Serv Bus (2013) 7:167–182
DOI 10.1007/s11628-012-0147-9

A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models

Gitae Kim · Bongsug Kevin Chae · David L. Olson

Received: 27 April 2012 / Accepted: 3 May 2012 / Published online: 24 May 2012

Abstract Customer response is a crucial aspect of service business. The ability to accurately predict which customer profiles are productive has proven invaluable in customer relationship management. An area that has received little attention in the literature on direct marketing is the class imbalance problem (the very low response rate). We propose a customer response predictive model approach combining recency, frequency, and monetary variables and support vector machine analysis. We have identified three sets of direct marketing data with a different degree of class imbalance (little, moderate, high) and used random undersampling method to reduce the degree of the imbalance problem. We report the empirical results in terms of gain values and prediction accuracy and the impact of random undersampling on customer response model performance. We also discuss these empirical results with the findings of previous studies and the implications for industry practice and future research.

Keywords Class imbalance · Customer response model · RFM · Support vector machine · Random undersampling · Marketing data mining

G. Kim
Department of Industrial and Manufacturing Systems Engineering, Kansas State University,
Manhattan, KS, USA
e-mail: gitaekimemail@gmail.com

B. K. Chae (✉)
Department of Management, Kansas State University, Manhattan, KS, USA
e-mail: Kevinbschae@gmail.com

D. L. Olson
Department of Management, University of Nebraska, Lincoln, NE, USA
e-mail: Dolson3@unl.edu

1 Introduction

Identification of potential donors or customers is critical for company's sales, profits, and fundraising. This area, commonly known as direct, interactive, targeted, and database marketing, has been a keen interest among practitioners and researchers in marketing and business research in the retail service sector (McCarthy and Hastak 2007). A key characteristic of this area is a heavy use and analysis of customer's transactional and behavioral data to predict customer's future behaviors (e.g., donation, purchase) (Hughes 2005). Over the past decades, academic research and industry practice of analyzing customer data has been driven by the increase of corporate databases and the advancement of data analysis techniques. As a result, in the literature there are a large number of customer response models using various statistical and machine learning techniques, including RFM score models, decision tree, neural network, and logistic regression (Bose and Chen 2009). Olson (2007) reviewed the applications of data mining in the service industry including customer relationship management (CRM).

Despite this trend, one area has received relatively less attention in the literature. This is known as the class imbalance problem. Most real-world datasets tend to be imbalanced in that when a binary variable (1 for purchase and 0 for no purchase) is considered for prediction, there is only a small percentage of 1 in those datasets. Examples include insurance fraud detection, oil spill prediction, pre-term birth prediction, system failure prediction, student dropout prediction, and disease prediction (Ling and Li 1998; Weiss 2004). In these datasets, data analysis techniques tend to be biased toward the majority cases, thus the minority cases are easily misclassified as the majority cases, leading to poor performance of predictive models. This problem is clearly present in the most real-world marketing datasets (Cui et al. 2008; Ling and Li 1998) such as customer churn prediction (Burez and Van den Poel 2009) and customer response prediction (McCarthy and Hastak 2007).

In this article, we investigate the class imbalance problem in developing customer response models in direct marketing contexts. We are particularly interested in the application of support vector machine (SVM) models using recency, frequency, and monetary (RFM) variables and the comparison of the SVM–RFM model performance with other techniques (C5, neural network, logistic regression, RFM score model). The reasons are twofold. First, SVM is a machine learning technique with a proven record of good performance in engineering and science domains and an increasing interest from the marketing community (Cui and Curry 2005). Second, while various customer-related variables have been used in developing customer response model in direct marketing, RFM variables are the commonly used in diverse models (Bose and Chen 2009). Despite some limitations (Joo et al. 2011), RFM variables have been found to be reliable for predicting customer responses in many previous studies (Baesens et al. 2002; McCarthy and Hastak 2007; Verhaert and Van den Poel 2011). They allow the development of a parsimonious and economically efficient customer response model (Hughes 2005; McCarthy and Hastak 2007).

The motivation of this article is to find a customer response model considering three important factors: classifier, variables, and data imbalance. The classifier needs to perform well in practice and to be proved to find good solution and generalization. Variables should play an important role in finding the pattern of the data. The method also must consider the data imbalance. Therefore, this article chose the SVM, RFM variables, and random undersampling.

We identify three sets of direct marketing data with a different degrees of class imbalance (little, moderate, high) and use the random undersampling method to reduce the degree of the imbalance problem in an effort of improving the model performance. Random undersampling is relatively easy to implement and also beneficial for large datasets such as most customer-related data (Drummond and Holte 2003; Khoshgoftaar et al. 2011). The results show that SVM is severely affected by class imbalance. Overall SVM performs equally with other techniques for the dataset with little class imbalance, but underperforms for moderately and highly class imbalanced datasets. This requires class balancing and random undersampling method is shown to significantly improve the overall performance of SVM. For highly class imbalanced datasets, SVM is better off with moderate undersampling. Excessive undersampling appears to be removing important majority instances from the training dataset, negatively affecting the SVM performance. Overall, logistic regression and neural network show robust performance across different datasets and different ratios of majority and minority classes. They are less benefited from random undersampling than SVM and C5 decision tree. The RFM score model, even though overall it underperforms other models, also appears to be robust across all three datasets. In the following sections, first, we briefly review some relevant literature on customer response model and the class imbalance problem. Next, we propose an approach to imbalanced direct marketing data using SVM, RFM, and random undersampling method. Then, the research design and the experimental results are presented. Finally, we discuss these results with the findings of previous studies and the implications for industry practice and future research.

