VOL 1 (2017) NO 4 - 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Customer Profiling using Classification Approach for Bank Telemarketing

Shamala Palaniappan[#], Aida Mustapha[#], Cik Feresa Mohd Foozy[#] Rodziah Atan[*]

[#] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia*
[*] *Software Engineering and Information Systems,Faculty of Computer Science & Information Technology, Universiti Putra Malaysia*

E-mail: shamala@uthm.edu.my, aidam@uthm.edu.my, feresa@uthm.edu.my, rodziah@upm.edu.my

Abstract— **Telemarketing is a type of direct marketing where a salesperson contacts the customers to sell products or services over the phone. The database of prospective customers comes from direct marketing database. It is important for the company to predict the set of customers with highest probability to accept the sales or offer based on their personal characteristics or behaviour during shopping. Recently, companies have started to resort to data mining approaches for customer profiling. This project focuses on helping banks to increase the accuracy of their customer profiling through classification as well as identifying a group of customers who have a high probability to subscribe to a long-term deposit. In the experiments, three classification algorithms are used, which are Naïve Bayes, Random Forest, and Decision Tree. The experiments measured accuracy percentage, precision and recall rates and showed that classification is useful for predicting customer profiles and increasing telemarketing sales.**

Keywords— **Decision tree, classification, data mining, customer profiling.**

I. INTRODUCTION

Telemarketing is a type of direct marketing where a salesperson contacts the customers to sell products or services over the phone. The database of prospective customers comes from direct marketing database and is used mostly for communication, advertisement and analysis [1]. To ensure the success of telemarketing, the company must focus on their potential customer database by predicting the list of customers with higher probability to use the product or service that the company is trying to sell. At present, many banks have adopted the predictive approach or business analytics that use data mining techniques to classify potential customers before they invest on making the calls. Many predictive models have been proposed that each model has its own advantages and disadvantages [2].

The customer database consists of customer information that will be used as input feature during the data mining task. One important factor that affects the performance of a prediction task is the number of input feature. Because customer database has many features, both relevant and irrelevant as prospective customers, many work in literature focused on feature selection, a method where only the relevant features will be selected, discarding the irrelevant or weak features in the dataset [3, 4]. The research attempted to find the minimum set of features that is close enough to represent the original dataset but gives a good result.

Other than feature selection, another data mining task used in telemarketing business is classification. Classification is one of the most popular data mining techniques that focus on building a classification model or function, called as a classifier, and predict the class of objects whose class label that is unknown. Examples of classification applications include pattern recognition, medical diagnosis, detecting faults in industry application, and classifying financial market trends.

[5] performed customer prediction using the bank telemarketing dataset based on a correlation-based feature subset selection algorithm and a dataset balancing technique. The dataset balancing technique is a method to to make the label of dataset equivalent before applying the correlation-based feature subset selection algorithm to select the robust feature. This is carried out by randomly selecting any data of each label out of a dataset equally. The correlation-based feature subset selection algorithm is a heuristic method worth of a subset of features [6].

This project focuses on helping banks to increase the accuracy of their customer profiling through classification as well as identifying a group of customers who have a high probability to subscribe to a long term deposit. The remainder of this paper proceeds as follows. Section II presents the materials and methods used to achieve the objective of this research, Section III presents the experimental results, and finally Section IV concludes with some indication for future work.

II. Introduction

This paper proposes a classification approach to customer profiling using a bank telemarketing dataset. The objective is to identify a group of potential customers who have a high probability to subscribe to a long term deposit. Customer profiling is also important for the banks to assess whether they can trust on the customers' profiles or whether they should offer any services to the customers.

Dataset

This study considers real data on bank telemarketing provided by the UCI Machine Learning Repository. The data was collected from a Portuguese retail bank within the duration of five years (2008 to 2013). The telemarketing instances consist of a total of 41,188 phone contacts used in various direct marketing campaigns in a Portuguese banking institution. Table 1 shows the dataset that is composed of 21 attributes including a class label.

TABLE 1
ATTRIBUTES IN BANK TELEMARKETING DATASET

Attributes	Type
Age	Numeric
Job	Categorical
Marital	Categorical
Education	Categorical
Default	Categorical
Housing	Categorical
Loan	Categorical
Contact	Categorical
Month	Categorical
day of week	Categorical
Duration	Numeric
Campaign	Numeric
Pdays	Numeric
Previous	Numeric
Pountcome	Categorical
emp.var.rate	Numeric
cos.price.idx	Numeric
cons.conf.idx	Numeric
euribor3m	Numeric
nr.employed	Numeric
Label	Categorical

Pre-Processing

Given the dataset, pre-processing focused on normalization to keep data consistent and to check that no loss of data [10]. Normalization is a technique used during the design of database tables in order to minimize data duplication and to ensure the database does not have any logical and structural anomalies. Table 2 shows the normalization parameters used during the pre-processing.

