41th EMAC Conference

Marketing to Citizens
Going beyond Customers and Consumers

22-25 May 2012
ISCTE Business School

ISBN: 978-989-732-004-0

Conference Secretariat
ISCTE Business School
Av.ª das Forças Armadas
1649-026 Lisbon
Portugal

EMAC 2012 Conference Website administration:
EMAC Secretariat
c/o EIASM, Place de Brouckère Plein 31
1000 Brussels
Belgium
Tel: +32 2 2266660
Fax: +32 2 5121929

Enhancing Bank Direct Marketing through Data Mining

Author: SÉRGIO MORO - Email: scmoro@gmail.com
University: LISBON UNIVERSITY INSTITUTE (ISCTE)

Track: Modelling and Forecasting

Co-author(s): Raul M. S. Laureano (Instituto Universitário de Lisboa (ISCTE-IUL) , UNIDE, Lisboa, Portugal)
/ **Paulo Cortez** (Universidade do Minho)

Enhancing Bank Direct Marketing through Data Mining

The financial crisis created pressure on banks due to credit restriction, increasing competition for deposits retention and demanding efficiency improvements of direct marketing campaigns.

Our research conducted a data mining project on direct marketing campaigns for deposits subscriptions by using recent data of a Portuguese retail bank. We used the Support Vector Machine (SVM) data mining technique for modeling and evaluated it through a sensitive analysis.

The findings revealed previously unknown valuable knowledge, such as the best months for campaigns to occur, and optimal call duration. Such knowledge can be used to improve campaign efficiency.

Keywords: direct marketing, data mining, business intelligence, targeting, contact management, retail banking.

Track: Modeling and Forecasting.

1. Introduction

The execution of direct marketing campaigns over time produces data and information in the form of reports that need to be analyzed by managers in order to support decision making. However, it is a difficult task for humans to analyze all this vast and often complex data. This difficulty led to the development of business intelligence techniques, which aim at the extraction of useful knowledge to support decision making (Turban et al., 2010).

Although every profitable business intends an increase of the return on investment (Keen & Digrius, 2003), huge drivers can potentiate this need. The global economic crisis is a great example: it is paving its way by triggering new ideas about financial management and new thoughts and points of view (Hodgson, 2009). Considering the recent effects of the crisis on Europe, one consequence for banks and, in particular, for those more affected due to harshness of the countries public debts, is the credit restriction. Moreover, to prevent an uncontrolled increase in the offered rates, some national banks imposed limits (Penty, 2011). That led the retailing banks to invest in products and campaigns to gather and retain financial assets through deposits. In this context, the following research question is pertinent: can we extract knowledge through available data from past campaigns in order to improve future long-term deposit campaigns efficiency?

In this paper, we answer such question under a data mining approach. Using recent data from direct marketing campaigns ran between May 2008 and November 2010, a good tuned model is achieved allowing for a better understanding of how call duration, months and past successes influences future results. We emphasize on the most relevant knowledge discovered with this research, in a belief that it can effectively benefit business by helping to support management decisions.

The paper begins with a brief review of some background on the concepts of direct marketing and business intelligence, followed by the methodological approach and the available data. Then, the model is evaluated and the knowledge discovered is analyzed in a results section. Finally, conclusions are drawn and guidelines for future improving are suggested.

2. Background

Direct marketing is the process of identifying likely buyers of certain products and promoting the products accordingly (Ling & Li, 1998). Targeting a set of selected clients allows choosing those that supposedly will be keener to acquire the product or service being offered (Ou et al., 2003). Still, the use of telemarketing phone calls is generating a growing number of complaints due to an increase in its use (Tapp, 2008).

One effective way of analyzing reports from previous and similar campaigns in the search for trends and patterns is through business intelligence and data mining techniques, to build models and then extract knowledge. Business intelligence is a vast concept that includes data mining which consists in knowledge extraction from raw data, according to Turban et al. (2010) and Witten and Frank (2005).

