# Predict Success of Bank Derict Marketing By ISAS

ID:773950464

NAME: Yong Wan

UPI: wany991

## 1.Business and Situation understanding.

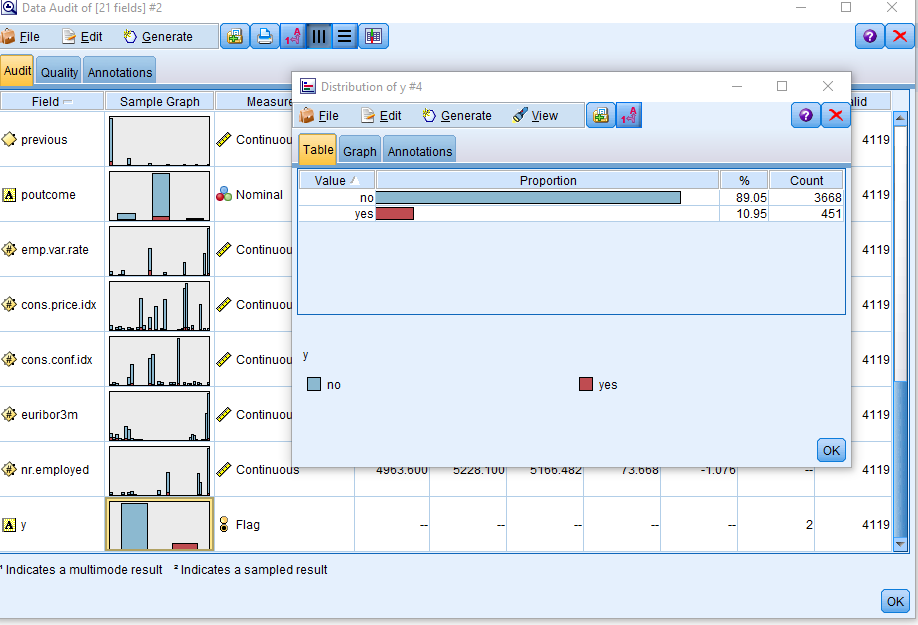
“First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer’s viewpoint.” (Fayyad et al., 1996)

### 1.1 Identify the objectives of the business and/or situation

Marketing sales activities are a typical promotion strategy business. For bank,it is not an exception.Banks use direct marketing when targeting segments Customers connect with them 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.

### 1.2 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.So our aim is to decrease 40% of number of phoning out and at max decease 5% of number of successful call. This means we will get at least 20% of success of phoning ,double than before.Also means we should get at least 90% correct prediction on customer who’s original answer is Yes and nearly 60% correct prediction on customer who’s original answer is No.



