Predicting Alumni Engagement Using Institutional Educational Data: A Machine Learning Approach

Assemgul Baieshova
Master in Computer Science
Paragon International University
Phnom Penh, Cambodia
abaieshova@paragoniu.edu.kh
ORCID: 0009-0006-5816-5934

***Abstract*—Alumni engagement is a strategic asset for higher education institutions because it supports mentorship, institutional reputation, graduate networking, and long-term sustainability. Despite this importance, many alumni management systems still function mainly as descriptive record systems rather than predictive systems that help institutions identify which graduates are more likely to remain actively connected. This study proposes a machine learning-based Alumni Engagement Prediction System (AEPS) that uses institutional educational data and participation history to predict alumni engagement status. Using institutional alumni records from Paragon International Scool Cambodia, the analytical dataset comprised 474 alumni records and 676 columns, of which 655 were event-participation fields used to derive attendance behavior. The study first transforms raw event attendance into a behavioral participation count and then constructs a binary engagement target using a median split at 12 attended activities. To avoid target leakage, the attendance count is used only to define the target variable and is excluded from the predictor set during model training.**

The final modeling pipeline applies missing-value imputation, one-hot encoding, stratified train–test splitting, five-fold stratified cross-validation, and hyperparameter tuning across Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost models. Baseline evaluation shows that Logistic Regression achieved the highest cross-validated ROC–AUC among untuned models at 0.7517, while the tuned XGBoost model delivered the strongest held-out test performance with 0.6737 accuracy, 0.6804 F1-score, and 0.7473 ROC–AUC. SHAP-based interpretability analysis further indicates that faculty, university background, generation, and country of studies are the most influential predictors of engagement, while local-versus-abroad study location contributes comparatively little. The findings show that institutional alumni records contain meaningful predictive signals and that machine learning can support more targeted, data-driven alumni relationship management.

***Index Terms*—alumni engagement, machine learning, educational data mining, predictive analytics, XGBoost, SHAP, institutional data analysis**

I. Introduction

A. Background

Alumni engagement plays a central role in higher education ecosystems because active alumni contribute to mentorship, institutional promotion, recruitment, networking, and long-term sustainability. In many institutions, however, alumni databases are still used primarily for storing records rather than for predictive decision support [1], [8]. As educational data mining and machine learning have matured, institutions have gained the ability to move from descriptive reporting toward predictive analytics that can identify behavioral patterns and support more efficient outreach strategies [1], [5], [13].

Recent work in educational data mining has shown that data-driven approaches can reveal actionable patterns in learner and graduate behavior, especially when institutional data are systematically collected and modeled using supervised learning techniques [1], [6], [13]. In the alumni context, this means institutions can move beyond asking how many alumni participated in past activities and instead estimate which alumni are more likely to remain engaged in the future based on academic and institutional background variables. Such a transition is particularly important for institutions with limited outreach resources because predictive segmentation can help prioritize communication, events, and mentoring opportunities [14], [15].

B. Research Problem

Although universities and schools collect rich alumni-related data, there remains limited empirical work on predicting alumni engagement from structured institutional variables such as faculty, university background, study location, responsibility level, and international mentor connection. Existing alumni systems are often descriptive [8], while predictive modeling in alumni studies has more commonly focused on outcomes such as employment or income rather than institutional engagement itself [4], [12]. This creates an important research gap: institutions need evidence-based methods to estimate engagement likelihood using the information already available in their alumni records.

This study addresses that gap by asking whether alumni engagement can be predicted from institutional educational data using supervised machine learning models. It also examines which models perform best under a reproducible evaluation pipeline, which institutional factors contribute most strongly

to the resulting predictions, and how the resulting outputs can support alumni relationship management in practice.

C. Research Objectives

The objective of this study is to design and evaluate a machine learning-based Alumni Engagement Prediction System that predicts alumni engagement using institutional educational variables and participation-derived engagement labels. More specifically, the study seeks to:

1) construct a defensible engagement target using participation frequency;
2) compare multiple machine learning models under a reproducible evaluation pipeline;
3) identify the most influential predictors of engagement using explainable AI techniques; and
4) demonstrate how predictive outputs can support institutional alumni strategy.

