ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/236231158

Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology

Conference Paper · October 2011

CITATIONS
95

READS
3,043

3 authors:

Sérgio Moro
ISCTE-Instituto Universitário de Lisboa
41 PUBLICATIONS **421** CITATIONS
SEE PROFILE

Paulo Cortez
University of Minho
217 PUBLICATIONS **2,973** CITATIONS
SEE PROFILE

Raul Laureano
ISCTE-Instituto Universitário de Lisboa
70 PUBLICATIONS **198** CITATIONS
SEE PROFILE

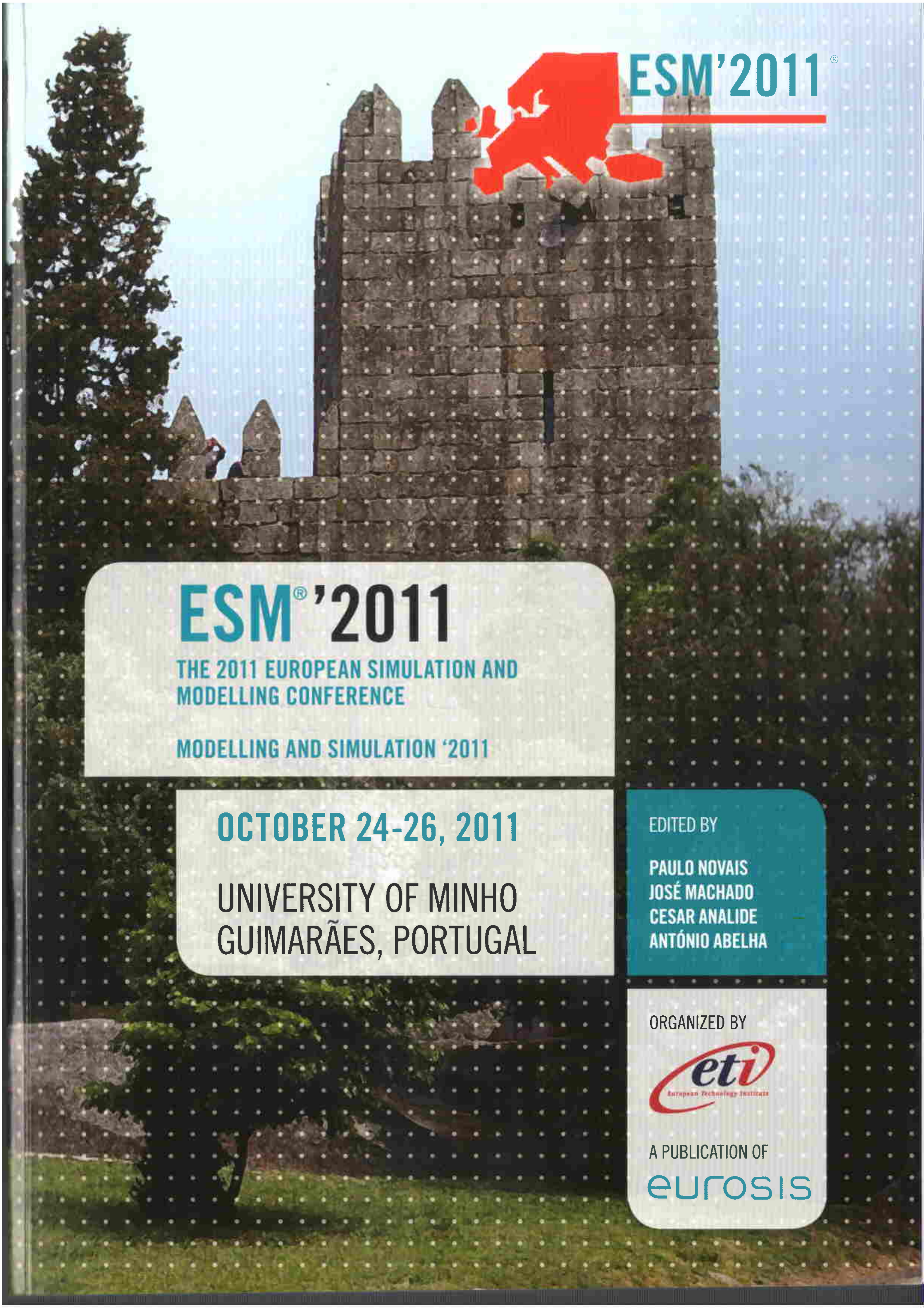
Some of the authors of this publication are also working on these related projects:

Project Modern Optimization with R, Springer, 2014 (book) View project

Project A decision support system for slope management and repair View project

All content following this page was uploaded by Paulo Cortez on 23 November 2015.

