BRU-IUL
BUSINESS RESEARCH UNIT (UNIDE-IUL)
Instituto Universitário de Lisboa

Working Paper – 13/06

A Data Mining Approach for Bank Telemarketing Using the rminer Package and R Tool

Sérgio Moro

Paulo Cortez

Raúl M. S. Laureano

ISCTE - Instituto Universitário de Lisboa
BRU-IUL (Unide-IUL)
Av. Forças Armadas
1649-126 Lisbon-Portugal
http://bru-unide.iscte.pt/

FCT Strategic Project UI 315 PEst-OE/EGE/UI0315

A Data Mining Approach for Bank Telemarketing
Using the rminer Package and R Tool

Sérgio Moro (scmoro@gmail.com)
Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal

Paulo Cortez (pcortez@dsi.uminho.pt)
Universidade do Minho, Centro Algoritmi, Dep. Sistemas de Informação, Portugal

Raul M. S. Laureano (raul.laureano@iscte.pt)
Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Lisboa, Portugal

Abstract

Due to the global financial crisis, credit on international markets became more restricted for banks, turning attention to internal clients and their deposits to gather funds. This driver led to a demand for knowledge about client's behavior towards deposits and especially their response to telemarketing campaigns.

This work describes a data mining approach to extract valuable knowledge from recent Portuguese bank telemarketing campaign data. Such approach was guided by the CRISP--DM methodology and the data analysis was conducted using the rminer package and R tool. Three classification models were tested (i.e., Decision Trees, Naïve Bayes and Support Vector Machines) and compared using two relevant criteria: ROC and Lift curve analysis. Overall, the Support Vector Machine obtained the best results and a sensitive analysis was applied to extract useful knowledge from this model, such as the best months for contacts and the influence of the last campaign result and having or not a mortgage credit on a successful deposit subscription.

Keywords

Telemarketing, Direct Marketing, Long-term Deposits, Data Mining, CRISP-DM, Classification Problem, Banking, R.

JEL classification: C88, C38, M31

Introduction

In times of financial crisis, distrust on banks becomes an important issue to consider for these institutions. Suspiciousness leads to withdraws, frozen investments wait for more optimistic scenarios, and credit becomes restricted not only for individual clients and companies, but also between financial institutions (Gerali et al., 2010).

Such a context constitutes by its own a huge driver that potentiates a pursuit for efficiency. The global economic crisis that emerged in 2007 in the United States and spread world-wide, affecting Europe in particular, 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, which led to competition for client's deposits. Moreover, some national banks imposed limits to prevent an uncontrolled increase in the offered rates (Penty, 2011).

Those drivers led retail banks to invest in products and campaigns to gather and retain financial assets by selling long-term deposits to internal clients, taking the advantage of knowing their profile.

There are two main approaches for enterprises to promote products and/or services: through mass campaigns, targeting general indiscriminate public, or direct marketing, targeting a specific set of clients (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, direct 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, direct marketing has some drawbacks, for instance it may trigger a negative attitude towards banks due to the intrusion on privacy (Page and Luding, 2003).

Telemarketing can be defined as marketing conducted through remote communication channels, such as telephone, for targeting a set of selected clients thus allowing choosing

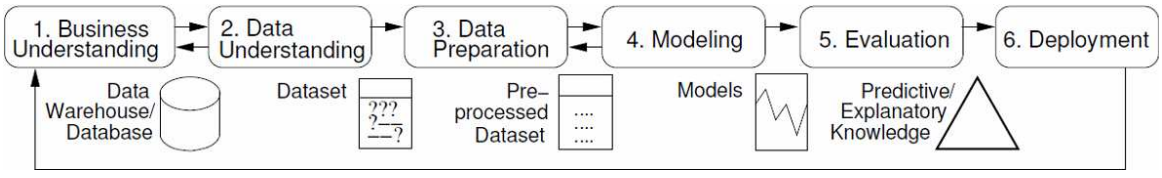
those that supposedly will be keener to acquire the product or service being offered (Tapp, 2008). With the automation of telemarketing through Computer-Telephony-Integration techniques, it became quite common and easy to generate a wide variety of reports from marketing campaigns and so to add-up other types of information available for the organizations.

One effective way of analyzing all those sources of information in order to discover trends and patterns is through Business Intelligence and Data Mining (DM) techniques, to build data-driven models and then extract useful knowledge (Witten and Frank, 2005; Turban et al., 2010).

There are several works that use DM techniques to improve bank marketing campaigns. Ling and Li (1998) address the problem of direct marketing having a very low positive number of responses (approximately 1% for the cases they studied). Li et al. (2010) applied DM techniques to define clusters of clients oriented with the goal of conducting direct marketing campaigns to sell credit cards.

A DM project encompasses all those steps needed to extract useful knowledge and apply or use it to produce some kind of benefit, typically, when concerning business, by improving the ROI (Return-On-Investment). The CRoss-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology for increasing the success of data mining projects (Chapman et al., 2000). The methodology defines a non-rigid sequence of six phases, which allow the building and implementation of a data mining model to be used in a real environment, helping to support business decisions (Figure 1). CRISP-DM defines a project as a cyclic process, where several iterations can be conducted for tuning DM towards business goals.

