Expert Systems with Applications 36 (2009) 5445–5449

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Customer churn prediction using improved balanced random forests

Yaya Xie [a], Xiu Li [a,*], E.W.T. Ngai [b], Weiyun Ying [c]

[a] *Department of Automation, Tsinghua University, Beijing, PR China*
[b] *Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, PR China*
[c] *School of Management, Xi'an Jiaotong University, Xi'an, PR China*

ARTICLE INFO

Keywords:
Churn prediction
Random forests
Imbalanced data

ABSTRACT

Churn prediction is becoming a major focus of banks in China who wish to retain customers by satisfying their needs under resource constraints. In churn prediction, an important yet challenging problem is the imbalance in the data distribution. In this paper, we propose a novel learning method, called improved balanced random forests (IBRF), and demonstrate its application to churn prediction. We investigate the effectiveness of the standard random forests approach in predicting customer churn, while also integrating sampling techniques and cost-sensitive learning into the approach to achieve a better performance than most existing algorithms. The nature of IBRF is that the best features are iteratively learned by altering the class distribution and by putting higher penalties on misclassification of the minority class. We apply the method to a real bank customer churn data set. It is found to improve prediction accuracy significantly compared with other algorithms, such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM). Moreover, IBRF also produces better prediction results than other random forests algorithms such as balanced random forests and weighted random forests.

1. Introduction

Customer churn, which is defined as the propensity of customers to cease doing business with a company in a given time period, has become a significant problem and is one of the prime challenges many companies worldwide are having to face (Chandar, Laha, & Krishna, 2006).

In order to survive in an increasingly competitive marketplace, many companies are turning to data mining techniques for churn analysis. A number of studies using various algorithms, such as sequential patterns (Chiang, Wang, Lee, & Lin, 2003), genetic modeling (Eiben, Koudijs, & Slisser, 1998), classification trees (Lemmens & Croux, 2003), neural networks (Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000), and SVM (Zhao, Li, Li, Liu, & Ren, 2005), have been conducted to explore customer churn and to demonstrate the potential of data mining through experiments and case studies.

However, the existing algorithms for churn analysis still have some limitations because of the specific nature of the churn prediction problem. This has three major characteristics: (1) the data is usually imbalanced; that is, the number of churn customers constitutes only a very small minority of the data (usually 2% of the total samples) (Zhao et al., 2005); (2) large learning applications will inevitably have some type of noise in the data (Shah, 1996); and (3) the task of predicting churn requires the ranking of subscribers according to their likelihood to churn (Au, Chan, & Yao, 2003).

Several approaches have been proposed to address this problem. Decision-tree-based algorithms can be extended to determine the ranking, but it is possible that some leaves in a decision tree have similar class probabilities and the approach is vulnerable to noise. The neural network algorithm does not explicitly express the uncovered patterns in a symbolic, easily understandable way. Genetic algorithms can produce accurate predictive models, but they cannot determine the likelihood associated with their predictions. These problems prevent the above techniques from being applicable to the churn prediction problem (Au et al., 2003). Some other methods, such as the Bayesian multi-net classifier (Luo & Mu, 2004), SVM, sequential patterns, and survival analysis (Lariviere & Van den Poel, 2004), have made good attempts to predict churn, but the error rates are still unsatisfactory.

In response to these limitations of existing algorithms, we present an improved balanced random forests (IBRF) method in this study. To the best of our knowledge, only a few implementations of random forests (Breiman, 2001) in a customer churn environment have been published (Buckinx & Van den Poel, 2005; Burez & Van den Poel, 2007; Coussement & Van den Poel, 2008; Lariviere & Van den Poel, 2005). Our study contributes to the existing literature not only by investigating the effectiveness of the random forests approach in predicting customer churn but also by integrating sampling techniques and cost-sensitive learning into random forests to achieve better performance than existing algorithms.

