Data Mining for Student Success and Perseverance

Sameer Bhatnagar
Jonathan Guillemette
Micheal Dugdale
Sahir Bhatnagar
Nathaniel Lasry

2017-12-02

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

2.1 Context

This report outlines the results from a three year intercollegiate research project funded by the**PAREA** agency (*Programmme d'Aide à la Recherche en Enseignment et Apprentissage*) from the *Ministère de l'Education* of the provincial Government of Quebec.

In the province of Quebec, students finish their secondary education at what is the equivalent of Grade 11 in other parts of North America. Students are then able to attend **CEGEP** (*Collège d'enseignment général et professionel*) for either

- two years, as part of pre-university program, e.g Science, Social Science, Liberal Arts
- three years, as part of technical program, meant specifically to lead directly to the job market, e.g. Nursing, Civil Engineering Technology, Diagnotsic Imaging Technology

There are 48 CEGEPs in the Quebec network, and public or private, they all fall under the purview of the *Ministère de l'Education et Enseignment Superieur.* Over the past twenty years, there has been significant work(Jorgensen et al., 2003, 2005, 2009a; Rivière, 1995; Shaienks et al., 2008) and media (Breton, 2016; Dion-Viens, 2015; Duchaine, 2017) on the topic of student attrition in CEGEP. The scholarly work done has often focused on determing predictors of attrition through surveys, or focused on specific vulnerable sub populations. The media has often reported on government figures, which rely on data that looks at information at a very coarse level of granularity (of students graduated from high schoolhow many obtain diplomas from CEGEP)

2.2 Objectives

Almost all of the CEGEP's use the same database system, known as **CLARA** (developped by the company Skytech) in order to manage the data related to student admission, registration and graduation. Our research team's main objective is to leverage this uniformity of how data is automatically generated and stored, in order to determine if, in this wealth of data, there might be predictors of student attrition. This effort stands apart from previous work and reports in that:

- the data analyzed is much finer-grained: the unit of analysis is down to the semester registration records for each student
- we look at the general population of students
- to our knowledge, this is the first ever such study to span multiple CEGEPs, which we hope adds a greater relibility and validity to our findings.

This project has two specific objectives:

1) find predictors of students dropping out, whether they be demographic, or based in academic performance, on a term by term basis.
2) evaluate methods by which students can be automatically flagged as being at risk of dropping out, without so much a focus on understanding why, but for the purpose of "general offers of support" or further investigation

This report is structured to reflect these two objectives. Namely, we begin with - a standard review of previous related work, - a section on descriptive statistics which gives an overview of the dataset.

We then move on to our methods and models over two chapters: - the first addresses objective 1, outlining our efforts to find explanatory predictors of student attrition in classical statistical models. - The following chapter addresses objective 2, describing modern methods from the field of machine learning, which, at the expense of model interpretability, are fit to provide maximum predictive accuracy in identifying students at-risk.

The report then concludes with - comparisons of these methods with each other, and to current intervention frameworks in place at participating colleges - an auxilliary chapter looking at students from the division of Continuing Education - a concluding chapter, with reccomendations for future directions

Chapter 3

Literature

Under Construction

3.1 Learning Analytics and Educational Data Mining

Data Mining and Statistical Machine Learning methodologies can be roughly subdivided into *Predictive* and *Explanatory* modelling(Shmueli, 2010). These two paradigms manifest themselves as well in the domain on *Educational* data mining(Liu and Koedinger, 2017).

3.2 Modeling Student Attrition in Higher Education

Some of the earliest theoretical work on persistence in higher education (Tinto, 1975) models dropouts in higher education as a social process, resulting from poor student integration within their institution; that the probability of a student staying in school is shaped by the level of matching between the student's motivation and academic ability, with the social and academic culture of the institution (Cabrera et al., 1993). This "Student Integration Model" goes as far as claiming that attrition can serve as a "barometer of the social and intellectual health of college life"(Tinto, 1987).

A competing theoretical model, known generally as the "Student Attrition Model", borrows from work on "turnover"" in industry and organizations, wherein student satisfaction, and perceptions of value in their own work, directly influence their intention and decision to dropout (Bean, 1980),(Bean, 1983). While the "Student Integration Model" posits more importance on variables *internal to the institution*, the "Student Attrition Model" assigns what is observed at the college as an *outcome*, shaped by external factors, such as parental and social support, or presence of opportunities elsewhere. In Tinto's work, college grades are seen as a result of a student's integration into the institution. In contrast, (Bean, 1985) models these same grades as a *cause* of "dropout syndrom".

Longitudinal models of student attrition are important if we are to take into account the evolving nature of a student's ultimate decision to stop school(Tinto, 1975), and how the timing of interventions can have an impact on the outcome(Ishitani and DesJardins, 2002). (Simons, 2011) Research on attrition from institutions of higher education (Singell and Waddell, 2010)

3.3 Modeling success and attrition in CEGEP

The most important relevant work for this project is (Jorgensen et al., 2009b), wherein the aim was to determine what factors were most important in attrition at an English Montreal CEGEP. Querying the academic records of over 40,000 students, the authors found that by-and-large, the most important predictor of attrition were the grades attained in high school. The effect of gender was more pronounced for weaker students, with male drop-out rates up to 10% higher than females, when their secondary overall average was below 80%.