Figure 1 - The CRISP-DM process model



Source: adapted from Chapman et al. (2000)

The global crisis driver means that investment funds of all types need to be reduced. As far as DM projects are concerned, being IT projects, one of such ways is to adopt open source technologies, such as the R statistical tool. While not directly addressed for DM, the R tool can be extended through the installation of packages and several of these packages implement DM techniques. In particular, rminer facilitates the use of supervised DM tasks (i.e. classification or regression), by offering a small and coherent set of functions (Cortez, 2010). This is a package that has already been applied to distinct domains, such as meat quality (Cortez et al., 2006), intensive care medicine (Silva et al., 2008), wine preferences (Cortez et al., 2009a), spam email detection (Cortez et al., 2009b) and civil engineering (Tinoco et al., 2009). The rminer is available at CRAN (cran.r-project.org/web/packages/rminer) which allows downloading and installing it directly from the R prompt environment.

This paper describes how the problem of understanding success drivers in telemarketing campaigns for selling deposits of a Portuguese bank was addressed using R/rminer tool.

Business Problem

A Portuguese retail bank uses its own contact-center to do direct marketing campaigns, mainly through phone calls (telemarketing). Each campaign is managed in an integrated fashion and the results for all calls and clients within the campaign are gathered together, in a flat file report concerning only the data used to do the phone call. The agents were all human, thus no automatic calls through Interactive Voice Response (IVR) or Voice Response Unit (VRU) were performed, and they had a campaign script that helps to conduct the conversation with the client. The computer application to execute the telemarketing campaigns is in use since the end of 2007, and it performs several tasks, such as launching automatic client calls and producing the reports with the final campaign results, in a nightly batch process.

In this context, in September of 2010, a research project was conducted to evaluate the efficiency and effectiveness of the telemarketing campaigns to sell long-term deposits. The primary goal was to achieve previously undiscovered valuable knowledge in order to redirect managers' efforts to improve campaign results. In other words the objective was, on the one hand, decrease the number of phone calls (efficiency dimension – cost reduction) and, on the other hand, increase or at least not to decrease the total number of deposits subscriptions (effectiveness dimension – retain financial assets from clients for longer periods). The research project was to be conducted within one year, ending by the mid/end of 2011. Since this project started being analyzed in detail in September of 2010, it meant that there were available reports for about three years of telemarketing campaigns, encompassing already the effects of the global financial crisis and banks' responses to it in order to improve deposits subscriptions.

Data Extraction

Data was collected mainly through the reports of previously executed campaigns. Since the telemarketing application reports were improved right shortly after they started to be produced and stabilized in the available information by April 2008, data collected included 17 campaigns executed between May 2008 and November 2010, corresponding to a total of 79354 contacts and 58 attributes to characterize each of them (Table 1). During these phone campaigns, an attractive long-term deposit application was offered.

As stated previously, the reports contained only information used in contact execution, meaning that they had attributes related specifically to the contact and also some client information needed to conduct the dialogue (such as the name, to introduce the conversation, e.g., "Good evening, Mr. John. I am calling from Bank…").

Considering the information available, there urged a need for other types of attributes regarding other information features, like specific bank client information (e.g., average account balance) and personal information (e.g., age at the date of the phone call).

Table 1 - Attributes

Name	Description	Type	Missing values (NA=not available)
Personal information (11 attributes)			
Age	Years at date of contact	Numeric	- - -
Profession	1727 enumerated possible values	Nominal	11470 unknown and 5597 NA
Employment status	27 enumerated possible values	Nominal	9144 unknown and 2 NA
Marital status	Married, Divorced, Separated, Single and Widowed	Nominal	65 unknown
Parish	Residence address	Nominal	4210 NA
Zip code		Nominal	247 NA
Zip code town		Nominal	247 NA
County		Nominal	1923 NA
Honorific title	22 enumerated possible values	Nominal	23 NA
Sex	Male/Female	Nominal	- - -
Educational qualifications	Basic, Primary, Secondary, University degree, unknown	Nominal	1319 unknown and 27 NA
Bank client information (13 attributes)			
Generic block	Triggers a generic block usage for bank clerks to be aware of (Yes/No)	Nominal	- - -
Informational block	Triggers an informational alert for bank clerks (Yes/No)	Nominal	- - -
Check block	Triggers an alert once the client tries to use checks (Yes/No)	Nominal	- - -
Check inhibition	Inhibits check usage (Yes/No)	Nominal	- - -
Bank associated	If the client is bank associated (Yes/No)	Nominal	- - -
Loans in delay	Client has loans in delay (Yes/No)	Nominal	- - -
Annual balance	Average annual current account balance	Numeric	7537 NA
Debt card	Client has a debit card (Yes/No)	Nominal	- - -
Salary account	Client has a salary account (Yes/No)	Nominal	- - -
Credit card	Client has a credit card (Yes/No)	Nominal	- - -
Mortgage credit	Client has a mortgage credit (Yes/No)	Nominal	- - -
Individual credit	Client has an individual credit (Yes/No)	Nominal	- - -
Domiciliation of payments	Client holds direct debits payments using domiciliation (Yes/No)	Nominal	- - -
Generic contact information (1 attribute)			
Number of calls	Number of phone calls for contacts that required more than one call until the client finally decides to subscribe or not	Numeric	- - -
Information for the last call (7 attributes)			
Agent ID	Agent identification for who made the call	Nominal	- - -
Phone type	Landline/Mobile	Nominal	10783 unknown
Day of week	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday	Nominal	- - -
Day of month	From 1 to 31	Nominal	- - -
Month	From 1 to 12 (January to December)	Nominal	- - -
Hour	From 7 to 22 (24 hour clock)	Nominal	- - -
Duration	In seconds	Numeric	- - -

