CLASSIFICATION TREES OUTPERFORM LOGISTIC REGRESSION PREDICTIONS OF ATTRITION IN THE U.S. MARINE CORPS

BY

JUAN MANUEL ALZATE VANEGAS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Adviser:

Professor Fritz Drasgow

ABSTRACT

The present study compared the performance of machine learning classification models against logistic regression in the context of predicting training attrition from the Delayed Enlistment Program in the United States Marine Corps (UMSC) with scores from the Tailored Adaptive Personality Assessment System (TAPAS). The base-rate of attrition was low which made the model training process difficult, but the random-forest model outperformed logistic regression in predicting cases of attrition in a stratified 50% attrition sample.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The attrition rate of new recruits is a metric of high importance for efficient resource management in the U.S. Military, namely because it is associated with wasted time and resources during training. For this reason, all branches of the U.S. Military maintain records of the attrition rates and recruiter performance for their training programs and employ a variety of methods to predict and minimize attrition as part of an effort to cut unnecessary costs (Halstead, 2009).

Loosely speaking, this process involves inputting an array of recruit data (e.g., demographics, personality/cognitive test scores, etc.) into a classifier algorithm to predict whether the individual will indeed follow through with training until completion. However, attrition is a dichotomous decision variable (i.e., *did attrit*, *did not attrit*) with a low base-rate (i.e., observed instances of attrition are relatively infrequent, about 12% in the sample analyzed in this study), which can make classification with generalized linear models (GLMs) (i.e., a probit or logit regression model) difficult due to large fluctuations in the point-biserial correlations between low base-rate dichotomous variables and continuous predictors (e.g., Berkson's fallacy).

A popular approach to classification is statistical learning ("*machine learning*"), which includes algorithms that can sometimes yield greater predictive accuracy than GLMs (Vijayakumar & Cheung, 2018). The present study explored the feasibility of incorporating facet scores from the Tailored Adaptive Personality Assessment System (TAPAS) into a predictive model of U.S. Marine recruit attrition using machine learning classification techniques.

CHAPTER 2: METHOD

Participants

Records from a total of 39,043 recruits to the United States Marine Corps (USMC) were analyzed. Of these, only 238 (0.61%) records contained missing data for at least one variable. The data were standardized and missing values were imputed during preprocessing using *k*-nearest neighbor (*k*NN) imputation (Bokhari & Hubert, 2018; Witten, Frank, & Hall, 2011).

The sample was 89.13% male and the mean age was 22.09 (*SD* = 2.40 years). Demographic statistics are displayed in *Tables 2.1* and *2.2*. Finally, the recruits' target role in the USMC and reason for discharge is reported in *Table 2.3*. Notably, not every target role is represented by a case of attrition from the DEP program, and the proportion of each type of case differed by target role according to the likelihood-ratio test, $\chi^2(144, N = 39{,}043) = 1{,}143.55$, $p < .0001$, Cramér's $V = .06$.

Table 2.1: Sample demographics.

	Male	Female	Other			
BR	12.25%	13.72%	0.00%	Attrition	BR	Totals
African American	2,750	487		404 (8.18%)	12.48%	3,237 (8.29%)
American Indian	154	25		25 (0.51%)	13.97%	179 (0.46%)
Asian	994	104	1	169 (3.42%)	15.38%	1,099 (2.81%)
Biracial	767	117		137 (2.78%)	15.50%	884 (2.26%)
Caucasian	22,006	2,500		3,728 (75.53%)	15.21%	24,506 (62.77%)
Hawaiian/Pacific Islander	205	117		30 (0.61%)	13.10%	322 (0.59%)
Unknown	7,865	969		434 (8.79%)	4.91%	8,834 (22.63%)
Declined to respond	59	16		9 (0.18%)	12.00%	75 (0.19%)
Attrition	4,354 (88.21%)	582 (11.79%)	 (0.00%)			4,936
Total	34,800 (89.13%)	4,242 (10.87%)	1 (<.01%)			39,043

BR refers to the base rate of attrition within each group.
The values along the Attrition row and column refer to the proportions of each group represented in the subsample of USMC recruits who are discharged from the DEP program.

Instruments

Demographic records

Recruit age and of years of education were specified as covariates in each of the models.

AFQT score

The Armed Forces Qualification Test (AFQT) score is the percentile score obtained from four subtests (i.e., *Arithmetic Reasoning*, *Mathematics Knowledge*, *Word Knowledge*, and *Paragraph Comprehension*) of the Armed Services Vocational Aptitude Battery (ASVAB), the test commonly used for assessing recruit enlistment eligibility by branches of the U.S. Military (Drasgow, Embretson, Kyllonen, & Schmitt, 2006).

Table 2.2 Means, standard deviations, and correlation matrix.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1) Attrition	N/A	[-.16, .06]	[-.16,- .12]	[2.27, 3.39]	[-.03, -.00]	[-*.02*, *.00*]	[-.10, -.06]	[-*.01*, *.02*]	[-*.01*, *.02*]	[-*.01*, *.02*]	[.00, .03]	[-*.01*, *.01*]	[-.08, -.04]	[.01, .03]	[*.01*, *.01*]	[-*.01*, *.02*]	[-.03,- .00]	[-*.02*, *.01*]	[-*.02*, *.01*]
(2) Age	-.03**	22.09 (2.40)																	
(3) Years of education	-.05**	.66**	11.69 (087)																
(4) AFQT	.04**	.06**	.15**	59.29 (21.34)															
TAPAS facets																			
(5)*Achievement*	-.01*	.09**	.09**	.01**	0.14 (0.50)														
(6)*Adjustment*	*-.01*	.02*	.02**	.08**	.17**	0.05 (0.61)													
(7)*Commitment to serve*	*.01*	-.02**	-.03**	-.09**	.25**	.19**	0.05 (0.61)												
(8)*Courage*	*.00*	*.01*	.02**	.11**	.33**	.27**	.32**	0.36 (0.55)											
(9)*Dominance*	*.00*	.06**	.08**	.08**	.29**	.14**	.18**	.26**	0.27 (0.50)										
(10)*Even-tempered*	*.00*	.08**	.07**	.07**	.14**	.24**	.09**	.13**	.01*	0.32 (0.51)									
(11)*Ingenuity*	.01*	.07**	.06**	.06**	.19**	.09**	.03**	.17**	.24**	.03**	-0.02 (0.50)								
(12)*Optimism*	*-.00*	.02*	.04**	.05**	.17**	.28**	.13**	.16**	.11**	.19**	.05**	0.17 (0.44)							
(13)*Physical condition*	-.04**	*-.01*	.03**	.02**	.24**	.09**	.15**	.20**	.18**	-.02**	.06**	.08**	0.26 (0.55)						
(14)*Responsibility*	.01**	.07**	.07**	.14**	.32**	.17**	.18**	.31**	.21**	.19**	.09**	.17**	.13**	0.30 (0.47)					
(15)*Selflessness*	*-.00*	.01*	.02**	-.05**	.17**	-.02**	.06**	.07**	.08**	.11**	.07**	.08**	.03**	.16**	0.03 (0.42)				
(16)*Sociability*	*.00*	*.02*	.01**	-.13**	.14**	.12**	.11**	.09**	.25**	.03**	.18**	.11**	.07**	.04**	.14**	-0.35 (0.56)			
(17)*Team orientation*	-.01*	.02**	.03**	-.09**	.08**	.04**	.07**	.06**	.10**	.10**	.02**	.07**	.05**	.05**	.12**	.22**	-0.09 (0.45)		
(18)*Tolerance*	*-.00*	.07**	.08**	.04**	.06**	.04**	*.01*	.06**	.03**	.14**	.12**	.06**	.01**	.05**	.21**	.15**	.10**	0.001 (0.51)	
(19) *Virtue*	*-.00*	.10**	.07**	*<.01*	.27**	.07**	.18**	.19**	.08**	.23**	.04**	.14**	*<.01*	.28**	.25**	.04**	.07**	.12**	0.59 (0.22)

*p < .05 **p < .01 are bolded. Statistically non-significant correlation coefficients and mean differences are italicized.
Note: For row (1), the values in brackets are 95% confidence intervals for the pairwise mean difference (relative to the no-attrition group) using the Welch two-sample t-test. Analyses were repeated using the corresponding nonparametric test, the Wilcoxon rank-sum-test, and results were identical.

TAPAS

The Tailored Adaptive Personality Assessment System (TAPAS; Drasgow, Stark, Chernyshenko, Nye, & Hulin, 2012) was developed as a large-scale, fake-resistant assessment of the Big Five taxonomy of personality (including 21 lower-order personality facets and physical condition), using computer adaptive testing (CAT) techniques to reduce the risk of test exposure and compromise. The test has shown promise in predicting important outcomes in military applications, such as in-role performance and recruiter performance (Drasgow et al., 2012; Nye, White, Horgen, Drasgow, Stark, & Chernyshenko, 2018).

Attrition

Attrition was coded as a dichotomous outcome variable to indicate whether the recruit had ultimately been discharged from the Delayed Enlistment Program (DEP), regardless of the target role or the reason for discharge. Originally, we attempted to predict the reason for discharge as a multinomial outcome variable, but the lack of observations in several of the reason categories resulted in several of the machine learning models (i.e., CART, random forests) making no classifications into the attrition group, even with stratification.

Table 2.3: Recruits' target role in the USMC and reason for discharge from the DEP program.

