Ray Anthony Roderos

C004089342

APAN K5200

Data Mining

Do distribution – univariate analysis (frequency distribution)

Data cleaning

-missing data

-test for data that doesn’t make sense

-test for correlation and if there’s a need to correct if highly correlated

Assessing for Assumptions

- test for nonconstant variance

- test for nonnormality

- test for dependence

Tabulate (slicing and dicing data)

Scatterplot Matrix (correlation between pairs of co-variate)

Graph builder (visual exploration)

Sorting

**Proposal 1: Telemarketing of Special Bank Rate Deposits**

**Introduction**

The study aims to find the socio-economic, operational and marketing determinants of a client availing of a financial product in the form of a special bank term deposit through telemarketing. It will seek to establish the relationship of various variables both from the client and the bank based on the data obtained from the research of Moro and Cortez on using Data for Bank Direct Marketing [Moro et al., 2014].

The original researchers used a technique called the Cross-Industry Standard Process for Data Mining (CRISP-DM) in R which is a popular methodology for increasing the success of data mining. Their main objective was to find the best model out of the 3 that was tested to determine successful calls. Their methodology entailed using Naive Bayes, Decision Trees and Support Vector Machines using R.

This paper will use a broader and more abstract set of features in SAS JMP and will mainly focus on using logistic regression as its empirical model to analyze the relationship of the variables defined by Moro and Cortez’s study. The conclusion and recommendation may benefit not only bankers on which clients to target but also call centers on how to operate efficiently while maximizing caller impact.

**Background**

Business process outsourcing (BPO) involves the contracting out of certain business operations to third party providers but for the past many years, the BPO industry has been booming in the Philippines, generating over USD 25 billion in revenue accounting for about 10% of the Philippine economy (Lee 2015). The industry is contributing to the country’s annual growth rate of 8-10%. This is mainly in customer and technical support in the form of voice and non-voice communication but also increasingly, in more sophisticated work such as accounting, legal and medical secretariat services. Manila has now been called the call center capital of the world, overtaking India’s New Delhi with its cheap labor, customer-centric and friendly attitude and Americanized English (Lee 2015). With so many call center agents in the Philippines, telemarketing is an area that the researcher wishes to explore more. The research will analyze the data and see how clients are listed and agent’s key performance indicators are set. In the professional world, data scientists and analysts who have processed these types of data and gave their recommendations have impacted business policies that are felt by the Filipinos who are working in the BPO companies.

**Problem Statement**

1. Client Side: What are the Socio-economic characteristics of the potential client that will entail him/her to avail of the special bank term deposit if telemarketed?
2. Agency Side: What operational and marketing tactics should the bank use to increase their chances of successfully gaining a client through telemarketing?

**Dataset**

The Portuguese bank’s direct marketing dataset obtained is multivariate with real numbers. There are 45,200 instances with 17 variables dated 2012. The variables are categorized into 4: bank client data, data on the last contact with the client of the current campaign, social and economic context such as employment rate and the consumer price index and lastly, other attributes such as total number of contacts for this campaign and the outcomes of previous marketing campaigns with the dependent variable being a “yes” or a “no” from the client. The paper hypothesizes that bank client data and call information from previous contacts on the current and past campaigns have a significant effect on the client for him/her to accept the offer.

**Summary Statistics**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Distributions**  **age**    **Summary Statistics**   |  |  | | --- | --- | | Mean | 40.93621 | | Std Dev | 10.618762 | | Std Err Mean | 0.0499404 | | Upper 95% Mean | 41.034094 | | Lower 95% Mean | 40.838326 | | N | 45211 | | **job**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | admin. | 5171 | 0.11437 | | blue-collar | 9732 | 0.21526 | | entrepreneur | 1487 | 0.03289 | | housemaid | 1240 | 0.02743 | | management | 9458 | 0.20920 | | retired | 2264 | 0.05008 | | self-employed | 1579 | 0.03493 | | services | 4154 | 0.09188 | | student | 938 | 0.02075 | | technician | 7597 | 0.16803 | | unemployed | 1303 | 0.02882 | | unknown | 288 | 0.00637 | | Total | 45211 | 1.00000 | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **education**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | primary | 6851 | 0.15153 | | secondary | 23202 | 0.51319 | | tertiary | 13301 | 0.29420 | | unknown | 1857 | 0.04107 | | Total | 45211 | 1.00000 |   N Missing 0 | **contact**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | cellular | 29285 | 0.64774 | | telephone | 2906 | 0.06428 | | unknown | 13020 | 0.28798 | | Total | 45211 | 1.00000 | | N Missing 0 |  |  | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Y – decision of potential client**    **Frequencies**   | **Level** | **Count** | **Prob** | | --- | --- | --- | | no | 39922 | 0.88302 | | yes | 5289 | 0.11698 | | Total | 45211 | 1.00000 |   N Missing 0  2 Levels |  |

**Sources**

Lee, Don, The Philippines has become the call-center capital of the world, LA Times, http://www.latimes.com/business/la-fi-philippines-economy-20150202-story.html

S. Moro, P. Cortez and P. Rita. A Data­Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22­31, June 2014

UC Irvine Machine Learning Repository**,** https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

archive.ics.uci.edu/ml/datasets/Online+News+Popularity

Introduction  
The study aims to determine the operational circumstances and key content that is written that affects  
the popularity of articles on news sites. It will seek to establish the relationship of the features of the  
articles and its shares in social networks based from the data obtained from Mashable.com and from  
the research of Fernandes, Vinagre and Cortez (August 2015).  
This paper will use a either a logistic regression or a random forest analysis to seek the link of the  
variables and the impact or effect it makes on the article’s potential popularity. The conclusion and  
recommendation may benefit journalists and content creators to determine the popularity of its article  
before it will be published online.  
Background  
With the web continuing to expand to the realm of news, many news outfits are moving online to post  
their news. It has simply become cheaper and more convenient for users to just check the news on a  
topic they are interested in on their computers or on their phones rather than pay for a newspaper in  
which they won’t read many of the sections there. And if the most important metric for print news  
outlets is papers sold, the most important metric for online news websites is the interactions of their  
readers on the articles they post, most especially sharing the article on their social media. In the  
Philippines, only recently have long established newspaper starting to take seriously their online  
presence when venture-funded start ups are already sustainable solely on the traffic of their websites,  
analysis such as this will be very helpful to accelerate their online brand equity.  
Problem Statement  
1. What are the characteristics of the online news article that will increase the number of shares?  
2. When should be the right time to post a news article given the time of date and current  
sentiment of the readers to increase the number of shares?  
Dataset  
The Mashable.com news data obtained is multivariate with the variables either in integer or real  
numbers dated 2015. It has 39,700 instances spread across 61 variables. 25 of the variables are  
characteristics of the articles such as titles, number of words in title, number of words in the content.,  
number of images, number of videos, type of article and day it was published. The rest are shares of the  
article, keywords that were shared, number of shares of reference articles, rating, sentiment analysis  
and polarity of content with the dependent variable being the number of shares of the article. The paper  
hypothesizes that number of words, number of links, images, videos and date of posting will determine  
the number of shares.