mining models and human interaction to provide a decision making tools to analysis the data and information. Decision support system and data mining models combined together can be appearing as "the spectrum of analytical information technologies" providing a unifying platform in order to obtain an optimal combination of data dictated and human driven analytics. In other words, data mining process cover many activities: from the identification of business problem to the visualization of the results up to the interpretation and evaluation of the findings. In contrast, statistics is only the science of learning from the data. In fact the main goal of statistics is just to interpret the data and not to understand the causes of the results obtained. In terms of a data description and inference, about the parameters under study, it includes everything from the data collection to the data processing. "The intersection between statistics and data mining enables the user to make effective utilization of the available data, to gain a better understating of the past, and predict the future through better decision making" (Murthy 2010, p. 2). Analysing this point of view emerge that statistical data analysis and data mining analysis are complementary. The statistical data analysis emphasizes and removes the major part of data variation before that data mining analysis is used. This explains why the data warehousing tool not only stores data but also contains and executes some statistical analysis programs. In addition, statistical theory and methods are central to the classification, clustering, and modelling issues involved in most data mining applications especially for driving the quality of the variables used and to test final results. Statistical data mining applications, which nearly always involve making use of information draw from multiple databases, are particularly subject to limitations of data and methods. Also, the procedures used to combine the individual data sources may themselves introduce error and uncertainly. The different features of the datasets involved can give rise to multiple sources of error that may interact with one another in unknown ways. The underlying sources of error that may include, among others, coverage and content errors, the possibly different time references of individual datasets, and the additionally uncertainly introduced when some of the datasets are based on samples. These sources of error and uncertainly emphasize the importance of ensuring that the necessary statistical expertise is involved in data mining application. Data mining tools and applications are proved to be an asset to the organization. In fact, according to Murthy (2010) global companies using the data mining techniques and tools to answer at the following questions:

- Which segment of population is most likely to respond to a particular advertising campaign?
- How many clusters or bubbles demand it is possible to find in huge databases?
- How many products are bought at the same time and in contemporary?
- How many customers could become churners in a given time?
- How many customers are in an insolvency state in a given period?

- What is the level of satisfaction related to a given product or service?
- What are the ideal conditions for launching a new product or service?

On the other hand, statistics can helps greatly in this process by helping to answer several important questions about the data:

- What hidden patterns are there in the databases?
- What is the probability that an event will occur?
- What patterns are significant?
- What is the high level summary of the data that gives some idea of what is contained in the database?

To sum up, it is clear that statistic and data mining are complementary. In fact data mining is useful with large quantities of data while statistical inference when there are small quantities of data. Indeed, the domain knowledge is useful in either case.

In global markets news generations of statistical models are required to analyse vast amount of data collected in huge databases. The main problem for global companies is to understand which raw data could be contains competitive information thus relevant knowledge for getting strategic decisions. But, often this potential competitive knowledge remains stored in dormant databases without to create new competitive customer value for enterprises. In fact "databases are frequently a dormant potential resource that, tapped, can yield substantial benefits" (Fayyad, Piatetsky-Shapiro and Smyth 1996, p. 28). Before the advent of globalization, due to the limited volume of data, for turning raw data into knowledge, decision makers use traditional methods based on manual analysis and interpretations. This approach is too much slowing, expensive, and highly subjective. Obviously, today this approach makes no sense. "Databases continue to grow both in terms of the number N of records, or objects, and the number d of fields, or attributes, per object" (Fayyad, Piatetsky-Shapiro and Smyth 1996, p. 28). As a results of this, the knowledge discovery in database process and the data mining methodology are always more important both in managerial and academic contest in order to extract competitive knowledge from enormous databases.

It is really important to discern data mining from online analytical processing. This latter are quite different from data mining because they provide only a really good view of what is happening but cannot predict what will happen in the future or why it is happening (K Pal 2011). For this reason, data mining is considered an emergency methodology that has made a revolutionary change in the information society (K Pal 2011). In literature, sometimes the knowledge discovery in database and the data mining have the same meaning. The terms of data mining and knowledge discovery in database could be interchangeably (Fayyad 1996; K Pal 2011). Data Mining could be seen as an integration of more disciplines as statistics, computer science and artificial intelligence, machine learning and data base management (Murthy 2010).

