Bank Marketing Data Project

By Eric Drew

Executive Summary

The bank clients can be described in 5 different dimensions:

- Dimension 1 - deal seekers - This dimension separates clients by interest rate and consumer price index.
- Dimension 2 – repeat customer - This dimension separates clients that have previously purchased the product being marketed vs. people who have not been marketed the product previously.
- Dimension 3 – young starting out savers - This dimension separates clients by young and student vs. older and retired.
- Dimension 4 – old frequent contacts - This dimension separates clients by older clients that have been contacted many times before the current campaign vs. younger clients that have not been contacted many times before the current campaign.
- Dimension 5 - old married retired and poor economy - This dimension separtes clients who are old, married, sometimes divorced, retired, and consumer confidence is low, with clients that are young, single, often students, and consumer confidence index is high.

The clients of the bank can be segmented into three different groups:

- Cluster 1 – high rate no contact - Where interest rates are high and the clients have not been contacted in previous marketing campaigns.
- Cluster 2 – previous customers - the clients that had been contacted in previous campaigns, current interest rate are low, and they had previously purchased the product being marketed.
- Cluster 3 – low rate no contact - Where interest rates are low, the client had not been previously marketed the product, and the client had not purchased the product in previous marketing campaigns.

By segmenting the clients, we have achieved a better understanding of the types of people that are clients at the bank. This allows us to better understand the needs of the current types of clients and to better know the types of people that might join the bank in the future.

Introduction and Goals

This project is based on a dataset containing bank marketing data. It contains information about the demographics of the bank customers and information about the marketing campaign for a term deposit product. The analysis that I will be performing is PCA into clustering. I will first reduce the dimensionality using PCA and then use the new transformed PCA dimensions to perform clustering on the dataset. The goal of the PCA analysis is to reduce the dimensions in the dataset, identify the principal components of the data, and analyze the loadings on those components to identify what the new dimensions represent. The goal of the clustering analysis will be to use the new dimensions from the PCA analysis to identify the major clusters in the data. By doing this, I will be able to segment the bank customers into different groups. This type of analysis will allow the bank to better understand the types of people that are customers at their bank. By better understanding the different types of customers that use the bank, the bank would be able to better target their marketing campaigns to the right groups of customers. Also by understanding the types of people that are customers at the bank they can know the types of people the bank should target to join their bank.

The dataset was originally created with the purpose of creating a model to predict a binary response variable. The binary response variable is whether a customer purchased a term deposit. Both PCA and clustering are unsupervised approaches to data analysis, so in this analysis I am not interested in creating a model to predict the response variable, rather I am interested in reducing the dimensionality in the data, understanding those dimensions, and then using the new dimensions to create a model to cluster the data into different groups. The goal is not to create a classification model for the response variable.

Dataset

The following pertains to the specifics of the dataset and its variable definitions. The dataset comes from a Portuguese banking institution. The dataset was created by the bank to understand how to better market their term deposit product. There are 20 variables in our dataset. There is also a binary response variable, which states whether the customer purchased the term deposit product or not. Each row in the dataset is a client at the bank. There are 41,188 clients in the dataset. There is a mixture of categorical variables and numeric variables. Specifically, there are 10 of each type.

The variables in the dataset can be broken down into 4 categories. The first category is information about the client and demographics. This category contains information about the client's age, job, education, marital status, whether they have a home loan, whether they have a personal loan, and whether they have credit in default.

The second category is related to the current marketing campaign. This category contains information about whether the client was contacted by telephone or cellphone, the month of the last contact with the client, the day of the week of the last contact, and the duration of the last contact with the client.

The third category is more information about the current marketing campaign and information about previous marketing campaigns. This category contains information about the number times the client was contacted during the current campaign, the number of days since the client was last contacted from a previous campaign, the number of times the client was contacted in previous campaigns, and whether the client purchased the product in previous marketing campaigns.

The fourth category is information about the economy when the product was marketed. This category contains information about the employment variation rate, the level of the consumer price index, the level of the consumer confidence index, the euribor 3 month rate, and the number of employees at the bank. In summary, there are 20 variables that describe 41,188 bank clients. The variables describe the demographics of the clients, information about the current marketing campaign, and information about the state of the economy.

Exploration of Dataset

This section is an examination of the specifics of each variable and their distributions. This will help inform us as to how to proceed with the analysis. Age is skewed to the right. The average client is 40 years old. The median client is 38 years old. There are also a number of clients above 75 years old. The marital status variable is broken down into 4 categories: single, married, divorced, or unknown. The majority of clients are married. 60.52% are married, 28.09% are single, 11.2% are divorced, and .19% is unknown.

The majority of client occupations are administrators and blue-collar workers. 25.3% are admins, 22.47% are blue-collar, 3.53% are entrepreneurs, 2.57% are housemaids, 7.1% are management, 4.18% are retired, 3.45% are self-employed, 9.64% are services, 2.12% are students, 16.37% are technicians, 2.46% are unemployed, and .8% is unknown.

Most of the clients are well educated and have a university degree. 10.14% have a basic 4 year education, 5.56% have a basic 6 year education, 14.68% have a basic 9 year education, 23.1% have a high school education, .04% of clients are illiterate, 12.73% have a professional course education, 29.54% have a university degree, and 4.2% is unknown.

Most clients do not have a loan in default. Only .007% has a loan in default and 20.87% is unknown. About half the clients have a housing loan. 52.38% yes, 45.21% do not have a loan, and 2.4% is unknown. Most clients do not have a personal loan. 15.17% have a personal loan, 82.43% do not have a personal loan, and 2.4% is unknown. Most clients are contacted by cellphone. 63.47% Were contacted on a cellphone, and 36.52% were contacted by telephone. The most common months for contacting clients were May, July, and August. Clients were contacted pretty evenly each day of the week.

