2023-03

WHO LEAVES? INDIVIDUAL-BASED PREDICTIVE MODELING OF NON-END OF ACTIVE SERVICE ATTRITION FOR ENLISTED MARINES

Falk, Aaron

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

WHO LEAVES? INDIVIDUAL-BASED PREDICTIVE MODELING OF NON-END OF ACTIVE SERVICE ATTRITION FOR ENLISTED MARINES

by

Aaron Falk

March 2023

Thesis Advisor: Marigee Bacolod
Co-Advisor: Chad W. Seagren

Approved for public release. Distribution is unlimited.

This project was funded in part by the NPS Naval Research Program.

THIS PAGE INTENTIONALLY LEFT BLANK

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

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

4. TITLE AND SUBTITLE WHO LEAVES? INDIVIDUAL-BASED PREDICTIVE MODELING OF NON-END OF ACTIVE SERVICE ATTRITION FOR ENLISTED MARINES	**5. FUNDING NUMBERS** NPS-22-M205-A
6. AUTHOR(S) Aaron Falk	
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. This project was funded in part by the NPS Naval Research Program.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

Talent Management 2030 posits that the United States Marine Corps' manpower system hails from the industrial era and calls for broad modernization. This thesis serves as a proof of concept designed to implement modern predictive machine-learning algorithms and techniques to an age-old military manpower problem. Current Marine Corps attrition modeling is conducted using historical averages and does not account for individual attributes of each Marine. This study employs two machine-learning models, a Random Forest classifier and a multinomial logistic regression with least absolute shrinkage and selection operator predictor selection. It uses individual, disaggregated data and compares the prediction results to current Marine Corps attrition modeling processes. Two key findings are reported. First, the Random Forest classifier models outperform the current trailing average models at predicting aggregate attrition. One caveat is that these models have difficulty at correctly classifying non-end of active service attrition at the Marine level, achieving an average of 45% correct individual classification. Second, even though the machine-learning models provide superior prediction, they may not be managerially relevant because of the opportunity cost of construction due to the current database structure, data systems, and capabilities employed by Marine Corps manpower entities.

14. SUBJECT TERMS manpower, USMC, predictive, machine learning, enlisted, attrition, end of active service, random forest, classifier, logistic regression, LASSO	**15. NUMBER OF PAGES** 105
	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.

WHO LEAVES? INDIVIDUAL-BASED PREDICTIVE MODELING OF NON-END OF ACTIVE SERVICE ATTRITION FOR ENLISTED MARINES

Aaron Falk
Major, United States Marine Corps
BA, Hillsdale College, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 2023

Approved by: Marigee Bacolod
Advisor

Chad W. Seagren
Co-Advisor

Marigee Bacolod
Academic Associate, Department of Defense Management

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Talent Management 2030 posits that the United States Marine Corps' manpower system hails from the industrial era and calls for broad modernization. This thesis serves as a proof of concept designed to implement modern predictive machine-learning algorithms and techniques to an age-old military manpower problem. Current Marine Corps attrition modeling is conducted using historical averages and does not account for individual attributes of each Marine. This study employs two machine-learning models, a Random Forest classifier and a multinomial logistic regression with least absolute shrinkage and selection operator predictor selection. It uses individual, disaggregated data and compares the prediction results to current Marine Corps attrition modeling processes. Two key findings are reported. First, the Random Forest classifier models outperform the current trailing average models at predicting aggregate attrition. One caveat is that these models have difficulty at correctly classifying non-end of active service attrition at the Marine level, achieving an average of 45% correct individual classification. Second, even though the machine-learning models provide superior prediction, they may not be managerially relevant because of the opportunity cost of construction due to the current database structure, data systems, and capabilities employed by Marine Corps manpower entities.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AFQT	Armed Forces Qualification Test
AUC	Area Under the Curve
CFT	Combat Fitness Test
CMC	Commandant of the Marine Corps
Con	Conduct Score
CSV	Comma Separated Values
DC	Deputy Commandant
DEP	Delayed Entry Processing
DOD	Department of Defense
EAS	End of Active Service
ECC	Expiration of Current Contract
ENTNAC	Entrance National Agency Check
FitRep	Fitness Report
FY	Fiscal Year
HPC	High Performance Computing
HRDP	Human Resource Development Process
LASSO	Least Absolute Shrinkage and Selection Operator
M&RA	Manpower and Reserve Affairs
MCMAP	Marine Corps Martial Arts Program
MCO	Marine Corps Order
MinN	Minimum Node Size
ML	Machine Learning
MMIB	Manpower Management Integration Branch
MPP	Manpower Plans and Policy
Mtry	Number of Randomly Drawn Candidate Variables
NDAA	National Defense Authorization Act
NEAS	Non-End of Active Service
NPS	Naval Postgraduate School

PFT	Physical Fitness Test
PMCC	Present Monitored Command Code
PMOS	Primary Military Occupational Specialty
Pro	Proficiency Score
ROC	Receiver Operator Characteristic
SMOTE	Synthetic Minority Oversampling Technique
TFDW	Total Force Data Warehouse
USMC	United States Marine Corps

ACKNOWLEDGMENTS

This thesis is dedicated to my late father, John Falk. His intellect, wisdom, spirit of adventure, and devotion to his Creator continue to inspire me to take challenges head on, with confident determination. I thank my wife, Mandie, and my kids, Kenzi, Asher, and Olivia, for enduring the countless hours spent on this project. Dr. Marigee Bacolod and Dr. Marine Chad Seagren have my sincerest thanks for making this thesis possible. Your insights, willing instruction, and selflessness speak volumes about your character and dedication to the Department of Defense. Thanks to Dr. Jeff Haferman for his patient instruction and assistance on employing high-performance computing for this project. Thanks to John Forbes for his support with obtaining the large amount of data needed. A special thanks to my Hillsdale College buddy and roommate, Josh Trojniak. We've come a long way from the barracks, and your data science skills and friendship greatly contributed to my successful tenure at NPS and this thesis. Finally, I thank God for his redemptive Grace.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The United States Marine Corps (USMC) has been urged to modernize its manpower systems to meet the operational demands of the 21st century, according to recent strategic guidance. As part of this effort, this thesis explores the modernization of pre-contract, or, non-end of active service (NEAS), attrition prediction within the USMC. The cost of NEAS attrition to the institution is significant, and current Manpower and Reserve Affairs (M&RA) attrition prediction modeling relies on historical averages, which fails to account for specific attributes of individual Marines or predict attrition classification at the individual level. I implement modern predictive machine learning (ML) algorithms and techniques to classify individual Marines within the current enlisted inventory using two modeling techniques, a Random Forest classifier and a least absolute shrinkage and selection operator multinomial logistic regression. I compare the prediction results to current USMC attrition modeling processes, thus providing a proof of concept for the implementation of modern ML techniques to improve NEAS attrition prediction.

The models are trained on separate monthly data pooled between fiscal year (FY) 17 and 19, while the model is validated using data from the same month in FY20. Each monthly model training data consists of roughly 450,000 observations. Figure 1 depicts an example of the generalized model design for the month of October. Total Force Data Warehouse and Manpower Information Systems Branch supplied all data. The model defines and classifies each individual Marine into three predictive states within the specific month: continue service, end of active service, and non-end of active service.

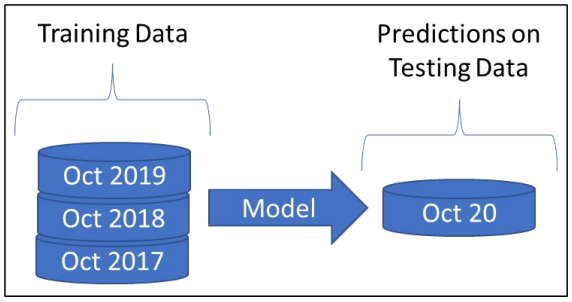


Figure 1. Model Design

Two key findings are reported. First, the Random Forest classifier series of models generally outperform currently used trailing average models at predicting aggregate NEAS attrition. The Random Forest classifier models underestimate aggregate annual NEAS attrition by 4% compared to the current modeling process that underestimates by 12% on the FY20 validation data. Figure 2 and Figure 3 compare NEAS predictions using the current modeling technique and the Random Forest classifier by month.

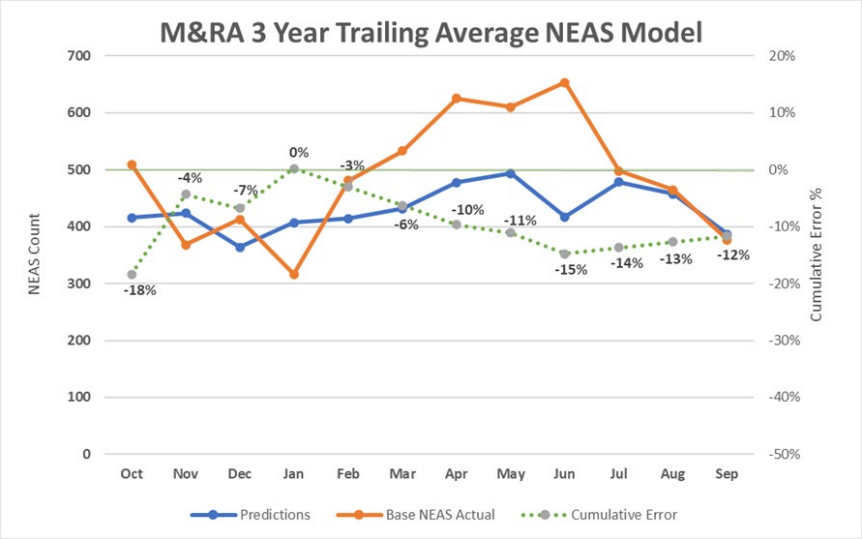


Figure 2. M&RA Current Non-End of Active Service (NEAS) Model

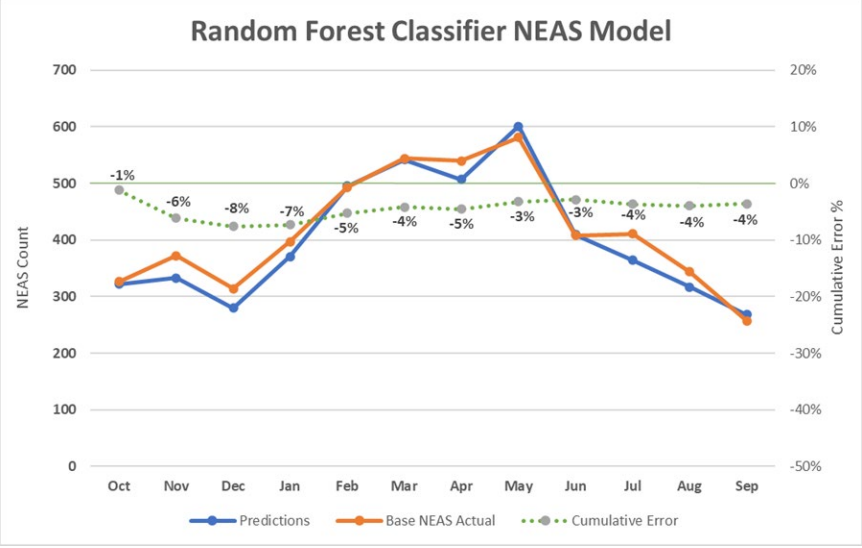


Figure 3. Random Forest Classifier Machine Learning NEAS Model

Second, even though the ML models provide superior aggregate prediction, they may not be immediately managerially relevant. Currently, the database structure, data systems, and capabilities employed by the services' manpower planners are not designed to support the use of these nascent ML techniques. For example, there is a significant effort and cost in construction of training data, from pulling data tables across multiple systems to cleaning and verification of data. Training ML models also requires significant computing power that is not easily accessible.

The results of this thesis indicate that harnessing ML to modernize manpower models can yield better predictive results, at least for aggregate NEAS attrition. Additionally, even though correct NEAS classification at the individual level hovered around 45% accuracy, value can be obtained from correctly identifying this half of Marines for potential targeting. More broadly, having demonstrated the proof of concept in this thesis, there are likely many other applications for ML within the USMC manpower system.

The institution should focus on three areas. First, it should optimize its data systems and data infrastructure to better facilitate the use of ML. Second, it should continue to look for opportunities to employ ML in areas that currently use legacy systems or techniques. It should also explore using ML techniques in conjunction with legacy modeling in a hybrid fashion. Finally, the institution should continue to strive to collect better data that can help build more accurate future ML models.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Personnel attrition remains a significant concern for both public and private organizations of all types and all sizes. The Department of Defense (DOD) has a vested interest in studying and understanding attrition and its causes. This is especially true given the DOD's constant manpower resource constraints; it must internally grow and develop its own talent. Additionally, the DOD must be able to effectively operate within an environment of constant manpower turnover. The services must be able to predict future attrition to accurately reach their annually mandated end strength goals, and more importantly, meet their operational missions.

Manpower modeling and data analysis techniques have continued to advance in the current data driven environment, specifically within the private sector. The DOD, and the United States Marine Corps (USMC), has not been able to maintain pace with these advancements. This thesis explores nascent manpower modeling approaches to try and better predict enlisted attrition, specifically those who will depart the service before their contract expires, using data that is already collected on the individual Marine.

A. BACKGROUND

The Fiscal Year (FY) 2022 National Defense Authorization Act (NDAA) is the primary legislative document that "authorize(s) appropriations for fiscal year 2022 for military activities of the Department of Defense, for military construction, and for defense activities of the Department of Energy, to prescribe military personnel strengths for such fiscal year, and for other purposes" (*National Defense Authorization Act for Fiscal Year 2022*, 2021). For the purposes of my thesis, the primary applicable component of the NDAA is the mandated end strength for the USMC. This end strength number must be met at the conclusion of each FY. Manpower planners within the USMC work throughout the fiscal year adjusting various components within the system to achieve the mandated end strength number.

Figure 1 depicts the end strength formula. USMC manpower planners begin with the current strength of the force, subtract the losses that occur, add the gains, and conclude

with an overall force end strength. This process is conducted monthly and subdivided between officers and enlisted personnel (J. Cruz, personal communication, August 23, 2022). One difficultly that manpower planners face is accurately predicting the institutions' enlisted losses, especially when they occur outside of contract. Loss prediction has direct impacts on the number of accessions needed, which in turn, affects the institution's ability to meet end strength.

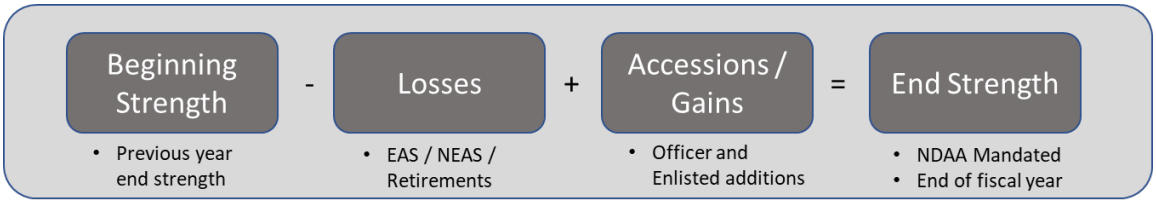


Figure 1. End Strength Defined

The Commandant of the Marine Corps (CMC) has authored strategic level guidance, Talent Management 2030, aimed at modernizing the overall Human Resource Development Process (HRDP) that the organization employs (Berger, 2021). This guidance seeks to reexamine the manpower system in light of the institution's new focus on Great Power Competition and the changing expectations and views of human capital within the United States (Berger, 2021). This document also calls for an improvement of the current manpower planning models. The CMC states that "our manpower model was devised in an era before personal computers, mobile phones, and the Internet, when Marines received paper orders and paper paychecks" (Berger, 2021).

The USMC maintains a robust database of personnel information within the Total Force Data Warehouse (TFDW), and other databases, that can support more advanced manpower and attrition modeling. The breadth of this data and the lack of employment of advanced modeling makes it difficult for manpower planners to meet the CMC's modernization intent. My thesis leverages existing data coupled with emerging machine learning (ML) modeling techniques to attempt to meet the CMC's intent in one small area of research.

B. INSTITUTIONAL BACKGROUND

The primary institution responsible to the CMC for manpower issues within the USMC is the Deputy Commandant (DC), M&RA. DC, M&RA manages "the current inventory of Marines, builds plans for the creation and distribution of future inventory, and assigns available inventory against billets manned" (MCO 5250.1). Within M&RA, the Manpower Plans and Policy (MPP) division is responsible to DC M&RA for the future inventory of the force to include "manpower plans for force development and implementation of HRDP initiatives, provides manpower policy support, and prepares manpower budget estimates and justifications" (MCO 5250.1). MPP is further subdivided. MPP 20, Enlisted Plans Section, is the department primarily responsible for developing the enlisted accession plans for each fiscal year. A component of this accession planning is forecasting enlisted attrition.

Attrition remains a cyclical and necessary component of the USMC's HRDP. Within the institution, attrition possesses a certain degree of predictability and seasonality. The inflow of new Marines into the organization is necessary due to the current force structure requirements; grade shaping requires a large base of junior enlisted Marines from which to promote. Simply stated, the Marine Corps has the most need at the junior enlisted levels (E1-E3). These population levels also comprise most of the enlisted attrition. Attrition that occurs prior to the contractual end of active service (EAS), is considered especially harmful to the institution as it creates fiscal costs, readiness costs, and personnel costs.

Manpower Planners at M&RA describe this type of loss as non-EAS (NEAS) attrition (J. Cruz, personal communication, August 23, 2022). They have categorized NEAS attrition into three areas. The first is Base NEAS, the second is Recruit NEAS, and the final is Retirements. Base NEAS attrition represents all attrition that is not EAS departures, attrition of Recruits at the two enlisted Recruit Depots, or attrition due to retirements. Base NEAS attrition, as defined, represents Marines who have already completed a portion of their training but depart the service prior to their contractual obligation. Planners divide enlisted attrition in this manner because it allows them to better model and forecast these rates. Additionally, understanding these numbers in their

appropriate context aids in policy and process improvement. Base NEAS attrition is the primary subject of this thesis.

Base NEAS attrition for enlisted Marines is currently tabulated by the Enlisted End Strength Planners residing within MPP 20. The current forecasting model for Base NEAS attrition relies heavily on manual inputs from the planners. The processes involves manually moving spreadsheets and data across multiple platforms and between multiple people (J. Cruz, personal communication, August 23, 2022). The computation of the specific projections are done manually using spreadsheet modeling.

The accuracy of the current model requires a relatively stable composition and disposition of the force (J. Cruz, personal communication, August 23, 2022). The process amounts to compiling a three-year trailing average of the ratio of monthly Base NEAS attrition to monthly end strength goals (J. Cruz, personal communication, August 23, 2022). The planners forecast the Base NEAS for subsequent months and provide those predictions to the recruiting apparatus of the institution. Variations of this method have existed as far back as 2008, as evidenced by a previous NPS thesis on the same subject authored by Sanford Orrick. Each subsequent planner has approached the modeling technique with slight variations but have ultimately used the same trailing average technique.

The planner's goal is to achieve an accurate forecast to meet the legally mandated enlisted end strength numbers by the close of the fiscal year. This is best achieved by accurately forecasting losses at the monthly interval. This allows for incremental adjustments to be made on both the accession and retention sides of the institution. Accurate prediction of Base NEAS attrition is especially useful to the institution's Enlisted End Strength Planners because it represents the attrition variable that is the costliest to the institution.

Historical data shows that retirement and recruit attrition exhibit generally predictable behavior. Additionally, the institution has limited control over retirements. This control can allow the institution the ability to influence its losses by artificially stabilizing them through management of who is allowed to retire and when they are allowed to retire.

Within the last 5 years, roughly 20% of NEAS attrition is credited to retirements (R. Johnson, personal communication, June 8, 2022). Recruit attrition tends to be predicable in nature as well. It accounts for roughly 30% of NEAS attrition over the last five years.

C. RESEARCH QUESTIONS

The following research questions are proposed.

1. Primary Research Questions

a. Can a manpower model be developed using ML that predicts monthly Base NEAS attrition using individual Marine data contained in existing USMC personnel databases?

In this thesis, I successfully develop multiple models using two ML techniques that predict Base NEAS attrition at the individual Marine level.

b. Are ML modeling approaches using individual data better at predicting aggregate Base NEAS attrition compared to current models?

The ML models that I develop often outperform the current modeling techniques. Across the entirety of FY20, they dramatically improve upon the current process underestimating Base NEAS attrition by 4% compared to the current process which underestimates by 12%. Additionally, the cumulative error observed at each month within the FY remains more stable with the ML models.

2. Secondary Research Questions

a. Does using disaggregated data and an ML model better account for various shocks to the USMC manpower system?

The use of disaggregated data and prediction at the individual Marine level enables the ML models to account for individual differences in Marines. It achieves greater success than current processes at accounting for shocks within the manpower system. This thesis posits that the individual differences in Marines can be exploited by the ML algorithms employed.

b. Is the predictive ML model developed feasible for Marine planners to implement given their current systems, software, and programs?

Further research is needed to determine the feasibility of employing the ML modeling techniques from this thesis. Current USMC data systems and collected data increase the opportunity costs associated with employing these techniques.

D. SCOPE

This thesis focuses on the enlisted population of Marines. It uses currently collected data from FY10–FY21 to construct a series of monthly models that predicts whether or not a Marine departs the service, with a given month, before their contractual end of active service (EAS). This type of attrition is only a small subset of total losses within the institution, as will be discussed in more detail. I compare the output of these more advanced models to the currently employed models and examine the benefits and shortfalls of both.

E. THESIS ORGANIZATION

This thesis contains six chapters. Chapter I contains the introduction and institutional background on the currently employed enlisted end strength modeling processes. Chapter II contains the literature review of four relevant studies along with a small meta-analysis of civilian and military personnel attrition. Chapter III outlines the data and methodologies I employ for this study. It also includes further details on data cleaning and the specific models chosen. Chapter IV is an analysis of results of the methodologies I employ. Chapter V contains recommendations and topics for future research. Finally, Chapter VI contains the appendices consisting of useful summary statistics and additional model information.

II. LITERATURE REVIEW

This chapter reviews previous attrition prediction literature from both the military and civilian sectors. The studies selected for this thesis were chosen because of the differing analytical techniques employed including traditional regression and more modern ML. These techniques are relevant because end strength planners can employ them within the USMC HRDP. This literature review is divided into three categories. First, traditional attrition modeling techniques that employ older regression and probability modeling. Second, modern attrition modeling techniques that employ compute intensive ML. And third, outcomes and explanatory factors that influence attrition that are consistent throughout the reviewed literature.

