05\_Program\_DoubleMajor

Angel Huang

12/5/2020

# 1 Prepare data

## Load libraries

# Install EDA related libraries  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'stringr' was built under R version 3.6.2

## Warning: package 'forcats' was built under R version 3.6.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(emmeans)

## Warning: package 'emmeans' was built under R version 3.6.3

library(ggplot2)  
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 3.6.3

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

#library(plyr)  
library(dplyr) # for group by  
#library(tidyr)

## Load data from revised\_df.Rdata

load("../../RData/revised\_df.RData")  
dim(revised\_df)

## [1] 14171 10

We recall the variable names.

names(revised\_df)

## [1] "Retained" "Class" "Residency" "DoubleMajor"   
## [5] "Enrollment" "Credits" "CreditsFailed" "Program"   
## [9] "CumulativeTerms" "TestCredits"

df = revised\_df

## Variable engineering: From previous 03\_tree models, we discovered that credits failed and attempted are the first two main split, based on those split, we will convert failed credits into categories.

### Classification and Regression Tree (CART)

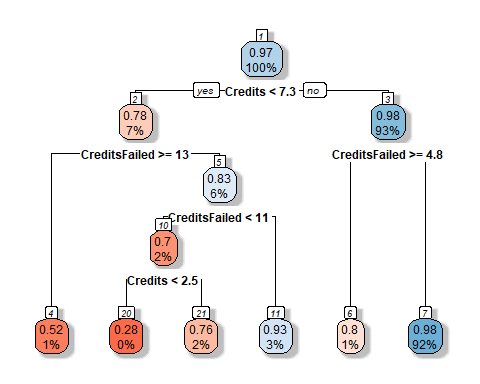
library(rpart)

## Warning: package 'rpart' was built under R version 3.6.3

library(rpart.plot)

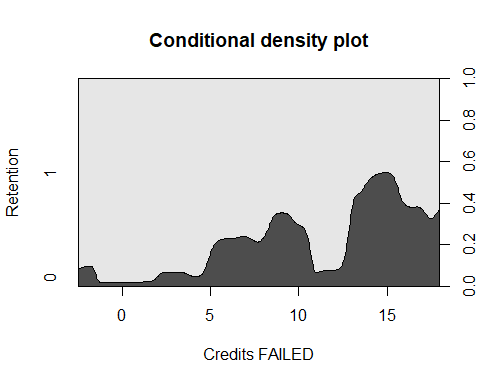
## Warning: package 'rpart.plot' was built under R version 3.6.3

#Build the model  
# For classification tree, response needs to be factor  
# tree doesn't need log transform  
ml\_formula = formula(Retained ~ Class +  
 Residency +  
 DoubleMajor +   
 Enrollment +   
 Credits +  
 CreditsFailed +  
 Program)  
# This gives the same result  
#ml\_formula = formula(Retained ~ Credits +  
# CreditsFailed)  
  
ml\_tree = rpart(ml\_formula,data=df,cp=0.008) # cp for complexity, smaller cp =more branches  
  
#Summarize the model  
#summary(ml\_tree)  
# Visualize the decision tree with rpart.plot  
rpart.plot(ml\_tree, box.palette="RdBu", shadow.col="gray", nn=TRUE)

 We fitted a decision tree model with retention rate being the dependent variable and class, residency, double major, enrollment, credits, credits failed, and program as the independent variables. Decision tree is a classification algorithm and uses gini impurity criteria to select splits, which will divide dataset into most homogenous sets. The fact that this tree splits mainly on credits and credits failed, indicating that these are the two of the most important variables for classification. In addition, we discovered the splits used for classification, which is Credits <2.5 and <7.3; splits for CreditsFailed are >=4.8, >=13. The highest retention group are students who earn >=7.3 credits and fail <4.8 credits (retention rate = 0.98, 92% of students are in this group). The lowest retention group (reddest) are students who earn <2.5 credits and fail 11~13 credits (retention rate = 0.28).

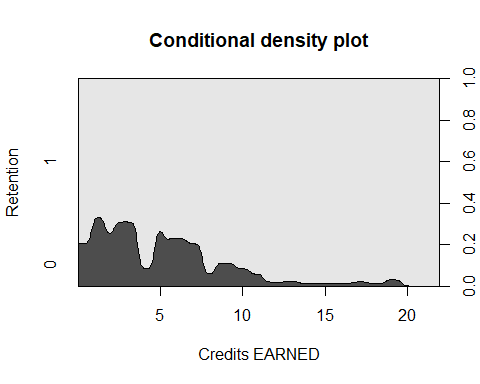
First, we do a density plot to visually verify the split

cdplot(factor(Retained) ~ CreditsFailed,   
 df,  
 #col=c("lightgoldenrod", "lightcyan"),   
 ylab = "Retention", xlab ="Credits FAILED", main = "Conditional density plot")

 More failed credits, more dropout. The split happens at >=4.8, >=13, consistent with the tree split.

Then density plot on credits attempted.

cdplot(factor(Retained) ~ Credits,   
 df,  
 #col=c("lightgoldenrod", "lightcyan"),   
 ylab = "Retention", xlab ="Credits EARNED", main = "Conditional density plot")

 The split happens at <4, <7.3, mostly consistent with the tree split.

## Create new variables

# convert variables to factor  
df$Failed = cut(  
 df$CreditsFailed,  
 breaks = c(-Inf, 4.8, 13, Inf), # If use 0 for the first one will make all entries with 0 into NA  
 labels = c("low (-3,4]", "med (4,12]", "high (12,18]"),  
 right = TRUE # include right hand number  
)  
df$Earned = cut(  
 df$Credits,  
 breaks = c(-Inf, 2.5, 7.3, Inf),  
 labels = c("low (0,2]", "med (2,7]", "high (7,22]"),  
 right = TRUE # include right hand number  
)

# 3. GLM model

## Use a full model to predict retention rate

df$DoubleMajorF = factor(df$DoubleMajor) # convert to factor  
glm1 = glm(Retained ~ (Class +  
 Residency +  
 DoubleMajorF +   
 Enrollment +   
 Credits +  
 CreditsFailed +  
 Program)^2,   
 data = df,  
 family = binomial(link = "logit"))

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glm1)

