



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**FORECASTING ARMY RESERVE OFFICER TRAINING
CORPS COMMISSIONS USING MACHINE LEARNING
TECHNIQUES**

by

Daniel R. Hudalla

June 2019

Thesis Advisor:
Second Reader:

Robert A. Koyak
Robert L. Bassett

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 2019	3. REPORT TYPE AND DATES COVERED Master's thesis	
4. TITLE AND SUBTITLE FORECASTING ARMY RESERVE OFFICER TRAINING CORPS COMMISSIONS USING MACHINE LEARNING TECHNIQUES			5. FUNDING NUMBERS	
6. AUTHOR(S) Daniel R. Hudalla				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (maximum 200 words) The U.S. Army Cadet Command (USACC) produces about two-thirds of the Army officer corps in any given year. The command currently forecasts commissions 24 months before the completion of a fiscal year despite limited ability to influence production within this timeframe. Consequently, USACC must make recruiting decisions to shape its cohorts at least six months before current forecasts begin. This study explores the use of statistical machine learning models to forecast the number of currently enrolled cadets who will commission in a cohort nearly three years out. The developed forecasts can be used to determine the number of new cadets USACC must recruit to accomplish future recruiting missions. We find that a machine learning model can identify predictors that increase or decrease the likelihood of commissioning, and offer insight related to scholarship and contracting policies based on model outputs.				
14. SUBJECT TERMS forecasting, attrition, Cadets, ROTC, USACC, commission, machine learning, random forest			15. NUMBER OF PAGES 87	
			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.

**FORECASTING ARMY RESERVE OFFICER TRAINING CORPS
COMMISSIONS USING MACHINE LEARNING TECHNIQUES**

Daniel R. Hudalla
Major, United States Army
BBA, University of Georgia, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
June 2019**

Approved by: Robert A. Koyak
Advisor

Robert L. Bassett
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The U.S. Army Cadet Command (USACC) produces about two-thirds of the Army officer corps in any given year. The command currently forecasts commissions 24 months before the completion of a fiscal year despite limited ability to influence production within this timeframe. Consequently, USACC must make recruiting decisions to shape its cohorts at least six months before current forecasts begin. This study explores the use of statistical machine learning models to forecast the number of currently enrolled cadets who will commission in a cohort nearly three years out. The developed forecasts can be used to determine the number of new cadets USACC must recruit to accomplish future recruiting missions. We find that a machine learning model can identify predictors that increase or decrease the likelihood of commissioning, and offer insight related to scholarship and contracting policies based on model outputs.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
A.	MOTIVATION AND OBJECTIVES	1
B.	RESEARCH FOCUS.....	2
C.	THESIS ORGANIZATION.....	2
II.	BACKGROUND	5
A.	U.S. ARMY CADET COMMAND.....	5
B.	PROGRESSION IN AROTC	6
1.	Military Science Levels.....	7
2.	The Basic Course.....	8
3.	The Advanced Course.....	9
4.	Military Junior Colleges and the Early Commissioning Program	10
5.	Green to Gold Program.....	10
6.	Simultaneous Membership Program	11
C.	FORECASTING CHALLENGES OF AROTC	11
D.	LITERATURE REVIEW	13
III.	DATA AND METHODOLOGY	17
A.	DATA AND FORMATTING	17
1.	Data Files and Variables	17
2.	Data Processing	24
B.	METHODOLOGY	30
1.	Random Forest Models for Predicting Outcome Classification	30
2.	CART Model Comparisons.....	32
3.	Forecasting Population Totals Based on Probability Estimates.....	32
4.	Bootstrapping Random Forest Models to Determine Variability.....	35
IV.	ANALYSIS	39
A.	INDIVIDUAL MODEL PERFORMANCE	39
B.	MODEL VERIFICATION	40
C.	VARIABLE IMPORTANCE.....	42
D.	MISSION SET 2021 COMMISSION ESTIMATE	44
E.	PROFILE OF HIGHLY LIKELY ATTRITERS	45

F.	PROFILE OF CADETS WITH HIGH COMMISSION PROBABILITIES	50
G.	EFFECTS OF MIGRATION BEHAVIOR.....	55
V.	CONCLUSIONS AND RECOMMENDATIONS.....	59
A.	CONCLUSIONS	59
1.	Which Cadets Are Most Likely to Attrite, and What Factors Indicate that a Cadet Is Not Likely to Complete the Program?	59
2.	Which Cadets Are Most Likely to Successfully Complete the Program and Commission?	59
3.	What Insights can We Gain about the Cadets who Migrate from One Fiscal Year Cohort to Another?	60
B.	POTENTIAL APPLICATIONS.....	60
C.	RECOMMENDATIONS FOR FURTHER RESEARCH	60
	LIST OF REFERENCES	63
	INITIAL DISTRIBUTION LIST	67

LIST OF FIGURES

Figure 1.	Map of AROTC Brigades. Adapted from Tatro (2019b).	5
Figure 2.	ROTC Pipeline. Source: Haupt (2018a).	7
Figure 3.	Methodology Concept.....	30
Figure 4.	Histogram of Bootstrapped Commission Estimates for the Mission Set 2021 Cohort	36
Figure 5.	Model Performance Stratified by School Year.....	41
Figure 6.	Variable Importance for Cadets Enrolled Two Years.....	43
Figure 7.	Contracted vs. Conditionally Contracted Attrition Probability, Enrolled Two Years	47
Figure 8.	Attrition Categories by School Difficulty, Contracted Cadets Enrolled One Year	48
Figure 9.	Actual Attrition Rate by School Difficulty over Time	49
Figure 10.	Comparison of Commission Probability, All vs. Scholarship Cadets, Two-Year Enrolled	51
Figure 11.	Actual Commission Percentages (within Two Years) by Years Enrolled and Status	52
Figure 12.	Boxplots of Contracted Cadet Commission Probabilities, Enrolled Three Years.....	53
Figure 13.	Comparison of Estimated Attrition Probabilities by Migration Category, Two-Year Enrollees	57
Figure 14.	Comparison of Estimated Attrition Probabilities by Migration Category, One-Year Enrollees	57

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1.	Opening Enrollment Variables. Adapted from Haupt (2018b).....	18
Table 2.	ADM Descriptions. Adapted from Haupt (2018b).	18
Table 3.	The Relationship between Scholarship Award Codes and Scholarship Category. Adapted from Haupt (2018b).	19
Table 4.	Enrollment Status Code Values. Adapted from Haupt (2018b).	20
Table 5.	Disenrollment variables. Adapted from Haupt (2018b).	21
Table 6.	School variables. Adapted from Haupt (2018b).	22
Table 7.	Peterson’s Database Entrance Difficulty Codes. Adapted from Haupt (2018b).....	23
Table 8.	Statuses Based on Enrollment Status Codes.....	23
Table 9.	Administrative Variables Used for Model Building.....	26
Table 10.	Demographic Variables Used for Modeling.....	27
Table 11.	AROTC Variables Used for Modeling.....	27
Table 12.	School-related Variables Used for Modeling.....	28
Table 13.	The Dependent Variable Used in Modeling.....	29
Table 14.	Comparison of Random Forest and Classification Tree Model Performance (Misclassification Rates).....	39
Table 15.	Model Forecast Accuracy by School Year and Outcome.....	41
Table 16.	Variable Importance for the Random Forest Models.....	42
Table 17.	Mission Set 2021 Commissioning Population Forecast Based on November 2018 Enrollment Data.....	44
Table 18.	Actual Attrition Rates, by Cadet Status.....	47
Table 19.	Actual Commission Rates by Academic Class Code, Cadets Enrolled Three Years.....	54
Table 20.	Proportion of Cadets by School Difficulty, All Cadets vs. Migrators.....	56

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ADM	academic discipline mix
ADSO	active duty service obligation
APFT	Army physical fitness test
ARNG	Army National Guard
AROTC	Army Reserve Officer Training Corps
CART	classification and regression tree
CBEF	cadet based enrollment form
CCIMM	Cadet Command Information Management Module
DA	Department of the Army
ECP	early commissioning program
FICE	Federal Interagency Committee on Education
GRFD	guaranteed Reserve Forces duty
HBCU	historically Black college and university
HSI	Hispanic serving institution
JROTC	Junior Reserve Officer Training Corps
MJC	military junior college
MSL (I thru IV)	mission set level (one through four)
OCS	officer candidate school
PMS	professor of military science
RC	Reserve Component
ROTC	Reserve Officer Training Corps
SMC	senior military college
SMP	simultaneous membership program
STEM	science, technology, engineering, and math
SY	school year
TRADOC	Training and Doctrine Command
USACC	United States Army Cadet Command
USAR	United States Army Reserve
USMA	United States Military Academy

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The Army Reserve Officer Training Corps (AROTC) produces about two-thirds of active component officers each year and the majority of new Army Reserve and Army National Guard Lieutenants as well. To facilitate recruiting, the U.S. Army Cadet Command (USACC) maintains relationships with over 1,100 undergraduate colleges and universities from which it recruits and manages a large scholarship budget. Due to enrollment requirements, USACC must recruit new cadets before their junior year of studies, and it must retain a sufficient number of cadets to complete a given mission set two years later. However, current commission forecasts begin 24 months before the completion of each cohort, and do not provide insight with enough lead time to affect recruiting policies.

Our study seeks to develop a model that can be used to forecast the number of currently enrolled cadets who will commission nearly three years out, during the fall of the sophomore year. We explore the use of basic machine learning models, which use data from all cadets enrolled in AROTC, to forecast the aggregate number of commissions expected in a target fiscal year. Specifically, we develop and compare random forest models and classification and regression tree (CART) models to forecast commissions. We find that both models provide adequate capability for this purpose. Our models are used to produce aggregate forecasts, such as the total number of commissions in a future fiscal year, by combining predictions from a collection of models for cadets based on how long they have been enrolled. In addition to model building, we use a statistical concept known as bootstrapping to produce uncertainty estimates with our forecasts.

We use our models to assess which factors are most important to forecast commissions, and we use our results to provide insight to USACC regarding cadets who are likely to attrite or commission. The models indicate that factors such as the number of days a cadet has been contracted, a cadet's enrollment status, and the number of days until the expected commission date, among other variables, improve forecasting ability. We also analyze the population of cadets who ultimately commission in a different fiscal year than initially planned. Factors such as academic school difficulty and enrollment status impact

the likelihood of cadet attrition, and the influence of these factors differs depending on the length of time a that cadet has been enrolled.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Robert Koyak, who has been as selfless and supportive an academic mentor as a student could ask for. Professor Koyak has impressed me through his professionalism, his dedication to helping me improve as a researcher and writer, and the sheer breadth of his statistical knowledge.

I cannot help but mention my old colleagues in the Operations Analysis Division of the U.S. Army Cadet Command Headquarters. Gary Tatro has provided me as much insight and support in this and past research as he has laughs along the way. My first Operations Research supervisor, LTC (Ret.) Eric Jamison, modeled for me how analysts are supposed to think and influenced my graduate education immensely before it ever started. Additionally, my former colleagues MAJ Simon Boyd, Donna Dorminey, Tim Borgerding, and Mike Sutton each offered insight in years past that I was able to draw on during this research.

Most importantly, I am grateful for my teammate and wife, Beverly, and the love that she and our four kids provided during my graduate school experience. The sacrifices Beverly made to afford me the time to complete this research will not soon be forgotten. She may not have chosen the topic, but she shouldered more than her fair share of the hard work that went into making this thesis possible. As for Karis, Josiah, Deacon and Liberty, I hope that one day they stumble upon this work and are reminded of fond memories in Monterey, California.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This chapter discusses the primary purpose of this study and what we hope to achieve at the conclusion of the study. It also lays out the format of this thesis.

A. MOTIVATION AND OBJECTIVES

The primary mission of the U.S. Army Cadet Command (USACC), as it pertains to the Senior Army Reserve Officer Training Corps (AROTC), is to partner with colleges and universities to recruit, educate, train and commission leaders of character for the total Army (U.S. Army Cadet Command [USACC] 2018a). Each year, AROTC is responsible for producing about 60% of the Active Component Army Officers and over 5,000 total commissions, when U.S. Army Reserve and Army National Guard commissions are included (Department of the Army [DA] 2018, p. 2–1). As the largest producer of officers in the Army, AROTC provides a myriad of options for individuals to seek a commission and maintains a large network of schools from which it recruits.

Despite USACC’s recruiting potential, there are limitations which make it difficult for the organization to adapt to end-strength adjustments in the current or next cohort. If asked to produce more officers, for example, USACC cannot add new recruits to cohorts commissioning in the current fiscal year or the next fiscal year. Additionally, the organization currently lacks a tool to forecast the number of commissions it will produce for cohorts commissioning two fiscal years out and beyond. The aim of this thesis is to develop a model, or multiple models, to assist USACC in forecasting the number of commissions it will produce two or more years in the future.

In addition to providing a useful model, this thesis seeks to provide insight regarding the continuation behavior of various types of cadets in USACC’s many different commissioning programs. Specifically, we look at attrition rates and indicators of attrition, we attempt to quantify the rates at which cadets migrate from one fiscal year cohort to another, and we provide insight regarding other programs that are influenced by policy within the command. Lastly, we offer suggestions based on these insights to assist USACC in identifying helpful policy adjustments for recruiting and scholarship programs.

Our research is based on data containing over 197,000 observations of cadets, taken each November, over a period of six years. The nature of the AROTC program is such that a single cadet could be seen as few as one time or up to six times (six years) within the data. Features of the data include demographic variables, status variables which suggest a commitment level by each cadet, and academic and ROTC host school variables. A limitation of the data is that some academic and ROTC performance variables, such as grade point average (GPA), the most recent Army Physical Fitness Test (APFT) scores, and aptitude test scores are not consistently available.

B. RESEARCH FOCUS

The primary purpose of USACC is to provide commissioned officers to each component of the U.S. Army. To this end, the main focus of our research is to build an effective model that can be used by USACC to forecast the number of commissions two or more fiscal years out, based on the current population of cadets. The process of building this model also provides insight into which factors most affect retention, progression and completion of the program.

Our research focuses on answering the following questions:

1. Which cadets are most likely to attrite, and which factors increase the probability a cadet will not complete the program?
2. Which cadets are most likely to successfully complete the program and commission?
3. What insights can we gain about the cadets who migrate from one fiscal year cohort to another?

C. THESIS ORGANIZATION

The remainder of this thesis is organized as follows:

In Chapter II, we discuss details of the AROTC program that are useful to understanding how cadets enter, progress through and complete requirements to

commission through the AROTC program. We discuss scholarship programs and policies that influence recruitment and retention. We also outline related research others have completed.

In Chapter III, we describe the data used in our research, and we describe statistical methods used for analyzing the data.

In Chapter IV, we describe the analysis conducted with our data, using the methods defined in the previous chapter.

In Chapter V, we present our conclusions and recommendation for further research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. U.S. ARMY CADET COMMAND

The modern era of AROTC began when President Wilson signed the National Defense Act of 1916, creating the Reserve Officer Training Corps along with a formal Reserve Component (Brown 2001, p. 40). The formation of USACC as its own headquarters under the guidance of the Training and Doctrine Command (TRADOC) officially took place 15 April 1986 (Coumbe et al. 1999, p. 73). Several reorganization initiatives have led to its current organization, which includes eight brigades, seven of which are regionally aligned and one of which oversees AROTC activities in all Military Junior Colleges (MJC) and Senior Military Colleges (SMCs). A map of AROTC brigades is displayed in Figure 1. Enrollment in AROTC typically includes over 31,000 students (Tatro 2019b) at over 1,100 colleges and universities (USACC 2018c).

