ISCTE IUL

Instituto Universitário de Lisboa

Department of Information Science and Technology

Feature Selection Strategies for Improving Data-Driven Decision Support in Bank Telemarketing

Sérgio Miguel Carneiro Moro

A Thesis presented in partial fulfillment of the Requirements for the Degree of

Doctor of Information Science and Technology

Supervisors:

Doctor Paulo Cortez, PhD, Associate Professor,
University of Minho

Doctor Paulo Rita, PhD, Full Professor,
ISCTE – University Institute of Lisbon

April, 2015

ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

Feature Selection Strategies for Improving Data-Driven Decision Support in Bank Telemarketing

Sérgio Miguel Carneiro Moro

A Thesis presented in partial fulfillment of the Requirements for the Degree of

Doctor of Information Science and Technology

Jury:

Doctor Luís Ducla Soares, PhD, Assistant Professor (President), ISCTE – IUL
Doctor Manuel Filipe Santos, PhD, Associate Professor, Universidade do Minho
Doctor Pedro Quelhas Brito, PhD, Assistant Professor, Universidade do Porto
Doctor Nuno Marques, PhD, Assistant Professor, Universidade Nova de Lisboa
Doctor Raul Laureano, PhD, Assistant Professor, ISCTE – IUL
Doctor Paulo Cortez, PhD, Associate Professor, University of Minho
Doctor Paulo Rita, PhD, Full Professor, ISCTE – University Institute of Lisbon

April, 2015

Abstract

The usage of data mining techniques to unveil previously undiscovered knowledge has been applied in past years to a wide number of domains, including banking and marketing. Raw data is the basic ingredient for successfully detecting interesting patterns. A key aspect of raw data manipulation is feature engineering and it is related with the correct characterization or selection of relevant features (or variables) that conceal relations with the target goal.

This study is particularly focused on feature engineering, aiming at the unfolding features that best characterize the problem of selling long-term bank deposits through telemarketing campaigns. For the experimental setup, a case-study from a Portuguese bank, ranging the 2008-2013 year period and encompassing the recent global financial crisis, was addressed. To assess the relevance of such problem, a novel literature analysis using text mining and the latent Dirichlet allocation algorithm was conducted, confirming the existence of a research gap for bank telemarketing.

Starting from a dataset containing typical telemarketing contacts and client information, research followed three different and complementary strategies: first, by enriching the dataset with social and economic context features; then, by including customer lifetime value related features; finally, by applying a divide and conquer strategy for splitting the problem in smaller fractions, leading to optimized sub-problems. Each of the three approaches improved previous results in terms of model metrics related to prediction performance. The relevance of the proposed features was evaluated, confirming the obtained models as credible and valuable for telemarketing campaign managers.

Keywords: telemarketing, banking, bank deposits, decision support, decision making, feature selection, feature engineering, data mining.

Resumo

A utilização de técnicas de *data mining* para a descoberta de conhecimento tem sido aplicada nos últimos anos a uma grande variedade de domínios, incluindo banca e *marketing*. Os dados no seu estado primitivo constituem o ingrediente básico para a deteção de padrões de informação. Um aspeto chave da manipulação de dados em bruto consiste na "engenharia de atributos", que compreende uma correta definição e seleção de atributos relevantes (ou variáveis) que se relacionem com o alvo da descoberta de conhecimento.

Este trabalho foca-se numa abordagem de "engenharia de atributos" para definir as variáveis que melhor caraterizam o problema de vender depósitos bancários a prazo através de campanhas de *telemarketing*. Sendo um estudo empírico, foi utilizado um caso de estudo de um banco português, abrangendo o período 2008-2013, que inclui os efeitos da crise financeira internacional. Para aferir da importância deste problema, foi realizada uma inovadora análise da literatura recorrendo a *text mining* e ao algoritmo *latent Dirichlet allocation*, confirmando a existência de uma lacuna nesta matéria.

Utilizando como base um conjunto de dados de contactos de *telemarketing* e informação sobre os clientes, três estratégias diferentes e complementares foram propostas: primeiro, os dados foram enriquecidos com atributos socioeconómicos; posteriormente, foram adicionadas características associadas ao valor do cliente ao longo do seu tempo de vida; finalmente, o problema foi dividido em problemas mais específicos, permitindo abordagens otimizadas a cada subproblema. Cada abordagem melhorou as métricas associadas à capacidade preditiva do modelo. Adicionalmente, a relevância dos atributos foi avaliada, confirmando os modelos obtidos como credíveis e valiosos para gestores de campanhas de telemarketing.

Palavras-chave: *telemarketing*, banco, depósitos bancários, sistemas de suporte à decisão, apoio à decisão, seleção de atributos, engenharia de atributos, *data mining*.

This work is dedicated to the beloved memory of Hermínia Claudina Silva.

Acknowledgments

The PhD journey is a lifetime challenge for everyone who embarks it. In my personal case, it was a valuable chance to continue research efforts following my MSc experience. This meant three years of a major investigation project immediately after the two years dedicated to my master degree, which could disincentive any head of family and full time working professional.

It all began in August of 2011, when I was still finishing my MSc, impelled by its success, decided to question Dr. Paulo Cortez, my MSc supervisor, about the feasibility for a PhD level research using the same bank telemarketing case study addressed for the masters' experiments. I was thrilled with his enthusiastic positive response, and started on the same day to plan my efforts. To complement supervision and obtain top-level Marketing experience, I invited Dr. Paulo Rita, thus forging the triumvirate for leading this doctoral research effort toward success. During this period, both of them launched me challenges which largely increased my academic maturity, namely for reviewing articles for ISI indexed journals and for lecturing during a semester in an MSc course.

There are not enough adjectives to express my gratitude to both my supervisors. Dr. Paulo Cortez is a data mining technical expert with excellent teaching skills for conveying his ideas. I even admire more his personal qualities: always available, with high motivation, eager to provide valuable advice, still open-minded for accepting his lower rank student's choice, and a natural research leader, with very ambitious however focused and realistic goals, making of him the perfect supervisor. I could not achieve this stage without his guidance. Dr. Paulo Rita, who is one of the very few Full Professors of Marketing in Portugal, leveraged my PhD through his inestimable advices not only in marketing and business management, but also through his vast experience. He is a fantastic leader always keen to provide his valuable knowledge. I deeply thank both for their top-level supervision!

While covering this hard-working path, I got helpful suggestions and recommendations from a few high esteemed Professors at ISCTE – IUL. I would like to thank Dr. Raul Laureano, from the ISCTE Business School, Dr. Pedro Ramos and Dr. Luís Botelho, from the Department of Information Science and Technology. Considering this research focused on a banking problem with different strategies benefiting from machine learning techniques, but also employing human expert knowledge, I was grateful for counting with a few experienced colleagues, whom I enumerate: Hugo Carrilho, who provided valuable help in every stages of the research, and Hugo Gonçalves, who lend his expertise on a later stage.

Finally, some words of gratitude for my family. My brother, Artur, who is a Post-Doctoral researcher in Chemistry, lend me his experience, while my parents Jorge and Deolinda kept encouraging me to continue. My grandparents, Abel who, unfortunately could not live to see this

accomplishment from his grandson, and Conceição provided me the needed inspiration through a life full of sacrifices and hard work. I have saved the most important people of my life for last: my son Tomás, who was born in the midst of this challenge, my older daughter Marta, and the love of my life, my beautiful wife Tânia. This work is only a reality through their sacrifice and encouragement!

Contents

List of Tables

Chapter I - Introduction

Chapter II - Articles

2.1 Article nr. #1

2.2 Article nr. #2

2.3 Article nr. #3

2.4 Article nr. #4

Chapter III - Conclusions

List of Figures

Chapter I - Introduction

Chapter II - Articles

2.1 Article nr. #1

2.2 Article nr. #2

2.3 Article nr. #3

2.4 Article nr. #4

Chapter III - Conclusions

List of Acronyms

ALIFT – Area of the LIFT cumulative curve

AUC – Area of the Receiver Operating Characteristic Curve

BI – Business Intelligence

CBR – Case-based Reasoning

CRISP-DM – Cross Industry Standard Process for Data Mining

CRM – Customer Relationship Management

DM – Data Mining

DSA – Data-based Sensitivity Analysis

DSS – Decision Support Systems

DT – Decision Trees

ECB – European Central Bank

ESWA – Expert Systems with Applications

FN – False Negatives

FP – False Positives

FPR – False Positive Rate

IDSS – Intelligent Decision Support Systems

INPREP – In Preparation

IS – Information Systems

JCR – Journal Citation Report

LDA – Latent Dirichlet Allocation

LR – Logistic Regression

LTV – Customer Lifetime Value

NCAA – Neural Computing and Applications

NN – Neural Network

PDSS – Personal Decision Support Systems

RFM – Recency, Frequency and Monetary

ROC – Receiver Operating Characteristic

SMO – Sequential Minimal Optimization

SVM – Support Vector Machine

TM – Text Mining

TP – True Positives

TPR – True Positive Rate

VEC – Variable Effect Characteristic

Chapter I - Introduction

This chapter describes the foundations which provided the starting point for our research project. These encompass, in the first sections, the catalysts and reasons for conducting this research to improve bank telemarketing (personal, professional and academic) and the main goals that driven this work.

Afterwards, the backgrounds on the domains involved are addressed. Considering this is a multi-disciplinary empirical research, several subsections are included, comprehending banking, telemarketing, decision support and data mining, and feature selection. However, it should be noted that such subsections are not completely exhaustive, in order to avoid repeating background sections from the articles published, which are fully copied from each journal to this document. It should also be noted that no background on similar banking applications to this research is referred in this chapter, considering one of the articles published and listed in the next chapter focus precisely in discovering the relevance and research gaps involving business intelligence applications to banking.

We also include a section for explaining the case-study used for validating the proposed strategies. While it is briefly mentioned in each article, a detailed description of the problem helps to elucidate the empirical analyses conducted. The next section describes the software and tools used for the experiments. Finally, the last section addresses the document organization. Since this document adopts a thesis by articles structure, such section is of paramount relevance for easy readability and understanding of the research flow which tightens the full work toward the common goal of identifying an interesting set of features for improving bank telemarketing.

1.1 Motivation and goals

The banking industry sets the roots and branches that sustain entire economies. High level competition on democracies makes of banks institutions thriving with innovation, in order to prosper and excel in a constantly changing world, always facing new challenges for economies to overcome. I have been lucky to work in such a demanding and formidable environment. Since I joined a major Portuguese retail bank, back in 2001, I have faced numerous tasks and projects such as home banking, phone banking, contact center management applications, campaign management, and asset-liability management, operational risk management for distinct areas such as Contact-Center, Marketing and Risk. My career has been evolving so far from a software engineer to a project coordinator in the information systems development department, allowing me to embrace a wide variety of challenges directly with the respective business related areas, which has provided me valuable insight knowledge for understanding some of the problems faced by banks.

An interesting domain which has kept my attention for a large portion of my banking career is customer related interaction, including marketing campaigns and contact-center management. Therefore, the problem associated with optimizing targeting for bank telemarketing campaigns emerged naturally as a potentially interesting research problem to address. Furthermore, I designed an entirely new solution implemented for operational campaign contact management, leaving to the business managers the definition of the customers who to target at for a specific product. Such gap in supporting managers' decisions led my ideas toward optimizing client selection recurring to analytical data-driven solutions. Considering the bank I work in is of a conservative type, a large number of telemarketing campaigns are conducted for convincing clients to subscribe long-term deposit accounts, in a hope to retain their assets for a longer period. Such fact made those campaigns excellent candidates for the experiments hereby described, by providing a large dataset to test the proposed approaches.

Within this research, our first goal was to assess the relevance of the proposed problem application, business intelligence applications to improve bank telemarketing campaigns for selling deposits. While we could perform a traditional literature review of a fraction of relevant publications on those domains, it was our intention to guarantee as much as possible that the extensiveness of this review covered a large number of articles. Considering the enormous number of publications available online, an exhaustive literature analysis can prove to be a demanding task. Such challenge constituted our first goal, for which we intended to use an automated literature analysis approach in order to deal with the large number of articles needed to consider.

It should be noted that this particular banking problem of assessing best clients to target at for subscribing long-term deposits was previously studied, under a post-graduate master degree research level. By using a data mining approach, explanatory knowledge was extracted from the same case study, with the findings being of interest for the scientific community, leading to two papers in two distinctive international conferences (Moro et al., 2011; Moro et al., 2012). However, the knowledge found could not be used for predictive purposes, as it used features which could only be known after each contact was already executed, such as the phone call duration. In fact, the call duration was considered the most relevant feature for the best predictive model, which comes as no surprise, since a longer call is a direct consequence of a more interested client, keener to hear what is being offered.

The data mining approach undertaken in previous research consisted in three CRISP-DM (Cross Industry Standard Process for Data Mining) iterations for a knowledge discovery project (Chapman et al., 2000). The first of those used the full dataset, revealing computational limitations which should be addressed. One of the issues identified was that there were twelve possible outcomes for a call (Table 1). A better business understanding suggested that it could not be inferred a conclusive result whenever the client did not specified it clearly, i.e., when there is an uncertainty about client's real intentions regarding the contact offer. This happens not only for unestablished contacts (e.g., the phone number does not belong to the client) but also when clients keep rescheduling the contact and managers decide to end the campaign, not finishing such contacts. For the second iteration, only the concluded contacts were kept (grayed in Table 1). It should be stressed that this principle was also followed in the present research. Finally, in the third iteration, a feature selection took place, by visually analyzing influence of each individual input on the outcome, and discarding those features which seemed to have little to none influence on the result.

Table 1 – Enumerated outcomes of a telemarketing contact.

Contact Result	Type of Contact	Total
Unsuccessful	Concluded contact	46387
Successful		6557
Not the owner of the phone	Cancelled contact	1056
Did not answer		5090
Fax instead of phone		5259
Abandoned call		2347
Aborted by the agent		79
Scheduled by other than the client	Scheduled contact	9848
Scheduled by the client himself		588
Scheduled – deposit presented to the client		3915
Scheduled – deposit not presented		2308
Scheduled due to machine answer		572
		81126

The feature strategy procedure used in the third CRISP-DM iteration from previous research suffered from two flaws: first, it considered only individual influence of a feature on the outcome – it happens quite often that several input features hide each other's' influence; then it is based on visual analysis and does not consisted in definition of metrics for threshold of influence. Both these issues are addressed with the strategies proposed in the present thesis.

It should be noted that a significant subset from the dataset used for previous research was made publicly available, removing only sensitive confidential information. This led to an exhaustion of data exploration with several published results from researchers across the globe for this specific dataset. Considering such fact and also the limitations on the feature selection procedure used in previous research, the challenge posed for the present research was to push the scientific boundaries to a new level, with novel feature selection strategies that could effectively enhance previous results. While the dataset used for empirical research was updated with more recent contacts (from November 2010, in previous work, to June 2013), emphasis should be given on effectively characterizing the problem in its plenitude. With this general research challenge in mind, two other more specific challenges associated with it arose:

- Usage of features that could be known prior to the initiation of the contact, thus discarding those such as the call duration which is only known after the call is terminated; such issue is quite relevant, for it means that we can discover predictive knowledge which can be used to actively select the most likely subscribers, instead of just explanatory knowledge, as in previous research;
- The feature selection strategies proposed should meet two requirements, namely:
    - Effectively improve bank telemarketing prediction performance, in terms of suggesting the right clients to target with the deposits, translating it into a direct benefit for decision support when managing campaigns, with an improvement in efficiency; this should be a direct result from the empirical analysis on the case study addressed;
    - While the features characterize the specific problem of bank telemarketing, with an emphasis on the case study used, the design of the feature selection strategies should allow to be extended to other domains, therefore providing valuable insights for future research on other problems using similar strategies.

The described goals define the main achievements and contributions for pushing the boundaries of this specific field of applied research.

1.2 Background

1.2.1.Banking

Banking is by its own a challenging industry and has always been on the verge of new technologies and information systems (Pennings & Harianto, 1992; Baets, 1996). There has been a long record of innovation in this industry. Examples include technology acceptance by end-users (Yousafzai et al., 2010), credit risk evaluation solutions (Wang et al., 2011), new forms of banking using newly conceived communication channels (Luo et al., 2010) and adaptation of widely used expert systems to address specific banking issues (Aburrous et al., 2010; Bahrammirzaee, 2010). Furthermore, the mere fact that it provides a basic but crucial service for the entire economy enables a head advantage in the availability for investing resources in research, as it can be observed by the groundbreaking products and services overtime, such as the automatic teller machines and credit cards which began to globally spread in the nineteen sixties (Consoli, 2005) and electronic secured transactions, in the eighties (Barber et al., 2012).

The recent global financial crisis, which arose in 2008 and widespread globally, is an example of the relevance of the banking industry. When some of the largest banks in the world collapsed, a mass hysteria took place worldwide, confirming the huge impact of banking management in every economy (Hurlburt et al., 2009). The crisis showed the systems used before for predicting systemic risk were inadequate for modeling an economic shock scenario (Demyanyk & Hasan, 2010). Since then, new systems have been developed and tested taking into account the variables which catapulted the crisis (Kerstein & Kozberg, 2013).

Marketing strategies in banking have suffered with the crisis. Targeting customers with bank products and services resulted in an even lower successful rate, demanding for new systems which could effectively understand customer behavior and detect the probability for a customer acquiring a certain product (Yada et al., 2010).

1.2.2.Telemarketing

Contacting directly a given list of clients for selling a product or service is named direct Marketing. It requires a directed communication channel for enabling individual contacts, as opposed to mass Marketing, with advertisements being publicized through the mass media. Usually such contacts are encompassed within a campaign context, which encloses the global strategy for the product being offered, the target audience, and selling approach (Krafft et al., 2007). The usage of the telephone for performing calls to clients in a Marketing campaign is denominated telemarketing. It allows an individual bidirectional and synchronous means of interacting with customers, enabling to present the product being offered, counter argue with negative aspects mentioned by the customer, and close a successful sale, all within the same contact. However, one of the major disadvantages is the intrusiveness associated with such contacts, which led to specific legislation with databases for opted-out customers (Milne & Rohm, 2000). Nevertheless, telemarketing has proven in past years its usefulness in a wide variety of industries (Moutinho, 2000).

Research focusing on improving targeting for telemarketing campaigns has been prolific for a quite large period. Back in the nineteen sixties, some pioneers such as Cox & Good (1967) were already working on designing architectures for Marketing Information Systems that could effectively provide better decision support on several related areas. However, it was only a few decades later, with more matured business information systems contemplating structured and efficient databases, that such systems evolved for real decision support customer-based systems (Abraham & Lodish, 1993; Van Bruggen et al., 1998). One of the most widely spread terms in such domain is Customer Relationship Management (CRM) systems, benefiting largely from database analysis techniques such as data mining (Berry & Linoff, 1999). In fact, the more recent work of Ngai et al. (2009) evaluated the vast literature published on data mining applications to CRM. Specifically for targeting efficiency, Young (2002) proposed a choice-based segmentation approach, while the work of Rotfeld (2004) shows interesting insights for opportunities emerged with the apparent negative effect of the opt-out registries for telemarketing. New mobile devices introduced other forms of complementary telemarketing, such as the usage of mobile Short Message Services (Rettie et al., 2005). Another different trend of research is the analysis of customers' receptivity toward telemarketing campaigns, using such knowledge for improving future campaigns (Mehrotra & Agarwal, 2009). On a very recent work with a similar goal of improving telemarketing through feature selection, Tan et al. (2014) proposed a completely different approach, using a single-feature evaluator specifically for addressing the imbalance class associated with telemarketing problems. Their study tested the method proposed on a large online employment advertising company, being able to improve a standard classification approach. However, it is limited to a standard database, not focusing on improving its value in terms of business knowledge. Also their single feature selection evaluation does not seem to apprehend the relations between different features which affect the final outcome. Finally, it follows a solely automated procedure, failing to use valuable and irreplaceable expert domain knowledge.

1.2.3.Decision support and data mining

The usage of information for decision support has become a common practice, tending to replace pure intuition or human experience based decision making. Classical reporting tools have proven to be insufficient in a global competitive economy and thus decision support and expert systems have evolved to take advantage of new data mining techniques for knowledge extraction based on statistics and artificial intelligence (Witten & Frank, 2005). In the 1990s, the concept of Business Intelligence spread globally, encompassing every decision support related concepts for benefiting business, not only data mining and knowledge discovery but also including information systems and architectures designed for the purpose of providing better decision support, such as data warehouses and both classical and interactive reporting tools (Turban et al., 2011).

A decision support system aims at providing managerial decision support through an information system specifically conceived for this goal (Turban et al., 2011). These systems are drawn based on past experiences for providing knowledge to leverage decision making. Arnott and Pervan (2008) analyzed literature in an attempt to identify fields of research in decision support systems. Two of these major groups are personal decision support systems and intelligent decision support systems. The first are small-scale systems that support a decision task of one manager, while the later use artificial intelligence techniques to support decisions. In such literature analysis, the authors show that the decision support systems discipline has been mostly driven by personal decision support systems. This could be motivated by two factors: the higher level of implementation of decision support systems in smaller companies; and the easier to access information by academics, considering data was collected from academic articles for that review. The same authors also allege that the relative lack of exposure of academics to contemporary professional practice is a particular problem for this discipline.

Sprague Jr. and Watson (1976) were amongst the first authors to publish about the relevance of decision support systems for banks. Since then, much has evolved with the literature reflecting this evolution. The study of Zopounidis et al. (1997) conducted a survey on the use of knowledge based decision support systems in financial management, showing that even then the subject was already of core relevance, with much of the financial industry adopting these decision support systems. They argued that these intelligent systems are the combination of classical decision support and expert systems, providing two crucial advantages: the capacity to analyze large amounts of data; and high levels of performance. In the recent years, the implementation of intelligent decision support systems in banks has been expanded (Skulimowski, 2011). This is a direct result from the large quantity of data that is stored by banks, driving the need to harvest this data for building knowledge to support decisions.

Data mining is a broad concept that involves methods and techniques associated with the extraction of valuable knowledge from raw data. Such knowledge can take the form of explanatory knowledge, by providing interesting insights for leveraging decision making, or predictive knowledge, to be used directly for forecasting the result of future occurrences (Witten & Frank, 2005). Classification is the most popular data mining task (Domingos, 2012), where the goal is to discover a discrete labeled outcome according with the knowledge found from the features that characterize an item.

Decision support using data mining has been applied to an immensely variety of problems. Koh & Tan (2011) argued that data mining has been increasingly essential in healthcare, ranging from discovering the best treatment for patients to diagnosis services based on patients and diseases knowledge. Moreover, advanced scientific groundbreaking research areas such as aerospace engineering, where errors must be kept below to just a slightly minimum threshold, have been benefiting from the usage of data mining (Monroe et al., 2012), proofing the reliability of some data mining approaches. And of course, business related areas are also using data mining to provide leverage for managers' decisions. These include detection of financial fraud (Ngai et al., 2011) and marketing-related areas such as customer relationship management (Hosseini et al., 2010).

While data mining demands for a structured dataset of problem events characterized by a list of common features, there are other types of problems in which such structured dataset is not easily available, thus the raw data is displayed in an unstructured fashion, harshening the process of mining for knowledge. If the problem instances are characterized by an unstructured text, or if they consist in text documents, then text mining provides an interesting approach for extracting knowledge (Fan et al., 2006). Indeed, the fact that the raw data is unstructured is what distinguishes a text mining problem from a data mining problem (Feldman & Sanger, 2007). Once the data is somewhat converted into a structured format for it can be handled by a typical data mining technique, then data mining and text mining may follow similar approaches. However, usually the data structure resulting from the first stage of a text mining procedure appears similar in its format for a large range of domains, and contains a high number of features (e.g., relevant words), although there are a few techniques for reducing this dimension to a manageable size. One of the most widely used formats is the document term matrix, which comprises of a bi-dimensional matrix with the columns referring to each term analyzed, and the rows representing the instances of the problem (e.g., the text documents), with each cell containing the number of times each term appears for each instance (Fortuna et al., 2005). When categorizing text documents, clustering techniques can be quite effective, by defining segments or topics of documents according to the keywords mined from the texts. The latent Dirichlet allocation is a typical example of such techniques (Blei, 2012).

1.2.4.Feature selection

Features are the characteristics that define an event for a given problem (Domingos, 2012). Considering typical data-driven approaches for solving problems, features are also denominated variables or attributes. For the remaining document, we adopt the term feature, in compliance with the thesis title.

Feature engineering is the most important factor in data mining projects and where most of the effort goes (Domingos, 2012): "if you have many independent features that each correlated well with the class, learning is easy". Feature engineering includes both feature design and feature selection. Feature engineering is also considered more difficult and costly because it is domain-specific (Witten and Frank, 2006; Domingos, 2012).

In real-world problems surrounded by complex contexts where infinitude of features may potentially influence the outcome of an event, the task of finding the best features, i.e., the ones that best characterize the problem becomes a quite challenging problem (Kohavi & John, 1997). Nevertheless, finding such features is of paramount relevance for any machine learning algorithm and technique, since the characterization of the problem sets the baseline for discovering the best approach for solving such problem. Quite often, a feature selection is the first sub-problem to be solved in order for proceeding with broader problem analysis. Feature selection emerged as a major machine learning research problem in the late nineties, when problem solving algorithms began to be fed with a large number of features, in some cases, with an order of magnitude of tens of thousands (Guyon & Elisseeff, 2003).

Several computerized techniques have been developed for dealing with feature selection. Liu & Yu (2005) presented a generalized wrapper algorithm for selecting the set of features that optimize a data mining problem. Their proposal starts from a training dataset with a set of features and generates a subset of those features, evaluating the performance of the data mining task; if it outperforms previous subset of features, then it becomes the best subset found so far. Thus the emphasis is on reducing the number of features. This suits fine for large sets of features which already include those that best characterize the problem; however, real-world problem often encompass intrinsic relations with several context variables that are hardly included in a common dataset available for data mining (Ciskowski, 2006). Feature selection techniques range from the more basic randomization (e.g., the Monte Carlo algorithm) to procedures which attribute a score of relevance for each feature for removing the least relevant (Liu & Motoda, 2007). Nevertheless, the performance of the later method is totally dependent on the function which defines feature relevance. In compliance with the common sense “no free lunch” theorem, measuring feature relevance is very difficult because of a limited training dataset, the curse of dimensionality and required computational effort, and noise in the data. Furthermore, evaluation of the distribution of the input variables and the response always relies on some model. Also masking poses another challenge for feature relevance, for it occurs when one feature can effectively represent others in a model, hiding each other’s influence (Tuv et al., 2009).

A sensitivity analysis enables to measure the effect of feature variation on a given model when predicting a desired outcome (Guyon & Elisseeff, 2003). If the model predictions vary largely when going through the range of a certain feature, it means such feature is highly relevant for the model. However, it should be stressed that a model with several relevant features does not necessarily indicates it is a model with good prediction performance. Sensitivity analysis has been used for assessing feature relevance in a wide variety of problems, such as medicine (Hunter et al., 2002; Abeel et al., 2010), handwritten recognition (Oliveira et al., 2002), biology (Sharma et al., 2012), topography (Guo et al., 2011), churning (Coussement et al., 2010) and profiling users (Mansingh et al., 2013). Furthermore, research on sensitivity analysis as a method for computing feature relevance has been the subject of study for a few studies (Saltelli, 2002; Guyon et al., 2006; Liu et al., 2010). Cortez & Embrechts (2013) presented a novel data-based sensitivity analysis, which is computationally efficient when compared with previous sensitivity analysis methods considering feature inter-relations. It unfolds those inter-relations by using samples collected from the training set for simultaneous feature variation. For our experiments with automated feature selection, we adopted the data-based sensitivity analysis to compute feature relevance and provide guidance on the least relevant and thus best candidate features for discarding.

While automated approaches for feature selection have been extensively studied, one of the most common methods remains the usage of domain knowledge taken from human experts for assessing which are the relevant features for the problem, thus discarding the rest. Most notably, this is considered one of the best feature selection methods (Witten & Frank, 2006). It is completely straightforward, provided that the degree of expertise allows a glimpse on the real goal of the problem and its specificities, which are hardly discovered by any machine learning technique in a reasonable amount of computational effort and time. Furthermore, recent advances in machine learning and artificial intelligence continue to lack true human cognition, and it seems there will still be a while before automatic problem solvers can spare such valuable knowledge (Breuker, 2013). However, for a very large set of features, a manual analysis of such amount of features may render unfeasible, thus automated approaches are useful for pruning the set to a reasonable number.

Within our work, all strategies that are proposed are composed of a mixed approach, including first a human expert knowledge for assessing the relevance of features and pruning them to a reasonable sized list of promising feature candidates, and then an automated data-based sensitivity analysis for further reducing such feature set to the ones which optimize prediction performance. The usage of a combined expert and automated knowledge was also proposed by Gebus & Leiviskä (2009), who conducted empirical analysis in an assembly line quality control problem and concluded that expert knowledge improved drastically the performance of data mining methods for feature selection. However, their strategy consisted in using the expert knowledge for helping in the feature selection process from the available database, not trying to enrich it with new helpful information for characterizing the problem.

1.3 Case study description

The case used for the empirical research consists in several telemarketing campaigns executed from May 2008 to June 2013, in a total of 52,944 phone contacts. It should be stressed that all of these contacts have reached a final concluded stage, meaning their result is either a success or an unsuccessful outcome. Therefore, every other result different from the concluded result type (Table 1) were previously discarded due to the reasons mentioned in section 1.1. A typical characteristic of targeting problems is the low successful rate (Verhoef &

Leeflang, 2009). Hence it comes as no surprise that the dataset for our case is unbalanced, as only 6,557 (12.38%) records are related with successes.

Some remarks must be made about this case. First, the concept of a campaign for this specific bank institution defines a given period of time where a group of agents try to sell the campaign deposit through the execution of phone contacts. The main differences between campaigns are two: the time frame in which the calls were made; and the dialog script defined by the campaign manager for agents to conduct conversation. While the later may play a role in influencing the client, the exact conversation which took place is not stored; therefore, it could not be accounted for in this study. Also, while the deposit type may be different for each campaign, sometimes the same deposit is communicated in different periods. Furthermore, the agents may vary, but this is more dependent on human resources management than the specific campaign (i.e., agents are not hired for specific campaigns). Second, all calls are executed through human agents (i.e., no automated calls are made), with a large fraction of the contacts being executed in outbound, where the institution takes the initiative of contacting the client, while the remaining are inbound calls where the client is calling for any other reason and the agent gets an alert indicating that client is targeted for the campaign, giving the choice to the agent of deciding or not to approach the client for selling the deposit. Third, while a campaign may communicate the same or a different deposit than previous campaigns, the characteristics of the deposit type are incorporated in terms of features (e.g., interest rate offered, term period) into the dataset.

The dataset for this problem is characterized by 119 features. The strategies undertaken focus on the features and include enrichment with newly proposed features and feature selection procedures. Each of the strategies will change the original dataset, adding value in the form of problem characterization, which can be translated into insightful knowledge.

1.4 Research methodology

In general terms, the present thesis followed a design science research approach (Kuechler & Vaishnavi, 2008), by producing scientific artifacts in the form of data mining models, which were improved in each of the proposed strategies. The project plan that drove this research is rooted on the CRISP-DM knowledge discovery methodology (Chapman et al., 2000).

Figure 1 shows both the design science research and CRISP-DM approaches. We find interesting to notice the similarities between the more general design science research method and the specifically conceived framework of CRISP-DM. It is exactly for this reason that we chose to show both methodologies side by side. While our research is mainly focused on feature engineering, it followed data mining procedures for building and validating models, in a pursuit for the best set of features. Thus, within each strategy, the goal needed to be defined as well as the adequacy of the proposed features for enriching the dataset. These two objectives are the aim of the two stages of both methodologies. The next stage is the development of the artifact, whereas in CRISP-DM it includes two specific data-driven phases which are tightly coupled: data preparation and modeling.

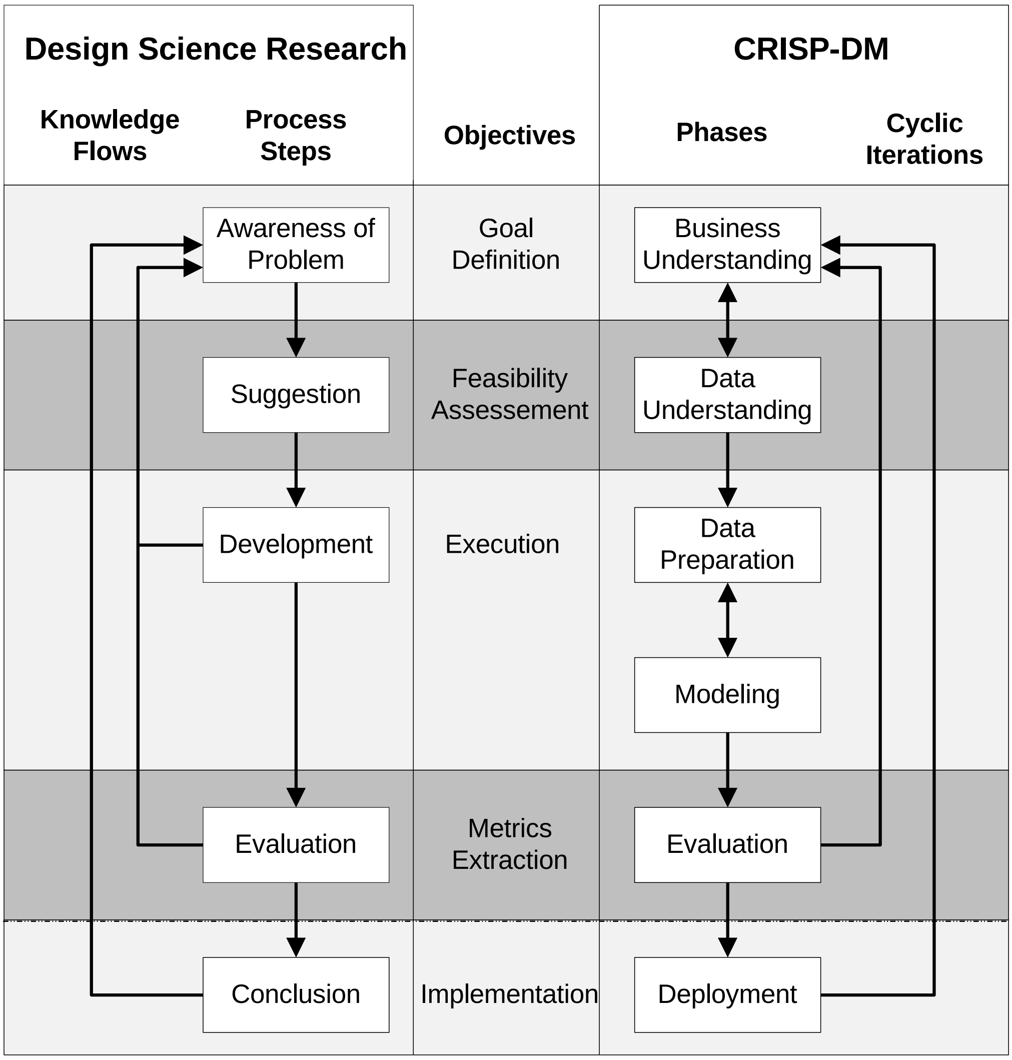
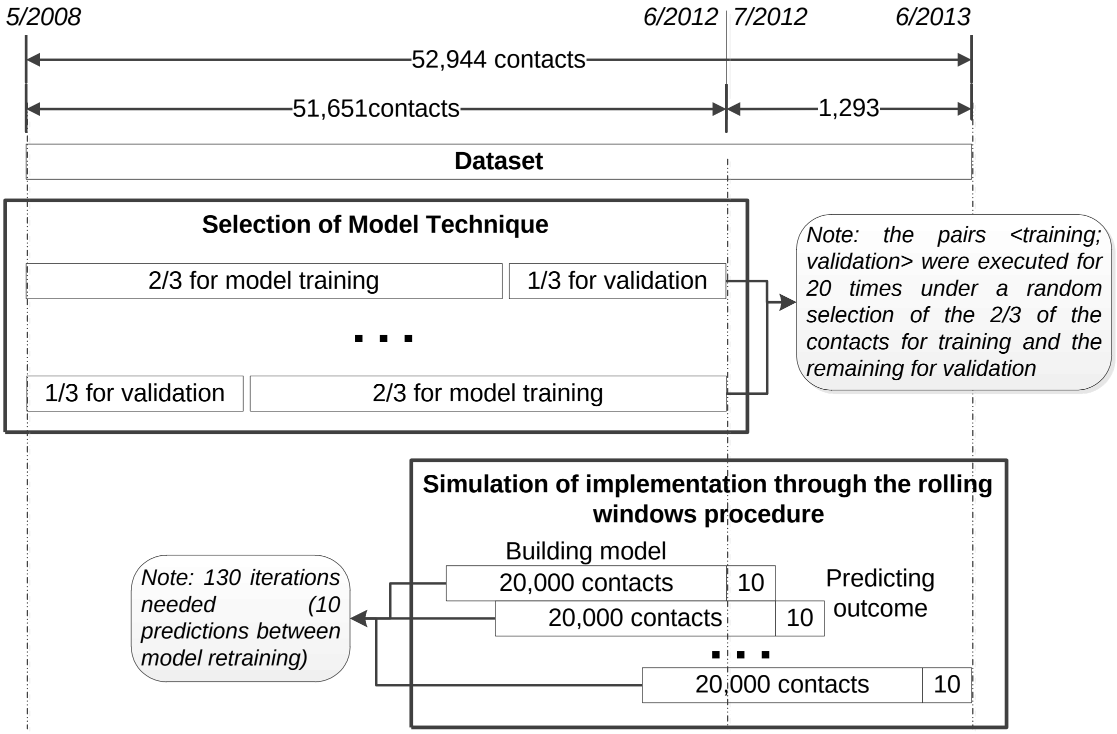


Figure 1 – Design Science Research and CRISP-DM.