##   
## Call:  
## glm(formula = Retained ~ (Class + Residency + DoubleMajorF +   
## Enrollment + Credits + CreditsFailed + Program)^2, family = binomial(link = "logit"),   
## data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.6592 0.1063 0.1484 0.2149 1.8067   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error z value  
## (Intercept) -1.195e+00 1.030e+00 -1.161  
## ClassSO 1.290e+00 1.079e+00 1.195  
## ClassJR 2.278e+00 1.100e+00 2.071  
## ClassSR 1.914e+00 1.074e+00 1.782  
## ResidencyOut-of-State -2.749e-01 7.434e-01 -0.370  
## DoubleMajorF1 2.801e+00 1.295e+00 2.163  
## EnrollmentTransfer Student -3.103e+00 1.426e+00 -2.177  
## Credits 3.654e-01 7.285e-02 5.015  
## CreditsFailed 7.599e-02 7.456e-02 1.019  
## ProgramBACH 1.524e+00 1.239e+00 1.230  
## ProgramBAED 1.168e+01 1.459e+03 0.008  
## ProgramBAMJ 5.225e+00 2.210e+00 2.364  
## ProgramBFA 1.828e+01 1.443e+04 0.001  
## ProgramBMUS 2.083e+02 5.523e+03 0.038  
## ProgramBS 1.284e-01 7.778e-01 0.165  
## ProgramBSBA 3.247e+00 2.501e+00 1.298  
## ProgramBSIS 2.203e+01 2.796e+03 0.008  
## ProgramBSN 3.429e-01 2.307e+00 0.149  
## ProgramBSPH 1.007e+00 3.841e+00 0.262  
## ProgramBSPHS 7.567e+00 1.059e+01 0.714  
## ClassSO:ResidencyOut-of-State -3.511e-01 4.189e-01 -0.838  
## ClassJR:ResidencyOut-of-State -4.921e-01 4.252e-01 -1.157  
## ClassSR:ResidencyOut-of-State -4.591e-01 4.406e-01 -1.042  
## ClassSO:DoubleMajorF1 3.105e-01 6.360e-01 0.488  
## ClassJR:DoubleMajorF1 1.216e+00 6.591e-01 1.845  
## ClassSR:DoubleMajorF1 1.052e+00 6.158e-01 1.708  
## ClassSO:EnrollmentTransfer Student 1.126e+00 1.257e+00 0.896  
## ClassJR:EnrollmentTransfer Student 1.079e+00 1.230e+00 0.877  
## ClassSR:EnrollmentTransfer Student 2.387e+00 1.256e+00 1.901  
## ClassSO:Credits -6.660e-02 7.668e-02 -0.869  
## ClassJR:Credits -1.210e-01 7.844e-02 -1.542  
## ClassSR:Credits -1.877e-01 7.688e-02 -2.441  
## ClassSO:CreditsFailed -1.012e-02 7.982e-02 -0.127  
## ClassJR:CreditsFailed -4.321e-02 8.118e-02 -0.532  
## ClassSR:CreditsFailed -6.619e-02 7.954e-02 -0.832  
## ClassSO:ProgramBACH -9.520e-02 5.364e-01 -0.177  
## ClassJR:ProgramBACH -7.910e-01 5.022e-01 -1.575  
## ClassSR:ProgramBACH 4.938e-01 1.130e+00 0.437  
## ClassSO:ProgramBAED -1.549e+01 1.459e+03 -0.011  
## ClassJR:ProgramBAED -1.369e+01 1.459e+03 -0.009  
## ClassSR:ProgramBAED 1.270e+00 1.742e+03 0.001  
## ClassSO:ProgramBAMJ 5.278e-01 7.620e-01 0.693  
## ClassJR:ProgramBAMJ 8.066e-01 6.387e-01 1.263  
## ClassSR:ProgramBAMJ NA NA NA  
## ClassSO:ProgramBFA -4.847e-01 6.270e+03 0.000  
## ClassJR:ProgramBFA -6.222e-01 6.054e+03 0.000  
## ClassSR:ProgramBFA 2.491e-01 5.621e+03 0.000  
## ClassSO:ProgramBMUS 8.001e+01 3.583e+03 0.022  
## ClassJR:ProgramBMUS 2.675e+01 2.782e+03 0.010  
## ClassSR:ProgramBMUS 6.896e+01 3.299e+03 0.021  
## ClassSO:ProgramBS -1.466e-01 4.269e-01 -0.343  
## ClassJR:ProgramBS 4.110e-01 4.461e-01 0.921  
## ClassSR:ProgramBS 8.564e-01 4.330e-01 1.978  
## ClassSO:ProgramBSBA -1.425e-01 6.892e-01 -0.207  
## ClassJR:ProgramBSBA 9.693e-01 9.178e-01 1.056  
## ClassSR:ProgramBSBA 1.543e+01 4.154e+02 0.037  
## ClassSO:ProgramBSIS -1.734e+01 2.796e+03 -0.006  
## ClassJR:ProgramBSIS -8.453e-01 2.947e+03 0.000  
## ClassSR:ProgramBSIS -1.514e+01 2.796e+03 -0.005  
## ClassSO:ProgramBSN -6.131e-01 1.020e+00 -0.601  
## ClassJR:ProgramBSN -2.723e+00 1.086e+00 -2.506  
## ClassSR:ProgramBSN 3.225e-01 1.129e+00 0.286  
## ClassSO:ProgramBSPH 4.593e-01 1.066e+00 0.431  
## ClassJR:ProgramBSPH 1.481e+01 1.031e+03 0.014  
## ClassSR:ProgramBSPH 1.401e+00 1.087e+00 1.289  
## ClassSO:ProgramBSPHS -4.132e+00 4.335e+00 -0.953  
## ClassJR:ProgramBSPHS 1.173e+01 2.420e+03 0.005  
## ClassSR:ProgramBSPHS NA NA NA  
## ResidencyOut-of-State:DoubleMajorF1 -2.164e-01 3.758e-01 -0.576  
## ResidencyOut-of-State:EnrollmentTransfer Student 6.899e-01 3.608e-01 1.912  
## ResidencyOut-of-State:Credits 6.253e-03 4.784e-02 0.131  
## ResidencyOut-of-State:CreditsFailed 3.571e-02 5.316e-02 0.672  
## ResidencyOut-of-State:ProgramBACH -1.218e-01 4.697e-01 -0.259  
## ResidencyOut-of-State:ProgramBAED -1.876e+00 1.375e+00 -1.364  
## ResidencyOut-of-State:ProgramBAMJ 2.989e-01 7.231e-01 0.413  
## ResidencyOut-of-State:ProgramBFA -3.719e+01 1.007e+04 -0.004  
## ResidencyOut-of-State:ProgramBMUS 5.393e+01 2.376e+03 0.023  
## ResidencyOut-of-State:ProgramBS -5.088e-01 3.030e-01 -1.679  
