Predicting University Students' Academic Success and Choice of Major using Random Forests

Cédric Beaulac Jeffrey S. Rosenthal

August 31,2017

Abstract

In this paper, a large data set containing every course taken by every undergraduate student in a major university in Canada over 10 years is analyzed. 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 choice is predicted using once again the first few courses they've registered to. A classification tree is an intuitive and powerful classifier and building a random forest of trees lowers the variance of the classifier and also prevents overfitting. Random forests also allow for reliable variable importance measurements. These measures explain what variables are useful to both of the classifiers and can be used to better understand what is statistically related to the students' choices. The results are two accurate classifiers and a variable importance analysis that provides useful information to the university.

Keywords : Higher Education, Students' Success and Choice, Machine Learning, Classification Tree, Random Forest, Variable Importance

1 Introduction

As the demand for qualified labour increases it becomes more and more important to understand what motivates students to complete their program and how they select their majors. In parallel, universities are continuously trying to improve their programs and attract more students. It would be useful for a university to be able to predict whether or not a student that begins a program will complete it. It would also be useful for a university administration to have the ability to predict which major or specialization students are going to choose. Understanding which variables are useful in both of these predictions is important as it might help understand what drives student through their academic career. More importantly, are grades in some departments more useful than grades in other departments? If so, what is the effect of these grades on the student's academic path.

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 and grades obtained by students in order to classify them. To prevent overfitting and to reduce the variance of 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 35 000 students. This data set was studied in [1] 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 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. The variable importance analysis revealed the great importance of grades, but more precisely, grades in departments that are considered low-grading departments. This result is interesting as many researchers are still trying to understand the repercussion of the distorted grade inflation that is happening in many universities.

2 Program Completion, Course Selection and Grades

In this article a model is established to predict if a student succeeds at completing an undergraduate program and to predict what major is chosen. To classify a student, the algorithm uses information about the first few courses taken by them. The benefits of having those precise classifiers for a university administration are obvious as it can only help the administration to allocate their resources accordingly, understand the needs of some students and a variable importance analysis might help them understand certain strengths or weaknesses of their programs.

Identifying and interpreting the variables that are useful to those predictions are important problems as well. The precise effect of grades on a student motivation lead to many debates and publications over the years (more recently [19] [20]). Other than the trivial relation between having good grades and completing a program, understanding the importance of the grades in predicting if a student will complete their program or not is still a relevant question. Random forest mechanisms produce variable importance techniques that will be useful in understanding how grades affect student choices.

The phenomenon of *grade inflation* and some of its effect has been already discussed in many papers ([21], [13], [2]) and it is consensual that this inflation differs from one department to another. According to Sabot and Wakeman-Linn, [21] 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-grade departments and low-grade departments. Economics, Chemistry and Mathematics are examples of low-grading departments while English, Philosophy, Psychology and Political Science are considered high-grading.

As Johnson mentions [13], 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 discussed by Bailey, Rosenthal and Yoon [1] 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. It seems important to analyze if the importance of a grade variable is somehow tied to whether it is coming from a high-grading or a low-grading department.

3 Classification Tree and Random Forest : Techniques and notations

3.1 Fitting a Classification Tree

A typical supervised statistical learning problem is defined when the relationship between a response variable $\mathbf{Y}$ and an associated m-dimensional covariate 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 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. $\varphi(D, \mathbf{X}) = \sum_{j=1}^{k} j\mathbf{1}\{\mathbf{X} \in B_j\}$.

A classification tree [6] is an algorithm that forms these regions by recursively dividing the covariate space, more precisely, this procedure performs recursive binary partition. Beginning with the entire feature space, the algorithm selects the variable to split upon and the location of the split that minimizes some impurity measurement. Then the resulting two regions are split into two more regions until some stopping rule is applied. The classifier will label each region with one of the k possible classes. Depending on the stopping rule there can be either more or less than k regions. Therefore, multiple regions can be associated with the same class and, similarly, some classes might not be represented by any region.