A. CONVENTIONAL ATTRITION PREDICTION

In his Rand study on military attrition behavior, Richard Budding (1984) employs four service specific multivariate logistic regression models designed to deduce the significance of collected predictors associated with early attrition. His study represents one of the earlier attempts at quantifying and identifying the differences between enlisted military members and their civilian counterparts. He finds that an individual's work history prior to enlistment bares significant weight on their likelihood of attrition, increasing specifically in those service members nineteen years of age having more than four previous employers (Buddin, 1984). He also finds that non-high school graduates and older services members are more prone to early attrition. His study represents an early attempt at identifying problematic military attrition at the individual service member level.

Sanford Orrick (2008) examines the Marine Corps' NEAS enlisted attrition with the goal of designing a model using existing, in service, observational data to predict NEAS attrition. At the time of Orrick's study, the Marine Corps was employing two methods of modeling NEAS attrition. First, a steady state weighted average approach of past monthly attrition numbers, and second, a Monte Carlo simulation that also incorporated weighted averages of attrition numbers. These models are very similar to what Marine Corps manpower planners are currently employing.

Orrick employs logistic regression on USMC personnel data from TFDW, spanning from 1998 – 2008. He designed the outcome variable as a Marine being either a NEAS or EAS attrite (Orrick, 2008). His model uses aggregated data from 1998 to 2004 to predict the probability of NEAS attrition on 2005 data. This framework was replicated for the years 2006, 2007, and 2008 using the same training and validation method. The cross validation on the 2005 data yielded an overall accuracy rate of 76.18%, which includes accurately predicting NEAS attrition and EAS attrition (this is based upon the construction of the outcome variable NEAS and EAS as the outcome options). Orrick selected a 0.5 classifier that identified a Marine as NEAS attrition if the predicted value was greater than 0.5 and an EAS attrition if the predicted value was less than 0.5 (Orrick, 2008). His model performed better at identifying specificity, or correctly classifying Marines who would execute normal EAS attrition, achieving 85.25% accuracy (Orrick, 2008). This is compared to a sensitivity rate, or correctly assessing a Marine as a NEAS loss, of 61.11%. Statistically, this is only marginally better than a random guess, which would yield a 50% probability of either outcome.

Orrick's study has several limitations. First, his model solely focuses on predicting EAS or NEAS on those Marines who had already separated and not on the total inventory within each year or month. His data also suffered severely from duplicate entries and from missing separation codes. His cleaned dataset was reduced by 40%, from over 500,000 to approximately 300,000 observations. His study also used aggregated results for prediction. It is likely that some of the seasonality within USMC attrition was lost in employing this technique. Additionally, predicting at the annual interval is less useful to end strength and recruiting planners because of the monthly operating cycle they employ. Finally, the predictors selected did not exhaust all possible options that exists within TFDW.

Another study that employs similar methods is authored by James Marrone (2020). He develops a predictive model for attrition within the first 36 months of a new military member's career. Using pre-accession data, he identifies predictors of first term attrition with the goal of constructing a model that can indicate attrition risk prior to a service members entry into the institution.

This research focuses on comparison of recruit characteristics at accession, probability of attrition using probit regression modeling, and the marginal effect of specific variables that his regression identified as significant. Additionally, he grouped variables together to attempt to extract any additional significance from their interaction. His model proved to be marginally effective at predicting recruit attrition, ultimately achieving a 55–60% sensitivity rate, or properly identifying service members who attrite before the expiration of their first 36 months.

Marrone's research consists of aggregated observations across all four services. This is noteworthy because Marrone's probit model was trained on the complete dataset and then applied individually to each service. This may have led to improper characterization of attrition within each service, given that each service has its own unique characteristics and personality traits of its recruits. Additionally, Marrone's study only employed a probit regression model. Other modern ML techniques can be applied, such as predictor selection, before modeling to help draw out statistical significance. Finally, the model failed to incorporate cross validation, or a test-train split. This is crucial to ensuring that a model can accurately predict future events and helps guard against overfitting.

B. NASCENT ATTRITION PREDICTION

Patrick Gallagher (2020) examines the use of ML techniques to predict retention among enlisted Marines. The goal of his research is twofold. First, he employs ML techniques to predict retention, and second, he seeks to determine if individually targeted bonuses for enlisted Marines could prove successful at increasing retention. His study uses data prior to the reenlistment decision to predict retention, in his case, reenlistment.

Gallaher employs nine different ML and regression techniques on his data to determine the best model for predicting reenlistment, representing a form of retention. He evaluates his results using the F1 score, a common ML evaluation metric that employs the harmonic mean of precision, the total percentage of correct predictions, and recall, the actual positives that were correctly identified (Gallagher, 2020).

Gallagher finds that the Random Forest classifier, logistic regression, and Naïve Bayes algorithms are best at predicting reenlistment with F1 scores of 0.686, 0.672, and

0.668 respectively. His modeling is unable to facilitate targeting Marines at the individual level, however, application of his models prove to be beneficial when basic cost analysis of retention is performed. If used, his models could provide cost savings by identifying Marines that would reenlist regardless of bonus incentive; the institution would not need to offer a bonus to secure their reenlistment.

Attrition prediction in manpower systems is not limited to the DOD. Private firms and organizations employ human resources arms to assist with retention and attrition related evaluation. In a recent study by Fallucchi et al. (2020), researchers employ several ML techniques to predict attrition and identify indicators that could be useful for managers to incentivize retention.

The authors conclude that the most effective model is a Gaussian Naïve Bays classifier followed by logistic regression. The associated F1 score for these models is 0.446 and 0.445 respectively. These two techniques have the best performance at true positive identification, or simply stated, correctly predicting the probability of an employee leaving a company (Fallucchi et al., 2020).

Fallucchi's research is distinct from the other literature reviewed because it focuses on non-military personnel data. The variables available to the researchers fall into three distinct bins. First, the dataset includes demographic information, second, the data includes current employment information, and third, it includes environmental factors that influence both the individual and the job, including things like satisfaction derived from the job, interoffice relationships, and workplace environment.

In their paper, Alduayj et al. (2018) use synthetic data to develop a series of predictive ML models designed to identify individual civilian employee attrition. They first employ three separate ML techniques, including a Random Forest classifier, on their unbalanced training dataset (Alduayj & Rajpoot, 2018). They then employ a data balancing technique and run the same set of models. The balanced training data provides significantly better results for all ML algorithms, with an F1 score increase on the Random Forest classifier from 0.269 to 0.921 (Alduayj & Rajpoot, 2018). They also employ feature selection techniques to reduce the size of the predictor dataset. This paper represents a

convergence of computer science and statistical analysis employed on a manpower problem. The paper lacks the formal development of a proper training and validation data split and is employed on synthetically created data. However, the employment of the selected techniques and the logical process used to properly leverage ML contribute to the development of my research.

C. OUTCOMES AND EXPLANATORY FACTORS

Table 1 is a meta-analysis of independent variables that are consistent across the reviewed literature. These predictors are found to have a statistically significant relationship with attrition. Data from Gustavo Terrazas' (Terrazas, 2020) study on implementing ML techniques on USMC reenlistment data and an additional study from Marrone et al. (2021) on U.S. Army attrition, are also analyzed to develop the table below. Similar variables between civilian and military datasets are combined together within the table, such as time in service and occupational tenure. It is important to note that results from ML techniques do not always indicate the direction of the relationship between the feature and the outcome, only that the predictor is significant. In these cases, the notes section will indicate ML.

Table 1. Meta-analysis of Reviewed Literature on Attrition Related Outcomes

Metanalysis of Predictors: Relationship with Attrition			
	Relationship	**Notes**	**Study**
Age	Negative	- Negative as age increases - Most positive at 26–30 years old	Marrone Fallucchi
Gender	Positive	- Positive for women	Orrick Marrone
Race	Negative	- Negative for non-whites	Orrick Marrone
Marital Status	Negative	- Negative for married - Positive for widowed	Orrick Marrone

Metanalysis of Predictors: Relationship with Attrition			
Citizenship	Varies	- Negative when non-U.S. Citizen	Marrone
Dependents	Negative	- More than one dependent is negative - Negative with children	Orrick Marrone
Education	Negative	- Negative as education increases	Orrick Marrone
Rank/Income	Negative	- Significant Feature - Negative as rank increases	Marrone Fallucchi
Commute Distance	Varies	- Significant Feature	Fallucchi
Overtime	Varies	- Significant Feature - More overtime, higher attrition	Fallucchi
AFQT	Negative	- Higher scores, lower attrition - Positive as scores decrease	Orrick Marrone
MOS Category	Varies	- Positive towards sales jobs - Positive toward combat jobs	Marrone Fallucchi Terrazas
Unit Type	Varies	- Positive toward combat units	Marrone
Bonus	Negative	- Negative	Marrone
Contract Length	Positive	- Positive as term length increases	Orrick Marrone
Duty Location	Varies	- Varies and determined by MOS	Marrone
Leadership	Negative	- Negative towards good leadership	Marrone
Unit Culture	Significant	- Significant Feature (ML)	Marrone Terrazas Fallucchi
Accession Month	Varies	- Less likely at beginning of quarters (Jan, Apr, Jul, Aug)	Marrone
Prior Active Duty	Positive	- Positive	Marrone
ENTNAC	Positive	- Positive toward waivers	Marrone

Metanalysis of Predictors: Relationship with Attrition			
DEP	Negative	- Negative if time spent in DEP	Marrone
Combat Tour	Negative	- Significant for Reenlistment outcome (ML)	Orrick Terrazas
Deployment	Significant	- Significant for Reenlistment outcome (ML)	Terrazas
Time in Service	Negative	- Positive for > 4 years TIS Negative for those with >11yrs	Orrick Fallucchi
Pro/Con Composite Score	Significant	- Significant for Reenlistment outcome (ML)	Terrazas
Legal	Significant	- Significant with Reenlistment (ML)	Terrazas

The literature shows a positive relationship between attrition and gender (women), marital status of widowed, lower AFQT scores, combat designated units, contract length (increasing as length increases), prior Active-Duty status, time in service, and receipt of an entrance national agency check (ENTAC) waiver. Intuitively, these predictors' positive relationship with attrition make sense and are consistent with common understanding and theory. The literature in Table 1 also shows a negative relationship between attrition and numerous predictors, most notably, marital status, having dependents, receiving a bonus, having good unit leadership, and time spent in a delayed entry processing (DEP) status. These predictors are also congruent with theory and common understanding of human behavior. Finally, the ML algorithms used in the reviewed literature indicate that deployment, proficiency and conduct score, legal action, unit culture, commute distance, and overtime worked, are significant predictors. While these predictors are not given any directional relationship with attrition, theory and literature agrees with their selection.

D. CONCLUSION

My research benefits from the reviewed literature in three distinct areas. First, my models have a deliberately constructed training dataset broken down at the monthly level. Second, I employ traditional multinomial logistic regression techniques, and two more modern ML techniques, least absolute shrinkage and selection operator (LASSO) and Random Forest classifier. Finally, this thesis employs feature engineering and pulls in a large set of predictors, many of which are indicated as significant in the reviewed literature.

III. DATA AND METHODOLOGY

A. DATA SOURCES

The data for this study is drawn from various data tables contained in the USMC TFDW and other data provided by Manpower Information Systems Branch. The study population includes all enlisted Marines in the active duty component in service from October 2009 through October 2021, corresponding to FY10 to FY21. The data is longitudinal, where an individual Marine is observed over this period and followed every month until that Marine attrites or leaves active duty service. The primary data cleaning program I use is Stata 17 MP and the primary modeling program I use is R 4.2.2.

B. DATA CLEANING

1. TFDW Data Structure

TFDW operates the primary Sequel database for historical data for the USMC. It stores data based upon monthly snapshots of the total force. These snapshots are referred to as sequence numbers (J. Forbes, personal communication, July 26, 2022). There are multiple data tables within each sequence number that store specific types of information based upon their category. This structure allows users to access specific historical information without having to access the entire database. Additionally, it allows researchers the ability to construct longitudinal datasets for detailed analysis.

2. Data Preparation

The data extracts from TFDW for this study are Microsoft Excel .CSV files of nine distinct data tables. The process of data cleaning and merging these tables is best illustrated using a hub and spoke structure. The primary data hub captures the total inventory or population of Marines in that month or sequence number, coming from two TFDW data tables that describe general characteristics of each Marine in service during that given sequence number or month. All other data tables are merged to this hub table using a combination of unique identification numbers for individual Marines and the associated sequence number. Figure 2 depicts the data table merge process.

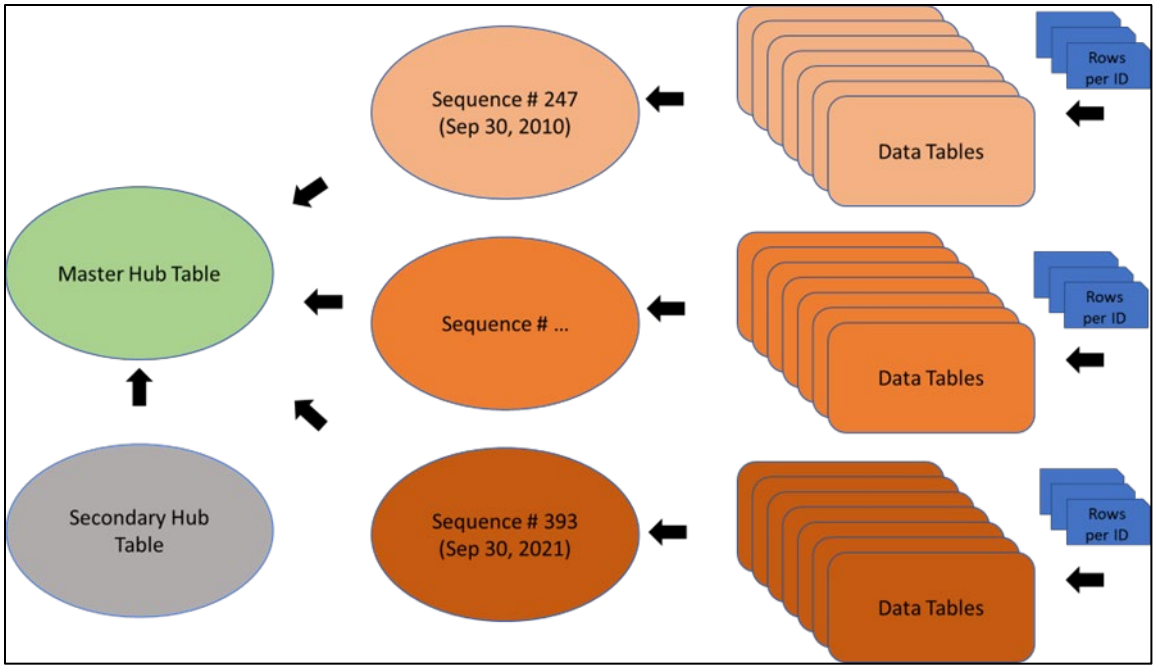


Figure 2. Data Tables Merge Process

The specific data tables merged to the hub contain useful attributes that prior literature has shown to have a relationship with attrition. Table 2 contains a list and description of these data tables. Most of these data tables are entered per transaction rather than longitudinal; that is, individual Marines are not followed over time but data are entered when an event or transaction, such as a Legal Action, occurs. In some instances, multiple entries can occur in a given month or sequence number, such as when a Marine acquires more than one Award in a month. Thus, data within each of the spoke tables are first collapsed to just one row per sequence number for each unique individual Marine, prior to merging with the hub. An example of this is Physical Fitness Test (PFT) scores. When a Marine has multiple PFT entries performed in that sequence number, I kept the maximum PFT score a Marine achieves in that sequence number or month. Other similar instances of multiple entries use the most recent non-missing value or date. The data tables also contain a significant amount of overlap, and I kept the variable from the table with the most fidelity. Finally, I drop predictors that are not relevant to the research.

Table 2. Merged Data Tables

Data Table	Description
Hub	Contains all basic information about each Marine in inventory
Separations / Losses	Contains all separations / losses from the USMC inventory
Awards	Contains all types of awards given to individual Marines
Legal Action	Contains all judicial and non-judicial actions for individual Marines
PFT	Contains all Physical Fitness Test (PFT) scores
MCMAP	Contains all Marine Corps Martial Arts Program (MCMAP) belts achieved
FitRep	Contains all Fitness Report data
Accessions	Contains all data on Marines collected at time of accession

Numerous predictors contain missing values. This is not uncommon in USMC longitudinal datasets. For example, if a Marine is an E4 or below, they will not receive a Fitness Report (FitRep). In this instance, that Marine would have a missing value in the FitRep Cumulative Value variable. The ML methods I employ in this thesis require a complete dataset with no missing values. To overcome this challenge, I create indicator predictors for each predictor that has more than one missing value. This indicator is binary with 0 equating to no missing data and 1 equating to missing data. This method allows the algorithm to quantify the presence or absence of a value in the predictor in question. After I create indicators for missing data, all missing values are replaced with zeros and no appreciable change in distributions are noted. In very few cases, values are imputed from adjacent or closely related predictors. In its entirety, the merged raw dataset includes over 22 million individual Marine-month observations and over 700 predictors.

The predictor formats within the data are numeric, continuous, categorical, and calendar date. These require little preparation prior to executing the selected modeling techniques. In certain cases, the date predictors are required to be in numeric form, ascending from zero with an origin date of January 1, 1970. The primary approach I use is to compute dates as the number of days between the predictor calendar date and the sequence number date equivalent. This gives unique values of the count of days between each date predictor and each sequence number. Categorical predictors that contain levels with zero as a value are consolidated to remove the level with no value or data. I also remove predictors contained in the data that have zero variance in their values as these will not impact the outcome variable.

3. Response Variable

The primary goal of this thesis is to predict aggregate Base NEAS attrition. Upon separation from the USMC, each Marine is given a specific separation code corresponding to the type of separation they receive. TFDW maintains a lookup table to correspond with these separation codes. This lookup table consists of 708 unique separation codes. Broadly, NEAS contains 322 different separation codes. I further narrow this list of separation codes down to clearly define Base NEAS attrition, which is separate from attrition at Recruit Training and attrition due to Retirement. I identify Recruit Training attrition using the Recruit Depot present monitored command code (PMCC) and the Recruit assignment primary military occupational specialty (PMOS). This ensures that only recruits who attrite while assigned to the two Recruit Depots are removed, and not permanent personnel assigned to the Recruit Depot parent units. Separately, I remove Retirements using specific separation codes. This allows me to create a cleanly defined Base NEAS response variable that is the closest possible replication to how End Strength Planners at M&RA define Base NEAS. I also assign these categories based off current Fleet Marine Force and M&RA processes, streamlining where appropriate (C. Dowling, personal communication, October 12, 2022). For example, a Marine who deserts his unit and is separated is included in Base NEAS attrition.

Over 75% of the separations data entered in TFDW appears in the sequence number after a Marine has separated. This is due to the time of the month the database sequence number snapshots are taken. For example, if a Marine separates in sequence number 350, their separation could be recorded in sequence number 351. Because of this, I merge separations data that spans from sequence 247–394 (FY10–FY21 plus October FY22), even though my primary hub dataset only spans from sequence number 247–393. The additional sequence number captured by the separations table allows me to capture the separations that occur in the last sequence number of the hub dataset, sequence number 393.

Thus, there are three possible states or outcomes for Marines within this data. First, they can remain in service during the course of the specific sequence number under observation, known as continue service. In this case, they do not have a separation code in that month. For example, if a Marine does not receive a separation code in the month of October, they are classified as continue service for that month. Second, they can attrite via normal EAS within the sequence number or month under observation, which I identify based upon the set of associated separation codes. Finally, they can attrite via Base NEAS within the sequence number or month under observation, based upon the set of associated separation codes discussed above.

Thus, the final composition of my response variable is categorical and looks at a Marines state only within the specific month or sequence number under observation. It takes on three forms: Continue Service, Normal EAS, and Base NEAS. This design facilitates the multinomial logistic regression and the multiclass Random Forest classifier models with the goal of giving the models more options for classification. It properly isolates each of the three states described above. The underlying logic behind this approach is that predictors and indicators between the various subcategories are different. Figure 3 depicts the categorical outcome variable.

$$Outcome = \begin{cases} 0 \ if \ Marine \ Continued \ Service \\ 1 \ if \ Marine \ EAS \ normally \\ 2 \ if \ Marine \ NEAS \end{cases}$$

Figure 3. Outcome Variable Defined

C. DESCRIPTIVE STATISTICS

1. Predictors

The cleaned dataset for each model consists of over 470 predictors used for determining the classification of each Marine. These predictors fit into five broad categories. First, pre-service data represents information pulled while an individual undergoes the recruitment and screening process. Second, personal attributes data describes individual background characteristics that do not frequently change. Third, performance data depicts all aspects of a Marine's performance while in service. Fourth, date data captures all time related events. Fifth, separations data captures attributes of a Marine at the time of separation from service. Table 3 lists example predictors from each of the five broad data categories. A full list of summary statistics for the training and testing data for the October model can be found in the appendix.

Table 3. Predictor Categories and Examples

Pre-service	Personal Attributes	Performance	Dates	Separations
Home of Record	Education	FitRep Cumulative Value	Armed Forces Active-Duty Base Date	Separations Code
Source of Entry Code	CONUS	Proficiency Score	End of Active Service	-
AFQT Score	Ethnicity	Conduct Score	Promotion Restriction date	-
Recruiter Rank	Number of Dependents	Legal Action	Date of Birth	-
Medical Disposition	Deployments	PFT	Extension Date	-
Ship To	MOS Category	Awards	Crisis Participation Date	-

D. METHODOLOGY

Predictive modeling involves assigning a classification or probability to an outcome variable based upon observed attributes found in the predictors. The two modeling methods I estimate are multinomial logistic regression and Random Forest classifier. As with ML algorithms, how well the models predict the outcomes is validated by splitting the data into training and testing data.

1. Model Design

Both modeling techniques I employ revolve around a calendar month dataset. To enable comparison with the current trailing average model, I compile monthly datasets comprising of that specific month over a three-year historical period. For example, the October model training data is comprised of data from October 2017, October 2018, and October 2019. Each monthly dataset consists of roughly 450,000 observations. This structure mimics the current trailing average model, in that the average Base NEAS attrition from the last three years for that month is generally used to predict current FY month's attrition. Additionally, this structure allows for faster data training and computation. Figure 4 is a graphical depiction of the model design using October as a representative example.

