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).

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.

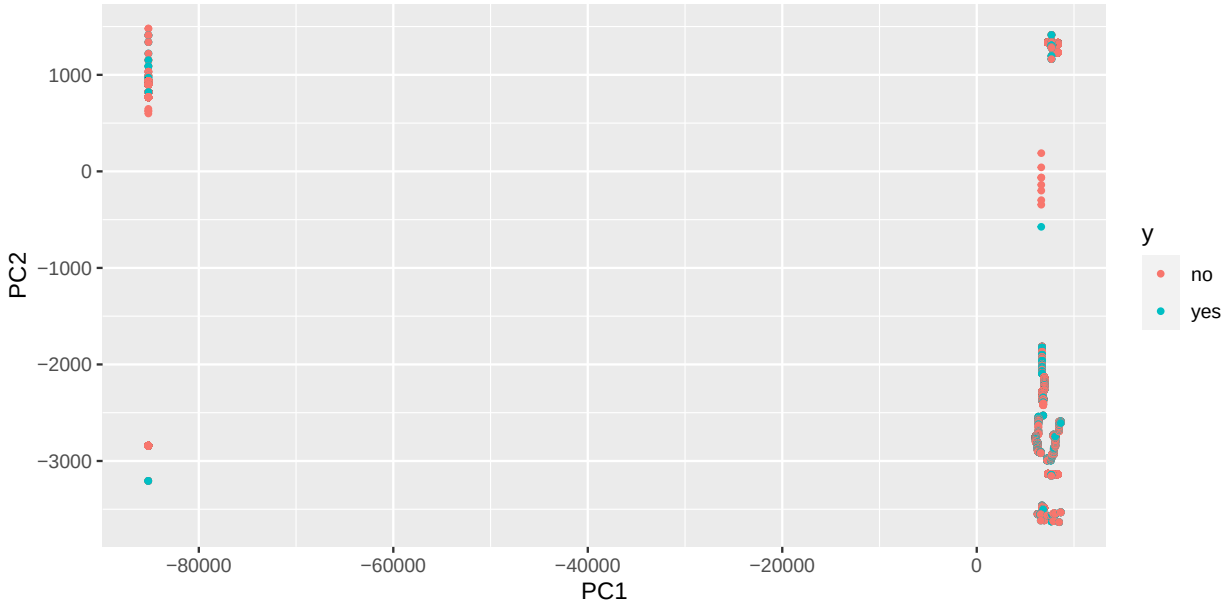

Figure 1: Results of the Principal component analysis without scaling in the context of y variable.

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.

