Ant-X Student Academic Risk Early-Warning Model

Shengdi Lu

Centre for Student Affairs

Supported by LM

1. Background

(1) Limitations of Ant-X Phase 1 (2017)

- **Limited Coverage**: Only targeted mainland undergraduate students, excluding overseas students and postgraduate (master’s/doctoral) groups.
- **Insufficient Data Governance**: Incomplete university data collection scope; risk warnings relied on post-hoc statistics (e.g., delayed course failure alerts) with no advance prediction; insufficient cross-departmental data collaboration led to underutilization of data value.
- **Risks from External Vendors**: External development introduced issues like data off-domain (privacy leakage) and uncontrollable algorithm black boxes.

(2) Reasons for Suspension & Restart

- **Suspension:** Pandemic-related loss of external vendors' technical backbone left no follow-up development capacity, halting the project temporarily.
- **Restart Opportunity:** In 2025, the university's mature data governance provided a foundation for risk warning; with support from two vice presidents, the project was approved as a university-level special research project, resolving historical barriers to budget transfer and cross-departmental collaboration.

2. Solutions

(1) Cross-Center Collaboration Mechanism

A cross-center working group was formed with:

- **Business Experts:** SAO & Registry, demand definition & business validation;
- **Technical Experts:** MITS, data security, system deployment & API development;
- **Model Experts:** Prof. Wang Qiufeng's team from SAT ,algorithm R&D and model optimization.

(2) University Special RDF Funded Research Project Approval

Approved as a Special RDF Funded Research Project to avoid external vendor risks, ensure data security & algorithm controllability; regular project meetings were established for efficient development.

(3) Model Development & Functional Design

- **Algorithm Models:**
  - **Unsupervised Learning**: Local Outlier Factor (LOF), Isolation Forest (supports parameter adjustment for anomaly detection);
  - **Supervised Learning**: Binary classification (risk existence) & regression (risk probability) tasks with fixed, high-precision model structures.
- **Early Warning Mechanism:**

Built a "early warning-intervention-feedback" closed loop—risk alerts are auto-assigned; Student Development Advisor(DA) record interventions & feedback on model accuracy.

(4) Launch Steps

Phased implementation:

- Phase 1: Complete academic risk model development (course failure probability, group deviation);

- Phase 2: Expand to psychological warning (short-term stress anomalies);
- Phase 3: Plan employment difficulty prediction;

Gradual system launch: Include data encryption, permission management, auto-ticket assignment, etc.

3. Outcomes and Benefits

(1) Outstanding Model Performance

Validated with historical data from AY2023-2024, all 4 core models achieved over 80% in F1-score (classification tasks) and R^2 (regression tasks), exceeding industry benchmarks and laying a solid algorithmic foundation for subsequent system application.