The user has requested enhancement of the downloaded file.



ESM'2011®

ESM®'2011

THE 2011 EUROPEAN SIMULATION AND MODELLING CONFERENCE

MODELLING AND SIMULATION '2011

OCTOBER 24-26, 2011

UNIVERSITY OF MINHO
GUIMARÃES, PORTUGAL

EDITED BY

PAULO NOVAIS
JOSÉ MACHADO
CESAR ANALIDE
ANTÓNIO ABELHA

ORGANIZED BY
eti
European Technology Institute

A PUBLICATION OF
eurosis

© 2011 EUROSIS-ETI

Responsibility for the accuracy of all statements in each peer-referenced paper rests solely with the author(s). Statements are not necessarily representative of nor endorsed by the European Simulation Society. Permission is granted to photocopy portions of the publication for personal use and for the use of students providing credit is given to the conference and publication. Permission does not extend to other types of reproduction nor to copying for incorporation into commercial advertising nor for any other profit-making purpose. Other publications are encouraged to include 300- to 500-word abstracts or excerpts from any paper contained in this book, provided credits are given to the author and the conference.

All author contact information provided in this Proceedings falls under the European Privacy Law and may not be used in any form, written or electronic, without the written permission of the author and the publisher.

All articles published in these Proceedings have been peer reviewed

EUROSIS-ETI Publications are ISI-Thomson and INSPEC referenced

A CIP Catalogue record for this book is available from the Royal Library of Belgium under nr.12620

For permission to publish a complete paper write EUROSIS, c/o Philippe Geril, ETI Executive Director, Greenbridge NV, Wetenschapspark 1, Plassendale 1, B-8400 Ostend Belgium.

EUROSIS is a Division of ETI Bvba, The European Technology Institute, Torhoutsesteenweg 162, Box 4, B-8400 Ostend, Belgium

Printed in Belgium by Reproduct NV, Ghent, Belgium
Cover Design by Grafisch Bedrijf Lammaing, Ostend, Belgium

ESM® is a European registered trademark of the **European Technology Institute** under nr: 002433290

EUROSIS-ETI Publication

ISBN: 978-90-77381-66-3
EAN: 9789077381663

CONTENTS

ECLIPSE-DDML: A Graphical Modelling Tool for Dynamic System Simulation
Ufuoma Bright Ighoroje and Mamadou Kaba Traoré 61

DISTRIBUTED PARALLEL SIMULATION

HLA High Performance and Real-Time Simulation Studies with CERTI
Jean-Baptiste Chaudron, Martin Adelantado, Eric Noulard and Pierre Siron 69

Warp-Level Parallelism: Enabling Multiple Replications In Parallel on GPU
Jonathan Passerat-Palmbach, Jonathan Caux, Pridi Siregar, Claude Mazel
and David R.C. Hill 76

Simulation Based Evaluation of QOS Guided Scheduling for Distributed Computing in Large Environments
Florin Pop 84

A Generic Methodology to DEVS Parallel and Distributed Simulation
Adedoyin Adegoke and Mamadou K. Traoré 91

BUSINESS SIMULATION

Storyboards Method and Structured Shot
David Ježek and Svatopluk Štolfa 99

Trust Model for Social Network
Arnostka Netrvalova and Jiri Safarik 102

Competency Based Human Resources Allocation in Business Process Simulations
Štěpán Kuchař, Jan Martinovič and Tomáš Kocyan 108

Market Segmentation through Conjoint Analysis using Latent Class Models
Liberato Camilleri and Lara Marie Azzopardi 112