According to Gartner Group, world leading information technology research and advisory company, and Larose (2010), data mining is the process of discovering meaningful new relations, models and trends through the analysis of a large amount of data collected in huge databases, using statistical and mathematical techniques, models and methods. In other words, the main role of data mining is to discover, in advance, unknown relations in enormous databases in order to predict actions, behaviours and outcomes related to the market players. In global scenario data mining is one of the main methodologies able to forecasting with high accuracy business performance. In addition, "data mining is the process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large database" (Turban, Aronson, Liang and Sharda 2007, p.305). Data mining models are helpful in the decision making process because they surmount the limits of traditional statistical models. In fact, before to get a strategic business decision, data miners evaluate the quantitative results in a qualitative way (opinion mining or expert judgment). Broadly, "data mining could be useful to answer the queries on: forecasting with regard to what may happen in the future, classifying things into groups by recognizing patterns, associating similar events that are likely to occur together, clustering the peoples into group based on their attributes and making the sequence what events are likely to lead to whom" (K Pal 2011, p.10). Finally, data mining is based on the multiplicity of applications

especially in marketing area such as Market-Basket[4], Web-Click-Stream[5], Customer Profiling[6], Retention/Churn[7] Analysis (Figini and Giudici 2009).

Data and Methodology

The dataset analysed was provided by a global company that offers digital data driven marketing solutions across all interactive channels: digital, direct response, relationship based media and design. Currently this company operates in most different countries across Europe, North America, South America, Asia, Africa and Australia. Unfortunately, for privacy reasons, it is impossible gives other details about the company.
A data mining analysis is used to accomplish the research goals. The original whole database contained more than 1.463.199 prospect and 41 variables related to them. In order to ensure maximum accuracy of the results the database has not been sampled. The dataset analysed is composed both quantitative and qualitative variables. The qualitative variables (for instance: the name of the advertiser, the type of banner, the timestamp for the activity on the advertiser's website, the keywords categories, etc.) have been treated as a quantitative because they were categorized into groups.

Given the huge size of the database, before to aggregate the data by "user id" (code that identifies each potential customer), it has been necessary to perform a preliminary screening of the variables for detecting possible outliers and anomalies in the data. In our case at each row in the spreadsheet correspond to a potential customer. After the aggregation the database contains 1.463.199 potential customers and 276 quantitative variables. The number of variables increases dramatically because of the creation of dummy variables. Before starting the exploratory analysis it was fundamental a second preliminary screening with the aim to eliminate new irrelevant and redundant variables through correlation index and multiple linear regressions. After this second screening, the number of variables decries at 27. The quantitative nature of the research aims at conducting rigorous theory testing using predictive data mining models.

The final dataset contains 1.463.199 potential customers and 27 quantitative variable related to their purchase behaviour. The target variable is dichotomous thus it assumes only two values: 0 when the potential customer does not buy the service (churner) and 1 when the potential customer buys the service. Given the form of the target variable a logistic regression model based on "Enter Method" with a significance level of p-value has been implemented in this research. P-value is the estimate probability of rejecting the null hypothesis of a study question when hypothesis is true (Lehmann and Romano 2005). With the "Enter Method" decision makers inserts all variable at the same time. In fact, the results of the analysis are based both on mathematical and statistical algorithms and human judgment. The main reason of this methodological choice has been the dimension of the dataset in terms of variables and the nature of the dependent variable. If the target variable is quantitative continuous a multiple regression model must be developed to accomplish the research goals (Giudici 2003; Berry and Linoff 2011).
Before showing the main results of this research a brief description of the variables analysed in this research is provided by the following table.

Table 1 Database Description

Variables	Description
Dynamic Click	Banner Moving. This click is able to generate the click through rate.
Standard Click	Fixed Banner.

[4] The Market-Basket Analysis identifies which products are jointly purchased with others in order to improve the layout of goods on the shelves and to increase, through promotions on items associated with them, sales of determinate product or group of products.
[5] The main aim of this application is to understand the dynamics of the web navigation for improving the navigation on the site and, in e-commerce sites, to accelerate the paths that promote purchases.
[6] The Customer Profiling Analysis allows creating homogenous profiles of customers, based on their past behavior, in order to involve them in target promotions.
[7] This application identifies homogeneous group of customers, in terms of behavior and personal characteristics, in order to retain existing customers and to identify new potential customers.

