

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,000

Open access books available

125,000

International authors and editors

140M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.

For more information visit www.intechopen.com



Data Mining in Banking Sector Using Weighted Decision Jungle Method

Derya Birant

Abstract

Classification, as one of the most popular data mining techniques, has been used in the banking sector for different purposes, for example, for bank customer churn prediction, credit approval, fraud detection, bank failure estimation, and bank telemarketing prediction. However, traditional classification algorithms do not take into account the class distribution, which results into undesirable performance on imbalanced banking data. To solve this problem, this paper proposes an approach which improves the decision jungle (DJ) method with a class-based weighting mechanism. The experiments conducted on 17 real-world bank datasets show that the proposed approach outperforms the decision jungle method when handling imbalanced banking data.

Keywords: data mining, classification, banking sector, decision jungle, imbalanced data

1. Introduction

Data mining is the process of analyzing large data stored in data warehouses in order to automatically extract hidden, previously unknown, valid, interesting, and actionable knowledge such as patterns, anomalies, associations, and changes. It has been commonly used in a wide range of different areas that include marketing, health care, military, environment, and education. Data mining is becoming increasingly important and essential for banking sector as well, since the amount of data collected by banks has grown remarkably and the need to discover hidden and useful patterns from banking data becomes widely recognized.

Banking systems collect huge amounts of data more rapidly as the number of channels (i.e., Internet banking, telebanking, retail banking, mobile banking, ATM) has increased. Banking data has been currently generated from various sources, including but not limited to bank account transactions, credit card details, loan applications, and telex messages. Hence, data mining can be used to extract meaningful information from these collected banking data, to enable banking institutions to make better decision-making process. For example, *classification*, which is one of the most popular data mining techniques, can be used to predict bank failures [1–3], to estimate bank customer churns [4], to detect frauds [5], and to evaluate loan approvals [6].

In many real-world banking applications, the distribution of the classes in the dataset is highly skewed. A bank data is *imbalanced*, when its target variable is categorical and if the number of samples in one class is significantly different from those of the other class(es). For example, in credit card fraud detection, most of the instances in the dataset are labeled as “non-fraud” (majority class), while very few are labeled as “fraud” (minority class). Similarly, in bank customer churn prediction, many instances are represented as negative class, whereas the minorities are marked as positive class. However, the performance of classification models is significantly affected by a skewed distribution of the classes; hence, this imbalance problem in the dataset may lead to bad estimates and misclassifications. Dealing with imbalanced data has been considered as one of the 10 most difficult problems in the field of data mining [7]. With this motivation, this paper proposes a class-based weighting strategy.

The main contribution of this paper is that it improves the decision jungle (DJ) method by a class-based weighting mechanism to make it effective in handling imbalanced data. In the proposed approach, a weight is assigned to each class based on its distribution, and this weight value is combined with class probabilities. The experimental studies conducted on 17 real-world banking datasets confirm that our approach generally performs better than the traditional decision jungle algorithm when the data is imbalanced.

The rest of this paper is organized as follows. Section 2 briefly presents the recent and related research in the literature. Section 3 describes the proposed approach, class-based weighted decision jungle method, in detail. Section 4 is devoted to the presentation and discussion of the experimental results, including the dataset descriptions. Finally, Section 5 gives the concluding remarks and provides some future research directions.

2. Related work

As a data-intensive sector, banking has been a popular application area for data mining researchers since the information technology revolution. The continuous developments in banking systems and the rapidly increasing availability of big banking data make data mining one of the most essential tasks for the banking industry.

Banking industries have used data mining techniques in various applications, especially on bank failure prediction [1–3], possible bank customer churns identification [4], fraudulent transaction detection [5], customer segmentation [8–10], predictions on bank telemarketing [11–14], and sentiment analysis for bank customers [15]. Some of the classification studies in the banking sector have been compared in **Table 1**. The objectives of the studies, years they were conducted, algorithms and ensemble learning techniques they used, the country of the bank, and obtained results are shown in this table.

The main data mining tasks are classification (or categorical prediction), regression (or numeric prediction), clustering, association rule mining, and anomaly detection. Among these data mining tasks, classification is the most frequently used one in the banking sector [16], which is followed by clustering. Some banking applications [8, 10] have used more than one data mining techniques, among which clustering before classification has shown sufficient evidence of both popularity and applicability.

