Student Dropout Prediction

Aarushi Majumder
Department of Computer Science
Stanford University
aaru@stanford.edu

Abstract

Early and accurate prediction of children dropping out from is a serious problem in education, especially in developing countries. Several factors can influence truancy. Additionally, a traditional classification approach to solve the problem might not be exigent enough, given that it would have to be performed as close to the end of school as possible for optimal results. I trained several machine learning algorithms and a neural network in order to come up with the best prediction model of student dropout as soon as possible. The data used was gathered from 460 high schools students in India.

1 Introduction

An important issue, especially in developing countries, is truancy from school. In order to address this, it is critical to understand the causes and recognize the signs. This project will aim to accurately predict the probability of a student dropping out from school. I will measure prediction accuracy and analyze aspects of the students' data so as to recognize the most important factors leading to high dropout rates. Machine learning techniques can effectively facilitate determination of at-risk students and timely planning for interventions. I will implement several classification algorithms as well as train a neural network in order to find the best predictor.

2 Related work

Since machine learning is one of the most effective ways to predict student dropouts, I looked at several academic journals, books and case studies, using several machine learning algorithms. However, most of those algorithms have been developed and tested in developed countries, such as in Europe.[1] Studies have also been conducted in the areas of higher education[2] and online education.[3] Hence, developing countries are facing lack of research on the use of machine learning on addressing this problem. It is especially critical in these countries to identify the students likely to dropout at the high school level, as that is when most of the dropouts occur. Additionally, there are several factors, specifically lack of sanitation, and availability of clean drinking water at school that affect student dropouts exclusively in developing countries. Therefore, this paper presents an overview of machine learning in education with the focus on techniques for student dropout prediction in a developing country (India). The data set used reflects these differences.

[1] Shahidul, SM and Karim, AHMZ. 2015. Factors contributing to school dropout among the girls: a review of literature. European Journal of Research and Reflection in Educational Sciences, 3(2): 25–36.

[2] Aulck, L, Velagapudi, N, Blumenstock, J and West, J. 2016. Predicting Student Dropout in Higher Education. In: ICML Workshop on Data4Good: Machine Learning in within the Open Polytechnic of New Zealand, relying Social Good Applications. New York, NY, USA.

[3] Wang, W, Yu, H and Miao, C. 2017b. Deep Model for Dropout Prediction in MOOCs. Proceedings of the 2nd International Conference on Crowd Science and Engineering – ICCSE'17, 26–32.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

3 Data set and Features

The data was gathered from an Indian government database - data.gov.in. The data set comprises 460 high school students in the year 2015-16. The student data used includes 33 features that relate to demographic and student behavior data, information related to the school's education processes and infrastructure, and data corresponding to the academic processes and socioeconomic information of the students. The target feature is a 0 or 1 indicating dropouts.

The first step was to clean the data obtained, in order to determine that there is no information redundancy and blank fields or data that may affect the prediction process. The data was cleaned according to the following criteria: - Students with 5 or more missing values were removed from the original data set. - Students with a mean grade inferior to 0.2 were removed from the original data set.

Each student was represented in the data set using an 34-dimensional vector consisting of the student's high school data, demographic and situational factors. In order to properly train the classifiers I used Synthetic Minority Over-sampling Technique (SMOTE) to balance the data set.

For dropout classification the data set was split in 60% train and 40% test, training the models using grid search and cross-validation on the training set and evaluating them on the test set. Additionally, the most important factors were extracted and considered as predictors in the input layer of the neural network model.

4 Methods

Part 1: Classification

Four classifiers were trained: Boosted Decision Tree (BDT), Random Forest Classifier (RF), Boosted Decision Tree with SMOTE and Random Forest Classifier with SMOTE using the feature vector of the training set samples. I chose these four models to establish a fair baseline value for the prediction. A brief explanation of each of the algorithms is provided below.

1. A Boosted Decision Tree (BDT) is a structure based on a sequential decision process. Starting from the root, a feature is evaluated and one of the two branches is selected. This procedure is repeated until a final leaf is reached, which normally represents the classification target. [4] Boosting means that each tree is dependent on prior trees. The algorithm learns by fitting the residual of the trees that preceded it. This improves the accuracy. [5]

2. Random Forests Classifiers (RF) are an ensemble learning technique that works by constructing a multitude of Decision Trees and outputs the mode of the classes of the individual trees. This model is trained using Feature Bagging. [6]

3. Boosted Decision Tree with SMOTE. SMOTE is a technique for countering imbalance in a dataset, in the boosting procedure. After each boosting round, we apply the SMOTE algorithm in order to create new synthetic examples from the minority class. SMOTE creates synthetic instances of the minority class by operating in the "feature space" rather than the "data space". By synthetically generating more instances of the minority class, the decision tree is able to broaden its decision regions for the minority class. [7]

