At-Risk students prediction using Machine Learning

Author: Reginald Ledain
Advisor: Dr. Jeffrey Duffany
Electrical and Computer Engineering and Computer Science Department

Abstract

This article intends to discover how machine learning can be used to predict at-risk students during the school year. Different algorithms were tested within a common framework to compare their accuracy and their interpretability. Using some education expert knowledge, we examined each model relevance in relation to the most important features they used. Attendance, language proficiency and interim test completion were found to be very deterministic in the models prediction capabilities; not a surprise but a validation of the adequacy of the technology for this difficult task.

Introduction

The school districts in the K-12 education domain usually rely on descriptive after-the-fact analytics to take actions on students' performance. Those actions come usually too late for many students and the hope is to implement corrective changes for the next cohort. Beyond obtaining updated reports during the school year, channeling the most urgent information to school leaders and teachers to intervene and help the students at-risk of failing in the end-of-year standardized tests, would be ideal.

Background

While formative and interim assessments are a good way of measuring the students learning process[1], they usually don't offer that 360 view that can predict the actual student performance on their standardized tests. For that reason, we consider that pairing a mid-term, fall or winter, scores for the different strands with all the other indicators will create a more deterministic dataset for our intended goal.

Existing work in this field mostly try to predict broader impact in terms of the district graduation rates for instance. Some others delve into the learning sciences at a level where sophisticated data compilation and tools are necessary to try to understand students learning process. We will investigate using commonly available metrics in school systems like unit testing in math in combination with a variety of well-known indicators like attendance, behavior and demographics, to create a possibly early warning for those students susceptible to fail at the end of the year for a particular subject.

Problem

Machine Learning predictions tend to be black box implementations of the technology. In the field of education, knowing what exactly is going on is crucial to take corrective actions. Beyond being able to predict whether a student will fail or not, we also need to know what factors are influencing such an outcome.

Methodology

Cluster Analysis was used as an initial exploration method to try to identify patterns in the data collected. Various features presented some interesting clusters notably the Limited Language Proficiency (lepstatus), the Disability flag (hasdisability) and the Unit 1 statistics results in the fall. Those patterns will help us validate the machines effectiveness later.

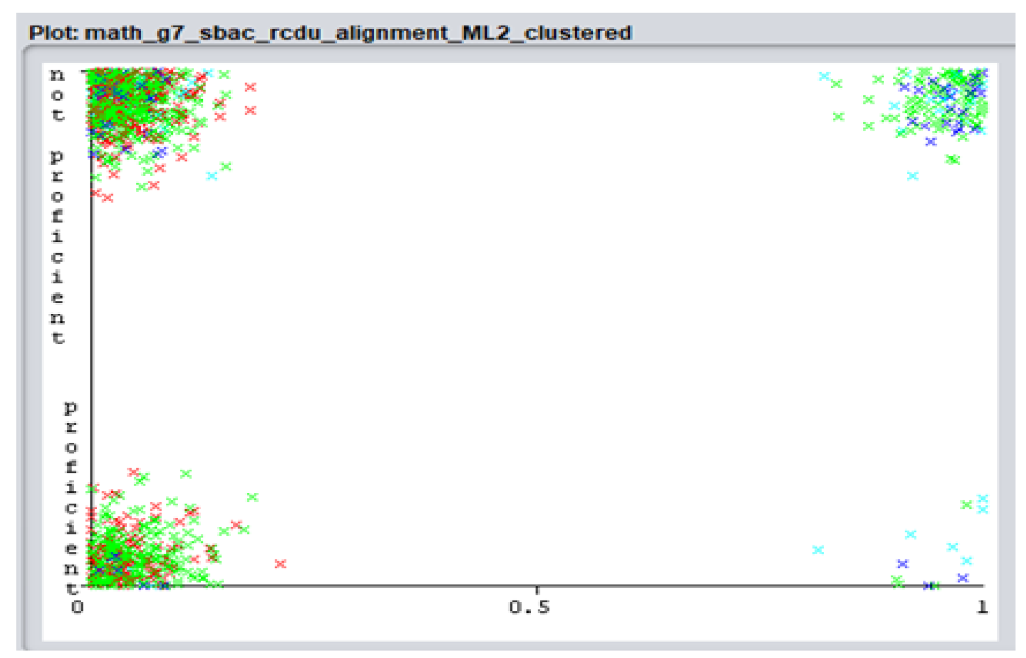

Figure 1
Hasdisability clusters

Various machine learning models can be used for classifying students at-risk. We explore the efficacy and convenience of three popular models:

- Multi-layer perceptron or neural network
- Classification and Regression Tree
- Random Forest

Python has the popular scikit-learn library[4] which implements a large variety of those models and is the main framework for this project. To help identify the optimal configurations of those models we will use some specialized functions to find those parameters that yield the best accuracy starting with a generic model as our baseline. While we compare the three models' performance, we will also identify the most important features as rated by each one. We will also investigate each model interpretability and capacity to provide insight into the subject matter to help identify the root causes.

Results and Discussion

The Decision Tree model performed at the same accuracy as the Multilayer Perceptron but gave us the outstanding capability to trace the rules that determine the prediction as shown in figure 2 and 3

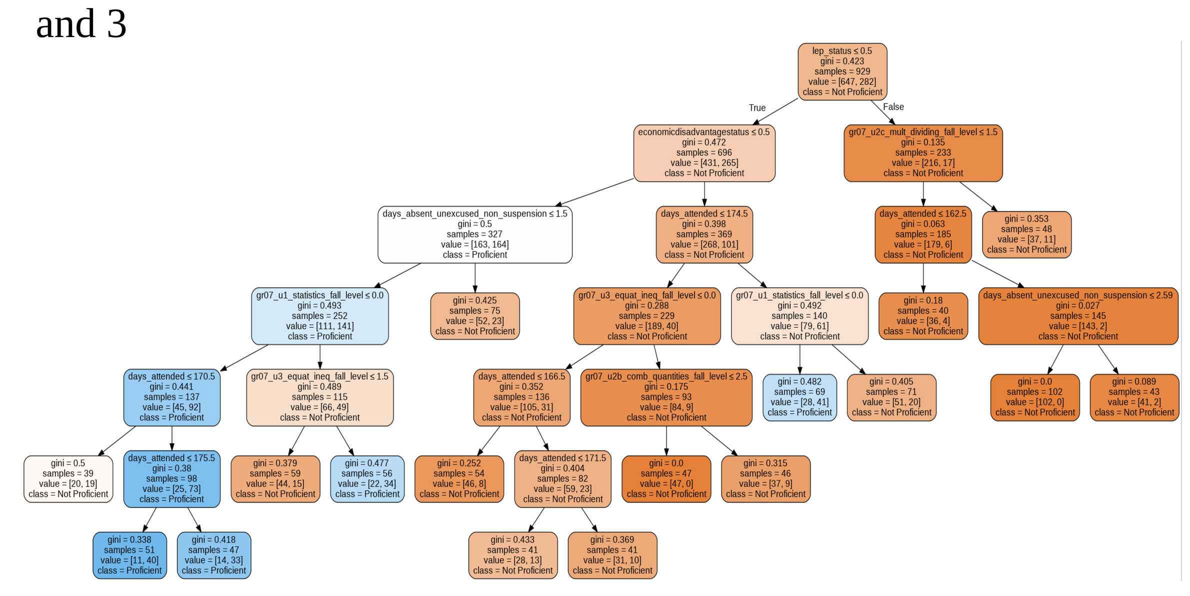

Figure 2
Decision Tree Classifier Visualization

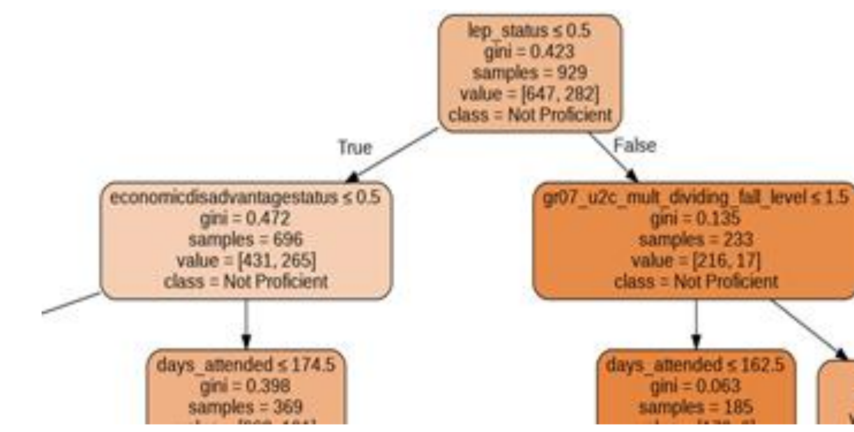

Figure 3
Decision Tree Initial Nodes

However, the Random Forest model had the best prediction as seen in the table below with an 83% of accuracy.