Apart from novel task-specific algorithms proposed by the authors, the most commonly used classification algorithms in the banking sector are decision tree (DT), neural network (NN), support vector machine (SVM), k-nearest neighbor

Ref	Year	Algorithms						Ensemble learning		Description	Country of the bank	Result
		DT	NN	SVM	KNN	NB	LR	Bagging (i.e., RF)	Boosting (AB, XGB)			
Manthoulis et al. [1]	2020			√			√		√	Bank failure prediction	USA	AUC >0.97
Ilham et al. [11]	2019	√	√	√	√	√	√	√		Long-term deposit prediction	Portugal	ACC 97.07%
Lv et al. [5]	2019		√							Fraud detection in bank accounts	—	ACC 97.39%
Krishna et al. [15]	2019	√	√	√	√	√	√	√	√	Sentiment analysis for bank customers	India	AUC 0.8268
Farooqi and Iqbal [12]	2019	√	√	√	√	√				Prediction of bank telemarketing outcomes	Portugal	ACC 91.2%
Carmona et al. [2]	2019						√	√	√	Bank failure prediction	USA	ACC 94.74%
Jing and Fang [3]	2018		√	√			√			Bank failure prediction	USA	AUC 0.916
Lahmiri [13]	2017		√							Prediction of bank telemarketing outcomes	Portugal	ACC 71%
Marinakos and Daskalaki [8]	2017	√	√	√	√		√			Customer classification for bank direct marketing	Portugal	AUC 0.9
Keramati et al. [4]	2016	√								Bank customer churn prediction	—	AUC 0.929
Wan et al. [6]	2016	√		√	√			√	√	Predicting nonperforming loans	China	AUC 0.965
Ogwueleka et al. [10]	2015		√						√	Identifying bank customer behavior	Intercontinental	AUC 0.94
Moro et al. [14]	2014	√	√	√			√			Prediction of bank telemarketing outcomes	Portugal	AUC 0.8
Smeureanu et al. [9]	2013		√	√						Customer segmentation in banking sector	Romania	ACC 97.127%

Table 1.
Classification applications in the banking sector.

(KNN), Naive Bayes (NB), and logistic regression (LR), as shown in **Table 1**. Some data mining studies in the banking sector [1, 2, 6, 11, 15] have used ensemble learning methods to increase the classification performance. Bagging and boosting are the most popular ensemble learning methods due to their theoretical performance advantages. Random forest (RF) [2, 6, 11, 15], AdaBoost (AB) [6], and extreme gradient boosting (XGB) [2, 15] have also been used in the banking sector as the most well-known bagging and boosting algorithms, respectively. As shown in **Table 1**, accuracy (ACC) and area under ROC curve (AUC) are the commonly used performance measures for classification.

Dealing with class imbalance problem, various solutions have been proposed in the literature. Such methods can be mainly grouped under two different approaches: (i) application of a data preprocessing step and (ii) modifying existing methods. The first approach focuses on balancing the dataset, which may be done either by increasing the number of minority class examples (over-sampling) or reducing the number of majority class examples (under-sampling). In the literature, synthetic minority over-sampling technique (SMOTE) [17] is commonly used as an over-sampling technique. As an alternative approach, some studies (i.e., [18]) focus on modifying the existing classification algorithms to make them more effective when dealing with imbalanced data. Unlike these studies, this paper proposes a novel approach (class-based weighting approach) to solve imbalanced data problem.

3. Methods

3.1 Decision jungle

A *decision jungle* is an ensemble of rooted decision *directed acyclic graphs* (DAGs), which are powerful and compact distinct models for classification. While a traditional decision tree only allows one path to every node, a DAG in a DJ allows multiple paths from the root to each leaf [19]. During the training phase, node splitting and merging operations are done by the minimization of an objective function (the weighted sum of entropies at the leaves).

Unlike a decision forest that consists of several evolutionary induced decision trees, decision jungle consists of an ensemble of decision directed acyclic graphs. Experiments presented in [19] show that decision jungles require significantly less memory while significantly improving generalization, compared to decision forests and their variants.

3.2 Class-based weighted decision jungle method

In this study, we improve the decision jungle method by a class-based weighting mechanism to make it effective in dealing with imbalanced data.

Given a training dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ that contains N instances, each instance is represented by a pair (x, y) , where x is a d -dimensional vector such that $x_i = [x_{i1}, x_{i2}, \dots, x_{id}]$ and y is its corresponding class label. While x is defined as input variable, y is referred as output variable in the categorical domain $Y = \{y_1, y_2, \dots, y_k\}$, where k is the number of class labels. The goal is to learn a classifier function $f: X \rightarrow Y$ that optimizes some specific evaluation metric(s) and can predict the class label for unseen instances.

