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.

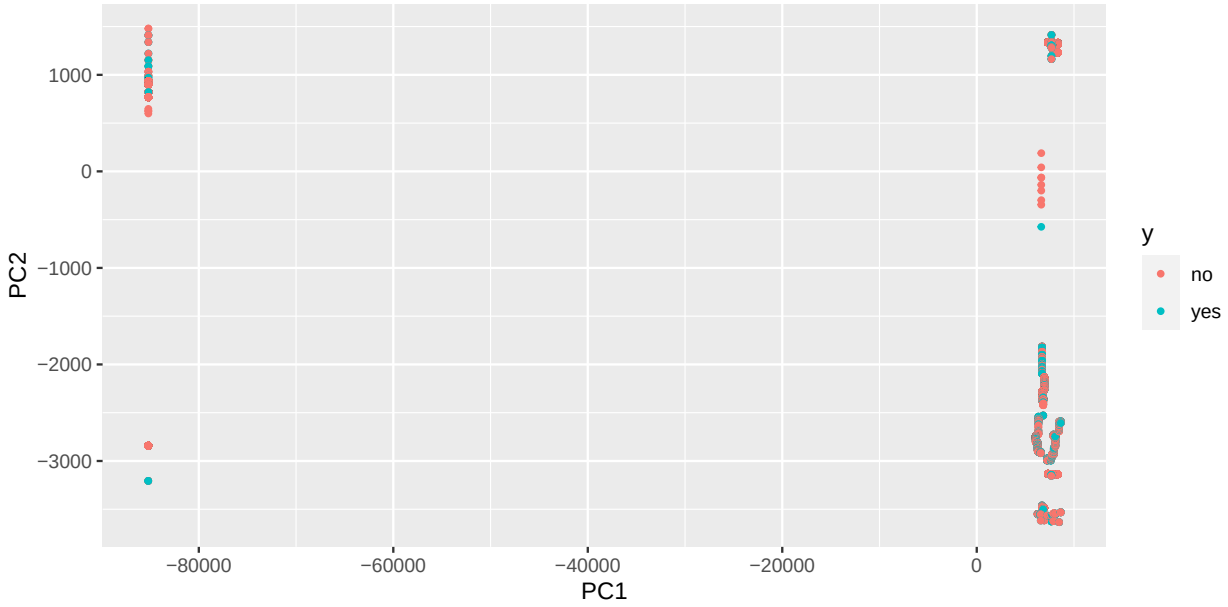

Figure 1: Results of the Principal component analysis without scaling in the context of y variable.

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.

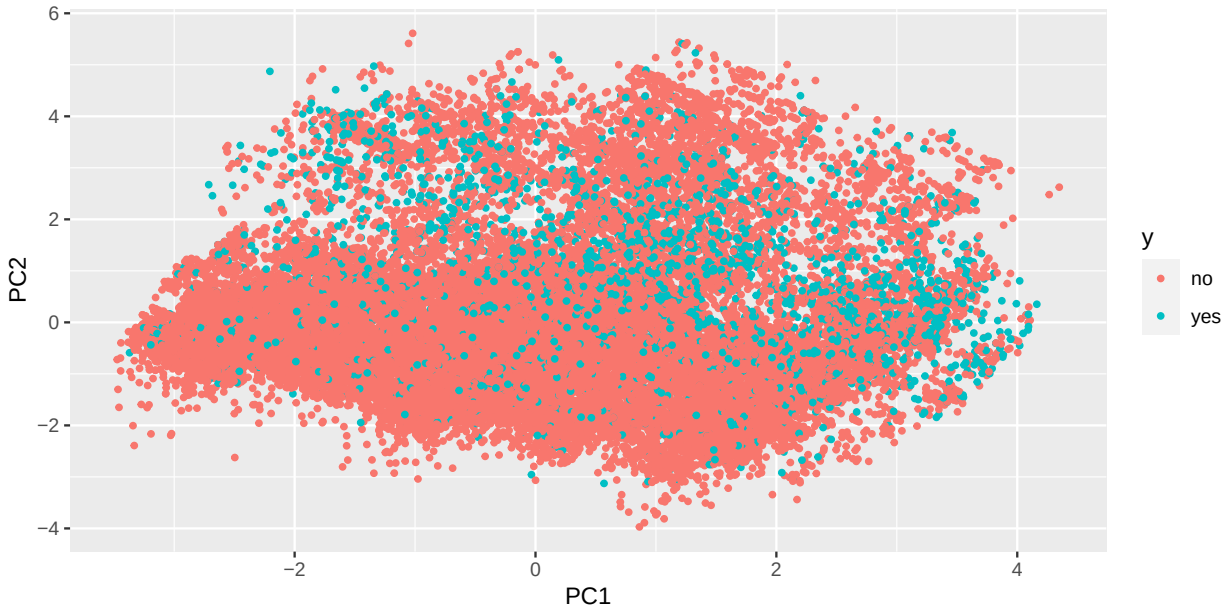

Figure 2: Results of the Principal component analysis with scaling in the context of y variable.

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.

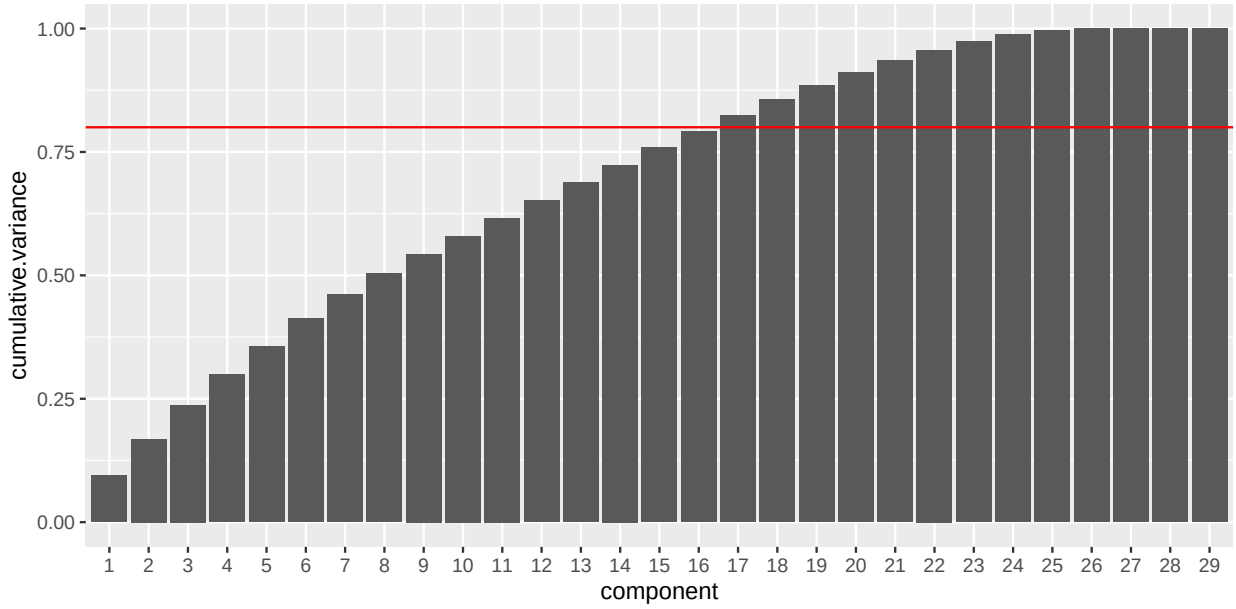

Figure 3: Cummulative variance plot for PCA with data scaling.

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.