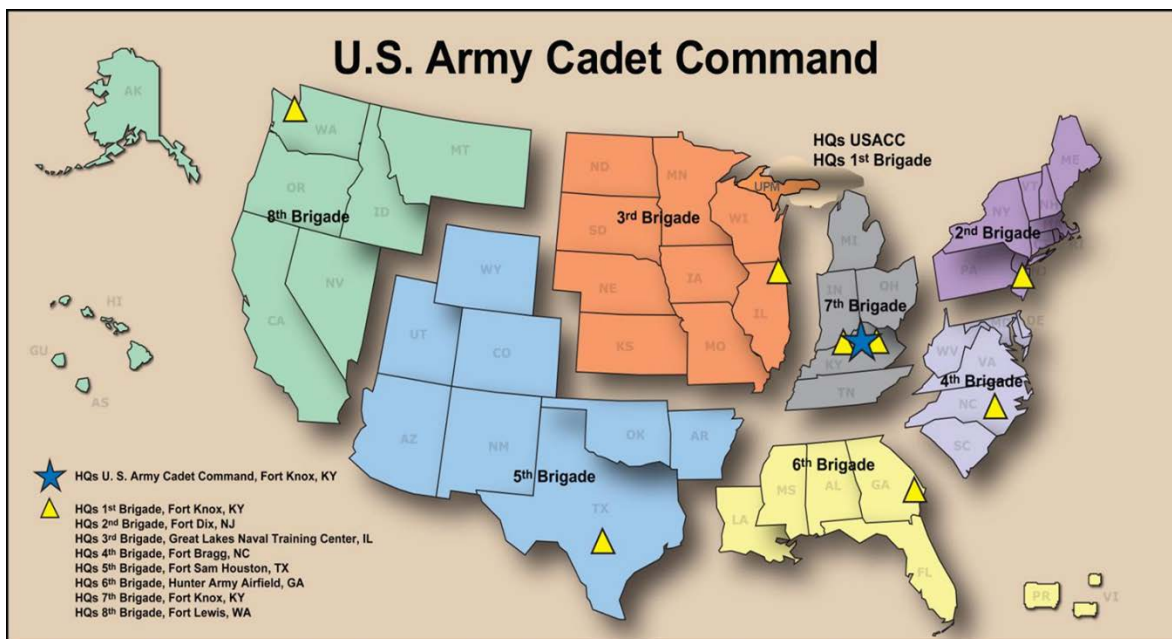


Figure 1. Map of AROTC Brigades. Adapted from Tatro (2019b).

Each AROTC Brigade oversees a varying number of host AROTC programs, also called Battalions, of which there are currently 273. Students wishing to participate in

AROTC enroll in military science courses at either a host program or an affiliated program that has a designated relationship with a nearby host. Programs are overseen by a Professor of Military Science (PMS) and are staffed by varying numbers of active duty military officers and non-commissioned officers, called “cadre,” as well as civilian administrative staff. The military cadre is responsible for teaching military science courses, mentoring and conducting training for cadets.

B. PROGRESSION IN AROTC

At its core, AROTC is a recruiting and training organization for the U.S. Army and Reserve Components (RCs). Many students without ambition to join the military enroll in basic military science courses at their academic institutions to learn more about the Army. While these students are potential prospects for becoming officers in the Army, most will not complete the program or commission. In order to gain an understanding of how cadets complete the program and become commissioned officers, it is essential to define the varying relationship levels students have with the AROTC program. Those who merely attend basic military science courses are referred to as “enrolled.” Those who have entered into a contractual relationship with the AROTC program are “contracted.” At times, those pending a contract are referred to as “conditional” or “conditionally contracted.”

Contracted cadets are also identified as either being a cadet on scholarship or as a non-scholarship cadet. As a recruiting organization for the active Army and RCs, USACC maintains a large scholarship budget to incentivize students to consider contracting with the Army. Many scholarship programs are available to AROTC prospects, some of which are described on the Cadet Command website (USACC 2018c). While both scholarship and non-scholarship cadets receive a monthly stipend once contracted, scholarship cadets also receive financial support for tuition and books, for room and board, or for campus meal plans. Regardless of scholarship status, all cadets who are contracted must show progress toward a commission.

Cadets in the AROTC program progress toward a commission by completing successive military science levels (MSLs). The path to producing a cohort is similar to a pipeline as depicted in Figure 2 (Haupt 2018a). There must be enough supply on the front

end of the pipeline to ensure that demand on the back end can be met. There are also specific entrance points in which cadets may begin the program and specific points at which individuals in the pipeline are likeliest to attrite. The primary sources of cadets for AROTC include the national scholarship process, which is aimed at high school students, campus-based recruiting efforts directed at students in their first or second year of school, and lateral entry cadets. Lateral entry cadets are those who typically join the program between their sophomore and junior year of school.

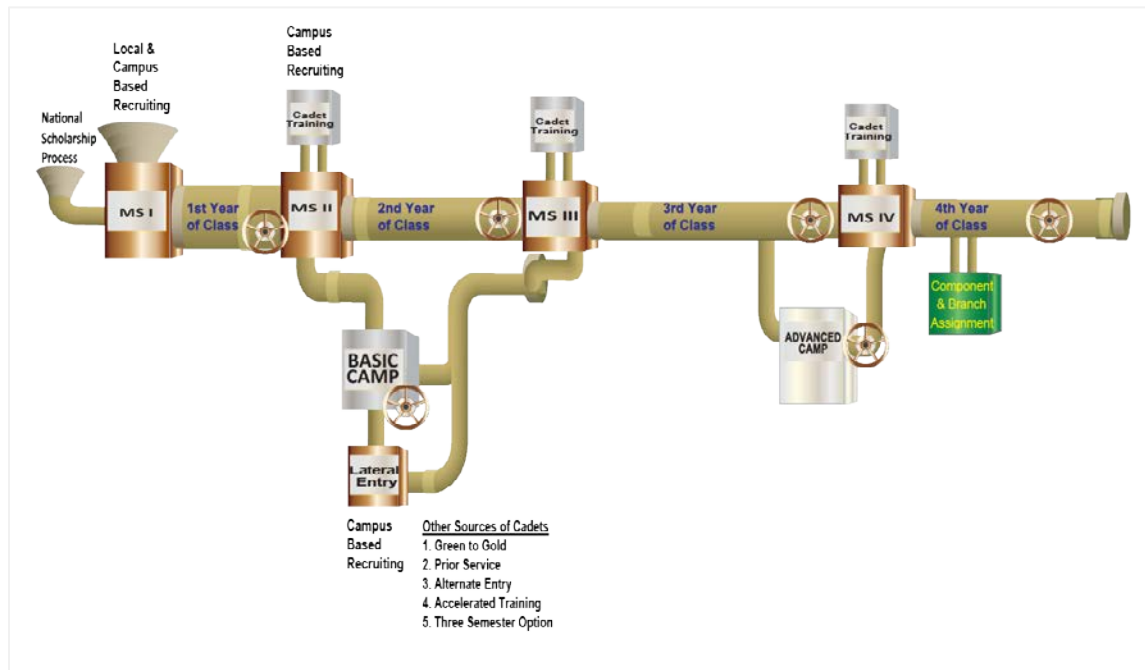


Figure 2. ROTC Pipeline. Source: Haupt (2018a).

1. Military Science Levels

In order to complete the program and commission as an officer in the U.S. Army, a cadet must have credit for the basic course (MSLs I and MSL II) in order to enroll in the advanced course (MSLs III and IV), and must complete the advanced course (U.S. Army 1996, p. 37). Every cadet who commissions will take and receive credit for the advanced course, but there are a number of ways in which cadets may receive credit for the basic course (DA 2011, p. 15). In addition to MSL I, II, III, and IV, there are codes for what are

known as “completion cadets.” Completion cadets are those who have completed ROTC requirements but not academic requirements, which often occurs for engineering majors or students who cannot take required classes on schedule. In such cases, the cadet is listed as an MSL V or MSL VI.

2. The Basic Course

Approximately 1,200 high school students per year are awarded four-year national scholarships, and another 1200 to 1500 high school students are typically awarded three year “advanced designee” scholarships (Borgerding 2019). These cadets often form the basis of an ROTC cohort, and are joined on many campuses by students recruited during freshmen orientations to take MSL I. Any cadet who starts the coursework by taking MSL I and continues on to the advanced course is considered a “progression” cadet. Cadets beginning at MSL I are the first to join a cohort’s pipeline.

While a majority of cadets enter the program during their first year of school and begin with MSL I, many will enter through one of a few other entry points. Students often take AROTC classes for the first time as an academic sophomore. These students are typically listed as enrolled in MSL II. The host Professor of Military Science (PMS) may deem the cadet as having met the requirements to enroll in the advanced course after having only completed MSL II. These cadets are generally campus-based recruits added to the pipeline of a cohort during their sophomore year. It is important to note that not all MSL I or MSL II cadets are contracted with the Army. Many cadets in the first two years of the curriculum are taking the course for academic credit only. Cadets who contract with the Army during this timeframe have increased their commitment to the program.

Four non-exhaustive examples of different paths a cadet may take to enter the AROTC program illustrate the variety of entrance points. The examples also demonstrate the challenge of modeling AROTC cadet behavior.

- A cadet who is a recipient of a four-year national scholarship out of high school enrolls into AROTC immediately as an academic freshman. He is classified as a progression cadet since his formal AROTC training includes the basic course classroom and field lab instruction.

- A second cadet first learns of AROTC from a peer during her academic freshmen year and decides to take AROTC beginning her sophomore year. She later becomes interested in contracting with AROTC and completes additional, compressed, training to gain credit for the basic course.
- A third cadet, who is a varsity athlete, takes AROTC his sophomore year for academic credit only. Because he has no ambition to commit to the Army, he completes two semesters of military science coursework and does not re-enroll in future courses or contract with the Army.
- The fourth cadet decides he wants to join the Army toward the end of his sophomore year, and, because he has not attended any basic military science courses and does not meet any criteria for receiving credit for the basic course, he must complete the basic camp at Fort Knox, Kentucky in order to enroll in the advanced course. After completing the requirement, he receives a two-year scholarship and enrolls for the first time as an academic Junior.

3. The Advanced Course

Requirements for enrollment into the advanced course (MSL III or above) are listed in chapter 103 of U.S. Code Title X and can also be found in Army Regulation 145–1, Section 3–12. In order to be enrolled in the advanced course, a cadet must be contracted or pending contract with the Army, and must have completed or received credit for the basic course. Cadets who have completed three years of Junior ROTC (JROTC) in high school, for example, are given credit for the basic course, as are students who have completed basic training as part of prior military service in an Active, Reserve or National Guard capacity. Additionally, some ROTC host programs have compressed curriculum opportunities within which cadets receive instruction for both MSL I and MSL II in an accelerated format.

Finally, students who are interested in contracting with AROTC and who have two academic years of school remaining, but who have not had the ability to complete the basic

course, may contract only after attending the basic camp. Basic camp is an accelerated training program that takes place in the summer at Fort Knox, Kentucky, and is designed to teach the same principles taught in the classroom and at field labs during the basic course. The basic camp is also the last campus-based recruiting effort available for schools to complete an ROTC cohort.

There are several AROTC scholarship programs that make it both a flexible commissioning source and a difficult process to model. Some of the unique scholarship programs offered by AROTC are discussed in the next sections.

4. Military Junior Colleges and the Early Commissioning Program

Students who attend a Military Junior College (MJC) receive training prior to their first year of college, which affords them credit for the basic course. Basic information about MJC programs can be found in Army Regulation 145–1 (DA 2011). After completing the training, these students are immediately contracted and enrolled in the advanced course at MSL III as first year college students. These cadets generally complete ROTC requirements in two years, commission into either the Army Reserve or Army National Guard and go on to complete their degrees at another institution as Second Lieutenants (DA 2011, p. 40). Program participants who receive their bachelor's degrees have the opportunity to gain an Active Duty commission through the Early Commissioning Program (ECP). For four of the six school years covered in our research there were five MJCs: Georgia Military College, Marion Military Institute, New Mexico Military Institute, Valley Forge Military Academy and College and Wentworth Military Academy. Wentworth Military Academy closed in 2017 (The Kansas City Star 2017).

5. Green to Gold Program

Each year USACC affords high-performing enlisted soldiers who have earned some college credit the opportunity to complete their degree on a path to becoming officers in what is known as the “Green to Gold” program (USACC 2016). Soldiers are nominated by their chains of command and selected by USACC. This off-campus recruiting program is effective for filling in recruiting voids in cohorts because a high percentage of cadets complete the program. Cadets in this program enter the ROTC pipeline at the MSL III

level. Requirements for the program include having a letter of acceptance to a school with an Army ROTC program, a letter of acceptance from an ROTC PMS and a current college GPA of 2.5. Once in the program, soldiers continue to receive both their basic pay and tuition, paid for by the Army. USACC has a limit of about 200 cadets it can accept into this program per year, but generally it awards fewer scholarships than that. More information regarding the Green to Gold Program can be found in CC 145–1 (USACC 2016) or on the program website (U.S. Army 2018).

6. Simultaneous Membership Program

Some students take advantage of a program called the Simultaneous Membership Program (SMP) in which they enlist into the Army National Guard (ARNG) or U.S. Army Reserve (USAR), and participate in training one weekend per month with a local National Guard or Reserve unit, while fully enrolled in the ROTC program. The SMP program affords students an opportunity to gain additional training and pay while pursuing a commission into the same Army component with which they drill (USACC 2018b). Multiple scholarship programs fall under the broader SMP title. These programs include the Guaranteed Reserve Forces Duty (GRFD) Scholarship Program, the Minuteman Scholarship program and the ECP, mentioned previously. Specific details for the program are covered in Cadet Command Regulation 145–1, Army ROTC Incentives Policy (USACC 2016).

C. FORECASTING CHALLENGES OF AROTC

Army end-strength authorizations change often, due to the inherent uncertainty involved in worldwide conflict, changing political climates and changes in military requirements. These changes to end-strength authorizations impact the Army's operational readiness for several years, especially in the officer corps, where end-strength is like a funnel. Nearly all of the officers enter at the beginning of a cohort and limited capability exists to add officers to the cohort later (Henning 2006, p. 19). If the Army decides it needs to increase authorizations for new Second Lieutenants in the current fiscal year, for example, the only major accessions source capable of immediate increased throughput is the Officer Candidate School (OCS). Neither AROTC nor the U.S. Military Academy

(USMA) can recruit new officers in the senior year of the cohort. If the Army fails to produce the required number of officers, the resulting officer cohort may be under-manned for years.

When additional accessed officers are needed over an extended period, AROTC is an ideal commission source for several reasons. First, congressional limitations in Chapter 753, Section 7442 of U.S. Code Title X prevent increased officer production at USMA by capping enrollment (John S. McCain National Defense Authorization Act for Fiscal Year 2019). Enrollment caps prevent USMA from substantially increasing commissions from current levels. Second, depleting the Non-Commissioned Officer corps by increasing OCS commissions during times of end-strength growth is counter-productive to increasing enlisted end-strength, since those selected for OCS leave the Non-Commissioned Officer corps. Additionally, officers produced through OCS are a short term addition to officer end-strength since OCS officers are less likely to remain in service for the ten years needed to develop into Army Majors (Demirel, 2002). In short, OCS is the short-term option while AROTC can be relied on to meet long-term accession needs.

While AROTC is not able to substantially increase output for the current accessions cohort, its recruiting potential for the long-term is vast, given its presence on over 1,100 campuses nation-wide and the array of scholarship programs offers. For this reason, AROTC, which produces about 60% of the Army's active component officers in any year, is the long-term commission source of choice for growing the officer corps (DA 2018, p. 2–1). If the Army needs to grow the officer corps in future years, it must have confidence that USACC can produce these officers. Since current law stipulates that newly contracted cadets must have at least two years of studies remaining in order to contract, AROTC cannot substantially increase its current or next fiscal cohorts. This limitation means that AROTC can significantly adjust its production only two or more years out.

Currently, AROTC does not have a tool to project commissions more than two years out. It is able to forecast for the current and next fiscal year cohorts only, with no tool to forecast the outcome of the current sophomore or freshmen cohorts (Tatro 2019a). Therefore, the forecasting capability does not fully support the forecasting need. If the

Army needs more officers in the future, no tool exists to assist USACC in identifying how well it is postured to produce the requirements or how to address any projected shortfalls.

An ideal time to forecast for future cohorts is during the spring of the cohort's sophomore year. This is because the last time to begin recruiting for a cohort is about 31 months before the cohort is completed, or March of the sophomore year. The spring semester of the sophomore year is a key decision point for recruiting "lateral entry" cadets, and it is the final opportunity to increase the number of Green-to-Gold scholarship selections. A reliable forecast during this time would assist USACC in deciding how many new prospects must be added to the ROTC pipeline, and which programs are best suited for an increase.

