2022-09

OPTIMIZING FIRST-TERM RETENTION OF SAILORS

Hong, Young S.

Monterey, CA; Naval Postgraduate School

http://hdl.handle.net/10945/71069

NAVAL
POSTGRADUATE
SCHOOL

MONTEREY, CALIFORNIA

THESIS

OPTIMIZING FIRST-TERM RETENTION OF SAILORS

by

Young S. Hong

September 2022

Thesis Advisor:	Louis Chen
Co-Advisor:	Ruriko Yoshida
Second Reader:	Samuel E. Buttrey

THIS PAGE INTENTIONALLY LEFT BLANK

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

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

4. TITLE AND SUBTITLE OPTIMIZING FIRST-TERM RETENTION OF SAILORS	**5. FUNDING NUMBERS**
6. AUTHOR(S) Young S. Hong	
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)

In many cases, a low retention rate of first-term sailors indicates that unsatisfactory sailors who are struggling to find guidance also cannot find a career path that fits them. Helping each sailor find their best fit can improve the retention rate in the Navy. The Navy recently developed the Job Opportunities In the Navy (JOIN) program to help sailors find their career paths on the Bureau of Naval Personnel Online. However, there are not enough data to support the effectiveness of JOIN.

Based on a dataset obtained from the Navy Enlisted System, we first analyze which factors correlate to sailors’ stays in the Navy. Then using the results from the first part of the analysis, we set up a probability distribution model to maximize the retention rate of enlisted sailors in the Navy. The result from this study can be used to help first-term sailors with JOIN. First, we conduct an analysis using a binomial logistic regression model and then calculate the model’s accuracy using a confusion matrix. Second, using the variables we select in the first part of our analysis, we set up an optimization model, specifically a probability distribution model, to maximize the retention rates of enlisted sailors. Our model produces a list of rates, from the highest probability of retention rate to the lowest probability for recruits.

14. SUBJECT TERMS attrition rate, retention rate, recruiting, recruit, first-term sailor	**15. NUMBER OF PAGES** 61
	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

OPTIMIZING FIRST-TERM RETENTION OF SAILORS

Young S. Hong
Lieutenant, United States Navy
BS, United States Naval Academy, 2017

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2022

Approved by: Louis Chen
Advisor

Ruriko Yoshida
Co-Advisor

Samuel E. Buttrey
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In many cases, a low retention rate of first-term sailors indicates that unsatisfactory sailors who are struggling to find guidance also cannot find a career path that fits them. Helping each sailor find their best fit can improve the retention rate in the Navy. The Navy recently developed the Job Opportunities In the Navy (JOIN) program to help sailors find their career paths on the Bureau of Naval Personnel Online. However, there are not enough data to support the effectiveness of JOIN.

Based on a dataset obtained from the Navy Enlisted System, we first analyze which factors correlate to sailors' stays in the Navy. Then using the results from the first part of the analysis, we set up a probability distribution model to maximize the retention rate of enlisted sailors in the Navy. The result from this study can be used to help first-term sailors with JOIN. First, we conduct an analysis using a binomial logistic regression model and then calculate the model's accuracy using a confusion matrix. Second, using the variables we select in the first part of our analysis, we set up an optimization model, specifically a probability distribution model, to maximize the retention rates of enlisted sailors. Our model produces a list of rates, from the highest probability of retention rate to the lowest probability for recruits.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ABE	Aviation Boatswain's Mate-Equipment
AC	Navy Air Traffic Controllers
AO	Assembling Objects
AR	Arithmetic Reasoning
AS	Auto and Shop Information
ASVAB	Armed Services Vocational Aptitude Battery
BLR	Binomial Logistic Regression
BUPERS	Bureau of Naval Personnel
BUPERS-34	Metrics and Analytic Branch
CA	Construction Apprentice
CR	Construction Recruit
CTN	Cryptologic Technician Network
CTO	Cryptologic Technician Communication
DLI	Defense Language Institute
DM	Draftsman/Illustrator
DN	Dentalman
EDVR	Enlisted Distribution Verification Reports
EI	Electronics Information
FN	False Negative

FP	False Positive
GS	General Science
HR	Human Resource
JOIN	Job Opportunities in the Navy
LI	Lithographer
MC	Mechanical Comprehension
MEPS	Military Entrance Process Station
MK	Mathematics Knowledge
MS	Mess Specialist
MU	Musician
NES	Navy Enlisted System
NPS	Naval Postgraduate School
PC	Paragraph Comprehension
TN	True Negative
TP	True Positive
VE	Verbal Skills
VK	Verbal Knowledge
WK	Word Knowledge

Executive Summary

Many first-term sailors decide to leave the Navy every year. Although it is expected to have some first-term sailors attrite or separate, an increased number of first-term sailors leaving can lead to a high separation rate. A high separation rate can be expensive for the Navy and impact combat readiness of the Navy. To prevent a high separation rate, we develop a match-making model to help the Navy retain first-term sailors.

Previously, several studies have attempted to reduce the attrition rate of first-term sailors instead of the separation rate of first-term sailors. First-term attrition occurs when a first-term sailor fails to meet the initial enlistment contract. First-term separation occurs when a first-term sailor fails to complete the initial contract or does not re-enlist to serve another term. Therefore, we focus on improving the first-term separation rate to retain more first-term sailors.

The Navy collects administrative information of all active duty Navy enlisted personnel annually. This data is collected in the Navy Enlisted System (NES). We use BLR to analyze NES to find factors causing separation, and with these identified factors, we build a match making model to help first-term sailors choose the fittest career path.

We analyze NES data from 1998 to 2021. In these data, 96 unique ratings (jobs) are identified. Then, the BLR model is used to find the separation factors in each rating. Lastly, these factors are used in the probability distribution equation to calculate the retention probability for each rating for a first-term sailor.

Our match making model produces a list of ratings from the highest probability of retention to the lowest probability of retention for a first-term sailor. This rating list can help first-term sailors and detailers at Military Entrance Process Station (MEPS) find the fittest ratings for first-term sailors, decreasing the separation rate while maintaining the strength of the Navy.

Recently, the Navy implemented a new program called Job Opportunities In the navy (JOIN) to help sailors to find the fittest ratings. Our match-making model can be used to support JOIN by providing a list of retention probability of ratings.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I would like to thank my parents for their lifelong unconditional love and support. I am who I am today because of my parents.

I thank the United States Navy for granting me an opportunity to obtain a master's degree at the Naval Postgraduate School.

I sincerely appreciate Dr. Yoshida Ruriko and Dr. Louis Chen's guidance and insight throughout this journey. They generously held a weekly meeting to keep me on track. Their wealth of knowledge didn't cease to amaze me throughout the process. Without their contribution and help, I would not be able to complete this study. I do not have the words to express my appreciation for my support and care.

I thank the United States Navy Human Resource community for this opportunity and especially appreciate CDR. Patrick Saluke for providing the dataset for this study.

Lastly, I would like to expand my appreciation for all the sailors and dedicate this study to the brave young men and women who voluntarily join the armed services to protect our nation.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1: Introduction

Many young Americans decide to join the United States Navy with a dream of serving the country. It is safe to assume that the majority of them do not join thinking that they quit during the first-term contract. However, many of them end up leaving the Navy during or upon completion of their first-term contract. Although it is not necessarily a bad thing that sailors are leaving, an excessive number of sailors leaving can be expensive for the Navy. According to a technical report on attrition cost, in the fiscal year 2008, it was estimated that first-term attrition cost about $209 million to $219 million. About 10% of the training and recruiting budget was lost due to attrition given that the training and recruiting budget for the fiscal year 2008 was about $2,075 million (Enns 2008). More importantly, a high attrition rate of first-term sailors is causing a shortage of sailors in various Navy communities, which has a direct impact on the combat readiness of the Navy. To maintain the Navy's mission and capabilities, it is necessary to improve the retention rate of first-term sailors.

First-term attrition has been an ongoing problem for the Navy. This attrition occurs when a sailor fails to complete his or her initial contract. Over the years, the Navy has investigated reasons for attrition and found that many recruits end up leaving while they are at basic training facility, which is designed to test recruits' physical and mental strength. Understandably, many end up quitting due to high physical and mental stress. To address this issue the Navy remodeled the structure of the basic training to decrease the attrition rate. However, the threat of attrition does not stop at basic training. Even after recruits become sailors upon the completion of basic training, many still quit, for several reasons. According to a study conducted by G.E. Larson and S.B Kewley (2000), poor career opportunities are one of the important factors in attrition. Poor career opportunities not only affect attrition but also affect separation. Separation occurs when first-term sailors do not re-enlist after their initial contract. Although separation is not as costly as attrition, it is still affecting the combat readiness of the Navy. Therefore, it is important to find a way to decrease both attrition and separation rates.

1.1 Purpose

The purpose of this study is to develop a match-making model to help first-term sailors to choose the fittest ratings for their careers. Choosing the right career path improves the retention rate of first-term sailors. This improvement will result in saving recruiting costs as well as maintaining the strength of each Navy community. Before this study, the Navy has been using the same method to match ratings to recruits for the last 50 years. In 2018, a new match-making system was introduced to the Navy, but its effectiveness has not been verified. This study offers a match-making model that can be easily implemented in the rating selection process.

1.2 Retention Rate

First-term attrition is used to describe those who failed to complete the initial contract. Attrition does not capture those who decided to leave the Navy upon completion of their initial contract; therefore, to capture all the sailors who left the Navy during or after the initial contract, retention rate should be used. The one major benefit of focusing on retention rate compared to attrition rate is that the Navy can potentially retain more sailors after their first-term contract. Retaining more sailors leads to keeping the strength of each Navy community as well as saving on the recruiting budget.

1.3 Reason for Focusing First-Term Enlisted Sailors

Focusing on first-term enlisted sailors is important because the separation of first-term enlisted sailors puts a monetary burden on the Navy and it weakens the strength of each Navy community. Therefore, by focusing on first-term enlisted sailors, the Navy can benefit from saving recruiting budget while strengthening each Navy community

1.4 Benefit of this Study

This study offers a supplementary match-making model to help first-term sailors to find the fittest ratings for their careers. Successfully matching sailors to the right ratings decreases both attrition and separation rates. A reduction in both rates decreases the Navy's recruiting costs and improves the strength of each navy community. Furthermore, the results of this study can be easily implemented into the current match-making process without additional costs. The Navy will benefit greatly from the results of this study.

