TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 4052 - 4057

Bank Marketing Data Mining

Amjaad Aljadani
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah, 22332, Saudi Arabia
amaljedani@effatuniversity.edu.sa

Hatoun Mukhtar
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah, 22332, Saudi Arabia
hamukhtar@effatuniversity.edu.sa

Bayan Alzanbaqi
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah, 22332, Saudi Arabia
baalzanbaqi@effatuniversity.edu.sa

Abdulhamit Subasi
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah, 22332, Saudi Arabia
absubasi@effatuniversity.edu.sa

Nada Aljehani
College of Engineering,
EFFAT University
AnNazlah Al Yamaniyyah,
Jeddah, 22332, Saudi Arabia
naaljehani@effatuniversity. edu.sa

Article Info
Volume 81
Page Number: 4052 - 4057
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *19 December 2019*

Abstract:

A lot of businesses include banks has use direct marketing strategies to reach customers in order to maximize the return rate and minimize the campaigning cost. In the bank marketing, the application of data mining classification is essential to determine the valuable customer and support the effective usage of Customer Relationship Management (CRM) system. And the unbalanced data mostly affect the accuracy rate of the results. Therefore, this study determine the preferred classification model based on the accuracy ration and other classification matrices after convert it to SMOTE "balanced", for enhancing the efficiency of bank marketing. The obtained results demonstrated RANDOM FOREST (RF) with MultiBoostAB recorded highest accuracy rate of 95.2996 from, which can be applied for direct marketing application.

Keywords:-*CRM; Data mining; RANDOM FOREST (RF); SMOTE*

1. INTRODUCTION

Companies mostly rely on direct marketing for target segments of customers by contacting them to achieve a specific goal. Customer centralization has remote interaction to the contact center to facilitate operational management of marketing campaigns. And in term of the bank sector they are like any type of business differentiates themselves through different direct marketing strategies as well as; using phone call to promote for new bank services or products.

However, a lot marketing campaigns have been failed because of the random classification strategies, which lead the banks to be under competition and economic pressures. and Technology such as; data mining techniques consider as excellent strategies for improving the efficiency of the business campaigns and helping the marketer in rethinking about the marketing strategies by focusing on maximizing valuable customer lifetime through the evaluation of obtainable information and customer metrics.

According to (Raorane& Kulkarni, 2011), studying consumer's behaviour, mind-set, psychology and motivation help organizations to improve their marketing strategy [1]. In other word it is allowing us to build longer and tighter relations in bias with in CRM-based bank marketing applications.

It is obvious from the statement before those organizations have to hold data on the customer, and promote capacity in order to analyse and use the transactional data in the perfect way. One of the transactions that have been automated is customer relationship management (CRM) [2].Data classification methods are essential in selecting the target customers for the direct marketing for specific bank's services and products.

Ensemble techniques are used in approximately all the classification problems [3]. Choosing the right ensembles determines the level of accuracy in classifiers. The methods are effective in the classification of the data, and they are very beneficial in the implementation of complex and sensitive applications. They are applied in healthcare where any slight machine learning improvement accuracy improvement can save a life [4].

Therefore in this study we aiming to implement and test different data mining methods including ANN, SVM, KNN, Rotation Forest, CART, C4.5 and RF Decision Trees classifiers and applying them to the data set of "bank marketing dataset" that taken from UCI as it is real data collected from a Portuguese retail bank, from May 2008 to June 2013, contains a total of 52,944 phone contact in order to determine the efficient classification as data mining model that can bring effective results in CRM-based banking marketing applications.

2. Data Mining Methods

2.1.1 Support Vector Machines (SVM)

It is a type of data mining classifier which applies an algorithm of supervised machine languages in regression and data classification challenges. SVM algorithm in the case of banking marketing data is such that the classifier plots the data as a point in a dimensional space whose size depends on the scale of the data. SVM as a classifier transforms series of input data onto multidimensional feature space through the process of non-linear mapping. The segregation of the classes of data depends on the identification of the appropriate hyperplane and kernel alignment. The support vectors show the data coordinates which rely on the observation of a researcher hence the definition of the concept leads to the description of the Support Vector Machine to be the critical data classification technique that is useful in the segregation of data into different classes on the two sides of the hyperplane [5]

2.1.2 Decision Trees

Classifiers that fall under the Decision Tree categories include CART, C4.5, ID3, CHAID and Conditional Inference Trees. They are the forms of supervised learning algorithms which have pre-defined target variables for solving issues in data classification. The principle of DTs is such that they use splitting, root, decision and terminal nodes to work with continuous and categorical output and input variables. The classifier split the data sample into different sets of classes which can be homogenous and the classification procedure is dependent on the type of differentiator or splitter of the input variables. After the classification of data using the decision trees, the other appropriate technique is the application of Random Forests which average the output of the decision trees across the sets of data with the aim of decreasing variance of the data [6].

