NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

ANALYSIS OF MARINE CORPS RECRUITS IN THE DELAYED ENTRY PROGRAM

by

Johnathan L. Thornton

September 2023

Thesis Advisor: Ruriko Yoshida
Second Reader: Robert A. Koyak

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2023	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE ANALYSIS OF MARINE CORPS RECRUITS IN THE DELAYED ENTRY PROGRAM	**5. FUNDING NUMBERS**
6. AUTHOR(S) Johnathan L. Thornton	
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)

This thesis aims to assist the Marine Corps in solving the problems associated with achieving the objectives outlined in Talent Management 2030, a supplement to Force Design 2030. We analyzed 35 variables derived and originated from the Marine Corps Recruiting Command's applicant information system spanning fiscal year (FY) 15–22. This study utilized random forest models to identify the essential attributes contributing to the accurate classification of Marine Corps recruits that ship to basic training. We cross-validated nine separate random forest models against each FY testing data. The analysis revealed that our added variables significantly affected the model's accuracy. Our added variables included the days in the Marine Corps delayed entry program, days between the date of enlistment and the initial strength test, and the applicant weight difference between the initial strength test day and ship day. Although our model demonstrated high specificity, it tended to make false positive errors when classifying applicants as shippers. When cross-analyzed, most models proved robust against each FY test data, with the exception of FY 22. Future research could explore alternative analysis methods, such as time series or survival analysis, to refine the model's accuracy. Overall, our findings provide valuable insights into our added variables that aided in accurately classifying a Marine Corps recruit who ships to basic training and highlight areas for further investigation.

14. SUBJECT TERMS Marine Corps Recruiting Command, MCRC, Delayed Entry Program, DEP, recruiting, Talent Management 2030, Force Design 2030, random forest, fiscal year, FY	**15. NUMBER OF PAGES** 83
	16. PRICE CODE

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

ANALYSIS OF MARINE CORPS RECRUITS IN THE DELAYED ENTRY PROGRAM

Johnathan L. Thornton
Lieutenant, United States Navy
BS, United States Naval Academy, 2014

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2023

Approved by: Ruriko Yoshida
Advisor

Robert A. Koyak
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This thesis aims to assist the Marine Corps in solving the problems associated with achieving the objectives outlined in Talent Management 2030, a supplement to Force Design 2030. We analyzed 35 variables derived and originated from the Marine Corps Recruiting Command's applicant information system spanning fiscal year (FY) 15–22. This study utilized random forest models to identify the essential attributes contributing to the accurate classification of Marine Corps recruits that ship to basic training. We cross-validated nine separate random forest models against each FY testing data. The analysis revealed that our added variables significantly affected the model's accuracy. Our added variables included the days in the Marine Corps delayed entry program, days between the date of enlistment and the initial strength test, and the applicant weight difference between the initial strength test day and ship day. Although our model demonstrated high specificity, it tended to make false positive errors when classifying applicants as shippers. When cross-analyzed, most models proved robust against each FY test data, with the exception of FY 22. Future research could explore alternative analysis methods, such as time series or survival analysis, to refine the model's accuracy. Overall, our findings provide valuable insights into our added variables that aided in accurately classifying a Marine Corps recruit who ships to basic training and highlight areas for further investigation.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AFQT	Armed Forces Qualification Test
ASVAB	Armed Services Vocational Aptitude Battery
AVF	All-Volunteer Force
DEP	Delayed Entry Program
DLI	Defense Language Institute
DOD	Department of Defense
EAS	End of Active Service
FY	Fiscal Year
GPA	Grade Point Average
IST	Initial Strength Test
MCRC	Marine Corps Recruiting Command
MCRD	Marine Corps Recruit Depot
MEPS	Military Entrance Processing Station
MOS	Military Operational Specialty
NPS	Naval Postgraduate School
NROTC	Naval Reserve Officer Training Corps
PDE	Person-Event Data Environment
PFA	Physical Fitness Assessment
PROM-WED	Planned Resource Optimization Model with Experimental Design

ROC	Receiver Operating Characteristic
SAT	Scholastic Aptitude Test
TAPAS	Tailored Adaptive Personality Assessment System
USN	U.S. Navy

Executive Summary

Talent Management 2030 prioritizes the use of machine learning and big data to optimize the skills of individual marines as a portion of its efforts to modernize the Corps. The use of machine learning to assist decision-makers in recruiting and retention efforts is a subsection of the Marine Corps modernization described in Force Design 2030. There are few studies conducted to predict the success of a Marine Corps recruit shipping to basic training.

Of the few studies over our research topic Vanegas et al. (2022) used Tailored Adaptive Personality Assessment System and Armed Services Vocational Aptitude Battery scores along with machine learning methods to make predictions on non-End of Active Service (EAS) attrition among Marine Corps recruits. We conducted an in-depth analysis to determine the attributes that contribute to the accurate classification of a Marine Corps Recruit that ships to basic training from the delayed entry program. Vanegas et al. (2022) found that of the machine learning methods used, the random forest classifier was the most accurate for non-EAS attrition. With data provided by the Marine Corps Recruiting Command's Information Support System, we trained a random forest model to classify a shipper for eight fiscal years (FY) spanning FY 15-22. We added another model trained from a random sample of the aggregate data for a total of nine models. We analyzed each model's accuracy, cross-analyzing against each FY testing data. We found several variables that impacted the likelihood of a recruit shipping to basic training. The most significant variables were:

1. Difference in weight between the Military Entrance Processing Station (MEPS) contract date and ship date or final recorded weight.
2. Number of days between the initial strength test date and the (MEPS) contract date.
3. Number of days an applicant spent in the delayed entry program.

We created these variables by merging information from the original data. Each random forest model demonstrated high specificity, meaning that an applicant classified not to ship was very likely not to ship. This could aid decision-makers to either apply more resources to applicants of higher risk or use resources elsewhere. Conversely, the models were biased towards false positive errors when classifying applicants as shippers. As such, future research could explore alternative analysis methods, such as a time series or survival analysis, to

further refine the model's accuracy. Overall, this study provides valuable insights into the key attributes that can aid in accurately classifying Marine Corps applicants and highlights several areas for further investigation.

REFERENCES

U.S. Marine Corps (2020) Force design 2030. Force Design 2030, Washington, DC, https://www.marines.mil/Force-Design-2030/.

U.S. Marine Corps (2021) Talent management 2030. Talent Management 2030, Washington, DC, https://www.marines.mil/Talent-Management-2030/.

Vanegas JMA, Wine W, Drasgow F (2022) Predictions of attrition among U.S. Marine Corps: Comparison of four predictive methods. *Military Psychology* 34(2):147–166, https://doi.org/10.1080/08995605.2021.1978754.

Acknowledgments

I am offering a huge thank you for the patience and support offered to me by the entire Operations Research Department.

Specifically, I would like to express my gratitude to Professor Yoshida for her guidance, patience, and gentle nudges to complete this.

Also, to my second reader, Professor Koyak, for a meticulous review.

Soli Deo Gloria.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 Force Design 2030

The U.S. Marine Corps is modernizing, as evidenced by its stated goals in Force Design 2030 (U.S. Marine Corps 2020). This is a strategic initiative by the Marine Corps to restructure and reshape its combat power to address future near-peer threats and challenges. The redesign process is informed by threats, based on specific concepts, and is accountable to a campaign of learning (U.S. Marine Corps 2020). The primary goal of Force Design 2030 is to transform the Marine Corps' existing model to better contend with the evolving nature of warfare. Some significant changes proposed under this initiative include the Marine Corps transitioning to be a "stand-in force" comprised of "small but lethal forces" designed to operate across various stages of competition (Cancian 2022).

Relation to this Thesis

We introduce the Marine Corps Talent Management 2030 initiative here but further expand upon it in the next chapter. Its aim is to match the skills of Marines with the changing demands of the Service (U.S. Marine Corps 2021). This is achieved through the use of modern computing techniques like machine learning in order to optimize the skills of each Marine and the Corps as a whole (U.S. Marine Corps 2021). This complements the Marine Corps Force Design 2030 initiative, which focuses on restructuring combat power for future challenges (Cancian 2022). Our thesis examines the portion of a Marine's career that starts when they enter the Delayed Entry Program (DEP) as a recruit.

1.2 Problem Statement

This thesis aims to identify the key factors that lead to a successful shipment of Marine Corps recruits to basic training after they enter the DEP. To achieve this, we have employed machine learning techniques via a random forest classification model. Our findings will assist the Marine Corps Recruiting Command (MCRC) to advance towards their goals outlined

in Talent Management 2030, which supplements broader objectives in Force Design 2030 (U.S. Marine Corps 2020, 2021). The Marine Corps is currently exploring several avenues using machine learning and big data to identify potential attrition risks associated with each DEP applicant (Eckstein 2022; Athey 2022). Our thesis aims to enhance these methods by incorporating a random forest classifier, a well-regarded approach that utilizes the inputs already gathered by MCRC as part of its standard recruiting data collecting.

1.3 Structure of Thesis

This paper is composed of five chapters. Chapter 2 delves into the history of recruiting in the United States, exploring the reasons behind the need for a volunteer force and how machine learning has contributed to the optimization goals of Talent Management 2030. Additionally, Chapter 2 demonstrates how our thesis aligns with Talent Management 2030 objectives. In Chapter 3, we focus on the data used for our research, how we preprocessed it and how we formulated our model. Chapter 4 presents the study's outcomes and graphically displays the importance of each variable. Lastly, Chapter 5 summarizes our results and offers suggestions for future research.

CHAPTER 2:
Background and Literature Review

In Section 2.1, we discuss the history of the U.S. Military's All-Volunteer Force (AVF) and its importance. In Section 2.2, we look at challenges to recruiting identified by previous work. In Section 2.3, we observe methods observed in previous work to overcome the challenges facing Marine Corps recruiting. Finally, Section 2.4 introduces the importance of the random forest classification method.

2.1 Motivation for an AVF

The U.S. military has a history of both volunteer service and conscription. Volunteers essentially made up the first Continental Army during the American Revolutionary War (Royster 1979). During the American Revolution, volunteering to fight in the war was demanded at a spiritual level, and it served as a unifying force for the future United States (Hutson 1998). Following the American Revolution, conscription in the U.S. was implemented during times of significant conflict, such as the Civil War, World War I, and World War II (Royster 1979). Post-World War II, The Selective Service Act of 1948 required all men aged 18-26 to register for potential conscription, although the last draft call was in 1972 during the Vietnam War (Royster 1979).

2.1.1 Exploring Conscription Impacts

There is some debate on the effects of conscription within the U.S. to combat the recruitment challenges noted in this chapter. The impacts of conscription are explored below to motivate this thesis's importance. Based on the U.S. Constitution, a primary requirement for the U.S. to wage war is the support of the American people via Congress (U.S. Constitution 1776). The last time that Congress had waged war was World War II, but the U.S. has been in many conflicts since. This section assumes that obtaining the backing of the American public is

crucial for participating in a foreign conflict, as it serves as a preventative measure against a repeat of the resistance to conscription witnessed during and after the Vietnam War (Foley 2003). The rationale behind this opposition will be further scrutinized in the upcoming sections.

2.1.2 Self-Interest or Inconvenience?

According to Bergan, the self-interest of conscripted individuals is the primary factor that drives support by the American people for the conflict in which they are conscripted (Bergan 2009). Heidi Urben and Peter Feaver assert that drafting individuals to combat the lack of volunteers does not inspire future volunteers Urben and Feaver (2023), The evidence presented by Bergan reinforces this assertion (Urben and Feaver 2023). Since the end of the Vietnam Conflict, the U.S. military has relied on an AVF. While some experts (NPR 2023) agree with Bergan that a draft or mobilization of reserves could lead to greater public attention and constraint on foreign policy decisions, Blankshain's research suggests that the relationship is more complicated (Blankshain et al. 2022).

Blankshain argues that although conscription could potentially decrease the likelihood of the U.S. joining a conflict, there is no conclusive evidence linking personal cost expectations with reluctance to support military action (Blankshain et al. 2022). Additionally, Blankshain proposes that the disruptive nature of conscription, independent of the conflict, is the primary factor to dampen support from the American people, not Bergan's self-interest (Blankshain et al. 2022). Blankshain proposes that if conscription were implemented as a continuous policy, its impact could be reduced over time, similar to the decreased effects of Guard and Reserve mobilization since the 1960s (Blankshain et al. 2022). This position is not unique; Figure 2.1 shows that many countries support policies of varied levels of conscription. Some countries implement mandatory military service, but only a small percentage of eligible individuals are actually drafted (Buchholz 2023). Exceptions and bribes are common. Less than 30 countries require entire age groups to serve, with a higher incidence in the Middle East and Asia. Normalized conscription policies exist in Taiwan, South Korea, and Israel despite differences in size, location, and country origin and history.

2.1.3 Forward with the AVF

While there are varying perspectives on the factors behind negative attitudes towards a conscription policy, Bergan and Blankshain concur that it can harm the level of support that the American people have towards it (Blankshain et al. 2022; Bergan 2009). As a result, this study strives to advance the long-standing U.S. policy of an AVF and contribute to recruiting committed volunteers willing to serve in the U.S.'s current and future wars and engagements.

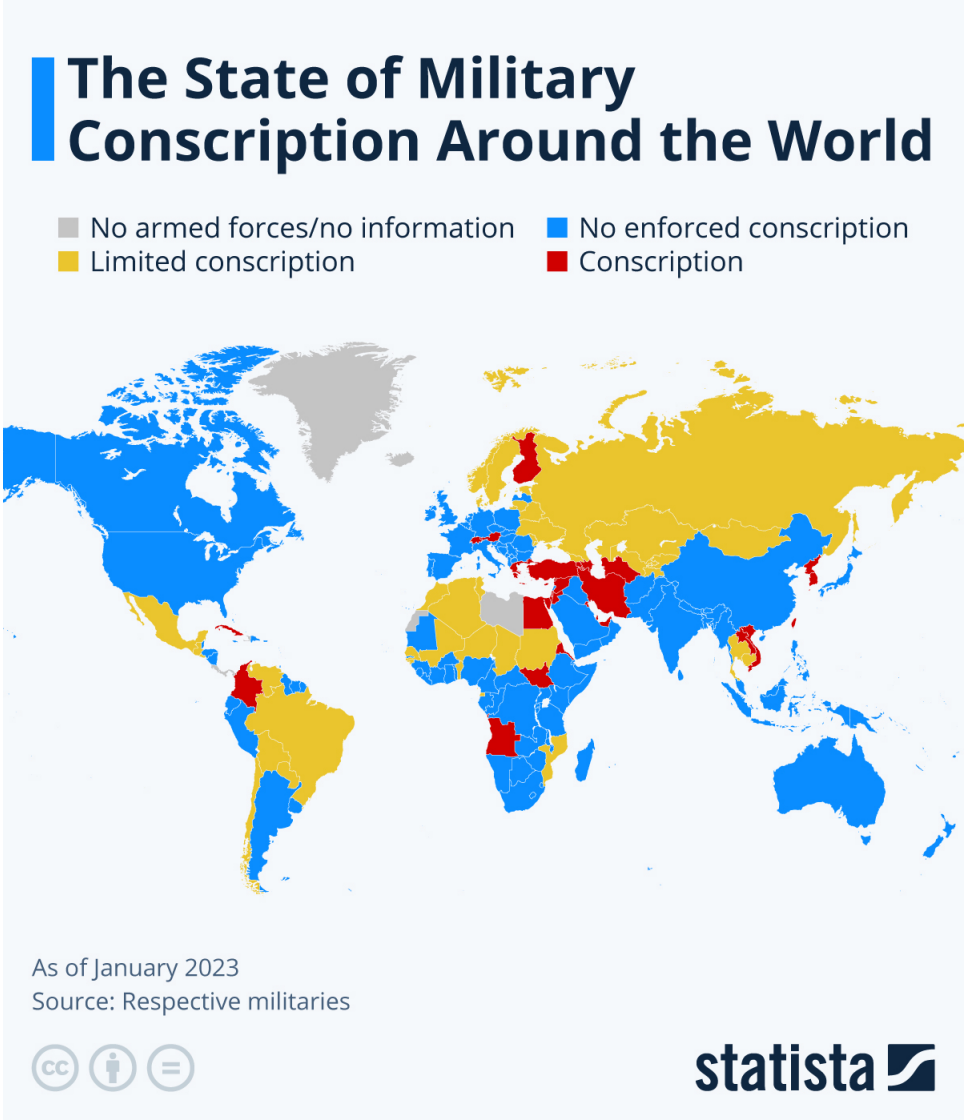


Figure 2.1. World map and conscription policies. Source: Buchholz (2023).