While this work did include survey results, the methodological approach followed (DesJardins et al., 1999), wherein the core analysis was done on large data sets that avoid expensive surveying and that are subject to low response rates, and instead used only variables that are already stored in instituional academic databases. The objective of our project is to continue this work, but with several important extensions. This will include how we: - examine attrition in the broader context of multiple institutions in the CEGEP Network, - investigate the application of a more varied set of statistical modelling and machine learning techniques, - leverage the affordances of longitudinal data: namely, what added insight can be offered by modelling attrition based on the repeated observations of student academic performance at every semester? - study the impact of refined definitions of "attrition" on this research, as suggested by (Tinto, 1975): how can voluntary withdrawals be seperated from those dropouts resulting from academic failure? How are temporary absences and program transfers taken into account?

Chapter 4

Descriptive Statistics

** Under Construction **

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

4.1 Demographics across the colleges and major programs

4.1.1 Dawson

	AN	AU	FR	Sum
F	0.33	0.15	0.12	0.6
M	0.23	0.1	0.07	0.4
Sum	0.56	0.25	0.19	1

4.1.2 John Abbott College

	AN	AU	FR	Sum
F	0.31	0.09	0.12	0.52
M	0.3	0.07	0.11	0.48
Sum	0.61	0.16	0.23	1

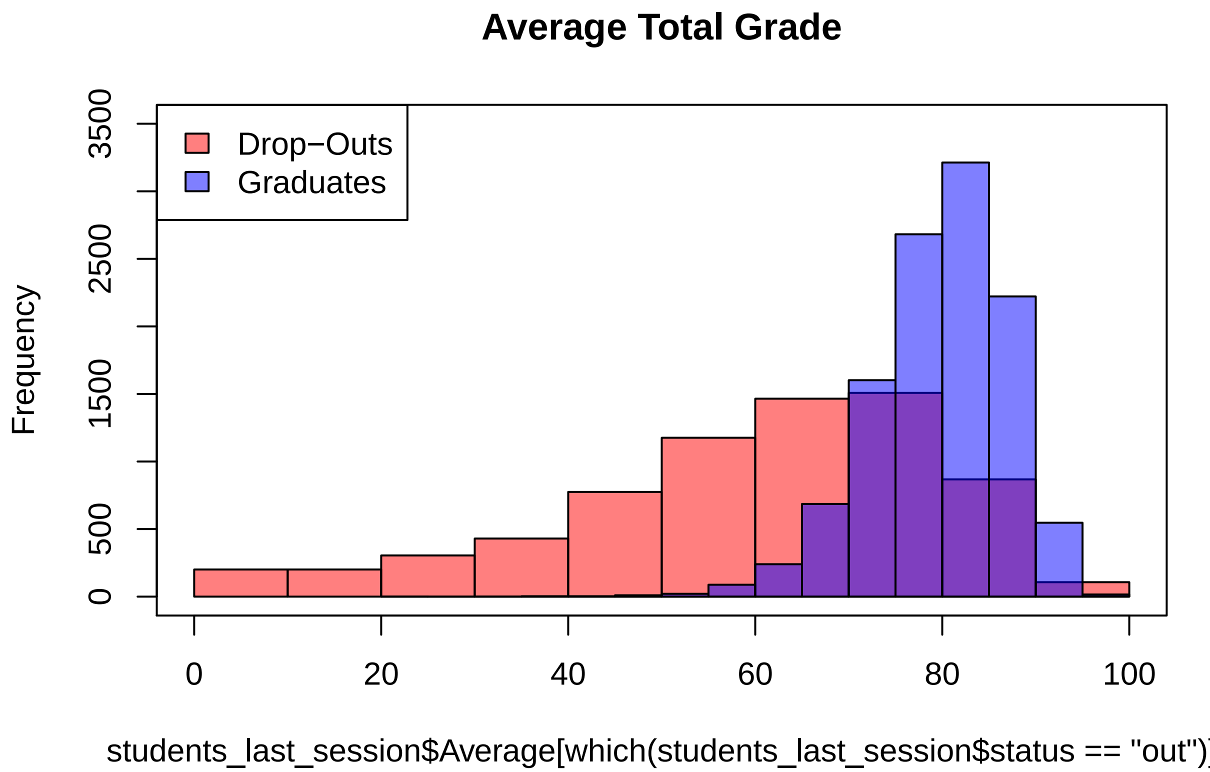
4.1.3 Vanier

	AN	AU	FR	Sum
F	0.26	0.17	0.1	0.53
M	0.24	0.16	0.08	0.48
