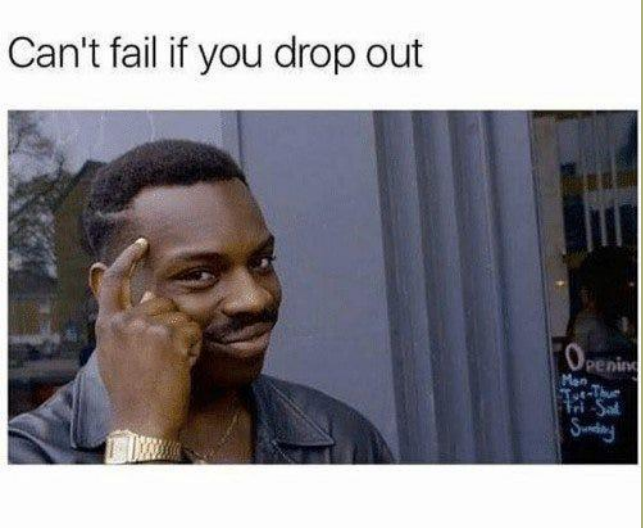
Predicting Student Dropout Risk in Early University Stages

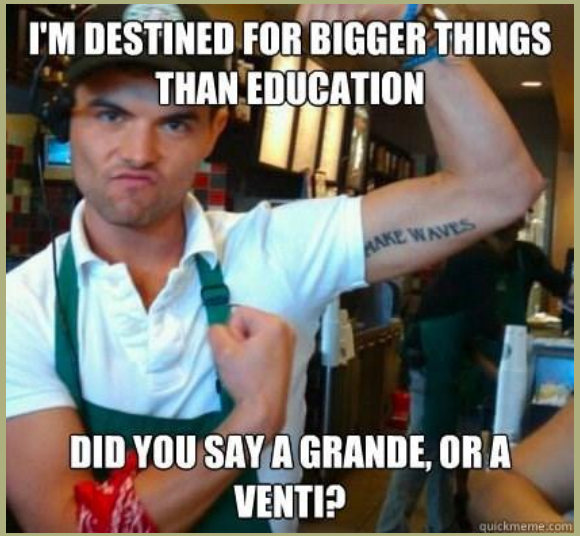
Impact on economy and society.
Why this even matter?

Asset Mukayev
3036384608

But before I will start...

Have you ever thought about dropping out?





Did you know?

$\frac{1}{3}$ Of students who enroll in higher education do not complete their degree programs

01

The importance

Statistics

Lose-Lose-Lose situation:

- students face lower earnings and FOMO
- universities face revenue loss and reputation damage
- economy faces increased unemployment rate

The National Dropout Prevention Center (NDPC) in the US claims:

Economy

Each year's class of dropouts will cost the country over $200 billion during their lifetimes in lost earnings and unrealized tax revenue (Catterall, 1985).

Existing Data

kaggle

A 2021 dataset from Portugal University containing records for 4,400 students with 35 features.

Existing Data Analysis works in GitHub

Research Papers

Existing Research Papers with Statistics and Application

Loaded Dataset

```
data <- read.csv("C:\\Users\\Asset\\Downloads\\dataset.csv")

head(data)
names(data)

library(tidyverse)
library(dslabs)
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
library(caret)
library(HistData)
library(xgboost)
library(Matrix)
library(pROC)
library(recipes)
library(randomForest)
library(smotefamily)
library(gridExtra)
library(ggcorrplot)
```

```
 [1] "Marital.status"                                "Application.mode"
 [3] "Application.order"                             "Course"
 [5] "Daytime.evening.attendance"                    "Previous.qualification"
 [7] "Nacionality"                                   "Mother.s.qualification"
 [9] "Father.s.qualification"                        "Mother.s.occupation"
[11] "Father.s.occupation"                           "Displaced"
[13] "Educational.special.needs"                     "Debtor"
[15] "Tuition.fees.up.to.date"                       "Gender"
[17] "Scholarship.holder"                            "Age.at.enrollment"
[19] "International"
"Curricular.units.1st.sem..credited."
[21] "Curricular.units.1st.sem..enrolled."
"Curricular.units.1st.sem..evaluations."
[23] "Curricular.units.1st.sem..approved."
"Curricular.units.1st.sem..grade."
[25] "Curricular.units.1st.sem..without.evaluations."
"Curricular.units.2nd.sem..credited."
[27] "Curricular.units.2nd.sem..enrolled."
"Curricular.units.2nd.sem..evaluations."
[29] "Curricular.units.2nd.sem..approved."
"Curricular.units.2nd.sem..grade."
[31] "Curricular.units.2nd.sem..without.evaluations." "Unemployment.rate"
[33] "Inflation.rate"                                "GDP"
[35] "Target"
```