Several data inputs could be used to make better forecasts. Current commission forecasts only account for three factors, time until the end of the cohort and the total number of cadets contracted accounted for by brigade (Tatro 2019a). The primary driver of the forecast is the total number of contracted cadets for each AROTC brigade, since attrition rates generally differ between brigades due to differences in policy. It is likely that including gender, scholarship type, school type, major type, the number of conditional contracts pending, and any number of additional factors could improve forecasts and provide valuable insight to policy makers in USACC.

D. LITERATURE REVIEW

Literature involving ROTC, and specifically Army ROTC, generally falls into one of three categories: studies involving resource allocation, studies of career retention or performance comparison, and studies of the behavioral attitudes of ROTC cadets. The literature involving prediction of ROTC commissions is scarce. Studies related to forecasting attrition or retention have been conducted in populations that have already entered service for both active, reserve and civilian defense personnel.

In a Naval Postgraduate School master's thesis, Hopkinson (1988) uses regression models to identify which of several predictors is significant in predicting overall commission levels. The research does not use cadet data, but instead uses predictors such as unemployment rate, average yearly tuition costs, and AROTC enrollment, among others.

Among Hopkinson's conclusions are that unemployment and average yearly college tuition are positively correlated with commissions and that cadets with AROTC scholarships are retained in the program at higher rates than those without scholarships.

Goldman et al. (1999) published the report *Allocating Scholarships for Army ROTC*, examining USACC scholarship policy changes and their effect on scholarship acceptance rates. The report also explores options for restructuring the scholarship program and compares the retention rates of officers, by school type, who completed the AROTC program. While the purpose of the report is not directly related to the purpose of this thesis, the handling of USACC data and how RAND dealt with AROTC policies is very applicable. The findings of the study suggest that additional information such as standardized test scores, which are not available for our study, likely affect retention. Prospects with high standardized test scores, for example, were less likely to accept AROTC scholarships if they did not cover most education expenses (Goldman et al. 1999, p. 16). Additionally, the authors compare promotion rate differences of commissioned cadets based on school type, which encouraged us to compare schools similarly in our study (Goldman et al. 1999, p. 25).

Many studies have examined officer or enlisted attrition in the U.S. armed forces. Ugurbus and Korkmaz (2015) for example, examine first term attrition factors for both enlisted and officer Selected Marine Corps Reservists. The authors use one logistic regression model for officers and another for enlisted Marines to identify which factors are correlated with a subject failing to complete their initial service commitment. One of the challenges of the methods used is that dataset sizes are reduced substantially due to assumptions and missingness. What also makes such studies distinguishable from ours is that AROTC cadets have the ability to enter at different points in a cohort, whereas the data for subjects in attrition studies conducted after accessions typically begins at the time of accession.

A master's thesis by Streetzel (2018) uses a decision-tree model to forecast accession and attrition in the Selected Marine Corps Reserve. Like our study, Streetzel's goal is to forecast aggregated outcomes of his study population in order to inform policy

decisions. The study successfully modeled the ability to predict outcomes 12 months into the future.

What distinguishes our study from those reviewed is the unique aspect of AROTC that allows cadets to enter at varying points in time. We did not find any attrition studies that accounted for individuals who could join cohorts at separate points in time. Our approach is novel in that we explore attributes of both AROTC cadets and the educational institutions they attend. Additionally, we are not aware of any studies or models which attempt to forecast ROTC commissions using basic machine learning techniques such as random forest models or decision trees. We also use bootstrapping to account for uncertainty in our commission forecast, and in the development of a prediction interval.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA AND METHODOLOGY

A. DATA AND FORMATTING

In this section, we outline the data sets used in this study, we describe each variable used, and we provide a detailed summary of steps taken to prepare the data for analysis.

1. Data Files and Variables

All data used in our research was provided by USACC, as was a detailed data dictionary. The data provided for research includes opening enrollment snapshots, historical commission data, disenrollment data, and data describing academic and AROTC host institutions.

a. Opening enrollment snapshots

Opening enrollment data is recorded by USACC the first Tuesday of November every year. The data used for this analysis includes opening enrollment snapshots from each school year beginning in 2013 and ending in 2018. Each snapshot includes information about cadets that are enrolled in AROTC on the snapshot date. The opening enrollment data consists of demographic information, academic information, and information related to each cadet's relationship with AROTC. Individual cadets are observed at least once and up to six times in the data, resulting in 194,492 observations of 103,483 unique cadets.

The opening enrollment snapshots include the variables listed in Table 1. Cadet data was linked using an anonymous employer identification number (*emplid*) generated by USACC. The *emplid* also allows separate datasets to be linked.

The *sy*, *emplid*, *sex_cd*, and *redcat* variables indicate the school year of the observation, the *emplid*, the gender code of each cadet and the race-ethnicity code of each cadet, respectively.

The *rotc_sch_cd* and *acad_sch_cd* variables represent the academic Federal Interagency Committee on Education (FICE) codes associated with the AROTC host

institution and a cadet's academic institution, respectively. There are 1190 school code values in our data, 275 of which are associated with host AROTC programs.

Table 1. Opening Enrollment Variables. Adapted from Haupt (2018b).

Variable	Description	Type
<i>sy</i>	Academic school year	Categorical
<i>emplid</i>	Unique employer ID for each cadet	Numeric
<i>sex_cd</i>	Sex code of the cadet	Categorical
<i>redcat</i>	Race-ethnicity category	Categorical
<i>rotc_sch_cd</i>	The academic FICE code of the host AROTC program	Categorical
<i>acad_sch_cd</i>	The academic FICE code of the academic school attended	Categorical
<i>brigade_cd</i>	The Brigade aligned with a cadet's host AROTC program	Categorical
<i>acad_disc_mix_cd</i>	Academic discipline mix describing the general field of study	Categorical
<i>ms_init_cd</i>	Initial military science level the cadet entered	Ordinal
<i>ms_cls_enrl_cd</i>	Military Science Level the cadet is currently enrolled in	Ordinal
<i>acad_cls_enrl_cd</i>	Academic level of the cadet	Ordinal
<i>sclr_award_cd</i>	The length of scholarship (if applicable)	Categorical
<i>sclr_award_cat_cd</i>	Specific scholarship program code (if applicable)	Categorical
<i>enrollment_dt</i>	Date cadet initially enrolled in AROTC	Date
<i>contract_dt</i>	Date AROTC contract became effective	Date
<i>graduation_dt</i>	Projected graduation date	Date
<i>comm_dt</i>	Projected commission date	Date
<i>grfd_nmbr_id</i>	Unique GRFD contract number (iff applicable)	Alphanumeric
<i>enrl_stat_cd</i>	Enrollment status code	Categorical

The *acad_disc_mix_cd* variable groups each cadet's major into one of five categories, called the academic discipline mix (ADM). The ADM categories are broad, and degree fields are generalized considerably. Each category is listed in Table 2.

Note that ADM codes 3 and 4 are considered science, technology, engineering or math (STEM) majors by USACC and are used to determine the percentage of students commissioning with STEM degrees each year.

Table 2. ADM Descriptions. Adapted from Haupt (2018b).

<i>academic mix code</i>	Description
1	Generalist
2	Technical & Management
3	Physical Sciences
4	Engineering
5	Nursing

The *ms_init_cd* variable represents the first MSL assigned to a cadet. This variable may be blank if the cadet is enrolled in AROTC for the first time, and takes values one through three.

The *ms_cls_enrl_cd* variable indicates the MSL in which a cadet is enrolled, and takes on values one through six or “C,” for completion cadets.

The *acad_cls_enrl_cd* variable indicates the academic level of a cadet. Because students often enroll in undergraduate studies with academic credits from high school advanced placement courses, there is some ambiguity when defining which academic class a cadet is associated with. Values for *acad_cls_enrl_cd* include one through five and “G,” for graduate students.

The *sclr_award_cd* and *sclr_award_cat_cd* variables describe the length of scholarship awarded to a cadet and the scholarship program the cadet is participating in, if applicable. Values for these variables and the corresponding scholarship programs are described in Table 3.

Table 3. The Relationship between Scholarship Award Codes and Scholarship Category. Adapted from Haupt (2018b).

<i>sclr_award_cd</i>	<i>sclr_award_cat_cd</i>	Scholarship Award Category Description
2	2A	2-YR GREEN TO GOLD SCHOLARSHIP
2	2C	2-YR CB ENROLLED SCHOLARSHIP
2	BC	BASIC CAMP SCHOLARSHIP
2	HP	GREEN TO GOLD HIP POCKET SCHOLARSHIP
2	MJ	2-YR DEDICATED MJC SCHOLARSHIP
2	N2	2-YR CB NON-ENROLLED SCHOLARSHIP
3	2D	2-YR ADVANCE DESIGNEE SCHOLARSHIP
3	2H	2H - 2 1/2 YR SCHOLARSHIP
3	3A	3-YR GREEN TO GOLD SCHOLARSHIP
3	3C	3-YR CB ENROLLED SCHOLARSHIP
3	3U	3-YR AD USMA SCHOLARSHIP
3	N3	3-YR CB NON-ENROLLED SCHOLARSHIP
4	3D	3-YR ADVANCE DESIGNEE SCHOLARSHIP
4	3H	3H - 3 1/2 YR SCHOLARSHIP
4	4A	4-YR GREEN TO GOLD SCHOLARSHIP
4	4R	4-YR SCHOLARSHIP
4	4U	4-YR HQCC SCHOLARSHIP
4	J2	MJC 2+2 SCHOLARSHIP
4	QE	4-YR HBCU SCHOLARSHIP

The scholarship-related variables help to identify the specific scholarship program a cadet participates in. These variables also hint at additional information about scholarship cadets, like the point in the cadet’s college career in which they began pursuing a contract and whether or not they are enrolled in AROTC when the scholarship is awarded.

There are four date-related variables in the opening enrollment data: *enrollment_dt*, *contract_dt*, *graduation_dt*, and *comm_dt*. The enrollment date refers to the date a cadet is first enrolled in AROTC. The contract date exists only for cadets who are contracted with AROTC. The graduation and commission date variables are the projected graduation and commission dates, respectively, at the time of the opening enrollment data pull. While the enrollment date rarely changes for a cadet, projected graduation and commission date changes are not uncommon. Each of the date variables are used to derive time-related predictors, which will be discussed in Section 2 of this chapter.

The *grfd_nmbr_id* variable is a scholarship identification code assigned to cadets who elect to participate in the GRFD program.

The *enrl_stat_cd* variable is the primary means by which AROTC identifies the enrollment status of cadets. We also use these codes, listed in Table 4, to generalize each cadet’s relationship with AROTC by creating a status variable, which will be discussed in Section 2 of this chapter. The opening enrollment datasets provided do not include cadets in a commissioned (“M”) status. Data for these cadets is provided in the historical commissions dataset.

Table 4. Enrollment Status Code Values. Adapted from Haupt (2018b).

w_enrl_stat_cd Value	Description
B	Conditional - pending waiver or admin action
C	Completion - All ROTC Requirements Met
E	Enrolled and Contracted
E\$	Enrolled and Contracted, Scholarship
I	Not a U.S. Citizen. Not Contracted.
L	LOA - Pending Return
M	Commissioned
R	Scholarship Award Pending
S	Conditional - Advanced Scholarship Designee
TR	Transferring from one program to another
U	Enrolled (Non-Contracted) in MS1 or MS2 at a MJC
V	Enrolled (Non-Contracted) in MS 1,2,3 or 4 at a SMC
X	Green to Gold Active Duty Option
Z	Not Contracted - Enrolled in MS1 or MS2

b. Historical Commission Data

The historical commission data includes the final AROTC commissioning information for 27,332 cadets who completed the program between the beginning of fiscal year 2013 and January 2019. For the purposes of our study, we use only the actual commission date from this dataset. Commission dates in the opening enrollment data are projected dates, whereas the actual commission date is the date a cadet successfully completed the program.

c. Disenrollment Data

The disenrollment data provided for this study includes a disenrollment date for 60,813 individuals who were disenrolled during the period covered by our research. The data provided is from two separate data tables in the Cadet Command Information Management Module (CCIMM). Disenrollment dates for cadets who were once contracted but later dropped out of the program are placed into a disenrollment table in CCIMM to ensure benefits are no longer paid. Cadets who have never been contracted with the Army are not in this table since non-contracted students do not receive monetary benefits. Disenrollment dates for non-contracted cadets were taken from an enrollment status table and the disenrollment date provided for these cadets is the date in which their enrollment status changed from an active to a non-active status in CCIMM. Table 5 shows the fields used from the disenrollment data.

Table 5. Disenrollment variables. Adapted from Haupt (2018b).

Variable	Description	Type
<i>emplid</i>	Unique employer ID for each cadet	Numeric
<i>w_dis_drop_dt</i>	Date Cadet was dis-enrolled (if contracted)	Date
<i>effdt</i>	Effective date Cadet changed to a non-enrolled status (if never contracted)	Date

d. AROTC and Academic Institution Data

USACC provided data related to each academic institution it maintains a relationship with. Variables used from the institutional data are listed in Table 6, and include unique identifiers for each host and academic institution, the brigade each school is aligned with, whether the school is recognized as serving a minority population or distinguished as a military college, and an academic difficulty code from Peterson's Database®. Peterson's Database is a comprehensive academic database that includes information about undergraduate and graduate academic institutions. USACC purchases a license to use Peterson's Database each year and provided the most recently available admission difficulty category for each academic institution. Table 7 defines the entrance difficulty codes from Peterson's Database provided for our study.

Table 6. School variables. Adapted from Haupt (2018b).

Variable	Description	Type
<i>host_fice</i>	Academic FICE code of the AROTC host school	Categorical
<i>academic_fice</i>	Academic FICE code of the academic school attended	Categorical
<i>school_type_description</i>	Description of Public or Private	Categorical
<i>military_type_code</i>	Description of military school type	Categorical
<i>hbcu_hsi_code</i>	Description of whether school is Historically Black College and University (HBCU) or Hispanic Serving Institution (HSI)	Categorical
<i>petersons_difficulty_code</i>	Entrance difficulty of academic school from Petersons Database	Categorical

The *host_fice* and *academic_fice* variables match the *rotc_sch_cd* and *acad_sch_cd* previously discussed when describing the opening enrollment data.

The *school_type_description* variable identifies whether an academic school attended by a cadet as either public or private.

The *military_type_code* variable identifies each academic school as being either a SMC, MJC or neither.

The *hbcu_hsi_code* variable identifies academic schools that are considered historically black colleges and universities (HBCUs), Hispanic serving institutions (HSIs) or neither.

Table 7. Peterson's Database Entrance Difficulty Codes. Adapted from Haupt (2018b).

Peterson's Difficulty	Description
MOST	Most Difficult - More than 75% of all freshmen were in the top 10% of their high school class and scored over 1310 on SAT (critical reading and math scores combined) or over 29 on ACT; about 30% or fewer of all applicants accepted.
VERY	Very Difficult - More than 50% of all freshmen were in the top 10% of their high school class and scored over 1230 on SAT (critical reading and math scores combined) or over 26 on ACT; about 60% or fewer of applicants accepted.
MOD	Moderately Difficult - More than 75% of freshmen were in the top 50% of their high school class and scored over 1010 on SAT (critical reading and math scores combined) or over 18 on the ACT; about 85% or fewer of applicants accepted.
MIN	Minimally Difficult - Most freshmen were not in the top 50% of their high school class and scored somewhat below 1010 on SAT (critical reading and math scores combined) or below 19 on ACT; up to 95% of applicants accepted (not 100%).
NONC	Noncompetitive - Virtually all applicants were accepted regardless of high school rank or test scores.
NR	Reporting requirements are voluntary and many schools do not report for various reasons.

e. Derived variables

We derive several variables for predictive purposes during our study. The most basic of these variables is a generic status of each cadet observation. The *status* variable groups cadets by enrollment status code to identify the basic relationship the cadet has with AROTC. Cadets with an enrollment status code of E, E\$, C, L or X, for example, are contracted cadets. Likewise, several enrollment status codes indicate whether a cadet is conditionally contracted, a prospect, or transferring between programs. A summary of the enrollment status codes that comprise each status group is given in Table 8.

