NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

U.S. ARMY RESERVE RETENTION MODELING FOR MID-LEVEL LEADERS

by

Jordan T. Thomas

June 2023

Thesis Advisor:	Ruriko Yoshida
Co-Advisor:	Candice Farney
Second Reader:	Peter A. Nesbitt

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2023	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE U.S. ARMY RESERVE RETENTION MODELING FOR MID-LEVEL LEADERS			**5. FUNDING NUMBERS**
6. AUTHOR(S) Jordan T. Thomas			
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)

The Army Reserve wants to retain more mid-level leaders (defined as E6–E7 and O3–O4). In order to do this, the Reserve would like to know what influences a service member to stay or leave the Reserve military. Retaining mid-level leaders will greatly reduce the monetary costs of attrition, but will also increase readiness across the Army Reserve. This problem is constantly being evaluated by Human Resources Command and U.S. Army Recruiting Command; however, a new model consisting of multi-dimensional personnel data may provide new insight on what factors are predictive of a service member staying in the Army Reserve and may also provide insight on what programs are most effective at affecting retention. Once the model was created using logistic regression, it was evaluated using a test set to see how effective the model was at prediction as well as evaluate how influential each predictor was in forecasting retention of a service member.

14. SUBJECT TERMS Army, Reserve, retention, modeling, leaders			**15. NUMBER OF PAGES** 63
			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.

U.S. ARMY RESERVE RETENTION MODELING FOR MID-LEVEL LEADERS

Jordan T. Thomas
Major, United States Army Reserve
BS, North Carolina State University, 2011

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2023

Approved by: Ruriko Yoshida
Advisor

Candice Farney
Co-Advisor

Peter A. Nesbitt
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Army Reserve wants to retain more mid-level leaders (defined as E6–E7 and O3–O4). In order to do this, the Reserve would like to know what influences a service member to stay or leave the Reserve military. Retaining mid-level leaders will greatly reduce the monetary costs of attrition, but will also increase readiness across the Army Reserve. This problem is constantly being evaluated by Human Resources Command and U.S. Army Recruiting Command; however, a new model consisting of multi-dimensional personnel data may provide new insight on what factors are predictive of a service member staying in the Army Reserve and may also provide insight on what programs are most effective at affecting retention. Once the model was created using logistic regression, it was evaluated using a test set to see how effective the model was at prediction as well as evaluate how influential each predictor was in forecasting retention of a service member.

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

ADSO Active Duty Service Obligation

AFQT Armed Forces Qualification Test

AR Army Reserve

AUC Area Under the Curve

CART Classification and Regression Tree

CPT Captain

CSV Comma Separated Value

DoD Department of Defense

IDA Institute for Defense Analysis

IPERMS Interactive Personnel Elective Records Management System

IRB Institutional Review Board

IRR Individual Ready Reserve

ITAP Integrated Total Army Personnel Database

LASSO Least Absolute Shrinkage and Selection Operator

LT Lieutenant

MAJ Major

MOS Military Occupational Specialty

MSE Mean Squared Error

NCO Non-commissioned Officer

NG	National Guard
NPS	Naval Postgraduate School
OCS	Officer Candidate School
OMIP	Officer Menu of Incentives Program
ORSA	Operations Research Systems Analyst
OTRA	Other Than Regular Army
PCS	Permanent Change of Station
PDE	Person-Event Data Environment
RA	Regular Army
RCCPDS	Reserve Components Common Personnel Data System
ROC	Receiver Operating Characteristic
ROTC	Reserve Officer Training Corps
RPM	Retention Prediction model
SRB	Selective Retention Bonuses
SQL	Structured Query Language
TAPAS	Tailored Adaptive Personality Assessment
USMA	United States Military Academy

Executive Summary

All organizations find it difficult to retain well-trained, highly capable leaders. The Army Reserve is no different. Most studies in the past have taken a qualitative or quantitative approach that has had little effect on creating a predictive model that identifies the most meaningful predictors for retention or attrition. It is the goal of this study to use supervised statistical learning techniques that can provide senior leaders and commanders predictive power to identify factors that can have effect on the retention of Staff Sergeants, Sergeants First Class, Captains, and Majors within the Army Reserve. These ranks are the most needed and most under-filled billets in the Army Reserve.

In order to create our predictive models, we requested data from the Army's PDE system. After receiving the data from 11 different data sets, we merged data into one large data set. Predictors (variables) that were not meaningful to predicting attrition were dropped. Then we applied logistic regression and random forest models and we evaluated their performance using Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC).

This thesis provides predictors for both the enlisted and officer populations that indicate correlations within the specified populations. The model information is then analyzed and recommendations are given in order to inform senior leaders on where efforts of future research should be focused.

Based on the results of the models, logistic regression proved to be superior to random forest in producing greater ROC and AUC. Additionally, logistic regression has the added benefit of identifying correlation of predictors that affects attrition. The data suggests that the deployments, the amount of civilian education, award count, and birth country have the strongest effect on enlisted attrition. Gender, the amount of civilian education, and birth country proved to have the strongest effects on officer attrition.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I would like to thank my thesis advisors Dr. Ruriko (Rudy) Yoshida, MAJ Candice Farney, my second reader LTC Dr. Peter Nesbitt, and the rest of the OR faculty at Naval Postgraduate School. The last two years have been arduous and I couldn't have completed this degree without you. I also want to thank my family for their unwavering support.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Employee attrition has always been a wicked problem for government, the corporate world, and volunteer organizations alike. A wicked problem is defined as a social or cultural problem that is difficult or impossible to solve because of its complex and interconnected nature (Camillus 2008). According to the U.S. Secretary of the Army, Christine Wormuth in October 2022, "The nation risks falling behind in the race against China if it can't recruit enough Americans into the service" (Thomas 2022). Retaining exceptional talent is a challenge for all organizations because of the very fact that talented individuals are highly sought after. Furthermore, it is not always clear why individuals leave organizations. There is also a tendency to use anecdotal stories with small sample sizes that do not capture the whole story for the entire organization. This is where data becomes helpful in solving this problem of identifying how and why people leave the military. Although this may seem obvious, the very nature of identifying factors that accurately predict an individual's likelihood of attrition and their likelihood of staying at their current place of employment can be very challenging for organizations. This is not to say that identifying accurate predictors for potential attrition cannot be done, but it is clear that personnel attrition is not an easy or intuitive problem to solve.

An even more important subset of the problem of attrition has arisen within the Army Reserve (AR). According to "Operation Shaping Tomorrow -Building the AR End Strength," personnel end strength was the AR's highest priority (Jones 2022). Lieutenant General Daniels, Chief of the Army Reserve, in a written statement to Congress on June 7, 2022, said, "In an extremely challenging recruiting environment, the Army Reserve is tackling mid-grade officer and enlisted deficits" (Daniels 2022). There is an even greater need to identify factors that contribute to the attrition of mid-level leaders, which is defined in this context as staff sergeant to sergeant first class for enlisted personnel and captain to major for officers. There is a particular problem of retaining personnel around 10 years of service within the military. The midway point in a military career is such a tenuous time due to the Solider in question having several years of leadership experience and extensive time practicing their specific Military Cccupational Specialty (MOS) skills. These individuals are

greatly desired by corporate America. The fact that they are only half way to a full pension retirement, and given the right incentives, these highly skilled Soldiers have a tendency to leave the military.

Therefore, the AR has a particular desire to retain highly trained, effective leaders that are at the midpoint of their Army careers. Not only does the AR have hundreds of mid-level leader positions that remain vacant, but also there is a large cost associated with skilled Soldiers leaving the force. According to U.S. Army Recruiting Command, "the cost of training a Soldier in the first year can be up to $70,000 depending on the occupation" (Hughes et al. 2020). "Furthermore, based on accession and attrition rates from 2005 to 2012 across education tiers, a decrease by as little as 0.1% could save the Army as much as $4,550,000 per year" (Hughes et al. 2020). Given the current difficulties of recruiting and missing the recruitment goal by 25% in 2022, the cost to the Army becomes more than just monetary.

1.1 A Brief History

The Armed Forces are interested in retaining talent. There are, however, ranks that are more susceptible to attrition than others. In recent history, the retention of mid-level leaders has become an ever-increasing priority. A 2011 article titled "The Effectiveness of the Recent Army Captain Retention Program," which was published in the journal Armed Forces and Society stated, "Since 2001, Army officer shortages have received scrutiny because of increased attrition rates, particularly at the ranks of Captain (CPT) and Major (MAJ)" (Coates et al. 2011). In 2006, the Congressional Research Service predicted the Army would have a shortage of "at least 2,554 majors" by year 2013 (Coates et al. 2011).

