NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

EVALUATION OF MACHINE LEARNING APPLICABILITY FOR USMC REENLISTMENT

by

Gustavo A. Terrazas

March 2020

Thesis Advisor: Sae Young Ahn
Co-Advisor: James J. Fan

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
EVALUATION OF MACHINE LEARNING APPLICABILITY FOR USMC REENLISTMENT

5. FUNDING NUMBERS

6. AUTHOR(S) Gustavo A. Terrazas

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)

This research examines the applicability of machine learning algorithms to best predict the probability of reenlistment of enlisted first-term Marines. Given the availability of data today, machine learning can be a useful tool to make policy decisions that can impact the future Fleet Marine Force. This thesis uses demographic data, pre-boot-camp data, performance indicators, legal data, awards data, and selective reenlistment bonus indicators to identify factors that contribute to the prediction of reenlistment. This thesis applies data from the Total Force Data Warehouse (TFDW) and fits machine learning algorithms to assess their prediction accuracy. Measuring machine learning models by accuracy alone is not sufficient. An evaluation of top predictors is conducted to choose the best-preforming machine learning algorithm. Given the data used in this thesis, the machine learning algorithm that best predicts the probability of reenlistment is the C5 algorithm. Variables associated with deployment and performance are among the top ten predictors of importance.

14. SUBJECT TERMS
machine learning, talent management, reenlistment, Marine Corps

15. NUMBER OF PAGES
83

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

EVALUATION OF MACHINE LEARNING APPLICABILITY FOR USMC REENLISTMENT

Gustavo A. Terrazas
Captain, United States Marine Corps
BS, California State University San Marcos, 2013

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 2020

Approved by: Sae Young Ahn
Advisor

James J. Fan
Co-Advisor

Marigee Bacolod
Academic Associate, Graduate School of Defense Management

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This research examines the applicability of machine learning algorithms to best predict the probability of reenlistment of enlisted first-term Marines. Given the availability of data today, machine learning can be a useful tool to make policy decisions that can impact the future Fleet Marine Force. This thesis uses demographic data, pre-boot-camp data, performance indicators, legal data, awards data, and selective reenlistment bonus indicators to identify factors that contribute to the prediction of reenlistment. This thesis applies data from the Total Force Data Warehouse (TFDW) and fits machine learning algorithms to assess their prediction accuracy. Measuring machine learning models by accuracy alone is not sufficient. An evaluation of top predictors is conducted to choose the best-preforming machine learning algorithm. Given the data used in this thesis, the machine learning algorithm that best predicts the probability of reenlistment is the C5 algorithm. Variables associated with deployment and performance are among the top ten predictors of importance.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AFQT	Armed Forces Qualification Test
ASR	Authorized Strength Report
CART	Classification and Regression Tree
CHAID	Chi-squared Automatic Integration Detection
CPG	Commands Planning Guidance
EAS	End of Active Service
FTAP	First Term Alignment Program
HRDP	Human Resource Development Process
M&RA	Manpower and Reserve Affairs
MARADMIN	Marine Administrative Message
MCO	Marine Corps Order
MCRISS	Marine Corps Recruiting Information Support System
MEPS	Military Entry Processing Station
MM	Manpower Management
MOS	Military Occupational Specialty
MPP	Manpower Plans and Policy
NJP	Non-judicial Punishment
PFT	Physical Fitness Test
PMOS	Primary Military Occupational Specialty
ROC	Receiver Operating Characteristic
RPM	Retention Prediction Model
SRB	Service Reenlistment Bonus
STAP	Subsequent Term Alignment Program
SVM	Support Vector Machine
T/O	Table of Organization
TAPAS	Tailored Adaptive Personality Assessment System
TFDW	Total Force Data Warehouse
TIS	Time in Service

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I first want to thank the staff that manages the TFDW, who made this thesis possible.

Next I want to thank my advisors, Dr. Tom Ahn and Dr. James Fan, for their knowledge, patience, and support while completing this thesis. I also thank Dr. Christopher Griffin, of Pennsylvania State University, for his machine learning advice.

To the all my professors, student peers, and friends, I thank you for your shared moments with me. I thank you for challenging me to always do better.

To my family, my wife, Maira, and children, thank you for your continued support during my career, especially here at NPS.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The Marine Corps has been operating with antiquated manpower systems; even the Commandant's Planning Guidance (CPG) states that the Marine Corps can no longer afford the inefficiencies of old legacy systems (U.S. Marine Corps, 2019a). The CPG also aligns its priorities in the investment in artificial intelligence and machine learning. The CPG states that the Marine Corps should focus its retention efforts just as the Corps does with precision fires. This can be achieved through data science: improving how we collect data, how we use the data collected, and how we make sense of the data collected. This is critical in improving the talent management process within the Marine Corps.

There are several challenges for the Marine Corps today. As the Corps' top planners align their priorities to those outlined in the 2019 CPG, they will be faced with several tradeoffs between investments in equipment and human capital to fulfill readiness requirement while modernizing the force according to the National Defense Strategy (U.S. Marine Corps, 2019a). The CPG states that talent retention is critical to realize future capabilities, such as the cyber capability, and it is not just a Marine Corps problem but a joint force problem.

A. BACKGROUND

To better understand how the Marine Corps provides retention goals or even how many billets are available in each Military Occupational Specialty (MOS) each year, one must first understand the Human Resource Development Process (HRDP). As the focus of this thesis is on enlisted Marines, this section focuses on the enlisted HRDP. The Marine Corps HRDP can be broken into four main quadrants: requirements, programming, planning, and execution (Barry & Gilikin, 2005).

1. Human Resource Development Process

Each fiscal year, the Marine Corps determines what the force structure should look like. The force structure is derived from manpower requirements and past structures are determined by several stakeholders, such as occupational field sponsors and operational

commanders. The Deputy Commandant for Combat Development and Integration (DC, CD&I) receives guidance from the Commandant of the Marine Corps (CMC). The CMC conducts an analysis of the National Security Strategy and the National Defense Strategy to best provide his guidance. The DC, CD&I then approves the new requirements, which then become the new Table of Organization (T/O). The T/O is then the Marine Corps wartime requirement and is what the Marine Corps tries to build and sustain by buying billets each year. Once planners have the updated T/O, it is used as an input for the next quadrant to account for fiscal constraints.

The programming quadrant accounts for fiscal constraints. Each year the Marine Corps end strength is provided by the National Defense Authorization Act via an Authorized Strength Report (ASR). When a difference between the T/O and the ASR exists, then the Marine Corps Order (MCO) for Manning and Staffing Precedence is taken into account. MCO 5320.12 prioritizes and allocates the planned and available inventory against T/O requirements (U.S. Marine Corps, 2012). The outputs of this quadrant are ASR and the end strength. The ASR is the main document Manpower and Reserve Affairs (M&RA) receives from the Total Force Structure Division. However, this is a multi-directional process. According to MCO 5311.1E, M&RA provides TFSD with the end strength controls in order to produce the ASR (U.S. Marine Corps, 2015). Both of these organizations continuously work with each other.

The planning quadrant is where planners consider the needs of the Marine Corps. Headed by M&RA, planners take into account budgetary constraints, inventory costs, title X constraints, and the Training Input Plan. Additional manning controls are implemented within M&RA to account for patients, prisoners, transients, and trainees. Once these billets are taken out of the initial ASR, a Grade Adjusted Recapitulation (GAR) is developed. The GAR serves as the target for each MOS by rank for the fiscal year. It is in this quadrant where most of the forecasting takes place. Since the Marine Corps must grow or shape its inventory, it plans for accessions, promotions, reenlistments, and losses. Manpower Plans and Policy Division (MPP) heads this portion of the HRDP. MPP takes the ASR and develops plans to meet end-strength requirements. It does this by manipulating the entire human resource life cycle; from recruiting and retention to separations. For example, If the

Marine Corps forecast that it is approaching its authorized end strength requirement, then the Marine Corps might tighten recruitment efforts, and loosen retention or separation requirements. MPP develops plans to meet end strength and provides the staffing goals for the Manpower Management (MM) division to execute. After the initial ASR is adjusted for political and budgetary constraints and a staffing goal is developed, it is then ready for assignment.

The last quadrant is execution, which focuses on the distribution and assignment of personnel. The main input for this quadrant is the staffing goals and available inventory. MM takes the assignable inventory and matches it with available billets to meet every unit's staffing goal. The assignable inventory are the actual Marines available to fill the staffing goal requirements. It is during this portion of the HRDP where marines are given assignment orders to execute.

The assignment of personnel to billets is one of many functions of the planning and execution quadrants. Specifically, the MPP division develops the retention forecasts and goals for each occupational field through the input of several shareholders and occupational field sponsors. Career planners use that information to retain a certain amount of Marines in each MOS.

2. Reenlistment Overview

The terms retention and reenlistments are used interchangeably within the Marine Corps, and they are broken down into two sub-categories: First Term Alignment Plan (FTAP) and the Subsequent Term Alignment Plan (STAP). The reenlistment plans are developed in conjunction with accession plans, promotion plans, and losses plans. These four areas (accessions, reenlistments, promotions, losses) are all connected, and a change in one affects another. For example, a Marine who is executing an End of Active Service (EAS) or received an officer commission is a loss that has to be back filled with either a reenlistment or a new accession.

Each year the Marine Corps publishes details on enlisted retention guidelines for the FTAP through the Total Force Retention System (U.S. Marine Corps, 2019b). Every first term Marine is highly encouraged to submit for retention. The submission goes

through the Marine's chain of command and is scrutinized if the MOS is a fast filling occupation as the Marine Corps wants to ensure it maintains the most qualified. For FTAP, the number of available billets for reenlistments are calculated each year. These available billets are known as "boat spaces". This simply translates to how many first-term Marines, by MOS, the Marine Corps needs to retain in order to sustain the fleet Marine force while meeting end strength requirements. To calculate this, the Marine Corps has a computer model that is managed by MPP. In general, the computer model incorporates regression models and optimization models, taking into consideration career paths, end strength, current inventory, and historical retention rates. The output is the number of recommended boat spaces by MOS and by retention zones (A, B, C, D, or E). Therefore, the Marine Corps, at this point, has a billet list to fill and creates the targets for each MOS. If an MOS is historically difficult to fill, then bonuses are designed to incentivize Marines into those specialties. However, it is unknown if the current model incorporates a prediction of the probability of reenlistment of current inventory. If it does, then an analysis of the predictors is essential to get an accurate prediction. If it does not, then this thesis serves as a starting point to identify what available data can be used in the prediction of the reenlistment decision.

B. PURPOSE OF THIS STUDY

There are various supervised and unsupervised machine learning algorithms that can be used to predict the probability of reenlistment. Many studies use econometric techniques, such as probit or logit, to answer a similar question, but most are limited to the number of variables they use. This thesis addresses the following research questions:

- Which machine learning algorithm best identify the predictors of reenlistment?
- What machine learning algorithm works best in predicting the probability of reenlistment?

C. MOTIVATION

The topic of talent management is aligned with the current CPG and is supported by several stakeholders. For example, unit commanders are interested in retaining the right talent in their units, manpower planners are interested in forecasting, as accurately possible, the potential losses and reenlistments each year to meet the end strength requirement set by the National Defense Authorization Act; and the Marine Corps' Programs and Resources department is interested because it will have to execute budgets that include reenlistment bonuses. Leveraging machine learning and big data can help identify which tools are most effective at predicting if a Marine will reenlist. Any improvement in forecasting the retention of Marines is beneficial for the Marine Corps force planners. The focus of this thesis is on evaluating how the use of machine learning can contribute to the Marine Corps' effort of first-term reenlistment forecasting.

D. SCOPE

This thesis is covers a ten-year period, from fiscal year 2008 to 2018, with over 400,000 first-term Marines. During these years, the Marine Corps end-strength increased from 189,000 to 202,100 in 2010 (U.S. Senate Committee on Armed Services, 2009). The Marine Corps also experienced several drawdowns to a level of 184,000 in 2016 (U.S. Senate Committee on Armed Services, 2015). Bottom line, the drawdowns and increases in end strength affected every occupational field in retaining talented Marines. The availability of available billets either decreased significantly and became more competitive or increased and allowed many Marines to reenlist.

In addition to the ten-year period, this thesis' data source is the Total Force Data Warehouse (TFDW). The TFDW is the Marine Corps' record of historical manpower data and is composed of many sub-databases, such as the database for Marine Corps Recruiting Information Support Systems (MCRISS) and Marine Corps Training Information System. A single source of pre-collected data is used to evaluate the usefulness of the data on various machine learning algorithms. Economic data from the Bureau of Labor and Statistics will not be used to simplify the analysis of the reenlistment decision using common machine learning techniques. Additionally, the new Tailored Adaptive

Personality Assessment System (TAPAS) data will not be used in this thesis since the Marine Corps did not begin collecting data until mid 2018. This means no data is available for the target population of this thesis.

E. THESIS ORGANIZATION

This thesis is divided into five chapters. Chapter II provides a literature review for previous research on reenlistment, prediction of reenlistment, machine learning and manpower research using machine learning. Chapter III provides a detailed description of the data collected from the TFDW, the cleaning process, and methodology used in this thesis. Chapter III will also provide a detailed description of the data collected and cleaning process. Chapter IV describes the analysis and results. Lastly, Chapter V provides a summary of the thesis, recommendations for future research, and a conclusion.

II. LITERATURE REVIEW

This chapter provides a comprehensive literature review related to reenlistment research to identify what algorithms have been attempted in relation to reenlistment. Additionally, the literature review explores different studies that aim to either predict or explaining causation. To better understand how to evaluate machine learning algorithms against the prediction of reenlistment, this chapter explores the available machine learning techniques. Lastly, this chapter looks into recent research involving manpower and machine learning.

A. REENLISTMENT RESEARCH

Many studies have researched reenlistment. The research either aims to forecast a number or identify factors affecting reenlistment. Marine Corps First Term Alignment Program (FTAP) has also been studied in detail by the Center for Naval Analysis. For example, Hattiangadi et al. (2005) provide a great overview of the Marine Corps enlisted manpower plans model. An EAS model was researched and concluded that even though EAS losses of active duty enlisted Marines were easy to plan for. According to Hattiangadi et al. (2005) the Marine Corps uses a steady state model to determine the required number of first term reenlistments by PMOS each year. However, as cited by Hattiangadi et al., EAS losses have been difficult to predict. The authors of this study also commented on the importance of accurately forecasting enlisted losses because the enlisted force is much larger than the officer corps.