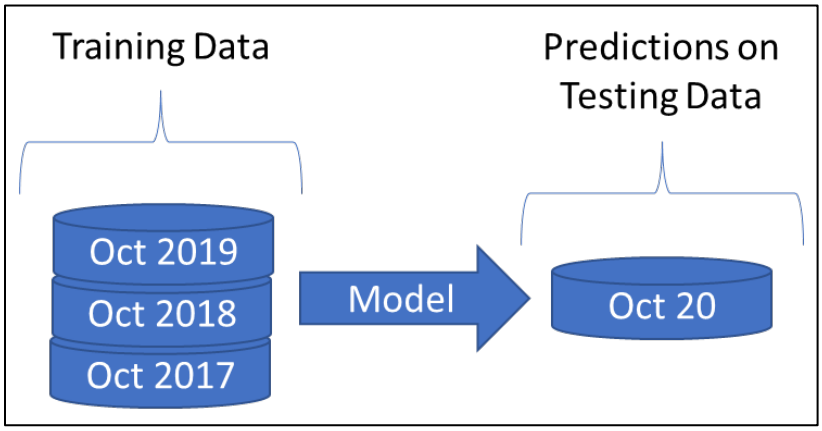


Figure 4. Model Design

2. Data Imbalance

The nature of this prediction problem results in an imbalanced dataset. This occurs because in a given month, there are far more Marines who continue service than Marines who attrite due to EAS or NEAS. This poses a difficult problem for the classification algorithms and is not easily solved.

The balancing procedure I use is Synthetic Minority Oversampling Technique (SMOTE). This technique creates additional data of the minority class, in this case, Base NEAS attrition, and adds it to the training dataset. It creates the additional minority class by using elements of the existing minority class data, employing randomly chosen k-nearest neighbors technique (Chawla et al., 2002). Table 4 and Table 5 depict the unbalanced and balanced training data for each monthly model. This balancing technique is used to better train the Random Forest classifier algorithm by exposing it to additional cases of the minority classes, EAS, and Base NEAS. The testing data is not subjected to this data balancing technique and is depicted in Table 6. Figure 5 shows the response variable for October FY17-19 training data before SMOTE was applied. Figure 6 shows after SMOTE is applied. The multinomial logistic regression models are not trained using SMOTE.

Table 4. Imbalanced Observations Training Data by Classification (Response Variable)

Month	Continue Service	EAS	NEAS	Predictors
October (FY17-19)	419,603	4,185	1,098	471
November (FY17-19)	419,386	4,216	922	466
December (FY17-19)	416,425	5,725	976	465
January (FY17-19)	421,144	4,356	1,021	466
February (FY17-19)	420,944	4,477	1,136	465
March (FY17-19)	424,538	3,851	1,225	464
April (FY17-19)	422,472	4,268	1,232	465
May (FY17-19)	418,576	8,569	1,045	465
June (FY17-19)	418,733	7,894	1,168	466
July (FY17-19)	418,020	8,252	1,122	464
August (FY17-19)	416,596	8,458	979	465
September (FY17-19)	416,913	5,386	1,183	466

Table 5. Balanced Observations Training Data by Classification (Response Variable)

Month	Continue Service	EAS	NEAS	Predictors
October (FY17-19)	419,603	209,801	209,801	471
November (FY17-19)	419,386	209,693	209,693	466
December (FY17-19)	416,425	208,212	208,212	465
January (FY17-19)	421,144	210,572	210,572	466
February (FY17-19)	420,944	210,472	210,472	465
March (FY17-19)	424,538	212,269	212,269	464
April (FY17-19)	422,472	211,236	211,236	465
May (FY17-19)	418,576	209,288	209,288	465
June (FY17-19)	418,733	209,366	209,366	466
July (FY17-19)	418,020	209,010	209,010	464
August (FY17-19)	416,596	208,298	208,298	465
September (FY17-19)	416,913	208,456	208,456	466

Table 6. Testing Data Metrics, Observations, and Predictor Count (Response Variable)

Month	Continue Service	EAS	NEAS	Predictors
October (FY20)	140,378	1,352	326	471
November (FY20)	141,639	1,653	372	466
December (FY20)	139,034	2,076	314	465
January (FY20)	142,457	1,696	397	466
February (FY20)	141,542	1,276	493	465
March (FY20)	142,375	1,159	544	464
April (FY20)	142,500	1,419	540	465
May (FY20)	138,412	2,642	581	465
June (FY20)	135,050	2,274	408	466
July (FY20)	132,659	3,137	411	464
August (FY20)	132,015	2,759	344	465
September (FY20)	133,641	1,910	257	466

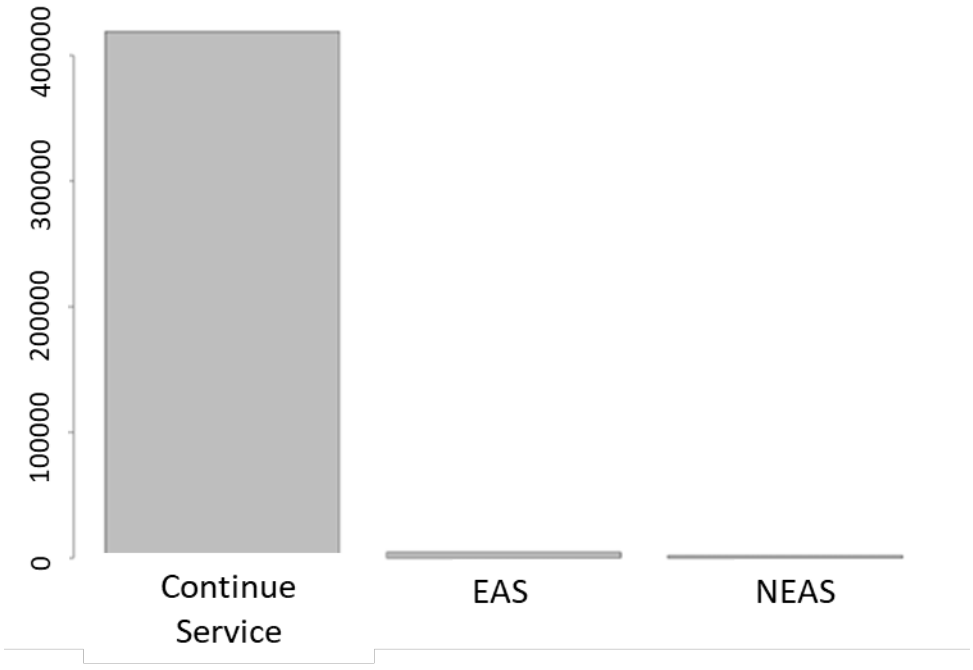


Figure 5. Pre-SMOTE Response Variable (October FY17–19)

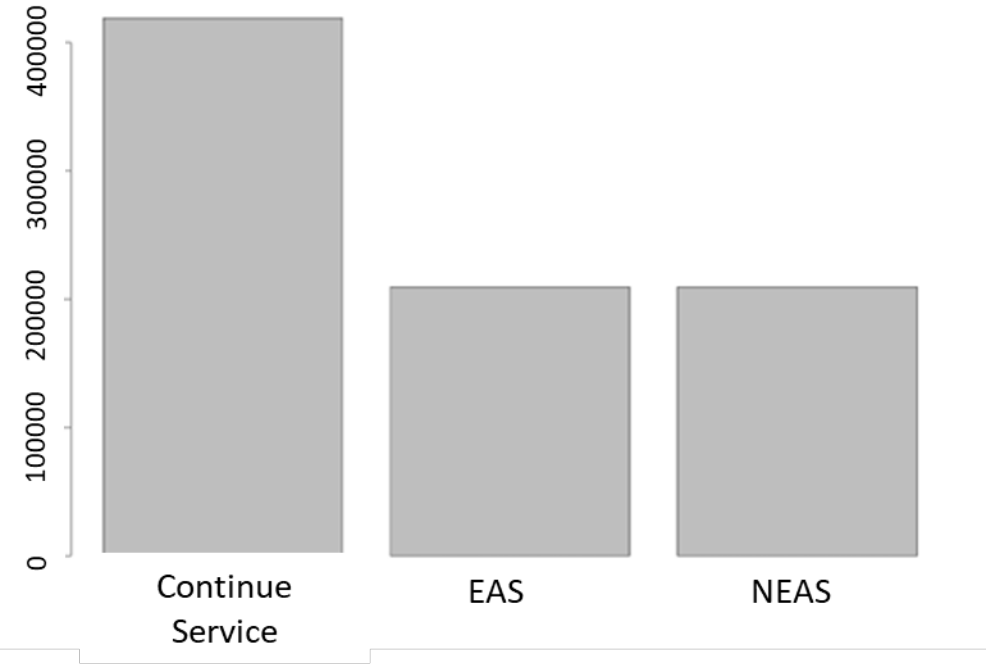


Figure 6. Post-SMOTE Response Variable (October FY17–19)

3. M&RA Trailing Average Replication

I recreate the current Base NEAS trailing average model in order to compare it to the ML techniques I develop. I average the historical Base NEAS attrition for FY17-19, monthly, and project that value for FY20 monthly Base NEAS attrition (J. Cruz, personal communication, August 23, 2022). The difference between the trailing average projection and the actual observed values makes up the accuracy of the model. Table 7 depicts this process. The specific trailing average approach that M&RA has employed has varied slightly depending on the staff's preference but ultimately, it has revolved around this basic trailing average since 2008 (Orrick, 2008). In addition to employing this model, I also replicate this trailing average model on all training data to further lend credibility to how I define my response variable and build my model.

Table 7. Base NEAS Model Replication

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
FY17 NEAS-Base	282	500	307	432	439	503	356	539	463	348	458	328
FY18 NEAS-Base	465	394	365	413	405	444	561	509	449	549	540	316
FY19 NEAS-Base	500	376	419	376	398	348	515	432	339	537	374	517
3 Year Average	416	423	364	407	414	432	477	493	417	478	457	387

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
FY20 NEAS-Base	509	368	413	316	481	533	625	610	653	498	465	377

Variation between how I define Base NEAS and how current End Strength Planners define it does exist, but it is minimal and due in part to the changing nature of the USMC's separation codes. Certain codes that are used in earlier instances of the data, are not seen in current data. This creates a difficult problem, one that I solve by cataloging separations according to the separations data table definitions, the Marine Corps Separations manual, and by consulting with subject matter experts within the Fleet Marine Force. A specific example of this is separation due to homosexual misconduct. This is not currently considered a separation offense but within certain times in my dataset, it is. Table 8 depicts the trailing averages pulled from M&RA and the trailing averages I compute from my training data.

Table 8. FY20 NEAS M&RA and Testing Data Variation

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
M&RA NEAS Data	509	368	413	316	481	533	625	610	653	498	465	377
Testing Data NEAS	326	372	314	397	493	544	540	581	408	411	344	257

4. Managerial Relevance

Statistician George Box is credited with the adage that all models are wrong, some are just useful. In order to be useful, end strength planning models have to meet certain success thresholds. Managerial relevance refers to the idea that a model can be realistically implemented by end strength planners. The complexity of data collection, computation, and interpretation must be considered as factors when determining this. For this thesis, managerial relevance is defined by the current Enlisted End Strength Planner's process. If the currently employed processes can be replicated or improved within reasonable measures, managerial relevance is achieved. This supports the secondary research question.

5. Prediction versus Causality

The objective of this thesis is the prediction of Base NEAS attrition within the USMC enlisted population. Since the desired end state is predictive in nature, I do not attempt a causal interpretation of any predictor's effect on Base NEAS attrition. I seek to develop the most accurate model possible and identify predictors that correlate significantly to the response variable, Base NEAS attrition. I use multiple individual characteristics as predictors with the only goal of increasing the likelihood of correct predictions. The specific methodologies I discuss in the follow-on paragraphs are chosen to support the desired predictive end state of the model. Additionally, interpretation of marginal effects of specific predictors are not always possible with certain ML techniques that I employ.

6. Logistic Regression

Broadly defined, logistic regressions also estimate a probability of an outcome that can be interpreted as a classification. The primary advantage of this model is that the fitted values, or predictions, are constrained between 0% and 100% (Lemeshow et al., 2013). Additionally, the non-constant partial effects that the model allows for make it more effective at predicting non-linear outcomes. This allows for the model to perform better at predicting individual differences between individual Marines. Figure 7 depicts the basic logistic regression formula (Lemeshow et al., 2013).

$$prob(y_i = j) = \frac{e^{b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}}}{1 + e^{b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}}}$$

Figure 7. Logistic Regression Formula

Multinomial logistic regression is identical to binary logistic regression but with a categorical response variable. The interpretation of the outcome in a multinomial logistic regression is relative to one of the levels, in the case of this thesis, the base level is set to continue service. The classification threshold for the multinomial logistic regression I employ is the greatest probability. This means that that the response variable class with the largest probability becomes the observation classification. Figure 8 depicts the multinomial logistic regression formula I employ (Lemeshow et al., 2013).

$$prob(y_i = j) = \frac{e^{b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}}}{1 + e^{b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}}}$$

$$j = \begin{cases} 0 \ \ if\ Marine\ Continued\ Service \\ 1 \ \ if\ Marine\ EAS\ normally \\ 2 \ \ if\ Marine\ NEAS \end{cases}$$

Figure 8. Multinomial Logistic Regression Formula

7. LASSO Machine Learning

In this thesis, I also employ the LASSO ML technique (Tibshirani, 1996). This technique allows for a plethora of predictors to be added into the initial stages of model creation. The LASSO algorithm derives the most significant predictors to the desired outcome, in this case, Base NEAS attrition. The LASSO technique also helps to avoid incorrect correlations or overfitting of the predictors to the outcome by penalizing the addition of uncorrelated predictors (Tibshirani, 1996). I employ the LASSO technique with the primary goal of predictor selection for the multinomial logistic regression model.

8. Random Forrest Classifier

The Random Forrest classifier is a supervised ML algorithm that operates using multiple decision trees. The algorithm compiles numerous and randomly selected subsets of the data, known as bagging, sampling with or without replacement, and then compiles a classification using decision trees (Breiman, 2001). The majority classification decision is then selected as the observation prediction. Random Forest classifiers have numerous benefits, including protection from overfitting, due to the random process associated with bagging, and relatively little required data preparation (Wright & Ziegler, 2015).

Random Forrest classifier algorithms require hyperparametric tuning to maximize their accuracy. The primary parameters that can be tuned are the number of randomly drawn candidate variables (Mtry), the number of trees grown, and the minimum node size (MinN) (Probst et al., 2019). I optimize each model by systematically changing Mtry, which the literature shows is the most influential parameter in a Random Forest classifier (Probst et al., 2019). Table 9 depicts the optimized hyperparameters per model.

Table 9. Optimized Hyperparameter Values

Model	Tree Count	Mtry	MinN	Variable Importance Mode	Split Rule
October	500	350	5	Impurity	Gini
November	500	385	5	Impurity	Gini
December	500	385	5	Impurity	Gini
January	500	385	5	Impurity	Gini
February	500	165	5	Impurity	Gini
March	500	375	5	Impurity	Gini
April	500	385	5	Impurity	Gini
May	500	350	5	Impurity	Gini
June	500	375	5	Impurity	Gini
July	500	350	5	Impurity	Gini
August	500	325	5	Impurity	Gini
September	500	300	5	Impurity	Gini

9. Test Train Split

I employ a separate testing dataset using a separate year of data in order to ensure that models are not overfit to the training data. The training data consists of 75% of the total data and represents three monthly snapshots of the month being modeled using data from FY17–FY19, as shown in Figure 1. The testing data equates to 25% of the total data and it is a single snapshot of the month being modeled from FY20 data. This validation data represents completely new data for the model to be run against. These methods ensure that the reported metrics are a true representation of how each algorithm used will perform on new, real-world data.

10. Feature Engineering

To help overcome the lack of specific types of in-service data, I conduct feature engineering to create variables to attempt to deduce statistical significance that otherwise would go unnoticed (Patel, 2021). Example features that I engineer include time to EAS, change in education status (did a Marine receive more education while in service), and change in marital status during service. Creating categorical variables based upon perceived and calculated variance also assists in engineering statistically significant predictors. The step of feature engineering was not specifically noted in any of the reviewed literature but remains an important component as borne out by the results of this thesis.

11. Supervised ML

Supervised ML describes the framework of how a computer algorithm computes data to train a specified model. In this case, a supervised ML model is given the correct outcomes to train it (Marsland, 2009). In my application, this equates to providing the separation status of each Marine or row of data; either continue service, EAS, or Base NEAS. Both the logistic regression and Random Forest classier algorithms are considered supervised ML models.

12. Evaluation Metrics

I report several evaluation metrics to analyze the model results. The most relevant for my primary research question is the model's deviation from the test data. I evaluate this

as the total number of Base NEAS predictions minus the number of actual Base NEAS observations, divided by the observed number of Base NEAS observations, all within the testing dataset. Figure 9 depicts this definition. I predict aggregate Base NEAS attrition monthly, and this simple metric is easily interpreted, evaluated, and understood. The output is understood as the percent difference, plus or minus, between the model and the testing data, with 0% representing a model no deviation, or perfect fit. This definition generates the most utility for Enlisted End Strength Planners at M&RA. I also report conventional metrics found in the confusion matrix, Receiver Operating Characteristic (ROC) curve, area under the ROC curve (AUC), precision and recall, and F1 score (Zheng, 2015). I report the model deviation of the currently employed M&RA trailing average technique using the same method outlined in Figure 9.

$$Model\ Deviation = \frac{Base\ NEAS\ Predictions - Base\ NEAS\ Observations}{Base\ NEAS\ Observations}$$

Figure 9. NEAS Model Deviation

13. High Performance Computing

I employ the Naval Postgraduate School (NPS) supercomputer, Hamming, as my primary computing resource. Hamming is a Linux based, multi-node computing cluster that contains thousands of cores and terabytes of memory. The size of the data and computing complexity of the algorithms I use require these resources. In numerous instances, I employ over 500 gigabytes of high-performance computing (HPC) memory, far beyond the capacity of a normal personal computer. The size of the composited training and validation dataset requires significant memory allocation to clean and manipulate. The HPC also allows for a file batching process, enabling multiple R coding scripts to be run simultaneously. I leverage this capability to reduce computation time by simultaneously running all twelve Random Forest classifier and multinomial logistic regression models at once.

Computing time varied greatly depending on the type of model being run. The finial series of Random Forest classifier models require between six and twelve hours to run. The

multinomial logistic regression models require between two and six hours. These finalized models had to be intentionally limited in scope to enable computation as initial runs using additional optimization techniques and algorithms take over seven days to compute. Further discussion on Hamming and other computing shortfalls is discussed in the limitations section of this thesis.

IV. RESULTS AND ANALYSIS

The results and analysis chapter explores the outputs of the ML models and compares them against the currently employed trailing average Base NEAS models. This section also includes a discussion of the predictors that are found to be significant within each modeling technique. The two ML techniques are also compared against each other using the previously discussed accuracy metrics. The ML techniques are also evaluated against the currently employed modeling technique. The current modeling process cannot be evaluated using all the metrics used for the ML techniques as it does not employ classification at the individual level.

A. RESULTS

Table 10 depicts the results from the optimized Random Forest classifier models. The F1 score, and balanced accuracy employ an averaging approach that reports the aggregate results from each class of the response variable. The F1 score is computed using the harmonic mean of precision and recall (Zheng, 2015). The balanced accuracy is calculated by summing the sensitivity scores, true positives divided by true positives plus false negatives, and dividing them by the total number of response variable classes. All models report a high degree of overall accuracy. However, this is a poor metric for evaluation due to the imbalance of the majority class, continue service. The models perform exceptionally well at classifying continue service, and this skews the overall accuracy results. F1 score and balanced accuracy represent a better metric for comparison and interpretation. All twelve Random Forest classifier models perform in a generally uniform manner. Table 10 depicts the metrics for each model including the AUC broken down by the class of the response variable. These results are also consistent with the F1 score and balanced accuracy metrics. Generally, the models do an excellent job at correctly classifying continue service, and EAS. The models have a more difficult time at correctly predicting Base NEAS returning the lowest scores of all the classes. This is supported by the confusion matrix discussed later in this section.

Table 10. Random Forest Classifier Evaluation Results

Model	Overall Accuracy	Balanced Accuracy	F1 Score	AUC: Continue Service	AUC: EAS	AUC: NEAS
October	0.995	0.804	0.782	0.972	0.749	0.501
November	0.996	0.786	0.789	0.972	0.717	0.559
December	0.996	0.790	0.793	0.970	0.735	0.533
January	0.996	0.810	0.816	0.973	0.754	0.510
February	0.995	0.793	0.794	0.962	0.735	0.515
March	0.992	0.784	0.744	0.953	0.732	0.519
April	0.991	0.793	0.744	0.967	0.732	0.526
May	0.993	0.798	0.788	0.972	0.732	0.527
June	0.995	0.809	0.802	0.979	0.753	0.498
July	0.993	0.822	0.819	0.975	0.772	0.541
August	0.995	0.793	0.796	0.978	0.723	0.551
September	0.994	0.807	0.783	0.972	0.744	0.502

Table 11 depicts the results from the multinomial logistic regression models. Overall accuracy is again a poor metric for evaluation due to the class imbalance resident within the response variable. F1 score and balanced accuracy is defined identically as the Random Forest models. AUC is also reported for each class of the response variable. The twelve multinomial logistic models also perform in a uniform manner.

Table 11. Multinomial Logistic Regression Evaluation Results

Model	Overall Accuracy	Balanced Accuracy	F1 Score	AUC: Continue Service	AUC: EAS	AUC: NEAS
October	0.989	0.599	0.616	0.757	0.681	0.504
November	0.989	0.534	0.602	0.702	0.615	0.454
December	0.989	0.652	0.691	0.827	0.655	0.574
January	0.990	0.634	0.690	0.750	0.713	0.437
February	0.990	0.620	0.663	0.792	0.662	0.548
March	0.990	0.591	0.630	0.787	0.617	0.613
April	0.984	0.480	0.510	0.596	0.640	0.568
May	0.985	0.599	0.648	0.807	0.613	0.637
June	0.985	0.601	0.626	0.820	0.617	0.650
July	0.983	0.654	0.699	0.826	0.673	0.553
August	0.985	0.644	0.671	0.852	0.628	0.644
September	0.987	0.668	0.671	0.790	0.734	0.434

The metrics reported in Table 10 and Table 11 are not the primary interest of this thesis. They do not completely capture the accuracy of the aggregate Base NEAS prediction results. However, the Random Forest classifier models do consistently outperform the multinomial logistic regression models at individual classification using the comparison metrics of both the F1 score and balanced accuracy.

Both ML techniques report the output of each predictive model in the form of a multiclassification confusion matrix. Figure 10 and Figure 11 report the confusion matrix for the Random Forest classifier and multinomial logistic regression October models. The rows in Figure 10 and Figure 11 represent the model predictions, and the columns represent the truth data, or actual observations of each classification type. The values in each square indicate the proportion of the overall classification, the count of the classification, and the accuracy of the count in relation to the predicted and observed values. The summation of predicted and observed values is found in the farthest right column and in the bottom row. The summation of predicted Base NEAS compared to the summation of observed Base NEAS is the primary method I use to evaluate the overall performance of the models. It best represents the Base NEAS aggregate prediction desired by end strength planners. The Random Forest classifier model in Figure 10 predicts a Base NEAS attrition total of 322 comparted to an actual Base NEAS attrition of 326. Likewise, the multinomial logistic regression model in Figure 11 predicts Base NEAS attrition of 313 compared to the observed value of 326.