* Corresponding author. Tel.: +86 10 62771152.
E-mail address: lixiu@tsinghua.edu.cn (X. Li).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.121

The proposed method incorporates both sampling techniques and cost-sensitive learning, which are two common approaches to tackle the problem of imbalanced data. By introducing "interval variables", these two approaches alter the class distribution and put heavier penalties on misclassification of the minority class. The interval variables determine the distribution of samples in different iterations to maintain the randomicity of the sample selection, which results in a higher noise tolerance. It allows ineffective and unstable weak classifiers to learn based on both an appropriate discriminant measure and a more balanced dataset. It therefore can achieve a more precise prediction.

To test our proposed method, we apply it to predict churn in the banking industry. Banks are thought to be appropriate because extensive customer behavior data are available which enable the prediction of future customer behavior. Moreover, data can be easily collected. Although our study is limited to a specific bank, the method can be applied to many other service industries as well as to various engineering applications.

The remainder of this paper is structured as follows. In Section 2, we explain the methodological underpinnings of random forests. The dataset preparation and various experiments using IBRF and their results are presented in Sections 3 and 4. Some concluding remarks and ideas for future work are given in Section 5.

2. Methodology

In this study, we use the IBRF technique to predict customers' churn behavior. In this section, we present the methodological underpinnings of the technique and the evaluation criteria we use to analyze the performance of the method.

2.1. Random forests and their extensions

The random forests method, introduced by Breiman (2001), adds an additional layer of randomness to bootstrap aggregating ("bagging") and is found to perform very well compared to many other classifiers. It is robust against overfitting and very user-friendly (Liaw & Wiener, 2002).

The strategy of random forests is to select randomly subsets of m_{try} descriptors to grow trees, each tree being grown on a bootstrap sample of the training set. This number, m_{try}, is used to split the nodes and is much smaller than the total number of descriptors available for analysis (Lariviere & Van den Poel, 2005). Because each tree depends on the values of an independently sampled random vector and standard random forests are a combination of tree predictors with the same distribution for all trees in the forest (Breiman, 2001), the standard random forests do not work well on datasets where data is extremely unbalanced, such as the dataset of customer churn prediction.

Chen et al. (2004) proposed two ways to handle the imbalanced data classification problem of random forests. The first, weighted random forests, is based on cost-sensitive learning; and the second, balanced random forests, is based on a sampling technique. Both methods improve the prediction accuracy of the minority class, and perform better than the existing algorithms (Chen et al., 2004). However, they also have their limitations.

Weighted random forests assign a weight to each class, and the minority class is given a larger weight. Thus, they penalize misclassifications of the minority class more heavily. Weighted random forests are computationally less efficient with large imbalanced data, since they need to use the entire training set. In addition, assigning a weight to the minority class may make the method more vulnerable to noise (mislabeled class) (Chen et al., 2004).

Balanced random forests artificially make class priors equal by over-sampling the minority class in learning extremely imbalanced data. However, changing the balance of negative and positive training samples has little effect on the classifiers produced by decision-tree learning methods (Elkan, 2001). In summary, according to Chen et al.'s study (2004), there is no clear winner between weighted random forests and balanced random forests.

2.2. Improved balanced random forests

We propose improved balanced random forests by combining balanced random forests and weighted random forests. On one hand, the sampling technique which is employed in balanced random forests is computationally more efficient with large imbalanced data and more noise tolerant. On the other hand, the cost-sensitive learning used in weighted random forests has more effect on the classifiers produced by decision-tree learning methods.

To combine these two methods, we introduce two "interval variables" m and d, where m is the middle point of an interval and d is the length of the interval. A distribution variable α is randomly generated between $m - d/2$ and $m + d/2$, which directly determines the distribution of samples from different classes for one iteration. The main reason for introducing these variables is to maintain the random distribution of different classes for each iteration, which results in higher noise tolerance. By contrast, balanced random forests draw the same number of samples from both the majority and minority class so that the classes are represented equally in each tree. Thus, the different classes are no longer randomly distributed across different iterations, making the method more vulnerable to noise.

The algorithm takes as input a training set $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i, i = 1, \ldots, n$ is a vector of descriptors and Y_i is the corresponding class label. The training set is then split into two subsets D^+ and D^-, the first of which consists of all positive training samples and the second of all negative samples. Let $h_t : X \mapsto \mathbb{R}$ denote a weak hypothesis.

