Final Project COMP 4442

Michael Wise/Jairus Martinez/Sarah Buckingham

2024-04-30

```
# Libraries
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(vcd)
```

```
## Loading required package: grid
```

```
library(nnet)
library(caret)
```

```
## Loading required package: lattice
```

Introduction

UCI ML Repo: Predicting Student's Dropout and Academic Success

In our project, we will demonstrate the use of Multinomial Logistic Regression (MLR) using a dataset from the UCI Machine Learning Repository. The dataset classifies student outcomes into three categories: dropout, enrolled, and graduate. We will walk through the entire process of data preparation, model fitting, evaluation, and interpretation of the results.

Learning Objectives

1. **Understanding key terms and concepts**: Students should be able to recall the definition of multinomial logistic regression, differentiate it from other types of logistic regression, and understand the necessary dataset structure for this analysis.
2. **Recognize when multinomial logistic regression is appropriate**: Students will understand what types of problems multinomial logistic regression solves, when it is appropriate, and what assumptions must be met.
3. **Application of MLR**: Students will learn to apply multinomial logistic regression to predict categorical outcomes with more than two levels.

4. **Interpretation of the model's output**: Students will learn to interpret the results, including coefficients and the confusion matrix, and make informed conclusions about the relationships between predictors and the outcome.

R Code Implementation

Using Multiclass Logistic Regression to Predict Student Outcomes

Regressions in general are statistical methods that aim to understand the relationship between independent variables and dependent variables. Often, this method is used to make predictions based off that learned relationship. In traditional regression (like linear regression), the dependent variable to predict is continuous. In logistic regression, the dependent variable to predict is binary and categorical. A special case of logistic regression is multinomial logistic regression.

Multinomial logistic regression is used to solve problems that involve predicting categorical outcomes that have more than two categories. Many different research questions investigate the relationship between multiple predictors and multiple, categorical outcome variables. Some examples include predicting which political candidates will get elected to office or which diseases a patient is predicted to have. While multinomial logistic regression can be used for prediction, it also gives insight into the relationships between predictor variables and the dependent variables, making it suitable for inference.

The data from the provided student_success.csv file is designed to address the problem of academic success in higher education. With its 4424 instances representing individual students and 35 features covering various aspects like college application attributes, course/academic attributes, demographics, and socio-economic factors, it provides a rich source of information for predicting students' outcomes. The dependent outcome variables we will be predicting fit into three classification categories - dropout, enrolled, graduate.

Therefore, we can use multinomial logistic regression to:

1. Predict 3 categorical outcome variables (dropout, enrolled, graduate)
2. Get insight into how the different features affect the likelihood of dropping out, enrolling, or graduating (inference)

Data Preparation

First, we load the dataset and check its structure and for any missing values.

```
# Reads CSV file into a data frame
academic_success <- read.csv("student_success.csv")

# Structure of the data frame
str(academic_success) # We may want to trim this down or only look at a bit of this for our final code
```

```
## 'data.frame':    4424 obs. of  35 variables:
##  $ Marital.status                                : int  1 1 1 1 2 2 1 1 1 1 ...
##  $ Application.mode                              : int  8 6 1 8 12 12 1 9 1 1 ...
##  $ Application.order                             : int  5 1 5 2 1 1 1 4 3 1 ...
##  $ Course                                        : int  2 11 5 15 3 17 12 11 10 10 ...
##  $ Daytime.evening.attendance                    : int  1 1 1 1 0 0 1 1 1 1 ...
##  $ Previous.qualification                        : int  1 1 1 1 1 12 1 1 1 1 ...
##  $ Nacionality                                   : int  1 1 1 1 1 1 1 1 15 1 ...
##  $ Mother.s.qualification                        : int  13 1 22 23 22 22 13 22 1 1 ...
##  $ Father.s.qualification                        : int  10 3 27 27 28 27 28 27 1 14 ...
##  $ Mother.s.occupation                           : int  6 4 10 6 10 10 8 10 10 5 ...
##  $ Father.s.occupation                           : int  10 4 10 4 10 8 11 10 10 8 ...
##  $ Displaced                                     : int  1 1 1 1 0 0 1 1 0 1 ...
##  $ Educational.special.needs                     : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ Debtor                                        : int  0 0 0 0 0 1 0 0 0 1 ...
##  $ Tuition.fees.up.to.date                       : int  1 0 0 1 1 1 1 0 1 0 ...
##  $ Gender                                        : int  1 1 1 0 0 1 0 1 0 0 ...
##  $ Scholarship.holder                            : int  0 0 0 0 0 0 1 0 1 0 ...
##  $ Age.at.enrollment                             : int  20 19 19 20 45 50 18 22 21 18 ...
##  $ International                                 : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Curricular.units.1st.sem..credited.           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Curricular.units.1st.sem..enrolled.           : int  0 6 6 6 6 5 7 5 6 6 ...
##  $ Curricular.units.1st.sem..evaluations.        : int  0 6 0 8 9 10 9 5 8 9 ...
##  $ Curricular.units.1st.sem..approved.           : int  0 6 0 6 5 5 7 0 6 5 ...
##  $ Curricular.units.1st.sem..grade.              : num  0 14 0 13.4 12.3 ...
##  $ Curricular.units.1st.sem..without.evaluations.: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Curricular.units.2nd.sem..credited.           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Curricular.units.2nd.sem..enrolled.           : int  0 6 6 6 6 5 8 5 6 6 ...
##  $ Curricular.units.2nd.sem..evaluations.        : int  0 6 0 10 6 17 8 5 7 14 ...
##  $ Curricular.units.2nd.sem..approved.           : int  0 6 0 5 6 5 8 0 6 2 ...
##  $ Curricular.units.2nd.sem..grade.              : num  0 13.7 0 12.4 13 ...
##  $ Curricular.units.2nd.sem..without.evaluations.: int  0 0 0 0 0 5 0 0 0 0 ...
##  $ Unemployment.rate                             : num  10.8 13.9 10.8 9.4 13.9 16.2 15.5 15.5 16.2 8
##  $ Inflation.rate                                : num  1.4 -0.3 1.4 -0.8 -0.3 0.3 2.8 2.8 0.3 1.4 .
##  $ GDP                                           : num  1.74 0.79 1.74 -3.12 0.79 -0.92 -4.06 -4.06 -
##  $ Target                                        : chr  "Dropout" "Graduate" "Dropout" "Graduate" ..
```

The dataset comprises 4424 observations and 35 variables, including character (factors, although not of the factor type) and numerical variables. The outcome variable, "Target," has three levels: "Dropout," "Enrolled," and "Graduate."