Training dataset is usually considered as a set of samples from a probability distribution F on $X \times Y$. An instance component x is associated with a label class y_j of Y such that:

$$\frac{P(y_j|x)}{P(y_m|x)} > threshold, \forall m \neq j \quad (1)$$

where $P(y_j|x)$ is the predicted conditional probability of x belonging to y_j and threshold is typically set to 1.

In this paper, we focus on imbalanced data problem, where the number of instances in one class (y_i) is much larger or less than instances in the other class (y_j). Like many other classification algorithms, the decision jungle method is also affected by a skewed distribution of the classes, because the traditional classifiers tend to be overwhelmed by the majority class and ignore the rare samples in the minority class. In order to overcome this problem, we locally adapted a class-based weighted mechanism, where weights are determined depending on the distribution of the class labels in the dataset. The main idea is that the minority class receives a higher weight, while the majority class is assigned with a lower weight during the combination class probabilities. According to this approach, the weight over a class is calculated as follows:

$$W_c = \frac{1}{\sum_{i=1}^k \frac{1}{\text{Log}(N_i+1)}} \quad (2)$$

where W_c is the weight assigned to the class c , N is the total number of instances in the dataset, N_c is the number of instances present in the class c , and k is the number of class labels. In the proposed approach, Eq. (1) is updated as follows:

$$\frac{W_j * P(y_j|x)}{W_m * P(y_m|x)} > threshold, \forall m \neq j \quad (3)$$

Figure 1 shows the general structure of the proposed approach. In the first step, various types of raw banking data are obtained from different sources such as account transactions, credit card details, loan applications, and social media texts. Next, raw banking data is preprocessed by applying several different techniques to provide data integration, data selection, and data transformation. The prepared data is then passed to the training step, where weighted decision jungle algorithm is used to build an effective model which accurately maps inputs to desired outputs. The classification validation step provides feedback to the learning phase for adjustment

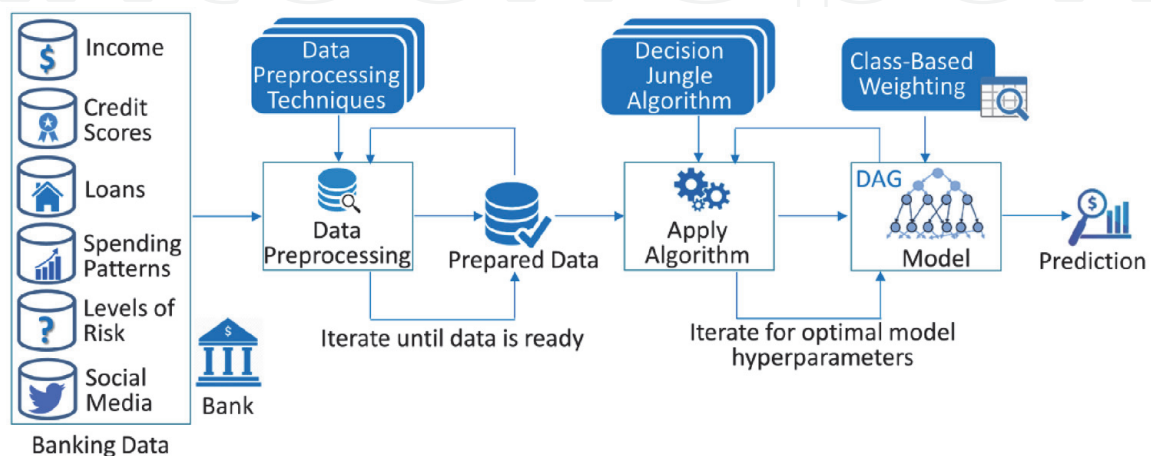


Figure 1.
 General structure of proposed approach.

to improve model performance. The training phase is repeated until a desired classification performance is achieved. Once a model is build, after that it can be used to predict unseen data.