Clients Are contacted a median of 2 times and a mean of 2.568 times during the current campaign. The max was 56 times. The variables titled p-days contains information about the number of days that passed by after the client was last contacted from a previous campaign. 96.32% of the observations are 999. This is a placeholder for clients that were not contacted in previous campaigns. It can be thought of as representing infinity. This is a problem that will have to be resolved later in the preprocessing step. Of the clients that were contacted in the previous campaign, most clients Are only contacted one or two times. 86.34% of the clients were not contacted in previous marketing campaigns.

The mean employment variation rate is .082 and the range is -3.4 – 1.5. The distribution is not normal, there are regions of higher density, in the -2 - -1.5 range, and the 1 – 1.4 range. The consumer price index ranges from 92.2 to 94.77 and the mean is 93.58. The consumer confidence index ranges from -50.8 to -26.9 with a mean of -40.5. The euribor rate ranges from 5.045 to .634 and the mean is 3.621. The number of bank employees ranges from 4964 to 5228 with a mean of 5167. Overall there is a pretty good distribution of varying economic states and varying interest rate environments in the data.

Preprocessing

Now that the variables have been identified and explored, the next step is to preprocess the data. First we have to do what's recommended in the readme file that is associated with the dataset. It is recommended that the duration variable be removed because the variable has a very high correlation on the target variable and it is not known before the marketing call is made. This makes sense we don't know beforehand how long the client will engage with the marketing call beforehand and the longer that they end up on the phone, the more likely they are to purchase the product. So I removed that variable.

The next step is to transform the P-days variable. P-days is defined as containing information about the number of days that passed after the client was last contacted from a previous campaign. The variable needs to be transformed because 96.32% of the values are 999, which is a placeholder for infinity. Most of the clients were not contacted in the previous campaign so there is no value for the number of days that passed after the client was last contacted from a previous campaign. I decided to transform this into a binary variable. 1 indicates that the client was contacted in a previous campaign and 0 indicates that the client was not contact in a previous campaign.

The next step is to deal with missing values. There aren't any missing values in our dataset but there are a number of variables that had values that were unknown. The variables that contained unknown values were marital status .194%, job .8%, education 4.2%, default 20.87%, housing 2.4%, and personal loan 2.4% unknown. To replace the unknown values I created classification models for each variable. For the personal loan variable, I created a logistic regression the used job, employment variation rate, consumer price index, and consumer confidence index to predict whether the client had personal loan.

For the housing loan variable, I created a logistic regression to classify whether the client had a housing loan. The variables used in the logistic regression were job, loan, month, day of the week, employment variation rate, consumer price index, consumer confidence index, euribor rate, and number of employees.

I used LDA to predict education, job, and marital status. Now that I have created classification models for the variables that contained unknown values, I can use the models to predict those values for the clients and replace all the unknown variables with our predictions. If I did not do this step about 20% of the rows would contain an unknown value.

PCA – Analysis

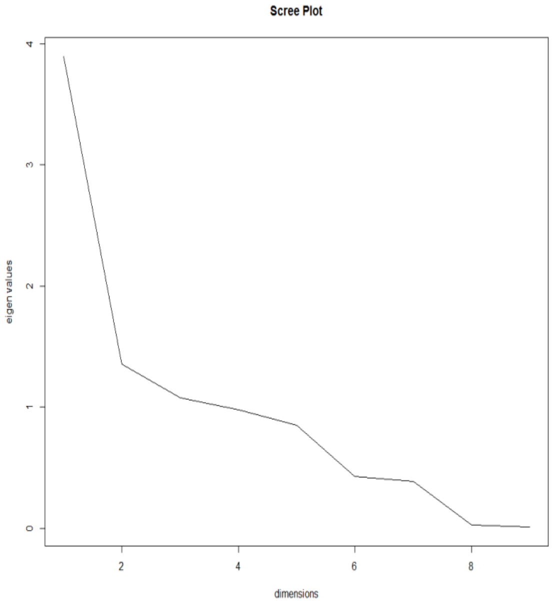
I have now explored, preprocessed, and transformed the necessary variables. I now have a dataset that I can use to perform principal component analysis (PCA). I used the whole dataset excluding the binary response variable to perform the PCA analysis.

PCA does not work with categorical variables, and about half of the variables are categorical. Therefore, I will perform PCA on the numeric variables to extract the new features and then I will use the categorical variables as supplementary variables.

The first step in PCA is to calculate the transformed PCA dimensions. The variables are automatically scaled when the PCA function is applied to the dataset. There are 9 numeric variables so there will be 9 new PCA dimensions. As I would expect, the first dimension captures the most variance at 43.262%. Also, as I would expect, the 9 PCA components account for 100% of the cumulative variance.

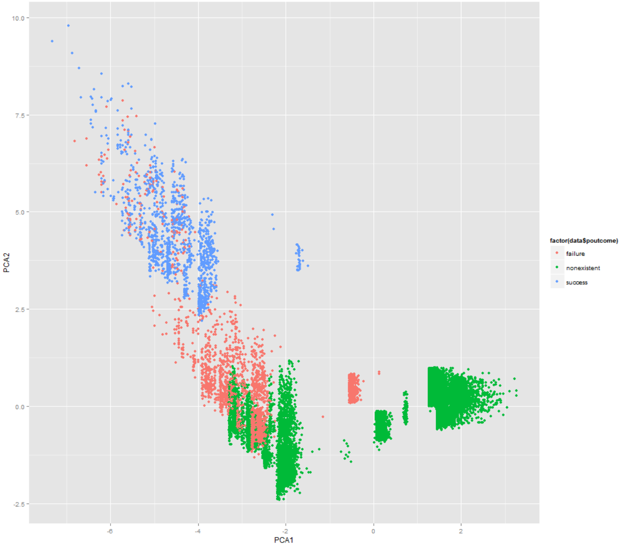
	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
Variance	3.894	1.357	1.078	.976	.849	.426	.0387	.025	.011
% of var.	43.262	15.072	11.974	10.841	9.429	4.732	4.295	.2277	0.118
Cumulative var.	43.262	58.334	70.308	81.49	90.579	95.310	99.606	99.882	100.00