02

Key Questions

1. Which factors affect students’ drop out?
2. Is there any impact on economy?
3. Which model is the best?

Data Preprocessing

Was lucky enough to get dataset without NAs

Changing data type from string to integer:

```
unique(data$Target)
```

```
[1] "Dropout"  "Graduate" "Enrolled"
```

```
data_updated <- data |>
  mutate(Target = trimws(Target)) |>
  mutate(Target = case_when(
    Target == "Dropout"  ~ 0,
    Target == "Enrolled" ~ 1,
    Target == "Graduate" ~ 2
  ))

unique(data_updated$Target)
```

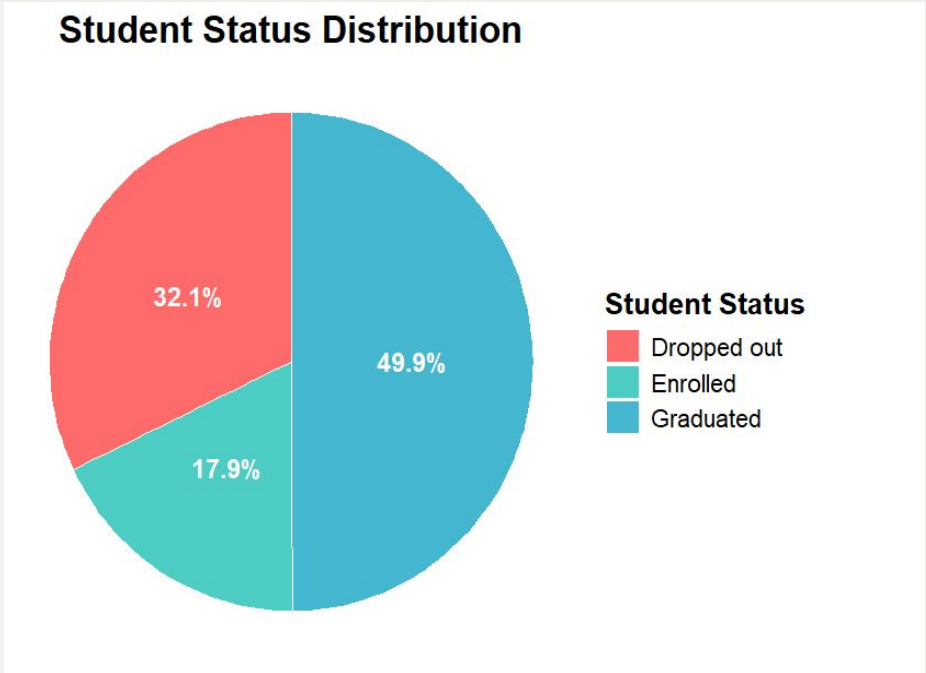
```
[1] 0 2 1
```

Statistics

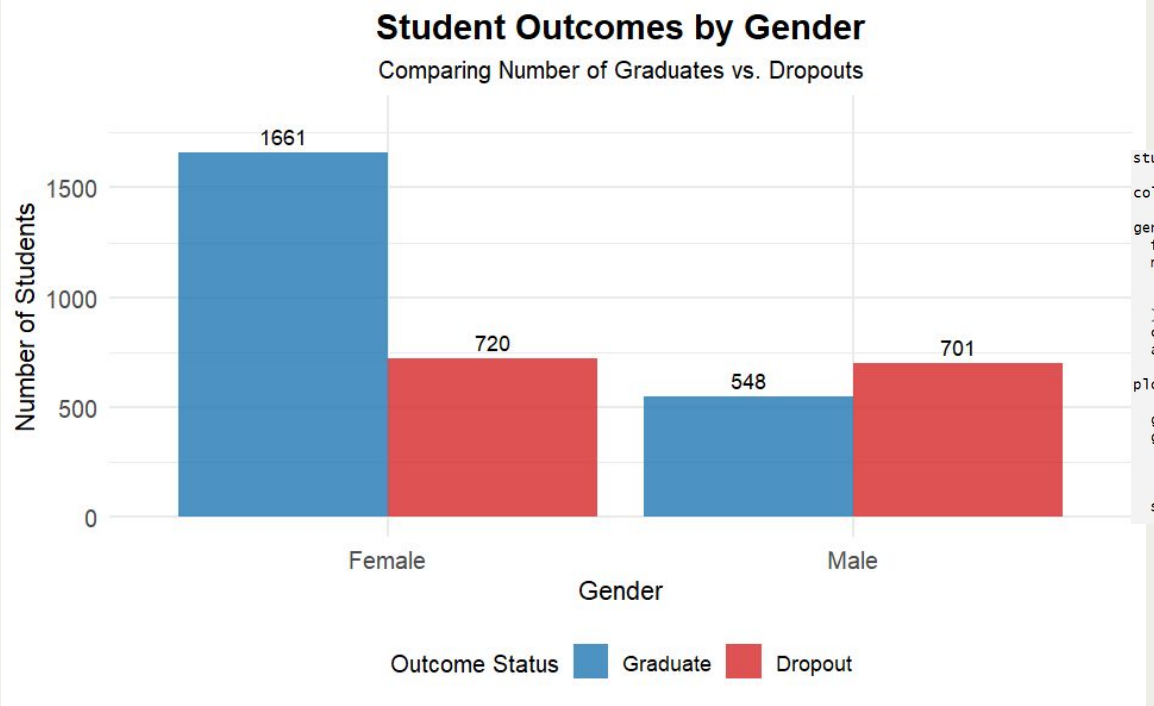
```r
data_distribution <- data |>
  mutate(Target = factor(Target,
                         levels = c("Dropout", "Enrolled", "Graduate"),
                         labels = c("Dropped out", "Enrolled", "Graduated")))

status_counts <- data_distribution |>
  count(Target) |>
  mutate(percentage = n/sum(n)*100,
         label = paste0(round(percentage, 1), "%"))

ggplot(status_counts, aes(x = "", y = percentage, fill = Target)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  geom_text(aes(label = label),
            position = position_stack(vjust = 0.5),
            size = 4.5, color = "white", fontface = "bold") +
  scale_fill_manual(values = c("#FF6B6B", "#4ECDC4", "#45B7D1")) +
  labs(title = "Student Status Distribution",
       fill = "Student Status") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
        legend.position = "right",
        legend.text = element_text(size = 12),
        legend.title = element_text(size = 14, face = "bold"))
```



Statistics




```r
student_data_raw <- data

colnames(student_data_raw) <- make.names(colnames(student_data_raw), unique = TRUE)

gender_outcome_data <- student_data_raw %>%
  filter(Target %in% c("Dropout", "Graduate")) %>%
  mutate(
    Gender_Label = factor(ifelse(Gender == 1, "Male", "Female")),
    Outcome = factor(Target, levels = c("Graduate", "Dropout"))
  ) %>%
  count(Gender_Label, Outcome, name = "Count") %>%
  arrange(Gender_Label, Outcome)

plot_gender_outcome <- ggplot(gender_outcome_data,
                              aes(x = Gender_Label, y = Count, fill = Outcome)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9), alpha = 0.8) +
  geom_text(aes(label = Count),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5) +
  scale_fill_manual(values = c("Graduate" = "#1f77b4", "Dropout" = "#d62728")) +
```