## ResidencyOut-of-State:ProgramBSBA -1.054e+00 6.075e-01 -1.735  
## ResidencyOut-of-State:ProgramBSIS -2.411e+00 2.516e+00 -0.958  
## ResidencyOut-of-State:ProgramBSN 6.084e-01 1.468e+00 0.414  
## ResidencyOut-of-State:ProgramBSPH 6.509e-01 1.220e+00 0.533  
## ResidencyOut-of-State:ProgramBSPHS 1.704e+01 2.163e+03 0.008  
## DoubleMajorF1:EnrollmentTransfer Student -5.181e-01 4.793e-01 -1.081  
## DoubleMajorF1:Credits -1.827e-01 7.778e-02 -2.349  
## DoubleMajorF1:CreditsFailed -1.831e-01 8.848e-02 -2.070  
## DoubleMajorF1:ProgramBACH 1.380e+01 6.523e+03 0.002  
## DoubleMajorF1:ProgramBAED -3.051e-01 1.413e+00 -0.216  
## DoubleMajorF1:ProgramBAMJ -8.607e-01 6.414e-01 -1.342  
## DoubleMajorF1:ProgramBFA -6.638e-01 4.025e+03 0.000  
## DoubleMajorF1:ProgramBMUS -7.308e+01 1.934e+03 -0.038  
## DoubleMajorF1:ProgramBS -2.300e-01 3.884e-01 -0.592  
## DoubleMajorF1:ProgramBSBA 1.766e-01 1.159e+00 0.152  
## DoubleMajorF1:ProgramBSIS -9.486e-01 1.990e+00 -0.477  
## DoubleMajorF1:ProgramBSN 1.387e+01 1.560e+03 0.009  
## DoubleMajorF1:ProgramBSPH -1.024e+00 1.255e+00 -0.816  
## DoubleMajorF1:ProgramBSPHS -4.141e+00 2.055e+00 -2.015  
## EnrollmentTransfer Student:Credits 1.584e-01 5.235e-02 3.026  
## EnrollmentTransfer Student:CreditsFailed 2.804e-02 5.420e-02 0.517  
## EnrollmentTransfer Student:ProgramBACH 1.406e+00 1.214e+00 1.158  
## EnrollmentTransfer Student:ProgramBAED -1.635e+00 1.358e+00 -1.204  
## EnrollmentTransfer Student:ProgramBAMJ -2.096e+00 6.605e-01 -3.173  
## EnrollmentTransfer Student:ProgramBFA -4.593e-01 4.075e+03 0.000  
## EnrollmentTransfer Student:ProgramBMUS -4.921e+01 6.721e+03 -0.007  
## EnrollmentTransfer Student:ProgramBS -1.428e+00 3.763e-01 -3.795  
## EnrollmentTransfer Student:ProgramBSBA -5.492e-01 1.224e+00 -0.449  
## EnrollmentTransfer Student:ProgramBSIS 1.223e+01 1.822e+03 0.007  
## EnrollmentTransfer Student:ProgramBSN 1.895e+00 1.004e+00 1.887  
## EnrollmentTransfer Student:ProgramBSPH 1.284e+01 1.477e+03 0.009  
## EnrollmentTransfer Student:ProgramBSPHS 1.441e+01 6.523e+03 0.002  
## Credits:CreditsFailed -2.056e-02 2.945e-03 -6.982  
## Credits:ProgramBACH -1.328e-01 8.484e-02 -1.565  
## Credits:ProgramBAED 2.226e-01 2.492e-01 0.893  
## Credits:ProgramBAMJ -3.981e-01 1.489e-01 -2.674  
## Credits:ProgramBFA -2.488e-01 1.075e+03 0.000  
## Credits:ProgramBMUS -1.080e+01 2.753e+02 -0.039  
## Credits:ProgramBS 1.393e-02 4.981e-02 0.280  
## Credits:ProgramBSBA -1.669e-01 1.677e-01 -0.995  
## Credits:ProgramBSIS -3.728e-01 6.764e-01 -0.551  
## Credits:ProgramBSN -5.336e-03 1.559e-01 -0.034  
## Credits:ProgramBSPH -7.513e-02 2.613e-01 -0.287  
## Credits:ProgramBSPHS -2.567e-01 7.076e-01 -0.363  
## CreditsFailed:ProgramBACH -7.637e-02 9.276e-02 -0.823  
## CreditsFailed:ProgramBAED 2.848e-01 2.889e-01 0.986  
## CreditsFailed:ProgramBAMJ -2.526e-01 1.824e-01 -1.385  
## CreditsFailed:ProgramBFA 1.292e-01 1.260e+03 0.000  
## CreditsFailed:ProgramBMUS -1.598e+01 4.188e+02 -0.038  
## CreditsFailed:ProgramBS -2.616e-02 5.404e-02 -0.484  
## CreditsFailed:ProgramBSBA -1.148e-01 1.828e-01 -0.628  
## CreditsFailed:ProgramBSIS -2.895e-01 6.276e-01 -0.461  
## CreditsFailed:ProgramBSN -8.359e-02 1.670e-01 -0.501  
## CreditsFailed:ProgramBSPH -1.780e-01 2.751e-01 -0.647  
## CreditsFailed:ProgramBSPHS -4.857e-01 7.381e-01 -0.658  
## Pr(>|z|)   
## (Intercept) 0.245766   
## ClassSO 0.231951   
## ClassJR 0.038343 \*   
## ClassSR 0.074760 .   
## ResidencyOut-of-State 0.711523   
## DoubleMajorF1 0.030549 \*   
## EnrollmentTransfer Student 0.029506 \*   
## Credits 5.29e-07 \*\*\*  
## CreditsFailed 0.308103   
## ProgramBACH 0.218719   
## ProgramBAED 0.993612   
## ProgramBAMJ 0.018089 \*   
## ProgramBFA 0.998989   
## ProgramBMUS 0.969919   
## ProgramBS 0.868911   
## ProgramBSBA 0.194245   
## ProgramBSIS 0.993712   
## ProgramBSN 0.881865   
## ProgramBSPH 0.793244   
## ProgramBSPHS 0.475086   
## ClassSO:ResidencyOut-of-State 0.401881   
## ClassJR:ResidencyOut-of-State 0.247141   
## ClassSR:ResidencyOut-of-State 0.297394   
## ClassSO:DoubleMajorF1 0.625378   
## ClassJR:DoubleMajorF1 0.065103 .   
## ClassSR:DoubleMajorF1 0.087604 .   
## ClassSO:EnrollmentTransfer Student 0.370279   
## ClassJR:EnrollmentTransfer Student 0.380491   
## ClassSR:EnrollmentTransfer Student 0.057300 .   
