Report I

Anna Szymanek (230042), Patryk Wielopolski (234891)

1 Introduction

Cluster analysis with quality assessment Application of the selected dimension reduction method in connection with classification and cluster analysis

2 Methods

2.1 Data description

Telemarketing is one of the forms used to encourage clients to buy new bank's product. If we imagine real-world scenario it may be very hard to decide to which customer we should call in order to achive our goal (in this case bank term deposit subscription) because it's not possible to call them all as we have limited human resources of telemarketers. Such in this case we could use historical data about calls done in previous marketing campaigns to formulate conclusions about what type of clients are our target group and what part of the day / week / year is a good time for such a projects. Moreover after that we can create classification models which will learn to give us good recomendations which clients we should call in the first order.

In general our dataset can be splitted in five categories: bank client data, last contact of the current campaign, other attributes, social and economic context, outcome as we can observer in Table 1. First category describes general information about client - age, job, etc. Second category describes how the last contact with client was performed. There is also important note that Duration 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 only be included for benchmark purposes and we will discard it because our intention is to have a realistic predictive model. Third category in general tell us about previous campaigns and previous contacts with given person. There is also a note from dataset authors that Pdays equal 999 means client was not previously contacted. We will also note that in our data. Fourth category is about social and economic context attributes such as employment rate. This might give us information how was economy in this time and might be driving factor for some people. The last category is our outcome variable, i.e. flag if the client subsribed a term deposit.

2.2 Data analysis

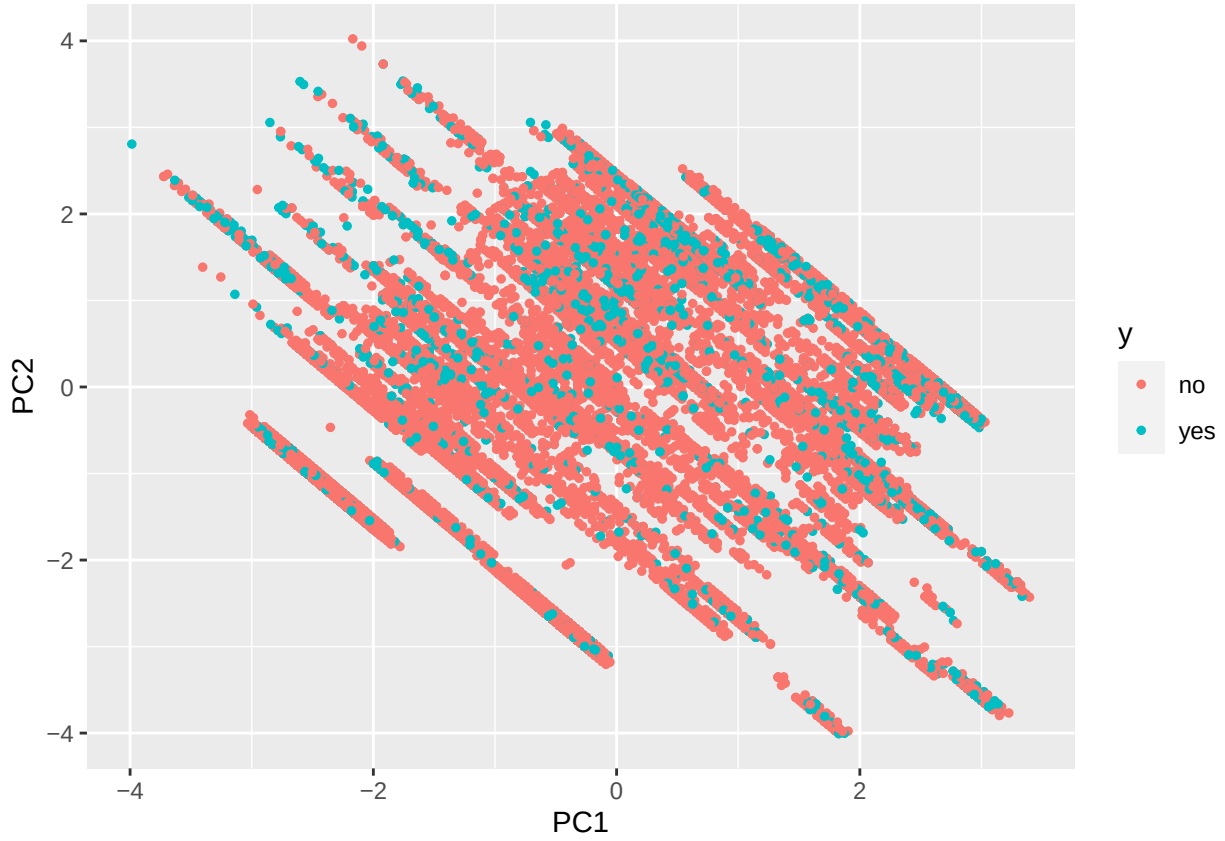
2.3 Dimension reduction

Objective: feature extraction / visualization Methods: - PCA - MDS - ICA - NMF

Variable name	Type	Description
Bank client data		
Age	Numeric	Age of client
Job	Categorical	Type of job
Marital	Categorical	Marital status of client
Education	Categorical	Education status of client
Default	Categorical	Has credit in default?
Housing	Categorical	Has housing loan?
Loan	Categorical	Has personal loan?
Variables related with the last contact of the current campaign		
Contact	Categorical	Contact communication type
Month	Categorical	Last contact month of year
Day of week	Categorical	Last contact day of the week
Duration	Numeric	Last contact duration, in seconds
Other attributes		
Campaign	Numeric	Number of contacts performed during this campaign
Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign
Previous	Numeric	Number of contacts performed before this campaign
Poutcome	Categorical	Outcome of the previous marketing campaign
Social and economic context attributes		
Emp.var.rate	Numeric	Employment variation rate - quarterly indicator
Cons.price.idx	Numeric	Consumer price index - monthly indicator
Cons.conf.idx	Numeric	Consumer confidence index - monthly indicator
Euribor3m	Numeric	Euribor 3 month rate - daily indicator
Nr. employed	Numeric	Number of employees - quarterly indicator
Outcome variable		
y	Categorical	Has the client subscribed a term deposit?

Table 1: Input variables.

2.3.1 Principal components analysis



2.3.2 Multidimensional scalling

Not possible to calculate . . .

2.3.3 FastICA

Same as PCA

2.3.4 Non-negative matrix factorization

Explore more...

2.4 Classification

Similar to the 1. project

2.5 Clustering

Objective: group objects according to their similarity Methods: - k-means - PAM - AGNES - Other (HDBSCAN?) Quality assessment of cluster analysis results: - Average silhouette vs different K - Separation / Compactness / Other ideas

3 Results and discussion