Churning of Bank Customers Using Supervised Learning

Chapter · January 2020

DOI: 10.1007/978-981-15-3172-9_64

3 authors:

Hemlata Dalmia
Sreyas Institute Of Engineering and Technology
13 PUBLICATIONS **64** CITATIONS
SEE PROFILE

Ch V S S Nikil
4 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Sandeep Kumar
Sreyas Institute Of Engineering and Technology
102 PUBLICATIONS **2,295** CITATIONS
SEE PROFILE

Churning of Bank Customers Using Supervised Learning

Hemlata Dalmia, Ch V S S Nikil and Sandeep Kumar

Abstract In the current challenging era, there is prominent competition in bank industry. To improve quality and level of service, bank concentrates on customer retention as well as customer churning. This paper discusses the classification problem of banking industry. It focuses on the customers of a bank concerns towards churning, predicting the departing customers from potential customers. Machine learning is the cutting edge technology that is practical and handy to solve such problems. Using supervised machine learning, a proprietary algorithm (a typical machine learning model) is created to forecast and inform the bank about the customers who are at the highest risk in leaving the bank. A customer churn prediction can be used here as churn and nonchurn customers are to be defined. Using ML, gap is to be resolved between churn and nonchurn customers. Different accuracy levels are achieved by classifiers using different data sheets. A novel approach K-nearest neighbor algorithm (KNN) is presented in which dataset is suitably grouped into training and testing models depending on weighted scales along with XGBooster algorithm for high and improved accuracy.

Keywords Customer churning · Machine learning · XGBooster · KNN

1 Introduction

Customers are surrounded by number of resources of information in today's digitized environment, and they have all resources at the tip of their fingers. Smartphones, e.g., provide instant access to various branded products, mbanking, comparative information. And customer perspectives demand based on advanced technology,

H. Dalmia (✉) · S. Kumar
ECE Department, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India
e-mail: dalmiahemlata@gmail.com

S. Kumar
e-mail: er.sandeepsahratia@gmail.com

Ch V S S Nikil
Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India
e-mail: saisainikil@gmail.com

H. S. Saini et al. (eds.), *Innovations in Electronics and Communication Engineering*,
Lecture Notes in Networks and Systems 107,
https://doi.org/10.1007/978-981-15-3172-9_64

convenience reasons, price sensitivity, service quality reasons and socio factors [1]. Due to availability of different options as a boon of advanced technologies, customers keep on changing from one service provider to another. That is why for companies, it is difficult to retain and attract the customers and loses their wealth due to switching action by their customers. The procedure of customers leaving their service providers is called churn [2].

For banking sector, this customer churn prediction [3] is the serious issue and gargantuan impact on the profit line of bankers. Thus, customer retention scheme can be targeted on high-risk customers who wish to discontinue their custom and switch to another competitor. To minimize the cost of bank sectors customer retention marketing scheme [4], an accurate and prior identification of these customers is hypercritical.

Customer churning [5, 6] is the estimate or analysis of degree of customers who turn to shift to an alternative. It is the most common problem witnessed in any industry. Banking is one such industry that focuses a lot on customer's behavior by tracking their activities. It is very extortionate to add a new customer to the bank when compared to retention [7]. Companies can raise their profits by handling these customers. Hence, there is a need to keep up the existing customers, which will be achieved only by understanding the customer's grievances of changing the bank. The paper presents a model to churn the bank customers using *k*-nearest neighbor (KNN) algorithm. This simple KNN algorithm is used to classify the customers into two classes, those who will leave the bank and those who will not leave. To enhance the accuracy, XGBooster algorithm is applied, whereas many research papers are available from various journals based on bank customer churn prediction, but techniques applied are decision tree [8], logistic regression [9], random forest [10], unsupervised learning [11], artificial neural network (ANN), data mining, [12, 13] neurocomputing [14]. Next section explains the literature survey based on different algorithms used in various papers.

2 Related Work

2.1 Literature Review

From the above discussion, it is clear that customer retention is important for a company and for its business strategy. Customer churning becomes business intelligence to know which customers will shift or who will get retained. To achieve customer churning, companies started adapting machine learning techniques for customer churn prediction models. In this section, a few techniques are compared considering churn prediction.