	Marine Regular	Marine Corps Regular	Marine Reserve	Marine Corps Reserve	Marine Non-applicant	Marine Corps Non-applicant	Merchant Marines	
BR	15.03%	4.47%	18.91%	8.71%	0.59%	0.84%	0%	Totals
Attrition	3,876 (78.53%)	335 (6.79%)	624 (12.64%)	97 (1.97%)	3 (0.06%)	1 (0.02%)	(0%)	4,936 (12.64%)
Apathy/personal problem	1,586	107	248	37	3			1,981 (40.13%)
Officer program	406	45	39	9				499 (10.11%)
(Non-EPTS) Medical disqualification	286	45	95	21		1		448 (9.08%)
Enlisted in another service	355	17	55	5				432 (8.75%)
Dependency disqualification	278	17	36	6				337 (6.83%)
(EPTS) Moral disqualification	154	13	40	7				214 (4.34%)
Disqualified for option	179	14	15	1				209 (4.23%)
Marriage	101	24	10	3				138 (2.80%)
Temporarily disqualified	112	4	10	2				128 (2.59%)
Death	95	13	13	1				112 (2.47%)
Did not report on ship date	71	15	14	2				102 (2.07%)
Failed to graduate	46	1	9					56 (1.13%)
Religious training/appointment	44	6	4					54 (1.09%)
Enlistment misunderstanding	28	3	18	1				50 (1.01%)
Pursuit of higher education	30	2		21				35 (0.71%)
Recruiting error	27							27 (0.55%)
Refused to enlist	14		6					20 (0.41%)
(EPTS) Medical delinquency	16	2						18 (0.36%)
(Non-EPTS) Moral disqualification	9	2	4	1				16 (0.32%)
Personal hardship	12		1					13 (0.26%)
Component code change	10	1						11 (0.22%)
Pregnancy	8		3					11 (0.22%)
Positive direct antiglobulin test (DAT)	4	3	1					8 (0.16%)
Other	4	1	1					6 (0.12%)
No attrition	21,905	7,166	3,299	1,114	503	119	2	34,108 (87.36%)
Totals	25,780 (66.03%)	7,501 (19.21%)	3,923 (10.05%)	1,211 (3.10%)	506 (1.30%)	120 (0.31%)	2 (0.01%)	39,043

BR refers to the base rate of attrition within each target role.

The percentages along the Attrition row are the proportion of cases of attrition from recruits seeking the corresponding role, while the percentage in the Attrition row of the totals column is the overall base rate.

Note: EPTS refers to conditions that existed prior to military service.

Classifiers

CART model

Classification and regression trees (CART) are a family of statistical learning models that are relatively easy to interpret, but often suffer from high variance and low accuracy (James, Witten, Hastie, & Tibshirani, 2013). The classification tree model predicts a categorical outcome variable by using a set of predictor variables to generate a binary decision tree via recursive partitioning of the dataset (Breiman, Friedman, Olshen, & Stone, 1984).

The CART model is generated by a top-down, greedy algorithm that splits the dataset into two disjoint subsets based on the values of one predictor. For each predictor, we then consider a number of possible splits to branch off. For a dataset with *n* observations, there are $n - 1$ possible splits between adjacent values of a continuous predictor variable, while for a categorical variable with *c* categories, there are 2^{c-1} possible splits. The split that is selected to form a new tree branch is the one that minimizes node impurity—a global measure of error—of the leftover sets. Continuous variables use the sum of squared errors within groups as a measure of node impurity, while categorical variables use the sum of the Gini diversity indexes (gdi) obtained from the proportions of groups formed by the split: $gdi = 1 - \sum_{c=1}^{C} p_c^2$, where p_c indicates the proportion of the sample observed in the *c*th category of the outcome variable (Witten et al., 2011).

Moreover, the CART algorithm is a *stagewise* (i.e., “*myopic*”) greedy algorithm, which means it selects the best possible split (i.e., the one that generates the least node impurity) at every iteration, and once a split is made the algorithm does not revisit it. As a result, the final tree structure given by the algorithm is not guaranteed to be an optimal solution (Breiman et al., 1984). Note that the Gini indexes are maximized when group proportions are equal, and minimized when the proportion of one class is 100%. Thus, the base-rates of the subsets have direct influence over the generation of the classification tree structure at every iteration of the algorithm.

Of course, as with other forms of supervised learning, there is a dilemma between *overfitting* and *underfitting* the model to the training sample. An *underfitted* model produces an overly shallow, inaccurate tree. On the other hand, excessively complex trees are associated with increased resubstitution error, as well as *shrinkage* (i.e., increased testing error) (Witten et al., 2011). Attempting to fit saturated trees onto a dataset produces potentially spurious results from excessively large and volatile estimates of the irreducible error component of mean-squared error (MSE) (i.e., the variance of the error residuals; $\mathcal{V}(\boldsymbol{\varepsilon})$) obtained during training, as well as unstable tree structures generated by the learning algorithm; however, *overfitting* may even occur in model selection (Bokhari & Hubert, 2018; Cawley & Talbot, 2010; Hastic et al., 2009; Yarkoni & Westfall, 2017). The effect of overfitting is comparable to using a less powerful machine learning algorithm (Cawley & Talbot, 2010).

There are a variety of methods for managing the problems caused by overfitting, such as regularization or early stopping rules (Cawley & Talbot, 2010). In the classification tree algorithm, for example, the resulting tree can be simplified ("*pruned*") back to an optimal level using a training control procedure (Witten et al., 2011). *k*-fold cross-validation (CV) is a type of training control procedure that involves sampling observations without replacement into *k* equally-sized subsets (without overlap) by using each subset as the testing set and the remainder as the training set. The statistical learning model estimated from all but one of the subsets is then applied to each of the *k* subsets to yield an average estimate of testing error (Bokhari & Hubert, 2018; Witten et al., 2011, Chapter 5; Yarkoni & Westfall, 2017).

The most complex form of this procedure, *n*-fold cross-validation (also called leave-one-out cross-validation), generates models using each individual observation as the testing sample (i.e., *k* = *n*) (Bokhari & Hubert, 2018). This procedure is approximately unbiased and is the most efficient *k*-fold CV procedure; however, it is also the most computationally-expensive, does not allow for stratification, and may not work well for all datasets (Bokhari & Hubert, 2018). For instance, classifying datasets with unbalanced group proportions may be particularly difficult to achieve, especially if the training set contains fewer than 5 observations in each cell. Thus, stratification is sometimes combined with CV procedures to create models using subsets with equivalent proportions of each group; however, this procedure slightly reduces estimated error variance (Breiman, Friedman, Olshen, & Stone, 1984; Yarkoni & Westfall, 2017).

A good compromise is stratified 10-fold CV, which retains some of the most desirable properties, but with slightly smaller variance and larger bias than the leave-one-out CV procedure (Bokhari & Hubert, 2018; Breiman & Spector, 1992; Breiman et al., 1984). As mentioned, both *k*-fold CV and stratification can be applied to other types of classifiers besides CART.

Random forests

Sometimes, certain statistical learning methods like decision trees suffer from too much variance to such an extent that the model may exhibit differential performance when fit onto two randomly-selected halves of the dataset. This occurs because classification trees have high variance and low bias, which often results in overfitting (James et al., 2013). A remedy for the large variance is bootstrap aggregation ("*bagging*"), which improves prediction accuracy by generating *B* trees from *B* random training samples taken with replacement from the dataset, instead of a single tree. In bagging, individual trees are grown deep and not pruned, which reduces variance when averaging their predictions. However, trees may differ significantly in structure, and they make predictions less interpretable (James et al., 2013).

The random forests model is an extension of bagged classification trees that is constructed by a similar algorithm (Breiman, 2001), except that decision trees are instead grown using q random subsets of p predictors at each node (usually q is set equal to $\sqrt{p}$, to reduce test error and over bagging) (James et al., 2013). This algorithm is advantageous for generating models that do not rely on any one variable as the dominant predictor of the outcome variable, except possibly the root node, and it reduces error variance in the trees (Breiman et al., 1984); however, each random forests model uses a smaller portion of the data, reducing accuracy (Witten & Frank, 2011). Unlike the CART model or OLS regression, the random forests model is not easily interpretable within the context of the original problem because it consists of a multitude of deep trees, each with a large number of nodes, each generated from only a random subset of predictors in a subset of the dataset. In this study, we used a random forests model with $B = 500$ classification trees and 2,717 nodes ($M = 5.434$ nodes per tree). The attrition classification is predicted by each of the individual decision trees in the forest as described in the CART model, and the most commonly occurring category is selected as the result.

Generalized linear model

In the present study, a logistic regression (logit) model was trained to classify individuals on the attrition variable for comparison, using the same model specification and training controls as the other models. However, there are several problems worth mentioning that make traditional ordinary least squares (OLS) regression an inadequate training model. In OLS regression, even a predictor that is unrelated to the outcome will have a nonzero coefficient due to statistical noise, which leads to overfitting particularly in models with a high ratio of predictors relative to sample size (Yarkoni & Westfall, 2017).

Regularization techniques are thus often used to mitigate overfitting and improve statistical prediction by penalizing the model's objective function (e.g., in OLS, the sum of squared errors) *a priori* (Cawley & Talbot, 2010). One such example is a widely-used alternative to OLS regression, least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996), which works by including a shrinking penalty in the objective function (i.e., the sum of squared errors) based on the absolute coefficient magnitudes (the ℓ_1 norm), $\|\boldsymbol{\beta}\|_1 \leq t$, to produce intentionally biased coefficient estimates. This is known as an L_1 penalty, and it relies on the size of a tuning parameter, λ ($\lambda \geq 0$), which controls the amount of shrinking. As λ increases, model bias increases and forces small coefficient estimates to 0 (to thus eliminate predictors that do not contribute to predicting the outcome) (James et al., 2013). Generally, coefficient estimates produced by LASSO regression outperform OLS and generalize better to new datasets (Yarkoni & Westfall, 2017), and the resulting models are simpler to interpret.

The LASSO regression model is included in this study for a comparison between the machine learning models and (regularized) linear regression. When selecting the tuning parameter λ for the LASSO model based on fit criteria, using the Bayesian Information Criterion (BIC) yields the true model more consistently than using cross-validated error (Wang et al., 2007; Zhang et al., 2010). However, for consistency with the other classification models, we selected a tuning value that minimized cross-validated AUC, not BIC.

Test performance indices

A variety of indices were used to assess aspects of model performance. They are discussed here within the context of the UMSC dataset.

Accuracy

Accuracy simply refers to the total proportion of recruits who were assigned by a model to the correct attrition classification.

Phi coefficient

The phi coefficient ($\hat{\phi}$) is a measure of association between two dichotomous variables that can be interpreted similarly to the Pearson correlation coefficient. It is related to the goodness-of-fit chi-square statistic obtained from the 2 × 2 confusion table by $\phi = \sqrt{\frac{\chi^2}{n}}$. In this study, we investigate the strength of the association between model predictions of attrition and their actual values as an index of test performance.

Sensitivity

Sensitivity (also called the true-positive rate; TPR) refers to the proportion of true-positive results that are correctly identified (Bokhari & Hubert, 2015). In this context, it refers to the proportion of recruits who attrit the DEP program and are correctly classified by a model into the yes-attrition group.

