Predicting online student performance in an Open University course

Ahmad Al-Dhalaan

January 3, 2021

Contents

List of Tables

List of Figures

Abstract

We analyze an educational dataset with categorical and continuous independent variables in order to predict student success in an online course at The Open University. Independent variables consist of demographic characteristics and behavioral values with a binary outcome of completed or not completed. We fit a multiple logistic regression model. The model fits best with five independent variables: a student's previous education, presence of a previous attempt, current credits enrolled, score on the first marked assignment and number of clicks during the course. All estimated variable coefficients are significant at the 95% confidence level with the model have a residual deviance of 1783. Classification shows strong model predictive power with the ROC having an area under the curve of 0.81.

Introduction

In this report, we will be analyzing datasets provided by the Knowledge Media Institute at The Open University. The Open University is the largest university in the United Kingdom for undergraduate education. In our analysis, we use methods described in “An Introduction to Categorical Data Analysis” by Agresti (2019) and “Applied Linear Statistical Models” by Kutner et al. (2014).

Analysis Objectives

Online education has recently proliferated and matured in both content and delivery. This ubiquitousness has resulted in large amounts of data, readily accessible for analysis of student populations and their performance. Analysis of independent variables that are associated with student success is important, as it will help educators in building a profile of a successful student. Building a profile for each course can help educators in understanding the success factors for that particular course.

Demographics might help a course administrator in predicting before the commencement of a course the absence/presence of factors that have shown to be associated with success in a course. Factors such as age or highest education attained might shed light on how a student is predicted to perform, warranting extra attention as the course progresses. Student engagement and early results might also be strong indicators of success in a course. Since students are not attending in person, clicks or early registration might indicate engagement in the course. On the other hand, an early result on an assignment might predict a student’s performance in the course overall. These behaviors could be of interest to a course administrator in order to intervene early in the course and provide additional support or guidance. These variables could also lead to changes in course structure or the virtual environment.

In this explanatory report we are interested in determining whether successful completion of an online course can be predicted by a number of demographic and behavioral measurements in students enrolled in an online course. We seek to build a model that can predict the probability of a student completing their course given their accessible measurements. We will determine which independent variables are significant in predicting success.

Method

Data

Data in this report has been provided by the Knowledge Media Institute at The Open University. The dataset, called the Open University Learning Analytics dataset, contains six files (all in csv format) and are linked by students’ ID numbers. The dataset contains data about courses, students and their interactions with the university’s Virtual Learning Environment (VLE). The dataset may be accessed here.

We use four of the six files and narrow down the data to the February 2013 offering (2013J) of the “BBB” module. The outcome variable i our analysis is the student’s final result. In the dataset, the final result is reported as either withdraw, fail, pass or distinction. However, since our interest is in predicting whether a student successfully completes the module or not, we group withdraw and fail as "Not Completed" and pass and distinction as "Completed". This simplifies the interpretation of our model and fulfills the objectives of this report.

The *studentInfo* file has demographic information for each student and their final result in the module. As independent variables in our analysis, we use the student's gender, highest education, age, disability, previous attempts of this module and number of current credits enrolled in. The final result is also found in this file and is used as our outcome variable.

The *studentRegistration* file has the day of registration/unregistration for each student. A day of registration of 0 refers to the starts date of the module. We use the day of registration as an independent variable.

The *studentAssessment* file has the results of each assessment taken during the module. We use the result of the first tutor-marked assignment (TMA) as an independent variable. This is the first major assignment in the module and could be used as an early indicator of student success or lack thereof.

The *studentVLE* file has student virtual interaction information during the module. We use the number of clicks the student performs in the virtual environment for the module as an independent variable.

The total number of students used in this analysis is 1675 students. Distributions of the categorical variables are shown in Figure 1 and distributions of the continuous variables are shown in Figure 2.

