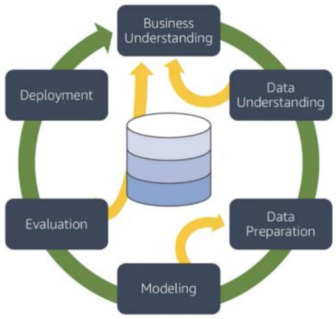


1. Data
2. Business Understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

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.

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.

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	

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	

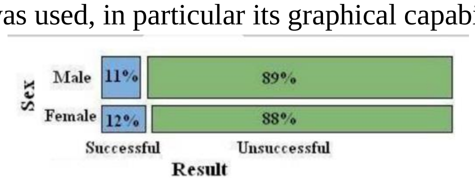
Second iteration – goal redefinition

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



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**

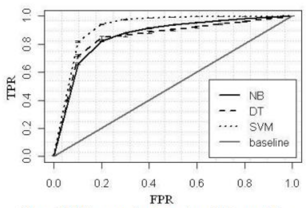
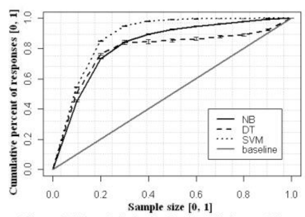


Figure 3 ROC curves for the best predicting models



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.

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.

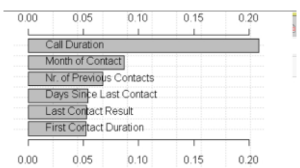
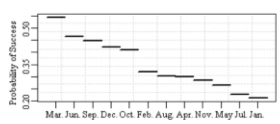


Figure 5 Most relevant inputs of the SVM model



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).