Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 4 No.2, December 2019, pp. 57 – 66

Prediction on Deposit Subscription of Customer based on Bank Telemarketing using Decision Tree with Entropy Comparison

Ardytha Luthfiarta[1], Junta Zeniarja[2],
Universitas Dian Nuswantoro (UDINUS)
Semarang, Indonesia
E-mail : [1]ardytha.luthfiarta@dsn.dinus.ac.id,[2]junta@dsn.dinus.ac.id
**Corresponding author*

Edi Faisal[3], Wibowo Wicaksono[4]
Universitas Dian Nuswantoro (UDINUS)
Semarang, Indonesia
E-mail : [3]edi.faisal@dsn.dinus.ac.id, [4]wibowo.wicaksono@dsn.dinus.ac.id

Abstract—Banking system collect enormous amounts of data every day. This data can be in the form of customer information, transaction details, risk profiles, credit card details, limits and collateral details, compliance Anti Money Laundering (AML) related information, trade finance data, SWIFT and telex messages. In addition, Thousands of decision are made in Banking system. For example, banks everyday creates credit decisions, relationship start up, investment decisions, AML and Illegal financing related decision. To create this decision, comprehensive review on various reports and drills down tools provided by the banking systems is needed. However, this is a manual process which is error prone and time consuming due to large volume of transactional and historical data available. Hence, automatic knowledge mining is needed to ease the decision making process. This research focuses on data mining techniques to handle the mentioned problem. The technique will focus on classification method using Decision Tree algorithms. This research provides an overview of the data mining techniques and procedures will be performed. It also provides an insight into how these techniques can be used in deposit subscription in banking system to make a decision making process easier and more productive.

Keywords - Telemarketing, bank deposit, decision tree, classification, data mining, entropy.

1. INTRODUCTION

Nowadays, in the era of globalization and competition, companies struggles to compete with each other. This phenomenon also happens in the field of banking. In order to win the competition, each banks are expected to come with excellent strategy tools. It means that, the success of banks not only depends on the execution of business processes, but also more importantly on the creation of knowledge. The creation of knowledge, or knowledge mining, in banking system is supported by the banks' ability to generate, capture and store enormous in recent years. The information that can be extracted from this data can be very useful. The huge amounts of raw data available and the urgency to transform those data into knowledge encourage IT industry to use data mining. In banking fields, Leading banks are already use Data Mining (DM) tools to perform knowledge mining[1][2][3][4]. For example, banks use DM to create customer segmentation and profitability data, to decide credit scoring and approval, to predict payment default to target best marketing segment, and to detect fraudulent transactions[5][6].

How Data Mining can contribute in solving business problems is by finding patterns, associations and correlations, which are hidden in the databases . Because banking is in the service industry, the task of maintaining a strong and effective Customer Relationship Management is a critical issue[5][7][8]. This task can be performed by several ways, one of which is by performing direct marketing. Direct marketing is used to target segments of customers which meet a specific criteria. Direct marketing can be performed by either using fixed-line or mobile phone. Because of the remote characteristic of this marketing, it is also called telemarketing[2]. It should be stressed that the task of selecting the best set of clients, i.e., that are more likely to subscribe a product, is considered NP-hard in Ref. [3]

Research in the field of banking client targeting by telemarketing have been performed previously[4]. The research of Martens et all [9] identified clients for targeting at a major bank using pseudo-social networks based on relations (money transfers between stakeholders). Their approach offers an interesting alternative to traditional usage of business characteristics for modeling. Other work by Sérgio Moro, Raul Laureano, Paulo Cortez [10], explored data-driven models for modeling bank telemarketing success. Yet, in the mentioned research, good models is only achieved when using attributes that are only known on call execution, such as call duration[11][12]. Thus, while providing interesting information for campaign managers, such models cannot be used for prediction.[13][14]

The main aim of this research is to predict the customer deposit subscription in telemarketing based on the possible attributes. The main contributions of this work are focus on feature engineering, which is a key aspect in DM. In addition, we propose generic social and economic indicators to be added to the more commonly used bank client and product attributes. We analyze a 10% dataset (4119 records) randomly selected from a Portuguese bank. The data were collected from 2008 to 2013, thus including the effects of the global financial crisis that peaked in 2008. The sources of data is from Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) and Paulo Rita (ISCTE-IUL).

We are using a Decision Tree model with a different criterion to get a better accuracy for the customer deposit subscription in a bank telemarketing business. The paper is organized as follows: Section 2 presents the literature review; Section 3 describes the proposed work; Section 4 is experimental result; finally, conclusions and future work in Section 5.

