NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MODELING OFFICER SELECTION FOR NAVAL SPECIAL WARFARE

by

Keith V. Campion

September 2021

Thesis Advisor:	Robert A. Koyak
Second Reader:	Samuel E. Buttrey

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2021	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
MODELING OFFICER SELECTION FOR NAVAL SPECIAL WARFARE

5. FUNDING NUMBERS

6. AUTHOR(S) Keith V. Campion

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Every year, the SEAL Officer Community Management (OCM) receives approximately 300 applications from candidates who want to attend Basic Underwater Demolition/SEAL (BUD/S) training to become Navy SEAL Officers. The applications include multiple data elements such as university attendance, grade point average (GPA) and physical fitness test scores. From this data, the SEAL OCM selects approximately 80% of the candidates to participate in SEAL Officer Assessment and Selection (SOAS). At SOAS, the candidates are assessed by the cadre at Naval Special Warfare (NSW) Basic Training Command (BTC), who provide evaluation data and recommendations to the SEAL OCM for candidate selection to BUD/S.

In total, the process of assessing, selecting and training candidates to become SEAL Officers is resource-intensive, incurring a financial cost to the Navy, manning challenges for NSW and the time and energy of the candidates themselves. Removing candidates who have a low probability of success at BUD/S early in the process reduces the costs and allows them to be reassigned within the Navy to a community more appropriate to their abilities.

This thesis aims to analyze data collected on the candidates to train statistical models capable of predicting a candidate's probability of success in the first phase of BUD/S and inform data collection to capture information on candidates for future data analysis.

14. SUBJECT TERMS NSW, BUD/S, selection, recruiting, Officers, SEAL, training, OCM	**15. NUMBER OF PAGES** 67
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

MODELING OFFICER SELECTION FOR NAVAL SPECIAL WARFARE

Keith V. Campion
Lieutenant, United States Navy
BA, Virginia Military Institute, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2021

Approved by: Robert A. Koyak
Advisor

Samuel E. Buttrey
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Every year, the SEAL Officer Community Management (OCM) receives approximately 300 applications from candidates who want to attend Basic Underwater Demolition/SEAL (BUD/S) training to become Navy SEAL Officers. The applications include multiple data elements such as university attendance, grade point average (GPA) and physical fitness test scores. From this data, the SEAL OCM selects approximately 80% of the candidates to participate in SEAL Officer Assessment and Selection (SOAS). At SOAS, the candidates are assessed by the cadre at Naval Special Warfare (NSW) Basic Training Command (BTC), who provide evaluation data and recommendations to the SEAL OCM for candidate selection to BUD/S.

In total, the process of assessing, selecting and training candidates to become SEAL Officers is resource-intensive, incurring a financial cost to the Navy, manning challenges for NSW and the time and energy of the candidates themselves. Removing candidates who have a low probability of success at BUD/S early in the process reduces the costs and allows them to be reassigned within the Navy to a community more appropriate to their abilities.

This thesis aims to analyze data collected on the candidates to train statistical models capable of predicting a candidate's probability of success in the first phase of BUD/S and inform data collection to capture information on candidates for future data analysis.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
AUC	Area Under the (ROC) Curve
BO	Basic Orientation
BTC	Basic Training Command
BUD/S	Basic Underwater Demolition/SEAL
CART	Classification and Regression Tree
COMPC	Complete Cases
DIM	Data Imputed with MICE
DLIFLC	Defense Language Institute Foreign Language Center
DLPT	Defense Language Proficiency Test
DOR	Drop On Request
GPA	Grade Point Average
MAR	Missing At Random
MCAR	Missingness Completely At Random
MICE	Multivariate Imputation by Chained Equations
ML	Machine Learning
MNAR	Missing Not At Random
NATRACOM	Naval Air Training Command
NDE	Naval Special Warfare Data Environment
NSW	Naval Special Warfare
NSWO	Naval Special Warfare Orientation
OCM	Officer Community Management
OCS	Officer Candidate School
OCSAD	Officer Candidate School Active Duty
OIC	Officer in Charge
OOB	Out-of-Bag
PST	Physical Screening Test
PTRR	Physical Therapy Rest and Recovery
RF	Random Forests
ROC	Receiver Operating Characteristics

SEA	Senior Enlisted Advisor
SOAS	SEAL Officer Assessment and Selection
SQT	SEAL Qualification Training
UDWM	Unimputed Data with Missingness
USNA	United States Naval Academy

EXECUTIVE SUMMARY

Every year the SEAL Officer Community Management (OCM) receives approximately 300 applications from candidates who want to become Navy SEAL officers. Basic Underwater Demolition/SEAL (BUD/S) training is the main obstacle that the candidates must complete to meet their objective, but it is one stage in a rigorous and multifaceted training process that begins with the application. The applications include multiple data elements such as university attendance, grade point average (GPA) and physical fitness test scores. From this data, the SEAL OCM selects approximately 80 percent of the candidates to participate in SEAL Officer Assessment and Selection (SOAS). At SOAS, the candidates are assessed by the cadre at Naval Special Warfare (NSW) Basic Training Command (BTC), who provide evaluation data and a recommendation to the SEAL OCM for candidate selection to BUD/S, to which fewer than half can be selected.

In total, the process of assessing, selecting and training candidates to become SEAL officers is resource-intensive, incurring a financial cost to the Navy, manning challenges for NSW and the time and energy of the candidates themselves. Removing candidates who have a low probability of success at BUD/S early in the process reduces the costs and allows for them to be reassigned within the Navy to a community more appropriate to their abilities.

In this thesis, we analyze data collected on the candidates to train statistical models capable of predicting a candidate's probability of success in the first phase of BUD/S. We focus on success in the first phase because that is the primary test of a candidate's will to succeed in the program. Subsequent phases are designed to assess a candidate's ability to meet more technical performance requirements.

We find that the data collected at each stage of the training process is informative of a candidate's probability of transitioning to the next stage of training but loses the power to predict outcomes beyond that. Models trained on application data are 85 percent accurate at predicting candidates selected for SOAS but only 76 percent accurate at predicting which candidates will ultimately be successful in the first phase of BUD/S. Additionally, models

trained on both application data and data collected at SOAS are 82 percent accurate in predicting a candidate's probability of being selected for BUD/S, but only 75 percent accurate at predicting success in the first phase.

The discounted value of data in predicting subsequent event outcomes suggests that the SEAL OCM is selecting candidates effectively based on the information available. If this were not the case and a data element collected in the applications were predictive of candidate success at BUD/S, it would suggest that the SEAL OCM was not fully utilizing the data to make informed decisions. In addition, if application data were reliably predictive of candidate success in the first phase of BUD/S, it would suggest that SOAS is unnecessary for making this determination and could be discontinued, saving considerable cost.

Particularly challenging is predicting which candidates will fail in the first phase of BUD/S. Our models gain most of their accuracy from predicting positive outcomes in the first phase relying on the high (80 percent) success rate of the officers that make it to that point. Because the model for predicting candidate success in the first phase of BUD/S achieves 75 percent accuracy, a naïve model that predicts success for all candidates achieves a higher rate of correct classification. This suggests that the first phase of BUD/S is an indispensable part of the training process and it correspondingly collects data on a candidate's determination that is unknown before that point.

Our research provides NSW with useful insights from data collected on candidates to influence future data collection efforts as well as statistical models to predict candidate outcomes at each stage of training that can be used to inform candidate selection criteria. Of particular interest is the feature importance outputs from the models that suggest that the subjective candidate assessments provided by the SOAS cadre are the most important factors in predicting candidate selection to BUD/S and success in the first phase.

ACKNOWLEDGMENTS

Thank you to my advisor, Professor Koyak, and my second reader, Professor Buttrey, for your guidance. And thank you to the many individuals at NSW who helped shape this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. THESIS OBJECTIVE

Naval Special Warfare (NSW) is the U.S. military's premier maritime special operations force. Within NSW, the Navy's SEAL Teams are comprised of highly trained and motivated operators capable of conducting special operations missions from the sea, air or land. Basic Underwater Demolition/SEAL (BUD/S) training is the fundamental course to prepare students, as they are commonly referred to, for service in the SEAL Teams and to reject those unqualified for that service. At its core, BUD/S training is designed to induce physical and mental exhaustion to test students' ability to adapt to challenging conditions, their will to succeed, and their determination to commit themselves to a rigorous career in the SEAL Teams. Over the years, BUD/S training has evolved, but an attrition rate of approximately 70 percent for all students has remained relatively constant despite programs instituted by NSW to better select and prepare candidates for training (Atlamazoglou 2021). For officers, the attrition rate is lower, approximately 30 percent, due to the more rigorous training that officer candidates are subjected to before BUD/S. Increasing the number of assessment and screening phases prior to BUD/S may further decrease the attrition rate but at a financial cost to the Navy and an expenditure of time and energy by the instructor staff and the candidates themselves. Because these costs are higher when candidates matriculate into a BUD/S class than during selection, rejecting candidates who are less likely to complete training improves efficiency but increases the risk of rejecting those who would have completed training and provided needed manning to the force.

The goal of this thesis is to develop a statistical model for predicting the success of a SEAL candidate at various points in the training process, using information that is available about the candidate at the point. Such a model could be used by the SEAL Officer Community Management (OCM) to inform candidate selection for attendance to SEAL Officer Assessment and Selection (SOAS), the primary screening program for officers, based on initial data collected from candidate applications, and to BUD/S based on additional data collected at SOAS.

B. THE PROCESS OF BECOMING A NAVY SEAL OFFICER

A Navy SEAL officer candidate must endure a rigorous and multifaceted training process in order to succeed. We now explain this process in more detail.

1. Application

Officers apply for BUD/S training by submitting a package to the SEAL OCM. Each package includes the candidate's medical screening results, letters of recommendation, educational information such as university attendance and grade point average (GPA), prior service (if applicable), and his or her Physical Screening Test (PST) score described in Appendix B.

The SEAL OCM typically receives applications from over 300 candidates annually, of which approximately 70 matriculate into SEAL training. The SEAL OCM conducts an initial "down select" on the applications to eliminate roughly one-third of the candidates who are physically or academically uncompetitive with peers within their accession source. The approximate 200 candidates remaining are invited to SOAS.

2. SOAS

During one of the four two-week SOAS blocks offered in May, June, July, and August, candidates undergo various training evolutions and tests designed to rank the students by performance and reveal aspects of their character (U.S. Navy 2021). Some evolutions, such as the PST, are held every year for every block, but most are not; these evolutions vary as the SOAS cadre rotates into and out of the command or as NSW leadership prioritizes different characteristics. This poses a challenge to data analysis but may be a feature that provides value to the organization.

Based on a candidate's initial packages, the data collected during SOAS, and the recommendations of the SOAS cadre, the SEAL OCM selects approximately 70 candidates to receive orders to NSW Basic Training Command (BTC) and matriculate into SEAL training as students.

3. NSWO

Upon arrival at BTC, candidates report for Naval Special Warfare Orientation (NSWO). In NSWO, officer and enlisted candidates are brought together for the first time and become a class. Over the course of three weeks, the NSWO cadre instructs the class on cleaning and maintaining their gear, completing the various training evolutions to standard, and upholding the Navy and NSW values. NSWO is not designed to down select students but rather to prepare them for BUD/S training. However, not all students proceed from NSWO to BUD/S.