Name	Description	Type	Missing values
Information for the first call (12 attributes)			
Result	Result of the call – one of the scheduled values presented ahead in Table 2	Nominal	17069 unknown
Agent ID	Agent identification for who made the call	Nominal	
Phone type	Landline/Mobile	Nominal	
Day of week	Monday to Sunday	Nominal	
Day of month	From 1 to 31	Nominal	
Month	From 1 to 12 (January to December)	Nominal	
Hour	From 7 to 22 (24 hour clock)	Nominal	
Scheduled day of week	Information for cases when the client decided to schedule the contact for another date	Nominal	
Scheduled day of month		Nominal	
Scheduled month		Nominal	
Scheduled hour		Nominal	
Duration	In seconds	Numeric	
Web page hits information (4 attributes)			
Phone banking date	Last web page hit for phone banking agent	Date	37819 NA
Number of hits	Number of hits for phone banking agent	Numeric	
Home banking date	Last web page hit for home banking	Date	39029 NA
Number of hits	Number of hits for home banking	Numeric	
Historic information (10 attributes)			
Days since first contact	Number of days since first contact for any other deposit campaign	Numeric	- - -
Days since last contact	Number of days since first contact for any other deposit campaign	Numeric	- - -
Previous contacts	Previous contacts for other campaigns	Numeric	- - -
Previous successes	Successful previous contacts	Numeric	- - -
Previous failures	Unsuccessful previous contacts	Numeric	- - -
Last result	Last campaign contact result	Nominal	45462 NA
Last result value	Last campaign contact amount subscribed if a successful contact	Numeric	- - -
Total value	Total previous subscriptions value	Numeric	- - -
Total phone page hits	Number of hits for phone banking agent for previous campaigns	Numeric	- - -
Total home banking page hits	Number of hits for home banking for previous campaigns	Numeric	- - -

Additionally, more recent campaigns could benefit from previously executed campaigns knowledge for each specific client. Moreover, some hypotheses arose: perhaps if the client previously subscribed a deposit he maintains the interest in our offer, or maybe he prefers to try another deposit from the competition.

Furthermore, 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.

Since all the data were available in the form of reports and files with specific client information, we needed to join all those together, as a base dataset to conduct the DM project.

Data Exploration and Preprocessing

At the initial stage, there is a need to explore the dataset supplied. First of all, the main goal is to know the outcome of a telemarketing campaign call to sell a deposit: the client subscribed it or not. Hence, at a first glance, the outcome to predict is a nominal value making this a classification problem.

Analyses of the dataset allowed identifying the names of all the attributes and confirmed information supplied relating to the dataset, including the name of the outcome attribute. Thus, an exploration of the outcome attribute is mandatory to understand the effectiveness of the campaigns.

The possible outcomes for the dataset are grouped in Table 2. Unlike the expected, there are 12 different final results for a contact. There are two main reasons that account for this situation. First, in some cases the client could not be contacted at all, corresponding to the "Cancelled contact" group (perhaps the phone number was wrong, or the phone device was disconnected every time a call was attempted). The other cases belong to the "Scheduled contact" group, in which a client keeps postponing an answer and asks to be contacted later, and meanwhile the campaign ends, leading to a final scheduled result, where a call is scheduled but is never executed since the campaign already ended.

Considering that everything else besides successful contacts are unsuccessful, since the client did not subscribed the deposit, there are about 8.19% of successes from the total of 79354 contacts.

Table 2 - Enumerated values for the final call result

Contact result	#	Group	#
Success (subscribe the deposit)	6 499	Concluded contact	55 817
Failure (reject the offer)	49 318		
Not the owner of the phone	1 011	Cancelled contact	8 365
Did not answer	5 091		
Fax instead of phone	151		
Abandoned call	2 059		
Aborted by the agent	53		
Scheduled by other than the client	9 640	Scheduled contact	15 172
Scheduled by the client himself	622		
Scheduled – deposit presented to the client	2 916		
Scheduled – deposit not presented	1 763		
Scheduled due to machine answer	231		
Total:	79 354		