Table 8. Statuses Based on Enrollment Status Codes

status assigned	Enrollment Status Codes
prospect	I, U, V, Z
conditional	B, R, S
contracted	E, E\$, C, L, X
transfer	TR

In order to construct numeric variables based on the amount of time enrolled or contracted and the amount of time until the projected graduation or commission date, the

dates that each opening enrollment snapshot occurred are added to the opening enrollment data. After dates are added, we derive four time-related variables for predictive modeling. The derived variables are *since_enroll*, *since_contract*, *til_grad* and *til_cmsn* and represent the time, in days, a cadet has been enrolled or contracted, or the number of days until the projected graduation or commission date. All cadets have enrollment dates, projected graduation dates and projected commission dates. Non-contracted cadets do not have a contract date, and therefore do not have an original value for the *since_contract* variable.

We also construct variables to determine if a cadet has changed schools or majors, and additional variables to indicate if the changes take place within the year prior to the observation used to build our models. Additionally, variables are designed to identify when a commission or graduation date has changed more than 180 days, and whether a cadet has migrated from one fiscal year cohort to another.

2. Data Processing

Data processing includes four primary phases. The first and second phases include preparing the data for use and combining like data sets, respectively. The third phase involves using the combined data set to create subsets of individuals based on the length of time since enrollment and converting the information into model-ready data. Lastly, we establish rules for dealing with missing values and ensure variables are of the correct type for modeling.

a. Initial data preparation

The first step in preparing our data is to merge each opening enrollment data set from different years into one complete dataset. We combine the six separate opening enrollment snapshots into a single data frame. While combining the datasets, we convert dates into date format. Observations having blank *emplid* variables are dropped. We also add the data pull dates to the corresponding observations and derive the four time-related variables mentioned at the end of Section 1.e.

b. *Wide format conversion*

In order to estimate the probability an individual in the AROTC program commissions, it is necessary to structure all observations of an individual longitudinally, in wide format. We use the reshape function from the *stats* library, included in the base distribution of R software (R Core Team 2018). The time variable used to convert our data into wide format is the school year. The resulting data frame includes a column for each variable in our original data for every school year in the data set. Variables that do not change in time, such as sex and race-ethnicity are not replicated. The resulting wide format includes one row for each *emplid*, and all known variable values for each of the years the *emplid* is observed. A cadet appearing in only one school year will have variable values for the columns associated with only that particular year.

After converting to wide format, several variables are added to each observation to facilitate the creation of an outcome variable during model preparation. Values of *emplid* from the wide format data are matched with values of *emplid* in the historical commissions data to return actual commission dates for cadets who have commissioned. The actual commission dates are added to the wide format data.

Next, we add disenrollment dates to the data by matching *emplid* values in our wide format data to those in the disenrollment data. The disenrollment date and disenrollment code for cadets with matching records in the two data sets are recorded for each observation. Disenrollment dates prior to November 2013 indicate that a cadet in our dataset disenrolled and re-enrolled in AROTC since all cadet observations in our data occurred after November 2011. Therefore, disenrollment dates prior to November 2013 are excluded and these cadets are assumed not to have left the AROTC program.

c. *Model preparation*

In order to avoid using multiple observations of any cadet when building each model, we develop a function that builds model-prepared datasets based on the length of time cadets have been enrolled in AROTC. Our function, *MakeY*, takes three inputs: our wide format dataset, the number of years since enrollment, and the number of years ahead we wish to predict. We label the second input *years_enrolled*, and the third input

years_ahead. As an example, if we wish to make a dataset including all observations of individuals who have been enrolled for two years (between 18 and 30 months) and identify the outcome of each individual two years later, we call our function: *MakeY(dataset, 2, 2)*. The resulting data includes only the observations of cadets who have been enrolled for about two years.

We categorize variables returned by our *MakeY* function into the following categories: administrative variables, AROTC variables, school variables and the dependent variable. The specific variables returned in the *MakeY* function are listed in Tables 9 through 12. Once the user defines the inputs, the function identifies the *since_enrolled* variable for each school year for each *emplid*, and returns all variables corresponding to the applicable observation, if one exists. For example, if *MakeY(dataset, 1,2)* is entered, the function identifies all observations of each cadet and if, at the time of one of those observations, the cadet was enrolled for one year (between six and 18 months), the data for that school year are returned for the corresponding cadets.

Table 9. Administrative Variables Used for Model Building

Variable Name	Description
<i>use</i>	TRUE if observation is useable
<i>emplid</i>	Unique employer ID
<i>init.cohort</i>	Initial ROTC cohort
<i>current.cohort</i>	Current ROTC cohort
<i>sy</i>	School Year of observation
<i>school.rotc</i>	Host ROTC school code
<i>school.acad</i>	Academic School

Administrative variables are included in the creation of our model data for verification that our *MakeY* function works as intended and to assist us when analyzing model outputs. The *use* variable is a Boolean indicator signifying whether or not each *emplid* has an observation that meets the requirements of the function inputs. The initial and current cohorts are the cohorts a cadet is associated with at the time of enrollment and during the observation returned during the *MakeY* function, respectively. The school year variable is necessary to check the output of our functions with the original variable values for accuracy. School variables correspond to the FICE codes of the host and academic schools a cadet attends.

Table 10. Demographic Variables Used for Modeling

Variable Name	Description
<i>sex</i>	Sex code
<i>raceeth</i>	Race Ethnicity Code

The demographic variables used, listed in Table 10, indicate the sex and race-ethnicity of each cadet.

The AROTC variables, listed in Table 11, describe a cadet’s involvement and relationship with the AROTC program at the time of the observation. Several of the variables are constructed during the *MakeY* function. Specifically, the *extend.comm.date*, *reduce.comm.date*, *migration* and *start.month* variables are determined during construction of the model-ready data. These variables indicate that a cadet’s commission date has changed more than 180 days, or whether the cohort a cadet is currently a part of has changed since initial enrollment. The migration variable is numeric, with a value of “1,” for example, indicating the cadet is scheduled to commission in a cohort one fiscal year later than originally expected. The *start.month* variable indicates the month of the year a cadet originally enrolled in AROTC.

Table 11. AROTC Variables Used for Modeling

Variable Name	Description
<i>initMSL</i>	Initial Military Science Level
<i>initBrigade</i>	Initial Brigade
<i>mil.School</i>	Military school type (MJC, SMC, None)
<i>grfd</i>	Active GRFD scholarship or None
<i>msstatus</i>	MSL at time of observation
<i>enrstatus</i>	Enrollment code at time of
<i>status</i>	Status at time of observation
<i>days.enrolled</i>	Days since enrollment
<i>days.contracted</i>	Days since contracting
<i>days.tocomm</i>	Days until expected commission
<i>scholar.code</i>	Length of scholarship (if applicable)
<i>scholar.catcode</i>	Specific Scholarship type
<i>extend.comm.date</i>	True if commission date extended
<i>reduce.comm.date</i>	True if commission date reduced 180+
<i>migration</i>	Indicates number of cohorts a cadet
<i>start.month</i>	Month initially enrolled in AROTC

Each of the AROTC variables listed in Table 11 are predictor variables, which we use to assist in forecasting the likelihood a cadet commissions in the future.

Modeling variables associated with a cadet's school are listed in Table 12. The variables *change.sch.acad*, *change.sch.acad.lastyr*, *change.adm* and *change.adm.lastyr* are derived during the *MakeY* function and serve to identify cadets whose school or academic discipline has changed, and the timing of the change. The *stem* and *nurse* variables are also determined during the *MakeY* function and are based on the ADM code used by USACC. Additionally, the *extend.grad.date* and *reduce.grad.date* variables identify cadets whose graduation dates have changed 180 or more days since enrollment.

Table 12. School-related Variables Used for Modeling

Variable Name	Description
<i>change.sch.acad</i>	True if cadet has changed academic
<i>change.sch.acad.lastyr</i>	True if school change within last year
<i>pubpriv</i>	Cadet attends public or private school
<i>school.diff</i>	Peterson's difficulty code of academic
<i>minority.sch</i>	Minority school type attending (if
<i>acadclass</i>	academic class at time of observation
<i>adm</i>	ADM at time of observation
<i>change.adm</i>	True if cadet has changed ADM
<i>change.adm.lastyr</i>	True if ADM has changed in the year prior to the observation
<i>stem</i>	True if ADM = 3 or 4
<i>nurse</i>	True if ADM = 5
<i>days.tograd</i>	Days until expected graduation at the
<i>extend.grad.date</i>	True if graduation date extended 180+ days since enrollment
<i>reduce.grad.date</i>	True if graduation date reduced 180+ days since enrollment

The outcome variable is determined during the execution of the *MakeY* function and possible values are given in Table 13. If a commission date exists for a cadet, the function determines if that date is between the current observation and the number of *years_ahead* the user input. If a commission date occurs between the two observations, the outcome for the cadet is labelled as having commissioned. If the commission date does not occur between the two observations, the cadet outcome is set to "continue," since we know the cadet eventually commissions after the *years_ahead* timeframe.

If no true commission date exists for a cadet, the *MakeY* function identifies whether a disenrollment date exists. If a disenrollment date exists for the cadet between the current observation and the observation the number of *years_ahead* the user input, then the outcome for the cadet is attrite. If the disenrollment date takes place after the number of *years_ahead* input by the user, that cadet’s outcome is labelled as “continue.”

For cadet observations having neither a commission date or a disenrollment date, it is possible the record has gone “stale,” or that it has been forgotten about. This is evident when a cadet is observed during one year, but not in subsequent years. If a cadet has no disenrollment or commission date and the cadet’s status variable is blank for the observation in the *years_ahead* look, we assume the cadet attrited. The only additional scenario is such that we continue to observe the cadet in the *years_ahead* observation, but have no final commission or disenrollment date. This occurs frequently in the last several years of our data since most cadets who enrolled in the last year or two of our data have not had time to complete the program or attrite. In such cases, we assign an outcome of “continue,” since we know the cadet is still participating in AROTC.

Table 13. The Dependent Variable Used in Modeling

Variable Name	Description
<i>outcome</i>	"Attrite", "Commissioned" or "Continue"

d. Post data preparation

After executing the *MakeY* function, we complete data preparation tasks. We drop unusable observations, blank or missing values are adjusted for use with the models we build. For factors such as military school or scholarship type, we replace blanks with “no” values, for example. For non-scholarship cadets we replace missing values with “0” for the length of scholarship. Some predictive variables produce missing values that cannot easily be addressed and are small in number. For each dataset we create, we drop observations with missing values for *msstatus*, *initmsl*, *adm* or *start.month*. For all models, the total number of observations dropped is less than one per thousand observations. Lastly, several

factors are converted to ordered factors. Specifically, we order *initmsl*, *school.diff*, *msstatus*, *acadclass* and *status*.

B. METHODOLOGY

In this section, we explain the methodology used for predictive modeling of an AROTC cohort two years from now. We also discuss comparisons to other statistical models using the same data.

1. Random Forest Models for Predicting Outcome Classification

Our study utilizes five random forest models to forecast the total number of cadets that will commission two years in the future. Our dependent variable is a multinomial response with three classes: attrite, commission, and continue. The amount of time a cadet has been enrolled determines which model will be used to predict the likelihood the cadet commissions. After tuning and training each model, we test the results on an AROTC cohort for which we already know the outcome, and finally make a prediction for a cohort based on the most recent year of data provided. Figure 3 depicts our methodology visually.

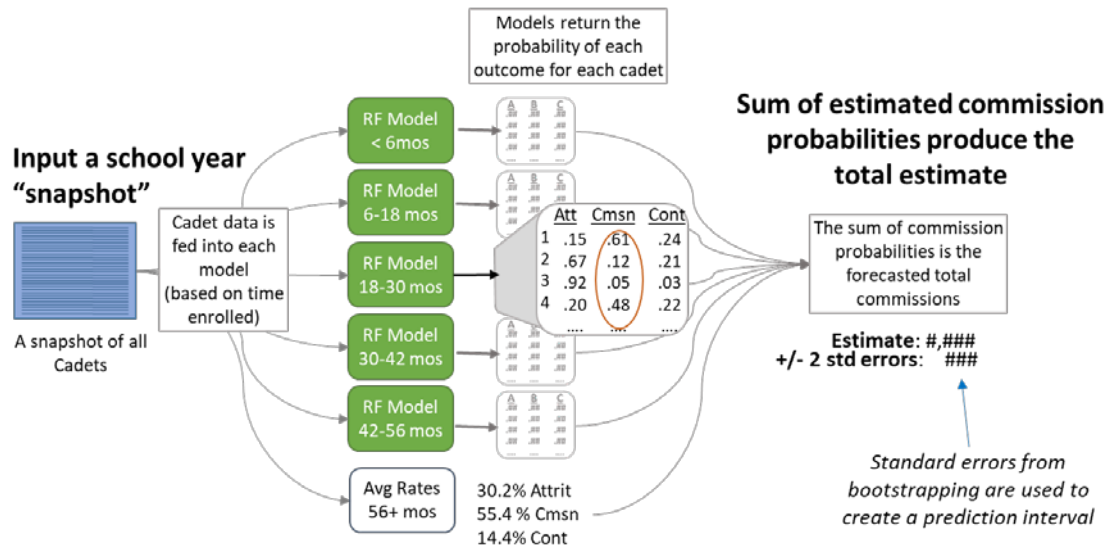


Figure 3 presents our methodology in visual form. The sums of estimated commission probabilities are combined for all models to form a prediction. For cadets enrolled longer than 56 months, the average outcome probabilities are used.

Figure 3. Methodology Concept

Random forest models are statistical models extending from the concept of classification and regression trees (CART). The statistical concept of decision trees, as they are often called, was introduced by Brieman et al. in 1984, as discussed in Faraway (2006, p. 278). Trees use a recursive partitioning algorithm to predict or classify a target variable. The goal of a tree-based model when used for classification is to create leaves, or terminal nodes, which are as pure as possible. Purity of tree leaves can be measured in several ways, but each measure of purity essentially is minimized when all class predictions in a leaf are of the same type (Faraway 2006, p.287). If i is a terminal node and there are k possible classes, then p_{ik} is the observed proportion of class k in node i , and D_i , the measure of impurity for node i , can be calculated using the Gini index as a measure of deviance. The formulation of the Gini index (Faraway 2006, p. 287) is given in (1) as

$$D_i = 1 - \sum_k p_{ik}^2 \quad (1)$$

Random forest models extend the idea of trees by growing many trees from randomly sampled predictors on bootstrap samples as described by Brieman (2001). Random forests are effective models because they reduce variance in the classifications and keep the model bias low. The method achieves this by using a technique called bootstrap aggregating, or bagging. Bagging consists of selecting a random sample of predictor variables from the training data, with replacement, and fitting a tree based on the sampled factors. The process is repeated many times while training, and the overall classification for each observation is made by predicting the classification outcome that receives the most “votes” from all trees grown. The typical default value for the number of randomly selected predictors in a random forest model is the square root of the number of predictors available, but this parameter can be adjusted.

We train five random forest models for the purposes of our study. Each model is designed to forecast the outcome of individuals based on the amount of time they have been enrolled in AROTC. We tune the models by testing multiple combinations of two input parameters, namely the number of predictor variables to try in each decision tree (*mtry*), and the number of trees to predict (*num.trees*). The *mtry* and *num.trees* parameters

that produce the lowest misclassification rates during the tuning process are used to build our final models.

2. CART Model Comparisons

While random forest models have many advantages, their primary drawback is that they are not easily interpretable. In cases where a simple tree performs nearly as well as a random forest model, it is often preferable to use the former due to interpretability. For this reason, we build separate classifications using trees to compare to each of our random forests. We use the *ranger* package in R to build our random forest models, due to the fast implementation it provides (Wright and Ziegler 2017).

The first section of this chapter discusses the general concept of CART models. We compare each of our random forest models to corresponding classification trees built using the *rpart* package in R (Therneau and Atkinson 2018). Using the same training data as for our random forest models, we build large trees using an initial complexity parameter, (*cp*), equal to 0.005. To prevent growing too large of a tree, we use the process of tree pruning described by Faraway (2006, p. 283) in which we select the smallest size of the tree that is within one standard deviation of the *cp* with the smallest cross-validated error.

We build five classification tree models to compare to the corresponding random forest models discussed previously. Each tree is trained using the same training data used to build the corresponding random forests, and is therefore a model to predict the outcome of a cadet based on how long the cadet has been enrolled. Each model is pruned, and the performance of the pruned tree is determined using the misclassification rate on the same test set used to assess the performance of the similarly purposed random forest model.