1.5 Thesis Organization

This study contains five chapters. Each chapter explains the development and result of this study. Chapter 2 presents a literature review that discusses previous studies that are similar to this study and how they are different. It also talks about the current match-making process for first-term sailors. Chapter 3 presents the methodology of match-making model development. Chapter 4 describes the result of this study and demonstrates how to apply it to the current system. Chapter 5 offers ideas for future studies that can improve this study.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2: Literature Review

There are many prior research and studies on attrition rate of first-term Navy enlistees. Flyer and Elster (1983) examine attrition rate by various demographics such as gender, education level, and age. They simply calculated attrition rate by dividing the total number of sailor undergoing attrition by the total number of service members for each category. Flyer and Elster were able to identify reasons for first-term attrition. In their study, they found that large numbers of female enlistees were separated for pregnancy during their first three years of service. They also found that many enlistees failed to meet the minimum behavior and performance standards. Besides these two findings, they found many other factors that were affecting attrition (see Flyer and Elster 1983 for more details on their findings). These are certainly important findings that can potentially help improve the attrition rate. However, this study over-simplifies the factors important to attrition. The authors describe only the direct relationship of factor to attrition rather than the correlation among factors. Flyer and Elster's study views attrition as a complicated problem that has solutions (Flyer and Elster 1983); however, this study views attrition as a complex problem that has no simple solution but exhibits a combination of many interrelated factors. In actual application, it is difficult to predict the probability of a first-term sailor separating from the service derived from a single factor. There are many potential factors when it comes to deciding to leave; therefore, this study reflects on the correlation of factors. This certainly benefits the Navy by understanding and reducing the attrition rate. Moreover, the data set that they used is from 1973 to 1978 which makes this study outdated. Since the 1970s the United States had experienced many civil rights movements and other forces, which have changed the demographic of the military. We use recent data that reflect the current Navy.

Besides Flyer and Elster's study, there are other studies that focused on more specific factors. Carroll (2008-09) conducted a study on Navy enlisted attrition based on race, ethnicity, and type of occupation. This study examines the attrition of racial-ethnic groups by using cross-tabulation analysis. This method is similar to what Flyer and Elster did for their research. Cross-tabulation analysis help summarize and observe patterns of data, which can be useful for a simpler problem. The disadvantage of using cross-tabulation analysis is that it typically shows relationships among two or three variables; therefore, when there are more than three variables, cross-tabulation becomes ineffective. For Carroll's study, cross-tabulation analysis is appropriate since the study only focuses on two or three variables.

We have a different approach than previous studies. Instead of using cross-tabulation analysis, we use Binomial Logistic Regression (BLR) analysis. BLR is a model to analyze significant factors that affect the status of sailors. This method allows us to analyze multiple factors.

Our study focuses on the separation rate instead of the attrition rate. The separation rate includes all the sailors who left the Navy during or after the initial contract; therefore, decreasing the separation rate has more impact on improving the retention rate than decreasing the attrition rate. Broadening the scope of study to the separation rate can help to see factors that can improve the overall retention rate of the first-term sailor. Improving the retention rate has a bigger impact on the Navy's recruiting issues as well as reduce the financial burden. To find a correlation between various categories, the binomial logistic regression analysis method is used. This method provides a more dynamic approach to improving the retention rate.

2.1 Attrition vs. Separation Rate

First-term attrition occurs when a first-term sailor fails to meet the initial enlistment contract. First-term separation occurs when a first-term sailor either fails to complete the initial contract or does not re-enlist to serve another term. Therefore, the attrition rate is a part of the separation rate.

Table 2.1 shows the attrition and separation rates from 1998 to 2021. These rates are calculated from data provided by the Metrics and Analytic Branch (BUPERS-34) in Millington, Tennessee

The separation rate is about 4% higher than the attrition rate each year. Therefore, decreasing the separation rate can potentially have a bigger improvement on the retention rate than decreasing the attrition rate. Improving the retention rate will strengthen the combat capability of the Navy while decreasing the budget for recruiting. To improve the

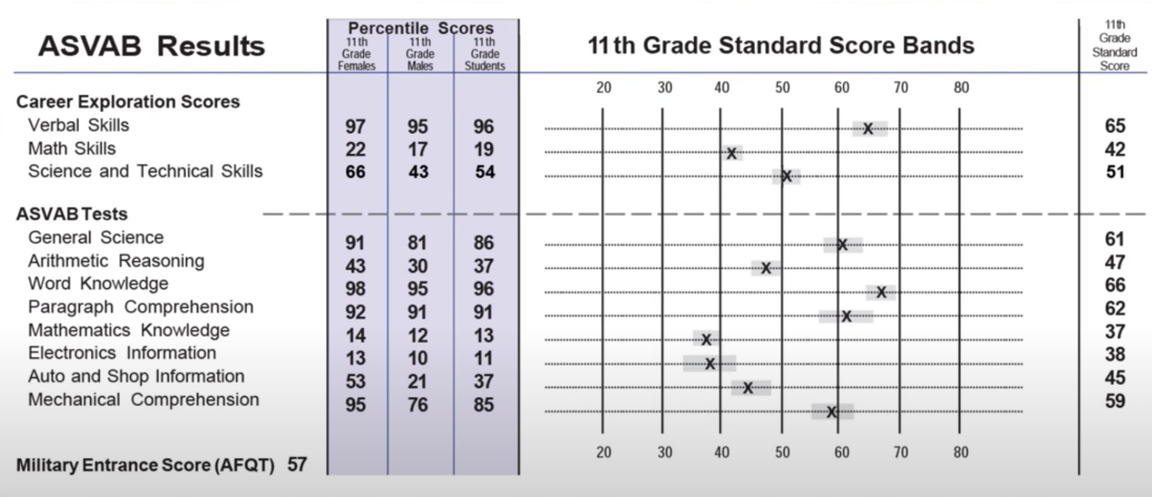
year	Attrition Rate(%)	Separation Rate(%)
1998	13.9	18.4
1999	12.7	16.8
2000	11.3	15.3
2001	10.2	14.2
2002	8.3	11.6
2003	7.4	11.0
2004	7.7	12.3
2005	7.7	13.2
2006	8.2	12.8
2007	8.9	13.7
2008	8.5	13.0
2009	7.8	12.0
2010	7.3	10.5
2011	7.0	10.5
2012	6.8	11.4
2013	5.8	9.8
2014	5.5	9.6
2015	5.8	9.8
2016	6.6	11.4
2017	6.7	11.8
2018	8.0	11.1
2019	7.0	10.5
2020	6.3	8.6
2021	5.9	7.9

Table 2.1. Annual Attrition Rate vs Separation Rate. These rates are calculated from Navy Enlisted System (NES) data from 1998 to 2021. The attrition rate is calculated by dividing the total number of first-term sailors by first-term sailors who left prior to completing their initial contract. The separation rate is calculated by diving the total number of first-term sailors by first-term sailors who left prior or after their initial contract.

retention rate, it is important to find the causes of separation.The current rate match-making is outdated and fail to effectively assign rates to sailors. Not having an effective match-making system will assign sailors to rate sub-optimally, which might cause sailors to leave the Navy. The next section discusses the current method of match-making in the Navy.

2.2 Armed Services Vocational Aptitude Battery (ASVAB)

The Navy relied on the ASVAB test to determine a recruit's rating qualification. The ASVAB test is designed to test the recruit's general knowledge and comprehension skill. Based on the ASVAB score, a recruit is qualified for certain ratings. Figure 2.1 shows an example of an ASVAB testing result.



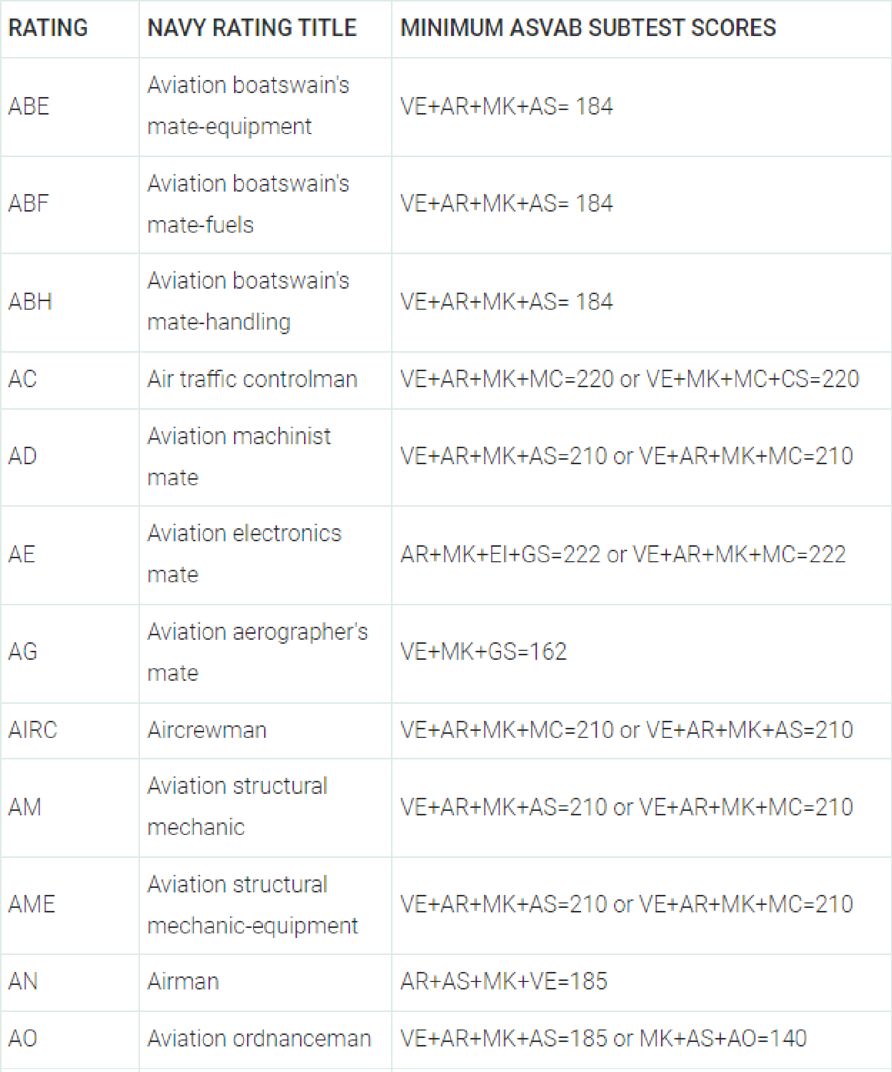
ASVAB Results	Percentile Scores 11th Grade Females	 11th Grade Males	 11th Grade Students	11th Grade Standard Score Bands 20 30 40 50 60 70 80	11th Grade Standard Score
Career Exploration Scores					
Verbal Skills	97	95	96		65
Math Skills	22	17	19		42
Science and Technical Skills	66	43	54		51
ASVAB Tests					
General Science	91	81	86		61
Arithmetic Reasoning	43	30	37		47
Word Knowledge	98	95	96		66
Paragraph Comprehension	92	91	91		62
Mathematics Knowledge	14	12	13		37
Electronics Information	13	10	11		38
Auto and Shop Information	53	21	37		45
Mechanical Comprehension	95	76	85		59
Military Entrance Score (AFQT) 57				20 30 40 50 60 70 80	

Figure 2.1. Sample of ASVAB Score sheet. ASVAB Results are separated into two categories: Career Exploration Scores and ASVAB Tests. The second column shows the percentile scores for 11th grade students. And the last column shows the score that this particular recruit received. The figure is adopted from ASVAB Career Exploration Program (2022).