## ClassSO:Credits 0.385112   
## ClassJR:Credits 0.122979   
## ClassSR:Credits 0.014644 \*   
## ClassSO:CreditsFailed 0.899071   
## ClassJR:CreditsFailed 0.594576   
## ClassSR:CreditsFailed 0.405302   
## ClassSO:ProgramBACH 0.859127   
## ClassJR:ProgramBACH 0.115208   
## ClassSR:ProgramBACH 0.662133   
## ClassSO:ProgramBAED 0.991531   
## ClassJR:ProgramBAED 0.992514   
## ClassSR:ProgramBAED 0.999418   
## ClassSO:ProgramBAMJ 0.488550   
## ClassJR:ProgramBAMJ 0.206584   
## ClassSR:ProgramBAMJ NA   
## ClassSO:ProgramBFA 0.999938   
## ClassJR:ProgramBFA 0.999918   
## ClassSR:ProgramBFA 0.999965   
## ClassSO:ProgramBMUS 0.982187   
## ClassJR:ProgramBMUS 0.992329   
## ClassSR:ProgramBMUS 0.983321   
## ClassSO:ProgramBS 0.731288   
## ClassJR:ProgramBS 0.356830   
## ClassSR:ProgramBS 0.047967 \*   
## ClassSO:ProgramBSBA 0.836216   
## ClassJR:ProgramBSBA 0.290910   
## ClassSR:ProgramBSBA 0.970367   
## ClassSO:ProgramBSIS 0.995052   
## ClassJR:ProgramBSIS 0.999771   
## ClassSR:ProgramBSIS 0.995679   
## ClassSO:ProgramBSN 0.547780   
## ClassJR:ProgramBSN 0.012194 \*   
## ClassSR:ProgramBSN 0.775188   
## ClassSO:ProgramBSPH 0.666452   
## ClassJR:ProgramBSPH 0.988536   
## ClassSR:ProgramBSPH 0.197498   
## ClassSO:ProgramBSPHS 0.340468   
## ClassJR:ProgramBSPHS 0.996132   
## ClassSR:ProgramBSPHS NA   
## ResidencyOut-of-State:DoubleMajorF1 0.564752   
## ResidencyOut-of-State:EnrollmentTransfer Student 0.055871 .   
## ResidencyOut-of-State:Credits 0.895994   
## ResidencyOut-of-State:CreditsFailed 0.501681   
## ResidencyOut-of-State:ProgramBACH 0.795315   
## ResidencyOut-of-State:ProgramBAED 0.172455   
## ResidencyOut-of-State:ProgramBAMJ 0.679369   
## ResidencyOut-of-State:ProgramBFA 0.997054   
## ResidencyOut-of-State:ProgramBMUS 0.981892   
## ResidencyOut-of-State:ProgramBS 0.093161 .   
## ResidencyOut-of-State:ProgramBSBA 0.082824 .   
## ResidencyOut-of-State:ProgramBSIS 0.338003   
## ResidencyOut-of-State:ProgramBSN 0.678529   
## ResidencyOut-of-State:ProgramBSPH 0.593705   
## ResidencyOut-of-State:ProgramBSPHS 0.993714   
## DoubleMajorF1:EnrollmentTransfer Student 0.279688   
## DoubleMajorF1:Credits 0.018807 \*   
## DoubleMajorF1:CreditsFailed 0.038464 \*   
## DoubleMajorF1:ProgramBACH 0.998312   
## DoubleMajorF1:ProgramBAED 0.829047   
## DoubleMajorF1:ProgramBAMJ 0.179614   
## DoubleMajorF1:ProgramBFA 0.999868   
## DoubleMajorF1:ProgramBMUS 0.969859   
## DoubleMajorF1:ProgramBS 0.553686   
## DoubleMajorF1:ProgramBSBA 0.878926   
## DoubleMajorF1:ProgramBSIS 0.633554   
## DoubleMajorF1:ProgramBSN 0.992909   
## DoubleMajorF1:ProgramBSPH 0.414420   
## DoubleMajorF1:ProgramBSPHS 0.043943 \*   
## EnrollmentTransfer Student:Credits 0.002478 \*\*   
## EnrollmentTransfer Student:CreditsFailed 0.604890   
## EnrollmentTransfer Student:ProgramBACH 0.246771   
## EnrollmentTransfer Student:ProgramBAED 0.228676   
## EnrollmentTransfer Student:ProgramBAMJ 0.001508 \*\*   
## EnrollmentTransfer Student:ProgramBFA 0.999910   
## EnrollmentTransfer Student:ProgramBMUS 0.994158   
## EnrollmentTransfer Student:ProgramBS 0.000148 \*\*\*  
## EnrollmentTransfer Student:ProgramBSBA 0.653639   
## EnrollmentTransfer Student:ProgramBSIS 0.994644   
## EnrollmentTransfer Student:ProgramBSN 0.059111 .   
## EnrollmentTransfer Student:ProgramBSPH 0.993065   
## EnrollmentTransfer Student:ProgramBSPHS 0.998238   
## Credits:CreditsFailed 2.92e-12 \*\*\*  
## Credits:ProgramBACH 0.117657   
## Credits:ProgramBAED 0.371797   
## Credits:ProgramBAMJ 0.007502 \*\*   
## Credits:ProgramBFA 0.999815   
## Credits:ProgramBMUS 0.968718   
## Credits:ProgramBS 0.779809   
## Credits:ProgramBSBA 0.319677   
## Credits:ProgramBSIS 0.581512   
## Credits:ProgramBSN 0.972701   
## Credits:ProgramBSPH 0.773739   
## Credits:ProgramBSPHS 0.716746   
## CreditsFailed:ProgramBACH 0.410328   
## CreditsFailed:ProgramBAED 0.324326   
## CreditsFailed:ProgramBAMJ 0.165981   
## CreditsFailed:ProgramBFA 0.999918   
## CreditsFailed:ProgramBMUS 0.969569   
## CreditsFailed:ProgramBS 0.628347   
## CreditsFailed:ProgramBSBA 0.530240   
## CreditsFailed:ProgramBSIS 0.644539   
## CreditsFailed:ProgramBSN 0.616714   
## CreditsFailed:ProgramBSPH 0.517578   
## CreditsFailed:ProgramBSPHS 0.510461   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4051.6 on 14170 degrees of freedom  
## Residual deviance: 3027.0 on 14040 degrees of freedom  
## AIC: 3289  
##   
## Number of Fisher Scoring iterations: 17