The steps of IBRF algorithm are

- **Input**: Training examples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$; interval variables m and d; the number of trees to grow n_{tree};
- **For** $t = 1, \ldots, n_{\text{tree}}$:
  - Randomly generate a variable α within the interval between $m - d/2$ and $m + d/2$;
  - Randomly draw $n\alpha$ sample with replacement from the negative training dataset D^- and $n(1-\alpha)$ sample with replacement from the positive training dataset D^+;
  - Assign w_1 to the negative class and w_2 to the positive class where $w_1 = 1 - \alpha$ and $w_2 = \alpha$;
  - Grow an unpruned classification tree;
- **Output** the final ranking. Order all test samples by the negative score of each sample. The negative score of each sample can be considered to be the total number of trees which predict the sample to be negative. The more trees predict the sample to be negative, the higher negative score the sample gets.

In this method, the normalized vote for class j at X_i equals

$$\frac{\sum_k I(h(X_i) = j) w_k}{\sum_k w_k}. \tag{1}$$

In addition, we grow an unpruned classification tree with the following modification: at each node, rather than searching through all descriptors for the optimal split, only randomly sample m_{try} of the descriptors and choose the best split. Here m_{try} is the number of input descriptors which is used to split on at each node. The error of this method is

$$\varepsilon = \Pr_{i \sim n_{\text{tree}\, i}}[h_t(X_i) \neq Y_i]. \tag{2}$$

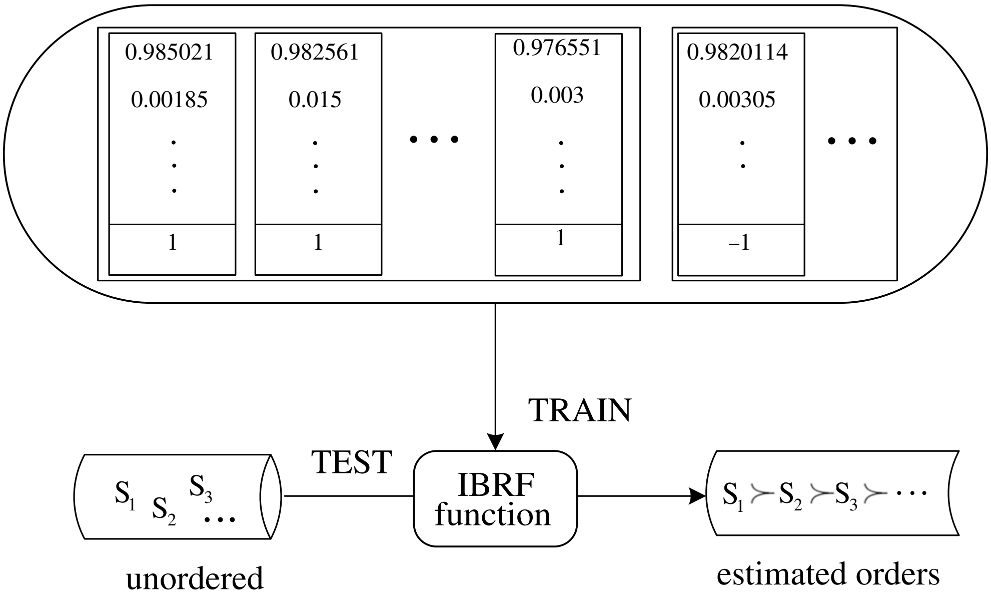

Fig. 1. Framework of IBRF.eps.

The framework of IBRF is shown in Fig. 1. Let S_i denote the testing sample. The training inputs of the ranking problem are samples from D^+ and D^- provided with the information that the negative samples should be ranked higher than positive ones. The samples which are most prone to churn are ranked higher in output.

