



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

**MACHINE LEARNING ANALYSIS OF THE ALCOHOL
AND DRUG MANAGEMENT INFORMATION
TRACKING SYSTEM**

by

Vanny Mae O. Angeles

September 2021

Thesis Advisor:
Second Reader:

Ruriko Yoshida
Samuel E. Buttrey

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE September 2021	3. REPORT TYPE AND DATES COVERED Master's thesis	
4. TITLE AND SUBTITLE MACHINE LEARNING ANALYSIS OF THE ALCOHOL AND DRUG MANAGEMENT INFORMATION TRACKING SYSTEM			5. FUNDING NUMBERS	
6. AUTHOR(S) Vanny Mae O. Angeles				
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) Substance abuse remains a significant problem among Sailors within the U.S. Excessive use of alcohol and drugs can be detrimental to a Sailor's health, safety, and naval service. The U.S. Navy recognizes that substance abuse is preventable and treatable and aims to eliminate these destructive behaviors through continuous training, intervention, and treatment. However, the need to address alcohol and drug abuse using an alternative approach arises as Sailors continue to exhibit these destructive behaviors. We propose the use of supervised machine-learning methods complemented by social network analysis to explore alcohol- or drug-incident data from the Alcohol and Drug Management Information Tracking System (ADMITS). Under our method, we build prediction, classification, forecasting, and generalized network autoregressive (GNAR) models to provide insights on the correlations between Sailors' backgrounds and their alcohol- or drug-related incident information. We utilize performance assessments for regression, classification, and time-dependent data to measure the accuracy of our models and to identify which perform best. Our results strongly demonstrate that the use of supervised machine-learning methods provide an accurate and effective approach to modeling and further understanding of the different aspects of substance abuse among Sailors.				
14. SUBJECT TERMS substance abuse, alcohol abuse, drug abuse, machine learning, supervised learning, predictive modeling, time series forecasting, network analysis, network time series, social network analysis			15. NUMBER OF PAGES 77	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

**MACHINE LEARNING ANALYSIS OF THE ALCOHOL AND DRUG
MANAGEMENT INFORMATION TRACKING SYSTEM**

Vanny Mae O. Angeles
Lieutenant, United States Navy
BS, Jacksonville University, 2014

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
September 2021**

Approved by: Ruriko Yoshida
Advisor

Samuel E. Buttrey
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Substance abuse remains a significant problem among Sailors within the U.S. Excessive use of alcohol and drugs can be detrimental to a Sailor's health, safety, and naval service. The U.S. Navy recognizes that substance abuse is preventable and treatable and aims to eliminate these destructive behaviors through continuous training, intervention, and treatment. However, the need to address alcohol and drug abuse using an alternative approach arises as Sailors continue to exhibit these destructive behaviors. We propose the use of supervised machine-learning methods complemented by social network analysis to explore alcohol- or drug-incident data from the Alcohol and Drug Management Information Tracking System (ADMITS). Under our method, we build prediction, classification, forecasting, and generalized network autoregressive (GNAR) models to provide insights on the correlations between Sailors' backgrounds and their alcohol- or drug-related incident information. We utilize performance assessments for regression, classification, and time-dependent data to measure the accuracy of our models and to identify which perform best. Our results strongly demonstrate that the use of supervised machine-learning methods provide an accurate and effective approach to modeling and further understanding of the different aspects of substance abuse among Sailors.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.2	Objective	3
1.3	Structure.	4
2	Background	5
2.1	Machine Learning Methods	5
2.2	Time Series	9
2.3	Networks	14
2.4	Performance Evaluation.	15
2.5	Literature Review	20
3	Models and Methodology	25
3.1	Methodology Overview.	25
3.2	Data	26
3.3	Model Fitting and Implementation	30
4	Results and Analysis	37
4.1	Model Results and Analysis	37
5	Conclusion	47
5.1	Summary	47
5.2	Future Work	48
	List of References	53
	Initial Distribution List	57

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

Figure 2.1	Time Series Components	10
Figure 2.2	Time Series Training and Testing Data Sets	11
Figure 2.3	Example of Seasonal Decomposition	12
Figure 2.4	Confusion Matrix for Binary Classification	17
Figure 2.5	Example an Ideal ROC Curve	18
Figure 2.6	Example of Rolling-Horizons Design	19
Figure 3.1	Overview of Thesis Methodology.	26
Figure 3.2	Summary of Missing Values in ADMITS Data Set	29
Figure 3.3	Regression and Classification Workflow	32
Figure 3.4	Time Series Workflow	34
Figure 3.5	GNAR Workflow	36
Figure 4.1	Classification Results	38
Figure 4.2	Variable Importance for Poisson Regression	39
Figure 4.3	Variable Importance for Logistic Regression	40
Figure 4.4	Seasonal Decomposition of ADMITS Data Set	41
Figure 4.5	Time Series Forecasting Results	42
Figure 4.6	Residuals from the $ARIMA(2, 1, 2)(1, 0, 1)_{12}$ Model	43
Figure 4.7	GNAR MAPE Results	45
Figure 4.8	GNAR MASE Results	46

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Table 2.1	MAPE Recommendations	20
Table 3.1	ADMITS Data Set Attributes	27
Table 4.1	Regression and Classification Model Summary	38
Table 4.2	Time Series Forecasting Model Summary	42
Table 4.3	GNAR Model Summary	44

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

ACF	Autocorrelation Function
ADMITS	Alcohol and Drug Management Information Tracking System
AIC	Aikake Information Criterion
ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under the (ROC) Curve
BIC	Bayesian Information Criterion
CNN	Convolutional Neural Network
DAR	Drug and Alcohol Report
DON	Department of the Navy
ETS	Error, Trend, Seasonal or Exponential Smoothing
FPR	False Positive Rate
GLM	Generalized Linear Model
GNAR	Generalized Network Autoregressive
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
ML	Machine Learning
MSE	Mean Squared Error
OPNAV	Office of the Chief of Naval Operations
OPNAVINST	Office of the Chief of Naval Operations Instruction

PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SARIMA	Seasonal Autoregressive Integrated Moving Average (ARIMA)
SARP	Substance Abuse Rehabilitation Program
SNA	Social Network Analysis
STL	Seasonal and Trend Decomposition using Loess
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
USN	U.S. Navy

Executive Summary

Alcohol and drug abuse continues to be a concern among Sailors within the U.S. Navy (USN). These destructive behaviors are inconsistent with USN core values and can be detrimental to a Sailor's health, safety, and naval service. Through the Office of the Chief of Naval Operations (OPNAV) N17 branch, the USN aims to eliminate substance abuse using continuous training, intervention, and treatment. A priority for the N17 organization is to be able to identify Sailors who exhibit at-risk behaviors of substance abuse and to understand the underlying causes of these destructive behaviors. Alcohol- and drug-related incident information is collected from commands through Drug and Alcohol Reports (DARs) and populated in a database called the Alcohol and Drug Management Information Tracking System (ADMITS).

In this thesis, we apply a combination of machine learning (ML) methods and social network analysis (SNA) to historical data collected from ADMITS. We build prediction, classification, forecasting, and generalized network autoregressive (GNAR) models to gain insights on the correlations between Sailors' backgrounds and their alcohol- or drug-related incident information. To identify which models perform best, we use the appropriate predictive performance measures for regression, classification, and time-dependent data. Our results demonstrate that supervised ML methods are an accurate and effective approach to modeling the different aspects of substance abuse, particularly among Sailors.

This study establishes a foundation that allows for subsequent application of ML methods in substance abuse research to gain additional insights on destructive behaviors among Sailors in the USN. Opportunities for future research extend to improving the models built in this study, the incorporation of these models into practical applications, and the exploration of more complex ML methods and supplementary data sources.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

First and foremost, I would like to thank Professor Ruriko Yoshida for all of her time, guidance, and patience throughout the thesis-writing process. This process would have been much more difficult without her, and I am especially grateful that I had the chance to work with someone so passionate and energetic about their field of expertise. I would also like to thank Professor Samuel Buttrey for his guidance, mentorship, and witty sense of humor that helped make this process a little less daunting.

Many thanks to my friends for their constant support and reinforcing encouragement, especially to those whom I have had the pleasure of meeting at the Naval Postgraduate School. This journey was difficult, but I will always be grateful that I had the opportunity to work alongside such an outstanding group of individuals.

Finally, to my family, a very heartfelt thank you for always believing in me. In particular, I must thank my husband, Kenneth, and our dog, Po, for their endless love, support, patience, and coffee runs. Achieving this goal would not have been possible without them, and I am forever grateful that they were by my side during this process.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:

Introduction

This chapter introduces the motivation behind this thesis and the organizational impact of substance abuse in the U.S. Navy (USN). The subsequent sections present the objectives this study aims to achieve, the significance of our research, and the organization for the remainder of this thesis.

1.1 Motivation

Alcohol and drug abuse is “detrimental to operational readiness and is inconsistent with USN core values and initiatives to promote personal excellence and healthy lifestyles among Navy members” (Office of the Chief of Naval Operations [OCNO] 2009, p. 3). According to OPNAVINST 5350.4D (2009, p. 3), alcohol and drug abuse by USN members “can seriously damage their physical and mental health, jeopardize their safety and the safety of others, and can lead to criminal prosecution and separation from naval service.” Members who choose to consume alcoholic beverages must drink responsibly and abide by the minimum legal drinking age. In addition, the USN strictly enforces a zero-tolerance drug policy. Navy members shall “never wrongfully possess, distribute or abuse drugs, be in possession of drug abuse paraphernalia, or be under the unauthorized influence of prescribed drugs” (OCNO 2009, p. 5).

The Department of the Navy (DON) recognizes that alcohol and drug abuse are preventable and treatable conditions (OCNO 2009). Screening, referral, early intervention, treatment, and continuing care services for alcohol and drug abuse are services available to all USN members. On June 2013, the DON established the Office of the Chief of Naval Operations (OPNAV) N17 branch, also known as the 21st Century Sailor Office (Vice Chief of Naval Operations [VCNO] 2013). OPNAV N17 comprises of several departments that aim to establish a more efficient way of providing USN members support regarding Sailor resiliency and readiness programs, such as substance abuse prevention, suicide prevention, and the Substance Abuse Rehabilitation Program (SARP) (VCNO 2013).

A priority for the N17 organization, among many others, is to be able to identify Sailors who

exhibit at-risk behaviors of substance abuse and to understand the underlying causes of these destructive behaviors. OPNAV N17 is able to monitor alcohol and drug abuse and misuse Navy-wide using a database known as the Alcohol and Drug Management Information Tracking System (ADMITS). ADMITS is the “Navy and Marine Corps repository for alcohol and drug incident, screening, treatment, and training information” (OCNO 2021, para. 1). USN commands submit Drug and Alcohol Reports (DARs) that feed Sailor and incident information into ADMITS. The database supports the USN’s ongoing approach to eliminating alcohol and drug abuse by providing statistical reports regarding the efficacy of current substance abuse programs and the urinalysis testing policy (OCNO 2021).

The USN’s ongoing approach to eliminating substance abuse consist of “enhanced detection, deterrence, prevention, education, intervention, and treatment when necessary” (OCNO 2009, p. 2). While all USN personnel are educated about substance abuse policies, programs, and resources, data released by the RAND Corporation in 2015 as part of a health-related behavioral study revealed that 34.2 percent of Sailors reported binge drinking within the past 30 days, the second highest percentage among the military services (Harkins 2018; OCNO 2009). As a result, an alternative approach is needed to aid in identifying the potential risk factors contributing to destructive behaviors. Barenholtz et al. (2020) present several substance abuse studies that utilize machine learning (ML) techniques and have shown promising results. The growing application of ML methods in substance abuse research among civilian populations motivates the study of these techniques to assist the 21st Century Sailor Office in identifying, understanding, and predicting addictive behavior among Sailors.

Substance abuse “continues to accelerate towards becoming the most severe public health problem in the United States,” and identifying their determinants is an urgent priority (Hu et al. 2019, p. 1). In clinical studies, exposure to various forms of stress has been identified as one of many risk factors for alcohol and drug abuse (Uhart and Wand 2008). Social interactions have also been recognized as a risk factor as “most substance-use researchers would agree that people who misuse substances are often surrounded by friends, family members, and associates that also misuse substances or tacitly approve of doing so” (Valente et al. 2004, p. 1693).

More recently, applications of ML methods in substance abuse research have emerged as an effective approach to studying substance-related incidents and the risk behaviors associated with them (Hu et al. 2019). Specifically, social networks, time series forecasting, and other supervised ML algorithms have been leveraged to predict alcohol- or drug-related incidents and identify key attributes that potentially influence these destructive behaviors (Hu et al. 2019). The ability to identify, predict (or classify), and forecast substance use and risk behavior among a group of individuals can help with monitoring the trend of alcohol- or drug-related incidents. Forecasting and monitoring are particularly important in identifying individuals at risk for substance abuse and informing policy changes within organizations.

1.2 Objective

In this thesis, we apply a combination of supervised ML methods and social network analysis (SNA) to historical data obtained from ADMITS to gain an improved understanding of the potential risk factors contributing to destructive behaviors among Sailors within the USN. In particular, the following research questions are addressed:

1. Using ML methods, with what degree of certainty can we predict, classify, or forecast
 - whether or not a Sailor will repeat an alcohol- or drug-related incident?
 - the number of alcohol- or drug-related incidents committed by a particular Sailor?
 - the total number of monthly alcohol- or drug-related incidents for a given period of time?
2. Using a social network time series structure, with what degree of certainty can we predict when a particular Sailor will commit and/or repeat an alcohol- or drug-related incident?
3. Through the implementation of ML models and feature importance, which attributes (or features) are most useful in classifying whether or not a Sailor will repeat an alcohol- or drug-related incident?

Currently, substance abuse studies that apply ML methods have only analyzed destructive behaviors among a civilian group or population. It remains a challenge to isolate specific risk factors, social interactions, or stressors that influence or possibly trigger destructive behaviors in USN personnel. This study aims to understand these behaviors to better inform

potential policy changes within the DON and provide a foundation for further application of ML methods in substance use and abuse research.

1.3 Structure

The subsequent chapters in this thesis develop the data science problem and present an analytical framework for research and analysis. Chapter 2 provides ML and networks background information and reviews relevant literature intended to compare studies' methodologies, results, and relevance to the research questions in this thesis. Chapter 3 describes the ADMITS data and the prediction, classification, forecasting, and generalized network autoregressive (GNAR) models used in this study. Chapter 4 presents the results and deductions from our analysis. Finally, Chapter 5 summarizes key findings, recommendations, and areas for continued research.