3. Forecasting Population Totals Based on Probability Estimates

Consider a “target” population of AROTC cadets at a fixed moment in time. This population is a mixture of individuals who are in the program for their initial year, second year, etc. One of our objectives is to forecast how many of these individuals will commission, attrite, or will still be enrolled (“continue”) a given number of years into the future. We also seek to place uncertainty bounds on our forecasts. We now describe how

we construct forecasts and prediction intervals for population totals from the probability estimates that we obtain using random forests, classification trees, or any other model.

We begin by stratifying our target population by the number of years that a cadet has been enrolled in the AROTC program. We assume that we are at or near the beginning of an academic year, so that each cadet has t years of enrollment completed, where $t = 0, 1, \dots, 7$. A cadet at $t = 0$ is just starting the AROTC program; $t = 4$ implies that the cadet is starting a fifth year of study. Because many students require more than the traditional four years to complete an undergraduate degree, we consider cadets up to seven years of enrollment, although the strata rapidly diminish in size past $t = 4$ years. Let n_t denote the number of individuals in stratum t . The status of cadet i in stratum t at the end of y fiscal years from the current one can be described using trinomial random variables $(X_{1,t,y,i}, X_{2,t,y,i}, X_{3,t,y,i})$, where

$$X_{k,t,y,i} = \begin{cases} 1, & \text{if the cadet has outcome } k \text{ at } y \text{ years into the future} \\ 0, & \text{if the cadet does not have outcome } k \text{ at } y \text{ years into the future} \end{cases}, \quad k = 1, 2, 3$$

The outcomes are commissioning ($k = 1$), attriting ($k = 2$) or continuing ($k = 3$). We note that the three variables sum to one for each individual because the outcomes are mutually exclusive and exhaustive. The total number of cadets who have outcome k at y fiscal years from the current time, which we denote $S_{k,y}$, is the sum of random variables for that outcome across the entire population.

$$S_{k,y} = \sum_{t=1}^7 \sum_{i=1}^{n_t} X_{k,t,y,i}, \quad k = 1, 2, 3 \quad (2)$$

A reasonable forecast for $S_{k,y}$ is its expected value, which is found by summing the expected values of each Bernoulli random variable in the sum

$$\mu_{k,y} = E(S_{k,y}) = \sum_{t=1}^7 \sum_{i=1}^{n_t} p_{k,t,y,i}, \quad k = 1, 2, 3, \quad (3)$$

where $p_{k,t,y,i}$ is the probability that cadet i in stratum t has outcome k before the end of y fiscal years from the current time. If each cadet's disposition is independent of others, the variance of $S_{k,y}$ is given by

$$\sigma_{k,y}^2 = \text{Var}(S_{k,y}) = \sum_{t=1}^7 \sum_{i=1}^{n_t} p_{k,t,y,i} (1 - p_{k,t,y,i}), \quad k = 1, 2, 3 \quad (4)$$

If the probabilities in these expressions were known, an approximate 95% prediction interval for $S_{k,y}$ would take the form

$$\mu_{k,y} \pm 2 \cdot \sigma_{k,y}, \quad k = 1, 2, 3, \quad (5)$$

based on a normal approximation. The probabilities, however, are not known but are estimated using statistical models, which introduces additional uncertainty. Our forecasts take the form

$$\hat{\mu}_{k,y} = \sum_{t=1}^7 \sum_{i=1}^{n_t} \hat{p}_{k,t,y,i}, \quad k = 1, 2, 3 \quad (6)$$

to reflect the use of estimates. It is important to understand that for each stratum t a separate model is used to obtain probability estimates, using data from cadets across all years that share the same profile (t, y values). Because the same model is used to estimate all probabilities in a stratum, its probability estimates are correlated. And, because cadets contribute data to several models as they progress along their scholastic paths, probability estimates are correlated across models. As a result, a naïve prediction interval (4) that replaces $p_{k,t,y,i}$ with estimates $\hat{p}_{k,t,y,i}$ in (3) cannot be regarded as reliable.

One final note of importance is that our model produces estimates for the total number of cadets currently enrolled who will commission between the snapshot date and the end of the fiscal year for which a prediction is being made. Because AROTC receives a commission mission for each fiscal year, our estimate should quantify the number of cadets projected to commission in the target fiscal year rather than the number expected to commission between the snapshot and end of the target year. In order to provide the fiscal year estimate, we build models to forecast two fiscal years into the future as well as one

fiscal year into the future. The difference of the two estimates is the number of currently enrolled cadets expected to commission in the target fiscal year.

4. Bootstrapping Random Forest Models to Determine Variability

Our approach to producing prediction intervals is to use the bootstrap. The bootstrap method is a means by which the sampling properties of estimators can be ascertained by iteratively resampling the data with replacement (Devore 2016, p. 260). In particular, we can use the bootstrap to estimate the standard errors of our forecasts. In a bootstrap sample we randomly select samples of cadets from each entry year with replacement, and as many as there are in the samples that are used to produce the estimated stratum models. After the samples are taken, the stratum models are estimated using the bootstrap samples, and a new set of estimated probabilities are obtained for the target population. An individual cadet i in stratum t has three estimated probabilities $(\hat{p}_{1,t,y,i}, \hat{p}_{2,t,y,i}, \hat{p}_{3,t,y,i})$ for commissioning, attriting, and continuing respectively. We use these probabilities to generate a single realization of a multinomial random variable for that individual which we denote $(X_{1,t,y,i}^{(b)}, X_{2,t,y,i}^{(b)}, X_{3,t,y,i}^{(b)})$, of which exactly one element is equal to one and the others equal to zero. Summing across all individuals in the target population, this produces one set of realizations of the population totals:

$$S_{k,y}^{(b)} = \sum_{t=1}^7 \sum_{i=1}^{n_t} X_{k,t,y,i}^{(b)}, \quad k=1,2,3 \quad (7)$$

By repeating this process many times ($b=1, \dots, B$ where B is large) a sense of variability in the forecasted population totals may be obtained. Importantly, this variability comes both from the inherent uncertainty of individual outcomes and from the estimation of probabilities.

There are two main approaches to obtaining prediction intervals using the bootstrapped quantities: the quantile method, and the normal-approximation method. With the quantile method, the sample 2.5th and 97.5th sample percentiles constitute an approximate 95% prediction interval for the population total of interest. This method usually is preferred but it requires many bootstrap replications (typically $B=1000$ or

larger) in order to estimate the quantiles reliably. With the normal-approximation method, the sample standard deviations of the bootstrap samples ($\hat{\sigma}_{k,y}^{\text{boot}}$) are used with the original point estimates ($\hat{\mu}_{k,y}$) to produce intervals of the form

$$\hat{\mu}_{k,y} \pm 2 \cdot \hat{\sigma}_{k,y}^{\text{boot}}, \quad k = 1, 2, 3. \quad (8)$$

The normal approximation requires many fewer bootstrap samples than the quantile method (usually $B = 100$ or 200 is sufficient) but it requires an additional assumption due to its reliance on the validity of the normal approximation. Figure 4 is a histogram of commission estimates derived from 1,000 bootstrapped samples derived using our methodology and overlaid with a normal curve. It shows our model estimates are approximately normally distributed.

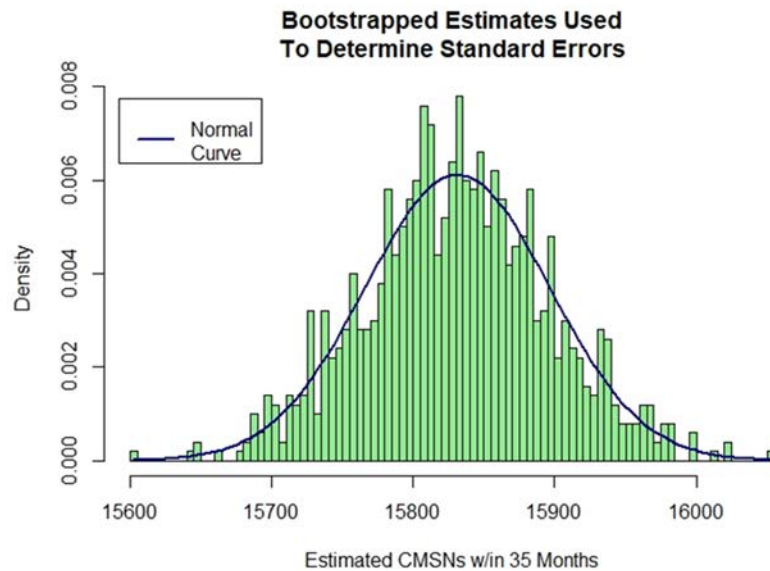


Figure 4. Histogram of Bootstrapped Commission Estimates for the Mission Set 2021 Cohort

We have found that the normal-approximation method ($B = 200$) produces prediction intervals that are almost identical to those produced by the quantile method ($B = 1000$) in about one-fifth of the computation time. Because each bootstrap sample

requires the re-estimation of several probability models each of which is computationally intensive, we recommend the use of the normal-approximation method for constructing prediction intervals.

Finally, we address how the bootstrap may be used to produce prediction intervals for changes in numbers of cadets associated with an outcome group between two successive years. Let $\delta_{k,y} = \mu_{k,y} - \mu_{k,y-1}$ denote the change in the number of cadets in the target population belonging to outcome category k looking y years ahead from the number for the previous year. We expect to see positive values of $\delta_{k,y}$ for $k=1,2$ (commissioning and attriting, respectively) and negative values for $k=3$ (continuing). We estimate this change using

$$\hat{\delta}_{k,y} = \hat{\mu}_{k,y} - \hat{\mu}_{k,y-1} \quad (9)$$

which is the difference of sums of probability estimates from two separate models. To estimate a prediction standard error we again use bootstrapping to produce simulated realizations $\hat{\delta}_{k,y}^{(b)}, b=1, \dots, B=200$ where again sampling is stratified by entry year, and an additional entry year is available for obtaining $\hat{\mu}_{k,y-1}$.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS

In this chapter, we discuss the results of our models and analysis conducted based on model predictions.

A. INDIVIDUAL MODEL PERFORMANCE

We use misclassification rates to compare the performance of each of our random forest models and classification tree models. A cadet is predicted to belong to the outcome category (commission, attrite, or continue) that has the highest estimated probability. The misclassification rate for each model is given in Table 14. Note that the two model types perform similarly regardless of the subset of cadets on which the models are trained.

Table 14. Comparison of Random Forest and Classification Tree Model Performance (Misclassification Rates)

	Random Forest Model	Classification Tree Model
0-6 months	0.274	0.269
6-18 months	0.245	0.251
18-30 months	0.142	0.145
30-42 months	0.086	0.086
42-56 months	0.121	0.112

Both random forest models and classification trees perform best for cadets enrolled two to four years. Lower misclassification rates indicate more accurate models.

It is an accepted practice to choose classification trees over random forests when the models perform similarly. This is due to the interpretability advantages that a classification tree offers. But because we suggest the future addition of predictors such as GPA, Army Physical Fitness Test score, standardized test scores and similar numeric variables, we have chosen to use the random forest models for our forecasts.

The models used in our research seem reasonable and useful for USACC because they are designed to forecast the number of current cadets who will commission two fiscal years after the current year. Effectively, we forecast commissions nearly three years from

an opening enrollment snapshot, which is the input to the models. Since current models forecast commissions 24 months ahead of time, our models increase the forecasting ability to 35 months, with a prediction standard error of about 77 commissions.

It is important to note that our model predicts only the number of currently enrolled cadets who will commission in the selected fiscal year. We do not model commissions of cadets who will enroll in AROTC between the enrollment snapshot date used as input to the model and the selected fiscal year. For that reason, the estimate is reasonably lower than AROTC's commission mission for the same cohort. The difference between the commission mission and our estimate is the number of new cadets AROTC should plan to recruit between the opening enrollment snapshot and the beginning of the junior year of the target cohort.

B. MODEL VERIFICATION

Our models are challenging to verify for a number of reasons. Most prominently, model verification should be performed on data that has not been used to train the models. Ideally, we would hold out an entire snapshot from our data to verify the outputs of our models with the true disposition of cadets. Due to right censoring in our data, we are limited to training our models using the first three years of our data, and model accuracy is diminished if the training data are reduced further. However, we verify the accuracy of our methodology across school years by training our models and then testing them against a holdout set for each school year. This practice allows us to assess whether our models are better at predicting specific years or whether they may be applied broadly to any school year in our data sets.

Figure 5 shows the predicted estimates for each of our multinomial outcomes and compares them to the actual numbers in our holdout sets, stratified by school year. The predicted outcomes of enrolled cadets in the school years listed are similar to the actual number of cadets in each category. Beginning in school year 2016–2017 (SY16–17), there are large numbers of cadets whose final dispositions are unknown, which reduces the accuracy of the models if data in these years is used to train them. For this reason, data for SY16–17 and later years are not used to train the models.

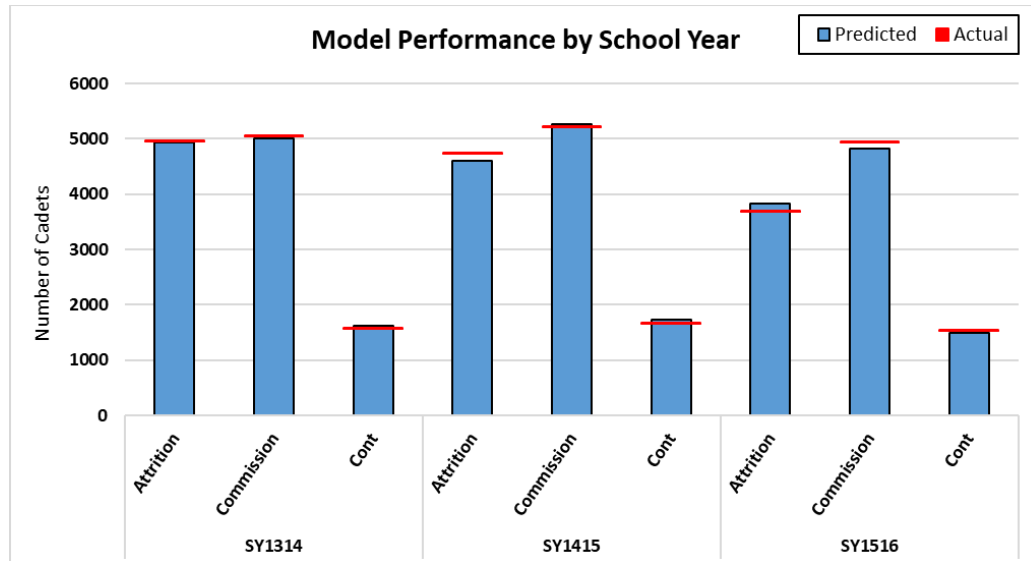


Figure 5. Model Performance Stratified by School Year

A table of the model performance by outcome and school year is given in Table 15. The percentages shown are the relative errors observed for each SY used to train our models. For example, for SY13–14 the total number of enrolled cadets that our model predicts will attrite is 4,941, while the actual number of enrolled cadets who attrite is 4,958. Therefore, our model for SY13–14 attrition under estimates by just 17 cadets.

Table 15. Model Forecast Accuracy by School Year and Outcome

		Predicted	Actual	Rel Error
SY1314	Attrition	4941	4958	-0.3%
	Commission	5006	5053	-0.9%
	Cont	1629	1566	4.0%
SY1415	Attrition	4600	4731	-2.8%
	Commission	5258	5201	1.1%
	Cont	1729	1655	4.5%
SY1516	Attrition	3821	3690	3.6%
	Commission	4831	4928	-2.0%
	Cont	1497	1531	-2.2%

The model forecasts the number of commissions with at least 98% accuracy in each of the school years given. Additionally, differences in the data between school years does not appear to affect the accuracy of our methodology. This fact is evident in how no individual school year forecast seems to be substantially easier or more difficult to predict.

C. VARIABLE IMPORTANCE

A natural analytical question to ask when modeling is which variables have the greatest effect on the outcome. We previously discussed how the Gini index is a measure of deviance for classification trees, and how random forests are based on many trees created from randomly sampled predictor variables. For random forest models, we measure the importance of a variable by how much node purity improves when the tree is split using that variable over the trees that comprise the forest. The ten most important variables for each of our random forest models are listed in Table 16.