Retention rates are a perennial problem and are extremely important to the Army. Even in the 1950s, studies were being conducted to identify how to affect retention rates. Nadel and Mowbray identified a large disparity in regards to retention within the officer corps. United States Military Academy (USMA) graduates had significantly different attrition rates as compared to Reserve Officer Training Corps (ROTC) and Officer Candidate School (OCS) graduates. The largest disparity found, however, were the retention rates between Regular Army (RA) and Other Than Regular Army (OTRA), which we today would call AR and National Guard (NG). Between the years of 1957 and 1965, the highest percent of OTRA junior officers retained after obligated service, which at that time was only two years, capped

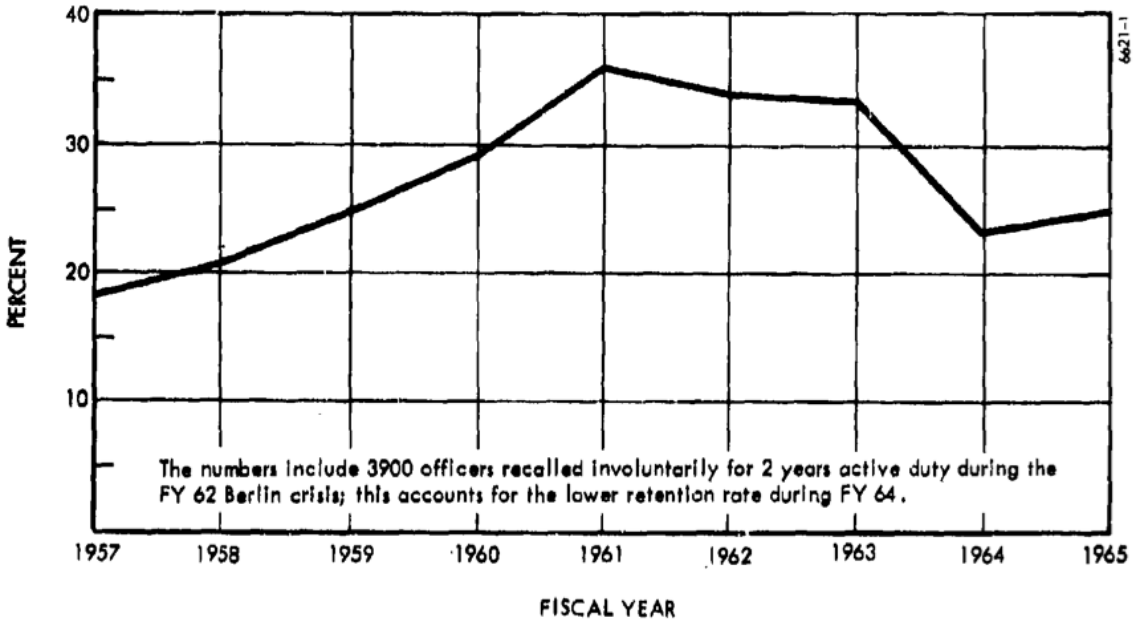

Figure 1.1. Retention rates of OTRA junior officers after obligated service (2 years). Historical percent of officers retained in any given year hovers between 20% and 35%. Source: Nadel and Mowbray (1966).

at roughly 35% as depicted in Figure 1.1. (Nadel and Mowbray 1966).

In September 2007, the Army enacted a retention program that targeted retention of captains called the Officer Menu of Incentives Program (OMIP) that gave several training or monetary incentives, which can be seen in Table 1.1, in order to retain officers through officers incurring an Active Duty Service Obligation (ADSO).

Table 1.1. Officer Menu of Incentives Program Options available in 2007. Adapted from Coates et al. (2011).

Option	Incenctive	ADSO
Critical Skills Retention Bonus	$25K-$35k	3 years
Graduate School Education	12-18 months	3 years
Military Training	Ranger, Airborne, etc.	1 year
Defense Language Training	Defense Language Institute	3 days/1 day in training

According to Coates et al., "The intent of the OMIP was to retain 14,000 officers; however, it fell short of the goal by more than 2,100 captains during its first year" (Coates et al. 2011). This is but one example of the many incentives programs that the Department of Defense (DoD) and the military branches of services have used in order to retain personnel. We have identified many factors that were not addressed in the OMIP program as a result of subsequent studies. Intrinsic rewards, extrinsic rewards, Permanent Change of Station (PCS) incentives, skills training, occupational achievements, generational idiosyncrasies (Generation X values versus Generation Y values), job security, recognition, social responsibility, competence, individual personality, and many more predictors have been identified as providing influence on a service member's tendency to remain in the military.

1.2 Purpose

It is the intent of this thesis to identify predictors of retention and attrition of mid-level leaders within the AR, provide predictive models that have a high level of accuracy, and provide a framework for future personnel analysis through supervised learning statistical models . The predictors that will be used are qualitative, quantitative, ordinal, and nominal.

This thesis work focuses on predicting losses. We have produced several models that can provide predictive power for future policies and decision making in order to increase mid-level leader retention. As of 2020, "only 5% of enlisted soldiers will reach 20 years of service, while 30% of officers" (Coates et al. 2011) will make it to the 20-year mark.

Because of these attrition rates, filling specific rank-based billets can be a challenge and forecasting these numbers continues to be difficult.

1.3 Benefit of this Study

This study provides decision makers with current, up-to-date models that can inform policy decisions and help with the structuring of retention programs that are effective in retaining targeted personnel. Though studies have been conducted in the past with the desired end-state of having a set of explanatory predictors that would provide high predictive power on retention, few have been conducted using random forest analysis with a focus on predictor sparsity, especially within the AR.

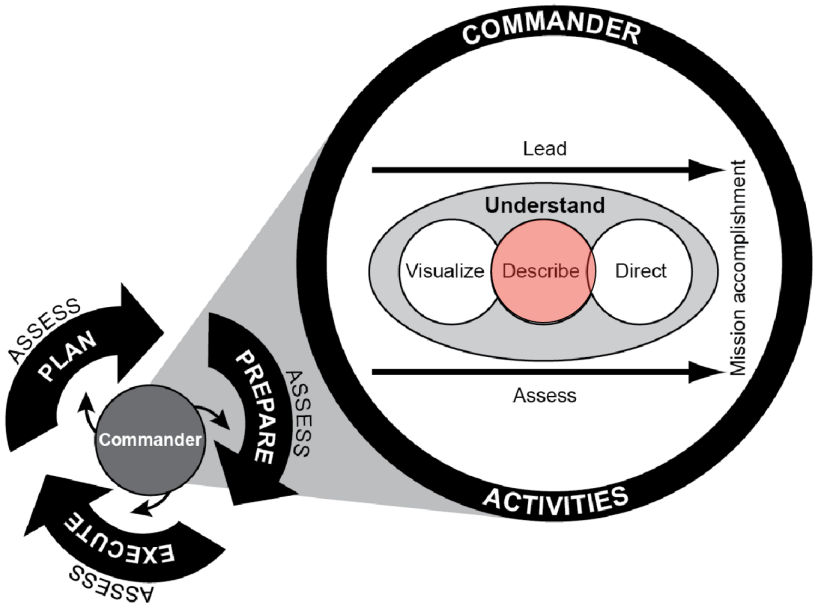

Figure 1.2. Operations process. This figure shows where attrition modeling helps commanders make decisions by describing the problem. This leads to better understanding, which in turn helps commanders to better plan, prepare, and execute a plan of action. Adapted from ATP 5-0.1 (2015).

"Developing a force that is lethal, efficient, and ready requires that leaders anticipate upcoming changes in the size and shape of the military workforce at a detailed level" (Pechacek 2022). This study will help with decision making within the AR by properly and precisely

forecasting attrition and retention and will fit into the operations process to help leaders describe and understand the problem, as is seen in Figure 1.2.

This study also strives to provide foundational ground work for the Army Reserve to determine future model selection and the selection of predictor data. Due to the lean structure of operations research within the AR, even tasks of great importance such as the next year's mission forecasting for how many recruits are needed does not receive the care and attention that is required in order to maximize total benefit to the Army Reserve. This thesis aspires to provide tangible feedback to AR decision makers on model selection for attrition models.

1.4 Reason for Focusing on Mid-level Leaders

The AR currently, as of December 2022, has a 26% vacancy among the ranks of Staff Sergeant (E6), Sergeant First Class (E7), Captain (O3), and Major (O4) and the historical trend between 2010 and 2020 can be seen in Figure 1.3. Vacancy in these mid-level leadership positions present a high level of risk to the Army Reserve, both monetarily and operationally. Mid-level leaders take national and strategic objectives and create operational plans in order to achieve national interest goals. These leaders are the conduit between high-level plans and low-level achievement.

Furthermore, this population has historically been the most difficult for the Army to retain as can be seen in Figure 1.4, which depicts how the Army develops talent and the most likely point of exit within an officer's career. As would be expected, this population is highly sought after by the private sector of business due to the mid-level Army leader's unique experiences and training.

1.5 Why Attrition Is So Difficult to Affect

Many models, both within the Department of Defense and the corporate world, have focused on increasing the retention of their most talented employees. Most, in essence, attempt to simplify an extremely complex system into a dozen or fewer meaningful predictors. The problem is that we are attempting to take a highly dimensional system and over simplifying it, which can be summarised as the "curse of dimensionality." As the number of predictors

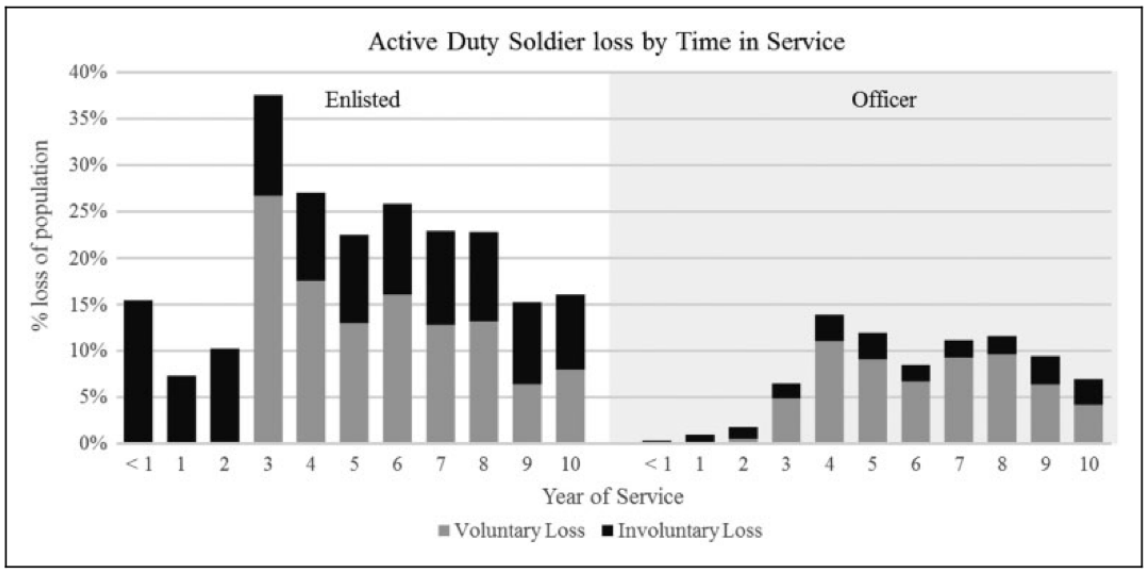

Figure 1.3. Percentage of soldiers (enlisted and officer) voluntary (End Term of Service, retirement, and resignation) and involuntary (training failure, medical, misconduct, and performance) separated by time in service between FY14 and FY17. Source: Hogue and Miller (2020).

grows and the dimensions of those predictors increase, the potential number of unique observations increases exponentially. Therefore, the question remains, how do you take an inherently complex system and create an accurate predictive model? With the ability of modern computing, the increasing size of accessible databases, and the development of more sophisticated algorithms, it is possible to create a model with hundreds of predictors where all the predictors, interactions between the predictors, and higher-order interactions can all be analyzed. Then a penalizing coefficient can be applied in order to find the most important predictors, as is done with Least Absolute Shrinkage and Selection Operator (LASSO) regression.