We will check for any nulls in the data:

```
# Check for missing values in specific columns
colSums(is.na(academic_success))
```

```
##                                  Marital.status
##                                               0
##                                Application.mode
##                                               0
##                               Application.order
##                                               0
##                                          Course
##                                               0
##                       Daytime.evening.attendance
##                                               0
##                           Previous.qualification
##                                               0
##                                     Nacionality
##                                               0
##                           Mother.s.qualification
##                                               0
##                           Father.s.qualification
##                                               0
##                              Mother.s.occupation
##                                               0
##                              Father.s.occupation
##                                               0
##                                       Displaced
##                                               0
```

```
##                       Educational.special.needs
##                                               0
##                                          Debtor
##                                               0
##                         Tuition.fees.up.to.date
##                                               0
##                                          Gender
##                                               0
##                              Scholarship.holder
##                                               0
##                               Age.at.enrollment
##                                               0
##                                   International
##                                               0
##            Curricular.units.1st.sem..credited.
##                                               0
##            Curricular.units.1st.sem..enrolled.
##                                               0
##         Curricular.units.1st.sem..evaluations.
##                                               0
##            Curricular.units.1st.sem..approved.
##                                               0
##               Curricular.units.1st.sem..grade.
##                                               0
## Curricular.units.1st.sem..without.evaluations.
##                                               0
##            Curricular.units.2nd.sem..credited.
##                                               0
##            Curricular.units.2nd.sem..enrolled.
##                                               0
##         Curricular.units.2nd.sem..evaluations.
##                                               0
##            Curricular.units.2nd.sem..approved.
##                                               0
##               Curricular.units.2nd.sem..grade.
##                                               0
## Curricular.units.2nd.sem..without.evaluations.
##                                               0
##                               Unemployment.rate
##                                               0
##                                  Inflation.rate
##                                               0
##                                             GDP
##                                               0
##                                          Target
##                                               0
```

```r
# Trim possible leading/trailing whitespace from all character data
academic_success[] <- lapply(academic_success, function(x) if(is.character(x)) trimws(x) else x)
```

Next, we ensure that categorical variables are correctly formatted as factors.

```r
# Converts necessary variables to factor type; convert categorical columns from numeric to factor
categorical_columns <- c("Marital.status", "Application.mode", "Daytime.evening.attendance",
                         "Displaced", "Educational.special.needs", "Debtor",
```

```
                    "Tuition.fees.up.to.date", "Gender", "Scholarship.holder", "International")

# Ensure correct data type and then convert to factors
academic_success[categorical_columns] <- lapply(academic_success[categorical_columns], function(x) facto

# Convert 'Target' to factor
academic_success$Target <- factor(academic_success$Target)

# Check the structure
str(academic_success)
```

```
## 'data.frame':    4424 obs. of  35 variables:
##  $ Marital.status                                : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 2 1
##  $ Application.mode                              : Factor w/ 18 levels "1","2","3","4",..: 8 6 1 8 1
##  $ Application.order                             : int  5 1 5 2 1 1 1 4 3 1 ...
##  $ Course                                        : int  2 11 5 15 3 17 12 11 10 10 ...
##  $ Daytime.evening.attendance                    : Factor w/ 2 levels "0","1": 2 2 2 2 1 1 2 2 2 2 .
##  $ Previous.qualification                        : int  1 1 1 1 1 12 1 1 1 1 ...
##  $ Nacionality                                   : int  1 1 1 1 1 1 1 1 15 1 ...
##  $ Mother.s.qualification                        : int  13 1 22 23 22 22 13 22 1 1 ...
##  $ Father.s.qualification                        : int  10 3 27 27 28 27 28 27 1 14 ...
##  $ Mother.s.occupation                           : int  6 4 10 6 10 10 8 10 10 5 ...
##  $ Father.s.occupation                           : int  10 4 10 4 10 8 11 10 10 8 ...
##  $ Displaced                                     : Factor w/ 2 levels "0","1": 2 2 2 2 1 1 2 2 1 2 .
##  $ Educational.special.needs                     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 .
##  $ Debtor                                        : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 2 .
##  $ Tuition.fees.up.to.date                       : Factor w/ 2 levels "0","1": 2 1 1 2 2 2 2 1 2 1 .
##  $ Gender                                        : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 1 2 1 1 .
##  $ Scholarship.holder                            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 2 1 .
##  $ Age.at.enrollment                             : int  20 19 19 20 45 50 18 22 21 18 ...
##  $ International                                 : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 .
##  $ Curricular.units.1st.sem..credited.           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Curricular.units.1st.sem..enrolled.           : int  0 6 6 6 6 5 7 5 6 6 ...
##  $ Curricular.units.1st.sem..evaluations.        : int  0 6 0 8 9 10 9 5 8 9 ...
##  $ Curricular.units.1st.sem..approved.           : int  0 6 0 6 5 5 7 0 6 5 ...
##  $ Curricular.units.1st.sem..grade.              : num  0 14 0 13.4 12.3 ...
##  $ Curricular.units.1st.sem..without.evaluations.: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Curricular.units.2nd.sem..credited.           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Curricular.units.2nd.sem..enrolled.           : int  0 6 6 6 6 5 8 5 6 6 ...
##  $ Curricular.units.2nd.sem..evaluations.        : int  0 6 0 10 6 17 8 5 7 14 ...
##  $ Curricular.units.2nd.sem..approved.           : int  0 6 0 5 6 5 8 0 6 2 ...
##  $ Curricular.units.2nd.sem..grade.              : num  0 13.7 0 12.4 13 ...
##  $ Curricular.units.2nd.sem..without.evaluations.: int  0 0 0 0 0 5 0 0 0 0 ...
##  $ Unemployment.rate                             : num  10.8 13.9 10.8 9.4 13.9 16.2 15.5 15.5 16.2 8
##  $ Inflation.rate                                : num  1.4 -0.3 1.4 -0.8 -0.3 0.3 2.8 2.8 0.3 1.4 .
##  $ GDP                                           : num  1.74 0.79 1.74 -3.12 0.79 -0.92 -4.06 -4.06 -
##  $ Target                                        : Factor w/ 3 levels "Dropout","Enrolled",..: 1 3 1
```

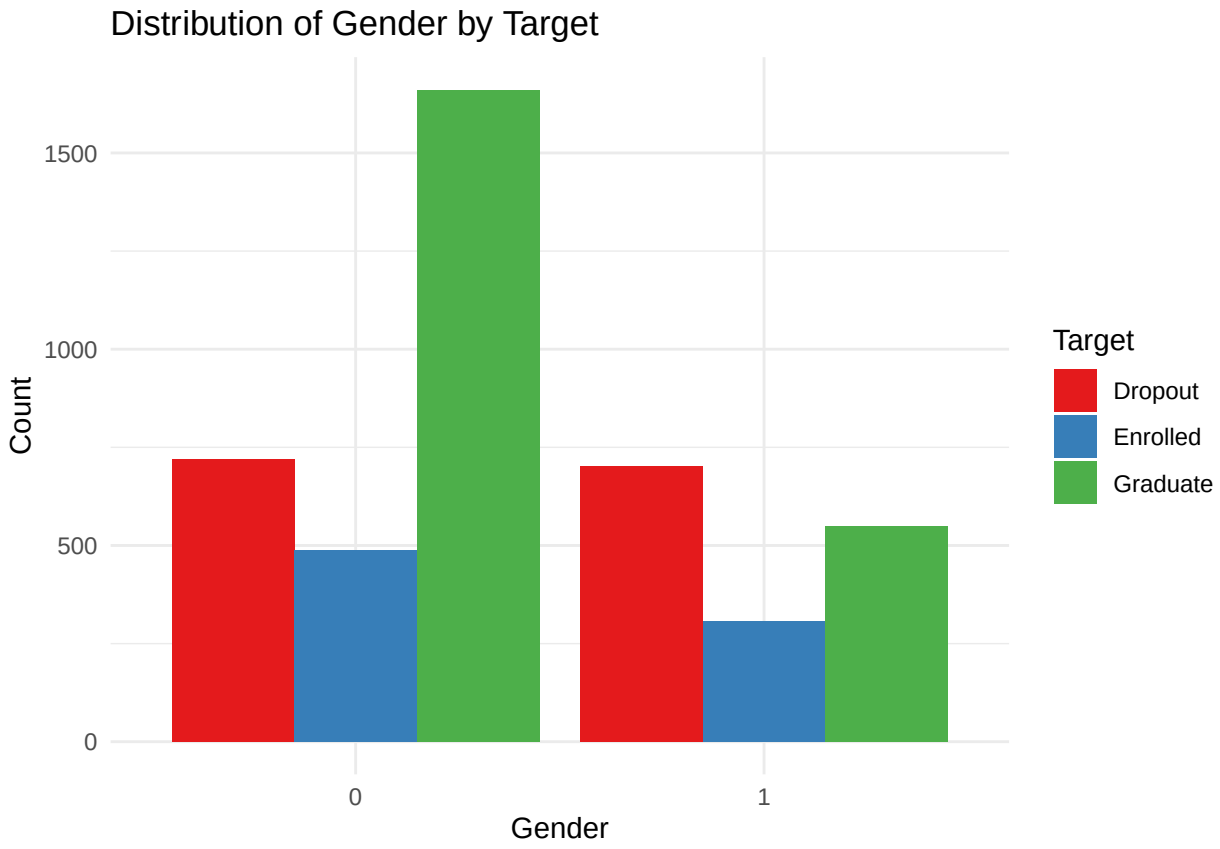
Exploratory Data Analysis (EDA)