### 1.3 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>

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 Due to there is not primary key for records, So we decide to add a RecordID by function @Index.  It will be convenient to do merging operation.  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 “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.   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)  the data mining methods we chose included  Python Model Group:Rondam Forest,Rondam Trees, Bayesian Network;  Decision Tree:C5.0 and C&R Tree  lassification and regression technique:Support Vector Machine (SVM),KNN  other mothods: XGBoost Tree,  For the whole result,scores list in the Table 1-Seven modelling result comparing above:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Order | Modeller Name | Total Training | Correct | Wrong | The Rate of Correct | | 1 | KNN | 14293 | 10295 | 3998 | 72.03% | | 2 | Random Tree | 14372 | 10235 | 4137 | 71.21% | | 3 | C5.0 | 14235 | 10122 | 113 | 71.11% | | 4 | Bayes Net | 14154 | 10003 | 4151 | 70.67% | | 5 | Random Forest | 14418 | 10187 | 4231 | 70.65% | | 6 | SVM | 14208 | 10006 | 4202 | 70.43% | | 7 | CRT | 14238 | 9747 | 4491 | 68.46% | | 8 | Neural Net | 14302 | 7516 | 6786 | 52.55% |   ( Table 1-Seven modelling result comparing)  Select the appropriate data-mining method(s) based on discussion  Basing on Table 1-Seven modelling result comparing,only Neural Net got a very low result,it totally did not match this task, so we decided choosing other seven methods including KNN,and Random Tree,C5.0,Bayes Net,Random Forest and SVM , CRT to do the next step research. 6.Data-mining algorithm(s) selection “Sixth is exploratory analysis and model and hypothesis selection: choosing the datamining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).” (Fayyad et al., 1996)  For this task,as we know,there are only 10% of customers will be attracted  by marketing activities. So comparing to the correct rate of the whole data, the correct rate of customers who labelled success is more important. So we decide to look the differents between these parts each modeling got.  We create a new field result\_label,combined from the original exception field “Y” and the calculating exception field , which has labelled “No-No”,”No-Yes”,”Yes-Yes”,”Yes-No”,  “No-NO” means original is “NO”,and the calculation result also is “NO”  “Yes-Yes” means original is “Yes”,and the calculation result also is “Yes”  “No-Yes” means original is “NO”,and the calculation result is “Yes”  “Yes-No” means original is “Yes”,and the calculation result is “No”      From Table-2 Yes-No comparing,KNN ,Random Forest and SVM got the best three score on “YES”,they only miss 14.38%,15.99% and 17.29% respectively. So we decided these three methods will be able to enter the next step.  Table-2 Yes-No comparing   |  |  |  |  | | --- | --- | --- | --- | | Order | Modeller Name | Number | The Rate of incorrect | | 1 | KNN | 2541 | 14.38% | | 2 | Random Forest | 2849 | 15.99% | | 3 | SVM | 3085 | 17.29% | | 4 | C5.0 | 3246 | 18.25% | | 5 | Bayes Net | 3597 | 20.23% | | 6 | Random Tree | 3958 | 22.36% | | 7 | CRT | 4439 | 25.02% |     From Table-3 No-Yes comparing,CRT ,Random Forest and Bayes Net got the best three score on original “NO”,they only miss 5.56%,7.14% and 8.94% respectively.It means If we add these prediction data to the waiting phoning pool,it will not increase the number of failures.So we decided these three methods also will be able to enter the next step.  Table-3 No-Yes Comparing   |  |  |  |  | | --- | --- | --- | --- | | Order | Modeller Name | Number | The Rate of incorrect | | 1 | CRT | 986 | 5.56% | | 2 | Random Forest | 1263 | 7.14% | | 3 | Bayes Net | 1590 | 8.94% | | 4 | C5.0 | 1932 | 10.86% | | 5 | SVM | 2205 | 12.36% | | 6 | Random Tree | 2347 | 13.17 | | 7 | KNN | 2437 | 13.69% |   So Totally,Five Modeling are still in the pool we want to choose.they are KNN,Random Forest,SVM,CRT and Bayes Net.But there is also no any modelling which match the original target in the beginning of this report. Due to the accuracy on Yes is in 70%-80%,lower thant our target, and the accuracy on No is also between 70-80%,higher than our target. So we could do something to improve the accuracy on Yes by decreasing the accuracy on No.  It is our final decision to combine the five models result, if any of these five models predict it is Yes, we predict it is “Yes”, in other words, if all of these models predict a customer will answer NO, we set it is No, otherwise it will be Yes.  Let us have a look at the final data.   7.Data Mining “Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.” (Fayyad et al., 1996) This is, of course, the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data.   8.Interpretation “Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.” (Fayyad et al., 1996) We assess and evaluate the models and the results and their reliability. “You are ready to evaluate how the data mining results can help you to achieve your objectives.” (SPSS, 2007)  8.1 Study and discuss the mined patterns  8.2 Visualize the data, results, models, and patterns  8.3 Interpret the results, models, and patterns  8.4 Assess and evaluate results, models, and patterns  8.5 Iterate prior steps (1 – 7) as required | |
|  |