Data mining by author 'Sen K' aims to analyze large dataset by converting the sets of data into useful data. And a customer churn prediction model is developed and is measured using accuracy, sensitivity and specificity and Kappa's statistics [15].

The support vector machine (SVM) is the popular technique providing guide to the bank for customer strategy. SVM has larger probability of customer churns in the samples. With good number of plenty vectors, SVM provides good precision in predicting technique models. SVM gives high fitting accuracy rate of 0.59 by the author 'Zhao Jing' [16]

'Guoxun Wang' focuses on the comparison of all techniques used to build credit card holder churn model for the banks in China based on multi-criteria decision algorithm and constructing techniques using PROMETHEE and TOPSIS methods [17]. In MCDM algorithm, decision tree methods are implemented. 'Shaoying Cui' presents [18] improved FCM algorithm as data mining algorithm to facilitate the banks with a new idea for predicting customer churn. It achieved accuracy rate of 80% for high-value customers and 83% for low-value customers. 'Pradeep B' proposed to construct a model for churn prediction for a company using logistic regression and decision trees techniques. In Pradeep's approach there is a trial to retrieve the important factors of the customer churn that provides additional and useful knowledge which supports decision making [19].

Alisa Bilal Zorić applied a data mining technique 'neural network' in the software package ANN to predict churn in bank customer. Using this model, the reason of customer leaving the bank can be easily acquainted by entering the parameters [20].

'Abinash Mishra' proposed methodology of ensemble classifiers comprising bagging, boosting and random forest [21] to predict customer churn for telecom industry. Random forest achieves high accuracy of 96% with low specificity and high sensitivity and low error rate [21].

'Ning Lu' presented a paper in which an experimental evaluation proves that the boosting provides a good source of churn data, efficiently providing the customer churn model. The measures for churn prediction are calculated using a training set of customers over a period of six months [22].

'Hend Sayed' presented a methodology of decision tree in which two packages ML and MLib were conducted, to evaluate accuracy, model training and model evaluation. They got effective result with ML package [23].

2.2 Data Acquisition

Dataset used for this supervised prediction is acquired from an online source. The target dataset is subjected to churning of customers of bank containing information about 10,000 customers with 14 features for each customer. The customers of the bank are identified as churn or loyal based on the potential features like credit score, age, gender, estimated salary, etc. A user of the bank is classified as loyal if he/she is active and remains with the bank. Customers are classified as churners if they switch to another bank. The variable exited in the dataset gives the actual status of the customer if he/she had switched to another bank.

2.3 Data Preprocessing

The process to identify the required independent variables for predicting the exit status of a customer and to predict the binary dependent variable 'EXITED' using the independent variables is data preprocessing. The dataset used for predicting churning of customers of a bank contains information about 10,000 customers with 14 features for each customer. These features include row number, customer id, surname, credit score, geography, gender, age, tenure, balance, number of products, has cr card, is active member, estimated salary, exited [24].

To predict the churning of customers, dataset is split suitable for training and testing. At this instance, splitting has 80% training rate and 20% testing rate (Table 1).

The value of this attribute will be 1 if the customer has left the bank and 0 if remained there.

Feature scaling or data normalization is a technique used to standardize the range of independent variables in the dataset.

3 Methodology

In this paper, whole focus is using flexible technique to boost the accuracy in customer churning process. So, along with K-nearest neighbors (KNN) algorithm, XGBoost algorithm is implemented. The block diagram is represented below to describe the whole process (Fig. 1).

