



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**STATISTICALLY PREDICTING NAVY ENLISTED
SAILORS LENGTH OF STAY**

by

Wonbin Ko

September 2023

Thesis Advisor:

Ruriko Yoshida

Second Reader:

Jefferson Huang

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 September 2023	3. REPORT TYPE AND DATES COVERED Master's thesis		
4. TITLE AND SUBTITLE STATISTICALLY PREDICTING NAVY ENLISTED SAILORS LENGTH OF STAY			5. FUNDING NUMBERS	
6. AUTHOR(S) Wonbin Ko				
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) Recruiting and retaining enlisted sailors has become an increasingly daunting task for the United States Navy. A significant number of Navy Enlisted sailors prematurely end their service after completing their initial term, leading to substantial financial implications for the Navy. This study aims to address this issue by analyzing a comprehensive dataset obtained from the Navy Enlisted System (NES) using survival analysis and various statistical techniques to conduct model selections. The primary objective is to predict years of their service and to identify the factors that correlate with the duration of enlisted sailors' tenure in the Navy. By accurately predicting the length of their service and understanding the influential factors, this research provides valuable insights for informed decision-making in recruitment-related matters. Through survival analysis and modeling techniques, we identify the factors associated with the likelihood of sailors leaving the Navy and use these insights to predict their length of stay. The findings of this research hold potential for the development of proactive measures to improve retention and reduce attrition rates, particularly considering the Navy's recruitment challenges.				
14. SUBJECT TERMS USN, Navy recruiting, enlisted sailor, retention, length of stay, Navy Enlisted			15. NUMBER OF PAGES 59	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**STATISTICALLY PREDICTING NAVY ENLISTED SAILORS
LENGTH OF STAY**

Wonbin Ko
Lieutenant, United States Navy
BA, The George Washington University, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
September 2023**

Approved by: Ruriko Yoshida
Advisor

Jefferson Huang
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Recruiting and retaining enlisted sailors has become an increasingly daunting task for the United States Navy. A significant number of Navy Enlisted sailors prematurely end their service after completing their initial term, leading to substantial financial implications for the Navy. This study aims to address this issue by analyzing a comprehensive dataset obtained from the Navy Enlisted System (NES) using survival analysis and various statistical techniques to conduct model selections. The primary objective is to predict years of their service and to identify the factors that correlate with the duration of enlisted sailors' tenure in the Navy. By accurately predicting the length of their service and understanding the influential factors, this research provides valuable insights for informed decision-making in recruitment-related matters. Through survival analysis and modeling techniques, we identify the factors associated with the likelihood of sailors leaving the Navy and use these insights to predict their length of stay. The findings of this research hold potential for the development of proactive measures to improve retention and reduce attrition rates, particularly considering the Navy's recruitment challenges.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
1.1	Background and Motivation	1
1.2	Significance of the Study	2
1.3	Research Objectives	2
1.4	Scope and Organization.	2
2	Literature Review	5
2.1	Previous Studies.	5
2.2	Method Used	8
2.3	Annual Active Enlisted Inventory Data.	10
3	Methodology	11
3.1	Data Set	11
3.2	Model Fitting and Implementation	19
4	Result	23
4.1	Model Selection.	23
4.2	Survival Analysis	24
4.3	Random Forest	32
5	Conclusion	33
5.1	Summary	33
5.2	Future Work	34
	List of References	37
	Initial Distribution List	39

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

Figure 3.1	Snapshot of Navy Enlisted System (NES) Data (Randomly Selected)	11
Figure 3.2	Work Flow of Data Cleaning	13
Figure 3.3	Final Dataset Sample	18
Figure 3.4	The Survival Probability at Time t	20
Figure 4.1	Likelihood Ratio Test Result	23
Figure 4.2	Multiple Likelihood Ratio Test Result	24
Figure 4.3	Kaplan-Meier Estimate Results	25
Figure 4.4	Kaplan-Meier Estimate Results Graph	25
Figure 4.5	Cox Proportional Hazard Initial Model	26
Figure 4.6	Cox Proportional Hazard Initial Model Graph	27
Figure 4.7	Cox Proportional Hazard Alternate Model Results	28
Figure 4.8	Cox Proportional Hazard Alternate Model Graph	28
Figure 4.9	Aalen's Additive Regression Model Results	30
Figure 4.10	Aalen's Additive Regression Model Results Graph	31
Figure 4.11	Random Forest Results Graph	32

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Table 3.1	Variable Names in the Dataset.	12
Table 3.2	Column Names in the Dataset.	15

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AC	Active Component
ADMITS	Alcohol and Drug Management Information Tracking System
BLR	Binomial Logistic Regression
BRS	Blended Retirement System
BUPERS	Bureau of Naval Personnel
BUPERS-34	Bureau of Naval Personnel (BUPERS) Metrics and Analytic Branch
CPH	Cox Proportional Hazards
DOD	Department of Defense
FTS	Full Time Support
JOIN	Job Opportunities in the Navy
KM	Kaplan-Meier
MASINT	measurement and signature intelligence
NES	Navy Enlisted System
NPS	Naval Postgraduate School
RF	Random Forest
RSF	Random Survival Forest
SRWBR	short range wide band radio
TCP	Transmission Control Protocol
TNS	Terminal Node Statistics
USN	United States Navy

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

This thesis is a comprehensive exploration of the intricate dynamics influencing the attrition rates of first-term Navy enlistees. It delves into the factors that contribute to their decisions to either continue their military service or leave the Navy, thereby providing valuable insights for effective personnel management and enhancing the Navy's operational readiness.

The landscape of military service has evolved significantly over the years, making the retention of first-term enlistees a critical concern. This research employs a variety of advanced statistical techniques, including Survival Analysis, Cox Proportional Hazards Regression Analysis, Aalen's Additive Regression Model, and Random Forest. By analyzing data from the Annual Active Enlisted Inventory, the study identifies significant factors that impact the retention decisions of first-term sailors.

The study's findings are presented in detail, highlighting the diverse range of factors that influence first-term sailors' choices. These factors encompass demographic attributes, service-related activities, external influences, and more. The comprehensive analysis sheds light on the complex interplay of variables that contribute to the attrition rates within the Navy.

The implications of the findings are discussed within the context of naval personnel management. Practical applications of the research outcomes are emphasized, demonstrating how the insights can be utilized to enhance retention strategies and inform workforce planning. The research not only addresses academic curiosity but also holds practical value by providing actionable recommendations for the Navy's personnel policies.

In conclusion, this thesis offers a holistic understanding of the multifaceted factors affecting the attrition rates of first-term Navy enlistees. By combining advanced analytical techniques with real-world implications, this research contributes to the Navy's ability to maintain a skilled and resilient force. The insights derived from this study can shape policies and strategies that optimize the Navy's personnel composition, adapt to evolving challenges, and ensure readiness for the future.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I'd like to begin by extending my heartfelt gratitude to my wonderful and unwaveringly supportive wife, Hailey Hyerin Ko. Hailey, your steadfast presence has been my anchor, accompanying me through both the highs and lows of this academic journey. I must also express my appreciation for our loyal companion, Lia Ko, whose comforting presence provided solace throughout this endeavor. I pledge to continue offering the same level of support and care to both of you.

I'm deeply thankful to my parents and in-laws for their unwavering love and support throughout my life. Their constant encouragement and belief in my abilities have been the driving force behind my accomplishments. Being able to say that I've graduated from the Naval Postgraduate School (NPS), just as my father did, fills me with immense pride and sentiment. This shared educational legacy signifies more than just an academic achievement; it represents a profound and meaningful connection that binds generations of our family to this esteemed institution.

I want to express my heartfelt appreciation to the distinguished faculty of the Operations Research (OR) department and my remarkable cohort. Your contributions have been pivotal to my academic journey at NPS. Without the strong foundation of knowledge and the invaluable guidance you provided during my time here, completing this thesis and earning my degree would have been an insurmountable challenge. The experiences we've shared have been truly enriching, and they will remain treasured memories for me throughout my life.

Lastly, I reserve a special place of gratitude for Professor Yoshida, whose willingness to guide me through this thesis has been instrumental. Your extensive mentorship and unwavering commitment to your students, as evidenced by the multitude you have guided, is a testament to your teaching prowess. I consider myself fortunate to be among your advisees, benefiting from your profound knowledge and invaluable guidance. Thank you, Professor Yoshida, for your continued support and expertise.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:

Introduction

Attrition among first-term Navy enlistees is a multifaceted issue with significant implications for both military readiness and personnel management. The decision-making process of first-term sailors regarding whether to continue their service or seek different paths is not only a subject of statistical analysis but a fundamental concern for ensuring the effectiveness and resilience of our armed forces.

1.1 Background and Motivation

Within the intricate and ever-changing landscape of military service, the question of retaining first-term Navy enlistees has emerged as a critical concern. Deciding whether to continue military service or leave the Navy is a complex outcome shaped by a multitude of factors. This study aims to untangle these factors and understand their broader implications. This pursuit is not just important for effective personnel management, but also essential for enhancing the Navy's overall operational readiness.

Given the evolving nature of both the Navy and the broader societal context, it's important to reevaluate attrition within today's dynamics. As modern military environments experience shifts in demands, expectations, and opportunities, there's a compelling need to thoroughly examine the factors influencing the decisions of first-term Navy enlistees.

The insights gained from this research go beyond mere academic interest. They have the potential to guide practical strategies for managing personnel, contributing directly to the Navy's strategic efforts in workforce planning. By uncovering the intricate factors that shape enlistees' decisions to stay or leave, this study aligns with the Navy's overarching goal of maintaining a capable and resilient force, ready to meet the evolving demands of its mission.

1.2 Significance of the Study

Efforts to address attrition are crucial for the Navy's operational success and long-term sustainability. Previous studies have offered insights into factors such as demographics, career expectations, and individual experiences that might influence enlistees' choices. However, these analyses often fail to account for the complex interplay of numerous variables. With the evolution of the military environment, characterized by changing demographics, new policies, and shifting societal norms, it is imperative to undertake a more comprehensive investigation that considers the intricate relationships among various factors.