D. Research Questions

This study is guided by the following questions:

1) Can alumni engagement be predicted using institutional educational data?
2) Which machine learning model performs best for this dataset?
3) Which institutional variables contribute most strongly to the model's predictions?
4) What do the results imply for institutional alumni management and engagement strategy?

E. Research Contributions

This study makes five main contributions.

First, it presents a practical machine learning framework for alumni engagement prediction using institutional educational data.

Second, it operationalizes engagement as a behavioral target derived from event attendance counts rather than as an arbitrary subjective label.

Third, it compares four supervised learning algorithms under a common preprocessing and validation pipeline.

Fourth, it applies SHAP-based interpretability to recover institutionally meaningful insights from the final model.

Fifth, it provides a deployable analytical foundation for more targeted alumni outreach and relationship management.

II. Related Work

A. Alumni Tracking Systems

Earlier alumni tracking systems have focused mainly on data management, record updating, and communication support rather than predictive modeling. Jaiswal et al.(2021) [8] described a web-based alumni tracking system intended to improve institutional record management and communication workflows, but the focus remained on system architecture and data organization rather than behavioral prediction. Similarly, the Makerere University alumni tracking project emphasized centralized record access and communication infrastructure without incorporating predictive analytics capabilities [16]. These systems establish the importance of digital alumni data infrastructure, yet they remain largely descriptive.

B. Artificial Intelligence and Alumni Engagement

Recent practitioner-oriented discussions indicate a broader transition from static alumni directories toward more intelligent engagement ecosystems powered by artificial intelligence and analytics. Guerassimov (2025) argued that AI-enhanced alumni systems can support automated updating, intelligent matching, and personalized communication workflows [15]. Sivakumar (2025) similarly highlighted the value of predictive segmentation for more effective alumni campaigns and relationship management [14]. However, these discussions are mostly conceptual and do not provide empirical evidence using structured institutional alumni datasets.

C. Machine Learning in Educational and Alumni Data

Machine learning has already been applied successfully in adjacent alumni and educational prediction contexts. Gomez-Cravioto et al. (2022) [4] used supervised machine learning to predict alumni income outcomes and demonstrated that machine learning models can outperform traditional approaches when relationships are nonlinear and feature interactions matter. In educational data mining more broadly, Baker and Yacef (2009), Romero and Ventura (2013), and Hernández-López et al. (2019) showed that machine learning can model engagement-related behavior, identify hidden patterns, and support better decision-making in institutional settings [1], [6], [13]. These studies justify the use of predictive models in alumni research, but engagement forecasting based specifically on institutional educational attributes remains underexplored.

D. Research Gap

The literature shows three main gaps.

First, many alumni systems remain descriptive instead of predictive.

Second, existing studies often focus on outcomes other than institutional engagement itself.

Third, prior methodological approaches may rely on limited model comparisons or simplistic preprocessing choices.

This study addresses those gaps by using a behavior-based engagement target, a leakage-aware predictor design, and a multi-model comparison pipeline using modern preprocessing and validation practices.

III. Data and Problem Formulation

A. Data Source

The study uses primary alumni data from Paragon International School Cambodia, maintained by institutional coordinators and updated periodically as part of routine alumni administration and event-participation tracking. The raw analytical file contained 474 alumni records and 676 columns before feature engineering, of which 655 were event-participation fields labeled as event variables. Each row represents one alumnus or alumna record in the institutional dataset used for analysis. These records reflect real institutional alumni

TABLE I
OVERVIEW OF VARIABLES USED FOR ALUMNI ENGAGEMENT PREDICTION.