Table 1 Utilization of features of dataset

Serial No.	Features	Utilisation
1.	RowNumber	Unused attribute
2.	CustomerId	Unused attribute
3.	Surname	Unused attribute
4.	Credit score	Unused attribute
5.	Geography	Used as input after encoding
6.	Gender	Used as input after encoding
7.	Age	Used as input
8.	Tenure	Used as input
9.	Balance	Used as input
10.	NumOfProducts	Used as input
11.	HasCrCard	Used as input
12.	IsActiveMember	Used as input
13.	EstimatedSalary	Used as input
14.	Exited	Used as target

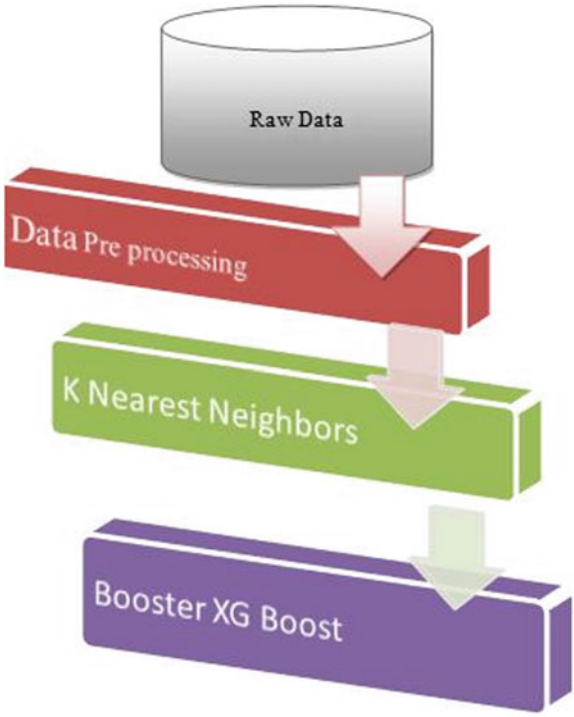


Fig. 1 Block diagram to build customer churn model

3.1 K-Nearest Neighbors

K-nearest neighbors is a machine learning and data mining algorithm used to address classification and regression problems. It uses Euclidean distance to find the similarity between the classes.

3.2 Boosting

The idea of using Boosting algorithm for customer churn prediction is to train a series of classifier simultaneously and keep updating the model accuracy for improving the performance of the classifier.

3.3 XGBoost

XGBoost is extreme gradient boosting. It is the mostly used predominant and advanced algorithm to solve machine learning problems. It dispenses support in most

of the developing environments (C++, python, R). It is mainly used for increased performance and high speed.

Installing XGBoost: To implement XGBoost model, use XGBoost function from XGBoost package in RStudio by importing the package. Fit the model to the data and predict the churning customers using the function XGBoost by supplying the training dataset, dependent variable and the number of iterations to which the classifier is to be trained. Use the *k*-fold cross-validation technique to find the average accuracy of XGBoost classifier.

The customers are classified into two classes those who will leave the bank and who those who will not leave the bank; we apply an effective algorithm and predict the churning customers whose probability of leaving the bank is very high.

3.4 KNN Algorithm

Step 1: The dataset is imported and preprocessed. Preprocessing is needed to get good quality results. The dataset is split suitable for training and testing. We have 80% for training and 20% testing.

Step 2: Training data is fitted to the KNN classifier, and the exit status of customer is predicted as follows:

y_pred = knn (train = training set[, −11], test = test

Step 3: Step 3: Find the accuracy obtained using the KNN classifier. The K value can be identified by checking with multiple values. Accuracy is improved by tuning algorithm with different K values.

_set [, −11], cl = training set [, 11], $k = 5$, prob = TRUE)

The working of K-nearest neighbor is explained as follows:

To find the exit status of a customer (record) X, the Euclidean distance of the record with respect to all the other records X_i $i = \{1, 2, 3 \ldots n\}$ is calculated using the following equation

$$\text{Euclidean distance} = \text{Sqrt}\left(\sum (X_i - X)^2\right), \quad i = \{1, 2, 3 \ldots n\}$$

The distances obtained are arranged in ascending order. And the first K distances from the obtained distances ($K > 0$) are selected. The records (points) corresponding to the distances are identified, and the exit status for each of the records is observed. The exit status of record X based on majority voting has to be evaluated at last.

3.5 XGBoost Algorithm

Step 1: XGBoost classifier is installed from XGBoost package and fitted to the training set by specifying the maximum number of iterations.