An important question is how to determine the values of "interval variables" m and d. Variable α, which determines the balance of negative and positive training samples in each iteration, directly depends on m and d. The larger α is the greater the number of negative samples in the training set, which means the larger the weight we assign to the negative class. In our experiment with IBRF, we set the initial value of m to be 0.1 and d to be 0.1, then searched the values of m and d for the best performance. In more detail, we first held the value of m and searched for the optimal value of d using a fix step length 0.1. Then we change the value of m using a fix step length 0.1 and searched d again continuously until $d = 1$ and $m = 1$. As a result, we find that the results are not sensitive to the values of m and d.

2.3. Evaluation criteria

In this study, we use the lift curve and top-decile lift to quantify the accuracy of the predictive model.

The lift curve is preferred for evaluating and comparing model performance in customer churn analysis. For a given churn probability threshold, the lift curve plots the fraction of all subscribers above the threshold against the fraction of all churners above the threshold. It is related to the ROC (relative operating characteristic) curve of signal detection theory and the precision-recall curve in the information retrieval literature (Mozer, Wolniewicz, & Grimes, 2000). The lift curve indicates the fraction of all churners who could be included if a certain fraction of all subscribers were to be contacted (Zhao et al., 2005).

Another measurement we use is top-decile lift. This is the percentage of the 10% of customers predicted to be most likely to churn who actually churned, divided by the baseline churn rate. The higher the lift, the more accurate the model is and, intuitively, the more profitable a targeted proactive churn management program will be (Neslin, Gupta, Kamakura, Lu, & Mason, 2004). It demonstrates the model's power to beat the average performance or random model.

3. Empirical study

To evaluate the performance of the proposed method, IBRF, we apply it to a real-world database. A major Chinese bank provided the database for this study. The data set, as extracted from the bank's data warehouse, included records of more than 20,000 customers described by 27 variables.

We remove descriptors that obviously have nothing to do with the prediction, such as identification card number. Descriptors with too many missing values (more than 30% missing) are also removed. Fifteen descriptors remain after these two operations. We explore three major descriptor categories that encompass our input potential explanatory descriptors. The three categories are personal demographics, account level, and customer behavior. They are identified as follows:

(1) Personal demographics is the geographic and population data of a given customer or, more generally, information about a group living in a particular area.
(2) Account level is the billing system including contract charges, sales charges, mortality, and expense risk charges.
(3) Customer behavior is any behavior related to a customer's bank account.

In this study, the descriptors that we consider in the personal demographics category include age, education, size of disposable income, employment type, marital status, number of dependants, and service grade. The account level category includes account type, guarantee type, length of maturity of loan, loan data and loan amount. Finally, the customer behavior category includes account status, credit status, and the number of times the terms of an agreement have been broken.

The distribution of the data used in training and testing is shown in Table 1. A total of 1524 samples (762 examples for the training dataset and 762 examples for the testing dataset) are randomly selected from the dataset, which consists of 20,000 customers. There are a total of 73 potential churners in the selected samples.

Although many authors emphasize the need for a balanced training sample in order to differentiate reliably between churners and nonchurners (Dekimpe & Degraeve, 1997; Rust & Metters,

Table 1
Distribution of the data used in the training and the test in the simulation

Training dataset	
Number of normal examples	724
Number of churn examples	38
Testing dataset	
Number of normal examples	727
Number of churn examples	35

1996), we use a training dataset which contains a proportion of churners that is representative of the true population to approximate the predictive performance in a real-life situation. We construct all variables in the same way for each dataset.

4. Findings

We apply IBRF to a set of churn data in a bank as described above. To test the performance of our proposed method, we run several comparative experiments. A comparison of results from IBRF and other standard methods, namely artificial neural network (ANN), decision tree (DT), and CWC-SVM (Scholkopf, Platt, Shawe, Smola, & Williamson, 1999), is shown in Table 2 and Fig. 2.

One can observe a significantly better performance for IBRF in Table 2 and Fig. 2. To evaluate the performance of the novel approach further, we also compare our method with other random forests algorithms, namely balanced random forests and weighted random forests. Fig. 3 is the cumulative lift gain chart for identifying the customers who churned when $n_{\text{tree}} = 50$.

