Serv Bus (2017) 11:831–849
DOI 10.1007/s11628-016-0332-3

EMPIRICAL ARTICLE

Predicting direct marketing response in banking: comparison of class imbalance methods

Vera L. Miguéis[1] · Ana S. Camanho[1] · José Borges[1]

Received: 13 May 2016 / Accepted: 20 December 2016 / Published online: 2 January 2017

Abstract Customers' response is an important topic in direct marketing. This study proposes a data mining response model supported by random forests to support the definition of target customers for banking campaigns. Class imbalance is a typical problem in telemarketing that can affect the performance of the data mining techniques. This study also contributes to the literature by exploring the use of class imbalance methods in the banking context. The performance of an undersampling method (the EasyEnsemble algorithm) is compared with that of an oversampling method (the Synthetic Minority Oversampling Technique) in order to determine the most appropriate specification. The importance of the attribute features included in the response model is also explored. In particular, discriminative performance was enhanced by the inclusion of demographic information, contact details and socio-economic features. Random forests, supported by an undersampling algorithm, presented very high prediction performance, outperforming the other techniques explored.

Keywords Direct marketing · Response modelling · Customer targeting · Random forests · Class imbalance

✉ Vera L. Miguéis
vera.migueis@fe.up.pt

Ana S. Camanho
acamanho@fe.up.pt

José Borges
jlborges@fe.up.pt

[1] Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

1 Introduction

Most companies engaged in commercial business adopt strategies to promote their products and services. Banking, insurance and retailing are examples of highly competitive sectors, in which the promotional activity is intense. Both mass marketing and direct marketing strategies are common in these sectors, although nowadays companies are increasingly adopting direct marketing approaches, which involve effective communications with customers. Direct marketing can be defined as a data-driven, cross-media, interactive, multichannel process for building and cultivating mutually beneficial relationships between companies and their current or potential customers (Direct Marketing Association 2012). This approach implies the study of customers' characteristics and needs, leading to the selection of a target group for direct promotions. The communication between companies and customers can be done by e-mail, SMS, phone calls or post. Direct marketing aims at increasing the response rate of the campaigns.

Marketers have identified some advantages and disadvantages of direct marketing. Vriens et al. (1998) highlight the flexibility of direct marketing in terms of formats, timing and testing. Verhoef et al. (2003) refer the ability of involving the customers, the possibility of communicating customized offers and the lack of direct competition for the attention of the customer. While increasing the response rate is not an easy task, the impact of a small increment can result in a significant profit increase. Baesens et al. (2002) observed that a 1% increase in the response rate of a mail-order company generated an additional profit of 500,000 €. Direct marketing can also strengthen customer loyalty (Sun et al. 2006). When properly targeted, customers are more likely to appreciate the initiates of the company and feel satisfied. However, the relatively high cost of direct mail, when compared with traditional media, may compromise its use. Thus, increasing the response rate of direct marketing campaigns is essential to limit their costs whilst maintaining effectiveness. Moreover, customers may develop negative attitudes towards direct communications if the promotions become too frequent and are not of their interest (Ansari et al. 2008).

Particularly in banking, due to competition and the current financial crisis, the sector has been forced to increase financial assets. Therefore, a relatively new trend in banking is to offer products through direct marketing campaigns. In fact, due to the pressure for a reduction in costs, banks have realized that fewer but more personal contacts may represent a more efficient strategy to promote their offers. In this context, direct marketing campaigns in banking have gained relevance. Due to the popularity of direct marketing, the literature on this topic is also growing, but can still be considered scarce (Gzquez-Abad et al. 2011).

Nevertheless, banks are also investing in technology (Lu et al. 2015; Abroud et al. 2015), which enables the collection of a huge number of customer-related records (Chen et al. 2014). In this sector, data mining techniques have a huge potential to explore the data and turn them into useful knowledge for customer retention and attraction. Data mining can be very effective in supporting direct marketing in banking, as it enables to predict which customers will respond positively to the campaigns (Ling and Li 1998). For example, Ngai et al. (2009) provide a literature review of the application of data mining to customer relationship management contexts. It proposes a framework describing the most appropriate data mining techniques to address specific customer relationship management dimensions, namely customer attraction through direct marketing. Amini et al. (2015) and Govindarajan (2015) are examples of successful applications of data mining to guide direct marketing actions.

In this context, the current study addresses the following questions:

(1) Are random forests an appropriate data mining technique to predict customers' response to direct telemarketing campaigns in a banking context?
(2) Given the usual imbalance nature of the data used to support direct marketing decisions, what is the best class imbalance method (the EasyEnsemble algorithm or the Synthetic Minority Oversampling Technique) to enhance the predictive performance of the data mining techniques?
(3) What are the dimensions that determine the propensity of customers to answer positively to direct marketing campaigns?

The remainder of the paper is organized as follows. The next section explores the use of data mining in the banking sector. Section 3 presents the methodology used in this study, including the class imbalance methods and the method used to infer the importance of variables. Section 4 presents the case study and the empirical results and the last section draws some conclusions.

2 Data mining in banking

The banking sector has changed significantly in recent years. Currently, banks follow the technological trends and are equipped with new banking technologies. Due to the use of the web and automated software, the basic concepts and operations of banking business are changing. This allows managers to focus on the improvement of financial performance and customer relationship, which nowadays is a critical dimension of banking strategy (e.g. Wang et al. 2014; Chih et al. 2014).