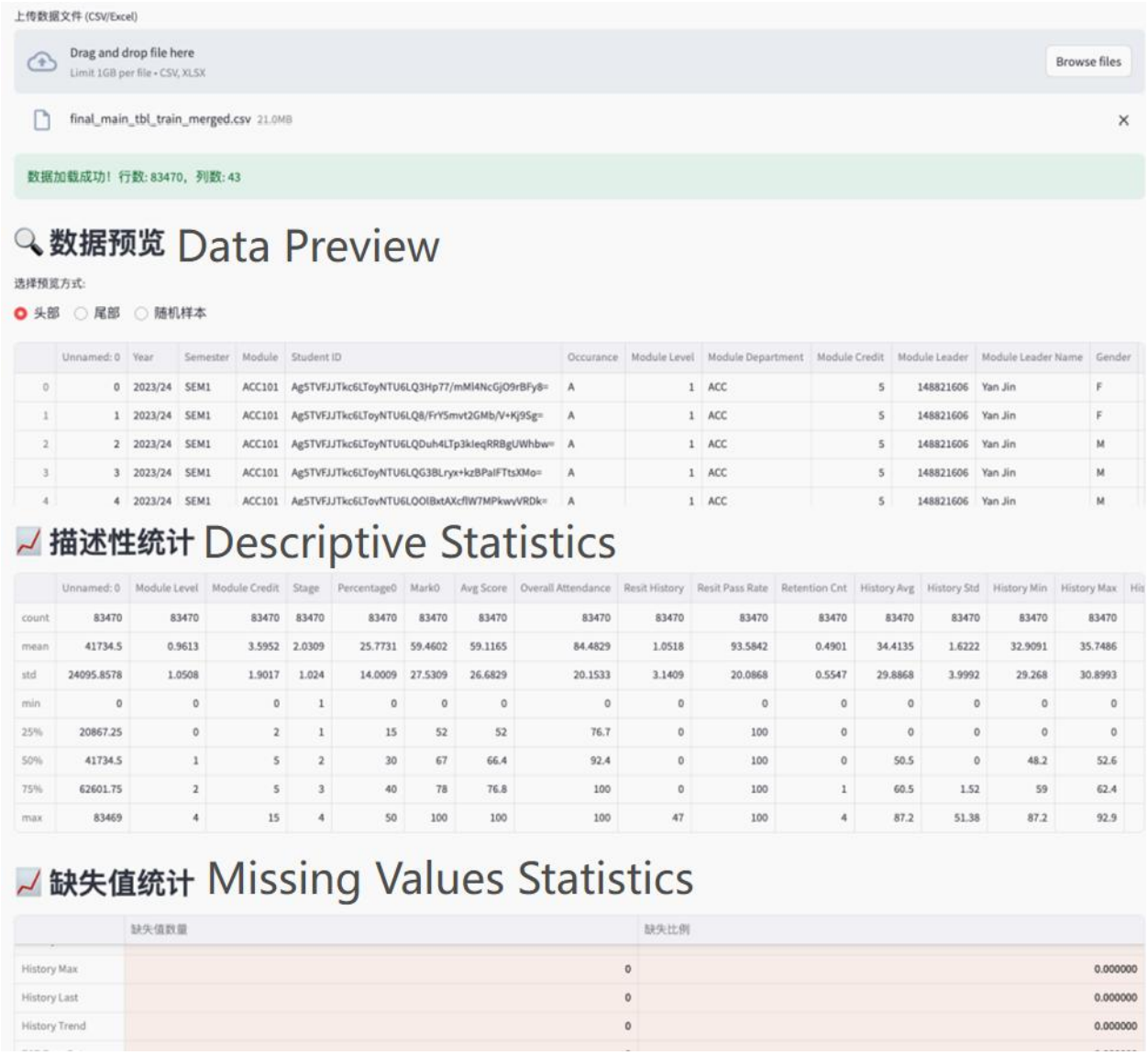

Figure 1. Model Data Upload, Preview, and Descriptive Statistics

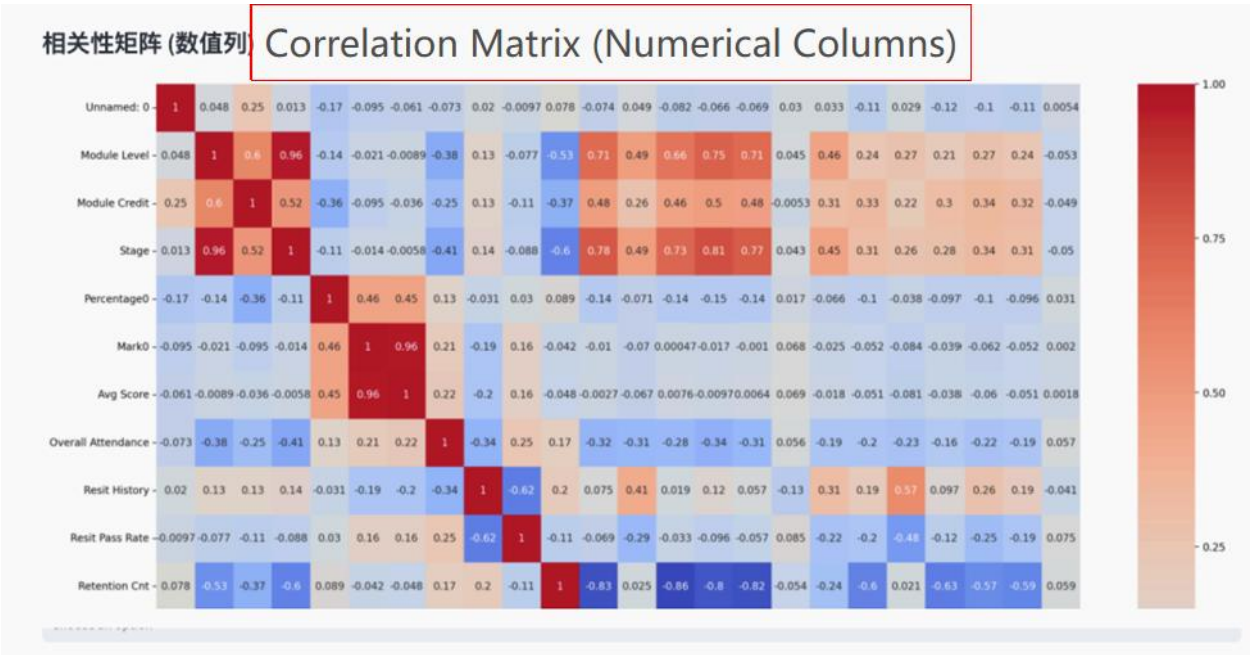

Figure 2. Correlation Matrix

```
False}. Best is trial 11 with value: 0.8686840104061164.
2025-11-13 00:09:57.419 | INFO     | cf_train:optimize_hyperparameters:142 - best model value macf1 8.6868e-0
2025-11-13 00:09:57.437 | INFO     | cf_model:create_model:140 - use model cat for regression? False ....
2025-11-13 00:12:23.220 | INFO     | cf_train:optimize_hyperparameters:146 - save the model to ./model/cat#tr
in_risk_classification.csv#mdl.joblib
2025-11-13 00:12:23.244 | INFO     | cf_train:test:206 - Loading model from ./model/cat#train_risk_classifica
ion.csv#mdl.joblib...
2025-11-13 00:12:23.646 | INFO     | cf_train:test:234 - class dist in test set: #feat 49 #num 10504
2025-11-13 00:12:23.647 | INFO     | cf_train:test:241 - Class distribution in test set:
2025-11-13 00:12:23.648 | INFO     | cf_train:test:242 - PassModuleProb
0    7808
1    2696
Name: count, dtype: int64
2025-11-13 00:12:24.222 | INFO     | cf_train:test:256 - Confusion Matrix:
2025-11-13 00:12:24.222 | INFO     | cf_train:test:257 - [[7631  177]
 [ 336 2360]]
2025-11-13 00:12:24.276 | INFO     | cf_train:test:266 - Accuracy: 0.9512
2025-11-13 00:12:24.276 | INFO     | cf_train:test:267 - Micro F1-score: 0.9512
2025-11-13 00:12:24.284 | INFO     | cf_train:test:268 - Macro F1-score: 0.9347
2025-11-13 00:12:24.386 | INFO     | cf_train:test:274 - Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      7808
           1       0.93      0.88      0.90      2696

    accuracy                           0.95     10504
   macro avg       0.94      0.93      0.93     10504
weighted avg       0.95      0.95      0.95     10504
```

Figure 3. Model Performance (Recall and F1 Score)

(2) Data Security Assurance

The in-house development and localized data deployment design completely avoids data off-domain risks from external vendors

4. Replicability and Promotion Value

(1) Model Innovation

- **Research-Driven Management**: Special RDF projects integrate cross-departmental resources to resolve budget/collaboration barriers, serving as a reference for similar university projects;
- **Interdisciplinary Collaboration**: Integrates education, psychology, and computer science frameworks, applicable to other management scenarios (e.g., psychological warning, employment prediction).

(2) Promotion Scenarios

Suitable for student risk management system construction in various universities, especially those needing multi-source data integration and external vendor risk avoidance.