



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**PREDICTING U.S. ARMY FIRST-TERM ATTRITION
AFTER INITIAL ENTRY TRAINING**

by

Karey J. Speten

June 2018

Thesis Advisor:
Second Reader:

Andrew T. Anglemeyer
Jonathan K. Alt

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
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 2018	3. REPORT TYPE AND DATES COVERED Master's thesis		
4. TITLE AND SUBTITLE PREDICTING U.S. ARMY FIRST-TERM ATTRITION AFTER INITIAL ENTRY TRAINING		5. FUNDING NUMBERS RNLQ8		
6. AUTHOR(S) Karey J. Speten				
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) U.S. Army TRADOC, Monterey, CA 93943		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 United States Army recently announced a reduction of its 2018 recruiting goal due to a challenging recruiting environment and a shrinking population of eligible candidates. However, the Sergeant Major of the Army has stated that the current improvement in the retention of existing soldiers should mitigate the loss of new recruits. The goal of this research is to identify demographic and administrative factors of active-component, first-term enlisted soldiers who have completed their Initial Entry Training to construct predictive models capable of identifying soldiers with high chances of failure in completing their initial contractual obligation. We construct a binary logistic regression model and a random forest classification model to predict a soldier's probability of first-term attrition based on the individual's unique service record. We find that a soldier's deployment history and the duration of the initial contract are significant predictors of whether a soldier will complete his or her first term. Knowledge of the key factors and other influencing variables assists the Army Resiliency Directorate in creation of models and tools to better advise Army leadership and develop intervention strategies and preventative measures to prevent the loss of first-term soldiers.				
14. SUBJECT TERMS Army, attrition, logistic regression, random forest, enlisted, retention, first-term, classification, predict, tree			15. NUMBER OF PAGES 107	
			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	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**PREDICTING U.S. ARMY FIRST-TERM ATTRITION AFTER INITIAL ENTRY
TRAINING**

Karey J. Speten
Major, United States Army Reserve
BA, Lawrence University, 2010
MPA, Bellevue University, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
June 2018**

Approved by: Andrew T. Anglemeyer
Advisor

Jonathan K. Alt
Second Reader

Patricia A. Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The United States Army recently announced a reduction of its 2018 recruiting goal due to a challenging recruiting environment and a shrinking population of eligible candidates. However, the Sergeant Major of the Army has stated that the current improvement in the retention of existing soldiers should mitigate the loss of new recruits. The goal of this research is to identify demographic and administrative factors of active-component, first-term enlisted soldiers who have completed their Initial Entry Training to construct predictive models capable of identifying soldiers with high chances of failure in completing their initial contractual obligation. We construct a binary logistic regression model and a random forest classification model to predict a soldier's probability of first-term attrition based on the individual's unique service record. We find that a soldier's deployment history and the duration of the initial contract are significant predictors of whether a soldier will complete his or her first term. Knowledge of the key factors and other influencing variables assists the Army Resiliency Directorate in creation of models and tools to better advise Army leadership and develop intervention strategies and preventative measures to prevent the loss of first-term soldiers.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
A.	OVERVIEW.....	1
B.	PURPOSE OF THE RESEARCH.....	2
C.	PREVIOUS ATTRITION RESEARCH.....	2
1.	Government Accountability Office	3
2.	RAND Corporation.....	5
D.	SCOPE AND ORGANIZATION	6
II.	DATA AND METHODOLOGY	9
A.	PERSON-EVENT DATA ENVIRONMENT	9
B.	RESEARCH DATA DESCRIPTION	10
1.	Datasets Used.....	10
2.	Variables Used.....	11
C.	LIMITATIONS AND ASSUMPTIONS	19
D.	METHODOLOGY	21
1.	Building the Response Variable.....	21
2.	Merging the Predictor Variables.....	24
3.	Data Stratification.....	24
4.	Modeling Techniques.....	24
III.	DESCRIPTIVE STATISTICS.....	27
A.	COHORT DATASET OVERVIEW	27
B.	NUMERIC FEATURES SUMMARY	28
C.	BINARY FEATURES SUMMARY	30
D.	CATEGORICAL FEATURES SUMMARY.....	33
E.	UNIVARIATE ANALYSIS.....	36
1.	Numeric Features.....	37
2.	Binary and Categorical Features.....	38
IV.	MULTI-VARIATE MODELING AND ANALYSIS.....	41
A.	BINARY LOGISTIC REGRESSION.....	41
1.	Purposeful Variable Selection.....	41
2.	Model Diagnostics	45
3.	Regression Analysis	47
4.	Prediction.....	53
B.	CLASSIFICATION TREE	57
C.	RANDOM FOREST	59

D.	MODEL SELECTION	63
V.	SUMMARY	67
A.	CONCLUSIONS	67
1.	Data Preparation.....	67
2.	Data Analysis.....	67
B.	RECOMMENDATIONS.....	68
1.	Implementation	68
2.	Future Research	69
	APPENDIX A. SEPARATION CODES	71
	APPENDIX B. COHORT DATASET SUMMARY	73
	APPENDIX C. UNIVARIATE MODEL RESULTS.....	80
	LIST OF REFERENCES.....	84
	INITIAL DISTRIBUTION LIST	87

LIST OF FIGURES

Figure 1.	Unemployment Rates versus Military Enlistments. Source: Syeed and Whiteaker (2018).	1
Figure 2.	Classification Summary of the Response Variable.....	23
Figure 3.	Average Enlistment Age by Attrition Category.....	28
Figure 4.	Average ASVAB GT Score by Attrition Category.....	29
Figure 5.	Average Number of Days Deployed by Attrition Category	30
Figure 6.	Proportion of Gender Levels by Attrition Category	31
Figure 7.	Proportion of High School Certification by Attrition Category	32
Figure 8.	Proportion of Prior Service by Attrition Category.....	33
Figure 9.	Proportion of AFQT Category by Attrition Category.....	34
Figure 10.	Attrition Rate by Home of Record.....	35
Figure 11.	Proportion of Marital Status by Attrition Category	36
Figure 12.	Binned Predicted Probabilities and Observed Proportions. Adapted from Faraway (2016).	46
Figure 13.	Attrition Probability—Days Deployed by Gender with Education Tier Panels	50
Figure 14.	Attrition Probability—Days Deployed by Gender with AFQT Category Panels	51
Figure 15.	Attrition Probability—Days Deployed by Gender with Military Occupation Group Panels	52
Figure 16.	Attrition Probability—Days Deployed by Gender with Marital Status Panels	53
Figure 17.	Attrition Classification by Logistic Regression—ROC Curve (Training Dataset)	55
Figure 18.	Attrition Classification by Logistic Regression—ROC Curve (Test Dataset)	56

Figure 19.	Attrition Classification Tree	58
Figure 20.	Attrition Classification by Classification Tree—ROC Curve (Test Dataset)	59
Figure 21.	Random Forest Error by Tree Quantity	61
Figure 22.	Random Forest Variable Importance	62
Figure 23.	Attrition Classification by Random Forest—ROC Curve (Test Dataset)	63
Figure 24.	Model Comparison ROC Curves	64

LIST OF TABLES

Table 1.	Feature Summary and Data Source Mapping	12
Table 2.	Numeric Feature Definition	14
Table 3.	Binary Feature Definition	15
Table 4.	Categorical Feature Definition.....	16
Table 5.	Military Occupation Map.....	18
Table 6.	Full Cohort Dataset—Attrition Rate by Fiscal Year of Enlistment.....	27
Table 7.	Training and Test Datasets—Counts by Fiscal Year of Enlistment	37
Table 8.	Univariate Summary of Numeric Features: Mean (Std Dev)	38
Table 9.	Max Time in Grade Importance Comparison	43
Table 10.	Logistic Regression Variable Importance.....	44
Table 11.	Regression Model Variable Inflation Factors	46
Table 12.	Regression Model Coefficients Matrix	47
Table 13.	Logistic Regression Training Dataset Confusion Matrix	54
Table 14.	Logistic Regression Test Dataset Confusion Matrix	56
Table 15.	Significant Variable Utilization by Model.....	64
Table 16.	Successful Separation Code Definitions.....	71
Table 17.	Numeric Feature Summary: Mean (Std Dev) by Fiscal Year of Enlistment	73
Table 18.	Binary Feature Summary: Counts (Proportion of Attrition Category) by Fiscal Year of Enlistment.....	74
Table 19.	Categorical Feature Summary: Counts (Proportion of Attrition Category) by Fiscal Year of Enlistment.....	75
Table 20.	Univariate Summary of Binary/Categorical Features: Count and Proportion by Attrition Category	80

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AAG	Army Analytics Group
AFQT	Armed Forces Qualification Test
ARD	Army Resiliency Directorate
ASVAB	Armed Services Vocational Aptitude Battery
AWD	Army Waiver Database
AUC	Area Under the Curve
BASD	Basic Active Service Date
CMF	Career Management Field
CTS-OCO	Contingency Tracking System – Overseas Contingency Operations
DA	Department of the Army
DCIPS	Defense Casualty Information Processing System
DMDC	Defense Manpower Data Center
DOD	Department of Defense
GAO	Government Accountability Office
GED	General Education Diploma
HOR	Home of Record
HRC	Human Resources Command
IET	Initial Entry Training
MEPCOM	Military Entrance and Processing Command
MOS	Military Occupation Specialty
MTOE	Modified Table of Organization and Equipment
PDE	Person-Event Data Environment
PID	Person Identifier
TAPDB	Total Army Personnel Database
TSC	Test Score Category
TDA	Table of Distribution and Allowances
TRAC	TRADOC Analysis Center
TRADOC	Training and Doctrine Command
RF	Random Forest
ROC	Receiver Operating Characteristic

USAREC
VIF

United States Army Recruiting Command
Variable Inflation Factor

EXECUTIVE SUMMARY

The United States Army recently reduced its 2018 recruiting goal. A challenging recruiting environment and a shrinking population of eligible candidates requires increased efforts to retain and reenlist enlisted personnel to meet long-term manning level sustainability and growth. Historical research shows that the Army loses 28% of new recruits during their initial contractual period after expending significant effort and resources throughout the enlistment and initial entry training (IET) process. The goal of our research is to identify demographic and administrative factors of active component, first-term enlisted soldiers who have completed IET to construct predictive models capable of identifying soldiers with high chances of failure in completing their initial contractual obligation. We recommend a model to be used by the Army Resiliency Directorate (ARD) in their development of a prediction tool. The tool could be leveraged to inform policy making and intervention strategy development intended to support soldiers throughout the first term and reduce attrition rates.

Using the Person-Event Data Environment, a cloud-based virtual data repository, we had access to numerous decentralized datasets containing sensitive Army personnel data and the research tools necessary for statistical analysis. Our research examined the personnel records of active-component enlistees from FY2005 to FY2010. The data consisted of quarterly snapshots from six sources including Human Resources Command (HRC), Military Entrance Processing Command (MEPCOM), and U.S. Army deployment records. After wrangling the data and selecting the relevant time-based soldier records, our final cohort included 418,000 observations with 11 numeric predictor variables and 21 categorical variables encompassing 94 factor levels.

We describe a unique method of identifying whether a soldier completed his or her first term by examining career status codes and identification of separation codes that signify a success. The remaining personnel were categorized by the absence of personnel records after their calculated obligation end date. We calculated attrition rates for all variables and found that women and soldiers who received a General Education Diploma (GED) versus a high school diploma are more likely to fail to complete their first terms

than their peers. Additional differences are noted between military occupation specialties (MOS) and Armed Services Vocational Aptitude Battery (ASVAB) test scores.

Three predictive models were constructed for a comparison of predictive accuracy: simple logistic regression, classification tree, and a random forest. The models each predicted soldier attrition with accuracy rates greater than 80%. The number of days a soldier has deployed, soldier marital status, and the length of the initial contract are the most influential factors in predicting the success of a soldier in completing their first term. A soldier deployed more than three months has a 90% chance of completing his or her first term. Longer contract periods increase the risk of attrition while married or divorced soldiers are less likely than single soldiers to be discharged prior to completing their first term. Though attrition rates differ, a soldier's gender and ASVAB score have very little influence in predicting attrition. Additionally, we saw no evidence that enlistment waivers affect the probability of a soldier's attrition.

Our attrition rate findings based on demographic and administrative soldier data may help to inform U.S. Army force strength requirements, recruiting goals, and retention efforts. Knowledge of these key factors, and other influencing variables, can assist the ARD in further development of models and tools to better advise U.S. Army executives and unit leadership. Individual soldier probability assessments may also offer quantitative justification for funding prioritizations and lines of effort critical to sustaining and strengthening the U.S. Army's most valuable resource: Soldiers.

ACKNOWLEDGMENTS

I first offer thanks to my family for their love, encouragement, and support throughout my Naval Postgraduate School adventure and thesis completion. Their ability to sleep through my late nights of feverishly loud typing outside their bedroom doors is truly amazing!

This thesis would not have been possible without the expertise and guidance of Dr. Andrew Anglemeyer and Dr. Jon Alt. Their commitment to this research and unwavering support of my efforts were remarkable and appreciated. A sincere thank you is also extended to MAJ Tony Smith and MAJ Jarrod Shingleton with TRAC-Monterey. Their tireless contributions were critical in acquiring the data and pushing the project forward at a record pace given the challenges.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. OVERVIEW

As reported in a recent Associated Press article, Sergeant Major of the Army Daniel Dailey describes a positive development in first-term attrition rates: “Retaining current soldiers has been more successful this year than in the past, with 86% staying on, compared with 81% in previous years” (Baldor, 2018, p. 1). On the other hand, the Army has announced its projected failure to meet its recruiting goal of 80,000 active soldiers in FY2018. The U.S. Army faces a significant challenge in recruiting and retention for the foreseeable future due to the low unemployment rate (Figure 1). According to Syeed and Whiteaker (2018), “high obesity rates, drug use, criminal records and failing grades on the Army’s aptitude test” complicate the challenge even further by reducing the eligible recruiting pool to “29% of the available population of 17-to 24-year-olds” (p. 1).

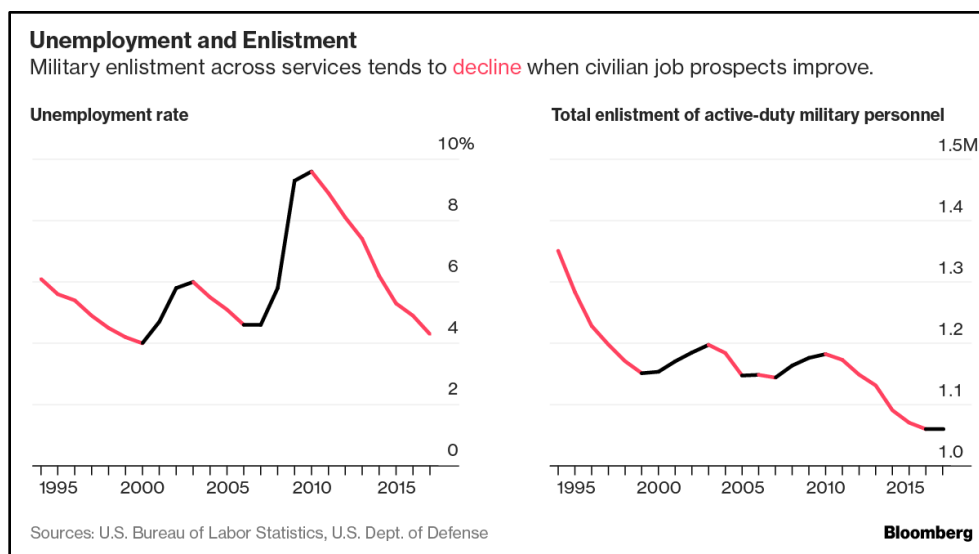


Figure 1. Unemployment Rates versus Military Enlistments.
Source: Syeed and Whiteaker (2018).

As highlighted by Sergeant Major Dailey’s comment, reenlistments and retention of existing soldiers is key to long-term manning level sustainability and growth.

B. PURPOSE OF THE RESEARCH

The goal of our research was to identify and explore demographic and administrative factors of active component, first-term enlisted soldiers that can be used to significantly predict soldiers' probability of failing to complete their initial contractual obligation. We used the predictor variables to construct a logistic regression model and a random forest classification model to predict a soldier's probability of first-term attrition based on the individual's unique service record. Comparisons of the models' predictive accuracy are discussed, and recommendations are provided to inform the creation of a web-based prediction tool utilizing our model and improve future U.S. Army attrition research.

The Army Resiliency Directorate (ARD) is the Army agency tasked with synchronizing and driving the cultural shift for the Ready and Resilient Campaign. The U.S. Training and Doctrine Command (TRADOC) Analysis Center (TRAC) is leading the research efforts in support of ARD and sponsored our research. ARD hopes to gain current and relevant insights into first-term attrition to assist in policy development and stakeholder engagement. Moreover, the agency seeks a sustainable predictive modeling tool to assist Army leaders at all levels to better understand the individual soldier needs within their formations and aid soldiers that could benefit from preventative measures.

C. PREVIOUS ATTRITION RESEARCH

Significant research examining Army attrition has been conducted by two organizations: RAND Corporation and the Government Accountability Office (GAO). Nearly four decades of their research from as early as 1984 consistently demonstrate that the most significant factors linked to attrition are a soldier's gender and the means in which they attained their high school certification. Specifically, females and General Education Diploma (GED) recipients have a higher probability of failing to fulfill their initial service obligations than males and high school graduates. The general methods of analysis and data sources are similar across all the studies; however, the specific data sources and data manipulation methodologies are not well documented. A brief overview of their research will assist in better understanding U.S. Army attrition.

1. Government Accountability Office

The GAO has conducted several studies on military attrition across all branches of service. The first report to examine military attrition rates after the enlistee population had stabilized for a decade following the elimination of the draft was published in 1997 and studied data collected by the Defense Manpower Data Center (DMDC) on service members from 1986 to 1994 (Government Accountability Office [GAO], 1997). The study concluded that first-term attrition across all military services consistently averaged approximately 30% but cautioned the services against drawing any definitive conclusions from the results because of the lack of consistency in the available data and arbitrary attrition goals created by each service. In his testimony before the Subcommittee on Military Personnel of the 105th Congress in 1997, Mark Gebicke of the GAO commented specifically on the inadequate quality of data available to researchers:

DOD's [Department of Defense] [sic] current data on attrition is inconsistent and incomplete for two reasons. First, the services interpret DOD's definitions of separation codes differently and therefore place enlistees with identical situations in different discharge categories. ...Second, DOD's separation codes—which represent DOD's primary source of service-wide data on why people are leaving the services—capture only the official reason for discharge. ...In an attempt to standardize the services' use of these codes, DOD issued a list of the codes with their definitions. However, it has not issued implementing guidance for interpreting these definitions (Military Attrition, 1997, p. 6)

Gebicke testified again before the committee in 1998 and stressed the need for better analysis of separations by the DoD to improve military recruiting efforts. Additionally, he referenced specific findings from other GAO studies that “consistently [showed] that persons with high school diplomas [vice GED holders] and Armed Forces Qualification Test [AFQT] scores in the upper 50th percentile have lower first-term attrition rates” (Military Attrition, 1998, p. 3).

One of the studies referenced by Gebicke provides numerous descriptive summaries of the attrition data allowing for a baseline to check against the constructed dataset for our current research (Government Accountability Office [GAO], 1998):

- First-term attrition rate from 1986 to 1998 was 31%

- Eighty-seven percent of enlistees signed 2-year, 3-year, or 4-year contracts
- In FY1993, female attrition rates were 51% compared to 37% for males

Many of the descriptive summaries are used by the authors to specifically highlight the inconsistency of the separation codes applied across the services as mentioned in GAO's previous studies. As an example, the separation codes were grouped into broad categories: misconduct, medical conditions, unsatisfactory performance, drug use, and pregnancy. The large volume of approximately 1.7 million annual enlistees encouraged the researchers to assume statistical similarity across the services when comparing the attrition rates in each of the categories. However, significant differences in attrition rates within categories were seen between the services across all fiscal years (GAO, 1998) as previously suspected. One last key piece of information was included in this study—it was the only document we reviewed for our current research that clearly defined “attrition.” The authors state that DoD has defined attrition “as the failure of an enlistee to complete his or her contractual obligation” (GAO, 1998, p. 16).

Gebicke returned to the congressional stage in 1999 with some positive news. The recommendations from the 1997 GAO report had been adopted into the National Defense Authorization Act for Fiscal Year 1998 (Military Attrition, 1999). Moreover, the DoD had standardized the separation codes and created definitions through the efforts of a joint planning group. While the DoD had still not provided specific guidance to ensure a universal application of the codes across the services, Gebicke was optimistic that the data challenges surrounding attrition analysis would be fixed within 18 months (Military Attrition, 1999, p. 2). Unfortunately, Gebicke's successor, Norman Rabkin, presented testimony to Congress in 2000 concluding that “we [GAO] see no evidence that the problems we previously identified with these [separation] codes have been solved” (Military Personnel, 2000, p. 9).

The most recently published report on overall DoD attrition rates is found within a 2017 GAO study focused on the attrition of first-term enlistees due to medical separations. The study sought a comparison of first-term medical separations with overall first-term

separation totals and reports that the overall attrition rate was steady from FY2005 to FY2015 across all services at an average of approximately 28% (Farrell, 2017, p. 14).

2. RAND Corporation

Three RAND Corporation documents encompass nearly three decades of military enlistee research. While the earliest work examines attrition across all military services, the other two studies provide specific insight into first-term attrition solely within the Army.

In a summary research brief published in 1985, Buddin utilized new survey data from the 1979 Survey of Personnel Entering Military Service to enrich the typical service personnel record with additional variables (e.g., employment history, job match, job satisfaction, entry point decisions). Initially, his findings mirrored results from other studies such as the link between attrition and the lack of a high school diploma and lower AFQT scores. However, the new survey data illuminated unique findings as well. The probability of attrition increased by 1% as the age of an enlistee at enlistment increased from 17. The results also indicated that unemployment prior to enlistment increased the attrition rate. Significantly, there was no relationship found between job match or satisfaction and attrition.

An extensive description of Army attrition research is presented in a doctoral dissertation by Martin (1995) of the RAND Graduate School. The goal of the research was to construct a predictive logistic regression model with Army personnel data provided by the DMDC. Martin's approach was to factor all numerical predictors (e.g., age, aptitude test scores) and collapse the factors into broad categories as to "maximize the difference in attrition between the dummy variable classes" (p. 35). Although his model ultimately had very poor predictive power, he was satisfied that it was at least no worse than other modeling attempts. Martin's work also provides interesting thoughts regarding the selection of variables in most attrition research. He points out that many variables linked to attrition, such as gender, are unable to be affected by any change in policy. However, there is a strong reason to believe that the variables are proxies for other correlated

variables that may not be represented elsewhere in the data and should be considered as valid predictors even though the “true” variable is still unknown (Martin, 1995, p. 19).

In 2005, Buddin focused his attention specifically on Army first-term attrition. The study data are very similar to the dataset used in our current research. He used DMDC personnel data from FY1995 to FY2001 consisting of 550,000 observations (i.e., individual soldier enlistments). Unlike our study, he incorporated extensive recruiting station and individual recruiter data provided by the U.S. Army Recruiting Command. He found that none of the recruiting information had any impact on attrition rates. Interestingly, he also did not report any evidence of his unique finding regarding age at enlistment from his previous work. While confirming overall attrition rates of 34% similar to other research, he found a higher attrition rate (51%) among women compared to men (31%) and a higher rate (50%) among GED holders versus high school graduates (32%) (p. 74). Unlike previous research, he also detected higher rates of attrition among African-American and white non-Hispanic enlistees compared to Asian and Hispanic recruits. Unfortunately, Buddin does not provide a detailed explanation of his definition of “attrit” or provide a roadmap of his classification methodology of a soldier’s attrition status. Despite these limitations, his research is an excellent comparison document as we prepare our dataset and attempt to validate our classification technique.

