



Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data

Tuba Parlar¹, Songul Kakilli Acaravci^{2*}

¹Vocational School of Antakya, Department of Computer Technology, Mustafa Kemal University, Antakya, Hatay, Turkey, ²Faculty of Economics and Administrative Sciences, Department of Finance and Accounting, Mustafa Kemal University, Hatay, Turkey.

*Email: sacaravci@mku.edu.tr

ABSTRACT

Collection of customer information is seen necessary for development of the marketing strategies. Developing technologies are used very effectively in bank marketing campaigns as in many field of life. Customer data is stored electronically and the size of this data is so immense that to analyse it manually with a team of human analysts is impossible. In this paper, data mining techniques are used to interpret and define the important features to increase the campaign's effectiveness, i.e., if the client subscribes the term deposit. The bank marketing dataset from the University of California at Irvine Machine Learning Repository has been used for the proposed paper. We consider two feature selection methods namely information gain and Chi-square methods to select the important features. The methods are compared using a supervised machine learning algorithm of Naive Bayes. The experimental results show that reduced set of features improves the classification performance.

Keywords: Bank Marketing, Feature Selection, Machine Learning Methods, Data Mining, Chi-square, Information Gain

JEL Classifications: C80, C50, Y10, M31

1. INTRODUCTION

The global banking sector is rapidly changing and developing in which innovations are applied widely. Especially, in recent years, redesigning of working manner and activity structure in the banking sector has become widespread in all around the world. The developments in behavior and preferences of consumers, competition from different sectors and continuously changing legislations have created serious pressure on banks. Today, customer satisfaction is more important key factor on being ahead of this highly competitive sector. To increase customer satisfaction, banks should improve with creative products and distribution channels to make differentiation in the powerful competition environment (Karaslar, 2014. p. II).

Banks store huge amount of information about their customers to offer them for several campaigns or products. Banks reach the customers information a multitude of channels, including mail, e-mail, phone, in person for sharing information about products or services. Meeting the demands of the customers directly through

a variety of media is called direct marketing. This term is first introduced and defined by Lester Wunderman in 1967. 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 (Elsalamony and Elsayad, 2013. p. 392; Elsalamony, 2014. p. 12).

Direct marketing is the process of identifying possible customers of products and promoting these products to this specified customer mass. Recently, due to the fact that mass marketing campaigns targeting general public are not successful, firms give more importance to direct marketing campaigns targeting a specific set of customers. Direct marketing methods are more successful especially in banking sector where there is more pressure and competition according to other sectors. Data mining methods are used to increase the success of direct marketing campaigns by identifying the factors that affect these campaigns. Thus, these methods provide to direct available resources and to create a reasonable and true set of potential customers (Dolgun and Ersel, 2014. p. 1).

Quite a few studies have been conducted on bank direct marketing domain. Some researchers have used machine learning algorithms to classify the model according to data mining approaches (Moro et al, 2014; Elsalamony and Elsayad, 2013; Sharma et al., 2015). Data mining approaches aim to build a predictive model that labels data into a predefined class (for example “yes” or “no”). Because of all bank marketing strategies are dependent to analyze huge electronic data of customers. It is impossible for a human analyst to evaluate the meaningful knowledge from the vast amount of data. Many researchers use some of popular data mining techniques such as Naïve Bayes (NB), decision trees, support vector machines, and logistic regression. The purpose is increasing the campaign effectiveness by identifying the main characteristics that affect the success (Elsalamony and Elsayad, 2013. p. 400; Elsalamony, 2014. p. 21; Sharma et al., 2015. p. 7141, 7147). The success of such campaigns is directly proportional to the effective customer participation. The success rate of such campaigns that especially conducted by banks may be enhanced by data mining methods (Akçetin, 2015. p. 5).

This paper aims to improve the efficiency of the marketing campaigns and helping the decision makers by reducing the number of features of bank direct marketing data. We use different features selecting methods to select the most important features. Thus, this study may also make meaningful contribution to literature. The rest of the paper is organized as follows: Section 2 introduces methodology used, section 3 explains dataset used and experimental setup, finally section 4 shows the conclusion.

2. METHODOLOGY USED

In this study, we use data mining method. Information technology is evolving and growing rapidly. Data mining mines meaningful knowledge from large amounts of data. Actually, data mining is used as a synonym for knowledge discovery from data (KDD) term. Data mining is one of the important phase of this KDD process. Data become meaningful when they are processed for finding knowledge. Today, using data mining techniques which can process large amount of data is becoming more important. Finding the meaningful information is possible with data mining (Savaş, et al., 2012. p. 1). Data mining include descriptive and predictive techniques for meaningful knowledge which is unknown early from data (Albayrak and Yılmaz, 2009. p. 31-52). We use Weka data mining software¹ for analyzing the bank direct data. Weka is a collection of machine learning algorithms for data mining tasks (Witten et al., 2011). Weka is the product of the University of Waikato (New Zealand) and written in the Java programming language. It uses the General Public Licence.