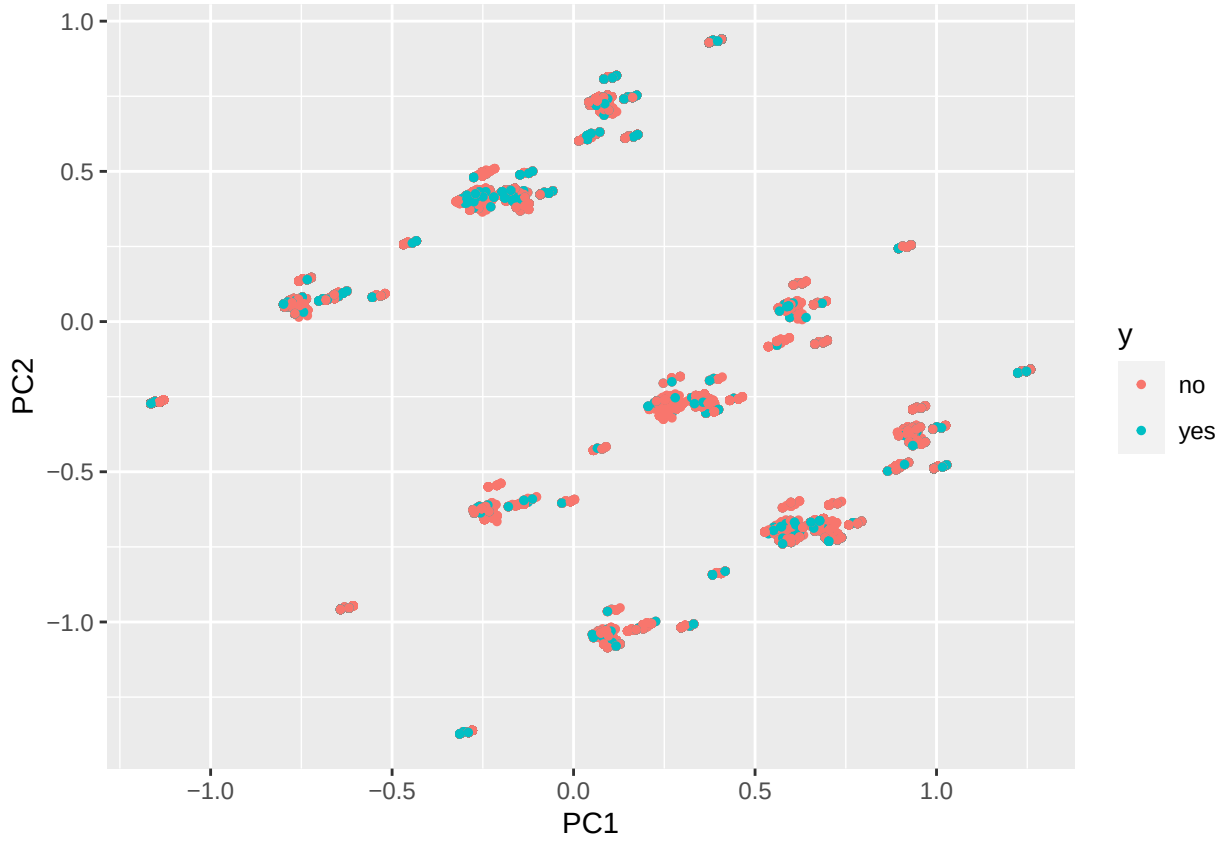

Figure 4: Results of the Principal component analysis in the context of y variable using age, job, marital, education, housing, loan variables.

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.

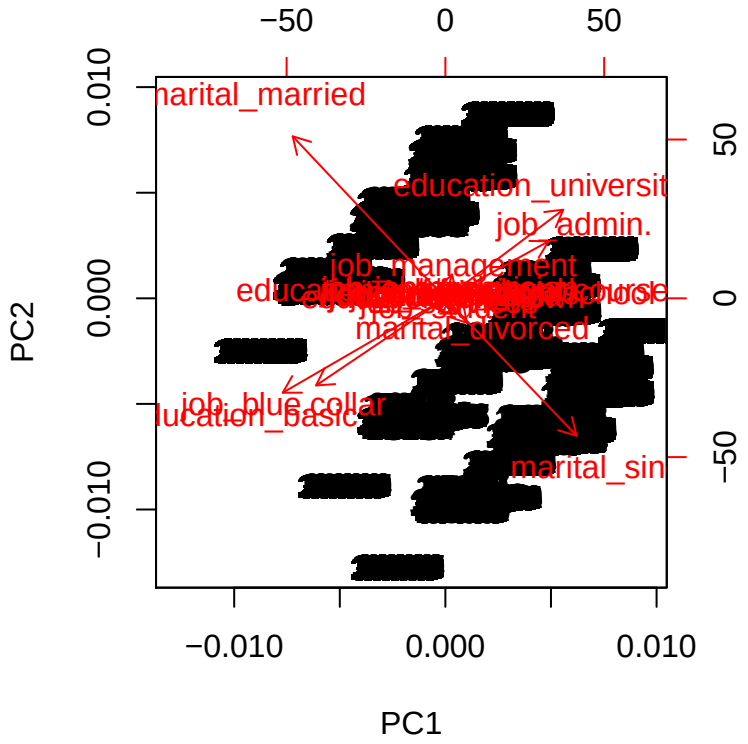

Figure 5: Biplot of the Principal component analysis using job, marital, education, housing, loan variables.

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.