TABLE 2
NORMALIZATION PARAMETER

Parameter	Value
Create view	No
Attribute filter type	All
Invert selection	No
Include special attributes	No
Method	Z-transformation

Classification

For the classification experiments, three classification algorithms are chosen, which are Naïve Bayes, Decision Tree, and Random Forest. Naive Bayes is a simple probabilistic classifier that assumes that all nodes or features are independent from one another. The decision function measures the probability $P(c \mid D)$ as the output of a probabilistic classifier whether an instance D belongs to the class c [7]. Apart from being known as a statistical method for classification, Naives Bayes is also known as a type of supervised learning method. The Bayes Theorem is as follows.

$$\mathrm{P(h/D)\ =\ P(D/h)\,P(h)\ /\ P(D)}$$

where $P(h)$ is the prior probability of hypothesis h, $P(D)$ is the prior probability of training data D, $P(h/D)$ is the probability of h given D, and finally $P(D/h)$ is the probability of D given h.

A decision tree is a classifier a tree-structured classifier that uses decision rules from the large amount of initial data to extract knowledge. Classifying using a decision tree is straightforward as the algorithm concisely stored and efficiently classifies new data. The advantages of decision tree in data mining as compared to Random Forest and naïve Bayes includes its ability to handle different input data types such as, numerical, textual and nominal, it can even take care of datasets whose instances have missing values and errors, and it is also available in various packages of data mining and a number of platforms.

The third classification algorithm is the Random Forest, which is part of a collection of decision trees. It is presented independently with some controlled modification. Basically, the trees built in a Random Forest are based on majority voting, which is already a representation of an accurate output. In a Random forest, the instances or cases in the training dataset will be sampled randomly but with replacement from the original data. This sample will then act as a training set for growing the tree. In order to split the nodes, a constant value is chosen during the entire growth of the forest. Each tree is made to grow to the largest extent possible. Unlike the normal decision tree, pruning in the forest is restricted unless higher classification accuracy is needed at the expense of higher execution time.

Evaluation Metrics

In evaluating the performance of the classification algorithms in customer profiling, the classifiers are measured for their accuracy percentage, precision, and recall rates using RapidMiner tool. RapidMiner is a data mining platform enables concentrated effort on machine learning and data mining.

i. Accuracy: Accuracy of a classifier is measured by the percentage of the test set tuples that are correctly classified by the classifier. The formula for accuracy is as follows where TP is the true positives, FN is the false negatives, TN is the true negatives, and FP is the false positive rates.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

ii. Precision: Precision is the ratio of all true positives against the sum of all results; both negative and positive.

$$precision = \frac{TP}{TP + FP + TN + FN}$$

iii. Recall: Recall is the ratio of positive tuples that are correctly identified or true positives, against the sum of all true positives and false negatives.

$$recall = \frac{TP}{TP + FN}$$

III. Experiments and Results

In the classification experiment, one attribute was taken as label or class attribute, which is the subscription of the product (bank term deposit) with values of would be ('yes') or not ('no'). The RapidMiner has been used to classify the product using three classification algorithms, which are Naive Bayes, Random Forest and Decision Tree algorithms. In the case of Random Forest and Decision Tree, pruning and pre-pruning is applied for better results by compromising the execution time. Table 3 and Table 4 show the parameters used for Decision Tree and Random Forest in RapidMiner.

TABLE 3
DECISION TREE PARAMETER

Criterion	Gain Ratio
Maximal Depth	20
Confidence	0.25
Minimal gain	0.1
Minimal leaf size	2
Minimal size for split	4
Number of pre-pruning alternatives	3

TABLE 4
RANDOM FOREST PARAMETER

Parameter	Value
Number of Trees	10
Criterion	Gain_ratio
Maximal Depth	20
Confidence	0.25
Minimal gain	0.1
Minimal leaf size	2
Minimal size for split	4
Number of pre-pruning alternatives	3

Validation

In the experiments, cross-validation approach was used to obtain the classification accuracy, precision and recall rates of the three classification algorithms. From the literature, the most common validation approaches include the hold-out method as well as the cross-validation approach [8]. In the hold-out method, a portion of the dataset is held out for testing and the remaining parts are used for training the classifier. The cross-validation approach applies the same concept, however, it repeats the process by *x* number of times so that all instances in the dataset will eventually used for both training and testing.

For example, in a 10-fold cross validation method, the data is divided into 10 parts or division. From the 10 parts, 9 parts will be used for training the classifier in the first run, while the last part is used during testing the classifier. In following run, different part is used for training and testing but keeping the same 10 parts divided at the beginning of the experiment. The classification experiments will continue to run until all 10 parts are used as part of either training or testing dataset. Table 5 shows the cross-validation parameter as in RapidMiner.

TABLE 5
CROSS-VALIDATION PARAMETER.

Parameter	Value
Average performance	Yes
Leave one out	No
Number of validation	10
Sampling type	Automatic
Use local random seed	No

Results

Three algorithms were sourced from the RapidMiner data mining tool in order to perform the classification experiment, which are Naïve Bayes, Random Forest, and Decision Tree. The results from the classification experiments on the bank telemarketing dataset is given in Table 6, consisting the detailed percentage for accuracy, precision, and recall percentage of each classifier.

