NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING INDIVIDUAL USNR ENLISTED ATTRITION

by

Bria N. Rand

September 2022

Thesis Advisor:	Samuel E. Buttrey
Second Reader:	Clark Petri

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2022	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE PREDICTING INDIVIDUAL USNR ENLISTED ATTRITION	**5. FUNDING NUMBERS**
6. AUTHOR(S) Bria N. Rand	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

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

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

13. ABSTRACT (maximum 200 words)

U.S. Navy Reserve sailors are trained to conduct critical operational missions and support the Active-Duty component. They also manage the administration and training of the Reserve program. Despite the importance of these personnel, in many recent years end-strength levels have not been met. This problem has arisen because the current end strength model has not accurately predicted these shortfalls. The variability in the accuracy of the attrition prediction input, a four-year weighted average, presents the difficulty of predicting Reserve attrition. While this thesis does not aim to replace the current aggregate model, it does aim to forecast individual attrition by using medical, administrative, and demographic factors to fit binary logistic regression models that predict whether a service member will attrit in the following year. This study differs from other individual attrition models in that they focus solely on first-term and early attrition that directly impacts recruiting. The results show that improvements to the model are required to increase accuracy. Inclusion of medical variables, as seen in prior theses, and inclusion of Navy Reserve specific variables may be beneficial to identify a subset of variables that can improve the model's predictive power.

14. SUBJECT TERMS reserve, manning, manpower	**15. NUMBER OF PAGES** 73
	16. PRICE CODE

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

NSN 7540-01-280-5500

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

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

PREDICTING INDIVIDUAL USNR ENLISTED ATTRITION

Bria N. Rand
Lieutenant Commander, United States Navy
BS, Jacksonville University, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2022

Approved by: Samuel E. Buttrey
Advisor

Clark Petri
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

U.S. Navy Reserve sailors are trained to conduct critical operational missions and support the Active-Duty component. They also manage the administration and training of the Reserve program. Despite the importance of these personnel, in many recent years end-strength levels have not been met. This problem has arisen because the current end strength model has not accurately predicted these shortfalls. The variability in the accuracy of the attrition prediction input, a four-year weighted average, presents the difficulty of predicting Reserve attrition. While this thesis does not aim to replace the current aggregate model, it does aim to forecast individual attrition by using medical, administrative, and demographic factors to fit binary logistic regression models that predict whether a service member will attrit in the following year. This study differs from other individual attrition models in that they focus solely on first-term and early attrition that directly impacts recruiting. The results show that improvements to the model are required to increase accuracy. Inclusion of medical variables, as seen in prior theses, and inclusion of Navy Reserve specific variables may be beneficial to identify a subset of variables that can improve the model's predictive power.

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

AAG	Army Analytics Group
AD	Active Duty
AGR	Active Guard Reserve
AUROC	Area Under the Receiver Operating Characteristics Curve
DEERS	Defense Enrollment Eligibility Reporting System
DMDC	Defense Manpower Data Center
DOD	Department of Defense
EAS	End of Active Service
ETP	Exception to Policy
FTS	Full Time Support
FY	Fiscal Year
JRIC	Joint Reserve Intelligence Center
IMR	Individual Medical Readiness
IRR	Individual Ready Reserve
LASSO	Least absolute shrinkage and selection operator
MOS	Military Occupational Specialists
MRRS	Medical Readiness Reporting System
NA	Not Applicable
NEAS	Non-End of Active Service
NRC	Navy Reserve Center
NPS	Naval Postgraduate School
NSIPS	Navy Standard Integrated Personnel System
PDE	Person Event Data Environment
PID	Personal Identifier
PII	Personally Identifying Information
PO2	Petty Officer 2nd Class
RCCPDS	Reserve Components Common Personnel Data System
SELRES	Selected Reserve
TAR	Training and Administration of the Reserve
TFDW	Total Force Data Warehouse

USMC	United States Marine Corps
USNR	United States Naval Reserve
VIF	Variance Inflation Factor

EXECUTIVE SUMMARY

U.S. Navy Reserve service members are trained to perform critical operational missions and support the Active-Duty component. They also manage the administration and training of the Reserve program. Despite their importance, in previous fiscal years end strength levels have not been met (Office of the Chief of Navy Reserve, Strength Planner, Personal Communication, 2022). A problem identified by reserve strength planners is that the current end strength model did not accurately predict these shortfalls and subsequent adjustments to accessions lagged behind the losses. The variability in the accuracy of the attrition prediction input, a 4-year weighted average, presents the difficulty of predicting reserve service member attrition. While this thesis does not aim to replace the current aggregate model, it does aim to forecast individual attrition by using medical, administrative, and demographic factors to fit a logistic regression model that will predict if a service member will attrit in the following year. This study differs from other individual attrition models in that they focus solely on first-term and early attrition that directly impacts recruiting.

Demographic, medical and unit-related factors provided in datasets within the Person-Event Data Environment (PDE) are used to build a binary logistic regression model that predicts attrition among enlisted Selected Reserve service members. The dataset includes 65,541 observations that are split into a training data set containing 70%, 24,823 observations, and a test set containing 30%, 10,718 observations. A full logistic regression model using 25 variables, a Least absolute shrinkage and selection operator (LASSO) model, and a logistic regression model using ten random-forest selected variables of importance are fit. The coefficients of the models show demographic variables as factors that are most important for predicting attrition with the LASSO logistic regression model and full logistic regression model, while unit and service-related variables are most important in the random forest model. A comparison of the performance of the models shows the LASSO logistic regression model as the best model. However, with suboptimal Area under the ROC curve (AUC) values, all models require significant improvement in order to be used in predicting enlisted Selected Reserve service member attrition.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

1. United States Naval Reserves (USNR)

The mission statement of the United States Naval Reserves is to provide strategic depth and deliver operational capabilities to the Navy and Marine Corps team and Joint Forces in times of peace or war (COMNAVRESFOR, 2022). These service members maintain readiness and training in order to fill both planned and unplanned gaps within the Active-Duty component. In addition, they manage the administration of the reserve program. They also provide manpower for operational units. These operational units retain organic equipment and manning and mobilize as a distinct unit. As shown in Figure 1, the Navy Reserve consists of the Ready Reserve, Standby Reserve, and Retired Reserve. The Ready Reserve includes the Selected Reserves (SELRES) and Individual Ready Reserves (IRR). The SELRES are drilling reserve service members who are required to attend one weekend per month and two weeks a year at a Navy Reserve Center (NRC), squadron or Joint Reserve Intelligence Center (JRIC). This group also includes the Training and Administration of the Reserves (TAR), formerly Full-Time Support (FTS). TAR are indistinguishable from Active Duty sailors in that they perform full-time service. IRR have previous training, but no obligation to drill. The Standby Reserves is a much smaller component. This group includes personnel who are temporarily assigned for reasons such as medical issues, disciplinary actions, failure to maintain security clearances, or being key employees in the workforce. The Retired Reserves includes inactive reserve service members who are eligible for retired pay. At the end of fiscal year (FY) 2021, there were 96,297 Ready Reserve service members, 1,075 Standby Reserve service members, and 28,930 Retired Reserve service members (DMDC, 2022).

Reserve service members are accessed via the same methods as Active-Duty members. For Officers, commission is of indefinite duration. Discharge, after completion of the initial eight-year military service obligation (MSO), must be requested. However, additional obligation can be incurred following acceptance of financial incentives or

secondary education. For enlisted members, initial contract length varies depending on additional factors such as component or prior service. Commissioned officers and enlisted service members share methods of attrition. This can include planned reasons like retirement, resignation, and transfer to active duty, or unplanned reasons like medical or disciplinary.

NAVY RESERVE CATEGORIES BASED ON RESERVE STATUS

<table>
<tr><td colspan="4">ACTIVE STATUS
*Officers are members on the Reserve Active Status List (RASL)</td><td>INACTIVE STATUS</td><td>RETIRED STATUS</td></tr>
<tr><td colspan="3">READY RESERVE</td><td rowspan="2">S-1 Standby Reserve Active

Key Federal Employee - Hardships</td><td rowspan="3">S-2 Standby Reserve Inactive

Can't earn Retirement points or promote</td><td rowspan="3">Qualified for Non-regular Retirement or Regular Retirement</td></tr>
<tr><td>SELRES (Selected Reserve)</td><td colspan="2">IRR (Individual Ready Reserve)</td></tr>
<tr><td>Drilling Reservists (With Pay)

**
- Full-Time Support
- Canvassing Recruiter
- Active Duty Recall</td><td>VTU (Voluntary Training Unit)

** Drilling Reservists (Non-Pay)</td><td>ASP (Active Status Pool)</td></tr>
<tr><td colspan="3">Qualifying Years of Service via approved Correspondence Courses, Non-Pay Drills (S-1 in VTU) ADT without pay (ASP and S-1)</td></tr>
</table>

*Member on the RASL are eligible for promotion
*Enlisted members are eligible for advancement while a SELRES or in the VTU and not subject to HYT.

Figure 1. Navy Reserve Categories. Source: NAVPERSCOM (2005).

2. USNR Current Attrition Modeling

A current model used by USNR strength planners to forecast fiscal year end strength utilizes four-year weighted averages of attrition and accession. Table 1, NAVPERSCOM (2005), shows the model's attrition prediction compared to the actual attrition observed for FY 21. Attrition was underestimated by 534 servicemembers. Table 2 shows the performance of the current USNR strength planner's model for preceding fiscal years.

Table 1. FY21 Monthly Attrition Prediction vs. Observed Attrition. Adapted from: Office of the Chief of Navy Reserve, Strength Planner, Personal Communication (2022)

RPN Total	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	Total
OPLAN[a]	763	713	426	805	666	663	703	641	678	628	658	818	8,162
FY21 Observed[b]	783	575	586	701	600	725	935	710	680	583	912	906	8,696
Delta Year to Date	+20	− 138	+160	−104	−66	+62	+232	+69	+2	−45	+254	+88	+534

[a]OPLAN – Model Prediction of Attrition for each month.

[b]Observed– Actual Attrition Values recorded each month.

(+) Recorded attrition higher than model prediction.

Table 2. FY17 – FY21 Attrition Prediction vs Actual Attrition. Adapted from: Office of the Chief of Navy Reserve, Strength Planner, Personal Communication (2022).

FY	Oplan Total	FY Observed	Delta
17	8611	8909	+298
18	8590	8709	+119
19	8459	7970	−489
20	8129	7914	−215
21	8162	8696	+534

In three of the past five years, FY17, FY18, and FY21, attrition was underestimated. Accurate predictions of attrition are important in this sense because they are being used as an input in the end strength model.

Underestimates of attrition means that the Reserve Component loses more service members than expected. The risk of underestimates is that end strength, as mandated by Congressional controls, is not met. This impedes the ability of the Navy to provide strategic depth. Overestimates of attrition means that the Reserve Component loses fewer members than expected. This results in over-execution of end strength. This may require reallocation of resources from other programs.

Another source that provides projections of USNR losses is The Defense Manpower Data Center (DMDC). These monthly projections are based solely on the member's Expiration Term of Service (ETS) (DMDC, 2022). This means they do not account for any unexpected losses.

B. PURPOSE

The variance in prediction accuracy throughout each month and through each fiscal year hints at the inherent difficulty of predicting servicemembers' attrition. While this thesis does not focus on creation of a more accurate end-strength model, it aims to forecast individual attrition more accurately than the models in place, by comparing logistic regression techniques that predict a binary response: whether a service member will or will not attrit in the next year.

C. LITERATURE REVIEW

Literature and previous studies specific to the reserves among branches focus mostly on aggregate attrition models, Markov models, and smoothing models, and the effectiveness in forecasting end-strength. The studies discussed below solely on first-term attrition and do not include Reserve Sailors; however, they provide relevant information for a starting point and for building a model. This thesis focuses on USNR attrition occurring among all personnel specifically within one year from the date the data set was taken.