Similarly to other sectors, the banking industry is trying to obtain competitive advantages using customers' data. The process of extracting knowledge from data and use this knowledge to support the design of strategic plans is essential to succeed in the increasingly competitive market.

Data mining has become a widely accepted process to improve organizational performance in several contexts by transforming data into useful knowledge (e.g. Zarnani et al. 2009; Burton et al. 2014; Seret et al. 2015). It provides the ability to access huge volumes of raw data and obtain useful information at the right time. The application of modern techniques, routed in statistics, mathematics, artificial intelligence and machine learning, allows to extract information from incomplete, noisy and vague data (Jingbiao and Shaohong 2010) and find hidden patterns that may be valuable for decision making.

In recent years, banking companies have adopted data mining technologies for several purposes. Examples include customer segmentation and profitability estimation, credit scoring, identification of fraudulent transactions, cross-selling and customer retention (Hu 2005; Yeh and Lien 2009; Jayasree 2013). Customer segmentation consists in establishing customer groups and including each new customer in the right group, or updating the group a customer belongs to. For example, Li et al. (2010) classified credit card customers into four segments based on data from a Chinese commercial bank credit card. This information was used to search for potential customers and to implement target marketing. Hsieh (2004) identified three major groups of profitable customers based on repayment behaviour, recency, frequency and monetary value. This study aimed at supporting the inference of the profiles of customers and the design of appropriate strategies suitable to the characteristics of each group of customers. The prediction of customer profitability encompasses the identification of indicators that may be behind the profit levels associated to each customer. The potential profitability of new customers can then be estimated based on those indicators. For example, Khajvand and Tarokh (2011) estimated, for each segment of customers in the retail banking scope, their future value based on recency, frequency and monetary value. Ras and Wieczorkowska (2000) classified customers into groups of different profitability and showed which actions should be taken to improve their profitability.

Credit scoring represents a process to convert the characteristics of the applicants into numbers that are combined in order to obtain a score that represents the risk profile of the applicant. Migueis et al. (2013) used data from a German bank to provide accurate predictions of which customers may default in the future and highlighted the factors behind the probability of default. Furthermore, this study proposed a segmentation scheme of the customers in terms of risk of default and the corresponding uncertainty about the prediction. Xiong et al. (2013) investigated the use of the sequence pattern of clients' behaviour to identify bad accounts and to mine bankruptcy features. Fraud detection consists in anticipating and quickly detecting fraud in order to support immediate action to minimize costs. For example, Quah and Sriganesh (2008) focused on real-time fraud detection and presents an innovative approach to understand spending patterns to recognize potential fraud cases. Libana-Cabanillas et al. (2013) explored several methods to determine which variables, including socio-demographic, economic, financial and behavioural strategic variables, are the most important to predict the likely levels of trust among electronic banking users. Cross-selling implies the development of strategies to sell different products, especially in additional product categories, to a customer who intends to purchase an initial product. Cohen (2004) explored what products, if any, should be offered to each customer in order to maximize the marketing return on investment. Liao et al. (2009) developed an approach to establish potential cross-selling through personalized product mix analysis, considering customers' needs and making suggestions for new product developments. Finally, customer retention consists in identifying the customers who may leave the bank and adjust the product portfolio, the pricing and the promotions offered in order to retain those customers. Nie et al. (2011) built a model to identify the customers who may leave the company using credit card data collected from a Chinese bank. This study explores the contribution of four variable categories: customer information, card information, risk information and transaction activity information. Gür Ali and Aritürk (2014) proposed a dynamic framework to identify the defectors, the early defectors and the indicators of customer abandonment.

A relatively new trend in banking is to offer products through direct marketing campaigns. The application of data mining techniques to identify target customers has a huge potential in this field, as it enables the identification of those customers who are more likely to respond to a campaign. Although the introduction of response models supported by data mining is frequent in certain sectors, such as insurance (e.g. Schwartz and Lauridsen 2007; Chan and Loh 2004), retail (e.g. Ha et al. 2005; Olson and Chae 2012) and telecommunications (e.g. Chen et al. 2011), there is room for further developments in the banking sector. Indeed, the study developed by American Banker Research in 2012 American Banker (2012) revealed that 71% of the 170 bankers included in the survey do not use customer analytics and that only 6% of these non-users were planning to use analytics in the following 12 months. The literature also reinforces the scarcity in the use of data mining to support marketing actions in banking sector. Using the keywords “data mining”, “marketing” and “banking” in the search platform Elsevier’s Scopus, the results obtained only represent 0.1% of those obtained when using the keyword “banking”. The number of publications involving data mining and marketing in banking sector has remained stable over the last years and their authors mostly originate from India, United States of America, China and France.

Examples of the use of data mining to support direct marketing in banking include Ling and Li (1998) who explored the dimension of the dataset and the predictive accuracy involved in the process of identifying the likely subscribers of a loan product from a major bank in Canada. Elsalamony and Elsayad (2013) evaluated the classification performance of two different data mining techniques in order to establish a model able to increase the effectiveness of a deposit campaign. This study also identified the main characteristics that affect the success of the direct marketing campaign. Moro et al. (2014) used the same dataset to predict customer response. This paper used a feature selection procedure and compared the classification performance of several data mining techniques. Ayetiran and Adeyemo (2012) explored the use of customers’ historic purchases and demographic information to predict the probability that a customer will respond to a promotion of a Nigerian bank.