The name *classification tree* comes from the ability to represent the final set of regions as leaves in a tree as represented in Figure 1. Let p_{rk} be the proportion of the class k in the region r, if the region r contains n_r observations then :

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

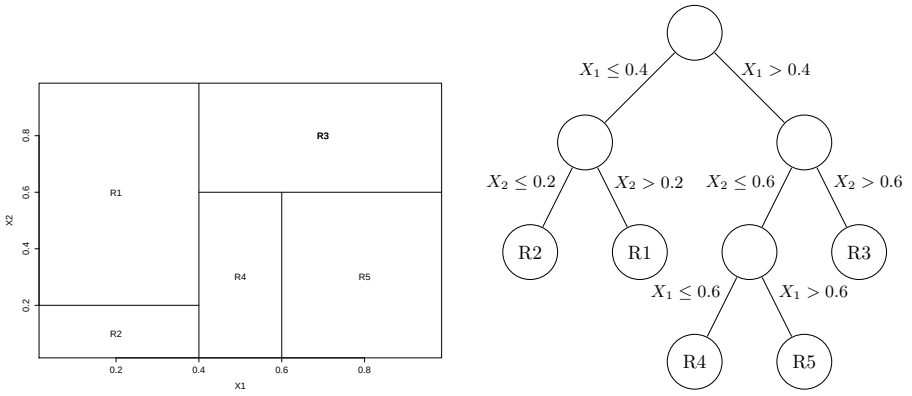

Figure 1: Illustration of the decision tree partitioning process and the resulting tree

The class prediction for a new observation that shall fall in the region r is the majority class in that region, i.e. if $\mathbf{X} \in R_r$, $\varphi(D, \mathbf{X}) = \text{argmax}_k(p_{kr})$. Hastie & al. [11] introduces many possible region impurity measurements, in this project, the *Gini index* has been chosen :

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

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 covariate 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, we usually consider all of the $2^{q-1} - 1$ possible splits.

The partitioning continues until a stopping rule is applied. Most algorithms stop whenever every terminal node of the tree contains less than β observations. This β is a tuning parameter; a small β will require more computing time and will create deeper trees that will have low classification error on the training set but might overfit. To prevent overfitting, a deep tree is built by setting β to be really small, 5 or 10, and then the tree can be pruned. Tree-pruning is a cost-complexity procedure that relies on considering that each leaf, region, is associated with a cost α. Rigorously, the tree pruning procedure consists of minimizing the cost complexity criterion defined as :

$$C_\alpha = \sum_{r=1}^{|R|} n_r Q_r + \alpha |R|, \tag{3}$$

where $|R|$ is the number of regions, i.e. the number of leaves in the tree. The procedure begins by collapsing regions that produce the smallest increase in total impurity $\sum_{r=1}^{R} n_r Q_r$ and this technique will collapse leaves as long as the increase in impurity is less than the cost α of the additional region. The α

parameter can be determined by cross-validation or with the use of a validation set.

3.2 Constructing a Random Forest

By constructing a pruned tree a powerful and easy to interpret classifier is obtained. One way to improve the stability of this classifier while also preventing overfitting is to build a forest of trees using bootstrap samples of the training set.

Bootstrap aggregating (*bagging*) was introduced by Breiman [3] as a way to reduce the variance of unstable predictors and to prevent overfitting. For the classification problem, the procedure consists of using an ensemble of classifiers that will each cast a vote towards a certain class. In bagging, each classifier in our ensemble is built upon a different bootstrap sample of our training set. Regarding decision trees, the bagging procedure will in fact creates multiple trees. Breiman [5] defines a *random forest* as a classifier consisting of a set of tree-structured classifiers build upon independent identically distributed random vectors where each tree casts a unit vote for the most popular class at one input.

Rigorously, suppose $\{D_q\}$ is an ensemble of learning sets. The goal is to find a way to use the entire set of classifiers $\{\varphi(D_q)\}$ to get a new classifier 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 trees. More precisely, let $T_k = |\{q : \varphi(D_q, \mathbf{X}) = k\}|$ then, the aggregating classifier becomes $\varphi_a(\mathbf{X}) = \text{argmax}_k(T_k)$.

Since having multiple training sets is unusual; bootstrap samples of the data set D can be drawn to form our ensemble of learning sets $\{D_B\}$. This can be used to create a similar aggregating classifier $\varphi_B(\mathbf{X})$ out of this ensemble of learning sets. 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 unpruned tree is fitted and together they form a random forest.

A critical factor in whether the bagging procedure will improve the accuracy or not is the stability of the individual classifier φ. If small variations in the learning set D have almost no effect on the original classifier φ then generating bootstrap samples and multiple classifiers 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 and the aggregation will greatly improve both the stability and accuracy of the procedure. Procedure stability was studied by Breiman [4]; classification trees are unstable and thus, greatly benefit from bagging.