Adapted from Kuechler & Vaishnavi (2008) and Chapman et al. (2000).

The artifact (or model, in data mining) evaluation is a critical stage in both methodologies: it aims to validate the results from the execution stage. Several metrics are available for evaluating a data mining model. We used two of them: the area under the receiver operating curve (AUC), which provides a measure for the degree of discrimination that can be obtained with the model across the range of possible probabilistic outcome thresholds (Fawcett, 2006); and the area under the Lift cumulative curve (ALIFT), for measuring the performance by a decrease ordering of the events in terms of the probability of occurring the desired outcome (Coppock, 2002). Both metrics range from 0 to 1, with the ideal model presenting metrics closer to 1. Also based on the Lift cumulative curve, we extracted an objective criterion for measuring efficiency: how many successes from the total dataset would be reached if a reduction of half of the contacts was needed. This corresponds to one point of the Lift cumulative curve: the vertical axis value that corresponds to the middle of the horizontal axis (i.e., 0.5).

Considering the feature reduction and feature selection procedures included several modeling stages, the dataset was divided in the first four years, from May 2008 to June 2012, for conducting experiments on the data, and the later year, from July 2012 to June 2013 for testing the impact on the results (Figure 2). The first fraction was used for analyzing modeling techniques in order to assess which provided the best predictions for our specific case. Each of the models was computed 20 times on each of the data mining techniques tested (logistic regression, decision tree, neural network and support vector machine), with a different random selection of the 2/3 of the contacts for building the model, and 1/3 for measuring its performance.



Note: the pairs <training; validation> were executed for 20 times under a random selection of the 2/3 of the contacts for training and the remaining for validation

Note: 130 iterations needed (10 predictions between model retraining)

Figure 2 – Modeling and testing methods

Since the present thesis focus on feature engineering, a significant large amount of effort was spent in evaluating model impact, for providing solid ground proof that justified discarding some of the features while retaining others.

To address this issue, the most recent year of contacts, from July 2012 to June 2013, was used for a realistic simulation following a fixed size rolling window procedure (Tashman, 2000), where a window of ten predictions for the next ten contacts was set, and the model being re-trained with the most recent past 20,000 contacts, from which the 10 predicted are included after evaluating prediction results, while the oldest 10 contacts are removed, keeping the window size constant, and providing a simple method for simulating a real execution.

The last stage for both the design science research and CRISP-DM methodologies aims to implement the usage of the produced artifact/model for benefiting business. However, while our procedure was extensively validated, no deployment was made which resulted in a full working decision support system. By comparing both methodologies followed, we emphasize the resemblances as if the CRISP-DM is an instantiation of the design science research for the specific case of knowledge discovery projects. Such comparison is in alignment with the very recent analysis of Arnott and Pervan (2014), which emphasizes that decision support systems research is evolving from a field based on statistical hypothesis testing and conceptual studies to one where design science research is the most popular method.

Considering the artifacts were developed for solving the specific problem of selling long term deposits through bank telemarketing, and since the dataset was collected from a specific real bank, this approach can be viewed as a case study based research.

One of the first methodologies for designing decision support systems is the model defined by Simon (1960). This model was later extended to four stages (Sprague Jr., 1980), as shown in Figure 3, and is nowadays considered a classical baseline model for designing decision support systems.

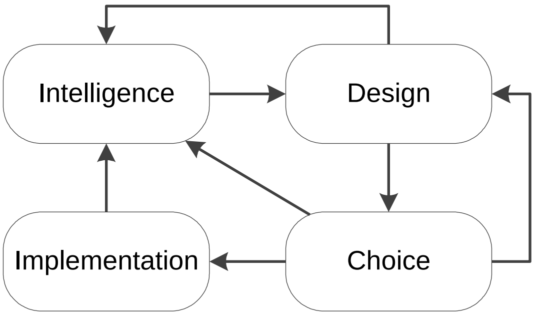


Figure 3 – Decision support systems model

Adapted from Simon (1960) and Sprague Jr. (1980)

The main focus of this thesis is on unveiling new features through novel feature selection procedures, thus a full working decision support system was not implemented. Nevertheless, it is worth of noticing the paradigm of the model of Simon applies in several relevant premises defined through two of the four stages of the model. This research focused in finding and selecting relevant features, which suits the purpose of the “Intelligence” stage of defining and characterizing the problem being addressed, for which, in a data mining research, is highly related with the features that characterize each occurrence of the problem. Also it can be considered that the "Design" stage is addressed by the present research, mainly for defining the prediction model that can be used to support bank telemarketing decisions, although it was tested on a simulation run. As such, the final bridge to the real decision model that would select the most likely subscribers for our problem was not addressed. Thus, it can be argued that our research does not address the full decision support system development cycle, but nevertheless it sets a base from which a real operating decision support system can be evolved.

1.5 Contributions

At a first stage, this thesis addressed a novel semi-automated literature analysis approach using text mining, including in particular four main contributions:

1. employment of expert knowledge for defining the dictionaries that intersected both domains reviewed, business intelligence and banking;
2. application of the latent Dirichlet allocation in literature analysis by defining topics grouping similar articles in the terms addressed, when considering the dictionaries used;
3. analysis in a timeline of the articles published through the twelve years of the time frame considered;
4. validation of the findings through the selection and in-depth manual analysis of the most representative article for each topic, providing an interesting although quite simple method of confirming the trend.

Turning to the main research thread of this thesis, we highlight the three strategies for feature selection:

5. Feature selection strategy consisting in unveiling novel features related to problem context (e.g., social and economic characteristics);
6. Feature selection strategy consisting in computing new historic features from the initial dataset for enriching problem characterization (e.g., customer lifetime value characteristics);
7. Feature selection strategy consisting in dividing the problem in smaller and more manageable sub-problems to provide a better characterization with novel features optimized for the sub-problems.

The three above mentioned strategies provided a steady improvement with each of them adding value to the dataset in terms of modeling prediction performance, thus were materialized in a framework for improving problem characterization.

1.6 List of Publications

The work related with this thesis resulted in the publication of the following articles:

- Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications, 26, 131–139;
- Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation, Expert Systems with Applications, 42(3), 1314-1324;
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

And the following article is in the process of submission:

- Moro, S., Cortez, P., & Rita, P. A divide and conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing (**In Preparation**).

1.7 Software and tools

Several tools are available for data-driven analysis, providing statistical functions and advanced data mining techniques. Competition is fierce in the software industry for this specific fast growing market. Back in 2007, Watson & Wixom (2007) indicated that major global vendors such as Microsoft, IBM, Oracle, and SAP were increasing their commitment and investment in business intelligence. The entry of such huge players emphasizes competition in a market with already global scale companies dedicated specifically to business intelligence and data analysis, such as SAS and Microstrategy. However, those World dimension players face themselves enormous competition from open source software, supported by large communities of contributors and fans, such as R (Ihaka & Gentleman, 1996) and WEKA (Hall et al., 2009). Furthermore, these open source solutions are free of cost, compared to the high expensiveness associated with the former solutions mentioned. In a highly competitive environment, open source software is increasing its penetration in large enterprises and markets. Muenchen (2012) evaluated a large number of business intelligence solutions under several metrics. Considering the number of scholarly articles found, R comes in the fourth place, with the giants IBM SPSS and SAS occupying the top two positions; a similar result occurs when the criterion is the number of books published, with SPSS and SAS in the top spots and R in third. However, R is by far the more relevant software when the metrics consider e-mail traffic and posts in forums and discussion groups, proving the free and enthusiastic support associated with this open source solution. Moreover, the very recent analysis of Horwitz (2014) about the impact of the open software machine learning tools of R and Hadoop emphasizes the fast growing pace of these solutions, fueled by the community for making the core software useful and relevant by providing answers to common questions. For all the experiments conducted in this thesis, the R statistical tool was used, benefiting from the several advantages mentioned above.

The R statistical tool enables any contributors to supply their independent packages and upload them on to the *Comprehensive R Archive Network*[1] (CRAN) repository, which is public and freely available for any user. For the automated literature analysis, a text mining approach was the logical choice, with the **tm** package offering an interest kit of functions to manage the articles we needed to review (Meyer et al., 2008). Also, for creating clusters of articles for assessing the most relevant subjects and possible research gaps, the **topicmodels** provides an excellent choice in combination with the **tm** package, for it is fed with inputs originated from the resulting outputs of the **tm** package (Hornik & Grün, 2011). Other helpful packages were also included, such as the **wordcloud** package.

For the experiments with the suggested feature selection strategies, a data mining related package was in demand. We adopted the **rminer** package for it implements a very simple and coherent set of functions which provide an abstraction layer over the specificities and technical issues associated with advanced data mining algorithms (Cortez, 2010). Besides the usual set of data mining algorithms, it also provides feature relevance assessment through sensitivity analysis, providing an interesting set of graphic displays for easier interpretation of feature relevance and influence on the desired outcome. Furthermore, such package was implemented by the same author who proposed the novel data based sensitivity analysis, which was adopted for the experiments presented.

1.8 Thesis organization

Following a possibility allowed by the regulations of this PhD program, and that is only an option when at least three Journal Citation Reports (JCR) ISI indexed journal articles are published, this document is presented as a thesis by articles type. We opted for such possibility since it allowed us to put a higher effort in the writing (and publishing) of high quality scientific papers. Therefore, the main part of the thesis, corresponding to Chapter II, is composed of three accepted articles and one submitted and currently under review. All of the three articles are published and available online, with two of them having already a volume number attributed. As stated by Hong (2013), the most straightforward and objective form of evaluation for research progress and findings is through research publications. While several communication types are available, such as conferences and working papers, scientific peer-reviewed journals remain the major form for communicating certified results, whereas the former mentioned types

[1] Available at: http://cran.r-project.org/

provide a fast means of communication (Halpern & Parkes, 2011). Nevertheless, with the proliferation of journal titles, it can become quite difficult to assess the relevance of a given journal title for a research domain, when comparing to other journals. The most widely used methods all come down to metrics related with citation numbers. While it can be argued that such criterion can be a deceiving metric (Parnas, 2007), simplicity associated with such metrics makes of rankings based on them very popular, with several rank lists available, including new access free Google Scholar (Butler, 2008). One of the most widely used of such metrics is the JCR compiled by Thomson Reuters (known as ISI) and available at the Web of Knowledge portal. The fact that major journals are indexed in this report makes of it very reliable (Glänzel & Moed, 2002; Bornmann et al., 2011). While there are a few studies that criticize the influence and relevance attributed to this report, such as Bar-Ilan (2012), in practical terms JCR is a standard index for ranking the quality of journals. Using JCR, four quartiles are commonly defined, ranging from Q1 (top quartile) to Q4 (last quartile).

The presented research can be logically divided into four major stages: first, the quest for relevance and gaps for the problem addressed; the next three stages focus in three different strategies of feature selection for improving results. Each strategy started from the outcome of the previous strategy in terms of features, thus benefiting from previously improved results, in an iterative procedure for adding value to the overall dataset. This research project plan was therefore composed of four independent stages, which constituted sub-problems within the main goal (Figure 4). Nevertheless, all of them are connected toward the main goal.

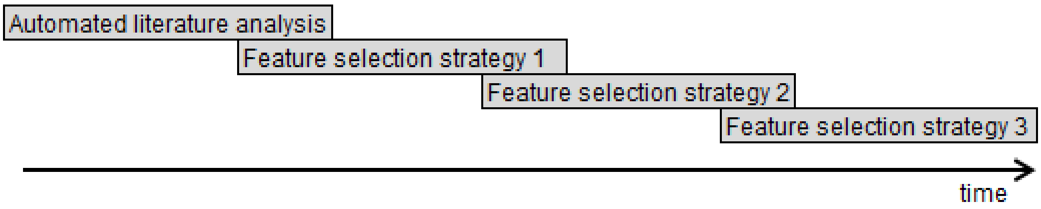

Automated literature analysis
Feature selection strategy 1
Feature selection strategy 2
Feature selection strategy 3
time

Figure 4 – Research project plan.

With the dawn of the research for one of the stages, the next was previously being prepared, allowing optimizing the timeframe available and benefiting from earlier insights, with the later serving the purpose of tightening the entire research project.

Considering the four stages defined, the goal was that each of them resulted in an independent publishable work. We chose to bet on quality instead of quantity, with four articles in four major journals. This was achieved for the three first stages, with the fourth resulting in an article still currently under review. It should also be said that every article published is authored by the doctoral candidate and his two supervisors. Nevertheless, the candidate is always the first author and the main impeller of the work, with the remaining authors contributing in the usual supervision process. The resulting articles are displayed in Table 2.

Table 2 – Articles produced.

	Title	**First Available**	**Journal (initials)**	**Publisher**	**JCR Impact Factor***	
					2013 5 year	**2013 Rank****
#1	Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation	29-Sep-2014	Expert Systems with Applications (ESWA)	Elsevier	1.965 2.254	Q1(30/121) CS-AI Q1(11/79) OR&MS
#2	A data-driven approach to predict the success of bank telemarketing	13-Mar-2014	Decision Support Systems (DSS)	Elsevier	2.036 2.651	Q1(27/121) CS-AI Q1(20/135) CS-IS Q1(9/79) OR&MS
#3	Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns	2-Sep-2014	Neural Computing & Applications (NCAA)	Springer	1.763 1.502	Q2(40/121) CS-AI
#4	A divide and conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing	NA	In Preparation (INPREP)	NA	NA	NA

* Impact factor of Journal Citation Report (JCR) published by Thomson Reuters.
** Qn = Quartile number, ordered from the highest to lowest impact factor; Ranks are by subject category – in the table, only relevant categories for subjects addressed are shown:
CS-AI = Computer Science - Artificial Intelligence;
CS-IS = Computer Science - Information Systems;
CS-IA = Computer Science - Interdisciplinary Applications;
OR&MS = Operations Research and Management Science.

As stated previously, while the JCR metrics have some inconveniences, those are nevertheless the most widely accepted means for measuring journal impact, using its globally known impact factor. Taking a deeper look into the journals chosen and corresponding impact factor rankings, it can be observed that those are considered major journals in their respective fields of research, providing not only an effective way of communicating results, but also a reliable means of certifying the work through each journal double (sometimes triple) blinded international peer-review process.

The abbreviations for journal titles in terms of their capital letters are used in the remaining document (Chapter III) for referring to each of the articles. They are also numbered for allowing their position in the research plan. It should be stressed that the publication date is dependent on each journals' staff, thus dates shown do not reflect the chronological order of work.

To better present each individual work, we choose to retain for each article the complete version as published in the respective publication. This leads us to two different referencing systems in the thesis. For each article in Chapter II, its own references are provided together with the article, in the specific journal format, while in the bibliography section at the end of the thesis are supplied the cited publications in the remaining document, i.e., Chapters I and III. The journal branding logos and other specificities are copied exactly to this document (e.g., the DSS article supplies a short biography of the authors). Also there are two different pagination numbering systems, considering each article is contained in a volume of articles within each journal. We chose to add our pagination numbering to the footer of each article's page, thus making this small change to the original articles. Since some journals formats already provide small margins, for some cases our page number is placed in a location below the remaining thesis page numbers. However, by proceeding with this change, we are allowed to use a unique referencing page format.

As stated previously, each article can be considered an independent work on its own, hence the publications achieved, with each containing typical sections such as introduction/background, materials and methods, experiments and results, and conclusions. Still, each of them contributes to the major thesis goal of improving the data-driven bank telemarketing results through novel feature selection strategies. Thus a final chapter is focused on drawing the conclusion of the thesis. It encompasses the relevance of the problem emerged from the first article, and the discussion for the three results exhibited in the three strategy articles, for designing a common framework which is validated through the experiments conducted in those articles.

The present thesis is organized as follows:

In Chapter II, the main research effort and results are presented, in the form of the three articles published and one submitted. Each section inside this chapter corresponds to an article. The first section comprehends the literature analysis work, whereas the following three sections are the experiments and results associated to each of the three suggested strategies.

Chapter III starts with a thorough discussion on the results exhibited in the previous chapter, and draws the framework as a result of the application of the three strategies. Finally, the last two sections mention the final remarks on the thesis and the future work which may follow this research.

Chapter II - Articles

2.1 Article nr. #1

This article presents a literature analysis on Business Intelligence applications to the banking industry, encompassing a recent time period (2002 to 2013). The goal was to assess research trends and gaps while covering a large number of representative manuscripts, extracted from highly relevant publications.

The main contribution to the present thesis was to uncover the existence of a research gap on the specific problem of selling bank deposits through telemarketing campaigns in terms of Business Intelligence related applications. Such gap drove the following steps of the present research. Also, other relevant contributions were made, namely related to the procedures undertaken using text mining and the latent Dirichlet allocation and the approach for validating the findings.

Article details:

- Title: Business Intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation;
- Date: February 2015 (made available online since 29 September 2014);
- Journal: Expert Systems with Applications;
- Publisher: Elsevier.

Expert Systems with Applications 42 (2015) 1314–1324

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation

Sérgio Moro [a,*], Paulo Cortez [b], Paulo Rita [c]

[a] Business Research Unit (UNIDE-IUL), Department of Information Science and Technology, ISCTE – University Institute of Lisbon, 1649-026 Lisbon, Portugal
[b] ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal
[c] Business Research Unit (BRU-UNIDE), ISCTE – University Institute of Lisbon, 1649-026 Lisbon, Portugal

ARTICLE INFO

Article history:
Available online 29 September 2014

Keywords:
Banking
Business intelligence
Data mining
Text mining
Decision support systems

ABSTRACT

This paper analyzes recent literature in the search for trends in business intelligence applications for the banking industry. Searches were performed in relevant journals resulting in 219 articles published between 2002 and 2013. To analyze such a large number of manuscripts, text mining techniques were used in pursuit for relevant terms on both business intelligence and banking domains. Moreover, the latent Dirichlet allocation modeling was used in order to group articles in several relevant topics. The analysis was conducted using a dictionary of terms belonging to both banking and business intelligence domains. Such procedure allowed for the identification of relationships between terms and topics grouping articles, enabling to emerge hypotheses regarding research directions. To confirm such hypotheses, relevant articles were collected and scrutinized, allowing to validate the text mining procedure. The results show that credit in banking is clearly the main application trend, particularly predicting risk and thus supporting credit approval or denial. There is also a relevant interest in bankruptcy and fraud prediction. Customer retention seems to be associated, although weakly, with targeting, justifying bank offers to reduce churn. In addition, a large number of articles focused more on business intelligence techniques and its applications, using the banking industry just for evaluation, thus, not clearly acclaiming for benefits in the banking business. By identifying these current research topics, this study also highlights opportunities for future research.

1. Introduction

Banking has been a prolific industry for innovation concerning information systems and technologies (Shu & Strassmann, 2005). For example, new technologies have enabled new communication channels which were quickly adopted by banks. Also, advanced data analysis techniques are currently used to evaluate risk in credit approval (Huang, Chen, Hsu, Chen, & Wu, 2004) and fraud detection (Ngai, Hu, Wong, Chen, & Sun, 2011).

Business intelligence (BI) is an umbrella term that includes architectures, tools, databases, applications and methodologies with the goal of analyzing data in order to support decisions of business managers (Turban, Sharda, & Delen, 2011). Banking domains, such as credit evaluation, branches performance, e-banking, customer segmentation and retention, are excellent fields for application of a wide variety of BI concepts and techniques, including data mining (DM), data warehouses and decision support systems (DSS). For bank firms to survive and even excel in today's turbulent business environment, bank managers need to have a continuous focus on solving challenging problems and exploiting opportunities. That demands a need for computerized support of managerial decision making thus implying the need of decision support and business intelligence systems.

There are several surveys/reviews of the banking domain. Wilson, Casu, Girardone, and Molyneux (2010) published a recent literature review covering the impact of the global financial crisis in the banking business. Their results put the risk domain as a subject that deserves a deeper attention in order to achieve a systemic stability. The review of Ngai et al. (2011) devoted attention to financial fraud detection, and classified 49 articles depending on the type of fraud. The findings suggest a lack of research on mortgage fraud, money laundering, and securities and commodities fraud, by contrast to a large number of articles on credit fraud. More related with this paper, Fethi and Pasiouras (2010) presented a survey on bank branches performance based on 196 articles which employ operational research and artificial intelligence

* Corresponding author.
E-mail addresses: scmoro@gmail.com (S. Moro), pcortez@dsi.uminho.pt (P. Cortez), paulo.rita@iscte.pt (P. Rita).

http://dx.doi.org/10.1016/j.eswa.2014.09.024
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

techniques, concluding that profit efficiency and capacity efficiency have received limited attention in the studies evaluated.

A large research attention has been given toward credit. In fact, although credit is traditionally related to banking, it has long spread to other industries. Therefore, some recent reviews and surveys are naturally available on the subject. Abdou and Pointon (2011) reviewed 214 articles/books/thesis on credit scoring applications, searching for the statistical techniques used for evaluation and found that there is not an overall best technique for building models. The review of Marqués, García, and Sánchez (2012) reports over the use of evolutionary computation for credit scoring. Another subject of interest is e-banking, specifically customer acceptance toward a new communication channel. Dahlberg, Mallat, Ondrus, and Zmijewska (2008) reviewed publications on mobile payments and found through their framework lacking of research on social and cultural factors impacting mobile payments, as well as traditional payment services.

The enlisted surveys and reviews cover some themes in banking. However, within the authors' knowledge, there is a lack of a recent literature analysis for BI applications in the main subjects related to the banking industry, thus motivating the present research. Furthermore, none of the discussed reviews adopted an automated text analysis, by using Text Mining (TM) techniques such as the ones presented in this study, thus facilitating the analysis of a much larger set of sources.

This paper presents an automated text mining literature analysis, from 2002 to 2013, of BI applications within the banking domain, allowing the identification of current research trends and interesting future applications, thus highlighting opportunities for further research. Although BI has been extensively studied, recent years and particularly the last decade have experienced a huge increase in BI applications through the industry, especially in the banking sector, therefore stimulating research. This article is organized as follows. Section 2 introduces the main concepts related with both banking and BI domains, and presents also other references of literature analyses. Next, Section 3 presents the methods used for analyzing the literature. Then, the results are discussed in Section 4. Finally, conclusions are summarized in Section 5, which also presents future research directions.

2. Background

2.1. Text mining

Data mining (DM) aims to extract useful knowledge (e.g., patterns or trends) from raw data (Witten & Frank, 2005). Text mining (TM) is a particular type of DM that is focused on handling unstructured or semi structured data sets, such as text documents (Fan, Wallace, Rich, & Zhang, 2006). Delen and Crossland (2008) proposed the application of TM for analyzing the literature and identify research trends, thus helping researchers in conducting state of the art reviews on a given research subject. Their research focused on three major journals in management information systems, although they argue that their TM approach can be valuable in virtually any research field.

Within a literature analysis, searching with individual words is often not enough, since many searchable terms can be composed of a sequence of words, such as "data mining" or "decision support systems". Those sequences, which can be made of n words, are called n-grams. When extracted from large texts, n-grams constitute a valuable asset, in particularly when analyzing publications, such as the study of Soper and Turel (2012) showed by analyzing eleven years (from 2000 to 2010) of publications in the Communications of the ACM journal.

When conducting TM over text documents, relevant words and terms are often extracted in order to produce a categorization that can help building a body of knowledge over the literature considered (Delen & Crossland, 2008). An interesting approach is modeling a certain number of distinct topics defined according to the number and distribution of terms across the documents, which can be achieved through the latent Dirichlet allocation (LDA) model (Blei, Ng, & Jordan, 2003). For each document, it is determined the probability of belonging to each of the topics, allowing to group documents to the more likely matching topics. This organization structure can help identifying which topics are capturing more attention from researchers and also to find gaps for future research. TM can be used indiscriminately, by looking for the most overall referred words, or through the use of specific dictionary words. Since this work is about a focused literature analysis, a dictionary of terms in both banking and BI domains is used.

2.2. Banking

Banks are institutions that operate in the financial business domain, concerning activities such as loaning, deposits management and investments in capital markets, among others. The banking industry is crucial for the economy and thus it is a subject of great interest for researchers in a widespread of different domains, such as management science, marketing, finance and information technologies. Berger (2003) found evidence of a relation between technological progress and productivity in banking. The same author also emphasizes that banks employ statistical models based on their financial data for different purposes, such as credit scoring and risk evaluation.

Financial sector reforms allowed an increase in competition, turning bank lending an important source of funding. Credit risk evaluation is by its own a vast domain, encompassing a large number of research publications within banking and spread through the last twelve years (e.g., Marqués et al. (2012)). Other banking related subject where research has been active is fraud prevention and detection in traditional banking services (e.g., Abbasi, Albrecht, Vance, & Hansen (2012)) and in new communication channels that support e-banking services (e.g., Shuaibu, Norwawi, Selamat, & Al-Alwani (2013)), from which electronic mail spamming in order to illicitly obtain private financial information is a specific case of interest (e.g., Amayri & Bouguila (2010)). E-banking is also subject of another research domain related to technology acceptance regarding new communication channels adopted by banks (e.g., Vatanasombut, Igbaria, Stylianou, & Rodgers (2008), Lin (2011)). A not so recent theme that however has boomed in research, driven by the global financial crisis, is bankruptcy and related subjects such as systemic risk and contagion (e.g., Hu, Zhao, Hua, & Wong (2012)). Competition had also an effect on client related areas, with banks increasing investment in customer retention, customer relationship management (CRM) and targeting (e.g., Karakostas, Kardaras, & Papathanassiou (2005)).

Research in banking is currently an interesting domain of research. Due to advances in information technology, virtually all banking operations and procedures are automated, generating large amounts of data. Therefore, all the subjects mentioned above can potentially benefit from BI solutions.

2.3. Business intelligence

BI involves several distinct areas and technologies that converge in the common goal of having access to data in order to help businesses by facilitating knowledge and supporting better management decisions. One way to accomplish this is by predicting a certain behavior or result based on data-driven models, in what is known as DM or predictive analytics, thus providing the most likely outcomes to managers (Han, Kamber, & Pei, 2006; Turban et al., 2011; Witten & Frank, 2005).

Intersecting several fields of research, such as artificial intelligence, statistics and databases, several supervised learning DM algorithms have been proposed for building data-driven models. These predictive DM models are classified into two main types: classification, if the output target is a categorical value, and regression, if the target variable is a numeric value. Examples of popular DM models that can be applied to both classification and regression are decision trees, neural networks and support vector machines (Witten & Frank, 2005). There are also other DM goals, such as clustering, which uses unsupervised learning to group similar items. Self-organizing maps is an example of a popular clustering technique. Data warehouses (DW) are another popular BI concept that consists in data repositories for accessing data from different sources, organized in a unique schema and place in order to facilitate information extraction to produce knowledge.

A DSS is an information system that provides assistance in supporting business decision making (Turban et al., 2011). While often used as a synonym of BI, DSS can also use expert knowledge rather than data-driven models (e.g., group DSS). New concepts are emerging related to DSS and BI, such as the adaptive business intelligence, which aims to reduce the gap between supporting and making the decision by adding adaptive prediction and optimization modules to classical BI systems (Michalewicz et al., 2008).

2.4. Literature analysis

A literature review of a set of articles enables to analyze a given subject and identify trends of research and possible gaps that can lead to new studies and discoveries (Levy & Ellis, 2006). In fact, it is considered a critical step and a baseline to unveil new insights on a research subject, thus an enabler and driver of new findings. Such relevance is expressed through the numerous publications on conducting literature reviews across the different sciences (Cronin, Ryan, & Coughlan, 2008; Jesson & Lacey, 2006).

Traditionally, exhaustive literature analyses demand considerable amount of efforts from researchers, in pursuit for the state of the art on a given subject which may serve as a driver on new research. New technologies enabled online library databases, easy to access from any location, offering researchers an enormous amount of available articles. The usage of search queries provided by such libraries facilitates the retrieval of articles on a given subject; however, the high volumes of articles returned present the challenging task of reading the contents of each paper, even though smaller parts of the articles (e.g., title, abstract, keywords) may provide a lead on the research conducted. To address this difficulty, a few TM approaches have been proposed recently for analyzing literature.

Table 1 summarizes four frameworks for literature analysis that use different techniques. The first (Jourdan, Rainer, & Marshall, 2008) uses a traditional human effort approach, while the remaining three conducted TM literature analyses. Finally, in the last row, the characteristics of the present approach are also displayed, to allow a straightforward comparison. The four frameworks were chosen to represent different and recent literature analysis methodologies on research areas closely related to BI, which is the focus of the present research, here applied to the banking industry. It was also taken into account that each of those frameworks should mention the criteria and methods of research, expressed in the columns of Table 1, to enable comparing different approaches with the proposed method.

The work of Jourdan et al. (2008) provides a general review on BI and requires that at least two humans (sometimes three, in cases of different opinion from the two authors) manually read every of the 167 articles. One main advantage of such approach is the fact that a human reader can readily understand the meaning of a word by the context of the remaining text (e.g., "senior" may refer to older people, or to senior professionals, which may not be so old), while an automated approach cannot. However, the time needed to conduct such a manual analysis prohibits it from being applied to a large number of manuscripts.

The remaining three frameworks use TM approaches, analyzing a number of articles greater than a thousand. The work of Sunikka and Bragge (2012) focus on two subjects, still it performs a separate analysis of both results, while the remaining two focus

Table 1
Examples of relevant frameworks for literature analysis and the proposed approach.

Reference	Areas of research	Nr. articles	Nr. journals	Search period	Search query	Description of the techniques used
Jourdan et al. (2008)	Business intelligence	167	10	1997–2006	Business analytics OR business intelligence OR data mining OR data warehousing	Classification by research strategy by 2–3 human coders; Classification of articles by topic using brainstorming and discussions
Delen and Crossland (2008)	Management information systems	1123	3	1994–2005	All articles from the 3 journals	TM on title and abstract of articles, using singular value decomposition to reduce the size of the document term matrix, and then clustering using an expectation–maximization algorithm
Sunikka and Bragge (2012)	Customization and personalization	883 + 1544 (customization + personalization)	457 + 664 (customization + personalization)	1986–2009	Customization OR personalization (two separate searches)	TM (tool: VantagePoint) on articles keywords, using Aduna cluster map of the keywords used; Autocorrelation map of authors with some selected keywords
Bragge et al. (2012)	Multiple criteria decision making	15198	Usage of the Web of Science database (not mentioned the different publication titles found)	1970–2007	Multiple criteria OR multiple attribute OR multiple objective OR goal programming OR vector optimization	TM on articles keywords, using auto-correlation maps based on the 60 most cited authors per decade
Proposed approach	BI in banking	219	14	2002–2013	Described in Section 3.2	TM, using dictionaries to reduce the size of search-space, and then the LDA to group articles

on just a subject of analysis. As a result, the present work, which analyzes BI applications in banking, is the only one from Table 1 using a search query with a conjunction ("AND") element (explained further ahead in Section 3.2). This justifies the significantly smaller number of articles analyzed in the present article, even though the procedure presented is scalable.

Another difference between the TM approaches is the procedure used to reduce the search space to a manageable number of terms: Delen and Crossland (2008) analyzed the abstract, discarding the keywords, and used a singular value decomposition, while the remaining two frameworks considered only the keywords. The former authors argued that the keywords are generally mentioned in the abstract, and even that some authors select keywords that they would like to be associated with their work. However, it can be argued that the approach of Delen and Crossland (2008) discards relevant terms composed of more than one word such as "data mining" or "decision support systems", which are included in the present work through the usage of a specific domain dictionary, overcoming both this limitation (Han, Wang, & El-Kishky, 2014) and the one associated with the usage of just the keywords mentioned above. It should be noted also that while all the TM analysis in Table 1 perform clustering analyses, none of the mentioned works used the LDA clustering algorithm. Also no evidence was found of literature analysis using this technique. Its computational complexity of probabilistic inference for finding a large number of topics is considered NP-hard, while unsupervised learning using expectation–maximization requires the repeated computation of marginal probabilities of what topics are present in the documents. LDA model is considered one of the most important probabilistic models in widespread use today (Sontag & Roy, 2011).

3. Materials and methods

3.1. Journal selection

Given the emphasis on technology aspects of BI applications to the banking industry, the articles were chosen from journals more related with technology rather than management. Nevertheless, with the popularity increase of BI (in both industry and research), the corresponding publications have boomed, making a literature review in this domain a challenging task. To select the relevant publications where to search, the focus was set on finding the most influential peer-reviewed journals on BI applications to business and management, within a recent time frame that includes around one decade (last twelve years, 2002–2013).