Avgpos_best5	Number of times that the site name appears in the top best 5 in the search engine.
Brandfl	The potential customer digits one of the brand company in the search engine.
Impclkhour_13	Total number of impressions and/or click at 1pm.
Impclkhour_14	Total number of impressions and/or click at 2pm.
Impclkhour_15	Total number of impressions and/or click at 3pm.
Impclkhour_16	Total number of impressions and/or click at 4pm.
Number of Impression or Click 2010	Total number of impressions and/or click 2010.
Number or Impression or Click 2011	Total number of impressions and/or click 2011.
Matchtype_Broad	The potential customer digits a similar name of keywords on the search engine.
Matchtype_Exact	The potential customer digits the exact keywords on the search engine.
Number of Purchases 2010	Total purchases in 2010.
Number of Purchases 2011	Total purchases in 2011.
Sales Quantity	Total number of sales.
Search Engine on Google	Number of times that a potential customer searches a keyword on Google.
Search Engine	It represents an exposure that is a search click.
Affiliate Marketing: MCUK	Number of times that the potential customer is exposed to a specific banner (MCUK).
Affiliate Marketing: ADJUG6	Number of times that the potential customer is exposed to a specific banner (ADJUG6).
AffiliateWindow	Number of times that the potential customer is exposed to a specific banner (AffiliateWindow).
Affiliate Marketing: DRIVEPM	Number of times that the potential customer is exposed to a specific banner (DRIVEPM).
Affiliate Marketing: MCUKQUIDCO	Number of times that the potential customer is exposed to a specific banner (MCUKQUIDCO).
Affiliate Marketing: MCUKYAHOO	Number of times that the potential customer is exposed to a specific banner (MCUKYAHOO).
Avgcpc	Mean cost par click.
Avgctr	Mean click through rate.
Avgpos	The mean average position of a search term in.
Rank	Auto incrementing value representing all touch point of a user journey. Does reset on a conversion/activity. Maximum number of the activity subjected at the user.
Number of Purchases (Target Variable)	Dichotomous variables: 1 if potential customers buy the service on-line. 0 when potential customers not buy the service on-line.

Findings

First of all, due to the huge amount of data, we have developed a Bivariate Pearson Correlation Analysis to identify the relationship between the number of purchases and one or more variables collected in the dataset. This stage is really important because of it achieves to discover the variables related to the marketing activities thus to draw accurate predictive data mining model and strategies. Table 2 provides the main descriptive measures related to the data studied. Table 3 shows the values of the Pearson Correlation among the number of purchases and the variable “Rank”.

Table 2 Bivariate Pearson Correlation Analysis (Dependent Variable: Number of Purchases)

Variables	Mean	Std. Deviation	Correlation Values	Collinearity Situations	Recommendations
Avgcpc	.64	.63	-.07**		No predictive information
Avgctr	.00	.01	.29**		
Avgpos	1.39	.82	-.17**		
Rank	4.18	45.55	.07**	X	See Table 3
Dynamic Click	.03	.31	.62**		
Standard Click	4.07	31.45	.09**	X	No predictive information
Avgpos_best5	.01	.15	.38**		
Brandfl	.01	.14	.39**		
Impclkhour_13	.28	2.20	.10**		
Impclkhour_14	.30	2.26	.10**		
Impclkhour_15	.31	2.32	.10**		
Impclkhour_16	.30	2.30	.10**		
Number of Impression or Click 2010	2.35	25.51	.06**	X	Decomposition of the Number of Purchases variable
Number of Impression or Click 2011	1.80	19.98	.08**	X	Decomposition of the variable Number of Purchases
Matchtype_Broad	.00	.06	.16**		
Matchtype_Exact	.01	.13	.38**		
Sales Quantity	.02	.14	1.0**	X	Field valorised only for 12.816 Potential Customer
Number of Purchases 2010	.02	.13	.95**	X	Redundancy with the Number of Purchases variable
Number of Purchases 2011	.00	.04	.30**	X	Redundancy with the Number of Purchases variable
Search Engine on Google	.01	.16	.39**		
Search Engine	.01	.16	.40**	X	Redundancy with the Search Engine on Google variable
Affiliate Marketing: MCUK	.22	7.15	.09**		
Affiliate Marketing: ADJUG6	.41	7.14	.05**		
Affiliate Windows	.01	.14	.41**		
Affiliate Marketing: DRIVEPM	.29	8.72	.05**		
Affiliate Marketing: MCUKQUIDCO	.02	.27	.49**	X	Redundancy with the variable Dynamic Click
Affiliate Marketing: MCUKYAHOO	.25	6.67	.09**		
Number of Purchases	.02	.13	1		

