Enhancing the Quality of Predictive Modeling on College Enrollment

Feyzi R. Bagirov

Acknowledgements

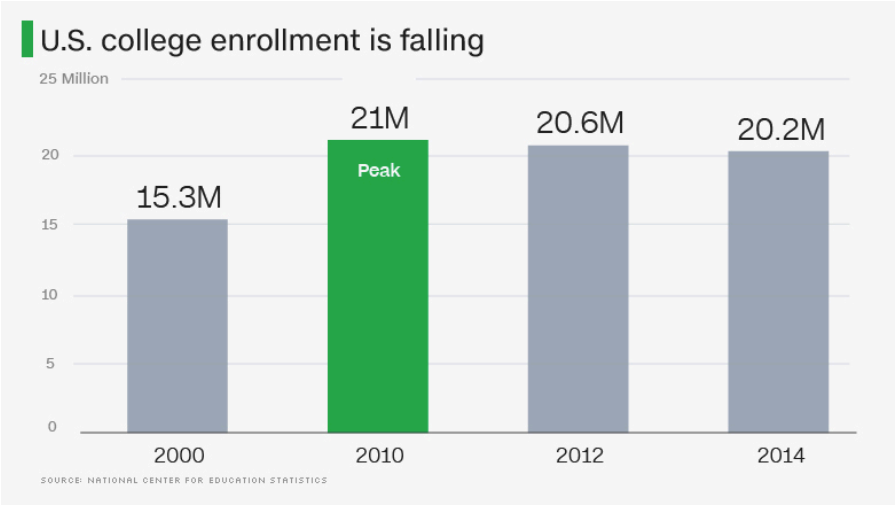
- Yun Xiang, Director of Institutional Research and Assessment at University of New Hampshire

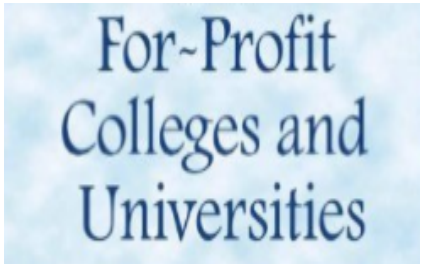
Agenda

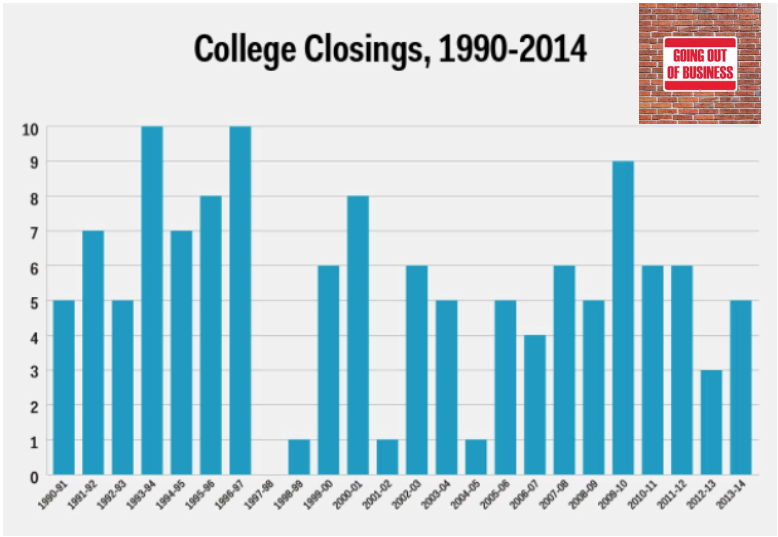
- Enrollment in the US
- Background
- Predictive Analytics Workflow
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment
- Next Steps
- Take Away Messages
- Q&A

- Founding Director of a Bachelor of Science in Data Science program at Becker College
- Faculty of Analytics at Harrisburg University of Science and Technology
- CBO at 529 LLC (Educational infrastructure and analytics)
- Data Science Advisor at Metadata.io (ABM, B2B Demand Generation)

Enrollment in the US







The two types of colleges with the biggest declines in enrollment are: community colleges and for-profit universities. Those schools draw heavily from low-income and minority households.

Source: http://money.cnn.com/2016/05/20/news/economy/college-enrollment-down/

Enrollment in the US

- The 10-year average for college closures is five annually.
- The main struggle for many small colleges is declining enrollment
- Moody's report predicts that inability of small colleges to increase revenue will result in triple the number of closures and double the number of mergers in the coming years.

Exhibit 1

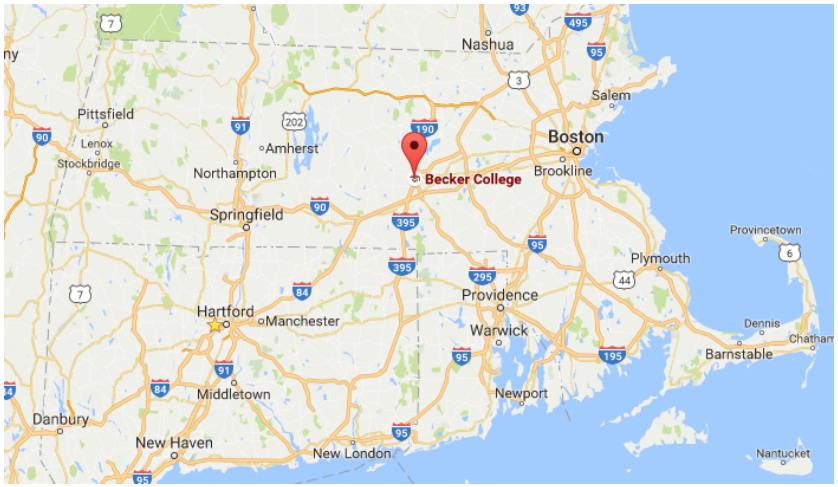
Revenue Stress Will Drive Higher Closure Rates Among Small Colleges Through 2017

Number of college closures by calendar year

Sources: US Department of Education National Center for Education Statistics, Moody's Investors Service for expected 2014 total and projections