D. SCOPE AND ORGANIZATION

The scope of our research includes the creation of a dataset containing personnel record data of Army soldiers who enlisted in the Active Component in any of six fiscal years from FY2005 to FY2010. The analysis of the data provides insights into the significant factors that can be used to predict the probability of a soldier to leave the Army before completing his or her initial contractual obligation term. A detailed description of the methodology used to categorize a soldier as “attrit” or “non-attrit” is provided to assist in future research. Both logistic regression and random forest models are constructed, and their validity is examined. The best model is then used after stratifying the data by fiscal year to explore the ability of the model to use the previous year’s data to forecast the next year accurately.

The thesis consists of five chapters. Chapter II provides information regarding the data source, summarizes the cohorts, and describes the variables. Additionally, we address the research limitations and assumptions before presenting the methodology. In Chapter III, we present descriptive statistical summaries and key findings. Chapter IV illustrates the modeling techniques and we provide our findings regarding the predictive power of the final model. Finally, Chapter V concludes our study results and recommends future research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA AND METHODOLOGY

A. PERSON-EVENT DATA ENVIRONMENT

Typically, researchers spend a great deal of time and effort simply assembling and collating data before any research can even begin. In fact, history shows that 60% of research time and funding is spent on these critical tasks (Jensen, 2016). The data collection process is often labor-intensive and may involve in-person survey execution, written agreements between researchers and data holders to allow data sharing, and extensive data wrangling to combine the final data received in a meaningful way. For our research, the Person-Event Data Environment (PDE) remedied the pain and offered relatively easy access to sensitive Army personnel data.

The Army Analytics Group (AAG), located in Fairfield, CA, created the Person-Event Data Environment in coordination with the DMDC to provide “a consolidated repository for manpower, service, personnel, financial, health, and medical data that cost-effectively brings the researcher to the data, rather than ... bringing data to the researcher” (Jensen, 2016, p. 2). The PDE is a cloud-based environment that not only offers researchers access to numerous decentralized datasets, but also the software tools required to interact with the data and perform analysis. Additionally, analyst data permissions are limited solely to the approved projects to which they have been formally assigned.

One feature particularly useful to researchers is the system generated, persistent person identifier (PID) field contained within all of the datasets that allows researchers to easily merge tables from different data sources on the individual soldier represented. The PID represents each “person” contained in the datasets while anonymizing personal health information and personally identifiable information in accordance with DoD and federal regulations. There is also a rigorous review process prior to the release of any research output outside of the secure PDE virtual private network.

The most crucial element pertaining to our study is the ability of the PDE to provide access to the project’s data and tools to analysts at ARD. This allows us to perform the research independently and share the source code and specific data methodologies with

ARD analysts at the conclusion of our research. Additionally, the analysts are assured matching data access to tables and data fields required to implement subsequent interactive tools based on our model results.

B. RESEARCH DATA DESCRIPTION

1. Datasets Used

Our research employed six datasets within the PDE. The primary dataset used to construct the cohort of enlisted soldiers for analysis is the Active Duty Military Personnel Master. The data consists of the demographic information contained in a soldier's service record and detailed information from a soldier's personnel file maintained by the Army Human Resources Command in the Total Army Personnel Database (TAPDB). Next, we merged the Active Duty Military Personnel Transaction dataset with our newly created Cohort Dataset. The transaction table captures the changes in a soldier's record such as enlistment into the Army, separation from the Army, and reenlistments. The table provided the documented separation codes used in our analysis for determining if a soldier left the Army before the completion of the initial contractual obligation.

Next, the MEPCOM-700 file was added. The data is derived from the system of record utilized by the Military Entrance and Processing Command (MEPCOM) during the recruitment process of a civilian applicant in all branches of service. The data contains additional demographic data not found in the Army's Master file such as scores from the Armed Services Vocational Aptitude Battery (ASVAB) testing and the number of dependents at the time of enlistment. Related to the MEPCOM file, we also included the Army Waiver Database (AWD) maintained by the U.S. Army Recruiting Command (USAREC) containing information about administrative and medical waiver events granting a soldier admission into the Army.

The final two datasets consisted of events that occurred during a soldier's initial contractual period. The injury table reports data contained in the Defense Casualty Information Processing System: Injury file (DCIPS) and lists any injuries that occurred to a soldier while in a deployed status: both hostile and non-hostile types. Finally, we joined on the Contingency Tracking System – Overseas Contingency Operations (CTS-OCO)

dataset for a count of the number of deployments and number of days deployed for each soldier. Unlike the previously described datasets which provided mostly standardized codes, the variables created from the injury and deployment tables represent logic-based calculations in determining the counts of events and ensuring only events that occurred prior to either the completion of the contractual period or early discharge were credited.

2. Variables Used

After joining the tables together into a dataset containing one observation for each of the enlisted soldiers that joined the Army from FY2005 to FY2010, our Cohort Dataset contained 418,000 observations (soldiers) with 11 numerical predictor variables, seven binary variables, and 14 categorical variables encompassing 80 factor levels. Despite the volume of previous attrition research that provided consistent recommendations for variable selection, we decided to initially explore all of the variables available to us within the PDE deemed as even “possible” predictors of attrition. During the analysis and modeling stages of the research, some of the original numeric variables were converted to categorical variables and many of the factor levels within the original categorical variables were collapsed; however, a brief description of the variables selected for examination provides a better understanding of both the original source data and the challenges faced during data wrangling.

Many of the variables used in our research required a determination of which record from the original data tables would be selected for use in the cohort dataset based on the time-related nature of the data. The *marital status (max)* variable provides an illustrative example. Is a new recruit’s marital status a predictor of whether he or she will successfully complete their first-term, or is marital status later in the first-term a better predictor? It seems logical to account for the fact that a soldier may get married during his or her first-term and that having a family may have more impact on whether they complete their first-term successfully than the fact that they decided to join the military when they were single. Selecting the later record may increase the predictive power of a model; however, information only available later in a soldier’s first term may not be useful for MEPCOM personnel seeking to better understand the recruit population. Many of the variables present

in our dataset required this specific selection because the data was stored as quarterly snapshots of soldier records. As our research focused on the development of a predictive tool intended for the Army resiliency initiative of ARD, we were not limited in selecting only the data points available prior to a soldier’s enlistment and completion of IET.

A data mapping is provided to summarize this data selection process (Table 1). The mapping provides a list of the variables (features) used in the analysis to the authoritative data table used from within the PDE. The features that may have changed over time include an identifier in the name of “enlistment” or “max” to indicate if the record used was from the first time we saw a soldier in the data tables or if we selected a later record. The variables requiring an identification of “max” records were first sorted to identify the last data point for each soldier that was recorded prior to the end of their initial contractual obligation period. The data is categorized by data type: continuous numeric values, binary “yes” or “no,” or categorical variable consisting of categories such as gender “male” or “female.” Additionally, descriptions are provided identifying the features that were engineered by an algorithm applied to the original data (e.g., enlistment age calculated by examining a soldier’s birthdate and comparing against the date of enlistment) and whether factor levels were collapsed within the categorical features to provide more statistical power to the modeling process. Categorical variables also include the number of factor levels present for modeling and analysis.

Table 1. Feature Summary and Data Source Mapping

Feature	Data Source	Data Type	Engineered	Collapsed	Factor Levels
AFQT Category	Master	Categorical	N	Y	5
Age at Enlistment (Years)	Master	Numeric	Y	N/A	N/A
ASVAB GT Score (Scale 3-150)	MEPCOM	Numeric	Y	N/A	N/A
Citizenship Origination	Master	Categorical	N	Y	3
Citizenship Status (Enlistment)	Master	Binary	N	N	2
Contract Duration (Years)	Master	Numeric	N	N/A	N/A

Feature	Data Source	Data Type	Engineered	Collapsed	Factor Levels
Days Deployed (Qty)	CTS-OCO	Numeric	Y	N/A	N/A
Dependents (Max Qty)	Master	Numeric	N	N/A	N/A
Deployments (Qty)	CTS-OCO	Numeric	Y	N/A	N/A
Education Level (Enlistment)	Master	Categorical	N	Y	4
Education Level (Max)	Master	Categorical	N	Y	4
Education Tier	Master	Categorical	N	N	3
Fiscal Year (Enlistment)	Master	Categorical	Y	N	6
Gender	Master	Binary	N	N	2
Height at Enlistment (Inches)	MEPCOM	Numeric	N	N/A	N/A
Home of Record Region	Master	Categorical	N	N	5
Hostile Injuries (Qty)	DCIPS	Numeric	Y	N/A	N/A
Marital Status (Max)	Master	Categorical	N	Y	3
Max Time-in-Grade (Months)	Master	Numeric	Y	N/A	N/A
Military Occupation (Max)	Master	Categorical	N	Y	21
Military Occupation Group	Master	Categorical	N	N	3
Non-Hostile Injuries (Qty)	DCIPS	Numeric	Y	N/A	N/A
Prior Service	Transaction	Binary	Y	N	2
Rank (Enlistment)	Master	Categorical	N	N	6
Rank (Max)	Master	Categorical	N	N	6
Unit Region (Max)	Master	Categorical	Y	N	5
Unit Type (Max)	Master	Categorical	N	Y	5
Waiver (Admin)	AWD	Binary	N	N	2
Waiver (Conduct)	AWD	Binary	N	N	2
Waiver (Drug)	AWD	Binary	N	N	2
Waiver (Medical)	AWD	Binary	N	N	2
Weight at Enlistment (Pounds)	MEPCOM	Numeric	N	N/A	N/A

a. *Numeric Feature Definition*

Unlike the previous research conducted by Martin (1995), we chose to retain the continuous nature of our numeric variables instead of converting them all to categorical variables. The variables were left to “stand for themselves” and smooth the underlying structure of the predictive model. Additionally, the numeric property recognizes that a variable such as age would have a continuous effect on the probability that a soldier fails to complete their initial contractual period (e.g., for age, the probability to attrit does not change only at the time of a birthday, but continuously throughout the year). The cohort dataset consisted of 12 numeric features for inclusion in modeling and analysis (Table 2).

Table 2. Numeric Feature Definition

Feature	Description
Age at Enlistment (Years)	Enlistee age in years at time of enlistment
ASVAB GT Score (Scale 3-150)	Sum of lines scores: word knowledge, paragraph comprehension, and arithmetic reasoning
Contract Duration (Years)	Initial contractual obligation duration
Days Deployed (Qty)	Sum of days deployed during the first-term: consecutive deployed days > 30 required to be considered a deployment
Dependents (Max Qty)	Number of dependents at end of first-term
Deployments (Qty)	Number of deployments completed during first-term: consecutive deployed days > 30 required for credit
Height at Enlistment (Inches)	Enlistee height at time of enlistment
Hostile Injuries (Qty)	Number of hostile injuries received in a deployed status
Max Time-in-Grade (Months)	Maximum number of months a soldier spent in any one paygrade during the first term
Non-Hostile Injuries (Qty)	Number of non-hostile injuries received in a deployed status
Weight at Enlistment (Pounds)	Enlistee weight at time of enlistment

b. Binary Feature Definition

A binary variable is a special case of categorical variables consisting of only two factor levels. Binary features within our cohort dataset represent two methods of variable assignment: categorical or computed. As *citizenship status* and *gender* only had two original categories, one of the categories was coded as a “0” and the other a “1.” We constructed the waiver variables based on the lack of a soldier’s presence in the Army Waiver Database as the source data only reported an enlistee who received a waiver. The classification of a soldier as having prior service in a branch other than the Army resulted from an examination of the separation codes present in the Transactions table. Four Active Duty Personnel Transaction Type Codes representing prior service were identified: 115-Gain Prior Service Reserve, 117-Gain Prior Service Retired, 120-Gain Prior Service Military Control, and 123-Gain Prior Service Other. A summary of the final binary designations is provided in Table 3.

Table 3. Binary Feature Definition

Feature	Description
Citizenship Status (Enlistment)	0: U.S. Citizen 1: Non- U.S. Citizen
Gender	0: Female 1: Male
Prior Service	0: Non-Prior Service 1: Prior Service – soldiers with gain codes 115, 117, 120, and 123 considered gains from prior service
Waiver (Admin)	0: No Waiver 1: Waiver Received – if soldier was in the data table, then waiver received, else no waiver
Waiver (Conduct)	0: No Waiver 1: Waiver Received – if soldier was in the data table, then waiver received, else no waiver
Waiver (Drug)	0: No Waiver 1: Waiver Received – if soldier was in the data table, then waiver received, else no waiver
Waiver (Medical)	0: No Waiver 1: Waiver Received – if soldier was in the data table, then waiver received, else no waiver

c. Categorical Feature Definition

Categorical features consist of variables with more than two categories (levels). Many of the features required collapsing of category levels to ensure that minimum count

requirements were met for logistic modeling and contingency table analysis. In general, categorical levels with a total less than 1,000 were collapsed due to the overall size of our population. The resulting features and levels are summarized in Table 4 followed by a description of the variables requiring clarification.

Table 4. Categorical Feature Definition

Feature	Description	Levels	Description
AFQT Category	Army test score category (TSC) representing compiled scores from the ASVAB test of quantile achievement	TSC-I	AFQT 93-99
		TSC-II	AFQT 65-92%
		TSC-IIIA	AFQT 50-64%
		TSC-IIIB	AFQT 31-49%
		TSC-IVA	AFQT 16-30%
		TSC-IVB+	AFQT 0-15%
Citizenship Origination	Identifies non-citizens, U.S. nationals, and naturalized citizens	A	Born in United States
		C	Born outside United States
		N	Citizen by naturalization
Education Level (Enlistment)	Education level of recruit at enlistment	HS	High School or GED
		CLG	Some college
		BAC	Baccalaureate Degree
		GRAD	Graduate Degree
Education Level (Max)	Education level of soldier at end of first-term	HS	High School or GED
		CLG	Some college
		BAC	Baccalaureate Degree
		GRAD	Graduate Degree
Education Tier	DMDC-derived variable representing source of secondary school credit	1	High School diploma
		2	GED or equivalent
		3	No secondary school
Fiscal Year (Enlistment)	Fiscal year in which an enlistee reported to basic training as derived from the Basic Active Service Date	2005	Fiscal Year (1 Oct – 30 Sep)
		2006	
		2007	
		2008	
		2009	
		2010	
Home of Record Region	Soldier Home of Record (HOR) from one of five regions of the United States, as defined by the U.S. Postal Service	Midwest	Derived from HOR state
		Northeast	
		South	
		West	
		Territory	

Feature	Description	Levels	Description
Marital Status (Max)	Marital status of soldier at end of first-term	D	Divorced or separated
		M	Married
		N	Never Married
		Other	Widow or unknown
Military Occupation (Max)	Soldier career management field (CMF) at end of first-term – the first two numbers found in a soldier’s primary service occupation code	Multiple	See Table 5
Military Occupation Group (Max)	Functional categories of CMF codes grouping similar roles and functions	OPNS	Operations
		OS	Operations Support
		FS	Force Sustainment
Rank (Enlistment)	Soldier rank at enlistment	PV1	Private
		PV2	Private
		PFC	Private First Class
		CPL	Corporal or Specialist
		SGT	Sergeant
		SSG	Staff Sergeant
Rank (Max)	Soldier rank at end of first-term	PV1	Private
		PV2	Private
		PFC	Private First Class
		CPL	Corporal or Specialist
		SGT	Sergeant
		SSG	Staff Sergeant
Unit Region (Max)	Soldier unit location at end of first-term grouped into one of five regions of the United States	Midwest	Derived from assigned unit state location
		Northeast	
		South	
		West	
		Territory	
Unit Type (Max)	The type of unit authorization document governing unit manning and organization at end of first-term	MTOE	Modified Table of Organization and Equipment
		TDA	Table of Distribution and Allowances
		MULTI	Multi-Component

The *AFQT Category* represents the classification of an enlistee based on the percentile scores on the ASVAB. It is used by the U.S. Army Recruiting Command to determine if a recruit is eligible to join the Army and identify the military occupational specialties for which an enlistee can pursue. Test score categories are defined by

MEPCOM: however, the Army determines which categories will be accepted (Department of the Army [DA], 2016, pp. 11-12). The lowest score categories (TSC-IVB and TSC-V) were combined as neither category is allowable for entrance into the Army.

Education Level reports the level of education achieved. The codes from the Master data table were consolidated to capture the four key educational milestones: high school completion, some college, undergraduate degree, and graduate degree. Since education levels can change over time, two features were constructed to consider both education level at enlistment and whether final education level impacted first-term attrition.

The military occupation variables capture the job skill of the soldier. The Career Management Field (CMF) is the first two numbers in a soldier's primary occupational specialty code. The CMF groups similar, but unique, specialties into broad categories such as infantry or supply. The *military occupation* variable is the individual soldier skill. Factor levels with small counts were collapsed into an *LD* (low density) category. The *military occupation grp* variable grouped the CMFs into much broader categories as defined in the Army force structure regulation (Department of the Army [DA], 2014, p. 11). The military occupation definitions and groupings are shown in Table 5.

Table 5. Military Occupation Map

Military Occupation Group	Military Occupation (CMF)	Description
Operations	11	Infantry
	12	Engineer
	13	Field Artillery
	14	Air Defense Artillery
	15	Aviation
	18	Special Operations Forces
	19	Armor
	31	Military Police
Operations Support	74	Chemical
	25	Signal
	35	Military Intelligence

Military Occupation Group	Military Occupation (CMF)	Description
Force Sustainment	42	Human Resources
	68	Health Services
	88	Transportation
	91	Ordnance
	92	Quartermaster
LD (Low Density)	27	Judge Advocate General
	29	Electronic Warfare
	36	Finance
	37	Psychological Operations
	38	Civil Affairs
	46	Public Affairs
	51	Acquisitions
	56	Chaplain
	71	Health Services (Lab)
	79	Recruiting

The *unit type* variable represents the difference in the mission and function of the unit to which a soldier was assigned at the end of their first-term. Generally, a unit authorized by a Modified Table of Organization and Equipment (MTOE) document is part of the Operating Force responsible for deployed warfighting functions while a unit defined by a Table of Distribution and Allowances (TDA) document belongs to the Generating Force responsible for non-deployable administrative, training, and strategic functions (Department of the Army [DA], 2013). A multi-component unit is manned by a mix of Active Component, Army Reserve, and National Guard members.

C. LIMITATIONS AND ASSUMPTIONS

The most critical limitation in our research was attempting to define the response variable of “attrit” or “non-attrit”. As mentioned in every study reviewed, interpretation of the separation codes proved very challenging for researchers due to the lack of clear definitions and non-standardized manual data entry. Though the validity of the data was questioned, none of the previous work detailed the eventual use of the data or how it was used to determine whether an enlistee successfully completed their first-term. Without a definitive list of the 172 unique separation codes and their classification as either “attrit”

or “non-attrit”, exact duplication of the studies is impossible. Therefore, our construction of the response variable relied on a complex methodology and assumptions as detailed in the Methodology section of the thesis.

Another limitation is that the race code derived by DMDC in the Master data table does not match the data definitions available within the PDE. Additionally, the ethnicity code available consists of 75% missing values, which prove nearly useless for analysis. Therefore, our study is unable to assess the impact that race may have on attrition rate and cannot be compared to previous findings by Buddin (2005, p. 74).

A required assumption of our research is that the data maintained within the PDE is accurate and represents complete soldier information for the six fiscal years analyzed. Additionally, we assume that no significant changes occurred in a soldier’s record that were not accurately captured by the quarterly snapshot dates available.

Two critical assumptions were made in the creation of our response variable. First, less than 1% of our total population had contractual obligation durations with “odd” values of one, two, seven, or eight years. The standard enlistment contract is between three to six years. Since the odd values represented such a small percentage of the data, we removed these observations. Later in the creation of the predictor variables, 2.5% of the soldiers had the same odd values, which prevented selection of the soldier record with the correct end date of their first-term. To prevent bias in the analysis caused by removing these classified observations, the average contractual period contained in the full population was computed, and the odd contractual obligation values were assigned this value of 4 years.

Second, an initial review of the data revealed that over 10% of the soldiers classified as having failed to meet their contractual obligation had first-term end dates that fell within the last three months before the next snapshot date contained within the PDE. In other words, soldiers were assumed to have been discharged less than three months before they would have completed their first-term. We believed this result to be a function of only having quarterly snapshot data available. Therefore, we adjusted our classification algorithm and assumed that a soldier successfully completed their initial obligation if they were within three months of their first-term end date.

D. METHODOLOGY

To provide transparency of our research, a detailed description of the method used in building our cohort dataset is required. The classification of the soldiers into the “attrit” and “non-attrit” response variable required considerable research into the data available in the PDE. Once soldiers were classified, each predictor variable was layered onto each of the observations with consideration given to the moment in time with which we were concerned. Finally, the data was split into a training and test dataset to allow for future model validation.

1. Building the Response Variable

Instead of relying solely on the separation codes to classify soldiers as “attrit” or “non-attrit”, we sought a unique method made possible by the wealth of data available in the PDE. Instead of merely examining the separation codes and identifying codes that meant a soldier failed to complete the first-term, we worked the problem in reverse by classifying a soldier as having completed the first-term by inspecting data fields other than the separation codes. Once the soldier attrition classification was completed, we verified consistency of the data by comparing our results across fiscal years and with previous research findings.

We started with 429,908 unique soldier records representing all enlisted soldiers that arrived at basic training in FY2005 to FY2010. Of those soldiers, 11,704 were removed due to an Initial Service Separation Code (ISVC_SEP_CD) of “1087” indicating that they were discharged from the Army prior to the completion of their Initial Entry Training. Next, the Enlisted Career Status Code (ENL_CRER_STAT_CD) was examined. The code consists of one of two values representing a soldier still in the first-term (“1”) and a soldier that has reenlisted (“3”). Since we had snapshot data through FY2017, we classified 42.1% of the soldiers as “non-attrit” by seeing a “3” anywhere in their service record. We considered these observations accurately classified and removed them from the Master table.

Next, we turned to the Separation and Discharge Code (SPD_CD) and Initial Service Separation Code (ISVC_SEP_CD) from the Transaction table. After joining the

table to the new Master table, 1,869 observations were missing any Transaction table data. This prevented an examination of separation codes and another method was required. The complete snapshot records allowed us to determine if a soldier completed their first-term if they simply had any record after their first-term end date. Unfortunately, there was no end date present in the data. Therefore, we developed an algorithm to create the Calculated Obligation Date (CALC_OBL_DT) by adding the initially contracted number of years to the Basic Active Service Date (AFMS_DT) of each soldier. Once this value was calculated and adjusted for the assumption of successful completion if a record existed within the final three months of the first-term, 1,528 soldiers were classified as having failed to meet their obligation (“attrit”) and were removed from the Master table.

Next, we examined the remaining 240,365 observations for “good” separation codes. Of the 172 unique SPD_CD categories and 55 unique ISVC_SEP_CD categories, 46 of them were determined to represent successful completion of the first-term of service (Appendix A). The separation codes allowed us to classify another 26% of the total population as having successfully completed their contractual obligation period. At this point, we had classified over two-thirds of our total population as successful without even referencing the extensive and questionable list of “bad” separation codes.

Finally, the remaining 131,239 soldiers required a classification determination without referencing the separation codes. Instead, we used the derived CALC_OBL_DT as used for soldiers with no transaction data. We removed the 3,299 observations having no value for an initial obligation duration. Of the remaining observations, 83% had no record in the master data table beyond their CALC_OBL_DT. We classified the soldiers as having failed to meet their initial contractual obligation (“attrit”).

Figure 2 summarizes our methodology of classification as a flowchart. Along with the classification walk-through is a final summary table of our cohort dataset that includes the total number of observations and attrition rates per fiscal year. At a glance, the data appears valid as the number of accessions across fiscal years is steady and the attrition rates are consistent and near the 27-30% attrition rates reported in previous studies.

