Predicting University Students' Academic Success and Major using Random Forests

Cédric Beaulac Jeffrey S. Rosenthal

January 11, 2019

Abstract

In this article, a large data set containing every course taken by every undergraduate student in a major university in Canada over 10 years is analysed. Modern machine learning algorithms can use large data sets to build useful tools for the data provider, in this case, the university. In this article, two classifiers are constructed using random forests. To begin, the first two semesters of courses completed by a student are used to predict if they will obtain an undergraduate degree. Secondly, for the students that completed a program, their major is predicted using once again the first few courses they have registered to. A classification tree is an intuitive and powerful classifier and building a random forest of trees improves this classifier. Random forests also allow for reliable variable importance measurements. These measures explain what variables are useful to the classifiers and can be used to better understand what is statistically related to the students' situation. The results are two accurate classifiers and a variable importance analysis that provides useful information to university administrations.

Keywords : Higher Education, Student Retention, Academic Success, Machine Learning, Classification Tree, Random Forest, Variable Importance

1 Introduction

Being able to predict if a student is at risk of not completing its program is valuable for universities that would like to intervene and help those students move forward. Predicting the major that will be completed by students is also important in order to understand as soon as possible which program attracts more students and allocate resources accordingly. Since gathering data can be an expensive procedure, it would be useful being able to predict both of these things using data the university already possesses such as student records. Understanding which variables are useful in both of these predictions is important as it might help understand what drives student in taking specific classes.

Formally, these two prediction problems are classification ones. To solve these, a popular machine learning algorithm is used, a classification tree. A classification tree is an easy to interpret classification procedure that naturally allows interactions of high degree across predictors. The classification tree uses the first few courses attempted and grades obtained by students in order to classify them. To improve this classifier, multiple trees are grown and the result is a random forest. A random forest can also be used to assess variable importance in a reliable manner.

The University of Toronto provided a large data set containing individual-level student grades for all undergraduate students enrolled at the Faculty of Arts and Science at the University of Toronto - St. George campus between 2000 and 2010. The data set contains over 1 600 000 grades and over 65 000 students. This data set was studied by Bailey et al. (2016) and was used to build an adjusted GPA that considers course difficulty levels. Here, random forest classifiers are built upon this data set and these classifiers are later tested.

The contribution in this article is two-fold. First, classifiers are built and the prediction accuracy of those classifiers exceeds the accuracy of the linear classifiers thus making them useful for universities that would like to predict where their resources need to be allocated. Second, the variable importance analysis contains a lot of interesting information. Among many things, the high importance of grades in low-grading departments was noted and might be a symptom of grade inflation.

2 Literature review

2.1 Predicting success

In this article a statistical learning model is established to predict if a student succeeds at completing an undergraduate program and to predict what major was completed. This statistical analysis of a higher education data set shares similarities with recent articles by Chen and Desjardins (2008, 2010) and Leeds and DesJardins (2015) as a new statistical approach will be introduced, a data set will be presented and policy making implications will be discussed. The task of predicting student academic success has already been undertaken by many researchers. Recently Kappe and van des Flier (2012) tried to predict academic success using personality traits. In the meanwhile, Glaesser and Cooper (2012) were interested in the role of parents' education, gender and other socio-economic metrics in predicting high school success.

While the articles mentioned above use socio-economic status and personality traits to predict academic success, many researchers are looking at academic-related metrics to predict graduation rates. Johnson and Stage (2018) use High-Impact Practices, such as undergraduate research, freshman seminars, internships and collaborative assignments to predict academic success. Using regression models, they noted that freshman seminars and internships were significant predictors. Niessen and al. (2016) discuss the significance of trial-studying test in predicting student dropouts. This test was designed to simulate a representative first-year course and student would take it before admission. The authors noted that this test was consistently the best academic achievement predictor.

More recently, Aulck and al. (2016) used various machine learning methods to analyse a rather large data set containing both socio-economic and academic metrics to predict dropouts. They noted similar performances for the three methods compared; logistic regression, k-nearest neighbours and random forests. The proposed analysis differs from the above-mentioned as it takes on the challenge to predict academic success and major using strictly academic information available in student records. The benefits of having classifiers built upon data they already own is huge for university administrations. It means university would not need to force students to take entry tests or relies on outside firms in order to predict success rate and major which is useful in order to prevent dropout or to allocate resources among departments. As noted by Aulck and al. (2016) machine learning analysis of academic data has potential and the uses of random forest in the following article aims at exploiting this potential.

2.2 Identifying important predictors

Identifying and interpreting the variables that are useful to those predictions are important problems as well. It can provide university administrator with interesting information. The precise effect of grades on a student motivation lead to many debates and publications over the years (more recently (Mills & Blankstein, 2000; Ost, 2010)). Because grades should be indicators of a student's abilities, evaluating the predictive power of grades in various departments is important. University administrators might want to know if grades in a department are better predictors than grades in other departments. Continuing on the point, it is also important to understand what makes the evaluations in a department a better indicator of students' success. Random forest mechanisms lead to variable importance assessment techniques that will be useful to understand the predictive power of grades variables.

Understanding the importance ranking of grades in various departments can also enlighten us regarding the phenomenon of *grade inflation.* This problem and some of its effect has been already discussed in many papers ((Sabot & Wakeman-Linn, 1991; V. E. Johnson, 2003; Bar, Kadiyali, & Zussman, 2009)) and it is consensual that this inflation differs from one department to another. According to Sabot and Wakeman-Linn, (1991) this is problematic since grades serve as incentives for course choices for students and now those incentives are distorted by the grade inflation. As a consequence of the different growths in grades, they noted that in many universities there exist a chasm in grading policies creating high-grading departments and low-grading departments. Economics, Chemistry and Mathematics are examples of low-grading departments while English, Philosophy and Political Science are considered high-grading.

As Johnson mentions (V. E. Johnson, 2003), students are aware of these differences in grading, openly discuss them and this may affect the courses they select. This inconsistency in course difficulty is also considered by Bailey and al. (2016) as they built an adjusted GPA that considers course difficulty levels. The accuracy of that adjusted GPA in predicting uniform test result is a great demonstration that courses do vary in difficulty. If some departments suffer from grade inflation, the grades assigned in that department should be less tied to the actual student ability and therefore they should be less predictive of student success. A thorough variable importance analysis will be performed in order to test this assumption.

Understanding which predictors are important can also provide university administrators with feedback. For example, some of the High-Impact Practices identified by Randall Johnson and King Stage (2018) are part of the University of Toronto's program. The variable importance analysis could be a useful tool to assess the effect of such practices.