Variable	Description	Type	Used as Predictor?
Local Abroad	Whether the alumnus or alumna studied locally or abroad	Categorical	Yes
Country of Studies	Country where education was completed	Categorical	Yes
University	University attended by the alumnus or alumna	Categorical	Yes
Faculty	Academic faculty or discipline	Categorical	Yes
Responsibility	Institutional or leadership role	Categorical	Yes
Overseas Mentor Connected	Whether the alumnus or alumna is connected to international mentoring	Categorical	Yes
Generation	Alumni cohort or generation	Categorical	Yes
Current Year	Current academic or graduation year representation	Categorical	Yes
Number of Attended Activities	Sum of binary event participation indicators across 655 event columns	Numerical	No, target construction only
Engaged	Binary target derived from attendance count using median split	Binary	Target

information rather than synthetic or externally compiled data, which strengthens the practical relevance of the analysis.

B. Engagement Target Construction

A core methodological decision in this study is the operational definition of engagement. Instead of relying on subjective labels, the study derives engagement from participation behavior by converting raw event columns into binary indicators and summing them to create a new variable called *Number of Attended Activities*. For each alumnus, this attendance count is computed as the row-wise sum of 655 binary event indicators. The reported mean (14.87), median (12), and maximum (81) were then calculated across the 474 alumnus-level attendance counts in the cleaned dataset.

The binary target variable, denoted y_i, is constructed using a median split on the attendance count. Let A_i denote the number of activities attended by the alumnus i, and let $\tilde{A}$ denote the sample median attendance. The engagement target is defined as:

$$y_i = \begin{cases} 1, & \text{if } A_i \geq \tilde{A}, \\ 0, & \text{if } A_i < \tilde{A}. \end{cases}$$

In this study, $\tilde{A} = 12$, 250 cases engaged and 224 less-engaged cases. The median split was used because it provides a behavior-based and relatively balanced classification target from observed institutional participation records, avoiding arbitrary manual thresholds while preserving clear interpretability for binary classification. This construction grounds engagement in observed alumni behavior rather than intuition.

C. Target Leakage Prevention

Because the engagement label is derived directly from the attendance count, including *Number of Attended Activities* as an input feature would allow the model to learn from the same signal that defines the target. To avoid this target leakage, the final predictor set excludes the attendance count entirely and uses only institutional educational and role-based attributes. This design makes the resulting model more valid for practical prediction and more defensible in academic evaluation.

IV. METHODOLOGY

A. Data Preprocessing

The modeling pipeline begins by cleaning column names and detecting all event-related columns programmatically. Event-related columns were identified programmatically from the institutional spreadsheet structure and verified against the event-participation fields in the original dataset before binary cleaning was applied. A total of 655 event columns are identified and converted into binary 0/1 indicators using a custom cleaning function that maps values such as `TRUE`, `YES`, and `Y` to 1, while `FALSE`, `NO`, missing, and unsupported values are mapped to 0. The attendance count is then computed as the row-wise sum of these event indicators.

After target construction, the modeling dataframe is restricted to eight predictors:

- Local Abroad
- Country of Studies
- University
- Faculty
- Responsibility
- Overseas Mentor Connected
- Generation
- Current Year

Missing values are standardized and all selected predictors are treated as categorical in the final model implementation. This reflects the mixed symbolic nature of the institutional variables and ensures that the encoding procedure is consistent across all predictors.

B. Feature Transformation

The final notebook uses a preprocessing pipeline with imputation and one-hot encoding. Numerical variables, if present, are imputed with the median and scaled, while categorical variables are imputed using the most frequent category and then transformed using `OneHotEncoder(handle_unknown="ignore")`. Because the selected predictors are treated as categorical in the final pipeline, the the one-hot branch carries the main transformation load.

This approach is preferable to label encoding because label encoding imposes artificial ordinal relationships among nominal categories, which may distort model behavior. One-hot encoding avoids that issue and is especially appropriate for institution- and identity-related variables such as faculty, university, and country of studies.