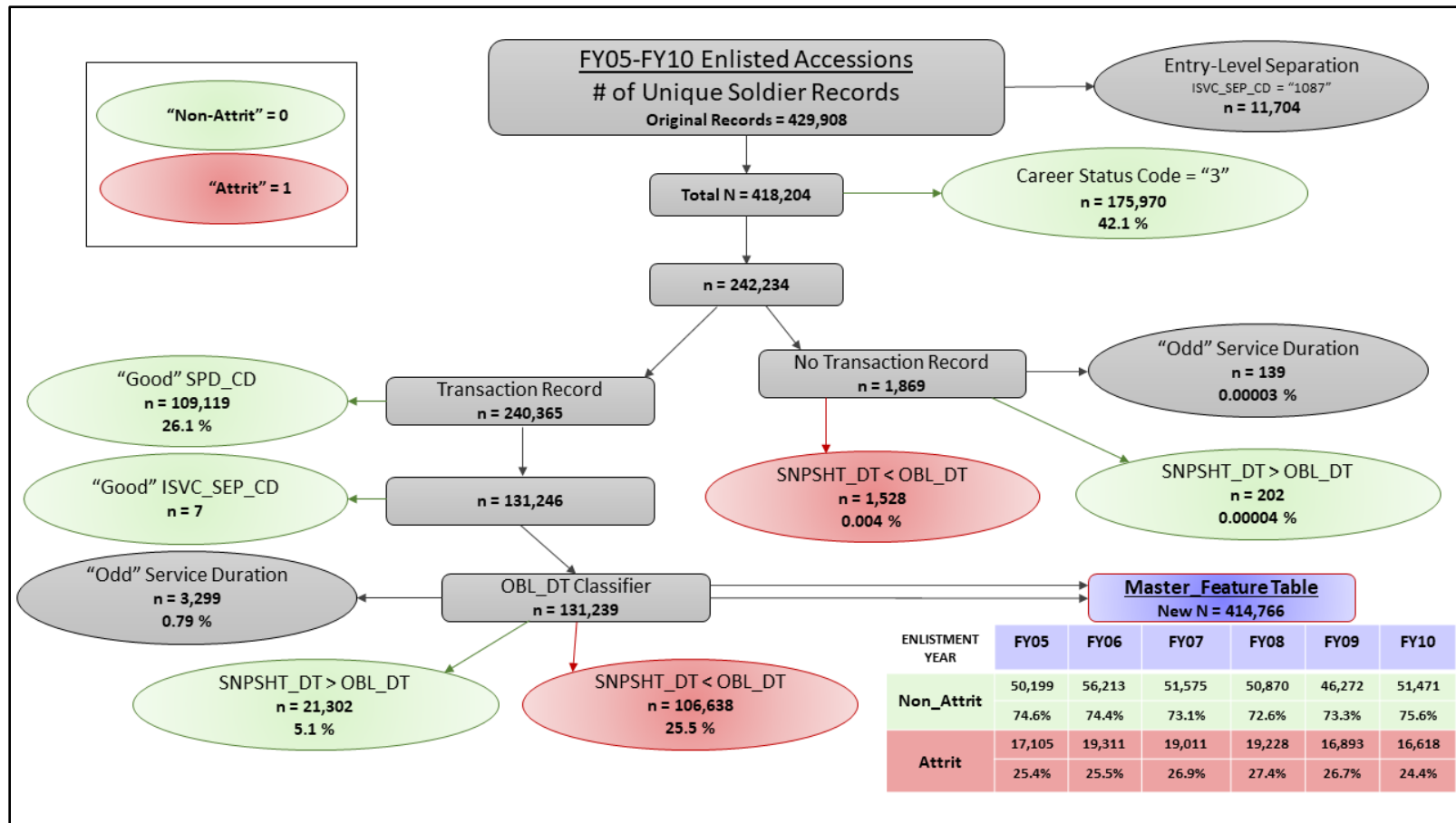


Figure 2. Classification Summary of the Response Variable

2. Merging the Predictor Variables

Having created our cohort dataset with a response variable, the next step was to determine the predictor variables to be included and select the appropriate time-based snapshot record from the additional data tables. The process involved identifying whether the variable required selection of the earliest snapshot date of a soldier (assumed to be the date at time of enlistment) or the record closest to, but not greater than, the calculated end of the initial contractual service obligation. The latter determination accounted for soldiers that completed their service obligation and had many snapshot dates in the data tables past their first-term or for those soldiers that failed to complete their initial obligation and had a final snapshot date any time prior to the first-term end date.

We made a distinct choice to identify variables that changed over time (e.g., *Rank*, *Marital Status*, *Education Level*) versus those that stayed constant (e.g., *Gender*, *Home of Record*, *Age at Enlistment*). For consistency, we joined constant variables with the record having the earliest snapshot date. To determine the correct end date record for the time-varying factors, we sliced the data tables containing predictor variables for each soldier by snapshot date and merged the applicable record with the cohort dataset.

3. Data Stratification

The last step of data preparation was to create training and test datasets allowing for the eventual validation of the “best” model. Since the research plan included an analysis of fiscal year differences, we stratified the data by fiscal year of accession. Additionally, we stratified the data by the response variable. This formal stratification process ensured that both the training dataset and the test dataset would have enough observations to allow for independent modeling over the spectrum of fiscal years and an appropriate number of observations for model training.

4. Modeling Techniques

After stratifying the data, a random 80/20 split kept 80% of the data as the training dataset for model building and selected 20% of the observations as the testing dataset to be “put in the vault” and not touched until the final model fits had been identified. We

constructed a binary logistic regression model utilizing an adapted backward stepwise regression with purposeful selection (Zhang, 2016). Also, we developed classification trees and random forest classification models for predictive power comparison utilizing the test dataset. The receiver operating characteristic (ROC) curves and Area Under the Curve (AUC) calculations served as the measures of performance for model comparison.

Our statistical methods included descriptive statistics and univariate analysis. Additionally, we performed multivariate logistic regression analysis utilizing an adapted backward stepwise regression with purposeful selection (Zhang, 2016). The analysis included verification of model assumptions regarding confounding variables and collinearity. Due to our large dataset, we addressed the issue of the “ p -value problem.” In large-sample studies, miniscule effects can be found as statistically significant. Previous research published in the *Information Systems Research* journal cautions against relying solely on p -values as they “can lead to claims of support for hypotheses of little or no practical significance” (Lin, Lucas Jr., & Shmueli, 2013, p. 906). In other words, results may have extreme statistical significance but no real-world applicability to the problem. The authors note that some statisticians favor the lowering of the p -value downward from the standard 0.05 to a level that retains “many” of the predictor variables. The analysis should then focus more on variable proportions, differences in effect sizes represented by the variable coefficients (marginal analysis), and charts representing descriptive statistic relationships rather than statistical significance. We selected a p -value threshold of 0.001 and carefully considered each variable throughout the purposeful selection.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DESCRIPTIVE STATISTICS

A. COHORT DATASET OVERVIEW

We stratified the data by fiscal year of enlistment and summarized by variable to allow initial examination of the distribution of variables. While the numerical values were summarized by averages and standard deviations, we grouped the binary and categorical variables by each of the factor levels to examine the differences in the rates of attrition. The examination of the variable summaries provided a method of confirming the proportions of our population against the results of previous studies.

Table 6 provides count and relative frequencies of the response variable across all fiscal years. The summaries indicate a consistent attrition rate of approximately 26% across the fiscal years. Though our attrition rates were a bit below the bulk of the GAO and RAND research findings, we were very close to the 28% steady attrition rate reported in the most recent study by Farrell (2017). Most importantly, the consistency of our total attrition rate findings was a very positive sign that our methodology for classifying the response variable was sound.

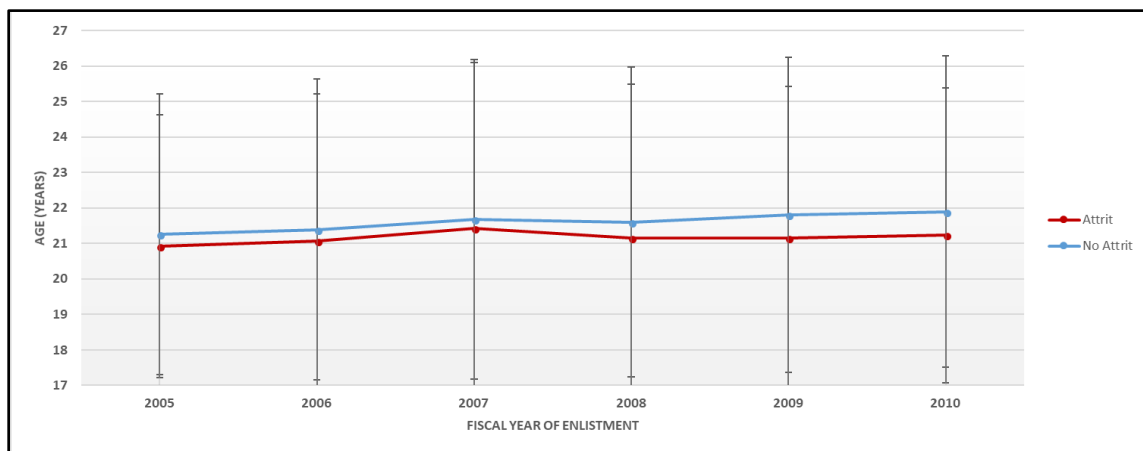
Table 6. Full Cohort Dataset—Attrition Rate by Fiscal Year of Enlistment

	2005	2006	2007	2008	2009	2010
Non-Attrit	50,199 (74.6%)	56,213 (74.4%)	51,575 (73.1%)	50,870 (72.6%)	46,272 (73.3%)	51,471 (75.6%)
Attrit	(25.4%) 17,105	(25.5%) 19,311	(26.9%) 19,011	(27.4%) 19,228	(26.7%) 16,893	(24.4%) 16,618

Due to the lack of publicly available data from previous studies and for transparency in our research, we provide a full summary of our cohort dataset in Appendix B. The numeric features include the mean and standard deviation calculated for each of the fiscal years of accession. The binary and categorical factors contain data counts and proportions stratified by the fiscal year of accession for all factors and levels. Missing data is reported by a factor level of “NA” for the variables that had incomplete cases.

B. NUMERIC FEATURES SUMMARY

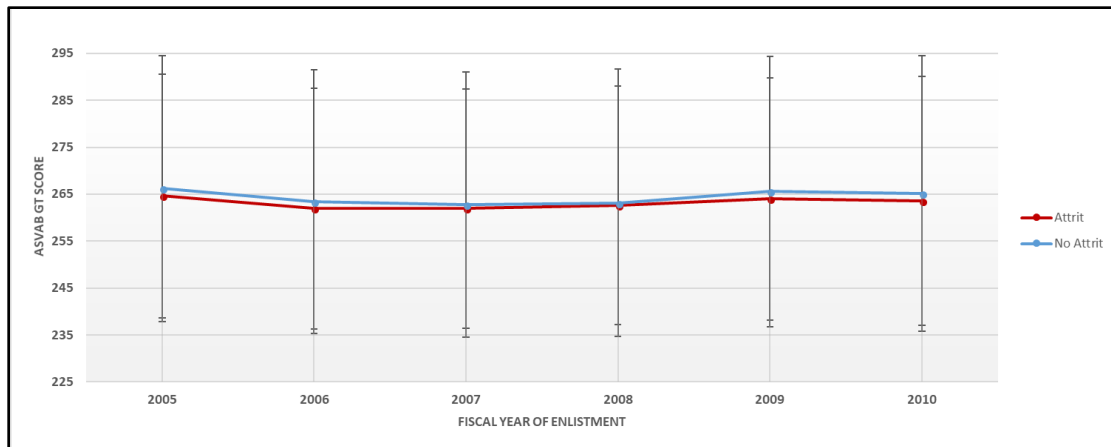
We summarized the numeric features by stratifying the calculated mean of the data for each of the fiscal years of enlistment. In his earliest research, Buddin reported finding a significance in the age of recruits at the time of enlistment (1985). Our data showed a higher average enlistment age of soldiers who successfully completed their first term (Figure 3) and the trend was consistent across all fiscal years of enlistment; however, the differences were all well within overlapping standard deviation bars.



Note: Error bars represent standard deviation.

Figure 3. Average Enlistment Age by Attrition Category

Unlike Martin's 1995 research, our data indicated very little difference in ASVAB scores between the attrition categories. Whereas Martin collapsed the AFQT percentiles into categories less than 65 and those greater than or equal to 65, our use of the ASVAB GT score similarly quantified a recruit's performance. Once again, any difference in average GT scores was minor and well within the margin of error between the attrition categories (Figure 4).



Note: Error bars represent standard deviation.

Figure 4. Average ASVAB GT Score by Attrition Category

The average number of days deployed was different between the attrition categories (Figure 5). The difference could be an indicator of the sense of purpose and accomplishment that can result from performing the mission for which a soldier has trained. However, it may only reflect the length of the initial contract and the related amount of time a soldier was eligible for possible deployment prior to discharge. Further analysis is required of the full deployment history of the soldiers to better understand this relationship. Incidentally, the sharp decrease in the average number of days deployed for soldiers completing their first-term matched expected results from the creation of our cohort dataset as U.S. Army deployment schedules slowed after the Iraq surge in 2007.

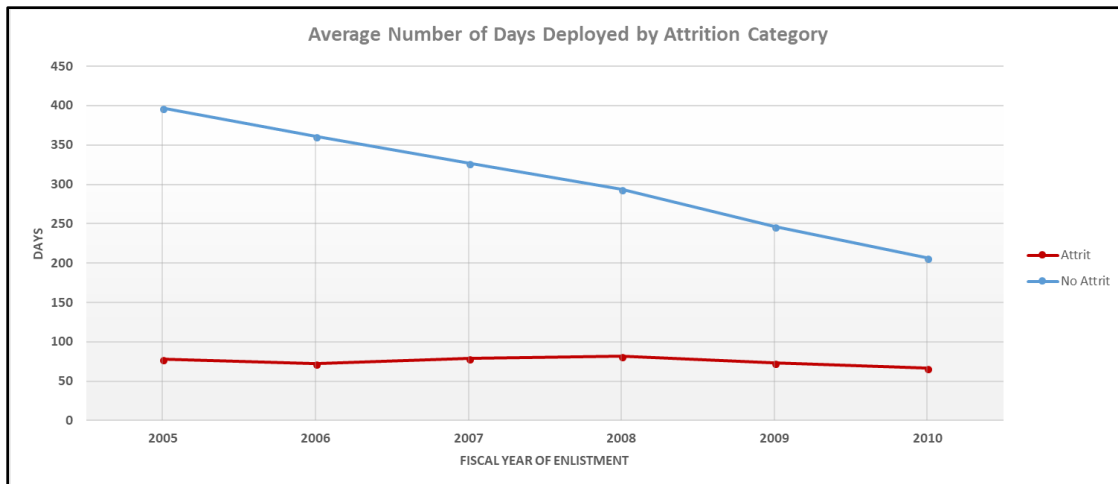


Figure 5. Average Number of Days Deployed by Attrition Category

The remainder of the numeric variables showed even less of a difference between categories and provided little insight. However, they were retained in the cohort dataset for further statistical analysis.

C. BINARY FEATURES SUMMARY

Every identified, published attrition study noted that gender and the source of a high school certification were significant factors for determining a soldier's probability of successfully completing the first-term of enlistment. Therefore, our initial analysis of the binary data examined the proportion of soldiers in each of these categories. Specifically, we were expecting to see a higher proportion of women and GED high school certifications among the population of soldiers who failed to complete their first term.

In fact, our data revealed a difference in the proportion of women that fail to complete their first-term compared to men. However, our average attrition rates across all fiscal years of enlistment for women of 40% and 24% for men were less than the results of 51% and 31% for women and men, respectively, reported in the 2005 attrition research of Buddin. Figure 6 summarizes our findings for each of the fiscal years of enlistment for males versus females. The female attrition rate ranged from 34% to 42% while the male attrition rate fell between 22% and 25%. The results consistently showed a higher attrition rate among women. The inability to duplicate the results of the previous research are likely

due to vague separation discharge codes and the lack of specific data methodology descriptions in the previous research.

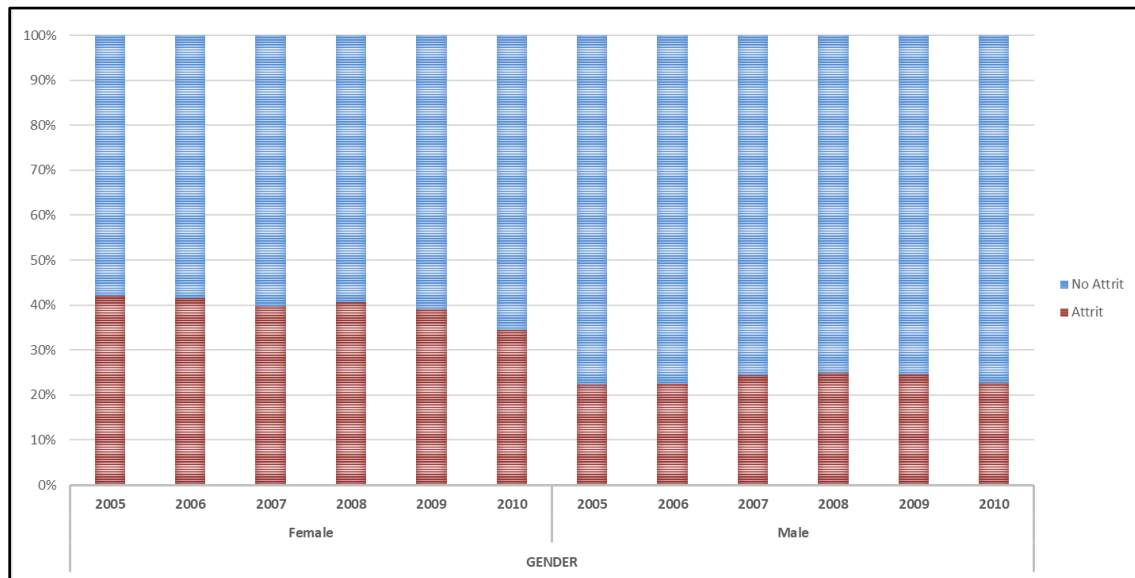
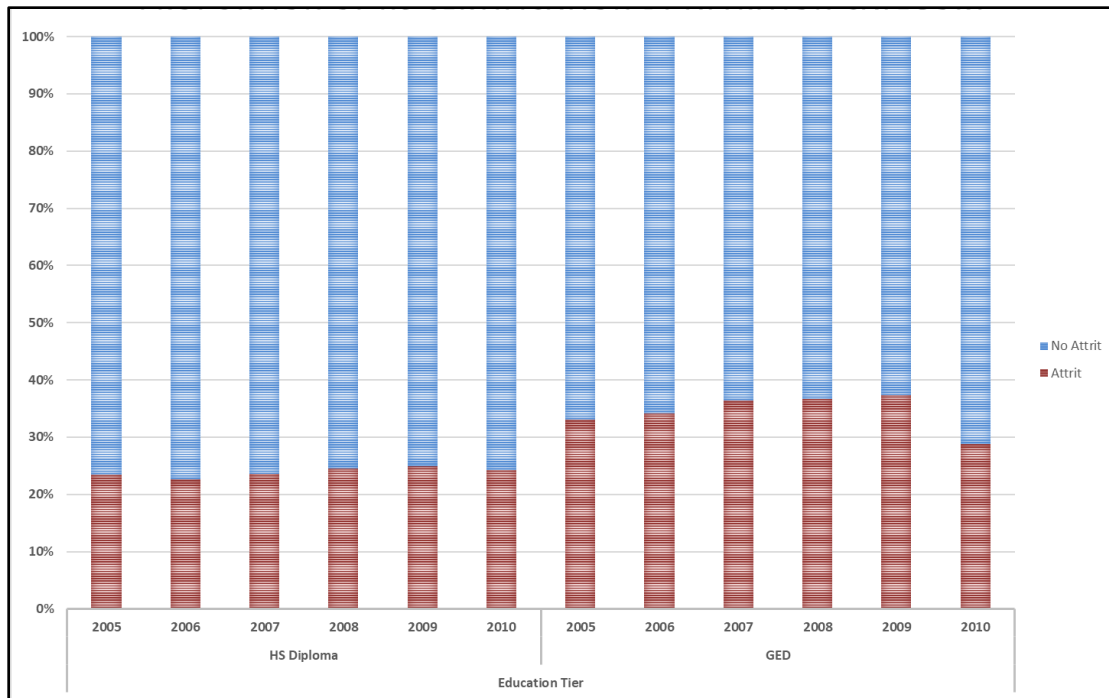


Figure 6. Proportion of Gender Levels by Attrition Category

As with our findings for gender, soldiers with a high school diploma had a lower attrition rate compared to soldiers holding a GED throughout our cohort dataset, though our rates were lower than the proportions found by Buddin (2005). High school graduates had a steady attrition rate of approximately 23% throughout the period of study while the attrition rate of enlistees receiving a GED fluctuated between 28% and 36% (Figure 7).



Note: Missing data and other high school certification levels have been omitted for clarity.

Figure 7. Proportion of High School Certification by Attrition Category

During the construction of our cohort dataset, soldiers with prior service were identified by the gain codes found within the Transaction table. These soldiers appeared to have a significantly higher attrition rate than non-prior service soldiers (Figure 8). However, there is cause for concern in this finding regarding its validity. First, the result is very counter-intuitive, as it would be assumed that a soldier with prior-service experience is joining the Army with a full understanding of the commitment required and a willful choice to fulfill that obligation. Additionally, the variable was constructed from the notoriously ill-defined separation and gain codes (Military Personnel, 2000). Even if the gain codes are assumed to be accurate, the question remains of how their initial contractual obligation was handled with regard to the Basic Active Service Date (BASD) assignment as 87% of the soldiers classified as prior-service have less than two months of Active Federal Service in their personnel record when they are first seen within the Master data table within the PDE. Since our methodology uses the BASD to classify soldiers who fail to complete their first-term, an accurate BASD is critical. Though we initially included the prior service indicator in our analysis, the soldiers identified as prior service were excluded

from the final models to prevent bias as these soldiers comprised less than 5% of our total population and we were unable to clarify the underlying data.