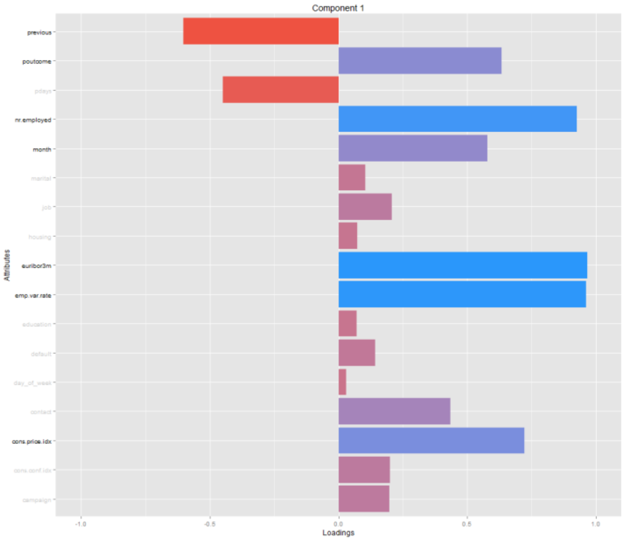
Now that I have the new PCA components from the numeric variables, I looked at the scree plot to decide how many PCA dimensions should be used. The cutoff was the "knee" in the scree plot. By looking at the scree plot I decided that 5 dimensions was the appropriate number to retain because with 5 dimensions it captures 90.579% of the variance in the dataset.



Now that I have decided to use 5 dimensions from the transformed PCA matrix, I can investigate the loadings on those dimensions to understand what the 5 new dimensions are representing.

PCA dimension 1 captures 43.262% of the variance. The variables with strong loadings are previous, p-outcome, number of employees, month, euribor, employment variation rate, and consumer price index.

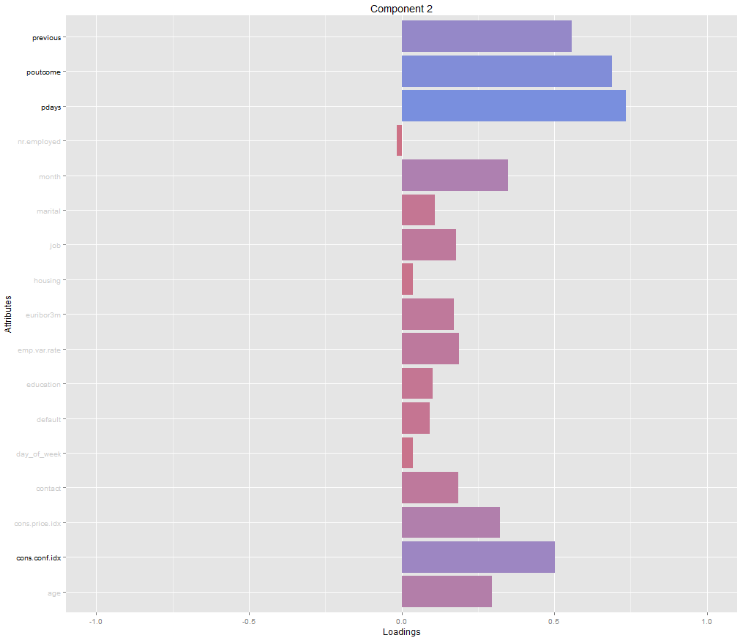




P-outcome was a categorical variable so it was not included in the calculation of the PCA dimensions but by looking at the graph above and on the left I can see that there is separation between the categories of p-outcome in dimension 1. The x-axis in the graph is dimension 1 and the y-axis is dimension 2 which also separates the categories of p-outcome. From the graph you can see that high values for dimension 1 is often associated with clients that have a non-existent outcome in previous marketing campaigns. Low values in dimension 1 are often associated with a successful outcome in previous marketing campaigns. This means that the client purchased the product in a previous campaign.

By further investigating the loadings for the numeric variables, I can conclude that high dimension 1 values are associated with zero contacts with client before this campaign, high number of bank employees, the month of June, high euribor rate and consumer price index, and high employment variation rate.

Based on this information I have concluded that the title for dimension 1 should be, the deal seekers dimension. These are clients that have not been contacted previously, have not previously been sold the product, and euribor rates and consumer price index is high. If interest rates are high then the product would be more appealing because the interest rate would be higher on the term deposit product.



PCA dimension 2 captures 15.072% of the variation in the data. The variables with strong loadings are previous, p-outcome, p-days, and consumer confidence index. Again the we see the p-outcome variable which is categorical and was not used to compute the PCA dimensions. However, from the p-outcome graph was saw when discussing dimension 1, I can see that dimension 2 also separates the client by the categories of p-outcome. High values of dimension 2 are associated with the previous marketing campaign outcome being a success. Low values of dimension 2 are associated with non-existent outcomes in the previous marketing campaigns.

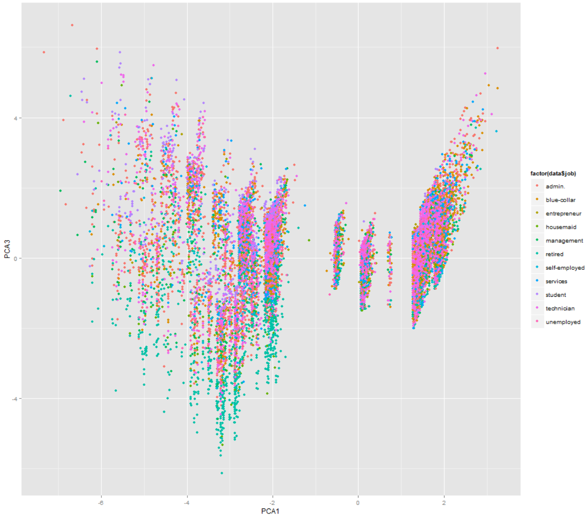
Further investigation of the numeric variables with strong loadings on dimension 2 reveal that high values of dimension 2 are associated with high number of times contacted in previous marketing campaigns, a successful sale in previous marketing campaigns, clients that were contacted in previous campaigns, and a low consumer confidence index.