There are eleven subjects in ASVAB testing. A recruit receives scores in individual subjects[1], then these subjects are combined in different combinations to determine qualified rates. Figure 2.2 shows an example of how ASVAB scores are used to determine qualified rates.

Detailers use this ASVAB Qualification table to list all the rates that recruits are qualified for, then recruits get to put in their preference rates. Once detailers receive recruits'

[1] Verbal Skills (VE),Mathematics Knowledge (MK),Verbal Knowledge (VK),General Science (GS),Arithmetic Reasoning (AR),Word Knowledge (WK),Paragraph Comprehension (PC),Electronics Information (EI),Auto and Shop Information (AS),Mechanical Comprehension (MC),and Assembling Objects (AO).



RATING	NAVY RATING TITLE	MINIMUM ASVAB SUBTEST SCORES
ABE	Aviation boatswain's mate-equipment	VE+AR+MK+AS= 184
ABF	Aviation boatswain's mate-fuels	VE+AR+MK+AS= 184
ABH	Aviation boatswain's mate-handling	VE+AR+MK+AS= 184
AC	Air traffic controlman	VE+AR+MK+MC=220 or VE+MK+MC+CS=220
AD	Aviation machinist mate	VE+AR+MK+AS=210 or VE+AR+MK+MC=210
AE	Aviation electronics mate	AR+MK+EI+GS=222 or VE+AR+MK+MC=222
AG	Aviation aerographer's mate	VE+MK+GS=162
AIRC	Aircrewman	VE+AR+MK+MC=210 or VE+AR+MK+AS=210
AM	Aviation structural mechanic	VE+AR+MK+AS=210 or VE+AR+MK+MC=210
AME	Aviation structural mechanic-equipment	VE+AR+MK+AS=210 or VE+AR+MK+MC=210
AN	Airman	AR+AS+MK+VE=185
AO	Aviation ordnanceman	VE+AR+MK+AS=185 or MK+AS+AO=140

Figure 2.2. Example of ASVAB Minimum Score for Rate Qualification. Each Rating has a minimum ASVAB Subtest scores. The subtest can be calculated by combining different subject results from ASVAB test. For example. to qualify for Aviation Boatswain's Mate-Equipment (ABE), a recruit must have a minimum score of 184 from combining VE,AR,VK,and AS scores. The chart is adopted from Military.com (2022).

preferences, they make sure each Navy community is meeting its minimum requirement for recruits then try their best to match each recruit with the order of preference. The Navy has been able to sustain its combat readiness and capabilities by relying on the ASVAB match-making method for decades.

However, there are many criticisms of the effectiveness of ASVAB testing. It does not take an account one's ability to adapt. In the Navy, things are moving fast and learning curves are steep; therefore, sailors must be able to adapt. Many sailors fail to adapt to the high intensity of the Navy despite their ASVAB score. For instance, recruits who scored high in ASVAB testing are qualified to attend the Defense Language Institute (DLI) to become linguists. However, DLI has an attrition rate of around 25 percent. Many sailors fail to adapt to the high intensity of DLI despite scoring high in ASVAB testing.

ASVAB testing is outdated. The test was created back in 1968. This type of standardized testing is being phased out in the civilian sector because of its inability to capture an applicant's true potential. Therefore, a new type of match-making is desperately needed in the Navy. The Navy hired Dr. Stephen E. Watson to develop a new match-making program for the Navy. The next section discusses the new Navy match-making program.

2.3 Job Opportunities in the Navy (JOIN)

Dr. Stephen E. Watson (2020) developed a software program called JOIN. Traditionally, many software developers used cognitive testing such as employee aptitude testing to develop job matching software. However, according to Watson, cognitive testing fails short when it comes to predicting performance over time; moreover, Navy jobs changes frequently which makes predicting performance even harder. Due to the limiting capabilities of cognitive testing, Watson created a taxonomy for approximately 80 Navy job/ratings with hundreds of words representing the description of each rating. The taxonomy of each rating is then matched with the recruit's preference of their work style. The goal of JOIN is to meet nine criteria designed for the Navy's needs. JOIN was approved to be used in all accessions in October 2018; therefore, there is not yet any actual data to show its effect on meeting these criteria, but using a combination of quantitative and qualitative evidence, JOIN promises to meet the Navy's criteria. The psychometric properties of JOIN are derived from data of 6,988 U.S. Navy Sailors, as well as gender differences and factor structure. JOIN also examined the association between JOIN scores and five Navy criterion variables to validate the JOIN model.

Watson (2020) used a combination of quantitative and qualitative evidence to show how JOIN meet the Navy's needs according to the nine design criteria.

2.3.1 Nine Design criteria for JOIN

According to Watson the nine design criteria for JOIN are integrated parts of developing JOIN. The following list explains each criterion in detail.

1. Discriminative. JOIN's "goodness of fit" scores provide the Navy's 80 distinctive ratings with the additional potential to inform Navy policy for rating mergers[2].
2. Sustainable. The requirement of the Navy enlisted forces change constantly, therefore ratings constantly change as well. JOIN is able to create standardized procedures by using interest-based job analysis to minimize resources.
3. Parsimonious. JOIN is relatively inexpensive to operate and maintain due to its taxonomy for ratings. JOIN can create unique profiles for each Navy rating with the few number of descriptions.
4. Fast. JOIN is able to run the test within 20 minutes.
5. Modular. JOIN can be used as a standalone source of vocational preference, but it can also be paired with the Rating Identification Engine[3]. JOIN is an input to RIDE and provides unique ratings.
6. Intuitive. Many subject matter experts participated in creating JOIN. Upon testing with recruits, 93.7 percent of them rated JOIN's ease-of-use as good or very good.
7. Educational. JOIN's job description is simple and direct and it helps recruits form an understanding of what the job entails.
8. Honest. JOIN used realistic and honest pictures that depict a diversity of tasks and environments that each rating is expected to face, thus removing ambiguity for recruits.
9. Predictive. JOIN can predict job performance, which can increase the retention rate.

In summary, JOIN is a test of human vocational interest that captures people's interest in Navy rates. JOIN is created with inputs from subject matter experts and each rating description to help recruits to find their best match. JOIN addresses the current problem with

[2] Rating mergers happen when ratings with similar job description merged to reduce additional training requirements.

[3] Navy's centralized classification system.

ASVAB testing and provides a potential solution to improve the retention rate. Unfortunately, there is no actual data to support JOIN's claim at this moment. An additional couple of years, or possibly a decade, will be needed until there are meaningful data to analyze the effectiveness of JOIN. Regardless of its actual effectiveness, JOIN offers a new way to improve the current match making method and implements many interesting ideas. Contrary to JOIN, our research is driven by actual data and provides an analytical solution. Our research can be used alongside JOIN to help detailers to maintain the health of each Navy community while assigning the most appropriate rates for recruits.

In the next chapter, we discuss the methodology of our approach to help improve the retention rate of first-term sailors.

CHAPTER 3:
Methodology

In this study, we are interested in finding factors that are associated with separation among first-term sailors. The separation rate can be calculated by dividing the number of sailors separated in their first term by the total number of first-term sailors; the Navy denotes the status of each sailor as either separated or active. This is “Yes or No” which is a binary problem; therefore, using the BLR method is appropriate to find factors that are causing separation.

We primarily used the statistical software R Core Team (2020) for building our model. R is a programming language and environments specialized in statistical computing. W e also used Python for data cleaning process.

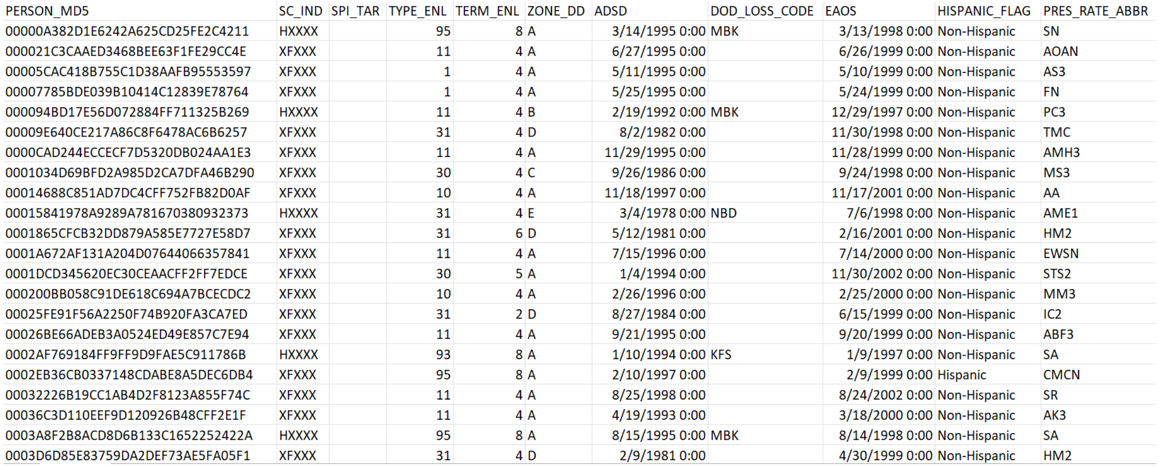
3.1 Binomial Logistic Regression Method

BLR is a great method to predict a binary response variable from one or more independent (explanatory) variables. Unlike linear regression, which expects a numerical response variable, the BLR predicts the probability of being in a particular category of the dependent variable given the independent variables (James et al. 2013). This property of BLR is very useful for our research since we are interested in creating a new match-making model to improve the retention rate. In our case, the BLR identifies factors that are associated with sailors leaving or staying in the Navy. With the information that we infer from BLR, we can build a model that helps recruits and detailers to find the best matching rates with the highest probability of retaining recruits.

3.2 Data Set

The data for this study was provided by the BUPERS-34 in Millington, Tennessee. They are annual active enlisted inventory data from the NES. According to the Navy (2022), NES is the Navy’s authoritative database for all active duty Navy enlisted personnel. The system generates and maintains the official personnel records for all active and reserve enlisted personnel. NES is mainly used for calculating enlisted strength and authorizing the establishment of a payment record at the Defense Finance Accounting Center. NES is also used to prepare Enlisted Distribution Verification Reports (EDVR) for distribution to field actives. The enlisted distribution and promotion processes depend on the quality of NES data, as are managerial and congressional groups overseeing accumulated information about the active enlisted population (Navy 2022). Therefore, NES is a perfect data set for this research since it includes all the active Navy enlisted personnel as well as their information in regard to their service. More detailed information about the data set is discussed in a later section.