The imbalance in the data is an issue that has deserved attention in other settings, such as fraud detection, heart failure and customer churn (e. g. Ling and Li 1998; Burez and Van den Poel 2009; Srinivas et al. 2014), as non-respondents usually outnumber respondents. When the imbalance is not properly mitigated, the models developed may lead to poor prediction. This fact has been disregarded in direct marketing response (Kim et al. 2013), particularly in banking. This paper aims to fill this gap by comparing different methods to deal with class imbalance in response modelling in banking. Furthermore, the purpose of this study is to develop an accurate model to predict customers’ response and to identify the customer’s characteristics that motivate the subscription of the offer.

3 Methodology

Following the framework introduced by Ngai et al. (2009), the methodology followed in this paper aims to explore the use of data mining classification techniques to support customers' attraction. More specifically, we propose the use of random forests to predict customers' response to direct marketing campaigns. Random forests is a classification technique that is used to predict categorical class labels (see Han and Kamber 2006, for further details). In our case, the class is binary and corresponds to a positive or negative response to the campaign. The performance of random forests (see Breiman 1996, for further details) is compared with some of the most popular classification techniques used in the literature: logistic regression, neural networks (Mcculloch and Pitts 1943) and support vector machines (Vapnik 1995) (e.g. Ben Ishak 2016; Hosseini and Bideh 2014; Olson et al. 2009). This paper also aims to address data imbalance in response prediction by exploring the use of EasyEnsemble and Synthetic Minority Oversampling Technique (SMOTE).

In the remainder of this section, we present the methods selected to deal with class imbalance, the evaluation criteria adopted to estimate classification performance and the method used to determine the importance of the variables included in the final model.

3.1 Class imbalance problems

Most classification algorithms expect an approximately even distribution of the instances among the different classes. The results obtained suffer when this is not the case (Garca-Pedrajas et al. 2012). Therefore, handling class imbalance constitutes a very relevant task, as the majority of direct marketing prediction models use unbalanced samples.

There are several methods and algorithms available to mitigate the effect of class imbalance on the accuracy of the predictions. The literature suggests three main approaches to deal with this issue: internal approaches acting on the algorithm, external approaches acting on the data and combined approaches based on boosting (see Sun et al. 2007; Garca and Herrera 2009; Freund and Schapire 1996, for further details). The first approach consists in adapting the learning algorithm to deal with the imbalance problem. For example, the decision threshold of the algorithm may be modified in order to create a bias towards the minority class. The second approach involves adjustments to the data, such as oversampling the minority class or undersampling the majority class. The third approach modifies the basic boosting method to account for the underrepresentation of the minority class.

The utilization of an external approach based on data sampling has several advantages over the other approaches. Sampling does not depend on the possibility of adapting a certain algorithm and is more general. The modification of the learning algorithms is usually a very difficult task (Garca-Pedrajas et al. 2012). In this paper, we adopted an external approach based on data sampling.

As mentioned before, the data sampling approach can generally be classified into two groups: undersampling the majority class and oversampling the minority class. There are some algorithms that combine both methods. Undersampling and oversampling methods can be supported by a random procedure or can involve the identification of the most or least useful instances using more sophisticated procedures.

3.1.1 Undersampling methods

Undersampling consists in creating a subset of the original dataset by selectively or randomly disregarding some of the samples of the majority class while keeping the original population of the minority class (Japkowicz and Stephen 2002). In general, undersampling algorithms lead to better results than oversampling algorithms, especially when they use sophisticated data elimination methods (Chawla et al. 2002). In addition, according to (Ling and Li 1998) the combination of undersampling with oversampling approaches does not provide better results than simply undersampling the majority class.

The literature encompasses several undersampling methods. For example, Yen and Lee (2009) propose cluster-based undersampling approaches in order to improve the classification accuracy for the minority class. Garca and Herrera (2009) propose an evolutionary undersampling method for classification with imbalanced datasets. Liu et al. (2009) proposed two informed undersampling methods: EasyEnsemble and BalanceCascade. The EasyEnsemble method was adopted in this study because it is easy to use in real-world contexts.

EasyEnsemble consists in developing an ensemble learning system by independently sampling several subsets from the majority class and developing multiple classifiers based on the combination of each subset with the minority class data. This method can be considered as an unsupervised learning algorithm that explores the majority class by using independent random sampling without replacement (He 2011). In this study, we used five balanced subproblems sampling from the majority class.

3.1.2 Oversampling methods

Oversampling methods produce a superset of the original dataset by replicating some of the samples from the minority class or creating new samples from the original minority class samples. The replication of the minority class has proven not to be efficient, as it consists in merely making copies of the minority class samples, without adding new information to the dataset. Introduced by Chawla et al. (2002), Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling technique, which overcomes this problem by creating new synthetic samples from the minority class by randomly interpolating pairs of the closest neighbours in the minority class.

Consider a subset S, and the K-nearest neighbours for each example $x_i \in S$ belonging to the minority class dataset are defined as the K elements of S whose Euclidean distance between itself and the x_i under consideration exhibits the smallest magnitude along the n-dimensions of feature space X. To create a synthetic sample x_{new}, one of the K-nearest neighbours $\hat{x}_i$ is randomly selected and then the corresponding feature vector difference is multiplied by a random number δ between 0 and 1, which is added to x_i [see Eq. (1)] He (2011).

$$x_{new} = x_i + \delta\left(\hat{x}_i - x_i\right) \tag{1}$$

In this study, the number of nearest neighbours that are used to generate the new examples of the minority class was 1000.