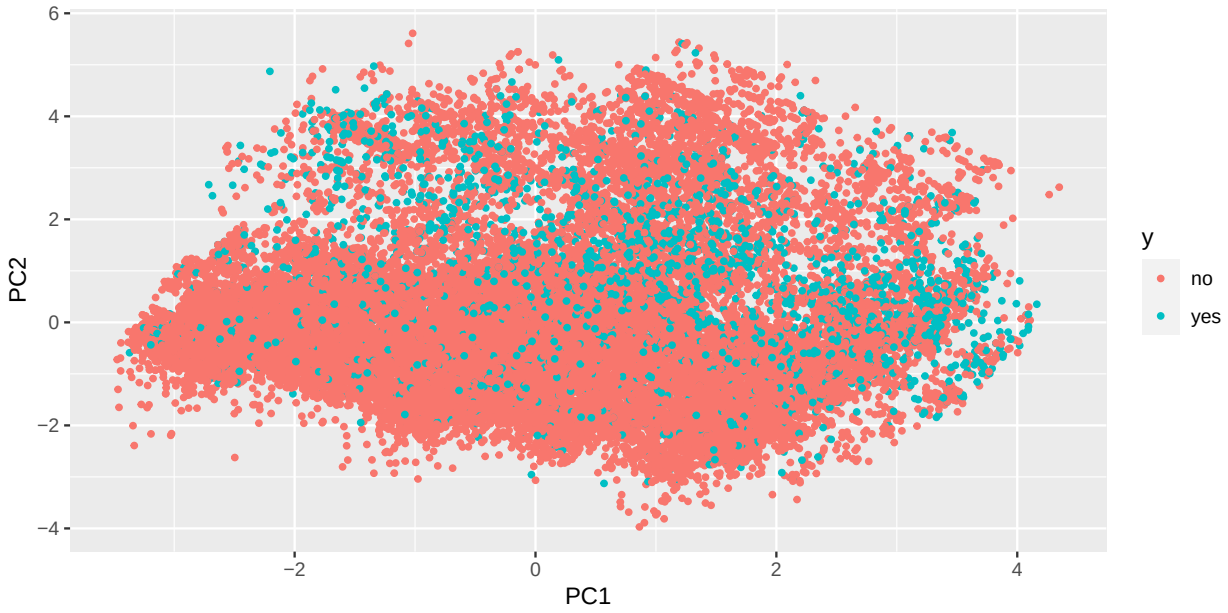

Figure 2: Results of the Principal component analysis with scaling in the context of y variable.

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. Let's explore more deeply these results and find out what number of the components should be optimal and what variables are in these components.

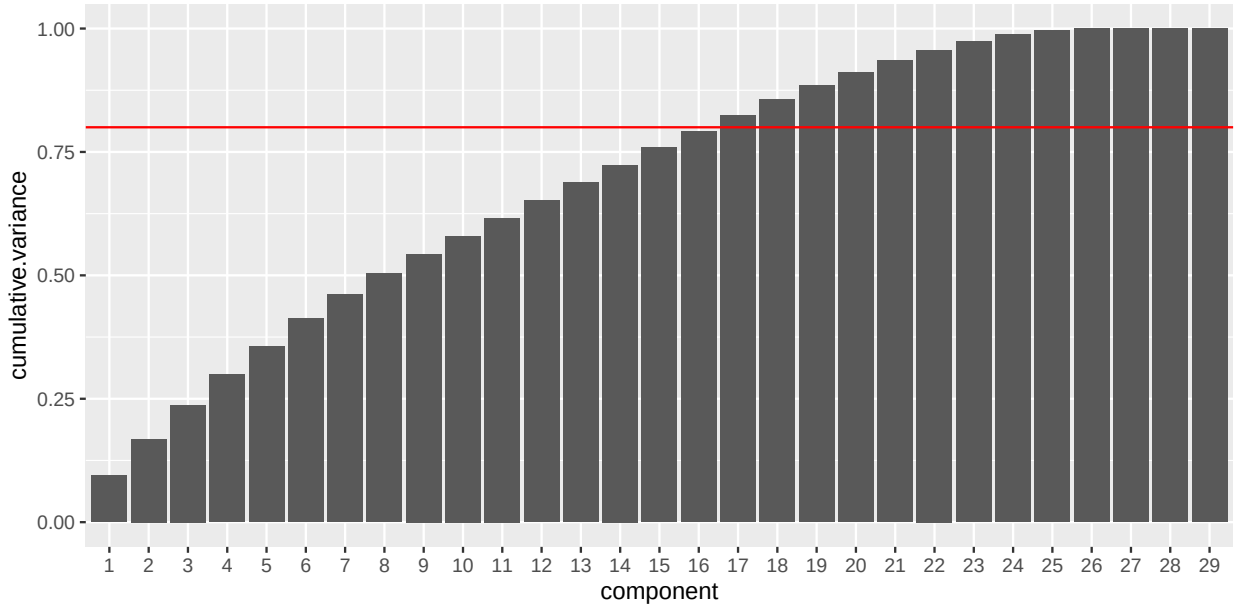

Figure 3: Cummulative variance plot for PCA with data scaling.

We can observe on Figure 3 that our first component explains only around 10% of the variance. Moreover we require first seventeen component to explain more than 80% of variance. That could potentialy explain why we don't have such a clear groups in the 2D plot. Based on these observations we will take these components for classification task.

	.
education_basic	0.41
marital_single	0.37
marital_married	0.37
job_blue.collar	0.35
education_university.degree	0.30
job_admin.	0.27
age	0.25
emp.var.rate	0.21
contact_cellular	0.20
euribor3m	0.18

Table 2: Top 10 most influential variables in the first component of PCA with scaling.

Let's also take a look into Table 2 where we have information about content of the first PCA's component. It contains mostly information about basic/univesristy education, marital status and blue collar / administration job.