Even though NES contains many important and useful data, it also contains administrative information that is not useful for this study. Because NES is a large file with many columns, Figure 3.1 below shows only an example of raw NES data.



PERSON_MD5	SC_IND	SPI_TAR	TYPE_ENL	TERM_ENL	ZONE_DD	ADSD	DOD_LOSS_CODE	EAOS	HISPANIC_FLAG	PRES_RATE_ABBR
00000A382D1E6242A625CD25FE2C4211	HXXXX		95	8	A	3/14/1995 0:00	MBK	3/13/1998 0:00	Non-Hispanic	SN
000021C3CAAED3468BEE63F1FE29CC4E	XFXXX		11	4	A	6/27/1995 0:00		6/26/1999 0:00	Non-Hispanic	AOAN
00005CAC418B755C1D38AAFB95553597	XFXXX		1	4	A	5/11/1995 0:00		5/10/1999 0:00	Non-Hispanic	AS3
00007785BDE039B10414C12839E78764	XFXXX		1	4	A	5/25/1995 0:00		5/24/1999 0:00	Non-Hispanic	FN
000094BD17E56D072884FF711325B269	HXXXX		11	4	B	2/19/1992 0:00	MBK	12/29/1997 0:00	Non-Hispanic	PC3
00009E640CE217A86C8F6478AC6B6257	XFXXX		31	4	D	8/2/1982 0:00		11/30/1998 0:00	Non-Hispanic	TMC
0000CAD244ECCECF7D5320DB024AA1E3	XFXXX		11	4	A	11/29/1995 0:00		11/28/1999 0:00	Non-Hispanic	AMH3
0001034D69BFD2A985D2CA7DFA46B290	XFXXX		30	4	C	9/26/1986 0:00		9/24/1998 0:00	Non-Hispanic	MS3
00014688C851AD7DC4CFF752FB82D0AF	XFXXX		10	4	A	11/18/1997 0:00		11/17/2001 0:00	Non-Hispanic	AA
00015841978A9289A781670380932373	HXXXX		31	4	E	3/4/1978 0:00	NBD	7/6/1998 0:00	Non-Hispanic	AME1
0001865CFCB32DD879A585E7727E58D7	XFXXX		31	6	D	5/12/1981 0:00		2/16/2001 0:00	Non-Hispanic	HM2
0001A672AF131A204D07644066357841	XFXXX		11	4	A	7/15/1996 0:00		7/14/2000 0:00	Non-Hispanic	EWSN
0001DCD345620EC30CEAACFF2FF7EDCE	XFXXX		30	5	A	1/4/1994 0:00		11/30/2002 0:00	Non-Hispanic	STS2
000200BB058C91DE618C694A7BCECDC2	XFXXX		10	4	A	2/26/1996 0:00		2/25/2000 0:00	Non-Hispanic	MM3
00025FE91F56A2250F74B920FA3CA7ED	XFXXX		31	2	D	8/27/1984 0:00		6/15/1999 0:00	Non-Hispanic	IC2
00026BE66ADEB3A0524ED49E857C7E94	XFXXX		11	4	A	9/21/1995 0:00		9/20/1999 0:00	Non-Hispanic	ABF3
0002AF769184FF9FF9D9FAE5C911786B	HXXXX		93	8	A	1/10/1994 0:00	KFS	1/9/1997 0:00	Non-Hispanic	SA
0002EB36CB0337148CDABE8A5DEC6DB4	XFXXX		95	8	A	2/10/1997 0:00		2/9/1999 0:00	Hispanic	CMCN
00032226B19CC1AB4D2F8123A855F74C	XFXXX		11	4	A	8/25/1998 0:00		8/24/2002 0:00	Non-Hispanic	SR
00036C3D110EEF9D120926B48CFF2E1F	XFXXX		11	4	A	4/19/1993 0:00		3/18/2000 0:00	Non-Hispanic	AK3
0003A8F2B8ACD8D6B133C1652252422A	HXXXX		95	8	A	8/15/1995 0:00	MBK	8/14/1998 0:00	Non-Hispanic	SA
0003D6D85E83759DA2DEF73AE5FA05F1	XFXXX		31	4	D	2/9/1981 0:00		4/30/1999 0:00	Non-Hispanic	HM2

Figure 3.1. Snapshot of NES data from 1998 (only shown columns 1-10). The original NES data contains total of 30 columns. Each column is listed in Table 3.1.

This initial data is not suitable for logistic regression analysis, since many columns are not useful. To take a closer look at the data, Table 3.1 shows each column.

1.DAY-DATE	2.PERSON-MD5	3.SC-IND
4.SPI-TAR	5.TYPE-ENL	6.TERM-ENL
7.ZONE-DD	8.ADSD	9.DOD-LOSS-CODE
10.EAOS	11.HISPANIC-FLAG	12.PRES-RATE-ABBR
13.PAYGRADE	14.RACE-GROUP	15.SEX
16.SOFT-EAOS	17.STR-CLASS	18.TIME-IN-RATE-DATE
19.YEARS-ACTIVE-SERVICE-FOR-PAY	20.YEARS-ACTIVE-SERVICE	21.ENL-MGMT-COMMUNITY
22.LOSS-CHANGE-DATE-OF-OCCURRENCE	23.ONBOARD-ACTY-NAME	24.ONBOARD-SS
25.ONBOARD-DATE-REC	26.ONBOARD-REC	27.PRES-RATE-CODE-PG
28.PROS-RATE-CODE-PG	29.EFFECTIVE-DATE-OF-PAYGRADE	30.PRES-RATE-AUTH

Table 3.1. Names of the columns in the data set.

One of the biggest challenges in this research is to clean the initial data to be suitable for BLR. Due to the dataset's large size, it is more practical to have many steps to clean the initial data. Figure 3.2 shows the master workflow of the methodology for this research. The next subsections discuss the detailed procedure of each step in the workflow.

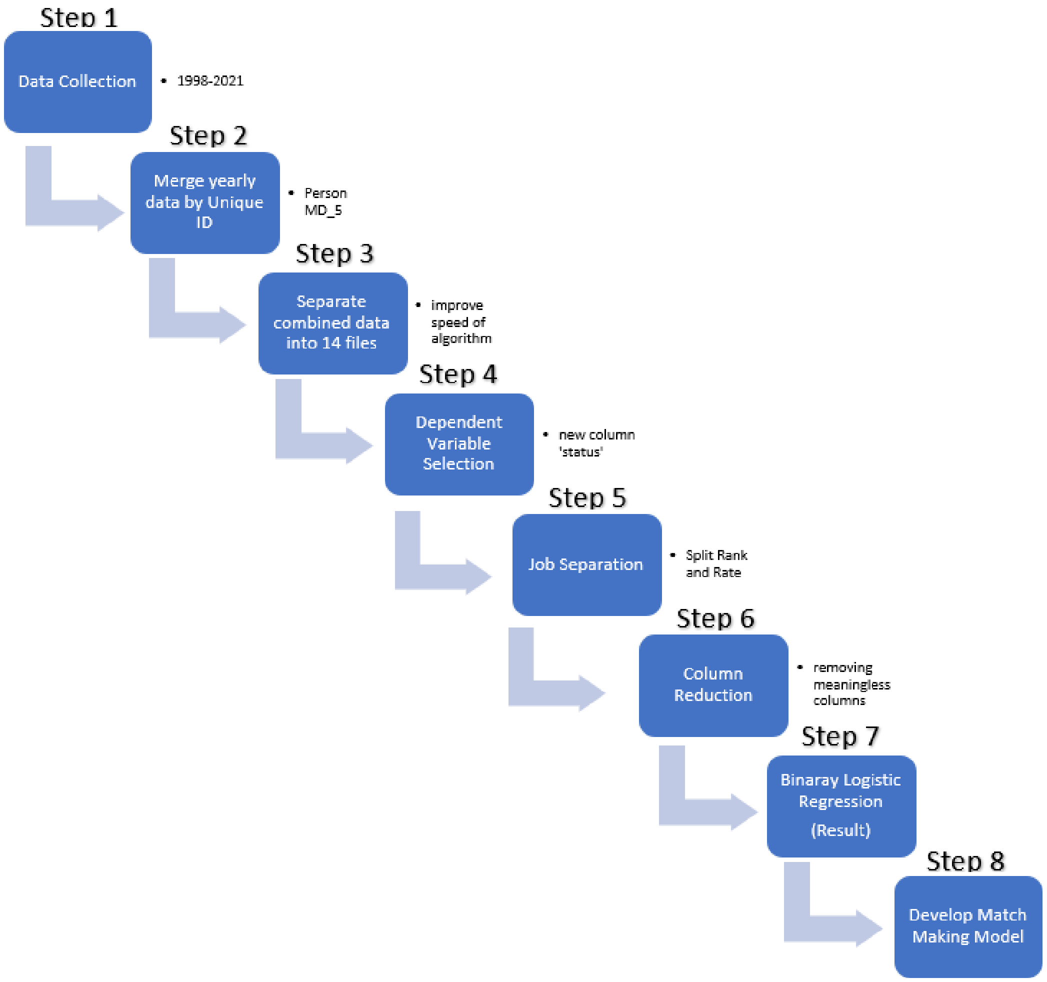

Step 1
Data Collection
1998-2021
Step 2
Merge yearly data by Unique ID
Person MD_5
Step 3
Separate combined data into 14 files
improve speed of algorithm
Step 4
Dependent Variable Selection
new column 'status'
Step 5
Job Separation
Split Rank and Rate
Step 6
Column Reduction
removing meaningless columns
Step 7
Binaray Logistic Regression
(Result)
Step 8
Develop Match Making Model

Figure 3.2. Master Work Flow of Building the Match Making Model. There are total 8 steps in building the match making model. Steps 1 to 6 show in detail how the original NES data is cleaned to be suitable for analysis. Step 7 explains results of BLR and the last step shows the development of the match-making model.

3.2.1 Data Merging

NES generates data on every Navy active enlisted personnel annually; therefore there is a duplication of the same individual through the years until he or she separates from the Navy. For instance, a sailor who enlisted in 1998 and separated in 2005, appears in each year from 1998 to 2005. This duplicated data leads to data redundancy which gives an inconsistent result. To avoid duplication, data from 1998 to 2021 are merged by their unique identification code (SCI-ID, column 2). Since this research is interested in the separation rate of the first-term sailor, the data is merged with the most updated data. This means that the sailor in the example above is merged with his or her 2005 data. An accumulated single file is created after this process.

3.2.2 Data Separation

Merging all the data files created one large file. To reduce data cleaning processing time, the data is separated into 14 data files, each file containing about 100,000 Navy enlisted personnel. Once all the files gets properly cleaned, they consolidated once again. This process can be skipped for high-performing computers.

