Modeling Air Force Retention with Macroeconomic Indicators

THESIS

Michelle K. McGee, 2d Lt, USAF

AFIT-ENS-MS-21-M-176

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-21-M-176

MODELING AIR FORCE RETENTION WITH MACROECONOMIC INDICATORS

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Operations Research

Michelle K. McGee, BS

2d Lt, USAF

March 25, 2021

AFIT-ENS-MS-21-M-176

MODELING AIR FORCE RETENTION WITH MACROECONOMIC INDICATORS

THESIS

Michelle K. McGee, BS
2d Lt, USAF

Committee Membership:

Dr. Raymond Hill, PhD
Chair

Dr. Bruce Cox, PhD
Member

AFIT-ENS-MS-21-M-176

Abstract

Officer retention has been a longstanding problem for Air Force leadership. Both intuition and previous research suggest economic and demographic factors play important roles in an officers decision to separate from service. Leading economic indicators are nationally reported statistics that tend to be predictive of where the economy is heading. This work targets the research gap of how leading economic indicators explain and predict attrition. Due to the noisy, complex nature of the data, the model had varied success in accurately predicting future attrition rates. As a result of this research, the current models can incorporate these findings to become more targeted and precise. Leadership who make key decisions regarding critically-manned career fields, career-specific retention bonuses, force shaping, consolidation measures, etc. will have higher-quality, better scoped results.

This research is dedicated to my family, especially my parents, for their constant love and support all of my life.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my thesis advisor, Dr. Raymond Hill, for his mentorship and guidance throughout my research process. Additionally, I would like to thank my classmates and friends for their assistance and camaraderie throughout the program.

Michelle K. McGee

Table of Contents

Page

List of Figures

List of Tables

MODELING AIR FORCE RETENTION WITH MACROECONOMIC INDICATORS

I. Introduction

1.1 Background and Problem

The United States Air Force has been battling the challenge of shaping the size and strength of its personnel since its inception in 1947. As an all-volunteer force, the Air Force must maintain the health and readiness of its force to be fully mission-capable, while also balancing the ebb and flow of the accession, promotion, separation, and retirement cycles for its officer corps. Upon accession, every officer incurs an Active Duty Service Commitment (ADSC) of between four to five years depending on the commissioning source. The structure of the officer workforce is comprised of Air Force Specialty Codes (AFSCs), which define specialties, particular skill sets, and experiences throughout an officer's career path. Furthermore, more specialized career fields, such as Pilot or Doctor, require additional years of service commitment for the longer training and financial investment.

Headquarters Air Force (HAF) oversees the entire process of officer recruitment, training, development, and career progression, using sustainment models as a key tool. Their current sustainment models seek to manage each specific AFSC's health and readiness based on commissioned year groups for a time horizon of 20 to 30 years. Historical attrition rates feed into the model to identify problematic trends for specific AFSCs in order to dictate the need for more or less recruitment, retention bonuses, or general force shaping. The longstanding challenge has been maintaining healthy

levels of officers, especially as years of service (YOS) increase and officers complete their mandated ADSC. Moreover, turnover occurs regularly with promotions and retirements since the military structure is strictly hierarchical based on YOS. The time at which obligations are completed is a flash point for eligible officers to determine if they want to remain in the service or separate. Ongoing research to precisely identify the best variables and conditions that can confidently predict future retention trends has been a priority for HAF. A goal of this work is to identify factors to anticipate attrition so the sustainment models can become more accurate and targeted.

1.2 Research Questions

Both intuition and previous research suggest economic and demographic factors play important roles in the personal decision to separate from service. Intuition follows that economic booms lead to higher officer attrition since there are readily available jobs that incentivize officers to leave Active Duty. Conversely, officers are less likely to separate during economic recessions or downturns since there are less available jobs in the private sector. It would be useful for sustainment models to include economic factors. This research explores the predictive capability of various external economic and demographic factors that are hypothesized to influence officers' decision to separate from service. Moreover, it seeks to predict future retention rates for officers by taking into account such external information. The key external variables are called leading economic indicators, which tend to be predictory in nature of future behavior or trends. Current sustainment models use YOS as the primary variable for explaining and predicting officer retention. This thesis determines if there is an alternative variable, or set of variables, to YOS as a metric in retention models to help produce more useful results. This research proposes that if some or all of the examined variables provide better insight into attrition, then it would be possible to

better anticipate future retention trends and adjust recruitment and sustainment as needed. Leadership can capitalize on that information when determining career sizes, retention bonuses, and overall force shaping programs.

1.3 Limitation of Work

This work attempts to capture and predict human behavior and decision-making via a model containing personal and economic attributes. Naturally, this is difficult due to the innate irrationality and unpredictability of people. There are too many underlying factors in real-life that are nearly impossible to pinpoint and define in a mathematical model. However, leading economic indicators and certain demographic patterns aim to capture some of those unobservable traits, and give insight at least at some macro level.

1.4 Thesis Organization

This chapter outlined the general motivation and scope of this thesis. Chapter 2 provides a literature review on relevant past research done on officer retention. Economic-focused research is particularly highlighted. Additionally, it discusses important economic concepts that play a key part in this thesis. Chapter 3 focuses on the methodology, data attainment and preparation, and explains the techniques used to conduct analysis. The two main techniques employed are regression analysis and multivariate analysis. Chapter 4 presents the results and analysis of both the regression and multivariate techniques. Chapter 5 concludes the thesis and recommends follow-on work for future research.

II. Literature Review

2.1 Overview

Maintaining a cadre of quality officers with the proper mix of experience and rank is a perpetual concern for the Air Force. As such, there have been ongoing studies aimed at identifying the underlying causes and recommending solutions to rectify the retention problems. More recently, an economic-focused approach has emerged as an important part of the conversation regarding retention and the modeling of retention. The intuition is that the health of the economy is a factor in individual officers' decision to stay in or separate from military service. Thus, understanding the trajectory of the economy may shed light on upcoming attrition trends over the officer corps or subsets of that corps. However, the extent of this relationship has not been fully explored. This section summarizes previous studies, and their models, focused on military officer retention, both in a general sense and from an economic perspective.

2.2 Officer Retention Problem

Maintaining a strong and healthy officer corps is essential to the success of the Air Force. Not only does this ensure combat readiness and the continued projection as a world superpower abroad, but also it ensures American freedoms and way of life domestically. Officer personnel sustainment has long been a concern for Air Force leadership because there is the constant ebb and flow of officers joining, separating, and retiring that impacts the total size of the officer corps. The longstanding challenge has been to influence and predict future fluctuations in officer retention to ensure a strong and healthy force. The research on external factors' impact on officer retention is a more recent shift that is reviewed in this chapter as the basis for the

subsequent work. Such external factors include the various measures of the health of the economy which may be combined with individual demographics. The bulk of this recent research focuses on the health and future projection of the economy, and how these measures have been examined as influencers of retention.

2.3 Macroeconomics Review

The National Bureau of Economic Research is a private, non-partisan research organization that facilitates research and publication of major economic issues [24]. Economists at the NBER use the business cycle as a measure of the state and strength of the economy. A business cycle is defined by NBER as "peaks and troughs that frame economic recessions and expansions" [25]. In this context, economic low points are defined as troughs and the economic high points are called peaks. A chronology of the U.S. business cycles, characterized by unemployment rate, is shown in Figure 1. The gray shaded areas represent periods of recessions.

Figure 1. Unemployment Rate Business Cycle. Graph acquired from NBER's business cycle web page [25]

Macroeconomists, experts in the U.S. economy at large, have struggled to precisely model and forecast the business cycle, made difficult due to its complexity and

innate unpredictability. However, indicator variables are a measurable and consistent proxy for evaluating economic strength and projection. Unemployment rate is a top indicator variable for NBER's chronology of U.S. business cycles because it is reported on a monthly basis. Statistics reported on a monthly basis are preferred over those reported quarterly.

In economics, indicator variables are categorized as lagging, coincident, or leading indicators. Lagging indicators have a delay before their effects are observed in real-life. Unemployment rate is one such example because the measurable economic impacts via applications for unemployment benefits are observable after-the-fact. Generally, lagging indicators are not desirable for prediction purposes because their observable impacts occur too late. Coincident indicators are observed in real-time and provide perspective to the current health of the economy. Coincident indicators are best used in conjunction with both lagging and leading indicators to identify economic trends or patterns. A leading economic indicator can be used to predict future economic trends and the direction the economy is heading. If the goal is prediction, leading economic indicators are preferred over coincident or lagging indicators [1]. Leading economic indicators are particularly important right before a drastic shift in the economy since they act as the first data points of a new business cycle [10]. Intuitively, the state of the economy plays a role in officer retention because it is a strong factor in the availability of jobs in the civilian job market. Previous work has already concluded the positive relationship between a strong national economy and increased officer attrition. Such work is discussed next.

2.4 Relevant Previous Work

An early modeling effort that examined economic influences is the agent-based model introduced by Gaupp [11] and published in Hill and Gaupp [19]. Their agent-

based model was used to examine the propensity of pilots to remain in the Air Force or separate based on personal as well as socio-economic influences. The model was primarily a proof-of-concept model. The computer simulation findings implied retention is largely determined at an individual level and that retention efforts should focus on the training and work environment, rather than targeting individual attitudes.

Schofield [26] explored attrition for non-rated Air Force officer AFSCs based on logistic regression and survival analysis using personnel data from January 1999 to December 2013. Logistic regression is a predictive modeling tool that uses a logit transformation of the binary (stay in service, get out of service) response variable. In general, logistic regression is deemed superior to ordinary least squares regression for binary responses because it involves the more appropriate binomial assumption on the data. Survival analysis is useful when the time of a specific event is of interest [9]. For retention, the event of interest is the service member's decision to separate or remain in service upon ADSC completion.

Schofield [26] used a binary response variable of officer retention and applied logistic regression to identify the important underlying variables. YOS is an important factor in her model. ADSCs often expire at the four to five year mark or ten to twelve year mark depending on non-rated or rated career fields. The results indicated that more YOS, distinguished graduate status, and career field were all indicators of a higher probability of retention. These significant factors were then used in a survival analysis model. The main takeaways from Schofield's research were that females are more likely to separate, and distinguished graduate status, and number of years of enlisted service are significant indicators for predicting retention [26]. There was no inclusion of economic variables in Schofield's work.

Jantscher [20] studied economic factors that were possibly influential to officer retention. She focused on economic metrics using a correlation analysis to determine

the relationship between those variables and AFSC retention rates. Her data sources were personnel data from 2010 to 2014 for 100 different AFSCs in conjunction with nationally maintained economic databases, such as the Bureau of Economic Analysis (BEA) and the global data aggregating company Quandl Financial and Economic Data. Her conclusions followed intuition in that retention rates tend to decrease when the economy is strong, except for the career fields of Intelligence and Chaplains. However, her goal of creating a regression model to predict future AFSC retention with updated data was not successful. There was severe multicollinearity in her model, hindering its application for future use [20].

Franzen [10] used a survival analysis regression model to identify six demographic factors (marital status, gender, commission source, prior service, distinguished graduate, and dependents) and one economic factor (New Orders value from the Advance Durable Goods Report) to help predict rated officer retention. Realizing that the point of ADSC expiration is when economic factors have the most impact on the service member's decision to remain in service or separate, Franzen created a model to help predict and anticipate future retention rates of Air Force rated officers [10]. Franzen's recommendation for future work included examining the New Orders values from the Durable Goods Report from the U.S. Census Bureau because it is a leading economic indicator [5]. Thus, one could extrapolate future New Order values as a proxy for economic strength [10].

Elliott [7] conducted an econometric analysis on Air Force officer retention, focused on the formal relationship between various indicators of the economic environment and officer personnel attrition in the Air Force. Elliott examined non-rated officer retention using AFSC attrition data from October 2004 to September 2017 and various economic indicators from the Federal Reserve Bank of St. Louis - Federal Reserve Economic Data (FRED) database. He used Auto-Regressive Integrated Moving

Average (ARIMA), exponential smoothing, and regression to facilitate his analysis. He accounted for the time variables such as unemployment rate by introducing lag-periods of 0, 6, 12, 18, and 24 months. All of the top performing forecasting models included the unemployment rate variable lagged by 24 months. His results reinforced the notion of the relationship between the lagging unemployment rate variable and officer attrition (such as suggested in Jantscher [20] and Franzen [10]). He also identified that that relationship has a two year time lag. He suggested future work in this area as investigating possible differences across AFSC groupings [7].

Pujats [23] created a model to forecast Air Force officer retention by AFSC as an extension to Elliott's work. His study focused on how the civilian sector hiring cycle predicts and forecasts upcoming officer retention. The data spanned the time frame from November 2004 to September 2017 and included attrition by AFSC, internal personnel data, and external economic data. Eight AFSCs (11X Pilots, 17D Cyber, 31P Security Forces, 32E Civil Engineers, 61A OR Analysts, 62E Engineers, 63A Acquisitions, and 64P Contracting) were chosen based on their relevance to jobs in the civilian sector, how critically manned they are, and if they are an emerging career field expected to have rapid growth in coming years. Methods of analysis included multiple linear regression and ARIMA forecasting. His analysis showed varied results in predicting attrition by AFSC and suggested further research is necessary to come to a concrete conclusion. Pujats identified two key indicators for predicting attrition: YOS and individual demographics. He emphasized quality economic data as of paramount importance. The AFSC attrition data was monthly but the economic data was often quarterly or yearly. This lack of data alignment leads to the problem of incorrect or exaggerated aggregation. Overall, he suggests that future attrition can be better modeled and predicted with the combination of individual demographic factors and introducing more factors into the model to explain more of the variability in

attrition.

Patterson et al. [22] studied the effects of retention efforts and retirement programs on labor supply in the U.S. Army for mid-career soldiers on both the enlisted and officer sides. Their results validated the concern that the highest educated, most skilled, and talented soldiers are separating at a rate much higher than their counterparts. Their results showed that low-ability soldiers, as designated by their Armed Forces Qualification Test (AFQT) score, and a proxy variable for the speed of a soldier's promotions, are more responsive to lump-sum retention bonuses and early retirement benefits. The AFQT score is regarded as indicative of overall military performance and individual knowledge in a particular career field, but fails to account for all dimensions of military aptitude. The proxy variable for speed of promotions helps provide a more comprehensive perspective for aptitude.

Patterson et al. [22] emphasized the structure of commonly accepted retention and retirement programs seeking to maintain a highly skilled military workforce is not reaching all echelons of talent in the Army. This study concludes that the Army is failing to retain its highest-performing and most talented soldiers because this subset of officers is largely unaffected by seemingly lucrative retention incentives and retirement benefits. Their results emphasize the important fact that there are no one-size-fits-all incentive retention and retirement programs, particularly among career fields that require more technical skills or higher-level education. The authors suggest that retention programs should be more targeted and refined for particular groups to retain higher-ability and skilled service members.

Hanson and Nataraj [12] present and discuss results from research projects relating to economic trends in the private sector and U.S. Army officers' perceived notions of civilian versus military employment. Their findings underscored the necessity of adequate education on the unique differences in employment between the public and

private sector. There is a general concern that active duty military members experience "optimism bias" when considering the trade-offs of transitioning from active duty military to private sector employment. In psychology, optimism bias occurs when an individual overestimates the probability of positive outcomes and underestimates the probability of negative outcomes [12]. This materializes in officers having unrealistic expectations about easily finding comparable civilian careers, the additional costs of those jobs that the military covers (i.e. healthcare, relocation, etc), and retirement benefit programs. Essentially, officers are ignorant of the socioeconomic differences between military and civilian employment because they are unaware of the additional costs of civilian employment. Hanson and Nataraj [12] explain such costs as health-care benefits, underemployment risks, general unemployment, and retirement programs. The total magnitude of such differences is much greater than one presumes due to lack of understanding. The goal is to eliminate the "grass is greener on the other side" mentality. The authors conclude that effective and more complete communication before an Army officer candidate commissions is the best way to ensure a more informed perspective and, ultimately, help improve officer retention.