At any time and in any phase, a student will be removed from training if that student fails to meet the standards set in NSWO. Typically, students who fail to meet a physical standard but show strong will and character can stay in BTC, work on their area of weakness, and receive an opportunity to join a later class. This is called a Performance Roll and typically sets students back by a single class. But students who continue to fall short on physical standards or fail to meet the ethical standard are removed from training entirely. This is called a Performance Drop and results in the student's reassignment within the Navy or termination of service.

In prior years, NSWO went by the name Basic Orientation (BO) or Indoctrination ("Indoc"). The data available on these legacy programs is sparse. BTC captured little data prior to 2013 and has not maintained it since that time. In addition, the purpose of the training block has changed over time from arduous assessment and selection to the current focus of preparing the class for BUD/S. For this reason, data from the legacy programs does not align with the data collected more recently by NSWO.

4. BUD/S

BUD/S training is a 21-week training program consisting of three seven-week phases. Each phase has a unique focus, building on the foundations of the preceding phase with the overall goal of preparing students for service in the SEAL Teams and removing those who do not meet the requirements for that service.

The first phase is designed to test students physically to assess their mental grit and the strength of their will to be in the SEAL Teams. Some tests are meant to specifically

assess students' water competency, strength, endurance, capacity to work as part of a team, and other characteristics, with the purpose of inducing exhaustion and discomfort so that students learn to overcome those temporary conditions. Whether from a self-realized lack of ability or will, a BUD/S student can, at any time and in any phase, ask to be removed from training. This is called Drop on Request (DOR) and typically occurs within the first four weeks of the first phase.

The fourth week of the first phase, known as Hell Week, is more similar to a single weeklong evolution than a week with many individual evolutions. Over five days, students run over 200 miles, sleep for less than four hours total and spend little time dry or warm. While most DORs take place in the first four weeks of the first phase, most of those occur within the first three days of Hell Week.

Those students who made it through Hell Week but are physically unable to proceed in training due to injury go to Physical Therapy Rest and Recovery (PTRR). PTRR gives injured students access to medical treatment, physical therapy, and time to heal. Many students pass through PTRR during their time in BUD/S, most commonly after Hell Week, but an injury sustained in any training phase will send a student to PTRR for recovery before they return to training. Those students who cannot fully recover and return to training are removed from BTC for medical reasons. This is known as a medical drop and, although it is less common than the other reasons for students being removed from training, it is observed in the data.

In the second phase, students are trained in combat diving and tested on their ability to execute procedures under stressful, underwater conditions. Attrition tends to be lower in the second phase than in the first, but performance rolls and drops take over for DORs as the leading causes for attrition and continue to be for the remainder of BUD/S.

The third phase is the final seven weeks of BUD/S and mainly takes place on San Clemente Island. This phase is commonly referred to as land warfare and incorporates land navigation, weapons training and instruction on the foundational tactics used by Navy SEALs.

After the third phase, students graduate from BUD/S but not from training. They continue on to SEAL Qualification Training (SQT), where, over the course of a year, they develop the tactical and technical proficiency required to join the SEAL Teams. As with BUD/S, students who fail to meet performance and ethical standards are dropped from SQT. Those who complete SQT are assigned to either an East or West Coast-based SEAL Team.

C. ORGANIZATION OF THE THESIS

The remainder of this thesis is focuses on exploring the data we utilize and the techniques we employ to develop statistical models for predicting candidate success and analyzing model performance. Chapter II is a review of a selection of published work of relevance to this thesis. Chapter III explains the data available for this research and the methodology we use to train and assess statistical models, as well as a brief explanation of the modeling techniques used. Chapter IV presents an analysis of the models and provides empirical examples to demonstrate model performance. Lastly, Chapter V summarizes our findings and suggests areas for future research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

Although the SEALs are a unique warfighting force in the U.S. military, their multistage training and development have structural similarities to the training processes of other organizations, such as the naval aviation community. We examine several studies from the technical literature that have focused on these organizations.

A. NAVAL AIR TRAINING COMMAND PRIMARY FLIGHT TRAINING

Analyzing data from Naval Air Training Command (NATRACOM), Erjavec (2019) develops statistical models of a flight student's probability of completing primary flight training ("primary") at three points in the training pipeline prior to beginning primary. Because primary has a much higher per-student cost ($200,000) than the preceding two stages of pilot training, there are significant cost savings associated with removing students who have a low probability of success at primary (Erjavec 2019). Similar to this thesis, Erjavec utilizes techniques based on decision trees to develop statistical models to predict a student's probability of success. The models that Erjavec presents offer a valuable tool for NATRACOM to assess flight students and reduce costly attrition at primary.

B. U.S. MARINE CORPS FORCE RECONNAISSANCE

Using data from the U.S. Marine Corps Force Reconnaissance School, Nowicki (2017) applies logistic regression to predict student outcomes based on students' entry data, such as demographic information and test scores. The author can identify statistically significant predictors that include physical fitness scores, Marine Corps enlistment cognition test scores and completion of one or more semesters of college courses as positively correlating with candidate success. In addition, he finds that repeating training tasks after failure increases candidates' likelihood of success at given tasks, suggesting candidates can improve over time. Nowicki cautions, however, that the school's mission is to select those Marines who are suited for force reconnaissance and repeat attempts should be allocated sparingly.

C. DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER

Bermudez-Mendez (2020) uses four logistic regression models to predict a student's ability to meet the new aptitude standard on the Defense Language Proficiency Test (DLPT), which is the culmination of their time at The Defense Language Institute Foreign Language Center (DLIFLC). The first model considers a student's entry data, such as prior education and the number of languages the student knows before DLIFLC. This model has mediocre performance for predicting students that will meet the aptitude standard, correctly identifying 25 percent of those that would, but moderate performance at predicting those that will fail, correctly identifying 87 percent (Bermudez-Mendez 2020, p. 21). He trains the other three models at the end of the first, second and third semesters incorporating academic performance data from those semesters. The models' performance increases at each subsequent milestone, with the third model having the most data inputs and the highest accuracy, correctly classifying 69 percent of the students who pass and 82 percent who fail (Bermudez-Mendez 2020, p. 18-19).

III. DATA AND METHODOLOGY

> We continue to modernize our enterprise data engineering services to accommodate an expansion of Naval Special Warfare AI/Machine Learning (ML) capabilities. The Naval Special Warfare Data Environment (NDE) is the foundational plumbing to support *structured and unstructured authoritative data* from disparate, disconnected, and internal data sources to access, ingest, cleanse, curate, store, model, and analyze data efficiently and effectively. The NDE includes Operational Data Stores, Warehouses, and Lakes in order to connect business systems, data analytics, and machine learning initiatives with authoritative data. (Rear Admiral Howard before the 117th Congress Senate Armed Service Committee. April 28, 2021)

Naval Special Warfare Command has established a central repository for data from throughout the NSW community to enable data analysis and Artificial Intelligence (AI) and Machine Learning (ML) application development (Howard 2021). We worked with the data engineers leading that effort, who provided key insights into the data to help guide our research. However, it is a new initiative and older data, particularly unstructured, older training data, is unavailable in the Naval Special Warfare Data Environment. Instead, we received most of the data from the SEAL OCM, SOAS and BTC as approximately 30 Microsoft Excel files, each in a unique format.

Buttrey and Whitaker (2018) estimate that 80 percent of data analysis is time spent preparing data that in real life is "messy." Of particular relevance to this thesis are the effects of missingness, duplication, and inconsistency that occur when merging data from multiple sources (Buttrey and Whitaker 2018). These challenges are the focus of the remainder of the chapter.

A. DATA COLLECTION AND MERGING

The first task for preparing the data was to consolidate the disparate files to merge information on SEAL candidates. This is important because candidates may apply multiple times; and, while their basic demographic and educational history information does not change from one attempt to the next, other information such as PST score are subject to change. One challenge to this task is that the data each command maintains changes over time to fit the needs of the decision-makers for that year. So, the SEAL OCM collects PST

scores for each event some years, but only the cumulative score for others. In addition, the separate commands may collect the same data on candidates, but the values that each command records do not always agree.

B. DATA CLEANING

Once the data is consolidated so that each row represents an individual candidate's unique attempt, the task shifts to dropping those columns with too many missing values or combining similar columns to form a composite metric. Working closely with the SOAS leadership to understand the individual elements, we were able to combine the data from similar events that were held during different years. Figure 1 shows an example of how this is implemented with multiple events, each aimed at assessing a candidate's ability to think critically. Once the observations are consolidated into the data element "CRITICALTHINKING," they are grouped by cohort and scaled so that this single data element provides a common attribute to assess candidates from different cohorts.

Sometimes the value of the missing data can be deduced from what is available, as is the case when one value from the PST scores is missing because that value can be calculated as a linear combination of the other five. Given enough data and assuming that the contradictions and missingness are randomly distributed, deletion would be preferable for those affected observations that cannot be reliably inferred. However, with this data, missingness, in particular, affects a large percent of observations, so we used our best judgment to delete only the most problematic observations or, in the case of contradiction, select one of the values based on the best information available. At the end of the data cleaning and merging phase, missingness is still prevalent, with 136 of the 166 total attributes missing values for more than 50 percent of the observations.

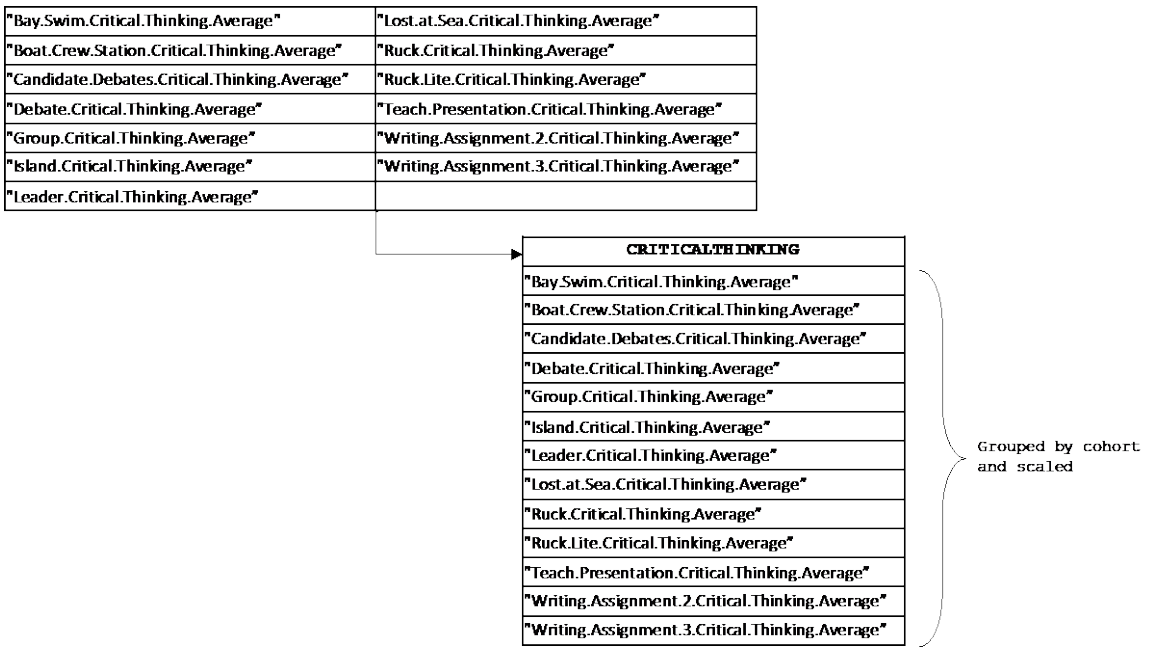

Figure 1. Consolidating Like Data Elements to Form a New Predictor

C. MISSINGNESS

Before discussing methods for resolving missingness in data, it is important to understand the reasons for missingness. In multilevel data, such as NSW's candidate data, attribute missingness may be a feature of level dependence. For example, in the candidate data, observations are present for variables collected during SOAS only if that candidate was selected for SOAS on that specific attempt. It would be nonsensical to fill SOAS variables for those candidates/attempt observations that did not make it to SOAS, so we will ignore the type of missingness that is entirely a feature of level-based attrition for the remainder of the thesis and focus on missingness within observations inherent to each level.