Using Data Mining for Bank Direct Marketing: An Application of the Crisp DM Methodology
Sérgio Moro, Raul M. S. Laureano and Paulo Cortez 117

INDUSTRIAL DECISION MAKING

A Simulation Study of a Multi-Site Production Plant using ARENA
António Carvalho Brito, Carlos Bragança de Oliveira
and Paulo Sá Marques 125

USING DATA MINING FOR BANK DIRECT MARKETING: AN APPLICATION OF THE CRISP-DM METHODOLOGY

Sérgio Moro and Raul M. S. Laureano
Instituto Universitário de Lisboa (ISCTE – IUL)
Av.ª das Forças Armadas
1649-026 Lisboa, Portugal
E-mail: scmoro@gmail.com, raul.laureano@iscte.pt

Paulo Cortez
Centro Algoritmi, Dep. Sistemas de Informação
Universidade do Minho
4800-058 Guimarães, Portugal
E-mail: pcortez@dsi.uminho.pt

KEYWORDS

Directed Marketing, Data Mining, Contact Management, Targeting, CRISP-DM.

ABSTRACT

The increasingly vast number of marketing campaigns over time has reduced its effect on the general public. Furthermore, economical pressures and competition has led marketing managers to invest on directed campaigns with a strict and rigorous selection of contacts. Such direct campaigns can be enhanced through the use of Business Intelligence (BI) and Data Mining (DM) techniques.

This paper describes an implementation of a DM project based on the CRISP-DM methodology. Real-world data were collected from a Portuguese marketing campaign related with bank deposit subscription. The business goal is to find a model that can explain success of a contact, i.e. if the client subscribes the deposit. Such model can increase campaign efficiency by identifying the main characteristics that affect success, helping in a better management of the available resources (e.g. human effort, phone calls, time) and selection of a high quality and affordable set of potential buying customers.

BACKGROUND

Bank direct marketing

There are two main approaches for enterprises to promote products and/or services: through mass campaigns, targeting general indiscriminate public or directed marketing, targeting a specific set of contacts (Ling and Li 1998). Nowadays, in a global competitive world, positive responses to mass campaigns are typically very low, less than 1%, according to the same study. Alternatively, directed marketing focus on targets that assumable will be keener to that specific product/service, making this kind of campaigns more attractive due to its efficiency (Ou et al. 2003). Nevertheless, directed marketing has some drawbacks, for instance it may trigger a negative attitude towards banks due to the intrusion of privacy (Page and Luding 2003).

It should be stressed that due to internal competition and current financial crisis, there are huge pressures for European banks to increase a financial asset. To solve this issue, one adopted strategy is offer attractive long-term deposit applications with good interest rates, in particular by using directed marketing campaigns. Also, the same drivers are pressing for a reduction in costs and time. Thus, there is a need for an improvement in efficiency: lesser contacts should be done, but an approximately number of successes (clients subscribing the deposit) should be kept.

Business Intelligence and Data Mining

According to Turban et al. (2010), BI is an umbrella term that includes architectures, tools, databases, applications and methodologies with the goal of using data to support decisions of business managers. DM is a BI technology that uses data-driven models to extract useful knowledge (e.g. patterns) from complex and vast data (Witten and Frank, 2005).

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology for increasing the success of DM projects (Chapman et al., 2000). The methodology defines a non-rigid sequence of six phases, which allow the building and implementation of a DM model to be used in a real environment, helping to support business decisions (Figure 1).