1. United States Army (USA)

Various methods have been used to predict individual attrition in the military. A series of studies was completed by Naval Postgraduate School (NPS) students with the aim of addressing the Army reduction of its 2018 recruiting goals. This reduction required insight to improve retention and provide strategies to mitigate first-term attrition. The goal of the thesis completed by Speten (2018) was to identify demographic and administrative factors of enlisted soldiers to create logistic regression models that could effectively predict this attrition. It analyzed Army soldiers that enlisted between 2005 and 2010. The results showed that the most significant predictors of attrition for this group of soldiers are the duration of initial contract longer contracts increasing risk of attrition and deployment history, where longer deployments increase risk of attrition. The two logistic regression models in this study performed similarly on a test-set with an accuracy of roughly 80%. A shortfall noted by Speten is that his study did not include race or ethnicity, as he chose to remove the variable due to the large percentage of missing entries. Another thesis, conducted by Gobea (2019), builds on Speten's study by including medical factors such as presence of medical conditions, hearing and dental classes, height, and weight. With this addition, Gobea finds the most important variables to be deployment information, with non-deployable status increasing risk of attrition, contract duration, with longer length increasing risk of attrition and Dental Class 4. His study compares a lasso - logistic regression model and a random forest - logistic regression model. In this case, both models

perform similarly with an accuracy of 90%, which is a 10% improvement from Speten's model.

A third NPS student, Cammack (2020), adds on to the study by including the most recent time-varying factors. An example is that Gobea's study utilized weight at initial enlistment, while Cammack utilized the most recent weight on file. This impacted medical variables as well as demographic variables such as marital status. His study also included Periodic Health Assessment (PHA) data. Cammack excludes Dental Class 4, although it was identified as a variable of importance from Gobea (2019) and attributes this significance to issues outside service member's control. The reasoning excludes it as a good indicator of actual dental health. A similar approach is taken in building Cammack's models, but this study separates analysis by year in contract, resulting in 6 models. Across all years, contract duration and previous service maintain variable importance. Her analysis also showed that as the length of the contract increased, demographic variables lose importance, while medical variables gain. The authors noted that limitations of the studies include the assumption of data accuracy, the methods for handling missing data, and the fact that they include only active-duty soldiers.

2. RAND Corporation

A military attrition study was published by RAND Corporation in 2020 (RAND, 2020). This study, sponsored by Office of the Secretary of Defense, was conducted to address the costs imposed on all the services that are incurred when service members attrit prior to completion of first-term. The study builds on past research similar to the theses mentioned above. The data analyzed was provided by the DMDC and included those servicemembers who entered service between FY 2002 and 2013. The study follows each member out to the end of the initial service contract, assumed in this case to be 36 months. Demographic, administrative, and medical variables are used but an addition unique to this study is the inclusion of economic variables. The analyst builds a probit regression model that predicts the binary outcome; attrit or non-attrit within the 36-month period. The difference in the RAND study is that it compares attrition across the four services. The study showed that the economic and medical variables had the least ability to effectively

distinguish between attrit and non-attrit. These were most useful when combined with the demographic and administrative variables. The model itself across the four services could predict first-term attrition only 60% percent of the time. This means that 40% of first-term attrition was either based on unobservable variables or variables that occur or change after accession. The recommendation for future work and policy was to focus more on a population-level approach instead of focusing on individual recruit characteristics for retention.

3. United States Marine Corps (USMC)

Another NPS student, Orrick (2008), conducted a study pertaining to USMC active-duty attrition. More specifically, it differentiates between End of Active Service (EAS) and non-end of active service (NEAS) losses and credited a logit regression model for predicting the latter to compare to the weighted average model in use. NEAS losses include recruit losses, retirement losses, at category losses. Category losses include "Convenience of the government, physical disability, misconduct, unsatisfactory performance, deserter status, and death" (Orrick, 2008). The significance is that they are unplanned and at the time, NEAS losses accounted for 46% of total USMC losses.

Orrick uses data from the Total Force Data Warehouse (TFDW) that includes enlisted accessions from 1997 to 2007. The model forecasts losses by comparing EAS to NEAS losses and performs with 76% prediction accuracy. Shortfalls of the study are that it may be based on a misrepresentation of the population in that the data set was reduced from 500,000 observations to 167,000 due to missing separation codes. Because this variable was used as the dependent variable all observations with missing entries had to be removed from the data set. Another shortfall of this study was that it failed to include female service members. Recommendations from the study are that future studies may benefit from inclusion of variables such as unemployment rates and fitness report data. It also recommends inclusion of military occupation variables, which were removed due to missing values.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA AND METHODOLOGY

A. DATA

1. PDE

This thesis utilizes the Person-Event Data Environment (PDE) for all data collection and analysis. The PDE is an Army and Department of Defense (DOD) platform that provides a repository of information pulled from numerous databases. These databases contain information on all active-duty service members, reserve service members, and DOD employees. Each data set is scrubbed for variables containing personally identifiable information (PII). This information is encoded according to Army Analytics Group (AAG) standard operating procedures and requires additional approvals for usage. Social security numbers are removed and replaced with a 12-digit PDE person identifier (PID). The PDE PID creates the ability to link data sets from different databases and to ability to track a person through time, without identifying who they are. Each request for information goes through multiple steps of approval, including a Human Protections Manager. A justification for data use must be submitted and approved by the data owner. Once approved, data is provisioned for analysis. Access to data with PDE is limited to a predetermined time frame and any exports of data from the PDE require additional approval. All analytic tools and data are contained within CITRIX, a virtual desktop. This is another measure in place to protect the information being used. As noted in Gobea (2019), a drawback of this standalone system also seen in this project is that analysts are subjected to planned and unplanned maintenance periods, lengthy delays in data approvals, and required routing of documentation such as Data Use Agreements and Exception To Policy (ETP). Clarification on variable definitions may require additional time as not all data owners provide this upfront.

2. Datasets

The datasets used in this project came from multiple sources. The first is the Reserve Components Common Personnel Data System (RCCPDS) Master. It contains administrative information and serves as an inventory for all USNR service members. The

data is pulled from the Defense Enrollment Eligibility Reporting System (DEERS) and is owned by the Defense Manpower Data Center (DMDC). Updated data is uploaded to PDE every four months. This updating process is referred to as a "snapshot."

The next data set used is the RCCPDS Transaction. This data is also specific to USNR members and provides information specific to personnel transactions. Transactions include entrance, separation, transfer, reenlistment, and retirement information. The Reserve Duty Family is the last RCCPDS data set used. It provides the number of servicemember's dependents and children, as well as their ages. However, this data set is only updated once a year in March. The next data sets were taken from the Navy Medical Readiness Reporting System (MRRS) data base. The Postpartum data set provides all pregnant female service members' expected due dates and postpartum dates. The Individual Medical Readiness (IMR) data set provides medical variables such as allergies, blood type, etc. Lastly, the Personnel Medical File lists known conditions, dental class, and audio/visual status. To note, all data sets provided from MRRS were updated only once in 2019, and updates are not currently maintained. Data pertaining to drill attendance, active-duty service, mobilization requests, civilian occupations, and waivers was available in the Navy Standard Integrated Personnel System (NSIPS). Similar to the MRRS data, the NSIPS file was last updated in 2019 but being that it provided information specific to reserve service members it was imperative to include in the study.

Data was pulled from each of the mentioned data sets on the snapshot date closest to September 2018. This was done to get the most recent data that would have medical and NSIPS information available. The MRRS and NSIPS data was extracted from the March 2019 snapshot but only included updates up until September 2018. The RCCPDS Master File snapshot extracted was from September 2018. The datasets mentioned were merged using the PDE_PID as the common variable. The RCCPDS Master file and Transaction file from September 2019 were used for creation of the response variable.

B. METHODOLOGY

The initial data set contained 137,638 people and 81 variables. To start data cleaning, all variables with only null entries were removed. All observations for Officers

were removed, as the focus was on enlisted attrition. This study focuses solely on SELRES. This is because the number of SELRES members identified by code assignment in PDE closely mirrored that in the DMDC reporting system. In addition, variables missing more than 30% of observations were removed. This included variables such as AFQT (ASVAB) score, accession source, and many medical entries. The resulting data set included 35,541 observations and 25 variables. Though not all variables will be discussed in detail, Table 3 provides additional information for each. This study includes both constant and time-varying variables.

Table 3. Variable Definitions

Variable	Type	Definition	Levels
RANK_PDE	Categorical	Servicemember Rank	1. SR 2. SA 3. SN 4. PO3 5. PO2 6. PO1 7. CPO
ASG_UNT_LOC_ISO_A3_CTRY_CD	Categorical	Assigned Unit Country Code	1. Africa 2. Asia 3. Europe 4. Latin America and the Caribbean 5. North America 6. Middle East
ASG_UNT_LOC_US_ST_CD	Categorical	Assigned Unit State Code	1. Northwest Everett 2. Southwest San Diego 3. Mid-Atlantic Great Lakes 4. Southeast Fort Worth 5. Mid Atlantic Norfolk 6. Southeast Jacksonville 7. Not US
ASG_UNT_NV_ASHR_AFLT_CD	Categorical	Navy Ashore Afloat Code. Type of duty assigned.	1. Shore Duty 2. Sea Duty-CONUS Ships 3. Non-rotated Sea Duty – Ships Homeported Overseas

EDU_LVL_CD	Categorical	Represents the highest level of education that a person has attained.	1. No HS Diploma 2. HS Graduate 3. Some College 4. Bachelor's Degree 5. Master's Degree 6. Doctorate Degree
ETH_AFF_CD	Categorical	Ethnic Affinity Code. Self-identified cultural background.	1. AJ Asian Descent 2. AK Hispanic Descent 3. AR US or Canadian Indian Tribes 4. AV Other Pacific Island Descent 5. ZZ Other/ None/Not Applicable
FAITH_GRP_CD	Categorical	Faith Code.	1. Indigenous/Christian 2. Western/Christian 3. Eastern/Christian 4. Restorationist/Christian 5. Non-Denominational/Christian 6. Fundamentalist 7. Islam 8. Judaism 9. Buddhism/Hindu/Bahai 10. Pagan 11. Atheist 12. Agnostic 13. None
MRTL_STAT_CD	Categorical	Service member marital status.	1. M, Married 2. N, Never married, Unknown 3. D, Divorced 4. L, Legally separated 5. A, Annulled, Divorced 6. W, Widow or widower
PN_AGE_QY	Numeric	Servicemember Age	
PN_SEX_CD	Binary	Sex Code	0. Male 1. Female

RACE_CD	Categorical	Race	1. C. White 2. M. Asian or Pacific Islander 3. N. Black 4. R. American Indian or Alaskan Native 5. X. Other, Unknown
RSBI_TYP_CD	Categorical	The type of bonus or stipend for service member who is appointed, enlists, reenlists, affiliates, or extends in a Reserve Component Incentive Program.	1. A. Enlistment bonus (3-yr, prior service) 2. B. Enlistment bonus (6-yr, prior service) 3. C. Enlistment bonus (6-yr, non-prior service) 4. E. Reenlistment bonus (3-yr, SELRES) 5. F. Reenlistment bonus (6-yr, SELRES) 6. Z. Unknown/Not Applicable/No bonus or Stipend
RSV_RET_ELIG_NTFCN_IND_CD	Binary	The code represents whether certain criteria are met for eligibility of retirement pay.	0. No, Not Applicable 1. Yes
RSV_RET_ELIG_SVC_YR_QY	Numeric	The number of qualifying retirement years.	
RSV_RET_PT_EARN_CRER_QY	Numeric	The number of retirement points earned.	
US_CTZP_STAT_CD	Binary	Indicates whether a person is a US Citizen.	0. No 1. Yes