2. RESEARCH METHOD

Vivek Bhambri [11] proposed that creation of knowledge base and its utilization for the benefit of the organization is becoming a strategy tool to compete. Hence, the need of the techniques like data mining can help them to compete in the market especially in banking sector. Besides, trends can also be analyzed and predicted with the availability of historical data and the data warehouse assures that everyone is using the same data at the same level of extraction, which eliminates conflicting analytical results and arguments over the source and quality of data used for analysis.

According to Dr. K. Chitra, B. Subashini [15], data mining is becoming a strategically important area for many business organizations including banking sector. It is a process of analyzing the data from various perspectives and summarizing it into valuable information. Data mining assists the bank to look for a hidden pattern in a group and discover unknown relationship in the data. The researchers proposed that by using data mining techniques it is simple to build a successful predictive model and visualize the report into meaningful information to the user since Banks nowadays realized the important factor of their success is Customer Relationship Management (CRM). CRM is the strategy that can help them build long-lasting relationships with their customers and increase their revenues and profits. The challenges that bank faced were how to retain the most profitable customers and how to do that at a lower cost. At the same time they need to find and implement this solution quickly and the solution to be flexible. By using traditional method of data analysis required complex and time- consuming investigations that deals with different domains of knowledge like financial, economics, business practices and law.

Safia Abbas [16] proposed two classification Data Mining techniques, decision tree (DT) and rough set theory (RST) using real world set collected from Portuguese marketing campaign and concerned about customer deposit subscription. The research utilize those two techniques aiming to find the redact set, the minimal set of attributes that can discriminate between objects with respect to the approximations of the information spaces, extract predictive rules with accuracy to aid in decision making and avoiding risks, compare between the early analysis techniques and DM techniques results. In other words, the aim of this paper is to improve the efficiency of the marketing campaigns and helping the decision makers by reducing the number of features, that describes the dataset and spotting on the most significant ones, and predict the deposit customer retention criteria based on potential predictive rules.

Kazi Imran Moin, Dr. Qazi Baseer Ahmed [1] proposed to implement the concept of data warehousing and data mining in banking sector. In the financial services industry throughout the world, the traditional face-to-face customer contacts are being replaced by electronic points of contact to reduce the time and cost of processing an application for various products and ultimately improve the financial performance. The computerization of financial operations, use of internet and automated software has completely changed the basic concept of business and the way the business operations are being carried out. The amount of data collected by banks has grown rapidly in recent years. Existing statistical data analysis techniques find it difficult to manage with the large volumes of data now available. This explosive growth has lead to the need for new data analysis techniques and tools in order to find the information hidden in this data. Banking area is an area where vast amount of data are collected. This data can be generated from bank account transactions, loan repayments, credit card repayments, etc. It is assumed that valuable information on the financial profile of customers is hidden within these massive operational databases and this information can be used to improve the performance of the bank. Researcher suggest Data mining algorithm that can be used is classification in fraud detection and credit risk applications using decision tree (DT) or Neural Network (NN), Clustering techniques for cross-selling activities, Prediction techniques like Linear Regression for product failure rate.

Karl D. Majeske, Thomas W. Lauer [17] proposed a probability model to evaluate the predictive validity of two- way classification schemes in the context of personal credit scoring and bank loan applications. The researchers suggest a Bayesian decision model provides a structure for identifying classification rules that lead to optimal-maximum expected payoff or minimum expected cost-classifications. Using payoffs from multiple perspectives allows identifying conditions where the various perspective produce contradictory classifications generating either profit premiums or cost penalties depending on the perspective. The problems are framed as a decision maker who must place an individual into one of two categories based on a set of data or attributes.

In this research, the main aim is to predict the customer deposit subscription in the direct marketing field which is telemarketing based on the possible attributes. In this research we find out the useful knowledge from the real dataset related to the particular domain. We are using a Cross-Industry Standard Process for the implementation. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology for increasing the success of DM projects [18]. This methodology defines a sequence of six phases, which allow the building and implementation of a DM model to be used in a real environment. CRISP-DM defines in a cyclic process, where several iterations can be used to allow final result more tuned towards the business goals. This methodology can be shown in Figure 1.