C. Train–Test Split and Cross-Validation

The dataset is split into training and test sets using an 80:20 stratified split with a fixed random state of 42, producing 379 training records and 95 test records. An 80:20 split was selected to preserve sufficient data for model training while retaining an independent holdout set for final evaluation, and stratification preserves the class balance between engaged and less-engaged alumni across both subsets. Model development then uses 5-fold stratified cross-validation, again with a fixed random state for reproducibility. This evaluation design follows standard supervised learning practice in applied predictive modeling [2], [7], [11].

D. Classification Models

The study compares four supervised learning algorithms:

- Logistic Regression,
- Random Forest,
- Support Vector Machine (SVM),
- XGBoost.

Logistic Regression serves as a strong and interpretable linear baseline that is widely used in classification research [2], [11]. Random Forest captures nonlinear relationships through ensemble decision trees and is suitable for structured tabular data with potential interaction effects [5], [7]. SVM provides a margin-based alternative for high-dimensional encoded feature spaces and has long been used in supervised classification settings [5]. XGBoost adds a boosted-tree framework capable of learning flexible feature interactions and handling sparse one-hot encoded design matrices efficiently [3].

These four models were selected to provide a balanced comparison between interpretable linear modeling, bagged tree ensembles, margin-based classification, and gradient-boosted trees under a common preprocessing pipeline. Class-balanced settings are used where appropriate, and XGBoost uses a `scale_pos_weight` parameter derived from the training distribution. Each algorithm is wrapped in a common preprocessing-and-model pipeline so that evaluation remains methodologically consistent.

E. Hyperparameter Tuning

After baseline comparison, two stronger candidate models are tuned using `GridSearchCV` with ROC–AUC as the selection metric. Logistic Regression is tuned over regularization strength and solver choice, while XGBoost is tuned over number of estimators, maximum depth, and learning rate. The best XGBoost parameter combination is

- learning_rate = 0.05,
- max_depth = 3,
- n_estimators = 50,

and the best Logistic Regression combination is

- C = 0.1,
- solver = `liblinear`.

These tuned models are then retrained on the training data and evaluated on the held-out test set.

F. Evaluation Metrics

Model performance is evaluated using Accuracy, Precision, Recall, F1-score, and ROC–AUC. For binary classification:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

ROC–AUC is emphasized as the main ranking metric during model selection because it is less sensitive than accuracy to threshold choice and gives a clearer view of discrimination performance.

G. Experimental Environment

The experiments were implemented in Python using Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and XGBoost in a notebook-based environment. Fixed random states were used to support reproducibility, and the full workflow was structured as a consistent preprocessing and evaluation pipeline.

H. Ethical Considerations

The study uses institutional data for predictive analysis and therefore requires careful handling of privacy, fairness, and responsible use. Personally identifiable information was anonymized and excluded before analytical deployment, and predicted engagement scores used only for constructive institutional planning rather than exclusionary decision-making. Explainability is also important: by using SHAP [9], the study improves transparency and enables institutional stakeholders to understand which features drive model predictions.

V. Results

A. Engagement Distribution

The event-processing stage shows that alumni participation varies substantially across records, with attendance counts ranging from 0 to 81 and a median of 12 activities. After median splitting, the binary target is reasonably balanced, with 250 engaged alumni (52.74%) and 224 less-engaged alumni (47.26%). This near-balanced outcome reduces the severity of class imbalance and supports the use of standard classification metrics without extreme skew correction.

As shown in Fig. 1, label 0 denotes less-engaged alumni and label 1 denotes engaged alumni.