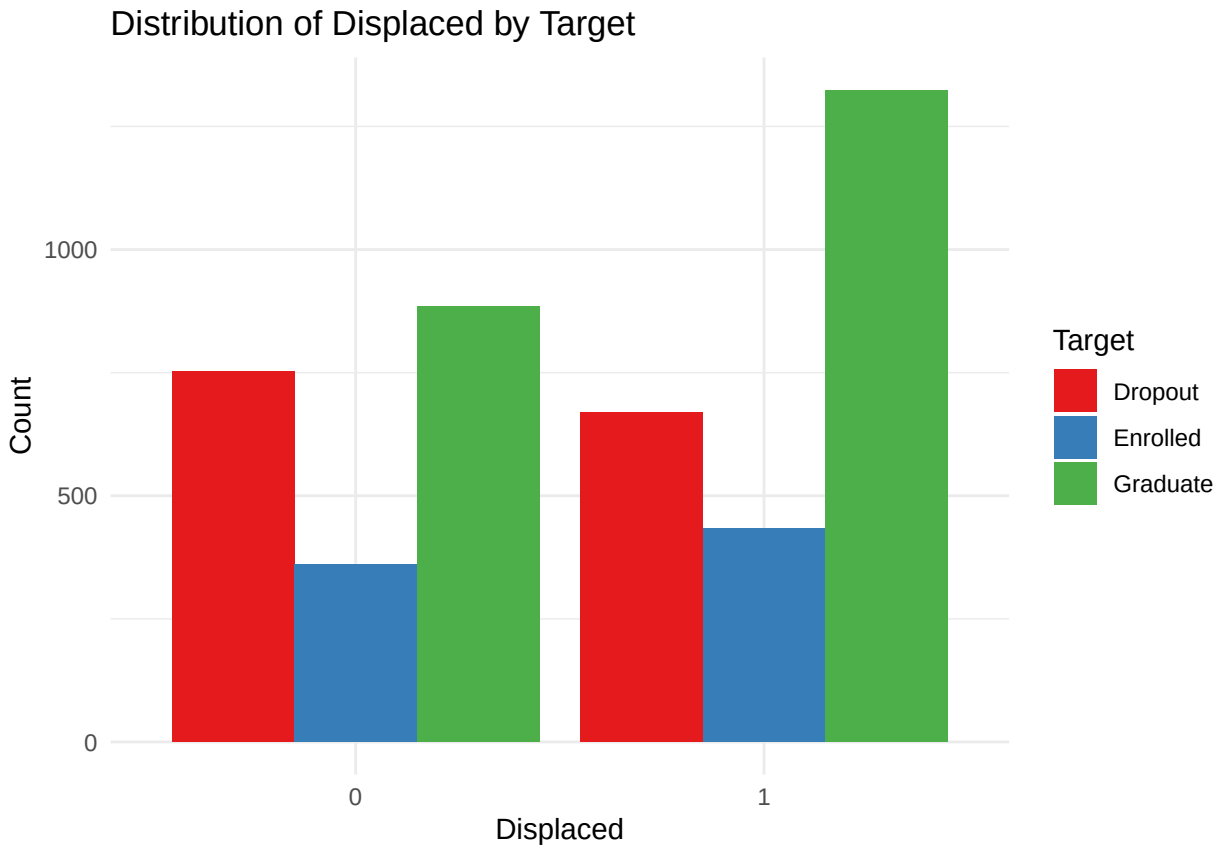
We will now create some visualizations to explore the distribution of various categorical and continuous variables by the target outcome.

```r
categorical_vars <- c("Gender", "Displaced", "Scholarship.holder")

# Loop through each categorical variable to create a side-by-side bar charts
for (var in categorical_vars) {
    p <- ggplot(academic_success, aes_string(x = var, fill = "Target")) +  # Use aes_string if using va
        geom_bar(stat = "count", position = "dodge") +  # position="dodge" makes the bars side by side
        labs(title = paste("Distribution of", var, "by Target"), x = var, y = "Count") +
        scale_fill_brewer(palette = "Set1") +  # Optional: uses a color palette for the fill
        theme_minimal()
    print(p)
}
```

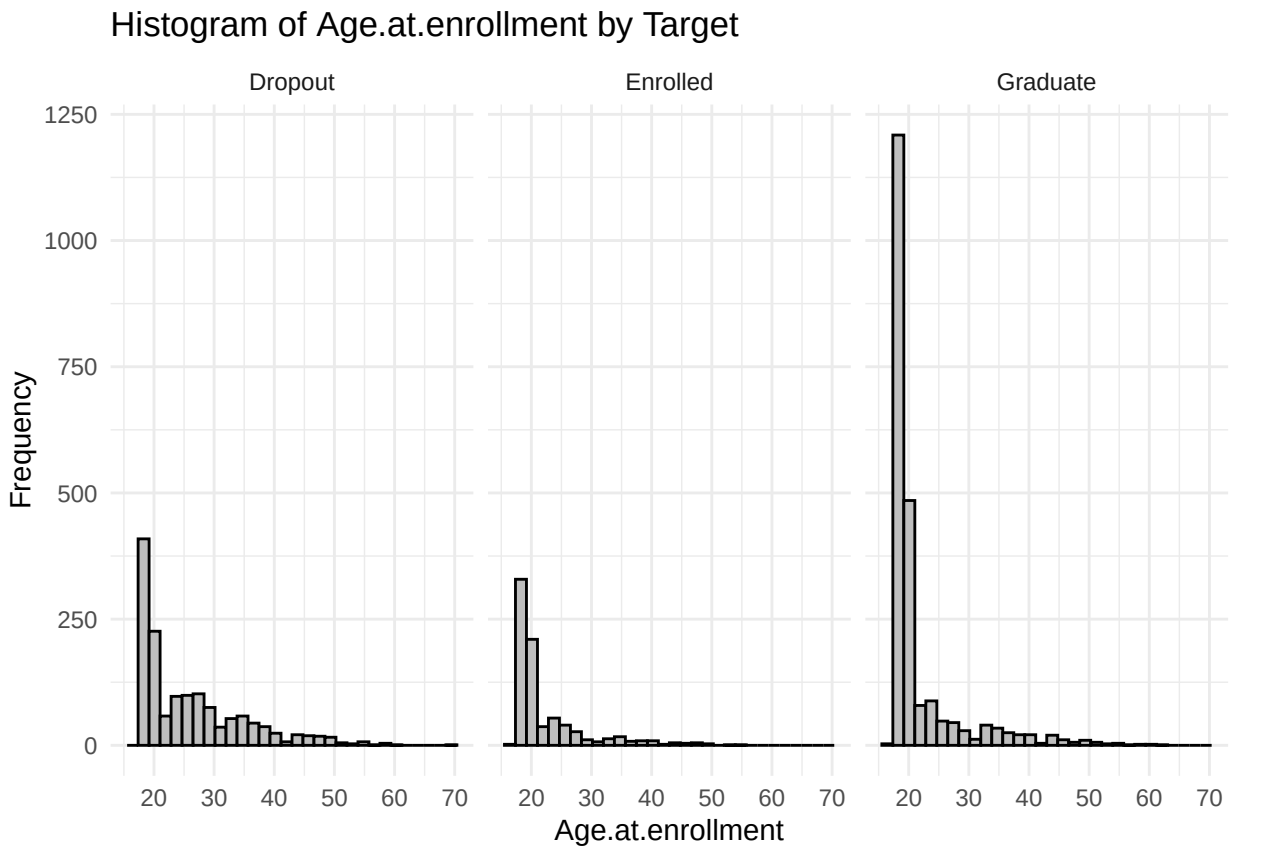
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