Instead of defining one specific publication metric criterion (e.g., by using impact factor or number of citations) for selecting journals, the choices were based on literature reviews and publication analysis. It should be noted that there are studies that criticize impact factor rankings accuracy, such as Andersen, Belmont, and Cho (2006) that analyzed the impact factor of the journal citation reports (JCR) published by the Institute of Scientific Information (ISI). The value of survey and review studies on the subjects in analysis is that the journals selected through those were already validated through a deeper analysis of publications rather than just citations considerations. Few articles evaluate the influence of journals on the information systems (IS) domain. To assist in the selection of journals, three review articles were chosen, one from each third of the time frame (i.e., 2002–2005, 2006–2009 and 2010–2013). The criteria for such review article selection included: consider only journal articles but with no restriction regarding journal title; consider reviews on related areas to this study (i.e., BI and banking); consider articles with a list of journals used in their review and the number of articles retrieved for each journal in such list.

The oldest of them (Huang & Hsu, 2005) analyzed publication productivity in IS from 1999 to 2003. Their study also used three other journal reviews as a base of work, and selected 12 reference journals on the field of IS. Ngai, Xiu, and Chau (2009) analyzed literature from 2000 to 2006 on a more specific field related to the research presented here, DM and its applications to CRM, by reviewing 87 articles from 28 different journals. Finally, the more recent study of Chen, Chiang, and Storey (2012) focused in BI and analytics and its impact to business through a literature review on those subjects in the past decade (2000–2011). Those three studies were also selected in order to be complementary in terms of the domains of IS, CRM and BI, thus providing with a vaster choice of journals.

The criteria for selection of journals is to include every journal used in at least one of the three reviews mentioned, except for:

- non technical journals, which are more related to business and management, were excluded, such as the MIT Sloan Management and the Harvard Business Review; and
- since the review of Ngai et al. (2009) presented a very large list of references, journals cited only once or twice in this review were discarded, except for Information & Management, which was selected since it was also used in the review of Huang and Hsu (2005).

The final result is a list of fourteen journals from eight different publishers (Table 2) that set the sources in this study for searching relevant articles.

Table 2
Journals selected and search results.

Journal	Publisher (search engine)	[1]☆	[2]☆	[3]☆	Hits
Expert systems with applications	Elsevier (SciVerse Science Direct)		X		126
Decision support systems		X	X	X	25
European journal of operational research			X		48
Information & management		X	X		2
IEEE Trans. knowledge and data engineering	IEEE (IEEE Xplore)		X		2
IEEE intelligent systems			X		2
Information systems research	INFORMS (Informs online)	X		X	0
Journal on computing				X	1
Journal of the association for IS	Association for IS (AIS elect. library)	X		X	1
Communications of the association of IS				X	1
Data mining and knowledge discovery	Springer (Springer link)		X		4
Communications of the ACM	ACM (ACM digital library)	X			1
Journal of management IS	M. E. Sharpe (jmis-web.org)	X		X	3
MIS Quarterly	MIS research center (misq.org)	X		X	3
					Total: 219

A "X" indicates the journal was used as a source in the corresponding column reference.
☆ [1] = Huang and Hsu (2005); [2] = Ngai et al. (2009); [3] = Chen et al. (2012).

3.2. Article search

The searches were performed through each of the publishers online search engine. Most of the search engines are optimized, allowing complex search queries through the use of specific fields and Boolean operators "AND" and "OR". It should be noted that a few of the search engines did not provide the flexibility needed (e. g. Boolean AND/OR operators, search field specification), thus the search was partitioned in searches within the main search.

The query used is the same for every journal, and consists in a Boolean expression containing two OR connected expressions, one for banking terms and another for BI, and both are connected through an AND, meaning that any article should include at least one banking term and another BI term:

> (banking OR bank OR credit) AND ("business intelligence" OR "data mining" OR "decision support system" OR "knowledge discovery" OR "business analytics" OR forecasting OR "modern optimization" OR modeling OR "machine learning" OR "artificial intelligence" OR prediction OR predictive).

The composition of such query is always subjective. To reduce such subjectivity, the authors and two banking domain experts conducted several broader searches with single keywords such as "banking" and "business intelligence", reaching to a consensus consisting in the query presented above. Some remarks should be mentioned. First, credit is a subject on its own, although closely related to banking, so it is considered in the search. For BI terms, the choice is on high-level concepts, discarding specific methods and techniques such as data warehouses, neural networks and decision tables.

All searches were performed in 2014, with the corresponding journal 2013 volumes already published, and included only the article title, abstract and keywords, since those are the most visible article areas where, if a certain concept is relevant, should be mentioned. It should also be noted that some online databases search engines only allow searching in these types of contents, rendering unfeasible a full-text search.

The first search results included a total of 240 articles. A manual analysis, consisting in reading each title, abstract and keywords, detected several articles where the terms occurred with a different meaning, such as "blood bank" or "credit" mentioned in a non-financial context. This manual pruning led to a pool of 219 articles. Table 2 shows each journal contribution in terms of search hits (where each hit denotes an article).

3.3. Text mining for literature review

Since 219 articles is quite a large number for a manual analysis, in this study TM was used to facilitate in producing organized information to analyze the literature. Considering the goal is set specifically on applications of BI to banking, in order to keep the scope within a manageable list of terms, it makes sense to define a dictionary that encompasses both BI and banking more common terms and concepts, rather than let the TM algorithms to search, group and count words indiscriminately. Hence, two dictionaries were defined, one for banking and another for BI, each of them containing a list of terms composed of one or more words (n-grams).

Stemming is a technique often applied in TM, in order to reduce similar words to a unique term (e.g., "banking" and "banks" are transformed in "bank"). Rather than just performing usual stemming, an extended list of related terms was created that includes other concepts in the same domain. For example, "loyalty" and "lifetime value" are the opposite of "defection" and "churning", but all of them concern with the problem of customer "retention", thus all of them were grouped under this reduced term.

Both the definition of dictionaries and the grouping of terms under a unique reduced term are subjective. To reduce this subjectivity, the three authors of this paper analyzed all decisions. It should be mentioned that, while all three authors are experienced in information systems and BI, one of them is a full-time information systems manager in a retail bank since 2001, having coordinated projects in distinct areas such as marketing and risk. Additionally, two experienced banking professionals in different areas were consulted (one of them has 3 years as a technical Contact Center support, and 10 years as a technician in Marketing, while the other has 6 years in the Commercial Area, plus one year in the Risk Department).

To further extend the validation of the dictionaries, considering these will guide the entire TM approach, and also the relatively small number of articles, each of the articles was analyzed in terms of title, abstract and keywords for prospecting the adequacy of the terms in each dictionary for the articles. For a large number of articles, an alternative would be to pick up a reasonable randomly selected number of articles for validating the dictionary.

The resulting dictionaries and grouping of terms defined for banking and BI are shown in Tables 3 and 4, respectively.[1] Some considerations should be made about the dictionaries. First, both the terms "banking" and "business intelligence" were not included, since are the two broader terms that characterize every article found. An also relevant term that was not included is risk and its variations, since it is a research subject by its own and it is implicit to other specific banking domains such as credit scoring, fraud and bankruptcy detection and churning (considering the risk of losing customers).

For the literature analysis of the articles collected, the full-text is considered. Since this analysis encompasses two distinct areas, BI and banking, it is likely that some of the terms from the dictionaries may not be present in the title + abstract + keyword, for they are not the main focus of the research (e.g., certain BI techniques applied). Also the full-text analysis allows a better evaluation of term frequencies, since a term expressed numerous times through an article is probably more relevant than another that is only mentioned in the abstract. The exception is the references section, which was pruned from all articles. By proceeding this way, it is assured that no term from the dictionary will match any from publication titles cited in the article. If some term in the dictionary is relevant for some study, then it is likely mentioned through the article text.

The TM procedure adopted included several steps over the corpus of documents, for striping extra whitespaces, converting all words in lowercase, reducing the terms of the dictionary to a common term, and finally defining the document term matrix, which is a bi-dimensional representation used as an input for the LDA (the dimensions are the articles and the terms, and each cell contains the frequency which $term_x$ appears in $article_y$).

There are a wide variety of tools and software that can be used to perform TM. For this review, the **R** statistical tool was chosen (<www.r-project.org>), since it is open source and provides a high flexibility through the installation of packages. In particular, the **tm** package chosen was adopted, since it offers a large number of functions for managing text documents and provides an abstraction of the process of document manipulation (Meyer, Hornik, & Feinerer, 2008).

For demonstration purposes, part of the R code is exposed (Code 1). This code was used first to create the corpus of documents based on a path containing all documents (line 1), perform cleaning by removing extra spaces (line 2) and converting all words to lowercase (line 3). Then the list of equivalent terms for reducing them to a common unique term (Tables 3 and 4) are loaded into

[1] Also available online at: <https://fenix.iscte.pt/homepage/smcmo@iscte.pt/BIinBankingReview>.

Table 3
Dictionary for the banking domain.

Reduced term	Similar terms or from the same domain☆
Bankruptcy	systemic risk, crisis, contagion, financial distress, solvency
Branches	bank branch, banking center, financial center
Central bank	central banks
Credit	loan
crm	customer relationship management
Deposit	savings, bank account, bank accounts, deposits
e-bank	e-banking, electronic banking, electronic bank, homebanking, homebank, home banking, home bank, internet banking, internet bank, online banking, online bank, netbanking, net banking, netbank, net bank, mobile banking, m-banking, m-bank, sms banking, sms bank, mobile bank, technology acceptance, tam
Fraud	fraud detection, fraud evaluation, fraud detect, fraud prevention, fraud risk, money laundering
Interest rate	interest rates, annual percentage rate, annual percentage rates, bank rate, bank rates, borrowing rate, borrowing rates, lending rate, lending rates, prime rate, prime rates, rate of interest, rates of interest
Investment	investments
Retention	defection, churning, churn, loyalty, lifetime value
Segmentation	client segment, profiling, client profile, client profiles, customer profile, customer profiles
Stocks	stock price, stock exchange, stock market, commodity, commodities
Targeting	direct marketing, database marketing, telemarketing, cross-selling

☆ All terms are in lower case and separated by commas.

a lookup table (line 5) and the reduced terms (first element of the R lookup table list) are checked against the dictionaries previously loaded through the intersect function, constituting the reduced terms dictionary (line 6). Next follows a computationally expensive mapping to perform a stem function which uses the terms in the lookup table to reduce them to a common term (line 7). Finally line 10 defines a function to allow tokens up to three words (the maximum words for the terms in the considered dictionaries) and line 11 builds the document term matrix (Delen & Crossland, 2008; Meyer et al., 2008).

R Code 1: Creating the corpus, cleaning it and build the document term matrix

```
1  articles <- Corpus(DirSource(pathOut), readerControl = list(
       language = "en"))
2  articles <- tm_map(articles, stripWhitespace) # remove spaces
3  articles <- tm_map(articles, tolower) # lower case
4
5  termDomains <- stemFromFileLoad("equivalent.txt")
6  reducedDictionary <- as.vector(intersect(unique(termDomains
       [[1]]),dictionary))
7  articles <- tm_map(articles, function(x) stemFromFile(doc=x,
       equivTerms=termDomains))
8
9  # create the document term matrix
10 phraseTokenizer <- function(x) RWeka::NGramTokenizer(x,
       Weka_control(min = 1, max = 3)) # terms up to 3 words
11 dtm <- DocumentTermMatrix(articles, control = list(tokenize =
        phraseTokenizer, dictionary = reducedDictionary))
```

For a general characterization of the literature, the frequency of each term was obtained for the combined dictionary including both Tables 3 and 4. Also a word cloud was designed to allow a visual interpretation of the obtained results.

Table 4
Dictionary for the BI domain.

Reduced term	Similar terms or from the same domain☆
Adaptive	adaptive
Analytic	analytics, data science, data sciences
Artificial intelligence	machine learning, intelligent agent, intelligent agents
Association rule	association rules
Big data	terabytes, massive data
cbr	case-based reasoning
Classification	classifier, classifiers
Cluster	clusters, clustering, clusterings
Data mining	data miner, datamining
Data warehouse	datawarehouse, datawarehouses, data warehouses
Decision support system	decision support systems, expert system, expert systems
Decision table	decision tables
Decision tree	decision trees, random forest, random forests
Genetic algorithm	genetic algorithms, genetic programming
Knowledge discovery	knowledge discovering
Modeling	modeling, data model
Naive bayes	naivebayes, bayesian
Neural network	neural networks, artificial network, artificial networks, multilayer perceptron, multilayer perceptrons
Optimization	optimize
Predict	prediction, predictive, predicting, forecasting, forecast
Regression	time series, time series
Self-organizing map	self-organizing feature map, self organizing map, sofm, kohonen map, kohonen network
Set theory	rough set, rough sets, fuzzy set, fuzzy sets, sets theory
Support vector machine	support vector machines

☆ All terms are in lower case and separated by commas.

3.4. Classification of topics

To obtain a structure that groups articles in order to allow a deeper analysis, the R package **topicmodels** is a logical choice, since it takes advantage of the data structures produced by the **tm** package in order to provide basic infrastructure for fitting topic models (Hornik & Grün, 2011).

Within the **topicmodels** package, the latent Dirichlet allocation (LDA) algorithm (Blei et al., 2003) is implemented and can be applied by receiving just two parameters, the document term matrix created for the TM and the desired number of topics. The result is a complex structure from which can be obtained the topics and terms that define it, characterized through a beta (β)

distribution computed for each term for a given topic. Also for each article, it can be obtained the likelihood of matching it to each of the topics. In this study, only the most probable topic according to LDA for a given article was considered. Also, the three most significant terms for characterizing each topic according to the β distribution will be analyzed.

As stated previously, one necessary parameter for LDA is the number of topics. Following the approach of Delen and Crossland (2008), this value was set to half of the terms considered. To simplify the analysis conducted, the topics will be presented in tables referring the number of articles in each topic published through the considered period of the last twelve years.

4. Results and analysis

The results are presented in two Sections: in the first, the results are analyzed based on term frequencies for the whole 219 articles collected. The respective results are shown using a table and a word cloud, which uses a larger font size for the most frequent terms. After the global analysis, the topics generated with LDA are displayed and analyzed. In the second Section, a representative article for each topic is selected and scrutinized with the goal to understand if the trend suggested by the topic characterization is aligned with such article.

4.1. Text mining and latent Dirichlet allocation topics

The global results are presented in Table 5, with a total of 38 terms. The respective word cloud is shown in Fig. 1. Overall, the BI terms are much more evenly distributed: credit is the top term, followed by four BI terms, and next comes two banking related terms, fraud and bankruptcy. This is an expected result, since banking defines the problems being addressed, to which many different BI solutions can be applied, including more specific algorithms and tools or more general approaches, such as modeling and knowledge discovery. Only three of the fourteen bank terms are among the eleven most cited. This global analysis allows taking a glimpse on what seems to be an interesting hypothesis to test: most of the BI research efforts are directed toward a few (and probably more relevant) of the banking domains. The word cloud on Fig. 1 seems to help support this claim, since it makes more visible that credit is the dominant term, followed by several BI terms, and only then comes the next two banking problems: fraud and bankruptcy. The second level analysis, using a LDA parameterized to 19 topics and presented in Table 6, is more interesting for this study, as it allows to relate BI terms to banking problems, thus identifying research trends and eventually gaps for further research. Each topic is presented in horizontal lines, with the column labeled "topics" presenting the most relevant terms and β distribution values (converted to positives, since they are used only for comparison purposes) in respect to a given topic (defined by the row). The number of articles column presents the number of articles that were included in the topic and that were published through the analyzed twelve year period.

The results of Table 6 show an increasing although not steady interest in BI applied to banking. For each topic, there is always a dominant term, with a β value that matches it to closer to a certain banking problem or to a type of BI technique, tool or context. Given that the three most relevant terms are shown for each topic, most of them have at least one of the top 3 terms belonging to banking and another to BI, which enables to analyze each topic as a BI application to banking. Still, there are four topics that focus specifically on BI (topics 3, 9, 14 and 16), with the three dominant terms matching all BI terms, and one equivalent topic for banking (topic 5).

Table 5
Most relevant term frequencies for the BI applied to banking.

#	Term	Frequency
1.	Credit	7299
2.	Predict	4053
3.	Regression	2022
4.	Decision support system	1765
5.	Neural network	1735
6.	Fraud	1358
7.	Bankruptcy	1152
8.	Decision tree	997
9.	Data mining	874
10.	Cluster	839
11.	Modeling	793
12.	e-bank	621
13.	Investment	545
14.	Retention	536
15.	Interest rate	493
16.	Artificial intelligence	444
17.	Deposit	389
18.	Optimization	382
19.	Support vector machine	379
20.	Classification	336
21.	Set theory	335
22.	Data warehouse	267
23.	Naive bayes	261
24.	cbr	260
25.	Genetic algorithm	235
26.	Adaptive	204
27.	Association rule	168
28.	Branches	158
29.	Segmentation	157
30.	Stocks	139
31.	Targeting	118
32.	crm	108
33.	Central bank	67
34.	Knowledge discovery	64
35.	Decision table	47
36.	Analytic	42
37.	Self-organizing map	33
38.	Big data	3

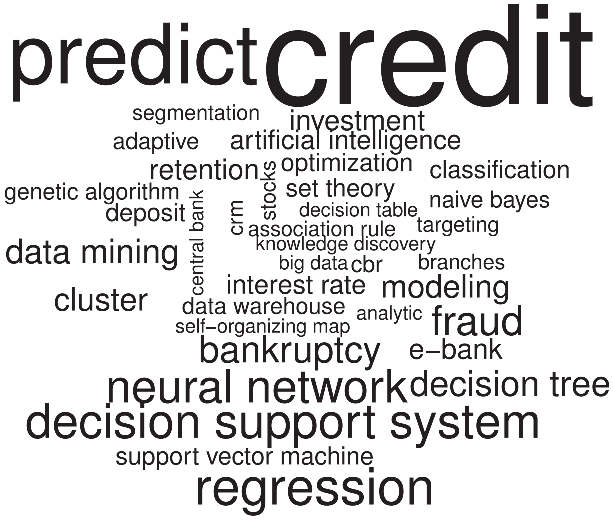


Fig. 1. Word cloud for BI applied to banking.

The topic best identified with credit gets 70 matching articles, although second and third terms for this topic, predict and segmentation, have a significantly higher β value (greater than 3.3), meaning that its relation is not so tight. This puts emphasizes on numerous applications of BI to benefit credit business and risk evaluation. In fact, credit gets into the top 3 of six more topics

Table 6
Relevant topics for BI applied to banking.

Topic	#	1st Term		2nd Term		3rd Term		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
		term	*β*	*term*	*β*	*term*	*β*												
1.	70	Credit	0.08	Predict	3.34	Segmentation	4.37	2	0	2	4	4	6	3	6	13	13	11	6
2.	25	Predict	0.15	Set theory	2.58	Stocks	3.65	0	0	1	3	1	2	1	8	1	2	5	1
3.	22	Neural network	0.85	Predict	1.18	Support vector machine	2.43	0	0	2	1	1	3	0	6	0	4	4	1
4.	12	Credit	0.80	Neural network	1.59	Adaptive	2.41	1	0	0	1	1	2	0	2	1	0	3	1
5.	12	Retention	0.89	Interest rate	1.07	Targeting	2.50	1	0	2	1	1	1	1	0	0	2	2	1
6.	9	Fraud	0.26	Classification	3.06	Regression	3.43	0	0	1	0	0	0	0	2	0	2	3	1
7.	8	Optimization	0.96	Deposit	1.19	Branches	1.67	0	0	0	1	0	0	0	2	4	0	1	0
8.	7	Decision tree	0.57	Classification	1.83	Credit	3.20	0	0	2	1	0	1	1	1	0	0	1	0
9.	7	Decision support system	0.19	Naive bayes	2.06	Adaptive	4.39	1	1	0	0	0	0	0	3	1	0	1	0
10.	7	Bankruptcy	0.25	Predict	2.18	Deposit	2.83	0	0	0	0	2	1	0	0	0	2	1	1
11.	7	Regression	0.09	Predict	3.10	Credit	4.14	0	1	0	0	1	1	0	1	0	1	1	1
12.	6	Cluster	0.13	Credit	3.08	Neural network	4.09	0	0	0	0	1	1	1	1	2	0	0	0
13.	5	e-bank	0.09	Decision support system	3.09	predict	4.26	0	0	0	0	1	0	0	1	2	0	0	1
14.	5	Artificial intelligence	0.69	Association rule	1.20	Decision table	2.75	0	1	1	0	0	0	0	1	0	0	0	2
15.	5	Modeling	0.36	Credit	1.40	optimization	3.96	0	0	0	0	0	0	0	0	0	2	2	1
16.	5	Data mining	0.28	Decision support system	2.56	Knowledge discovery	2.88	0	0	0	0	0	0	0	2	0	2	1	0
17.	4	Investment	0.03	Predict	4.35	Analytic	4.70	0	0	0	1	0	0	0	0	0	1	2	0
18.	2	cbr	0.35	Credit	2.00	Decision support system	2.68	0	0	0	0	0	0	0	1	1	0	0	0
19.	1	Data warehouse	0.18	Decision support system	2.71	Investment	3.56	0	0	1	0	0	0	0	0	0	0	0	0

β corresponds to the correlation between the topic and term; # is the number of articles in the topic.

while being also the top term for the fourth topic, confirming the diversity of this subject.

The year of 2008 seems to be an outlier, containing a smaller number of articles when compared to its surrounding years (only seven articles). Probably the global financial crisis, which culminated in 2008 with the failure of major financial institutions, also helped to boom research in the following year of 2009, with a total of 37 articles for the set analyzed. The second topic, with 25 articles in total, had eight publications just for 2009, the highest number for the topic in the twelve years studied. Furthermore, the topic includes stocks as the third more relevant term, while predict and set theory are the first and second, respectively.

Concerning the banking domain, fraud and bankruptcy prediction get a match of nine (topic 6) and seven articles (topic 10) respectively, although most of them are recent, which can be also a result of the financial crisis. Neural networks are the dominant specific learning technique adopted, topping the third topic with more articles (22). Topic 5, with 12 articles, has the three most relevant terms for banking only: retention, interest rate and targeting. This is an interesting topic, since it shows an evenly distributed publication number for the period considered, with most years having just one or two articles, with the exception of the years 2003, 2009 and 2010. Considering that the three terms have significantly close β values, one can hypothesize that by targeting customers with attractive interest rates in the products offered may also serve the purpose of retaining them, thus reducing churn.

DSS are a thematic rather old, but far from outdated. From the topics in Table 6, it is possible to confirm the wide reference to DSS by counting five occurrences of the term decision support systems in different topics, with an apparent even distribution in the years considered. On the other hand, data mining has only one reference in the top 3 terms for every topics, which is on topic 16, with just 5 articles. This an unexpected result, since the state of the art for prediction is the application of data mining techniques. Nevertheless, it should be noted that dominant data mining techniques include neural network and regression, which have several references spread through the 19 topics.

In respect for the four topics which are best identified by three terms all related to BI, and some other topics, one may hypothesize that it is probably an indication that the main focus is on BI applications, not evaluating in-depth benefits to banking.

Looking at the end of the table, data warehouse is surprisingly low in publications, although banks continue to invest on those systems as a way to unify data otherwise spread through an organization. Other recently proposed terms for BI, such as adaptive (Michalewicz, Schmidt, Michalewicz, & Chiriac, 2005) (mentioned in topics 4 and 9, but only as the third most relevant term in both cases), still get few publications and others are not even on any of the top 3 terms list (e.g., big data).

4.2. Analysis of representative articles per topic

In previous Section, LDA was applied to unveil topics which group articles, characterized by the terms identified on Table 6, suggesting the major trends of research concerning BI applications to banking. However, such automated approach has a significant limitation (Thomas, McNaught, & Ananiadou, 2011): document clustering is completely dependent on the technique used for creating the clusters, which is based on term identification; the problem consists in terms with different meanings based on the remaining text (e.g., risk may refer to credit default risk or to bankruptcy risk). In this Section, this issue is addressed by identifying the most representative articles for each topic. Then a full text manual analysis of each of the nineteen articles is performed in order to confirm or not the hypotheses suggested by the topics found. Table 7 identifies the articles chosen.

Considering the fact that the three most relevant terms were selected for characterizing the topics (Table 6), in order to select the most relevant article two metrics were considered, by the following order of relevance: the number of different terms mentioned in each article (from one to the whole three most relevant terms, displayed for each topic), and the total number of times each of the three terms occurred, regardless of the specific term. Such procedure is best explained through an example: topic 10

Table 7
Core article per topic.

Topic	Article	Different terms	Frequency
1.	Chi and Hsu (2012)	3	186
2.	Kumar and Ravi (2007)	3	189
3.	Huang et al. (2004)	3	221
4.	Malhotra and Malhotra (2002)	3	175
5.	Prinzie and Van den Poel (2006)	3	27
6.	Abbasi et al. (2012)	3	426
7.	Azadeh et al. (2012)	3	27
8.	Sinha and May (2004)	3	208
9.	Ben-David and Frank (2009)	2	89
10.	Hu et al. (2012)	3	167
11.	Zhao et al. (2011)	3	179
12.	Lim and Sohn (2007)	3	71
13.	Gu et al. (2009)	3	162
14.	Hsieh (2004)	1	52
15.	Liu et al. (2012)	3	67
16.	Chen et al. (2011)	3	64
17.	Soper et al. (2012)	3	190
18.	Park et al. (2009)	3	123
19.	Hwang et al. (2004)	3	255

grouped a total of seven articles; for each of those articles, the number of occurrences for each of the three most relevant terms was extracted from the document term matrix (hence resulting in a sub-matrix limited to the seven articles and the terms "bankruptcy", "predict" and "deposit" from topic 10); next, the sub-matrix was ordered decreasingly by the number of different terms and the number of times each of the three terms occurred, resulting in four articles which mentioned the three terms, and three referring only two terms; from those four articles, the one with higher frequency of terms was selected, which was the study of Hu et al. (2012).

Topic 1 is best represented by the work of Chi and Hsu (2012), which is a typical research for predicting credit risk of default, suiting perfectly in the two most relevant terms, "credit" and "predict" (Table 6), whereas "segmentation" (the third most relevant term) is also used in their work for defining homogeneous risk groups. This study confirms the hypothesis arose from the previous Section, which pointed this research trend of predicting credit behavior as the major application for BI to banking.

By looking at the title of the article best identified with topic 2, the work of Kumar and Ravi (2007), "Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review", one could argue why it did not match topic ten, which focus precisely on predicting bankruptcy. However, a deeper analysis of such article revealed it is a work more focused in applying set theory as well as other techniques for comparing their performance when addressing a prediction problem, which happens to be bankruptcy. In fact, the third term for topic ten is "deposit", while for topic two is "stocks", which is much more related to bankruptcy: it is mentioned several times through the text.

Topic 3 is more focused on the techniques applied rather than the banking problem itself, which fits perfectly with the chosen article (Huang et al., 2004). This study is focused on comparing machine learning techniques, using corporate credit rating for benchmarking their performance.

While seventeen of the nineteen topics were best matched by one article which referred the three most relevant terms (from Table 6), there remain two for which the best article only contained two of the three most relevant terms (topic 9) or just one (topic 14). In case of topic 9, the significantly higher β value for "adaptive", the third term, more than twice as the second term, may justify the result. However, topic 14 shows clearly a weakness of this approach: although it groups five articles, none is related with more than one term of the three most relevant (e.g., the work of Hsieh (2004) is dedicated to association rules, without even mentioning the remaining two terms). One may hypothesize that this is a direct consequence of the ill-posed problem of clustering: the data-driven nature of clustering makes it very difficult to correctly find clusters in the given data (Jain, 2010). LDA faces the same challenge of other clustering algorithms, implying that there will inevitably exist articles that cannot match to any of the existing topics, leading to issues such as the one in topic 14.

5. Conclusions

This literature analysis paper focused on the main banking problems and BI solutions used to solve them. Banking is a competitive industry where innovation thrives, due to the importance of this sector for the economy, thus making it an attractive field for researchers. Banking is also a domain that generates large amount of data and where BI applications can potentially benefit business, increasing the visibility and recognition of research achievements. This recent analysis encompassed the last twelve years (2002–2013), being a period that includes the effect of the global financial crisis and its impact on research on this sector. Thus, this study can potentially benefit researchers by allowing the identification of new research trends and possible gaps for future research.

For analyzing literature, a text mining approach using the latent Dirichlet allocation was performed. As a result, several topics grouping articles were found, being each of those topics characterized by the three most relevant terms. Such topics suggested several research trends. However, intrinsic limitations of clustering algorithms such as LDA have lead efforts toward validating the hypotheses for relations between the several terms and corresponding trends. To address this issue, the most relevant article for each topic was exhaustively analyzed, in order to confirm the hypothesized trends. Expert systems conducting automated text analyses may benefit from the suggested procedure with little additional human effort for analyzing the few more relevant documents.

The most relevant conclusion is that credit maintains its status as the dominant field of research in the banking industry. Other relevant banking subjects are fraud and bankruptcy, mainly for detection and prevention, thus mitigating risks taken by banks. Concerning BI, the main goal consists in prediction, rather than modeling and knowledge discovery, which emphasizes the importance of estimating what is going to happen on the future in order to better support decision-making. There are some studies that use banking problems to test and evaluate BI techniques and tools, but possibly not accounting for real business benefits for banks, since banking terms are lesser relevant for those articles. Regarding the evolution of publications per year, 2009 is a milestone year, triggering a boom in the research publications on the domains analyzed. Most likely, this effect was motivated by business pressures due to the global financial crisis. Still, through the time period studied, publications related with BI approaches applied to banking had a steady increase until 2012, indicating this is a domain application much studied. Nevertheless, research has diminished in 2013, although some lack of research in newer concepts such as big data may suggest there is still open room for research.

The results highlight some possibly interesting research gaps. DSS in banking is a subject far from exhaust. Emerging concepts such as adaptive BI and optimization can be applied to enhance DSS and improve banking efficiency in several areas. For example, targeting customers to sell deposits is an application domain where there seems to be a lack of research. Although some articles mention deposits and others targeting, none of these words top any of the topics computed. Still concerning customer domains, it is interesting to verify that CRM is not a top banking domain for BI applications. This comes as a surprise, since CRM is a subject where research has been quite active, although the results here presented show it is not the case for the banking industry.

With the intensifying global competition within the financial sector and namely involving the banking industry, CRM has become critical. Thus, future studies in this topic are paramount in order to understand clearly what is more successful according to the size and nature of the financial organizations being at stake. E-banking offers a wide spectrum of services to customers. Some involve non-transactional tasks such as viewing of account balances or recent transactions, downloading account and bank statements. Others demand real transactions like fund transfers, bill payments, loan applications and transactions, investments in stocks and bonds. Banks offering all these services online are becoming financial "supermarkets" and demand further research in this area. Mobile devices such as smartphones and tablets are at the forefront of electronic consumer products. Their penetration is increasing rapidly in a diverse range of markets across the globe. Mobile banking solutions are nowadays a major challenge to the banking sector in order to be able to adapt its approach to new customer demands and expectations. Hence, this is another important focus for further research.

Bankruptcy associated with systemic risk is also a recent interesting subject, with its visibility set to a high level thanks to the global financial crisis which is far from over. Furthermore, it is now known that prior to the crisis, systems failed to predict it, and prediction is precisely a top keyword for BI, thus applications for this case must also be enhanced in the next years in order to try to prevent future financial crisis.

Acknowledgments

The authors would like to thank the valuable effort of the two bank professionals who helped in defining both the search query and the dictionary used. Also, the authors wish to thank the anonymous reviewers for their helpful comments.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, 36.

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management, 18*, 59–88.

Amayri, O., & Bouguila, N. (2010). A study of spam filtering using support vector machines. *Artificial Intelligence Review, 34*, 73–108.

Andersen, J., Belmont, J., & Cho, C. T. (2006). Journal impact factor in the era of expanding literature. *Journal of Microbiology, Immunology, and Infection/Wei Mian yu Gan Ran za Zhi, 39*, 436–443.

Azadeh, A., Saberi, M., & Jiryaei, Z. (2012). An intelligent decision support system for forecasting and optimization of complex personnel attributes in a large bank. *Expert Systems with Applications, 39*, 12358–12370.

Ben-David, A., & Frank, E. (2009). Accuracy of machine learning models versus "hand crafted" expert systems–A credit scoring case study. *Expert Systems with Applications, 36*, 5264–5271.

Berger, A. N. (2003). The economic effects of technological progress: Evidence from the banking industry. *Journal of Money, Credit, and Banking, 35*, 141–176.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.

Bragge, J., Korhonen, P., Wallenius, H., & Wallenius, J. (2012). Scholarly communities of research in multiple criteria decision making: A bibliometric research profiling study. *International Journal of Information Technology & Decision Making, 11*, 401–426.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36.

Chen, J., Wu, G., Shen, L., & Ji, Z. (2011). Differentiated security levels for personal identifiable information in identity management system. *Expert Systems with Applications, 38*, 14156–14162.

Chi, B.-W., & Hsu, C.-C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications, 39*, 2650–2661.

Cronin, P., Ryan, F., & Coughlan, M. (2008). Undertaking a literature review: A step-by-step approach. *British Journal of Nursing, 17*, 38–43.

Dahlberg, T., Mallat, N., Ondrus, J., & Zmijewska, A. (2008). Past, present and future of mobile payments research: A literature review. *Electronic Commerce Research and Applications, 7*, 165–181.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems With Applications, 34*, 1707–1720.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49*, 76–82.

Fethi, M. D., & Pasiouras, F. (2010). Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey. *European Journal of Operational Research, 204*, 189–198.

Gu, J.-C., Lee, S.-C., & Suh, Y.-H. (2009). Determinants of behavioral intention to mobile banking. *Expert Systems with Applications, 36*, 11605–11616.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Han, J., Wang, C., & El-Kishky, A. (2014). Bringing structure to text: Mining phrases, entities, topics, and hierarchies. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1968). ACM.

Hornik, K., & Grün, B. (2011). Topicmodels: An r package for fitting topic models. *Journal of Statistical Software, 40*, 1–30.

Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications, 27*, 623–633.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*, 543–558.

Huang, H.-H., & Hsu, J. S.-C. (2005). An evaluation of publication productivity in information systems: 1999 to 2003. *Communications of the Association for Information Systems*, 15.

Hu, D., Zhao, J. L., Hua, Z., & Wong, M. (2012). Network-based modeling and analysis of systemic risk in banking systems. *MIS Quarterly*, 36.

Hwang, H.-G., Ku, C.-Y., Yen, D. C., & Cheng, C.-C. (2004). Critical factors influencing the adoption of data warehouse technology: A study of the banking industry in Taiwan. *Decision Support Systems, 37*, 1–21.

Jain, A. K. (2010). Data clustering: 50 Years beyond k-means. *Pattern Recognition Letters, 31*, 651–666.