2.2 Recruiting Challenges

Maintaining an AVF for the U.S. military requires identifying highly qualified individuals. However, recent years have presented recruitment challenges, with 2021 marking one of the most difficult years for recruitment since 1973 (Pyne 2023). The Army failed to meet its 2022 recruitment goal by 15,000 soldiers, and the Army, Air Force, and Navy are all expected to miss their 2023 goals (Pyne 2023). This shortage is blamed on domestic issues, including a competitive job market, limited in-person recruiting during the pandemic, and a young adult population lacking the necessary knowledge, interest, and qualifications for military service (Pyne 2023). Although the Marine Corps is projected to meet its recruiting targets for 2023 (DOD 2023), its leaders acknowledge that the recruitment process has become more challenging (Mongilio 2023).

2.2.1 Traditional Recruiting Practices

Paredes (2020) conducted research in a thesis titled, Making the Marine Corps Recruiting Process More Efficient, to explore innovative ways for the Marine Corps to enhance its recruitment process. Paredes researched enlisted recruitment and explained the eight steps involved with the Marine Corps method of recruiting. Systematic recruiting is a method recruiters use to efficiently locate and process potential applicants for enlistment, as illustrated in Figure 2.2.

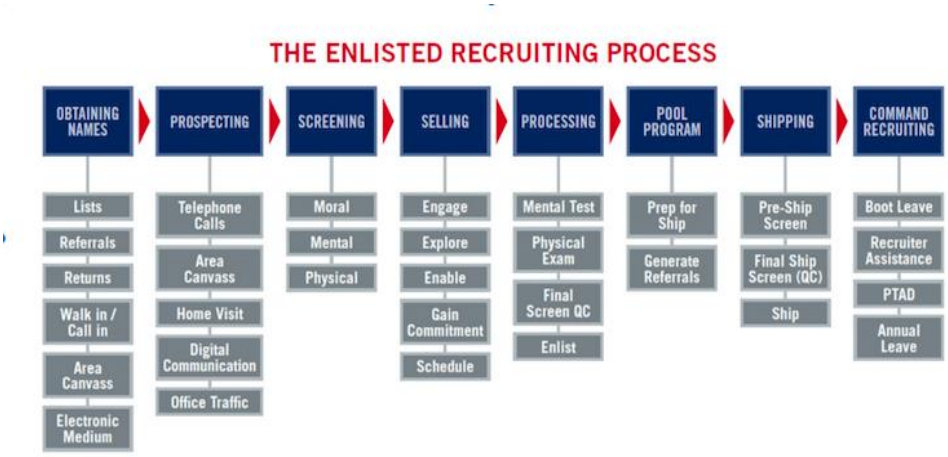


Figure 2.2. Systematic Recruiting. Source: Paredes (2020).

Referrals

Recruiters gather many potential candidates from high schools, referrals, walk-ins, and the Armed Services Vocational Aptitude Battery (ASVAB). They then contact each applicant through phone calls, in-person visits, digital communication, or inviting them to their office (Paredes 2020). Referrals play a crucial role in securing contracts for the Marine Corps. From fiscal years 2018 to 2019, 8 percent of all contracts were obtained through referrals (Paredes 2020). These referrals originate from different sources, including command recruiters, new enlistees, local Marine Corps reserve units, community contacts, and former Marines. Referrals are considered efficient and reliable because the person making the referral knows the applicant's interest in the Marine Corps, and they are likely to meet the recruitment requirements (Paredes 2020). Referrals also save time compared to other potential uninterested leads that may require more screening (Paredes 2020).

Area Canvassing

Another effective prospecting activity stated by Paredes is area canvassing, which accounts for 25 percent of all contracts during the given time frame (Paredes 2020). Although inefficient, area canvassing takes up 52 percent of a recruiter's working hours (Paredes 2020). This activity capitalizes on the Marine Corps' reputation for professionalism and appearance to be successful (Paredes 2020). Recruiters spend significant ime n various locations, such as schools, malls, stores, and athletic events, to increase their chances of interacting with potential applicants. Area canvassing has proven to be an effective prospecting method, compared to walk-ins and telephone calls, which are inefficient forms of prospecting (Paredes 2020).

Walk-Ins and Telephone Calls

Walk-ins accounted for only two percent of contracts written in fiscal years 2018 and 2019, despite recruiter efforts to enhance visibility and accessibility. Although walk-ins are ineffective, they are a passive recruiting practice that do not require much investment of time (Paredes 2020).

Paredes states that recruiters spend 34 percent of total recruiting time on telephone calls, resulting in 15 and 14 percent of total contracts in 2018 and 2019, respectively. Typically, it takes around 225 phone calls to obtain a single contract, indicating that this approach is highly ineffective (Paredes 2020). The recruiting practices listed above are outdated, and Marine Corps leaders agree that they must be improved (Athey 2022).

2.2.2 Social Media

Paredes' research resulted in a recommendation to the Marine Corps to put more effort into recruiting over social media platforms (Paredes 2020). Paredes highlighted that the existing Marine Corps' social media presence resulted in 12 percent of total enlistments generated with two percent investment in recruiter time (Paredes 2020). Paredes' findings are overall valuable for the motivation in developing this thesis as we look to add other methods to enhance recruiting efficiency. We strive to add to the efforts to improve the outdated processes listed above. The following section will discuss how data analytics can enhance the recruitment process for the nation's AVF, making it more efficient.

2.3 Overcoming Recruiting Challenges with Data

Buttrey et al. (2018) utilized various data-driven approaches beyond social media to enhance recruitment and effectively target promising candidates. They focused on optimizing the geographic placement of recruiters based on the number of potential recruits available. Although Buttrey et al. (2018) focus on the Navy, the Marine Corps already invests in data analytics to streamline their recruitment process (Barnett 2022) and uses data-driven strategies to understand trends and effectively target potential recruits (Athey 2022).

2.3.1 Prescriptive Modeling

Prescriptive modeling is a type of analytical technique that involves creating models that provide recommendations or prescriptions for making optimal decisions based on a set of inputs, constraints, and objectives. The goal of prescriptive modeling is to help individuals or organizations make informed choices that lead to the best possible outcomes.

In agreement with Paredes, Buttrey et al. (2018) argue that maximizing recruitment efficiency in the U.S. military can lead to significant cost savings. The financial benefits could be directed toward other pressing military needs, such as investing in new equipment or personnel (Buttrey et al. 2018).

Optimization

Buttrey et al. (2018) explain that assigning recruiters based on both the propensity and eligibility of potential recruits in a particular zip code could result in a saving measure for Navy recruiters. The proposal by Buttrey et al. (2018) aligns with the Marine Corps ongoing efforts to leverage modern computing resources for resolving a mixed-integer linear program to find the optimal number of recruiting stations filled with an optimal number of recruiters. The result would minimize the number of recruiters, leading to significant savings in time and resources (Buttrey et al. 2018).

Planned Resource Optimization Model with Experimental Design (PROM-WED)

Allison Hogarth conducted research in a thesis titled, Improving Navy Recruiting with the New Planned Resource Optimization Model with Experimental Design, which combines the Buttrey et al. (2018) cost-saving proposal using optimization tools with data farming (Hogarth 2017). Hogarth introduced Navy workforce analysts to a powerful tool called the PROM-WED, which enables scenario-based explorations and tradeoff analyses, providing valuable insights that help optimize recruitment resources. With this innovative capability, Navy decision-makers can confidently make informed choices that yield robust results (Hogarth 2017). PROM-WED uses Near Orthogonal Latin Hypercube designs for effective space-filling and design flexibility. Users can set up designs for model inputs spread across multiple fiscal years based on decision factors under the Navy's control (e.g., number of production recruiters, advertising budget) and uncontrollable factors (unemployment rate, number of qualified military available) (Hogarth 2017). PROM-WED helps decision-makers evaluate choices, apply restrictions, and generate data for further analysis in statistical software. PROM-WED is a quick and efficient way for analysts to gather experimental data across a broader range of input variables than ever before. At the time of Hogarth's thesis, its capabilities were in the evaluation phase (Hogarth 2017). PROM-WED helps Navy analysts

allocate resources efficiently for recruiting by exploring different scenarios and conducting trade-off analyses (Hogarth 2017). It is a valuable tool for decision-making and workforce planning that the Marine Corps could also adopt as part of its modernization efforts.

2.3.2 Predictive Modeling

Predictive modeling is a data science and statistics process that involves creating a model to predict future outcomes based on historical data. It is a type of machine learning that aims to make predictions or forecasts about unknown future events by analyzing patterns and relationships within existing data. Buttrey et al. (2018) and (Hogarth 2017) successfully established recruitment resource allocation utilizing prescriptive modeling in the articles we reviewed. The following thesis by Juan Vanegas discusses using predictive modeling to identify potential attrition among Marine Recruits (Vanegas et al. 2022).

The Marine Corps' Talent Management 2030

Vanegas' research is motivated by a Marine Corps publication titled Talent Management 2030 (U.S. Marine Corps 2021). Talent Management 2030 supports the Marine Corps modernization efforts outlined in Marine Corps Force Design 2030 (2020). The Marine Corps has set the goal of modernizing its recruitment and retention strategies through Talent Management 2030. As part of this initiative, the Marine Corps seeks to use predictive modeling to identify those who may not successfully complete their first enlistment, known as non-End of Active Service (EAS) attrition, (U.S. Marine Corps 2021). Around 20 percent of recruits are classified as a non-EAS attrite, resulting in losses of millions of dollars annually to the Corps (U.S. Marine Corps 2021). The updated 2023 Talent Management 2030 states that Tailored Adaptive Personality Assessment System (TAPAS) gathers information from Marine applicants in order to predict their chances of success at different stages in their respective careers in hopes of lowering non-EAS attrition (U.S. Marine Corps 2023).

Predictions of Non-EAS Attrition Among Marine Corps Recruits in the DEP

Vanegas's research in a thesis titled, "Predictions of Non-EAS Attrition Among Marine Corps Recruits", aimed to investigate Marine recruit attrition, primarily using the TAPAS facet scores while considering ASVAB scores and demographic information. Predicting non-EAS attrition is crucial for cost-saving purposes, particularly for those who have already

enlisted in the Marine Corps but have yet to commence basic training known as the DEP. The DEP plays a vital role in the recruitment process (Vanegas et al. 2022) because the Marine Corps has invested time into the recruitment. The study aimed to compare different statistical models and understand the psychological drivers of early attrition from the military using machine learning classification techniques (Vanegas et al. 2022).

The analysis included records from 39,043 recruits, with missing data imputed using k-nearest neighbor imputation (Vanegas et al. 2022). The TAPAS assessed facets such as achievement, adjustment, commitment to serve, courage, dominance, even-temperedness, ingenuity, optimism, physical condition, responsibility, selflessness, sociability, team orientation, tolerance, and virtue (Vanegas et al. 2022). Vanegas conducted a complete sample comparison of logistic regression models with classification trees and random forests for predicting attrition from the DEP. The goal was to find the most accurate predictive tool (Vanegas et al. 2022). The machine learning models outperformed logistic regression in predicting voluntary attrition in a stratified 50 percent attrition sample. In his research, Vanegas discovered that random forests were the most effective method for classifying cases of attrition. However, these models were less interpretable compared to other methods. Nonetheless, they still managed to minimize false negatives and accurately identify all cases of attrition (Vanegas et al. 2022). The study also identified personality predictors, such as a higher Armed Forces Qualification Test (AFQT) score indicating a higher likelihood of attrition (Vanegas et al. 2022).

This research is significant because it demonstrates that machine-learning models can be more effective than traditional classification methods in predicting attrition among military recruits. Vanegas's research is in line with the goals of the Marine Corps Talent Management 2030 initiative. Based on the results of Vanegas's analysis, this thesis builds on the use of machine learning to predict Marine recruit attrition in the DEP using the random forest classification method to predict whether an applicant ships to basic training after enlisting in the Marine Corps.

Identifying Factors of Attrition to Effect Retention of Soldiers attending the Defense Language Institute (DLI)

In a thesis entitled "Defense Language Institute Survival Analysis of Army Enlisted Defense Language Institute Graduate Attrition Factors," Philip J. Lukanich researched the factors that cause attrition. Lukanich utilized Kaplan-Meier, Cox Proportional Hazard, and random survival forest models to determine which characteristics of a soldier contributed to attrition from DLI (Lukanich 2023). The Marine Corps aims to reduce non-EAS attrition as part of Talent Management 2030, and identifying these contributing features can help achieve this goal. According to Lukanich, it costs approximately 320 thousand dollars for each of the roughly three thousand students enrolled annually at DLI (Lukanich 2023). Identifying factors that contribute to attrition saved the Army financial resources. This analysis used the Person-Event Data Environment (PDE) to examine enlisted Army soldiers who joined enlisted between the beginning of 2010 and the end of 2012 (Lukanich 2023). The PDE is a centralized database that holds confidential manpower, training, financial, health, and medical information of U.S. Army personnel (Vie et al. 2015). Features of the data that significantly contributed to attrition shown in Figure 2.3 are age, ethnicity, marital status, language difficulty, education level, and ASVAB (labeled AFQT) percentile (Lukanich 2023). Slightly more than half of the observed service members were already discharged from the Army, with just under 70 percent leaving between four and six years of service (Lukanich 2023).

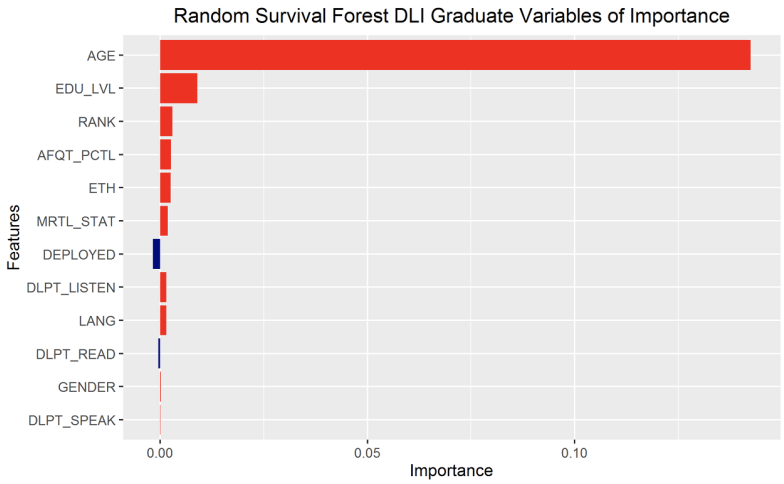


Figure 2.3. Variables contributing to attrition. Source: Lukanich (2023).

Future decision-makers can utilize the results of Lukanich's research to adjust and develop policies focused on retention. The results of our thesis focus on the features that make up attrition among Marine recruits in the DEP, in line with efforts of Talent Management 2030.

Identifying Factors of Attrition to Effect Retention Within Naval Reserve Officer Training Corps (NROTC)

Zachary H. Swenson researched factors leading to attrition within NROTC midshipmen in a thesis entitled "Predictive Statistical Modeling of Naval Reserve Officers Training Corps Attrition." This study looks to answer why midshipmen drop out of NROTC by analyzing demographic, academic, and performance factors from 2013-2020. Predictive models were used to classify NROTC commissions and drop requests (Swenson 2020). The Naval Education and Training Command provided the necessary data for the research. The data encompassed demographic information like race, sex, ethnicity, scholarship status, and Navy or Marine option, coupled with academic information such as Grade Point Average (GPA), Scholastic Aptitude Test (SAT) scores, and choice of major, along with performance-based measures such as aptitude score, Physical Fitness Assessment (PFA) score, and Nuclear Submarine Officer eligibility (Swenson 2020). The data spanned 2013-2020, observing 21,496 NROTC midshipmen (Swenson 2020).

