Data Mining Techniques and Mathematical Models for the Optimal Scholarship Allocation Problem for a State University

Shuai Wang

Wright State University
Department of Biomedical, Industrial, and Human Factors Engineering

December 6th, 2016

Presentation Outline

Background

- In the United States, the 2012-2013 academic year, there were a total of 20.4 million students in degree-granting institutions.
- More than 80% of them received financial aid.
- Studies have shown that financial aid is one of the most important factors in attracting student and is vital to enrollment management.

Background Cont.

Enrollment management consisted of approaches to help university to meet the established goal such as:

- Attract more high-caliber students.
- Diversify student body.
- Increase retention.
- Improve graduation rate.

WRIGHT STATE UNIVERSITY

Scholarship

- Scholarship is the focused type of financial aid of the study.
- Scholarship is the major marketing tool for targeting students.
- Scholarship helps the university:
  1. Giving more access to families who need help.
  2. Stimulating more students to major in area having labor shortage.
  3. Diversifying the student body.
- Merit-based scholarship.

Motivation

1. **Non-optimal** usage of scholarship budget at university:
   - Over-spending scholarship budget would reduce revenue.
   - Under-spending scholarship would potentially undermine enrollment number and revenue.
2. The tuition income accounts for 48% of the yearly revenue of the university under study.
3. The optimal allocation of scholarship problem has **not** been widely studied in literatures.

Research Questions and Contribution

- Research questions:
  1. How does scholarship affect student's decision?
  2. What is the optimal scholarship for each student under the overall budget?
  3. What is the ideal scholarship budget for school?
- Contribution:
  Utilized various data mining and optimization techniques to solve the above questions.

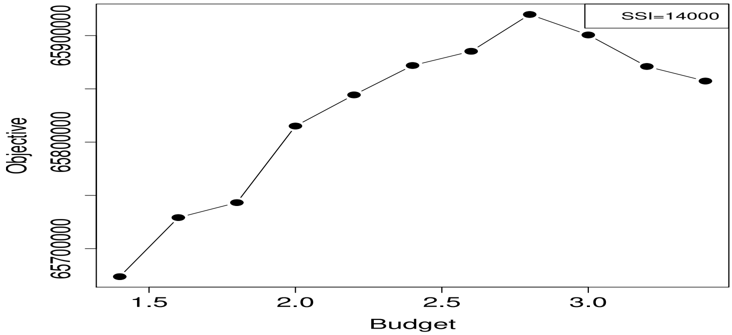
Results example

Enrollment probability of students with various background under different scholarship level.

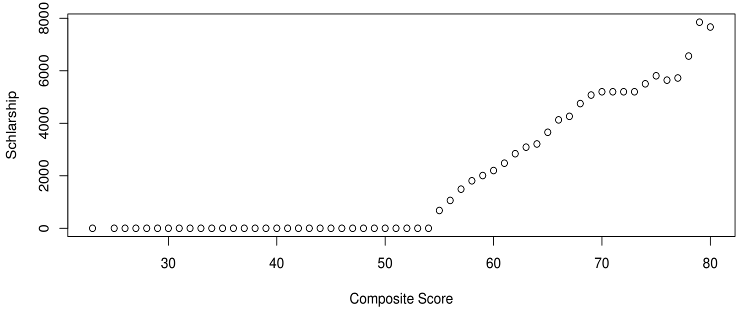
	GPA 2.9, ACT 19									
	Student	0	1000	2000	3000	4000	5000	6000	7000	8000
1	2.9-Tier1-19-White	59.55	64.63	69.39	73.77	77.73	81.24	84.31	86.96	89.22
2	2.9-Tier5-19-White	36.96	40.20	43.53	46.92	50.34	53.75	57.13	60.45	63.67
	GPA 3.3, ACT 25									
3	3.3-Tier1-25-Hispanic	23.80	27.44	31.42	35.69	40.20	44.88	49.65	54.43	59.13
4	3.3-Tier1-25-White	55.60	59.32	62.94	66.42	69.72	72.84	75.75	78.43	80.90
	GPA 3.8, ACT 28									
5	3.8-Tier1-28-White	42.29	46.05	49.85	53.65	57.41	61.08	64.63	68.03	71.25
6	3.8-Tier4-28-White	20.54	22.87	25.37	28.05	30.89	33.89	37.02	40.26	43.60

Results example

Using prediction of enrollment, graduation and number of years as input, and the optimization model yields:



Results example



Average scholarship allocation vs combine score when SSI=14,000

WRIGHT STATE UNIVERSITY

Macro and micro level of enrollment study

Traditional enrollment studies:

- Macro level student demand studies.
- Micro level students college choice models.