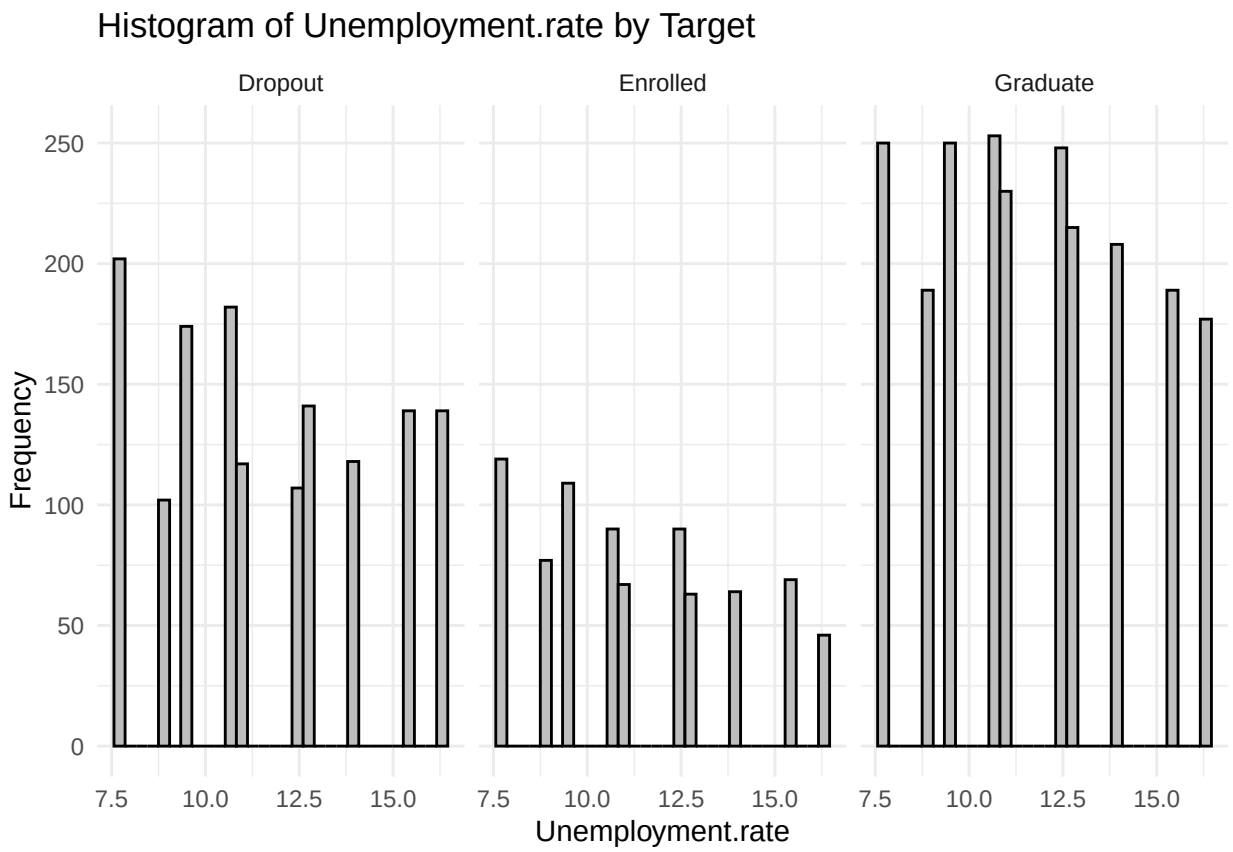


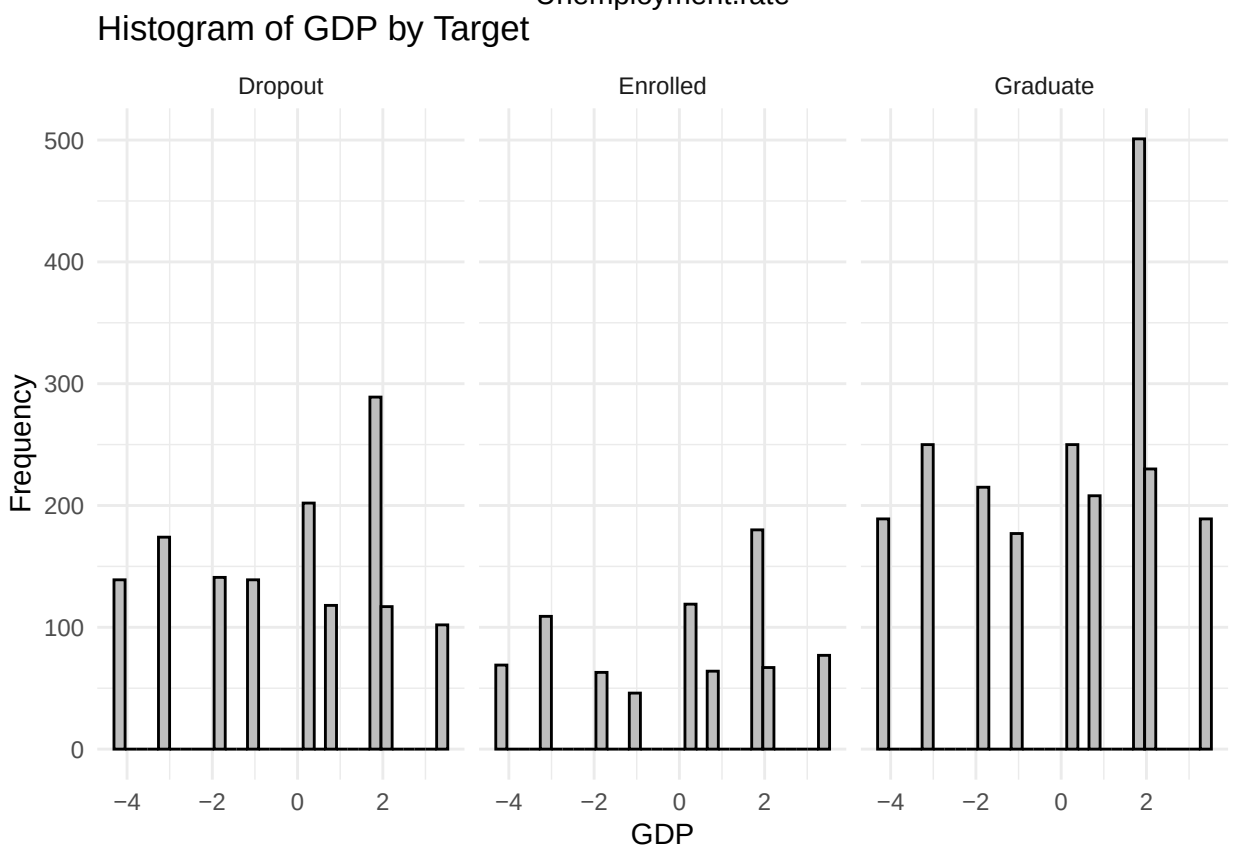

Distribution of Displaced by Target
Count
1000
500
0
0
1
Displaced
Target
Dropout
Enrolled
Graduate