Enrollment in the US

Background



BECKER COLLEGE
1887 1784
LEICESTER • WORCESTER

A Private, 4-year College

Enrollment: 2,000 undergraduates

Location: Central Massachusetts

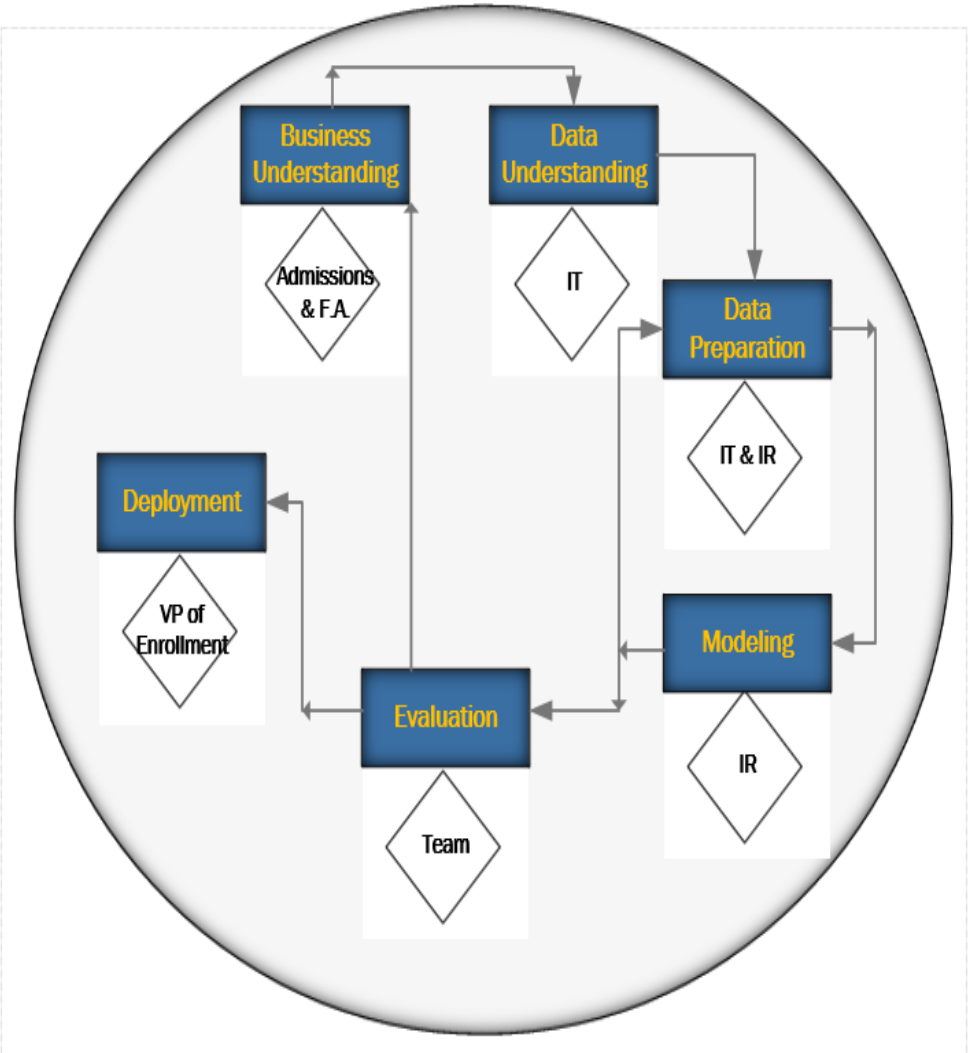
Expectation:

- 1st year—Build the data structure
- 2nd year—Run the predictive models

“Skunk” Team

- Vice President of Enrollment Management
- Chief Information Officer
- Director of Institutional Research
- Dean of Admissions
- Director of Financial Aid
- Director of Data Science
- Director of Enterprise Applications
- Data Engineer

Predictive Analytic Work Flow



The Cross-Industry Standard Process Model of Data Mining (CRISP-DM)

Step One—Business Understanding

Often overlooked!

Initial Requirement

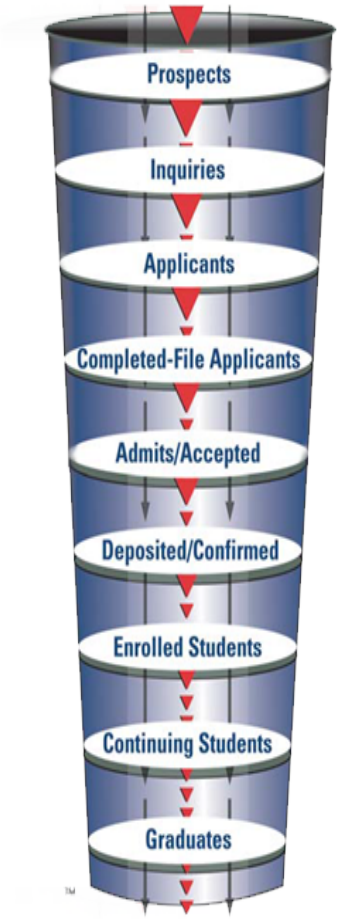
“Utilize Big Data to increase the enrollment

Start with a high-level idea

- Who is my customer?
- What is making my customer complain so much?

Enrollment Management & Admission Funnel

Definition: Enrollment Management is the organizational integration of functions such as academic advising, admissions, financial aid, and orientation into a comprehensive institutional approach designed to enable college and university administrators to exert greater influence over the factors that shape their enrollments (Hossler & Bontrager, 2015, p. 7-8).



Rules of Progression

Response %

Conversion %

Completion %

Acceptance %

Confirmation %

Capture %

Persistence %

Graduation %

Enrollment Management Process

Goals:

- Maintain or increase class size
- Increase ethnic diversity
- Improve academic profile
- Increase net tuition revenue
- Lower the tuition discount rate
- Strengthen weak academic programs
- Maximize the return of strong academic programs
- Support athletic or other specialized programs on campus

Process:

- Prospect/Inquiry Generation
- Applicant Management
- Deposit/Confirmation Management
- Event Management
- Marketing
- Registrar

Hossler, D., & Bontrager, B. (2015). Handbook of strategic enrollment management. San Francisco, CA: Jossey-Bass.

Industry use cases

Marist College Identifies at Risk Students with Predictive Analytics, Increases Degree Completion Rates

The Open Academic Analytics Initiative (OAAI) at Marist College positively impacts higher education with Pentaho big data integration and analytics

September 24, 2014, —

Marist College, a liberal arts college located in Poughkeepsie, New York, has created a portable, open source, predictive analytics model utilizing big data to address the degree completion crisis trending in higher education across America. By utilizing Pentaho's open-source Business Analytics Platform, Marist students, faculty and staff were able to engage in the research, development, and implementation of an early alert system, which is able to identify at-risk students.