CHAPTER 2: Background

This chapter begins by providing an overview of the three classes of ML methods and the algorithmic modeling workflow adapted in this study. Next, Section 2.2 describes the basic terminology of time series analysis and the various time series models and forecasting algorithms. Section 2.3 describes the basic terminology of network analysis and the four classes of networks. Section 2.4 explains the performance assessments used to measure the accuracy of our prediction, classification, forecasting, and GNAR models. Lastly, Section 2.5 reviews and analyzes the similarities and differences of studies that apply ML methods in substance abuse research. Moreover, we describe how their methodologies and conclusions motivate the approach used in this research to gain insights on the correlations between Sailors’ backgrounds and their alcohol- or drug-related incident information.

2.1 Machine Learning Methods

ML is an application of data analysis that learns relations and patterns using computational statistical model building. ML methods are often partitioned into three broad classes: supervised, unsupervised, and reinforcement learning. There tends to be overlap between these classes, and ML methods are typically combined to solve analytic problems. As a general rule, “supervised learning methods are designed to be predictive, unsupervised learning methods are designed to be descriptive, and reinforcement learning methods are designed to be prescriptive” (Brown and Huddleston 2018, p. 239).

Supervised learning uses predictor variables—often referred to as features, explanatory variables, or attributes—to predict at least one response variable associated with new observations. Conversely, unsupervised learning identifies patterns and hidden structures in data sets that are neither classified or labeled. As a result, it requires no response variable. Reinforcement learning seeks to identify what actions or behaviors maximize reward by evaluating trial-and-error interactions.

In this thesis, we only consider supervised ML algorithms to develop prediction, classification, and forecasting statistical models. The subsequent sections of this thesis provide

brief summaries of the supervised ML algorithms applied in this study. Methods described include:

- Logistic Regression
- Poisson Regression
- Support Vector Machine (SVM)
- Random Forest
- Seasonal Decomposition
- Time Series Forecasting
- GNAR Processes

2.1.1 Algorithmic Modeling Overview

This study utilizes an adaptation of the algorithmic modeling overview described by Brown and Huddleston (2018). The authors present the following steps to outline a standard workflow for a supervised learning problem:

1. **Data Acquisition and Cleaning** – The first step in the process is to acquire and clean the data. Considerable time is spent cleaning, formatting, and standardizing any data acquired. In addition, it is important for data scientists to communicate with subject-matter experts to ensure that the data acquired is relevant to the problem(s) being addressed.
2. **Feature Engineering and Scaling** – Feature engineering involves constructing new attributes from the existing data through interaction variables or variable transformations. Feature scaling, or data normalization, is the process of adjusting all attributes to a standard scale. The objective of this step is “to generate attributes that provide predictive power for a particular ML algorithm” (Brown and Huddleston 2018, p. 237).
3. **Model Fitting (Training) and Feature Selection** – Model fitting comprises of setting all of the parameters of a model by using the appropriate training data set for each ML algorithm. Feature selection methods include the stepwise algorithm—which measures a model’s goodness-of-fit using statistics such as the Aikake Information Criterion (AIC) or Bayesian Information Criterion (BIC)—and cross-validation. In ML, the idea of overfitting or underfitting the data arises when fitting models. Over-

fitted models are the result of an algorithmic approach that is too sensitive to the observations in the data set used to train them (Yoshida 2020a). Moreover, they miss the general trends in the data in favor of closely matching the training observations (Yoshida 2020a). If the model is underfitted, the model has not learned enough from the training data, resulting in unreliable predictions (Al-Masri 2019). Overfitting and underfitting negatively impacts any model's ability to be generalized to new data. In this process, the outcome of the model fitting and feature selection step is a set of models that will be compared for final selection.

4. **Model Selection** – In this step, predictive performances for each model are compared based on their predictive ability using the validation data set. The model that provides the best predictive performance is identified as the best model. Of note, the trade-off between predictive performance and model interpretability may have to be considered.
5. **Model Performance Assessment** – A new predictive model is generated by applying the best performing algorithm to the combined training and validation data set. This model is used to make predictions on the testing data set and gives an approximation of how well it generalizes and performs on unseen data.
6. **Model Implementation** – The last step in the process is to apply the final model to the full data set. This step also involves the development of visualizations, reports, and procedures so that the model can be applied or reproduced.

2.1.2 Supervised Machine Learning Algorithms

Supervised ML is further sub-categorized into two categories: regression and classification. In regression, the response variables take on quantitative values, continuous or integer. Response variables in classification, however, are categorical.

Logistic Regression

Logistic regression is a classification algorithm that aims to identify decision boundaries among the different classes, or categories, of the response variable (Gudivada et al. 2016). The parameters of a logistic regression model are estimates of the weights for the features of the data set, where “each weighted feature vector is mapped to a value between 0 and 1 using the S-shaped logistic function” (Gudivada et al. 2016, p. 183). This value, which represents the probability of an observation belonging to a specific category, is tuned to effectively classify the observations in a training data set (Gudivada et al. 2016).

Poisson Regression

Poisson regression is a type of generalized linear model (GLM) that models count data. This type of regression is often applied when the errors of the data are not normally distributed (Lord et al. 2015). In Poisson regression, Lord et al. (2015) states, the count of the response variable is modeled as a Poisson random variable, with the probability of occurrence being:

$$P(y_i) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}, \quad (2.1)$$

where y_i is the count for one group or class i , λ is the mean count over a certain time period, and e is the base of the natural logarithm.

SVM

SVMs can be used for both regression and classification problems. An SVM tries to “find a plane or line that can separate the data into different classes” (Burger 2018, p. 39). In doing so, the algorithm uses vectors that maximize these separations. According to Burger (2018), SVMs work by employing the “kernel trick.” Under this method, “data can be transformed where a decision boundary is to be drawn, and then hyperplane separation is applied on the transformed data” (Burger 2018, p. 173). The method allows for curves to be drawn around the data, resulting in a better model fit (Burger 2018).

The downside, Burger (2018) states, comes with the interpretability of an SVM model. While it is easy to “pass data through an SVM and get a meaningful output, describing the process by which the transformations occur can often be swept under the moniker of a ‘black box’ operation” (Burger 2018, p. 176).

Random Forest

A random forest is a type of tree-based model that can be used in both regression and classification problems. Like SVM models, a random forest is also a type of “black box” model due to the difficulty in interpreting the complex interactions of the algorithm. A single decision tree works using the “if-then” logic of “starting with a feature, splitting based on a value of ranges in that feature, and then moving down the tree to a final result” (Burger 2018, p. 155). According to Burger (2018, p. 135), random forests “can be complicated mathematically, but generally boil down to a collection of different tree models being asked

to vote on a result.”

In a random forest model, instead of growing a single tree, N different trees are grown by randomizing outputs to the algorithm. Each tree will produce an output based on the feature splits in the data. The output for the model is selected by identifying which output was produced the most (Burger 2018).

2.2 Time Series

A time series is a set of observations indexed by time. Time series data is encountered in many fields, including engineering, science, sociology, and economics. In time series analysis, the objectives are to draw inferences from the sequence of observations by building a statistical model to represent the data and to forecast future observations (Brockwell and Davis 2016).

In this study, we consider several time series models and forecasting algorithms. Section 2.2.2 provides brief summaries of the time series and forecasting models applied in this thesis, which include:

- Seasonal Decomposition
- Naïve and Seasonal Naïve
- Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA)
- Error, Trend, Seasonal or Exponential Smoothing (ETS)
- Holt-Winters

2.2.1 Definitions

Since time series data can exhibit underlying patterns and behaviors, it is useful to decompose the data into four components. Yoshida describes these components, illustrated in Figure 2.1, as:

- **trend**, long-term (not necessarily linear) increases or decreases in the data,
- **seasonal (periodic)**, increases or decreases in data with a fixed/known period,
- **cycle**, rises and falls that are not of fixed period, and

- **noise**, the remaining variance in the data after we've accounted for the components above. (Yoshida 2020b)

Furthermore, Yoshida describes the different properties of time series data as they can vary consistently or irregularly over time. She defines these properties as:

- **Homoscedastic**, observed variance remains the same over time,
- **Heteroscedastic**, observed variance changes over time, and
- **Stationarity**, [a stationary time series is] one whose properties do not depend on the time when the series is observed. A stationary time series can exhibit cyclic behavior, but any series with trend, seasonality, or heteroscedasticity is non-stationary. (Yoshida 2020b)

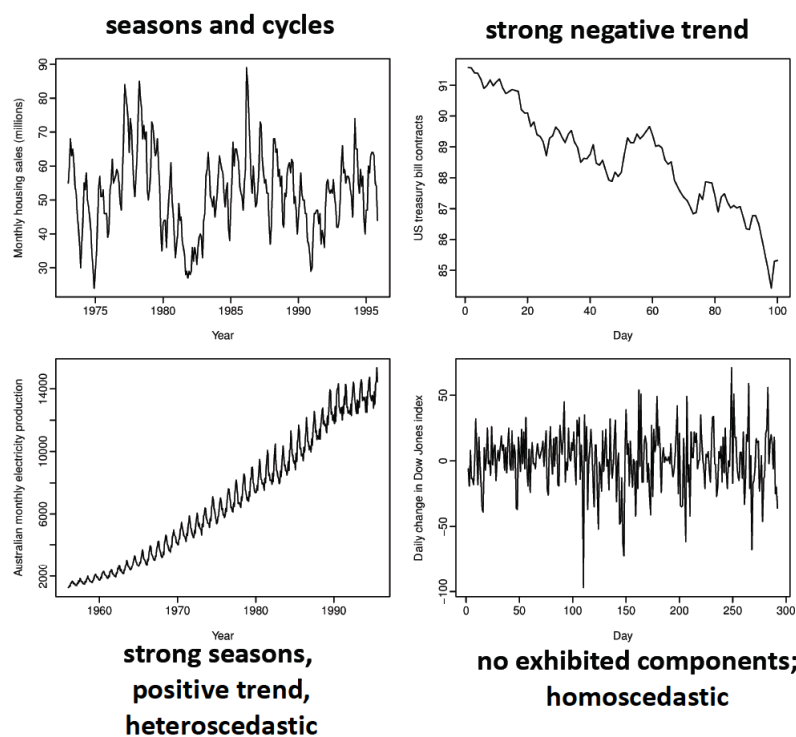


Figure 2.1. **Time Series Components.** The time series data in the upper-left exhibits seasons and cycles. The upper-right exhibits a strong, negative trend over time. The bottom-left exhibits a strong seasonality with a positive trend with heteroscedasticity over time. The bottom-right exhibits none of these components but shows homoscedasticity. Source: Yoshida (2020b).

2.2.2 Time Series Models and Forecasting Algorithms

Time series models and forecasting algorithms can be simple or highly complex. According to Hyndman and Athanasopoulos (2018, “What can be Forecast?” section), “good forecasts capture the genuine patterns and relationships which exist in historical data, but do not replicate past events that will not occur again.” In choosing and fitting time series models, Hyndman and Athanasopoulos (2018, “The Basic Steps in a Forecasting Task” section) state that “the best model to use depends on the availability of historical data, the strength of the relationships between the forecast variable and explanatory variables, and the way in which the forecasts are to be used.”

When fitting time series forecasting models, it is common to compare several potential models. Similar to data in algorithmic modeling, time series data sets are often partitioned into training and testing data sets to validate a model’s forecasting accuracy. The difference is that instead of observations being split into random subsets, the training data set of time series data consists of observations up to a specific point in time. The remaining data of the time series then becomes the testing data set. Figure 2.2 demonstrates the division of a time series data set into training and testing subsets.



Figure 2.2. **Time Series Training and Testing Data Sets.** In time series analysis, model parameters are estimated from training data, and model performance is evaluated against testing data. Source: Hyndman and Athanasopoulos (2018).

Seasonal Decomposition

Seasonal decomposition attempts to extract the trend-cycle, seasonal, and remainder components of time series data. There are various types of time series decomposition models, but one commonly used is seasonal and trend decomposition using Loess (STL) due to its modeling capabilities (Hyndman and Athanasopoulos 2018). According to Hyndman and Athanasopoulos (2018), an advantage of STL over other seasonal decomposition models is its ability to handle any type of seasonality and fluctuations in the seasonal component of time series data. In this study, note that we only utilize an STL time series model to identify

any potential seasonal or trend patterns and did not apply this model to forecasting. Figure 2.3 illustrates an example of time series seasonal decomposition.

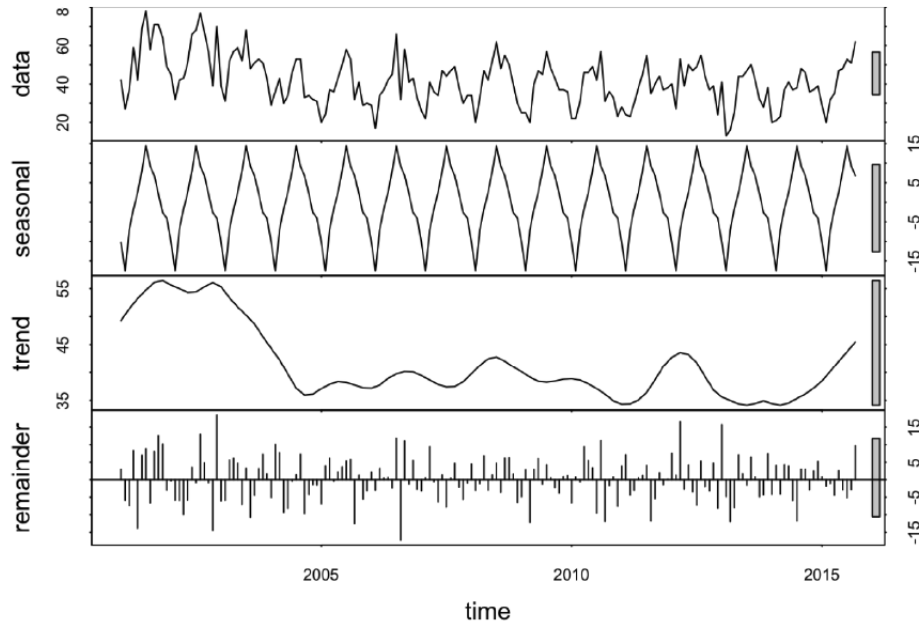


Figure 2.3. **Example of Seasonal Decomposition of Time Series.** Seasonal decomposition separates time series data into three components: trend, seasonal, and remainder. These components can be added together to reconstruct the data shown in the top panel. Source: Yoshida (2020b).