Distribution of Scholarship.holder by Target
Count
1000
500
0
0
1
Scholarship.holder
Target
Dropout
Enrolled
Graduate

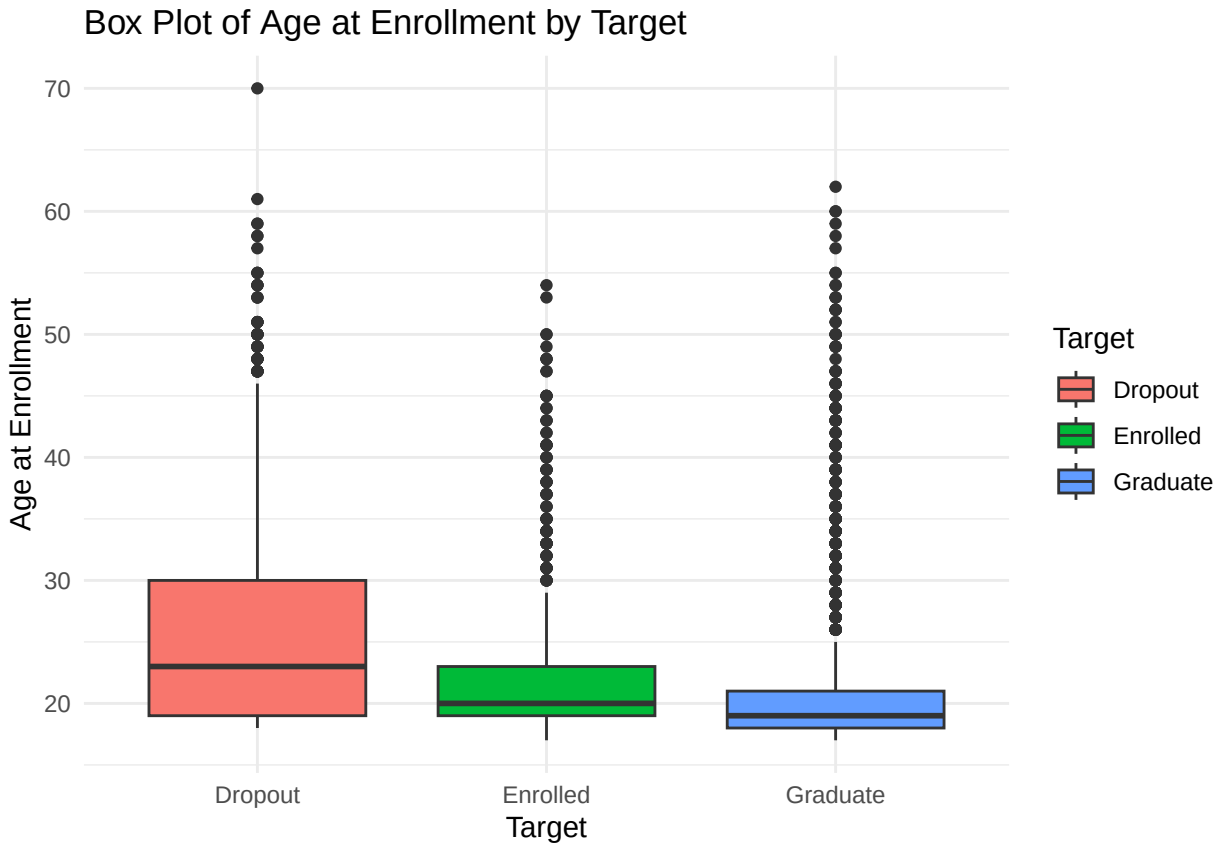
```r
# Histograms for Continuous Variables
continuous_vars <- c("Age.at.enrollment", "Unemployment.rate", "GDP")
for (var in continuous_vars) {
    p <- ggplot(academic_success, aes_string(x = var)) +
        geom_histogram(bins = 30, fill = "gray", color = "black") +
        facet_wrap(~Target) +
        labs(title = paste("Histogram of", var, "by Target"), x = var, y = "Frequency") +
        theme_minimal()
    print(p)
}
```




Histogram of Unemployment.rate by Target
Dropout
Enrolled
Graduate
Frequency
0
50
100
150
200
250
7.5
10.0
12.5
15.0
Unemployment.rate


Histogram of GDP by Target
Dropout
Enrolled
Graduate
Frequency
0
100
200
300
400
500
−4
−2
0
2
GDP

```
# Box Plot for 'Age.at.enrollment' by 'Target'
ggplot(academic_success, aes(x = Target, y = Age.at.enrollment, fill = Target)) +
    geom_boxplot() +
    labs(title = "Box Plot of Age at Enrollment by Target", x = "Target", y = "Age at Enrollment") +
    theme_minimal()
```




```
# Check if there are any issues with 'Target' or other variables before plotting
str(academic_success$Target)
```

```
##  Factor w/ 3 levels "Dropout","Enrolled",..: 1 3 1 3 3 3 3 1 3 1 ...
```

```
summary(academic_success$Target)
```

```
##  Dropout Enrolled Graduate
##     1421      794     2209
```

Categorical Variables:
- Gender (1 = male, 0 = female): The distribution of Gender by Target indicates variation primarily in graduation rates between genders.
- Displaced: The distribution of Displaced by Target indicates variation primarily in graduation rates between genders.
- Scholarship Holder: Our visual comparison of scholarship holders' outcomes shows a significant difference in the distributions of academic success.

Continuous Variables:
- Age at Enrollment: Histogram and box plots show that there appear to be some slight differences in the distribution and spread of age among different target categories.
- Unemployment Rate: The histogram illustrates the different distributions of unemployment across the target outcomes, which have similar shapes but different means.

- GDP: The histogram illustrates the different distributions of GDP across the target outcomes, particularly among the graduate group.

Multinomial Logistic Regression Assumptions to Consider:

1. **Dependent/response variable is categorical (3+)**
   - Yes. Dropout, enrolled, and graduate.
2. **Little or no multicollinearity between the predictor/explanatory variables**
   - Multicollinearity may exist, but we do not have strong evidence of it.
3. **Linear relationship of independent variables to log odds**
   - As our data is virtually ALL binary, categorical data, we cannot check this assumption for the majority of our data.
4. **Prefers large sample size**
   - Yes, the dataset is about ~4,4k rows long.
5. **Problem with extreme outliers**
   - None identified.
6. **Independent observations**
   - Yes, assuming independence of individual outcomes. It is reasonable to say that one student's success should not affect another's.