3.2 Evaluation criteria

To estimate the performance of the classification models, we used the area under the receiver operating characteristic curve (ROC) (Lessmann et al. 2008). In the context of this study, ROC illustrates the trade-off between the proportion of false responders and the proportion of true responders for every possible cut-off of the predicted probability. An area under the ROC curve (AUC) close to 1.0 reveals that the model has perfect discrimination, while an AUC close to 0.5 suggests poor discrimination.

In the context of marketing, a decile analysis is also of utmost importance (Ratner 2004). Marketers are usually interested in targeting the segment of customers with the highest propensity to respond positively to direct marketing campaigns. Therefore, companies group the customers with the highest probability of taking advantage of the campaign and estimate how much higher the response rate can be within that selection over the expected response rate from a random selection.

Consider that a company is interested in the top p-th percentile of most likely responders, based on predicted response probabilities. The top p-th percentile lift is equal to the ratio of the proportion of responders in the top p-th percentile of ordered posterior response probabilities to the response rate of the population composed by all customers. For example, a p-th percentile lift equal to 2 means that the model identifies two times more customers who respond positively in the top p-th percentile than a random assignment would do. Given the fact that the proportion of customers that a company intends to target depends on several factors, namely budget available for the promotional campaign; this paper includes the lift for different percentiles (10 and 20%).

Following the literature recommendations (Alpaydin 2009), we used a 10-fold cross-validation to estimate the performance. Each dataset under consideration (imbalanced, balanced through SMOTE or balanced through EasyEnsemble) was randomly split into ten disjoint subsets. Ten runs were conducted, and in each run nine of the subsets were combined to form the training set, while the remaining subset formed the testing set.

3.2.1 Variable importance

The identification of relevant predictors is important to get insights into the customers' characteristics that affect the propensity to accept a direct marketing offer. By means of variable importance indicators, the predictor variables can be compared with respect to their impact in predicting the response to the campaigns. An advantage of random forest is the ability to accommodate easily a procedure to reflect the importance of the variables. Consequently, variable importance measures have been receiving increased attention in many classification settings involving the use of random forests.

A random forest variable importance measure is related to the following key idea: when a variable that contributes to the prediction of accuracy is "noised up", i.e. replaced with random noise, the classification performance should decrease. In contrast, if a variable is irrelevant, "noising" it up should have little effect on performance. A number of specific variable importance measures were proposed based on this idea (Friedman 2001).

In this paper, we use a popular measure computed by permuting the data of the instances that are left out of the bootstrap sample used to construct each tree (usually called "out-of-bag" (OOB) data). For each tree, the prediction error on the OOB data is recorded. Then, the error is estimated after permuting each predictor variable. The variable importance indicator results from the average over all trees of the difference between the two error measures, normalized by the standard deviation of the differences.

4 Case study

4.1 Data

The empirical component of this study is centred in a Portuguese banking institution that uses its contact centre to develop directed marketing campaigns. This study is focused on a direct marketing campaign developed to sell long-term deposits with attracting interest rates, from May 2008 to November 2010. This period encompasses the acute international crises faced by economies worldwide. During this period, credit tightened and international trade declined. In Portugal, the crisis was experienced in an intense way, triggering capital injections into some banks. In such context, most banks tried to increase their financial resources by attracting long-term deposit applications.

The bank used in the empirical application uses as the main communication channel with their customers the telephone, with a human agent as the interlocutor. The bank's online platform also provides information regarding bank's offers to specific target customers. The dataset gathered encompasses 41,188 contacts conducted in order to persuade customers to subscribe attractive long-term deposit applications with good interest rates. It is important to note that sometimes more than one contact to the same customer was required, in order to be sure whether or not the offer would be subscribed.

For each customer, a large number of attributes were stored, including demographic characteristics (e.g. the age of the customer and marital status), contact details (e.g. duration of the last call and the number of contacts conducted before this campaign for the same customer), customers' financial standing (e.g. existence of personal loan and existence of credit in default) and employment (e.g. type of job). Due to privacy issues, some of the information collected was not made available for this study. Furthermore, the information available was enriched by the addition of social and economic features, such as consumer confidence index and consumer price index. This dataset is publicly available for research. The details are described in Moro et al. (2014). The set of explanatory variables used in this study and their descriptive statistics are shown in Table 3 of the Appendix. The dependent variable reveals subscription of the deposit (value = 1) or non-subscription (value = 0).

The dataset is mostly composed of married people with a university degree, who work in administration and have a housing loan. The average age of these customers is about 40 years. The dataset is imbalanced, as only 4640 customers (11.2%) subscribed the deposit (see Table 3).

4.2 Results and discussion

4.2.1 Random forest models

The performance of response models is of utmost importance in direct marketing. This section compares the predictive performance of random forests when disregarding the imbalance nature of the problem with the performance of models when considering an undersampling and an oversampling approach. The undersampling method adopted was EasyEnsemble, and the oversampling method adopted was SMOTE. The performance evaluation was based on the AUC, top 10% lift and top 20% lift (see Table 1).