Interestingly, there does not appear to be any widespread effort to disseminate information about military versus civilian career benefits among Air Force officer candidates in a formal educational program. The creation and adoption of such a program among Reserve Office Training Corps (ROTC), Officer Training School (OTS), and United States Air Force Academy (USAFA) military development and education may lead to greater success in retaining the highest-performing officers. As a result of such education, officers will possess well-rounded knowledge of the realities of military versus civilian careers before even entering active duty and will have that critical perspective long before making the personal decision to separate or remain in service.

This literature review examined work that specifically linked military retention

studies with economic impacts or explored the relationship between economic health and retention. While retention studies have long been a topic of military interest, often to determine widespread policies and programs, the economic factors that may be driving or underlying such retention trends have remained largely unexplored. Nevertheless, the work conducted to date in this vein are useful and provide recommendations for continuation.

This work is a follow-on to previous work conducted on macroeconomic focused impacts on officer retention. Moreover, inclusion of demographic variables to identify trends in attrition are considered. This work continues the work of Elliott [7] and Pujats [23] work by combining an economic and demographic focus on officer attrition. The Patterson et al. [22] study suggests more tailored, individualized retention programs according to career fields, so categorization according to AFSC is a focus in this research. Certain leading economic indicators are used as proxies for economic strength and examined for their effects on future retention, as suggested by Franzen [10]. This approach is further explained in the next chapter.

III. Data Preparation and Methodology

3.1 Overview

This chapter outlines the methodology, data attainment and preparation, and an explanation of the techniques applied to conduct the linear regression and multivariate analysis. Multiple sources provided relevant data sets so extensive cleaning and manipulation led to a uniform master data set. All subsequent analysis used this master data set. Detailed, specific results and conclusions are presented in later chapters.

3.2 Data Attainment and Preparation

One goal of this analysis was to examine which, if any, leading economic indicators may be applied as a predictor for Air Force Officer retention models. The focus variables are leading economic indicators because they provide reliable insights regarding both the state of the current economy and where the economy may be headed in the near future.

Leading economic indicators considered in this research include the New Orders on Durable Goods, Consumer Confidence Index (CCI), Military Expenditures, and the Business Formation Statistic. CCI is a survey administered to American households regarding their economic expectations both in the present and for the near future. CCI is widely regarded as a leading economic indicator because it measures the health of the U.S. economy based on peoples' perceptions and attitudes towards businesses, employment, and income both presently and for the next six months. Economists from the Organization for Economic Cooperation and Development (OECD) regard CCI as the most credible gauge of U.S. consumer confidence [8]. The Advance Report on Durable Goods Manufacturers' Shipments, Inventories, and Orders is a survey pro-

vides the monthly figures on manufacturing activities [21]. The New Durable Goods Orders is used as a proxy for economic strength because increased goods tends to mean higher productivity of labor, increased employment, and more consumption. New Durable Goods Orders is regarded as a reputable leading economic indicator because it provides an indication of future business trends. Intuition follows that increased New Durable Goods Orders indicate economic growth and decreased New Durable Goods Orders signify economic stagnation or downtown. New Business Formations is a survey that reports the number of new U.S. business applications [4]. It is also a proxy of economic strength because more applications tend to mean higher confidence in the economy and vice versa, and it captures the entrepreneurial spirit of the people and the risks people are willing to take based on how they perceive the economy. Lastly, military expenditures are the total amount of spending spent annually on all defense-related ventures. Ongoing research is inconclusive on whether the link between economic growth and military expenditures is significant [2]. Pujats [23] found the variable insignificant in his research. Based on his findings, military expenditures was not included in this analysis. The three leading economic indicators were evaluated in conjunction with Air Force officer personnel demographic data. Evaluating the existence and strength of any relationship between them, and the dependent variable of separation rate, was one of the goals of this work. There are multiple sources of data used in this analysis:

- 2004-2019 personnel data from Headquarters Air Force/A1 is the primary source for the bulk of the data. This data contains important separation and attrition information including date of separation, AFSC, grade, etc. Hundreds more demographic metrics were available from the personnel data. In previous research, obtaining this data was a challenge. Access to this individualized big data is valuable to this research. All personally identifiable information was scrubbed

from the analysis. The demographic features included in this research were age, sex, race, marital status, number of dependents, degree groups 1-4, source of commissioning, long overseas tours, short overseas tours, Masters' degree, and PhD.

- St. Louis Fed - FRED is a nationally maintained database of major economic data. Both CCI and military expenditures are found from this source. CCI is a normalized value around 100. Another major economic indicator from St. Louis Fed - FRED database is military expenditures. However, that data were excluded from this research.

- The U.S. Census contains the seasonal Advance Report on Durable Goods Manufacturers' Shipments, Inventories, and Orders [21] and New Business Formations [4]. For this analysis, both statistics were seasonally adjusted, which means they are computed as compared to the previous quarter's activity. Monthly data are preferred since it is more detailed but that data were not accessible with these two indicators.

- Elliott [7], followed by Pujats [23], compiled a master data set including separation rate by month by AFSC, monthly national macroeconomic indicators, and employment rates for civilian sectors most comparable to the key AFSCs studied from 2004-2017. Pujats [23] accounted for missing observations in this data set using the k-Nearest Neighbor algorithm for data imputation. This algorithm uses 31 of the missing observation's nearest neighbors and a weighted average to estimate the missing data. This technique did not introduce significant bias since there were very few missing observations and the dimensionality was relatively small. Moreover, Pujats [23] accounted for irregular economic events in the time period between 2004-2017, such as the Great Recession from

2007-2009, by curtailing data for specific AFSCs. This thesis implements and builds upon that master data set. The AFSCs, separation rates, and time span from the previous 2004-2017 data set are useful. However, based on Pujats [23] conclusions, all of the economic variables and civilian hiring data are eliminated. He did not find significant trends among the economic factors he included, but he did confirm the suspicion that comparable civilian hiring trends play a significant role in retention. The leading economic indicators discussed earlier in this section were added to this master data set. Moreover, the timeline expands to include the most recent 2018-2019 retention and economic statistics. Data imputation on a large scale was not necessary since there were not many missing observations.

The main aspect of the data cleaning process was modifying the Elliott [7] and Pujats [23] data and properly formatting the HAF/A1 data to curate a master data set. The external economic statistics were added by month-year to the HAF/A1 data. The HAF/A1 data contained hundreds of demographic statistics, personal identifiable information, and career progression variables. The initial cleaning removed any variables containing sensitive personal identifiable information such as names, dates of births, addresses, social security numbers, etc. Although the focus was on macroeconomic variables, a select group of demographic variables were kept for analysis. These are grade, age, sex, race, marital status, number of dependents, degree type and number, source of commissioning, number of short and long overseas tours, and higher-level education degree attainment. Further elimination was necessary before formulating the initial model. Those measures are discussed later in this chapter.

Elliott [7] and Pujats [23]'s data were categorized by specific AFSC separation rates while the HAF/A1 data had individual observations that represented whether an officer separated or remained in the service. The monthly AFSC separation rates

were calculated for 2018-2019 and added to the Elliott [7] and Pujats [23] data. All observations that were not separations were eliminated from the HAF/A1 data. The date were reported as date-month-year, but it was modified to month-year to match the Elliott [7] and Pujats [23]'s date format. To combine the various data sets into the master data set, observations were matched by both month-date separation and AFSC. Thus, the resulting master data set contained individual observations of separations, grouped by AFSC and month-year dates. The individual observations included the associated demographic data while the economic data remained uniform across all AFSCs for that month-year time period.

3.3 Data Manipulation

The DoD recommends Airmen begin their separation planning at least 12 months prior to separation. Based on this timeline, all of the leading economic indicators were lagged by 12 months. The month-year separation for each observation was associated with the previous year's economic indicator statistics. For example, the September 2005 separations had the economic statistics from September 2004.

The 2004-2019 master data set was a combination of the HAF/A1 personnel data with the separation rate data set from Pujats [23] and Elliott [7]. The data were partitioned into training and validation sets. The training data set was the data from 2004-2017 and the validation data set was the data from 2018-2019. The training master data set contained 19,799 observations and the validation master data set contained 3,022 observations. The validation master data set was withheld from the analysis until a subset model from the training data was selected. Then the training data subset model was used to create a prediction interval for the validation data and evaluated.

The variables of the master data set were categorized as date, numeric, or char-

acter to ensure proper scaling and linearization. The Masters and PhD variables were changed to indicator variables, where a zero means no degree and one equals degree achievement. The scales for the numeric regressors were standardized to provide better interpretations of raw coefficient estimates and to understand the relative impacts of the units. While not necessary in linear regression, this does help with interpretation. This was confirmed by checking the ranges of the variables before and after normalization. The normalized standardization was applied to all numeric variables. This type of feature scaling for numeric variables enabled simpler and more straightforward interpretation of results.

Closer consideration of the independent variables from the HAF/A1 data revealed missing observations for certain variables initially deemed interesting and important enough to include in the regression. There was also some clear redundancies among the variables. There were two variables that reported source of commissioning. The first was the full source and the second was an assigned number representing the full source. Thus, only the latter was included in the regression as an indicator variable. Additionally, there were five variables measuring the number of degrees an individual attained. However, the pertinent information for what type of degree and the subject matter of the degree was sporadic and inconsistent across years. Thus, while this level of degree detail is useful in the analysis, it was not complete enough. Therefore, these degree variables were eliminated from model inclusion.

3.4 Linear Regression

Linear regression was the first step to determining the best model to forecast attrition trends. Such analysis evaluated the existence of any linear relationships between predictor variables and retention data. While the precise relationships among the variables may be unknown, regression provided an adequate approximation to build

upon for further study. Good data are critical input for linear regression. Regression techniques should not be applied until all data preparation and manipulation are complete. Model building began once the data preparation and manipulation step described above was complete.

This regression involved several regressors. An associated regression equation for the model roughly approximated relationships between the k regressors and the response variable, y. The general form for a multiple regression equation is shown in Equation 1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i \tag{1}$$

The equation structure is a linear function of the unknown β parameters. Thus, it is possible to apply linear regression even if the relationships among the variables are not perfectly linear. An example of interpretation of the β parameters when the variables are scaled and standardized the same way is as follows: *the parameter β_1 indicates the expected change in the response variable, y, per unit change in its associated predictor variable, x_1, when all other regressors are held constant.*

Real-life data are often very complex. When the data are based on human behavior or attributes, it becomes even more difficult to account for, and explain, unknown causes of variability among the variables. Regression analysis simplifies such features by requiring certain generalities be satisfied in order to apply its techniques and interpret results. This essential step of regression is called model adequacy checking. Another way to think about model adequacy checks is as a sanity check because it reveals inconsistencies or problems with both the individual variables and the overall model. Regression assumes normality and constant variance in the data. Normality means that the residuals are reasonably symmetrical with diminishing tails of the distribution. Constant variance means that there no noticeable changes in how well the model fits the observations over time. If one or both of these assumptions are

violated, they must be remedied. The entire model building process is iterative. Several approaches to addressing violations of regression assumptions are discussed in this chapter and applied in the following chapter.

One cannot conclude a cause-and-effect relationship between variables based solely on regression results. Closely related variables determined in a model environment does not always translate to real-life outcomes. Rather, a seemingly-close relationship among data simply implies that there is correlation and requires additional observation outside the model environment to confidently conclude results.

This research's methodology involved analyzing all variables, systematically eliminating variables deemed unnecessary, checking model adequacy, using a standard selection technique to generate three subset models, performing detailed analysis on those three models, and finally concluding the advantages and disadvantages of each subset model before selecting one for the final forecasting ability and multivariate analysis for predictive power.

3.4.1 Full Model Evaluation

The first step was to analyze a full model including the macroeconomic and demographic variables. This process included finding the summary statistics, checking model adequacy, and examining the variance inflation factors (VIFs). The summary statistics include R^2, t-statistics, and p-values. R^2 is the coefficient of multiple determination, which is a measure of the fit of a regression model. It describes how much the regressors explain the response variable. The regressors' t-statistics are used in a t-test to determine individual regressors' significance in the model. The model p-value and individual regressors' p-values further determine fit and significance. A p-value less than 0.05 generally indicates significance. Interpreting the collective statistics convey a multitude of important takeaways of the model. The magnitude of the

variables' coefficients indicate how significant they are in the model. The VIFs are a multicollinearity diagnostic. Multicollinearity means there is correlation among the independent variables, which, among other issues, decreases the degree to which the individual regressors explain the variation of the dependent variable. The conservative VIF cutoff of five is used for this analysis. Multiple variables with large VIFs indicate multicollinearity in the model. Multicollinearity requires removing some independent variables from the model.

Residual analysis is a primary method for checking model adequacy. Model adequacy diagnoses the basic regression assumptions of linearity, zero mean error terms, constant variance, and uncorrelated errors. The normal Q-Q plot and various residual plots provide a check on the error normality and constant variance assumptions. For normal Q-Q plots, the observations and the normal distribution line should be closely aligned. A straight, approximately 45 degree line supports the hypothesis that the data are normally distributed. The residuals versus predicted values plot shows the relationship between the residuals and predicted response. The variance of the errors, or spread of the residuals, should be constant across the observations. There are common patterns of concern to check for in the residual plot, such as double-bowing or fanning, that suggest a violation of the constant variance assumption or some effects not yet defined in the model. When such patterns are discovered, remedial measures are used to resolve the assumption violations.

3.4.2 Transformations and Skewness

Variable transformations are necessary when there is nonlinearity and/or nonconstant variance in the model. This is indicated by unusual patterns in the normal Q-Q and residual plots. If the issue is a violation of linearity, the response variable can be manipulated with a function such as square root or natural log to attempt to

reconcile the response and regressors to form a more linear relationship. Skewness in the model is generally part of deviations from normality. Skewness is identified by examining the residual quantiles from the model summary statistics. The mean of the residuals should be close to zero with the distance to the minimum and maximum similar in magnitude for skewness to not be present. If this is not the case, there are likely outliers skewing the residuals. Additionally, it is possible to find the skewness value for the overall model. The widely accepted cutoff for skewness is a value below 1.0. If the skewness is greater than 1.0 or less than -1.0, this informs the analyst of whether the skewness is positive or negative and how to proceed to remedy the skewness. For nonconstant variance, variance stabilizing transformations result in a more precise estimate of the model parameters and increases sensitivity for the statistical tests. After a transformation is applied, it is necessary to examine the improvements to the model by examining common plots again and R^2 value. In general, transforming the response is considered when there are both linearity and constant variance violations. Conversely, transforming the regressors is considered when the only issue is the linearity assumption.

3.4.3 Outlier Analysis

Outlier analysis is important. Unidentified outliers may negatively influence the model. An outlier is defined as an observation with an usual x and/or y value. As a result, these type of points noticeably impact the model coefficients since they "pull" the regression model in their direction [6]. Influence points have unusual x and y values. Leverage points are observations with seemingly unusual x values. Not every unusual observation is an outlier requiring removal. Analyst expertise is necessary to determine the best course of action to treat, eliminate, or leave potential outliers in the data. Outlier analysis identifies the existence of potentially problematic points and

helps determine if those can be attributed to issues with data collection or external factors outside of the model environment.