3.2.3 Dependent Variable Selection

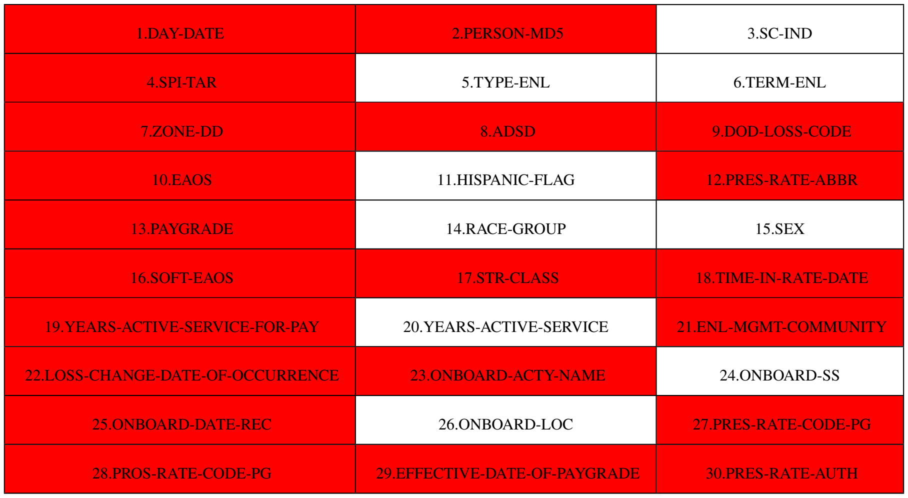
The goal of BLR is to find any significant factors that may impact the outcome of the dependent variable. To reflect the cause of separation, a dependent variable has to represent the status of each sailor. SC-IND, column 3, of the data set has two variables: XFXXX, HXXXX. The variable XFXXX represents an active Navy enlisted member while HXXXX represents a separated Navy enlisted member. Therefore, this column is the dependent variable. For the easier representation of status, a new column ‘status’ is created to reflect the status of each sailor using 0 and 1. The variable XFXXX is converted to 0 while HXXXXX is converted to 1.

3.2.4 Job Separation

In the Navy, ratings and rank are combined when it stores the data. For instance, a yeomen, a Navy rating, with a rank of petty officer first class, is stored as YN1. Combing ratings and rank is practical while on duty, but it creates many issues when it is analyzed since each combination is treated as unique factors despite having the same ratings. Therefore, it is necessary to separate the rating and rank. We created a new column 'job' and stored each sailor's ratings in that column. In this process, 96 unique ratings are identified from 1998 to 2021. Then the data is grouped by ratings and stored in separate files under each rating. This allows us to run BLR analysis on individual ratings. This way we are able to gather significant factors that are associated with the separation rate.

3.2.5 Column Reduction

As stated above, NES generates information for administrative purposes. Therefore, there is information that is not related to the status of individual sailors. This information is not included in our analysis. The highlighted cells in Table 3.2 depict columns that are not be included.



🟥 1.DAY-DATE	🟥 2.PERSON-MD5	3.SC-IND
🟥 4.SPI-TAR	5.TYPE-ENL	6.TERM-ENL
🟥 7.ZONE-DD	🟥 8.ADSD	🟥 9.DOD-LOSS-CODE
🟥 10.EAOS	11.HISPANIC-FLAG	🟥 12.PRES-RATE-ABBR
🟥 13.PAYGRADE	14.RACE-GROUP	15.SEX
🟥 16.SOFT-EAOS	🟥 17.STR-CLASS	🟥 18.TIME-IN-RATE-DATE
🟥 19.YEARS-ACTIVE-SERVICE-FOR-PAY	20.YEARS-ACTIVE-SERVICE	🟥 21.ENL-MGMT-COMMUNITY
🟥 22.LOSS-CHANGE-DATE-OF-OCCURRENCE	🟥 23.ONBOARD-ACTY-NAME	24.ONBOARD-SS
🟥 25.ONBOARD-DATE-REC	26.ONBOARD-LOC	🟥 27.PRES-RATE-CODE-PG
🟥 28.PROS-RATE-CODE-PG	🟥 29.EFFECTIVE-DATE-OF-PAYGRADE	🟥 30.PRES-RATE-AUTH

Table 3.2. Names of the columns in the data set. Red represents deleted columns. The number in each cell represents column number

The following list explains the meaning and usefulness of each columns. The numbers on the list correspond with column numbers.

1. Data collection date (DAY-DATE) is removed since data collection date is irrelevant to the separation rate.
2. Personal Identification Code (PERSON-MD5) is removed because the unique identification numbers are simply used to identify unique individuals. However, this column is used to identify unique individuals that appear through different years. This column is used to merge all the NES data over the years to identify every unique individual that ever served in the Navy since 1998.
3. SC-IND is used to determine the duty status of each sailor. This column is used to create a new column "Status" which is the dependent variable.
4. Special program Indicator (SPI-TAR) is removed because this code represents Reserve or Full-Time Support (FTS) forces which is irrelevant for this analysis.
5. Type of Enlistment (TYPE-ENL) provides information on how individual sailors enlisted as well as whether they received signing bonuses. This column is directly related to the financial aspect of enlistment; thus it is used as one of the independent variables.
6. Term of Enlistment (TERM-ENL) represents the number of years of the sailor's current enlistment. This column is used to distinguish between first-term sailors and non-first-term sailors.
7. Selective Reenlistment Bonus Zone (ZONE-DD) is irrelevant to this study since this study is only interested in first-term enlisted sailors.
8. Active Duty Service Date (ADSD) provides redundant information as TERM-ENL, therefore it is removed.
9. The type and reason for separation (DOD-LOSS-CODE) provide information on why sailors are separated. This can provide good insight into the main cause of separation, unfortunately, there were too many missing data in this column which makes this column unusable.
10. Expiration of Active Obligated Service - Hard (EAOS) is removed since it does not provide relevant information for this study.
11. Hispanic or non-Hispanic (HISPANIC-FLAG) has two variables. This column is similar to Race, but with a broader scope. This column is included in our study.

12. Present Rate (PRES-RATE-ABBR) is removed because the present rate is irrelevant to the first-term sailor.
13. Paygrade (PAYGRADE) is removed because paygrade is irrelevant to the first-term sailor.
14. Race (RACE-GROUP) is used to see if there is any correlation between race and separation rate in the Navy.
15. Gender (SEX) is used to see if gender display any roles in the separation rate in the Navy.
16. Expiration of Active Obligated Service - Soft (SOFT-EAOS) is removed. This data is primarily used for record-keeping and does not provide any meaningful information.
17. Type of service (STR-CLASS) is removed due to too many missing data.
18. Time in Rate date (TIME-IN-RATE-DATE) is irrelevant for this study.
19. Years of active duty pay (YEARS-ACTIVE-SERVICE-FOR-PAY) is removed for redundancy. Years of active duty is used instead.
20. Years of active duty (YEARS-ACTIVE-SERVICE) is used to determine first-term sailors by comparing to Term of Enlistment (Column 6).
21. Community manager specification code (ENL-MGMT-COMMUNITY) is irrelevant for this study. This information is exclusively used by Navy Human Resource officers to manage community health.
22. Date of loss from Navy (LOSS-CHANGE-DATE-OF-OCCURRENCE) is removed since years of active duty provides enough information to separate first-term sailors and the rest.
23. Rate Attachment (ONBOARD-ACTY-NAME) shows the exact location of individual sailors. This information is useful but not plausible for analysis since there are too many inputs that is hard to group. Instead of sailors exact location, their approximate location (ONBOARD-LOC, column 26) is used.
24. Onboard activity sea/shore code (ONBOARD-SS) is used to see sailor's job at their duty station.
25. Date of assigned duty (ONBOARD-DATE-REC) is irrelevant for this analysis.
26. Duty location (ONBOARD-LOC) is used to determine each sailor duty location.
27. Present rate (PRES-RATE-CODE-PG) shows the rank of each sailor and rank of each sailors is irrelevant for this study.
28. Prospective Rate:paygrade (PROS-RATE-CODE-PG) is irrelevant for this study.

29. Effective date of paygrade (EFFECTIVE-DATE-OF-PAYGRADE) is irrelevant for this study.
30. Present Rate: Rate change authority (PRES-RATE-AUTH) is irrelevant for this study.

After reviewing each column, we determined that 9 out of 30 columns are determined useful or relevant for BLR analysis for first-term enlisted sailors. The next section describes the cleaning process of the selected columns.

3.2.6 Further Cleaning Of Selected Columns

SC-IND shows the duty status of individual sailors. The Navy used XFXXX and HXXXX notation to represent active and separated sailors respectively. As it is mentioned above, this is the dependent variable for this analysis. This column is converted to the ‘status’ column.

TYPE-ENL represents the different ways that individual sailors enlisted in the Navy. The Navy used numeric values from 01 to 95. It also includes whether sailors received a signing bonus or not, which provides useful information for the analysis. To reduce the number of factors in this column, they are grouped by similar characteristics and are combined into nine factors[4]. To capture the impact of signing bonuses, a new column ‘bonus’ is added to separate enlistment types with bonuses and without bonuses. This new column has three levels representing bonus, no bonus, or unknown.

TERM-ENL is important information that separates first-term sailors from the rest. This column was strictly used in the process of separating sailors but was not used in the analysis.

YEARS OF ACTIVE-SERVICE is used alongside TERM-ENL to separate first-term sailors and the rest.

[4] 01-10 is grouped as ’1’, 11-20 is ’2’ and so on.

ONBOARD-LOC is the most challenging column to reduce its factors. There are more than a couple hundred factors each representing a different duty location in this column. It is not plausible to analyze because there are too many levels of factors, thus location in this column is categorized by U.S Navy Region of command. Figure 3.3 shows the U.S Navy Region map.