Jesson, J. K., & Lacey, F. M. (2006). How to do (or not to do) a critical literature review. *Pharmacy Education, 6*, 139–148.

Jourdan, Z., Rainer, R. K., & Marshall, T. E. (2008). Business intelligence: An analysis of the literature. *Information Systems Management, 25*, 121–131.

Karakostas, B., Kardaras, D., & Papathanassiou, E. (2005). The state of crm adoption by the financial services in the UK: An empirical investigation. *Information & Management, 42*, 853–863.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research, 180*, 1–28.

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline, 9*, 181–212.

Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications, 32*, 427–431.

Lin, H.-F. (2011). An empirical investigation of mobile banking adoption: The effect of innovation attributes and knowledge-based trust. *International Journal of Information Management, 31*, 252–260.

Liu, Y., Zhang, H., Li, C., & Jiao, R. J. (2012). Workflow simulation for operational decision support using event graph through process mining. *Decision Support Systems, 52*, 685–697.

Malhotra, R., & Malhotra, D. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research, 136*, 190–211.

Marqués, A., García, V., & Sánchez, J. (2012). A literature review on the application of evolutionary computing to credit scoring. *Journal of the Operational Research Society, 64*, 1384–1399.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in r. *Journal of Statistical Software, 25*, 1–54.

Michalewicz, Z., & Michalewicz, M. (2008). Machine intelligence, adaptive business intelligence and natural intelligence. *Computational intelligence magazine* (Vol. 3, pp. 54–63). IEEE [research frontier].

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2005). Case study: An intelligent decision support system. *Intelligent systems* (Vol. 20, pp. 44–49). IEEE.

Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*, 559–569.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems With Applications, 36*, 2592–2602.

Park, Y.-J., Choi, E., & Park, S.-H. (2009). Two-step filtering datamining method integrating case-based reasoning and rule induction. *Expert Systems With Applications, 36*, 861–871.

Prinzie, A., & Van den Poel, D. (2006). Investigating purchasing-sequence patterns for financial services using Markov, mtd and mtdg models. *European Journal of Operational Research, 170*, 710–734.

Shuaibu, B. M., Norwawi, N. M., Selamat, M. H., & Al-Alwani, A. (2013). Systematic review of web application security development model. *Artificial Intelligence Review*, 1–18.

Shu, W., & Strassmann, P. A. (2005). Does information technology provide banks with profit? *Information & Management, 42*, 781–787.

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems, 21*, 249–280.

Sontag, D., & Roy, D. (2011). Complexity of inference in latent Dirichlet allocation. In *Advances in neural information processing systems* (pp. 1008–1016).

Soper, D. S., Demirkan, H., Goul, M., & St Louis, R. (2012). An empirical examination of the impact of ict investments on future levels of institutionalized democracy and foreign direct investment in emerging societies. *Journal of the Association for Information Systems, 13*, 116–149.

Soper, D. S., & Turel, O. (2012). An n-gram analysis of communications 2000–2010. *Communications of the ACM, 55*, 81–87.

Sunikka, A., & Bragge, J. (2012). Applying text-mining to personalization and customization research literature–who, what and where? *Expert Systems with Applications, 39*, 10049–10058.

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods, 2*, 1–14.

Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligence systems*. USA: Prentice Hall Press.

Vatanasombut, B., Igbaria, M., Stylianou, A. C., & Rodgers, W. (2008). Information systems continuance intention of web-based applications customers: The case of online banking. *Information & Management, 45*, 419–428.

Wilson, J. O., Casu, B., Girardone, C., & Molyneux, P. (2010). Emerging themes in banking: Recent literature and directions for future research. *The British Accounting Review, 42*, 153–169.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zhao, H., Sinha, A. P., & Bansal, G. (2011). An extended tuning method for cost-sensitive regression and forecasting. *Decision Support Systems, 51*, 372–383.

2.2 Article nr. #2

This article explores the bank telemarketing problem by enriching the problem dataset with context features, namely related to social and economic variables. It follows a feature reduction by selecting the most relevant features according to a mixed approach based on expert knowledge and automated data-based sensitivity analysis for assessing feature relevance. It also sets the baseline methods for the following experiments by comparing the procedures proposed against other well-known data analysis techniques, thus validating the benefits of the approach undertaken.

There are two main contributions of this work to the present thesis. First, the highly tuned model achieved valuable performance in terms of improving business metrics: by selecting the half most likely subscribers of the deposits, the model allows to reach around 79% of the successes. Also, some of the proposed context features were considered highly relevant for the model. Namely, the three month Euribor rate was considered the most relevant. Such finding proves the value of the context feature selection strategy proposed.

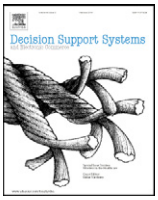
Article Details:

- Title: A data-driven approach to predict the success of bank telemarketing;
- Date: June 2014 (made available online since 13 March 2014);
- Journal: Decision Support Systems;
- Publisher: Elsevier.

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



A data-driven approach to predict the success of bank telemarketing

Sérgio Moro [a],*, Paulo Cortez [b], Paulo Rita [a]

[a] ISCTE-IUL, Business Research Unit (BRU-IUL), Lisboa, Portugal
[b] ALGORITMI Research Centre, Univ. of Minho, 4800-058 Guimarães, Portugal

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form 28 February 2014
Accepted 4 March 2014
Available online 13 March 2014

Keywords:
Bank deposits
Telemarketing
Savings
Classification
Neural networks
Variable selection

ABSTRACT

We propose a data mining (DM) approach to predict the success of telemarketing calls for selling bank long-term deposits. A Portuguese retail bank was addressed, with data collected from 2008 to 2013, thus including the effects of the recent financial crisis. We analyzed a large set of 150 features related with bank client, product and social-economic attributes. A semi-automatic feature selection was explored in the modeling phase, performed with the data prior to July 2012 and that allowed to select a reduced set of 22 features. We also compared four DM models: logistic regression, decision trees (DTs), neural network (NN) and support vector machine. Using two metrics, area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT), the four models were tested on an evaluation set, using the most recent data (after July 2012) and a rolling window scheme. The NN presented the best results (AUC = 0.8 and ALIFT = 0.7), allowing to reach 79% of the subscribers by selecting the half better classified clients. Also, two knowledge extraction methods, a sensitivity analysis and a DT, were applied to the NN model and revealed several key attributes (e.g., Euribor rate, direction of the call and bank agent experience). Such knowledge extraction confirmed the obtained model as credible and valuable for telemarketing campaign managers.

1. Introduction

Marketing selling campaigns constitute a typical strategy to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. Centralizing customer remote interactions in a contact center eases operational management of campaigns. Such centers allow communicating with customers through various channels, telephone (fixed-line or mobile) being one of the most widely used. Marketing operationalized through a contact center is called telemarketing due to the remoteness characteristic [16]. Contacts can be divided into inbound and outbound, depending on which side triggered the contact (client or contact center), with each case posing different challenges (e.g., outbound calls are often considered more intrusive). Technology enables rethinking marketing by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics, thus allowing us to build longer and tighter relations in alignment with business demand [28]. Also, it should be stressed that the task of selecting the best set of clients, i.e., that are more likely to subscribe a product, is considered NP-hard in Ref. [31].

Decision support systems (DSSs) use information technology to support managerial decision making. There are several DSSs sub-fields, such as personal and intelligent DSSs. Personal DSSs are related with small-scale systems that support a decision task of one manager, while intelligent DSSs use artificial intelligence techniques to support decisions [1]. Another related DSS concept is Business Intelligence (BI), which is an umbrella term that includes information technologies, such as data warehouses and data mining (DM), to support decision making using business data [32]. DM can play a key role in personal and intelligent DSSs, allowing the semi-automatic extraction of explanatory and predictive knowledge from raw data [34]. In particular, classification is the most common DM task [10] and the goal is to build a data-driven model that learns an unknown underlying function that maps several input variables, which characterize an item (e.g., bank client), with one labeled output target (e.g., type of bank deposit sell: "failure" or "success").

There are several classification models, such as the classical Logistic Regression (LR), decision trees (DTs) and the more recent neural networks (NNs) and support vector machines (SVMs) [13]. LR and DT have the advantage of fitting models that tend to be easily understood by humans, while also providing good predictions in classification tasks. NN and SVM are more flexible (i.e., no a priori restriction is imposed) when compared with classical statistical modeling (e.g., LR) or even DT, presenting learning capabilities that range from linear to complex nonlinear mappings. Due to such flexibility, NN and SVM tend to provide accurate predictions, but the obtained models are difficult to be understood by humans. However, these "black box" models can be opened by using a sensitivity analysis, which allows us to measure the importance and effect of particular input in the model output response [7]. When comparing DT, NN and SVM, several studies have shown different classification performances. For instance, SVM provided better

* Corresponding author.
E-mail address: scmoro@gmail.com (S. Moro).

http://dx.doi.org/10.1016/j.dss.2014.03.001
0167-9236/© 2014 Elsevier B.V. All rights reserved.

results in Refs. [6,8], comparable NN and SVM performances were obtained in Ref. [5], while DT outperformed NN and SVM in Ref. [24]. These differences in performance emphasize the impact of the problem context and provide a strong reason to test several techniques when addressing a problem before choosing one of them [9].

DSS and BI have been applied to banking in numerous domains, such as credit pricing [25].

However, the research is rather scarce in terms of the specific area of banking client targeting. For instance, Ref. [17] described the potential usefulness of DM techniques in marketing within Hong-Kong banking sector but no actual data-driven model was tested. The research of Ref. [19] identified clients for targeting at a major bank using pseudo-social networks based on relations (money transfers between stakeholders). Their approach offers an interesting alternative to traditional usage of business characteristics for modeling.

In previous work [23], we have explored data-driven models for modeling bank telemarketing success. Yet, we only achieved good models when using attributes that are only known on call execution, such as call duration. Thus, while providing interesting information for campaign managers, such models cannot be used for prediction. In what is more closely related with our approach, Ref. [15] analyzed how a mass media (e.g., radio and television) marketing campaign could affect the buying of a new bank product. The data was collected from an Iran bank, with a total of 22,427 customers related with a six month period, from January to July of 2006, when the mass media campaign was conducted. It was assumed that all customers who bought the product (7%) were influenced by the marketing campaign. Historical data allowed the extraction of a total of 85 input attributes related with recency, frequency and monetary features and the age of the client. A binary classification task was modeled using a SVM algorithm that was fed with 26 attributes (after a feature selection step), using 2/3 randomly selected customers for training and 1/3 for testing. The classification accuracy achieved was 81% and through a Lift analysis [3], such model could select 79% of the positive responders with just 40% of the customers. While these results are interesting, a robust validation was not conducted. Only one holdout run (train/test split) was considered. Also, such random split does not reflect the temporal dimension that a real prediction system would have to follow, i.e., using past patterns to fit the model in order to issue predictions for future client contacts.

In this paper, we propose a personal and intelligent DSS that can automatically predict the result of a phone call to sell long term deposits by using a DM approach. Such DSS is valuable to assist managers in prioritizing and selecting the next customers to be contacted during bank marketing campaigns. For instance, by using a Lift analysis that analyzes the probability of success and leaves to managers only the decision on how many customers to contact. As a consequence, the time and costs of such campaigns would be reduced. Also, by performing fewer and more effective phone calls, client stress and intrusiveness would be diminished. The main contributions of this work are:

- We focus on feature engineering, which is a key aspect in DM [10], and propose generic social and economic indicators in addition to the more commonly used bank client and product attributes, in a total of 150 analyzed features. In the modeling phase, a semi-automated process (based on business knowledge and a forward method) allowed to reduce the original set to 22 relevant features that are used by the DM models.
- We analyze a recent and large dataset (52,944 records) from a Portuguese bank. The data were collected from 2008 to 2013, thus including the effects of the global financial crisis that peaked in 2008.
- We compare four DM models (LR, DT, NN and SVM) using a realistic rolling window evaluation and two classification metrics. We also show how the best model (NN) could benefit the bank telemarketing business.

The paper is organized as follows: Section 2 presents the bank data and DM approach; Section 3 describes the experiments conducted and analyzes the obtained results; finally, conclusions are drawn in Section 4.

2. Materials and methods

2.1. Bank telemarketing data

This research focus on targeting through telemarketing phone calls to sell long-term deposits. Within a campaign, the human agents execute phone calls to a list of clients to sell the deposit (outbound) or, if meanwhile the client calls the contact-center for any other reason, he is asked to subscribe the deposit (inbound). Thus, the result is a binary unsuccessful or successful contact.

This study considers real data collected from a Portuguese retail bank, from May 2008 to June 2013, in a total of 52,944 phone contacts. The dataset is unbalanced, as only 6557 (12.38%) records are related with successes. For evaluation purposes, a time ordered split was initially performed, where the records were divided into training (four years) and test data (one year). The training data is used for feature and model selection and includes all contacts executed up to June 2012, in a total of 51,651 examples. The test data is used for measuring the prediction capabilities of the selected data-driven model, including the most recent 1293 contacts, from July 2012 to June 2013.

Each record included the output target, the contact outcome ({“failure”, “success”}), and candidate input features. These include telemarketing attributes (e.g., call direction), product details (e.g., interest rate offered) and client information (e.g., age). These records were enriched with social and economic influence features (e.g., unemployment variation rate), by gathering external data from the central bank of the Portuguese Republic statistical web site[1]. The merging of the two data sources led to a large set of potentially useful features, with a total of 150 attributes, which are scrutinized in Section 2.4.

2.2. Data mining models

In this work, we test four binary classification DM models, as implemented in the **rminer** package of the **R** tool [5]: logistic regression (LR), decision trees (DTs), neural network (NN) and support vector machine (SVM).

The LR is a popular choice (e.g., in credit scoring) that operates a smooth nonlinear logistic transformation over a multiple regression model and allows the estimation of class probabilities [33]:$p(c|\mathrm{x}_k) = \frac{1}{1+\exp\left(w_0+\sum_{i=1}^{M} w_i x_{k,i}\right)}$,where $p(c|\mathrm{x})$ denotes the probability of class c given the k-th input example $\mathrm{x}_k = (x_{k,1}, \ldots, x_{k,M})$ with M features and w_i denotes a weight factor, adjusted by the learning algorithm. Due to the additive linear combination of its independent variables (x), the model is easy to interpret. Yet, the model is quite rigid and cannot model adequately complex nonlinear relationships.

The DT is a branching structure that represents a set of rules, distinguishing values in a hierarchical form [2]. This representation can be translated into a set of IF–THEN rules, which are easy to understand by humans.

The multilayer perceptron is the most popular NN architecture [14]. We adopt a multilayer perceptron with one hidden layer of H hidden nodes and one output node. The H hyperparameter sets the model learning complexity. A NN with a value of $H = 0$ is equivalent to the LR model, while a high H value allows the NN to learn complex nonlinear relationships. For a given input x_k the state of the i-th neuron (s_i) is computed by: $s_i = f\left(w_{i,0} + \sum_{j \in P_i} w_{i,j} \times s_j\right)$,where P_i represents the set of nodes reaching node i; f is the logistic function; $w_{i,j}$ denotes the weight of the connection between nodes j and i; and $s_1 = x_{k,1}, \ldots, s_M = x_{k,M}$. Given that the logistic function is used, the

[1] http://www.bportugal.pt/EstatisticasWeb/Default.aspx?Lang=en-GB

output node automatically produces a probability estimate (ϵ[0,1]). The NN final solution is dependent of the choice of starting weights. As suggested in Ref. [13], to solve this issue, the **rminer** package uses an ensemble of N_r different trained networks and outputs the average of the individual predictions [13].

The SVM classifier [4] transforms the input $\mathrm{x} \in \Re^M$ space into a high m-dimensional feature space by using a nonlinear mapping that depends on a kernel. Then, the SVM finds the best linear separating hyperplane, related to a set of support vector points, in the feature space. The **rminer** package adopts the popular Gaussian kernel [13], which presents less parameters than other kernels (e.g., polynomial): $K(\mathrm{x}, \mathrm{x}') = \exp(-\gamma||\mathrm{x} - \mathrm{x}'||^2)$, $\gamma > 0$. The probabilistic SVM output is given by Ref. [35]: $f(\mathrm{x}_i) = \sum_{j=1}^{m} y_j \alpha_j K(\mathrm{x}_j, \mathrm{x}_i) + b$ and $p(i) = 1/(1 + \exp(Af(\mathrm{x}_i) + B))$, where m is the number of support vectors, $y_i \in \{-1, 1\}$ is the output for a binary classification, b and α_j are the coefficients of the model, and A and B are determined by solving a regularized maximum likelihood problem.

Before fitting the NN and SVM models, the input data is first standardized to a zero mean and one standard deviation [13]. For DT, **rminer** adopts the default parameters of the **rpart R** package, which implements the popular CART algorithm [2] For the LR and NN learning, **rminer** uses the efficient BFGS algorithm [22], from the family of quasi-Newton methods, while SVM is trained using the sequential minimal optimization (SMO) [26]. The learning capabilities of NN and SVM are affected by the choice of their hyperparameters (H for NN; γ and C, a complex penalty parameter, for SVM). For setting these values, **rminer** uses grid search and heuristics [5].

Complex DM models, such as NN and SVM, often achieve accurate predictive performances. Yet, the increased complexity of NN and SVM makes the final data-driven model difficult to be understood by humans. To open these black-box models, there are two interesting possibilities, rule extraction and sensitivity analysis. Rule extraction often involves the use of a white-box method (e.g., decision tree) to learn the black-box responses [29]. The sensitivity analysis procedure works by analyzing the responses of a model when a given input is varied through its domain [7]. By analyzing the sensitivity responses, it is possible to measure input relevance and average impact of a particular input in the model. The former can be shown visually using an input importance bar plot and the latter by plotting the Variable Effect Characteristic (VEC) curve. Opening the black-box allows in explaining how the model makes the decisions and improves the acceptance of prediction models by the domain experts, as shown in Ref. [20].

2.3. Evaluation

A class can be assigned from a probabilistic outcome by assigning a threshold D, such that event c is true if $p(c|\mathrm{x}_k) > D$. The receiver operating characteristic (ROC) curve shows the performance of a two class classifier across the range of possible threshold (D) values, plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) [11]. The overall accuracy is given by the area under the curve ($AUC = \int_0^1 ROCdD$), measuring the degree of discrimination that can be obtained from a given model. AUC is a popular classification metric [21] that presents advantages of being independent of the class frequency or specific false positive/negative costs. The ideal method should present an AUC of 1.0, while an AUC of 0.5 denotes a random classifier.

In the domain of marketing, the Lift analysis is popular for accessing the quality of targeting models [3]. Usually, the population is divided into deciles, under a decreasing order of their predictive probability for success. A useful Lift cumulative curve is obtained by plotting the population samples (ordered by the deciles, x-axis) versus the cumulative percentage of real responses captured (y-axis). Similarly to the AUC metric, the ideal method should present an area under the LIFT (ALIFT) cumulative curve close to 1.0. A high ALIFT confirms that the predictive model concentrates responders in the top deciles, while an ALIFT of 0.5 corresponds to the performance of a random baseline.

Given that the training data includes a large number of contacts (51,651), we adopt the popular and fast holdout method (with R distinct runs) for feature and model selection purposes. Under this holdout scheme, the training data is further divided into training and validation sets by using a random split with 2/3 and 1/3 of the contacts, respectively. The results are aggregated by the average of the R runs and a Mann–Whitney non-parametric test is used to check statistical significance at the 95% confidence level.

In real environment, the DSS should be regularly updated as new contact data becomes available. Moreover, client propensity to subscribe a bank product may evolve through time (e.g., changes in the economic environment). Hence, for achieving a robust predictive evaluation we adopt the more realistic fixed-size (of length W) rolling window evaluation scheme that performs several model updates and discards oldest data [18]. Under this scheme, a training window of W consecutive contacts is used to fit the model and then we perform predictions related with the next K contacts. Next, we update (i.e., slide) the training window by replacing the oldest K contacts with K newest contacts (related with the previously predicted contacts but now we assume that the outcome result is known), in order to perform new K predictions, and so on. For a test set of length L, a total number of model updates (i.e., trainings) are $U = L/K$. Fig. 1 exemplifies the rolling window evaluation procedure.

2.4. Feature selection

The large number (150) of potential useful features demanded a stricter choice of relevant attributes. Feature selection is often a key DM step, since it is useful to discard irrelevant inputs, leading to simpler data-driven models that are easier to interpret and that tend to provide better predictive performances [12]. In Ref. [34], it is argued that while automatic methods can be useful, the best way is to perform a manual feature selection by using problem domain knowledge, i.e., by having a clear understanding of what the attributes actually mean. In this work, we use a semi-automatic approach for feature selection based on two steps that are described below.

In the first step, business intuitive knowledge was used to define a set of fourteen questions, which represent certain hypotheses that are tested. Each question (or factor of analysis) is defined in terms of a group of related attributes selected from the original set of 150 features by a bank campaign manager (domain expert). For instance, the question about the gender influence (male/female) includes the three features, related with the gender of the banking agent, client and client–agent difference (0 – if same sex; 1 – else). Table 1 exhibits the analyzed factors and the number of attributes related with each factor, covering a total of 69 features (reduction of 46%).

In the second step, an automated selection approach is adopted, based an adapted forward selection method [12]. Given that standard forward selection is dependent on the sequence of features used and that the features related with a factor of analysis are highly related, we first apply a simple wrapper selection method that works with a

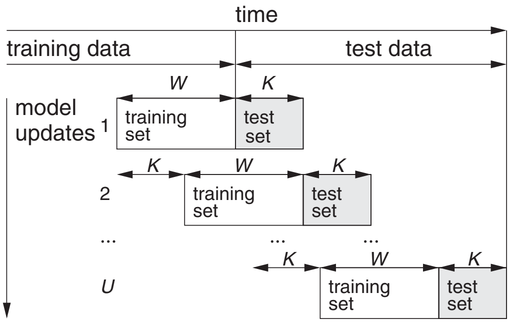


Fig. 1. Schematic of the adopted rolling window evaluation procedure.

Table 1
Analyzed business questions for a successful contact result.

Question (factor of analysis)	Number of features
1: Is offered rate relevant?	**5**
2: Is gender relevant?	**3**
3: Is agent experience relevant?	**3**
4: Are social status and stability relevant?	5
5: Is client–bank relationship relevant?	**11**
6: Are bank blocks (triggered to prevent certain operations) relevant?	6
7: Is phone call context relevant?	**4**
8: Are date and time conditions relevant?	**3**
9: Are bank profiling indicators relevant?	**7**
10: Are social and economic indicators relevant?	**11**
11: Are financial assets relevant?	3
12: Is residence district relevant?	1
13: Can age be related to products with longer term periods?	3
14: Are web page hits (for campaigns displayed in bank web sites) relevant?	4
Number of features after business knowledge selection	69
Number of features after first feature selection phase	**22**

DM fed with combinations of inputs taken from a single factor. The goal is to identify the most interesting factors and features attached to such factors. Using only training set data, several DM models are fit, by using: each individual feature related to a particular question (i.e., one input) to predict the contact result; and all features related with the same question (e.g., 3 inputs for question #2 about gender influence). Let AUC_q and $AUC_{q,i}$ denote the AUC values, as measured on the validation set, for the model fed with all inputs related with question q and only the i-th individual feature of question q. We assume that the business hypothesis is confirmed if at least one of the individually tested attributes achieves an $AUC_{q,i}$ greater than a threshold T_1 and if the model with all question related features returns an AUC_q greater than another threshold T_2. When a hypothesis is confirmed, only the m-th feature is selected if $AUC_{q,m} > AUC_q$ or $AUC_q - AUC_{q,m} < T_3$, where $AUC_{q,m} = max(AUC_{q,i})$. Else, we rank the input relevance of the model with all question related features in order to select the most relevant ones, such that the sum of input importances is higher than a threshold T_4.

Once a set of confirmed hypotheses and relevant features is achieved, a forward selection method is applied, working on a factor by factor step basis. A DM model that is fed with training set data using as inputs all relevant features of the first confirmed factor and then AUC is computed over the validation set. Then, another DM model is trained with all previous inputs plus the relevant features of the next confirmed factor. If there is an increase in the AUC, then the current factor features are included in the next step DM model, else they are discarded. This procedure ends when all confirmed factors have been tested if they improve the predictive performance in terms of the AUC value.

3. Experiments and results

3.1. Modeling

All experiments were performed using the **rminer** package and **R** tool [5] and conducted in a Linux server, with an Intel Xeon 5500 2.27 GHz processor. Each DM model related with this section was executed using a total of $R = 20$ runs. For the feature selection, we adopted the NN model described in Section 2.2 as the base DM model, since preliminary experiments, using only training data, confirmed that NN provided the best AUC and ALIFT results when compared with other DM methods. Also, these preliminary experiments confirmed that SVM required much more computation when compared with NN, in an expected result since SMO algorithm memory and processing requirements grow much more heavily with the size of the dataset when compared with BFGS algorithm used by the NN. At this stage, we set the number of hidden nodes using the heuristic $H = \text{round}(M/2)$ (M is the number of inputs), which is also adopted by the WEKA tool [34] and tends to provide good classification results [5]. The NN ensemble is composed of $N_r = 7$ distinct networks, each trained with 100 epochs of the BFGS algorithm.

Before executing the feature selection, we fixed the initial phase thresholds to reasonable values: $T_1 = 0.60$ and $T_2 = 0.65$, two AUC values better than the random baseline of 0.5 and such that $T_2 > T_1$; $T_3 = 0.01$, the minimum difference of AUC values; and $T_4 = 60\%$, such that the sum of input importances accounts for at least 60% of the influence. Table 1 presents the eight confirmed hypothesis (question numbers in **bold**) and associated result of 22 relevant features, after applying the first feature selection phase. This procedure discarded 6 factors and 47 features, leading to a 32% reduction rate when compared with 69 features set by the business knowledge selection. Then, the forward selection phase was executed. Table 2 presents the respective AUC results (column *AUC*, average of $R = 20$ runs) and full list of selected features. The second phase confirmed the relevance of all factors, given that each time a new factor was added, the DM model produced a higher AUC value. An additional experiment was conducted with the LR model, executing the same feature selection method and confirmed the same eight factors of analysis and leading to a similar reduced set (with 24 features). Yet, the NN model with 22 inputs got better AUC and ALIFT values when compared with LR, and thus such 22 inputs are adopted in the remainder of this paper.

After selecting the final set of input features, we compared the performance of the four DM models: LR, DT, NN and SVM. The comparison of SVM with NN was set under similar conditions, where the best hyperparameters (H and γ) were set by performing a grid search under the ranges $H\epsilon\{0,2,6,8,10,12\}$ and $\gamma \in 2^k : k \in \{-15, -11.4, -7.8, -4.2, -0.6, 3\}$. The second SVM parameter (which is less relevant) was fixed using the heuristic $C = 3$ proposed in for $\mathbf{x}$ standardized input data [5]. The **rminer** package applies this grid search by performing an internal holdout scheme over the training set, in order to select the best hyperparameter (H or γ) that corresponds to the lowest AUC value measured on a subset of the training set, and then trains the best model with all training set data.

The obtained results for the modeling phase (using only training and validation set data) are shown in Table 3 in terms of the average (over $R = 20$ runs) of the AUC and ALIFT metrics (Section 2.3) computed on the validation set. The best result was achieved by the NN model, which outperformed LR (improvement of 3 pp), DT (improvement of 10 pp) and SVM (improvement of 4 and 3 pp) in both metrics and with statistical confidence (i.e., Mann–Whitney p-value < 0.05). In the table, the selected NN and SVM hyperparameters are presented in brackets (median value shown for H and γ). It should be noted that the hidden node grid search strategy for NN did not improve the AUC value (0.929) when compared with the $H = \text{round}(M/2) = 11$ heuristic (used in Table 2). Nevertheless, given that a simpler model was selected (i.e., $H = 6$), we opt for such model in the remainder of this paper.

To attest the utility of the proposed feature selection approach, we compared it with two alternatives: no selection, which makes use of the all 150 features; and forward selection, which adopts the standard forward method. The latter alternative uses all 150 features as feature candidates. In the first iteration, it selects the feature that produces the highest AUC value, measured using the validation set (1/3 of the training data) when considering the average of 20 runs. Then, the selected feature is fixed and a second iteration is executed to select the second feature within the remaining 149 candidates and the obtained AUC is compared with the one obtained in previous iteration. This method proceeds with more iterations until there is no AUC improvement or if all features are selected. Table 4 compares the three feature selection methods in terms of number of features used by the model, time elapsed and performance metric (AUC).

The obtained results confirm the usefulness of the proposed approach, which obtains the best AUC value. The proposed method uses

Table 2
Final set of selected attributes.

Factor	Attributes	Description	AUC
1: Interest rate	nat.avg.rate	National monthly average of deposit interest rate	0.781
	suited.rate	Most suited rate to the client according to bank criteria	
	dif.best.rate.avg	Difference between best rate offered and the national average	
2: Gender	ag.sex	Sex of the agent (male/female) that made (outbound) or answered (inbound) the call	0.793
3: agent experience	ag.generic	If generic agent, i.e. temporary hired, with less experience (yes/no)	0.799
	ag.created	Number of days since the agent was created	
5: Client–bank relationship	cli.house.loan	If the client has a house loan contract (yes/no)	0.805
	cli.affluent	If the client is an affluent client (yes/no)	
	cli.indiv.credit	If the client has an individual credit contract (yes/no)	
	cli.salary.account	If the client has a salary account (yes/no)	
7: Phone call context	call.dir	Call direction (inbound/outbound)	0.809
	call.nr.schedules	Number of previously scheduled calls during the same campaign	
	call.prev.durations	Duration of previously scheduled calls (in *s*)	
8: Date and time	call.month	Month in which the call is made	0.810
9: Bank profiling indicators	cli.sec.group	Security group bank classification	0.927
	cli.agreggate	If the client has aggregated products and services	
	cli.profile	Generic client profile, considering assets and risk	
10: Social and economic indicators	emp.var.rate	Employment variation rate, with a quarterly frequency	0.929
	cons.price.idx	Monthly average consumer price index	
	cons.conf.idx	Monthly average consumer confidence index	
	euribor3m	Daily three month Euribor rate	
	nr.employed	Quarterly average of the total number of employed citizens	

lesser features (around a factor of 7) when compared with the full feature approach. Also, it is also much faster (around a factor of 5) when compared with the simple forward selection.

3.2. Predictive knowledge and potential impact

The best model from previous section (NN fed with 22 features from Table 2, with $H = 6$ and $N_r = 7$) was tested for its predictive capabilities under a more realistic and robust evaluation scheme. Such scheme is based on a rolling window evaluation (Section 2.3) over the test data, with $L = 1293$ contacts from the most recent year. Taking into account the computational effort required, the rolling window parameters were fixed to the reasonable values of $W = 20000$ (window size) and $K = 10$ (predictions made each model update), which corresponds to $U = 130$ model updates (trainings and evaluations). We note that a sensitivity analysis was executed over W, where other W configurations were tested (e.g., 19,000 and 21,000) leading to very similar results. For comparison purposes, we also tested LR, DT and SVM (as set in Section 3.1).

The results of all $U = 130$ updates are summarized in Table 5. While a trained model only predicts $K = 10$ contact outcomes (in each update), the AUC and ALIFT metrics were computing using the full set of predictions and desired values. Similarly to the modeling phase, the best results are given by the NN model and for both metrics, with improvements of: 2.7 pp for SVM, 3.7 pp for DT and 7.9 pp for LR, in terms of AUC; and 1.6 pp for SVM, 2.1 pp for DT and 4.6 pp for LR, in terms of ALIFT. Interestingly, while DT was the worse performing technique in the modeling phase, prediction tests revealed it as the third best model, outperforming LR and justifying the need for technique comparison in every stage of the decision making process [9].

The left of Fig. 2 plots the ROC curves for the four models tested. A good model should offer the best compromise between a desirable high true positive rate (TPR) and low false positive rate (FPR). The former goal corresponds to a sensitive model, while the latter is related with a more specific model. The advantage of the ROC curve is that the domain user can select the best TPR and FPR trade-off that serves its needs. The NN ROC curve is related with the highest area (AUC) and outperforms all other methods within most (75%) of the FPR range (e.g., NN is the best method for FPR within [0.00,0.10], [0.26,0.85] and [0.94,1.00]).

Focusing on the case studied of bank telemarketing, it is difficult to financially quantify costs, since long term deposits have different amounts, interest rates and subscription periods. Moreover, human agents are hired to accept inbound phone calls, as well as sell other non-deposit products. In addition, it is difficult to estimate intrusiveness of an outbound call (e.g., due to a stressful conversation). Nevertheless, we highlight that current bank context favors more sensitive models: communication costs are contracted in bundle packages, keeping costs low; and more importantly, the 2008 financial crisis strongly increased the pressure for Portuguese banks to increase long term deposits. Hence, for this particular bank it is better to produce more successful sells even if this involves loosing some effort in contacting non-buyers. Under such context, NN is the advised modeling technique, producing the best TPR and FPR trade-off within most of the sensitive range. For the range FPR within [0.26,0.85], the NN gets a high TPR value (ranging from 0.75 to 0.97). The NN TPR mean difference under the FPR range [0.45,0.62] is 2 pp when compared with SVM and 9 pp when compared with DT. For demonstrative purposes, the right of Fig. 2 shows the confusion matrix related with the NN model and for $D = 0.5$.

The left of Fig. 3 plots the Lift cumulative curves for the predictions using the four models, while the right of Fig. 3 shows examples of cumulative lift response values for the best three models (NN, SVM and DT) and several sample size configurations (e.g., 10% and 50%). Under the cumulative lift analysis, the NN model is the best model within a large portion (77%) of the sample size range. In effect, NN outperforms the SVM model for the sample size ranges of [0.06;0.24] and [0.27;0.99], presenting an average difference of 2 pp within the range [0.27:0.9]. Also, NN is better than DT for the sample size ranges of [0,0.22],

Table 3
Comparison of DM models for the modeling phase (bold denotes the best value).

Metric	LR	DT	SVM $\left(\widetilde{\gamma} = 2^{-7.8}, C = 3\right)$	NN $\left(\widetilde{H} = 6, N_r = 7\right)$
AUC	0.900	0.833	0.891	**0.929**[a]
ALIFT	0.849	0.756	0.844	**0.878**[a]

[a] Statistically significant under a pairwise comparison with SVM, LR and DT.

Table 4
Comparison of feature selection methods for the modeling phase using NN model (bold denotes the best AUC).

Method	#Features	Time elapsed (in *s*)	AUC metric
no selection	150	3223	0.832
forward selection	7	97,975	0.896
proposed	22	18,651[a]	**0.929**

[a] Includes interview with domain expert (5400 *s*) for Table 1 definition.

Table 5
Comparison of models for the rolling window phase (bold denotes the best value).

