Feature Engineering and Ensemble Machine Learning in the Navy Reserve: Using Holistic Behavioral Profiles to Predict Mobilization Cancellation

Robert Milletich, Andrew Turscak*, Daniel Miller,
Mark Moreno*, Eric White, Robert Green, Shirley Bergstrom

Commander, Navy Reserve Forces Command
Norfolk, VA

***Corresponding Authors: Mark.A.Moreno1@navy.mil, Turscak_Andrew@bah.com**

ABSTRACT

Mobilization cancellation in the Navy Reserve drives mission gaps and compresses pre-mobilization timelines. The effect is felt most directly by those tagged to backfill the orders, with the window of preparation constrained by the time of cancellation. The mission of N36, Force Analytics, is to deliver data-driven insights that allow actionable decision-making to enhance the strategic depth and operational capability of the Navy Reserve Forces. N36 provides evidentiary data analysis supporting the reasoning behind decisions and insights made by the N3 Operations Directorate, CNRFC, and the Navy Reserve Force as a whole. In this paper, a machine learning pipeline is developed that predicts the individual likelihood of mobilization cancellation for all Navy Reserve Sailors. Holistic profiles are extracted for each Reservist that contain data about behavioral history in the Navy Reserve Forces, personal information, and administrative details tied to the Reservist. Using these profiles as inputs to a series of machine learning models, two experiments are presented that focus on optimizing a model's predictive performance and identifying an optimal subset of profile information to accurately identify Reservists at risk for mobilization cancellation. Preliminary results highlight the potential benefit of using a machine learning pipeline to assist with data-driven decisions regarding mobilization readiness. The study also indicates that a comprehensive data strategy, weighing manifold behavioral and personal records, is essential to predict behavior accurately, with no single factor or set of factors responsible for mobilization cancellation.

ABOUT THE AUTHORS

Dr. Robert Milletich (Booz Allen Hamilton) is Chief Data Scientist for N36, providing both technical and administrative analytical support. Current research interests are in ensemble learning and feature selection. He has a M.S. in Experimental Psychology, M.S. in Applied Mathematics, and a Ph.D. in Quantitative Psychology, all from Old Dominion University.

Mr. Andrew Turscak (Booz Allen Hamilton) is a Data Scientist for N36, providing analytical support and managing the CNRFC data science training curriculum. Professional interests include natural language processing and behavioral analytics. He holds a M.S. in Computational Operations Research and B.A. in Economics, both from the College of William & Mary.

Mr. Daniel Miller (Booz Allen Hamilton) is the N36 project lead, helping to empower the Navy Reserve to make data-driven decisions through the establishment of a data science capability. Professional interests include applied economics and ensemble learning. He is an M.S. candidate in Business Analytics at Indiana University and holds a B.A. in Economics and Business from the Virginia Military Institute.

Mr. Mark Moreno (CNRFC) is Deputy Director of Navy Reserve Force Analytics, code N36 responsible for establishing and growing the data analytics capability for the Navy Reserve Force and providing subject matter expertise and continuity for those efforts. He serves as a conduit between the contractors and military staff across all areas of the data analytics capability for each project. He holds a B.S. in Medical Technology from Univ. of WI Madison, retired from the Navy as a Submarine Full Time Support Officer, and has performed as a Program/Data Analyst at CNRFC since 2011.

Mr. Eric White (Booz Allen Hamilton) is a Data Engineer for N36, providing analytical support and managing the N36 web presence. He specializes in building and optimizing scalable ETL pipelines from unstructured data and disaggregate systems.

He holds a M.A. in Economics and Business Modelling and Simulation as well as a B.S. in International Business, both from Old Dominion University.

LCDR Robert Green (CNRFC) is the Technical Director of N36. He holds a M.S. in Operations Research from the University of Alabama in Huntsville, a graduate certificate in Modeling and Simulation Engineering from Old Dominion University, and a B.S. in Political Science from the United States Naval Academy.

Mrs. Shirley Bergstrom (Booz Allen Hamilton) is Strategy Lead for N36, guiding the direction of establishing a mature analytics department at CNRFC. She holds a B.S. in Mathematical Sciences from Virginia Commonwealth University with a concentration in Operations Research.