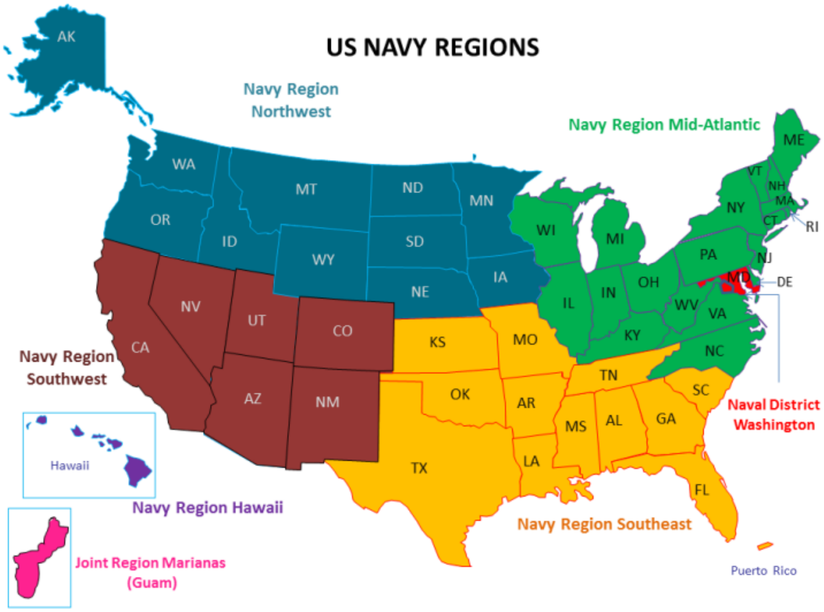
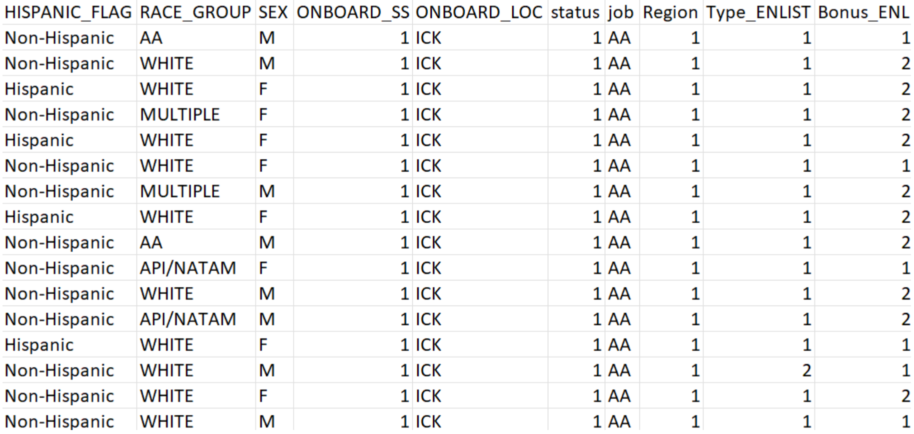


Figure 3.3. U.S Navy Regional Command. Each regions is shown in different color. Oversea duty locations and undefined duty locations are not shown in the figure. The figure is adopted from cnic.navy.mil (2022).

Each sailor's duty location is assigned to a U.S Navy Region and is stored in a new column 'region'. There is a total of nine factors in a 'region'. Seven out of nine regions are shown Figure 3.3, and the last region is an oversea duty station.

HISPANIC-FLAG, RACE-GROUP, SEX, and ONBOARD-SS columns do not require any type of alteration for analysis. A sample of the final data set is shown in Figure 3.4.



HISPANIC_FLAG	RACE_GROUP	SEX	ONBOARD_SS	ONBOARD_LOC	status	job	Region	Type_ENLIST	Bonus_ENL
Non-Hispanic	AA	M	1	ICK	1	AA	1	1	1
Non-Hispanic	WHITE	M	1	ICK	1	AA	1	1	2
Hispanic	WHITE	F	1	ICK	1	AA	1	1	2
Non-Hispanic	MULTIPLE	F	1	ICK	1	AA	1	1	2
Hispanic	WHITE	F	1	ICK	1	AA	1	1	2
Non-Hispanic	WHITE	F	1	ICK	1	AA	1	1	1
Non-Hispanic	MULTIPLE	M	1	ICK	1	AA	1	1	2
Hispanic	WHITE	F	1	ICK	1	AA	1	1	2
Non-Hispanic	AA	M	1	ICK	1	AA	1	1	2
Non-Hispanic	API/NATAM	F	1	ICK	1	AA	1	1	1
Non-Hispanic	WHITE	M	1	ICK	1	AA	1	1	2
Non-Hispanic	API/NATAM	M	1	ICK	1	AA	1	1	2
Hispanic	WHITE	F	1	ICK	1	AA	1	1	1
Non-Hispanic	WHITE	M	1	ICK	1	AA	1	2	1
Non-Hispanic	WHITE	F	1	ICK	1	AA	1	1	2
Non-Hispanic	WHITE	M	1	ICK	1	AA	1	1	1

Figure 3.4. Sample of Final Data. The sample shown in this figure is from the Navy Air Traffic Controllers (AC) rating data set. The final data has a total of 10 columns. There are a total of 96 data sets representing each identified rating.

3.3 Analyzing Steps

3.3.1 Data Splitting

In a machine learning algorithm, it is important to evaluate the model performance. The data splitting technique is an effective method to test the model's performance. The data splitting technique is also easy to perform. For our research, we randomly select 70 percent of the sample for the training set and the rest for the test set. The training set is used to build the BLR model, then the model is applied to the test set to evaluate the performance of the model.

3.3.2 List of Binomial Logistic Regression Model

As stated in the earlier section, the data has been separated by ratings. This way, we can potentially observe different factors that are associated with separation for each rating; therefore, a BLR model is built for every rating. The data splitting technique is applied to every rating, then a BLR analysis is performed with factors that are selected[5]. The result of each analysis is stored in a list. This list is used to evaluate the accuracy of each model.

3.3.3 Validating the Model

Once the list of the BLR models is obtained, the next step is to use the test set to validate the model. In R, the predict() function allows users to predict the outcome based on the new input data. This predicted value can measure the performance of the BLR model by comparing the predicted value with the actual value. In our study, test sets from each rating are used as new input data and the model predicts either 0 or 1 to indicate the duty status of the test set. These predicted values then are compared with the actual values using the confusion matrix method. A confusion matrix is a table that is used to validate the performance of a classification model. Each rating produces a confusion matrix to evaluate the performance of the individual BLR model (James et al. 2013).

[5] Refer to table 3.2

3.3.4 Calculating the Probability of retention

Once the model is validated, then the next step is calculating the ratings with the highest probability of retention for individual active enlisted sailors. In the earlier step, BLR is performed to obtain significant factors that are affecting the separation rate. R computes coefficients for these factors which can be used to calculate the model's estimates probability of retention in each rating. The result is an expression which calculates the probability of retention in each rating.

$$\Pr(Y_i = 1|X_1, X_2, X_3 \ldots, X_p) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p}}$$

Figure 3.5. Probability distribution of the response variable Y defined under the binomial logistic regression model.

In this equation, $Y \in \{0, 1\}$ is the response variable where $Y = 1$ means the person leaves a position within one year. In addition, $X_1, \ldots, X_p$ represent explanatory variables which are individual ratings, thus there are $p = 96$ explanatory variables in this equation. $\beta_0, \beta_1, \ldots, \beta_p$ represent regression coefficients from each BLR. The output of the equation is a probability of retention at each rating. To test this model's effectiveness, we are using the data of separated enlisted sailors as our sample to find out the ratings with the highest probability of retention. Finally, this model applies to recruits as well;R therefore increasing the retention rate in the Navy.

In the next Chapter, we discuss the result of the BLR analysis as well as the interpretation of the results.

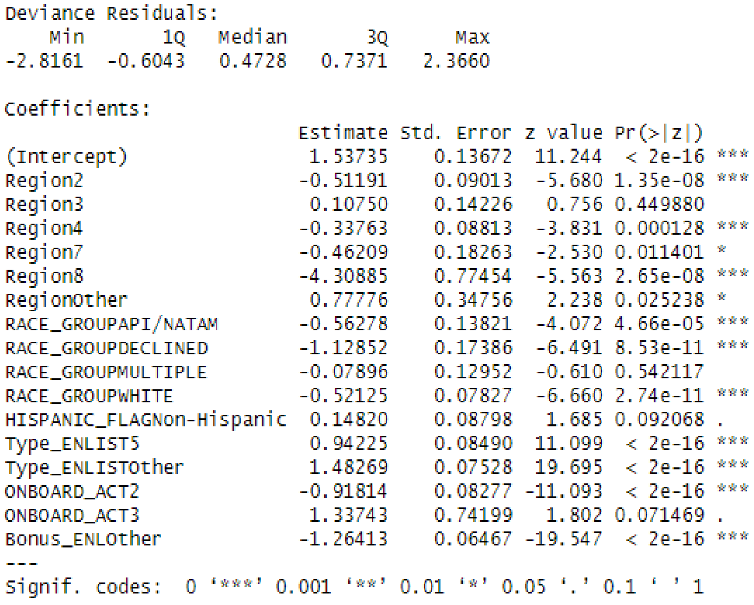
THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Result

In this chapter, we discuss the findings of our research. The final result of our analysis can be useful for both recruits and detailers. Recruits can utilize this model to find the best-fit ratings detailers can use it to effectively maintain the strength of each rating community. Since there are 96 BLR results, one for each of 96 identified ratings, we are going to use one of the ratings as an example to explain the result.

4.1 Result of Binomial Logistic Regression

The BLR is a great method to predict dichotomous dependent variables with multiple independent variables. Figure 4.1 shows a summary of BLR of AC. The BLR summary for the AC is rating selected to demonstrate the logic behind building the final model. There are 95 BLR summaries similar to AC's, however, each BLR summary has a different result.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8161  -0.6043   0.4728   0.7371   2.3660

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.53735    0.13672  11.244  < 2e-16 ***
Region2                   -0.51191    0.09013  -5.680 1.35e-08 ***
Region3                    0.10750    0.14226   0.756 0.449880
Region4                   -0.33763    0.08813  -3.831 0.000128 ***
Region7                   -0.46209    0.18263  -2.530 0.011401 *
Region8                   -4.30885    0.77454  -5.563 2.65e-08 ***
RegionOther                0.77776    0.34756   2.238 0.025238 *
RACE_GROUPAPI/NATAM       -0.56278    0.13821  -4.072 4.66e-05 ***
RACE_GROUPDECLINED        -1.12852    0.17386  -6.491 8.53e-11 ***
RACE_GROUPMULTIPLE        -0.07896    0.12952  -0.610 0.542117
RACE_GROUPWHITE           -0.52125    0.07827  -6.660 2.74e-11 ***
HISPANIC_FLAGNon-Hispanic  0.14820    0.08798   1.685 0.092068 .
Type_ENLIST5               0.94225    0.08490  11.099  < 2e-16 ***
Type_ENLISTOther           1.48269    0.07528  19.695  < 2e-16 ***
ONBOARD_ACT2              -0.91814    0.08277 -11.093  < 2e-16 ***
ONBOARD_ACT3               1.33743    0.74199   1.802 0.071469 .
Bonus_ENLOther            -1.26413    0.06467 -19.547  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.1. BLR Summary of AC. Look at the paragraph below for the detail explanation.