Sum	0.5	0.33	0.18	1.01

4.2 Academic Performance in Semesters leading to drop-out

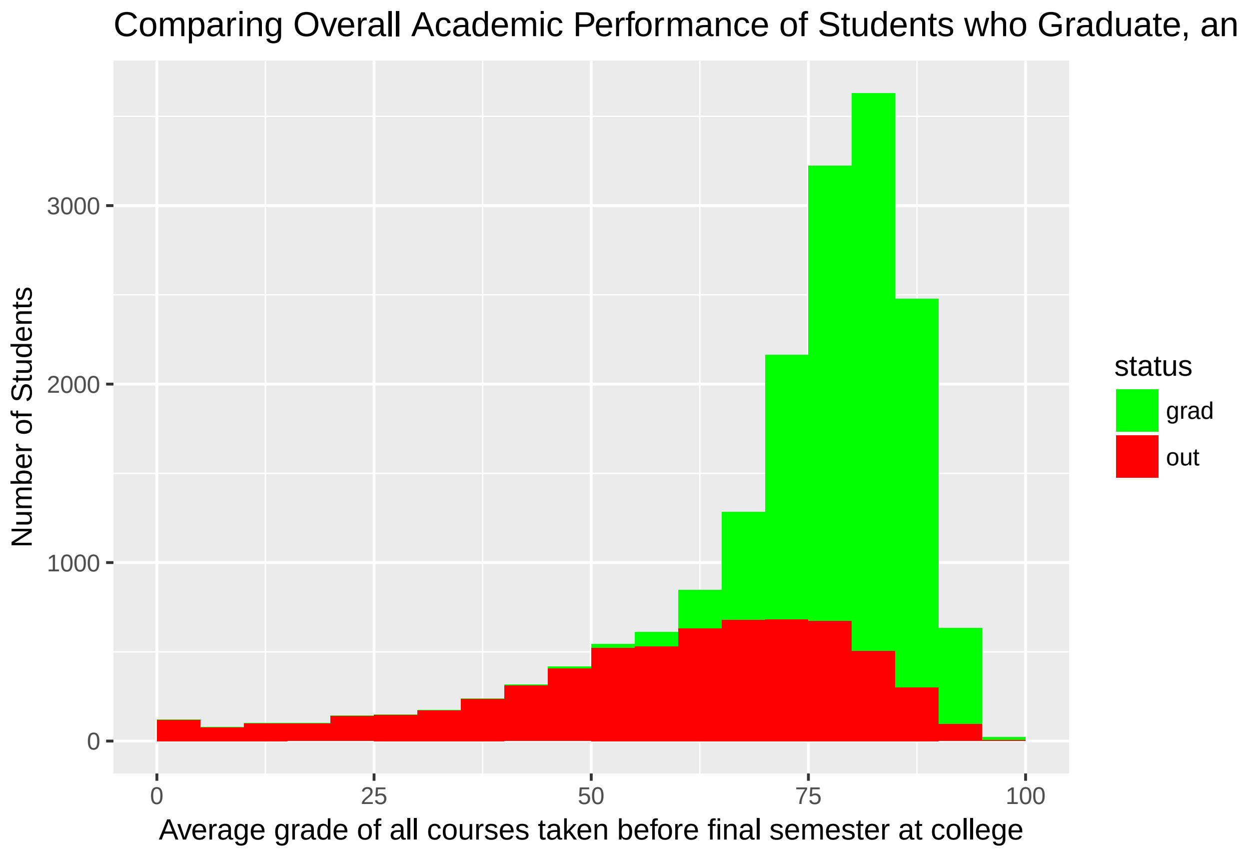
4.2.1 Grades

Do students who drop out do so because of poor grades? What fraction of students are counted year after year as drop-outs and labeled as problems to be solved by the system while being exemplary students in terms of academic performance. Armed with this dataset, we can get the answer to that question. Let us begin by looking at the average grades of students who eventually dropped out compared to grades of students who haven't. The following set of graphs will look at that comparison for 3 different semesters: the semester in which they dropped out (or graduated), the semester before that and the one before that. Let us see what the data says.



From the average total grade between the drop outs and the graduates, we can clearly see that the distributions are significantly different, but what is surprising is that NaN% of students have an average grade above 75% and that more than NaN% of the students who dropped out had an average grade above 60%. In other words, most students who drop out have passing averages. One hypothesis for this effect is that students who end up dropping out start with good semesters and have their performance decline closer to the final semester. Let us verify this hypothesis.

Let us now look at the performance of drop outs vs graduates on a semester by semester basis. Let us start by looking at the average grades of students in their last semester during which they are either graduating or