Feature Engineering and Ensemble Machine Learning in the Navy Reserve: Using Holistic Behavioral Profiles to Predict Mobilization Cancellation

Robert Milletich, Andrew Turscak*, Daniel Miller,
Mark Moreno*, Eric White, Robert Green, Shirley Bergstrom

Commander, Navy Reserve Forces Command
Norfolk, VA

***Corresponding Authors: Mark.A.Moreno1@navy.mil, Turscak_Andrew@bah.com**

INTRODUCTION

Problem Statement

Mobilization cancellation in the Navy Reserve Forces drives mission gaps and compresses pre-mobilization timelines. The effect is felt most directly by those tagged to backfill the orders, with the window of preparation constrained by the time of cancellation. Sailors selected for mobilization are currently chosen at random, provided they meet certain mission-essential qualifications and are available for mobilization. A thorough assessment of each individual sailor's risk of cancellation provides the opportunity for the Navy Reserve to better prepare for the possibility of cancellation, assess cancellation trends, and empower sailors to mobilize.

This paper hypothesizes that a Navy Reservist's risk of cancellation can be accurately predicted with machine learning, using a holistic profile of historical behavior and details known prior to initial contact with the sailor. Preliminary results highlight the potential benefit of using a machine learning pipeline to assist with data-driven decisions regarding mobilization readiness. The study also indicates that a comprehensive data strategy, weighing manifold behavioral and personal records, is essential to predict behavior accurately, with no single trait or set of related features responsible for mobilization cancellation.

Approach Summary

Navy data systems were assessed to identify opportunities and constraints posed by the data available. The fields available, history of data collection, and feasibility of data access were heavily influential in the choice of which features to include in a Reservist's mobilization risk profile. Cancellation rates across relevant populations in the dataset were visualized over time to determine if any factors exhibited correlation with cancellation. Variables assessed included but were not limited to prior cancellations, medical history, order and travel history, personal information, and drill history. Sparse, ambiguous, and/or irrelevant fields were removed from the analysis. To prevent data leakage, data were truncated to include only those details known prior to the initiation of a sailor's most recent mobilization orders. Sailors with no mobilization or cancellation history were not included in the training set.

Features were engineered to parse out statistical information hidden in the data using a variety of techniques including manual aggregation techniques, natural language processing, and unsupervised machine learning. A variety of supervised machine learning models were then trained and evaluated using fiscal year 2017 mobilizations as the training dataset and fiscal year 2018 as the test set. An ablation study was then performed to examine the effect of dropping various sets of variables from the model. Models were evaluated using the area under the curve (AUC) score, and other portions of the model building process revisited as necessary. Figure 1 details the iterative approach used in the construction of the machine learning model.