Macro level student demand studies

Impact of the increasing tuition on enrollment decisions.

Questions include [Leslie and Brinkman, 1987, 1988, Heller, 1997, Ehrenber, 2004, Crouse, 2015]:

- What happens to the enrollment when universities raise their tuition?
- What is the net impact of higher tuition and reduced enrollment upon institutional finance?
- Who will not attend university because of the increasing tuition?

Macro level student demand studies

- The increasing tuition leads to reduction in enrollment:
  1. Leslie and Brinkman [1987] found that every $100 tuition increase came with a 0.7% drop in enrollment.
  2. Crouse [2015] found a $100 increase in tuition would lead to a decline in enrollment of about 0.88%.

Macro level student demand studies

However:

- Tuition kept increasing in the past decades.
- Enrollment still grows.

Ameliorating effects of financial aid [Leslie and Brinkman, 1987].

Macro level student demand studies

Student demand theory on financial aid:

How does student response to the changes of financial aid.

- Braunstein et al. [1999] found that every $1,000 increase in financial aid, the probability of enrollment increased between 1.1% and 2.5%.

Targeting Effect of Financial Aid on Students with Various Characteristics

Targeting effects of financial aid vs broader effects of tuition.

1. The following group of students are more sensitive to changes of tuition:
   - Low income students [Crouse, 2015].
   - African American and Latino students [Hossler et al., 1989].
2. High-caliber students are less sensitive to the increase of tuition [Heller, 1999].

Micro level student choice model

- Student demand studies, at macro level, investigates the enrollment decisions associated with **groups** of students.
- Student choice model, at micro level, studies the individual student's decision to enrollment at particular university.

Enrollment Prediction at Micro Level

1. Predicting individual student's enrollment decision.
2. Type of variables used in the prediction model: Paulsen [1990], Hossler et al. [1998]:
   - Academic (GPA, ACT/SAT, Percentile)
   - Financial (Income, Pell Grant, etc,.)
   - Demographic (Ethnicity, First language, High school, Home distance to campus, etc,.)

Response to Financial Aid and Optimization at Micro Level

Objective of the mathematic modeling in the literature:

1. Maximize index: SAT [Ehrenberg and Sherman, 1984, Sugrue, 2010].
2. Maximize tuition revenue [Thanh and Haddawy, 2007].

Response to Financial Aid and Optimization at Micro Level

Constraints of the mathematic modeling in the literature:

1. Capacity and faculty-student ratio [Thanh and Haddawy, 2007].
2. Availability of students in expected SAT range, budget, and capacity [Sugrue, 2010].

Response to Financial Aid and Optimization at Micro Level

Issues of these studies:

1. Did not address the allocation of financial aid for students with various socioeconomic characteristics.
2. Did not address how to optimally allocate scholarship at the aggregate level.

Objective of the optimization model

- Objective of the optimization model: Revenue income + SSI income
- Revenue income: Prob E * (Tuition - Scholarship)* NumYears
- SSI income: Prob E * Prob G * SSI

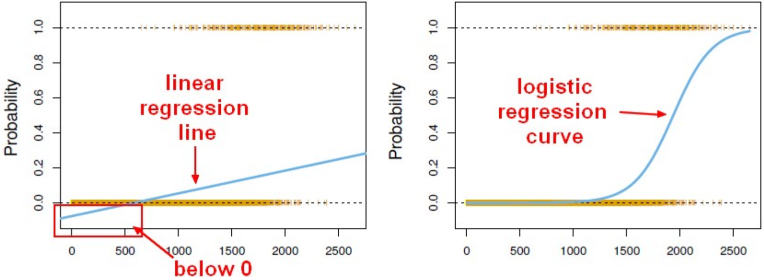
Enrollment and graduation outcome prediction

- The outcome of enrollment and graduation problem is binary (yes or no).
- It is a two-class classification problem.
- Common methods for two-class classification problem: logistic regression, decision tree, svm, etc,.

Logistic Regression

Logistic regression was used as the base line model.

Why linear regression does not work?



Logistic Regression

The goal of logistic regression is to find:

- Best fitting model to describe the relationship between the binary outcome and a set of independent variables.

Logistic Regression: Variables

1. Academic: HS GPA, ACT/SAT, HS Percentile.
2. Financial: Pell Grant, EFC, Out-of-pocket, Scholarship, Unemployment rate.
3. Demographic: High School, Tier, Ethnicity.

