**Classifying MBA Student Success Risk**

The MBA program staff has observed that students who are admitted to the program with a low GPA or GMAT score will often have difficulty completing the program. Students who do no complete the program or have difficulty degrade the reputation of the program. In an effort to guide admissions decisions in regard to questionable GPA and GMAT scores, the staff needs a way to rate the students into three risk groups – group 1 is “low risk” and group 2 “med risk” group 3 “high risk.” The file MBAStudents.xlsm has the data.

1. Create a scatterplot of GPA and GMAT scores with different colors for the “satisfactory” and “unsatisfactory” ratings of current and 90% confidence ellipses for these two groups.
   1. Roughly how would you classify the data into three risk groups? What is your basis for this?

> ggplot (MBAStudents, aes (x = GPA, y = GMAT, colour = Rating, group = Rating)) + geom\_point() + stat\_ellipse(l=.9)



The intersection may roughly represent the medium risk since in this range students seem to about equally likely to be satisfactory or unsatisfactory.

* 1. Use k-means with 3 groups to determine the candidate clusters. Assign the current students to the clusters. Try to interpret to what degree the three clusters are meaningful.

Because k-means uses distance, we again must normalize the data because of different units in the variables.

> N<-as.data.frame(scale(rbind(MBAStudents[,3:4],NewApplicants[,2:3])))

> row.names(N)[71:75] = paste("A", 1:nrow(NewApplicants), sep="")

> mdl\_3means <- kmeans(N[1:70,],centers=3)

> MBAStudents$cluster <- mdl\_3means$cluster

> ggplot (MBAStudents, aes (x = GPA, y = GMAT, colour = factor(cluster), group = factor(cluster) )) + geom\_point() + stat\_ellipse(level =.9) + scale\_colour\_manual(values = c("red", "blue", "green")) 

> ggplot (MBAStudents, aes (x = GPA, y = GMAT, colour = factor(cluster), group = Rating )) + geom\_point() + stat\_ellipse(level =.9) + scale\_colour\_manual(values = c("red", "blue", "green"))



> ggplot (MBAStudents, aes (x = GPA, y = GMAT, colour = Rating, group = cluster )) + geom\_point() + stat\_ellipse(level =.95)



Cluster 1 seems to have mostly satisfactory, cluster 3 mostly unsatisfactory, while cluster 2 seems to have about equal satisfactory and unsatisfactory. A reasonable interpretation then for these clusters is

Cluster 1 = low risk of unsatisfactory

Cluster 2 = med risk of unsatisfactory

Cluster 3 = high risk of unsatisfactory

* 1. Use a classification methods to determine a classification rule for the clusters.

> tree <- rpart(factor(cluster) ~ GPA + GMAT, MBAStudents)

> decisionplot(tree, as.data.frame(cbind(MBAStudents[1:70,3:4], MBAStudents[,11])), class = "cluster", main = "3-means clustering", predict\_type="vector")



> plot(tree)

> text(tree ,all=TRUE,digits=3,use.n=TRUE,cex=0.8,xpd=TRUE)



So we would predict that an applicant with GPA < 2.925 is a high risk and should not be admitted while an applicant with a GPA > 2.925 and GMAT > 607 is a low risk and should be admitted. Applicants with GPA > 2.925 but GMAT < 607 are a medium risk and should be evaluated with other factors that may indicate success. For our new applicants, we get:

> predict(tree, newdata = NewApplicants, type="class")

A1 A2 A3 A4 A5

1 1 1 3 1

Levels: 1 2 3

This is consistent with the supervised 2-group classification except that A1 always was predicted to be unsatisfactory while here it’s rated low risk. It wasn’t that unconfident. We might expect that something like A5 which has low confidence might be placed into the med risk group.

* 1. Can this be improved by changing the number of groups or using a semi-supervised approach or a different clustering method (e.g. hierarchical)?