Utilizing Exploratory Data Analysis and Machine Learning to Predict and Summarize College Student Academic Performance

Garrett King, Tim Kaufman
Faculty Advisors: Keith Hubbard (Ph.D.) | Dipak Singh (Ph.D.)
Department of Computer Science, Stephen F. Austin University

This research was funded in part through Stephen F. Austin State University's *President's Innovation Fund*.

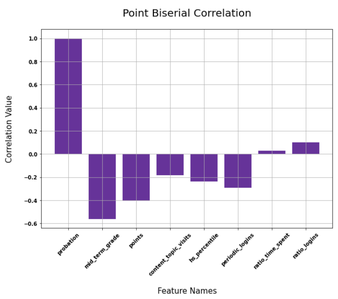
Objective

We aim to use machine learning on university data to create predictions for student success. These predictions are made available through a website aiming to provide additional aid to advisors.

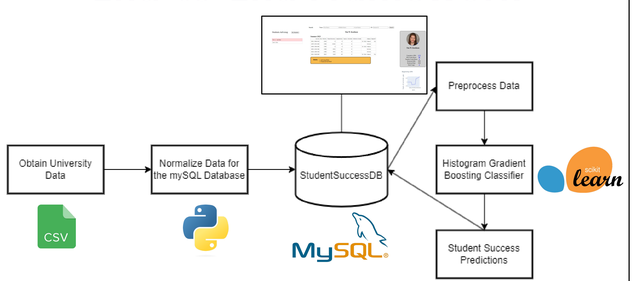
Data Exploration

Before processed through a machine learning model, data must first be explored and cleaned. This includes handling missing values, standardizing and normalizing continuous variables, as well as handling categorical data. Throughout our dataset there were many discrepancies with differing date formats and course codes that had to be rigorously cleaned before model fitting.

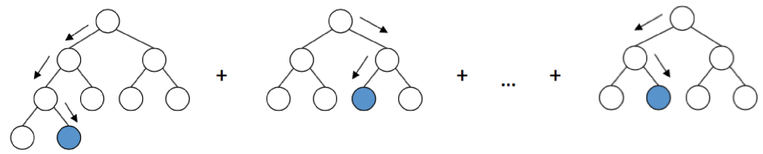
semester_gpa	974
begin_of_semester_gpa	6688
total_course_hours_attempted	0
hs_percentile	15210
points	3557
periodic_logins	5132
mid_term_grade	4883
points_vs_median	2830
hs_sfa_gap	704
total_sfa_hours_bt	0
total_time_spent_in_content	171



End-to-End Framework



Histogram Gradient Boosting (HGB)

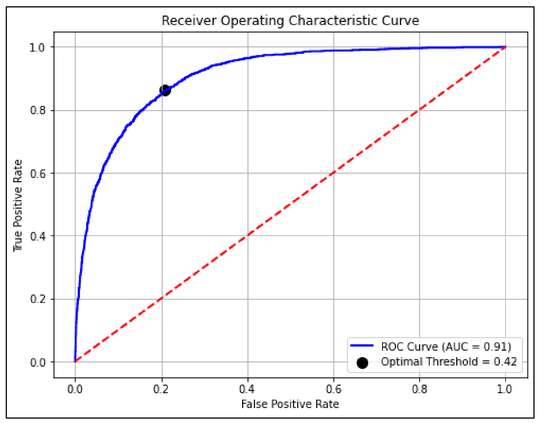


https://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html

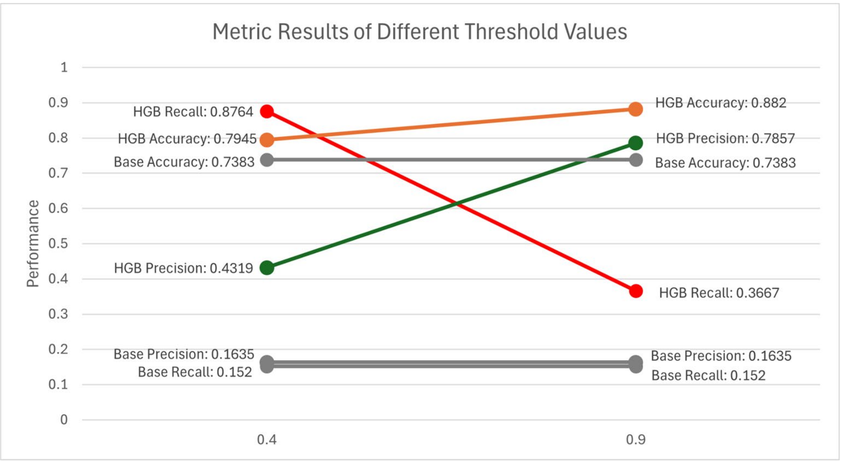
Histogram Gradient Boosting is a special type of traditional gradient boosting. Its most important feature that pertains to our data is its ability to handle missing data. When determining the best splitting points in the decision trees, it will categorize the continuous features into bins, and more importantly, a bin will be reserved for missing values. Also, the model is capable of handling imbalanced data by giving more weight to the minority class .

Our model was tuned using 5-fold cross validation as well as calculating the optimal threshold by using an ROC curve. After proper tuning, it outperformed our Random Forest model as well as Logistic Regression.

Results



Receiver Operating Characteristic (ROC) Curve
The optimal threshold for our classification model was determined by evaluating the area under the ROC curve. We were able to maximize our true positive rate and minimize our false positive rate with a threshold value of 0.42±0.02 giving an AUC of 0.91.



Threshold 0.4
Based on a threshold value of 0.4, our model has a precision of 43.70%, accuracy of 79.80%, and a recall of 87.60%. This threshold value would be optimal for trying to identify as many at-risk students as possible while also attempting to minimize false positives.

Threshold 0.9
With a threshold value of 0.9, the model will have to be much more confident to predict a student may go on probation. This leads to a precision of 79.90%, accuracy of 88.20%, and recall of 36.10%

Improvements

Incorporating behavioral data, or data that can be reasonably substituted, has a high chance of positively affecting our results. Behavioral data has been shown to strongly link with student engagement in other studies.

In the future, maintaining and improving the website and its database becomes a continual, important task. We are looking towards implementing deep learning models to increase the complexity of our modeling. This includes recurrent neural networks (RNN) as well as fully-connected neural networks.