**Correlation is significant at the 0.01 level

Table 2 shows some general descriptive statistics such as: arithmetic mean, standard deviation and Person Correlation. Particular attention must be paid on the Pearson Correlation Coefficient (PCC) because of it measures the strength and direction (decreasing or increasing, depending on the coefficient sign) of a linear relationship between two variables without identifying causes and effects (Ahlgren, Journeving, Rousseau 2003). From a statistical point of view only the variables with p-value < 0.005 are significant correlated to the target variables (Baum 2006). The value of the p-value represents a decreasing index of the reliability of a result (Moody 2009). In addition, the PCC provides information about the collinearity of the variables. An important remark is that in order to establish the variables to put in the logistic regression model, decision makers used both quantitative and qualitative approach. Sometimes, a qualitative approach or better human judgment overcomes the quantitate results. In other words, they used a data mining approach. Additionally, high and similar correlation values suggestion that some variables are redundant between them. For instance the variable "Sales Quantity" has the meaning of the variables "Number of Purchases". Also, the variables "Search Engine on Google" and "Search Engine" are redundant between them.

Table 3 Pearson Correlation Analysis (Dependent Variable: Rank)

Variables	Correlation Value
Avgcpc	-.00
Avgctr	.00
Avgpos	.01
Rank	1
Dynamic Click	.090**
Standard Click	.84**
Avgpos_best5	.06**
Brandfl	.058**
Impclkhour_13	.62**
Impclkhour_14	.56**
Impclkhour_15	.53**
Impclkhour_16	.53**
Number of Impression or Click 2010	.82**
Number of Impression or Click 2011	.77**
Matchtype_Broad	.04**
Matchtype_Exact	.05**
Sales Quantity	.07**
Number of Purchases 2010	.06**
Number of Purchases 2011	.04**
Search Engine on Google	.06**
Search Engine	.07**
Affiliate Marketing: MCUK	.24**
Affiliate Marketing: ADJUG6	.30**
Affiliate Windows	.09**
Affiliate Marketing: DRIVEPM	.26**
Affiliate Marketing: MCUKQUIDCO	.05**
Affiliate Marketing: MCUKYAHOO	.37**
Number of Purchases	.07**

**Correlation is significant at the 0.0 level

The previously table confirms that the variable "Rank" is correlated with many marketing activities. For this reason, in the predictive data mining model we will not insert this variable.

A predictive data mining model is needed to better understand the strength and sign of these relationships: the following tables show the interaction between the target variables and the explanatory variables related to the potential customer behaviour. The attention has been paid only on the variables with high predictive value. The "Enter Method" has generated five models before to identify an accurate predict model able to detect the main marketing activities generated by potential customers that bough the service online.

The following tables provide a representation of the Logistic Regression Models developed in this analysis. Only the variables with a significance level < of 0.05 can be consider a good predictors for our research goals. On the other hand, variables with p-value > 0.05 must be removed from the model because irrelevant for the analysis. Additionally, particular attention must be paid on the last column (Exp (β) Odds-Ratio[8]) because of it measures the strength of association between each explanatory variable and the target variable (Guidici 2003). Very high value of the Odds-Ratio could indicate some anomalies in the variable studied.

From Table 8 it emerges that the variables "Affiliate Windows and Affiliate Marketing: DRIVEPM are redundant between them. In fact, these variables have been eliminated from the final data mining predictive model.