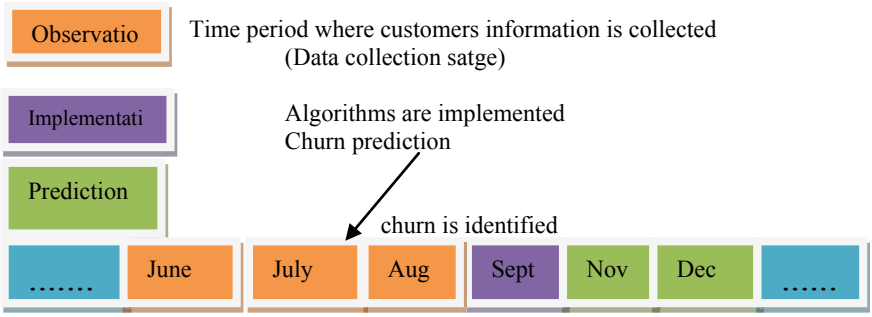


Fig. 2 Time line diagram for customers of bank

Step 2: Evaluate the XGBoost classifier accuracy using *K*-fold cross-validation by mentioning the number of folds. The ultimate accuracy of the XGBoost model is the mean of all the folds.

In training and testing, the training data is fit to the classifier. The data frame is constructed as matrix to pass it as an argument, and then, the maximum number of iterations is specified. In testing, the predict exit status of the customer is tested with the actual status to determine the accuracy of the classifier. Here, the confusion matrix shows an accuracy of 86.85%.

K-fold cross-validation method is applied by specifying number of folds (sample 10), and the average accuracy 88.07% is evaluated.

The performance is measured using the following parameters:

1. Accuracy 2. Specificity 3. Sensitivity 4. Error rate

Along with performance matrix, confusion matrix is also selected to prove the efficiency of model on the dataset for which the true values are familiar. The values for different classifiers are known using *R*Studio tool. Using confusion matrix, all the above-mentioned performance parameters are calculated (Fig. 2).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

$$\text{Error Rate} = 1 - \text{Accuracy}$$

TP—True Positive TN—True Negative FP—False Positive FN—False Negative

The flowchart depicts the systematic approach for classifying the customers of a bank by implementing predictive machine learning algorithm 'K-nearest neighbor' and a booster 'XGBoost' (Fig. 3).

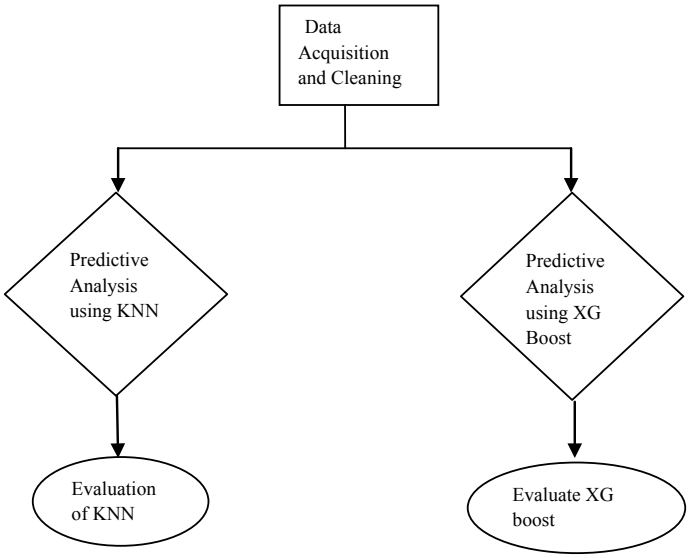
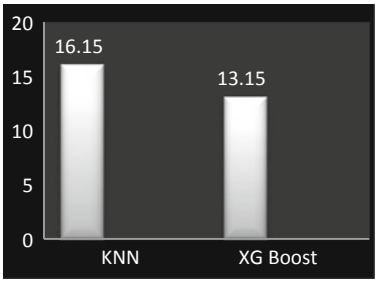


Fig. 3 Flowchart for customer churning system

4 Result

To assess the performance of the classifier model for churn prediction, bank data is trained for a specific period, and then, the customers of the bank are classified into loyal or churn based on their activities. This prediction gives useful insights to the bank officials regarding its customers and functioning of bank. The performance of the prediction model is the capability to identify customers exit status accurately. We use confusion matrix for evaluation (Figs. 4, 5, 6 and 7; Tables 2, 3 and 4).