Outlier analysis is conducted on individual observations or in groupings. This is related to attrition data because it is more insightful to look for general clustering of leverage or influential points that seem to capture groups of observations. Once identified, the analyst can determine the best action to take such as discarding such points altogether, weighing them differently than the well-behaved observations, or keeping them in the model as is. Nevertheless, valuable insight is gained from this effort in the model building and checking process.

In this thesis, the goal was identifying if clusters of outliers exist and what (if any) external factors coincide with them. For example, perhaps unusual patterns are observed during distinct economic downturns or booms, AFSC consolidation measures, or targeted retention bonuses. The inclusion of outlier analysis benefited the regression because it allows attribution of variability.

The R feature that reports influence diagnostics for any given model was the primary tool for outlier analysis. The R output, along with general knowledge of the data set, helped to decide what action to take on those potential outliers.

3.4.4 t-tests

Once the model is fit with all necessary regressors, a thorough residual and outlier analysis completed, and any necessary transformations applied, the final step is the examination of the t-tests on the individual regressors is the final step to the full model analysis. Each t-test considers a specific null hypothesis, H_0, defined as a regressor's coefficient being equal to zero. The alternative hypothesis, H_A, is defined as each regressor's coefficient not being equal to zero. Equation 2 provides the t-test

hypothesis.

$$H_0 : \beta_i = 0 \qquad \forall\ i \tag{2}$$

$$H_A : \beta_i \neq 0 \qquad \text{for some } i$$

The t-critical value was established and then each individual regressors' t-value was compared to the t-critical value. If the absolute value of the regressor's t-value was greater than the t-critical value, that suggested that the regressor was statistically significant in the model and helps to explain the dependent variable of *Separation_Rate*.

3.5 Linear Regression Equation

Table 1 contains the full list of data considered in the model building process. The full analysis finds a smaller subset of data that best predicts retention. The linear regression evaluated if and how much each of the independent variables influence the dependent variable of *Separation_Rate*. The last three variables of the table were included for data cleaning and manipulation purposes only, but they were not included in the regression as independent variables. Variables that were eliminated from the regression because there were too many missing observations or redundancies are noted as well. These are the five degree variables, x_{13} through x_{17}, and the long form variable for source of commissioning, x_{18}. The regression model summary will include the coefficient, β_i, for each variable.

Table 1. Full Variable List

Name: Class	Variable Definition
AFSC: categorical	x_1
Date of Separation: date	x_2
Grade: numeric	x_3
Years of Service: numeric	x_4
New Business Formations: numeric	x_5
New Durable Goods Orders: numeric	x_6
Consumer Confidence Index: numeric	x_7
Age: numeric	x_8
Sex: categorical	x_9
Race: categorical	x_{10}
Marital Status: categorical	x_{11}
Number of Dependents: numeric	x_{12}
Deggrp: eliminated	x_{13}
Deggrp1: eliminated	x_{14}
Deggrp2: eliminated	x_{15}
Deggrp3: eliminated	x_{16}
Deggrp4: eliminated	x_{17}
Source of Commissioning Long: eliminated	x_{18}
Source of Commissioning: categorical	x_{19}
Long Overseas Tour: categorical	x_{20}
Short Overseas Tour: categorical	x_{21}
Master's Degree: categorical	x_{22}
PhD Degree: categorical	x_{23}
Fiscal Year	x_{24}
Number Assigned	x_{25}
Number Separated	x_{26}
Separation Rate	$Separation\_Rate$

Equation (3) is the regression equation produced from the full model analysis.

$$Separation\_Rate = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{23} x_{23} + \epsilon_i \tag{3}$$

3.6 Reduced Model Analysis

Various subset models containing various combinations of regressors were examined next. The R function dredge() effectively compares and selects a subset model of the most effective variables based on a given criteria. The output automatically computes several important statistics in evaluating subset regression models. R^2 measures how much variability in the response is explained by the regressors. When

comparing models, R^2_{adj} is a better statistic than R^2. The R^2_{adj} penalizes the R^2 value when insignificant terms are added to the model. According to the R^2_{adj} criterion, the model with the maximum value is the best. Mallow's Cp is another useful statistic that measures model capability and parsimony. Smaller values of Cp, close to the number of parameters, p, in the model are preferred. The Akaike Information Criterion (AIC) is a measure of model parsimony, where a lower value indicates a model with less variables. The key insight for AIC is the nuances of the number of regressors versus the expected information that each contributes to the model. The number of regressors in the model is included in calculating AIC. The AIC will penalize models that include more regressors that bring minimal valuable information compared to other models [6]. Since the end goal of the final model is maximum predictive power, the PRESS statistic is another important factor in evaluating subset models. The PRESS residual is the residual, e_j, that results when x_j is not included in the model, sort of a hold-out validation measure. The PRESS statistic involves all PRESS residuals. PRESS sheds light on which regressors have small effects that are not very valuable to the model. Often times, PRESS is used to compare alternative subset models because it discriminates between alternative models. For PRESS, a lower value means a better predictive model.

Additionally, the F-test is a global test to check if the model has significant variables. It is used to evaluate the subset models. Each model's F-statistic is compared to a critical value and their corresponding p-values indicate whether the overall model is significant. This indicates significance of some or all of the terms in the model.

In addition to the subset regression model statistics, model adequacy checks ensure there are no violations of the common regression assumptions. Such model adequacy checks are performed for each of the three subset models to confirm they are well-behaved. Any number of models could be selected for comparison. This work

evaluated the top three models generated by the dredge command.

3.7 Prediction for Validation Data

Model validation helped determine the success of the final model in correctly predicting retention. The data were split to include a training and validation set. The training set was the data for 2004-2017. The model validation data was for 2018-2019. The final step of the methodology tested if the validation data correctly identifies the predicted *Separation_Rate* within the 95% prediction interval. Prediction intervals provide a range of reasonable values that the true value likely falls into 95% of the time [3]. For example, in this research, the prediction interval provided a range of values based on the model that likely encompasses the true separation rate with 95% confidence.

A prediction interval was produced for 20 different AFSCs for 2018-2019. Breakout by AFSC is insightful because often there are patterns or trends that are career field specific. Pujats [23] emphasized that future work should be AFSC specific as he found that attrition varies by AFSC. Examining tailored relevant attrition data patterns and trends allows leadership to make the best possible informed decisions on incentive programs and force structuring by career field.

3.8 Multivariate Analysis

Multivariate analysis explores relationships among data consisting of multiple variables. The model types can be explanatory, predictive, prescriptive, or descriptive. The scope of this analysis was predictive and built on the explanatory linear regression analysis. Some of JMP's predictive techniques are applied to the master validation data set to test the power of the selected subset model. The multivariate analysis tools in JMP Pro 15 are used. For this analysis, classification trees and forests, neural

networks, and predictor screening are considered. This section summarizes each of the methods used in this analysis.

For each method, the 2018-2019 data was split into training and validation data based on a holdback methodology. The holdback methodology randomly splits the entire data set into either training or validation. JMP's interface allows the analyst to specify the proportion of the data set to be held as the validation set. The holdback proportion was 0.2 for the entirety of the analysis.

The multivariate analysis focused on the model chosen from the model selection process from the linear regression. The overall goal of the multivariate analysis was to examine the prediction power of the chosen model from the regression analysis. Moreover, examining the model in a variety of predictive techniques further pinpoints the driving variables in the model for predictive power.

3.8.1 Classification Trees and Forests

The primary objective to using a tree-classification approach with ensemble methods is to determine how well the methods classify the data and if specific improvements can be applied. Decision trees are visual depictions that recursively partition data according to the dependent and independent variables. The algorithm continually performs the splits of the predictors until the best solution to predict the response is reached. If necessary, the analyst can define the desired number of splits. When the independent variables are continuous, such as *Separation_Rate*, the partition algorithm splits at the "cutting value" [17]. That is, the splitting criterion is based on where the significant shifts happen for the response variable. A disadvantage of decision trees are their tendency to over-fit. Over-fitting can be thought of similarly to being so precise to the data that future predictive power becomes obsolete. Decision trees that overfit the data yield lower quality predictions. Nevertheless, decision

trees are useful to explore, filter, and classify data [17]. Bagging overcomes the single decision tree's downfalls. It is also known as bootstrap aggregation. Bagging creates random samples of observations, with replacement, which are then used to build a tree. Each individual decision tree is fit using the recursive partitioning process [13]. Each decision tree is used to create the model predictions. Alternatively, boosting is another superior technique to the single decision tree. It is the additive process of building a large decision tree made up of smaller layers of decision trees [14]. As the layers build upon one another, bad fitting data is corrected to ensure higher accuracy.

Ensemble methods like bootstrap forests are the explicit use of combining various techniques to provide the most accurate, targeted results. Ensemble methods improve upon decision trees by combining multiple techniques to mitigate the single decision tree's overfitting tendency [15]. The result is better predictions than the single decision tree model. Random forests is an example ensemble method that further builds upon bagging, but ensures each tree difers in the variables used to create the splits during the building process. It creates a multitude of random trees based on the bagging technique. This ensemble of decision trees is used to create the model predictions. Ultimately, this approach results in greater accuracy and computation speed. The bootstrap forest method is based on the boosting technique and is used in this analysis.

3.8.2 Predictor Screening

Predictor screening further builds upon the classification trees and forest method. It uses a bootstrap forest to identify the best predictive variables upon the response variable [18]. For each bootstrap forest, 100 decision trees are built. The resulting report identifies the predictors that are strongest in the model. Predictor screening was used in this analysis.

3.8.3 Neural Networks

A neural network (NN) maps given inputs to a single output similarly based on an analogy to how the human brain functions as a sequence of neurons [16]. It is commonly used in classification and regression problems. NNs are useful when data are noisy and computationally intensive. For this analysis, NNs can capture the unobserved non-linear relationships in the data that a model may not be able to capture. Further, the JMP features allows examining individual variable importance within the determined NN model. This naturally leads to further steps regarding which variables to further examine in continuing work.

The NN approach for this analysis consisted of a baseline and a complex model. Activation functions are applied at the nodes of the hidden layers. JMP provides three default activation functions: TanH, Linear, and Gaussian. In simple terms, these are linear combination transformations of the regressors. The number of tours is how many times to restart the NN fitting. The penalty method is squared because this is the best method to apply when determining the regressor variables' contribution to predictive power. The baseline model uses three nodes in one layer using only the TanH function. The Linear and Gaussian functions are left at zero nodes. The number of tours was left at one for the baseline model. This baseline NN serves as a benchmark to compare to the complex model. The complex model aims to capture the complexity and non-linear nature of the real-life data by using all three functions available in two different layers. Increased layers means more flexibility within the NN. Three nodes for both the first and second layers for TanH, Linear, and Gaussian are set. Moreover, the number of tours is increased to 10.

3.8.4 Multivariate Analysis Conclusion

Many of these predictive techniques overlap in purpose. This redundancy is valuable because it confirms patterns and trends that present themselves and provides greater confidence in the legitimacy of the results. Overall, applying multiple techniques is often more insightful and helps confirm results for greater confidence.

IV. Results and Analysis

4.1 Introduction

This chapter presents the analysis conducted on the curated master data set, the prediction interval, and discusses the predictive power of the model results. The analysis goal was to provide further insight on the AFSC specific patterns and trends of officer separation from 2004-2019. Multiple linear regression selected the best possible model by highlighting the most important regressors among a variety of economic and demographic variables for the training data set between 2004-2017. Once the best subset model was identified using the classical model selection techniques, that model was used to predict the validation data's separation rate point prediction intervals. This portion of the analysis was conducted in R. The exhaustive list of economic and demographic variables is summarized in Table 1. The final model was then examined using a battery of JMP Pro 15's predictive features from its multivariate platform.

4.1.1 Full Model Evaluation

The full model evaluation with the macroeconomic and demographic variables deemed important was the first step of the regression process. Equation 3 was the basis for the full model evaluation. The summary statistics, model adequacy checks, and VIFs are presented next. Interpreting the model regression statistics in a holistic manner led to several important insights of the model.

The summary statistics are shown in Figure 2. The regression results showed the overall model was significant with the F-statistic of 146.3 and p-value less than 0.05. The coefficient estimates for the intercept, x_4, x_5, x_6, x_7, x_8, x_9, x_{12}, x_{20}, x_{22}, and x_{23} had significant impacts for predicting $Separation\_Rate$. Since all data were sealed, the magnitude of the variables' coefficients indicated how significant they were in the

model. The top three largest coefficients in terms of magnitude were x_5, x_6, and x_9.

The VIF values provided in Figure 3 indicated whether multicollinearity is a problem in the model. Any VIF values greater than five were a cause for concern. For the full model, only two variables had VIF values of concern: x_4, YOS, and x_8, Age. Intuitively, one expected these two variables to be correlated. Since YOS has been used in past studies, it was the variable retained in the model.

```
lm(formula = Separation_Rate ~ x3 + x4 + x5 + x6 + x7 + x8 +
    x9 + x12 + x19 + x20 + x21 + x22 + x23, data = all_training)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9445 -0.4832 -0.1855  0.1875 18.0238

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.133302   0.100624   1.325  0.18527
x3          -0.009844   0.013170  -0.747  0.45481
x4          -0.112609   0.020351  -5.533 3.18e-08 ***
x5          -0.236762   0.009275 -25.528  < 2e-16 ***
x6           0.218009   0.008130  26.817  < 2e-16 ***
x7           0.099399   0.008242  12.060  < 2e-16 ***
x8          -0.093735   0.019953  -4.698 2.65e-06 ***
x9M         -0.233165   0.019789 -11.782  < 2e-16 ***
x12         -0.022769   0.007644  -2.979  0.00290 **
x192        -0.033861   0.099386  -0.341  0.73333
x193         0.136007   0.098546   1.380  0.16756
x194         0.174963   0.099146   1.765  0.07763 .
x19M        -0.188775   0.669789  -0.282  0.77807
x19O        -0.022007   0.159896  -0.138  0.89053
x19P         0.056114   0.103767   0.541  0.58868
x19X         0.634223   0.254261   2.494  0.01263 *
x20          0.046215   0.006672   6.927 4.45e-12 ***
x21          0.097860   0.060675   1.613  0.10679
x221        -0.067355   0.018534  -3.634  0.00028 ***
x231         0.046423   0.018525   2.506  0.01222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9368 on 19779 degrees of freedom
Multiple R-squared:  0.1232,    Adjusted R-squared:  0.1224
F-statistic: 146.3 on 19 and 19779 DF,  p-value: < 2.2e-16
```

Figure 2. R Output of Full Model Summary Statistics

```
Analysis of Variance Table

Response: Separation_Rate
             Df  Sum Sq Mean Sq   F value    Pr(>F)
x3            1   939.9  939.92 1070.9684 < 2.2e-16 ***
x4            1   203.9  203.91  232.3412 < 2.2e-16 ***
x5            1    73.3   73.26   83.4798 < 2.2e-16 ***
x6            1   714.8  714.82  814.4847 < 2.2e-16 ***
x7            1   141.3  141.35  161.0542 < 2.2e-16 ***
x8            1    19.3   19.33   22.0301 2.702e-06 ***
x9            1   161.8  161.80  184.3629 < 2.2e-16 ***
x12           1    10.2   10.25   11.6751 0.0006347 ***
x19           7   114.8   16.41   18.6926 < 2.2e-16 ***
x20           1    41.4   41.44   47.2206 6.532e-12 ***
x21           1     2.3    2.29    2.6061 0.1064667
x22           1    10.5   10.50   11.9600 0.0005447 ***
x23           1     5.5    5.51    6.2797 0.0122207 *
Residuals 19779 17358.8    0.88
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> vif(lmall)
        GVIF Df GVIF^(1/(2*Df))
x3  3.912626  1        1.978036
x4  9.342703  1        3.056584
x5  1.940410  1        1.392986
x6  1.490896  1        1.221022
x7  1.532405  1        1.237903
x8  8.981207  1        2.996866
x9  1.113640  1        1.055291
x12 1.317957  1        1.148023
x19 1.569113  7        1.032703
x20 1.004253  1        1.002124
x21 1.002706  1        1.001352
x22 1.371188  1        1.170977
x23 1.372624  1        1.171590
```

Figure 3. R Output of Full Model ANOVA Statistics and VIFs

Both the normal Q-Q plot and residual versus fit plot raised suspicions regarding the normality and constant variance assumptions - see Figures 4 and 6. The normal Q-Q plot in Figure 4, showed a non-linear line. Figure 6 is the residuals versus fit plot, which showed a funnel pattern sloping slightly upwards to the right. The data also appeared in very apparent clusters, instead of randomly bouncing around the residual equals zero line. This suggested heteroscedasticity. A transformation of the data is one way to address the possible heteroscedasticity.