Missingness Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) are essential ideas to understanding the nature of missingness in data and the methods available to handle it. MCAR implies that missingness is entirely unrelated to the data itself so that the occurrence of a missing value in a variable or group of variables is independent across observations and randomly distributed (Van Buuren 2018). This is rare in that it requires unique circumstances to produce such missingness at random that is not influenced by time or observation groupings.

MAR is less restrictive than MCAR in that the grouping of observations influences the probability of missingness by some attribute or the time at which the operations took place so that within each group, the probability of missingness is the same.

Finally, MNAR describes missingness that is influenced by some factor that is unknown and cannot be accounted for; for instance, if the data consists of measurements from a scale that malfunctions with greater frequency over time, but the time and order of observations were lost, then the specific feature that influences missingness is unknown to the researcher but still influential (Van Buuren 2018). In this data, MCAR, MAR and MNAR may all be applicable. In 2018 some variables were not recorded, implying MAR, while other values are missing without explanation. This presents a challenge for analysis, but there are techniques for overcoming the issue. For a more technical explanation of MCAR, MAR and MNAR, refer to Van Buuren (2018).

D. IMPUTATION

Imputation is a process of replacing missing values with estimates. Commonly used methods for imputation include using either the mean, median or mode of the complete cases of a variable to replace all missing values with the same estimate. This method of imputation is easy to implement. In addition, using the mean for imputation does not change the sample mean, which has some strengths for statistical practice, but it does have some important drawbacks, including reducing variance and obscuring relationships among variables (Van Buuren 2018).

Regardless of which of these basic imputation methods are used, the result is an overly exact estimation that does not accurately quantify uncertainty surrounding the estimate or correlations among variables that may be important to the value being approximated (Azur et al. 2011). Regression techniques take into account correlations but are difficult to implement when missingness is distributed throughout the data and not limited to one variable. Additionally, these techniques do not account for the ambiguity surrounding missingness. Finally, most basic methods for imputation are valid only under the assumption of MCAR, which often do not apply.

1. When and Why

Imputation is an attractive idea for data analysis as it offers an option for saving observations from being discarded due to missingness. This is especially true on smaller datasets as many modeling techniques only work on complete cases and those may be too few to draw both a training and test set from. Separating training and test sets is essential for providing a true means for evaluating a model's performance by training it on a randomly sampled subset of about 70 to 80 percent of the original data and then testing its performance on the held-out data. If an honest assessment of a modeling technique on data with few complete cases is required, then imputation may be necessary to provide a more substantial set of complete cases.

However, imputation is not a replacement for good data collection and management. There is a cost to imputation in the form of a loss of degrees of freedom and a reduction of the model's accuracy. In our research, we evaluated the performance of models with imputed values for variables that had as much as 49 percent missingness but found that model accuracy notably deteriorated with more than 20 percent missingness.

2. MICE

Missingness is a ubiquitous feature of human observational data (Rubin 1996). Therefore, it follows that researchers interested in fields such as medical research, where observations may be few, have pioneered methods for dealing with that missingness that go beyond mean imputation or deletion. One such method is Multivariate Imputation by Chained Equations (MICE), which builds simulated sets of imputed data, perhaps 5, 20 or 100, that reflect the uncertainty of the estimate (Van Buuren 2018).

The MICE algorithm begins by imputing the data using a variable-dependent, basic imputation method and then trains a model on a randomly sampled training set using user-selected modeling techniques appropriate for each variable (King et al. 2001). To produce multiple sets of imputed values, the algorithm uses random resampling. Figure 2 demonstrates this process for one missing value, but the same procedure is followed for every missing value, in this case using five imputed values for each missing one.

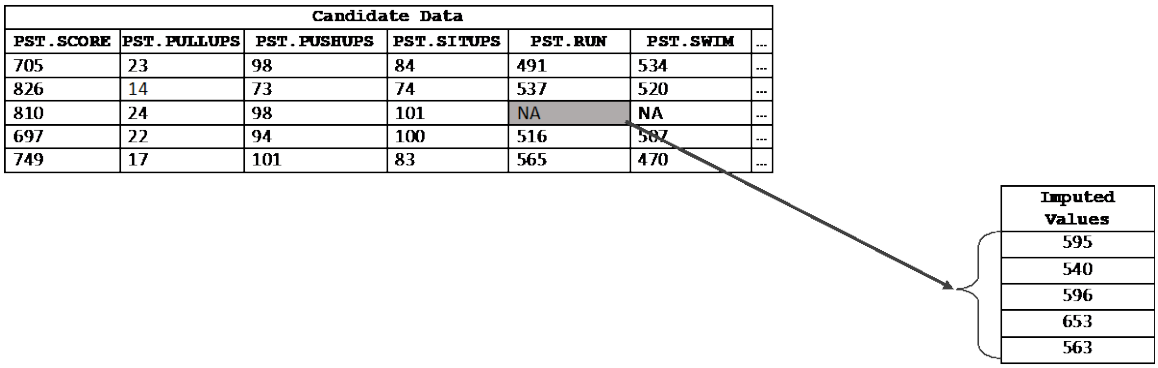

Figure 2. MICE on One Value Using Five Imputations

Researchers can then explore features of this simulated data, such as variance, to determine the fit of the imputed values, higher variance indicating a lack of specificity in the model and reason to doubt the imputed values. In addition, the simulated data should approximately follow the distribution of the present values. Figure 3 depicts the density plots of simulated data compared to the observations using ten imputations. The simulated data is plotted in red, while the observed data is plotted in blue. That the simulated data holds close to the observed data indicates a good fit in line with what is expected for MAR values. If the variance is low and the imputed values appear plausible, then the models which produced those values can be "pooled" to produce a single estimate to replace the missing value.

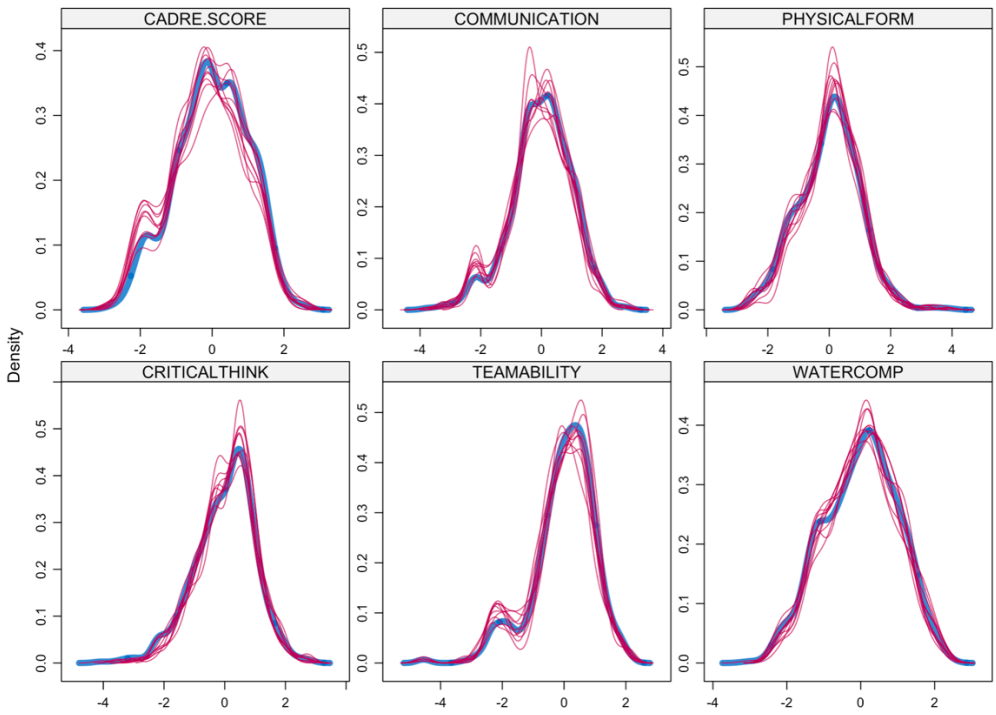

Figure 3. Density Plot of Ten Sets of Simulated Data (Red) and True Data (Blue)

One strength of MICE is that it can be utilized under the assumption of either MCAR, MAR or MNAR (Van Buuren 2018). This makes MICE the ideal technique for this data, as the modeling techniques for each variable can be adjusted to accommodate for missingness assumptions.

E. DATASETS

Given the techniques discussed, three options are available: to impute missing values, delete observations with missing values, or do neither and analyze data with some portion of missingness. We consider each of these options: Complete Cases (COMPC), Data Imputed with MICE (DIM), and Unimputed Data with Missingness (UDWM). This enables us to explore the benefits and limitations, as well as compare the outputs of models trained from each. All three datasets have the same 22 variables, defined in Table 1 and expanded on in Appendix A, that were chosen, in part, because they were the only variables of the original 166 with less than 50 percent missingness.

Table 1. List and Description of Variables

Variable	Description
ACCESSION	Source the candidate came from
COLLEGE.FACT	Type of college the candidate attended
PST.SCORE	PST score submitted by the candidate
PST.PULLUPS	Number of pullups performed
PST.SITUPS	Number of sit-ups performed
PST.PUSHUPS	Number of pushups performed
PST.RUN	Number of seconds to run 1.5 miles
PST.SWIM	Number of seconds to swim 500 yards
GPA	College grade point average
CADRE.SCORE	Overall score given by SOAS cadre
COMMUNICATION	Ability to communicate clearly and concisely
PHYSICALFORM	Average form while conducting exercises
CRITICALTHINK	Average performance during critical thinking drills
TEAMABILITY	Willingness and ability to work as part of a team
WATERCOMP	Performance and perceived comfort level in the water
SOAS.SELECT	Selected for SOAS or not
BUDS.SELECT	Selected for BUD/S or not
BROWNSHIRT	Completed Hell Week or not

Each dataset can also be broken down by those variables collected from the candidates' applications and those collected during SOAS. This mostly impacts COMPC because there are fewer complete cases when considering all 22 variables than with the 12 application variables. Figure 4 depicts the dimensions of the datasets. COMPC is the smallest of the three, and DIM is approximately the same length as UDWM except for two observations that were removed because they could not be reliably imputed.

