Predict Success of Bank Derict Marketing By ISAS

ID:773950464

NAME: Yong Wan

UPI: wany991

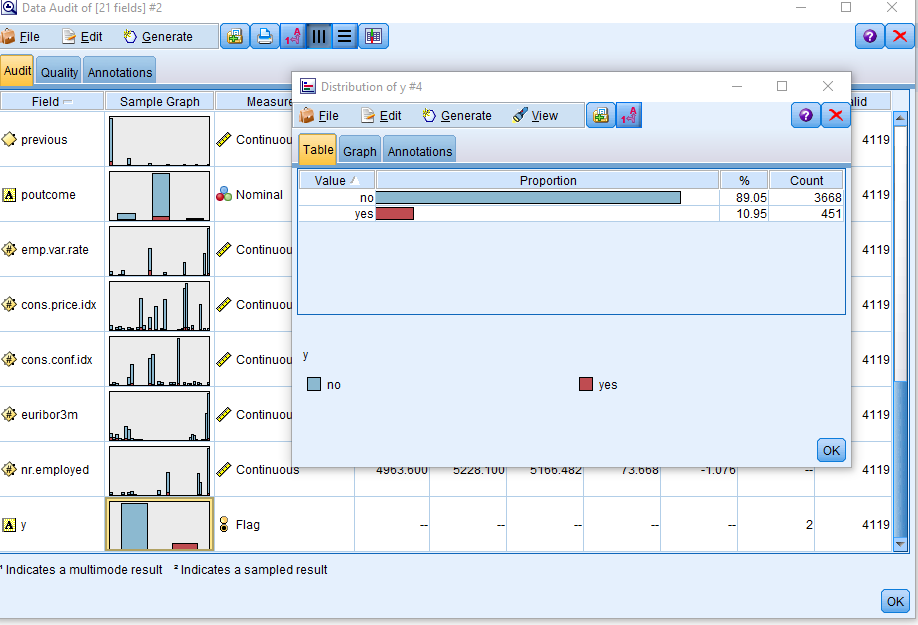
**1. Business and/or Situation understanding.** “First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer’s viewpoint.” (Fayyad et al., 1996)

* 1. **Identify the objectives of the business and/or situation**

**Marketing sales activities are a typical promotion strategy business. For bank,it was not an exception.Banks use direct marketing when targeting segments Customers connect with thme to achieve specific goals.But how to improve the success of these kinds of marketing activities is a big problem for bank managers. This paper aims to improve it by analysis the data a bank collected from** May 2008 to November 2010 through methods data mining and KDD.

* 1. **Assess the situation**

**Base on the data, there were only 10% of success in the past three years, it means there are a huge gap between success and unsuccess. Also we should do some thing to improve it to decrease the phone call times due to our customers may be annoyed and increase the profit for banks due to the phone agents get more success by same phone outs as before.**



* 1. **Determine data mining objectives, and**

**This objective is to predict the success of phoning a customer then dediced whether he was worth to call.**

**Task list**

1.Describe the methods for model assessment (for example, accuracy, performance, etc.).

2.Define benchmarks for evaluating success. Provide specific numbers.

3.Define subjective measurements as best you can and determine the arbiter of success.

4.Consider whether the successful deployment of model results is part of data mining success. Start planning now for deployment.

**1.4 Produce a project plan**

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Time | Resources | Risk |
| Business understanding | 0.5 week | All analysts | Economic Crisis |
| Data understanding | 2 weeks | All analysts | Economic Crisis |
| Data preparation | 4 weeks | All analysts | Economic Crisis |
| Modeling | 3.5 weeks | All analysts | Economic Crisis |
| Evaluation | 1 week | All analysts | Economic Crisis |
| Deployment | 1 week | All analysts | Economic Crisis |

**2. Data understanding.** Data provides the “raw materials” of data mining. This phase addresses the need to understand what your data resources are and the characteristics of those resources. “Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.” (Fayyad et al., 1996)

**2.1 Collect initial data**

**Link** [**https://archive.ics.uci.edu/ml/datasets/bank+marketing**](https://archive.ics.uci.edu/ml/datasets/bank+marketing)

**Citation Request:**

This dataset is public available for research. The details are described in [Moro et al., 2014].   
Please include this citation if you plan to use this database:   
  
