week5 teamwork

Zhaoqi Liu

12/7/2020

load("~/Downloads/STOR 765 consulting/STOR765\_FALL2020\_TeamDataScience/RData/revised\_df.RData")

library(ggplot2)  
library(dplyr)

df<-revised\_df  
df$Residency<-as.factor(revised\_df$Residency)  
df$DoubleMajor<-as.factor(revised\_df$DoubleMajor)  
df$Enrollment<-as.factor(revised\_df$Enrollment)  
df$Program<-as.factor(revised\_df$Program)

#if include program's interaction, the algorithm did not converge   
fit\_all <- glm(Retained ~ (Class +  
 Residency +  
 DoubleMajor +   
 Enrollment +   
 Credits +  
 CreditsFailed+  
 TestCredits)^2+Program,   
 data = df,  
 family = binomial(link = "logit"))

#summary(fit\_all)

anova(fit\_all, test = "LR")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Retained  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 14170 4051.6   
## Class 3 24.21 14167 4027.4 2.255e-05 \*\*\*  
## Residency 1 12.74 14166 4014.7 0.0003574 \*\*\*  
## DoubleMajor 1 35.18 14165 3979.5 3.003e-09 \*\*\*  
## Enrollment 1 32.46 14164 3947.0 1.214e-08 \*\*\*  
## Credits 1 572.57 14163 3374.5 < 2.2e-16 \*\*\*  
## CreditsFailed 1 1.02 14162 3373.4 0.3122221   
## TestCredits 1 30.20 14161 3343.2 3.900e-08 \*\*\*  
## Program 13 65.53 14148 3277.7 5.274e-09 \*\*\*  
## Class:Residency 3 1.44 14145 3276.3 0.6951671   
## Class:DoubleMajor 3 9.99 14142 3266.3 0.0186307 \*   
## Class:Enrollment 3 13.97 14139 3252.3 0.0029493 \*\*   
## Class:Credits 3 22.32 14136 3230.0 5.608e-05 \*\*\*  
## Class:CreditsFailed 3 2.01 14133 3228.0 0.5704824   
## Class:TestCredits 3 10.36 14130 3217.6 0.0157059 \*   
## Residency:DoubleMajor 1 0.85 14129 3216.8 0.3564197   
## Residency:Enrollment 1 4.52 14128 3212.2 0.0335477 \*   
## Residency:Credits 1 0.38 14127 3211.9 0.5401969   
## Residency:CreditsFailed 1 0.05 14126 3211.8 0.8203053   
## Residency:TestCredits 1 2.07 14125 3209.7 0.1505674   
## DoubleMajor:Enrollment 1 1.33 14124 3208.4 0.2493221   
## DoubleMajor:Credits 1 5.52 14123 3202.9 0.0188311 \*   
## DoubleMajor:CreditsFailed 1 9.83 14122 3193.1 0.0017168 \*\*   
## DoubleMajor:TestCredits 1 0.03 14121 3193.0 0.8721912   
## Enrollment:Credits 1 16.63 14120 3176.4 4.536e-05 \*\*\*  
## Enrollment:CreditsFailed 1 0.04 14119 3176.4 0.8327921   
## Enrollment:TestCredits 1 0.77 14118 3175.6 0.3807325   
## Credits:CreditsFailed 1 45.95 14117 3129.6 1.212e-11 \*\*\*  
## Credits:TestCredits 1 4.76 14116 3124.9 0.0291826 \*   
## CreditsFailed:TestCredits 1 1.69 14115 3123.2 0.1931599   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

###drop some interaction terms choose interaction term based on the anova test above and test if the reduced model has lack of fit

fit\_reduced <- glm(Retained ~ Class + Residency + DoubleMajor + Enrollment +   
 Credits + CreditsFailed+TestCredits + Program +  
 DoubleMajor\*(Class+CreditsFailed) +   
 Credits\*(Class+DoubleMajor+Enrollment+  
 CreditsFailed+TestCredits) +  
 Class\*Enrollment,  
   
 data = df,  
 family = binomial(link = "logit"))

summary(fit\_reduced)