PREGNANCY	Binary	Expected date of delivery for expecting service members between 30-Sep-17 and 30-Sep-19.	0. No 1. Yes
ALLERGY_CD	Binary	Servicemember documented allergies	0. No documented allergies 1. Has documented allergies
DEPLOYING_FLAG	Binary		0. No, N/A 1. Yes
TOTAL_DEP_QY	Numeric	Number of service member dependents; includes children, spouse, and others.	
ATTRIT	Binary	Service member separated prior to next year	0. No 1. Yes
SPD_CD	Categorical	Defines the reason for service member separation	1. Unknown 2. Expiration of Service 3. UNSAT Performance 4. Drugs/Alcohol 5. Medical 6. Early Release 7. Other
AFMS_YR_QY	Numeric	Length in years a service member has been in USN, includes active-duty service	
PRI_DOD_OCC_CD	Categorical	Job type within the USNR	1. Combat Systems 2. Builders 3. Administration 4. Medical

			5. Security 6. Intel 7. Other 8. Services 9. Aviation 10. Specwar/EOD 11. Surface Nav 12. Seamanship 13. Unqualified
SVC_AGMT	Numeric	Length of current contractual obligation	
ACS_SCRTY_CLRNC_CD	Categorical	Security Clearance Code	1. N, No Clearance 2. C, Confidential 3. S, Secret 4. T, Top Secret 5. Y, None

1. Variables

a. Categorical

The data set consists of six numeric, six binary, and thirteen categorical variables. The difficulty in including the categorical variables was that many have large numbers of possible outcomes. To manage this, the variable outcomes were grouped for easier implementation into analysis platforms and to increase the number of observations per category. The 51 categories for unit state locations were grouped by US Navy Reserve force breakdown; Northwest Everett (AK, WA, OR, IO, WY, MT, ND, SD, NE, MN, IA), Southwest San Diego (GU, CA, NV, AZ, UT, NM, CO, HI), Mid-Atlantic Great Lakes (WI, IL, IN, KY, OH, WV, PQ), Southeast Fort Worth (TX, OK, KS, MO, AR, LA, MS), Mid Atlantic Norfolk (NC, VA, MD, DE, NJ, NY, CT, RI, MA, NH, VT, ME), Southeast Jacksonville (TN, AL, GA, SC, FL). Unit country code was simplified grouping by region; North America (PR, IT, GU), Africa, Asia, Europe, Latin America and the Caribbean/ Oceania.

The 28 categories of education level are grouped as well. The groups include servicemembers with no high school diploma (includes less than high school diploma, those attending high school, secondary school credentials), with a high school diploma (test based equivalency diploma, occupational program certificate, correspondence school diploma, high school certificate of attendance, home study diploma, adult education diploma, GED, high school diploma), some college (includes associates and cert programs, no degree), a bachelor's degree (baccalaureate, 1 + years master), a master's degree, or a doctorate degree (post doctorate, first professional degree, doctorate).

Accession source codes are grouped into induction, voluntary enlistment (reserve or regulator component), service academy (US Naval Academy, Air Force, Coast Guard, Merchant Marine, Aviation Cadet), ROTC/NROTC (scholarship), ROTC/NROTC (non-scholarship), officer candidate school (OCS), direct appointment (Officer, Warrant Officer), LDO Program, and Other. The Unit Afloat/Ashore code is regrouped as shore duty, sea duty CONUS (partial sea credit, double sea credit), shore duty CONUS, and sea duty CONUS.

Faith group codes are grouped into Indigenous/Christian, Western/Christian (Protestant, Adventist, Anglican, Baptist, Evangelical, Holiness, Lutheran, Methodist), Eastern/Christian (Orthodox, Eastern Protestant), Restorationist/Christian (Jehovah's Witnesses, Latter day Saints), Non-Denominational/Christian, Catholic, Fundamentalist, Islam, Judaism, Buddhism/Hindu/Bahai, Pagan, Atheist, Agnostic, and None (No preference, Unknown, None). This decreases the number of categories from 93 to 13. Ethnic Affinity Code is minimized from 22 to 18 by grouping Pacific Asians (Melanesian, Micronesian, Polynesian, Other Pacific Island Descent), and Hispanic Descent (Mexican, Puerto Rican, Cuban, Latin American with Hispanic Descent, Other Hispanic Descent). DOD Occupational Code had the largest number of possible responses, at 250 unique entries. The data and was grouped into nine categories that included security, spec war/Divers/EOD, services, combat systems/weapons, engineering, surface, unknown, administration, and aviation. Specific Navy enlisted classifications (NEC) and ratings were not available in the datasets selected.

For rank group, senior enlisted are defined in PDE as PO2, PO1, and CPO. Junior enlisted are defined as SR, SA, SN, and PO3. 3,314 personnel were assigned Rank "EEE" in PDE. No clarification was provided at the time of this study, but the assumption is that these service members were assigned the ranks as a measure of privacy. Individuals that can be easily identified by the factors provided may have these factors encoded by the PDE team. Because these personnel were also assigned the "Senior Enlisted" variable for rank group, they were manually assigned the mode of the "Senior Enlisted" variable, which is Petty Officer Second Class.

For those assigned "Attrit," discussed in depth below, a new single variable was constructed, Separation Type. The RCCPDS Transaction File contains multiple variables that indicate separation type. These included separation code, interservice separation code, selected reserve involuntary loss code, and selected reserve misconduct loss type code. These were compared against each other for each member and consolidated into a single category called separation type.

b. Binary

The response variable in this thesis is a binary variable with the response of either Attrit or Non-Attrit as the possible outcomes. There were two methods that could be used in determining attrition. The first was to compare a RCCPDS master file with the master file of the next year. Because only presently serving members are listed, if the corresponding PDE identification appeared on the 2018 data set, but not on the 2019 data set, this member could be assigned an "attrit" code. The other method was to compare the September 2018 master file to the 2019 transaction file. If a reserve service member separated from service within the previous year, the 2019 transaction file annotates the manner in which the member separated. Because of this, the member could be coded as "attrit." For this study, the first method was used. This was done because of recommendations by the Army thesis students mentioned in the literature review. They indicated large numbers of flaws in data entry and missing values which required removal of observations or best guesses as to method of separation. The transaction file variables for separation were annotated for servicemembers identified as attrits, based on the master file record presence. Because there were many possible variables expressing separation, mentioned previously, with more than 50% missing values, they were merged as one "Separation Category." This was assigned as a separate variable. Therefore, there are members listed as attrit who do not have a documented reason for separation. Using the second method to identify attrition would have excluded all the records without separation code, even though the master file shows they no longer have an active record.

The pregnancy status variable was constructed from the column "Due Date" on the MRRS postpartum dataset. If the service member had a date listed that fell within the past year or predicted in the next 9 months, she was assigned as "Y." If a not date was listed, or service member sex is male, they were assigned "N."

c. Numeric

Retirement points represent the number of creditable retirement points that are earned as a service member completes training in accordance with the Secretary of the Navy Manpower and Reserve Affairs approved correspondence course list, attends

required drill, or serves additional years. The maximum number of points that can be earned each year is 365. 50 retirement points is credited as one year of service that qualifies toward the 20-year retirement minimum. Retirement years are different from the variable that expresses total years of service. This is because a service member may not achieve a qualifying retirement year in a year of service. Total years of service includes all years that a service member has been under contract.

2. Missing Data

The data set used required careful manipulation of missing data and entries with "not applicable" entries. Orrick (2019) alluded to the possible misrepresentation of the population in his study caused by removing observations with missing variables. Best effort was made to keep as many variables as possible and only the variables with more that 75% missing values were removed. Few individual observations were removed from the group. The methods of handling missing data for each variable are as follows:

- Rank observations were assigned PO2 as mentioned previously.
- Faith groups assigned as unknown/none.
- Service Agreement was assigned four years.
- Education level assigned high school diploma.
- Marital status assigned single.

3. Test/Training Set

The data was then randomly split into a training set of 24,871 observations, and test set of 10,670 observations.

4. Limitations and Assumptions

As mentioned in previous studies, the accuracy of information available in the PDE system relies on user input. It would be very difficult to verify each PDE PID entry and variables as accurate. The management of missing data also incurs some inherent inaccuracies. Because PDE data providers update at different frequencies and on different

dates for each database, it is safe to assume the dataset does include outdated information that may not fully capture the present status of each service member in the dataset.

Including a single year of data prevents the possibility of identifying attrition trends over time. Another limitation is that although PDE is very well maintained for Army analysis, data for the Navy is very limited. Army attrition studies by Gobea (2019) and Cammack (2020) noted that analysis was more accurate with inclusion of medical and physical readiness data, compared to Speten (2018) where only demographic and administrative variables were included. However, the Navy Physical Readiness Management system (PRMS) utilized member identification numbers that did not directly correlate to the PDE PID, so the data associated to physical readiness was not included.

An assumption made is that the methodology for defining the response variable of attrit or non-attrit is accurate. As discussed, other studies relied on specific assignments of separation codes on the RCCPDS transaction file to identify those that left service.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DESCRIPTIVE STATISTICS

A. DATASET OVERVIEW

This chapter provides an overview of the dataset used to construct the logistic regression model. In summarizing the data set, it is interesting to note both the similarities and differences in the population of enlisted reserve sailors in comparison to that of active-duty sailors. The dataset is composed of 75% males which is just slightly lower than active duty for the same year, which consisted of 80% (DMDC, 2022). The racial demographic mirrored that of the active-duty component with white service members making up 60% of the population. With the average age 32.46, enlisted reserve service members average 5 years older than the active duty (AD) community. The data set also shows that the education level is lower amongst the reserve service members in that 77% of the enlisted population have only high school diplomas, 10% have completed some college, 10% have attained a bachelor's degree, but only 1% have attained a master's degree or higher. In the active-duty community, 11% have attained a bachelor's degree, and 7% have attained a master's degree or higher. Mirroring that of the AD community, the majority of servicemembers in this data set are either married or single, and never married. Only about 8% are divorced or widow(er)s. 30% of selected reserve member unit locations are based in the Southwest region. Of the 35,541 observations, 7,218 are defined as having left service in the observed year. This is an attrition rate of 20.3% and is much lower than the AD enlisted attrition rate of 29.2% for FY 19.

B. NUMERIC VARIABLE SUMMARY

Table 4 shows the summary statistics for data set numeric variables compared to the subset of members that left service. The statistics of the age variable are consistent amongst the full data set and the subset of those who left service the following year. When analyzing the ages by groups, we see that the group of 25–29-year-old sailors left service at a higher proportion than others, at a rate of 23%.

Table 4. Numeric Variables

Full Data Set					Attrition Subset		
	Min	Max	Mean		Min	Max	Mean
Age	17	60	32.46		17	60	31.94
Qualifying retirement years	0	31	7.69		0	29	7.41
Retirement Points	0	7804	1597		0	7804	1564
Years of Service	0	21	3.37		0	21	3.31
Length of Contract	0	8	5.70		0	8	5.57
Number of Dependents	0	10	1.27		0	10	.99

C. BINARY VARIABLE SUMMARY

Figure 2 shows the rate of attrition for males and females in this data set. Males have an attrition rate of 21%, while female attrition rate is slightly lower at 19%. This differs from Navy active duty and other services in that females have a higher attrition rate in those populations. In this case, analysis of the separation reason would provide added insight.