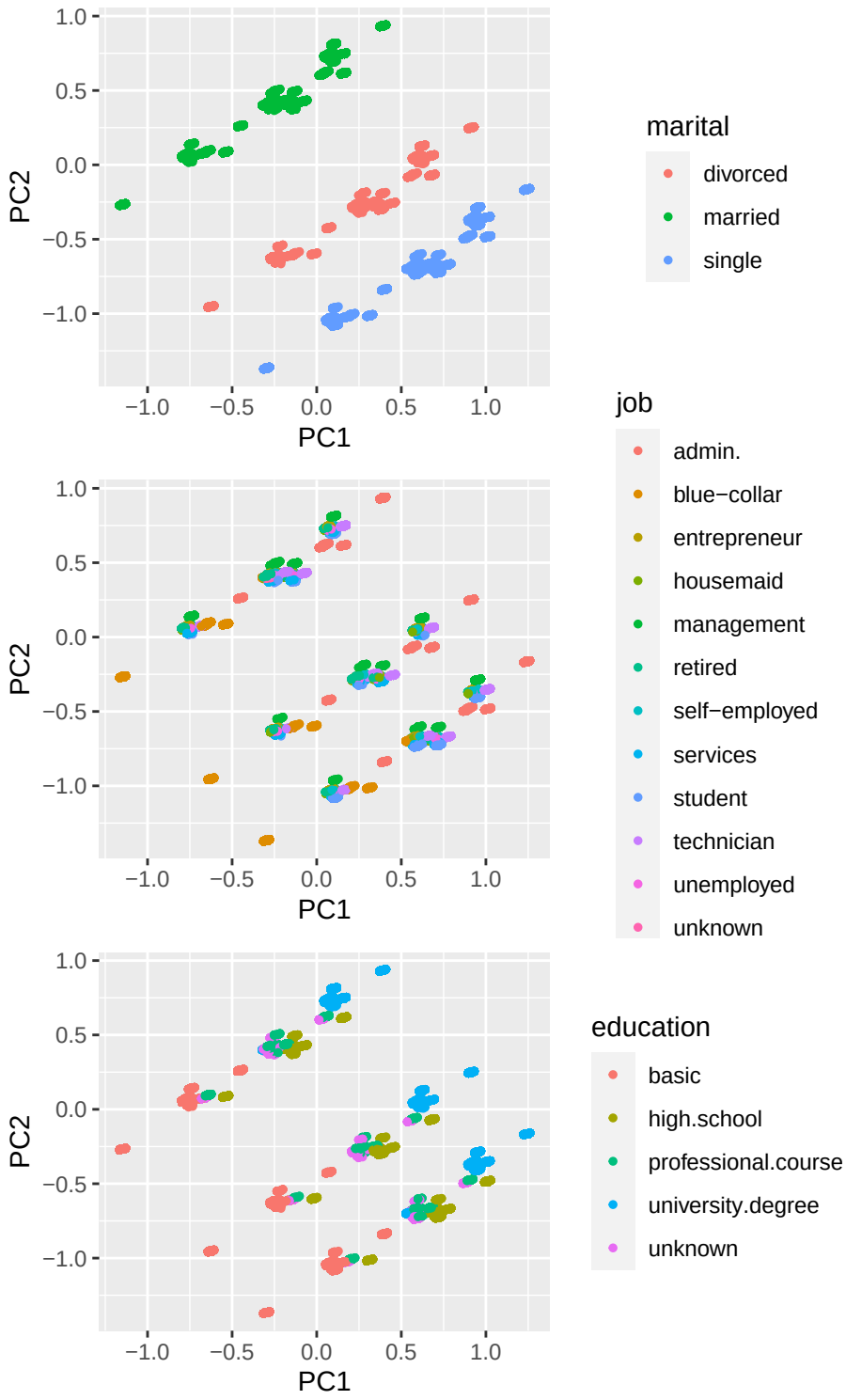

Figure 6: Results of the Principal component analysis in the context of marital, *job*, *education* variable using age, job, marital, education, housing, loan variables.

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.

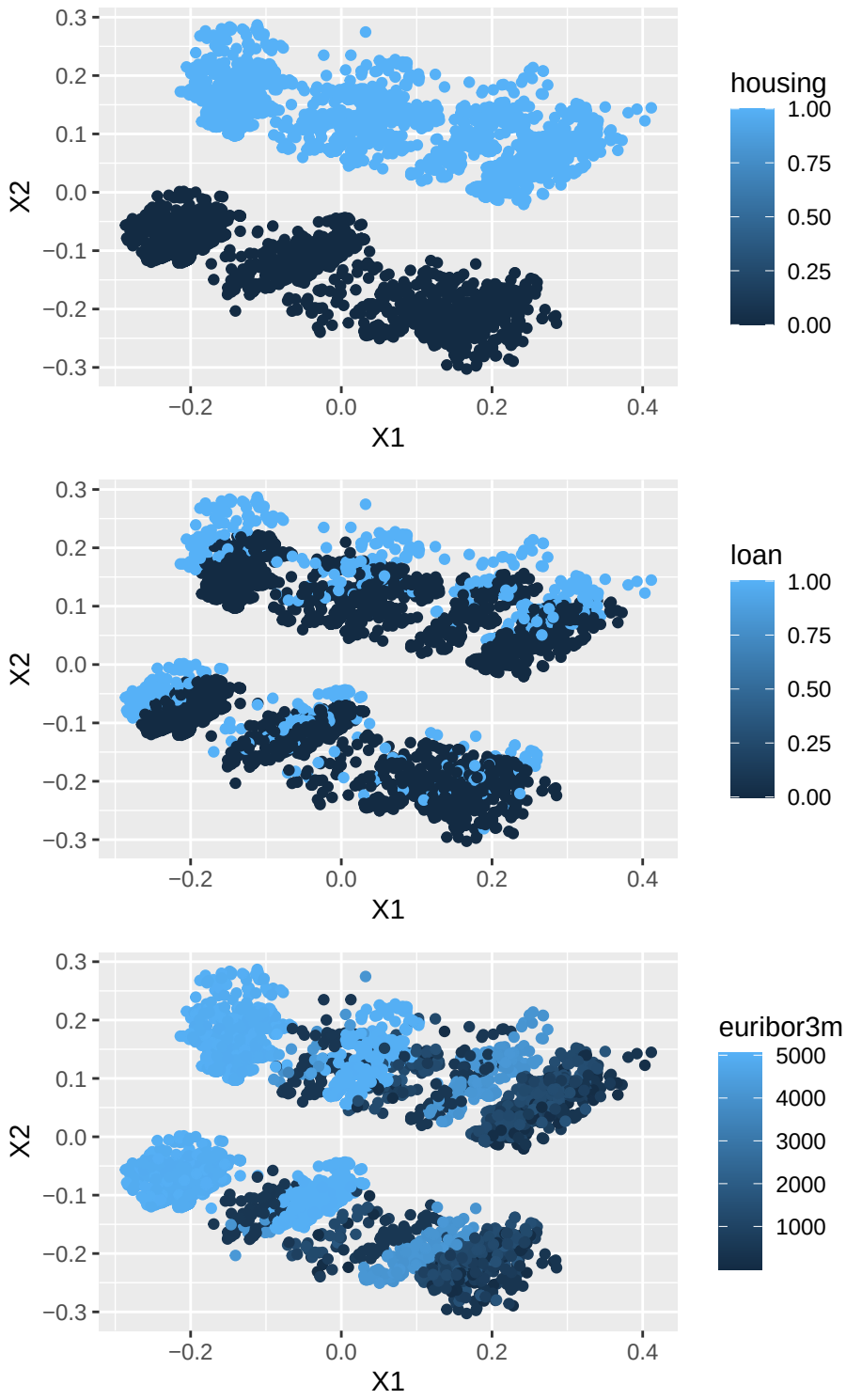

Figure 8: Results of the Multidimensional scaling in the context of housing, *loan*, *euribor3m* variables.

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.

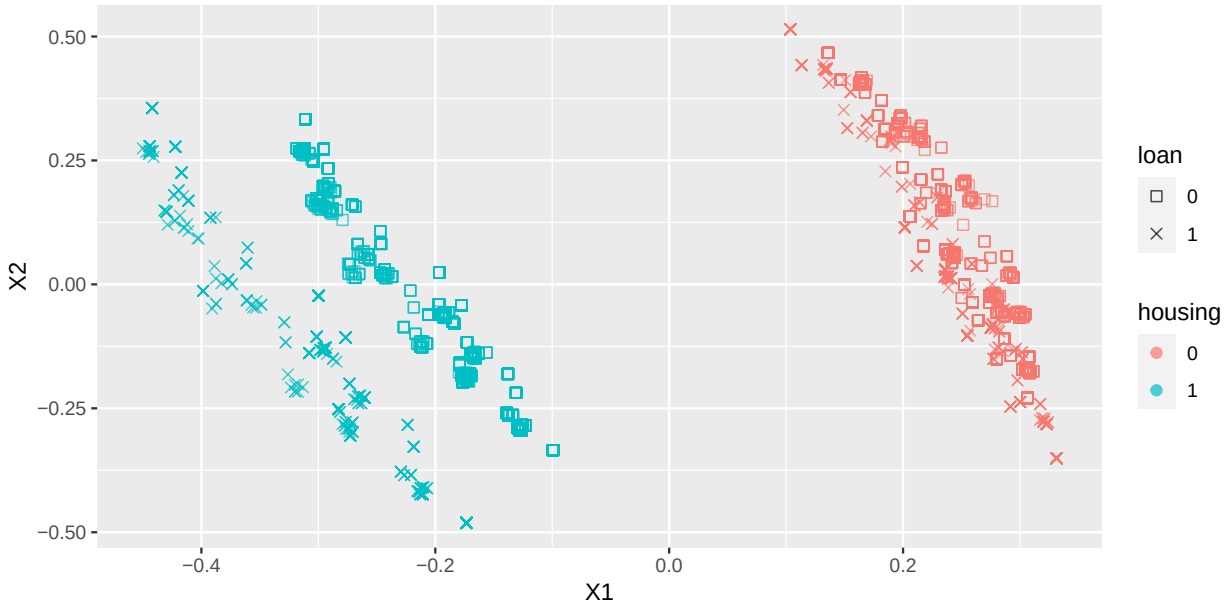

Figure 10: Results of the Multidimensional scaling in context of *housing* and *loan* variables for only client data.