2 Relevant literature

2.1 Response model

Response or scoring models, a key area in CRM research and practice, are important for both sales and membership organizations to increase profits and donations and has a long history of industry practice (Baesens et al. 2002; Hughes 2005; Verhoef et al. 2010). Various statistical and machine learning techniques have been proposed for response models (Bose and Chen 2009). Among them, the RFM model clearly stands out in terms of popularity and easy of use (Hughes 2005; McCarthy and Hastak 2007; Verhoef et al. 2003). Some limitations of RFM are well described in the literature. However, an extensive list of previous studies has used RFM either as a standalone response model or with other marketing variables (e.g., demographics, psychology) and data mining techniques for more sophisticated response models (Bose and Chen 2009; Joo et al. 2011; Verhaert and Van den Poel 2011).

Lately, companies are increasingly deluged by data and sophisticated data mining techniques are available to marketers (Ngai et al. 2009). For example, McCarty and

Hastak (2007) used both simple (RFM model) and more sophisticated data mining techniques (e.g., CHAID, logistic regression) with three variables (recency, frequency, monetary) and compared the performance of these models. Similarly, many studies have explored the potentials of other data mining techniques for response or scoring models. A study shows that neural network, decision tree, and regression are the most widely adopted techniques in CRM (Ngai et al. 2009). This trend is also found in customer segmentation or response model (Bose and Chen 2009). On other hand, the use of SVM is rare in both CRM and customer response model, with exceptions (Viaene et al. 2001).

2.2 Class imbalance

Class imbalance occurs when the number of the majority cases significantly outnumbers that of the minority cases. Most real-world data tend to contain this problem (Ling and Li 1998; Weiss 2004). Examples include the prediction of fraud insurance claim, hospitality fatality, heart failure, actual purchase after promotion, and customer churn. In these examples, researchers or marketers are interested in the minority cases of “fraud”, “fatality”, “failure”, “purchase”, and “churn”. However, these cases are rare in real-world datasets. The challenge is that statistical and machine learning techniques favor the majority cases thus are likely to predict all cases as the majority cases. This leads to the increase in prediction accuracy but gives little useful information about the class of interest (e.g., fraud, fatality).

The class imbalance problem is strongly present in most real-world marketing datasets (Bose and Chen 2009; Cui et al. 2008; Ha et al. 2005; Ling and Li 1998; Wang et al. 2005), including customer churn prediction (Burez and Van den Poel 2009) and customer response prediction (Ling and Li 1998). For example, McCarty and Hastak (2007) reported a direct marketing dataset with less than 3 % of response rate. Ling and Li (1998) used a dataset with less than 1.5 % response rate in their direct marketing study. When this imbalance is not handled properly, the result is that the model offers a poor prediction of true response rates.

Accordingly, it is essential for prediction methods to consider the problem of imbalanced data in direct marketing. Several methods have been proposed to deal with class imbalance (Weiss 2004). Two basic approaches for balancing data are the classifier change method and the data change method. The classifier change method modifies the predictor with different weights on the cost function to avoid the bias of the prediction. The data change method generates or reduces the dataset to balance the data using oversampling or undersampling (He and Garcia 2009).

3 Support vector machine

SVM is a statistical machine learning method and has been applied to various regression and classification problems such as text/face recognition, medical diagnosis, and market analysis (Schölkopf et al. 2000; Vapnik 1995). To classify non-separable dataset, the SVM nonlinearly maps the data into another or a higher dimensional space so that it can be classified by a linear separating hyperplane.

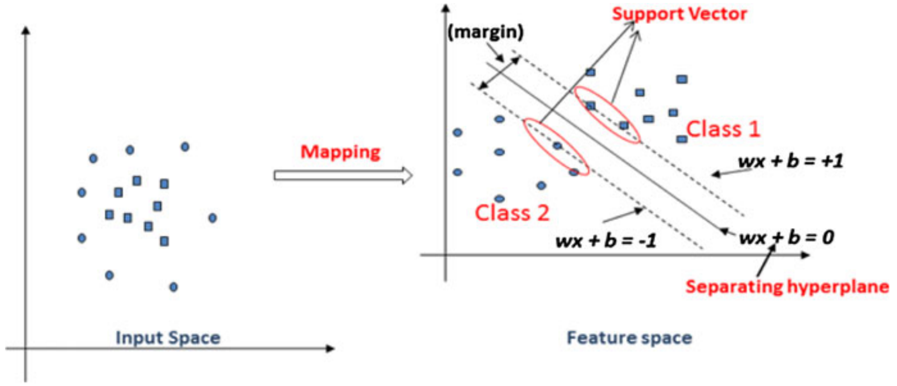


Fig. 1 SVM classification

Finding the best separating hyperplane is the basic concept of the SVM classification. The SVM finds the hyperplane that can maximally separates two classifications (Han et al. 2011). The data points that only effect to form the hyperplane are called support vectors (See Fig. 1). These support vectors are the data points from each class which are the closest to the hyperplane. Thus, they are the data points that only affect to build the classifier.