The BLR summary displays many information about the result of BLR. However, our research is only interested in the section where it displays the coefficients of factors. This section explains the factors that are significant and by how much. In the figure above, there is an 'Estimate' column in a predictor section. This estimate indicates how much the predicted odds are getting affected for each unit change. And this estimate is measured in log odds. For instance, the BLR summary of AC indicates that Region2 decreased log odds of retaining first-term sailors by 0.51 compared to the baseline Region1. Therefore, a first-term sailor who is stationed in Region 2, which is the Southeast command, has a log-odds of separating that is higher by 0.51. The symbols next to the P-value (indicated as Pr(>|z|)) indicate the p-value for a significance test for the null hypothesis which states that the coefficient of the particular given explanatory variable is zero; the alternative hypothesis states that the coefficient of the particular given explanatory variable is not equal to zero. No symbol means that the p-value is larger than 0.05 for the null hypothesis,while three "*" means that the p-value for the significance test is very small. These two indicators are important to building an algorithm that can create the fittest match making for first-term sailors. In the next few paragraphs, we break down the BLR summary of AC to demonstrate the logical process of building the final model.

The BLR summary of AC indicates that Region2 is associated with decreased log odds of retaining first-term sailors by 0.51 and the p-value is extremely small which means the null hypothesis is rejected; thus, the effect is statistically significant. Region 2 is Navy Southeast command. Unfortunately, BLR summary can not explain the reasoning behind this behavior. It can only show the relationship between factors and separation. Regions 4 and 7 which are Southwest and D.C are negatively impacting retention for AC as well. However, Region 8 which is oversea locations show the most significant impact in this BLR. It estimates that first-term AC sailors stationed oversea have the highest log odds of leaving the Navy and the p-value is less than 0.001 which means that we have enough evidence to reject the null hypothesis, thus is significant. Since the purpose of our research is not to find reasons for separation, we are simply going to extract the coefficients from the BLR summaries without explaining the cause of this behavior. However, the Navy can utilize this BLR summary to conduct further studies to find the cause of attrition/separation for overseas sailors and sailors from other regional commands.

The race aspect of this summary indicates that all the racial groups outside the baseline are associated with increased separation. The group of sailors who declined to reveal their race shows the strongest impact. The Hispanic-flag which is a broader category of race is not that significant and can be disregarded in this model.

A couple of enlistment type indicates that they are positively associated with retention rate. Type 5 enlistment indicates enlistment over 3 months. This could mean that sailors who served more than 3 months are more likely to stay in the Navy. This indication logically makes sense since sailors are in boot camp in their first three months. During boot camp, many sailors drop out or get separated from the service. The "other enlistment" is a miscellaneous enlistment type that encompassed different codes.

Some of the activities on board show indications that they are significant. Onboard activity 2 represents sea duty and it is negatively associated retention. Perhaps this is due to being in a stressful and intense environment for a while. Interestingly, Onboard activity 3, which is overseas shore duty, shows a great positive impact on retention. On one hand, is associated with an increase chance of separation, but if it is overseas shore duty, the retention increases.

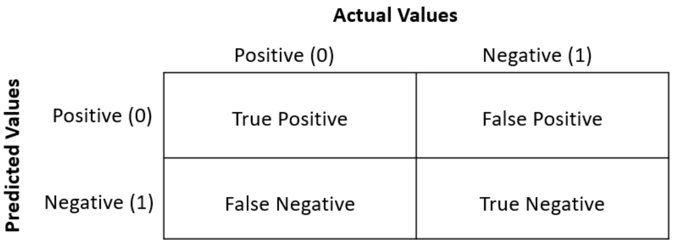
Lastly miscellaneous bonus status of enlistment shows a negative associated with predicted probability retention.

In conclusion, the BLR summary for AC shows factors that are affecting separation for first-term ACs. The coefficients section shows how some categorized factors are affecting the retention of ACs. It is important to note that this is exclusive to first-term ACs. There are 95 other BLR summaries just like this one which indicated different factors that are associated with the retention of their first-term sailors.

For our research, it is not important to understand what each factor represents, but it is important to extract coefficients for significant factors to develop an algorithm that can perform match-making for recruits. Before using the extracted data from BLR, it is important to validate the accuracy of the BLR. The next section explains how we validated the BLR before proceeding to build the match-making algorithm.

4.2 Confusion Matrix

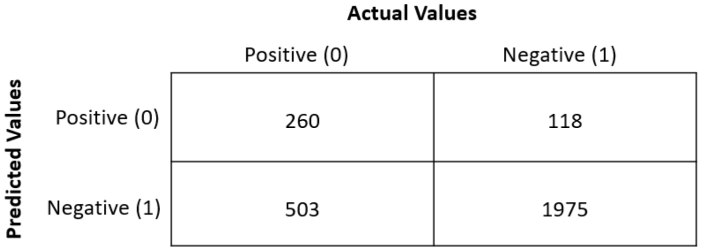
A confusion Matrix measures the performance of a machine learning classification (James et al. 2013). The concept of a confusion matrix is straightforward. It compares predicted values with actual values and then measures various characteristics of machine learning classification including accuracy. There are four elements to the matrix: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). Figure 4.2 displays the confusion matrix table for visual understanding.



		Actual Values	
		Positive (0)	Negative (1)
Predicted Values	Positive (0)	True Positive	False Positive
	Negative (1)	False Negative	True Negative

Figure 4.2. Confusion Matrix Diagram. The columns represent the actual values with 0 as a positive and 1 as a negative. The rows represent the predicted values with 0 as a positive and 1 as a negative. TP happens when both the actual and the predicted values are positive. TN happens when both values are negative. FP happens when the actual value is negative while the predicted value is positive. Lastly FN occurs when the actual value is positive while the predicted value is negative

TP means the model predicted positive and the actual value is positive. TN is when the model predicted negative and the actual value is negative. FP is when the model is predicted positive but the actual value is negative, which is false. Finally, FN is when the model is predicted negative and the actual value is positive, which is false. To put it in context, TP occurs if the model predicted that AC stays after the first term and the actual data indicates that he or she indeed reenlisted. TN is the opposite of TP for which it predicted that AC separates during the first term and he or she indeed is separated. FP and FN are both errors in which predicted values and actual values are not matching. Therefore, the accuracy of the model is measured by the number of TP and TN over the total sample. There is no particular accuracy required to validate the model, but higher accuracy is desirable. Figure 4.3 shows the confusion matrix for AC.



		Actual Values	
		Positive (0)	Negative (1)
Predicted Values	Positive (0)	260	118
	Negative (1)	503	1975

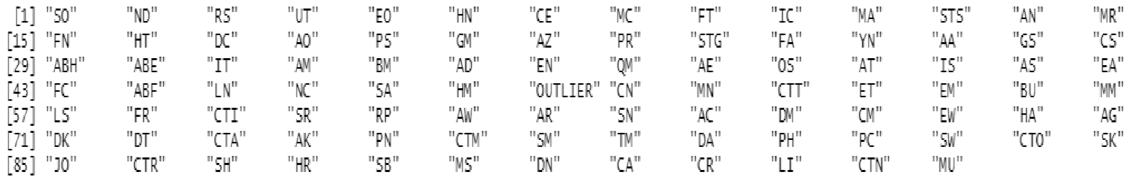
Figure 4.3. Confusion Matrix of AC. There are total of 2856 first-term sailors in AC ratings. The accuracy of confusion matrix is measured by sum of TP and TN divided by the total number of sample. In this case there are 260 TP and 1975 TN; therefore the sum of them is 2235. 2235 divided by 2856 is about 79%.

To calculate the accuracy of BLR model of AC, the sum of TP and TN is divided by the total number of samples. The accuracy of the AC BLR model is about 79%. This is not an overwhelming accuracy, but it is acceptable. This confusion matrix is applied to every BLR model and yields an average of 83% accuracy. 83% overall is an acceptable rating. To improve the accuracy of the model, we also used the Random Forest machine learning method for classification, but the accuracy rating is almost identical to BLR model. Since BLR models have acceptable accuracy ratings, the next step is to create a match-making algorithm.

4.3 Match-Making Algorithm

Our match-making algorithm is straightforward. It simply takes inputs from a sailor and loops through every rating, then lists the ratings with the highest probability of retention in order. The significant coefficients that are stored are the bases of this calculation. To test this model, we used all the attrited and separated sailors as our sample. Figure 4.4 shows the order of rates with the highest probability of retention for a random sample.

This figure indicates that the sailor has the highest probability of staying in the Navy if he or she was a MU. This is just an example of how our match-making model produces the result. This match-making model can be applied to any recruit to help him or her choose the fittest career path.



```
 [1] "SO"      "ND"      "RS"      "UT"      "EO"      "HN"      "CE"      "MC"      "FT"      "IC"      "MA"      "STS"     "AN"      "MR"
[15] "FN"      "HT"      "DC"      "AO"      "PS"      "GM"      "AZ"      "PR"      "STG"     "FA"      "YN"      "AA"      "GS"      "CS"
[29] "ABH"     "ABE"     "IT"      "AM"      "BM"      "AD"      "EN"      "QM"      "AE"      "OS"      "AT"      "IS"      "AS"      "EA"
[43] "FC"      "ABF"     "LN"      "NC"      "SA"      "HM"      "OUTLIER" "CN"      "MN"      "CTT"     "ET"      "EM"      "BU"      "MM"
[57] "LS"      "FR"      "CTI"     "SR"      "RP"      "AW"      "AR"      "SN"      "AC"      "DM"      "CM"      "EW"      "HA"      "AG"
[71] "DK"      "DT"      "CTA"     "AK"      "PN"      "CTM"     "SM"      "TM"      "DA"      "PH"      "PC"      "SW"      "CTO"     "SK"
[85] "JO"      "CTR"     "SH"      "HR"      "SB"      "MS"      "DN"      "CA"      "CR"      "LI"      "CTN"     "MU"
```

Figure 4.4. Order of the lowest to highest probability of retention for a Random Sample. The match-making algorithm lists rates from the lowest to highest probability of retention. In this example, Musician (MU) has the highest chance of retaining the sample sailor. The match-making algorithm produce a rating list for every recruit similar to the figure above.

4.4 Limitation

This match-making algorithm can be applied to any recruit. The idea behind this match-making model is to help recruits find the best fitting rates for them as well as for detailers to manage the health of each rate community more dynamically. However, this model is more useful for detailers than recruits. The input of this model requires certain information that is not readily available to recruits. For instance, recruits are not able to choose or know exactly where they are going before the rate assignment. Also, this model does not take into consideration the interest of each recruit; disregarding the interest of each sailor can have a more negative impact on the retention rate. Therefore, this match-making model should be used as a supplementary tool for detailers to make decisions. For instance, the output from tool should compared to the recruit's preference lists. Each recruit gets to request their preferences for qualified rates. When detailers receive these preferences, they should run a match-making model to see which rates, among these preferences, which rates have the highest probability of retention. This dynamic method can help the Navy capitalize on the benefit of our match-making model while respecting the interest of recruits.