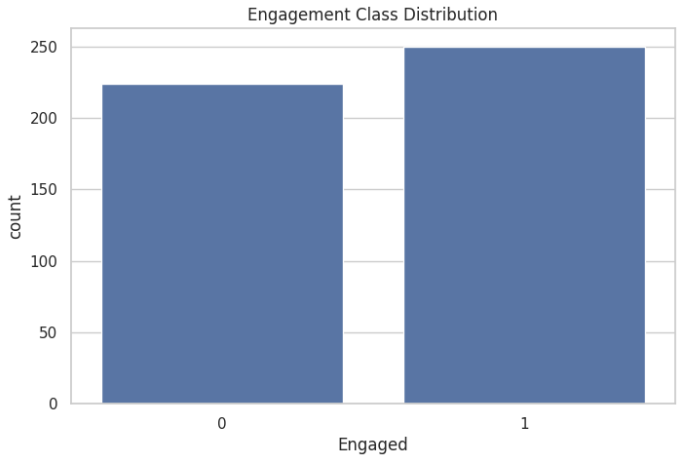


Fig. 1. Distribution of alumni engagement classes, where 0 denotes less-engaged alumni and 1 denotes engaged alumni.

B. Baseline Model Comparison

Cross-validated results reveal that Logistic Regression performs best among the untuned baseline models, achieving a mean ROC–AUC of 0.7517 and a mean F1-score of 0.6863. Random Forest, SVM, and XGBoost all perform moderately but remain below Logistic Regression in baseline ROC–AUC. As presented in Table II and Fig. 2, the baseline comparison suggests that the dataset contains substantial linear predictive structure even before hyperparameter tuning.

TABLE II
CROSS-VALIDATED PERFORMANCE OF BASELINE MODELS.

Model	Acc.Mean	Acc.Std	Prec.Mean	Recall Mean
Logistic Regression	0.6808	0.0475	0.6709	0.7517
Random Forest	0.6361	0.0492	0.6585	0.6408
SVM	0.6545	0.0482	0.6951	0.6109
XGBoost	0.6281	0.0484	0.6489	0.6358

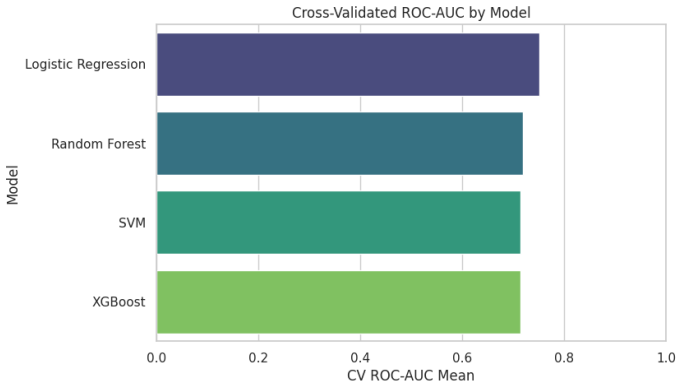


Fig. 2. Cross-validated ROC–AUC by model.

C. Hyperparameter-Tuned Final Models

After tuning, the ranking on the held-out test set changes. The tuned XGBoost model achieves the best test ROC–AUC at 0.7473, with 0.6737 test accuracy and 0.6804 test F1-score, surpassing the tuned Logistic Regression model across these summary metrics. As shown in Table III, tuning changes the ranking on the held-out test set, with XGBoost overtaking Logistic Regression on all reported summary metrics.

TABLE III
TEST-SET PERFORMANCE OF TUNED FINAL MODELS.

Model	Acc	Prec	Rec	F1	ROC–AUC
Tuned XGB	0.6737	0.7021	0.6660	0.6804	0.7473
Tuned Log. Reg.	0.6421	0.6739	0.6209	0.6458	0.7118

D. Confusion Matrix and ROC Curve Analysis

To further assess predictive behavior, a confusion matrix and ROC analysis are conducted on the held-out test set. The tuned XGBoost model achieves a test accuracy of 0.67 and a ROC–AUC of approximately 0.75, outperforming the tuned Logistic Regression model.