Naïve and Seasonal Naïve

Naïve models are often used as a benchmark, or baseline, model. A simple naïve model sets the forecast value to be the value of the last observation (Yoshida 2020b). To account for seasonal data, a seasonal naïve model can be used. In a seasonal naïve model, “each forecast value is set to be equal to the last observed value from the same season of the year, i.e., for a monthly forecast, the same month of the previous year is used” (Yoshida 2020b, “Key TS Ideas: Naïve Models” section).

ARIMA and SARIMA

ARIMA models aim to describe the autocorrelations in time series data (Hyndman and Athanasopoulos 2018). Non-seasonal ARIMA models are commonly designated $ARIMA(p, d, q)$ as defined by Hyndman and Anthanasopoulos, where:

- p is the order of the autoregressive part,
- q is the degree of first differencing involved, and
- d is the order of the moving average part. (Hyndman and Athanasopoulos 2018)

An autocorrelation function (ACF) is the coefficient of correlation between two values in a time series that measures the linear relationship between lagged values in a time series (Yoshida 2020b). This measure, along with the partial autocorrelation function (PACF), are used to “identify the order of the autoregressive (p) and moving average (q) components for the ARIMA model” (Yoshida 2020b, “Key Ideas: Time Series Auto-Regression” section). ARIMA models are also capable of modeling seasonal data. A seasonal ARIMA (SARIMA) model is comprised of seasonal terms that are added to the ARIMA model and is designated by $ARIMA(p, d, q)(P, D, Q)_m$, where m is the number of observations per year and P , D , and Q are the seasonal variants of p , d , and q (Hyndman and Athanasopoulos 2018).

ETS

Exponential smoothing, or ETS, methods forecast by weighing averages of past observations, where the weights decay exponentially over time (Hyndman and Athanasopoulos 2018). In ETS models, each model “consists of a measurement equation that describes the observed data, and some state equation that describe how the unobserved components or states (level, trend, seasonal) change over time” (Hyndman and Athanasopoulos 2018, “Innovations State Space Models for Exponential Smoothing” section). These are referred to as state space models, which are labeled as $ETS(\cdot, \cdot, \cdot)$ for (Error, Trend, Seasonal) (Hyndman and Athanasopoulos 2018). These models consider different combinations of the error, trend, and seasonal components. The possibilities for each component as defined by Hyndman and Athanasopoulos (2018) are:

- **Error** = $\{A, M, Z\}$
- **Trend** = $\{N, A, M, Z\}$
- **Seasonal** = $\{N, A, M, Z\}$

In all cases, N is none, A is additive, M is multiplicative, and Z is automatically selected (Hyndman and Athanasopoulos 2018).

Holt-Winters

Holt-Winters models apply an exponential smoothing technique to the trend, seasonality,

and cycle components of a time series. Holt’s method was expanded to capture seasonality, resulting in the Holt-Winters seasonal method, which comprises of “the forecast equation and three smoothing equations with corresponding smoothing parameters” (Hyndman and Athanasopoulos 2018, “Holt-Winters’ Seasonal Method” section).

Seasonality within the data can be modeled by either an additive or multiplicative method. According to Hyndman and Athanasopoulos (2018, “Holt-Winters’ Seasonal Method” section), “the additive method is preferred when the seasonal variations are roughly constant through the series, while the multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series.”

2.3 Networks

The field of network science has become increasingly relevant in computer science, biology, statistics, and others, as networks are now more often being used to model systems in these disciplines. Patterns of relationships between variables within the network structure can be derived through network analysis techniques. In this thesis, we focus on incorporating social networks and time series data to represent connections between Sailors and their associated alcohol- and drug-incident information.

2.3.1 Definitions

In the study of networks, there are several terms that need to be introduced, as usages differ depending on the field. This research applies the following definitions described by Newman (2003) and Valente et al. (2004):

- **Network** – A set of nodes joined by edges. Systems taking the form of a network can also be defined as a graph.
- **Node** – A point where a connection or interaction is established. Nodes, also known as vertices, may have a variety of properties associated with them.
- **Edge** – The connection or interaction between two nodes. Like nodes, edges may have a variety of properties associated with them, e.g., weights.
- **Degree** – The number of edges that connect to a node.
- **Adjacency Matrix** – A square matrix that describes a network or graph. The elements of the matrix represent whether pairs of nodes are connected by an edge. Entries can

also represent the weight of an edge.

2.3.2 Social Networks

Newman (2018) divides the most commonly studied networks into four general classes: social, information, technological, and biological networks. Similar to ML classes, network classes experience overlap, meaning some networks belong to more than one class. In this study, we utilize social network structures to depict connectedness among Sailors—where a node represents a Sailor, and an edge represents the connection between two Sailors.

Social networks model “a set of people or groups of people with some pattern of contacts or interactions between them” (Newman 2003, p. 174). In social networks, nodes represent people and edges represent their connections. Examples of this type of network include friendships between individuals and professional relationships. SNA goes beyond social media connections through online platforms such as Facebook and Twitter and examines patterns between family, professional, friendship, and acquaintance interactions.

2.3.3 GNAR Processes

The GNAR model and subsequent R package developed by Knight et al. (2020a, p. 1) “fits, predicts, and simulates from a powerful new class of generalized network autoregressive processes.” These processes include a multivariate time series and a network that contains correlations between nodes through network edges and their associated weights. To forecast future observations, GNAR models utilize the autoregressive nature of the multivariate time series and compares values of a given node and time to its previous values and values of its neighboring nodes (Knight et al. 2020a). In addition, GNAR processes offer much modeling flexibility as it is able to model different types of networks and network structures, and fit models where time series data is missing (Knight et al. 2020a).

2.4 Performance Evaluation

Many different predictive performance measures exist, so it is important to utilize the appropriate performance functions based on the type of data being predicted. In this study, we consider several performance measures for regression, classification, and time-dependent data. Performance functions described in this section include:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Receiver Operating Characteristic (ROC) Curve
- Area Under the (ROC) Curve (AUC)
- Mean Absolute Percentage Error (MAPE)
- Mean Absolute Scaled Error (MASE)

Regression Performance Evaluation

In regression, the response variable is numerical. Accordingly, predictive performance is measured by calculating the residual error, or the difference between observed and predicted values, for that particular observation. The regression performance functions used in this study are the MSE and RMSE. As described by Brown and Huddleston (2018), MSE is defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (2.2)$$

where n is the sample size, y_i is the response variable of the i^{th} observation, and $\hat{y}_i$ is the estimate of y_i using a given model. It follows that RMSE is defined as

$$\text{RMSE} = \sqrt{\text{MSE}}. \quad (2.3)$$

Classification Performance Evaluation

In classification, the response variable is categorical. While there exists several classification algorithms, often called classifiers, binary classifiers are most commonly used. Response variables take on values of 1 or 0 to indicate the probability of a “positive” or “negative” case, respectively. Performance measures are derived from values from a confusion matrix, a table that records the counts for true positives, true negatives, false positives, and false negatives for a given binary classifier at a specified threshold. Figure 2.4 depicts the confusion matrix for binary classification.

		Actual Class	
		Positive (1)	Negative (0)
Predicted Class	Positive (1)	True Positives	False Positives
	Negative (0)	False Negatives	True Negatives

Figure 2.4. **Confusion Matrix for Binary Classification.** A confusion matrix for binary classification records the counts of occurrence for each of the four outcomes for prediction in classification. Adapted from: Brown and Huddleston (2018, p. 250).

Two types of errors associated with binary classification problems arise from the confusion matrix: false positive errors, or type I errors, and false negative errors, or type II errors. In addition, two accuracy measures derived from the confusion matrix are sensitivity and specificity. Sensitivity, also known as the true positive rate (TPR), measures a classifier's ability to correctly identify true positives and is calculated as

$$\text{Sensitivity} = \text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \quad (2.4)$$

Conversely, specificity, also known as the true negative rate (TNR), measures a classifier's ability to correctly classify true negatives and is calculated as

$$\text{Specificity} = \text{TNR} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}. \quad (2.5)$$

ROC curves are “another tool used to evaluate the performance of binary classifiers because they allow investigation of sensitivity and specificity over the full range of possible thresholds” (Brown and Huddleston 2018, p. 250). ROC curves plot a classifier's TPR against its false positive rate (FPR) at all classification thresholds. The false positive rate (FPR) is calculated as

$$\text{FPR} = 1 - \text{TPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}. \quad (2.6)$$

The overall performance of a classifier, summarized over all possible thresholds, is measured by the AUC. An ideal ROC curve will hug the top left corner, as depicted in Figure 2.5, indicating a high TPR and low FPR (James et al. 2013). Therefore, the larger the AUC, the better the performance of the classifier.

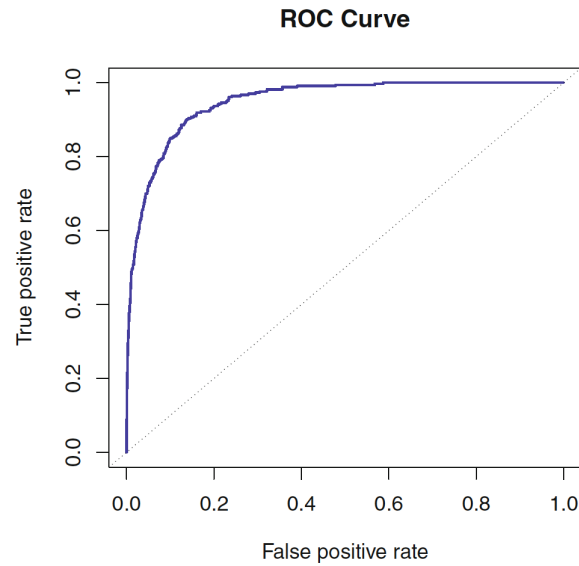


Figure 2.5. **Example of an Ideal ROC Curve.** An ideal ROC curve hugs the top left corner, indicating a high TPR and low FPR. Source: James et al. (2013, p. 147).

Performance Evaluation for Time-Dependent Data

When factors affecting prediction exhibit time dependency, model development and performance assessments used to measure prediction accuracy need to account for these changes. This type of performance evaluation is used in time series forecasting and many other regression or classification problems. While traditional statistical performance measures such as RMSE and adjusted R^2 can still be calculated, they are often not scale-free and are generally inferior (Guo et al. 2012). Often, a rolling-horizons design, used in concurrence with the MAPE and MASE, is employed for model fitting and performance evaluation when

there is time-dependent data.

In a rolling-horizons design, a forecasting model is repeatedly fit to a designated rolling period that increments forward in time at each iteration. In this process, the model learns from each training period to forecast observations in a future testing period (Brown and Huddleston 2018). Figure 2.6 illustrates a rolling-horizons design that estimates model performance by averaging performance over four periods (Brown and Huddleston 2018).

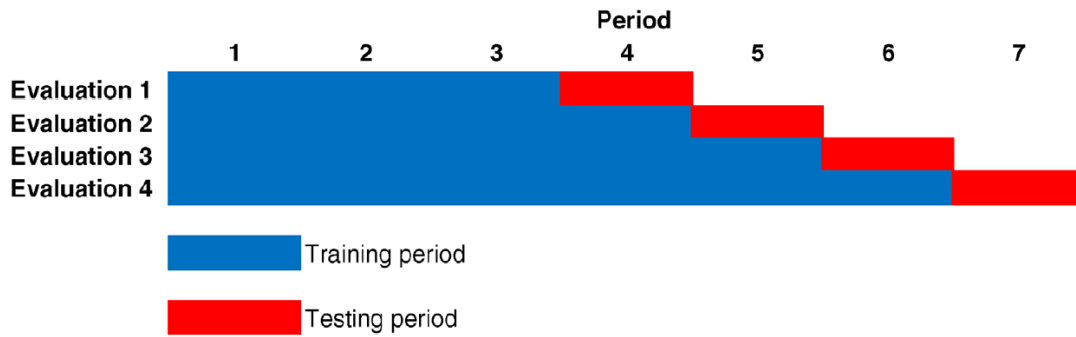


Figure 2.6. **Example of Rolling-Horizons Design.** Model performance is estimated by averaging performance over periods four through seven. Source: Brown and Huddleston (2018, p. 254).

Two performance statistics that can be computed from a rolling-horizons design are the MAPE and MASE. MAPE measures the accuracy of a forecasting algorithm by calculating the average absolute percent error for each time period minus the actual values divided by the actual values (Glen 2021). The formula for MAPE is defined as

$$\text{MAPE} = \frac{\sum_{t=1}^N \left| \frac{y_t - \hat{f}_t}{y_t} \right|}{N}, \quad (2.7)$$

where y_t is the observed value at time t , $\hat{f}_t$ is the forecast value produced by the model at time t , and N is the number of observations in the time series (Hornbuckle 2020). MASE, in comparison, measures the accuracy of a forecasting algorithm by comparing the absolute error of the forecast values and the baseline's average error (Glen 2019). In this study, we utilize the one-step naïve forecasting model as the baseline model. The formula for MASE

is defined as

$$\text{MASE} = \frac{\sum_{t=1}^N \left| \frac{y_t - f_t}{y_t - y_{t-1}} \right|}{N}, \quad (2.8)$$

where y_t is the observed value at time t , f_t is the forecast value produced by the model at time t , y_{t-1} is the observed value from the previous time period $t - 1$, and N is the number of observations in the time series (Hornbuckle 2020).

For a model to have a strong forecasting capability, it must surpass several standards. First, a model's MASE needs to exceed the accuracy of its comparative naïve forecasting model. Second, a model's MAPE needs to be at a minimum of below 50% as shown in Table 2.1 (Lewis 1982). It is also important to consider MAPE and MASE values that are closer to zero as they represent more accurate models (Hyndman and Athanasopoulos 2018).

Table 2.1. **MAPE Recommendations.** Guidance and recommendations for interpreting a given time series' MAPE. Adapted from: Lewis (1982).

