Unveiling Risk Factors for College Sucess

Junio de 2017

Contenido

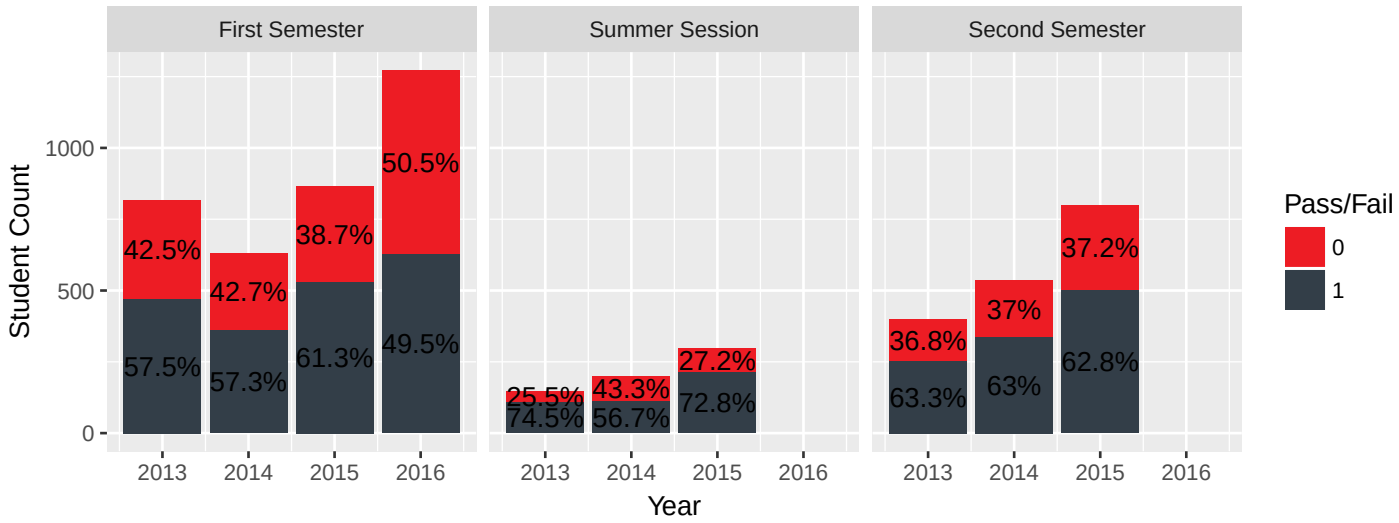
Introducción

- The educational environment is transitioning into the era of Big Data producing very large data sets of diverse nature at an impressive rate.
- Understanding what factors contribute to chances of success in college has unquestionably numerous applications:
  1. Accurate admission's decisions through.
  2. Early detection of course failure
  3. Automatic professor rating.
- We estimate the risk factors of course failure over three different periods of the semester; identity factors' relative importance and influence using a l_1-regularized generalized linear model;
- Data comes from the historical performance of students at Universidad de los Andes..

Introduction

- Approximately 2500 students at Universidad de los Andes enroll in basic courses in science each year. A reasonably high percentage of those students fail to approve these courses

Pass Fail Statistics



Score Sucess Correlation

Cuadro: *Saber 11* Score Sucess Correlation

Period	201310	201320	201410	201420	201510	201520	201610
Total *Saber 11* Score	0.273	0.133	0.274	0.292	0.231	0.178	0.182
Math *Saber 11* Score	0.262	0.165	0.305	0.274	0.229	0.154	0.128

Contenido

Data set

- The first period for training classifiers is immediately before classes begin, that is, as soon as the newly admitted and retakers enroll: we call this period Ex-Ante (EA).
- The second period is the moment right after the first partial exam has taken place and its results are known; we call it Ex-Post P1 (EP1).
- The last period of prediction takes into account all information available up to the results of the second partial exam, this last period is called Ex-Post P2 (EP2).

Data set

Cuadro: Features characteristics

Feature	Type	Description
Final Grade	Real	Range: $[0, 5]$
Exam 1	Real	Range: $[0, 5]$
Exam 2	Real	Range: $[0, 5]$
Age	Real	Range: $[16, 25]$
Math *Saber 11*	Real	Range: $\mathbb{R}$
Total *Saber 11*	Real	Range: $\mathbb{R}$
Math Classification Exam	Real	Range: $\{0, \ldots, 15\}$
Highschool popularity	Real	Range: $\{0, \ldots, 7\}$
Class Size	Real	Range: $\{25, \ldots, 100\}$
Credits Registered	Real	Range: $\{3, \ldots, 25\}$
Taking Physics also	Binary	–
Has taken Pre-calculus Before	Binary	–
Has taken Calculus Before	Binary	–
Socio-economic Status	Categorical	6 Categories
Funding	Categorical	5 Categories
Entrance Period	Categorical	3 Categories
School	Categorical	9 Categories
Place of Birth	Categorical	7 Categories
Gender	Categorical	2 Categories

Contenido

Risks

- The first step is finding the optimal threshold of t in p_t.
- Most students with a final grade around the passing mark have similar distribution of their predictor variables. Sweeping t in the range $[2{,}5, 3{,}5]$, we tested how different the covariate distributions of the classes after thresholding were.
- The resulting optimal was $t = 3{,}3$.