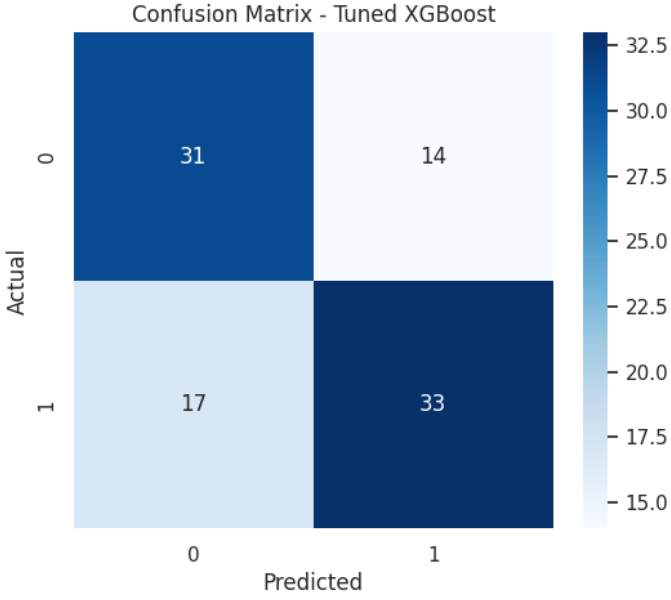


Fig. 3. Confusion matrix for the tuned XGBoost model.

The confusion matrix presented in Fig. 3 shows balanced performance across classes, with precision around 0.65 and recall around 0.69 for non-engaged alumni, and precision of 0.70 and recall of 0.66 for engaged alumni.

Likewise, the ROC comparison in Fig. 4 shows that the tuned XGBoost curve lies consistently above that of tuned Logistic Regression, and an AUC close to 0.75 indicates acceptable discrimination for an educational dataset driven by heterogeneous socio-academic factors.

E. SHAP-Based Model Interpretability

To make the final model more interpretable, SHAP (SHapley Additive exPlanations) is applied using `shap.TreeExplainer` on the tuned XGBoost model [9]. SHAP values are computed on the transformed feature

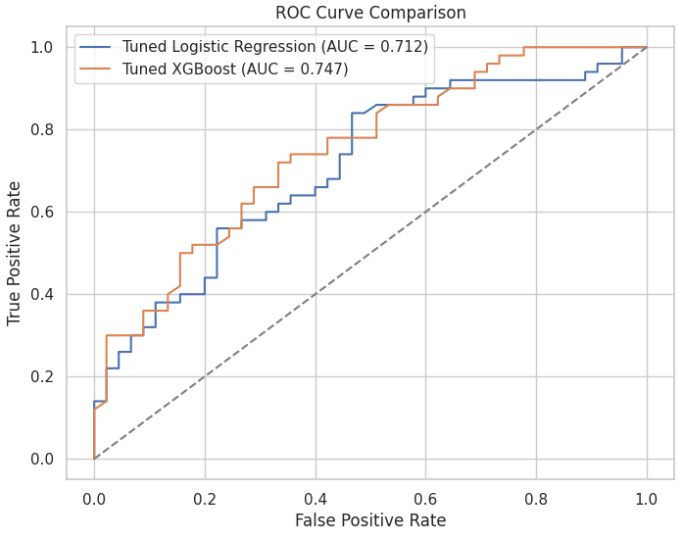


Fig. 4. ROC curve comparison of tuned Logistic Regression and tuned XGBoost.

space and aggregated back to the original feature names so that interpretation remains meaningful for institutional stakeholders.

The grouped SHAP results shown in Table IV and Fig. 5 indicate that *Faculty* is the most influential predictor, followed by *University*, *Generation*, and *Country of Studies*, *Responsibility*, *Currently Year*, and *Overseas Mentor Connected* contribute more modestly, while *Local Abroad* has the smallest aggregate contribution. These findings suggest that alumni engagement in this institutional setting is shaped primarily by academic community membership and cohort structure rather than by simple geographic mobility alone.

TABLE IV
AGGREGATED SHAP IMPORTANCE BY ORIGINAL FEATURE FOR THE TUNED XGBOOST MODEL.