A random forest classifier is more precise than a single classification tree in the sense that it has lower mean-squared prediction error [3]. 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. This process might increase the bias of the individual classifiers in exchange. Breiman introduced in 2001 random forests with random inputs [5] which is the most commonly used version of random forests. The major difference in this new random forest model is in the tree-growing procedure. Instead of finding the best variable and split point among all the variables, the algorithm will now randomly select $p < m$ random covariates and will find the best split among those p covariates.

These random forests are explained in detail in [11] and it is pointed out that these models require more tuning. Overfitting can be a problem and thus it is recommended to fit rather shallow trees by fixing a much large β parameter. It is also crucial to select an appropriate p parameter. Smaller p will result in less correlated tree and thus lower total variance but also higher bias. The default value for p is $\sqrt{m}$ and aims to be a good trade-off creating rather uncorrelated trees without increasing the bias too much.

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 variables throughout the entire prediction process.

4 Variable Importance in Random Forests

A random forest provides multiple interesting variable importance computations. When building a tree, the algorithm picks the split variable that reduces the most of a pre-specified impurity measurement. If a tree has access to all the variables, the first split should be done on the variable that grants the largest increase in prediction accuracy. Unless predictors are randomly selected, the first few splits performed by the algorithm are on variables that give a high amount of information on their own. Compiling the appearances of variables among the first few splits is an easy way to use the mechanisms of classification trees to assess variable importance. A problem regarding this technique is that it doesn't represent appropriately the total effect of a variable in the model.

Starting with the entire training set, the variable and split point that produces the largest decrease in impurity measurement is selected by the algorithm. If a variable isn't producing the largest decrease right away but instead is frequently picked by the algorithm in the deeper levels of a tree to a point where it produces a large total decrease in impurity over multiple splits, that variable should be considered an important one. The *Gini decrease importance* sums the total Gini decrease caused by splitting on a variable 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.

Finally, the *permutation decrease importance* was introduced by Breiman in 2001 [5]. Intuitively, if an input has a significant effect on the output we should lose a lot of prediction accuracy if the values of that covariate are mixed up in our data set. One way to disrupt the predictors values is by permutation. 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 points, run this permuted data through the forest and compute the new accuracy. If the input j is significant, we should lose a lot of 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. [22] published an article in 2007 where these techniques are analyzed and compared. According to this paper, the selection bias of the decision tree procedure might lead to misleading variable importance. Numerous papers ([6], [14], [15]) noticed a selection bias within the decision tree procedure when the predictors are of different nature. If the predictors are a mix of continuous and categorical variable, there exist more potential split points for the continuous variables, therefore, there are higher chances that the variable is picked for splitting by the algorithm.

The simulation studies produced by Storbl & al. [22] show that the Gini decrease importance is not a reliable variable importance measure when predictors are of varying types. Because of the selection bias towards variable with more possible split points, the algorithm will pick those variable more frequently and the Gini decrease attributed to those will grow very large. In other words, the Gini decrease importance measure tends to overestimate the importance of variable with more possible split points. It seems that the permutation decrease importance is more reliable has it is unbiased. The permutation decrease importance measure has a higher variance for predictors with a high number of split points but it is still unbiased and with a sufficient number of trees this measurement is reliable.

It is shown in [22] that the variable importance techniques described above can give misleading results due to the selection bias within classification trees and 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 ([18], [14], [17], [12]). If a classic tree-building procedure is used, predictors should be of the same type or only the permutation decrease importance is reliable.

5 Experimental set up

5.1 Data formatting

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. This data set will require formatting in order for a classification tree to be used. 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 chose.

To begin, the data needs to be formatted in a way that one observation represents one student. Then, the predictors need to represent the first two semesters of courses. To answer the first research question a binary response indicating whether or not a student completed their program is needed. Finally, for the second research question, a categorical response representing the major completed by the student is required.

At the University of Toronto, a student must complete 20 credits with a GPA of 1.85 or more in order to obtain an Honours B.A. or B.Sc [23]. A student must also either complete 1 Specialist, 2 Majors or 1 Major and 2 Minors. The first five credits attempted by a student represent about 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.

Courses were taken by students in 71 different departments in the data set so the covariate vector is of length 142. The covariate identified simply by the department name represents the number of credits a student completed in that department. On the other hand, a department name followed by the letter **G** represents the average grade across all courses taken by a student in that department. This variable is set to 0 if no course were taken in that department.

