03. Full Logistic Regression Model

Angel Huang modified from Perry Haaland

11/29/2020

In this document, I’m going to look at fitting a GLM model with all of the terms that have been suggested so far by other team members.

## Load the data

We are all working from a common data set this week.

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

## [1] 14171 9

We recall the variable names.

names(revised\_df)

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

## Logistic Regression with all terms

In this analysis, I am going to start with the “big” model that has all of the terms and interactions. Unfortunately, the model containing MajorDiv failed to converge (result not shown). The model with the remaining terms is fit below.

fit\_all <- glm(Retained ~ (Class +  
 Residency +  
 DoubleMajor +   
 Enrollment +   
 Credits +  
 CreditsFailed +  
 Program)^2,   
 data = revised\_df,  
 family = binomial(link = "logit"))

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

Noted that fitted probabilities 0 or 1 occurred.

We can see from the table below.

aov\_all <- anova(fit\_all, test = "LR")

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

aov\_all

## 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   
## Program 11 46.98 14151 3326.5 2.168e-06 \*\*\*  
## Class:Residency 3 1.67 14148 3324.8 0.6426238   
## Class:DoubleMajor 3 9.61 14145 3315.2 0.0221862 \*   
## Class:Enrollment 3 11.43 14142 3303.7 0.0096057 \*\*   
## Class:Credits 3 25.37 14139 3278.4 1.290e-05 \*\*\*  
## Class:CreditsFailed 3 0.94 14136 3277.4 0.8168193   
## Class:Program 31 58.24 14105 3219.2 0.0021636 \*\*   
## Residency:DoubleMajor 1 0.68 14104 3218.5 0.4078977   
## Residency:Enrollment 1 3.81 14103 3214.7 0.0510077 .   
## Residency:Credits 1 0.74 14102 3214.0 0.3910688   
## Residency:CreditsFailed 1 0.01 14101 3213.9 0.9313944   
## Residency:Program 11 10.72 14090 3203.2 0.4670282   
## DoubleMajor:Enrollment 1 1.64 14089 3201.6 0.1996873   
## DoubleMajor:Credits 1 7.12 14088 3194.5 0.0076382 \*\*   
## DoubleMajor:CreditsFailed 1 10.68 14087 3183.8 0.0010802 \*\*   
## DoubleMajor:Program 11 16.55 14076 3167.2 0.1218386   
## Enrollment:Credits 1 17.84 14075 3149.4 2.405e-05 \*\*\*  
## Enrollment:CreditsFailed 1 0.04 14074 3149.3 0.8330797   
## Enrollment:Program 11 33.70 14063 3115.6 0.0004042 \*\*\*  
## Credits:CreditsFailed 1 48.48 14062 3067.2 3.333e-12 \*\*\*  
## Credits:Program 11 29.84 14051 3037.3 0.0016793 \*\*   
## CreditsFailed:Program 11 10.36 14040 3027.0 0.4980682   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Pareto Plot of Effects

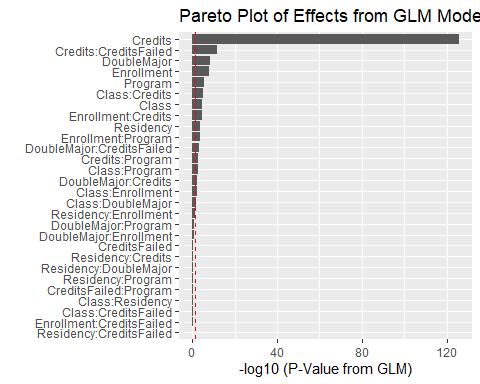
We are going to use the AOV output to make a graph. Note that P-values and Deviance give the same ordering when not including “Program”, but different ordering when including “Program”.