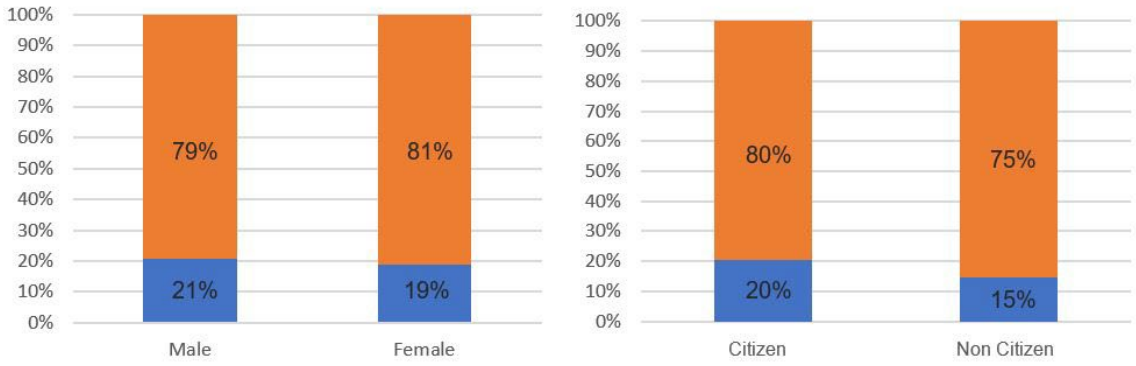

Figure 2. SELRES Enlisted Attrition by Sex and Citizenship

The biggest difference in attrition rates among the binary variables was seen in citizenship status. Figure 2 shows that selected reserve service members who are US citizens attrit at 6% higher rate than those who are not. However, it is worth noting that there are only 192 observations of non-citizens in this data set. The remaining binary variables, retirement eligibility, deploying flag, pregnancy, and major allergies, did not display a significant difference in attrition rates amongst levels.

D. CATEGORICAL VARIABLE SUMMARY

The highest attrition rate, by rank, is seen among the seaman recruit group at 24%. Figure 3 shows this in comparison to the other ranks. Interesting to note is that the attrition rate decreases as rank increases, with the exception of the petty officer third class rank. At the Chief Petty Officer rank, the attrition rate decreases to 17%. Amongst the studies referenced that included rank as a variable, the E1 paygrade tended to show the highest rate of attrition.

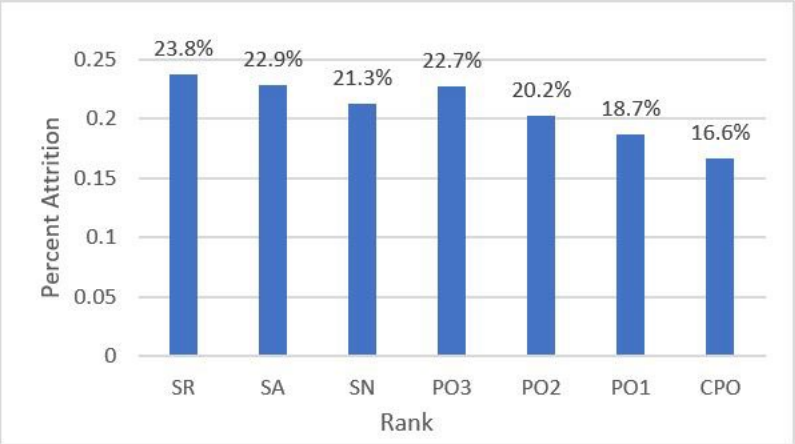

Figure 3. SELRES Enlisted Attrition by Rank.

Although the Southwest reserve region contains the largest proportion of reserve service members, 30%, it accounts for the lowest attrition rate at only 14%. The Mid Atlantic - Norfolk region has the highest rate 29%. The second highest attrition rate is seen in the Mid Atlantic Great Lakes region. The attrition rates of all regions are displayed in

Figure 4. Speten (2018) also uses unit location in his first-term USMC attrition study. Across the five-year span in his data set, the South averages 60% attrition while the Northeast, which covers the Mid Atlantic – Norfolk region, only averages 3% attrition.

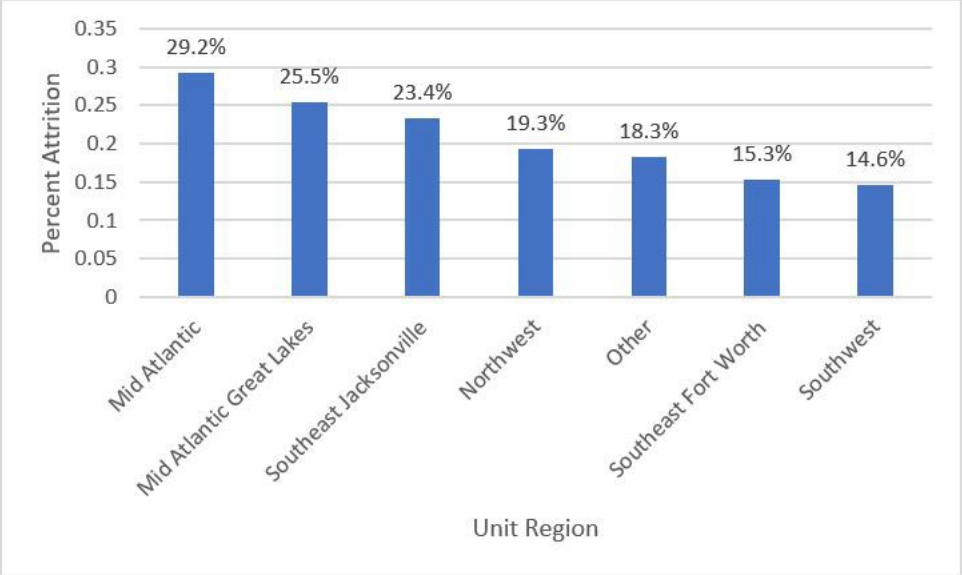

Figure 4. SELRES Enlisted Attrition by Unit Location.

As seen in Figure 5, attrition is highest among those servicemembers with only high school diplomas. Attrition rate decreases as education level increases.

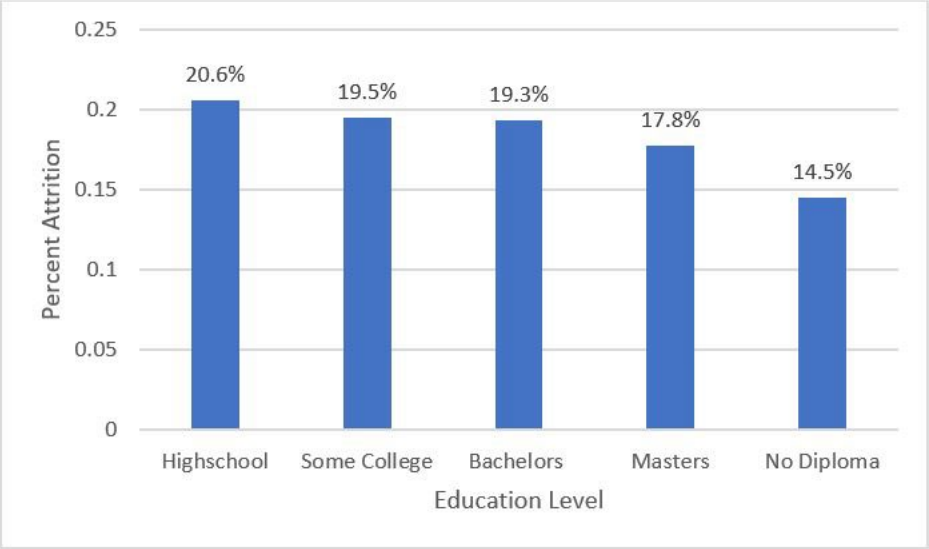

Figure 5. SELRES Enlisted Attrition by Education Level.

Similar to what was observed in the active-duty studies reference in the literature review, married service members in this data set have the lowest rate of attrition; however, as seen in Figure 6, the difference across the four the attrition rates is only 4%. Also note, this data set only includes 64 observations that are coded as widows.

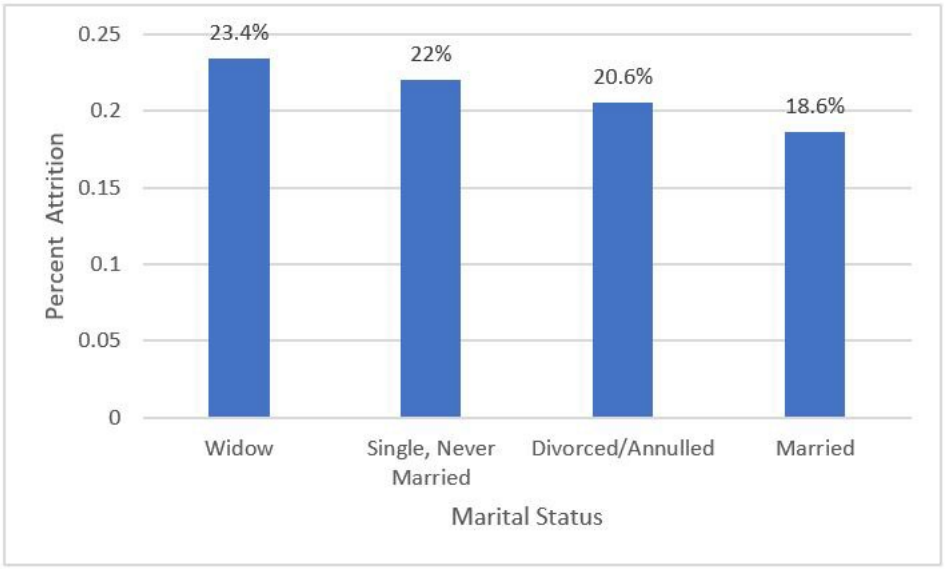

Figure 6. Attrition by Marital Status.

In contrast to the active-duty studies referenced, the Black and African American demographic showed the highest rate of attrition at 5% higher than that of whites. As seen in Figure 7, the attrition rate is approximately 20% for Whites and American Indians and 14% for Asians and Native Hawaiians /Pacific Islanders.

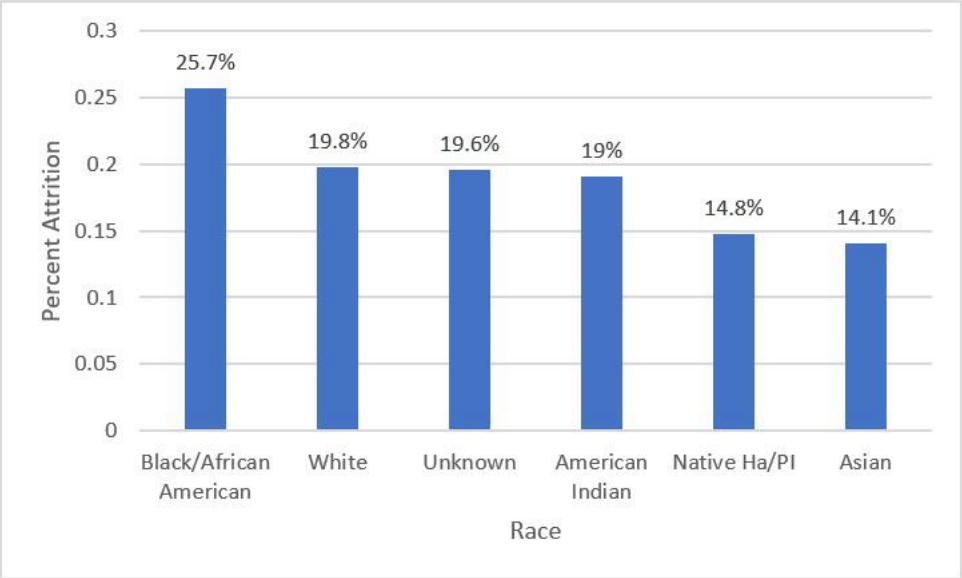

Figure 7. SELRES Enlisted Attrition by Race

Figure 8 displays the attrition rates for each DOD occupation category. Reserve service members with occupations that fall under service, seamanship, and unqualified have the highest three attrition rates, all above 22%. These categories all include very junior personnel. Noted previously, those with lower rank and less time in service had higher attrition rates.

