ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #44608

Student Dropout Prediction in Regional Universities Using Automated Machine Learning

Bin Chen, Purdue University Fort Wayne

Student Dropout Prediction in Regional Universities Using Automated Machine Learning

Bin Chen
Department of Electrical and Computer Engineering
Purdue University Fort Wayne
Fort Wayne, IN, 46805
Email: chenb@pfw.edu

Introduction

According to the latest statistics, over two million people received bachelor's degrees in the United States in 2019, but only 6.3% of them were engineering graduates[1]. NSF's National Survey of College Graduates in 2017 found that only 55% of engineering bachelor's graduates work as engineers, and less than 6% of employees in the engineering workforce hold degrees from non-engineering disciplines [2]. This indicates that engineering jobs require specialized training in engineering, but engineering graduates have more job opportunities in other fields. Other studies have shown that approximately 50-60% of students drop out of their engineering programs in the first two years in the United States [3]. Therefore, only approximately 25% first-year engineering students ultimately graduate and work in engineering-related fields. The high dropout rate causes significant educational resource waste and shortage in the engineering workforce [4,5].

Regional universities and campuses have experienced much higher rates of student attrition in engineering programs. Given the high retention and graduation rates at research one universities and the nationwide engineering dropout rate of approximately 50%, it is clear that the high dropout rate in engineering primarily occurs at regional universities. Reducing the dropout rate at regional universities is therefore the most effective and economical way to increase retention and graduation nationwide [6-8]. However, few reports or publications study the dropout problem in regional universities. Most studies focus on large public universities as research target. The dropout problem in regional universities has not been systematically studied.

Engineering dropout has been studied from educational and social science perspectives for years. Recently, data analysis methods have been used to investigate the problem. Most data analysis methods use academic data (nominal or numerical values) to identify the most influential factors of dropout, or cluster students into different groups. For example, Dorris *et. al.* [6] used statistical testing, cluster analysis and logistic regression to compare or predict engineering student dropouts. Statistical testing for comparisons found that female and full-time students had lower proportions of dropouts. The regression model found that GPA and race were significant predictors in the first and second years respectively. Chen *et. al.* [7] developed a survival analysis framework to predict who will dropout and estimate when they dropout. They compared the performance of their method with basic machine learning algorithms and tested it with Aalen's Additive model and Cox's Proportional Hazard model.

More recently, machine learning has emerged as a powerful tool for achieving breakthroughs in a wide range of challenging problems. This has led to a surge of interest in applying machine learning to identify students at high risk of dropping out. Most classification and regression methods in machine learning, such as neural networks, support vector machines (SVMs), logistic regression, Bayesian networks, clustering methods, decision trees, random forests, and boosting methods, have been applied to dropout estimation and prediction [8]. For example, Aulck *et. al.*[9] used demographics and transcript records at the University of Washington to predict dropout with three machine learning models (regularized logistic regression, k-nearest neighbors, and random forests) and found the strongest predictors of attrition. Niyogisubizo *et. a*l [10] used random forest, extreme gradient boosting (xgboost) [11] and gradient boosting to predict a dropout/not-dropout binary classification problem. Most research in dropout prediction aims to identify the most significant factors of student dropout, and provide statistical information on the likelihood of dropout.

While more and more research started to use the latest progresses in machine learning to predict student academic performances, most of studies are limited to use one classifier and adjust hyperparameters based on researchers' preference and experience. This study uses automated machine learning (autoML) based approach [12] by combining the latest top performers in tabular data analysis: xgboost, lightgbm [13] and catboost [14] for optimal performance in predicting student dropout. These gradient boost based algorithms are currently the most powerful machine learning techniques for regression and classification on structural data. The targeted student group is the first-year engineering students who contribute the most significant dropout rate among the whole college students.

Methods

Heterogeneous data from following main data sources were used to identify students at risk of dropout

1) high school information,
2) demographic information,
3) college and department program information,
4) academic information

The first three categories change less frequently and have stable and long-term effects on the decision of dropout. The data in the last category have higher variations. The combination of all above data includes long-term to short-term influences on dropout decisions in a static, dynamic, and cumulative manner.

Before sending data to machine learning algorithms, the raw data were preprocessed and organized to create input datasets.

- Student home address, residential address and high school address were converted into the distance to the university in miles.
- All information of date and time, such as student high school graduation date, academic program enrollment date, planned and actual graduation date, and dropout date were converted to the student's ages at that time.

The data were preprocessed and then split into training and testing datasets respectively for machine learning. The predicted risk of a student dropping out from an engineering program will be a probability between 100% for graduation and 0% for dropout. If a student dropped out of the engineering program the following semester, the probability of graduation would be 0, otherwise, a new graduation probability will be assigned to the student until that student either graduates or drops out.

The baselines were established by applying each of the algorithms (xgboost, lightgbm and catboost) separately using default settings of these algorithms for classification. The fields, such as major, gender, pursued degree, and race etc.. were designated as categorical data, while other variables such as GPAs were kept as numerical data. The objective was to predict the probability of dropout, which was then classified into either dropout class or enroll (not dropout) class.