Logistic Regression: Enrollment prediction results

	Student	0	1000	2000	3000	4000	5000	6000	7000	8000
	GPA 2.9, ACT 19									
1	2.9-Tier1-19-White	59.55	64.63	69.39	73.77	77.73	81.24	84.31	86.96	89.22
2	2.9-Tier5-19-White	36.96	40.20	43.53	46.92	50.34	53.75	57.13	60.45	63.67
	GPA 3.3, ACT 25									
3	3.3-Tier1-25-Hispanic	23.80	27.44	31.42	35.69	40.20	44.88	49.65	54.43	59.13
4	3.3-Tier1-25-White	55.60	59.32	62.94	66.42	69.72	72.84	75.75	78.43	80.90
	GPA 3.8, ACT 28									
5	3.8-Tier1-28-White	42.29	46.05	49.85	53.65	57.41	61.08	64.63	68.03	71.25
6	3.8-Tier4-28-White	20.54	22.87	25.37	28.05	30.89	33.89	37.02	40.26	43.60

Number of years prediction

- It is a regression problem.
- The following methods are compared to predict the years of stay in school:

	10-Fold Cross Validation		Test Data	
Model	RMSE	MAE	RMSE	MAE
GLM	1.40	1.2	1.53	1.26
SVM (Linear Kernel)	1.44	1.20	1.62	1.32
Decision Tree	1.43	1.24	1.43	1.23
Stochastic Gradient Boosting	1.40	1.19	**1.40**	**1.19**

Why prediction model is not enough?

1. The prediction of enrollment and graduation provide some insights of how students response to the various scholarship.
2. They have not addressed the allocation of limited scholarship budget to students fundamentally.

Objective Model

Sets:

- I: set of applicants, indexed by i and j
- M: different levels of scholarship awards, indexed by m
  $m \in M = \{0, 1000, 2000, \ldots, 8000\}$

Variables:

- x_{im} : binary, whether a scholarship award m is allocated to applicant i or not

Math notation

Parameters:

- p^{e}_{im}: probability of enrollment for applicant i, if given award m
- p^{g}_{im}: probability of graduation for applicant i, if given award m
- N_{im}: expected number of years student i stays at the institution, if given award m
- $d(i,j)$: 1 if applicant i dominates applicant j; 0 otherwise
- B: total budget for financial aid
- A_m: monetary value of award m
- T_i: tuition paid by applicant i
- SSI_i: government compensation for applicant i graduates

Objective

Maximize the total revenue: Tuition revenue + SSI income:

$$\max \quad \sum_{i \in I} \sum_{m \in M} x_{im} \cdot p^{e}_{im} \cdot (T_i - A_m) \cdot N_{im} + \sum_{i \in I} \sum_{m \in M} x_{im} \cdot p^{e}_{im} \cdot p^{g}_{im} \cdot SSI_i$$

Constraints

1. Each student only gets one scholarship:
   $\sum_{m \in M} x_{im} = 1 \quad \forall i \in I$
2. Total budget constraint:
   $\sum_{i \in I} \sum_{m \in M} x_{im} \cdot p^e_{im} \cdot A_m \leq B$
3. Dominance constraint:
   $\sum_{m \in M} x_{im} \cdot A_m \geq \sum_{m \in M} x_{jm} \cdot A_m \quad \forall (i,j) | d(i,j) = 1$

Pair-wise Dominance Constraints

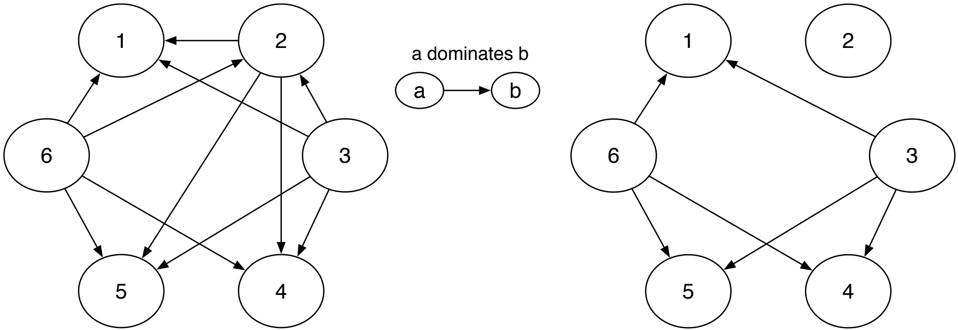
Around 5,500 applicants each year.

For the dominance constraints:

There are $(5,500 \times 5,500)/2$ or more than 15 million constraints.

