Mining a Marketing Campaigns Data of Bank

Yasemin Gultepe[1]; Wisam Gwad[2]; Yuosra Aljamel[3]; Yossf Ahmed[4]

[1,2,3,4] Computer Engineering & KASTAMONU University, Turkey

[1] yasemingultepe@ogr.kastamonu.edu.tr; [2] wgwad@ogr.kastamonu.edu.tr
[3] usraalhade@gmail.com, [4] yghyth2@gmail.com

Abstract— In this paper, we propose a data mining approach to predict the success of telemarketing. We are applying the algorithms for the first time on the dataset. The dataset obtained from UCI, which contain the most common machine learning datasets. The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The number of the instance is 45212 with 15 input variables and the output variable. Classification is a data mining techniques used to predict group membership for a data instance. we present the comparison of different classification techniques in open source data mining software which consists of a One-R algorithm methods and Naïve-Bayes algorithm The experiment results show are a bout classification sensitivity, specificity, accuracy. The results on bank marketing data discovered that the One-R algorithm is better in classifying the data comparing with the Naïve-Bayes algorithm; where the error rate is lower.
Keywords— Bank Marketing, Data mining, One-R Algorithm Naïve-Bayes algorithm.

I. INTRODUCTION

Data mining (Han, Kamber, & Pei, 2011b) is the process of extracting previously unknown information from a large dataset. Today, data mining is being used by several industries including finance and banking. The bank is marketing department can use data mining to analyse customer datasets and develop statistical profiles of individual customer preference for product and service.

In bank direct marketing domain, there are several data mining techniques can be used for classifying marketing service such *One-R Algorithm*, naive Bayes classifier, classification and association rule mining.

Exploratory data analysis on variables will be used, to discover the relation between the variables and the class variable, the relation between two variables according to the class variable, and data mining algorithms "classification" will be used to classify the bank client's data. The classification goal is to predict if the client will subscribe to a term deposit (variable outcome).

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (data objects whose class label is known).

II. LITERATURE SURVEY

Today, many applications in a wide and various ranges of the business founded and worked in this regulation
In fact, direct marketing is in the main a strategy of many of the banks and insurance companies for interacting with their customers.
The listed researchers are most famous researches in this field. The proposed method different from all techniques which used deep auto-encoders also presented more satisfactory results when compared with previous works.

In 1996, U. Fayyad, G. Shapiro defined the general knowledge discovery process as an interactive and iterative process involving more or less the following steps: understanding the application field, data selecting, preprocessing and cleaning data, integration of data, data reduction and transformation, selecting algorithms of data mining, interpretation and description of the results and using the discovered knowledge.(Elsalamony & Elsayad, 2013)

In fact, the data mining can be classified into two categories descriptive and predictive.
Actually, in the recent years, data mining is occupying great position of attention area in the society of business or banking because its elasticity in working with a large amount of data and turning such data into clear information and knowledge.(Han, Kamber, & Pei, 2011a).

Most of the people may be confused in understanding between the terms "knowledge discovery" and "data mining" in different areas. Knowledge discovery in databases is the process of identifying valid, novel, probably useful, and finally understandable patterns/models with data. On the other hand, data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that under some acceptable computational efficiency limitations, finds patterns or models in data.(Ian H. Witten, Eibe Frank, 2013)

Usually, the selected customers are contacted directly through personal contact, telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. In fact, direct marketing is in the main a strategy of many of the banks and insurance companies for interacting with their customers.(Ayetiran & Adeyemo, 2012).

Some of the banks and financial services companies may depend only on strategy of mass marketing for promoting a new service or product to their customers. In this strategy, a single communication message is broadcasted to all customers through media such as television, radio or advertising firm, etc.(Elsalamony & Elsayad, 2013).

From the literature, the direct marketing is becoming a very important application in data mining these days. The data mining has been used widely in direct marketing to identify prospective customers for new products, by using purchasing data, a predictive model to measure that a customer is going to respond to the promotion or an offer.(A.Elsalamony, 2014).