1.3 Research Objectives

The primary objective of this study is to provide a thorough understanding of the factors affecting the retention of first-term Navy enlistees. By employing advanced statistical and analytical techniques, we aim to identify not only the key drivers of attrition but also to elucidate the time-dependent effects of these factors. This study examines the dynamic relationship between attributes such as race, gender, education, and service history, and the likelihood of enlistees leaving the Navy. Additionally, the study aims to contribute insights into the potential influence of factors that have evolved over time, such as changes in retirement systems and shifting duty station assignments.

1.4 Scope and Organization

This thesis is structured to provide a comprehensive exploration of the factors that influence the attrition of first-term Navy enlistees. Chapter 2 reviews relevant literature, highlighting previous studies on attrition, their limitations, and the gaps that this research aims to address. Chapter 3 presents the methodology employed in the analysis, detailing the data sources, statistical models, and analytical techniques used to study retention factors. Chapter 4 discusses the results of the survival analyses and presents the findings obtained from the different models. The concluding chapter, Chapter 5, summarizes the key findings, discusses their implications, and proposes avenues for future research. By following this

organizational structure, we aim to present a coherent and insightful exploration of the attrition phenomenon in the United States Navy (USN).

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:

Literature Review

Our research employs a range of survival analytical techniques, including Survival Analysis, Cox Proportional Hazards Regression Analysis, Aalen's Additive Regression (Aalen 1989), and Random Forest. These methods are utilized to identify significant factors predictive of enlisted sailors' retention in the Navy. Survival analysis, commonly applied to assess the time until certain events occur, such as biological organism death or mechanical system failure, will be adapted in our study to predict sailors' separation from the Navy (David and Mitchel 2005). Our study will utilize this methods to find out which factors are highly correlated to the sailor's separation from the Navy. By identifying such factor in this study, we hope to assist Navy to figure out which factors are important in retention of the sailors and it may can assist in improvement of retention.

2.1 Previous Studies

2.1.1 Optimizing First-Term Retention of Sailors

According to Hong (2022), the Navy has employed the same approach for matching ratings to recruits over the past five decades. However, the effectiveness of a novel match-making system introduced in 2018 has not been substantiated (Hong 2022). His study aimed to construct a match-making model tailored to aid first-term sailors in selecting the most suitable career ratings. This approach intends to enhance retention rates, minimize recruitment expenses, and uphold the potency of various Navy communities. Concentrating on first-term enlisted sailors is pivotal due to the financial strain their separation imposes on the Navy and the consequent depreciation of Navy community resilience. An effective match-making model has the potential to substantially curtail attrition and separation rates, delivering substantial benefits to the Navy without incurring supplementary costs (Hong 2022).

Hong's study aims to identify factors associated with separation among first-term sailors by using the Binomial Logistic Regression (BLR) method (Hong 2022). The study utilized

the statistical software R and Python for data cleaning. The BLR method is appropriate for this study as it predicts the probability of a binary response variable, which in this case is whether a first-term sailor is separated or active in the Navy. The research gathered data from the official database of the Navy, known as the Navy Enlisted System (NES), encompassing comprehensive information about the service of all active duty Navy enlisted personnel. The NES data was cleaned to exclude administrative information not relevant to the study. The study's objective involves leveraging insights gained from the analysis of BLR to create a model for matchmaking. This model is designed to assist both recruits and detailers in identifying optimal matching rates characterized by the greatest likelihood of retaining recruits.

The logistic regression analysis of the Navy's first-term Active Component (AC) retention rates revealed that several factors affect separation. In terms of race, it was observed that all racial categories beyond the baseline exhibited a higher likelihood of separation. Notably, individuals who chose not to disclose their race displayed the most pronounced impact. Additionally, enlistment types that displayed a positive correlation with retention rates encompassed Type 5 enlistment, signifying enlistment duration exceeding three months, as well as the category labeled "other enlistment," which encompasses various enlistment codes (Hong 2022). Onboard activities that negatively impacted retention included sea duty, while overseas shore duty had a positive impact. The miscellaneous bonus status of enlistment showed a negative association with predicted probability retention. The coefficients derived from the analysis of significant factors have the potential to be employed in the creation of a matchmaking algorithm for recruits. However, it is crucial to emphasize the necessity of validating the accuracy of the logistic regression analysis results before utilizing the extracted data for this purpose (Hong 2022).

2.1.2 Machine Learning, Time Series, and Survival Analysis for Describing Substance Abuse and Relapse Among U.S. Navy Personnel

The thesis by Minnick (2022) focuses on understanding substance abuse among U.S. Navy service members in order to provide support and empower those involved in addressing the issue. The study utilizes data from the USN Alcohol and Drug Management Information

Tracking System (ADMITS) and employs machine learning models such as classification, time series, and survival analysis to achieve its research objectives.

Minnick's study pinpoints harmful behaviors linked to relapses and forecast incidents of substance abuse (Minnick 2022). Furthermore, it evaluates the duration of survival following relapses to guide decisions on the timing of enrollment in treatment programs. In addition, the research undertakes supplementary examination of U.S. duty locations, USN ranks, and USN occupation subgroups with the purpose of detecting patterns in the prevalence of substance abuse events.

The findings of Minnick suggest that individuals with previous relapses play a crucial role in predicting subsequent relapses. The examination uncovers pronounced cyclic patterns in instances of alcohol abuse and identifies a specific group of individuals who experience relapses at a faster rate than the observed treatment schedules. Furthermore, specific U.S. duty locations, USN ranks, and USN occupation subgroups demonstrate variations in representation concerning occurrences of alcohol abuse across time periods (Minnick 2022).

The survival analysis conducted in the study focuses on time-to-event, specifically the time between substance abuse occurrences. The survival curves demonstrate relatively high survival probabilities for alcohol and other-drug relapses, stabilizing at approximately 80% and 77% respectively. However, closer examination reveals a significant reduction in survival probabilities within the first two months of survival time, followed by a steady decline up to 12 months.

One concern raised in the data is the occurrence of relapses within the first 0-7 days, which may be attributed to duplicate entries or data entry errors. Validation is required to address this issue and potentially increase the survival probabilities over time.

Overall, the study highlights that while the majority of USN service members with at least one substance abuse incident do not relapse, there is a subset of the population that relapses quickly. Additionally, a significant portion of repeat incidents occurs before prior incident screenings, suggesting a need for timely support following substance abuse incidents. The findings offer new perspectives and opportunities to address substance abuse issues among USN service members proactively.

2.2 Method Used

Statistical machine learning refers to the utilization of mathematical algorithms to analyze data and develop models that can detect patterns, structures, and correlations. These models are then used to assist in making forecasts, predictions, and data-driven decisions. The selection of algorithms can vary, and in some cases, multiple algorithms may be used together. However, the specific choice of algorithms typically depends on the analytical objectives and the requirements of the decision-maker.

2.2.1 Survival Analysis

Survival analysis is a statistical approach used to analyze data where the main focus is on the time until a specific event occurs (David and Mitchel 2005). The time variable in survival analysis refers to the duration in years, months, weeks, or days from the beginning of observation until the event takes place. The event can cover diverse outcomes, such as mortality, disease occurrence, relapse, recuperation, or any pertinent individual experience (David and Mitchel 2005).

While it is possible to consider multiple events in the analysis, we will assume a single event of interest for simplicity. If multiple events are involved, such as death from different causes, the analysis may fall into the categories of recurrent event or competing risk problems, which are discussed separately. In the context of survival analysis, the time variable is frequently denoted as the “survival time,” representing the duration for which an individual has persisted or remained under observation during the specified follow-up period. The event is typically referred to as failure, although it can also represent a positive outcome, such as the time to return to work after a surgical procedure (David and Mitchel 2005).

2.2.2 Cox Proportional Hazards Regression Analysis

Survival analysis is a statistical method used to study the timing of events, with death being the typical event of interest (David and Mitchel 2005). However, survival analysis has a broader scope and is employed in various fields under different names, such as “event-history analysis” in sociology and “failure-time analysis” in engineering. The main

focus of survival analysis is to examine the distribution of survival times and understand the relationship between survival and predictor variables, often referred to as covariates. The Cox proportional-hazards regression model, initially proposed by Cox (1972), is a commonly employed technique in survival analysis. For this research, we will consider the departure of sailors from the Navy as an analogous event to “death,” facilitating the application of survival analysis to our dataset.

2.2.3 Aalen’s Additive Regression Model

Aalen’s additive regression model, also known as the Aalen model or the Aalen–Johansen estimator, is a flexible and versatile statistical model used to analyze survival or event time data (Aalen 1989). Introduced by Odd Aalen and Ornulf Borgan in the 1980s, the model estimates the cumulative hazard function, representing the instantaneous risk of an event occurring at any given time.

By assuming that the hazard function is a sum of time-varying regression coefficients multiplied by covariates, the Aalen model accommodates complex survival data without making assumptions about hazard function shape or event time distribution (Aalen 1989). It employs partial likelihood estimation, considering observed event times and censoring information, to estimate regression coefficients and the cumulative hazard function. Widely applied in medical research, particularly in survival analysis, the Aalen model enables the examination of covariate effects, assessment of time-varying effects, and prediction of future events, such as death or disease recurrence. Overall, Aalen’s additive regression model offers a powerful framework for analyzing survival or event time data and estimating time-varying covariate effects on the cumulative hazard function.

2.2.4 Random Forest

Random Forest Regression is a predictive modeling technique operating within the realm of supervised learning. It employs a unique ensemble approach to address regression tasks (James et al. 2013). It combines multiple decision trees with bootstrapping and selecting subset of predictors, forming a “forest,” to make predictions. During training, each tree is built on a different subset of the data, with random subsets of features considered at each

split. This introduces diversity and reduces overfitting. During prediction, the trees independently generate outcomes, which are combined through majority voting (classification) or averaging (regression).

Random Forest can handle numerous features, including categorical variables, without feature scaling, and it provides feature importance estimates for selection and interpretation. With its accuracy, flexibility, and robustness, Random Forest finds applications in finance, healthcare, and ecology, serving as a powerful algorithm for predictive modeling and feature analysis.