Regarding the response variable, since it is assumed that students can take classes in other universities or faculties, every student who completed 18 credits are considered students who completed a program. Students who registered to 5 credits worth of courses and succeeded at fewer than 18 credits worth of courses are considered students who began a program but did not complete it. With this formatting, the first research question data set contains 38 842 students of which 26 488 completed an undergraduate program and 12 294 did not. The training set contains 90% of the total data set while both the validation and test set contains 5% each.

In order to answer the second research question, the 26 448 students who completed a program are kept. This time, the response will represent the choice of major made by the student. Since this information is not contained in the data set, the department in which a 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, one for each department. 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 our 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.

5.2 Algorithms

A classification tree using the Gini impurity as split measurement was coded in the C++ language using the Rcpp library [10]. The code is available upon request from the first author. The algorithm proceeds as explained in Section 3.1, the tree it produces is unpruned and regions are partitioned until they contain only 50 observations. Three versions of the random forest algorithm are going to be used. **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 400 trees but randomly selects the variable to be split upon in each region.

Finally, the popular R RandomForest package [16] 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 $p = \sqrt{m}$ inputs for each region. Linear models were fit for both of the classification problems serving as benchmarks.

6 Results

6.1 First research question : Predicting if a student completes an undergraduate program

Predicting if a student will complete their undergraduate program is the first task that the algorithm was asked to complete. Out of the 904 students who completed a program in the test set, the **random forest # 1** predicts accurately that 91.03% completed their program. Among the students who did not complete their program in the test set, the algorithm achieves a 49.92% accuracy. The result is a combined 77.75% accuracy on the entire test set.

Variable importance is measured using the permutation decrease since it is more reliable as explained in Section 4. The top 15 variables according to their mean decrease were kept and ordered in Figure 2. Boxplots were also used in order to visualize the variance of these measurements.

As expected the variance of the grade variables are a bit larger. A surprisingly low number of variables are statistically significant. It seems important to point out that among the top three grades related variables are Mathematics (MAT) and Economics (ECO) which are considered low-grade departments.

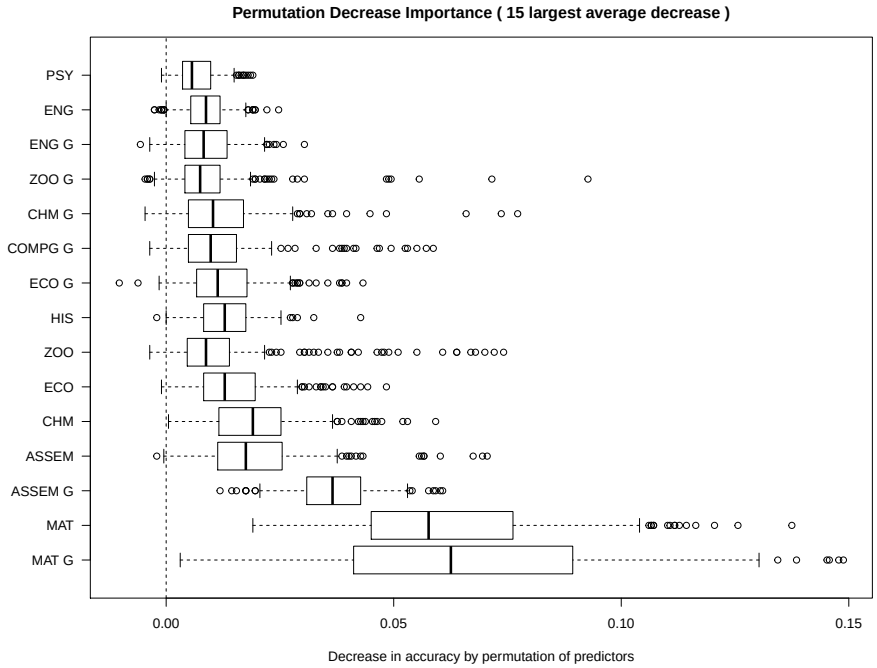

Figure 2: Variables importance boxplots for the **random forest # 1**. Predictors are ordered by their average decrease in accuracy from the bottom to the top in decreasing order.

Perhaps the strict marking of these departments helps in better distinguishing students among themselves. The ASSEM code represents a special type of first year seminar course. It seems that the type of student that registers in theses courses are easy to classify as both grades and the number of credits are considered important.