Direct marketing is very effective and widely used strategy of contacting customers or potential customers rather than having an indirect channel, especially for banking sector.(Nachev, 2014).

However, it is of vital importance to find the "right" customers to make sure the success of bank telemarketing because if the contacted customer did not want the product, the outbound calls would be considered intrusive and inbound calls loaded with too much campaign content would also be annoying. Thus, more focus should be put on the task of selecting the best set of clients or targeting the right segments of customers, i.e., those who are more likely to subscribe a Product.(Cortez, n.d.)

By using these data mining techniques, organizations and institutions can mine information that does not exist about their customers and products and by this can simply define the ideals of customers and forecast their future behaviours and requirements.(Moin & Ahmed, 2012).

Recently, technological progress deeply affects Marketing domain with specific demands for specific targets. Thus, many companies prefer target marketing campaign instead of mass marketing campaign, which become very less effective because of the intensive competition when facing challenges from a rapidly changing market situation. (Su, Chen, & Sha, 2006).

III. Data SET AND PERFORMANCE MEASUREMENT

Dataset used:

The dataset used for experiments in this paper was related with direct marketing campaigns of a Portuguese banking institution and is available at UCI Machine Learning Repository The marketing campaigns(Cortez, n.d.). Were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed results of direct bank marketing campaigns. It includes 17 campaigns of a Portuguese bank conducted between May 2008 and November 2010. The details of 17 attributes are shown in Table 1.

TABLE 1. DATASET SUMMARY

Variable	Description
age	numeric, age of client
job	categorical, type of job (admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services)
marital	Categorical, marital status (married, divorced, single. Here "divorced" states the both divorced or widowed)
education	categorical (unknown, secondary, primary and tertiary)
default	binary, customer credit is in default (yes, no)
balance	numeric, average yearly balance (in euros)
housing	binary, status of housing loan (yes, no)
loan	binary, clients personal loan (yes, no)
contact	categorical, contact communication type (unknown, telephone, cellular)
day	numeric, the last contact day of the month range (1-31)
month	categorical, last contact month of the year
duration	numeric, last contact duration (in seconds)
campaign	numeric, number of contacts performed during this campaign
pdays	numeric, number of days that passed by after the client was last contacted from a previous campaign
previous	numeric, number of contacts which are made before this campaign
poutcome	categorical, result or outcome of the previous marketing campaign (unknown, other, failure, success)
y	binary, (desired target) client subscribed a term deposit or not

PERFORMANCE MEASUREMENT:

a. The relation between the variables and the class variable:

• Duration:

When the contact time duration increased the client tends mostly to subscribe a term deposit; When the contact time duration between 3250 to 3499 the client tends mostly to not subscribe a term deposit; When the contact time duration between 4750 to 4999 the client tends at all to not subscribe a term deposit.

This attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. In addition, after the end of the call y is obviously known, as shown in Fig. 1.
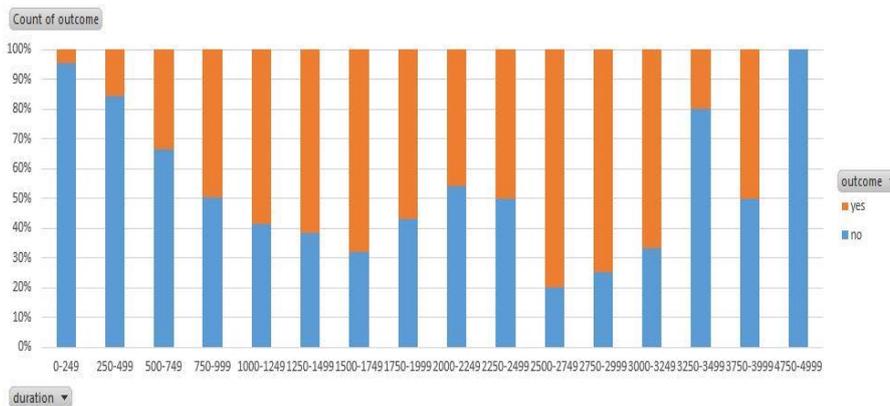

