

BA 820 - Group Project

*Zhaoying Chen (zychen96), Hui Jiang (hjiang97), Yiyang Wang (wangy97), Rui Xu (xurr),
Tyler McMurray (tjm), Yanyue Fu (yanyuefu)*

12/9/2019

Who Are We & What Are We Doing?

Our group works for a financial services company. We have collected data on our customers. Our data is basic information of demographics including variables like age, marital status, job, and education. We also have information regarding their relationship to the bank like their balance at the bank and loan status if they have some or not. Finally, we have information related to our last marketing campaign where the end goal was to have them open up a term deposit. There are multiple uses with this dataset. We can gain insight and benefits through multiple analyses that we will conduct. We will utilize exploratory data analysis, cluster analysis, and supervised learning to benefit our company.

About Our Data

As discussed above we have multiple different variables we want to utilize in our analysis. Some of the variables are categorical variables but unfortunately some of the methods we want to utilize requires only using numerical variables; instead of deleting useful information we encoded them as dummy variables. After converting our numerical values, the amount of variables changed to 29 in total, due to the dummy variables increasing the amount. We discussed utilizing a Principal Component Analysis, we will abbreviate it to PCA from here on out, to deal with the increase as we can utilize this method for dimension reduction. However, after further consideration we realized utilizing a PCA to reduce the dimensions is inherently flawed as the binary variables are not true continuous variables like those needed for the PCA to be fully effective.

We also used an Exploratory Factor Analysis(EFA) method to our dataset. In the beginning, the factor analysis (EFA) failed for the dataset because there are too many dummy variables that identify the jobs and marital status to give us a NA p-value when we use the Bartlett's test for correlation accuracy. We need to have a good dataset for EFA, so I replaced the job column with the numeric value, such as a rating for different jobs by their average balance value in the dataset. Values in the marital column also replaced by 1, 2 and 3 [1: single, 2: married, 3:divorced].