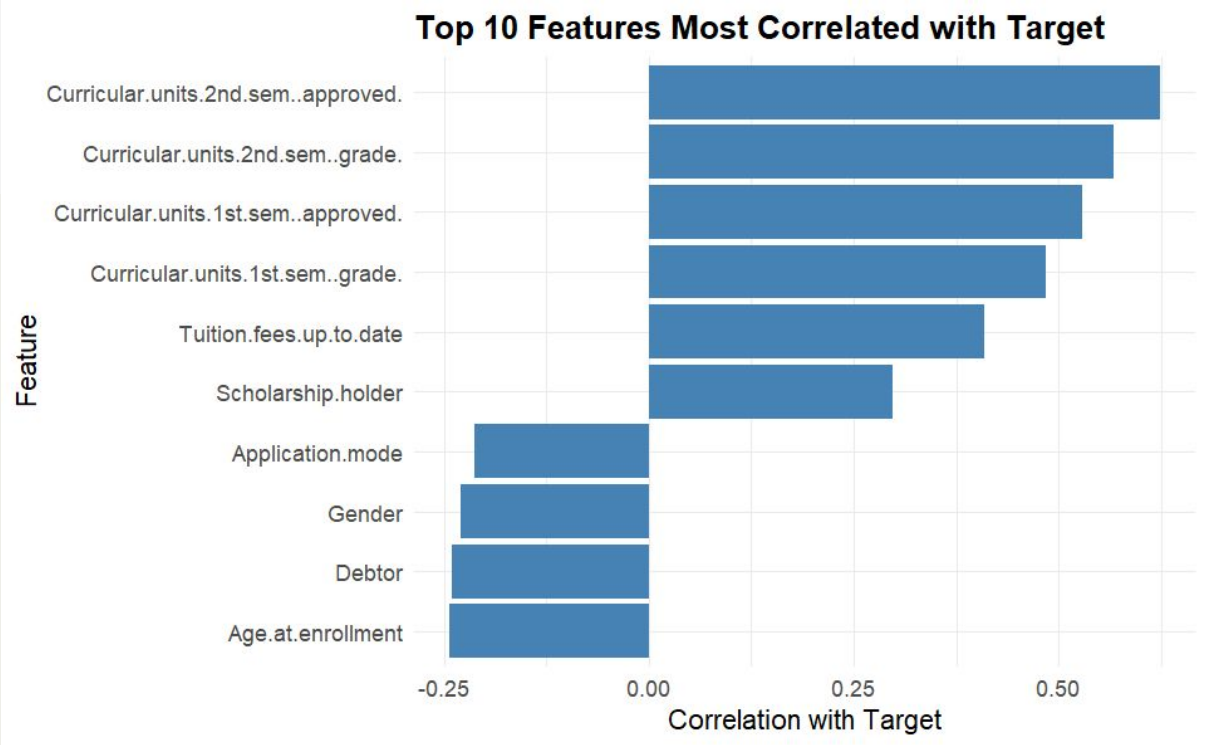
Correlations

```
data_updated |>
  cor(data_updated$Target, use = "complete.obs")
```

Marital.status	-0.0898035316
Application.mode	-0.2120247730
Application.order	0.0897909053
Course	0.0078410376
Daytime.evening.attendance	0.0751065009
Previous.qualification	-0.0913651925
Nacionality	-0.0047403981
Mother.s.qualification	-0.0383464474
Father.s.qualification	0.0003288101
Mother.s.occupation	0.0484242495
Father.s.occupation	0.0517017935
Displaced	0.1139855700
Educational.special.needs	-0.0073530732
Debtor	-0.2409989028
Tuition.fees.up.to.date	0.4098267547
Gender	-0.2292695743
Scholarship.holder	0.2975952914
Age.at.enrollment	-0.2434375114
International	0.0039339935
Curricular.units.1st.sem..credited.	0.0481497148
Curricular.units.1st.sem..enrolled.	0.1559739685
Curricular.units.1st.sem..evaluations.	0.0443615535
Curricular.units.1st.sem..approved.	0.5291232554
Curricular.units.1st.sem..grade.	0.4852073909
Curricular.units.1st.sem..without.evaluations.	-0.0687018194
Curricular.units.2nd.sem..credited.	0.0540038083
Curricular.units.2nd.sem..enrolled.	0.1758468240
Curricular.units.2nd.sem..evaluations.	0.0927206488
Curricular.units.2nd.sem..approved.	0.6241574640
Curricular.units.2nd.sem..grade.	0.5668272799
Curricular.units.2nd.sem..without.evaluations.	-0.0940277704
Unemployment.rate	0.0086266814
Inflation.rate	-0.0268740649
GDP	0.0441346899

