2024-03

AN AUTOMATED MACHINE LEARNING APPROACH FOR MORE EFFICIENT MARINE CORPS RECRUITER PROSPECTING

Born, Andrew A.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AN AUTOMATED MACHINE LEARNING APPROACH FOR MORE EFFICIENT MARINE CORPS RECRUITER PROSPECTING

by

Andrew A. Born

March 2024

Thesis Advisor: Maxim Massenkoff
Second Reader: Sae Young Ahn

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2024	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE AN AUTOMATED MACHINE LEARNING APPROACH FOR MORE EFFICIENT MARINE CORPS RECRUITER PROSPECTING	**5. FUNDING NUMBERS**
6. AUTHOR(S) Andrew A. Born	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The military recruiting environment is facing significant challenges, making recruitment goals more difficult to obtain. Due to these difficulties, the Marine Corps must find new ways to target the right demographics effectively. This thesis serves as a proof of concept for recruiting: can we employ automated machine learning to accurately prioritize public high schools using publicly available data? Current methods by Marine Corps Recruiting Command to prioritize high schools are largely unsystematic, potentially leading to inefficient allocation of recruiting resources. This study employs Microsoft Azure to demonstrate how we can use automated machine learning to enhance the efficiency of recruiting efforts.

I find that automated machine learning using publicly available data may be an effective tool for predicting which public high schools to prioritize. Additionally, the automated machine learning predictions produced more contracts than the Marine Corps' choices of priority schools. I recommend that the Marine Corps and other branches of service further explore the use of automated machine learning and open-source data to enhance their recruitment strategies. Additionally, the key predictive variables identified by the automated machine learning model align closely with the criteria used by Recruiting Station leaders. However, the model provides a more granular analysis, enabling the identification of subtle patterns and interactions between each variable.

14. SUBJECT TERMS recruiting, Marine Corps, AutoML, machine learning, open source, enlisted, recruiting, talent management, manpower, labor economics, behavioral economics, public education, end strength, MCRC	**15. NUMBER OF PAGES** 57
	16. PRICE CODE

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

NSN 7540–01-280-5500

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

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

AN AUTOMATED MACHINE LEARNING APPROACH FOR MORE EFFICIENT MARINE CORPS RECRUITER PROSPECTING

Andrew A. Born
Captain, United States Marine Corps
BA, Davidson College, 2018

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 2024

Approved by: Maxim Massenkoff
Advisor

Sae Young Ahn
Second Reader

Marigee Bacolod
Academic Associate, Department of Defense Management

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The military recruiting environment is facing significant challenges, making recruitment goals more difficult to obtain. Due to these difficulties, the Marine Corps must find new ways to target the right demographics effectively. This thesis serves as a proof of concept for recruiting: can we employ automated machine learning to accurately prioritize public high schools using publicly available data? Current methods by Marine Corps Recruiting Command to prioritize high schools are largely unsystematic, potentially leading to inefficient allocation of recruiting resources. This study employs Microsoft Azure to demonstrate how we can use automated machine learning to enhance the efficiency of recruiting efforts.

I find that automated machine learning using publicly available data may be an effective tool for predicting which public high schools to prioritize. Additionally, the automated machine learning predictions produced more contracts than the Marine Corps' choices of priority schools. I recommend that the Marine Corps and other service branches further explore the use of automated machine learning and open-source data to enhance their recruitment strategies. Additionally, the key predictive variables identified by the automated machine learning model align closely with the criteria used by Recruiting Station leaders. However, the model provides a more granular analysis, enabling the identification of subtle patterns and interactions between each variable.

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

ACS	American Community Survey
AGI	adjusted gross income
AUC	area under the curve
AutoML	automated machine learning
CCD	Common Core of Data
DSA	Direct Selling Association
FBI	Federal Bureau of Investigation
HS/CC	high school/community college
IRS	Internal Revenue Service
MCRC	Marine Corps Recruiting Command
MCRISS	Marine Corps Recruiting Information Support System
NCES	National Center for Education Statistics
NELS:88	National Education Longitudinal Study of 1988
RS	Recruiting Station
RSS	Recruiting Substation
SAMA	Segmentation Analysis and Market Assessment
SNCOIC	Senior Staff Noncommissioned Officer in Charge
USPS	United States Postal Service
XO	Executive Officer

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. INTRODUCTION

Since the implementation of the all-volunteer force in 1973, the Marine Corps has continuously evolved its recruitment strategies and policies to attract a diverse and skilled pool of recruits. Despite these efforts, the military recruiting environment today is facing significant challenges, including a shrinking pool of eligible and interested prospects, increased competition with the civilian job market, and changing perceptions about military service among younger generations (Berger, 2022; Kleykamp et al., 2023). These factors have made it increasingly challenging to meet recruitment goals and maintain the necessary readiness levels. Due to these difficulties, the Marine Corps must find new ways to target the right demographics effectively.

The nation's public high schools are the cornerstone of the Marine Corps' recruiting efforts and are a significant source of recruiting prospects (Headquarters U.S. Marine Corps, 2015). However, the abundance of high schools in each Recruiting Station's (RS) area often exceeds the capacity of recruiters to engage with them all effectively. For this reason, each RS assesses the potential of all schools and assigns a priority status to each school. The RS leadership sets these priorities to focus recruiter efforts and maximize the results of their efforts.

The most important decision in the high school recruitment plan is determining a school's priority status. Currently, each school's total senior population is given primary consideration when choosing a school's priority status. This approach, however, might overlook other crucial factors, such as socioeconomic backgrounds, which could be instrumental in predicting contract results.

As General Berger emphasized, "Current recruiting practices across the joint force are not producing adequate numbers of soldiers, sailors, airmen, Marines, and Coast Guardsmen. Meeting recruiting goals will require the military to reevaluate and adjust its methods, not double down on existing approaches in hopes of achieving different results" (Berger, 2022). This thesis aims to build a model using Automated Machine Learning that

recruiters can use to optimize the allocation of recruitment resources and identify critical factors that determine the potential and success rates of contracting recruits in public high schools.

B. PROBLEM STATEMENT

The Marine Corps faces a tightening recruiting environment while restricted to limited resources. As the Marine Corps' largest producer of recruits, public high schools are critical in meeting recruitment targets. The Marine Corps must strategically allocate its recruiters to maximize recruitment efficiency by ensuring high schools are assigned the correct priority status. Current methods for prioritizing high schools may result in underrecruiting some schools while overrecruiting at others.

C. RESEARCH QUESTIONS

Primary Research Question 1: Using open-source demographic data, can a recruitment tool be developed using machine learning that accurately assigns high school priority status for public high schools?

Primary Research Question 2: Are machine learning approaches using open-source demographic data better at predicting and assigning high school priority status than current methods?

Secondary Research Question: What combination of factors and characteristics are most predictive in determining which high schools produce the most recruits?

D. THESIS ORGANIZATION

The remaining chapters of this thesis are structured as follows: Chapter II will provide a concise overview of the Marine Corps' high school recruiting process and an introduction to Automated Machine Learning. Chapter III presents a summary of relevant literature. Chapter IV details the data used and outlines the methodology employed. Chapter V discusses the findings, and Chapter VI concludes with recommendations for implementation and further research.

II. BACKGROUND

A. MARINE CORPS RECRUITING HIGH SCHOOL PROGRAM

The Marine Corps Recruiting Command is divided into two regions: the Eastern Recruiting Region and the Western Recruiting Region. It is further organized into six districts, each covering several states. Within these districts are RSs, usually situated in large urban centers, and these stations are further broken down into Recruiting Substations (RSSs), responsible for outreach in smaller towns and rural areas.

While all major Marine Corps recruiting programs are important, the High School/ Community College program (HS/CC) is the cornerstone of the Marine Corps recruiting efforts because high school seniors make up around half of all recruit contracts (Headquarters U.S. Marine Corps, 2015). The XO at each RS is appointed as the program manager for the HS/CC program, and they are responsible for the overall administration of the program. While the HS/CC program includes all high schools and community colleges within an RS, this thesis will focus on public high schools due to data availability.