Table 4 Logistic Regression Analysis (First Model)

VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β) Odds-Ratio
Avgctr	.673	.209	10.341	.001	1.959
Avgpos	-.055	.029	3.568	.059	.947
Dynamic Click	2.503	.213	138.364	.000	12.222
Avgpos_best5	-.544	.075	52.390	.000	.580
Brandfl	.943	.057	272.236	.000	2.569
Impclkhour_13	-.003	.006	.222	.637	.997
Impclkhour_14	-.005	.007	.618	.432	.995
Impclkhour_15	.000	.005	.004	.947	1.000
Impclkhour_16	-.010	.005	3.738	.053	.990
Matchtype_Broad	.364	.059	38.596	.000	1.439
Matchtype_Exact	.947	.062	230.470	.000	2.579
Search Engine on Google	-.284	.061	21.967	.000	.753
Affiliate Marketing: MCUK	.047	.006	61.160	.000	1.048
Affiliate Marketing: ADJUG6	.002	.002	1.705	.192	1.002
AffiliateWindow	-1.120	.239	22.017	.000	.326
Affiliate Marketing: DrivePM	-.003	.001	12.055	.001	.997
Affiliate Marketing: MCUKYAHOO	.004	.001	7.338	.007	1.004

*Significance Level < 0.05

Table 5 Logistic Regression (Second Model)

VARIABLE IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β) Odds-Ratio
Avgctr	.674	.209	10.447	.001	1.963
Avgpos	-.055	.029	3.569	.059	.947

[8] If the value of the Odds-Ratio is >1 there is a positive association between the variables while, when the 0 < Odds-Ratio < 1 it means that there is a negative association between them (Giudici 2003).

Dynamic Click	2.504	.213	138.486	.000	12.237
Avgpos_best5	-.545	.075	52.511	.000	.580
Brandfl	.944	.057	272.868	.000	2.571
Impclkhour_16	-.016	.003	34.081	.000	.985
Matchtype_Broad	.365	.059	38.910	.000	1.441
Matchtype_Exact	.949	.062	231.410	.000	2.582
Searchengnm_Google	-.285	.061	22.209	.000	.752
Affiliate Marketing: MCUK	.046	.006	60.970	.000	1.047
Affiliate Marketing: ADJUG6	.002	.001	1.818	.178	1.002
AffiliateWindow	-1.137	.238	22.882	.000	.321
Affiliate Marketing: DrivePM	-.003	.001	16.679	.000	.997
Affiliate Marketing: MCUKYAHOO	.003	.001	6.720	.010	1.003

*Significance Level < 0.05

Table 6 Logistic Regression (Third Model)

VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β) Odds-Ratio
Avgctr	.655	.208	9.912	.002	1.925
Avgpos	-.055	.029	3.650	.056	.946
Dynamic Click	2.505	.213	138.669	.000	12.250
Avgpos_best5	-.544	.075	52.346	.000	.581
Brandfl	.943	.057	272.632	.000	2.568
Impclkhour_16	-.015	.002	36.135	.000	.985
Matchtype_Broad	.366	.059	39.071	.000	1.442
Matchtype_Exact	.950	.062	232.254	.000	2.586
Search Engine on Google	-.286	.061	22.395	.000	.751
Affiliate Marketing: MCUK	.047	.006	62.889	.000	1.048
AffiliateWindow	-1.142	.237	23.141	.000	.319
Affiliate Markeitng: DRIVEPM	-.003	.001	16.598	.000	.997
AffiliateMarketing: MCUKYAHOO	.003	.001	6.487	.011	1.003

*Significance Level < 0.05

Table 7 Logistic Regression (Fourth Model)

Variables in the Equation	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β) Odds-Ratio
Avgctr	4.251	.246	299.276	.000	70.144
Dynamic Click	8.032	.059	18427.986	.000	3078.246
Avgpos_best5	.969	.102	89.560	.000	2.635
Brandfl	1.889	.085	494.237	.000	6.611
Impclkhour_16	.019	.003	45.070	.000	1.019
Matchtype_Broad	1.752	.085	421.894	.000	5.766
Matchtype_Exact	2.714	.095	810.470	.000	15.096
Search Engine on Google	.352	.096	13.488	.000	1.422
Affiliate Marketing: MCUK	.013	.001	144.151	.000	1.013
AffiliateWindow	-.510	.074	47.662	.000	.601
Affiliate Marketing: DRIVEPM	-.005	.001	49.566	.000	.995
Affiliate Marketing: MCUKYAHOO	.001	.001	.747	.387	1.001