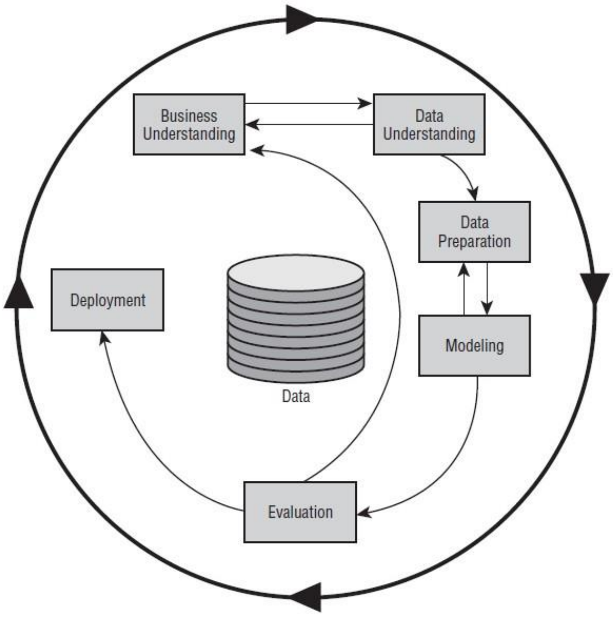


Figure 1: The CRISP-DM process model (adapted from Chapman et al., 2000)

2.1. Business Understanding

Business understanding is about defining the objective of the project and converting this knowledge into a data mining problem definition. For this research, the goal is to predict the customer deposit subscription in the direct marketing field which is telemarketing based on the possible attributes from database.

2.2. Data Understanding

Data understanding phase is about identify the data quality problems or detect the interesting subsets to form hypothesis for hidden information. In this research, a dataset total of 4119 records are randomly selected from Portuguese bank were analyzed. The potential attribute which related to the predicted label which is customer bank deposit subscription were analyzed.

2.3. Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. For this research, only the nominal result was accounted, thus the goal is to predict if a customer will subscribe the deposit or not. The possible categorized attributes that related to the customer deposit subscription are demographics, last contact of the current campaign, and social and economic context attributes. There were 20 total regular attributes and one nominal label which is customer deposit subscription (“Yes” or “No”). In addition to input attribute, there were also several instances with missing values that are stated as “Unknown”.

2.4. Modeling

In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Most of the dataset for this research are nominal. The suitable model technique for this research is classification because classification technique using a nominal data. For this research we have considering a various classification algorithm such as Decision Tree (DT), Naïve Bays (NB), Random Forest (RF) and k-nearest neighbors (K-NN). We have selected a Decision Tree, because it is most suitable model for the nominal data and also it is understandable by human. Simplification was needed if knowledge was to be extracted successfully.

2.5. Evaluation

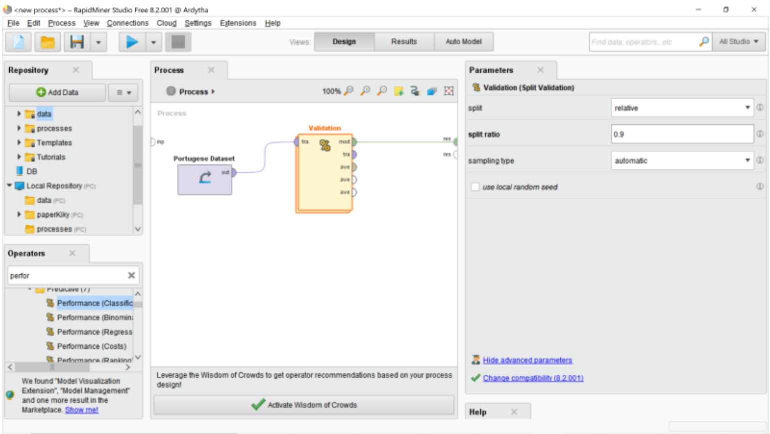
At this stage the models obtained are more thoroughly evaluated and the steps executed to construct the model are in terms of its performance and utility. For this research, the data accuracy was tested using Rapid Miner. The result shows that the Decision Tree accuracy is the most highest compared to Naïve Bayes (NB), Random Forest (RF) and k-nearest neighbors (K-NN). We also compare the accuracy value for the Decision Tree with a different criterion which is gain ratio, information gain, gini index and accuracy.

2.6. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained need to be organized in a way which everyone can understand. If the obtained model is not good enough for use to support business, then a new iteration for the CRISP-DM is defined. Else, the model is implemented in a real time environment. For this research, it can be seen that the Decision Tree algorithm is suitable for implementation and also suitable with the aim of this research.

