Using Fractional Factorial Designs for Variable Importance in Random Forest Models

Ewa M. Sztendur and Neil T. Diamond

Abstract—Random Forests are a powerful classification technique, consisting of a collection of decision trees. One useful feature of Random Forests is the ability to determine the importance of each variable in predicting the outcome. This is done by permuting each variable and computing the change in prediction accuracy before and after the permutation. This variable importance calculation is similar to a one-factor-at-a time experiment and therefore is inefficient. In this paper, we use a regular fractional factorial design to determine which variables to permute. Based on the results of the trials in the experiment, we calculate the individual importance of the variables, with improved precision over the standard method. The method is illustrated with a study of student attrition at Monash University.

Keywords—Random Forests, Variable Importance, Fractional Factorial Designs, Student Attrition.

I. INTRODUCTION: ATTRITION AT MONASH UNIVERSITY

Student attrition is a major problem at most universities. It has a major economic effect on the university, as well as having a social and economic effect on the students. Monash University, the largest university in Australia, has over the last few years begun to address this problem. As part of this, an attempt to build a predictive model for student attrition at Monash was undertaken, summarized in this paper.

II. DATA AND ANALYSIS

A. Data

Student Attrition is only determined on an annual basis. A student is said to have discontinued if they were enrolled on the census date in one year, but were not enrolled on the census date in the following year, and had not graduated. Detailed enrolment records were available for the past 5 years. The major emphasis is on first year students, as this is the most common time when students drop out.

B. Analysis

Since the factors involved in attrition seemed to vary from faculty to faculty, and from stage to stage, separate models were built for each stage and faculty. The majority of students, coming from secondary schools, have an ATAR (Australian Tertiary Admission Rank) score, while students who come from non-traditional routes may not have an ATAR score. Separate models were built for students with and without ATAR scores. The models were based on Random Forests with three response categories: continued, graduated, and discontinued.

III. DECISION TREES AND RANDOM FORESTS

A. Decision Trees

Random Forests are a variation on decision trees. In a tree (see, for example, [3, pages 305–313]) the data is divided into homogeneous segments by sequentially choosing the best binary splits of the variables. The process is continued until no further splits are possible due to lack of data. The resultant tree is pruned using cross-validation.

An example of a decision tree for Stage 1 students with an ATAR score at one of the faculties at Monash is given in Figure 1. Trees are easy to understand and explain. However, a disadvantage of trees is that they are quite variable: a small change in the data leads to a quite different tree. One solution is Random Forests.

B. Random Forests

Random Forests ([2]), available from Salford Systems, are a powerful classification technique, consisting of a collection of decision trees. There is also an R ([7]) package randomForest ([4],[5]), which implements Breiman’s random forest algorithm (based on Breiman and Cutler’s original Fortran code), which has been used in this paper.

By default, Random Forests are constructed as follows: 500 bootstrap samples are taken from the original sample, and for each bootstrap sample a tree is constructed, but with only a sample of three variables used as candidate variables at each split. No pruning of the trees is done. To classify a new observation, the observation is classified by each of the 500 trees, and then a majority vote is used. The accuracy of each of the trees is assessed on the “out-of-bag” data, that is the data not used to build the tree.

Table I gives the confusion matrix for the default Random Forest model applied to the Stage 1 students with an ATAR score from one of the faculties at Monash University. Without

<table>
<thead>
<tr>
<th></th>
<th>Continue</th>
<th>Attrite</th>
<th>Grad</th>
<th>class.error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continue</td>
<td>1044</td>
<td>15</td>
<td>5</td>
<td>0.02</td>
</tr>
<tr>
<td>Attrite</td>
<td>182</td>
<td>39</td>
<td>2</td>
<td>0.83</td>
</tr>
<tr>
<td>Grad</td>
<td>93</td>
<td>2</td>
<td>5</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table I gives the confusion matrix for the default Random Forest model applied to the Stage 1 students with an ATAR score from one of the faculties at Monash University. Without
Fig. 1. A Decision Tree for Stage 1 students with an ATAR score at one of the faculties at Monash University. At branches, true statements go to the left and false statements go to the right. At terminal nodes, the count of continuing, attriting, and graduating students are given, as well as the predicted outcome. The tree shows that Stage 1 full-time students with an ATAR score attending Caulfield or Gippsland campuses for whom there is some information on their new to higher education status are more likely to continue, but that younger part-time students are more likely to attrite.

optimization, the Random Forest does not work well at all: although only 2% of continuing students are miss-classified, the miss-classification rate for attriting and graduating students are 83% and 95% respectively. The maximum miss-classification rate is 95% and this is used as a performance measure for the Random Forest. The problem, in this example, is that the three classes are represented unequally in the data. There are almost 10 times as many “continuing” students as there are “graduating” students.

