Report I

Anna Szymanek (230042), Patryk Wielopolski (234891)

1 Introduction

In this report we will continue the analysis of the dataset [1] connected to direct marketing campaigns of a Portuguese banking institution. This time we will explore dimension reduction techniques and cluster analysis.

We will begin our analytical journey with dimension reduction techniques during which we would like to better understand our dataset via visualization and extract useful features for classification and clustering. During exploration of the second task we would like to reveal the hidden structure of the data. Moreover we would like to find out how it is related to the response variable y - information whether bank's product (term deposit) would be subscribed or not by given client. Furthermore we would like to find dependencies between clients in our dataset and identify some specific group of clients. Finally, we would like to utilize results of the dimension reduction techniques in the classification task and compare our results with previously obtained in the first report.

2 Methods

In the following sections we will go through all mentioned in introduction tasks - dimension reduction, classification and cluster analysis. We will use all the transformations used in the first part of the project in context of the classification, i.e. we will focus only on the new clients who has never been targeted in previous campaigns and additionaly we performed data transformations connected to missing data, rare values and categorical variables encoding.

2.1 Dimension reduction

In this section we will go through a few dimension reduction techniques in order to visualize our dataset from different perspectives and look for interesting patterns which we hope to utilize in next section connected to classification and cluster analysis. We will use and compare following methods:

- Principal component analysis (PCA),
- Multidimensional scaling (MDS),
- Independent component analysis (ICA),
- Non-negative matrix factorization (NMF).

2.1.1 Principal components analysis

We will begin with principal components analysis. Firstly we will try naive approach and take whole dataset, conduct one hot encoding for categorical variables, remove target variable and analyse the obtained results in context of target variable. Moreover we will only center our data without scaling and treat it as a experiment

how does scaling influence the results. We already known from the first report that socio-economic variables has huge values compared to the other variables and we expect that results may be higlhy influenced by these variables.

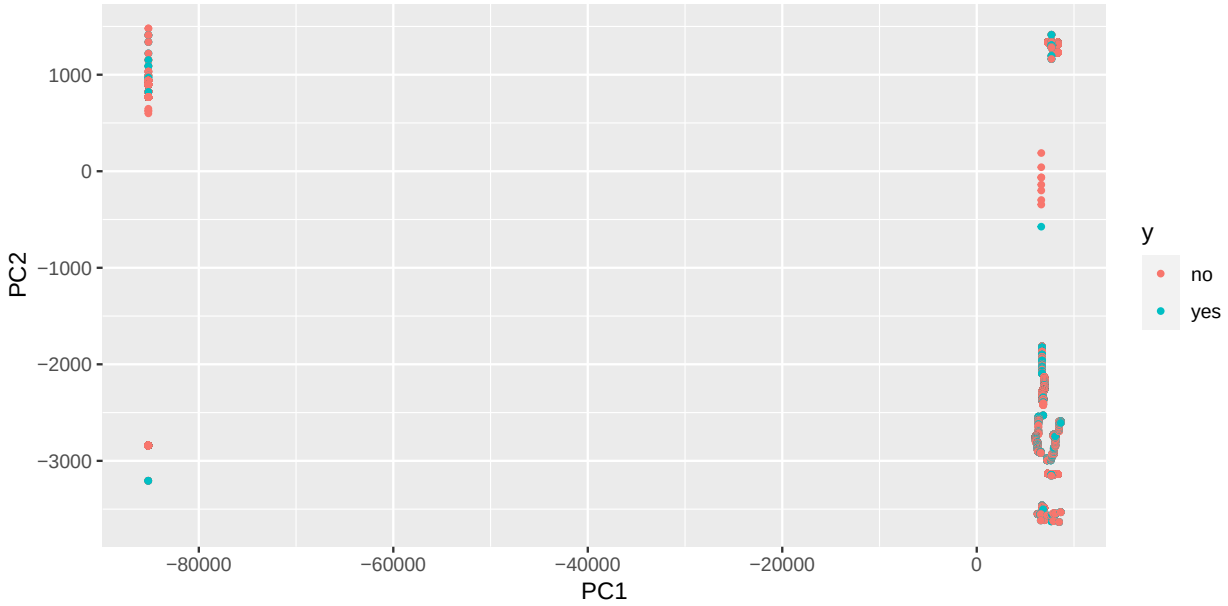


Figure 1: Results of the Principal component analysis without scaling in the context of y variable.

The results of the PCA without scaling we can observe on Figure 1. As we expected the principal components have big values which are probably influenced by socio-economic variables. Let's explore the formulated cluster with *PC1 < -80000* and *PC2 > 0* values and find out what data is in this subspace.