3. RESULT & DISCUSSION

All experiments were performed using the rapid miner software. Each DM model related with this section was executed using rapid miner. For the feature selection, we adopted the Decision Tree (DT) model described in section (1 C) as the base of DM methods. The 10% from dataset (4119 records) randomly selected from a Portuguese bank. The data were collected from 2008 to 2013, thus including the effects of the global financial crisis that peaked in 2008.



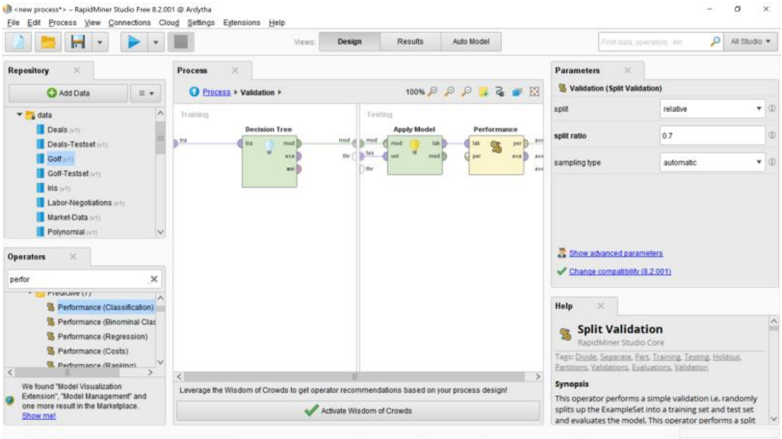
Validation (Split Validation)
split: relative
split ratio: 0.9
sampling type: automatic
use local random seed

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

We found "Model Visualization Extension", "Model Management" and one more result in the Marketplace. Show me!

Figure 2: The validation process of the dataset

Based on **Figure 2** shown above, the process begin by choosing the dataset that want to be tested. In this research we are using dataset of Bank Marketing [19].The dataset should be upload first into the Rapid miner Software. To test the dataset, we must validate the dataset by using Evaluation operator (Evaluation > X-Validation) as shown in the figure below. Connect dataset with the Validation operator and from the Validation operator to the result (res).From Validation to result, it should have 3 line connected each other. Each line correspond to data training, data testing section. Double click the Validation operator to run the validation process as shown in **Figure 3** below.



Validation (Split Validation)
split: relative
split ratio: 0.7
sampling type: automatic

Split Validation
RapidMiner Studio Core
Tags: Divide, Separate, Part, Training, Testing, Holdout, Partitions, Validations, Evaluations, Validation
Synopsis
This operator performs a simple validation i.e. randomly splits up the ExampleSet into a training set and test set and evaluates the model. This operator performs a split

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

We found "Model Visualization Extension", "Model Management" and one more result in the Marketplace. Show me!

Figure 3: Validation Process of dataset include in two section; Training section (Decision Tree model) and Testing section(Apply model and Performance).

From **Figure 3**, it show how the dataset will be validate to obtain the result (Validation process). There are two section for validation process; Training and Testing. In Training section the operator will be the algorithm that are chosen to be tested. For this research, we choose Decision tree Algorithm (Modelling >

Classification and Regression > Tree Induction > Decision Tree) . In the Testing Section, the operator will be the Apply Model operator (Modelling > Model Application > Apply Model) and Performance operator (Evaluation > Performance Measures >Performance). Connect each operator as shown in the figure above. Then click to run the process.

In this research, the dataset were tested to find the accuracy among the criterion under decision tree (DT) algorithm in order to find the best criterion that suitable with the dataset and the algorithm chosen. The criterion are as listed below and as shown in **Table 1**:

i. gain_ratio
ii. information_gain
iii. gini_index
iv. accuracy

Table 1: Criterion involve with the DT algorithm with the accuracy.

Criterion	Accuracy(%)
Gain_ratio	89.59
Information_gain	89.17
Gini_index	90.09
Accuracy	89.15

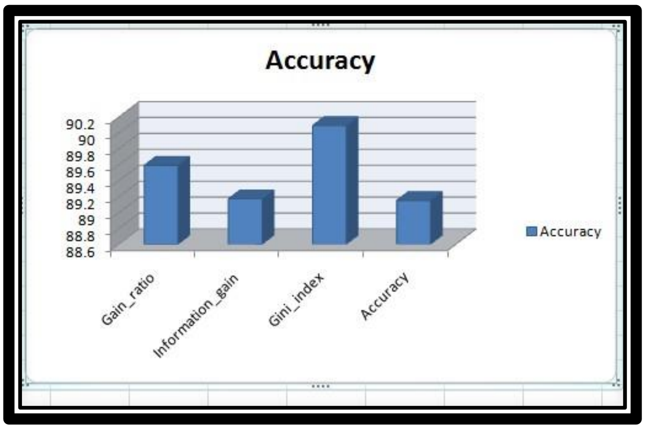


Figure 4: Accuracy of criterion for Decision Tree algorithm for dataset.