Recent decades provided more advanced modeling methods, being one of the most recent the Support Vector Machine (Cortes & Vapnik, 1995). It defines optimal hyper planes through the definition of a subset of representative cases, the support vectors (Hearst et al., 1998). By applying a non-linear function to the inputs of a new case, it defines its location in the space divided by the hyper planes. One of the more used functions is the Gaussian kernel, since it needs lesser parameters and so is easier to use. This technique revealed high performance results on classification problems, compared to other techniques (Wu et al.,

2008), like Naïve Bayes (NB) (Zhang, 2004) and Decision Trees (DT) (Aptéa & Weiss, 1997). Nevertheless, one of the relevant issues that favors older and more common data mining techniques like Logistic Regression and DT in detriment of the more recent and advanced techniques like the Artificial Neural Network (ANN) and the SVM is that the models built through the later are hard to understand and their interpretation is not intuitive for business managers (Witten & Frank, 2005). Considering this issue, to help to extract knowledge from models, a sensitive analysis method can be used to characterize the inputs influence in success (Cortez & Embrechts, 2011). With this method, we measure the effects on the output of a given model by varying the inputs through their range of values.

To evaluate a model, some techniques can be used, like the confusion matrix (Kohavi & Provost, 1998) and the Receiver Operating Characteristic (ROC) curve (Fawcett, 2005), both based on testing the model to predict outcome and compare it to the observed result. Another model evaluation technique quite popular in marketing analysis is the Lift cumulative curve, which shows how much positive answers would be achieved from a partial selection of the most likely positive answers considered by the model (Coppock, 2002).

Application of data mining to marketing is a subject of research. For example, Kim and Street (2004) applied ANN to predict which households are interested in purchasing an insurance policy for recreational vehicles. Hu (2005) and Li et al. (2010) also applied data mining to bank clients, in the first case for client retention, and the later for clustering clients based on their profile. Relating to bank credit, Shen et al. (2007) studied credit card fraud detection models and Nwulu and Oroja (2011) researched on credit scoring models using both SVM and ANN.

Nobibon et al. (2011) studied optimization models for targeted banking offers. Their research provided excellent prediction results through the use of heuristic algorithms for large datasets.

3. Data and Methods

The dataset used was supplied by a Portuguese bank and related to direct marketing reports for contacts to sell long-term deposits using the telephone as the communication channel. The corresponding campaigns were executed between May 2008 and November 2010, and encompassed a total of 79354 contacts, 6499 of which were successful (8% of the total), i. e., the product was subscribed.

Some initial relevant flattening of the contact information was required, since the reports included one independent line per call, but a client could receive numerous calls within the same campaign (meaning that only the final call was a terminal state). To solve this issue, we decided to keep just the first and last call information, and save a counter for the total number of calls to the client in that campaign. For the whole set of contacts, only 27870 of them had more than two calls, so no information was lost for about 65% of the contacts. Intuitively, the most relevant information belongs to the final call, since that holds the dialog that conducted the contact to a result. We also considered the first call important since it introduced the campaign or contact attempt to the client, even though the call was rescheduled. To characterize each specific client contacted, further data besides the report results were gathered. The result of joining both client and contact characteristics was a dataset with 59 attributes.

A data mining project was conducted using the CRISP-DM methodology (Chapman et al., 2000) through its several phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. During this process, the dataset was reduced both in length (number of contacts) and dimension (number of attributes). All the

contacts for which the last call result was different from success or unsuccessful were eliminated (that happened for calls that were scheduled but, for some reason, the client could not be contacted again, or if the call could not be made at all – e. g. if the phone number was wrong). Also all the contacts with missing data were excluded. For the feature selection, an exploratory data analysis procedure was conducted, using graphics generated using the rattle tool (Williams, 2009) that showed the relation between the predictor attribute and the contact outcome (success or unsuccessful). The attribute was excluded in case of an inexistent relation. The result was a dataset with 45211 contacts (5289 of which were successes) with 29 characterization attributes, which were grouped according to its type and/or origin (Table 1).

The modeling of the contact's success was done through the rminer package (Cortez, 2010), based on the R statistical environment, both of which are open source. The model was obtained with the SVM by using 2/3 of the contacts for model training and another 1/3 for testing.