Metric	LR	DT	SVM	NN
AUC	0.715	0.757	0.767	**0.794**
ALIFT	0.626	0.651	0.656	**0.672**

[0.33,0.40], [0.46,0.9] and [0.96,1]. The largest NN difference when compared with DT is achieved for the sample size range of [0.46,0.9], reaching up to 8 pp. Since for this particular bank and context the pressure is set towards getting more successful sells (as previously explained), this is an important sample size range. Currently, the bank uses a standard process that does not filter clients, thus involving a calling to all clients in the database. Nevertheless, in the future there can be changes in the bank client selection policy. For instance, one might imagine the scenario where telemarketing manager is asked to reduce the number of contacts by half (maximum of the bank's current intentions). As shown in Fig. 3, without the data-driven model conceived, telemarketing would reach expectedly just 50% of the possible subscribers, while with the NN model proposed here would allow to reach around 79% of the responses, thus benefiting from an increase of 29 percentage points of successful contacts. This result attests the utility of such model, which allows campaign managers to increase efficiency through cost reduction (less calls made) and still reaching a large portion of successful contacts.

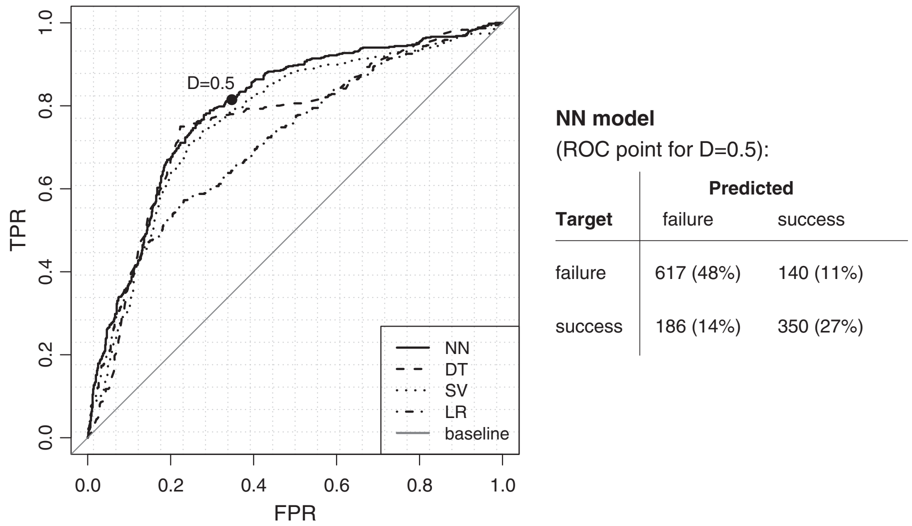
When comparing the best proposed model NN in terms of modeling versus rolling window phases, there is a decrease in performance, with a reduction in AUC from 0.929 to 0.794 and ALIFT from 0.878 to 0.672. However, such reduction was expected since in the modeling phase the feature selection was tuned based on validation set errors, while the best model was then fixed (i.e., 22 inputs and $H = 6$) and tested on completely new unseen and more recent data. Moreover, the obtained AUC and ALIFT values are much better than the random baseline of 50%.

3.3. Explanatory knowledge

In this section, we show how explanatory knowledge can be extracted by using a sensitivity analysis and rule extraction techniques (Section 2.2) to open the data-driven model. Using the *Importance* function of the **rminer** package, we applied the Data-based Sensitivity Analysis (DSA) algorithm, which is capable of measuring the global influence of an input, including its iterations with other attributes [7]. The DSA algorithm was executed on the selected NN model, fitted with all training data (51,651 oldest contacts). Fig. 4 exhibits the respective input importance bar plot (the attribute names are described in more detail in Table 2). A DT was also applied to the output responses of the NN model that was fitted with all training data. We set the DT complexity parameter to 0.001, which allowed us to fit a DT as a low error, obtaining a mean absolute error of 0.03 when predicting the NN responses. A large tree was obtained and to simplify the analysis, Fig. 5 presents the obtained decision rules up to six decision levels. An example of an extracted rule is: if the number of employed is equal or higher than 5088 thousand and duration of previously scheduled calls is less than 13 min and the call is not made in March, April, October or December, and the call is inbound then the probability of success is 0.62. In Fig. 5, decision rules that are aligned with the sensitivity analysis are shown in **bold** and are discussed in the next paragraphs.

An interesting result shown by Fig. 4 is that the three month Euribor rate (euribor3m), computed by the European Central Bank (ECB) and published by Thomson Reuters, i.e., a publicly available and widely used index, was considered the most relevant attribute, with a relative importance around 17%. Next comes the direction of the phone call (inbound versus outbound, 15%), followed by the number of days since the agent login was created, which is an indicator of agent experience in the bank contact center, although not necessarily on the deposit campaign, since each agent can deal with different types of service (e.g., phone banking). The difference between the best possible rate for the product being offered and the national average rate is the fourth most relevant input attribute. This feature was expected to be one of the top attributes, since it stands for what the client will receive for subscribing the bank deposit when compared to the competition. Along with the Euribor rate, these two attributes are the ones from the top five which are not specifically related to call context, so they will be analyzed together further ahead. Last in the top five attributes comes the duration of previous calls that needed to be rescheduled to obtain a final answer by the client. It is also interesting to notice that the top ten attributes found by the sensitivity analysis (Fig. 4) are also used by the extracted decision tree, as shown in Fig. 5.

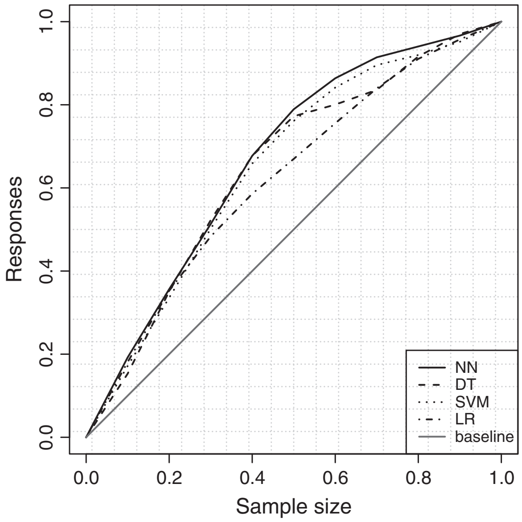
Concerning the sensitivity analysis input ranking, one may also take into consideration the relevance of the sixth and eighth most relevant attributes, both related to social quarterly indicators of employment, the number of employees and the employment variation rate, which reveal that these social indicators play a role in success contact modeling. While client attributes are specific of an individual, they were considered less relevant, with six of them in the bottom of the input bar plot (Fig. 4). This does not necessarily mean that these types of attributes have general few impact on modeling contact success. In this particular



NN model
(ROC point for D=0.5):

	Predicted	
Target	failure	success
failure	617 (48%)	140 (11%)
success	186 (14%)	350 (27%)

Fig. 2. ROC curves for the four models (left) and example confusion matrix for NN and $D = 0.5$ (right).



Examples of DT, SVM and NN response values:

Sample size	DT	SVM	NN
5%	8.0%	9.0%	10.4%
10%	15.3%	17.2%	19.2%
20%	35.1%	33.6%	35.6%
30%	52.2%	50.2%	51.3%
40%	67.7%	65.9%	67.7%
50%	77.2%	76.1%	78.9%
60%	80.0%	84.1%	86.4%
70%	83.6%	89.6%	91.4%

Fig. 3. Lift cumulative curves for the four models (left) and examples of NN and DT cumulative lift response values (right).

case, the profiling indicators used were defined by the bank and the obtained results suggest that probably these indicators are not adequate for our problem of targeting deposits.

The sensitivity analysis results can also be visualized using a VEC curve, which allows understanding the global influence of an attribute in the predicted outcome by plotting the attribute range of values versus the average sensitivity responses [7]. We analyzed the top five most relevant attributes, with the corresponding VEC curves being plotted in the left (Euribor and product offered interest rates) and right (remaining top 5 relevant attributes) of Fig. 6.

When considering the Euribor rate, one might think that a lower Euribor would result in a decline in savings rate since most European banks align their deposit interest rate offers with ECB indexes, particularly with the three month Euribor [27]. Still, the right of Fig. 6 reveals the opposite, with a lower Euribor corresponding to a higher probability for deposit subscription, and the same probability decreasing along with the increase of the three month Euribor. A similar effect is visible in a decision node of the extracted DT (Fig. 5), where the probability of success decreases by 10 pp when the Euribor rate is higher than 0.73. This behavior is explained by a more recent research [30], which revealed that while prior to 2008 a weak positive relation could be observed between offered rate for deposits and savings rate, after 2008, with the financial crisis, that relation reversed, turning clients more prone to savings while the Euribor constantly decreased. This apparent contradiction might be due to clients' perception of a real economic recession and social depression. Consumers might feel an increased need to consider saving for the future as opposed to immediate gratification coming from spending money in purchasing desired products or services. This observation emphasizes the inclusion of this kind of information on similar DM projects. Concerning the difference between best product rate offered and national average, Fig. 6 confirms our expectation that an increase in this attribute does increase the probability for subscribing a deposit. Still, once the difference reaches 0.73%, the influence on the probability of subscription is highly reduced, which means that an interest rate slightly above the competition seems to be enough to make the difference on the result. It is also interesting to note that the extracted DT reveals a positive effect of the rate difference with a successful contact (Fig. 5).

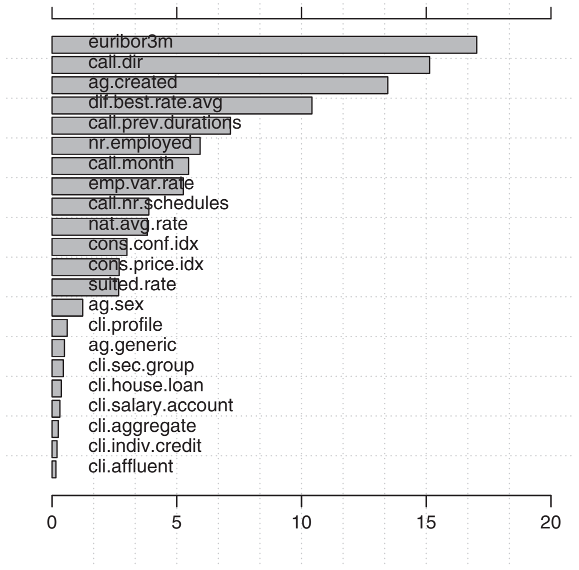


Fig. 4. Relative importance of each input attribute for the NN model (in %).

The right of Fig. 6 shows the influence of the second, third and fifth most relevant attributes. Regarding call direction, we validate that clients contacted through inbound are keener to subscribe the deposit. A similar effect is measured by the extracted DT, where an inbound call increases the probability of success by 25 pp (Fig. 5). Inbound is associated with less intrusiveness given that the client has called the bank and thus he/she is more receptive for a sell. Another expected outcome is related with agent experience, where the knowledge extraction results show that it has a significant impact on a successful contact. Quite interestingly, a few days of experience are enough to produce a strong impact, given that under the VEC analysis with just six days the average probability of success is above 50% (Fig. 6) and the extracted DT increases the probability of successful sell by 9 pp when the experience is higher or equal than 3.3 days (Fig. 5). Regarding the duration of previously scheduled calls, it happens often that the client does not decide on the first call on whether to subscribe or not the deposit, asking to be called again, thus rescheduling another call. In those cases (63.8% for the whole dataset), a contact develops through more than one phone call. The sensitivity analysis (Fig. 6) shows that more time already spent on past calls within the same campaign increases probability of success. Similarly, the extracted DT confirms a positive effect of the duration of previous calls. For instance, when the duration is higher or equal than 13 min (left node at the second level of Fig. 5), then the associated global probability of success is 0.3, while the value decreases to 0.05 (25 pp difference) if this duration condition is false.

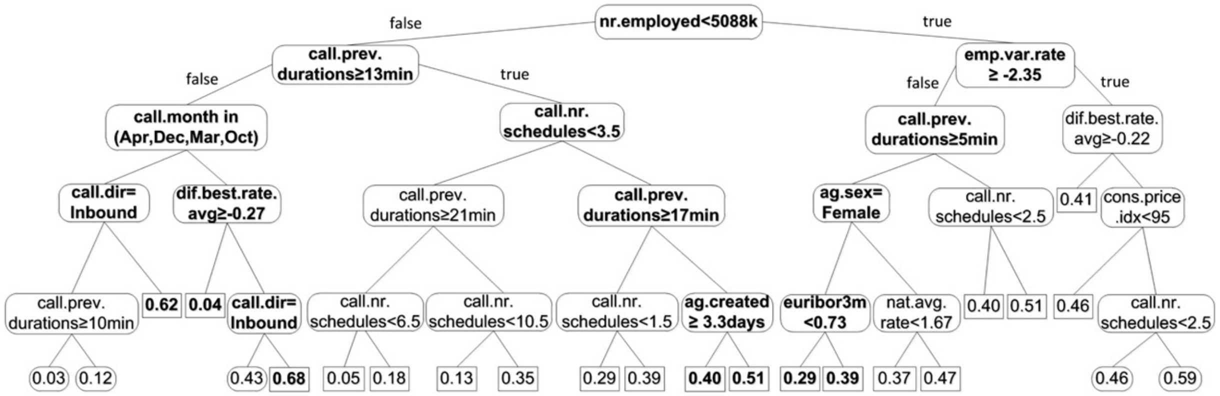


Fig. 5. Decision tree extracted from the NN model.

It is interesting to note that some explanatory variables are uncontrolled by the commercial bank (e.g., three month Euribor rate) while others are partially controlled, i.e., can be influenced by bank managers' decisions (e.g., difference between best offered and national average rates, which also depends on competitors' decisions), and other variables can be fully controlled (e.g., direction of call, if outbound; agent experience — ag.created; duration of previously scheduled calls). Given these characteristics, telemarketing managers can act directly over some variables, while analyzing expectations influenced by uncontrolled variables. For instance, managers can increase campaign investment (e.g., by assigning more agents) when the expected return is high, while postponing or reducing marketing campaigns when a lower success is globally predicted.

4. Conclusions

Within the banking industry, optimizing targeting for telemarketing is a key issue, under a growing pressure to increase profits and reduce costs. The recent 2008 financial crisis dramatically changed the business of European banks. In particular, Portuguese banks were pressured to increase capital requirements (e.g., by capturing more long term deposits). Under this context, the use of a decision support system (DSS) based on a data-driven model to predict the result of a telemarketing phone call to sell long term deposits is a valuable tool to support client selection decisions of bank campaign managers.

In this study, we propose a personal and intelligent DSS that uses a data mining (DM) approach for the selection of bank telemarketing clients. We analyzed a recent and large Portuguese bank dataset, collected from 2008 to 2013, with a total of 52,944 records. The goal was to model the success of subscribing a long-term deposit using attributes that were known before the telemarketing call was executed. A particular emphasis was given on feature engineering, as we considered an initial set of 150 input attributes, including the commonly used bank client and product features and also newly proposed social and economic indicators. During the modeling phase, and using a semi-automated feature selection procedure, we selected a reduced set of 22 relevant features. Also, four DM models were compared: logistic regression (LR), decision trees (DTs), neural networks (NNs) and support vector machines (SVMs). These models were compared using two metrics, area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT), both at the modeling and rolling window evaluation phases. For both metrics and phases, the best results were obtained by the NN, which resulted in an AUC of 0.80 and ALIFT of 0.67 during the rolling window evaluation. Such AUC corresponds to a very good discrimination. Moreover, the proposed model has impact in the banking domain. For instance, the cumulative LIFT analysis reveals that 79% of the successful sells can be achieved when contacting only half of the clients, which translates in an improvement of 29 percentage points when compared with the current bank practice, which simply contacts all clients. By selecting only the most likely buyers, the proposed DSS creates value for the bank telemarketing managers in

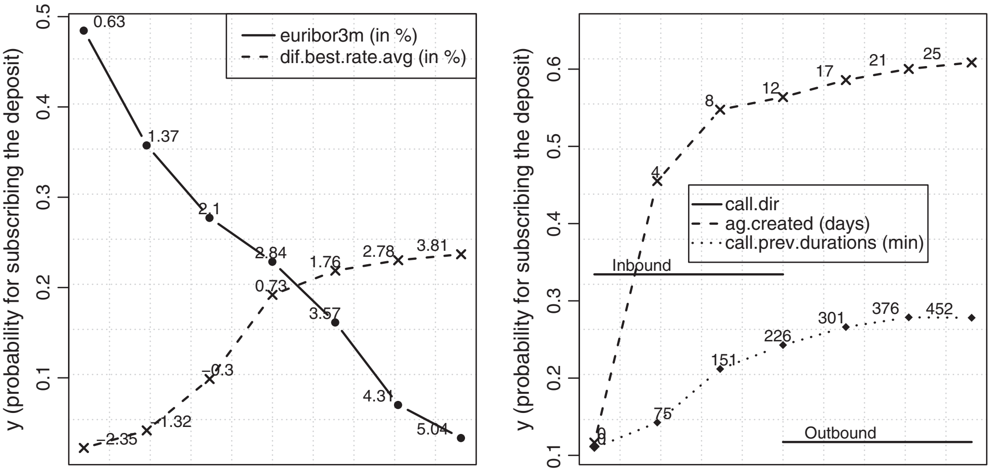


Fig. 6. VEC curves showing the influence of the first and fourth (left) and second, third and fifth (right) most relevant attributes.

terms of campaign efficiency improvement (e.g., reducing client intrusiveness and contact costs).

Two knowledge extraction techniques were also applied to the proposed model: a sensitivity analysis, which ranked the input attributes and showed the average effect of the most relevant features in the NN responses; and a decision tree, which learned the NN responses with a low error and allowed the extraction of decision rules that are easy to interpret. As an interesting outcome, the three month Euribor rate was considered the most relevant attribute by the sensitivity analysis, followed by the direction call (outbound or inbound), the bank agent experience, difference between the best possible rate for the product being offered and the national average rate, and the duration of previous calls that needed to be rescheduled to obtain a final answer by the client. Several of the extracted decision rules were aligned with the sensitivity analysis results and make use of the top ten attributes ranked by the sensitivity analysis. The obtained results are credible for the banking domain and provide valuable knowledge for the telemarketing campaign manager. For instance, we confirm the result of Ref. [30], which claims that the financial crisis changed the way the Euribor affects savings rate, turning clients more likely to perform savings while Euribor decreased. Moreover, inbound calls and an increase in other highly relevant attributes (i.e., difference in best possible rate, agent experience or duration of previous calls) enhance the probability for a successful deposit sell.

In future work, we intend to address the prediction of other telemarketing relevant variables, such as the duration of the call (which highly affects the probability of a successful contact [23]) or the amount that is deposited in the bank. Additionally, the dataset may provide history telemarketing behavior for cases when clients have previously been contacted. Such information could be used to enrich the dataset (e.g., computing recency, frequency and monetary features) and possibly provide new valuable knowledge to improve model accuracy. Also it would be interesting to consider the possibility of splitting the sample according to two sub-periods of time within the range of 2008–2012, which would allow us to analyze impact of hard-hit recession versus slow recovery.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions.

References

[1] David Arnott, Graham Pervan, Eight key issues for the decision support systems discipline, Decision Support Systems 44 (3) (2008) 657–672.
[2] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA, 1984.
[3] David S. Coppock, Why lift? Data modeling and mining, Information Management (2002) 5329-1 2002, (Online; accessed 19-July-2013).
[4] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (3) (1995) 273–297.
[5] Paulo Cortez, Data mining with neural networks and support vector machines using the r/rminer tool, Advances in Data Mining. Applications and Theoretical Aspects, 6171, Springer, 2010, pp. 572–583.
[6] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis, Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems 47 (4) (2009) 547–553.
[7] Paulo Cortez, Mark J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, Information Sciences 225 (2013) 1–17.
[8] Dursun Delen, A comparative analysis of machine learning techniques for student retention management, Decision Support Systems 49 (4) (2010) 498–506.
[9] Dursun Delen, Ramesh Sharda, Prajeeb Kumar, Movie forecast guru: a Web-based DSS for Hollywood managers, Decision Support Systems 43 (4) (2007) 1151–1170.
[10] Pedro Domingos, A few useful things to know about machine learning, Communications of the ACM 55 (10) (2012) 78–87.
[11] Tom Fawcett, An introduction to roc analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.
[12] Isabelle Guyon, André Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition Springer-Verlag, NY, USA, 2008.
[14] S.S. Haykin, Neural networks and Learning Machines, Prentice Hall, 2009.
[15] Sadaf Hossein Javaheri, Mohammad Mehdi Sepehri, Babak Teimourpour, Response modeling in direct marketing: a data mining based approach for target selection, in: Yanchang Zhao, Yonghua Cen (Eds.), Data Mining Applications with R, Elsevier, 2014, pp. 153–178, (chapter 6).
[16] Philip Kotler, Kevin Lane Keller, Framework for Marketing Management, 5th edition Pearson, 2012.
[17] Kin-Nam Lau, Haily Chow, Connie Liu, A database approach to cross selling in the banking industry: practices, strategies and challenges, Journal of Database Marketing and Customer Strategy Management 11 (3) (2004) 216–234.
[18] William Leigh, Russell Purvis, James M. Ragusa, Forecasting the nyse composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support, Decision Support Systems 32 (4) (2002) 361–377.
[19] David Martens, Foster Provost, Pseudo-social network targeting from consumer transaction data, NYU Working Papers Series, , CeDER-11-05, 2011.
[20] David Martens, Foster Provost, Explaining data-driven document classifications, MIS Quarterly 38 (1) (2014) 73–99.
[21] David Martens, Jan Vanthienen, Wouter Verbeke, Bart Baesens, Performance of classification models from a user perspective, Decision Support Systems 51 (4) (2011) 782–793.
[22] M. Moller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks 6 (4) (1993) 525–533.
[23] Sérgio Moro, Raul Laureano, Paulo Cortez, Enhancing bank direct marketing through data mining, Proceedings of the Forty-First International Conference of the European Marketing Academy, European Marketing Academy, 2012, pp. 1–8.
[24] David L. Olson, Dursun Delen, Yanyan Meng, Comparative analysis of data mining methods for bankruptcy prediction, Decision Support Systems 52 (2) (2012) 464–473.
[25] Robert Phillips, Optimizing prices for consumer credit, Journal and Review Pricing Management 12 (2013) 360–377.
[26] John Platt, Sequential minimal optimization: a fast algorithm for training support vector machines, Technical Report MSR-TR-98-14, , Microsoft Research, 1998.
[27] Gerard O. Reilly, Information in financial market indicators: an overview, Quarterly Bulletin Articles 4 (2005) 133–141 (Central Bank of Ireland).
[28] Roland T. Rust, Christine Moorman, Gaurav Bhalla, Rethinking marketing, Harvard Business Review 1 (2010) 1–8.
[29] R. Setiono, Techniques for extracting classification and regression rules from artificial neural networks, in: D. Fogel, C. Robinson (Eds.), Computational Intelligence: The Experts Speak, IEEE Press, 2003, pp. 99–114.
[30] P. Stinglhamber, Ch. Van Nieuwenhuyze, M.D. Zachary, The impact of low interest rates on household financial behaviour, Econometric Reviews 2 (2011) 77–91 (National Bank of Belgium).
[31] Fabrice Talla Nobibon, Roel Leus, Frits CR Spieksma, Optimization models for targeted offers in direct marketing: exact and heuristic algorithms, European Journal of Operational Research 210 (3) (2011) 670–683.
[32] Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, 9th edition Pearson, 2011.
[33] W. Venables, B. Ripley, Modern Applied Statistics with S, 4th edition Springer, 2003.
[34] Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition Morgan Kaufmann, 2005.
[35] T.F. Wu, C.J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, Journal of Machine Learning Research 5 (2004) 975–1005.

Sérgio Moro holds a 5 year degree diploma in Computer Engineering from Instituto Superior Técnico (Technical University of Lisbon), and a MSc in Management Information Systems from ISCTE — University Institute of Lisbon, where he is also a PhD researcher in the fields of business intelligence and decision support systems. He currently works in the Business Intelligence & Analytics team of a Portuguese bank, where he is responsible for projects such as GRC (Governance, Risk and Compliance), and ALM (Asset–Liability Management). Previously he was responsible for software development for the bank's contact center.

Paulo Cortez is an Associate Professor at the Department of Information Systems at University of Minho and Coordinator of the Information Systems and Technologies R&D group of ALGORITMI Research Centre. He completed his PhD (2002) in Computer Science and Habilitation (2013) in Information Systems and Technologies at the same university. His research interests include: business intelligence, data mining, neural networks, evolutionary computation and forecasting. Currently, he is associate editor of the journals Expert Systems and Neural Processing Letters. He has published more than 92 indexed (ISI or Scopus) papers. His research has appeared in Journal of Heuristics, Decision Support Systems, Information Sciences and others (see http://www3.dsi.uminho.pt/pcortez).

Paulo Rita holds a PhD in Marketing from Cardiff Business School, UK and has a Post-Doctorate in E-Marketing from the University of Nevada Las Vegas, United States. Dr Rita is Professor of Marketing at IBS – ISCTE Business School, Lisbon – Portugal where he is Director of the PhD in Marketing, Director of the PhD in Tourism Management in partnership with the European University and Director of the Master in Hospitality and Tourism Management which is a double degree with University of Central Florida, Orlando, United States. Paulo Rita is currently a member of the Executive Committee of EMAC (European Marketing Academy) and Vice President of EDAMBA (European Doctoral Programmes Association in Management and Business Administration). His areas of scientific research interest and expertise are in Consumer Behavior, E-Marketing, Business Intelligence/Analytics and Tourism Marketing.

2.3 Article nr. #3

In this paper, the bank telemarketing problem dataset is re-evaluated for assessing if, within the available data, is there any more relevant information that can be extracted for using it to improve previous model performance. The emphasis is on including past history information in the dataset, considering the time period 2008 to 2013 includes clients that were contacted more than once. The new features proposed are related to customer lifetime value assessment.

The main contributions to the thesis include both the improvement in model performance, and the assessment of the relevance of some of the newly proposed history features. The new model can now reach around 83% of the successes by selecting the half most likely subscribers. While the improvement in this metric is just 4%, it is worthy of noticing that such enhancement was achieved with the same information that was initially available. Two of the newly proposed features ranked in the third (the last result for a previous campaign selling the same deposit) and the fourth (number of previous successes) positions of the most relevant features for the model. These findings validate the value of the proposed history feature selection strategy.

Article details:

- Title: Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns;
- Date: January 2015 (made available online since 2 September 2014);
- Journal: Neural Computing & Applications;
- Publisher: Springer.

Neural Comput & Applic (2015) 26:131–139
DOI 10.1007/s00521-014-1703-0

ORIGINAL ARTICLE

Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns

Sérgio Moro · Paulo Cortez · Paulo Rita

Received: 24 February 2014 / Accepted: 17 August 2014 / Published online: 2 September 2014

Abstract Customer lifetime value (LTV) enables using client characteristics, such as recency, frequency and monetary value, to describe the value of a client through time in terms of profitability. We present the concept of LTV applied to telemarketing for improving the return-on-investment, using a recent (from 2008 to 2013) and real case study of bank campaigns to sell long-term deposits. The goal was to benefit from past contacts history to extract additional knowledge. A total of twelve LTV input variables were tested, under a forward selection method and using a realistic rolling windows scheme, highlighting the validity of five new LTV features. The results achieved by our LTV data-driven approach using neural networks allowed an improvement up to 4 pp in the Lift cumulative curve for targeting the deposit subscribers when compared with a baseline model (with no history data). Explanatory knowledge was also extracted from the proposed model, revealing two highly relevant LTV features, the last result of the previous campaign to sell the same product and the frequency of past client successes. The obtained results are particularly valuable for contact center companies, which can improve predictive performance without even having to ask for more information to the companies they serve.

S. Moro (✉)
Department of Information Science and Technology, ISCTE - University Institute of Lisbon, 1649-026 Lisbon, Portugal
e-mail: scmoro@gmail.com

P. Cortez
Department of Information Systems/ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal

P. Rita
Business Research Unit (BRU-UNIDE), ISCTE - University Institute of Lisbon, 1649-026 Lisbon, Portugal

Keywords Customer lifetime value (LTV) · Multilayer perceptron · Recency, frequency and monetary value (RFM) · Telemarketing · Bank deposits · Data mining

1 Introduction

Customer lifetime value (LTV) stands for the value of a customer in terms of expected benefits considering likely future interactions with the customer [9]. LTV can be regarded as a relevant construct for every decision to improve customer relationship and profitability.

The identification of the most profitable customers, aiming to redirect a larger amount of the marketing effort toward those customers, has been a holy grail of marketing [21]. It is directly associated with predicting future behavior of customers, meaning that a good computation of LTV can help serving such purpose.

Typically, one way of characterizing a database of customers is by computing their recency, frequency and monetary (RFM) characteristics. These allow to capture customer behavior in a very small number of features [18]. In effect, RFM can be used as a base to compute LTV [10]. Still, the relative importance among RFM varies with the characteristics of the product and industry. In fact, feature weight has been a subject of research in order to improve classification accuracy [1]. There are several approaches to determine the relevance of each of the three RFM features. Liu and Shih [17] used the analytic hierarchy process to determine the relative weights of RFM variables in the evaluation of LTV. Considering more recent studies, the work of Cheng and Chen [5] used the K-means algorithm to build clusters by RFM attributes, resulting in an enhancement of classification accuracy when applied to a Taiwanese company operating in the electronic industry.

Kwon and Lee [16] measured loyalty by using a case-based reasoning method to compute the relationship of a user through its RFM and a smart object that attributes weights to RFM features based on its goal.

While most studies focus on optimizing the relation between each RFM parameter, Chen et al. [4] targeted the discovery of future RFM values by applying sequential pattern mining. The research used a Taiwanese supermarket chain and the results of the proposed algorithm were compared with the generalized sequential pattern a priori algorithm, showing special evidence of improvement in RFM values by cutting off more uninteresting patterns.

Another trend is to increase RFM modeling knowledge by adding other problem characterizing features. Yeh et al. [27] included two additional parameters, time since first purchase and churn probability, to model the likelihood that a customer will buy next time. Their research used a blood transfusion service for empirical analysis, and the results showed greater predictive accuracy than using single RFM traditional approaches.

The evaluation of LTV is a subject widely studied. Keramati et al. [15] conducted an assessment of weaknesses and risks in customer relationship management implementations when considering the return in terms of LTV. Customer analysis and segmentation may help to enhance future targeting in terms of customer responses to marketing campaigns [26], increasing LTV. Still, it is remarkably difficult to predict future behavior of customers with effective accuracy [19]. Simply there are too many variables to account for. Each customer is a specific individual and will take any opportunity to get more value for his/her money, considering his/her own personal benefit, which varies according to each person. Furthermore, each market and company has specific contexts which are huge in its influencing characteristics making them hard to be computationally modeled. Therefore, every research that shows an improvement in predictions for targeting is relevant in its context environment.

It is worth of noticing that decision support systems and neural networks (NN) usage and its effect on real business by enabling better decision making aligned with business needs have been considered a differentiating factor to increase the return-on-investment, particularly in the banking industry [2, 3, 25]. This emphasizes the value of such systems and justifies the research in an attempt of improvement the accuracy of the baseline system.

Previously, we reported the research for a data mining approach to predict the success of telemarketing to sell bank deposits [20]. The goal for this problem was to reduce the number of contacts while at the same time minimizing the loss of successful contacts due to model error. Thus, we worked toward a reduced number of false positives (FP) while maintaining a high number of true positives (TP), to the detriment of a higher number of false negatives (FN). The results achieved were considered of very good quality: By selecting the half better classified clients in terms of probability of subscribing the deposit, the best model would reach 79 % of successful contacts from the total subscribers in the whole dataset. Still, the analysis did not include historical information from previous marketing contacts to the same clients. Rather than proposing new machine learning algorithms, this paper focuses on feature engineering, which is argued in Domingos [8] as a key issue for providing better predictive capabilities in real-world applications. In particular, in this paper, we study the utility of customer telemarketing historical data, with a focus on LTV and related features, such as RFM. The main contributions of the research are as follows:

- performing a feature forward selection procedure, in an attempt to enhance the original model (from the previous research) by adding client history features, such as RFM.
- comparing the proposed approach with previous research (used as a baseline) and analyze the improvement achieved in the light of LTV relevance.
- finally, showing how the model using customer history information might benefit the bank telemarketing business in terms of targeting.

This paper is organized as follows: Sect. 2 presents the case study used for comparison and the techniques used; in Sect. 3, the experimental design is described and the obtained results are analyzed; finally, conclusions are drawn in Sect. 4.

2 Materials and methods

2.1 Data mining

For proper comparison purposes, we adopt the same methodology that was followed in our previous work [20] and that is described in detail in this section. For the experimental setup, we chose the rminer package of the R tool, which provides a simple set of coherent functions designed specifically for conducting data mining computation in a very intuitive way [6]. In Moro et al. [20], four data mining techniques were explored: logistic regression, decision trees, support vector machines and NN. The best result was achieved by the NN; thus, this is the only technique used for the experiments reported in this paper.

The NN is based on the popular multilayer perceptron with one hidden layer with H hidden nodes and one output node [14]. The input layer holds the input vector and then propagates the activations in a feedforward fashion, via weighted connections, through the entire network.

For a given input $\mathbf{x}_k$ the state of the ith neuron (s_i) is computed by:

$$s_i = f\left(w_{i,0} + \sum_{j \in P_i} w_{i,j} \times s_j\right) \tag{1}$$

where P_i represents the set of nodes reaching node i; f is the logistic function; $w_{i,j}$ denotes the weight of the connection between nodes j and i; and $s_1 = x_{k,1}$, $\ldots$, $s_M = x_{k,M}$. The logistic function allows to model the output response as a probability, where the output response s_o should be one for a successful contact. The training of multilayer perceptron is not optimal, since the final solution is dependent of the choice of NN starting weights. To solve this issue, the *rminer* package uses an ensemble of N_r different trained networks and outputs the average of the individual predictions [13]. In [20], the NN ensemble is composed of $N_r = 7$ distinct networks, each trained with 100 epochs of the BFGS algorithm and the final ensemble response is given by $\sum_{j=1}^{7} s_{o,j}/7$, where $s_{o,j}$ denotes the output success probability for the jth multilayer perceptron of the ensemble.

To evaluate the performance of predictions, two popular client targeting classification metrics are adopted: area under the curve (AUC) of the receiver operating characteristic (ROC) graph and area of the lift cumulative curve (ALIFT). The ROC curve shows the performance of a two class classifier across the range of possible threshold ($D \in [0, 1]$) values, where the NN predicted class is interpreted as positive if $s_o > D$, else it is interpreted as negative [11]. The ROC curve plots the false-positive rate (FPR), in the x-axis, versus the true-positive rate (TPR), in the y-axis As stated in Sect. 1, our main goal was to achieve a high TPR while reducing the total number of contacts, thus the D value point in the ROC graph should be in the lower values of FPR. The overall performance of the model can be measured by computing the area of the curve ($AUC = \int_0^1 ROC dD$), where a classifier is better if its AUC value is closer to 1.0.

Regarding the lift analysis, it sorts the records in a decreasing order of the predicted probability of success by dividing the population in deciles to facilitate selection of the most likely buyers, resulting in a popular measure in marketing [22]. The lift cumulative curve plots incrementally selected fractions of the population versus the real results, ordered from the most likely to the least likely buyers. Thus, a lift cumulative curve area (ALIFT) closer to 1.0 presents a better model that is capable of selecting more buyers in a smaller fraction of the population.