MAPE	Interpretation
< 10	Highly accurate forecasting
10 - 20	Good forecasting
20 - 50	Reasonable forecasting
> 50	Inaccurate forecasting

2.5 Literature Review

The implementation of supervised ML methods and techniques to alcohol- and drug-abuse data has recently emerged as an effective approach to substance abuse research. Barenholtz et al. (2020) assert that the application of ML models in substance abuse research is likely to be successful in the future, as evidenced by several studies showing high predictive accuracy in their models. This literature review examines the relevancy of current substance abuse studies that apply ML methods to this thesis' research objectives. Consequently, we analyze

key similarities and differences in their methodologies and results to build a foundation for the approach of this study.

As presented by Barenholtz et al. (2020, “Recent Findings” section), recent studies have applied ML methods and techniques to substance abuse research for various predictive applications including “current abuse, assessing future risk, and predicting treatment success.” Islam et al. (2021) administered 36 questionnaire items to a group ($n = 486$) from the young urban population of Dhaka, Bangladesh, and trained models to predict an individual’s present vulnerability towards substance abuse. Out of six different ML algorithms, the logistic regression classifier performed best in distinguishing between healthy and addicted classes ($AUC = 0.98$) (Islam et al. 2021). In another study that utilized a similar data collection method, Jing et al. (2020) aimed to predict adolescents at risk for developing substance abuse. Under their method, they administered 1,000 questionnaire items to children and their parents ($n = 700$) at the ages of 10–12, 12–14, 16, 19, and 22. At each stage, seven ML algorithms were used, and the random forest algorithm was found to be the most effective model ($AUC = 0.74 - 0.86$) (Jing et al. 2020).

Implementing an alternative data collection approach, Afzali et al. (2019) aimed to predict alcohol use in mid-adolescence by analyzing samples obtained from the Canadian Co-Venture cohort ($n = 3826$) and the Australian Climate Schools and Preventure cohort ($n = 2190$). Their results showed that, out of the seven ML algorithms employed, the elastic-net regression model performed best on both the Canadian ($AUC = 0.869$) and Australian ($AUC = 0.855$) samples (Afzali et al. 2019). Park et al. (2021) pursued a different approach to substance abuse research and analyzed data from outpatient treatment centers with the aim to effectively predict the risk of patients dropping out. Six models—logistic regression, SVM, k-nearest neighbor (KNN), random forest, neural network, and adaptive boosting (AdaBoost)—were implemented, and AdaBoost was selected due to its prediction performance ($AUC = 0.72$) (Park et al. 2021). Additionally, the study applied feature importance and identified four variables that had the most effect in model prediction: length of hospitalization, age, residential area, and diabetes (Park et al. 2021).

The key similarity found within these studies was their common implementation of the ML process. In all cases, these studies implemented a similar process when conducting their research. This suggests that this methodology is widely accepted in the field of ML.

Each study collected data for its analysis, either through a database or questionnaires. Then, the studies utilized various methods of feature selection—such as literature reviews or correspondence with professionals—and built models using multiple algorithms. Models with the best performance were identified and selected by their ROC and AUC measures.

The differences in these studies include their objective—what they aimed to achieve with their models—their data collection methods, and the types of ML algorithms implemented in their research. Each study aimed to predict (or classify) something different within substance research. Additionally, their methods of collecting data varied. Two of the four studies acquired data by use of questionnaires, while the other two studies obtained cohort and medical data. While these studies used different ML algorithms to build their models, there was a general commonality in what algorithms were used. The most common ML algorithms applied were logistic regression, random forest, SVM, KNN, AdaBoost, and naïve Bayes.

The availability of data in building ML models is key. A general trend observed among these studies was the use of smaller data sets in their model fitting processes. While the data acquisition process is time-consuming and complex, as seen in the study conducted by Jing et al. (2020), uneven results are a likely outcome of developing models with a data set with limited observations. In particular, Jing et al. (2020, p. 5) noted that “the standard deviations of the accuracy across the 10-fold cross-validation were large, indicating that their prediction accuracy can be improved by using a larger data set with a more balanced distribution.” According to Barenholtz et al. (2020), future models will likely benefit from data sets that are much larger.

The foundation of our research incorporates different aspects of the studies analyzed and seeks to address the shortcomings of the studies aforementioned. We utilize a similar methodology in which we obtain data from a database, develop and train ML models, measure their predictive performances, and compare predictive performance measures to identify the best models. Our processes differ with the addition of time series data and network structures, and the application of these models to Sailors in the USN rather than a civilian population. Attributes of Sailors provide a different perspective on risk factors that could potentially contribute to substance abuse and destructive behaviors. Additionally, we use a much larger data set that incorporates historical alcohol or drug incident data for our

model fitting process. This study incorporates time series forecasting algorithms and social network analysis structures in our analysis, which provide additional insights in substance abuse research, particularly among Sailors.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:

Models and Methodology

This chapter presents our thesis methodology and describes the ADMITS data set and its properties and limitations. Section 3.3 explains the model fitting and implementation process for each supervised ML algorithm applied in this study.

3.1 Methodology Overview

The purpose of ML is to develop an approach to discover patterns in data and make predictions based on these patterns to answer research questions. Our methodology incorporates supervised ML methods complemented by SNA to explore historical data from the ADMITS database. By taking this approach, we build prediction, classification, forecasting, and GNAR models to provide insights on the correlations between Sailors' backgrounds and their alcohol- or drug-related incident information.

This thesis utilizes R and various R packages to execute the steps in our methodology (R Core Team 2019). R is “both a statistical environment and programming language that is widely used in both commercial and academic settings” (Buttrey and Whitaker 2016, p. 2). It allows us the ability to “read, manage, and clean data; fit statistical models; and build visualizations” (Buttrey and Whitaker 2016, p. 2). The detailed steps of our methodology, summarized in Figure 3.1, are:

1. **Gather** the data needed to conduct the analysis. For this research, we obtained historical ADMITS data from OPNAV N17.
2. **Prepare** the raw data for model fitting and implementation. In this step, we cleaned and manipulated the raw ADMITS data, and partitioned the resulting data set into training and testing data sets based on the type of supervised ML algorithm used.
3. **Train** supervised ML algorithms. In this step, we fit regression, classification, time series, and GNAR models using the appropriate training data sets. Additionally, feature selection methods such as stepwise selection and cross-validation were implemented.
4. **Implement** supervised ML algorithms. We applied the supervised ML models trained in the previous step to the appropriate testing data sets.

5. **Evaluate** model performance for each supervised ML algorithm. We compared performance measures such as RMSE, AUC, MAPE, and MASE to select the optimal performing regression, classification, forecasting, and GNAR model.
6. **Analyze** the output of the each optimal supervised ML model. In this final step, we analyzed the outputs of optimal models identified in the previous step, to include feature importance. In addition, we transformed our results into visualizations to aid in model output interpretability.

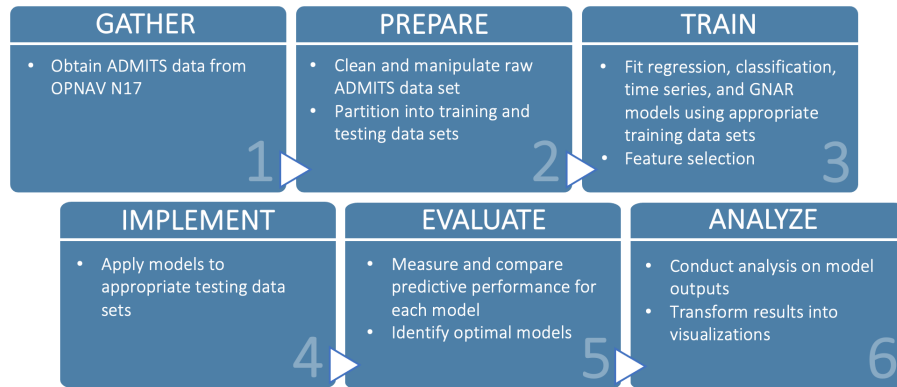


Figure 3.1. **Overview of Thesis Methodology.** Our methodology is comprised of six steps. In step 1, we obtained the ADMITS data from OPNAV N17. In step 2, we cleaned, manipulated, and partitioned the data set. In steps 3 and 4, we trained and implemented our models. In step 5, we measured predictive performance and identified the optimal models. In step 6, we analyzed the outputs and transformed our results into visualizations.

3.2 Data

Data comes in many forms. In almost every case, data is expected to be comprised of several different data types (Buttrey and Whitaker 2016). The ADMITS data analyzed in this study consists of three data types: **character**, **date**, and **logical**. Text is represented by a sequence of characters, also known as strings. The **character** data type is used to represent attributes that consist of categorical data. In R, categories are often converted into factors, where the different categories are stored as levels for each categorical attribute. The simplest date data type in R is called **date**, and “an object of this data type is represented internally as an integer representing the number of days since a particular ‘origin’ date”

(Buttrey and Whitaker 2016, p. 80). The usual origin date in R is January 1, 1970. Logical data types store logical or Boolean values of either TRUE or FALSE.

3.2.1 Data Overview

The raw ADMITS data received from OPNAV N17 contains a total of 101,963 observations and 39 attributes. The raw data set is comprised of personal, demographic, command, and alcohol and drug incident information for 74,761 Sailors from October 2008 to December 2020. Prior to any analysis, the raw data set was modified to remove each Sailor’s personally identifiable information (PII), such as name and social security number, and was assigned a random unique identification number. Available observations from June 2020 to December 2020—a total of 9—were removed as a result of inaccurate reporting of alcohol- or drug-related incidents according to OPNAV N17. In addition, Sailor department, general rate, and treatment facility type attributes were added, and attributes that were determined unhelpful (e.g., PlainLanguageAddress) were omitted.

The resulting modified ADMITS data set used for analysis consists of 101,954 observations and 36 attributes for 74,752 Sailors, with alcohol and drug incident information dating from October 2008 to May 2020. Table 3.1 provides a description of each attribute in the modified data set. Subsets of the modified data set were manipulated based on the type of supervised ML algorithm being applied in the model fitting process. Modifications include the addition of an alcohol- or drug-related incident counter and whether or not a Sailor committed a repeat incident. Throughout the remainder of this thesis, the modified ADMITS data set will be referred to as the ADMITS data set.

Table 3.1. **ADMITS Data Set Attributes.** Adapted from: OCNO (2020).

Attribute	Description
ID	Member’s unique identifier
UICRUC	Member’s command 5-digit identifier for each DAR
CommandName	Member’s command name
SerBranchID	Member’s branch of service
Current.UIC	Member’s current command 5-digit identifier
Continued on next page	

Attribute	Description
ReportDate	Date the member was entered into ADMITS
IncidentDate	Date of DAR incident or referral
SubstanceName	DAR substance
IncidentInfoName	Incident information for DAR
DUIDWI	Driving under the influence/while intoxicated
City	Command location by city
State	Command location by state
Zip	Command location by ZIP code
cSerBranchID	Command's branch of service
Sex	Member's gender
BirthDate	Member's date of birth
Race	Member's race identification
EthnicGroup	Member's ethnic group identification
mSerBranchID	Member's branch of service
CurrentMember	Member still on active service in the military
Rank	Member's pay grade
Rate	Member's Navy occupation with rank
GenRate	Member's Navy occupation without rank
Dept	Member's command department
ScreeningDate	Date member was screened by SARP facility
sFacilityName	SARP facility name
ReferralReason	Reason member was referred to SARP for screening
RecommendedAction	SARP recommendation for treatment
CompletionDate	Treatment completion date
ProgramType	SARP program type or name
ReferralSource	Source of referral for DAR
ProgramAction	SARP treatment recommended program action
PrognosisName	SARP recommendation for successful recovery
Counselor	SARP facility counselor
tFacilityName	SARP treatment facility name
tFacilityType	SARP treatment facility type

3.2.2 Data Limitations

The ADMITS data set comes with several limitations. A common problem found in real-world data is missing values. According to Buttrey and Whitaker (2016), an essential skill for a data scientist is to be able to identify and understand the effects of missing values on the analytical problem, and to know what resources are available to counter them. In the ADMITS data set, 25.1 percent of observations contained missing values. Figure 3.2 provides a summary of missing values in the ADMITS data set using the `vis_miss()` function from the `naniar` package (Tierney et al. 2021). Much of the missing values were found within attributes used to describe a Sailor’s alcohol or drug incident and treatment information (e.g., IncidentInfoName and tFacilityName).

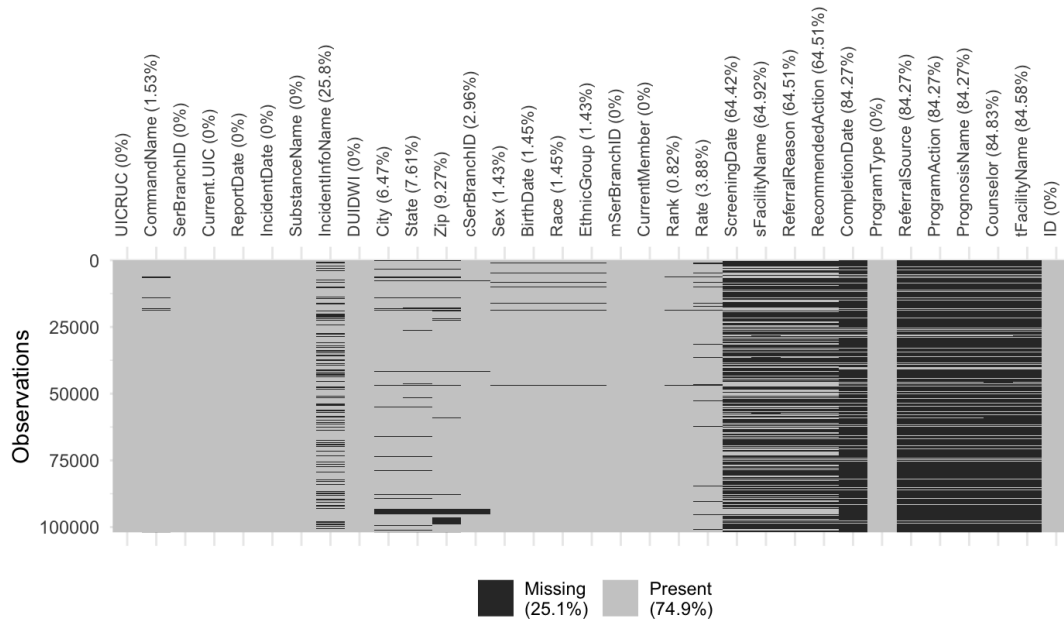


Figure 3.2. **Summary of Missing Values in ADMITS Data Set.** 25.1 percent of observations within the ADMITS data set contains missing values. Adapted from: OCNO (2020) and Tierney et al. (2021).