Recruiters in most RSs have more high schools than they can effectively work. Because of this, the most important decision in the HS/CC plan involves determining the priority status of each school. Additionally, the recruiters at each Recruiting Substation (RSS) have varying distributions of school types and sizes, so it is important for schools to be ranked accordingly. In recruiting, priority indicates the level of emphasis on a specific school compared to others within each RS. The RS XO, in collaboration with the Senior Staff Noncommissioned Officer in Charge (SNCOIC), is responsible for assigning each school a priority status based on the following criteria: the size of the senior class, Armed Services Vocational Aptitude Battery pass rates, and the number of contracts from prior years. Priority schools have the potential to be productive, and recruiters are required to conduct all elements of the high school program at these schools. The three levels of priority designations are below (Headquarters U.S. Marine Corps, 2015):

1. Priority 1: A school where the combination of these same factors indicates the recruiter should get excellent results from efforts.

2. Priority 2: A school where the combination of these same factors indicates the recruiter should get good results from efforts.

3. Priority 3: A school where the combination of these same factors indicates the recruiter should get sufficient results from efforts to justify the school's working status.

A Non-Priority school is one in which the XO and SNCOIC determine it is ineffective to conduct all high school program elements. Although a school might be considered Non-Priority, recruiters may still schedule recruiting events, and an Initial Visit is still required for all working high schools.

B. AUTOMATED MACHINE LEARNING

This thesis uses cloud-based automated machine learning or AutoML to build and deploy a model to predict the probability of high schools generating recruit contracts. Below is a brief background on what AutoML and cloud-based AutoML are.

AutoML represents a significant evolution in machine learning and data science. It automates the process of machine learning methods, which has become increasingly important due to the complexity and diversity of different methods. In traditional machine learning, data analysts must manually perform various steps. These include feature selection, where they identify the most relevant variables to speed up training times and improve a model's accuracy (Solorio-Fernández et al., 2022); feature engineering, which involves creating, selecting, and refining variables to improve model performance (Dong & Liu, 2018); and hyperparameter tuning, adjusting the model settings to optimize performance before training beings (Bartz et al., 2023). As De Bie et al. (2022) explain, AutoML significantly enhances efficiency by automating these tedious tasks, saving time and effort for seasoned data analysts while enabling individuals with limited machine learning knowledge to develop and deploy complex models.

Cloud-based AutoML is another advancement in machine learning, further expanding its accessibility. Cloud-based AutoML services, offered by major cloud providers like Google, Amazon, and Microsoft, provide an integrated environment where

analysts can easily access AutoML capabilities without requiring extensive computational resources or advanced technical expertise in machine learning. These platforms leverage the massive computational resources of cloud infrastructures, allowing for the efficient processing of large datasets and rapid evaluation of numerous machine learning models. In addition, cloud-based AutoML services provide a relatively user-friendly interface. The platforms offer a graphical interface where users can upload datasets, initiate the model training process, and evaluate the performance of the results, all without writing a single line of code.

THIS PAGE INTENTIONALLY LEFT BLANK

III. LITERATURE REVIEW

This chapter examines previous academic studies that provide insights into machine learning, independent predictor variables for enlistment, and military recruiting models. I aim to identify key trends, methodologies, and outcomes of these studies, offering an understanding of how machine learning and predictive analytics can enhance military recruitment strategies. I will also identify gaps in current knowledge and illustrate how my work can improve and add to the existing literature.

A. MACHINE LEARNING

Although there is limited research available on machine learning for military recruiting, researchers have studied machine learning extensively for sales and marketing applications. While there are many such studies, here I review a few examples.

Glackin and Adivar (2023) address the potential benefits of applying data mining and machine learning algorithms in sales and marketing research, mainly focusing on the direct-selling industry. Using the 2018 Direct Selling Association (DSA) National Salesforce Survey, which included demographic, behavioral, and attitudinal information from direct-selling independent representatives, they found that machine learning models significantly outperform traditional statistical methods in predicting sales performance. Their findings reveal that machine learning algorithms can more accurately identify patterns and predictors of success within the salesforce, such as key behavioral traits and demographic factors that correlate with higher sales figures and customer retention rates. Similar to how Glackin and Adivar demonstrated machine learning's ability to identify successful salesforce characteristics, my thesis hypothesizes that similar machine learning approaches can significantly improve the prediction of high schools' potential to produce recruits.

In exploring the potential of AutoML, Shahriyar et al. (2022) offer an example of AutoML's application for predicting employability and recruiting. The study used student demographic data to predict whether students possessed a desirable level of employability. The results were compared with traditional machine learning models. The models were run

on two different datasets, and the AutoML model achieved the highest accuracy with a value of 72.45%, beating the next highest model by 1.17%. In the second dataset, the AutoML model had an accuracy of 84.65%, beating the next highest model by .46%. Although the AutoML models performed similarly to other machine learning models, the hyperparameter optimization is automatic, which takes much less human time to run. Although my thesis focuses on high schools, not individual students, Shahriyar et al. demonstrate that AutoML can accurately predict hiring outcomes using demographic data.

Rezazadeh (2020) took a novel approach to predict sales outcomes using Azure AutoML, a cloud-based machine learning program. While my thesis predicts the likelihood of a school producing a recruit, Rezazadeh predicted won and lost sales opportunities and compared it to human performance in predicting those outcomes. A total of 20 relevant variables were used for each sales opportunity, encompassing aspects of the sales project and customer detail. The model demonstrated an accuracy of 85% and an Area Under the Curve (AUC) of 0.87, indicating a strong capability to distinguish between won and lost sales. This was much better than human performance, which had an accuracy of just 67%. Additionally, the machine learning predictions excelled in monetary performance and had a higher cumulative value of correctly classified sales opportunities. My research aims to produce similar outcomes but with the total number of recruits as the outcome rather than monetary value.

These articles highlight the potential of machine learning in various fields and lay a foundation for its application in military recruitment. The studies demonstrate how machine learning can significantly enhance predictive accuracy and efficiency, leading to more targeted and effective strategies. My use of AutoML will take these approaches and automate selecting and tuning the most suitable algorithms.

B. MILITARY RECRUITING (SOCIOECONOMIC AND DEMOGRAPHIC FACTORS)

Having established the effectiveness of machine learning, it is essential to examine its applicability within a more specific area. This brings us to the role of socioeconomic and demographic factors in military recruiting.

Lutz (2008) analyzed demographic trends in the U.S. military, focusing on race, class, and immigration status. Initially, she examines representation across race-ethnicity by comparing the percentage of different racial-ethnic groups in the military with the general population. Using data from the 1980, 1990, and 2000 Integrated Public Use Microdata Series and Defense Equal Opportunity Management Institute, she finds that in 1981, blacks were disproportionately represented while whites and Hispanics were underrepresented. By 2000, the disparities in representation among these groups remained unchanged.

In the second part of her study, Lutz (2008) uses data from the National Education Longitudinal Study (NELS:88) to examine trends on how race, class, and immigration status affect decisions to serve in the military. She finds that the probability of joining the military is comparable across various racial and ethnic groups. This suggests that military service is accessible to individuals from all backgrounds, challenging common misconceptions and statistics from 2000 regarding active duty service members and suggesting that further analysis is warranted. However, she did find that family income is inversely related to military service, and lower-income individuals are more likely to serve.

While Lutz (2008) uses comprehensive data, the NELS:88 only partially represents current or broader trends. We cannot apply changes in socio-political contexts and population demographics to today's Marine Corps. Additionally, the study does not address why these factors may influence an individual's decision to enlist. Furthermore, the study's reliance on correlational data limits its ability to establish causal relationships between variables such as race, class, and military service. Determining if these factors directly influence enlistment choices or are just associated with other unmeasured variables is challenging.

Unlike Lutz (2008), who provides a broad demographic analysis, Kleykamp (2006) offers a more focused study on the influence of community and educational factors on military enlistment decisions. Using data from 2002 on Texas high school graduates, Kleykamp (2006) uses multinomial logistic regression to identify factors that influence high school graduates' decisions to enlist in the military. The study includes variables for educational aspirations, military presence in the community, socioeconomic status, racial

and ethnic background, family characteristics, and number of recruiters in the area. She finds that having a desire to attend college reduces the likelihood of preferring military service over pursuing higher education but also increases the likelihood of opting for the military instead of entering the workforce. Additionally, she finds that military presence in communities positively influences enlistment decisions. Lastly, people from families with lower incomes, bigger family sizes, and parents who have received less education are more inclined to enlist in the military.

