Using Fractional Factorial Designs for Variable Importance in Random Forest Models

Ewa. M. Sztendur and Neil T. Diamond

Open Science Index, Physical and Mathematical Sciences Vol:6, No:11, 2012 publications.waset.org/3038.pdf

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 began 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
Confusion Matrix for default Random Forest model for Stage 1 students with an ATAR score at one of the faculties at Monash University.

	Continue	Attrite	Grad	class.error
Continue	1044	15	5	0.02
Attrite	182	39	2	0.83
Grad	93	2	5	0.95

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

Ewa Sztendur is with the Office of the Pro Vice-Chancellor (Academic & Students) and with the Office of the Pro Vice-Chancellor (Research & Research Training) at Victoria University, Melbourne, Australia. e-mail: ewa.sztendur@vu.edu.au.

Neil Diamond is with the Department of Econometrics and Business Statistics at Monash University, Melbourne, Australia and also with the the Office of the Pro Vice-Chancellor (Research & Research Training) at Victoria University, Melbourne, Australia.

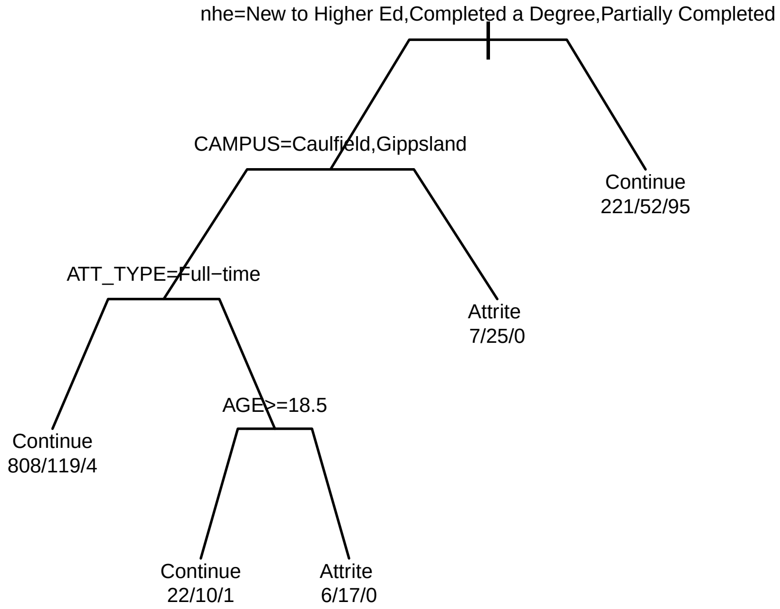

Fig. 1. A Decision Tree for Stage 1 students with an ATAR score at one of the faculties at Monash University. At branches, true statements go to the left and false statements go to the right. At terminal nodes, the count of continuing, attriting, and graduating students are given, as well as the predicted 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:

$$\begin{aligned} c_1 &= \gamma_1 \\ c_2 &= (1-\gamma_1)\gamma_2 \\ c_3 &= 1-\gamma_1-(1-\gamma_1)\gamma_2 \end{aligned}$$

where

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

To optimize the random forest, a grid of values of γ_1 and γ_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 miss-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 γ_1 and γ_2 leading to the minimum fitted miss-classification rate was found.

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

mtry 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 γ_1 and γ_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.

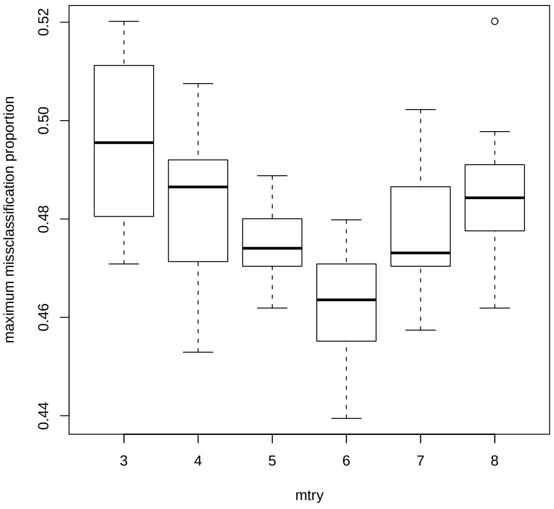

Fig. 2. Box-plots of maximum missclassification proportions for various values of `mtry`

TABLE II
CONFUSION MATRIX FOR OPTIMIZED RANDOM FOREST MODEL FOR STAGE 1 STUDENTS WITH AN ATAR SCORE AT ONE OF THE FACULTIES AT MONASH UNIVERSITY.