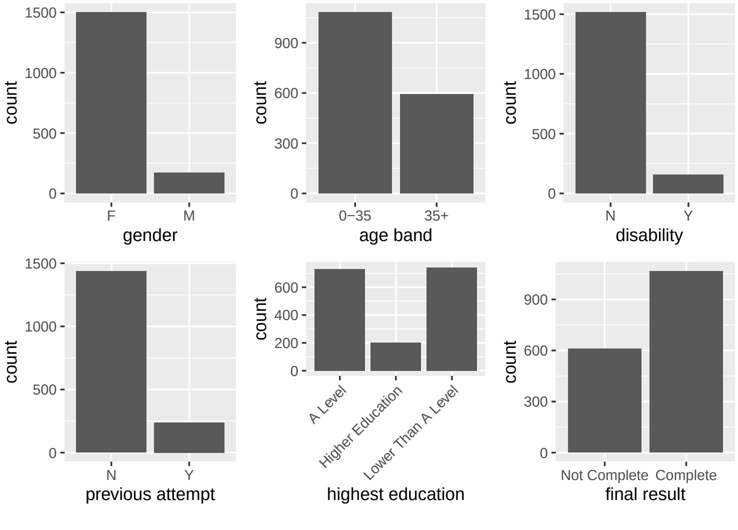

Figure 1: Distribution of categorical variables

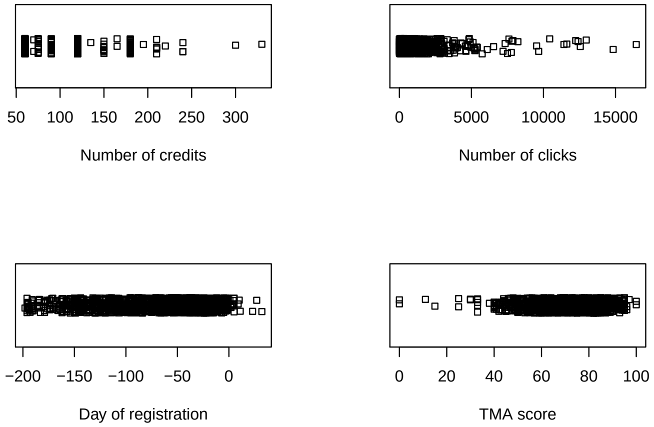

Figure 2: Distribution of continuous variables

Statistical Analysis

The analysis uses multiple logistic regression to predict a binary categorical variable from a combination of nine continuous and categorical independent variables. We summarize our initial model as:

$$logit[P(Y=1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 c_2 + \beta_{10} c_3$$

x_1 = number of current credits enrolled
x_2 = day of registration
x_3 = score for first TMA assignment
x_4 = total number of clicks during the module
x_5 = gender of student (Female = 0/Male = 1)
x_6 = age band of student (0-35 = 0/35+ = 1)
x_7 = disability (No = 0/ Yes = 1)
x_8 = previously attempted this module (No = 0/ Yes = 1)
c_2 = highest education (Lower than A Level = 0/ A Levels = 1)
c_3 = highest education (Lower than A Level = 0/ Higher than A Levels = 1)

The β_j parameters represent the effect of x_j on the log odds that $Y = 1$ (a student successfully completes the module), while holding all other x values constant.

Model Assumptions

All inferences are conducted using $\alpha = 0.05$ unless stated otherwise. No adjustments for multiplicity are made as this is an exploratory analysis. Categorical variables are summarized with proportions and frequencies. Continuous variables are summarized using the following statistics: mean, median, standard deviation, minimum, maximum and quantiles.

The null hypothesis is that there is no association between the outcome variable and independent variables:

$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$

That is, the probability of completing the module is independent of all the nine variables.

All statistical analyses in this document were performed using R version 4.0.3 (2020-10-10). R packages used were maintained using the packrat dependency management system.

Objective Analysis

Variable Diagnostics

We first examine the distribution of the five categorical prediction variables: gender, age, disability, previous attempts, previous education.

Tables 1-5 show completion rates for all categorical variables. For gender, males and females complete the module at 65.3% and 63.4%, respectively. For age bands, students over the age of 35 complete the module at 69.6%, above students 35 and under at 60.3%. Students with a disability have a lower completion rate at 56.3% compared to students without a disability at 64.3%. Students who have attempted the module before have a completion rate of 46.0%, much lower than students who have not attempted the module at 66.5%. Students with a lower than A Level education have lower completion rates at 53.9% compared to higher educated students.