For **random forest # 2**, the algorithm achieves a 94.90% accuracy among the students who completed a program. Among the students who did not complete an undergraduate program in the test set, the algorithm predicts correctly that 42.74% did not complete their program. It achieves a 78.06% accuracy over the entire test set. Using the random forest where the split variable is randomly selected greatly increases the variance of the permutation decrease of all of the variables, this is noticeable in Figure 3.

Seminar department courses are important but also Mathematics, Zoology/Biology (ZOO), Chemistry (CHM) and Economics grades are important. As a matter of fact, the grades in departments are always of greater importance than their number of credit counterparts. Once again, most of these are among the low-grade departments of the university.

Finally, the RandomForest package was used for the **random forest # 3**. Among the students who completed their program in the test set, the package

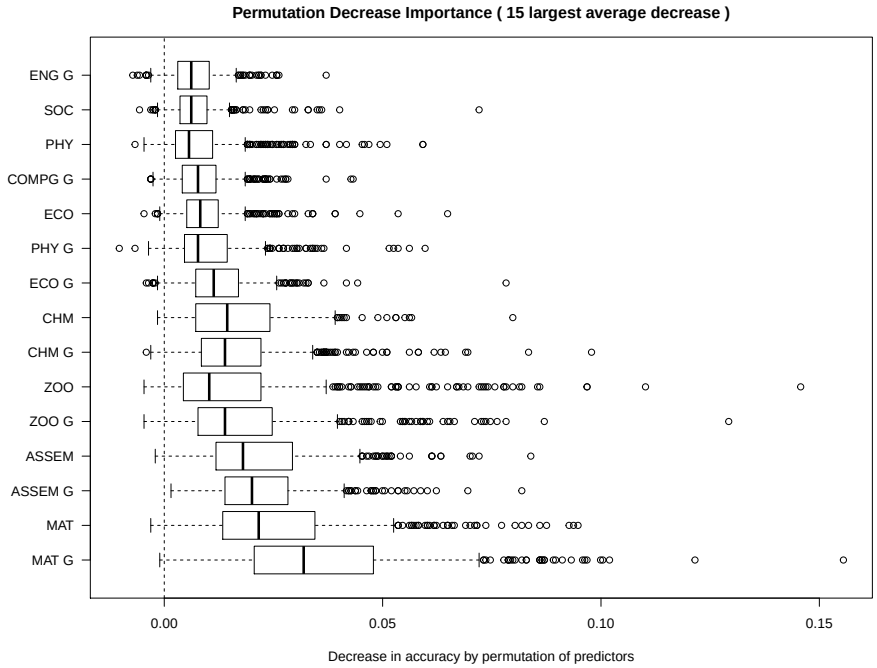

Figure 3: Variables importance boxplots for the **random forest # 2**. Predictors are ordered by their average decrease in accuracy from the bottom to the top in decreasing order.

achieves a 91.19% accuracy. Out of the 418 students who did not complete their program, the **random forest # 3** achieves a 52.95% accuracy. The combined result it a 78.84% accuracy over the complete test set.

Figure 4 contains the variable importance plot produced by the function included with the package. Not also the grades in Seminars, Mathematics and Economics are the top three grades variables but they are in truth the three most important variable according to this iteration of a random forest. The package does not provide the variance of these measurements. This is somewhat problematic since it is expected and has been observed in previous graphics that the grades variable importance measurements are likely to have a much larger variance than their credits counterpart and even though the grade variables seem more important they might not be significantly more important.

To summarize, theses different forests all achieve reasonable accuracy located around 78% which is slightly above the 74% accuracy achieved with a logistic regression. They all appear to overestimate the probability of completion as they all have great accuracy for predicting who will complete but low accuracy for predicting who will drop out.

For variable importance, the results are rather interesting as the average grade in a department seems to be almost consistently more important than