	Continue	Attrite	Grad	class.error
Continue	583	368	113	0.452
Attrite	84	118	21	0.471
Grad	6	26	68	0.320

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_2, 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 6]) 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:

$$\begin{array}{rclrcl}
E & = & -AB & F & = & -AC \\
G & = & -AD & H & = & -BC \\
J & = & -BD & K & = & -CD \\
L & = & ABC & M & = & ABD \\
N & = & ACD & & &
\end{array}$$

TABLE III
VARIABLE IMPORTANCE MEASURES

	Continue	Attrite	Grad	Mean Decrease Accuracy	Mean Decrease Gini
	i_1	i_2	i_3	i_4	i_5
ATT_MODC	-0.000	0.000	0.000	-0.000	0.27
ATT_TYPE	0.014	0.018	-0.009	0.014	4.72
nhe	-0.009	-0.003	0.441	0.005	26.72
ABOR	-0.000	0.000	-0.000	-0.000	0.09
AGE	-0.013	-0.011	0.285	-0.003	21.80
OS	0.001	-0.000	0.002	0.001	0.54
DISAB	0.000	0.002	0.000	0.001	1.94
GENDER	-0.001	0.005	0.010	0.000	5.04
sch_type	-0.004	0.005	0.003	-0.003	8.33
REGION	-0.001	-0.000	-0.002	-0.001	2.45
SES2	-0.003	0.003	-0.000	-0.002	7.31
ATAR	0.000	-0.006	0.001	-0.000	31.73
CAMPUS	0.008	0.050	0.002	0.013	6.29

A	Attendance Mode
B	Attendance Type
C	New to Higher Education
D	Aboriginal and Torres Strait Islander Status
E	Age
F	Overseas Status
G	Disability Status
H	Gender
J	School Type
K	Region
L	Socio-economic status
M	ATAR score
N	Campus

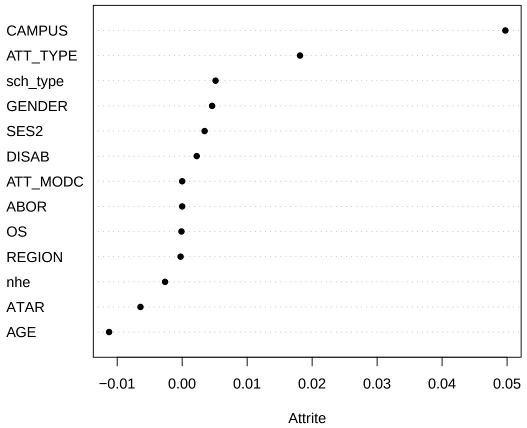

Fig. 3. Variable Importance diagram for Stage 1 students in the Art and Design faculty at Monash University.

TABLE IV
ONE-FACTOR AT A TIME DESIGN FOR 13 FACTORS. VARIABLES WHICH ARE NOT TO BE PERMUTED ARE DENOTED BY $-$, WHILE VARIABLES TO BE PERMUTED ARE DENOTED BY $+$. THE MEAN MISS-CLASSIFICATION RATE FOR ATTRITING STUDENTS IS GIVEN IN THE COLUMN LABELED m_2, WHILE THE CALCULATED IMPORTANCE IS GIVEN IN THE COLUMN LABELED i_2.

A	B	C	D	E	F	G	H	J	K	L	M	N	m_2	i_2
−	−	−	−	−	−	−	−	−	−	−	−	−	0.548	
+	−	−	−	−	−	−	−	−	−	−	−	−	0.548	0.000
−	+	−	−	−	−	−	−	−	−	−	−	−	0.565	0.017
−	−	+	−	−	−	−	−	−	−	−	−	−	0.545	−0.003
−	−	−	+	−	−	−	−	−	−	−	−	−	0.548	0.000
−	−	−	−	+	−	−	−	−	−	−	−	−	0.534	−0.014
−	−	−	−	−	+	−	−	−	−	−	−	−	0.548	0.000
−	−	−	−	−	−	+	−	−	−	−	−	−	0.550	0.002
−	−	−	−	−	−	−	+	−	−	−	−	−	0.553	0.005
−	−	−	−	−	−	−	−	+	−	−	−	−	0.551	0.003
−	−	−	−	−	−	−	−	−	+	−	−	−	0.547	−0.001
−	−	−	−	−	−	−	−	−	−	+	−	−	0.549	0.001
−	−	−	−	−	−	−	−	−	−	−	+	−	0.542	−0.006
−	−	−	−	−	−	−	−	−	−	−	−	+	0.598	0.050

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

TABLE V
FRACTIONAL FACTORIAL DESIGN FOR 13 FACTORS. VARIABLES WHICH ARE NOT TO BE PERMUTED ARE DENOTED BY −, WHILE VARIABLES TO BE PERMUTED ARE DENOTED BY +. THE MEAN MISS-CLASSIFICATION RATE FOR ATTRITING STUDENTS IS GIVEN IN THE COLUMN LABELED m_2.