Figure 1 presents the basic concept of the SVM classification. SVMs based on the concept of the structural risk minimization (SRM) principle have become a popular predictor with many attractive features such as statistical background, good generalization, and promising performance. Unlike the empirical risk minimization (ERM) that only considers the error of data, SRM considers both how the data chosen is good and the complexity of the model. To find the best separating hyperplane in Fig. 1, one solves a convex optimization problem that is simpler than other complicated statistical models. The decision function is noted as a decision boundary, the hyperplane, in the feature space.

4 A proposed SVM approach using RFM variables

The customer data contains the information of purchase history with 91 variables of each customer. Using the general coding scheme (Hughes 2005), these raw data are converted into RFM variables. The data is randomly sampled into 80 % of training set and 20 % of testing set. Before SVM trains the data, random undersampling is used to balance the data set. SVM finds the pattern of the data and predicts the test data set.

There are several emerging advanced data mining techniques for business research in general and for customer response model in particular (Bose and Chen 2009). Among them, it is SVM that has received considerable attention recently. It shows good performance with practical classification problems in various science and engineering fields (Lessmann and Voß 2009) and nonlinear datasets (Olson and

Delen 2008) and an increasing interest from the marketing community (Cui and Curry 2005). But, there is a limited application of SVM for practical marketing problems (Cui and Curry 2005; Lessmann and Voß 2009).

The model uses well-known RFM variables (recency, frequency, monetary) as the primary predictors. In previous studies these three variables were found to be powerful enough to develop parsimonious yet reliable response models (Blattberg et al. 2008; Hughes 2005; McCarthy and Hastak 2007).

Our approach also considers the class imbalance issue. Like other data mining techniques, SVM tends to suffer from the class imbalance. In SVM, the imbalanced data makes the decision boundary to be shifted toward the minority class so that most data points are classified into the majority class, which leads to poor model performance.

As noted earlier in Sect. 2, different methods can be considered to deal with that issue (Weiss 2004). One method to avoid this problem is to modify the classifier with different weights to different class points. Another method uses sampling, either over or under sampling. Oversampling is to generate the more minority data randomly to balance data while undersampling is to reduce the majority data. However, different weighting and oversampling use equal or greater amounts of data. Thus, when the size of data is large, these two methods are less efficient. Also oversampling may result in over-fitting since the method creates the copies of the minority cases (Weiss 2004). Undersampling has been widely used for balancing the data with a smaller amount of data. While it has drawbacks, including removing potentially useful majority cases (Weiss 2004), the method has been shown to perform well with large datasets (Burez and Van den Poel 2009; Drummond and Holte 2003). Due to its simplicity with good performance, the random undersampling is used in this article to balance the data and reduces the majority class data to the same size or the double size as the minority class. A set of sample points from the training data can be randomly selected and removed to meet the proper ratio between the majority and the minority classes. The method can easily obtain the class balance in the dataset.

4.1 Research design

The proposed model is validated with three datasets (1, 2, 4) from the Direct marketing education foundation (DMEF). Dataset 1 is from a non-profit organization that relied on direct mailing to solicit contributions from past donors. The dataset includes the past behavior information of almost 100,000 donors. 27,208 responded to the direct mailing campaign so the response rate was 27.42 %. Dataset 2 is from a business that used catalogs to its customers. The dataset is with the information of 96,551 customers and only 2,371 customers responded to the catalogs (2.46 % response rate). Dataset 4 is the information of customer purchase behavior of an upscale business using catalog mailing. It contains over 100,000 customers' transactions. The response rate was 9.42 %. Given that the response rate in a typical response model dataset is very low, we describe these datasets as little (27.41 %—dataset 1), moderate (9.42 %—dataset 4), and High (2.46 %—dataset 2) in terms of the degree of class imbalance.

Each of the raw datasets contains a large number of transaction-related variables such as amounts of purchase and dates of purchase. There are several transactional records per each donor or customer, meaning multiple donations or purchases made by a single person. For each dataset, we split into 80 and 20 % for training and testing, respectively. In the effort for data balancing, we applied two random undersampling rules—2:1 (33 % of 1 or responses) and 1:1 (equal number of 1 and 0)—into the training dataset of those three datasets. This produced three training sets with different 0 and 1 ratios (original, 2:1, 1:1) per each dataset.

Data preparation was necessary. We followed the general coding scheme to develop the values of R, F, and M variables (Hughes 2005). Various data exploration and preparation techniques were employed during this phase. These RFM values were used as the input variables and the result of response (0 or 1) as the predicted value in our SVM-based customer response model. These same variables and datasets were used with other data mining techniques (DT, LR, NN). These are among the most popular data mining techniques in CRM. We do not describe them in this article since it is assumed that the reader is familiar with them.

5 Experimental results

In this section, the experimental results are presented in the light of those measures. The performance of our customer response model is measured in three effectiveness measures (accuracy, sensitivity, and specificity) as well as gain values. We also demonstrate the impact of random undersampling on customer response models in terms of those performance measures.