Conatser (2006) also looks at FTAP and creates a forecasting model that combines the FTAP and STAP to determine the number of required reenlistment by MOS and grade. Conatser's (2006) forecasting model predicts the number of reenlistment required by MOS each year. He uses a logistic regression and classification trees to predict the reenlistment of a Marine. He concluded that logistic regression provided varying results, and that no one model best predicted reenlistment.

Cole (2014) and Fletcher (2018) both aim to identify the factors affecting reenlistment. Cole (2014) evaluates how retention is affected by the Marine Corps enlisted

reenlistment computerized process. The Marine Corps implemented a 4-tier computerized scoring system around 2011 to aid the leadership to compare Marines against their peers (Cole, 2014). This system uses, as mentioned by Cole 2014, performing metrics such as marksmanship, physical fitness, time in service, and proficiency and conduct mark. She employs a linear regression to evaluate the relationship between EAS and the reenlistment application time. She finds that prior to the computer tier systems, the Marine Corps had a first-come first-served model, and the reenlistment was approved dependent on when the application was submitted. In essence, the new computer tier system aids the Reenlistment Extension Lateral Move process to help identify the quality of the Marine and not just accept the first to reenlist. Fletcher (2018) uses logistic regression and random forest techniques to identify the factors associated with a successful completion of a first term reenlistment. He concludes that the logistic regression and the random forest algorithms both correctly predicts at about the same rate, above 80 percent. He then concludes that the quality of the data collected contributes to the results.

Variable selection varied among these selected studies. Conatser (2006) uses demographic variables to include Armed Forces Qualification Test (AFQT), dependents, ethnicity, marital status, and sex. For service-related variables, he includes grade, selective reenlistment bonus eligibility, Primary Military Occupational Specialty (PMOS), and years of service. Cole (2014) uses basic demographic variables but does not include performance indicators relevant in studying reenlistment behavior such as Physical Fitness Test, Combat Fitness Test, Marine Corps Martial Arts Program, meritorious promotion, and legal misconduct. Fletcher (2018) uses a more robust set of data to assess the successful completion on an enlistment. Random forest is used in this study to identify feature importance and to compare the results found using the logistic regression. He uses more data variables than other research, such as civilian education, marital status at entry, age at entry, and home of record state. No one can really say how much data is needed to accurately predict human behavior, but with higher computational power and various machine learning algorithms, we can apply these algorithms to larger datasets. Therefore, this thesis will use a more robust set of variables to address the research question.

B. CAUSALITY VERSUS PREDICTION

Manpower research is commonly conducted to either answer the question of causality or predict an outcome. Causality is a fundamental concept in economics and social sciences. Arkes (2019) describes that a causation is when one variable has an effect on another. A correlation is when we observe two or more variables move together, positively or negatively. He asserts that the important thing to keep in mind is that a correlation can exist without causation. The reason is that there could be some other variable that is affecting both the independent and dependent variable.

There is research that aims to predict the probability of Marines making a decision to stay in the Marine Corps. For example, Scarfe (2016) assesses significant factors in the decision of junior Marine officers to leave the Marine Corps. He employs a probit model with a dataset of over 3,900 officers. He aims to predict the probability that officer will remain in the Marine Corps after their first obligated contract. A probit model is a regression model that aims to predict the probability of an outcome, in this case staying in the Marine Corps. Arkes (2019) says that regression analysis is most commonly used to quantify how one factor causally affects another. Another use of regression analysis is to forecast or predict an outcome. For example, this can be seen in probit and logit regressions to predict the probability of retention given certain variables. He also says that if the goal is to forecast or predict something, he recommends including as many predicting variables, because getting the true causal effects are not important. However, he recommends avoiding variables subject to reverse causality and variables that are not readily available if frequent replication and forecasting is the goal.

In contrast, some research that uses econometric techniques to identify the causal relationship between a key variable and retention. For example, Ugurbas and Korkmas (2015) employ a multivariate model to find key "determinants of first-term attrition for enlisted and officer Selected Marine Corps Reservists" (p. 19). This is an example of common research done with the focus on trying to get at the causal effect of one variable on another within the retention or attrition realm. In this research, Ugurbas and Korkmas find that being married and having an education level higher than high school is correlated with a higher attrition probability.

C. MACHINE LEARNING

Forecasting Marine reenlistments, attrition, and retention is a constant process. Machine Learning can aid decision makers by identifying patterns in Marine's behaviors that lead to a reenlistment decision, before Marines even make that decision. Additionally, the Marine Corps could use the results of machine learning to fine-tune existing forecasting models and implement service-wide policies. Machine learning algorithms can be organized by either their learning style or likeness. When looking at machine learning algorithms by learning style, we can group them as supervised learning, unsupervised learning, and semi-supervised learning.

1. Supervised Machine Learning

With supervised learning algorithms, input data serves as training data and there is a result, such as reenlist or not reenlist. A model is then created. If the model's prediction accuracy is low, the model may be adjusted to achieve the desired accuracy through a training process (Brownlee, 2019). Uses for supervised learning include regression and classification problems. Regression and classification are problems in which there is an input, an output, and the task to learn the mapping from input to output (Alpaydin, 2014). An example of a supervised learning algorithm is logistic regression.

2. Unsupervised Machine Learning

With unsupervised learning algorithms, there is no objective, just input data. Unlike supervised learning, where the goal is to learn the mapping from an input to an output, there is no supervision and there is only input data (Alpaydin, 2014) The unsupervised learning model is created by taking any input data, and automatically making statistical correlations among the data (Brownlee, 2019). This can be done through mathematical processes or simply by organizing the data. Uses for unsupervised learning include clustering and association rule learning problems (Brownlee, 2019). An example of an unsupervised learning algorithm is K-means.

3. Semi-supervised Machine Learning

Semi-supervised learning algorithms are a hybrid of supervised and unsupervised learning in which there is or is not a defined objective. There still are data inputs to create a model that is used to answer a prediction problem, but the model must first organize the data inputs (Brownlee, 2019). Uses for semi-unsupervised learning are the same as supervised, regression and classification problems.

4. Types of Algorithms

Machine learning can also be organized by their likeness. Literature provides more than 10 types of algorithms. Brownlee (2019) points out some of the commonly known algorithms. He lists common algorithms by their function. He groups the algorithms into regression, instance-based, regularization, decision trees, Bayesian, clustering, deep learning, dimensionality reduction, and ensemble. This thesis will use an even more concise list of machine learning algorithms that are applicable to answering the research question: What machine learning algorithm best predicts the probability of reenlistment?

(1) Regression Algorithms

Regression is a statistical method to show the relationship between two variables, an input and an output. Regression techniques are heavily used in manpower planning and forecasting such as modeling the relationship between a bonus and reenlistment The outputs can either be a continuous variable or a category. For example, if we are interested in predicting the cost of something, then we could implement a linear regression to estimate the coefficients of each factor that contributes to the cost. However, if we are interested in predicting a yes or no answer, then we can use a classifier type of regression like logistic regression. Common regression algorithms include Ordinary Least Squares regression, Linear regression, and logistic regression (Brownlee, 2019).

(2) Instance-Based Algorithms

Instance-based algorithms generate predictions using only specific instances. This type of learning does not maintain abstractions from the specific instances (Aha et al., 1991). Instead this model constructs a hypothesis of the training instance themselves.

Unlike other supervised machine learning algorithms, these type of algorithms take current data, store it in memory, and uses that data to make a decision. Other algorithms, like linear regression, use data to find a function that later is used to predict or classify something, and the data is no longer used. For this reason, instance-based algorithms require large amount of storage (Aha et al., 1991). Some common instance-based algorithms include k-nearest neighbor (kNN) and support vector machines (SVM) algorithms (Brownlee, 2019).

(3) Regularization Algorithms

Regularization algorithms aim to address the issue of overfitting of regression methods. This type of algorithm favors simpler models (Brownlee, 2019). Overfitting occurs when there is data with no apparent pattern or trend, but a function is still fitted to it. This results in large amount of variation in predictions. A way to address this overfitting is by the use of a regularization parameter. This parameter controls the tradeoff between fitting the existing training data while also keeping the parameter small. Therefore, a simpler hypothesis and avoids overfitting. Common regularization algorithms, as cited by Brownlee, include ridge regression and least absolute shrinkage and selection operator.

(4) Decision Trees Algorithms

Decision trees construct models of decisions made creating decision forks until a choice is made (Brownlee, 2019). A popular classification algorithm, decision trees are used in regression and classification problems and are often quick and accurate. These algorithms typically pick the best attributes, meaning the data can be split in half, often with a yes or no decision. Once the best attributes are selected, then we can ask a question, like will a Marine reenlist? We then follow the patch of all the yes or no answers and keep doing this until we get the final answer. Some common decision tree algorithms, as cited by Brownlee, are classification and regression tree (CART), chi-squared automatic interaction detection (CHAID), and conditional decision trees.

(5) Bayesian Algorithms

These types of algorithms apply the Bayes Theorem, which focuses on the probability that event X will occur given Y occurs. Naïve Bayes is simply an extension of

this question that once we have too many events, we ask ourselves, are there any naïve assumptions that can be made to make the math work much easier? This type of algorithm can be applied to both regression or classification type problems to problems such as regression and classification. Some common Bayesian algorithms are Naive Bayes, Gaussian Naive Bayes, and Bayesian network (Brownlee, 2019).

(6) Clustering Algorithms

Clustering algorithms are hierarchal in nature and organizes that data into the best fitting groups given certain attributes (Brownlee, 2019). An example of unsupervised learning, clustering aims to understand the data, versus trying to predict or classify something. It concerns itself with identifying patterns in the data. We typically want to know if there are any subgroups in the data, how big are these groups, and are there any commonalities. Some common clustering algorithms, as cited by Brownlee, are hierarchical clustering and K-means.

(7) Deep Learning Algorithms

Deep learning algorithms build complex neural networks and are heavily used with large amounts of data (Brownlee, 2019). Deep learning is a subset of machine learning that draws inspiration from the human brain structure. An example of this algorithm would be the use of a neural network to classify a type of fruit. We as human could probably tell the difference between an orange or apple and we could even probably apply a supervised learning algorithm to classify appropriately. However, a neural network can do it on its own, it will just require time and more data to train on than supervised learning. Some common deep learning algorithms that Brownlee cites are convolutional neural network and recurrent neural networks.

(8) Dimensionality Reduction Algorithms

Dimensionality reduction algorithms aim to summarize data using less information. Similar to clustering, it uses the natural structure of the data; however, it is done without supervision (Brownlee, 2019). These types of algorithms can be applied to regression and classification problems. This type of algorithm compresses data to reduce the amount of

computer disk space it requires and speeds up calculations. For example, there might be several variables in a dataset that provide similar information, and these types of algorithm, reduces the amount of variable to include the most effective. The result of too many variables could be collinearlity between the independent and dependet variables. Some common dimensionality reduction algorithms, as cited by Brownlee, include principal component analysis, linear discriminant analysis, and principal component regression.

(9) Ensemble Algorithms

Ensemble algorithms are those that use models composed of several weaker models, each trained independently, and in the end, all predictions are combined to provide one unified prediction based on the variables chosen (Brownlee, 2019). There might be a classification problem, which can be address with a regression, a decision tree, or even with a K-Nearest Neighbor algorithm. However, ensemble algorithms implement simple rules, where each individually might not be as effective at, say classifying, but together they can provide a better classification. We can think of this an several trees in a forest, each of them with their optimal solution, but together as a forest only one answer comes out. Some common ensemble algorithms that Brownlee includes are Adaboost, random forest, boosting, and bootstrapped aggregation (Bagging).

D. MANPOWER RESEARCH AND MACHINE LEARNING

The goal of predicting manpower numbers has been attempted using machine learning techniques. However, most military theses only employ one or two algorithms to address their research questions. For example, Conatser (2006) and Fletcher (2018) both use logistic regression, a supervised machine learning algorithm, to predict reenlistment. Cole (2014) employs a linear regression model. Scarfe (2016) and Ugurbas and Korkmas (2015) both employ regression models to examine factors that lead to an output.

Outside of military theses, there is civilian research applying machine learning techniques to predict customer loss. Sabbeh (2018) offers a civilian perspective at getting at predicting churn. Although the focus is customer churn, he employs 10 analytical techniques which include: decision trees (CART), instance-based learning (k-nearest neighbors), support vector machines, and logistic regression. This study is confirmation

that several civilian organizations are using machine learning to predict turnover. Similarly, we can apply machine learning to predict the reenlistment of first term Marines.

Another study predicted whether a service member will separate from the military. Pechacek et al. (2019) developed the Retention Prediction Model (RPM). As part of the Institute for Defense Analysis, the RPM applies machine learning to a very large set of data to predict when a service member will separate from the military. This study is relatively new, and the data range is from 2000 to 2018. There were over 4 million individuals in the dataset used in this analysis. Based on a service member's characteristics, the RPM can estimate the probability that the service member will choose to reenlist. The RPM produces individual level predictions. In their test, two randomly selected service members were selected. Knowing prior to the test that one will separate within a year but not the other, the model correctly identified the right service member 88% of the time. In this study, the authors checkout four models: a feed-forward neural network, gradient boosted trees, a random forest, and a logistic regression. In the end, they compared these four models using the area under the curve and the neural network outperforming the other three. The literature review of this thesis found that there is a lack of research similar to the one conducted by the Institute for Defense Analysis. This thesis aims to provide a comprehensive evaluation of applicable machine learning algorithms using readily available data from the Marine Corps.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA AND METHODOLOGY

This chapter discusses the data collection and preparation process required to conduct an evaluation of machine learning and its application with reenlistment. Additionally, this chapter provides the methodology used for model selection and the testing design.

A. DATA SOURCE

This thesis uses two sources, the TFDW and published Marine Administrative Messages (MARADMINs; U.S. Marine Corps, 2006b, 2007, 2008, 2009, 2010, 2011). To get data from the TFDW, one must have access to the CAC-enabled site where you can submit a Manpower Information Request (MIR). Once your MIR is approved, the data will be made available to download. The TFDW data used in this thesis was made available in December 2019.

To be able to predict whether a Marine decides to reenlist, there are demographic, recruiting, performance, deployment, legal, and reenlistment incentive variables to consider. Data includes the entire active duty, first-term, enlisted population from fiscal years 2008 to 2018.

Demographic data used includes the traditional information of gender, race, ethnicity, date of birth, and marital status. This is important information to identify whether there is a statistical correlation between demographic variables and reenlistment. Reenlistment data is important since every Marine goes through this process, and it is abundant with information like citizenship status, number of dependents, and Scholastic Aptitude Test scores, if any, that can be correlated with the decision to reenlist after their first contract.

