2020 Enterprise Computing Community (ECC) Conference

A two-stage boosted classification framework for improving early detection of freshmen attrition

Eitel J. M. Lauría [1]
Edward Presutti [2]
Eric Stenton [1,2]

[1] School of Computer Science & Mathematics, Marist College
[2] Data Science & Analytics Group, Marist College

{Eitel.Lauria, Edward.Presutti, Eric.Stenton1}@marist.edu

June 7-9, 2020 (Hosted on a virtual platform)

MARIST

Why try to predict attrition ?

- One of the most critical problems in higher education
- Institutional initiative: improve retention
- New research avenue, building on our previous experience
- Very challenging in comparison

Predicting Freshmen attrition

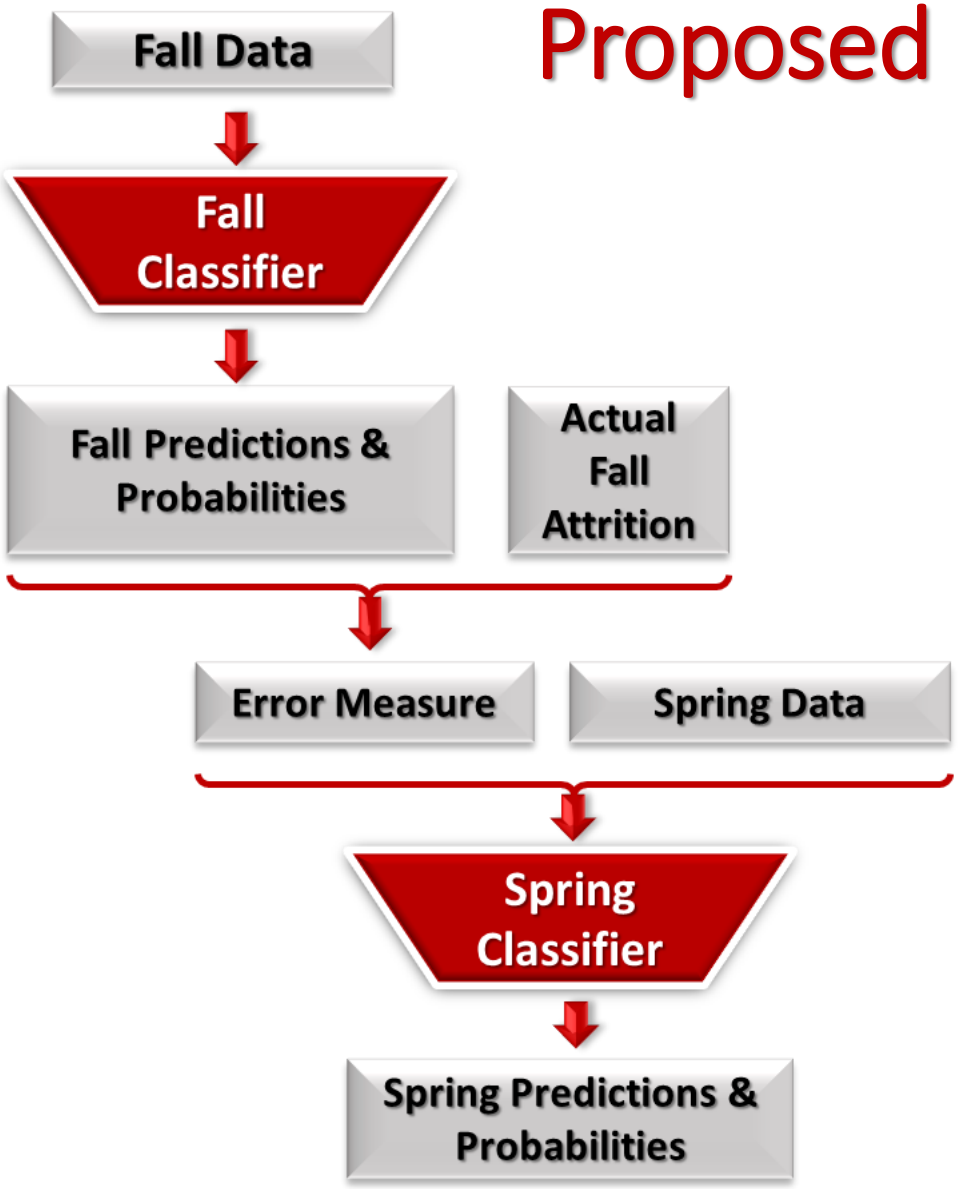
- Freshmen attrition rates are the largest ones in comparison.
- Making predictions of freshmen retention is especially challenging
- Predicting Spring term freshmen attrition
  - The focus of this presentation
  - Fall GPA helps some
  - Are we missing something?