Table 16. Variable Importance for the Random Forest Models

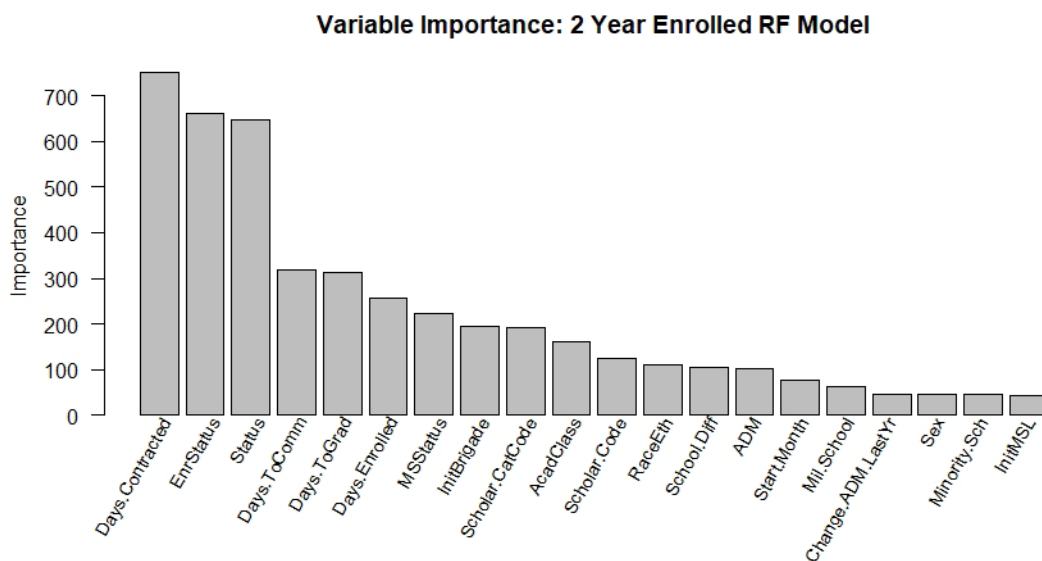
Importance Rank	0 Year Model (0-6 months)	1 Year Model (6-18 months)	2 Year Model (18-30 months)	3 Year Model (30-42 months)	4 Year Model (42-56 months)
1st	Days to Comm	Days Contracted	Days Contracted	Days Contracted	Days Contracted
2nd	Days to Grad	Status	Enrollment Status	Enrollment Status	Enrollment Status
3rd	Days Enrolled	Days to Comm	Status	Status	Days to Comm
4th	Status	Enrollment Status	Days to Comm	Days to Comm	Status
5th	Scholar Code	Days to Grad	Days to Grad	Days to Grad	Days to Grad
6th	Scholar Cat Code	Days Enrolled	Days Enrolled	Academic Class	Days Enrolled
7th	Enrollment Status	Scholar Cat Code	MS Status	Days Enrolled	Initial BDE
8th	Initial BDE	Scholar Code	Initial BDE	MS Status	MS Status
9th	Initial MSL	MS Status	Scholar Cat Code	Initial BDE	Academic Class
10th	Days Contracted	Initial BDE	Academic Class	Scholar Cat Code	Scholar Cat Code

Unsurprisingly, several variables are important in every model. These variables include the number of days contracted or enrolled, the number of days until the estimated commission and graduation dates, the enrollment status code or the general status category (prospect, conditional, contracted), initial brigade, and the scholarship type (scholar category code). Variables such as initial MSL and scholarship code (length of scholarship) are also important for predicting the status of those who have been enrolled fewer than six months, but are less important in the other models. Academic class, on the other hand, seems to be helpful for predicting those who have been enrolled about three or more years, but is not as useful for cadets enrolled only a short time.

The four time-related variables, which we derive from the data, each prove helpful for prediction. We believe that this is due to the relative robustness of these variables as predictors. By that, we mean that if there is some uncertainty about a graduation or

commission date for a cadet, the estimated date may still prove useful for prediction as long as the final date is reasonably close to the estimated date. Commission and graduation dates are somewhat less ambiguous than the academic class or military science level, and provide our models a tangible means of estimating the likelihood of commissioning within a timeframe. For example, if a cadet is an academic freshman, but has an estimated commission date one year out, a trained model should be able to adjust the probability of commissioning despite the discrepancy.

The predictor variables shown in Table 16 does not indicate the degree to which one predictor is more important than another. Figure 6 shows the relative importance of the variables for cadets enrolled two years.



The number of days contracted, enrollment status and the derived *status* variable contribute most to the ability to predict the status of cadets enrolled two years than do other predictors.

Figure 6. Variable Importance for Cadets Enrolled Two Years

There is less parity between variables in models utilizing cadets with a longer enrollment history. By that, we mean that the models trained with cadets who have been enrolled for longer periods of time each have a few variables that influence model outcomes greatly, while the model based on cadets enrolled fewer than six months, for example, had much more parity. The difference in the number of highly important variables is expected

for several reasons. First, the population of cadets enrolled fewer than six months is very diverse, and includes cadets on two-year scholarships who are expected to commission in less than two years, as well as cadets who are academic freshmen with no ambition to join the Army. Therefore, different variables may prove important for each of the two cadets described. Models trained based on cadets enrolled three or more years, however, contain mostly contracted cadets. In such cases, we expect to see similar variables impact the likelihood of commissioning.

D. MISSION SET 2021 COMMISSION ESTIMATE

The main objective of our study is to propose a means for estimating the number of commissions earlier in a cohort than AROTC currently has the ability to forecast. We make our commission forecast for the cohort of cadets that is scheduled to commission in 2021 (MS21). Cadet observations in the opening enrollment data from November 2018, are subset by length of time enrolled and used as inputs to each of our five random forest models. For cadets enrolled longer than 54 months, we use average commission rates to forecast the number that will commission. The total number of currently enrolled cadets estimated to commission in MS21 is 4,492, with a standard deviation of 77 commissions. Using formula (2) from Chapter III, the 95% prediction interval is given in Table 17.

Table 17. Mission Set 2021 Commissioning Population Forecast Based on November 2018 Enrollment Data

	Attrite	Commission	Continue
Within Two Years	10,884 (10,743 , 11,026)	15,833 (15,713 , 15,954)	5,464 (5,357 , 5,582)
Within One Year	8,996 (8,865 , 9,128)	11,348 (11,255 , 11,443)	11,835 (11,718 , 11,966)
FY21 Estimate	1,888 (1,701 , 2,074)	4,485 (4,334 , 4,636)	

Numbers in parentheses are 95% prediction interval bounds. FY21 predictions for “Continue” are not given since the difference between the two-year and one-year estimates are accounted for in the “Attrite” and “Commission” estimate.

The 95% prediction interval suggests that after running our model, we expect that 19 times out of 20 the interval provided would contain the actual number of currently enrolled cadets who will commission.

Given that the commission mission for USACC in FY21 is 6,000 total cadets, the implication of this estimate is that USACC must add enough new cadets to the 2021 cohort to achieve approximately 1,515 additional commissions in FY21. The additional commissions must come from cadets not currently enrolled in AROTC. Many of the students USACC needs to recruit will not be eligible to enroll in the advanced course (MSL 3 and MSL4) without credit for the basic course. Cadets needing basic course credit can attend the basic camp, as discussed in Chapter II, to earn credit and qualify to enroll in the advanced course. Therefore, our estimate also informs USACC regarding the potential number of cadets that can be expected to attend the basic camp.

E. PROFILE OF HIGHLY LIKELY ATTRITERS

Cadets who contract early in their college careers are those whom AROTC deems worthy of significant investment, since the Army will develop them and provide benefits for several years. We consider cadets with more than 50% probability of attriting from AROTC within two years to be likely attriters, and classify them according to our model outputs. We focus primarily on the population of cadets who have been enrolled two years or less.

When examining the population of cadets enrolled fewer than six months, it is clear that most of the cadets we classify as likely to attrite are prospects (97% of 25,634 cadets). This is not surprising, as most prospects do not wish to join the Army. When we eliminate prospects from our analysis and focus on cadets who are conditionally contracted or contracted, we observe that conditional cadets make up the vast majority of this group (96.8% of 777 cadets). More than half of this same subset of cadets (51.6% of the 777) have been awarded four-year scholarships. This demonstrates that the models identify the fact that a cadet on a four-year scholarship, who is still in a conditional status, is likelier to attrite. Scholarship recipients who have not resolved their conditional status, perhaps

because they are awaiting waivers or medical eligibility determinations, are at a higher risk of attrition.

Cadets with one year of enrollment are generally entering their second year of AROTC. Many of these cadets are still classified as prospects, and once again we see that prospects are highly likely to attrite (96.1% of over 12,200 cadets whom our models deem highly likely to attrite). When we focus on the contracted and conditionally contracted population of one-year enrolled cadets, we continue to see that conditionally contracted cadets make up the majority of the remaining likely attriters. Over 94% of the 463 cadets who are not prospects, and who our models considered highly likely to attrite, are conditionally contracted. The remaining 27 of 463 cadets are contracted. While this is a small number for analytical exploration, we observe that 15 of the 27 (56%) contracted cadets are on a leave of absence, which the model identifies as an indicator of attrition.

The trend of prospects occupying the largest proportion of likely attriters continues in the two-year enrolled population. Nearly 87% of the 2,956 likely-to-attrite cadets enrolled two years are prospects. When we examine only contracted and conditionally contracted cadets in this population, we find nearly 62% of the remaining 381 cadets to be in a conditional status. Most of the remaining cadets, who are fully contracted, seem likely to attrite because they are in a leave of absence status. Of the 381 contracted cadets classified as probable attriters 130 are on leave, and the two-year model identifies this as an indicator of attrition. In all, 96% of the two-year enrolled likely-to-attrite population are either not fully contracted or are on a leave of absence.

Figure 7 highlights the fact that cadets conditionally contracted still have very uncertain futures in the AROTC program. Note that cadets who have been enrolled two years are entering their third year of AROTC and would be expected to commission within two years under most circumstances. The histogram on the left shows that most contracted cadets have smaller probabilities of attriting, whereas the models assign higher probabilities to cadets whose statuses remain conditional, pending some administrative action or waiver.

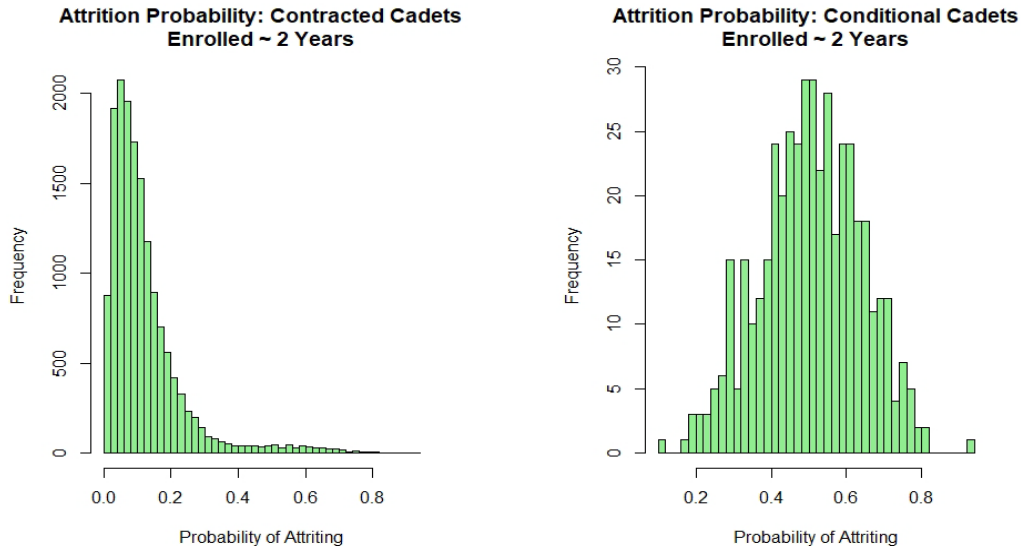


Figure 7. Contracted vs. Conditionally Contracted Attrition Probability, Enrolled Two Years

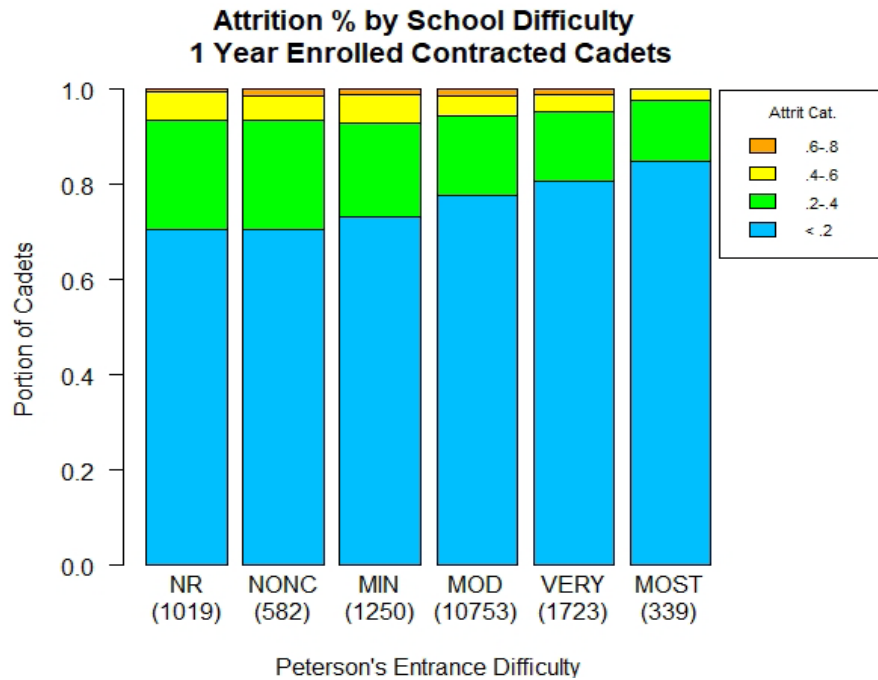
When we consider the observations above and explore the true attrition rates of cadets, we see the same trends. Note that above we described the populations of high attriters, those who our models assessed more than 50% chance of attriting. Table 18 summarizes the actual attrition rates of cadets by status and time enrolled in AROTC.

Table 18. Actual Attrition Rates, by Cadet Status

Enrollment Status	Years Enrolled				
	0	1	2	3	4
Prospects	57.1%	56.3%	64.1%	83.4%	90.1%
Conditional	25.1%	31.1%	37.7%	44.4%	56.7%
Contracted	11.4%	8.8%	8.0%	6.7%	12.4%

The outcomes in Table 18 reflect whether or not a cadet actually attrited within two fiscal years after being observed in our data. Our models distinguish cadets that are prospects or conditionally contracted as being more likely to attrite than contracted cadets. Additionally, while the attrition rates are consistently low for contracted cadets no matter how much time they have spent enrolled, our models are able to identify factors that lead to a higher risk of attrition for these cadets—namely, whether or not a contracted cadet is on leave of absence.

Another observation regarding attrition is that our models estimate slightly higher attrition probabilities for less competitive schools, as categorized by Peterson's Database. Contracted cadets generally have low attrition probabilities, but contracted cadets at schools categorized as non-competitive, minimally competitive, or that are not rated have slightly higher percentages of cadets that the models deem likely to attrite. We categorized cadets into five attrition categories based on the estimated attrition probabilities of our models: a less than 20% chance of attriting, a 20–40% a chance, a 40–60% chance, a 60–80% chance, or an 80% or better chance of attriting. This schema allows us to broadly compare the distribution of cadets based on attrition likelihood. Figure 8 shows the distribution of cadets between attrition categories by school difficulty.



Less competitive schools had smaller proportions of cadets whose attrition probability was very low. Labels are NR (not rated), NONC (non-competitive), MIN (minimally competitive), MOD (moderately competitive), VERY (very competitive), and MOST (most competitive).