Specificity

Specificity (also called the true-negative rate; TNR) refers to the proportion of true-negative results that are correctly rejected (Bokhari & Hubert, 2015). In this context, it refers to the proportion of recruits who do not attrit the DEP program and are correctly classified by a model into the no-attrition group.

PPV

Positive predictive value (PPV; aka. precision) is the proportion of true-positive results among all positive predictions (Bokhari & Hubert, 2015). In this context, it refers to the proportion of recruits who attrit amongst those assigned by a model into the yes-attrition group.

NPV

The proportion of true negatives refers to the proportion of true-negative results among all negative predictions (Bokhari & Hubert, 2015). In this context, the negative predictive value (NPV) refers to the proportion of recruits who do not attrit amongst those assigned by a model into the no-attrition group.

Sensitivity indices

Based on signal detection theory, we used the normalized hit and false alarm rates from the receiver operating characteristic (ROC) curve to calculate the sensitivity index, *d*-prime (*d′*), which corresponds to the distance between the homoscedastic noise and signal distributions (Green & Swets, 1966). In this context, a larger value of *d′* indicates the model is better able to discriminate between the attrition groups. We also calculated *A′* (Grier, 1971), a nonparametric alternative that relaxes the normality distributional assumptions of *d′* and is independent of response bias. *A′* is particularly useful for evaluating sensitivity when signal and noise distributions are heteroscedastic, or when the difference between the sensitivity and the false-alarm rate is small (Pollack & Norman, 1964).

Clinical efficiency

We examined whether any of the specified models met the criteria for accurate classification prediction beyond base rates (i.e., clinical efficiency) (Bokhari & Hubert, 2015). Notably, however, Bokhari and Hubert (2015) caution that when base rates are low, meeting clinical efficiency is difficult and requires high test specificity.

The criteria are as follows: the Meehl-Rosen criterion is met when the positive-predictive value is at least as large as the base rate (i.e., $PPV \geq base\ rate$) (Meehl & Rosen, 1955). Assuming false positive and negative errors are equally undesirable, the Dawes criterion is met either when *PPV* $\geq$ 50% (for *base rates* $\leq$ 50%) or when *NPV* $\geq$ 50% (for *base rates* > 50%) (Dawes, 1962). Lastly, the Bokhari-Hubert (B-H) criterion, which implies the former criteria (although the reverse is not necessarily true), specifies that the use of a test over base rates is justifiable if and only if the test generates a confusion matrix such that $n_{\text{True Positives}} > n_{\text{False Positives}}$ and $n_{\text{True Negatives}} > n_{\text{False Negatives}}$ (Bokhari & Hubert, 2015). Accordingly, satisfying this criterion corresponds to finding a pair of confidence intervals around sensitivity and specificity that both exceed 50%.

AUC

The area under the receiving operating curve (AUC), also known as the concordance index, is a popular indicator of diagnostic test reliability. In general, AUC is independent of the underlying base rates (Bokhari & Hubert, 2015, Chapter 1), but when calculating a single decision threshold, AUC is exactly equivalent to accuracy when the base-rates of both categories are equivalent (i.e., 50%) (Hanley & McNeil, 1982).

Despite its usefulness, AUC has several problems of its own which merit discussion, namely variability across populations with different base rates, spectrum bias within groups (Ransohoff & Feinstein, 1978), as well as other biases traditionally associated with sensitivity and specificity (Begg, 1971; Moons & Harell, 2003; Witten et al., 2011, Chapter 5). For a dichotomous outcome variable such as attrition, the low base-rate (12.28%) can make predictions (e.g., classification tree structures) especially volatile because of its influence on the greedy algorithms used for generating the trees. This makes AUC by itself inadequate for assessing the accuracy of classification models generated from low-base rate signals (Bokhari & Hubert, 2015; Witten et al., 2011). In these cases, establishing clinical efficiency requires high test specificity or class stratification (Bokhari & Hubert, 2015; Dawes, 1962).

Unfortunately, the low-base rate of attrition posed a problem for the present analysis because the CART algorithm was unable to generate a tree beyond the root node. Similarly, the model generated by the random forests algorithm had a selection rate of 0%. These conditions make comparisons between models meaningless as many of the indices of test performance cannot be compared or even computed.

As such, we repeated the same analysis following the stratified 10-fold cross-validation procedure with random subsamples of size *N* ~ 8,885 (approximately 22.76% of the original dataset), such that the attrition rate was fixed at 50%. Here, the subset was obtained by using all of the yes-attrition records, and randomly sampling an equal amount of no-attrition records. This procedure has a strong empirical basis as the best choice for obtaining reliable estimates of diagnostic accuracy, particularly in cases where one class is unbalanced (Witten, et al., 2011, Chapter 5).

CHAPTER 3: RESULTS

The specified models were trained using the dataset in R. Confusion matrices and errors resulting from this analysis are displayed in *Table 3.2*. Unique contributions from each model are shown in the Venn diagram of correct predictions in *Figure 3.6*.

The full classification tree model generated by the CART algorithm is drawn in *Figure 3.1*. Individuals are classified by the decision tree according to the predictors, beginning at the root node (the uppermost node on the tree). Each node involves a yes-no decision based on standardized values of a predictor, ending at the leaf node (the lowest node on the tree) where a prediction is made. In *Figure 3.1*, as we descend the tree, if the condition on a node is met, we move left; otherwise, we move right. The percentages at the bottom of each node are the proportions of recruits (out of the entire sample) that reach this part of the tree, with the majority category represented as a “Yes” for attrition or “No” for no attrition (in the leaf node, this also represents the prediction that is made), while the two decimal numbers at the center represent proportions of the no-attrition and yes-attrition categories respectively within this part of the tree.

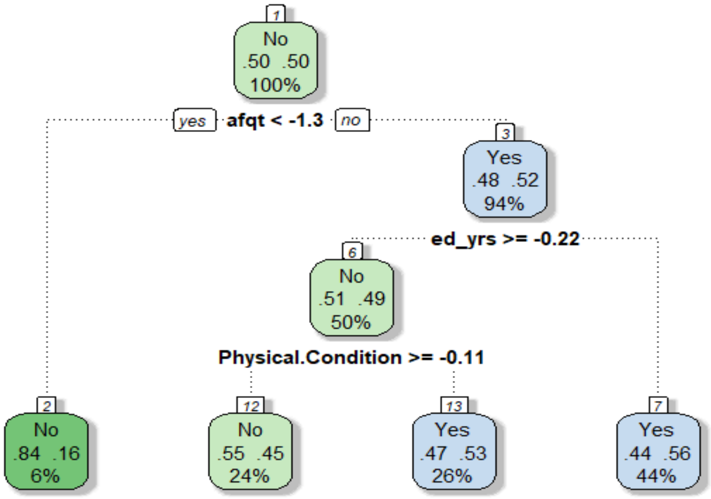


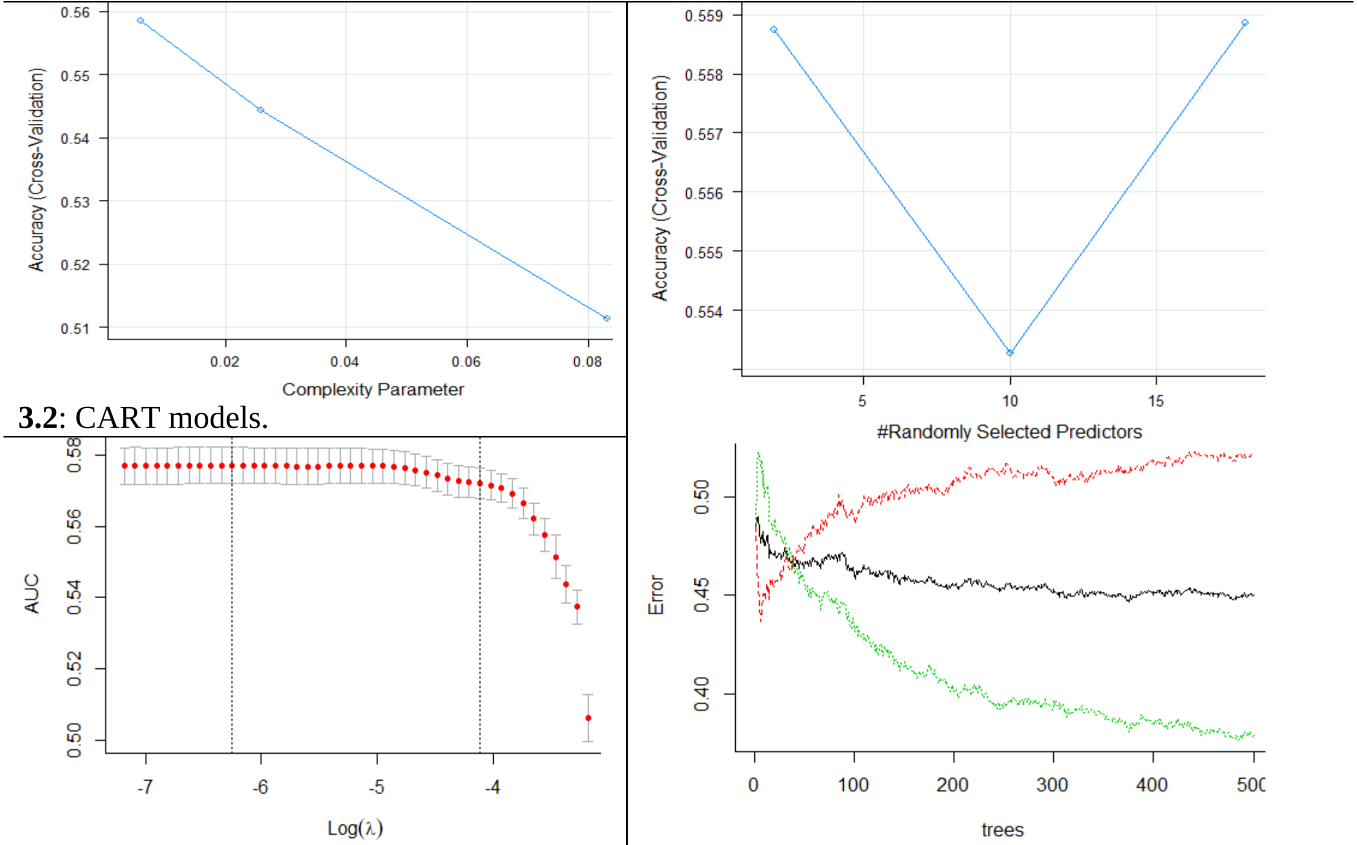
Figure 3.1*:* Attrition classification tree for the 50% stratified sample.