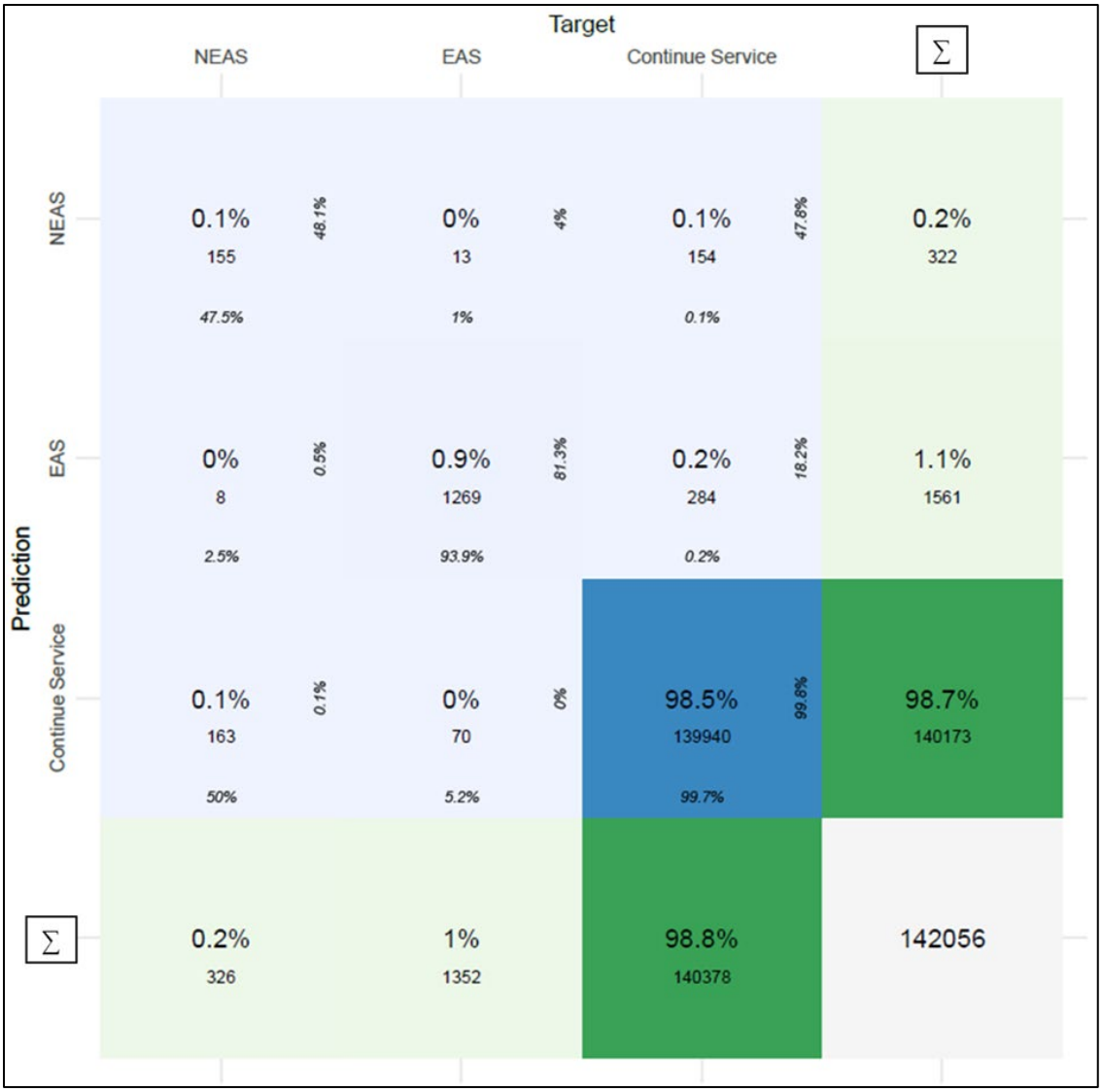


Figure 10. Random Forest Classifier October Model Confusion Matrix

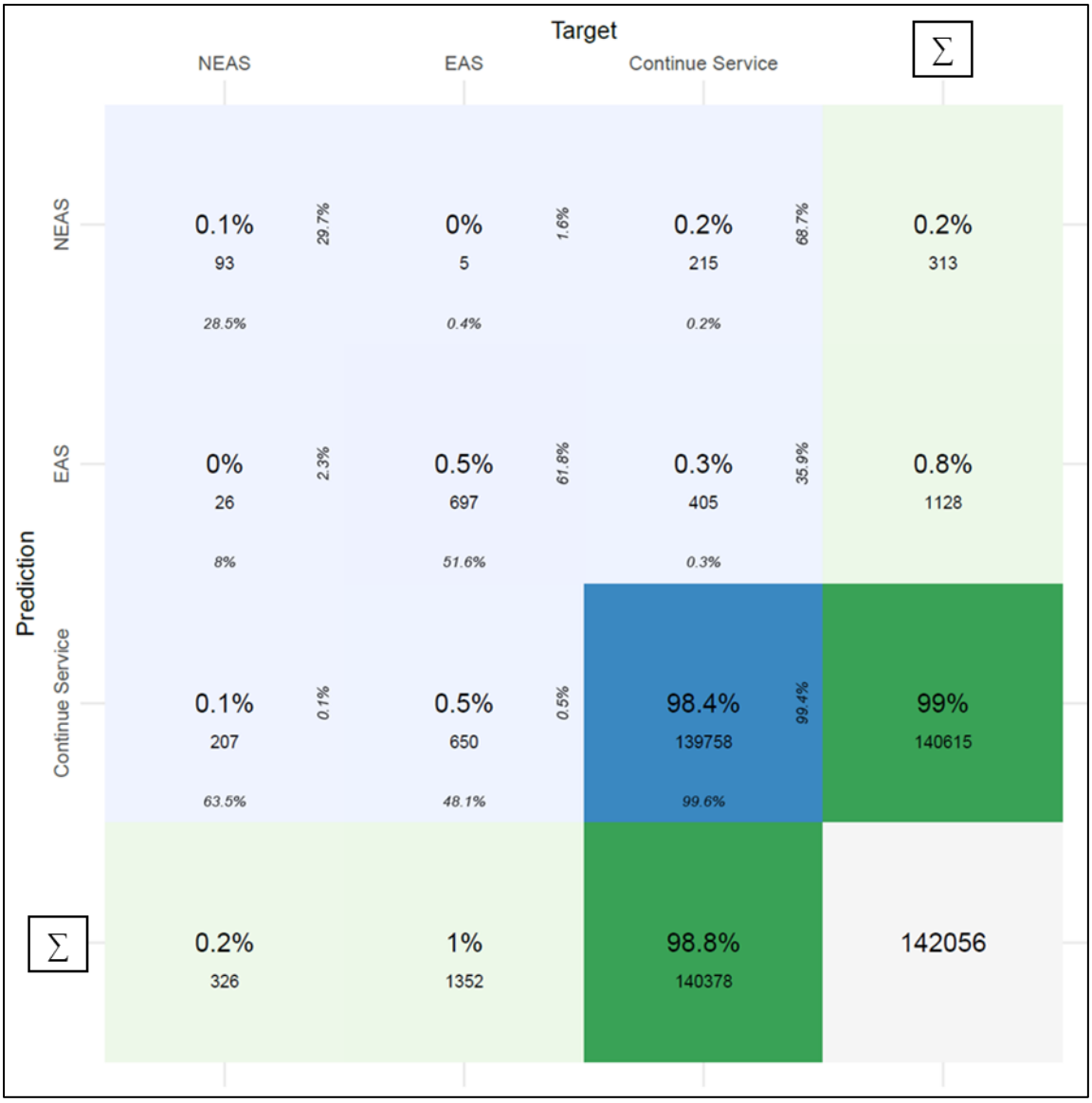


Figure 11. Multinomial Logistic Regression October Model Confusion Matrix

Figure 12 and Figure 13 report the ROC and AUC for the Random Forest classifier and multinomial logistic October models. Each specific class of the response variable has its own ROC and AUC reported. These curves indicate that both ML modeling techniques perform better at predicting continue service, and EAS than they do at predicting Base NEAS.

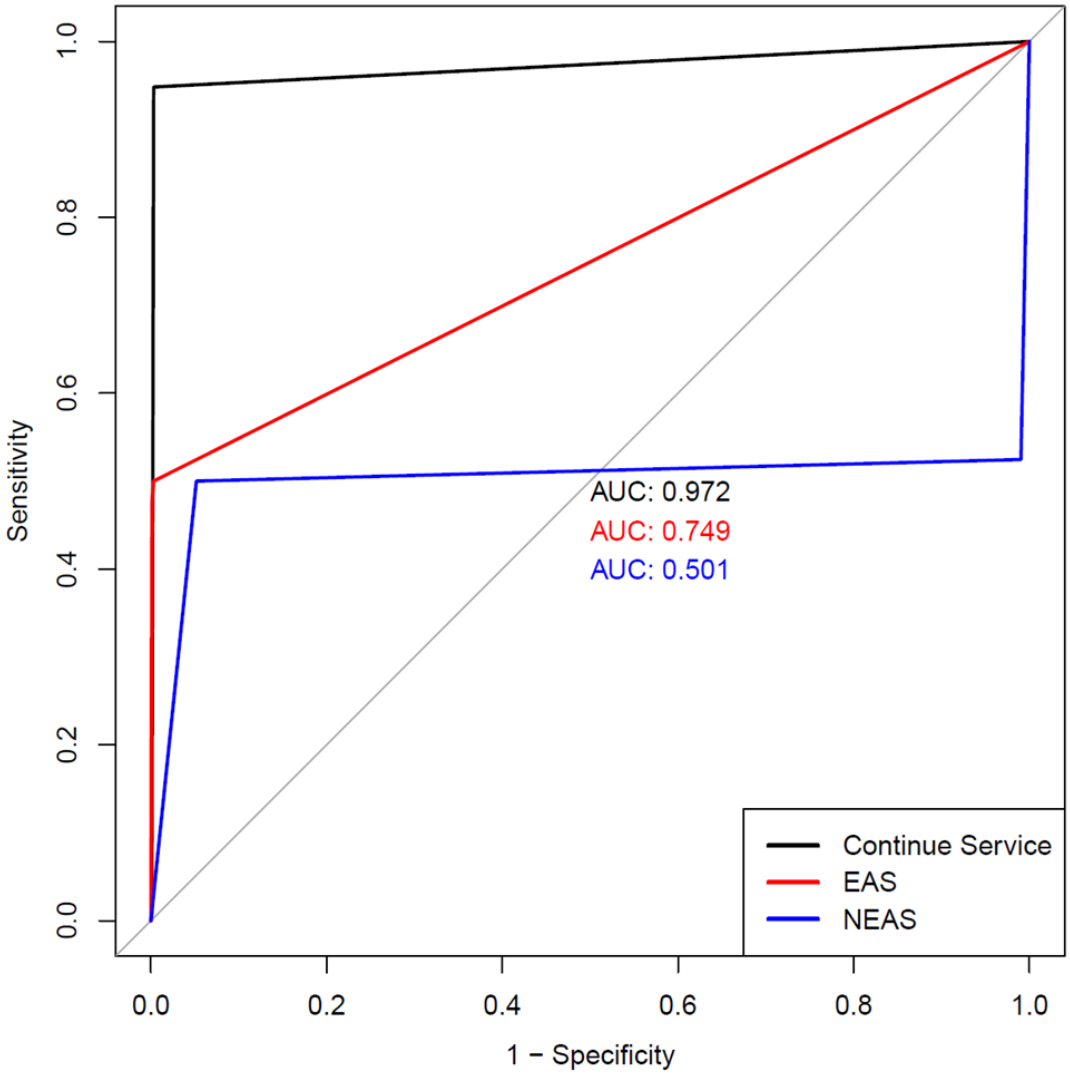


Figure 12. ROC / AUC Random Forest Classifier October Model

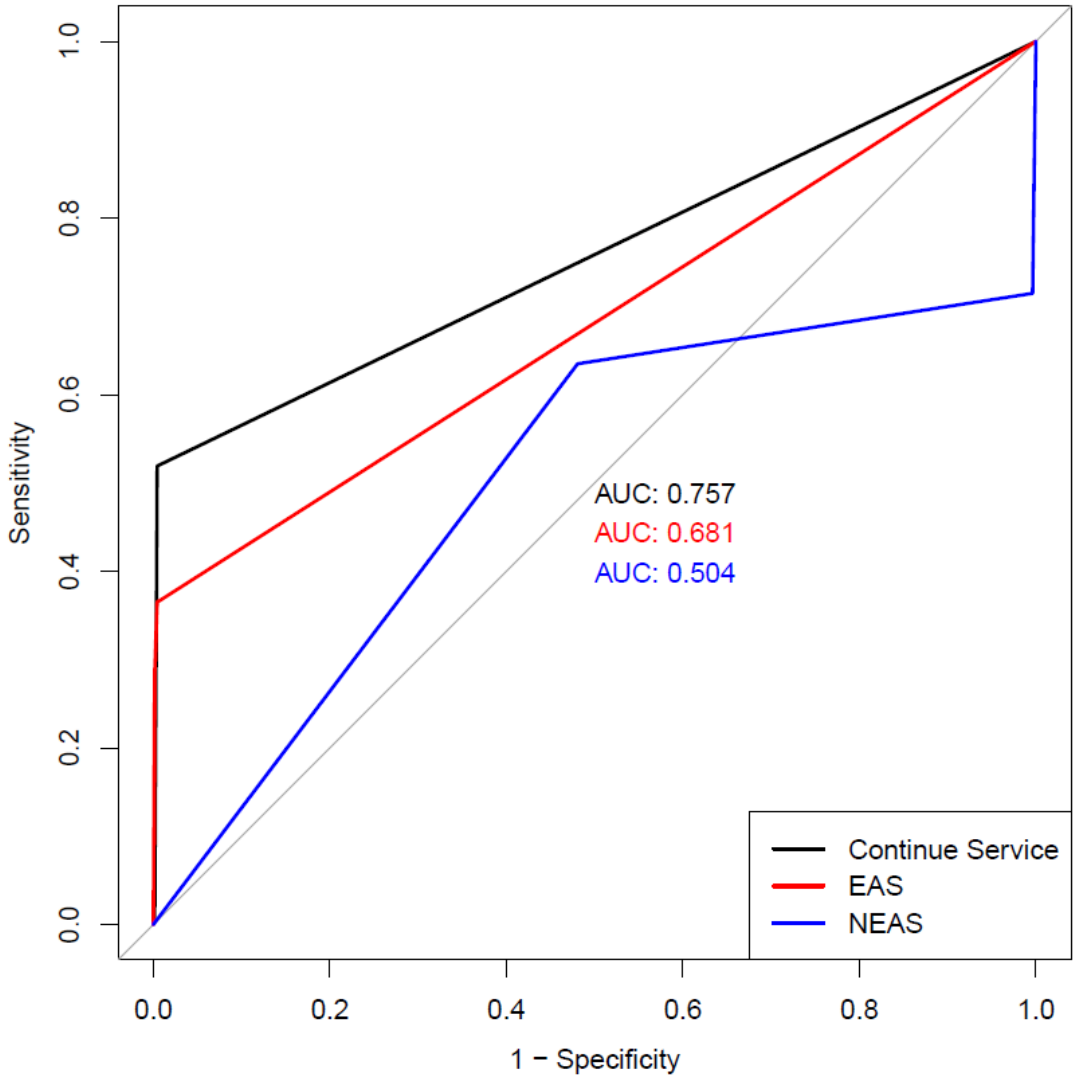


Figure 13. ROC / AUC Multinomial Logistic Regression October Model

B. MODEL COMPARISON

Model comparison is accomplished by calculating the difference between the total predicted value of Base NEAS and the actual observed values of Base NEAS attrition for each model. This model deviation calculation is conducted for each monthly model and plotted for the entirety of FY20. The summation of FY20 predicted values compared to the summation of FY20 actual values is used to describe the effectiveness of the entirety of all twelve models. The distance between the prediction and actual lines represents the error between the model's prediction of Base NEAS and the actual observed values. This applies across all models run.

These evaluation methods facilitate the best evaluation of aggregated monthly and annual Base NEAS prediction. Additionally, the cumulative model error is compounded and reported at each monthly model. This metric is useful because it indicates the attrition error up to that specific point in time, for the given FY. For example, Figure 14 shows that as of December 2020, the FY20 series of models has underestimated Base NEAS attrition by 7%. This gives manpower planners the ability to understand and adjust their retention and accession numbers accordingly.

Figure 14 depicts the trailing average models currently employed by M&RA planners. The current modeling process has difficulty responding to various shocks to the manpower system as evidenced by the divergence seen between February to July. The model predictions generally exhibit a smoothed horizontal trend. This is expected as the averaging technique removes major outliers and spikes. The cumulative error of the model generally trends downward, indicated that the aggregate effect of the model is to underestimate Base NEAS attrition.

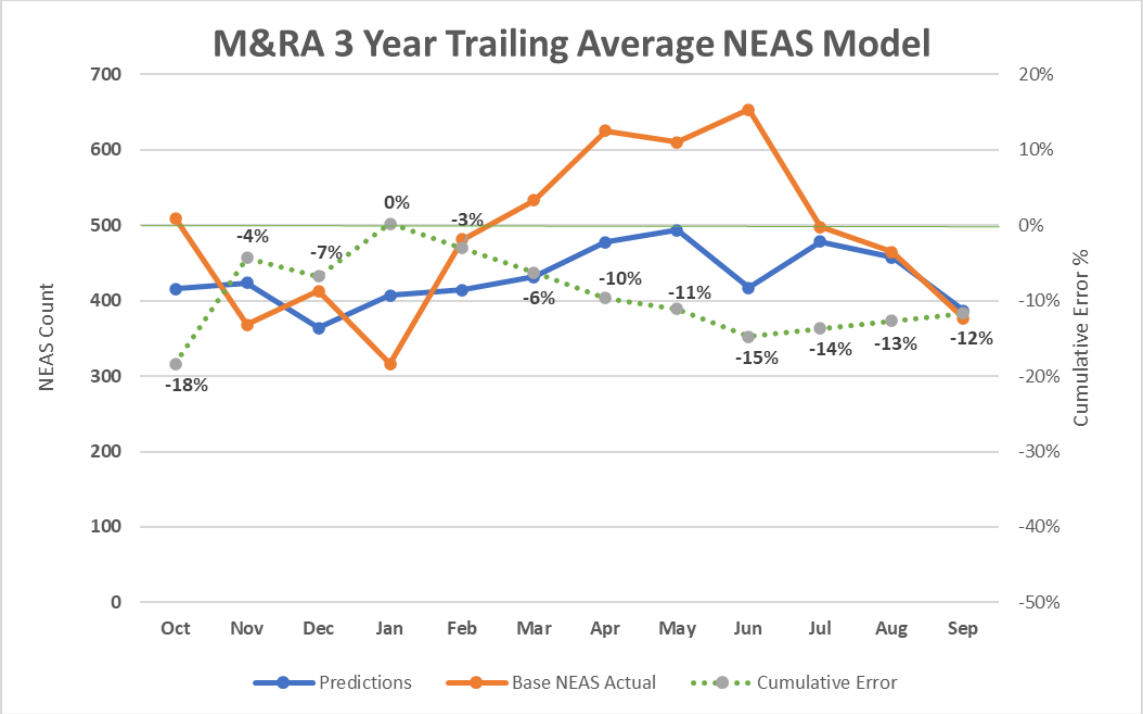


Figure 14. M&RA Trailing Average Base NEAS Models FY20

Figure 15 depicts the three-year trailing average model run on the training data. The results from these models are consistent with the results seen in the M&RA trailing average models in Figure 14. This provides further evidence that although I am not able to replicate the separations data to an exact figure, the underlying process and results are consistent and correct. The trailing average model behaves as expected providing a relatively smooth and linear prediction of Base NEAS. It suffers from the same inability to respond to shocks or spikes in the manpower system and tends to underestimate Base NEAS attrition.

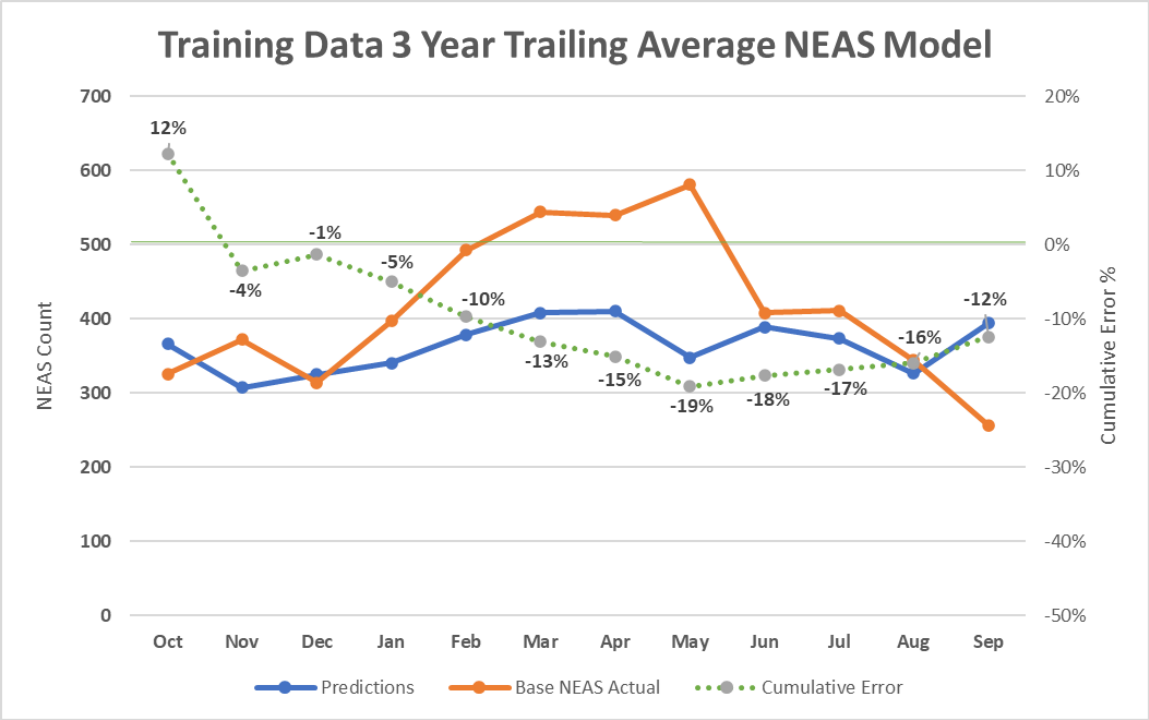


Figure 15. Training Data Trailing Average Base NEAS Models FY20

Figure 16 depicts the results from the multinomial logistic regression models. The models perform poorly and consistently underestimates aggregate Base NEAS attrition. The models may suffer from a smaller predictor pool, due to the employment of the LASSO technique, required for timely computation. Additionally, this regression technique has limited ability to alter the algorithms parameters making it difficult to find an optimum.