It isn't just the difficulty of the data that hinders progress in retention. Even if a model is created that is highly accurate at predicting retention and the predictors that most influence retention are identified, other challenges still arise in creating an effective strategy that increases retention. For example, legal factors can play a large role. What if it is found that young females with children who enter the military tend to stay longer than young, single males without children? Is it legal or ethical to only target one demographic within

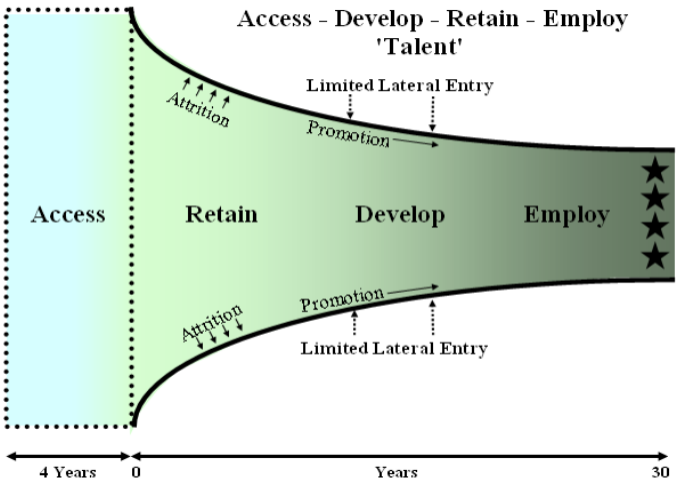

Figure 1.4. Army Officer Career Model. The OCM Model is used to describe the development and employment of talent within the Army. Source: Dabkowski et al. (2010).

a population? Answering these questions are beyond the scope of this study, but one can quickly see that reducing attrition is not always as easy as just providing another sign-on bonus..

Additionally, there is the difficulty of trying to quantify intangible attributes such as talent. It is intuitively assumed that it is extremely important for identifying and retaining the best officers. The difficulty arises when trying to quantify and assess talent across the Army. Dabkowski and his colleagues, from a paper in 2010 written about shaping senior leader officer talent, define talent as being the totality of an individuals skills, knowledge, and behaviors. One can quickly see how it would be difficult to quantify these dimensions across the entire Army accurately (Dabkowski et al. 2010).

Thus, attrition and retention modeling is of keen interest to the AR and other DoD forces, but the problem is quite difficult. High dimensional data, legal concerns, and quantifying intangible attributes are some of the reasons why identifying the causes of attrition can prove to be arduous.

CHAPTER 2:
Literature Review

There has been much research and study dedicated to identifying factors that influence attrition and retention within organizations. Though a great body of data exists on the subject, much of its helpfulness is questionable, especially in the context of Army Reserve Soldiers. Many studies are very narrow in focus, looking at one or two predictive tools, or are so broad and sweeping in general statements that sound decisions cannot be made based on the conclusions of the studies.

Hughes et al. (2020), looked at the effect of Armed Forces Qualification Test (AFQT) and Tailored Adaptive Personality Assessment (TAPAS) to predict attrition and reenlistment. Problematically, in the discussion of results, it was stated, "Generally, higher ability Soldiers were less likely to attrit" (Hughes et al. 2020). Also discussed was the fact that "Soldiers who placed greater emphasis on physical conditioning were less likely to attrit over time" (Hughes et al. 2020). Conclusions like these are common in attrition and retention data. Words used like "most likely," "indicate," and "suggest" are frequently used in studies on attrition data. Soldiers that have higher levels of physical fitness are intuitively more likely to stay in the military. What is so problematic about this answer, however, is that there is no quantification of how strongly physical fitness may predict attrition, or, at the very least, a comparison of how well physical fitness predicts it in relation to any other predictor.

In contrast to non-specificity, Coates et al. (2011) provide very telling results in their discussion of the Army's officer Menu of Incentives Program. They state, "Although 95 percent of the retained officers selected the financial incentive, military schooling and defense language training were not preferred options and thus probably not true incentives" (Coates et al. 2011). Such analysis provides commanders with direct, actionable information that proves to be valuable. Conclusions like these can guide Army policy for which incentives to provide in order to retain the best qualified Soldiers.

Similar to Hughes et al., Langkamer and Ervin (2008) show relationships of how command climate, morale, and organizational commitment play roles in the decision of a captain to stay in the Army until retirement, which can be seen in Figure 2.1. Arguably, these factors

play a very large role in retention, however, the study was a qualitative analysis that provided insight in direction, but not magnitude.

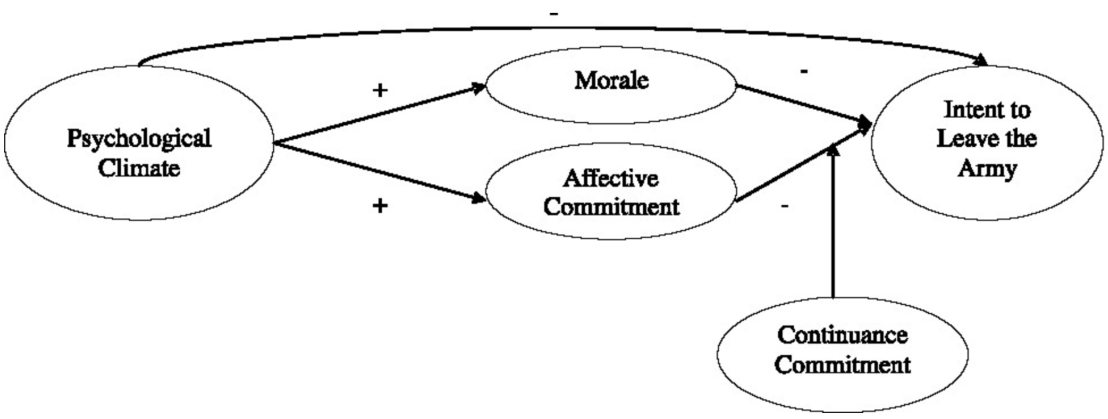

Figure 2.1. Illustration of hypothesized relationships between command climate and likelihood to leave the Army. The + and - symbols shows positive or negative relationships One shortcoming of this graphic is the lack of the ability to quantify relationships. Source: Langkamer and Ervin (2008).

2.1 Retention Research

The Army Reserve and other non-active components of the army employ more than 1 million soldiers, which makes it the largest employer of Americans besides Walmart (Hogue and Miller 2020). As would be expected, the personnel management and human resources machine within the Army is a colossal enterprise. Previous research on retention within the DoD, and specifically the AR, has focused on a number of common categorical approaches: "individual-level approaches (including demographically based and rational-economic actor models); institutional/organizational approaches; and approaches examining conflicts of primary and secondary commitments between reserve service and civilian employment, and between reserve service and family life" (Perry et al. 1991). Although these frameworks have produced fruitful outputs, a more meaningful and computationally intensive approach is available with today's technology.

Perry et al. (1991) argue that the typical attempt to increase retention is to take an individual-level selection approach that seeks to recruit applicants into the armed forces that have higher retention probabilities based on personal characteristics. Although much time and effort has

gone into attempting to identify strong predictors of attrition or retention, using personal characteristic models "have resulted in very low explanatory power; the R^2 ranged from .008 to .088" (Perry et al. 1991). This is not a very predictive result.

> Complicating manning challenges beyond sheer size are a myriad of laws, DoD directives, and Army policies that further constrain personnel decisions. A few examples include mandatory end strengths, limited lateral entry, rank distribution limits, mandatory promotion opportunities, and enlistment and reenlistment length requirements. Each of these limits push manning decisions further from an optimal solution. (Hogue and Miller 2020)

As is true in any optimization problem, the more constraints that are placed on a problem set, the higher the likelihood that the feasible solution will become worse and at best will stay the same.

A unique and state-of-the-art approach was taken by the Institute for Defense Analysis (IDA). They created a machine-learning model in 2019 for the same purpose of this thesis, which is to predict when military servicemembers will separate from the military. "The Retention Prediction model (RPM) uses machine learning algorithms and extensive personnel records to capture rich interactions in service characteristics and predict when individual servicemembers will separate from the military" (Pechacek 2022). This is a novel approach to DoD data analysis, however, the model created by IDA was specifically for Active Duty Soldiers and utilized a different modeling technique than this thesis work. Furthermore, this model was focused on comparing two separate randomly selected Soldiers and predicting which one would leave the service. This thesis focuses on aggregate data and total number of Soldiers leaving the service, not comparisons between two Soldiers. The IDA model was successful at predicting which of two Soldiers would leave active duty with a successful prediction rate of 88 percent.

