Office of Admissions
Enrollment Prediction Project

Jienian Zhang
Li Zhang
Yichi Zhang
Variable Selection

• Variables that we think are important for predicting enrollment:
  • Residency*
  • Application Month*
  • ACT Choice*
  • Legacy
  • Campus Visits
  • RAI (we will limit this presentation to RAI-present students)
  • GPA
  • Scholarship

* Variables that required some re-coding for model building
Variable Manipulation/Transformation

- **ACT Choice** (1, 2, 3, 4, 5, 6, C, S, or nothing sent)
  - New coding has 4 classes:
    - ‘nothing sent’, 1, S, or ‘other’ ('C' was never observed)

- **Application Month** (1 through 12)
  - New coding has 3 classes:
    - Early (May-July)
    - Normal (August – January)
    - Late (February – April)
Variable Manipulation/Transformation

- **Residency** (many possible entries, 54 levels)
  - Information used for modeling is in the three new variables below:
    - Iowa resident (yes or no)
    - Illinois resident (yes or no)
    - Wisconsin resident (yes or no)

- Variables Not used in the model building process:
  Gender, Major, Admission Score, Contact and Some Scholarships
Some Data Demographics

2011 - RAI present subset

Sample size: 9,576
35% enrolled (65% did not)
34% from Iowa
49% from Illinois
3% from Wisconsin
49.15% went on a campus visit

The ‘model building’ data set

GPA

Minimum: 2.140
Maximum: 5.32
Mean: 3.683
Std deviation: 0.4013

RAI

Minimum: 181
Maximum 410.0
Mean: 303.4
Std deviation: 30.89
2011 - RAI present subset (n=9,576)

ACT Choice
19% chose Iowa as the 1st ACT score recipient
20% chose Iowa as the 2nd ACT score recipient
46% chose Iowa as the supplementary ACT score recipient

Application Month
Early: 3.8% (May-Sept)
Normal: 91.0% (Aug.-Jan.)
Late: 5.2% (Feb.-April)

10.6% was legacy

Exploring relationships with Residency...

Iowa residents (n=3,272)
57.39% visited Campus
21.4% was legacy
5.38% applied early
7.77% applied late
55.66% enrolled

Non-Iowa residents (n=6,304)
42.96% visited Campus
6.19% was legacy
3.00% applied early
3.83% applied late
23.54% enrolled
### 2011 - RAI present subset (n=9,576)

*Exploring relationships with residency and campus visits...*

<table>
<thead>
<tr>
<th>Iowa residents (n=3,272)</th>
<th>Non-Iowa residents (n=6,304)</th>
</tr>
</thead>
<tbody>
<tr>
<td>55.66% enrolled</td>
<td>23.54% enrolled</td>
</tr>
</tbody>
</table>

#### Iowa residents who visited campus

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>64.26% enrolled</td>
<td></td>
</tr>
<tr>
<td>If was a legacy, 64.64% enrolled</td>
<td></td>
</tr>
<tr>
<td>If applied early, 70.08% enrolled</td>
<td></td>
</tr>
<tr>
<td>If applied normal time, 64.27% enrolled</td>
<td></td>
</tr>
<tr>
<td>If applied late, 57.46% enrolled</td>
<td></td>
</tr>
</tbody>
</table>

#### Non-Iowa residents who visited campus

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>37.33% enrolled</td>
<td></td>
</tr>
<tr>
<td>If was a legacy, 39.41% enrolled</td>
<td></td>
</tr>
<tr>
<td>If applied early, 45.14% enrolled</td>
<td></td>
</tr>
<tr>
<td>If applied normal time, 36.59% enrolled</td>
<td></td>
</tr>
<tr>
<td>If applied late, 50.00% enrolled</td>
<td></td>
</tr>
</tbody>
</table>
A classification tree (example at right) can be used to categorize students on numerous variables, then calculate a percent enrolled.

The tree automatically chooses the ‘best’ variables on which to split the data. Here, it chose Iowa residency (yes vs. no), ACT school choice (1, S vs. others), Campus Visit (yes vs. no).

The tree suggests that Campus Visit matters for the non-Iowa residents, but not for the Iowa residents.
Model building – cross validation

• We used a cross-validation technique to build our model (i.e. choose which set of variables was relevant for predicting enrollment)

• This process splits the data set into a training set and test set. The model is built using the training data and tested on the test data.