Based on the model performance and inference results, random forest models resulted in the most efficient classifiers for both NROTC commissions and drop requests in the dataset (Swenson 2020). The scholarship status, aptitude score, GPA, and PFA score were the key predictors for the RF classification models. A finding from Swenson (2020) was that performance-based data were superior predictors for both commissions and drop requests compared to demographic data.

In concurrence with Vanegas, Swenson concluded that the classification process of random forests, composed of numerous decision trees, is difficult to interpret or graphically display (Swenson 2020). Despite their hard-to-interpret nature, random forests aid in accurately classifying target variables (commission and drop requests) by determining the significance of variables involved (Swenson 2020). Swenson's research adds to the growing efforts to maximize the usage of modern machine learning analytics to minimize attrition in its recruits by the Marine Corps (Eckstein 2022). This thesis builds on the use of machine learning to predict Marine recruit attrition in the DEP using the random forest classification method to predict whether an applicant ships to basic training after enlisting in the Marine Corps.

2.4 The Random Forest Classification Algorithm

In the field of data analysis, the random forest algorithm has become a go-to tool. Leo Breiman developed this ensemble machine-learning method in 2001 (Boulesteix et al. 2012). It is particularly useful in bioinformatics and computational biology, where there are more variables than observations. The algorithm manages complex interaction structures and highly correlated variables well, providing reliable measures of variable importance (Boulesteix et al. 2012). Even when faced with a large number of predictor variables and complicated interactions, this algorithm has proved to be highly effective (Hooker and Mentch 2021). It has been applied to both regression and classification problems, showing promise in various applications (Rose and Hassen 2019). This study will use the random forest classifier to identify the most critical features in predicting whether an applicant in the Marine Corps DEP will ship off to basic training.

CHAPTER 3:
Methodology

This thesis aims to utilize the random forest classifier to determine he key attributes that aid in classifying an applicant for the Marine Corps who has successfully finished the processing phase of the Marine Corps' systematic recruiting process, as depicted in Figure 2.2. This chapter details our methodology. First, we review the dataset and explain our pre-processing and cleaning process. After that, we discuss how we developed our models.

Overall, our research was executed in the following steps:

1. Acquired Marine applicant data.
2. Pre-processed the data to allow for analysis.
3. Added categorical response variables.
4. Split the data into training/testing sets and fit random forest classifiers on training datasets.
5. Cross analyzed models.

Steps 1-4 are discussed in this chapter, while the final step is discussed in Chapter 4.

3.1 Acquiring the Data

The data used in this study was provided by the MCRC through their Information Support System. The data included information on Marine Corps applicants, from the processing stage to the shipping stage, as illustrated in Figure 2.2. The data covered the period from Fiscal Year (FY) 2015 to FY 2022. Although the data was compiled in November 2022, it was not provided until February 2023 due to administrative approval delays beyond the control of MCRC. The data was supplied in eight separate .csv files representing FYs 15-22 with a total of 437,989 rows of observations along with 70 named columns with 13 additional error columns discussed further in this section. In the given data set, every observation row pertained to a distinct applicant to the Marine Corps.

To provide more information about the variables and their respective data types, a data dictionary consisting of the 70 named columns can be found in Appendix A.

3.2 Data Pre-Processing

The Pandas package was used to import each .csv file into Python. However, we became aware that about 2 percent of the total observations in FY 17, FY 18, FY 19, FY 20, and FY 21 had mixed data columns.

3.2.1 Mixed Datatypes

Upon closer examination, we determined that the "IST.Notes" column, which was of character datatype, had been parsed over multiple columns, resulting in additional rows, mixed data columns, and sparsely populated columns.

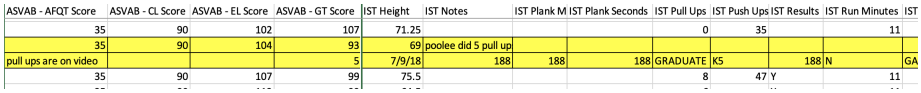
Figure C.32 shows the FY 18 dataset with the most egregious frequency of error. To rectify this error, we created a Python function to check that each column matched with the correct column data types identified in Appendix A. To retain the maximum data, each FY dataset was processed using a for-loop function that iterated over each row. During each iteration, the keywords were matched with their respective column names in order to determine their data type.

IST Notes	IST Plank M	IST Plank Seconds	IST Pull Ups	IST Push Ups	IST Results	IST Run Minutes	IST Run Seconds	Ship To	District	Region	RS Name	RSS Name	Ship Date	Weight At Co	Weig
5			10		Y	12	22	San Diego	9MCD	WRR	DES MOINES	CEDAR FALLS	11/27/17	176	
5			2	35		13	21	Parris Island	4MCD	ERR	COLUMBUS	SOUTH COLU	7/8/19	149	
5	MAKING AMAZING	SHOULD E/	SOULD BE AN AM.		14	Y	11	45	San Diego	9MCD	WRR	ST. LOUIS	FAIRVIEW HE		

Figure 3.1. Mixed data shown in FY 18 dataset. Source: MCRC data. Figure generated using Microsoft Excel.

To reiterate for further clarification, we carried out the following steps to fix the issue of extra rows caused by the Initial Strength Test (IST) Column parsing error, as seen in Figure 3.2. According to Appendix A, any column containing the word "Score" should have a numerical value. To ensure this, we checked each row to confirm that all column names with "SCORE" had corresponding numerical values ("ASVAB...AFQT.Score," "ASVAB...CL.Score," "ASVAB...EL.Score," and "ASVAB...GT.Score."

During this process, the loop kept track of the row number where mixed data was found and the preceding row. In case of an incorrect data type, the value was substituted with an "NA" value and the row, as well as the previous row of items were removed.



ASVAB - AFQT Score	ASVAB - CL Score	ASVAB - EL Score	ASVAB - GT Score	IST Height	IST Notes	IST Plank M	IST Plank Seconds	IST Pull Ups	IST Push Ups	IST Results	IST Run Minutes	IST
35	90	102	107	71.25				0	35		11	
35	90	104	93	69	poolee did 5 pull up							
pull ups are on video			5	7/9/18	188	188	188	GRADUATE	KS	188	N	GA
35	90	107	99	75.5				8	47	Y	11	

Figure 3.2. Mixed data shown in FY 18 dataset showing extra row. Source: MCRC data. Figure generated using Microsoft Excel.

After correcting for erroneously created rows, we repeated the steps above and tracked only the single row number where the mixed value was located using the following keywords: "EXTENSION," "WEIGHT," "DATE," and "IST" resulting in removing the single row of data. After eliminating the extra rows and mixed data type columns, we resolved the next issue of mistakenly created columns by removing them.

3.2.2 Minor Data Entry Errors

In this section, we will discuss how we addressed minor errors that occurred during data entry. General pre-processing of the data consisted of correcting the date-time format of columns with majority date values. Next, we created a fiscal year column so we could compile the datasets into one large Pandas data frame for output into a .csv file for loading into the statistical programming language R, for further processing (R Core Team 2022). Once the large data file was loaded into R, we used the table function to confirm that the remaining entries were usable for proceeding to the next pre-processing step. We used the R table function to identify columns that had errors. Minor errors mainly consisted of nonstandard "Race..MEPCOM..[1-5]," "Ethnic..MEPCOM.," "DEP.Discharge.Code," "DEP.Discharge.Reason," rectified by standardization. For example, we will show a lack of standardization in the variables "DEP.Discharge.Code" and "DEP.Discharge.Reason." The code for an applicant "DEP.Discharge.Reason" equivalent to "COMPONENT CODE CHANGE" was "ZKC," but some observations in the dataset had "COMP CODE CHANGE" as the "DEP.Discharge.Reason." We standardized all other dataset ethnic, race, and pairs of code and reason in the same matter.

3.2.3 Empty and Erroneous Values

There were several variables that had missing data, with Figure 3.3 only showing the variables out of the total of 70 with missing values. The variables "ASVAB...CL.Score" and "ASVAB...EL.Score" had values near zero percent. The variables related to the IST had the most missing data. The "IST Results" had all missing data, so it was removed. The female-only event of the IST flexed arm hang had the second-highest percentage of missing data, likely due to most of the data being male. A U.S. Marine Corps (2019) MARADMIN stated that on January 1, 2020, the IST Plank was introduced as an optional alternative to traditional crunches, which contributed to missing values for both the IST Plank and IST Crunch variables. Missing weights in both ship and inspect values could have been due to the applicant dropping out of the DEP. There was no hypothesized explanation for the absence of certain data, such as IST run results, height, and weight, except that they were never added to the MCRC database.

Figure 3.3. Percent of missing data. Source: MCRC data. Figure generated using R.

Imputation

Any obvious errors in the data for IST events, such as hanging, push-ups, planks, and crunches, were limited to world-record values. For example, in cases where multiple values for 1.5 mile run times exceeded the world record pace for one mile, the variable values were strictly confined to the pace of the world record mile (Guinness World Records 2023). Missing applicant weights and zero values were replaced with the mean. Missing "DEP.Extension" values were set to 0. There were 4,150 observations where the "Weight.At.Contract" values were missing. In these cases, the applicant weight was either set to the "Weight.At.Inspect," "IST.Weight," or "Weight.At.Ship." If none of these were available, the weight was set to the mean. The "IST.Height" was constrained to greater than 55 inches to account for "0" values.

3.3 Variables Added

In order to run an effective analysis, we decided to add a few features to help us better classify a Marine Applicant who ships to basic training.

3.3.1 Variables Added Using Python

We derived a binary variable to analyze the features of a Marine Applicant who is shipping to basic training. If the applicant had a ship date listed in the "Ship Date" column, the result was "1" and "0" otherwise. The "Ship" column was used as the thesis response variable, and the count in the final dataset is shown in Figure 3.4. It is worth noting that the "Disposition" category provided a lot of information on the applicant's status, as seen in Figure 3.5 from the FY 17 dataset. However, we decided that recording the applicant's MCRD graduation status went beyond the scope of this thesis, so it was ignored.

```
True     265965
False     60682
Name: Ship., dtype: int64
```

Figure 3.4. Final ship count. Source: MCRC data. Figure generated using Python.

```
GRADUATE             31611
DEP DISCHARGE         8385
MCRD DISCHARGE        3614
RECRUIT               1026
POOLEE                  13
DQ-OTHER                12
                         1
DQ-MORAL                 1
DEP AND HELD             1
DQ-MEDICAL               1
PENDING DISCHARGE        1
```

Figure 3.5. Disposition data. Source: MCRC data. Figure generated using Python.

3.3.2 Variables Added Using R

To simplify the dataset, we merged all categories with a race into a single unique race code. To analyze the impact of the DEP variable on shipping applicants, we created the "Days in DEP" variable. This was calculated by subtracting the "Ship Date" column from "MEPS Contract date," the difference resulted in the "Days in DEP" variable. If the applicant did not have a ship date, we used the DEP discharge date instead. Any members who lacked both a ship date and DEP discharge date were excluded from the dataset. The distribution of days in DEP is illustrated in Figure 3.6.

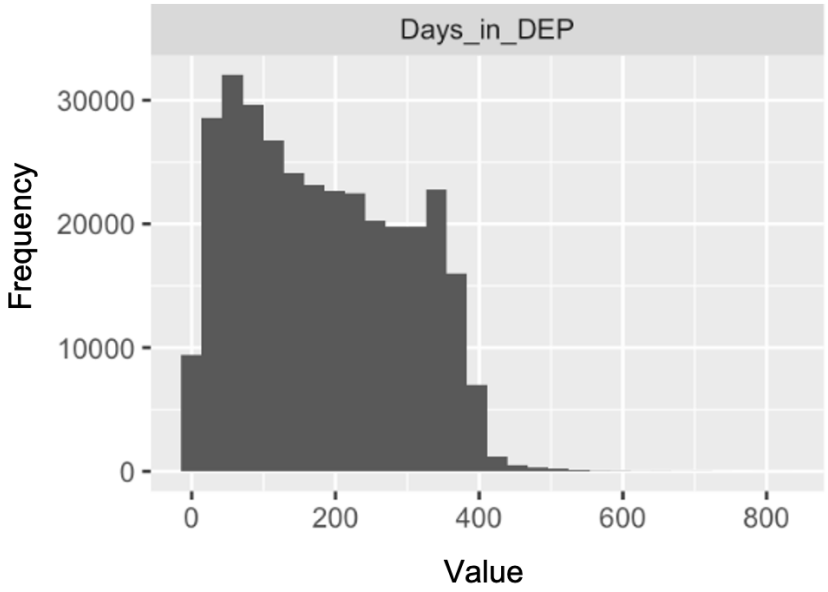


Figure 3.6. Final days in DEP. Source: MCRC data. Figure generated using R.

In order to gauge the impact of weight on shipping likelihood, we derived a weight difference variable. This metric measured the difference between the weight at the time of contract and the weight at the time of ship. In instances where the applicant didn't ship, we evaluated the weight difference between the contract and the inspection. If the inspection weight was not available, we assumed the difference value to be zero. The distribution of days in DEP is illustrated in Figure 3.7.

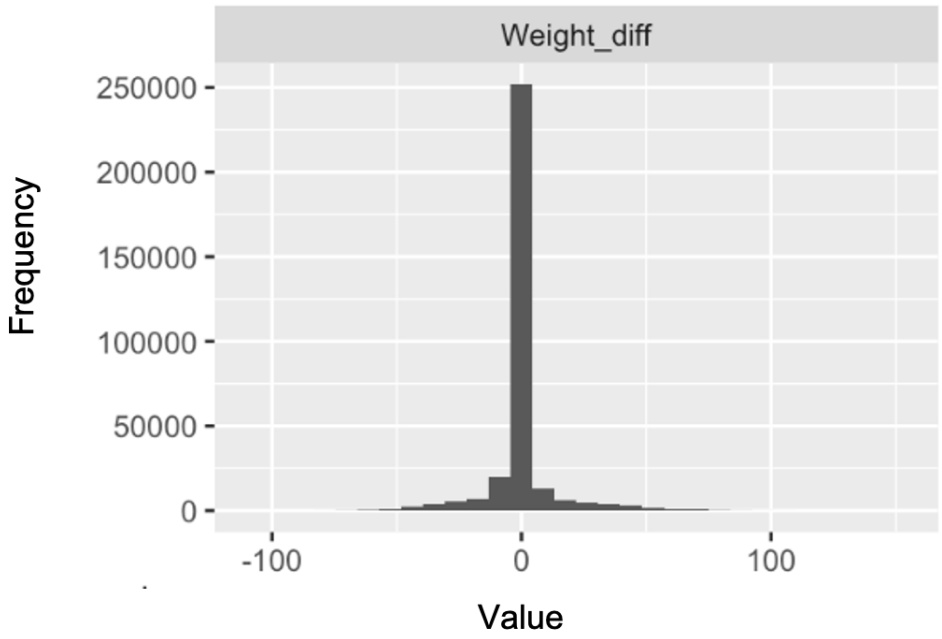


Figure 3.7. Final weight difference variable. Source: MCRC data. Figure generated using R.

In addition, we established a variable to assess the impact of the duration between inspection and contract dates. Our assumption was that a prolonged absence of communication with the applicant could potentially affect their shipping decision. We subtracted the inspection date from the contract date to calculate this variable, resulting in an integer day value. The distribution of days in DEP is illustrated in Figure 3.8.