So far we have prepared PCA for classification task and now we will focus on PCA in context of clusters analysis. In this case 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.

We've tested the different combinations of PCA - with / without scaling and with / without age variable (which is the only one numeric variable). The most interesting results we've obtained for PCA without scaling and without age variable. The results are presented on Figure 4.

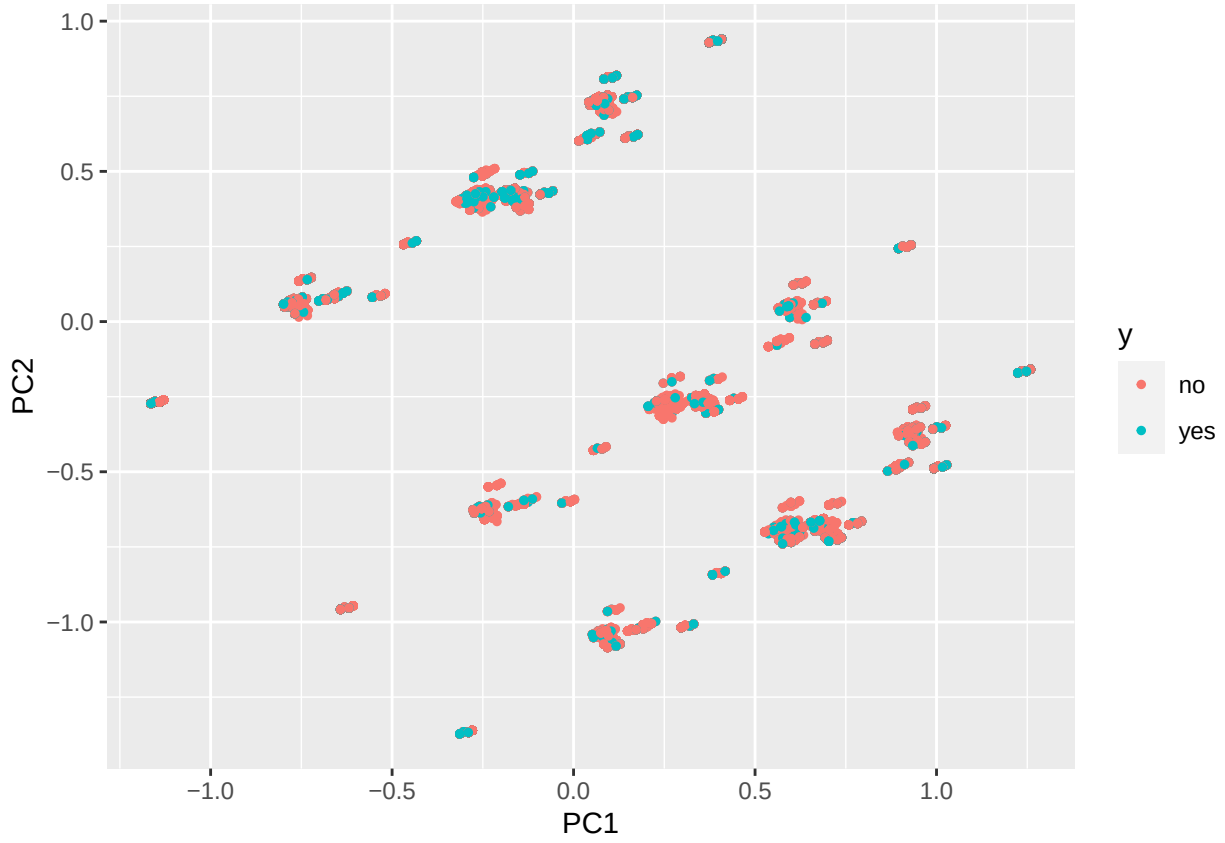

Figure 4: Results of the Principal component analysis in the context of y variable using age, job, marital, education, housing, loan variables.

As we can observe we've obtained clearly separable clusters for only 2D results which is very interesting and may be very useful for clustering task. Similarily to the previous PCA analysis let's explore one cluster, for example $PC1 < -1$.

