BA 820 - Group Project

Zhaoying Chen (zychen96), Hui Jiang (hjiang97), Yiying Wang (wangy97), Rui Xu (xurr), Tyler Mcmurray (tfm), 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 there 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 and 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 material column also replaced by 1, 2 and 3 [1: single, 2: married, 3:divorced].

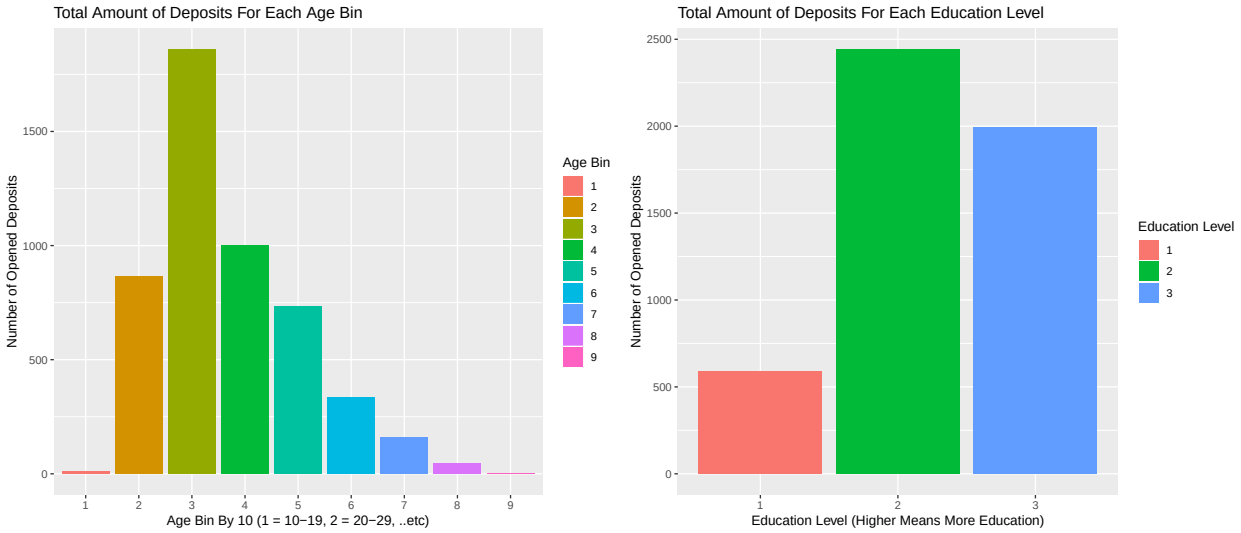
The EFA model works for the replaced dataset with a Bartlett's test p-value equal to 0. After adjusting some variables. The outcome shows that these variables could be divided into 5 factors. The deposit and duration variable could be named as the phone-call effect. Age and marital could be named as the client background. The pdays and pervious 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. Another metric was job, an example of which is someone who job is in management also has the highest number of deposits opened and also the highest balance than people in other jobs. This is important for our marketing team, because maybe they should start advertising to people in those jobs. The two graphics below show the amount of deposits opened by Age groups and then Education levels.



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.