Different types of data were collected during a Marine's service: performance, deployment, legal, reenlistment, and acceptance of reenlistment bonuses. After initial training (boot camp), performance information is collected throughout a Marine's time in service. This information includes scores for marksmanship, physical performance, proficiency and conduct marks, and awards received. This performance information can

uncover useful insights as to what type of Marines are choosing to reenlist. As some Marines deploy, it is useful to know whether a deployment to a hostile area is a good predictor of reenlistment. Some Marines also face disciplinary action, and for this there is legal data that captures information about Marines who undergo legal ramifications. Information of what type of legal interactions they have had, non-judicial punishment or court martial, and how many of those they had is useful in predictions. Lastly, there are also ways the Marine Corps incentivizes Marines to reenlist. One of which is reenlistment bonuses. Information of whether a Marine did not reenlist even with a bonus is useful to consider.

The TFDW data received came in 11 different datasets with a total of 171 variables. MARADMINS provided the information for the 12th dataset for this thesis, which adds another 41 variables.

(1) Demographics

The first dataset was Marine demographic data. This dataset yielded 441,434 observations and 51 variables. Data from this dataset includes accession date, MOS, civilian education level completed prior to entry, conduct and proficiency marks in service, race, ethnicity, number of dependents, years of service completed, and marital status.

(2) Recruiting

The MCRISS dataset is the recruiting data, which yielded 438,777 observations and 48 variables. This dataset included location of the military entry processing station (MEPS), citizenship information, date of birth, gender, and information about having a driver's license.

(3) Test Scores

This dataset yielded 441,474 observations and 21 variables, including AFQT scores, cyber test scores, and TAPAS test scores. However, there were no results for the target population for the cyber test scores nor TAPAS scores, therefore those variables were dropped from the data.

(4) Awards

This dataset came in a panel form and yielded over 2.1 million observations with 4 variables, that included the type of award and how many times that award was rewarded to the Marine. There are 99 unique awards in this dataset, ranging from a Letter of Appreciation to a Silver Star Medal.

(5) Physical Fitness

The physical fitness test (PFT) dataset came in panel form and yielded over 1.3 million observations. The data includes multiple entries for Marines for every PFT they were administered during the period from 2008 to 2018.

(6) Rifle Marksmanship

This dataset had four variables and included the date a Marine qualified on the range, their score, and rifle class. The rifle dataset yielded 409,820 observations. However, this is also panel data and there is more than one observation for each Marine. Marines at a minimum will have one qualification on record from boot camp training.

(7) Pistol Marksmanship

This dataset had four variables and included the date a Marine qualified on the range, their score, and pistol class. The dataset yielded 140, 240. This makes sense, since not all junior Marines are required to qualify with the service pistol.

(8) Deployment

The deployment data came from the Global War on Terrorism (GWOT) dataset. This dataset yielded 168,489 observations and 7 variables. The variables provide information on the number of days deployed to a hostile and non-hostile area of every Marine in the sample size.

(9) Legal Data

This dataset yielded 112,805 observations and 3 variables that included type of legal action of every Marine during this period and the date. This is panel data, so there are observations for some Marines in the sample data.

(10) Waiver Data

Any waiver a Marine received to enter the Marine Corps. This dataset yielded 339,245 observations and 8 variables. These variables include information about the type of waiver, how many, and at what level (from the recruiting station to Marine Corps Recruiting Command).

(11) Reenlistment Bonus

This data was also included in this analysis to see whether a Marine that reenlisted did so because of a bonus. The reenlistment data yielded 1,060 observations and 10 variables. These variables include information about the fiscal year the bonus was accepted, the bonus zone, and the unit the Marines was in during the time of accepting the bonus.

(12) Selective Reenlistment Bonus Amounts

The published MARADMINs provided information on the bonus amounts from fiscal years 2007 to 2012. This range was chosen because it provides information of available bonuses for certain occupational fields. It was not until 2008 where the Marine Corps began to publish bonus incentives with dollar amounts on the MARADMINS. Prior to 2008, MARADINS published a multiple next to the MOSs that rated a bonus. The multiple ranged from 0.5 to 5. The multiple is used to calculate a Marine's bonus amount. The other factors that go into the calculations include the new service obligation incurred and the base pay of the Marine's rank. In addition to the range of years used, the focus of the data was on Zone A bonuses. Zone A applied to "those Marines with 17 months to 6 years of active service" (U.S. Marine Corps, 2006a, Para. 2B). Table 1 shows a summary of the data provided gathered from TFDW and MARADMINS.

Table 1. Data Totals by Dataset

Dataset	Source	Number of Observations	Number of Variables
marine	TFDW	441,434	51
mcriss	TFDW	438,777	48
test scores	TFDW	441,474	21
awards	TFDW	2,109,643	4
gwot	TFDW	168,489	7
legal	TFDW	112,805	3
pft	TFDW	1,366,770	4
pistol	TFDW	140,240	4
rifle	TFDW	409,820	4
reenlistment bonus	TFDW	1,060	10
waivers	TFDW	339,245	8
SRB amounts	MARADMINS	961	41
Total		5,970,718	205

TFDW data provided in December 2019; see Chapter III Section A.

MARADMIN data from U.S. Marine Corps (2006b, 2007, 2008, 2009, 2010, 2011)

B. DATA PREPARATION

The statistical software STATA was used to import, merge, and clean the data. In order to merge the datasets together, they all had to be in the same format. For this thesis, the format of choice was to have all of the datasets in long format with one observation representing an individual Marine. To do this, datasets were cleansed to remove any duplicate entries and reshaped from wide to long formats. After the data was merged, the results were 441,474 observations with 430 variables. The variable increase is a result of reshaping the datasets. For example, the awards dataset originally came with four variables. After reshaping the data, the new dataset now has 199 variables.

Dummy variables were created for most of the variable categories. The MEPS is information of where in the United States a Marine accessed from. The variable for MEPS was recoded into dummy variables. A dummy variable is a new variable that takes on the values of 1 or 0; 1 means something is true. Another way to think of a dummy variable is as binary. Dummy variables were created for variables MEPS, citizenship, driver's license state, education tier at contract, ethnicity, marital status, number of dependents, civilian

education level, race, reenlistment recommendation, legal action type, waiver type, and waiver level. For example, the variable “marital status code” was recoded as six dummy variables, one for each type of marital status. Figure 1 shows the marital status code variable transformation.

MARITAL_STATUS_CODE	Freq.	Percent	Cum.
A	124	0.03	0.03
D	7,857	1.79	1.82
L	621	0.14	1.96
M	158,546	36.14	38.10
S	271,481	61.88	99.98
W	100	0.02	100.00
Total	438,729	100.00	

Variable	Obs	Mean	Std. Dev.	Min	Max
Anulled	438742	.000283	.016809	0	1
Divorced	438742	.017908	.132617	0	1
Legal Separated	438742	.001415	.037595	0	1
Married	438742	.361365	.480397	0	1
Single	438742	.618771	.485689	0	1
Widowed	438742	.000228	.015095	0	1

Data from TFDW, December 2019 (see Chapter III Section A)

Figure 1. Dummy Variable Creation Example

There is only one dependent variable used in this thesis. The dependent variable "reenlist" was derived using the variable "separation code." This variable provides a code for the type of separation. If the Marine did not have a separation, then the Marine would have a "0000" entry. Therefore, the dependent variable "reenlist" was coded to equal to 1 if the Marine's separation code was "0000" otherwise, it was coded with a 0.

Summary statistics were derived from current data points. For example, performance data such as physical fitness test scores were used to derive the average, highest, and lowest test scores for each Marine in in the dataset. This similar method was used for rifle and pistol marksmanship scores. Legal instances of data were used to derive the sum of each of the four main types of legal action a Marine received during this period: Non-judicial Punishment (NJP), special court martial, summary court martial, and court martial. Waiver data was transformed to provide a better picture of how many and of what type of waivers each Marine received. The waiver categories range from age to drug waivers. In addition to the type of waiver, the data includes what level the waiver was granted. Figure 2 is an example of the various types of waivers in the original data.

1 WAIVER_CATEGORY_DESC	Freq.	Percent	Cum.
AGE	211	0.09	0.09
ALIEN/HOSTILE COUNTRY	12	0.01	0.10
DEPENDENCY	6,778	3.01	3.11
DRUG INVOLVEMENT	108,472	48.12	51.22
LAW VIOLATIONS	30,622	13.58	64.81
MEDICAL/PHYSICAL/DQ	42,787	18.98	83.79
MENTAL QUALIFICATIONS	44	0.02	83.81
MINIMUM EDUCATION LEVEL	153	0.07	83.88
NOT APPLICABLE	32	0.01	83.89
PRIOR MILITARY SERVICE	745	0.33	84.22
USMC ADMIN/UNIQUE WVR	35,571	15.78	100.00
Total	225,427	100.00	

Data from TFDW, December 2019 (see Chapter III Section A)

Figure 2. Waiver Categories

Some variables were left in their original form. Variables that served as indicators that a Marine reenlisted were Time in Service (TIS) and reenlistment recommendations. TIS captures the duration a Marine is in service and was kept in the form of years completed. This data ranges from 0 to 12 years. For example, if a Marine separated before their contract ended, they would have a number up to 3. Reenlistment codes are a good indicator of the future potential of a Marine. There are 20 different types of reenlistment recommendations in the dataset.

C. DATA SUMMARY

Even though the focus of this thesis is not to identify causal variables, it is beneficial to understand the data this thesis is dealing with. The focus of this section is to provide a summary of the data.

The final dataset used in this thesis has 441,474 observations with 430 variables. The target variable is "reenlist" and is a binary variable, where 1 represents that a Marine reenlisted, and 0 otherwise. According to the dataset, there were over 100 thousand Marines that reenlisted, that is about 25% of the population. Figure 3 shows the actual amounts.

reenlist	Freq.	Percent	Cum.
0	304,323	75.25	75.25
1	100,072	24.75	100.00
Total	404,395	100.00	

Data from TFDW, December 2019 (see Chapter III Section A)

Figure 3. Target Variable Summary Statistics

Of the Marines that reenlisted, less than 10% (9,208) were females. Looking even further, there were more Asians reenlisting compared to blacks, but whites tremendously had the most reenlistments, approximately 8 times more. Specifically, reenlistments comprised of 1,099 (1.1%) Indian or Alaskan native, 10,646 (3.1%) Asian, 10,175 (10.2%) black, 1,121 (1.1%) were Hawaiian or islander, and 83,557 (83.5%) white. Figure 4 is a sample of the summary statistics.

Variable	Obs	Mean	Std. Dev.	Min	Max
id	404,395	213579.7	124249.2	1	436208
reenlist	404,395	.247461	.4315374	0	1
birth_year	404,395	90.36514	8.599425	0	99
female	404,395	.0791503	.2699736	0	1
mcc_	404,395	553.118	480.7042	0	1304
ruc	404,395	16242.9	14195.2	0	87297
fmf_extent~t	404,395	.3705634	2.018531	0	48
TIS	404,395	3.242562	1.398462	0	12
marital_~led	404,395	.0003338	.018268	0	1
marital_~ced	404,395	.0017631	.0419526	0	1
marital_st~c	404,395	2.47e-06	.0015725	0	1
marital_st~p	404,395	.0000618	.0078624	0	1
marital_~ied	404,395	.021662	.1455775	0	1
marital_st~e	404,395	.9759196	.1532991	0	1
marital_~wed	404,395	.0000569	.0075414	0	1
marital_st~n	404,395	.0002003	.0141513	0	1
dependents~c	404,395	.0168449	.1286904	0	1

Data from TFDW, December 2019 (see Chapter III Section A)

Figure 4. Sample Summary Statistics

The dataset captures marital status at the time of recruitment and at the time of reenlistment decision. About 1.2% of Marines that reenlisted were married at the time of recruitment, later that number increased to over 24% at the time of reenlistment. In contrast, about 98.6% of the Marines that reenlisted were single at recruiting, but that number decreases to 75% at the time of reenlistment.

Number of dependents also experienced similar results as marital status. Only about 1.6% of Marines had any dependents at recruiting and about 21.2% had a dependent at reenlistment. Also, the number of dependents ranged from 1 to 7, however, only one Marine had 6 and another had 7 dependents.

TIS for Marines in the dataset ranges from 0 to 12 years, with the max number of observations at 4 years. However, TIS range decreases by one year, to 11 years. The number of Marines between 0 and 3 years of service have the most frequencies due to Marines not having made a reenlistment.

As for Primary MOS (PMOS), over 11% of all Marines that reenlist had the PMOS 0311 Rifleman, followed by PMOS 3531 Motor Vehicle Operator at 5.8%, and PMOS 0621 Field Radio Operator at 4%. Appendix A lists all the variables used in this thesis.

D. METHODOLOGY

Once the final dataset is cleaned and prepared for analysis, a target variable is clearly identified, a model is selected, the test design is implemented, and the model performance is evaluated.

1. Target

Some assumptions were made with the target variable. For instance, Marines in the dataset that did not have a separation code was coded with a "0000" entry, and the condition to determine if someone reenlisted or not was this separation code. If a Marine had a "0000" then the Marine would have a 1 under the "reenlist" variable, 0 otherwise. However, there are Marines in the dataset that have less than 4 years in service, which means these Marine have not yet arrived at the point in their careers where they have to reenlist. So, reenlist means the Marine either reenlisted or have yet to make the decision. The target variable is "reenlist" and is a binary variable.

$$f(x) = \begin{cases} 1 \; if \; reenlist \\ 0 \; if \; not \; reenlist \end{cases}$$

2. Model Selection

Several algorithms could be fitted for classification problems. Logistic Regression, K-Nearest Neighbors, Bayesian Network, Random Trees, XG-Boost Linear, CART, are examples. This thesis fits several models to the dataset with various training-testing splits and conducts a performance evaluation of each. Table 2 shows the applicable machine learning algorithms of interest and initial fitting at different levels of observations. Once the list of applicable machine learning algorithms is narrowed, a detailed evaluation of those algorithms will be conducted using one performance metric at a time.

Table 2. Initial Machine Learning Algorithm Selection

	Number of Observations		
Algorithm	**1,000**	**10,000**	**404,395**
Bayesian Network	X	X	X
C&RT			
C5	X		
CHAID		X	
Decision List		X	
Discriminant			
KNN			
Logistic Regression	X	X	X
LSVM			
Neural Net		X	
Quest			
Random Forest			
Random Trees	X	X	X
Tree-AS			
XGBoost Linear	X		
XGBoost Tree	X		
SVM			

3. Test Design

Before trying to apply any model, the data is split into a training and test set. Normally, the training set is used to fit a model and the test set is kept aside until the end. A common split is 75–25 split, where 75 percent of the data is used for training, and the remainder 25% is reserved for testing (Friedman et al., 2001). This thesis initially follows this split and explores an 80–20 split for analysis.