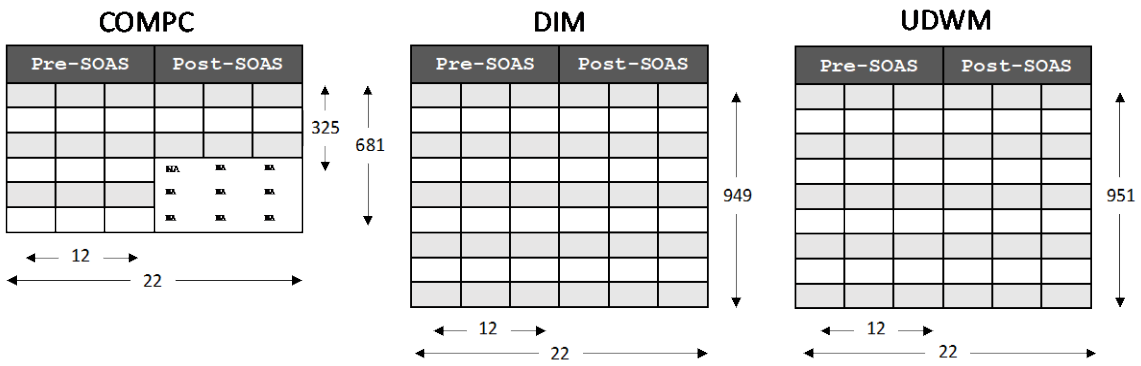

Figure 4. Dataset Dimensions

Across the datasets, the models should generally agree, but some dissimilarity is expected based on the unique artifacts of each dataset. For instance, the COMPC data does not include any observations for 2018 because no SOAS data was collected in that year. While the SOAS data for that year is not used in the estimation of any model that is within the scope of our analysis, the models trained on DIM and UDWM data are influenced by application data on outcomes for that same year. Comparing the models builds our trust in their outputs and gives a sense of the cost that missingness imposes on our analysis.

F. SCALING

Because SOAS is conducted in four blocks each summer, the data collected during these blocks should be scaled so that values are adjusted to account for the influences unique to those blocks. For example, tides and currents may impact the swim times of candidates during a swimming evolution. Therefore, it would not be reasonable to compare candidates from different blocks that are affected by dissimilar circumstances. Instead, the candidates are grouped by their respective cohorts and the values are scaled by subtracting the cohort mean and dividing by the standard deviation (RDocumentation 2021). After this transformation, candidates from different cohorts can be assessed on a common scale. Unfortunately, a value for which block candidates attended is not reliably present in the data. This, along with significant missingness, influenced our decision to remove variables such as the 1.5 nautical mile swim from consideration.

Other variables that need to be scaled by cohort are the subjective scores assigned by cadre. The grading rubric for candidates remains the same for all SOAS blocks, but different cadres oversee the various blocks and their scoring styles will influence the data if not accounted for. At first glance, it appears that these variables, like 1.5 nautical mile swim performance, must be rejected, but unlike the latter, the subjective variables can be scaled by year. While not all of the cadre members will be the same for each block of SOAS, the Officer in Charge (OIC) and Senior Enlisted Advisor (SEA) do not change within the year and therefore provide the continuity needed to scale the subjective variables such as CADRE.SCORE.

G. PREDICTIVE MODELING USING DECISION TREES

We used tree-based modeling methods, commonly referred to as decision tree methods, due to the interpretability of the models produced with those methods and the advantages that some methods offer to deal with missingness (James et al. 2017, p. 303). To demonstrate how these methods work, we apply them to an example classification problem. Consider the task of classifying items as either class A or class B from a set of observations with three variables *x*, *y* and *z*. For both A and B, the variable *x* is randomly sampled from the normal distribution centered at 50 with a standard deviation of 12 for A and 10 for B, making *x* more spread out for A than B, and the variable *y* is randomly sampled from the uniform distribution between 0 and 100 for both. Variable *z* is defined as follows. For class A, *z* is equal to 1 when *x* is outside of one standard deviation from the mean and *z* is equal to 0 otherwise. For class B, *z* is equal to 1 if x is within one standard deviation of the mean and *z* is equal to 0 otherwise.

This formulation ensures that the interaction between the *x* and *z* variables is the primary way to determine class. Class B will also tend to have *x* values closer to 50 than A, and the *y* variable acts as random noise. However, to impose uncertainty into the models, we also made *z* equal to 0 for those observations for which the variable *y* is a multiple of 13. This will allow us to explain the concept of impurity. 10,000 observations of each class were generated for a total of 20,000. From this, training and test sets of 80 percent and 20 percent were drawn at random.

Classification and Regression Trees (CART) is a commonly used method for using data to estimate a decision tree. Because this thesis aims to classify candidates by success or failure, we focus on the classification tree and refer to the method as its generic form, CART. Figure 5 shows the CART model for the example problem. The top node in the CART model, which contains all of the data, is referred to as the root. The root is classified as the most common class present in the node. From the root, the model splits the observations into two child nodes based on the criterion stated below the node. In this model, the criterion is z equal to 0. Observations for which the criterion is true go to the child node 6

on the left, and those for which it is false go to the right. The criterion is chosen to maximize the contrast between the two child nodes. The model proceeds recursively in the same manner on the unsplit nodes until only terminal nodes ("leaves") remain, as shown at the bottom of Figure 5.

Node color is determined by the proportion of observations in the node belonging to each class. The class determination in this example is binary, with class B as the positive case and class A as the negative. For the binary classification, the colors scale from dark blue, indicating few or no positive observations, to dark green, indicating most or all positive observations, with lighter blues and greens representing a mix of classifications.

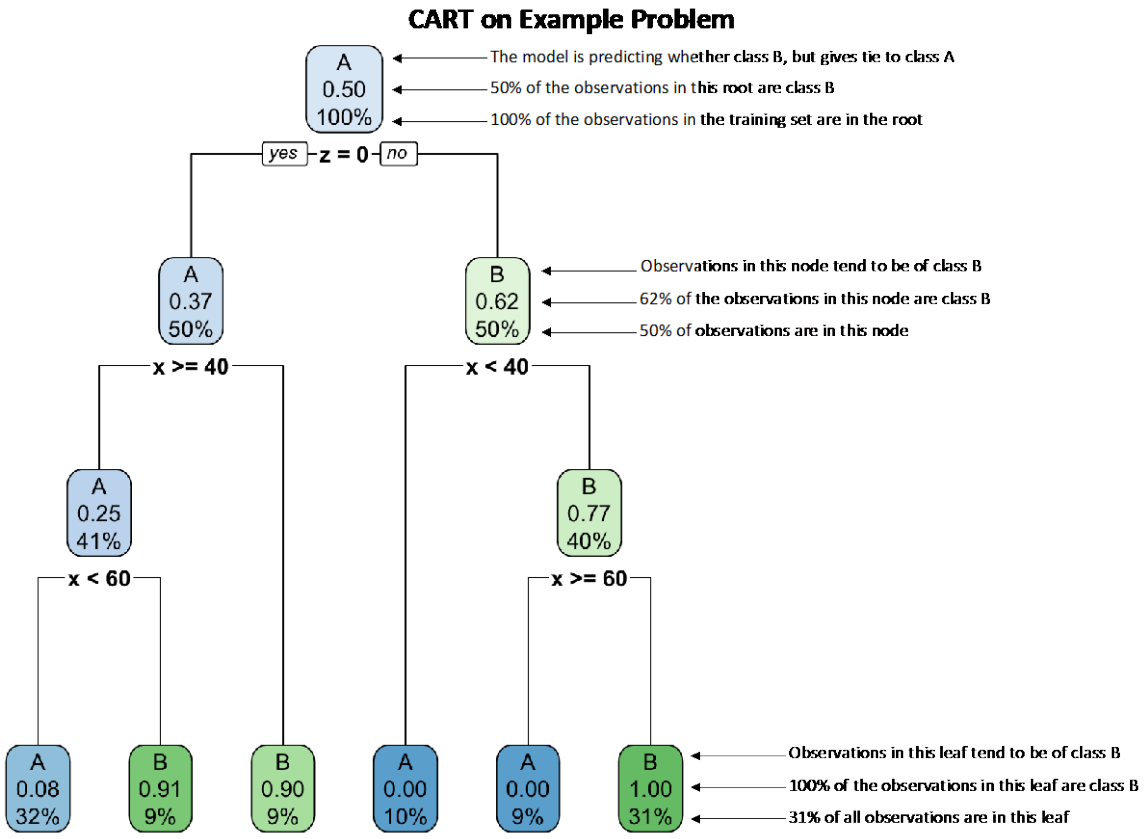

Figure 5. CART Model Applied to Example Problem

To determine which variable to split on, CART uses Gini impurity to select a split for which the two resulting child nodes have the greatest contrast (James et al. 2017, p. 312). Impurity is a measure of how cleanly the classes are separated for that split. The Gini impurity is calculated as

$$Gini\ Impurity = \sum_{i=1}^{2} \left(Gini\ Value_i \right) \times \left(\frac{Total\ Observations\ in\ Node_i}{Total\ Observations\ in\ Both\ Nodes} \right)$$

where the Gini value is calculated for both child nodes and then combined after weighting the values by the proportion of observations from the parent node present in the respective child.

$$Gini\ Value_i = 1 - \left(\frac{Number\ of\ \ Class\ A\ in\ Node_i}{Total\ Observations\ in\ Node_i} \right)^2 - \left(\frac{Number\ of\ \ Class\ B\ in\ Node_i}{Total\ Observations\ in\ Node_i} \right)^2$$

Figure 6 shows the Gini impurity calculated for the first split of the CART model, but the model does this for all possible splits and selects the one with the lowest Gini impurity, and therefore the one that will result in two child nodes with the greatest contrast.

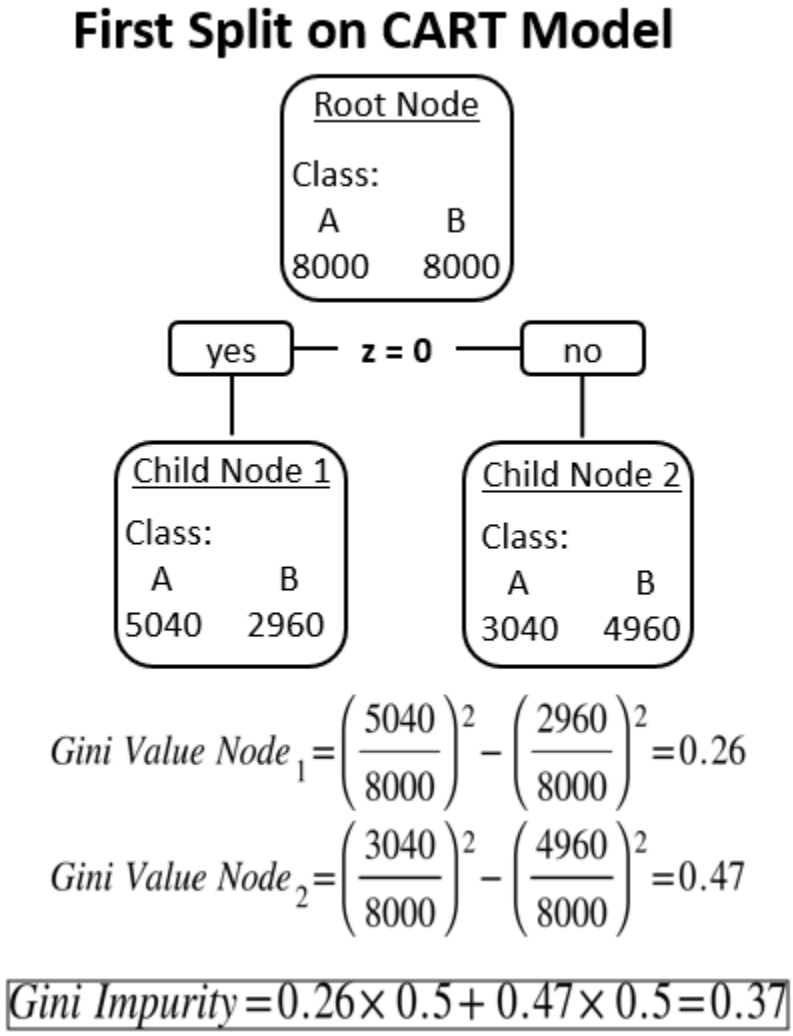

Figure 6. Calculating Gini Impurity to Decide Split in CART Model

Because *z* equal to 0 is the first split, we know that it produces the lowest impurity. The model continues to identify the next best split in the same way until a user-specified cutoff point, usually either the maximum number of splits or a minimum number of observations in the leaves is met (James et al. 2017, p. 308). Once a tree is built using this process, it is pruned to reduce overfitting on the training data. In our investigation, we apply

pruning so that the models have fewer splits; this strategy favors models with greater interpretability and guards against overfitting, but it also limits the depth of interactions that the models can detect.