Fig. 1 The relation between Duration and output target

B. The relation between two variables according to the class variable:
- Duration and Campaign:

When the X-axis is duration and the Y-axis campaign, we discover the following:
The most of the clients didn't accept the offer when the duration under the 300 seconds in all the campaigns
Whereas the most of the clients accept the offer when the duration more than 300 seconds and the campaign between 1 and 10 as shown in Fig. 2.
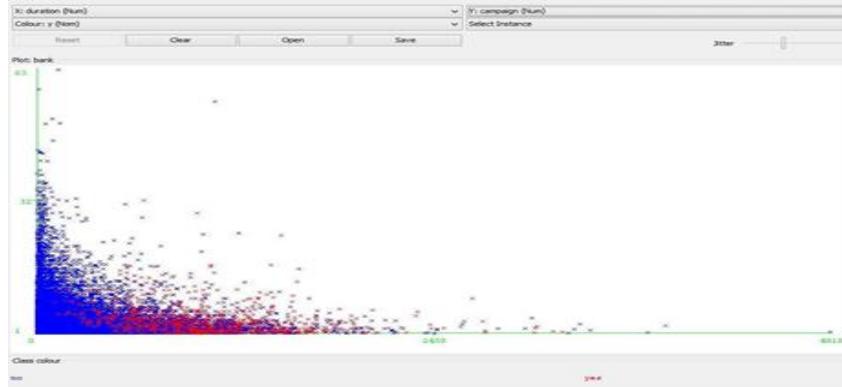

Fig. 2 The relation between Duration and Campaign

IV. EXPERIMENTAL RESULTS

This section presents the results obtained after running the two classification techniques for bank direct marketing data set. To construct the algorithms, we use Waikato Environment for Knowledge Analysis, an open source data mining too. Which was developed at University of Waikato New Zealand WEKA is an open source application that is freely available under the GNU general public license agreement. All experiment were performed on Duo Core with 1.8GHz CPU and 4G RAM. The result for each classification algorithms. We applied two algorithms on the dataset; One-R and Naïve-Bayes and we get the following:

A. Results for classification using Naive-Bayes Algorithm:
Naïve baiyes (Bouckaert, 2008) assumed the variables are independence, After applying the algorithm on the data by "Weka" program at the first time, the result was:
As we see in the following figures, the percentage of every variable how it effect the class variable are shown in Table 2:

TABLE 2 SUMMARY

Correctly Classified Instances	39991	88.4541 %
Incorrectly Classified Instances	5220	11.5459 %
Kappa statistic	0.4524	
Mean absolute error	0.1514	
Root mean squared error	0.3012	
Relative absolute error	73.2828 %	
Root relative squared error	93.7234 %	
Total Number of Instances	45211	

Confusion Matrix:
The confusion matrix One of the popular performance evaluation methods in data mining, machine learning artificial, intelligence and statistics is confusion matrix. In order to quantify the performance of a problem that contains two classes, confusion matrix is usually used (L.Gupta, K. Malviya, & Singh, 2012). That mean, the algorithm classified 37187 instance as nowhere they are no and that correct and classified 2735 instance as yes and they are no and that false
Also, the algorithm classified 2485 instance as nowhere they are yes and that false and classified 2804 as yes where they are yes and that true are shown in Table 3:

Table 3 Confusion Matrix

classified as	A	B
A = NO	37187	2735
B = YES	2485	2804

B. *Results for classification using* One-R Algorithm:

When applying the algorithm One-R Algorithm (Garg & Khurana, 2014) for the first time we find from the result that "Duration" attribute will be the attribute which determines the value of the class variable are shown in Table 4:

TABLE 4 SUMMARY

Correctly Classified Instances:	40413	(89.3875%),
Incorrectly Classified Instances:	4798	(10.6125%).
Kappa statistic	0.2157	
Mean absolute error	0.1155	
Root mean squared error	0.3399	
Relative absolute error	55.903 %	
Root relative squared error	105.742 %	
Total Number of Instances	45211	