Applicant	GPA	ACT
1	2.9	18
2	3.7	21
3	3.8	30
4	2.7	21
5	3.3	17
6	3.9	27

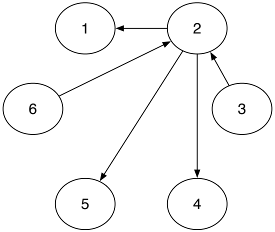
Full and redundant dominance



$$\sum_{m\in M} x_{im}\cdot A_m \geq \sum_{m\in M} x_{jm}\cdot A_m \qquad \forall (i,j)|d(i,j)=1$$

There are total 11 constraints in this case, 6 of them are redundant.

Minimum dominance

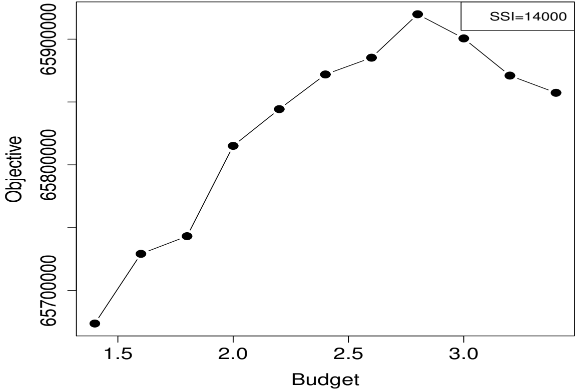


	1	2	3	4	5	6
1	0	0	0	0	0	0
2	1	0	0	1	1	0
3	0	1	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	1	0	0	0	0

Size of the optimization models

Model Components		Original Model	Reduced Model
Variables	Allocation (binary) x_i	57,860	57,860
Constraint	One Award per ID	5,260	5,260
	Dominance	**13,833,800**	**191,497**
	Total Budget	1	1
	Total Number of Constraints	13,839,061	196,758

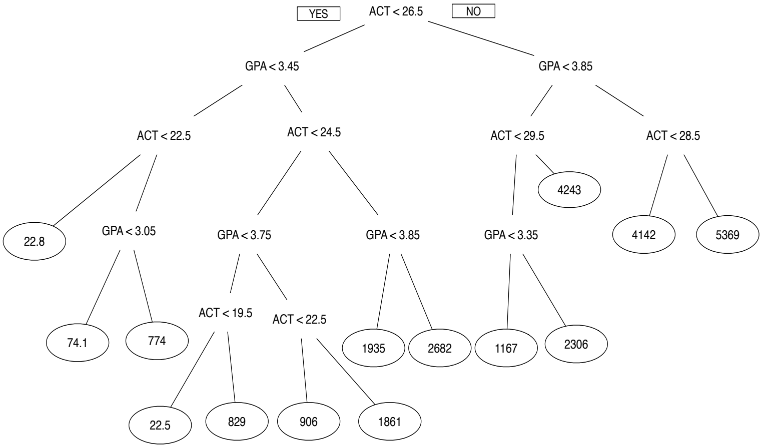
Optimization Results



Policy and Implementation

- First phase solves the prediction problems.
- Second phase solves optimal allocation problem.
- Enrollment administration needs a simple policy to implement.
- Decision tree and piecewise linear regression were used for this task.