The Open Academic Analytics Initiative (OAAI) was founded by Marist College with the support of a $250,000 grant from the EDUCAUSE's Next Generation Learning Challenges program, funded primarily by the Bill and Melinda Gates Foundation. OAAI developed, deployed, and researched the open-source "academic early alert system" that predicts which students in specific courses are not likely to complete the course successfully within two weeks of the start of the course. Pentaho's open-source BI platform was in alignment with Marist's strategic plan to implement and influence new open-source strategies and promote community-source projects, such as OAAI. Further, Pentaho's BI platform is completely open-source and eliminates licensing fees which is a massive cost avoidance in comparison to the available commercial products.

MARIST

At the heart of the OAAI system is a predictive model that "mines" three historical data sets: student aptitude data, learning management system event-log data, and electronic gradebook data. A rich set of plug-ins and filters to perform various data integration and data mining operations are some of the features in the Pentaho BI platform. These features allow an automated sourcing of data from multiple sources that feed the academic predictive model to achieve accurate scoring. Course-specific Academic Alert Reports are generated, and with this actionable

Retention

Industry use cases

For the 2,200 students at community colleges and historically black universities, the Open Academic Analytics Initiative in 2014 — a program developed by Marist College and business analytics firm Pentaho —tracked habits like clicking on online reading materials, whether they posted to online forums, and how long they needed to complete their homework.

By the end of some students' first two weeks in a college course, an analytical model can determine with 75% accuracy rate how well they'll end up doing.

"We know the day before the course starts which students are highly unlikely to succeed," Marie Cini, the provost at UMUC, told the Chronicle of Higher Education.

Industry use cases

"We are entering a new era of data and data responsibility," Mitchell Stevens, an associate professor in Stanford University's Graduate School of Education, told the Chronicle. "Are we acting responsibly as educators? What values are we trying to pursue and preserve?"

- Data collection on students should be considered a joint venture, with all parties — students, parents, instructors, administrators — on the same page about how the information is being used.
- Data-analytics programs need to be transparent, especially when they're making a decision about what will happen to a particular student.
- Colleges using data analytics have to make sure their students have "open futures" — that their programs create educational opportunities, not the other way around.

Data Privacy

Use cases studied prior to design

- A major university in the south of the US
- Started experimenting with analytics in 2003
- Used all available data in modeling
- Leveraged Clearing House data to see where they were losing students; this allowed to have a better idea of true competitors. This led to a completely restructured understanding of their competition.
- Used ACT datasets for profiling

Use cases studied prior to design

- Game-changers and take-away messages:
  - Visits to campus made a difference in predicting which students would enroll
  - Freshmen were required to live on campus and that made students persist better
  - Targeting of the out-of-state students based on geographies where they had a large populations already. By visualizing large populations with maps it showed concentrations by the area

Use cases studied prior to design

- Determining target markets, the Admission's business model changed drastically:
  - Hiring regional recruiters in areas of high prospect concentration
  - Changing target prospects (high-probability prospects require less time, more efforts on 60-80% probability prospects)
  - Changed the targeting process. Customize targeting messages.
  - Changed the types of recruiters they've hired – the ones that better understand data and use that data to bring in the prospects
  - Hired a campaign director to track the success of the campaigns
  - Used in-house callers to spend a lot of time on calling campaigns talking to students
  - Overlaid prospective students with alumni data sets and reached out to alumni to host lunches in areas outside of the home state to engage prospects.

Use cases studied prior to design

- Outcomes:
  - Launched predictive models in 2003. Most information was in Excel. Overtime, acquired a CRM system and leveraged that for consolidation and accrual of applicant and event data.
  - Redundant and irrelevant data in CRM was removed over time (for example, all students wanted a scholarship, so they did not include that in their model)
  - Enrollment increase from 20k to 35k over a targeted period, predominantly due to out-of-state students

Be Realistic, Expect Changes!

Reality

Goal: Increase freshmen yield rate

Question: What admitted students are more likely to deposit?

Step Two—Data Understanding

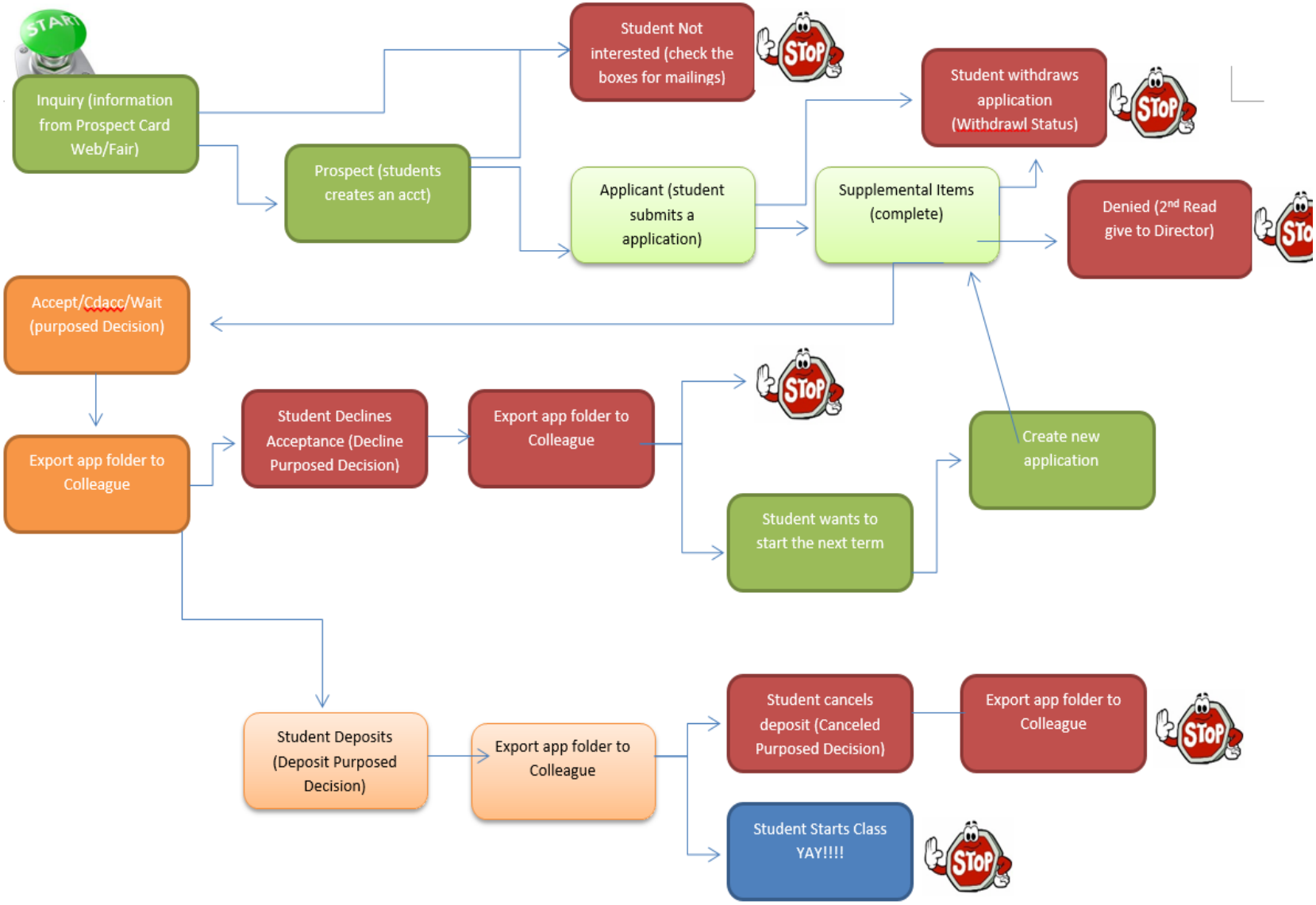
Often overlooked!