Note: the threshold values for each of the variables on the tree refer to standardized values.

For example, after the first split, approximately 6% of all recruits in the stratified sample have an AFQT score lower than 1.30 *SD* below the mean and are immediately classified into the leftmost node in the tree which corresponds to the no-attrition group. Accordingly, 84% of this group is indeed part of the no-attrition group.

The 10-fold cross-validated testing errors and tuning parameters are in *Figures 3.2-3.4*. In this study, we only investigated the models with superior cross-validated testing error.

Figures 3.2-3.4: Gradient descent of cross-validated testing errors by model.



3.2: CART models.

3.3: LASSO models.

3.4: Random forests models.

Model Comparison

The logit model was an adequate fit for the dataset, $\chi^2(18, N = 9{,}872) = 216.07$, $p < .0001$, $\chi^2/df = 12.00$, $MSE = 4.10$, $AIC = 13{,}507.43$, $BIC = 13{,}644.18$, $-2\ln\hat{\lambda}_n = -6{,}734.715$. The only statistically-significant ($p < .05$) predictors of attrition were years of education ($\hat{\beta} = -0.25$, $SE = 0.03$), the AFQT score ($\hat{\beta} = 0.15$, $SE = 0.02$), and the *Commitment to serve* ($\hat{\beta} = -0.12$, *SE* = 0.02), *Physical condition* ($\hat{\beta}$ = -0.11, *SE* = 0.02), *Sociability* ($\hat{\beta}$ = 0.07, *SE* = 0.02), and *Responsibility* ($\hat{\beta}$ = 0.07, *SE* = 0.02) facets of TAPAS. The solution paths for the LASSO model are included in *Figure* 3.5. Accordingly, the coefficients corresponding to the TAPAS facets of the *Dominance*, *Optimism*, *Selflessness*, and *Team Orientation* were shrunk to 0 at λ = 1.926614 × 10^{-3}, the value of the tuning parameter that maximized cross-validated AUC. Regularization by the LASSO model shrunk the magnitude of the coefficients by at most .03, for *Age*. The coefficients for both models are summarized in *Table 3.1*.

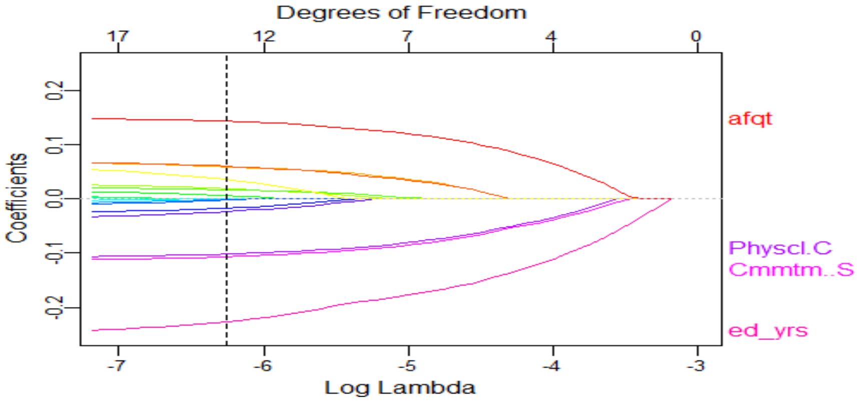


Figure 3.5: Effect of LASSO regularization on regression coefficients.

Table 3.1: Coefficients for logit and LASSO regression.

Source	$\hat{\beta}_{\text{Logit}}$	$\widehat{\text{OR}}_{\text{Logit}}$	$\widehat{\Delta p}_{\text{Logit}}$	$\hat{\beta}_{\text{LASSO}}$	$\widehat{\text{OR}}_{\text{LASSO}}$	$\widehat{\Delta p}_{\text{LASSO}}$	$SE(\hat{\beta})$
Intercept	-.034	0.966	--	-.032	0.969	--	.021
Age	.066	1.068	+1.65%	.036	1.037	+0.91%	.036
Years of education	**-.253**	**0.777**	**-6.29%**	**-.227**	**0.797**	**-5.64%**	.030
AFQT	**.151**	**1.163**	**+3.77%**	**.144**	**1.155**	**+3.59%**	.023
TAPAS							
Achievement	-.011	0.989	-0.29%	--	--	--	.024
Adjustment	-.040	0.961	-0.99%	-.023	0.977	-0.57%	.023
Commitment to serve	**-.115**	**0.891**	**-2.87%**	**-.107**	**0.899**	**-2.67%**	.022
Courage	.018	1.019	+0.46%	.006	1.006	+0.16%	.023
Dominance	.006	1.006	+0.15%	--	--	--	.023
Even-tempered	.031	1.031	+0.76%	.019	1.019	+0.48%	.022
Ingenuity	.022	1.023	+0.56%	.017	1.017	+0.43%	.022
Optimism	.008	1.008	+0.21%	--	--	--	.022
Physical condition	**-.109**	**0.896**	**-2.73%**	**-.101**	**0.904**	**-2.53%**	.021
Responsibility	**.071**	**1.074**	**+1.78%**	**.060**	**1.061**	**+1.49%**	.024
Selflessness	-.005	0.995	-0.12%	--	--	--	.021
Sociability	**.072**	**1.075**	**+1.80%**	**.061**	**1.063**	**+1.52%**	.022
Team orientation	-.003	0.997	-0.70%	--	--	--	.021
Tolerance	-.027	0.973	-0.69%	-.017	0.983	-0.42%	.021
Virtue	-.013	0.987	-0.00%	-.002	0.998	-0.05%	.023

Standardized regression coefficients that are significant at p < .05 are bolded. $\widehat{\Delta p}$ refers to the estimated change in the probability of attrition, per 1-*SD* increase in the predictor.

By contrast, recruits were predicted by the CART model to attrit from the DEP program if the AFQT score was 1.30 *SD* above the mean (~87th percentile) and they received at least 0.22 *SD* below the mean number of years of education (*M* = 11.5), or if their physical condition was at or below 0.11 standard deviations below the mean (*M* = 0.14).

Table 3.2: Confusion matrices (50% attrition sample).

	Logit model			LASSO model			CART model			Random forests model		
	Attrition	No attrition	Error	Attrition	No Attrition	Error	Attrition	No attrition	Error	Attrition	No attrition	Error
Attrition	2,838	2,098	42.50%	2,881	2,055	41.63%	**3,484**	**1,452**	**29.42%**	3,145	1,791	36.28%
No attrition	2,273	2,663	46.05%	**2,249**	**2,687**	**45.56%**	2,907	2,029	58.89%	2,564	2,372	51.94%

Note: the optimal values for each cell are bolded.

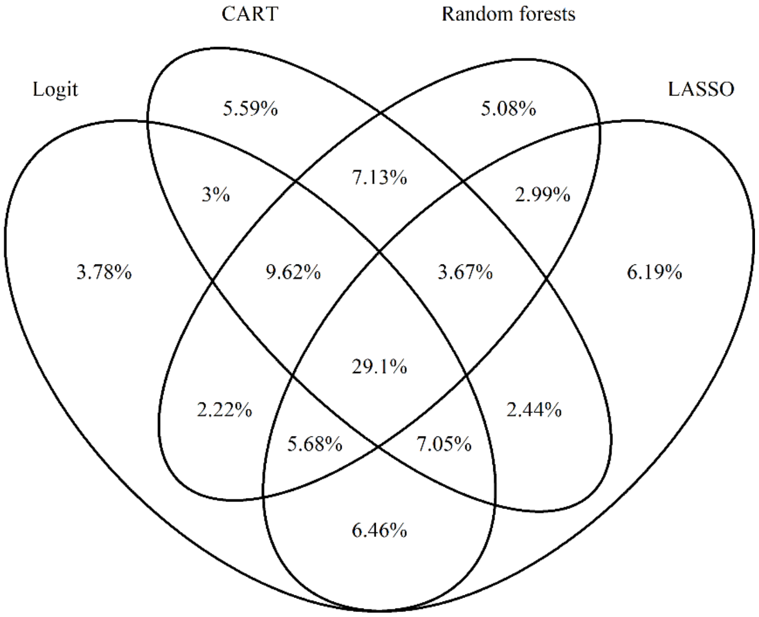


Figure 3.6: Venn diagram of correct predictions of attrition by model.

Table 3.3: Phi coefficients between model predictions.

	(1)	(2)	(3)	(4)
(1) Logit	(***.0201***)			
(2) LASSO	**.9308**	(***.0204***)		
(3) CART	**.3982**	**.4170**	(*.0093*)	
(4) Random forests	**.3416**	**.3528**	**.4086**	(*.0007*)

All Phi coefficients significant at *p* < .05 are bolded.
Note: For comparison, italicized values in parentheses along the diagonal are the largest Phi coefficients calculated between each model's predictions and one of 10,000 randomly-selected permutations of possible classifications for the sample of *N* = 9,872 participants, out of all $2^{9,872}$ permutations. The performance for the optimal randomized classifier was: ACC = 52.19%, SENS = 52.07, SPEC = 51.80, PPV = 51.93, NPV = 51.94.

All models outperformed base-rate prediction and shared about 29.10% of correct predictions, as shown in *Figure 3.6.* Phi coefficients were calculated between model predictions in *Table 3.3*. Approximately 86.63% of the variability in predictions was shared between the logit and LASSO models, compared with 16.70% between the CART and random forests model.

A goodness-of-fit test revealed the proportion of observations assigned to each cell by the logit model differed from chance assignment (no-information rate = 50%), $\chi^2(1, N = 9{,}872) = 129.05$, $p < .0001$, Cramér's $V = .11$. Predictions from the LASSO model also outperform base-rate assignment, $\chi^2(1, N = 9{,}872) = 161.58$, $p < .0001$, Cramér's $V = .13$. Finally, the CART model ($\chi^2(1, N = 9{,}872) = 147.22$, $p < .0001$, Cramér's $V = .12$) and the random forests model generated confusion matrices that differed significantly from chance assignment, $\chi^2(1, N = 9{,}872) = 139.73$, $p < .0001$, Cramér's $V = .12$. A full comparison of performance indices is provided in *Table 3.5* and *Figure 3.7*.

