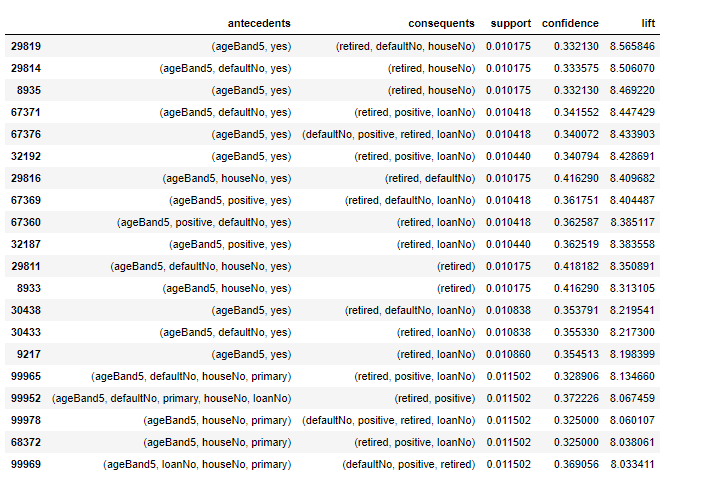
Ben Tennant

COMP 4321

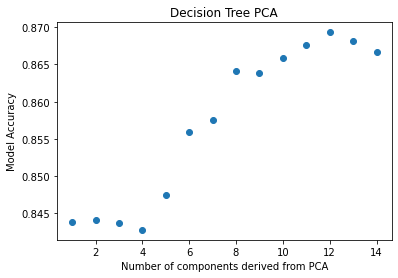
I started my analysis by looking for common association rules within the dataset. I found that I could derive some valuable information from the associations with the parameters below. You will also see the results of my rule mining. Before that I think it important to mention that my analysis may not be as thorough because I have zero subject matter knowledge on the dataset itself. For instance, the outcome ‘y’ represent whether or not the customer subscribed to term deposit. Frankly, I don’t know if that is the desired choice or not but we can still gather some information, but I may make assumptions about what the data represents.

Min\_supp = 0.01, min\_conf = 0.3, all lifts are greater than 8.05



One thing that I think is clear from these results is the most common demographic of people that did subscribe to the term deposit. We clearly see that ageBand5, the highest ageBand, representing people older than 50, most often do subscribe to the term deposit. We also see the ‘retired’ value present a lot implying again, that the older customers often do subscribe to the term deposit. We can also interpret the results to see that subscribing to the term deposit is associated with having no loans whatsoever and a positive current balance. Again, I believe that this can again be explained by age. Often older, retired people have paid off their loans already. One may also deduce that someone of any age category that has no loans will subscribe to the term deposit. One way I think this information can be made useful is in the marketing department. If the bank is hoping to get as many people to subscribe to the term deposit, then they may want to run some sort of marketing project and rather than paying to have the marketing strategy, maybe a promotion or just an ad of some sort, sent to every customer, they can focus on people that fit the criteria found above which will likely result in a better outcome for the marketing project.

I then used the two classification techniques to see if I could find the features that most often lead to a desired result. Again, I will assume that the desired result is subscribing to the term deposit. My analysis found that the decision tree model and Naïve-Bayes model produce rather different results. I felt the best way to determine the features that explain the most is run a principal component analysis for 1 feature through all features. After I found the number of features that produce the most accurate results I looked to determine which particular features those are. Below you will find the result of the PCA which is scatter plot with the number of components on the x-axis and the accuracy of the model built on those components on the y-axis.



From the result of the PCA on the decision tree model, I decided that the optimal number of components would be 12 because the number of components yielded the highest accuracy. I then used the feature importance’s of the tree model fit on all the variables. Those results are below.

Feature: 0, FName: age, Score: 0.10677

Feature: 1, FName: job, Score: 0.04859

Feature: 2, FName: marital, Score: 0.02260

Feature: 3, FName: education, Score: 0.02638

Feature: 4, FName: default, Score: 0.00184

Feature: 5, FName: balance, Score: 0.11072

Feature: 6, FName: housing, Score: 0.02515

Feature: 7, FName: loan, Score: 0.01009

Feature: 8, FName: contact, Score: 0.01894

Feature: 9, FName: day, Score: 0.09354

Feature: 10, FName: month, Score: 0.08723

Feature: 11, FName: duration, Score: 0.29259

Feature: 12, FName: campaign, Score: 0.03752

Feature: 13, FName: previous, Score: 0.04776

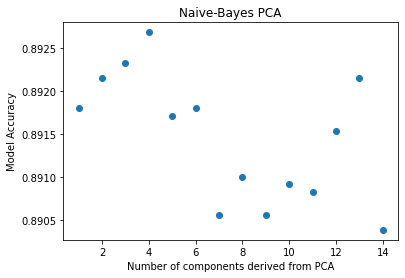
Feature: 14, FName: poutcome, Score: 0.07029

Based on the feature importances of the tree model fit on all components I would elect to drop 'default', 'loan' and 'contact' from the model. Default has the lowest feature importance so that is dropped for sure. 'Contact' appears to be the method used to contact the customer so I am not surprised that it is not very useful. The 'loan' however is surprising because I would imagine that a loan on the account would have a considerable effect on the predictive ability of the model. However, the model shows that 'loan' has very little importance to the model so I would still drop it. We can also see from these results that age is one of the most important features for determining the correct classification which I like to see because it agrees with the association rule conclusions. We can also see that balance is a strong feature which is a feature that I expected to have a strong importance. It makes sense that someone who has a larger

balance may be more likely to subscribe to the term deposit than someone with a low balance in their

account.

Next is the Naïve-Bayes model. Below you will again find the accuracy vs number of components graph and a list of the feature importances.



[0.01082898 0.00118553 0.00185791 0.00088472 0.00012386 0.00185791

0.00997965 0.00076086 0.03820225 0.01698664 0.0583208 0.04303282

0.00268955 0.04220119 0.03354862]

From the above results, I determined that four features produces the most accurate Naïve-Bayes model and those were 'contact', 'month', 'duration', 'previous'. It seems to find a lot of importance on timing and communication with the customer. 'Contact' is just the method of communication used to reach the customer, 'month' is the last month that the customer was contacted (I think), 'duration' is how long in

seconds the communication with the customer lasted and 'previous' is the number of times the

customer had been contacted before the current campaign. This is interesting because 'contact' was one of the components we removed from the decision tree model, however it is one of the most significant

components in the Naive-Bayes. Based on this analysis, I would conclude that the Naive-Bayes classifier would be the better model for this dataset. For every iteration of the PCA, its accuracy was higher than the decision tree's and the optimal Naive-Bayes model only used 4 components rather than 12 and the Naive-Bayes still had a better accuracy. I believe this also tells us that marketing is very important factor

to consider when trying to increase the number of subscriptions.