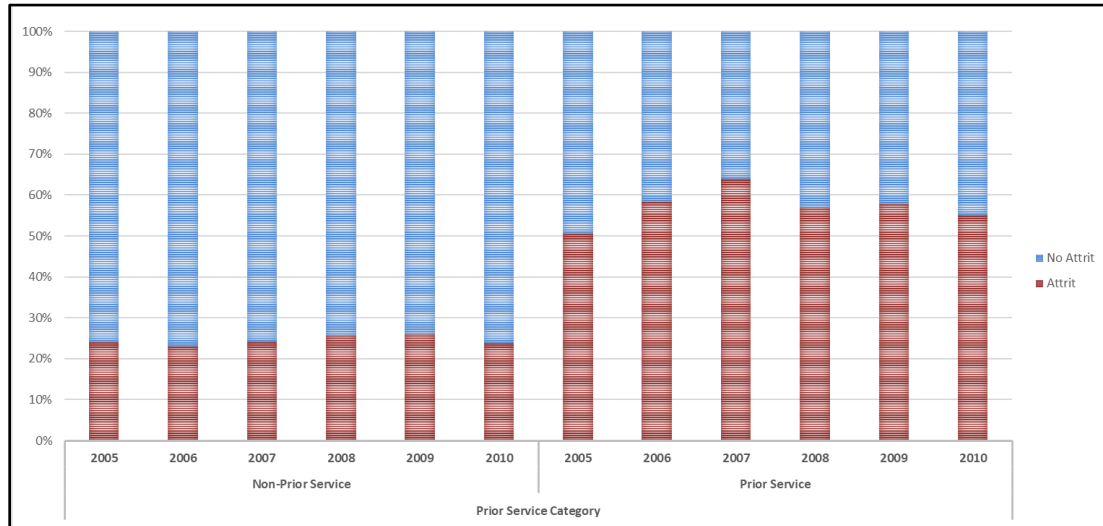


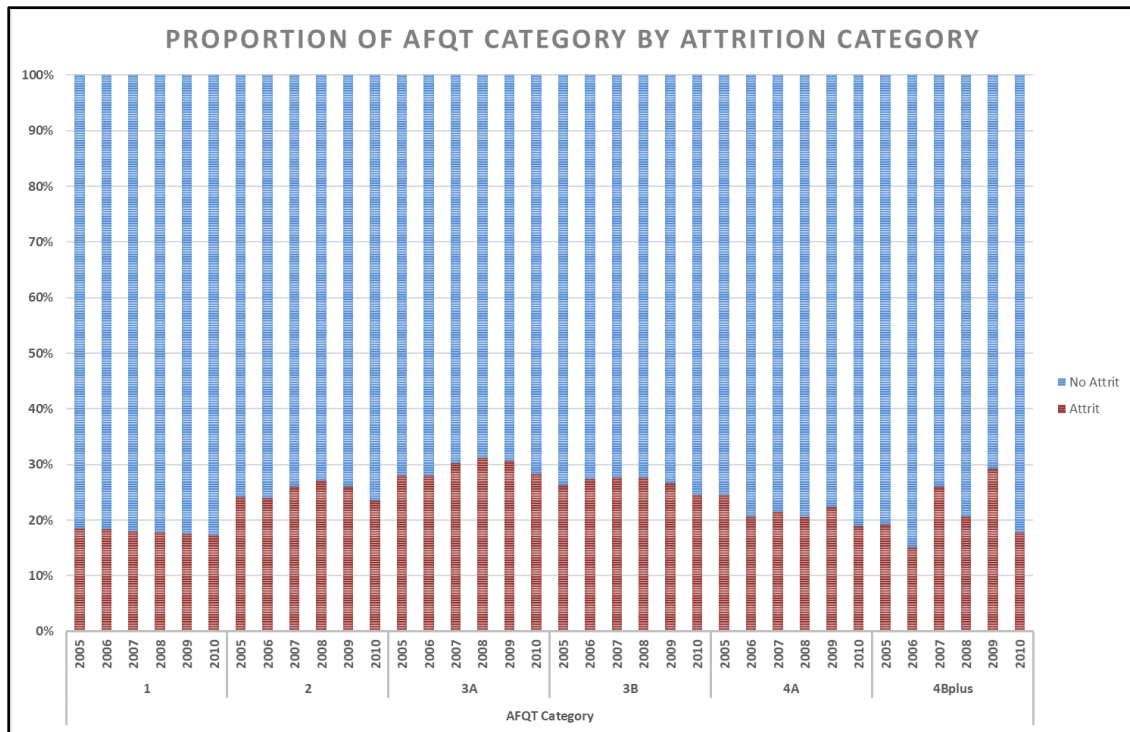
Figure 8. Proportion of Prior Service by Attrition Category

Analysis of the waiver data revealed insignificant differences in the attrition rates between soldiers receiving waivers to allow enlistment and soldiers without waivers. Moreover, soldiers who received a waiver and failed to complete their initial contractual obligation period represented less than 6% of the total population within the cohort dataset.

D. CATEGORICAL FEATURES SUMMARY

Our research purposely retained all possible predictor variables thought to remotely influence first-term attrition based on background research. Consequently, many of the categorical variables were found to show trivial differences in attrition rates among factor levels. However, a few variables were notable as possible influential predictors.

The AFQT category indicated a consistently lower attrition rate throughout the time frame of analysis in both the highest aptitude category (TSC-I) and the lower category (TSC-IVA). Though the data was not collapsed as in previous studies, our results are similar to Buddin's (1995) and Martin's (1985) research. However, the number of soldiers who fell into these categories was relatively small and represented approximately 10% of the population.



Note: Missing values have been omitted for clarity.

Figure 9. Proportion of AFQT Category by Attrition Category

The original personnel data reported a soldier's Home of Record state, which was collapsed in our research to a region of the United States for modeling purposes. However, a look at the raw Home of Record state information is insightful. Figure 10 displays the overall attrition rate of individual states within the continental United States and illustrates an attrition rate 10% higher in the southern portion of the U.S. compared to other locations. Enlistees from West Virginia appeared to be particularly prone to discharge from the Army before the end of their first term.

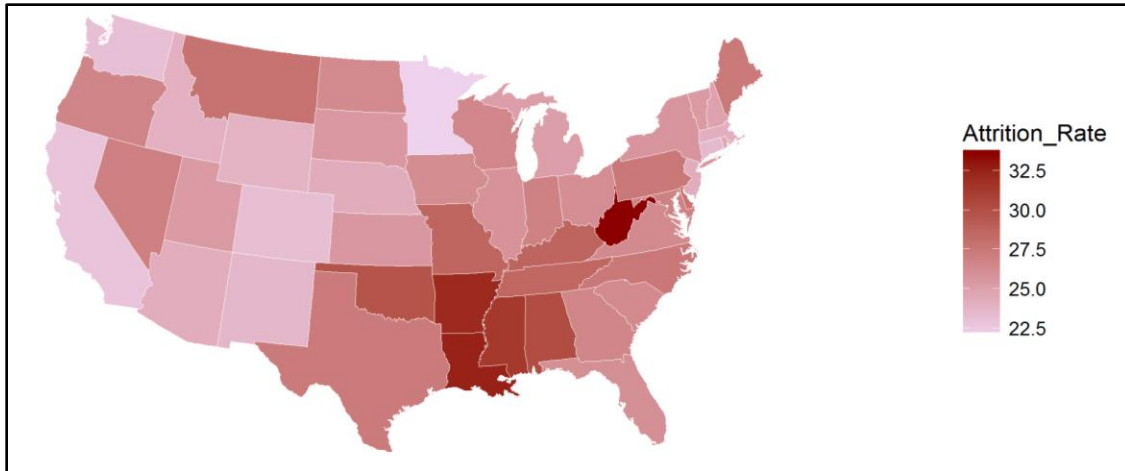
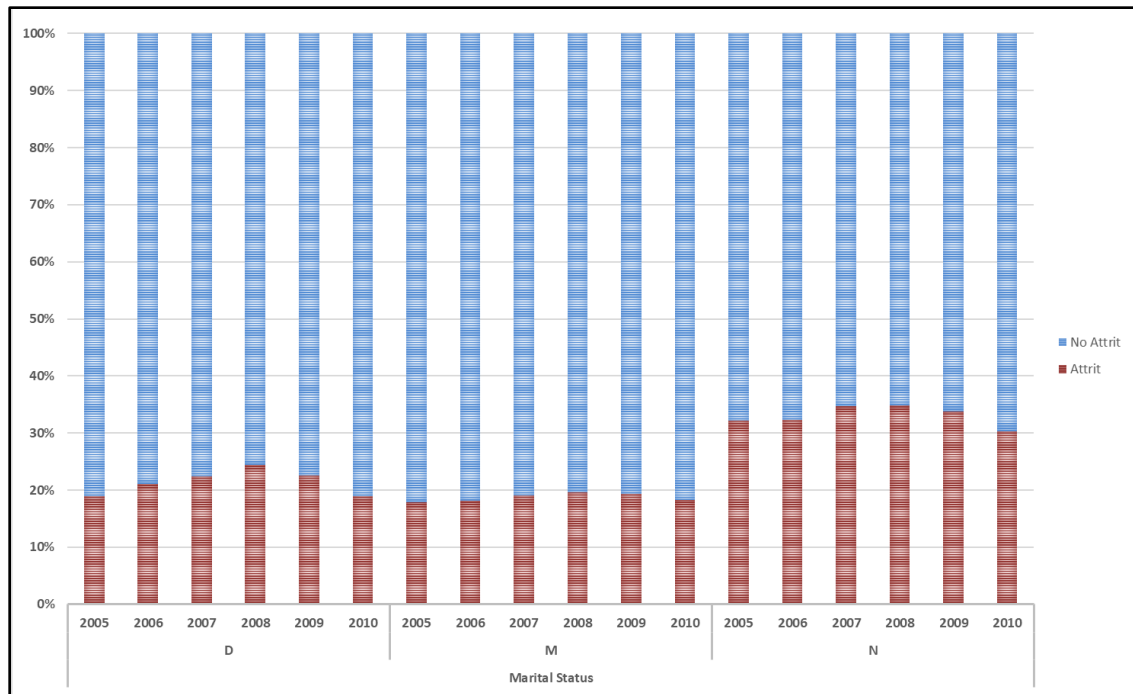


Figure 10. Attrition Rate by Home of Record

Our research considered the marital status of a soldier at the end of a successful first-term enlistment or at the point-of-failure for those soldiers who failed to complete the obligation period. The attrition rate of married or divorced soldiers averaged nearly 10% less than soldiers never having married (Figure 11). Unlike many of the categorical variables where differences in attrition rates are seen primarily in factor levels that include only a small amount of the total population, married and divorced soldiers made up nearly half of the cohort. The finding may be explained by a married soldier's sense of commitment to his dependents in providing for their welfare. Additionally, divorced soldiers frequently have children requiring support whether they are still living with the soldier or being provided financial support from alimony and child support payments.



D-Divorced, M-Married, N-Never Married. “Other” category removed for clarity.

Figure 11. Proportion of Marital Status by Attrition Category

A notable exclusion from the list of insightful categorical variables includes the rank of a soldier at enlistment. The Army offers an incentive of higher ranks (paygrade and responsibility) to incoming recruits having certain academic achievements or referring additional enlistees. While this incentive may offer the benefit of higher recruitment volume, it does not appear that enlistees with a higher rank are more likely to complete their first-term than any other recruits.

E. UNIVARIATE ANALYSIS

Prior to a univariate analysis, the training and testing datasets were created. The stratified observation counts for both the training and test datasets are consistent and represent the 80/20 split of the data (Table 7). The totals provide a large enough sample size for predictive modeling and eventual validation. The univariate tables provide summaries of data contained only in the training dataset.

Table 7. Training and Test Datasets—Counts by Fiscal Year of Enlistment

		2005	2006	2007	2008	2009	2010
Non-Attrit	Train	40187	45044	41314	40704	37082	41123
	Test	10012	11169	10261	10166	9190	10348
Attrit	Train	13657	15376	15155	15375	13450	13349
	Test	3448	3935	3856	3853	3443	3269

For univariate analysis, a logistic regression model was built for each predictor variable to explore the relationship between the variable and the first-term attrition outcome. Due to the sizable population in our research, we purposely selected a significance level of 0.001 for the identification of significant variables. Unfortunately, the statistical testing identified nearly every variable as significant due to the population size. Despite the lack of key variable identification provided by the summary, the tables are presented in full to provide the research community with a complete list of our cohort dataset. Most importantly, the tables provide evidence of attrition rates for all variables considered in our research.

1. Numeric Features

The numeric features are summarized by their averages and standard deviations (Table 8). Interestingly, the number of dependents does not appear to be associated with a soldier's success in completing their first-term. The insignificance of the number of dependents is interesting based on our earlier finding of possible significance in the marital status of a soldier. The mean ASVAB GT Score is nearly identical among soldiers regardless of their first-term success. Soldier deployment history appears very different. Soldiers who attrit had four times fewer deployments and total number of days deployed on average than soldiers successfully completing their first term. While the log-likelihood *p*-values provide limited utility in our research, the summary data alone is useful for better understanding the characteristics of new Army recruits.

Table 8. Univariate Summary of Numeric Features: Mean (Std Dev)

	No Attrit	Attrit	Total	p-value
Age at Enlistment Mean (Std Dev)	21.6 (4.32)	21.16 (4.23)	21.49 (4.3)	< 0.001
ASVAB GT Score	264.31 (28.54)	263.1 (25.77)	263.99 (27.84)	< 0.001
Contract Duration	3.76 (0.91)	4.03 (1.02)	3.83 (0.95)	< 0.001
Days Deployed	307.25 (225.17)	75.65 (151.14)	246.97 (231.9)	< 0.001
Dependents	0.33 (0.86)	0.33 (0.85)	0.33 (0.85)	0.1839
Deployments	1.98 (1.51)	0.53 (1.04)	1.6 (1.54)	< 0.001
Height at Enlistment	68.73 (3.23)	68.2 (3.49)	68.59 (3.31)	< 0.001
Hostile Injuries	0.02 (0.16)	0.01 (0.11)	0.02 (0.15)	< 0.001
Max Time-in-Grade	1.69 (0.89)	0.9 (0.81)	1.48 (0.93)	< 0.001
Non-Hostile Injuries	0.01 (0.08)	0.01 (0.09)	0.01 (0.09)	< 0.001
Weight at Enlistment	168.29 (31.7)	165.97 (33.8)	167.68 (32.28)	< 0.001

2. Binary and Categorical Features

Among the binary and categorical features, due to our large population, nearly all factor levels were significant based on the log-likelihood *p*-values. However, the summary table serves as a simple comparison tool to relate our attrition rate findings with both previous and future attrition studies (Appendix C). Raw data counts and their associated proportions are made available for all variables considered in our research. Additionally, the totals stratified by only attrition category accurately reflect the descriptive statistics previously reviewed.

Notably, the attrition rate of prior service soldiers was more than 30% higher. This finding is extremely counter-intuitive and may only result from the use of personnel gain codes. However, the finding may have merit as it duplicates related findings by Smith that showed lower graduation rates amongst trainees that repeated portions of the Ranger

Assessment course compared to first-time students (Smith, 2017). As in previous research findings, female soldiers and GED recipients had attrition rates more than one and a half times that of males and soldiers with high school diplomas.

Both current U.S. citizens and soldiers born in the United States failed to complete their first term at twice the rate of non-citizens and those born outside the United States. Regarding a soldier's education level at the time of enlistment, soldiers having only high school experience failed at a higher rate (27%) than those with even some college experience. However, soldiers holding advanced degrees are half as likely to fail in their first term. Over half of soldiers assigned to a unit authorized by a TDA document failed to complete their first term compared to soldiers in deployable organizations. Without specific unit data for verification, this may be more a function of medically discharged soldiers being placed in medical holding units prior to release and administrative discharges being moved to installation personnel action centers for processing. Despite the question of unit assignment, the variable was considered in our models to better understand the potential intervention focus for the ARD. The attrition rates among the individual military occupation specialties ranged from 5% within the highly specialized Special Operations community to 33% among the Military Police though most of the military occupations lost soldiers in their first term at a rate near 25%.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. MULTI-VARIATE MODELING AND ANALYSIS

Having split our cohort dataset into training and test sets, our initial modeling technique was a binary logistic regression to identify statistically significant variables and to generate model coefficients that may be used for a prediction tool. Additionally, we constructed both a binary classification tree and random forests and compared the discrimination power of the models by examining their ROC curves.

A. BINARY LOGISTIC REGRESSION

Research using logistic regression methodology in a 2017 Naval Postgraduate School thesis was recently completed that examined failure rates of soldiers attending the 75th Ranger Regiment assessment course (Smith, 2017). The research utilized a binary logistic regression model, which calculated the probability that a soldier will fail to graduate from the course. Like Smith's research, we chose to model only the main effects within our cohort dataset for the sake of simplicity and initial data exploration. Variable interactions and polynomial terms were not considered. Due to the binary response variable, we required predicted probabilities prior to performing certain model diagnostics; therefore, we constructed a model before performing checks on our underlying assumptions.

1. Purposeful Variable Selection

The first step of the binary regression we employed was to construct a full model with all variables in order to identify the non-statistically significant variables: $\alpha = .01$ level. Though we used a significance level of 0.001 in the univariate model, by allowing for more candidate variables to be selected in the multivariate analyses using a more conservative significance level, we aimed to reduce bias in our model by identifying all possible valid predictors. The model resulted in the identification of variables with multivariate p -values greater than 0.01: *unit region*, *home of record region*, *military occupation group*, *citizenship status*, *waiver (medical)*, and *waiver (drug)*. These variables were removed, and a new model was created.

The next step is described by Zhang (2016) in the *Annals of Translational Medicine* in which he provides a strategy for purposeful selection of variables. In his words, “the coefficients of variables should be compared to coefficients in the original one. If a change of coefficients...is more than 20%, the deleted variables have provided important adjustment of the effect of remaining variables” (p. 3). These variables are called confounding variables and can provide an indication of correlation while potentially increasing the variance or introducing bias. The difference in the coefficients between our models was negligible for nearly all variables except for the military occupation variables.

Since the *military occupation group* is a collapsed form of the *military occupation*, it is natural to believe that removal of the term caused instability in the model. To test the theory, we reconstructed the model with *military occupation group* added back. As expected, the coefficients were nearly unchanged between models. Whereas most confounding variables should be kept in a model, our knowledge of the relationship between these variables led us to believe that the difference in coefficients was indicative of correlation. In the interest of a more parsimonious model, we removed the 17-level *military occupation* variable in favor of the 3-level *military occupation group*. Similarly, we identified the relationship between the *ASVAB GT Score* and the *AFQT Category* variables, which are derived from raw ASVAB line scores. Though the numeric *ASVAB GT Score* would provide the simplest model, there were 11,527 missing values compared to only 1,757 missing values in the *AFQT Category* predictor. The *ASVAB GT Score* variable was thus removed.

At this stage of the research, we were forced to make choices about variable selection based on known variable relationships and predictive use. For instance, the *Fiscal Year of Enlistment* indicator may be a statistically significant predictor in our training and test datasets. However, the historical data from FY2005 to FY2010 required for the determination of the response variable does not assist in predicting whether a current soldier who enlisted in FY2016 will complete the first term. Likewise, the rank of a soldier and the education level at the end of the first term is information that will not be available when using the model to predict a current soldier. Therefore, we removed the variables *Fiscal Year of Enlistment*, *Rank (Max)*, and *Education Level (Max)* and generated a new

model to identify any variables that were no longer significant. *Age (Enlistment)* was no longer statistically significant and was removed.

Our model now consisted of only the statistically significant variables. The five variables with the highest importance metric were very different from the other variables (Table 9). Unfortunately, the most influential variable, *Max Time in Grade*, was a variable created during the feature engineering of our cohort dataset and proved to be suspect under scrutiny. This concern combined with the challenge for a soldier to provide this input to a predictive tool led us to remove the variable from the model.

Table 9. Max Time in Grade Importance Comparison

Model with Max Time in Grade		Model without Max Time in Grade	
Variable	Score	Variable	Score
Max Time in Grade	173.66	Contract Duration	97.50
Contract Duration	158.91	Days Deployed	82.58
Days Deployed	70.07	Unit Type (“MTOE”)	71.31
Prior Service (“Yes”)	51.80	Education Tier (“GED”)	40.34
Education Tier (“GED”)	31.12	Gender (“Male”)	32.44

Next, the *Contract Duration* variable was converted to a categorical factor from a numeric variable as it represented one of four discrete values from three to six years. Additionally, the variables *Education Level (Enlistment)*, *Hostile Injuries*, and *Deployments*, were no longer statistically significant and were removed.

The final model consisted of five numeric predictor variables and 40 levels of 12 categorical variables with importance scores assigned for non-reference variables. The variable importance table (Table 10) provides the feature names sorted in decreasing order of importance. The number of days a soldier spends deployed was clearly the most influential feature in the model. Soldiers with three- or four-year enlistment contracts may have very little time left in their first term after the initial entry training period and a deployment. The length of the initial contract was also very influential in the model. It makes intuitive sense that a soldier’s probability of leaving the Army before the end of the first term would be impacted by the length of time a soldier spends in the first term.

Interestingly, unlike previous research, the most influential variables did not include *gender* and *education tier*. We believe our broader variable selection may account for this.

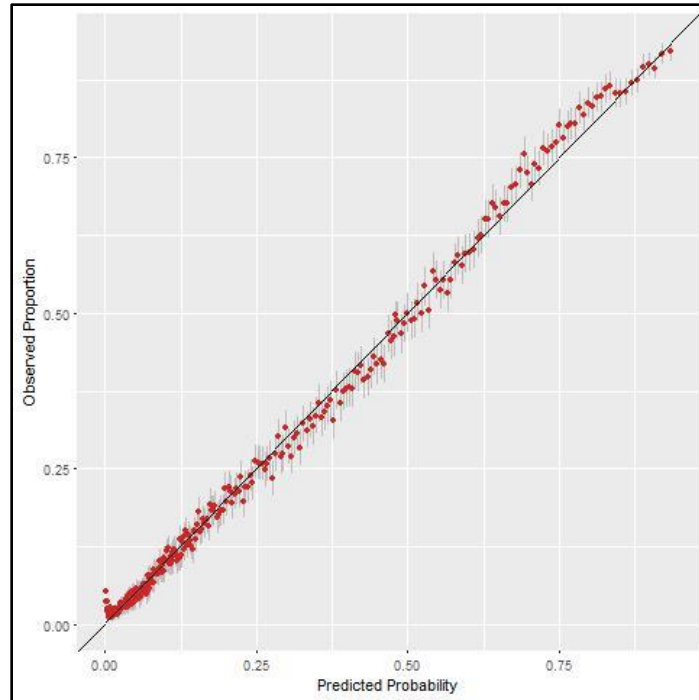
Table 10. Logistic Regression Variable Importance

Feature Name	Variable Code	Importance Score
Days Deployed	DPLY_DAYS_QY	195.3798
Contract Duration - 6 years	ASVC_AGMT_DRTN_YR_QY6	88.09909
Unit Type (Max) - MTOE	UNIT_TYPE_MAX_CLPSMTOE	71.87137
Contract Duration - 5 years	ASVC_AGMT_DRTN_YR_QY5	68.46465
Contract Duration - 4 years	ASVC_AGMT_DRTN_YR_QY4	50.40338
Rank (Enlistment) - PV1	RANK_MINPV1	41.07629
Education Tier - GED	EDU_TIER_CD2	41.02178
Gender - Male	GENDERM	32.85909
Marital Status - Never Married	MRTL_STAT_CD_CLPSN	32.68731
Military Occupation Group - Operations Support	CMF_FUNC_GRPPOS	31.07285
Rank (Enlistment) - PV2	RANK_MINPV2	31.0492
Prior Service - Yes	PRIOR_SRVC1	30.14718
Military Occupation Group - Force Sustainment	CMF_FUNC_GRPFS	25.81135
Dependents	DEP_QY_MEPS	23.2742
Rank (Enlistment) - PFC	RANK_MINPFC	20.11481
Citizenship Origination - Naturalized	US_CTZP_ORIG_CD_CLPSN	19.03031
AFQT Category - IIIB	AFQT_CAT_CD_CLPS3B	18.77884
Weight (Enlistment)	PN_WGHT_QY	18.23063
Unit Type (Max) - Multi-Component	UNIT_TYPE_MAX_CLPSMULTI	17.24774
Waiver (Conduct) - Yes	WAIVER_CONDUCT_YN1	16.80668
Waiver (Admin) - Yes	WAIVER_ADMIN_YN1	15.18946
AFQT Category - IIIA	AFQT_CAT_CD_CLPS3A	14.78308
AFQT Category - I	AFQT_CAT_CD_CLPS1	13.12183
Rank (Enlistment) - SSG	RANK_MINSSG	12.79808
Rank (Enlistment) - SGT	RANK_MINSGT	11.86484
AFQT Category - IVA	AFQT_CAT_CD_CLPS4A	11.61668
Education Tier - Other	EDU_TIER_CD3	9.688365
Height (Enlistment)	HGT_DM	6.240778
Non-Hostile Injuries	INJ_NON_HOSTILE_CNT	6.231266
Citizenship Origination - Outside U.S.	US_CTZP_ORIG_CD_CLPSC	6.146431
AFQT Category - IVB+	AFQT_CAT_CD_CLPS4Bplus	3.67796
Marital Status - Married	MRTL_STAT_CD_CLPSM	2.014206
Marital Status - Other	MRTL_STAT_CD_CLPSOther	1.277019

2. Model Diagnostics

Prior to the examination of the model coefficients, we performed diagnostics to assess the underlying model assumptions. Binary logistic regression models do not face the restrictive assumptions required in linear models; however, we examined the model results for outliers and the presence of multicollinearity.

