Retention Risk Prediction
Pilot Project

Office of Institutional Research and Analysis
October 2009
Office of Institutional Research and Analysis (OIRA)
- Involved in coordinating government and external reporting
- Engages in multilateral institutional analysis projects
  - e.g. with COU, MTCU, HEQCO

Higher Education Quality Council of Ontario (HEQCO)
- Interested in supporting work on alternatives to “raw” retention rates
- Opportunity for York University

HEQCO Multi-institution Retention Project: currently ramping up
Pilot Project In A Nutshell

- Explore use of newer analysis techniques
  - Estimating student-level retention risks
  - Predicting stop-out events at student level

- Ultimate goals
  - Retention forecasting tool
  - Tool to help move multilateral discussions towards a consideration of retention risks and activities

- Related benefits
  - e.g. tool for use in retention practice evaluation?
Response variable: (time-to-event, stop-out indicator)
- Time-to-event measured in academic years
- Non-stop-outs also included in data
  - Graduated without ever stopping-out
  - Still enrolled at end of study period
- Only first event per student considered

Predictor variables
- Admissions, biographical, financial and academic

Data for 83,593 undergraduate students
- New to York University, entering 1st year, direct entry programs
- Study period: fall 1996 through fall 2006
- Both full-time and part-time students
Predictor variables eventually used in model

- initial year: OAC grade group (60-69.9%, 70-74.9%, 75-80.9%, 80-89.9%, 90-100%)
- initial year: full-time status (at least 80% of full course load)
- initial year: full-time status (at least 60% of full course load)
- initial year: age (years)
- initial year: living in residence / commuting status
- initial year: domestic / international student status
- initial year: admissions adjudication type (101 or 105)
- current year: previous stop-out flag
- current year: academic year level repeater
- current year: domestic / international student status
- current year: domestic / international status switch flag (1 if any switch since initial year)
- current year: living in residence / commuting status
- current year: honours / general curriculum status
- current year: cumulative 1st major count (initially 1, increments with every switch)
- current year: cumulative 2nd major count (initially 1, increments with every switch)
- current year: home faculty
- current year: faculty switch flag (0 if no switch since initial year, otherwise 1)
- current year: degree objective
- current year: degree objective switch flag (0 if no switch since initial year, otherwise 1)
- current year: full-time status (at least 80% of full course load)
- current year: full-time status (at least 60% of full course load)
- current year: fiscal full time equivalent value
Data

- Stopping-out:
  - 34% of those entering as full-time students (≥80% full course load)
  - 60% of those entering as part-time students
Data

Distribution of stop-outs by time to stop-out for students entering between Fall 1996 and Fall 2000
Technique:
- “Random survival forests” algorithm
  - Extension of tree-based methods to time-to-event (survival) data
  - Generates hundreds of trees to create a “forest” of predictive models
    - Each tree is a weak model on its own
    - Combining output of all models provides very good prediction performance

Advantages:
- Easy to implement in prognostic setting
- Prediction performance compares favourably with other techniques
- Handles many predictor variables, interaction effects, collinearities, nonlinearities
- Provides forecasts at the level of individual students

Disadvantage:
- Black box (dark grey?)

Output:
- Survival probability function prediction for each student
Five training data sets
- Each 10,000 records
- Drawn randomly, without replacement, from initial 83,593
  • A student might occur in more than one training set, but only once in any single training set

RSF algorithm applied to each training set, and output compared

Survival probability estimates from one training set analyzed further:
- Compared (visually) with known times-to-event
- Input to a binary classification tree algorithm
  • Output: threshold survival probability values
  • Estimates below threshold interpreted as predicting stop-outs

One test data set
- 10,000 records
- Drawn randomly, without replacement, from initial 83,593 records minus those occurring in training set

Survival probability function estimates generated for test data
- Compared to threshold values, stop-outs predicted
Results: Retention Risk Estimation

- **Stop-out status**
  - 1 = 1st year stop-out
  - 0 = continue past 1st year

- **Colour**
  - Red = stop-outs (any year)
  - Blue = non stop-outs
Results: Retention Risk Estimation

- Comparing estimated risks against actual rates
  - Actual rates localized via moving window

- Survival probability (retention risk) estimates differ substantially from actual rates
  - Particularly for those most at risk of stopping-out

- But... broadly speaking, estimates are well behaved
Results: Stop-out Event Prediction

- **Stop-out status**
  - 1 = stop-out at any point
  - 0 = no stop-out

- **Colour**
  - Red = 1\(^{\text{st}}\) year stop-out
  - Green = 2\(^{\text{nd}}\) year stop-out
  - Yellow = >2\(^{\text{nd}}\) year stop-out

- **Binary classification threshold**
  - < 0.67 = predicted stop-out
Results: Stop-out Event Prediction

- Actual stop-outs correctly predicted
  - Overall: 52.5%
  - 1st year stop-outs: 90.3%
  - Upper year stop-outs: 24.6%
    - 35.3% of 2nd year stop-outs

- False positives (efficiency)
  - 17.6% of predicted stop-outs
Next Steps

- Further testing of the approach
  - Update the data set and models
  - Predict stop-out status for current students

- Options for further data development
  - Finer temporal resolution
  - “Initial conditions” data
    - Goals, commitments, preparedness
    - Social and cultural capital
    - Emotional intelligence
  - “In-stream” data
    - Preliminary grades in selected courses
    - Integration and engagement
    - Intentions to persist

- HEQCO Multi-institution Retention Project
  - Mixed methods approach

- Opportunities here at York?
  - Populations-at-risk
  - Knowledge and data

- Objective: Accurate retention risk estimates