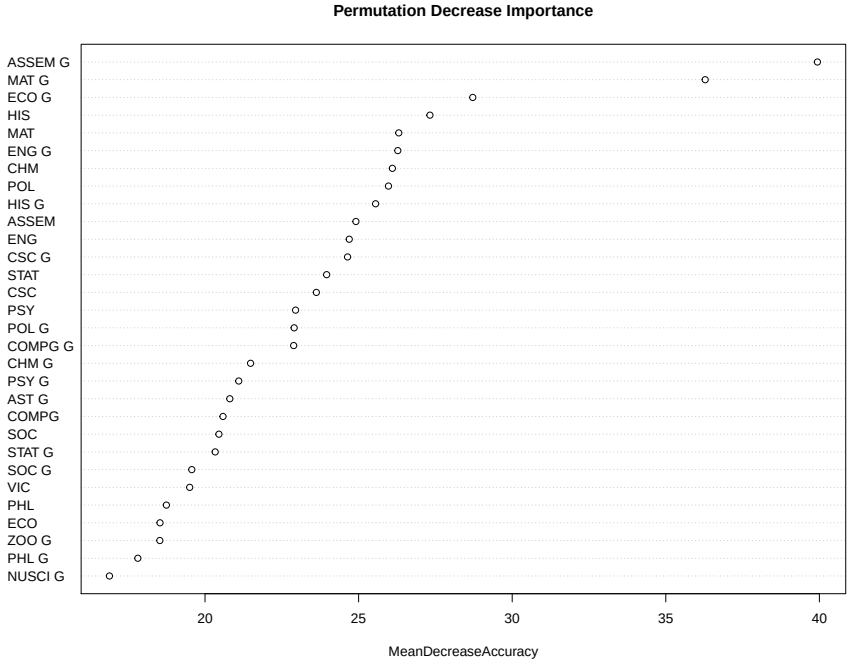

Figure 4: Variable importance plot produced by the RandomForest package for the **random forest # 3**. By default it contains the 30 variables with the highest mean decrease in accuracy.

the number of credit in that department. Grades from low-grade departments are always more important than their counterpart and in some cases, they were the only significant grades. 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.

6.2 Second research question : Predicting the major choice

The second task that random forests need to be complete is to predict the major chosen by a student. For the first implementation, **random forest # 1**, the algorithm predicts the correct major choice for 46.31% of the students in the test set. 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. Once again, the variable importance boxplots will be computed and looked at.

A decrease in importance for the grades variable is noted in Figure 5. This was to 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 chosen by the student these variable importance measures aren't surprising. Actually, if all the courses were included, the amount of credit in every department

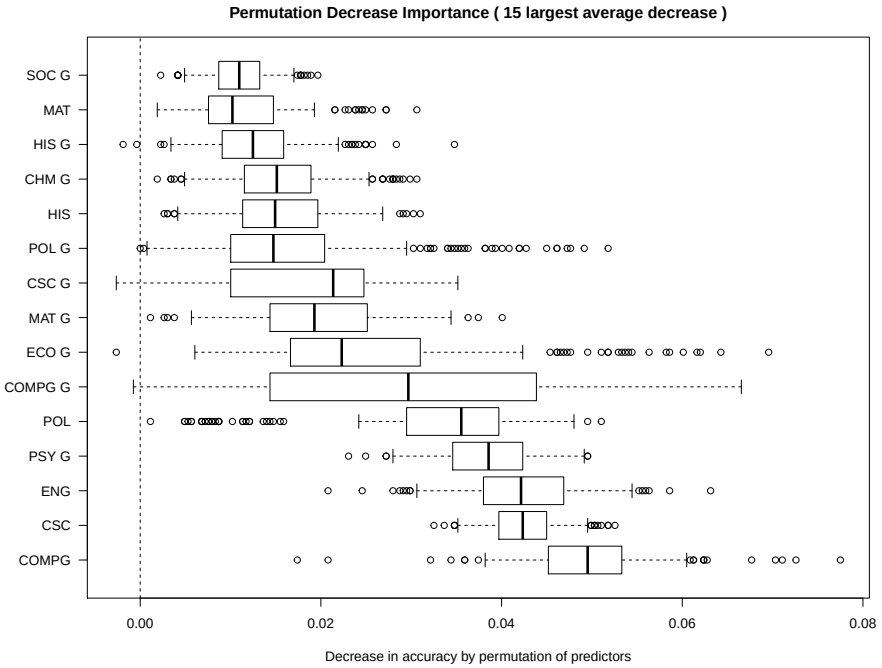

Figure 5: Variables importance boxplots for the **random forest # 1**.Predictors are ordered by their average decrease in accuracy from the bottom to the top in decreasing order.

precisely defines the response variable. Considering this weakness in the data formatting, the grades still have a relatively high importance.

Among the 4 most important grade variables, Mathematics, Economics and Finance (COMPG) are found, indicating the high importance of these grades, not only to predict if a student will succeed at a program or not, but to predict what he will major in. The three most important variables are the number of credits in the Finance department, the computer science (CSC) department and in the English (ENG) department. In the training set, it seems that only students registered in those programs actually take courses in those departments. Therefore, these three variables are really important to classify the students registered in these respective departments, while grades related variable affect more students but have a lesser importance overall.