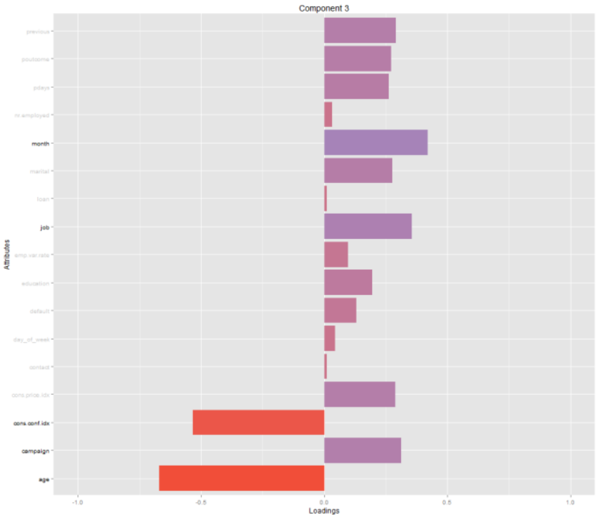
From this information I have concluded that dimension 2 should be called, the repeat customer dimension. This dimension separated clients that have previously purchased the product being marketed vs. people who have not been marketed the product previously.

PCA dimension 3 captures 11.974% of the variance in the data. The variables with strong loadings are month, job, consumer confidence index, campaign, and age. Job is a categorical variable. By looking at the graph below, I can see that dimension 3 is mostly separating jobs by student and retired.

When the value of dimension 3 is high, the job category is often student. In the graph below, dimension 3 is the y-axis and dimension 1 is the x-axis.

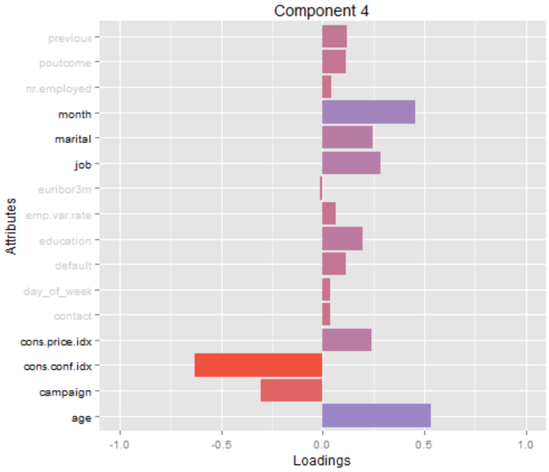
The numeric variable loadings reveal that clients that score high in dimension 3 have been contacted before the current campaign, and are young in age. This makes sense since they are also often students. Based on this information, I would call this dimension, the young starting savers. The dimension separates clients by old vs young and student vs retired.

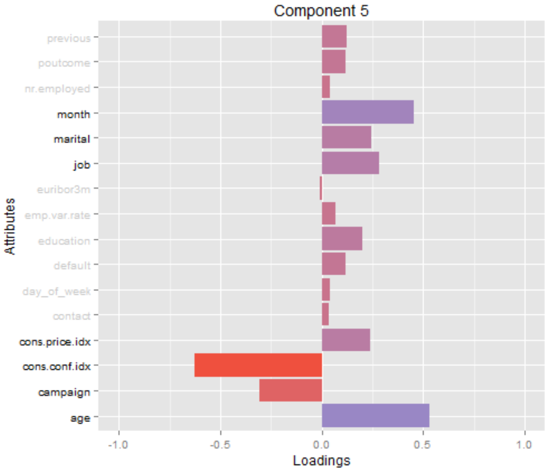


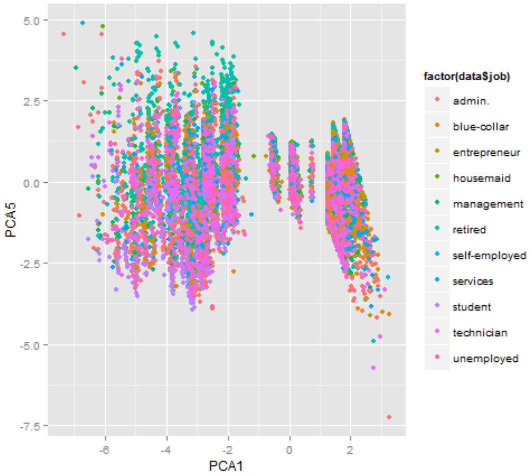


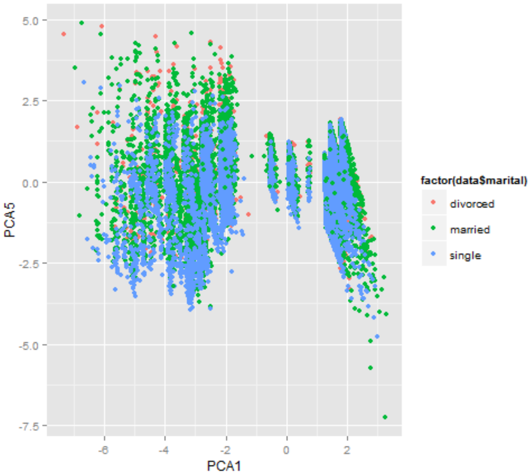
PCA dimension 4 captures 10.841% of the variance. The variables with strong loadings are campaign and age. Both of these variables are numeric. Further investigation reveals that, high values of dimension 4 are associated with older age and contacted many times before the current marketing campaign. Based on this information, I would call this dimension the older frequent contacts. This dimension is separating clients that are older and have been contacted many times before the current campaign vs. younger clients that have not been contacted many times before the current campaign.