2.1.3 K-Nearest Neighbour (k-NN)

The k-nearest-neighbor method is when given large training groups. It has been quite used to know the patterns. Nearest-neighbor classifiers are depend on learning by identification, that is, by matching a given trial tuple with training tuples that are analogous, to it. The trial tuple is defined by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n-dimensional pattern space. When given an undefined tuple, a k-nearest-neighbor classifier investigates the pattern area for the k trial tuples that are closest to the undefined tuple. These k training tuples are the k “nearest neighbors” of the unknown tuple. “Closeness” is defined based on distance metric, like Euclidean distance [7].

2.1.4 REPTree

The REPTree objective is to build regression tree or decisions using information gain/variance reduction and prunes it using reduced-error pruning. Vermin choice for speed, it sorts only values for numeric attributes once. Also, deals with missing values through splitting examples into pieces, like what C4.5 does, so you can set the minimum and the maximum

numbers of instances per leaf. The minimum proportion of training set variance for a split the numeric classes only is the number of folds for pruning [7].

2.1.5 NBTree

The NBTreeobjectives is to create trees with leaves which are Naïve Bayes classifiers is a hybrid between decision trees and Naïve Bayes. It creates trees with leaves that are Naïve Bayes classifiers for the instances that reach the leaf. When building the tree, cross-validation is used to decide whether a node should be split further or a Naïve Bayes model used instead [7].

2.1.6 ADTree

ADTree use boosting to build an alternating decision tree for two-class problems. The number of boosting iterations is a parameter that can be tuned to suit the dataset and the desired complexity–accuracy trade off. Each of the iteration adds three nodes to the tree (one split node and two prediction nodes) unless nodes can be merged. The default search method is the exhaustive search; the others are heuristics and are much faster. You can determine whether to save instance data for visualization. LADTree is an alternating decision tree algorithm that can handle multiclass problems based on the LogitBoost algorithm [8].

2.1.7 Random Forests (RF)

Imagine that each of the classifiers in the ensemble is a decision tree classifier so that the collection of classifiers is a "forest." To determine the spilt the individual decision trees are generated using a random selection of attributes. During classification, each tree votes and the most popular class is returned. Random forests can be built using bagging in tandem with random attribute selection. The CART methodology is used to grow the trees; to maximum size and are not pruned. Random forests formed this way, with random input selection, are called Forest-RI. Another form of random forest, called Forest-RC, uses random linear combinations of the input attributes. Instead of randomly selecting a subset of the attributes, it creates new attributes/features that are a linear combination of the existing attributes [7].

2.1.8 Bagging

Bagging is a learning technique in the basic ensemble methods. It improves classifiers' accuracy by combining individual models which are better compared to random guess model. The name is derived from Bootstrap aggregating constructed from a different dataset. This method enjoins both aggregating and bootstrapping. A classifier with better properties is achieved when the estimate of the bootstrap data distribution parameters is robust and accurate than the traditional one. Bagging is vital especially in building a better classifier in the case of noisy observations in the training set [9]. Better results are achieved in the ensemble as compared to single classifiers to build the final classifier [6].

2.1.9 MultiBoostAB

MultiBoostAB utilizes the use of classifiers by improving its accuracy. The classifier is used as a subroutine to build accurate classifier in the training set. In this technique, the classification system is repeated on the training data, with each step, the attention is given to the various examples of the set. On completion of the process, the individual model obtained are integrated into an accurate classifier in the training set. Hence, the final classifier obtained has a better accuracy. There are different versions of boosting algorithms, but AdaBoost is the best preferred [10].