Multinominal Logistic Regression

We now can split the dataset into training and testing sets, fit the model, and evaluate its performance.

We will use a 70/30 train-test split for our model.

```
# Split the dataset into training and testing sets
set.seed(123)  # for reproducibility
train_index <- createDataPartition(academic_success$Target, p = 0.7, list = FALSE)
train_data <- academic_success[train_index, ]
test_data <- academic_success[-train_index, ]
```

```
# Fits the multinomial logistic regression model
model <- multinom(Target ~ ., data = train_data)
```

```
## # weights:  168 (110 variable)
## initial  value 3403.500870
## iter  10 value 2216.787487
## iter  20 value 2106.677282
## iter  30 value 2056.780581
## iter  40 value 1955.015713
## iter  50 value 1833.881161
## iter  60 value 1746.611589
## iter  70 value 1707.759619
## iter  80 value 1685.902086
## iter  90 value 1683.759433
## iter 100 value 1683.678285
## final  value 1683.678285
## stopped after 100 iterations
```

```
# Summary of the model
summary(model)
```

```
## Call:
## multinom(formula = Target ~ ., data = train_data)
##
## Coefficients:
```

```
##          (Intercept) Marital.status2 Marital.status3 Marital.status4
## Enrolled   0.6117102      0.14123780        2.061327       0.7126513
## Graduate   0.6282169      0.06669225        1.385022       0.7445702
##          Marital.status5 Marital.status6 Application.mode2 Application.mode3
## Enrolled      -0.3882554    -0.005482948        -2.9159462          2.318624
## Graduate      -0.3878728    -2.026409248        -0.7848863          1.350059
##          Application.mode4 Application.mode5 Application.mode6
## Enrolled        -1.5264665         -3.010398          1.845225
## Graduate        -0.3880277          3.480042          2.903569
##          Application.mode7 Application.mode8 Application.mode9
## Enrolled          1.876892        -0.1291556        -0.4408973
## Graduate          1.997456        -0.2117890        -0.6749251
##          Application.mode10 Application.mode11 Application.mode12
## Enrolled          -2.552551          -3.678994         -0.2546112
## Graduate          -4.118079          -7.915526         -0.6880046
##          Application.mode13 Application.mode14 Application.mode15
## Enrolled        -0.01151383          0.1311526          0.4648808
## Graduate        -0.35957586         -0.1478074          0.1752612
##          Application.mode16 Application.mode17 Application.mode18
## Enrolled           1.088254           2.830659         -0.9621697
## Graduate           0.356518           1.676483          5.8180068
##          Application.order     Course Daytime.evening.attendance1
## Enrolled       -0.12320166 -0.0532918                  -0.2733244
## Graduate       -0.08112028 -0.1106046                  -0.4494906
##          Previous.qualification Nacionality Mother.s.qualification
## Enrolled            -0.03167466  -0.1933118            -0.02289109
## Graduate             0.00990219  -0.2756723            -0.01187137
##          Father.s.qualification Mother.s.occupation Father.s.occupation
## Enrolled           0.0005176216          0.06674579          0.01688713
## Graduate           0.0060969574          0.09828177         -0.01264109
##          Displaced1 Educational.special.needs1     Debtor1
## Enrolled -0.4491404                 0.08897317 -0.08675309
## Graduate -0.4095232                -0.16606205 -0.76309593
##          Tuition.fees.up.to.date1    Gender1 Scholarship.holder1
## Enrolled                 2.142218 -0.1216737           0.0309521
## Graduate                 3.049752 -0.4395557           0.6845036
##          Age.at.enrollment International1 Curricular.units.1st.sem..credited.
## Enrolled       -0.04507800       1.593485                            -0.1039940
## Graduate       -0.04741784       2.583663                            -0.1834319
##          Curricular.units.1st.sem..enrolled.
## Enrolled                          -0.02595306
## Graduate                          -0.33157835
##          Curricular.units.1st.sem..evaluations.
## Enrolled                             0.01277956
## Graduate                            -0.04727035
##          Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## Enrolled                           0.1174672                       -0.02237252
## Graduate                           0.7895827                       -0.10983475
##          Curricular.units.1st.sem..without.evaluations.
## Enrolled                                      0.1529560
## Graduate                                      0.2660186
##          Curricular.units.2nd.sem..credited.
## Enrolled                          -0.1954114
## Graduate                          -0.1905555
```

```
##          Curricular.units.2nd.sem..enrolled.
## Enrolled                          -0.3611963
## Graduate                          -1.0034565
##          Curricular.units.2nd.sem..evaluations.
## Enrolled                            0.105762254
## Graduate                           -0.002326231
##          Curricular.units.2nd.sem..approved. Curricular.units.2nd.sem..grade.
## Enrolled                           0.3546539                         0.06626025
## Graduate                           1.1577482                         0.17720438
##          Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## Enrolled                                     -0.008969982        -0.1125389
## Graduate                                      0.017624457        -0.1287036
##          Inflation.rate         GDP
## Enrolled    -0.03504659 -0.01829377
## Graduate     0.06516400 -0.06994250
##
## Std. Errors:
##          (Intercept) Marital.status2 Marital.status3 Marital.status4
## Enrolled   0.6762239       0.3138385        1.568217       0.4659508
## Graduate   0.7483582       0.3376813        2.180127       0.5598520
##          Marital.status5 Marital.status6 Application.mode2 Application.mode3
## Enrolled       0.9292028        1.474162        0.02715501          1.215525
## Graduate       0.8314856        1.746867        2.01321185          1.420647
##          Application.mode4 Application.mode5 Application.mode6
## Enrolled         0.6663927       0.003625742          1.103448
## Graduate         0.5207819       0.019308268          1.104628
##          Application.mode7 Application.mode8 Application.mode9
## Enrolled          1.195294         0.1800934         0.4377064
## Graduate          1.221391         0.1881893         0.4474970
##          Application.mode10 Application.mode11 Application.mode12
## Enrolled        0.002240288        0.002470727          0.2722676
## Graduate        0.002881384        0.002858224          0.3100953
##          Application.mode13 Application.mode14 Application.mode15
## Enrolled          0.5074659          0.2858451          0.4517132
## Graduate          0.5700368          0.3328168          0.5015635
##          Application.mode16 Application.mode17 Application.mode18
## Enrolled          0.5716630          0.9971045       0.0001195355
## Graduate          0.6613596          1.1069848       0.0027410785
##          Application.order     Course Daytime.evening.attendance1
## Enrolled        0.05710490 0.01804388                   0.2544531
## Graduate        0.05744225 0.01958483                   0.2814199
##          Previous.qualification Nacionality Mother.s.qualification
## Enrolled             0.02651742   0.1115101            0.009451079
## Graduate             0.02834893   0.1125175            0.010091439
##          Father.s.qualification Mother.s.occupation Father.s.occupation
## Enrolled            0.007181247          0.02564359          0.02039697
## Graduate            0.007785445          0.02787260          0.02257048
##          Displaced1 Educational.special.needs1   Debtor1
## Enrolled  0.1542603                  0.5681664 0.2110660
## Graduate  0.1674303                  0.6314622 0.2618276
##          Tuition.fees.up.to.date1  Gender1 Scholarship.holder1
## Enrolled                0.2495188 0.141238           0.1901920
## Graduate                0.3362177 0.153096           0.1839805
##          Age.at.enrollment International1 Curricular.units.1st.sem..credited.
```

```
## Enrolled        0.01642814       1.329514                              0.1216621
## Graduate        0.01739912       1.480545                              0.1279007
##          Curricular.units.1st.sem..enrolled.
## Enrolled                           0.1380328
## Graduate                           0.1646898
##          Curricular.units.1st.sem..evaluations.
## Enrolled                             0.03011193
## Graduate                             0.03910303
##          Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## Enrolled                           0.06784628                        0.02858896
## Graduate                           0.08867583                        0.05286981
##          Curricular.units.1st.sem..without.evaluations.
## Enrolled                                      0.1272293
## Graduate                                      0.1735258
##          Curricular.units.2nd.sem..credited.
## Enrolled                           0.1286822
## Graduate                           0.1290968
##          Curricular.units.2nd.sem..enrolled.
## Enrolled                           0.1347907
## Graduate                           0.1618619
##          Curricular.units.2nd.sem..evaluations.
## Enrolled                             0.02837932
## Graduate                             0.03699517
##          Curricular.units.2nd.sem..approved. Curricular.units.2nd.sem..grade.
## Enrolled                           0.06109639                        0.02623831
## Graduate                           0.08258083                        0.05395088
##          Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## Enrolled                                      0.1001623        0.02882107
## Graduate                                      0.1409894        0.03140293
##          Inflation.rate        GDP
## Enrolled     0.04871151 0.03403048
## Graduate     0.05264825 0.03686740
##
## Residual Deviance: 3367.357
## AIC: 3587.357
```