Decision tree policy



Optimization Mean Scholarship vs GPA and ACT

GPA/ACT	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Total
1																			0
1.1																			0
1.2		0																	0
1.3	0																		0
1.4		0																	0
1.5		0							0										0
2.5	0	0	0	0	0	0	0	0	0	0	0	0		0					0
2.6	0	0	0	0	0	0	0	0	0	0		1250	1500			1500			25.8
2.7	0	0	0	0	0	0	0	0	0	0	1500		2000						24.2
2.8	0	0	0	0	0	0	0	0	0	300	1500	2000				2500			39.4
2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	818.2	1875.0	2000.0								63.4
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1166.7	1500.0	2000.0	2000.0	2200.0		3200.0			5200.0		216.3
3.1	0.0	0.0	0.0	0.0	0.0	0.0	523.8	1500.0	1727.3	2000.0	2000.0	2500.0	2500.0					7300.0	276.7
3.2	0.0	0.0	0.0	0.0	0.0	781.3	1000.0	1750.0	2000.0	2250.0	2500.0	2500.0	2780.0	3950.0	5200.0				573.7
3.3	0.0	0.0	0.0	0.0	647.1	1000.0	1613.6	2000.0	2315.8	2500.0	2920.0	3200.0	3866.7	4200.0		5200.0	5200.0		962.8
3.4	0.0	0.0	0.0	694.4	1000.0	1550.0	2000.0	2000.0	2500.0	2990.0	3200.0	3200.0	4200.0	4200.0					1185.1
3.5	0.0	0.0	695.7	1000.0	1724.1	2000.0	2000.0	2216.7	2500.0	3200.0	3200.0	3200.0	4200.0	4200.0		5200.0			1696.7
3.6	0.0	545.5	1000.0	1000.0	2000.0	2000.0	2357.1	2500.0	2864.0	3200.0	3200.0	3800.0	4200.0	4200.0	5200.0	5200.0	5200.0		2148.1
3.7	0.0	1000.0	1000.0	1500.0	2000.0	2285.7	2500.0	2945.5	3200.0	3200.0	3700.0	4200.0	4200.0	5200.0	5200.0		5200.0		2532.8
3.8	0.0	1000.0	1000.0	2000.0	2000.0	2500.0	3006.9	3200.0	3200.0	4014.8	4200.0	4200.0	4900.0	5200.0	5200.0	5200.0			3007.5
3.9	666.7	1000.0	1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	3680.0	4200.0	4200.0	4644.4	5200.0	5200.0	5200.0	5200.0	5200.0		3360.4
4		1000.0	1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	4200.0	4808.7	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	6542.9	7300.0	4186.1
4.1		1000.0	1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		3982.1
4.2	1000.0		1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		4175.0
4.3					2000.0	2500.0	3200.0	3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		4607.9
4.4							3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		5125.0
4.5			1000.0		2000.0	2925.0		5200.0	5200.0		5200.0	5200.0		5200.0	5200.0	6233.3		7300.0	4605.3
4.6						4200.0			5200.0	5200.0	5200.0		5200.0		5900.0	7300.0	7300.0		5730.0
4.7				5200.0					5200.0	5200.0	6200.0			8400.0	8400.0	8400.0			6377.8
4.8										5200.0	6200.0		6200.0						5866.7
Grand Total	7.2	92.7	166.2	428.7	787.1	1206	1664.8	2112	2642.1	3407.5	3835.6	3981.3	4561.4	4784.9	5323.2	5258.8	6247.6	7300	1134.7

Piecewise linear regression based policy

Piecewise linear regression using composite score:

Composite Score	# of Applicants	Scholarship Amount
0-53.9	2,897	0
54-68.9	2,103	$309 \times CS - 16,380$
69-76.9	241	$101 \times CS - 2,024$
77-80	19	$711 \times CS - 48,910$

Piecewise linear regression based policy

Simplified version of policy in piecewise regression form.

Composite Score	Scholarship Amount	# of Applicants
0-54.9	0	3,074
55-59.9	1,500	872
60-65.9	2,500	812
66-69.9	3,500	298
70-74.9	4,500	166
75+	6,000	38

Note: 41.6% applicants receive scholarship.

Business impact and Application

- The result of the study has been successfully implemented in the state university and has resulted in millions of financial benefits.
- The research would be applicable to many other institutions and offers a methodology, tools and insights into the solution of financial aid problems.

Business impact and Application

	2013	2014	# Increase	% Increase
Application	6,101	6,068	-43	-0.7%
Admitted	4,541	4,773	232	5.1%
Non-Scholarship	2,166	2,157	-9	-0.4%
Scholarship Award	2,375	2,616	241	**10.1%**
Matriculated	2,001	2,222	221	**11.0%**

Business impact and Application

	2014					2013				
	(O)	(Y)	(Amount)	(total)	Mean	(O)	(Y)	(Amount)	(Total)	Mean
Wright Promise	669	346	1500	$519,000		1072	493	1500	$739,500	
Raider Excellence	787	366	2500	$915,000		673	281	2500	$702,500	
Paul Dunbar	442	190	2500	$475,000		436	190	2500	$475,000	
APS3500	249	88	3500	$308,000						$1,988
APS4500	186	71	4500	$319,500						
APS6000	34	8	6000	$48,000	**$2,417**					
Commended	28	**12**	14000	$168,000						
Val/Sal	184	**119**	8750	$1,041,250		155	**80**	8540	$683,200	
		1200		**$3,793,750**	$ 3,161		**1044**		**$ 2,600,200**	$2,490

Total Merit Number Increase:
(1200-12-119-190)-(1044-80-190)=105

One Year Tuition Increase:
105*(8750-2417)=664,965

Four Year Revenue Increase(50% graduation rate):
(664,965*4 + 105*12000)*0.5 = $1,959,930

Summary

1. A series of models are developed to predict:
   - Enrollment probability
   - Graduation probability
   - Number of years of stay
2. Developed minimum cardinality dominance table to reduce the model size.
3. An optimization model is developed with the objective to maximize the revenue.
4. A regression analysis is developed to translate the optimization results to managerial insights and derive a policy for implementation.

WRIGHT STATE UNIVERSITY

Limitation

1. The prediction of student's enrollment decision is hard.
2. The research is also limited by the availability of data provided by the institution under study.