Table 1 - Attributes used to implement the model

Group	Name	Description and Values	
Personal Client Information	Age	In years and at the date of the last contact made (Mean=41, SD=10.6)	
	Professional Status	27 possible values (Mode="specialized operative")	
	Marital Status	Married (60%), Divorced (10%), Separated (1%), Single (28%), Widowed (1%)	
	Title	22 possible values (Mode= "Without specific title")	
	Academic Qualifications	10 possible values (Mode="University degree")	
Bank Client Information	Loans in Delay	Indicates that the client has loans in delay – (Y)es=98%/(N)o=2%	
	Average Annual Balance	Average annual balance of all the current accounts that the client owns (Mean=1362, SD=3045)	
	Debt Card	Indicates that the client has debt card – Y=77.4%/N=22.6%	
	Credit Card	Indicates that the client has credit card – Y=54%/N=46%	
	Mortgage Credit	Indicates that the client has a mortgage account – Y=56%/N=44%	
	Individual Credit	Indicates that the client has an individual credit – Y=16%/N=84%	
	Domiciliation	Indicates that the client has domiciliation for automatic payment of one or more authorized debts – Y=79%/N=21%	
Contact	Nr. of Calls	Number of calls for the same campaign (Mean=2.76, SD=3.10)	
		First Call	**Last Call**
	Result	Mode= "*no first call made*"	*The outcome (Successes=11.7%)*
	Human Agent	102 different agents made all the calls	
	Phone Type	Mobile=34%,Fixed=4%,NA=62%	Mobile=65%,Fixed=29%,NA=6%
	Day of Month	*No first call was made for 43.3% of the contacts, meaning that this percentage of contacts were ended in just one call*	*Even distribution through days*
	Month		Mode=May (30.45%)
	Hour		*Even distribution (10am-9pm)*
	Duration in seconds		Mean=258.2, SD=257.5
History	Days Since Last Contact	Number of days since the last contact for any other campaign (Mean=41, SD=100)	
	Nr. of Contacts	Total number of previous contacts (Mean=0.58, SD=2.30)	
	Last Result	Result for the last campaign (81.7% did not have been yet contacted)	

4. Results

The model was validated through 20 runs of the modeling code to rebuild the model and validating it by a different selection of the training and test samples (cross-validation) resulting in a confusion matrix, presented in Table 2, and an area under the ROC curve of 0.938 (the higher the better, with an ideal model having a 1.0 value), meaning that the model can be used to extract knowledge about contacts. We also used the Lift cumulative curve, which resulted in an area under the curve of 0.887 and, considering the predictions for the test set, if we select the 30% most likely subscribers of the deposits, we would obtain the total

number of successes from the whole test set. Those Lift and ROC values obtained were compared to two other models obtained through the classic techniques of NB and DT and the results also sustain our choice of the SVM (Table 3).

Table 2 - Confusion matrix

Predicted / Observed	Unsuccessful	Success	Correctly Classified
Unsuccessful	225016	41144	83.60%
Success	3287	31973	90.68%

The overall prediction accuracy is 85.26%

Table 3 – Models Comparison

Model	NB	DT	SVM
ROC	0.870	0.868	0.938
Lift	0.827	0.790	0.887

Both the ROC and Lift values refer to the area under the respective curve graphics

Through some functions provided by the rminer package, it is possible to obtain a graphic that shows the importance of each attribute in defining the model, as shown in Figure 1 for the 10 most relevant attributes, thus summarizing the sensitive analysis.

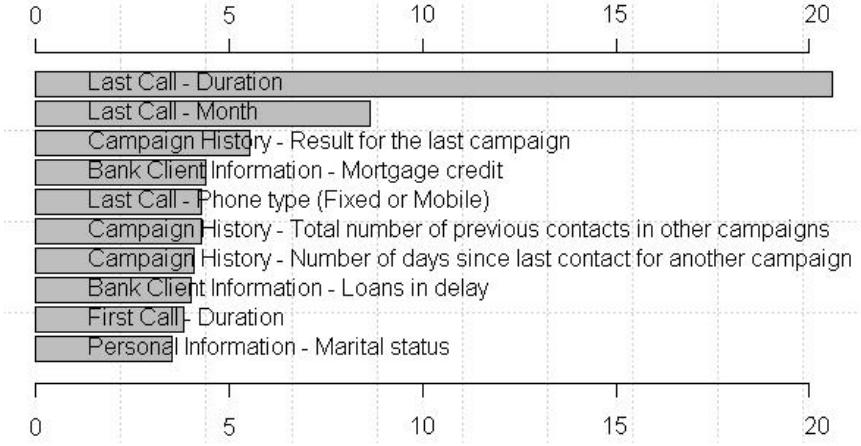

Figure 1 - The most relevant attributes (in percentage) for the SVM model

The two most important attributes are related to "Last Call". This emphasizes how important the runtime execution call information is for success comparing to the other more static client related information. Clearly the managers should have an in-loco management style, making adjustments after the campaign is opened for execution.

