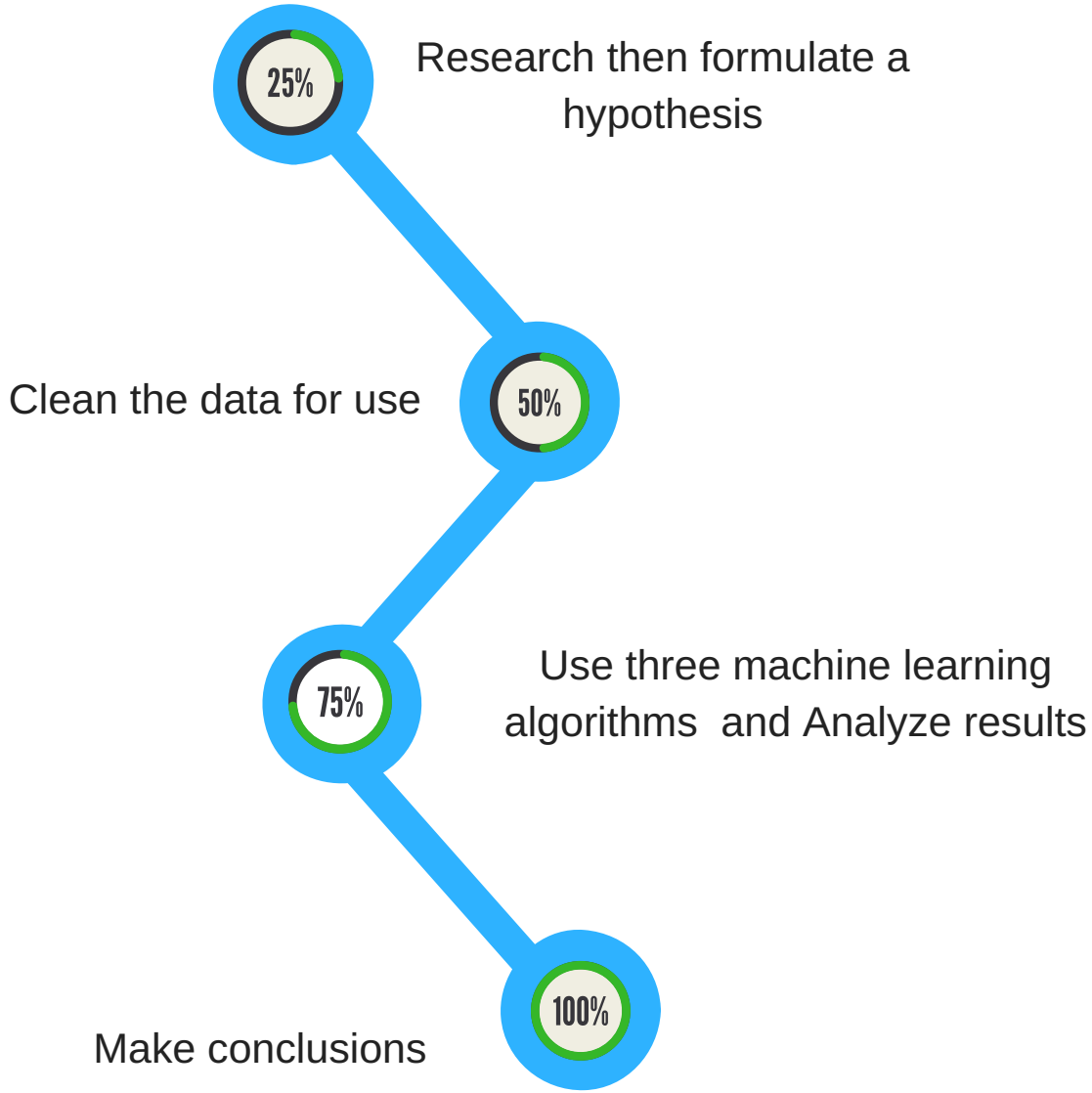
STUDENT ATTRITION

Tereza Shterenberg
Ryan Kallicharran

July 28, 2016

Method


25%
Research then formulate a hypothesis
50%
Clean the data for use
75%
Use three machine learning algorithms and Analyze results
100%
Make conclusions

Hypothesis & Goal

Student's age, GPA, and admission score predict students attrition

Our Goal:

- Build a prediction model for student attrition

Why age, GPA and Admission Scores?