##   
## Call:  
## glm(formula = Retained ~ Class + Residency + DoubleMajor + Enrollment +   
## Credits + CreditsFailed + TestCredits + Program + DoubleMajor \*   
## (Class + CreditsFailed) + Credits \* (Class + DoubleMajor +   
## Enrollment + CreditsFailed + TestCredits) + Class \* Enrollment,   
## family = binomial(link = "logit"), data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1228 0.1291 0.1613 0.2142 1.5458   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.400e-01 4.712e-01 -0.934 0.350352   
## ClassSO 6.887e-01 3.831e-01 1.798 0.072228 .   
## ClassJR 1.213e+00 4.185e-01 2.898 0.003758 \*\*   
## ClassSR 6.183e-01 4.287e-01 1.442 0.149191   
## ResidencyOut-of-State -5.421e-01 1.237e-01 -4.383 1.17e-05 \*\*\*  
## DoubleMajor1 3.315e+00 1.238e+00 2.679 0.007389 \*\*   
## EnrollmentTransfer Student -1.880e+00 1.195e+00 -1.572 0.115856   
## Credits 3.335e-01 3.452e-02 9.662 < 2e-16 \*\*\*  
## CreditsFailed 3.137e-02 2.385e-02 1.315 0.188469   
## TestCredits 4.385e-02 1.099e-02 3.990 6.62e-05 \*\*\*  
## ProgramASBCH -1.715e-01 2.255e-01 -0.760 0.447109   
## ProgramASBFA 1.414e+01 5.588e+02 0.025 0.979816   
## ProgramASBM -9.563e-01 1.050e+00 -0.910 0.362604   
## ProgramASBS 2.597e-01 1.666e-01 1.559 0.119087   
## ProgramILBSI 1.044e-01 7.635e-01 0.137 0.891200   
## ProgramKFBSB 3.451e+00 1.011e+00 3.412 0.000646 \*\*\*  
## ProgramMJBAB 1.382e+01 4.158e+02 0.033 0.973491   
## ProgramMJBAJ 2.125e-01 2.751e-01 0.773 0.439770   
## ProgramPHBSP 9.148e-01 7.392e-01 1.238 0.215861   
## ProgramSDBS 1.270e+01 3.625e+02 0.035 0.972058   
## ProgramSEBAE 4.978e-01 7.294e-01 0.682 0.494950   
## ProgramSMBS 1.210e+01 4.284e+02 0.028 0.977476   
## ProgramSNBSN 1.909e-01 4.992e-01 0.383 0.702070   
## ClassSO:DoubleMajor1 2.445e-01 5.778e-01 0.423 0.672210   
## ClassJR:DoubleMajor1 1.135e+00 5.948e-01 1.909 0.056277 .   
## ClassSR:DoubleMajor1 7.567e-01 5.605e-01 1.350 0.177039   
## DoubleMajor1:CreditsFailed -2.321e-01 8.398e-02 -2.763 0.005723 \*\*   
## ClassSO:Credits -4.621e-02 3.319e-02 -1.392 0.163861   
## ClassJR:Credits -8.505e-02 3.280e-02 -2.593 0.009523 \*\*   
## ClassSR:Credits -1.084e-01 3.351e-02 -3.234 0.001220 \*\*   
## DoubleMajor1:Credits -2.354e-01 7.353e-02 -3.202 0.001367 \*\*   
## EnrollmentTransfer Student:Credits 8.907e-02 2.849e-02 3.127 0.001768 \*\*   
## Credits:CreditsFailed -2.046e-02 2.857e-03 -7.163 7.89e-13 \*\*\*  
## Credits:TestCredits -2.383e-03 9.492e-04 -2.510 0.012065 \*   
## ClassSO:EnrollmentTransfer Student 3.735e-01 1.195e+00 0.312 0.754719   
## ClassJR:EnrollmentTransfer Student 5.010e-01 1.167e+00 0.429 0.667741   
## ClassSR:EnrollmentTransfer Student 1.740e+00 1.192e+00 1.460 0.144347   
## ---  
## 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: 3142.1 on 14134 degrees of freedom  
## AIC: 3216.1  
##   
## Number of Fisher Scoring iterations: 15