Organizational Silos

Initial Questions for Becker

- Determining data sources:
  - ACT, what data elements are included and can be collected
  - SAT, what data elements are included and can be collected
  - Clearing House, how long will it take to perform validation of students' enrollment to Becker
  - Accessing population data to be used in targeting areas
  - Do we know who are our true competitors?
  - What are we doing to target students differently?

Life Cycle of an Inquiry in Recruiter



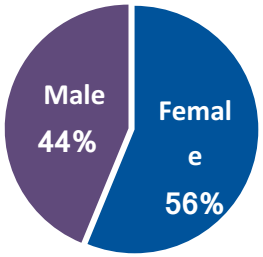
Building In-house Data – IR & IT Work Together, Happily

The process took 75% of time of the whole project!

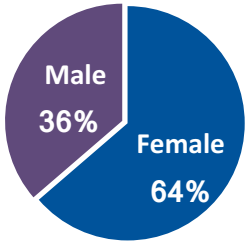
Miller, W. (2016). Conflict Resolution.

Student Profile

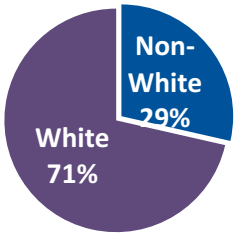
Gender 2014/15



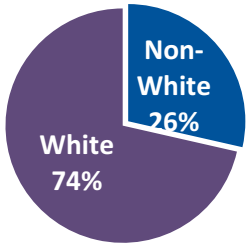
Gender 2016



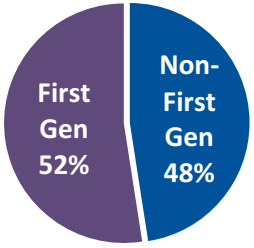
Ethnicity 2014/15



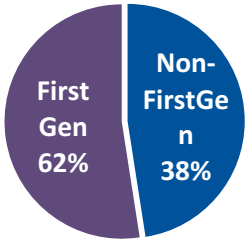
Ethnicity 2016



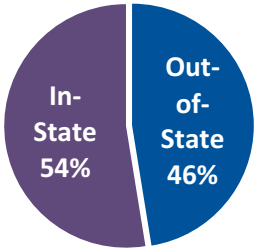
FirstGen vs. Non FirstGen 2014/15



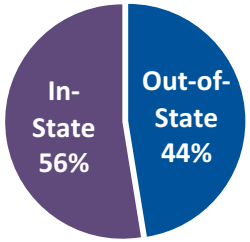
FirstGen vs. Non FirstGen 2016



In-State vs. Out-of-State 2014/15



In-State vs. Out-of-State 2016

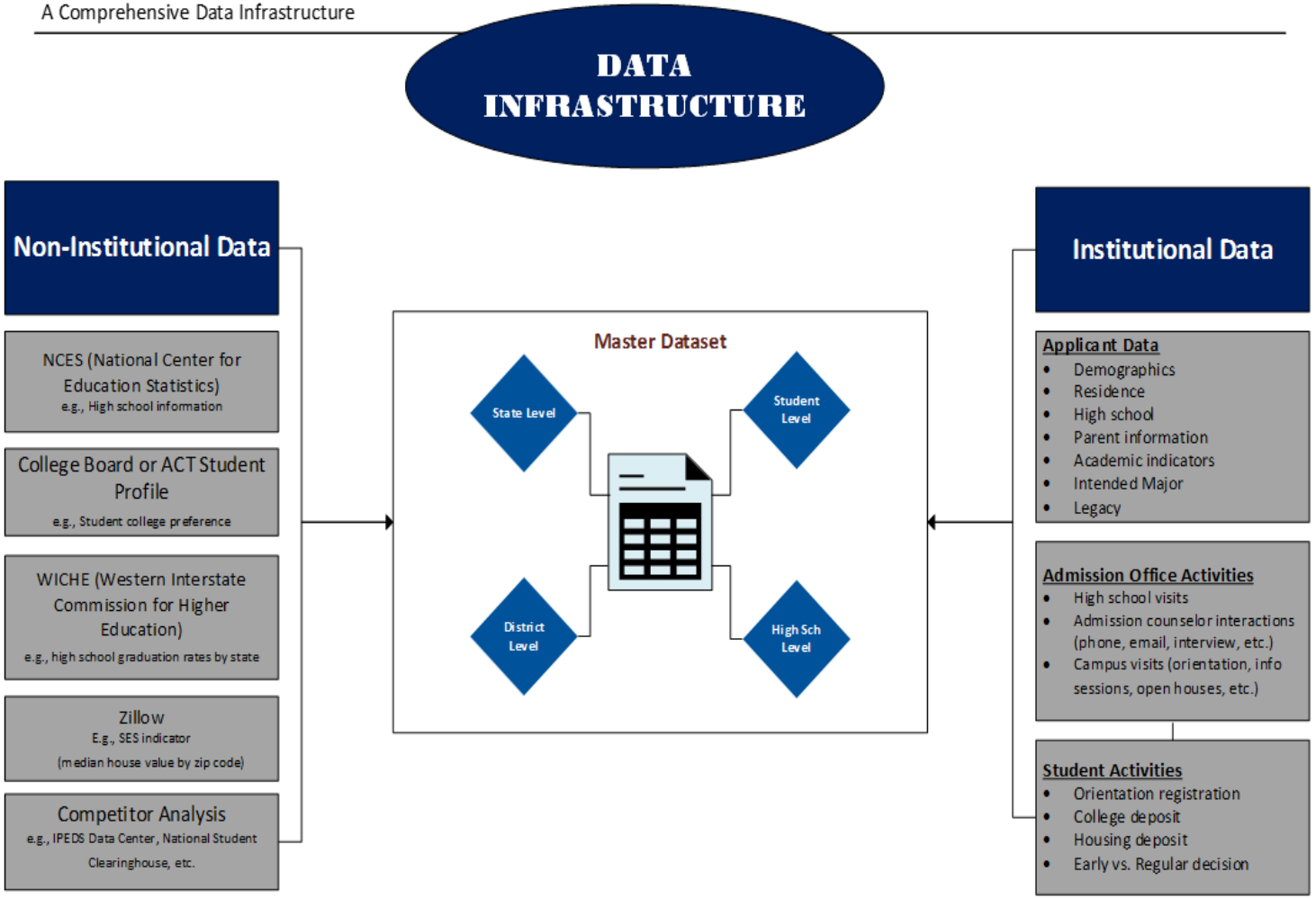


Step Three—Data Preparation

Be Realistic, Expect Changes!

Plan

Appendix A:
A Comprehensive Data Infrastructure



Note: 1) The final data set will consist of four data levels based on multiple data sources (non-institutional data and institutional data).

Step Four—Modeling

Predictive Modeling Details

Data

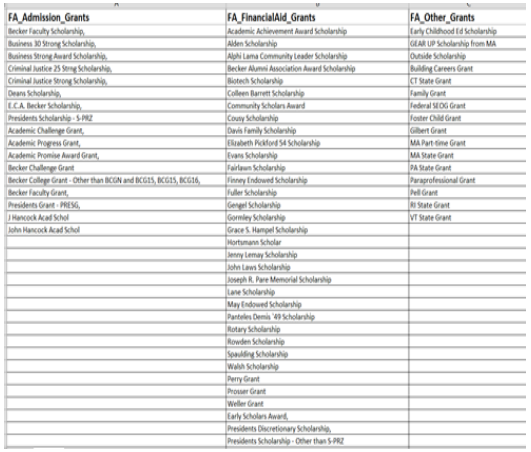
- Two years of historical data
- Descriptive: Characteristics and demographics (gender, race/ethnicity, geography, first-gen, Pell grant receiver, etc.)
- Financial aid: Need-based, merit-based, other grant, loan, work-study, etc.
- Behavioral: Application date; deposited date; Admission activities (phone call, email, campus visits, etc.)

Model

- Logistic regression
- Predict “deposit” probability for each student (values from 0 to 1)
- Test each predictor separately first vs putting everything in the model

Tools

- R/RStudio
- Learning curve, but it’s worth



FA_Admission_Grants	FA_FinancialAid_Grants	FA_Other_Grants
Becker Faculty Scholarship,	Academic Achievement Award Scholarship	Early Childhood Ed Scholarship