2.1. Feature Selection Methods

Feature selection is an important procedure that involves two main tasks. First, decreasing the size of the features is important because of the efficiency of the training process.

Second, eliminating invaluable features increases the accuracy of the classification. Feature selection algorithms are based on statistical measures that assign a scoring value to each feature then the features are ranked by this value. Information gain (IG) and Chi-square methods are the most common feature selection methods.

2.1.1. Chi-square feature selection method

Chi-square measures the lack of independence between a feature and a class. We can define Chi-square statistic using the 2-by-2 contingency table given in Table 1.

In Table 1, A is the frequency in class c that feature f occurs in, B represents the frequency in the other class that feature f occurs in, C represents the frequency in class c that feature f does not occur in, and D is the frequency in the other class that feature f does not occur in.

The Chi-square statistic is computed as follows (Witten et al., 2011).

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{1}$$

Where, i and j are columns and rows in a contingency table. o_{ij} is the observed frequency and e_{ij} is the expected frequency of tuple (i,j). Then the Chi-square can be defined as follows:

$$\chi^2 = \frac{N(AD-BC)^2}{(A+B)(A+C)(B+D)(C+D)} \tag{2}$$

2.1.2. Information gain feature selection method

IG is one of the important feature selection methods that measures the content of information by analysing the feature is known. IG of a feature can be calculated as follows (Yang and Pederson, 1997):

$$IG(f) = - \sum_{i=1}^m P_r(c_i) \log P_r(c_i) + P_r(f) \sum_{i=1}^m P_r(c_i|f) \log P_r(c_i|f) + P_r(\bar{f}) \sum_{i=1}^m P_r(c_i|\bar{f}) \log P_r(c_i|\bar{f}) \tag{3}$$

Where m is the number of classes and p_i is the probability of a feature f in class c_i .

2.2. Naive Bayes Machine Learning Algorithm

Naive Bayes method is based on Bayes’ rule to compute conditional probabilities and assumes that attributes are independent. The

Table 1: Contingency table

	Frequency in class	Frequency not in class
Feature in document	A	B
Feature not in document	C	D

¹ <http://www.cs.waikato.ac.nz/ml/weka/>.

probability $P(c|d)$ that the document d in the class c and we can calculate this probability as follows:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{4}$$

The probability $P(d|c)$ is calculated with some assumptions related to the structure of document d . In a document feature vector $d = (f_1, f_2, \dots, f_i)$, it is assumed that all features are independent of each other, and the classifier resulting from this assumption is called Naive Bayes:

$$P(d | c) = \prod_i P(f_i|c) \tag{5}$$

We chose NB classifier for our experiments included in Weka Data Mining software (Witten et al., 2011).

3. DATASET USED AND EXPERIMENTAL SETUP

3.1. Dataset

We use bank marketing dataset that is publicly available for research from the University of California at Irvine Machine Learning Repository in our experiments². The bank direct marketing dataset is introduced by Moro et al. (2014). This data based on direct marketing campaigns of a Portuguese banking institution. The marketing campaigns are based on phone calls and related to 17 campaigns, which occurred from May 2008 to November 2010. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. As you can see from Table 2, Seventeen attributes are stored for each contact.

3.2. Measuring Performance

We compare the performance of the classifier based on precision, recall, and F measure values. According to Han and Kamber

2 <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>.

(2006), a set of documents relevant to a query be denoted as $\{\text{Relevant}\}$, and the set of documents retrieved be denoted as $\{\text{Retrieved}\}$. The set of documents that are both relevant and retrieved is denoted as $\{\text{Relevant}\} \cap \{\text{Retrieved}\}$. Precision (P) is the percentage of retrieved documents that are relevant to the query and defined as follows:

$$P = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|} \tag{6}$$

Recall (R) is the percentage of documents that are relevant to the query and retrieved. Recall is defined as follows:

$$R = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|} \tag{7}$$

Thus, F measure is defined as follows (Han and Kamber, 2006):

$$F = 2 \frac{P \times R}{P + R} \tag{8}$$

3.3. Baseline Results

Our aim to evaluate applicability of feature selection methods on bank direct marketing dataset. Therefore, first we obtain the baseline results by using NB classifier on the bank marketing dataset to compare the further analysis. The experiment is validated using five-fold cross validation. Table 3 presents performance of baseline settings in terms of precision, recall, and F measure of the NB classifier. The average weighted F measure result are shown in boldface in Table 3.