Her work highlights the relationship between education, familial background, and military enlistment. This relationship is critical for devising more effective and targeted military recruitment strategies. Although the study is geographically and temporally specific, focusing on Texas high school graduates in 2002, its insights offer an understanding of the socioeconomic and demographic drivers behind enlistment decisions. This article, along with Lutz (2008), provides a guide on developing predictive models for recruitment by suggesting variables to be considered.

C. MILITARY RECRUITING MODELS

Building on factors influencing military enlistment, we now analyze military recruiting models. Using data on Army recruit contracts from 2010 to 2014, Marmion (2015) examines the Segmentation Analysis and Market Assessment (SAMA) tool that the Army uses to evaluate the recruitment potential of recruiting centers within market segments. He finds that the SAMA model overestimates recruitment opportunities for 96% of recruiting centers. Marmion (2015) uses additional factors and modeling approaches as an alternative to the SAMA. The other variables he includes are unemployment rates, number of recruiters assigned to each center, responsibility area for each recruiting center, distance to the nearest Medical Entrance Processing Station, and total qualified military available in the area. He finds that the new modeling techniques enhance predictive power and only overestimated average recruiting potential by 3.8%. Additionally, he finds that including data on past performance results in an R-squared of .89 on the test set, while the model without past performance data results in an R-squared of .65. R-squared is a measure of the percentage of the dependent variable's variance that the independent variables

explain. A higher R-squared indicates that a model can better predict the dependent variable using the independent variables.

While Marmion's (2015) models offer insightful approaches, they are not without limitations. The reliance on historical data (the four-year average of enlistments) poses a risk of lag in response to real-time changes, potentially leading to less accurate forecasts in evolving markets. My model uses a combination of historical and real-time data to predict recruitment outcomes.

Using open-source data from 2011 to 2013, Intrater (2015) aims to predict the number of Navy enlisted accessions based on past results. He uses data from various sources, including the Integrated Postsecondary Education Data System, the Internal Revenue Service, the Federal Bureau of Investigation, and the U.S. Census Bureau. The data is at the ZIP code and county level and includes 71 variables partitioned into six categories: military influence and recruiter workload, crime, population characteristics, economic stability, education opportunities, and veteran population. Intrater (2015) used a multiple linear regression model at the recruiting station level and a zero-inflated negative binomial model to account for the large number of 0s, as 64% of the ZIP codes have 0 accessions. These models' performance was checked using sample data to predict 2012 accessions. Based on these models, he found that recruiter strength (total number of recruiters in an area), areas with adjusted gross income (AGI) below $25,000, and areas with violent crime showed the strongest positive indicators for recruitment. Additionally, wealthier areas (AGI > $200,000), with higher educational opportunities, like concentration of Division I universities, produce fewer recruits.

Next, Intrater (2015) analyzed open-source data from the IRS, FBI, and Census Bureau, employing two models to predict annual Navy accessions. Using data from 2013–2015, he used multiple linear regression and zero-inflated negative binomial models. He found that recruiter strength, low-income areas, and high crime rates positively influenced Navy enlistment accessions, while areas with higher incomes and educational opportunities yielded fewer recruits. The models included 71 independent county and ZIP code-level predictor variables. Intrater grouped the predictor variables into six categories: military influence and recruiter workload, crime, population characteristics, economic stability,

education opportunities, and veteran population. The zero-inflated negative binomial models consider these 0s and allow for a more accurate data analysis by distinguishing between zeros that occur as part of the natural variation in the data and zeros that arise due to specific reasons.

Intrater's (2015) models identified significant predictors influencing Navy accessions and showed improved performance from previous zero-inflated models. However, a shortcoming of the data is that crime is self-reported and is only reported at the county level. The data was linked to the ZIP code level by multiplying the crime statistics reported for each county with the population proportion of each ZIP code within these counties. This relies on the assumption that crime is uniformly distributed within a county, which is usually untrue.

Using Army recruit leads from 2011–2013 and variables from publicly available data, Fulton's (2016) models aim to improve the efficiency of Army recruiting using a Tree Clustering algorithm. He finds that economic data, which includes variables on the total number of businesses, types of businesses in each ZIP code, and individual income tax returns, was the most predictive. Fulton found that cluster assignments based on economic data outperformed the previous model the United States Army Recruiting Command used and achieved an R-squared of 0.69. The clustering process involved 347 publicly available variables categorized into five groups to ensure each type of dataset did not mask the other. The categories included demographics, health, education, economic, and military. He clustered ZIP codes and then applied Poisson regression models to the clusters, using the number of national leads—individual requests for military service information—as the response variable.

Instead of analyzing actual Army accessions, Fulton (2016) uses national leads as the dependent variable, a metric that may not fully capture the outcomes of the Army recruitment process. While initial inquiries can provide valuable insights into public perception and interest levels, they do not necessarily translate into actual enlistments.

D. CONCLUSION

This literature review highlights the landscape of military recruitment, where machine learning and predictive analytics are emerging as essential tools in enhancing recruitment efficacy. Prior literature, such as those by Glackin and Adivar (2023), Shahriyar et al. (2022), and Rezazadeh (2020) offer insights into how machine learning can be used to predict outcomes accurately. Lutz (2008) and Kleykamp (2006) discuss socioeconomic and demographic factors that influence military enlistment. These factors provide insights into which predictive variables should be included in recruiting models. Marmion (2015) and Intrater (2015) demonstrate the potential of military recruiting models to accurately predict recruitment outcomes.

In this thesis, I aim to extend these insights by proposing a novel machine learning-based approach to identify priority high schools for Marine Corps recruitment. Drawing on Marine Corps accession data, public high school data, and American Community Survey demographics and socioeconomic variables, this research aims to develop an AutoML model, integrating demographic, educational, and community variables like the abovementioned models. This approach is expected to refine recruitment efficiency by accurately predicting high schools' potential for recruit contracts, a significant leap from the current methods limited to employing and testing a couple of models at a time. The integration of AutoML in this context allows for the rapid testing and deployment of multiple models, overcoming the constraints of traditional methods. The anticipated outcome is a more dynamic and responsive recruitment process that adapts to changing socioeconomic landscapes and demographic shifts.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. DATA AND METHODOLOGY

In this chapter, I discuss where the data came from and how it was merged and managed. Then I explain the methodology. A list of the variables and summary statistics from the training and test datasets are listed in the appendix.

A. DATA SOURCES

1. Marine Corps Recruiting Command

I received the recruiting data from Marine Corps Recruiting Command. The data was obtained from the Marine Corps Recruiting Information Support System (MCRISS) and contained information on all high schools and community colleges where the Marine Corps recruits. The dataset spans from 2016 to 2022 and has variables for school code, region, district, recruiting station, substation, priority code, number of male and female seniors, location, and number of high school seniors who signed enlistment contracts per year.

2. Common Core of Data

The Common Core of Data (CCD) is a national database for all public elementary and secondary schools in the United States. It is managed by the National Center for Education Statistics (NCES), which is part of the U.S. Department of Education's Institute of Education Sciences. The CCD collects yearly data on schools, school districts, students, and staff and includes enrollment numbers and demographics. The NCEN provides a web application that allows for the download of specific variables for each school by year. The datasets I downloaded are from the school years 2016–2021.

3. American Community Survey

The American Community Survey (ACS) is a yearly, ongoing survey conducted by the U.S. Census Bureau that includes information about the nation and its people. I used the 2021 ACS 5-Year dataset, which is a period estimate with data collected over 5 years. The datasets include social, economic, housing, and demographic data by ZIP Code Tabulation Area. ZIP code tabulation areas are generalized representations of the United States Postal Service (USPS) ZIP code areas. They are intended to mirror USPS ZIP codes as closely as

possible (U.S. Census Bureau, 2023). I used four datasets from the 2021 ACS. The datasets are the DP02 (social characteristics), DP03 (selected economic characteristics), DP04 (selected housing characteristics), and DP05 (demographic and housing estimates).

B. DATA MERGE AND MANAGEMENT

The Marine Corps dataset started with 24,528 schools. The final dataset had 17,051 different schools. However, the number of schools varies within each year's datasets. Table 1 shows the number of schools I dropped and for what reason. Merging the data was one of the most time-consuming processes of this thesis. The Marine Corps and CCD had many high schools with different nomenclatures for the same schools, making the matching process tedious. Additionally, despite having ZIP codes and addresses in both datasets, there were many discrepancies, with some school's data being off by a few digits between the two datasets. To streamline future data integration, I recommend that the Marine Corps incorporate a CCD school ID variable into MCRISS. This will make adding information from the CCD much easier.