The lowest AUC obtained through a 10-fold cross-validation was 0.945, corresponding to both the imbalanced dataset and the SMOTE dataset. This result reveals that the proposed model has very good discrimination power since it significantly exceeds the random classifier corresponding to an AUC of 0.5. Furthermore, the lowest 10% lift was 5.542, corresponding to the SMOTE approach, and the lowest 20% top lift was 3.271, corresponding to an imbalanced approach. These results reveal that the model proposed is able to identify 5.542 and 4.386 more applicants than a random model would do in the top 10 and 20%, respectively.

Table 1 Random forests' performance

	Random forests		
	AUC	Lift 10	Lift 20
Imbalanced	0.945	5.678	4.371
SMOTE	0.945	5.542	4.386
EasyEnsemble	0.989	7.937	4.960

For all metrics considered, the best results were obtained when considering the EasyEnsemble method to overcome data imbalance, which resulted in an AUC of 0.989 and a 10 and a 20% top lift of 7.937 and 4.960, respectively. This AUC corresponds to a very good discrimination power, very close to a perfect model. It is important to note that the model based on imbalanced data performs similarly to the model based on data including synthetic samples.

The proposed model has a significant impact in the banking direct marketing actions. For instance, the top percentile lift reveals that the number of successful contacts can increase significantly when targeting only a relatively small number of customers. In particular, if the response model supported by EasyEnsemble is implemented, the company can achieve a positive response rate of 89.41% by contacting 10% of the customers, which is significantly higher than the value of 11.26% concerning the current bank practice, based on contacts to all customers.

4.2.2 Comparative performance of logistic regression, neural networks and SVM

The use of random forests to estimate the propensity of a customer to respond to an offer was validated through the comparison with the performance of logistic regression, neural networks and support vector machines. The same indicators of performance presented in the previous section were utilized (see Table 2).

The analysis of Tables 1 and 2 shows that random forests technique outperforms all the other techniques considered. This happens in the case of the utilization of imbalanced data, data balanced through SMOTE and EasyEnsemble. The AUC results are confirmed by the top percentile lifts.

From Table 2, it can also be revealed that the EasyEnsemble is not always the most appropriate method. For example, logistic regression performed better when applied to data previously adjusted using SMOTE, although the differences between the performances of these methods is not very pronounced. In the case of neural networks, the best results were obtained using imbalanced data. In the case of support vector machines, the best results were obtained using SMOTE. However, when the company is interested in targeting the top 10% customers, the best results were obtained using imbalanced data.

These results reveal that the identification of the best approach to deal with class imbalance issues is not straightforward. While in some cases the adoption of an EasyEnsemble approach contributes to significantly increasing the prediction

Table 2 Comparative performances

	Logistic			NN			SVM		
	AUC	Lift 10	Lift 20	AUC	Lift 10	Lift 20	AUC	Lift 10	Lift 20
Imbalanced	0.928	5.488	4.052	0.935	5.660	4.166	0.925	5.300	4.092
SMOTE	0.930	5.445	4.119	0.900	4.544	3.720	0.928	4.112	4.112
EasyEnsemble	0.890	4.365	4.094	0.933	5.309	4.302	0.877	3.764	3.765

performance (e.g. using random forests), in other cases the use of data sampling approaches does not contribute significantly to obtaining better prediction results (e.g. Logistic, NN and SVM).

In line with these results, academics and practitioners should consider using other techniques, such as random forests, an alternative to the traditional approaches, such as logistic regression. Nevertheless, a trade-off has to be made between the time allocated to the modelling procedure and the performance achieved.

4.2.3 Variable importance

In this section, we provide an overview of the most important variables for the prediction of customer response in this context. This is done based on the outcome of the random forest importance measure introduced in Section 3.2.1. In this study, we will elaborate on the top 20 most important response predictors.

From the analysis of Table 4 in the Appendix, it is clear that the duration of the previous call is the variable that influences the most this model of direct marketing response prediction. The importance of this variable could be anticipated, as it is acceptable that if in the past a customer had a long conversation with the interlocutor, he/she may be sensitive to the company offers. Consequently, this customer may be prone to subscribe the new offer. The opposite also seems to be valid. Moreover, the top 20 most explaining response variables include all socio-economic variables representing the conjuncture at the time of the offer. Furthermore, the age of the customer is the demographic variable that influences the model most strongly. Finally, recency, measured by the number of days since last contact, is also included in the top 20 most relevant variables.

Summing up, the response variables with the greatest explanatory power reflect the following dimensions: demographic characteristics of customers, contact details and socio-economic features of the conjuncture. Customers' financial standing and employment are not part of this group of variables considered the most relevant for prediction purposes in the context studied.

5 Conclusions

This study proposed a model to predict the response to a direct marketing campaign using real data from a Portuguese bank. It aimed at increasing the subscription rate of a long-term deposit. The paper contributes to the existing literature in three distinct ways. First, it explores the use of random forests for response prediction, as the application of this technique in this domain is still incipient. Secondly, this study explores the impact of class imbalance in the prediction performance, using SMOTE and EasyEnsemble methods. Finally, this paper identifies the dimensions that contribute the most to distinguishing the customers who subscribe from the others.

Similarly to what happens in other settings, random forests showed very good performance in this study. Its performance was better when used in combination with the EasyEnsemble method, as this guarantees a balanced distribution on the data used to train the model. To validate the performance of random forests, we statistically compared its predictive performance with that of logistic regression, support vector machines and neural networks. It was shown that random forests technique outperforms the other techniques when using both balanced and imbalanced data. However, these results cannot be generalized to all situations, due to the data-driven nature of this technique.

