How Machine Learning and Multiple Measures are Reshaping College Placement

Terrence Willett, John Hetts, Craig Hayward

CAIR Conference
Concord
November 8, 2017
Acknowledgements

• Current and former MMAP team members
• Academic Senate for California Community Colleges
• Common Assessment Initiative Steering Committee
• Pilot colleges
• High School participants
• California Community College Chancellor’s Office MIS team
Transitions and intersegmental trust

- Within systems: highly reliable progression after successful completion
- Between systems – different story
- HS to CSU
  - 38% repeat previously completed coursework, ~60% African Americans, 45% of Hispanics
- HS to CCC transition
  - ~3/4 repeat ≥ 1 level, ~1/2 repeat ≥ 2 levels of math
  - African Americans & Hispanics ~60% more likely, Female students ~20% more likely
- Noyce Foundation report
  - Algebra in 8th grade, ~2/3 repeat including 50% of students with B or better
  - Algebra in 7th grade advance to Geometry in 8th grade
Data Set for the Models

• California Community College (CCC) students enrolled in an English, Math, Reading or ESL class with matching high school data in California Partnership for Achieving Student Success (CalPASS) statewide intersegmental database
  • ~1 M cases for Math & English; ~200k for Reading & ESL
• Bulk of first CCC enrollments from 2008 through 2014
• Rules were developed with the subset of students who had four years of high school data (about 25% of total sample)
• Used machine learning \textit{rpart} package in R to create decision trees
  – \url{http://rpgroup.org/Our-Projects/All-Projects/Multiple-Measures/PilotCollegeResources} see Decision Rules and Analysis Code -> Using R for Creating Predictive Models
• R4IR Tutorial \url{https://drive.google.com/drive/folders/0Bz-jqwGzLQjJajA5YUIxUjdETzA?usp=sharing}
Variables Explored in the Models

- High School Unweighted Cumulative GPA
- Grades in high school courses
- CST scores
- Advanced Placement course taking
- Taking higher level courses (math)
- Delay between HS and CCC (math)
- HS English types (expository, remedial, ESL)
- HS Math level (Elem. Algebra, Integrated Algebra, Pre-Calculus)
What are Decision Trees?

• Howard Raiffa explains decision trees in *Decision Analysis* (1968).

• Ross Quinlan invented ID3 and introduced it to the world in his 1975 book, *Machine Learning*.

• CART popularized by Breiman et al. in mid-90’s
  – Based on information theory rather than statistics; developed for signal recognition
Engineering Flowchart

DOES IT MOVE?
Increasing Homogeneity with each split
How is homogeneity measured?

\[ D = 1 - \sum_{i=1}^{n} p_i^2 \]

- Gini-Simpson Index
- p-square = probability of two items taken at random from the set being of same types; D=dissimilarity/diversity
- Proposed by Corrado Gini in 1912 as a measure of inequality of income or wealth; used in demographics and ecology as diversity index
- If selecting two individual items randomly from a collection, what is the probability they are in different categories.
- Other indices such as Shannon-Wiener can also be used
Key considerations

• Splitting criterion: how small should the leaves be? What are the minimum # of splits?
• Stopping criterion: when should one stop growing the branch of the tree?
• Pruning: avoiding overfitting of the tree and improving
• Understanding classification performance
Loading Data in R

# set working directory for location of data
setwd("C:/Users/Me/Documents/MMAPData")

# Load data
MMAPMath <- read.csv("C:/Folder/MMAPMath.csv", header=T)

# save data and analyses to working directory
save.image("MMAPMath.RData")

Basic Classification Decision Tree

#CART packages

library(rpart)
library(rpart.plot)

#set control parameter
ctrl <- rpart.control(minsplit = 100, cp = 0.0015, xval=10)  # control specs here

cartfit_m5statpoisson <- rpart(formula = CC_FIRST_COURSE_SUCCESS_IND ~ HS_11_GPA_CUM + PRE_ALG_ANY_C + ALG_I_ANY_C + ALG_II_ANY_C + GEO_ANY_C + TRIG_ANY_C + PRE_CALC_ANY_C + CALC_ANY_C + STAT_ANY_C + STAR_MATH_EAP_IND + HS_EXIT_SUBJ_TO_CC_ENTRY_SUBJ + AP_ANY_C + [CST score and subscale variables]
  ,data = m5stat
  ,method="poisson"  # Change method here to test different distributions
  ,control=ctrl)  # Change control specs here
Splitting Methods

- Class = used for categorical dependent var
- ANOVA = used for continuous dependent var
- Poisson = used for count of events in time frame such as survival data
- Exponential = can also be used for survival with different distributional assumptions
CART Output and Diagnostics