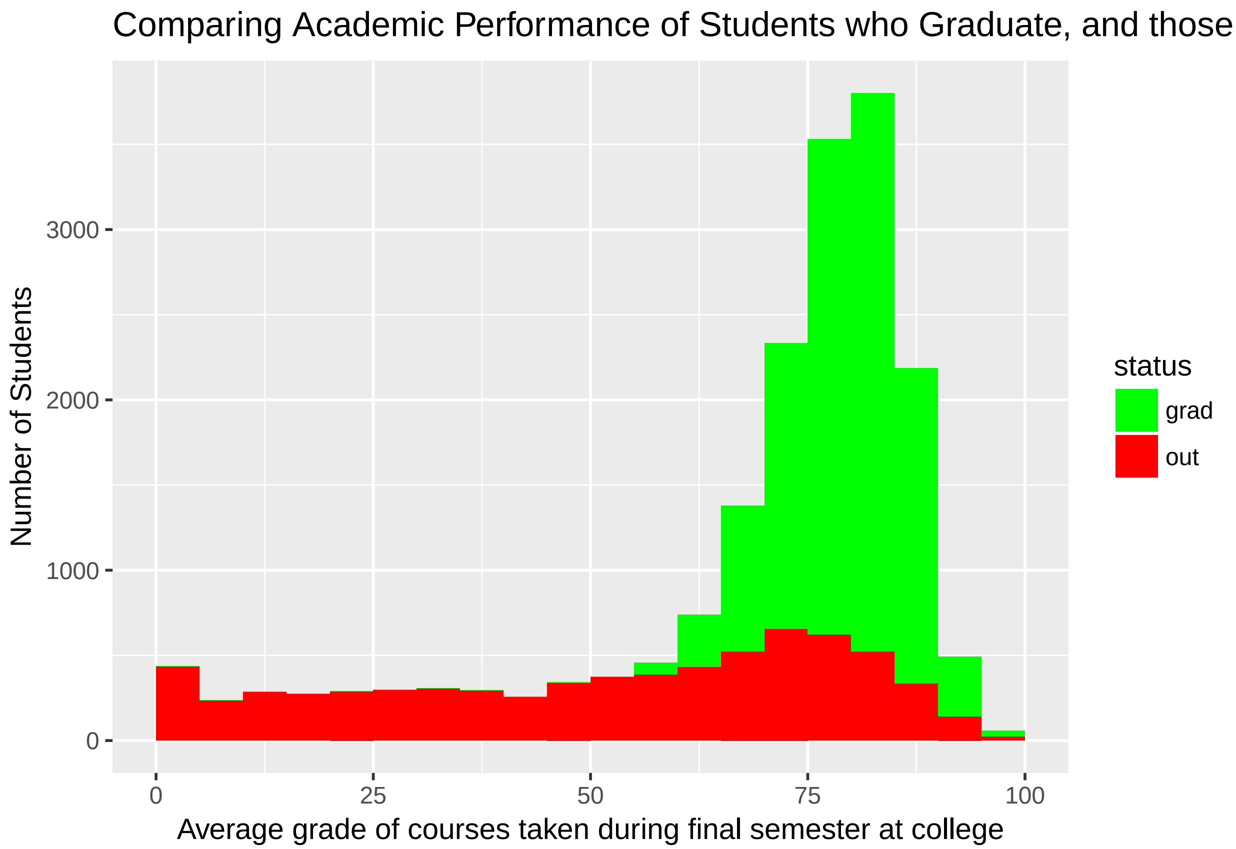


dropping-out.

```
##
##  Welch Two Sample t-test
##
## data:  Average_all_courses by status
## t = 78.251, df = 7349.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  20.36923 21.41600
## sample estimates:
## mean in group grad  mean in group out
##           79.97370           59.08109
```

From the average total grade between the drop outs and the graduates, we can clearly see that the distributions are significantly different, but what is surprising is that 22% of students have an average grade above 75% and that more than 52% of the students who dropped out had an average grade above 60%. In other words, most students who drop out have passing averages. One hypothesis for this effect is that students who end up dropping out start with good semesters and have their performance decline closer to the final semester. Let us verify this hypothesis.

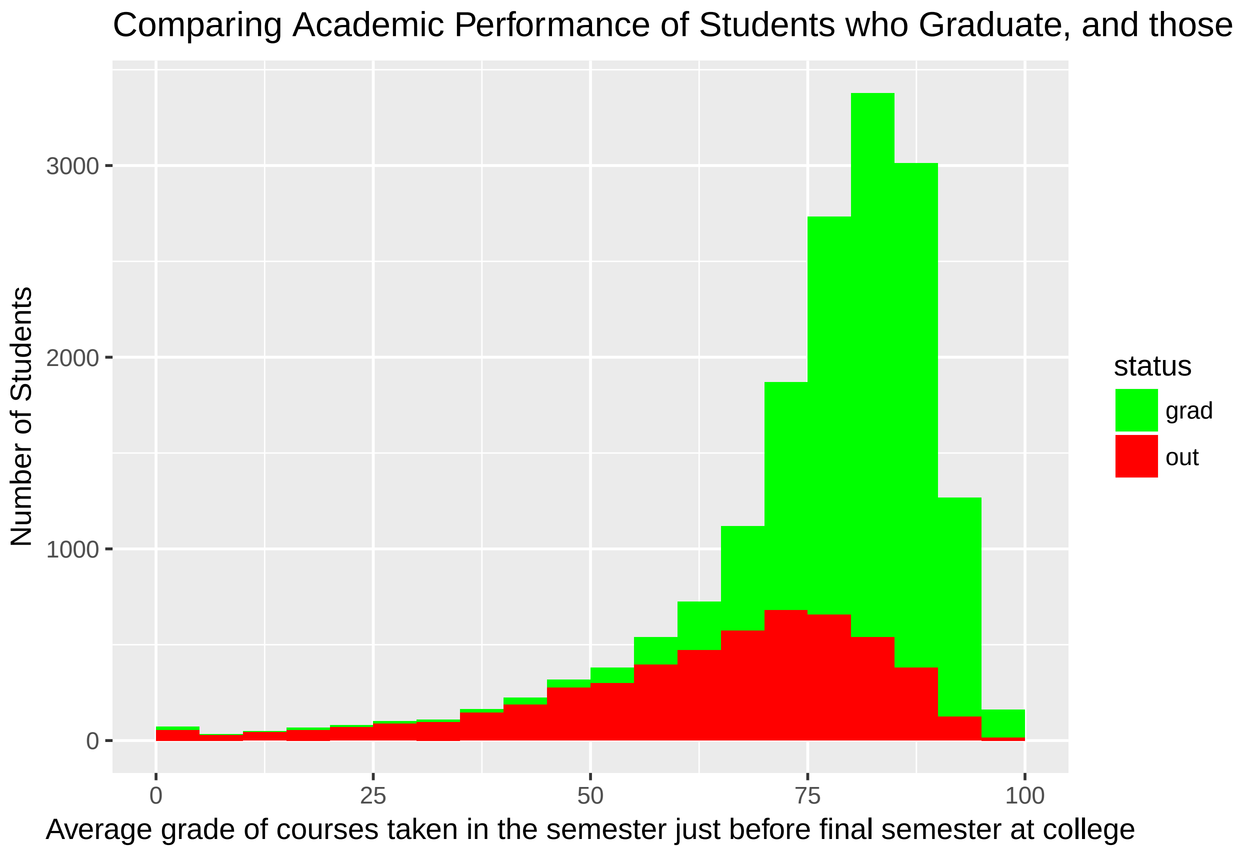
Let us now look at the performance of drop outs vs graduates on a semester by semester basis. Let us start by looking at the average grades of students in their last semester during which they are either graduating or dropping-out.




```
##
##  Welch Two Sample t-test
##
## data:  avg_grade by status
## t = 84.649, df = 7619.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  27.00250 28.28279
## sample estimates:
## mean in group grad  mean in group out
##           79.17656           51.53391
```