2.3 Annual Active Enlisted Inventory Data

The information utilized in this study was acquired from the Bureau of Naval Personnel (BUPERS) Metrics and Analytic Branch (BUPERS-34) situated in Millington, Tennessee. This dataset includes annual records of active enlisted inventory obtained from the NES. The NES is tasked with the duty of producing and upkeeping personnel records for both active and reserve enlisted personnel, playing a pivotal role in determining enlisted strength, validating payment records, and distributing records to field activities.

Ensuring the accuracy and quality of NES data is paramount for the enlisted distribution and promotion processes, as well as for overseeing bodies at the managerial and congressional levels that require information on the active enlisted population. Consequently, NES provides a comprehensive dataset for this research, encompassing all active Navy enlisted personnel and their relevant service-related information. Further elaboration on the dataset will be provided in the next chapter.

CHAPTER 3: Methodology

For our model development, we primarily utilized the statistical software R (R Core Team 2023). Additionally, we employed Python for the data cleaning process.

3.1 Data Set

In our study, the dataset employed was previously utilized in the research conducted by Hong (2022). Referred to as the “annual active enlisted inventory data,” this dataset was sourced from the NES. It encompasses comprehensive details concerning active Navy enlisted personnel, encompassing their service history. However, it’s essential to mention that the NES dataset contains some data that isn’t directly pertinent to our study. This additional information is administrative in nature and doesn’t contribute to our specific research objectives. Due to the large size of the NES dataset, it contains numerous columns, but in Figure 3.1, we display only a sample of the raw data for illustration purposes.

TERM_ENL_ZONE_DD	ADSD	DOD_LOSS_CODE	EAOS	HISPANIC_FLAG	PRES_RAT	PRES_RAT	RACE	GRCSEX	SOFT_EAOS	STR_CLASS	TIME_IN_RATE_DATE	YEARS_ACTIVE_SERVICE_FOR_PAY	
4	B	11/7/2006 0:00	JKQ	11/6/2012 0:00	Non-Hispanic	FC2	E5	WHITE	M	11/6/2012 0:00	Active	1/1/2008 0:00	6
4	B	11/7/2006 0:00	JKQ	11/6/2012 0:00	Non-Hispanic	FC2	E5	WHITE	M	11/6/2012 0:00	Active	1/1/2008 0:00	6
3	E	11/3/2004 0:00	JBK	10/27/2013 0:00	Non-Hispanic	HM3	E4	WHITE	M	10/27/2013 0:00	Active	7/1/2009 0:00	8
4	B	9/21/1992 0:00	NBD	10/28/2012 0:00	Non-Hispanic	LS2	E5	AA	M	10/28/2012 0:00	Active	1/1/2001 0:00	21
4	B	8/25/2004 0:00	JCC	8/22/2014 0:00	Non-Hispanic	STG2	E5	WHITE	M	8/22/2014 0:00	Active	1/1/2008 0:00	9
4	A	6/30/2008 0:00	KCC	6/29/2014 0:00	Non-Hispanic	MMFN	E3	WHITE	M	6/29/2014 0:00	Active	7/5/2012 0:00	5
6	C	10/18/1999 0:00	SFJ	10/15/2012 0:00	Non-Hispanic	QM1	E6	AA	F	10/15/2012 0:00	Active	7/1/2006 0:00	13
4	A	1/27/2009 0:00	MBK	1/26/2013 0:00	Non-Hispanic	YN3	E4	AA	M	1/26/2013 0:00	Active	1/18/2013 0:00	4
6	B	9/13/2005 0:00	KGM	3/30/2014 0:00	Non-Hispanic	EM1	E6	WHITE	M	2/29/2016 0:00	Active	7/1/2011 0:00	8
2	E	11/29/1992 0:00	NBD	6/9/2013 0:00	Non-Hispanic	MA1	E6	WHITE	M	6/9/2013 0:00	Active	7/1/2007 0:00	21
2	B	4/28/2005 0:00	LGH	1/11/2013 0:00	Non-Hispanic	AMEAA	E2	WHITE	M	1/11/2013 0:00	Active	4/19/2012 0:00	8
4	C	6/29/2000 0:00	JCC	3/17/2015 0:00	Non-Hispanic	RP1	E6	WHITE	M	3/17/2015 0:00	Active	7/1/2010 0:00	13
5	A	3/3/2010 0:00	MGU	6/24/2017 0:00	Hispanic	LS2	E5	AA	M	6/24/2017 0:00	Active	1/1/2012 0:00	3
2	E	8/23/1985 0:00	NBD	9/23/2013 0:00	Non-Hispanic	SBCM	E9	WHITE	M	9/23/2013 0:00	Active	7/1/2008 0:00	29
4	A	11/24/2008 0:00	SFJ	12/23/2012 0:00	Non-Hispanic	STS3	E4	MULTIPLE	M	6/23/2013 0:00	Active	7/1/2012 0:00	4
4	C	10/24/2000 0:00	MBK	4/23/2013 0:00	Non-Hispanic	MM1	E6	WHITE	M	4/23/2013 0:00	Active	7/1/2007 0:00	12
4	A	2/11/2013 0:00	JFV	2/10/2017 0:00	Non-Hispanic	SN	E3	MULTIPLE	M	2/10/2019 0:00	Active	2/1/2013 0:00	0
4	A	7/8/2013 0:00	JFC	7/7/2017 0:00	Non-Hispanic	SR	E1	WHITE	F	7/7/2017 0:00	Active	7/8/2013 0:00	0
4	A	9/1/2011 0:00	JFV	8/31/2015 0:00	Non-Hispanic	IC3	E4	AA	M	8/31/2017 0:00	Active	1/1/2013 0:00	2

Figure 3.1. Snapshot of NES Data (Randomly Selected). The Complete NES Data Comprises a Total of 30 Columns, With Each Column's Specific Contents Detailed in Table 3.1.

The initial data sourced from the NES requires preprocessing before being suitable for survival analysis and other modeling tasks, as it contains several extraneous columns. To enhance data comprehension, Table 3.1 offers a comprehensive depiction of each column within the dataset.

Table 3.1. Variable Names in the Dataset

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

Addressing the data pre-processing requirements is a pivotal aspect of this research, ensuring the data's suitability for subsequent survival analysis and modeling. Given the extensive size of the dataset, a pragmatic approach involves breaking down the data cleaning process into a series of distinct steps. Figure 3.2 provides a visual representation of the methodology workflow adopted in this study. The subsequent subsections will delve into an in-depth exposition of the procedural intricacies inherent in each individual step within this workflow.

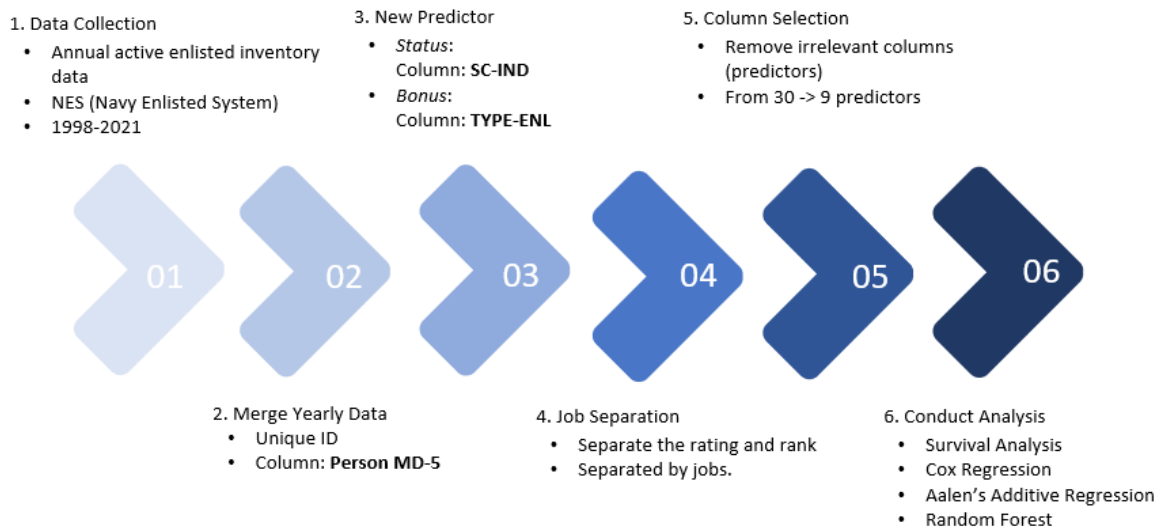


Figure 3.2. The Data Cleaning Workflow Consists of Six Steps Aimed at Preparing the Original NES Data for Analysis. Steps 1 to 6 Provide a Detailed Account of How the Data is Processed to Ensure Its Suitability for Analysis.

3.1.1 Merge Yearly Data

The NES generates annual data for each active Navy enlisted individual, resulting in recurring entries for the same person across years until their Navy separation. For instance, a sailor who enlisted in 2002 and left in 2015 would be present in datasets from 2002 to 2015. This replication causes data redundancy, potentially leading to incongruous outcomes. To mitigate this, data spanning 1998 to 2021 is combined using the distinctive identification code (Column 2: SCI-ID) to prevent duplication and enhance the accuracy and coherence of the analysis outcomes.

3.1.2 Adding New Predictors

The core aim of this study is to identify notable factors impacting Survival analysis and related modeling methods. To denote the distinct circumstances of individuals, the variable requiring attention must indicate their position. In the dataset, the SC-IND column holds two categories: XFXXX, signifying active Navy enlisted personnel, and HXXXX, denoting Navy enlisted members who have left. Consequently, this column serves as the relevant variable. For a more lucid depiction, a column named 'status' is introduced. This new

column utilizes values 0 and 1 to reflect the position of each individual. To elaborate, XFXXX is altered to 0 to denote active enlisted members, while HXXXXX is adjusted to 1 to signify those who have departed. This modification simplifies the understanding of the 'status' variable for subsequent analysis.