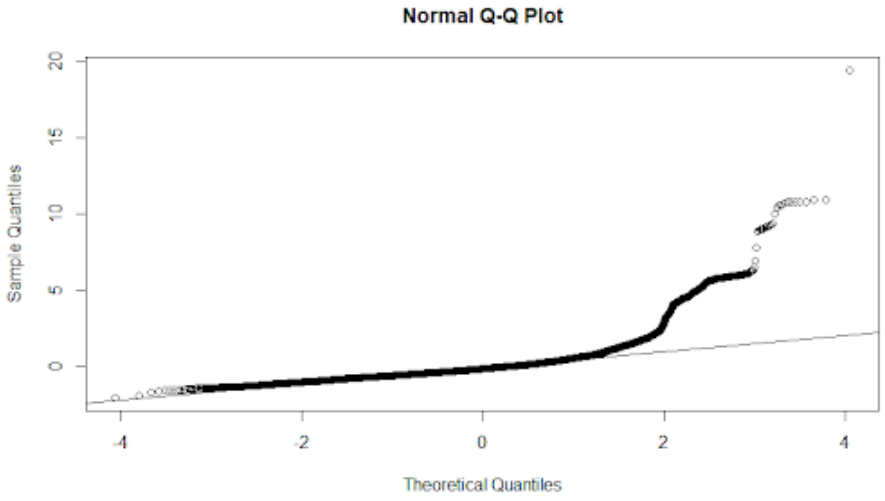


Figure 4. R Output for Full Model Normal Q-Q Plot

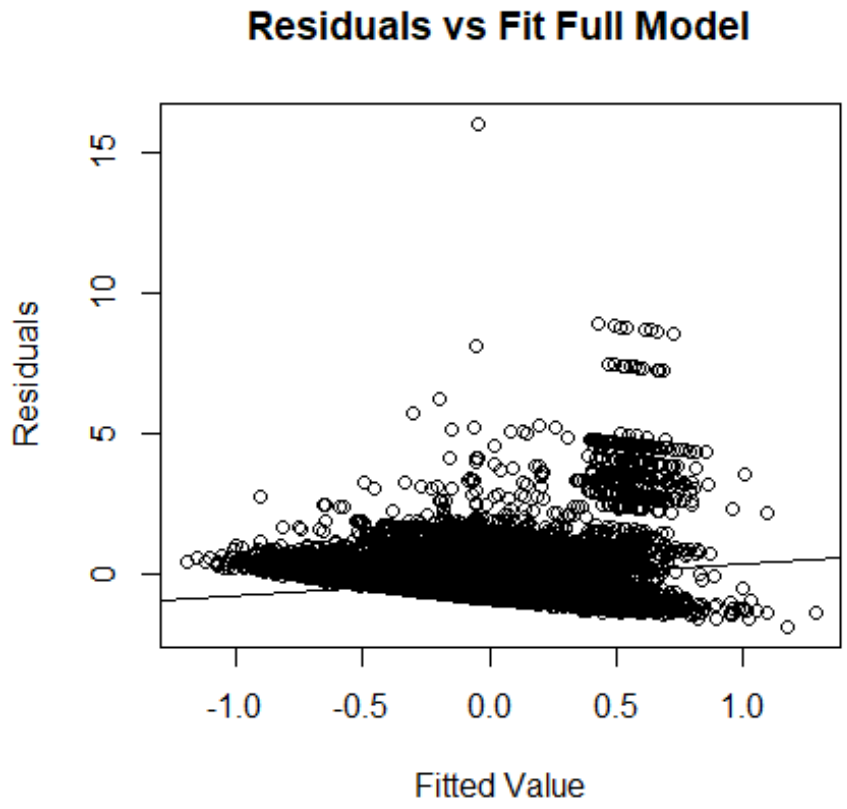


Figure 5. R Output for Full Model Residuals

4.1.2 Transformations

The residual quantiles in Figure 5 revealed apparent skewness in the full model. The absolute value of the maximum residual is significantly greater than the absolute value of the minimum. The minimum of -1.944 and the maximum of 18.024 indicate severe positive skewness in the model.

```
lm(formula = Separation_Rate ~ x3 + x4 + x5 + x6 + x7 + x8 +
    x9 + x12 + x19 + x20 + x21 + x22 + x23, data = all_training)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9445 -0.4832 -0.1855  0.1875 18.0238
```

Figure 6. R Output for Full Model Residual Plot

The dependent variable of *Separation_Rate* had a skewness greater than 1, indicating positive or right-skewness. The skewness values were determined in R using the skewness() command. The untransformed skewness of the dependent variable had a value of 4.53, significantly greater than 1.0.

Skewness among the regressors also appeared to contribute to the violation of the normality assumption. Two variables in particular, x_5 and x_6, exceeded the 1.0 threshold. x_5 had a skewness of 1.10, and x_6 had a skewness of -1.12. The rest of the numerical variables skewness values were within the acceptable range of -1.0 to 1.0.

Transformations are used to address violations of the standard linear regression assumptions and high skewness. To remedy the normality assumption violation, a log(y+1) transformation on the response was applied. Since *Separation_Rate* was scaled from -1.0 to 1.0, there are negative values among the observations. Adding the constant to the observation before applying the logarithm transformation ensures all observations are strictly positive. The logarithm transformation helps to both linearize and reduce heteroscedasticity in a model. The subsequent summary statistics and model adequacy checks are shown in Appendix A, Figures 24 through

27 for the reciprocal and logarithmic response only transformation. The reciprocal transformation on the response variable was not strong enough, so a logarithmic transformation was applied. The logarithmic response only transformation model still did not remedy the biggest issues with skewness. The individual regressors were examined for skewness as well; x_5 was transformed with its reciprocal and x_6 with the formula $\sqrt{x_{6_{max}} - x_6} = \sqrt{3.613 - x_6}$. These were applied in addition to the logarithmic transformation on the response variable. The remaining skewness values were 1.0 for *Separation_Rate*, 0.98 for x_5, and -0.22 for x_6. The resulting model is shown in Figure 7 and 8. Its residual analysis revealed a slightly upward sloping pattern among the residuals.

```
lm(formula = log(1 + Separation_Rate) ~ x3 + x4 + recx5 + (sqrt(3.61338 -
    all_training$x6)) + x7 + x9 + x12 + x19 + x20 + x21 + x22 +
    x23, data = all_training, na.action = "na.fail")

Residuals:
     Min       1Q   Median       3Q      Max
-1.94506 -0.39474 -0.08038  0.33355  3.08033

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.3512470  0.0701257   5.009 5.52e-07 ***
x3                            -0.0399661  0.0078454  -5.094 3.53e-07 ***
x4                            -0.1146799  0.0079799 -14.371  < 2e-16 ***
recx5                         -0.0125380  0.0007165 -17.500  < 2e-16 ***
sqrt(3.61338 - all_training$x6) -0.2847123  0.0172807 -16.476  < 2e-16 ***
x7                             0.0022037  0.0045627   0.483  0.62912
x9M                           -0.1823546  0.0122286 -14.912  < 2e-16 ***
x12                            0.0253438  0.0045777   5.536 3.13e-08 ***
x192                           0.0076585  0.0612747   0.125  0.90054
x193                           0.1146976  0.0609173   1.883  0.05974 .
x194                           0.1735945  0.0613557   2.829  0.00467 **
x19M                          -0.7600294  0.4153707  -1.830  0.06730 .
x19O                           0.0058981  0.0990418   0.060  0.95251
x19P                           0.1323706  0.0641388   2.064  0.03905 *
x19X                           0.4269187  0.1576022   2.709  0.00676 **
x20                            0.0295076  0.0041355   7.135 1.00e-12 ***
x21                            0.1612353  0.0376275   4.285 1.84e-05 ***
x221                          -0.0365043  0.0114723  -3.182  0.00147 **
x231                           0.0933267  0.0114054   8.183 2.94e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.581 on 19780 degrees of freedom
Multiple R-squared:  0.1069,    Adjusted R-squared:  0.1061
F-statistic: 131.6 on 18 and 19780 DF,  p-value: < 2.2e-16
```

Figure 7. R Output for Final Full Model Summary Statistics

```
Analysis of Variance Table

Response: log(1 + Separation_Rate)
                                  Df Sum Sq Mean Sq   F value    Pr(>F)
x3                                 1  344.4  344.38 1020.1890 < 2.2e-16 ***
x4                                 1   68.2   68.17  201.9382 < 2.2e-16 ***
recx5                              1   81.0   81.04  240.0808 < 2.2e-16 ***
sqrt(3.61338 - all_training$x6)    1  102.1  102.06  302.3392 < 2.2e-16 ***
x7                                 1    0.1    0.10    0.3055  0.580463
x9                                 1   80.3   80.30  237.8784 < 2.2e-16 ***
x12                                1    9.7    9.67   28.6370 8.827e-08 ***
x19                                7   66.5    9.50   28.1512 < 2.2e-16 ***
x20                                1   16.2   16.17   47.9161 4.585e-12 ***
x21                                1    6.3    6.27   18.5747 1.642e-05 ***
x22                                1    2.3    2.30    6.8125  0.009059 **
x23                                1   22.6   22.60   66.9560 2.943e-16 ***
Residuals                      19780 6677.0    0.34
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> vif(trans_all3)
                                     GVIF Df GVIF^(1/(2*Df))
x3                               3.609867  1        1.899965
x4                               3.734698  1        1.932537
recx5                            1.055234  1        1.027246
sqrt(3.61338 - all_training$x6) 1.211704  1        1.100774
x7                               1.220958  1        1.104970
x9                               1.105569  1        1.051461
x12                              1.229014  1        1.108609
x19                              1.203204  7        1.013301
x20                              1.003057  1        1.001527
x21                              1.002604  1        1.001301
x22                              1.365901  1        1.168718
x23                              1.352725  1        1.163067
```

Figure 8. R Output for Final Full Model ANOVA Statistics and VIFs

The updated skewness values were within the cutoff of -1.0 and 1.0 at 0.998 and -0.216. The effects were further confirmed by checking the residual quantiles and normality plot in Figures 9 and 10. The normality plot was now closely aligned, suggesting reasonable the normality. While the residual quantile showed the positive skewness was not completely rectified, the transformations significantly reduced the magnitude the first model's maximum from being 9.269 larger to the final full model only being 1.584 times larger than the minimum.

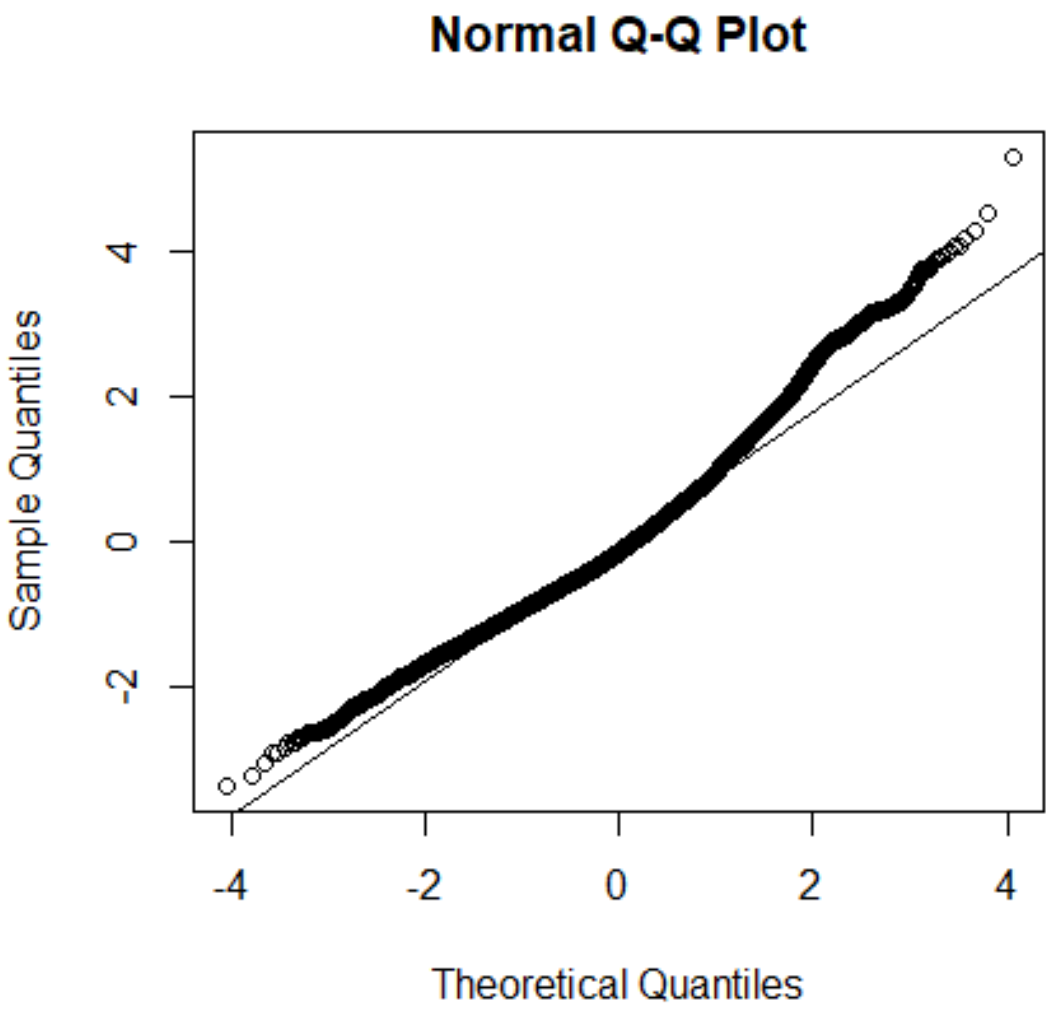


Figure 9. R Output for Final Full Model Normal Q-Q Plot

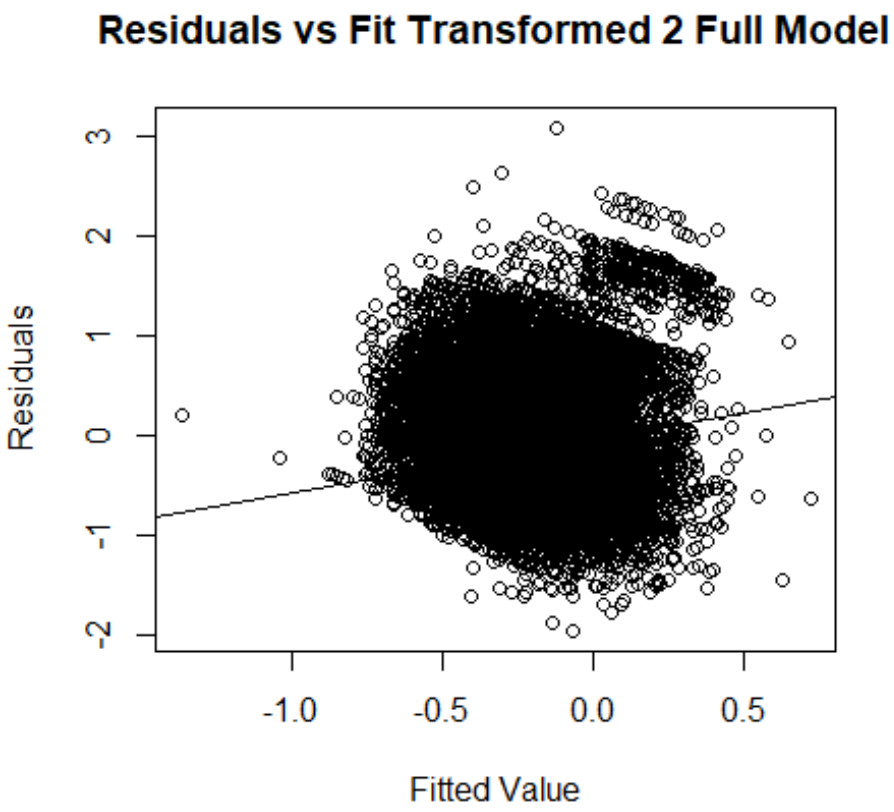


Figure 10. R Output for Final Full Model Residual Plot

The final full model contained a log(y+1) transformation on the response and the appropriate x_5 and x_6 regressor transformations. All resulting skewness values were within the cutoff of 1.0 and the model adequacy checks passed.