In this data, we can clearly observe a stark difference between the graduates and the drop outs. First of all, note that 23% of students still have a term average over 75% in the semester in which they drop out. To push it further 15% of students who drop out have an average grade over 80%. The data clearly suggests that some of the drop outs aren't dropping out because of academic performance. Furthermore, the long tail of the data on the low end of performance suggests that some of the students stopped coming to class prior to the end of the semester resulting in very low grades that serve to drive their cegep average grades from the graph above even lower. 34% of students have a grade below 40 suggesting that they have indeed stopped coming to school some time during the semester. What if these students hadn't stopped coming, would their performance be similar to that of graduates?

Let us now turn our attention to the semester before the one where they graduate or drop out.



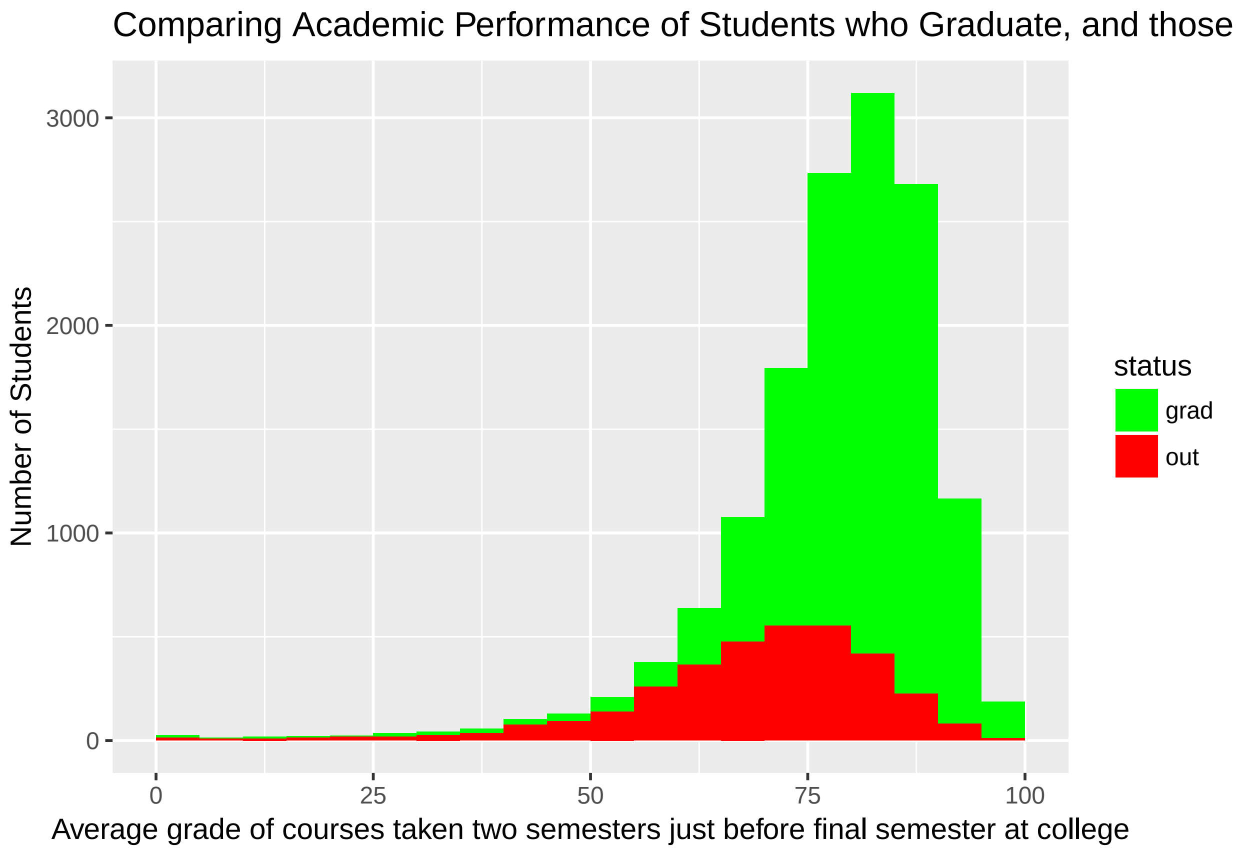
In this data, we can clearly observe a stark difference between the graduates and the drop outs. First of all, note that 22% of students still have a term average over 75% in the semester in which they drop out. To push it further 14% of students who drop out have an average grade over 80%. The data clearly suggests that some of the drop outs aren't dropping out because of academic performance. Furthermore, the long tail of the data on the low end of performance suggests that some of the students stopped coming to class prior to the end of the semester resulting in very low grades that serve to drive their cegep average grades from the graph above even lower. 34% of students have a grade below 40 suggesting that they have indeed stopped coming to school some time during the semester. What if these students hadn't stopped coming, would their performance be similar to that of graduates?