TABLE 6
ANALYSIS ON DATASET.

Algorithm	Accuracy	Precision	Recall
Naive Bayes	86.27%	68.65%	75.54%
Random Forest	88.81%	82.34%	50.45%
Decision Tree	90.68%	77.84%	69.71%

In total, three different classification algorithms have been used to classify the data datasets and the following results have been obtained. No special modification were made to each classifier operators in RapidMiner. Fig. 1 shows the comparison of performance across all three classifiers. Note there is a considerable difference between the percentage of Precision and especially of Recall in the dataset.

The results for classification accuracy showed that Naïve Bayes has the lowest accuracy of 86.27%. Random Forest has an average accuracy percentage of 88.81%, and Decision Tree has the highest accuracy with 90.68%. In terms of precision percentage for precision, the results revealed that the Naïve Bayes classifier produced the lowest precision percentage of 68.65%. Random forest has the highest precision percentage of 82.34%, and Decision Tree has an average precision value of 77.84%. In this application, Decision tree comes second with a high success percentage of 77.84%. As for the recall percentage Random Forest produced the lowest recall value with 50.45%, Naïve Bayes has the highest recall value of 75.54%, and Decision Tree has an average recall percentage of 69.71%. In this application, Decision tree is the second recall values with high success values of 69.71%.

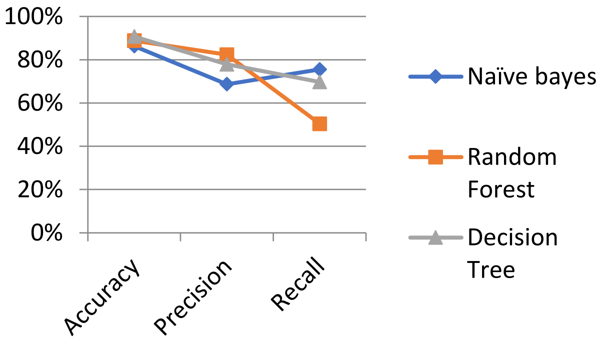

Fig. 1 Comparison of performance for each classifier.

In conclusion, the Random Forest algorithm produced the second highest accuracy percentage but with lower recall value and higher precision values. It is an average classification algorithm with unstable precision values and recall values as compared to other algorithms. Naïve Bayes has the lowest accuracy percentage with lower precision value but higher recall percentage. Overall, the Decision Tree is the best algorithm for classification of customer profiling as compared to other classification algorithm. In addition, in terms of precision and recall, Decision Tree is also the second highest algorithm with best precision and recall.

IV. Conclusions

This paper presented a classification approach for customer profiling in banking telemarketing. The classification experiments compared three classification algorithms, which are Naïve Bayes, Decision Tree, and Random Forest.

The experimental results showed that the selected algorithms in the proposed classification approach are capable to improve the prediction performance even when working with smaller number of features. The proposed method is able to enhance the performance of the classification model while employing smaller storage space. This approach also reduces the computation time and gains higher prediction.

Acknowledgment

This research is funded by Universiti Tun Hussein Onn Malaysia.

References

[1] Kadir, H. M. G. A. Garis Panduan Penasihatan Akademik Di Politeknik (KADIR, H.). Kuala Lumpur (2012).

[2] Asuncion, A., Newman, D. CA: University of California, School of Information and Computer Science, UCI Machine Learning Repository. Irvine, (2012).

[3] Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann (2016).

[4] John, G., Kohavi, R., and Pfleger. Irrelevant features and the subset selection problem. Int. Conf. on Machine Learning, Morgan Kaufman, San Francisco. (1994), 121-129.

[5] Duch, W., Winiarski, T., Biesiada, J., and Kachel, A. Feature Ranking Selection and Discretization. Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), Istanbul. (2003), 251-254.

[6] Xinguo, L. et al., A Novel Feature Selection Method Based on CFS in Cancer Recognition, IEEE 6th International Conference on System Biology IISB), (2012).

[7] Yogesan, K., Eikelboom, R.H., Barry, C.J. Centre for Ophthalmology and Visual Science. University of Western Australia, 2 Verdun Street, Netherlands, WA 6009, Australia.

[8] Zahlmann, G., Scherf', M., Wegner, A. Neurofuzzy and EUBAFES as tools for knowledge discovery in visual field data. National Research Centre for Environment and Health, Proceedings of the 20th Annual International Conference of the ZEEE Engineering in Medicine and Biology Society, Vol. 20, No 3 (1998).

[9] Teli, S., Kanikar, P. A Survey on Decision Tree Based Approaches in Data Mining. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, (2015).

[10] Ting, S.L., IP, W.H., Tsang, A.H.C. Is Naïve Bayes a Good Classifier for Document Classification? International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, (2011).

[11] Gupte, A., Joshi, S., Gadgul, P., Kadam, A. Comparative Study of Classification Algorithms used in Sentiment Analysis. International Journal of Computer Science and Information Technologies, Vol. 5 (5), (2014).

[12] Moro, S., Cortez, P. Rita. P. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, (2014), 62:22-31.