*Significance Level < 0.05

Table 8 Logistic Regression (Fifth Model)

VARIABLES IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level *	Exp (β) Odds-Ratio
Avgctr	4.245	.246	298.819	.000	69.722
Dynamic Click	8.035	.059	18483.196	.000	3086.543
Avgpos_best5	.971	.102	90.027	.000	2.641
Brandfl	1.888	.085	494.367	.000	6.606
Impclkhour_16	.019	.003	49.044	.000	1.019
Matchtype_Broad	1.753	.085	422.188	.000	5.769
Matchtype_Exact	2.716	.095	811.705	.000	15.113
Search Engine on Google	.351	.096	13.386	.000	1.420
Affiliate Marketing: MCUK	.013	.001	144.343	.000	1.013
AffiliateWindow	-.508	.074	47.389	.000	.602
Affiliate Marketing: DRIVEPM	-.005	.001	49.640	.000	.995

*Significance Level < 0.05

Being the relationship too strong from a quantitative standpoint, the variable "Dynamic Click" has been removes from the predictive model in order to avoid polluting the other variables. As showed by table 3, the result can be erroneously due to a dynamic banner that links to the purchase page or to a genuine loyalty of customers to Quidco.

Table 9 shows the maximum likelihood estimates corresponding to the final model and the statistical significance of the parameters. For the entire explanatory variable we obtain a significance level lower than 0.05. This means that all explanatory variables are significantly associated with the number of purchases and they are useful in explaining whether a variable is a good predictor for the customer conversion.

Table 9 Logistic Regression Model (Optimal Model)

VARIABLE IN THE EQUATION	Estimate Value	Standard Error	Wald Statistic	Significance Level*	Exp (β) Odd-Ratio
Avgctr	2.923	.223	171.955	.000	18.605
Avgpos_best5	.597	.086	47.973	.000	1.818
Brandfl	1.676	.071	549.417	.000	5.343
Impclkhour_16	.109	.002	2909.043	.000	1.115
Matchtype_Broad	1.289	.071	327.345	.000	3.631
Matchtype_Exact	2.089	.079	700.440	.000	8.073
Search Engine on Google	.502	.080	39.553	.000	1.652

*Significance Level <0.05

Table 10 Confusion Matrix

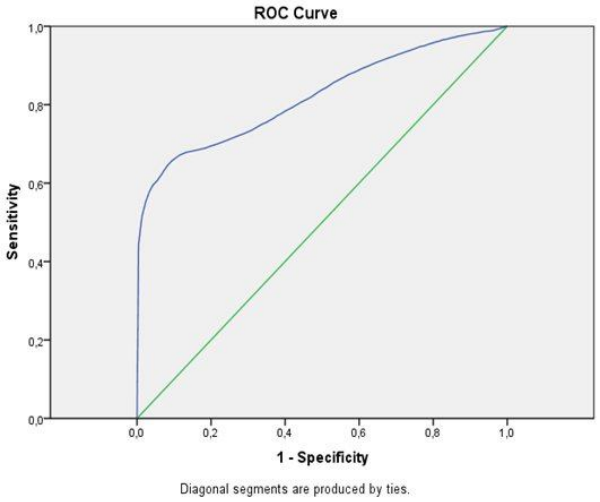
OBSERVED VALUES		PREDICTED VALUES		
		Number of Purchases		Percentage Correct
		0	1	
Number of Purchases	0	1.434.637	2.694	**99.8**
	1	1.3017	6.556	**33.5**
Total Accuracy of the Predictive Model				**98.9***

*The cut-off value is 0.05

Table 10 confirms that the logistic regression model based on the "Enter Method" is a good predictive tool for estimating the main marketing activities in which the company must be maximize its future investments. (Total Accuracy of the Predictive Model: 99.8%).