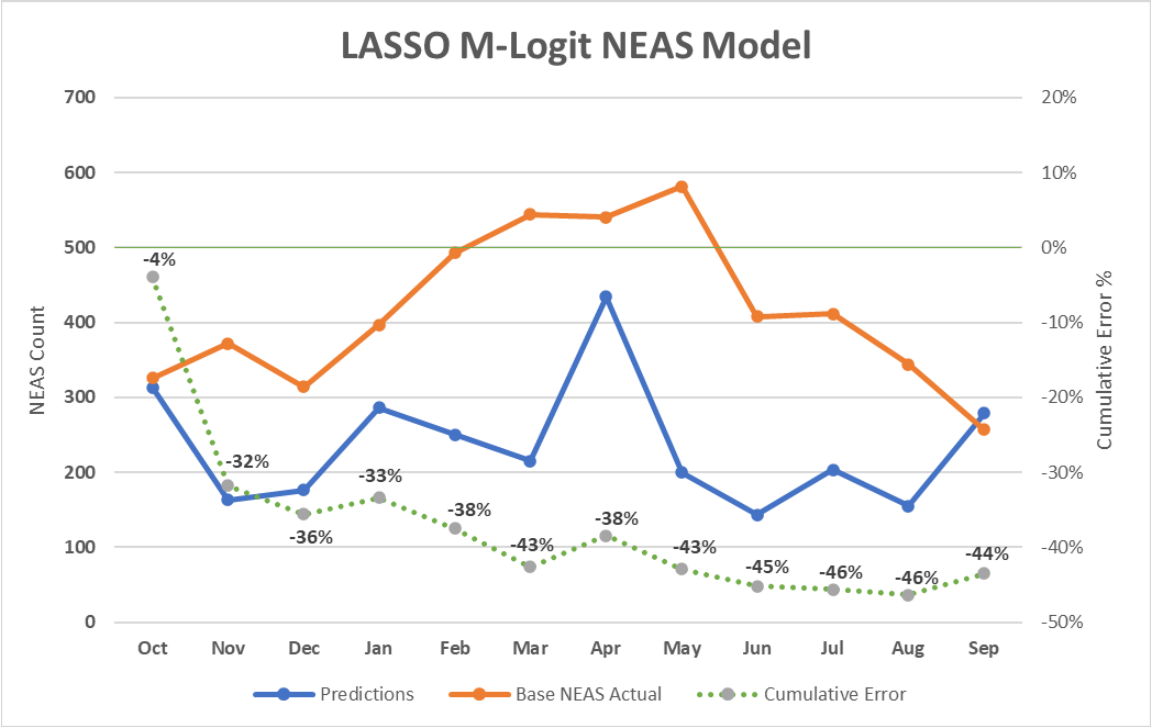


Figure 16. Multinomial Logistic Regression Base NEAS Models FY20

Figure 17 displays the Random Forest classifier models. These models consistently track the validation data and demonstrate an improved ability to accurately account for manpower shocks or spikes in the system. This is specifically seen in the February to July models and in contrast to the other models. The Random Forest technique is particularly suited to adjust for differences at the individual Marine level. The cumulative error remains relatively stable throughout FY20. The models still consistently underestimate Base NEAS attrition, however, by a reduced margin.

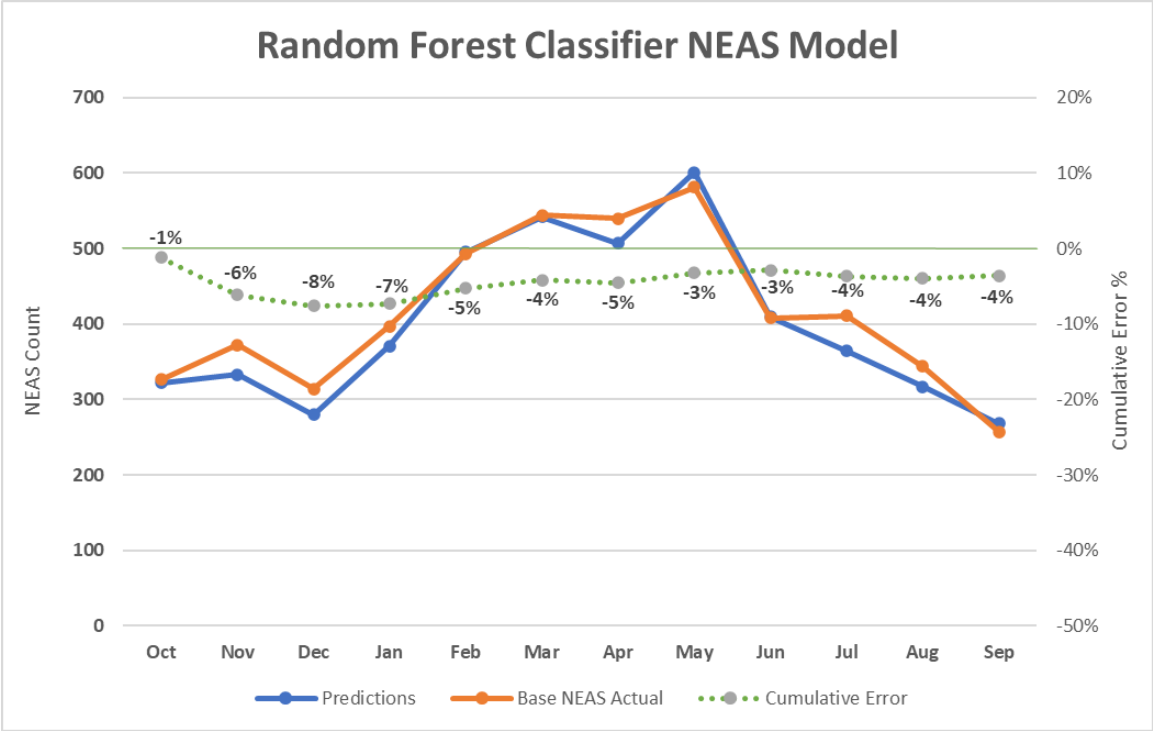


Figure 17. Random Forest Classifier Base NEAS Models FY20

Table 12 depicts the comparison results, by monthly model, of each technique I employ. It reports the model deviation calculations in terms of the percent, plus or minus, difference between each model's predictions and the actual Base NEAS observations. For example, for the month of October, model projections using the M&RA Monthly Trailing Average resulted in underestimated NEAS attrition by 18%. Likewise, the Random Forest Classifier model for the same month underestimates Base NEAS attrition by 1%. The results indicate that the Random Forest classifier models generally outperform the other monthly models that are employed. The Random Forest classifier models attain significantly higher annual Base NEAS prediction accuracy achieving an underestimation of 4% on the entire FY20 validation data. This is compared to the current M&RA modeling process which achieves an underestimation of annual Base NEAS attrition by 12%.

Table 12. Base NEAS Models Deviation Comparison

Modeling Technique	Oct	Nov	Dec	Jan	Feb	Mar
M&RA Monthly Trailing Average	-18%	15%	-12%	29%	-14%	-19%
Multinomial Logistic Regression	-4%	-56%	-44%	-28%	-49%	-60%
Training Data Trailing Average	12%	-17%	4%	-14%	-23%	-25%
Random Forest Classifier	-1%	-10%	-11%	-7%	0%	0%

Modeling Technique	Apr	May	Jun	Jul	Aug	Sep
M&RA Monthly Trailing Average	-24%	-19%	-36%	-4%	-2%	3%
Multinomial Logistic Regression	-20%	-66%	-65%	-51%	-55%	9%
Training Data Trailing Average	-24%	-40%	-5%	-9%	-5%	53%
Random Forest Classifier	-6%	3%	0%	-11%	-8%	4%

Modeling Technique	Cumulative Annual Totals
M&RA Monthly Trailing Average	-12%
Multinomial Logistic Regression	-44%
Training Data Trailing Average	-12%
Random Forest Classifier	-4%

C. PREDICTORS

The multinomial logistic regression models require a condensed pool of predictor variables for efficient computation. I employ the LASSO technique for each monthly model to select the most significant predictors associated with the response variable. LASSO is applied on the training dataset. Figure 18 reports top fifty LASSO selected predictors for the multinomial logistic regression October model. If the selected variable is categorical, the LASSO algorithm I use reports the specific category of that variable that is selected. I use the entire categorical variable as I am not able to remove the less predictive individual categories. I then estimate the multinomial logistic regression on the set of predictor variables that were optimally selected using the LASSO technique. The color of the bars in Figure 18 represents the direction of the relationship between the predictor and the response variable. Blue indicates a positive directional relationship and red represents a negative one. For example, the red bar for years of service indicates that the longer a Marine has served, the more likely that Marine is to NEAS.

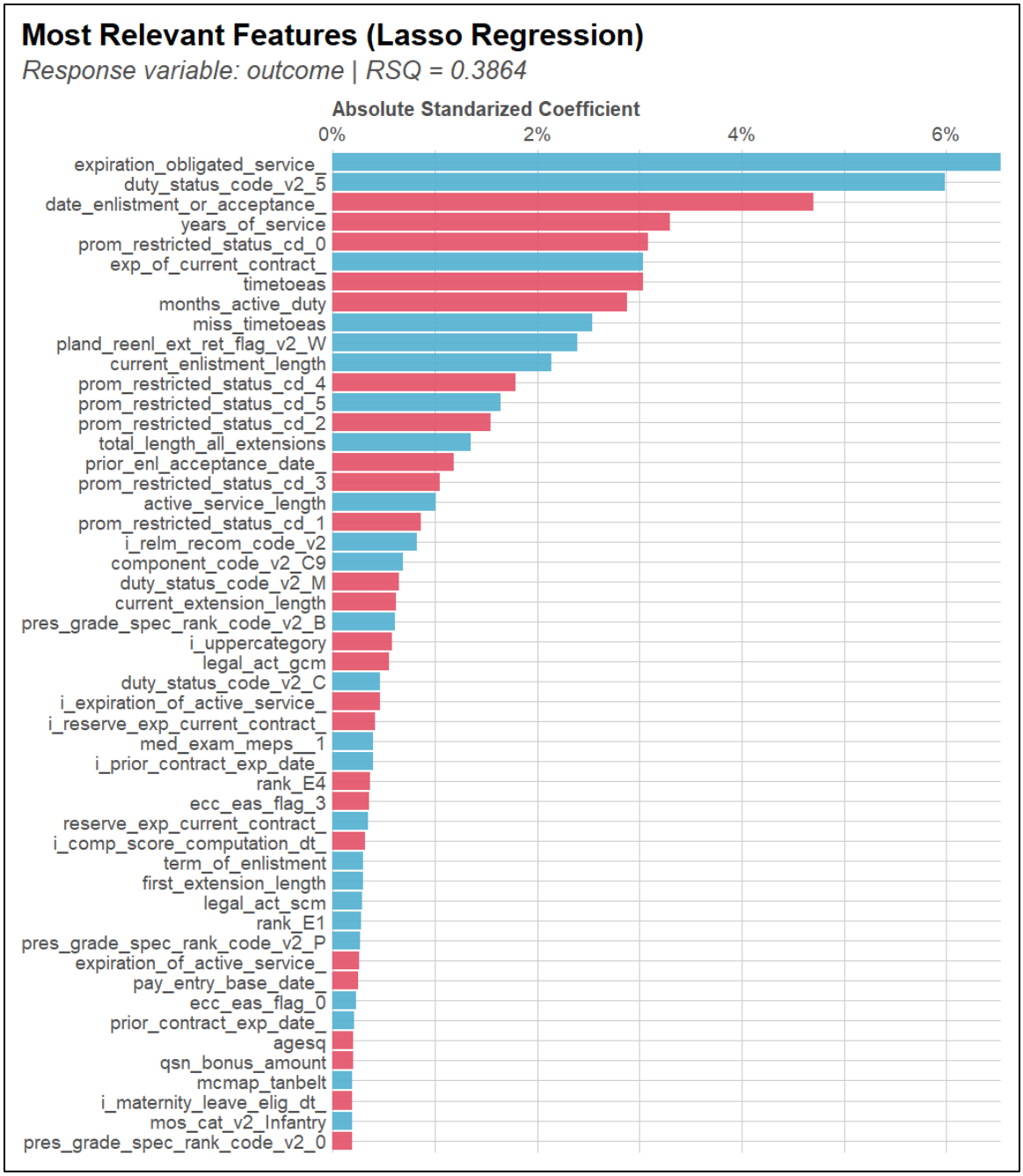


Figure 18. October Model LASSO Predictor Variables

Meanwhile, the Random Forest classifier variable importance plot indicates the most significant predictor variables for the response variable. Figure 19 depicts the top fifty predictor variables associated with the October Random Forest classifier model. Random Forest models do not report the direction of the relationship between the predictor and the

response variable, nor can they report an interpretable predictor coefficient. In both Figure 18 and Figure 19 all date predictors represent the count of days between the date predictor and the sequence number, or month, under observation.

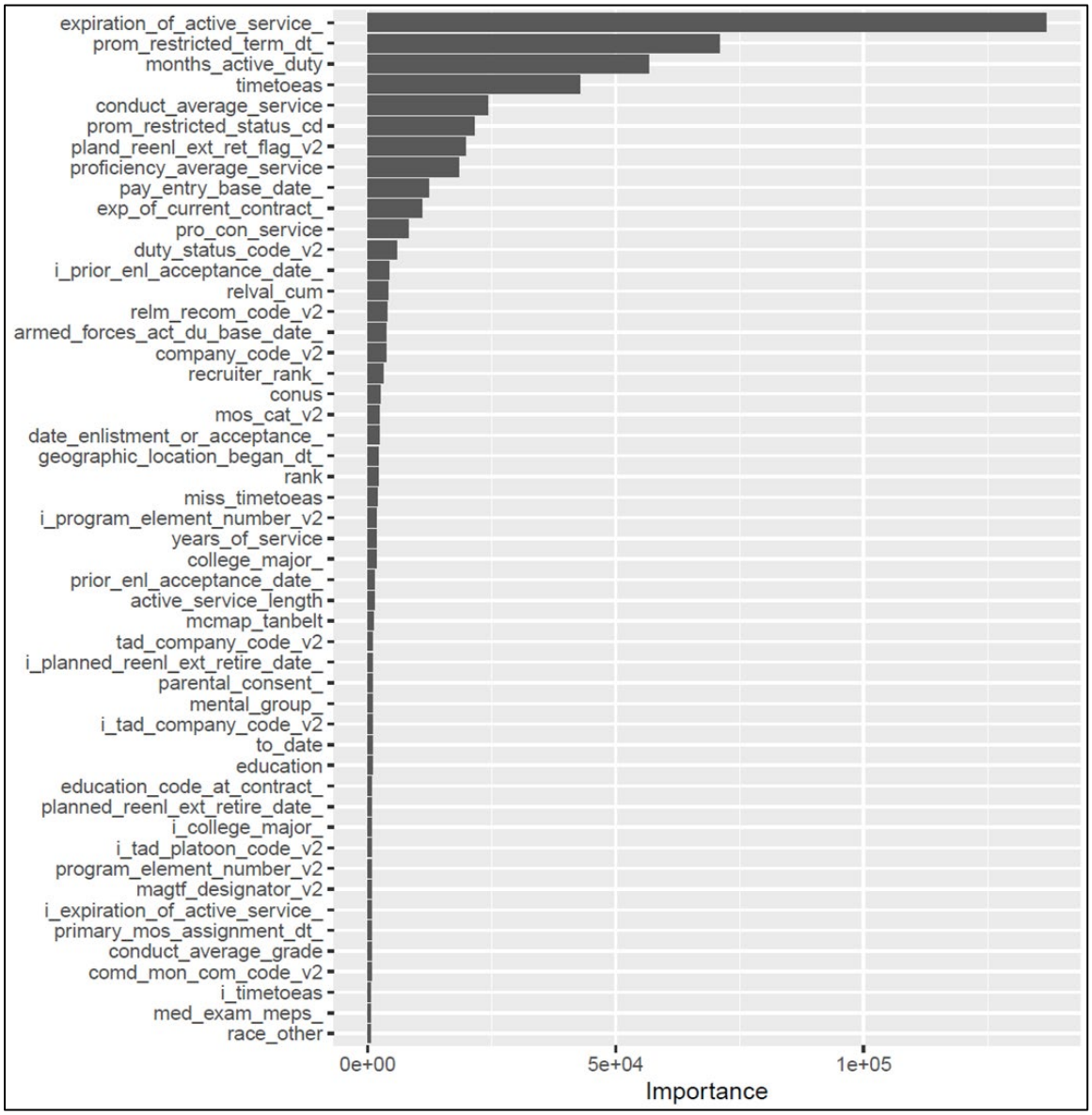


Figure 19. October Model Random Forest Variable Importance Plot

Each monthly model within the Random Forest classifier has a unique variable importance plot. Each model within the multinomial logistic regression also possesses a

unique LASSO selected predictor variable set. I observe substantial overlap in the set of predictors between all of the models from both techniques. Figure 20 reports the predictor overlap for the Random Forest classifier monthly models. Figure 20 shows predictors that occurred eight or more times across all twelve models. The Average Predictor Rank column represents the summation of the rank order within each model, divided by the total number of models, with the lowest score equating to the highest rank. It represents the most significant predictor in the training dataset, across all twelve Random Forest classifier models. For example, the predictor expiration_of_active_service_ occurred in all twelve models and has an average rank of 1.5, the most significant predictor in the family of Random Forest classifier models. The Count Of Occurrence column indicates the number of times that the specified predictor occurs in total over the twelve different models. The variables timetoeas, miss_timetoeas, and conus are all examples of feature engineered predictors.

Predictor	Count Of Occurance	Average Predictor Rank
expiration_of_active_service_	12	1.5
prom_restricted_term_dt_	12	2.2
timetoeas	12	3.5
months_active_duty	12	4.9
prom_restricted_status_cd	12	6.3
conduct_average_service	12	6.5
proficiency_average_service	12	8.8
pland_reenl_ext_ret_flag_v2	12	9.0
exp_of_current_contract_	12	10.9
armed_forces_act_du_base_date_	12	11.3
pay_entry_base_date_	12	11.6
duty_status_code_v2	12	13.3
pro_con_service	12	14.7
company_code_v2	12	17.3
mcmap_tanbelt	12	17.4
miss_timetoeas	12	18.2
relm_recom_code_v2	12	20.8
rank	12	20.9
geographic_location_began_dt_	12	25.8
years_of_service	12	33.1
i_program_element_number_v2	11	24.9
conus	11	26.2
i_planned_reenl_ext_retire_date_	11	28.8
primary_mos_assignment_dt_	11	30.0
mos_cat_v2	10	21.5
date_enlistment_or_acceptance_	10	22.3
prior_enl_acceptance_date_	10	28.1
mental_group_	10	28.7
i_prior_enl_acceptance_date_	9	19.7
i_career_status_flag_v2	9	25.8
comd_mon_com_code_v2	9	29.9
conduct_average_grade	8	17.8
planned_reenl_ext_retire_date_	8	21.1
to_date	8	21.3

Figure 20. Random Forest Classifier Predictor Overlap

The Random Forest classifier models do also contain many unique predictors specific to each monthly model. These models exhibit a non-parametric relationship. This supports the premise that each monthly model be trained independently on past data from that same month, not on aggregated yearly data. Figure 21 depicts the relationship between each model and its predictors.

$$Response\ Variable\ = f_m(x)$$
$$m = unique\ month$$
$$x = predictors$$

Figure 21. Non-parametric Modeling Relationship

D. LIMITATIONS

Several limitations are present within my research. The first limitation is the size of the training dataset that I employ. The data represents three years of data for a given month. ML techniques are equipped to operate on much larger datasets. I am unable to employ the full ten years of data for this research due to computing power issues.

An additional limitation is the inability to employ a hyperparametric tuning algorithm to find the optimal model parameters. I use a manual method of finding the optimal value of one hyperparameter, Mtry, because of the computing limitations of the HPC and software packages used. More advanced k-fold cross validation or grid search methods find the optimal values for all algorithm parameters. The multinomial logistic regression models are also employed on training data that is highly imbalanced. The SMOTE technique is not employed on the training data for the multinomial logistic regression models.

General computing power provided by the HPC represents another limitation. The memory required to load and clean entire datasets, preprocess data, and run models, overwhelmed the HPC. Because of this, I was limited in the size and scope of the modeling and data used.

The inability to exactly replicate the M&RA Base NEAS attrition numbers is another limitation. This limitation is not of material concern to the results of this thesis as I am able to overcome it by employing an identical process on the training data.

Data collection, storage, and retrieval remains a constant theme through the related USMC manpower literature (Orrick, 2008). The data I obtained from TFDW is incomplete. Numerous additional data tables that were requested were not able to be found.

Additionally, certain data tables were incomplete and had to be discarded. The USMC manpower data systems and infrastructure did not provide smooth facilitation of the ML modeling techniques that I employ.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION AND RECOMMENDATIONS

Current Base NEAS modeling uses aggregate data and trailing averages to predict attrition. This approach is simple, efficient, and generally accurate. However, it fails to account for various shocks to the system and relies on constant composition and disposition of the force. My modeling approach seeks to classify Base NEAS attrition at both the individual and aggregate level using modern ML techniques. This approach is significantly more compute intensive but can result in a more accurate prediction of Base NEAS attrition.

A. SUMMARY

1. Primary Research Questions

a. Can a manpower model be developed using ML that predicts monthly Base NEAS attrition using individual Marine data contained in existing USMC personnel databases?

b. Are ML modeling approaches using individual data better at predicting aggregate Base NEAS attrition compared to current models?

Manpower data and attrition classification lends itself to ML, specifically Random Forest and logistic regression models. I find that the Radom Forest classifier series of models outperforms all other models that I employ and that are currently used by USMC End Strength Planners. The construction of an ML model that predicts Base NEAS attrition using existing data is possible. The results of these ML models, specifically the Random Forest classifier, are generally better at predicting Base NEAS Attrition than current processes. The ability of the ML algorithms to identify individual attributes at the Marine level give them the ability to more accurately account for the various shocks to the manpower system.

2. Secondary Research Questions

a. Does using disaggregated data and an ML model better account for various shocks to the USMC manpower system?