5.1 Performance results based on accuracy, sensitivity, and specificity

Table 1 summarizes the performance of the response models based on SVM and other data mining techniques in terms of more general measures: accuracy, sensitivity, and specificity. The summary table indicates that SVM's accuracy performance has an inverse relationship with the percentage of minority class. The SVM accuracy is 73.4 % for dataset 1 (27.4 % minority class). This accuracy is increased to 90.6 and 95.3 % for dataset 4 (9.4 % minority class) and dataset 2 (2.4 % minority class), respectively. This overall pattern is also found in other models based on DT, LR, and NN.

For dataset 1 (little class imbalance) and dataset 4 (moderate class imbalance), there is no significant difference between SVM and other response models, except the C5-based (decision tree) model's high accuracy and sensitivity for dataset 4. For example, for dataset 1 all those models produce the accuracy rates between 73 and 74 %. For dataset 2 (high class imbalance), those models based on C5, LR, and NN predict all cases as the majority class (0 or no response). This led to a high prediction rate (97.64 %) and 100 % specificity but no sensitivity at all, showing the tendency of typical response models being biased toward the majority class. It is only the SVM model that predicted some cases as the minority class (1 or response), producing a positive sensitivity rate (7.3 %) and 95.3 % accuracy rate.

Table 1 Effectiveness measures of the SVM approach and other models

Data sets	% of minority or 0 response	Accuracy (%)				Sensitivity (%)				Specificity (%)			
		SVM	DT	LR	NN	SVM	DT	LR	NN	SVM	DT	LR	NN
1	27.42 % (original)	73.4	74.0	73.7	74.0	8.3	14.9	12.5	16.8	98.1	96.5	97.0	95.8
	33 % (2:1)	73.4	73.8	73.7	73.8	13.1	28.7	18.3	21.0	96.3	91.0	94.8	93.9
	50 % (1:1)	72.4	63.1	60.2	63.6	19.7	69.1	73.5	67.9	92.5	60.8	55.1	61.9
4	9.42 % (original)	90.6	98.0	90.5	90.6	4.6	89.2	5.9	6.8	99.7	98.9	99.6	99.5
	33 % (2:1)	86.2	97.9	86.8	85.0	45.7	95.1	45.5	58.8	90.5	98.2	91.2	87.8
	50 % (1:1)	71.0	96.6	68.3	73.0	73.5	97.9	75.8	71.3	70.7	96.5	67.5	73.2
2	2.46 % (original)	95.3	*97.6	*97.6	*97.6	7.3	0	0	0	97.4	100	100	100
	33 % (2:1)	95.2	92.6	94.8	93.3	23.8	16.1	8.8	13.5	97.4	94.4	96.9	95.2
	50 % (1:1)	93.9	75.9	62.3	55.6	9.5	41.1	56.5	62.9	95.9	76.8	62.5	55.4

Note These models with asterisks predicted all as 0. Due to this degeneracy, sensitivity is 0 % and specificity is 100 %

Overall, the sensitivity (true positive rate) of all models for the original datasets is very low as compared to specificity, indicating high error rates in predicting the minority class. This also demonstrates that the minority class prediction significantly underperforms the majority class prediction. The minority class is likely to be misclassified as the majority class. All the response models are negatively affected by class imbalance.

5.2 Impact of random undersampling on accuracy, sensitivity, and specificity

Two ratios (33 and 50 %) of undersampling were used in this study. As the class imbalance is reduced through 33 and 50 % undersampling, the sensitivity (predicting actual donors or customers correctly) is increased. With the original dataset 4 (moderate class imbalance), for example, the models based on SVM, LR, and NN produce relatively low sensitivity (4.6–6.8 %). However, 33 % undersampling significantly improves the sensitivity to over 45 %. This increase of sensitivity comes with the decrease in the overall accuracy and specificity. Across all models, undersampling tends to increase sensitivity while accuracy and specificity are decreased.

For dataset 2 (high class imbalance) the impact of the moderate 33 % undersampling on those models based on DT, LR, and NN is shown by significantly improving their sensitivity from zero to 16.1 % for DT, 8.8 % for LR, 13.5 % for NN. The SVM-based model remains robust, maintaining the highest accuracy (95.2 %), sensitivity (23.8 %), and specificity (97.4 %) among all the models. Overall, 50 % (1:1) undersampling has a positive impact on the sensitivity of SVM, DT, LR, and NN. However, an exception exists with the SVM model for dataset 2 (high class imbalance): 50 % undersampling, producing the training dataset with an equal number of 0 and 1 cases (1,918 cases per each class), slightly lowered the SVM model's accuracy, sensitivity, and specificity.

5.3 Performance results based on gain values

Gain score or value is another method we used to analyze the performance of the SVM-based model and also compare it with other models. Gain chart is quite well known to marketers (Linoff and Berry 2011) and has been often used in customer response model evaluation (McCarthy and Hastak 2007; Olson et al. 2009). In this analysis, we also considered RFM Score model as another alternative response model since this relatively simple model has been with marketers for a long history of success and is also known as the most popular response model in database marketing. In practice, marketers prefer target marketing due to the high costs of mass marketing. Thus, the gain values of the first four or five deciles (10, 20, 30, 40, and 50 %) (McCarthy and Hastak 2007), rather than of the entire customer dataset, are of interest in this section.