4. Evaluation Metrics

The type of algorithm that could fit changes dependent on the performance metric that is evaluated. There are several types of performance metrics to evaluate classification problems, such as accuracy, Receiver Operating Characteristic (ROC) curve, and the area under the ROC curve (AUC) (Zheng, 2015). This thesis uses accuracy as its performance evaluation metric since its the common performance metric used in classification problems with binary outcomes. The accuracy of a model is a ratio between the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{\#\ of\ correct\ predictions}{\#\ of\ total\ predictions}$$

IV. ANALYSIS

A. SETUP

The model selection process included one master dataset with demographic, performance, legal, and deployment data of active-duty, first-term enlisted Marines from fiscal years 2008 to 2018. This dataset was further reduced from the original 373 variables to the lowest 64-variable dataset. Each model attempted applied the filtered datasets that comprised of 373, 367, 348, 347 and 64 variables, while keeping all 403,385 observations. The second dataset of 367 variables omits variables that are obvious for predicting reenlistment, such as PEBD, TIS, separation code, EAS.

After observing the collinearity amongst the top predictors, variables associated with a reenlistment recommendation were removed, bringing the dataset down to 348 variables. However, this 348-variable dataset also provided with a collinear top predictor, thus the variable for having a Good Conduct Medal was omitted. Furthermore, a feature selection model was used to filter the data and only using the most important variables with statistical significance greater than 90%, using Pearson's chi-squared test. After filtering the data, the data selection has only 64 variables. A 75% training, 25% testing and a 50% training, 50% testing partition was utilized by every model attempted.

Of the 17 models initially identified as a potential to answer the research question, four were selected for further refinement and comparison with the original dataset. These four are: CART, CHAID, Linear SVM, C5 and k-means. Additionally, after using a feature selection process to filter the data to only include those fields with a statistically significance of 90 percent or higher, the models applied to this dataset with 64 variables are Logistic Regression, CART, CHAID, Bayes Network, C5, and Random Trees. Table 3 summarizes the model selection at various amounts of variables. An "X" on this table means that the model was selected and applied to the data. Only models that could successfully apply the algorithms to the dataset were selected. If an algorithm was unsuccessful at using vast amount of the data, it was not selected.

All models were tested for their prediction accuracy and compared against each other to see which model performs the best. Additionally, each model provides a set of important predictors. The k-means model is an unsupervised machine learning model that uses clusters to segregate the data without using a target variable like the other models. The result is in number of clusters, the cluster quality, and the cluster size. The next section provides the results of every model and its prediction accuracy comparison.

Table 3. Model Selection Table

	Dataset				
Model	372 (all variables)	367 (excludes PEBD, TIS, separation code, and EAS)	348 (excludes reenlistment recommendations)	347 (excludes Good Conduct Medal)	64 (only variables with statistical significance > 0.9)*
C5			X	X	X
Logistic Regression					X
Decision List					
Bayesian Network					X
Discriminant					
KNN					
Linear SVM	X	X	X	X	
Random Trees					
SVM					
Trees-AS					
XGBoost Linear					
XGBoost Tree					
CHAID	X	X	X	X	X
Quest					
CART	X	X	X	X	X
Random Forest					X
Neural Net					
K-Means (Unsupervised)		X	X	X	

* Pearson chi-squared test is used to rank the importance of predictors

B. MODEL EVALUATION

The amount and type of variables influence which algorithm is appropriate to predict reenlistment. The data partition also has an effect on how each model performs. This section will evaluate each algorithm against the five variations of the dataset and on two data partition ratios. All datasets maintained their original number of observations.

1. CART

The CART models were among the most consistent models across all variable amounts. When looking at a 75:25 data partition ratio, the CART model performs best with both the original dataset with 372 variables and the dataset after a feature selection process was applied. The feature selection process is a way to reduce the number of variables to those only statistically significant. In this case, only variables with a statistical significance greater than 0.90 were included. The prediction accuracy for CART with these two datasets is 98.83 percent. Similarly, when a 50:50 data partition ratio is applied, the CART model performs best with the original dataset and the 64-variable dataset with an accuracy of 98.84%. Table 4 summarizes the CART models prediction accuracies.

Table 4. CART Model Prediction Accuracy

Variable Structure	CART Prediction Accuracy with 75:25 Data Partition Ratio	CART Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	98.83%	98.84%
367 (excludes PEBD, TIS, separation code, and EAS)	97.50%	96.96%
348 (excludes reenlistment recommendations)	90.25%	90.19%
347 (excludes good conduct medal)	82.48%	82.44%
64 (only variables with statistical significance > 0.9)	98.83%	98.84%

Looking at predictor importance, variables associated with separation, such as a reenlistment recommendation, EAS date, or even when a Marine started their active duty service, ranked high in predictor importance. These top variables are collinear to the target variable and is why they are the top predictors. To address this, those variables were omitted in subsequent models. Figure 5 shows the top 10 predictors using the dataset with 347 variables and a 75:25 data partition. The top predictor "COMBATDA" represents the number of days a Marine have served in combat. This variable ranges from 0 to 1,056 days. The relationship between combat deployments and reenlistment is complex. A Marine might be satisfied as a result of serving their country in a combat zone, however, there are also many serious risks of combat. The next top predictor is "AWARDSSD" which represents a Sea Service Deployment Ribbon. This ribbon is personal award that is awarded after being deployed for 90 days consecutively (U.S. Marine Corps, 2003). The sea service deployment ribbon is correlated with a deployment, since a Marine had to be deployed just to rate the ribbon. The third predictor, in Figure 5, "CON_AVE_" represents a Marine's average conduct marks in the current enlistment. Proficiency and conduct marks are given semi-annually to all junior Marines and contribute to their composite score, which is used for promotion. Conduct marks range from 0 to 5.0 and are a metric used in a RELM's tier calculation. The majority of Marines in the dataset have average conduct marks between 4.1 and 4.7. The rest of the top predictors are associated with a Marine's age, whether they are married and have any dependents. Marital status and children can influence a Marine to reenlist if the alternative is less desirable.

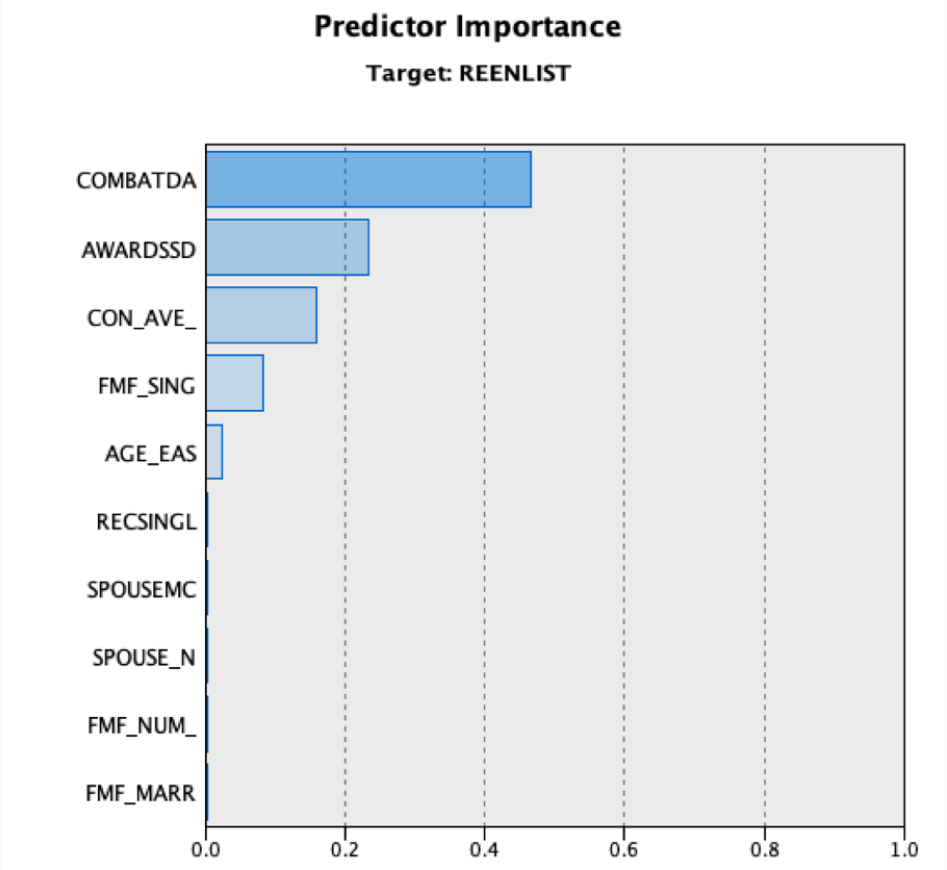

Figure 5. Top 10 Predictors for CART Model with a 75:25 Partition on 347-Variable Dataset

2. CHAID

The CHAID models were selected and applied to all datasets similar to the CART models. Using a 75:25 data partition ratio, the CHAID model performs the best with the original dataset of 372 variables. Similarly, the CHAID model performs the best with the original dataset under a 50:50 data partition ratio. Table 5 summarizes the CHAID models prediction accuracies.

Table 5. CHAID Model Prediction Accuracy

Variable Structure	CHAID Prediction Accuracy with 75:25 Data Partition Ratio	CHAID Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	98.48%	98.46%
367 (excludes PEBD, TIS, separation code, and EAS)	97.78%	97.74%
348 (excludes reenlistment recommendations)	89.12%	89.06%
347 (excludes good conduct medal)	82.49%	82.46%
64 (only variables with statistical significance > 0.9)	97.51%	97.51%

The CHAID model's predictor importance tell a similar story as the CART models. Variables associated with deployment rank the highest for CHAID models. However, unlike CART, CHAID top three predictors varied from one model to another. The first two CHAID models included a Marine's service entry date, separation code, and a reenlistment recommendation. After omitting PEBD and separation code variables, the top three predictors were all reenlistment recommendations. After feature selection, the CHAID models favor PEBD, EAS, and a reenlistment recommendation. Figure 6 shows the top 10 predictors using the dataset with 347 variables and a 75:25 data partition. The top three predictors here are similar to that of the CART model, just in a different order. Here, variable "AWARDSSD" which represents the Sea Service Deployment Ribbon ranks the highest. The next variable is "CON_AVE_" represents a Marines average conduct marks and is understandable why this would be in the top three. However, proficiency marks are not near the top ten. It could be that a Marine's good behavior is an indicator of them reenlisting. The third variable in Figure 6 is "COMBATDA" represents the number of days a Marine has served in combat. Unlike CART, whose top predictors number four through ten are associated with marital status and dependency, CHAID includes performance and award variables for the rest of the top ten variables.

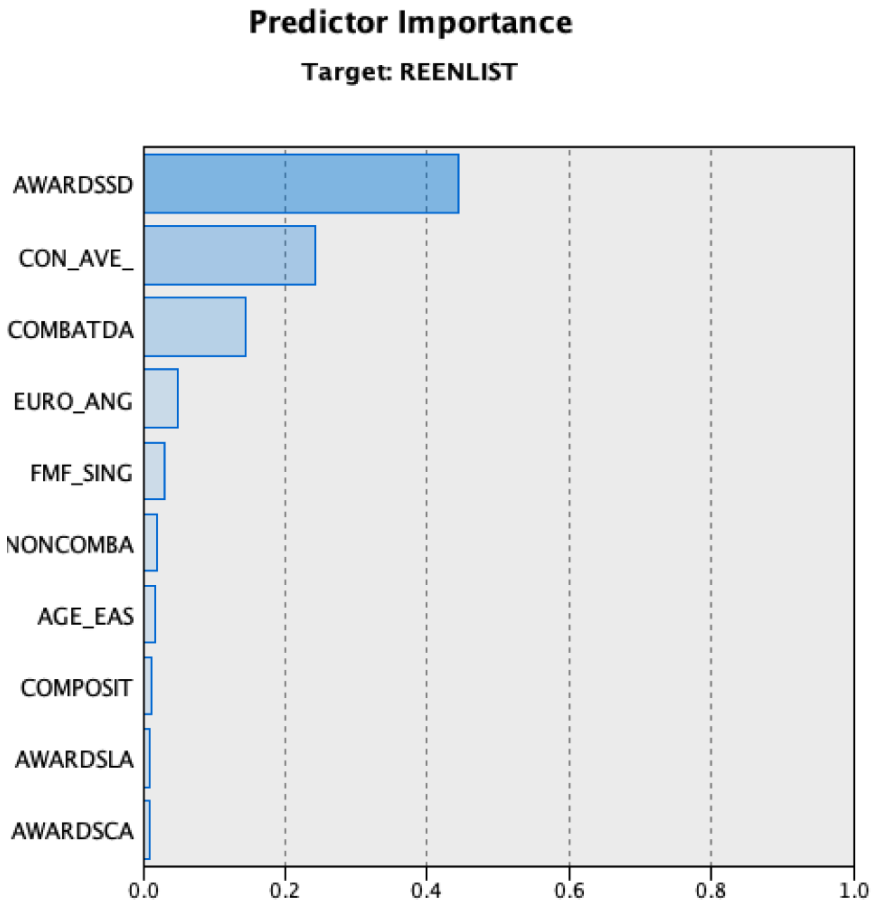

Figure 6. Top 10 Predictors for CHAID Model with a 75:25 Partition on 347-Variable Dataset

3. Linear SVM

Linear SVM models were applied to four of the five datasets. Using a 75:25 data partition ratio, a Linear SVM model performs slightly better with the second, third, and fourth. When using a 50:50 data partition ratio, there is no difference on the performance of the Linear SVM model. Table 6 summarizes the Linear SVM models prediction accuracies.