Let us now turn our attention to the semester before the one where they graduate or drop out.

```
##
##  Welch Two Sample t-test
##
## data:  avg_grade by status
## t = 57.702, df = 6604.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15.65690 16.75814
## sample estimates:
## mean in group grad  mean in group out
##           80.56981           64.36230
```

The data clearly shows that even if there are significant differences between the groups, the drop out student population is getting closer to the graduate population. A section of 33% is observed to have an average grade above 75%.

Finally, let us look at 2 semesters before they graduate or drop-out. We are again expecting the same trend.

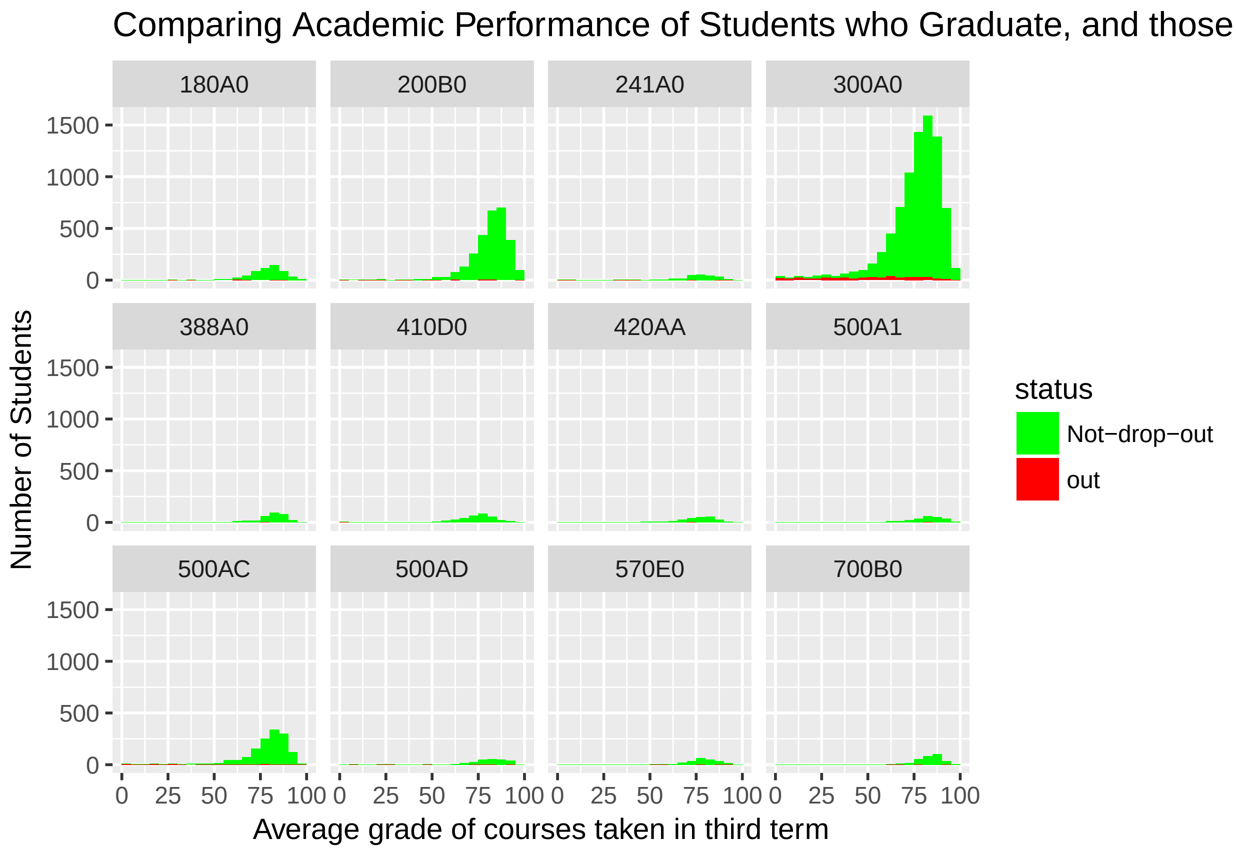



```
##
##  Welch Two Sample t-test
##
## data:  avg_grade by status
## t = 42.166, df = 4390.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  10.95430 12.02262
## sample estimates:
## mean in group grad  mean in group out
##           80.42574           68.93729
```

4.3 Conclusion

In conclusion, the data strongly suggests that there are approximately 20% of students who consistently have averages above 75%, but still drop out. Therefore, that section of the drop out population is a section that will always go undetected if only traditional academic failure metrics are used to assess which students are likely to drop out. Students who move out to the US or the rest of Canada after their second semester and those who drop out by lack of interest are two potential student types that will always drop out no matter what remedial solutions are offered for them.