When using logistic regression, support vector machines and neural networks, the EasyEnsemble method was not the most appropriate method to deal with class imbalance. The adoption of data sampling approaches to overcome the class imbalance problem was only advantageous when using random forests. Therefore, analysts should be aware that for each particular case a detailed exploratory analysis of the impact of the use of different class imbalance methods should be developed.

In this study, we identified the most important response predictors, which include demographic characteristics of customers, contact details and socio-economic features of the conjuncture. Despite the importance of customers' financial standing and type of job for several banking decisions, these features did not play an important role in explaining response in the context of this study.

The results of the empirical application emphasize the potential of data mining techniques, to support the design of direct marketing campaigns. These techniques, particularly random forests, allow to predict customer behaviour and consequently may be used to effectively specify the target of the campaigns. Being linked to the effectiveness of the contacts established, the model proposed allows a reduction in marketing campaigns' costs, as fewer contacts are required to increase significantly the response rate. Furthermore, this decision support tool enables the creation of stronger relationships with customers, as it avoids disturbing those who are not interested in the offer, and consequently increases their likelihood to respond positively to future communications.

Appendix

See Tables 3 and 4.

Table 3 Dataset variables

Attribute	Variables	Type	Name	Description	Mean (SD)/Most (freq.) and Least (freq.)
1	1	Quantitative	Age (years)		40.0 (10.4)
2	2–12	Dummy	Type of job	Housemaid, services, administration, blue-collar, retired technician, entrepreneur, self-employed, student, management, unemployed	Administration (10422) and Student (875)
3	13–15	Dummy	Marital status	Married, single, divorced	Married (24928) and Divorced (4612)
4	16–21	Dummy	Educational level	Basic (4th year), basic(6th year), basic (9th year), professional course, high school, university degree	University degree (12168) and Basic (6th year) (2292)
5	22	Dummy	Credit default	No	No (32588) and Yes (8600)
6	23–24	Dummy	Housing loan	No, Yes	Yes (21576) and No (18622)
7	25	Dummy	Personal loan	No	No (33950) and Yes (7238)
8	26	Dummy	Contact communication type	Mobile phone	Mobile phone (15044) and Telephone (26144)
9	27	Quantitative	Last contact month of year		6.6 (2.0)
10	28	Quantitative	Last contact day of the week		3.0 (1.4)
11	29	Quantitative	Last contact duration (seconds)		258.3 (259.3)
12	30	Quantitative	# contacts performed during the analysis period		2.6 (2.8)
13	31	Quantitative	# days since last contact		962.5 (186.9)
14	32	Quantitative	# previous contacts		0.2 (0.5)
15	33–34	Dummy	Outcome of the previous marketing campaign	Failure, success	Failure (4252) and Success (1373)
16	35	Quantitative	Quarterly average employment variation rate		0.1 (1.6)

Table 3 continued

Attribute	Variables	Type	Name	Description	Mean (SD)/Most (freq.) and Least (freq.)
17	36	Quantitative	Monthly average consumer price index		93.6 (0.6)
18	37	Quantitative	Monthly average consumer confidence index		−40.5 (4.6)
19	38	Quantitative	Daily three-month Euribor rate		3.6 (1.7)
20	39	Quantitative	Quarterly average of the total number of employed citizens		5167.0 (72.2)

Table 4 Variable importance measure

	Variable importance measure
Last contact duration (seconds)	199,269
Daily three-month Euribor rate	30,896
Quarterly average of the total number of employed citizens	25,964
# Days since last contact	24,768
Age (years)	22,599
Monthly average consumer confidence index	22,470
Quarterly average employment variation rate	22,288
Monthly average consumer price index	21,564
Outcome of the previous marketing campaign (success)	19,840
Last contact month of year	19,710
Last contact day of the week	14,993
# Contacts performed during the analysis period	12,916
Credit default	12,396
Type of job (blue-collar)	9274
Contact communication type	8942
Educational level (university degree)	8922
Outcome of the previous marketing campaign (failure)	7482
# Previous contacts	7376
Marital status (single)	5690
Type of job (housemaid)	5527
Type of job (student)	5051
Type of job (technician)	4448
Educational level (Basic (4th year))	4117
Type of job (administration)	3592
Marital status (married)	2987
Educational level (professional course)	2434
Educational level (high school)	2420
Educational level (basic (9th year))	2362
Type of job (retired)	1896
Marital status (divorced)	1781
Educational level (basic (6th year))	1663
Type of job (unemployed)	0548
Housing	0184
Type of job (self-employed)	0015
Type of job (services)	−1055
Type of job (entrepreneur)	−1178
Personal loan	−1528
Type of job (management)	−2547

References

Abroud A, Choong YV, Muthaiyah S, Fie DYG (2015) Adopting e-finance: decomposing the technology acceptance model for investors. Serv Bus 9(1):161–182

Alpaydin E (2009) Introduction to machine learning, 2nd edn. The MIT Press, Cambridge

American Banker (2012) Customer analytics growing in banks. http://www.americanbanker.com/btn/25_11/customer-analytics-growing-in-banks-1053866-1.html

Amini M, Rezaeenour J, Hadavandi E (2015) A cluster-based data balancing ensemble classifier for response modeling in Bank Direct Marketing. Int J Comput Intell Appl 14(04):1550,022. doi:10.1142/S1469026815500224