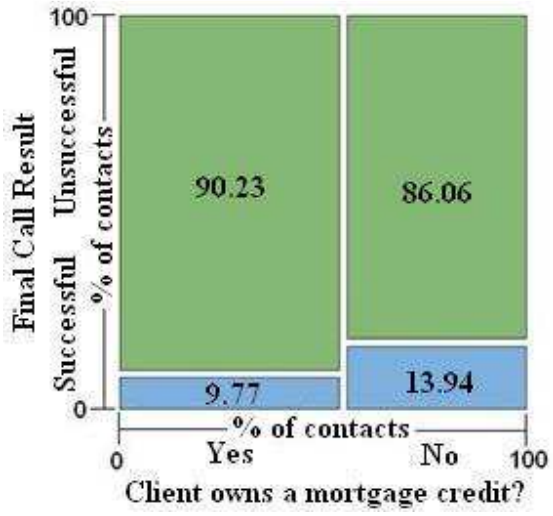
For this project, we tested three DM classifiers: Naïve Bayes (NB) (Zhang, 2004), Decision Trees (DT) (Aptéa and Weiss, 1997) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995). In the first CRISP-DM iteration, initial attempts to run the rminer *mining* function with the DT and SVM did not produce the desired result (i.e. the rminer got out of memory or the processing did not end despite waiting for several hours). One of the hypotheses for such difficulty was the high number of possible output values, i.e. class labels (Table 2). With this in mind, we went back to the Business Understanding phase of CRISP-DM (Figure 1), constituting a second iteration of the cycle. We opted to redefine the output as a binary task by using only the conclusive results of Table 2: success and failure. It should be noted that for all the other results, there is always an uncertainty about the client's real intentions regarding the contact offer. Hence, the non-conclusive contacts were discarded, leading to a total of 55817 contacts (the same 6499 successes).

At the second CRISP-DM iteration, we were able to test NB and DT methods but not SVM (due to computational memory problems). Since there was a large number of inputs (58) and missing data, we addressed these issues in the data preparation phase. For variable selection, we adopted the *rattle* tool, in particular its graphical capabilities. *Rattle* is a graphical user interface that runs on the R environment and which facilitates DM analysis

through visualization techniques by plotting input attributes possible values versus the corresponding outcomes for every record in the dataset (Williams, 2009). While these graphics do not add new information since they are based on simple statistics that count the possible inputs of an attribute for each outcome, they add visual information that can easier be identified by humans (Martinez, 2011).

For all attributes, graphics were plotted that related each of them with the target attribute (the final call result). For some input variables, we identified a clear difference in the histogram frequencies related with call result outcome. For instance, Figure 2 shows that not owning a mortgage credit increases the probability of a successful contact. This visual analysis paved the way for a backward elimination (Guyon and Elisseeff, 2003), by providing guidance on which attributes to remove. Thus, for each attribute considered, it was removed from the model achieved with NB at the first CRISP-DM iteration, and measurements of prediction capabilities through accuracy (which contact results were correctly classified) allowed deciding if the attribute should be eliminated from the dataset.

Figure 2 - Influence of housing /owning a mortgage credit on the final call result



In addition to input attribute reduction, there were also several contacts with missing values. While some DM algorithms (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 contacts that contained missing values, leading to a dataset with 45211

contacts (5289 of which were successful – 11.7% success rate). The remaining attributes were logically grouped according to the type of information, resulting in 29 explanatory attributes (Table 3).

Table 3 - Explanatory attributes used in the modeling phase

Group	Name	Description and Values	
Personal Client Information	age	In years and at the date of the last contact made (Mean=40.9, SD=10.6)	
	job	27 possible values (Mode="specialized operative")	
	marital	Married (60%), Divorced (10%), Separated (1%), Single (28%), Widowed (1%)	
	title	22 possible values (Mode= "Without specific title")	
	education	10 possible values (Mode="University degree")	
Bank Client Information	default	Indicates that the client has loans in delay – (Y)es=98%/(N)o=2%	
	balance	Average annual balance of all the current accounts that the client owns (Mean=1 362, SD=3 045)	
	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%	
	housing	Indicates that the client has a mortgage account – Y=56%/N=44%	
	loan	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	campaigns	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	
	contact	Mobile=34%,Fixed=4%,NA=62 %	Mobile=65%,Fixed=29%,NA=6%
	day	*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	pdays	Number of days since the last contact for any other campaign (Mean=41, SD=100)	
	previous	Total number of previous contacts (Mean=0.58, SD=2.30)	
	poutcome	Result for the last campaign (81.7% did not have been yet contacted)	

Modeling

In this work, we explore three distinct DM classifiers: NB, DT and SVM. The last model was more recently proposed (Cortes and Vapnik, 1995; Hearst et al., 1998). When compared with the previous methods (NB and DT), SVM tends to produce higher predictive performances (Wu et al., 2008). However, the interpretation of the data-driven SVM model is not intuitive for business managers (Witten and Frank, 2005). Nevertheless, rminer uses a sensitive analysis procedure to assess input attribute importance and characterize the average influence in target output (Cortez and Embrechts, 2013). Such procedure is based on measuring the effects on the output of a fitted model when one attribute is varied through its domain of values and other attributes are fixed at their average values. The SVM adopted by rminer uses a Gaussian kernel and a simple grid search for setting the SVM hyperparameters (Cortez, 2010).

To evaluate a classification model, popular metrics are based on the confusion matrix (Kohavi and Provost, 1998) and the Receiver Operating Characteristic (ROC) curve (Fawcett, 2005). Another 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 cases, the ones with the most likely positive answers estimated by the model (Coppock, 2002).