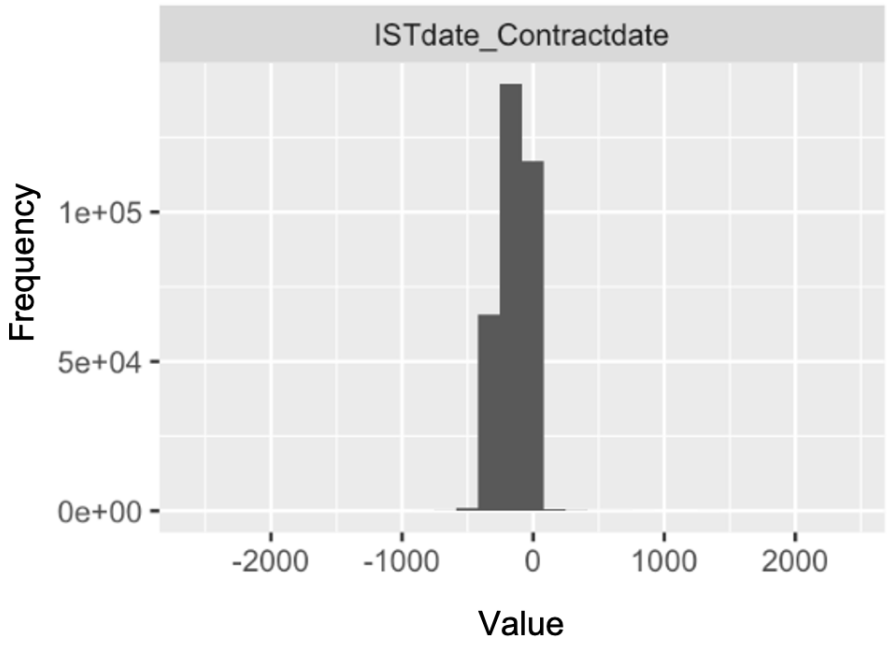


Figure 3.8. Final days between IST date and contract date. Source: MCRC data. Figure generated using R.

Variable Removal

We removed missing days from the DEP values as they were our response variable. Additionally, we removed the variables "Current.City" and "Current.Zipcode" as they were not used in our analysis. The geographic location data of each applicant was captured in our "Current.State" variable.

The data fields "MEPS.CONTRACT.DATE," "DEP.Date," "IST.Date," "Ship.Date," and "DEP.Discharge.Date" were not required for our analysis, as we had already calculated the necessary dates and had determined that a time-series analysis was beyond the scope of this thesis. "Weight.At.Contract," "Weight.At.Ship" and "Weight.At.Inspect" data was captured in our weight difference variable calculated above. "MCRD.Discharge.Reason.Description", "MCRD.Graduation.Date", "MCRD.Discharge.Date," were removed, as they did not serve in our analysis as they involved applicant data beyond shipping. "DEP.Discharge.Reason" variable information was captured in our DEP.Discharge.Code variable. "Region", "RS.Name", and "RSS.Name", recruiter information was captured in the MCRC dis-

trict variable. “IST.Notes”, “IST.Results”, “Disposition”, “Ship.To ‘’, “Grad boot”, “DEP.Discharge.Code” had no purpose in our analysis and the “Four.Ten.Program” variable was outside the scope of this thesis. All other race column data was captured in our created single combined race column, resulting in the removal of All other race columns. Waiver category code and status description variables were removed as the data was captured in separate similar columns.

3.4 Splitting the Data

The final pre-processed dataset consisted of and can be explored in-depth by referencing Appendix B. The final dataset consisted of 326,647 observations spanning 35 explanatory variables. Twenty variables contained discrete data, and 15 variables contained continuous data. A data dictionary is provided in Appendix B.

After splitting the data into different fiscal years, we encountered an issue with the large size of the entire dataset and limitations with R. To address this, we decided to sample a total of 30,000 observations using the random number generator in R. This resulted in a single dataset representing the full observations of each fiscal year and a sampled dataset containing observations from all FYs. We proceeded with an 80-20 split for training and testing on each fiscal year and sampled the full dataset utilizing the rnorm R function. This resulted in nine models that will be examined in the next chapter.

3.4.1 Packages Used

We utilized the randomForest() function from the randomForest package by Liaw and Wiener (2002) to construct a random forest model. To do this, we used the training set as input and set the ntrees parameter equal to 500. This parameter corresponds to the number of bootstrap sampled trees to grow (Liaw and Wiener 2002). After constructing the model, we then used the randomForestExplainer package by Biecek and Burzykowski (2020) to visualize the results. Specifically, the utilized functions were plot_multi_way_importance() and plot_min_depth_distribution() from the randomForestExplainer package to produce Figure 4.2 and Figure 4.3 (Biecek and Burzykowski 2020).

We used the ROCR package from Sing et al. (2021) to construct Receiver Operating Characteristic (ROC) curves aiding in model comparison shown in Figure 4.4 and Figure 4.5. Finally, the caret package by Kuhn (2021) was used to display the confusion matrices shown in Figure 4.1 (Kuhn 2021).

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Results

The results of the random forest models that we fit the data are presented through confusion matrices, two sets of multi-way importance plots, and ROC curves. Additional model results can be found in Appendix C, which closely match the aggregate model with any exceptions identified below.

4.1 Confusion Matrices

To evaluate the effectiveness of our classification model, the confusion matrix provided us with valuable information. The confusion matrix summarized the model's accuracy by showing the numbers of true positives, false positives, true negatives, and false negatives. Using the confusion matrix to assess our model's effectiveness in determining an applicant's shipping potential, we considered the following metrics. Sensitivity measured the probability of correctly predicting an applicant as a shipper, given that they do actually ship (Kuhn 2021). On the other hand, specificity represented the probability of a correct prediction of an applicant as a non-shipper when they truly did not ship. Precision measured the accuracy of predictions by displaying the percentage of applicants predicted to ship against applicants who actually shipped (Kuhn 2021). Recall gave us insight into the percentage of actual shippers that were correctly predicted as such. The F1 score measured the combination of precision and recall, while the Kappa statistic compared actual accuracy to random chance (Kuhn 2021). Finally, observed accuracy gave us information on the number of correctly classified instances throughout the confusion matrix. Our resultant confusion matrix for the aggregate FY is shown in Figure 4.1. The y-axis displays the model predicted ship and non ship classifications of an applicant in the testing dataset. The horizontal axis represents the actual non ship and ship applicants in the testing dataset.

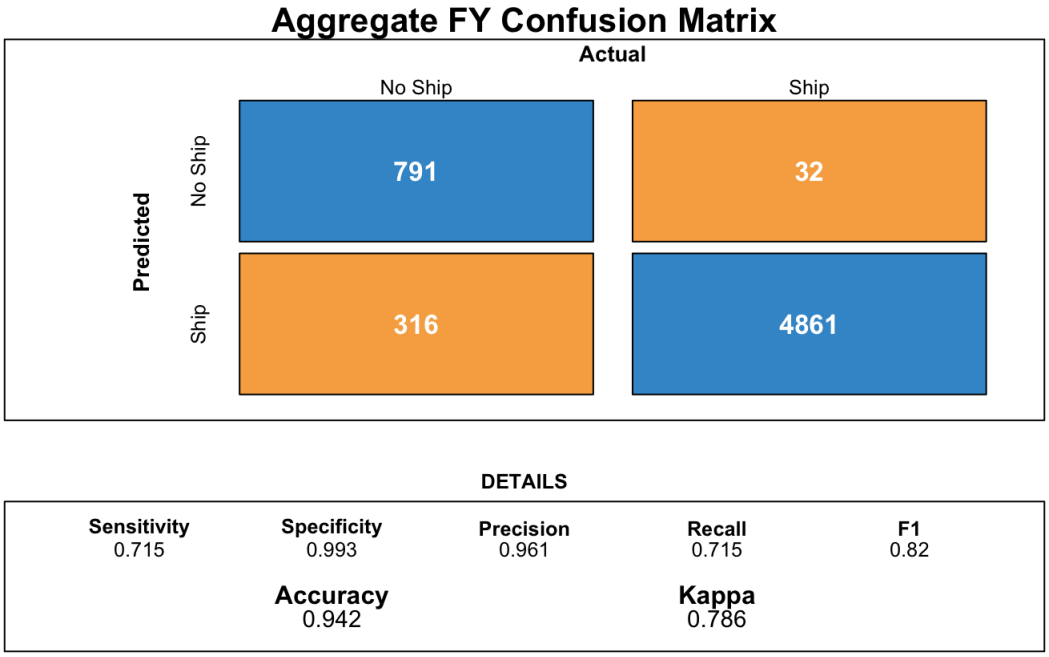


Figure 4.1. Aggregate model confusion matrix. Source: MCRC data. Figure generated using R.

Our model tended to make false positive errors based on our analysis, meaning there was a higher likelihood of classifying an applicant as a shipper even if the applicant did not actually ship to basic training. This is evidenced by a lower sensitivity at 71% when compared to specificity at 99%. Our model had high specificity, indicating that an applicant classified as a non-shipper was highly unlikely to ship.

4.2 Multi-way Importance Plots

This section covers the two multi-way importance plots used to visualize our results. Our first multi-way importance plot measured root versus mean minimum depth and the next multi-way importance plot measured Gini versus accuracy loss.

4.2.1 Multi-way Importance Plot: Root versus Mean Min Depth

It was helpful for us to look at two measures to determine the importance of a variable When analyzing our random forest model for the aggregate fiscal year sampled data. The

first measure was the number of trees the root was split on a particular variable. This count showed how many trees began with a particular variable as the root node in the forest ensemble. The second measure was the total number of nodes in the forest that split on the particular variable. This count included all nodes, both root and subsequent splits in the tree structure, and provided insight into the cumulative importance of the variable in the decision-making process (Biecek and Burzykowski 2020). Variables with high counts in both measures were considered influential in the model's predictions. On the other hand, if the counts were low, the variable had less impact on the model's predictions. Another helpful measure was the mean minimum depth, This was the average distance a single data point traveled from the starting point to reach the final prediction in the decision tree (Biecek and Burzykowski 2020). A lower mean minimum depth indicated a simpler decision tree was formed on a split of the variable, which suggested that the variable had a significant role in classifying a shipper. Conversely, a higher mean minimum depth implied a more complex decision tree and the variable had less importance in classifying a shipper. Figure 4.2 displays our first multi-way importance plot.

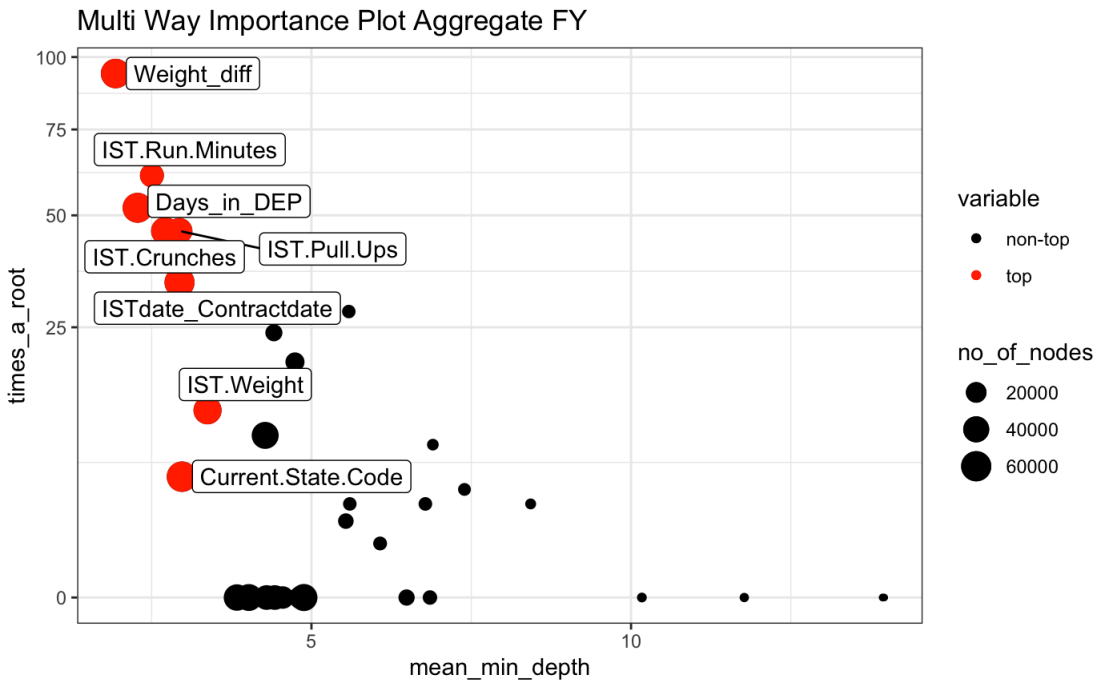


Figure 4.2. Aggregate model multi-way importance plot root versus mean min depth. Source: MCRC data. Figure generated using R.

Upon analyzing the data, we found that our model heavily relied on three variables: "Weight Difference," "IST.Run.Minutes" and "Days in DEP." These variables were represented by large plotted points on the graph, indicating that many nodes were split based on them. Furthermore, they were located in the upper left corner of the graph, which suggests that the random forest classifier used them as the primary root node as indicated on the y-axis, and decision trees branched off from there as indicated by the horizontal axis. Based on the results, it appears that these three variables are the most significant factors affecting the accuracy of the random forest classifier.

4.2.2 Multi-way Importance Plot: Gini versus Accuracy Loss

Various features or attributes were evaluated to determine the best splitting criterion for each node when constructing a decision tree. The Gini impurity is a measure of the disorder or heterogeneity within a set of data samples and represents the probability of a randomly selected data point to be miss-classified (Biecek and Burzykowski 2020). A higher Gini decrease indicated that the factor contributed more to make the data more organized and separated based on their categories or labels (Biecek and Burzykowski 2020). This decrease is calculated by the comparison between the Gini impurity of the current node with the weighted average of the impurities of the resulting child nodes after the split. The Gini decrease measured the extent to which the split criterion reduced the impurity in the data and signified an improved separation of the classes or labels in the data (Biecek and Burzykowski 2020). A larger Gini decrease indicated a more effective split. Each split aims to find the criterion that maximizes the decrease or minimizes the impurity to achieve informative subsequent splits (Biecek and Burzykowski 2020). On the other hand, a decrease in accuracy suggested that removing a particular variable from the analysis negatively affected the model's ability to make accurate classifications. It indicated that the variable carried valuable information which contributed to the model's overall accuracy. Our second multi-way importance plot is shown in Figure 4.3.

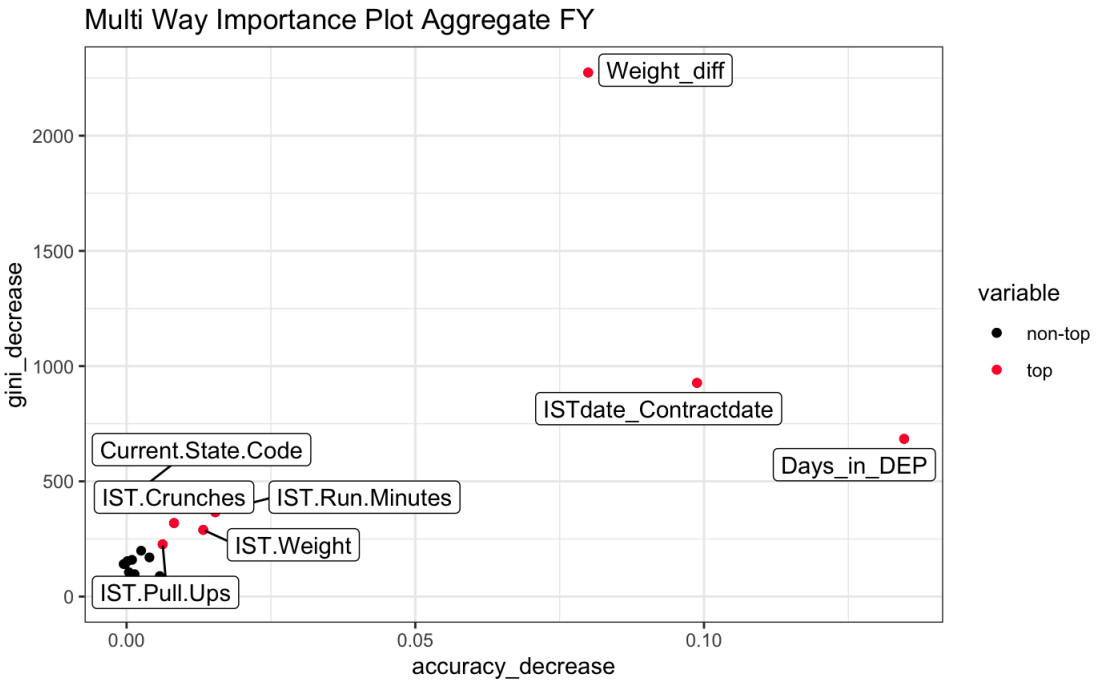


Figure 4.3. Aggregate model multi-way importance plot Gini versus accuracy loss. Source: MCRC data. Figure generated using R.