df$pred = predict(glm1,df[,c("Class","Residency", "DoubleMajorF", "Enrollment","Credits","CreditsFailed","Program")],type="response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

# Use Credits and creditsFailed for a better prediction

Some interesting significant interactions are: Credits:ProgramBAMJ, DoubleMajor:Credits, DoubleMajor:CreditsFailed.

# 4. Plot for variables of interest (Earned, Failed, Program, DoubleMajor)

Using a decision tree model, we have identified critical splits of variable “FailedCredits” and “Credits”, based on this, we divided FailedCredits into 3 intervals (i.e. Failed): low=0-5, med=5.5-13, high=13.5-Inf, and Credits into 3 intervals (i.e. Earned): low=0-2, med=2.5-7, high=7-Inf, for easier interpretation.

Here we use box plots to visualize interaction effects between the 4 variables of interest: Earned, Failed, Program, DoubleMajor on (full-GLM-model estimated) probability of return.

## 4.1 Retention rate by Earned and Failed

### Violin plot – CDF too narrow (dispersed), can’t see shape

#ggplot(df, aes(x=Earned, y=pred, fill=Failed)) +  
# geom\_violin() +  
 #geom\_boxplot(width=0.1) + theme\_minimal()

### Helper function to get n for each category

getN <- function(x){  
 return(c(y = 0, label = length(x)))  
}  
getMedian <- function(x){  
 return(c(y = median(x)\*1.05, label = round(median(x),digit=2)))  
}  
# experiment with the multiplier to find the perfect position

### Real retention rate by Earned and Failed

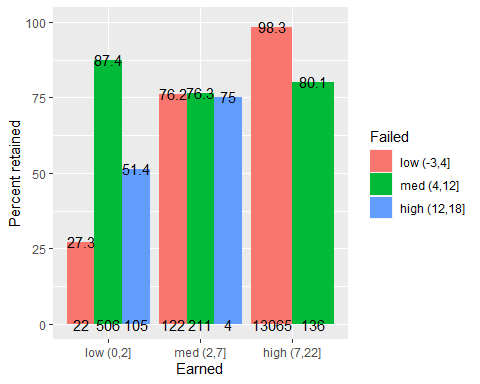
yLim = c(0,100)  
df\_plot = df %>%  
 group\_by(Earned, Failed) %>%  
 summarise(retained = mean(Retained), n=n())

## `summarise()` regrouping output by 'Earned' (override with `.groups` argument)

df\_plot

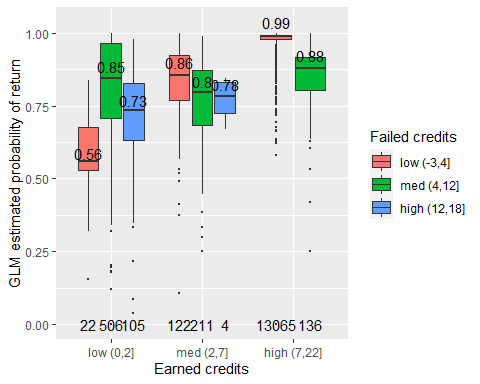
## # A tibble: 8 x 4  
## # Groups: Earned [3]  
## Earned Failed retained n  
## <fct> <fct> <dbl> <int>  
## 1 low (0,2] low (-3,4] 0.273 22  
## 2 low (0,2] med (4,12] 0.874 506  
## 3 low (0,2] high (12,18] 0.514 105  
## 4 med (2,7] low (-3,4] 0.762 122  
## 5 med (2,7] med (4,12] 0.763 211  
## 6 med (2,7] high (12,18] 0.75 4  
## 7 high (7,22] low (-3,4] 0.983 13065  
## 8 high (7,22] med (4,12] 0.801 136

ggplot(df\_plot,aes(x = Earned, y = retained\*100, fill=Failed)) +   
 geom\_bar(position="dodge", stat="identity") +  
 ylab("Percent retained") +  
 coord\_cartesian(ylim=yLim) + # has to use coord\_cartesian, which doesn't filter on data like ylim does  
 #theme(axis.text.x = element\_text(angle =45)) + # tilt x label  
   
 # add student count  
 geom\_text(data=data.frame(df\_plot),   
 aes(Earned, yLim[1], label=df\_plot$n),   
 position = position\_dodge(width=0.9),  
 size=4) +  
 geom\_text(data=data.frame(df\_plot),   
 aes(Earned, retained\*100.5, label=round(df\_plot$retained\*100,digit=1)),  
 position = position\_dodge(width=0.9),  
 size=4)

 In low credit group, 27% of student with <=4 failed credits returned, which is very low, whereas 87% of student with 4-12 failed credits returned, indicating students who registered for more credits are more likely to have higher motivation for return. However, when too many credits are failed (>12), only 51% of students returned. In the median credit group, regardless of failed credits, around 75% students returned. In the high credit group, there is no one failed more than 12 credits, and the majority of students failed <=4 credits, and 98% of these students returned. The rest of the students failed 4-12 credits and have 80% returning rate. The group that needs attention most is the one with <=2 credits earned and <=4 credits failed, then is the one with <=2 credits earned and >12 credits failed.

### Predicted retention rate by Earned and Failed

ggplot(data=df, aes(x=Earned, y=pred, fill=Failed)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Earned credits") +   
 labs(fill = "Failed credits") +  
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=4) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=4)



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1) #

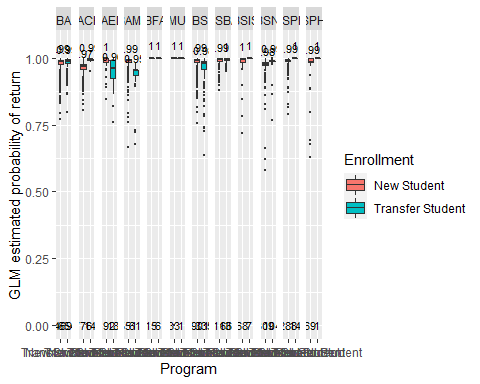
Since we identified a significant interaction effect between earned and failed credits on retention rate in both a GLM model and a decision tree model, now we will investigate further into this interaction effect. The categories for earned and failed credits were defined by the decision tree split shown above, in this way, we can split the data into the most homogeneous groups while enabling easier interpretation. When earn less than 2 credits, retention rate is low. Within this group, if failed credit is <=4, the median return rate is the lowest, around 56%, probably because so few credits were attempted in total. Interestingly, when failed 4~12 credits, the return rate increases to 85%, then when more credits are failed, the return rate drops again to around 73%. This implies a combination effect of how many credits are attempted in total versus how many are failed. When earns 2-7 credits, the return rate is similar regardless of failed credits, with a slight decreasing trend with increased failed credits. When earn more than 7 credits, there is no student fail more than 12 credits. The lower failed credit results in highest return rate ~99%, which is also the category with the most number of students. This trend is consistent with the real retention rate by earned and failed credits, providing support for the GLM model predictions.