At the third iteration of the CRISP-DM methodology, the dataset consisted of 45211 samples (i.e. client contacts), that maps 29 explanatory variables into a binary target. To create and test models, the rminer provides a single function, *mining*, that can be parameterized in order to obtain the desired model. The R code executed for the NB was:

```
# 2/3 of contacts for training and the remaining 1/3 for testing
MNB3=mining(y~.,DF,method=c("holdout",2/3), model="naivebayes",
     Runs=20)
# save the MNB object for future use
savemining(MNB3, "mining_nb.output",ascii=TRUE)
```

As an example, we present also the *mining* command used for SVM:

```
MSVM3=mining(y~.,DF,method=c("holdout",2/3), model="svm",Runs=20)
```

The simplicity of rminer commands usage is evident: with the same command, through extensive parameterization, one can access a wide variety of combinations for modeling/validation techniques.

With the command *mmetric* also from rminer, it is possible to obtain several evaluation metrics directly from the structure returned by *mining*. Some of them are the confusion matrix, the accuracy and the true positive rate, the Receiver Operating Characteristic (ROC) curve and the Lift curve. For the experiments with the *mmetric* command, the target class of a Successful contact (deposit subscription) was considered. The results of the *mmetric* commands executed for each model are shown on Table 4 and the NB R code is:

```
MNB3=loadmining("mining_nb.output") # read MNB3 from file
# TC = Target Class (2=Successful contact)
print(mmetric(MNB3,metric="CONF",TC=2)) # confusion matrix
print(mmetric(MNB3,metric="AUC",TC=2)) # area under the ROC curve
print(mmetric(MNB3,metric="ACC",TC=2)) # accuracy
print(mmetric(MNB3,metric="TPR",TC=2)) # true positive rate
print(mmetric(MNB3,metric="ALIFT",TC=2)) # area under Lift curve
```

Table 4 - Metrics for the third CRISP-DM iteration

NB

Observed \ Predicted	Unsuccessful	Successful
Unsuccessful	235 740	30 420
Successful	10 512	24 748

AUC = 0.870
ACC = 0.864
TPR = 0.702
ALIFT = 0.827

DT

Observed \ Predicted	Unsuccessful	Successful
Unsuccessful	256 783	9 377
Successful	19 136	16 124

AUC = 0.868
ACC = 0.905
TPR = 0.457
ALIFT = 0.790

SVM

Observed \ Predicted	Unsuccessful	Successful
Unsuccessful	258 242	7 918
Successful	19 233	16 077

AUC = 0.938
ACC = 0.910
TPR = 0.455
ALIFT = 0.887

The results are shown in terms of the average of all runs, except the Confusion Matrix, which includes the sum of all runs. The metrics shown in the table are: Area Under the ROC curve (AUC), confusion matrix accuracy (ACC), True Positive Rate (TPR), Area under the Lift accumulative curve (ALIFT).

Model Evaluation

At this stage, there was a need to do a more in-depth evaluation of the predictive models in order to consider its value for business. Both the values obtained for AUC and ALIFT can help to check if some model is better on some aspect (overall discriminatory power for the ROC analysis and selecting the most likely buyers in case of the Lift). The rminer function *mgraph* can be used to plot both the ROC and Lift curves. Furthermore, it is possible to plot several curves in the same graphic by passing a vector of structures. The R code used to plot both the ROC and Lift curves for the three DM models obtained in the third CRISP--DM iteration is presented here:

```
# Results for the third CRISP-DM iteration saved in:
# MNB3 – NB; MDT3 – DT and; MSVM3 – SVM:
L=vector("list",3); L[[1]]=MNB3; L[[2]]=MDT3; L[[3]]=MSVM3
mgraph(L,graph="ROC",TC=2,leg=list(pos=c(0.65,0.55),
      leg=c("NB","DT","SVM")), baseline=TRUE, Grid=15,
      main="ROC curves") # ROC graph

L=vector("list",3); L[[1]]=MNB3; L[[2]]=MDT3; L[[3]]=MSVM3
mgraph(L,graph="LIFT",TC=2,leg=list(pos=c(0.65,0.4),
      leg=c("NB","DT","SVM")), baseline=TRUE, Grid=15,
      main="Lift curves") # Accumulative Lift graph
```

The plots generated by these commands are shown on Figures 3 and 4. In both ROC and Lift curve analysis, SVM presents the best predictive results.

Figure 3 - ROC Curves for the models Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM)

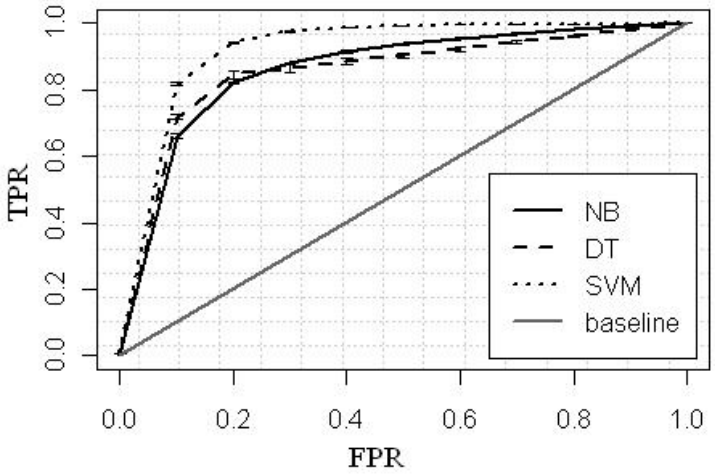


Figure 4 - Lift Curves for the models Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM)