4.1.3 Outlier Analysis

The influence measures output reported 1,294 influential points out of the 19,799 total training data observations. This report listed each influential observation and its corresponding measures, with asterisk for which measure is deemed unusual. At this point, analyst judgment is important to identify the best course of treatment for these potential outliers.

Upon further investigation, the majority of the outliers correspond with identifiable events including the Great Recession from 2007-2009, drastic force shaping changes, and career field consolidations. This analysis sought to identify but not completely eradicate outliers among the data since the focus years contained particularly high amounts of outliers. These instances were identified but untouched because they are realistic. Models that assume or explain away all undesirable features are not realistic or applicable for real-life implementation. The goal was to create a model that is resilient enough to handle the realities of attrition and fluctuations to economic health and drastic force shaping.

4.1.4 t-tests

The t-tests on the individual regressors helped determine the best variables to explain the dependent variable of *Separation_Rate* in the model. At the 95% significance level, the t-critical value used is 1.968. Each of the regressors' t-values were compared to the t-critical value. The absolute value of every regressor's t-value, except for x_7, was greater than the t-critical value of 1.968. This suggested that nearly every regressor in the model was significant. The next step was to evaluate different subset reduction models based on this final full model.

4.2 Reduced Model Analysis

The next step produced all possible regressions from the full model and then selected three subset models for detailed analysis to finalize the best predictive model. Model A was defined by dredge as

$$\begin{aligned} Separation\_Rate =& 0.356 - 0.040x_3 - 0.115x_4 - 0.013x_5 - 0.288x_6 + \beta_9 x_9 + 0.025x_{12} \\ &+ \beta_{19}x_{19} + 0.029x_{20} + 0.161x_{21} + \beta_{22}x_{22} + \beta_{23}x_{23} + \epsilon_i \end{aligned} \tag{4}$$

.

Model B was defined by dredge as

$$\begin{aligned} Separation\_Rate =& 0.351 - 0.040x_3 - 0.115x_4 - 0.013x_5 - 0.285x_6 + 2.204e-03x_7 \\ &+ \beta_9 x_9 + 0.025x_{12} + \beta_{19}x_{19} + 0.029x_{20} + 0.161x_{21} + \beta_{22}x_{22} \\ &+ \beta_{23}x_{23} + \epsilon_i \end{aligned} \tag{5}$$

.

Model C was defined by dredge as

$$\begin{aligned} Separation\_Rate =& 0.325 - 0.043x_3 - 0.118x_4 - 0.013x_5 - 0.285x_6 + \beta_9 x_9 \\ &+ 0.024x_{12} + \beta_{19}x_{19} + 0.029x_{20} + 0.161x_{21} + \beta_{23}x_{23} + \epsilon_i \end{aligned} \tag{6}$$

.

Four of the variables, x_9, x_{19}, x_{22}, and x_{23}, do not have coefficients in Equations 4-6 because their effects were negligible in the model compared to other variables. Nevertheless, they are still included in the model comparison process and are represented symbolically. The key statistics for model comparison were R^2, R^2_{adj}, Mallow's Cp, AIC, and PRESS. The most important statistic for this analysis was PRESS because the overall goal was maximum predictive power. The top three models as recommended by the dredge() function are summarized in Table 2.

Table 2. Dredge Output Statistics

Model	R^2	R^2_{adj}	$MS_{Residual}$	**Mallow's Cp**	**df**	**AICc**	**PRESS**
A	0.107	0.106	0.338	491.351	19	34704.700	6689.587
B	0.107	0.106	0.338	493.113	20	34706.470	6690.076
C	0.106	0.106	0.338	499.538	18	34712.660	6692.179

The subset regression model statistics and model adequacy checks are shown in the Reduced Model Analysis section of the Appendix. The normal Q-Q and residual plots did not raise suspicions of violations of the normality or constant variance

assumptions. The complete model adequacy checking process is shown in Appendix B, Figures 28 through 36.

The best subset model from the comparative modeling analysis was Model C - Equation 6. All three models performed almost exactly the same for R^2, $Adjusted\ R^2$, and $MS_{Residual}$. As emphasized earlier, the objective of the final model was maximum predictive power. Thus, selecting the model with the highest PRESS statistic was the priority. Model C's PRESS is the highest among the models at 6692.179. Model C was the only one of the three evaluated to include the variable for PhD attainment but not the one for for Master's attainment. Intuitively, it follows that multicollinearity would be present between these two variables. Both measure higher education degree attainment. However, in many fields, a Master's degree is a prerequisite to a PhD. Although the VIF criterion did not reveal severe multicollinearity between the variables in the full model, closer consideration and analyst judgment confirmed the suspicion to only include one for a more parsimonious model.

4.3 Prediction for Validation Data

The final step in the regression analysis was testing the predictive power of the chosen model. The subset model C from the multiple regression was applied to the validation data set aside from 2018-2019. Model C tested if the prediction intervals contained the true 2018-2019 separation rates for specific AFSCs. Prediction intervals were produced for 20 individual AFSC. The prediction interval for the 35B was excluded because of the consolidation of the 35B and 35P fields as one grouping of Public Affairs.

The process for model validation included the same data cleaning, manipulation, and transformation process as the training data set. The same variable transformations were applied to the response and regressors. Skewness in particular was a main

issue to address. The same reciprocal and $\sqrt{x_{6_{max}} - x_6}$ transformations were applied to mitigate skewness among the regressors. Additionally, the response variable was changed to be strictly positive values so that a logarithm transformation could be applied. The prediction interval was the final output into how accurately Model C predicted the *Separation_Rate*.

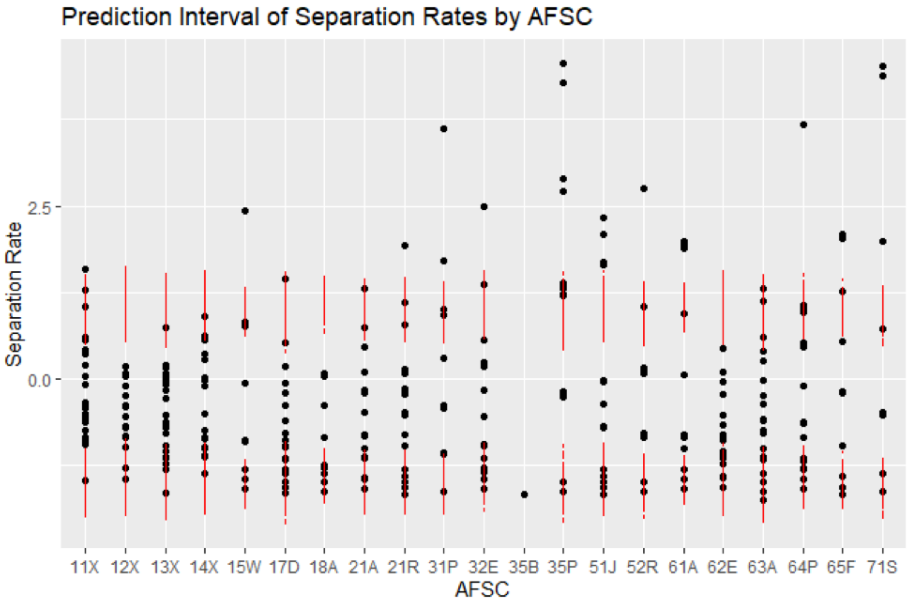


Figure 11. R Prediction Interval for Validation Data

Figure 11 shows a summary for the individual AFSC prediction intervals. The black plotted points are the observed separations, by AFSC. The red lines are Model C's prediction intervals for the validation data in Figure 11. 9 out of the 20 AFSCs were accurately predicted with Model C. Those AFSCs are 11X, 12X, 13X, 14X, 17D, 18A, 21A, 62E, and 63A. Thus, for each of these AFSCs, it was possible to confirm their individual separation rate with 95% prediction confidence. That is, the true mean of the AFSC's separation rate lies within the 95% prediction interval.

The inability to successfully predict the other 11 AFSCs likely stemmed from a low R^2. This may be due to unobservable factors the model did not account for. Further research, such as including more significant regressors to increase R^2 or an

alternative approach to explaining *Separation_Rate* altogether, is needed to address the issues with this portion of the analysis.

4.3.1 Regression Analysis Conclusion

In conclusion, select economic leading indicators and demographic variables do improve the quality of regression analysis models of officer attrition. New Business Formations and New Durable Goods Orders (x_5 and x_6) particularly proved valuable in explaining the *Separation_Rate*. YOS proved significant and should not be replaced as the primary metric in retention models. Overall, the regression analysis concluded that the two leading economic indicators should be used in addition to YOS to measure attrition. New Business Formations and New Durable Goods Orders are not strict replacements for YOS.

Due to the noisy, complex nature of data that was based on erratic, irrational human behavior, the selected model requires more quality inputs to confidently predict future attrition rates for various AFSCs. In short, there is too much unaccounted for variability among the data. The validation data confirmed this notion as it was possible to confirm the true mean of the separation rate of the prediction interval with 95% confidence for only 9 of the 20 evaluated AFSCs. The primary takeaway was the confirmation that two leading economic indicators, along with YOS, are significant variables in attrition models. The multivariate analysis that measured the predictive power of the model is discussed next.

4.4 Predictive Multivariate Analysis

The multivariate analysis featured key predictive modeling techniques available with JMP Pro 15. Specifically, it focused on classification trees and forests, predictor screening, and neural networks. Model C was applied to the 2018-2019 validation

data. This section reports the results for each of the multivariate techniques used.

4.4.1 Initial Examination

The 2018-2019 validation data was the data set for this portion of the analysis. Model C's, Equation 6, regressors were the focus. The initial examination consisted of confirming the proper variable types and transformations, fitting the model with the relevant variables, and examining the output for the proper patterns that should match the linear regression conclusions. Figures 12 and 13 display the fit model output.

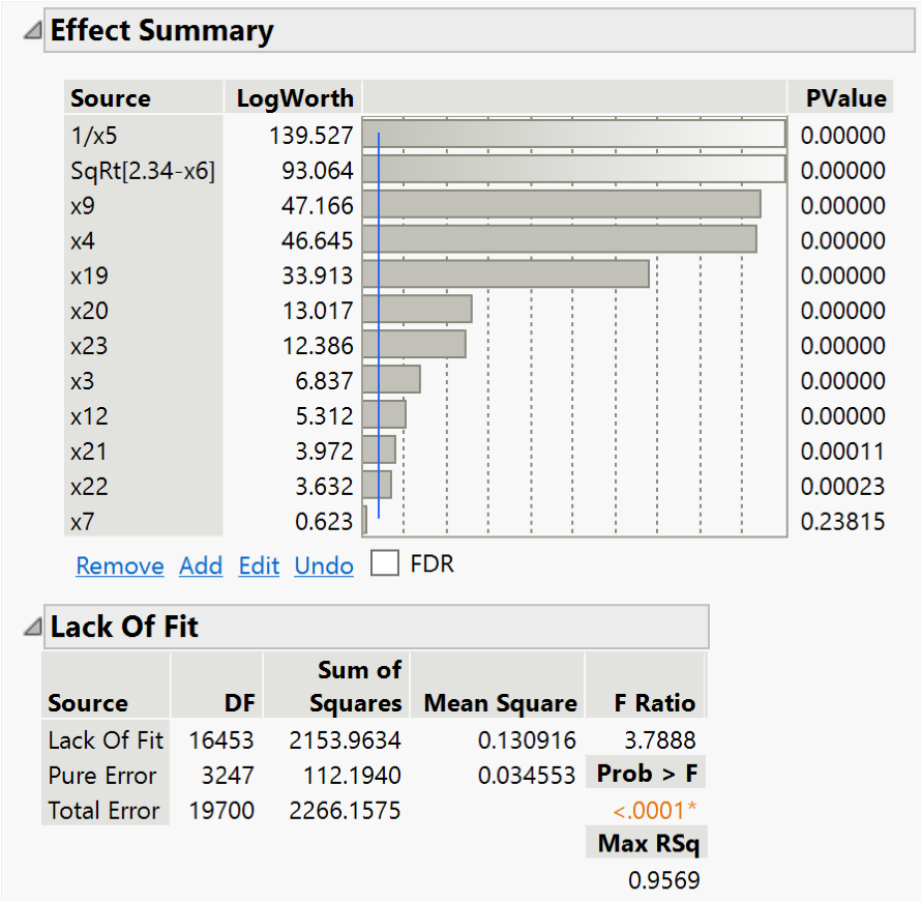


Figure 12. JMP Fit Model Output

green.