3.4. Feature Selection Results

We tested two feature selection methods namely IG and Chi-square for reduced feature sizes to see the classification performances. We tried four feature sizes (5, 8, 10, 15) for the

Table 2: Explanations of features of the bank dataset

Features	Explanations
“Age”	Numeric
“Job”	“Type of job:” Categorical: “Administrative,” “unknown,” “unemployed,” “management,” “housemaid,” “entrepreneur,” “student,” “blue-collar,” “self-employed,” “retired,” “technician,” “services”
“Marital”	“Marital status:” Categorical: “Married,” “divorced,” “single,” (note: “divorced” means divorced or widowed)
“Education”	Categorical: “Unknown,” “secondary,” “primary,” “tertiary”
“Default”	“Has credit in default?” (binary: “Yes,” “no”)
“Balance”	“Average yearly balance, in euros” (numeric)
“Housing”	“Has housing loan?” (binary: “Yes,” “no”)
“Loan”	“Has personal loan?” (binary: “Yes,” “no”)
“Contact”	“Contact communication type:” Categorical: “Unknown,” “telephone,” “cellular”
“Day”	“Last contact day of the month” (numeric)
“Month”	“Last contact month of year” (categorical: “Jan,” “Feb,” “Mar,” ..., “Nov,” “Dec”)
“Duration”	Last contact duration, in seconds (numeric)
“Campaign”	“Number of contacts performed during this campaign and for this client” (numeric, includes last contact)
“Pdays”	“Number of days that passed by after the client was last contacted from a previous campaign” (numeric, -1 means client was not previously contacted)
“Previous”	“Number of contacts performed before this campaign and for this client” (numeric)
“Pout come”	“Outcome of the previous marketing campaign” (categorical: “Unknown,” “other,” “failure,” “success”)
“Y”	“Has the client subscribed a term deposit?” (“yes” or “no”)

bank marketing dataset. Experiments are run using NB classifier with five-fold cross validation. The experiment is validated using five-fold cross validation that means the whole dataset is divided into five equal sized sets and classifier is trained on 4 train sets and tested on one test set. This process is repeated five times and then an average of all folds is taken as a result. Tables 4 and 5 present performance of IG and Chi-square feature selection methods in terms of precision, recall, and F measure values using NB classifier when trained and tested on the reduced size of features. As can be seen in Tables 4 and 5, the best result is achieved with IG feature selection method with 5 features: (1) Duration, (2) pout come, (3) month, (4) pdays, and (5) contact. The F measure is increased from 0.873 to 0.883. However, as can be seen from Tables 4 and 5, both IG and Chi-square methods are achieved 0.880 in F measure for ten features. These highest ranked ten features are (1) duration, (2) pout come, (3) month, (4) pdays, (5) contact, (6) previous, (7) age, (8) job, (9) housing, and (10) balance.

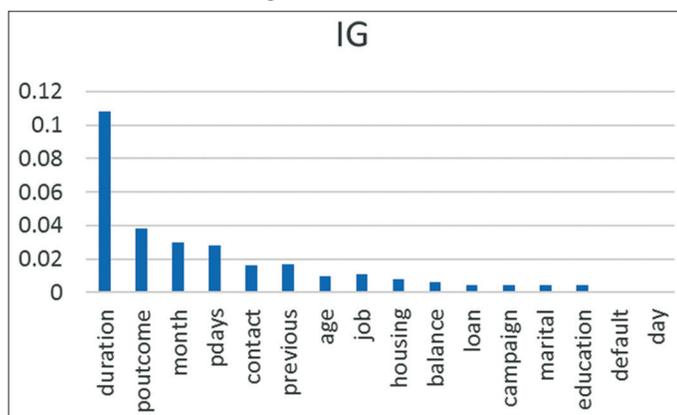
Also observed in Tables 4-6 and Figure 1 is that IG and Chi-square feature selection methods have almost the same performance for the bank direct marketing dataset. The results of IG and Chi-square feature selection methods are presented in Table 6 in a descending order. "Duration" that means the contact duration time is most important feature according to both methods. Second key feature is the "pout come" that means outcome of the previous marketing campaign.

4. CONCLUSIONS

Economic conditions affect business organizations and banking sectors. Accordingly, marketing managers are in need to increase marketing campaigns. The most challenging issues appear due to the huge amount of data records. Data mining has a key role to analyze this huge amount of data. Managers can reshape their businesses and campaign strategies using data mining tools.