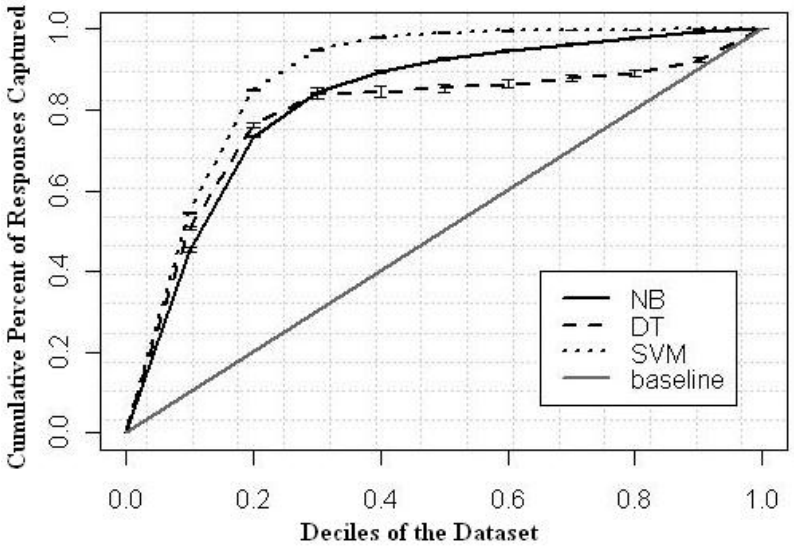


Table 5 compares classification results for test set predictions of the multiple models obtained across the three iterations. In particular, a comparison of the three NB models shows a clear improvement in its prediction capabilities, thus justifying the three CRISP--DM iterations carried out.

Table 5 - Predictive metrics for all the DM algorithms and CRISP-DM iterations

CRISP-DM Iteration	**1st**	**2nd**		**3rd**		
Instances × Attributes (Nr. Possible Results)	79 354×58 (12)	55 817 × 58 (2)		45 211 × 29 (2)		
Algorithm	NB	NB	DT	**NB**	**DT**	**SVM**
Number of executions (runs)	1	20	20	20	20	20
Area Under the ROC Curve	0.776	0.823	0.764	**0.870**	**0.868**	**0.938**
Area Under the LIFT Curve	0.687	0.790	0.591	**0.827**	**0.790**	**0.887**

Overall, at the third CRISP-DM iteration, SVM obtained the best AUC (0.938) and ALIFT (0.887) values. In the next section, we describe what knowledge can be extracted from this model.

Findings

Complex data-driven models, such as SVM, can be more easily understood by humans by adopting a sensitive analysis procedure (Cortez and Embrechts, 2013). To simplify the analysis, we first fitted SVM using the full dataset, through the *fit* rminer function, and then performed the sensitivity analysis, using the rminer *Importance* function:

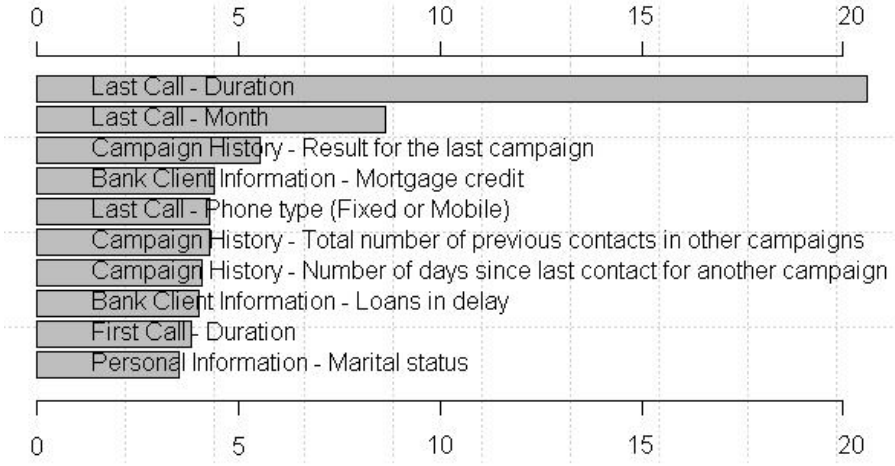
```
# DF denotes the whole dataset
M=fit(y~,DF,model="svm")
I=Importance(M, DF)
```

Using the returned *I* object, the *mgraph* function can be invoked to show the sensitivity analysis input importance bar plot, sorted from most to least important:

```
S=sort.int(I$imp,decreasing=TRUE,index.return=TRUE)
N=10 # choose the 10 most relevant attributes
L=list(runs=1,sen=t(I$imp[S$ix[1:N]]))
LEG=names(DF)
mgraph(L,graph="IMP",leg=LEG[S$ix[1:N]],col="gray",Grid=10)
```

A sensitivity analysis measures how a model is influenced by each of its input attributes in percentage of the remaining. In this way, it is possible to quantify the contribution of a given attribute for the model. The 10 most relevant input attributes (in percentage) are shown in Figure 5.