4. Random Forest Classifier with SMOTE. Similar to BDTs, SMOTE helps RFs deal with unbalanced datasets.

Part 2: Training a Neural Network
Artificial Neural Networks (ANNs) can model complex non-linear relationships (Mun et al., 2017; Ramachandra and Way, 2018). The ANN is composed of input layer units, hidden layer units, output layer units and connections between these layers. The input layer unit corresponds to each variable of

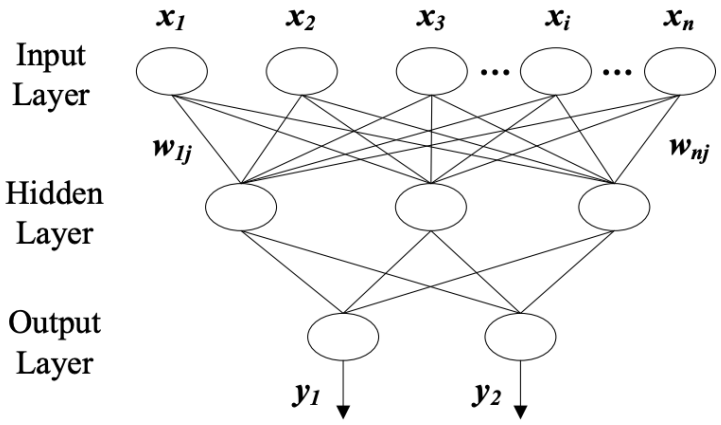
[4](Binary Decision Trees, 2020)

[5](Boosted Decision Tree Regression: Module Reference - Azure Machine Learning, 2020)

[6]Ho TK. Random Decision Forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition. vol. 1 of ICDAR'95. Washington, DC, USA: IEEE Computer Society; 1995. p. 278–282.

[7]Www3.nd.edu. 2020. [online] Available at: <https://www3.nd.edu/ nchawla/papers/ECML03.pdf>.

the input attributions, while the output layer corresponds to the variables of the category attributions. The ANN that was used is based on the multilayer feed-forward error back propagation algorithm.[8]



5 Experiments/Results/Discussion

Evaluation Metrics
The performance of the classifiers is assessed using the standard measures of accuracy, recall, precision and F1.

The classifier metrics are defined as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1 = \frac{2tp}{2tp + fp + fn}$$

where tp is true positive (dropout), tn true negative (not dropout), fp false positive and fn false negative. We consider dropout as the positive class and non-dropout as the negative class. Because we want to minimize false negatives (students who drop out are predicted as students who do not drop out) we will select models with the high recall over those with better precision. We will analyze the trade-off between these metrics using F1.[9]

Results
The plots below (Figure 1) represent the density plots for some of the features. They show that the dropout students are more likely to be problematic in attendance and achievement, and are less likely to participate in the school activities.

Figure 2 below presents the ROC curves for the four binary classifiers used in this study. The AUC of the random forest (RF), and random forest with SMOTE (SMOTE + RF) were the same. The boosted decision tree (BDT) performed better than both the aforementioned models and the boosted decision tree with SMOTE (SMOTE + BDT) performed the best.

Figure 3 below presents the PR curves for the four binary classifiers used in this study. The AUC of the random forest (RF) was the least, followed by random forest with SMOTE (SMOTE + RF), boosted decision tree with SMOTE (SMOTE + BDT) and boosted decision tree (BDT).

[8]TAN, Mingjie; SHAO, Peiji. Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. International Journal of Emerging Technologies in Learning (iJET), [S.l.], v. 10, n. 1, p. pp. 11-17, feb. 2015. ISSN 1863-0383.

[9](Rovira, Puertas and Igual, 2020)

