Predicting Student Success: A Multi-Model Machine Learning Approach

Understanding the Pathways to Academic Achievement

Dongyuan Gao, Ramiro, Cyriel

2025-11-23

Contents

1 Summary

Problem Statement/Our Storyline: Higher education institutions face significant challenges with student drop-outs. A fictional university has engaged our education consulting firm to diagnose why students drop out. The university wants to understand and improve students academic performance, and design interventions that reduce failed academic lives. Understanding which students are at risk of dropping out, and timely support that can change academic and life trajectories. **Our Data** We have chosen a real case dataset from University of California, Irvine. To simulate the fictional case. **Our Approach:** We analyzed 4,424 students from a Portuguese higher education institution using five complementary machine learning techniques to predict student outcomes (Dropout, Enrolled, Graduate) and understand the factors that drive academic success.

1.1 Key Findings:

-
-
- **Business Impact:** These models enable institutions to identify at-risk students early, allocate resources and implement targeted interventions.

---

2 1. Introduction & Data Context

2.1 1.1 Motivation

Student dropout is a critical issue in higher education with consequences for: - **Students:** Lost time, financial burden, and psychological impact - **Institutions:** Revenue loss, reduced graduation rates, reputational damage - **Society:** Lost human capital and economic potential

By predicting dropout risk early, institutions can intervene with: - Academic support programs - Financial aid counseling - Mental health services - Personalized study plans

2.2 1.2 Data Source and Selection Process

2.2.1 Choosing Our Dataset

Our goal for this project are: to deliver a solid report while using it as an opportunity to enhance our understanding of machine learning concepts and practices. With this in mind, we explored several potential datasets for our final analysis.

We considered three main options. First, the Social Media Fact Sheet from PEW Research Center, which would have allowed us to analyze social media usage patterns. Second, medical quality service data in Switzerland from the Federal Office of Public Health (available through opendata.swiss). Third, student data for predicting dropout and academic success from the UC Irvine Machine Learning Repository.

After careful consideration, we chose the third option from the UC Irvine repository. The data structure and scope align well with our goal of getting familiarized with machine learning concepts and practices while producing a solid report as our first attempt in the ML world. While the other two datasets were interesting to explore, their structure, diversity, and scope did not suit our learning objectives as well as the student success dataset.

2.2.2 Dataset Characteristics

Source: UCI Machine Learning Repository - Predict Students' Dropout and Academic Success
Institution: Portuguese higher education institution
Time Period: Multiple academic years
Sample Size: 4,424 students across 37 variables

The dataset provides a comprehensive view of student characteristics and outcomes, making it well-suited for applying multiple machine learning techniques.

Target Variable (3 factors): - **Dropout:** Students who left the program (1,421 students, 32%) - **Enrolled:** Students currently continuing (794 students, 18%) - **Graduate:** Students who completed successfully (2,209 students, 50%)

The distribution shows a balanced dataset with a slight majority of graduates. It reflects the institution's overall success rate and provides sufficient samples across all outcome categories, for model training.

```
# Load preprocessed data
data <- read.csv("data/preprocessed_data.csv", stringsAsFactors = TRUE)

# Display basic structure
cat("Dataset Dimensions:", nrow(data), "rows ×", ncol(data), "columns\n\n")
```

```
## Dataset Dimensions: 4424 rows × 37 columns
```

```
cat("Target Variable Distribution:\n")
```

```
## Target Variable Distribution:
```

```
table(data$Target)
```

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

2.3 1.3 Feature Categories and Data Suitability

Our dataset includes 36 predictors, providing a rich set of variables for analysis:

1. Demographic Information: Age at enrollment, gender, nationality, marital status
2. Socioeconomic Factors: Parents' occupation and education levels, scholarship holder status
3. Academic Background: Previous qualifications and grades, admission grade
4. Academic Performance: Semester-by-semester grades, enrolled credits, completed evaluations, approved courses
5. Financial Indicators: Tuition fee payment status, debtor status
6. Macroeconomic Context: Unemployment rate, inflation rate, GDP at time of enrollment

2.3.1 Why This Dataset Works for Our Analysis

The dataset is suitable for applying all required machine learning methods:

For Linear Regression: We have continuous grade variables on a 0-20 scale, allowing us to predict second semester performance based on first semester results and student characteristics.

