Predicting Student Fee Default in a Ghanaian Private School: A Logistic Regression Approach with Ethical Deployment

Abstract

Fee default is a critical financial sustainability challenge for private schools in sub-Saharan Africa, where tuition fees constitute the primary revenue stream. This study develops and deploys a logistic regression model to predict student fee default at a private basic school in Ghana's Volta Region, using 2,280 term-level administrative payment records spanning three academic years (2023–24 to 2025–26). Fourteen predictor variables were identified from routine school management records, including payment compliance indicators, first payment behaviour, and student-level characteristics. The model achieved 89% accuracy and a ROC-AUC of 0.942, with compliance with the final payment policy emerging as the strongest protective predictor. To operationalise the findings, an interactive web-based dashboard was built using Streamlit, backed by a PostgreSQL database, and containerised with Docker, enabling real-time default risk monitoring by school administrators. All student records were anonymised prior to deployment in adherence to institutional data protection obligations, demonstrating that ethical AI principles can be applied in resource-constrained educational settings. This work shows that standard administrative data, when responsibly modelled, can transform reactive fee collection into a proactive early warning system accessible to under-resourced private schools across Ghana and West Africa.

keywords: fee default prediction, logistic regression, educational data science, ethical AI, Ghana

1 Introduction

Private schools in Ghana depend almost entirely on student fees for operational sustainability. Fee default, defined as the partial or complete non-payment of academic fees within a stipulated term, threatens institutional viability by disrupting teacher salaries, infrastructure maintenance, and educational investment (1). In the Volta Region, where private schools serve communities with variable public school quality, this challenge is particularly acute. Despite its severity, fee default in Ghanaian private basic schools remains largely unaddressed by data-driven approaches; administrators continue to rely on informal, experience-based collection strategies.

This study addresses that gap by applying logistic regression to administrative payment records from a private basic school in Ho, Ghana, to build a predictive model of fee default. The school recorded an overall default rate of 23.2% across three academic years, meaning roughly one in four term-level fee obligations goes partially or fully unmet. Beyond model performance, this work prioritises responsible deployment: student records were anonymised before any analytical or public use, and the resulting tool is designed to support, not replace, administrator judgment. In doing so, the study illustrates how data science, when grounded in *local context and applied responsibly, can meaningfully address real institutional challenges.*

2 Methods

2.1 Data and Preprocessing

The dataset comprises 2,280 term-level records representing 360 unique students across up to nine terms each. It was derived from administrative fee invoices and payment records, augmented with synthetic data generation techniques that preserved distributional realism while protecting student identities. The binary target variable `defaulted` encodes 1 (Partially Paid or Unpaid) and 0 (Fully Paid), with a class distribution of 23.2% defaulted and 76.8% fully paid.

Fourteen predictor variables were selected spanning four domains: payment behaviour (`first_payment_amount`, `total_payments_made`, `days_to_first_payment`), policy compliance (`complied_part_payment_policy`, `complied_final_payment_policy`), fee structure (`fee_amount`, `discount_type`, `post_restructuring`), and student characteristics (`level`, `siblings_enrolled`, `is_staff_child`, `has_ucmas`). Categorical variables were label-encoded and numeric features were standardised using `StandardScaler` prior to model fitting. Missing values in `days_to_first_payment` for unpaid records (11.58%) were imputed with a sentinel value of 999, encoding the absence of any payment as a strong default signal.

2.2 Modelling and Evaluation

A logistic regression classifier (2) was trained on an 80/20 stratified train-test split (1,824 training, 456 test records), preserving the 23.2% default rate in both partitions. The logistic regression model takes the form:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{1}$$

where p is the predicted probability of default and β_i are coefficients estimated by maximum likelihood. Principal Component Analysis (PCA) (3) was applied for exploratory dimensionality analysis; the original 14 features were retained for modelling to preserve coefficient interpretability. Shapiro-Wilk normality tests confirmed non-normal distributions for all numeric predictors ($p < 0.001$), validating logistic regression as the appropriate method. Model performance was evaluated on accuracy, precision, recall, F1-score, and ROC-AUC.

To operationalise the model, an interactive dashboard was developed using Streamlit and Plotly, backed by a PostgreSQL 15 database and containerised with Docker Compose. The dashboard provides five modules: an overview with KPI cards and default trend charts, exploratory feature analysis, default risk factor breakdowns, live model training and evaluation, and a student risk lookup. All student names and identifiers were anonymised prior to any analytical or deployment use, with the raw dataset excluded from the public repository.

3 Results

The logistic regression model achieved 89% overall accuracy and a ROC-AUC of 0.942 on the held-out test set, indicating strong discriminative ability. Table 1 summarises the full evaluation metrics. The model correctly identified 65% of actual defaulters (recall), with 82% precision for the default class.

Table 1: Logistic Regression Model Performance on Test Set ($n = 456$)

Metric	Value
Overall Accuracy	89%
ROC-AUC Score	0.942
Precision (Default)	0.82
Recall (Default)	0.65
F1-Score (Default)	0.73
F1-Score (Weighted Avg)	0.88

Compliance with the final payment policy was the strongest predictor (coefficient $= -2.91$), followed by `days_to_first_payment` as the strongest risk factor (coefficient $= +1.63$). Notably, the raw Pearson correlation between `days_to_first_payment` and default was only $r = 0.02$, yet the multivariate model revealed a coefficient of $+1.63$, illustrating the limitation of bivariate screening for feature selection. Exploratory analysis revealed a default rate of approximately 40% in post-fee-restructuring periods, compared to 12% pre-restructuring, and Term 2 consistently exhibited the highest default rate ($\approx$27%), consistent with post-Christmas household expenditure pressure (5).

4 Discussion and Conclusion

This study demonstrates that routine administrative records from a Ghanaian private basic school contain sufficient predictive signal to build a high-performing fee default model, achieving results that meet or exceed comparable benchmarks in the credit risk literature (4; 6). The deployed dashboard translates model outputs into actionable intelligence: administrators can flag at-risk students within the first two weeks of a term, before financial shortfalls materialise, transforming fee management from reactive collection into proactive intervention.

The study's ethical design is a deliberate feature, not an afterthought. Student data was anonymised at source, the model is transparent and interpretable, and the system is framed as a decision-support tool that augments rather than supplants human judgment. This approach reflects a broader argument: that ethical, inclusive AI in Ghana does not require large datasets, proprietary infrastructure, or external expertise. It requires locally grounded problems, responsible data handling, and tools built to serve the communities that generate the data.

Limitations include the use of synthetic augmentation for privacy compliance, which may constrain generalisability, and a default-class recall of 65%, which future work could improve through ensemble methods such as gradient boosting (4). Retraining the model annually as new payment data accumulates will further strengthen its operational utility.

The complete source code and dashboard implementation are publicly available upon request.

References

[1] Acheampong, F. (2021). Determinants of school fee payment behaviour in private basic schools in Ghana: Evidence from the Ashanti Region. *African Journal of Educational Management*, 19(2):45–63.

[2] Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[3] Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202.

[4] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 247(1):124–136.

[5] Quaye, E., Asante, B., & Owusu, G. (2019). Fee collection challenges in private basic schools in Greater Accra, Ghana. *International Journal of Educational Administration and Policy Studies*, 11(1):1–11.

[6] Tripathi, S., Bhardwaj, A., & Poonia, R.C. (2021). Prediction of student loan default using machine learning: A comparative study. *Education and Information Technologies*, 26(5):5675–5694.