We also examined whether the models satisfied the criteria for clinical efficiency (Bokhari & Hubert, 2015). A proportion test on the PPV and NPV of each model was used to obtain 95% confidence intervals for the Meehl-Rosen and Dawes criteria as measures of the incremental predictive power of each model. The results for these analyses are summarized in *Table 3.4.* In general, the LASSO and logit models best fulfilled the Meehl-Rosen criterion (and the PPV part of the Dawes criterion), while the CART and random forests model best fulfilled the NPV part of the Dawes criterion; however, the differences between models were small, and the confidence intervals overlap. For the B-H criterion, a pair of exact binomial tests were conducted using $n_{\text{True Positives}}$ and $n_{\text{True Negatives}}$ as the number of successes within each attrition group to obtain 95% confidence intervals for sensitivity and specificity against the base-rate also in *Table 3.4*. Almost every model met each of the target criteria for superior accuracy beyond base-rate prediction in the stratified subsample. In the case of the CART and random forests model, specificity did not outperform change assignment.

Finally, the LASSO model had the largest AUC (best accuracy), followed by the random forests model; however, the 95% CIs closely overlap. In general, the LASSO and logit models minimized classification error in the no-attrition groups (i.e., maximized specificity, PPV), while the CART and random forests models minimized classification error in the yes-attrition groups (i.e., maximized sensitivity, NPV). This result suggests the choice of model involves a tradeoff between two kinds of misclassification errors that can be minimized. Although the LASSO and logit models had superior specificity, the sensitivity indexes of both the CART and random forests model were much better, suggesting the machine learning models may be better at correctly identifying instances of attrition. Notably, regularization (i.e., LASSO regression) slightly improved performance across all criteria by 0.49-0.87%.

The random forests model had the largest PPV and was best able to discriminate instances of attrition from noise according to the detection indices *d*' and *A*'. The confidence intervals of the sensitivity index are non-overlapping, suggesting that both CART and random forests models outperform logistic regression in correctly identifying cases of attrition, at a cost of lower specificity.

Table 3.4: Statistical criteria for clinical efficiency.

	Logit	LASSO	CART	Random forests
SENS – 50%	7.50%	8.37%	**20.58%**	13.72%
	[.0610, .0888]	[.0698, .0975]	[.1929, .2185]	[.1236, .1506]
SPEC – 50%	3.95%	**4.44%**	-8.89%	-1.94%
	[.0255, .0535]	[.0304, .0583]	[-.1027, -.0751]	[-.0335, -.0054]
PPV – 50%	5.55%	**6.16%**	4.51%	5.09%
	[.0415, .0690]	[.0479, .0752]	[.0329, .0572]	[.0379, .0638]
NPV – 50%	5.93%	6.66%	**8.29%**	6.98%
	[.0451, .0735]	[.0524, .0808]	[.0663, .0993]	[.0546, .0849]

SENS = sensitivity, SPEC = specificity, PPV = positive predictive value, NPV = negative predictive value.
Note: Brackets represent 95% confidence intervals. Optimal values are bolded for each row.

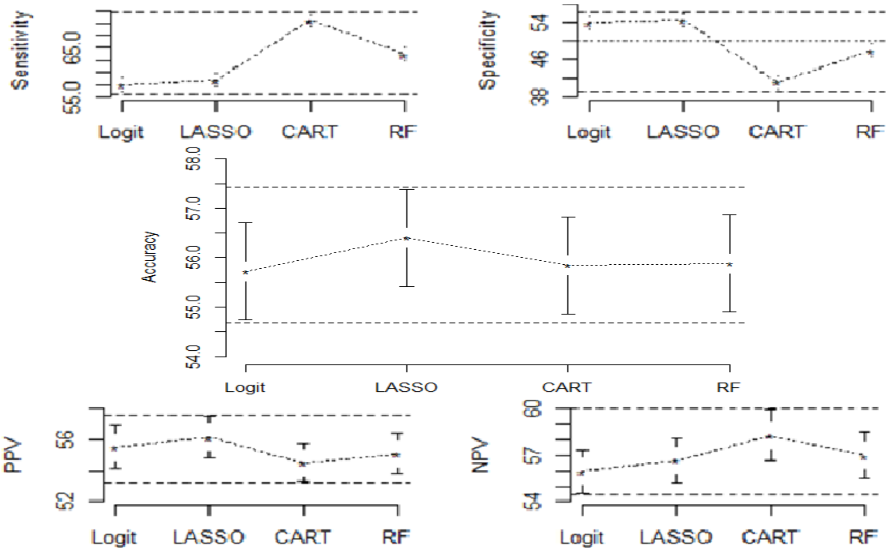


Figure 3.7: Comparisons of model performance.

Table 3.5: Model performance indices (50% attrition sample).

	Logit	LASSO	CART[a]	Random forests[b]
SR	51.77%	51.97%	64.74%	58.04%
ACC	55.72%	**56.40%**	55.84%	55.89%
SENS	57.50%	58.37%	**70.58%**	63.72%
SPEC	53.95%	**54.44%**	41.11%	48.06%
PPV	55.53%	**56.16%**	54.41%	55.09%
NPV	55.93%	56.66%	**58.29%**	56.98%
d'	.2881	**.3226**	.3163	.3020
A'	.6028	**.6136**	.6125	.6074
ϕ	.1145	**.1281**	.1223	.1192

SR = selection rate, ACC = accuracy (equivalent to area under the receiving operating curve (AUC)), SENS = sensitivity, SPEC = specificity, PPV = positive predictive value, NPV = negative predictive value, *d*' = sensitivity index, *A*' = nonparametric sensitivity index, ϕ = Phi coefficient between attrition predictions and observed values.

Note: The selection rate considers an occurrence of attrition as a positive result. The no-information rate is 50%. Finally, the optimal values for each row are bolded where applicable.

[a]The optimal CART model was tuned with 3 splits using a complexity parameter of 0.006077796.

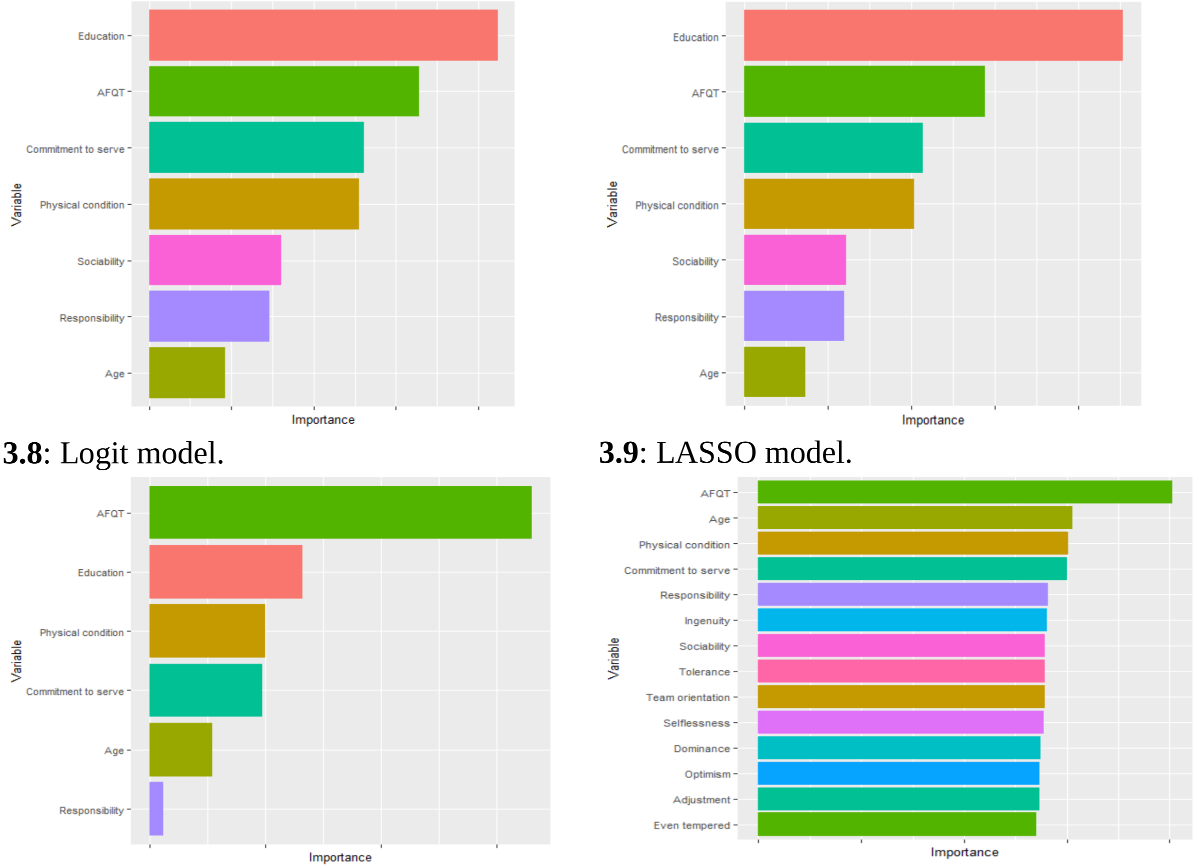
[b]The optimal random forests model was tuned using 2 variables randomly sampled as candidates at each branch split.

All proportions and Phi coefficients are significant at $p < .0001$.

Predictor importance

The importance ratings (given by the log-loss function for the regression models; Gini node impurity for the machine learning models) of the most important predictors of attrition in the stratified sample (i.e., those that accounted for ≥80% of the total importance ratings) are given for the highest-accuracy models in *Figures 3.8-3.11*. Namely, the AFQT score, years of education (and/or possibly age), physical condition, as well as the *Commitment to serve*, *Sociability*, *Responsibility*, and *Courage* facets of TAPAS were important predictors of attrition common to most models. Interestingly, higher AFQT scores were associated with increased probability of attrition. Unsurprisingly, variable importance was nearly identical between the LASSO and logit models; by comparison, the CART model emphasized the AFQT score and physical condition over years of education, *Commitment to Serve*, and *Responsibility*. The model generated by the random forests algorithm had a considerably more uniform and varied distribution of predictor importance than the other models, as evidenced by *Figure 3.11*; however, the AFQT was still the most important predictor, as with the CART model.