As we can observe in the Table 3 we have found cluster of people who work as a blue collars, are married and have basic education. It's very interesting result as we will be able to found very homogenous group of people.

	job	marital	education	housing	loan
1	blue-collar	married	basic	1.00	0.00
2	blue-collar	married	basic	1.00	0.00
3	blue-collar	married	basic	1.00	1.00
4	blue-collar	married	basic	1.00	1.00
5	blue-collar	married	basic	1.00	0.00
6	blue-collar	married	basic	0.00	1.00
7	blue-collar	married	basic	1.00	0.00
8	blue-collar	married	basic	0.00	0.00
9	blue-collar	married	basic	1.00	0.00
10	blue-collar	married	basic	1.00	0.00
11	blue-collar	married	basic	0.00	0.00
12	blue-collar	married	basic	0.00	0.00
13	blue-collar	married	basic	1.00	0.00
14	blue-collar	married	basic	0.00	1.00
15	blue-collar	married	basic	0.00	0.00
16	blue-collar	married	basic	0.00	0.00
17	blue-collar	married	basic	1.00	0.00
18	blue-collar	married	basic	1.00	1.00
19	blue-collar	married	basic	1.00	0.00
20	blue-collar	married	basic	0.00	0.00

Table 3: Example data from PCA's for cluster analysis one cluster.

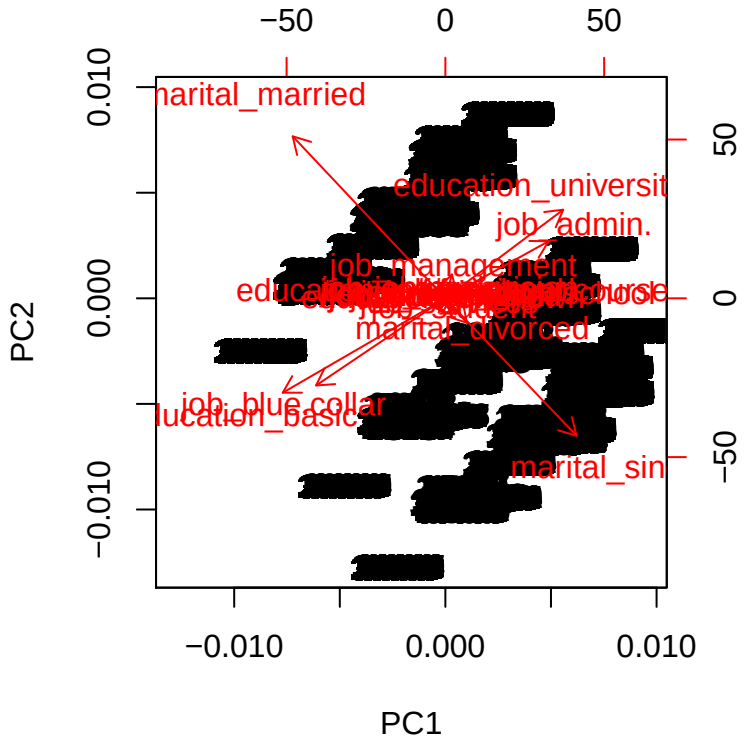

Figure 5: Biplot of the Principal component analysis using job, marital, education, housing, loan variables.

Let's explore a little bit more this PCA resutls and observe Figure 5. We can cleary see that on the left upper corner we've got married people and on the other side we've got respectively divorced and single people which in general create group of the single people. On the opposite diagonal we have got encoded information about job and education. On the lower left corenr we've got correlated blue collar jobs and basic education and on the opposite side we've got administration jobs with university education. To our mind this is very interesting and exciting results as it was possible to so clearly distinguish groups of people.