Missing values and observations in the ADMITS data set can arise in several ways. Data could be missing due to unforeseen circumstances or incomplete and inaccurate reporting from commands submitting a DAR. The affect of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) on the data set is evident in the significant decline of incidents

reported by commands in 2020. In addition, ADMITS experienced a system outage in late 2013 that affected alcohol-related incident data. This resulted in a significant decrease in the total number of incidents reported at the end of 2013 into 2014. Generally, missing values can be left alone, replaced, or removed (Buttrey and Whitaker 2016). In this study, we omitted observations with missing values for regression, classification, and time series forecasting models during the model fitting process. The resulting ADMITS data set with no missing values still provided 74,752 observations for analysis. Another common problem found in real data is human error in manual data entry databases, such as ADMITS. The database is prone to spelling, grammar, or punctuation mistakes. In the data cleaning process, much time was spent correcting misspellings, unnecessary spacing, and non-standardized inputs of attributes like command names, cities, or states. For example, “California” and “california” are considered to be two different entities in R.

3.3 Model Fitting and Implementation

The model fitting process entails training an ML algorithm to predict, classify, or forecast information of interest from attributes in a training data set, tuning it as necessary, and validating it by implementing the model on a testing data set. Four general types of models were trained in this study: regression, classification, time series forecasting, and GNAR, each providing various insights on the correlations between Sailors’ backgrounds and their alcohol- or drug-related incident information.

3.3.1 Regression and Classification

By applying regression and classification algorithms to ADMITS data, this thesis seeks to identify with what degree of certainty we can (1) predict the number of alcohol- or drug-related incidents committed by a particular Sailor, and (2) classify whether or not a particular Sailor will repeat an alcohol- or drug-related incident. Furthermore, through the implementation of feature importance, we explore which attributes prove most useful in classifying whether or not a Sailor commits a repeat incident. Figure 3.3 illustrates the regression and classification modeling workflow applied in this study.

Regression and classification data sets were derived by modifying the ADMITS data set. Additionally, they were manipulated to construct the target variable. For regression, the target

variable is `NumIncidents`, or the total number of incidents for each Sailor. For classification, the target variable is `RepeatIncidents`, or whether or not a Sailor committed a repeat incident (represented by the logical `TRUE` or `FALSE`). The regression and classification data sets implemented an “80/20 train-test split,” in which a random selection of 80 percent of observations was used for the training data set and the rest for the testing data set. Of note, observations with missing values were omitted. Additionally, at least five observations were present for each factor level to ensure a balanced train/test split. To assess the predictive performance of the regression models, we used the RMSE. For classification models, we evaluated classification performance by comparing each model’s ROC and AUC.

Fitting Regression Models

This study implements a single regression algorithm, Poisson regression, to model counts of alcohol- or drug-related incidents for each Sailor. The `predict()` function, from the `stats` package in R, was used to predict values based on the testing data set and the Poisson regression model (R Core Team 2019). The following function and individual parameters were used to build the regression model:

- **Poisson Regression** – This model was fit using the training data set and the `glm()` function from the `stats` package in R. The “`glm()` function is used to fit GLMs, specified by giving a symbolic description of the linear predictor and a description of the error distribution” (R Core Team 2021a, Description section). To indicate a Poisson family object, `poisson(link='log')` was used in the `family` argument.

Fitting Classification Models

This thesis fits three different classification algorithms to model whether a particular Sailor commits a repeat incident: logistic regression, SVM, and random forest. Like the Poisson regression model, we applied the `predict()` function to classify classes based on the testing data set and the particular classifier being used. The following functions and individual parameters were used to build each classification model:

- **Logistic Regression** – Similar to fitting the Poisson regression model, the logistic regression model was built using the training data set and the `glm()` function from the `stats` package in R. To specify a logistic family object, `binomial(link='logit')` was inputted into the `family` argument (R Core Team 2021a). We performed feature selection by using the `step()` function, which selects a model by different criteria

using a stepwise algorithm (the AIC criterion by default) (R Core Team 2021b). A forward and backward selection process for the stepwise algorithm was specified by indicating both for the `direction` argument.

- **SVM** – To build this model, we used the training data set and the `tune.svm()` function from the `e1071` package (Meyer et al. 2019). The function employs 10-fold cross-validation and “returns a tuning object including the best parameter set obtained by optimizing over the specified parameter vectors” (Meyer et al. 2019, p. 62). The best SVM model was extracted using the `$best.model` operator. In our model, we applied a linear kernel and weights to balance the data. The function takes arguments of `cost`, or the cost of misclassification, and `gamma`, or the parameter of a Gaussian Kernel to handle non-linear classification (Meyer et al. 2019). The sets of values `(0.1, 1, 10, 100, 1000)` and `(0.5, 1, 2, 3, 4)` were used for these parameters.
- **Random Forest** – This model was built using the training data set and the `randomForest()` function from the `randomForest` package (Liaw and Wiener 2002). Classification was specified using the `type` argument. Additionally, the function takes arguments of `ntree`, or the number of trees to grow, and `importance`, or whether the importance of predictors should be assessed (Liaw and Wiener 2002). Inputs of `500` and `TRUE` were used for these parameters.

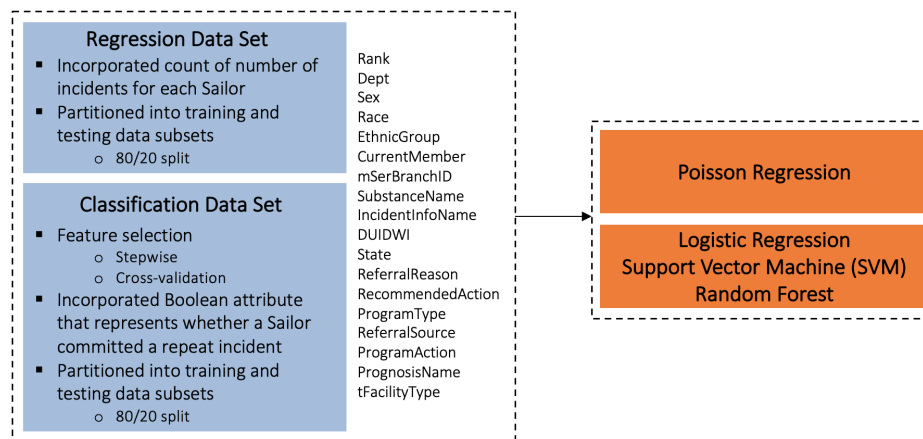


Figure 3.3. **Regression and Classification Workflow.** First, the regression and classification data sets were structured by modifying the ADMITS data set based on the response variable being predicted or classified. Next, the data sets were further split into training and testing data sets for both methods. Lastly, the regression and classification models were fit using the appropriate functions in R.

3.3.2 Seasonal Decomposition and Time Series Forecasting

By applying time series forecasting algorithms to ADMITS data, this study aims to identify with what degree of certainty we can forecast the total number of alcohol- or drug-related incidents for a given period of time. Figure 3.4 provides the time series forecasting modeling workflow applied in this thesis. To assess the forecasting performance of each algorithm, we employed MAPE and MASE measures in conjunction with a rolling-horizons design to compare each model's performance.

Fitting Time Series Models

Prior to fitting the different time series models, we first manipulated the ADMITS data set into a data set with a time series structure. The resulting time series data set consists of the total monthly count of alcohol- or drug-related incidents from October 2008 to May 2020. The time series data set was further partitioned into training and testing data sets, from October 2008 to December 2016 and January 2017 to May 2020, respectively.

This study fits five different time series forecasting algorithms to model the total monthly count of alcohol- or drug-related incidents: naïve, seasonal naïve, SARIMA, ETS, and Holt-Winters (multiplicative and additive). In addition, as previously mentioned in Section 2.2.2, we fit an STL model to identify potential seasonal or trend patterns in the monthly count of incidents. All time series forecasting models were fit using the training data set and functions from the `stats` and `forecast` packages in R (Hyndman et al. 2020). Forecasts for each model were measured using the testing data set and the `forecast()` function from the R package previously mentioned. Additionally, the naïve models were used as benchmarks to determine whether or not a particular forecasting model has predictive power. The following functions and individual parameters were used to build each time series forecasting model:

- **Seasonal Decomposition** – This model was built using the `stl()` function. The function takes arguments of `y`, or the numeric vector or time series, and `s.window`, which takes the string “periodic” (Hyndman et al. 2020). The time series data set was used for `y`.
- **Naïve and Seasonal Naïve** – These models were built using the `naive()` and `snaive()` functions. Both functions take arguments of `y`, or the numeric vector or time series, and `h`, or the number of periods for forecasting (Hyndman et al. 2020). The training data set and the length of the testing data set were used for these

parameters.

- **SARIMA** – To build this model, we first applied the `auto.arima()` function to the training data set to extract the best ARIMA model based on the optimal AIC or BIC value (Hyndman et al. 2020). Then, the model was built using the parameters extracted from the best model. The default behavior of the `auto.arima()` function is to model seasonal behavior (Hyndman et al. 2020).
- **ETS** – This model was built using the `ets()` function. The function takes arguments of `y`, or the numeric vector or time series, and `model`, or the three-character method-naming convention discussed in Subsection 2.2.2 (Hyndman et al. 2020). To indicate that the parameters were to be automatically selected, `ZZZ` was utilized. The parameters chosen and implemented for the model were `A,N,A`.
- **Holt-Winters** – Two types of Holt-Winter models, multiplicative and additive, were built using the `hw()` function based on the seasonality in the model. The function takes arguments of `y`, or the numeric vector or time series, and `h`, or the number of periods for forecasting (Hyndman et al. 2020). The training data set and the length of the testing data set were used for these parameters. Additionally, the `seasonal` argument of the function was used to indicate the type of seasonality (multiplicative or additive).

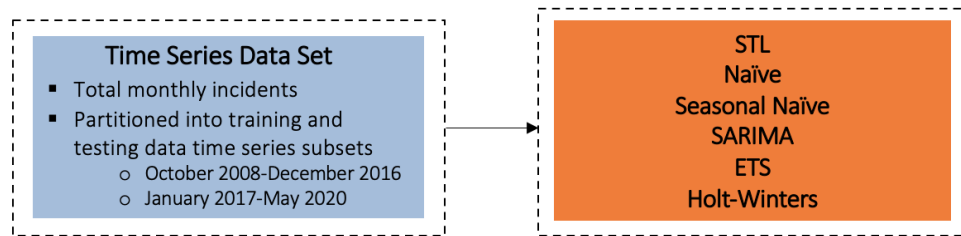


Figure 3.4. **Time Series Workflow.** First, a modified ADMITS data set, called time series data set, was structured using the total monthly incidents from October 2008 to May 2020. Then, the data set was further split into training and testing data sets, from October 2008 to December 2016 and January 2017 to May 2020, respectively. Lastly, the seasonal decomposition and forecasting models were fit using the appropriate functions in R.

3.3.3 GNAR

With data regarding the number of alcohol- or drug-related incidents collected and social network adjacencies identified, we can fit a social network time series model using the GNAR package in R (Knight et al. 2020b). Using GNAR statistical models, we seek to identify with what degree of certainty we can predict when a particular Sailor will commit and/or repeat an alcohol- or drug-related incident. Figure 3.5 describes the GNAR modeling workflow applied in this study. Under our method, we built a time series on social networks that we constructed from a random selection of 500 Sailors from the ADMITS data set, where a node represents a Sailor and an edge represents the connection between each Sailor.

Since the social network on all Sailors in the data set was a very dense network, i.e., nodes were connected to almost all other nodes, it was infeasible to train the GNAR models using the entire data set. As a result, we utilized a subset of randomly chosen 500 Sailors to increase computational efficiency. In addition, a sub-network of a random sample of 500 Sailors would be similar to a sub-network of any other random sample of 500 Sailors in the data set. The social networks were built based on values in an adjacency matrix, where any (i, j) entry greater than zero establishes an edge (or connection) between Sailor i and Sailor j . An edge was established between Sailor i and Sailor j if they shared a common attribute. Weighted edges were utilized to represent the number of attributes Sailors have in common. Furthermore, each node at network time t was assigned a number that represents the number of alcohol- or drug-related incidents committed by a particular Sailor. This number can change depending on whether the Sailor commits a repeat incident over time. MAPE and MASE measures in conjunction with a rolling-horizons design were employed to evaluate the predictive performance of each GNAR model.

GNAR Assumptions

The GNAR package's most basic assumption is that the time series data set has an underlying network structure (Knight et al. 2020a). Since the ADMITS data set contains time series data of alcohol- or drug-related incidents for each Sailor, and Sailors can be organized into a social network structure, this assumption is satisfied.

The GNAR package also assumes that the multivariate time series follows an autoregressive framework at each node and is influenced by its neighboring nodes (Knight et al. 2020a). In our social network time series structure, we utilized the number of cumulative incidents

over the given time-range for each Sailor. Consequently, the cumulative number of incidents depended on the previous number of incidents for that particular Sailor. Furthermore, given the nature of a social network, it follows that the number of alcohol- or drug-related incidents a Sailor commits is likely influenced by a Sailor’s adjacent neighbors. Therefore, this modeling assumption is relevant and applicable.

Fitting GNAR Models

To build the GNAR models, we adjusted individual parameters to create three different model fits using the `GNARfit()` function in R (Knight et al. 2020b). The function takes the social networks and multivariate time series as inputs. The first model was fit with no parameter modifications. This was used as a benchmark to measure predictive performance. For the second model fit, we applied a non-negative integer, `alphaOrder = 1`, that specifies a maximum time-lag of 1 to model along with a vector of length `betaOrder = 0`, which specifies the maximum neighbor set to model at each of the time-lags (Knight et al. 2020b). We utilized the same `alphaOrder` parameter for the third model fit, but applied a `betaOrder = 1`.