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m
1	-1.00	932.00	-42.00	4733.00
2	-1.00	932.00	-42.00	4733.00
3	-1.00	932.00	-42.00	4663.00
4	-1.00	932.00	-42.00	4663.00
5	-1.00	932.00	-42.00	4663.00
6	-1.00	932.00	-42.00	4663.00
7	-1.00	932.00	-42.00	4663.00
8	-1.00	932.00	-42.00	4663.00
9	-1.00	932.00	-42.00	4663.00
10	-1.00	932.00	-42.00	4663.00
11	-1.00	932.00	-42.00	4592.00
12	-1.00	932.00	-42.00	4592.00
13	-1.00	932.00	-42.00	4592.00
14	-1.00	932.00	-42.00	4592.00
15	-1.00	932.00	-42.00	4474.00
16	-1.00	932.00	-42.00	4474.00
17	-1.00	932.00	-42.00	4474.00
18	-1.00	932.00	-42.00	4406.00
19	-1.00	932.00	-42.00	4406.00
20	-1.00	932.00	-42.00	4406.00

Table 1: Example data from PCA's (without scaling) one cluster.

We can observe part of the results in the Table 1. We only present a small subset of the extracted cluster however we can easily observe that the socio-economic values were indicating character of this group. We expect that rest of the formulated clusters have a similar structure. In context of the extracting knowledge it may be very interesting result because it's possible to extract some correlated periods in economics however in context of clients clustering or term deposit subscription it's not a direction we want to follow (because we cannot see any particular structure in y variable). Let's explore if something will change when we scale our dataset.

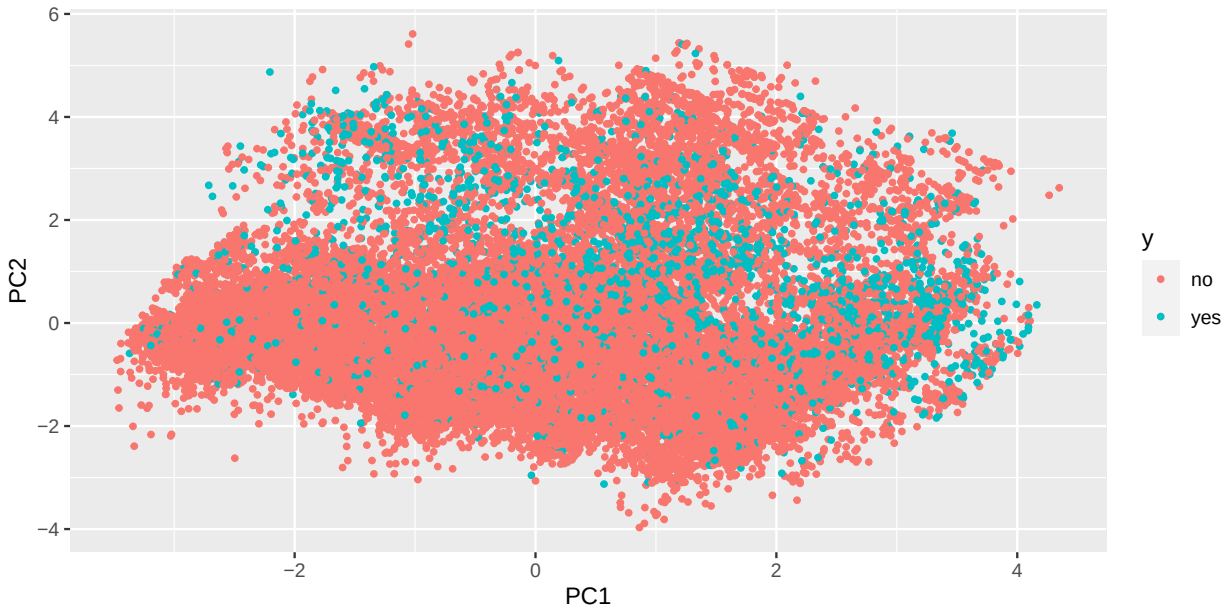


Figure 2: Results of the Principal component analysis with scaling in the context of y variable.

We can observe obtained results on the Figure 2. We can distinguish some area of the plot which is mostly covered by *no* response - lower and upper part of the plot, and middle one with the advantage of *yes* response. That's a very good information in context of our classification task where such a structure may be very helpful for a model.

However for clusters analysis we would like to analyse information only about clients so we will use only variables connected to them and skip variables connect to socio-economic factors. The resulting variables are as follows:

- age,
- job,
- marital
- education,
- housing,
- loan.