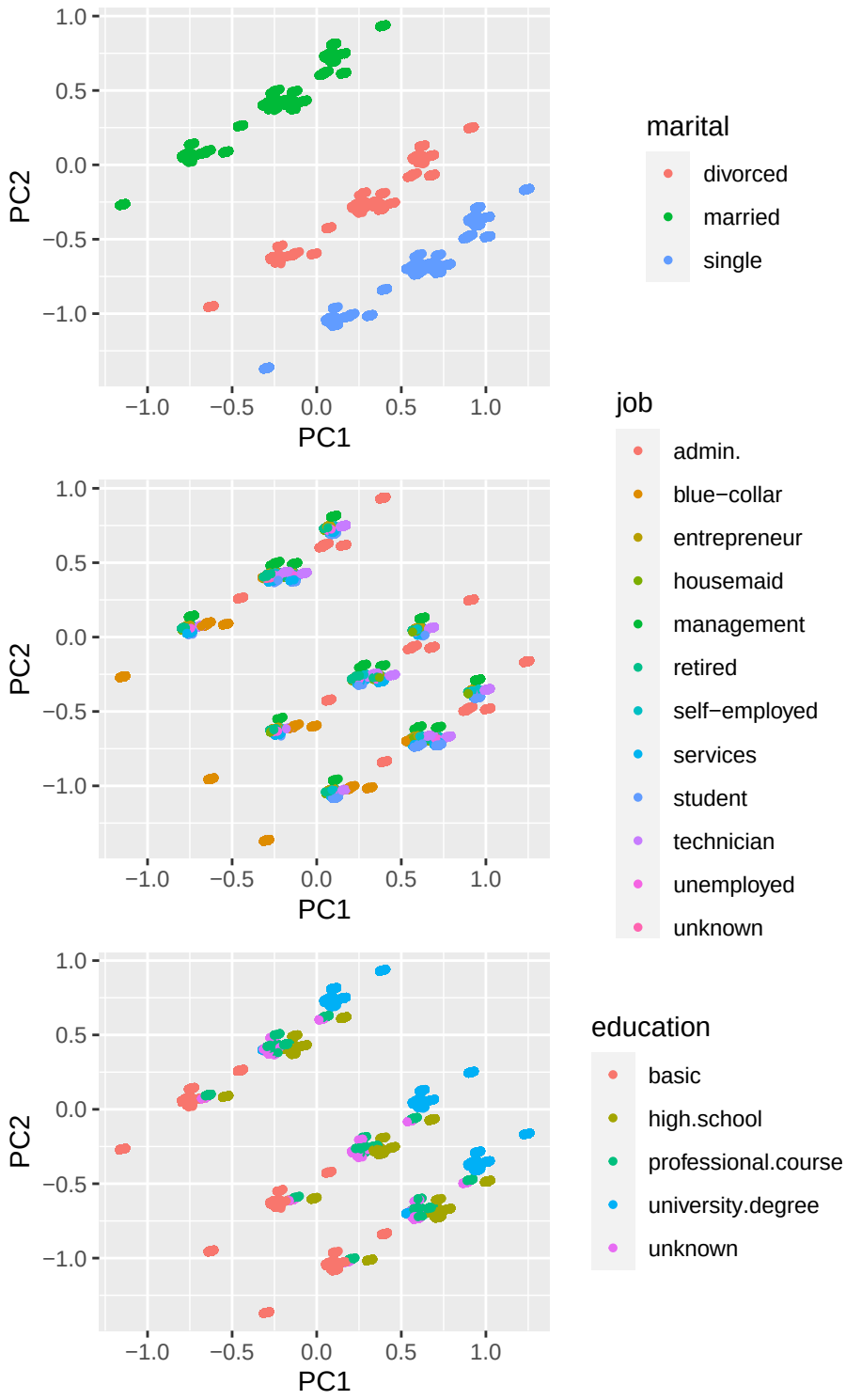

Figure 6: Results of the Principal component analysis in the context of marital, *job*, *education* variable using age, job, marital, education, housing, loan variables.

We checked these hypothesis on Figure 6. We can clearly confirm our observation about marital status and

seperate groups between three formulated lines. In the context of job we can also confirm that we have separate groups of administration and blue collar jobs. Moreover we can observe some kind of the structure in center clusters. At the last part of this plot we can observe some structure in education where on the bottom part we've got people with basic education, in the middle mixed but with majority of high school and professional courses, lastly we have at the top people with university degree.

At this point we will end up our PCA analysis. We've performed extensive analysis with very interesting results which will be definitely utilized during classification and cluster analysis. In context of the cluster analysis we could even say that we've partialy performed it with very satisfactory results.