Confusion Matrix:

The confusion matrix mean that the algorithm classified 39062 instance as nowhere they are no and that correct and classified 860 instance as yes and they are no and that false

Also, the algorithm classified 4362 instance as nowhere they are yes and that false and classified 927 as yes where they are yes and that true are shown in Table 5:

Table 5 Confusion Matrix

CLASSIFIED AS	A	B
A = NO	39062	860
B = YES	4362	927

V. CONCLUSIONS

Bank direct marketing and business decisions are more important than ever for preserving the relationship with the best customers. For success and survival in the business, there is a need for customer care and marketing strategies.

Data mining and predictive analytics can provide help in such marketing strategies. After implementing the two algorithms on the dataset for the first time. The dataset obtained from UCI, which contain the most common machine learning datasets. We discovered that the One-R algorithm is better in classifying the data comparing with the Naïve-Bayes algorithm; where the error rate is lower.

We suggest to use the "Duration" attribute as a classifier to the data set, in the future when a new data come, we can see the "Duration" attribute and according to it we can detect the class variable if will be "yes" or "no".

REFERENCES

[1]   A.Elsalamony, H. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. *International Journal of Computer Applications*, *85*(7), 12–22. https://doi.org/10.5120/14852-3218

[2]   Ayetiran, E. F., & Adeyemo, A. B. (2012). A data mining-based response model for target selection in direct marketing. *IJ Information Technology and Computer Science*, *1*, 9–18.

[3]   Bouckaert, R. R. (2008). Bouckaert - Bayesian Nets in Weka, 23. Retrieved from papers2://publication/uuid/3A4EABD9-258F-4C97-981E-2D70365FDE28

[4]   Cortez, P. (n.d.). A Data-Driven Approach to Predict the Success of Bank Telemarketing, (February 2014).

[5]   Elsalamony, H. A., & Elsayad, A. M. (2013). Bank Direct Marketing Based on Neural Network and C5. 0 Models. *International of Engineering and Advanced Technology*, *2*(6), 392–400.

[6]   Garg, T., & Khurana, S. S. (2014). Comparison of classification techniques for intrusion detection dataset using WEKA. *International Conference on Recent Advances and Innovations in Engineering, ICRAIE*

2014, (May). https://doi.org/10.1109/ICRAIE.2014.6909184

[7]   Han, J., Kamber, M., & Pei, J. (2011a). Concepts etc techniques.

[8]   Han, J., Kamber, M., & Pei, J. (2011b). *Data Transformation by Normalization. Data Mining: Concepts and Techniques*. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

[9]   Ian H. Witten, Eibe Frank, M. A. A. H. (2013). Data Mining Practical Mchine Learning Tools and Techniques. *Morgan Kaufmann*. Retrieved from http://ir.obihiro.ac.jp/dspace/handle/10322/3933

[10] L.Gupta, D., K. Malviya, A., & Singh, S. (2012). Performance Analysis of Classification Tree Learning Algorithms. *International Journal of Computer Applications*, *55*(6), 39–44. https://doi.org/10.5120/8762- 2680.

[11] Moin, I. K., & Ahmed, B. Q. (2012). Use of Data Mining in Banking. *International Journal of Engineering Research and Applications*, *2*(2), 1–5. Retrieved from http://ijera.com/papers/Vol2_issue2/DU22738742.pdf

[12] Nachev, A. (2014). Application of Data Mining Techniques for Direct Marketing. *Ithea*, 86–95.

[13] Reutemann, P., Hall, M., Frank, E., Witten, I. H., Holmes, G., & Pfahringer, B. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10. https://doi.org/10.1145/1656274.1656278

[14] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[15] Su, C. T., Chen, Y. H., & Sha, D. Y. (2006). Linking innovative product development with customer knowledge: a data-mining approach. *Technovation*, *26*(7), 784–795. https://doi.org/10.1016/j.technovation.2005.05.005