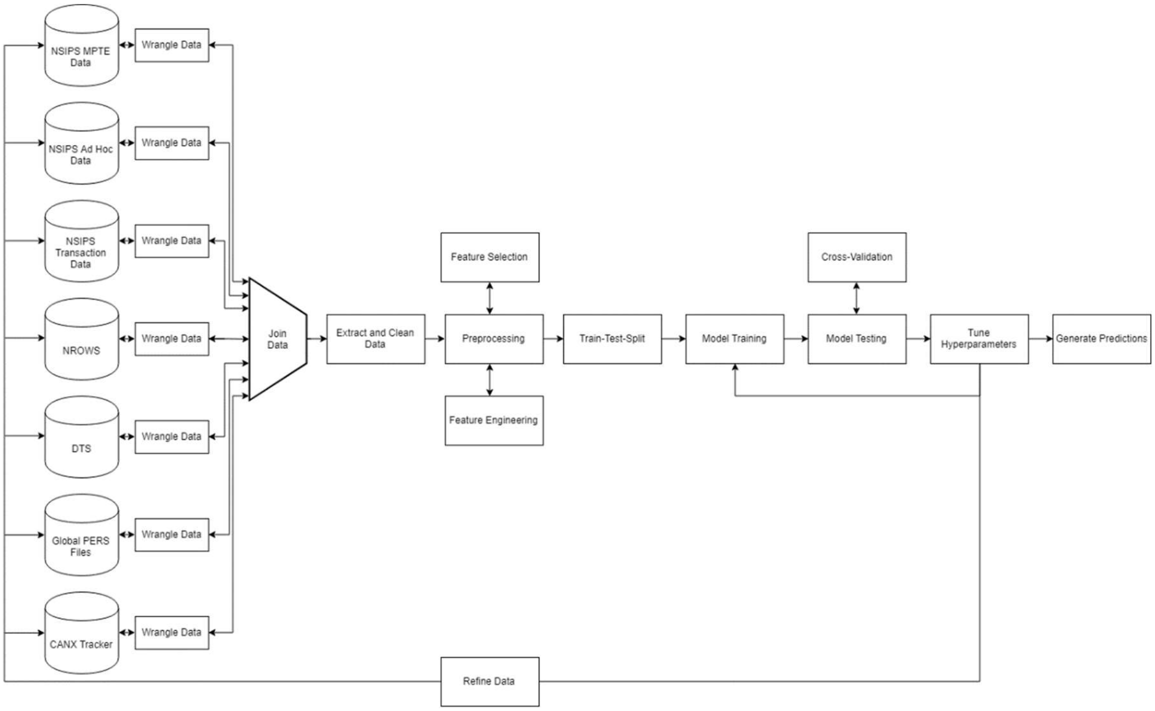

Figure 1. Iterative Analytic Approach

DATA

The sample consisted of Navy Reservists tagged for at least one mobilization in fiscal years 2017 and 2018. For each Reservist, we gathered one year of historical data, including mobilizations, manpower availability, work-related orders and travel, drills, and demographics, based on their most recent mobilization start date. Data were obtained from authoritative Department of Defense (DoD) sources, including the Navy Standard Integrated Personnel System, Defense Travel System, and the Navy Order Writing System. The raw data required substantial preprocessing and manipulation to make it numerically interpretable, and an array of innovative strategies were adopted to maximally parse out information hidden in the feature space.

Data Manipulation & Engineering

Prior to feature engineering, data were preprocessed to clean free text and impute missing values. Free text fields were parsed and aggregated together into higher-level categories using a fuzzy string matching algorithm. Specifically, dictionaries were built with appropriate aggregate categories as keys along with key words that map to those categories as values. The algorithm computes the Levenshtein Distance between all dictionary values and free text strings, using features such as length and shared text characters, and returns the value with the best match to the text string. The process was necessary to account for large disparities in manually entered fields and is depicted in Figure 2 in the context of civilian job categories.

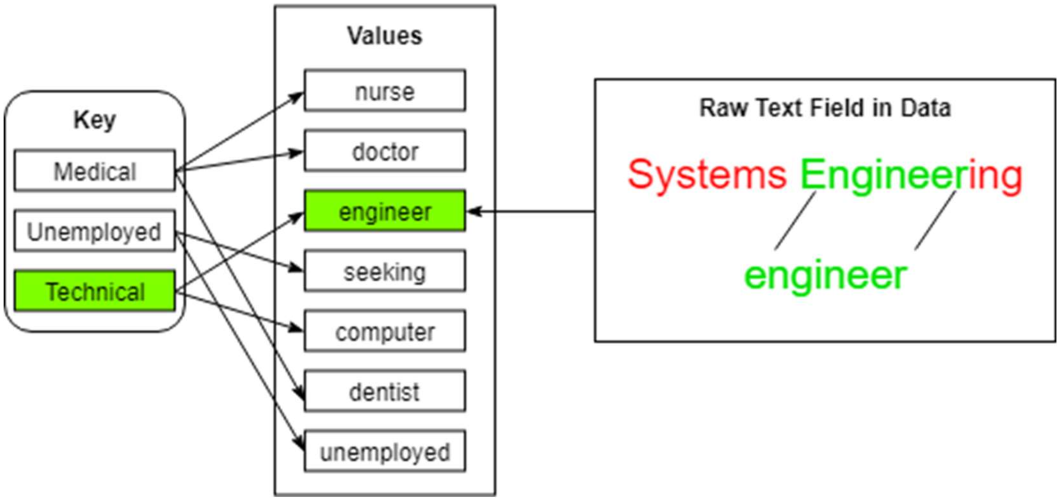

Figure 2. Fuzzy String Matching for Category Aggregation

For missing values, variables that contained meaningful zeros were imputed using zero, whereas, other variables were imputed using the mean, median, mode, maximum, or minimum of the observed data where appropriate. Several of the fields in the raw data came in a format that was minimally interpretable or wholly non-interpretable from a modeling perspective. For example, the field for rank was concatenated with 2-character and 3-character job codes, resulting in a large number of sparse levels in the factor. Additionally, the hierarchical nature of rank could be more appropriately modeled as an ordinal numeric value. The peculiarities of individual data fields were wide-ranging, from dated sequences of events, to relevant factors hidden in encoded substrings, to numeric and categorical variables described in free text fields. As such, a rigorous feature engineering pipeline was constructed to handle the idiosyncrasies of each feature.