anova(fit\_reduced, test="LR")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Retained  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 14170 4051.6   
## Class 3 24.21 14167 4027.4 2.255e-05 \*\*\*  
## Residency 1 12.74 14166 4014.7 0.0003574 \*\*\*  
## DoubleMajor 1 35.18 14165 3979.5 3.003e-09 \*\*\*  
## Enrollment 1 32.46 14164 3947.0 1.214e-08 \*\*\*  
## Credits 1 572.57 14163 3374.5 < 2.2e-16 \*\*\*  
## CreditsFailed 1 1.02 14162 3373.4 0.3122221   
## TestCredits 1 30.20 14161 3343.2 3.900e-08 \*\*\*  
## Program 13 65.53 14148 3277.7 5.274e-09 \*\*\*  
## Class:DoubleMajor 3 9.99 14145 3267.7 0.0186751 \*   
## DoubleMajor:CreditsFailed 1 3.02 14144 3264.7 0.0823841 .   
## Class:Credits 3 21.10 14141 3243.6 0.0001004 \*\*\*  
## DoubleMajor:Credits 1 16.27 14140 3227.3 5.481e-05 \*\*\*  
## Enrollment:Credits 1 13.22 14139 3214.1 0.0002775 \*\*\*  
## Credits:CreditsFailed 1 50.43 14138 3163.7 1.237e-12 \*\*\*  
## Credits:TestCredits 1 5.75 14137 3157.9 0.0164814 \*   
## Class:Enrollment 3 15.83 14134 3142.1 0.0012290 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Compare reduced model and full model: there is no lack of fit. The reduced model contains less interaction terms: doublemajor with class\_level, credits and credits\_failed. Credits with all the other features, and Class\*Enrollment

anova(fit\_reduced,fit\_all,test="LR")

## Analysis of Deviance Table  
##   
## Model 1: Retained ~ Class + Residency + DoubleMajor + Enrollment + Credits +   
## CreditsFailed + TestCredits + Program + DoubleMajor \* (Class +   
## CreditsFailed) + Credits \* (Class + DoubleMajor + Enrollment +   
## CreditsFailed + TestCredits) + Class \* Enrollment  
## Model 2: Retained ~ (Class + Residency + DoubleMajor + Enrollment + Credits +   
## CreditsFailed + TestCredits)^2 + Program  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 14134 3142.1   
## 2 14115 3123.2 19 18.903 0.4631

Add prediced retation rate from reduced model to df.

pred<-predict(fit\_reduced,type="response")  
df$pred<-pred

Create new columns for categorical credits and creditsfailed Since the minimum course load for undergraduates for a single semester is 12 academic credit hours. For students who have less than 12 credits for a semester, they need the permission from their dean. So we classify credits(actual earned) to two categories: below 12 and above 12. We also classify creditsfailed into sequence with an interval 3 since most classes are 3 credits.

cat\_creditfail<-cut(revised\_df$CreditsFailed,breaks = seq(-3,18,3),include.lowest = TRUE)  
table(cat\_creditfail)

## cat\_creditfail  
## [-3,0] (0,3] (3,6] (6,9] (9,12] (12,15] (15,18]   
## 11662 1356 361 140 499 117 36

cat\_credits<-ifelse(revised\_df$Credits<12,"less than 12 credits","greater than or equal to 12 credits" )  
#levels(cat\_credits)  
cat\_credits<-factor(cat\_credits, levels = c("less than 12 credits","greater than or equal to 12 credits"))  
cat\_credit2<-cut(revised\_df$Credits,breaks = c(seq(0,18,3),22),include.lowest = TRUE, right=FALSE)  
table(cat\_credit2)