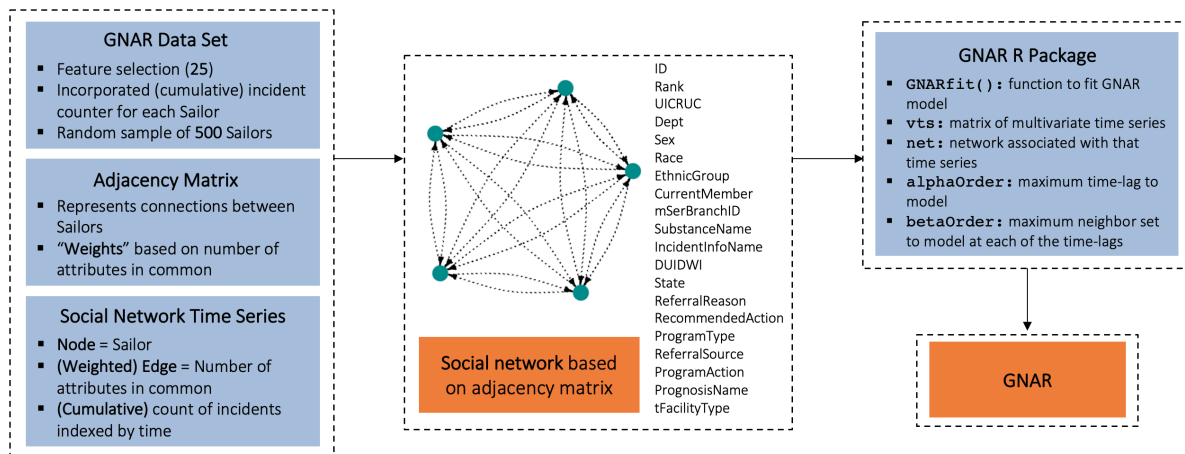


Figure 3.5. **GNAR Workflow**. First, a random sample of 500 Sailors was taken from a modified ADMITS data set called GNAR data set. Second, an adjacency matrix—in which nodes represent Sailors and edges represent connections between them—was created to construct social networks. Then, time series data was constructed using the cumulative count of incidents for each Sailor indexed by time. Lastly, three GNAR models with varying parameters were fit using the constructed social networks, multivariate time series, and `GNARfit()` function in R.

CHAPTER 4:

Results and Analysis

This chapter presents results of the ML models built and the key findings that emerged from their analyses. In addition, we compare our regression, classification, time series forecasting, and GNAR models' predictive performance using their respective performance measures introduced in Section 2.4 to identify the optimal model for each category.

4.1 Model Results and Analysis

4.1.1 Regression and Classification

Table 4.1 provides the RMSE and AUC predictive performance measures for the Poisson regression model and the three classification model fits. Due to the response variable—the number of alcohol- or drug-related incidents committed by each Sailor—being a non-negative integer, we did not consider any other regression models beyond Poisson regression. Although the model's results provided insight to what attributes were most useful in predicting the number of alcohol- or drug-related incidents committed by a particular Sailor, to accurately predict the number of incidents, further exploration of other regression models is needed for comparison.

By contrast, we notice that the classification models performed reasonably well. Figure 4.1 depicts the ROC curves for all three classification models. Out of the three classification model fits, logistic regression yielded the highest AUC and is therefore identified as the optimal classification model. As illustrated in Figure 4.1, we see that the curve representing logistic regression is closer to the axes, indicating a higher AUC. The predictive performances for all three classifiers were fairly close and could effectively be applied in follow-on substance abuse research.

Table 4.1. **Regression and Classification Model Summary.** RMSE and AUC values for regression and classification models, respectively. The logistic regression model outperformed the SVM and random forest models as indicated by the highest AUC. Adapted from: OCNO (2020).

Model	RMSE	AUC
Regression		
Poisson Regression	0.5464	
Classification		
Logistic Regression		0.6903
SVM		0.6788
Random Forest		0.6359

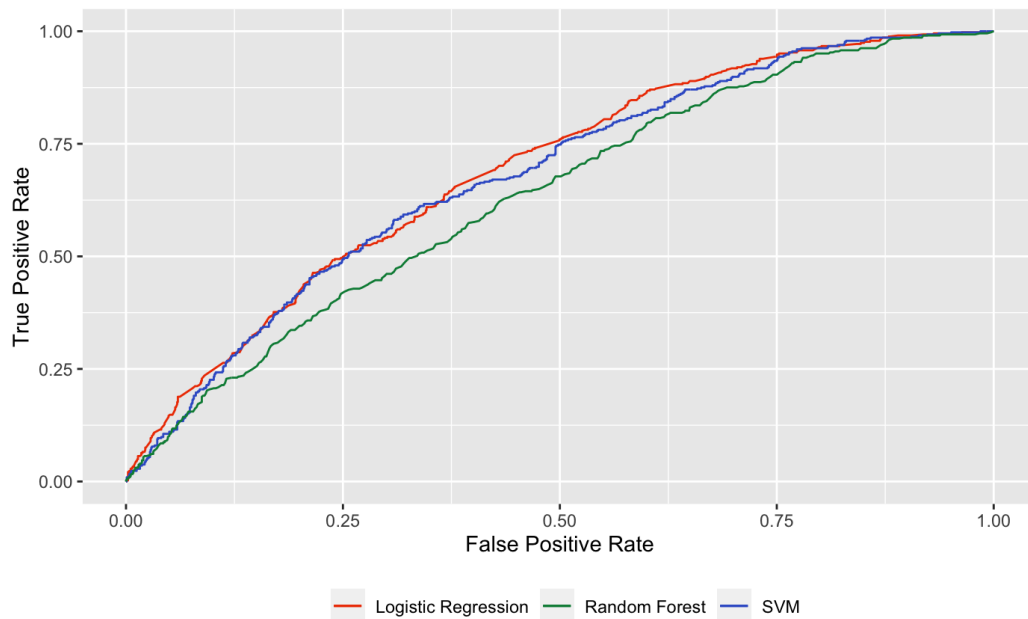


Figure 4.1. **Classification Results.** ROC curves for all three classification model fits. The logistic regression model outperformed the SVM and random forest models as evidenced by its curve being closest to the axes. Adapted from: OCNO (2020).

Figures 4.2 and 4.3 present permutation-based variable importance for the Poisson and logistic regression models using the DALEX package in R (Biecek 2018). Permutation-based variable importance is “a model-agnostic approach to the assessment of the influence of an explanatory variable on a model’s performance” (Biecek and Burzykowski 2021, “Variable-importance Measures” section). The importance of a particular variable is measured based on its effect on the model’s performance after permuting the values of the variable or group of variables (Biecek and Burzykowski 2021).

Figures 4.2 and 4.3 show outcomes of 50 permutations and the loss functions of RMSE (for regression) and $1 - \text{AUC}$ (for classification), respectively. The plots indicate that for both the Poisson and logistic regression models, the most important variable is DUIDWI. Figure 4.2 suggests that there exists a correlation between the number of alcohol- or drug-related incidents committed by a Sailor and whether or not an incident was classified as a DUIDWI. Similarly, Figure 4.3 suggests that there exists a correlation between whether a Sailor has committed a repeat incident and whether or not an incident was classified as a DUIDWI.

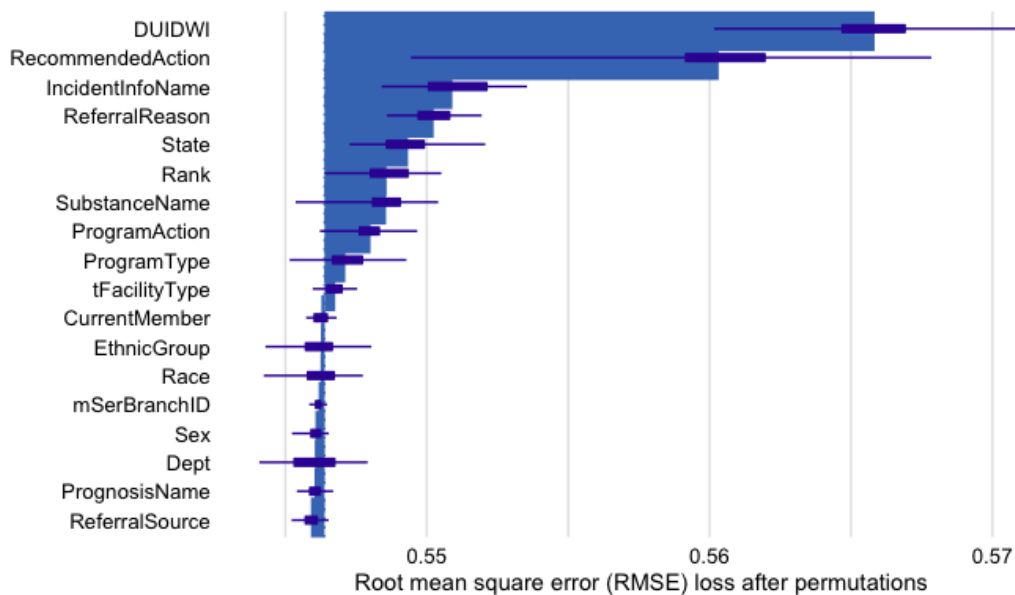


Figure 4.2. **Variable Importance for Poisson Regression.** The most useful attribute in predicting the number of alcohol- or drug- related incidents committed by a particular Sailor is DUIDWI, followed by RecommendedAction, IncidentInfoName, ReferralReason, and State. Adapted from: Biecek (2018) and OCNO (2020).

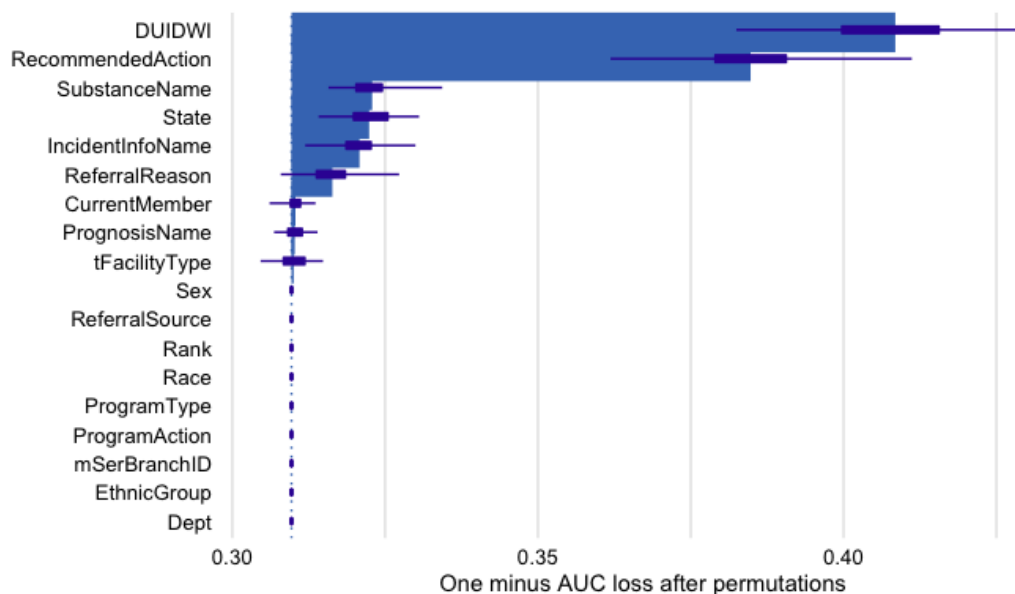


Figure 4.3. **Variable Importance for Logistic Regression.** The most useful attribute in classifying whether or not a particular Sailor will repeat an alcohol- or drug-related incident is DUIDWI, followed by RecommendedAction, SubstanceName, State, and IncidentInfoName. Adapted from: Biecek (2018) and OCNO (2020).

4.1.2 Seasonal Decomposition and Time Series Forecasting

Figure 4.4 illustrates the seasonal and trend decomposition using Loess (STL) of the ADMITS time series data for the total monthly count of alcohol and drug incidents from October 2008 to May 2020. We notice a downward trend from 2008 to 2014, an upward trend from 2014 to 2019, and finally a downward trend from 2019 to 2020. The significant drop in the number of incidents in late 2013 was a result of the database outage, as previously discussed in Section 3.2.

Furthermore, we notice an oscillating pattern in the seasonal component of the STL model. These fluctuations indicate the presence of seasonality in the data. We can deduce a pattern from the plot in which the number of alcohol- or drug-related incidents decreases before the end of each year, spikes several months into the year, and then gradually decreases again. The STL model is robust to outliers and, as a result, “occasional unusual observations do not affect estimates of the trend-cycle and seasonal component” (Hyndman and

Athanasopoulos 2018, “STL Decomposition” section). This is evident in the STL model’s remainder component as we see it account for the database outage that occurred at the end of 2013, indicated by the large spikes in late 2013 to early 2014.

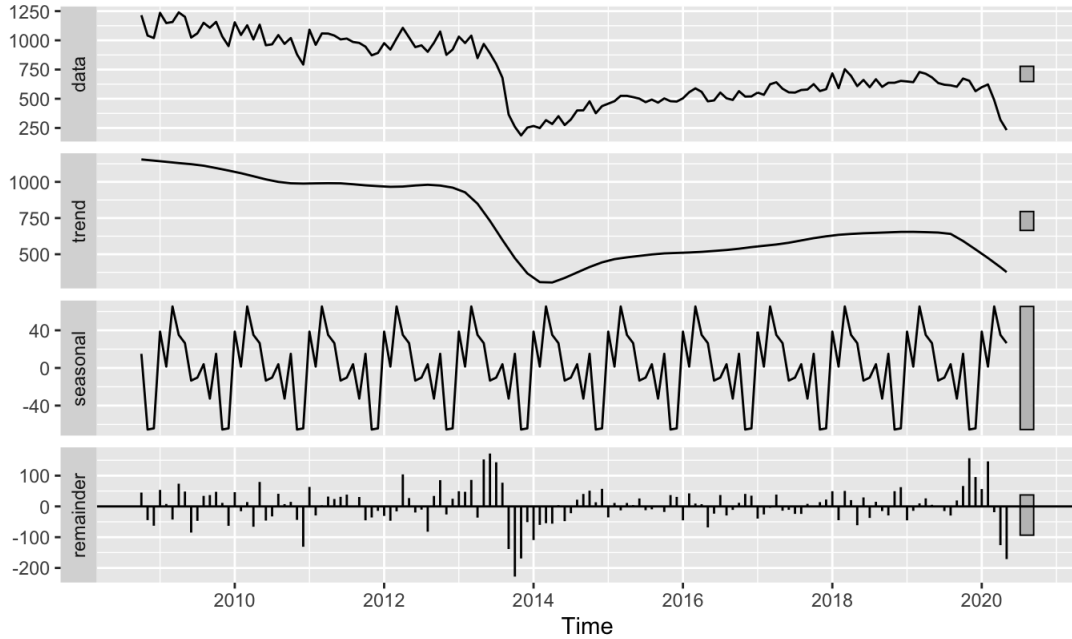


Figure 4.4. **Seasonal Decomposition of ADMITS Data Set.** The seasonal and trend decomposition using Loess (STL) of the ADMITS data set indicates that the data has an alternating decreasing and increasing trend and seasonal fluctuations. Adapted from: OCNO (2020).