Figures 3.8-3.11: Variable importance by model.



3.8: Logit model.

3.9: LASSO model.

3.10: CART model.

3.11: Random forests model.

Reason-specific models

Following the same imputation and cross-validation procedures, we generated reason-specific prediction models from stratified subsets of the original sample. Like before, we calculated 95% confidence intervals for accuracy and other performance indices. Sample sizes for each of the 24 reason-specific subsets ranged from *N* = 12 to *N* = 3,962 (i.e., double the frequencies counts for reasons for attrition in *Table 2.3*). Aggregate results by model are presented in *Table 3.6*. Confidence intervals ordered by increasing sample size are reported for accuracy in *Figures 3.12-3.15* and the other metrics in *3.16-3.19*.

Table 3.6: Standard deviations, weighted means, and medians for performance indices across models generated to predict various reasons for attrition.

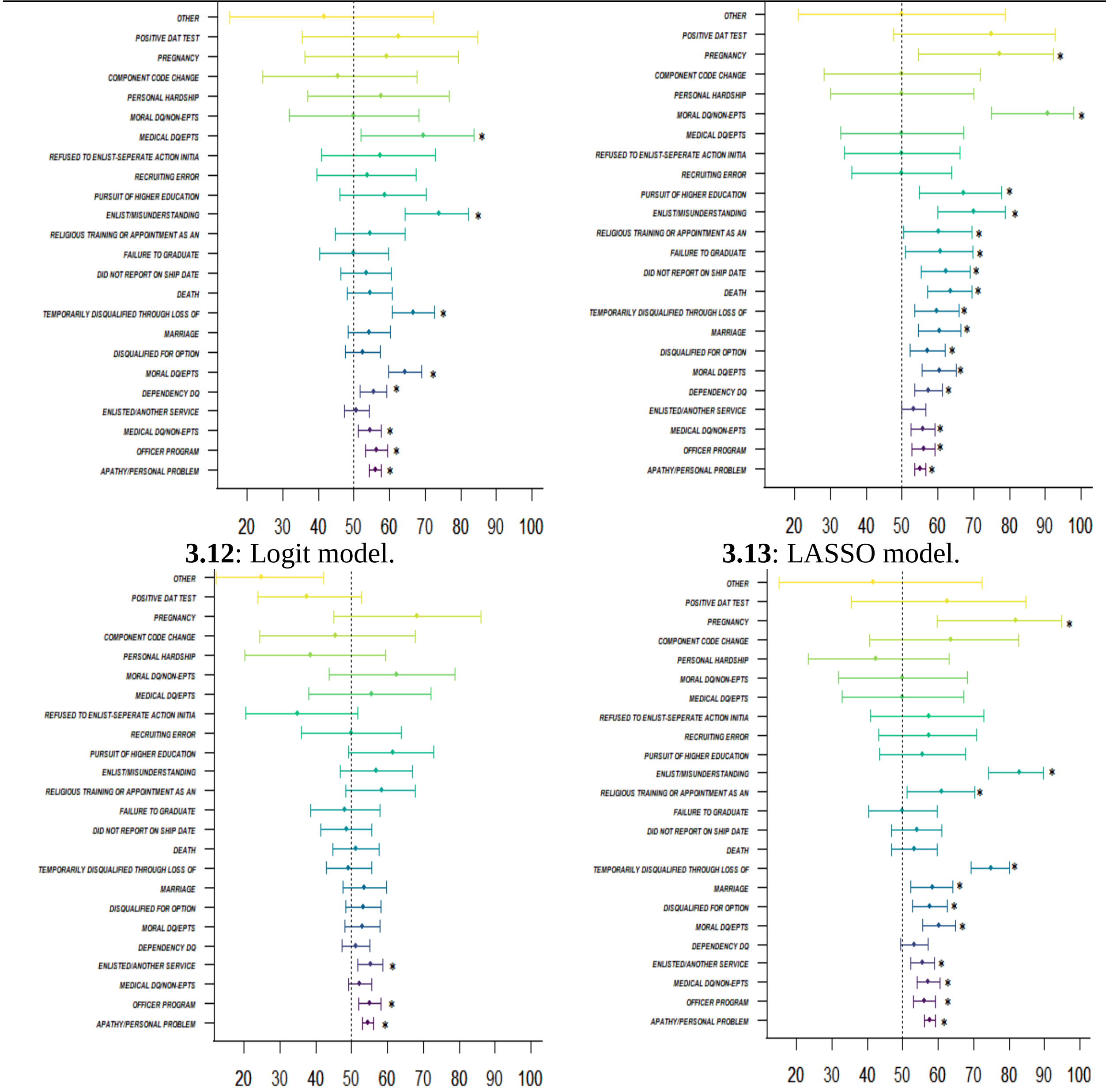
		Logit	LASSO	CART	Random forests
ACC	M_{Weighted} (*Mdn*)	55.89% (55.06%)	56.62% (**58.59%**)	53.64% (52.69%)	**57.59%** (57.33%)
	SD	**7.23%**	10.10%	9.40%	10.06%
SENS	M_{Weighted} (*Mdn*)	56.27% (54.55%)	54.63% (**57.49%**)	53.15% (51.51%)	**60.61%** (57.03%)
	SD	**9.45%**	30.00%	18.42%	12.89%
SPEC	M_{Weighted} (*Mdn*)	55.50% (55.99%)	**58.62%** (**63.31%**)	54.12% (55.01%)	54.57% (55.53%)
	SD	**7.50%**	17.54%	14.23%	8.35%
PPV	M_{Weighted} (*Mdn*)	55.79% (55.15%)	56.89% (**61.01%**)	53.61% (55.28%)	**57.08%** (57.02%)
	SD	**7.19%**	8.31%	17.12%	9.51%
NPV	M_{Weighted} (*Mdn*)	56.05% (55.00%)	56.54% (**58.36%**)	53.86% (52.64%)	**58.27%** (57.24%)
	SD	**7.67%**	11.66%	7.82%	11.14%

ACC = accuracy (equivalent to area under the receiving operating curve (AUC)), SENS = sensitivity, SPEC = specificity, PPV = positive predictive value, NPV = negative predictive value

Note: Weighted means are calculated for 24 reasons for attrition using individual sample sizes ranging from *N* = 12 to *N* = 3,962. Optimal values are bolded for each row.

Overall, none of the models performed significantly better than chance at predicting cases of attrition due to personal hardship (*N* = 13), positive DAT test (*N* = 8), recruiting errors (*N* = 27), or component code change (*N* = 11). Unlike the previous analysis with the 50-50 stratified sample, the model-building algorithms used in this aggregate analysis are heavily impacted by the small sample sizes for some of the reasons of attrition. The LASSO model performed better than chance in 15 reasons for attrition, compared to 11 by the random forests model and 8 by logit models. Regularization from logit to LASSO improved accuracy for most reasons for attrition, while only 3 CART models produced better accuracy than base-rate prediction.

Figures 3.12-3.15: 95% confidence intervals around prediction accuracy by attrition reason.



3.12: Logit model.

3.13: LASSO model.

3.14: CART model.

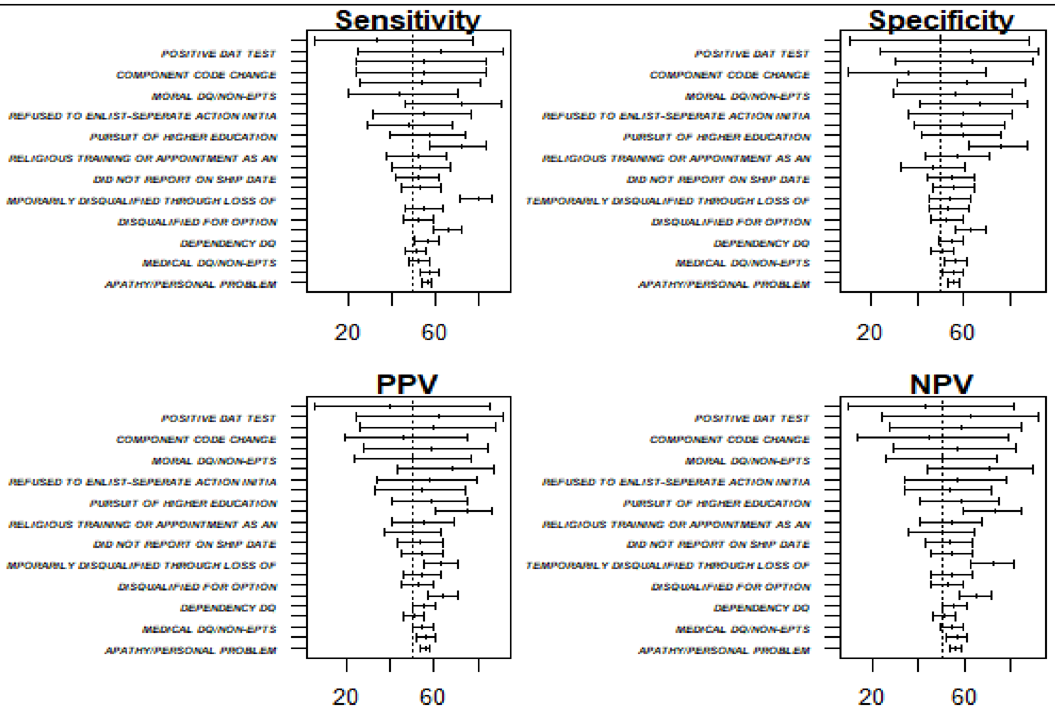
3.15: Random forests model.

In general, the random forests models had the best weighted mean accuracy, sensitivity, PPV, and NPV, followed closely by the LASSO models, which had superior specificity. This pattern of results is similar to the one we obtained for the 50-50 stratified sample between the LASSO and CART models: the LASSO models seem to be best at maximizing specificity, while the machine learning models are best at maximizing sensitivity. In this case, the aggregate results showed the random forests models barely surpassing the LASSO models in mean accuracy and

PPV, while individually, the LASSO models at the 50th percentile of each performance index performed better than any other type of model. It is worth noting, as before, that the confidence intervals overlap closely between performance metrics of these models, and because the intervals are calculated from random subsamples (for which the selection rate may be zero), it is possible these differences may be statistically spurious. Moreover, the LASSO and CART models were the most variable in performance, while the indices of performance in the logit models had the least variability.