Figure 2 shows the gain percentages of the SVM-based model for three original datasets. For example, this model for dataset 1 (little class imbalance) can capture 34 % of actual respondents from the 20 % sample, which is 14 % higher than what the random model would afford. However, SVM's performance dwindles as the class imbalance increases in datasets (26 and 23 % of actual respondents for dataset 4 and dataset 2, respectively), showing SVM is severely affected by class imbalance.

Fig. 2 Gain values of the SVM-based model for three original datasets

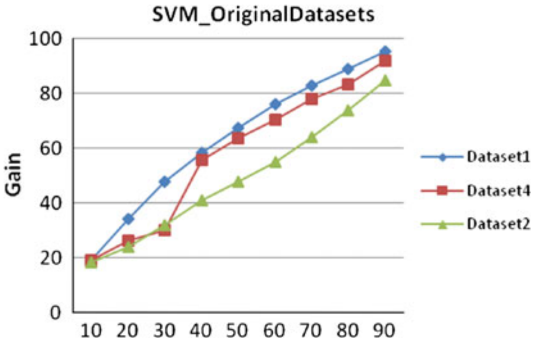


Figure 3a, b, c provides the gain values of all five models (including RFM score model) for the original class distribution of three datasets. Figure 3a shows that in the 10–40 % samples of dataset 1, there is no significant difference in the gain values of those models. However, Figs. 3b and c demonstrate that the gain performance of those models varies significantly as the datasets become moderately (dataset 4) and highly (dataset 1) imbalanced. What is noticeable is that overall SVM performs worse than other models for dataset 4 (Fig. 3b). For dataset 2, SVM outperforms C5 (DT) in the first four deciles, but underperforms all the other models (Fig. 3c).

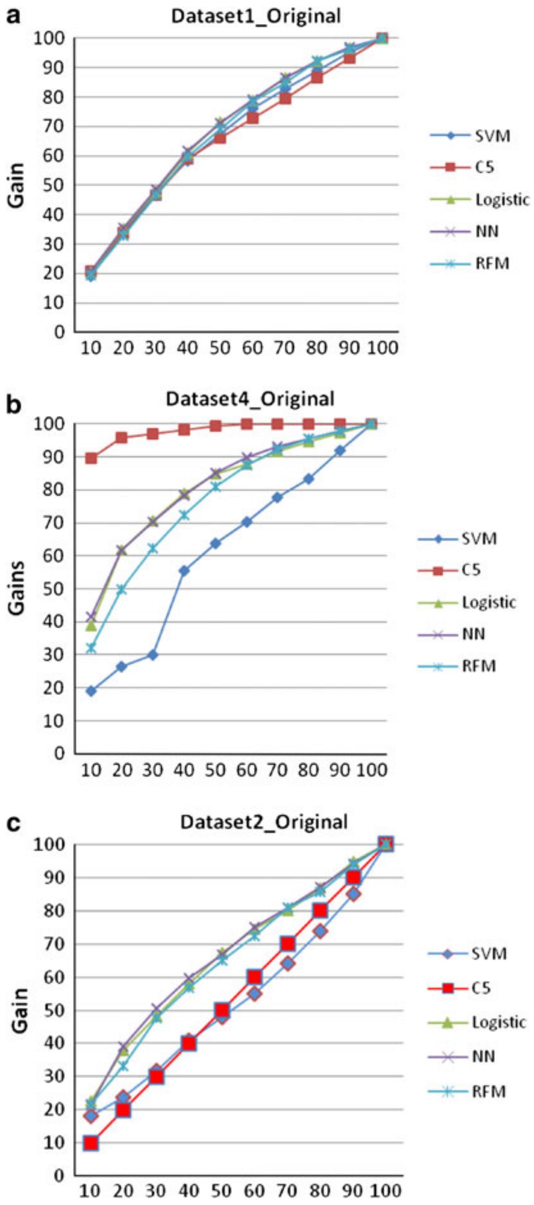
5.4 The impact of random undersampling on gain values

Figures 4a, b, and c show the impact of random undersampling on the SVM-based model's performance. Figure 4a shows no significant change in gain values from random undersampling. On the other hand, Figs. 4b and c demonstrate a significantly positive impact of random undersampling on the gain values for the SVM-based model. In the case of dataset 4, for example, if the 10 % test sample is selected for direct marketing, the model based on the original class distribution can capture 19 % of actual respondents. However, 30 and 50 % undersampling can increase the model performance so the model can capture over 36 and 45 % of actual respondents, respectively (Fig. 4b).

Figures 4b and c also show that overall there is a slight difference in the SVM-based model's performance between 33 % (2:1) and 50 % (1:1) undersampling. Figure 4b (dataset 4-moderate class imbalance) demonstrates that in the 10 % test sample there is a much improvement (about 10 % increase in gain value) from 2:1 to 1:1 undersampling; on the other hand, Fig. 4c (dataset 2-high class imbalance) shows there is a decrease (about 5 %) in gain values from 2:1 to 1:1 undersampling in the 30 % test sample. This decreasing trend is shown in other upper deciles (e.g., 40 %). These similar results are also observed with C5 (DT).