> printcp(cartfit_m5statpoisson) ← shows relative error by cp value
> print(cartfit_m5statpoisson) ← indented text print out of tree
> rsq.rpart(cartfit_m5statpoisson) ← graph showing error by # splits
> prp(cartfit_m5statpoisson,main="Transfer Level Statistics"
    ,extra=100,varlen=0,left=FALSE) ← graph tree
Pros and Cons of Decision Trees

**Strengths**
- Visualization
- Easy to understand output
- Easy to code rules
- Model complex relationships easily
- Linearity, normality, not assumed
- Handles large data sets
- Can use categorical and numeric inputs

**Weaknesses**
- Results dependent on training data set – can be unstable esp. with small N
- Can easily overfit data
- Out of sample predictions can be problematic
- Greedy method selects only ‘best’ predictor
- Must re-grow trees when adding new observations
Statistics Decision Tree for Direct Matriculants

Root Node

HS_11_GPA_CUM >= 3

Node 1

Branch

Internal node/split

HS_11_GPA_CUM >= 3.3

Node 2

Node 4

PRE CALC_UP11_C >= 0.5

Node 8

Node 9

ALG II_UP11_C >= 0.5

Node 10

Node 11

Node 12

Node 13

Predicted success rate

Percent of students in node

Branch
## MMAP Transfer-Level Placement Recommendations

<table>
<thead>
<tr>
<th>Transfer Level Course</th>
<th>Direct Matriculant</th>
<th>Non-Direct Matriculant</th>
</tr>
</thead>
<tbody>
<tr>
<td>College Algebra (STEM)</td>
<td>HS 11 GPA &gt;=3.2 OR</td>
<td>HS 12 GPA &gt;=3.2 OR</td>
</tr>
<tr>
<td>Passed Algebra II (or better)</td>
<td>HS 11 GPA &gt;=2.9 AND Pre-Calculus C (or better)</td>
<td>HS 12 GPA &gt;=3.0 AND Pre-Calculus or Statistics (C or better)</td>
</tr>
<tr>
<td>Statistics (General Education/Liberal Arts)</td>
<td>HS 11 GPA &gt;=3.0 OR</td>
<td>HS 12 GPA &gt;=3.0 OR</td>
</tr>
<tr>
<td>Passed Algebra I (or better)</td>
<td>HS 11 GPA &gt;=2.3 AND Pre-Calculus C (or better)</td>
<td>HS 12 GPA &gt;=2.6 AND Pre-Calculus (C or better)</td>
</tr>
<tr>
<td>English</td>
<td>HS 11 GPA &gt;=2.6</td>
<td>HS 12 GPA &gt;=2.6</td>
</tr>
</tbody>
</table>

“Under our previous policies, African American and Latino students were far less likely to place into transfer-level math. Under the new policies, African American students’ access to transfer-level math increased eightfold, Latino students’ access increased four-fold, and the disproportionate impact in placement was eliminated for all racial groups.”
– Santa Monica College

“MMAP is a COMPLETION initiative, not a SUCCESS initiative.”
– Santa Monica College

“There are thousands of reasons to do this; each one has a name.”
– Bakersfield College
Transfer level placement by year/method in Math at Cuyamaca

<table>
<thead>
<tr>
<th>Group</th>
<th>Fall 2015 (Any transfer)</th>
<th>Fall 2016 (STEM ONLY)</th>
<th>Fall 2016 (STEM + Statistics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>36%</td>
<td>90%</td>
<td>79%</td>
</tr>
<tr>
<td>African American</td>
<td>9%</td>
<td>49%</td>
<td>73%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>21%</td>
<td>85%</td>
<td>62%</td>
</tr>
<tr>
<td>White</td>
<td>27%</td>
<td>84%</td>
<td>62%</td>
</tr>
<tr>
<td>All</td>
<td>24%</td>
<td>84%</td>
<td>62%</td>
</tr>
</tbody>
</table>
Gateway momentum in Math at Cuyamaca

Successful completion of transfer-level math before and after change by assessment level

<table>
<thead>
<tr>
<th>Assessment Level</th>
<th>Fall 2013 Cohort</th>
<th>Fall 2016 Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three+ Levels Below</td>
<td>4%</td>
<td>56%</td>
</tr>
<tr>
<td>Two Levels Below</td>
<td>19%</td>
<td></td>
</tr>
<tr>
<td>One Level Below</td>
<td>36%</td>
<td>66%</td>
</tr>
<tr>
<td>All</td>
<td>23%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Successful completion of transfer-level math before and after change by ethnicity

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Fall 2013 Cohort</th>
<th>Fall 2016 Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>33%</td>
<td>75%</td>
</tr>
<tr>
<td>African American</td>
<td>6%</td>
<td>55%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>15%</td>
<td>65%</td>
</tr>
<tr>
<td>White</td>
<td>16%</td>
<td>76%</td>
</tr>
<tr>
<td>All</td>
<td>15%</td>
<td>69%</td>
</tr>
</tbody>
</table>
Gateway momentum in English at Skyline