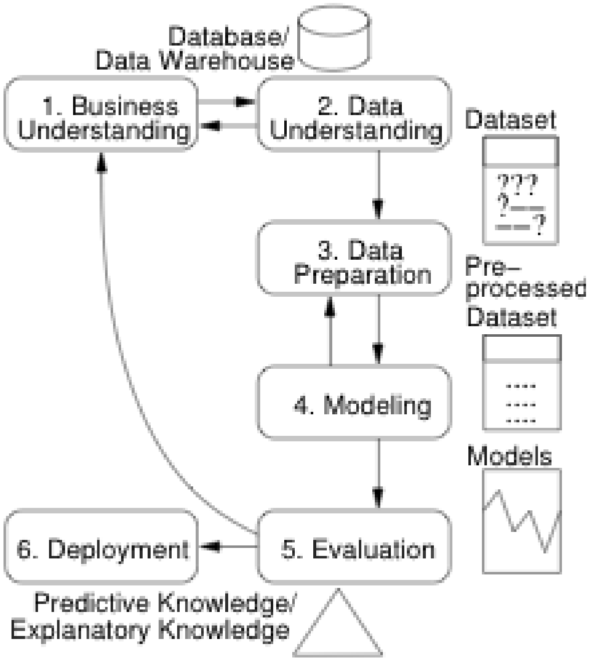


Figure 1 The CRISP-DM process model (adapted from Chapman et al., 2000)

CRISP-DM defines a project as a cyclic process, where several iterations can be used to allow final result more tuned towards the business goals. After identifying the goal to achieve (Business Understanding phase), the data needs to be analyzed (Data Understanding) and processed (Data Preparation). According to Witten and Frank (2005), data has concepts (what needs to be learned), instances (independent records related to an occurrence) and attributes (which characterize a specific aspect of a given instance). The Modeling phase builds the model that represents the learned knowledge (e.g. given an instance, the model can be used to predict the target value that represents the goal defined). Next, the model is analyzed in the Evaluation phase, in terms of its performance and utility. For instance, in classification tasks (to predict a discrete target), common metrics are the confusion matrix (Kohavi and Provost 1998) and the Receiver Operating Characteristic (ROC) curve (Fawcett 2005). If the obtained model is not good enough for use to support business, then a new iteration for the CRISP-DM is defined. Else, the model is implemented in a real time environment (Deployment phase).

Data Mining on Marketing Campaigns

Given the interest in this domain, there are several works that use DM to improve bank marketing campaigns (Ling and Li, 1998)(Hu, 2005)(Li et al, 2010). In particular, often these works use a classification DM approach, where the goal is to build a predictive model that can label a data item into one of several predefined classes (e.g. “yes”, “no”). Several DM algorithms can be used for classifying marketing contacts, each one with its own purposes and capabilities. Examples of popular DM techniques are: Naïve Bayes (NB) (Zhang, 2004), Decision Trees (DT) (Aptéa and Weiss, 1997) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995).

To access the classifier performance, classification metrics, such as accuracy rate or ROC curve, can be used. Yet, for marketing campaigns, the Lift is the most commonly used metric to evaluate prediction models (Coppock 2002). In particular, the cumulative Lift curve is a percentage graph that divides the population into deciles, in which population members are placed based on their predicted probability of response. The responder deciles are sorted, with the highest responders are put on the first decile. Lift can be effectively used as a tool for marketing managers to decide how many contacts to do (from the original set) and also to check if, for some goal of target responses, there is an alternate better model.

MATERIALS AND METHODS

Bank direct marketing data

We collected data from a Portuguese bank that used its own contact-center to do directed marketing campaigns. The telephone, with a human agent as the interlocutor, was the dominant marketing channel, although sometimes with an auxiliary use of the Internet online banking channel (e.g. by showing information to specific targeted client). Furthermore, each campaign was managed in an integrated fashion and the results for all channels were outputted together.

The dataset collected is related to 17 campaigns that occurred between May 2008 and November 2010, corresponding to a total of 79354 contacts. During these phone campaigns, an attractive long-term deposit application, with good interest rates, was offered. For each contact, a large number of attributes was stored (e.g. see Table 2) and if there was a success (the target variable). For the whole database considered, there were 6499 successes (8% success rate).