Summary of Fit

RSquare	0.128501
RSquare Adj	0.127705
Root Mean Square Error	0.339166
Mean of Response	0.474572
Observations (or Sum Wgts)	19719

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	18	334.1406	18.5634	161.3738
Error	19700	2266.1575	0.1150	**Prob > F**
C. Total	19718	2600.2981		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8483686	0.037448	22.65	<.0001*
x3	-0.024127	0.004587	-5.26	<.0001*
SqRt[2.34-x6]	-0.189643	0.009181	-20.66	<.0001*
x9[F]	0.0521206	0.003575	14.58	<.0001*
x7	-0.003182	0.002697	-1.18	0.2382
1/x5	-0.010666	0.00042	-25.42	<.0001*
x19[1]	-0.022108	0.044788	-0.49	0.6216
x19[2]	-0.01985	0.032989	-0.60	0.5474
x19[3]	0.0429029	0.032806	1.31	0.1910
x19[4]	0.073901	0.033012	2.24	0.0252*
x19[M]	-0.339238	0.210311	-1.61	0.1068
x19[O]	-0.017647	0.051348	-0.34	0.7311
x19[P]	0.0407702	0.03436	1.19	0.2354
x20[-0.301412352]	-0.032611	0.004376	-7.45	<.0001*
x21[0]	-0.042651	0.011005	-3.88	0.0001*
x22[0]	0.0123412	0.003353	3.68	0.0002*
x23[0]	-0.024221	0.003338	-7.26	<.0001*
x4	-0.067636	0.004666	-14.50	<.0001*
x12	0.0122505	0.00268	4.57	<.0001*

Figure 13. JMP Fit Model Output

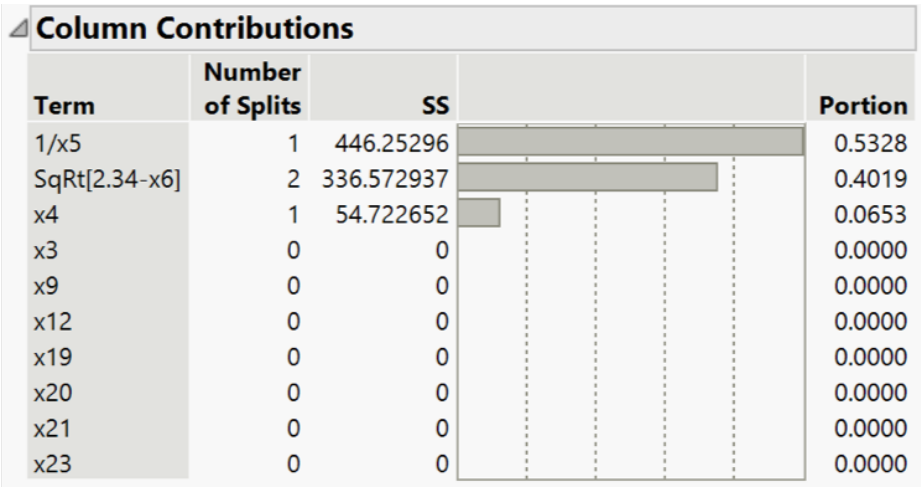
4.4.2 Classification Tree and Forests

Decision trees are the first predictive technique applied to the validation data. Splits were manually determined at both three and four trees. The three split decision tree had a R^2 value of 0.38, and the four split tree improved with a R^2 of 0.40. Figures 14 and 15 show the column contribution results for both of these splits. The most important conclusion from the decision trees was that the majority of the portion of the trees are contributed by the x_5 and x_6 variables. Combined, x_5 and x_6 contributed approximately 93% of the entire tree predictability. The next highest contributing variable was x_4, YOS, at approximately 6.6%.

Column Contributions

Term	Number of Splits	SS	Portion
1/x5	1	446.25296	0.5606
SqRt[2.34-x6]	1	295.066206	0.3707
x4	1	54.722652	0.0687
x3	0	0	0.0000
x9	0	0	0.0000
x12	0	0	0.0000
x19	0	0	0.0000
x20	0	0	0.0000
x21	0	0	0.0000
x23	0	0	0.0000

Figure 14. JMP Decision Tree Splits=3 Column Contributions

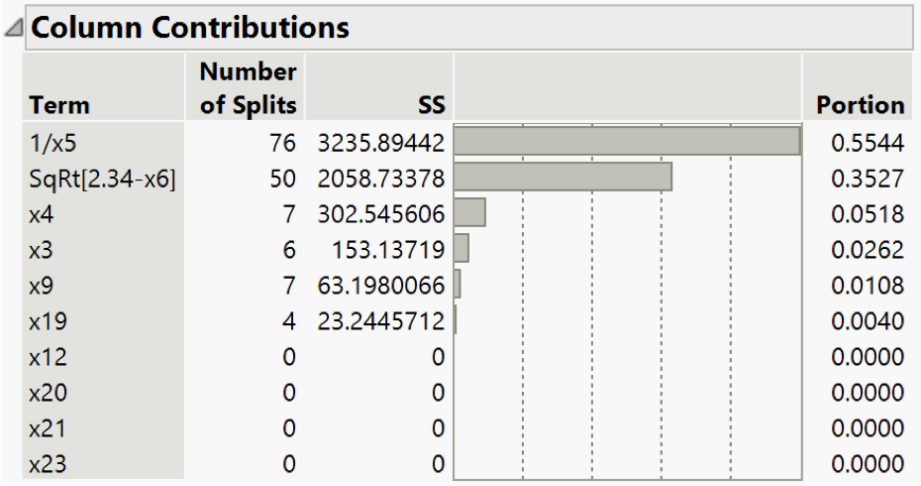


Column Contributions

Term	Number of Splits	SS	Portion
1/x5	1	446.25296	0.5328
SqRt[2.34-x6]	2	336.572937	0.4019
x4	1	54.722652	0.0653
x3	0	0	0.0000
x9	0	0	0.0000
x12	0	0	0.0000
x19	0	0	0.0000
x20	0	0	0.0000
x21	0	0	0.0000
x23	0	0	0.0000

Figure 15. JMP Decision Tree Splits=4 Column Contributions

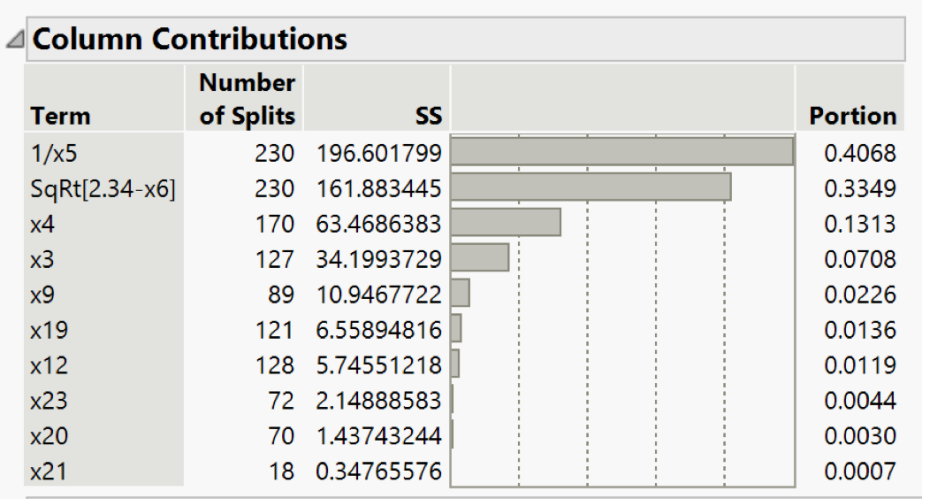
The boosted trees approach eliminated some of the overfitting of the data that decisions trees are prone to. The boosted tree had 50 layers with three splits per tree. This technique improved upon the overall statistics compared to the single decision tree method, as demonstrated by an increased R^2 of 0.53. Figure 16 shows the column contributions. x_5 and x_6 contributed approximately 87% together. However, compared to the single tree, this technique concluded x_4, x_3, x_9, and x_{19} jointly contribute less than 5%. The bootstrap forest ensemble method, shown in Figure 17 further improved the accuracy of the model. Although the R^2 drops to 0.47, the ensemble method approach of random forest was more accurate in explaining the data. Variable x_4s contribution increased from a portion of only 5% to 13%. Again, x_5 and x_6 contributed a majority of the portion at approximately 74%. Variable x_3 contributed 7%. Variables x_9, x_{19}, x_{12}, x_{23}, x_{20}, and x_{21} contributed the remaining 5.6%. The conclusion from the ensemble method was that the leading economic indicators contributed the most significance at approximately 74%, followed by YOS and grade with 13%, and the various demographic variables contributed minimally with only approximately 6%.



Column Contributions

Term	Number of Splits	SS	Portion
1/x5	76	3235.89442	0.5544
SqRt[2.34-x6]	50	2058.73378	0.3527
x4	7	302.545606	0.0518
x3	6	153.13719	0.0262
x9	7	63.1980066	0.0108
x19	4	23.2445712	0.0040
x12	0	0	0.0000
x20	0	0	0.0000
x21	0	0	0.0000
x23	0	0	0.0000

Figure 16. JMP Boosted Tree Splits=3 per Tree Column Contributions



Column Contributions

Term	Number of Splits	SS	Portion
1/x5	230	196.601799	0.4068
SqRt[2.34-x6]	230	161.883445	0.3349
x4	170	63.4686383	0.1313
x3	127	34.1993729	0.0708
x9	89	10.9467722	0.0226
x19	121	6.55894816	0.0136
x12	128	5.74551218	0.0119
x23	72	2.14888583	0.0044
x20	70	1.43743244	0.0030
x21	18	0.34765576	0.0007

Figure 17. JMP Bootstrap Forest Column Contributions

4.4.3 Predictor Screening

The predictor screening report ranked the variables based on their contribution of predicting the response variable. Since this technique is based on a bootstrap forest approach with 100 decision trees in each forest, it makes sense the report mimics the results from the classification tree and forest methods in the previous section. Figure 18 shows the JMP predictor screening report.

Predictor Screening

Log[Valid_Separation_Rate+1.75] — Copy Selected

Predictor	Contribution	Portion	Rank
1/x5	176.168	0.4638	1
SqRt[2.34-x6]	94.546	0.2489	2
x4	51.737	0.1362	3
x3	40.769	0.1073	4
x9	6.565	0.0173	5
x12	4.595	0.0121	6
x19	3.273	0.0086	7
x23	1.226	0.0032	8
x20	0.924	0.0024	9
x21	0.056	0.0001	10

Figure 18. JMP Predictor Screening Report

Figure 19 reports the conclusion of the predictor screening report from JMP. Variables x_5 and x_6 together contributed approximately 71%, and x_4 and x_3 account for 24% of the model's total predictive power. The remaining demographic variables contributed less than 5% of predictive power to the model. The economic indicators, YOS, and grade attributed 95% of the total predictive power of the model. This further solidified the conclusion that the two leading economic indicators and YOS should be used.

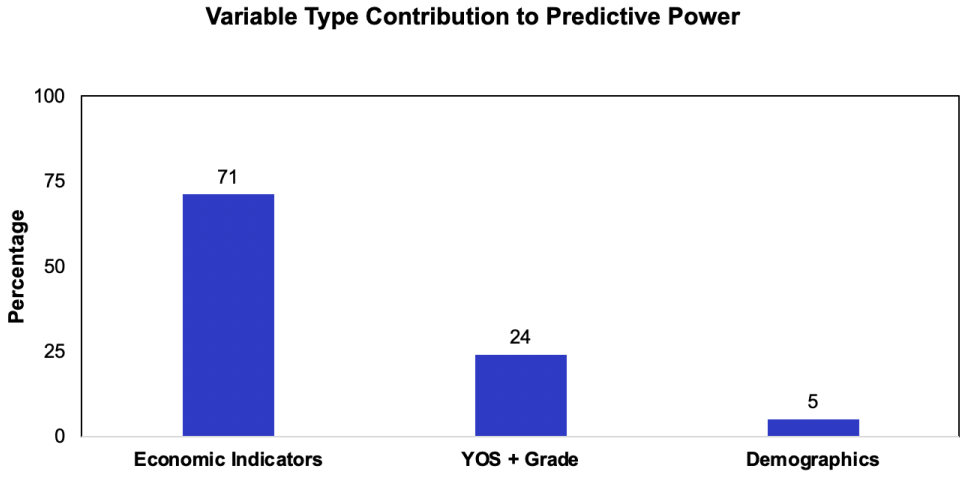


Figure 19. Variable Type Contribution to Predictive Power

4.4.4 Neural Networks

Neural networks was the final multivariate technique applied to this data. It captured the complex non-linear relationships in this data, which was very difficult for the linear regression and multivariate techniques to complete. The baseline model served as a benchmark and was based on three nodes in the first of two layers using only the TanH function. Figure 20 reports the baseline neural network summary.

The complex model took advantage of all three activation functions in the NN feature within JMP. Using all three functions was valuable because certain models may benefit more from one function than another, but it was difficult to determine which was the best without a complicated deep-dive into the data. Thus, all three functions were used as a mix for the complex model to attempt to cover as much ground as possible. Figure 21 reports the complex neural network summary. Figures 22 and 23 show the baseline and complex diagrams. Note how much more intricate the complex models diagram was and its improved model summary statistics. The R^2 increased from 0.44 for the baseline to 0.61 for the complex model. That was the highest R^2 seen in this entire analysis and demonstrates perhaps a linear regression approach is not adequate for the complex data of modeling retention. Suggestions building on this notion are discussed Chapter 5.