IV. TUNING THE RANDOM FOREST

There are a number of parameters in the randomForest package that can be varied to improve the performance of the Random Forest. These are listed below with some details on how we have changed them from the default values:

- strata The samples were stratified by the response variable attrition.
- sampsize In a bootstrap sample about 63.2% of cases are represented at least once. Since there are 100 graduating cases, the strata sample sizes were chosen to be 63 continuing students, 63 attriting students, and 63 graduating students.
- replace Rather than sampling with replacement, sampling without replacement was done.
- cutoff The cutoff is a vector of length equal to the number of classes. From the help menu it is explained that

The ‘winning’ class for an observation is the one with the maximum ratio of proportion of votes to cutoff. Default is $1/k$ where $k$ is the number of classes (i.e., majority vote wins).

Examining the code, the sum of the cutoff values equals 1 i.e. $c = (c_1, c_2, c_3)$ where $0 < c_i < 1$ and $\sum c_i = 1$. Since the Random Forest is not a deterministic function, usual methods of optimization cannot be used. To find the best values of $c_i$, the following parameterization was used:

- $c_1 = \gamma_1$
- $c_2 = (1 - \gamma_1)\gamma_2$
- $c_3 = 1 - \gamma_1 - (1 - \gamma_1)\gamma_2$

where $0 < \gamma_1, \gamma_2 < 1$.

To optimize the random forest, a grid of values of $\gamma_1$ and $\gamma_2$ between 0.05 and 0.95 in steps of 0.05 was used and for each value of the grid, the Random Forest was applied and the maximum mis-classification rate was determined. A local weighted quadratic surface was fitted to this data, using the R function loess, and the convex hull of the points with fitted miss-classification rates less than the 5th percentile was determined. In the interior of the convex hull on a grid of steps of 0.01 the random
Forest was again applied and the maximum miss-classification was again determined. The two sets of results were combined and the loess surface was refitted and the values of $\gamma_1$ and $\gamma_2$ leading to the minimum fitted miss-classification rate was found.

The number of trees to grow was set at the default value of 500 although this could be increased if desired.

**ntree**  
This parameter is the number of variables randomly sampled as candidates at each split. The default value is the square root of the number of variables. We optimized the random forest for each value of mtry from 3 to 8, and then applied the optimum cutoffs twenty times. The boxplots of the calculated miss-classification rates are given in Figure 2. For this example, the best value of mtry was 6, with best values of $\gamma_1$ and $\gamma_2$ equal to 0.22 and 0.23, respectively, corresponding to a cutoff of $c = (0.22, 0.1794, 0.6006)$. The corresponding confusion matrix is given in Table II, with a maximum miss-classification rate of 47.1%, much improved over the maximum miss-classification rate shown in Table I.

For this example, the best value of mtry was 6, with best values of $\gamma_1$ and $\gamma_2$ equal to 0.22 and 0.23, respectively, corresponding to a cutoff of $c = (0.22, 0.1794, 0.6006)$. The corresponding confusion matrix is given in Table II, with a maximum miss-classification rate of 47.1%, much improved over the maximum miss-classification rate shown in Table I.

**V. VARIABLE IMPORTANCE**

One useful feature of Random Forests is the ability to determine the importance of each variable in predicting the outcome. This is done by permuting each variable and computing the prediction accuracy of the out-of-bag portion of the data before and after the permutation. An example of the variable importance calculation, together with the variables considered, is given in Table III.

In the table, there are five different importance measures. The first three are class-specific measures; for example, the miss-classification rates for attriting students when the campus variable is permuted is increased by 5%. Similar calculations are done for the other explanatory variables and for the other classes. The fourth measure is the average of the first three measures, while the fifth measure is the total decrease in node impurities, measured by the Gini index (see, for example, [3, page309]), from splitting on the variable, averaged over all trees.

Since our major interest here is attrition, we have focused on the second measure. It is possible to scale this measure by dividing by the standard error of the miss-classification increases over all the trees, but we prefer to focus on the unscaled measure. From the plot in Figure 3, it appears that there are two important explanatory variables, Campus and Attendance Type.

It is instructive to examine how this variable importance measure is done. The actual design, given in Table IV, is identical to a one-factor at a time design (see, for example, [6]), where a $-$ sign corresponds to not permuting the variable, and a $+$ sign corresponds to permuting the variable. The average miss-classification rate for attriting students is given in the column labeled $m_{22}$, while the calculated importance is given in the column labeled $i_2$. Note that the results in Tables 3 and 4 do not correspond exactly because the importance calculation is based on a random permutation. Similarly, note that the average miss-classification rate for the trees is higher than for the random forest itself.