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

In this section we would like to evaluate if dimension reduction techniques could improve results in classification task. The whole set up will be the same as in previous report, i.e. AUCPR will be our main objective metric with 5-Fold Stratified Cross Validation. We will test the following configurations of data:

- dataset with all columns with One Hot Encoding on *job*, *marital*, *education* variables,
- first 17 columns of the PCA on the whole dataset,
- first 5 columns of the PCA on the whole dataset,
- first 2 columns of the PCA on the whole dataset

and we will try the following models:

- Logistic Regression,
- Decision Tree,
- Random Forest (with 50 and 100 trees).

We will skip testing results obtained during multidimensional scaling as it's not possible to use whole dataset and the results between different transformations won't be comparable. Moreover we're testing different number of PCA components to check if we can compress our dataset while perserving similar accuracy.

During the testing it turned out that our set up from first project was wrong because R treated '*no* in target variables as {1} and because of that our results was so superior. In the following experiments we've fixed this bug.

The results of the classification can be found in Table 4. It turned out that the best model is Random Forest with 100 trees on dataset without PCA transformation. The model looks very overfitted however we've set a lot of hyperparameters as defaults and it would be possible to obtain better results. We can also observe that almost all models trained on PCA transformed dataset were worse. Moreover the smaller number of dimensions of PCA transformation the worse results were.

As we remember from our first report our dataset was very unbalanced with 3070 *yes* values and 31581 *no* values. In the next experiment we will try undersampling and oversampling with ratio $\frac{1}{10}$ and 10 respectively. We will also only try Logistic Regression and Random Forest with 100 trees as they were best performing and also we will only use whole dataset and transformed on PCA with first 17 components as lower number of components always performed worse.

The results of balancing experiment can be found in Table 5. Please note that *ranger* in \textit{learner_id} is alias for Random Forest model with 100 trees. We can observe that also this time Random Forest on not transformed and unbalanced dataset was the best one. However the results of the undersampled Random Forest on not transformed dataset has quite close results in terms of AUCPR however it's not so overfitted. Unfortunately also this time the PCA transformation has not improved the results. As a potential next step we would probably take Random Forest model with undersampling technique and tune hyperparameters to obtain even better results.

	task_id	learner_id	precision_train	precision_test	recall_train	recall_test	aucpr_train	aucpr_
1	standard	random-forest-50	0.95	0.48	0.32	0.11	0.75	
2	standard	random-forest-100	0.95	0.49	0.32	0.11	0.76	
3	pca_17	log-reg	0.50	0.49	0.03	0.03	0.27	
4	standard	log-reg	0.51	0.46	0.03	0.03	0.27	
5	pca_17	random-forest-100	0.99	0.38	0.58	0.10	0.98	
6	pca_17	random-forest-50	0.99	0.40	0.59	0.11	0.97	
7	pca_5	random-forest-100	0.99	0.35	0.54	0.08	0.98	
8	pca_5	random-forest-50	0.98	0.36	0.55	0.08	0.97	
9	pca_5	log-reg			0.00	0.00	0.21	
10	pca_2	random-forest-100	0.98	0.30	0.51	0.06	0.97	
11	standard	decision-tree	0.55	0.29	0.34	0.18	0.43	
12	pca_2	random-forest-50	0.98	0.31	0.53	0.06	0.97	
13	pca_2	log-reg			0.00	0.00	0.15	
14	pca_17	decision-tree	0.57	0.20	0.55	0.21	0.51	
15	pca_5	decision-tree	0.57	0.19	0.53	0.19	0.50	
16	pca_2	decision-tree	0.57	0.17	0.51	0.15	0.49	

Table 4: Results of the classification accuracy of the models trained on dataset with and without PCA transformations.

	task_id	learner_id	precision_train	precision_test	recall_train	recall_test	aucpr_train
1	standard	random-forest-100	0.96	0.48	0.32	0.11	0.76
2	standard	undersample.classif.ranger	0.28	0.21	0.81	0.59	0.45
3	standard	log-reg	0.50	0.52	0.03	0.03	0.27
4	pca_17	log-reg	0.50	0.48	0.03	0.03	0.27
5	standard	oversample.classif.ranger	0.70	0.33	0.97	0.39	0.93
6	standard	oversample.classif.log_reg	0.21	0.21	0.59	0.59	0.26
7	pca_17	oversample.classif.log_reg	0.21	0.21	0.59	0.59	0.27
8	pca_17	undersample.classif.log_reg	0.21	0.21	0.59	0.59	0.26
9	standard	undersample.classif.log_reg	0.21	0.21	0.59	0.59	0.26
10	pca_17	undersample.classif.ranger	0.26	0.17	0.97	0.62	0.51
11	pca_17	random-forest-100	0.99	0.37	0.58	0.09	0.98
12	pca_17	oversample.classif.ranger	0.90	0.26	1.00	0.20	1.00

Table 5: Results of the classification accuracy of the models trained on dataset with and without PCA transformations with usage of undersampling and oversampling.

2.3 Clustering

In this section we will focus on cluster analysis task. Our goal will be to group clients according to their similarity. We've done some part of the work in Dimension Reduction section as we were able to find some groups of clients after PCA and MDS transformations however in this part we will do that in more quantitative manner. We will test the following methods:

- k-means,
- Partition Around Medoids (PAM),
- Aglomerative Nesting (AGNES)

and perform quality assesment of cluster analysis results. As we mentioned we want to group clients, so we will use only the features connected to clients, i.e.

- job,

- marital
- education,
- housing,
- loan.

2.3.1 K-means

In this section we will perform cluster analysis using the K-means algorithm. We will begin our journey with this algorithm by looking for within-cluster and between-cluster dispersion depending on number of K. Then we will try to choose optimal number of clusters and analyse results for selected K. For this algorithm we also performed one hot encoding on our dataset.