PCA dimension 5 captures 9.429% of the variance in the data. The variables with strong loadings are month, marital, job, consumer price index, consumer confidence index, campaign, and age. Marital and job are categorical variables. From the graphs below, I can see that dimension 5, which is the y-axis on both graphs, shows that high values of dimension 5 are associated with married or divorced, and retired or blue-collar.









The numeric variables in dimension 5 are consumer price index, consumer confidence index, campaign, and age. Further investigation reveals that high values of dimension 5 are associated with high consumer price index, low consumer confidence index, low number of contacts before the current campaign, and older in age.

I would characterize dimension 5 as the old, married, sometimes divorced, retired, clients when consumer confidence is low vs. the young, single, often student, and consumer confidence index high. I would call this dimension, the old, married, retired, and poor economy dimension.

PCA - Technical Summary

The following is a summary of the PCA analysis. The first step was to transform the data into the new transformed PCA dimensions. The PCA was only performed on the nine numeric variables and not the categorical variables. This resulted in 9 new PCA dimensions. By looking at the scree plot, I selected 5 dimensions to be of significance. These 5 dimensions explained 90.579% of the variance in the data. So I did lose about 10% of the variance in the data but I reduced the dimensionality of the data from 9 to 5 dimensions.

The next step was to investigate the loadings of the variables on the 5 components. Although the categorical variables were not used in the calculation of the PCA dimensions, I still can identify if the categorizations explain the PCA dimensions by looking at the categorization of the dimensions by the categorical variables.

I found that dimension 1 could be described as the 'deal seeker' dimension. This dimension separated clients by interest rate and consumer price index. Dimension 2 was the 'repeat customer' dimension. I saw that dimension 2 clients were separated by whether or not the client had purchased the product being marketed in previous marketing campaigns. Dimension 3 was the 'young starting out savers' dimension. This dimension separated clients by young and student vs. older and retired. Dimension 4 was described as the 'older frequent contacts' dimension. This dimension separated clients by older clients that have been contacted many times before the current campaign vs. younger clients that have not been contacted many times before the current campaign. Lastly, dimension 5 is the 'old married retired and poor economy' dimension. This dimension separtes clients who are old, married, sometimes divorced, retired, and consumer confidence is low, with clients that are the young, single, often student, and consumer confidence index high. Overall I extracted 5 new dimensions from the data, and by looking at the loadings of the variables on those new dimensions I was able to define what the new dimensions represented.

PCA – Non-Technical Summary

The goal of PCA was to reduce the dimensionality of the data. I was trying to use less dimensions than our dataset has while still preserving as much of the variance as possible in the data. To do this, I identified a vector that explains as much of the variance as possible and then found subsequent vectors, which are orthogonal to the previous vector and explain less and less of the data. I was trying to find new dimensions that explain as much of the variance as possible in fewer dimensions then the number of original dimensions in the data. This results in a new matrix with transformed values. I was not just deleting dimensions from the data, but instead I was extracting new dimensions from the data.

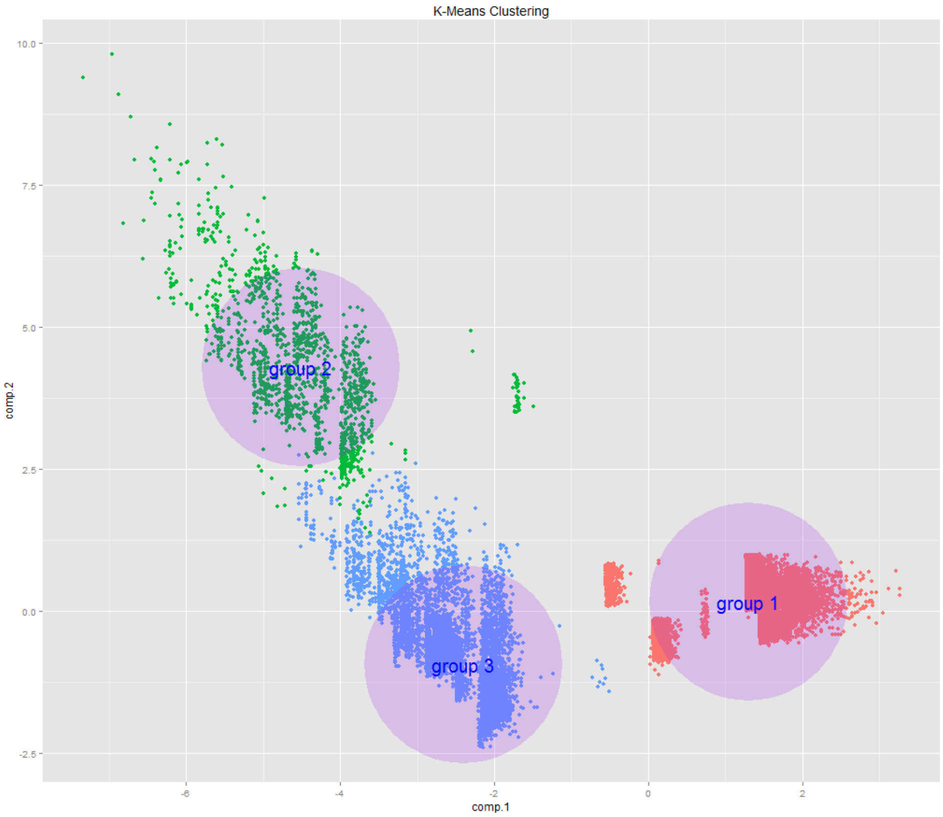
Once I calculated the new dimensions, I was able to identify that 4 of the dimensions can be removed because they only explain a small amount of the variance in the data. That leaves us with 5 dimensions. I investigated these new dimensions and looked at which variables have the most influence on the new dimensions to identify what these new dimensions represent.