The use of individual data and ML techniques does help to ensure accurate predictions when dealing with various shocks to the USMC enlisted end strength model. The results of the Random Forest Classifier models from February through July indicate this. Consistent with prior literature, disaggregated data and individual attributes can be used as predictors to improve model accuracy.

b. Is the predictive ML model developed feasible for Marine planners to implement given their current systems, software, and programs?

The feasibility of implementing ML models for USMC End Strength Planners is not able to be deduced from this research. The opportunity costs associated with the development and implementation of ML modeling may or may not be feasible given the time, financial, and data systems constraints currently placed on Marine planners. Additionally, the structure of the current databases and data systems do not facilitate easy employment of the techniques employed in this thesis.

B. RECOMMENDATIONS

The results of this thesis serve as a proof of concept and indicate that using ML to modernize manpower models can yield better predictive results. Additionally, even though the individual correct classification of Base NEAS hovered around 45%, value can be obtained from identifying even half of these Marines at the individual level. The institution should focus on three areas. First, it should optimize its data systems and data infrastructure to better facilitate the use of ML. Second, it should continue to look for opportunities to employ ML in areas that currently use legacy systems or techniques. It should also explore using ML techniques in conjunction with legacy modeling in a hybrid fashion. Finally, the institution should continue to strive to collect better data that can help build more accurate future ML models.

C. FUTURE RESEARCH

Areas for future research include the following:

1. Employment of larger, aggregated training dataset and alternate methods to define the response variable.
2. Employment of alternate ML classification techniques.
3. Development of additional in-service data to be collected to better facilitate attrition at the individual Marine level.
4. Interpretation of the marginal effects associated with significant predictors found within this thesis to help inform policy affecting attrition.
5. Explore opportunities to employ ML and legacy systems and techniques in a hybrid fashion that meets managerial relevance.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX

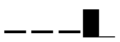
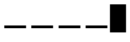
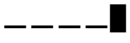
A. SUMMARY STATISTICS OCTOBER MODEL TRAINING DATA

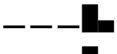
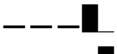
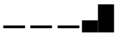
	Name: oct_model		
	Number of Rows	rows	424886
	Number of Columns	columns	472
	_____		
	Type	frequency:	
	Date	1	
	factor	68	
	numeric	403	
	_____		
	Group Variables	None	
	Variable	type:	Date
	skim_variable	n_missing	min	max
1	filedate	0	10/31/2016	10/31/2018
		n_unique		
		3		

	Variable	type:	Factor	
	skim_variable	n_missing	ordered	n_unique
1	ecc_eas_flag	0	FALSE	7
2	reserve_reporting_unit_code	0	FALSE	155
3	career_status_bonus_elect_cd	0	FALSE	2
4	component_code_v2	0	FALSE	4
5	res_component_code_v2	0	FALSE	7
6	res_record_status_code_v2	0	FALSE	3
7	civ_educ_graduation_cd_v2	0	FALSE	3
8	civ_ed_major_subject_cd_v2	0	FALSE	252
9	relm_recom_code_v2	0	FALSE	2
10	dod_trn_cat_pay_group_v2	0	FALSE	12
11	grade_select_v2	0	FALSE	14
12	religion_v2	0	FALSE	271
13	home_of_rec_st_cntry_v2	0	FALSE	87
14	comd_mon_com_code_v2	0	FALSE	298
15	service_spouse_code_v2	0	FALSE	2

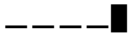
16	bonus_prgm_enl_for_cd_v2	0	FALSE	2
17	res_mon_command_code_v2	0	FALSE	175
18	mobilization_status_v2	0	FALSE	5
19	crisis_remark_uniq_id_v2	0	FALSE	4
20	crisis_remk_code_desc_tx_v2	0	FALSE	4
21	duty_status_code_v2	0	FALSE	31
22	program_element_number_v2	0	FALSE	241
23	contract_legal_agrment_v2	0	FALSE	1
24	pland_reenl_ext_ret_flag_v2	0	FALSE	15
25	addl_first_mos_code	0	FALSE	2
26	addl_second_mos_code	0	FALSE	2
27	addl_temp_reporting_unit_code	0	FALSE	190
28	temporary_add_duty_excess_flag	0	FALSE	6
29	fap_reporting_unit_code	0	FALSE	2
30	operation_reporting_unit_code	0	FALSE	65
31	prom_restricted_status_cd	0	FALSE	7
32	plead_city_cd	0	FALSE	10
33	plead_county_cd	0	FALSE	10
34	plead_state_country_cd	0	FALSE	10
35	plead_zip_cd	0	FALSE	11
36	reenlistment_bonus_type_v2	0	FALSE	3
37	custody_status_code_v2	0	FALSE	5
38	separation_incentive_code_v2	0	FALSE	3
39	mobiliz_mont_cmd_cd_v2	0	FALSE	42
40	officer_candidate_cd_v2	0	FALSE	2
41	rank	0	FALSE	10
42	pres_grade_spec_rank_code_v2	0	FALSE	9
43	additional_temporary_mcc_v2	0	FALSE	200
44	citzen_cntry_orig_geo_code_v2	0	FALSE	244
45	place_of_birth_v2	0	FALSE	247
46	separation_document_type_v2	0	FALSE	10
47	fap_monitored_com_cd_v2	0	FALSE	2
48	company_code_v2	0	FALSE	36
49	cohesion_future_mc_code_v2	0	FALSE	63
50	magtf_designator_v2	0	FALSE	77
51	fap_company_code_v2	0	FALSE	24
52	fap_platoon_code_v2	0	FALSE	229
53	operation_identifier_v2	0	FALSE	48
54	operation_montd_cmd_code_v2	0	FALSE	69
55	tad_company_code_v2	0	FALSE	36

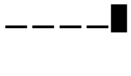
56	education_tier_at_contract	0	FALSE	4
57	mos_cat_v2	0	FALSE	46
58	activity_description_	0	FALSE	7
59	asvab_version_	0	FALSE	92
60	citizenship_desc_	0	FALSE	2
61	college_major_	0	FALSE	219
62	current_state_code_	0	FALSE	81
63	district_	0	FALSE	7
64	education_code_at_contract_	0	FALSE	81
65	hor_state_code_	0	FALSE	81
66	medical_disposition_	0	FALSE	3
67	source_code_	0	FALSE	22
68	outcome	0	FALSE	3

—	Variable	type:	Numeric		
	skim_variable	n_missing	mean	sd	hist
1	education	0	1.21E+01	1.24	
2	age	0	2.43E+01	5.76	
3	months_active_duty	0	5.93E+01	63.2	
4	current_enlistment_length	0	4.20E+00	0.559	
5	first_extension_length	0	1.53E+00	5.35	
6	current_extension_number	0	1.31E-01	0.389	
7	current_extension_length	0	1.55E+00	5.35	
8	home_of_record_county	0	9.41E+01	114	
9	number_of_dependents	0	7.57E-01	1.21	
10	years_of_service	0	4.60E+00	5.36	
11	total_satisfactory_year	0	1.05E-01	0.896	
12	maternity_leave_bal_qy	0	1.70E-03	0.0412	
13	billet_mos_v2	0	2.58E+02	220	
14	self_education_bonus_points	0	1.45E+01	32.7	
15	cmd_recruiter_bonus_points	0	2.64E-02	1.05	
16	not_recommended_for_promotion	0	4.43E-02	0.206	
17	conduct_average_grade	0	2.19E+00	2.17	
18	proficiency_average_grade	0	2.20E+00	2.17	
19	conduct_average_service	0	4.26E+00	0.675	
20	proficiency_average_service	0	4.27E+00	0.675	
21	special_duty_bonus_points_qy	0	6.66E-01	8.13	
22	retired_category_code	0	1.20E-04	0.019	
23	active_service_length	0	1.51E+01	22.7	
24	total_length_all_extensions	0	1.69E+00	5.83	

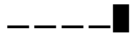
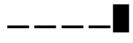
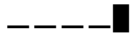
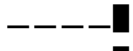
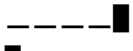
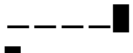
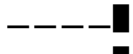
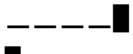
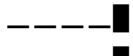
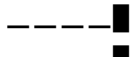
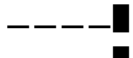
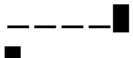
25	total_number_of_extensions	0	1.32E-01	0.39	
26	responsibility_center_number	0	0	0	
27	composite_score	0	4.11E+02	693	
28	first_depl_ttl_day_waiver_qy	0	3.92E+01	66.8	
29	prospectiv_off_srce_cd_v2	0	3.39E-04	0.0184	
30	selective_relm_bonus_zone_v2	0	4.95E-01	1.08	
31	selres_trans_prgm_code_v2	0	8.47E-05	0.0159	
32	career_status_flag_v2	0	6.28E-01	0.93	
33	cohesion_number_id_v2	0	2.12E-05	0.00797	
34	flight_status_identifier_v2	0	1.30E-02	0.113	
35	incurred_obligd_service_cd_v2	0	0	0	
36	natl_call_srv_waiver_code_v2	0	1.41E-05	0.00651	
37	home_of_selection_flg_v2	0	2.38E-03	0.0487	
38	final_payment_flg_v2	0	0	0	
39	mplp_caregiver_cd_v2	0	0	0	
40	apc_date	0	-4.25E+01	108	
41	received_date	0	-3.88E+01	104	
42	from_date	0	5.41E+01	112	
43	to_date	0	-2.47E+01	91.2	
44	avg_this_rpt	0	1.21E+00	1.76	
45	relval_at_proc	0	1.09E+04	31091	
46	relval_cum	0	7.95E+03	27013	
47	uppercategory	0	1.10E+01	18.8	
48	withcategory	0	9.38E+00	14	
49	belowcategory	0	1.27E+01	20.6	
50	num_fitrep	0	3.74E+00	6.3	
51	months_fitrep	0	2.59E+00	4.15	
52	adverse_fitrep	0	9.99E-03	0.0995	
53	commendatory_fitrep	0	1.16E-01	0.321	
54	rec_4_promote	0	2.68E-01	0.443	
55	sat_score	0	5.31E+00	78.1	
56	act_score	0	2.45E-01	2.4	
57	afqt_score	0	5.92E+01	21.3	
58	dlab_score	0	2.03E+02	128867	
59	el	0	1.04E+02	24.6	
60	gt	0	1.03E+02	24.5	
61	ist_pull_ups	0	1.05E+01	7.64	
62	term_of_enlistment	0	4.17E+00	1.01	
63	qsn_bonus_amount	0	1.18E+04	2471	
64	phys_fit_score_qy	0	2.26E+02	77.7	

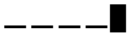
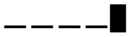
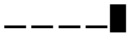
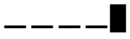
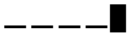
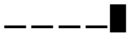
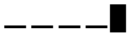
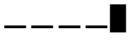
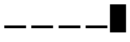
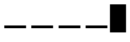
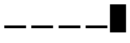
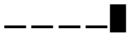
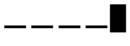
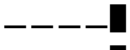
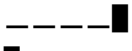
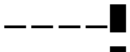
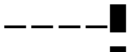
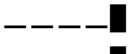
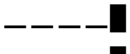
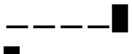
65	phys_fit_pull_up_scr_qy	0	1.66E+01	10.8	
66	phys_fit_crunch_scr_qy	0	9.76E+01	28.9	
67	phys_fit_run_scr_tm	0	2.12E+03	662	
68	phys_fit_push_up_scr_qy	0	1.26E+00	8.25	
69	phys_fit_rowing_scr_tm	0	3.29E+00	85.5	
70	pft_third_class	0	7.91E-02	0.27	
71	pft_second_class	0	2.14E-01	0.41	
72	pft_first_class	0	6.63E-01	0.473	
73	award_unit	0	3.68E-02	0.19	
74	award_service	0	9.03E-01	0.401	
75	award_personal	0	7.03E-02	0.257	
76	legal_act_gcm	0	9.77E-04	0.0312	
77	legal_act_njp	0	6.14E-02	0.24	
78	legal_act_scm	0	8.00E-04	0.0283	
79	legal_act_spcm	0	4.78E-04	0.0219	
80	mcmap_instructor	0	3.47E-02	0.183	
81	mcmap_above_blackbelt	0	1.08E-03	0.0328	
82	mcmap_blackbelt	0	1.03E-01	0.304	
83	mcmap_brownbelt	0	1.04E-01	0.306	
84	mcmap_greenbelt	0	1.96E-01	0.397	
85	mcmap_graybelt	0	1.93E-01	0.394	
86	mcmap_tanbelt	0	2.92E-01	0.455	
87	mcmap_no_belt	0	7.96E-03	0.0888	
88	mcmap_revoke	0	2.74E-03	0.0523	
89	present_grade_effective_date_	0	4.62E+02	497	
90	pay_entry_base_date_	0	1.84E+03	1954	
91	extension_enlist_effect_date_	0	5.18E+02	1750	
92	armed_forces_act_du_base_date_	0	1.82E+03	1924	
93	retirement_date_	0	1.81E-01	30.1	
94	initial_active_duty_date_	0	6.04E+01	555	
95	date_joined_smcr_	0	4.31E+01	478	
96	date_enlistment_or_acceptance_	0	8.19E+02	479	
97	date_of_birth_	0	9.05E+03	2102	
98	prior_contract_exp_date_	0	8.29E+00	61.4	
99	accession_date_	0	1.33E+03	1309	
100	current_tour_begin_date_	0	5.81E+02	551	
101	present_rank_date_	0	4.63E+02	498	
102	planned_reenl_ext_retire_date_	0	-1.86E+00	41.1	
103	dep_location_begin_date_	0	2.85E+02	597	
104	present_unit_joined_date_	0	4.96E+02	444	

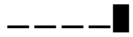
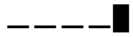
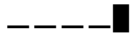
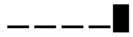
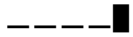
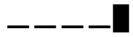
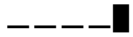
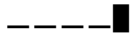
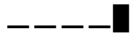
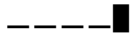
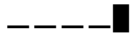
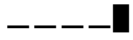
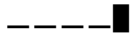
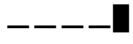
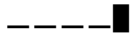
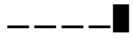
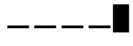
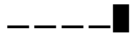
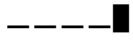
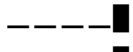
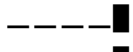
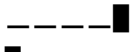
105	strength_cat_effective_date_	0	2.82E+02	284	
106	date_initial_entry_reserve_	0	7.62E+01	633	
107	service_spouse_effective_date_	0	7.84E+01	424	
108	selected_grade_date_	0	-1.09E-02	3.56	
109	special_duty_bonus_date_	0	-3.93E+00	53.2	
110	intended_transfer_date_	0	-3.32E+01	154	
111	drop_discharge_unit_date_	0	5.98E+01	545	
112	anniversary_date_	0	4.96E+01	478	
113	officer_candidate_eff_dt_	0	4.16E-01	48.3	
114	prior_enl_acceptance_date_	0	7.09E+02	1109	
115	permanent_rank_date_	0	4.63E+02	498	
116	good_conduct_medal_date_	0	5.01E+02	313	
117	deployed_calculation_date_	0	3.98E+00	2.83	
118	flight_status_date_	0	2.63E+01	274	
119	natl_call_service_waiver_date_	0	2.17E-02	10	
120	operation_effective_date_	0	7.71E+00	50.2	
121	asvab_date_	0	2.09E+03	1913	
122	civ_educ_left_school_dt_1_	0	2.15E+03	1762	
123	crisis_participation_begin_dt_	0	3.41E+00	22.2	
124	primary_mos_assignment_dt_	0	1.19E+03	1322	
125	geographic_location_began_dt_	0	6.12E+02	577	
126	career_status_bonus_elect_dt_	0	1.19E+02	539	
127	maternity_leave_elig_dt_	0	2.88E-01	8.17	
128	addl_first_mos_assignment_dt_	0	5.67E+02	1274	
129	addl_second_mos_assignment_dt_	0	2.37E+02	851	
130	cohesion_proj_train_compl_dt_	0	7.66E+02	1355	
131	comp_score_computation_dt_	0	8.13E+02	1613	
132	prom_restricted_term_dt_	0	1.16E+02	609	
133	law_enforce_ci_id_dt_	0	1.15E+01	129	
134	final_payment_flg_dt_	0	0	0	
135	mplp_leave_elig_dt_	0	0	0	
136	exp_of_current_contract_	0	-7.48E+02	440	
137	expiration_of_active_service_	0	-7.76E+02	430	
138	reserve_exp_current_contract_	0	7.07E-01	176	
139	expiration_obligated_service_	0	-8.91E+02	1951	
140	sda1	0	9.00E-02	0.286	
141	sda2	0	2.72E-02	0.163	
142	sda_all	0	1.15E-01	0.319	
143	DI_1	0	1.31E-02	0.114	
144	DI_2	0	6.03E-03	0.0774	

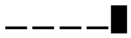
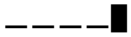
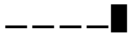
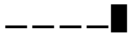
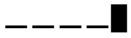
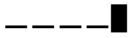
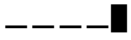
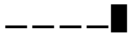
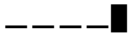
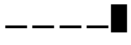
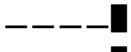
145	drill_instructor_sda	0	1.91E-02	0.137	
146	combatIns1	0	6.96E-03	0.0832	
147	combatIns2	0	4.28E-03	0.0653	
148	combat_instructor_sda	0	1.12E-02	0.105	
149	recruiter1	0	3.84E-02	0.192	
150	recruiter2	0	1.21E-02	0.109	
151	recruiter_sda	0	5.05E-02	0.219	
152	MSG1	0	1.45E-02	0.119	
153	MSG2	0	3.32E-03	0.0576	
154	msg_sda	0	1.78E-02	0.132	
155	security1	0	1.71E-02	0.13	
156	security2	0	1.48E-03	0.0384	
157	security_forces_sda	0	1.79E-02	0.133	
158	demoted	0	2.75E-03	0.0523	
159	time_in_grade	0	1.54E+01	16.6	
160	promo_competitive	0	4.63E-02	0.21	
161	pro_con_grade	0	2.20E+00	2.17	
162	pro_con_service	0	4.27E+00	0.674	
163	timetoeas	0	2.59E+01	14.3	
164	miss_timetoeas	0	1.57E-03	0.0396	
165	agesq	0	6.23E+02	337	
166	female	0	8.37E-02	0.277	
167	male	0	9.16E-01	0.277	
168	married	0	4.01E-01	0.49	
169	divorced	0	2.98E-02	0.17	
170	single	0	5.67E-01	0.495	
171	leg_separated	0	1.06E-03	0.0325	
172	race_white	0	5.66E-01	0.496	
173	race_black	0	9.52E-02	0.293	
174	race_hisp	0	4.30E-02	0.203	
175	race_mex	0	1.30E-01	0.336	
176	race_latin	0	3.18E-02	0.175	
177	race_declined	0	5.32E-02	0.224	
178	race_other	0	8.11E-02	0.273	
179	dep_change	0	1.37E-02	0.116	
180	dep_loss	0	2.41E-03	0.049	
181	dep_gain	0	1.13E-02	0.105	
182	marital_change	0	8.56E-03	0.0921	
183	citiz_change	0	6.52E-04	0.0255	
184	conus	0	9.24E-01	0.266	

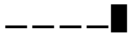
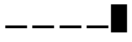
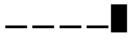
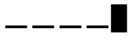
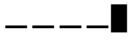
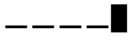
185	crisis_experience	0	3.57E-02	0.186	
186	combat_service	0	1.90E-01	0.392	
187	hor_county_	0	9.24E+02	560	
188	med_exam_meps_	0	3.71E+01	22.3	
189	mental_group_	0	6.52E+00	1.65	
190	parental_consent_	0	1.17E+00	0.559	
191	pay_grade_at_contract_	0	4.77E-01	1.07	
192	race_desc_	0	1.76E+01	5.45	
193	recruiter_rank_	0	2.25E+01	6.02	
194	rs_long_name_	0	2.83E+01	16	
195	ship_to_	0	3.31E+00	0.865	
196	educ_	0	3.61E+00	2.01	
197	i_addl_first_mos_assignment_dt_	0	2.88E-01	0.453	
198	i_addl_second_mos_assignment_dt_	0	1.23E-01	0.328	
199	i_anniversary_date_	0	1.80E-02	0.133	
200	i_apc_date	0	3.41E-01	0.474	
201	i_career_status_bonus_elect_dt_	0	7.51E-02	0.264	
202	i_civ_educ_left_school_dt_1_	0	9.83E-01	0.128	
203	i_cohesion_proj_train_compl_dt_	0	4.89E-01	0.5	
204	i_comp_score_computation_dt_	0	7.06E-01	0.456	
205	i_crisis_participation_begin_dt_	0	3.57E-02	0.186	
206	i_current_tour_begin_date_	0	1.00E+00	0.0169	
207	i_date_initial_entry_reserve_	0	1.76E-02	0.131	
208	i_date_joined_smcr_	0	1.09E-02	0.104	
209	i_dep_location_begin_date_	0	3.86E-01	0.487	
210	i_drop_discharge_unit_date_	0	1.81E-02	0.133	
211	i_expiration_obligated_service_	0	1.00E+00	0.00485	
212	i_expiration_of_active_service_	0	9.98E-01	0.0432	
213	i_exp_of_current_contract_	0	1.00E+00	0.00153	
214	i_extension_enlist_effect_date_	0	1.13E-01	0.316	
215	i_final_payment_flg_dt_	0	0	0	
216	i_flight_status_date_	0	1.30E-02	0.113	
217	i_from_date	0	3.41E-01	0.474	
218	i_geographic_location_began_dt_	0	9.81E-01	0.135	
219	i_good_conduct_medal_date_	0	9.95E-01	0.0719	
220	i_initial_active_duty_date_	0	1.78E-02	0.132	
221	i_intended_transfer_date_	0	5.83E-02	0.234	
222	i_law_enforce_ci_id_dt_	0	1.27E-02	0.112	
223	i_maternity_leave_elig_dt_	0	1.98E-03	0.0444	