My first step was to create a database with both the CCD school IDs and the Marine Corps school codes. This would allow me to merge yearly data from the CCD and Marine Corps later. To do this, I downloaded a dataset with all public high schools from the most recent CCD data from the 2022 school year. The dataset had variables for school name, address, and school ID. Using the school name, ZIP code, and street address, I first attempted to merge the two datasets using Google Places API within R. Using these variables, Google Places queried the information and returned a unique Place ID for the schools within both datasets. Using this Place ID, I merged the datasets, which returned a match of around 50%.

For the remaining unmatched schools, I used fuzzy matching functions using a combination of school names, zip codes, and addresses. After running these functions, I was left with just 1,252 unmatched schools. At this point, I went through the remaining schools and manually matched them using the CCD school search tool on the NCES website. Finally, I checked for any duplicate matches that occurred during the fuzzy matching. Table 1 shows the number of observations that were dropped and the reason why.

Table 1. Dropped Observations

Observations Dropped	**Reason**
2,275	Marked as closed
1,813	Listed as private schools
1,525	Marked as community colleges
1,453	Schools dropped due to not matching with CCD data or ACS data. Many of these schools were private schools but were not listed as such in MCRISS.
358	Schools outside of 50 U.S. states
42	Marked as 2 year or 4 year colleges
11	Had CC or community college in school name
7,477	**Total Dropped**

Schools dropped from Marine Corps Recruiting Information Support System data (MCRISS).

Once I had a CCD school ID and a Marine Corps school code for each of the schools, I was able to merge in the selected variables for each year to create a combined Marine Corps and CCD dataset that had 17,051 different schools. The datasets from the CCD were from the prior school year, to replicate how an analyst would access and use the data for prediction for the following year. For example, the final 2018 dataset has CCD data from the 2017 school year and includes a Number of Grade 11 Students variable, which represents the number of seniors for the 2018 school year. Finally, I merged the ACS data into the combined Marine Corps and CCD data by ZIP code.

Since some of the predictor variables contain missing values, I created indicator variables to identify variables with missing information for each school. These variables take the value of 1 if the data for the predictor is missing for a given school and 0 if the data is present. This method ensures that the model does not lose valuable information due to incomplete records and can recognize patterns related to the absence of data. Once the indicator variables were added, the missing data was filled with a 0.

For this thesis, I use data from 2018, 2019, and 2022. There were 16,071 schools in the 2018 data, 16,282 in the 2019 data, and 16,901 in the 2022 data. Table 2 shows statistics for some notable variables. The rest of the summary statistics are in the Appendix.

Table 2. Key Variable Statistics

Year	Number of Schools	Number of Priority 1 Schools	Number of Priority 2 Schools	Number of Priority 3 Schools	Average Number of Senior Contracts per School	Number of Schools with At Least One Contract
2018	16,071	3,698	2,896	2,784	1.1	8,054
2019	16,282	3,711	2,909	2,923	1.2	8,411
2022	16,901	3,621	2,961	3,065	.94	7,472

C. METHODOLOGY

This thesis develops a predictive model that employs AutoML to estimate the probability of high schools producing at least one enlistment contract. Predictive modeling assigns classifications or probabilities to a target variable based on characteristics in predictor variables. As with typical machine learning models, the data is split into a training set, used to develop the model, and a test set, used to evaluate its predictive accuracy. The model's performance is determined by its ability to correctly classify the recruiting potential of each school. Performance is also measured by ranking the schools by the probability of producing an enlistment contract and then comparing the model's top schools to the Marine Corps' choices of priority schools.

The selection of independent variables for my model was guided by the literature and demographic factors known to influence military enlistment. These variables include socioeconomic status, educational attainment, and data on the incoming senior class for each high school. The goal was to use a wide spectrum of variables that might influence a high school student's decision to enlist.

The dependent variable for this model is binary, representing whether a school produced at least one recruit contract for that school year. The probability of a school producing a contract will be used to rank schools and assign a priority code.

As Figure 1 shows, the model was trained using data from the 2018 and 2019 school years and tested using data from the 2022 school year. The reason for the gap in years from 2019 to 2022 is due to COVID. I wanted to use data more representative of a typical recruiting environment, so I chose not to test on these years. However, I also ran a model using data from 2017 and 2018 to train and 2019 to test. The results are in the appendix and are very similar.

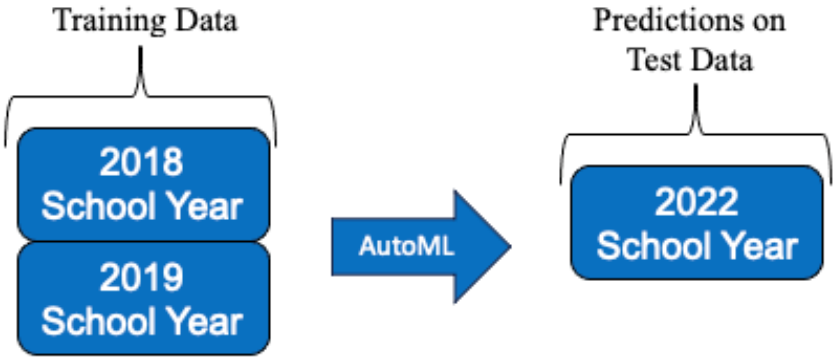

Figure 1. Model Design

I chose to use Microsoft Azure AutoML, a cloud-based machine learning service, to conduct my research. AutoML automatically performs feature selection, model selection, and hyperparameter turning. The training data was input into Azure AutoML, which ran multiple types of machine learning models using the program's default settings. The models were evaluated using the AUC to measure their accuracy in distinguishing between schools likely to produce a contract and those that are not. Higher AUC values indicate better model performance.

Next, I selected the best performing model, based on highest AUC, and tested the algorithm using data from the 2022 school year. I downloaded the results and then analyzed how well the model performed compared to the Marine Corps' priority designations for that year. The results are discussed in the next chapter.

THIS PAGE INTENTIONALLY LEFT BLANK

V. RESULTS AND ANALYSIS

A. MODEL RESULTS

When tested using the 2022 data, the model exhibited an accuracy of 74.3% and an AUC of 0.818. An accuracy of 74.3% is particularly significant when considering just 44.2% of schools in the test data produced at least one contract. A higher AUC value, closer to 1, indicates better model accuracy and a value of 0.818 is generally considered good (Hosmer Jr., et al., 2013). Additional metrics are listed below in Table 3. Sensitivity refers to the ability to correctly identify true positives. In this case, it is the percentage of high schools that were correctly identified as having at least one contract. Specificity measures how accurately true negatives are identified, or in this case, the percentage of high schools that were correctly identified as not having any contracts. Overall, these metrics indicate that the model has a strong ability to distinguish between schools that produce at least one recruit and those that do not. However, the metrics are somewhat subjective and the next section will solidify the model's performance.

Table 3. Model Results Metrics

AUC:	0.818
Accuracy:	0.743
Sensitivity:	0.740
Specificity:	0.746

To further illustrate the model's predictive ability, Figure 2 is a histogram that shows the true positives (in green) and the tree negatives (in red). The peaks in the green and red areas represent areas of high confidence in correct predictions and make up most of the distribution. If the predictions were perfect, there would be no overlap between the red and green bars.