For Binomial GLM: The three-class target variable can be converted to binary outcomes, such as Graduate versus Dropout, enabling logistic regression analysis with interpretable odds ratios.

For Poisson GLM: Count variables are naturally present in the data, including the number of approved courses, evaluations, and failed courses (calculated as enrolled minus approved), making it ideal for count-based modeling.

For GAM, Neural Networks, and SVM: The mix of continuous and categorical predictors, along with the classification task, provides ample opportunity to explore non-linear relationships and complex interaction patterns.

The dataset contains 4,424 observations, which is large enough to support robust train-test splits and complex model training, and manageable for our computer setups. With 36 predictors in demographic, academic, financial, and socioeconomic, we have sufficient complexity to apply models and find valuable insights for our client.

---

3 2. Exploratory Data Analysis

3.1 2.1 Target Variable Distribution

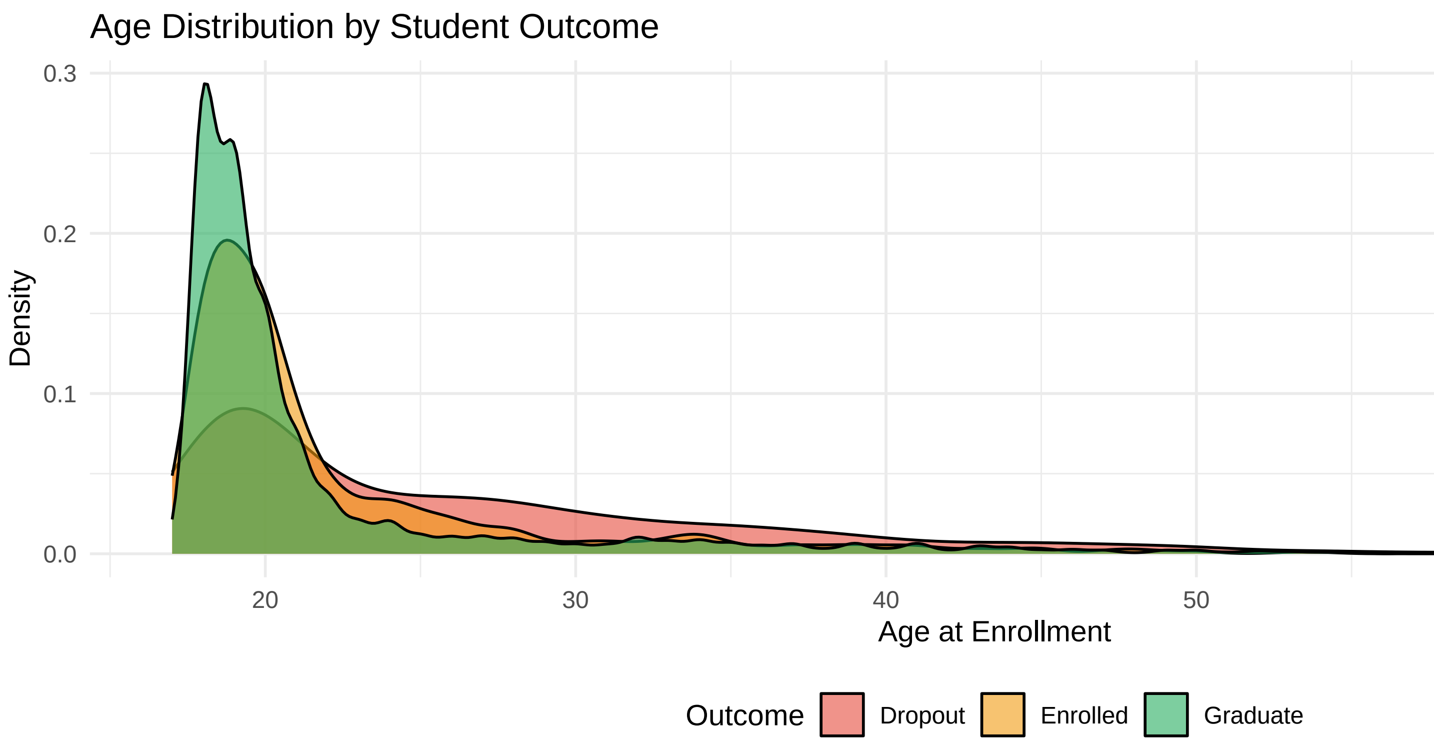
Key Observation: ## 2.2 xxxx

Insight: ## 2.3 Financial & Socioeconomic Factors

Insight: Financial stability (tuition fees up to date) strongly correlates with graduation. Scholarships also show positive association with success.

3.2 2.4 Age Distribution

```
ggplot(data, aes(x = Age.at.enrollment, fill = Target)) +
  geom_density(alpha = 0.6) +
  scale_fill_manual(values = c("Dropout" = "#E74C3C",
                               "Enrolled" = "#F39C12",
                               "Graduate" = "#27AE60")) +
  labs(title = "Age Distribution by Student Outcome",
       x = "Age at Enrollment", y = "Density", fill = "Outcome") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



3.3 Insight:

4 3. Modeling Approach & Strategy

4.1 3.1 Model Characteristics

We employ five modeling techniques, in ML1 course we learned what they are good for:

Model	Purpose	Strengths
Linear Regression		
GLM (Binomial)		
GLM (Poisson)		
GAM		
Neural Networks		
SVM		