Computational Environment

All experiments reported in this work were conducted using the **rminer** library (Cortez 2010), which is an open source package for the R tool that facilitates the use of DM techniques (and in particular, classification tasks), and **rattle**, a graphical user interface for DM in R (Williams, 2009). The rminer main advantage is that only a few set of coherent functions are required for a complex DM analysis. As an example, the following R/rminer code was used to build a DT and obtain a ROC curve (the rminer functions are underlined):

```
library(rminer) # load the library
# read the data:
D=read.table("data.csv",sep=";",header=TRUE)
AT=c(6,7,9:12,17:20,22:25,27,30,31,36,43:46,48:51
,57,58,61,5)
D=D[,AT] # select some attributes
DF=na.omit(DF) # remove missing data
M=mining(y~.,DF,method=c("holdout",2/3),model="dt
",Runs=20) # model several decision trees
# plot the ROC curve
mgraph(M,graph="ROC",Grid=10,baseline=TRUE)
```

In this paper, we used three DM rminer models: NB, DT and SVM (with Gaussian kernel), while the rattle tool was used for graphical data exploration, in Data Understanding phase.

EXPERIMENTS AND RESULTS

In this section, we explain the experiments performed and analyze the results obtained over a set of three CRISP-DM iterations.

First iteration – project viability and goal definition

Starting on the *Business Understanding* phase (of the CRISP-DM), it was clear that the goal was to increase efficiency of directed campaigns for long-term deposit subscriptions by reducing the number of contacts to do.

During the Data Understanding phase, we analyzed the data main characteristics. The output presented in the reports of previous campaigns was composed of two values: the result (nominal attribute with the possible values enumerated in Table 1) and the amount of money invested (numeric value in *euro*). For this research, only the nominal result was accounted for, thus the goal is to predict if a client will subscribe the deposit, not regarding which amount is retained, turning it a classification task. Results are grouped together taking into account the type of contact, as shown in Table 1.

Table 1 Enumerated values for the output result

Contact result	**Group**
Successful	Concluded contact
Unsuccessful	
Not the owner of the phone	Cancelled contact
Did not answer	
Fax instead of phone	
Abandoned call	
Aborted by the agent	
Scheduled by other than the client	Scheduled contact
Scheduled by the client himself	
Scheduled – deposit presented to the client	
Scheduled – deposit not presented	
Scheduled due to machine answer	

The data attributes obtained from the campaign reports were most of them related directly to contact information (e. g. time and duration of contact, type of phone used, agent who made the call). Although this specific domain information may be of great value, clearly there was a need for collecting client information. Therefore, using internal databases of the bank institution, we collected extra attributes that better describe the clients' characteristics (Table 2), resulting in a total number of 59 attributes (including the output target).

Table 2 Examples of some of the 59 client attributes

Name	**Description and Values**
Personal Client Information	
Age	Age at the contact date (Numeric ≥18)
Marital status	Married, single, divorced, widowed, separated (Nominal)
Sex	Male or Female (Nominal)
Bank Client Information	
Annual balance	in *euro* currency (Numeric)
Debt card?	Yes or No (Nominal)
Loans in delay?	Yes or No (Nominal)
Last Contact Information	
Agent	Human that answered the call
Date and time	Referring to when the contact was made
Duration	Of the contact (in seconds)
First Contact Information	
Agent	Human that answered the call
Date and time	Referring to when the contact was made
Duration	Of the contact (in seconds)
Visualization's Information	
Number of times the client has seen the product in the home banking site	
History Information	
Result of the last campaign if another contact was made	
Days since last contact in other campaign	

Having into account that each instance relates to a client being contacted in a campaign context, a client can be contacted more than once (hence the scheduled contact result – which can be the final result, if the campaign is suddenly closed). For this reason, there is a group of attributes related to the first contact (which are not filled if only one contact was made) and another group for the final contact. Still, there may be intermediate contacts (if a client was contacted more than two times) for which the information is lost. Since the total number of contacts may be relevant, it was also saved.

After this analysis, some simple data pre-processing (Data Preparation phase) was performed and then the first execution of rminer algorithms (Modeling phase) was conducted, to get some preliminary rough models. At this stage, only one model was obtained, through the use of the NB algorithm. For DT and SVM, the rminer got out of memory or the processing did not end despite waiting several hours. Simplification (e.g. better data selection) was needed if knowledge was to be extracted successfully.