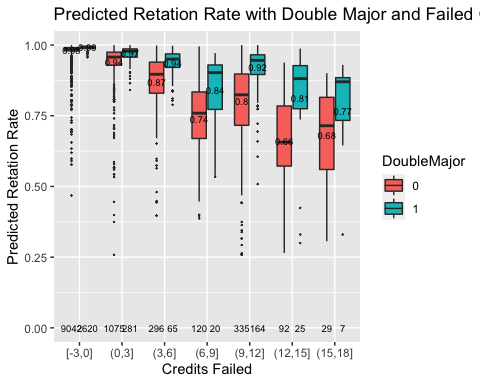
## cat\_credit2  
## [0,3) [3,6) [6,9) [9,12) [12,15) [15,18) [18,22]   
## 633 130 243 526 6437 5833 369

df$cat\_creditfail<-cat\_creditfail  
df$cat\_credit<-cat\_credits  
df$cat\_credit2<-cat\_credit2

#return the position and text lable   
getN <- function(x){  
 return(c(y = 0, label = length(x)))  
}  
getMean <- function(x){  
 return(c(y = mean(x), label = round(mean(x),digit=2)))  
}

###Double Major with Credit Fails add the total number of each category to the plot

df%>%ggplot(aes(x=cat\_creditfail, y=pred,fill=DoubleMajor))+  
 geom\_boxplot(outlier.size = 0.1) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "Predicted Retation Rate with Double Major and Failed Credits")+  
 xlab("Credits Failed")+  
 ylab("Predicted Retation Rate")

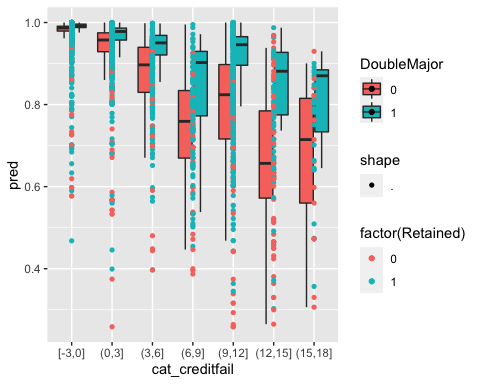


For students who did’t fail any credits, whether they have a second major doesn’t affect the predicted rate much, but students with one major have more outliers.

#don't know how to change it   
is\_outlier <- function(x) {  
 return(x < quantile(x, 0.25) - 1.5 \* IQR(x) | x > quantile(x, 0.75) + 1.5 \* IQR(x))  
}  
df1<-df  
df1%>%  
 group\_by(CreditsFailed,DoubleMajor) %>%  
 mutate(outlier = ifelse(is\_outlier(pred), pred, as.numeric(NA)))

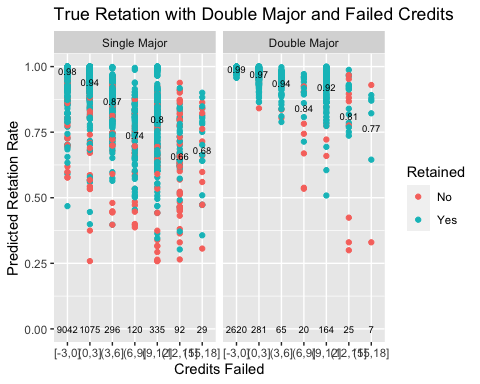
## # A tibble: 14,171 x 15  
## # Groups: CreditsFailed, DoubleMajor [50]  
## Retained Class Residency DoubleMajor Enrollment Credits CreditsFailed Program  
## <int> <fct> <fct> <fct> <fct> <dbl> <dbl> <fct>   
## 1 1 SO Out-of-S… 1 New Stude… 16 0 ASBCH   
## 2 1 SR In-State 0 New Stude… 17 0 ASBS   
## 3 1 JR Out-of-S… 1 Transfer … 12 4 ASBCH   
## 4 1 SO In-State 0 New Stude… 16 0 ASBCH   
## 5 1 JR In-State 0 Transfer … 12 0 ASBA   
## 6 1 SO Out-of-S… 0 New Stude… 14 0 ASBCH   
## 7 1 SO In-State 0 New Stude… 15 0 ASBCH   
## 8 0 SO In-State 1 Transfer … 10 7 ASBCH   
## 9 1 SO In-State 0 New Stude… 18 0 ASBCH   
## 10 1 SO In-State 0 New Stude… 12 0 ASBCH   
## # … with 14,161 more rows, and 7 more variables: CumulativeTerms <dbl>,  
## # TestCredits <dbl>, pred <dbl>, cat\_creditfail <fct>, cat\_credit <fct>,  
## # cat\_credit2 <fct>, outlier <dbl>

df1%>%ggplot(aes(x = cat\_creditfail, y = pred, fill=DoubleMajor)) +  
 geom\_boxplot(outlier.shape = NA) +  
 geom\_point(aes(x = cat\_creditfail,y = pred, fill=DoubleMajor, color = factor(Retained),shape ="."), na.rm = TRUE)



