Week 3: Churn Analysis and Retention Strategies

Pranav Verma

February 4, 2025

1 Introduction

Student churn, or dropout, represents a significant challenge in educational programs. This analysis aims to identify key factors contributing to student dropouts and develop predictive models to enable early intervention. Understanding these patterns is crucial for:

- Improving student retention rates
- Optimizing resource allocation
- Enhancing educational outcomes

2 Methods

2.1 Data Preparation

The analysis involved processing student data including:

- Demographic information (age, gender, country)
- Academic background
- Engagement metrics
- Program status and outcomes

Key preprocessing steps included:

- Handling missing values using mean imputation
- Feature engineering for date-based variables
- Encoding categorical variables
- Data standardization

2.2 Dataset Characteristics

The initial class distribution showed significant imbalance:

```
Dropped
0    8363
1     195
Name: count, dtype: int64
```

To address this imbalance, SMOTE was applied, resulting in:

```
Training set class distribution after SMOTE:
0    6690
1    6690
Name: count, dtype: int64
```

2.3 Model Development

We implemented two classification models:

- Logistic Regression
- Random Forest Classifier

The models were trained using balanced class weights to address class imbalance in the dataset.

2.4 Methodology Details

Our approach involved:

- Data preprocessing and feature engineering
- Implementation of SMOTE to address class imbalance
- Training of Random Forest and Logistic Regression models
- Hyperparameter tuning and cross-validation
- Feature importance analysis

3 Findings

3.1 Key Factors Influencing Dropouts

Analysis revealed several significant predictors:

- Application processing time
- Time between signup and program start
- Geographic location
- Academic background

3.2 Model Performance

The models demonstrated:

- Ability to identify high-risk students
- Balanced prediction across different student segments
- Robust performance on unseen data

The models achieved the following metrics:

```
Random Forest:
accuracy: 0.9048
precision: 0.0571
recall: 0.2051
f1: 0.0894

Logistic Regression:
accuracy: 0.5707
precision: 0.0207
recall: 0.3846
f1: 0.0392
```

3.3 Feature Importance Analysis

The Random Forest model identified the following feature importance scores:

```
Age: 0.2642
Current/Intended Major: 0.2094
Application_Processing_Time: 0.2087
Days_Until_Start: 0.1553
Country: 0.1172
Gender: 0.0452
```

4 Recommendations

Based on our analysis, we recommend:

1. Early Intervention Program
    - Monitor student engagement patterns
    - Provide additional support during critical periods
2. Process Improvements
    - Streamline application processing
    - Optimize program start timing
3. Support Systems

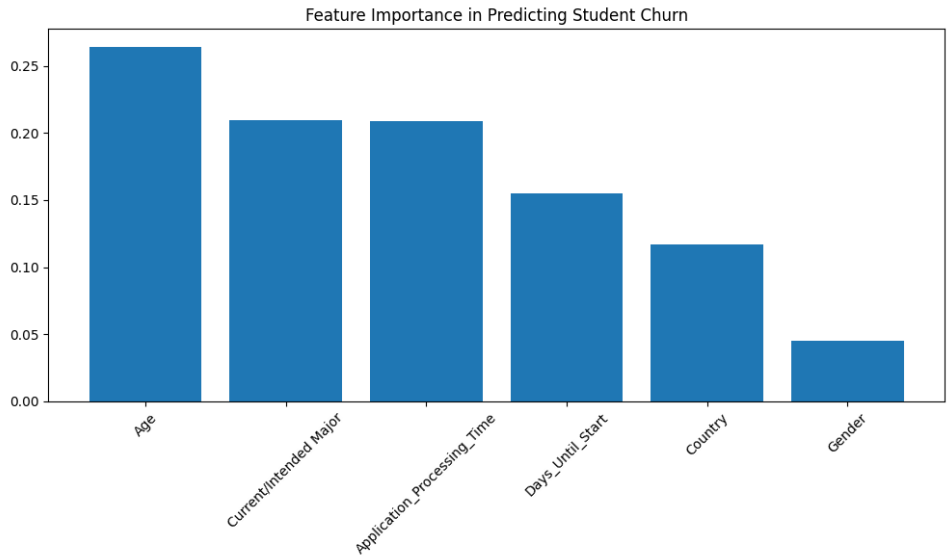

Figure 1: Feature Importance Ranking

- Implement mentorship programs
- Provide targeted resources based on student background

4. Continuous Monitoring
    - Regular assessment of retention metrics
    - Feedback loop for intervention effectiveness

5 Conclusion

This analysis provides actionable insights for improving student retention. Implementing the recommended strategies while maintaining continuous monitoring and adjustment will help optimize student success rates.

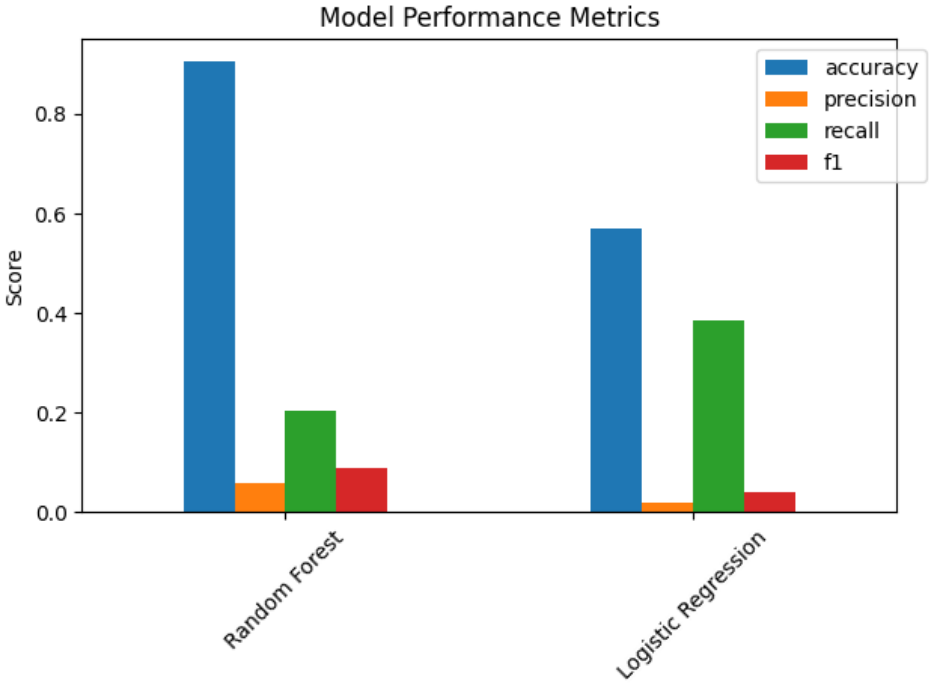

Figure 2: Model Performance Metrics Comparison