For dataset 1 (little class imbalance), Fig. 5 shows that there is no major difference in the performance of all five response models including RFM score model. However, this changes with the datasets with class imbalance. For example, the performance of those response models for the original dataset 4 and dataset 2 varies quite significantly. For the original dataset 4 (Fig. 6), SVM shows the lowest gain values, even lower than the simple model, RFM Score model. For the original

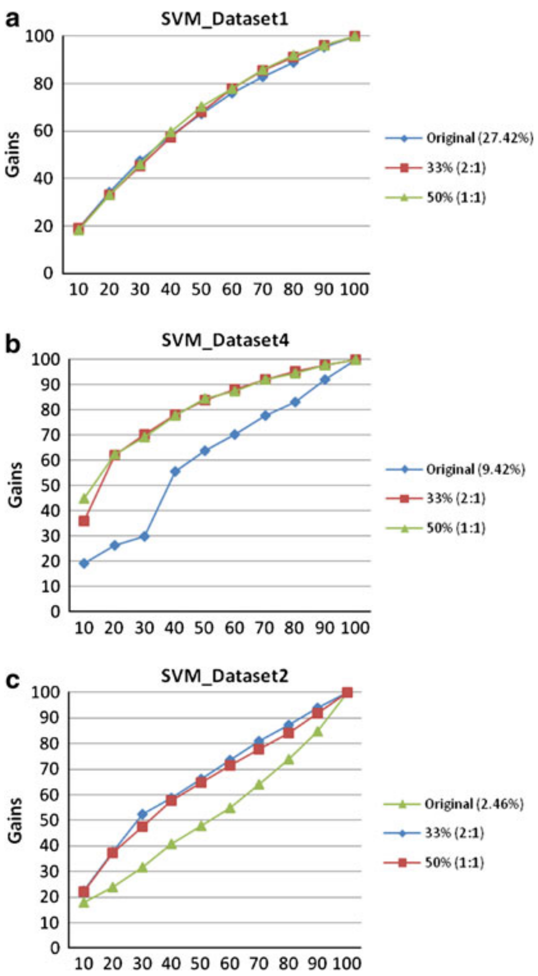
Fig. 3 **a** Gain values of five methods for dataset 1 original. **b** Gain values of five methods for dataset 4 original. **c** Gain values of five methods for dataset 2 original



dataset 2 (Fig. 7), C5 and SVM perform lower than other models. However, these performance gaps are nearly closed as random undersampling is applied. The impact of random undersampling is clear in the 30 % test sample that for 33 % dataset 2 SVM predicts 52 % of actual respondents, higher than DT (46 %), LR (49 %), NN (50 %), and RFM score model (47 %).

Overall, LR and NN show robust gain performance across different datasets and different ratios of majority and minority classes. This also means that random

Fig. 4 **a** Gain values of SVM for dataset 1. **b** Gain values of SVM for dataset 4. **c** Gain values of SVM for dataset 2



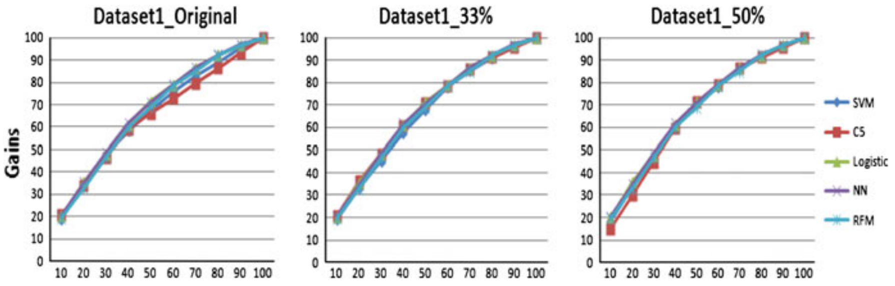


Fig. 5 Gain values of five methods for dataset 1

undersampling has a relatively less impact on the models based on LR and NN. RFM score model, even though it overall underperforms other models, appears to be robust across all three datasets. In contrast, it is shown that C5 is severely affected