Second iteration – goal redefinition

One of the hypotheses for the difficulty in obtaining models was the high number of possible output values, i.e. class labels. With this in mind, in the Business Understanding we transformed the output into a binary task, by using only the conclusive results of Table 1: successful and unsuccessful. It should be noted that for all the other results, there is always an uncertainty about client's real intentions regarding the contact offer. Hence, the non-conclusive instances were discarded, leading to a total of 55817 contacts (the same 6499 successes).

After this goal redefinition, we were capable of testing the NB and DT algorithms in the R/rminer tool in the Modeling phase. However, there was still a large number of inputs to be considered (58), missing data (not handled yet), thus the predictive performances could be improved in another CRISP-DM round.

Third iteration – variable and instance selection

We first assumed that there were several irrelevant input attributes that difficult the DM algorithm learning process (e.g. by increase of noise). To test this hypothesis, we went back to the Data Understanding phase and analyzed which attributes could influence the target. For this purpose, the rattle tool was used, in particular its graphical capabilities. For example, Figure 2 shows that the Sex attribute can be discarded, since the rate of successes for Male and Female is almost the same. With a similar analysis, we deleted half (29) of the considered inputs, leading to 29 input variables and 1 target output. We should note that while this graphical analysis may have flaws (e.g. there could be interactions of a particular Sex value with other attributes), the results of Table 3 do back this approach.

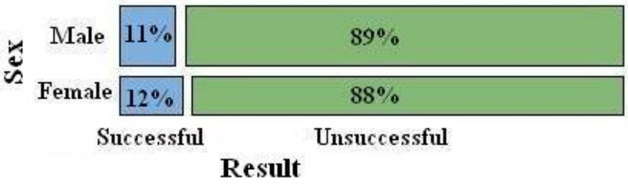


Figure 2 Influence of Sex in the contact target

Preliminary experiments, using NB and removing the attributes suggested by the Rattle analysis, also backed our manual feature selection procedure.

In addition to input attribute reduction, there were also several instances with missing values that were dealt during the Data Preparation phase. While some DM models (e.g. DT) work well with missing data, there are others (e.g. SVM) that require missing data substitution or deletion. Since we had a large dataset, we opted to discard the examples that contained missing values, leading to a dataset with 45211 instances (5289 of which were successful – 11.7% success rate).

In this third iteration of CRISP-DM, during the Modeling phase, we managed to successfully test the three DM algorithms (i.e. NB, DT and SVM). Also it is worth to notice that in all CRISP-DM iterations, the models were validated by using a holdout split, where the whole dataset was randomly divided into training (2/3) and test (1/3) sets. Given the large number of instances, 2/3 of them were considered good enough to build the models. To get more robust estimates of the performances, we applied 20 runs for each DM model in the second and third CRISP-DM iterations.

Results

Table 3 shows the predictive results for the test data during the three CRISP-DM iterations. The results are shown in terms of the mean value of the runs considered.

The AUC plots the False Positive Rate (FPR) versus the True Positive Rate (TPR) and allows identifying how good is the class discrimination: the higher the better, with the ideal model having a value of 1.0.

Table 3 Predictive metrics for all the DM algorithms and CRISP-DM iterations

CRISP-DM Iteration	**1st**	**2nd**		**3rd**		
Instances × Attributes (Nr. Possible Results)	79354×59 (12)	55817 × 53 (2)		45211 × 29 (2)		
Algorithm	NB	NB	DT	**NB**	**DT**	**SVM**
Number of executions (runs)	1	20	20	20	20	20
AUC (Area Under the ROC Curve)	0.776	0.823	0.764	**0.870**	**0.868**	**0.938**
ALIFT (Area Under the LIFT Curve)	0.687	0.790	0.591	**0.827**	**0.790**	**0.887**

Overall, the results show that there was a clear evolution in prediction capabilities for models obtained, with each iteration resulting in better models than the previous one. In particular, the best predictive model is SVM, which provides a high quality AUC value, higher than 0.9.

To complement the ROC analysis, we also used the cumulative Lift curve area. Similar to the AUC, the random baseline classifier produces a 0.5 Area under the Lift curve (ALIFT), and the perfect ALIFT value is 1.0.