3 Methodology

3.1 Data

The data set provided by the University of Toronto contains 1 656 977 data points, where each observation represents the grade of one student in one course. A data point is a 7 dimensions observation containing the student ID, the course title, the department of the course, the semester, the credit value of the course and finally the numerical grade obtained by the student. As this is the only data obtained, some pre-processing is required in order for algorithms to be trained. The **first research question** is whether it is possible to design an algorithm which accurately predicts whether or not a student will complete their program. The **second research question** is whether it is possible to design an algorithm which accurately predicts, for students who complete their program, which major they will complete. These two predictions will be based upon first-year student records.

The data has been pre-processed for the needs of the analyses. At the University of Toronto, a student must complete 20 credits in order to obtain an Honours B.A. or B.Sc (University of Toronto, 2017). A student must also either complete 1 Specialist, 2 Majors or 1 Major and 2 Minors. The first five credits attempted by a student roughly represent one year of courses. Therefore, for each student every semester until the student reaches 5 attempted credits are used for prediction. It means that for some students, the predictors represent exactly 5 attempted credits and for some other students, a bit more. The set of predictors consists of the number of credits a student attempted in every department and the average grade across all courses taken by the student in each department. Since courses were taken by students in 71 different departments, the predictor vector is of length 142. Of course, many other predictors could also be computed from the data set, but these are the most appropriate ones for the purpose of the variable importance analysis.

To answer the first research question, a binary response indicating whether or not a student completed their program is needed. Students that completed 18 credits were labelled as students who completed their program. Students who registered to 5 credits worth of courses, succeeded at fewer than 18 credits worth of courses and stopped taking courses for 3 consecutive semesters are considered students who began a program but did not complete it. All other students were left out of the analysis. Since some students take classes in other faculties or universities, 18 credits was deemed a reasonable threshold. It is possible that some students did not complete their program even though they completed 18 credits, but it is more likely that they took courses in other faculties or universities. To be considered dropouts, only students who registered to at least 5 credits worth of courses were considered. It was assumed that students that registered to fewer credits were registered in another faculty, campus, university or were simply auditing students. After this pre-processing was performed, the data set contains 38 842 students of which 26 488 completed an undergraduate program and 12 294 did not.

To answer the second research question a categorical response representing the major completed by the student is required. To do so, the 26 448 students who completed a program are kept. The response will represent the major completed by the student. Since this information is not available in the data set, the department in which the student completed the largest number of credits is considered the program they majored in. Therefore, the response variable is a categorical variable that can take 71 possible values. This formatting choice might be a problem for students who completed more than 1 major. Some recommendations to fix that problem can be found in the conclusion.

Regarding the various grading policies of this university it was noticed that Mathematics, Chemistry and Economics are the three departments with the lowest average grades. As grades do vary widely across the data set there is no statistically significant difference between the departments but it is still interesting to observe that departments that were defined as low-grading departments in many papers do appear as the lowest grading departments in this data set too. Finally, the data set was divided in three parts as is it usually done. The algorithm is trained upon the training set, which contains 90% of the observations in order to learn from a large portion of the data set. 5% of the data set is assigned to the validation set which is utilized to select various optimization parameters. Finally, the rest of the data set is assigned to the test set, which is a data set totally left aside during training and later used to test the performances of the trained classifier.

3.2 Classification Tree

A typical supervised statistical learning problem is defined when the relationship between a response variable and an associated set of predictors (used interchangeably with inputs) is of interest. The response is what needs prediction, such as the program completion, and the predictors, such as the grades, are used to predict the response. When the response variable is categorical, this problem is defined as a classification problem. One challenge in classification problems is to use a data set in order to construct a classifier. A classifier is built to emit a class prediction for any new observation with unknown response. In this analysis, classifiers are built upon the data set described in section 3.1 to predict if a new student will complete its program and what major will be completed using information related to its first year of courses.

A classification tree (Breiman, Friedman, Olshen, & Stone, 1984) is a model that classifies new observations based on set of conditions related to the predictors. For example, a classification tree could predict a student is on its way to complete a program because it attempted more than 2 Mathematics courses, obtained an averaged grade in Mathematics above 80 and attempted fewer than 2 Psychology courses. The set of conditions established by a decision tree partitions in multiple regions the space defined by possible predictors values. Intuitively, a classification tree forms regions defined by some predictors values and assign a response label for new observations that would belong in those regions. Figure 1 illustrates an example of a predictor space partition, its associated regions and its associated classification tree for observations defined by two predictors. The final set of regions can be defined as leaves in a tree as represented in Figure 1, hence the name classification trees.

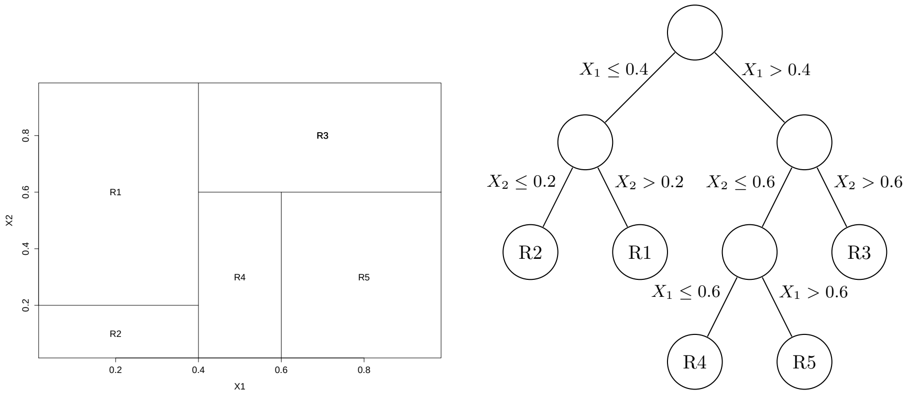

Figure 1: Illustration of a decision tree partition of the predictor space in 5 regions and the associated decision tree

Now that the model has been established, an algorithm that creates the classification tree using a training set of labelled observations needs to be defined. The algorithm creates the regions by recursively establishing the conditions. It aims at building regions that contains a high concentration of observations of the same class. Usually a measure of impurity is defined; the further the region is from containing only observations with the same label, the bigger this measure is. Intuitively, it is desired to obtain a set of conditions under which all students either completed their programs or not. Therefore, the algorithm analyses how mixed are the labels according to all possible conditions and selects the condition that minimizes the measure of impurity. For example, the algorithm will look at all conditions of the form : ”did the student attempt more or less than 1 Mathematics course ?” and select the condition that best divides students that completed a program from students that did not.

Once a condition is selected, the training observations are effectively divided in two sets of training observations based upon the condition. The process is repeatedly applied on the two resulting training sets. The algorithm divides the training observations in smaller sets until each resulting set contains few observations. When the partitioning process is completed, each region is labelled with the class representing the majority of observations respecting the conditions defining the region. A more formal definition of the algorithm is included in the appendix.