Table 1 Initial set of features used for modeling a successful telemarketing contact

Attribute	Description
nat.avg.rate	National monthly average of deposits interest rate
suited.rate	Most suited rate to the client according to its characteristics (e.g., if it holds a credit card, a mortgage account, assets above a certain threshold, etc.)
dif.best.rate.avg	Difference between best rate offered for the deposit (independent of client characteristics) and the national average
ag.gender	Gender of the agent (male/female) that made (outbound) or answered (inbound) the call
ag.generic	If generic agent, i.e., temporary hired, with less experience (yes/no)
ag.created	Number of days since the agent was created
cli.house.loan	If the client has a house loan contract (yes/no)
cli.affluent	If the client is an affluent client (yes/no)
cli.indiv.credit	If the client has an individual credit contract (yes/no)
cli.salary.account	If the client has a salary account (yes/no)
call.dir	Call direction (inbound/outbound)
call.nr.schedules	Number of previously scheduled calls during the same campaign
call.prev.durations	Duration of previously scheduled calls (in *s*)
call.month	Month in which the call is made
cli.sec.group	Security group bank classification
cli.agreggate	If the client has aggregated products and services
cli.profile	Generic client profile, considering assets and risk
emp.var.rate	Employment variation rate, with a quarterly frequency
cons.price.idx	Monthly average consumer price index
cons.conf.idx	Monthly average consumer confidence index
euribor3m	Daily three month Euribor rate
nr.employed	Quarterly average of the total number of employed citizens
outcome	Unsuccessful or successful contact (output target)

Complex data-driven models, such as NN, tend to provide accurate predictions, but the obtained models are difficult to be understood by humans. To open the "black box," we adopt the data-based sensitivity analysis (DSA) algorithm [7], which is a sensitivity analysis technique that works by analyzing the responses of a model when a given input is varied through its domain. The analysis of the sensitivity results allows to rank the input attribute influence and also show the average effect of the most relevant features in the model responses. Visually, the former can be analyzed using a input importance bar plot, while the later can be inspected using variable effect characteristic (VEC) curve.

2.2 Bank telemarketing data

For our experiments, we use real data collected from a Portuguese bank, consisting in telemarketing campaigns to sell long-term deposits and encompassing a period of 5 years, from May 2008 to June 2013. All contacts are executed through phone calls with a human agent as the interlocutor. The vast majority of the contacts are outbound, while the few inbound contacts are used when the client calls the bank for any other reason and the agent takes advantage of the contact to try to sell the deposit.

The dataset consists of 52,944 contacts executed through phone calls where only 6,557 of them resulted in successful deposit subscriptions; thus, it is an unbalanced dataset. The previous research [20] initially analyzed 150 characterizing features, but after careful selection, adopted only a subset of 22 features for prediction (listed on Table 1). This subset will serve as our baseline, for enrichment with additional LTV-related features.

As stated previously, this study is an attempt of improving the previous work results [20]; thus, the data evaluation procedures are exactly the same, to allow a direct comparison of the results. The whole dataset is divided into two subsets: a training set (from May 2008 to June 2012), with 51,651 of the oldest contacts and a test set (from July 2012 to June 2013), which includes the most recent 1,293 contacts, i.e., the ones that will effectively be used for prediction evaluation. To simulate a real runtime execution environment, a rolling window realistic scheme (of fixed-length of size W) is used, which performs several model updates and discards oldest data [24]. This method represents an ongoing process that defines the model with the latest data, performs K predictions and feeds back the model with the results from the most recent K predictions, discarding the oldest K, adapting the model to the most recent reality and keeping the process time to fit the model constant by maintaining the number of contacts used to build the model equal to W. The usage of only the most recent window of contacts in each prediction ensures that every modeling is computed with records that reflect a very recent past. In particular, we use a rolling window of size W = 20,000, with the most recent W training examples at time t, in order to fit a NN and then predict the next $K = 10$ future contact outcomes. In the next iteration, the training examples are updated by discarding the oldest K contacts and incorporating K new real outcome values (available at time $t + K$). Then, a model update is fit and K new predictions are executed. This procedure is repeated, until there is a total of $U = L/K = 130$ model updates (i.e., NN trainings and predictions), where L = 1,293 is the length of the test set. Using this procedure, the results achieved for the two metrics with the baseline method were 0.794 for AUC and 0.672 for ALIFT.

Table 2 RFM telemarketing features analyzed

Factor	Reference	Citation	Application to telemarketing
recency	[18]	How recent is the last purchase?	Months since the last purchase up to date (for our case, 2008–2013, we choose months since they are enough for discriminating the records)
	[10]	Time of most recent purchase	
	[17]	Period since the last purchase	
	[23]	The total days between the day of the latest purchase and analysis (days)	
	[4]	The period since a customers last purchase	
frequency	[18]	How often does a customer buy a product?	Number of times the client subscribed the deposit previously
	[10]	Number of prior purchases	
	[17]	Number of purchases made within a certain period	
	[23]	Consuming frequency (times)	
	[4]	The number of purchases made within a certain period	
monetary	[17]	The money spent during a certain period	Total amount of money the client subscribed in previous contacts
	[23]	Amount of money of total consuming	
	[4]	The amount of money that a customer spent during a certain period	
monetary. successes. avg	[18]	How much money does the customer spend per order?	Average value subscribed per success (up to date)
	[10]	Average purchase amount per transaction	

2.3 Customer lifetime value features

As stated previously, the model built with the 22 input features listed in Table 1 constitutes the baseline model. To enrich it, we listed all possibly LTV characterizing features that can be made available based on the telemarketing history records from previous contacts. Table 2

Table 3 Other telemarketing client historical features analyzed

Feature	Description
last.result	Last campaign result in which the client was contacted
last.result.prod	Similar to last.result but considering only campaigns where the same product was being sold
prev.contacts.durations	Total call durations for every previous contacts
prev.contacts.durations.avg	Similar to prev.contacts.durations but considering average
total.contacts	Total number of contacts
successes.per.contacts	Total successes/total contacts
successes.minus.unsuc	Total successes minus total unsuccessful contacts

defines the groups of features analyzed for RFM characteristics. Those are well documented and explored in the literature, thus are natural candidates for incorporating LTV in our decision support system. Only one attribute is used to represent the recency and frequency factors, while two input features are proposed to describe the monetary component. While the RFM concepts are widely known in the literature, authors have slightly different definitions for each of them. Some of the definitions are quoted from the references cited in the second column of Table 2. Besides the RFM-related attributes, we also consider seven other LTV features that are described in Table 3. We should note that some of those features depend upon each other, representing variations which we wanted to test whether they affected the model.

To select which of these features do in fact provide added value to the predictive model, we first group the features into logical blocks. Then, we adopt a feature selection approach using the popular forward selection technique [12]. In each iteration of this forward selection, we add a few features (related to logical blocks, described in Sect. 3.1) to the original model. If its predictive performance improves (in terms of AUC and ALIFT metrics), the features are kept. Otherwise, these are discarded. Then, a next iteration is executed, in order to test a few more features. The procedure is repeated until all LTV features have been tested.

It is worth to be noticed that while some clients had previously been contacted which allows to provide history information, others did not. Thus, in order to fully evaluate the difference between using or not history information, we compute the metrics of AUC and ALIFT for all clients (overall, total of 1,293 contacts in the test set), for clients with history (the ones with previous contacts, 353 of 1,293 telemarketing calls) and with no history (950 of 1,293 contacts).

3 Experiments and results

3.1 Customer lifetime value feature analysis

All experiments described in this article were executed using the rminer package and R tool [6]. The NN ensemble is composed of $N_r = 7$ distinct networks, each trained with 100 epochs of the BFGS algorithm. For setting the number of hidden nodes (H), we performed a grid search where the number of hidden nodes was searched within the set $H \in \{0, 2, 6, 8, 10\}$. The rminer package applies this grid search by performing an internal holdout scheme over the training set (with 2/3 of the data), in order to select the best H value, that corresponds to the lowest AUC value measured on a subset of the training set, and then trains the best model with all training set data.

The rolling windows procedure was executed for the baseline model (with 22 features from Table 1) in order to extract the values of AUC and ALIFT metrics for both groups of contacts (with and without history). The results are shown on the baseline row of Table 4. We found only a very slight difference (0.0116 in terms of AUC) between the prediction results for clients with and without telemarketing history. For comparison purposes, we also show the results for predictions without using the rolling windows procedure, to assess the benefits of adapting the model iteratively with the most recent data. The difference in the performance measurements of both AUC and ALIFT is quite significatively. Considering this remark, the rolling windows procedure was adopted for the remaining experiments.

The rows of Table 4 show the effect of adding each new group of features to the initial set of 22 features (from 1). The table is divided by thicker lines according to the logical blocks (after the baseline). The first block includes the standard RFM features, under two configurations (with monetary or monetary.success.avg); next, comes the last campaign result, followed by the duration of phone calls; finally, three isolated features appear, as each tries to add specific telemarketing-associated knowledge.

The best results for each logical group of features are signaled through a gray background cell and white digits. As expected, the results of both AUC and ALIFT metrics do not change much for the contacts without history information, when compared with the baseline values (variation below 0.0129 for AUC and 0.0078 for ALIFT). However, for the group of 353 contacts with history information, the metrics are consistently increased as each group of features is added. In effect, AUC increases from 0.8002 to 0.8609, while ALIFT is enhanced from 0.6701 to 0.7044. The exceptions are the last three blocks, which included a single new feature each, and resulted in poorer performances for both

Table 4 Results of the forward selection procedure used to select LTV features

Feature	**Overall**		**With history**		**No history**	
	AUC	ALIFT	AUC	ALIFT	AUC	ALIFT
baseline model [20]	0.7935	0.6718	0.8002	0.6701	0.7886	0.6746
baseline (no rolling windows)	0.7443	0.6718	0.7513	0.6444	0.7393	0.6472
recency, frequency, monetary	0.8142	0.6841	0.8382	0.6917	0.7868	0.6731
recency, frequency, **monetary.success.avg**	**0.8170**	**0.6857**	**0.8423**	**0.6940**	0.7886	0.6744
recency, frequency, monetary.success.avg, **last.result**	0.8166	0.6853	0.8398	0.6924	0.7904	0.6760
recency, frequency, monetary.success.avg monetary, **last.result.prod**	**0.8233**	**0.6893**	**0.8480**	**0.6972**	0.7943	0.6779
recency, frequency, monetary.success.avg, last.result, last.result.prod	0.8197	0.6873	0.8423	0.6940	0.7933	0.6773
recency, frequency, monetary.success.avg, last.result.prod, **prev.contacts.durations**	**0.8247**	**0.6901**	**0.8609**	**0.7044**	0.7825	0.6709
recency, frequency, monetary.success.avg, last.result.prod, **prev.contacts.durations.avg**	0.8227	0.6887	0.8553	0.7013	0.7853	0.6725
recency, frequency, monetary.success.avg, last.result.prod, prev.contacts.durations, **total.contacts**	0.8197	0.6871	0.8525	0.6998	0.7814	0.6701
recency, frequency, monetary.success.avg, last.result.prod, prev.contacts.durations, **successes.per.contacts**	0.8193	0.6868	0.8473	0.6966	0.7871	0.6738
recency, frequency, monetary.success.avg, last.result.prod, prev.contacts.durations, **successes.minus.unsuc**	0.8215	0.6882	0.8507	0.6986	0.7874	0.6740

metrics, thus being discarded. The selected model got enriched with new LTV features: recency, frequency, monetary value considering average of successful contacts, the result for the last campaign to sell the same product and the total time spent on previous contacts for past campaigns.

To visually compare the quality of prediction results using client history information versus without history, we plot for both groups the ROC curves [11] and the cumulative Lift curves [22], respectively, in Figs. 1 and 2.

Considering the ROC curve, Fig. 1 shows that the usage of the LTV history features benefits the results particulary for the lower values of FPR, meeting our goal, as stated in Sect. 2.1. The Lift figure shows that both curves are next to each other, although the largest curve area, which represents the prediction capability for contacts with history information, stands consistently above the baseline (without LTV features). For some client selection sample sizes' (x-axis), such as 60 and 70 %, the difference is higher than 4 pp. This in an interesting discovery that directly benefits business, as we used solely telemarketing history information that is easily available at telemarketing service operators. Furthermore, contact center companies can use this type of information to enhance telemarketing campaigns without even having to ask for more information to their clients.

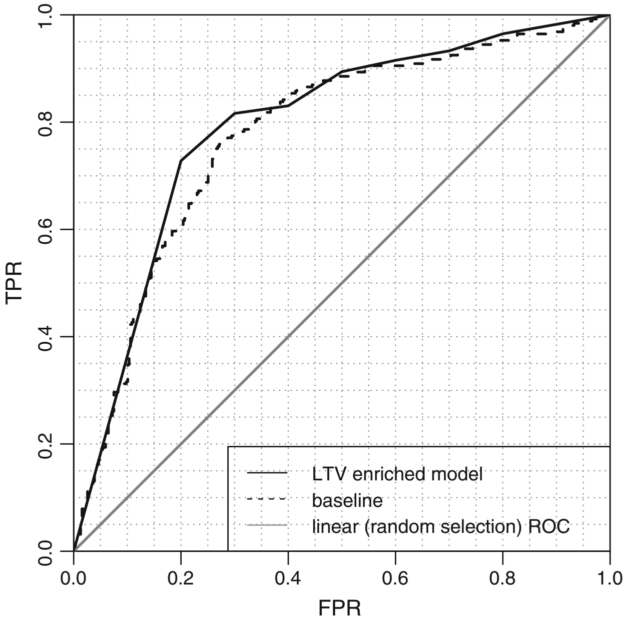


Fig. 1 Bank telemarketing ROC curves (with and without LTV history)

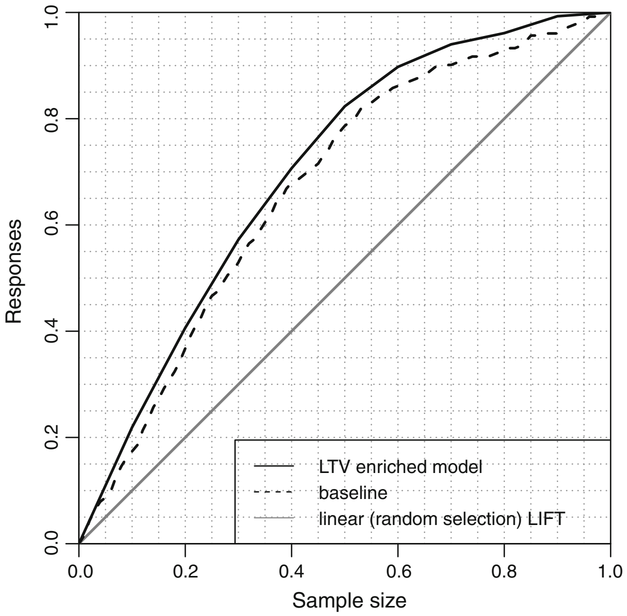


Fig. 2 Bank telemarketing lift cumulative curves (with and without LTV history)

3.2 Explanatory knowledge

In this section, we extract explanatory knowledge using a sensitivity analysis procedure. First, the best model (with 27 features, including the novel five LTV inputs) is fit with all data contacts. Then, the DSA algorithm [7], which is capable of measuring the global influence of an input, including its iterations with other attributes, is executed with its default parameters (e.g., use of seven levels for numeric inputs) on the best selected NN. The respective DSA input importance bar graphic is plotted in Fig. 3.

We detail our analysis to the top five most relevant features. Figure 4 compares the top five input features for the baseline and LTV enhanced model. The difference between the best rate offered for the deposit and the national average is now considered the most relevant feature, while previously it was the fourth in the rank [20]. This is a significant change for the model especially since this feature alone got a relevance higher than 15 %, that is, 5 pp increase when compared to the initial baseline model. On the other hand, the euribor rate dropped from a relevance of 15 % to slightly above 10 %, dropping to second place in the rank. We note that the sensitivity method adopted (DSA) is capable of measuring the global influence of a feature in a predictive model, including its interactions with other features. Given that the main difference between the two predictive models is the inclusion of the new LTV inputs, obtained results suggest that the best predictive model performs a higher degree of interaction between the added LTV features with the difference rate attribute.

More importantly, two of the newly proposed LTV inputs are highly ranked (denoted by a star symbol in Fig. 4): The last result for previous campaign to sell the same product (last.result.prod) is ranked third, while frequency of successes is ranked as the fourth most relevant input. From the features added, it is worth to notice that while frequency is the fourth most relevant, recency only comes in eleventh place (Fig. 3), with a relevance that is roughly half of the one obtained by frequency. This is a different result when compared with the work of [17], which points to recency as being more relevant than frequency.

Next, we analyze the average influence of the two most relevant LTV features. Figure 5 plots the respective VEC curves, where the x-axis denotes the range of values of the input and y-axis represents the expected average change in the output response. As suggested in [7], the x-axis is scaled, in order to compare the influence of two distinct inputs in the same graph. The obtained results are aligned with our expectations: If the client has subscribed the deposit in the last campaign through which he/she was contacted, then it is much more likely that he will subscribe it again. Indeed, there is an improvement of around 20 pp in the success probability when the last product result changes from unsuccessful to successful. Considering frequency, an increase in the number of successes also improves the probability for a next successful result. It should be noted that the improvement in the subscription probability is not linear, with the highest increase (around 10 pp) being set between 0 and 1 frequency interval.

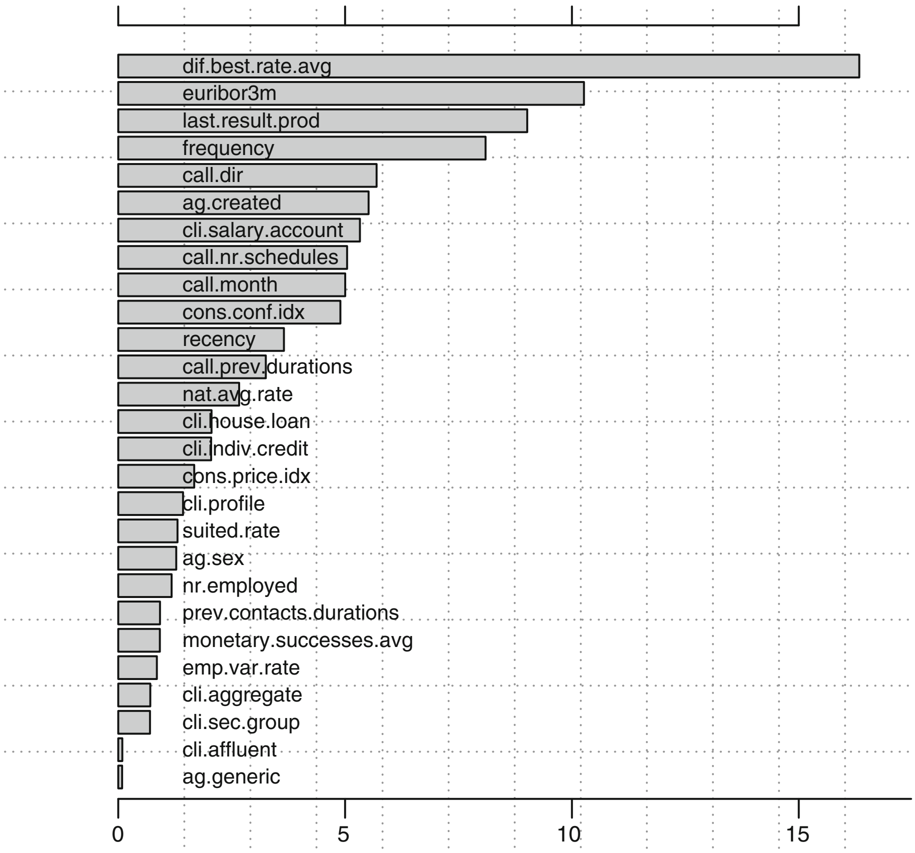


Fig. 3 Relative importance of input features to the data-driven model (in %)

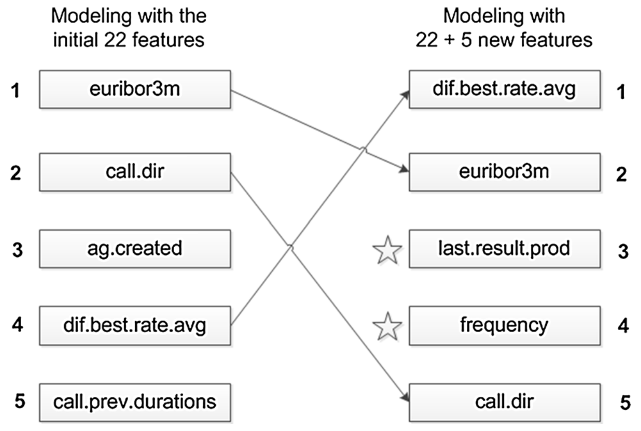


Fig. 4 Feature relevance evaluation (baseline and LTV enhanced models)

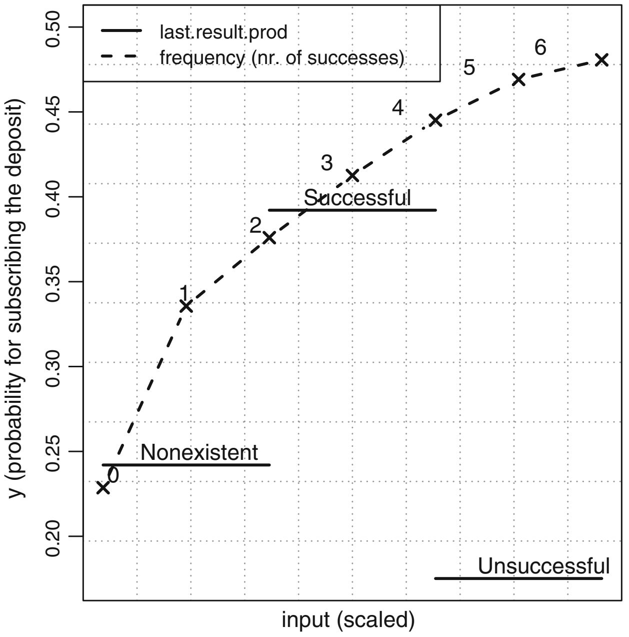


Fig. 5 Influence of the two LTV most relevant features (last.result.prod and frequency)

4 Conclusions

In a mature ongoing marketing business, usually it is available inside the company raw information that can potentially increase the LTV of customers. The usage of LTV history information, such as RFM characteristics to enhance data-driven models, is thus a key issue to improve prediction accuracy in marketing applications.

In this paper, we applied the concept of LTV by incorporating history information to enhance prediction capabilities of an already robust baseline decision support system using NN to sell bank deposits in a telemarketing campaign context. A forward selection technique was conducted, where twelve LTV candidate input features were tested. The evaluation procedure, using a robust and realistic rolling window scheme, and two metrics, favored a data-driven model that included five LTV features. When compared with the baseline model (with no LTV features), the enhanced LTV model produced an improvement of 6 pp in the area of ROC curve, with a total AUC = 0.86, and 4 pp in the cumulative lift curve, with a total ALIFT = 0.70, for clients with previous telemarketing history. Moreover, the improvement is perfectly consistent, meaning that whatever is the portion of clients selected for targeting, the estimated number of successes is always better than the baseline system. Furthermore, such LTV history information is easily available at marketing service operators and thus can be immediately used to benefit business (as opposed to external information, which might require requests and time for its acquisition and integration).

We also extracted explanatory knowledge from the LTV enhanced model, by using a sensitivity analysis procedure that allows to rank the inputs and show the global influence of each input in the data-driven model. Two of the newly LTV input variables were included in the top five most relevant features, confirming the utility of the proposed approach, namely the last result for previous campaign to sell the same product and the frequency of successes.

In future research, we intend to gather additional more recent data and perform new tests to address the limitation of using predictions of only 353 contacts. As time goes by and new telemarketing campaigns are executed, one can expect that more clients will be contacted again, allowing to benefit from past history information. It would be interesting to understand whether the results achieved are time proof or additional history provides different results in terms of LTV features.

References

1. Ahn H, Kim K, Han I (2006) Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system. Expert Syst 23(5):290–301
2. Bahrammirzaee A (2010) A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Comput Appl 19(8):1165–1195
3. Burrell PR, Folarin BO (1997) The impact of neural networks in finance. Neural Comput Appl 6(4):193–200
4. Chen Y-L, Kuo M-H, Tang K (2009) Discovering recency, frequency, and monetary (RFM) sequential patterns from customers purchasing data. Electron Commer Res Appl 8(5):241–251
5. Cheng C-H, Chen Y-S (2009) Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst Appl 36(3):4176–4184
6. Cortez P (2010) Data mining with neural networks and support vector machines using the r/rminer tool. In: Perner P (ed) Advances in data mining. Applications and theoretical aspects, vol 6171. Springer, Berlin, pp 572–583
7. Cortez P, Embrechts MJ (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci 225:1–17. doi:10.1016/j.ins.2012.10.039
8. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10):78–87
9. Dwyer FR (1997) Customer lifetime valuation to support marketing decision making. J Interact Mark 11(4):6–13
10. Fader PS, Hardie BGS, Lee KL (2005) RFM and CLV: using iso-value curves for customer base analysis. J Mark Res 42(4):415–430
11. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874. doi:10.1016/j.patrec.2005.10.010
12. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. doi:10.1016/j.ejor.2010.10.019
13. Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, NY
14. Haykin SS, Haykin SS, Haykin SS, Haykin SS (2009) Neural networks and learning machines, vol 3. Prentice Hall, New York
15. Keramati A, Nazari-Shirkouhi S, Moshki H, Afshari-Mofrad M, Maleki-Berneti E (2013) A novel methodology for evaluating the risk of CRM projects in fuzzy environment. Neural Comput Appl 23(1):29–53
16. Kwon O, Lee N (2011) A relationship-aware methodology for context-aware service selection. Expert Syst 28(4):375–390
17. Liu D-R, Shih Y-Y (2005) Integrating AHP and data mining for product recommendation based on customer lifetime value. Inf Manag 42(3):387–400
18. Madeira S, Sousa JM (2002) Comparison of target selection methods in direct marketing. In: European symposium on intelligent technologies, hybrid systems and their implementation on smart adaptive systems. Citeseer
19. Malthouse EC, Blattberg RC (2005) Can we predict customer lifetime value? J Interact Mark 19(1):2–16
20. Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 62:22–31
21. Nevid JS (2010) Introduction to the special issue: implicit measures of consumer response the search for the holy grail of marketing research. Psychol Market 27(10):913–920. ISSN 1520–6793. doi:10.1002/mar.20365
22. Olson DL, Chae BK (2012) Direct marketing decision support through predictive customer response modeling. Decis Support Syst 54(1):443–451
23. Shen C-C, Chuang H-M (2009) A study on the applications of data mining techniques to enhance customer lifetime value. WSEAS Trans Inf Sci Appl 6(2):319–328
24. Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. Int J Forecast 16:437–450
25. Wu DD, Hall JG (2011) Special issue: business decision support systems. Expert Syst 28(3):197–198
26. Yao Z, Sarlin P, Eklund T, Back B (2012) Combining visual customer segmentation and response modeling. Neural Comput Appl 25(1):123–134
27. Yeh I, Yang K-J, Ting TM et al (2009) Knowledge discovery on RFM model using bernoulli sequence. Expert Syst Appl 36(3):5866–5871

2.4 Article nr. #4

This article proposes a different strategy for unveiling new features. It focuses on a divide and conquer approach by assessing the most logical division of the bank telemarketing problem in smaller and more manageable problems. The five most relevant features for previous model are re-assessed by a domain expert for deciding on where to split the problem. Then a feature selection procedure is undertaken focusing on the new sub-problem.

The present thesis benefits from two main contributions: the procedure proposed for finding the division of the problem and the impact in terms of model performance; and the high relevance of some of newly included features. Using a domain expert, the bank telemarketing was divided in inbound and outbound. The new model can now reach the total number successes from the initial contacts by selecting the half most likely buyers. Also, the second and third most relevant features were not included in previous model (the time since previous inbound call and the number of previously outbound calls, respectively), showing the value of the divide and conquer strategy proposed.

Article details:

- Title: A divide and conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing;
- Date: In preparation.

A divide and conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing

Sérgio Miguel Carneiro Moro[a,*], Paulo Alexandre Ribeiro Cortez[b], Paulo Miguel Rasquinho Ferreira Rita[c]

[a] *Business Research Unit (UNIDE-IUL), Dep. Information Science and Technology, ISCTE - University Institute of Lisbon, 1649-026 Lisbon, Portugal*

[b] *ALGORITMI Research Centre/Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal*

[c] *Business Research Unit (BRU-UNIDE), ISCTE University Institute of Lisbon, 1649-026 Lisbon, Portugal*

Abstract

The discovery of knowledge through data mining provides a valuable asset for addressing decision making problems. While a list of features may characterize a problem, it is often the case that a subset of those features may influence more a certain group of events constituting a sub-problem within the original problem. We propose a divide and conquer strategy for data mining using both feature relevance and expert evaluation for splitting the problem of characterizing telemarketing contacts to sell long-term bank deposits. As a result, the call direction was considered the most suitable candidate feature, dividing the telemarketing dataset in inbound and outbound contacts. The inbound telemarketing sub-problem was then re-evaluated, leading to a different selection of relevant features. The results validate the benefits of such approach, with a large increase in the area under the Receiver Operating Characteristic curve, from 0.75 (original solution) to 0.89 (enhanced dataset).

Keywords:

[*] Corresponding author (S. Moro).

Email addresses: scmoro@gmail.com (Sérgio Miguel Carneiro Moro), pcortez@dsi.uminho.pt (Paulo Alexandre Ribeiro Cortez), paulo.rita@iscte.pt (Paulo Miguel Rasquinho Ferreira Rita)

January 12, 2015

Data mining, Telemarketing, Feature relevance, Banking, Modeling, Bank deposits

1. Introduction

Data mining (DM) enables to unveil previously undiscovered knowledge, providing a leverage for decision making (Witten & Frank, 2005). The basic ingredient for DM is raw data, representing known instances of an interesting phenomenon to be modeled (i.e., the problem) that can be characterized by a list of features. A key aspect for a successful DM project is feature engineering, since data features need to be related with the problem modeled (e.g., desired outcome to predict) (Chen et al., 2009).

DM has been successfully applied to a wide number of domains, ranging from medicine to managerial problems, including banking and marketing (Turban et al., 2011). Classification is one of the most relevant DM tasks, where the goal is to predict a certain discrete outcome (output target) for a problem instance (characterized by a set of input features). One of the most prolific research contexts for DM applications is credit risk, comprehending several classification problems such as the profiling of credit card holders for defining credit limit and credit risk scoring to assess the risk of default (Hsieh, 2004; Wang et al., 2011). DM can assist banks with the automation of the loan application process by accurately predicting the most likely defaulters. In marketing, the problem of targeting the right customers with certain products or services is especially tailored for DM (Kim & Street, 2004; Ngai et al., 2009). Thus, every bank product or service offered in the context of marketing campaigns can potentially benefit from a DM approach.

In previous works, we have addressed the telemarketing problem of predicting the result from phone call campaigns to sell bank deposits with interesting results. First, a bank telemarketing dataset was enriched with social and economic context features, leading to a tuned model that enabled to reach 79% of the deposit subscribers by selecting the half better classified clients (Moro et al., 2014). Such a model was then improved by including customer lifetime value related features, increasing the performance to 83% of subscribers with the same number of contacts (half from the total dataset) (Moro et al., 2015b). In this paper, we further enhance the classification model by presenting a divide and conquer strategy for splitting the problem of selecting the most likely subscribers of a bank long-term deposit offered in the context of telemarketing campaigns, where contacts were conducted through phone calls within the bank's contact center. It addresses two main issues: the proposed approach and selection of the best candidate feature for splitting the problem; and the analysis of the new sub-problem for assessing the benefits of the division suggested from the first issue results. This article is organized as follows. Section 2 describes the telemarketing and DM background related with this paper. Next, Section 3 presents the bank dataset, the methods used for the divide and conquer strategy proposed and for assessing the impact on the new sub-problem. Then, the results are discussed in Section 4. Finally, conclusions are summarized in Section 5.

2. Background

2.1. Bank Telemarketing

Direct marketing is the method of targeting specific customers allowing companies to promote products or services on an individual basis. The usage of a customer database can enhance the process, turning it into database marketing (Tapp, 2008). The telephone is still one of the major communication channels, emphasized by the proliferation of mobile devices. Marketing promotions and campaigns conducted using mainly the telephone are defined as telemarketing (Kotler & Keller, 2012).

Companies with the need to support a large number of customers typically concentrate communication in contact centers, where automated or human agents can answer customers. Such centers are also used to conduct telemarketing campaigns, which include performing outbound calls to a list of customers, or by taking advantage of inbound calls from customers to approach them in order to promote campaign products. The difference between inbound and outbound telemarketing has been reported by a few researches for problems such as simulation modeling and staffing (Mehrotra & Fama, 2003; Lin et al., 2010).

One of the major problems in telemarketing is to specify the list of customers which present a higher likelihood of buying the product being offered (Talla Nobibon et al., 2011). Decision support systems using predictive models can provide better informed decision making, increasing telemarketing managers' awareness of the impact resulting from predicted outcomes (Turban et al., 2011). DM modeling techniques allow unveiling patterns of information, translating it into knowledge which can be incorporated in such predictive systems (Witten & Frank, 2005). Despite the potential gain of using DM for modeling bank telemarketing success, few works have adopted such approach (Moro et al., 2015a). As explained by Moro et al. (2014), in 2012, we initially explored several data-driven models for targeting the subscription of long-term deposits but we only achieved accurate predictive models when using features that were only known on call execution, such as call duration. In 2014, such drawback was solved by adding features that could be known before call execution, namely: social and economic context related (e.g., Euribor rate) (Moro et al., 2014); and lifetime value features (e.g., frequency of past client successes) (Moro et al., 2015b). Also in 2014, Javaheri et al. (2014) followed a distinct approach, modeling the effect of mass media campaigns (e.g., television) on the buying of new bank products. As explained in the Section 1, this paper provides a purely novel approach to enhance the modeling bank telemarketing success, where a divide and conquer strategy, supported by expert domain knowledge, is used to split the dataset instances according to the direction of the call, with a particular focus on inbound call specific features.

2.2. Data Mining Learning Model

DM projects entail several relevant phases such as data cleaning, preparation, modeling and evaluation (Witten & Frank, 2005). Concerning the modeling phase, in Moro et al. (2014), four DM learning techniques were explored: logistic regression, decision trees, support vector machines and a neural network ensemble. The best result was achieved by the neural network, thus this is the only technique used for the experiments reported in this paper. It should be mention that neural network is a popular DM learning model, which attempts, to a certain degree, mimic human neural processing (Levine & Leven, 2014). It has been applied in various domains, including the forecasting of stock market index price and targeting customers in direct marketing campaigns (Enke & Thawornwong, 2005; Guido et al., 2011). Neural networks can be used for segmenting and targeting consumers as well as predicting sales. The most common architecture for a neural network algorithm is the multilayer perceptron (Haykin et al., 2009).