Table 1
Models Performance

#	Model	Accuracy
1	Baseline	0.71
2	Neural Network	0.78
3	Decision Tree	0.78
4	Random Forest	0.83

Despite being a black box in a way as we cannot visually interpret the Random Forest as the visualization of a single tree, we were able to extract its most important features as listed in Table 2

Table 2
Random Forest Classifier Features Importance Ranking

#	feature	weight
1	days_attended	0.15
2	days_absent_excused_non_suspension	0.12
3	lep_status	0.11
4	gr07_u1_statistics_fall_level	0.09
5	days_absent_unexcused_non_suspension	0.08
6	economicdisadvantagestatus	0.07
7	gr07_u3_equat_ineq_fall_level	0.05
8	hasdisability	0.04
9	disciplinary_incidents	0.04
10	gr07_u2a_the_number_line_fall_level	0.04
11	gr07_u2b_comb_quantities_fall_level	0.04
12	gr07_u2c_mult_dividing_fall_level	0.04
13	gr07_u5_unit_rates_etc_fall_level	0.03
14	days_absent_out_of_school_suspension	0.02
15	total_action_duration_days	0.02
16	gr07_u6_probability_fall_level	0.02
17	gr07_u7_geometry_fall_level	0.02
18	gender_male	0.01
19	gender_female	0.01
20	gender_undefined	0
21	migrantstatus	0
22	days_in_attendance_in_school_suspension	0

With the increased accuracy we can attest that some of the evidence-based[5] indicators are ranking higher as expected in the Random Forest compared to the other models. In practice we know attendance is of utmost importance and limited English proficiency of English language learners adversely impact their performance, validating the model behavior in cleverly ranking these related features the highest.

Conclusions

We have seen how the initial exploration of the data through clustering identified marked patterns for some features. Those features importance was also confirmed in their use by the 3 models we tested, where the Random Forest model was the most accurate.

Despite the use of different types of algorithms with varying techniques, we were able to validate some general known assumptions about the data and obtain each model accuracy and its adequacy to generate warnings for students at-risk of not passing their end of year test.

Future Work

Additional data preparation to scale some features like the tests scores may greatly improve the models accuracy. More features, as well as a larger dataset spanning multiple years can also contribute to better train the models.

Future work is planned to leverage the models interpretability into individualized students prediction as a warning system, while providing the teachers and administrators with the specific details about the weights or the rules that are used in the classification. End users enabled with this information will be able to take more specific actions to help the students succeed instead of relying on a black box prediction.

References

[1] M. Perie, S. Marion, and B. Gong, "Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments," Educational Measurement: Issues and Practice, vol. 28, no. 3, pp. 5–13, Feb. 2009.

[2] A. Haghighi, "Numerical Optimization: Understanding L-BFGS," aria42. [Online]. Available: http://aria42.com/blog/2014/12/understanding-lbfgs. [Accessed: 03-Sep-2019].

[3] R. Roy, "ML: Stochastic Gradient Descent (SGD)," GeeksforGeeks, 17-Feb-2019. [Online]. Available: https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/. [Accessed: 03-Sep-2019].

[4] Pedregosa et al., "Scikit-learn: Machine Learning in Python," Python® Machine Learning, pp. 2825–2830, Aug. 2019.

[5] Koon, Sharon, et al. "Using Evidence-Based Decision Trees Instead of Formulas to Identify at-Risk Readers." Institute of Education Sciences (IES) Home Page, a Part of the U.S. Department of Education, Regional Educational Laboratory Program (REL), June 2014, https://ies.ed.gov/ncee/edlabs/projects/project.asp?ProjectID=414
.