Within the dataset, the TYPE-ENL column delineates the diverse manners in which individual sailors joined the Navy. Numeric values from 01 to 95 are employed by the Navy to categorize these enlistment modalities. This column also contains relevant information regarding the availability of signing bonuses for sailors, which is pertinent to our analysis. To enhance the column's structure and diminish the number of factors, akin attributes are amalgamated into nine distinct groups. Furthermore, to incorporate the influence of signing bonuses, a column labeled 'bonus' is introduced. This 'bonus' column classifies enlistment types into three tiers: 'bonus', 'no bonus', or 'unknown', contingent on whether the sailor received a signing bonus. This restructuring of the TYPE-ENL column and the introduction of the 'bonus' column render the portrayal of enlistment types and the influence of signing bonuses more manageable and conducive for subsequent analysis.

The Years column is derived from YEARS-ACTIVE-SERVICE to determine the duration of active duty for each sailor, allowing the study to analyze the influence of years of active duty on separation rates.

3.1.3 Separate Ratings and Ranks

In the Navy, data regarding ratings and ranks are combined and stored together, resulting in entries like "BM1" for a Boatswain's Mate with the rank "Petty Officer First Class." Although this combination is practical during active duty, it poses challenges during analysis because each unique combination is treated as a distinct factor, even if they share the same rating. To overcome this issue, we separated the rating and rank information.

A new column named 'job' was created to store each sailor's rating. This process revealed a total of 96 unique ratings from the dataset covering the years 1998 to 2021. By separating the rating and rank, we can effectively identify significant factors associated with the separation rate, eliminating the problems related to duplication and uniqueness inherent in the previous combined representation. This separation enables a more precise analysis of the data and

facilitates the identification of pertinent factors influencing the separation rate.

3.1.4 Remove Unnecessary Columns

As previously stated, the NES produces data primarily for administrative use, encompassing details not directly relevant to the individual sailors' status. Hence, extraneous information of this nature is omitted from our analysis. In Table 3.2, the cells marked for emphasis signify the columns that will be excluded from our study. These specific columns harbor data that does not align with our research objectives and will consequently be disregarded in the investigation.

Table 3.2. The Table Displays the Column Names in the Dataset, Where Red Highlights Represent Deleted Columns. Each Cell is Labeled With Its Respective Column Number.

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

Here is a list delineating the significance and utility of each dataset column, along with the corresponding column numbers:

1. DAY-DATE: The date of data collection has been omitted due to its lack of relevance to our study.
2. PERSON-MD5: The Personal Identification Code has been excluded, given that the unique identification numbers are solely employed for identifying individuals across various years.
3. SC-IND: SC-IND is harnessed to ascertain the duty status of each sailor. This information is then utilized to generate a column named “Status.”
4. SPI-TAR: The Special Program Indicator has been excluded, as it denotes Reserve or Full Time Support (FTS) forces, which lacks relevance for our study.
5. TYPE-ENL: The Type of Enlistment column furnishes insights into the enlistment methods of individual sailors, along with indications of whether they received signing bonuses.
6. TERM-ENL: The Term of Enlistment column signifies the duration, in years, of the sailor’s ongoing enlistment.
7. ZONE-DD: Selective Reenlistment Bonus Zone divides into 5 Zones, A to E, and separated by years of services.
8. ADSD: The Active Duty Service Date has been eliminated due to its redundancy with the TERM-ENL column.
9. DOD-LOSS-CODE: The type and cause of separation offer insights into the reasons for sailors’ departures. Nonetheless, the high prevalence of missing data in this column renders it unsuitable for analysis purposes.
10. EAOS: The Expiration of Active Obligated Service-Hard has been left out of the study since it does not offer pertinent information relevant to the study.
11. HISPANIC-FLAG: Hispanic or non-Hispanic column, serving the purpose of exploring potential correlations between ethnicity and the separation rate within the Navy.
12. PRES-RATE-ABBR: Present Rate is omitted from the analysis since it is irrelevant for first-term sailors.
13. PAYGRADE: Paygrade has been removed from the study due to its lack of relevance to this study.

14. RACE-GROUP: This column is used to explore whether any correlation exists between race and the separation rate within the Navy.
15. SEX: Gender is employed to scrutinize whether it bears any significance in the context of separations within the Navy.
16. SOFT-EAOS: This column is omitted from the analysis due to its lack of relevance for the current study.
17. STR-CLASS: Type of service is utilized even though there were few missing data because it showed statistical relevance in our study.
18. TIME-IN-RATE-DATE: This column has been excluded from consideration, as it does not hold relevance for this study.
19. YEARS-ACTIVE-SERVICE-FOR-PAY: This column has been disregarded for analysis purposes due to its redundancy.
20. YEARS-ACTIVE-SERVICE: This column is used for identification of individuals within their initial term of service in the Navy.
21. ENL-MGMT-COMMUNITY: Community Manager Specification Code is exclusively used by Navy Human Resource officers to manage community health and does not have any direct impact on the analysis of separation rates among first-term sailors. Therefore, it is not included in the study.
22. LOSS-CHANGE-DATE-OF-OCCURRENCE: Date of loss from Navy is omitted from analysis, as the YEARS-ACTIVE-SERVICE gives redundant data.
23. ONBOARD-ACTY-NAME: Rate Attachment, while potentially valuable for pinpointing individual sailor locations, is unsuitable for analysis due to the challenge of managing numerous distinct inputs.
24. ONBOARD-SS: Onboard Activity Sea/Shore Code identifies sailors' duty station roles but wasn't utilized in the analysis.
25. ONBOARD-DATE-REC: The Date of Assigned Duty is omitted from the study due to its lack of relevance to the analysis.
26. ONBOARD-LOC: This column shows duty location of sailors but not utilized in the analysis.
27. PRES-RATE-CODE-PG: This column shows the rank but it is omitted from the study due to irrelevance.
28. PROS-RATE-CODE-PG: Prospective Rate:paygrade is not needed for our study and is therefore excluded from the analysis.

29. EFFECTIVE-DATE-OF-PAYGRADE: This column is also not relevant for this study so it was excluded.
30. PRES-RATE-AUTH: This column is not needed for our study and was removed from the analysis.

After undergoing various data cleaning methods, the final dataset is presented in Figure 3.3.

TYPE_ENL	TERM_ENL	ZONE_DD	HISPANIC_FL	RACE_GROUP	SEX	STR_CLASS	YEARS_ACTIV	status	Bonus	years
1	4	A	Non-Hispanic	AA	M	Active	0	0	1	0
51	4	E	Non-Hispanic	AA	M	Active	22	1	1	21
50	5	C	Non-Hispanic	API/NATAM	F	Active	13	1	2	13
2	4	A	Non-Hispanic	WHITE	M	Active	0	0	2	1
41	6	C	Non-Hispanic	AA	M	Active	11	0	1	12
2	4	A	Non-Hispanic	WHITE	M	Active	3	0	2	4
41	4	E	Non-Hispanic	WHITE	M	Active	25	1	1	25
96	4	C	Non-Hispanic	WHITE	M	Active	13	0	3	14
1	4	A	Non-Hispanic	DECLINED	M	Active	5	0	1	5
50	5	B	Non-Hispanic	AA	M	Active	7	0	2	7
2	4	A	Non-Hispanic	WHITE	F	Active	0	0	2	0
51	2	B	Non-Hispanic	WHITE	F	Active	9	0	1	9
51	6	E	Non-Hispanic	AA	M	Active	22	0	1	23
2	4	A	Non-Hispanic	WHITE	M	Active	2	0	2	2
2	4	A	Non-Hispanic	WHITE	F	Active	0	0	2	1
51	4	B	Non-Hispanic	WHITE	M	Active	8	0	1	9
51	5	E	Non-Hispanic	WHITE	M	Active	25	0	1	26
1	4	A	Non-Hispanic	AA	M	Active	0	0	1	0
2	4	A	Non-Hispanic	WHITE	F	Active	1	0	2	2
50	6	A	Non-Hispanic	AA	M	Active	3	0	2	4
51	6	D	Non-Hispanic	API/NATAM	M	Active	16	0	1	17
40	3	A	Hispanic	WHITE	M	Active	6	0	2	7
51	6	C	Non-Hispanic	WHITE	M	Active	12	0	1	12
1	4	A	Non-Hispanic	WHITE	M	Active	3	0	1	4
1	4	A	Non-Hispanic	WHITE	M	Active	2	0	1	3
51	6	A	Non-Hispanic	WHITE	M	Active	5	0	1	5
2	4	A	Non-Hispanic	WHITE	M	Active	0	1	2	0

Figure 3.3. Final Dataset Sample. The Final Data Has a Total of 11 Columns. For Analysis Purposes, Three Additional Columns Were Created in the Final Dataset: Status, Bonus, Years.

3.2 Model Fitting and Implementation

In this section, we delve into the statistical methodologies, procedures, and parameters applied in each analysis conducted as part of this study to meet our research objectives.

3.2.1 Hypothesis Test

Likelihood Ratio Test

In statistics, the likelihood-ratio test is a powerful tool used to compare the goodness of fit of two competing statistical models. The test evaluates whether a more constrained model, obtained by imposing specific restrictions on the parameters, fits the data significantly worse than a more flexible model that is fit over the entire parameter space. It is based on the ratio of the likelihoods of the two models (King 1989).

In the context of the likelihood-ratio test, the null hypothesis assumes that the constrained model is true (i.e., the restrictions hold), while the alternative hypothesis assumes that the more flexible model is true (King 1989). The test assesses whether the observed data provides evidence to reject the null hypothesis in favor of the alternative hypothesis.

In R, the likelihood-ratio test can be performed using the `anova()` function (Girden 1991). When applied to models fit using maximum likelihood estimation, the `anova()` function computes the likelihood ratio test statistic along with its corresponding p-value. (Soetewey 2020). The test statistic gauges the discrepancy in log-likelihoods between the two models, while the p-value signifies the likelihood of encountering a test statistic as extreme as the observed one under the null hypothesis.

If the p-value is small (typically below a significance threshold, such as 0.05), it suggests that the more constrained model provides a significantly worse fit to the data compared to the more flexible model, leading to the rejection of the null hypothesis (Soetewey 2020). On the other hand, a large p-value indicates that the data does not provide enough evidence to reject the null hypothesis, implying that the constrained model may be an acceptable representation of the data.