I found that dimension 1 could be described as the 'deal seeker' dimension. This dimension separated clients by interest rate and consumer price index. Dimension 2 was the 'repeat customer' dimension. I saw that dimension 2 clients were separated by whether or not the client had purchased the product being marketed in previous marketing campaigns. Dimension 3 was the 'young starting out savers' dimension. This dimension separated clients by young and student vs. older and retired. Dimension 4 was described as the 'older frequent contacts' dimension. This dimension separated clients by older clients that have been contacted many times before the current campaign vs. younger clients that have not been contacted many times before the current campaign. Lastly, dimension 5 is the 'old married retired and poor economy' dimension. This dimension separtes clients who are old, married, sometimes divorced, retired, and consumer confidence is low, with clients that are the young, single, often student, and consumer confidence index high. Overall I extracted 5 new dimensions from the data, and by looking at the loadings of the variables on those new dimensions I was able to define what the new dimensions represented.

K-Means Clustering - Analysis

Now that I had the new transformed matrix from PCA, where there are 41,188 rows and 5 columns, I used these new dimensions to perform k-means clustering. The goal of k-means clustering is to identify the distinct groups of bank clients within the dataset.

The first step in k-means is to decide which value of k to select. I tried a few different values for k but it seems pretty obvious, by visually inspecting the data, that there are three main clusters in the data. By looking at the graph below, where the x-axis is component 1 and the y-axis is component 2, I can see separation of the three groups in the first two components. Group 1 is the red values in the bottom right of the graph. Group 2 is the green values in the top left of the graph. Group 3 is the blue values in the bottom middle of the graph. Visually, I can see that there are three distinct clusters and the clusters have pretty good separation between the groups. The centers of the purple circles are the centers for each cluster.



	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
Group 1	1.29	.16	-.02	-.05	-.02
Group 2	-4.48	4.23	1.31	-.22	.07
Group 3	-2.4	-0.95	-.13	.15	.04

To identify what the three clusters represent, I looked at where the cluster centers were positioned in each of the 5 dimensions. Above is a table of the positions of the cluster centers in each of the 5 dimensions. Group 1's mean is high in dimension 1. I can describe this cluster as the cluster where interest rates are high and the clients have not been contacted in previous marketing campaigns.

Group 2 scored low in dimension 1, high in dimension 2, and somewhat high in dimension 3. Scoring low in dimension 1 means that interest rates are low and the clients have been previously contacted in previous marketing campaigns. A high score in dimension 2 means that the clients in this cluster had been contacted multiple times in previous marketing campaigns and had previously purchased the product being marketed. A high score in dimension 3 means that the clients in this cluster are younger and have been contacted in the previous campaign. From this information I concluded that group 2 represents clients that had been contacted in previous campaigns, current interest rates were low, and they had previously purchased the product being marketed.

Group 3's mean scored low in dimension 1 and scored low in dimension 2. Scoring low in dimension 1 means that interest rates are low and the clients have been previously contacted in previous marketing campaigns. A low score in dimension 2 means that the clients were previously marketed the product and they had not previously purchased the product. Based on this information I concluded that group 3 is where interest rates are low, the client had not been previously been marketed the product, and the client had not purchased the product in previous marketing campaigns.

K-Means Clustering – Technical Summary

Using the 5 new dimensions from the PCA analysis, I performed k-means clustering on the data. I decided that the data could be best segmented into three different groups. The centers of the groups are the mean values of the clients that are segmented in that group. I used the centers of the clusters to identify what the groups represent. I determined that group 1 is the cluster where interest rates are high and the clients have not been contacted in previous marketing campaigns. Group 2 is the clients that had been contacted in previous campaigns, current interest rate were low, and they had previously purchased the product being marketed. Group 3 is where interest rates are low, the client had not been previously marketed the product, and the client had not purchased the product in previous marketing campaigns.

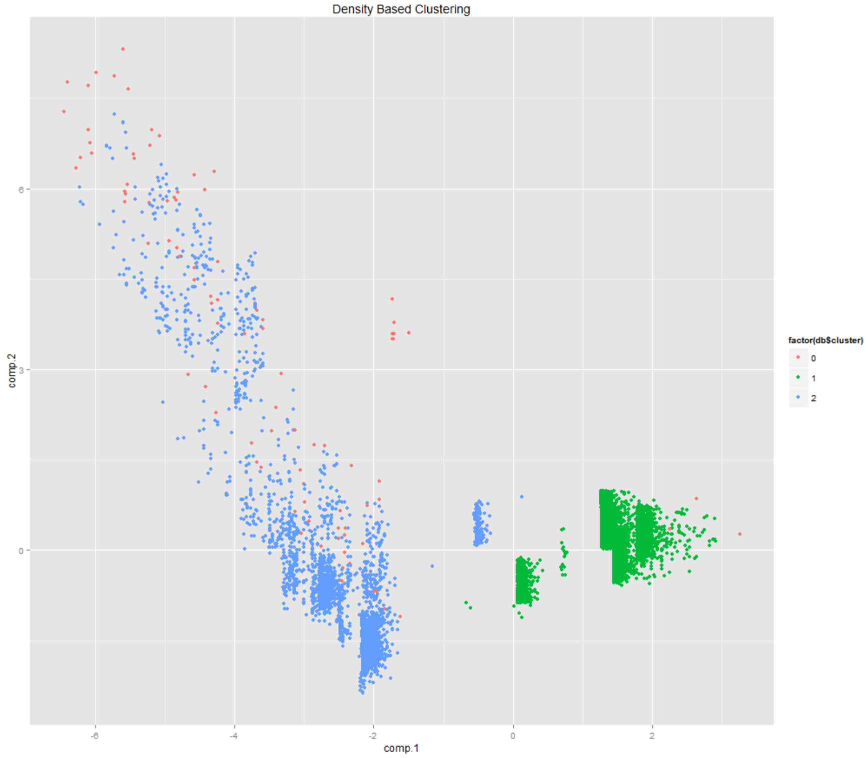
K-Means Clustering – Non-Technical Summary

Using our newly transformed dataset, I segmented the clients into separate groupings. From the analysis, I discovered that there are three distinct segments of clients. I used the knowledge of the results from our PCA analysis to come to a conclusion about what the three client segments represent. I determined that group 1 was the cluster where interest rates are high and the clients have not been contacted in previous marketing campaigns. Group 2 was the clients that had been contacted in previous campaigns, current interest rate were low, and they had previously purchased the product being marketed. Group 3 was where interest rates are low, the client had not been previously marketed the product, and the client had not purchased the product in previous marketing campaigns. This analysis has allowed us to understand what types of people are clients at the bank. In this case, we can segment the client into three distinct groups.