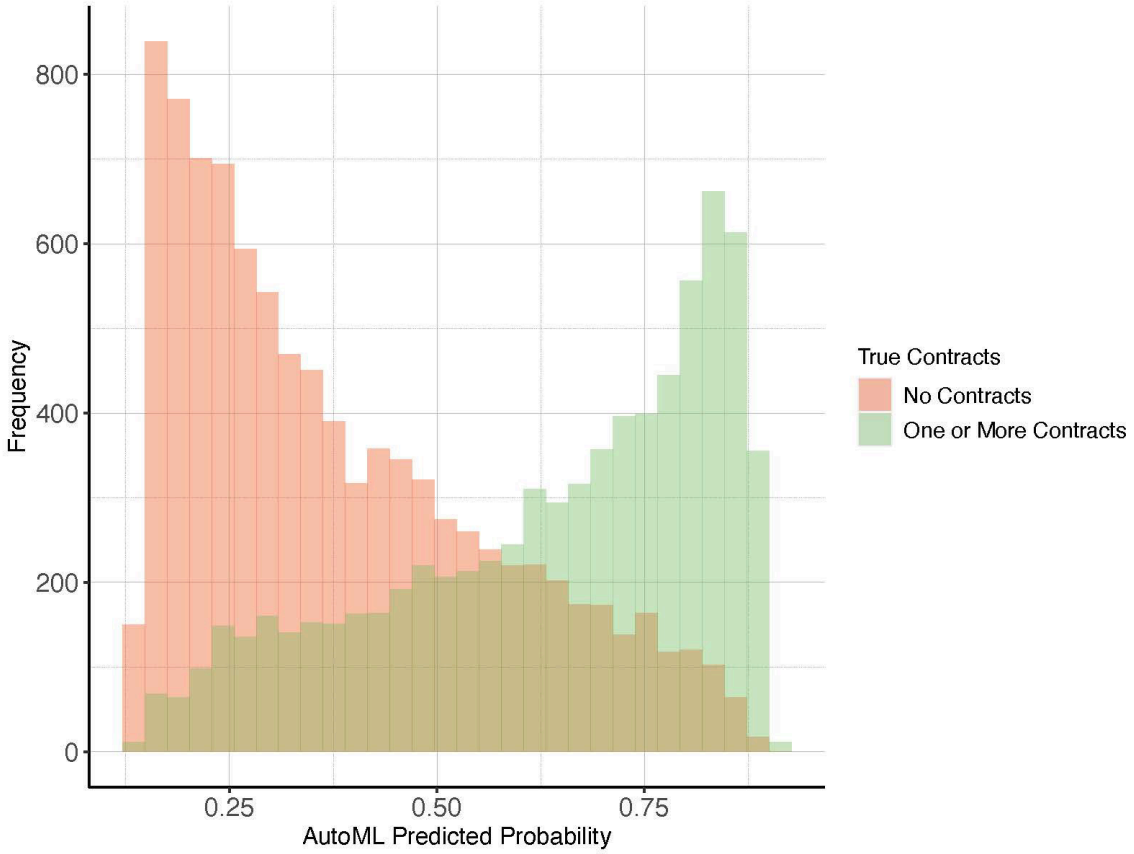

Figure 2. AutoML Results: True Positives and Negatives

B. AUTOML MODEL AND MARINE CORPS COMPARISON

Next, I analyzed how well the model performed compared to the Marines' choices of priority high schools. In the 2022 test data, there were 9,647 schools with a priority code of 1, 2, or 3. Of these priority schools, 63.3% of them generated at least one contract. Of the top 9,647 schools from the AutoML model, ranked by predicted probability, 63.8% generated at least one contract. So for this large group, the AutoML is not significantly better at predicting which schools will generate contracts.

What if the AutoML model was used to generate priority codes like the Marines? To answer this, I found the number of schools the Marine Corps had assigned for each priority code within the 2022 test data. Then I ranked the schools from the AutoML predictions based on predicted probability and assigned new priority codes to those schools, matching the original amount of schools within each priority code. The results are

shown in Figure 3. The blue bars show the original priority code results while the red bars show the inferred priority codes from the AutoML model's results. Success rates means the percentage of schools that were accurately identified as producing at least one recruit. For Priority 1 schools, the AutoML model had a success rate of 83.4%, while MCRC had a success rate of 79.0%. For Priority 2 schools, the AutoML model had a success rate of 63.0%, while MCRC had a success rate of 62.8%. For Priority 3 schools, the AutoML model had a success rate of 41.0% and MCRC had a success rate of 45.4%. Although it appears there is little difference between the AutoML predictions and MCRC's choice of priority schools, the results indicate that the AutoML's algorithm can be used as an effective tool.

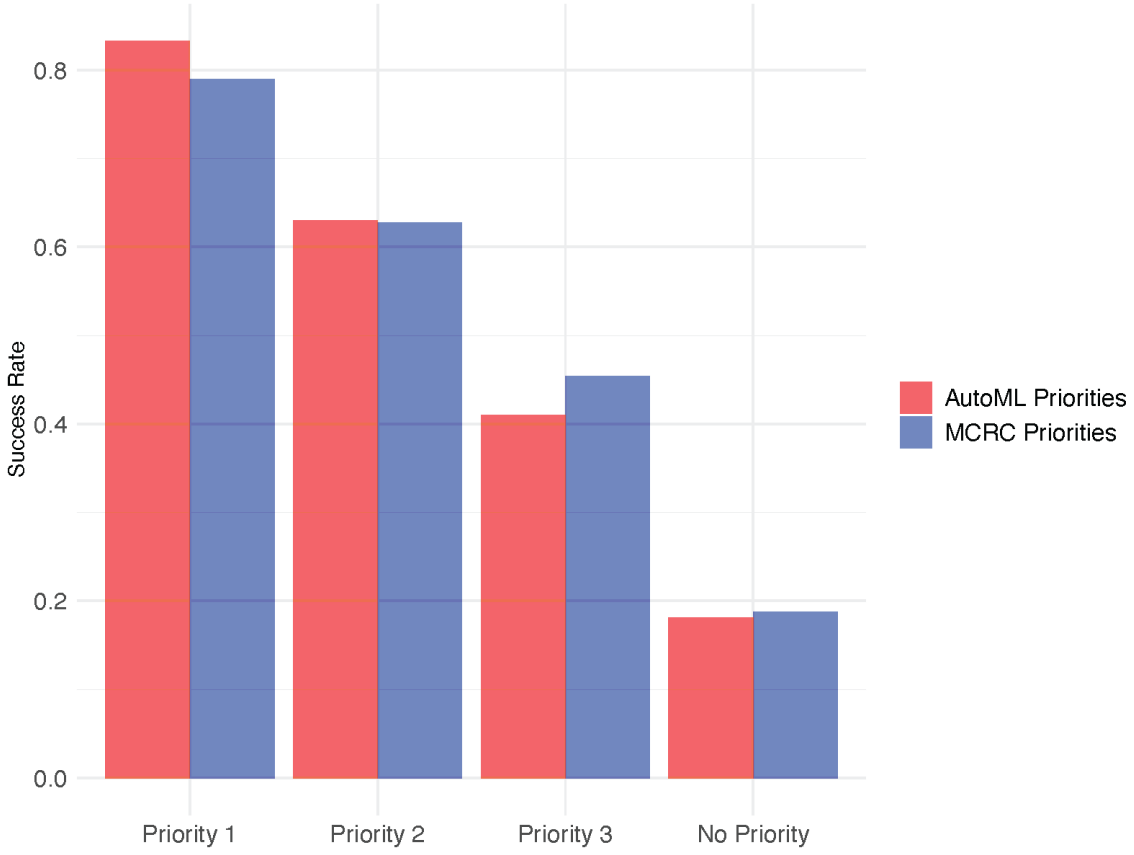

Figure 3. AutoML vs. MCRC Success Rates by Priority Code

The results above only compare the success rates of the number of schools with at least one contract produced, not the total number of contracts produced from those schools.

When the number of contracts is regressed on the priority code, as shown in Table 4, the R-squared is 0.250. When the number of contracts is regressed on the predicted probability from the AutoML model, the R-squared increases to 0.310. The higher R-squared indicates that the more granular AutoML predictions do a better job at predicting variation in the number of recruits. Additionally, when the number of contracts is regressed on the AutoML model's predicted probability and MCRC's priority codes, predicted probability retains its significance and the R-squared increases. These findings suggest that while the current prioritization generally aligns with the model's contract production outcomes, integrating AutoML predictions could potentially lead to a more efficient allocation of recruiting resources and a higher overall number of contracts generated.

Table 4. Number of Contracts Regressed on Priority and Predicted Probability

	Dependent variable:		
	Actual Contracts		
	(1)	(2)	(3)
Priority 1	1.912***		0.436***
	(0.026)		(0.042)
Priority 2	1.088***		-0.047
	(0.028)		(0.037)
Priority 3	0.524***		-0.203***
	(0.028)		(0.031)
Predicted Probability		3.591***	3.177***
		(0.041)	(0.073)
Constant	0.245***	-0.840***	-0.683***
	(0.015)	(0.023)	(0.026)
Observations	16,901	16,901	16,901
R^2	0.250	0.310	0.327

Note: *p<0.1; **p<0.05; ***p<0.01

The binscatter plots below show the AutoML model's predictive capability on contract production. The X-axis represents the model's predicted probability of a school

producing at least one contract and the Y-axis represents the average number of contracts those schools produce. Figure 4 includes schools with all priority codes while Figure 5 separates each of the different schools by priority code. They all show a clear upward trend in the average number of actual contracts per school as predicted probability increases. This shows that the model is adding information, otherwise the slope would be zero as probability increases.