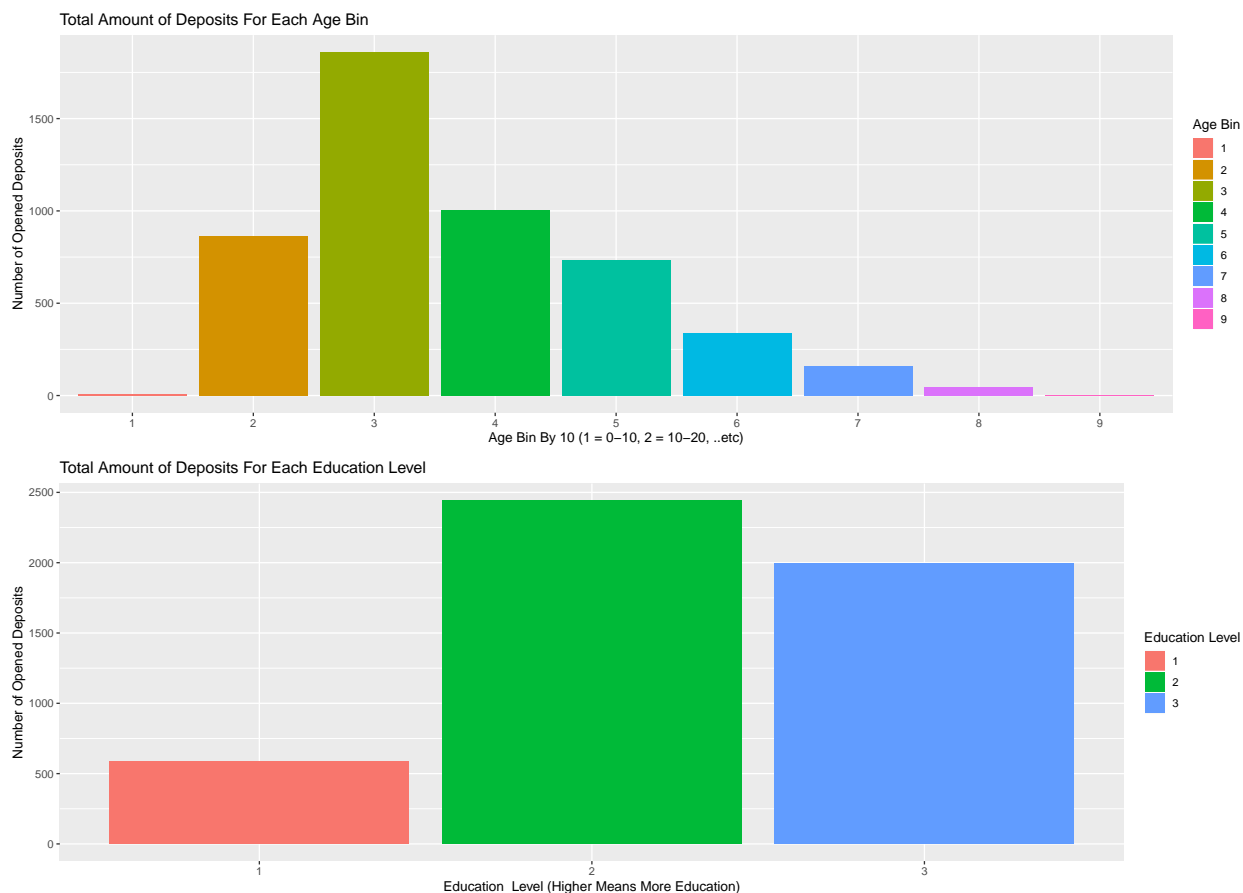
The EFA model works for the replaced dataset with a Bartlett's test p-value equal to 0. After adjusting some variables. The outcome shows that these variables could be divided into 5 factors. The deposit and duration variable could be named as the phone-call effect. Age and marital could be named as the client background. The pdays and previous variable could be named as history record. The loan and housing variable in two factors, but could see them as the financial background. The four potential patterns are affecting clients' behavior, we could take them into consideration for bank campaign designing.

We also realize we have 10634 for 26, so we do not think this level of dimensionality would cause serious problems. For the rest of our analysis we did not apply any of the mentioned dimension reduction techniques like EFA and PCA.

Exploring Our Data

We wanted to get a better understanding of our data and gain some insight through exploring our data. We believe some of this might help guide our understanding and analysis in the future. One insight we gained was that most of the deposits are from the phone calls smaller than 2 phone calls. This could mean after

a certain point it might be a waste of resources to keep trying to get to open a term deposit .Also most of the deposits are from the duration smaller than 2000 secs for the last phone call. If it goes above this time it could mean either our staff is taking too long or that they may not be as able to clearly communicate and need some more help with training. People who have an education level of 2 or 3 are most likely to deposit, which in laymen terms is completing high school and completing a higher education or trade school, respectively for 2 and 3. People who have an age from 30-50 are more likely to have a deposit than other age groups, especially the people who are and between the ages of 30-40. Depending on the job plots, it is very similar that the number of job who do the deposit and also the balance they deposit.



Cluster Analysis

We tried two methods for segmenting our data into groups. Those methods were the K-Means and hierarchical clustering. To determine the amount of clusters we utilized both the silhouette and WSS methods to help us better determine the amount of clusters to for K-Means and the cuts for the hierarchical clustering. The amount was between 2 and 4 clusters. We will try 2, 3, and then 4 clusters for our data through the two methods we will be applying to our data.