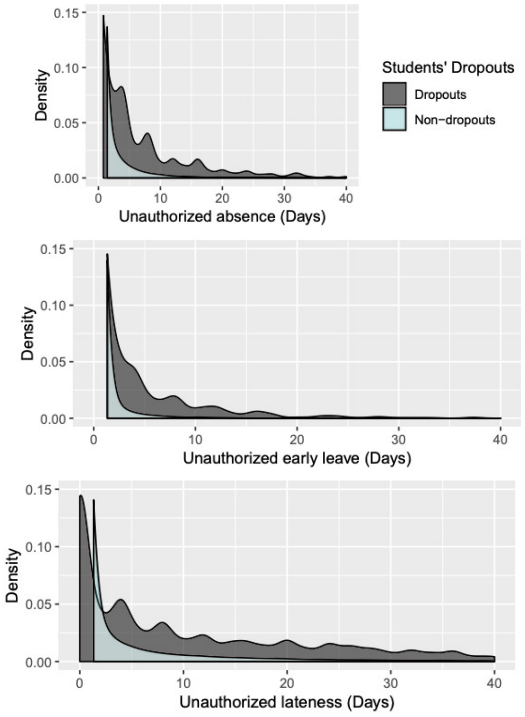


Figure 1: Density plots for some features

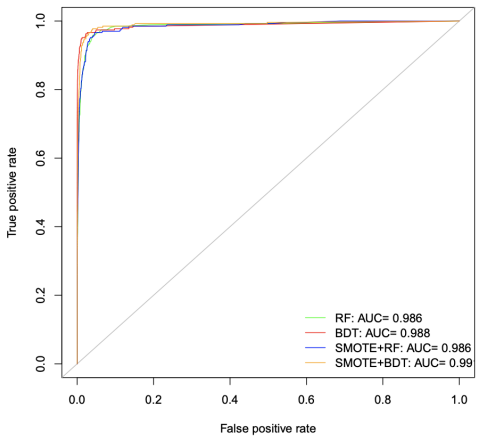


Figure 2: ROC curves for the 4 classifiers

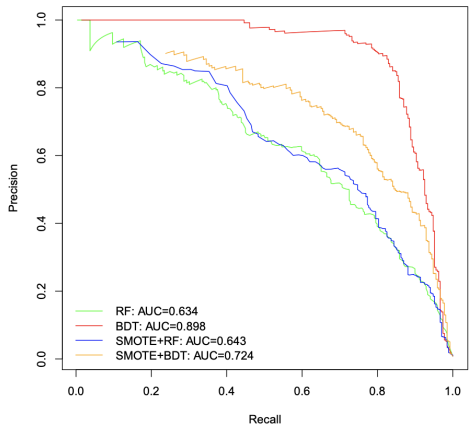


Figure 3: PR curves for the 4 classifiers

Based on both, the ROC and PR curves, the boosted decision tree showed the best performance. The four ROC curves indicate that all four models were excellent in terms of AUCs. However, this is not very informative as the values are quite close together. The PR curves were more useful because their corresponding AUC values were more distinctive. According to the AUC values of the PR curves, the BDT showed the best performance indicating that, among the four tested classifiers, the dropout classification based on BDT was optimal.

The ANN had a true positive rate of 0.994, which is again quite similar to the classifiers. The precision score of the ANN was 0.936, which is a significant improvement in performance over the classifiers.

Overall, there was not too much difference in the performance of the classifiers and the ANN; however, there was significant over-fitting in the ANN which might be attributed to the high number of parameters as well as low number of data points.

6 Conclusion/Future Work

Student drop out prediction is an important and challenging task. In this paper, I attempted to evaluated the effectiveness of several classification techniques as well as a neural network in student dropout prediction. The result was that a the neural network performed the best, followed by the boosted decision tree. Further, the strongest predictors of dropouts were 'Unauthorized absence', 'Unauthorized early leave' and 'Unauthorized lateness'.

Some improvements that can be made to the experiment include a more advanced solution dealing with missing values rather than mapping unknown values to 0. Additionally, most existing studies ignore the fact that the dropout rate is often low in existing data sets; this means that future research should consider developing a student dropout algorithm with consideration of data imbalance problem. Finally, using a bigger data set might be more useful.

An interesting extension could be ranking all students according to risk of dropping out in order to prioritize staging timely interventions.

References

[1]Chollet, F., others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

[2]Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] 4Peng CYJ, Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and Reporting. TheJournal of Educational Research. 2002;96(1):3–14.

[4] Rennie JDM, Shih L, Teevan J, Karger DR. Tackling the Poor Assumptions of Naive Bayes Text Classifiers.In: Proceedings of the Twentieth International Conference on Machine Learning; 2003. p. 616–623.

[5](Binary Decision Trees, 2020)7Ho TK. Random Decision Forests. In: Proceedings of the Third International Conference on DocumentAnalysis and Recognition. vol. 1 of ICDAR'95. Washington, DC, USA: IEEE Computer Society; 1995. p.278–282.

[6] Shahidul, SM and Karim, AHMZ. 2015. Factors contributing to school dropout among the girls: a reviewof literature. European Journal of Research and Reflection in Educational Sciences, 3(2): 25–36.

[7] Aulck, L, Velagapudi, N, Blumenstock, J and West, J. 2016. Predicting Student Dropout in Higher Education. In: ICML Workshop on Data4Good: Machine Learning in within the Open Polytechnic of New Zealand, relying Social Good Applications. New York, NY, USA.

[8] Wang, W, Yu, H and Miao, C. 2017b. Deep Model for Dropout Prediction in MOOCs. Proceedings of the2nd International Conference on Crowd Science and Engineering – ICCSE'17, 26–32.

[9] TAN, Mingjie; SHAO, Peiji. Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. International Journal of Emerging Technologies in Learning (iJET), [S.l.], v. 10, n. 1, p. pp. 11-17, feb. 2015. ISSN 1863-0383.