2.1.2 Multidimensional scaling

In this section we will test another dimension reduction technique - multidimensional scaling. At this moment we recall that our dataset has 34651 rows. It's important in this moment because this method uses similarity matrix which has to compare all rows to all rows. In our case it would produce very big matrix which may be not possible to handle. Because of that fact we will use randomly selected 15% of our dataset with stratification on y variable.

Figure 7: Results of the Multidimensional scaling in context of y variable.

We've performed dissimilarity calculation using gower metric without standarization (using standarization doesn't change anything) using all variables except y and then used multidimensional scaling with rank equal 2. The results can be found on Figure 7. We can easily see two disjoint groups however unfortunately without clear separation on y. Let's find out by which variables our transformed data was separated.

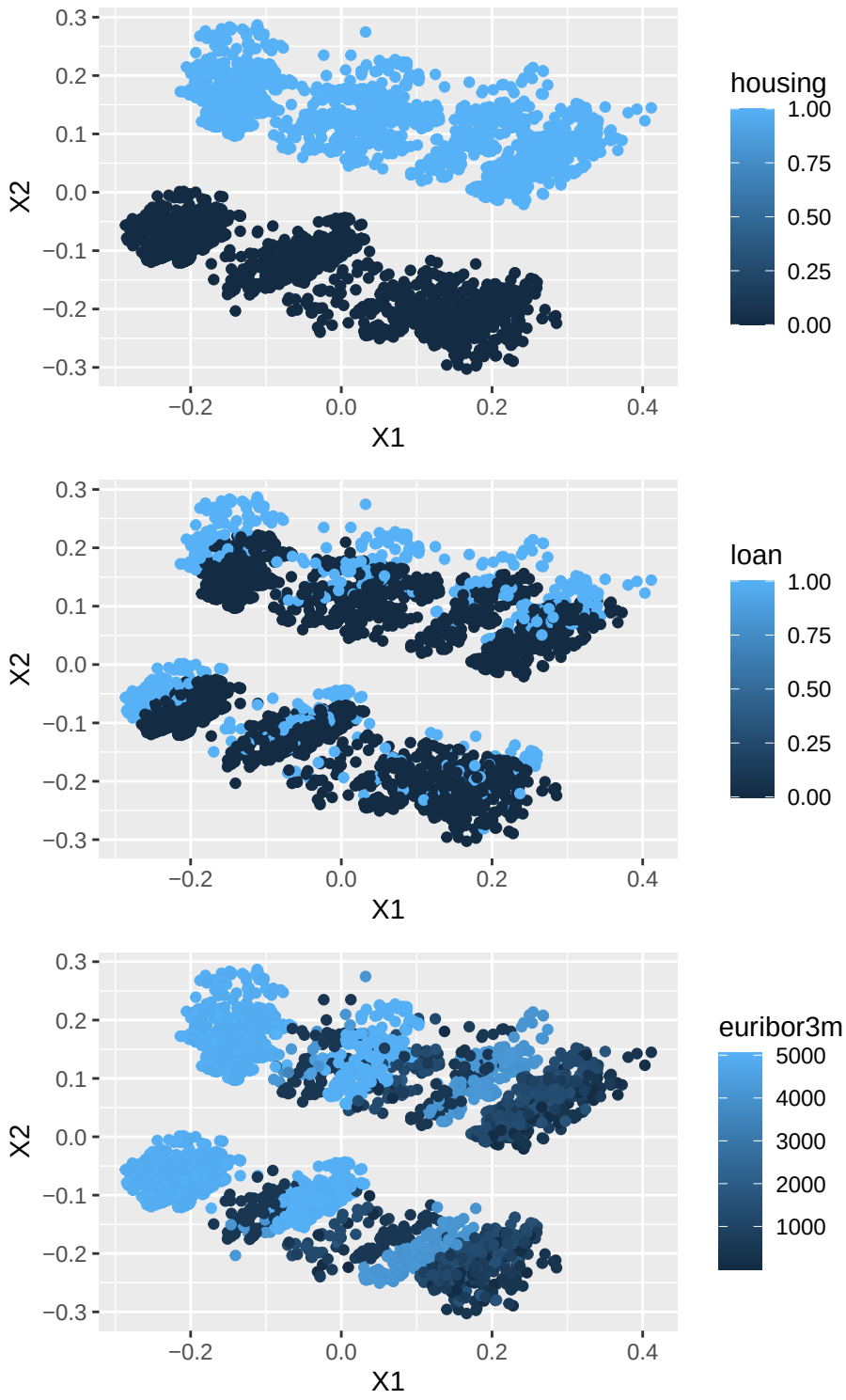

Figure 8: Results of the Multidimensional scaling in the context of housing, *loan*, *euribor3m* variables.