Overall, the likelihood-ratio test performed with `anova()` in R helps researchers make informed decisions about the appropriateness of different statistical models and identify which model provides a better fit to the observed data.

3.2.2 Survival Analysis

Kaplan-Meier Method

The Kaplan-Meier (KM) method is utilized for calculating survival probabilities, drawing from actual survival duration (Kaplan and Meier 1958).

$$S(t_i) = S(t_{i-1}) \left(1 - \frac{d_i}{n_i}\right)$$

Where,

- $S(t_{i-1})$ = the probability of being alive at t_{i-1}
- n_i = the number of patients alive just before t_i
- d_i = the number of events at t_i
- $t_0 = 0, S(0) = 1$

Figure 3.4. The Survival Probability at Time t . Source: Kaplan and Meier (1958).

The estimated probability ($S(t)$) undergoes discrete changes solely at the time of each event occurrence. Furthermore, confidence intervals for the survival probability can be computed. Represented as a plot of the KM survival probability against time, the KM survival curve provides a succinct overview of the data. This curve is particularly useful for estimating metrics like the median survival time (Clark et al. 2003).

The KM model was fitted using the `survfit()` and `Surv()` functions from the survival package (Therneau 2023). As part of our approach, we utilized the `Surv()` function within the formula input for `survfit()`. Within the `Surv()` function, we specified various parameters including

survival time (time), censoring method (event), type of censoring, confidence interval type, and the dataset (Therneau 2023).

Cox Proportional Hazards Regression Analysis

To fit Cox models in the survival package, the `coxph()` function is used, which assumes a semi-parametric approach and does not rely on a specific distribution for survival times. Although its syntax is similar to the previously discussed `survreg()` function, the `coxph()` function does not require a `dist` argument. Utilizing the `summary()` method with the fitted `coxph()` model provides detailed output, offering valuable insights from the analysis.

For constructing our Cox Proportional Hazards (CPH) model, we made use of the `coxph()` and `Surv()` functions from the survival package (Therneau 2023). The `coxph()` function shares common parameters with the `survfit()` function. We set the formula parameter of `coxph()` to be the `Surv()` function (Therneau 2023). In the `Surv()` function, we utilized the same parameters as in the KM model, except for the `conf.type` parameter, which is handled by `coxph()`. Unlike the KM model, the `Surv()` function enables us to utilize dataset features to predict the hazard probabilities of the observations.

3.2.3 Aalen's additive regression model

In this non-parametric model, the effects of covariates, whether constant or variable, are additive to the baseline hazard. The relationship between the hazard function and the failure time (T_i) with a p -dimensional vector X (representing covariates) can be expressed as shown below:

$$\lambda_t(t) = \lambda_0(t) + \gamma_1(t)x_{t1}(t) + \dots + \gamma_p(t)x_{tp}(t) \quad (3.1)$$

Here, the baseline hazard function is denoted as $\lambda_0(t)$. The coefficients γ_i , $i = 1, \dots, p$, represent time-dependent additive effects. Aalen's model allows us to plot cumulative regression functions (Aalen's plots). These plots illustrate the variations in the impact of covariates over time. In simpler terms, Aalen's plots visually depict whether the effect of each covariate remains constant or changes as time progresses (Aalen 1989).

3.2.4 Random Forest

Random Forest (RF) were initially used for regression and classification problems, but the introduction of Random Survival Forest (RSF) extended their capabilities to handle right-censored survival data. Just as with RF, the implementation of RSF operates on similar principles: survival trees are formed through bootstrapped data, random feature selection guides node splitting, trees are commonly grown to substantial depths, and the survival forest ensemble is created by averaging Terminal Node Statistics (TNS) (Lu et al. 2023). RSF proves to be a powerful tool for survival analysis and related applications, effectively handling survival data with censored observations (Lu et al. 2023).

CHAPTER 4:

Result

Our discoveries are the outcome of a comprehensive survival analysis, involving the Cox proportional hazard model, Aalen's additive regression model, and the robust random forest approach.

4.1 Model Selection

4.1.1 Likelihood Ratio Test

To assess the goodness of fit for different models, we conducted likelihood ratio tests using the function `anova()` in R (Girden 1991). Our objective was to select the best fitting model among the multiple models tested. The likelihood ratio test allowed us to compare the competing models and determine which one provides the most suitable fit for our data.

```
##                                     Terms
## 1          DF.HISPANIC_FLAG + DF.SEX + DF.RACE_GROUP + DF.STR_CLASS + DF.Bonus
## 2 (DF.HISPANIC_FLAG + DF.SEX) * (DF.RACE_GROUP + DF.STR_CLASS + DF.Bonus + I(DF.Bonus^2))
##  Resid. Df    -2*LL Test Df Deviance      Pr(>Chi)
## 1    329354 348561.7      NA      NA      NA
## 2    329335 347204.7    = 19  1356.96 2.171376e-276
```

Figure 4.1. Likelihood Ratio Test Result.

Figure 4.1 shows the result of our likelihood ratio test, comparing two different model. The likelihood ratio test was performed to compare two models: the null model and the alternate model. The test statistic yielded a value of 19, with a corresponding p-value of approximately $2.171376e-276$. This low p-value provides strong evidence against the null hypothesis, indicating that the extended model offers a significantly better fit compared to the null model. Consequently, including complex terms such as interactions and nonlinear terms in the model is justified, as they significantly contribute to explaining the survival outcome.

##	Terms					
## 1	DF.HISPANIC_FLAG + DF.SEX + DF.RACE_GROUP + DF.STR_CLASS + DF.Bonus					
## 2	(DF.HISPANIC_FLAG + DF.SEX) * (DF.RACE_GROUP + DF.STR_CLASS + DF.Bonus + I(DF.Bonus^2))					
## 3	DF.HISPANIC_FLAG * DF.SEX * DF.RACE_GROUP * DF.STR_CLASS + DF.Bonus + I(DF.Bonus^2)					
## 4	DF.HISPANIC_FLAG * DF.SEX * DF.Bonus + DF.RACE_GROUP + DF.STR_CLASS + I(DF.STR_CLASS^2)					
## 5	DF.HISPANIC_FLAG + DF.SEX + DF.RACE_GROUP * DF.STR_CLASS * DF.Bonus + I(DF.HISPANIC_FLAG^2)					
## 6	DF.HISPANIC_FLAG * DF.SEX + DF.RACE_GROUP + DF.STR_CLASS + I(DF.STR_CLASS^3) + DF.Bonus					
## 7	DF.HISPANIC_FLAG + DF.SEX + DF.RACE_GROUP * DF.STR_CLASS * DF.Bonus					
## 8	DF.HISPANIC_FLAG * DF.SEX * DF.RACE_GROUP + sqrt(as.numeric(DF.STR_CLASS)) + DF.Bonus					
## 9	DF.HISPANIC_FLAG * DF.SEX * DF.RACE_GROUP * DF.STR_CLASS * DF.Bonus					
##	Resid.	Df	-2*LL Test	Df	Deviance	Pr(>Chi)
## 1	329354	348561.7		NA	NA	NA
## 2	329335	347204.7	=	19	1.356960e+03	2.171376e-276
## 3	329291	347199.8	=	44	4.953864e+00	1.000000e+00
## 4	-17	0.0	=	329308	3.471998e+05	2.440188e-104
## 5	-40	0.0	=	23	0.000000e+00	1.000000e+00
## 6	-14	0.0	=	-26	0.000000e+00	1.000000e+00
## 7	329327	347554.6	=	-329341	-3.475546e+05	5.734421e-108
## 8	329339	348373.5	=	-12	-8.189388e+02	1.435334e-167
## 9	329221	351767.1	=	118	-3.393597e+03	NA

Figure 4.2. Multiple Likelihood Ratio Test Result.

Figure 4.2 shows the result of vigorous likelihood ratio test, comparing multiple models. Among the tested models, Model 2 still stands out as the best choice, as it exhibits the lowest p-value, indicating a significantly better fit compared to the null model. Model 2 includes interactions and nonlinear terms for certain covariates, allowing it to more accurately explain the survival outcome. Therefore, Model 2 is the recommended model for predicting and understanding the survival rates based on the provided covariates so we conducted the analysis using model 2.

4.2 Survival Analysis

4.2.1 Kaplan-Meier Estimate

The KM estimate is calculated using the `survfit()` function. When we provide "1" as the argument, a single survival curve is estimated based on all the data. The `survfit()` function generates essential outcomes, including the tally of subjects and events, along with calculations for the median survival time. Additionally, it provides a 95% confidence interval for the estimated median survival time.

```
call: survfit(formula = Surv(years, status) ~ 1, data = DF2)

      n events median 0.95LCL 0.95UCL
[1,] 329366 33280    28      27      28
call: survfit(formula = Surv(years, status) ~ 1, data = DF2)

  time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10  94525   26120   0.868 0.000817   0.866   0.869
   30    542    7083   0.352 0.009836   0.333   0.372
$quantile
30 40
22 25

$lower
30 40
22 24

$upper
30 40
22 25
```

Figure 4.3. Kaplan-Meier Estimate Results.

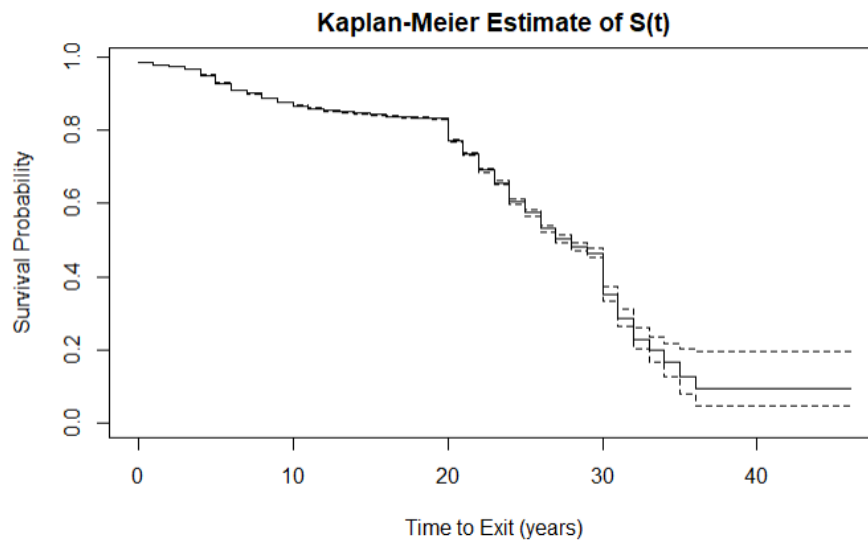


Figure 4.4. Kaplan-Meier Estimate Results Graph.