As Fig. 3 indicates, discriminability is clearly higher for IBRF than for the other algorithms. The chart shows that the top-decile lift captures about 88% of churners, while the top-four-decile lift capture 100% of churners.

As mentioned in Section 2.3, one would expect to have to determine the values of "interval variables" m and d that give the best performance. However, the results turn out to be insensitive to the values of these two variables. Fig. 4 shows the results when we vary the value of m and set $d = 0.1$. Fig. 5 shows the results when we vary the value of d and set $m = 0.5$.

Note that when $d = 0.1$ and $m = 0$, which means that almost all the training samples are selected from the positive training dataset, the performance of the proposed IBRF drops sharply. The same happens when $d = 0.1$ and $m = 1$, meaning that almost all the training samples are selected from the negative training dataset. Furthermore, when $m = 0.5$, the performance of IBRF drops with increasing d, especially when the value of d is close to 1. We conclude that IBRF achieves better performance than other algorithms

Table 2
Experimental results of different algorithms

Algorithm	ANN	DT	CWC-SVM	IBRF
Accuracy rate	78.1%	62.0%	87.2%	93.2%
Top-decile lift	2.6	3.2	3.5	7.1

ANN, artificial neural network; DT, decision tree; IBRF, improved balanced random forests.

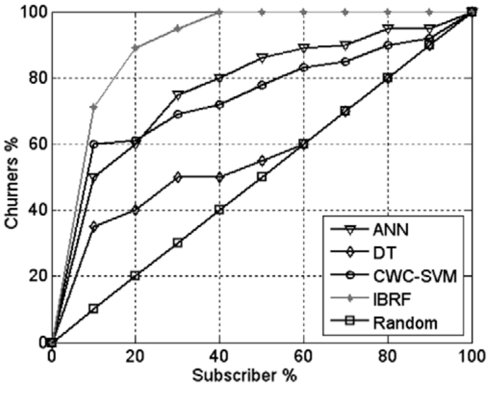

Fig. 2. Lift curve of different algorithms.eps.

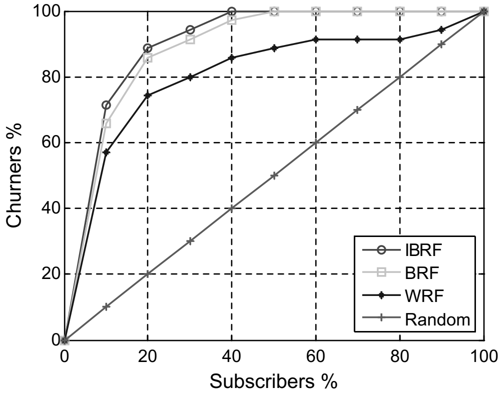

Fig. 3. Lift curve of different random forests algorithms.eps.

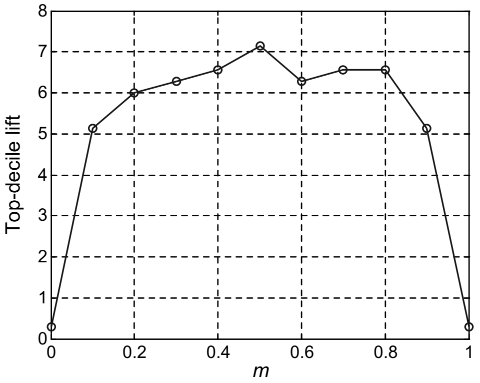

Fig. 4. Experiment result when d equals 0.1.eps.

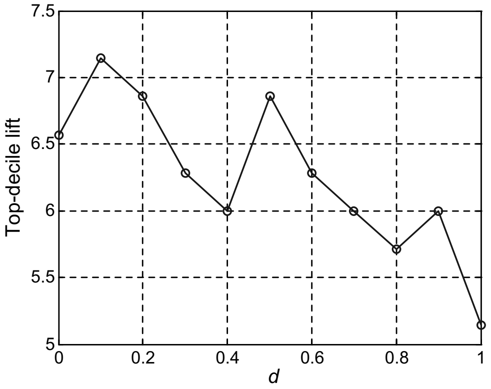

Fig. 5. Experiment result when m equals 0.5.eps

because the distribution of different classes in training dataset for each iteration is designed to be relative balance.

