3rd International Conference on Big Data, IoT and Machine Learning (BIM 2025)

Paper ID: 499

Paper Title:

A Socio-Economic machine Learning Framework for Predicting Programmer Retention

Presenter: Md. Mehedi Hasan

Date: 26/09/2025

Organizer

Publication Partner

Support Partner

Authors

Md. Mehedi Hasan, *Dept. of CSE, Dhaka International University.*

Rasheduzzaman Rakib, *Dept. of CSE, Dhaka International University.*

Md. Asraful Molla, *Dept. of CSE, Dhaka International University.*

Rownak Borhan, *Dept. of CSE, Dhaka International University.*

Md. Abdul Based, *Dept. of CSE, Dhaka International University.*

Contents

Abstract

The retention of computer science graduates in the technology workforce remains a critical challenge, especially in developing countries like Bangladesh. This study proposes a machine learning-based framework to predict programmer retention by analyzing socio-economic, academic and motivational factors. Using data from 188 students across 21 Bangladeshi universities, we evaluate multiple classification models, with Logistic Regression showing the best performance as 0.91 accuracy with 0.94 Area Under the ROC Curve (AUC). The analysis highlights the impact of factors such as age, prior programming exposure, income and geographical location on long-term tech career intent. This research offers actionable insights for education policymakers and contributes to data-driven approaches in tech workforce development.

Introduction

- Digital transformation driving demand for tech professionals.
- Retention of CS graduates is critical, especially in LMICs.
- Many leave early- talent leakage.
- Our study: ML framework for predicting programmer retention.

Problem Statement

- Despite high enrollment, many graduates exit tech workforce prematurely.
- Causes: socio-economic, motivational, academic, and cultural barriers.
- Lack of predictive models for retention in LMICs like Bangladesh.

Related Work

Why these works are important?

- Tasmin et al. (2019): Identified socio-cultural & parental barriers in Bangladesh CS education.
- Ahmed et al. (2022): Showed long-term institutional & workplace biases in Bangladesh.
- UNESCO (2019): Highlighted motivation & supportive environment as critical in STEM globally.

Difference Between Their Work and Ours

1. Tasmin et al. (2019)

- Their Work: Qualitative survey → parental support, gender norms, safety concerns.
- Our Work: Quantitative ML model → socio-economic + motivational + academic features.
- Key Difference: They explained "why barriers exist," we predict who stays/who leaves with measurable accuracy.

2. Ahmed et al. (2022)

- Their Work: Longitudinal study → institutional/workplace biases, exclusion in CS.
- Our Work: Predictive modeling → focuses on career retention intention using real dataset.
- Key Difference: They described challenges, we provide data-driven foresight + policy simulations.

3. UNESCO (2019)

- Their Work: Global perspective → motivational/environmental support factors.
- Our Work: Bangladesh-specific dataset → localized socio-economic realities (income, rural–urban gap).
- Key Difference: UNESCO offered broad policy guidance, we deliver contextual, actionable, and reproducible ML framework.

Limitations of Their Work

- Tasmin et al. (2019): Only small-scale, qualitative → no predictive power.
- Ahmed et al. (2022): Context-specific, descriptive → limited generalizability.
- UNESCO (2019): Global analysis → lacks Bangladesh-specific insights.

Research Questions

- What socio-economic, academic, and motivational factors influence retention?
- Can ML predict who stays or leaves?
- How can predictive insights guide policy?

Objectives

- Build ML framework to forecast programmer retention.
- Identify key determinants.
- Simulate policy interventions (scholarships, rural access, early coding).

Outcomes & Impacts

- Logistic Regression: 91% accuracy, AUC 0.94.
- Key predictors: family encouragement, coding experience, CGPA, income.
- Actionable insights for policymakers.
- Supports Digital Bangladesh goals.

Methodology

- Survey of 188 students from 21 universities.
- Features: demographics, family, academics, motivation, community factors.
- Preprocessing: one-hot encoding, imputation, binarization.
- Correlation analysis: age, gender, family encouragement, CGPA, location.
- Algorithms: LR, RF, SVM, KNN, NB, XGBoost, LGBM, CatBoost, MLP.
- Validation: 5-fold CV + bootstrapped CI.