3.3 Random Forest

By constructing a decision tree, a powerful and easy to interpret classifier is obtained. As will be demonstrated in this section, one way to improve this classifier is to build a set of classifiers using samples of the training set.

Suppose there is a way to obtain a set of classifiers. The goal is to find a technique that uses the entire set of classifiers to get a new classifier that is better than any of them individually. One method of aggregating the class predictions is by *voting*: the predicted class for a new observation is the most picked class among individual classifier. A critical factor in whether the aggregating procedure will improve the accuracy or not is the stability of the individual classifiers. If a small variation in the training set has almost no effect on the classifier, this classifier is said to be stable, and utilizing a set of classifiers based upon similar training sets will result in a set of almost identical classifiers. For unstable procedures, the classifiers in the set are going to be very different from one another. For such classifiers, the aggregation will greatly improve both the stability and accuracy of the procedure. Procedure stability was studied by Breiman (1996b); classification trees are unstable.

Bootstrap aggregating (*bagging*) was introduced by Breiman (1996a) as a way to improve unstable classifiers. In bagging, each classifier in the set is built upon a different bootstrap sample of the training set. A bootstrap sample is simply a random sample of the original training sets. Each of the samples are drawn at random with replacement from the original training set and are of the same size. Doing so will produce a set of different training sets. For each of these training set a decision tree is fitted and together they form a random forest. Overfitting is a problem caused when a classifier identifies a structure that corresponds too closely to the training set and generalizes poorly to new observations. By generating multiple training sets, fitting multiple trees and building a forest out of these tree classifiers it greatly reduces the chances of overfitting. Breiman (2001) defines a *random forest* as a classifier consisting of a set of tree-structured classifiers where each tree casts a unit vote for the most popular class at one input.

Breiman introduced in 2001 random forests with random inputs (Breiman, 2001) which is the most commonly used random forest classifier. The novelty of this random forest model is in the tree-growing procedure. Instead of finding the best condition among all the predictors, the algorithm will now randomly select a subset of predictors and will find the best condition among these, this modification greatly improved the accuracy of random forests.

Random forests are easy to use and are stable classifiers with many interesting properties. One of these interesting properties is that they allow for powerful variable importance computations that evaluate the importance of individual predictors throughout the entire prediction process.

3.4 Variable Importance in Random Forests

A variable importance analysis aims at understanding the effect of individual predictors on the classifier output. A predictor with a great effect is considered an important predictor. A random forest provides multiple interesting variable importance computations. The *Gini decrease importance* sums the total impurity measure decrease caused by partitioning upon a predictor throughout an entire tree and then computes the average of this measure across all trees in a forest. This technique is tightly related to the construction process of the tree itself and is pretty easy to obtain as it is non-demanding computationally.

The *permutation decrease importance* was introduced by Breiman (2001). Intuitively if a predictor has a significant effect on the response, the algorithm should lose a lot of prediction accuracy if the values of that predictor are mixed up in the data set. One way to disrupt the predictors values is by permutations. The procedure computes the prediction accuracy on the test set using the true test set. Then, it permutes the values of one predictor, j, across all observations, run this permuted data through the forest and compute the new accuracy. If the input j is important, the algorithm should lose a lot of its prediction accuracy by permuting the values of j in the test set. The process is repeated for all predictors, then it is averaged across all trees and the averaged prediction accuracy decreases are compared. The larger the decrease in accuracy the more important the variable is considered.

Storbl & al. (2007) recently published an article where these techniques are analysed and compared. According to this paper, the selection bias of the decision tree procedure might lead to misleading variable importance. Numerous papers (Breiman et al., 1984; Kim & Loh, 2001; Kononenko, 1995) noticed a selection bias within the decision tree procedure when the predictors are of different nature. The simulation studies produced by Storbl & al. (2007) show that the Gini decrease importance is not a reliable variable importance measure when predictors are of varying types. The Gini decrease importance measure tends to overestimate the importance of continuous variables.

It is also shown (Strobl et al., 2007) that the variable importance techniques described above can give misleading results due the replacements when drawing bootstrap samples. It is recommended that researchers build random forests with bootstrap samples without replacements and use an unbiased tree-building procedure (Loh & Shih, 1997; Kim & Loh, 2001; Loh, 2002; Hothorn, Hornik, & Zeileis, 2006). If a classic tree-building procedure is used, predictors should be of the same type or only the permutation decrease importance is reliable.

3.5 Algorithms

A classification tree using the Gini impurity as split measurement was coded in the C++ language using the Rcpp library (Eddelbuettel & Francois, 2011). The code is available upon request from the first author. The algorithm proceeds as explained in Section 3.2, the tree it produces is unpruned and training sets are partitioned until they contain only 50 observations. Three versions of the random forest algorithm are going to be used. Even though one of these models will outperform to two other in terms of prediction accuracy, the variable importance analysis of all three models will be considered and aggregate. For clarity and conciseness purposes, only the best model's performance will be assessed. **Random forest # 1** consists of 200 trees and can split upon every variable in each region. Bootstrap samples are drawn without replacement and contain 63% of the original training set. **Random forest # 2** fits 200 trees but randomly selects the variable to be partitioned upon in each region.

Finally, the popular R RandomForest package (Liaw & Wiener, 2002) was also used. It is an easy to use and reliable package that can fit random forests and produce variable importance plots. Using this package, **random forest # 3** was built. It contains 200 trees. Once again, bootstrap samples are drawn without replacement and contain about 63% of the size of the original training set. By default, this algorithm randomly selects a subset of inputs for each region. Regarding the impurity measure, the Gini impurity was selected because it has interesting theoretical properties, such as being differentiable, and has been performing well empirically.

Linear models were trained for both of the classification problems serving as benchmarks. In order for the comparison to be as direct as possible, the linear model classifiers were constructed upon the same set of predictors; it may be possible to improve both the random forest and the linear model with different predictors. As the problems are two classification ones, the linear models selected were logistic regression models and details regarding their parametrizations are included in the appendix.

4 Results

4.1 First research question : Predicting program completion

Random forest # 3 produced the best accuracy on the test set. Among the students who completed their program in the test set, the classifier achieves a 91.19% accuracy. Out of the 418 students who did not complete their program, the classifier achieves a 52.95% accuracy. The combined result it a 78.84% accuracy over the complete test set.