Table 1: Gender Proportions

	Not Complete	Complete
F	0.366	0.634
M	0.347	0.653

Table 2: Age Band Proportion

	Not Complete	Complete
0-35	0.397	0.603
35+	0.304	0.696

Table 3: Disability Proportion

	Not Complete	Complete
N	0.357	0.643
Y	0.437	0.563

A generalized Cochran-Mantel-Haenszel test of association of each categorical independent variable and a student's final result was performed. The test shows significant nonzero correlation between a student's final result and age band, disability, previous attempts and previous education. There is no evidence of an association between final result and gender.

Turning to the continuous independent variables: total clicks, result on first TMA, day of registration and number of enrolled credits. We notice that students who completed the module had on average enrolled in 6.75 less credits, clicked 711.53 times more and scored 6.76 higher on their first TMA.

Table 4: Previous Attempts Proportion

	Not Complete	Complete
N	0.335	0.665
Y	0.540	0.460

Table 5: Highest Education Proportion

	Not Complete	Complete
A Level	0.289	0.711
Higher Education	0.282	0.718
Lower Than A Level	0.461	0.539

Model Selection

Fitting the initial multiple logistic regression model with all the variables using the glm() function in R, we are presented with this model:

Table 6: Estimated Coefficients - Full Model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.5845	0.395	-4.012	0.000
Gender: M	0.2000	0.188	1.063	0.288
Highest Education: Higher Education	-0.1240	0.198	-0.627	0.531
Highest Education: Lower Than A Level	-0.6414	0.124	-5.179	0.000
Age band: 35+	0.0442	0.128	0.347	0.729
Previous Attempts: Y	-0.6307	0.162	-3.888	0.000
Studied Credits	-0.0087	0.002	-4.976	0.000
Disability: Y	-0.2228	0.199	-1.118	0.264
Registration Day	0.0004	0.001	0.338	0.736
TMA Score	0.0332	0.005	6.615	0.000
Total Clicks	0.0016	0.000	10.512	0.000

Table 6 shows that gender, age band, disability and registration day are not significant. This aligns well with our variable diagnostics and the Cochran-Mantel-Haenszel test of association. Highest education is significant for students with a lower than A Levels education but not for students with an education higher than A Levels.

We use a stepwise model selection process, using AIC to select a model in a backward manner. The process starts with all nine independent variables and removes one variable in each step until AIC is minimized. AIC measures how well the model fits the data by measuring a particular expected distance. Smaller AIC values are better. The model with all nine independent variables has a AIC of 1802. The stepwise model selection process shows that AIC is minimized at 1797 with five independent variables: highest education, previous attempts, number of enrolled credits, TMA score and total clicks. This agrees with our assessment of significance of the independent variables in the full model above.

Therefore, we fit the model with the five independent variables. Table 7 shows the coefficient estimates for each variable with all being significant at the 95% confidence interval.

Table 7: Estimated Coefficients - Reduced Model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.5613	0.386	-4.048	0.000
Highest Education: Higher Education	-0.0843	0.195	-0.432	0.666
Highest Education: Lower Than A Level	-0.6427	0.123	-5.210	0.000
Previous Attempts: Y	-0.6520	0.161	-4.056	0.000
Studied Credits	-0.0089	0.002	-5.097	0.000
TMA Score	0.0329	0.005	6.571	0.000
Total Clicks	0.0016	0.000	10.686	0.000

Model Checking

We perform a likelihood-ratio test for each independent variable, with results shown in Table 8. The LR statistic and associated p-values show strong evidence of association between each variable and a student's final result.

Table 8: Analysis of Deviance Table

	LR Chisq	Df	Pr(>Chisq)
Highest Education	29.23	2	0.0000
Previous Attempts: Y	16.59	1	0.0000
Studied Credits	26.42	1	0.0000
TMA Score	45.80	1	0.0000
Total Clicks	210.67	1	0.0000

A 95% confidence interval, shown in Table 9, shows no values of 0, further confirming evidence of association between the variables and a student's final result. As discussed earlier the higher than A Level category is not significant and has a 0 in its confidence interval.