A	B	C	D	E	F	G	H	J	K	L	M	N	m_2
−	−	−	−	−	−	−	−	−	−	−	−	−	0.548
+	−	−	−	+	+	+	−	−	−	+	+	+	0.592
−	+	−	−	+	−	−	+	+	−	+	+	−	0.566
+	+	−	−	−	+	+	+	+	−	−	−	+	0.633
−	−	+	−	−	+	−	+	−	+	+	−	+	0.586
+	−	+	−	+	−	+	+	−	+	−	+	−	0.549
−	+	+	−	+	+	−	−	+	+	−	+	+	0.630
+	+	+	−	−	−	+	−	+	+	+	−	−	0.574
−	−	−	+	−	−	+	−	+	+	−	+	+	0.596
+	−	−	+	+	+	−	−	+	+	+	−	−	0.542
−	+	−	+	+	−	+	+	−	+	+	−	+	0.625
+	+	−	+	−	+	−	+	−	+	−	+	−	0.564
−	−	+	+	−	+	+	+	+	−	+	+	−	0.541
+	−	+	+	+	−	−	+	+	−	−	−	+	0.598
−	+	+	+	+	+	+	−	−	−	−	−	−	0.590
+	+	+	+	−	−	−	−	−	−	+	+	+	0.626

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 σ^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.

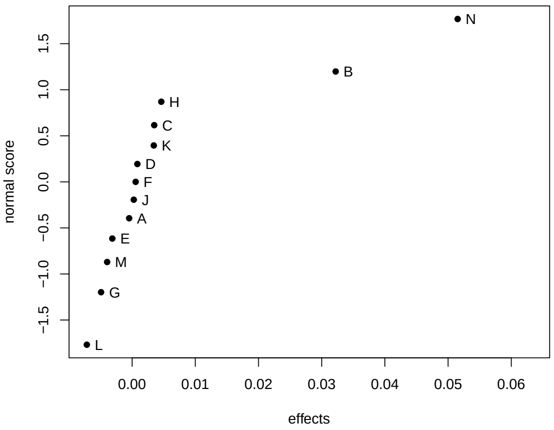

Fig. 4. Daniel Plot based on regular fractional factorial design. The plot shows that B (Attendance Mode) and N (Campus) certainly seem to be important variables.

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

[1] Box, G.E.P. and Hunter, J.S. and Hunter, W.G., *Statistics for Experimenters*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, 2005.

[2] Breiman, L. and Cutler, A., "Random Forests", Salford Sytems, www.salfordsystems.com, 2008.

[3] Hastie, T. and R.Tibshirani and J.Friedman, *The Elements of Statistical Learning,* 2nd. Ed., New York: Springer, 2009.

[4] Liaw, A. and M.Wiener, "Classification and Regression by randomForest", *R News*, **2**(3), 18-22, 2002.

[5] Liaw, A. and M.Wiener, *randomForest: Breiman and Cutler's random forests for classification and regression*, R package version 4.6-12., http:/CRAN.R-project.org/package=randomForest, 2012.

[6] Margolon, B.H., "Results on factorial designs of resolution IV for the 2^n and 2^n3^m series", *Technometrics*, **10**, 431-444, 1969.

[7] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org, 2012.

Ewa Sztendur Dr Ewa Sztendur is a Senior Statistical Consultant with the Office of the Pro Vice-Chancellor (Research and Research Training) and a Research Fellow with the Office of the Pro Vice-Chancellor (Academic and Students) at Victoria University, Melbourne, Australia. She has a PhD in statistics in the area of experimental design. Dr Sztendur has consulted and lectured in statistics at the University of Melbourne, Monash University and Victoria University in Australia. She has been involved in numerous research and evaluation projects in the field of education in the capacity of a statistical expert.

Neil Diamond Neil Diamond is Associate Professor and Co-ordinator of Statistical Support at Victoria University, Melbourne, Australia and Director of Statistical Consulting in the Department of Econometrics and Business Statistics at Monash University, Melbourne, Australia. He has over 30 years experience as a statistician having worked in industry, business and academia and has a Ph.D. in experimental design from the University of Melbourne. Neil has served as president of the Victorian Branch of the Statistical Society of Australia and is an accredited statistician of the Statistical Society of Australia.