In his seminal text, *Extending the Linear Model with R*, Faraway presents a method of visualization to subjectively gauge the goodness-of-fit of a logistic regression model (Faraway, 2016). The challenge arises from the fact that we cannot use the deviance in a binary logistic regression model to test the fit, as the deviance is a function of the fitted probabilities. Our analysis required the adaptation of his code to examine our model (Faraway, 2016, pp. 40-41). If the model is a good fit, we would expect the observed proportions of binned predictions to match the frequency of the event occurring within the bin. After the linear predictor was generated by our model, the predictors were grouped in 300 quantile bins. The first-term attrition events (response variable) were counted, and the mean of the linear predictors was calculated with a 95% confidence interval inside each bin. Once plotted, a slight variation was discernible at the higher values of predicted probability, but the line mostly fell within the confidence interval hashes (Figure 12). Without consistent or excessive deviation, there was no evidence that the model residual variance was excessive.



Since the linear predictor line mostly falls within the 95% confidence interval hashes, there is no evidence that the model residual variance is excessive.

Figure 12. Binned Predicted Probabilities and Observed Proportions.
Adapted from Faraway (2016).

To check for multicollinearity within a binary logistic regression model, the variable inflation factor (VIF) was examined. The variable inflation factor is a measure of the effect of correlation among two or more predictor variables within a regression model. The typical rule of thumb is variables with a VIF value greater than 10 are highly correlated and may cause problems within the model. In our model, none of the variables had a VIF value exceeding 1.5 and we proceeded with our analysis of the model (Table 11).

Table 11. Regression Model Variable Inflation Factors

Variable Code	VIF Value
ASVC_AGMT_DRTN_YR_QY	1.057994
PRIOR_SRVC	1.01333
CMF_FUNC_GRP	1.062777
GENDER	1.333007
RANK_MIN	1.021972

Variable Code	VIF Value
WAIVER_CONDUCT_YN	1.017749
WAIVER_ADMIN_YN	1.115067
EDU_TIER_CD	1.03149
INJ_NON_HOSTILE_CNT	1.001179
DPLY_DAYS_QY	1.076537
DEP_QY_MEPS	1.175652
HGT_DM	1.430239
PN_WGHT_QY	1.250949
AFQT_CAT_CD_CLPS	1.026545
US_CTZP_ORIG_CD_CLPS	1.004612
MRTL_STAT_CD_CLPS	1.042209
UNIT_TYPE_MAX_CLPS	1.02462

3. Regression Analysis

We begin our analysis with a look at the coefficient matrix output by our logistic regression model. The linear predictor estimates, log-odds, and probability are listed in the model coefficient matrix (Table 12). The simplest interpretation is by understanding that variables with positive linear predictor estimate values increase the probability of first-term attrition, while negative estimate values decrease the probability. Thus, we see many results that match our intuition. For instance, as the *contract duration* levels increased, the attrition probability increased. The most influential variable in the model, *days deployed*, was inversely related to the probability of attrition. Utilizing the probability, for each 30-day period that a soldier has deployed, his probability of first-term attrition is reduced by 30% if all other variables are held fixed. Other findings of note include the increased probability of attrition for heavier enlistees and decreased probability for taller enlistees.

Table 12. Regression Model Coefficients Matrix

Variable Name	Linear Predictor Estimate	Linear Predictor Error	Log-Odds	Prob.	Pr(> z)
Contract Duration - 3 years	Ref.				
Contract Duration - 4 years	0.64	0.01	1.90	0.66	< 0.001
Contract Duration - 5 years	1.25	0.02	3.50	0.78	< 0.001

Variable Name	Linear Predictor Estimate	Linear Predictor Error	Log-Odds	Prob.	Pr(> z)
Contract Duration - 6 years	1.75	0.02	5.74	0.85	< 0.001
Prior Service - No	Ref.				
Prior Service - Yes	0.68	0.02	1.97	0.66	< 0.001
Military Occupation Group - Operations	Ref.				
Military Occupation Group - Operations Support	-0.55	0.02	0.58	0.37	< 0.001
Military Occupation Group - Force Sustainment	-0.31	0.01	0.73	0.42	< 0.001
Gender - Female	Ref.				
Gender - Male	-0.57	0.02	0.56	0.36	< 0.001
Rank (Enlistment) - CPL	Ref.				
Rank (Enlistment) - PFC	0.51	0.03	1.67	0.63	< 0.001
Rank (Enlistment) - PV1	1.02	0.02	2.77	0.73	< 0.001
Rank (Enlistment) - PV2	0.78	0.03	2.19	0.69	< 0.001
Rank (Enlistment) - SGT	0.95	0.08	2.59	0.72	< 0.001
Rank (Enlistment) - SSG	1.84	0.14	6.28	0.86	< 0.001
Waiver (Conduct) - No	Ref.				
Waiver (Conduct) - Yes	0.29	0.02	1.34	0.57	< 0.001
Waiver (Admin) - No	Ref.				
Waiver (Admin) - Yes	-0.41	0.03	0.66	0.40	< 0.001
Education Tier - High School diploma	Ref.				
Education Tier - GED	0.55	0.01	1.74	0.64	< 0.001
Education Tier - No secondary school	0.61	0.06	1.83	0.65	< 0.001
Non-Hostile Injuries	0.34	0.05	1.41	0.58	< 0.001
Days Deployed	-0.01	0.00	0.99	0.50	< 0.001
Dependents	0.16	0.01	1.18	0.54	< 0.001
Height (Enlistment)	-0.01	0.00	0.99	0.50	< 0.001
Weight (Enlistment)	0.00	0.00	1.00	0.50	< 0.001
AFQT Category - II	Ref.				
AFQT Category - IIIA	0.20	0.01	1.23	0.55	< 0.001
AFQT Category - IIIB	0.26	0.01	1.29	0.56	< 0.001
AFQT Category - I	-0.35	0.03	0.71	0.41	< 0.001
AFQT Category - IVA	0.39	0.03	1.47	0.60	< 0.001
AFQT Category - IVB+	0.75	0.21	2.13	0.68	< 0.001
Citizenship Origination - Born in U.S.	Ref.				
Citizenship Origination - Naturalized	-0.74	0.04	0.48	0.32	< 0.001
Citizenship Origination - Outside U.S.	-0.23	0.04	0.80	0.44	< 0.001
Marital Status - Divorced	Ref.				
Marital Status - Married	0.06	0.03	1.06	0.51	0.044
Marital Status - Never Married	0.97	0.03	2.65	0.73	< 0.001

Variable Name	Linear Predictor Estimate	Linear Predictor Error	Log-Odds	Prob.	Pr(> z)
Marital Status - Other	0.25	0.20	1.29	0.56	0.202
Unit Type - TDA	Ref.				
Unit Type - MTOE	-0.89	0.01	0.41	0.29	< 0.001
Unit Type - Multi-Component	-1.91	0.11	0.15	0.13	< 0.001

Charts may prove more useful to examine a few of the variable effects. We selected the variable, *days deployed*, as the x-axis continuous discriminator due to its high influence in the model. First, we look at the chart that displays the typical significant variables found in other research (Figure 13). The higher first-term attrition probability is clearly visible in the female population until a soldier reached two years of total deployment where the probabilities converge. Like other research, a higher maximum attrition probability was apparent in soldiers with a GED instead of a high school diploma.

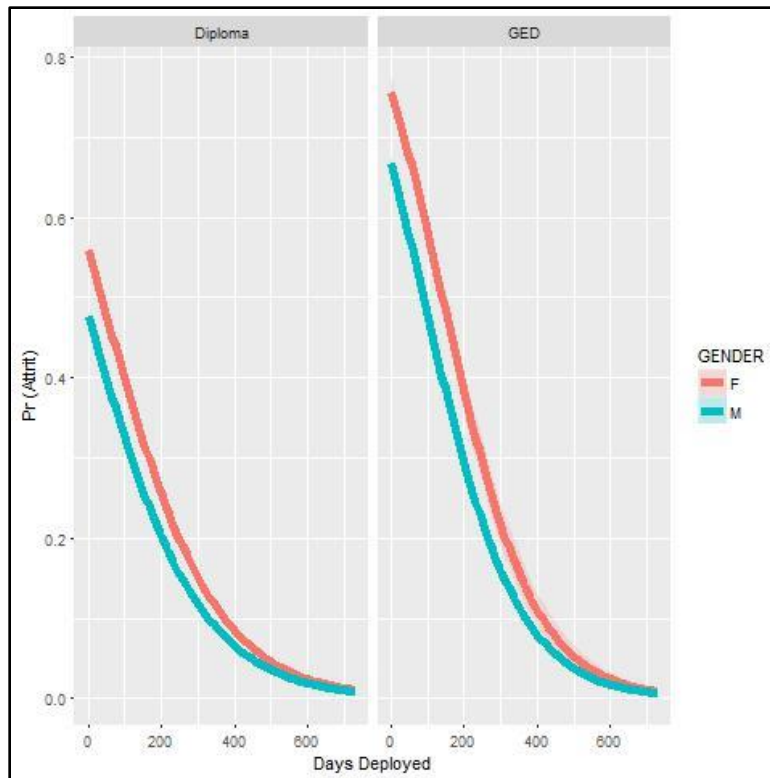


Figure 13. Attrition Probability—Days Deployed by Gender with Education Tier Panels

Another significant predictor often cited is a soldier's performance on the enlistment aptitude exam. Again, higher attrition probability was observed in women while the rate of decreased probability based on deployment days was consistent across all levels of AFQT Categories (Figure 14). The attrition probability of a male soldier arriving to his first unit after training averaged approximately 53% unless the soldier was in the highest percentile AFQT Category whereas his attrition probability was reduced to only 38%.

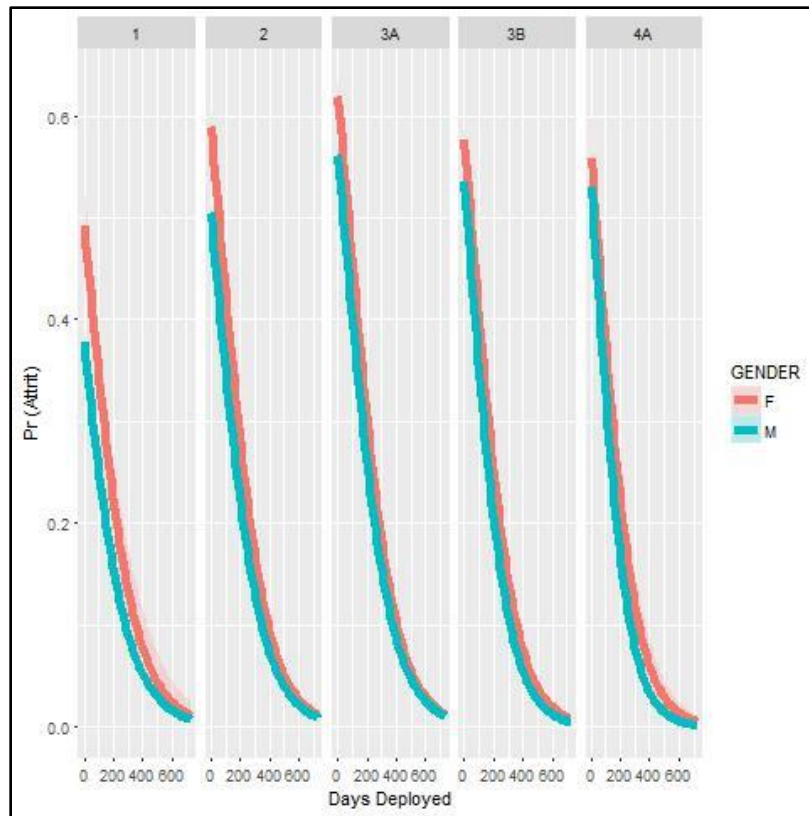


Figure 14. Attrition Probability—Days Deployed by Gender with AFQT Category Panels

There is a particularly striking difference observed in the attrition probabilities between the collapsed military occupation groups. Though individual military occupations were removed from the model, the combat arms specialties in the Operations group reflected a higher attrition probability of 59%: 10% higher than the other military occupation groups (Figure 15). Despite the lower density of women in the Operations specialties, their attrition probability closely mirrored that of men. However, the attrition probability for women did not reflect the probability reduction seen for men in the other military occupation groups.

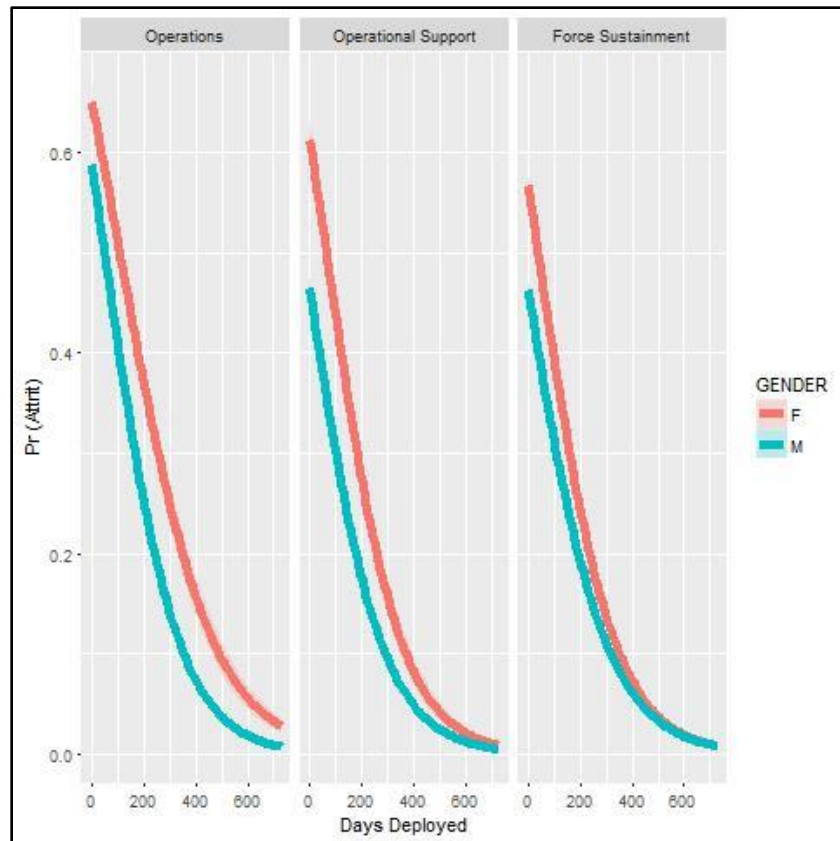


Figure 15. Attrition Probability—Days Deployed by Gender with Military Occupation Group Panels

Finally, we will look at probability differences among marital statuses. Married and divorced soldiers had similar attrition probabilities that never exceed 50%. However, divorced, female soldiers saw a reduction in their attrition probability much faster than married soldiers, as the number of deployed days increase. Both male and female single soldiers had a significantly higher probability of first-term attrition at 64% and 71%, respectively.

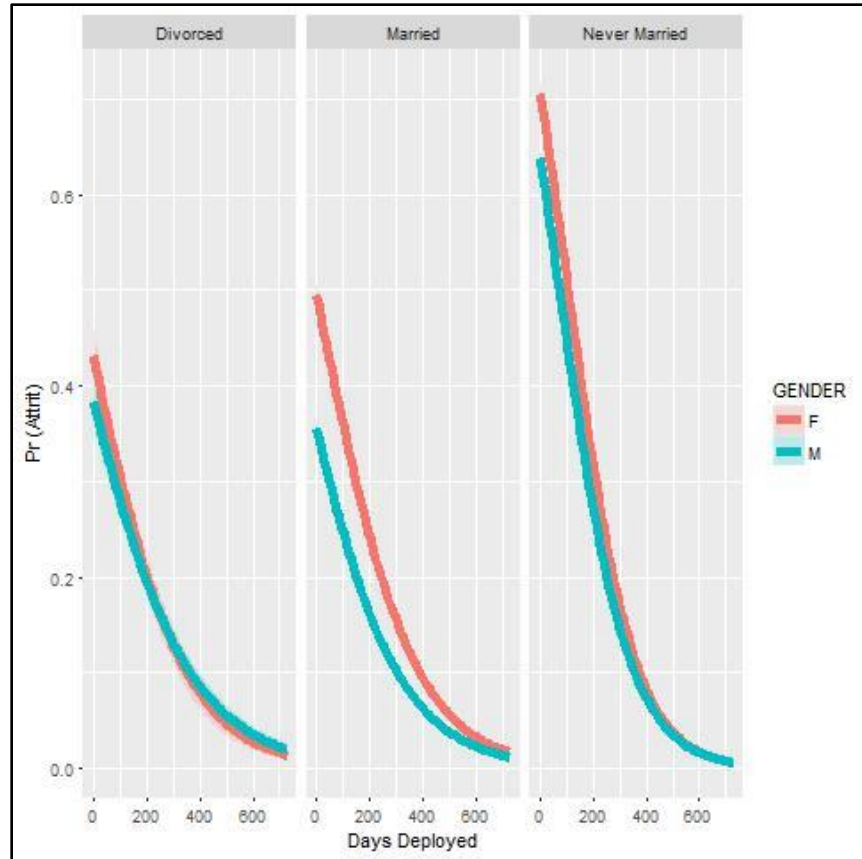


Figure 16. Attrition Probability—Days Deployed by Gender with Marital Status Panels

4. Prediction

Since our research goal was to develop models useful in a predictive tool, we utilized the confusion matrix and the ROC curve to assess the quality of our predictions. The confusion matrix provides the quantities of correct predictions, false negatives, and false positives (Table 13). The *specificity* of the model is a measure of the proportion of soldiers correctly predicted to complete their first term where the probability cut-off for prediction of attrition is greater than 0.5. The specificity of the model on our training dataset was $\frac{206,519}{206,519+33,569} = 86\%$. The *sensitivity* of the model measures the proportion of soldiers that will fail to complete their first term that are correctly predicted by the model. Our model sensitivity was calculated as $\frac{45,864}{45,864+18,255} = 71\%$. Additionally, the confusion matrix facilitates the calculation of the misclassification rate:

$$1 - \frac{206,519 + 45,864}{206,519 + 45,864 + 18,255 + 33,569} = 17\%.$$

Our misclassification rate on the training observations was low; accurately predicting 83% of the observations is very promising.

Table 13. Logistic Regression Training Dataset Confusion Matrix

Predicted Attrition Category	Observed Attrition Category	
	Non-Attrit	Attrit
Non-Attrit	206,519	33,569
Attrit	18,255	45,864

The confusion matrix is a static depiction of the predictive performance of our model. A better option is to vary the prediction cut-off threshold and examine the changes in the specificity and sensitivity in a ROC curve plot (Figure 17). The plot displays the false positive rate (1 - specificity) on the x-axis and the true positive rate (sensitivity) on the y-axis. The curve represents the change in the relationship between the two rates as the threshold varies. A very good test results in a curve pulled toward the top-left corner of the plot while a test performing no better than random chance will fall on the $y = x$ (diagonal) line. The color scale of the curve indicates the probability threshold assignment that produced the parametric point on the curve.

Another measure of performance of the model is the Area Under the Curve (AUC). As its name implies, this is a value signifying the approximated area of the polygon under the ROC curve. Since a worthless test falls on the diagonal line and the best test pulls to the upper left corner, the range of AUC is [0.5,1.0]. Obviously, a higher AUC value indicates a better performing model; moreover, an AUC value greater than 0.8 is typically considered very respectable.

Since the observations used to build the model are now predicted by the model and we had a low misclassification rate, the training dataset ROC curve produced a high AUC value of 0.866 (Figure 17). The best probability threshold to balance specificity and sensitivity may be 0.59 depending on policy decisions to balance the costs and benefits of

attrition intervention programs offered to soldiers that would complete their first term without preventative measures. The AUC value and misclassification rate were used throughout our modeling as the measures of performance for model comparison.

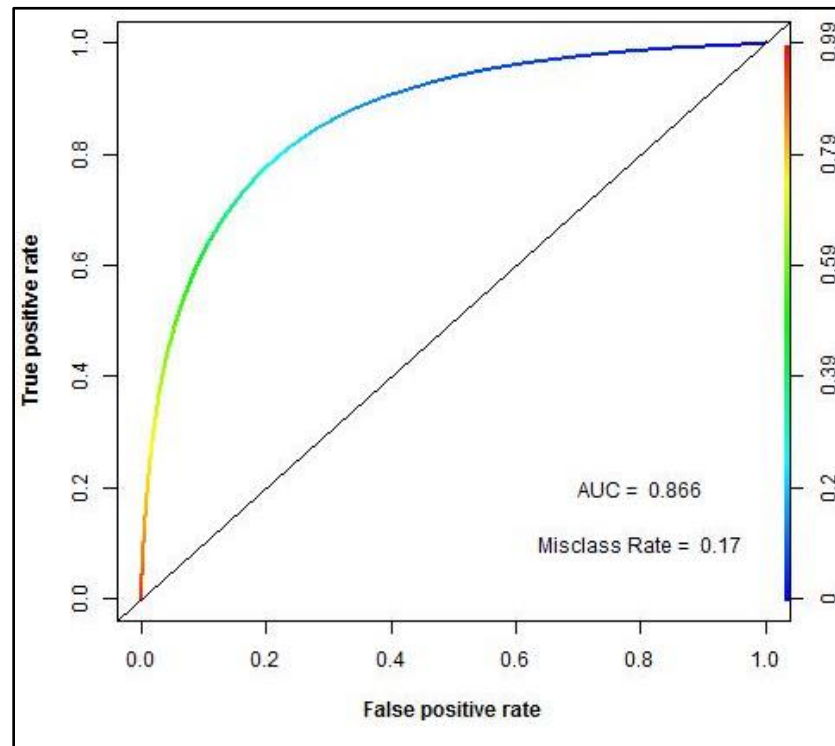


Figure 17. Attrition Classification by Logistic Regression—ROC Curve (Training Dataset)

Despite our large dataset, a 10-fold cross-validation technique was used to provide a better estimate of the classification rate (accuracy) of our model and a 95% confidence interval. 10-fold cross-validation consists of splitting the dataset into 10 folds, calculating the coefficients based on our modeling decisions with nine of the folds, and predicting the response variable for the fold kept out. After performing these actions for each of the folds, we averaged the accuracies and calculated the confidence interval. Our model resulted in an overall accuracy rate of 0.830 (95% confidence interval 0.827-0.833).

The ultimate step of our regression analysis was to take the test dataset “out of the vault” and use our model to predict the response variable. The model performed extremely

well on the test dataset with a misclassification rate of 17.2% (Table 14) and an AUC value of 0.8719 (Figure 18).