Table 9: Confidence Interval

	2.5 %	97.5 %
Intercept	-2.3237	-0.8107
Highest Education: Higher Education	-0.4631	0.3023
Highest Education: Lower Than A Level	-0.8854	-0.4016
Previous Attempts: Y	-0.9686	-0.3379
Studied Credits	-0.0123	-0.0055
TMA Score	0.0232	0.0428
Total Clicks	0.0013	0.0019

Results

Having checked the model, we report the model equation as:

$$logit[P(Y=1)] = -1.5613 - 0.0088x_1 + 0.0328x_2 + 0.0016x_3 - 0.6520x_4 - 0.0842c_2 - 0.6427c_3$$

x_1 =number of current credits enrolled
x_2 =score for first TMA assignment
x_3 =total number of clicks during the module
x_4 =previously attempted this module (No = 0/ Yes = 1)
c_2 =highest education (A Levels = 0/ Higher than A Levels = 1)
c_3 =highest education (A Levels = 0/ Lower than A Level = 1)

Table 10 presents the marginal effects of each independent variable.

The estimated probability of a student completing the course is 0.12 lower when the student's education level is lower than A Levels. There is no evidence of probability of success being influenced by having an education higher than an A Levels.

The estimated probability of a student completing the module is 0.123 lower when the student is repeating the module.

An increase in 1-credit enrolled decreases the average rate of change in the estimated probability of completion by 0.0016. Given that the majority of modules at The Open University are 30 credits, an increase in 1-module registered would decrease the estimated probability of completion by 4.8%.

An increase in 1-percent scored on the first TMA increases the average rate of change in the estimated probability of completion by 0.0059. Therefore, an increase in the TMA score of 10% would increase the estimated probability of completion by 5.95%.

An increase in 1-click in the VLE increases the average rate of change in the estimated probability of completion by 0.00029. Given the module is presented for 268 days, an increase of 1-click per day will increase the estimated probability of completion by 7.77%.

Table 10: Marginal Effects

	dF/dx	Std. Err.	z	P>|z|
Highest Education: Higher Education	-0.01528	0.035	-0.431	0.667
Highest Education: Lower Than A Level	-0.11956	0.023	-5.200	0.000
Previous Attempts: Y	-0.12259	0.031	-3.997	0.000
Studied Credits	-0.00160	0.000	-4.891	0.000
TMA Score	0.00595	0.001	6.134	0.000
Total Clicks	0.00029	0.000	9.236	0.000

Discussion

Given the significance of the independent variables introduced in the previous section, we have a model that predicts whether a student will successfully complete the BBB module. To further summarize the model and to determine the model's fit, we investigate the predictive power of the model.

The classification table in Table 11 show the predictions of the model's binary outcome (Complete = 1, Not complete = 0) against the actual outcome. Table 12 uses a cutoff of

$\pi_0 = 0.64$, the sample proportion of y=1 outcomes, with probabilities over this amount indicating $\hat{y} = 1$.

Table 11: Classification Table - February 2013

	Not Complete	Complete
Not Complete	474	136
Complete	323	742

The estimated sensitivity of the model is $P(\hat{y} = 1|y = 1) = 0.697$ and the estimated specificity is $P(\hat{y} = 0|y = 0) = 0.777$. Since we are interested in studens predicted to not complete the module, we are more interested in the specificity of the model. 77.7% of the students that the model predicts will not complete the module, in fact will not complete the module. This high specificity might warrant further examination of this model in tracking student performance. In addition, a reasonably high specificity ensures that we won't "misdiagnose" many students as completing when they are not.

Further examining specificity and sensitivity, we plot the ROC curve which shows the specificity and sensitivity for all π_0 cutoff values, as seen in Figure 3. The bigger the area under the ROC curve, the stronger the predictive power of the model, with an area of 0.5 being equivalent to random prediction. The area under the curve of the ROC curve is 0.81, showing strong predictive power of the model.