Next, a random forest with random inputs is built as a classifier for the major choice. A 44.57% accuracy was achieved by the **random forest #2** on the test set. Figure 6 contains the 15 largest mean decreases in accuracy after permuting the predictors.

Here, in Figure 6, grades in Finance, Mathematics, Economics and Chemistry are 4 of the 5 most important grades. Once again the grade in Psychology

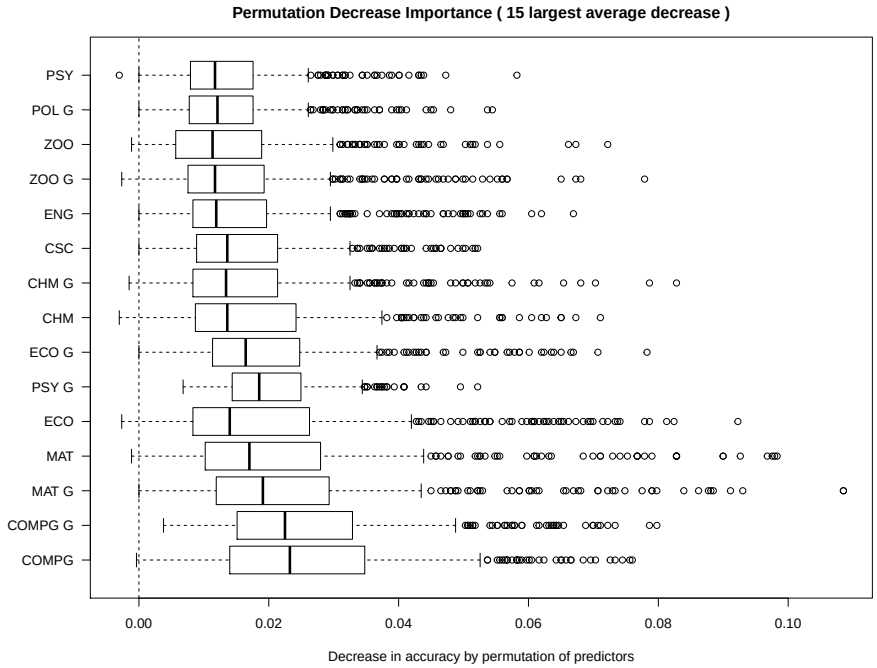

Figure 6: Variables importance boxplots for the **random forest # 2**.Predictors are ordered by their average decrease in accuracy from the bottom to the top in decreasing order.

(PSY) is the most important grade coming from a department considered high-grading by Sabot & Wakeman-Linn. Like in Figure 3, it seems like the random forest with random inputs produces variable importance measurements where it is harder to judge statistically significant any of the variables.

Using the R RandomForest package to produce the **random forest # 3**, a 47.41% accuracy in predicting the major choice is achieved. Once again, one time out of two, the algorithm selects the correct program within the list of 71 programs. Figure 7 contains the variable importance plot produced by the package.

Figure 7 contains results that aren't truly in line with the previous ones. Grades in Psychology is the most important variable while grades in Economics and Mathematics fall rather low. Once again, the lack of boxplots or any other detail regarding the variance makes this plot a bit less unreliable.

Overall, the results are less convincing for the prediction of the students' major choice. A linear multinomial model was fit and its accuracy is 42%. The three different forests had a prediction accuracy located around 45% which is a slight increase from the linear model. Across the three models the grades are slightly less important than they were for the completion prediction. Even