Table 14. Logistic Regression Test Dataset Confusion Matrix

Predicted Attrition Category	Observed Attrition Category	
	Non-Attrit	Attrit
Non-Attrit	43,394	7,266
Attrit	3,934	10,661

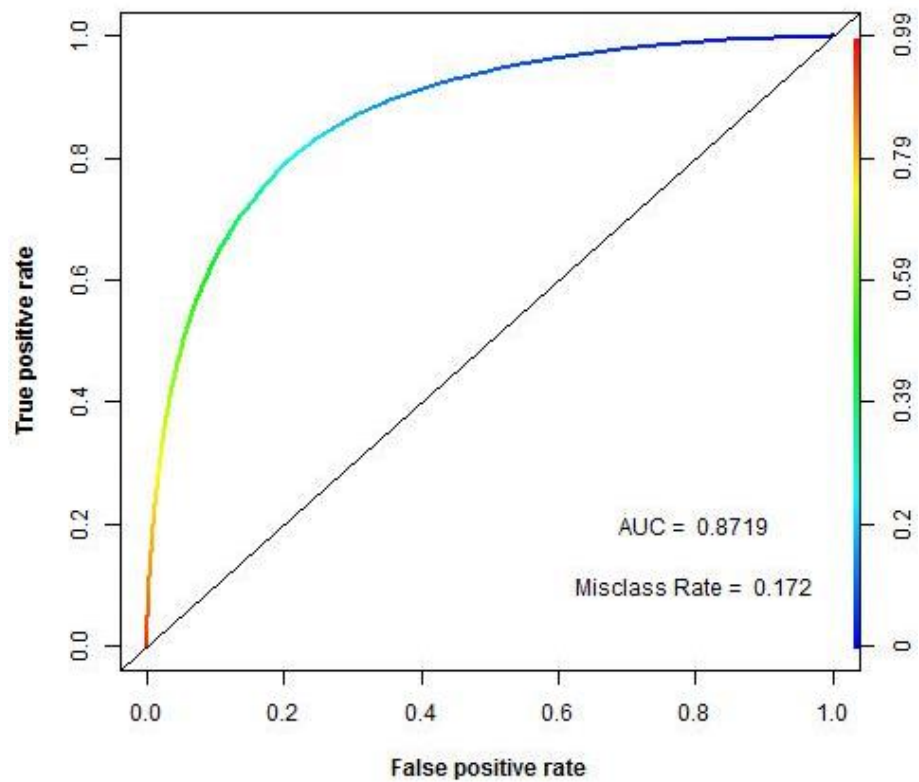


Figure 18. Attrition Classification by Logistic Regression—ROC Curve (Test Dataset)

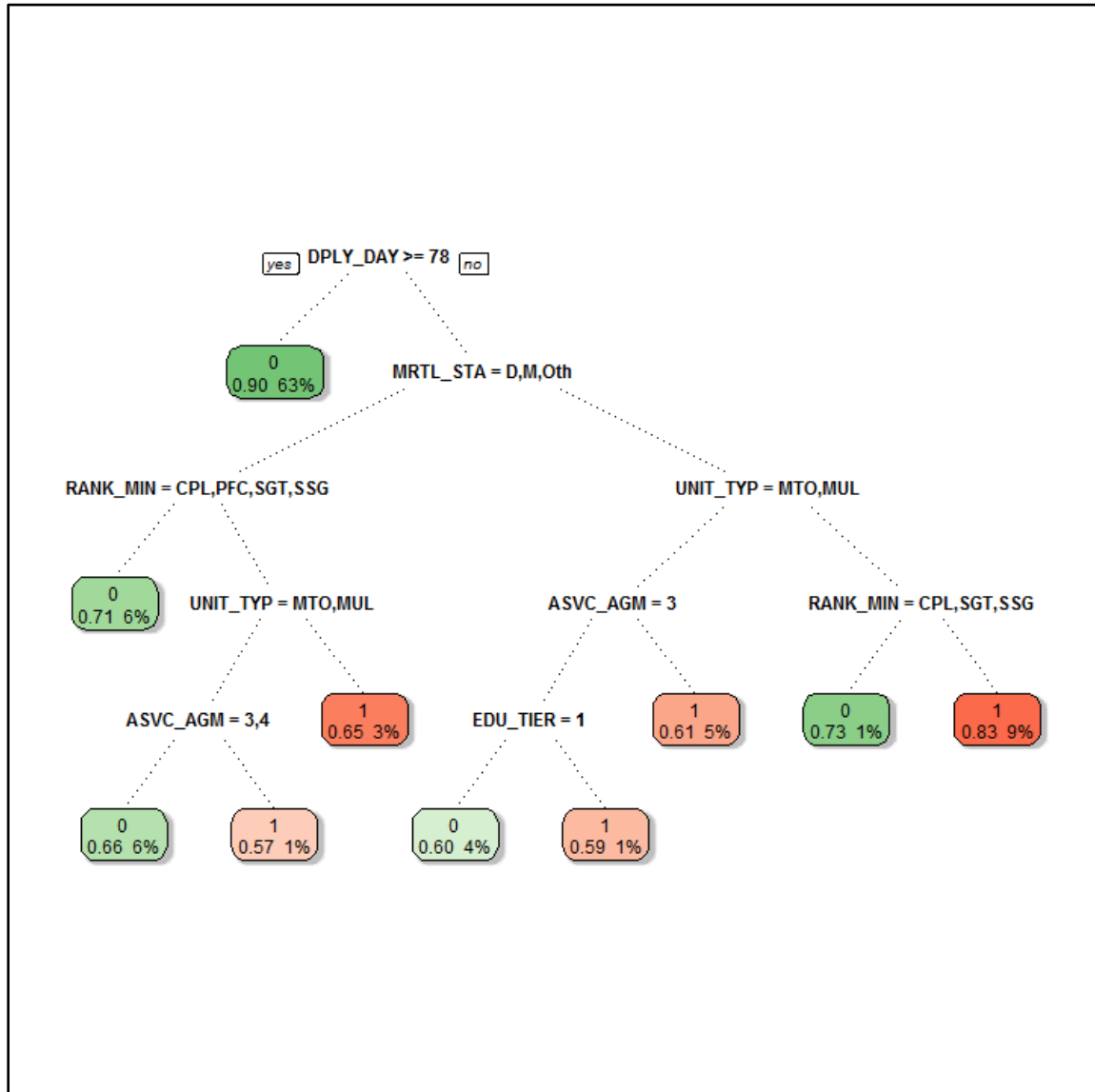
B. CLASSIFICATION TREE

We constructed a classification tree for comparison to the logistic regression model. Like nonparametric statistics, a tree model requires few assumptions regarding the distribution of the data and is constructed by partitioning (splitting) the variables at the point that minimizes the residual sum of squares in the two branches of the nodes. Trees also handle missing values well and provide an analyst with an easily explained graphical representation of the relationships among the predictor variables. The leaves of the tree provide information about the observations that were classified into each leaf. The purity of the leaf is the proportion of observations within the leaf that match the “winning” class and the level of purity is indicated by color shading. This value can be thought of as the probability a soldier with matching predictor values leading to the leaf is going to match the leaf classification. The percentage data reports the proportion of observations contained each leaf.

Though trees require fewer assumptions and much less effort in variable selection, they still must be pruned to determine the best size of the tree. We fit a tree on our training dataset and selected the smallest tree with a cross-validated error within one standard error of the minimum. In our tree, the complexity parameter was set to 0.0025 resulting in nine splits of our data. The first split was on *days deployed* which matched the most influential variable in the regression model (Figure 19). If a soldier has deployed for 78 days or more, she moves down the left branch into the leaf and has a 90% chance of completing her first term and the leaf contains 63% of the total dataset. Soldiers answering “Yes” to the implied questions posed at each node will always move down the left branch until reaching a leaf. Thus, a married soldier in the rank of PFC with no deployed days has a 71% chance of successfully completing his or her first term. Conversely, a single soldier in the rank of PFC with no deployed days belonging to an MTOE unit has an 83% chance of failing to complete his or her first term.

The tree mirrors the resulting influential variables identified in our regression analysis. Like the logistic regression, gender was a negligible factor in the classification tree and was not even included in the classification methodology. Also, the leaves

representing a soldier's high school diploma source (*EDU_TIER*) only represented 5% of the observations and provided minimal classification information.



Terminology: *DPLY_DAY* - Days Deployed, *MRTL_STA* - Marital Status (Divorced, Married, Other)
RANK_MIN - Rank at Enlistment, *UNIT_TYP* - Unit Type (MTOE, Multi-Compo)
ASVC_AGM - Contract Duration (years), *EDU_TIER* - High School Method (Diploma)

Figure 19. Attrition Classification Tree

The tree was used to predict the response variable of our test dataset. Though the performance of the model was very good, the ROC curve and AUC value of 0.8129 indicated that the classification tree model performed slightly worse than the logistic regression model.

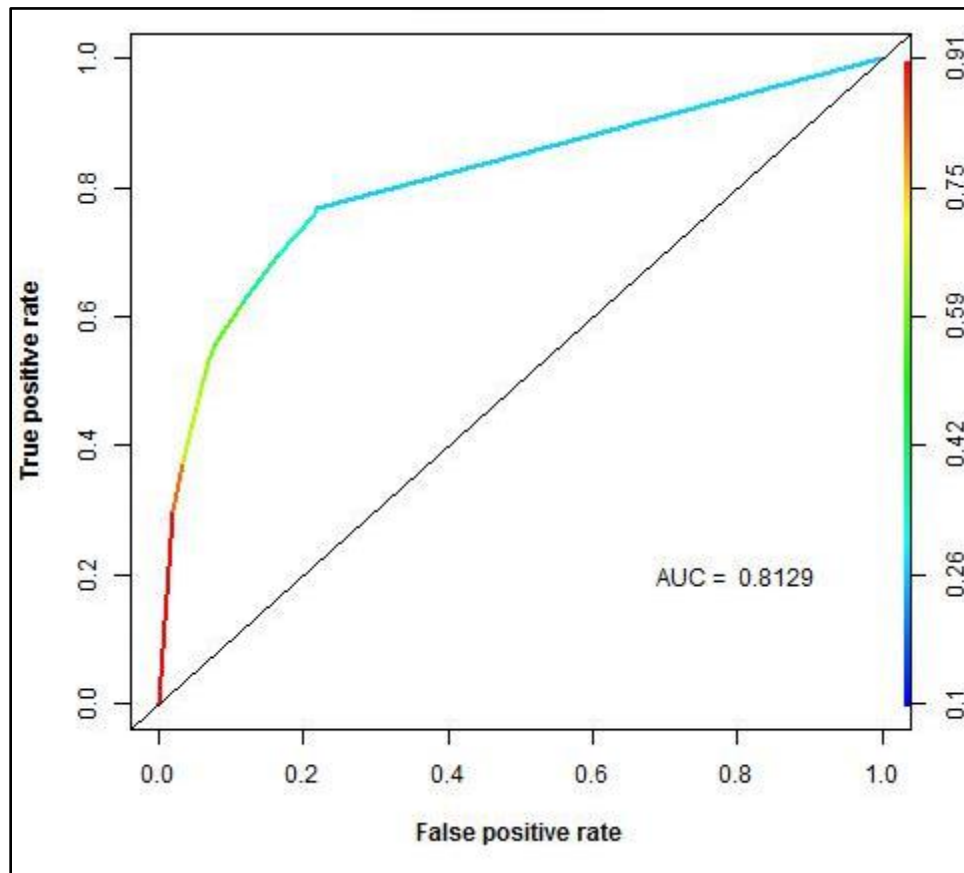


Figure 20. Attrition Classification by Classification Tree—ROC Curve (Test Dataset)

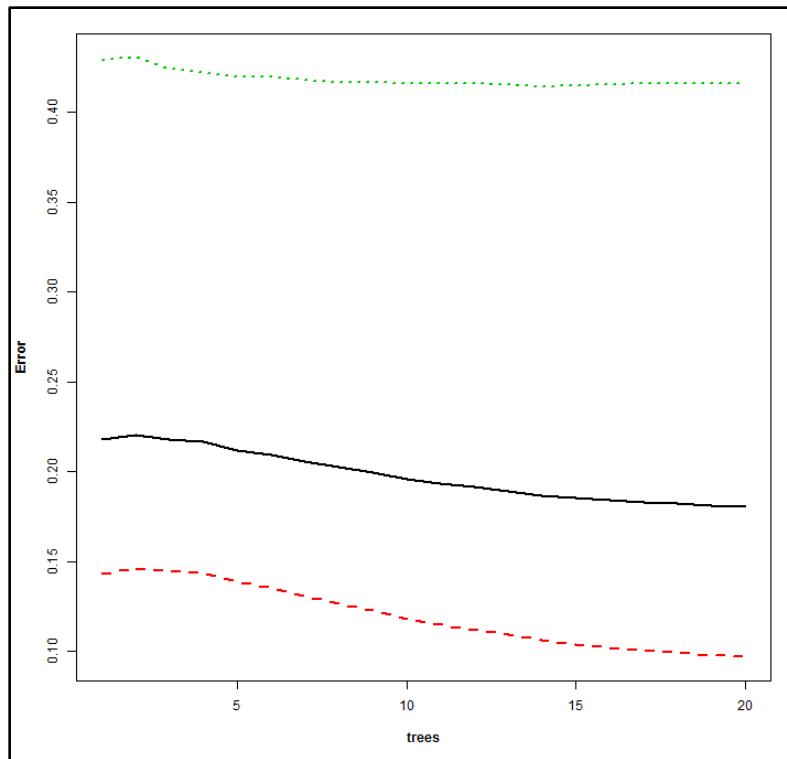
C. RANDOM FOREST

The final model we considered was a classification random forest (RF). The random forest method consists of building numerous trees by randomly sampling the data with replacement. The trees are typically larger than a sole classification tree constructed by an analyst. At each split node, a random sample of predictors is chosen to ultimately test each of the predictor variables and prevent correlation between the trees in the forest. Finally,

the new predictions are made by pushing each predictor variable through each tree and averaging the predictions throughout the forest.

Two modeling parameters must be tuned in the implementation of the random forest method. First, we need to identify the number of trees required in our forest to balance the computational complexity of the model while achieving the highest leaf purity. Second, we must select the number of predictor variable candidates considered at each split to capture the variability within the dataset without degrading the efficiency of the model. Random forest modeling does not allow for missing values within the data without the use of estimate-generating algorithms to impute missing values. Since our dataset was very large and our examination of missing values indicated randomness, we removed the observations. Our final training dataset consisted of 275,789 records for random forest generation.

Our tuning process initially ran the model with 250 trees; however, the computation took approximately 15 minutes. After analysis of the model error versus tree quantity plot, we found the number of trees required was much smaller. Ultimately, the prediction error stabilized with only 20 trees (Figure 21). The smaller forest greatly improved the speed of model construction, which allowed for an efficient process to vary the number of predictor candidates considered at each split. Cross-validation techniques showed the stabilization of model accuracy with five predictor candidates considered at each split.



The solid line represents the overall model error by the number of trees in the forest. Dotted and dashed lines depict the error of unique “yes” and “no” responses.

Figure 21. Random Forest Error by Tree Quantity

Unlike classification tree models, which offer an intuitive visualization tree output, random forests are often considered as “black box” solutions offering very little insight into the underlying data relationships. However, the variable importance chart provides a window into the noteworthy features within our data (Figure 22). Unsurprisingly, the most influential variables of the random forest model matched the variables identified in the classification tree splits. A soldier’s number of deployed days strongly influenced the prediction of first-term attrition while contract duration, marital status, unit type, and the soldier’s rank at enlistment had similar influential effects.

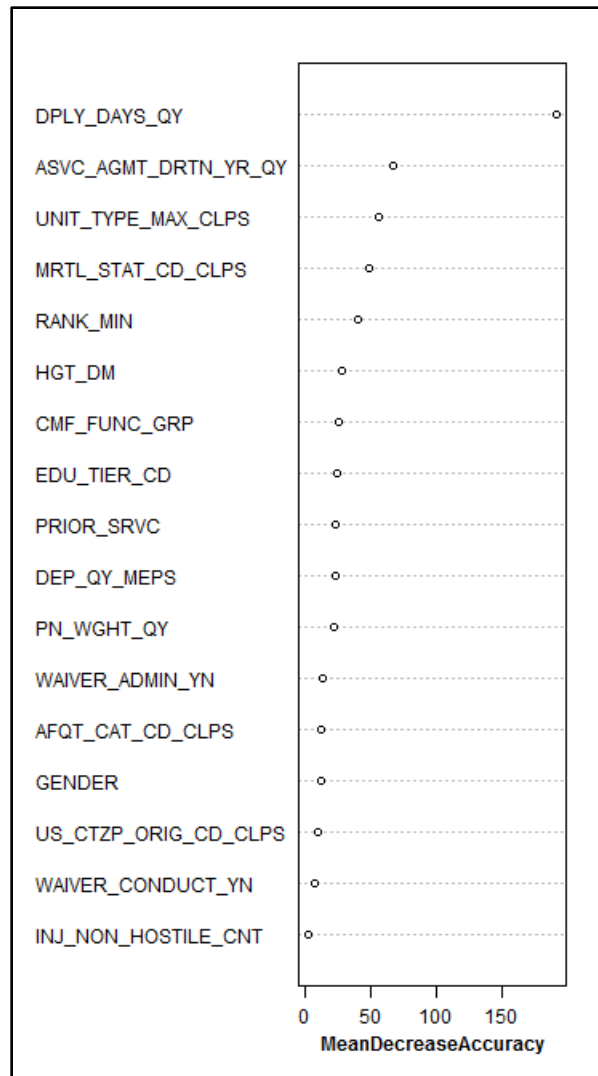


Figure 22. Random Forest Variable Importance

As expected, the predictive power of our random forest model was higher than with a single classification tree. The predictive ability of a random forest most often exceeds a single tree as multiple classification results are averaged for each set of predictor variables put through the forest. After predicting the observations in the test dataset, the superior ROC curve and AUC value of 0.8539 encouraged using the random forest instead of a classification tree in any future predictive tool (Figure 23) with a cut-off threshold near 0.63.

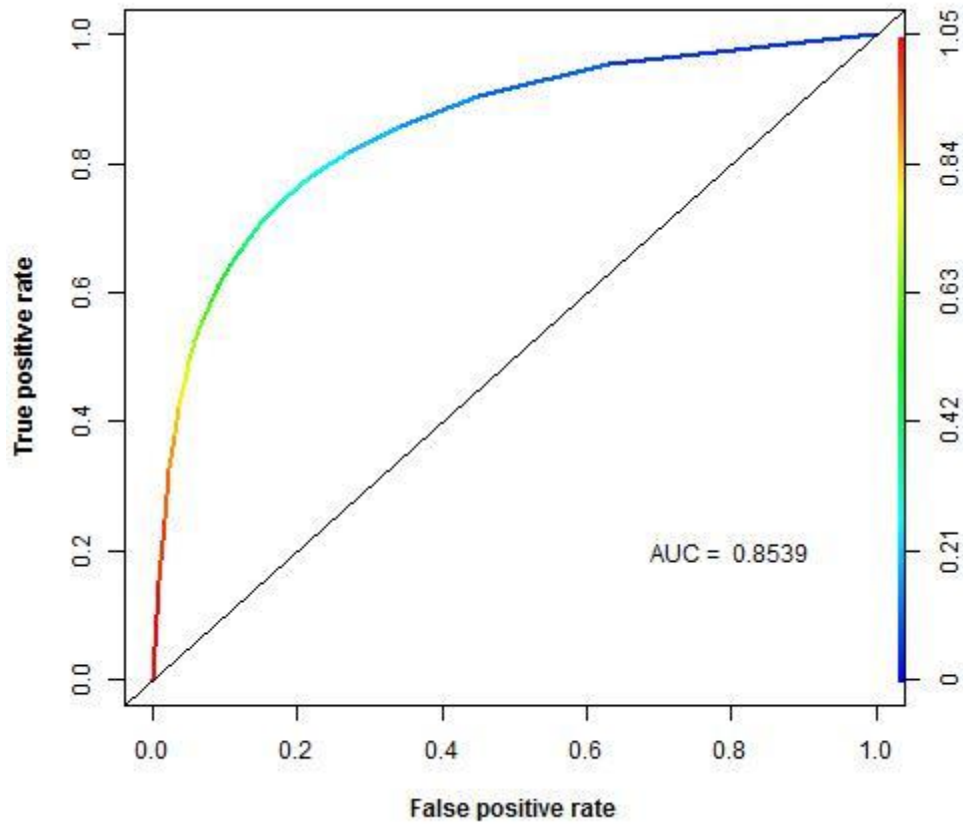


Figure 23. Attrition Classification by Random Forest—ROC Curve (Test Dataset)

D. MODEL SELECTION

We used the ROC curves and AUC calculations to determine the “best” model (Figure 24). Though the predictive power of our logistic regression model outperforms the other classification methods, it is noted that the construction of the random forest model was considerably faster.

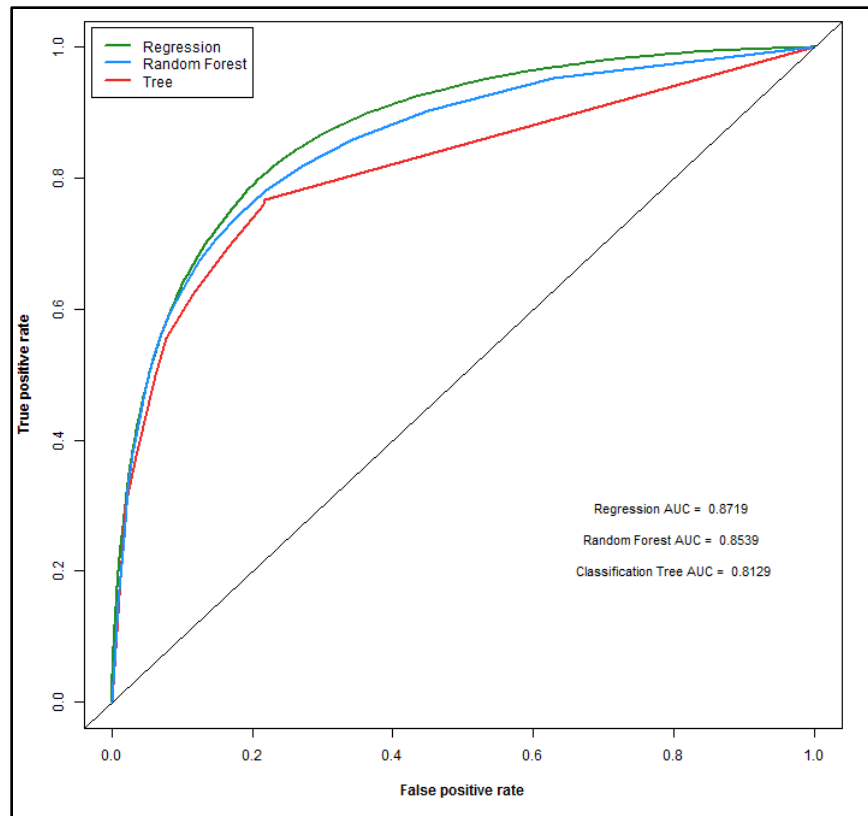


Figure 24. Model Comparison ROC Curves

The most influential predictor variables are consistent in each of our three models. While the logistic regression model utilizes the most predictors and factor levels, the classification tree model closely matches the accuracy performance with less than half of the predictors (Table 15). Though the random forest model considers all of the same predictors as the logistic regression model, the most influential random forest variables identified in the table provide the majority of information determining the accuracy of the model.

Table 15. Significant Variable Utilization by Model

Variable	Logistic Regression	Classification Tree	Random Forest
AFQT Category	x		
Citizenship Origination	x		
Contract Duration	x	x	x
Days Deployed	x	x	x

Variable	Logistic Regression	Classification Tree	Random Forest
Dependents	x		x
Education Tier	x	x	x
Gender	x		
Height (Enlistment)	x		x
Marital Status	x	x	x
Military Occupation Group	x		x
Non-hostile Injuries	x		
Prior Service	x		x
Rank (Enlistment)	x	x	x
Unit Type	x	x	x
Waiver (Admin)	x		
Waiver (Conduct)	x		
Weight (Enlistment)	x		x

In the interest of simplicity and accuracy, the random forest model provides ARD analysts increased flexibility as the data catalog grows within the PDE and more soldier data is made available. Most importantly, an analyst can extract a single tree from a random forest model and use it to provide stakeholders a visualization that is easy to understand and interpret. Additionally, ARD analysts could create simple flow charts that quickly allow Army leaders to assess the probability of a first-term soldier leaving prior to the completion of his or her contractual obligation.