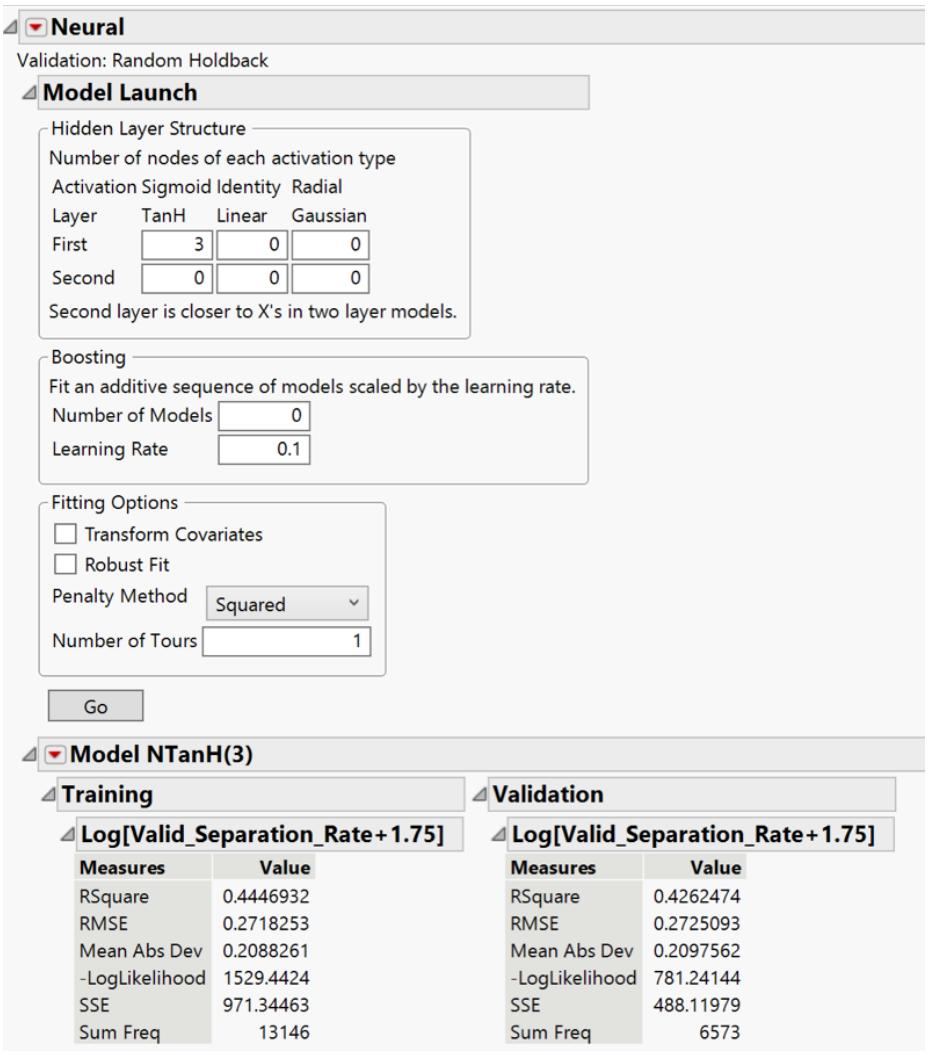


Figure 20. JMP NN Baseline Summary

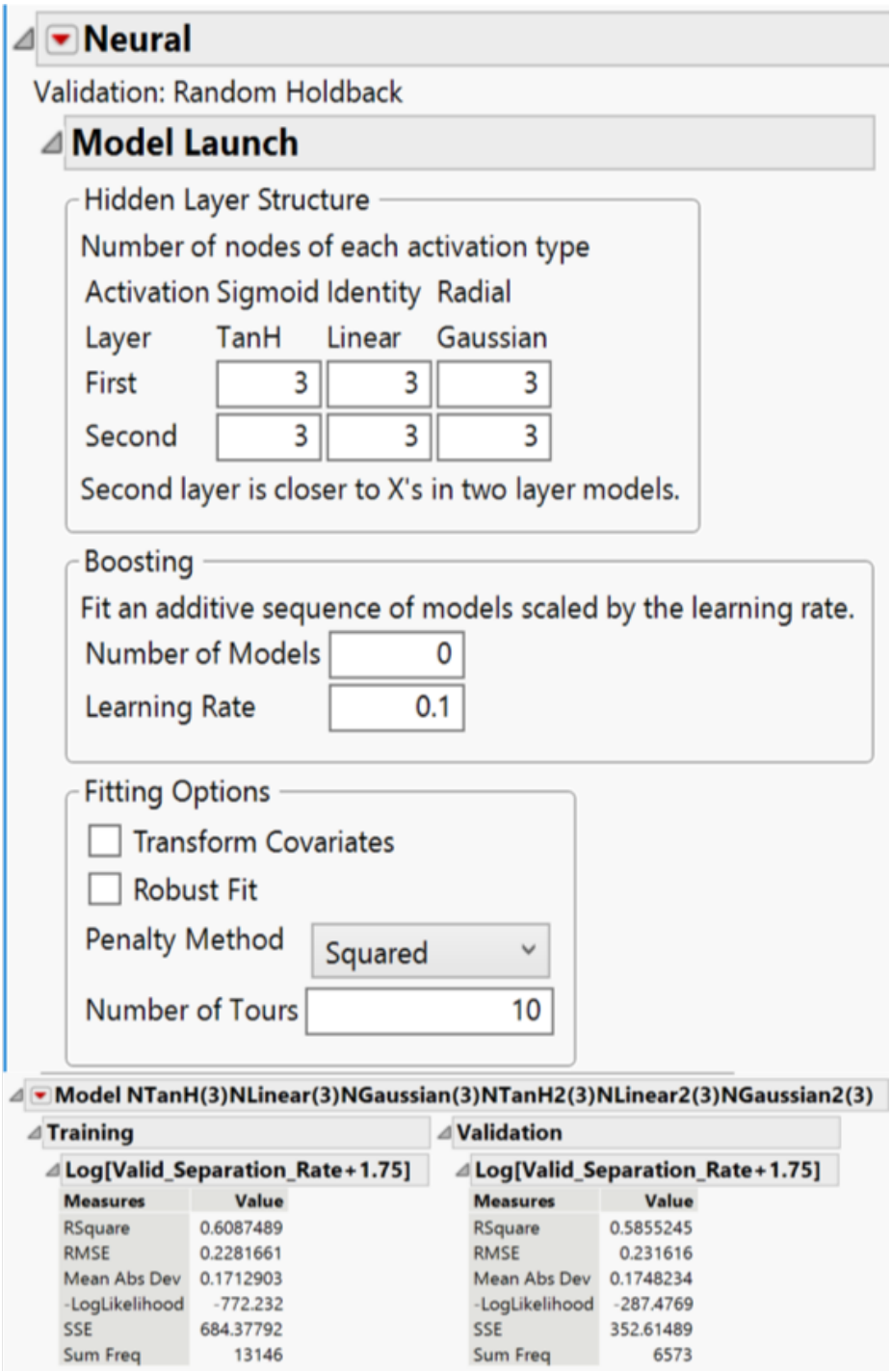


Figure 21. JMP NN Complex Summary

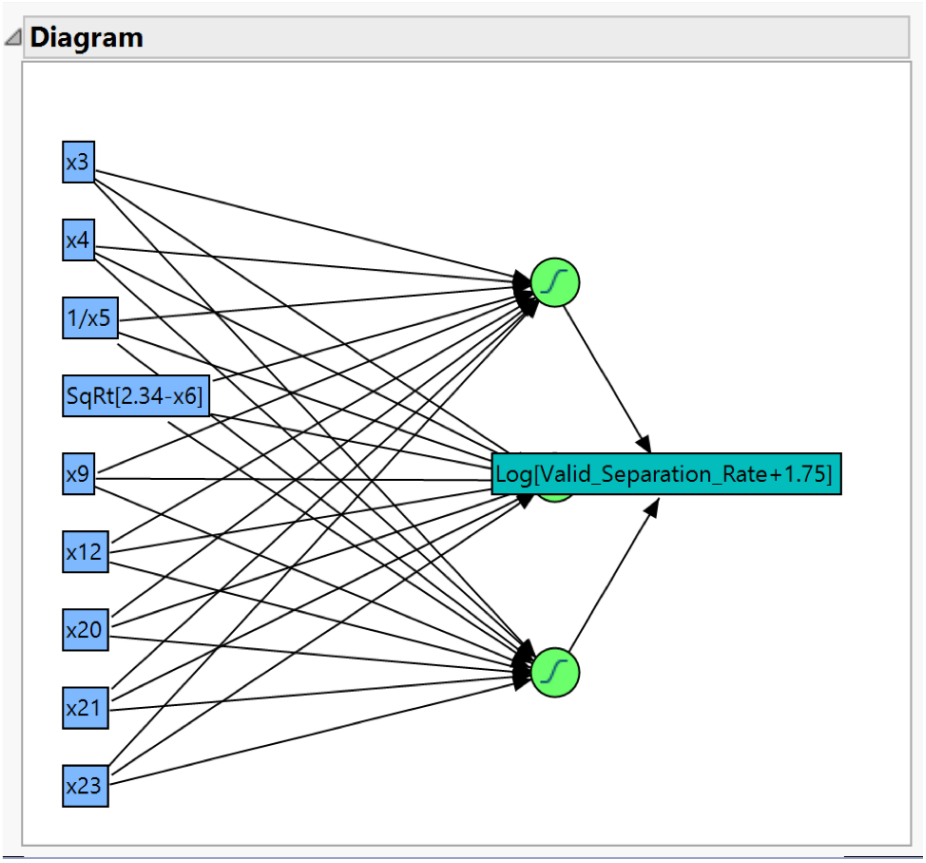


Figure 22. JMP NN Base Diagram

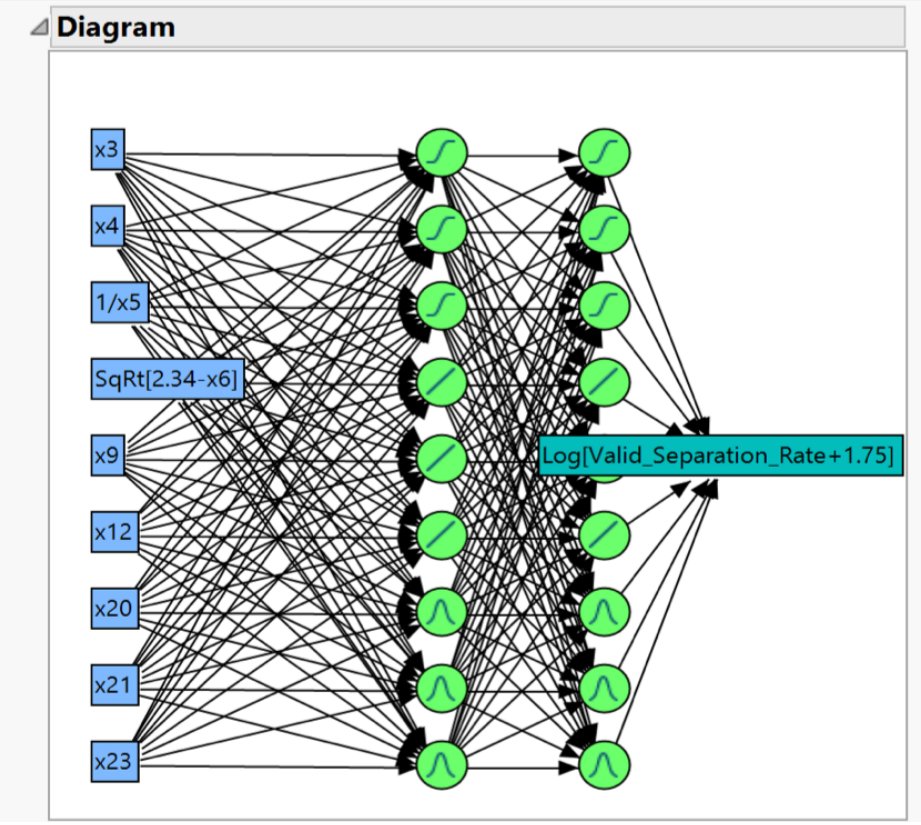


Figure 23. JMP NN Complex Diagram

4.4.5 Multivariate Conclusion

The multivariate portion of the analysis further explored the predictive power of the regressors in Model C to both explain the variability and predict *Separation_Rate*. All three multivariate techniques concluded that New Business Formations and New Durable Goods Orders are the two variables to best predict *Separation_Rate* in Model C. Therefore, it was possible to conclude with confidence that these leading economic indicators should be included in the retention models the Air Force uses to track and predict officer attrition.

V. Conclusions and Recommendations

5.1 Conclusions

This research began with an exploratory analysis to examine any significant relationships between the leading economic variables and the personnel data. Several linear regression models were evaluated using a curated master data set of combined personnel and external economic data sets. From there, the question of the best fitting predictive model was answered. The best model contained the leading economic indicators of New Business Formations and New Durable Goods Orders, as well as YOS. All three variables should be used in future force sustainment models due to their strong significance and potential for prediction. However, the model's predictive success to forecast future attrition was varied. With 95% individual confidence, Model C only successfully predicted 9 out of the 20 AFSC separation rates for the validation data. While issues of multicollinearity, transformations, and adding seemingly important regressors were properly addressed, there still remained too much unaccounted-for variability in the model. A multivariate analysis was then applied to the best model from the regression analysis. That multivariate analysis addressed the model's potential for prediction and which specific variables contribute the most predictive power. The findings of that portion of the research underlined the importance of identifying statistically significant variables to include in sustainment models. The two variables found the most statistically significant in both the linear regression and multivariate analysis were New Business Formations and New Durable Goods Orders. In addition, YOS should remain as the primary metric for measuring attrition in sustainment models.

5.2 Recommendations

Sustainment models that are clearly defined with statistically significant variables are critical for HAF to maintain a fully-mission capable officer corps. This analysis included more demographic variables as recommended by Pujats [23], years of service, and various leading economic indicators to examine their power in explaining and predicting attrition. This research concluded that two specific leading economic indicators, New Business Formations and New Durable Goods Orders, improve attrition models when combined with years of service. While years of service remains a key indicator for attrition, it is not the only insightful one. As a result of such model implementation, leadership has higher-quality, more scoped results to better inform their decision-making when determining career sizes, retention bonuses, and widespread force shaping.

There are many unexplored avenues for future and follow-on work for this research. Some recommendations are to perform logistic regression on individual observations, focus on the demographic features, introduce more lagging variables, incorporate the Auto-Regressive Integrated Moving Average (ARIMA) model forecasting and apply robust regression technique.

This research performed strict linear regression on the continuous numeric dependent variable of AFSC-specific separation rate. Individual attrition should be considered a priority in continuing this work. Examining and predicting the likelihood of an individual separating would be one way to scope this model down even further. While this research did identify key demographics that improve the accuracy of the model, such as PhD attainment, assumptions were generalized across career fields. A binary classification of separate versus remain in force using logistic regression may identify patterns of behavior among similar groups.

Inclusion of more demographic variables could help increase R^2 value of the model.

The final attrition prediction intervals were not successful for all AFSCs. One method to remedy this is to identify and include better regressors in the model. Moreover, increased focus on identifying more significant demographic variables could reveal insightful conclusions not yet explored. A couple demographic variables that were identified as minimally significant from this analysis were sex, number of dependents, source of commissioning, and PhD attainment.

Based on the DoD's recommendation for Airmen to begin planning their separation 12 months in advance, 12 months was the time frame used to lag the leading economic indicators. The next iteration of this research could include more lag times for both the economic and demographic variables. Including 6 and 18 month lags on the leading economic indicators could better capture the accuracy of how economic health and demographic features influence attrition.

Lastly, ARIMA was commonly used in previous related research to more accurately forecast based on the trends in the data itself. It can also provide prediction and confidence intervals of forecasted attrition. Robust regression is a technique to use when data are contaminated with outliers and influential points. Attrition data is riddled with such points because it is real-life data. Robust regression is an alternative to least squares regression because it identifies where such influential observations occur and weighs them differently than the well-behaved observations. As a result, the impact of outlier points is drastically reduced. Since the bulk of the outliers and influential points in this analysis were identified but untouched, applying robust regression would be valuable to explore.