Figure 4.3 presents the results of `survfit()` for our dataset, allowing us to observe the survival probabilities over time. By using the `summary()` method with the `times` argument set to 10 and 30, we obtained the survival probabilities at 10 and 30 years, respectively. The `times` argument is specified in the scale of the original time variable, which in our study was measured in years. Figure 4.4 displays the estimated survival curve using the `plot()` method,

including the 95% confidence interval by default. However, it's important to note that the KM method does not incorporate the complex models we selected during the likelihood ratio test, limiting its ability to utilize the factors identified as important for survival analysis. As a result, we were unable to fully utilize the model we selected, and the KM method may not fully leverage all the relevant factors for survival analysis.

4.2.2 Cox Proportional Hazard Model

The CPH model was applied to our dataset using the `coxph()` function in R. The model included various interaction terms between *DF.HISPANIC_FLAG*, *DF.SEX*, *DF.RACE_GROUP*, *DF.STR_CLASS*, and *DF.Bonus*, along with nonlinear terms. The coefficients, standard errors, z-values, and corresponding p-values for each term are presented in the output.

```
Call:
coxph(formula = Surv(years, status) ~ (DF.HISPANIC_FLAG + DF.SEX) *
      (DF.RACE_GROUP + DF.STR_CLASS + DF.Bonus + I(DF.Bonus^2)),
      data = DF2)

n= 329366, number of events= 33280
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
DF.HISPANIC_FLAGNon-Hispanic	-0.81527	0.44252	0.21467	-3.798	0.000146 ***
DF.SEXM	-0.76443	0.46560	0.17656	-4.330	1.49e-05 ***
DF.RACE_GROUPPAA	5.72842	307.48190	79.21956	0.072	0.942355
DF.RACE_GROUPAPI/NATAM	5.43799	229.97866	79.21957	0.069	0.945273
DF.RACE_GROUPDECLINED	5.29327	198.99272	79.21957	0.067	0.946727
DF.RACE_GROUPMULTIPLE	5.61664	274.96525	79.21957	0.071	0.943478
DF.RACE_GROUPWHITE	5.99208	400.24628	79.21954	0.076	0.939706
DF.STR_CLASSReserve	2.14032	8.50220	0.56198	3.809	0.000140 ***
DF.STR_CLASSReserve(FTS)	0.30389	1.35512	0.19132	1.588	0.112202
DF.Bonus	2.24696	9.45895	0.32838	6.843	7.78e-12 ***
I(DF.Bonus^2)	-0.71055	0.49137	0.10493	-6.771	1.28e-11 ***
DF.HISPANIC_FLAGNon-Hispanic:DF.RACE_GROUPPAA	-0.07074	0.93171	0.04768	-1.484	0.137912
DF.HISPANIC_FLAGNon-Hispanic:DF.RACE_GROUPAPI/NATAM	0.11066	1.11701	0.06772	1.634	0.102246
DF.HISPANIC_FLAGNon-Hispanic:DF.RACE_GROUPDECLINED	0.50526	1.65742	0.06083	8.306	< 2e-16 ***
DF.HISPANIC_FLAGNon-Hispanic:DF.RACE_GROUPMULTIPLE	0.10331	1.10884	0.06126	1.687	0.091698 .
DF.HISPANIC_FLAGNon-Hispanic:DF.RACE_GROUPWHITE	NA	NA	0.00000	NA	NA
DF.HISPANIC_FLAGNon-Hispanic:DF.STR_CLASSReserve	-1.16860	0.31080	0.52034	-2.246	0.024715 *
DF.HISPANIC_FLAGNon-Hispanic:DF.STR_CLASSReserve(FTS)	0.40412	1.49799	0.18354	2.202	0.027678 *
DF.HISPANIC_FLAGNon-Hispanic:DF.Bonus	1.17337	3.23287	0.30678	3.825	0.000131 ***
DF.HISPANIC_FLAGNon-Hispanic:I(DF.Bonus^2)	-0.33117	0.71808	0.09848	-3.363	0.000771 ***
DF.SEXM:DF.RACE_GROUPPAA	0.27234	1.31303	0.03206	8.495	< 2e-16 ***
DF.SEXM:DF.RACE_GROUPAPI/NATAM	0.12409	1.13212	0.05041	2.462	0.013827 *
DF.SEXM:DF.RACE_GROUPDECLINED	0.37764	1.45884	0.06749	5.596	2.20e-08 ***
DF.SEXM:DF.RACE_GROUPMULTIPLE	0.17236	1.18810	0.05157	3.342	0.000832 ***
DF.SEXM:DF.RACE_GROUPWHITE	NA	NA	0.00000	NA	NA
DF.SEXM:DF.STR_CLASSReserve	0.55262	1.73780	0.46067	1.200	0.230298
DF.SEXM:DF.STR_CLASSReserve(FTS)	-0.08202	0.92125	0.13918	-0.589	0.555643
DF.SEXM:DF.Bonus	0.22276	1.24952	0.24980	0.892	0.372543
DF.SEXM:I(DF.Bonus^2)	-0.02078	0.97944	0.07964	-0.261	0.794169

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 4.5. Result of Initial Cox Proportional Hazard Model.

From the results shown in figure 4.5, we can observe the significance of each term, denoted by the number of asterisks (e.g., "*" for $p < 0.001$, etc.). The hazard ratio, `exp(coef)`, indicates the proportional change in the hazard for a unit increase in the covariate. For example, *DF.HISPANIC_FLAGNon-Hispanic* has a hazard ratio of 0.4425, indicating a 55.75% reduction in the hazard for non-Hispanic individuals compared to Hispanic individuals. In

contrast, the term for *DF.RACE* in the Cox proportional hazard model is showing unusual values for $\exp(\text{coef})$ and is not statistically significant, as indicated by the high p-value. As a result, it may not be a relevant predictor for the survival outcome. Figure 4.6 displays the graph of the CPH model, but it appears to have some issues and may not be an ideal representation. Further investigation and refinement of the model may be necessary to achieve more accurate and meaningful results.

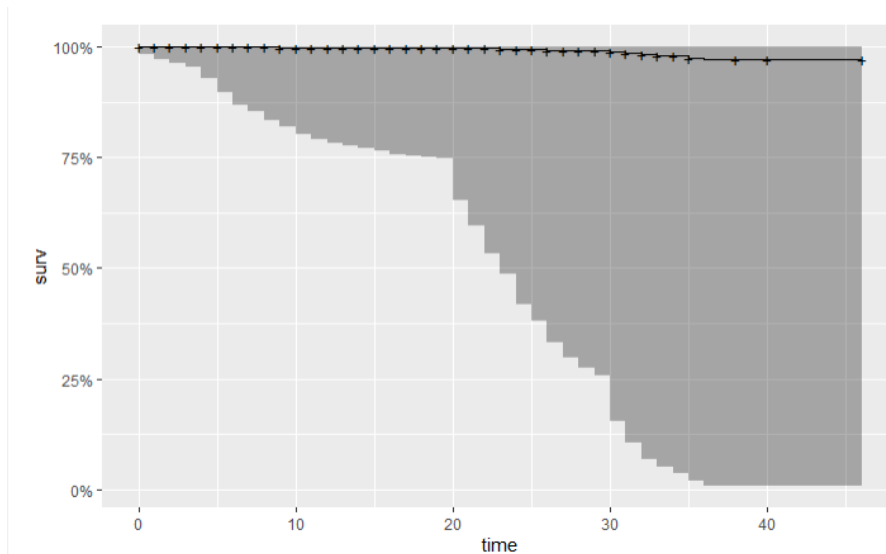


Figure 4.6. Graph of Initial Cox Proportional Hazard Model.

In the model, we made the decision to exclude the factor *DF.RACE_GROUP* to assess whether removing it would lead to improved results.

```

Call:
coxph(formula = Surv(years, status) ~ (DF.HISPANIC_FLAG + DF.SEX) *
      (DF.STR_CLASS + DF.Bonus + I(DF.Bonus^2)), data = DF2)

n = 329366, number of events = 33280

              coef exp(coef) se(coef)      z Pr(>|z|)
DF.HISPANIC_FLAGNon-Hispanic -0.754681 0.470160 0.213958 -3.527 0.000420 ***
DF.SEXM -0.593971 0.552131 0.175465 -3.385 0.000711 ***
DF.STR_CLASSReserve 1.998017 7.374417 0.568268 3.516 0.000438 ***
DF.STR_CLASSReserve(FTS) 0.289282 1.335468 0.192555 1.502 0.133011
DF.Bonus 2.370429 10.701978 0.329478 7.194 6.27e-13 ***
I(DF.Bonus^2) -0.741751 0.476279 0.105327 -7.042 1.89e-12 ***
DF.HISPANIC_FLAGNon-Hispanic:DF.STR_CLASSReserve -0.937594 0.391569 0.516663 -1.815 0.069568
DF.HISPANIC_FLAGNon-Hispanic:DF.STR_CLASSReserve(FTS) 0.400568 1.492672 0.134269 2.174 0.029719 *
DF.HISPANIC_FLAGNon-Hispanic:DF.Bonus 1.110512 3.035913 0.306806 3.620 0.000295 ***
DF.HISPANIC_FLAGNon-Hispanic:I(DF.Bonus^2) -0.313106 0.731172 0.098550 -3.177 0.001487 **
DF.SEXM:DF.STR_CLASSReserve 0.553405 1.742646 0.460108 1.207 0.227387
DF.SEXM:DF.STR_CLASSReserve(FTS) -0.066860 0.935326 0.138992 -0.481 0.630491
DF.SEXM:DF.Bonus 0.172364 1.188110 0.249291 0.691 0.489304
DF.SEXM:I(DF.Bonus^2) -0.008313 0.991721 0.079489 -0.105 0.916708