To have a more specific knowledge from the model, it is necessary to find how each attribute influences the probability of success. In Figures 2 to 5, we plot the variable effect characteristic (VEC) curve, which plots the average influence of a given attribute (x-axis) on the SVM probability of success (y-axis) (Cortez & Embrechts, 2011).

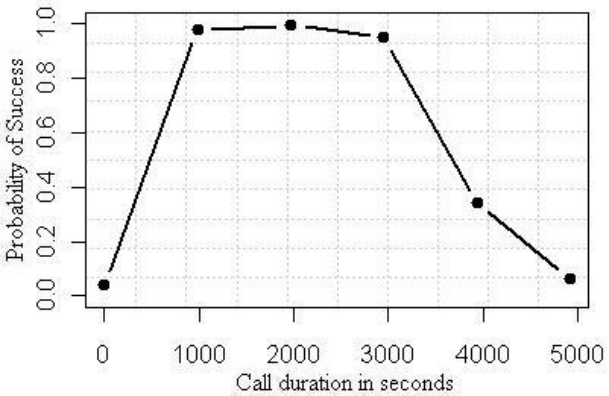

Figure 2 - Influence of the last call duration on success

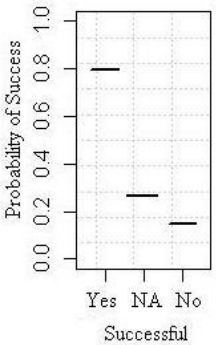

Figure 3 - Influence of the last campaign result on success (NA – no previous contact was made)

Rechecking Figure 1, one understands that the call duration by its own explains more than 20% of the success, regarding the model achieved. Although maybe this is common sense, since successful calls require a deeper dialog to describe the product, we can check in Figure 2 that after a certain call duration (3000 seconds, or 50 minutes), the probability of success starts to decrease. Analysis to the month influence (Figure 4) reveales that success is

more likely to occur in the last month of each quarter (March, June, September and December). This can be a valuable knowledge, since managers can try to shift campaigns for those months. Regarding the result for the last campaign, it is quite obvious from Figure 3 that a previous success means it is much more likely that the next campaign is also successful for the same client. For the mortgage credit ownership, the influence difference is not so remarkable (Figure 5), but still not having that credit makes more likely a deposit subscription as a result of a direct marketing campaign.

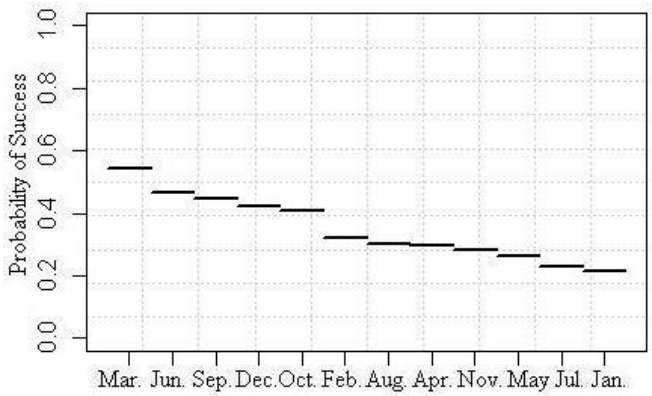

Figure 4 - Influence of the month the last contact was executed on the success

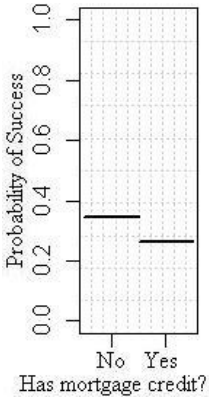

Figure 5 - Influence of having a mortgage credit on success

5. Conclusions

In bank direct marketing, sets of reports results from executed campaigns are used to identify trends of behavior to a better management and more judged decisions. Managers are shifting from traditional statistics analysis towards new more sophisticated techniques of data mining. Still, extracting valuable knowledge from explanatory models requires simple tools and easy to understand. This research focused on data mining through the SVM technique with the goal of enhancing bank direct marketing. For instance, we discovered that the best months for contact marketing are the last of each quarter. It is our belief that campaign management can benefit from this knowledge. On the other hand, we confirmed that clients previously contacted with success are keener to repeat a future deposit subscription.