2.1.10 SMOTE

In high dimensional data, the number of variables supersedes the number of samples. This type of predicament can be solved by oversampling or undersampling, as a result, we will have balanced data. Undersampling is helpful compared to oversampling. SMOTE, Synthetic Minority Oversampling Technique is the most popular oversampling technique whose main function was to improve random oversampling. However, this method's high dimensional data has not yet been thoroughly investigated [11].

2.1.11 NO SMOTE

NO SMOTE classification technique learns the classification category features. Data fields in the dataset are correctly labeled, and hence there are no options to choose the fields as is the case with the SMOTE classification technique. NO SMOTE is the exact opposite of SMOTE method [11].

2.1.12 F MEASURE

Record linkage involves linking and identifying records with same entities from different databases. This is considered as a classification since the purpose is to decide whether records are a match or not. This

method is not efficient because of imbalance in record linkage problem. Rather precision and recall are used and combined into F-measure. Although the weights depend on linkage method used, it can still be presented regarding recalls and precisions [9].

2.1.13 ROC

Receiver Operating Characteristics (ROC) is usually used for visualizing and organizing classifiers performance [5]. They are commonly applied to medical decision making. The area under the ROC curve is mainly used to measure the performance of supervised classification rules. It is however applicable to the case of two classes, which is extended by averaging pairwise comparison [5].

2.1.14 Kappa

Kappa coefficient of the agreement is used for evaluating concordance agreement for tagging tasks. Cohen's kappa and weighted kappa are used to measure accuracy in classification in data mining [6]. This technique is used for data with ordinal or nominal scales. It measures weighted kappa by quantifying the classification errors as weights [6].

3. Methodology

3.1 Sources of Data

One data set for bank marketing is used to test the performances of models. It is real data collected from a Portuguese retail data mining bank, from May 2008 to June 2013, in a total of 52,944 phone contacts and the dataset is unbalanced. The sources of this data set are taken from UCI [12].

3.2 Data Management

Data management concerns the safe storage data before use. The data storage will be electronically in hard copy format stored for security purposes. Every collection method will have its data stored in respective columns in the computer for easy retrieval.

3.3 Data Analysis Strategies

The analysis will entail the calculation of the frequency of data occurrence basing in the variables. For example, the analysis of some banks will be using Support Vector Machines and Decision Trees in regarding the percentage of banks that prefer each method. For the quantitative analysis of data mining classifiers, the useful measures will include classification model's specificity, sensitivity, and accuracy of the data classification [13]. The classification instances will be as false negative (FN), false positive (FP), true negative (TN) and false negative (PN) which will aid in the computation of specificity, accuracy, and sensitivity of the classifier chosen. N indicates the frequency of cases. The formulas are:

Specificity = (NT)/ (NT+PF)

Accuracy = (NT+PT)/N

Sensitivity = (PT)/ (PT+NF)

3.4 Data Classification software

WEKA software is used to apply the data mining and process the data set to come up with needed results to conduct the study. It is a machine learning toolkit that supports the decision making. WEKA software equipped with set of algorithms tools for data mining purpose.

Versions used in this study are Weka 3.7.4 and Weka 3.6, and Weka 3.8.

4. Result and discussion

In this study, only one data set "bank marketing" to apply different data mining methods such as ANN, KNN, SVM, RF, C4.5,Random Tree, REPIREE, LAB Tree and PART.

There are three different methods are used for testing and implementation which are single, bagging and multiboostAB.

Table 2 and 3 tabulate the result of the data set with SMOTE and NO SMOTE. From the tables, the final result of the balanced data (SMATE) shows higher accuracy than the imbalance data (NO SMATE) in all aspects.

November-December 2019
ISSN: 0193-4120 Page No. 4052 - 4057

Table 1. Final results of the imbalanced data (NO SMOTE)

	Single	Bagging	MultiBoostAB	Random Subspace	Rotation Forest	AdaBoost
Classifiers	Accuracy (%)					
ANN (MLP)	89.1236	90.3132	89.1236	90.5317	90.2403	89.1236
k-NN	89.2207	89.3178	87.5698	89.9247	89.682	89.2207
SVM	91.1872	91.2357	91.3814	90.2403	91.1629	91.0658
RF	91.0415	90.8716	91.0415	90.7016	90.7259	91.0415
C4.5 J48graft	90.8958	91.1872	90.8958	90.4103	90.993	90.0218
Random Tree	88.1039	90.556	88.1525	90.4831	90.4831	88.1525
REPTREE	90.7259	90.7987	90.386	90.4346	91.3328	89.9005
LAD Tree	91.6242	91.43	91.1629	90.6773	91.5999	90.8716
PART	89.512	89.682	90.2646	89.8519	90.993	89.8276

