

**A Seminar Report on**

**BANKING DATA PROCESSING**

**FOR FLEXIBLE TARGET PREDICTION**

In partial fulfillment of requirements for the degree of Bachelor of

Science In Mathematics and Computer Science

SUBMITTED BY:

Nguyen Thi Thu Duyen

Under the Guidance of

Dr. Tran Anh Tuan

**ACKNOWLEDGEMENT**

This project would not have been completed without the help of many people. Many thanks to my lecturer, Dr. Tran Anh Tuan for guiding, reading, and correcting the report for me many times. Thank you to the University of Science, HCMC for giving me the opportunity to work on this project. Thank you to my classmates who are always dedicated to helping me in difficult times. Thank my family for always supporting me during the past.

**INDEX PAGE**

Abstract 4

**List of Figures 5**

**List of Tables 6**

**CHAPTER 1 : INTRODUCTION 7**

1.1 Introduction 7

1.2 Motivation 7

1.3 Problem Description 7

1.4 Organization 7

**CHAPTER 2 : LITERATURE REVIEW 8**

2.1 Related Works 8

2.2 Challenges 8

2.3 Proposed Method 8

2.4 Contributions 9

**CHAPTER 3 : METHODOLOGY 10**

3.1 Collecting available data 10

3.2 Grouping available data 10

3.3 Identifying target feature 10

3.4 Analysing features’ relationship 10

3.5 Selecting features level 1 10

3.6 Selecting features level 2 11

3.7 Predicting 11

**CHAPTER 4 : EXPERIMENTAL RESULTS 12**

4.1 Database overview 12

4.2 Each step evaluation 12

4.3 Results 14

**CHAPTER 5 : DISCUSSION 16**

5.1 Benefits of Proposed Approach 16

5.2 Limitations of Proposed Approach 16

5.3 Future works 16

**CONCLUSIONS 17**

**REFERENCES 18**

**Abstract**

In business, the collected data sometimes does not meet the requirement, such as arising from the lack of a feature after collecting data or the need to add a new feature without re-collecting the data. More specifically, in banking, sometimes the lack of a feature can cause a significant reduction in profits, but re-collecting data is not feasible or too expensive. Therefore, we propose a method of feature proposition in order to predict any new feature based on the existing dataset. The available features will be divided into different subgroups, we will filter these features at two different levels to obtain features that are directly related to the feature to be predicted and make the prediction.

*Keywords - customer segmentation* *; data preprocessing ; data mining ; banking.*

**List of Figures**

[Figure 1. Overview of our method](#_heading=h.2et92p0) 9

[Figure 2. Distribution of nr.employed](#_heading=h.2et92p0) 13

[Figure 3. Correlation matrix](#_heading=h.2et92p0) 13

[Figure 4. Feature importance for job prediction](#_heading=h.2et92p0) 14

**List of Tables**

[Table 1. Features groups 12](#_heading=h.2et92p0)

[Table 2. The accuracies of our method using different prediction methods with feature selection levels](#_heading=h.2et92p0) 15

**CHAPTER 1: INTRODUCTION**

* 1. **Introduction**

Customer segmentation is the process of dividing customers into groups with similar characteristics. Customer segmentation helps companies identify the customer groups, their needs, and the profitability of each group. There are several popular customer segments such as demographic segmentation (considering characteristics such as age, gender, education, income, occupation, marital status ...), geographic segmentation (grouping customers by country, state, region ...), behavioral segmentation (segmentation based on customer's interact) and so on.

* 1. **Motivation**

In business, especially the banking sector, using the collected data to maximize benefits, which is always focused. However, the features, which are required for each period and each project, are different. The collected dataset will become obsolete and some features need to be updated. Can we use the available data to predict a new feature based on the customer segmentation approach?

* 1. **Problem Description**

In practice, the bank's customer data is incomplete, and sometimes we are unable to collect these missing data. In addition, essential information may be hidden behind available features in the data. Therefore, we attempt to predict a new feature from the available data of the bank. However, using all the features in the data for prediction does not yield the desired result. Therefore, we propose a method of selecting data for prediction.

* 1. **Organization**

The rest of the report is organized as follows,

• In Chapter 2, we present the related work, the challenges, our approach and the contributions.

• In Chapter 3, we describe in detail the process of our proposed method.

• In Chapter 4, we present the dataset introduction and the result of our method.

• In Chapter 5, we discuss the benefits and limitations of our method, and the future work.