Future Study

The results from the optimization specify the scholarship awards to each applicant under a specific population and budget.

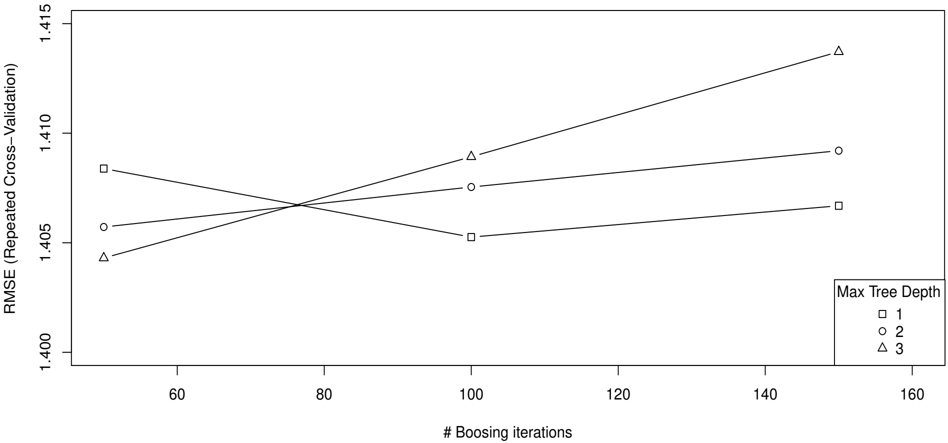
- The actual size and composition of the application pool could be affected by the unemployment rate and is random in nature.
- Uncertainty: Budget, SSI, etc,. Use scenario analysis.

Thank you!

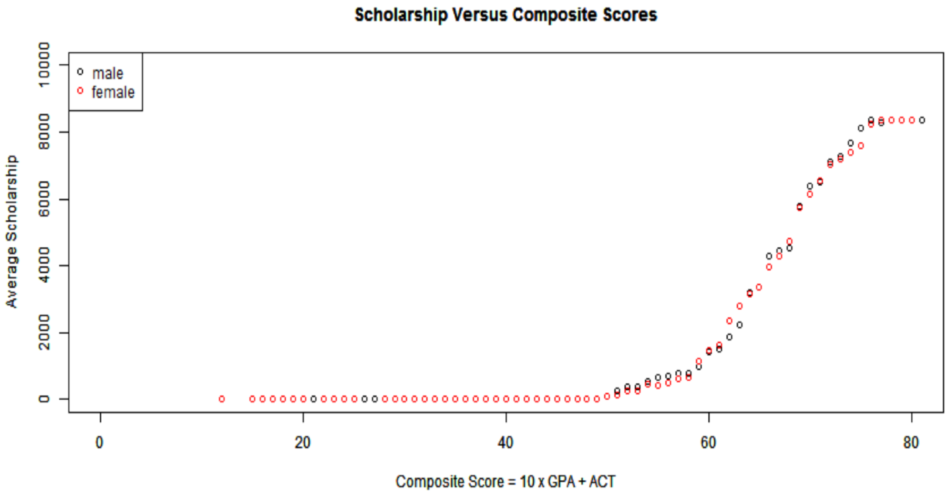
Dissertation Committee:

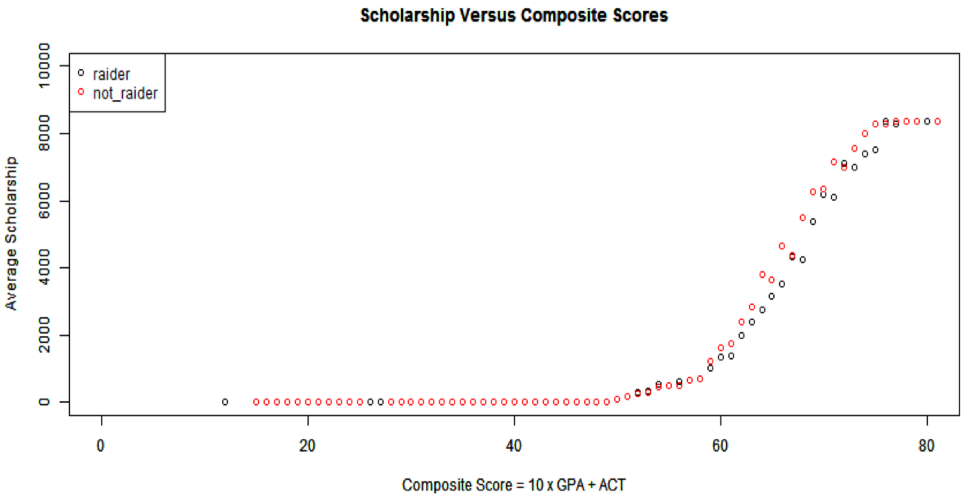
- Dr. Xinhui Zhang
- Dr. Pratik Parikh
- Dr. Caroline Cao
- Dr. Subhashini Ganapathy
- Dr. Nan Kong