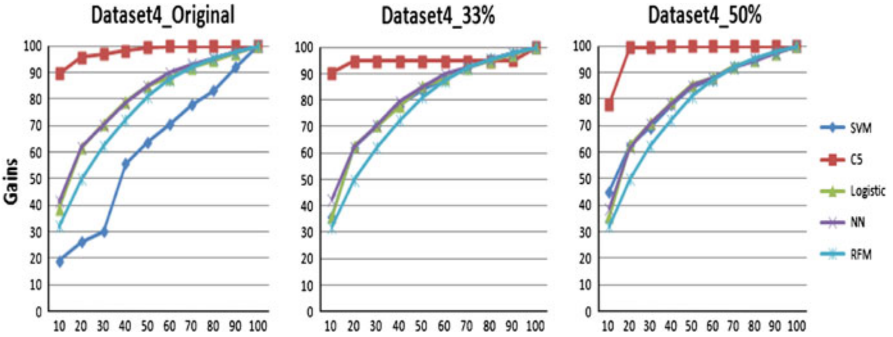


Fig. 6 Gain values of five methods for dataset 4

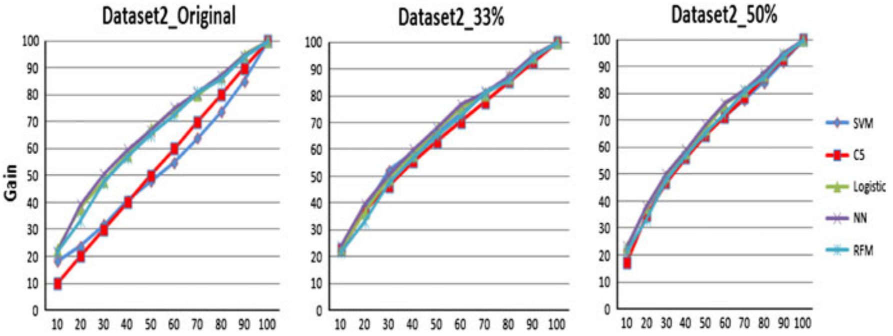


Fig. 7 Gain values of five methods for dataset 2

by high class imbalance and SVM by both moderate and high class imbalance. Thus, SVM and C5 are more benefited from random undersampling.

6 Discussion

This study has focused on three important topic areas in customer response model, including SVM and other data mining methods using popular RFM variables, class imbalance datasets, and the impact of random undersampling. We have used two kinds of performance measures to describe the experimental results, offering more findings and insights than using a single performance measure.

This study suggests that the predictive models of customer responses are affected by the class imbalance problem in general. The impact is severe as dataset is highly imbalanced (dataset 2 with only 2.4 % minority class). NN, LT, and C5 DT are shown to be biased toward majority class and thus their models predicted all cases as majority class. This led to high accuracy rates (97.6 %) but models are less useful in practice. While SVM avoids this degeneracy and produces positive sensitivity, it also suffers from class imbalance as demonstrated by gain charts.

In the case of moderate class imbalance (dataset 4), SVM performs worse than other techniques, including the simple RFM score model. In high class imbalance, SVM (and C5 DT) underperforms NN, LT, and RFM score model. Thus, class imbalance is a real issue in developing high-performing customer response model for target marketing. In general, all response models are negatively affected by high class imbalance. In particular, SVM may be more strongly affected by class imbalance than other traditional techniques. This similar finding of SVM is observed in medical research predicting heart disease (Wu et al. 2010).

On the other hand, for dataset 1 there is no significant difference in the performance of different response models, including SVM. This is demonstrated in gain charts as well as accuracy, sensitivity, and specificity. This indicates that when dataset is not affected by class imbalance, marketers could make use of relatively simple techniques (e.g., RFM, LT) for response models and these models can perform as equally as more sophisticated models.

The negative impact of class imbalance demands the rearrangement of class distribution in datasets. Random undersampling, a method for class balancing through reducing the size of majority class, has improved model performance in terms of gain values, which is the performance measure popularly used by marketers. SVM in particular and C5 DT are more positively affected from this class balancing method than NN and LR. Random undersampling is shown to be positive, but the degree of undersampling seems to be an important factor for model performance.

Two kinds of class distribution were created from two random undersampling methods: 2:1 ratio (of majority and minority class from 33 % undersampling) and 1:1 ratio (from 50 % undersampling). The result is that 1:1 class distribution is not necessarily better than 2:1 class distribution. SVM in particular shows that (1) for a dataset with moderate class imbalance, 50 % undersampling seems to outperform 30 % undersampling, (2) but the opposite is with high class imbalance. This also appears in the decrease of the SVM's sensitivity rate from 23.8 to 9.5 % as 50 % undersampling is used instead of 30 % undersampling. Then, the question is why SVM is disserved by 50 % undersampling. This seems to do with the manner SVM separates the data of two classes.

As discussed in Sect. 3, support vectors are critical since they are the data points that only are used to form the hyperplane which yields the decision function. Thus, the data points far away from the hyperplane (or non-support vectors) can be redundant information and removing them does not change the hyperplane as well as the decision function. This is one of the stronger points SVM has over other traditional data mining techniques (e.g., LR) (Cui and Curry 2005; Han et al. 2011). However, what is happening here is that 50 % undersampling could result in discarding the potentially useful majority class data for highly imbalanced datasets such as dataset 2. This is shown in the size of a new sample dataset of dataset 2 after 50 % undersampling. 50 % undersampling of dataset 4 creates a new dataset of 15314 records, which has an equal number of 1 and 0. On the other hand, 50 % undersampling of dataset 2 results in a dataset of only 3836 records. Removing such a large number of majority class seems to have eliminated useful information (Weiss 2004). Thus, it can be concluded that the degree of undersampling affects either positively or negatively model performance. Our research seems to indicate that 30 % undersampling is more helpful than 50 % undersampling when dataset is highly imbalanced.

The impact of class imbalance on LR and NN is less severe than on SVM and C5 DT. LR and NN are popularly used in predicting customer response model (Baesens et al. 2002; Bose and Chen 2009; McCarthy and Hastak 2007; Ngai et al. 2009; Verhaert and Van den Poel 2011). In terms of gain value, these techniques appear robust regardless types of datasets (e.g., ratio of majority and minority class). Thus, they are less benefited from random undersampling. Previous studies (Khoshgoftaar et al. 2010) show the similar finding that NN is relatively less affected by class imbalance than C5 DT. This seems to explain the position of LR and NN as the benchmark techniques of response model.