Another benefit of CART is its handling of missing values. Rather than requiring complete cases or imputation to fill missing data, CART uses what are known as surrogate splits. For a given node, surrogate splits are obtained by considering a set of splitting rules that come closest to mimicking the behavior of the chosen split, using different variables. When the model is later used for prediction with an observation for which the chosen split cannot be calculated due to a missing value, CART uses instead the best surrogate split that can be calculated in its place (Feelders 1999). In cases where there is no similar split based on the missingness of the observation, the observation is placed in the child with the greater quantity of observations and, therefore, a higher naive probability of being the correct choice (Feelders 1999). However, this approach to missingness is not superior to imputation. Feelders (1999) shows that surrogate splitting is often outperformed by multiple imputation.

H. RANDOM FORESTS

Random Forests (RF) is a decision tree method similar to CART, but instead of building one tree, it builds a forest of trees that are dissimilar in their construction based on the data and set of variables available to each (James et al. 2017, p. 319–323). If all data and variables were available when constructing each tree, the result would be a forest of identical trees built using the CART method. Instead, RF randomly samples the data before building each tree using a method called bagging and randomly samples the variables at each opportunity to split (James et al. 2017, p. 319). This reduces the correlation between trees and explores more variable interactions.

The trees in a RF usually are chosen to be of modest complexity, each using only a small, randomly selected subset of the available predictors. The power of a RF is derived from producing a large collection of such trees on bootstrapped samples, each of which is allowed to "vote" in order to make predictions. Aggregation of a large number of trees in this manner produces a predictor that can capture interaction structure with low bias and low variance (James et al. 2017, p. 319–320). In addition, because each tree is built using a

subset of the data that is selected by randomly sampling with replacement, approximately one-third of the data is not selected and is therefore available to provide an honest evaluation of each tree's performance. This is called out-of-bag (OOB) error estimation and is calculated by comparing an observation's true class with the class predicted by the model using the set of trees created without that observation. (James et al. 2017, p. 317–320).

One disadvantage to RF is low interpretability. Attempting to construct a "typical" tree from RF would violate the methods used to construct the forest, and discerning patterns among the many trees in the model is challenging. Fortunately, Gini impurity offers a technique for inferring variable importance. By averaging the decrease in Gini impurity achieved after each split for which a given variable is used across all trees in the forest, the mean decrease Gini impurity offers a measure of overall variable importance, with a higher value suggesting greater importance. The plot on the left of Figure 7 depicts the variable importance for the example problem as determined by the mean decrease in Gini impurity.

Another useful analytic tool offered by RF is the partial dependence plot. This plot shows the marginal effect of a variable on classification outcomes (James et al. 2017, p. 331). For the example problem, the marginal effect of variable x is shown by the partial dependence plot on the right of Figure 7. The RF model discovered that the values of x for class A tend to be less crowded around the mean (50) than for class B. This effect is depicted by the smooth line (blue) that averages the effect, as assessed by the likelihood of an observation being of class A, that is plotted in black.

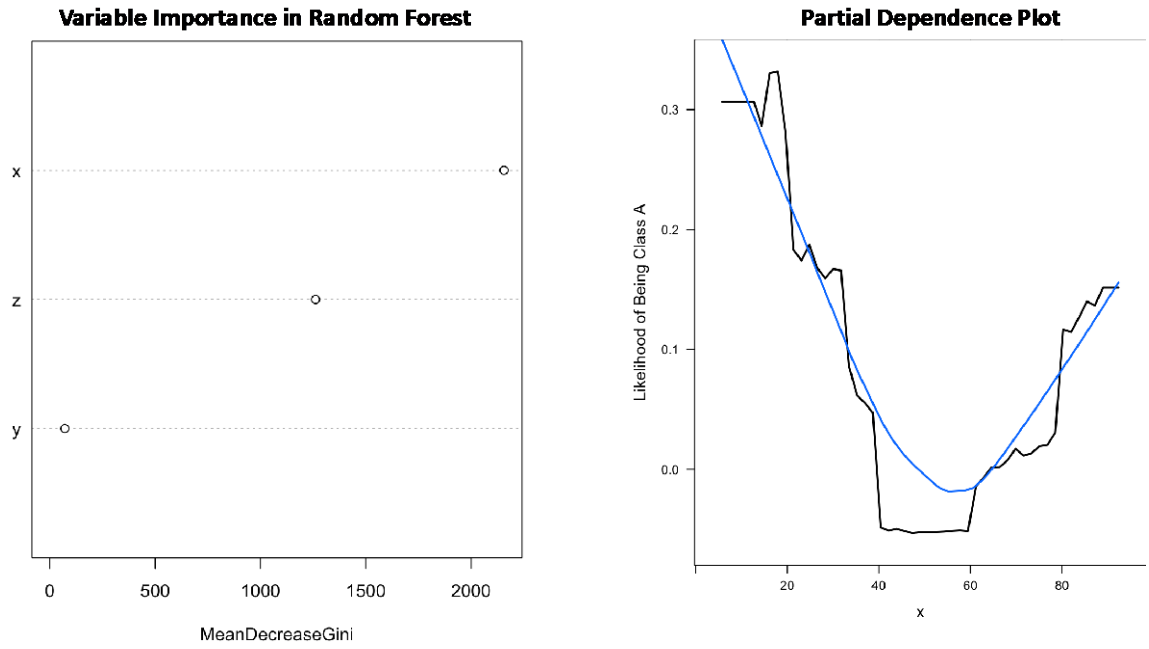

Figure 7. Analytical Tools Available for RF Models

I. MODEL SELECTION

Assuming that no model is perfect at classifying observations, we need to determine which performs the best for our application. A confusion matrix like the two shown in Figure 8 offers a method for contrasting the predicted and observed classifications. From the confusion matrices, the following measures can be calculated. For the sake of discussion, class A is associated with "positive" and class B with "negative."

Accuracy: percentage of correctly classified observations
Sensitivity: percentage of correctly classified As (or positive)
Specificity: percentage of correctly classified Bs (or negative)

Confusion Matrix for CART on Example Problem

Predicted/Observed	Class A	Class B
Class A	1887	98
Class B	45	1970

$$Accuracy = \frac{1887 + 1970}{4000} = 0.96$$

$$Sensitivity = \frac{1887}{1887 + 45} = 0.98$$

$$Specif\ icity = \frac{1970}{1970 + 98} = 0.95$$

Confusion Matrix for RF on Example Problem

Predicted/Observed	Class A	Class B
Class A	1648	543
Class B	284	1525

$$Accuracy = \frac{1648 + 1525}{4000} = 0.79$$

$$Sensitivity = \frac{1648}{1648 + 284} = 0.85$$

$$Specif\ icity = \frac{1525}{1525 + 543} = 0.74$$

Figure 8. Confusion Matrices and Performance Metrics

For this example, the confusion matrices in Figure 8 show that the CART model outperforms the RF model in accuracy, sensitivity and specificity, but this is not always the case. When modeling rare occurrences, for instance, the most accurate model may always predict a negative result and therefore have poor sensitivity. In addition, the models can be adjusted so that specificity or sensitivity are favored by tuning the probability threshold, usually chosen to be 0.5 by default, that an observation must meet to be classified as positive. If, for example, we care more about correctly classifying As and less about incorrectly classifying Bs, we could reduce the probability threshold so that classification as A is less restrictive. Receiver Operating Characteristics (ROC) curves such as the ones plotted in Figure 9 offer a method to discern model performance that considers these limitations of confusion matrices. As shown in Figure 9, the ROC curves are plot using true positive rate “sensitivity” and false positive rate (1-specificity).

Based on the plot shown in Figure 9, the CART model dominates at all points along the curve. More generally, Area Under the Curve (AUC) may be calculated as a numerical measure to determine which is the dominant model. The AUC for these models is 0.98 and 0.92, respectively. An AUC of 1 is perfect, and an AUC of 0.5 is the performance that can be expected from randomly guessing. Although both AUC values indicate strong performance in this example, the CART model dominates RF on both simplicity and performance.

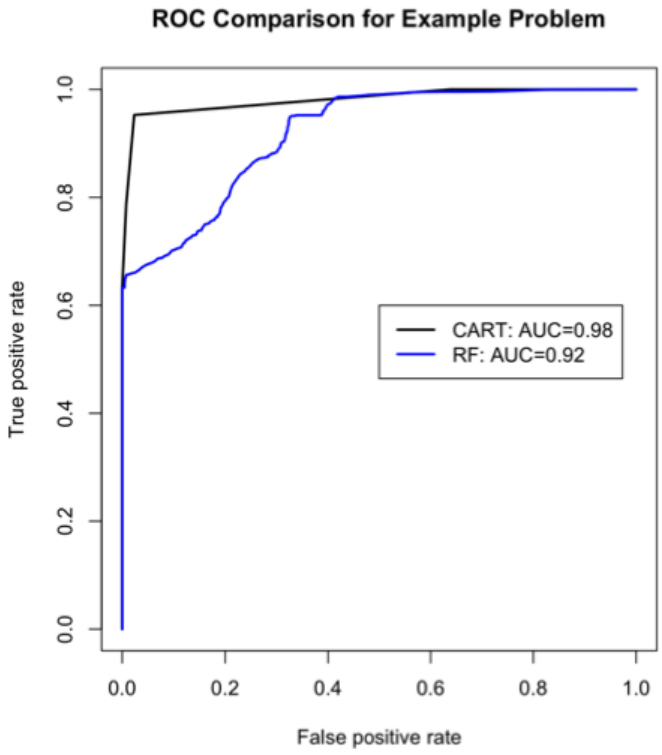

Figure 9. ROC Curves and AUC values

IV. ANALYSIS

This chapter describes the models used to predict candidate outcomes starting from the application phase and terminating at the first phase of BUD/S. The methodology discussed in Chapter III is applied to the three datasets that we examine: COMPC, DIM, and UDWM. We explore individual model strengths, variable interactions and variable importance. Figure 10 gives an overview of the models that we consider.