The neural network ensemble consists uses the popular multilayer perceptron as its base model, with a configuration consisting in one hidden layer of H hidden nodes and one output node. For a given input $\mathbf{x}_k$ the state of the i-th neuron (s_i) is computed by:

$$s_i = f(w_{i,0} + \sum_{j \in P_i} w_{i,j} \times s_j) \tag{1}$$

where P_i represents the set of nodes reaching node i; f is the logistic function; $w_{i,j}$ denotes the weight of the connection between nodes j and i; and $s_1 = x_{k,1}$, …, $s_M = x_{k,M}$. Given that the logistic function is used, the output node automatically produces a probability estimate (i.e., $s_{M+H+1} \in [0, 1]$). The ensemble consists of N_r different trained multilayer perceptrons, where the combined output is made in terms of the average of the individual neural network predictions.

2.3. Divide and Conquer Strategy

A wide number of divide and conquer strategies have been applied in the machine learning and DM domains. The division of a certain problem in smaller fractions is a natural means to reduce complexity to a manageable size. In fact, one of the most popular DM techniques, the decision trees, use a divide and conquer strategy for building the tree model, where each node divides a group of occurrences in smaller fractions (Quinlan, 1996).

Recent developments continue to use divide and conquer strategies for improving DM modeling tasks, such as: improving the performance of support vector machines (Hsieh et al., 2014); speeding up the genetic optimization of fuzzy cognitive maps (Stach et al., 2010); and handling large datasets by using an hierarchical classification for dividing the problem in smaller fractions which can be dealt with neural networks (Fritsch & Finke, 2012). While a large number of articles are devoted to DM algorithm design, a smaller number emphasizes the need to divide the problem to reduce the feature selection search space. However, so far the proposals are focused on automated solutions. The present study highlights a different mixed approach, which uses computed feature relevance combined with an expert knowledge.

2.4. Feature Relevance

In a DM project, the evaluation of feature relevance is quite useful in the pre-processing phase of data analysis. Some datasets are characterized by a large number of features, where only a few of them are relevant for that specific problem, burdening knowledge mining. Therefore, a feature selection step is often required in order to remove the least relevant features and/or to obtain final DM models that are easier to interpret (thus, enhancing explanatory knowledge) (Li et al., 2012).

While neural networks often provide better prediction results when compared to traditional models, such as logistic regression and decision trees, the difficulty to understand obtained data-driven model poses an obstacle to their adoption by managers. However, such model can be "opened" by using a sensitivity analysis procedure that analyzes the responses of a model when a given input is varied through its range. In particular, the data based sensitivity (DSA) approach can be adopted to rank the most relevant features and evaluate how they affect the classification target outcome (Cortez & Embrechts, 2013). DSA is a computationally feasible and enhanced sensitivity analysis variant that changes simultaneously several input features by using samples from the training set and thus can be used to detect the global effect of a feature even if such influence is only visible when it interacts with other input features.

2.5. Model Evaluation

An interesting approach to evaluate a classification problem, such as targeting customers for successful deposit subscriptions, is to use the probabilistic outcome computed by the data mining model. A successful result c is considered when $p(c|\mathbf{x}_k) > D$, where D is the probability threshold that defines achieved success and $p(c|\mathbf{x}_k)$ is the data-driven model response for the input client contact $\mathbf{x}_k$.

The advantage of using class probabilities is that a receiver operating characteristic (ROC) curve can easily be conducted by varying D from 0.0 to 1.0, displaying one minus the specificity (x-axis) versus the sensitivity (y-axis) (Fawcett, 2006). In effect, the area under the curve ($AUC = \int_0^1 ROCdD$) provides a popular classification metric to evaluate prediction results (Martens et al., 2011) and thus such metric was adopted in this work. A random classifier presents an AUC of 0.5. A higher AUC corresponds to a better discriminating model, with the perfect model achieving an AUC of 1.0.

Another adopted metric is based on the Lift analysis, which provides an interesting procedure to assess model prediction capabilities by dividing the dataset in deciles, ordered from the most to the least likely successful results (Witten & Frank, 2005). The cumulative Lift curve plots population samples (x-axis) versus the cumulative percentage of real responses captured ($y-$axis). Similarly to the AUC metric, the ideal method should present an area under the LIFT (ALIFT) cumulative curve close to 1.0.

3. Materials and Methods

3.1. Bank Case-study

This research type is case-study based using telemarketing contacts executed between May 2008 and June 2013 from a Portuguese bank for selling long-term deposit accounts, consisting in a total of 52944 records. In Moro et al. (2014) a feature selection and engineering procedure took place over the same dataset, from an initial base of 150 features. As a result, the final model used just 22 features. Subsequent work of Moro et al. (2015b) improved the model using five new customer lifetime value related features, increasing the total to 27.

For this work experiments, the five most relevant features for the best model (achieved by Moro et al. (2015b) and displayed in Table 1) are used for a human expert assessment of relevance, complementing previous feature relevance. Such expert is a telemarketing campaign manager, having worked three years as a technical Contact Center support and ten years as a technician in Marketing.

Table 1: Five most relevant features.

Feature	Description	Relevance
dif.best.rate.avg	difference between best rate offered and the national average	16.33%
euribor3m	daily three month Euribor rate	10.27%
last.result.prod	last campaign result for the same product in which the client was contacted	9.01%
frequency	number of times the client subscribed the deposit previously	8.10%
call.dir	call direction (inbound/outbound)	5.7 %

3.2. Data Mining

For computational experiments when modeling the sub-problem, the **rminer** package of the **R** tool was adopted, which provides a coherent and parameterizable small set of functions specifically designed for data mining computation (Cortez, 2010). To fit the data records, the choice was a neural network base learner, since this model provided the best results in previous studies with the same dataset, outperforming a support vector machine, a logistic regression and a decision tree. **Rminer** implements a multilayer perceptron ensemble with N_r learners, where each individual learner consists of a multilayer perceptron with several computing nodes organized in layers (Haykin et al., 2009). The input layer is fed with the input vector and propagates the activations in a feedforward fashion, via the weighted connections, through the entire network. The ensemble final response is set as the average of the distinct neural networks. The use of such ensemble turns the model less dependent on the random initialization of the neural network weights, as suggested by Hastie et al. (2009).

Regarding the configuration of the ensemble model, the neural network ensemble is composed of $N_r = 7$ distinct networks, each trained with 100 epochs of the BFGS algorithm. For setting the number of hidden nodes (H), a grid search was performed where the number of hidden nodes was searched within the set $H \in \{0, 2, 6, 8, 10\}$. The **rminer** package of the **R** tool applies this grid search by performing an internal holdout scheme over the training set (with 2/3 of the data), in order to select the best H value, that corresponds to the lowest AUC value measured on a subset of the training set, and then trains the best model with all training set data.

3.3. Divide and Conquer Strategy

For defining a division of the problem, first the five most relevant features from previous research were considered (Table 1). The telemarketing manager was asked to assess the relevance for each of those five features using his expertise, by selecting a quantitative metric for relevance from one (non-relevant) to ten (vital relevance). Next, an unstructured interview took place for discussing his assessment, including a comparison with feature relevance computed with the DSA sensitivity analysis method. The goal of such interview was specifically to find the most suitable candidate features for splitting the problem (and the dataset) in the light of telemarketing business management commitments. For justifying this approach, it can be argued that specific domain knowledge via human expertise remains the best method for characterizing a specific problem, as argued by Witten & Frank (2005). However, a quite important precaution was taken into account: the division of the dataset should not result in a workable dataset with fewer contacts than a thousand. Considering the modeling procedure explained in the next section, this was estimated as a reasonable minimum number of records for obtaining a reliable model. Thus, if the expert suggested a split leading to a very small dataset, the subject would have to be discussed in the light of feasibility of the experiments needed to validate such hypothesis. Figure 1 provides a simplified overview of the proposed procedure.

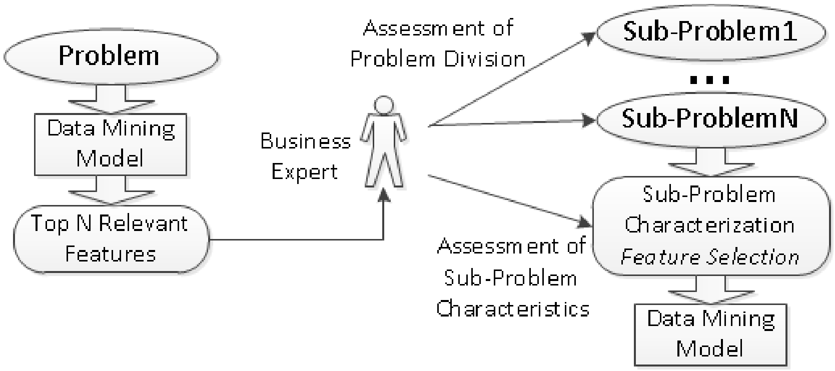


Figure 1: Proposed divide and conquer strategy

In the suggested strategy, a business expert intervenes in two phases: first, by assessing from a shortened list of the most relevant features for the DM model which is the more suitable candidate partition for the problem in analysis; then, for providing assistance in the characterization of the new sub-problem, which should be different than the main problem, justifying a new feature analysis. For the experiments hereby presented, the five most relevant features from the DM model were re-evaluated by the manager. However, this limitation of features was only imposed to save valuable time from the human expert who assisted the experiments. Such number could be increased if the intention was to allow the expert provide a broader analysis of problem characteristics available for modeling. This approach is an attempt to benefit from the best of two alternative strategies:

- usage of automated machine learning DM techniques, used to discover possible unknown yet useful knowledge from data; mixed with
- cognitive human thinking, which is known for making straightforward shortcuts for quick decision making.

Such approach addresses the complexity associated with the characterization of a problem using its features, which is highly application-dependent.

Furthermore, we highlight that intelligence behind many pattern recognition systems is supported by a human-engineered feature extraction process (Arel et al., 2010).

3.4. Modeling Procedure and Evaluation

After deciding which sub-problem should be addressed, a baseline model using the 27 features previously discovered was built, in order to allow a straightforward comparison (Moro et al., 2015b). At a first stage, a feature selection and feature engineering procedure was conducted, where the baseline features were merged with newly features selected by the telemarketing manager expert, drawn from the total number of features present in the original database (dataset with 150 attributes) (Moro et al., 2014). Then, a feature reduction was conducted in three iterative steps by using DSA to remove the least relevant features for modeling. The procedure was defined as follows:

(i) Firstly, asking the bank telemarketing manager to identify which other features (besides the 27 already identified), could better characterize the new sub-problem in which clients are approached to buy the campaign deposit. The telemarketing manager selected these extra features from the initial set of 150 features presented by Moro et al. (2014).

(ii) Then, modeling the subset telemarketing contacts considering:
   (a) the 27 original features;
   (b) the features selected by the telemarketing manager specifically for its potential of influencing the new sub-problem call results.

(iii) Afterwards, modeling using the whole features from (iia) and (iib).

(iv) Then proceeding with a three step iterative feature reduction phase, by excluding the least relevant features, as determined by a DSA, leaving only:
   (a) 1st step: the 20 most relevant features;
   (b) 2nd step: the 15 most relevant features;
   (c) 3rd step: the 10 most relevant features.

Thus the procedure includes first a feature selection, by adding features from phase (iia) to (iib), resulting in a list of the whole features (phase (iii)). Then, a feature reduction takes place from phase (iii) to phase (iv), including each inner step. By reducing to just as few as ten features, if the model achieves better performance as the baseline (27 features), it can serve as a proof of the large difference between addressing the telemarketing problem with the general solution and with the tuned model using specific sub-problem features.

In order to compare the results between every phase, both AUC and ALIFT metrics were computed, in an effort to improve AUC but also aligned with an increase in ALIFT. For assessing model performance, considering the new sub-problem dataset, two different methods were adopted:

- First, for every phases of the feature selection and reduction procedure, modeling was executed using a random holdout validation, with a selection of 2/3 of contacts for training and 1/3 for testing predictions. This procedure was performed for twenty times and computed average AUC and ALIFT metrics;

- Finally, a more realistic time ordered holdout was carried out with the most relevant features by selecting the 90% oldest contacts for constructing the model and then tested it on the more recent 10% of contacts. Such a different split from feature procedure intended to address the issue of having few contacts, thus we took a larger fraction for the training set for fitting more data. For comparison purposes, metrics were computed for modeling using the initial features, and the newly found characterizing features.

4. Experiments and Results

4.1. Divide and Conquer Strategy

The five most relevant features in the study of Moro et al. (2015b) were assessed by the human expert. It should be noticed that the expert was perfectly aware of the previous results, which could influence his judgment. However, Table 2 shows that the relevance experienced in the real environment is not the same as the computed from the DM model (shown in Table 1 and using DSA).

Call direction, the least relevant feature from the top five DM model DSA analysis, was considered the most relevant for the telemarketing manager. Such result pinpoints the need for a discussion to understand why human experience appointed to different direction from a computational result. First of all, the manager indicated the huge difference in the human agents handling the calls: while outbound agents are typically young university students (18-24 years) appealing for a part-time job to pay some of their

Table 2: Assessment of relevance by expert.

Model Result	Feature	Relevance by Expert (1 - 10)*
16.33%	dif.best.rate.avg	8
10.27%	euribor3m	6
9.01%	last.result.prod	7
8.10%	frequency	7
5.7%	call.dir	9

$\star$ - 1 - Non-relevant; 10 - Vital relevance.

day-to-day expenses, inbound agents are older, more experienced agents, with most of them having developed their career from outbound agents. In some cases of more workload, inbound agents may perform outbound telemarketing, but never the opposite. Furthermore, outbound agents usually do not stay for more than a year. In fact, their training reflects this situation, with highly focused and intensive sessions up to two weeks before beginning call handling, and smaller session for each new campaign, to understand the product or service being offered. On the other hand, inbound agents only receive briefing sessions and electronic documentation for a new distinct product. Also quality control is much more emphasized for outbound agents. It is interesting to note short-term outbound agents as opposed to mid or long-term inbound agents. In fact, outbound activities are pre-planned, more focused and can be better structured whereas inbound activities are likely to be less structured since the initiative does not come from the bank but rather from consumer. Hence, inbound interaction may demand more experience and breadth from the employee. Outbound looks more hard-sell, inbound more soft-sell.

The other big difference is customer entanglement versus intrusiveness. The manager states that customers are keener to listen the agent when they begin the interaction. The fact that it was the customer who first established the call is enough to make a difference, since if she/he is calling, it is because she/he has some available slot of time for conversation, while an outbound call may come in an inconvenient time. Then, there is the procedure for this particular contact-center, where customer requests are handled first and only then the agent tries to offer the product/service. Therefore, after the client is satisfied, it is more likely to be predisposed for listening the agent who happens to have just met her/his request.

Last but not least, there is a sense of insecurity in making banking transactions through phone. This is much more emphasized than a homebanking web site, where the client is not interacting with a human being, on the opposite of a phone contact, aggravated by the fact the client is not physically seeing the agent. Therefore, customers who are used to interact with the bank via the phone communication channel (thus perform phone calls) are much more prone to subscribing products/services using this channel. In fact, intrusiveness associated with outbound communication and specifically phone calls have been widely studied and regulated by legislation through opt-out registration (Woodside, 2005). Thus it is no surprise the conclusions of the telemarketing expert. However, the focus of this characteristic in comparison with the remaining features differs from the feature relevance determined through model analysis. Thus, a discussion about the second most relevant feature for the manager also should take place.

While it can be argued that the gain the customer achieves through the offered rate in comparison with the competition is a valuable driver toward deposit subscription, there is not such a remarkable difference such as for the type of phone call. The expert manager states that dialog script configured per campaign usually is different for each deposit, being adapted to its characteristics. This difference could suggest that mining a different model per product would improve targeting subscription performance. In fact, other reason may favor the split using the offered rate: the expert mentioned that eventually a DM model would optimally use different features associated with the deposit specification (e.g., a deposit with an offered rate higher than the competition may demand a high minimum amount of subscription, thus being more related with the client income). Nevertheless, the fact that interest rate is a numeric value (when compared with the binary direction call) also poses the difficulty of where the split should occur, or even how many splits.

In the end, the expert clearly indicated the inbound versus outbound as the most relevant split, although emphasizing that other divisions could occur considering the offered rate. Taking into account the size of the dataset analyzed, the inbound telemarketing sub-problem consisted in only 1915 instances, as opposed to the total 52944 contacts. This number suggests previous research was optimized for outbound telemarketing. Given such potential to substantially improve inbound telemarketing predictive results, in this work we adopt the call direction split, with a particular focus on inbound calls. In the next section, the sub-problem of inbound telemarketing is characterized by the telemarketing manager and re-evaluated. As a last remark, it may be argued that the knowledge emerging from this discussion solely based on human experience can hardly be inferred by the state of the art in machine learning. Simply, there are too many context factors to account for in each problem. Results explored in Section 4.2 attest the benefits of the proposed approach.

4.2. Modeling

In this section, modeling of the new problem takes place, first for tuning through feature selection, then by comparing the results from baseline to the newly defined model. Previous section suggested splitting the problem between inbound and outbound telemarketing, following advice from expert, also supported by the literature. As stated previously, the expert analyzed the new sub-problem of inbound telemarketing and its characteristics and selected a new list of features for adding value to the baseline 27 features (Table 3). It should be noted that two of the initially proposed features were discarded (call.dir and last.prod.result), since these attributes were constant for the small subset of just 1915 inbound contacts (but not for the whole 52944 records). Thus, further reference in this text mentions just 25 features to those used in the baseline study. Table 4 shows the aggregated results by the average for the twenty executions for each feature selection phase.

Table 3: List of features selected by a domain expert and that are expected to influence inbound call responses.

Feature	Description
phone.type	landline phone or mobile device from which the client called
inb.calls.n	number of inbound calls in which the agent received a system alert informing that the client has been selected for the current product campaign
phone.y.usage	phone banking yearly usage
internet.y.usage	Internet banking yearly usage
curr.account	client current account balance
term.accounts	client total term accounts balance, if available
ebank	subscriber of e-banking services (internet or phone)
prev.nr.outb	previous outbound attempts within the same campaign
inb.call.since	days passed since the last inbound contact for the same campaign

Recalling that a closer value of AUC and ALIFT to 1.0 represents a more accurate model, phase (iia), with the originally proposed features, achieved an AUC of 0.8818, while modeling using just the new features (phase (iib)) got an AUC of 0.7940. Nevertheless, the new features were an addition to the original characteristics, thus, an overall better result by merging all the features was expected, which was achieved through a model with an AUC of 0.9069 (phase (iii)), confirming the value of the manager's choices. To validate statistical significance, a Mann-Whitney non-parametric test was executed to check significance at the 95% confidence level (Molodianovitch et al., 2006).

Table 4: Feature selection results (best values in **bold**).

Phase	AUC	ALIFT	Features included
(iia) (25 features)	0.8818	0.8424	the 27 baseline features, except for call.dir and last.prod.result, which remained constant
(iib) (9 features)	0.7940	0.7649	features described in Table 3
(iii) (34 features)	**0.9069***	**0.8669***	both features included in (iia) and in Table 3

$\star$ - Statistically significant under a pairwise comparison with (iia) and (iib).

For the feature reduction procedure (Table 5), a first test was conducted for a reduction from the 34 features from phase (iii) to the most relevant twenty features, where these features were selected using DSA method applied to the model with the 34 features. Such a reduction resulted in an improvement in the AUC from 0.9069 to 0.9137.

Next, the choice was to perform two more steps by removing the five lesser relevant features in each iteration. The result is an achievement of an AUC of 0.9139 with fifteen features, and of 0.9247 with just ten features. These metrics confirm that the discarded features were less relevant, leading to a model with better predictive performances. This behavior often occurs when using feature selection methods (Guyon & Elisseeff, 2003).

Table 6 shows the results for the realistic simulation where the 90% older contacts were used for modeling and the newer 10% for testing the prediction accuracy, using the best set of 25 features proposed by Moro et al. (2015b) (original model, optimized for all types of contacts) and the ten new

Table 5: Feature reduction results (best values in **bold**).

Phase	AUC	ALIFT	Features included
(iva) (20 features)	0.9137	0.8730	ag.generic, suited.rate, phone.y.usage, inb.call.since, dif.best.rate.avg, ag.created, prev.nr.outb, call.nr.schedules, nr.employed, recency, call.prev.durations, cons.price.idx, cons.conf.idx, internet.y.usage, call.month, inb.calls.n, curr.account, cli.affluent, ebank, prev.contacts.durations
(ivb) (15 features)	0.9139	0.8714	suited.rate, dif.best.rate.avg, nr.employed, call.prev.durations, phone.y.usage, inb.call.since, cons.price.idx, call.month, internet.y.usage, ag.created, call.nr.schedules, prev.nr.outb, prev.contacts.durations, inb.calls.n, cons.conf.idx
(ivc) (10 features)	**0.9247***	**0.8802***	dif.best.rate.avg, ag.created, phone.y.usage, prev.nr.outb, nr.employed, suited.rate, call.nr.schedules, call.prev.durations, inb.call.since, cons.price.idx

$\star$ - Statistically significant under a pairwise comparison with (iva) and (ivb).

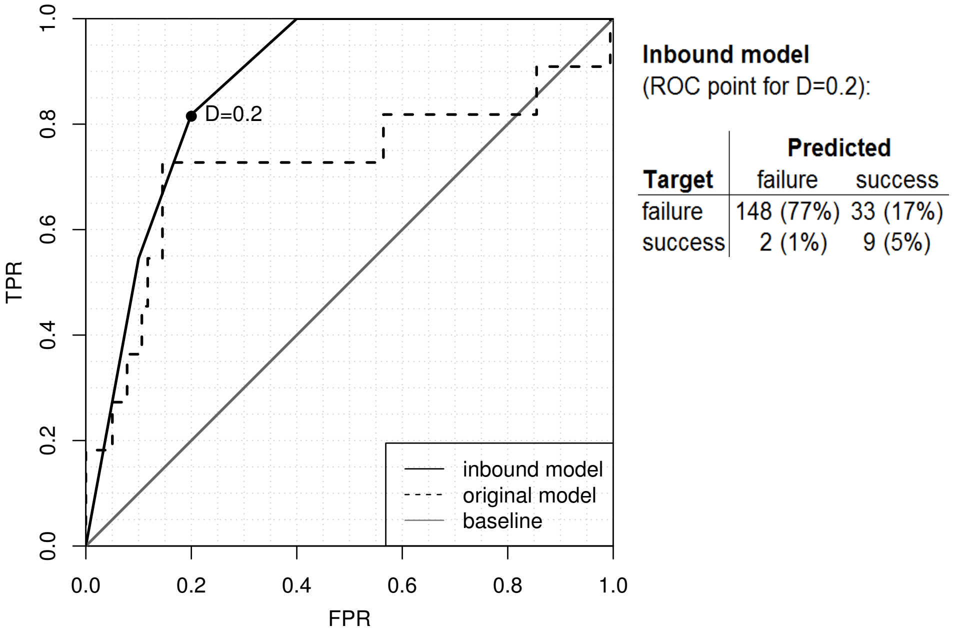
features discovered and presented in Table 5 (inbound optimized model, as set in phase (ivc)). The difference in the results attests the effectiveness of the suggested procedure. The inbound optimized model achieves the best AUC and ALIFT values, presenting an AUC of 0.89 and ALIFT of 0.87 that corresponds respectively to a 14 and 13 percentage point difference when compared with the original model (with 25 features).

Table 6: Realistic simulation test set results (best values in **bold**).

Model	AUC	ALIFT	Features included
original	0.7481	0.7341	25 features, described by Moro et al. (2015b)
inbound optimized	**0.8875***	**0.8641***	10 features, described in Table 5 (step (ivc))

$\star$ - Statistically significant under a pairwise comparison with the original model.

The curves plotted in Figure 2 for both the original model based on the 27 features previously discovered and the newly found features for the optimized inbound telemarketing model show the difference of both models, confirming the values shown on Table 6. Also the confusion matrix displayed for the predictions of the 10% of contacts used for validation exhibits the high level of accuracy, with a true positive rate of TPR=82% and true negative rate of TNR=82%, when considering a success if the model foresees more than 20% of probability of occurring a subscription. We should note that



Inbound model
(ROC point for D=0.2):

	Predicted	
Target	failure	success
failure	148 (77%)	33 (17%)
success	2 (1%)	9 (5%)

Figure 2: ROC curves for the inbound and original models (left) and example confusion matrix for $D = 0.2$ (right)

the D=20% threshold used to build the example confusion matrix (right of Figure 2) was chosen taking into account that this particular bank prefers to avoid losing business opportunities, which is translated into successful subscriptions. Hence, we opted for a low threshold and more sensitive example, above which the client is contacted.

Figure 3 shows the Lift cumulative curves for the simulation test, when using the 10% newest contacts. The plot visually confirms the large difference between both models in terms of prediction results. Table 7 exemplifies the portion of successful contacts reached from the total of successes for some sample sizes of the Lift curve. These results confirm that for the largest portion of the sample size range ($\in [0.07, 1.0]$), the inbound optimized model obtains the highest cumulative Lift (i.e., number of client responses). In particular, it should be stressed that when half of the most likely buyers are selected, the inbound model achieves the perfect performance, reaching all subscribers. The difference is large (27.2 percentage points) when compared with the original model. In effect, the original model needs to select a much larger sample (with 80% of the most likely buyers) to achieve the same level of performance.

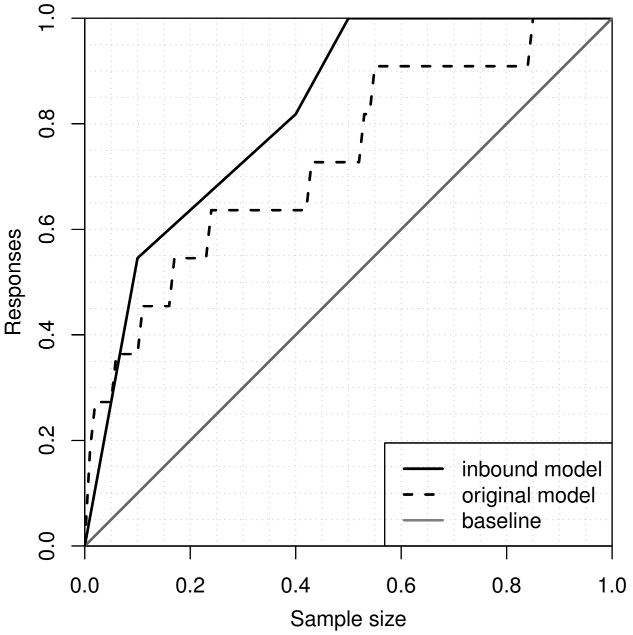


Figure 3: Lift cumulative curves for the original and inbound optimized models

Table 7: Comparison of results: Lift cumulative samples.

Sample size	Original model	Inbound model
10%	36.4%	54.5%
20%	54.5%	63.6%
30%	63.6%	72.7%
40%	63.6%	81.8%
50%	72.7%	100.0%
60%	90.9%	100.0%

To detail the relevance of the ten features used for defining the inbound optimized model, the DSA method was executed, which uses samples from the training data to analyze variations in the desired outcome. The higher the variation in the predicted response, the more relevant is the analyzed input feature. The respective DSA feature ranking is displayed in Figure 4.

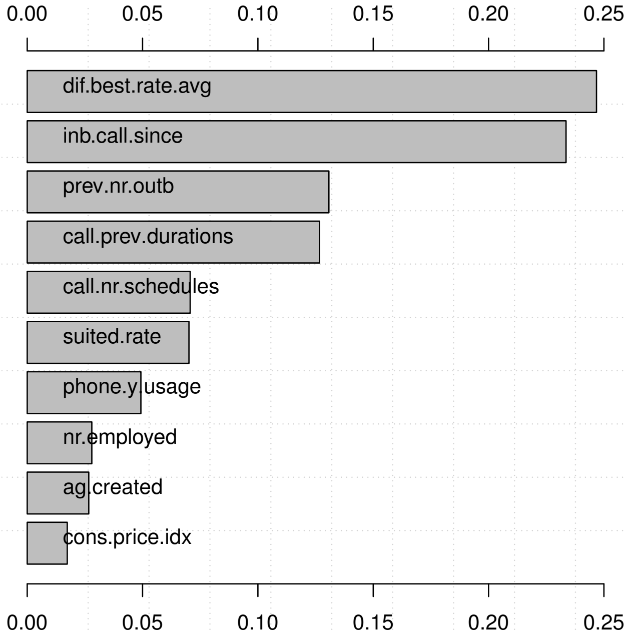


Figure 4: Feature importance bar plot of the inbound optimized model

The difference between the best offered rate and the national average is the most relevant feature for the model implemented. This is a confirmation of the results achieved by Moro et al. (2015b), which also stated this feature as the most relevant. Interestingly, the second and third most relevant features are specifically inbound and correspond to newly proposed features. The former represents the past time period since any previous call for the same campaign, while the later stands for the previous outbound attempts within the same campaign. The relevance of the second feature is just slightly below the top feature, and above 20%, emphasizing the influence for modeling inbound calls. The third feature has an influence of around 13%. The inclusion of these two new features in the top three is a confirmation that inbound telemarketing should deserve a specific attention, different from outbound telemarketing, proofing the value of the proposed divide and conquer strategy.

Considering the second and third most relevant and newly proposed features, we proceeded by using the sensitivity analysis DSA method to measure the global effect of the feature in the output response. Figure 5 presents the respective Variable Effect Characteristic (VEC) curve, which plots the input feature range of values (x-axis) versus the average sensitive model responses (y-axis).

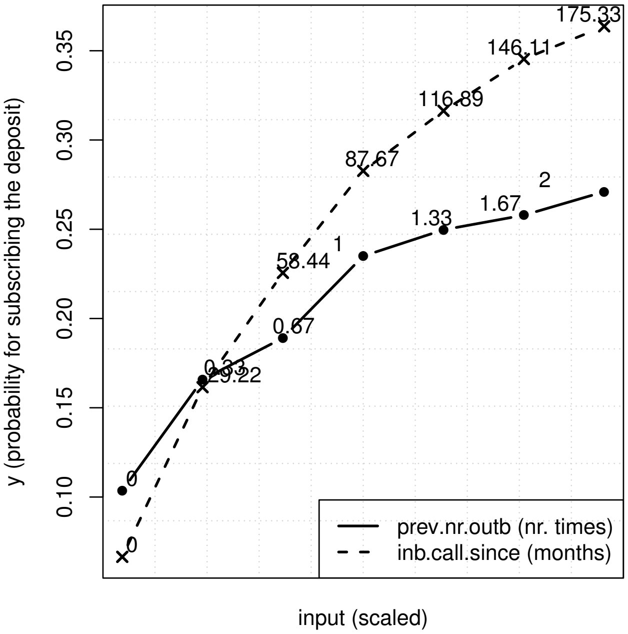


Figure 5: VEC curves showing the influence of the two most relevant inbound features in the predictive model

From the plot of Figure 5, it can be observed a direct influence of an increase in the number of days passed from another inbound contact for the same campaign with the likelihood of success. It should be noted that some of the inbound campaigns are resident, i.e., permanent since the corresponding product started to be sold through phone calls. This fact justifies the larger numbers in the curve, such as 175 months (approximately 14 years). Furthermore, each deposit has an expiration period, after which the value invested and interests earned are deposited in the main current account, meaning that the client may be contacted again after this date. Several reasons may justify the relationship found. One of the more plausible is the fact that the client may be bothered if few time passed since the last time he was asked previously to subscribe the same deposit. In fact, while the intrusiveness is more associated with outbound phone calls, inbound communication may also be affected by the customer perceived intrusiveness (Galloway & Brown, 2004; Jung, 2009)

The third most relevant feature is the number of previous outbound calls, with a higher value influencing positively the likelihood of success. This is not a surprising relationship. In fact, it often occurs that the client is contacted directly through an outbound call and asked to be contacted later, but sometimes it is the client himself that takes the initiative of contacting the bank to finally subscribe the deposit, resulting in a successful contact. However, the studied dataset does not provide information about which contacts are made by the clients specifically for subscribing the deposit.

5. Conclusions

In this paper, a divide and conquer strategy approach was presented, using a procedure based on feature relevance computed by a data mining (DM) model followed by an expert analysis in order to find the best feature for splitting the problem in smaller more manageable sub-problems. It is argued that human expertise remains the best strategy for dealing with real world problems, which are encompassed by complex contexts, with a large number of variables to account for.

This methodology was applied to a bank telemarketing problem consisting in campaigns conducted through phone calls for selling long-term deposits. This specific problem was addressed in previous studies, which resulted in a DM model defined by 27 features, from which the five most relevant were: the difference between best rate offered and the national average, the daily three month Euribor rate, the last campaign result for the same product in which the client was contacted, the number of times the client subscribed the deposit previously and the call direction (inbound/outbound). The expert evaluated these five features and assessed, based on his experience, which was the best candidate for splitting the problem. As a result, the call direction was considered the most suitable based mainly on three reasons: difference between inbound and outbound agents' profile, intrusiveness associated with outbound, and insecurity associated with transactions based on phone calls (clients used to inbound calls are keener to trust in this communication channel).

Considering most of the contacts in the dataset are from outbound calls (51029), the hypothesis arose is that inbound calls have been neglected in its characterization features. To test it, the 1915 inbound contacts were re-evaluated as a sub-problem from the original problem. First, the human expert helped in selecting additional features to increase value to the previously suggested 27. Then, a feature reduction procedure took place for improving the model by eliminating the least relevant features, leading to just ten.

The data-driven model was built based on a neural network ensemble model and the performance was measured using two criteria: the area under the Receiver Operating Characteristic curve (AUC) and the area under the Lift cumulative curve (ALIFT). As a baseline comparison, a model fed with 25 features that were previously identified as relevant was defined, but where there was no distinction between outbound and inbound contacts, with the vast majority of the calls being executed through outbound. Several experiments were held, using two validation schemes: a random holdout split, using 2/3 of the data for training and 1/3 for testing; and a more realistic time ordered holdout, where 10% of the newest contacts were used for the predictive performance comparison.

The obtained results confirm the inbound optimized model as the best solution, outperforming the baseline model by a large difference. For instance, for the realistic time ordered validation, the AUC increased from 0.75 for baseline to 0.89 for the inbound specific model (difference of 14 percentage points). Moreover, the inbound optimized model achieves the ideal Lift performance (i.e., reaching all potential buyers) when selecting only half of the most likely buyers, while the baseline model needs to select a much larger sample (80% of the clients) to reach the same performance.

These results clearly confirm the approach proposed for dividing a problem in a smaller sub-problem which is characterized by specific features, different than the ones which best represent the more global problem. Thus, using a "divide and conquer" approach, in which the inbound contacts are modeled with a distinct predictive model, leads to an added value for the studied telemarketing bank domain. Moreover, a sensitivity analysis was executed over the best predictive model, ranking two inbound specific features (i.e., number of days passed from another inbound contact for the same campaign and number of previous outbound calls), as the second and third most relevant features. Such knowledge is useful for bank campaign managers.

In the future, the authors intend to apply a similar divide and conquer to other problems in the banking context. Considering the specific telemarketing problem, it would be also interesting to validate if the proposed inbound features using data from other banks or countries attest the results achieved.