Table 2. Final results of the balanced data (SMOTE)

	Single	Bagging	MultiBoostAB	Random Subspace	Rotation Forest	AdaBoost
Classifiers	Accuracy (%)					
ANN (MLP)	93.6916	94.4887	93.6916	93.829	94.7774	93.6916
K-NN	92.3447	92.0148	91.2177	94.0902	92.5783	91.685
SVM	93.5267	93.7878	93.5404	93.6229	94.1176	93.4442
RF	95.3683	95.2309	95.2996	95.2446	95.1209	95.2996
C4.5 J48graft	93.4442	93.7741	94.0214	94.2826	93.8015	94.6811
Random Tree	91.74	95.0385	91.8087	94.8186	93.8015	91.8087
REPTREE	92.6058	93.4442	94.0352	94.2413	94.4475	94.3788
LAD Tree	89.3485	89.4173	87.7817	79.8378	91.6025	86.4623
PART	93.0731	91.3086	94.5162	94.7499	94.8186	94.3513

Based ontable 1 and table 2, the single classifiers and ensemble classifiers are tested at the same time for both balanced and imbalanced data. For ANN (MLP) classifier, the accuracy rate with imbalanced data is 89.1236 the accuracy positively increased about 5% after balancing the data and become 93.6916. Moreover, not only using balancing data can increase the accuracy rate, ensemble classifiers can increase it as well. In table 2, single-ANN (MLP) classifier gave 93.6916 even with balanced data. Additionally, using ensemble classifiers such as rotation forest increased the accuracy rate to become 94.7774. Thus, the obtained results conclude that an appropriate data-mining model that can be applied for direct marketing application is RANDOM FOREST (RF) at accuracy rate of 95.2996 with MultiBoostAB.

5. Conclusion

This study has to pass through several process and steps to test and implement many different methods in order to achieve the purpose of it which providing the banks with the appropriate data mining methods that would help them to improve and enhance their direct marketing. The obtained results demonstrated highest accuracy rate of 95.2996 from RANDOM FOREST (RF) with MultiBoostAB, which can be applied for direct marketing application.

6. References

[1] Raorane, A., & Kulkarni, R. (2011). Data mining techniques: A source for consumer behavior analysis. ArXiv Preprint ArXiv:1109.1202.

[2] Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, 25(4), 6.

[3] Kawaguchi, K., Suzuki, E., Yamaguchi, A., Yamamoto, M., Morita, S., & Toi, M. (2017). Altered expression of major immune regulatory molecules in peripheral blood immune cells associated with breast cancer. *Breast Cancer*, 24(1), 111–120.

[4] Kuncheva, L. I. (2004). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

[5] Bessis, N., &Dobre, C. (2014). Big data and internet of things: a roadmap for smart environments (Vol. 546). Springer.

[6] Nayak, J., Naik, B., &Behera, H. S. (2015). Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In Computational Intelligence in Data Mining-Volume 2 (pp. 133–149). Springer.

[7] Kamber, M., Han, J., & Pei, J. (2012). Data mining: Concepts and techniques. Elsevier.

[8] Holmes, S., & Featherstone, W. (2002). SHORT NOTE: Extending simplified high-degree synthesis methods to second latitudinal derivatives of geopotential. Journal of Geodesy, 76(8), 447–450.

[9] James, G., Witten, D., Hastie, T., &Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.

[10] Whitaker, C., &Kuncheva, L. (2003). Examining the relationship between majority vote accuracy and diversity in bagging and boosting. School of Informatics, University of Wales, Bangor.

[11] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

[12] UC Irvine Machine Learning Repository, Center for Machine Learning and Intelligent Systems. Available at: http://archive.ics.uci.edu/ml/index.php [Accessed 23 June 2017].

[13] Rivera, W. A., &Xanthopoulos, P. (2016). A priori synthetic over-sampling methods for increasing classification sensitivity in imbalanced data sets. *Expert Systems with Applications*, 66, 124–135.