Table 6. Linear SVM Model Prediction Accuracy

Variable Structure	Linear SVM Prediction Accuracy with 75:25 Data Partition Ratio	Linear SVM Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	75.28%	75.21%
367 (excludes PEBD, TIS, separation code, and EAS)	75.29%	75.21%
348 (excludes reenlistment recommendations)	75.29%	75.21%
347 (excludes good conduct medal)	75.29%	75.21%
64 (only variables with statistical significance > 0.9)	N/A	N/A

Linear SVM models applied produced very different top predictors, except for the top predictor, than those in CART and CHAID models. The first two Linear SVM models with a 75:25 data partition ratio had combat days, composite score, RUC, various marksmanship variables and primary MOS in their top 10 predictors. After omitting variables PEBD, TIS, EAS, and separation code, the result is similar. Both models with different data partitions also had combat days as the number one predictor. Figure 7 shows the top 10 predictors using the dataset with 347 variables and a 75:25 data partition. Again, the top predictor "COMBATDA" represents the number of days a Marine have served in combat. Being deployed to a combat zone can have positive or negative effects on a Marine's career. The deployment might reenergize a Marine's desire to serve their country, and a Marine might choose to reenlist. A Marine might also decide to separate because of the deployment. The next variable "MCC" represents a Marines' unit. A Marine's fist unit can have a positive or negative effect on reenlistment. Therefore, it makes sense that these predictors rank amongst the top. The next four variables in Figure 7 represents a Marine rifle marksmanship scores. Their minimum, maximum, mean, and last score is represented in these four variables. Marksmanship scores contributes to a Marine's composite score that is used for promotions. Junior Marines get promoted based on their composite score, which is a composite of various other metrics, such as physical fitness, marksmanship, proficiency and conduct marks, and TIS. So, the higher the score, the higher the chance of promotion and most likely is a direct correlation to the quality of Marine and their potential for reenlistment.

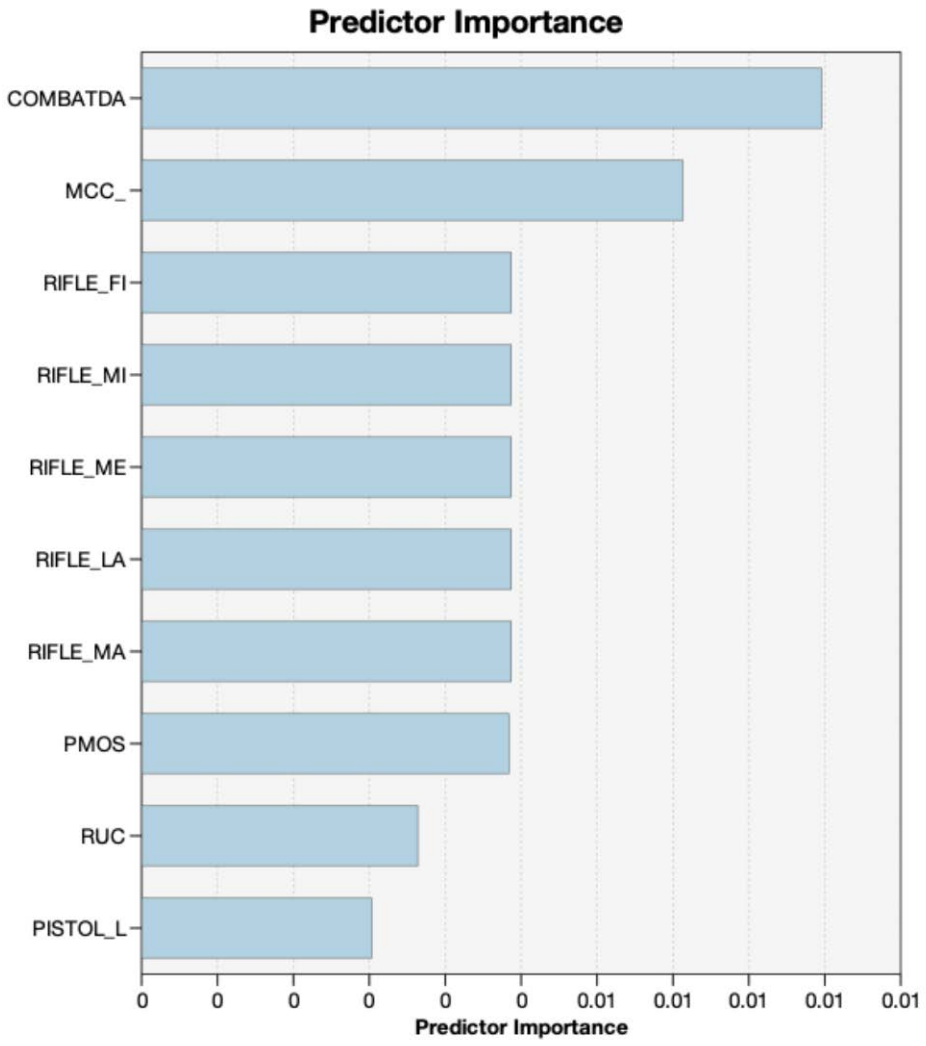

Figure 7. Top 10 Predictors for Linear SVM Model with a 75:25 Partition on 347-Variable Dataset

4. K-means

Being the only unsupervised model and a clustering algorithm, no solid evaluation metric can be used. This thesis applies a silhouette analysis to determine the degree of separation between clusters. The coefficient range for a silhouette ranges from -1 to 1, and we typically want to be as close to 1. The K-means models applied to the datasets resulted in the same average silhouette of 0.2, meaning its poor. All models used five clusters, with the largest clusters in models associated with the dataset with 347 variables. Table 7 summarizes the K-means models prediction accuracies.

Table 7. K-means Model Prediction Accuracy

Variable Structure	K-Means Prediction Accuracy with 75:25 Data Partition Ratio	K-Means Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	N/A	N/A
367 (excludes PEBD, TIS, separation code, and EAS)	Silhouete =0.2 Poor Largest Cluster = 25%	Silhouete =0.2 Poor Largest Cluster = 23.9%
348 (excludes reenlistment recommendations)	Silhouete =0.2 Poor Largest Cluster = 30.7%	Silhouete =0.2 Poor Largest Cluster = 24.4%
347 (excludes good conduct medal)	Silhouete =0.1 Poor Largest Cluster = 30.7%	Silhouete =0.1 Poor Largest Cluster = 26.5%
64 (only variables with statistical significance > 0.9)	Silhouete =0.2 Poor Largest Cluster = 27%	Silhouete =0.2 Poor Largest Cluster = 27.1%

About 22% (79 variables) of the dataset with 367 variables were deemed as important predictors at 100%. These predictors include marital status at recruiting, PMOS, unit, gender, composite score, awards, combat deployments, age at EAS, and service reenlistment bonus (SRB) amounts at a Marines EAS year. When using the smallest dataset post feature selection, over 57% (37 variables) of the dataset were deemed important predictors at 100%. These top predictors are no different than the other k-means models. Figure 8 shows the bottom 14 predictors while using a 75:25 data partition ratio. Even though this is the only unsupervised model, there is little confidence in the model outputs due to the poor average silhouette and the huge number of top predictors.

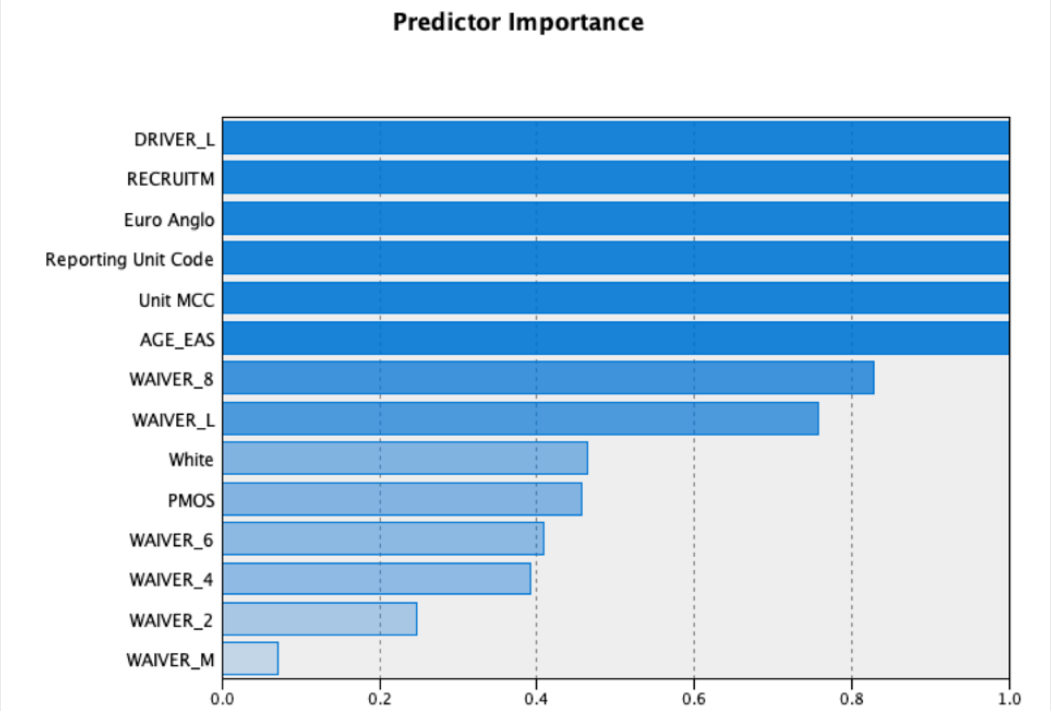

Figure 8. Bottom 14 Predictors for K-means Model with a 75:25 Partition

5. Logistic Regression

Logistic Regression models were only applied to the smallest dataset with 64 variables. When using a 75:25 data partition ratio, logistic regression has a prediction accuracy on 99.37%. Closely, the prediction accuracy with a 50:50 data partition ratio is 99.41%. Table 8 summarizes the Logistic Regression models prediction accuracies.

Table 8. Logistic Regression Model Prediction Accuracy

Variable Structure	Logistic Regression Prediction Accuracy with 75:25 Data Partition Ratio	Logistic Regression Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	N/A	N/A
367 (excludes PEBD, TIS, separation code, and EAS)	N/A	N/A
348 (excludes reenlistment recommendations)	N/A	N/A
347 (excludes good conduct medal)	N/A	N/A
64 (only variables with statistical significance > 0.9)	99.37%	99.41%

Logistic regression models only produce two predictors with an importance greater than 30%. In fact, the top four predictors at the different data partition ratios are identical. Figure 9 shows the top 10 predictors while using a 50:50 data partition ratio. The top predictor "EAS_FY" represents the fiscal years in which a Marine's EAS is. This is typically the year a Marine meets with a career counselor and applies for reenlistment. It is of no surprise that this variable is at the top. The next variable "REENLRE2" represents a reenlistment recommendation code 1A, meaning the Marine is recommended and eligible for reenlistment. This variable was the top predictor for both the CART and CHAID models.

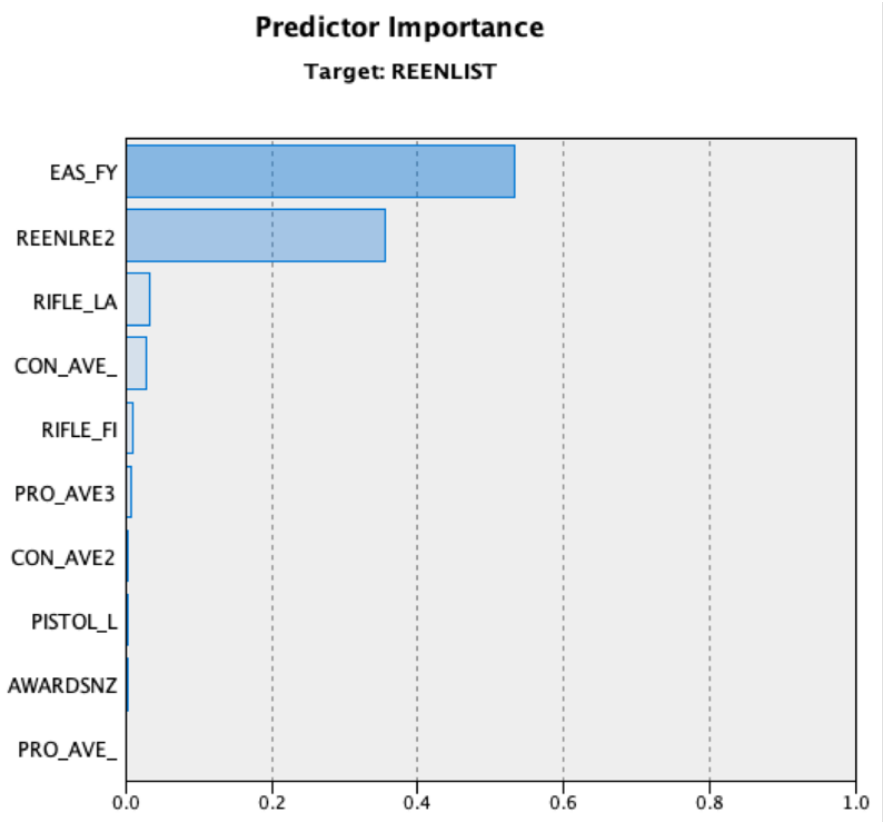

Figure 9. Top 10 Predictors for Logistic Regression with a 50:50 Partition

The next variable "RIFLE_LA" represents the most current rifle qualification score on record. Performance such as this is included into a Marine's composite score that is used for promotion. Lastly, the variable "CON_AVE" represents a Marines average conduct marking in their enlistment. Conduct marks range from 0 to 5, where 0 means a Marines conduct was unacceptable and a 5 means outstanding conduct by the Marine. Both, the rifle scores and conduct marks, contribute to a Marines composite score and is correlated to the quality of a Marine. The higher the quality, the more likely they will be recommended for reenlistment.

6. Bayesian Network

Bayesian Network models was also only able to be fitted to the smallest dataset. At a 75:25 data partition, the model's prediction accuracy is 98.21%. At a 50:50 data partition, the prediction accuracy drops to 97.7% Table 9 summarizes the Bayesian Network models prediction accuracies.

Table 9. Bayesian Network Model Prediction Accuracy

Variable Structure	Bayes Net Prediction Accuracy with 75:25 Data Partition Ratio	Bayes Net Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	N/A	N/A
367 (excludes PEBD, TIS, separation code, and EAS)	N/A	N/A
348 (excludes reenlistment recommendations)	N/A	N/A
347 (excludes good conduct medal)	N/A	N/A
64 (only variables with statistical significance > 0.9)	98.21%	97.70%

The Bayesian Network created arcs to identify correlation between the target variable "reenlist" and the predictors in the dataset. However, the model did not provide any variation in the predictor's importance. According the Bayesian model, no one predictor is important. There is little confidence in the Bayesian Network in predicting reenlistment with the given datasets. Although this model has a high prediction accuracy, it is not useful without seeing the top predictors to assess if indeed those predictors are legitimate and not a collinear with the target variable.