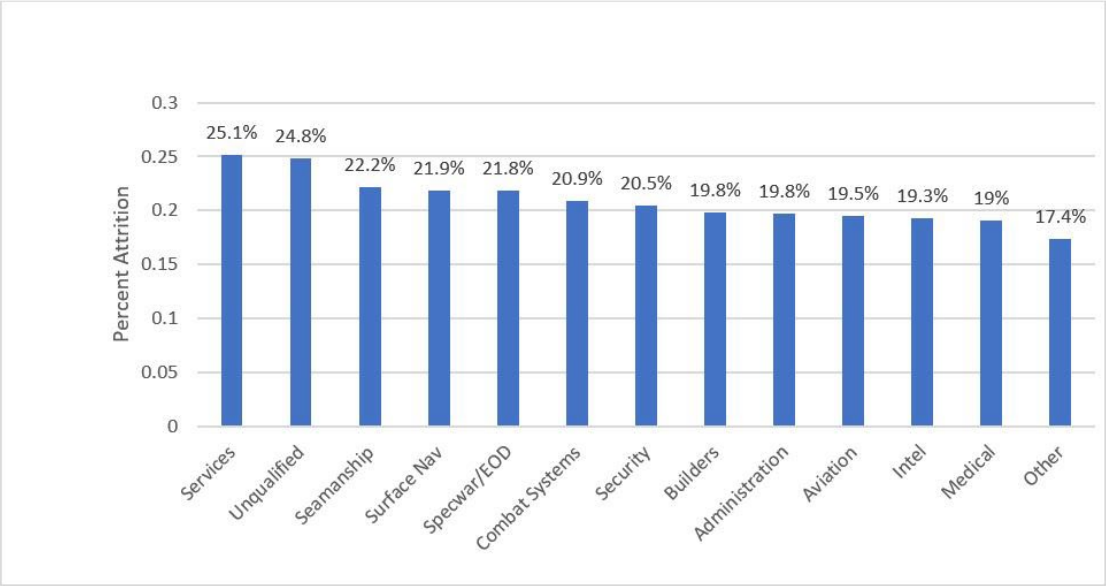

Figure 8. SELRES Enlisted Attrition by Job Type

THIS PAGE INTENTIONALLY LEFT BLANK

IV. MODELING AND ANALYSIS

A. FULL MODEL LOGISTIC REGRESSION

As the first approach to modeling, a logistic regression model was fit using all available predictors in the data set. Logistic regression uses the logit function to estimate probabilities between 0 and 1 that an event occurs. Figure 9 provides the logistic function and the log-odds formulation, where $X_1,\ldots,X_p$ represent p predictors and $\beta_0,\ldots,\beta_p$ represent the coefficients for p predictors.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p$$

Figure 9. Logit Function and Log Odds. Source: James (2013).

Following James (2013), the coefficients can be estimated using maximum likelihood estimation (MLE). The calculations result in coefficient estimates that maximize the likelihood function. For this thesis, the summary() function in R studio's <base> package (Rstudio 2021) was used to calculate these estimates. The results are shown in Appendix A.

A predictor with a positive value coefficient estimate signifies that an increase in that predictor is associated with an increase in the probability of attrition. Specifically, a one-unit increase in that predictor is associated with an increase in the log odds of attrition by the coefficient estimate amount. In interpreting *p* values, lower values represent greater statistical significance. In the output, the significance is indicated by use of asterisks.

The predictors with the largest negative value coefficients are ASG_UNT_LOC_ISO_A3_CTRY_CDAsia, PregnancyY, and MRTL_STAT_CDL

(legally separated), however, neither of the three have statistical significance. The predictors with the largest positive value coefficients estimates are ASG_UNT_NV_ASHR_AFLT_CDShore and Duty ASG_UNT_NV_ASHR_AFLT_CDSea Duty – Conus but similar to those with the largest negative values, they have no statistical significance. The predictors with statistical p-values less than 0.01 include all levels of rank, unit regions SE FW and SW, 11 of 14 faith levels, service member age, total dependent quantity, race levels Black and Asian, deployment flag-Y, and gender-Male. The only level of Education to show any significance is Master's and for DOD occupation the levels of significance are Intel and Other. The predictors with p-values equal to 1.0, which indicate no statistical significance, include all levels of security clearance, unit country location, marital status, years of service, service agreement and retirement points.

An optimal threshold value is used in the model to make predictions. Selection of an optimal threshold is done using the optimalCutoff() function in R studio's <InformationValue> package (Rstudio 2021). For the full logistic regression model, the value is calculated as 0.5242686. This means that if the probability a service member attrits is greater than 52.4%, the model will classify it as "Attrit," otherwise, it will be classified as "Not Attrit." This value is selected to minimize the overall error rate. However, choosing this value results in a near zero value of model sensitivity, meaning the model classifies almost every observation as non attrit. Following James (2013), while the default threshold is usually .50, he suggests that if the preference is to correctly assign more observations Y=1, in this case attrit, we may consider lowering the threshold. The threshold was changed to 0.2049, which reflects the proportion of service members in the training set that did in fact attrit. The confusion matrix, specificity, sensitivity, misclassification rate, and the area under the receiver operating characteristic (AUROC) curve were calculated to assess the quality of our predictions made using the test data set and the model's performance. A confusion matrix displays the possible results of the model predictions. The nomenclature and general organization of the results is seen Figure 10.

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Figure 10. Possible results when applying a classifier or diagnostic test to a population. Source: James (2013).

The true negative count represents the total number of observations whose outcomes were correctly classified as Attrit = No by the model. The true positives, represent the total number of observations whose outcomes were correctly classified as Attrit = Yes by the model. The false negatives represent the total number of observations whose outcomes were classified as Attrit = No by the model but were in fact attrits. False positives represent the total number of observations whose outcomes were classified as Attrit = Yes by the model and which were in fact not attrits. As shown in Figure 11, this information can be used to calculate other measures of classification.

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N	

Figure 11. Important measures for classification and diagnostic testing. Source: James (2013).

Measures and characteristics of the full logistic regression model for this thesis are displayed in Table 5. The model's sensitivity value of 0.5962 means it can correctly identify only 59% of those servicemembers who attrit. The model's specificity value of 0.596 means the model will also correctly identify 59% of servicemembers that do not attrit. The model's misclassification rate, which includes both false positives and false

negatives, is 0.4035. This means that approximately 40% of the time the model will incorrectly predict the outcome using this threshold.

Table 5. Logistic Regression Model Characteristics.

Logistic Regression Confusion Matrix (threshold = 0.2049)		
Predicted	**Observed**	
	Non Attrit	Attrit
Non Attrit	5123	860
Attrit	3465	1270
Measures of Performance		
Sensitivity	0.596	
Specificity	0.597	
Misclassification Rate	0.404	

Lastly, the ROC curve displays both the false positive and true positive rates across all thresholds. The area under the curve provides the overall model performance. This area can range from .5 to 1.0 and an AUC value closer to 1.0 indicates a better classifier. The closer the true positive rate is to 1 and the closer the false positive rate is to 0, the better the variables are at distinguishing those who attrit from those who do not. As seen in figure 12, the full logistic regression model AUC value is .6381. The overall performance can be interpreted as: if two Enlisted Reserve service members are selected at random from the population, such that one attrits and one does not, using this model, there is a 63.8% probability that the servicemember that actually does attrit will have a larger predicted probability of attrition than the one that does not.

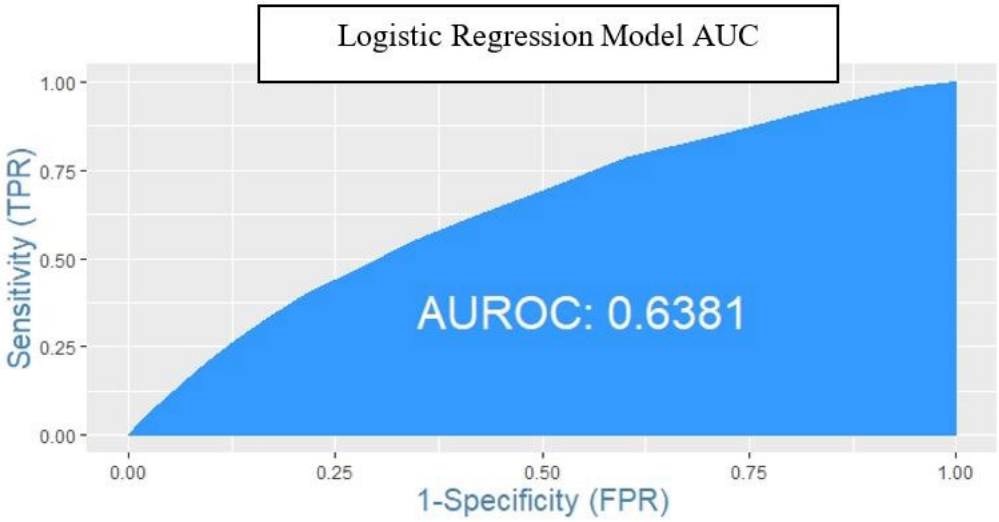

Figure 12. Logistic Regression Model AUC.

B. RANDOM FOREST VARIABLE SELECTION

The second modeling approach implements random forests as a method of variable selection to fit a logistic regression model. The model output provides a measure of variable importance from a random forest based on mean decrease Gini criterion. The mean decrease in the Gini criterion expresses how each variable contributes to the random forest. The variable importance for the random forest used in this thesis is displayed shown in Figure 13. From the output, Reserve retirement points, DOD occupation, and service member age are the three most important variables in classifying a service member as attrit or not-attrit. The least important include reserve retirement eligibility, citizenship status, and pregnancy.

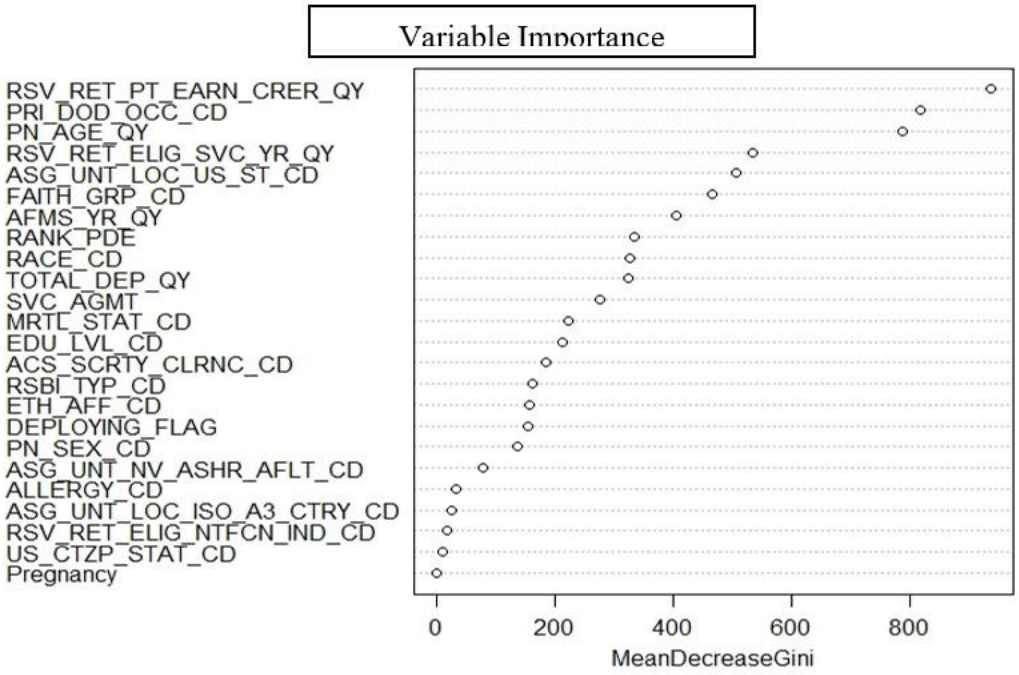

Figure 13. Random Forest Variable Importance.

