GradX – Graduation Success Predictor

Prathiksha Rumale Vishwanath

Introduction: Understanding Student Dropout Risk

In this analysis, we aim to predict whether a student will **graduate or drop out** based on various academic, financial, and demographic factors. The dataset contains information about student enrollment, previous qualifications, financial aid status, academic performance, and socioeconomic indicators. By applying **logistic regression**, we aim to identify key predictors of student success and help educational institutions implement data-driven retention strategies.

Objectives: - Identify the most significant factors affecting student dropout rates. - Train and evaluate a predictive model using real student data. - Provide actionable insights to improve student retention.

Load the Data

```
# Load student enrollment data set (CSV file)
data <- read_csv("data.csv", col_names = TRUE, col_types = cols())
head(data)
```

```
## # A tibble: 6 x 1
##   Marital status;Application mode;Application order;Course;"Daytime/evening at~1
##   <chr>
## 1 1;17;5;171;1;1;122.0;1;19;12;5;9;127.3;1;0;0;1;1;0;20;0;0;0;0;0;0.0;0;0;0;0;0~
## 2 1;15;1;9254;1;1;160.0;1;1;3;3;3;142.5;1;0;0;0;1;0;19;0;0;6;6;6;14.0;0;0;6;6;6~
## 3 1;1;5;9070;1;1;122.0;1;37;37;9;9;124.8;1;0;0;0;1;0;19;0;0;6;0;0;0.0;0;0;6;0;0~
## 4 1;17;2;9773;1;1;122.0;1;38;37;5;3;119.6;1;0;0;1;0;0;20;0;0;6;8;6;13.428571428~
## 5 2;39;1;8014;0;1;100.0;1;37;38;9;9;141.5;0;0;0;1;0;0;45;0;0;6;9;5;12.333333333~
## 6 2;39;1;9991;0;19;133.1;1;37;37;9;7;114.8;0;0;1;1;1;0;50;0;0;5;10;5;11.8571428~
## # i abbreviated name:
## #   1: `Marital status;Application mode;Application order;Course;"Daytime/evening attendance\t";Previ
```

Data Cleaning

```
# Check for missing values and display only columns with missing data
missing_values <- colSums(is.na(data))
missing_values[missing_values > 0]
```

```
## named numeric(0)
```

```
# Reload data set with correct delimiter once observed and analyzing presence or missing values
data <- read.csv("data.csv", sep = ";", header = TRUE)
head(data)
```

```
##   Marital.status Application.mode Application.order Course
## 1              1               17                 5    171
## 2              1               15                 1   9254
## 3              1                1                 5   9070
## 4              1               17                 2   9773
## 5              2               39                 1   8014
## 6              2               39                 1   9991
##   Daytime.evening.attendance. Previous.qualification
## 1                           1                      1
## 2                           1                      1
## 3                           1                      1
## 4                           1                      1
## 5                           0                      1
## 6                           0                     19
##   Previous.qualification..grade. Nacionality Mother.s.qualification
## 1                          122.0           1                     19
## 2                          160.0           1                      1
## 3                          122.0           1                     37
## 4                          122.0           1                     38
## 5                          100.0           1                     37
## 6                          133.1           1                     37
##   Father.s.qualification Mother.s.occupation Father.s.occupation
## 1                     12                   5                   9
## 2                      3                   3                   3
## 3                     37                   9                   9
## 4                     37                   5                   3
## 5                     38                   9                   9
## 6                     37                   9                   7
##   Admission.grade Displaced Educational.special.needs Debtor
## 1           127.3         1                        0      0
## 2           142.5         1                        0      0
## 3           124.8         1                        0      0
## 4           119.6         1                        0      0
## 5           141.5         0                        0      0
## 6           114.8         0                        0      1
##   Tuition.fees.up.to.date Gender Scholarship.holder Age.at.enrollment
## 1                       1      1                  0                20
## 2                       0      1                  0                19
## 3                       0      1                  0                19
## 4                       1      0                  0                20
## 5                       1      0                  0                45
## 6                       1      1                  0                50
##   International Curricular.units.1st.sem..credited.
## 1             0                                   0
## 2             0                                   0
## 3             0                                   0
## 4             0                                   0
## 5             0                                   0
## 6             0                                   0
##   Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.
## 1                                   0                                      0
## 2                                   6                                      6
## 3                                   6                                      0
## 4                                   6                                      8
```

