PREDICTING STUDENT DROPOUT RISK USING LOGISTIC REGRESSION

Executive Summary

This case study demonstrates how to use logistic regression to predict student dropout risk based on academic, behavioral, and demographic data. By applying data cleaning, feature selection, model building, and performance evaluation techniques, this project provides students with a practical example of how predictive analytics can solve real educational problems. It includes step-by-step code, interpretation of coefficients, and visualizations, helping learners build both statistical understanding and practical data science skills.

1. Introduction

Educational institutions face high dropout rates, especially in early semesters. Predicting which students are at risk can enable timely interventions and improve retention. This case study models dropout probability using logistic regression, one of the most interpretable and widely used classification techniques. The dataset includes variables such as GPA, attendance rate, parental education, course load, and previous semester performance.

2. Problem Statement

Build a binary classification model to predict whether a student is **likely to drop out (1)** or **continue (0)** based on their historical and demographic data.

3. Dataset Overview

- **Source**: Open Educational Data (simulated for academic use)
- **Observations**: 1,000 students
- **Features**:
    - age (numeric)
    - gender (categorical)
    - GPA_last_sem (numeric)
    - attendance_rate (numeric, %)
    - parental_education (ordinal)

- study_hours_per_week (numeric)
- course_load (categorical: full-time/part-time)
- dropout (target variable: 1 = dropped out, 0 = continued)

4. Methodology

Step 1: Data Cleaning

- Handled missing values using mean imputation for GPA and attendance
- Converted categorical variables using one-hot encoding (gender, course load)
- Normalized numerical features

Step 2: Exploratory Data Analysis

- Dropout rate = 23.4%
- Strong correlation observed between low GPA, low attendance, and higher dropout

Step 3: Model Building (Logistic Regression in Python)

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix

X = df.drop('dropout', axis=1)
y = df['dropout']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

model = LogisticRegression()
model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

5. Results and Interpretation

Confusion Matrix

[[187 13]

 [24 76]]

Accuracy: 87.7%

Precision (Dropout): 85.4%

Recall (Dropout): 76.0%

Key Coefficients:

Feature	**Coefficient (β)**	**Interpretation**
GPA_last_sem	-1.92	Each unit increase in GPA reduces dropout likelihood
Attendance Rate	-0.88	Higher attendance lowers dropout risk
Study Hours/Week	-0.43	More study hours reduce risk
Part-Time Course Load	+1.26	Higher risk of dropout for part-time students

6. Visualization

- **ROC Curve** shows AUC of 0.91
- **Bar Chart** of top 5 features ranked by coefficient magnitude
- **Heatmap** showing correlation between features and dropout

7. Conclusion

The logistic regression model accurately predicts student dropout with interpretable coefficients. Key indicators include GPA, attendance, study hours, and course load type. This type of model can be integrated into student monitoring dashboards or academic risk management systems.

8. Learning Outcomes for Students

- Understand how binary classification works using logistic regression
- Learn how to clean and encode real-world data

- Interpret model output beyond just accuracy (coefficients, odds, precision/recall)
- Gain exposure to real applications of predictive analytics in education

9. Suggested Enhancements

- Apply regularization (L1/L2) to improve generalization
- Test more complex models (Random Forest, XGBoost) for comparison
- Use SHAP values for detailed model interpretability
- Deploy as a simple web tool using Flask + Streamlit

10. References

- Hosmer, D.W., & Lemeshow, S. (2000). *Applied Logistic Regression*
- Scikit-Learn Logistic Regression Documentation
- UCI Educational Data Mining Repository
- Kaggle: Student Performance Dataset
- Feldman & Grossman (2021). *Data Science for Education Analytics*