In addition, Graph 1 shed light the attention on the Area Under the Roc Curve (AUC). In our case the AUC is equal to 82.10% thus according to Sweet (1998) the correctness of the model in predicting both probability of churn and customer conversion is moderate.

Graph 1 Roc Curve



In other words, just seven variables are particularly significant for forecasting the probability of customer conversion and maximize the company profitability.

Future Research Directions

This paper focused the attention on the GLMs and on the Data Mining Methodology to forecast accurate marketing performance in order to increase the company profitability in today global business. Whit new data the model created can be enhanced to predict future sales using current marketing activities. From a methodological point of view it could be interesting to develop a latent cluster analysis for identifying potential customer partition based on their behaviour and a survival analysis for estimating how many time customers remain so in the company.

Conclusions

From the results of the previous analysis it is clear that the data mining approach is a strategic methodology for forecasting accurate marketing performance especially in global organizations with an outside-in perspective. In fact, the research method developed in this article is effective in explaining both customer conversion and probability of churn. In our case, due to the dynamism and the large amount of the data, only a traditional statistical data analysis would have been inadequate to predict punctual marketing performance. In fact, decision makers must implement a new generation of computation techniques and tools in order to assist the extraction of useful knowledge from the rapidly growing volumes of data.

"Customer data is obtained through filtering, integrating, and extracting or formatting customer data" (Buchnowska 2011, p.27). After collecting data, companies transform customer data into customer information through various information systems. For instance, the main tools for extracting and analysing huge volume of data are the following: Customer Relationship Management, Business Intelligence and Customer Intelligence System. All this information systems are based on Data Mining Applications and Statistical Algorithms.

Nowadays, the term customer knowledge is frequently incorrectly confused with the concept of customer relationship management. This latter it is been defined as the business strategies, process, culture and technology that achieve firms to maximize their revenue and to rise their value through understanding and satisfying the individual customer's needs' (Reynolds 2002).

According to Goldemberg (2003) it is possible notes that this process integrates people, process and technology in order to maximize relationships with all customers. The main difference between customer relationship management and customer knowledge is that the first is broadly focused on the management of customer knowledge while the second is completely focused on knowledge from customers (Wilde 2011; Gibbert, Leibold, Probst 2002). "Knowledge from the customer is the knowledge that organizations receive from customers" (Buchnowska 2011, p.27). In this knowledge category it is possible to include the following types of knowledge: customer knowledge of products, supplier and markets (Gebert, Geib, Kolbe, Riempp 2002), the customer ideas and suggestions about the improvement of the product or the service (Triki and Zouaoui 2011), ideas thoughts and information related to the preferences, creativity or experience with products, services, processes or expectations (Peng, Lawrence and Lihua 2011).

Customer knowledge can be consider as an organized and analysed competitive customer information so that it becomes understandable and applicable in solving problems and making decisions in the area of relations between an organization and its customers. Often, information technology tools facilitate the gathering of customer data and its transformation into customer information. Unfortunately, these tools cannot convert customer information into customer knowledge, because knowledge is always related to a person or group of people (Rollins and Halinen, 2005; Ziemba and Minich 2005). From this point of view, it is obvious that the data mining approach and the knowledge discovery in databases are relevant for process for global companies in order to discover tapped knowledge and improve their business performance. Given the dynamism of the data mining models any process, from biotechnology to customer service, can be analysed using data mining. Finally, Foley and Russuel (1998) and K Pal (2011) said that the top three ends of uses of data mining are in business area with particular attention for the marketing sector.

Biography

Ahlgren, P., Jarneving, B. & Rousseau, R. (2003). Requirement for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54,(6), 550-60.

Baum, CF. (2006). *An introduction to modern econometric using Stata.* Stata Press.

Berry, J.A., & Linoff, G.S. (2011). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management.* Indianapolis. Wiley.

Buchnowska, D. (2011). *Customer Knowledge Management Models: Assessment and Proposal*. Department of Business Informatics. University of Gdansk, Sopot, Poland.

Bueren, A., Schierholz, R., Kolbe, L., & Brenner, W. (2004). Customer Knowledge management – improving performance of customer relationship management with knowledge management. *In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences. IEEE.*

Fayyad, U.M., Shapiro, P. & Smyth, P. (1996). From Data Mining to Knowledge Discovery: An Overview. *In Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. MIT Press, 471-494.