5. Conclusion and future work

In this paper, we propose a novel method called IBRF to predict churn in the banking industry. IBRF has advantages in that it combines sampling techniques with cost-sensitive learning to both alter the class distribution and penalize more heavily misclassification of the minority class. The best features are iteratively learned by artificially making class priors equal, based on which best weak classifiers are derived.

Experimental results on bank databases have shown that our method produces higher accuracy than other random forests algorithms such as balanced random forests and weighted random forests. In addition, the top-decile lift of IBRF is better than that of ANN, DT, and CWC-SVM. IBRF offers great potential compared to traditional approaches due to its scalability, and faster training and running speeds.

Continuing research should aim at improving the effectiveness and generalization ability. IBRF employs internal variables to determine the distribution of samples. Although the results are found to be insensitive to the values of these variables, imposing some limitations on them in future experiments may enhance the predictive effectiveness of the method. In addition, there is further research potential in the inquiry into the cost-effectiveness of this method. Experimenting with some other weak learners in random forests represents an interesting direction for further research. However, considering the large potential number of time-varying variables and the computation time and memory requirements, inappropriately chosen weak learners would not be cost-effective. Moreover, churning is not restricted to the banking industry but is also of great concern to other industries, suggesting interesting directions for future research with IBRF.

Acknowledgements

This research was partly supported by National Natural Science Foundation of China (NSFC, Project No.: 70671059) and The Hong Kong Polytechnic University (Project No.: G-YF02).

References

Au, W. H., Chan, K., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation, 7*(6), 532–545.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviorally-loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 154*(2), 508–523.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications, 32*(2), 277–288.

Chandar, M., Laha, A., & Krishna, P. (2006). Modeling churn behavior of bank customers using predictive data mining techniques. In *National conference on soft computing techniques for engineering applications (SCT-2006)*, March 24–26, 2006.

Chen, C., Liaw, A., & Breiman L. (2004). Using random forests to learn imbalanced data, Technical Report 666. Statistics Department of University of California at Berkeley.

Chiang, D., Wang, Y., Lee, S., & Lin, C. (2003). Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications, 25*(3), 293–302.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327.

Dekimpe, M. G., & Degraeve, Z. (1997). The attrition of volunteers. *European Journal of Operational Research, 98*(1), 37–51.

Eiben, A. E., Koudijs, A. E., & Slisser, F. (1998). Genetic modelling of customer retention. *Lecture Notes in Computer Science, 1391*, 178–186.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th international joint conference on artificial intelligence* (pp. 973–978).

Lariviere, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications, 27*, 277–285.

Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*, 472–484.

Lemmens, A., & Croux, C. (2003). Bagging and boosting classification trees to predict churn. DTEW Research Report 0361.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *The Newsletter of the R. Project, 2*(3), 18–22.

Luo, N., & Mu, Z. (2004). Bayesian network classifier and its application in CRM. *Computer Application, 24*(3), 79–81.

Mozer, M. C., Wolniewicz, R., & Grimes, D. B. (2000). Churn reduction in the wireless industry. *Advances in Neural Information Processing Systems, 12*, 935–941.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunication industry. *IEEE Transactions on Neural Networks, Special issue on Data Mining and Knowledge Representation*, 690–696.

Neslin, S., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2004). Defection detection: improving predictive accuracy of customer churn models. Working Paper, Teradata Center at Duke University.

Rust, R. T., & Metters, R. (1996). Mathematical models of service. *European Journal of Operational Research, 91*(3), 427–439.

Scholkopf, B., Platt, J. C., Shawe, J. T., Smola, A. J., & Williamson, R. C. (1999). Estimation the support of a high-dimensional distribution. Microsoft Research, Technical Report MSR-TR-99-87.

Shah, T. (1996). Putting a quality edge to digital wireless networks. *Cellular Business, 13*, 82–90.

Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. *Lecture Notes in Computer Science, 3584*, 300–306.