Data Mining for Student Success and Perseverance

Sameer Bhatnagar
Jonathan Guillemette
Micheal Dugdale
Sahir Bhatnagar
Nathaniel Lasry

2017-05-23

Contents

Chapter 1

Preface

This report summarizes the work done by our team on using college registration records at three diffrenet anglophone CEGEPS in Montreal in order to find predictors of attrition.

Chapter 2

Introduction

Chapter 3

Literature

The most important relevant work for this project is (Jorgensen et al., 2009).

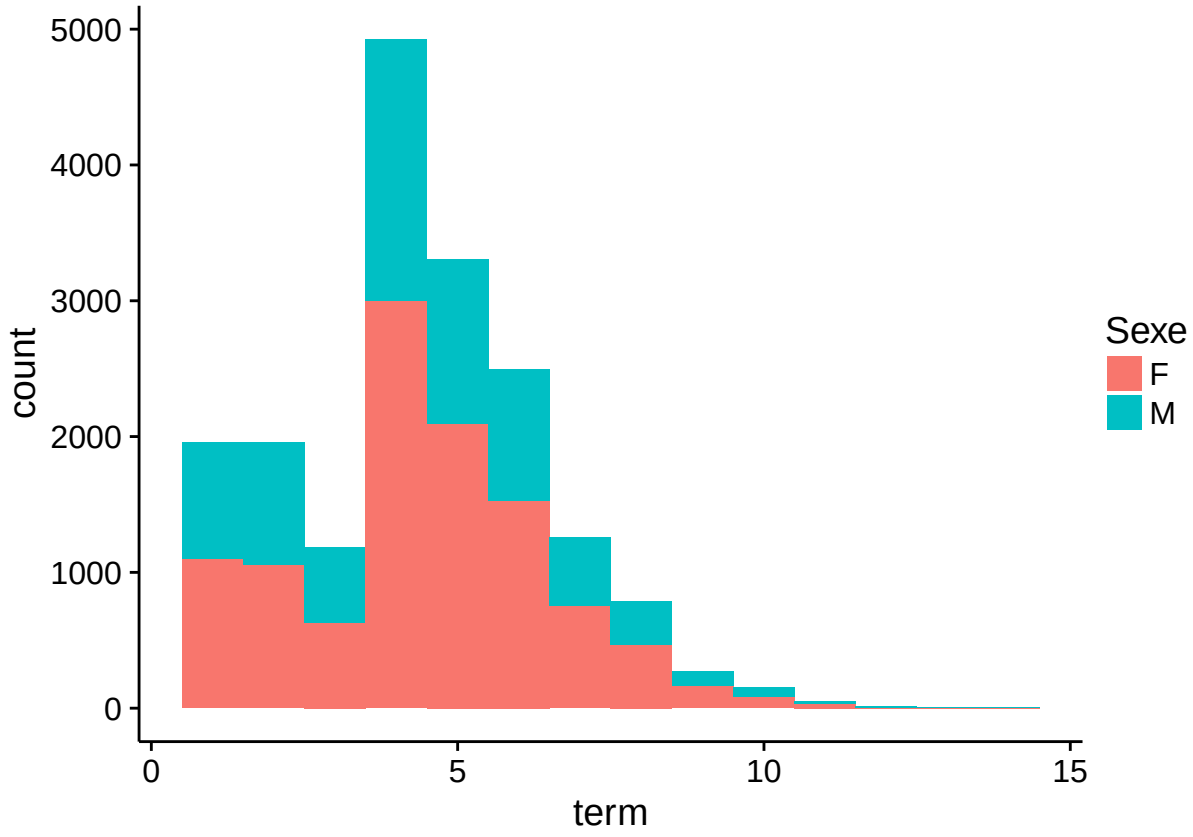
Chapter 4

Descriptive Statistics

Here in we will describe - the data set - the methods by which we label students at risk - the distributions of at-risk students by - demographic indicators - registration record indicators

For example, how many what is the distribution of terms spent by a student at Dawson College, disaggregated by gender?

```
library(ggplot2)
library(cowplot)
library(data.table)
load('bin/data/labelled_students.Rdata')
past_students=students_last_session[status!='current']
ggplot(data = past_students,aes(term,fill=Sexe))+geom_histogram(binwidth = 1)
```



Chapter 5

Methods Centered on Determining Predictive Factors

This chapter will be focused on methods which have a sound probabilistic framework, and allow for inference into the statistical importance of predictive factors.

- Logistic Regression
- Mixed Effects Models

Chapter 6

Methods centered on predicting at-risk students

This chapter will explore the use of machine learning algorithms whose primary focus is prediction. This comes at the expense of model interpretatbility, and this tradeoff will be discussed herein as well.

- Decision Trees and Random Forests
- Neural Networks

6.1 Example one

6.2 Example two

Chapter 7

Comparisons

This chapter will compare the effectiveness of the methods developped in the previous two chapters, and compare them to more basic approaches to identifying students at-risk.

For example, we know that some CEGEPs have implemented a policy whereby they identify students as being at risk based on their mid-term assessments: if the student receives a certain number of "at-risk" or "failing" results, they are automatically sent an email referring them to academic support services.

Based on this, we can ask the following research questions : - how effective is this approach at identifying students who drop-out? - how does this approach compare to our models from the previous chapters?

We begin with a basic logistic regression with demographic variables, and as well as the number of each type of results of mid-term assessment, for students in their last term at the college. With these predictors, we try to predict if students are about to graduate, or simply not register again.

	Estimate	Std. Error	z value	Pr(>|z|)
num_pass	0.5877	0.02321	25.32	1.747e-141
num_at_risk	1.455	0.03664	39.7	0
num_failing	2.193	0.04653	47.15	0
num_courses	-0.7442	0.02373	-31.36	6.315e-216
SexeM	0.2111	0.0386	5.469	4.537e-08
birth_placeQuebec	-0.5914	0.04815	-12.28	1.108e-34
LangueMaternelleAU	-0.193	0.05173	-3.731	0.0001907
LangueMaternelleFR	0.3504	0.04914	7.13	1.006e-12
(Intercept)	0.5654	0.0694	8.147	3.741e-16

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	24452 on 18370 degrees of freedom
Residual deviance:	17188 on 18362 degrees of freedom

Chapter 8

Final Words

We have finished a nice book.

Bibliography

Jorgensen, S., Fichten, C., and Havel, A. (2009). *Predicting the At Risk Status of College Students: Males and Students with Disabilities. Final Report Presented to PAREA, Spring 2009.*