System Architecture

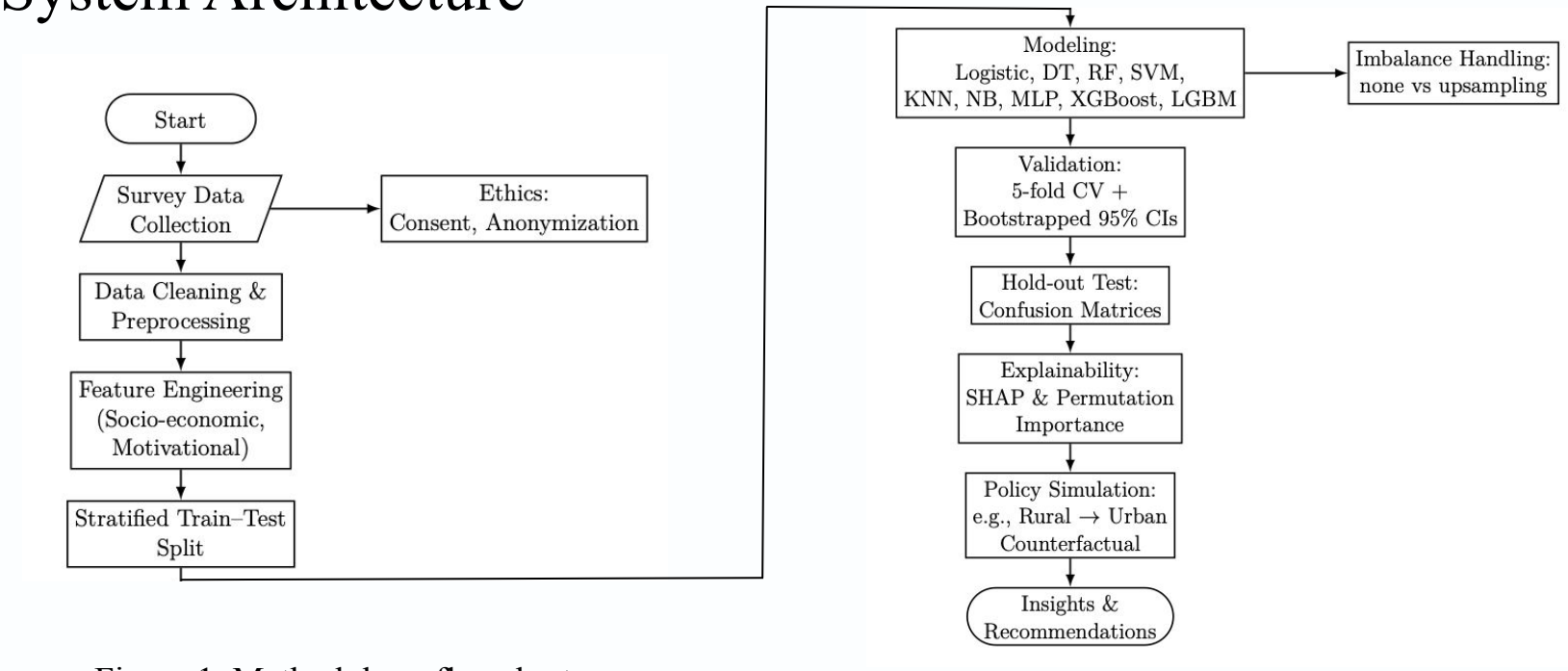

Figure 1. Methodology flowchart.

Summary of Dataset Attributes

Category	Attribute	Description
Demographics	Age	Respondent's age in years
	Gender	Male or Female
	Living Area	Urban / Suburban / Rural
Family Background	Household Income	Monthly family income bracket (e.g., 00-50k, 50k–100k BDT)
	Parental Education	Higest parental education attainment
	Parental Tech Background	Whether parents have technical-field backgrounds
Academic Background	University Type	Public, Private, or Technical institution
	Year of Study	Current academic year (e.g., 2nd year, final year)
	CGPA	Current cumulative grade point average
	Exposure to CS Courses	Number/type of CS courses taken
	Competitive Programming Experience	Participation in programming contests or platforms
Career Intentions	Career Goals	Aspirational roles (e.g., Software Engineer, Data Scientist)
	Expected Salary	Anticipated post-graduation salary
	Tech Career Intention (Target)	Intent to pursue a tech career (Yes/No)
Motivational Factors	Interest in Technology	Self-reported interest in the tech field
	Role Models	Identification with tech role models
	Motivation for Choosing CSE	Intrinsic interest vs external/family pressure
	Gender Barrier Perception	Perceived difficulty for women in tech
Prior Exposure	Pre-University Programming Experience	Coding experience before university
	Workshop/Seminar Participation	Attendance at tech training/events
	Self-Learning Engagement	Independent learning via online/offline resources
Community Factors	Women-in-Tech Involvement	Involvement or awareness of women-supportive tech groups
	Peer Encouragement	Peer support influence
Barriers	Transport Issues	Commuting difficulties to educational/tech events
	Time Constraints	Balancing academic and personal commitments
	Social Expectations	Societal or family pressures against tech careers
Programming Skills	Programming Proficiency	Self-assessed skill: Beginner, Intermediate, Advanced

Table 1. Summary of Dataset Attributes.