Our multinomial logistic regression model was fitted to predict the academic success categories using all predictors. The model converged after 100 iterations, with a final residual deviance of 3367.357 and an AIC of 3587.357, indicating a good fit.

Model Evaluation

Now, we predict the outcomes for the test set, construct a confusion matrix, and calculate accuracy.

```r
# Evaluates model performance

# Predict outcomes for the test set
predictions <- predict(model, newdata = test_data, type = "class")
# Construct confusion matrix
conf_matrix <- confusionMatrix(table(predictions, test_data$Target))
accuracy <- conf_matrix$overall['Accuracy']

# Prints confusion matrix and accuracy
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##
## predictions Dropout Enrolled Graduate
##    Dropout      339       55       27
##    Enrolled      39       79       27
##    Graduate      48      104      608
##
## Overall Statistics
##
##                Accuracy : 0.7738
##                  95% CI : (0.7503, 0.796)
##     No Information Rate : 0.4992
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.618
##
##  Mcnemar's Test P-Value : 1.2e-11
##
## Statistics by Class:
##
##                      Class: Dropout Class: Enrolled Class: Graduate
## Sensitivity                  0.7958         0.33193          0.9184
## Specificity                  0.9089         0.93934          0.7711
## Pos Pred Value               0.8052         0.54483          0.8000
## Neg Pred Value               0.9039         0.86537          0.9046
## Prevalence                   0.3213         0.17949          0.4992
## Detection Rate               0.2557         0.05958          0.4585
## Detection Prevalence         0.3175         0.10935          0.5732
## Balanced Accuracy            0.8523         0.63564          0.8448
```

```
print(paste("Accuracy:", round(accuracy, 4)))
```

```
## [1] "Accuracy: 0.7738"
```

Looking at our confusion matrix, we see our overall accuracy of the model is 77.38%, with a 95% CI of (0.7503, 0.796).

For further assessment of the model, we will focus on the sensitivity and the positive prediction value of the model for each outcome variable. Sensitivity (or recall) is the true positive rate. It gives the rate at which the model was able to correctly predict a certain class out of all the true instances of that class. The positive prediction value (or precision) gives the rate at which a model correctly predicts a certain class out of all the times it predicted that class (whether a true positive or a false positive).

Some takeaways regarding our specific outcomes include:

- Dropout: Moderate recall (79.58%) and precision (80.52%), indicating decent predictive power for dropouts.

- Enrolled: Poor recall (33.19%) and low precision (54.48%), suggesting it is harder to predict enrollment accurately.

- Graduate: High recall (91.84%) and moderate precision (80.00%), indicating strong predictive power for graduates.

Interpretation of Results

Finally, we discuss the model's coefficients and interpret the impact of each predictor on the likelihood of students falling into each outcome category.

```r
# Coefficients of the model
coefficients <- summary(model)$coefficients
print(coefficients)
```

```
##          (Intercept) Marital.status2 Marital.status3 Marital.status4
## Enrolled   0.6117102      0.14123780        2.061327       0.7126513
## Graduate   0.6282169      0.06669225        1.385022       0.7445702
##          Marital.status5 Marital.status6 Application.mode2 Application.mode3
## Enrolled      -0.3882554    -0.005482948        -2.9159462          2.318624
## Graduate      -0.3878728    -2.026409248        -0.7848863          1.350059
##          Application.mode4 Application.mode5 Application.mode6
## Enrolled        -1.5264665         -3.010398          1.845225
## Graduate        -0.3880277          3.480042          2.903569
##          Application.mode7 Application.mode8 Application.mode9
## Enrolled          1.876892        -0.1291556        -0.4408973
## Graduate          1.997456        -0.2117890        -0.6749251
##          Application.mode10 Application.mode11 Application.mode12
## Enrolled          -2.552551          -3.678994         -0.2546112
## Graduate          -4.118079          -7.915526         -0.6880046
##          Application.mode13 Application.mode14 Application.mode15
## Enrolled       -0.01151383          0.1311526          0.4648808
## Graduate       -0.35957586         -0.1478074          0.1752612
##          Application.mode16 Application.mode17 Application.mode18
## Enrolled           1.088254           2.830659         -0.9621697
## Graduate           0.356518           1.676483          5.8180068
##          Application.order     Course Daytime.evening.attendance1
## Enrolled       -0.12320166 -0.0532918                  -0.2733244
## Graduate       -0.08112028 -0.1106046                  -0.4494906
##          Previous.qualification Nacionality Mother.s.qualification
## Enrolled            -0.03167466  -0.1933118            -0.02289109
## Graduate             0.00990219  -0.2756723            -0.01187137
##          Father.s.qualification Mother.s.occupation Father.s.occupation
## Enrolled           0.0005176216          0.06674579          0.01688713
## Graduate           0.0060969574          0.09828177         -0.01264109
##          Displaced1 Educational.special.needs1     Debtor1
## Enrolled -0.4491404                0.08897317 -0.08675309
## Graduate -0.4095232               -0.16606205 -0.76309593
##          Tuition.fees.up.to.date1    Gender1 Scholarship.holder1
## Enrolled                 2.142218 -0.1216737           0.0309521
## Graduate                 3.049752 -0.4395557           0.6845036
##          Age.at.enrollment International1 Curricular.units.1st.sem..credited.
## Enrolled       -0.04507800       1.593485                           -0.1039940
## Graduate       -0.04741784       2.583663                           -0.1834319
##          Curricular.units.1st.sem..enrolled.
## Enrolled                         -0.02595306
## Graduate                         -0.33157835
##          Curricular.units.1st.sem..evaluations.
## Enrolled                             0.01277956
## Graduate                            -0.04727035
##          Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## Enrolled                           0.1174672                       -0.02237252
## Graduate                           0.7895827                       -0.10983475
##          Curricular.units.1st.sem..without.evaluations.
## Enrolled                                      0.1529560
```

```
## Graduate                                      0.2660186
##          Curricular.units.2nd.sem..credited.
## Enrolled                          -0.1954114
## Graduate                          -0.1905555
##          Curricular.units.2nd.sem..enrolled.
## Enrolled                          -0.3611963
## Graduate                          -1.0034565
##          Curricular.units.2nd.sem..evaluations.
## Enrolled                            0.105762254
## Graduate                           -0.002326231
##          Curricular.units.2nd.sem..approved. Curricular.units.2nd.sem..grade.
## Enrolled                           0.3546539                        0.06626025
## Graduate                           1.1577482                        0.17720438
##          Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## Enrolled                                     -0.008969982        -0.1125389
## Graduate                                      0.017624457        -0.1287036
##          Inflation.rate         GDP
## Enrolled    -0.03504659 -0.01829377
## Graduate     0.06516400 -0.06994250
```