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
DF.HISPANIC_FLAGNon-Hispanic 0.4702 2.12693 0.3091 0.7151
DF.SEXM 0.5521 1.81117 0.3915 0.7788
DF.STR_CLASSReserve 7.3744 0.13560 2.4211 22.4615
DF.STR_CLASSReserve(FTS) 1.3355 0.74880 0.9157 1.9478
DF.Bonus 10.7020 0.09344 5.6106 20.4135
I(DF.Bonus^2) 0.4763 2.09961 0.3874 0.5855
DF.HISPANIC_FLAGNon-Hispanic:DF.STR_CLASSReserve 0.3916 2.55383 0.1422 1.0779
DF.HISPANIC_FLAGNon-Hispanic:DF.STR_CLASSReserve(FTS) 1.4927 0.66994 1.0402 2.1420
DF.HISPANIC_FLAGNon-Hispanic:DF.Bonus 3.0359 0.32939 1.6639 5.5391
DF.HISPANIC_FLAGNon-Hispanic:I(DF.Bonus^2) 0.7312 1.36767 0.6027 0.8870
DF.SEXM:DF.STR_CLASSReserve 1.7426 0.57384 0.7072 4.2939
DF.SEXM:DF.STR_CLASSReserve(FTS) 0.9353 1.06915 0.7123 1.2282
DF.SEXM:DF.Bonus 1.1881 0.84167 0.7289 1.9367
DF.SEXM:I(DF.Bonus^2) 0.9917 1.00835 0.8487 1.1589

Concordance= 0.599 (se = 0.002 )
Likelihood ratio test= 2517 on 14 df, p=<2e-16
Wald test = 2422 on 14 df, p=<2e-16
Score (logrank) test = 2490 on 14 df, p=<2e-16

```

Figure 4.7. Results of the Alternate Cox Proportional Hazard Model.

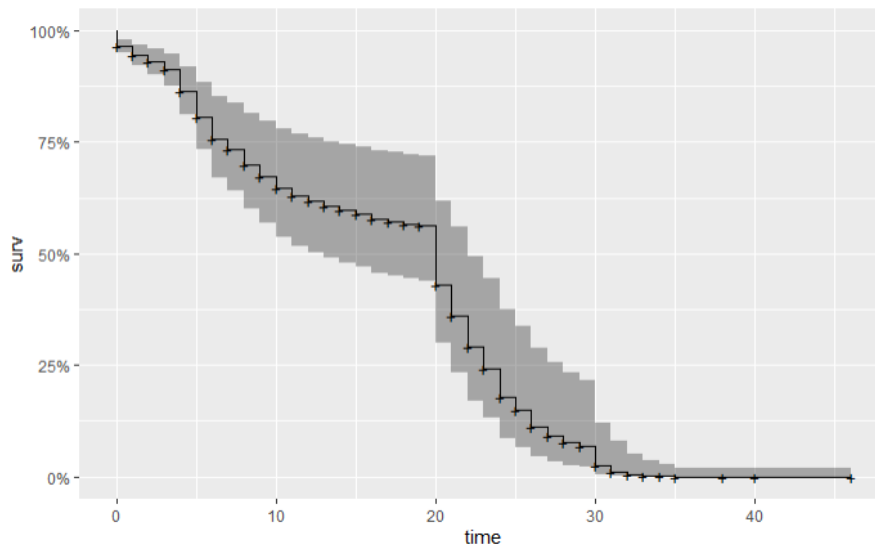


Figure 4.8. Graph of the Alternate Cox Proportional Hazard Model.

As we can see in figures 4.7 and 4.8 for alternate CPH model results, After removing the *DF.RACE_GROUP* factor from the CPH model, the results show some changes in the coefficients and their significance. The terms for *DF.RACE_GROUP* are no

longer present in the model. The remaining factors, *DF.HISPANIC_FLAG*, *DF.SEX*, *DF.STR_CLASS*, *DF.Bonus*, and $I(DF.Bonus^2)$, continue to show significant associations with the survival outcome. The likelihood ratio test indicates that the reduced model without *DF.RACE_GROUP* still provides a significantly better fit compared to the null model, confirming the importance of the remaining factors in explaining the survival outcome. The concordance index is approximately 0.599, suggesting moderate predictive accuracy for the CPH model.

The updated survival analysis with the reduced model reveals more meaningful insights. The survival probabilities show significant drops at around year 4 to year 10 and again at year 20. These periods may indicate critical events or factors influencing the survival outcome. The removal of the *DF.RACE_GROUP* factor seems to have improved the model's ability to capture these important survival patterns. This refined analysis provides a clearer understanding of the survival dynamics in the studied population.

4.2.3 Aalen's Additive Regression Model

The Aalen's additive regression model was employed to analyze the survival data in our dataset, investigating the relationship between various factors and the survival outcome. However, due to the large size of our original dataset, running the `aareg()` function in R proved to be computationally intensive and time-consuming, lasting over 48 hours without yielding results. In order to proceed with the analysis, we opted to take a sample from our original dataset and run the `aareg()` function on this smaller sample. By doing so, we were able to obtain meaningful results and gain insights into the relationship between various factors and the survival outcome within a reasonable amount of time. While using a sample introduces some potential limitations in terms of representativeness, it allowed us to overcome computational constraints and proceed with our survival analysis effectively.

```

Call:
aareg(formula = Surv(years, status) ~ (HISPANIC_FLAG + SEX) *
      (RACE_GROUP + STR_CLASS + Bonus + I(Bonus^2)), data = DF_Small)

n= 3293
13 out of 27 unique event times used


```

	slope	coef	se(coef)	z	p
Intercept	0.01540	8.72e-04	0.001200	0.72500	0.4680
HISPANIC_FLAGNon-Hispanic	-0.06080	-9.88e-04	0.000622	-1.59000	0.1120
SEXM	-0.01910	-1.26e-03	0.001120	-1.12000	0.2620
RACE_GROUPAPI/NATAM	-0.02580	-9.41e-04	0.000320	-2.94000	0.0033
RACE_GROUPDECLINED	0.01130	8.17e-05	0.000534	0.15300	0.8780
RACE_GROUPMULTIPLE	-0.01620	-8.02e-04	0.000366	-2.19000	0.0285
RACE_GROUPWHITE	0.00976	2.50e-06	0.000302	0.00827	0.9930
STR_CLASSReserve(FTS)	-0.01930	-6.24e-04	0.000923	-0.67600	0.4990
Bonus	0.01700	-1.57e-05	0.001700	-0.09250	0.9260
I(Bonus^2)	-0.00639	5.87e-05	0.000563	0.10400	0.9170
HISPANIC_FLAGNon-Hispanic:RACE_GROUPAPI/NATAM	0.01720	5.95e-04	0.000280	2.13000	0.0333
HISPANIC_FLAGNon-Hispanic:RACE_GROUPDECLINED	0.05000	1.12e-03	0.000464	2.42000	0.0154
HISPANIC_FLAGNon-Hispanic:RACE_GROUPMULTIPLE	0.00411	3.73e-04	0.000353	1.06000	0.2910
HISPANIC_FLAGNon-Hispanic:RACE_GROUPWHITE	0.00903	3.04e-04	0.000266	1.14000	0.2530
HISPANIC_FLAGNon-Hispanic:STR_CLASSReserve(FTS)	0.03480	4.13e-04	0.000396	1.04000	0.2980
HISPANIC_FLAGNon-Hispanic:Bonus	0.07730	1.09e-03	0.000821	1.33000	0.1830
HISPANIC_FLAGNon-Hispanic:I(Bonus^2)	-0.02140	-3.06e-04	0.000227	-1.35000	0.1770
SEXM:RACE_GROUPAPI/NATAM	0.00525	3.37e-04	0.000239	1.41000	0.1600
SEXM:RACE_GROUPDECLINED	-0.03440	-5.60e-04	0.000503	-1.11000	0.2650
SEXM:RACE_GROUPMULTIPLE	0.00826	4.57e-04	0.000249	1.83000	0.0669
SEXM:RACE_GROUPWHITE	-0.02150	-2.99e-04	0.000218	-1.37000	0.1700
SEXM:STR_CLASSReserve(FTS)	0.01740	8.19e-04	0.001060	0.77400	0.4390
SEXM:Bonus	0.02440	1.47e-03	0.001630	0.90600	0.3650
SEXM:I(Bonus^2)	-0.00545	-4.46e-04	0.000559	-0.79800	0.4250