In research conducted by Mackin and Mairs (1993), they explained that, "Information about the manpower costs and effects of personnel policy and other factors on the retention of high-quality active-duty commissioned officers, both for the aggregate Army and at the branch level, is critical to the development of the human resources necessary for an effective officer force." This is certainly true for the Reserve component as well. Continued research

in this ever-changing personnel landscape is necessary to maximize Army capability.

Another novel research approach was taken by Dabkowski et al. (2010) via creating a simulation model that focused on identifying ways to retain talent in the Army's higher ranks, specifically the rank of colonel (O-6). Instead of using data from the Army population, Dabkowski and his associates ran varying simulation models to see what affects would arise out of certain policy decisions, based on a set of predefined conditions. Their recommendations are as follows:

> If the Army adopted an incentive structure such that virtually no attrition occurred at the rank of Lieutenant Colonel, then the Army reaps marked improvements in all metrics. To implement such an incentive structure, the Army could announce the selection results for Colonel, before Lieutenant Colonels become eligible for retirement; each officer selected for Colonel would have very strong incentives to continue service for at least three years beyond his or her promotion date to lock in considerably higher retirement pay for the remainder of his or her life. (Dabkowski et al. 2010)

Although the results of their study are focused on changing outcomes at the O-6 level, it may be possible to implement similar analysis strategies to the mid-level leader problem.

2.1.1 Enlisted Retention and Attrition

Each year about 30 percent of the AR attrite and the majority of those losses come from the enlisted population (Perry et al. 1991). This has remained fairly constant over the past several decades. The highest achieved R^2 for Perry, Griffith, and White's study was 10 percent using all predictors. Although 10 percent is higher than previous studies, this figure still leaves much to be desired in terms of predicting retention and attrition.

A study conducted in 1999 showed that for enlisted personnel, "the most important variables by which to create these groups turn out to be race and gender. Generally white women have the lowest term completion and re-enlistment rates; those for non-white women and white men are similar; and those for non-white men are the highest" (Buttrey and Larson 1999). For many reasons, these findings may not be actionable or ethical to pursue.

A study conducted in 2005, which looked at Selective Retention Bonuses (SRB) for specific amounts of service already completed and per MOS found very meaningful data. "In general, the results for Zone A (between 17 months and 6 years of service) at all levels of occupational aggregation indicate that reenlistment bonuses have a positive and statistically significant effect on Zone A reenlistments. The magnitude of the effect varied by occupation, but a one-level increase in SRB at Zone A typically increases the reenlistment rate by three to seven percentage points, depending upon the occupation" (Hogan et al. 2005). These are helpful findings that are useful to commanders.

2.1.2 Officer Retention and Attrition

Officers may face unique circumstances that determine attrition and retention as compared to enlisted personnel, as can be surmised by Figure 2.2. "Statistical evidence suggests that Army battalion commanders are significant determinants of the retention of their lieutenants-especially high-potential lieutenants" (Spain et al. 2021). Professional relationships, nature of work, and perceived impact may be significantly different between officer and enlisted Soldiers. It is not, however, the goal of this analysis to address these differences. Each data point within the data set is assessed using the same predictors, whether enlisted or officer.

Enlisted and officer differences are, however, explored in relationship to predictors that were available within the Army's Person-Event Data Environment (PDE) system. Further research shows that "work experiences can influence Army captains' career intent. Specifically, psychological climate, affective and continuance commitment, and morale were predictors of intent to leave the Army before retirement" (Langkamer and Ervin 2008). As has been shown before, the three predictors that were used in Langkamer and Ervin's study may be helpful in explanatory analysis, however it is the goal of this thesis to create a model that has greater predictive power through analysing demographic data with higher dimensionality while utilizing supervised statistical learning techniques.

Much of the research up until this point that centers on retention and attrition have been qualitative in nature. In "Military Benefits that Retain Mid-Career Army Officers," a thesis written by Major Shane Roppoli, the eight common reasons why Soldiers stay in the Army are: "leadership, realistic expectations, quality healthcare, and mental services, military and civilian education, resilience to adversity, family support (housing), and pay and ben-

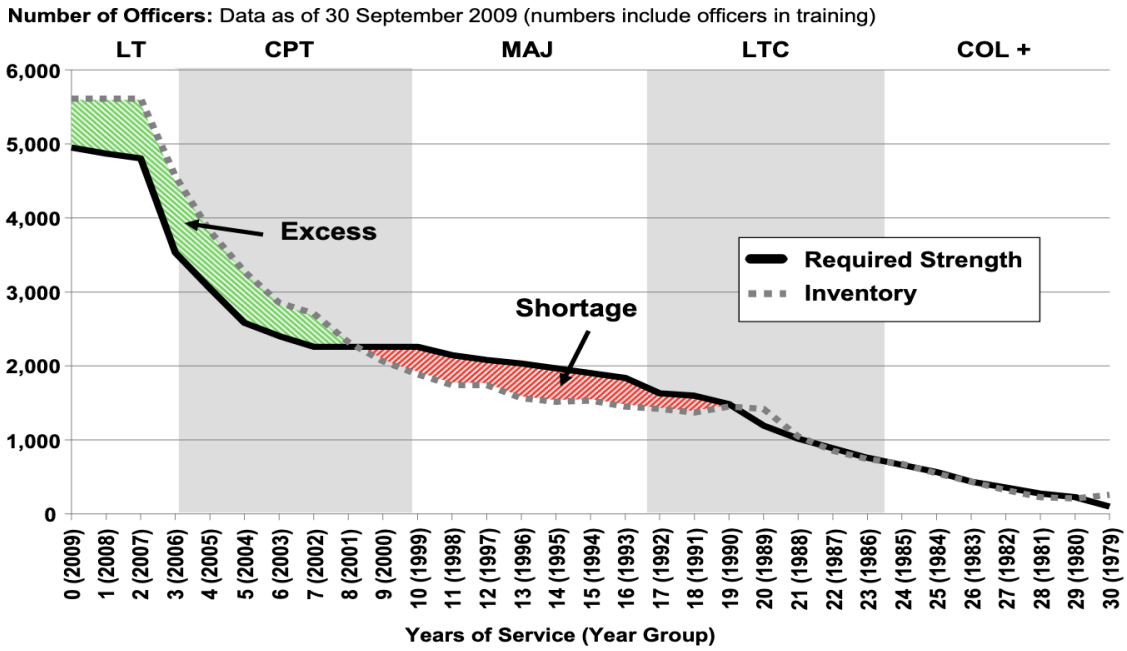

Figure 2.2. Officer imbalances between requirements and inventory as of 2009. Lieutenant (LT) and CPT ranks were over-strength, but MAJ have critical shortages. Source: Dabkowski et al. (2010).

efits" (Roppoli 2012). It would be prudent for the Army to track these predictors within an army officer's career, however, the data sets that are utilized in this study did not have access to data that captured these eight dimensions. Thus, we will be using demographic data that will be explained in more detail later.

In 2010, Dabkowski, Huddleston, and Kucik created a discrete event simulation to quantify the impacts of attrition on talent management within the Army officer corps. They posit that "Army Officers play a critical role in our nation's security strategy. Throughout a career of service, officers develop talents through a unique and rare set of experiences, education, and formal training. The demand by corporations for these talents, coupled with a distinct feature of the Officer Career Model, limited lateral entry, create significant retention challenges for the U.S. Army" (Dabkowski et al. 2010). The officer corps faces unique challenges to retention that are even more costly to the Army than the loss of enlisted Soldiers for a number of reasons. To begin with, there are much fewer Army officers. To compound the problem, the armed services do not have mid-career or senior-level lateral entries into service

except a few limiting circumstances (there are programs for medical doctors, psychologists, and niche technical fields to enter service mid-career). Unlike industry, which has the ability to pull mid- to upper-level talent from many sources, the Army has to grow their mid and upper-level leaders from the inside and from the bottom up. Lieutenants are the future leaders, but when officers leave the service the degradation to the talent pool is irreversible. Therefore, it is imperative for senior-level leadership to be able to properly forecast attrition rates in order to properly structure leadership demands.

2.2 The Current State

The demand for up-to-date retention and attrition models has been ever-changing and ever-present: "As the nation shifted from the industrial age to the information age in the mid 1980s, the Army experienced a decline in officer retention. Prior to the mid1980s, the Army could expect 60 percent of its officers to remain on active duty through eight years of service. Since the mid-1980s, approximately 40 percent of officers have continued past eight years of service" (Dabkowski et al. 2010). Though a corpus of models that have been produced in the past, there exists a necessity to produce an accurate model that predicts enlisted and officer attrition with present-day techniques that leverage current algorithms and computing power.

Though the current state of the Army Reserve is not as dire as it has been in years past, there still exists a need to improve end strength and retention. As is seen in Figure 2.3, there are some nuances that also must be considered. For example, it can be seen that O-5 (Lieutenant Colonel) end strength has hovered around 70 or 80 percent since 2015. However, it appears that the end strength and retention vastly improved in FY 21. This is not necessarily the case. The overall improvement comes from the extreme reduction in O-5 positions during FY 21, thus inflating the numbers in the years after.