We collected the top five rates with the highest probability of retention for all the separated and attrited first-term enlisted sailors to observe if there are any patterns. The result of this collection shows the limitation of our match making model. Figure 4.5 shows a table of the result.

This chart shows that our match making model heavily favors five ratings: Construction Apprentice (CA), Construction Recruit (CR), Cryptologic Technician Network (CTN),

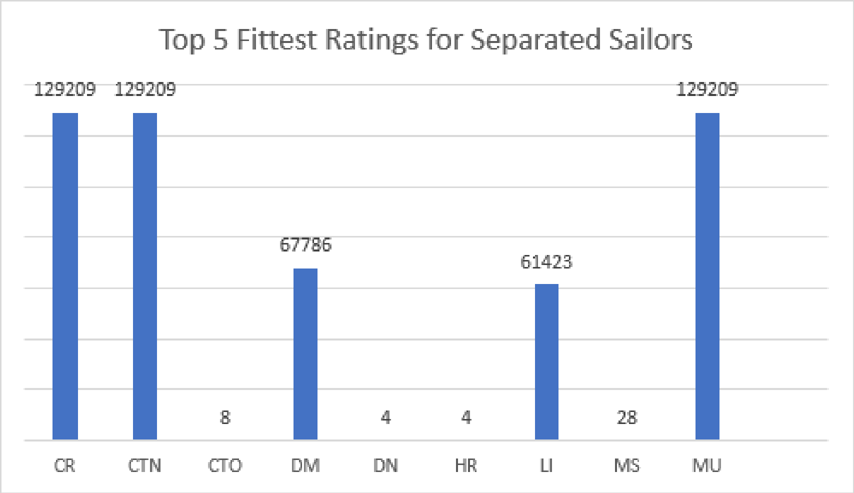


Figure 4.5. Top Five Fittest Ratings for each separated Sailors since 1998. The number on top of each bar represents number of sailors assigned to each rating.

Draftsman/Illustrator (DM), Lithographer (LI), and MU. The possible reason for the biased result is the size of these communities. For instance, the Navy Musician community is composed only about 200 active duty sailors and requires specific skills to join the community. Our model does not take into account the requirement for specific ratings, thus when the model runs it provides ratings with the highest probability of retention despite the requirement for each rating. However, this limitation can be overcome by proactively working with detailers at rating selection. As mentioned in the previous paragraph, this match making model is a secondary tool to help detailers to make a better decision with given ratings; therefore, they do not have to look at all 96 ratings that are ranked from the model.

Figure 4.6 demonstrates how our match-making model is implemented to improve the retention rate of first-term enlisted sailors.

First, a potential recruit visits a local recruiting depot to start the recruiting process. During this process, the potential recruit gets to discuss his or her interests with recruiters. Recruiters do their best to guide the potential recruit in the right direction as well as provide the required documentation to join the Navy. The potential recruit completes the required documents, including ASVAB and physical examination. The potential recruit

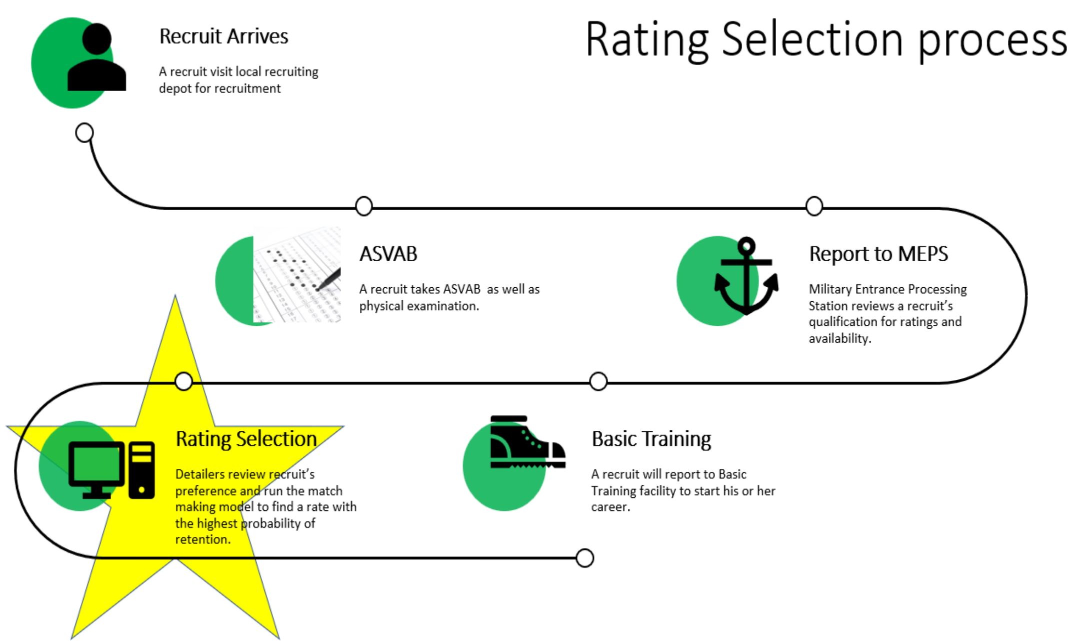

Rating Selection process
Recruit Arrives
A recruit visit local recruiting depot for recruitment
ASVAB
A recruit takes ASVAB as well as physical examination.
Report to MEPS
Military Entrance Processing Station reviews a recruit's qualification for ratings and availability.
Rating Selection
Detailers review recruit's preference and run the match making model to find a rate with the highest probability of retention.
Basic Training
A recruit will report to Basic Training facility to start his or her career.

Figure 4.6. Process of Rating Selection Using the match making Model. The star in the figure represents when the match making model is applied in rating selection process.

report to Military Entrance Process Station (MEPS) with rating preference. Detailers at MEPS review the qualification of the potential recruit as well as his or her preference. Once, the qualification is reviewed, detailers run match-making to determine which qualified ratings have the highest probability of retention. The potential recruit gets ratings from his or her preference list with the highest probability of retention. The potential recruit becomes a recruit and heads to the Navy Basic Training facility to become a fully active sailor.

The benefit of using our match-making model is how easily it can be implemented in the current recruiting process. The Navy does not have to change or fund more budget to incorporate our match-making model. As demonstrated in the previous paragraph, detailers need to run the model at MEPS. Also, it is easy to run our match-making model, thus detailers are able to run the model without having an extensive understanding of the model. This implementation can help improve the retention rate of first-term enlisted sailors while maintaining the current recruiting budget.

Our match-making model should be seen as an evolving model rather than a complete model. There is so much more potential that can make this model better with further research. The next chapter discusses some studies that can be conducted to improve the match making-model.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5: Further Studies

The NES data contains many useful data to develop our match-making model. Although the formatting of data as well as consistency of data input is not ideal for BLR analysis, we manage to clean the data for analysis. However, sailors' duty station[6] had too many factors, thus we had to compromise by grouping them by Navy regional command. This method decreases the accuracy of our model, but we did not have enough information to conduct any further research. The regional research increases the accuracy of this model thus improving the model overall.

5.0.1 Regional Study

As mentioned above, sailors' duty stations are grouped by Navy regional command. This grouping generalizes a large area that fails to capture characteristics of individual naval bases. The United States Navy is a large entity with global influence; therefore, it is a stretch to assign each sailor to only one of 8 different regions. For instance, a sailor stationed in Italy has a different experience than a sailor stationed in Japan, yet our model grouped them under 'overseas'. Therefore, to capture a more specific effect of the duty station, BLR analysis could be run regionally. This study can be performed by separating all the sailors regionally, then using data from ONBOARD-LOC columns to create a smaller area of focus. This way, BLR analysis can capture more significant factors; thus improving the accuracy of the match-making model.

5.0.2 Collaboration with JOIN

The Navy implemented JOIN to help sailors to find their career paths in 2018. Unfortunately, JOIN currently does not have sufficient data to test its efficiency. It will be a few more years, or possibly a decade, before a meaningful study on the effectiveness of JOIN can be conducted. When that analysis is performed, our match-making model should also be analyzed using

[6] Displayed as ONBOARD-LOC in NES data.

the same data. This will allow comparison of the results of each analysis, thus providing feedback to improve our model. It is important to note that our match-making model can also support the JOIN program by providing the rates with the highest probabilities of retention once the JOIN program produces a list of rates for a recruit.

5.0.3 Age

In future studies, it will be interesting to include the age of recruits and observe how they behave in BLR analysis. Generally, people think differently at different ages; therefore, we believe that an 18-year-old recruit thinks and has different goals than a 24-year-old recruit. This difference can lead to choosing different ratings despite having similar characteristics such as gender and race. The future study should include the age of recruits to capture behavior for each age group. This can also improve the accuracy of our match-making model.

Besides the three future studies discussed above, there is an unlimited number of studies to make this model better. And that is the true benefit of this model. As there are more studies and research done on this topic, this model will continuously evolve and help the Navy retain more first-term enlisted sailors. This will help the Navy save a tremendous amount of budget while maintaining its strength and capability of the Navy so that the United States Navy can remain a global power in the future.

List of References

ASVAB Career Exploration Program (2022) Understanding your ASVAB results. Https://www.military.com/join-armed-forces/asvab-and-navy-mos-jobs.html.

Carroll J (2008-09) *Study of Navy enlisted attrition race, ethnicity, and type of occupation*. Master's thesis, Naval Postgraduate School, Monterey, CA, http://hdl.handle.net/10945/3946.

cnicnavymil (2022) Regions. Https://www.cnic.navy.mil/Operations-and-Management/Base-Support/Command-and-Staff/Casualty-Assistance/Funeral-Honors/.

Flyer ES, Elster RS (1983) *First Term Attrition Among Non-Prior Service Enlisted Personnel: Loss Probabilities Based On Selected Entry Factors*. Naval Postgraduate School, Monterey, CA, https://apps.dtic.mil/sti/citations/ADA134365.

GE Larson and SB Kewley (2000) First-term attrition in the navy: Causes and proposed solutions. Technical Report No. 00-27, San Diego, CA, https://apps.dtic.mil/sti/pdfs/ADA388217.pdf.

James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning: With applications in R* (Springer, New York, NY).

Militarycom (2022) ASVAB scroees and navy jobs. Https://www.asvabprogram.com/media-center-article/46.

Navy (2022) Navy enlisted system (nes). Accessed August 28, 2022, https://www.mynavyhr.navy.mil/About-MyNavy-HR/Commands/Navy-Personnel-Command/Organization/NPC-Internal/Information-Management/Corporate-Systems/NES/.

Enns JH (2008) *Cost attrition: Army and Navy results for FY2008*. Master's thesis, Naval Postgraduate School, Monterey, ProQuest Dissertations and Theses database (AAT 3300426).

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. Vienna, Austria, https://www.R-project.org/.

Watson S (2020) Job opportunities in the navy. *Military psychology* (February 4).

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