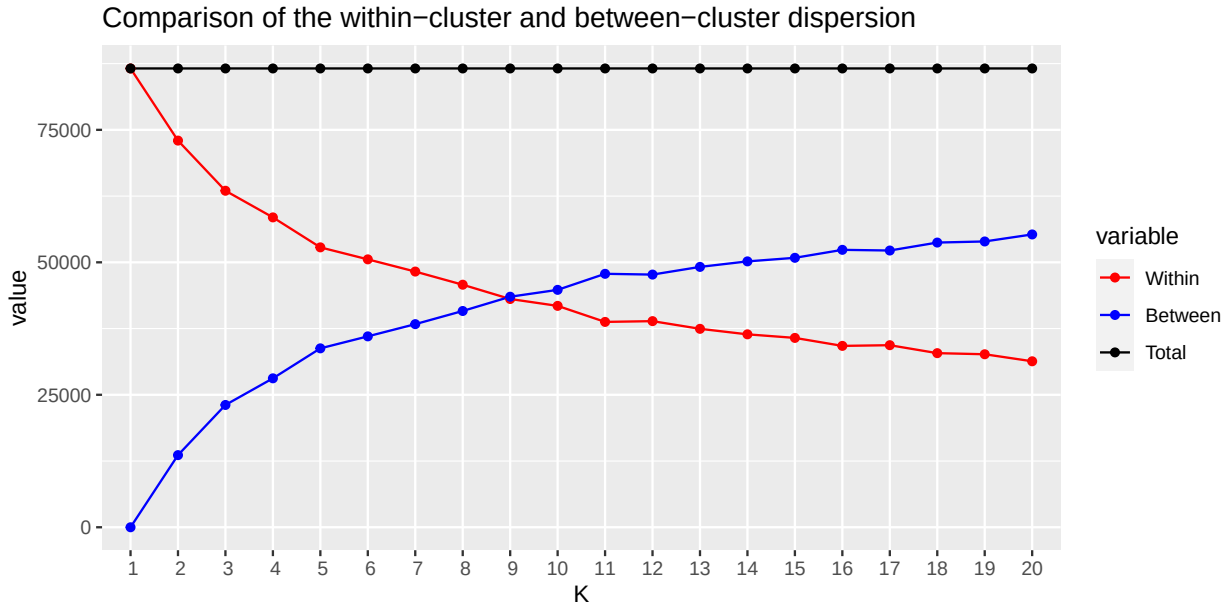

Figure 11: Values of the within-cluster and between-cluster dispersion for different K in K-means algorithm.

On the Figure 11 we can observe how does within-cluster and between-cluster changes depending on K. We can observe that we obtained standard elbow shape for both statistics however without a typical sharp value decrease in some point. Because of that we cannot easily choose optimal K. Let's try to find optimal value of K using silhouette statistic.

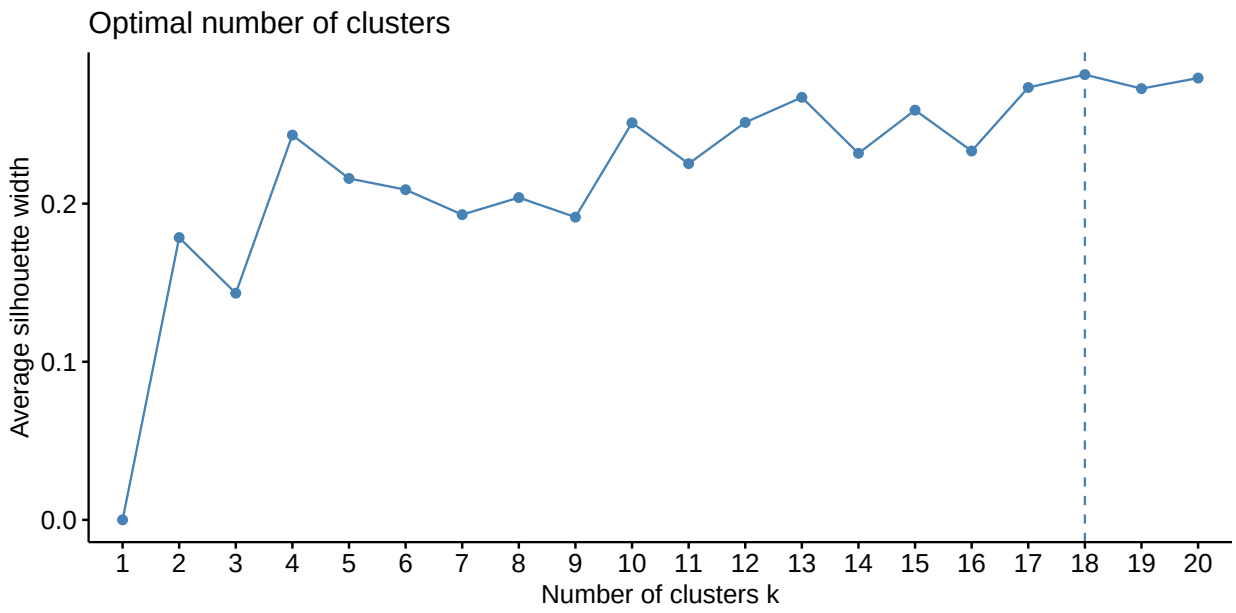

Figure 12: Values of average silhouette width for different number of cluster K in K-means algorithm.

The obtained results are presented on Figure 12. Also this time we cannot easily select one best value as the values doesn't vary much. However we will select K equal 13 as the optimal one as it's very close to the best value $K = 18$ however it's a bit smaller value which enables us to more easily analyse the obtained clusters.

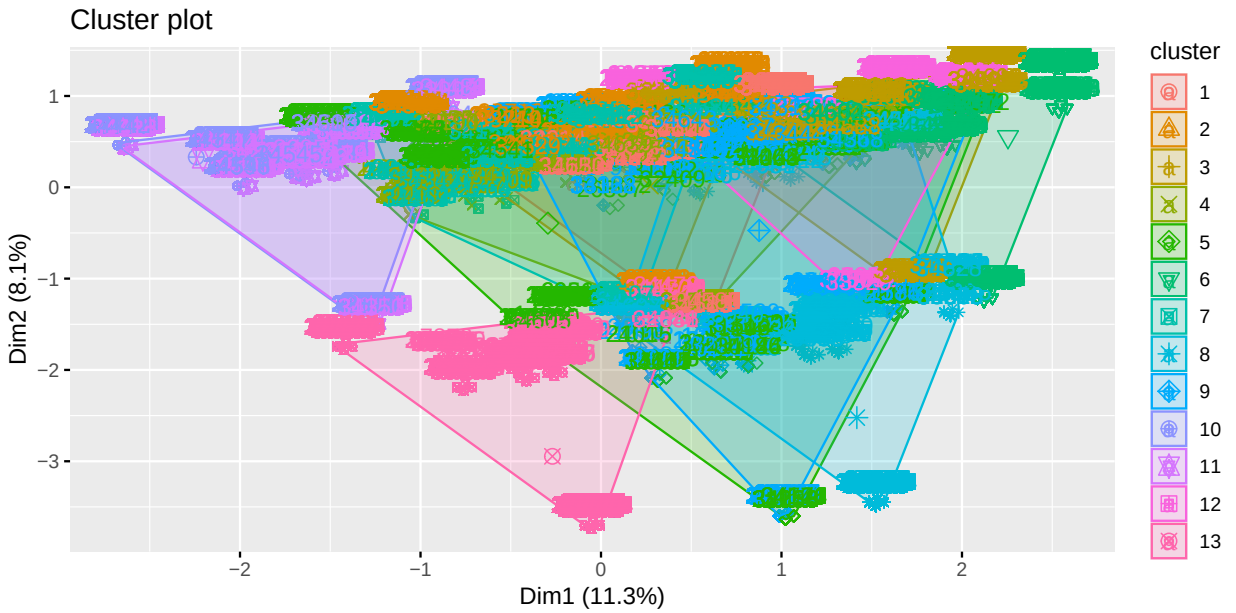

Figure 13: Clusters obtained by K-means algorithm presented in 2D PCA space.

In order to better understand our results we've plotted them in 2D PCA space on Figure 13. We can observe some distinct groups even in this 2D visualization which fill us with hope that obtained results will be interpretable.