Both the ROC and Lift curves are shown in Figure 3 and Figure 4, respectively, for the three models obtained in the third iteration. Lift also favors the SVM model, which provides the higher cumulative lift area and whose curve is always above the remaining methods, whatever the selection of contacts.

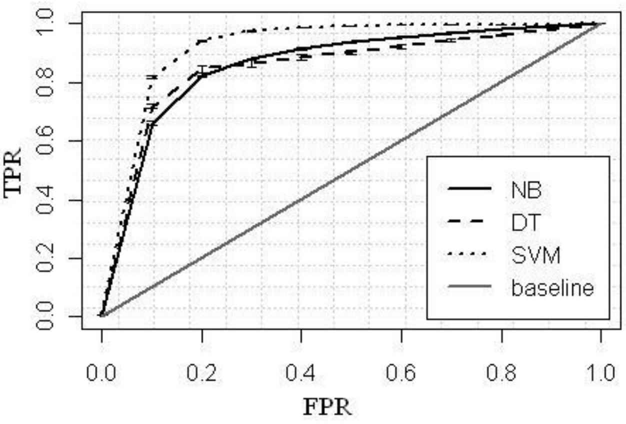


Figure 3 ROC curves for the best predicting models

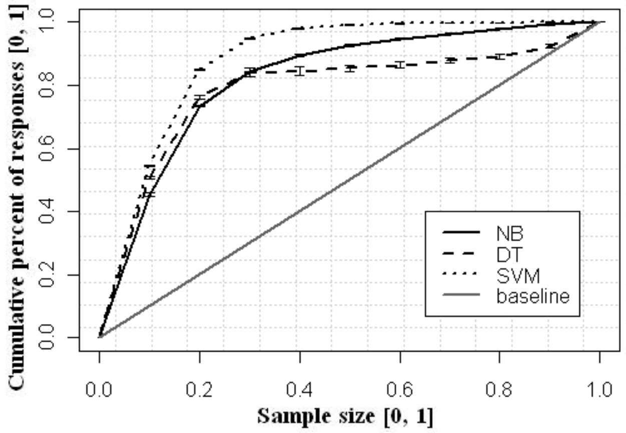


Figure 4 Lift analysis for the best predicting models

The good predictive SVM model cannot be used directly to the contact selection to be loaded into the campaign, since some inputs are related to runtime contact execution, after the campaign has began. Nevertheless, we still believe it is useful to improve marketing campaigns. For instance, using a sensitivity analysis method (Cortez and Embrechts, 2011), we can characterize the SVM inputs influence in success. Figure 5 shows the importance of the five most relevant inputs in the SVM model.

Call duration is the most relevant feature, meaning that longer calls tend increase successes. In second place comes the month of contact. Further analysis can show (Figure 6) that success is most likely to occur in the last month of each trimester (March, June, September and

December). Such knowledge can be used to shift campaigns to occur in those months.