Although the model evaluation results were very good, this means only that it provided a good explanation for campaigns already occurred in the past. For future work, the model should be built by using more recent data as an input and, ideally, by adding other characteristics non-related to the contact execution (e. g. residence county).

References

Aptéa, C. & Weiss, S. (1997). Data mining with decision trees and decision rules, *Future Generation Computer Systems, 13*(2-3), 197–210.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 – Step-by-step data mining guide*, CRISP-DM Consortium.

Coppock, D. (2002). *Why Lift? – Data Modeling and Mining*, Information Management Online (June).

Cortes, C. & Vapnik, V. (1995). Support Vector Networks, *Machine Learning,20*(3),273–297.

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. *Proceedings of the 10th Industrial Conference on Data Mining (Berlin, Germany, Jul.)*, Springer, LNAI 6171, 572–583.
Cortez, P. & Embrechts, M. (2011). Opening Black Box Data Mining Models Using Sensitivity Analysis. *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining (Paris, France)*, 341–348.
Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
Hearst, M., Dumais, S., Osman, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines, *IEEE Intelligent Systems, 13*(4), 18–28.
Hodgson, G. (2009). The Great Crash of 2008 and the Reform of Economics, *Cambridge Journal of Economics, 33*(6), 1205–1221.
Hu, X. (2005). A data mining approach for retailing bank customer attrition analysis. *Applied Intelligence, 22*(1), 47–60.
Keen, J. & Digrius, B. (2003). *Making Technology Investments Profitable: ROI Roadmap to Better Business Cases*, John Wiley & Sons, USA.
Kim, Y. & Street, W. (2004). An Intelligent System for Customer Targeting: A Data Mining Approach. *Decision Support Systems, 37*(2), 215–228.
Kohavi, R. & Provost, F. (1998). Glossary of Terms. *Machine Learning, 30*(2–3), 271–274.
Li, W., Wu, X., Sun, Y. & Zhang, Q. (2010). Credit Card Customer Segmentation and Target Marketing Based on Data Mining. *Proceedings of International Conference on Computational Intelligence and Security*, 73–76.
Ling, X. & Li, C., (1998). Data Mining for Direct Marketing: Problems and Solutions. *Proceedings of the 4th KDD conference*, AAAI Press, 73–79.
Nobibon, F., Leus, R. & Spieksma, F. (2011). Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms, *European Journal of Operational Research, 210*, 670–683.
Nwulu. N. & Oroja, S. (2011). A Comparison of Different Soft Computing Models for Credit Scoring, *World Academy of Science, Engineering and Technology 78*, 898–903.
Ou, C., Liu, C., Huang, J. & Zhong, N. (2003). On Data Mining for Direct Marketing. *Proceedings of the 9th RSFDGrC Conference*, 2639, 491–498.
Penty, C. (2011). Spain's Cabinet Approves Steps to Counter Bank Deposit War, *Bloomberg, 3 June 2011*, http://www.bloomberg.com/news/2011-06-03/spain-s-cabinet-approves-steps-to-counter-banks-deposit-war-.html.
Shen, A., Tong, R. & Deng, Y. (2007). Application of Classification Models on Credit Card Fraud Detection, *Proceedings of the International Conference on Service Systems and Service Management*, 1–4.
Tapp, A. (2008). Introducing direct marketing. In *Principles of direct and database marketing – a digital orientation, 4th edition* (pp. 1–52), Prentice Hall, USA.
Turban, E., Sharda, R. & Delen, D. (2010). *Decision Support and Business Intelligence Systems*, 9th edition, Prentice Hall Press, USA.
Williams, G. (2009). Rattle: a data mining GUI for R. *The R Journal, 1*(2), 45–55.
Witten, I. & Frank, E. (2005). *Data Mining – Practical Machine Learning Tools and Techniques*, 2nd edition, Elsevier, USA.
Wu, X., Kumar, V., Quinlan, J., Gosh, J., Yang, Q., Motoda, H., MacLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D. & Steinberg, D. (2008). Top 10 algorithms in Data Mining, *Knowledge and Information Systems, 14*(1), 1–37.
Zhang, H. (2004). The Optimality of Naïve Bayes. *Proceedings of the 17th FLAIRS conference*, AAAI Press.