- Paper 1: Student's age group have significant difference in reason for attrition
- Paper 2: Low course grades drive students away
- Paper 3: Good school admission scores lead to success in program

Data Filtering

Originally : 969,104 ins and 234 attributes

Collapsed students' classes: approx. 267,243 ins

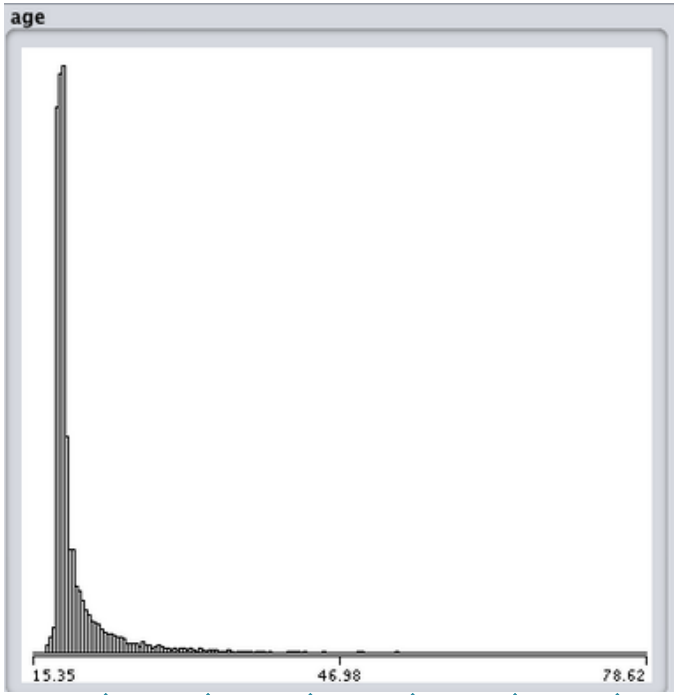
Reduced to 8 attributes:

studentid
age
start_dt_x
admission_sc
grade_num_x
graduated_2005_2015
semester
admissiontypedesc

* mean imputing technique

Stats

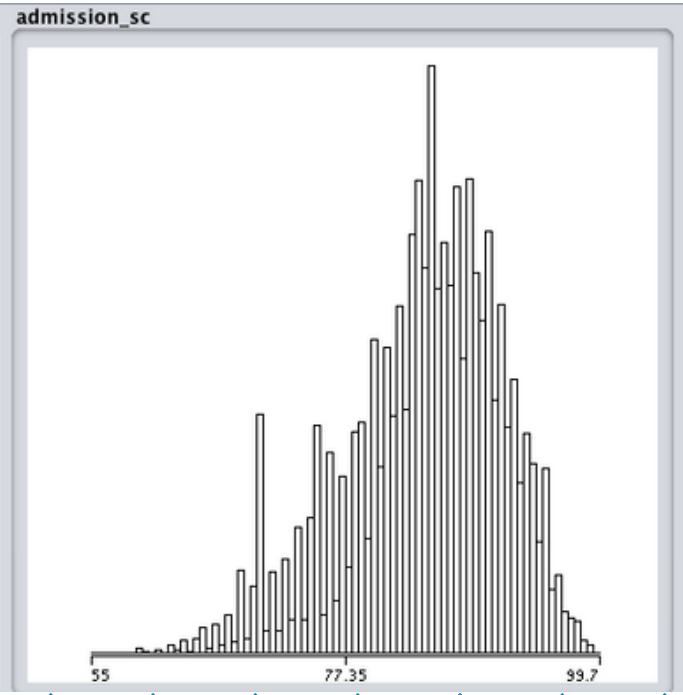
Age



Name: age
Missing: 2 (0%) Distinct: 3033 Type: Numeric
Unique: 142 (0%)

Statistic	Value
Minimum	15.35
Maximum	113.66
Mean	21.492
StdDev	5.689

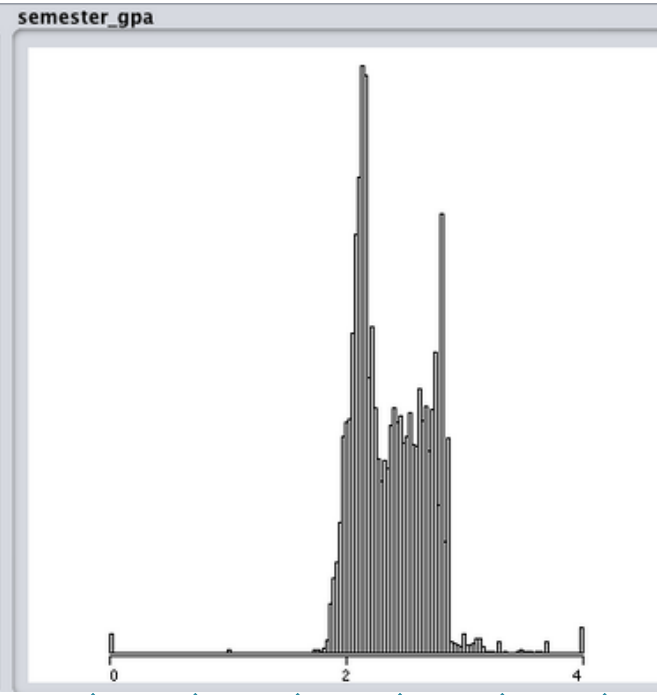
Admission score



Name: admission_sc
Missing: 0 (0%) Distinct: 229 Type: Numeric
Unique: 7 (0%)