Figure 8. Attrition Categories by School Difficulty,
Contracted Cadets Enrolled One Year

As contracted cadets spend more time in the AROTC program, the models show a reduced likelihood of attrition. This same observation occurred in contracted cadet populations enrolled two and three years. The blue sections of the stacked bars represent the portion of cadets with less than 20% chance of attriting. Non-rated, non-competitive and minimally competitive schools each had a smaller proportion of cadets in this low-attrition category. While the differences may seem small, the smaller proportions of less likely attriters show up in the one-, two- and three-year enrolled populations, and is further seen in Figure 9.

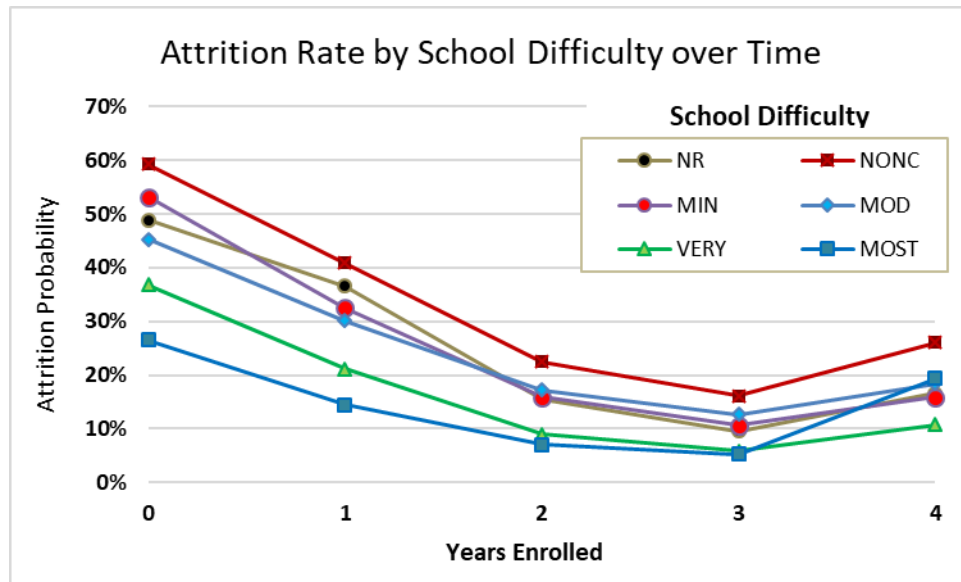


Figure 9. Actual Attrition Rate by School Difficulty over Time

Figure 9 shows that the true attrition rate of cadets enrolled in less competitive academic institutions is higher for all values of years enrolled. The only break from this pattern is the most difficult academic category for four-year enrolled cadets. The actual attrition rate for schools in the most difficult category rises to 19.3% in the four-year enrolled population from 5.2% in the three-year enrolled population. This increase indicates that four-year enrolled cadets at the most difficult institutions may have a reason to exit the AROTC program, or that they have a higher propensity to attrite.

In summary, our models successfully identify important indicators of attrition. The models identify major differences in attrition rates between non-contracted, conditionally

contracted and fully contracted cadets, as well as key combinations of enrollment statuses and contracting status (contracted cadets on leave of absence, for example). Additionally, smaller factors such as academic difficulty influence the attrition probability estimates of our models.

F. PROFILE OF CADETS WITH HIGH COMMISSION PROBABILITIES

Commissioning cadets is the ultimate goal of the AROTC program. In this section, we examine characteristics of cadets deemed by our models to be likely to commission. We call these individuals high commissioners.

Analysis of newly enrolled cadets shows that high commissioners from those enrolled less than six months are largely those starting in MSL 3. Nearly 1,900 of 2,622 (73%) probably commissioners from the early enrollees entered the program at the MSL 3 level, and only 10 of the 2,622 cadets started the program prior to MSL 2. Unsurprisingly, early enrollees who are also high commissioners are composed mostly of contracted cadets (nearly 81%). However, just over 11% of this population is made up of conditionally contracted cadets, and 8% of those likely to commission are prospects. The fact that the models assign a high commissioning probability to any newly enrolled prospects is intriguing, as this is not expected. A closer examination of this population shows that nearly 95% (200 of 211) of the prospects deemed likely to commission are not freshmen, and are also enrolled in MSL 2. This may indicate that new sophomore cadets (or upper classmen) are valuable students to recruit, as they are more likely than freshmen to remain in AROTC.

An analysis of the one-year enrolled population shows that once again cadets who are contracted are highly likely to commission. Among the high commissioners in this group, over 94% are either on scholarship, contracted with no scholarship (but still receiving a monthly stipend), or in the green-to-gold program (earning a salary). It also is interesting to observe that nearly no cadets with high commission probability estimates are in a leave of absence status, and almost none of these cadets are prospects at MJCs or SMCs.

The population of cadets enrolled two years is entering their third year of AROTC. Consequently, we expect these cadets to commission within two years, and therefore that

their commission probabilities will be high. Figure 10 shows this to be true, as the vast majority of cadets in the entire two-year enrolled population (left) have high commission probabilities. The histogram on the right shows how the majority of these cadets are on scholarship. Scholarships seem to raise the likelihood of a cadet commissioning according to our models.

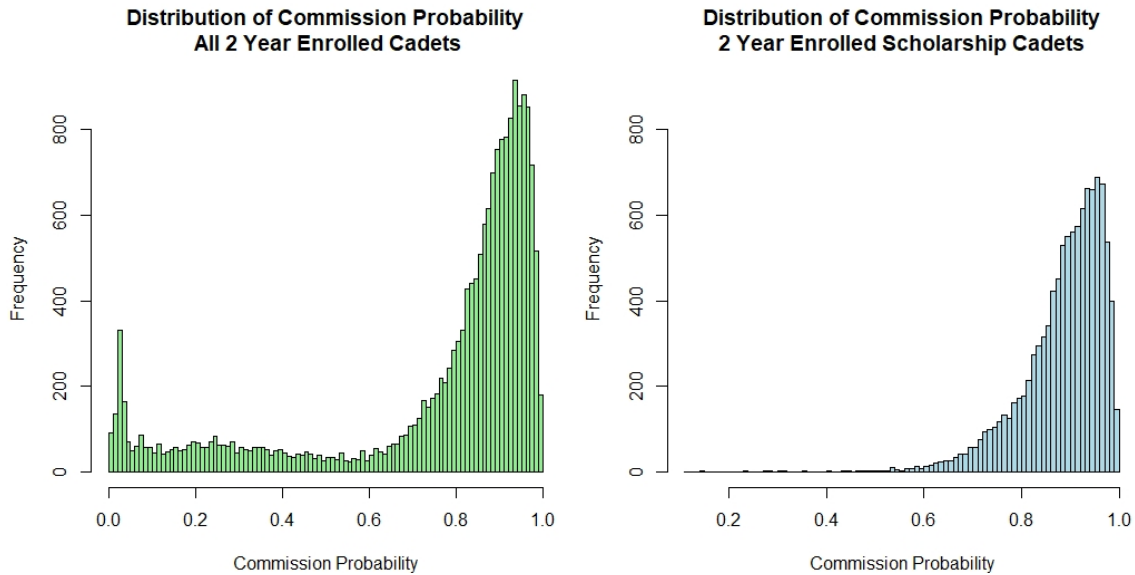


Figure 10. Comparison of Commission Probability, All vs. Scholarship Cadets, Two-Year Enrolled

When comparing the behavior of our models with reality, we see similar trends in the actual commissioning percentages between those cadets who are prospects versus those who are conditionally contracted or who are on scholarship. Figure 11 shows the actual commission rates by contract status and length of enrollment. Note that the actual commission outcomes are based on the known outcomes of cadets in our dataset as of January 2019. Due to some cadets' dispositions being unknown at the time of the study, the true commission rates are potentially slightly higher than shown. Nonetheless, the fact that contracted (non-scholarship), and contracted (with scholarship) cadets commission at a much higher rate than conditionally contracted cadets and prospects is apparent.

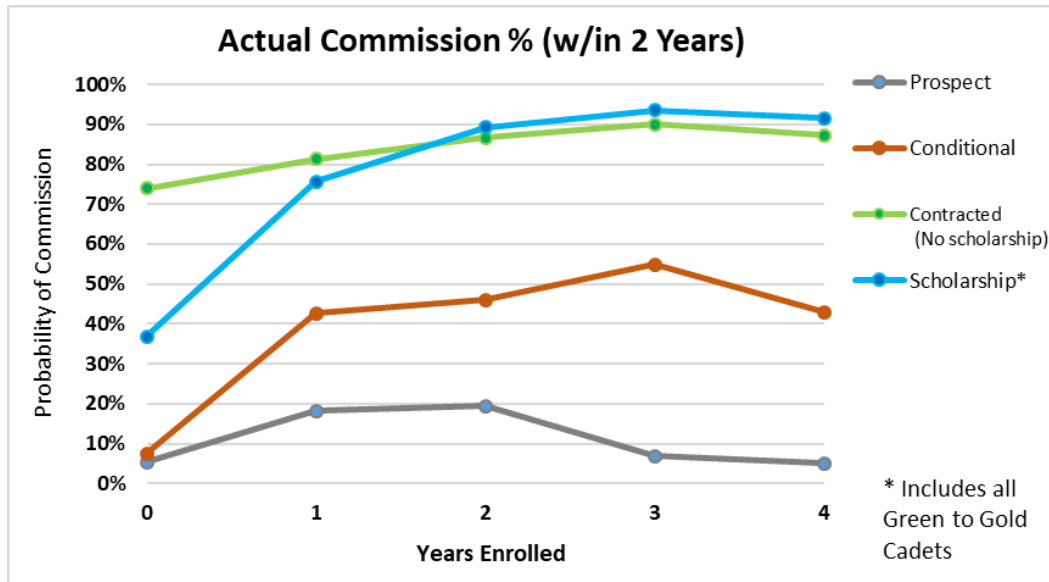


Figure 11. Actual Commission Percentages (within Two Years) by Years Enrolled and Status

A peculiar observation from Figure 11, in the population of cadets newly enrolled, is the fact that non-scholarship contracted cadets are more likely to commission than are scholarship cadets. This is likely due to the fact that most scholarship cadets (over 60% of 4,288 cadets in our data) who have been enrolled less than six months possess three- or four-year scholarships. Three- and four-year scholarship cadets who have just enrolled are not likely to commission within two years.

Another insight from our models is that the probability of a cadet commissioning is related to scholarship type. We have shown how scholarships are an important tool for USACC in that they are highly associated with successful commissions. AROTC offers a variety of scholarships. Using the commission probability outputs of our models, we examine the distributions of the probabilities for contracted cadets enrolled three years using boxplots, as shown in Figure 12. The scholarship category codes are displayed on the x-axis. Readers may wish to reference Table 3 in Chapter III to identify what each scholarship code means. Additionally, the number beneath the scholarship category is the sample size.

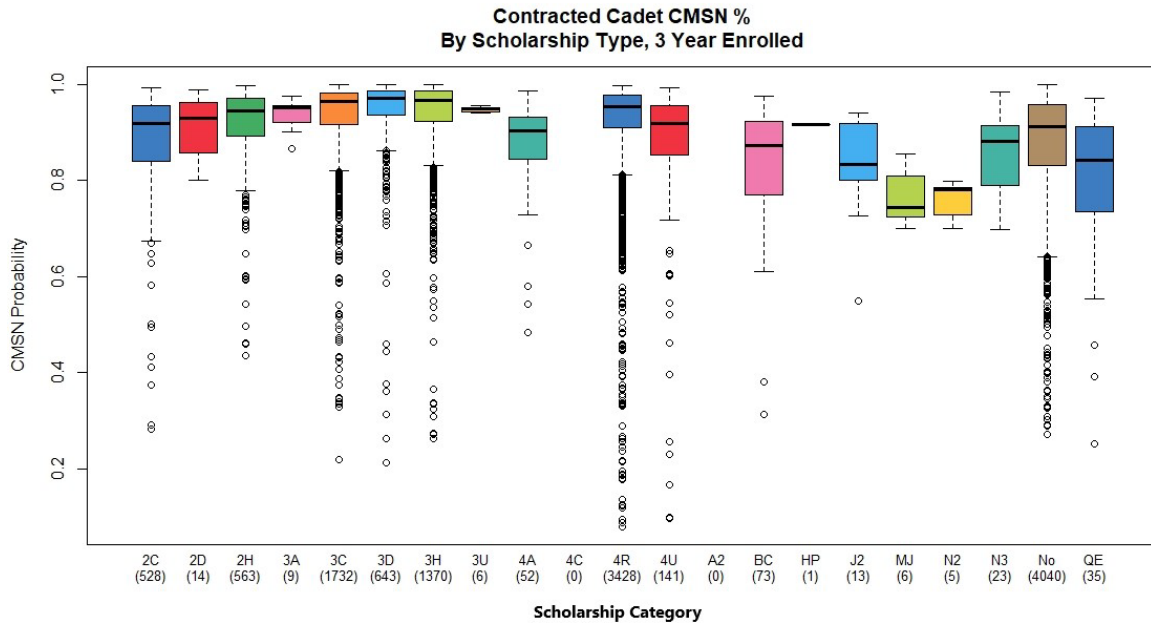


Figure 12. Boxplots of Contracted Cadet Commission Probabilities, Enrolled Three Years

As expected, the mean commission probability in nearly every scholarship category is fairly high. Because cadets enrolled three years are entering their fourth year of AROTC, most cadets should commission within two years. Aside from categories with a small number of observations (2D, 3A, 3U, and HP, for example), there are a few insights to be gained from these boxplots. In terms of forecasting, we observe that some scholarship types have more variability in the estimated probabilities than others. For example, BC and QE, a basic camp scholarship for cadets entering AROTC just prior to their junior year of college, and a four-year HBCU scholarship, respectively, each have wider commission probability distributions than some of the other categories. Category 3D, which is a three-year scholarship awarded to high school seniors, has a high mean commission probability and relatively little variability, as indicated by the small light blue box.

We also find more variability among non-scholarship cadets, labeled “No.” Despite over 4,000 observations, the brown non-scholarship boxplot is fairly wide and has long tails for the inner quartile range (IQR), which is the middle 50% of commission probability estimates for that group.

Boxplots also provide insight regarding specific scholarship types. For example, we analyze outliers for scholarship category codes 4R (four-year national scholarship winners), identifying all observations where our models estimate a commission probability less than 30%. Of the 22 observations, 14 of these cadets are in a leave of absence status. The eight cadets who are not on leave are identified as being either freshmen or sophomores and enrolled in MSL 1 or 2. The model is able to associate a lower probability of commissioning to these cadets since their academic and MSL statuses do not match the expected academic (junior) or MSL (3) values for cadets enrolled for three years.

We used the same methodology to analyze outliers in the non-scholarship population from Figure 12. Most of the outliers for the non-scholarship group have a commission probability below 60%. When we review information for cadets in this population, 19 of the 29 cadets are in a leave of absence status, and eight of the remaining 10 cadets have each passed their graduation date. The additional two cadets are listed as being academic sophomores (academic class 2), which is not the expected academic level of a cadet who has been enrolled for four years. Table 19 shows the actual commission rates by academic class for cadets enrolled three years.

Table 19. Actual Commission Rates by Academic Class Code,
Cadets Enrolled Three Years

Academic Class	Actual CMSN %
1	20.0%
2	36.9%
3	59.6%
4	92.9%
5	87.5%
Post-Grad	86.0%

The fact that cadets who have been enrolled for three years are not generally academic freshmen or sophomores is identified in the model. There is a clear difference in actual attrition rates between cadets listed in varying academic classes. Overall, the three-year model adjusts for academic class quite reasonably. When looking at the commission likelihood of all cadets in the three-year enrolled population, for example, only 2.1% of

cadets listed as academic freshmen are assigned commission probabilities of more than 80%. Over 81% of all cadets enrolled three years who were also indicated to be freshmen were assigned commission probabilities below 40%. On the other hand, cadets who are seniors or above are assigned high commission probabilities (above 80%) 92% of the time (10,022 cadets out of 10,162).

Because the goal of AROTC is to recruit future Army officers, and because the goal of our study is to forecast commissions, it is important that our models accurately assign probabilities based on cadet data. We find that our models identify factors, such as the differences between prospects, conditionally contracted cadets, and contracted cadets, that influence commission likelihood. Our models also pick up differences within these categories, such as the fact that contracted cadets on scholarship are more likely to commission than contracted cadets without a scholarship. Finally, our models are able to accurately identify outliers within the data, and they adjust commission probabilities accordingly.