Model	Predicted Outcome	Candidate Set	Trained On
A.	Success at First Phase	All Applicants	Application Data
B.	Selection to SOAS	All Applicants	Application Data
C.	Selection to BUD/S	Candidates Selected for SOAS	Application Data & SOAS Data
D.	Success at First Phase	Candidates Selected for BUD/S	Application Data & SOAS Data

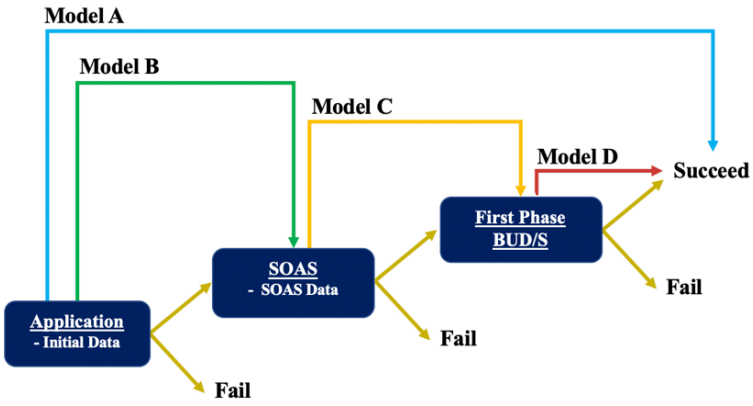

Figure 10. Overview of the Models Depicting the Data Available to Each and the Event Being Predicted

A. MODEL A (INITIAL MODEL)

A valuable model to increase the efficiency of recruiting would use initial data to predict candidates who are most likely to succeed in the first phase of BUD/S. However, this prediction is heavily influenced by deselection events before SOAS and BUD/S. We start by ignoring these deselection events and compare CART and RF models trained on COMPC and a CART model trained on the UDWM. This comparison is shown at the top of Figure 11. Both models show ACCESSION as the most important variable to split on.

On COMPC, CART does not include any additional splits, but on UDWM, the model splits a second time on PST.SCORE, identifying 810 as a threshold value. That these trees generally agree is consistent with the variable importance plot from RF in the lower left of Figure 11.

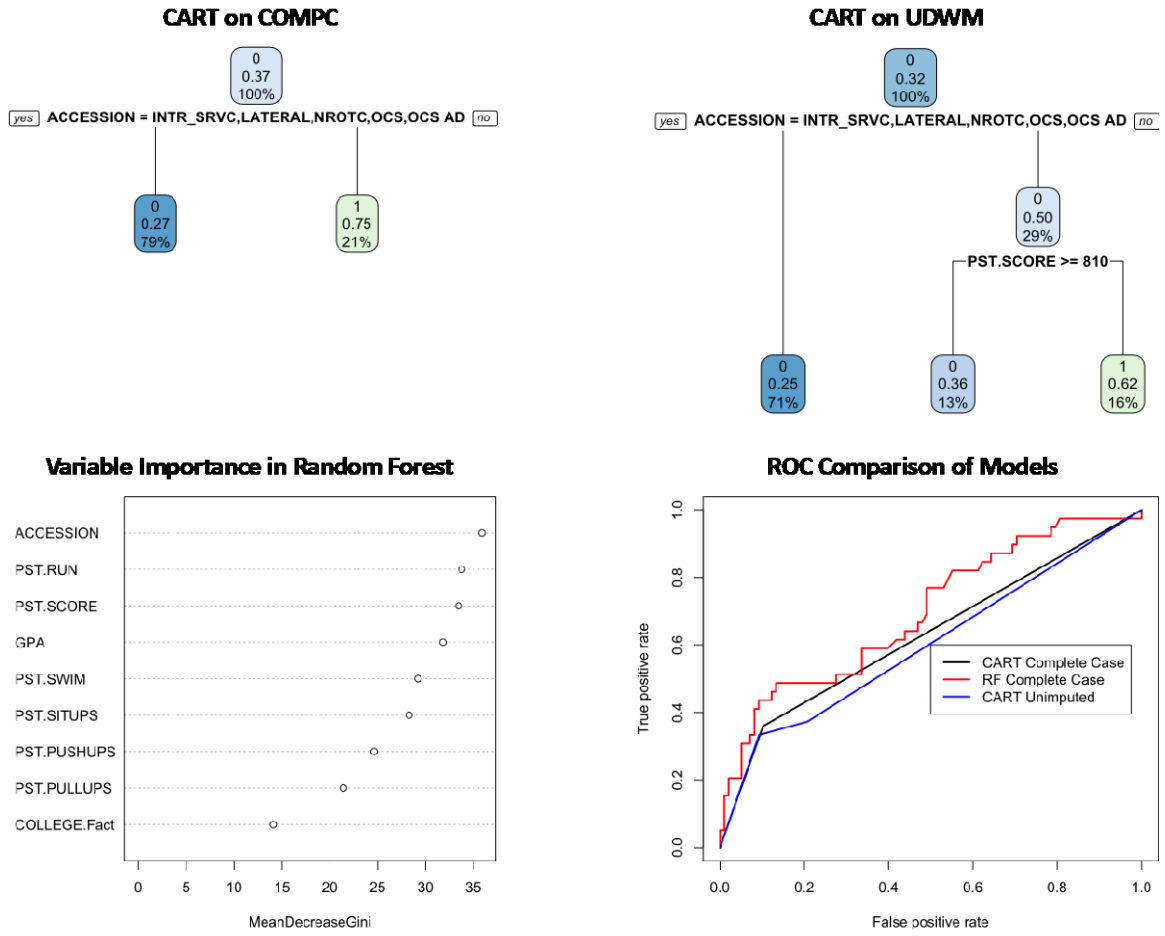

Figure 11. Model A for Predicting Success at the First Phase of BUD/S

The ROC plot in the lower right of Figure 11 indicates that the RF model trained on COMPC is the best performing model. As such, we explore its performance as shown in Table 2, which is used to calculate the following performance characteristics:

Accuracy (percentage of observations correctly classified): 76 percent
Sensitivity (percentage of accurately predicted failures): 44 percent
Specificity (percentage of accurately predicted successes): 89 percent

Table 2. Performance Characteristics of RF Model A

Predicted/Observed	Failure	Success
Failure	87	22
Success	11	17

One weakness of this approach is that the models give little consideration to variables other than ACCESSION. To gain a richer understanding of more variables and their importance, we can use the information available at each stage to predict an outcome at the next. Starting with predicting selection for SOAS based on initial data, then selection for BUD/S by including the data collected at SOAS, and finally predicting completion of the first phase of BUD/S. Understanding what influences success at each stage can help explain the evolution of the candidates that proceed through the stages and how the cohort's characteristics affect the models. For instance, if only candidates from one accession source are selected for SOAS, then ACCESSION will cease to be a valuable predictor at later stages.

B. MODEL B (SELECTION TO SOAS)

The SEAL OCM initial selection of candidates to attend SOAS can also be considered the first deselection of applicants. This means that models for predicting selection to SOAS are the last models we will consider trained and tested on the entire set of observed candidates. The reason for exploring these models is not to predict who will be selected but to update our understanding of the candidates considered at the next stage. The ROC plot in Figure 12 shows that CART and RF models trained on COMPC outperformed the CART model trained on UDWM. In particular, the RF model appears to be the dominant model at most points along the curve. The tree plot of the CART model in Figure 11 shows that candidates from accession sources other than INTR_SRVC, LATERAL and OCS AD account for 81 percent of the observations and have a 92 percent probability of being selected for SOAS.

The tree plot also shows some counterintuitive behavior regarding PST.SITUPS. According to the model, doing more than 110 situps in the PST is negatively associated with

selection to SOAS. The CART model trained on UDWM also detected this phenomenon. The partial dependence plot in the lower right of Figure 11 shows the log-likelihood of being selected for SOAS given the number of PST.SITUPS reported. According to the plot, the optimal number is approximately 90, with a steep decline after 100. There is no clear explanation for this phenomenon. It may be an artifact of the data arising as a feature from the relatively modest sample size or a correlation between PST.SITUPS and a variable not available to the research team.

Outside of PST.SITUPS, all other variable choices for splits make intuitive sense, with the primary split at ACCESSION indicating that as the most important predictor for the CART model. The variable importance plot in the lower right of Figure 12 shows that ACCESSION is also the most important variable as assessed by the RF model.

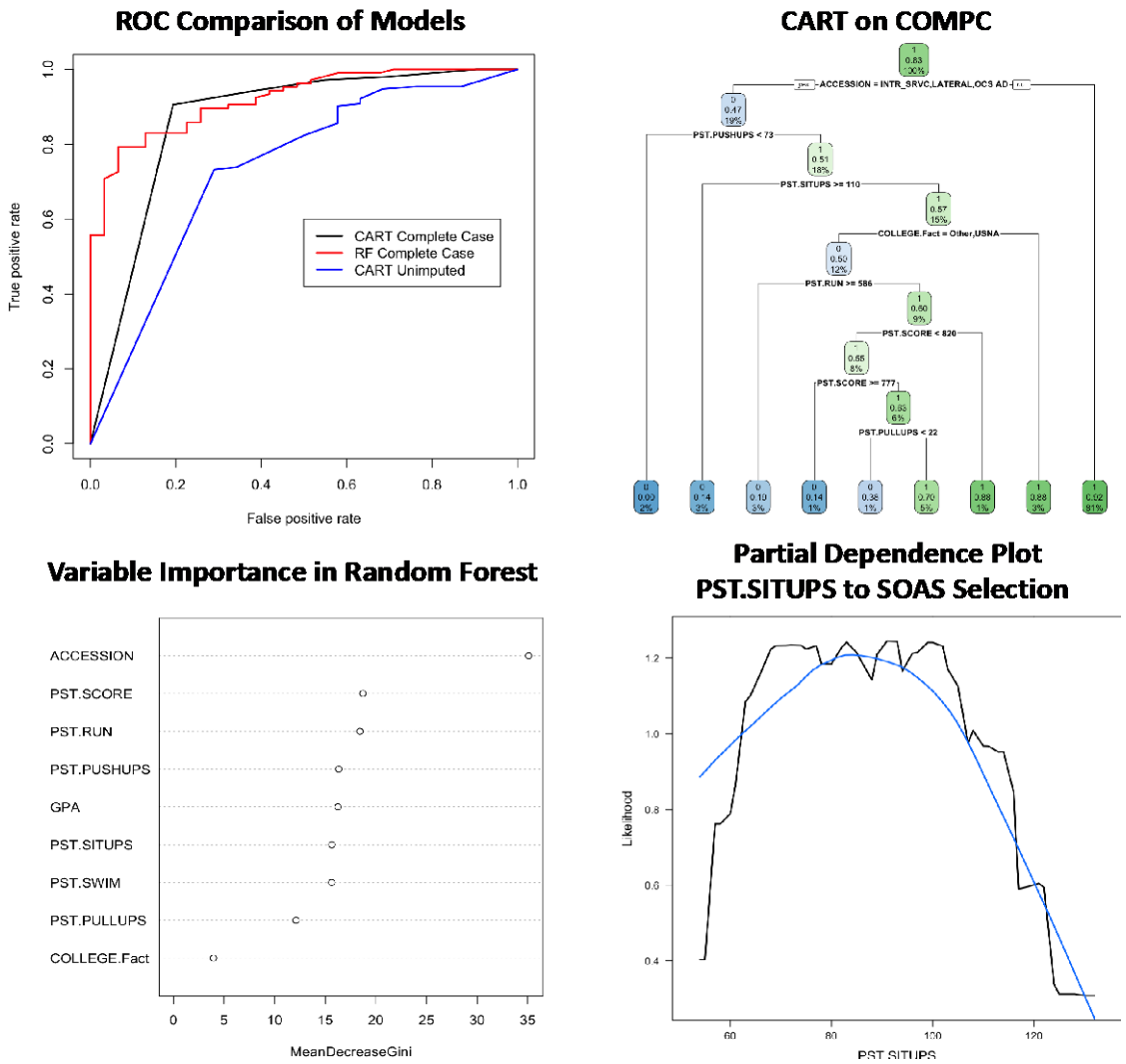

Figure 12. Models for Predicting Candidate Selection to SOAS

In addition to being the best model as indicated by the ROC plot, the RF model is also the top-performing model in regard to its performance characteristics outlined in Table 3, and summarized as follows:

Accuracy: 85 percent
Sensitivity: 95 percent
Specificity: 52 percent

Table 3. Performance Characteristics the RF Model B Predicting Candidate Selection to SOAS

Predicted/Observed	Failure	Success
Failure	16	5
Success	15	101

C. MODEL C (SELECTION TO BUD/S)

Next, we examine selection to BUD/S, given that a candidate was selected for SOAS. For this analysis, we no longer consider those candidates who were not selected for SOAS, and we gain the set of variables collected on candidates who attended SOAS. The ROC plot on the top left of Figure 13 shows RF on COMPC as the dominant model with TEAMABILITY, CADRE.SCORE and PHYSICALFORM providing the most significant mean decrease to the Gini score. The bottom of Figure 13 shows the partial dependence plots for each of these three top predictors, each shown as positively correlated to candidate selection to BUD/S. Of note, none of the initial variables were top predictors for candidate selection to BUD/S.

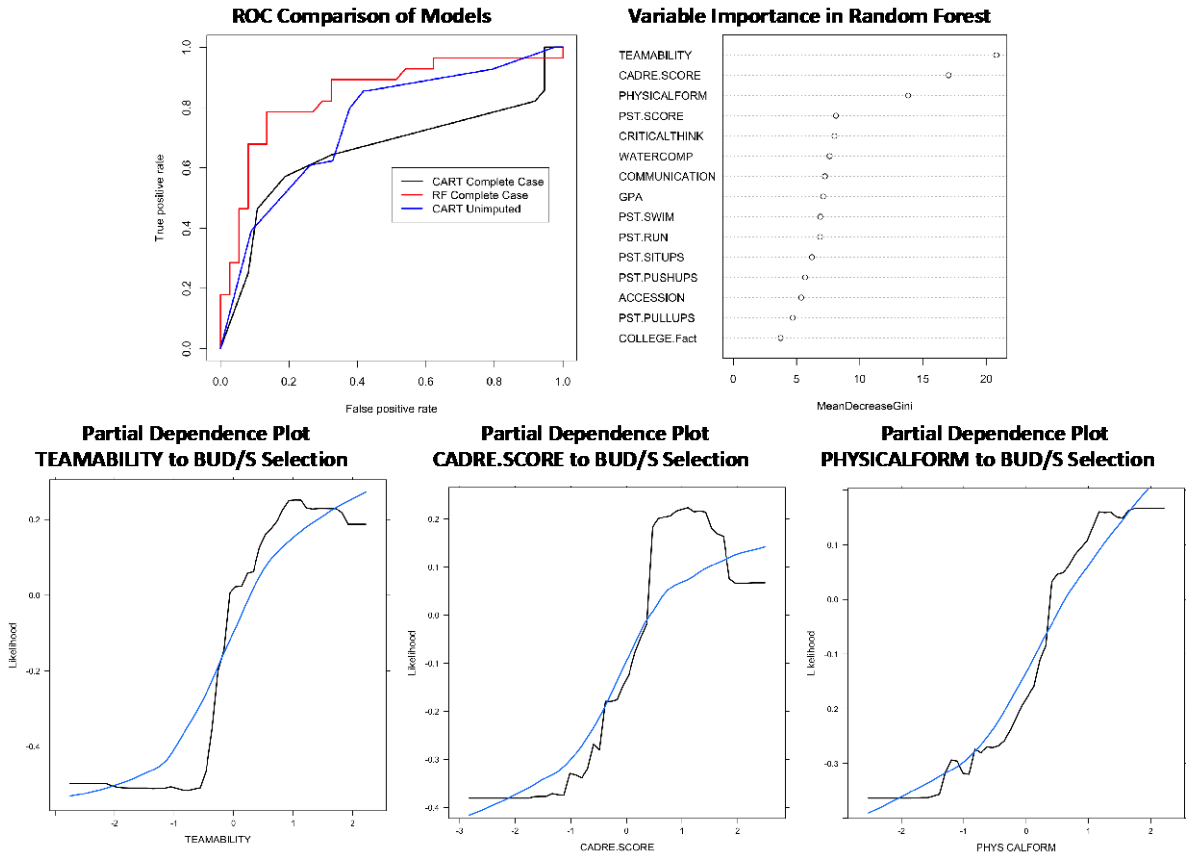

Figure 13. Models for Predicting Candidate Selection to BUD/S

As the top-performing model, the performance characteristics of the RF model trained on COMPC are outlined in Table 4 and summarized as follows:

Accuracy: 82 percent
Sensitivity: 79 percent
Specificity: 84 percent

Table 4. Performance Characteristics the RF Model C Predicting Candidate Selection to BUD/S

Predicted/Observed	Failure	Success
Failure	31	6
Success	6	22

D. MODEL D (PREDICTING COMPLETION OF THE FIRST PHASE OF BUD/S)

Finally, we predict success in the first phase of BUD/S. At this stage, there are only 152 observations in COMPC, resulting in 121 observations for an 80 percent training set. Models trained on such a small dataset can focus on artifacts unique to the sample and not representative of the population, inducing a bias. In addition, the 31 observations in the 20 percent test set cannot provide a realistic judgment of model performance. To reduce these effects, we utilized all 152 observations available in COMPC for training the models. We also pruned the CART Model so that no leaf had fewer than 5 percent of the observations (8 or more). This limits the depth of the model and may eliminate some valuable splits, but it also allows for a more general and transferable model.

With all 152 observations in COMPC being used to train the models, OOB error is used to evaluate the models' performance. This requires using bagging, discussed in Chapter III subsection I, to generate an OOB sample for each of the 100 trees produced with CART. The trees are restricted to the variables selected by the original CART Model D shown in Figure 14, and the same parameters set for that model. These conditions preserve the 100 bagged trees as honest representations of the original but allow for variance of splits and depth. As discussed in Chapter III subsection I, OOB error estimation is a key feature of RF; generating it does not require additional formulation. Upon comparing the OOB error estimates, the RF model emerges as dominant by providing greater accuracy and, in particular, greater specificity at 32 percent compared to 21 percent achieved by the CART model. The results of the RF model are shown in Table 5 and produce the following performance characteristics:

Accuracy: 75 percent
Sensitivity: 86 percent
Specificity: 32 percent

Table 5. Performance Characteristics the RF Model D Using OOB Samples

Predicted/Observed	Failure	Success
Failure	10	17
Success	21	104

While the performance of this model may appear to be worse than that of Model A, the comparison is not justified. That first model benefits from profiling candidates who did not make it to BUD/S based on the initial data and relying on the approximate 78 percent completion rate in the first phase for officers who make it to BUD/S. Model D, in contrast, looks only at those who made it to BUD/S and identified characteristics indicative of those who complete the first phase. The tree diagram on the left side of Figure 14 illustrates the splits chosen by the CART model, while the graphic on the right shows the variable importance as selected by a RF model trained on the same data.

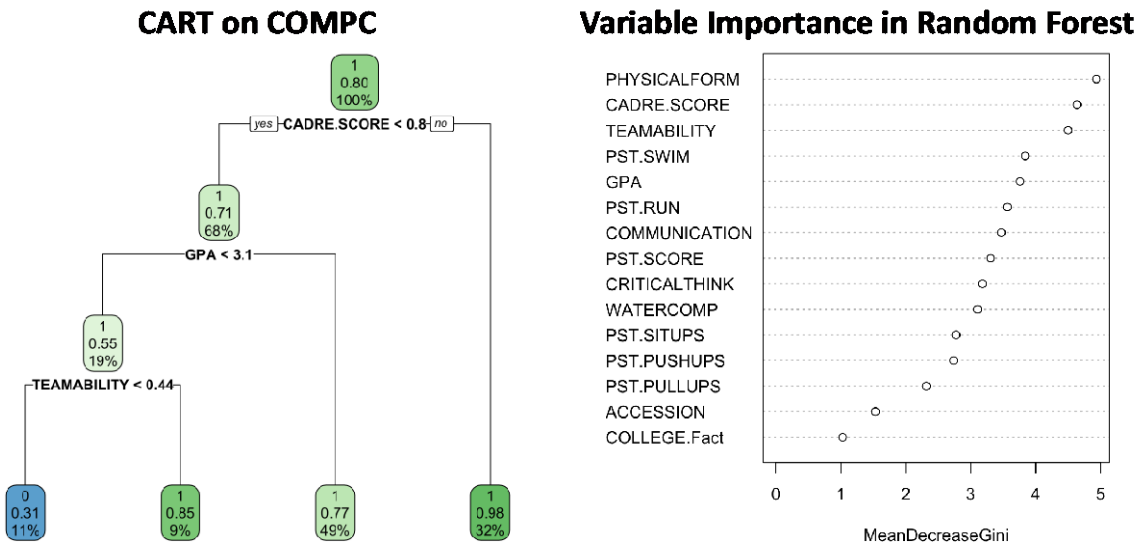

Figure 14. Models for Predicting Candidate Success in the First Phase of BUD/S

E. EMPIRICAL ANALYSIS

For illustration, we now examine the progress of five candidates through the stages and observe how the models assess them. These five candidates each offer an explanatory combination of outcome and probability of success in the first phase of BUD/S, as assessed by RF Model A. All five of the candidates were selected for SOAS, but not for BUD/S.

The cell colors used for tables in this section are meant to match the nodes of the CART models and indicate how a candidate's attributes affect the CART predictions. The color grey signifies that the CART model does not consider an attribute for that candidate. The colors give no indication as to the workings of the RF models, but ineffable structures are a feature of RF, so only the prediction probabilities are given.

Table 6 represents the assessment of the candidates with Model A. Notice the high predicted probability of success for candidate 34, who failed at BUD/S, and the low predicted probability for candidate 265, who was successful. While these models are not perfect, they are more accurate than suggested by these examples, which were chosen to demonstrate features of the models, not to stand as archetypal examples.