224	i_mplp_leave_elig_dt_	0	0	0	
225	i_natl_call_service_waiver_date_	0	4.71E-06	0.00217	
226	i_officer_candidate_eff_dt_	0	1.22E-04	0.0111	
227	i_operation_effective_date_	0	2.49E-02	0.156	
228	i_permanent_rank_date_	0	1.00E+00	0.00307	
229	i_planned_reenl_ext_retire_date_	0	1.28E-02	0.112	
230	i_primary_mos_assignment_dt_	0	1.00E+00	0.00376	
231	i_prior_contract_exp_date_	0	3.46E-02	0.183	
232	i_prior_enl_acceptance_date_	0	3.13E-01	0.464	
233	i_prom_restricted_term_dt_	0	9.65E-02	0.295	
234	i_received_date	0	3.41E-01	0.474	
235	i_reserve_exp_current_contract_	0	6.48E-03	0.0803	
236	i_retirement_date_	0	4.00E-05	0.00633	
237	i_selected_grade_date_	0	4.57E-02	0.209	
238	i_service_spouse_effective_date_	0	7.13E-02	0.257	
239	i_special_duty_bonus_date_	0	6.66E-03	0.0813	
240	i_strength_cat_effective_date_	0	1.00E+00	0.0161	
241	i_to_date	0	3.41E-01	0.474	
242	i_accession_date_	0	9.28E-01	0.259	
243	i_activity_description_	0	9.58E-01	0.2	
244	i_additional_temporary_mcc_v2	0	1.83E-02	0.134	
245	i_addl_first_mos_code	0	2.90E-01	0.454	
246	i_addl_second_mos_code	0	1.26E-01	0.332	
247	i_addl_temp_reporting_unit_code	0	9.81E-01	0.135	
248	i_adverse_fitrep	0	3.41E-01	0.474	
249	i_afqt_score	0	9.58E-01	0.2	
250	i_asvab_version_	0	9.58E-01	0.2	
251	i_avg_this_rpt	0	3.31E-01	0.47	
252	i_belowcategory	0	3.34E-01	0.472	
253	i_billet_mos_v2	0	9.89E-01	0.104	
254	i_bonus_prgm_enl_for_cd_v2	0	5.43E-01	0.498	
255	i_career_status_bonus_elect_cd	0	7.51E-02	0.264	
256	i_career_status_flag_v2	0	3.13E-01	0.464	
257	i_citizenship_desc_	0	9.57E-01	0.204	
258	i_citzen_cntry_orig_geo_code_v2	0	9.99E-01	0.0381	
259	i_civ_ed_major_subject_cd_v2	0	9.98E-01	0.0422	
260	i_civ_educ_graduation_cd_v2	0	9.84E-01	0.127	
261	i_cohesion_future_mc_code_v2	0	2.74E-03	0.0522	
262	i_cohesion_number_id_v2	0	7.06E-06	0.00266	
263	i_college_major_	0	1.08E-01	0.31	

264	i_comd_mon_com_code_v2	0	9.88E-01	0.108	
265	i_cmd_recruiter_bonus_points	0	1.00E+00	1	
266	i_commendatory_fitrep	0	3.41E-01	0.474	
267	i_company_code_v2	0	9.88E-01	0.11	
268	i_contract_legal_agrment_v2	0	1.00E+00	0.0202	
269	i_crisis_remark_uniq_id_v2	0	2.38E-02	0.153	
270	i_crisis_remk_code_desc_tx_v2	0	2.38E-02	0.153	
271	i_current_county_	0	9.44E-01	0.231	
272	i_current_state_code_	0	9.49E-01	0.22	
273	i_current_zipcode_	0	9.51E-01	0.215	
274	i_custody_status_code_v2	0	7.48E-01	0.434	
275	i_depend_geo_location_code_v2	0	7.62E-01	0.426	
276	i_district_	0	9.58E-01	0.201	
277	i_dlab_score	0	5.79E-02	0.234	
278	i_dod_trn_cat_pay_group_v2	0	1.76E-02	0.131	
279	i_duty_status_code_v2	0	1.00E+00	0.00153	
280	i_education_code_at_contract_	0	9.57E-01	0.202	
281	i_education_tier_at_contract	0	9.57E-01	0.202	
282	i_el	0	9.58E-01	0.2	
283	i_fap_company_code_v2	0	1.62E-02	0.126	
284	i_fap_monitored_com_cd_v2	0	1.65E-02	0.127	
285	i_fap_platoon_code_v2	0	1.62E-02	0.126	
286	i_fap_reporting_unit_code	0	9.81E-01	0.135	
287	i_final_payment_flg_v2	0	0	0	
288	i_flight_status_identifier_v2	0	1.30E-02	0.113	
289	i_grade_select_v2	0	5.93E-02	0.236	
290	i_gt	0	9.58E-01	0.2	
291	i_home_of_rec_st_cntry_v2	0	9.97E-01	0.0534	
292	i_home_of_record_county	0	9.97E-01	0.0501	
293	i_home_of_selection_flg_v2	0	2.38E-03	0.0487	
294	i_hor_area_code_	0	4.60E-01	0.498	
295	i_hor_county_	0	9.38E-01	0.241	
296	i_hor_state_code_	0	9.43E-01	0.232	
297	i_hor_zipcode_	0	9.45E-01	0.228	
298	i_incurred_obligd_service_cd_v2	0	0	0	
299	i_indiv_loc_county_code_v2	0	1.00E+00	0.00614	
300	i_individual_loc_city_code	0	1.00E+00	0.00614	
301	i_ist_pull_ups	0	9.06E-01	0.291	
302	i_legal_act_gcm	0	6.37E-02	0.244	
303	i_legal_act_njp	0	6.37E-02	0.244	

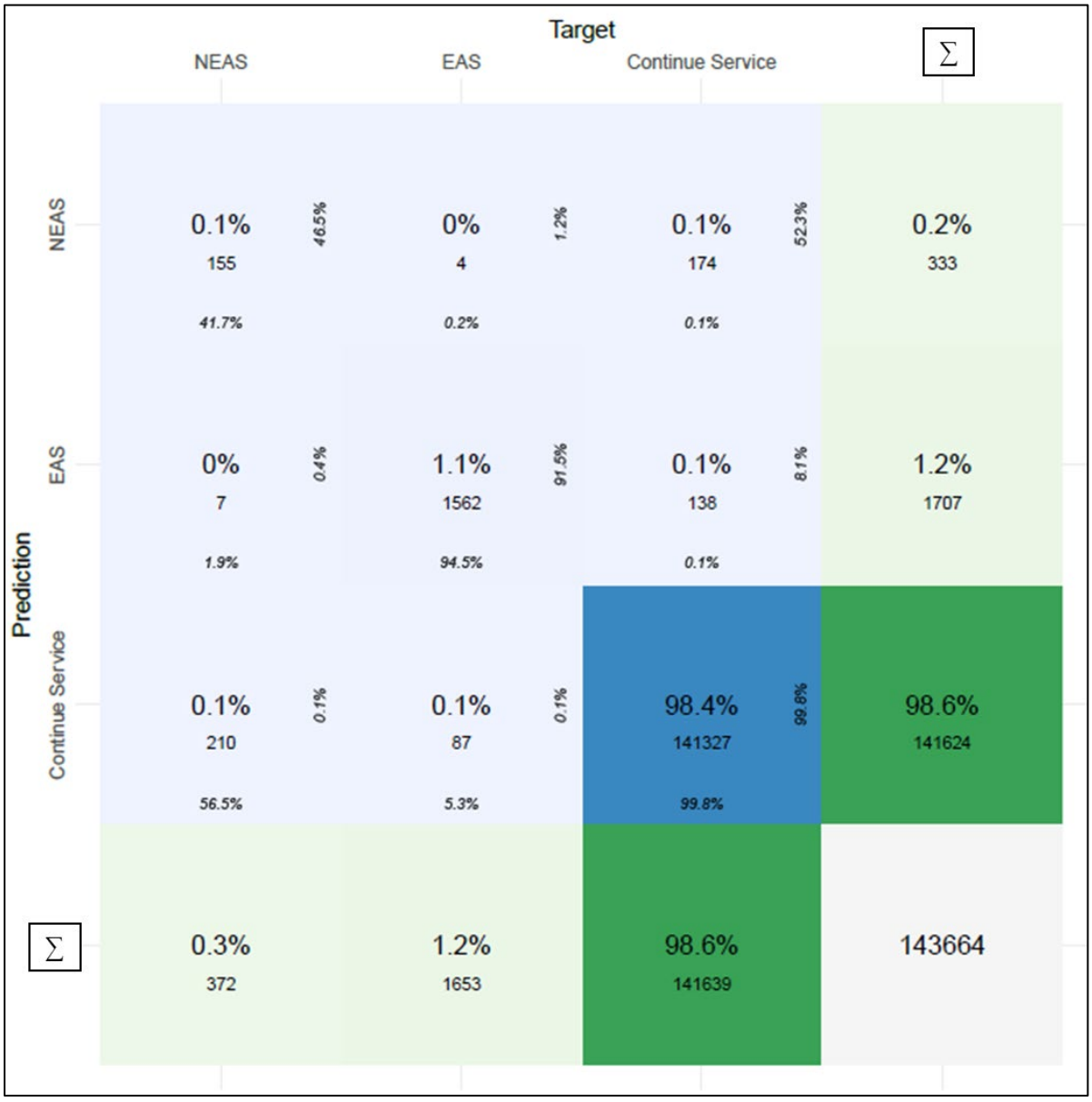
304	i_legal_act_scm	0	6.37E-02	0.244	
305	i_legal_act_spcm	0	6.37E-02	0.244	
306	i_magtf_designator_v2	0	8.92E-01	0.311	
307	i_maternity_leave_bal_qy	0	1.00E+00	1	
308	i_mcmap_above_blackbelt	0	9.17E-01	0.276	
309	i_mcmap_blackbelt	0	9.17E-01	0.276	
310	i_mcmap_brownbelt	0	9.17E-01	0.276	
311	i_mcmap_graybelt	0	9.17E-01	0.276	
312	i_mcmap_greenbelt	0	9.17E-01	0.276	
313	i_mcmap_instructor	0	9.17E-01	0.276	
314	i_mcmap_no_belt	0	9.17E-01	0.276	
315	i_mcmap_revoke	0	9.17E-01	0.276	
316	i_mcmap_tanbelt	0	9.17E-01	0.276	
317	i_med_exam_meps_	0	9.58E-01	0.201	
318	i_medical_disposition_	0	9.56E-01	0.205	
319	i_mental_group_	0	9.58E-01	0.2	
320	i_mobiliz_mont_cmd_cd_v2	0	1.30E-03	0.036	
321	i_mobilization_status_v2	0	3.17E-03	0.0563	
322	i_months_fitrep	0	3.41E-01	0.474	
323	i_mplp_caregiver_cd_v2	0	0	0	
324	i_natl_call_srv_waiver_code_v2	0	4.71E-06	0.00217	
325	i_not_recommended_for_promotion	0	4.43E-02	0.206	
326	i_num_fitrep	0	3.41E-01	0.474	
327	i_officer_candidate_cd_v2	0	1.13E-04	0.0106	
328	i_operation_identifier_v2	0	2.49E-02	0.156	
329	i_operation_montd_cmd_code_v2	0	2.49E-02	0.156	
330	i_operation_reporting_unit_code	0	2.49E-02	0.156	
331	i_parental_consent_	0	9.15E-01	0.279	
332	i_pay_grade_at_contract_	0	1.93E-01	0.395	
333	i_pft_first_class	0	9.51E-01	0.215	
334	i_pft_second_class	0	9.51E-01	0.215	
335	i_pft_third_class	0	9.51E-01	0.215	
336	i_phys_fit_crunch_scr_qy	0	9.51E-01	0.217	
337	i_phys_fit_pull_up_scr_qy	0	9.45E-01	0.228	
338	i_phys_fit_push_up_scr_qy	0	2.54E-02	0.157	
339	i_phys_fit_rowing_scr_tm	0	1.49E-03	0.0386	
340	i_phys_fit_run_scr_tm	0	9.51E-01	0.217	
341	i_phys_fit_score_qy	0	9.51E-01	0.215	
342	i_place_of_birth_v2	0	9.85E-01	0.123	
343	i_pland_reenl_ext_ret_flag_v2	0	3.07E-02	0.172	

344	i_plead_city_cd	0	3.36E-01	0.472	
345	i_plead_county_cd	0	3.36E-01	0.472	
346	i_plead_state_country_cd	0	2.35E-05	0.00485	
347	i_plead_zip_cd	0	2.35E-05	0.00485	
348	i_pres_grade_spec_rank_code_v2	0	9.39E-01	0.24	
349	i_program_element_number_v2	0	9.24E-01	0.265	
350	i_prom_restricted_status_cd	0	1.00E+00	1	
351	i_prospectiv_off_srce_cd_v2	0	3.39E-04	0.0184	
352	i_qsn_bonus_amount	0	9.58E-01	0.2	
353	i_race_desc_	0	9.58E-01	0.2	
354	i_rank	0	1.00E+00	0.00307	
355	i_rec_4_promote	0	3.41E-01	0.474	
356	i_recruiter_rank_	0	9.58E-01	0.201	
357	i_reenlistment_bonus_type_v2	0	1.90E-01	0.393	
358	i_religion_v2	0	9.99E-01	0.0253	
359	i_relm_recom_code_v2	0	3.17E-01	0.465	
360	i_relval_at_proc	0	3.41E-01	0.474	
361	i_relval_cum	0	3.41E-01	0.474	
362	i_res_mon_command_code_v2	0	4.80E-03	0.0691	
363	i_res_record_status_code_v2	0	5.27E-03	0.0724	
364	i_reserve_reporting_unit_code	0	1.31E-02	0.114	
365	i_responsibility_center_number	0	0	0	
366	i_rs_long_name_	0	9.58E-01	0.201	
367	i_sat_score	0	4.90E-03	0.0699	
368	i_selective_relm_bonus_zone_v2	0	1.89E-01	0.391	
369	i_selres_trans_prgm_code_v2	0	2.82E-05	0.00531	
370	i_separation_document_type_v2	0	1.95E-02	0.138	
371	i_separation_incentive_code_v2	0	2.19E-04	0.0148	
372	i_service_spouse_code_v2	0	8.74E-02	0.282	
373	i_ship_to_	0	9.58E-01	0.201	
374	i_source_code_	0	9.58E-01	0.2	
375	i_special_duty_bonus_points_qy	0	1.00E+00	1	
376	i_tad_company_code_v2	0	1.04E-01	0.305	
377	i_tad_platoon_code_v2	0	1.04E-01	0.305	
378	i_temporary_add_duty_excess_flag	0	1.42E-01	0.349	
379	i_temporary_mcc_v2	0	1.25E-01	0.331	
380	i_temporary_reporting_unit_code	0	9.81E-01	0.135	
381	i_term_of_enlistment	0	9.58E-01	0.2	
382	i_unit_id_cd_v2	0	0	0	
383	i_uppercategory	0	3.34E-01	0.472	

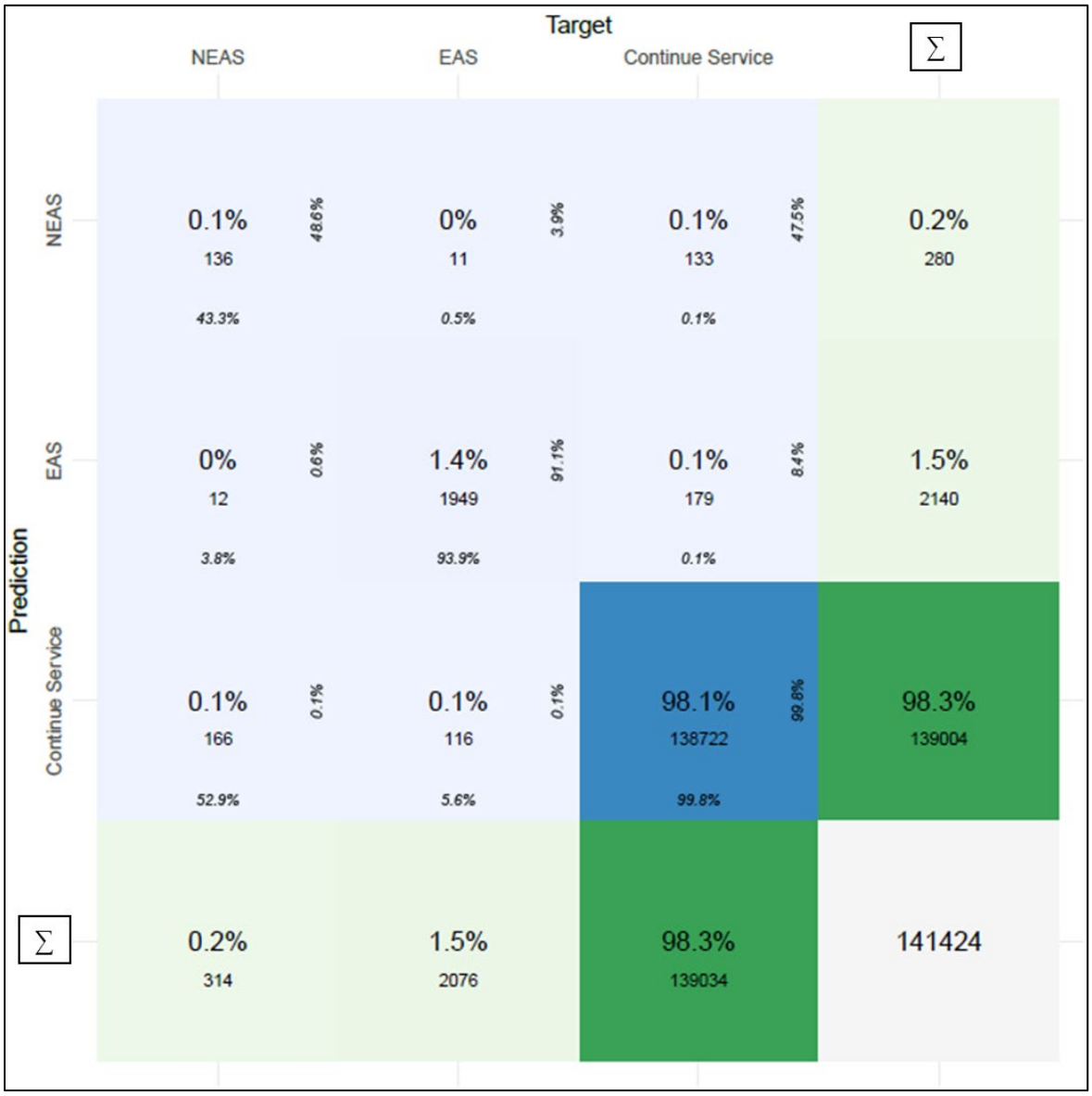
384	i_withcategory	0	3.34E-01	0.472	
385	i_educ_	0	9.97E-01	0.0557	
386	i_sda1	0	9.00E-02	0.286	
387	i_sda2	0	2.72E-02	0.163	
388	i_DI_1	0	1.31E-02	0.114	
389	i_DI_2	0	6.03E-03	0.0774	
390	i_combatIns1	0	6.96E-03	0.0832	
391	i_combatIns2	0	4.28E-03	0.0653	
392	i_recruiter1	0	3.84E-02	0.192	
393	i_recruiter2	0	1.21E-02	0.109	
394	i_MSG1	0	1.45E-02	0.119	
395	i_MSG2	0	3.32E-03	0.0576	
396	i_security1	0	1.71E-02	0.13	
397	i_security2	0	1.48E-03	0.0384	
398	i_timetoeas	0	9.98E-01	0.0432	
399	i_award_unit	0	9.61E-01	0.194	
400	i_award_service	0	9.61E-01	0.194	
401	i_award_personal	0	9.61E-01	0.194	
402	i_crisis_experience	0	1.00E+00	1	
403	i_combat_service	0	1.00E+00	0.00266	