Ordinal factors were mapped via manually created dictionaries to appropriate numeric values. Free text fields that were previously aggregated using the fuzzy string matching algorithm were converted to dummy variables. Many of these fields were further transformed into ordinal or numeric features where appropriate. Statistical information such as central tendency measures, dispersion measures, and other distributional measures (e.g., percentiles, entropy, coefficient of variation, kurtosis, skewness, minimum, maximum) were calculated from sequences of dated events, such as pre-mobilization readiness events, orders, travel, drill, and medical histories. Unsupervised learning methods were tested to extract hidden information from the feature space but were ultimately dropped for the sake of model interpretability and minimal predictive value added. The methods tested included cluster distances and labels using K-Means as well as latent feature representations using PCA and autoencoders. The final feature set consisted of 686 variables.

ANALYSIS

Train-Test-Split

Due to the temporal nature of the data, a temporal train/test split was used for model evaluation with the fiscal year 2017 mobilizations ($N = 2{,}869$) as the training set and fiscal year 2018 mobilizations ($N = 3{,}170$) as the testing set. Data only existed for sailors currently in the Navy Reserve, so dataset sizes decreased linearly going back each year due to sailors retiring or leaving the Force. For both training and testing sets, the class distribution was similar with approximately 11.5% cancellations and 88.5% mobilizations. To account for class imbalance, the AUC score was used to evaluate a model's predictive accuracy.

Importantly, given that each Reservist tagged for mobilization has a unique sequence of dates comprising their mobilization timeline, simple date thresholds and randomized splitting could not properly account for temporal effects known to artificially inflate model accuracy through posterior information leakage. To prevent this, feature data were truncated to include only those details known prior to the initialization of a sailor's most recent mobilization orders, hereafter referred to as the recent individual mobilization status (IMS) date.

Personal details that change with time, such as rank or marital status, were collected from legacy data stores corresponding to the recent IMS date. Tallied fields, such as the number of prior mobilizations, drill history, and average time between mobilization events, were calculated only up to the recent IMS date. All information occurring between the recent IMS date and the final mobilization or cancellation event –the response variable in the classification model– was purged to ensure that the risk scores could be known prior to assigning mobilization orders, as were any in-progress mobilization events trailing the response. This temporal control methodology is detailed in Figure 3.

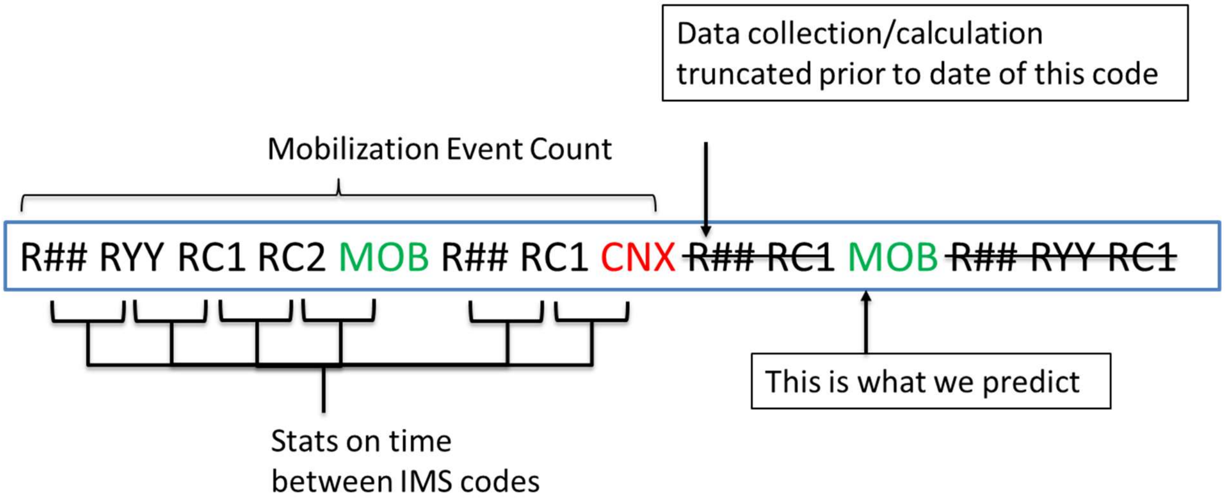
Data collection/calculation truncated prior to date of this code
Mobilization Event Count
R## RYY RC1 RC2 MOB R## RC1 CNX ~~R## RC1~~ MOB ~~R## RYY RC1~~
This is what we predict
Stats on time between IMS codes