7. C5

A C5 model was fitted to three of the five datasets. At a 75:25 data partition, the model's prediction accuracy is 99.54%. At a 50:50 data partition, the prediction accuracy increases slightly to 99.56%. Table 10 summarizes the C5 models prediction accuracies.

Table 10. C5 Model Prediction Accuracy

Variable Structure	C5 Prediction Accuracy with 75:25 Data Partition Ratio	C5 Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	N/A	N/A
367 (excludes PEBD, TIS, separation code, and EAS)	N/A	N/A
348 (excludes reenlistment recommendations)	92.79%	92.58%
347 (excludes good conduct medal)	89.59%	89.20%
64 (only variables with statistical significance > 0.9)	99.54%	99.56%

Similar to CART and CHAID, the C5 model number of combat days and the Sea Service Deployment Ribbon are in their top three predictors. Figure 10 shows the top 10 predictors using the dataset with 347 variables and a 75:25 data partition ratio. The top three predictors are associated with deploying. Days in combat and the Sea Service Deployment Ribbon have already been discussed in CART model section. The next four top predictors are associated with a Marine's performance. Predictors "PRO_AVE_" and "CON_AVE_" are metrics every junior Marine receives semi-annually. These scores go into the calculation of their composite score, which is represents by the next predictor "COMPOSIT". Similarly, predictor "LEGAL_NJ" represents the number of NJPs a Marine have received. This affects a Marine's chance for promotion and reenlistment.

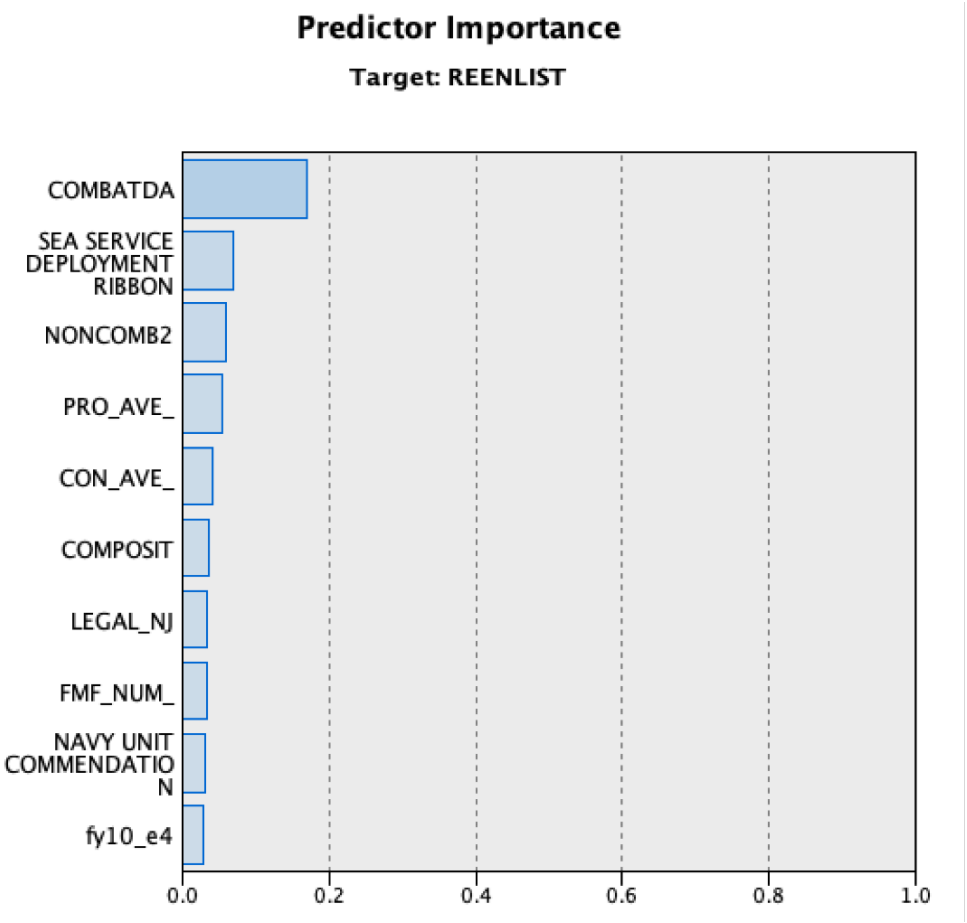

Figure 10. Top 10 Predictors for C5 Model Using a 75:25 Partition on 347-Variable Dataset

8. Random Trees

The Random Tree models were only applied to the down selected dataset with 64 variables. At a 75:25 data partition, the model's prediction accuracy is 99.4%. At a 50:50 data partition, the prediction accuracy decreases slightly to 99.37%. Table 11 summarizes the Random Trees models prediction accuracies.

Table 11. Random Trees Model Prediction Accuracy

Variable Structure	Random Trees Prediction Accuracy with 75:25 Data Partition Ratio	Random Trees Prediction Accuracy with 50:50 Data Partition Ratio
372 (all variables)	N/A	N/A
367 (excludes PEBD, TIS, separation code, and EAS)	N/A	N/A
348 (excludes reenlistment recommendations)	N/A	N/A
347 (excludes good conduct medal)	N/A	N/A
64 (only variables with statistical significance > 0.9)	99.40%	99.37%

There is no difference in the top 10 predictors with regards to the data partition ration used with the Random Trees models. The date a Marine entered the service, age, and unit seemed to have some effects on reenlistment. Figure 11 shows the top 10 predictors using the dataset with 64 variables and a 75:25 data partition ratio. At the top, "PMOS" represent a Marine's primary MOS and is determined early in a Marine's career. Typically, a Marine is classified into a PMOS and will remian with that PMOS for their remainder of their contracts or even careers. The next variable, and the most important predictor is "PEBD_YR", which represents the year a Marines career started. This could be looked at as the first day on the job. Since Marines sign contracts, it is understandable that this will be a top predictor of reenlistment. The next variable is "AGE_EAS" which represents a Marine's age on the year of their EAS. The age of the Marines in the dataset range from 17.4 to 48.8 years. Of the Marines that reenlisted between 2008 and 2018, the majority were between 22 and 23 years of age. The next two variables, "RUC" and "MCC" represents a Marines' unit. A Marine's fist unit can have a positive or negative effect on reenlistment. These top predictors are seen again and again with other models in this thesis.

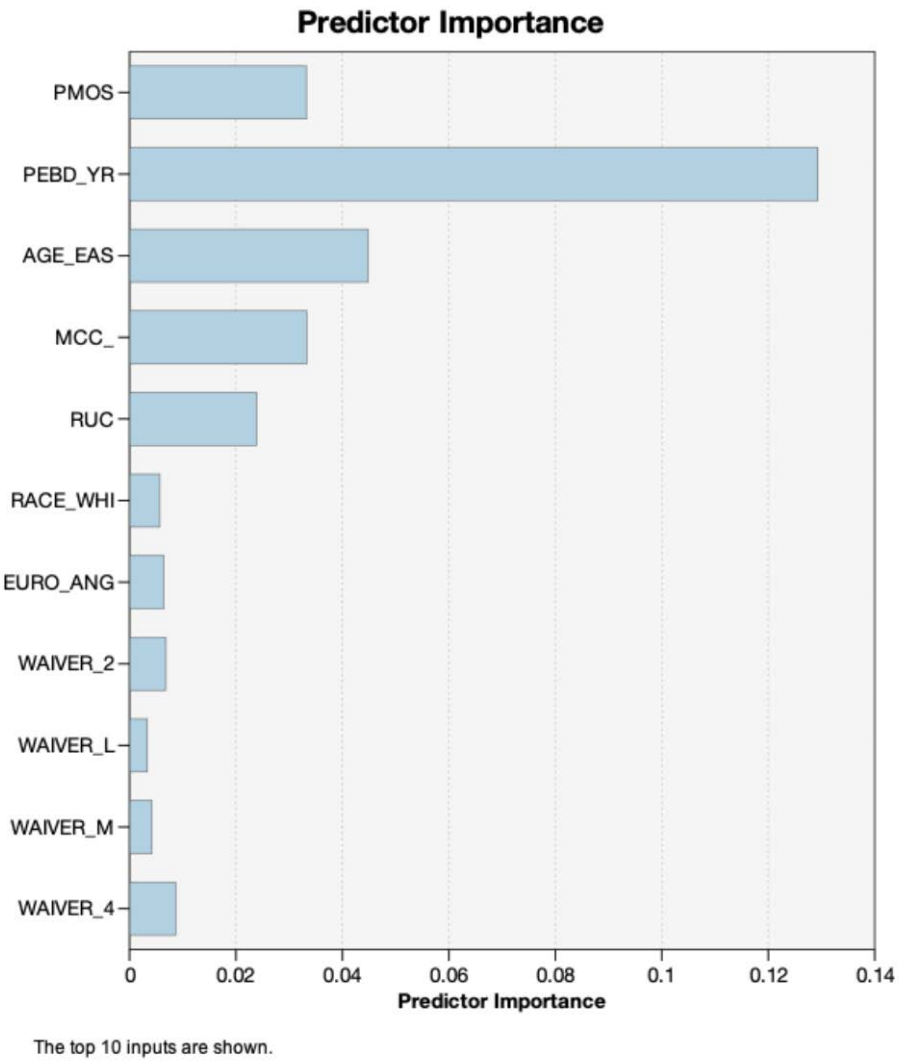

Figure 11. Top 10 Predictors for Random Trees Model Using a 75:25 Partition

C. MODEL COMPARISON

The best model performance with the original dataset is CART with a 50:50 data partition ratio. The CART model's highest prediction accuracy is 98.84%. When using the 367-variable dataset, CHAID with a 75:25 data partition ratio wins. After reducing the dataset to 348 and then 347 variables, the best performing model is C5 with a 75:25 data partition ration. After further reducing the dataset to 64 variables, the best performing

model is C5 with a 50:50 data partition ratio. The C5 model's prediction accuracy is 99.56%.

When comparing models across the different datasets, more models are applied and have higher prediction accuracy as the dataset shrunk to include only statistically significant variables. With a 75:25 data partition ratio, CART performs best under the original dataset and after feature selection with a prediction accuracy of 98.83%. CHAID performs best under the original dataset only with a prediction accuracy of 98.48%. Linear SVM performs best with the 367, 348, and 347-variable datasets at a prediction accuracy of 75.29%. K-means performed poorly with an average silhouette score of 0.2 across the board. C5 performed best under 348-variable dataset with a prediction accuracy of 92.19%. Logistic Regression, Bayes Network, and Random Trees were only applied to the last dataset with only 64 variables. Table 12 shows all the model's prediction accuracies.

Table 12. Combined Model Prediction Accuracy

Dataset	372 (all variables)	367 (excludes PEBD, TIS, separation code, and EAS)	348 (excludes reenlistment recommendations)	347 (excludes Good Conduct Medal)	64 (only variables with statistical significance > 0.9)
Model with 75:25 Data Partition Ratio	Model Prediction Accuracy	Model Prediction Accuracy	Model Prediction Accuracy	Model Prediction Accuracy	Model Prediction Accuracy
CART	98.83%	97.50%	90.25%	82.48%	98.83%
CHAID	98.48%	97.78%	89.12%	82.49%	97.51%
Linear SVM	75.28%	75.29%	75.29%	75.29%	N/A
K-Means	N/A	Silhouete =0.2 Poor. Largest Cluster = 25%	Silhouete =0.2 Poor Largest Cluster = 30.7%	Silhouete =0.1 Poor Largest Cluster = 30.7%	Silhouete =0.2 Poor Largest Cluster = 27%
Logistic Regression	N/A	N/A	N/A	N/A	99.37%
Bayes Net	N/A	N/A	N/A	N/A	98.21%
C5	N/A	N/A	92.79%	89.59%	99.54%
Random Trees	N/A	N/A	N/A	N/A	99.40%
Dataset	372 (all variables)	367 (excludes PEBD, TIS, separation code, and EAS)	348 (excludes reenlistment recommendations)	347 (excludes Good Conduct Medal)	64 (only variables with statistical significance > 0.9)
Model with 50:50 Data Partition Ratio	Model Prediction Accuracy	Model Prediction Accuracy	Model Prediction Accuracy	Model Prediction Accuracy	Model Prediction Accuracy
CART	98.84%	96.96%	90.19%	82.44%	98.84%
CHAID	98.46%	97.74%	89.06%	82.46%	97.51%
Linear SVM	75.21%	75.21%	75.21%	75.21%	N/A
K-Means	N/A	Silhouete =0.2 Poor Largest Cluster = 23.9%	Silhouete =0.2 Poor Largest Cluster = 24.4%	Silhouete =0.1 Poor Largest Cluster = 26.5%	Silhouete =0.2 Poor Largest Cluster = 27.1%
Logistic Regression	N/A	N/A	N/A	N/A	99.41%
Bayes Net	N/A	N/A	N/A	N/A	97.70%
C5	N/A	N/A	92.58%	89.20%	99.56%
Random Trees	N/A	N/A	N/A	N/A	99.37%

When using a 50:50 data partition ratio, CART continues to perform best under the original dataset and after feature selection with a prediction accuracy of 98.84%. CHAID performs best under the original dataset only with a prediction accuracy of 98.46%. Linear SVM performs the same with all datasets at a prediction accuracy of 75.21%. K-means performed poorly with an average silhouette score of 0.2 across the board. C5 performed best under 348-variable dataset with a prediction accuracy of 92.58%. Logistic Regression, Bayes Network, and Random Trees were only applied to the last dataset with only 64 variables. Table 12 shows all the model's prediction accuracies.

When comparing model prediction accuracy performance across different data partition ratios, models with a 50:50 partition ratio have higher prediction accuracies. CART model accuracy variation across the different data partitions is between 98.83 and 98.84%; CHAID model accuracy variation is between 98.46 and 98.48%; and Linear SVM model accuracy variation is between 75.21 and 75.28%.

After further filtering the dataset to 367 variables, the CART model's accuracy variation is between 96.96 and 97.5%; CHAID model accuracy variation is between 97.74 and 97.78%; Linear SVM model accuracy variation is between 75.21 and 75.29 percent; and there was no variation in the K-means average silhouette.

After reducing the variables to 347, the CART model's accuracy variation is between 82.44 and 82.48%; CHAID model accuracy variation is between 82.46 and 82.49%; Linear SVM model accuracy variation is between 75.21 and 75.29%; and there was no variation in the k-means model average silhouette.

After conducting a feature selection and reducing the dataset to 64 variables, the CART model's accuracy variation was between 98.83 and 98.84%; CHAID model had no variations is accuracy; K-means had no accuracy variation; Logistic Regression model accuracy variation is between 99.37 and 99.41%; Bayes Network model accuracy variation is between 97.7 and 98.21%; C5 model accuracy variation is between 99.54 and 99.56%; and Random Trees model accuracy variation is between 99.37 and 99.4%.