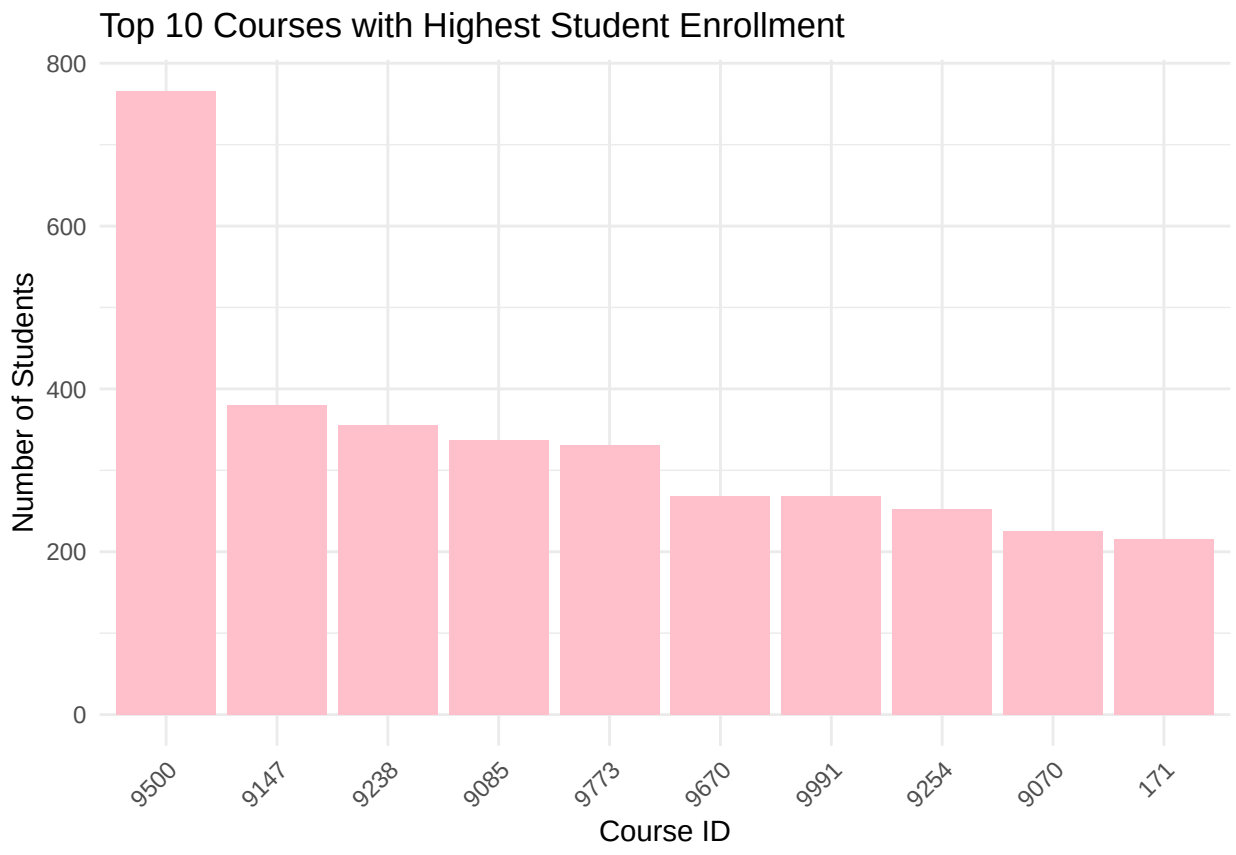
```
## 5                                    6                                     9
## 6                                    5                                    10
##   Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.
## 1                                    0                           0.00000
## 2                                    6                          14.00000
## 3                                    0                           0.00000
## 4                                    6                          13.42857
## 5                                    5                          12.33333
## 6                                    5                          11.85714
##   Curricular.units.1st.sem..without.evaluations.
## 1                                              0
## 2                                              0
## 3                                              0
## 4                                              0
## 5                                              0
## 6                                              0
##   Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.
## 1                                    0                                  0
## 2                                    0                                  6
## 3                                    0                                  6
## 4                                    0                                  6
## 5                                    0                                  6
## 6                                    0                                  5
##   Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.
## 1                                     0                                   0
## 2                                     6                                   6
## 3                                     0                                   0
## 4                                    10                                   5
## 5                                     6                                   6
## 6                                    17                                   5
##   Curricular.units.2nd.sem..grade.
## 1                          0.00000
## 2                         13.66667
## 3                          0.00000
## 4                         12.40000
## 5                         13.00000
## 6                         11.50000
##   Curricular.units.2nd.sem..without.evaluations. Unemployment.rate
## 1                                              0              10.8
## 2                                              0              13.9
## 3                                              0              10.8
## 4                                              0               9.4
## 5                                              0              13.9
## 6                                              5              16.2
##   Inflation.rate   GDP   Target
## 1            1.4  1.74  Dropout
## 2           -0.3  0.79 Graduate
## 3            1.4  1.74  Dropout
## 4           -0.8 -3.12 Graduate
## 5           -0.3  0.79 Graduate
## 6            0.3 -0.92 Graduate
```