The one-factor at a time design is known to be a very inefficient way of conducting experiments. It is much better to use fractional factorial designs (see, for example, [1, Chapter 61]) for this purpose. The key idea is to vary more than one factor at a time. The same idea was applied here: more than one variable at a time was permuted.

Table V shows a regular fractional factorial design involving the 13 factors in 16 runs. The first run corresponds to not permuting any of the variables, while the second run, for example, corresponds to permuting variables A, E, F, G, L, M, and N.

The design is easy to construct. Since there are 16 runs, columns A to D are every combination of $-$ and $+$ signs. The other columns are generated by multiplying columns A to D as follows:

$$
E = -AB \\
G = -AD \\
J = -BD \\
L = ABC \\
N = ACD
$$
### TABLE III

**VARIABLE IMPORTANCE MEASURES**

<table>
<thead>
<tr>
<th></th>
<th>Continue</th>
<th>Attrite</th>
<th>Grad</th>
<th>Mean Decrease</th>
<th>Mean Decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i_1$</td>
<td>$i_2$</td>
<td>$i_3$</td>
<td>$i_4$</td>
<td>$i_5$</td>
</tr>
<tr>
<td>ATT_MODEC</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.27</td>
</tr>
<tr>
<td>ATT_TYPE</td>
<td>0.014</td>
<td>0.018</td>
<td>-0.009</td>
<td>0.014</td>
<td>4.72</td>
</tr>
<tr>
<td>nhe</td>
<td>-0.009</td>
<td>0.003</td>
<td>0.441</td>
<td>0.005</td>
<td>26.72</td>
</tr>
<tr>
<td>AGE</td>
<td>0.013</td>
<td>-0.011</td>
<td>0.285</td>
<td>0.003</td>
<td>21.80</td>
</tr>
<tr>
<td>OS</td>
<td>0.001</td>
<td>-0.000</td>
<td>0.002</td>
<td>0.001</td>
<td>0.54</td>
</tr>
<tr>
<td>DISAB</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
<td>0.001</td>
<td>1.94</td>
</tr>
<tr>
<td>GENDER</td>
<td>-0.001</td>
<td>0.005</td>
<td>0.010</td>
<td>0.000</td>
<td>5.04</td>
</tr>
<tr>
<td>sch_type</td>
<td>-0.004</td>
<td>0.005</td>
<td>0.003</td>
<td>-0.003</td>
<td>8.33</td>
</tr>
<tr>
<td>REGION</td>
<td>-0.001</td>
<td>-0.000</td>
<td>0.002</td>
<td>-0.000</td>
<td>2.45</td>
</tr>
<tr>
<td>SES2</td>
<td>-0.003</td>
<td>0.003</td>
<td>0.000</td>
<td>-0.002</td>
<td>7.31</td>
</tr>
<tr>
<td>ATAR</td>
<td>0.000</td>
<td>-0.006</td>
<td>0.001</td>
<td>-0.000</td>
<td>31.73</td>
</tr>
<tr>
<td>CAMPUS</td>
<td>0.008</td>
<td>0.050</td>
<td>0.002</td>
<td>0.013</td>
<td>6.29</td>
</tr>
</tbody>
</table>

The negative signs for $E$ to $K$ ensure that the first row of the design has all $-$ signs, corresponding to not permuting any variable.

The average miss-classification rate for the trees is given in the column labeled $m^2$. The importance of each variable is calculated by obtaining the difference between the means for the runs where the variable is permuted and for the runs where the variable is not permuted. For $N$, (Campus), the calculated

![Variable Importance Diagram](image)
variable importance is 0.0515.

to determine the “real” effects, the calculated importances can be plotted on a Daniel plot ([1, page 203]). The Daniel plot in Figure 4 shows that \(B\) (Attendance Mode) and \(N\) (Campus) certainly seem to be important variables. In this case, the same conclusions have been reached using either Figure 3 or Figure 4. However, Figure 4 is much more conclusive because the variability of the variable importance effects is much less using a fractional factorial design than using a one-factor-at-a-time design. Assuming a variance of \(\sigma^2\), the one-factor-at-a-time importance measure has a variance of \(2\sigma^2\), while the fractional factorial design measure has a variance of \(\sigma^2/4\) at a cost, in this case, of two extra permutations.

![Daniel Plot](image)

**Fig. 4.** Daniel Plot based on regular fractional factorial design. The plot shows that \(B\) (Attendance Mode) and \(N\) (Campus) certainly seem to be important variables.

### VI. Concluding Remarks

Random Forests are powerful methods but require some form of optimization for them to work well. We have used stratified sampling as well as optimizing the cutoff values in order to improve the performance of the Random Forest. The variable importance calculation used in random forests is based on a one-factor-at-a-time experiment. The ideas of fractional factorial designs can be used to improve this calculation with very little additional effort.

### References


