FIGHTING COLLEGE ATTRITION THROUGH TIMELY INTERVENTIONS: A SEQUENTIAL MACHINE LEARNING MODEL

Juan Carlos Apitz, Department of Institutional Research and Analytics, California State University, Long Beach, Long Beach, CA 90840, 562-985-2285, juan.apitz@csulb.edu

Mahmoud Albawaneh, Department of Institutional Research and Analytics, California State University, Long Beach, Long Beach, CA 90840, 562-985-5230, mahmoud.albawaneh@csulb.edu

Santhiveeran, Department of Social Work, California State University Long Beach, Long Beach, CA 90840, 562-985-8339, janaki.s@csulb.edu

Tyler E Nakamura, Department of Institutional Research and Analytics, California State University Long Beach, Long Beach, CA 90840, 562-985-1778, tyler.nakamura@csulb.edu

Dhushy Sathianathan, Department of Academic Planning, California State University, Long Beach, Long Beach, CA 90840, 562-985-2389, Dhushy.Sathianathan@csulb.edu

INTRODUCTION

Attrition is defined as current college students not re-enrolling for a subsequent term. While it can occur for a variety of reasons, it serves as a detriment to the students who attrit. Using supervised Machine Learning techniques, this study aims to create a model and develop analytical tools to identify students at risk of attrition at California State University at Long Beach. The analysis in this paper focuses on attrition in the fourth semester, s_4, based on socio-demographic, pre-entry, and academic performance variables available during the first three semesters. The goal was to create an effective model to estimate the risk of student attrition in a given term.

MACHINE LEARNING METHODS

Four Machine Learning methods, Random Forest, XGBoost, CatBoost, and Light Gradient Boosting Machine (LGBM), were used to assign each student from the testing set a probability of attrition, based on demographic and academic performance features. If that probability was over 50%, attrition was predicted for the student. For each model, precision (percentage of true positives of all predicted positives), recall (percentage of true positives of all actual positives), F1 scores (weighted medium between precision and recall), and ROC-AUC were observed to evaluate how well the model predicted students in attrition. For this study, 10-fold cross validation was used.

Random Forest is a method that uses decision trees independently sampled from an identical distribution. Myriad independent iterations are performed, with features selected at random, and tested to find an error

term and the iteration with the lowest error term is chosen as the optimal model. Unlike other models, each iteration is independent, so features are not selected based on previous iterations, but rather sampled at random from the same distribution.

Unlike Random Forest, XGBoost uses Gradient Boosting, meaning the features are selected based on previous iterations, rather than randomly. When new tree structures are generated, features that minimize the error term are selected.

CatBoost is an open-source algorithm developed by Yandex researchers and engineers, which uses gradient boosting to combine weaker models to create strong models. It uses Ordered Boosting and introduces a new algorithm for dealing with categorical features.

LightGBM is a method used for large data sets. Two techniques, Gradient-Based One-Sided Sampling (GOSS) and Exclusive Feature Bundling (EFB), can be used to cut less relevant features, saving time and processing power.

RESULTS AND CONCLUSIONS

Random Forest yielded a recall of 51.90%, a precision of 19.25%, an F1 Score of 14.04%, and an ROC-AUC of 68.81%. Overall, the model was successful in predicting students who continued, but failed to predict those who attritted. This is due in large part to the smaller sample size from the students in attrition; less data means lower model performance.

XGBoost performed slightly better than Random Forest in all metrics, with a recall of 56.96%, a precision of 20.74%, an F1 Score of 15.20%, and an ROC-AUC of 71.34%. It served as the most balanced model, with the highest F1 score and ROC-AUC. Like Random Forest, XGBoost was more successful in predicting students who continued. Using F1 Score as a medium between precision and recall, XGBoost is objectively the best model. However, depending on an institution's individual goals, a different model may be more desirable.

CatBoost had a notably worse precision (12.46%), but had higher recall (77.85%), with an F1 Score of 10.74% and an ROC-AUC of 71.00%. Compared to Random Forest and XGBoost, it was better at identifying students in attrition, but also incorrectly predicted attrition in more students who continued their enrollment. This is a more expensive model, but ideal for institutions willing to invest more resources into combating attrition, as it would catch the most students at risk.

LightGBM had the highest precision (21.87%), but the lowest recall (43.04%), with an F1 Score of 14.50% and an ROC-AUC of 66.48%. Compared to all other models, it was best at predicting students who continued, at the cost of missing more students in attrition. This model is ideal for more frugal institutions looking to limit resources to students at the highest risk of attrition, saving money at the cost of missing students who attrit.