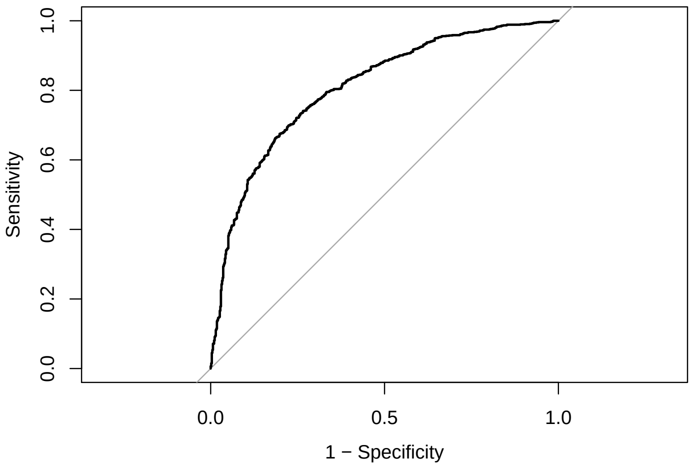

Figure 3: Receiver operating characteristic curve

The model was fit with data from the February 2013 offering of the "BBB" module. To test our model, we use data from the October 2013 offering of the same module. Table 12 shows the classification table of the predicted vs actual outcomes for the model for the October 2013 offering. The table uses a cutoff of $\pi_0 = 0.6$, the sample proportion of y=1 outcomes, with probabilities over this amount indicating $\hat{y} = 1$. The overall proportion of correct classifications is 0.69, indicating that 69% of predictions are correct.

Table 12: Classification Table - October 2013

	Not Complete	Complete
Not Complete	396	148
Complete	264	537

Limitations include the presence of a latent variable not collected in the dataset that strongly correlates with a student's final result. Also, the model holds only for the BBB module as factors of success could be different for each module. Changes in presentation style or instructors for the BBB module could affect the model predictions.

While the report limited the outcome to a binary "Complete" and "Not Complete", the dataset specifies the outcome as Fail/Withdraw/Pass/Distinction. A multicategory logit model, such as a cumulative logit model, could be used to predict varying levels of module completion. Also, a quasi-independence model could be used to observe a student's performance in another module for better observation of module completion when content is different. The dataset provides for this as it includes data for five other modules.