Obviously this is higher accuracy than if all students would be classified as students who competed their program, which would result in a 68.08% accuracy. The random forest accuracy is also slightly higher than the 74.21% accuracy achieved with a logistic regression based upon the same predictors. These predictions can be useful for university administrations that would like to predict the number of second-year students and prepare accordingly with a sufficient margin. About 75% of students identified as dropouts by the random forest classifier are true dropouts. Therefore students identified as dropouts by the algorithm could be considered higher-risk students and these predictions could be useful in order to target students in need of more support to succeed. The relatively high accuracy of the classifier is also an indicator that the variable importance analysis is reliable.

Variable importance is determined by the average decrease in accuracy in the test set caused by a random permutation of the predictor. This technique has been selected since it is more reliable as explained in Section 3.4. The top 15 variables according to the permutation decrease were kept and ordered in Figures 2,3 and 4. Since variable importance varies from one model to another, the three variable importance plots were included and the results will be aggregated.

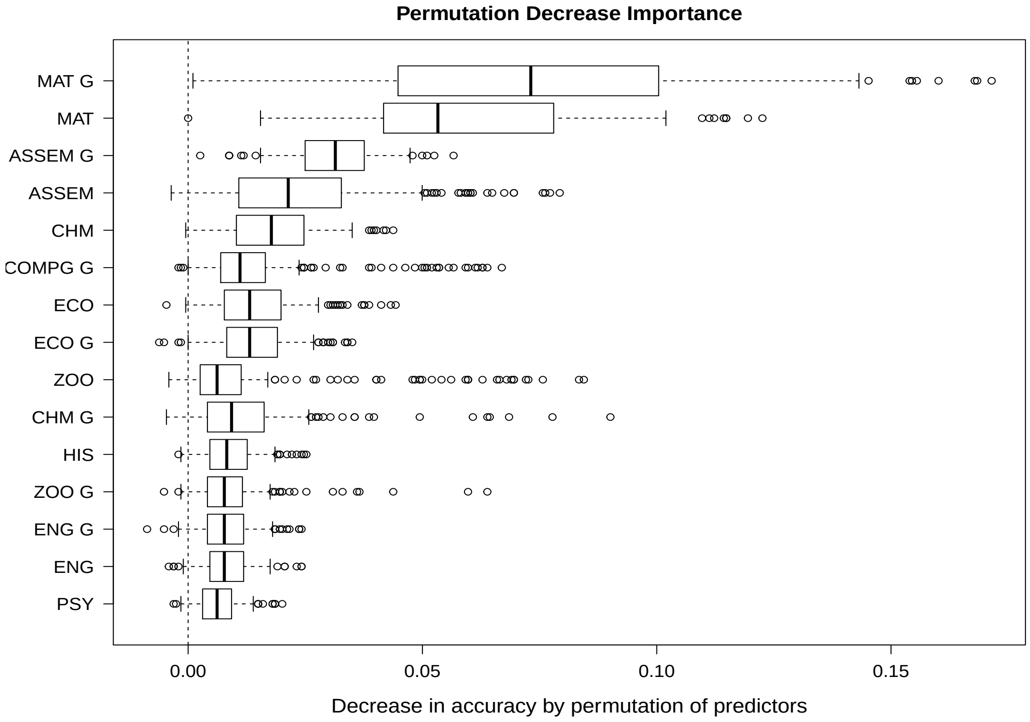

Figure 2: Variables importance boxplots for the **random forest # 1**.

In Figures 2,3 and 4 and for all the following figures, the variable representing the number

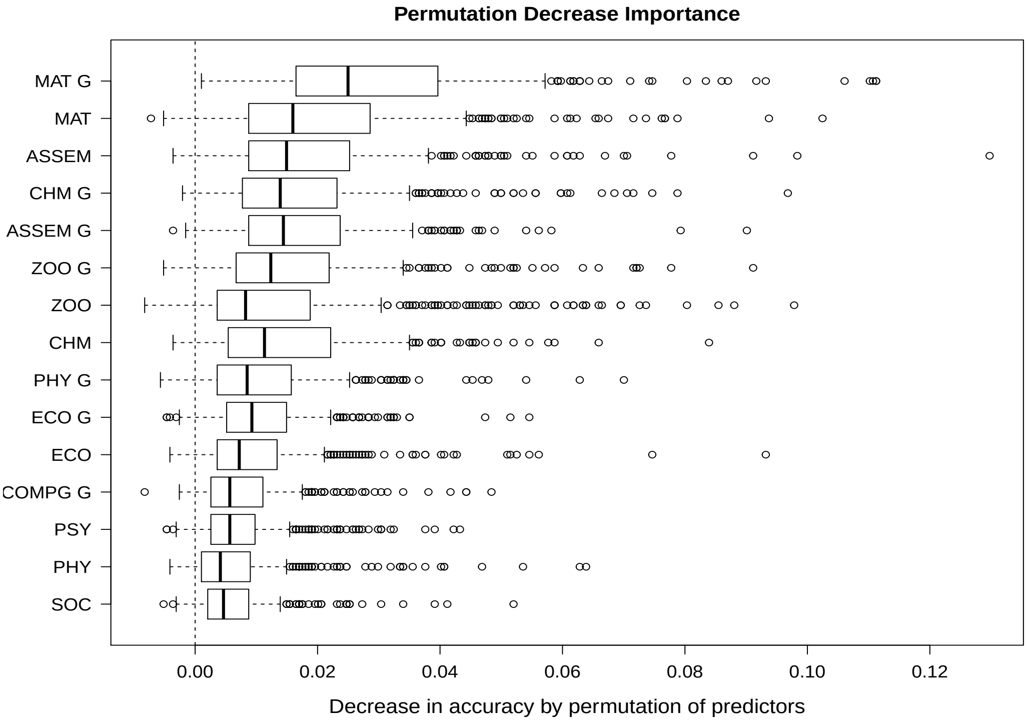

Figure 3: Variables importance boxplots for the **random forest # 2**.