Using the error measure of the Fall semester

- By the end of the Fall semester we know who left the College and who remained
- We also know what prediction errors we made

But wait: shouldn't Fall prediction errors also inform Spring attrition predictions ??

Proposed Classification Framework



The framework builds a Fall semester binary classifier, and subsequently attempts to improve the predictions of Spring attrition by including as a predictor in the Spring classifier the Fall prediction error produced by the Fall classifier.

Design Considerations

- Use cross-validation to tune models
- Use a sequence of years for training, follow by testing
  - ➢no compilation over several years followed by random splitting
- Error measure limited to Fall predictions' false positives
- Error measure computed as:

$$e^{Fall} = \text{abs}(y^{Fall} - \hat{p}^{Fall})$$

- Error measure treated as additional predictor of Spring data, instead of using a traditional boosting approach.

Experimental setup (i)

Goal:

- Investigate the use of a two-stage early detection framework (binary classifier) learnt from data for predicting Spring attrition of Freshman students.

Data:

- Freshmen data from three academic years (2016, 2017, 2018).
- Training dataset (2016+2017) included 2430 records, with 276 attritions distributed in 88 Fall attritions and 188 Spring attritions.
- The test (2018) dataset included 1303 records, with 150 attritions distributed in 50 Fall attritions and 100 Spring attritions
- Each record included the target variables (Attrited_Fall and Attrited_Spring) which were used alternatively for Fall and Spring predictions.

Experimental setup (ii)

Predictor	Description	Data Type
EARLYACTION	Applied for early action	Binary (1/0)
EARLYDECISION	Applied for early decision	Binary (1/0)
MERITSCHOLAMT	Merit Scholar Amount	Numeric
FINAIDRATING	Financial Aid Rating	Numeric
HSTIER	High School Tier	Numeric
FOREIGN	Foreign Student	Binary (1/0)
FAFSA	Applied for Federal Student Aid	Binary (1/0)
APCOURSES	Took AP courses	Binary (1/0)
MALE	Male	Binary (1/0)
MINORITY	Belongs to a minority group	Binary (1/0)
ATHLETE	Is a student athlete	Binary (1/0)
EARLYDEFERRAL	Applied for early deferral	Binary (1/0)
WAITLISTYN	Was waitlisted	Binary (1/0)
COMMUTE	Is a commuter student	Binary (1/0)
HS_GPA	High School GPA	Numeric
DISTANCE_IN_MILES	Distance from home (in miles)	Numeric
APTITUDE_SCORE	Aptitude Score (SAT/ ACT)	Numeric
FIRSTGENERATION	First Generation College Student	Binary (1/0)
SCHOOL	CC, CO, LA, SB, SI, SM	Categorical (6 categories)
RACE	Race (A, B, H, I, M, N, O, P, W)	Categorical (9 categories)
ISPELLRECIPIENT	Is recipient of Pell Grant	Binary (1/0)
ISDEANLIST	Joined Dean's List	Binary (1/0)
ISPROBATION	Is on probation	Binary (1/0)
OCCUPANTS_BLDG	No of occupants in dorm	Numeric
OCCUPANTS_ROOM	No of occupants in dorm's room	Numeric
IS_SINGLE_ROOM	Uses a single room	Binary (1/0)
FS_GPA_NUM	Fall semester GPA	Numeric
ERROR_MEASURE	error measure, used in Spring predictions	Numeric
Target features: Attrited_Spring - Binary (1/0); Attrited_Fall - Binary (1/0)		