### Selecting the high earned and low failed group

df\_goodCredit = subset(df, Failed=="low (-3,4]" & Earned=="high (7,22]")

### Look for where the outliers come from

ggplot(data=df\_goodCredit, aes(x=Enrollment, y=pred, fill=Enrollment)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Program") +   
 labs(fill = "Enrollment") +   
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=3) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=3) +  
 facet\_grid(col = vars(Program))



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

The outliers seem to come from many majors, especially BS, BAED, and more likely from transfer students (although they also have smaller sample size). I’ve also tried Residency, Class, non of them clearly separate out the outliers.

## 4.2 Real retention rate by Program

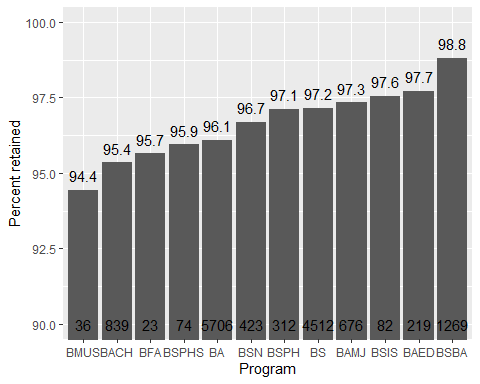
df\_plot = df %>%  
 group\_by(Program) %>%  
 summarise(retained = mean(Retained), n=n()) %>%  
 arrange(retained)

## `summarise()` ungrouping output (override with `.groups` argument)

df\_plot

## # A tibble: 12 x 3  
## Program retained n  
## <chr> <dbl> <int>  
## 1 BMUS 0.944 36  
## 2 BACH 0.954 839  
## 3 BFA 0.957 23  
## 4 BSPHS 0.959 74  
## 5 BA 0.961 5706  
## 6 BSN 0.967 423  
## 7 BSPH 0.971 312  
## 8 BS 0.972 4512  
## 9 BAMJ 0.973 676  
## 10 BSIS 0.976 82  
## 11 BAED 0.977 219  
## 12 BSBA 0.988 1269

ggplot(df\_plot,aes(x = reorder(Program, retained), y = retained\*100)) +   
 ylab("Percent retained") +  
 xlab("Program") +  
 geom\_col() +  
 coord\_cartesian(ylim=c(90,100)) + # has to use coord\_cartesian, which doesn't filter on data like ylim does  
 #theme(axis.text.x = element\_text(angle =45)) + # tilt x label  
 # add student count  
 geom\_text(data=data.frame(df\_plot),   
 aes(Program, 90, label=df\_plot$n),   
 position = position\_dodge(width=0.9),  
 size=4) +  
 geom\_text(data=data.frame(df\_plot),   
 aes(Program, retained\*100.5, label=round(df\_plot$retained\*100,digit=1)),  
 position = position\_dodge(width=0.9),  
 size=4)

 The programs with lowest retention rate is BMUS (Music Performance, 36 students), BACH (College of Arts and Sciences, 848 students), BFA (Studio Arts, 23 students), BSPHS (Pharmaceutical Sciences, 74 students), but all with smaller sample sizes. The programs with highest retention rate is BSBA (Business admin, 1270 students), BAED (School of education, 219 students).

### Pick 3 programs with the highest number of students as an example: science, art, and business

df\_program = subset(df, Program == 'BS' | Program == 'BA' | Program == 'BSBA')

### Predicted return by earned and program

ggplot(data=df\_program, aes(x=Program, y=pred, fill=Earned)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Program") +   
 labs(fill = "Earned credits") +   
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=3) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=3) +  
 facet\_grid(col = vars(DoubleMajorF), labeller = as\_labeller(c(`0`="Single Major",`1`="Double Major")))

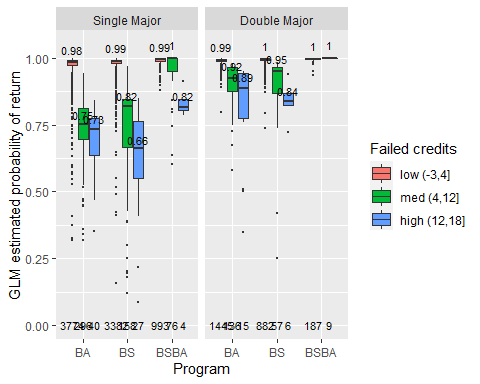


#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

Next, we are interested in how would earned credits and double major affects students’ return rate in different majors. We picked the three representative programs with the highest number of students: science, art, and business. Y axis indicates the probability of a student to return estimated by a full GLM model. Several interesting take-homes are: 1. There is a high return rate for business major, regardless of double major status and earned credits. This explains why there is such a high return rate (98.8%) in general. 2. For BA and BS, if a student earns high credits, there is a very high probability (98%) that they will return. 3. However, if the student earns low to median credits, double major students are more likely to return (~94%) than single major students (70~80%). In addition, single major students showed a gradient effect of increased credits will result in higher probability of return, whereas double major students remained a high probability of return for both low and median level of earned credits. Thus, the single major students are more sensitive to how many credits they earned, except for students from certain programs such as business.

### Predicted return by failed and program

ggplot(data=df\_program, aes(x=Program, y=pred, fill=Failed)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Program") +   
 labs(fill = "Failed credits") +   
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=3) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=3) +  
 facet\_grid(col = vars(DoubleMajorF), labeller = as\_labeller(c(`0`="Single Major",`1`="Double Major")))



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

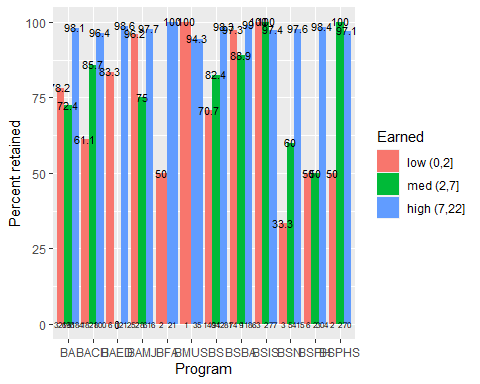
Unlike earned credits, when we look at failed credits, it seems like all students are demotivated (have lower probability of return) by increased failed credits, as shown by the gradient decrease in probability of return as failed credits increase. Consistent with previous findings, double major students are more resilient to failed credits in two ways. One is that they have a generally higher probability of return at each failed credit level, which is 10~20% higher than single major students. Second is that they are less sensitive to increased failed credits (less steep slope) compared to single major students. In addition, business students are still the most resilient and have higher probability of return. However we now see that even single major business students are demotivated by a high level of failed credits (>12 credits failed).