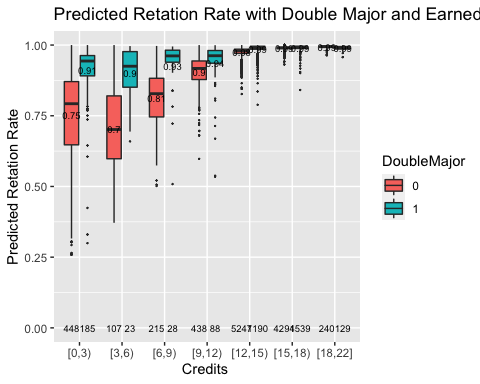
plot the retation rate with double major, credit fails and the true retation result

p<-df%>%ggplot(aes(x=cat\_creditfail, y=pred))+  
 geom\_point(aes(color=factor(Retained))) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "True Retation with Double Major and Failed Credits")+  
 scale\_color\_discrete(name = "Retained", labels =c("No","Yes"))+  
 xlab("Credits Failed")+  
 ylab("Predicted Retation Rate")  
p+facet\_grid(col = vars(DoubleMajor), labeller = as\_labeller(c(`0`="Single Major",`1`="Double Major")))



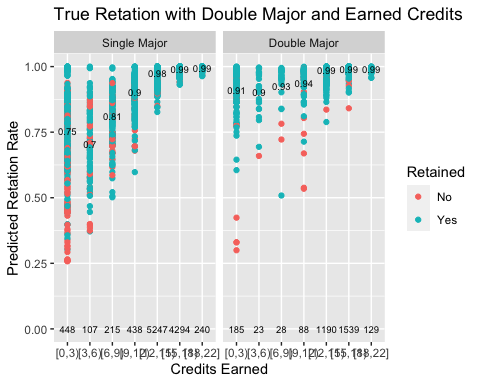
plot the predicted retation rate with double major and credits earned

df%>%ggplot(aes(x=cat\_credit2, y=pred,fill=DoubleMajor))+  
 geom\_boxplot(outlier.size = 0.1) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "Predicted Retation Rate with Double Major and Earned Credits")+  
 xlab("Credits ")+  
 ylab("Predicted Retation Rate")



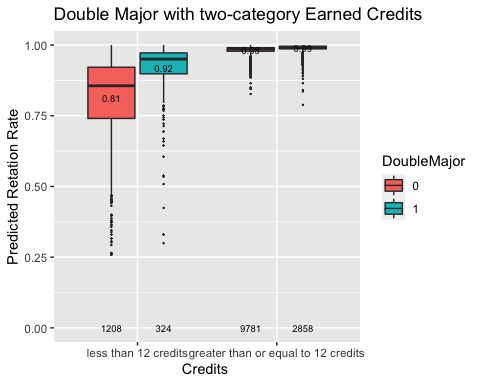
plot the predicted retation rate with double major, credits earned and the true retation result

p<-df%>%ggplot(aes(x=cat\_credit2, y=pred))+  
 geom\_point(aes(color=factor(Retained))) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "True Retation with Double Major and Earned Credits")+  
 scale\_color\_discrete(name = "Retained", labels =c("No","Yes"))+  
 xlab("Credits Earned")+  
 ylab("Predicted Retation Rate")  
p+facet\_grid(col = vars(DoubleMajor), labeller = as\_labeller(c(`0`="Single Major",`1`="Double Major")))