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY

A summary of the key findings of our data preparation and analysis is presented. Recommendations are also provided for the use of our model by the ARD and potential areas of study in future research.

A. CONCLUSIONS

1. Data Preparation

The PDE is an extremely valuable resource for Army personnel analytics due to the consolidated data tables created from disparate data sources and the collaborative environment. However, the lack of comprehensive data definitions challenges users to fully understand the variables.

The universal application of separation codes continues to be a concern nearly four decades after the issue was first raised. Our research highlighted the potential errors in the sole use of the codes to study attrition and offered a new methodology for identifying whether a soldier successfully completed his or her contractual obligation by examining the historical records of each soldier.

Careful selection of predictor variables is critical for accurately depicting a soldier's demographic and administrative history. The time-related snapshot data complicates the analysis and naïve variable assignment can lead to unexpected bias.

2. Data Analysis

U.S. Army first-term attrition rates were steady from FY2005 to FY2010 and averaged 26%. Our research confirms both the overall attrition rates and proportional rate differences between factor levels as reported in previous research. While 24% of male soldiers failed to complete their initial contract, 38% of female soldiers were discharged early. Enlistees with high school diplomas are 25% more likely to complete their first term than recruits holding a GED. Soldiers from Louisiana, Arkansas, and West Virginia fail to complete their first term at a 10% higher rate than states with the lowest attrition rates

mostly found in the western United States. Single soldiers are also discharged early at a 10% higher rate than married or divorced enlistees.

The most influential predictor of first-term attrition is a soldier's deployment history. A soldier deployed more than three months has a 90% chance of completing the first term. Additionally, the length of a soldier's initial contract and marital status are key discriminators in predicting the probability of completion of the first term. Longer contract periods increase the risk of attrition while married or divorced soldiers are less likely than single soldiers to attrit. A soldier's gender and ASVAB score have very little influence in predicting attrition. Additionally, whether a soldier received an enlistment waiver or not has a negligible effect on his or her probability of completing the first term.

Obviously, we do not advocate for the deployment of more soldiers, 10-year contracts, or marriage requirements. However, identifying soldiers with the highest probability of failure may significantly inform Army leadership in prioritizing resources and focusing intervention strategies at those soldiers most in need of assistance.

B. RECOMMENDATIONS

1. Implementation

The accuracy rate of our predictive models was 83% and provided enough fidelity to warrant consideration of its use by Army planners. General attrition rate findings based on demographic predictor variables may help to inform force strength requirements, recruiting goals, and retention efforts. The flexible and repeatable nature of the random forest modeling technique provides analysts the ability to react quickly to changes in data availability and shifts in both policies and priorities.

Most importantly, this research provides ARD insight as the agency continues its efforts to improve soldier resiliency and, by extension, first-term attrition rates. Application of our predictive model to the administrative records of current enlistees could provide policy makers with probability estimates of all first-term soldiers and facilitate the creation of intervention programs and prioritized resource strategies built upon a quantitative foundation.

A web-based predictive tool available to Army leaders at the lowest unit level would allow human resource professionals or junior Non-Commissioned Officers to engage new soldiers during annual record reviews and monthly professional counseling sessions. Once the attrition probability assessment is completed for each soldier, the appropriate training, administrative actions, or other intervention strategies could be leveraged to best assist the soldier.

2. Future Research

We were unable to secure access to comprehensive enlistee medical data within the PDE for inclusion in our research. Once the data is available, we urge the addition of medical factors to our cohort dataset for further analysis. Though the broad categories of medical waiver data used in our research were too generalized to use for prediction, detailed medical information of new recruits may provide better models and a deeper understanding of attrition tendencies.

Our research and models considered predictor variables that evolve over the time a soldier has spent in the service. Predictive research tailored to Army accession policy analysis and the recruiting mission of USAREC requires the selection of only variables known at the time of a soldier's enlistment. We recommend the adaptation of our research to identify these variables and construct predictive models useful in understanding pre-enlistment recruits.

Finally, we recommend exploration of statistical modeling techniques not attempted in our research. Specifically, unsupervised models such as cluster analysis and principal components may assist researchers in better understanding the relationships in the data. Additionally, supervised techniques such as support vector machine or neural network design may provide higher accuracy rates in the predictive models.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. SEPARATION CODES

Table 16. Successful Separation Code Definitions

Separation Code	Separation Code Type	Description
1001	ISVC_SEP_CD	Expiration Term of Service (ETS)
1002	ISVC_SEP_CD	ETS
1003	ISVC_SEP_CD	ETS
1008	ISVC_SEP_CD	ETS
1040	ISVC_SEP_CD	Transfer to Officer Program
1042	ISVC_SEP_CD	Enrollment in a service academy
1050	ISVC_SEP_CD	Retirement
1052	ISVC_SEP_CD	Retirement
1100	ISVC_SEP_CD	Reenlistment
948	SPD_CD	Enrollment in a service academy
FCA	SPD_CD	Resignation
FCB	SPD_CD	Resignation
FHC	SPD_CD	Reenlistment
FND	SPD_CD	Resignation
JBK	SPD_CD	ETS
JBM	SPD_CD	ETS
JCC	SPD_CD	ETS
JGH	SPD_CD	ETS
KBK	SPD_CD	ETS
KBM	SPD_CD	ETS
KCA	SPD_CD	Early Release Program (Voluntary Separation)
KCB	SPD_CD	Early Release Program (Special Separation Benefit)
KCC	SPD_CD	Early Release Program (Employment)
KCF	SPD_CD	ETS
KGM	SPD_CD	Transfer to Officer Program
KGX	SPD_CD	Transfer to Officer Program
KHC	SPD_CD	Reenlistment
KND	SPD_CD	Resignation
LBK	SPD_CD	ETS
LBM	SPD_CD	ETS
LCC	SPD_CD	ETS
LGH	SPD_CD	ETS
MBK	SPD_CD	ETS
MBM	SPD_CD	ETS

Separation Code	Separation Code Type	Description
MCA	SPD_CD	Early Release Program (Voluntary Separation)
MCB	SPD_CD	Early Release Program (Special Separation Benefit)
MCC	SPD_CD	Early Release Program (Employment)
MCF	SPD_CD	ETS
MDM	SPD_CD	ETS
MHC	SPD_CD	Reenlistment
RBD	SPD_CD	Retirement
RBE	SPD_CD	Retirement
RCC	SPD_CD	Retirement
SBB	SPD_CD	Retirement
SBC	SPD_CD	Retirement
VBK	SPD_CD	Retirement

APPENDIX B. COHORT DATASET SUMMARY

Table 17. Numeric Feature Summary: Mean (Std Dev) by Fiscal Year of Enlistment

	2005		2006		2007		2008		2009		2010	
	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
Age (Enlistment)	21.26 (3.96)	20.93 (3.7)	21.39 (4.24)	21.08 (4.14)	21.67 (4.5)	21.43 (4.66)	21.6 (4.36)	21.15 (4.34)	21.81 (4.44)	21.16 (4.26)	21.9 (4.39)	21.23 (4.16)
ASVAB GT Score	266.22 (28.31)	264.61 (25.97)	263.38 (28.06)	261.91 (25.62)	262.78 (28.21)	261.95 (25.45)	263.13 (28.45)	262.58 (25.41)	265.57 (28.82)	263.99 (25.76)	265.1 (29.32)	263.53 (26.53)
Contract Duration	3.82 (0.9)	3.96 (0.92)	3.8 (0.9)	3.99 (0.98)	3.84 (0.91)	4.11 (1.02)	3.79 (0.93)	4.11 (1.05)	3.68 (0.9)	4.03 (1.07)	3.64 (0.89)	3.99 (1.07)
Days Deployed	396.92 (273.41)	78.45 (162.01)	361.14 (242.74)	72.77 (161.56)	327.29 (219.08)	79.26 (163.51)	293.83 (194.2)	81.6 (150.09)	246.46 (174.88)	73.11 (136.69)	206.68 (166.16)	66.37 (124.04)
Dependents	0.32 (0.82)	0.3 (0.78)	0.32 (0.84)	0.33 (0.85)	0.35 (0.88)	0.38 (0.92)	0.35 (0.89)	0.35 (0.9)	0.36 (0.89)	0.33 (0.87)	0.31 (0.82)	0.27 (0.77)
Deployments	2.5 (1.77)	0.57 (1.14)	2.27 (1.62)	0.5 (1.08)	2.17 (1.52)	0.55 (1.12)	1.97 (1.38)	0.58 (1.07)	1.6 (1.19)	0.51 (0.94)	1.29 (1.12)	0.47 (0.87)
Height at Enlistment	68.56 (3.22)	67.87 (3.51)	68.5 (3.22)	67.81 (3.54)	68.58 (3.24)	68.09 (3.48)	68.9 (3.23)	68.42 (3.44)	68.96 (3.2)	68.52 (3.43)	68.88 (3.24)	68.52 (3.45)
Hostile Injuries	0.04 (0.2)	0.02 (0.13)	0.03 (0.16)	0.01 (0.12)	0.02 (0.13)	0.01 (0.08)	0.02 (0.15)	0.01 (0.09)	0.03 (0.17)	0.01 (0.11)	0.02 (0.15)	0.01 (0.12)
Max Time-in- Grade	1.67 (0.88)	0.77 (0.73)	1.7 (0.96)	0.85 (0.77)	1.75 (0.89)	0.87 (0.81)	1.71 (0.88)	0.9 (0.81)	1.65 (0.86)	0.96 (0.84)	1.66 (0.85)	1.06 (0.85)
Non-Hostile Injuries	0.01 (0.07)	0.01 (0.09)	0.01 (0.07)	0.01 (0.08)	0.01 (0.08)	0.01 (0.07)	0.01 (0.09)	0.01 (0.09)	0.01 (0.1)	0.01 (0.1)	0.01 (0.1)	0.01 (0.12)
Weight at Enlistment	167.26 (30.91)	162.89 (32.69)	168.13 (31.87)	164.61 (33.45)	168.14 (32.21)	165.92 (33.85)	168.32 (31.99)	166.13 (34.14)	169.49 (32.03)	168.15 (34.94)	168.14 (30.97)	168.45 (33.92)

Table 18. Binary Feature Summary: Counts (Proportion of Attrition Category) by Fiscal Year of Enlistment

		2005		2006		2007		2008		2009		2010	
		No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
Citizenship Status (Enlistment)	0	48750 (97.11)	16793 (98.18)	54693 (97.3)	19024 (98.51)	50112 (97.16)	18729 (98.52)	49297 (96.91)	18926 (98.43)	44813 (96.85)	16496 (97.65)	49463 (96.1)	16153 (97.2)
	1	230 (0.46)	42 (0.25)	320 (0.57)	54 (0.28)	440 (0.85)	44 (0.23)	871 (1.71)	69 (0.36)	1327 (2.87)	157 (0.93)	1788 (3.47)	236 (1.42)
	NA	1219 (2.43)	270 (1.58)	1200 (2.13)	233 (1.21)	1023 (1.98)	238 (1.25)	702 (1.38)	233 (1.21)	132 (0.29)	240 (1.42)	220 (0.43)	229 (1.38)
Gender	0	6264 (12.48)	4541 (26.55)	7189 (12.79)	5112 (26.47)	6823 (13.23)	4503 (23.69)	6476 (12.73)	4426 (23.02)	5748 (12.42)	3676 (21.76)	6734 (13.0)	3542 (21.3)
	1	43935 (87.52)	12564 (73.45)	49024 (87.21)	14199 (73.53)	44752 (86.77)	14508 (76.31)	44394 (87.27)	14802 (76.98)	40524 (87.58)	13217 (78.24)	44737 (86.9)	13076 (78.6)
Prior Service	0	48394 (96.4)	15255 (89.18)	53861 (95.82)	16013 (82.92)	49893 (96.74)	16021 (84.27)	49174 (96.67)	16985 (88.33)	45593 (98.53)	15966 (94.51)	51005 (99.09)	16047 (96.56)
	1	1805 (3.6)	1850 (10.82)	2352 (4.18)	3298 (17.08)	1682 (3.26)	2990 (15.73)	1696 (3.33)	2243 (11.67)	679 (1.47)	927 (5.49)	466 (0.91)	571 (3.44)
Waiver (Medical)	0	46708 (93.05)	15917 (93.05)	52497 (93.39)	18012 (93.27)	47582 (92.26)	17561 (92.37)	46845 (92.09)	17675 (91.92)	42773 (92.44)	15616 (92.44)	47769 (92.81)	15473 (93.11)
	1	3491 (6.95)	1188 (6.95)	3716 (6.61)	1299 (6.73)	3993 (7.74)	1450 (7.63)	4025 (7.91)	1553 (8.08)	3499 (7.56)	1277 (7.56)	3702 (7.19)	1145 (6.89)
Waiver (Conduct)	0	46457 (92.55)	15803 (92.39)	50662 (90.13)	17264 (89.4)	45042 (87.33)	16332 (85.91)	44964 (88.39)	16856 (87.66)	42487 (91.82)	15462 (91.53)	49210 (95.61)	15816 (95.17)
	1	3742 (7.45)	1302 (7.61)	5551 (9.87)	2047 (10.6)	6533 (12.67)	2679 (14.09)	5906 (11.61)	2372 (12.34)	3785 (8.18)	1431 (8.47)	2261 (4.39)	802 (4.83)
Waiver (Admin)	0	47264 (94.15)	16520 (96.58)	52934 (94.17)	18517 (95.89)	47888 (92.85)	17921 (94.27)	47682 (93.73)	18222 (94.77)	43936 (94.95)	16161 (95.67)	49305 (95.79)	16020 (96.4)
	1	2935 (5.85)	585 (3.42)	3279 (5.83)	794 (4.11)	3687 (7.15)	1090 (5.73)	3188 (6.27)	1006 (5.23)	2336 (5.05)	732 (4.33)	2166 (4.21)	598 (3.6)
Waiver (Drug)	0	49739 (99.08)	16882 (98.7)	55571 (98.86)	18978 (98.28)	50772 (98.44)	18577 (97.72)	50104 (98.49)	18837 (97.97)	46074 (99.57)	16782 (99.34)	51470 (100)	16618 (100)
	1	460 (0.92)	223 (1.3)	642 (1.14)	333 (1.72)	803 (1.56)	434 (2.28)	766 (1.51)	391 (2.03)	198 (0.43)	111 (0.66)	1 (0)	0 (0)

Table 19. Categorical Feature Summary: Counts (Proportion of Attrition Category) by Fiscal Year of Enlistment

		2005		2006		2007		2008		2009		2010	
		No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
Unit Region (Max)	Midwest	2650 (5.28%)	1525 (8.92%)	3003 (5.34%)	1516 (7.85%)	2598 (5.04%)	1678 (8.83%)	2721 (5.35%)	2057 (10.7%)	3187 (6.89%)	1574 (9.32%)	2735 (5.31%)	1305 (7.85%)
	Northeast	2391 (4.76)	722 (4.22)	2148 (3.82)	574 (2.97)	2985 (5.79)	921 (4.84)	2843 (5.59)	696 (3.62)	2370 (5.12)	515 (3.05)	2619 (5.09)	614 (3.69)
	South	27543 (54.87)	10622 (62.1)	33342 (59.31)	13080 (67.73)	28941 (56.11)	12415 (65.3)	29157 (57.32)	12313 (64.04)	26116 (56.44)	10983 (65.02)	29195 (56.72)	10161 (61.14)
	Territory	6 (0.01)	0 (0)	5 (0.01)	1 (0.01)	7 (0.01)	0 (0)	1 (0)	0 (0)	10 (0.02)	0 (0)	4 (0.01)	4 (0.02)
	West	11299 (22.51)	2792 (16.32)	11146 (19.83)	2679 (13.87)	11777 (22.83)	2861 (15.05)	11308 (22.23)	3130 (16.28)	9918 (21.43)	2463 (14.58)	12888 (25.04)	3423 (20.6)
	NA	6310 (12.57)	1444 (8.44)	6569 (11.69)	1461 (7.57)	5267 (10.21)	1136 (5.98)	4840 (9.51)	1032 (5.37)	4671 (10.09)	1358 (8.04)	4030 (7.83)	1111 (6.69)
Home of Record Region	Midwest	10857 (21.63)	3545 (20.72)	12423 (22.1)	4168 (21.58)	11035 (21.4)	3938 (20.71)	9814 (19.29)	3749 (19.5)	9113 (19.69)	3387 (20.05)	9790 (19.02)	3195 (19.23)
	Northeast	6005 (11.96)	2166 (12.66)	6772 (12.05)	2175 (11.26)	6152 (11.93)	2279 (11.99)	5936 (11.67)	2182 (11.35)	5646 (12.2)	2034 (12.04)	6033 (11.72)	1863 (11.21)
	South	20821 (41.48)	7652 (44.74)	23713 (42.18)	9060 (46.92)	22609 (43.84)	9223 (48.51)	22250 (43.74)	9046 (47.05)	19502 (42.15)	7756 (45.91)	22152 (43.04)	7812 (47.01)
	Territory	688 (1.37)	124 (0.72)	699 (1.24)	133 (0.69)	651 (1.26)	121 (0.64)	634 (1.25)	127 (0.66)	580 (1.25)	132 (0.78)	699 (1.36)	139 (0.84)
	West	11223 (22.36)	3500 (20.46)	11678 (20.77)	3545 (18.36)	10650 (20.65)	3287 (17.29)	10751 (21.13)	3726 (19.38)	10921 (23.6)	3491 (20.67)	12338 (23.97)	3537 (21.28)
	NA	605 (1.21)	118 (0.69)	928 (1.65)	230 (1.19)	478 (0.93)	163 (0.86)	1485 (2.92)	398 (2.07)	510 (1.1)	93 (0.55)	459 (0.89)	72 (0.43)
Military Occupation Group	OPNS	23978 (47.77)	7947 (46.46)	27478 (48.88)	8551 (44.28)	22746 (44.1)	8149 (42.86)	23359 (45.92)	8509 (44.25)	21256 (45.94)	7818 (46.28)	24456 (47.51)	7954 (47.86)
	OS	5960 (11.87)	1432 (8.37)	4993 (8.88)	1431 (7.41)	5297 (10.27)	2017 (10.61)	5892 (11.58)	2541 (13.22)	6570 (14.2)	2556 (15.13)	6829 (13.27)	2386 (14.36)
	FS	19012 (37.87)	6508 (38.05)	21836 (38.85)	8388 (43.44)	21970 (42.6)	8206 (43.16)	20028 (39.37)	7565 (39.34)	16736 (36.17)	6084 (36.01)	18484 (35.91)	5996 (36.08)
	NA	1249 (2.49)	1218 (7.12)	1906 (3.39)	941 (4.87)	1562 (3.03)	639 (3.36)	1591 (3.13)	613 (3.19)	1710 (3.7)	435 (2.58)	1702 (3.31)	282 (1.7)

		2005		2006		2007		2008		2009		2010	
		No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
Rank (Max)	PV1	1199 (2.39)	5265 (30.78)	1386 (2.47)	5912 (30.61)	1397 (2.71)	6202 (32.62)	1375 (2.7)	5828 (30.31)	820 (1.77)	4612 (27.3)	625 (1.21)	3658 (22.01)
	PV2	965 (1.92)	3844 (22.47)	1227 (2.18)	4743 (24.56)	1143 (2.22)	4658 (24.5)	1062 (2.09)	4799 (24.96)	602 (1.3)	3862 (22.86)	430 (0.84)	3215 (19.35)
	PFC	2728 (5.43)	4197 (24.54)	3325 (5.92)	4754 (24.62)	3078 (5.97)	4217 (22.18)	2799 (5.5)	4399 (22.88)	1959 (4.23)	4178 (24.73)	1828 (3.55)	4683 (28.18)
	CPL	30882 (61.52)	3466 (20.26)	36453 (64.85)	3505 (18.15)	34656 (67.2)	3496 (18.39)	35272 (69.34)	3938 (20.48)	32532 (70.31)	3971 (23.51)	37214 (72.3)	4802 (28.9)
	SGT	12920 (25.74)	302 (1.77)	12823 (22.81)	334 (1.73)	10501 (20.36)	350 (1.84)	9629 (18.93)	220 (1.14)	9695 (20.95)	248 (1.47)	10706 (20.8)	255 (1.53)
	SSG	1505 (3)	31 (0.18)	999 (1.78)	63 (0.33)	800 (1.55)	88 (0.46)	733 (1.44)	44 (0.23)	664 (1.43)	22 (0.13)	668 (1.3)	5 (0.03)
Rank (Enlistment)	PV1	20438 (40.71)	9026 (52.77)	23077 (41.05)	10221 (52.93)	20691 (40.12)	9936 (52.26)	19672 (38.67)	9618 (50.02)	15577 (33.66)	7497 (44.38)	14859 (28.87)	6539 (39.35)
	PV2	11598 (23.1)	3979 (23.26)	14423 (25.66)	4813 (24.92)	13985 (27.12)	5085 (26.75)	15381 (30.24)	5716 (29.73)	14529 (31.4)	5308 (31.42)	15788 (30.67)	5000 (30.09)
	PFC	10394 (20.71)	2902 (16.97)	11519 (20.49)	3022 (15.65)	10807 (20.95)	2876 (15.13)	10680 (20.99)	3072 (15.98)	11050 (23.88)	3311 (19.6)	14713 (28.59)	4191 (25.22)
	CPL	6342 (12.63)	1053 (6.16)	5872 (10.45)	1020 (5.28)	5352 (10.38)	848 (4.46)	4647 (9.14)	707 (3.68)	4749 (10.26)	722 (4.27)	5694 (11.06)	859 (5.17)
	SGT	1130 (2.25)	119 (0.7)	1023 (1.82)	175 (0.91)	600 (1.16)	183 (0.96)	383 (0.75)	71 (0.37)	302 (0.65)	36 (0.21)	317 (0.62)	24 (0.14)
	SSG	297 (0.59)	26 (0.15)	299 (0.53)	60 (0.31)	140 (0.27)	83 (0.44)	107 (0.21)	44 (0.23)	65 (0.14)	19 (0.11)	100 (0.19)	5 (0.03)
Education Tier	1	42689 (85.04)	13065 (76.38)	43852 (78.01)	12788 (66.22)	39095 (75.8)	12070 (63.49)	39530 (77.71)	12839 (66.77)	40236 (86.96)	13342 (78.98)	49204 (95.6)	15789 (95.01)
	2	7086 (14.12)	3492 (20.42)	11828 (21.04)	6124 (31.71)	12000 (23.27)	6863 (36.1)	10490 (20.62)	6077 (31.6)	4791 (10.35)	2848 (16.86)	1843 (3.58)	743 (4.47)
	3	47 (0.09)	13 (0.08)	83 (0.15)	40 (0.21)	77 (0.15)	52 (0.27)	205 (0.4)	175 (0.91)	856 (1.85)	669 (3.96)	110 (0.21)	59 (0.36)
	NA	377 (0.75)	535 (3.13)	450 (0.8)	359 (1.86)	403 (0.78)	26 (0.14)	645 (1.27)	137 (0.71)	389 (0.84)	34 (0.2)	314 (0.61)	27 (0.16)