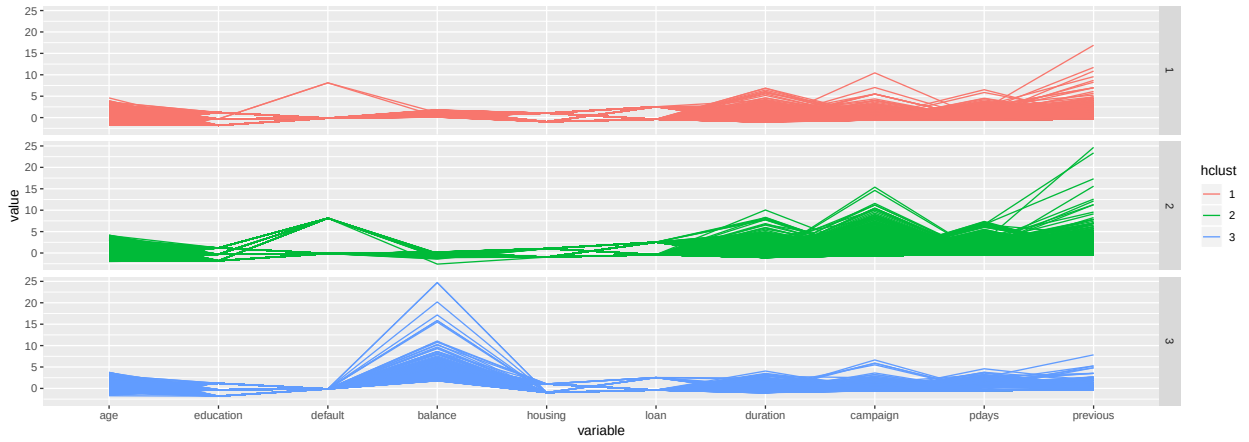
Hierarchical Clustering Analysis

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 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.



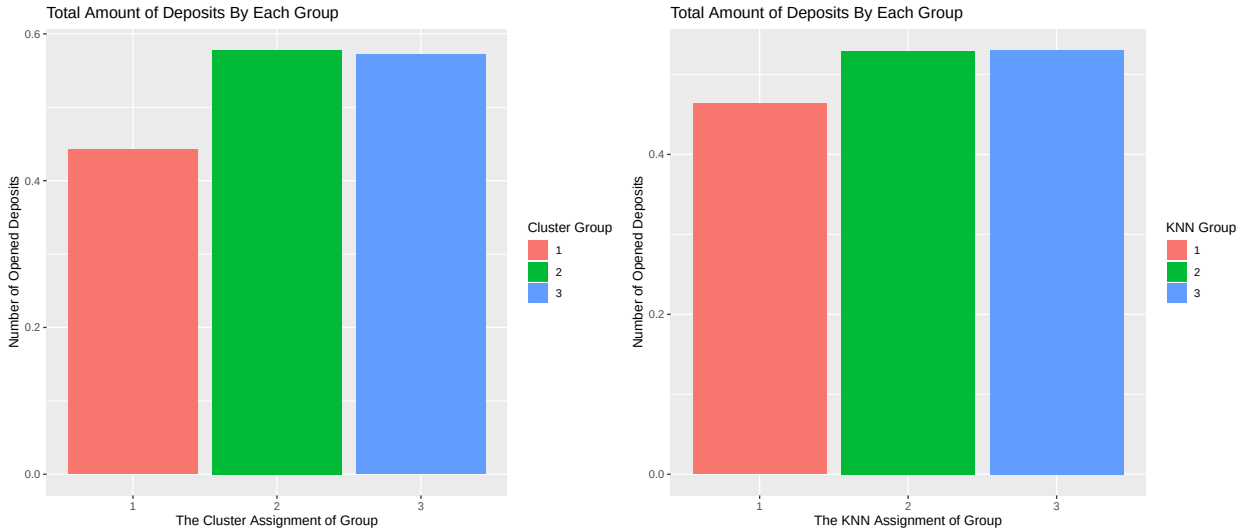
Using KNN To Improve Our Next Campaign

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 90% of our data. The remaining 10%, which is 1063.4 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.

The below graphics show us the identification of all groups through the hierarchical clustering from that train set, 90% of our data, and the KNN classifications from the test set, 10% of our data, respectively. The purpose of this is to validate our groups through the response metric that we want to focus on, not validating if it got correctly labeled which as we known our groups here are completely fictitious and do not hold any inherent value but are just a label. We can see that the KNN is doing pretty well to identify the quality of the group we care about. As the graphs have roughly the same average amount of deposits opened for their groups.

This would be helpful to our marketing team, as they get data about potential customers. They can utilize the KNN algorithm to easily and quickly classify people into groups. From those groups they can focus their

energy first on those that would reward them the most which are groups 2 and 3 as they are the most likely to open deposits on average. This should make our marketing team more profitable and efficient of opening more deposits for our company.