Table 4.2 presents the 1-Step and 12-Step MAPE and MASE forecasting performance measures for the six time series forecasting model fits over the date range in the testing data set using a rolling-horizons design. As noted in Section 2.4, a model that outperforms the naïve model has predictive power. We can see that the SARIMA model is the only forecasting model that outperformed the naïve model in all MAPE and MASE measures. As previously discussed in Section 2.2.2, we utilized the `auto.arima()` function, which extracts the best model. The optimal SARIMA model chosen by the function was the $ARIMA(2, 1, 2)(1, 0, 1)_{12}$ model. Figure 4.5 illustrates the forecasting results for each time series forecasting model.

Table 4.2. **Time Series Forecasting Model Summary.** 1-Step and 12-Step MAPE and MASE forecasting model fits for six time series forecasting model fits. The SARIMA model outperformed all other time series forecasting models. Adapted from: OCNO (2020).

Model	1-Step MAPE	12-Step MAPE	1-Step MASE	12-Step MASE
Naïve	0.095	0.1	1	1
Seasonal Naïve	0.17	0.1	1.5	0.98
Holt-Winters (A)	0.1	0.13	1	1.2
Holt-Winters (M)	0.095	0.13	0.98	1.3
SARIMA	0.09	0.089	0.91	0.83
ETS	0.1	0.092	1	0.86

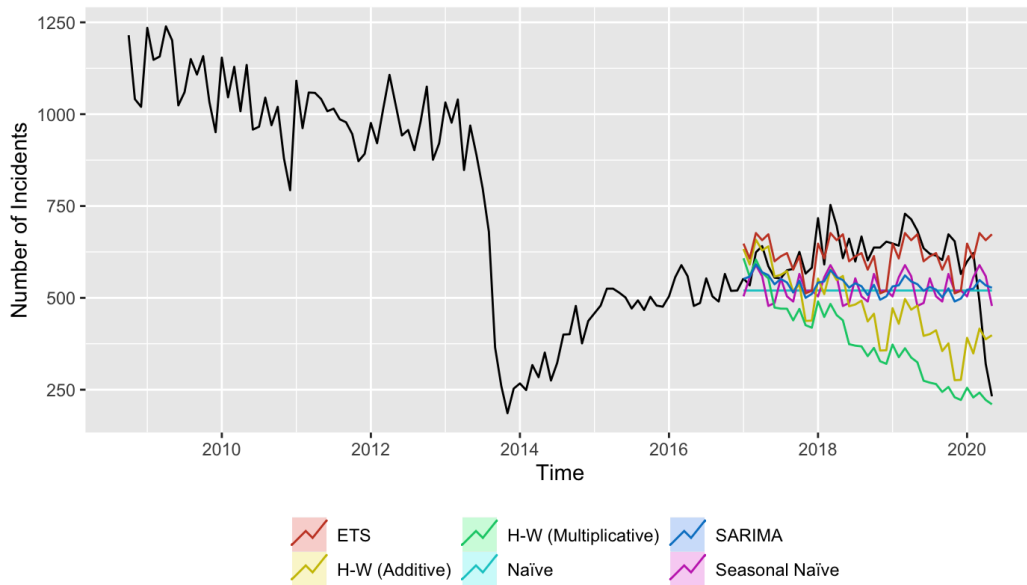


Figure 4.5. **Time Series Forecasting Results.** Forecasting results for all time series forecasting models. Visually, we are unable to identify which model performs best. However, as noted in Table 4.2, the SARIMA model outperformed all other time series forecasting models. Adapted from: OCNO (2020).

As demonstrated by Hyndman and Athanasopoulos (2018), we can analyze the residuals of a time series model fit using the `checkresiduals()` function to further assess its forecasting capability. Figure 4.6 presents a time plot of the residuals, the corresponding ACF, and a histogram for the $ARIMA(2, 1, 2)(1, 0, 1)_{12}$ model. We notice that all spikes in the ACF plot are within significance limits, indicated by the blue dashed lines. This suggests that the residuals in the model appear to be white noise, and that no remaining autocorrelations exist. As a result, the $ARIMA(2, 1, 2)(1, 0, 1)_{12}$ model surpasses performance standards, including those outlined in Section 2.4, and could effectively be applied to forecast the total number of monthly alcohol- or drug-related incidents.

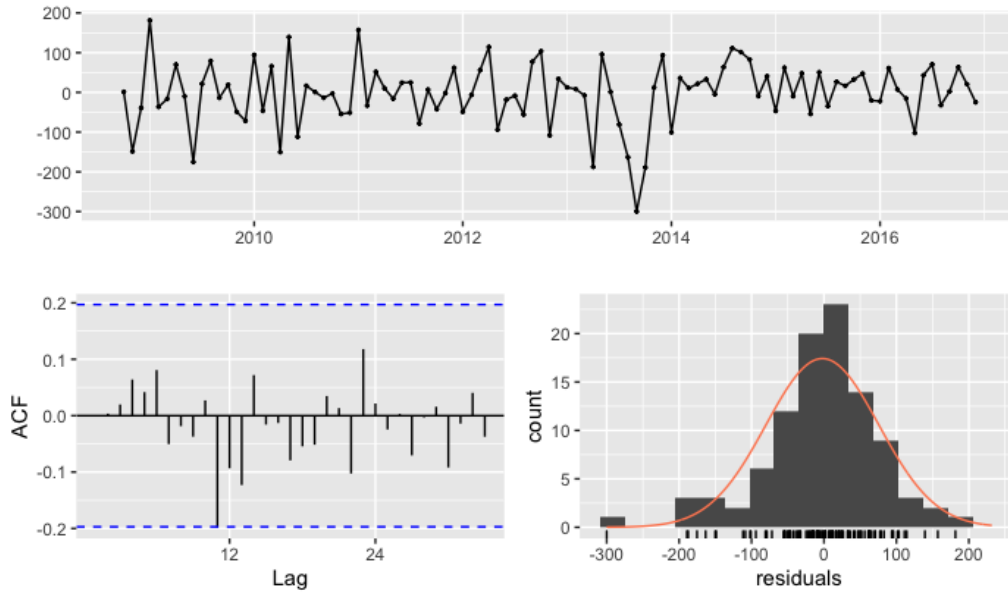


Figure 4.6. **Residuals from the $ARIMA(2,1,2)(1,0,1)_{12}$ Model.** ACF measures for the residuals of the model are all within significance limits, indicating no remaining autocorrelations. Adapted from: OCNO (2020).

4.1.3 GNAR

Table 4.3 presents the MAPE and MASE forecasting performance measures for the GNAR and naïve model fits. We see that all three GNAR models outperformed the naïve model in both MAPE and MASE measures, with GNAR Model 2 performing best. Since GNAR is a multivariate time series model that includes correlations between nodes in the given

network (through network edges and their associated weights), and the naïve model assumes that all nodes are independent (i.e., no edges in the network), it is expected that the GNAR models perform better than the naïve model. These results are further evidenced by Figures 4.7 and 4.8, which depict the MAPE and MASE measures, respectively, over the date range in the testing data set using a rolling-horizons design.

We see that the MAPE and MASE measures for all three GNAR models tend toward zero in Figures 4.7 and 4.8, respectively. MASE decreases at the start of 2018 and the end of 2019, but overall tends downward. This increase in forecasting accuracy is a direct result of the models effectively learning from previous observations to forecast future outcomes. In addition, we see that GNAR Model 2 is the most consistently accurate model out of the three parameter modifications. All three GNAR models surpass the performance standards outlined in Section 2.4, and could effectively be applied to forecast when a particular Sailor will commit and/or repeat an alcohol- or drug-related incident.

Table 4.3. **GNAR Model Summary.** MAPE and MASE forecasting performance measures for the GNAR and naïve model fits. All three GNAR models outperformed the naïve model, and GNAR Model 2 performed best overall. Adapted from: OCNO (2020) and Wren (2021).

Model	Mean	Median	Variance
MAPE			
GNAR Model 1	0.0645	0.0690	0.0004
GNAR Model 2	0.0246	0.0235	0.0001
GNAR Model 3	0.0794	0.0851	0.0005
Naïve	0.3143	0.3226	0.0043
MASE			
GNAR Model 1	0.2788	0.2850	0.0203
GNAR Model 2	0.2116	0.2074	0.0176
GNAR Model 3	0.3067	0.3149	0.0207
Naïve	1	1	0

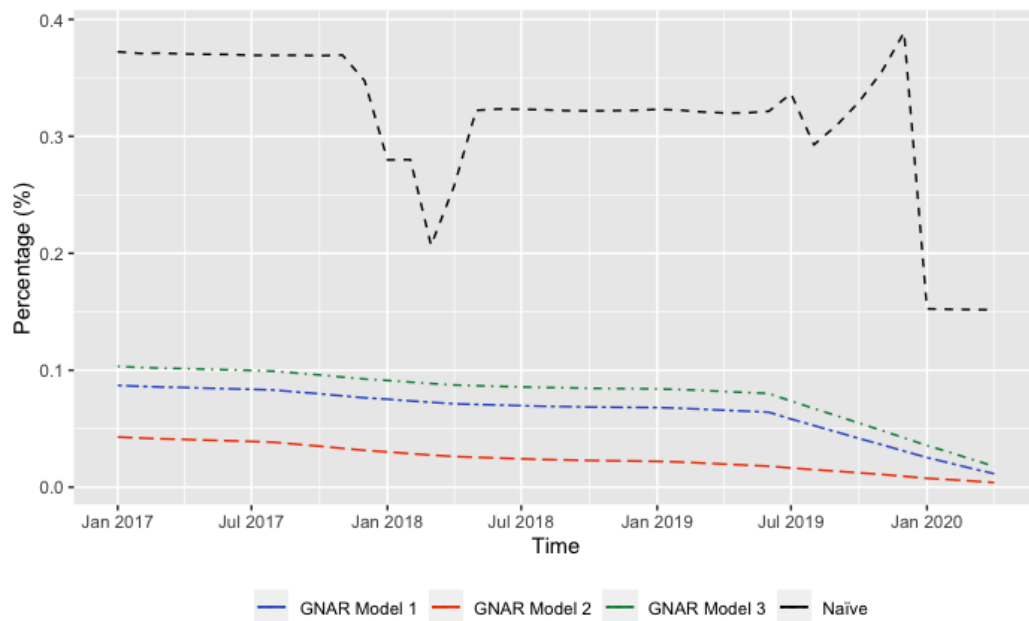


Figure 4.7. **GNAR MAPE Results.** MAPE forecasting performance measures for the GNAR and naïve models over the date-range in the testing data set using a rolling-horizons design. All three GNAR models tend toward zero in MAPE as the models learn and subsequently forecast. GNAR Model 2 outperformed all other models and yielded the highest accuracy over time. Adapted from: OCNO (2020).

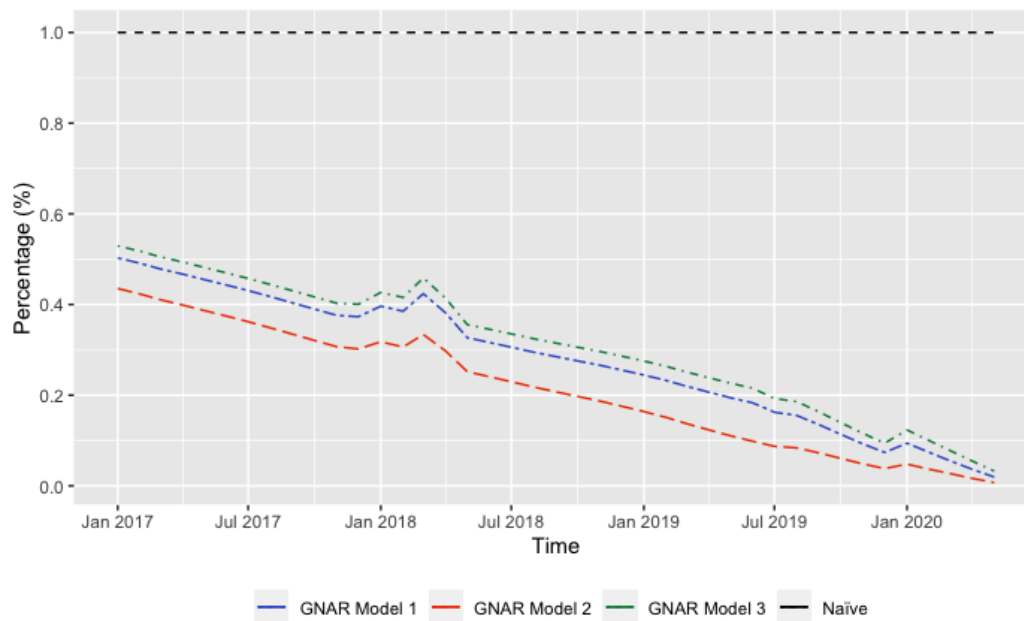


Figure 4.8. **GNAR MASE Results.** MASE forecasting performance measures for the GNAR and naïve models over the date-range in the testing data set using a rolling-horizons design. All three GNAR models tend toward zero in MASE as the models learn and subsequently forecast. MASE decreases at the start of 2018 and end of 2019, but overall tends downward. GNAR Model 2 outperformed all other models and yielded the highest accuracy over time. Adapted from: OCNO (2020).

CHAPTER 5:

Conclusion

This chapter summarizes our approach, findings, and analysis. Then, we introduce three recommended areas for future research regarding substance abuse among Sailors: models and model parameters, supplementary data collection, and sponsor interests. Continued research in these three areas could potentially increase each model's performance and ultimately their validity as decision-making tools for substance use and abuse training considerations and organizational changes.