4. Experimental studies

We implemented the proposed approach in Azure Machine Learning Studio framework on cloud platform. In all experiments, default input parameters of the decision forest algorithm were used as follows:

- Ensemble approach: Bagging
- Number of decision DAGs: 8
- Maximum width of the decision DAGs: 128
- Maximum depth of the decision DAGs: 32
- Number of optimization steps per decision DAG layer: 2048

Conventionally, *accuracy* is the most commonly used measure for evaluating a classifier performance. However, in the case of imbalanced data, accuracy is not sufficient alone since the minority class has very little impact on accuracy than the majority class. Using only accuracy measure is meaningless when the data is imbalanced and where the main learning target is the identification of the rare samples. In addition, accuracy does not distinguish between the numbers of correct class labels or misclassifications of different classes. Therefore, in this study, we also used several more metrics: *macro-averaged precision*, *recall*, and *F-measure*.

4.1 Dataset description

In this study, we conducted a series of experiments on 17 publically available real-world banking datasets which are described in **Table 2**. We obtained eight from the UCI Machine Learning Repository [20] and nine datasets from Kaggle data repository.

4.2 Experimental results

Table 3 shows the comparison of the classification performances of DJ and weighted DJ methods. According to the experimental results, on average, the weighted DJ method shows better classification outcome than its traditional version on the imbalanced banking datasets in terms of both accuracy and recall metrics. For example, the imbalanced dataset “bank additional” has an accuracy of 94.54% with the DJ method and 94.61% with the weighted DJ method. The accuracy is slightly higher with the weighted version because the classifier was able to classify the minority class samples better (0.8385, instead of 0.7914). The proposed method only disappointed in its accuracy and recall values for 4 of 17 datasets (with IDs 5, 9, 12, and 13).

It is observed from the experiments that the weighted DJ method failed in classifying only one dataset among 17 datasets in terms of macro-averaged recall values. This means that the proposed method generally can be able to build a good model to predict minority class samples.

No	Dataset	#Instances	#Features	#Class	Majority class (%)	Minority class (%)	Data source
1	Abstract dataset for credit card fraud detection	3075	12	2	85.4	14.6	Kaggle
2	Bank marketing [14]	4521	17	2	88.5	11.5	UCI
3	Bank full	45,211	17	2	88.3	11.7	UCI
4	Bank additional	4119	21	2	89.1	10.9	UCI
5	Bank additional full	41,188	21	2	88.7	11.3	UCI
6	Bank customer churn prediction	10,000	14	2	79.6	20.4	Kaggle
7	Bank loan status	100,000	19	2	77.4	22.6	Kaggle
8	Banknote authentication	1372	5	2	55.5	44.5	UCI
9	Credit approval	690	16	2	55.5	44.5	UCI
10	Credit card fraud detection [21]	284,807	31	2	99.8	0.2	Kaggle
11	Default of credit card clients [22]	30,000	25	2	77.9	22.1	UCI
12	German credit	1000	21	2	70.0	30.0	UCI
13	Give me some credit	150,000	12	2	93.3	6.7	Kaggle
14	Loan campaign response	20,000	40	2	87.4	12.6	Kaggle
15	Loan data for dummy bank	887,379	30	2	92.4	7.6	Kaggle
16	Loan prediction	614	13	2	68.7	31.3	Kaggle
17	Loan repayment prediction	9578	14	2	84.0	16.0	Kaggle

Table 2.
 The main characteristics of the banking datasets.

It can be deduced from the average precision and recall values that higher classification rates can be achieved with the weighted DJ method for minority classes, while more misclassified points in majority classes may also be detectable in the case of imbalanced data.

Figure 2 shows the comparison of the classification performances of two methods in terms of F-measure: decision jungle and class-based weighted decision jungle (weighted DJ). In principle, F-measure is defined as $F = (2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision})$, which is a harmonic mean between recall and precision. According to the results, for all banking datasets, the proposed method showed some increase or the same performance in the F-measure value.

It can be possible to conclude from the experiments that the minority and majority ratios are not the only issues in constructing a good prediction model. For example, the minority and majority ratios of the first and last datasets are very close, but the classification outcomes related to these datasets are not similar. Although the minority and majority class ratios are almost the same for these two datasets, there is a significant difference between the classification accuracy, precision, and recall values of the datasets, as can be seen in **Table 3**. There is also a need