In this paper, we have analysed two feature selection methods on the bank direct marketing dataset. Our results show that using feature selection methods, we can decide the most important

Figure 1: The importance levels of the features according to information gain feature selection method



features and rebuild our future decisions. IG and Chi-square feature selection methods can compute the importance of the features. The results for IG and Chi-square feature selection methods are very close, although they are different for first five highest ranked features. The performance of classification is improved with reduced size of features. According to IG and Chi-square feature selection methods, the highest ranked ten features are (1) duration, (2) pout come, (3) month, (4) pdays, (5) contact, (6) previous, (7) age, (8) job, (9) housing, and (10) balance.

In the future, we would like to compare the classification performances of different datasets from different domains with different methods.

Table 3: Baseline results for the dataset

	Has the client subscribed a term deposit?						
	Positive class			Negative class			Average
	P	R	F	P	R	F	F
Baseline	0.440	0.509	0.472	0.935	0.916	0.925	0.873

Table 4: Results of IG feature selection method

Number of top feature	Has the client subscribed a term deposit?						
	Positive class			Negative class			Average
	P	R	F	P	R	F	F
5	0.519	0.411	0.459	0.925	0.951	0.938	0.883
8	0.457	0.461	0.459	0.930	0.929	0.929	0.875
10	0.480	0.476	0.478	0.932	0.933	0.932	0.880
15	0.444	0.507	0.473	0.935	0.917	0.926	0.874

IG: Information gain

Table 5: Results of Chi-square feature selection method

Number of top feature	Has the client subscribed a term deposit?						
	Positive class			Negative class			Average
	P	R	F	P	R	F	F
5	0.472	0.424	0.447	0.926	0.938	0.932	0.876
8	0.457	0.461	0.459	0.930	0.929	0.929	0.875
10	0.480	0.476	0.478	0.932	0.933	0.932	0.880
15	0.444	0.507	0.473	0.935	0.917	0.926	0.874

Table 6: Ranking features according to feature selection methods

Features	IG	Chi-square
Duration	0.108	625.778
Pout come	0.038	310.434
Month	0.030	203.069
Pdays	0.028	202.034
Contact	0.016	98.163
Previous	0.017	70.744
Age	0.010	74.335
Job	0.011	57.986
Housing	0.008	39.799
Balance	0.006	27.91
Loan	0.004	18.124
Campaign	0.003	15.907
Marital	0.003	15.657
Education	0.002	12.657
Default	0	0.032
Day	0	0

IG: Information gain

REFERENCES

- Akçetin, E. (2015), Karınca kolonisi optimizasyonu sınıflandırma algoritması yöntemi ile telefon bankacılığında doğrudan pazarlama kampanyası üzerine bir sınıflandırma analizi. *İnternet Uygulamaları ve Yönetimi*, 6(1), 5-19.
- Albayrak, A.S., Yılmaz, Ş.K. (2009), Veri madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 14(1), 31-52.
- Dolgun, M.Ö., Ersel, D. (2014), Doğrudan pazarlama stratejilerinin belirlenmesinde veri madenciliği yöntemlerinin kullanımı. *İstatistikçiler Dergisi: İstatistik and Aktüerya*, 7, 1-13.
- Elsalamony, H.A. (2014), Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications*, 85(7), 12-22.
- Elsalamony, H.A., Elsayad, A.M. (2013), Bank direct marketing based on neural network. *International Journal of Engineering and Advanced Technology (IJEAT)*, 2(6), 392-400.
- Han, J., Kamber, M. (2006), *Data Mining: Concepts and Techniques*. 2nd ed. San Mateo, CA: Morgan-Kaufman Series in Data Management Systems.
- Karaslar, S. (2014), Bankacılık Sektöründe Pazarlama Uygulamaları: Garanti Bankası Örneği. *Trakya Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı Yüksek Lisans Dönem Projesi*.
- Liu, B. (2007), *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Data-Centric Systems and Applications*. Berlin: Springer.
- Moro, S., Cortez, P., Rita, P. (2014), A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22-31.
- Savaş, S., Topaloğlu, N., Yılmaz, M. (2012), Veri madenciliği ve Türkiye'deki uygulama örnekleri, *Istanbul ticaret üniversitesi. Fen Bilimleri Dergisi*, 11(21), 1-23.
- Sharma, N., Kaur, A., Gandotra, S., Sharma, B. (2015), Evaluation and comparison of data mining techniques over bank direct marketing. *International Journal of Innovative Research in Science, Engineering and Technology*, 4(8), 7141-7147.
- Witten, I.H., Frank, E., Hall, M.A. (2011), *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. San Francisco, California: Morgan Kaufmann.
- Yang, Y., Pederson, J.O. (1997), A comparative study on feature selection in text categorization. In: *Proceedings of the 14th International Conference on Machine Learning*. p412-420.