B. RANDOM FOREST CLASSIFIER CONFUSION MATRICES

1. November Model

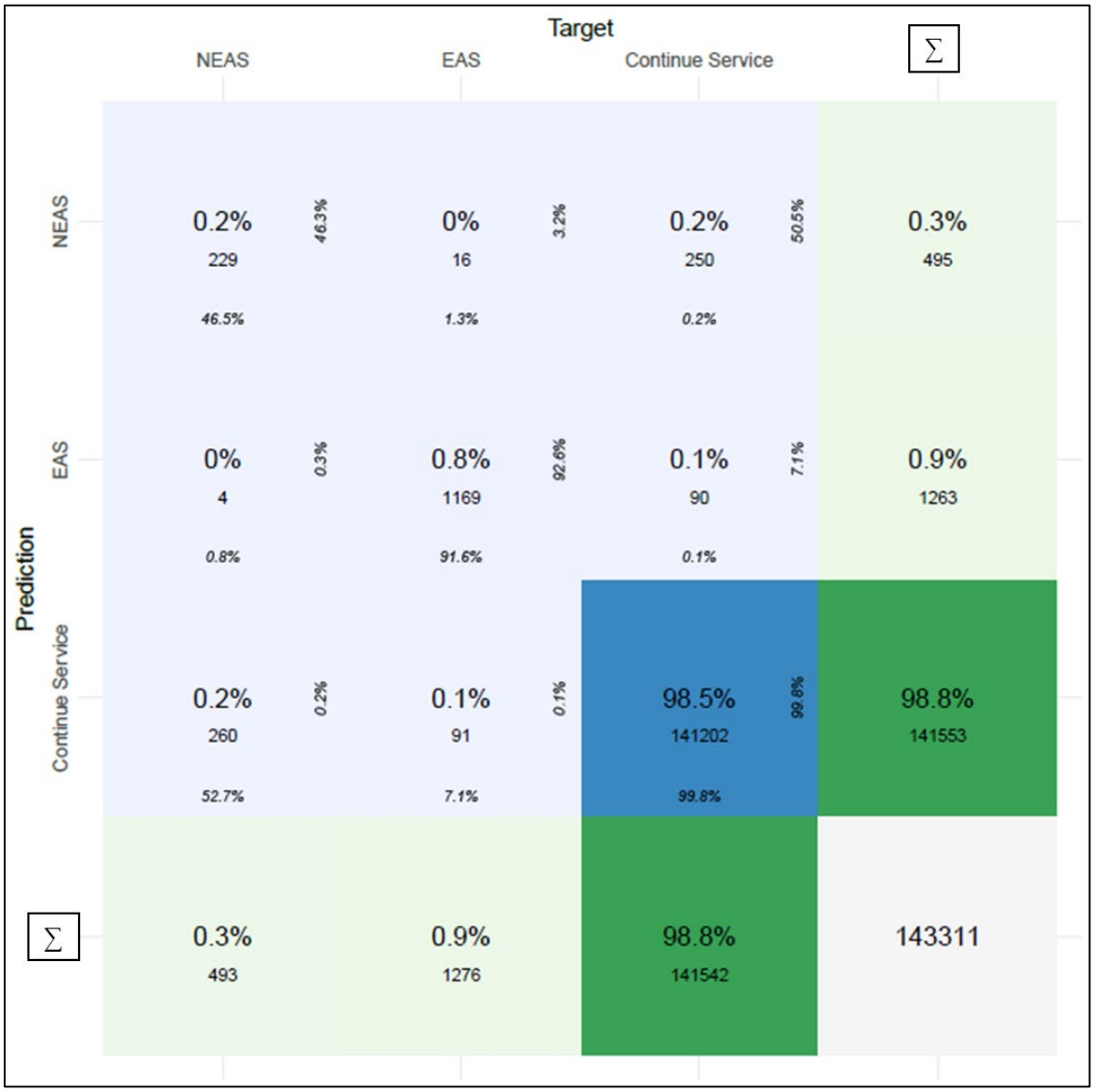


2. December Model

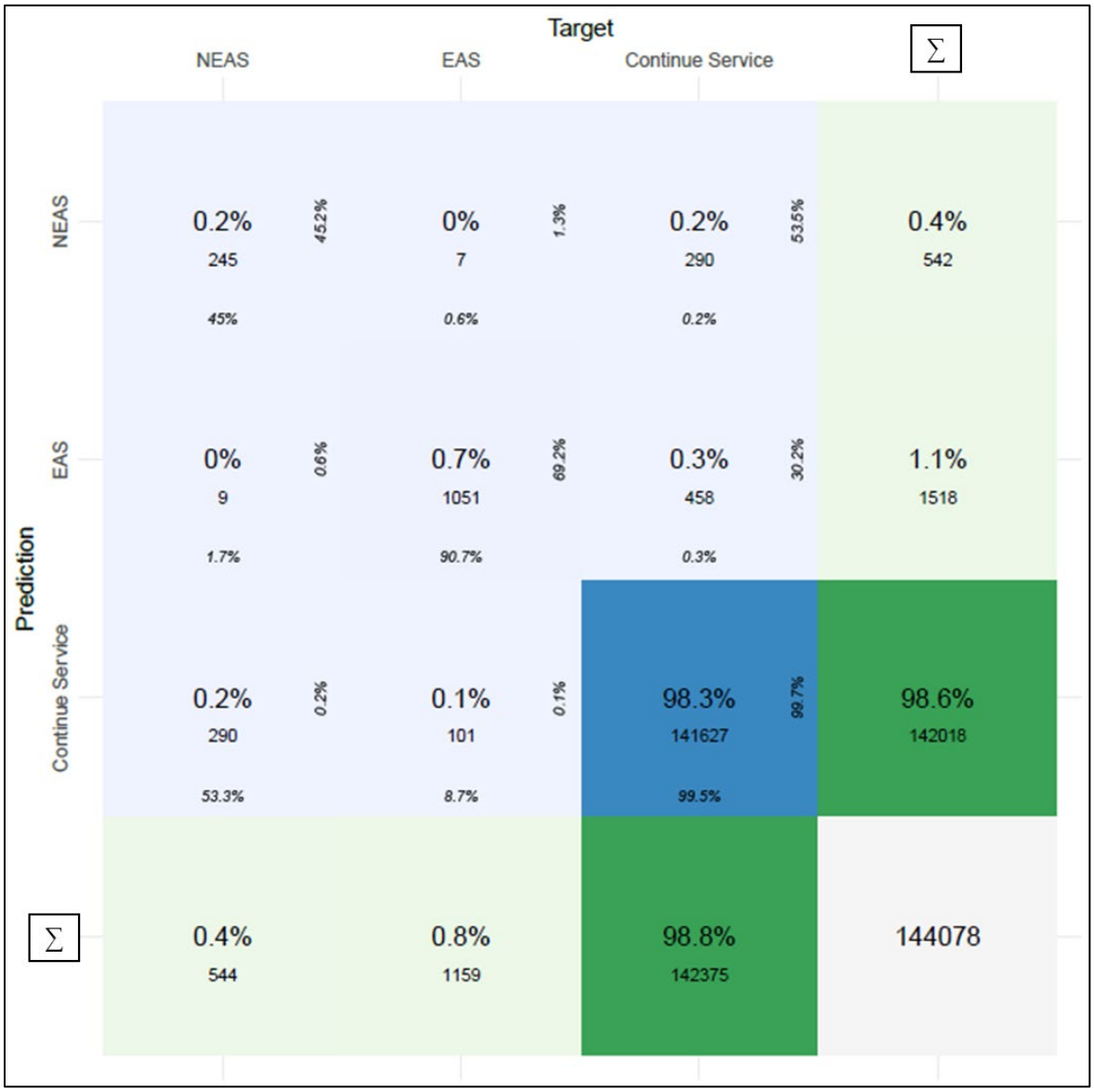


3. January Model

4. February Model

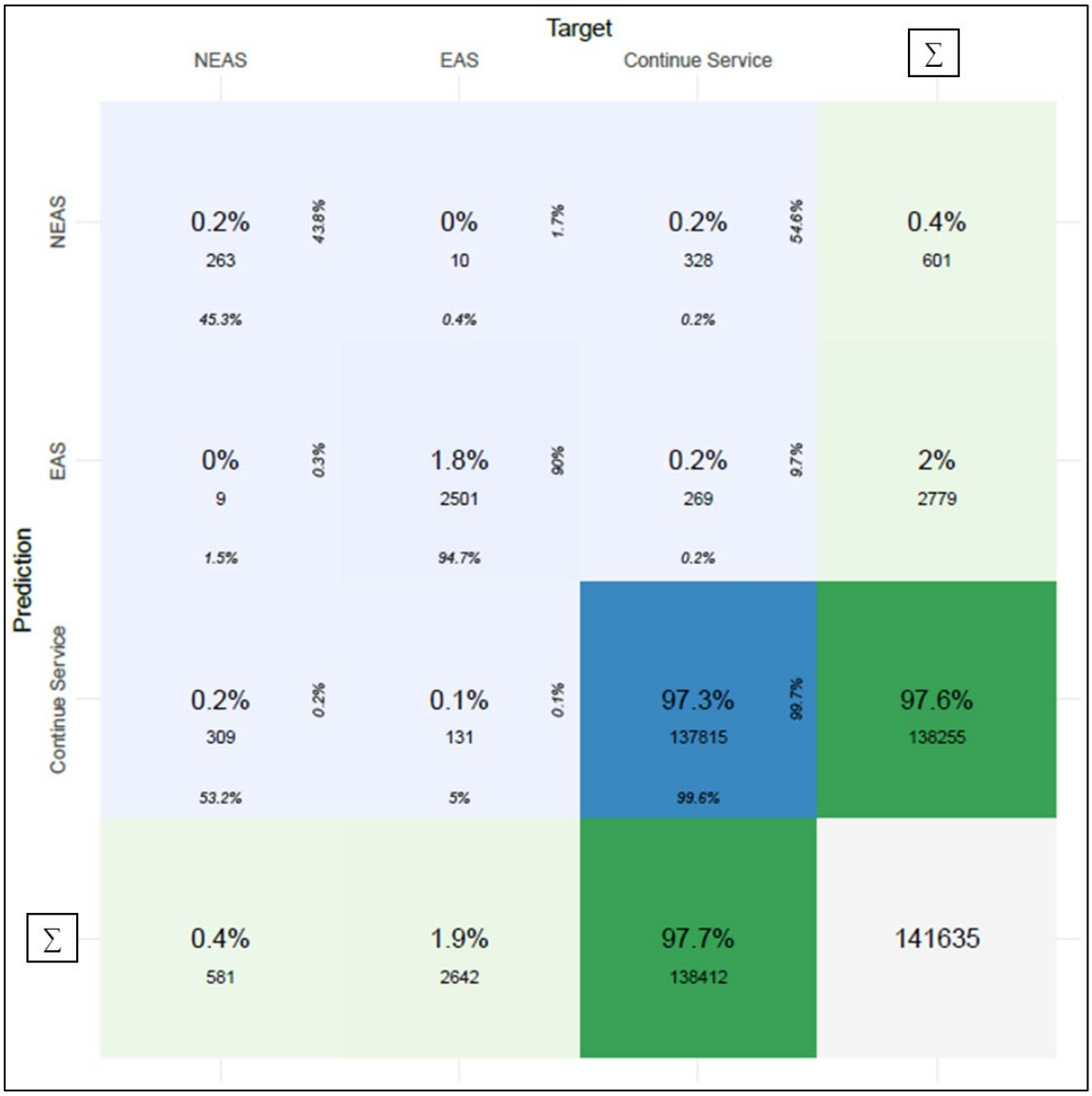


5. March Model



6. April Model

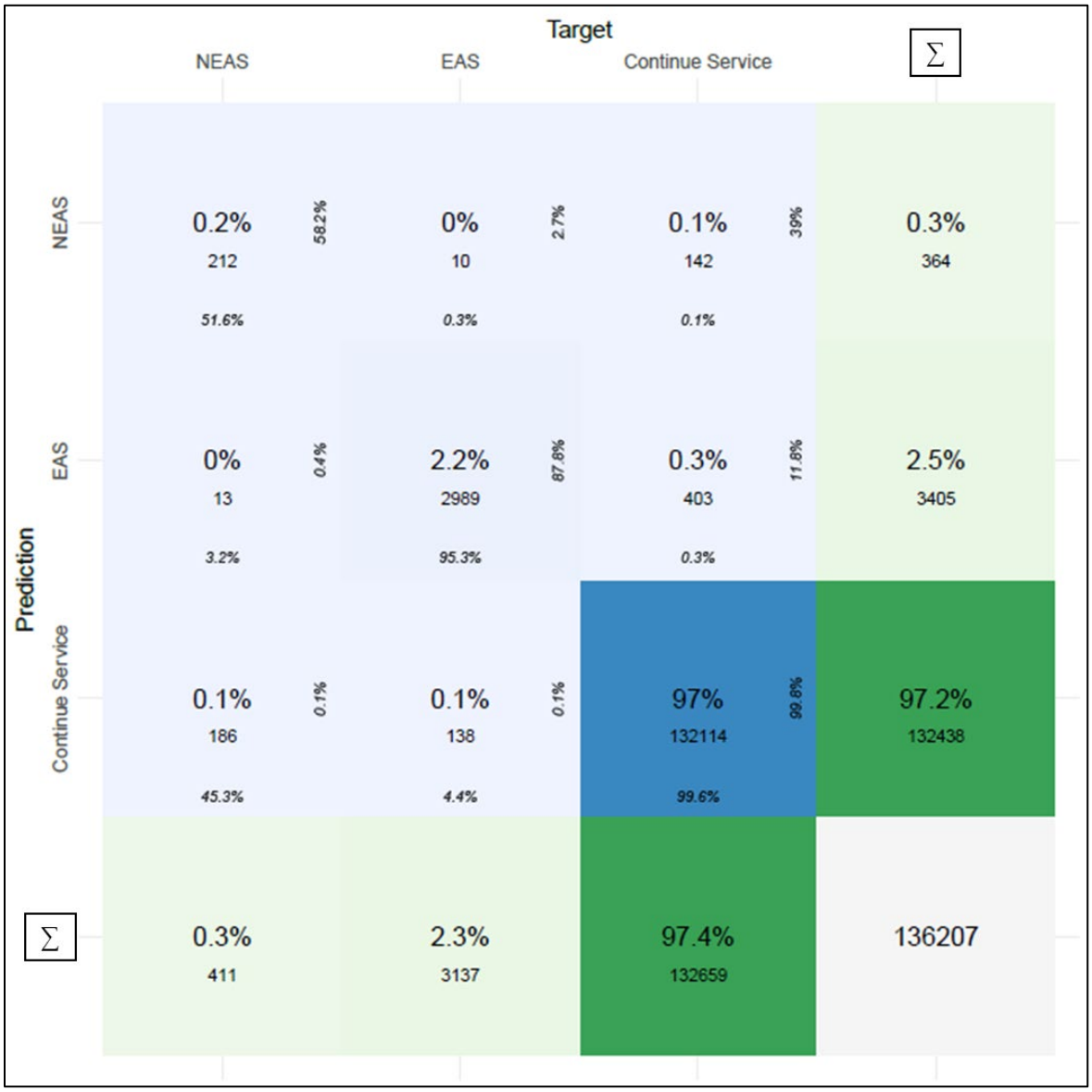
7. May Model


Target
NEAS
EAS
Continue Service
Σ
Prediction
NEAS
EAS
Continue Service
Σ
0.2%
263
43.8%
45.3%
0%
10
1.7%
0.4%
0.2%
328
54.6%
0.2%
0.4%
601
0%
9
0.3%
1.5%
1.8%
2501
90%
94.7%
0.2%
269
9.7%
0.2%
2%
2779
0.2%
309
0.2%
53.2%
0.1%
131
0.1%
5%
97.3%
137815
99.7%
99.6%
97.6%
138255
0.4%
581
1.9%
2642
97.7%
138412
141635

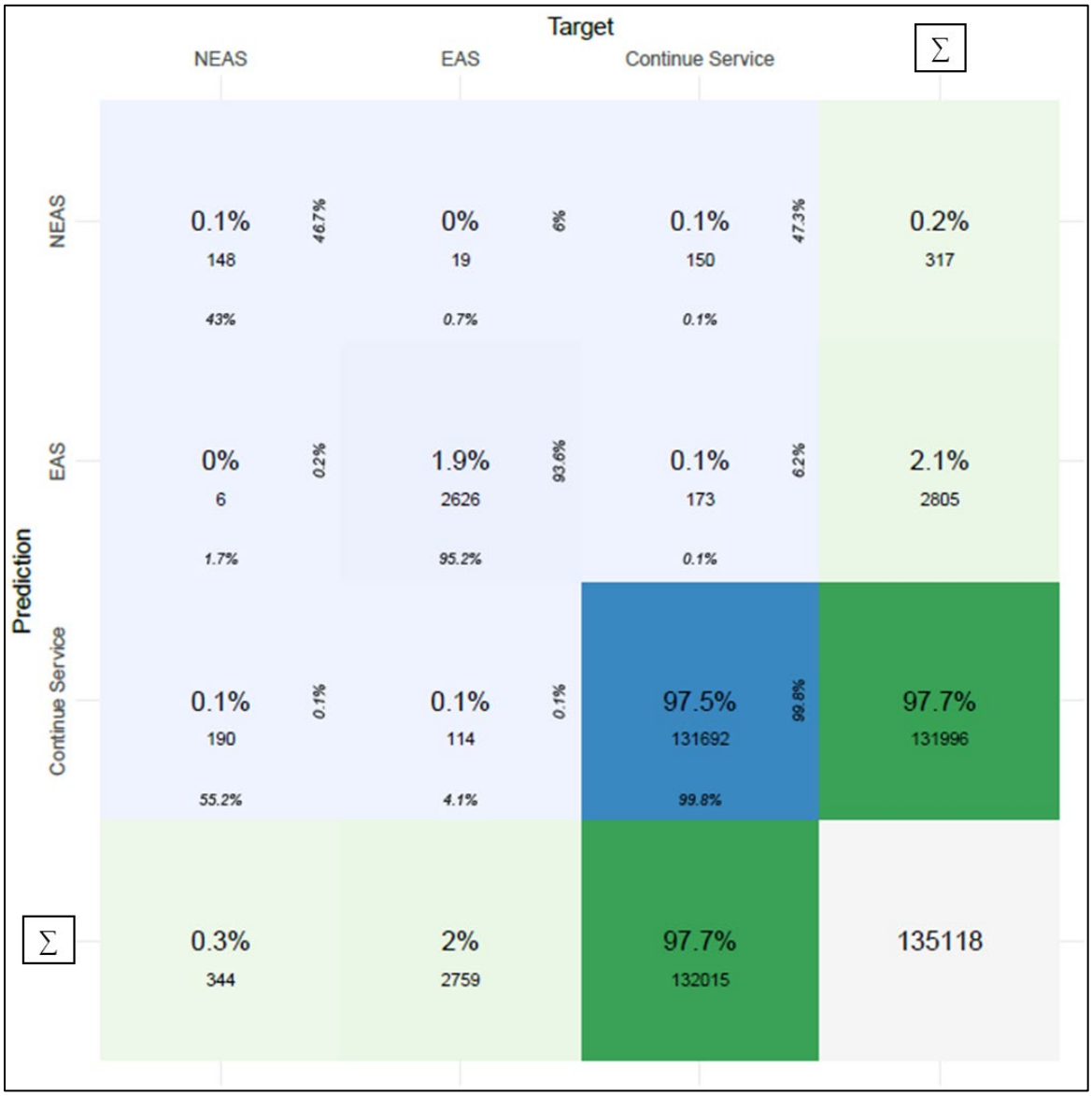
8. June Model


Target
NEAS
EAS
Continue Service
Σ
Prediction
NEAS
EAS
Continue Service
Σ
0.1%
192
46.9%
47.1%
0%
7
1.7%
0.3%
0.2%
210
51.3%
0.2%
0.3%
409
0%
15
0.6%
3.7%
1.6%
2182
91.9%
96%
0.1%
178
7.5%
0.1%
1.7%
2375
0.1%
201
0.1%
49.3%
0.1%
85
0.1%
3.7%
97.8%
134662
99.8%
99.7%
98%
134948
0.3%
408
1.7%
2274
98.1%
135050
137732

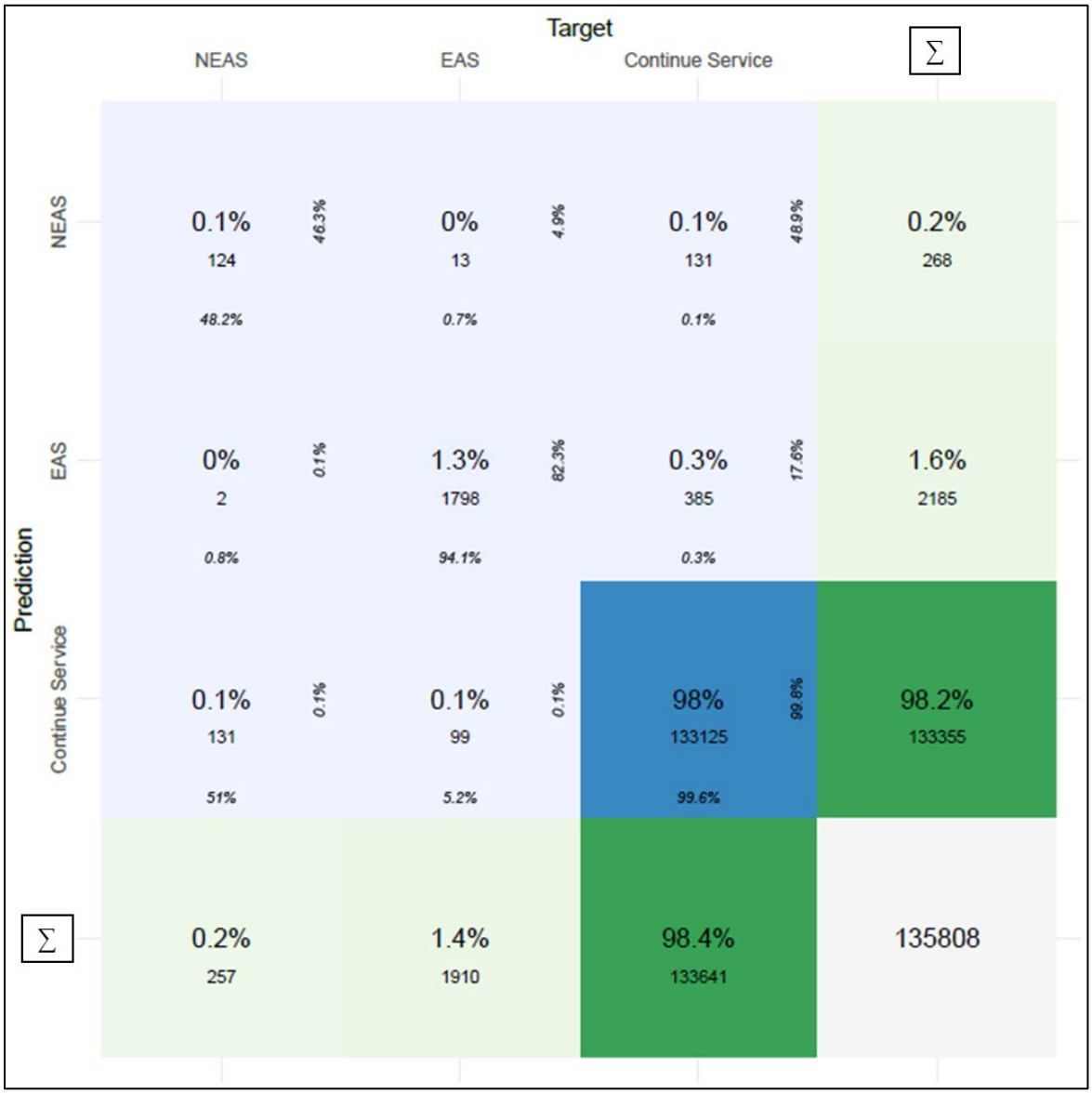
9. July Model


Target
NEAS
EAS
Continue Service
Σ
Prediction
NEAS
EAS
Continue Service
Σ
0.2%
212
58.2%
51.6%
0%
10
2.7%
0.3%
0.1%
142
39%
0.1%
0.3%
364
0%
13
0.4%
3.2%
2.2%
2989
87.8%
95.3%
0.3%
403
11.8%
0.3%
2.5%
3405
0.1%
186
0.1%
45.3%
0.1%
138
0.1%
4.4%
97%
132114
99.8%
99.6%
97.2%
132438
0.3%
411
2.3%
3137
97.4%
132659
136207

10. August Model



11. September Model



LIST OF REFERENCES

Alduayj, S. S., & Rajpoot, K. (2018). Predicting employee attrition using machine learning. *2018 International Conference on Innovations in Information Technology (IIT)*, 93–98. https://doi.org/10.1109/INNOVATIONS.2018.8605976

Berger, D. (2021). *Talent management 2030*. United States Marine Corps.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Buddin, R. J. (1984). *Analysis of early military attrition behavior*. Rand.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, *9*(4), Article 4. https://doi.org/10.3390/computers9040086

Gallagher, P. J. (2020). *Predicting Marine Corps retention behavior with machine learning* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. http://hdl.handle.net/10945/66074

Lemeshow, S., Sturdivant, R. X., & Hosmer, D. W. (2013). *Applied logistic regression*. Wiley. https://books.google.com/books?id=bRoxQBIZRd4C

Marrone, J. (2020). *Predicting 36-Month Attrition in the U.S. Military: A Comparison Across Service Branches*. RAND Corporation. https://doi.org/10.7249/RR4258

Marrone, J. (2021). *Organizational and cultural causes of army first-term attrition*. RAND Corporation. https://doi.org/10.7249/RR-A666-1

Marsland, S. (2009). *Machine learning, an algorithmic perspective*. Chapman & Hall/CRC.

National Defense Authorization Act for Fiscal Year 2022. (2021). Public Law 117. 81 Dec. 27, 2021.

Orrick, S. C. (2008). *Forecasting Marine Corps enlisted losses* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://hdl.handle.net/10945/4198

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

Terrazas, G. A. (2020). *Evaluation of machine learning applicability for USMC reenlistment* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://hdl.handle.net/10945/64888

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

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *ArXiv Preprint ArXiv:1508.04409*.

Zheng, A. (2015). *Evaluating machine learning models: A beginner's guide to key concepts and pitfalls*. O'Reilly Media.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California.