Appendix: R-code

```
library(knitr)
library(formatR)
library(stargazer)
library(xtable)
library(epiDisplay)
library("readxl")
library(ggplot2)
library(gridExtra)
library(plyr)
library(Hmisc)
library(MASS)
library(vcd)
library(dplyr)
library(psych)
library(leaps)
library(faraway)
library(car)
library(summarytools)
library(jtools)
library(vcdExtra)
library(VGAM)
library(pROC)
library(erer)
library(tidyverse)
library(caret)
library(bestglm)
library(mfx)
knitr::opts_chunk$set(echo = TRUE)
options(digits = 5, width = 60, xtable.comment = FALSE)
opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
out_type <- knitr::opts_knit$get("rmarkdown.pandoc.to")

clicks <- read.csv("studentVle.csv")
studentInfo <- read.csv("studentInfo.csv")
TMA <- read.csv("studentAssessment.csv")
register <- read.csv("studentRegistration.csv")

clicksbbb2013J <- subset(clicks, code_module == "BBB" & code_presentation ==
    "2013J", select = c(id_student, sum_click))
sumclicksbbb2013J <- summarise_at(group_by(clicksbbb2013J, id_student),
    vars(sum_click), funs(sum(., na.rm = TRUE)))
names(sumclicksbbb2013J) <- c("id_student", "sum_click")
```

```r
studentDembbb2013J <- subset(studentInfo, code_module == "BBB" &
    code_presentation == "2013J", select = c(id_student, gender,
    highest_education, age_band, num_of_prev_attempts, studied_credits,
    disability, date_registration, final_result))

TMAbbb2013J <- subset(TMA, id_assessment == 14996, select = c(id_student,
    score))
names(TMAbbb2013J) <- c("id_student", "TMA")

registerbbb2013J <- subset(register, code_module == "BBB" & code_presentation ==
    "2013J" & (date_unregistration > 30 | is.na(date_unregistration)),
    select = c("id_student"))

bbb2013J <- merge(registerbbb2013J, studentDembbb2013J, by = "id_student")
bbb2013J <- merge(bbb2013J, TMAbbb2013J, by = "id_student")
bbb2013J <- merge(bbb2013J, sumclicksbbb2013J, by = "id_student")
bbb2013J <- bbb2013J[, 2:11]

bbb2013J$final_result <- sub("Withdrawn", "Not Complete", bbb2013J$final_result)
bbb2013J$final_result <- sub("Fail", "Not Complete", bbb2013J$final_result)
bbb2013J$final_result <- sub("Distinction", "Complete", bbb2013J$final_result)
bbb2013J$final_result <- sub("Pass", "Complete", bbb2013J$final_result)

bbb2013J$highest_education <- sub("No Formal quals", "Lower Than A Level",
    bbb2013J$highest_education)
bbb2013J$highest_education <- sub("Lower Than A Level", "Lower Than A Level",
    bbb2013J$highest_education)
bbb2013J$highest_education <- sub("A Level or Equivalent", "A Level",
    bbb2013J$highest_education)
bbb2013J$highest_education <- sub("HE Qualification", "Higher Education",
    bbb2013J$highest_education)
bbb2013J$highest_education <- sub("Post Graduate Qualification",
    "Higher Education", bbb2013J$highest_education)

bbb2013J$age_band <- sub("55<=", "35+", bbb2013J$age_band)
bbb2013J$age_band <- sub("35-55", "35+", bbb2013J$age_band)
bbb2013J$age_band <- sub("0-35", "0-35", bbb2013J$age_band)

bbb2013J$num_of_prev_attempts <- sub("1", "Y", bbb2013J$num_of_prev_attempts)
bbb2013J$num_of_prev_attempts <- sub("2", "Y", bbb2013J$num_of_prev_attempts)
bbb2013J$num_of_prev_attempts <- sub("3", "Y", bbb2013J$num_of_prev_attempts)
bbb2013J$num_of_prev_attempts <- sub("4", "Y", bbb2013J$num_of_prev_attempts)
bbb2013J$num_of_prev_attempts <- sub("5", "Y", bbb2013J$num_of_prev_attempts)
bbb2013J$num_of_prev_attempts <- sub("0", "N", bbb2013J$num_of_prev_attempts)
```

```r
# reorder level
bbb2013J$highest_education <- factor(bbb2013J$highest_education)
bbb2013J$final_result <- ordered(factor(bbb2013J$final_result),
    levels = c("Not Complete", "Complete"))
bbb2013J$num_of_prev_attempts <- factor(bbb2013J$num_of_prev_attempts)
bbb2013J$age_band <- factor(bbb2013J$age_band)
bbb2013J$disability <- factor(bbb2013J$disability)
bbb2013J$gender <- factor(bbb2013J$gender)

plot1 <- ggplot(bbb2013J) + geom_bar(aes(x = gender))
plot2 <- ggplot(bbb2013J) + geom_bar(aes(x = age_band)) + labs(x = "age band")
plot3 <- ggplot(bbb2013J) + geom_bar(aes(x = disability))
plot4 <- ggplot(bbb2013J) + geom_bar(aes(x = num_of_prev_attempts)) +