**CHAPTER 2: LITERATURE REVIEW**

* 1. **Related Works**

Customer segmentation is an imperative requirement in many fields, especially commerce. Indeed, in commerce, dividing customers into groups with similar characteristics helps identify target customers and build more effective strategies, which can greatly increase profits. Therefore, customer segmentation techniques are a hot topic that interest data scientists to research and develop.

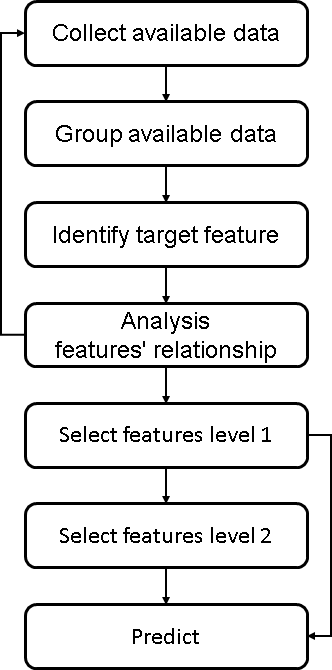
Customer segmentation is executed on the customer database. Studies have different points in the division and use of these data. For example, Colica [1] proposed four methods including Customer Profiling using Demographic and Purchase history to apply in case of small data; Customer Likeness Clustering using Demographic, Purchase history, and Data product to classify customers based on the target; RFM Cell Classification Grouping using Purchase history to map the three-dimension RFM efficiently; and Purchase Affinity Clustering using Purchase history and Data product to identify the most requested product. Baer [2] also proposed four methods including Business Rule using Demographic and Purchase history, which can use database query; Quantile membership using Purchase history to be able to be used with other data; Supervised Clustering with decision tree using Demographic and Purchase history to classify customers according to target; and Unsupervised Clustering using Purchase history that can use any number of customer attributes. However, the method widely applied by businesses is that Magento [3] uses Demographic, Purchase history, Data product, Data media, Data marketing, and Server log to have clear variable customer segmentation.

* 1. **Challenges**

In the banking sector, the lack of a few essential features in the data often occurs. The addition of missing features using prediction, which saves time and cost significantly. However, using all the available features for prediction does not get the expected results. Besides, some of the current data division methods are not really optimal because the target feature has relationships with features in different groups. Therefore, the grouping of data and the selection of features need to be further considered.

* 1. **Proposed Approach**

Some features may be predicted based on the features in the available data. We propose a banking data processing for flexible target prediction based on customer segmentation. The first step is to collect available data and divide it into appropriate groups. The next step is to identify the target feature to predict and analyze the relationship between the target feature and available features. The result of this analysis helps us decide whether to proceed with the next steps or not because sometimes features in the data do not contain the information of the target feature. Next, the features will be filtered at two different levels in order to find the features most necessary to make the prediction. Finally using some machine learning algorithms and the selected features as input for prediction.

****

**Fig 1.** Overview of our method

* 1. **Contributions**

The key contributions of this dissertation include:

* The introduction of the data division method.
* Proposing the data processing method to predict a new feature.
* Comparing the accuracies of the proposed method on machine learning algorithms.

**CHAPTER 3: METHODOLOGY**

* 1. **Collecting available data**

Our ultimate goal is to predict a new feature based on the available dataset. Thus, this dataset is the existing and feasible customer database, which means the information in the dataset is correct at present.

* 1. **Grouping available data**

We divide the data into four groups including:

* Profit potential: Variables that show the value of profit a customer can bring such as the number of transactions in a quarter, the average value of transactions...
* Past Purchases: Variables related to the last transaction such as last transaction value, last transaction date...
* Demographic: Variables related to personal information of customers such as age, education...
* Behavior: Variables related to customer feedback such as response to an advertisement, number of support contacts...
  1. **Identifying target feature**

In customer segmentation, we need to clearly define the business goals to help segmentation achieve efficiency. Likewise, it is important to identify an essential feature for a given project, which requires specialized knowledge and experience. We need to limit the predictable variables that are correlated because predicting correlated features wastes time and reduces the accuracy of the implementation. For example, we should not both predict the area code and the address.

* 1. **Analysing features’ relationship**

We perform data analysis to understand the data. Specifically, the inconsistent data are discarded. Then, we computed the available features' correlation for each other. Next, we analyze the correlation of the target feature with the existing features. In case the data contains prediction information, which means the available features are correlated with the target feature, we take the next steps. Conversely, if the data does not contain essential information, we will go back to the previous steps.