Figure 3. Temporal Control by Mobilization Event

Experiment 1: Optimizing Model Predictive Accuracy

The goal of the first experiment was to optimize the predictive accuracy of a machine learning classifier. A series of initial models were tested, including logistic regressions, support vector machines, tree-based models (i.e., extreme gradient boosted trees, random forests, extremely randomized trees), k-nearest neighbors, and deep neural networks. Hyperparameter optimization was performed using random search with successive halving. If possible, early stopping was implemented to reduce the computational cost of evaluating each hyperparameter configuration. Overall, results highlighted that the tree-based models tended to have the best predictive accuracy, and these results held across a range of different hyperparameter configurations. For a small number of hyperparameter configurations, the deep neural networks had similar predictive accuracy as some of the tree-based models, however, the accuracy of the models was highly variable across other hyperparameter configurations.

To further increase the predictive accuracy of the models evaluated, the best performing models were selected and combined into a stacking model. Stacking is an ensemble learning technique that uses k-fold cross-validation to train a series of level 1 models and make predictions on the k holdout sets. The predictions of each level 1 model are stacked and provided as input to the level 2 meta-model. Our initial selection of level 1 models included the top three performing tree-based models, namely the extreme gradient boosted trees, random forests, and extremely randomized trees. The fourth level 1 model was selected based on its performance in the stacked model. Among the models evaluated, we found that the K-nearest neighbors resulted in the largest gain in accuracy. The level 2 meta-model was a logistic regression with L2 penalty to handle multicollinearity in the level 1 model predictions. The final stacking model used in this experiment is pictured in Figure 4.

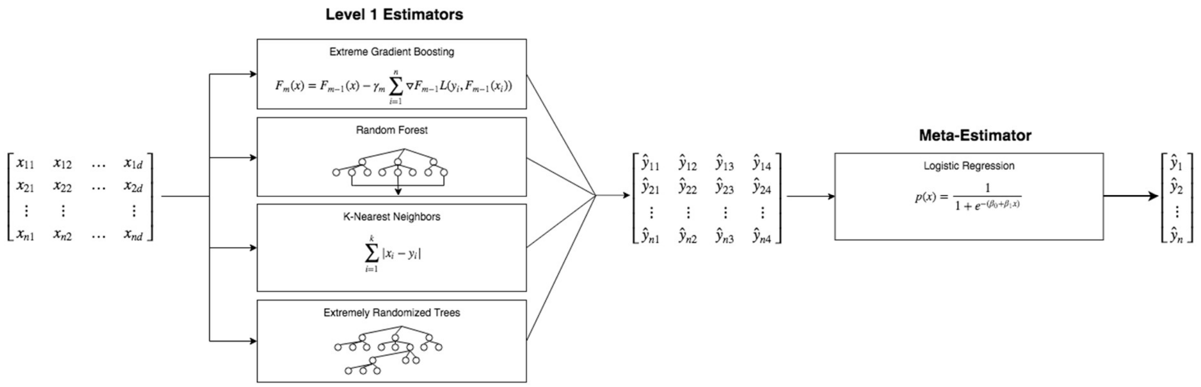
Level 1 Estimators
Extreme Gradient Boosting
$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla F_{m-1} L(y_i, F_{m-1}(x_i))$
Random Forest
K-Nearest Neighbors
$\sum_{i=1}^{k} |x_i - y_i|$
Extremely Randomized Trees
$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$
$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \hat{y}_{13} & \hat{y}_{14} \\ \hat{y}_{21} & \hat{y}_{22} & \hat{y}_{23} & \hat{y}_{24} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} & \hat{y}_{n3} & \hat{y}_{n4} \end{bmatrix}$
Meta-Estimator
Logistic Regression
$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$
$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$

Figure 4. Stacking Ensemble

Results showed that the best level 1 model was the extreme gradient boosted trees, with an AUC score marginally higher than both the random forests and extremely randomized trees. Although the K-Nearest Neighbors was less performant than the other level 1 models, its predictions combined with the models in the stacking ensemble led to approximately 2-12% gain in predictive accuracy compared to the level 1 models. Results of the level 1 models and stacking ensemble are summarized in Table 1.

Model	AUC
Extreme Gradient Boosted Trees	.804
Random Forests	.798
Extremely Randomized Trees	.795
K-Nearest Neighbors	.728
Stacking Ensemble	.818

Table 1. AUC Scores for Classifiers

Experiment 2: Ablation Study

To test the sensitivity of a model, should certain features become unavailable in a production environment, an ablation study was performed using extreme gradient boosted tree models. An ablation study removes features from the dataset in an iterative process, evaluating the model on each of the remaining feature sets until all combinations of features are tested and evaluated. We grouped similar features together (e.g., features about orders, features about drills) into mutually exclusive subsets, for a total of eight subsets. We evaluated all 256 possible feature subsets to gauge the minimal number of feature subsets needed before a model's predictive accuracy begins to drop. The results of the ablation study are detailed in Figure 4.