Ansari A, Mela CF, Neslin SA (2008) Customer channel migration. J Mark Res 45(1):60–76. doi:10.1509/jmkr.45.1.60

Ayetiran EF, Adeyemo AB (2012) A data mining-based response model for target selection in direct marketing. IJ Inf Technol Comput Sci 1:9–18

Baesens B, Viaene S, Van den Poel D, Vanthienen J, Dedene G (2002) Bayesian neural network learning for repeat purchase modelling in direct marketing. Eur J Oper Res 138(1):191–211

Ben Ishak A (2016) Variable selection using support vector regression and random forests: a comparative study. Intell Data Anal 20(1):83–104. doi:10.3233/IDA-150795

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Burez J, Van den Poel D (2009) Handling class imbalance in customer churn prediction. Expert Syst Appl 36:4626–4636

Burton SH, Morris RG, Giraud-Carrier CG, West JH, Thackeray R (2014) Mining useful association rules from questionnaire data. Intell Data Anal 18(3):479–494. doi:10.3233/IDA-140652

Chan KY, Loh WY (2004) LOTUS: an algorithm for building accurate and comprehensible logistic regression trees. J Comput Graph Stat 13(4):826–852. doi:10.1198/106186004X13064

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16(1):321–357

Chen WC, Hsu CC, Hsu JN (2011) Optimal selection of potential customer range through the union sequential pattern by using a response model. Expert Syst Appl 38(6):7451–7461. doi:10.1016/j.eswa.2010.12.078

Chen K, Hu YH, Hsieh YC (2014) Predicting customer churn from valuable B2B customers in the logistics industry: a case study. Inf Syst e-Bus Manag 13(3):475–494. doi:10.1007/s10257-014-0264-1

Chih WH, Liou DK, Hsu LC (2014) From positive and negative cognition perspectives to explore e-shoppers real purchase behavior: an application of tricomponent attitude model. Inf Syst e-Business Manag 13(3):495–526. doi:10.1007/s10257-014-0249-0

Cohen MD (2004) Exploiting response models optimizing cross-sell and up-sell opportunities in banking. Inf Syst 29(4):327–341. doi:10.1016/j.is.2003.08.001

Direct Marketing Association (2012) What is the direct marketing association? http://www.the-dma.org/aboutdma/whatisthedma.shtml

Elsalamony H, Elsayad A (2013) Bank direct marketing based on neural network. Int J Eng Adv Technol 2(6):392–400

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) Proceedings of the thirteenth international conference on machine learning (ICML 1996), Morgan Kaufmann, pp 148–156

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. doi:10.1214/aos/1013203451

Garca S, Herrera F (2009) Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evol Comput 17(3):275–306. doi:10.1162/evco.2009.17.3.275

Garca-Pedrajas N, Ortiz-Boyer D, Garca-Pedrajas MD, Fyfe C (2012) Class imbalance methods for translation initiation site recognition. In: Garca-Pedrajas N, Herrera F, Fyfe C, Bentez JM, Ali M (eds) Trends in applied intelligent systems, no. 6096 in lecture notes in computer science. Springer, Berlin, pp 327–336

Govindarajan M (2015) Comparative study of ensemble classifiers for direct marketing. Int Dec Tech 9(2):141–152. doi:10.3233/IDT-140212

Gür Ali Ö, Aritürk U (2014) Dynamic churn prediction framework with more effective use of rare event data: the case of private banking. Expert Syst Appl 41(17):7889–7903. doi:10.1016/j.eswa.2014.06.018

Gzquez-Abad JC, Cannire MHD, Martnez-Lpez FJ (2011) Dynamics of customer response to promotional and relational direct mailings from an apparel retailer: The moderating role of relationship strength. J Retail 87(2):166–181. doi:10.1016/j.jretai.2011.03.001

Ha K, Cho S, MacLachlan D (2005) Response models based on bagging neural networks. J Interact Market 19(1):17–30. doi:10.1002/dir.20028

Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd edn. Morgan Kaufmann, Amsterdam

He H (2011) Self-adaptive systems for machine intelligence. John Wiley & Sons, New Jersey

Hosseini SY, Bideh AZ (2014) A data mining approach for segmentation-based importance-performance analysis (SOM-BPNN-IPA): a new framework for developing customer retention strategies. Serv Bus 8(2):295–312. doi:10.1007/s11628-013-0197-7

Hsieh NC (2004) An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Syst Appl 27(4):623–633. doi:10.1016/j.eswa.2004.06.007

Hu X (2005) A data mining approach for retailing bank customer attrition analysis. Appl Intell 22(1):47–60. doi:10.1023/B:APIN.0000047383.53680.b6

Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal 6(5):429–449

Jayasree V (2013) A review on data mining in banking sector. Am J Appl Sci 10(10):1160–1165. doi:10.3844/ajassp.2013.1160.1165

Jingbiao R, Shaohong Y (2010) Research and improvement of clustering algorithm in data mining. In: 2010 2nd international conference on signal processing systems (ICSPS), vol 1, pp 842–845, doi:DOIurl10.1109/ICSPS.2010.5555239

Khajvand M, Tarokh MJ (2011) Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Comput Sci 3:1327–1332. doi:10.1016/j.procs.2011.01.011