aov\_df <- as.data.frame(aov\_all)  
## fix the last name to make it easier  
names(aov\_df)[5] <- "Pvalue"  
names(aov\_df)[4] <- "Resid.Dev"  
aov\_df$Term = rownames(aov\_df)  
rownames(aov\_df) <- NULL  
## first row is the intercept  
aov\_df <- aov\_df[-1,]  
## arrange from most to least significant  
aov\_df <- aov\_df %>%  
 select(Term, Deviance, Pvalue) %>%  
 arrange(Pvalue)   
aov\_df

## Term Deviance Pvalue  
## 1 Credits 5.725657e+02 1.553149e-126  
## 2 Credits:CreditsFailed 4.848203e+01 3.333295e-12  
## 3 DoubleMajor 3.518213e+01 3.002655e-09  
## 4 Enrollment 3.246405e+01 1.214177e-08  
## 5 Program 4.698055e+01 2.167667e-06  
## 6 Class:Credits 2.537407e+01 1.289518e-05  
## 7 Class 2.421243e+01 2.255462e-05  
## 8 Enrollment:Credits 1.783786e+01 2.405493e-05  
## 9 Residency 1.274255e+01 3.574310e-04  
## 10 Enrollment:Program 3.370420e+01 4.042442e-04  
## 11 DoubleMajor:CreditsFailed 1.068470e+01 1.080250e-03  
## 12 Credits:Program 2.983935e+01 1.679275e-03  
## 13 Class:Program 5.823989e+01 2.163644e-03  
## 14 DoubleMajor:Credits 7.116408e+00 7.638160e-03  
## 15 Class:Enrollment 1.143188e+01 9.605728e-03  
## 16 Class:DoubleMajor 9.610328e+00 2.218617e-02  
## 17 Residency:Enrollment 3.808022e+00 5.100765e-02  
## 18 DoubleMajor:Program 1.655215e+01 1.218386e-01  
## 19 DoubleMajor:Enrollment 1.644660e+00 1.996873e-01  
## 20 CreditsFailed 1.021253e+00 3.122221e-01  
## 21 Residency:Credits 7.356162e-01 3.910688e-01  
## 22 Residency:DoubleMajor 6.849200e-01 4.078977e-01  
## 23 Residency:Program 1.071979e+01 4.670282e-01  
## 24 CreditsFailed:Program 1.036292e+01 4.980682e-01  
## 25 Class:Residency 1.674473e+00 6.426238e-01  
## 26 Class:CreditsFailed 9.356416e-01 8.168193e-01  
## 27 Enrollment:CreditsFailed 4.441698e-02 8.330797e-01  
## 28 Residency:CreditsFailed 7.411589e-03 9.313944e-01

Now we make the Pareto Plot.

aov\_df %>%  
 ggplot() +  
 geom\_col(aes(y = -log10(Pvalue), x = reorder(Term, -Pvalue))) +  
 geom\_hline(yintercept = -log10(.05),   
 linetype = "dashed", color = "red") +  
 coord\_flip() +  
 labs(y = "-log10 (P-Value from GLM)",  
 x = "",  
 title = "Pareto Plot of Effects from GLM Model")

 The Pareto plot above shows the terms from the GLM logistic regression fit with Retention as the response. The terms are sorted from top to bottom from most to least significant. The model contained the main effects and all of the possible two-factor interactions:

* Credits = credit hours earned
* CreditsFailed = credit hours earned - credit hours attempted
* DoubleMajor = Yes = 1, No = 0
* Enrollment = Original enrollment status: New Student or New Transfer Student
* Program = First major degree
* Class = class year: FR, SO, JR, Sr
* Residency = In-state, Out-of-state

The red, vertical dashed line is the significance level of .05.

We note that the different levels of the descriptors Class, Residency, Enrollment, and DoubleMajor have different retention rates. So for example, In-state and Out-of-state residents have different retention rates, in general. The two closely related variables Credits and CreditsFailed seem to capture a dynamic that affects the retention rates overall and also differentially within the different student groups.

The model with MajorDiv did not converge, so we have not shown any results for that. It will need to be investigated separately.

We hypothesize that the two credits variables have an important impact on the student’s perceived progress toward a degree.