The graph illustrating schools without a priority code indicates that schools not labeled as recruitment priorities are experiencing a positive trend, particularly those with a predicted probability exceeding 0.75. This indicates the model has identified schools with contract-producing potential that the Marine Corps did not prioritize in 2022.

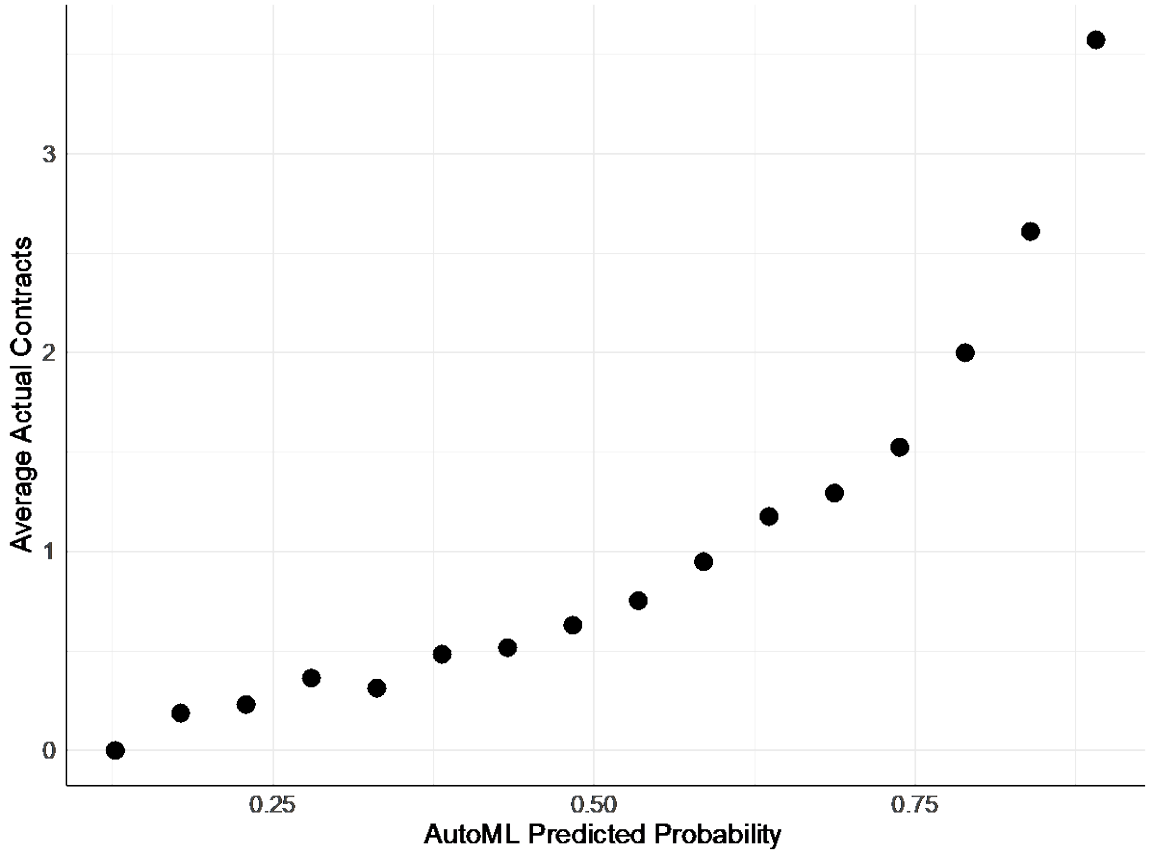

Figure 4. All Priority Schools: Average Number of Contracts vs. Predicted Probability of AutoML Model

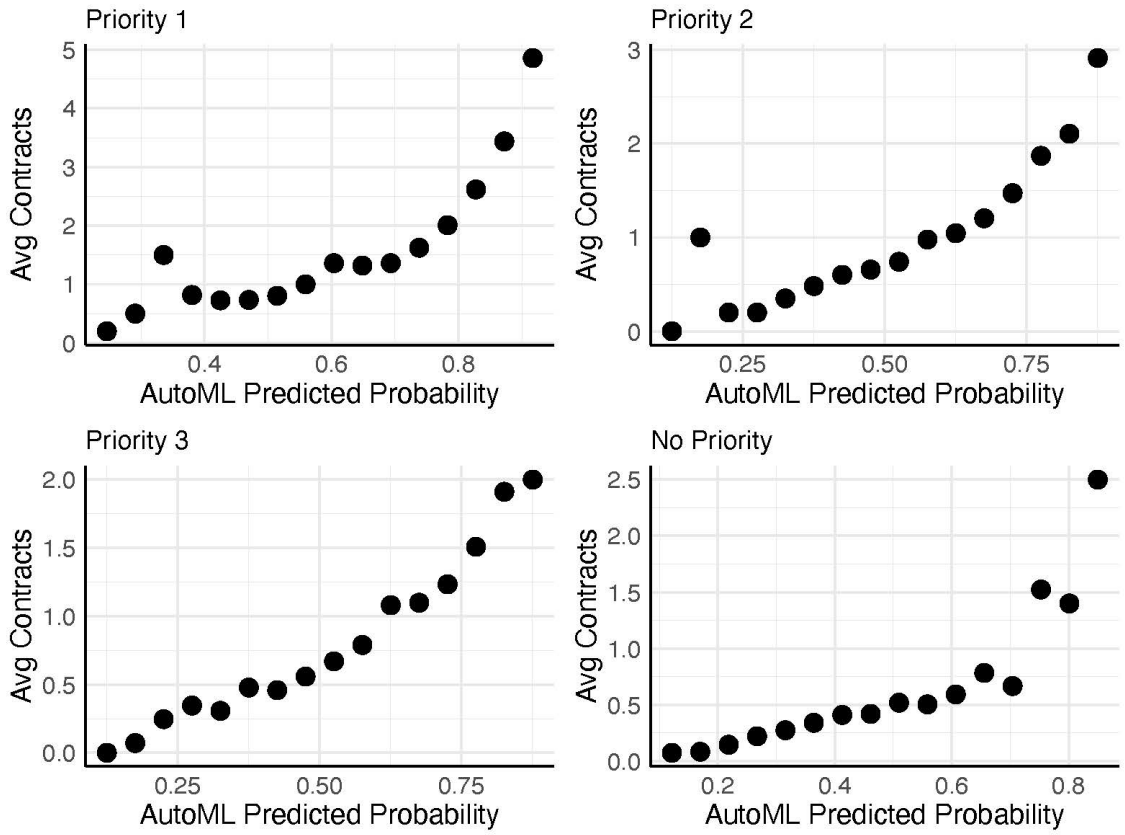

Figure 5. Average Number of Contracts vs. Predicted Probability of AutoML For Each Priority Code

C. COUNTERFACTUAL

What would happen if recruiters implemented the AutoML produced scores instead of the existing priority codes? To answer this, I compared the average number of contracts produced between the AutoML model's top schools and the Marines' priority schools.

The 2022 test data had around 3,621 Priority 1 schools, 6,582 Priority 1 and 2 schools, and 9,647 Priority 1, 2, and 3 schools. So, similarly to how I compared success rates, I ranked the AutoML model's predictions based on those numbers of schools. The results are in Table 5 and illustrated in Figure 6.

If the Marine Corps had recruited at just the top 3,621 schools based on the AutoML predictions, they would have produced about 12% more contracts compared to the Marine Corp's choices of Priority 1 schools. If they had visited just the top 6,582 schools from the

AutoML model, they would have produced 4.7% more contracts. If they visited the top 9,647 schools, they would have produced 0.85% more contracts.

This means recruiters would have been able to prioritize fewer schools while achieving similar or even better results. Also, it's reasonable to assume that by prioritizing fewer schools, more contracts could have been produced from those schools since they would have more time and resources devoted to them.