Upon analyzing the data, we discovered that our model heavily depended on three variables: "Weight Difference," "Days between IST Date and MEPS contract date," and "Days in DEP." These variables were situated in the top right corner of the graph, which implied that the random forest classifier experienced a significant decrease in organization when the variable was removed, as shown by the gini impurity decrease on the y-axis. Additionally, the model's accuracy decreased as these variables were removed, this indicates that they contained valuable information which contributed to the model's overall accuracy.

4.3 ROC Curves

To evaluate model robustness, we made used of ROC curves. Each model was used to predict the testing data of the remaining datasets. In order to verify that FY does not influence the response variable "Ship.", we cross-validated the prediction of a single FY model against all other FY test data, including the aggregate FY sample. The comparison of ROC curves is a widely accepted method for evaluation of the effectiveness of various classifiers or

predictive models (Sing et al. 2021). The ROC curve provided us with a graphical view of the trade-off between the true positive rate and false positive rate at several classification thresholds (Sing et al. 2021).

Several critical aspects were taken into consideration in comparing models with occurrences. The overall shape of the curves were compared first. Curves that were consistently closer to the top-left corner across all thresholds were generally considered superior. Curves that stayed consistently above the other curves were considered superior. Moreover, the smoothness and variability of the curves were analyzed. A smoother and less jagged curve indicated a more consistent and more dependable forecast (Sing et al. 2021).

We used the same cross-validation procedure for a fair and accurate comparison of all fiscal years. The aggregate model ROC plot is shown in Figure 4.4.

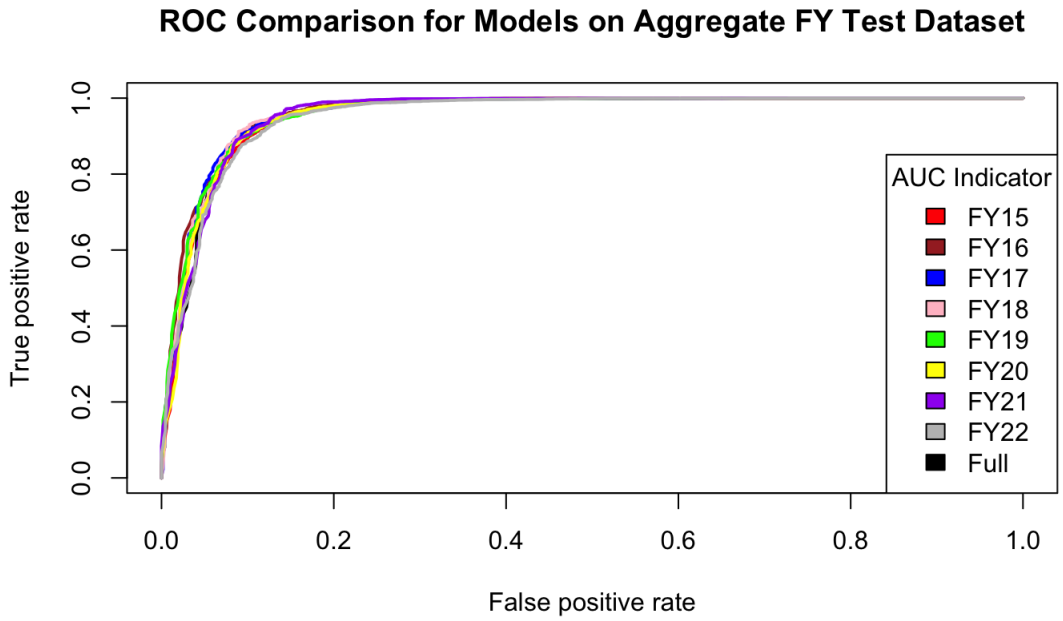


Figure 4.4. ROC curve for aggregate FY model. Source: MCRC data. Figure generated using R.

Based on our evaluation of the ROC curves, we discovered that, because all the lines represented for each FY were tightly joined, the aggregate model was robust in classifying shippers in the remaining FY test data. These results are similarly replicated for other FY models shown in Appendix C applied to all other test data sets with the exception to the FY 22 model. The FY 22 ROC curve is displayed in Figure 4.5.

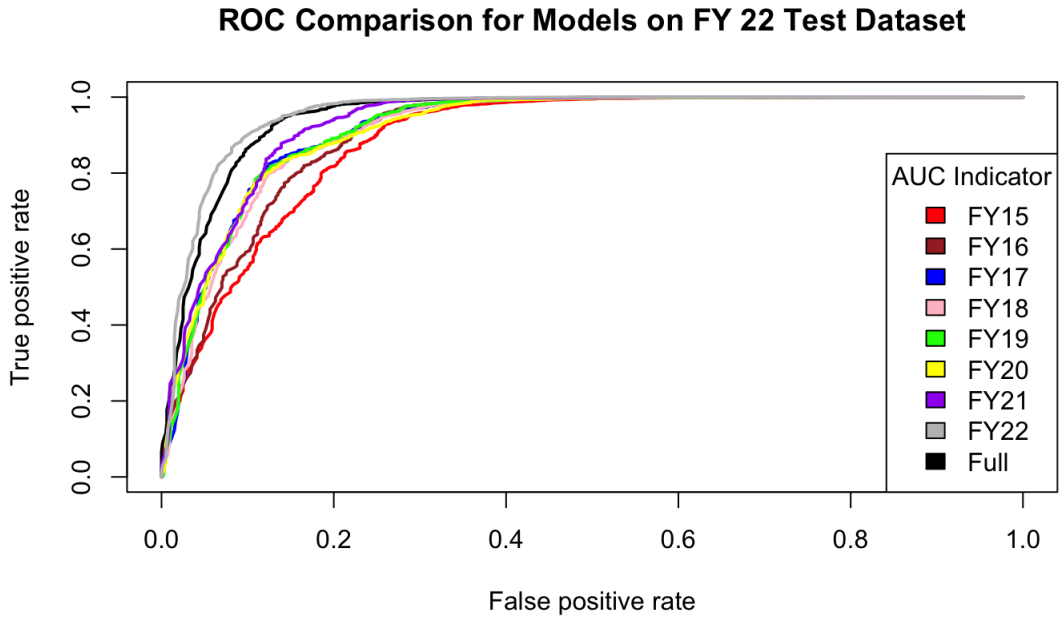


Figure 4.5. ROC curve for aggregate FY 22 model. Source: MCRC data. Figure generated using R.

Based on our evaluation of the ROC curves of Figure 4.5, we discovered that because all the lines representing each FY were less joined compared to Figure 4.4 and all other FY ROC curves, this indicated that further analysis was required to observe what separated the FY data from the other FY. We suspected that the FY 22 data set was incomplete based on the size of the FY 22 dataset compared to the other FY data. The comparison of the size of the datasets is observed in Figure 4.6.

FY	N
15	42272
16	44185
17	44052
18	44450
19	45188
20	38484
21	39549
22	28467

Figure 4.6. Number of observations of each FY. Source: MCRC data. Figure generated using R.

CHAPTER 5:
Summary and Future Research

This chapter brings the thesis to a close. We will summarize the results and discuss potential avenues for future research related to the topics covered in the thesis.

5.1 Summary

Our analysis of the Random Forest Model's performance shows that the model had high specificity but a tendency to make false positive errors when classifying applicants as shippers. Four variables heavily influenced the model's accuracy: "Weight Difference," "IST.Run.Minutes" "Days between IST Date and MEPS contract date," and "Days in DEP." Removing them decreased the model's accuracy. The ROC curve for FY 22 showed fewer joined lines, indicating incomplete data. Based on the results in the confusion matrices, it appears that applicants who were classified as not shipping to boot camp had a high likelihood of not shipping to boot camp. This information can be very useful in determining which applicants require additional attention when it comes to recruitment efforts, especially if they are deemed to pose a higher risk of not shipping while in DEP.

5.2 Future Research Opportunities

This section presents avenues for future research. Our suggestions include performing a Time Series Analysis, a Survival Analysis and thoroughly exploring the significant variables.

5.2.1 Time Series Analysis

It is suggested to observe how the season an applicant arrives in the DEP affects or influences the prediction of a hipper. Conducting time series analysis can uncover patterns and predict future values by utilizing data associated with a specific time (Brockwell and Davis 2016). For instance, in this research, it may be helpful to note that the model may behave differently in the winter compared to the summer months, just as they behave differently by fiscal year. Time series analysis is a specialized branch of statistics and data analysis that focuses on studying ordered, often temporal data (Brockwell and Davis 2016). The main purpose of

time series analysis is to extract meaningful statistics, patterns, and other characteristics from the data, and then use these insights for various applications (Brockwell and Davis 2016).

5.2.2 Survival Analysis

Time-to-event analysis, also known as survival analysis, is a statistical approach that involves examining the anticipated duration until one or more events occur (Kleinbaum and Klein 2005). It is commonly employed in medical research to determine the length of time until relapse, recovery, or death (Kleinbaum and Klein 2005). In this research, it would be beneficial to replace the time until death with the time until an applicant ships. Survival analysis employs several statistical techniques. The Kaplan-Meier estimator is a non-parametric method for predicting the survival function based on survival data. The Log-Rank test is used to test the hypothesis that there are no differences between the survival curves of two or more groups. Meanwhile, the Cox Proportional Hazards Model is a regression model that examines the impact of multiple variables on survival (Kleinbaum and Klein 2005).

5.2.3 Variable Analysis

Observing the target range of values that produce a shipper would be helpful in future research. For example, knowing the optimal days in DEP would be helpful to MCRC to be able to use as a target to aim for within recruiting each applicant, the same logic applies to each of the important variables identified.

APPENDIX A:
Original Data Dictionary

1. “ASVAB...AFQT.Score”: Integer value describing cumulative score on the ASVAB.
2. “ASVAB...CL.Score”: Integer value describing section score on the ASVAB for clerical related Military Operational Specialty (MOS) jobs.
3. “ASVAB...EL.Score”: Integer value describing cumulative score on the ASVAB for electronics related MOS jobs.
4. “ASVAB...GT.Score”: Integer value describing cumulative score on the ASVAB for general technical related MOS jobs.
5. “DEP.Date”: Date value originated by Marine Corps describing the date the applicant entered DEP.
6. “DEP.Discharge.Code”: Character value code describing the reason the applicant left the DEP.
7. “DEP.Discharge.Date”: Date value describing the date the applicant left the DEP.
8. “DEP.Discharge.Reason”: Character describing the reason the applicant left the DEP.
9. “DEP.Extension”: Integer value describing the number of days the applicant was extended in the DEP.
10. “Ethnic..MEPCOM.”: Character value describing whether the applicant is of Hispanic heritage or not.
11. “IST.Crunches”: Integer value totaling the number of crunches the applicant has completed on IST. The test ensures recruit is strong enough to meet the demands of boot camp and the physical fitness test of the Marine Corps is administered every six months. As of 01 January 2023, the Marine Corps exchanged this event of the IST to plank.
12. “IST.Date”: Date value describing date of the IST.
13. “IST.Hang.Seconds”: Integer value describing the total seconds an applicant can hang from the pull:up bar as part of the IST. (Female applicants only).
14. “IST.Height” Numeric value describing applicant height n date of IST.
15. “IST.Notes”: Character value for recruiter inputted comments on status of applicant.
16. “IST.Plank.Minutes”: Integer value describing the total minutes an applicant can stay

in the plank position during IST. Offered as an option to applicants prior to the Marine Corps change date of 01 January 2023.

17. "IST.Plank.Seconds": Integer value describing seconds an applicant can stay in the plank position during IST. Offered as an option to applicants prior to the Marine Corps change date of 01 January 2023.
18. "IST.Pull.Ups": Integer value describing number of completed pullups during IST. (Optional for female applicants).
19. "IST.Push.Ups": Integer value describing total number of push ups completed during IST.
20. "IST.Results" – Binary value describing whether the applicant participated in the IST.
21. "IST.Run.Minutes": Integer value describing the total minutes to complete 1.5 mile run during IST.
22. "IST.Run.Seconds": Integer value describing the total seconds to complete 1.5 mile run during IST.
23. "Ship.To '': Character value describing location applicant shipped for basic training.
24. "District": Character value describing 1 of 6 recruiting districts the applicant originated from.
25. "Region": Character value describing 1 of 2 recruiting regions the applicant originated from.
26. "RS.Name": Character value describing 1 of 48 recruiting stations the applicant originated from.
27. "RSS.Name": Character value describing 1 of 574 recruiting substations the applicant originated from.
28. "Ship.Date": Date value describing the date the applicant ships to basic training.
29. "Weight.At.Contract": Numeric value describing applicant weight on date of enlistment into DEP.
30. "Weight.At.Inspect": Numeric value describing applicant weight on date of periodic inspect date DEP.
31. "Weight.At.Ship": Integer value describing applicant weight on date on shipping to basic training.
32. "Disposition": Character value describing applicant status.
33. "Component.Code": Character value describing applicant.
34. "IST.Weight": Character value describing applicant.

35. “Four.Ten.Program”: Binary value describing whether applicant is a member of the four ten program.
36. “Current.City”: Character value describing City applicant originates.
37. “Current.State.Code”: Character value describing City applicant originates.
38. “Gender”: Binary value describing male or female applicant.
39. “MEPS.CONTRACT.DATE”: Date value describing date applicant entered the DEP.
40. “Waiver.1.Category.Code”: Single character code describing 1st waiver category.
41. “Waiver.1.Category.Description”: Character value describing 1st waiver category.
42. “Waiver.1.Status.Code”: Single Character value describing 1st waiver status.
43. “Waiver.1.Status.Description”: Character value describing 1st waiver status.
44. “Waiver.2.Category.Code”: Single character code describing 2nd waiver category.
45. “Waiver.2.Category.Description”: Character value describing 2nd waiver category.
46. “Waiver.2.Status.Code”: Single Character value describing 2nd waiver status.
47. “Waiver.2.Status.Description”: Character value describing 2nd waiver status.
48. “Waiver.3.Category.Code”: Single character code describing 3rd waiver category.
49. “Waiver.3.Category.Description”: Character value describing 3rd waiver category.
50. “Waiver.3.Status.Code”: Single Character value describing 3rd waiver status.
51. “Waiver.3.Status.Description”: Character value describing 3rd waiver status.
52. “Waiver.4.Category.Code”: Single character code describing 4th waiver category.
53. “Waiver.4.Category.Description”: Character value describing 4th waiver category.
54. “Waiver.4.Status.Code”: Single Character value describing 4th waiver status.
55. “Waiver.4.Status.Description”: Character value describing 4th waiver status.
56. “Waiver.5.Category.Code”: Single character code describing 5th waiver category.
57. “Waiver.5.Category.Description”: Character value describing 5th waiver category.
58. “Waiver.5.Status.Code”: Single Character value describing 5th waiver status.
59. “Waiver.5.Status.Description”: Character value describing 5th waiver status.
60. “Source”: Character value describing the recruiting source.
61. “Source.Code”: Character code describing the recruiting source.
62. “Current.Zipcode”: Character code describing applicant’s current zip code.
63. “MCRD.Discharge.Date”: Date value describing applicant DEP discharge date.
64. “MCRD.Discharge.Reason.Description”: Character value describing applicant Marine Corps Recruit Depot (MCRD) discharge reason.
65. “MCRD.Graduation.Date”: Date value describing applicant MCRD discharge reason.

66. “Race..MEPCOM..1”: Character value describing applicant 1st race value.
67. “Race..MEPCOM..2”: Character value describing applicant 2nd race value.
68. “Race..MEPCOM..3”: Character value describing applicant 3rd race value.
69. “Race..MEPCOM..4”: Character value describing applicant 4th race value.
70. “Race..MEPCOM..5”: Character value describing applicant 5th race value.