Once the system reaches equilibrium, however, it is hypothesized that the Army Reserve may continue to have difficulty retaining enough O-4's to fill O-5 positions. According to Dabkowski et al. (2010), "limited lateral entry, therefore, places a premium on the Army developing and retaining the officer talent it needs in its mid and senior level positions. Officer retention data shows that the Army has been hemorrhaging officer talent for much of the past quarter century" (Dabkowski et al. 2010). Using supervised statistical learning

Grade	FY15	FY16	FY17	FY18	FY19	FY20	FY21	FY22	FY23
E5	89%	104%	108%	106%	104%	86%	87%	91%	91%
E6	75%	75%	71%	68%	65%	74%	70%	68%	68%
E7	57%	54%	54%	50%	58%	65%	63%	58%	58%
W3	55%	66%	75%	86%	96%	89%	82%	72%	70%
O1-O2	181%	201%	220%	204%	204%	204%	213%	222%	222%
O3	78%	83%	79%	83%	86%	83%	91%	84%	83%
O4	59%	56%	62%	67%	79%	84%	84%	89%	87%
O5	82%	77%	66%	67%	70%	74%	96%	97%	95%

Figure 2.3. Reserve end strength percentages by year and grade (rank). These estimates were produced at the end of FY22 and FY23 numbers are predictions. Source: Gobea (2022).

models, such as random forests, can help identify predictors that affect attrition and will help leaders make decisions on how to retain talent.

2.2.1 Random Forest Models

Random forests have been used in many problems such as imaging bio-markers for Alzheimer's disease, groundwater potential mapping, landslide susceptibility assessments, and many more research areas. The reason why random forest models are useful in attrition modeling is because it is a non-parametric method that produces good predictions that are interpretable, it can handle large data sets efficiently, and they are very useful when analyzing unbalanced data sets that have missing data. Random forests proved to be the best analytical tool employed by Karey Speten in a thesis titled "Predicting U.S. Army First-Term Attrition After Initial Entry Training" (Speten 2018). Another paper published in Baltic Journal of Modern Computing titled "Employee Attrition Estimation Using Random Forest Algorith" concluded that the most efficient tool the authors employed was random forest (Pratt et al. 2021). Therefore, given this body of evidence, we used random forest in order to predict attrition.

CHAPTER 3:
Methodology

This chapter will walk through the steps that were taken in order to receive the data, create the models, and analyze the results.

3.1 Data Environment

The United States Army has been collecting personnel data for centuries. One of the more recent repositories of data is the PDE. "The PDE is a consolidated data repository that contains unclassified but sensitive manpower, training, financial, health, and medical records covering U.S. Army personnel (Active Duty, Reserve, and National Guard), civilian contractors, and military dependents" (Vie et al. 2015). The PDE is the sole source of data collected during this study in order to create the supervised statistical learning models.

3.2 Analysis Environment

Due to restrictions on information, the PDE does not allow for the retrieval, removal, or copying of personnel information. Therefore, a computing environment on PDE servers was required. On this computing environment, RStudio (now known as Posit) was used in order to perform the statistical analysis and the production of visualizations. Open source packages were used to include ggplot2 (Wickham 2016), caret (Kuhn 2008), class (Venables and Ripley 2002), rpart (Therneau et al. 2013), ISLR (James et al. 2021), ROCR (Sing et al. 2005), gridExtra (Auguie 2015), cowplot (Wilke 2020), cluster (Maechler et al. 2022), and dbscan (Hahsler et al. 2019).

3.3 Supervised Learning Techniques

Supervised learning techniques are an extremely powerful set of statistical modeling. There are many different techniques that fall under this umbrella. A few popular algorithms include linear regression, logistic regression, naive Bayes, decision trees, K-nearest neighbor, support-vector machines, ensemble models, Classification and Regression Tree (CART), and neural networks (multi-layer perceptron). "Broadly speaking, supervised statistical

learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs" (James et al. 2021). The supervised learning techniques in this study include logistic regression and random forests. As will be shown later, these models have the ability to create predictions with high accuracy, depending on the inputs that are given. Logistic regression and random forest modeling may not be appropriate for all situations, but they are among the most used modeling techniques and have great potential for analyzing attrition and retention data.

3.4 Assessing Model Accuracy

In order to assess model accuracy, we need a specific measure that we can use to quantify how well one of the models performs against the others. We will be using the most common measure of accuracy, the Mean Squared Error (MSE). In order to assess the MSE, the data will be randomly sorted into a training and test split; 80 percent consisting of training data and 20 percent consisting of testing data. The test data will be the sole data used in assessing the MSE performance. Only using test MSE data will simulate the testing of each model against new data and will protect against over-fitting models. The following mathematical definition of MSE is used:

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - f(x_i))^2,$$

where n is the sample size, y_i is the value for the response variable for the ith observation and $f(x_i)$ is the prediction based on x_i, the ith observation under the given model. Other metrics of model accuracy include the outputs of a confusion matrix, which will be used in this study. The confusion matrix is a 2 by 2 matrix which contains counts from predicted values based on the given classification model and true values. comparing model. Using a confusion matrix model, we can obtain accuracy, sensitivity, and specificity used to identify the performance of the model.

Finally, the most important metrics that will be used to measure accuracy during this study are the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC). The ROC curve is an extremely useful tool in classification problems because it shows the performance of a classification model at all classification thresholds. The ROC curve graphs

the true positive rate against the false positive rate. Then when the AUC is calculated, this provides an aggregate measure of performance across all possible classification thresholds.

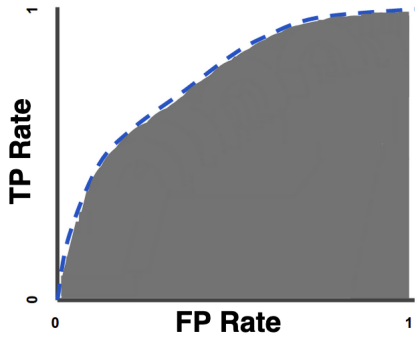

Figure 3.1. AUC example. The gray portion shows the cumulative area. As the area under the curve increases, the greater the accuracy of the model, in general. The ROC is a powerful tool in assessing a model's accuracy. The ROC plots the true positive rate of a model against the false positive rate. This provides an aggregate measure of performance across all possible classification thresholds. Here TP represents the true positive rate and FP is the false positive rate. Source: Google Developers: Machine Learning (2022).

AUC tends to be better for classification problems because it is scale-invariant and classification-threshold-invariant. Therefore, ROC AUC is a great tool in helping us determine how efficient our models are.

3.5 Requesting the Data

Before any manipulation of the data or analysis could be performed, several requests were made to the Army PDE repository (Knapp et al. 2018). Initial Institutional Review Board (IRB) approval was submitted through the Naval Postgraduate School IRB determination board, but since no individual persons' information would be linked to their name or any identification method, it was determined that an IRB would not be required. Then 11 data sets were identified within the PDE that contained the information that would be needed in order to perform a detailed analysis. The data set names can be seen below along with explanations of the data sets. All data utilized was specifically bounded by January 1, 2010 to December 31, 2021.

The following are explanations of the data sets, which are also shown in Figure 3.2:

1. MV_FOUO_UR_TRANS_RC_ARMY_V3-N1-11_21_2022 - Reserve Components Common Personnel Data System - Transaction [Army] Unit Identification Code and rank exposed. This data set has many metrics with information about Soldiers across the entire Army Reserve.

2. MV_INV_TRN_HIST_ARMY-N1-11_21_2022 - Individual training history of Soldiers across the Army Reserve.

3. MV_ITAPNOW_SOLDR_TRNG-N1-11_21_2022 - Individual Soldier training records according to the Integrated Total Army Personnel Database.

4. MV_ITAPNOW_SOLDR_AWD-N1-11_21_2022 - Awards records according to the Integrated Total Army Personnel Database (ITAP).

5. MV_ITAPNOW_MAJ_PERS_ACTN-N1-11_21_2022 - Major personnel action records according to the ITAP. This data set contains information concerning Soldier movement in and out of service as well as across components (Active, Reserve, Individual Ready Reserve (IRR), National Guard).

6. MV_ITAPNOW_MIL_EDUC-N1-11_21_2022 - Individual Soldier military education records according to the ITAP.

7. MV_ITAPNOW_PERSON_DEPLOY-N1-11_21_2022 - Records of movement of a person to an area of military operations.

8. MV_ITAPNOW_PERSON-N1-11_21_2022 - Demographic information about a person. Contains: -Date/location of birth and citizenship information -Number of child and adult dependents -Non-military education level -Demographics: Race, ethnicity, marital status, religion, sex -Prior military service and selective service class.

9. MV_ITAPNOW_CIV_EDUC-N1-11_21_2022 - The type of civilian education and when it was attained, the total number of years and credits completed, and the amount of money paid by the US towards post-secondary education. If the person is currently pursuing a degree, it contains an indicator for that and a projected date the degree will be received.

10. TA_IPERMS_DEROG_V2-N1-11_21_2022 - Interactive Personnel Elective Records Management System (IPERMS): Derogatory Reports. The data set contains the bad paper file indicating the time and type of bad paper (e.g. Article 15) that an Army soldier has been issued.

11. MV_MASTER_RC_ARMY_QTR_V3-N1-11_21_2022 - Reserve Components Common Personnel Data System (RCCPDS): Master. This is an encoded version of the Master File, where social security numbers are masked with 12 character PDE Identifiers. The entity provides an inventory of all individuals in the Army National Guard, Army Reserve at a point in time.