Business 30 Strong Scholarship,	Alden Scholarship	GEAR UP Scholarship from MA
Business Strong Award Scholarship,	Alphi Lama Community Leader Scholarship	Outside Scholarship
Criminal Justice 25 Strng Scholarship,	Becker Alumni Association Award Scholarship	Building Careers Grant
Criminal Justice Strong Scholarship,	Biotech Scholarship	CT State Grant
Deans Scholarship,	Colleen Barrett Scholarship	Family Grant
E.C.A. Becker Scholarship,	Community Scholars Award	Federal SEOG Grant
Presidents Scholarship - 5-PRZ	Crouy Scholarship	Foster Child Grant
Academic Challenge Grant,	Davis Family Scholarship	Gilbert Grant
Academic Progress Grant,	Elizabeth Pickford 54 Scholarship	MA Part-time Grant
Academic Promise Award Grant,	Evans Scholarship	MA State Grant
Becker Challenge Grant	Fairlawn Scholarship	PA State Grant
Becker College Grant - Other than BCGN and BCG15, BCG15, BCG16,	Finney Endowed Scholarship	Paraprofessional Grant
Becker Faculty Grant,	Fuller Scholarship	Pell Grant
Presidents Grant - PRESG,	Gengel Scholarship	RI State Grant
J Hancock Acad Schol	Gormley Scholarship	VT State Grant
John Hancock Acad Schol	Grace S. Hampel Scholarship	
	Hortsmann Scholar	
	Jenny Lemay Scholarship	
	John Laws Scholarship	
	Joseph R. Pare Memorial Scholarship	
	Lane Scholarship	
	May Endowed Scholarship	
	Panteles Denis '49 Scholarship	
	Rotary Scholarship	
	Rowden Scholarship	
	Spaulding Scholarship	
	Walsh Scholarship	
	Perry Grant	
	Prosser Grant	
	Weller Grant	
	Early Scholars Award,	
	Presidents Discretionary Scholarship,	
	Presidents Scholarship - Other than 5-PRZ	

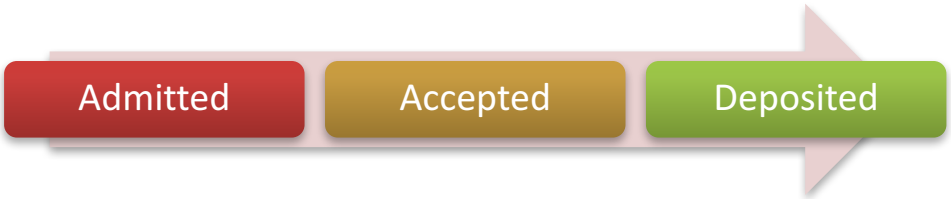
Data

- Descriptive (Characteristics and demographics)
  - Age
  - gender
  - race/ethnicity
  - Geography (regional – New England/Non-New England, MA/Non-MA, Worcester/Non-Worcester)
  - first generation (parental education)
  - Pell grant receiver
  - Sports activities count
  - External activities count

Data

- Financial aid:
  - need-based
  - merit-based
  - other grant
  - loan
  - work-study
  - family contribution

Data



- Behavioral:
  - Application date/applied term
  - deposited date
  - Admission activities
    - phone call count
    - email
    - campus visits (Acceptance Student Day, campus tours, etc.)