The top ten variables in order of importance were then used to fit the model. The coefficient estimates are listed in Appendix B. The predictors' statistical significance closely mirrors that seen previously. As discussed in the previous section, the optimal threshold, 0.2049, was used to calculate the model's accuracy in predictions made using the test data set. The confusion matrix and measures of performance are displayed in Table 6. The model's sensitivity value of 0.5854 means it correctly identifies 58.5% of the test set servicemembers who attrit. The model's specificity value of 0.6063 means the model correctly identifies 60.6% of servicemembers that do not attrit. The model's misclassification rate is 0.3978. This means that approximately 40% of the time the model incorrectly predicts the outcome in the test set.

Table 6. Random Forest Variable Selection Logistic Regression Model Characteristics.

Random Forest Logistic Regression Confusion Matrix		
Predicted	**Observed**	
	Non Attrit	Attrit
Non Attrit	5207	883
Attrit	3381	1247
Measures of Performance		
Sensitivity	0.585	
Specificity	0.606	
Misclassification Rate	0.398	

Displayed in Figure 14, the random forest variable selection logistic regression model AUC value is .6296, a slight decrease from the full logistic regression model.

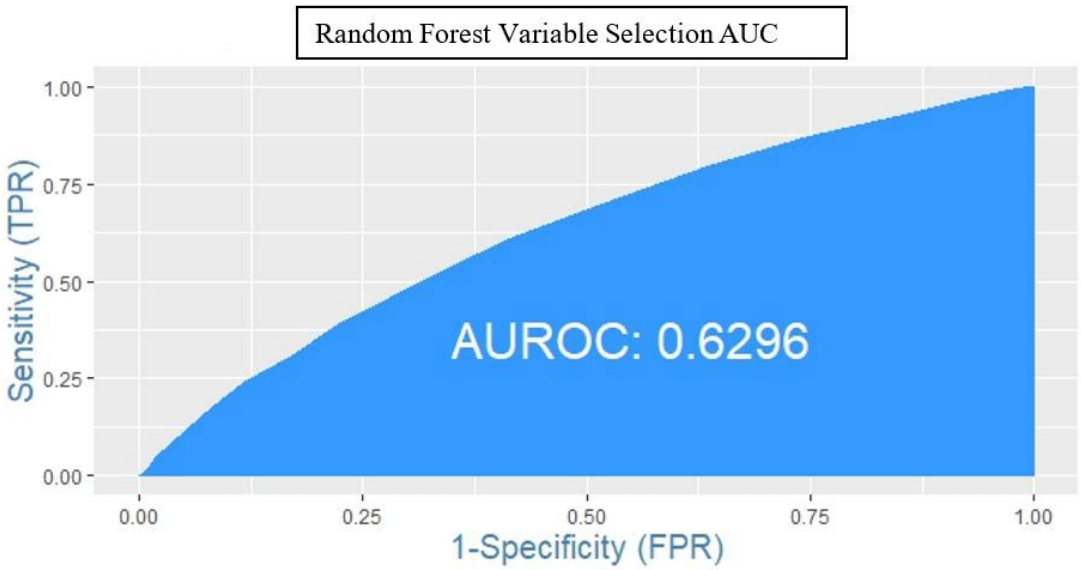

Figure 14. Random Forest Variable Selection Logistic Regression Model AUC.

C. LASSO VARIABLE SELECTION

The last approach to modelling used the least absolute shrinkage and selection operator (LASSO) variable selection to fit a logistic regression model. LASSO is a statistical method that applies a penalty to each predictor in order to shrink the estimates towards zero, in turn limiting variable inclusion and producing a subset of only those variables that are important to predicting the outcome. The outputs of the lasso model applied to the training data set are displayed in Appendix C.

Predictors whose coefficients were shrunk to zero included ASG_UNT_NV_ASHR_AFLT_CDShore Duty, EDU_LVL_CDSome College, ETH_AFF_CDPacific Island Descent, MRTL_STAT_CDD, PRI_DOD_OCC_CDAviation, and PRI_DOD_OCC_CDBuilder. Again, the threshold value, 0.2049, is used to make predictions using the test data set using the modeling consisting of the remaining predictors. The confusion matrix and measures of performance are displayed in Table 7. The model's sensitivity value of 0.5972 means it can correctly identify the test set servicemembers who attrit only 59.7% of the time. The model's specificity value of 0.598 means the model correctly identifies 59.8% of test set servicemembers who do not attrit. The model's misclassification rate is 0.4021. This means that approximately 40% of the time the model incorrectly predicts the outcome in the test set.

Table 7. LASSO Logistic Regression Model Characteristics.

LASSO Logistic Regression Confusion Matrix		
Predicted	**Observed**	
	Non Attrit	Attrit
Non Attrit	5136	858
Attrit	3452	1272
Measures of Performance		
Sensitivity	0.597	
Specificity	0.598	
Misclassification Rate	0.402	

Figure 15 displays the LASSO logistic regression ROC curve and AUC value .6384.

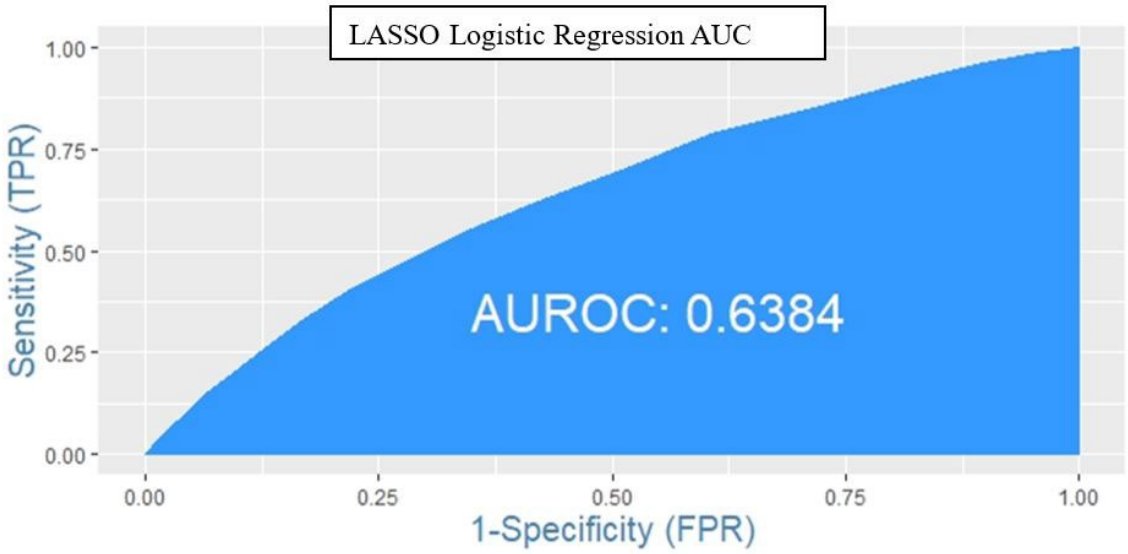

Figure 15. LASSO Logistic Regression Model AUC.

D. MODEL COMPARISON

Both sensitivity and specificity across all three are approximately 0.6. Also, all three have misclassification rates of about 0.40. In terms of area under the ROC curve, the model using LASSO variable selection provides the best overall performance. However, despite the slight differences in performance, all of the models produce suboptimal results.

To verify any underlying issues, the full model was checked for multicollinearity. Interpretation of the models' coefficients rely on all variables being uncorrelated. Assessing multicollinearity can be done by computing the variance inflation factor (VIF). Following James (2013) VIF values of 10 may cause problems with model fit and coefficient interpretability. Table 8 shows the model output from the VIF() function in R Studio's < VIF> package (Rstudio 2021).

Table 8. VIF Values for Logistic Regression Model

Variable Name	VIF Value
RANK_PDE	5.918
ACS_SCRTY_CLRNC_CD	2.178
ASG_UNT_LOC_US_ST_CD	1.555
ASG_UNT_NV_ASHR_AFLT_CD	1.054
EDU_LVL_CD	1.221
ETH_AFF_CD	2.134
FAITH_GRP_CD	1.370
MRTL_STAT_CD	2.108
PN_AGE_QY	2.891
PN_SEX_CD	1.074
RACE_CD	2.366
RSBI_TYP_CD	1.684
RSV_RET_ELIG_NTFCN_IND_CD	1.331
RSV_RET_ELIG_SVC_YR_QY	11.057
RSV_RET_PT_EARN_CRER_QY	24.132
US_CTZP_STAT_CD	1.059
Pregnancy	1.000
ALLERGY_CD	1.078
DEPLOYING_FLAG	1.233
TOTAL_DEP_QY	1.859
AFMS_YR_QY	10.744
PRI_DOD_OCC_CD	5.962
SVC_AGMT	1.538

The predictors AFMS_YR_QY, RSV_RET_PT_EARN_CRER_QY, and RSV_RET_PT_EARN_CRER_QY displayed VIF values greater than 10. Of the three predictors, RSV_RET_PT_EARN_CRER_QY was removed. After recalculation, VIF values for all remaining predictors were under 6.0. Another issue that may need be addressed is the presence of outliers that might be unduly influential to the dataset. Cook's distance is used to identify such outliers. The results of Cook's distance calculations, implemented using the cooks.distance() function in R Studio's <stats> package (Rstudio 2021) are shown in Figure 16.

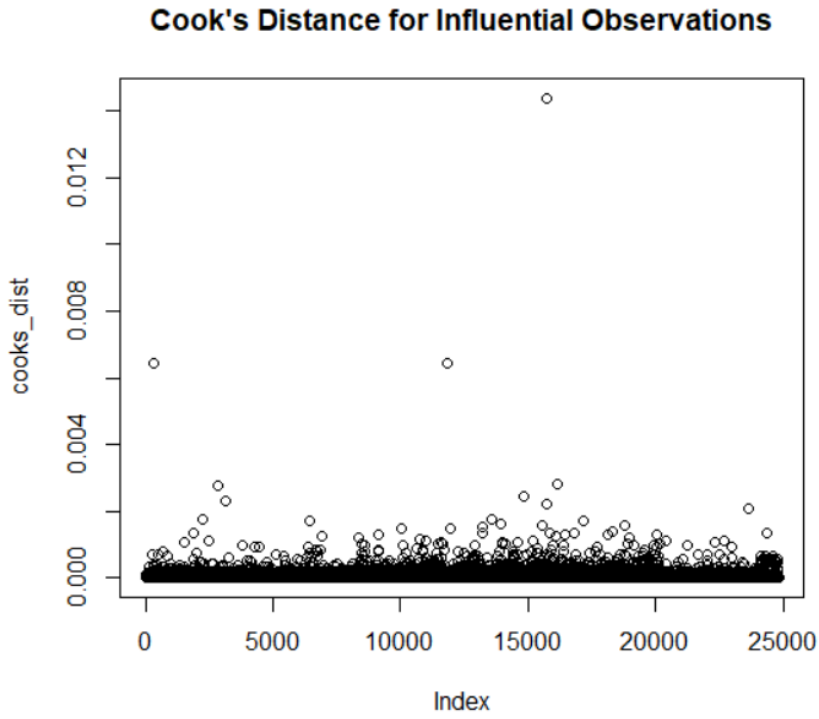

Figure 16. Cook's Distance for Dataset

There appears to be no observations with values greater than 1, which indicate outliers, following Cook (1977). When utilizing a newer cutoff, $4/n$ where n = the number of observations, we see three observations that may considered outliers. After adjusting to account for multicollinearity and outliers, all models were fit again; however, no change to the models' performance measures were observed.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION

A. SUMMARY OF FINDINGS

Overall, these models, with the variables used, cannot reliably classify attrition outcome. The models have little predictive power and poor discrimination. However, they do provide some insight. The variables excluded in the Lasso model do align with those variables of no significance in the full logistic regression model. However, these are not consistent with the most influential predictor variables determined in the random forest model. In the LASSO and full logistic regression model, demographic variables seemed to provide the most value. Some of the most significant included rank, ethnicity, faith, and race. Variables relating to the reserve job, such as location and occupation, provided little to the models. When interpreting odds ratios, conditions such as being Black, getting older, being unqualified, male, or occupying stations in the MidAtlantic Norfolk Region present greater odds of attrition the next year. Contrary to expectation, being married presents higher odds of attrition than being single. Being Asian or Pacific Islander/Hawaiian and having an increasing number of dependents present less odds of attrition. In the random forest model, the unit and service-related variables were most influential. This included reserve retirement points, time in service, unit location and occupation.