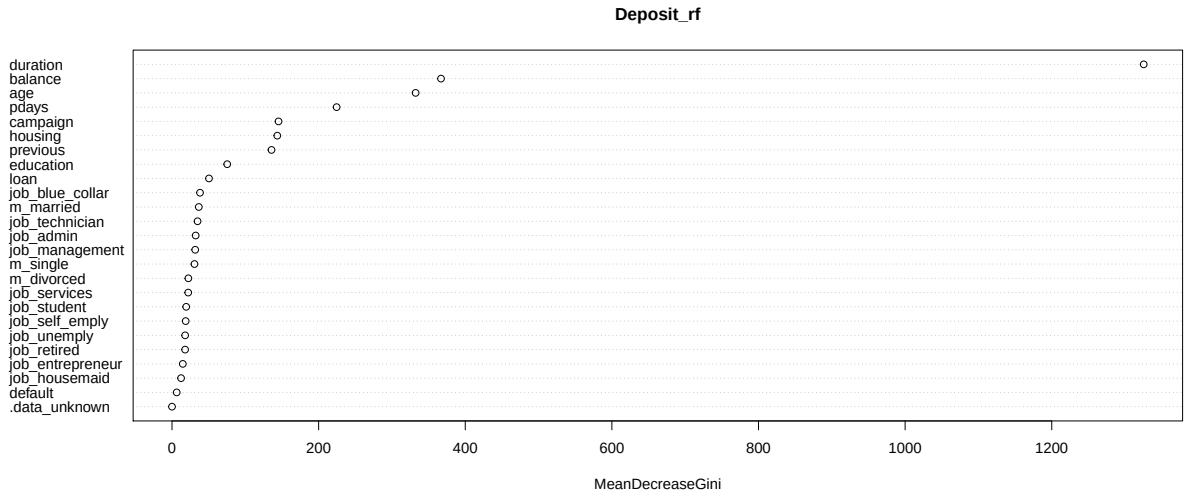
Using Supervised Machine Learning Methods

We already discussed some methods our company can use to become more effective and efficient but we did not include other effective methods which includes other supervised machine learning methods to predict variables. A variable we are very concerned with is "deposit", whether the person opened a term deposit or not. Since the outcome ("deposit") is a binary outcome, we need to use a classification method. Some of the best models are Logistic Regression, Random Forest, and eXtreme Gradient Boosting (XGBoosting) for this type of problem. In this section, XGBoost is applied instead of Boosting because it is an implementation of gradient boosted decision trees designed for speed and performance.

After those three models were trained and making predictions, we used a confusion matrix to describe the performance of our classification models. Two key performance metrics were chosen to compare model performance, AUC and F score. AUC (Area Under The Curve) - ROC (Receiver Operating Characteristics) curve is a performance measurement for classification problems at various thresholds settings. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at distinguishing between customers that will open a deposit and not open a deposit. Moreover, F Score is a weighted average of the true positive rate (recall) and precision. Higher the F score the better the model is.

As shown in the graph below, the XGBoosting has the highest AUC among them and the random forest has the highest F score among them. We can see that the logistics regression performed poorly for this dataset but we still cannot confirm which is the best between the remaining two models. Furthermore, we had run a t-test to compare model performance. P-value > 0.05 then the models are not statistically different, so we decided to run both models to explain the dataset.

We see that duration has a high correlation with term deposits meaning the higher the duration, the more likely it is for a client to open a term deposit. The mean duration time is 373s. People who were above the duration status, were more likely to open a term deposit. The factor of age and balance can have impact on deposit as well. The MSE train is 0.0369 and MSE test is 0.1376 which seems reasonable for our random forest.



From the XGBoost, we got that the five most important variables is `poutcome`, `contact`, `housing`, `loan`.

```
##                       top5   Overall
## poutcomesuccess poutcome 100.00000
## age                  age  72.91748
## contactunknown   contact  71.70119
## housingyes       housing  52.44418
## day                  day  47.23405
```

Our Conclusion

We have a few recommendations to our company. With our current data we would recommend that during our next campaign that we first focus on people in Group 3, then Group 2, finally Group 1. As Group 3 and 2 were the most likely to open up deposits. Also in the covariate plot, Group 3 had a higher balance and also took less time on the phone, according to the duration metric, making the team more efficient. When acquiring new data, we should apply the KNN algorithm to assign the group and follow the same order as above. However, we should predict the response variable using the supervised method for each new data observation in order to get an order within the group focusing first on those that have been predicted to the response we want. Our final and basic recommendation is that our marketing team utilize the groups to campaign differently to them based on other variables that we touched on in the exploratory analysis.