Table 6. Model A Empirical Analysis

Candidate ID	ACCESSION	Probability of Success CART	Probability of Success RF	Observed Outcome
1	USNA	0.746	0.976	Success
34	USNA	0.746	0.846	Failure (Failed at BUD/S)
86	USNA	0.746	0.740	Failure (Not Selected for BUD/S)
265	OCS	0.270	0.092	Success
315	OCS	0.270	0.026	Failure (Failed at BUD/S)

As all five candidates were selected for SOAS, Model B is not as informative but resolves as depicted in Table 7. As the table indicates, these five candidates came from United States Naval Academy (USNA) and Officer Candidate School (OCS) and therefore are selected by the CART model to pass into the leaf with the highest probability of success after one split. Had any of the candidates come from an alternate accession source, they would pass through more splits in the CART model shown in Chapter IV subsection B.

Table 7. Model B Empirical Analysis

Candidate ID	TEAMABILITY	PHYSICALFORM	Probability of Success CART	RF	Observed Outcome
1	0.157	1.401	0.873	0.928	Selected
34	-0.959	0.674	0.170	0.842	Selected
86	0.118	0.831	0.873	0.352	Not Selected
265	-0.394	-0.693	0.170	0.652	Selected
315	-0.023	-1.172	0.727	0.758	Selected

CART Model C is shown in Chapter IV subsection C, but the model's many splits are difficult to show in the table format. Table 8 shows the first two splits for the candidates. Assigning probabilities greater than 0.50 as predicted for selection, RF Model C accurately predicted the four candidates who were chosen for BUD/S. Meanwhile, CART Model C fails to predict selection for candidates 34 and 265. The candidates' low TEAMABILITY scores place them in the leaf with the lowest probability of selection without regard for other attributes. This demonstrates one weakness of CART models that RF resolves.

Table 8. Model C Empirical Analysis

Candidate ID	ACCESSION	Probability of Success CART	RF	Observed Outcome
1	USNA	0.918	0.999	Selected
34	USNA	0.918	0.998	Selected
86	USNA	0.918	0.966	Selected
265	OCS	0.918	0.924	Selected
315	OCS	0.918	0.958	Selected

Finally, Table 9 shows the three attributes considered by CART and the continued power of RF this time for Model D. Neither the CART nor the RF models were given selection to BUD/S as a predictor variable, so both consider candidate 86 as if he is eligible for success. Despite having a less competitive GPA, candidate 86 is assessed as having a high probability of success by both models especially compared to candidate 315. In addition, candidate 86's GPA is in the 15 percent percentile of candidates selected for BUD/S but appears to be competitive with, and in some cases, a stronger option than other candidates in that percentile regarding all attributes. This is one example that demonstrates

that information is likely missing from this data. That is not a criticism of this specific data, however. Even if the data were complete, there are aspects to human beings that are notoriously difficult to quantify. Moreover, the ultimate goal for the SEAL OCM is to select candidates who will make good SEALs, not just good BUD/S students. Therefore, the prudent assumption would be that the decision-makers are better informed than these models and generally make decisions with higher accuracy with respect to that objective.

Table 9. Model D Empirical Analysis

Candidate ID	CADRE.SCORE	GPA	TEAMABILITY	Probability of Success		Observed Outcome
				CART	RF	
1	1.277	3.88	0.157	0.980	0.992	Success
34	-0.028	3.20	-0.956	0.770	0.306	Failure
86	1.217	2.71	0.118	0.980	0.898	NA (Not Selected)
265	-0.476	3.42	-0.394	0.770	0.884	Success
315	-0.587	3.00	-0.023	0.313	0.196	Failure

V. CONCLUSION

This thesis aims to provide quantitative analysis of candidate training and selection data to inform officer selection criteria to SOAS and BUD/S, provide NSW with an appraisal as to the value of that data, and help shape the focus of data collection on candidates moving forward. We train CART and RF models to predict candidate success at making it to each stage of training from selection to SOAS, selection of BUD/S and finally completing the first phase of BUD/S.

We find that while the initial data collected from candidate applications can be used to reliably predict candidate selection to SOAS, those variables are less predictive of candidate selection to BUD/S and success at the first phase. Instead, the data elements collected at SOAS, particularly the subjective values assigned by the SOAS cadre, are the most important variables for predicting candidate selection to BUD/S and success in the first phase. This suggests that the methods currently employed to select candidates for SOAS and BUD/S are performing well at utilizing the data available at each stage, and there is little information remaining to inform later stages. If variables collected in earlier stages were predictive of success in later stages, that would suggest a suboptimal selection paradigm that does not exploit the information available.

Additionally, we explore techniques, such as MICE, for imputing values for missing data elements. In all cases, the models trained on complete data outperform those trained using the imputed data but generally agree on feature importance. This agreement builds our confidence in the model structure, but the dominance of models trained on complete cases highlights the importance of sound data collection and management practices.

The importance of the data collected at SOAS in predicting candidate selection to and success at BUD/S emphasizes the value of the SOAS program. Throughout the thesis process, we have worked with the SOAS cadre to inform data collection at SOAS. Based on the data features, including a high rate of missingness, we suggested a more flexible data collection plan that allows the SOAS cadre to modify the training plan without generating

new data elements or leaving values blank. The new data collection plan was instituted in time for the 2021 SOAS blocks.

Data collected under the new collection paradigm will vary from the data we analyze. This new data offers an area for future study. It would be interesting to compare models trained on the new data with the ones from this thesis. In addition, while officer and enlisted candidates differ in many ways, there is perhaps more commonality than difference between the two. For this reason, the features of importance identified in this thesis should be explored for applicability in assessing enlisted candidates.

APPENDIX A. DESCRIPTION OF VARIABLES

Variable	Description	Data Type	Factor Levels or Numeric Range
ACCESSION	Source the candidate came from	Categorical	INTERSERVICE ACADEMY: Commissioning in the Navy from U.S. Military Academy West Point, U.S. Air Force Academy or other service academy INTR_SRVC: Transferring from Army, Air Force, Marine Corps or Coast Guard LATERAL: Transfer from other community in Navy NROTC: Commissioning from a college NROTC Program OCS: Commissioning from OCS OCS AD: Prior enlisted service commissioning through OCS USNA: Commissioning from U.S. Naval Academy
COLLEGE.FACT	Type of college the candidate attended	Categorical	USNA / Other Military Schools / San Diego Schools / Other
PST.SCORE	PST score submitted by the candidate	Numeric	Min: 450, Max: 1127
PST.PULLUPS	Number of pullups performed	Numeric	Min: 10, Max: 40
PST.SITUPS	Number of sit-ups performed	Numeric	Min: 53, Max: 132
PST.PUSHUPS	Number of pushups performed	Numeric	Min: 50, Max: 146
PST.RUN	Number of seconds to run 1.5 miles	Numeric	Min: 424, Max: 654 (seconds)
PST.SWIM	Number of seconds to swim 500 yards	Numeric	Min: 345, Max: 711 (seconds)
GPA	College grade point average	Numeric	Min: 2.0, Max: 4.0 (4.0 scale)
CADRE.SCORE	Overall score given by SOAS cadre	Numeric	Min: -2.822, Max: 2.488
COMMUNICATION	Ability to communicate clearly and concisely	Numeric	Min: -3.741, Max: 2.779
PHYSICALFORM	Average form while conducting exercises	Numeric	Min: -2.634, Max: 3.872
CRITICALTHINK	Average performance during critical thinking drills	Numeric	Min: -4.081, Max: 2.746
TEAMABILITY	Willingness and ability to work as part of a team	Numeric	Min: -4.551, Max: 2.218
WATERCOMP	Performance and perceived comfort level in the water	Numeric	Min: -2.984, Max: 2.270
SOAS.SELECT	Selected for SOAS or not	Binary	{0,1}
BUDS.SELECT	Selected for BUD/S or not	Binary	{0,1}
BROWNSHIRT	Completed Hell Week or not	Binary	{0,1}

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. PST SCORING

Table 10. PST Scoring Rubric

Exercise	Time	Rest	Minimum Standard	Calculation
Swim 500 yards (breast or sidestroke)	Unlimited	10:00 minutes	12:30	+ Total seconds
Push-up	2:00 minutes	2:00 minutes	50	- Number Completed
Curl-up	2:00 minutes	2:00 minutes	50	- Number Completed
Pull-up	2:00 minutes	2:00 minutes	10	- 6 x Number Completed
Run 1.5 miles	Unlimited	Event over	10:30	+ Total seconds

Adapted from SEAL SWCC (2021)

$$PST\ Score = \begin{matrix} Total\ Seconds \\ (Run + Swim) \end{matrix} - \begin{pmatrix} Push\ Ups \\ + \\ Curl\ Ups \end{pmatrix} - 6 \times Pull\ Ups$$

Figure 15. Equation for Calculating PST Score

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Atlamazoglou, S (2021) How the Navy's 'Hell Week' reveals who has what it takes to be a SEAL. Accessed May 10, 2021, https://navyseals.com/5436/how-the-navys-hell-week-reveals-who-has-what-it-takes-to-be-a-seal/

Azur M, Stuart E, Frangakis C, Leaf P (2011) Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research* 20(1), U.S. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/.

Bermudez-Mendez (2020) Student success factors at Defense Language Institute Foreign Language Center. Master's thesis, Naval Postgraduate School, Monterey, CA, https://calhoun.nps.edu/handle/10945/64866

Buttrey, S E.; Whitaker, L R.. *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*. Wiley. Kindle Edition.

Feelders A (1999) Handling missing data in trees: surrogate splits or statistical imputation, Tilburg University, CentER for Economic Research, The Netherlands. https://webspace.science.uu.nl/~feeld101/pkdd99.pdf.

James G, Witten D, Hastie T, Tibshirani R (2017) *An Introduction to Statistical Learning with Applications in R* (Springer, New York, NY).

King G, Honaker J, Joseph A, Scheve, K (2001) Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 95(1), https://gking.harvard.edu/files/gking/files/evil.pdf.

Navy Personnel Command (2021) Navy SEAL Officer Community Manager. Accessed April 8, 2021, https://www.mynavyhr.navy.mil/Career-Management/Community-Management/Officer/Active-OCM/Unrestricted-Line/Special-Warfare-OCM/

Nowicki (2017) United States Marine Corps Basic Reconnaissance Course: Predictors of Success. Master's thesis, Naval Postgraduate School, Monterey, CA, https://calhoun.nps.edu/handle/10945/53025

Posture Statement of Rear Admiral H. W. Howard IIII, Usn Commander, Naval Special Warfare Command Before The 117th Congress Senate Armed Service Committee. https://www.armed-services.senate.gov/imo/media/doc/20210428_NSWC%20Posture%20Statement_SASC-ETC-FINAL.pdf

RDocumentation (2021) scale: Scaling and Centering of Matrix-like Objects. June 24, 2021, https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale.

Rubin D (1996) Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91:434, 473–489, https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476908?needAccess=true

SEAL SWCC (2021) SEAL PST Calculator. Accessed April 8, 2021, https://apps.sealswcc.com

U.S. Navy (2021) SEAL OCM, Sequence of Events For Interested SEAL Officer Applicant (OCS) Accessed April 8, 2021, https://www.mynavyhr.navy.mil/Portals/55/Career/OCM/Active/Unrestricted/NSW/OCS%20Package%20Flow%20Chart%20(1).pdf?ver=_yoJtdy1rydLm_JFWasCLQ%3d%3d

Van Buuren S (2018) *Flexible Imputation of Missing Data Second Edition*. Chapman & Hall/CRC*,* https://stefvanbuuren.name/fimd/.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California