Rank SHAP Importance	Original Feature	Mean Absolute
1	Faculty	5.1442
2	University	1.7103
3	Generation	1.3774
4	Country of Studies	0.7262
5	Responsibility	0.3303
6	Currently Year	0.2712
7	Oversea Mentor Connected	0.1610
8	Local Abroad	0.1492

Unlike Table IV, which aggregates SHAP importance at the original-variable level, Fig. 6 presents the most influential one-hot encoded feature levels within those variables, such as specific faculties, universities, countries of study, and graduation cohorts. The SHAP summary plot therefore complements the grouped analysis by showing which encoded category levels drive prediction at a more granular level.

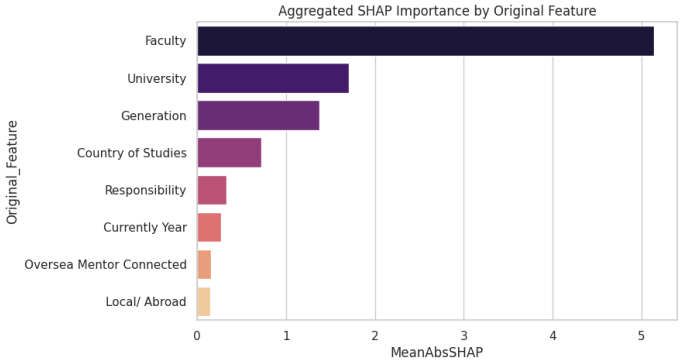


Fig. 5. Aggregated SHAP importance by original feature.

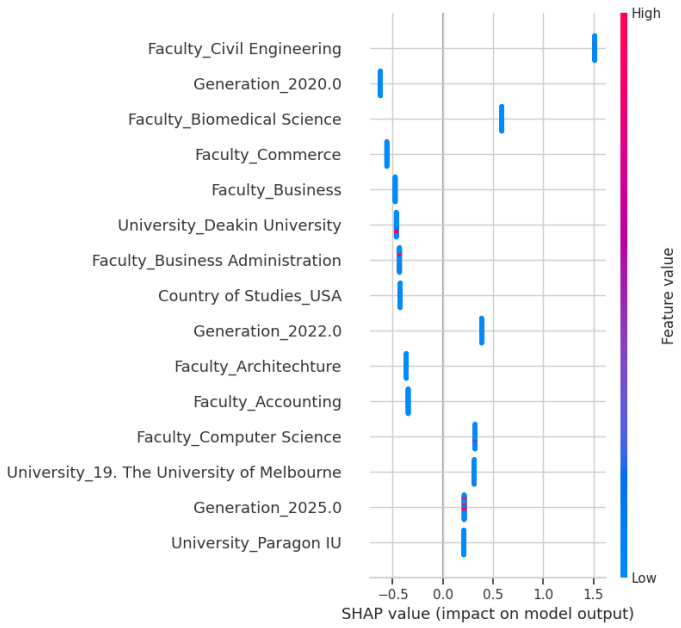


Fig. 6. SHAP summary plot of the top encoded feature levels for the tuned XGBoost model.

VI. DISCUSSION

The results answer the first research question by showing that alumni engagement can be predicted with moderate but meaningful performance using institutional educational data alone. The final tuned XGBoost model does not achieve near-perfect classification, but its ROC–AUC of 0.7473 shows that the model captures real discriminative signals beyond random classification and can therefore support institutional prioritization.

The second research question concerns model choice. An important finding is that the simplest baseline model, Logistic Regression, performs best before tuning, while tuned XGBoost emerges as the final winner on held-out data. This suggests that the predictor space contains both linearly useful signals and nonlinear interactions that become more visible once tree boosting is optimized.

The third research question concerns feature contribu-

tion. The SHAP findings strengthen institutional interpretation by showing that faculty and university background are the strongest predictors, followed by generation and country of studies. These patterns suggest that alumni engagement may be shaped by academic identity, peer networks, and social belonging formed during study. By contrast, the relatively weak contribution of Local Abroad suggests that simply distinguishing between domestic and international study experience is not sufficient to explain engagement behavior on its own.