B. RECOMMENDATIONS AND FUTURE WORK

While the result may indeed signify that the variables in the dataset are simply not effective predictors of attrition, Prakash (2016) provides recommendations for improving models with poor discrimination. One that may be applicable in this thesis would be to utilize a larger data set**.** Issues that limited the data set size were the amount of missing data as well as missing data dictionaries. As seen in this thesis, and theses referenced, all were required to address significant amounts of missing data when using the PDE. This could be addressed by reproducing the study outside of the PDE environment. This in turn may allow access to more updated information and quicker communication with the data owners. The PDE does contain datasets with more demographic, unit related and medical data, but at the time of this thesis completion, data dictionaries providing clarification on

many variable definitions and factors were not available. A takeaway is that it is incumbent upon data providers to provide usable information. Data owners and PDE system managers have taken on the task of locating such dictionaries so that they can be available to students for analysis in the future.

The next recommendation is to readdress the treatment of missing values and outliers. A total of 3,314 personnel were assigned Rank "EEE" in PDE. It was noted by the thesis advisor that these observations could have represented Senior Chief Petty Officers and Master Chief Petty Officers. The CPO variable did not distinguish if all CPO pay grades were included or if included on the E-7 pay grade. This could not be inferred and verified because inclusion of paygrades requires additional PDE permissions for data usage. This matters because these observations represent approximately 10% of the data of the dataset used. An incorrect assignment of rank could mean the dataset may be misrepresentative of the population.

A recommended extension of this study includes assignment of a risk scorecard value to each observation. The intended application of the risk scorecard is to provide a risk score, analogous to a credit score, expressing the likelihood of individual attrition within the next year. This would provide the opportunity, at a command level, to better plan for short-term unexpected and expected personnel losses.

APPENDIX A. LOGISTIC REGRESSION MODEL SUMMARY

Table 9. Logistic Regression Model Summary.

Variable	Estimated Coefficient	Odds Ratio	Std. Error	Z value	Pr(>|z|)
(Intercept)	-1.456e+01	4.759847e-07	3.247e+02	-0.045	0.964244
RANK_PDEPO1	2.269e-01	1.254762e+00	7.571e-02	2.997	0.002723 **
RANK_PDEPO2	3.061e-01	1.358076e+00	7.670e-02	3.990	6.60e-05 ***
RANK_PDEPO3	5.243e-01	1.689210 e+00	9.008e-02	5.820	5.89e-09 ***
RANK_PDESA	5.605e-01	1.751490 e+00	1.407e-01	3.983	6.80e-05 ***
RANK_PDESN	4.205e-01	1.522649 e+00	1.051e-01	3.999	6.37e-05 ***
RANK_PDESR	5.785e-01	1.783364e+00	1.810e-01	3.196	0.001396 **
ACS_SCRTY_CLRNC_CDSecret	-5.504e-02	9.46441e-01	6.371e-02	-0.864	0.387588
ACS_SCRTY_CLRNC_CDTop Secret	6.697e-02	1.069259e+00	7.836e-02	0.855	0.392758
ASG_UNT_LOC_ISO_A3_CTRY_CDAsia	-1.109e+01	1.527875e-05	2.295e+02	-0.048	0.961461
ASG_UNT_LOC_ISO_A3_CTRY_CDEurope	7.203e-01	2.054988e+00	4.809e-01	1.498	0.134202
ASG_UNT_LOC_ISO_A3_CTRY_CDLatin America	-2.076e-01	8.12549e-01	4.124e-01	-0.503	0.614736
ASG_UNT_LOC_ISO_A3_CTRY_CDMiddle East	9.312e-02	1.097592e+00	2.571e-01	0.362	0.717235
ASG_UNT_LOC_ISO_A3_CTRY_CDNorth America	5.855e-01	1.795824e+00	4.078e-01	1.436	0.151059
ASG_UNT_LOC_US_ST_CDMid Atlantic Great Lakes	-3.723e-01	6.891625e-01	3.545e-01	-1.050	0.293627
ASG_UNT_LOC_US_ST_CDMid Atlantic Norfolk	-1.108e-01	8.951510e-01	3.529e-01	-0.314	0.753637
ASG_UNT_LOC_US_ST_CDNorthwest	-7.183e-01	4.875933e-01	3.546e-01	-2.026	0.042812 *
ASG_UNT_LOC_US_ST_CDNot US	NA	NA	NA	NA	NA
ASG_UNT_LOC_US_ST_CDSoutheast Fort Worth	-9.635e-01	3.815703e-01	3.543e-01	-2.719	0.006542 **

ASG_UNT_LOC_US_ST_CDSoutheast Jacksonville	-4.459e-01	6.402705e-01	3.553e-01	-1.255	0.209526
ASG_UNT_LOC_US_ST_CDSouthwest	-9.659e-01	3.806461e-01	3.531e-01	-2.735	0.006235 **
ASG_UNT_NV_ASHR_AFLT_CDSea Duty - Conus	1.074e+01	4.596603e+04	3.247e+02	0.033	0.973628
ASG_UNT_NV_ASHR_AFLT_CDShore Duty	1.067e+01	4.302826e+04	3.247e+02	0.033	0.973790
EDU_LVL_CDHighschool Diploma	4.676e-02	7.212511e-01	5.574e-02	0.839	0.401530
EDU_LVL_CDMaster's Degree	-3.785e-01	6.848929e-01	1.860e-01	-2.035	0.041852 *
EDU_LVL_CDNo Highschool Diploma	-3.268e-01	7.212511e-01	3.932e-01	-0.831	0.405994
EDU_LVL_CDSome College	9.898e-04	1.000990e+00	7.291e-02	0.014	0.989168
ETH_AFF_CDHispanic Descent	-6.791e-03	9.932325e-01	1.201e-01	-0.057	0.954902
ETH_AFF_CDIndian (US, Canadian, Other)	2.662e-02	1.026974e+00	1.650e-01	0.161	0.871857
ETH_AFF_CDNone, Unk	3.085e-01	1.361411e+00	1.143e-01	2.699	0.006953 **
ETH_AFF_CDPacific Island Descent	4.891e-03	1.004903e+00	3.012e-01	0.016	0.987043
FAITH_GRP_CDAgnostic	1.380e+00	3.976782e+00	2.828e-01	4.882	1.05e-06 ***
FAITH_GRP_CDAtheist	1.018e+00	2.768812e+00	2.913e-01	3.496	0.000472 ***
FAITH_GRP_CDBuddaism/Hindu	1.373e+00	3.946025e+00	2.902e-01	4.731	2.24e-06 ***
FAITH_GRP_CDCatholic	1.043e+00	2.837244e+00	1.997e-01	5.223	1.76e-07 ***
FAITH_GRP_CDEastern, Christian	1.377e+00	3.962722e+00	4.136e-01	3.329	0.000872 ***
FAITH_GRP_CDIndigenous, Christian	8.333e-01	2.300951e+00	7.873e-01	1.058	0.289851
FAITH_GRP_CDIslam	1.021e+00	2.774709e+00	3.261e-01	3.130	0.001749 **
FAITH_GRP_CDJudaism	1.325e+00	3.762072e+00	3.951e-01	3.353	0.000799 ***
FAITH_GRP_CDNon Denom, Christian	9.382e-01	2.555266e+00	1.982e-01	4.734	2.20e-06 ***
FAITH_GRP_CDNone	1.016e+00	2.761752e+00	1.944e-01	5.225	1.75e-07 ***
FAITH_GRP_CDOther	6.030e-01	1.827566e+00	5.273e-01	1.144	0.252783
FAITH_GRP_CDPagan	6.076e-01	1.835941e+00	4.376e-01	1.389	0.164980
FAITH_GRP_CDReformative, Christian	8.047e-01	2.235988e+00	2.167e-01	3.713	0.000204 ***

FAITH_GRP_CDWestern, Christian	1.025e+00	2.787045e+00	2.012e-01	5.094	3.50e-07 ***
MRTL_STAT_CDD	4.713e-01	1.602022e+00	1.093e+00	0.431	0.666292
MRTL_STAT_CDL	-1.049e+01	2.771718e-05	3.247e+02	-0.032	0.974223
MRTL_STAT_CDM	6.135e-01	1.846884e+00	1.092e+00	0.562	0.574212
MRTL_STAT_CDN	3.241e-01	1.382718e+00	1.092e+00	0.297	0.766638
MRTL_STAT_CDW	7.586e-01	2.135184e+00	1.142e+00	0.665	0.506367
PN_AGE_QY	8.196e-03	1.008229e+00	3.098e-03	2.646	0.008152 **
PN_SEX_CDM	1.953e-01	1.215734e+00	3.943e-02	4.954	7.27e-07 ***
RACE_CDAsian	-3.208e-01	7.255422e-01	1.074e-01	-2.986	0.002825 **
RACE_CDBlack/African American	1.910e-01	1.210475e+00	7.377e-02	2.589	0.009620 **
RACE_CDNative Hawaiian/Pacific Islander	-1.985e-01	8.199518e-01	1.771e-01	-1.121	0.262269
RACE_CDUnknown	1.975e-02	1.019948e+00	9.882e-02	0.200	0.841571
RACE_CDWhite	-4.367e-02	9.572652e-01	6.768e-02	-0.645	0.518734
RSBI_TYP_CDEnlistment Bonus, Prior (3 yr)	1.872e-01	1.205834e+00	1.535e-01	1.219	0.222672
RSBI_TYP_CDEnlistment Bonus, Prior (6 yr)	-2.557e-01	7.743614e-01	1.013e-01	-2.525	0.011573 *
RSBI_TYP_CDNot Applicable, Unk	-1.258e-01	8.817495e-01	6.306e-02	-1.996	0.045960 *
RSBI_TYP_CDReEnlistment Bonus, (3 yr)	-2.502e-01	7.786412e-01	2.781e-01	-0.900	0.368199
RSBI_TYP_CDReEnlistment Bonus, (6 yr)	-1.459e-01	8.642750e-01	1.251e-01	-1.166	0.243493
RSV_RET_ELIG_NTFCN_IND_CDYes	-1.349e-01	8.733711e-01	1.052e-01	-1.282	0.199718
RSV_RET_ELIG_SVC_YR_QY	-1.060e-02	9.894573e-01	8.729e-03	-1.214	0.224699
RSV_RET_PT_EARN_CRER_QY	9.753e-05	1.000098e+00	5.463e-05	1.785	0.074201 .
US_CTZP_STAT_CDYes	3.180e-01	1.374383e+00	2.525e-01	1.259	0.207907
PregnancyY	-1.083e+01	1.982728e-05	1.280e+02	-0.085	0.932572
ALLERGY_CDY	6.780e-02	1.070154e+00	1.197e-01	0.567	0.571051
DEPLOYING_FLAGN/A	1.000e-01	1.105171e+00	4.168e-02	2.399	0.016429 *
DEPLOYING_FLAGY	-5.782e-01	5.608898e-01	1.666e-01	-3.470	0.000520 ***