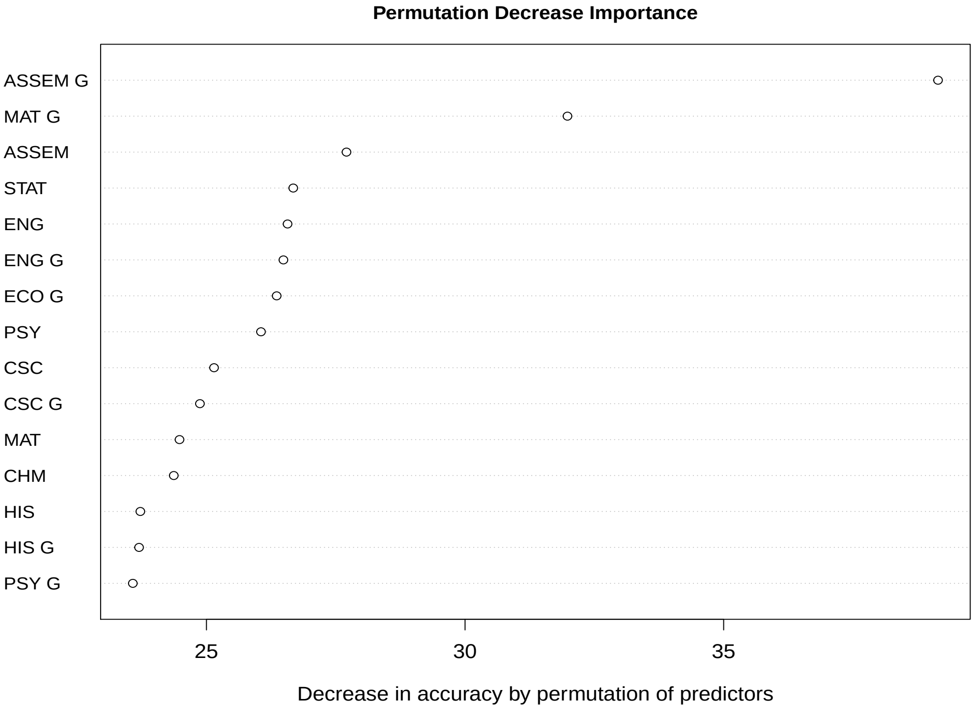

Figure 4: Variable importance plot produced by the RandomForest package for the **random forest # 3**.

of credits in a department is identified by the department code, i.e. the number of credits in Chemistry is identified by CHM. The variable representing the averaged grade in a department is identified by the department code followed by the letter G, i.e CHM G represents the averaged grade in Chemistry.

To begin, it was also noted that the variance for the grade variables were larger. Across all three random forests, the grades in Mathematics (MAT), Finance (COMPG), Economics (ECO) are consistently among the most important grade variable. These departments are considered low-grading departments and perhaps the strict marking of these departments helps to better distinguish students among themselves. A possible explanation is that the grade inflation that suffered the high-grading departments caused the grades to be no longer a reliable tool to distinguish students among themselves which could be a symptom of grade inflation as suggested in section 2.2. Other factors could have caused this phenomenon such as less sequential courses in Human Science fields, larger classes size, reduced access to a professor or other factors. It is impossible to claim for sure that these results are caused by the grade inflation problem, but these results could indicate such thing. Therefore, universities could use such technique to verify if grades in a department have more predictive power than grades in other departments and act accordingly since grades should represent students' abilities.

It is also important to notice the importance of ASSEM in the three variable importance plots. The ASSEM code represents a special type of first year seminar course. It seems that the students that registers in theses courses are easy to classify as both grades and the number of credits are considered important. This result agrees with the result obtained by Johnson and Stage (2018) about the importance of first year seminar courses. The first year seminar courses (ASSEM) were brand new at the University of Toronto and the analysis performed provided evidence of the merit of such courses in order to establish a student's profile and to predict success. In other words, such variable importance analysis could help university administrations assess the usefulness of new programs and courses.

4.2 Second research question : Predicting the major

The second task at hand is to build a random forest that predicts the student's major. Once again, from a prediction accuracy perspective, **random forest # 3** offered better performances with a 47.41% accuracy in predicting the major completed. This appears slightly lower than expected, but considering there are 71 different programs, being able to pin down the right program for about half of the students seems successful. This is a better result than the meager 4.75% obtained by assigning majors with probabilities weighted by the proportion of the majors completed. The 47.41% accuracy of the random forest is also above the 42.63% accuracy obtained by the multinomial logistic regression benchmark. For classification purposes, these classifiers could help individual departments predict the number of students registering to second, third or fourth year courses and graduate programs. Predicting the major could also help university administrations to allocate the financial resources among the departments or to decide the programs that require more advertisements.

Variable importance is also interesting for that research questions. Here is the variable importance analyses produced by the three random forests; once again, the 15 most important predictors are displayed. The importance of a predictor is determined by the average decrease in accuracy in the test set caused by a random permutation of the predictor.

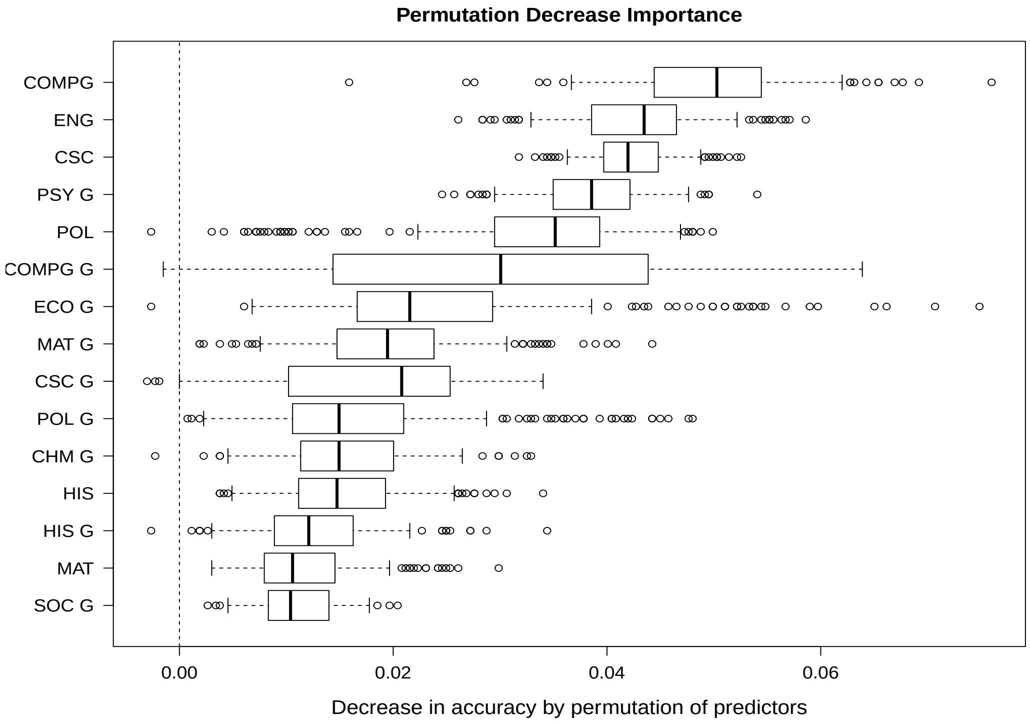

Figure 5: Variables importance boxplots for the **random forest # 1**.