APPENDIX B:
Cleaned Data EDA

B.1 Univariate Distribution

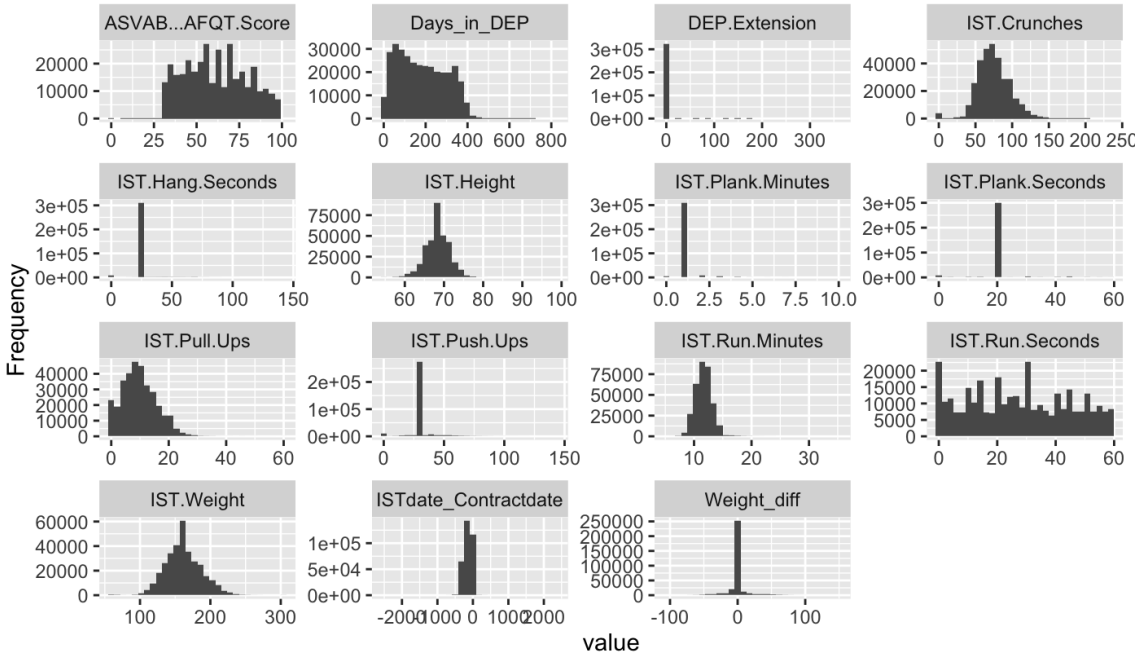


Figure B.1. Univariate distribution of full pre-processed dataset. Source: MCRC data. Figure generated using R.

B.2 Bar Charts (with Frequency)

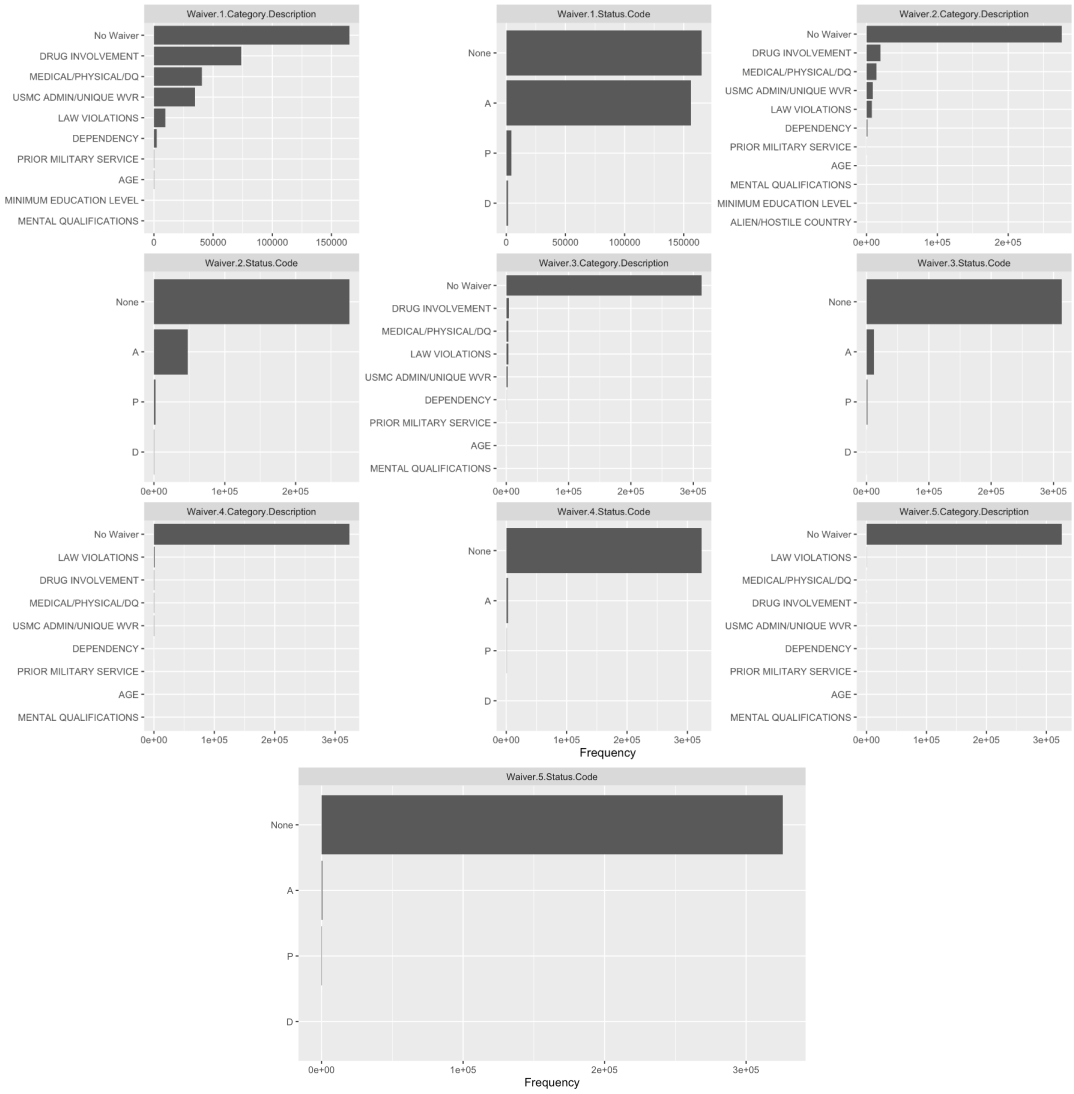


Figure B.2. Bar charts with frequency of full pre-processed dataset. Source: MCRC data. Figure generated using R.

B.3 Preprocessed Dataset Variable Data Dictionary

1. "ASVAB...AFQT.Score": Integer value describing cumulative score on the ASVAB.
2. "DEP.Extension": Integer value describing cumulative score on the ASVAB.
3. "IST.Crunches": Integer value totaling the number of crunches the applicant has completed on IST. The test ensures recruit is strong enough to meet the demands of boot camp and the physical fitness test of the Marine Corps is administered every six months. As of 01 January 2023, the Marine Corps exchanged this event of the IST to plank.
4. "IST.Hang.Seconds": Integer value describing the total seconds an applicant can hang from the pull:up bar as part of the IST. (Female applicants only).
5. "IST.Height" Numeric value describing applicant height n date of IST.
6. "IST.Plank.Minutes": Integer value describing the total minutes an applicant can stay in the plank position during IST. Offered as an option to applicants prior to the Marine Corps change date of 01 January 2023.
7. "IST.Plank.Seconds": Integer value describing seconds an applicant can stay in the plank position during IST. Offered as an option to applicants prior to the Marine Corps change date of 01 January 2023.
8. "IST.Pull.Ups": Integer value describing number of completed pullups during IST. (Optional for female applicants).
9. "IST.Push.Ups": Integer value describing total number of push ups completed during IST.
10. "IST.Run.Minutes": Integer value describing the total minutes to complete 1.5 mile run during IST.
11. "IST.Run.Seconds": Integer value describing the total seconds to complete 1.5 mile run during IST.
12. "District": Character value describing 1 of 6 recruiting districts the applicant originated from.
13. "Component.Code": Character value describing applicant.
14. "IST.Weight": Character value describing applicant.
15. "Four.Ten.Program": Binary value describing whether applicant is a member of the four ten program.
16. "Current.State.Code": Character value describing City applicant originates.
17. "Gender": Binary value describing male or female applicant.

18. “Waiver.1.Category.Description”: Character value describing 1st waiver category.
19. “Waiver.1.Status.Code”: Single Character value describing 1st waiver status.
20. “Waiver.2.Category.Description”: Character value describing 2nd waiver category.
21. “Waiver.2.Status.Code”: Single Character value describing 2nd waiver status.
22. “Waiver.3.Category.Description”: Character value describing 3rd waiver category.
23. “Waiver.3.Status.Code”: Single Character value describing 3rd waiver status.
24. “Waiver.4.Category.Description”: Character value describing 4th waiver category.
25. “Waiver.4.Status.Code”: Single Character value describing 4th waiver status.
26. “Waiver.5.Category.Description”: Character value describing 5th waiver category.
27. “Waiver.5.Status.Code”: Single Character value describing 5th waiver status.
28. “Source”: Character value describing the recruiting source.
29. “Source.Code”: Character code describing the recruiting source.
30. “RACE”: Character value describing the applicant’s race combination.
31. “Ship.”: Binary value representing applicant’s ship success.
32. “Days.in.DEP”: Integer value describing applicants total days in DEP.
33. “FY”: Numeric value describing FY Applicant was entered into the Information Support System.
34. “ISTdate.Contractdate”: Integer value describing applicants total days between applicant IST date and DEP entry date.
35. “Weight.diff”: Integer value describing applicants change in weight while in DEP.

APPENDIX C:
Expanded Result Figures

C.1 ROC Comparisons

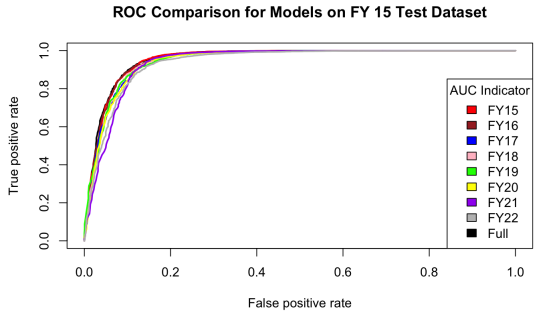


Figure C.1. ROC comparison for models on FY 15 test dataset. Source: MCRC data. Figure generated using R.

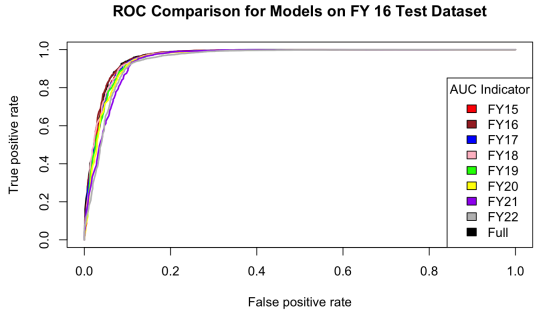


Figure C.2. ROC comparison for models on FY 16 test dataset. Source: MCRC data. Figure generated using R.

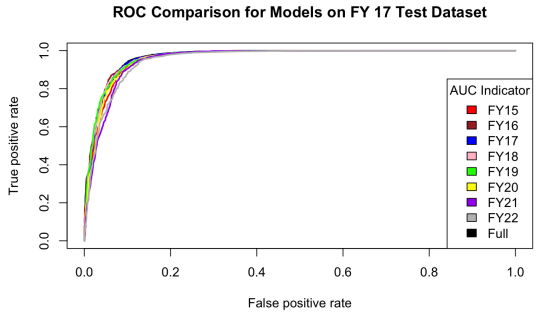


Figure C.3. ROC comparison for models on FY 17 test dataset. Source: MCRC data. Figure generated using R.

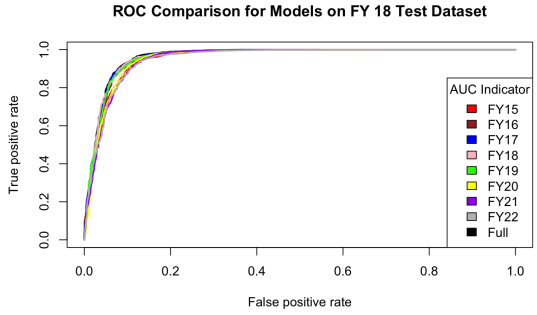


Figure C.4. ROC comparison for models on FY 18 test dataset. Source: MCRC data. Figure generated using R.

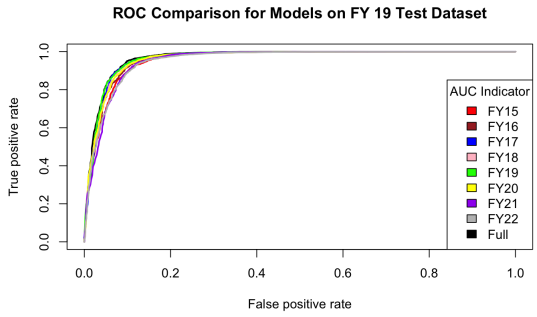


Figure C.5. ROC comparison for models on FY 19 test dataset. Source: MCRC data. Figure generated using R.

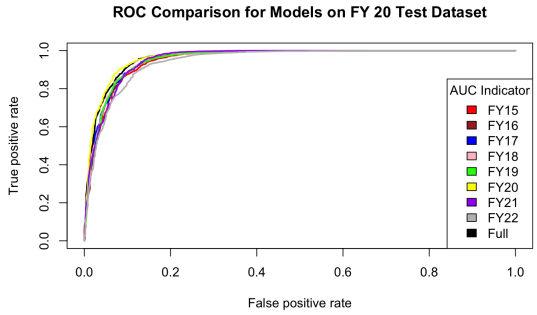


Figure C.6. ROC comparison for models on FY 20 test dataset. Source: MCRC data. Figure generated using R.

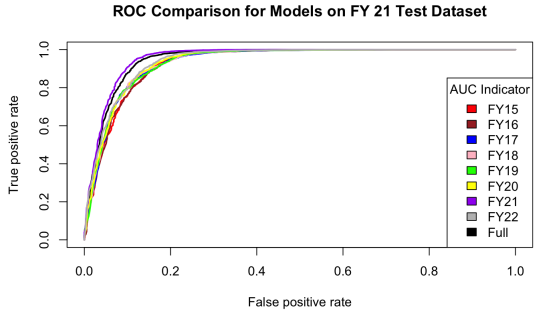


Figure C.7. ROC comparison for models on FY 21 test dataset. Source: MCRC data. Figure generated using R.

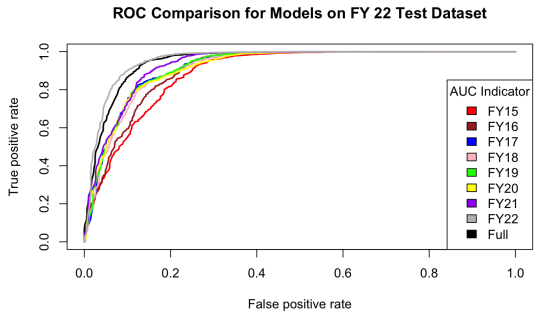


Figure C.8. ROC comparison for models on FY 22 test dataset. Source: MCRC data. Figure generated using R.

C.2 Confusion Matrices

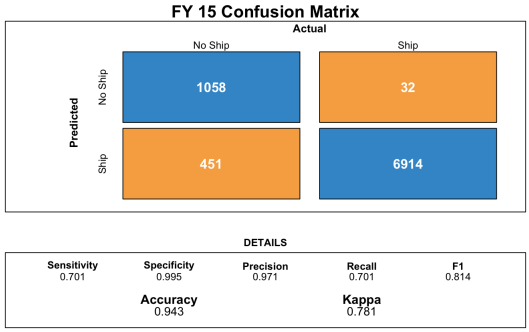


Figure C.9. Confusion matrix for FY 15 random forest model. Source: MCRC data. Figure generated using R.