Data Requests (11)

Request Name	Entity Count	Recurring	Requested On	Requested By	Request Status	Workflow Step	Cart
MV_FOUO_UR_TRANS_RC_ARMY_V3-N1-11_21_2022	1		11/21/2022 3:28 PM	Candice Farney	Approved	Data Provision	Add to Cart
MV_INV_TRN_HIST_ARMY-N1-11_21_2022	1		11/21/2022 1:49 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_SOLDR_TRNG-N1-11_21_2022	1		11/21/2022 12:28 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_SOLDR_AWD-N1-11_21_2022	1		11/21/2022 12:27 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_MAJ_PERS_ACTN-N1-11_21_2022	1		11/21/2022 12:26 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_MIL_EDUC-N1-11_21_2022	1		11/21/2022 12:23 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_PERSON_DEPLOY-N1-11_21_2022	1		11/21/2022 12:21 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_PERSON-N1-11_21_2022	1		11/21/2022 12:20 PM	Candice Farney	Approved	End	Add to Cart
MV_ITAPNOW_CIV_EDUC-N1-11_21_2022	1		11/21/2022 12:19 PM	Candice Farney	Approved	End	Add to Cart
TA_IPERMS_DEROG_V2-N1-11_21_2022	1		11/21/2022 12:17 PM	Candice Farney	Approved	End	Add to Cart
MV_MASTER_RC_ARMY_QTR_V3-N1-11_21_2022	1		11/21/2022 12:16 PM	Candice Farney	Approved	End	Add to Cart

Figure 3.2. Data sets requested from the Army PDE repository. These are the original data sets that were received for analysis. These data sets were then merged, cleaned, and extraneous predictor columns were removed in order to reduce dimensionality and increase efficiency of the models. The Request Name column is the name of the data within the PDE. Requested On shows the date the data was requested. Candice Farney helped with the collection of the data and is reflected in the Requested By column. The Request Status column shows the person who requested the data where it is at in the approval process. The Workflow Step shows where the data is in the data provisioning process.

3.6 Data Collection and Manipulation

A virtual machine was accessed through PDE using Citrix in order to be able to access the data without transferring the data to a local machine. The data was then moved from the PDE repository to the virtual machine through Structured Query Language (SQL) queries. Finally, the data was manipulated and analyzed through RStudio. During this time,

it became increasingly important to reduce the dimensionality of the data. After merging the 11 data sets based on each individual Soldiers' unique identification number, there remained 130,000 rows of data. There were many data columns out of the 53 that were present that mostly contained NA or NULL values. Several weeks were taken to remove the NA columns and to remove predictors that would be difficult to use for the study. Eventually the data was reduced from 53 to 15 columns and from 130,000 observations to 74,000.

3.6.1 Workflow

As can be seen in Figure 3.3, the process from requesting the data to model completion was pretty straight-forward, albeit time intensive. Data requests were made to the PDE for 11 data sets and the data was provisioned in 6 days.

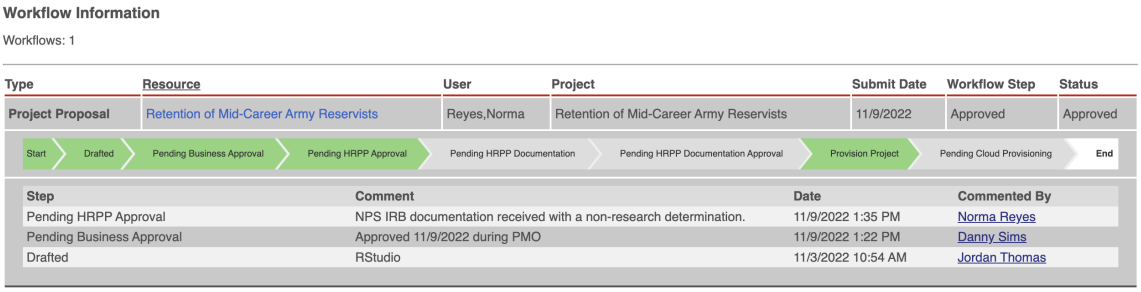

Figure 3.3. This is a depiction of the original data set request workflow. A request for resources and then a draft for the project was submitted to PDE. There was an IRB, approval process, and then the project was resourced with the requested data sets and a cloud computing environment was provided with RStudio.

Once the data was provisioned, SQL queries were necessary in order to sort through the data and create Comma Separated Value (CSV) files in order to load into RStudio, since the total number of rows in the initial data sets included 52 million observations. In order to reduce the number of observations, all were excluded except those that contained E-6, E-7, O-3 or O-4 data, which were the target populations of interest. This reduced the data down to 130,000 observations. To further decrease the number, only Soldiers within the Army Reserve were included, which reduced the data down to 74,000 observations. The data was then grouped into either Enlisted data or Officer data in order to create the supervised learning models.

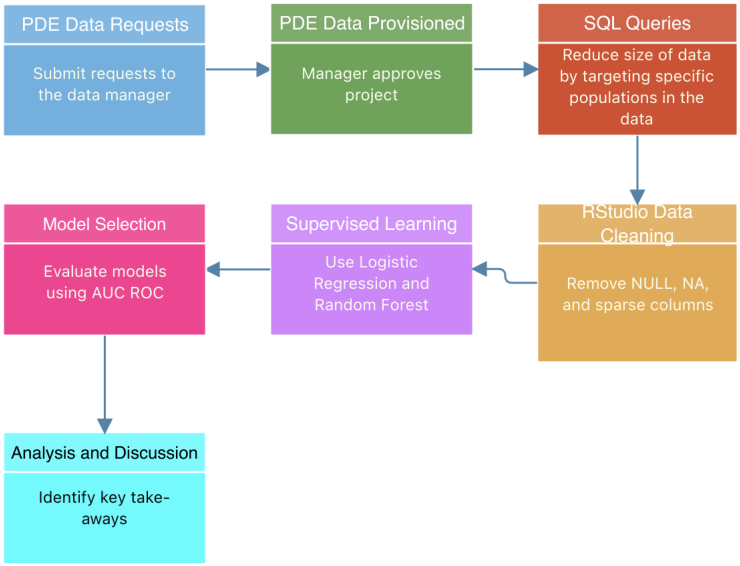

Figure 3.4. Pictorial depiction of the data workflow. First, data request and provisioning through PDE. Third, SQL queries for specific data that only included E-6, E-7, O-3, O-4 ranks. Additionally, only Army Reserve Soldiers were included. Fourth, RStudio cleaning. Fifth, implementation of supervised learning techniques. Finally, model selection and analysis.

3.6.2 Supervised Learning Algorithms

Two algorithms were used in the creation of predictive models using the data sets.

1. Logistic Regression – The first model that was used in order to produce predictions was a logistic regression model for both Enlisted Soldiers and Officers. We used the logistic regression model as a base-line even though we have many categorical variables in our data set.

2. Random Forests – Random forest was identified as being a good algorithm because it is a non-parametric method. We have many categorical predictors and its particular ability to identify predictors that produce a strong influence to the response is good.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Results

The results of the models employed are of utmost interest. Much like previous attrition or retention models utilized in other contexts or different industries, the random forest model performed very well. However, what is most surprising is how well the logistic regression performed in this classification context.

4.1 Model Variable Selection

We used a combination of tools to analyze model performance. We analyzed the confusion matrix, the ROC curve, and the AUC of the ROC in order to evaluate model performance. The goal was to maximize accuracy, sensitivity, specificity, and ROC AUC. In addition, it was important to identify the predictors that had the most effect on each model's performance.

Due to the high dimensionality of the data and the use of redundant columns that would introduce multicollinearity to the model, the following predictors were included in both the enlisted and officer models and the rest were removed from the data set:

1. Derogatory Documents in Soldier's File: Binary Classification
2. Deployment Type Code: Multiple level classification
3. Deployment Month Quantity: Multiple level classification
4. Civilian Education Years Quantity: Multiple level classification
5. Birth Country Code: Multiple level classification
6. Citizenship Country Code: Multiple level classification
7. Dependant Quantity: Integer
8. Ethnic Group Code: Multiple level classification
9. Marital Status Code: Multiple level classification
10. Sex Code: Binary Classification
11. Award Count: Integer
12. Paygrade: Binary classification
13. Service Entry Date: Multiple level classification
14. Known Deployments: Binary classification

15. Separated from military service: Binary classification

It became clear that other predictors that were discarded were not useful with the data set (filled with NA or NULL values) or the predictors created exceptional computational complexity with too many levels of classification. The goal was to create a sparse model that performed well, meaning that it would predict the proper classification for separation from the military with an accuracy of 75 percent or greater.

4.2 Enlisted Logistic Regression Performance

After creating a logistic regression model that included all the predictors, the confusion matrix depicted in Figure 4.1 was produced.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3467  312
         1 1179 1184

               Accuracy : 0.7572
                 95% CI : (0.7463, 0.7679)
    No Information Rate : 0.7564
    P-Value [Acc > NIR] : 0.4478

                  Kappa : 0.4494

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7462
            Specificity : 0.7914
         Pos Pred Value : 0.9174
         Neg Pred Value : 0.5011
             Prevalence : 0.7564
         Detection Rate : 0.5645
   Detection Prevalence : 0.6153
      Balanced Accuracy : 0.7688

       'Positive' Class : 0
```

Figure 4.1. Enlisted Logistic Regression Confusion Matrix. This shows a .75 accuracy, which will be considered the baseline for measuring future performance. The sensitivity and specificity were also between 75 and 80 percent.

This figure shows an accuracy of .75, a sesitivity of .75, and a specificity of .79. Also, an ROC curve was produced with an AUC of .836. The original logistic regression had a very

poor accuracy of 25 percent when the probability cutoff was set to 0.5, meaning, if the prediction of the logistic regression produced a prediction of 0.5 or higher, the classification of predicted separation from the military would be set to 1. Otherwise, the classification was set to 0. Since the results were very poor, the cutoff probability value was set to 0.3 and accuracy improved from 25 percent to 75 percent.

According to the model, the most important predictors are deployment type, number of years of civilian education, the birth country of the Soldier, the award count of the Soldier, and the pay-grade. Many of these predictors make intuitive sense, but we will discuss them more in the discussion of results section. With an AUC of .836, the logistic regression model performs well for the purposes of predicting whether an E-6 or E-7 will exit the military. We do, however, want to take a look to see if a random forest model would perform better.

4.3 Enlisted Random Forest Performance

A random forest model was produced using the randomForest library within R. This random forest model received the same inputs as the above logistic regression model and its confusion matrix is shown in Figure 4.2.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4291  765
         1  350  740

               Accuracy : 0.8186
                 95% CI : (0.8087, 0.8281)
    No Information Rate : 0.7551
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.459

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9246
            Specificity : 0.4917
         Pos Pred Value : 0.8487
         Neg Pred Value : 0.6789
             Prevalence : 0.7551
         Detection Rate : 0.6982
   Detection Prevalence : 0.8226
      Balanced Accuracy : 0.7081

       'Positive' Class : 0
```

Figure 4.2. Enlisted Random Forest Confusion Matrix. This shows an increase in accuracy above the baseline logistic regression model above. Sensitivity went up and specificity went down.