Experimental setup (iii)

Methods:

- 16 experiments

  2 different 1st stage classifiers
  - XGB, LOG

  4 different 2nd stage classifiers
  - XGB, RF, LOG, LDA

XGB	XGBTree (Chen and Guestrin, 2016)
LOG	Logistic Regression

XGB	XGBtree (Chen and Guestrin, 2016)
RF	Random Forests (Breiman, 2001)
LOG	Logistic Regression
LDA	Linear Discriminant Analysis

Software tools:

- Python 3.6 + scikit-learn
- SPSS Modeler 18.2
- Bayesian optimization libraries for hyperparameter tuning
  - scikit-optimize (skopt)
  - rfbopt library (https://rbfopt.readthedocs.io)

Hardware platform:

- Intel Xeon server, 2.90GHz, 8 processors, 64GB RAM
- Parallel processing was coded into the system to make use of all n cores during training and tuning

Predictive Performance Results

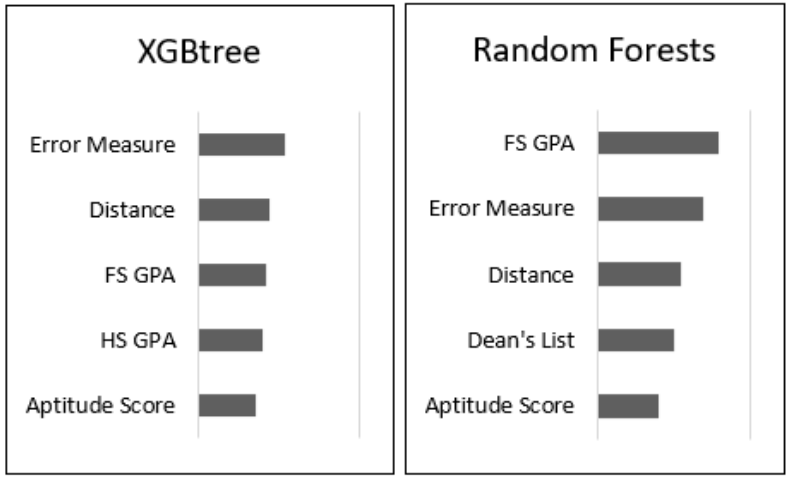
First Stage: Fall		Second Stage: Spring				
Classifier	LOG	Include error measure	XGB	RF	LDA	LOG
ROC AUC	0.66	▪ Accuracy	91.54%	93.06%	81.17%	92.42%
		▪ ROC AUC	0.78	0.802	0.817	0.808
		Exclude error measure	XGB	RF	LDA	LOG
		▪ Accuracy	91.54%	93.22%	70.31%	92.18%
		▪ ROC AUC	0.759	0.796	0.639	0.816

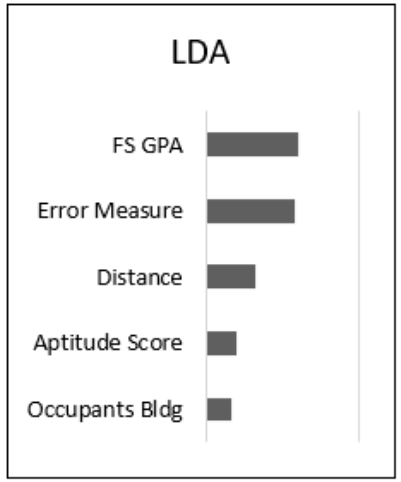
(a) Using Logistic Regression for Fall prediction