Table 5. Comparison of Average Contracts Per School by Priority

Number of Schools	AutoML Average Contracts Per School	Marine Corps Average Contracts Per School	Percentage Difference
3621 (Priority 1)	2.413	2.156	11.9%
6582 (Priority 1& 2)	1.870	1.786	4.7%
9647 (Priority 1, 2, & 3)	1.473	1.463	0.85%

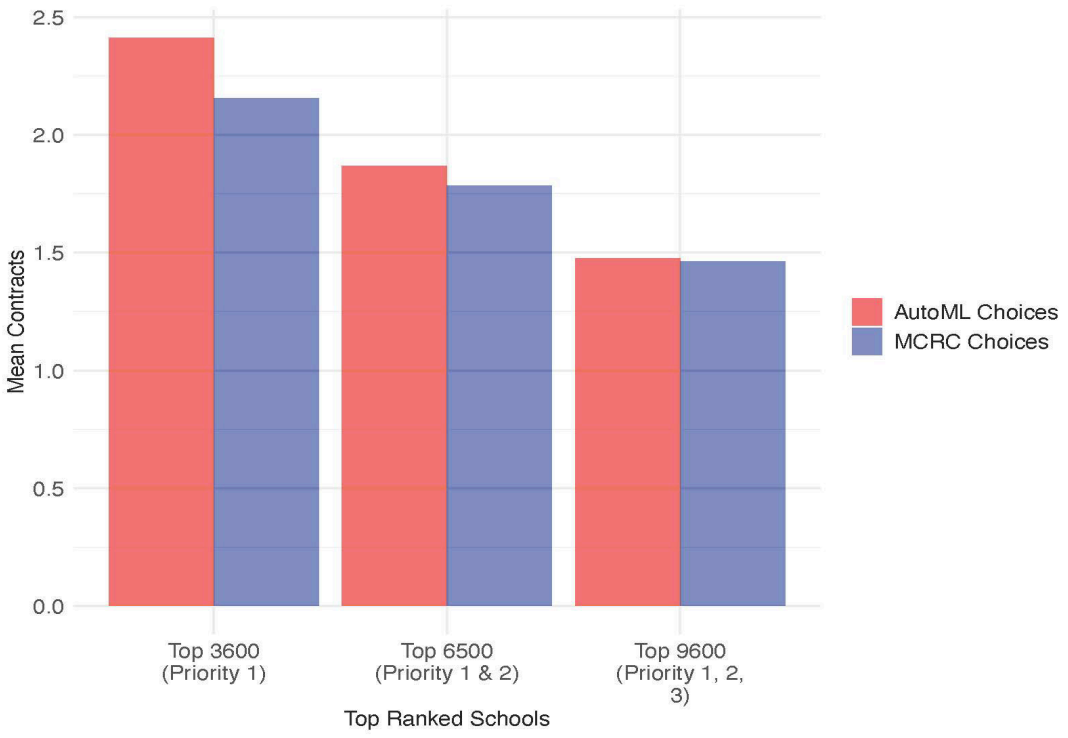

Figure 6. Counterfactual: Average Contracts Produced by Priority Code

Additionally, of the highest 9,647 predicted probability schools from the AutoML model, 1,049 were given no priority code by the Marine recruiters. However, over one-third of these schools generated at least one contract, and these schools averaged 0.522 contracts per school. While the exact level of effort invested in these schools is uncertain, it is reasonable to assume that they received significantly less attention. Had recruiters prioritized these schools, we can assume the outcomes would have been even higher. The next paragraph quantifies the potential impact of prioritizing these schools.

To estimate the potential impact of prioritizing schools the Marine Corps did not prioritize, I first calculated the number of schools within each priority code that the AutoML model assigned but the Marine Corps did not. The AutoML model identified 29 schools as Priority 1, 217 schools as Priority 2, and 803 schools as Priority 3, all of which had not been assigned any priority by the Marine Corps. Of these schools not considered a priority by the Marine Corps, Priority 1 schools averaged 1.66 contracts, Priority 2 schools averaged .664 contracts, and Priority 3 schools averaged .443 contracts. Schools assigned Priority 1 by the Marine Corps averaged 2.16 contracts, Priority 2 schools averaged 1.33 contracts, and Priority 3 schools averaged .769 contracts per school. By extrapolating from these findings and taking the difference between the average number of contracts per school of the AutoML and Marine Corps' priority designations, prioritizing these schools could potentially have resulted in an additional 421 contracts.

D. MOST PREDICTIVE VARIABLES

Below is a list of the top predictor variables according to Azure's model explanation of the AutoML algorithm. Table 6 is a regression table with these variables. The variables have been standardized, and although it only captures the linear relationship of whether a school produces at least one recruit, it is a quick way to look at the strength and direction of the variables in terms of the prediction. This list includes both negative and positive predictors.

- Number of grade 11 male students
- Number of grade 9–12 students

- Total number of grade 11 students
- Number of contracts from previous year
- Number of contracts 2 years ago
- Number of grade 11 white males
- Number of free and reduced lunch students
- Number of total male Students
- Percentage of population in ZIP code with a graduate or professional degree
- Number of grade 11 white females
- Percentage of population in ZIP code with a bachelor's degree
- Number of grade 11 Hispanic male students
- Percentage of population that are veterans

The top predictors are generally consistent with how the RS leadership currently assigns priority designation, but the additional variables add nuance and additional predictive power. It is not surprising to see that the top variables are related to the size of the school. Larger schools, especially those with more males, are most likely to yield more recruits. Additionally, the significance of a school's history of contracts emphasizes the expected correlation that prior success predicts future success. As discussed in the literature, individuals with less-educated parents are more likely to join the military (Kleykamp, 2006), which is consistent with the model's results. Finally, the percentage of the population that is veterans was ranked as the 12th most predictive variable and is also consistent with the literature (Kleykamp, 2006).

Table 6. Top Predictor Variables Regression Table

	Dependent variable:
	At Least One Contract Predicted
Number of Grade 11 Male Students	0.119***
	(0.032)
Number of Grade 9-12 Students	0.385***
	(0.019)
Number of Grade 11 Students	-0.319***
	(0.037)
Contracts from 2 Years Prior	0.055***
	(0.003)
Prior Year Contracts	0.049***
	(0.003)
Number of Grade 11 White Male Students	0.110***
	(0.016)
Number of Free and Reduced Lunch Students	0.058***
	(0.004)
Educational Attainment. Population 25 Years and Over. Graduate or Professional Degree	-0.029***
	(0.004)
Number of Grade 11 White Female Students	0.039**
	(0.016)
Educational Attainment. Population 25 Years and Over. Bachelors Degree	-0.004
	(0.004)
Number of Grade 11 Hispanic Males	-0.007
	(0.005)
Veteran Status. Percentage of Civilian Population 18 and Over. Civilian Veterans	0.012***
	(0.003)
Constant	0.475***
	(0.002)
Observations	16,901
R^2	0.588
Note:	*p<0.1; **p<0.05; ***p<0.01

E. LIMITATIONS

Currently, the RS leadership makes the decision of which schools to make a priority which means those choices are relative to that region. This thesis analyzes priority designations relative to the entire nation, not to individual RS and RSSs. This means some of the higher-ranked schools from the AutoML model might not make sense to recruit at

when compared to other schools within the RS. However, the model's results could be useful in determining recruiting force structure and mission share.

Additionally, although the Marine Corps also recruits at private schools and community colleges, the model does not include these schools because open-source data is not as easily accessible.

Finally, most cloud-based AutoML services are not free to use. However, Microsoft Azure charges less than $1.00 for each hour of model runtime. If the Marine Corps is unwilling to pay for this service, there are other AutoML packages within Python and R that are free.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

This thesis has demonstrated the significant potential of AutoML models in improving recruitment strategies in the Marine Corps. The results show that AutoML, using open-source data, can be an effective tool for predicting which public high schools to prioritize. Additionally, the AutoML predictions produced more contracts than the Marine Corps' choices of priority schools. This outcome indicates that the use of AutoML can provide more efficient and accurate insights into recruitment potential. As a result, resources can be allocated more strategically, ultimately leading to a higher return on recruitment efforts.

The key predictive variables identified by the AutoML model, especially the number of males in the incoming senior class as the most significant factor, align closely with the criteria used by RS leadership for their assessments. However, the AutoML model provides a more granular analysis, enabling the identification of subtle patterns and interactions between each variable.

B. RECOMMENDATIONS

I am sure even better models can be developed using additional datasets and other variables, but this thesis serves as a proof of concept that AutoML, using open-source data, can offer valuable support for decision-making processes within recruiting. I recommend that the Marine Corps and other branches of service further explore the use of AutoML and open-source data to enhance their recruitment strategies.