Fayyad, U.M. & Uthurusamy, R. (2002). Envolving data into mining solutions for insights. *Communication of the ACM,* 45 (8), 28-31.

Figini, S. & Giudici, P. (2009). Applied Data Mining for Business and Industry. New York. John Wiley & Sons, Inc.

Foley, J. & Russell, J.D. (1998). Mining your own Business. Retrivied on 10 July 2007 from http://www.informationweek.com/673/73iudat.htm.

Gebert, H., Geib, M., Kolbe, L. & Riempp, G. (2002). Towards Customer Knowledge Management: Integrating Customer Relationship Management and Knowledge Management Concepts. *In Proceedings of ICEB Conference, Taiwan*.

Gibbert, M., Leibold, M. & Probst, G. (2002). Five Styles of Customer Knowledge Management and How Smart Companies Put them into Actions. *European Management Journal,* 20(5), 459-460.

Giudici, P. (2003). *Applied Data Mining.* Chichester. Wiley.

Goldenberg, B.J. (2003). *CRM Automation.* Prentice Hall PTR.

Grossman, R.L., Kamath, C., Kegelmeyear, P., Kumar, V. & Novnburu, R. (2001). *Data Mining for scientific and engineering applications.*London. Springer-Verlag.

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York, USA: Farrar, Straus and Giroux.

KhakAbi, S. & Mohammad, R.G. (2010). Data Mining Applications in Customer Churn Management. International Conference on Intelligent Systems, Modelling and Simulation. *In Computer Society, IEEE.*

K Pal, J. (2011). Usefulness Applications of data mining in extracting information from different persoectives. *Annuals of Library and Information Studies*, 58, 7-16.

Hadden, J., Tiwari, A., Roy, R. & Ruta, D. (2005). Computer assisted customer churn management: State-of-the-art and future trends. *In Computers & Operations Research*, 34, 2902-2917.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining.* Cambridge. MIT Press.

Larose, D. (2010). *Data Mining Methods and Models*. New Jersey. John Wiley & Sons, Ltd.

Lehmann, E.L. & Romano, P.R. *Testing Statistical Hypotheses.* New York. Springer.

Lejeune, M. (2001). Measuring the impact of data mining on churn management. In *Internet Research: Elettronic Marketing Applications and Policy*, 11, 5, 375-387.

Murthy, I.K. (2010). *Data Mining-Statistics Applications: A Key to Managerial Decision Making*. Indiastat.com.

Moody, C. (2009). Basic Econometric with Stata. (http://www.scribd.com/doc/54549213/Stata-Manual-2009).

Peng, J., Lawrence, A. & Lihua, R. (2011). Customer Knowledge Management in International Project: A Case Study.

Reynolds, J. (2002). A practical guide to CRM: building more profitable customer relationships. New York. CMP Books.

Rollins, M. & Halimen, A. (2005). Customer Knowledge Management Competence: Towards a Theoretical Framework. *In: Proceedings of the 38th Hawaii International Conference on System Sciences*.

Sato, Y. (2000). Perspective on data mining from statistical viewpoints. Knowledge Discovery and Data Mining. *In: Current issues and New Applications, 4th Pacific-Asia Conference, PAKDD, Kyoto, Japan.*

Sweet, J.A. (1988). Measuring the accuracy of diagnostic system. *Science, 240*, 1285-1293

Triki, A. & Zouaoui, F. (2011). Customer Knowledge Management Competencies Role in the CRM Implementation Project. *Journal of Organizational Knowledge Management*.

Turban, E., Aronson, J.E., Liang, T.P. & Sharda, R. (2008). *Decision support and business intelligence systems (8th ed.).* Essex. Pearson Education.

Wilde, S. (2011). *Customer Knowledge Management. Improving Customer Relationship Through Knowledge Application.* Heidelberg. Springer.

Ziemba, E. & Minich, M. (2005). Informacja I wiedzy w przedsiebiorstwie. In: Olenski, J., Olejniczak, Z., Nowak, J. (Ed.), Informatyka. *Strategie I zarzadzanie wiedza, Polskie Towarzystwo Informatyczne, Katowice.*