Figure 5 - The 10 most relevant attributes (in percentage) for the best model



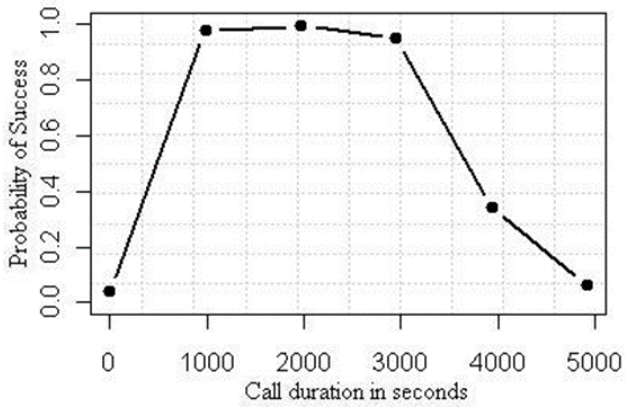
The two most important attributes are related to "Last Call". This emphasizes how important the runtime execution call information is. To get more input influence details, in Figures 6 to 9 we plot the variable effect characteristic (VEC) curve, which shows the average influence of a given attribute (x-axis) on the model probability of success (y-axis) (Cortez & Embrechts, 2013). This can be achieved through the rminer *vecplot* function, with the commands bellow:

```
# 6 = Last Call - Duration
vecplot(I,graph="VEC",xval=6,main="Call Duration Relevance",
Grid=10,TC=2,sort="decreasing") # result on Figure 6

# 4 = Last Call - Month
vecplot(I,graph="VEC",xval=4,main="Last Call Month",
Grid=10,TC=2,sort="decreasing") # result on Figure 7
```

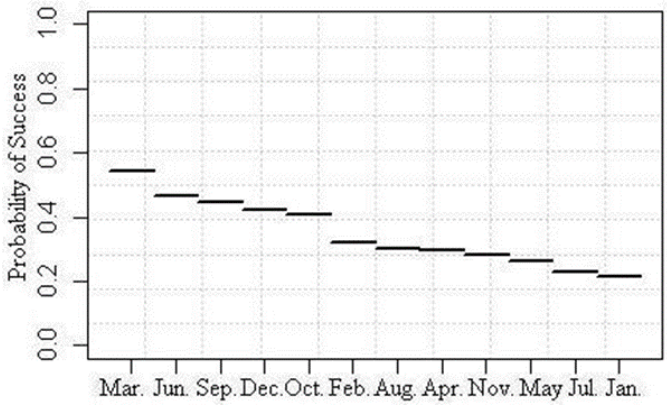
Figure 6 shows that the call duration by its own explains more than 20% of the success. This result makes sense, since a successful sell requires a deeper dialog to describe the product (and maybe create empathy with the client). However, as seen in Figure 6, after a certain threshold (3000 seconds, or 50 minutes), the probability of success starts to decrease (suggesting the client is only trying to be sympathetic but does not want to buy the product).

Figure 6 - Influence of the last call duration on success



Similarly, an analysis of the month's influence (Figure 7) reveals that a success is more likely to occur in the last month of each quarter (March, June, September and December). This can be valuable knowledge, since managers can try to shift campaigns for those specific months.

Figure 7 - Influence of the month the last contact was executed on the success



Regarding the result for the last campaign, it is clear from Figure 8 that a previous success increases the chances of performing a successful call. For the mortgage credit ownership, the influence is rather small (Figure 9). Nevertheless, not having that type of credit increases the chances of having a success.

Figure 8 - Influence of the last campaign result on success

(NA - no previous contact was made)

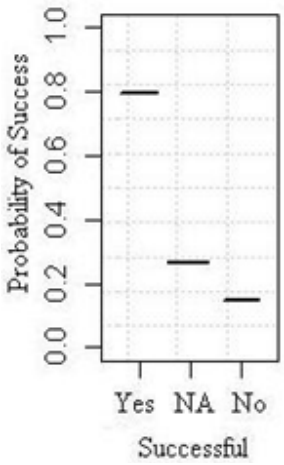
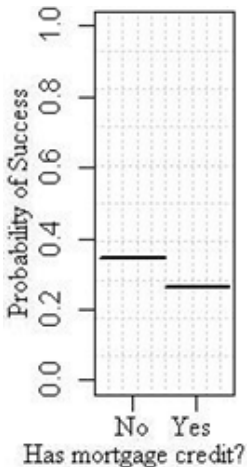


Figure 9 - Influence of having a mortgage credit on success



Lessons and Discussions

In bank direct marketing, reports results from executed telemarketing campaigns can be used to identify trends of the client's behavior. Managers are shifting from traditional statistics analysis towards more sophisticated DM techniques, in order to extract useful knowledge from raw data.

The real-world application case here presented shows that it is possible to analyze bank telemarketing data with open source tools, in particular, by using the R environment and rminer package. This DM project was guided by the CRISP-DM methodology, under three iterations. Three DM techniques were explored: NB, DT and SVM. Overall, for both ROC and Lift curve analysis, SVM obtained the best predictive results.