Risks

Cuadro: Risk factor's relative importance.

Order	EA	EP1	EP2
1	Classification Exam	Exam 1	Exam 2
2	Credits Registered	Classification Exam	Exam 1
3	Total Saber 11	Credits Registered	Classification Exam
4	Math Saber 11	Period	Credits Registered
5	Entrance Period	Total Saber 11	Period
6	Taking Physics	Taking Physics	Total Saber 11
7	Class Size	Math Saber 11	Math Saber 11
8	School	Region	Entrance Period
9	Socio-economic Status	School	Taking Physics
10	Region	Socio-economic Status	Region
11	Gender	Age	School
12	Highschool	Gender	Age

Risks

- One important aspect when explaining risk factors is achieving high interpretability.
- In order to enhance the interpretability, all Ex-Ante variables are going to be binned in the range $\{-1, 0, 1\}$ by finding optimal cutoff values in continuous variables and optimal subsets in their categorical counterparts.

Risks

Cuadro: Bining rules for Logistic Model

Variable Name	Variable Abbreviation	Bin −1	Bin 0	Bin 1
Classification Exam	ClEx	$[0, 8]$	-	$(8, 15]$
Credits	Crd	$[0, 10)$	$[10, 20{,}5]$	$(20{,}5, 25]$
Saber	Sb	$(-\infty, 0{,}5]$	-	$(0{,}5, \infty)$
Saber Math	SbM	$(-\infty, 1]$	-	$(1, \infty)$
Intersemetral Period	Int	No	-	Yes
Taking Physics	Phy	No	-	Yes
School	Sch	{ARQ, CPOL, DERE, DISO, EDCO, EDIR,GOBI, MEDI, MICR, MUSI, PSIC, ADMI, BIOL}	All others	{ICIV, IIND}
Graduation Period	Cal	A	-	B
Class Size	Siz	$[0, 40]$	-	$(40, 100]$
Socio-economic status	Se	$\{1, 2, 3, 4, 5\}$	-	$\{6\}$
Sex	Sex	Man	-	Woman
Highschool Popularity	Hgh	$\{1, 2, 3, 4, 5, 6\}$	-	$\{7\}$
Region	Reg	{Amz, Arc, Bbm, Cho, Ptm, SAn, Vic}	All other dptm.	{Boy, NStd, Std}
Age	Age	$[19, \infty)$	-	$[0, 19)$

Sub Risks

The resulting risk factores of our highly interpretable quantized model are shown in the tables.

Cuadro: Exante Risk Factors

		(β)	p-val
β_0	0.419	0.074	***
ClEx	0.561	0.033	***
Crd	0.767	0.104	***
Sb	0.252	0.037	***
SbM	0.230	0.048	***
Int	0.463	0.048	***
Phy	0.198	0.034	***
Siz	0.150	0.031	***
Sch	0.223	0.045	***
Cal	0.054	0.030	.
Se	0.119	0.035	***
Sex	0.087	0.030	**
Hgh	0.088	0.032	**
Reg	0.331	0.102	**
Age	0.096	0.034	**

Sub Risks

tiny

Cuadro: Risk Factors
after first exam

		(β)	p-val
β_0	-4.573	0.173	***
Exam 1	1.425	0.046	***
ClEx	0.342	0.039	***
Crd	0.634	0.117	***
Int	0.449	0.055	***
Sb	0.172	0.038	***
Phy	0.178	0.039	***
SbM	0.113	0.043	**
Siz	0.062	0.035	.
Sch	0.254	0.051	***
Se	0.123	0.040	**

Sub Risks

Cuadro: *

Risk Factors after second exam		(β)	p-val
β_0	-8.561	0.260	***
Exam 2	1.440	0.050	***
Exam 1	1.175	0.050	***
ClEx	0.297	0.042	***
Crd	0.632	0.130	***
Int	0.430	0.060	***
Sb	0.195	0.049	***
SbM	0.146	0.061	*
Phy	0.130	0.043	**
Reg	0.530	0.133	***
Sch	0.217	0.056	***

Sub Risks

- To make sure that our quantized model didn't sacrifice too much predicitive power in favor of interpretability, we trained and contrasted various Machine Learning machines
- We used ROC curve, AUC and accuracy to contrast our linear models against Gradient Boosting Machines and Support Vector Machines.
- The best results were achieved by the gradient boosting of trees, where the AUC reached 0.736 for the Exante Period, 0.838 after the First Exam Period, and 0.907 in the post-second exam period. Likewise, Logistic model performed almost as good: 0.719 Exante, 0.818 Expost first exam, and 0.887 Expost second exam.