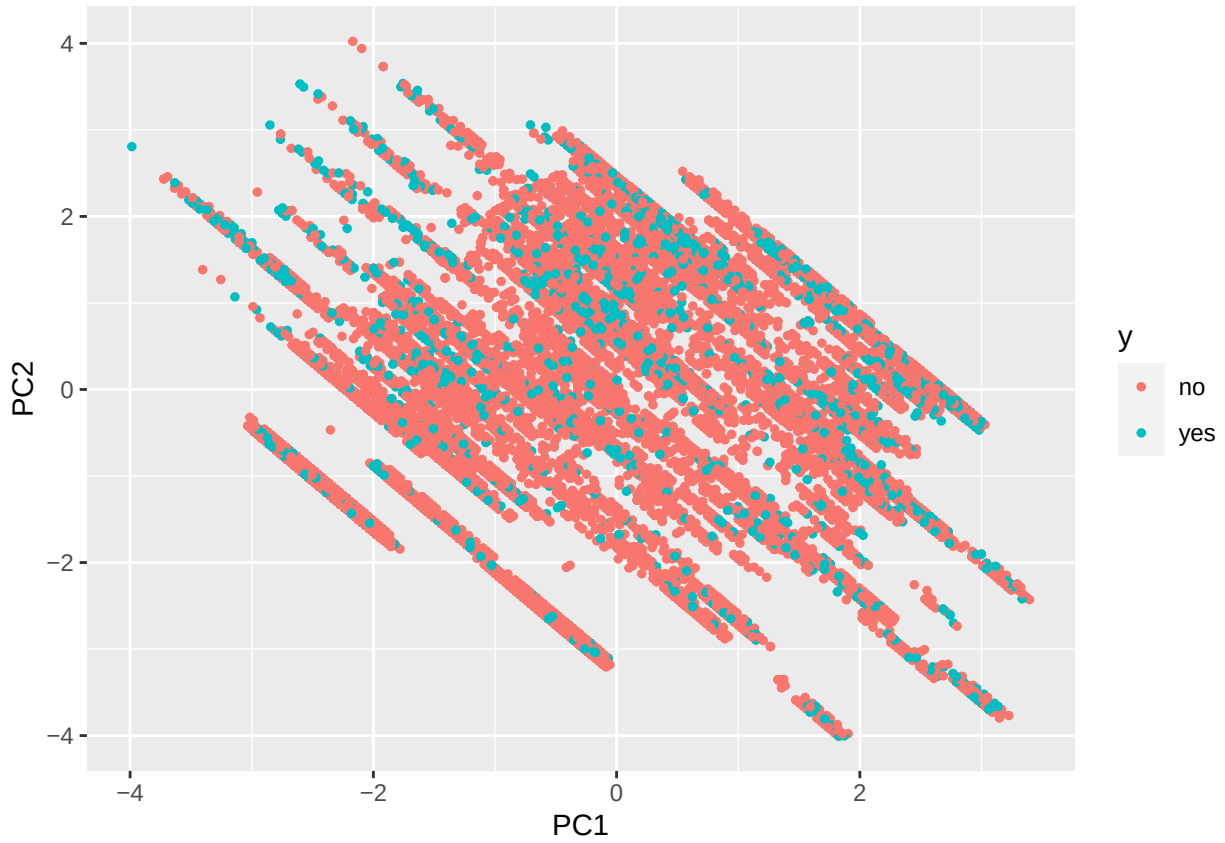
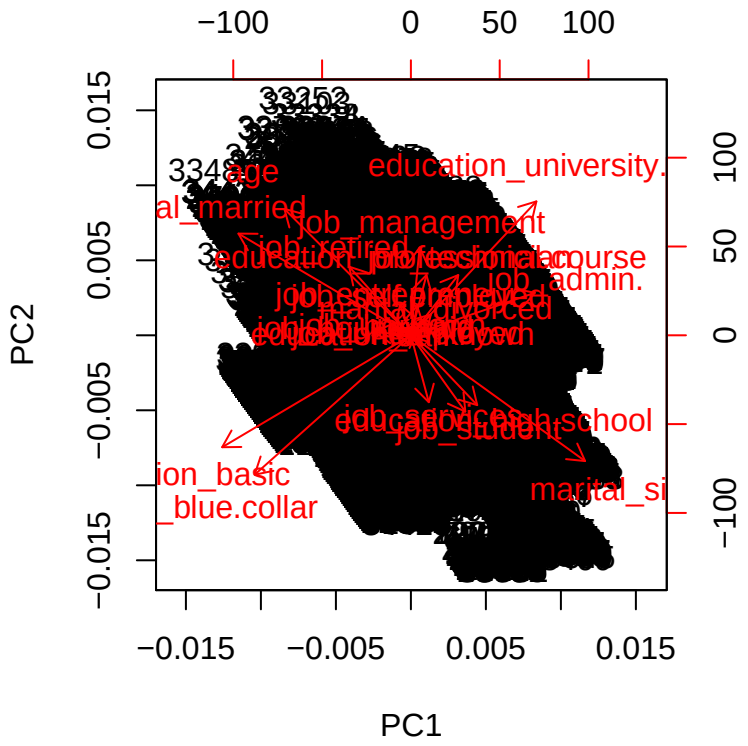


Figure 3: Results of the Principal component analysis in the context of y variable using age, job, marital, education, housing, loan variables.