Student Enrollment by Course

```
# Count number of students per course
course_enrollment <- data %>%
  group_by(Course) %>%
  summarise(Number_of_Students = n()) %>%
  arrange(desc(Number_of_Students))

# Display summary of top 10 enrolled courses
top_courses <- head(course_enrollment, 10)
print(top_courses)
```

```
## # A tibble: 10 x 2
##    Course Number_of_Students
##     <int>              <int>
##  1   9500                766
##  2   9147                380
##  3   9238                355
##  4   9085                337
##  5   9773                331
##  6   9670                268
##  7   9991                268
##  8   9254                252
##  9   9070                226
## 10    171                215
```

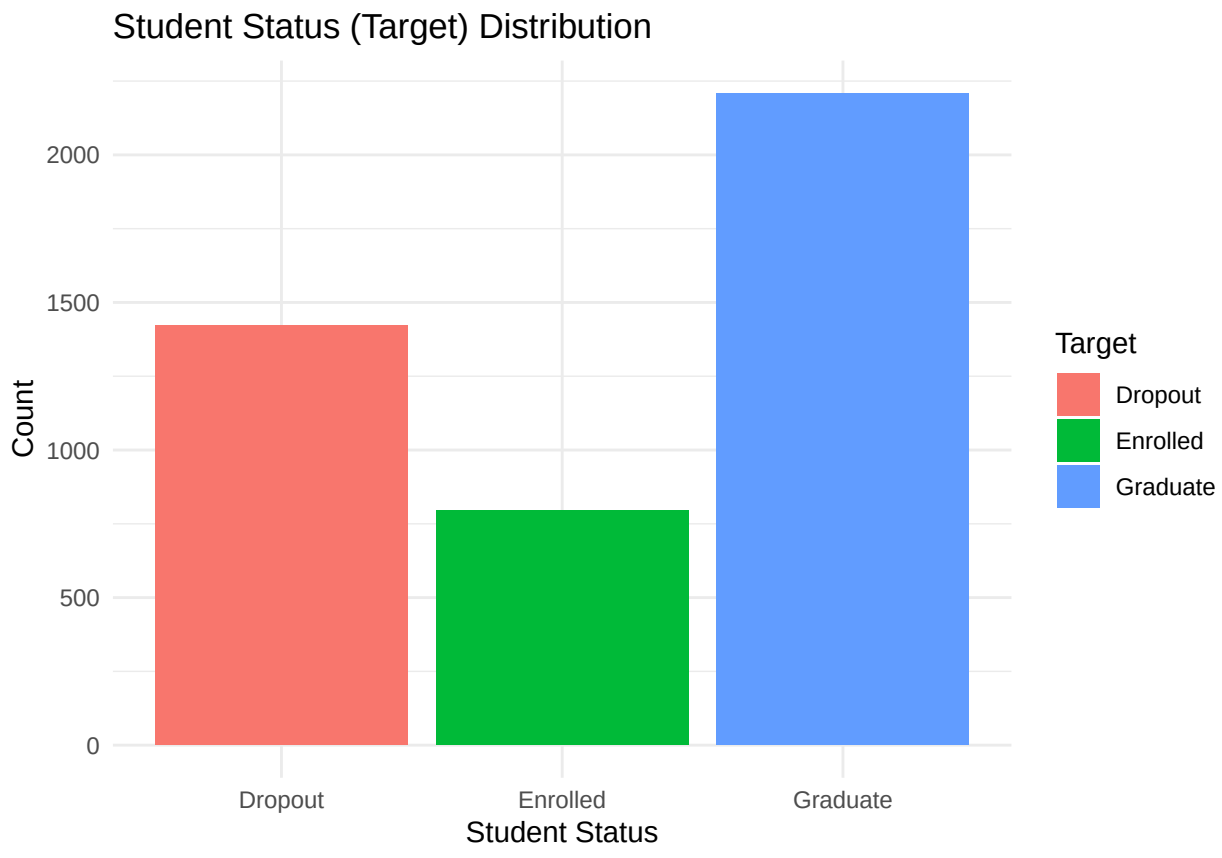


Dropout vs Graduation Rate

```
# Count students based on their academic status
dropout_distribution <- data %>%
  group_by(Target) %>%
  summarise(Count = n())

# Display summary
dropout_distribution
```

```
## # A tibble: 3 x 2
##   Target   Count
##   <chr>    <int>
## 1 Dropout   1421
## 2 Enrolled   794
## 3 Graduate  2209
```



Statistical Analysis: Predicting Dropout

```
#Analyzing all columns for feature engineering
colnames(data)
```

```
##  [1] "Marital.status"
##  [2] "Application.mode"
##  [3] "Application.order"
##  [4] "Course"
##  [5] "Daytime.evening.attendance."
##  [6] "Previous.qualification"
##  [7] "Previous.qualification..grade."
##  [8] "Nacionality"
##  [9] "Mother.s.qualification"
## [10] "Father.s.qualification"
## [11] "Mother.s.occupation"
## [12] "Father.s.occupation"
## [13] "Admission.grade"
## [14] "Displaced"
## [15] "Educational.special.needs"
## [16] "Debtor"
## [17] "Tuition.fees.up.to.date"
## [18] "Gender"
## [19] "Scholarship.holder"
## [20] "Age.at.enrollment"
## [21] "International"
## [22] "Curricular.units.1st.sem..credited."
## [23] "Curricular.units.1st.sem..enrolled."
## [24] "Curricular.units.1st.sem..evaluations."
## [25] "Curricular.units.1st.sem..approved."
## [26] "Curricular.units.1st.sem..grade."