A decrease in importance for the grades variable is noted in Figure 5,6 and 7. This was to

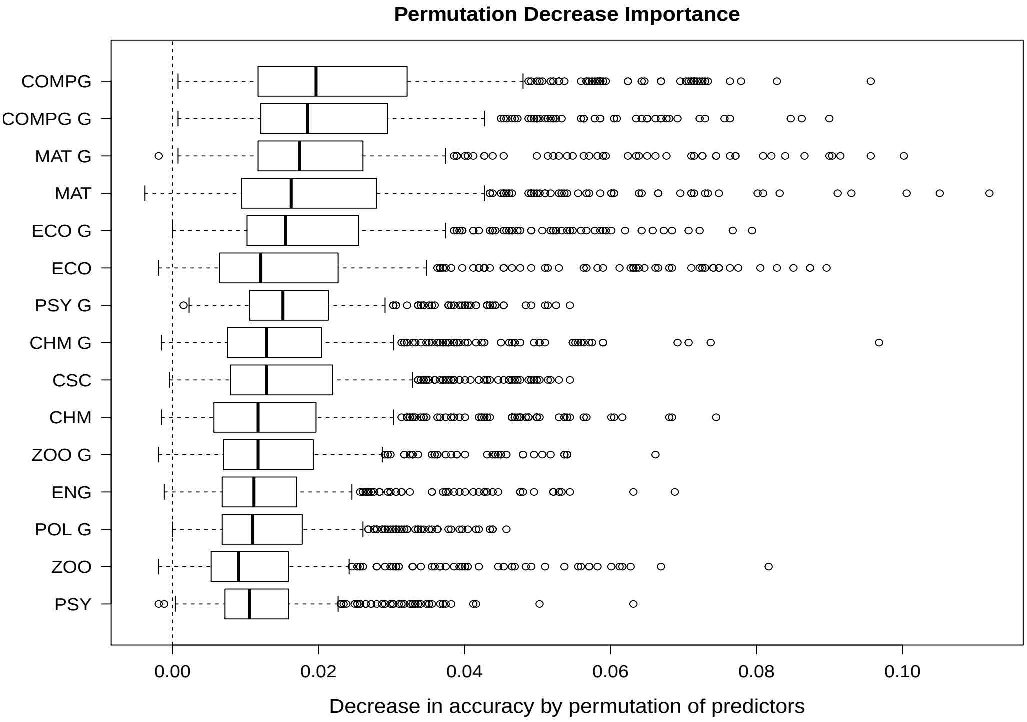

Figure 6: Variables importance boxplots for the **random forest # 2**.

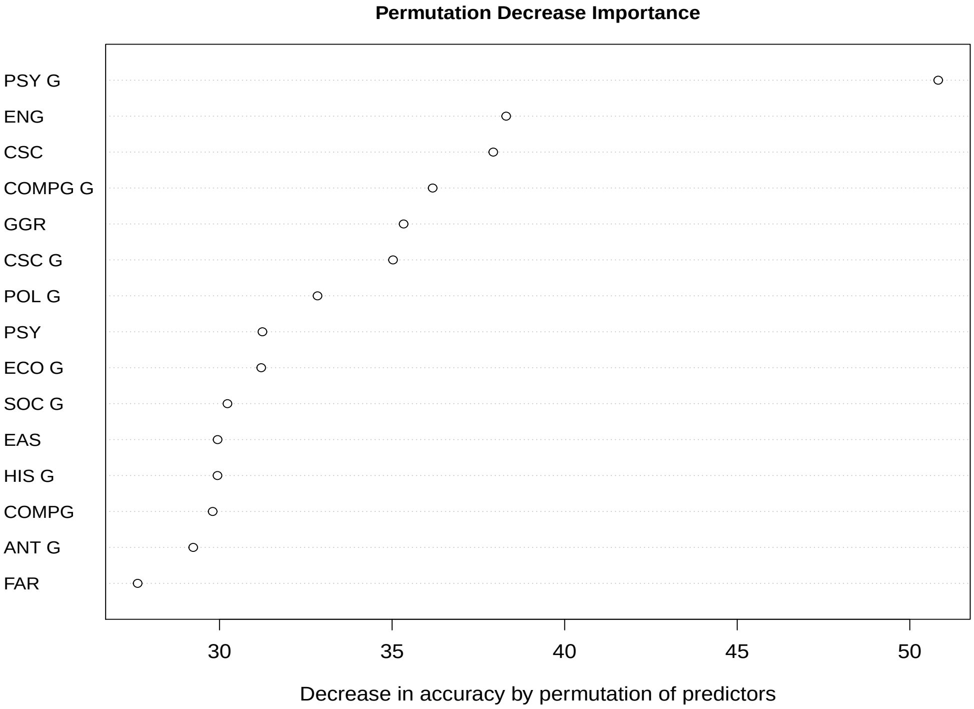

Figure 7: Variable importance plot produced by the RandomForest package for the **random forest # 3**.

be expected because of how the data was formatted. Since the department in which the highest amount of credit was obtained is considered the major completed by the student, these variable importance measures are not surprising. Actually, if all the courses were included, instead of only the first year, the amount of credit in every department precisely defines the response variable. Considering this weakness in the data formatting, the grades still have a relatively high importance. It seems hard to see any effect of grading policies in the predictive power of grades regarding that research question.

It seems like for some departments, such as English (ENG) and Computers Sciences (CSC), it is easy to predict students that will complete a major in those departments by almost solely looking at the number of courses attempted in those departments during the first year. This is caused by the fact that a vast majority of students that take courses in Computers Science or English during their first year end up completing an undergraduate program in these departments respectively. From a policy-making perspective, departments could use this information as they might want to adapt the content of their first-year courses now that they know more about the audience of these courses.

5 Conclusion

The first year's worth of courses and grades were used to build two classifiers; one that predicts if a student will complete their undergraduate program, the other that predicts the major of a student who completed a program. Random forests were used to build those classifiers. Random forests are easy to use with most statistical computing languages, fast to train, and they outperform linear logistic models in terms of prediction accuracy. For practitioners, random forests could be an alternative to typical linear models for various prediction tasks; to predict the number of students registered in second-year courses, the distribution of students across the many programs or to identify students at risk of failing or dropping out.

Evaluating the importance of each predictor is also something that offers random forest in comparison to the benchmark model. In this study, it was observed in Section 4 that grades were important for predicting if a student will complete their program. Grades in departments that were considered low-grading departments in some grades inflation research articles like Mathematics, Economics and Finance are consistently among the most important variables. These results indicate that a strong relationship exists between the grades in low-grading departments and the chance of succeeding at an undergraduate program, although this does not necessarily indicate a *causal* connection. Grades were somewhat less important predictors for predicting the students' major but even though they were less important, grades in Mathematics, Finance, Economics and Psychology (PSY) were still frequently significantly important.