### Real retention rate by Earned and Program

yLim = c(0,100)  
df\_plot = df %>%  
 group\_by(Program, Earned) %>%  
 summarise(retained = mean(Retained), n=n())

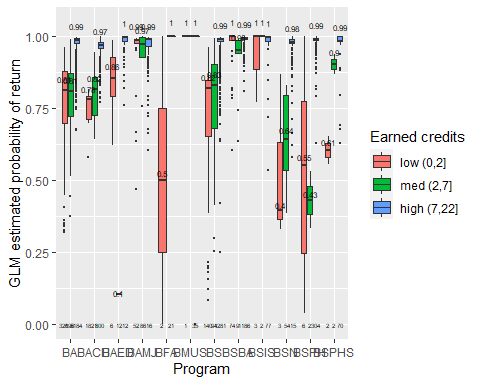
## `summarise()` regrouping output by 'Program' (override with `.groups` argument)

#df\_plot  
ggplot(df\_plot,aes(x = Program, y = retained\*100, fill=Earned)) +   
 geom\_bar(position="dodge", stat="identity") +  
 ylab("Percent retained") +  
 coord\_cartesian(ylim=yLim) + # has to use coord\_cartesian, which doesn't filter on data like ylim does  
 #theme(axis.text.x = element\_text(angle =45)) + # tilt x label  
   
 # add student count  
 geom\_text(data=data.frame(df\_plot),   
 aes(Program, yLim[1], label=df\_plot$n),   
 position = position\_dodge(width=0.9),  
 size=2) +  
 geom\_text(data=data.frame(df\_plot),   
 aes(Program, retained\*100.5, label=round(df\_plot$retained\*100,digit=1)),  
 position = position\_dodge(width=0.9),  
 size=3)

 Program with lowest % return is low credit students in BSN (1/3 returned), BFA (1/2 returned), BSPH(3/6 returned), BSPHS (1/2 returned).

## Predicted retention rate by Earned and Program

ggplot(data=df, aes(x=Program, y=pred, fill=Earned)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Program") +   
 labs(fill = "Earned credits") +   
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=1.7) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=2)



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

Interaction effect of Program and Earned on retention rate: For BA (5749 students), rate is >75%, lowest for students earned 2-7 credits. For BACH (College of Arts and Sciences, 848 students), For BAED (School of education, 219 students), retention rate is high, even for students earned <=2 credits, the retention rate is ~85%. For BAMJ (School of Media and Journalism, 676 students), retention rate is high in general, but lower (~75%) for students earned 2-7 credits. For BFA (Studio Arts, 23 students), low earned credits reduce retention rate to 50%. For BMUS (Music Performance, 36 students), sample size is small, but retention rate is almost 100%. For BS (4524 students), higher earned credits increases retention rate. For BSBA (Business admin, 1270 students), all retention rate is around 95%. For BSIS (Information Science, 84 students), sample size is small, but retention rate is almost 100%. For BSN (Nursing, 424 students), they have the lowest retention rate among programs, especially when earned credit is low, rate <25%, or medium, rate <50%. However, when earned high credits, return rate is very high >95%. For BSPH (Public Health, 312 students), for low and median earned credits, rate ~50%, for high credits, ~100%. For BSPHS (Pharmaceutical Sciences, 74 students), small sample size. When earned credits<=2, rate~50%, otherwise rate ~100%.

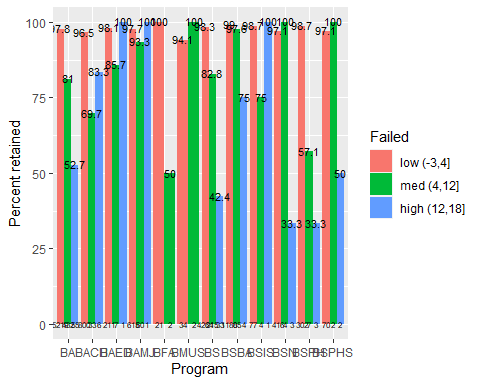
## 4.3 Retention rate by Failed and Program

### Real retention rate by Failed and Program

yLim = c(0,100)  
df\_plot = df %>%  
 group\_by(Program, Failed) %>%  
 summarise(retained = mean(Retained), n=n())

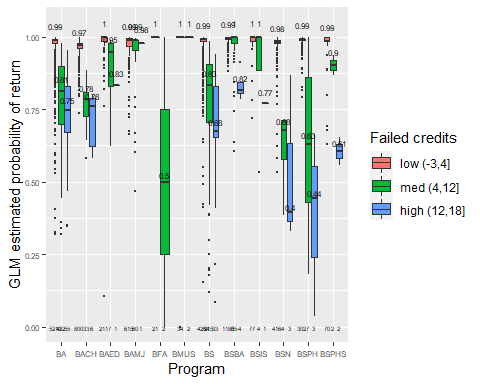
## `summarise()` regrouping output by 'Program' (override with `.groups` argument)

#df\_plot  
ggplot(df\_plot,aes(x = Program, y = retained\*100, fill=Failed)) +   
 geom\_bar(position="dodge", stat="identity") +  
 ylab("Percent retained") +  
 coord\_cartesian(ylim=yLim) + # has to use coord\_cartesian, which doesn't filter on data like ylim does  
 #theme(axis.text.x = element\_text(angle =45)) + # tilt x label  
   
 # add student count  
 geom\_text(data=data.frame(df\_plot),   
 aes(Program, yLim[1], label=df\_plot$n),   
 position = position\_dodge(width=0.9),  
 size=2) +  
 geom\_text(data=data.frame(df\_plot),   
 aes(Program, retained\*100.5, label=round(df\_plot$retained\*100,digit=1)),  
 position = position\_dodge(width=0.9),  
 size=3)



### Predicted retention rate by Failed and Program

ggplot(data=df, aes(x=Program, y=pred, fill=Failed)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Program") +   
 theme(axis.text=element\_text(size=6)) +  
 labs(fill = "Failed credits") +  
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=1.7) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=2)



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

Interaction effect of Program and Earned on retention rate: For all programs, if failed credits <=4, return rate is ~100%. When failed 4-12 credits, BFA and BSPH has the lowest return rate, ~50%. When failed >12 credits, BSN and BSPH has the lowest return rate, ~13% and ~38% respectively. Should be the focus of intervention.