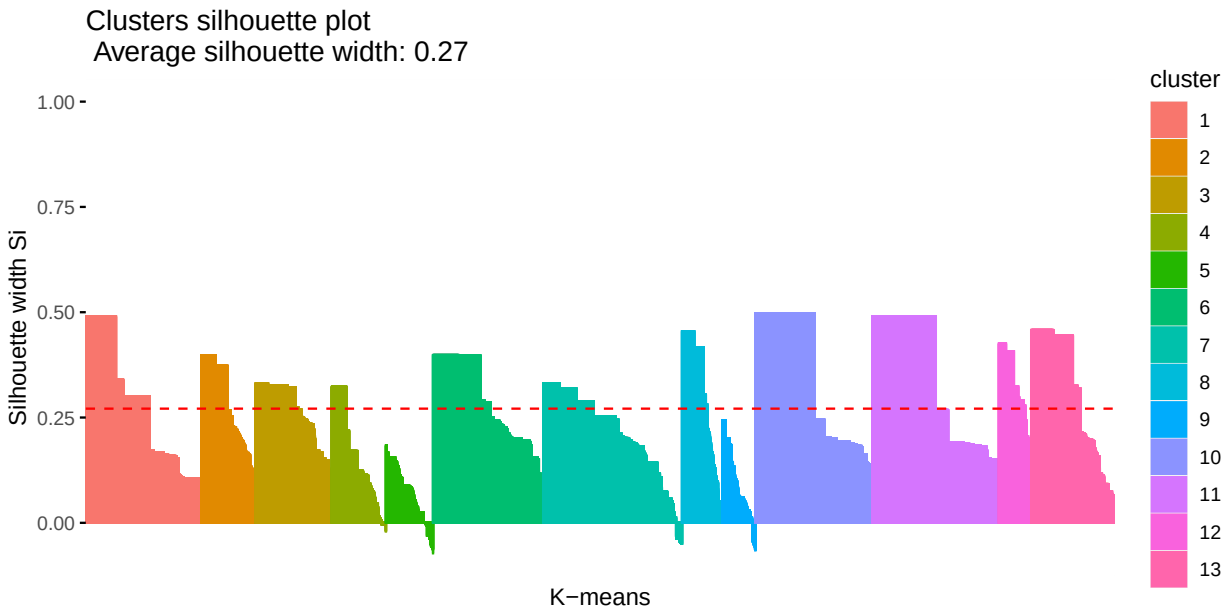

Figure 14: Clusters silhouette plot for K-means algorithm with K = 13.

The results of the silhouette for individual clusters are presented on Figure 14. We can observe that some of the clusters have very high silhouette values which means that they are properly assigned, but some of them have small results or even negative which suggests that some observations were incorrectly assigned. Especially the four negative peaks are probably the results of our choice of 13 clusters instead of 18.

Let us also compare the results in context of different categories and find some very homogenous groups. We will start with target variable y however we don't expect that we will find some specific cluster with only *yes* responses.

	no	yes
1	3500	393
2	1686	147
3	2255	307
4	1666	156
5	1447	138
6	3241	469
7	4314	362
8	1217	141
9	1033	80
10	3684	267
11	3971	281
12	992	92
13	2575	237

Table 6: Clusters vs. target variable

The results can be found in Table 6. As we expected there is no clear separation between clients' responses.

In context of cluster interpretation we can extract one more useful information from Figure 14. If cluster have very large silhouette it means that it should represent quite homogenuous group. Using that information we have found that

- cluster 6 represents clients with basic education, married, blue-collar job and mostly without loan (which is also cluster found during visualization of PCA transformed data),
- cluster 13 represents clients with university degree, married, managment job and mostly without loan (which also was previously recognized),
- cluster 2 represents clients with universtiy degree, married, mostyle administration job and mostly with house,
- cluster 12 represents clients with professional course, mostly working as a technicians.

In our opinion these results give us very interesting insigth into our dataset and may be very helpful in process of better understanding of clients in bank.

2.3.2 PAM

In this section we will very similar analysis to the previous one on K-means however this time we will use another technique - Partition Around Medoids (PAM). This is more robust generalization of K-means in which we will use dissimilarity matrix with gower distance. This will allow us to encode categorical values in different way and we expect to obtain a little bit different results as we obtained different results for PCA and MDS. Because in this section we will use dissimilarity matrix we will perform stratified sampling the same way as we did in MDS.

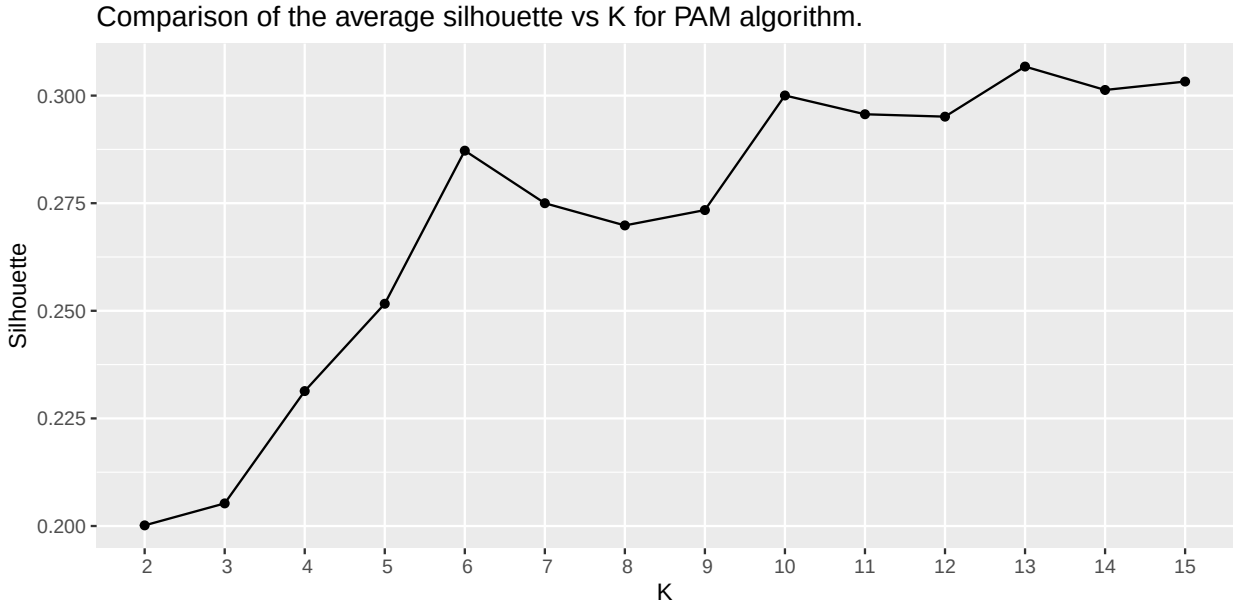

Figure 15: Values of average silhouette width for different number of cluster K in PAM algorithm.

Firstly, we begin our analysis with finding optimal number of clusters. As we can observe on Figure 15 the optimal values is equal 13 - the same as we selected in K-means algorithm. Let's check how does silhouette looks for selected K.