Acknowledgments

The authors would like to thank the valuable effort of the bank professional who contributed with his expertise for assessing feature relevance, selecting how the problem should be split, and providing assistance in the analysis of the sub-problem of inbound telemarketing.

References

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *Computational Intelligence Magazine, IEEE*, *5*, 13–18.

Chen, C.-M., Lee, H.-M., & Chang, Y.-J. (2009). Two novel feature selection approaches for web page classification. *Expert Systems with Applications*, *36*, 260–272.

Cortez, P. (2010). Data mining with neural networks and support vector machines using the r/rminer tool. In *Advances in Data Mining. Applications and Theoretical Aspects* (pp. 572–583). Springer.

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

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*, 927–940.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*, 861–874.

Fritsch, J., & Finke, M. (2012). Applying divide and conquer to large scale pattern recognition tasks. In *Neural Networks: Tricks of the Trade* (pp. 311–338). Springer.

Galloway, C. S., & Brown, S. P. (2004). Annoying, intrusive,... and constitutional: telemarketing and the national do-not-call registry. *Journal of Consumer Marketing*, *21*, 27–38.

Guido, G., Prete, M. I., Miraglia, S., & De Mare, I. (2011). Targeting direct marketing campaigns by neural networks. *Journal of Marketing Management*, *27*, 992–1006.

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

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* volume 2. Springer.

Haykin, S. S., Haykin, S. S., Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines* volume 3. Pearson Education Upper Saddle River.

Hsieh, C.-J., Si, S., & Dhillon, I. S. (2014). A divide-and-conquer solver for kernel support vector machines. In *Proceedings of the 31st International Conference on Machine Learning* (pp. 566–574). International Machine Learning Society.

Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications*, *27*, 623–633.

Javaheri, S. H., Sepehri, M. M., & Teimourpour, B. (2014). Response modeling in direct marketing: A data mining based approach for target selection. In Y. Zhao, & Y. Cen (Eds.), *Data Mining Applications with R* chapter 6. (pp. 153–178). Elsevier.

Jung, J. H. (2009). *Factors influencing consumer acceptance of mobile advertising*. Ph.D. thesis.

Kim, Y., & Street, W. N. (2004). An intelligent system for customer targeting: a data mining approach. *Decision Support Systems*, *37*, 215–228.

Kotler, P., & Keller, K. (2012). *A Framework for Marketing Management*. Painos. Edward Brother.

Levine, D. S., & Leven, S. J. (2014). *Motivation, emotion, and goal direction in neural networks*. Psychology Press.

Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications*, *39*, 424–430.

Lin, Y.-H., Chen, C.-Y., Hong, W.-H., & Lin, Y.-C. (2010). Perceived job stress and health complaints at a bank call center: comparison between inbound and outbound services. *Industrial health*, *48*, 349–356.

Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems*, *51*, 782–793.

Mehrotra, V., & Fama, J. (2003). Call center simulations: call center simulation modeling: methods, challenges, and opportunities. In *Proceedings of the 35th conference on Winter simulation: driving innovation* (pp. 135–143). Winter Simulation Conference.

Molodianovitch, K., Faraggi, D., & Reiser, B. (2006). Comparing the areas under two correlated roc curves: Parametric and non-parametric approaches. *Biometrical Journal*, *48*, 745–757.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31.

Moro, S., Cortez, P., & Rita, P. (2015a). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation. *Expert Systems with Applications*, *42*, 1314–1324.

Moro, S., Cortez, P., & Rita, P. (2015b). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, *26*, 131–139.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*, 2592–2602.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*, 71–72.

Stach, W., Kurgan, L., & Pedrycz, W. (2010). A divide and conquer method for learning large fuzzy cognitive maps. *Fuzzy Sets and Systems*, *161*, 2515–2532.

Talla Nobibon, F., Leus, R., & Spieksma, F. C. (2011). Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms. *European journal of operational research*, *210*, 670–683.

Tapp, A. (2008). *Principles of direct and database marketing*. Pearson Education.

Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems, 9th Edition*. Pearson.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*, 223–230.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Woodside, A. G. (2005). Consumer responses to interactive advertising campaigns coupling short-message-service direct marketing and tv commercials. *Journal of Advertising Research*, *45*, 382–401.

Chapter III - Conclusions

This final chapter outlines the main conclusions from this work. It is divided in three sections, with the first and lengthiest focusing on discussing and producing a common flow for the articles published.

First, the proposed approach for analyzing literature is discussed. While it constitutes a different branch from the remaining research, it provided the need for addressing the problem chosen for the empirical analysis. Therefore, the methods suggested for analyzing literature constitute by its own right an interesting subject for discussion. Furthermore, the proposed methodology can be applied to virtually any research area, providing a different and innovative means for evaluating and answering to the typical research questions of “where are we now?” and “where should we go from here?” which emerge on the beginning of every new research project. By answering to these questions for the addressed problem of selling long-term deposits through telemarketing campaigns, it laid the foundations and sparkles needed for pursuing novel feature selection strategies.

The majority of the discussion is about the suggested framework for conducting a flow of feature selection strategies that ultimately aim to improve data mining models built on a better characterized dataset of events. All the results used are collected from the three later articles, DSS, NCAA and INPREP, for providing solid ground proof on the suggested strategies. Then the framework is presented in a flowchart like figure and a discussion takes place on the procedures and complementarities of strategies and why they should be combined in the suggested framework. Next, follows a critical analysis of the combined results achieved for our instantiation of the problem of bank telemarketing, with a highlight on the most relevant findings for the case analyzed.

In section 3.2, the concluding remarks are drawn, focusing mainly on looking at each of the components that constitute our proposed framework and concluding on how all of them can work combined to produce a global feature selection schema for addressing data-driven problems to discover knowledge for better decision support.

Finally, possible future work is presented in section 3.3. On this subject, three different directions are proposed: the first is on testing the proposed framework on other different problems; the second points at designing an adaptive intelligent decision support system for predicting the likely outcome of every contact under a campaign context, benefiting from the strategies proposed on this work, while evaluating business impact; finally, the third tries to cover a wide spectrum of problems by finding different strategies for filling a really large informational gap on features for other situations and contexts.

3.1 Summary and discussion

The first goal of the literature analysis was to provide insights on the relevance of business intelligence applications to banking and ultimately find research gaps for targeting customers with bank products. In fact, the conclusions of the ESWA review detail that none of the articles analyzed focus specifically on targeting for selling deposits.

Nevertheless, the relevance of the ESWA article goes far beyond the findings on business intelligence applications to the banking domain. Indeed, the usage of an automated approach for analyzing articles allowed encompassing a wider number of articles, reaching to a larger universe of studies on the topics covered. Furthermore, several combined contributions were presented in the procedures suggested:

- employment of expert knowledge for defining the dictionaries that intersected both domains reviewed, business intelligence and banking;
- application of the latent Dirichlet allocation in literature analysis by defining topics grouping similar articles in the terms addressed, when considering the dictionaries used;
- analysis in a timeline of the articles published through the twelve years of the time frame considered;
- validation of the findings through the selection and in-depth manual analysis of the most representative article for each topic, providing an interesting although quite simple method of confirming the trends suggested by the topics found.

The procedures applied for this semi-automated literature review can be useful for virtually any literature analysis using electronic text documents. Considering the wide spectrum of focused in-depth research conducted world-wide and the initial requirements that each research project faces of finding what has been done so far on the domains addressed, one might have a glimpse on the bright future of automated approaches, for which our work has contributed with innovative procedures.

As stated previously, the three strategies for feature selection are the subject of each of the remainder three articles. Nevertheless, the first of these, the DSS article, and the corresponding experiments and findings, needed to extend the scope beyond suggesting and validating a feature selection strategy. Such demand emerged for validating the experimental procedures suggested. Thus, the DSS article is more extensiveness in several aspects related to validation. First, several data mining algorithms were tested and compared, all implemented in the **rminer** package (Cortez, 2010), including logistic regression, decision tree, support vector machines and neural networks. The results provided solid ground proof that for our case and specific problem instantiation, the ensemble of neural networks implemented in **rminer** provided the best results in terms of predictive performance. Afterwards, the extracted feature relevance with the data based sensitivity analysis was evaluated through the extraction of rules from the model with a decision tree. The comparison of both cleared that the most relevant features as determined by the data based sensitivity analysis appear in the top branches of the extracted decision tree. Those extensive validations, while time consuming, provided solid ground for using the same procedures in the following pursuits for feature selection strategies. Hence, the two later articles, NCAA and INPREP, focused only on the evaluation of the strategies proposed, using the validations of the DSS article for justifying the procedures applied.

The three feature selection strategies are complementary. We followed what seemed to be a logical order for the three approaches. Hence, we searched first for social and economic context features which provided additional relevant value to the dataset, and then we analyzed historical behavior, taking into account each event is a contact related to a specific client and that a significant portion of clients were already previously contacted (27.3% for the July 2012 to June 2013 timeframe, used for validation purposes), with that information being available on the dataset used. Nevertheless, considering the history features are client related, it can be argued that both the DSS and NCAA article approaches could be taken in any order, with the common goal of enriching contact characterization. Thus, both strategies are completely independent. Furthermore, the improvements we achieved with the history features over the already enriched dataset with context features provide the evidences that the benefits of the first feature selection strategy does not hide the gains of the second strategy. However, the last strategy requires that previous in-depth characterization of the problem is exhausted as much as possible, for this strategy focus in splitting the problem in smaller and more manageable fractions constituting specific sub-problems, for which a new full assessment of features may benefit from previous feature selection strategies and the enriched characterization of the main problem.

The resulting framework proposed, which compiles the three strategies, is shown on Figure 1. It should be stressed that such framework is not just an indistinct collection of strategies; it suggests a research flow based on the experiments reported in the three articles, DSS, NCAA and INPREP, as analyzed in the previous paragraph. This framework starts on a typical baseline for any data-driven research using models constructed through data mining techniques applied to a dataset of events characterized by a list of available features. The main goal is enriching problem characterization through the best possible feature selection approaches, incrementing iteratively the value of the dataset by providing more knowledge on portraying each of the events available in the dataset. As stated previously, although the work described on the NCAA article uses as a baseline the results achieved and reported in the DSS article, both the strategies proposed can run in parallel, contributing for a given number of highly relevant features, which were assessed through mixed procedures using both automated and human expert approaches. Finally, the most relevant features are evaluated by the expert for understanding if the problem can be distinctively split in smaller and more manageable problems. The list of features is again assessed in their relevance for tuning the final model which addresses only the sub-problem events.

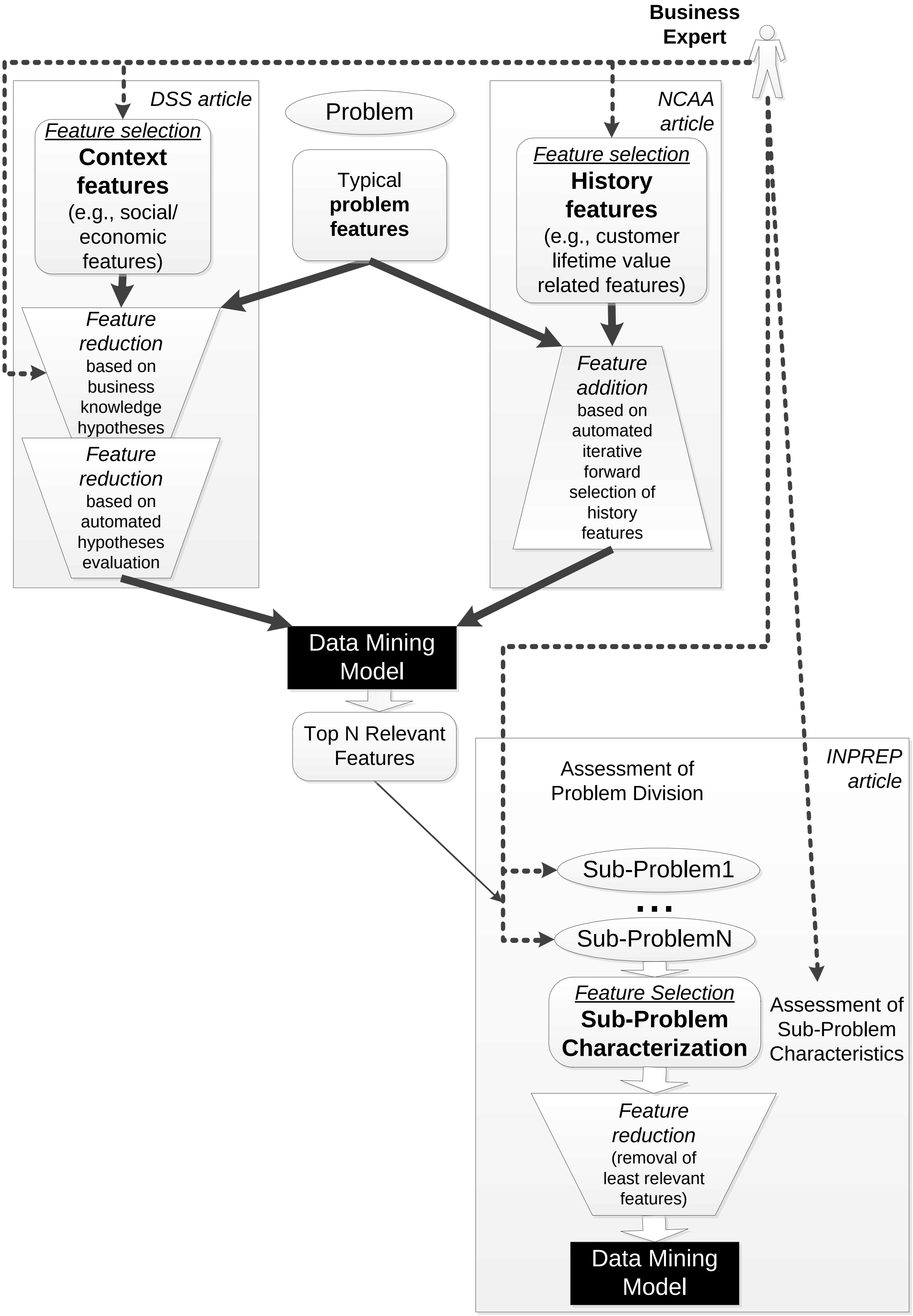

Business Expert
DSS article
Problem
NCAA article
Feature selection
Context features
(e.g., social/ economic features)
Typical problem features
Feature selection
History features
(e.g., customer lifetime value related features)
Feature reduction based on business knowledge hypotheses
Feature reduction based on automated hypotheses evaluation
Feature addition based on automated iterative forward selection of history features
Data Mining Model
Top N Relevant Features
INPREP article
Assessment of Problem Division
Sub-Problem1
...
Sub-ProblemN
Feature Selection
Sub-Problem Characterization
Assessment of Sub-Problem Characteristics
Feature reduction (removal of least relevant features)
Data Mining Model

Figure 1 – Framework for feature selection strategies.

The design displayed in Figure 1 emphasizes the similarities in the procedures undertaken between each of the individual approaches. The most distinguishable similarity is that every of the three strategies use business expert knowledge. As discussed in the articles and in section 1.2.4, the current state of the art in feature selection techniques is still no match for the insights provided by a human domain expert. Moreover, in the DSS article, we conducted an experiment for comparing three modeling strategies: with all the initial features, using an automated forward selection (Guyon & Elisseeff, 2003) and with the proposed feature strategy suggested in that article. The obtained results (see Table 4 from the same article) clearly benefit our two step feature selection strategy in terms of predictive performance. Human knowledge is often neglected in researches pursuing for improved automated feature selection procedures. Therefore, our framework tries to benefit from the best of both worlds: the automated approach and the insight knowledge from a human expert. In effect, all three strategies benefited from the combination of both these procedures. Another resemblance between the three strategies is the usage of the data based sensitivity analysis for assessing feature relevance, thus supporting automated feature selection procedures, whether it is for feature reduction, by removing least relevant features or for including new features in case of feature addition.

Considering the goal of iteratively improving predictive performance, the proposed strategies would necessarily have to result in better values that directly reflected the enhancements. In fact, Table 1, which summarizes performance metrics results, shows the clear impact benefiting the predictions. Such metrics are described in detail in each of the articles. It should be noticed that these metrics are the results for the simulation in the July 2012 to June 2013 time period, encompassing completely unforeseen data, thus strengthening the validation of the achievements.

Table 1 – Predictive performance improvement results.

	Strategies suggested in the articles		
Metrics	DSS	NCAA	INPREP
AUC (Area under the receiver operating characteristic curve)	79.4%	86.1%	88.8%
ALIFT (Area under the cumulative Lift curve)	67.2%	70.4%	86.4%
Percentage of successes reached by a selection of the 50% most likely subscribers (obtained through cumulative Lift analysis)	78.9%	82.8%	100.0%

While Table 1 shows a steady increase in performance, each of the strategies also unveiled the value of the newly proposed features. Table 2 lists these features, with the DSS article providing four social and economic features ranked in the top ten most relevant for the model, while the NCAA contributed with two customer lifetime value features, and the INPREP with three directly related with the sub-problem of inbound telemarketing.

Table 2 – Novel features proposed from the top ten in each article.

Article / Strategy	**Feature**	**Rank in terms of relevance**	**Description**
DSS	euribor3m	1	Daily three month Euribor rate.
	nr.employed	6	Quarterly average of the total number of employed citizens.
	emp.var.rate	8	Employment variation rate, with a quarterly frequency.
	nat.avg.rate	10	National monthly average of deposit interest rate.
NCAA	last.result.prod	3	Last campaign result in which the client was contacted where the same product was being sold.
	frequency	4	Number of times the client subscribed the deposit previously.
INPREP	inb.call.since	2	Days passed since the last inbound contact for the same campaign.
	prev.nr.outb	3	Previous outbound attempts within the same campaign.
	phone.y.usage	7	Phone banking yearly usage.

3.2 Final remarks

This study presents a novel framework for providing guidance on data-driven knowledge discovery projects through feature selection strategies. The suggested methodology ranges from finding new data for adding value to the database, computing values based on existing data and a divide and conquer strategy. Nevertheless, the common baseline for those methods consists in enriching the database, considering data are the key element for any data-driven research.

For validating the proposed framework, three empirical analyses were conducted for improving telemarketing campaigns to sell bank deposits, using data from campaign contacts occurred from 2008 to 2013 from a Portuguese bank. The relevance of this problem was assessed by applying a novel literature analysis using text mining and the latent Dirichlet allocation.

The findings show an alignment between each step and the consistent improvement of results, proving the value of such framework. First, the inclusion of context features, namely social and economic related features resulted in a system which could reach around 79% of successes by selecting the half most promising clients selected by the system model. The euribor rate, one of the proposed features, was evaluated through a data based sensitivity analysis as the most relevant feature.

Next, the feature enrichment proceeded with past history analysis by evaluating customer lifetime value (LTV) characteristics, culminating with an improvement of around 4%, i.e., the model could now reach almost 83% of subscribers by contacting the half best positioned clients. From the newly added LTV features, two of them were among the top five most relevant features, the result from previous campaign for the same product, and the number of previous successful contacts.

Finally, the divide and conquer approach based on expert knowledge and feature relevance split the problem in outbound and inbound telemarketing. While most of the dataset contacts were from outbound calls (96% of them), the newly problem of inbound telemarketing was re-evaluated, resulting in that the half better classified clients enclosed the total successful contacts, whereas previous generic model (with social and economic and LTV features) could only reach 72.7% of successes with the same fraction of inbound calls. In fact, the difference to the 83% of successes reached with previous model of NCAA supports that inbound contacts were being neglected by the mere fact they were only around 4% of the total contacts, justifying a different approach. One of the newly proposed inbound characteristics, the number of days passed since the last inbound contact for the same campaign, was considered the second most relevant feature for the model. We also note that the obtained predictive models (and selected features) were considered credible by two consulted bank experts. And the obtained predictive results, with high quality AUC and LIFT values, confirm the predictive models as valuable for telemarketing campaign managers.

Considering this is a multi-disciplinary problem addressed, it could be more difficult to publish in journals of any of the involved areas (expert systems, marketing, banking), since the gains and discoveries are divided through all those areas. Nevertheless, it is interesting to notice that we have achieved successfully to publish three articles, with a fourth submitted. This is a reliable indicator of the quality of the work produced, evidencing the relevance of our discoveries. In particular, we would like to stress that our feature selection strategies can potentially benefit any real-world classification problem. Moreover, the framework drawn on such strategies and supported by solid ground experiments certified by the articles published is based on concepts which are common to a wide spectrum of real-world problems, such as context environment, past history and division in more reduced and manageable sub-problems.

3.3 Future work

Several ideas have arisen from the present work that may lead to future research. One of the most obvious is addressing other problems with the proposed framework of combined domain knowledge and automated strategies. Given the fact that bank telemarketing is a very specific problem, from a peculiar industry, context that can hardly be transposed for another industry. Such a problem domain and industry change may present new challenges and possible need of adaptations. Nevertheless, if the new problem also verses on targeting, we envision that the NCAA approach may also be directly applied as long as the available data contains several contacts to the same customers. However, by changing the industry for applying the DSS approach, it would be needed to analyze the specific surrounding environment of that industry and the problem being addressed, which will necessarily result in a different set of features, although the general procedures may be similar. The same happens for the divide and conquer proposal of the INPREP strategy: the analysis for defining the split will lead to specific sub-problems, which may or not result in more tuned models.

The case study and the corresponding dataset used as a test bed for all the experiments may also serve other kind of research, also data-driven, but more focused on designing and implementing a real adaptive business intelligence and intelligent decision support system, which can automatically tune itself according to changes in the environment.. Besides the prediction if the client would subscribe or not the deposit, there are other two outcome variables that would be interesting on a more in-depth analysis: the call duration, and the value subscribed. While the call duration may have tight relation links with the call result (Moro et al., 2011), it may also conceal interesting knowledge to explore. However, it is the prediction of the value subscribed which should deserve the most attention. Although a value only exists if the result is a success, the relevance of which value is of paramount importance: for a real measurement of the campaign effectiveness, it is needed to take the value invested versus the value retained into action, for which the prediction of the amount subscribed plays a key role. The important thing to retain is that the case study chosen verses a real up-to-date problem, encompassing an interesting timeframe, which may lead to a wide variety of research goals, being far from exhausted within the current research.

Similarly to the importance and focus that Domingos (2012) puts on feature engineering, we also believe that the pursuit for feature selection strategies that can enhance the knowledge value concealed in a dataset is one of the most powerful tools for improving data mining results. In fact, there is no use for an advanced machine learning algorithm if it has no data from where it can learn. It is our belief that there is a huge informational gap in a large range of real-world which may lead to new feature selection strategies. This work contributed in reducing such gap in a very specific problem domain. As such, there is large room for more in depth research, in terms of: charactering classes of problems that might benefit from a hybrid domain knowledge automated feature selection; and also, proposing newer (and possible more generalizable) hybrid feature design and selection strategies.

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3), 392-398.

Abraham, M. M., & Lodish, L. M. (1993). An implemented system for improving promotion productivity using store scanner data. Marketing Science, 12(3), 248-269.

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Systems with Applications, *37*(12), 7913-7921.

Arnott, D. & Pervan, G. (2008). Eight key issues for the decision support systems discipline. Decision Support Systems, 44(3), 657-672.

Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: the rise of design science. Journal of Information Technology, 29(4), 269-293.

Baets, W. R. (1996). Some empirical evidence on IS strategy alignment in banking. Information & Management, 30, 155–177.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. Neural Computing and Applications, 19(8), 1165-1195.

Bar-Ilan, J. (2012). Journal report card. Scientometrics, 92(2), 249-260.

Barber, S., Boyen, X., Shi, E., & Uzun, E. (2012). Bitter to better - how to make bitcoin a better currency. Lecture Notes in Computer Science, 7397, pp. 399-414. Springer.

Berry, M., & Linoff, G. (1999). Mastering data mining: The art and science of customer relationship management. John Wiley & Sons, Inc..

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Bornmann, L., Marx, W., Gasparyan, A. Y., & Kitas, G. D. (2012). Diversity, value and limitations of the journal impact factor and alternative metrics. Rheumatology international, 32(7), 1861-1867.

Breuker, J. (2013). A cognitive science perspective on knowledge acquisition. International Journal of Human-Computer Studies, 71(2), 177-183.

Butler, D. (2008). Free journal-ranking tool enters citation market. Nature, 451(7174), 6.

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.

Ciskowski, P. (2006). Efficient Modeling of Contextual Mappings by Context-dependent Feedforward and Recurrent Neural Nets. In Neural Networks, 2006. IJCNN'06. International Joint Conference on (pp. 1098-1105). IEEE.

Consoli, D. (2005). The dynamics of technological change in UK retail banking services: An evolutionary perspective. Research Policy, 34, 461–480.

Coppock, D. S. (2002). Why Lift? Data Modeling and Mining. Information Management Online (Online; accessed 19-July-2013).

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

Cortez, P. (2010). Data mining with neural networks and support vector machines using the r/rminer tool, Advances in Data Mining. Applications and Theoretical Aspects, 6171 (pp. 572–583). Springer.

Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Systems with Applications, 37(3), 2132-2143.

Cox, D. F., & Good, R. E. (1967). How to build a marketing information system. Harvard Business Review, 45(3), 145-154.

Demyanyk, Y., & Hasan, I. (2010). Financial crises and bank failures: a review of prediction methods. Omega, 38(5), 315-324.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. Communications of the ACM, 49(9), 76-82.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2005). Visualization of Text Document Corpus. Special Issue: Hot Topics in European Agent Research I Guest Editors: Andrea Omicini, 497.

Gebus, S., & Leiviskä, K. (2009). Knowledge acquisition for decision support systems on an electronic assembly line. *Expert Systems with Applications*,*36*(1), 93-101.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. Scientometrics, 53(2), 171-193.

Guo, L., Chehata, N., Mallet, C., & Boukir, S. (2011). Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS Journal of Photogrammetry and Remote Sensing, 66(1), 56-66.

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

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. (2006). Feature extraction. Foundations and applications. Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Halpern, J. Y., & Parkes, D. C. (2011). Journals for certification, conferences for rapid dissemination. Communications of the ACM, 54(8), 36-38.

Hong, J. (2013). Ph. D. students must break away from undergraduate mentality. Communications of the ACM, 56(7), 10-11.

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37(7), 5259-5264.

Hunter, A., Kennedy, L., Henry, J., & Ferguson, I. (2000). Application of neural networks and sensitivity analysis to improved prediction of trauma survival. Computer methods and programs in biomedicine, 62(1), 11-19.

Hurlburt, G. F., Miller, K. W., & Voas, J. M. (2009). An ethical analysis of automation, risk, and the financial crises of 2008. IT professional, 11(1), 14-19.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. Journal of computational and graphical statistics, 5(3), 299-314.

Kerstein, J., & Kozberg, A. (2013). Using Accounting Proxies of Proprietary FDIC Ratings to Predict Bank Failures and Enforcement Actions During the Recent Financial Crisis. Journal of Accounting, Auditing & Finance, 28(2), 128-151.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of Healthcare Information Management, 19(2), 64-72.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1), 273-324.

Kuechler, B., & Vaishnavi, V. (2011). Promoting relevance in IS research: an informing system for design science research. Informing Science: the International Journal of an Emerging Transdiscipline, 14(1), 125-138.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1-30.

Horwitz, J. (2014). R and Hadoop make Machine Learning Possible for Everyone. Available at: http://www.kdnuggets.com/2014/11/r-hadoop-make-machine-learning-possible-everyone.html.

Krafft, M., Hesse, J., Höfling, J., Peters, K., & Rinas, D. (2007). International direct marketing. Springer.

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. Decision Support Systems, 32(4), 361-377.

Liu, H., & Motoda, H. (Eds.). (2007). Computational methods of feature selection. CRC Press.

Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining. In Proceedings of the Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining (pp. 4-13).

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. Knowledge and Data Engineering, IEEE Transactions on, 17(4), 491-502.

Luo, X., Li, H., Zhang, J., & Shim, J. P. (2010). Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services. Decision Support Systems, 49(2), 222-234.

Mansingh, G., Rao, L., Osei-Bryson, K. M., & Mills, A. (2013). Profiling internet banking users: A knowledge discovery in data mining process model based approach. Information Systems Frontiers, 1-23.

MAANZ, The Marketing Association of Australia and New Zealand (2014, November 8). Glossary of Marketing Terms. Available at: http://www.marketing.org.au/.

Mehrotra, A., & Agarwal, R. (2009). Classifying customers on the basis of their attitudes towards telemarketing. Journal of Targeting, Measurement and analysis for Marketing, 17(3), 171-193.

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in R. Journal of Statistical Software, 25(5), 1-54.

Milne, G. R., & Rohm, A. J. (2000). Consumer privacy and name removal across direct marketing channels: exploring opt-in and opt-out alternatives. Journal of Public Policy & Marketing, 19(2), 238-249.

Monroe, G. A., Freeman, K., & Jones, K. L. (2012). IT Data Mining Tool Uses in Aerospace. Proceedings of the 2012 NSBE Aerospace Systems Conference, NASA Ames Research Center.

Moro, S., Cortez, P., & Laureano, R. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011, EUROSIS.

Moro, S., Cortez, P., & Laureano, R. (2012). Enhancing Bank Direct Marketing through Data Mining. In Proceedings of the 41th European Marketing Academy Conference (EMAC), Lisbon, Portugal, May, 2012, European Marketing Academy.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

Moro, S., Cortez, P., & Rita, P. (2015a). Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation, Expert Systems with Applications, 42(3), 1314-1324.

Moro, S., Cortez, P., & Rita, P. (2015b). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications, 26, 131–139.

Moutinho, L. (2000). Segmentation, Targeting, Positioning and Strategic Marketing, in: Moutinho, L. (Ed.), Strategic management in tourism, CABI, pp. 121-165, (part two, 5).

Muenchen, R. A. (2012). The popularity of data analysis software. Available online at: http://r4stats.com/articles/popularity/.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*,*50*(3), 559-569.

Ngai, E. W. T., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592-2602.

Oliveira, L. S., Sabourin, R., Bortolozzi, F., & Suen, C. Y. (2002). Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 1, pp. 568-571). IEEE.

Parnas, D. L. (2007). Stop the numbers game. Communications of the ACM,50(11), 19-21.

Pennings, J. M., & Harianto, F. (1992). The diffusion of technological innovation in the commercial banking industry. Strategic Management Journal, 13, 29–46.

Rettie, R., Grandcolas, U., & Deakins, B. (2005). Text message advertising: Response rates and branding effects. Journal of Targeting, Measurement and Analysis for Marketing, 13(4), 304-312.

Rotfeld, H. J. (2004). Do-not-call as the US Government's improvement to telemarketing efficiency. Journal of Consumer Marketing, 21(4), 242-244.

Saltelli, A. (2002). Sensitivity analysis for importance assessment. Risk Analysis, 22(3), 579-590.

Sharma, A., Imoto, S., & Miyano, S. (2012). A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(3), 754-764.

Simon, H. A. (1960). The new science of management decision. Harper & Brothers.

Skulimowski, A. M. (2011). Future trends of intelligent decision support systems and models. In Future Information Technology (pp. 11-20). Springer Berlin Heidelberg.

Sprague Jr, R. H. (1980). A framework for the development of decision support systems. MIS Quarterly, 1-26.

Sprague Jr, R. H., & Watson, H. J. (1976). A decision support system for banks. Omega, 4(6), 657-671.

Tan, D. W., Liew, S. Y., & Yeoh, W. (2014). Improving telemarketing intelligence through significant proportion of target instances. Proceedings of the 2014 Pacific Asia Conference on Information Systems, Association for Information Systems, Paper 368.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting, 16(4), 437-450.

Turban, E., Sharda, R., Delen, D., & Efraim, T. (2011). Decision support and business intelligence systems, 9th Edition. Pearson.

Tuv, E., Borisov, A., Runger, G., & Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. The Journal of Machine Learning Research, 10, 1341-1366.

Van Bruggen, G. H., Smidts, A., & Wierenga, B. (1998). Improving decision making by means of a marketing decision support system. Management Science, 44(5), 645-658.

Verhoef, P. C., & Leeflang, P. S. (2009). Understanding the marketing department's influence within the firm. Journal of Marketing, 73(2), 14-37.

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications, 38(1), 223-230.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. Computer, 40(9), 96-99.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Yada, K., Washio, T., & Ukai, Y. (2010). Modelling deposit outflow in financial crises: application to branch management and customer relationship management. International Journal of Advanced Intelligence Paradigms, 2(2), 254-270.

Young, M. (2002). Choice-Based Segmentation and Targeting the ‘Switchable’Customer. Marketing Management Journal, 12(1), 98-106.

Yousafzai, S. Y., Foxall, G. R., & Pallister, J. G. (2010). Explaining Internet Banking Behavior: Theory of Reasoned Action, Theory of Planned Behavior, or Technology Acceptance Model? Journal of Applied Social Psychology, 40(5), 1172–1202.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, 5, 1205-1224.

Zopounidis, C., Doumpos, M., & Matsatsinis, N. F. (1997). On the use of knowledge-based decision support systems in financial management: a survey.Decision Support Systems, 20(3), 259-277.

Glossary

Machine Learning Terms

Data-based Sensitivity Analysis – A Sensitivity Analysis method which uses samples from the dataset used for training a model in order to assess changes to the outcome to predict when varying several input features (Cortez & Embrechts, 2013).

Feature Relevance – A measure of the impact that a certain feature has in the definition of a model (Yu & Liu, 2004).

Latent Dirichlet Allocation – A technique for modeling a certain number of distinct topics defined according to the number and distribution of terms across a given set of documents (Blei, 2012).

Lift Cumulative Curve – Plots the population samples (ordered by the deciles, x-axis) versus the cumulative percentage of real responses captured (y-axis) (Coppock, 2002).

ROC Curve – Shows the performance of a two class classifier across the range of possible threshold values, plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) (Fawcett, 2006).

Rolling Window – A scheme for selecting data for training a model which is updated with latest data and discards oldest, keeping a fixed sized window of training data (Leigh et al., 2002).

Sensitivity Analysis – Methods which measure the sensitivity of a given model to changes on the input features (Guyon & Elisseeff, 2003).

Variable Effect Characteristic Curve – Plots the feature range of values (x-axis) versus the effect on a given model on the outcome to predict (y-axis) (Cortez & Embrechts, 2013).

Marketing Terms

Campaign Contact – Contacting a potential customer within a campaign context (MAANZ, 2014).

Inbound Telemarketing – Telemarketing method in which a company receives telephone orders and enquiries from customers (MAANZ, 2014).

Marketing Campaign – Organized course of Marketing action, planned to achieve predefined objectives (MAANZ, 2014).

Outbound Telemarketing – Telemarketing method in which a firm uses trained agents for selling to customers by telephone (MAANZ, 2014).

Targeting – Narrowly focusing promotions to attract specific individuals and potential customers (MAANZ, 2014).

Telemarketing – Marketing communication method using telecommunication technology (e.g., telephone) and trained agents to conduct planned marketing activities directed at targeting customers (MAANZ, 2014).