We conclude from Figure 8 that main splitting factor was housing. The other two the most interesting variables in our opinion was loan and euribor. Separation and specific areas of subgroups for the other

variables also could be seen however not so clearly. Similarily to PCA example let's explore only client specific data.

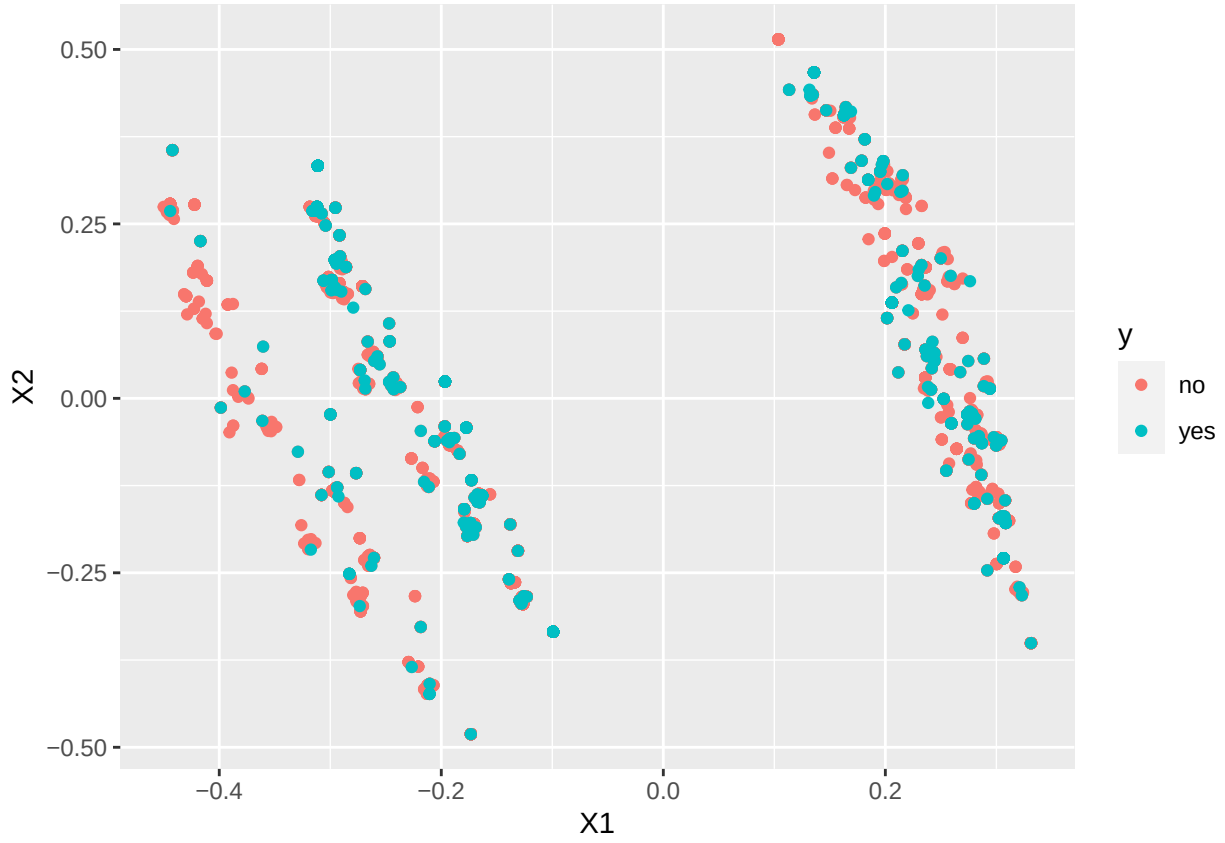

Figure 9: Results of the Multidimensional scaling in context of y variable for only client data.