* 1. **Selecting features level 1**

Data processing aims to reduce data dimension and execution time, while increasing prediction accuracy. At this step, we attempt to minimize redundant features and select those that are highly correlated with the target feature. Specifically, we compute the feature importance for each group of data divided at the Grouping available data step. Based on the values of feature importance value and correlation, we choose the features that contain the most necessary information. Note that the selected features are distributed across all four groups of data.

Some methods of Linear regression, Logistic regression, CART, Random forest, Permutation, XGBoost, CatBoost ... can be used for computing the feature importance.

Selected features can be used as input for prediction in the case the number of features is small or for a compact application.

* 1. **Selecting features level 2**

In order to further reduce the data dimension and choose the optimal features. With the input is the selected features in the previous step and the number of features we want to keep, we iterate to randomly select features, predict and record the results. After a certain number of iterations, we come up with the required groups and the specific features that are needed for prediction.

* 1. **Predicting**

To predict the target feature, we use machine learning techniques such as Linear regression, Random forest, Support vector machine, XGBoost, CatBoost, Donut clustering[2], K-means clustering... depending on target feature type.

**CHAPTER 4: EXPERIMENTAL RESULTS**

* 1. **Database Overview**

To demonstrate the effectiveness of our method, we applied our approach to a dataset from the Portuguese retail bank [4]. This dataset collects customer information after a direct marketing campaign in Portugue from May 2008 to June 2013, including the period affected by the economic crisis in 2008. The dataset includes 41188 rows (each row describes a customer information) and 21 columns.

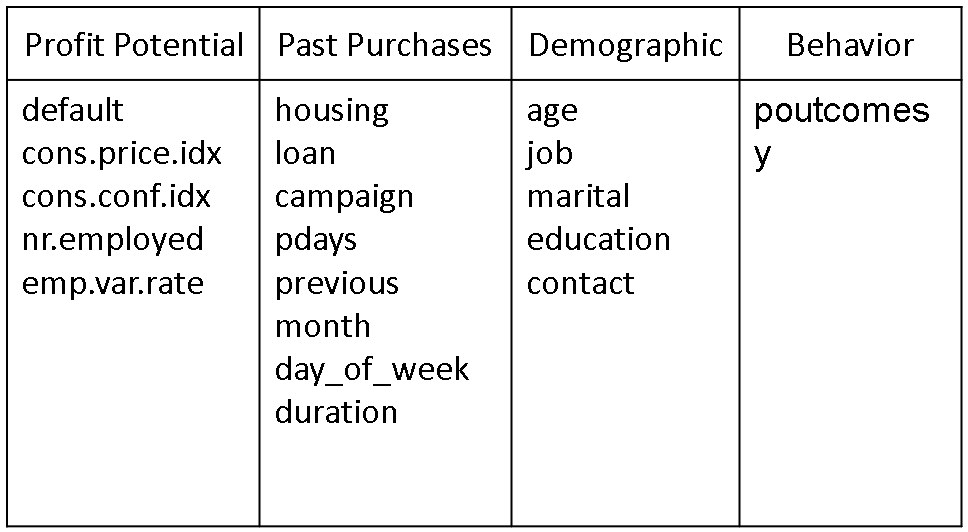
* 1. **Each Step Evaluation**

Step 1: Collect available data

We use bank marketing data to observe our method's effectiveness in this field.

Step 2: Group available data

We group the features as follows:

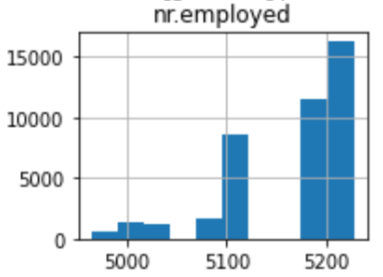
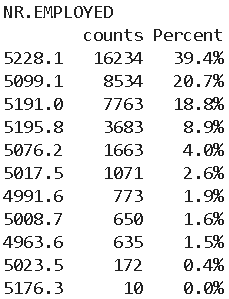
****