Correlations

```
ggplot(top10_cor, aes(x = reorder(Feature, Correlation), y = Correlation)) +
  geom_bar(stat = "identity", fill = "#4682B4") +
  coord_flip() +
  labs(title = "Top 10 Features Most Correlated with Target",
       x = "Feature",
       y = "Correlation with Target") +
  theme_minimal() +
  theme(text = element_text(size = 12),
        plot.title = element_text(size = 14, face = "bold"),
        axis.title = element_text(size = 12))
```



03

Solution

Random Forest

```r
# Split data
set.seed(123)
trainIndex <- createDataPartition(data_clean$Target, p = 0.8, list = FALSE)
train_data <- data_clean[trainIndex, ]
test_data <- data_clean[-trainIndex, ]

# Apply SMOTE to training data
smote_train <- SMOTE(
  X = train_data |> select(-Target),
  target = train_data$Target,
  K = 5,
  dup_size = 0
)$data |>
  rename(Target = class)

# Train Random Forest model
rf_model <- randomForest(
  as.factor(Target) ~ .,
  data = smote_train,
  ntree = 500,
  importance = TRUE
)
```

```r
# Evaluate on test set
test_pred <- predict(rf_model, test_data, type = "prob")[,2]
test_roc <- roc(test_data$Target, test_pred)
cat("Test Set AUC:", auc(test_roc), "\n")

# Generate predictions for all students
final_predictions <- predict(rf_model, data_clean, type = "prob")[,2] %>%
  {data.frame(
    Student_ID = rownames(data_clean),
    Dropout_Probability = .,
    Predicted_Dropout = ifelse(. >= 0.5, "High Risk", "Low Risk"),
    Actual_Status = ifelse(data_clean$Target == 1, "Dropout", "Retained")
  )}
```

	Student_ID <chr>	Dropout_Probability <dbl>	Predicted_Dropout <chr>	Actual_Status <chr>
1	1	1.000	High Risk	Dropout
2	2	0.062	Low Risk	Retained
3	3	1.000	High Risk	Dropout
4	4	0.094	Low Risk	Retained
5	5	0.014	Low Risk	Retained
6	6	0.192	Low Risk	Retained

6 rows

85.6%

accuracy

84%

balanced accuracy

Accuracy <dbl>	Balanced_Accuracy <dbl>	Precision <dbl>	Recall <dbl>	F1 <dbl>	AUC <dbl>
0.8563348	0.8401222	0.9016949	0.8851913	0.8933669	0.88916

Comparison with Research Papers

Research Papers from:

- Generalized Linear Mixed Models with 96% accuracy
- The number of credits earned in the first semester emerges as the most significant predictor for both early and late dropout.

- ⇧Number of failed courses → ⇧ dropout likelihood.
- Lower engagement levels, as indicated by reduced activity in the LMS, are associated with higher dropout rates.

- Moodle activity: frequency of logins, time spent on the platform, and interaction with course materials, were significant indicators of potential dropout.

What about HKU?

~98%

Graduation rate

Low amount of data

04

Conclusion

Limitations

1. Too many factors
2. Lack of context
3. Lack of balanced data

The main conclusion is:

believe in yourselves

exams are coming, but remain calm and strong, guys

Sources:

Data:

- https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data
- https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

GitHub:

- https://github.com/shivamsingh96/Predict-students-dropout-and-academic-success?tab=readme-ov-file

Research Papers:

- https://www.researchgate.net/journal/Studies-In-Higher-Education-1470-174X
- https://www.sciencedirect.com/science/article/pii/S0160791X24000228
- https://www.researchgate.net/publication/384977767_Predicting_Student_Dropout_Using_Machine_Learning_Algorithms
- https://www.nature.com/articles/s41598-025-93918-1
- https://www.mdpi.com/2306-5729/8/3/49

Thanks for attention!