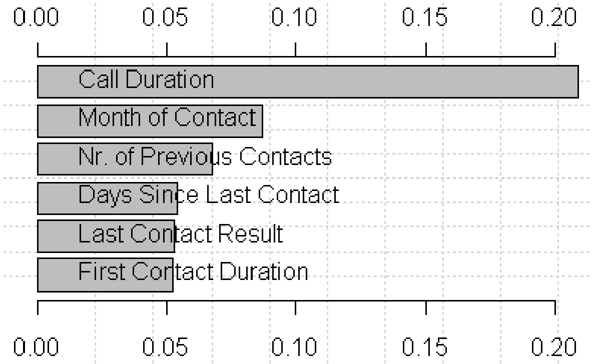


Figure 5 Most relevant inputs of the SVM model

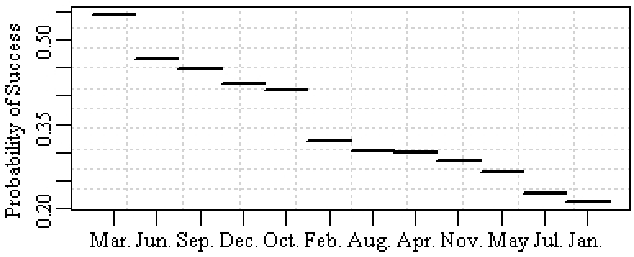


Figure 6 Month's influence on the SVM model

CONCLUSIONS

In this paper, we apply a Data Mining (DM) approach to bank direct marketing campaigns. In particular, we used real-world and recent data from a Portuguese bank and performed three iterations of the CRISP-DM methodology, in order to tune the DM model results. In effect, each CRISP-DM iteration has proven to be of great value, since obtained predictive performances increased. The best model, materialized by a Support Vector Machine (SVM), achieved high predictive performances. Using a sensitivity analysis, we measured the input importance in the SVM model and such knowledge can be used by managers to enhance campaigns (e.g. by asking agents to increase the length of their phone calls or scheduling campaigns to specific months). Another important outcome is the confirmation of open-source technology in the DM field that is able to provide high quality models for real applications (such as the rminer and rattle packages), which allows a cost reduction of DM projects.

In future work, we intend to collect more client based data, in order to check if high quality predictive models can be achieved without contact-based information. We also plan to apply the best DM models in a real setting, with a tighter interaction with marketing managers, in order to gain a valuable feedback.

REFERENCES

Aptéa, C. and Weiss, S. 1997. “Data mining with decision trees and decision rules”, *Future Generation Computer Systems* 13, No.2-3, 197–210.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and 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. and Vapnik, V. 1995. “Support Vector Networks”, *Machine Learning* 20, No.3, 273–297.

Cortez, P. 2010. “Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool”. In *Proceedings of the 10th Industrial Conference on Data Mining* (Berlin, Germany, Jul.). Springer, LNAI 6171, 572–583.

Cortez, P. and Embrechts, M. 2011. “Opening Black Box Data Mining Models Using Sensitivity Analysis”. In *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, No.8, 861–874.

Hu, X. 2005, “A data mining approach for retailing bank customer attrition analysis”, Applied Intelligence 22(1):47-60.

Kohavi, R. and Provost, F. 1998. “Glossary of Terms”, *Machine Learning* 30, No.2–3, 271–274.

Ling, X. and Li, C., 1998. “Data Mining for Direct Marketing: Problems and Solutions”. In *Proceedings of the 4th KDD conference*, AAAI Press, 73–79.

Li, W., Wu, X., Sun, Y. and Zhang, Q., 2010. “Credit Card Customer Segmentation and Target Marketing Based on Data Mining”, In Proceedings of International Conference on Computational Intelligence and Security, 73-76.

Ou, C., Liu, C., Huang, J. and Zhong, N. 2003. “On Data Mining for Direct Marketing”. In *Proceedings of the 9th RSFDGrC conference*, 2639, 491–498.

Page, C. and Luding, Y., 2003. “Bank manager’s direct marketing dilemmas – customer’s attitudes and purchase intention”. *International Journal of Bank Marketing* 21, No.3, 147–163.

Turban, E., Sharda, R. and 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. and Frank, E. 2005. *Data Mining – Pratical Machine Learning Tools and Techniques – 2nd edition*, Elsevier, USA.

Zhang, H. 2004. “The Optimality of Naïve Bayes”, In *Proceedings of the 17th FLAIRS conference*, AAAI Press.

BIOGRAPHY

SÉRGIO MORO (Computer Engineer from Instituto Superior Técnico) is a software engineer at a Portuguese bank and a MSc student at the Instituto Universitário de Lisboa (ISCTE-IUL).

RAUL M. S. LAUREANO is Assistant Professor at the Business School of the Instituto Universitário de Lisboa (ISCTE–IUL), Lisboa, Portugal and researcher at UNIDE (ISCTE-IUL), with research interests including Computer Skills Evaluation and Assessment, and Applied Statistics and Data Mining in Accounting and Management.

PAULO CORTEZ is Associate Professor at the Dep. of Information Systems of the University of Minho and researcher at centro Algoritmi, with interests in the fields of business intelligence and neural networks. His research has appeared in Journal of Decision Support Systems, Neurocomputing, and others (see http://www3.dsi.uminho.pt/pcortez).

View publication stats