		2005		2006		2007		2008		2009		2010	
		No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
AFQT Category	1	3545 (7.06)	803 (4.69)	3251 (5.78)	729 (3.78)	2967 (5.75)	648 (3.41)	2979 (5.86)	643 (3.34)	3266 (7.06)	692 (4.1)	3989 (7.75)	832 (5.01)
	2	17492 (34.85)	5587 (32.66)	18129 (32.25)	5703 (29.53)	16291 (31.59)	5705 (30.01)	16081 (31.61)	5976 (31.08)	15874 (34.31)	5573 (32.99)	17666 (34.32)	5451 (32.8)
	3A	12444 (24.79)	4854 (28.38)	12949 (23.04)	5050 (26.15)	11928 (23.13)	5181 (27.25)	11992 (23.57)	5418 (28.18)	11486 (24.82)	5084 (30.1)	11388 (22.13)	4488 (27.01)
	3B	14404 (28.69)	5148 (30.1)	19013 (33.82)	7142 (36.98)	17751 (34.42)	6790 (35.72)	17034 (33.49)	6487 (33.74)	14429 (31.18)	5265 (31.17)	17746 (34.48)	5739 (34.53)
	4A	2066 (4.12)	670 (3.92)	2453 (4.36)	641 (3.32)	2394 (4.64)	656 (3.45)	2038 (4.01)	528 (2.75)	841 (1.82)	241 (1.43)	407 (0.79)	95 (0.57)
	4Bplus	84 (0.17)	20 (0.12)	90 (0.16)	16 (0.08)	54 (0.1)	19 (0.1)	69 (0.14)	18 (0.09)	58 (0.13)	24 (0.14)	46 (0.09)	10 (0.06)
	NA	164 (0.33)	23 (0.13)	328 (0.58)	30 (0.16)	190 (0.37)	12 (0.06)	677 (1.33)	158 (0.82)	318 (0.69)	14 (0.08)	229 (0.44)	3 (0.02)
Citizenship Origination	A	46467 (92.57)	16358 (95.63)	52133 (92.74)	18560 (96.11)	47810 (92.7)	18223 (95.86)	46837 (92.07)	18374 (95.56)	41983 (90.73)	15864 (93.91)	45917 (89.21)	15458 (93.02)
	N	1370 (2.73)	217 (1.27)	1438 (2.56)	216 (1.12)	1294 (2.51)	222 (1.17)	1387 (2.73)	238 (1.24)	1578 (3.41)	258 (1.53)	2209 (4.29)	346 (2.08)
	C	926 (1.84)	218 (1.27)	1138 (2.02)	252 (1.3)	1082 (2.1)	293 (1.54)	1181 (2.32)	329 (1.71)	1257 (2.72)	395 (2.34)	1341 (2.61)	366 (2.2)
	NA	1436 (2.86)	312 (1.82)	1504 (2.68)	283 (1.47)	1389 (2.69)	273 (1.44)	1465 (2.88)	287 (1.49)	1454 (3.14)	376 (2.23)	2004 (3.89)	448 (2.7)
Education Level (Enlistment)	HS	42532 (84.73)	14597 (85.34)	45021 (80.09)	15935 (82.52)	46375 (89.92)	17781 (93.53)	45039 (88.54)	17867 (92.92)	39732 (85.87)	15552 (92.06)	42908 (83.36)	14824 (89.2)
	CLG	2458 (4.9)	836 (4.89)	2399 (4.27)	732 (3.79)	2237 (4.34)	760 (4)	2259 (4.44)	772 (4.01)	2539 (5.49)	804 (4.76)	3476 (6.75)	1094 (6.58)
	BAC	2368 (4.72)	415 (2.43)	2315 (4.12)	404 (2.09)	2611 (5.06)	424 (2.23)	2689 (5.29)	388 (2.02)	3330 (7.2)	453 (2.68)	4393 (8.53)	601 (3.62)
	GRAD	145 (0.29)	23 (0.13)	156 (0.28)	25 (0.13)	197 (0.38)	36 (0.19)	201 (0.4)	37 (0.19)	292 (0.63)	57 (0.34)	447 (0.87)	88 (0.53)
	NA	2696 (5.37)	1234 (7.21)	6322 (11.25)	2215 (11.47)	155 (0.3)	10 (0.05)	682 (1.34)	164 (0.85)	379 (0.82)	27 (0.16)	247 (0.48)	11 (0.07)

		2005		2006		2007		2008		2009		2010	
		No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
Education Level (Max)	HS	44341 (88.33)	15285 (89.36)	50046 (89.03)	17694 (91.63)	45499 (88.22)	17778 (93.51)	44226 (86.94)	17892 (93.05)	38696 (83.63)	15516 (91.85)	41216 (80.08)	14745 (88.73)
	CLG	2773 (5.52)	828 (4.84)	2913 (5.18)	770 (3.99)	2866 (5.56)	750 (3.95)	3095 (6.08)	781 (4.06)	3580 (7.74)	847 (5.01)	5175 (10.05)	1161 (6.99)
	BAC	2520 (5.02)	432 (2.53)	2611 (4.64)	454 (2.35)	2583 (5.01)	421 (2.21)	2681 (5.27)	377 (1.96)	3289 (7.11)	438 (2.59)	4270 (8.3)	591 (3.56)
	GRAD	188 (0.37)	25 (0.15)	193 (0.34)	34 (0.18)	224 (0.43)	36 (0.19)	223 (0.44)	41 (0.21)	318 (0.69)	58 (0.34)	496 (0.96)	94 (0.57)
	NA	377 (0.75)	535 (3.13)	450 (0.8)	359 (1.86)	403 (0.78)	26 (0.14)	645 (1.27)	137 (0.71)	389 (0.84)	34 (0.2)	314 (0.61)	27 (0.16)
		2051 (4.09)	481 (2.81)	2217 (3.94)	592 (3.07)	2069 (4.01)	598 (3.15)	1997 (3.93)	645 (3.35)	1621 (3.5)	472 (2.79)	1778 (3.45)	414 (2.49)
Marital Status (Max)	D	24078 (47.97)	5238 (30.62)	27500 (48.92)	6094 (31.56)	26613 (51.6)	6260 (32.93)	25818 (50.75)	6283 (32.68)	23459 (50.7)	5610 (33.21)	25611 (49.76)	5736 (34.52)
	M	24033 (47.88)	11367 (66.45)	26455 (47.06)	12621 (65.36)	22867 (44.34)	12139 (63.85)	23031 (45.27)	12281 (63.87)	21168 (45.75)	10805 (63.96)	24061 (46.75)	10457 (62.93)
	N	37 (0.07)	19 (0.11)	41 (0.07)	4 (0.02)	26 (0.05)	14 (0.07)	24 (0.05)	19 (0.1)	24 (0.05)	6 (0.04)	21 (0.04)	11 (0.07)
	OTHER	8267 (16.47)	7896 (46.16)	8608 (15.31)	9108 (47.16)	6617 (12.83)	8722 (45.88)	5726 (11.26)	8115 (42.2)	5277 (11.4)	6738 (39.89)	5719 (11.11)	5409 (32.55)
Unit Type (Max)	TDA	41673 (83.02)	8967 (52.42)	47387 (84.3)	10065 (52.12)	44768 (86.8)	10053 (52.88)	44882 (88.23)	10859 (56.47)	40849 (88.28)	10083 (59.69)	45610 (88.61)	11141 (67.04)
	MTOE	222 (0.44)	67 (0.39)	199 (0.35)	61 (0.32)	181 (0.35)	139 (0.73)	246 (0.48)	199 (1.03)	113 (0.24)	23 (0.14)	132 (0.26)	25 (0.15)
	MULTI	37 (0.07)	175 (1.02)	19 (0.03)	77 (0.4)	9 (0.02)	97 (0.51)	16 (0.03)	55 (0.29)	33 (0.07)	49 (0.29)	10 (0.02)	43 (0.26)
	NA												

		2005		2006		2007		2008		2009		2010	
		No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit	No Attrit	Attrit
Military Occupation	LD	439 (0.87)	105 (0.61)	513 (0.91)	112 (0.58)	577 (1.12)	137 (0.72)	514 (1.01)	150 (0.78)	629 (1.36)	183 (1.08)	585 (1.14)	158 (0.95)
	11	10048 (20.02)	3562 (20.82)	10657 (18.96)	3328 (17.23)	8037 (15.58)	2731 (14.37)	8543 (16.79)	2849 (14.82)	8586 (18.56)	2972 (17.59)	9725 (18.89)	3179 (19.13)
	12	2156 (4.29)	645 (3.77)	3046 (5.42)	903 (4.68)	2211 (4.29)	802 (4.22)	3606 (7.09)	1288 (6.7)	2881 (6.23)	1027 (6.08)	2949 (5.73)	887 (5.34)
	13	2902 (5.78)	792 (4.63)	4265 (7.59)	1267 (6.56)	3471 (6.73)	1189 (6.25)	3291 (6.47)	949 (4.94)	2658 (5.74)	857 (5.07)	3501 (6.8)	1013 (6.1)
	14	506 (1.01)	222 (1.3)	1209 (2.15)	623 (3.23)	1041 (2.02)	482 (2.54)	1102 (2.17)	398 (2.07)	573 (1.24)	221 (1.31)	1311 (2.55)	409 (2.46)
	15	2148 (4.28)	616 (3.6)	2176 (3.87)	659 (3.41)	1841 (3.57)	680 (3.58)	1651 (3.25)	722 (3.75)	1908 (4.12)	804 (4.76)	2051 (3.98)	846 (5.09)
	18	498 (0.99)	77 (0.45)	320 (0.57)	2 (0.01)	204 (0.4)	2 (0.01)	231 (0.45)	2 (0.01)	258 (0.56)	4 (0.02)	325 (0.63)	2 (0.01)
	19	2881 (5.74)	870 (5.09)	2694 (4.79)	652 (3.38)	2575 (4.99)	807 (4.24)	2139 (4.2)	740 (3.85)	1935 (4.18)	825 (4.88)	2658 (5.16)	828 (4.98)
	25	3081 (6.14)	1158 (6.77)	2929 (5.21)	1136 (5.88)	2959 (5.74)	1328 (6.99)	3712 (7.3)	1807 (9.4)	4131 (8.93)	1799 (10.65)	3612 (7.02)	1417 (8.53)
	31	1971 (3.93)	897 (5.24)	1826 (3.25)	651 (3.37)	2373 (4.6)	1205 (6.34)	2065 (4.06)	1270 (6.6)	1766 (3.82)	926 (5.48)	1101 (2.14)	559 (3.36)
	35	2879 (5.74)	274 (1.6)	2064 (3.67)	295 (1.53)	2338 (4.53)	689 (3.62)	2177 (4.28)	733 (3.81)	2432 (5.26)	757 (4.48)	3194 (6.21)	969 (5.83)
	42	920 (1.83)	282 (1.65)	1291 (2.3)	406 (2.1)	1864 (3.61)	654 (3.44)	1411 (2.77)	567 (2.95)	794 (1.72)	245 (1.45)	726 (1.41)	225 (1.35)
	68	3465 (6.9)	633 (3.7)	4005 (7.12)	1531 (7.93)	3824 (7.41)	1689 (8.88)	3274 (6.44)	1516 (7.88)	3285 (7.1)	1278 (7.57)	3521 (6.84)	1138 (6.85)
	74	759 (1.51)	254 (1.48)	1141 (2.03)	452 (2.34)	871 (1.69)	236 (1.24)	647 (1.27)	272 (1.41)	573 (1.24)	166 (0.98)	742 (1.44)	227 (1.37)
	88	2558 (5.1)	821 (4.8)	3362 (5.98)	1256 (6.5)	3198 (6.2)	1250 (6.58)	2726 (5.36)	984 (5.12)	2075 (4.48)	779 (4.61)	2308 (4.48)	774 (4.66)
	91	6225 (12.4)	2531 (14.8)	6513 (11.59)	2534 (13.12)	6350 (12.31)	2112 (11.11)	5381 (10.58)	1897 (9.87)	5415 (11.7)	1851 (10.96)	6405 (12.44)	1989 (11.97)
	92	5515 (10.99)	2149 (12.56)	6297 (11.2)	2563 (13.27)	6279 (12.17)	2379 (12.51)	6813 (13.39)	2471 (12.85)	4663 (10.08)	1764 (10.44)	5055 (9.82)	1716 (10.33)
	NA	1248 (2.49)	1217 (7.11)	1905 (3.39)	941 (4.87)	1562 (3.03)	639 (3.36)	1587 (3.12)	613 (3.19)	1710 (3.7)	435 (2.58)	1702 (3.31)	282 (1.7)

APPENDIX C. UNIVARIATE MODEL RESULTS

Table 20. Univariate Summary of Binary/Categorical Features:
Count and Proportion by Attrition Category

	Variable	No Attrit	Attrit	Total	p-value
Fiscal Year (Enlistment)	2005	40187 (74.64 %)	13657 (25.36 %)	53844	Ref
	2006	45044 (74.55)	15376 (25.45)	60420	0.7432
	2007	41314 (73.16)	15155 (26.84)	56469	< 0.001
	2008	40704 (72.58)	15375 (27.42)	56079	< 0.001
	2009	37082 (73.38)	13450 (26.62)	50532	< 0.001
	2010	41123 (75.49)	13349 (24.51)	54472	0.0011
Prior Service	0 - "No"	238527 (75.61)	76942 (24.39)	315469	Ref
	1 - "Yes"	6927 (42.37)	9420 (57.63)	16347	< 0.001
Unit Region (Max)	Midwest	13570 (63.87)	7676 (36.13)	21246	Ref
	Northeast	12216 (79.27)	3195 (20.73)	15411	< 0.001
	South	139473 (71.52)	55534 (28.48)	195007	< 0.001
	Territory	23 (88.46)	3 (11.54)	26	0.0169
	West	54787 (79.75)	13912 (20.25)	68699	< 0.001
Home of Record Region	Midwest	50235 (73.99)	17657 (26.01)	67892	Ref
	Northeast	29300 (74.28)	10143 (25.72)	39443	0.2926
	South	105092 (72.29)	40279 (27.71)	145371	< 0.001
	Territory	3182 (84.38)	589 (15.62)	3771	< 0.001
	West	54002 (76.24)	16827 (23.76)	70829	< 0.001
Military Occupation Group (Max)	Operations	114786 (74.59)	39095 (25.41)	153881	Ref
	Operational Support	28356 (74.26)	9827 (25.74)	38183	0.1845
	Force Sustainment	94559 (73.48)	34123 (26.52)	128682	< 0.001

	Variable	No Attrit	Attrit	Total	p-value
Gender	0 - Female	31355 (60.42)	20542 (39.58)	51897	Ref
	1 - Male	214099 (76.49)	65820 (23.51)	279919	< 0.001
Citizenship Status	0 - U.S. Citizen	237869 (73.73)	84747 (26.27)	322616	Ref
	1- Non-Citizen	3981 (89.3)	477 (10.7)	4458	< 0.001
Rank (Max)	PV1	5457 (17.81)	25176 (82.19)	30633	Ref
	PV2	4254 (17.56)	19972 (82.44)	24226	< 0.001
	PFC	12590 (37.29)	21176 (62.71)	33766	< 0.001
	CPL	165848 (89.99)	18458 (10.01)	184306	< 0.001
	SGT	53011 (97.47)	1376 (2.53)	54387	< 0.001
	SSG	4294 (95.46)	204 (4.54)	4498	< 0.001
Rank (Enlistment)	PV1	91659 (68.49)	42169 (31.51)	133828	Ref
	PV2	68568 (74.17)	23874 (25.83)	92442	< 0.001
	PFC	55203 (78.08)	15497 (21.92)	70700	< 0.001
	CPL	26230 (86.4)	4130 (13.6)	30360	< 0.001
	SGT	2968 (85.56)	501 (14.44)	3469	0.1734
	SSG	826 (81.22)	191 (18.78)	1017	< 0.001
Waiver (Medical)	0 - "No"	227569 (73.96)	80115 (26.04)	307684	Ref
	1 - "Yes"	17885 (74.11)	6247 (25.89)	24132	0.606
Waiver (Conduct)	0 - "No"	223280 (74.14)	77889 (25.86)	301169	Ref
	1 - "Yes"	22174 (72.35)	8473 (27.65)	30647	< 0.001
Waiver (Admin)	0 - "No"	231378 (73.71)	82513 (26.29)	313891	Ref
	1 - "Yes"	14076 (78.53)	3849 (21.47)	17925	< 0.001
Waiver (Drug)	0 - "No"	243147 (74.05)	85196 (25.95)	328343	Ref
	1 - "Yes"	2307 (66.43)	1166 (33.57)	3473	< 0.001

	Variable	No Attrit	Attrit	Total	p-value
Education Tier	1 - HS Diploma	203711 (76.16)	63762 (23.84)	267473	Ref
	2 - GED	38563 (64.84)	20913 (35.16)	59476	< 0.001
	3 - Other	1097 (57.59)	808 (42.41)	1905	< 0.001
AFQT Category	1	15966 (82.08)	3485 (17.92)	19451	Ref
	2	81260 (74.95)	27159 (25.05)	108419	< 0.001
	3A	57981 (70.69)	24045 (29.31)	82026	< 0.001
	3B	80247 (73.36)	29146 (26.64)	109393	< 0.001
	4A	8113 (78.22)	2259 (21.78)	10372	< 0.001
	4Bplus	318 (79.9)	80 (20.1)	398	0.0234
Citizenship Origination	A - Born in U.S.	225055 (73.25)	82179 (26.75)	307234	Ref
	N - Born outside U.S.	7450 (86.6)	1153 (13.4)	8603	< 0.001
	C - Naturalized	5539 (79.07)	1466 (20.93)	7005	< 0.001
Education Level (Enlistment)	HS - High School	209433 (73.09)	77106 (26.91)	286539	Ref
	CLG - Some College	12261 (75.46)	3987 (24.54)	16248	< 0.001
	BAC - Baccalaureate	14140 (86.69)	2171 (13.31)	16311	< 0.001
	GRAD - Graduate	1124 (84.26)	210 (15.74)	1334	< 0.001
Education Level (Max)	HS - High School	211427 (72.8)	78975 (27.2)	290402	Ref
	CLG - Some College	16319 (79.92)	4099 (20.08)	20418	< 0.001
	BAC - Baccalaureate	14345 (86.81)	2179 (13.19)	16524	< 0.001
	GRAD - Graduate	1280 (84.77)	230 (15.23)	1510	< 0.001
Marital Status (Max)	D - Divorced	9453 (78.59)	2576 (21.41)	12029	Ref
	M - Married	122658 (81.32)	28174 (18.68)	150832	< 0.001
	N - Never Married	113196 (67.08)	55556 (32.92)	168752	< 0.001
	Other	147 (72.41)	56 (27.59)	203	0.0347

	Variable	No Attrit	Attrit	Total	p-value
Unit Type (Max)	TDA	32112 (46.63)	36757 (53.37)	68869	Ref
	MTOE	212399 (81.31)	48824 (18.69)	261223	< 0.001
	MULTI	851 (68.24)	396 (31.76)	1247	< 0.001
Military Occupation (Max)	LD	2618 (79.48)	676 (20.52)	3294	Ref
	11	44576 (75)	14858 (25)	59434	< 0.001
	12	13543 (75.5)	4394 (24.5)	17937	< 0.001
	13	16091 (76.91)	4830 (23.09)	20921	0.0011
	14	4593 (70.79)	1895 (29.21)	6488	< 0.001
	15	9343 (72.98)	3460 (27.02)	12803	< 0.001
Military Occupation (Max)	18	1466 (95.19)	74 (4.81)	1540	< 0.001
	19	11932 (75.73)	3825 (24.27)	15757	< 0.001
	25	16331 (70.53)	6824 (29.47)	23155	< 0.001
	31	8906 (66.91)	4405 (33.09)	13311	< 0.001
	35	12001 (79.99)	3002 (20.01)	15003	0.506
	42	5658 (74.42)	1945 (25.58)	7603	< 0.001
	68	17064 (73.42)	6178 (26.58)	23242	< 0.001
	74	3796 (74.62)	1291 (25.38)	5087	< 0.001
	88	12922 (73.5)	4660 (26.5)	17582	< 0.001
	91	29140 (73.82)	10335 (26.18)	39475	< 0.001
	92	27725 (72.73)	10394 (27.27)	38119	< 0.001

LIST OF REFERENCES

- Baldor, L. C. (2018, April 22). Army lowers recruiting goal; more soldiers staying on. The Associated Press. Retrieved from <https://www.armytimes.com/news/your-army/2018/04/22/army-lowers-2017-recruiting-goal-more-soldiers-staying-on/>
- Buddin, R. (1985). *Analysis of early military attrition behavior*. (Report No. RB-2001-2). Retrieved from https://www.rand.org/pubs/research_briefs/RB2001-2.html
- Buddin, R. J. (2005). *Success of first-term soldiers: The effects of recruiting practices and recruit characteristics*. Santa Monica, CA: RAND Corporation.
- Department of the Army. (2013). *Force Development and Documentation* (AR 71-32). Washington, DC: Author. Retrieved from <https://armypubs.army.mil/ProductMaps/PubForm/AR.aspx>
- Department of the Army. (2014). *Commissioned Officer Professional Development and Career Management* (DA PAM 600-3). Washington, DC: Author. Retrieved from <https://armypubs.army.mil/ProductMaps/PubForm/PAM.aspx>
- Department of the Army. (2016). *Regular Army and Reserve Component Enlistment Program* (AR 601-210). Washington, DC: Author. Retrieved from <https://armypubs.army.mil/ProductMaps/PubForm/AR.aspx>
- Faraway, J. (2016). *Extending the linear model with r* (2nd ed.). Boca Raton, FL: Taylor & Francis Group, LLC.
- Farrell, B. S. (2017). *Military personnel: Improvements needed in the management of the enlistee medical early separation and enlistment information* (GAO-17-527). Washington, DC: Government Accountability Office.
- Government Accountability Office. (1997). *Military attrition: DOD could save millions by better screening enlisted personnel* (GAO-97-39). Washington, DC: Government Accountability Office.
- Government Accountability Office. (1998). *Military attrition: Better data, coupled with policy changes, could help the services reduce early separations* (GAO/NSIAD-98-213). Washington, DC: Government Accountability Office.
- Jensen, D. C. (2016) *Supplemental information: Person-Event Data Environment*. Unpublished technical report.
- Lin, M., Lucas Jr., H. C., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4), 906-917. <https://doi.org/10.1287/isre.2013.0480>

- Martin, T. J. (1995). *Who stays, who leaves? An analysis of first-term Army attrition* (Doctoral dissertation). Retrieved from https://www.rand.org/content/dam/rand/pubs/rgs_dissertations/2006/RGSD114.pdf
- Military attrition: DOD could save millions by better screening enlisted personnel*, 105th Cong. (1997) (testimony of Mark Gebicke, Director of Military Operations and Capabilities Issues, National Security and International Affairs Division).
- Military attrition: DOD needs to better analyze reasons for separation and improve recruiting systems*, 105th Cong. (1998) (testimony of Mark Gebicke, Director of Military Operations and Capabilities Issues, National Security and International Affairs Division, Government Accountability Office).
- Military attrition: DOD needs to follow through on actions initiated to reduce early separations*, 106th Cong. (1999) (testimony of Mark Gebicke, Director of Military Operations and Capabilities Issues, National Security and International Affairs Division, Government Accountability Office).
- Military personnel: First-term recruiting and attrition continue to require focused attention*, 106th Cong. (2000) (testimony of Norman Rabkin, Director of National Security Preparedness Issues, National Security and International Affairs Division, Government Accountability Office).
- Smith, A. D. (2017). *Predicting ranger assessment and selection program 1 success and optimizing class composition* (Master's Thesis). Retrieved from <https://calhoun.nps.edu/handle/10945/55538>
- Syeed, S. & Whiteaker, C. (2018, March 13). Low U.S. unemployment is making Army recruiting harder. *Bloomberg*. Retrieved from <https://www.bloomberg.com/news/articles/2018-03-13/trump-s-army-buildup-confronts-headwinds-of-tight-labor-market>
- Zhang, Z. (2016). Model building strategy for logistic regression: purposeful selection. *Annals of Translational Medicine*, (4)2, 111. <https://doi.org/10.21037/atm.2016.02.15>

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