According to the gain values, RFM score model (the simplest approach to build customer response model) can be considered relatively reliable across different datasets with varying class imbalance. This finding coincides with the conclusion of McCarty and Hastak (2007) that RFM score model can be "an inexpensive and generally reliable procedure" (p. 661).

In the end, class imbalance is an important issue which needs a greater attention in building customer response model. Marketers need to pay a keen attention to how to deal with this data issue. Random undersampling, among potentially other approaches, could be useful in this regard. This method is computationally efficient and relatively easy to apply in business situations. The overall performance of SVM is comparable with other techniques with help of random undersampling. A possible extension of this study can consider a large number of other variables than RFM in the dataset. Several recent studies are in this trend (Joo et al. 2011; Verhaert and Van den Poel 2011). SVM is considered strong in high-dimensionality problem for comparison with other competing methods (Clarke et al. 2008). That is, SVM is better at finding the hyperplane (or the boundary separating two classes) in a dataset with a large number of variables than other methods. Thus, applying SVM into datasets with high dimensionality could offer new insights to marketers and researchers.

References

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 138:191–211

Blattberg R, Kim B, Neslin S (2008) Database marketing: analyzing and managing customers, Chapt. 2 RFM analysis. Springer, New York

Bose I, Chen X (2009) Quantitative models for direct marketing: a review from systems perspective. Eur J Oper Res 195:1–16

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

Clarke R, Ressom H, Wang A, Xuan J, Liu M, Gehan E, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev 8:37–49

Cui D, Curry D (2005) Prediction in marketing using the support vector machine. Mark Sci 24:595–615

Cui G, Wong M, Zhang G, Li L (2008) Model selection for direct marketing: performance criteria and validation methods. Mark Intell Plan 26:275–292

Drummond C, Holte R (2003) C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on learning from imbalanced data sets at the 17th international conference on machine learning. Washington, DC, pp 1–8

Ha K, Cho S, Maclachlan D (2005) Response models based on bagging neural networks. J Interactive Mark 19:17–30

Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann, San Francisco

He H, Garcia E (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21:1263–1284

Hughes A (2005) Strategic database marketing, 3rd edn. McGraw-Hill, New York

Joo Y, Kim Y, Yang S (2011) Valuing customers for social network services. J Bus Res 64:1239–1244

Khoshgoftaar T, Van Hulse J, Napolitano A (2010) Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors. IEEE Trans Neural Netw 21:813–830

Khoshgoftaar T, Van Hulse J, Napolitano A (2011) Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans Syst Man Cybern Part A 41:552–568. doi: 10.1109/Tsmca.2010.2084081

Lessmann S, Voß S (2009) A reference model for customer-centric data mining with support vector machines. Eur J Oper Res 199:520–530

Ling C, Li C (1998) Data mining for direct marketing: problems and solutions. In: Proceeding of 4th international conference on knowledge discovery and data mining (KDD'98). AAAI Press, New York, pp 73–79

Linoff G, Berry M (2011) Data mining techniques, 3rd edn. Wiley, Indianapolis

McCarthy J, Hastak M (2007) Segmentation approaches in data-mining: a comparison of RFM, CHAID, and logistic regression. J Bus Res 60:656–662

Ngai E, Xiu L, Chau D (2009) Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst Appl 36:2592–2602. doi: 10.1016/j.eswa.2008.02.021

Olson D (2007) Data mining in business services. Serv Bus 1:181–193. doi:10.1007/s11628-006-0014-7

Olson D, Delen D (2008) Advanced data mining techniques. Springer, Heidelberg

Olson D, Cao Q, Gu C, Lee D (2009) Comparison of customer response models. Serv Bus 3:117–130

Schölkopf B, Smola A, Williamson R, Bartlett P (2000) New support vector algorithms. Neural Comput 12:1207–1245

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

Verhaert G, Van den Poel D (2011) Empathy as added value in predicting donation behavior. J Bus Res 64:1288–1295

Verhoef P, Spring P, Hoekstra J, Leeflang P (2003) The commerical use of segmentation and predictive modeling techniques for database marketing in the Netherlands. Decis Support Syst 34:471–481

Verhoef P, Venkatesan R, McAlister L, Malthouse E, Krafft M, Ganesan S (2010) CRM in data-rich multichannel retailing environments: a review and future research directions. J Interactive Mark 24:121–137

Viaene S, Baesens B, Van Gestel T, Suykens J, Van den Poel D, Vanthienen J, De Moor B, Dedene G (2001) Knowledge discovery in a direct marketing case using least squares support vector machines. Int J Intell Syst 16:1023–1036

Wang K, Zhou S, Yang Q, Yeung J (2005) Mining customer value: from association rules to direct marketing. Data Min Knowl Disc 11:57–79. doi:10.1007/s10618-005-1355-x

Weiss G (2004) Mining with rarity: a unifying framework. ACM SIGKDD Explor Newsl 6:7–19

Wu J, Roy J, Stewart W (2010) Prediction modeling using EHR data. Med Care 48:S106–S113