    labs(x = "previous attempt")
plot5 <- ggplot(bbb2013J) + geom_bar(aes(x = highest_education)) +
    labs(x = "highest education") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1))
plot6 <- ggplot(bbb2013J) + geom_bar(aes(x = final_result)) +
    labs(x = "final result")
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, ncol = 3)

par(mfrow = c(2, 2))
stripchart(bbb2013J$studied_credits, method = "jitter", xlab = "Number of credits")
stripchart(bbb2013J$sum_click, method = "jitter", xlab = "Number of clicks")
stripchart(bbb2013J$date_registration, method = "jitter", xlab = "Day of registration")
stripchart(bbb2013J$TMA, method = "jitter", xlab = "TMA score")

xtable(prop.table(xtabs(~gender + final_result, data = bbb2013J),
    1), caption = "Gender Proportions", digits = c(0, 3, 3))
xtable(prop.table(xtabs(~age_band + final_result, data = bbb2013J),
    1), caption = "Age Band Proportion", digits = c(0, 3, 3))
xtable(prop.table(xtabs(~disability + final_result, data = bbb2013J),
    1), caption = "Disability Proportion", digits = c(0, 3, 3))
xtable(prop.table(xtabs(~num_of_prev_attempts + final_result,
    data = bbb2013J), 1), caption = "Previous Attempts Proportion",
    digits = c(0, 3, 3))
xtable(prop.table(xtabs(~highest_education + final_result, data = bbb2013J),
    1), caption = "Highest Education Proportion", digits = c(0,
    3, 3))

GenderResult <- as.matrix(xtabs(~gender + final_result, data = bbb2013J))
CMHtest(GenderResult)
AgeResult <- as.matrix(xtabs(~age_band + final_result, data = bbb2013J))
CMHtest(AgeResult)
```

```r
DisabilityResult <- as.matrix(xtabs(~disability + final_result,
    data = bbb2013J))
CMHtest(DisabilityResult)
AttemptsResult <- as.matrix(xtabs(~num_of_prev_attempts + final_result,
    data = bbb2013J))
CMHtest(AttemptsResult)
HEResult <- as.matrix(xtabs(~highest_education + final_result,
    data = bbb2013J))
CMHtest(HEResult)

clM <- glm(final_result ~ ., family = binomial, data = bbb2013J)
clMcf <- summary(clM)$coefficients
rownames(clMcf) <- c("Intercept", "Gender: M", "Highest Education: Higher Education",
    "Highest Education: Lower Than A Level", "Age band: 35+",
    "Previous Attempts: Y", "Studied Credits", "Disability: Y",
    "Registration Day", "TMA Score", "Total Clicks")
xtable(clMcf, caption = "Estimated Coefficients - Full Model",
    digits = c(0, 4, 3, 3, 3))

stepAIC(clM)

cl2 <- glm(formula = final_result ~ highest_education + num_of_prev_attempts +
    studied_credits + TMA + sum_click, family = binomial, data = bbb2013J)
cl2cf <- summary(cl2)$coefficients
rownames(cl2cf) <- c("Intercept", "Highest Education: Higher Education",
    "Highest Education: Lower Than A Level", "Previous Attempts: Y",
    "Studied Credits", "TMA Score", "Total Clicks")
xtable(cl2cf, captions = "Estimated Coefficients - Reduced Model",
    digits = c(0, 4, 3, 3, 3))

cl2an <- Anova(cl2)
rownames(cl2an) <- c("Highest Education", "Previous Attempts: Y",
    "Studied Credits", "TMA Score", "Total Clicks")
xtable(cl2an, caption = "Analysis of Deviance Table")

cl2ci <- confint(cl2)
rownames(cl2ci) <- c("Intercept", "Highest Education: Higher Education",
    "Highest Education: Lower Than A Level", "Previous Attempts: Y",
    "Studied Credits", "TMA Score", "Total Clicks")
xtable(cl2ci, caption = "Confidence Interval", digits = c(0,
    4, 4))

lo <- logitmfx(cl2, atmean = FALSE, data = bbb2013J)
cl2log <- lo$mfxest
```

```
rownames(cl2log) <- c("Highest Education: Higher Education",
    "Highest Education: Lower Than A Level", "Previous Attempts: Y",
    "Studied Credits", "TMA Score", "Total Clicks")
xtable(cl2log, caption = "Marginal Effects", digits = c(0, 5,
    3, 3, 3))

prop <- sum(bbb2013J$final_result == "Complete")/nrow(bbb2013J)

fit <- glm(formula = final_result ~ highest_education + num_of_prev_attempts +
    studied_credits + TMA + sum_click, family = binomial, data = bbb2013J)

predicted <- as.numeric(fitted(fit) > prop)

cl2class1 <- xtabs(~bbb2013J$final_result + predicted)
colnames(cl2class1) <- c("Not Complete", "Complete")