## [27] "Curricular.units.1st.sem..without.evaluations."
## [28] "Curricular.units.2nd.sem..credited."
## [29] "Curricular.units.2nd.sem..enrolled."
## [30] "Curricular.units.2nd.sem..evaluations."
## [31] "Curricular.units.2nd.sem..approved."
## [32] "Curricular.units.2nd.sem..grade."
## [33] "Curricular.units.2nd.sem..without.evaluations."
## [34] "Unemployment.rate"
## [35] "Inflation.rate"
## [36] "GDP"
## [37] "Target"
```

Train-Test Split and Model Evaluation

```r
# Split data into 80% training and 20% testing set
set.seed(42)
trainIndex <- createDataPartition(data$Target, p = 0.8, list = FALSE)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]

# Ensure Target is a factor
trainData$Target <- factor(trainData$Target, levels = c("Dropout", "Graduate"))
testData$Target <- factor(testData$Target, levels = c("Dropout", "Graduate"))

# Train logistic regression model
model <- glm(Target ~ Age.at.enrollment + Scholarship.holder + Gender +
```

```
            Curricular.units.1st.sem..grade. + Curricular.units.2nd.sem..grade.,
            data = trainData, family = binomial)
```

##Summary

```
# Display model summary
summary(model)
```

```
##
## Call:
## glm(formula = Target ~ Age.at.enrollment + Scholarship.holder +
##     Gender + Curricular.units.1st.sem..grade. + Curricular.units.2nd.sem..grade.,
##     family = binomial, data = trainData)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.423370   0.228390  -6.232 4.60e-10 ***
## Age.at.enrollment              -0.049367   0.006315  -7.817 5.39e-15 ***
## Scholarship.holder              1.494828   0.143196  10.439  < 2e-16 ***
## Gender                         -0.569178   0.105685  -5.386 7.22e-08 ***
## Curricular.units.1st.sem..grade.  0.008027   0.022002   0.365    0.715
## Curricular.units.2nd.sem..grade.  0.274134   0.020976  13.069  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3889.0  on 2904  degrees of freedom
## Residual deviance: 2458.9  on 2899  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 2470.9
##
## Number of Fisher Scoring iterations: 5
```