Finally, the fourth research question concerns institutional meaning. Taken together, the results indicate that predictive modeling can support more strategic alumni outreach, particularly when institutions need to prioritize limited engagement resources across different alumni groups.

VII. Practical Implications

The proposed AEPS can support several institutional use cases.

First, it can help alumni offices identify likely high-engagement alumni for ambassador programs, mentoring initiatives, and targeted event invitations.

Second, it can support re-engagement campaigns by flagging alumni with lower predicted engagement scores who may require different communication strategies.

Third, the model's feature importance outputs can help institutions evaluate whether engagement disparities are associated with particular faculties, cohorts, or educational pathways.

The system should not be treated as a deterministic labeler, but rather as a ranking and segmentation mechanism. Probability-based prediction is especially useful in practice because it enables institutions to define tiers such as high-likelihood ambassadors, moderate-likelihood maintain groups, and lower-likelihood re-engagement candidates.

VIII. Limitations

This study has several limitations. It is based on data from a single institutional context, which limits generalizability. Although the dataset is meaningful for a master-level institutional analytics study, 474 records still represent a moderate sample size for machine learning, especially after high-dimensional one-hot expansion. The target variable is derived from participation counts using a median split, which is behaviorally grounded but still simplifies engagement into a binary outcome.

The model does not include potentially important psychosocial, financial, or communication-response variables that may influence alumni engagement beyond academic background. Because the predictors are primarily institutional and categorical, some underlying causal mechanisms remain unobserved. The model is therefore predictive rather than causal, and its outputs should be interpreted accordingly.

IX. Conclusion

This study developed and evaluated a machine learning-based Alumni Engagement Prediction System using institutional educational data and participation-derived engagement labels. In response to the first research question, the results show that alumni engagement can be predicted from institutional educational variables with moderate but meaningful performance. In response to the second question, Logistic Regression achieved the strongest cross-validated ROC–AUC among baseline models, but after tuning, XGBoost emerged as the final best model with 0.6737 test accuracy, 0.6804 F1-score, and 0.7473 ROC–AUC.

In response to the third question, SHAP-based interpretation showed that faculty, university background, generation, and country of studies are the most influential predictors of alumni engagement in the analyzed institutional context. In response to the fourth question, the findings suggest that institutional alumni records contain enough predictive structure to support more targeted, data-driven alumni relationship management, even when only institutional educational variables are used. The study therefore provides a practical analytical foundation for future institutional deployment, broader data integration, and multi-institutional validation.

References

[1] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.

[2] Boakye, N. F., O'Toole, C. C., Jalali, A., & Hannigan, A. (2025). Comparing logistic regression and machine learning for obesity risk prediction: A systematic review and meta-analysis. *International Journal of Medical Informatics*, 199, 105887.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[4] Gomez-Cravioto, D. A., et al. (2022). Machine learning approaches for alumni outcome prediction.

[5] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[6] Hernández-López, A., González-Fernández, M., & Muñoz-Merino, P. J. (2019). Predicting student engagement using machine learning techniques. *IEEE Transactions on Learning Technologies*, 12(3), 345–356.

[7] Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, 24(2), 1–11.

[8] Jaiswal, S., Kumar, A., & Patel, R. (2021). Alumni tracking system using web technologies. *International Research Journal of Engineering and Technology*, 8(4), 1234–1239.

[9] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*.

[10] McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[12] Pico-Saltos, R., et al. (2022). Alumni programs and career success in higher education.

[13] Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.

[14] K. E. Sivakumar, "AI applications in alumni engagement," white paper, 2025. [Online]. Available: https://cube84.com/blog/ai-for-alumni-engagement-a-practical-guide-for-what-you-can-do-today

[15] A. Guerassimov, "AI-enabled alumni engagement ecosystems," white paper, 2025. [Online]. Available: https://poetsandquantsforexecs.com/....

[16] P. C. Nanseera, F. Musisi, E. Baguma, and M. Jakisa, "Alumni management system," project report, Makerere University, Kampala, Uganda, 2021.