xtable(cl2class1, caption = "Classification Table - February 2013")

rocplot <- roc(final_result ~ fitted(cl2), data = bbb2013J)
plot.roc(rocplot, legacy.axes = TRUE)

auc(rocplot)

clicks_bbb2013B <- subset(clicks, code_module == "BBB" & code_presentation ==
    "2013B", select = c(id_student, sum_click))
sumclicks_bbb2013B <- summarise_at(group_by(clicks_bbb2013B,
    id_student), vars(sum_click), funs(sum(., na.rm = TRUE)))
names(sumclicks_bbb2013B) <- c("id_student", "sum_click")

studentDem_bbb2013B <- subset(studentInfo, code_module == "BBB" &
    code_presentation == "2013B", select = c(id_student, gender,
    highest_education, age_band, num_of_prev_attempts, studied_credits,
    disability, date_registration, final_result))

tma_bbb2013B <- subset(TMA, id_assessment == 14984, select = c(id_student,
    score))
names(tma_bbb2013B) <- c("id_student", "TMA")

register_bbb2013B <- subset(register, code_module == "BBB" &
    code_presentation == "2013B" & (date_unregistration > 30 |
    is.na(date_unregistration)), select = c("id_student"))

BBB2013B <- merge(register_bbb2013B, studentDem_bbb2013B, by = "id_student")
BBB2013B <- merge(BBB2013B, tma_bbb2013B, by = "id_student")
BBB2013B <- merge(BBB2013B, sumclicks_bbb2013B, by = "id_student")
```

```
BBB2013B <- BBB2013B[, 2:11]

BBB2013B$final_result <- sub("Withdrawn", "Not Complete", BBB2013B$final_result)
BBB2013B$final_result <- sub("Fail", "Not Complete", BBB2013B$final_result)
BBB2013B$final_result <- sub("Distinction", "Complete", BBB2013B$final_result)
BBB2013B$final_result <- sub("Pass", "Complete", BBB2013B$final_result)

BBB2013B$highest_education <- sub("No Formal quals", "Lower Than A Level",
    BBB2013B$highest_education)
BBB2013B$highest_education <- sub("Lower Than A Level", "Lower Than A Level",
    BBB2013B$highest_education)
BBB2013B$highest_education <- sub("A Level or Equivalent", "A Level",
    BBB2013B$highest_education)
BBB2013B$highest_education <- sub("HE Qualification", "Higher Education",
    BBB2013B$highest_education)
BBB2013B$highest_education <- sub("Post Graduate Qualification",
    "Higher Education", BBB2013B$highest_education)

BBB2013B$age_band <- sub("55<=", "35+", BBB2013B$age_band)
BBB2013B$age_band <- sub("35-55", "35+", BBB2013B$age_band)
BBB2013B$age_band <- sub("0-35", "0-35", BBB2013B$age_band)

BBB2013B$num_of_prev_attempts <- sub("1", "Y", BBB2013B$num_of_prev_attempts)
BBB2013B$num_of_prev_attempts <- sub("2", "Y", BBB2013B$num_of_prev_attempts)
BBB2013B$num_of_prev_attempts <- sub("3", "Y", BBB2013B$num_of_prev_attempts)
BBB2013B$num_of_prev_attempts <- sub("4", "Y", BBB2013B$num_of_prev_attempts)
BBB2013B$num_of_prev_attempts <- sub("5", "Y", BBB2013B$num_of_prev_attempts)
BBB2013B$num_of_prev_attempts <- sub("0", "N", BBB2013B$num_of_prev_attempts)

# reorder level
BBB2013B$highest_education <- factor(BBB2013B$highest_education,
    levels = c("Lower Than A Level", "A Level", "Higher Education"))
BBB2013B$final_result <- ordered(factor(BBB2013B$final_result),
    levels = c("Not Complete", "Complete"))
BBB2013B$num_of_prev_attempts <- factor(BBB2013B$num_of_prev_attempts)
BBB2013B$age_band <- factor(BBB2013B$age_band)
BBB2013B$disability <- factor(BBB2013B$disability)
BBB2013B$gender <- factor(BBB2013B$gender)

prop <- sum(BBB2013B$final_result == "Complete")/nrow(BBB2013B)

fit <- glm(formula = final_result ~ highest_education + num_of_prev_attempts +
    studied_credits + TMA + sum_click, family = binomial, data = BBB2013B)
```

```r
predicted <- as.numeric(fitted(fit) > prop)

cl2class2 <- xtabs(~BBB2013B$final_result + predicted)
colnames(cl2class2) <- c("Not Complete", "Complete")

xtable(cl2class2, caption = "Classification Table - October 2013")
```

References

Agresti, Alan (2019), An Introduction to Categorical Data Analysis, 3rd ed., John Wiley & Sons, Inc.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2014), Applied Linear Statistical Models, 5th ed., McGraw-Hill Irwin.

Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2010). Handbook of educational data mining. CRC press.