**Table 1.** Features Groups

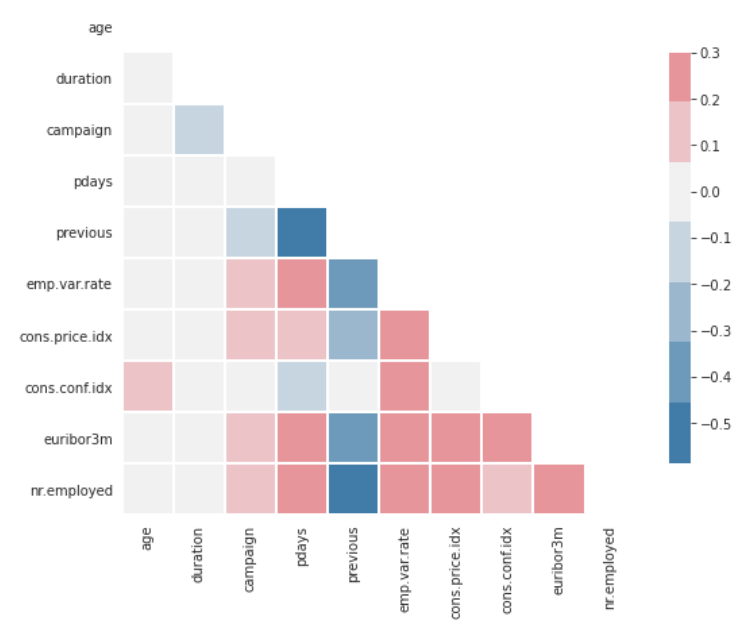
Step 3: Identify target feature

We choose a feature in this dataset to test prediction results.

Step 4: Analysis features’ relationship



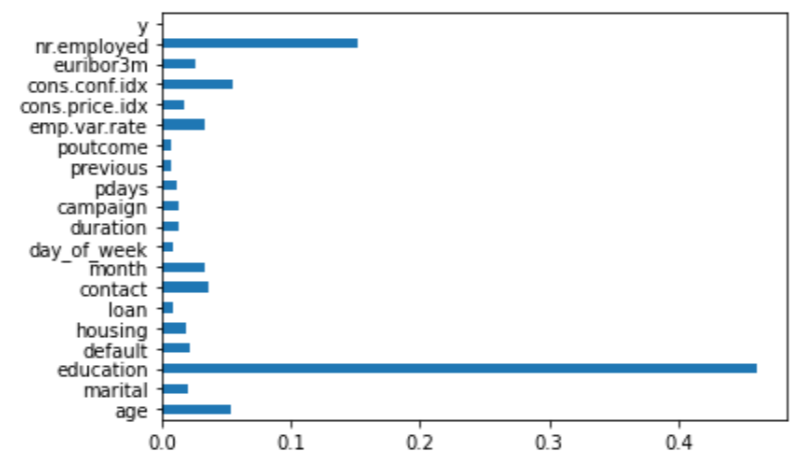
**Fig 2.** Distribution of nr.employed



**Fig 3.** Correlation matrix

Step 5: Select features level 1

We compute the feature importance as follows:



**Fig. 4.** Feature importance for job prediction

We retain 70% of the correlative features in each group.

Step 6: Select features level 2

We iterate to keep picking 70% of the features that are important for prediction. The list of selected features includes education, nr.employed, cons.conf.idx, emp.var.rate, month, contact, default, age.

Step 7: Predict

We use these selected features as input for prediction.

* 1. **Results**

We have found that using all the data for prediction has not been satisfactory. After the first feature selection step, the results are significantly improved. The prediction results in the second feature selection step are not improved, but we can further reduce the number of features in this step.



**Table 2.** The accuracies (%) of our method using different prediction methods with feature selection levels

**CHAPTER 5: DISCUSSION**

* 1. **Benefits of Proposed Approach**

Our methodology can reduce the number of features required for prediction without affecting the prediction results.In addition, our method separates the levels of feature selection, which is easy for compact application.

* 1. **Limitations** **of Proposed Approach**

The prediction results are not really as good as expected. In addition, we have not yet thoroughly dealt with the use of different types of data in a dataset because encrypting the data would otherwise cause loss of information.

* 1. **Future Works**

We are going to continue to develop this method to get more excellent results. Moreover, we are going to find solutions to be able to use different types of data more effectively.**CONCLUSIONS**

In this paper we propose a data processing method to reduce the number of data dimensions by selecting features that contain the most information to predict a new feature. This approach promises to achieve good predictive results while using a small number of features instead of the whole.

**REFERENCES**

*Basic format for books:*

1. Colica R. Customer Segmentation And Clustering Using SAS Enterprise Miner Part I The Basics. 2011:1-14.
2. Baer D. CSI : Customer Segmentation Intelligence for Increasing Profits. SAS Glob Forum.2012:1-13. http://support.sas.com/resources/papers/proceedings12/103-2012.pdf.
3. Magento. An Introduction to Customer Segmentation. 2014. An\_Introduction\_to\_Customer\_Segmentation
4. [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