While xgboost, lightgbm, or catboost performs well with default settings, achieving optimal performance typically requires hyperparameter tuning. Manual adjustment based on experience is still widely used and quite effective in real practice. Automated machine learning (autoML) offers a more appealing approach for searching larger hyperparameter spaces to achieve optimal performance improvements by using effective searching methods such as Bayesian Optimization. The optimization process typically includes the variables of

- Machine learning models,
- Learning rate
- Number of boosting rounds
- Early stopping conditions
- Max depth of trees
- Number of leaves
- Minimum data in a leaf
- Regularizations

In this study, the top-performing model for each classifier was adjusted manually, and then compared with the best estimator and optimal hyperparameter configuration identified by autoML. The effectiveness of dropout prediction was assessed by standard machine learning metrics in accuracy, precision, recall, and F1-score.

Results and Conclusions

Table 1 to Table 3 show the best results of each individual estimator under manual hyperparameter tuning. Table 4 shows the optimal outcome achieved by autoML. The automated machine learning framework included all three classifiers (xgboost, lightgbm and catboost) and used log-loss function as the loss for optimization. It automatically searched for the best combination of classifier and the settings that achieved the highest accuracy in classifying the data.

The scores highlighted in bold in the following tables represent the highest prediction score of all methods from Table 1 to Table 4.

Table 1: Best classification accuracy achieved by catboost
with manual tuning.

```
-------------------------------------------------
    Accuracy    0.8107

               Precision      Recall    F1-score
     Dropout      0.8309      0.7817      0.8056
      Enroll      0.7927      0.8400      0.8156
-------------------------------------------------
```

Table 2: Best classification accuracy achieved by xgboost
with manual tuning.

```
-------------------------------------------------
    Accuracy    0.7719

               Precision      Recall    F1-score
     Dropout      0.7710      0.7612      0.7661
      Enroll      0.7728      0.7823      0.7775
-------------------------------------------------
```

Table 3: Best classification accuracy achieved by lightgbm
with manual tuning.

```
-------------------------------------------------
    Accuracy    0.7837

               Precision      Recall    F1-score
     Dropout      0.7641      0.7693      0.7667
      Enroll      0.8007      0.7961      0.7984
-------------------------------------------------
```

Table 4: Best classification accuracy achieved by autoML.

```
---------------------------------------------------
    Best estimator            lightgbm
    Accuracy     0.8186

                 Precision       Recall     F1-score
     Dropout        0.8198       0.8181       0.8189
      Enroll        0.8173       0.8190       0.8182
---------------------------------------------------
```

In manually-tuned results, catboost achieved 81.07% accuracy which is about 3% higher than manually-tuned xgboost and lightgbm. On the other hand, automated machine learning took more time to discover the model and optimal hyperparameters for the best classification outcome. It achieved the highest scores across nearly all categories. Although the manually-tuned catboost is the top performer among all manually-tuned methods, autoML identified a set of hyperparameters for lightgbm that yielded the highest accuracy, albeit with a modest improvement. AutoML also streamlines the process by eliminating the need for manual design and hyperparameter selection. This allows researchers to concentrate on experimental design and feature engineering, which typically have greater impact on prediction performance.

In conclusion, the top three structured data classification algorithms yield similar results. This suggests that these algorithms are robust and effective for many tasks without extensive customization. However, further improvements in performance can be achieved through manual fine-tuning or automated machine learning techniques.

After a student is identified as being at high risk of dropping out, universities, communities, and the government can allocate resources to assist them by addressing the significant factors affecting the student through support programs such as wraparound services or scaffold support to enhance retention and graduation rates.

References:

1. NECS (2022) NCES Education Statistics. https://nces.ed.gov/programs/digest/d21/tables/dt21_322.10.asp
2. ASEE (2022) ASEE National Bench Mark Reports. https://ira.asee.org/national-benchmark-reports/workforce2019/

3. Aulck L, Velagapudi N, Blumenstock J, West J (2017) Predicting Student Dropout in Higher Education. http://arxiv.org/abs/1606.06364

4. Joppen L (2020) Shortage of Engineers Starting to Impact Industry. https://stainless-steel-world.net/shortage-of-engineers-starting-to-impact-industry/

5. Okrent A, Burke A (2021) NSF STEM Labor Force. https://ncses.nsf.gov/pubs/nsb20212

6. Dorris D, Swann J, Ivy J (2021) A Data-driven Approach for Understanding and Predicting Engineering Student Dropout. *2021 ASEE Virtual Annual Conference Content Access Proceedings*, :36575. https://doi.org/10.18260/1-2--36575

7. Chen Y, Johri A, Rangwala H (2018) Running out of STEM: a comparative study across STEM majors of college students at-risk of dropping out early. Proceedings of the 8th International Conference on Learning Analytics and Knowledge, :270–279. https://doi.org/10.1145/3170358.3170410

8. Shahiri AM, Husain W, Rashid NA (2015) A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, 72:414–422. https://doi.org/10.1016/j.procs.2015.12.157

9. Alturki S, Alturki N (2021) Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions. *Journal of Information Technology Education: Innovations in Practice*, 20:121–137. https://doi.org/10.28945/4835

10. Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC (2022) Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3:100066. https://doi.org/10.1016/j.caeai.2022.100066

11. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust (2016) 785–794

12. Wang C., Wu Q., Weimer M., Zhu E.. FLAML: A Fast and Lightweight AutoML Library. (2019). https://arxiv.org/abs/1911.04706

13. Ke G., Meng K., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu TY. LightGBM: a highly efficient gradient boosting decision tree. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. (2017) 3149–3157

14. Prokhorenkova L., Gusev G., Vorobev A., Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems (2018) 6639–6649