As can be seen from the figure, the random forest model performed marginally better than the logistic regression in classifying whether an E-6 or E-7 would separate from the military with an accuracy of .82. The sensitivity of the random forest was significantly better with a sensitivity of .93, but a lower specificity of .49. The ROC curve that was produced had an AUC of .695, which performed worse than the logistic regression model and can be seen in Figure 4.3.

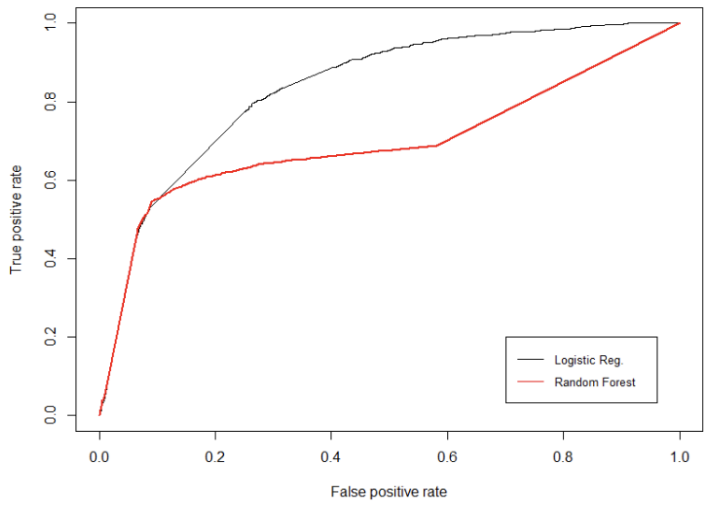

Figure 4.3. Enlisted Logistic Regression vs. Random Forest ROC Curves. Logistic regression had an AUC of .836 versus random forest with .695, creating a difference of .141. This shows that although the random forest had higher accuracy, in a general way the logistic regression model performs better.

4.4 Officer Logistic Regression Performance

The difference between the officer results and the enlisted were astounding. Where 32 percent of E-6's and E-7's separate from the Army Reserve, 98 percent of O-3's and O-4's separate from the Army Reserve, according to the collected data. This higher separation rate intrinsically made the officer models more accurate by virtue of the fact that almost all officers leave the military by the rank of O-4.

The officer logistic regression model had an accuracy of .997, a sensitivity of .98, and a specificity of .997. The ROC AUC for the logistic regression model was .997. These results are not surprising due to the high attrition rate of officers within the Army Reserve. Due to the high accuracy, sensitivity, and specificity of both the logistic regression and random forest models, the confusion matrix will not be shown.

4.5 Officer Random Forest Performance

The random forest model performs as well as the logistic regression model, which is what we would expect to see. The accuracy is .997, sensitivity is .984, and the specificity is .997, which matches the results of the logistic regression model.

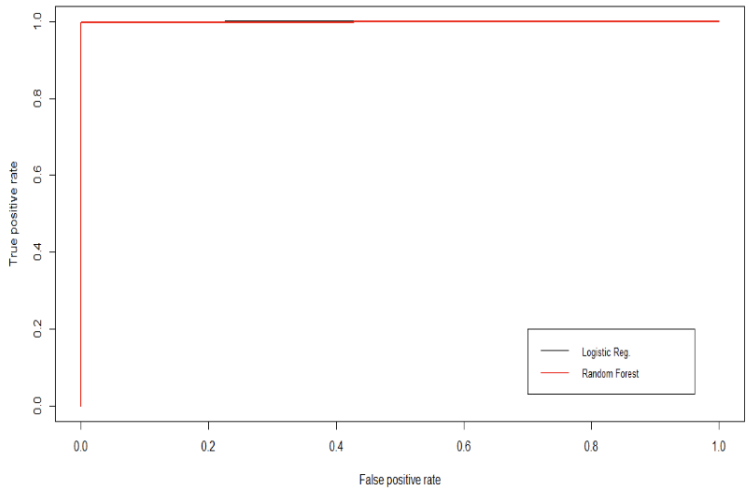

Figure 4.4. Officer Logistic Regression vs. Random Forest ROC Curve. Both models performed very well. Due to both models performing similarly, logistic regression is preferred due to its explainability and reduced computational complexity.

When taking a look at the ROC curve for the officer attrition predictions, it is unsurprising that the AUC is .998. Even though this is exceptional performance, there is virtually no difference between the models, therefore it makes sense to use the logistic regression model in order to model officer attrition due to it being less computationally intensive and the

ability to draw inference between the predictors and their effects on outcomes.

CHAPTER 5:
Discussion and Recommendations

The findings and recommendations of this research are specific to the models that we produced based on a given dataset. Inferences may be drawn from the results, although it is incumbent for leaders to help direct and guide future research and draw conclusions based on the results of the data.

5.1 Enlisted Models Discussion

When considering both enlisted models, the logistic regression model produces the greatest AUC and has the added benefit of interpretability. As shown in Table 5.1, the summary of the enlisted logistic regression model produced statistically significant variables:

Table 5.1. Enlisted Model Predictor Levels of Importance. This table shows the most important predictors that were identified in the enlisted logistic regression model. Deployment Type, Civilian Education Years, Birth Country, Award Count, and Pay Grade all had p-values less than .001.

Predictor	Value	P-value	Estimate	Importance
Deployment Type	Unknown	2.55e-05	-3.689	***
Civ Education Yrs	12	.0005	-2.321	***
Civ Education Yrs	Unknown	.0001	-2.491	***
Birth Country Code	Unknown	3.6e-06	1.277	***
Award Count	5 - 23	3.1e-09	-1.555	***
Pay Grade	E-7	6.6e-10	-0.419	***
Deployment Type	V	.0018	-3.044	**
Civ Education Yrs	15	.0019	-2.100	**

In terms of decision making, Table 5.1 shows that the type of deployment a Soldier has been on, the number of years of civilian education, their birth country, the number of awards the Soldier has received, and their current pay-grade are the most influential predictors to the outcome of separation from the military. What is the most surprising about results from our analysis is the fact that all but one predictor was negatively correlated. This means that the more education the Soldier has, the higher the rank of the Soldier, and the more awards that the Soldier has, the more likely the Soldier will not attrite and will promote to the next level or continue in service. This is very good news for the Army. This data shows that enlisted Soldiers who are more accomplished (and presumably more desirable to retain) are more likely to continue in service.

This does not, however, address the fact that historically speaking, E-7 billets are the ones that are least filled out of the ranks being analyzed, with an average fill rate of 57 percent over the past decade. The data suggests that retention may be affected by targeting Soldiers that have had specific deployments and also focusing on education benefits, which has been a long-standing tactic within the DoD.

The data also helps leaders to consider directions for future research and provides evidence to leaders on what may be most causal to attrition. Although there are many anecdotes or theories on what has caused the personnel shortages over the past decade, this data has concrete numbers that show what was most influential to the model. The data suggests that AR leaders should first focus on investigating the relationships between deployment, civilian education, and award counts before looking at theories such as the "wokeness" of the military as the causal agent of the reduction of personnel.

In addition to retention, the model can be used for forecasting attrition within these ranks so that better recruitment and retention decisions can be made. In much the same way that lieutenants are well over-staffed, it may be more beneficial for the AR to over-staff lower enlisted positions and to target them with educational incentives in order to produce an effect on E-7 numbers.

5.2 Officer Models Discussion

Looking at the logistic regression model that was produced, different predictors were important in comparison with the enlisted model. As can be seen below, gender, number of years of education, and birth country influenced the resulting prediction the most. Males were positively correlated with attrition, which means a larger proportion of female officers stay in comparison with male officers. Also, higher levels of education were negatively correlated with attrition. Finally, being a US born officer was positively correlated with attrition at the O-3 and O-4 level, all of which is shown in Table 5.2.

Table 5.2. Officer Model Predictor Levels of Importance. This table shows the most important predictors that were identified in the officer logistic regression model. Gender and Years of Civilian Education had a p-value less than .001.

Predictor	Value	P-value	Estimate	Importance
Gender Code	Male	2.54e-05	2.188	***
Civ Education Yrs	Unknown	1.54e-05	-3.612	***
Civ Education Yrs	16	.0018	-3.235	**
Birth Country Code	US	.0413	1.865	*

As was discussed with the enlisted data, not only can these predictors be used to make policy decisions about incentives programs that target at-risk populations, but a more accurate assessment of the force strength and manpower needs could help close gaps in personnel shortfalls and save money by properly resourcing Army Recruiting Command to meet their mission.

5.3 Recommendations

We must remember that correlation within the data does not necessarily mean that the given predictor is the root cause of attrition. When we look at the enlisted data, we can assume the reason why Soldiers with more awards stay in the Army Reserve is because they are hardworking, they are go-getters, and they excel in the military. We would not assume that

the reason why Soldiers with more awards stay longer in the military is because they have been given many awards. That being said, the kind of deployment (or maybe just the fact that the Soldier has deployed at all) has an effect on retention in a non-intuitive way. It is assumed that Soldiers who deploy have a lower likelihood of staying in the military, but the model does not agree with this assumption. Therefore, it would be helpful to conduct further research on the affects of deployment and retention.