Model

- Over 800 lines of code
- Majority of code is data transformation and cleanup
- Supervised modeling
- Logistic regression

Model

- Linear Regression != categorical dependent variable
- Logistic Regression, Decision Trees, SVM, Random Forest.
- A classification algorithm, that predicts a binary (1/0 or True/False) outcome

Model

- We are using log of dependent variable
- Predicts the probability of occurrence of an event to a login function
- Part of a larger class of algorithms known as Generalized Linear Models (glm)

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

- g() – the link function (to 'link' the expectation of y to the predictor)
- E(y) – expectation of target variable
- $\alpha + \beta x1 + \gamma x2$ is the linear predictor (α, β, γ to be predicted)

Model

- Important Points
  - GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
  - The dependent variable need not to be normally distributed.
  - It does not uses OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).
  - Errors need to be independent but not normally distributed.

Why R?

Tools

- Open-source (free to use, improve on, and redistribute)
- Runs on most standard OS
- Released frequently
- Graphics capabilities are better than in the most other analytical packages
- Huge and very active user community
- It provides analysts more control over what changes to make and what assumptions to test.

Tools

- Excel
- RDBMS database
- CRM

Why In-house?

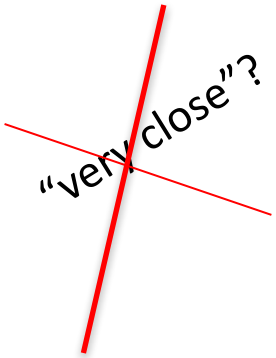
- Why in-house?
  - Most data cleaning can only be done in-house.
  - Colleges have a better control of the modeling process.
  - The process can be repeated once the codes are written.
  - More transparency and evaluation can be done in house.

Step Five—Evaluation

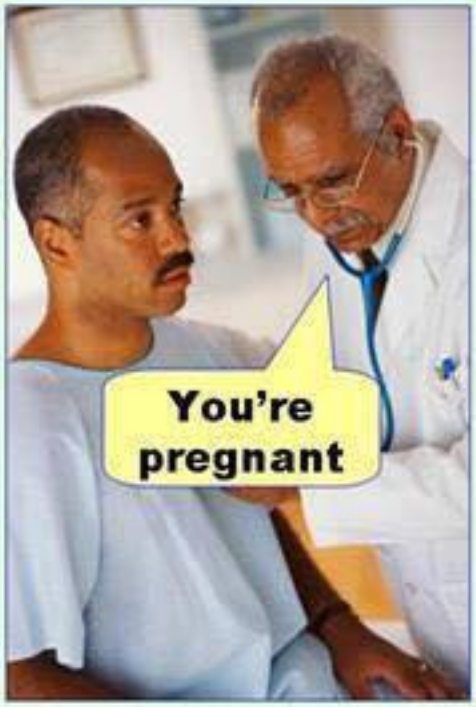
Evaluation step 1 – score models

- Is model good enough?
  - Likelihood Ratio test
  - Area under the ROC curve
- Are predictions accurate?
  - Confusion matrix
- Which Predictors are most important?

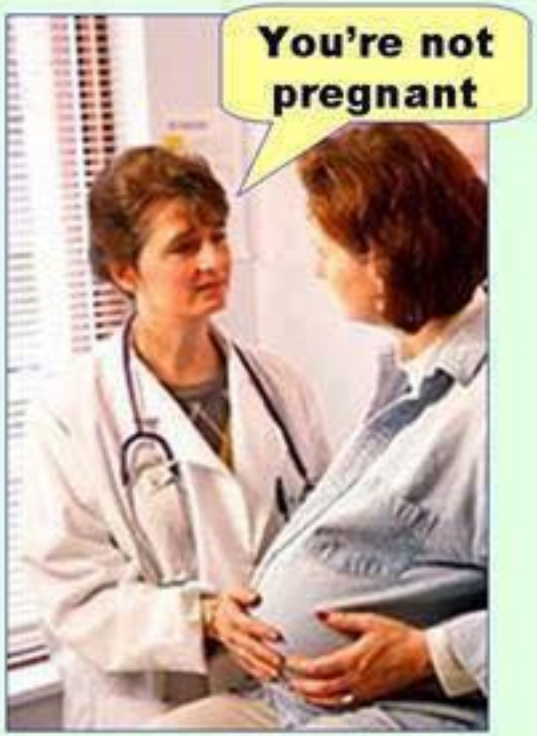
Evaluation Question – Are predictions accurate?



Type I error
(false positive)



Type II error
(false negative)



Evaluation Question – Are predictions accurate?

- Confusion matrix is a tabular representation of Actual vs Predicted.
- Helps to find the accuracy of the model and avoid overfitting

		Predicted	
		Good	Bad
Actual	Good	True Positive (d)	False Negative (c)
	Bad	False Positive (b)	True Negative (a)

Evaluation Question – Are predictions accurate?

- (True Positive + True Negatives)/(True Positive + True Negatives + False Positives + False Negatives)
- From confusion matrix, Specificity and Sensitivity can be derived as:

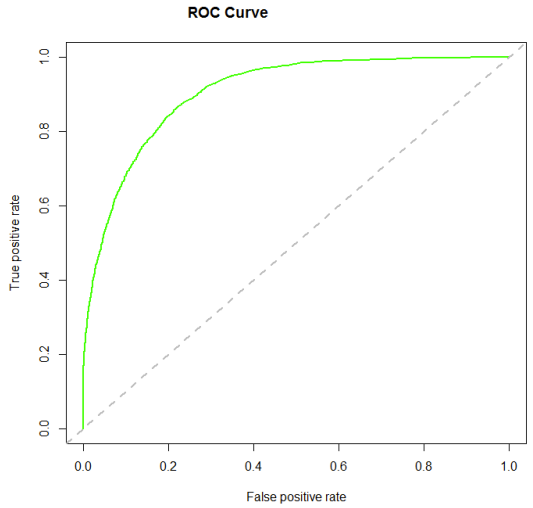
$$\left.\begin{aligned} \text{True Negative Rate (TNR), specificity} &= \frac{A}{A+B} \\ \text{False Positve Rate (FPR), 1 – specificity} &= \frac{B}{A+B} \end{aligned}\right\} \text{sum to 1}$$

$$\left.\begin{aligned} \text{True Positive Rate (TPR), sensitivity} &= \frac{D}{C+D} \\ \text{False Negative Rate (FNR)} &= \frac{C}{C+D} \end{aligned}\right\} \text{sum to 1}$$

Evaluation Question – Are predictions accurate?

- Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity).
- library(ROCR)
- Assume p > 0.5
- Performance metric for ROC curve is the area under curve (AUC)

-Higher the area under curve, better the prediction power of the model.
-The ROC of a perfect predictive model has TP equals 1 and FP equals 0.
-This curve will touch the top left corner of the graph.



Evaluation step 2 – review the model

- Did we miss anything?
- Any assumptions violated?

Evaluation step 3 – next step

- Deploy vs. recreate model

Step Six—Deployment

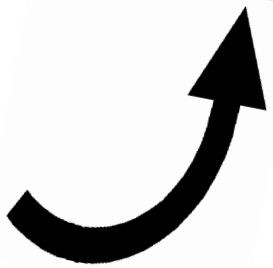
Actually using your model!

- Automation
- Getting feedback from model's output users
- Experimenting and feeding the new learning back to the model to improve its accuracy
- Monitoring outcomes of the model use

Deploy Results—
Be Responsive & Improve the Admissions Practice

	A	B	C	D	E
1	**Student ID**	**Student Name**	**Application..Source**	**likelihood**	**Category**
2	213628		Common Application	0.023829	Low
3	218268		Common Application	0.959735	High
4	215861		Common Application	0.837514	Medium
5	219100		Common Application	0.003928	Low
6	207435		UG Short Form	0.00519	Low
7	218509		Common Application	0.003123	Low
8	221393		Common Application	0.712934	Medium
9	229274		Common Application	0.00575	Low
10	226531		Common Application	0.448855	Low
11	221690		Common Application	0.519776	Medium
12	228830		Common Application	0.958078	High
13	225195		Common Application	0.990998	High
14	220969		Common Application	0.403066	Low
15	219295		Common Application	0.004229	Low
16	222623		Common Application	0.028212	Low
17	228026		UG Short Form	0.947711	High
18	227242		UG Short Form	0.965942	High
19	227239		UG Short Form	0.691129	Medium
20	228548		UG Short Form	0.653397	Medium

1. The admissions office will receive a file every two weeks.
2. The file will be uploaded to Recruiter.
3. The VP of Enrollment will take actions.



Challenges in Using Predictive Analytics

- **Obstacles in management**

No champion for the work

- **Obstacles with data**

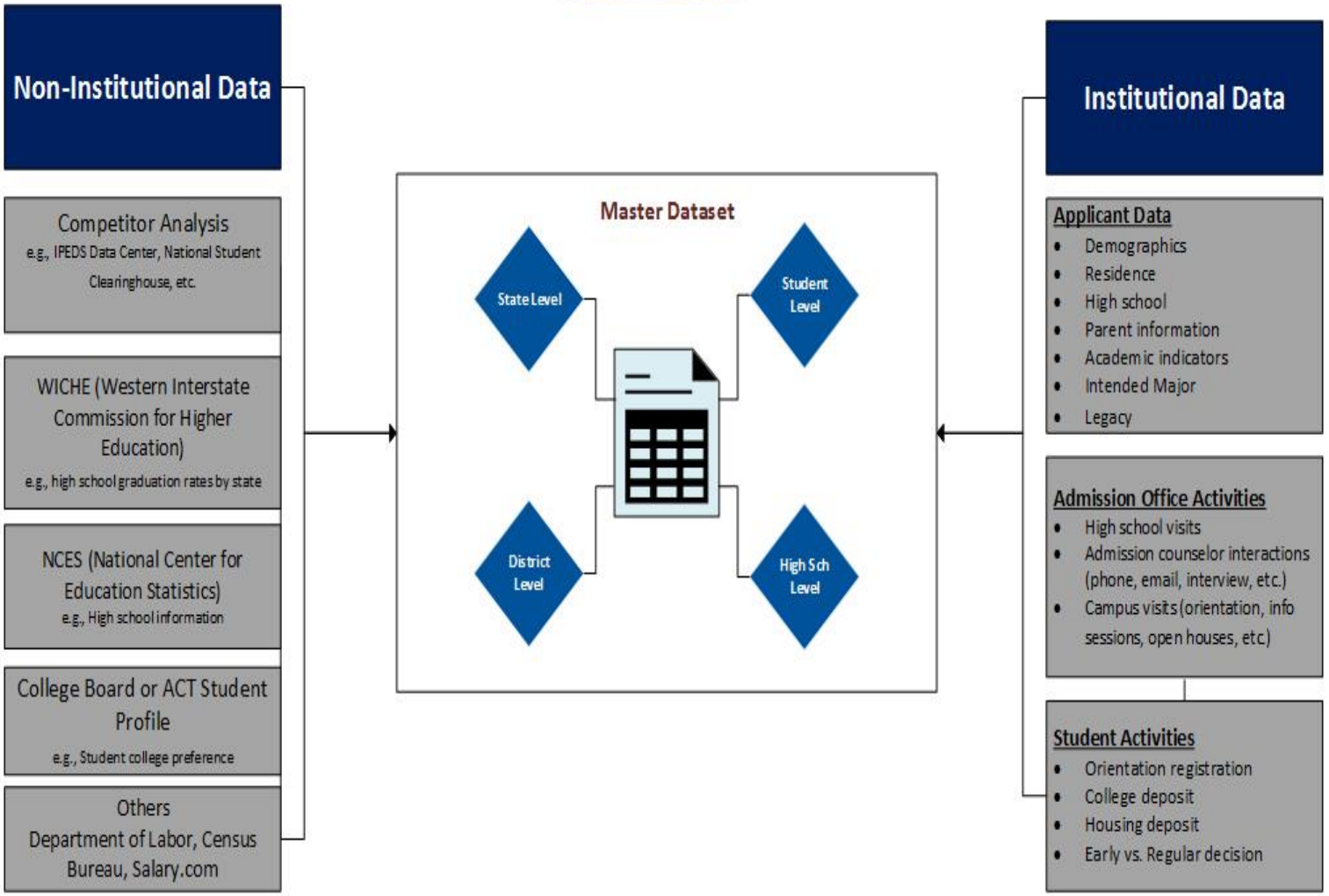
Data need → Data infrastructure → Data consistency