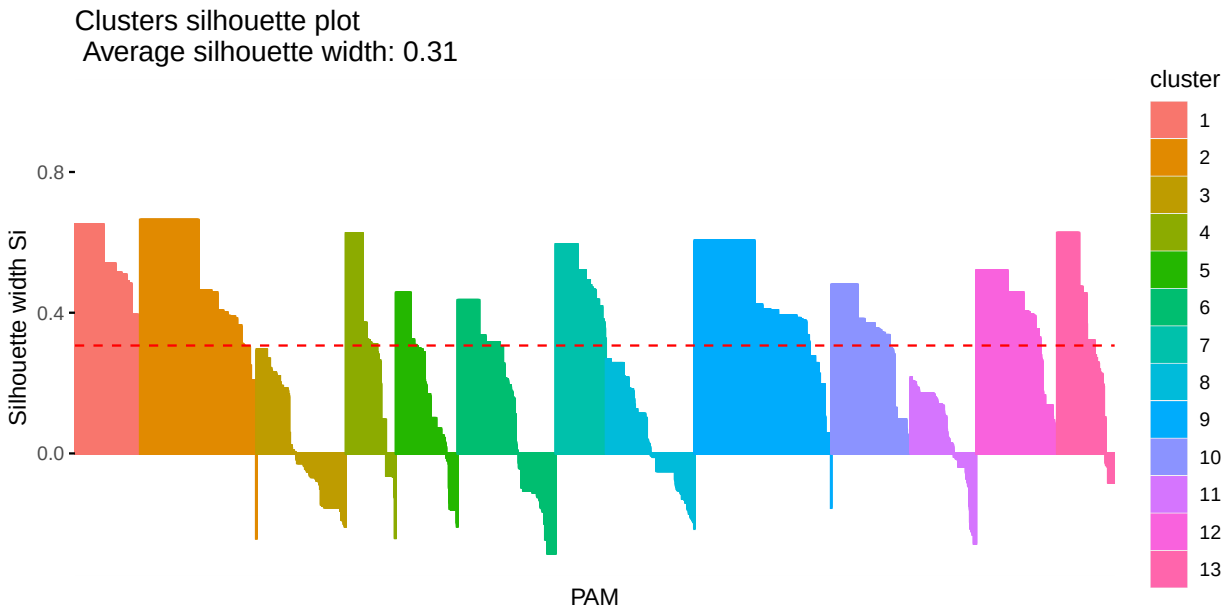

Figure 16: Clusters silhouette plot for PAM algorithm with K = 13.

The silhouette plot is presented on Figure 16. Most of the obtained clusters are quite consistent in terms of silhouette statistics and in the next step we will explore them in more details.

	no	yes
1	295	30
2	546	37
3	395	51
4	224	26
5	281	28
6	455	35
7	221	30
8	412	31
9	631	53
10	343	53
11	306	22
12	364	40
13	264	24

Table 7: Clusters vs. target variable

First of all, this time we also could find any structure in terms of target variable as we can observe in Table 7. Moreover we found that:

- cluster 1 represents clients with university degree, married, administration / management / self-employed job with house and without loan,
- cluster 2 represents clients with basic education, mostly married and working as blue-collars with house and without loan,
- cluster 4 represents clients with mostly high school degree, married, working in services, with house and without loan,
- cluster 7 represents clients with university degree, single, administrative job, without house and without

loan,

- cluster 9 represents clients with basic education, mostly married, working as blue-collars without house and without loan,
- cluster 13 represents clients with professional courses, mostly married, working as technicians, without housing and loans.

As we can observe the results are also very distinguishable however as we were analyzing the results we found that the main splitting criterion was housing and loans where for k-means it was mostly education, job and marital status. However it's quite natural as we used different type of measures for both algorithms. We have also seen differences between PCA and MDS representation and here we've corresponding results.

2.3.3 AGNES

In this section we will analyze our client dataset using AGNES algorithm which represents hierarchical methods of clustering, i.e. we firstly create clusters and then decide what number of clusters we want to obtain by looking on dendograms. This time we will also use dissimilarity matrix of clients with sampling.

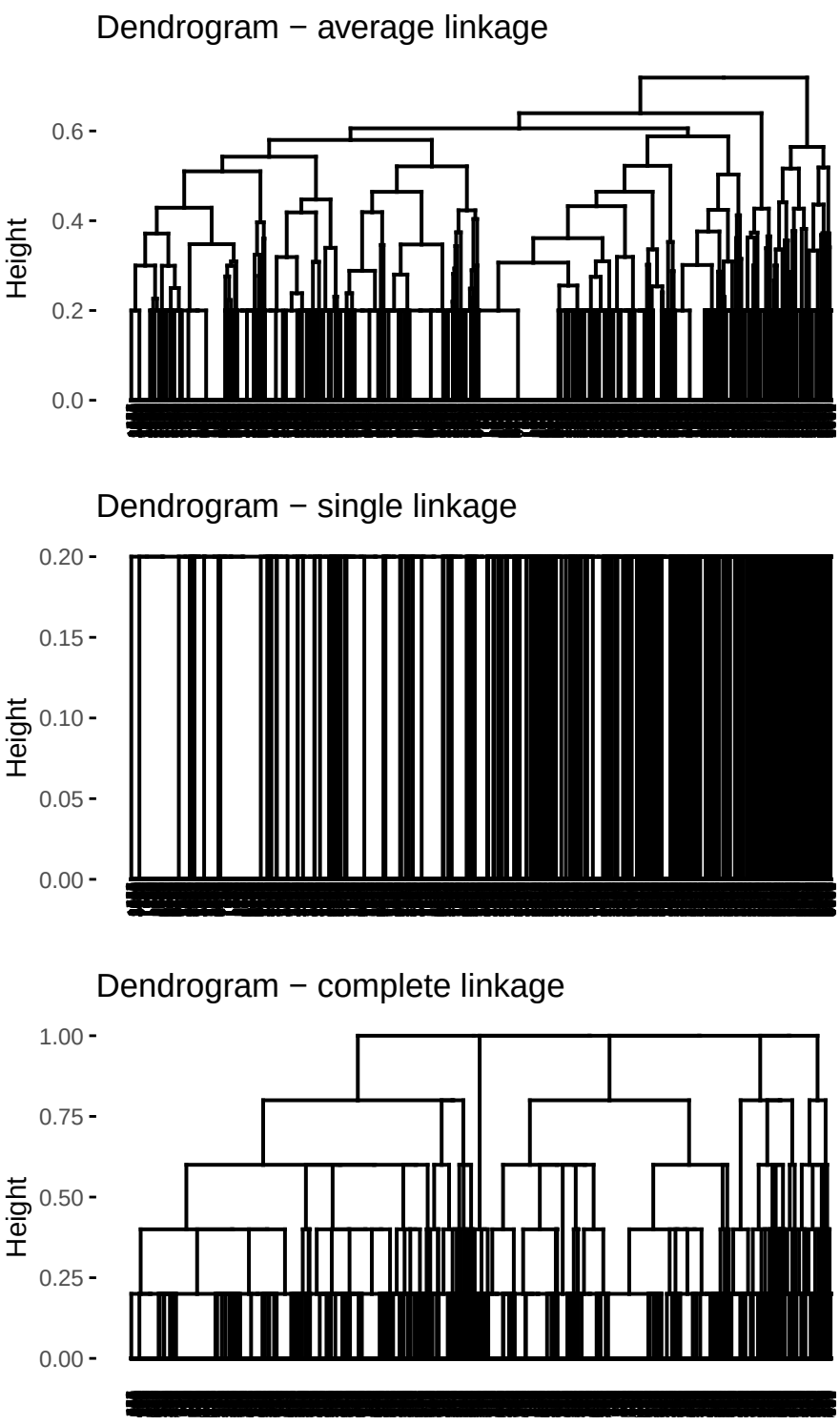

Figure 17: Dendograms obtained for AGNES algorithm using average, single and complete linkage.

The results of the clustering are presented on dendograms on Figure 17. As we can observe the results of single linkage are very unclear and it would be very hard to extract some meaningful clusters. However

complete linkage and average linkage are doing good job and we can recognize some groups. Let's explore in more details average linkage version of AGNES algorithm.