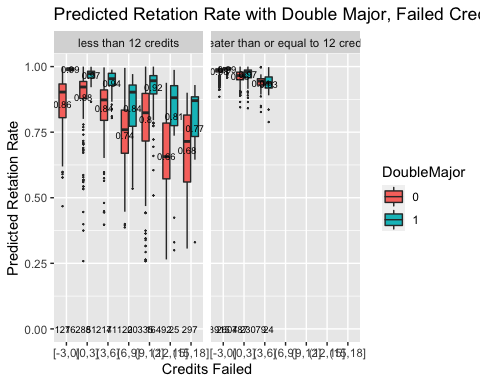
If treat credits earned as a two-category variable: <12 and >=12

df%>%ggplot(aes(x=cat\_credit, y=pred,fill=DoubleMajor))+  
 geom\_boxplot(outlier.size = 0.1) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "Double Major with two-category Earned Credits")+  
 xlab("Credits ")+  
 ylab("Predicted Retation Rate")



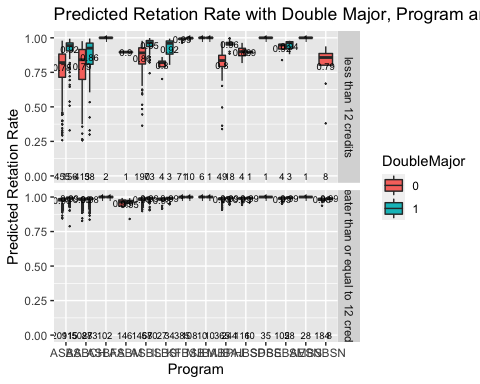
For students whose Credits <12, whether they have a second major plays an important role in determining whether they will come back.

p2<-df%>%ggplot(aes(x=cat\_creditfail, y=pred,fill=DoubleMajor))+  
 geom\_boxplot(outlier.size = 0.1) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "Predicted Retation Rate with Double Major, Failed Credits and two-category Credits")+  
 xlab("Credits Failed")+  
 ylab("Predicted Retation Rate")  
p2+facet\_grid(~cat\_credit)



For students who have earned less than 12 credits in this semester, which means undergraduate students cannot be regard as a full-time students at UNC, double major affects the retation. Those students who has a second major are more likely to return. For students who have earned more than 12 credits, double major doesn’t affect much on the retation rate prediction.

p3<-df%>%ggplot(aes(x=Program, y=pred,fill=DoubleMajor))+  
 geom\_boxplot(outlier.size = 0.1) +   
 stat\_summary(fun.data = getN, geom = "text",position = position\_dodge(width = 0.75), size=2.5)+  
 stat\_summary(fun.data = getMean, geom="text",position = position\_dodge(width = 0.75),size = 2.5)+  
 labs(title = "Predicted Retation Rate with Double Major, Program and Earned Credits")+  
 xlab("Program")+  
 ylab("Predicted Retation Rate")  
p3+facet\_grid(row=vars(cat\_credit))



ASBCH seems to be program related with natural science and mathematics. Considering the Programs, we see that students in some programs do not have a second major. The number of students in BA programs who obtain less than 12 credits in this semester, is larger than that of BS program. For students who obtain less than 12 credits this semester, students with a second major are more likely to return.

df%>%filter(cat\_credit=="less than 12 credits" )%>%  
 group\_by(Retained,Class,DoubleMajor)%>%  
 summarise(n = n())

## `summarise()` regrouping output by 'Retained', 'Class' (override with `.groups` argument)

## # A tibble: 16 x 4  
## # Groups: Retained, Class [8]  
## Retained Class DoubleMajor n  
## <int> <fct> <fct> <int>  
## 1 0 FR 0 31  
## 2 0 FR 1 3  
## 3 0 SO 0 48  
## 4 0 SO 1 9  
## 5 0 JR 0 80  
## 6 0 JR 1 5  
## 7 0 SR 0 69  
## 8 0 SR 1 15  
## 9 1 FR 0 114  
## 10 1 FR 1 5  
## 11 1 SO 0 203  
## 12 1 SO 1 34  
## 13 1 JR 0 398  
## 14 1 JR 1 123  
## 15 1 SR 0 265  
## 16 1 SR 1 130

We can see that for students has less than 12 credits, the number of students with a second major is less than the number of students with single major for all class level.