G. EFFECTS OF MIGRATION BEHAVIOR

Many circumstances cause students in an academic environment to delay graduation or commissioning. When a cadet's commission date changes from one fiscal year to another, the change is termed a "migration." Migrations make commission forecasting more challenging and may reduce USACC's ability to achieve its mission. We seek to answer questions about the number and type of migrations that occur each year. Understanding migration is made difficult by the fact that the expected commission date changes each year for many cadets. Because a complete study of migrations lends itself to entirely new models, we focus on summary statistics from our sample data to gain insight about the issue. An important note, as it pertains to migrations, is that when we state that a cadet is a migrator, we mean that we have already observed a change in the cadet's estimated commission date.

The data set provided for analysis limits our ability to assess the true number of migrations over time because many cadets take four or more years to commission. With only six snapshots of data and with right-censored outcomes and left-censored commission

dates, there are many cadets whose final migration status is not known. In a given snapshot, we observe that between 80% and 86% of cadets do not migrate. In each snapshot, between 10.7% and 14.7% of cadets migrate into the next (later) cohort, and between 2.3% and 3% of cadets migrate into the previous (earlier) cohort. Additionally, between 1.7% and 2.5% of cadets commissioned two or more years after originally planned.

There does not seem to be parity between schools of various academic entrance standards when considering migration of cadets. Schools described by Peterson’s Database as non-competitive, minimally competitive, or not rated were each over-represented in the population of cadets migrating to later cohorts, as shown in Table 20. Cadets enrolled at the less competitive schools also attrited at a higher rate, as discussed previously in this chapter.

Table 20. Proportion of Cadets by School Difficulty,
All Cadets vs. Migrators

	<u>Population</u> <u>Migrators</u>	
NR	7.1%	8.6%
NONC	3.6%	5.4%
MIN	8.2%	9.2%
MOD	69.6%	68.6%
VERY	9.7%	7.4%
<u>MOST</u>	<u>1.7%</u>	<u>0.9%</u>

The differences in proportions of the migrator population and the population at large may seem trivial. But when we visualize the effect that migrations have on attrition likelihood by comparing the distribution of attrition probabilities from our models between groups of migrators, we see that cadets who have migrated into later cohorts are more likely to attrite. We classify migrators as having commissioned “early,” not having migrated (None), migrated into the cohort one fiscal year later than initially planned (1), or having migrated two or more years later than planned (2+). For example, Figure 13 shows that cadets who migrate one fiscal year later (move their estimated commission date one year later than initially planned), and cadets who migrate two or more years, are assigned much higher attrition probabilities in our models. Note that the portion of cadets with more than 80% chance of attriting is sparse, and not visible to the eye, but six such cadets exist.

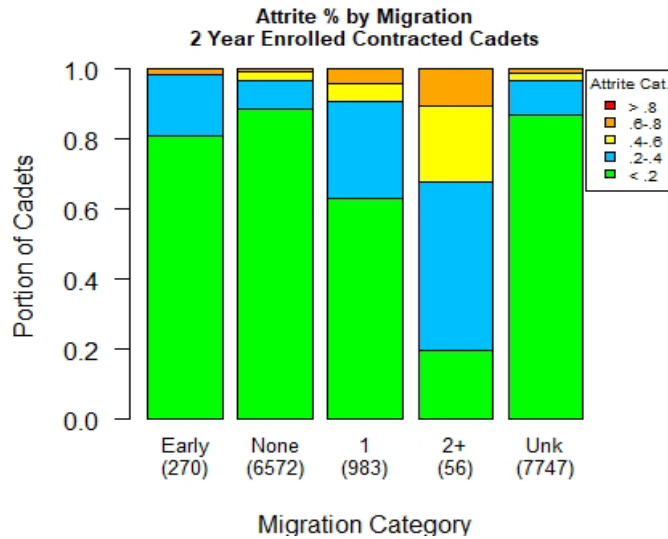


Figure 13. Comparison of Estimated Attrition Probabilities by Migration Category, Two-Year Enrollees

The cadets who have already migrated one cohort later are assigned slightly higher attrition probabilities. Cadets who have migrated two or more years have a substantially larger proportion of cadets to whom our models assign high attrition probabilities. This same trend is observed no matter how much time cadets have been enrolled, and is most prominent in the one-year enrolled population, shown in Figure 14.

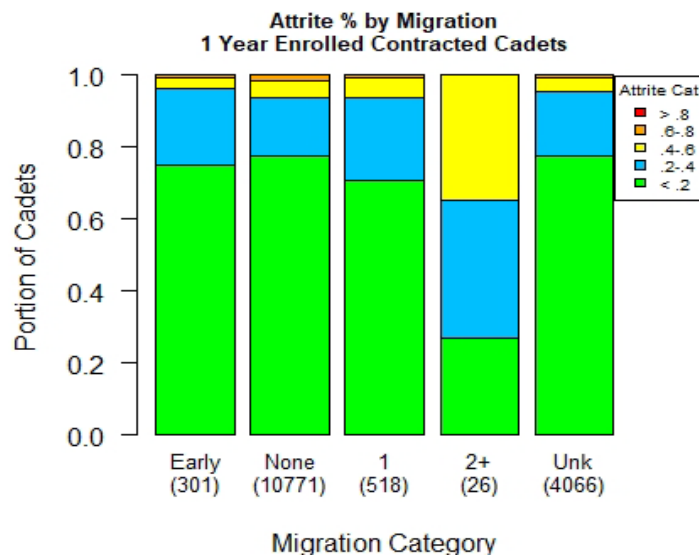


Figure 14. Comparison of Estimated Attrition Probabilities by Migration Category, One-Year Enrollees

Migrations clearly affect the likelihood of attriting within two fiscal years. An interesting observation when comparing Figure 13 with Figure 14 is that the effect of migration on attrition likelihoods seems more extreme when the migration occurs later in the college career. If a cadet has only been enrolled one year, and the cadet's commission date has already changed, our models assign less attrition probability than it does for cadets who have migrated, but who have been enrolled two years. Based on the insight that later migrations lead to likelier attrition, USACC may benefit from attempting to identify migrations early.

In summary, the effects of migrations are not trivial. Approximately one in five cadets migrate in a given year, based on the limited data we have to study the issue. Due to censoring in our data the number could be higher. Our models clearly indicate that migrations influence commission likelihood, which is not surprising. While our models identify migrations as increasing the attrition probability of cadets, a separate study may be warranted to explore the indicators or causes of migration.

V. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

This study demonstrates the use of machine learning models to forecast aggregate commissions of currently enrolled AROTC cadets up to 35 months out. We also demonstrate that bootstrapping can be used to establish a prediction interval around our prediction. Most importantly, the forecasts developed by our models inform key recruiting decisions by offering a target number of new contracts USACC must obtain to successfully meet mission in the fiscal year concerned.

We now address the study questions that we posed in Chapter 1:

1. Which Cadets Are Most Likely to Attrite, and What Factors Indicate that a Cadet Is Not Likely to Complete the Program?

We find that our models successfully identify characteristics of AROTC cadets that are predictive of their status (commissioning, attriting, or continuing) a specified length of time into the future. Contracted cadets are by far the least likely to attrite. Our models also identify which contracted cadets are most likely to attrite, namely those on leave of absence or whose academic or MSL does not match with their time of enrollment and contract status. Lastly, our models find slight differences in the attrition tendency of cadets from schools of varying difficulties. Specifically, less competitive academic institutions attrite at slightly higher rates, and our models do assign slightly higher probabilities to those students except in the model for cadets enrolled less than six months. For the cadets enrolled less than six months, our models reveal that the less competitive schools have a higher proportion of students who begin at the MSL 3 level of instruction, and that these cadets have slightly lower attrition probabilities.

2. Which Cadets Are Most Likely to Successfully Complete the Program and Commission?

When we examine highly likely commissioners, we find that our models assign the highest commission probabilities to contracted cadets, and especially those on scholarship. This tendency also is observed in the true commission rates. Our five models, which

forecast outcomes two years ahead and which are each based on length of time enrolled, identify differences in commission probabilities between each subset of cadets. For example, the models appropriately assign higher commission probabilities to cadets who have been enrolled two or three years, and slightly lower probabilities to cadets enrolled four years, which is also a trend observed in the true commission rates of cadets. Lastly, the models are able to identify cadets who have odd or misplaced status variables, such as being an academic freshman or sophomore despite being enrolled for four years. Identifying these discrepancies, the models adequately assign lower probabilities for these cadets.

3. What Insights can We Gain about the Cadets who Migrate from One Fiscal Year Cohort to Another?

While our data limited our ability to study migrations there are a few key insights that we can highlight. Approximately one in five cadets migrated in each of the school years available for study. Most importantly, migrations influence the likelihood of attrition considerably according to our models, and more so if the migration occurs later in a cadet's academic career. We suggest further studying the causes and indicators of attrition because it occurs so frequently.

B. POTENTIAL APPLICATIONS

The forecasts produced by our model provide USACC the ability to identify early recruiting shortfalls for a cohort. Additionally, the model affords USACC the ability to estimate throughput for the basic camp, since most new recruits will attend the training event during the summer before their junior year. The methodology we use should be applicable to other military ROTC programs as well. We have also demonstrated a unique means for determining variability in aggregated forecasting estimates by bootstrapping our models.

C. RECOMMENDATIONS FOR FURTHER RESEARCH

The accuracy of our models, and subsequently our estimates, are limited by the fact that several key academic variables were not included in the data that we analyze. The

addition of current grade point average (GPA), the most recent APFT score, whether cadets have passed height and weight requirements, and equivalent SAT scores are each likely to improve forecasting ability. Additionally, AROTC maintains a Cadet Based Enrollment Form (CBEF) score for scholarship cadets, which has been created by the Army Research Institute to predict the likelihood a cadet completes the AROTC program (Putka 2009, p. v). The incorporation of the CBEF scores for all cadets might prove useful for forecasting.

In addition to incorporating additional useful predictors, using an expanded dataset that incorporates earlier school years will improve model forecasts by eliminating left censoring. Modelers must balance the use of older data, which may have less relevance when predicting currently enrolled cadets, with the need to have complete information. We do not necessarily recommend using data from school years 2012–2013 and earlier for predictive purposes as much as we recommend using the data to determine whether or not a cadet changed majors or schools, for example. The nine variables in our study that explored the impact of changing majors, changing graduation or commission dates, or changing schools, were heavily censored due to the fact that we did not possess data, other than an enrollment date, for cadets prior to school year 2012–2013. Using data from prior years to inform predictors, and minimize left censoring, would maximize the effectiveness of all predictors in our models.

Lastly, our model is limited to forecasting commission likelihood for cadets enrolled at the time of the November opening enrollment snapshots. The models could be updated to include data from the peak enrollment period, occurring around March each year, which would provide information for last-minute recruiting decisions affecting the sophomore cohort. While our methodology provides a very early indicator of how a cohort is postured to meet the recruiting mission, adding the peak enrollment data would give USACC another useful indicator for recruiting purposes.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

- Borgerding T (2019) High school scholarship awardee numbers provided to the author via personal correspondence, April 26.
- Brieman, L (2001) Random forests. *Mach. Learn.* 45(1): 5–32.
- Brown, J (2001) *Historical Dictionary of the U.S. Army* (Greenwood Press, Westport, CT).
- Coumbe, A, Harford, L (1996) U.S. Army Cadet Command: The 10 Year History. Report. Office of the Command Historian, U.S. Army Cadet Command, Fort Monroe, VA. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a317940.pdf>
- Demirel, T (2002). A statistical analysis of officer retention in the U.S. Military. Master's thesis, Graduate School of Business and Public Policy, Naval Postgraduate School, Monterey, CA. <http://hdl.handle.net/10945/6078>.
- Department of the Army (2011) *Senior Reserve Officers' Training Corps Program: Organization, Administration, and Training (Rapid Action Revision)*. AR 145–1, Washington, DC, https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/r145_1.pdf.
- Department of the Army (2018) *Enclosure 2—Change #2 to FY18 and FY19 Officer Commission Missions (RA, USAR & ARNG)*. Memorandum, Office of the Deputy Chief of Staff, G–1, Washington, DC.
- Devore, J (2016) *Probability and statistics for engineering and the sciences, ninth edition* (Cengage Learning, Boston, MA).
- Faraway, J (2006) *Extending the Linear Model with R* (Chapman & Hall/CRC Taylor & Francis Group, Boca Raton, FL).
- Goldman, C, Mattock, M, Davis, J, McDonald, L, Orvis, B, Sollinger, J, Stone, A, Thomas, J, Tovar, N (1999) Allocating Scholarships for Army ROTC. Report MR-1069-A. RAND Corporation, Santa Monica, CA. https://www.rand.org/pubs/monograph_reports/MR1069.html.
- Haupt, A (2018a) USACC Intro to NPS. Presentation slides, October 11. U.S. Army Cadet Command.
- Haupt, A (2018b) CCIMM Data Dictionary, provide to the author via personal communication, 11 November 2018.
- Henning, C (2006) Army Officer Shortages: Background and Issues for Congress. Report for Congress RL33518, <https://fas.org/sgp/crs/natsec/RL33518.pdf>.

- Hopkinson, W (1988) An analysis and development of a production predictor model for the Army Reserve Officer Training Corps program. Master's thesis, Graduate School of Information and Policy Sciences, Naval Postgraduate School, Monterey, CA. <https://calhoun.nps.edu/handle/10945/22873>.
- Marvin N. Wright, Andreas Ziegler (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.doi:10.18637/jss.v077.i01.
- Putka, D (2009) Initial development and validation of assessments for predicting disenrollment of four-year scholarship recipients from the Reserve Officer Training Corps. Report FR-08-26, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a495510.pdf>.
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Streetzel, C (2018) Selected Marine Corps Reserve personnel status forecasting. Master's thesis, Graduate School of Operational and Information Sciences, Naval Postgraduate School, Monterey, CA. <https://calhoun.nps.edu/handle/10945/61279>
- Tatro, G (2018a) FY18 G2 Master File, provided to author via personal communication, 30 August 2018.
- Tatro, G (2018b) Host addresses and undergraduate populations, provided to author via personal communication, 30 August 2018.
- Tatro, G (2019a) USACC Commission Forecast Report (JAN 2019), provided to author via personal communication, 25 January 2019.
- Tatro, G (2019b) New USACC BDE Map OPR G5, provided to the author via personal communication, 03 April 2019.
- Therneau T, Atkinson B (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1–13. <https://CRAN.R-project.org/package=rpart>.
- The Kansas City Star (2017) After 137 years, Wentworth Military Academy in Missouri to close. Accessed April 25, 2019, <https://www.kansascity.com/news/state/missouri/article143518244.html>
- Ugurbas, U, Korkmaz, M (2015) Determinants of first-term attrition for enlisted and officer selected Marine Corps reservists. Master's thesis, Graduate School of Business and Public Policy, Naval Postgraduate School, Monterey, CA. <https://calhoun.nps.edu/handle/10945/45267>

- U.S. Army (1996) Senior Reserve Officers' Training Corps Program: Organization, Administration, and Training. AR 145–1, Washington, DC, https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/r145_1.pdf.
- U.S. Army (2018) Green to Gold. Accessed March 17, 2019, <https://www.goarmy.com/careers-and-jobs/current-and-prior-service/advance-your-career/green-to-gold/green-to-gold-active-duty.html>.
- U.S. Army Cadet Command (2016) Army ROTC Incentives Policy. CC 145–1, Fort Knox, KY, https://www.cadetcommand.army.mil/res/files/forms_policies/pamphlets/USACC%20Pamphlet%20145-1%20Reserve%20Officers%27%20Training%20Corps%20Army%20ROTC%20Incentives%20Procedures%20%2008-02-2016.pdf.
- U.S. Army Cadet Command (2018a) Our Mission. Accessed March 17, 2019, <http://www.cadetcommand.army.mil/>.
- U.S. Army Cadet Command (2018b) Army Reserve. Accessed March 17, 2019, <https://www.cadetcommand.army.mil/usar.aspx>.
- U.S. Army Cadet Command (2018c) Ways to Attend. Accessed March 19, 2019, <https://www.goarmy.com/rotc/ways-to-attend.html>.

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