Finally, for potential improvements in the data analysis, it is to be noted that some students might have completed more than one major or specialization. This might explain the relatively low accuracy for major choice prediction. Allowing for multiple major choices is a potential improvement for this model. This is in fact a multi-label classification problem and some solutions have already been proposed to adapt decision trees to accommodate this more complicated problem (Clare & King, 2001; Y.-L. Chen, Hsu, & Chou, 2003; Chou & Hsu, 2005). Some departments also share a great deal of similarities and might be considered equivalent by the university, thus combining some of them might increase the prediction accuracy. The missing values in the predictors were also problematic. Ideally, the algorithm would consider splitting on the grade variables for a certain department only to classify students who took courses in that department. Developing a new decision tree algorithm where new variables are added to the pool of potential split variables depending on previous partitioning should be a great way to improve the actual model in certain scenarios. Overall, implementing a new tree-building procedure where variable are added or discarded based upon previous partitioning and considering a multi-label classifier like suggested by Chen & al. (2003) could be great improvements for future work on that data set.

Acknowledgement

We are very grateful to Glenn Loney and Sinisa Markovic of the University of Toronto for providing us with students grade data. The authors also gratefully acknowledge the financial support from the NSERC of Canada.

References

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016, June). Predicting Student Dropout in Higher Education. *ArXiv e-prints*.

Bailey, M. A., Rosenthal, J. S., & Yoon, A. H. (2016). Grades and incentives: assessing competing grade point average measures and postgraduate outcomes. *Studies in Higher Education*, *41*(9), 1548-1562. Retrieved from http://dx.doi.org/10.1080/03075079.2014.982528 doi: 10.1080/03075079.2014.982528

Bar, T., Kadiyali, V., & Zussman, A. (2009). Grade information and grade inflation: The cornell experiment. *Journal of Economic Perspectivs*, *23*(3), 93–108.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *24*(2), 123–140. Retrieved from http://dx.doi.org/10.1007/BF00058655 doi: 10.1007/BF00058655

Breiman, L. (1996b, 12). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, *24*(6), 2350–2383. Retrieved from http://dx.doi.org/10.1214/aos/1032181158 doi: 10.1214/aos/1032181158

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. Retrieved from http://dx.doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, California, U.S.A.: Wadsworth Publishing Company.

Chen, R., & DesJardins, S. L. (2008, Feb 01). Exploring the effects of financial aid on the gap in student dropout risks by income level. *Research in Higher Education*, *49*(1), 1–18. Retrieved from https://doi.org/10.1007/s11162-007-9060-9 doi: 10.1007/s11162-007-9060-9

Chen, R., & DesJardins, S. L. (2010). Investigating the impact of financial aid on student dropout risks: Racial and ethnic differences. *The Journal of Higher Education*, *81*(2), 179–208. Retrieved from http://www.jstor.org/stable/40606850

Chen, Y.-L., Hsu, C.-L., & Chou, S.-C. (2003). Constructing a multi-valued and multi-labeled decision tree. *Expert Systems with Applications*, *25*(2), 199 - 209. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417403000472 doi: http://dx.doi.org/10.1016/S0957-4174(03)00047-2

Chou, S., & Hsu, C.-L. (2005, May). MMDT: A multi-valued and multi-labeled decision tree classifier for data mining. *Expert Syst. Appl.*, *28*(4), 799–812. Retrieved from http://dx.doi.org/10.1016/j.eswa.2004.12.035 doi: 10.1016/j.eswa.2004.12.035

Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. In L. De Raedt & A. Siebes (Eds.), *Principles of data mining and knowledge discovery: 5th european conference, pkdd 2001, freiburg, germany, september 3–5, 2001 proceedings* (pp. 42–53). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/3-540-44794-6 doi: 10.1007/3-540-44794-6

Eddelbuettel, D., & Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(1), 1–18. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v040i08 doi: 10.18637/jss.v040.i08

Glaesser, J., & Cooper, B. (2012). Gender, parental education, and ability: their interacting roles in predicting gcse success. *Cambridge Journal of Education*, *42*(4), 463-480. Retrieved from https://doi.org/10.1080/0305764X.2012.733346 doi: 10.1080/0305764X.2012.733346

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651-674. Retrieved from http://dx.doi.org/10.1198/106186006X133933 doi: 10.1198/106186006X133933

Johnson, S. R., & Stage, F. K. (2018). Academic engagement and student success: Do high-impact practices mean higher graduation rates? *The Journal of Higher Education*, *0*(0), 1-29. Retrieved from https://doi.org/10.1080/00221546.2018.1441107 doi: 10.1080/00221546.2018.1441107

Johnson, V. E. (2003). *Grade inflation : A crisis in college education.* Springer.

Kappe, R., & van der Flier, H. (2012, Dec 01). Predicting academic success in higher education: what's more important than being smart? *European Journal of Psychology of Education*, *27*(4), 605–619. Retrieved from https://doi.org/10.1007/s10212-011-0099-9 doi: 10.1007/s10212-011-0099-9

Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*, 589–604. Retrieved from http://www.stat.wisc.edu/~loh/treeprogs/cruise/cruise.pdf

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *Proceedings of the 14th international joint conference on artificial intelligence - volume 2* (pp. 1034–1040). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1643031.1643034

Leeds, D. M., & DesJardins, S. L. (2015, Aug 01). The effect of merit aid on enrollment: A regression discontinuity analysis of iowa's national scholars award. *Research in Higher Education*, *56*(5), 471–495. Retrieved from `https://doi.org/10.1007/s11162-014-9359-2` doi: 10.1007/s11162-014-9359-2

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18-22. Retrieved from `http://CRAN.R-project.org/doc/Rnews/`

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*, 361–386. Retrieved from `http://www.stat.wisc.edu/~loh/treeprogs/guide/guide02.pdf`

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815–840. Retrieved from `http://www3.stat.sinica.edu.tw/statistica/j7n4/j7n41/j7n41.htm`

Mills, J. S., & Blankstein, K. R. (2000). Perfectionism, intrinsic vs extrinsic motivation, and motivated strategies for learning: a multidimensional analysis of university students. *Personality and Individual Differences*, *29*(6), 1191 - 1204. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0191886900000039` doi: http://dx.doi.org/10.1016/S0191-8869(00)00003-9

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016, 04). Predicting performance in higher education using proximal predictors. *PLOS ONE*, *11*(4), 1-14. Retrieved from `https://doi.org/10.1371/journal.pone.0153663` doi: 10.1371/journal.pone.0153663

Ost, B. (2010). The role of peers and grades in determining major persistence in sciences. *Economics of Education Review*(29), 923–934.