Kim G, Chae BK, Olson DL (2013) A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models. Serv Bus 7(1):167–182. doi:10.1007/s11628-012-0147-9

Lessmann S, Baesens B, Mues C, Pietsch S (2008) Benchmarking classification models for software defect prediction: a proposed framework and novel findings. IEEE Trans Softw Eng 34(4):485–496. doi:10.1109/TSE.2008.35

Li W, Wu X, Sun Y, Zhang Q (2010) Credit card customer segmentation and target marketing based on data mining. In: 2010 international conference on computational intelligence and security (CIS), pp 73–76, doi:DOIurl10.1109/CIS.2010.23

Liao SH, Chen CM, Hsieh CL, Hsiao SC (2009) Mining information users' knowledge for one-to-one marketing on information appliance. Expert Syst Appl 36(3):4967–4979. doi:10.1016/j.eswa.2008.06.020

Libana-Cabanillas F, Nogueras R, Herrera LJ, Guilln A (2013) Analysing user trust in electronic banking using data mining methods. Expert Syst Appl 40(14):5439–5447. doi:10.1016/j.eswa.2013.03.010

Ling CX, Li C (1998) Data mining for direct marketing: Problems and solutions. In: Knowledge discovery and data mining, pp 217–225

Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B 39(2):539–550. doi:10.1109/TSMCB.2008.2007853

Lu MT, Tzeng GH, Cheng H, Hsu CC (2015) Exploring mobile banking services for user behavior in intention adoption: using new hybrid MADM model. Serv Bus 9(3):541–565. doi:10.1007/s11628-014-0239-9

Mcculloch W, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5(133):115

Migueis VL, Benoit DF, Van den Poel D (2013) Enhanced decision support in credit scoring using bayesian binary quantile regression. J Oper Res Soc 64(9):1374–1383. doi:10.1057/jors.2012.116

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 62:22–31. doi:10.1016/j.dss.2014.03.001

Ngai E, Xiu L, Chau D (2009) Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst Appl 36(2, Part 2):2592–2602

Nie G, Rowe W, Zhang L, Tian Y, Shi Y (2011) Credit card churn forecasting by logistic regression and decision tree. Expert Syst Appl 38(12):15,273–15,285

Olson DL, Chae B (2012) Direct marketing decision support through predictive customer response modeling. Decis Support Syst 54(1):443–451. doi:10.1016/j.dss.2012.06.005

Olson DL, Cao Q, Gu C, Lee D (2009) Comparison of customer response models. Serv Bus 3(2):117–130. doi:10.1007/s11628-009-0064-8

Quah JTS, Sriganesh M (2008) Real-time credit card fraud detection using computational intelligence. Expert Syst Appl 35(4):1721–1732. doi:10.1016/j.eswa.2007.08.093

Ras ZW, Wieczorkowska A (2000) Action-rules: how to increase profit of a company. In: Zighed DA, Komorowski J, Zytkow J (eds) Principles of data mining and knowledge discovery, no. 1910 in lecture notes in computer science. Springer, Berlin, pp 587–592

Ratner B (2004) Statistical modeling and analysis for database marketing: effective techniques for mining big data. CRC Press, Boca Raton

Schwartz B, Lauridsen JT (2007) Scoring of bank customers for a life insurance campaign. Technical Report 5/2007, University of Southern Denmark, Denmark

Seret A, Bejinaru A, Baesens B (2015) Domain knowledge based segmentation of online banking customers. Intell Data Anal 19:163–184. doi:10.3233/IDA-150776

Srinivas K, Rao GR, Govardhan A (2014) Adapting rough-fuzzy classifier to solve class imbalance problem in heart disease prediction using FCM. Int J Med Eng Inform 6(4):297–318. doi:10.1504/IJMEI.2014.065427

Sun B, Li S, Zhou C (2006) “Adaptive” learning and “proactive” customer relationship management. J Interact Market 20(3–4):82–96. doi:10.1002/dir.20069

Sun Y, Kamel MS, Wong AKC, Wang Y (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 40(12):3358–3378. doi:10.1016/j.patcog.2007.04.009

Vapnik VN (1995) The nature of statistical learning theory. Springer, New York

Verhoef PC, Spring PN, Hoekstra JC, Leeflang PS (2003) The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. Decis Support Syst 34(4):471–481

Vriens M, Van der Scheer HR, Hoekstra JC, Bult JR (1998) Conjoint experiments for direct mail response optimization. Eur J Market 32(3/4):323–339. doi:10.1108/03090569810204625

Wang YY, Luse A, Townsend AM, Mennecke BE (2014) Understanding the moderating roles of types of recommender systems and products on customer behavioral intention to use recommender systems. Inf Syst e-Bus Manag 13(4):769–799. doi:10.1007/s10257-014-0269-9

Xiong T, Wang S, Mayers A, Monga E (2013) Personal bankruptcy prediction by mining credit card data. Expert Syst Appl 40(2):665–676. doi:10.1016/j.eswa.2012.07.072

Yeh IC, Lien CH (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst Appl 36(2, Part 1):2473–2480. doi:10.1016/j.eswa.2007.12.020

Yen SJ, Lee YS (2009) Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 36(3):5718–5727. doi:10.1016/j.eswa.2008.06.108

Zarnani A, Rahgozar M, Lucas C, Taghiyareh F (2009) Effective spatial clustering methods for optimal facility establishment. Intell Data Anal 13(1):61–84