• This process tries to mimic the real-life use of the model which is prediction in a brand new set of data (i.e. we model using 2011 data set and then make the prediction with 2012 data set)

• The model chosen included all variables that were investigated: RAI+Legacy+In_state+CampusV+ACTchoice+OldG+NSA+Stealth+Month
Model using 2011 data with RAI

- `model1<-glm(Enll~RAI+Legacy+In_state+CampusV+ACTchoice+OldG +NSA+Stealth+Month, family=binomial(logit), data=subset.cc1)`

- Coefficients:

|                           | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------------------|----------|------------|---------|----------|
| (Intercept)               | 1.169348 | 0.344267   | 3.397   | 0.000682 *** |
| RAI                       | -0.012756| 0.001133   | -11.254 | < 2e-16 *** |
| Legacy                    | 0.128019 | 0.077636   | 1.649   | 0.099156 .  |
| In_state                  | 1.364138 | 0.078359   | 17.409  | < 2e-16 *** |
| CampusV                   | 1.097779 | 0.051030   | 21.513  | < 2e-16 *** |

Continued on next slide...
Model using 2011 data with RAI

- **ACTchoice1** 1.964073 0.105227 18.665 < 2e-16 ***
- **ACTchoiceother** 1.013916 0.105228 9.635 < 2e-16 ***
- **ACTchoiceS** 1.355831 0.096315 14.077 < 2e-16 ***
- **OldG** 0.559871 0.095310 5.874 4.25e-09 ***
- **NSA** 0.265973 0.084517 3.147 0.001650 **
- **Stealth** 0.195002 0.077804 2.506 0.012199 *
- **Monthlate** -0.052479 0.158767 -0.331 0.740992
- **Monthnormal** -0.571639 0.120623 -4.739 2.15e-06 ***
Predicting Enrollment

• Using the fitted model and an applicant’s specific characteristics, we can generate an individual PROBABILITY OF ENROLLMENT $\hat{p}$.

• For example, student ID=9794 did not enroll and has characteristics:

  RAI=296  Legacy=No  In_state=No
  CampusV=No  ACTchoice=S  OldG=No
  NSA=1  Stealth=No  Month=9(Normal)

  and has a predicted probability of enrollment of $\hat{p} = 0.1791$. 
Predicting Enrollment

• Just because we can create a predicted probability of enrollment $\hat{p}$ for each student, it doesn’t necessarily mean our predictions are very good.

• But certainly, we would expect most students with relatively low $\hat{p}$ values (e.g. less than 0.3) to not enroll.

• And we would expect most students with relatively high $\hat{p}$ values (e.g. greater than 0.7) to enroll.

So, how well do we do...
Plot of the true enrollment status vs. $\hat{p}$

Incorrectly predicted to not enroll

Correct prediction (not enrolled)

Correct prediction (enrolled)

Incorrectly predicted to enroll

Actual enrollment

Predicted probability of enrollment

0.0 0.2 0.4 0.6 0.8 1.0
Plot of the true enrollment status vs. $\hat{p}$

- **High probability** (above 0.7), then predicted to enroll. Correct rate = $\frac{770}{1029} = 75\%$.
- **Low probability** (below 0.3), then predicted to *not* enroll. Correct rate = $\frac{4006}{4712} = 85\%$.
- **Mediate Probability** (above 0.3 and below 0.7), then predict undecided. Correct rate = $\frac{2274}{3835} = 0.59$.
Conclusions from Model Building

• Who is the low-probability group:
  From out of state, did not visit campus and have a relatively high RAI
  
  \[ P(\text{From out of state}) = 0.9555 \]
  \[ P(\text{Did not make a campus visit}) = 0.9265 \]
  Relatively high RAI \( \rightarrow 309.0157 \)

• Who is the high-probability group:
  From Iowa, visited campus, have a relatively low RAI
  
  \[ P(\text{From Iowa}) = 0.9942 \]
  \[ P(\text{Did make a campus visit}) = 0.9591 \]
  Relatively low RAI \( \rightarrow 291.9328 \)
Follow-up: August 2014 predictions (N=6989)

Correct rate = 180/253 = 71%
(model building data set was 75%).

Correct rate = 4954/5632 = 88%
(model building data set was 85%).

Actual enrollment

predicted probability of enrollment
2014 predictions vs. 2011 model building data set

• Similarities
  • The proportion of correct predictions from the low and high groups
    • Low: 85% (2011) and 88% (2014)
    • High: 75% (2011) and 71% (2014)

• Differences
  • The proportion of students falling into the low and high groups
    • Low: 49% (2011) and 81% (2014)
    • High: 11% (2011) and 3% (2014)
  • The proportion that actually enrolled
    • 35% (2011), 19% (2014)

• NOTE: The low and high thresholds are somewhat arbitrary
1. Order the subjects by predicted probability of enrollment (high to low).
2. We expect high p-hats to be enrolled and low-phats to not be enrolled.
3. Move the threshold for who is declared ‘enrolled’ from p-hat=1 down to p-hat=0. We should do pretty well at first, then it gets worse as we eventually declare everyone ‘enrolled’.
4. An ROC with a steep ascent from left to right is desirable.
ROC curve (2014 data)

5. A method that randomly chose whether a student was enrolled would be expected to have a diagonal ROC curve.

6. In practice, you’ll choose a threshold (p-hat) for deciding on predicted enrollment (yes/not), usually p=0.5 and that will coincide with one specific false positive and true positive rate (similar to specificity and sensitivity).