When looking at the results of the officer model, there is a strong correlation between gender and attrition. It is unclear on why that may be. We do know, however, that there is an effect. It is recommended that this area be a focus for future research to see if causes can be established.

Education also plays a role in Soldiers attriting. Therefore, it would be prudent for the AR to conduct further analysis on how to affect these populations. It is unclear from the regression models that were created what the causal root, either for enlisted or officers, is for why education is negatively correlated with attrition. This does not mean that the models that we created are somehow a failure. Quite the opposite. These models now give us a direction to continue to look in order to discover causes instead of chasing every new theory that materializes out of popular media sources. This is the main value of these linear regression models. It gives the AR direction to start looking for cause and effect. There are many theories that are popular today, such as there isn't enough childcare, or the military is becoming too woke, or the military is no longer competitive in monetary compensation due to the rising minimum wage, but this research gives us a clear indicators on where to focus our efforts and our research dollars.

In addition, rather than pinpointing root causes, these models are useful at making an aggregate prediction on how many Soldiers will attrite in a given year and this number can help Recruiting Command determine how to set the recruiting mission for the year. For example, let's say that the AR is given a mission of maintaining 188,000 soldiers in the Reserve. Not only does the AR need to know how many soldiers they are currently short of that target, but also the number of soldiers that are expected to lose from attrition in a given year.

These models could also be used for both recruiters and retention Non-commissioned

Officers in order to estimate a risk score for future attrition given a new recruit or a Soldier that is seeking to re-enlist. Each model may help recruiters to make decisions on how to spend their time recruiting (seeking more desirous populations) or may help a retention Non-commissioned Officer (NCO) to utilize different incentive opportunities to better retain Soldiers, especially those of specific, high-demand MOS's.

5.4 Future Work

Though the models that were created performed fairly well (both performing above .80 AUC), there is always an opportunity to improve accuracy and an opportunity to use different predictors that were not present in the study in order to pinpoint causes of or correlates of attrition. It is recommended that even more data that is available in the PDE be used to see if other predictors are more explanatory than the ones looked at in this study. In addition, more models should be used to analyze the data. There is opportunity to us other supervised learning techniques such as support vector machines, k-nearest neighbor, time series analysis, and CART. Also, as was done by Pechacek and colleagues, there is ripe opportunity to use unsupervised learning techniques such as principle component analysis, k-means clustering, and hierarchical clustering to understand the structure of the data and to see whether there is a hidden data structure in order to improve classification accuracy for supervised learning models.

In addition, this work strongly points to the need to research why deployment, civilian education, birth country, and award count have such strong influence on the regression model for enlisted soldiers. It is also recommended to conduct further research on why gender, civilian education, and birth country have such a strong effect on the officer regression model. Though we can't yet establish causes for attrition, the models produced can point decision makers in the direction of where to look and to focus future efforts. Rather than investing money into determining if there are enough childcare programs for AR soldiers, or any other number of ideas, we have a direction in which to head in order to maximize our expectation of results.

5.5 Conclusion

The logistic regression models that were produced are good tools to use in order to predict attrition and retention within the ranks of E-6 to E-7, and O-3 to O-4. Both models have an AUC greater than .80. These prediction models can be used to help set the recruiting mission each year for the AR, to create a prediction of retention score for Soldiers that are re-enlisting, and can identify important predictors of attrition so that more research may be done to determine correlation or causation of attrition. However, it is important for leadership to provide direction based on these findings. Although these models performed well, we predict that these models will be fluid and ever-changing, creating a need to re-evaluate attrition models often. This thesis has set a foundational approach to creating and evaluating future attrition models.

List of References

Auguie B (2015) *gridExtra: Miscellaneous Functions for "Grid" Graphics*, http://CRAN.R-project.org/package=gridExtra, r package version 2.0.0.

Buttrey SE, Larson HJ (1999) Determining characteristic groups to predict Army attrition. Technical Report NPS-OR-99-003, Dept. of Operations Research, Monterey, CA, USA, https://calhoun.nps.edu/handle/10945/15397.

Camillus JC (2008) Strategy as a wicked problem. Harverd Business Review, https://hbr.org/2008/05/strategy-as-a-wicked-problem.

Coates HR, Silvernail TS, Fulton LV, Ivanitskaya L (2011) The effectiveness of the recent Army captain retention program. *Armed Forces and Society* 37(1):5–18.

Dabkowski MF, Huddleston SH, Kucik P, Lyle D (2010) Shaping senior leader officer talent: How personnel management decisions and attrition impact the flow of Army officer talent throughout the officer career model. *Proceedings of the 2010 Winter Simulation Conference* (IEEE), 1407–1418.

Daniels LJ (2022) On the 2022 posture of the United States Army Reserve: America's global operational reserve force. Subcommittee On Defense Committee On Appropriations United States Senate, Washington, DC, USA, https://www.usar.army.mil/Portals/98/Documents/CAR/LTG%20Daniels_Written%20Statement_FY23%20SAC-D%20Guard-Reserve%20Hearing%20(6_7_22).pdf?ver=auCuAdKMciGbVDpHWzaQ6A%3d%3d.

Gobea MG (2022) Strategic analysis branch – Strength report. Strategic Analysis Branch - Strength Report, Fort Belvoir, VA, USA.

Hahsler M, Piekenbrock M, Doran D (2019) dbscan: Fast density-based clustering with R. *Journal of Statistical Software* 91(1):1–30, https://doi.org/10.18637/jss.v091.i01.

Hogan PF, Espinosa J, Mackin PC, Greenston PM (2005) *A Model of Reenlistment Behavior: Estimates of the Effects of Army's Selective Reenlistment Bonus on Retention by Occupation*. Study report : 2005-02 (U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va).

Hogue L, Miller BJ (2020) Harnessing inertia to improve Army enlisted service length: A case for opt-out enlistment contracts. *Armed Forces and Society* 46(1):116–131.

Hughes MG, O'Brien EL, Reeder MC, Purl J (2020) Attrition and reenlistment in the Army: Using the tailored adaptive personality assessment system TAPAS to improve retention. *Military Psychology* 32(1):36–50.

James G, Witten D, Hastie T, Tibshirani R (2021) *An introduction to statistical learning: With applications in R*, second edition. ed. Springer texts in statistics (Springer, New York, NY).

Jones LK (2022) Operation Order 22-029 (Operation Shaping Tomorrow -Building the Army Reserve (AR) End Strength). U.S. Army Reserve Command and USARC G-1, Fort Bragg, NC, USA, https://www.usar.army.mil.

Knapp D, Asch BJ, DeMartini C, Ruder T, Hanley JM (2018) *Using the Person-Event Data Environment for Military Personnel Research in the Department of Defense: An Evaluation of Capability and Potential Uses* (RAND Corporation, Santa Monica, CA), https://doi.org/10.7249/RR2302.

Kuhn M (2008) Building predictive models in r using the caret package. *Journal of Statistical Software, Articles* 28(5):1–26, https://doi.org/10.18637/jss.v028.i05.

Langkamer KL, Ervin KS (2008) Psychological climate, organizational commitment and morale: Implications for Army captains' career intent. *Military Psychology* 20(4):219–236.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2022) *cluster: Cluster Analysis Basics and Extensions*, https://CRAN.R-project.org/package=cluster, r package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).

Nadel AB, Mowbray JB (1966) *Motivation and Retention in the U. S. Army* (United States Department of the Army. Office, Chief of Research and Development, Washington, D.C.).

Pechacek J (2022) *New Military Retention Prediction Model: Machine Learning for High-Fidelity Forecasting* (Institute for Defense Analyses, Alexandria, VA).

Perry S, Griffith J, White T (1991) Retention of junior enlisted soldiers in the all-volunteer Army Reserve. *Armed forces and society* 18(1):111–133.

Pratt M, Boudhane M, Cakula S (2021) Employee attrition estimation using random forest algorithm. *Baltic Journal of Modern Computing* 9:49–66.

Roppoli SA (2012) Military benefits that retain mid-career Army officers. M.A. thesis, Army Command And General Staff College, Fort Leavenworth, KS, USA, https://apps.dtic.mil/sti/citations/ADA569854.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) Rocr: Visualizing classifier performance in r. *Bioinformatics* 21(20):7881, http://rocr.bioinf.mpi-sb.mpg.de.

Spain E, Mukunda G, Bates A (2021) *The Battalion Commander Effect*, volume 51 (U.S. Army War College, Carlisle Barracks).

Speten KJ (2018) Predicting U.S. Army first-term attrition after initial entry training. M.A. thesis, Naval Postgraduate School, Monterey, CA, USA.

Therneau T, Atkinson B, Ripley B (2013) Rpart: Recursive partitioning. r package version 4.1-3. Http://CRAN.R-project.org/package=rpart.

Thomas I (2022) The U.S. Army is struggling to find the recruits its needs to win the fight over the future. CNBC, Oct. 26, https://www.cnbc.com/2022/10/26/us-army-struggles-to-find-recruits-its-needs-to-win-fight-of-future.html.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S*, Fourth ed. (Springer, New York), https://www.stats.ox.ac.uk/pub/MASS4/, iSBN 0-387-95457-0.

Vie LL, Scheier LM, Lester PB, Ho TE, Labarthe DR, Seligman MEP (2015) The U.S. Army person-event data environment: A military-civilian big data enterprise. *Big Data* 67–79.

Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York), https://ggplot2.tidyverse.org.

Wilke CO (2020) *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*, https://CRAN.R-project.org/package=cowplot, r package version 1.1.1.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California