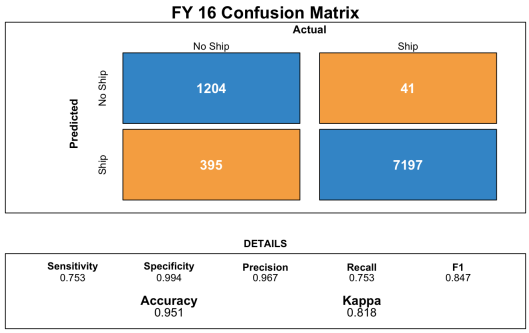


Figure C.10. Confusion matrix for FY 16 random forest model. Source: MCRC data. Figure generated using R.

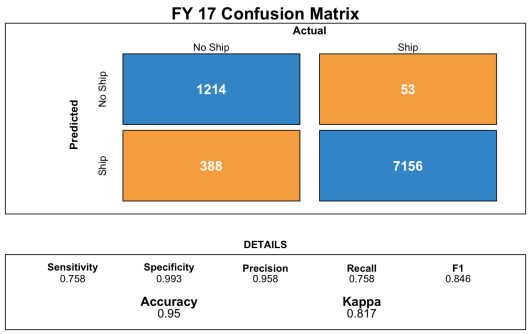


Figure C.11. Confusion matrix for FY 17 random forest model. Source: MCRC data. Figure generated using R.

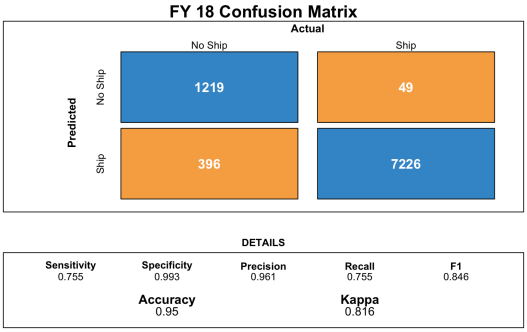


Figure C.12. Confusion matrix for FY 18 random forest model. Source: MCRC data. Figure generated using R.

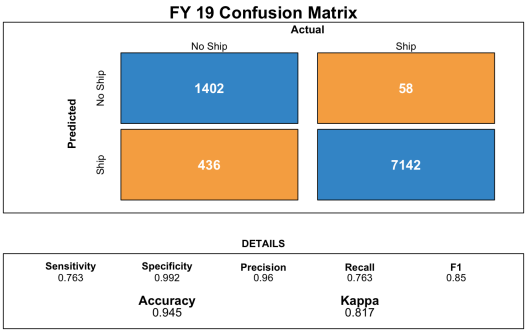


Figure C.13. Confusion matrix for FY 19 random forest model. Source: MCRC data. Figure generated using R.

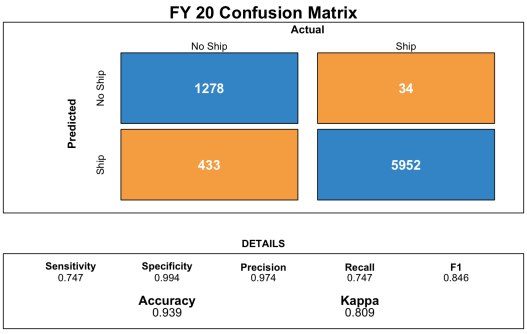


Figure C.14. Confusion matrix for FY 20 random forest model. Source: MCRC data. Figure generated using R.

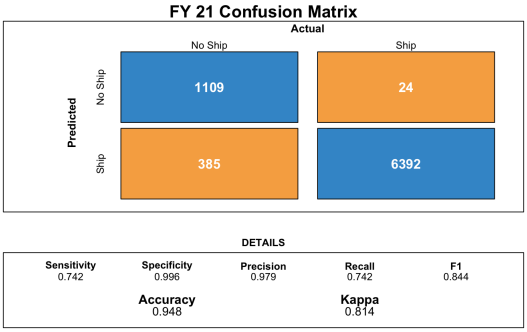


Figure C.15. Confusion matrix for FY 21 random forest model. Source: MCRC data. Figure generated using R.

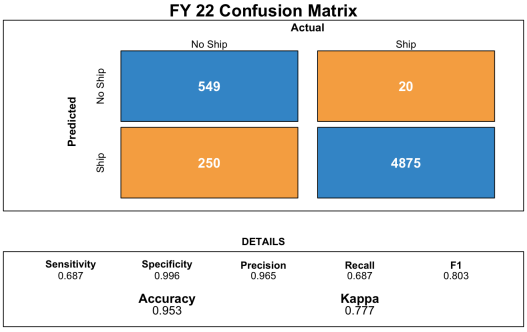


Figure C.16. Confusion matrix for FY 22 random forest model. Source: MCRC data. Figure generated using R.

C.3 Multi-way Importance Plots Root versus Mean Min Depth

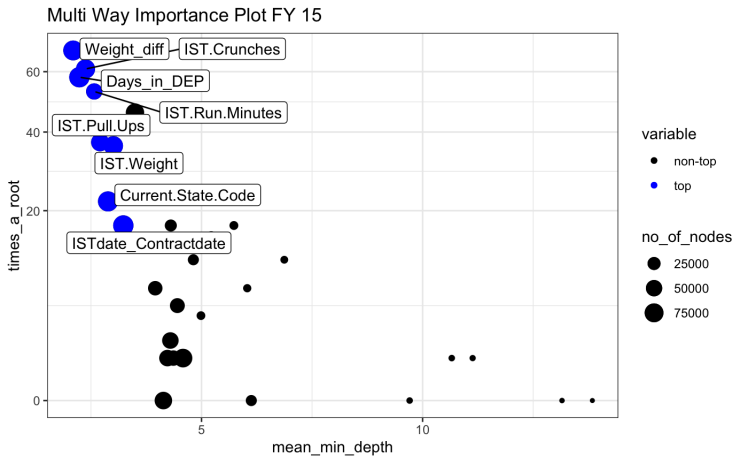


Figure C.17. Multi-way importance plot root versus mean min depth for the FY 15 random forest model. Source: MCRC data. Figure generated using R.

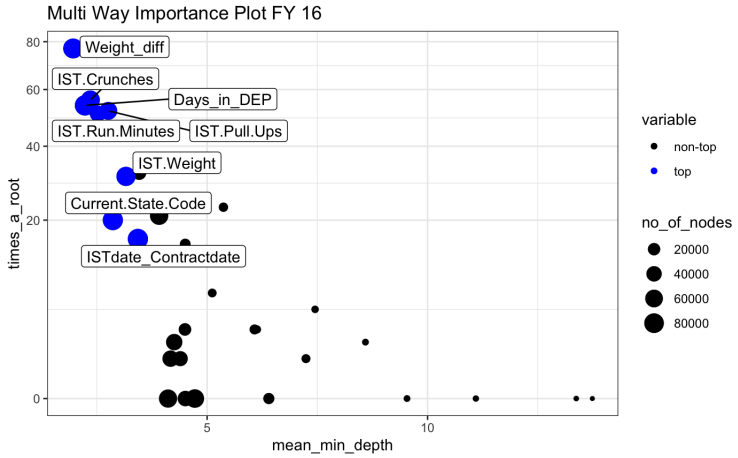


Figure C.18. Multi-way importance plot root versus mean min depth for the FY 16 random forest model. Source: MCRC data. Figure generated using R.

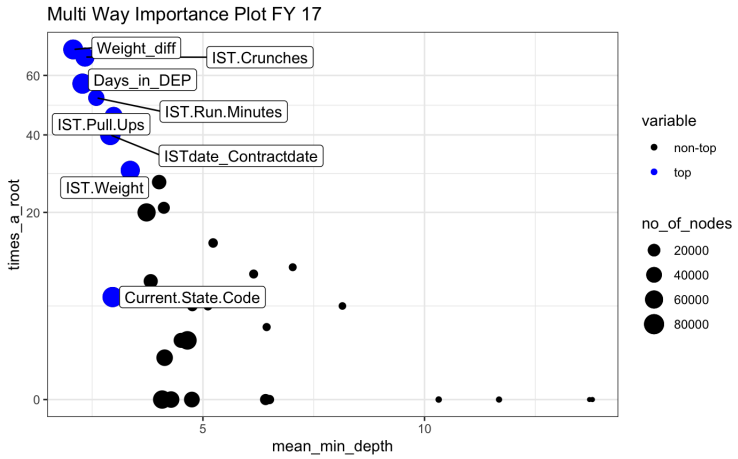


Figure C.19. Multi-way importance plot root versus mean min depth for the FY 17 random forest model. Source: MCRC data. Figure generated using R.

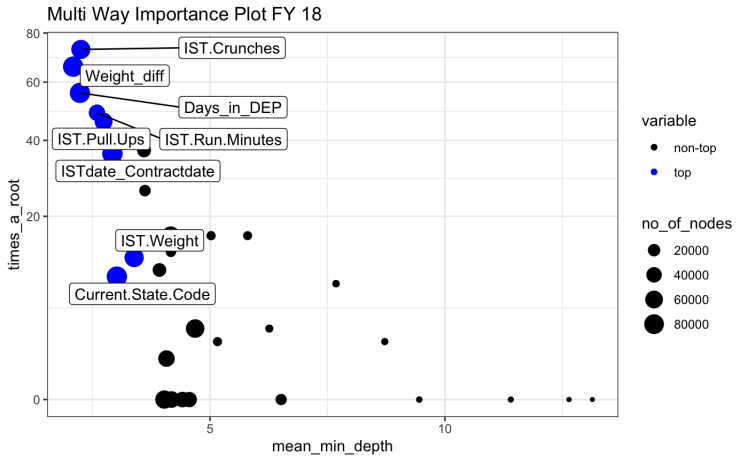


Figure C.20. Multi-way importance plot root versus mean min depth for the FY 18 random forest model. Source: MCRC data. Figure generated using R.

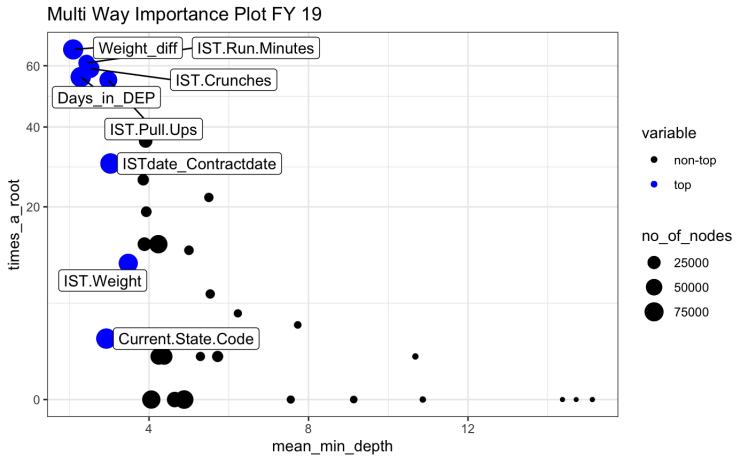


Figure C.21. Multi-way importance plot root versus mean min depth for the FY 19 random forest model. Source: MCRC data. Figure generated using R.

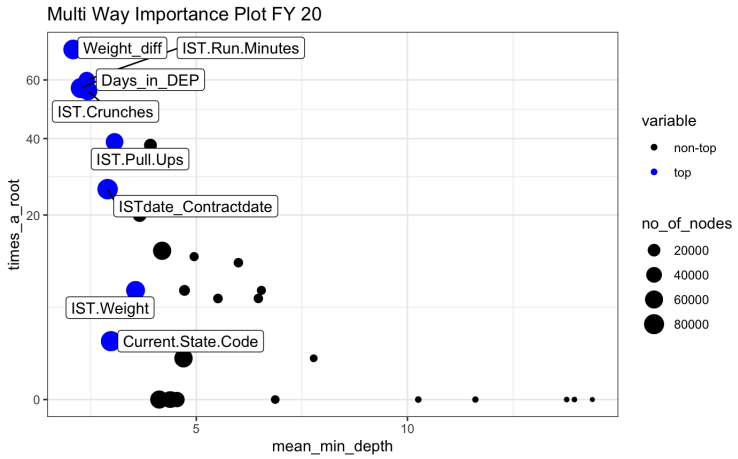


Figure C.22. Multi-way importance plot root versus mean min depth for the FY 20 random forest model. Source: MCRC data. Figure generated using R.

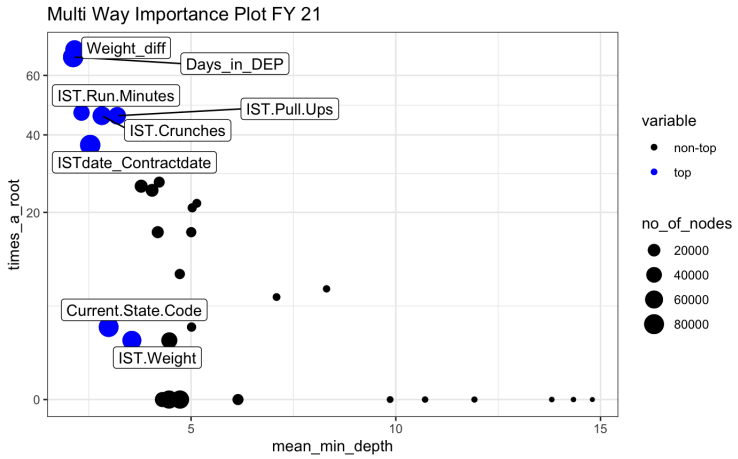


Figure C.23. Multi-way importance plot root versus mean min depth for the FY 21 random forest model. Source: MCRC data. Figure generated using R.

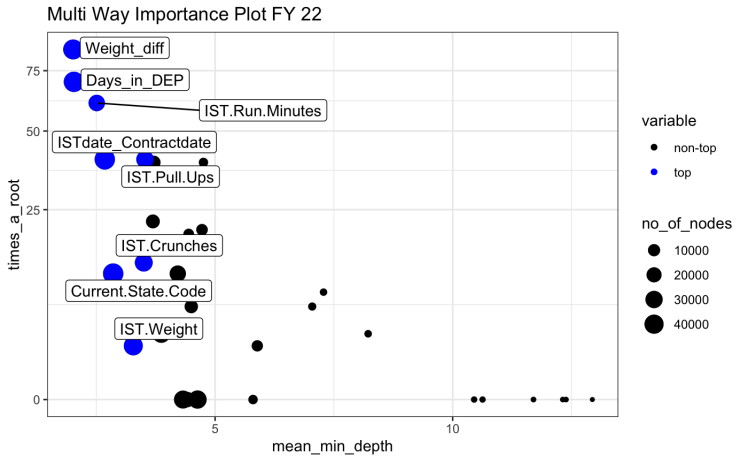


Figure C.24. Multi-way importance plot root versus mean min depth for the FY 22 random forest model. Source: MCRC data. Figure generated using R.

C.4 Multi-way Importance Plots Gini versus Accuracy Decrease

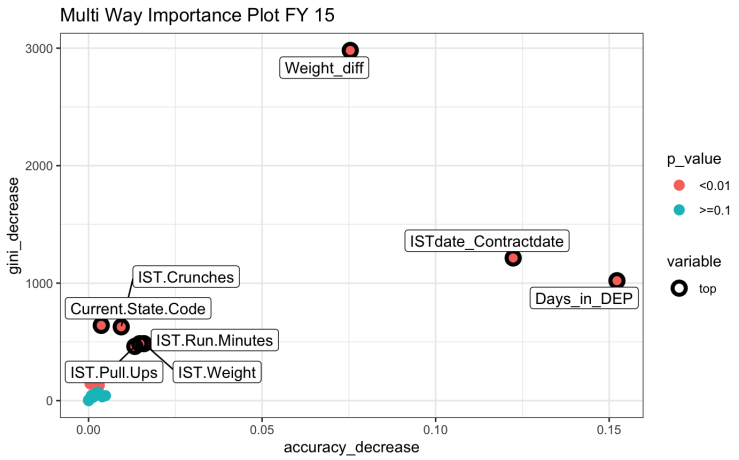


Figure C.25. Multi-way importance plot Gini versus accuracy decrease for the FY 15 random forest model. Source: MCRC data. Figure generated using R.