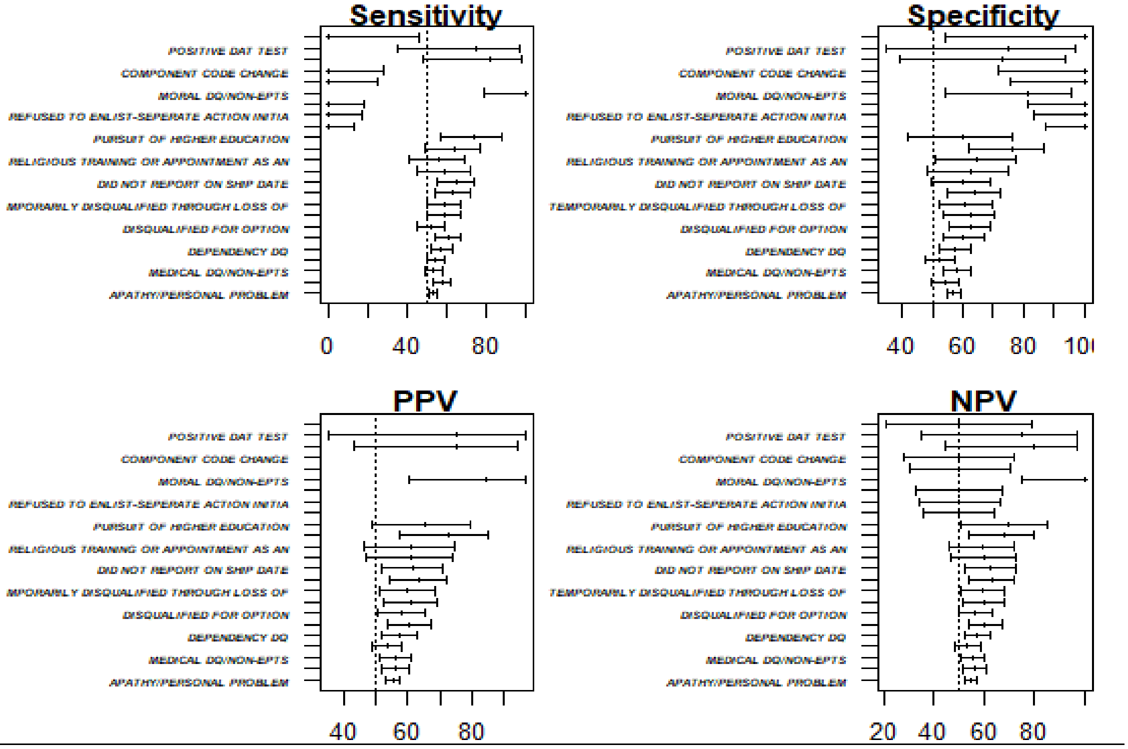
Figures 3.16-3.19: 95% confidence intervals by attrition reason for other metrics of performance.

3.16: Logit model.

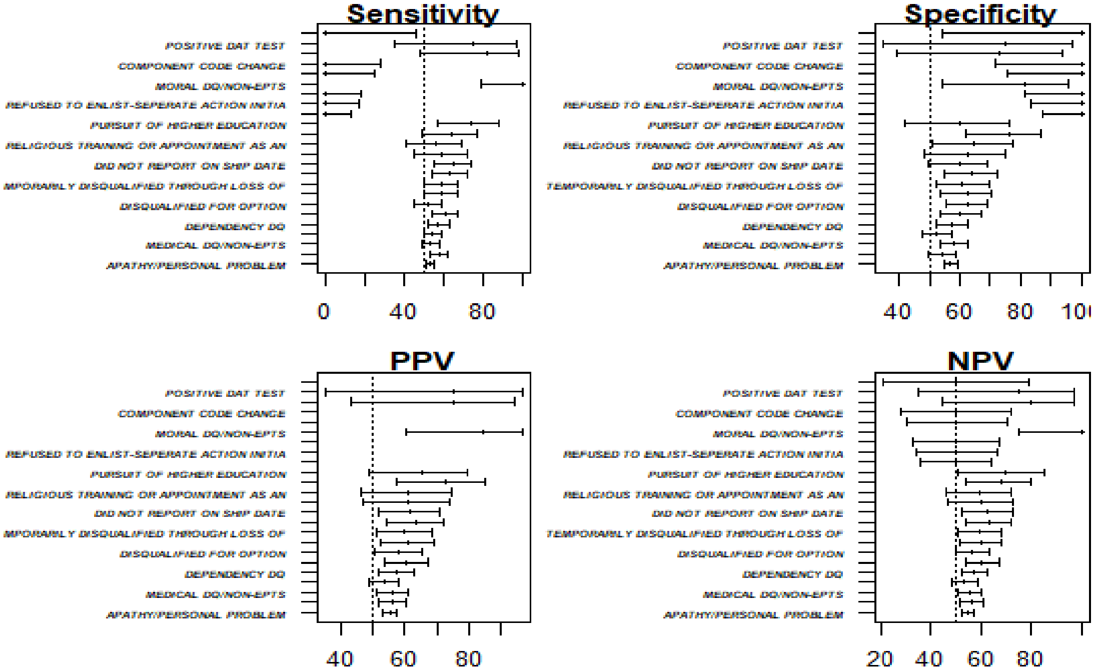


Note: The confidence intervals corresponding to each type of model appear in pages 24-26.

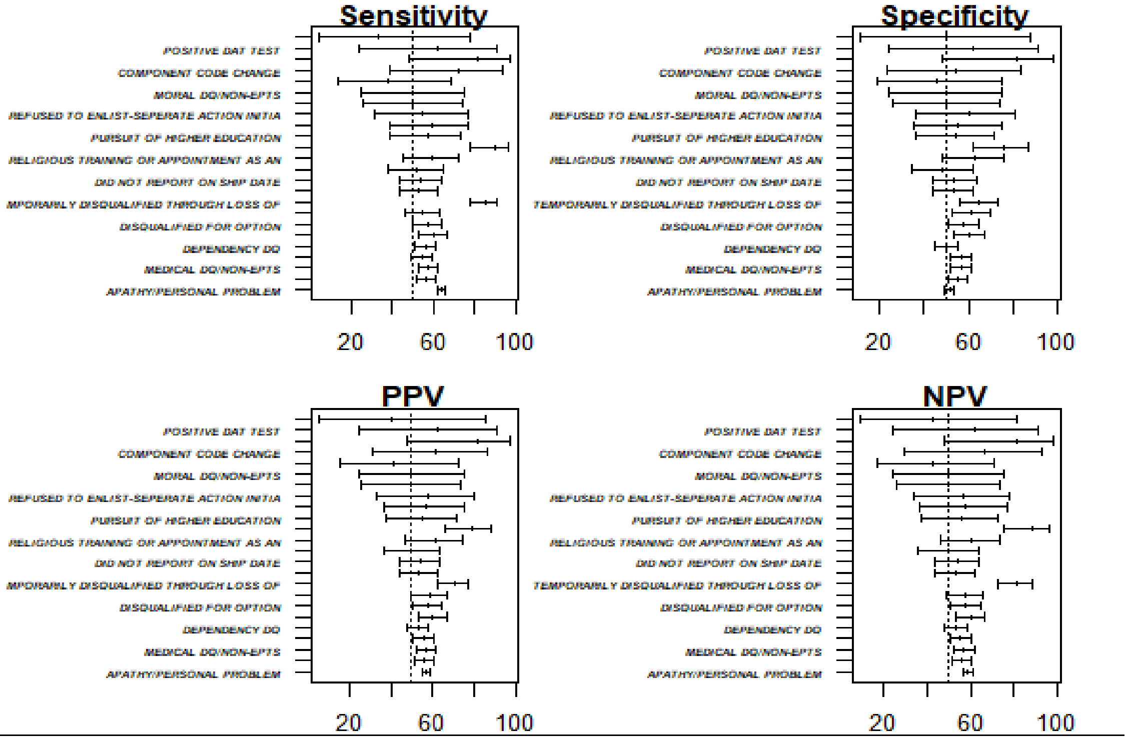
3.17: LASSO model.



3.18: CART model.



3.19: Random forests model.



Finally, we assessed model performance after applying each of the models generated from the stratified sample to the entire sample (with an attrition rate of 12.64%). Confusion matrices for this procedure are in *Table 3.7*, and full model performance is summarized in *Table 3.8*. In this analysis, the random forests model performed best out of all the models according to all of the indexes of performance, correctly classifying every case of attrition (i.e., perfect sensitivity) while also minimizing false-negative cases. The benefits of regularization over logit regression improved performance in the LASSO model, but not to the degree observed in the results with the stratified sample (in fact, unbalanced accuracy in the full sample was actually smaller in the LASSO model).

Table 3.7: Confusion matrices (Full sample).

	Logit model			LASSO model			CART model			Random forests model		
	Attrition	No attrition	Error	Attrition	No Attrition	Error	Attrition	No attrition	Error	Attrition	No attrition	Error
Attrition	2,854	2,082	42.18%	2,881	2,055	41.63%	3,762	1,174	23.78%	**4,936**	**0**	**0.00%**
No attrition	15,585	18,522	45.69%	15,577	18,530	45.67%	21,184	12,923	62.11%	**14,795**	**19,312**	**43.38%**

Note: the optimal values for each cell are bolded.

Table 3.8: Model performance indices (Full sample).

	Logit	LASSO	CART[a]	Random forests[b]
SR	47.23%	47.28%	63.89%	50.54%
ACC	57.82%	54.84%	42.73%	**62.11%**
ACC (Balanced)	56.06%	56.35%	57.05%	**78.31%**
SENS	57.82%	58.37%	76.22%	**100.00%**
SPEC	54.31%	54.33%	37.89%	**56.62%**
PPV	15.48%	15.61%	15.08%	**25.02%**
NPV	89.90%	90.02%	91.67%	**100.00%**
d'	.3053	.3200	.4047	**3.8826**
A'	.6082	.6128	.6393	**.8916**
ϕ	.0807	.0845	.0976	**.3763**

SR = selection rate, ACC = accuracy (equivalent to area under the receiving operating curve (AUC)), SENS = sensitivity, SPEC = specificity, ACC (Balanced) = mean of sensitivity and specificity, PPV = positive predictive value, NPV = negative predictive value, d' = sensitivity index, A' = nonparametric sensitivity index, ϕ = Phi coefficient between attrition predictions and observed values.

