Explorative Data Analysis Task – Predicting the Success of Term Deposit Marketing Activities

1. Preliminaries and Rules

Today's task is given to in written form. You should have received this text document as well as a file containing the data for the task. The data is a publically available data set, it is not generated by us in an artificial way. You are allowed to use any tools you are used to. However, please include any kind of source code into your final report.

2. Requirement and Expectations

Please analyse the data set described above and try to come up with a report that provides details on your findings. We are interested in customer groups that are likely to respond to marketing related selling activities. As these activities are expensive to carry out, we are looking for recommendations on customer segmentation or categories on which customers should participate in a campaign and which not in order to maximize a campaign's revenue.

Please describe how you approached this task, which analysis steps you do and how you evaluate your findings. Your report should allow an informed third person to understand your line of reasoning and approach. Please include plots, tables etc. in the report. It is however not required to care about the layout. We are not taking into consideration the presentational style of your report. We also do not expect the specification of a predictive model; please focus on the analysis of the data. You should however include some recommendations on what kind of predictive model we should build and why. Your line of reasoning is important to us.

Please try to set up a reasonable scope for the limited amount of time available. Please focus on traceability rather then maximizing insights. We value less, yet well grounded insights a lot! It is definitely impossible to do a full-blown analysis in the time frame available!

3. Introduction to the Problem

Marketing campaigns constitute a typical strategy to generate business revenue. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific business goal. Very often, call-centres are operated by businesses that allow communicating with customers via telephone – marketing activities carried out through a call-centre are often called telemarketing activities.

Information technology enables new marketing approaches by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics, thus allowing companies to build better and high value customer relations. However, the task of selecting the best set of customers for a telemarketing campaign, i.e., that are more likely to subscribe to a product, is considered very hard and is often done using statistical estimation and prediction methods which are used as input to optimization models for the selection of sets of customers that will receive an offer for one or more products during a campaign.

In this approach, data mining supports decision-making using business data, allowing for semi-automatic extraction of explanatory and predictive knowledge from raw data. In

particular, classification is the most common data-mining task. The goal of those data mining activities is to build a data driven model that learns an unknown underlying function that maps several input variables, which characterize an item, with one labelled output target.

4. Data Set Description

The enclosed data file contains real data from a retail bank, collected from May 2008 to June 2013. The marketing campaigns were based on phone calls. Often, more than one contact to the same customer was required in order to assess if the targeted product – a term deposit – is subscribed to by a customer, which is indicated by 'yes' or 'no' to subscribe to a deposit. The classification goal is to predict whether the customer will or will not subscribe to a term deposit.

A term deposit is held at a financial institution that has a fixed term. These are generally short-term with maturities ranging anywhere from a month to a few years. When a term deposit is purchased, the lender (the customer) understands that the money can only be withdrawn after the term has ended or by giving a predetermined number of days notice). Term deposits are an extremely safe investment and are therefore very appealing to conservative, low-risk lenders. By having the money tied up for a longer period of time, lenders generally get a higher rate with a term deposit compared with a demand deposit.

The input variables of the customer data are:

Customer Attributes

1. Age (Numeric)
2. Job, which is the type of the customer's job (Categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. Marital: The customer's marital status (Categorical: 'divorced','married','single','unknown';)
   Please note: 'divorced' means divorced or widowed
4. Education (Categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.cours', 'university.degree','unknown')
5. Default: Has the customer a credit in default? (Categorical: 'no','yes','unknown')
6. Housing: Has the customer a housing loan? (Categorical: 'no','yes','unknown')
7. Loan: Has the customer a personal loan? (Categorical: 'no','yes','unknown')

Attributes of the Last Contact of the Current Campaign

8. Contact: Contact communication type (Categorical: 'cellular','telephone')
9. Month: Last contact month of year (Categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. Day_Of_Week: Last contact day of the week (Categorical: 'mon','tue','wed','thu','fri')
11. Duration: Last contact duration, in seconds (Numeric).
    Important note: This attribute highly affects the output target (e.g., if duration=0 then subscription='no'). Yet, the duration is not known before a call is performed! Also, after the end of the call the customer's subscription decision is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other Attributes

12. Campaign: Number of contacts performed during this campaign and for this customer (Numeric, includes last contact)
13. Passed_Days: Number of days that passed by after the customer was last contacted from a previous campaign (Numeric; 999 means customer was not previously contacted)
14. Previous: Number of contacts performed before this campaign and for this customer (Numeric)
15. Previous_Outcome: Outcome of the previous marketing campaign (Categorical: 'failure','nonexistent','success')

Social and Economic Context Attributes

16. Emp_Var_Rate: Employment variation rate - quarterly indicator (Numeric)
17. Cons_Price_Index: Consumer price index - monthly indicator (Numeric)
18. Cons_Conf_Index: Consumer confidence index - monthly indicator (Numeric)
19. Euribor3m: Euribor 3 month rate - daily indicator (Numeric)
20. Nr_Employed: Number of employees - quarterly indicator (Numeric)
21. Subscription - Output variable, the desired target - has the customer subscribed to a term deposit? (Binary: 'yes','no')