Density Based Clustering – Analysis

K-Means clustering is just one type of clustering that I can perform on our data; I can also perform density based clustering. Specifically, I can perform a density based clustering method called DB-scan. The most difficult part of this application of this type of clustering is to determine the correct input parameters for the dataset because the results are very sensitive to the inputs. After much experimentation, I determined that the appropriate reachability distance (eps) was 1.27 and the most appropriate minimum number of points was 6. It should be noted that I had to use a random subset of the data, 10,000 values, because my computer ran out of memory if I used the entire dataset. In the graph below, I can see that there are two clusters. Group 1 is the blue points and group 2 is the green points. The red values are noise. Similar to K-Means clustering, I can calculate the mean values for clients in each group and look at where those means are positioned in each dimension of the data. Below is the table of mean values for the two groups in each dimension.

	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
Group 1	-2.52	-.39	.01	.08	.02
Group 2	1.34	.15	-.01	-.06	-.04



I can see that group 1 scored low in dimension 1 and somewhat low dimension 2. By looking back at our PCA analysis I can see that scoring low in dimension 1 means that interest rates are low and the clients have been previously contacted in previous marketing campaigns. A low score in dimension 2 means that the clients were previously marketed the product and they had not previously purchased the product. From this information, I concluded that this group is clients that sometimes were contacted in previous marketing campaigns and current interest rates are low.

By looking at the mean values of group 2, I can see that group 2 scored high in dimension 1. A high score in dimension 1 means that these clients have not been contacted previously, have not previously been sold the product, and interest rate are currently high. From this information, I concluded that group 2 represents clients that have not been contacted previously and have not previously been sold the product.

Density Based Clustering – Technical Summary

The results of Db-scan are very dependent on the input parameters. The two parameters are EPS and min-points. The EPS is maximum radius of the neighborhood and min-points are the minimum number of points that have to be in an EPS neighborhood. Changing these variables a small amount can result in very different results for the clustering. After experimenting with different parameters, I found that an eps of 1.27 and a min-points of 6 work best for the data. I had to use a random subset of 10,000 clients because otherwise my computer would run out of memory. The clusters are based on the density of the points.

This resulted in 2 clusters. The first cluster, based on the center of the cluster, can be described as representing clients that sometimes were contacted in previous marketing campaigns and current interest rates were low. The second cluster, based on the center of the cluster, represents clients that have not been contacted previously and have not previously been sold the product.

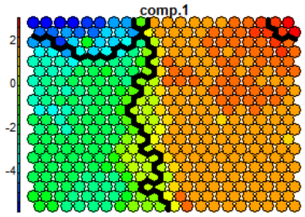
Density Based Clustering – Non-Technical Summary

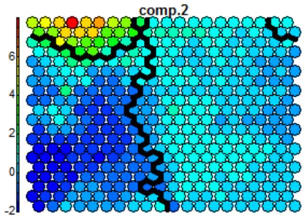
Density based clustering creates clusters based on the density of the observations. Its advantage over k-means clustering is that it can discover arbitrary cluster shapes and can handle noisy outliers. After clustering the points based on their density, the method found there are two clusters in the data. The two clusters can be defined as follows, The first cluster, based on the center of the cluster, can be described as representing clients that sometimes were contacted in previous marketing campaigns and current interest rates were low. The second cluster, based on the center of the cluster, represents clients that have not been contacted previously and have not previously been sold the product.

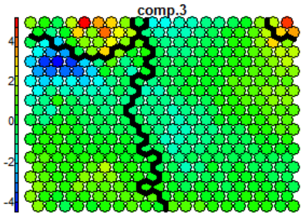
Self-organizing Maps – Analysis

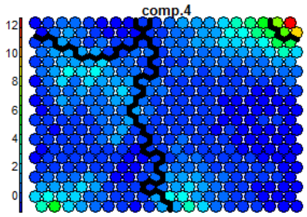
Another alternative way to cluster the transformed PCA dimensions is with a self-organizing map (SOM). SOM is way to transform the data into a low dimensional space. It also helps to see the relationship of the clusters among the dimensions and the subtleties among the groups.

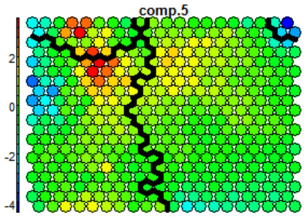
The first step is to train the model. The main argument is the number of nodes to train, in this case I chose 20 x 20 nodes. The main factor in deciding how many nodes to train is making sure that the number of samples mapped to each node is relatively evenly distributed. I can then perform hierarchical clustering on the nodes to determine which nodes are most similar. From there I can perform clustering on the nodes based on the hierarchical clustering. In this case, I found that 4 clusters were the most appropriate because if there are too many clusters then some of the clusters are very small in size compared to the other clusters. 4 clusters seem to give a good distribution of nodes among the clusters. The bottom right graph below shows how the nodes were clustered. Cluster 1 is the cluster in the top-left quadrant, cluster 2 is in the top-right quadrant, cluster 3 is in the left quadrant, and cluster 4 is in the right quadrant.