K-Means Clustering Analysis

We ran all three of those options for the amount of clusters, meaning using 2, 3, and then 4 clusters. We found that utilizing two clusters gave us the best silhouette width at .84 which was .1 higher than then the next closest option. We interpret a higher silhouette width as it means that the observations fall well into

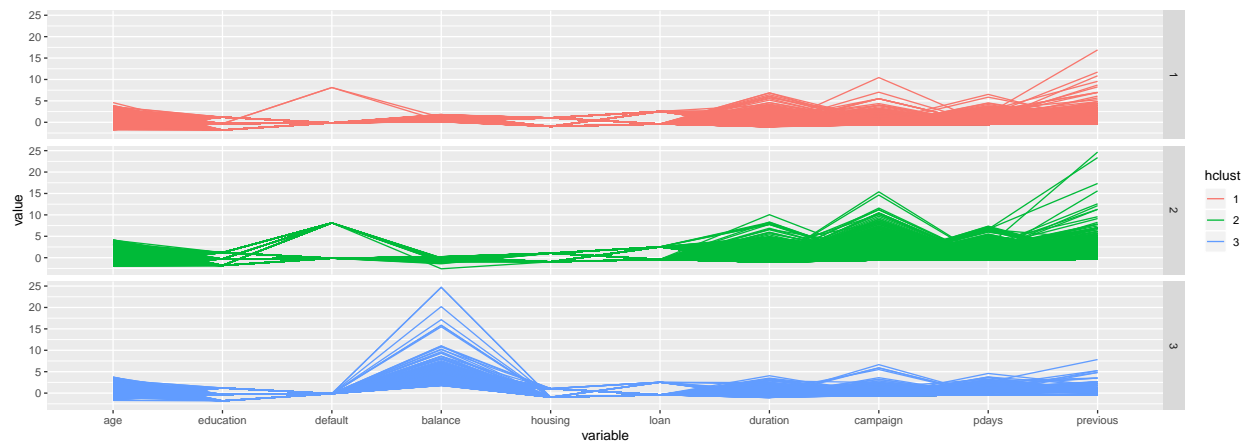
those groups we made. Utilizing two clusters we can see the amount of within each cluster which was 10,207 observations in one cluster and 427 in the other cluster.

Hierarchical Clustering Analysis

We again utilized the same amount of clusters on this graph by cutting the dendrogram tree by 2, 3, and then 4 clusters. However, unlike K-Means clustering, Hierarchical clustering takes a bit more of an interpretation to what is the right amount of clusters. We decided that 3 cuts were the best decision as it small enough amount that our marketing team can still handle. There are also enough clusters that we can have a more targeted strategy for our next marketing campaign, by dealing with the groups differently more targeted to their groups likes, desires, and needs so that we are more effective and efficient next time around. The sizes of the groups are more different than with the K-means, for the better. The sizes of the groups are 1765, 8483, and 386 for groups 1, 2, and 3 respectively.

The Groups

The below graphic is broken out by each group. this gives us some insight into the different groups. This only shows some of the data and doesn't have all variables. However, some distinctions that should be noted is that Group 1 and Group 2 are very similar, however, it seems Group 2 has increased values in regards to the last campaign compared to Group 1. A big distinction is that Group 3 has a lot higher balance variables than Group 1 and Group 2. The purpose of this graphic is to get a rough understanding of the groups and to verify that these groups do have differences and we are not just forcing a split in groups when really it should be less groups. We still feel comfortable with having three groups.



Using KNN To Improve Our Next Campaign

We used all of our data for our above cluster analysis. However, we would need new data to use a KNN method. Instead we are going to run another Hierarchical Cluster analysis to create the 3 groups but only using 95% of our data. The remaining 5%, which is 531.7 observations to use for our KNN method. The purpose of this is to show that we can take new observations, like new people utilizing our bank or data we derived from other methods, to assign into groups and prioritize and market different depending on where they now get assigned to. When we do get new data we would use all of the old data, like we did in the above clustering analysis to get groups. However, for this example we need to use some of our data to do the KNN method. What this means is we are inherently losing some of what made our groups our groups in the above analysis, meaning we are losing some information here compared to above. This means our analysis here should be taken more as a way and example of what we can do with future data for our campaigns strategies.