We also have shown how human understandable knowledge can be extracted from such SVM model through a sensitive analysis. An analysis of input attributes relative relevance for the model can provide guidance on which are the most relevant for the business the model is targeting at. Furthermore, by plotting an input attribute range of values versus the probability of the desired outcome, valuable knowledge can be extracted which contributes for a better understanding of the business. Such knowledge can be used to enhance future marketing campaigns.

References

Aptéa, C. and Weiss, S. (1997). Data mining with decision trees and decision rules, In "Future Generation Computer Systems", Vol. 13, No.2-3, pp. 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, In "Information Management Online, 5329–1[Online; accessed 19-August-2013]. URL http://www.information-management.com/news/5329-1.html.

Cortes, C. and Vapnik, V. (1995). Support Vector Networks, In "Machine Learning", Vol. 20, No.3, pp. 273–297.

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner (Ed.), Advances in Data Mining - Applications and Theoretical Aspects, In "Proceedings of the 10th Industrial Conference on Data Mining", LNAI 6171, pp. 572–583, Berlin, Germany, July, 2010. Springer.

Cortez, P. and Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. In "Information Sciences", Vol. 225, pp. 1–17.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009a). Modeling wine preferences by data mining from physicochemical properties, In "Decision Support Systems", Vol. 47, No.4, pp. 547–553.

Cortez, P., Lopes, C., Sousa, P., Rocha, M. and Rio, M. (2009b). Symbiotic Data Mining for Personalized Spam Filtering. In "Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence" (WI 2009), pp. 149–156. IEEE, Los Alamitos

Cortez, P., Portelinha, M., Rodrigues, S., Cadavez, V. and Teixeira, A. (2006). Lamb Meat Quality Assessment by Support Vector Machines. In "Neural Processing Letters", Springer, Vol. 4, No.1, 41-51, 2006. ISSN:1370-4621.

Fawcett, T. (2005). An introduction to ROC analysis, In "Pattern Recognition Letters", Vol. 27, No.8, pp. 861–874.

Gerali, A., Neri, S., Sessa, L. and Signoretti, F. (2010). Credit and Banking in a DSGE Model of the Euro Area. In "Journal of Money, Credit and Banking" (Wiley), Vol. 42(1), pp. 107–141. Blackwell Publishing Inc, The Ohio State University.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, The Journal of Machine Learning Research, Vol. 3, pp. 1157-1182.

Hearst, M., Dumais, S., Osman, E., Platt, J. and Scholkopf, B. (1998). Support Vector Machines, In "IEEE Intelligent Systems", Vol. 13, No.4, pp. 18–28.

Hodgson, G. (2009). The Great Crash of 2008 and the Reform of Economics, Cambridge In "Journal of Economics", Vol.33, No.6, pp. 1205–1221.

Kohavi, R. and Provost, F. (1998). Glossary of Terms, "Machine Learning", Vol. 30, No.2–3, pp. 271–274.

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", pp. 73-76, Lecture Notes in Computer Science, Springer.

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

Martinez, W. L. (2011). Graphical user interfaces. WIREs Computational Statistics, Vol. 3, pp. 119–133.

Ou, C., Liu, C., Huang, J. and Zhong, N. (2003). On Data Mining for Direct Marketing. In "Proceedings of the 9th RSFDGrC Conference", Vol. 2639, pp. 491–498.

Page, C. and Luding, Y., (2003). Bank manager's direct marketing dilemmas – customer's attitudes and purchase intention. In "International Journal of Bank Marketing", Vol. 21, No.3, pp. 147–163.

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.

Silva, A., Cortez, P., Santos, M., Gomes, L. and Neves, J. (2008). Rating Organ Failure via Adverse Events using Data Mining in the Intensive Care Unit. In "Artificial Intelligence in Medicine", Elsevier, Vol. 43, No.3, pp. 179-193. ISSN:0933-3657.

Tapp, A. (2008). Introducing direct marketing. In "Principles of direct and database marketing – a digital orientation", 4th edition, pp. 1–52, Prentice Hall, USA.

Tinoco, J., Correia, A. G. and Cortez, P. (2009). A Data Mining Approach for Jet Grouting Uniaxial Compressive Strength Prediction. In "World Congress on Nature and Biologically Inspired Computing" (NaBIC 2009), Coimbatore, India, December 2009, pp. 553–558. IEEE, Los Alamitos.

Williams, G. (2009). Rattle: a data mining GUI for R, In "The R Journal", Vol. 1, No.2, pp. 45-55.

Witten, I. and Frank, E. (2005). "Data Mining – Practical Machine Learning Tools and Techniques", 2nd edition, Elsevier, USA.

Wu, X., Kumar, V., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B. and Yu, P. (2008). Top 10 algorithms in data mining, In "Knowledge and Information Systems", Vol. 14, No.1, pp. 1-37, DOI: 10.1007/s10115-007-0114-2.

Turban, E., Sharda, R. and Delen, D. (2010). "Decision Support and Business Intelligence Systems", 9th edition, Prentice Hall Press, USA.

Zhang, H. (2004). The Optimality of Naïve Bayes, In "Proceedings of the 17th FLAIRS conference", Vol. 2, No. 3, pp. 3-9. AAAI Press.