Variance Inflation Factor (VIF) Analysis

```
# Check multicollinearity among predictor variables
vif(model)
```

```
##                Age.at.enrollment               Scholarship.holder
##                         1.032547                         1.028349
##                           Gender Curricular.units.1st.sem..grade.
##                         1.026532                         2.724981
## Curricular.units.2nd.sem..grade.
##                         2.739887
```

##Confusion Matrix

```r
# Make predictions on test data
predictions <- predict(model, newdata = testData, type = "response")

# Convert predicted probabilities into class labels
predicted_classes <- ifelse(predictions >= 0.5, "Graduate", "Dropout")

# Evaluate model performance using confusion matrix
confusionMatrix(factor(predicted_classes), testData$Target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Dropout Graduate
##   Dropout      179       32
##   Graduate     105      409
##
##                Accuracy : 0.811
##                  95% CI : (0.7806, 0.8389)
##     No Information Rate : 0.6083
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5845
##
##  Mcnemar's Test P-Value : 7.681e-10
##
##             Sensitivity : 0.6303
##             Specificity : 0.9274
##          Pos Pred Value : 0.8483
##          Neg Pred Value : 0.7957
##              Prevalence : 0.3917
##          Detection Rate : 0.2469
##    Detection Prevalence : 0.2910
##       Balanced Accuracy : 0.7789
##
##        'Positive' Class : Dropout
##
```

New Student Dropout or Graduation Prediction

```r
# Define new student data for prediction
new_data <- data.frame(
  Age.at.enrollment = c(21, 25, 30),
  Scholarship.holder = c(1, 0, 1),  # 1 = Has Scholarship, 0 = No Scholarship
  Gender = c(1, 0, 1),  # 1 = Male, 0 = Female
  Curricular.units.1st.sem..grade. = c(14, 10, 12),
  Curricular.units.2nd.sem..grade. = c(15, 9, 11)
)

# Predict dropout probability
predictions <- predict(model, newdata = new_data, type = "response")
```

```r
# Convert probabilities into class labels
predicted_classes <- ifelse(predictions >= 0.75, "Graduate", "Dropout")

# Display results
prediction_results <- data.frame(new_data, Predicted_Status = predicted_classes, Probability = predicti
print(prediction_results)
```

```
##   Age.at.enrollment Scholarship.holder Gender Curricular.units.1st.sem..grade.
## 1                21                  1      1                                14
## 2                25                  0      0                                10
## 3                30                  1      1                                12
##   Curricular.units.2nd.sem..grade. Predicted_Status Probability
## 1                               15         Graduate   0.9364314
## 2                                9          Dropout   0.4725101
## 3                               11         Graduate   0.7564042
```

Conclusion and Actionable Insights

This analysis provides insights into **student enrollment trends and dropout rates**, helping institutions understand key factors influencing student retention.

Key Takeaways:

- **VIF Analysis:** No severe multicollinearity issues detected, ensuring predictor stability.
- **Model Accuracy:** ~80.8% accuracy with a sensitivity of **62.3%** and specificity of **92.7%**.
- **Key Predictors:** Scholarship holders have significantly higher graduation rates, while older students are more likely to drop out.
- **Predictions:** The model successfully classifies new students based on their academic and demographic factors.

Actionable Insights:

1. **Scholarship Programs Significantly Reduce Dropout Rates**
   - Students who receive scholarships have a **much higher likelihood of graduating** than those without financial aid.
   - **Action:** Educational institutions should **increase scholarship availability** or introduce **financial assistance programs** to support at-risk students.
2. **Early Academic Performance is a Key Indicator of Retention**
   - **First-semester grades** strongly correlate with dropout likelihood.
   - **Action:** Universities should implement **early intervention strategies** such as **mentoring, tutoring, and academic support programs** for students struggling in their **first semester**.