**English placement by level and cohort**

- Transfer-Level: 80% (Fall 2013), 47% (Fall 2016)
- One Level Below: 40% (Fall 2013), 17% (Fall 2016)
- Two Levels Below: 14% (Fall 2013), 3% (Fall 2016)

**Successful rate by cohort and course type**

- Fall 2013: Transfer Level (F/Datamart) - 67%
- F2015-S2017 (traditional) - 65%
- F2015-2017 (w/support) - 69%
Cañada College Transfer-level Placements

<table>
<thead>
<tr>
<th>Course</th>
<th>Compass</th>
<th>MMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2014 Math</td>
<td>191</td>
<td></td>
</tr>
<tr>
<td>F2015 Math</td>
<td>192</td>
<td>53</td>
</tr>
<tr>
<td>F2014 Eng</td>
<td>123</td>
<td></td>
</tr>
<tr>
<td>F2015 Eng</td>
<td>123</td>
<td>134</td>
</tr>
</tbody>
</table>

Cañada College Transfer-level Success Rates

<table>
<thead>
<tr>
<th>Course</th>
<th>Compass</th>
<th>MMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>65%</td>
<td>68%</td>
</tr>
<tr>
<td>English</td>
<td>78%</td>
<td>75%</td>
</tr>
</tbody>
</table>

Rule set: English = 2.3 AND B- or better; Math = 3.2 AND C or better

[bit.ly/MMAPPilotLessons]
Various Placement Systems and Their Impact on Student Equity
Placement Error

- **Overplacement**: Student is placed above their ability to succeed. Highly visible.

- **Underplacement**: Student could have been successful at a higher level than where placed. Tends to be invisible.

- Current placement systems tend to result in much greater underplacement error.
Evaluating Placement Systems

**Disjunctive placement:**
Take the highest placement (Test or MMAP)
Recommended by MMAP

**Compensatory placement:**
Logistic regression (combines Test, MMAP simultaneously)
Run with two cut-values: 0.70, 0.50

**Conjunctive placement:**
Only if Test and MMAP in agreement
Highly restrictive
Not recommended by the CCCCCO
Accuracy: College Statistics Placement

Accurate Placement in College Statistics

- **Conjuctive**: 61.0%
- **Compensatory (0.70)**: 57.4%
- **Disjunctive (0.70)**: 64.5%
- **Compensatory (0.50)**: 68.1%

*Negatives are unknown for the disjunctive models, so accuracy cannot be completely calculated for disjunctive model.*
One Year Throughput Rate: College Statistics Course

Statistics Class Throughput rate by Placement System

Throughput Rate

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Conjunctive
Compensatory (0.70)
Disjunctive (0.70)
Compensatory (0.50)

13.3%
17.1%
32.2%
40.8%
### Percentage of Underrepresented Students of Color College-level Placements

#### Percentage College-level-placed Students who are URSC by Placement System

<table>
<thead>
<tr>
<th>Placement System</th>
<th>Percent of Transfer-placed who are URM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjunctive</td>
<td>42.3%</td>
</tr>
<tr>
<td>Compensatory (0.70)</td>
<td>47.5%</td>
</tr>
<tr>
<td>Disjunctive (0.70)</td>
<td>55.3%</td>
</tr>
<tr>
<td>Compensatory (0.50)</td>
<td>56.4%</td>
</tr>
</tbody>
</table>

**Note:** The bar chart visualizes the percentage of transfer-placed students who are underrepresented students of color (URSC) by placement system.
Summary of Modeling Placement Systems

- No single metric is sufficient but several well-chosen metrics (including throughput) can allow for a more informed decision
- Disjunctive models have higher access and throughput than compensatory models
- The conjunctive model was very restrictive and had the lowest throughput rates and URM placement rates
- Students placed via alternative methods
  - far more likely to be placed into college-level courses
  - successfully complete college-level courses at the same or higher rates when placed there
  - far more likely to complete the gateway course in the discipline
- Students should progress between systems as smoothly as within systems
MMAP Research Team

Terrence Willett
The RP Group
twillett@rpgroup.org

Mallory Newell
The RP Group
newellmallory@deanza.edu

Craig Hayward
The RP Group
chayward@rpgroup.org

Loris Fagioli
The RP Group
lfagioli@ivc.edu

Rachel Baker
UC Irvine
rachelbb@uci.edu

Nathan Pellegrin
The RP Group
nathan.pellegrin@gmail.com

Peter Bahr
University of Michigan
prbahr@umich.edu

John Hetts
Educational Results Partnership
jhetts@edresults.org

Ken Sorey
Educational Results Partnership
ken@edresults.org

Daniel Lamoree
Educational Results Partnership
dlamoree@edresults.org

---

EDUCATIONAL RESULTS PARTNERSHIP

theRPgroup
Research • Planning • Professional Development for California Community Colleges