- **Obstacles with modeling**

Analyst too zealous or too ambitious

Model is too complex (overfitting – significant relationshpis are just noise)

DATA INFRASTRUCTURE



Next Steps

Build more analytic work

- What areas or school districts should we spend more resources on (visits, calls, marketing campaign)?
- How do we improve the academic profile of incoming students?
- How can we create new aid models to leverage our intuitional aid to achieve higher yield?
- Retention - which students are most likely to drop out/transfer?
- Utilizing unstructured data to gain additional insights

Actually USING the results

- What actions need to take place based on scores generated from predictive models?
- How do we assess the effectiveness of our uplifting marketing campaigns (including customized emails, number of phone calls, and other marketing activities)?

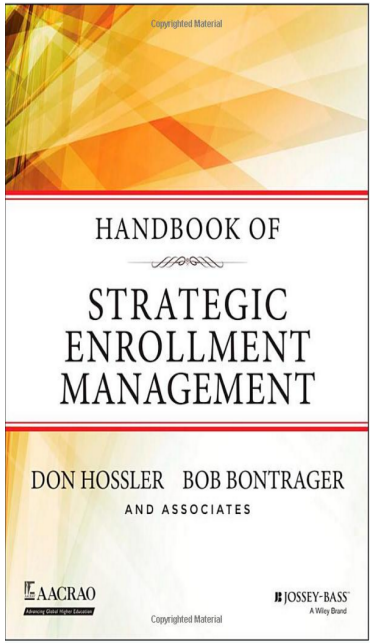
How else can data analytics help my school?

- Increasing students' retention
- Increasing students' graduation
- Increasing students'/teachers' performance
- Reducing students' absence

Take Away Messages

- Be realistic to narrow the scope of the predicative modeling work at the initial stage
- Evaluation of models should be in the process
- Building data structure is more important than methodology

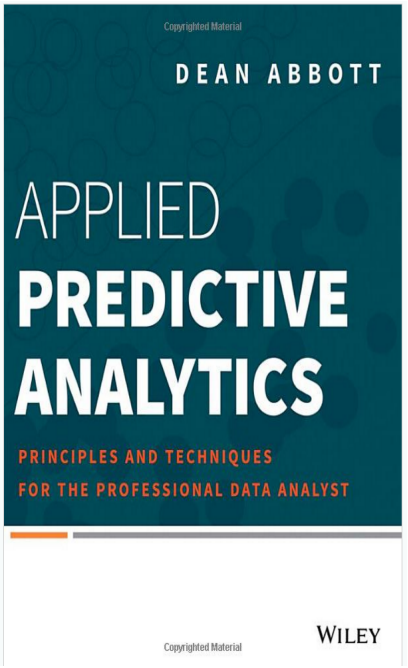
Predictive Modeling Steps and Details in the Context of College Enrollment



Details in strategic enrollment management process but very general guideline on modeling (Page 223-227).

Link the two sides

Build predictive analytics in the context of college enrollment



Practical suggestions on predictive analytics but focus on the economic field of fraud detection & customer satisfaction.

References

Abbot, D. (2014). Overview of predictive analytics. *Applied predictive analytics: Principles and techniques for the professional data analyst (17).* Indianapolis, IN: Wiley.

Abbot, D. (2014). Setting up the problem. *Applied predictive analytics: Principles and techniques for the professional data analyst (19).* Indianapolis, IN: Wiley.

Berg, B. (2012). Predictive modeling: A tool, not the answer: Benefits and cautions of using historical data to predict the future. *University Business*. Retrieved from: http://www.universitybusiness.com/article/predictive-modeling-tool-not-answer

Bergerson, A.A. (2010). College choice and access to college: Moving policy, research and practice to the 21st century. *ASHE Higher Education Report, 35*(4). San Francisco, CA: Jossey-Bass.

Cabrera, A.F. (1994). Logistic regression analysis in higher education: An applied perspective. In J.C. Smart (ed.), *Higher Education: Handbook of Theory and Research, 10,* (225-256). New York, NY: Agathon Press.

Davis, C.M., Hardin, J.M., Bohannon, T., Oglesby, J. (2007). Data mining applications in higher education. In K.D. Lawrence, S. Kudyba, & R.K. Klimberg (Eds.), *Data Mining Methods and Applications* (123-147). Boca Raton, FL: Auerbach Publications.

Hosmer, D.W., Lemeshow, S. (2000). Applied logistic regression. *(2).* New York, NY: John Wiley & Sons, Inc.

Hossler, D. (1991). Evaluating student recruitment and retention programs. *New Directions for Institutional Research, 70.* San Francisco, CA: Jossey-Bass.

Hossler, D., & Bontrager, B. (2015). *Handbook of strategic enrollment management.* San Francisco, CA: Jossey-Bass.

Hossler, D., Gallagher, K. S. (1987). Studying student college choice: A three- phase model and the implications for policymakers. *College and University, 62*(3), 207-221.

Hovland, M. (2004). Unraveling the mysteries of student college selection. *Paper presented at the 2004 ACT Enrollment Planner's Conference.* Chicago, IL.

Prescott B. & Bransberger, P. (2013). *Knocking at the College Door: Projections of High School Graduates by State, Income, and Race/Ethnicity,* Boulder, CO: Western Interstate Commission for Higher Education.

Luan, J. (2002). Data mining and its applications in higher education. *New Directions for Institutional Research, 113,*17-36.

McPherson, M.S. (1991). Does student aid affect college enrollment? New evidence on a persistent controversy. *The American Economic Review, 81*(1), 309-318.

Perna, L. (2006). Studying college access and choice: A proposed conceptual model. *Higher Education: Handbook of Theory and Research, 21,* 99-151.

Sigillo, A. (2015). *Predictive modeling in enrollment management: New insights and techniques.* Retrieved from: http://www.uversity.com/downloads/research/EI%20Whitepaper_R6.pdf


Questions

Thank you!

- Email: fbagirov@harrisburgu.edu
- Twitter: @FeyziBagirov