The Marine Corps should also have a database to upload and store open-source data for each high school. A database would improve the accessibility and organization of the data which will make further improvements and testing of AutoML models more efficient.

Finally, the use of AutoML should be extended to other manpower models in the Marine Corps. This could potentially lead to further efficiencies in manpower planning and structuring.

C. FURTHER RESEARCH

Areas for further research include:

- Relative Performance of RSs: I recommend that further research should be conducted to test how well the algorithm performs relative to the individual RS and RSS priority schools.

- Real-World Testing: Further testing should be done to see how well AutoML performs in the real world. One way to do this would be to assign control and test RSs, where the test RSs use the algorithm and the control continues to use their usual methods. Results would then be compared between the two groups.

- Model Refinement: Continuous improvement and refinement of the AutoML model should be ongoing. These refinements should consider feedback from real-world applications and incorporate additional data sources that may enhance the model's performance.

APPENDIX. SUMMARY STATISTICS AND RESULTS

The following tables have summary statistics for the training and test year data. The columns include the names of each variable, the count of non-missing observations, the mean value, standard deviation, minimum and maximum values, and the values at the 25th and 75th percentiles for each variable. Due to the large number of variables used in the model, I only included variables from MCRC and a selected number of CCD variables in these tables. The variables from the ACS are not included in the tables below. The ACS variables include population estimates and demographics such as gender breakdowns, age distributions, race and ethnicity percentages, and educational attainment. They also include household and economic indicators like income, employment status, occupations, industries, and poverty levels.

The reason for the gap in years from 2019 to 2022 is due to COVID. I wanted to use data more representative of a typical recruiting environment, so I chose not to test on these years.

Table 7. 2018 Training Data Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Number of Sr Year Contracts	16071	1.1	1.7	0	0	2	16
Schools with one or more contracts	16071	0.5	0.5	0	0	1	1
Number of Prior Year Contracts	16071	1.2	1.7	0	0	2	16
Number of Contracts from 2 Years Ago	16071	1.2	1.7	0	0	2	16
Priority Code	16071						
... 1	3698	23%					
... 2	2896	18%					
... 3	2784	17%					
... N	6693	42%					
Free.and.Reduced Lunch.Students.	16071	379	440	0	94	501	4397
Grades.9.12.Students	16071	839	769	13	232	1288	5664
Grade.11.Students.	16071	207	192	6	57	316	1594
Male.Students	16071	449	386	0	149	668	3324

Female.Students	16071	430	372	0	142	642	2437
Grade.11.Students. female.	16071	102	95	0	27	156	787
Grade.11.Students.. .male.	16071	105	98	0	28	160	807

Table 8. 2019 Training Data Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Number of Sr Year Contracts	16282	1.2	1.8	0	0	2	16
Schools with one or more contracts	16282	0.52	0.5	0	0	1	1
Number of Prior Ye ar Contracts	16282	1.1	1.7	0	0	2	16
Number of Contract s from 2 Years Ago	16282	1.2	1.7	0	0	2	16
Priority Code	16282						
... 1	3711	23%					
... 2	2909	18%					
... 3	2923	18%					
... N	6739	41%					
Free.and.Reduced Lunch.Students.	16282	384	440	0	96	511	4329
Grades.9.12.Studen ts	16282	834	770	12	232	1274	5839
Grade.11.Students.	16282	205	193	6	57	313	1650
Male.Students	16282	448	387	0	149	664	3478
Female.Students	16282	428	373	0	140	637	2572
Grade.11.Students. female.	16282	101	95	0	27	154	812
Grade.11.Students.. .male.	16282	104	98	0	28	158	851

Table 9. 2022 Test Data Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Number of Sr Year Contracts	16901	0.94	1.5	0	0	1	18
Schools with one or more contracts	16901	0.44	0.5	0	0	1	1
Number of Prior Ye ar Contracts	16901	1	1.5	0	0	2	17
Number of Contract s from 2 Years Ago	16901	0.98	1.5	0	0	1	15

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Priority Code	16901						
... 1	3621	21%					
... 2	2961	18%					
... 3	3065	18%					
... N	7254	43%					
Free.and.Reduced Lunch.Students.	16901	339	450	0	45	447	5421
Grades.9.12.Students	16901	817	770	11	221	1244	5943
Grade.11.Students.	16901	200	191	6	53	304	1702
Male.Students	16901	442	389	0	144	656	3542
Female.Students	16901	422	374	0	136	625	3215
Grade.11.Students. female.	16901	98	95	0	26	149	850
Grade.11.Students.. .male.	16901	102	98	0	27	154	899

Table 10. Results Using 2017 and 2018 as Training Data and 2019 as Test Data

Number of Schools	AutoML Average Contracts Per School	Marine Corps Average Contracts Per School	Percentage Difference	AutoML Success Rate	MCRC Success Rates
3700 (Priority 1)	3.056	2.734	11.7%	83.3%	78.9%
6620 (Priority 1& 2)	2.378	2.268	4.9%	63%	62.8%
9640 (Priority 1, 2, & 3)	1.911	1.899	0.67%	41.0%	45.4%

LIST OF REFERENCES

Bartz, Eva., Bartz-Beielstein, Thomas., Zaefferer, Martin., & Mersmann, Olaf. (2023). *Hyperparameter Tuning for Machine and Deep Learning with R : A Practical Guide.* (1st ed.). Springer. https://directory.doabooks.org/handle/20.500.12854/96206

Berger, D. (2022). Recruiting requires bold changes. U.S. Naval Institute. https://www.usni.org/magazines/proceedings/2022/november/recruiting-requires-bold-changes

De Bie, T., De Raedt, L., Hernández-Orallo, J., Hoos, H. H., Smyth, P., & Williams, C. K. I. (2022). *Automating data science*. https://doi.org/10.1145/3495256

Dong, G., & Liu, H. (2018). Feature Extraction and Learning for Visual Data Parag S. Chandakkar, Ragav Venkatesan, and Baoxin Li. In *Feature Engineering for Machine Learning and Data Analytics*. Taylor & Francis Group.

Fulton, B. M. (2016). Determining market categorization of United States ZIP codes for purposes of Army recruiting [Master's thesis, Naval Postgraduate School]. https://calhoun.nps.edu/handle/10945/49463

Glackin, & Adivar, M. (2023). Using the power of machine learning in sales research: process and potential. *The Journal of Personal Selling & Sales Management, 43*(3), 178–194. https://doi.org/10.1080/08853134.2022.2128812

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Assessing the fit of the model. In D.W. Hosmer, S. Lemeshow, & R.X. Sturdivant (Eds.), Applied Logistic Regression (3rd ed.). https://doi.org/10.1002/9781118548387.ch5

Headquarters U.S. Marine Corps (2015). Volume III, Guidebook for recruiting station operations, 2015 Edition.

Intrater, B. C. (2015). *Understanding the impact of socioeconomic factors on Navy accessions* [Master's thesis, Naval Postgraduate School]. https://calhoun.nps.edu/handle/10945/47279

Kleykamp. (2006). College, jobs, or the military? Enlistment during a time of war. *Social Science Quarterly*, *87*(2), 272–290. https://doi.org/10.1111/j.1540-6237.2006.00380.x

Kleykamp, M., Schwam, D., & Wenig, G. (2023). What Americans think about veterans and military service: Findings from a nationally representative survey. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA1363-7.html

Marmion, W. N. (2015). Evaluating and improving the SAMA (Segmentation Analysis and Market Assessment) recruiting model. [Master's thesis, Naval Postgraduate School]. https://calhoun.nps.edu/handle//10945/45894

Rezazadeh, A. (2020). A generalized flow for B2B sales predictive modeling: An Azure machine-learning approach. *Forecasting*, 2(3), 267–283. https://doi.org/10.3390/forecast2030015

Shahriyar, J., Ahmad, J. B., Zakaria, N. H., & Su, G. E. (2022). Enhancing prediction of employability of Students: Automated machine learning approach. *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 87–92. https://doi.org/10.1109/ICICyTA57421.2022.10038231

Solorio-Fernández, S., Ariel, C.-O. J., & Martínez-Trinidad, J. F. (2022). A survey on feature selection methods for mixed data. *The Artificial Intelligence Review, 55*(4), 2821–2846. https://doi.org/10.1007/s10462-021-10072-6

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Fort Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California