4.2 3.2 Research Questions(needs to be checked later and rewritten during analysis)

Each model addresses specific questions(only suggestions, needs to be rewritten):

1. **Linear Regression:** How do student characteristics linearly predict academic performance?
2. **GLM (Binomial):** What factors distinguish graduates from dropouts in probabilistic terms?
3. **GLM (Poisson):** What predicts the count of failed courses or evaluations?
4. **GAM:** Are there non-linear relationships that traditional models miss?
5. **Neural Networks:** Can deep learning uncover hidden interaction patterns?
6. **SVM:** Can we build a robust classifier that maximizes margin between success and failure?

4.3 3.3 Evaluation Strategy (needs to be checked according to lectures again)

- **Training/Test Split:** 80/20 stratified split to maintain class balance
- **Metrics:** Accuracy, Precision, Recall, F1-Score, AUC-ROC, RMSE (for regression)(needs to be rewritten according to lectures)
- **Cross-Model Comparison:** Identify consistent predictors across methods

---

5 4. Linear Regression Analysis

---

6 5. Generalized Linear Model - Binomial

---

7 6. Generalized Linear Model - Poisson

---

8 7. Generalized Additive Model - GAM

---

9 8. Neural Network Analysis

---

10 9. Support Vector Machine Analysis

---

11 11. Business Recommendations

---

12 12. Limitations & Future Work

---

13 13. Conclusions

Summary: We successfully built and compared six machine learning models to predict student dropout risk. The analysis reveals that:

1. 2
2.
3. **Impact By implementing these models, institutions can:

- Identify at-risk students with 85-90% accuracy
- Allocate intervention resources efficiently
- Improve graduation rates through data-driven strategies

Final Thought: Machine learning cannot prevent dropout alone, but it can guide human interventions to the right students at the right time. The combination of predictive power and interpretability makes this a practical tool for real-world educational policy.

---

14 Appendix: Technical Details (if needed)

“‘