TOTAL_DEP_QY	-2.382e-01	7.880638e-01	1.639e-02	-14.531	< 2e-16 ***
AFMS_YR_QY	1.215e-03	1.001216e+00	1.425e-02	0.085	0.932054
PRI_DOD_OCC_CDAviation	-1.604e-03	9.983976e-01	7.722e-02	-0.021	0.983431
PRI_DOD_OCC_CDBuilders	-2.716e-03	9.972572e-01	6.688e-02	-0.041	0.967601
PRI_DOD_OCC_CDCombatSystems/Weapon	-5.480e-02	9.466751e-01	6.785e-02	-0.808	0.419310
PRI_DOD_OCC_CDEngineering	1.278e-02	1.012861e+00	7.769e-02	0.164	0.869339
PRI_DOD_OCC_CDIntell/Cryp	-2.062e-01	8.136379e-01	8.510e-02	-2.424	0.015369 *
PRI_DOD_OCC_CDMedical/Dental	-5.328e-02	9.481181e-01	7.051e-02	-0.756	0.449925
PRI_DOD_OCC_CDOther	-4.624e-01	6.297816e-01	2.316e-01	-1.996	0.045890 *
PRI_DOD_OCC_CDSeamanship	8.105e-02	1.084429e+00	7.961e-02	1.018	0.308639
PRI_DOD_OCC_CDSecurity	1.449e-02	1.014591e+00	7.314e-02	0.198	0.842996
PRI_DOD_OCC_CDServices	2.372e-01	1.267642e+00	1.352e-01	1.754	0.079492 .
PRI_DOD_OCC_CDSpecWar/EOD	3.826e-02	1.039006e+00	1.388e-01	0.276	0.782811
PRI_DOD_OCC_CDSurface Nav	4.830e-02	1.049485e+00	1.122e-01	0.431	0.666763
PRI_DOD_OCC_CDUnqualified in Code	6.497e-02	1.067128e+00	1.805e-01	0.360	0.718847
SVC_AGMT	1.207e-02	1.012139e+00	8.663e-03	1.393	0.163661

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 25181 on 24822 degrees of freedom
Residual deviance: 24037 on 24738 degrees of freedom
AIC: 24207
Number of Fisher Scoring iterations: 11

APPENDIX B. RANDOM FOREST MODEL SUMMARY

Table 10. Random Forest Model Summary.

Variable	Estimated Coefficient	Odds Ratio	Std. Error	Z value	Pr(>|z|)
(Intercept)	-2.2556601426	0.1048043	4.235e-01	-5.327	1.00e-07 ***
RSV_RET_PT_EARN_CRER_QY	1.146e-04	1.0001146	5.338e-05	2.147	0.031820 *
PRI_DOD_OCC_CDAviation	5.401e-02	1.0554983	7.564e-02	0.714	0.475207
PRI_DOD_OCC_CDBuilders	7.642e-02	1.0794164	6.525e-02	1.171	0.241541
PRI_DOD_OCC_CDCombatSystems/Weapon	2.955e-02	1.0299947	6.537e-02	0.452	0.651188
PRI_DOD_OCC_CDEngineering	5.899e-02	1.0607668	7.648e-02	0.771	0.440533
PRI_DOD_OCC_CDIntell/Cryp	-5.225e-02	0.9490906	7.187e-02	-0.727	0.467226
PRI_DOD_OCC_CDMedical/Dental	-2.913e-02	0.9712855	6.927e-02	- 0.421	0.674064
PRI_DOD_OCC_CDOther	-3.766e-01	0.6861585	2.307e-01	-1.633	0.102487
PRI_DOD_OCC_CDSeamanship	1.095e-01	1.1157618	7.877e-02	1.391	0.164365
PRI_DOD_OCC_CDSecurity	7.820e-02	1.0813424	7.148e-02	1.094	0.273925
PRI_DOD_OCC_CDServices	2.610e-01	1.2982563	1.344e-01	1.942	0.052118 .
PRI_DOD_OCC_CDSpecWar/EOD	9.208e-02	1.0964554	1.350e-01	0.682	0.495215
PRI_DOD_OCC_CDSurface Nav	1.250e-01	1.1331801	1.113e-01	1.124	0.261152
PRI_DOD_OCC_CDUnqualified in Code	4.363e-02	1.0445971	1.763e-01	0.248	0.804486
PN_AGE_QY	8.846e-03	1.0088856	2.799e-03	3.160	0.001578 **
RSV_RET_ELIG_SVC_YR_QY	-9.866e-03	0.9901829	8.050e-03	-1.226	0.220365
ASG_UNT_LOC_US_ST_CD Mid Atlantic Great Lakes	-3.794e-01	0.6842867	3.531e-01	-1.075	0.282583
ASG_UNT_LOC_US_ST_CD Mid Atlantic Norfolk	-1.520e-01	0.8589690	3.515e-01	-0.432	0.665388
ASG_UNT_LOC_US_ST_CD Northwest	-7.258e-01	0.4839322	3.532e-01	-2.055	0.039909 *
ASG_UNT_LOC_US_ST_CD Not US	-6.796e-01	0.5068071	3.702e-01	-1.836	0.066388 .
ASG_UNT_LOC_US_ST_CD Southeast Fort Worth	-1.011e+00	0.3637567	3.529e-01	-2.866	0.004161 **
ASG_UNT_LOC_US_ST_CD Southeast Jacksonville	-4.902e-01	0.6125128	3.538e-01	-1.385	0.165963
ASG_UNT_LOC_US_ST_CD Southwest	-1.025e+00	0.3587840	3.517e-01	-2.915	0.003560 **

FAITH_GRP_CDAgnostic	1.401e+00	4.0589811	2.818e-01	4.971	6.66e-07 ***
FAITH_GRP_ CDAtheist	1.020e+00	2.7719976	2.900e-01	3.515	0.000439 ***
FAITH_GRP_CDBuddaism/Hindu	1.408e+00	4.0887241	2.888e-01	4.876	1.08e-06 ***
FAITH_GRP_CDCatholic	1.018e+00	2.7670285	1.987e-01	5.122	3.03e-07 ***
FAITH_GRP_CDEastern, Christian	1.454e+00	4.2782142	4.112e-01	3.535	0.000408 ***
FAITH_GRP_CDIndigenous, Christian	8.569e-01	2.3558331	7.867e-01	1.089	0.276074
FAITH_GRP_CDIslam	1.080e+00	2.9456430	3.250e-01	3.324	0.000887 ***
FAITH_GRP_CDJudaism	1.354e+00	3.8739097	3.936e-01	3.441	0.000579 ***
FAITH_GRP_CDNon Denom, Christian	9.624e-01	2.6180304	1.973e-01	4.878	1.07e-06 ***
FAITH_GRP_ CDNone	1.025e+00	2.7880776	1.937e-01	5.294	1.20e-07 ***
FAITH_GRP_CDOther	5.989e-01	1.8201640	5.259e-01	1.139	0.254731
FAITH_GRP_CDPagan	6.432e-01	1.9025401	4.354e-01	1.477	0.139640
FAITH_GRP_CDReformative, Christian	8.204e-01	2.2714619	2.158e-01	3.802	0.000144 ***
FAITH_GRP_CDWestern, Christian	1.061e+00	2.8901029	2.003e-01	5.298	1.17e-07 ***
AFMS_YR_QY	-3.420e-03	0.9965863	1.372e-02	-0.249	0.803187
RANK_PDEPO1	2.179e-01	1.2434366	7.485e-02	2.911	0.003604 **
RANK_PDEPO2	2.923e-01	1.3394420	7.563e-02	3.864	0.000111 ***
RANK_PDEPO3	4.974e-01	1.6444199	8.790e-02	5.658	1.53e-08 ***
RANK_PDESA	4.844e-01	1.6231535	1.375e-01	3.523	0.000427 ***
RANK_PDESN	3.587e-01	1.4315149	1.021e-01	3.514	0.000441 ***
RANK_PDESR	4.865e-01	1.6266556	1.781e-01	2.732	0.006289 **
TOTAL_DEP_QY	-1.798e-01	0.8354091	1.322e-02	-13.604	< 2e-16 ***
RACE_CDAsian	-3.896e-01	0.6773616	9.082e-02	-4.289	1.79e-05 ***
RACE_CDBlack/African American	2.372e-01	1.2676789	7.088e-02	3.346	0.000820 ***
RACE_CDNative Hawaiian/Pacific Islander	-2.668e-01	0.7657999	1.692e-01	-1.577	0.114726
RACE_CDUnknown	-1.647e-02	0.9836639	9.609e-02	-0.171	0.863899
RACE_CDWhite	4.766e-03	1.0047776	6.488e-02	0.073	0.941438

					Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
					(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 25181  on 24822  degrees of freedom
Residual deviance: 24225  on 24772  degrees of freedom
AIC: 24327
Number of Fisher Scoring iterations: 4
```

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Cammack, J. H. (2020). *Predicting Army first-term attrition using logistic regression and time varying covariates* [Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/65485

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression, Technometrics. https://doi.org/10.2307/1268249

COMNAVRESFOR. (2022, May 17). *Navy reserve fighting instructions 2022* (ALNAVRESFOR 020/22). Department of the Navy. https://www.navyreserve.navy.mil/Portals/35/2022%20alnavresfor%20020%20navy%20reserve%20fighting%20instructions.pdf?ver=onnkjypakflkvmppao0uzq%3d%3d

DMDC (2022). *DMDC reporting system.* [Guard and Reserve by Demographics, Service Component]. https://dmdcrs-pki.dmdc.osd.mil/dmdcrs/protected/self-service-reports/reports

Gobea, G. A. (2019). *Predicting U.S. Army first-term attrition after initial entry training, part II* [Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/64167

Hosmer, David et al., (2013). Applied Logistic Regression, Third Edition. DOI 10.1002/9781118548387

James, G. et al. (2013), An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7

Marrone, J. V. (2020). *Predicting 36-Month Attrition: A Comparison Across Service Branches*. The Rand Corporation. https://www.rand.org/pubs/research_reports/RR4258.html

NAVPERSCOM. (2015). *Navy reserve status and categories* (MILPERSMAN 1001-100). https://www.mynavyhr.navy.mil/Portals/55/Reference/MILPERSMAN/1000/1000General/1001-100.pdf?ver=At9NToZ-A6qlpaq9050D2Q%3D%3D

Orrick, S. C. (2008). *Forecasting marine corps enlisted losses* [Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/4198

Prakash, V. (2016, May 24). 8 way of boosting performance of machine learning models. Analytics India Mag. https://analyticsindiamag.com/8-way-boosting-performance-machine-learning-models/

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Speten, K. J. (2018). *Predicting U.S. Army first-term attrition after initial entry training* [Thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/59593

United States Naval Reserve (n. d.) *About the Navy Reserve.* America's Navy. Retrieved July 11, 2022, from https://preprod.navy.com/who-we-are/about-navy-reserve

Zhou, Lifeng & Wang, Hong. (2012). Loan Default Prediction on Large Imbalanced Data Using Random Forests. TELKOMNIKA Indonesian Journal of Electrical Engineering. 10. 1519-1525. 10.11591/telkomnika.v10i6.1323.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California