```

Chisq=138.21 on 23 df, p=<2e-16; test weights=aalen

```

Figure 4.9. Aalen's Additive Regression Model Results.

As we can see from the figure 4.9 coefficients in the model represent the impact of each variable on the rate of change in the cumulative hazard over time. Several variables displayed statistical significance, indicating their influence on survival. For instance, the *RACE_GROUPAPI/NATAM* and *RACE_GROUPMULTIPLE* categories showed negative coefficients, suggesting a lower hazard rate compared to the reference group. On the other hand, *HISPANIC_FLAGNon – Hispanic* and *HISPANIC_FLAGNon – Hispanic : RACE_GROUPDECLINED* exhibited positive coefficients, implying a higher hazard rate for these groups. Additionally, certain interactions, like *SEXM : RACE_GROUPMULTIPLE* and *SEXM : RACE_GROUPWHITE*, demonstrated borderline significance, indicating potential combined effects of *SEX* and specific racial groups on survival.

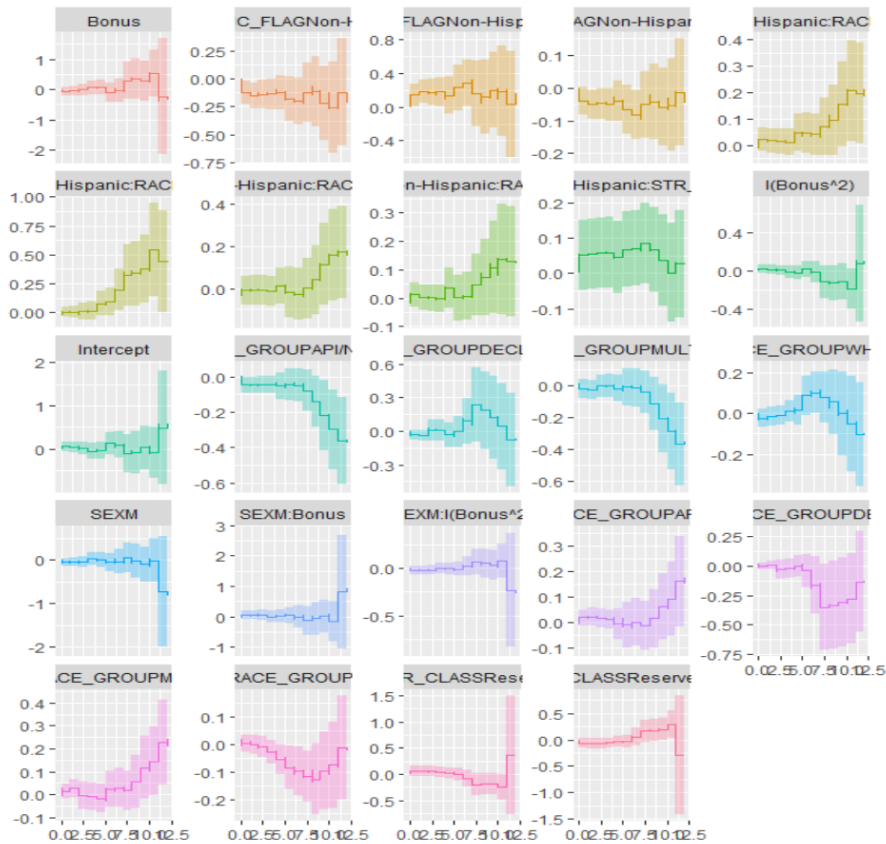


Figure 4.10. Aalen's Additive Regression Model Results Graph.

Figure 4.10 illustrates the graphical representation of our Aalen's additive regression results. This graph provides a clear visualization of the log hazard associated with each factor over time. The plot showcases the time-dependent effects of covariates, revealing how their impact on the hazard rate changes as time progresses. Some factors exhibit increasing effects on the hazard rate, while others show decreasing effects. This visual representation allows us to easily discern the dynamic relationship between the factors and the survival outcome, offering valuable insights into how these variables influence the risk of events occurring at different time points during the study period.

4.3 Random Forest

The RF survival model, created using the `ranger()` function, analyzes the survival data from our dataset. The model is constructed with predictors *Bonus*, *HISPANIC_FLAG*, *RACE_GROUP*, *SEX*, and *STR_CLASS*, while the survival outcome is given by `Surv(years, status)`. To better understand the results, we further calculated the mean survival probabilities for each unique event occurrence time and plotted them against these occurrences.

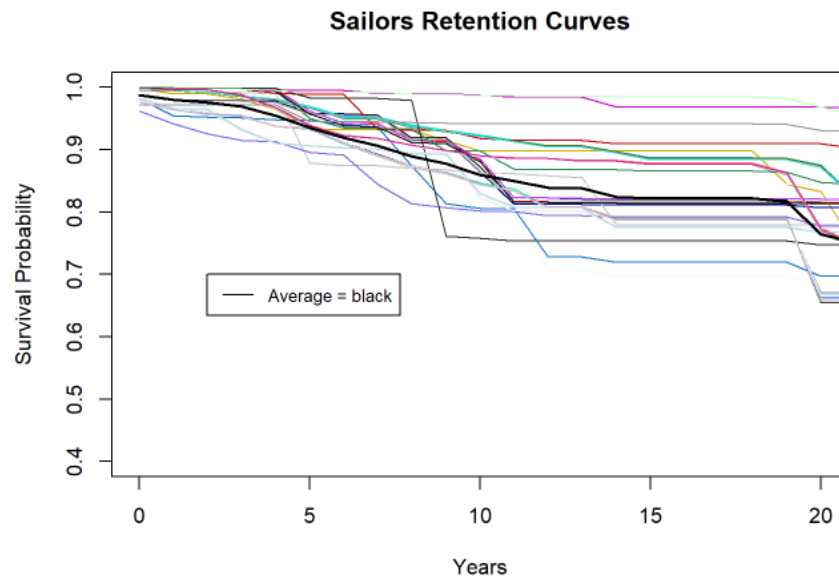


Figure 4.11. Graph of Random Forest Results.

As we can see in figure 4.11, we visualized individual survival curves for 250 randomly selected trees (representing enlisted sailors) using different colors, allowing us to observe the variability in survival probabilities across individuals. The resulting plot provides valuable insights into the survival trends for different enlisted sailors. Furthermore, a black line representing the average survival probability across all trees gives us an overall view of the dataset's survival pattern, helping us better understand the factors influencing sailors' retention.

CHAPTER 5:

Conclusion

5.1 Summary

In this research we focused on estimating probability distribution of the length of service time of first-term Navy enlistees, aiming to identify significant factors influencing their retention in the Navy. Through an extensive literature review, we observed that past studies often oversimplified the complexities of attrition, failing to capture interrelationships among factors. Moreover, many of these studies were based on outdated data, not accounting for the evolving demographic landscape of the military. Our approach differed by utilizing advanced survival analytical methods, including Survival Analysis, Cox Proportional Hazards Regression Analysis, Aalen's Additive Regression, and Random Forest, to gain a comprehensive understanding of retention predictors.

The utilization of these methods allowed us to uncover significant insights. The Survival Analysis revealed the dynamics of sailors' survival probabilities over time, while the Cox Proportional Hazards Regression Analysis identified critical factors influencing retention. Through refining the model and removing less relevant variables, we obtained a clearer picture of the associations between certain demographics, such as ethnicity and age, and the hazard rates of separation. Aalen's Additive Regression Model, though constrained by computational limitations, highlighted the relationship between various factors and the survival outcome, showing both positive and negative coefficients that signify their influence on retention.

The incorporation of the Random Forest model provided a unique perspective by visualizing individual survival curves for a subset of sailors, offering a comprehensive view of the survival trends and the overall average survival probability. Collectively, these findings provide a multi-faceted understanding of the complexities surrounding the retention of first-term Navy enlistees. This research underscores the need for a nuanced approach, considering interactions and time-dependent effects of various factors, in order to develop

more effective strategies for enhancing retention rates in the Navy. As the demographic and societal landscape continues to evolve, these insights will serve as a foundation for targeted interventions and informed decision-making within the Navy's retention policies.

5.2 Future Work

While this study provides valuable insights into the factors influencing the retention of first-term Navy enlistees, there are several avenues for future research that can further enhance our understanding and inform retention strategies.

5.2.1 Regional Study – Duty Station as a Factor

One intriguing avenue for future research involves delving into the impact of duty station on retention rates. The current dataset lumps overseas duty stations together without considering the substantial differences between various global locations. Conducting a regional study that analyzes how factors such as cultural differences, living conditions, and work environments influence attrition could yield valuable insights. For instance, the experiences of sailors stationed in Japan might significantly differ from those stationed in Djibouti. Exploring the nuanced effects of duty station on retention could offer a more granular understanding of the factors at play.

5.2.2 Age Factor

This analysis has yet to delve into the influence of sailors' ages on their retention. Recognizing that an 18-year-old recruit's perspective may differ significantly from that of a 22-year-old recruit, investigating the interplay between age and retention could be enlightening. Younger sailors might face unique challenges and opportunities compared to their older counterparts or vice versa. By incorporating age as a factor, researchers can uncover age-related trends that could aid in tailoring retention strategies based on different age groups.

5.2.3 Effects of Change in Pension System

As the Navy transitions between different pension systems, such as the High-3 retirement system and the Blended Retirement System (BRS), a crucial question emerges: how do these changes affect the survival probabilities of enlisted sailors? Understanding the potential impact of altered retirement incentives on retention is vital. Analyzing the survival probabilities under both systems and identifying any shifts or patterns could provide valuable insights into how changes in pension structure influence sailors' decisions to stay in the Navy.

Incorporating these future research directions will not only deepen our comprehension of the intricate factors influencing retention but also contribute to the development of more targeted and effective strategies for maintaining a robust and motivated enlisted workforce in the U.S. Navy.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

- Aalen G (1989) A linear regression model for the analysis of life times. *Stat Med.* 8(8):907–25.
- Clark TG, Bradburn MJ, Love SB, Altman DG (2003) Survival analysis part i: basic concepts and first analyses. *British Journal of Cancer* 89(2):232–238.
- Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2):187–220.
- David GK, Mitchel K (2005) *Survival Analysis: A Self-Learning Text* (Springer Science+Business Media, Inc., New York).
- Girden ER (1991) *ANOVA: Repeated Measures* (SAGE Publications, Inc., Thousand Oaks, CA).
- Hong S Young (2022) Optimizing first-term retention of sailors. Master’s thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, <https://hdl.handle.net/10945/71069>.
- James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning: With Applications in R* (Springer, New York, NY).
- Kaplan RE, Meier P (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):457–481.
- King G (1989) *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* (Cambridge University Press, New York, NY).
- Lu M, Ishwaran H, Lauer S Michael, Blackstone H Eugene, Kogalur B Udaya (2023) *Random Survival Forests*. <https://www.randomforestsrc.org/articles/survival.html#references>.
- Minnick CJ (2022) Machine learning, time series, and survival analysis for describing substance abuse and relapse among u.s. navy personnel. Master’s thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, <http://hdl.handle.net/10945/70754>.
- R Core Team (2023) *R: A Language and Environment for Statistical Computing*. Vienna, Austria, <https://www.R-project.org/>.
- Soetewey A (2020) *ANOVA in R*. <https://statsandr.com/blog/anova-in-r/#anova>.

Therneau M Terry (2023) *Survival: A Package for Survival Analysis in R*.
<https://github.com/therneau/survival>.

Initial Distribution List

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California



DUDLEY KNOX LIBRARY

NAVAL POSTGRADUATE SCHOOL

WWW.NPS.EDU

WHERE SCIENCE MEETS THE ART OF WARFARE