From the Table 1 above, among all criterion tested, the most high accuracy is gini_index with the percentage of accuracy 90.09% and the least accurate criterion for decision tree algorithm is the accuracy criterion with percentage 89.15% shown as in **Figure 4**. **Table 1** indicates that to classified accurately deposit subscription of customer based on bank telemarketing using Decision tree algorithm with the criterion gini_index.

4. CONCLUSION

In this paper, we apply a Data Mining approach to bank direct marketing campaigns which is to predict the customer deposit subscription. In particular, we used a recent data from a Portuguese bank[19] and performed a CRISP-DM methodology, in order to get a best Data Mining model results. Based on the experimental result, the best model, which is Decision Tree (DT), achieved high predictive performances among other tested model such as Naïve Bays (NB), Random Forest (RF) and k-nearest neighbors (K-NN). The Decision tree approach has been implemented on the data set and C4.5 classifier has been used in classification process. DT provides a wide number of decision and predictive rules associated with accuracy. Some of the DT extracted rules are summarized. The decision tree is easy to be implemented as a classifier. We measured importance input using Decision Tree model and the knowledge gain can be used by managers of the bank to enhance campaigns (for example increase a length of phone call) for the telemarketing predicting outcome of the customer deposit subscription. Another important thing is an open-source technology in the DM field that is able to provide high quality models for real applications (such as the rapid miner software), which will help to reduce cost of DM projects. Note that our proposed method can enhance the predictive model performance while it used a smaller storage space, reduces the computation time and gains the higher predictive performance.

5. FUTURE WORK

In future work, it is good to collect more customer data, in order to check if high quality predictive models can be achieved without contact-based information. For the future work, it is best to apply the Data Mining models in a real environment and having a good interaction with marketing managers in order to gain a valuable opinion and feedback.

REFERENCES

[1] Kazi Imran Moin, Dr. Qazi Baseer Ahmed, “Use of Data Mining in Banking”, 2012.

[2] Philip Kotler, Kevin Lane Keller, Framework for Marketing Management, 5th edition Pearson, 2012.

[3] Fabrice Talla Nobibon, Roel Leus, Frits CR Spieksma, Optimization models for targeted offers in direct marketing: exact and heuristic algorithms, European Journal of Operational Research 210 (3) (2011) 670–683.

[4] Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, 9th edition Pearson, 2011.

[5] Pedro Domingos, A few useful things to know about machine learning, Communications of the ACM 55 (10) (2012) 78–87.

[6] Paulo Cortez, Mark J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, Information Sciences 225 (2013) 1–17.

[7] David L. Olson, Dursun Delen, Yanyan Meng, Comparative analysis of data mining methods for bankruptcy prediction, Decision Support Systems 52 (2) (2012) 464–473.

[8] Robert Phillips, Optimizing prices for consumer credit, Journal and Review Pricing Management 12 (2013) 360–377.

[9] David Martens, Foster Provost, Pseudo-social network targeting from consumer transaction data, NYU Working Papers Series, , CeDER-11-05, 2011.

[10] Sérgio Moro, Raul Laureano, Paulo Cortez, Enhancing bank direct marketing through data mining, Proceedings of the Forty- First International Conference of the European Marketing Academy, European Marketing Academy, 2012, pp. 1–8.

[11] Vivek Bhambri "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011

[12] M Suman, T Anuradha, K Manasa Veena, "Direct Marketing with the Application of Data Mining", 2011.

[13] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau
Dan Lee, "Decision Trees for Uncertain Data", 2011.

[14] Pedro Domingos, A few useful things to know about machine learning, Communications of the ACM 55 (10) (2012) 78–87.

[15] Dr. K. Chitra, B. Subashini, "Data Mining Techniques and Its
Application in Banking Sector", 2013.

[16] Safia Abbas, "Deposit subscribe Prediction using Data Mining
Techniques based Real Marketing Dataset", 2015.

[17] Karl D. Majeske, Thomas W. Lauer, "The bank loan approval decision from multiple perspectives", 2013.

[18] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., "CRISP-DM 1.0 - Step-by-step data mining guide, CRISP-DM Consortium", 2000

[19][Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data- Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001