Fig. 4 Error rate



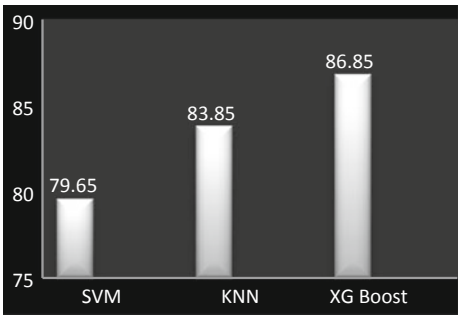


Fig. 5 Accuracy

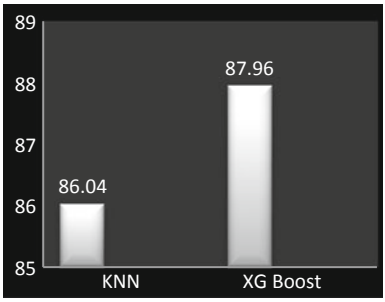


Fig. 6 Sensitivity

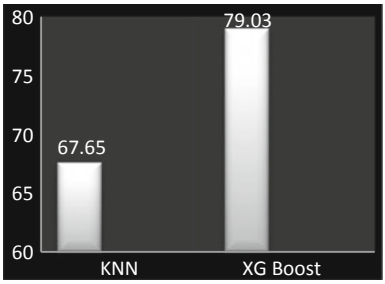


Fig. 7 Specificity

Table 2 Confusion matrix of KNN algorithm

Exit status	Predicted churn	Predicted retention
Actual churn	1516	77
Actual retention	246	161

Table 3 Confusion matrix of XGBoost algorithm

Exit status	Predicted churn	Predicted retention
Actual churn	1541	52
Actual retention	211	196

Table 4 Performance metrices

Classifier	% Accuracy	Error rate	Sensitivity	Specificity
KNN	83.85	16.15	86.04	67.65
XGBoost	86.85	13.15	87.96	79.03

5 Conclusion

In this paper, we propose an effective model of churn in bank industry. It combines the KNN with XGBoost algorithm to enhance the accuracy of the model; this proves the advantage of the technique used. XGBoost gives the best result in terms of accuracy, sensitivity and specificity. Boosting has given the increased accuracy of 86.85 with low error, high sensitivity and specificity.

Organizations periodically calculate customer churn in multiple aspects. Churning can be the number of customers lost, ratio or percentage of customers lost compared with total customers in bank. Churn can be calculated on quarter or annual basis. An accurate forecast can give insights on future using which a strategy can be formulated.

References

1. N. Hashmi, N.A. Butt, M. Iqbal, Customer churn prediction in telecommunication in a decade review and classification. Int. J. Comput. Sci. Issues (IJCSI) **10**(5), 271–281 (September 2013)
2. V. Mahajan, R. Mishra, R. Mahajan, Review of data mining techniques for churn prediction in telecom. JIOS **37**(2), 183–197 (2015)
3. L. Yan, R.H. Wolniewicz, R. Dodier, Predicting customer behavior in telecommunications. 1094-7167/04© 2004 IEEE Published by the IEEE Computer Society (2004)
4. B. Kaderabkora, P. Malecek, Churning and labour market flows in the new EU member states. Int. Inst. Soc. Econ. Sci. 372–378 (2015)
5. S.A. Qureshi, A.S. Rehman, A.M. Qamar, A. Kamal, *Telecommunication Subscribers' Churn Prediction Model Using Machine Learning* (IEEE, 2013), pp. 131–136
6. B. Mishachandar, K.A. Kumar, Predicting customer churn using targeted proactive retention. Int. J. Eng. Technol. **7**(2.27), 69–76 (2018)
7. K. Mishra, R. Rani, Churn prediction in telecommunication using machine learning, in *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)* (IEEE, 2017), pp. 2252–2257
8. E.M.L. Peters, G. Dedene, J. Poelmans, Understanding service quality and customer churn by process discovery for a multi-national banking contact center, in *Proceedings—IEEE 13th International Conference on Data Mining Workshops*, *ICDMW 2013* (2013), pp. 228–233. Art. no. 6753925
9. N. Wang, D.X. Niu, Credit card customer churn prediction based on the RST and LS-SVM, in *Proceedings of the 2009 6th International Conference on Service Systems and Service Management*, *ICSSSM '09* (2009), pp. 275–279. Art. no. 5174892
10. Y. Xie, X. Li, E.W.T. Ngai, W. Ying, *Customer Churn Prediction Using Improved Balanced Random Forests*. (Elsevier, Amsterdam, 2008), pp. 5445–5449
11. P. Spanoudes, T. Nguyen, Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors. arXiv:1703.03869v1 1–22 (2017)