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *Journal of Economic Perspectives*, *5*, 159-170.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. Retrieved from `http://dx.doi.org/10.1186/1471-2105-8-25` doi: 10.1186/1471-2105-8-25

University of Toronto. (2017). *Degree requirements (h.b.a., h.b.sc., bcom).* Retrieved 2017-08-30, from `http://calendar.artsci.utoronto.ca/Degree_Requirements_(H.B.A.,_H.B.Sc.,_BCom).html`

A Appendix

The following section contains some mathematical notations and definitions for readers who are interested in more a thorough explanation of sections' 3.2 and 3.3 content. Full understanding of the appendix is not needed in order to grasp the essential of the article but it serves as a brief but precise introduction to the mathematical formulation of decision trees and random forests.

Rigorously, a typical supervised statistical learning problem is defined when the relationship between a response variable $\mathbf{Y}$ and an associated m-dimensional predictor vector $\mathbf{X} = (X_1, ..., X_m)$ is of interest. When the response variable is categorical and takes k different possible values, this problem is defined as a k-class classification problem. One challenge in classification problems is to use a data set $D = \{(Y_i, X_{1,i}, ..., X_{m,i}); i = 1, ..., n\}$ in order to construct a classifier $\varphi(D)$. A classifier is built to emit a class prediction for any new data point $\mathbf{X}$ that belongs in the feature space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_m$. Therefore a classifier divides the feature space $\mathcal{X}$ into k disjoint regions such that $\cup_{j=1}^{k} B_l = \mathcal{X}$, i.e. $\varphi(D, \mathbf{X}) = \sum_{j=1}^{k} j \mathbf{1}\{\mathbf{X} \in B_j\}$.

As explained in section 3.2 a classification tree (Breiman et al., 1984) is an algorithm that forms these regions by recursively dividing the feature space $\mathcal{X}$ until a stopping rule is applied. Most algorithms stop the partitioning process whenever every terminal node of the tree contains less than β observations. This β is a tuning parameter that can be established by cross-validation. Let p_{rk} be the proportion of the class k in the region r, if the region r contains n_r observations then :

$$p_{rk} = \frac{1}{n_r} \sum_{x_i \in R_r} \mathbf{1}\{y_i = k\}. \tag{1}$$

The class prediction for a new observation that shall fall in the region r is the majority class in that region, i.e. if $\mathbf{X} \in R_r$, $\varphi(D, \mathbf{X}) = \text{argmax}_k(p_{kr})$. When splitting a region into two new regions R_1 and R_2 the algorithm will compute the total impurity of the new regions ; $n_1 Q_1 + n_2 Q_2$ and will pick the split variable j and split location s that minimizes that total impurity. If the predictor j is continuous, the possible splits are of the form $X_j \leq s$ and $X_j > s$ which usually results in $n_r - 1$ possible splits. For a categorical predictor having q possible values, it is common to consider all of the $2^{q-1} - 1$ possible splits. Hastie & al. (2009) introduces many possible region impurity measurements Q_r, in this project, the *Gini index* has been chosen :

$$Q_r = \sum_{j=1}^{k} p_{rj}(1 - p_{rj}). \tag{2}$$

Here is a pseudo-code of the algorithm :

```
Algorithm : DT($D$,$\beta$)

1. Starting with the entire data set $D$ as the first set of observations $r$.
2. Check ($n_r$ ¿ $\beta$).
3. if (false) :
      Assign a label to the node and exit.
   else if :
      for ($j$ in all predictors):
         for ($s$ in all possible splits) :
            Compute total impurity measure.
      Select variable $j$ and split $s$ with minimum impurity measure and split
      the set $r$ into two children sets of observations.
      Repeat steps 2 & 3 on the two resulting sets.
```

Since decision trees are unstable procedures (Breiman, 1996b) they greatly benefit from bootstrap aggregating (bagging) (Breiman, 1996a). In classifier aggregating, the goal is to find a way to use an entire set of classifiers $\{\varphi(D_q)\}$ to get a new classifier φ_a that is better than any of them individually. One method of aggregating the class predictions $\{\varphi(D_q, \mathbf{X})\}$ is by *voting*: the predicted class for the input $\mathbf{X}$ is the most picked class among the classifiers. More precisely, let $T_k = |\{q : \varphi(D_q, \mathbf{X}) = k\}|$ then, the aggregating classifier becomes $\varphi_a(\mathbf{X}) = \text{argmax}_k(T_k)$.

On way to form a set of classifiers is to draw bootstrap samples of the data set D which forms a set of learning sets $\{D_B\}$. Each of the bootstrap samples will be of size n drawn at random with replacement from the original training set D. For each of these learning set a classifier $\varphi(D_b)$ is constructed and the resulting set of classifiers $\{\varphi(D_b)\}$ can be used to create an aggregating classifier. If the classifier is an unpruned tree then the aggregating classifier is a random forest.

A random forest classifier is more precise than a single classification tree in the sense that it has lower mean-squared prediction error (Breiman, 1996a). By bagging a classifier, the bias will remain the same but the variance will decrease. One way to further decrease the variance of the random forest is by construction trees that are as uncorrelated as possible. Breiman introduced in 2001 random forests with random inputs (Breiman, 2001). In these forests, instead of finding the best variable and partitioning among all the variables, the algorithm will now randomly select $p < m$ random covariates and will find the best condition among those p covariates.

The fitted random forest classifiers were compared to two logistic regression models. A simple logistic model is used to predict if a student completes its program or not with the following parametrization :

$$P(Y_i = 1) = \frac{\exp(\sum_{i=0}^{m} \beta_i x_i)}{1 + \exp(\sum_{i=0}^{m} \beta_i x_i)}, \tag{3}$$

where $Y_i = 1$ means student i completed its program, m is the number of predictors, $\beta' s$ the parameters and $x_i' s$ the predictor values. To predict the major completed, a generalization of the logistic regression, the multinomial logistic regression is used with the following parametrization :

$$P(Y_i = p) = \frac{\exp(\sum_{i=0}^{m} \beta_i^{(p)} x_i)}{1 + \exp(\sum_{l=1}^{k} \sum_{i=0}^{m} \beta_i^{l} x_i)}, \tag{4}$$

where $Y_i = p$ means the student i completed the program p and where k is the number of programs.

Finally, here is a short example of code to fit random forests, get predictions for new observations and produce variable importance plots using the R language :

```
#Importing the randomForest package
require(randomForest)


#Fitting the random forest with 200 trees
#using bootstraps without replacement.
Fit <- randomForest(x=X,y=as.factor(Y),importance=TRUE,ntree=200,
replace=FALSE,sampsize=round(0.63*nrow(X)) )


#Prediction class labels for new observations newX
predictions <- predict(Fit,newX)
```

```
#Production variable importance plot

importance(Fit,type=1)
```