We can also look at these to interpret them as odds ratios:

```
# Exponentiate coefficients to interpret as odds ratios
exp_coefficients <- exp(coefficients)
print(exp_coefficients)
```

```
##          (Intercept) Marital.status2 Marital.status3 Marital.status4
## Enrolled    1.843582        1.151698        7.856386        2.039391
## Graduate    1.874266        1.068966        3.994912        2.105536
##          Marital.status5 Marital.status6 Application.mode2 Application.mode3
## Enrolled       0.6782391       0.9945321        0.05415277         10.161678
## Graduate       0.6784986       0.1318080        0.45617156          3.857653
##          Application.mode4 Application.mode5 Application.mode6
## Enrolled         0.2173021        0.04927207          6.329522
## Graduate         0.6783936       32.46110060         18.239122
##          Application.mode7 Application.mode8 Application.mode9
## Enrolled          6.533171         0.8788372         0.6434588
## Graduate          7.370279         0.8091354         0.5091945
##          Application.mode10 Application.mode11 Application.mode12
## Enrolled         0.07788272       0.0252483503          0.7752179
## Graduate         0.01627575       0.0003650317          0.5025779
##          Application.mode13 Application.mode14 Application.mode15
## Enrolled          0.9885522          1.1401418           1.591824
## Graduate          0.6979723          0.8625972           1.191557
##          Application.mode16 Application.mode17 Application.mode18
## Enrolled           2.969086          16.956627           0.382063
## Graduate           1.428347           5.346717         336.301064
##          Application.order    Course Daytime.evening.attendance1
## Enrolled         0.8840854 0.9481033                   0.7608459
## Graduate         0.9220828 0.8952927                   0.6379530
##          Previous.qualification Nacionality Mother.s.qualification
## Enrolled              0.9688217   0.8242249              0.9773689
## Graduate              1.0099514   0.7590617              0.9881988
##          Father.s.qualification Mother.s.occupation Father.s.occupation
## Enrolled               1.000518            1.069024           1.0170305
```

```
## Graduate               1.006116            1.103274           0.9874385
##          Displaced1 Educational.special.needs1   Debtor1
## Enrolled  0.6381765                  1.0930513 0.9169035
## Graduate  0.6639667                  0.8469937 0.4662208
##          Tuition.fees.up.to.date1   Gender1 Scholarship.holder1
## Enrolled                 8.518311 0.8854372            1.031436
## Graduate                21.110107 0.6443226            1.982787
##          Age.at.enrollment International1 Curricular.units.1st.sem..credited.
## Enrolled         0.9559229       4.920866                              0.9012307
## Graduate         0.9536888      13.245570                              0.8324086
##          Curricular.units.1st.sem..enrolled.
## Enrolled                           0.9743808
## Graduate                           0.7177899
##          Curricular.units.1st.sem..evaluations.
## Enrolled                             1.0128616
## Graduate                             0.9538295
##          Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## Enrolled                           1.124645                          0.9778759
## Graduate                           2.202477                          0.8959822
##          Curricular.units.1st.sem..without.evaluations.
## Enrolled                                       1.165274
## Graduate                                       1.304759
##          Curricular.units.2nd.sem..credited.
## Enrolled                           0.8224962
## Graduate                           0.8264999
##          Curricular.units.2nd.sem..enrolled.
## Enrolled                           0.6968422
## Graduate                           0.3666100
##          Curricular.units.2nd.sem..evaluations.
## Enrolled                             1.1115576
## Graduate                             0.9976765
##          Curricular.units.2nd.sem..approved. Curricular.units.2nd.sem..grade.
## Enrolled                           1.425687                           1.068505
## Graduate                           3.182758                           1.193875
##          Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## Enrolled                                       0.9910701         0.8935626
## Graduate                                       1.0177807         0.8792346
##          Inflation.rate       GDP
## Enrolled      0.9655604 0.9818725
## Graduate      1.0673341 0.9324474
```

Our coefficients provide insight into the relationship between predictors and the likelihood of being in each category (Enrolled, Graduate) compared to the reference category (Dropout). Here, we see that:

- **Marital statuses** show significant impacts on enrollment and graduation. Our model shows that being married (Marital Status 3) significantly increases the odds of being enrolled (7.86) or graduating (3.99) compared to dropping out.

- Different **application modes** significantly affect the likelihood of dropping out, enrolling, or graduating. Being an applicant that is over 23. years old (Application Mode 12) reduces the odds of being enrolled (0.78) and graduating (0.50), compared to dropping out.

- **Age** has a slight negative impact on enrollment and graduation compared to dropping out. Our model shows that a higher age at enrollment slightly decreases the odds of both being enrolled (0.96) and graduating (0.95), compared to dropping out.

- Our model suggests that students who have **up-to-date tuition fees** are more likely to graduate. Specifically, our model shows that being up-to-date with tuition fees significantly increases the odds of both being enrolled (8.52) and graduating (21.11), compared to dropping out.

- **Scholarship holders** increase their odds of graduating (1.98), as well as being enrolled (1.03) when compared to dropping out, highlighting the positive impact of financial support.

- **Gender** has a slight negative impact in our model. Males (coded as 1) have lower odds of enrolling (0.89) or graduating (0.64) compared to females (coded as 0).

Conclusion

Our model's accuracy suggests it is reliable, especially for predicting dropouts and graduates, though less effective for currently enrolled students.

Marital status, application mode, gender, age at enrollment, tuition fee status, and scholarship status are significant predictors of academic outcomes. Certain factors like being up-to-date with tuition fees have a strong positive impact on both enrollment and graduation probabilities, while others like age have negative impacts on academic success.

Some takeaways from this analysis is that targeted interventions could focus on financial support, especially in ensuring tuition fees are up-to-date to improve students' academic success rates. Additional support may be needed for students who attend evening classes or have higher ages at enrollment to reduce dropout rates. Overall, the multinomial logistic regression model provides valuable insights into the factors influencing academic success and can guide policies and interventions to improve student outcomes.

In this project, we demonstrated the use of multinomial logistic regression using a real-world dataset. We covered the steps of data preparation, exploratory data analysis, model fitting, and evaluation, as well as interpreting the results. By following these steps, you should now have a solid understanding of how to apply MLR and interpret its outputs in the context of predicting student success.

Reading List

Learning Resources:

- Introduction to MLR

- Advanced Regression Methods, Ch. 11: MLR

- Geek for Geeks: MLR Overview

- MLR Theory/Code From Scratch

- Wikipedia: MLR

R Implementation:

- UCLA: MLR Analysis w/ R

- Penguins: MLR w/ R Walk Through

- Kaggle MLR Walk-through Notebook

Python Implementation:

- MLR w/ Python

- MLR w/ Sklearn Walk-through