[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, Elsevier, 62:22-31, June 2014

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Title:**Bank Marketing Data Set** | | **Source:**  [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, Elsevier, 62:22-31, June 2014  **Relevant Papers:**  S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014  S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]  In the above references,two datasets were created.  1)bank-additional-full.csv with 41188 examples and 20 inputs,recorded between May 2008 and November 2010.  2)bank-full.csv has all examples with 17 inputs,is a older version dataset and with less inputs than 1).  **2.2 Describe the data**  1.Amout of data  45211 rows in bank\_full.csv and 41188 rows in bank\_additional\_full.csv  2.Value types    3.Coding schemes  1 - age (numeric in bank-additional-full.csv and string in bank-full.csv) 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') 3 - marital : marital status (categorical ) 4 - education (categorical ) 5 - default: has credit in default? (categorical) 6 - housing: has housing loan? (categorical 7 - loan: has personal loan? (categorical) # related with the last contact of the current campaign: 8 - contact: contact communication type (categorical)  9 - month: last contact month of year (categorical) 10 - day\_of\_week: last contact day of the week (categorical) 11 - duration: last contact duration, in seconds (numeric12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) 14 - previous: number of contacts performed before this campaign and for this client (numeric) 15 - poutcome: outcome of the previous marketing campaign (categorical) # social and economic context attributes 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric) 17 - cons.price.idx: consumer price index - monthly indicator (numeric)  18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)  19 - euribor3m: euribor 3 month rate - daily indicator (numeric) 20 - nr.employed: number of employees - quarterly indicator (numeric)  4.Output variable (based on sensory data):  y - has the client subscribed a term deposit? (binary: 'yes','no')  **2.4 Verify the data quality**  Missing Attribute Values: bank-full.csv do not have emp.var.rate , cons.price.idx, cons.conf.idx , euribor3m, nr.employed. So after appending,these five fields should be null.  Five fields have data labelled unknown, for example: job ,marital,default,housing，loan:  **3.Data preparation**  “Third is data cleaning and pre-processing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes” (Fayyad et al., 1996)  3.1 Select the data  Due we have more 80,000 rows data,so we just simple to discard the data who has unknown labels.    3.2 Clean the data  Because pdays=0 in bank-full.csv means the customer was not contacted before,but in bank-additional-full.csv,pdays=999 has same means,so we decide change to same number 999 as the customer was not contacted before.      3.3 Construct the data  As we known age has different data types between bank-full.csv and bank-additional-full.csv. it will not be able to append together. So we create a new field age\_i which it equals to\_integer(age)    3.4 Integrate various data sources  From here,We know the education is different between the two datasets, we should reclassify this variable after appending to education\_new.so we decided group “primary”, “basic.4y”,”basic.6y”,”illiterate” as primary, group “secondary”,”basic.9y”,”high.school” as “secondary”,then combine “tertiary”,” professional.course”,” university.degree” to “tertiary”    **4.Data transformation (5%).** “Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.” (Fayyad et al., 1996)  4.1 Reduce the data  When we generate a distribution graph of y, we saw that the data is  skewed (12% for Y and 88% for No). If a model did no thing,just sample say it is no for every record,it will be got 88% accuracy,I think it is useless for prediction. To fix this, we use”Balance Node (Reduce)” to make the data approximately 50/50.    As there are too many null data on fields( day\_of\_week,emp.var.rate,cons.price.idx,cons.conf.idx,euribor3m,nr.employed),we discard these six fields。    4.2 Project the data  Through “Feature Selection Model”, we decided to discard default field as it shows single category too large  For duration, 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. Also, after the end of the call y is obviously known. Thus, this input should be discarded as we want to have a realistic predictive model.  We use ”Partition” to random split the data to 80% for(training) ,10% to testing and 10% to validation.  After that, we used “Auto Data Prep” to convert all of categorical or nominal field to continuous data. It will be benefit on model building.   1. **Data-mining method(s) selection**   “Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on, are described later as well as in Fayyad, Piatetsky-Shapiro, and Smyth (1996).” (Fayyad et al., 1996)   * 1. Match and discuss the objectives of data mining (1.1) to data mining methods         5.2 Select the appropriate data-mining method(s) based on discussion | |
|  |