ID	Dataset	Decision jungle			Class-based weighted decision jungle		
		Acc (%)	Precision	Recall	Acc (%)	Precision	Recall
1	Abstract dataset for credit card fraud detection	99.09	0.9918	0.9715	99.19	0.9923	0.9749
2	Bank	92.70	0.8909	0.7175	92.70	0.8492	0.7593
3	Bank full	91.06	0.8181	0.6874	91.17	0.8039	0.7217
4	Bank additional	94.54	0.9082	0.7914	94.61	0.8739	0.8385
5	Bank additional full	92.21	0.8332	0.7347	92.19	0.8126	0.7762
6	Bank customer churn prediction	87.37	0.8514	0.7291	87.40	0.8394	0.7411
7	Bank loan status	84.37	0.9170	0.6328	84.38	0.9169	0.6332
8	Banknote authentication	99.85	0.9987	0.9984	100.00	1.0000	1.0000
9	Credit approval	92.80	0.9273	0.9275	92.65	0.9257	0.9261
10	Credit card fraud detection	99.97	0.9915	0.9167	99.97	0.9861	0.9309
11	Default of credit card clients	83.05	0.7833	0.6695	83.16	0.7793	0.6785
12	German credit	86.30	0.8545	0.8088	85.70	0.8338	0.8198
13	Give me some credit	93.88	0.8245	0.5986	93.77	0.7861	0.6240
14	Loan campaign response	89.34	0.9393	0.5763	90.34	0.9390	0.6178
15	Loan data for dummy bank	95.19	0.9753	0.6837	95.20	0.9753	0.6844
16	Loan prediction	83.54	0.8715	0.7443	83.54	0.8631	0.7481
17	Loan repayment prediction	84.82	0.9059	0.5266	85.35	0.8900	0.5453
	Average	91.18	0.8990	0.7479	91.25	0.8863	0.7659

Table 3. Comparison of unweighted and class-based weighted decision jungle methods in terms of accuracy, macro-averaged precision, and macro-averaged recall.

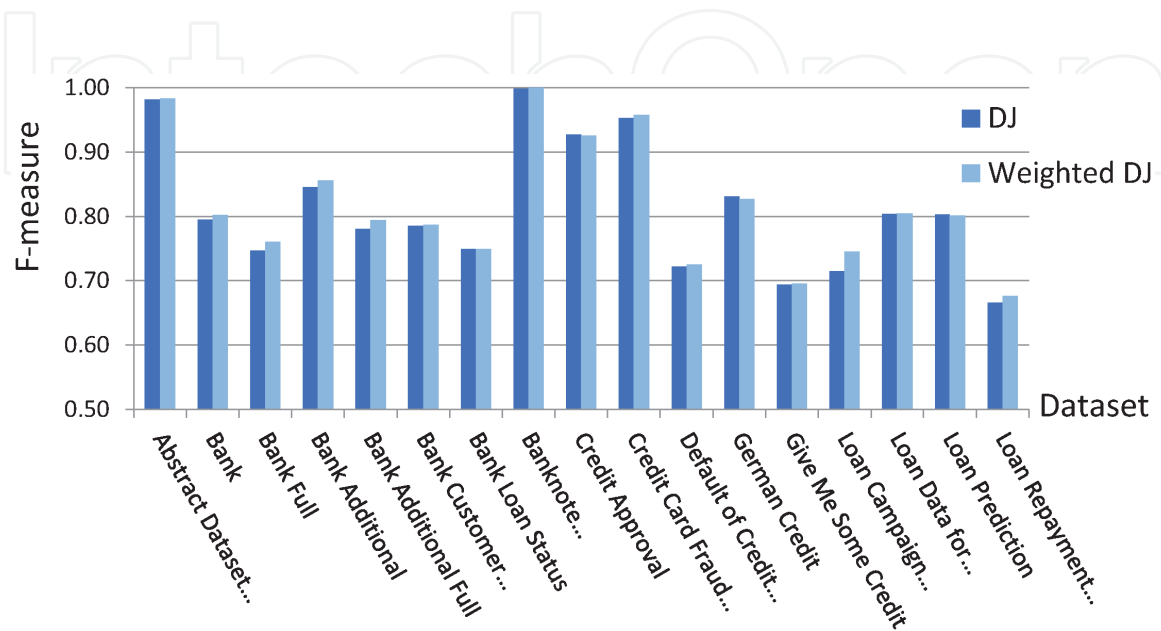


Figure 2. Comparison of unweighted and class-based weighted decision jungle methods in terms of F-measure.

for appropriate training examples that have data characteristics consistent with the class label assigned to them.