5.1 Summary

This thesis employed several ML methods to better understand substance abuse and the potential risk factors that contribute to destructive behaviors among Sailors within the USN. In particular, the following research questions were addressed:

1. Using ML methods, with what degree of certainty can we predict, classify, or forecast
 - whether or not a Sailor will repeat an alcohol- or drug-related incident?
 - the number of alcohol- or drug-related incidents committed by a particular Sailor?
 - the total number of monthly alcohol- or drug-related incidents for a given period of time?
2. Using a social network time series structure, with what degree of certainty can we predict when a particular Sailor will commit and/or repeat an alcohol- or drug-related incident?
3. Through the implementation of ML models and feature importance, which attributes (or features) are most useful in classifying whether or not a Sailor will repeat an alcohol- or drug-related incident?

We built prediction, classification, forecasting, and GNAR models using data from the ADMITS database to provide insights on the correlations between Sailors' backgrounds and their alcohol- or drug-related incident information. To assess our methodology, we utilized predictive performance measures for regression, classification, and time-dependent

data. We compared each model’s predictive performance using their respective performance measure to identify which models performed best.

Our results strongly indicate that supervised ML methods provide an accurate and effective approach to modeling the different aspects of substance abuse, particularly among Sailors within the USN. By employing feature importance, we were able to identify that `DUIDWI`, `RecommendedAction`, `IncidentInfoName`, `ReferralReason`, `SubstanceName`, and `State` were the most useful attributes in (1) predicting the number of alcohol- or drug-related incidents committed by a particular Sailor, and (2) classifying whether or not a particular Sailor will repeat an alcohol- or drug-related incident.

The SARIMA time series forecasting model results demonstrate its ability to accurately model and subsequently forecast the total number of monthly alcohol- or drug-related incidents for a given time range. Additionally, the logistic regression classification model performed the best overall in classifying whether or not a particular Sailor will repeat an alcohol- or drug-related incident. Finally, predictive performance measures for the GNAR model with modified parameters (`alphaOrder=1`, `betaOrder=0`) suggest its capability to effectively model when a particular Sailor will commit and/or repeat an alcohol- or drug-related incident. As a result, these models have exhibited predictive power and are viable for further analysis.

5.2 Future Work

Several existing studies, described by Barenholtz et al. (2020), apply ML methods to substance use and abuse research, however, this approach has not been leveraged to better understand substance-abuse behaviors among Sailors within the USN. Our research establishes a foundation for the continued study of substance abuse among this particular group of individuals, and provides preliminary insights on the correlations between Sailors’ backgrounds and their associated alcohol- or drug-related incident information in ADMITS. Further exploration of the underlying causes and correlations of substance abuse among Sailors using ML methods should be conducted to inform potential training considerations and organizational changes.

5.2.1 Models and Model Parameters

The models built in this thesis provide a baseline for future studies, and their accuracy demonstrates the effectiveness of ML methods in substance abuse research within the USN. ML models have much potential in what they can offer, especially with regard to providing insights that can influence command-wide decisions. To further leverage the advantages of ML, additional efforts could be made to improve this study's models or develop more advanced predictive applications that can detect current abuse, assess future risk, or predict treatment success.

Complex ML Methods

According to Barenholtz et al. (2020), future alcohol and drug abuse studies that incorporate ML methods will likely see promising results from the use of modern deep neural network models. Currently, modern deep neural networks, such as convolutional neural networks (CNNs), are not often used in substance abuse research (Barenholtz et al. 2020). These complex ML methods have shown success in modeling unstructured data, and provide the means for future studies to include multiple data sources, new data types, and new model architectures—all of which are not possible with previous ML techniques (Barenholtz et al. 2020).

Data Transformations

This study did not implement data transformations or steps to normalize attribute values during the model fitting process. Normalizing values allows equal weights to be distributed among each variable so that no single variable skews model performance due to its large value. For example, it would be useful to present alcohol- or drug-related incident counts normalized by the size of their respective department, since larger departments are expected to have greater numbers. Moreover, proper data transformations can improve data quality and the overall efficiency of the analytic models.

GNAR Parameters

When building the social network structure for the GNAR models, we considered a static network where we assumed all Sailors entered and exited the network at the same time. A future study could explore the effects of Sailors joining and leaving the social network based on their report date entered into ADMITS and whether or not they were still on active duty service. Additionally, the social network's adjacency structure can be modified

when establishing an edge (or connection) between two Sailors. Rather than drawing an edge based on whether two Sailors share a common attribute, a hierarchical approach can be utilized instead. For example, an edge would be established between two Sailors only if they are located within the same state, command, and department (in that order).

5.2.2 Supplementary Data Collection

To further improve model development, data from external sources could be introduced into the model fitting process to potentially increase its predictive performance or provide additional insights on the correlations between Sailors and alcohol- or drug-related incidents.

Ship Schedule Data

Taking into account ship schedule data could provide additional information that can be utilized in ML analyses. Potential correlations and trends between ships' underway or deployment schedules and the number of alcohol- or drug-related incidents can be explored. Acquiring and preparing ship schedule data for all commands within an ADMITS data set for analysis, however, could be time-consuming and must be extracted from a classified source.

Social Media Data

There are several substance abuse studies that incorporate social media data into their analyses. According to Barenholtz et al. (2020), studies that utilize ML methods—CNN models, in particular—in combination with social media data have shown promising results in identifying potential substance use risk behavior among individuals. Social media platforms, such as Twitter, Facebook, and Instagram, “represent a source of real-time insight into user trends and behaviors” (Barenholtz et al. 2020, p. 340). This behavioral insight could be used to predict the likelihood of a Sailor committing or repeating an alcohol- or drug-related incident.

5.2.3 Sponsor Interests

OPNAV N17 has expressed several areas of interest for further analysis regarding the ADMITS data set. First, the sponsor has communicated interest in independent analyses for drug and alcohol incidents as their programmatic efforts focus on addressing these issues separately. Initial analyses show that separating the two types of incidents improve

accuracy rates for predicting whether or not a particular Sailor repeats an alcohol- or drug-related incident. Second, OPNAV N17 is interested in seeing the possible influence of the database outage in 2013 on alcohol- or drug-related repeat incidents. Separate analyses can be conducted among the years before and after the outage to identify any potential differences in results.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

- Afzali MH, Sunderland M, Stewart S, Masse B, Seguin J, Newton N, Teesson M, Conrd P (2019) Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation. *Addiction* 114(4):662–671, <https://doi.org/10.1111/add.14504>.
- Al-Masri A (2019) What are overfitting and underfitting in machine learning? Towards-DataScience. Accessed July 29, 2021, <https://towardsdatascience.com/what-are-overfitting-and-underfitting-in-machine-learning-a96b30864690>.
- Barenholtz E, Fitzgerald ND, Hahn WE (2020) Machine-learning approaches to substance-abuse research: emerging trends and their implications. *Current Opinion in Psychiatry* 33(4):334–342, <https://doi.org/10.1097/YCO.0000000000000611>.
- Biecek P (2018) DALEX: Explainers for complex predictive models in r. *Journal of Machine Learning Research* 19(84):1–5, <https://jmlr.org/papers/v19/18-416.html>.
- Biecek P, Burzykowski T (2021) *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models* (Chapman & Hall/CRC, New York City, New York), <https://pbiecek.github.io/ema/>.
- Brockwell PJ, Davis RA (2016) *Introduction to Time Series and Forecasting* (Springer International Publishing, Basel, Switzerland), 3 edition.
- Brown GG, Huddleston SH (2018) Machine learning. Cochran JJ, ed., *INFORMS Analytics Body of Knowledge*, 231–274 (John Wiley and Sons, Inc., Hoboken, New Jersey).
- Burger SV (2018) *Introduction to Machine Learning with R* (O’Reilly Media, Inc., Sebastopol, California).
- Buttrey SE, Whitaker LR (2016) *A Data Scientist’s Guide to Acquiring, Cleaning, and Managing Data in R* (John Wiley and Sons, Inc., Hoboken, New Jersey).
- Glen S (2019) Mean absolute scaled error: definition, example. StatisticsHowTo: Elementary Statistics for the Rest of Us. Accessed July 20, 2021, <https://www.statisticshowto.com/mean-absolute-scaled-error-mase/>.
- Glen S (2021) Mean absolute percentage error (MAPE). StatisticsHowTo: Elementary Statistics for the Rest of Us. Accessed July 20, 2021, <https://www.statisticshowto.com/mean-absolute-percentage-error-mape/>.
- Gudivada VN, Irfan MT, Fathi E, Rao DL (2016) Cognitive analytics: going beyond big data analytics and machine learning. Rao CR, ed., *Handbook of Statistics*, 169–205 (Elsevier B.V., Oxford, United Kingdom).

- Guo ZX, Li M, Wong WK (2012) Sparsely connected neural network-based time series forecasting. *Information Sciences* 193:54–71, <https://doi.org/10.1016/j.ins.2012.01.011>.
- Harkins G (2018) Binge-drinking rates are highest in these military services. Military.com. Accessed July 26, 2021, <https://www.military.com/daily-news/2018/07/20/binge-drinking-rates-are-highest-these-military-services.html>.
- Hornbuckle LR (2020) *Time series analysis of naval aviation lubricant testing*. Master's thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, NPS Archive: Calhoun.
- Hu H, Phan N, Chun SA, Geller J, Vo H, Ye X, Jin R, Ding K, Kenne D, Dou D (2019) An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. *Computational Social Networks* 6(10), <https://doi.org/10.1186/s40649-019-0071-4>.
- Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeeen F (2020) *forecast: Forecasting Functions for Time Series and Linear Models*.
- Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (OTexts, Melbourne, Australia), 2 edition, <https://otexts.com/fpp2/>.
- Islam UI, Sarker IH, Haque E, Hoque MM (2021) Predicting individual substance abuse vulnerability using machine learning techniques. Abraham A, Hanne T, Castillo O, Gandhi N, Nogueira RT, Hong TP, eds., *Hybrid Intelligence Systems*, 412–421 (Springer International Publishing, Basel, Switzerland), https://doi.org/10.1007/978-3-030-73050-5_42.
- James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning: with Applications in R* (Springer-Verlag New York, Inc., New York City, New York).
- Jing Y, Hu Z, Fan P, Xue Y, Wang L, Tarter RE, Kirisci L, Wang J, Vanyukov M, Xie XQ (2020) Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. *Drug and Alcohol Dependence* 206, <https://doi.org/10.1016/j.drugalcdep.2019.107605>.
- Knight M, Leeming K, Nason G, Nunes M (2020a) Generalized network autoregressive processes and the GNAR package. *Journal of Statistical Software* 96(5), <https://doi.org/10.18637/jss.v096.i05>.
- Knight M, Leeming K, Nason G, Nunes M (2020b) Package 'GNAR'. CRAN. Accessed June 10, 2021, <https://cran.r-project.org/web/packages/GNAR/GNAR.pdf>.

- Lewis CD (1982) *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting* (Butterworth-Heinemann, Oxford, United Kingdom).
- Liaw A, Wiener M (2002) Classification and regression by randomforest. *R News* 2(3):18–22, <https://CRAN.R-project.org/doc/Rnews/>.
- Lord D, Park BJ, Levine N (2015) Poisson regression modeling. *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v4.02)* (Ned Levin & Associates, Houston, Texas and the National Institute of Justice, Washington, D.C.), <https://nij.ojp.gov/topics/articles/crimestat-spatial-statistics-program-analysis-crime-incident-locations>.
- Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2019) *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-3, <https://CRAN.R-project.org/package=e1071>.
- Newman M (2003) The structure and function of complex networks. *SIAM Review* 45(2):167–256, <https://doi.org/10.1137/S003614450342480>.
- Newman M (2018) *Networks* (Oxford University Press Inc., Oxford, United Kingdom), 2 edition.
- Office of the Chief of Naval Operations (2009) Navy alcohol and drug abuse prevention and control. Technical Report OPNAVINST 5350.4D, Washington, DC, <https://navytribe.com/wp-content/uploads/2015/11/53504c.pdf>.
- Office of the Chief of Naval Operations (2020) Alcohol and drug management information tracking system. Dataset, years 2008 to 2020. Millington, TN.
- Office of the Chief of Naval Operations (2021) Alcohol and drug management information tracking system (ADMITS). Department of the Navy. Accessed July 24, 2021, <https://www.mynavyhr.navy.mil/Support-Services/21st-Century-Sailor/NAAP/ADMITS/>.
- Park SJ, Lee SJ, Kim H, Kim JK, Chun JW, Lee SJ, Lee HK, Kim DJ, Choi IY (2021) Machine learning prediction of dropping out of outpatients with alcohol use disorders. *PLoS ONE* 16(8), <https://doi.org/10.1371/journal.pone.0255626>.
- R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.
- R Core Team (2021a) Fitting generalized linear models. Accessed June 3, 2021, R Documentation, <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html>.

- R Core Team (2021b) Step function. Accessed June 3, 2021, R Documentation, <https://search.r-project.org/CRAN/refmans/emdi/html/step.html>.
- Tierney N, Cook D, McBain M, Fay C (2021) *nanian: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.6.1, <https://CRAN.R-project.org/package=nanian>.
- Uhart M, Wand GS (2008) Stress, alcohol and drug interaction: an update of human research. *Addiction Biology* 14(1):43–64, <https://doi.org/10.1111/j.1369-1600.2008.00131.x>.
- Valente TW, Gallaher P, Mouttapa M (2004) Using social networks to understand and prevent substance use: a transdisciplinary perspective. *Substance Use & Misuse* 39(10-12):1685–1712, <https://doi.org/10.1081/JA-200033210>.
- Vice Chief of Naval Operations (2013) Establishment of Navy’s Twenty-first Century Sailor Office (OPNAV N17). Technical Report NAVADMIN 153/13, Washington, DC, <https://navadmin-viewer.fly.dev/NAVADMIN/153/133>.
- Wren DM (2021) *SARS-COV-2 Dissemination using United States county adjacencies*. Master’s thesis, Department of Operations Research, Naval Postgraduate School, Monterey, CA, NPS Archive: Calhoun.
- Yoshida R (2020a) Data science modeling overview. Class notes, Advanced Data Analysis, Naval Postgraduate School, July 7, Monterey, CA.
- Yoshida R (2020b) (Univariate) Time series overview. Class notes, Advanced Data Analysis, Naval Postgraduate School, July 21, Monterey, CA.

Initial Distribution List

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