Note: The selection rate considers an occurrence of attrition as a positive result. The no-information rate is 87.36%. Finally, the optimal values for each row are bolded where applicable.

[a] The optimal CART model was tuned with 3 splits using a complexity parameter of 0.006077796.

[b] The optimal random forests model was tuned using 2 variables randomly sampled as candidates at each branch split.

All Phi coefficients are significant at $p < .0001$.

CHAPTER 4: DISCUSSION

The choice of models in this study represented a spectrum of interpretability (i.e., face validity) for the predictive model of attrition; the most easily interpretable, the mutual independence logit model, was obtained by finding an optimal vector of parameters $\boldsymbol{\beta}$ that maximized the likelihood function (i.e., minimized the sum of squares). The size and direction of a parameter corresponding to a predictor can be interpreted as its effect on the outcome variable (attrition), independent of all other predictors in the model. By contrast, the random forests model (which offered the best performance) was the least interpretable, as mentioned earlier, because it consists of multiple layers of tree structures, generated from randomly-selected (i.e., dataset-sensitive) subsets of the predictors. Thus, these complex tree structures only yield an estimate of importance relative to other variables. As seen in *Figure 3.11*, its distribution of variable importance was relatively homogenous. Notably, the classification tree model was a reasonably accurate compromise that still presented a fairly interpretable structure (unlike the random forests model). However, unlike the random forests model, its distribution of variable importance was the least balanced.

A key finding of this study was that machine learning classification models can outperform logistic regression in correctly predicting instances of attrition (even when using regularization, as with the LASSO model), but model comparison is difficult in the presence of a low (12.28%) sample base-rate. Adequate model comparison required the 50% stratification procedure because the low base-rate prevented the machine learning algorithms from correctly generating prediction models. As mentioned, however, this procedure carries the consequence of slightly underestimating test error variance, which is a limitation inherent to this analysis (Breiman, Friedman, Olshen, & Stone, 1984; Yarkoni & Westfall, 2017). While the current study did not attempt to use leave-one-out cross-validation due to computational constraints, it would be interesting to verify if machine learning-based classification can be achieved despite the low base-rate of attrition using leave-one-out CV. The models generated in the stratified sample may also be used quasi-experimentally to predict attrition in a sample of unlabeled data.

The conclusions drawn from this study are subject to several other limitations stemming from the choice of predictors. The multicollinearity observed between several of the predictors (e.g., age and years of education) may have confounded the process of selecting the best split along the greedy algorithms. Also, evidence of adequate fit does not necessarily provide support for a model's validity without proper knowledge of the theory underlying the predictive model, the variability of the data, and the likelihood of other outcomes (in this case, the base rate of attrition). In fact, any model that can fit about 50% of the dataset will closely estimate the remainder by linear interpolation (Roberts & Pashler, 2000; Rodgers & Rowe, 2002). Moreover, there is also risk of leakage of information between observations (which can deflate estimates of testing error) when centering the values of a predictor, which was particularly high during preprocessing given that predictors were standardized and imputed using *k*NN imputation for missing values (Roberts & Pashler, 2000). Similarly, the chi-squared statistic used to assess fit is inflated in large (*N* > 1,000) samples like the one used in this study (Schreiber, Nora, Stage, Barlow, & King, 2006).

Notably, model selection strategies should ideally be embedded in the cross-validation steps. In this study, the optimal machine learning models we selected to investigate were those that were tuned to minimize cross-validated testing error (see note in *Table 3.5*). This choice is certainly worth questioning, as it can result in overfitted models (Cawley & Talbot, 2010; Yarkoni & Westfall, 2017). Incidentally, the models that minimized testing error in this study were often the least complex. However, as evidenced by *Figures 3.4* more complex models also offered similar (±4.5%) classification accuracy, and may be worth exploring to minimize certain kinds of misclassification errors or predict a single reason for attrition. It may also be possible to prune the complex models with access to another set of labeled data to yield more generalizable tree structures than those of the simpler models.

Nonetheless, the goal of training generalizable machine learning models on large datasets to outperform linear regression predictions of important outcome variables is certainly worth pursuing, as evidenced by the superior performance of the CART and random forests models in this study (Yarkoni & Westfall, 2017). Likewise, finding evidence of adequate model fit can be a good starting point for theory development (Rodgers & Rowe, 2002). Theoretically-oriented research may thus consider conducting an exploratory factor analysis (EFA) to narrow down the list of important predictors as they pertain to attrition prior to generating the models (which can be achieved with the covariance matrix presented in *Table 2*), or develop a model-building strategy for studying interaction effects relevant to attrition.

Overall, the pattern of results suggests the choice of a different training algorithm is useful for minimizing certain kinds of classification errors, even at the cost of lower overall accuracy. In our study, the choice of which model to endorse for the purpose of predicting attrition in the USMC largely depends on the severity of each misclassification error. Intuitively, classification in this context may be aimed at reducing false-negatives (i.e., recruits who go on to attrit); however, a false-positive result (i.e., incorrectly-identifying a recruit as someone who will attrit) may also incur costs. In terms of correctly identifying instances of recruit attrition (i.e., reducing false-negative errors) in the stratified sample, the CART and random forests models were much more sensitive to instances of attrition than logistic and LASSO regression. The random forests model retained its superior sensitivity even when applied to the full sample. In terms of correctly rejecting instances of no attrition (i.e., reducing false-positive errors), the LASSO and logit models outperformed the machine learning models. Ultimately, regularization of the logit model into the LASSO model resulted in better predictions than the logit model; given equal weight to both kinds of misclassification errors, the LASSO model best maximized cross-validated AUC. The bulk (29.10%) of correct predictions were common to all models, with each model contributing between 3.78%-6.19% unique correct predictions not covered by the others. A large portion (17.80%) of all correct predictions of attrition cases was uniquely predicted by the machine learning models, compared to 16.43% for the LASSO and logit models.

Individually, the random forests models outperformed logit regression (but not LASSO) in predicting most reasons for attrition, while only 3 CART models performed better than chance. Although the random forests model performed similarly to the LASSO model in this analysis, this result demonstrated the limitations of using machine learning when large sample sizes are not available. Several categories for which the logit and/or LASSO models outperformed the machine learning models involved sample sizes that were quite small ($N \leq 70$), suggesting the training algorithms may be overfitting at the expense of testing error (i.e., generalizability) in these occurrences of superior accuracy.

Further investigations may generate better models for budgeting purposes by assigning a cost to each false-positive and false-negative decision (or weigh these according to the discharge reason), and following our approach of using 95% confidence intervals around performance measures to determine which of the models best satisfies the generalized criteria for clinical efficiency under these conditions (Meehl & Rosen, 1955), as well as which criteria are most important to consider for the problem at hand. For clinical efficiency to hold, false negatives must be considered only between twice and 10.3 times as costly as false positives (Bokhari & Hubert, 2015). In the context of predicting USMC attrition, this type of analysis might help to generate an interpretable model that links TAPAS personality facets to important outcomes in the military, such as attrition, performance, and deviant behaviors.

References

Begg, C. B. (1987). Bias in the assessment of diagnostic tests. *Statistics in Medicine*, *6*, 411-423.

Bokhari, E. and Hubert, L. (2018). The lack of cross-validation can lead to inflated results and spurious conclusions: A re-analysis of the MacArthur Violence Risk Assessment Study. *Journal of Classification*, *35*, 147–171.

Bokhari, E. and Hubert, L. (2015). A new condition for assessing the clinical efficiency of a diagnostic test. *Psychological Assessment*, *27*, 745–754.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:/10.1023/A:1010933404324.

Breiman, L., and Spector, P. (1992). Submodel selection and evaluation in regression: The *X*-random case. *International Statistical Review*, 291–319.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall, Wadsworth, New York.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079‑2107.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. *20*(1): 37–46. doi:10.1177/001316446002000104.

Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology, 26,* 422–424.

Drasgow, F., Stark, S., Chernyshenko, O.S., Nye, C.D., and Hulin, C.L. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Army Selection and Classification Decisions* (Technical Report 1311). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB).* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of American Statistical Association, 49,* 732–762.

Grier, J. B. Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, 1971, *75*, 424–429.

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Halstead, J.B. (2009). Recruiter Selection model and implementation within the U.S. Army. *IEEE Transactions on System, Man, and Cybernetics*, *19*, 88-106.

Hanley, J.A., and McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148,* 839–43.

Hastic, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R.* Corrected edition. New York: Springer.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–215.

Moons, K. and Harrell, F. (2003). Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology, 10,* 670–672.

Nye, C.D., White, L.A., Horgen, K., Drasgow, F., Stark, S. & Chernyshenko, O.S., (2018, invited). Predictors of attitudes and performance in U.S. Army recruiters: Does personality matter? *Military Psychology*.

Pollack, I., & Norman, D. A. A non-parametric analysis of recognition experiments. *Psychonomic Science*, 1964, *1*, 125–126.

Ransohoff, D. F. and Feinstein, R. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, *299*, 926–930.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Rodgers, J. L., & Rowe, D. C. (2002). Theory development should begin (but not end) with good empirical fits: A comment on Roberts and Pashler (2000). *Psychological Review, 109*, 599–604.

Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A, & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review, *The Journal of Educational Research*, *99*(6), 323–338, doi:10.3200/ JOER.99.6.323-338.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B* (Methodological), *58*(1), 267–288.

Vijayakumar, R., & Cheung, M. W. (2018). Replicability of machine learning models in the social sciences: A case study in variable selection. *Zeitschrift Für Psychologie, 226*(4), 259–273. doi:10.1027/2151-2604/a000344.

Wang, H., Li, G. and Tsai, C. (2007). *Regression coefficient and autoregressive order shrinkage and selection via the lasso*. *Journal of the Royal Statistical Society: Series B* (Methodological), *69*(1), 63–78.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zhang, Y., Li, R. and Tsai, C.-L. (2010). *Regularization parameter selections via generalized information criterion*. Journal of the American Statistical Association, *105*, 312–323.