Statistic	Value
Minimum	55
Maximum	99
Mean	81.181
StdDev	3.45

GPA



Name: grade_num_x
Missing: 0 (0%) Distinct: 487 Type: Numeric
Unique: 63 (0%)

Statistic	Value
Minimum	0
Maximum	4
Mean	3.007
StdDev	0.943

Algorithms: Logistic Regression

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      200818               75.1443 %
Incorrectly Classified Instances     66425               24.8557 %
Kappa statistic                          0.4985
Mean absolute error                      0.3089
Root mean squared error                  0.4137
Relative absolute error                 62.1353 %
Root relative squared error             82.9722 %
Total Number of Instances           267243

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.788    0.292    0.759      0.788    0.773      0.499    0.825     0.821     yes
                 0.708    0.212    0.742      0.708    0.725      0.499    0.825     0.819     no
Weighted Avg.    0.751    0.255    0.751      0.751    0.751      0.499    0.825     0.820

=== Confusion Matrix ===

      a      b   <-- classified as
 113269  30389 |      a = yes
  36036  87549 |      b = no
```

Algorithms: Sequential Minimal Optimization

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      201841               75.5271 %
Incorrectly Classified Instances     65402               24.4729 %
Kappa statistic                          0.5033
Mean absolute error                      0.2447
Root mean squared error                  0.4947
Relative absolute error                 49.2234 %
Root relative squared error             99.2204 %
Total Number of Instances           267243

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.828    0.329    0.745      0.828    0.784      0.507    0.749     0.709     yes
                 0.671    0.172    0.770      0.671    0.717      0.507    0.749     0.669     no
Weighted Avg.    0.755    0.257    0.757      0.755    0.753      0.507    0.749     0.691

=== Confusion Matrix ===

      a      b   <-- classified as
 118958  24700 |      a = yes
  40702  82883 |      b = no
```

Algorithms: Bayes Network

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      202509               75.7771 %
Incorrectly Classified Instances     64734               24.2229 %
Kappa statistic                          0.5056
Mean absolute error                      0.3076
Root mean squared error                  0.4036
Relative absolute error                 61.8752 %
Root relative squared error             80.956  %
Total Number of Instances           267243

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.864    0.366    0.733      0.864    0.793      0.516    0.836     0.836     yes
                 0.634    0.136    0.801      0.634    0.708      0.516    0.836     0.840     no
Weighted Avg.    0.758    0.260    0.764      0.758    0.754      0.516    0.836     0.838