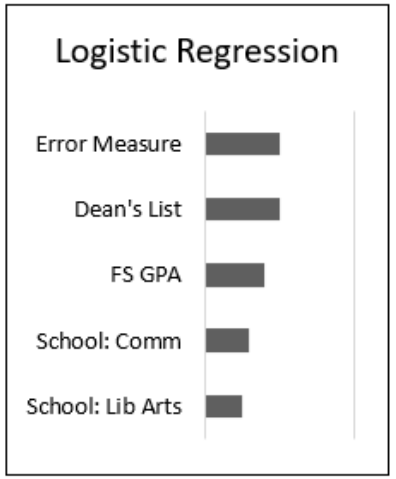
First Stage: Fall		Second Stage: Spring				
Classifier	XGB	Include error measure	XGB	RF	LDA	LOG
ROC AUC	0.64	▪ Accuracy	91.54%	93.30%	73.42%	92.34%
		▪ ROC AUC	0.782	0.792	0.684	0.816
		Exclude error measure	XGB	RF	LDA	LOG
		▪ Accuracy	91.94%	93.22%	70.31%	92.18%
		▪ ROC AUC	0.766	0.792	0.639	0.816

(b) Using XGBtree for Fall prediction

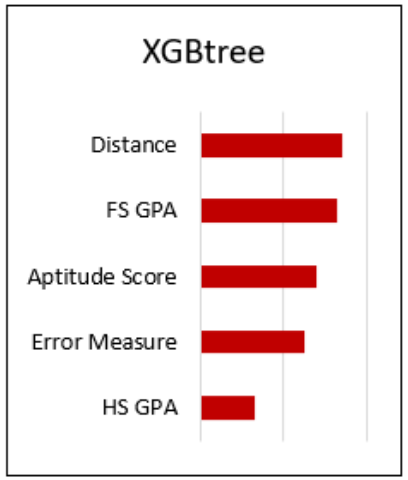
Feature Importance of Second Stage Classifiers

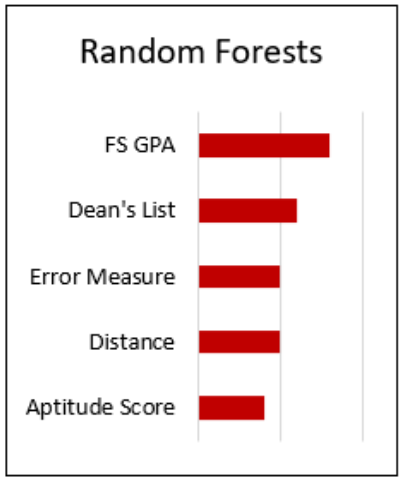


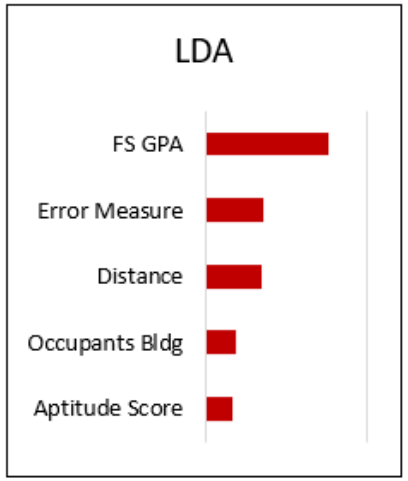


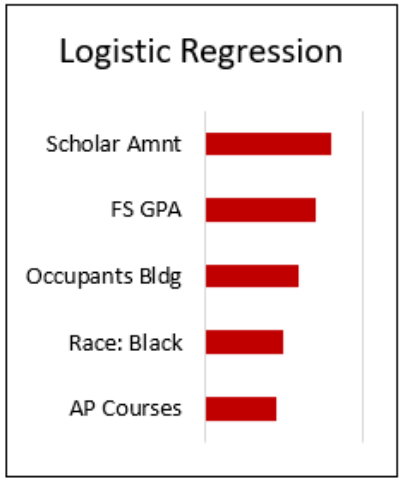


(a) Using Logistic Regression for Fall prediction









(b) Using XGBtree for Fall prediction

Summary

- We explored the use of a two-stage classification framework to improve predictions of freshmen attrition at the beginning of the Spring semester
- The study shows that improvement in predictive performance can be achieved when including the error measure as a predictor in the Spring dataset.
- The study imposed a limited group of classification algorithms.
- Although the experiments included state of the art algorithms, a broader, less discretionary analysis is probably necessary.
- The purpose of the study at this preliminary stage is not to identify an optimal architecture but rather to empirically test the validity and effectiveness of the proposed framework.

Thank you!