One of the profile level Key Performance Indicators that colleges are supposed to specifically keep track of is third semester retention. In this light, we can somewhat flip the line of questionning above, and look to see if the distribution of grades in the third semester students looks different for those who will drop out, and those who will continue on.



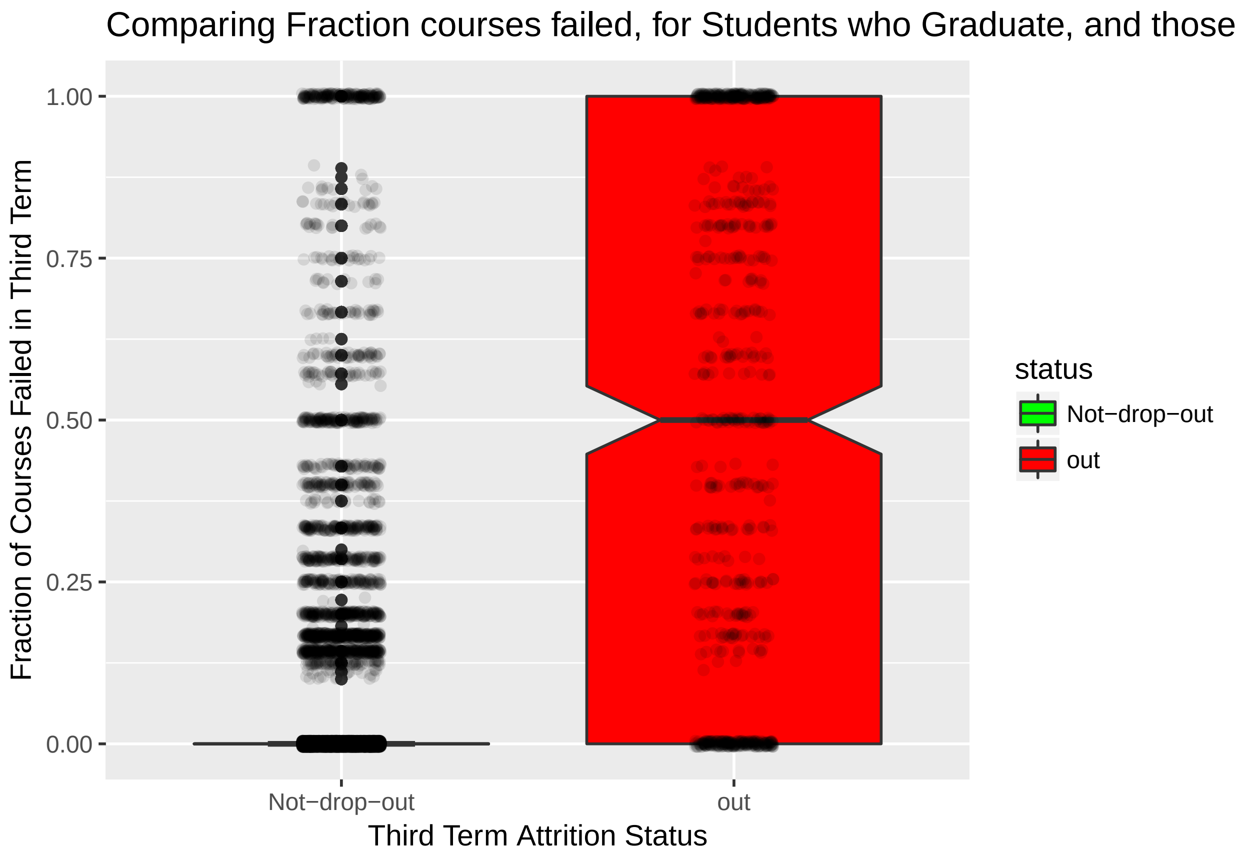
In conclusion, the data strongly suggests that there are approximately 20% of students who consistently have averages above 75%, but still drop out. Therefore, that section of the drop out population is a section that will always go undetected if only traditional academic failure metrics are used to assess which students are likely to drop out. Students who move out to the US or the rest of Canada after their second semester and those who drop out by lack of interest are two potential student types that will always drop out no matter what remedial solutions are offered for them.

4.3.1 Passed vs. Failed Courses

The line of questionning above can be repeated, but instead of looking at the average grade of courses taken in a term, we can instead look to see if the number of courses passed, or the proportion failed, might be different for students who drop out versus those who do not.

	0	1	2	3	4	5	6	7	8	10
Not-drop-out	11339	1320	447	216	116	66	34	8	2	0
out	263	113	93	80	134	109	59	36	15	1

What we remark in the table above is that the number of courses failed does not seem to differentiate students who will drop out after their third term. But perhaps, we should consider the fraction of courses that a student took in their third term, and failed?