=== Confusion Matrix ===

      a      b   <-- classified as
 124157  19501 |      a = yes
  45233  78352 |      b = no
```

Summary

Logistic

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.79	0.29	0.76	0.79	0.77	0.50	0.83	0.82	yes
0.71	0.21	0.74	0.71	0.72	0.50	0.83	0.82	no

SMO

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.83	0.33	0.75	0.83	0.78	0.51	0.75	0.71	yes
0.67	0.17	0.77	0.67	0.72	0.51	0.75	0.67	no

BayesNet

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.86	0.37	0.73	0.86	0.79	0.52	0.84	0.84	yes
0.63	0.14	0.80	0.63	0.71	0.52	0.84	0.84	no

Takeaways

- All 3 methods provided 75% accuracy confirming the proposed hypothesis

- GPA, Admission score, age of a student are factors contributing to their attrition

Future work

- Test this prediction model with other college data

- Try to see if we can find another attribute that can contribute to attrition prediction model

Thank you!

Github: https://github.com/ryankall/capstoneProject

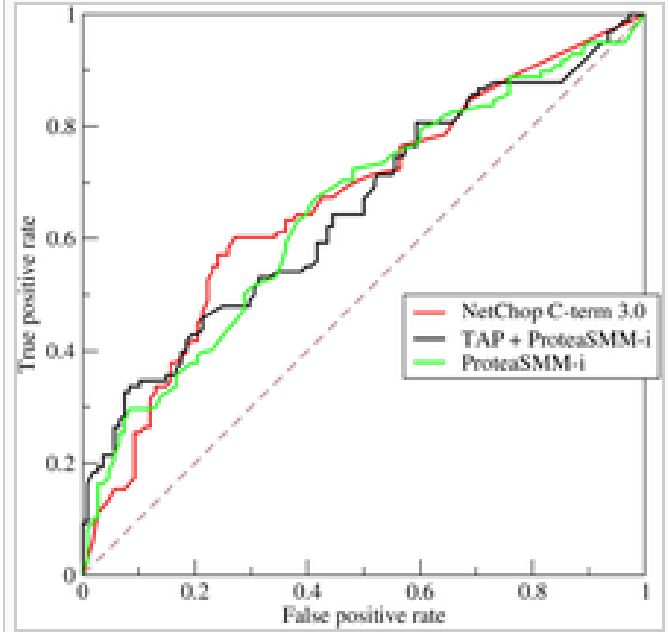
Additional Info

$$accuracy = \frac{\sum_{i=1\ldots N}\left(1-(target_i - threshold(f(\vec{x}_i)))\right)^2}{N}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

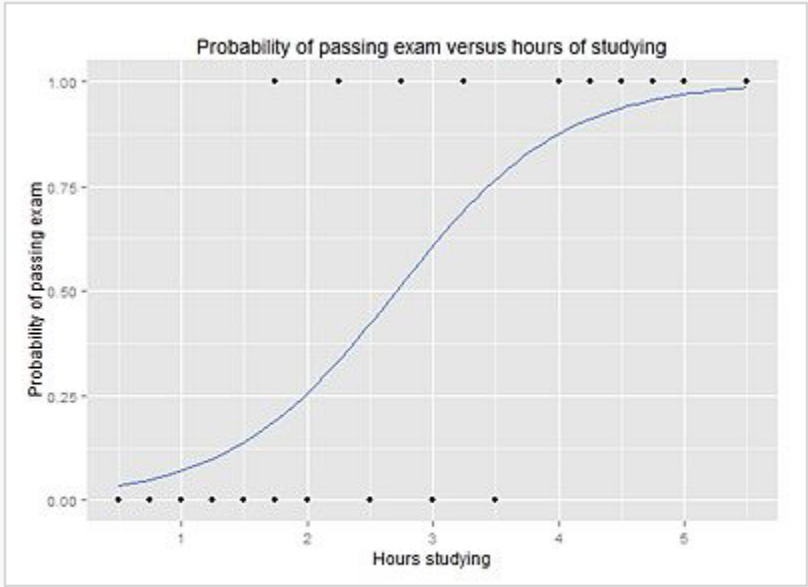
$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

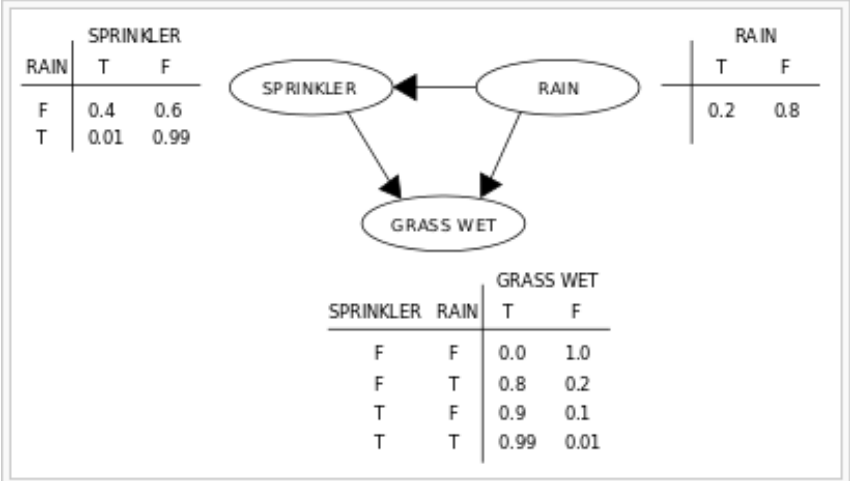


ROC curve of three predictors of peptide cleaving in the proteasome.

Models



Graph of a logistic regression curve showing probability of passing an exam versus hours studying



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

A simple Bayesian network with conditional probability tables

1. Find a Lagrange multiplier α_1 that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.
2. Pick a second multiplier α_2 and optimize the pair (α_1, α_2).
3. Repeat steps 1 and 2 until convergence.

Reference

Paper 1: Understanding Student Attrition in the Six Greater Toronto Area (GTA) Colleges by Tet S. Lopez-Rabson (Seneca College) et al.

Paper 2: Why They Leave: Understanding Student Attrition from Engineering Majors by D. Raj Raman & Brandi N. Geisinger

Paper 3: Student Attrition: Consequences, Contributing Factors, and Remedies by Ascend Learning, LLC