D. RESULTS

Given the down-selection process of the data and model accuracy comparisons, the preferred model is C5. Although the prediction accuracies of C5 models were lower than other models, the important predictors in Figure 12 are logical and intuitively make sense that they would be at the top of the list. Additionally, the 347-variable dataset accounts for collinearity among the top predictors present in the models with 373, 367, and 348 variables.

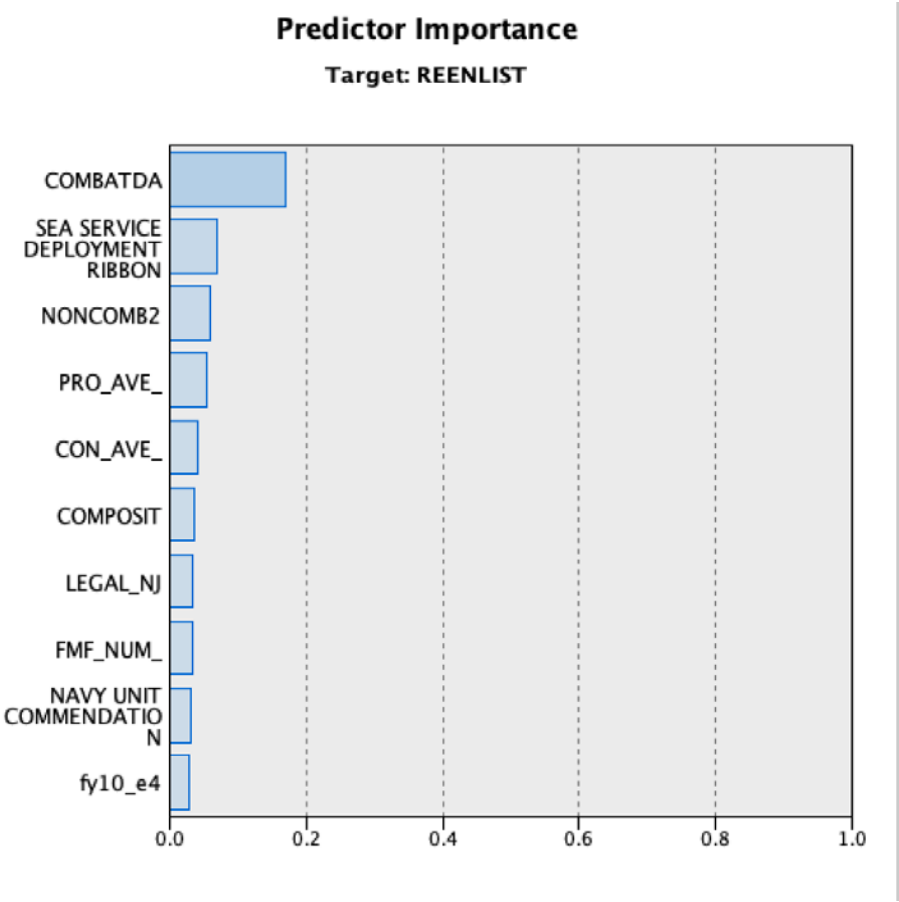

Figure 12. C5 Model Top 10 Predictors with 347-Variable Dataset and a 75:25 Data Partition Ratio

A manual feature reduction proved to be useful to discern from unhelpful top predictors that were collinear to the target variable of reenlistment. Table 13 displays the top four models under a reduced dataset to 347 variables that excludes variables such as PEBS, TIS, separation code, reenlistment recommendation, and the good conduct award.

Table 13. Top 4 Models

Model	Accuracy	Dataset	Data Partition Ratio
C5	89.59%	347 (excludes Good Conduct Medal)	75:25
CHAID	82.49%	347 (excludes Good Conduct Medal)	75:25
CART	82.48%	347 (excludes Good Conduct Medal)	75:25
Linear SVM	75.29%	347 (excludes Good Conduct Medal)	75:25

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY, CONCLUSION AND RECOMMENDATIONS

A. SUMMARY

Given the data used in this thesis, the best machine learning algorithm that works best at predicting the probability of reenlistment is the C5 algorithm. The C5 algorithm best prediction accuracy was 89.59%. The C5 algorithm also performed best with the 347-variable dataset and under a 75:25 data partition ratio. Looking at prediction accuracy is not enough. Therefore, a combination of other evaluation factors must be included, such as looking at the predictors and evaluate for collinearity. Among all the datasets, reducing the dataset to 347-variables was useful to know that machine learning can handle larger datasets, and not resort to the low 64-variable dataset. However, adding more data to the machine learning algorithm was not always useful.

Looking at an algorithm's prediction accuracy is not enough. There are many algorithms that provide high prediction accuracies, but this could be a result of overfitting. When using the original 372-variable dataset, the CART algorithm was able to produce a high 98.84% model prediction accuracy while using a 50:50 data partition ration. This seems to be good mathematically, but the top three predictors for CART with the 372-variable dataset were PEBD, EAS, and a positive reenlistment recommendation. Variables PEBD and EAS are associated with a Marine's separation from the service, and reenlistment recommendation is collinear with the target variable of reenlist. After removing these variables and reducing he dataset to 347 variables, the CART model prediction accuracy decreases to 82.44%. The predictors of importance, after reducing the data to 347 variables, are interesting and require further investigation as to why variables associated with deployment are in the top three.

There is no one algorithm that best identify the predictors of reenlistment. Removing variables associated with separation and reenlistment recommendations yields more useful top predictors of reenlistment. However, the top predictors vary among the algorithms applied in this thesis. The C5 algorithm's top predictors are associated with deployments and a proficiency and conduct marks.

B. CONCLUSION

This thesis demonstrates the usefulness of machine learning at supporting the Marine Corps' effort on talent management. Through this thesis, we learn several things about predicting reenlistment by using machine learning.

First, we learn that machine learning can use much more variables that those in traditional statistical models. Unlike traditional statistical modeling, a predictive machine learning model is generated via a computerized algorithm. Meaning that this allows us to add more variables than the standard convention and allow the machine to tell us what the optimal model in given a dataset. The algorithms used in this thesis were able to handle a 372-variable dataset and rapidly generated models and top predictors for each model. Due to the speed, we were able to look at the top predictors and test for collinearity between them and the dependent variable.

Second, we learn that machine learning typically does a better job at recognizing unexpected patterns in the data. Unlike traditional statistical modeling, where the analyst selects the variables that go in a model, machine learning can apply multiple algorithms to all the data and provide top predictors for an analyst to investigate further. A method used in this thesis was feature selection, where machine learning was applied to the data and reduced it from 373 to 64 statistically significant variables. In this 64-variable dataset, combat days and number of non-combat deployments ranked among the top 10 predictors.

Lastly, there are various types of algorithms that could be applied at a manpower problem. This thesis aimed to answer the question of the probability of a Marine reenlisting. This is a classification type of problem that machine can handle with various types of algorithms. We can start by using decision trees like CART, CHAID, or even a C5 algorithm to get at predicting the probability of reenlistment, and by extension separation.

There are several unexpected variables that leadership should focus on. With the data used in this thesis, the best preforming algorithm, in terms of accuracy and predictor importance, is the C5 algorithm. The C5 algorithm uncovered unexpected predictors, specifically number of combat days, the sea service deployment ribbon, and the number of

non-combat deployments. After removing obvious predictors of reenlistment, such as PEBD, TIS, reenlistment recommendations, and some not so obvious like the Good Conduct Medal, the top predictors across various algorithms circled around deployments. Leadership should focus on the effects of combat and non-combat deployments with reenlistment. With deployment opportunities decreasing, if this is a predictor of reenlistment, then the Marine Corps should consider alternatives to retain its required force.

C. FUTURE RESEARCH AND RECOMMENDATIONS

Future research could include enlisted data on TAPAS, exit survey data, and medical data. TAPAS data can uncover insights into how a Marine's personality and their experience in the Marine Corps influence the decision of reenlistment. Exit surveys are useful in getting insights as to why Marines decide to not reenlist. This survey data should be made accessible via the TFDW for easier access. Medical data was not used in this thesis, but data on injuries and limited duty status could provide a more wholistic view of what injuries are correlated with a separation.

It is recommended that a custom algorithm us used to be able to handle the ample amount of stored data in the TFDW and other Marine Corps database to best capture the best predictors.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX. DATA SUMMARY

Variable	Obs	Mean	Std. Dev.	Min
Unique Identifier	404395	213580	124249	1
reenlist	404395	.247461	.431537	0
Birth Year	404395	90.3651	8.59943	0
Gender	404395	.07915	.269974	0
Unit MCC	404395	553.118	480.704	0
Reporting Unit Code	404395	16242.9	14195.2	0
Contract Extention Length in Months	404395	.370563	2.01853	0
Time in Service in Years	404395	3.24256	1.39846	0
Marital Status at Recruiting_Annulled	404395	.000334	.018268	0
Marital Status at Recruiting_Divorced	404395	.001763	.041953	0
Marital Status at Recruiting_Interlock	404395	2.5e-06	.001573	0
Marital Status at Recruiting_Legal Separated	404395	.000062	.007862	0
Marital Status at Recruiting_Married	404395	.021662	.145577	0
Marital Status at Recruiting_Single	404395	.97592	.153299	0
Marital Status at Recruiting_Widowed	404395	.000057	.007541	0
Marital Status at Recruiting Unknown	404395	.0002	.014151	0
Dependents at Recruiting	404395	.016845	.12869	0
No Dependents at Recruiting	404395	.983155	.12869	0
One Dependent at Recruiting	404395	.011185	.105164	0
Two Dependents at Recruiting	404395	.0048	.069114	0
Three Dependents at Recruiting	404395	.000225	.014999	0
Four Dependents at Recruiting	404395	.000534	.023105	0
Five Dependents at Recruiting	404395	9.9e-06	.003145	0
Six Dependents at Recruiting	404395	.000069	.008321	0
Seven Dependents at Recruiting	404395	2.5e-06	.001573	0
Eight Dependents at Recruiting	404395	9.9e-06	.003145	0
Nine Dependents at Recruiting	404395	2.5e-06	.001573	0
Ten Dependents at Recruiting	404395	4.9e-06	.002224	0
Twelve Dependents at Recruiting	404395	2.5e-06	.001573	0
Indian or Alaskan Native	404395	.011922	.108533	0
Asian	404395	.026326	.160102	0
Black	404395	.094934	.293125	0
Hawaiian or Pacific Islander	404395	.010772	.103226	0
White	404395	.823742	.38104	0
No Response for Race	404395	.032305	.176809	0
No Reponse for Ethnicity	404395	.374468	.483986	0
Other Hispanic	404395	.030416	.171729	0
US or Canadian Indian	404395	.006227	.078663	0
Other Asian	404395	.006313	.079204	0
Puerto Rican	404395	.006133	.078071	0
Filipino	404395	.00613	.078055	0
Mexican	404395	.082976	.275846	0
Alaskan Native	404395	.001123	.033487	0
Cuban	404395	.001588	.039813	0
African	404395	.055658	.229261	0
Caribbean	404395	.002995	.054641	0
Indian	404395	.002826	.053089	0

Melanes	404395	.000225	.014999	0
Australasian	404395	.000309	.017579	0
Chinese	404395	.002339	.04831	0
Guamanian	404395	.000591	.024303	0
Japanese	404395	.000767	.027677	0
Korean	404395	.001949	.0441	0
Polynes	404395	.001113	.03334	0
Euro Anglo	404395	.392015	.488201	0
Other Pacific Islander	404395	.002468	.049617	0
Latin American	404395	.017579	.131417	0
Arabian	404395	.000912	.030193	0
Vietnamese	404395	.002606	.050986	0
Micrones	404395	.000274	.016565	0
US Citizen, abroad	404395	.011954	.108677	0
US Citizen, native	404395	.926995	.260146	0
US Citizen, derivative	404395	.00208	.045556	0
US Citizen, Naturalized	404395	.016481	.127318	0
Declined to answer	404395	.000072	.008468	0
Hispanic	404395	.002248	.047358	0
Immigrant Alien	404395	.02687	.161703	0
Foreign National	404395	.000052	.007206	0
Not Hispanic	404395	.010022	.099609	0
US, non-citizen	404395	.000727	.026953	0
asvab_meps_albany	404395	.013546	.115597	0
asvab_meps_albuquerque	404395	.005349	.072939	0
asvab_meps_amarillo	404395	.005277	.072451	0
asvab_meps_anchorage	404395	.002475	.049691	0
asvab_meps_atlanta	404395	.024147	.153506	0
asvab_meps_baltimore	404395	.028737	.167066	0
asvab_meps_beckley	404395	.007809	.088024	0
asvab_meps_boise	404395	.004679	.06824	0
asvab_meps_boston	404395	.01937	.137821	0
asvab_meps_brooklyn	404395	.036086	.186504	0
asvab_meps_buffalo	404395	.009723	.098126	0
asvab_meps_butte	404395	.004454	.066586	0
asvab_meps_charlotte	404395	.01816	.133532	0
asvab_meps_chicago	404395	.025898	.158831	0
asvab_meps_cleveland	404395	.018744	.13562	0
asvab_meps_jaxfla_closed	404395	4.9e-06	.002224	0
asvab_meps_columbus	404395	.012871	.112718	0
asvab_meps_dallas	404395	.031912	.175766	0
asvab_meps_denver	404395	.020703	.142387	0
asvab_meps_des_moines	404395	.011071	.104634	0
asvab_meps_detroit	404395	.020715	.142428	0
asvab_meps_el_paso	404395	.005	.070534	0
asvab_meps_fargo	404395	.004226	.064871	0
asvab_meps_fort_dix	404395	.01733	.130496	0
asvab_meps_ft_jackson	404395	.012317	.110297	0
asvab_meps_honolulu	404395	.006138	.078102	0