Results: Model Performance

- Logistic Regression best: Acc 0.91, AUC 0.94
- XGBoost competitive: Acc 0.88 AUC 0.90

Model	Precision (C1)	Recall (C1)	F1 (C1)	AUC
Logistic Regression	0.95	0.90	0.92	**0.94**
XGBoost	0.93	0.85	0.89	0.90
Random Forest	0.92	0.80	0.86	0.85
SVM	0.89	0.92	0.90	0.73
MLP	0.91	0.84	0.87	0.69
KNN	0.92	0.76	0.83	0.68
Naive Bayes	0.88	1.00	0.94	0.70
Decision Tree	0.87	1.00	0.93	0.44
LightGBM	0.90	1.00	0.95	0.79
CatBoost	0.90	1.00	0.95	0.61

Table 2. Evaluation Metrics for Class 1 (Retainers).

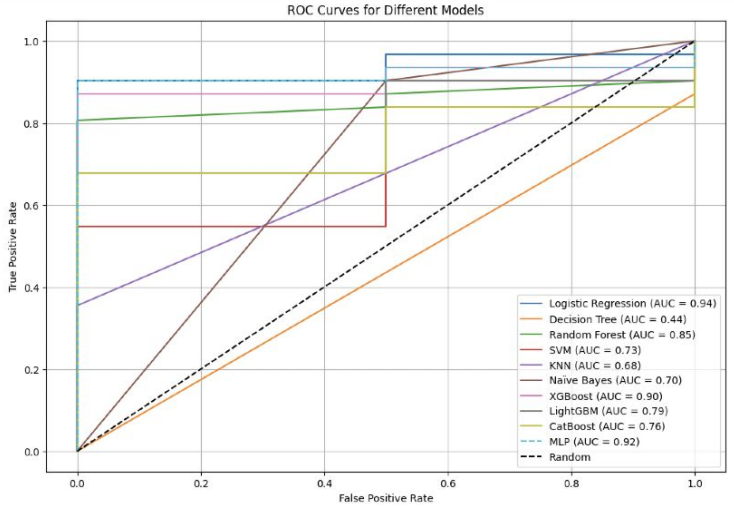

Figure 2. Applied machine learning models AUC.

Results: SMOTE Impact

- Class imbalance: 91.5% retainers, 8.5% non-retainers.
- SMOTE balancing improved minority detection.
- Trade-off: accuracy vs fairness.

Model	Before SMOTE				After SMOTE			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Logistic Regression	0.40	1.00	0.57	0.91	0.00	0.00	0.00	0.94
XGBoost	0.33	1.00	0.50	0.88	0.00	0.00	0.00	0.91
Random Forest	0.25	1.00	0.40	0.82	0.00	0.00	0.00	0.94
SVM	0.25	0.50	0.33	0.88	0.00	0.00	0.00	0.94
MLP	0.09	1.00	0.17	0.85	0.00	0.00	0.00	0.94
KNN	0.09	1.00	0.17	0.39	0.06	1.00	0.11	0.06
Decision Tree	0.00	0.00	0.00	0.82	0.00	0.00	0.00	0.82
Naive Bayes	0.00	0.00	0.00	0.88	0.25	0.50	0.33	0.88
LightGBM	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.94
CatBoost	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.94

Figure 3. Class 0 (Non-Retainers): Performance Before vs. After SMOTE.

Results

Feature Importance

- Family encouragement, coding experience, CGPA, income = strong predictors.
- Rural location = dropout risk.

Policy Simulation

- Rural access: retention 48% → 60%
- Scholarships: 52% → 66%
- Early coding: 72% → 83%
- Parental campaigns: 75% → 82%

Conclusion

- First ML-based predictive study on programmer retention in Bangladesh.
- Combines socio-economic + motivational features.
- Provides actionable insights for workforce development.

Future Directions

- Larger, longitudinal, multi-country datasets.
- Hybrid ML models for imbalanced data.
- Incorporate institutional & labor market dynamics.

References

- Tasmin, M., Ahmed, N., Motahar, T.: Gender disparity in computer science education in Bangladesh: A study of women's participation in computer science. In:2019 IEEE International Conference on Engineering, Technology and Education(TALE). https://doi.org/10.1109/TALE48000.2019.9225981 pp. 1-7 (2019).
- Ahmed, N., Iftekhar, L., Tasmin, M., Urmi, T., Ahmed, S., Motahar, T.: Challenges for women in computing in Bangladesh considering the learning and workplace environment over a period of eight years. SN Social Sciences https://doi.org/10.1007/s43545-022-00529-y 2(10), 227 (2022).
- UNESCO: Cracking the code: Girls' and women's education in STEM. https://unesdoc.unesco.org/ark:/48223/pf0000253479, accessed: 2025-07-14.

Acknowledgement

We would like to express our sincere gratitude to the 188 students from 21 universities in Bangladesh who generously shared their time and perspectives for this study. We are also grateful to the participating universities and faculty advisors for their valuable support during data collection and survey validation. Their contributions were essential to the development of this research framework.

Thank You.