For the above graphic, we calculate, for each student in their term, the fraction of courses that they took and *failed*, and we plot the distributions for the group of students we know who dropped out, and those we know stayed in the college for a fourth term. The red boxplot shows that,of the students who dropped out after their third term, 50% of them failed more than than half of their classes in that third term (the central notch represents a 95% confidence interval on the median). Conversely, the distribution on the left, squashed down at 0, shows that almost all students who stay on for a fourth term, pass all of their third term classes (The additional dots shown in this distribution are considered outliers to the distribution, meaning there are some students who failed all of their third term classes, and decided to stay for an additional fourth term).

Chapter 5

Methods Centered on Determining Predictive Factors

Under Construction

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

Under construction

Over the last twenty years, there has an increasing amount of work in the applied social sciences that explore the use of what (Breiman et al., 2001) refers to as “algorithmic modelling”“, as opposed to”data modelling“. He describes these as two cultures, the former being made up of mostly computer scientists, and the latter being made up statisticians (the methods therein are the ones explored in the previous chapter of this report). The key metric in such classical methods are **goodness of fit**, and such”explanatory“” modelling aims to find associative and causal relationships between predictors. Meanwhile, “algorithmic”“, or”predictive“” methods emphasize determining any function that maps input variables to output responses, with less regard for a probabilistic framework that allows for causation, focusing solely on emprirical precision(Shmueli, 2010). In the recently published **Handbok of Learning Analytics**(Lang et al., 2017), published by the Society of Learning Analytics Research, (Bergner, 2017) asserts that the researchers looking into educational data stand to gain from understanding the nuances of both methodologies, as previous work has shown the strengths and weakeness of either in this domain.

The previous chapter explored how classical statistical models can be built and used to determine what are the factors that influence dropout. This is useful for policy makers and admninstrators who want to dedicate resources in the most strategic places. However this chapter will explore models whose inner workings are less interpretable, but whose primary objective is prediction/identification of at-risk students. This is useful in the context where college administration has some blanket intervention that it would like to apply, and we just want to ensure that the students most in need are reached. Despite the less clear interpratability of factors in these *predictive* models (as compared to the *explanatory* models in the previous chapter), we will stil explore methods to “open up the black box”, and determine which features are most important in achieving both accrate and sensitive prediction.

6.1 Decision Trees and Random Forests

6.2 Neural Networks

Chapter 7

Comparisons

** Under Construction **

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

Conclusion

** Under Construction **

Bibliography

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education*, 12(2):155–187.

Bean, J. P. (1983). The application of a model of turnover in work organizations to the student attrition process. *The review of higher education*, 6(2):129–148.

Bean, J. P. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American educational research journal*, 22(1):35–64.

Bergner, Y. (2017). Measurement and its Uses in Learning Analytics. In Lang, C., Siemens, G., Wise, A. F., and Gaševic, D., editors, *The Handbook of Learning Analytics*, pages 34–48. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Breton, B. (2016). Décrocher du cégep et de l'université. *La Presse*.

Cabrera, A. F., Nora, A., and Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education*, 64(2):123–139.

DesJardins, S. L., Ahlburg, D. A., and McCall, B. P. (1999). An event history model of student departure. *Economics of education review*, 18(3):375–390.

Dion-Viens, D. (2015). Cégeps: à peine 31% des étudiants obtiennent un diplôme dans les délais.

Duchaine, G. (2017). Qu'est-ce qui cloche au cégep ? - La Presse+.

Ishitani, T. T. and DesJardins, S. L. (2002). A longitudinal investigation of dropout from college in the united states. *Journal of college student retention: research, theory & Practice*, 4(2):173–201.

Jorgensen, S., Ferraro, V., Fichten, C., and Havel, A. (2009a). Predicting college retention and dropout: Sex and disability. *Online Submission*.

Jorgensen, S., Fichten, C., and Havel, A. (2009b). Predicting the at risk status of college students: Males and students with disabilities. final report presented to parea, spring 2009. *Online Submission*.

Jorgensen, S., Fichten, C., Havel, A., Lamb, D., James, C., and Barile, M. (2003). *Students with Disabilities at Dawson College: Success and Outcomes. Final Report Presented to PAREA, Spring 2003.* ERIC.

Jorgensen, S., Fichten, C. S., Havel, A., Lamb, D., James, C., and Barile, M. (2005). Academic performance of college students with and without disabilities: An archival study. *Canadian Journal of Counselling*, 39(2):101.

Lang, C., Siemens, G., Wise, A., and Gasevic, D. (2017). *Handbook of learning analytics*. SOLAR.

Liu, R. and Koedinger, K. R. (2017). Going beyond better data prediction to create explanatory models of educational data. In Lang, C., Siemens, G., Wise, A. F., and Gaševic, D., editors, *The Handbook of Learning Analytics*, pages 69–76. Society for Learning Analytics Research (SoLAR), Alberta, Canada, 1 edition.

Rivière, B. (1995). Comprendre les décrocheurs afin de mieux les aider. *Pédagogie collégiale*, 9(2):11–15.

Shaienks, D., Gluszynski, T., and Bayard, J. (2008). *Les études postsecondaires, participation et décrochage: différences entre l'université, le collège et les autres types d'établissements postsecondaires.* Statistique Canada.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, 25(3):289–310.

Simons, J. M. (2011). *A national study of student early alert models at four-year institutions of higher education.* PhD thesis, ARKANSAS STATE UNIVERSITY.

Singell, L. D. and Waddell, G. R. (2010). Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education*, 51(6):546–572.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition.* ERIC.