12. W. Ying, X. Li, Y. Xie, E. Johnson, Preventing customer churn by using random forests modeling, in *IEEE International Conference on Information Reuse and Integration IEEE IRI-2008* (2008), pp. 429–434. Art. no. 4583069
13. Y. Chen, L. Zhang, Y. Shi, Post mining of multiple criteria linear programming classification model for actionable knowledge in credit card churning management, in *Proceedings—IEEE International Conference on Data Mining ICDM* (2011), pp. 204–211. Art. no. 6137381
14. A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, K. Huang, Customer churn prediction in telecommunication sector using rough set approach. Neurocomputing (2016). https://dx.doi.org/10.1016/j.neucom.2016.12.009
15. K. Sen, N.G. , in *Proceedings of 23rd Signal Processing and Communications Applications Conference, SIU* (2015), pp. 2384–2387. Art. no. 7130361
16. J. Zhao, X.H. Dang, Bank customer churn prediction based on support vector machine: taking a commercial bank's VIP customer churn as the example, in *2008 International Conference on Wireless Communications, Networking and Mobile Computing WiCOM 2008* (2008). Art. no. 4680698
17. G.Wang, L. Liu, Y. Peng, G. Nie, G. Kou, Y. Shi, Bayazit, Predicting credit card holder churn in banks of China using data mining and MCDM, in *Proceedings—2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology—Workshops, WI-IA 2010* (2010), pp. 215–218. Art. no. 5615798
18. S. Cui, N. Ding, Customer churn prediction using improved FCM algorithm, in *IEEE 3rd International Conference on Information Management (ICIM)*, (2017) Art. no. 16967234
19. B. Pradeep, S. Vishwanath Rao, & S. M. Puranik, Analysis of customer churn prediction in logistic industry using machine learning. Int. J. Sci. Res. Publ. **7**(11), 401–403 (2017)
20. A. Bilal Zorić, Predicting customer churn in banking industry using neural networks. 117–123 (2016)
21. A. Mishra, U.S. Reddy, A comparative study of customer churn prediction in telecom industry using ensemble based classifiers, in *Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)* (2017), pp. 721–725. Art. no. 17803488
22. N. Lu, H. Lin, J. Lu, G. Zhang, A customer churn prediction model in telecom industry using boosting. IEEE Trans. Ind. Inform. **10**(2), 1659–1665 (May 2014). https://doi.org/10.1109/TII.2012.2224355
23. H. Sayed, M.A. Abdel-Fattah, S. Kholief, Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **9**(11), 674–677 (2018)
24. A.R.K. Ahmad, A. Jafar, K. Aljoumaa, Customer churn prediction in telecom using machine learning in big data platform. J. Bigdata. (2019). https://doi.org/10.1186/s40537-019-0191-6
25. A. Keramati, H. Ghaneei, S.M. Mirmohammadi, Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innov. (2016). http://creativecommons.org/licenses/by/4.0/
26. J. Xiao, Y. Wang, S. Wang, A dynamic transfer ensemble model for customer churn prediction, in *2013 6th International Conference on Business Intelligence and Financial Engineering* (IEEE, 2014), pp. 115–119
27. S.F. Sabbeh, Machine-learning techniques for customer retention: a comparative study. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **9**(2), 273–281 (2018)
28. B. He, Y. Shi, Q. Wan, X. Zhao, Prediction of customer attrition of commercial banks based on SVM model. Procedia Comput. Sci. **31**, 423–430 (2014)