We can observe results of the MDS in context of only client data on Figure 9. Also here we can find some separate groups. However this time we also cannot see any particular pattern in context of target variable. In next step we will try to find which variables separates groups.

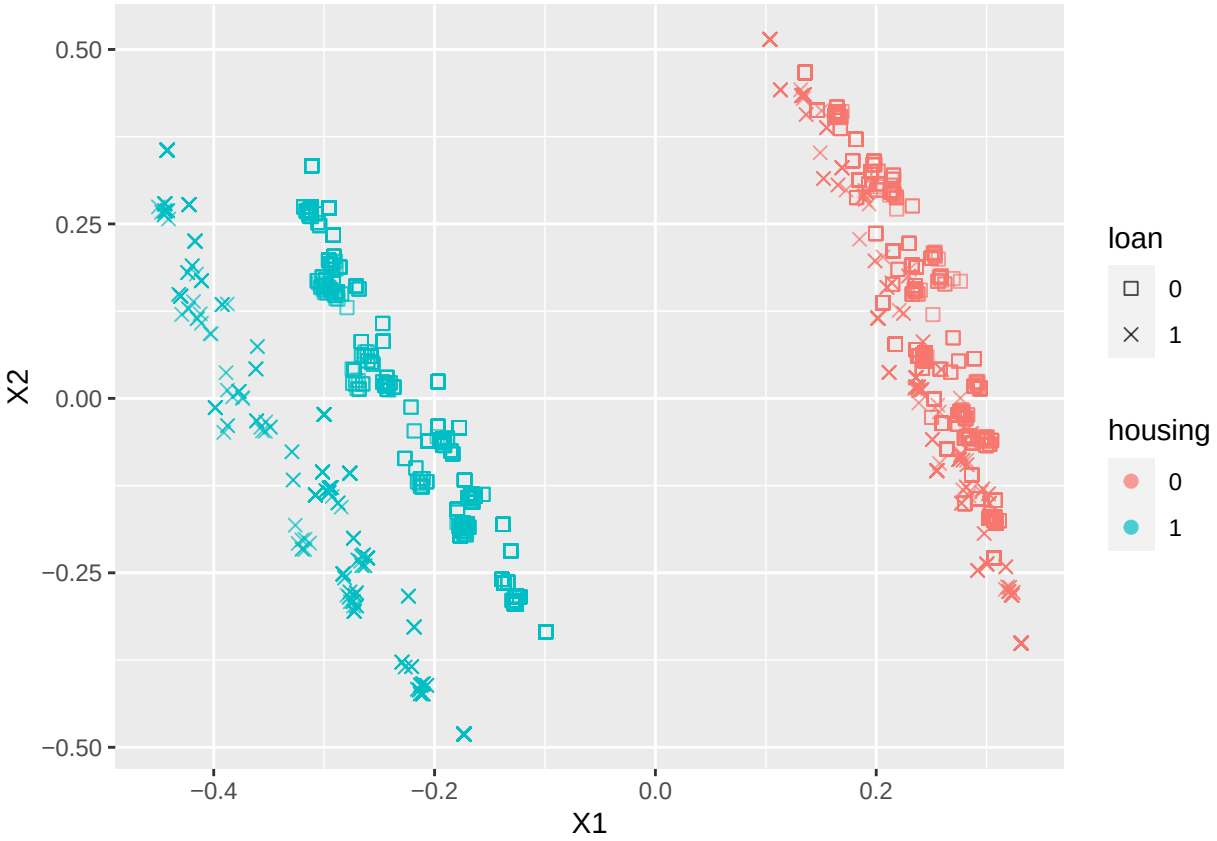

Figure 10: Results of the Multidimensional scaling in context of *housing* and *loan* variables for only client data.

After quick search we've obtained the following results which are presented also on Figure 10. Our main separating varialbe is *housing* and in the group of clients which have a house we can also distinguish people with and without loan. It's quite interesting pattern because it's very natural separation in real-world as people can have house on credit / on their own or doesn't have a house and may or may not have a loan.

Concluding, multidimensional scaling gave us also very interesting results, especially in context of data visualization / understanding and grouping. It also may be very powerful tool for us in the next sections connected to classification and cluster analysis.

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.