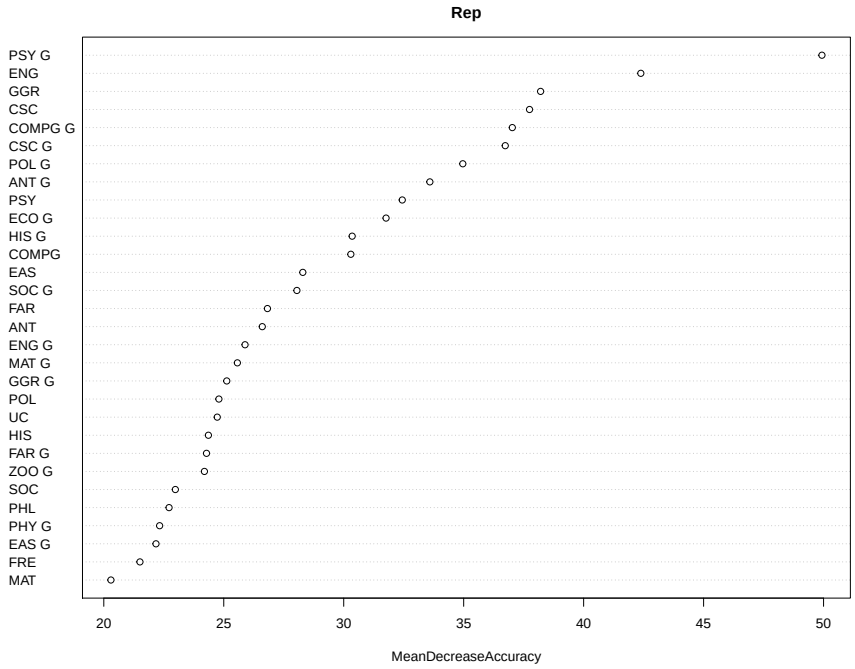

Figure 7: Variable importance plot produced by the RandomForest package for the **random forest # 3**. By default it contains the 30 variables with the highest mean decrease in accuracy.

though the grades in Mathematics and Economics are still among the most important they are a bit less important than in the previous set of classifiers. Some potential improvements regarding these classifiers are located in the conclusion.

7 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 choice of a student who completed a program. Random forests were used to build those classifiers. The prediction accuracy of those was evaluated but more importantly the importance of each predictor variable was assessed.

It was observed in Section 6 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, Finance, Biology and Chemistry are consistently among the most important variables. Grades in Psychology were also considered important in a lot of situations. These results indicate that a strong relationship exists between the grades in these 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 choice. The number of credits in Computer Science and Finance was important and it was noticed in the data set that almost every student who registered in courses in either of those two departments completed their program in that department. Even though they were less important, grades in Mathematics, Finance, Economics and Psychology 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 ([9],[7],[8]). The predictors were also problematic. Ideally, the algorithm would consider splitting on the grade variables for a certain department only for 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 splits should be easy to code and 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 splits and considering a multi-label classifier like suggested by Chen & al. [7] 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 or Canada.

References

[1] Bailey, M.A., Rosenthal, J.S., Yoon, A.H.: Grades and incentives: assessing competing grade point average measures and postgraduate outcomes. Studies in Higher Education **41**(9), 1548–1562 (2016). DOI 10.1080/03075079.2014.982528. URL `http://dx.doi.org/10.1080/03075079.2014.982528`

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

[3] Breiman, L.: Bagging predictors. Machine Learning **24**(2), 123–140 (1996). DOI 10.1007/BF00058655. URL `http://dx.doi.org/10.1007/BF00058655`

[4] Breiman, L.: Heuristics of instability and stabilization in model selection. Ann. Statist. **24**(6), 2350–2383 (1996). DOI 10.1214/aos/1032181158. URL `http://dx.doi.org/10.1214/aos/1032181158`

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

[6] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A. (1984)

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

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

[9] Clare, A., King, R.D.: Knowledge Discovery in Multi-label Phenotype Data, pp. 42–53. Springer Berlin Heidelberg, Berlin, Heidelberg (2001). DOI 10.1007/3-540-44794-6. URL http://dx.doi.org/10.1007/3-540-44794-6

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

[11] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2 edn. Springer (2009)

[12] Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics **15**(3), 651–674 (2006). DOI 10.1198/106186006X133933. URL http://dx.doi.org/10.1198/106186006X133933

[13] Johnson, V.E.: Grade Inflation : A Crisis in College Education. Springer (2003)

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

[15] Kononenko, I.: On biases in estimating multi-valued attributes. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95, pp. 1034–1040. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995). URL http://dl.acm.org/citation.cfm?id=1643031.1643034

[16] Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002). URL http://CRAN.R-project.org/doc/Rnews/

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

[18] Loh, W.Y., Shih, Y.S.: Split selection methods for classification trees. Statistica Sinica **7**, 815–840 (1997). URL http://www3.stat.sinica.edu.tw/statistica/j7n4/j7n41/j7n41.htm

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

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

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

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

[23] University of Toronto: Degree requirements (h.b.a., h.b.sc., bcom) (2017). URL http://calendar.artsci.utoronto.ca/Degree_Requirements_(H.B.A.,_H.B.Sc.,_BCom).html