5. Conclusion and future work

As a well-known data mining task, classification in real-world banking applications usually involves imbalanced datasets. In such cases, the performance of classification models is significantly affected by a skewed distribution of the classes. The data imbalance problem in the banking dataset may lead to bad estimates and misclassifications. To solve this problem, this paper proposes an approach which improves the decision jungle method with a class-based weighting mechanism. In the proposed approach, a weight is assigned to each class based on its distribution, and this weight value is combined with class probabilities. The empirical experiments conducted on 17 real-world bank datasets demonstrated that it is possible to improve the overall accuracy and recall values with the proposed approach.

As a future study, the proposed approach can be adapted for multi-label classification task. In addition, it can be enhanced for the ordinal classification problem.

IntechOpen


Author details

Derya Birant

Department of Computer Engineering, Dokuz Eylul University, Izmir, Turkey

*Address all correspondence to: derya@cs.deu.edu.tr

IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Manthoulis G, Doumpos M, Zopounidis C, Galariotis E. An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for US banks. *European Journal of Operational Research*. 2020;**282**(2):786-801
- [2] Carmona P, Climent F, Momparler A. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics and Finance*. 2019;**61**: 304-323
- [3] Jing Z, Fang Y. Predicting US bank failures: A comparison of logit and data mining models. *Journal of Forecasting*. 2018;**37**:235-256
- [4] Keramati A, Ghaneei H, Mirmohammadi SM. Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*. 2016; **2**(1):1-13
- [5] Lv F, Huang J, Wang W, Wei Y, Sun Y, Wang B. A two-route CNN model for bank account classification with heterogeneous data. *PLoS One*. 2019;**14**(8):1-22
- [6] Wan J, Yue Z-L, Yang D-H, Zhang Y, Jiao L, Zhi L, et al. Predicting non performing loan of business Bank with data mining techniques. *International Journal of Database Theory and Application*. 2016;**9**(12):23-34
- [7] Yang Q, Wu X. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*. 2006; **5**(4):597-604
- [8] Marinakos G, Daskalaki S. Imbalanced customer classification for bank direct marketing. *Journal of Marketing Analytics*. 2017;**5**(1):14-30
- [9] Smeureanu I, Ruxanda G, Badea LM. Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*. 2013; **14**(5):923-939
- [10] Ogwueleka FN, Misra S, Colomo-Palacios R, Fernandez L. Neural network and classification approach in identifying customer behavior in the banking sector: A case study of an international bank. *Human Factors and Ergonomics in Manufacturing*. 2015; **25**(1):28-42
- [11] Ilham A, Khikmah L, Indra A, Ulumuddin A, Iswara I. Long-term deposits prediction: A comparative framework of classification model for predict the success of bank telemarketing. *Journal of Physics Conference Series*. 2019; **1175**(1):1-6
- [12] Farooqi R, Iqbal N. Performance evaluation for competency of bank telemarketing prediction using data mining techniques. *International Journal of Recent Technology and Engineering*. 2019;**8**(2):5666-5674
- [13] Lahmiri S. A two-step system for direct bank telemarketing outcome classification. *Intelligent Systems in Accounting, Finance and Management*. 2017;**24**(1):49-55
- [14] Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*. 2014;**62**:22-31
- [15] Krishna GJ, Ravi V, Reddy BV, Zaheeruddin M, Jaiswal H, Sai Ravi Teja P, et al. Sentiment classification of Indian Banks' Customer Complaints. In: *Proceedings of IEEE Region 10 Annual International Conference*. India; 17–20 October 2019. pp. 429-434

[16] Hassani H, Huang X, Silva E. Digitalisation and Big Data Mining in Banking. *Big Data and Cognitive Computing*. 2018;2(3):1-13

[17] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357

[18] Cieslak D, Liu W, Chawla S, Chawla N. A robust decision tree algorithms for imbalanced data sets. In: *Proceedings of the Tenth SIAM International Conference on Data Mining (SDM 2010)*. Columbus, Ohio, USA; 29 Apr-1 May 2010. pp. 766-777

[19] Shotton J, Nowozin S, Sharp T, Winn J, Kohli P, Criminisi A. Decision jungles: Compact and rich models for classification. *Advances in Neural Information Processing Systems*. 2013; 26:234-242

[20] Dua D, Graff C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. 2019. Available from: <http://archive.ics.uci.edu/ml>

[21] Carcillo F, Borgne Y-A, Caelen O, Oble F, Bontempi G. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*. 2020 in press. DOI: 10.1016/j.ins.2019.05.042

[22] Yeh IC, Lien CH. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*. 2009;36(2): 2473-2480