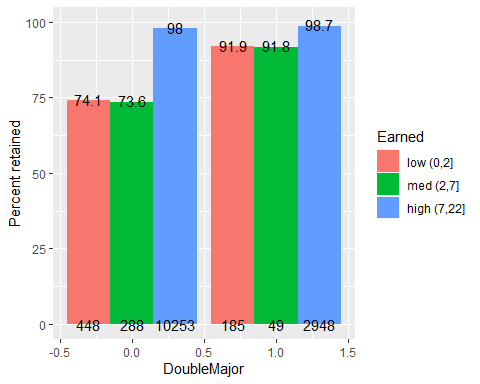
## 4.4 Retention rate by Earned and DoubleMajor

### Real retention rate by Earned and DoubleMajor

yLim = c(0,100)  
df\_plot = df %>%  
 group\_by(DoubleMajor, Earned) %>%  
 summarise(retained = mean(Retained), n=n())

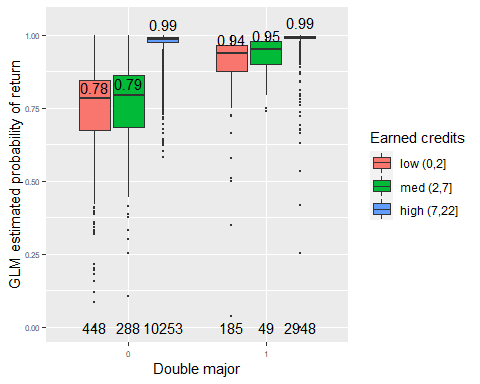
## `summarise()` regrouping output by 'DoubleMajor' (override with `.groups` argument)

#df\_plot  
ggplot(df\_plot,aes(x = DoubleMajor, y = retained\*100, fill=Earned)) +   
 geom\_bar(position="dodge", stat="identity") +  
 ylab("Percent retained") +  
 coord\_cartesian(ylim=yLim) + # has to use coord\_cartesian, which doesn't filter on data like ylim does  
 #theme(axis.text.x = element\_text(angle =45)) + # tilt x label  
   
 # add student count  
 geom\_text(data=data.frame(df\_plot),   
 aes(DoubleMajor, yLim[1], label=df\_plot$n),   
 position = position\_dodge(width=0.9),  
 size=4) +  
 geom\_text(data=data.frame(df\_plot),   
 aes(DoubleMajor, retained\*100.5, label=round(df\_plot$retained\*100,digit=1)),  
 position = position\_dodge(width=0.9),  
 size=4)

 Main difference is in low credit and med credit group, 92% of double major students returned whereas 74% of single major students returned.

### Predicted retention rate by Earned and DoubleMajor

ggplot(data=df, aes(x=DoubleMajorF, y=pred, fill=Earned)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Double major") +   
 theme(axis.text=element\_text(size=6)) +  
 labs(fill = "Earned credits") +  
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=4) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=4)



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

Interaction effect of Double major and Earned on retention rate: If a student earns >7 credits, retention rate is 99% regardless of double major status. If a student earned <=7 credits, retention rate is lower for single major than double major (79% vs. 95%).

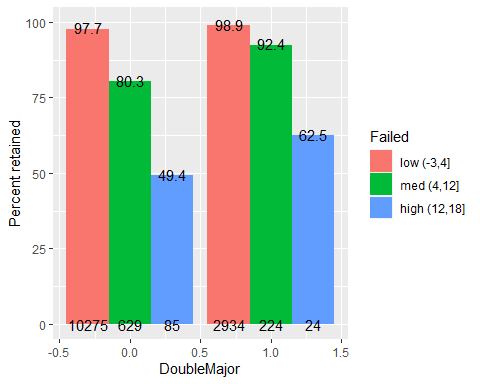
## 4.5 Retention rate by Failed and DoubleMajor

### Real retention rate by Failed and DoubleMajor

yLim = c(0,100)  
df\_plot = df %>%  
 group\_by(DoubleMajor, Failed) %>%  
 summarise(retained = mean(Retained), n=n())

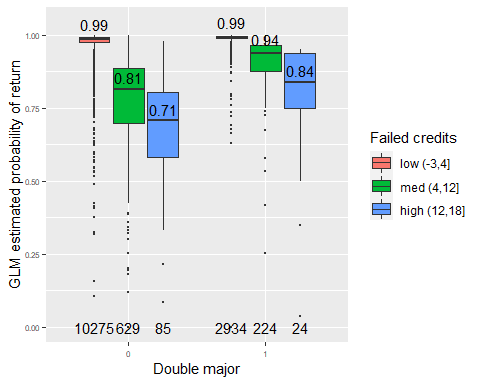
## `summarise()` regrouping output by 'DoubleMajor' (override with `.groups` argument)

#df\_plot  
ggplot(df\_plot,aes(x = DoubleMajor, y = retained\*100, fill=Failed)) +   
 geom\_bar(position="dodge", stat="identity") +  
 ylab("Percent retained") +  
 coord\_cartesian(ylim=yLim) + # has to use coord\_cartesian, which doesn't filter on data like ylim does  
 #theme(axis.text.x = element\_text(angle =45)) + # tilt x label  
   
 # add student count  
 geom\_text(data=data.frame(df\_plot),   
 aes(DoubleMajor, yLim[1], label=df\_plot$n),   
 position = position\_dodge(width=0.9),  
 size=4) +  
 geom\_text(data=data.frame(df\_plot),   
 aes(DoubleMajor, retained\*100.5, label=round(df\_plot$retained\*100,digit=1)),  
 position = position\_dodge(width=0.9),  
 size=4)

 Double major has higher return rate than single major, in all levels of failed credits. Students with single major are more sensitive to failed credits (have steeper slope of percent retained as failed credits increase), whereas double major students are more resilient.

### Predicted retention rate by Earned and DoubleMajor

ggplot(data=df, aes(x=DoubleMajorF, y=pred, fill=Failed)) +  
 geom\_boxplot(outlier.size = 0.1) +   
 ylab("GLM estimated probability of return") +  
 xlab("Double major") +   
 theme(axis.text=element\_text(size=6)) +  
 labs(fill = "Failed credits") +  
 stat\_summary(fun.data = getN, geom = "text",  
 position = position\_dodge(width = 0.75), size=4) +  
 stat\_summary(fun.data = getMedian, geom = "text",  
 position = position\_dodge(width = 0.75), size=4)



#geom\_dotplot(binaxis='y', stackdir='center',dotsize=0.1)

Interaction effect of Double major and Failed on retention rate: Single major students have lower return rate than double major, regardless of how many credits earned. When failed more credits, single major students are more sensitive and more likely to dropout.