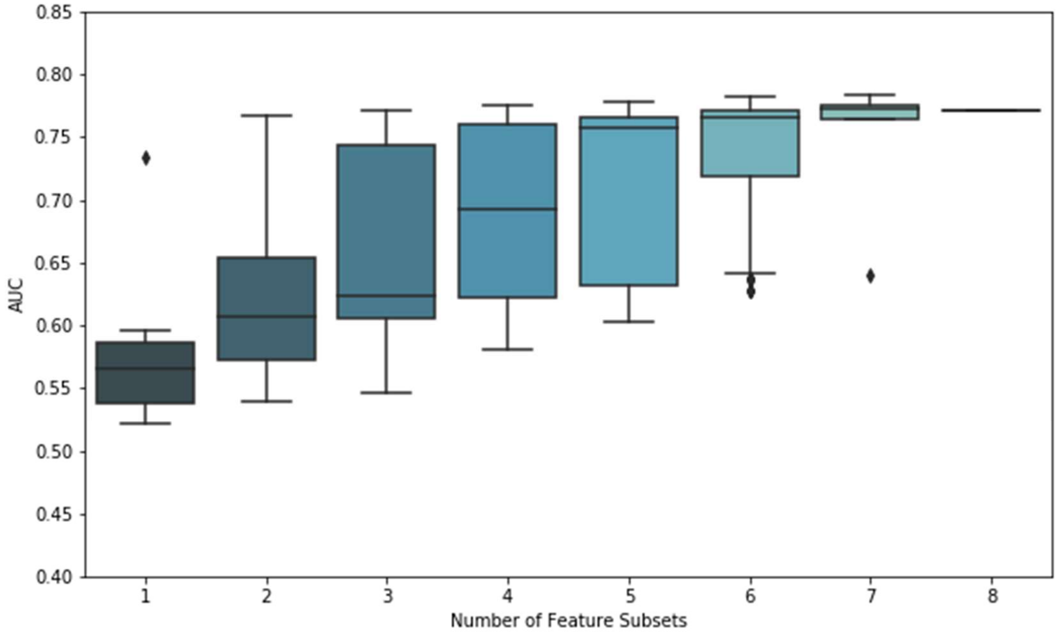

Figure 4. Distribution of AUC Scores by Feature Subset

From Figure 4, we can derive that nearly all combinations of 6 to 7 feature subsets achieve comparable accuracy to the full feature space. The one exception –the outlier in Figure 4 observed at 1 and 7 feature subsets, respectively– corresponds to personal information with minimal variance over time. Besides lending credence to the model's potential usefulness in a production environment, the study affirms that the feature engineering techniques used to control for temporal leakage were successful in removing individual features that could singularly and artificially inflate the AUC score.

DISCUSSION AND NEXT STEPS

The accuracy achieved in Experiment 1 highlights the potential of using machine learning to assist with mobilization readiness. This potential is even more remarkable given that most successful models are trained on datasets that are larger than the test dataset, which was not the case in this study. Any decisions made on the basis of this analysis, however, should factor in the narrow window during which the experiment was performed, as variation in mobilization trends between years may lead to changes in accuracy. The key benefit of the model is the strong foundation laid for additional years of back-testing and model training, with preliminary evidence suggesting that such continuation will increase model accuracy as new fiscal years of data are added to the training set.

The sturdiness the model observed in Experiment 2 mitigates concerns of feature dominance and data leakage. It also highlights that a comprehensive data strategy, weighing manifold behavioral and personal records, is essential to predict behavior accurately, with no subset of related factors substantially influencing predictive results. Finally, the novel feature-engineering techniques employed in the study comprise a robust framework for parsing signal hidden in otherwise indecipherable data fields, broadening the realm of the possible for the types of problems that can be solved using data science in the Navy Reserve.

ACKNOWLEDGEMENTS

The domain knowledge needed to conduct this project would have been unattainable without the guidance of experts in N35 (Force Mobilization). Additionally, the team would like to thank N6 (Communication and Information Systems) for guiding us through the proper channels to obtain the data for this analysis in an efficient, professional, and secure manner.