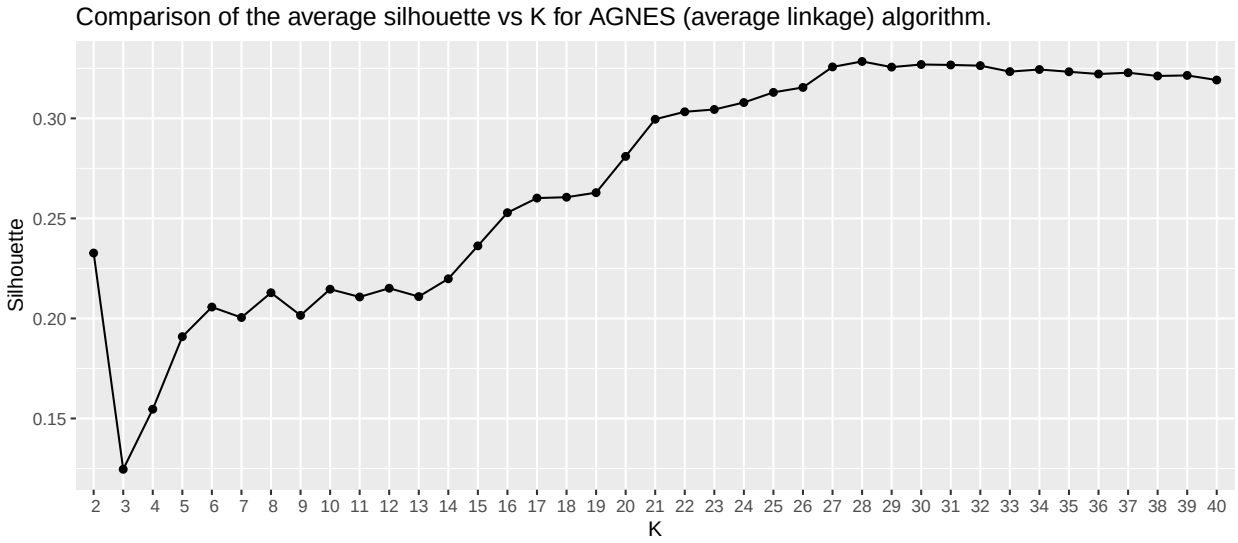

Figure 18: Values of average silhouette width for different number of cluster K in AGNES (average linkage) algorithm.

On the Figure 18 we present average silhouette depending on number of clusters. We can observe that it's from up K equal 3 it's nondecreasing up to around K equal 30. Because of that we selected K = 28 which is maxium in our selected range. Let's find out what values are inside corresponding clusters.

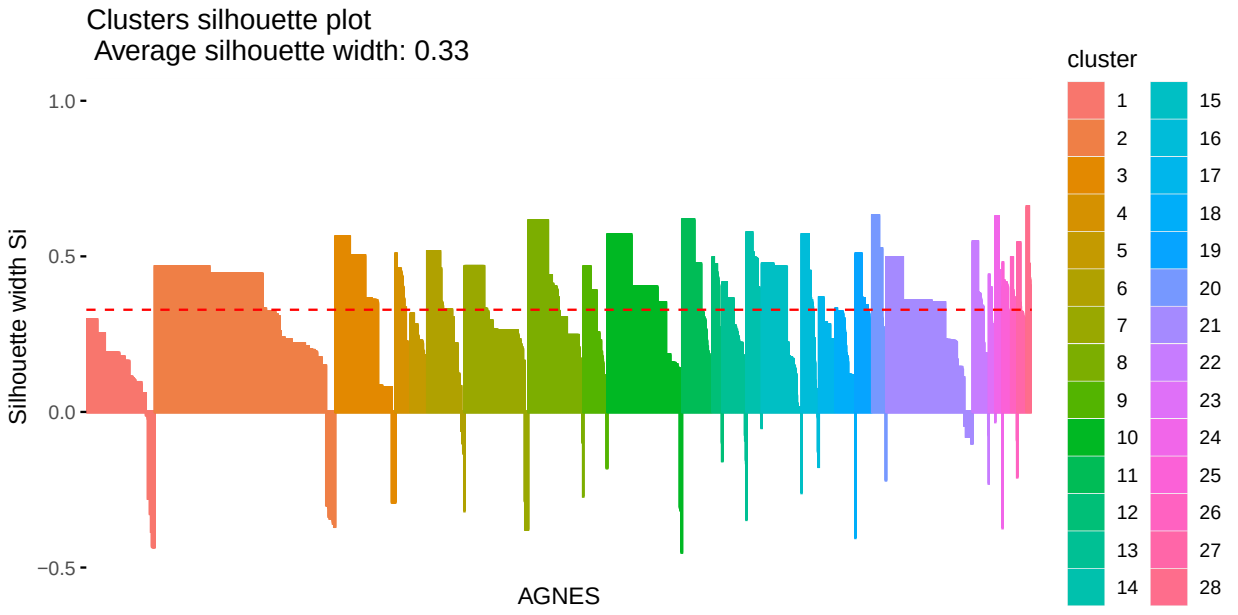

Figure 19: Clusters silhouette plot for AGNES algorithm with K = 28.

From Figure 19 we can observe that most of the clusters has very high value of silhouette and only few of them has some negative values. The results look very interesting and now check how the clusters look inside.

As previously we weren't be able to find groups of clients which were mostly subscribing term deposits. However we can procede with exploration of obtained clusters. Based on cluster assignments and values for corresponding clients' variables we found out following things.

- Cluster 8 represents clients with university degree, single, mostly working in administration, without house and loan.
- Cluster 10 represents clients with university degree, working in administration, with house and without loan.
- Cluster 11 represents clients with basic education, married, working as blue-collars, with loans.
- Other clusters have also interesting representations however provided ones were the most representative.

As we can observe the results are also quite similar to the previously obtained. Because of the higher number of cluster our groups are more granular and consistent which could be very useful for clients profiling.

During this section we've tested a few clustering algorihtm thanks to which we've able to better understand clients group in our dataset. We've found consistent groups of clients, splitted them into clearly separable groups and due to that we've gained better understanding of data structure. This information could be very beneficial for analysts for futher analysis process, e.g. clients segmentation / pricing or for stakeholders which by better understanding of their client base could take further actions related to the whole business.

3 Summary and discussion

During our second part of the journey with Marketing dataset we've gained even better understanding of the whole dataset.

Firstly, we've used dimension reduction techniques to visualize our dataset from different points of view. We were able to find some structure in the dataset, especially in the part connected to the clients where we found some specific groups as for example blue-collar workers, maried and with basic education. However in context of term subscription we weren't able to find such groups.

Secondly, we've prepared classification model for detecting if a new client will be interested in subscription of term deposit. During conducted experiments we've spotted mistake in our previous report which we fixed in this one. We've been testing if PCA transformation could improve results of the classification model however it didn't happened. Moreover we've tested one more time hypothesis if resampling techniques could improve our results but also this technique didn't yield better results. After all our best model was Random Forest which in next step could be improved by hyperparameter tuning.

Lastly, we've performed cluster analysis. We've tested different algorithm, both partitioning and hierarchical ones, which showed us our data from different perspectives. We were able to distinguish many different, interesting groups of clients which alowed us to better understand and profile customers of the bank. Moreover the challenging part was connected to the computations as we needed to sample data to compute all required informations. One of the further steps would be to overcome this limitation.

To conclude, this project showed us how many interesting informations can be found in the dataset and how important is to look into data from different perspectives. Moreover we are now familiar with many different, useful data mining techniques which will be very beneficial for us in the future.

References

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