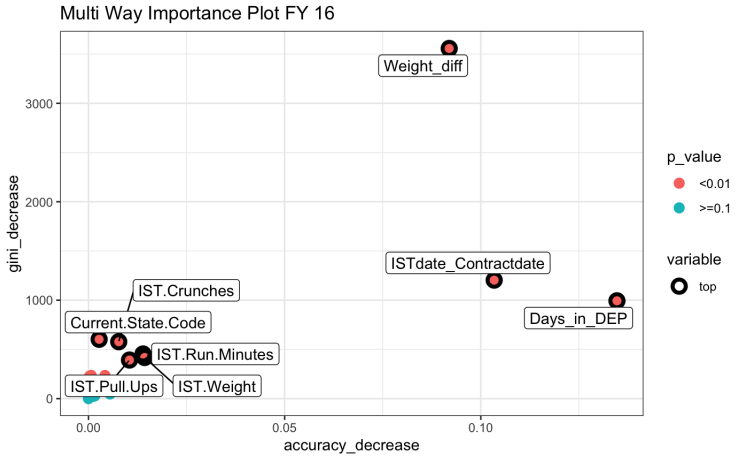


Figure C.26. Multi-way importance plot Gini versus accuracy decrease for the FY 16 random forest model. Source: MCRC data. Figure generated using R.

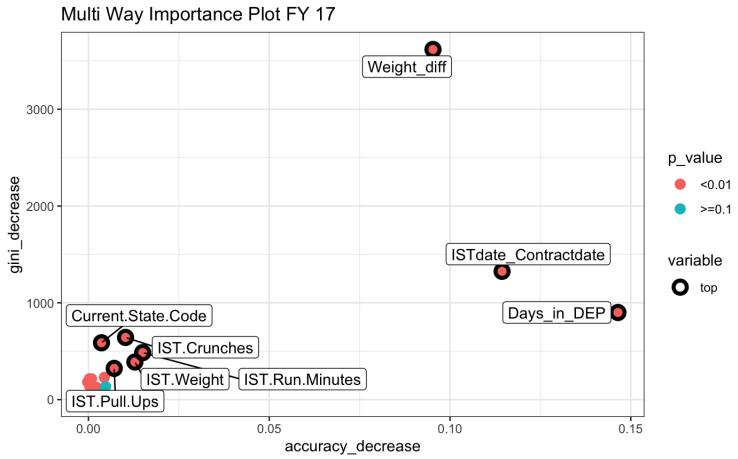


Figure C.27. Multi-way importance plot Gini versus accuracy decrease for the FY 17 random forest model. Source: MCRC data. Figure generated using R.

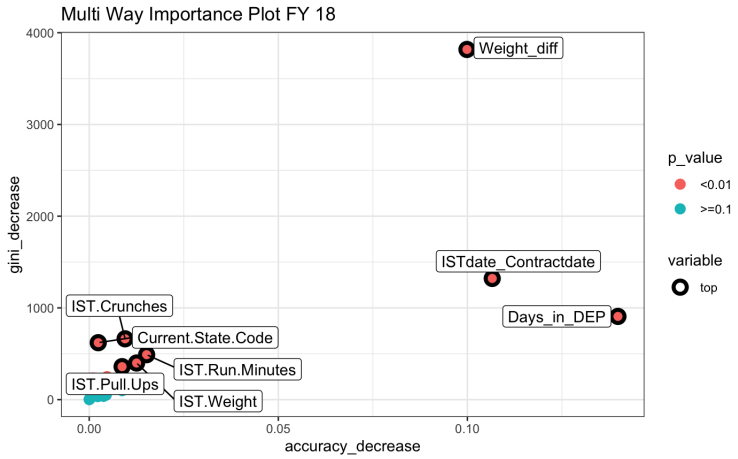


Figure C.28. Multi-way importance plot Gini versus accuracy decrease for the FY 18 random forest model. Source: MCRC data. Figure generated using R.

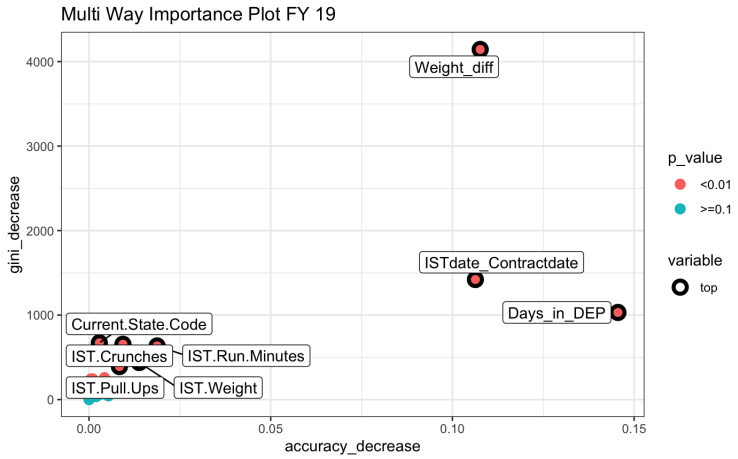


Figure C.29. Multi-way importance plot Gini versus accuracy decrease for the FY 19 random forest model. Source: MCRC data. Figure generated using R.

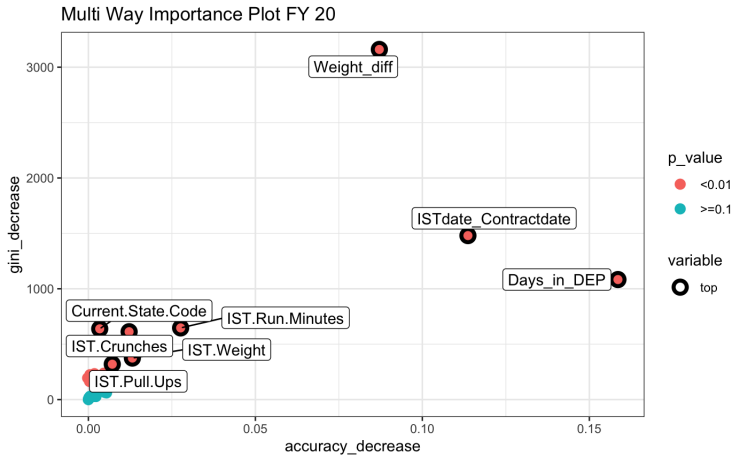


Figure C.30. Multi-way importance plot Gini versus accuracy decrease for the FY 20 random forest model. Source: MCRC data. Figure generated using R.

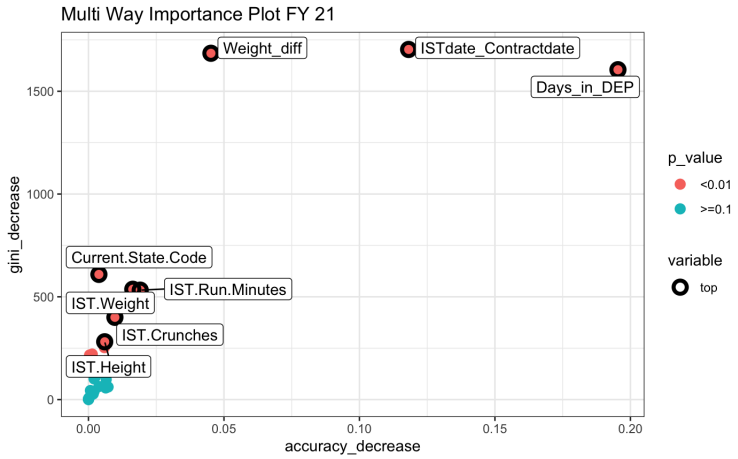


Figure C.31. Multi-way importance plot Gini versus accuracy decrease for the FY 21 random forest model. Source: MCRC data. Figure generated using R.

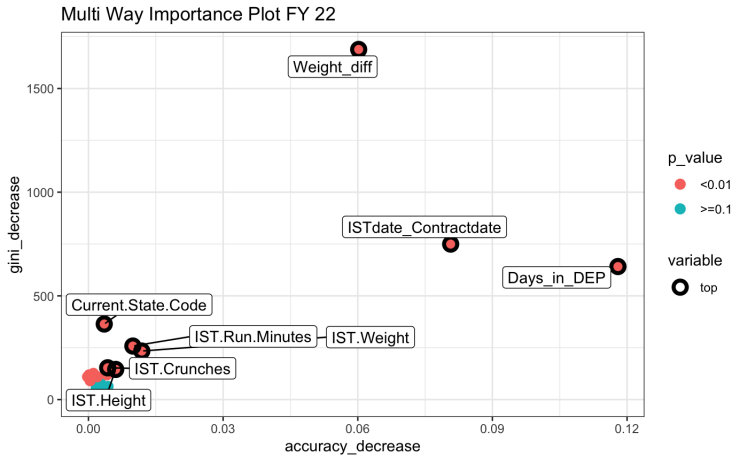


Figure C.32. Multi-way importance plot Gini versus accuracy decrease for the FY 22 random forest model. Source: MCRC data. Figure generated using R.

List of References

Athey P (2022) How big data might help the Marine Corps improve recruiting and retention. Marine Corps Times. Accessed August 4th, 2023, https://www.marinecorpstimes.com/news/your-marine-corps/2021/11/28/how-big-data-might-help-the-marine-corps improve-recruiting-and-retention/.

Barnett J (2022) Marine Corps Looks to machine learning for personnel retention. Defense Scoop. Accessed August 6th, 2023, https://defensescoop.com/2022/02/18/marine-corps-looking-to-machine-learning-for-personnel-retention/.

Bergan DE (2009) The draft lottery and attitudes towards the Vietnam War. *The Public Opinion Quarterly* 73(2), http://www.jstor.org/stable/25548087.

Biecek P, Burzykowski T (2020) Explaining and visualizing random forests in terms of variable importance. CRAN. Accessed August 4th 2023, https://CRAN.R-project.org/package=randomForestExplainer.

Blankshain JD, Cohn LP, Kriner DL (2022) Citizens to soldiers: Mobilization, cost perceptions, and support for military action. *Jurnalo of Global Security Studies* 7(4), ISSN 2057-3170, https://doi.org/10.1093/jogss/ogac017.

Boulesteix AL, Janitza S, Kruppa J (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining and Knowledge Discovery* 2(6):493–507, https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1072.

Brockwell PJ, Davis RA (2016) *Introduction to Time Series and Forecasting* (Springer).

Buchholz K (2023) The state of military conscription around the world. Statisitca. Accessed August 6th, 2023, https://www.statista.com/chart/3907/the-state-of-military-conscription-around-the-world.

Buttrey SE, Whitaker LR, Alt JK (2018) Developments in the statistical modeling of military recruiting. CHANCE. Accessed August 4th, 2023, https://hdl.handle.net/10945/61899.

Cancian M (2022) Analyzing the biggest changes in the marine corps force design 2030 update. Breaking Defense. Accessed August 4th 2023, https://breakingdefense.com/2022/06/analyzing-the-biggest-changes-in-the-marine-corps-force-design-2030-update/.

DOD (2023) Recruiting and retention numbers for fiscal year 2023 thru April 2023. Under Secretary for Personnel Readiness. Accessed August 4th, 2023, https://prhome.defense.gov/M-RA/Inside-M-RA/MPP/PR/.

Eckstein M (2022) Marine Corps will use AI to revamp recruiting and retention models. Defense News. Accessed August 4th 2023, https://www.defensenews.com/naval/2021/11/03/marine-corps-will-use-ai-to-revamp-recruiting-and-retention-models/.

Foley MS (2003) *Confronting the War Machine: Draft Resistance During the Vietnam War* (University of North Carolina Press).

Guinness World Records (2023) Guinness world records - fastest run one mile (male). Guinness World Records. Accessed August 4th 2023, https://www.guinnessworldrecords.com/world-records/fastest-run-one-mile-(male).

Hogarth AR (2017) Improving Navy Recruiting with the New Planned Resource Optimization Model with Experimental Design (PROM-WED). Master's Thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, https://hdl.handle.net/10945/52992.

Hooker G, Mentch L (2021) Bridging Breiman's Brook: From algorithmic modeling to statistical learning. *Observational Studies* 7(1):107–125, https://doi.org/10.1353/obs.2021.0027.

Hutson JH (1998) *Religion and the founding of the American Republic* (Washington, D.C: Library of Congress), ISBN 0844409480.

Kleinbaum DG, Klein M (2005) *Survival Analysis: A Self-Learning Text* (Springer Science & Business Media).

Kuhn M (2021) caret: Classification and regression training. CRAN. Accessed August 4th 2023, https://CRAN.R-project.org/package=caret.

Liaw A, Wiener M (2002) randomforest: Breiman and cutler's random forests for classification and regression. CRAN. Accessed August 4th 2023, https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

Lukanich PJ (2023) Survival analysis of army enlisted defense language institute graduate attrition factors. Master's Thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, https://hdl.handle.net/10945/72021.

Mongilio H (2023) Marines consulting outside experts for fixes to recruiting challenge. United States Naval Institute. Accessed August 4th, 2023,

https://news.usni.org/2023/01/25/marines-turning-to-outside-experts-for-fixes-to-recruiting-challenge.

NPR (2023) After Iraq, Mullen wants to prevent future Presidents from launching a War of Choice. National Public Radio. Accessed August 4th, 2023, https://www.npr.org/2023/03/20/1164641676/after-iraq-mullen-wants-to-prevent-future-presidents-from-launching-a-war-of-cho.

Paredes JA (2020) Making the Marine Corps Recruiting Process More Efficient. Master's Thesis, Defense Technical Information Center, Marine Corps University, Quantico, VA, https://apps.dtic.mil/sti/pdfs/AD1177814.pdf.

Pyne EM (2023) The uncertain future of the U.S. military's all-volunteer force. Council on Foreign Relations. Accessed August 4th, 2023, https://www.cfr.org/blog/uncertain-future-us-militarys-all-volunteer-force.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rose M, Hassen HR (2019) A Survey of Random Forest Pruning Techniques. Research Study, Department of Mathematical and Computer Sciences, Heriot Watt University, Dubai, UAE, https://www.semanticscholar.org/paper/A-Survey-of-Random-Forest-Pruning-Techniques-Rose-Hassen/1342eb0ade37bcdbaffa21e071b7fed7307e35f7.

Royster C (1979) *A Revolutionary People At War: The Continental Army and American Character* (University of North Carolina Press).

Sing T, Sander O, Beerenwinkel N, Lengauer T (2021) ROCR: Visualizing the performance of scoring classifiers. CRAN. Accessed August 4th 2023, https://CRAN.R-project.org/package=ROCR.

Swenson ZH (2020) Predictive statistical modeling of naval reserve officers training corps attrition. Master's Thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, https://hdl.handle.net/10945/66727.

Urben H, Feaver P (2023) The all-volunteer force at 50: Civil-military solutions in a time of partisan polarization. Just Security. Accessed August 4th, 2023, https://www.justsecurity.org/87053/the-all-volunteer-force-at-50-civil-military-solutions-in-a-time-of-partisan-polarization/.

U.S. Constitution (1776) Source of congress's war powers. Libary of Congress. Accessed August 6th, 2023, https://constitution.congress.gov/browse/essay/artI-S8-C11-1/ALDE$_0$0013587/.

U.S. Marine Corps (2019) Forthcoming change to the PFT. MARADMIN 330/19 June 06, 2019, https://www.marines.mil/News/Messages/MARADMINS/Article/1869148/forthcoming-change-to-the-physical-fitness-test-pft/.

U.S. Marine Corps (2020) Force design 2030. Force Design 2030, Washington, DC, https://www.marines.mil/Force-Design-2030/.

U.S. Marine Corps (2021) Talent management 2030. Talent Management 2030, Washington, DC, https://www.marines.mil/Talent-Management-2030/.

U.S. Marine Corps (2023) Talent management 2030 update march 2023. Talent Management 2030 Update March 2023, Washington, DC, https://www.marines.mil/Talent-Management-2030/.

Vanegas JMA, Wine W, Drasgow F (2022) Predictions of attrition among U.S. Marine Corps: Comparison of four predictive methods. *Military Psychology* 34(2):147–166, https://doi.org/10.1080/08995605.2021.1978754.

Vie LL, Scheier LM, Lester PB, Ho TE, Labarthe DR, Seligman ME (2015) The U.S. army person-event data environment: A military-civilian big data enterprise. *Big Data* 3(2):67–79, https://doi.org/10.1089/big.2014.0055.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California