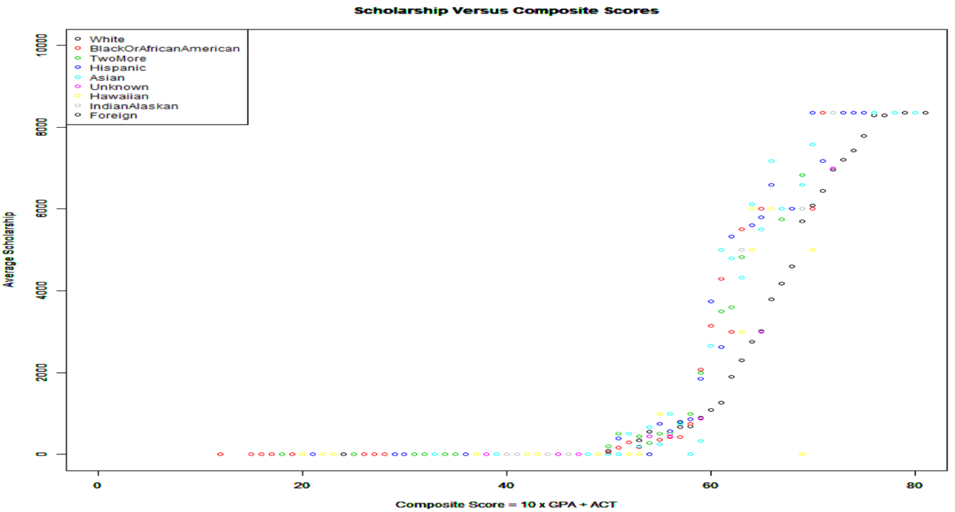
Backup
Stochastic Boosting

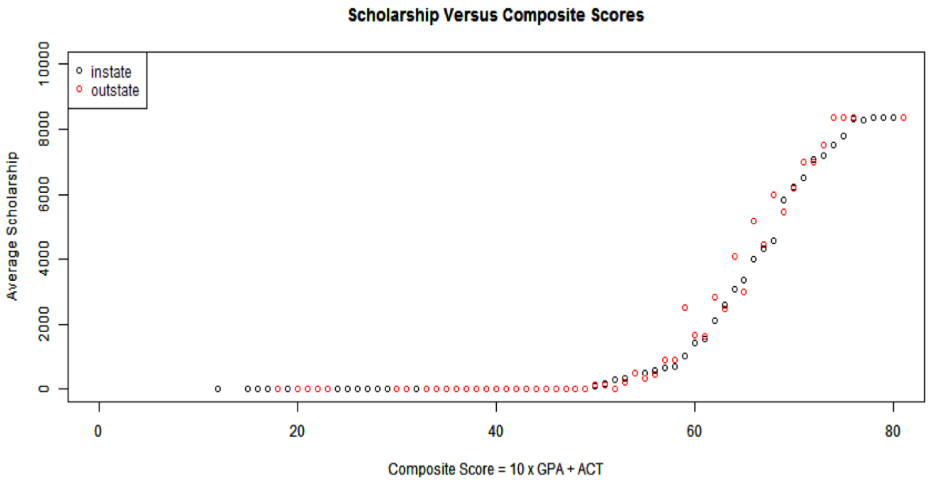


Backup


Scholarship Versus Composite Scores
male
female
Average Scholarship
0
2000
4000
6000
8000
10000
0
20
40
60
80
Composite Score = 10 x GPA + ACT


Scholarship Versus Composite Scores
raider
not_raider
Average Scholarship
0
2000
4000
6000
8000
10000
0
20
40
60
80
Composite Score = 10 x GPA + ACT


Scholarship Versus Composite Scores
White
BlackOrAfricanAmerican
TwoMore
Hispanic
Asian
Unknown
Hawaiian
IndianAlaskan
Foreign
Average Scholarship
0
2000
4000
6000
8000
10000
0
20
40
60
80
Composite Score = 10 x GPA + ACT


Scholarship Versus Composite Scores
instate
outstate
Average Scholarship
0
2000
4000
6000
8000
10000
0
20
40
60
80
Composite Score = 10 x GPA + ACT

What is the SSI formula?