The results we can observe on the Figure 3. We can

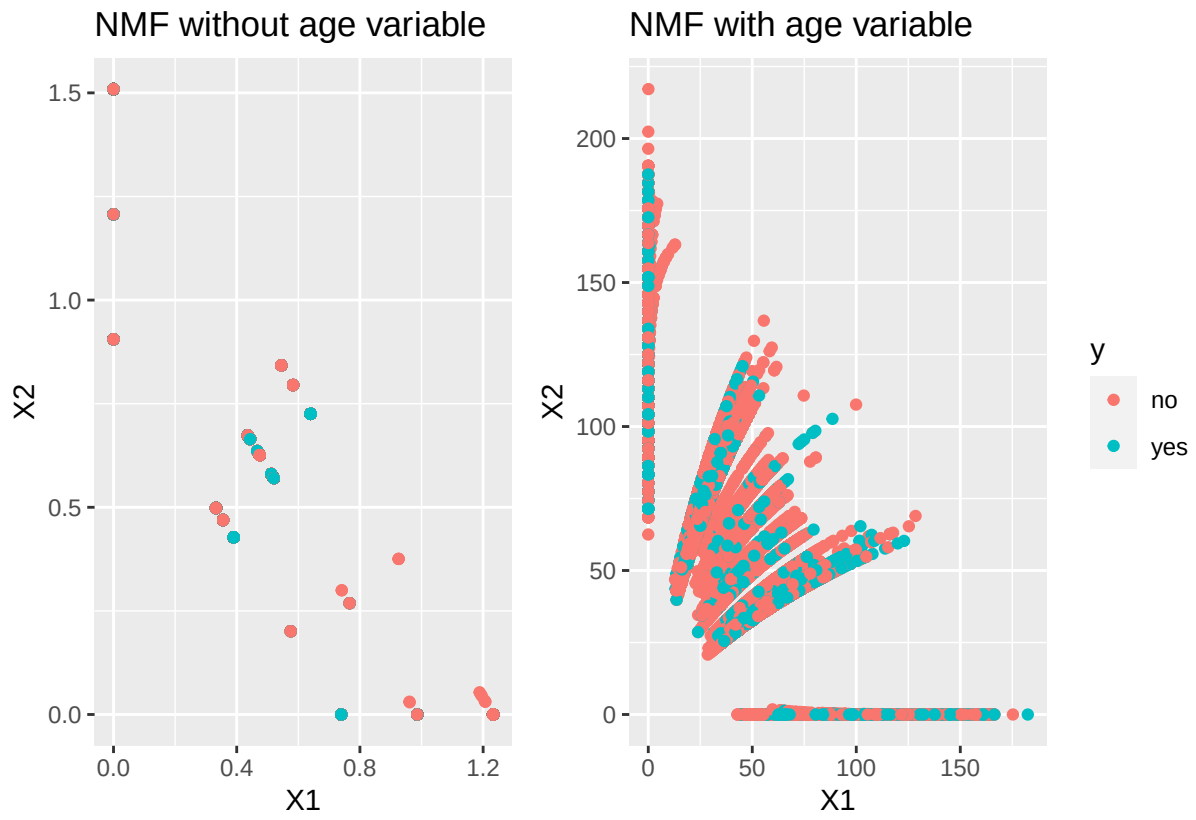
2.1.2 Multidimensional scalling

Not possible to calculate ... as they need to calculate all comparisons between ~34000 samples ? Maybe resampling ?

2.1.3 FastICA

Same as PCA

2.1.4 Non-negative matrix factorization



2.2 Classification

Similar to the 1. project

2.3 Clustering

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

Ideas: - Use Custering LARge Applications (CLARA) instead PAM if computations are slow - Different dissimilarity measures - Gower distance - Analyse what are the characteristic properties of objects thatwere assign to a given cluster (for example, in individualclusters you can analyse: average values for numerical featuresand counts for qualitative features). - Dendogram and banner plot for AGNES - Single / Complete / Average linkage for AGNES - Dunn index for cluster results analysis

Cluster results analysis:
- Internal validation:Use the averagesilhouette indexto comparethe results obtained for different clustering algorithms (e.g. PAMand AGNES) and for a different number of clustersK. Try to decideon the optimal number of clusters. - External validation:Use simple contingency table (confusionmatrix) to compare clustering results with real class membership. Compare results for different clustering algorithms, includingpartitioning and hierarchical methods. (Hint: you can use functionmatchClasses{e1071} to find anoptimal assignment (mapping) between two set of labels).

Notes: - AGNES will probably require data sampling

3 Results and discussion

References

[1] P. Cortez S. Moro and P. Rita. A data-driven approach to predict the success of bank telemarketin. *Decision Support Systems*, 2014.