asvab_meps_houston	404395	.02774	.164228	0
asvab_meps_i	404395	.000049	.007032	0
asvab_meps_ii	404395	.000524	.02289	0
asvab_meps_iiia	404395	.000485	.02201	0
asvab_meps_iiib	404395	.000361	.018997	0
asvab_meps_indianapolis	404395	.019595	.138603	0
asvab_meps_jackson	404395	.00364	.060223	0
asvab_meps_jacksonville	404395	.020062	.140213	0
asvab_meps_kansas_city	404395	.020871	.142952	0
asvab_meps_knoxville	404395	.007651	.087134	0
asvab_meps_lansing	404395	.018069	.133201	0
asvab_meps_las_vegas	404395	7.4e-06	.002724	0
asvab_meps_little_rock	404395	.005574	.074449	0
asvab_meps_los_angeles	404395	.047879	.21351	0
asvab_meps_louisville	404395	.016914	.12895	0
asvab_meps_mechanicsburg	404395	.020275	.140939	0
asvab_meps_memphis	404395	.006358	.079481	0
asvab_meps_miami	404395	.021694	.145683	0
asvab_meps_milwaukee	404395	.019859	.139517	0
asvab_meps_minneapolis	404395	.013727	.116354	0
asvab_meps_montgomery	404395	.016837	.128662	0
asvab_meps_nashville	404395	.011944	.108633	0
asvab_meps_new_orleans	404395	.010786	.103296	0
asvab_meps_oklahoma_city	404395	.01636	.126857	0
asvab_meps_omaha	404395	.006081	.077741	0
asvab_meps_phoenix	404395	.022626	.148709	0
asvab_meps_pittsburgh	404395	.010487	.101869	0
asvab_meps_portland	404395	.011422	.106262	0
asvab_meps_portland_me	404395	.004196	.064644	0
asvab_meps_portland_or	404395	.008581	.092234	0
asvab_meps_raleigh	404395	.014365	.118989	0
asvab_meps_richmond	404395	.017169	.1299	0
asvab_meps_sacramento	404395	.027854	.164554	0
asvab_meps_salt_lake_city	404395	.009592	.097469	0
asvab_meps_san_antonio	404395	.024867	.155719	0
asvab_meps_san_diego	404395	.027194	.162648	0
asvab_meps_san_jose	404395	.021964	.146565	0
asvab_meps_san_juan	404395	.003069	.055312	0
asvab_meps_seattle	404395	.015067	.121819	0
asvab_meps_shreveport	404395	.003494	.059008	0
asvab_meps_sioux_falls	404395	.004904	.069854	0
asvab_meps_spokane	404395	.006934	.08298	0
asvab_meps_springfield	404395	.014189	.11827	0
asvab_meps_st_louis	404395	.020018	.14006	0
asvab_meps_syracuse	404395	.007053	.083683	0
asvab_meps_tampa	404395	.02483	.155606	0
asvab_meps_unknown	404395	.000022	.004718	0
AFQT_SCORE	404395	61.6651	18.2269	0
DEFENSE_LANG_APT_BATTERY_SCORE	404395	5.19392	21.7451	0

DEPOT_COMP_GENERAL_TECH	404395	108.322	12.3147	0
DEPOT_COMP_MECH_MAINT	404395	106.427	14.1897	0
DEPOT_COMPOSITE_CLERICAL	404395	103.847	24.185	0
DEPOT_COMPOSITE_ELECTRONICS	404395	108.484	12.4423	0
waiver_age	404395	.000885	.030317	0
waiver_foreign	404395	.000049	.007032	0
waiver_dependency	404395	.026106	.169143	0
waiver_drug	404395	.337131	.51403	0
waiver_law	404395	.13195	.408745	0
waiver_medical	404395	.138246	.376367	0
waiver_mental	404395	.0002	.014325	0
waiver_educ	404395	.000581	.024202	0
waiver_priorservice	404395	.00321	.058244	0
waiver_admin	404395	.121354	.353827	0
waiver_level_district	404395	.110446	.350976	0
waiver_level_mcrc	404395	.149985	.390512	0
waiver_level_region	404395	.031605	.184201	0
waiver_level_rs	404395	.467676	.664515	0
recruitment_bonus	404395	.139218	.346174	0
recrtmnt_bonus_prgm	56299	55.5794	32.0373	1
DEP Extention Days	404395	.191249	3.46077	0
DLAB_SCORE	404395	4.14686	19.5108	0
driver_license	404395	.654034	.475683	0
Education Tier 1	404395	.987008	.113241	0
Education Tier 2	404395	.010502	.10194	0
Education Tier 3	404395	.000885	.02974	0
Education Tier Blank	404395	.001605	.040029	0
HEIGHT_AT_SHIP	404395	68.9477	3.04318	7
height_req_at_ship	404395	.999958	.006484	0
WEIGHT_AT_SHIP	404395	162.586	25.6546	0
weight_req_at_ship	404395	.999837	.012774	0
mental_psych0	404395	.000524	.02289	0
mental_psych1	404395	.989184	.103437	0
mental_psych2	404395	4.9e-06	.002224	0
mental_psych3	404395	.009318	.096077	0
mental_psych_unknown	404395	.000969	.031119	0
SAT_SCORE	404395	2.54496	54.4698	0
fmf_citizenship_unknown	404395	.000294	.017152	0
fmf_citizenship_usnat	404395	.002127	.046066	0
fmf_citizenship_us	404395	.940343	.236851	0
fmf_citizenship_dernat	404395	.001763	.041953	0
fmf_citizenship_derus	404395	.009901	.099011	0
fmf_citizenship_natlzd	404395	.029627	.169556	0
fmf_citizenship_resident	404395	.00134	.036585	0
fmf_citizenship_alien	404395	.014605	.119964	0
PMOS	404395	2537.54	2377.59	0
fmf_civ_educ_level0	404395	.004261	.065135	0
fmf_civ_educ_level1	404395	.000012	.003516	0
fmf_civ_educ_level7	404395	.000059	.007704	0

fmf_civ_educ_level8	404395	.000294	.017152	0
fmf_civ_educ_level9	404395	.000715	.026723	0
fmf_civ_educ_level10	404395	.003764	.061233	0
fmf_civ_educ_level11	404395	.00545	.073624	0
fmf_civ_educ_level12	404395	.936312	.244196	0
fmf_civ_educ_levelcoll_1	404395	.018326	.134128	0
fmf_civ_educ_levelcoll_2	404395	.017285	.130332	0
fmf_civ_educ_levelcoll_3	404395	.002703	.051918	0
fmf_civ_educ_levelcoll_4	404395	.010455	.101715	0
fmf_civ_educ_level_masters	404395	.000336	.018336	0
fmf_civ_educ_level_postmaster	404395	.00002	.004448	0
fmf_civ_educ_level_doctorate	404395	7.4e-06	.002724	0
fmf_civ_educ_cert_lessHS	404395	.000443	.021034	0
fmf_civ_educ_cert_nonTDhs	404395	.000277	.01664	0
fmf_civ_educ_cert_corspDip	404395	.001506	.038777	0
fmf_civ_educ_cert_1semColl	404395	.008957	.094215	0
fmf_civ_educ_cert_adultDip	404395	.012906	.112868	0
fmf_civ_educ_cert_OccCert	404395	.000129	.011339	0
fmf_civ_educ_cert_AssocDeg	404395	.009609	.097556	0
fmf_civ_educ_cert_GED	404395	.006548	.080655	0
fmf_civ_educ_cert_examFail	404395	.000371	.019256	0
fmf_civ_educ_cert_nurse	404395	.000015	.003852	0
fmf_civ_educ_cert_homeStudy	404395	.007357	.085455	0
fmf_civ_educ_cert_certATT	404395	.001118	.033414	0
fmf_civ_educ_cert_bachelors	404395	.00959	.097456	0
fmf_civ_educ_cert_HSdiploma	404395	.934913	.24668	0
fmf_civ_educ_cert_nearComp	404395	.000569	.023842	0
fmf_civ_educ_cert_masters	404395	.000334	.018268	0
fmf_civ_educ_cert_postMaster	404395	2.5e-06	.001573	0
fmf_civ_educ_cert_HSsenior	404395	.000616	.024806	0
fmf_civ_educ_cert_doctorate	404395	7.4e-06	.002724	0
fmf_civ_educ_cert_1stProf	404395	.000015	.003852	0
fmf_civ_educ_cert_GEDnatgd	404395	.000401	.020011	0
fmf_civ_educ_cert_unknown	404395	.000032	.00567	0
fmf_civ_ed_subj	404395	91.0067	10.4593	0
fmf_civ_educ_grad	404395	.985215	.120692	0
fmf_composite_score	404395	529.485	746.629	0
fmf_conduct_ave_enlistment	404395	36.5167	15.4557	0
fmf_conduct_ave_ingrade	404395	30.6498	19.6491	0
fmf_conduct_ave_service	404395	42.7923	4.31485	0
fmf_proficiency_ave_enlistment	404395	36.737	15.4851	0

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66. https://link.springer.com/article/10.1007/BF00153759

Alpaydin, E. (2014). *Introduction to machine learning*. MIT Press.

Arkes, J. (2019). *Regression Analysis: A Practical Introduction*. Routledge.

Barry, J.C., & Gilikin, P.L. (2005). *Comparative analysis of Navy and Marine Corps planning, programming, budgeting and execution systems from a manpower perspective* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/2322

Brownlee, J. (2019). *A tour of machine learning algorithms*. Machine Learning Mastery. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Cole, A. L. (2014). U.S. *Marine Corps enlisted retention: an analysis of stakeholder incentives for the retention of tier 1 first-term Marines* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/41360

Conatser, D. G. (2006). *Forecasting U.S. Marine Corps reenlistments by military occupational specialty and grade* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/2543

Fletcher, C. K. (2018). *Modeling first-term enlistment completion for the Selected Marine Corps Reserve* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/59658

Friedman J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* 2nd Edition. Springer.

Hattiangadi, A. U., Kimble, T. H., Lambert, W. B., & Quester, A. O. (2005). *Endstrength: Forecasting Marine Corps losses* (No. CRM-D0011188. A2). Center for Naval Analyses. https://apps.dtic.mil/dtic/tr/fulltext/u2/a434986.pdf

Pechacek, J., Gelder, A., Roberts, C., King, J., Bishop, J., Guggisberg, M., & Kirpichevsky, Y. (2019). *A new military Retention Prediction Model: Machine learning for high-fidelity forecasting*. Institute for Defense Analyses. https://www.ida.org/-/media/feature/publications/a/an/a-new-military-retention-prediction-model_machine-learning-for-high-fidelity-forecasting/d-10712.ashx

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications (IJACSA), 9*(2). https://pdfs.semanticscholar.org/2a9f/505e1ab148aa3d91810f509ee133272be554.pdf

Scarfe, J.C. (2016). *An analysis of departure behaviors of high-quality career designated first-term marine officer* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/48590

U.S. Marine Corps. (2003). *Sea service deployment ribbon criteria* (MARADMIN 582/03). https://www.marines.mil/News/Messages/Messages-Display/Article/891212/sea-service-deployment-ribbon-criteria/

U.S. Marine Corps. (2006a). *FY 2007 selective reenlistment bonus program* (MARADMIN 334/06). HTTPS://WWW.MARINES.MIL/NEWS/MESSAGES/MESSAGES-DISPLAY/ARTICLE/890510/MCBUL-7220-FISCAL-YEAR-2007-FY07-SELECTIVE-REENLISTMENT-BONUS-SRB-PROGRAM/

U.S. Marine Corps. (2006b). *MCBUL 7220. Fiscal year 2007 selective reenlistment bonus program* (MARADMIN 334/06). https://www.marines.mil/News/Messages/Messages-Display/Article/890510/mcbul-7220-fiscal-year-2007-fy07-selective-reenlistment-bonus-srb-program/

U.S. Marine Corps. (2007). *MCBUL 7220. Fiscal year 2008 selective reenlistment bonus program and FY08 broken service SRB program* (MARADMIN 349/07). https://www.marines.mil/News/Messages/Messages-Display/Article/893811/mcbul-7220-fiscal-year-2008-fy08-selective-reenlistment-bonus-srb-program-and-f/

U.S. Marine Corps. (2008). *MCBUL 7220. Fiscal year 2009 selective reenlistment bonus program and FY09 broken service SRB program* (MARADMIN 370/08). https://www.marines.mil/News/Messages/Messages-Display/Article/890146/mcbul-7220-fiscal-year-2009-fy09-selective-reenlistment-bonus-srb-program-and-f/

U.S. Marine Corps. (2009). *MCBUL 7220. Fiscal year 2010 selective reenlistment bonus program and FY10 broken service SRB program* (MARADMIN 0378/09). https://www.marines.mil/News/Messages/Messages-Display/Article/889440/mcbul-7220-fiscal-year-2010-fy10-selective-reenlistment-bonus-srb-program-and-f/

U.S. Marine Corps. (2010). *MCBUL 7220. Fiscal year 2011 selective reenlistment bonus program and FY11 broken service SRB program* (MARADMIN 341/10). https://www.marines.mil/News/Messages/Messages-Display/Article/888807/mcbul-7220-fiscal-year-2011-fy11-selective-reenlistment-bonus-srb-program-and-f/

U.S. Marine Corps. (2011). *MCBUL 7220. Fiscal year 2012 selective reenlistment bonus program and FY12 broken service SRB program* (MARADMIN 348/11). https://www.marines.mil/News/Messages/Messages-Display/Article/888119/mcbul-7220-fiscal-year-2012-fy12-selective-reenlistment-bonus-srb-program-and-f/

U.S. Marine Corps. (2012). *Precedence levels for manning and staffing*. (MCO 5320.12H). https://www.marines.mil/Portals/1/Publications/MCO%205320.12H.pdf

U.S. Marine Corps. (2015). *Total force structure process*. (MCO 5311.1E). https://www.marines.mil/portals/1/MCO%205311.1E%20z.pdf

U.S. Marine Corps. (2019a). *Commandant's Planning Guidance: 38th Command of the Marine Corps*. https://www.hqmc.marines.mil/Portals/142/Docs/%2038th%20Commandant%27s%20Planning%20Guidance_2019.pdf?ver=2019-07-16-200152-700

U.S. Marine Corps. (2019b). *FY 2020 enlisted retention campaign* (MARADMIN 277/19). https://www.marines.mil/News/Messages/MARADMINS/Article/1842655/fy-2020-enlisted-retention-campaign/

U.S. Senate Committee on Armed Services. (2009). *Conference report for the National Defense Authorization Bill for fiscal year 2010*. https://www.armed-services.senate.gov/imo/media/doc/NDAA-FY10-Conference-Press-Release1.pdf

U.S. Senate Committee on Armed Services. (2015). *National Defense Authorization Bill for fiscal year 2010* (ACT S.1356). https://docs.house.gov/billsthisweek/20151102/s1356_sus_xml.pdf

Ugurbas, U. & Kormaz, M. (2015). *Determinants of first-term attrition for enlisted and officer selected Marine Corps reservists* [Master's thesis, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/45267

Zheng, A. (2015). *Evaluating machine learning models*. O'Reilly.

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