SSI dollars are distributed directly to Ohio's public colleges and universities through a formula primarily driven by enrollments that are classified into models based on levels of instruction and the statewide average costs of each model. For example, it costs more for a campus to offer an advanced engineering course than it costs to offer an introductory history course. Therefore, enrollments in introductory courses are funded at a lower rate than enrollments in more advanced courses.

Source: https://www.ohiohighered.org/board-of-regents/legislative-services/faq

Logistic Regression

Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of the binary outcome:

$$p(1 \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{1}$$

where p is the probability of the binary outcome of interests.

Logistic Regression

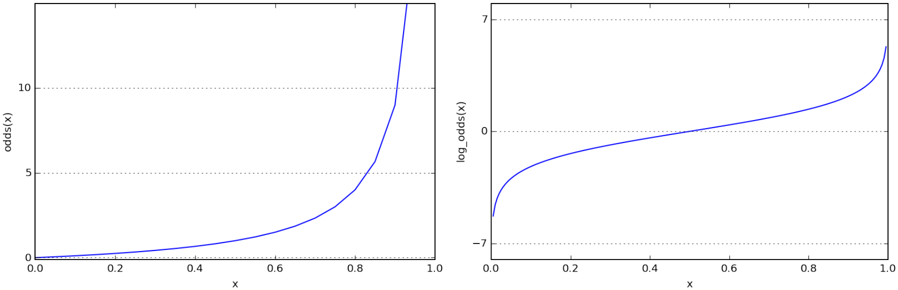
The logit transformation is defined as the logged odds:

$$Odds(p) = \frac{p}{1-p} = \frac{\text{Prob of yes}}{\text{Prob of no}} \tag{2}$$

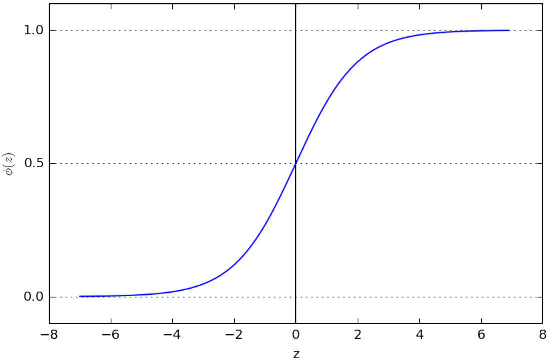
And

$$log\_xodds(p) = \ln \frac{p}{1-p} \tag{3}$$

odds and log odds



Sigmoid



$$log\_odds(P(y=1 \mid x)) = w_o + w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

$$P(y=1 \mid x) = \phi(z) = \frac{1}{1+e^{-z}}$$

A. Braunstein, M. Mcgrath, and D. Pescatrice. Measuring the impact of income and financial aid offers on college enrollment decisions. *Research in Higher Education*, 40(3):247–259, 1999.

J. T. Crouse. Estimating the average tuition elasticity of enrollment for two-year public colleges. *American Journal of Economics*, 5(3):303–314, 2015.

R. G. Ehrenber. Econometric studies of higher education. *Higher Education*, 1-2:19–37, 2004.

R. G. Ehrenberg and D. R. Sherman. Optimal financial aid policies for a selective university. *The Journal of Human Resources*, 19(2):202–230, 1984.

D. E. Heller. Student price response in higher education: An update to leslie and brinkman. *The Journal of Higher Education*, 68(6):624–659, 1997.

D. E. Heller. The effects of tuition and state financial aid on public college enrollment. *The Review of Higher Education*, 23(1):65–89, 1999.

D. Hossler, J. Braxton, and G. Coopersmith. Understanding student college choice. *Higher education: Handbook of theory and research*, 5:231–288, 1989.

D. Hossler, J. Schmit, and N. Vesper. *College Choice: Understanding Student Enrollment Behavior. ASHE-ERIC Higher Education Report No. 6.* Johns Hopkins University Press, 1998.

L. L. Leslie and P. T. Brinkman. Student price response in higher education: The student demand studies. *The Journal of Higher Education*, 58(2): 181–204, 1987.

L. L. Leslie and P. T. Brinkman. *The Economic Value of Higher Education. American Council on Education/Macmillan Series on Higher Education.* Collier Macmillan Publishers, 1988.

M. B. Paulsen. *College Choice: Understanding Student Enrollment Behavior. ASHE-ERIC Higher Education Report No. 6.* Ashe-Eric Higher Education Reports, 1990.

P. K. Sugrue. An optimization model for the allocation of university based merit aid. *Journal of Student Financial Aid*, 40(2), 2010.

L. V. Thanh and P. Haddawy. Deriving financial aid optimization models from admissions data. In *Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual*, pages F2A–7–F2A–12, Oct 2007.