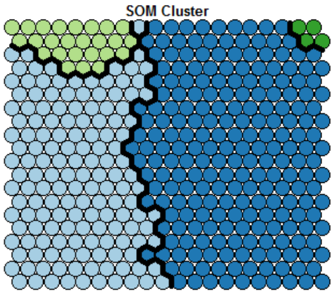












The 5 graphs above show the self-organizing map for each of the 5 dimension of the data and the cluster divisions. They are basically heat-maps for each cluster in the 5 dimensions. So for example, I can see that that cluster 1 had low values for component 1, high values in component 2 and 3, and low values in component 4.

By investigating the nodes in each clusters and looking their relationship to the 5 dimensions of the data, I can come to a conclusion about what the nodes in each cluster represent. Cluster 1 has low values in component 1. Low values for component 1 means that the clients have been contacted previously, have previously been sold the product, and euribor rates and consumer price index is low. Cluster 1 is also high in component 2. High values of component 2 are associated with a high number of times contacted in previous marketing campaigns, a successful sale in previous marketing campaigns, clients that were contacted in previous campaigns, and a low consumer confidence index. Cluster 1 is also high in component 3. High values in component 3 are associated with clients that have been contacted before the current campaign, and are young in age. So to summarize, cluster 1 is clients that have been contacted previously, previously been sold the product, and interest rate are low.

Cluster 2 is high in component 1. High values for component 1 means that the clients have not been contacted previously, have not previously been sold the product, and euribor rates and consumer price index is high. Cluster 2 is high in component 3. High values in component 3 are associated with clients that have been contacted before the current campaign, and are young in age. Cluster 2 is high in component 4. High values of component 4 are associated with older age and contacted many times before the current marketing campaign. Cluster 2 is low in component 5. Low values of component 5 are young, single, often student, and consumer confidence index high. To summarize, cluster 2 is clients that are young, single, often students, and have not previously been sold the product.

Cluster 3 is low in component 2. Low values of component 2 are associated with a low number of times contacted in previous marketing campaigns, a non-existent sale in previous marketing campaigns, clients that were not contacted in previous campaigns, and a high consumer confidence index. To summarize cluster 3 is the clients that have not been previously marketed to and therefore the previous outcome is non-existent.

Cluster 4 is high in component 1. High values for component 1 means that the clients have not been contacted previously, have not previously been sold the product, and euribor rates and consumer price index is high. To summarize cluster 4 is the clients that have not been previously sold the product and current interest rates are high.

Self-organizing Maps – Technical Summary

A self-organizing map creates a heat map for each dimension of the data. In this case the SOM uses 20 x 20 nodes. That number was chosen because it gave an even distribution of samples among the nodes. The nodes can then be clustered using hierarchical clustering. From the hierarchical clustering 4 clusters of nodes was chosen because it was the best representation of the data. By analyzing the values of the nodes in each cluster, I defined what the clusters represent.

Based on the SOM I concluded that cluster 1 is clients that have been contacted previously, previously been sold the product, and interest rate are low. Cluster 2 is clients that are young, single, often students, and have not previously been sold the product. Cluster 3 is the clients that have not been previously marketed to and therefore the previous outcome is non-existent. Cluster 4 is the clients that have not been previously sold the product and current interest rates are high.

Self-organizing Maps – Non-Technical Summary

A SOM is way of creating a heat map for each dimension of the data. It creates a bunch of nodes that are weighted in each dimension in a way that tries to best describe the data. I can then identify clusters in the map based on nodes that are similar to one and other. Based on the analysis, I concluded that 4 clusters was the correct number of clusters for this data. I can then look at the clusters and the heat maps for each dimension and define what the clusters represent.

Based on the SOM I concluded that cluster 1 is clients that have been contacted previously, previously been sold the product, and interest rate are low. Cluster 2 is clients that are young, single, often students, and have not previously been sold the product. Cluster 3 is the clients that have not been previously marketed to and therefore the previous outcome is non-existent. Cluster 4 is the clients that have not been previously sold the product and current interest rates are high.

Summary and Conclusions

After preprocessing and exploring the data, the first major step in the analysis was to perform PCA. PCA is a way of reducing the dimensionality of the data. Based on the analysis, I concluded that 5 PCA dimensions was the appropriate number of dimensions to retain from the 9 total PCA dimensions. The PCA dimensions were calculated using the numeric variables in that data and excluded the categorical variables.

I then analyzed the loadings on each dimension to define what the dimension represents.

- Dimension 1 - deal seekers - This dimension separated clients by interest rate and consumer price index.
- Dimension 2 – repeat customer - This dimension is separated by clients that have previously purchased the product being marketed vs. people who have not been marketed the product previously.
- Dimension 3 – young starting out savers - This dimension separated clients by young and student vs. older and retired.
- Dimension 4 – old frequent contacts - This dimension separated clients by older clients that have been contacted many times before the current campaign vs. younger clients that have not been contacted many times before the current campaign.
- Dimension 5 - old married retired and poor economy - This dimension separtes clients who are old, married, sometimes divorced, retired, and consumer confidence is low, with clients that are the young, single, often student, and consumer confidence index high.

I found that the K-means did the best job of partitioning the clients into separate groups.

- Cluster 1 – high rate no contact - Where interest rates are high and the clients have not been contacted in previous marketing campaigns.
- Cluster 2 – previous customers - the clients that had been contacted in previous campaigns, current interest rate were low, and they had previously purchased the product being marketed.
- Cluster 3 – low rate no contact - Where interest rates are low, the client had not been previously marketed the product, and the client had not purchased the product in previous marketing campaigns.

From the data I extracted new dimensions using principal component analysis. I then performed three different kinds of clustering on the transformed data. Each clustering method resulted in slightly different clusters. I found that the clustering method that produced the best clusters was k-means. K-means produced 3 clusters of clients. By analyzing the cluster centers I defined what the clusters represent. Based on this information the bank can cluster their clients into the three clusters that I defined. By segmenting the clients, I have achieved a better understanding of the types of people that are clients at the bank. This allows the bank to better understand the needs of the current clients and to better know the types of people that might join the bank in the future.