Appendix A: Transformations

```
lm(formula = (1/Separation_Rate) ~ x3 + x4 + x5 + x6 + x7 + x9 +
    x12 + x19 + x20 + x21 + x22 + x23, data = all_training)

Residuals:
    Min      1Q  Median      3Q     Max
-1910.5    -8.9    -1.1     6.6 30769.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2726    33.8402  -0.008   0.9936
x3            7.1989     4.2917   1.677   0.0935 .
x4           -2.0727     4.3378  -0.478   0.6328
x5            6.2451     3.1218   2.001   0.0455 *
x6           -2.1485     2.7401  -0.784   0.4330
x7           -0.4040     2.7787  -0.145   0.8844
x9M          -1.4348     6.6548  -0.216   0.8293
x12           3.6946     2.5760   1.434   0.1515
x192          7.4414    33.3115   0.223   0.8232
x193          4.7477    33.1155   0.143   0.8860
x194          2.9391    33.3549   0.088   0.9298
x19M         -2.1291   225.8049  -0.009   0.9925
x19O         -0.4684    53.8428  -0.009   0.9931
x19P          5.3717    34.8673   0.154   0.8776
x19X         12.4020    85.6772   0.145   0.8849
x20           4.4675     2.2489   1.987   0.0470 *
x21          -7.7135    20.4550  -0.377   0.7061
x221         -1.2976     6.2399  -0.208   0.8353
x231        -12.0699     6.2440  -1.933   0.0532 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 315.8 on 19780 degrees of freedom
Multiple R-squared:  0.0009421, Adjusted R-squared:  3.294e-05
F-statistic: 1.036 on 18 and 19780 DF,  p-value: 0.4136
```

Figure 24. R Output for Reciprocal Transformation lm Summary Statistics

```
Analysis of Variance Table

Response: (1/Separation_Rate)
             Df     Sum Sq Mean Sq F value  Pr(>F)
x3            1     336535  336535  3.3736 0.06626 .
x4            1      54487   54487  0.5462 0.45988
x5            1     308245  308245  3.0900 0.07879 .
x6            1      56595   56595  0.5673 0.45133
x7            1      29252   29252  0.2932 0.58816
x9            1       3323    3323  0.0333 0.85519
x12           1     206856  206856  2.0736 0.14988
x19           7      52534    7505  0.0752 0.99934
x20           1     413116  413116  4.1413 0.04186 *
x21           1      14425   14425  0.1446 0.70375
x22           1      12531   12531  0.1256 0.72303
x23           1     372758  372758  3.7367 0.05324 .
Residuals 19780 1973169448   99756
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vif(trans_all)
        GVIF Df GVIF^(1/(2*Df))
x3  3.655433  1        1.911919
x4  3.734463  1        1.932476
x5  1.934118  1        1.390726
x6  1.490050  1        1.220676
x7  1.532402  1        1.237902
x9  1.107945  1        1.052590
x12 1.316931  1        1.147576
x19 1.202046  7        1.013231
x20 1.003703  1        1.001850
x21 1.002613  1        1.001306
x22 1.367397  1        1.169357
x23 1.371915  1        1.171288
```

Figure 25. R Output for Reciprocal Transformation lm ANOVA Statistics and VIFs

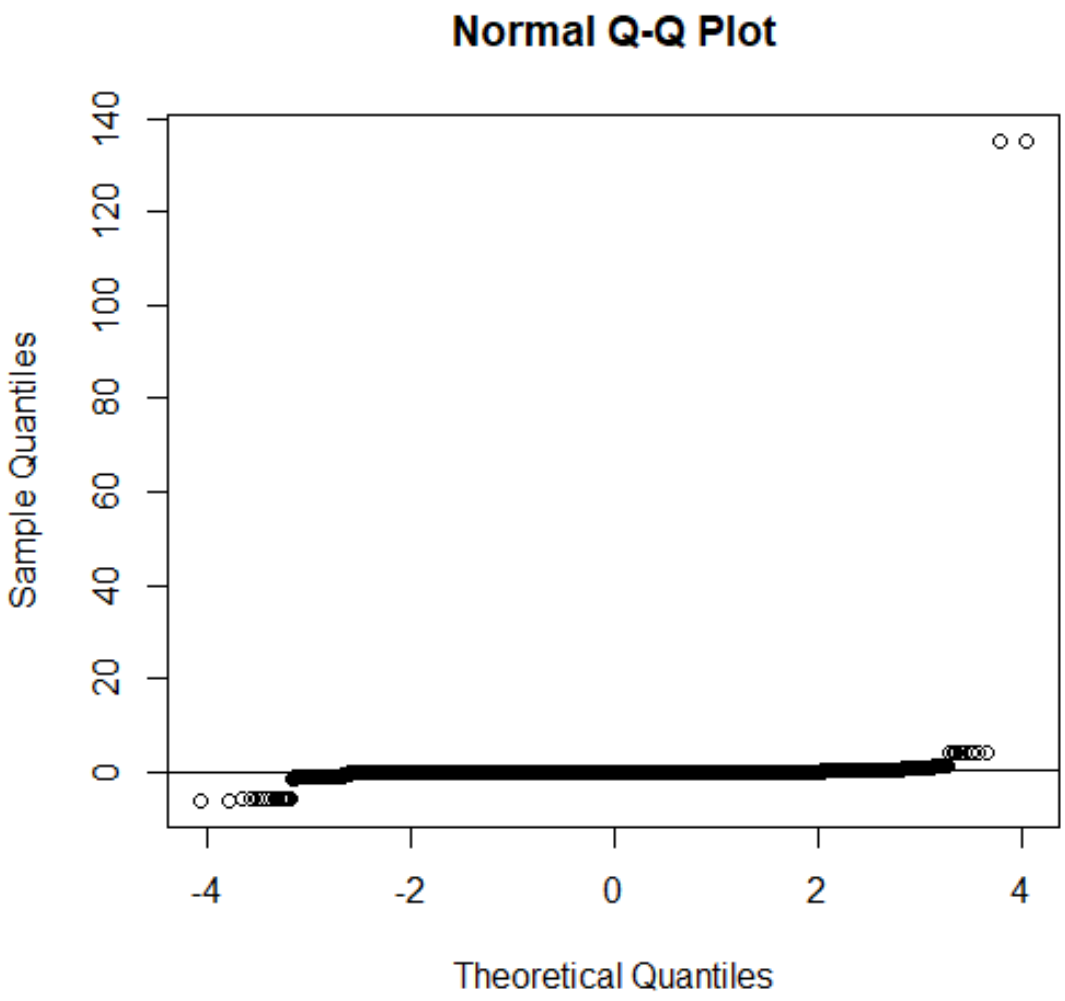


Figure 26. R Output for Reciprocal Transformation Normal Q-Q Plot

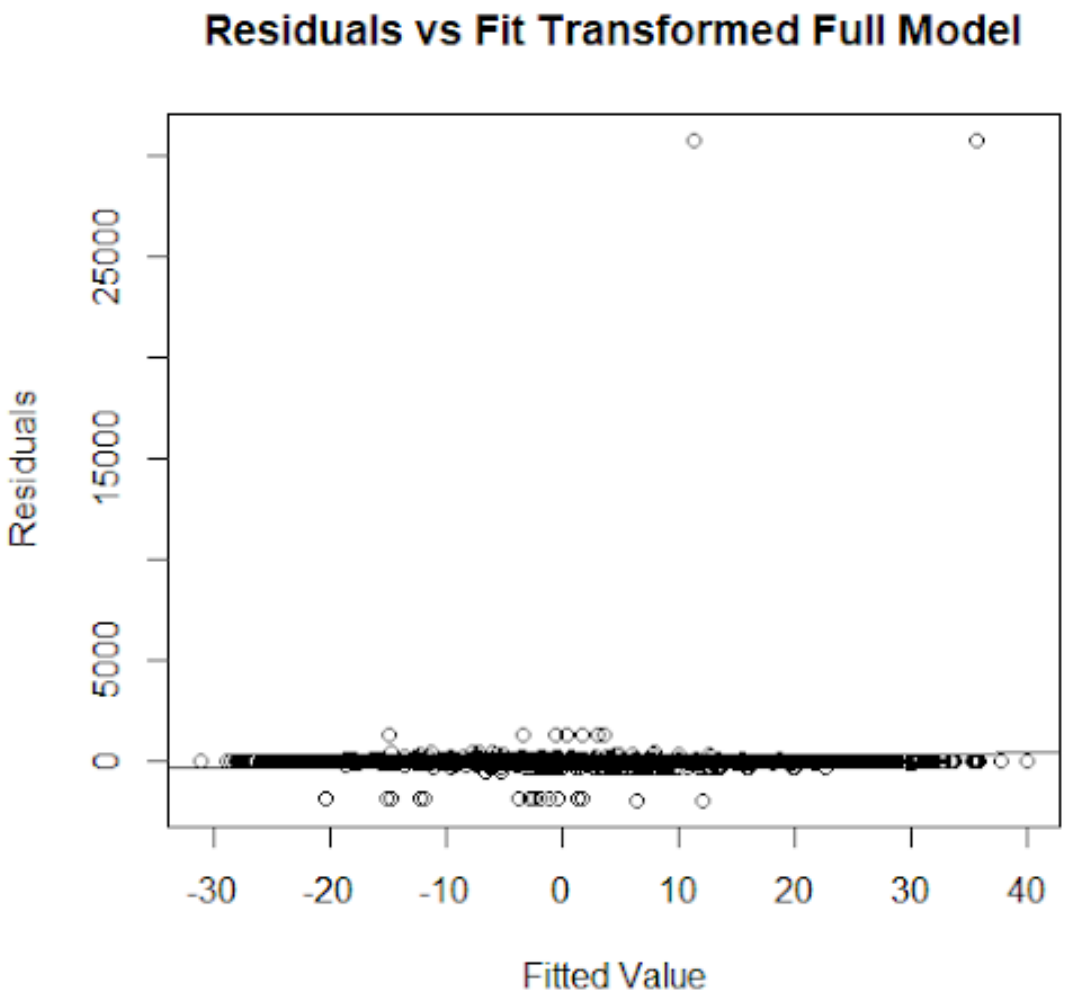


Figure 27. Reciprocal Transformation Residual Plot

Appendix B: Reduced Model Analysis

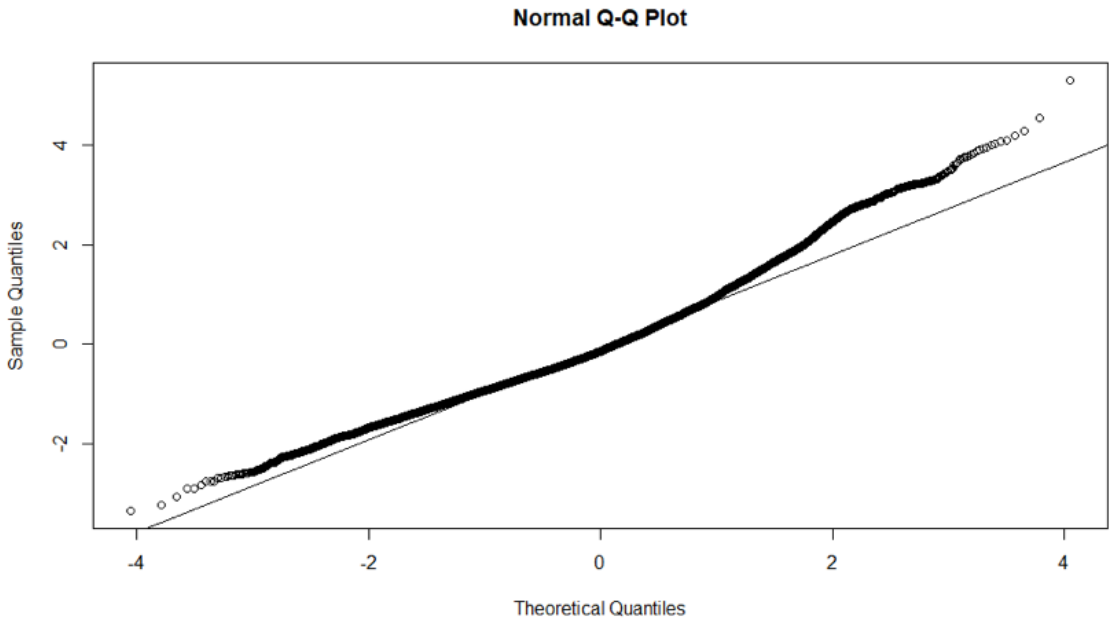


Figure 28. R Output for Model A Normal Q-Q Plot

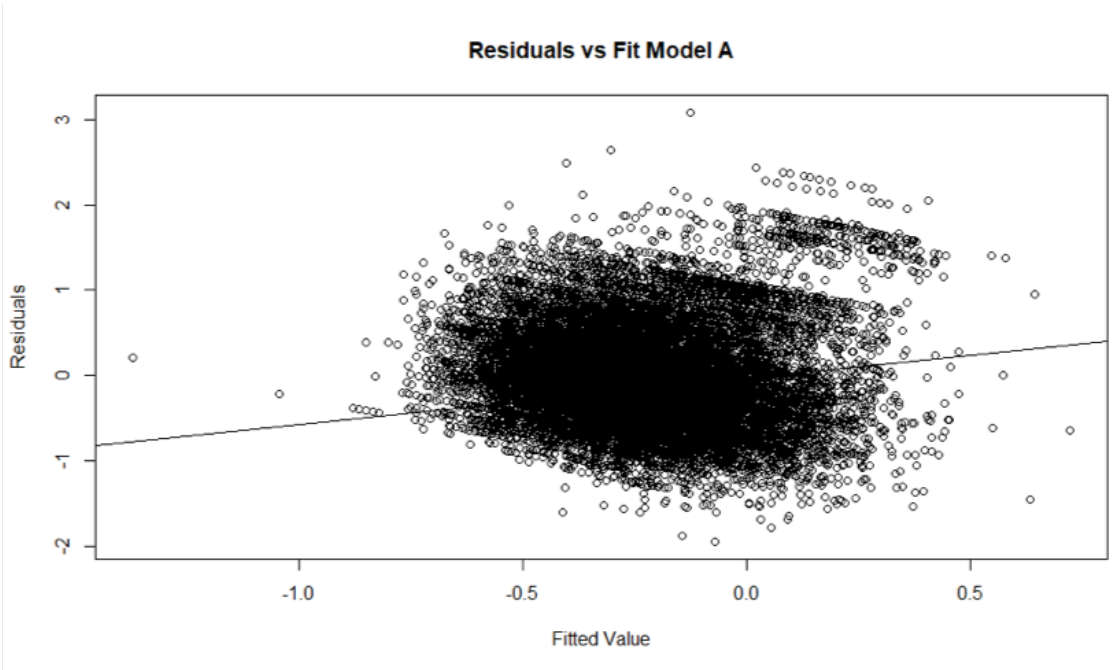


Figure 29. R Output for Model A Residual Plot

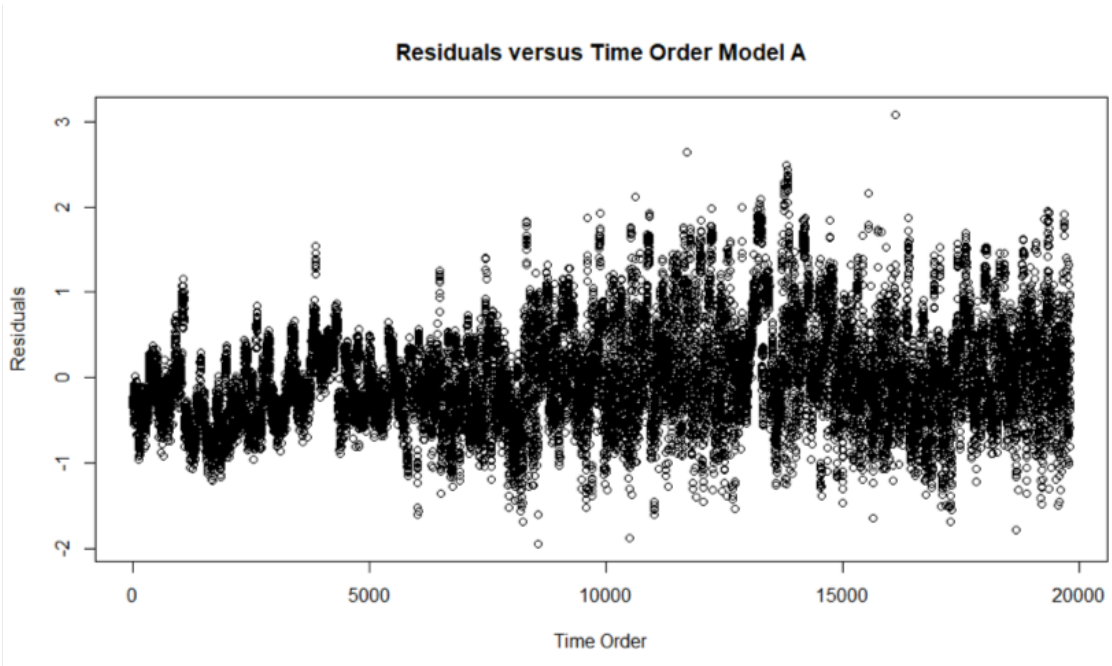


Figure 30. R Output for Model A Residuals vs Time Order

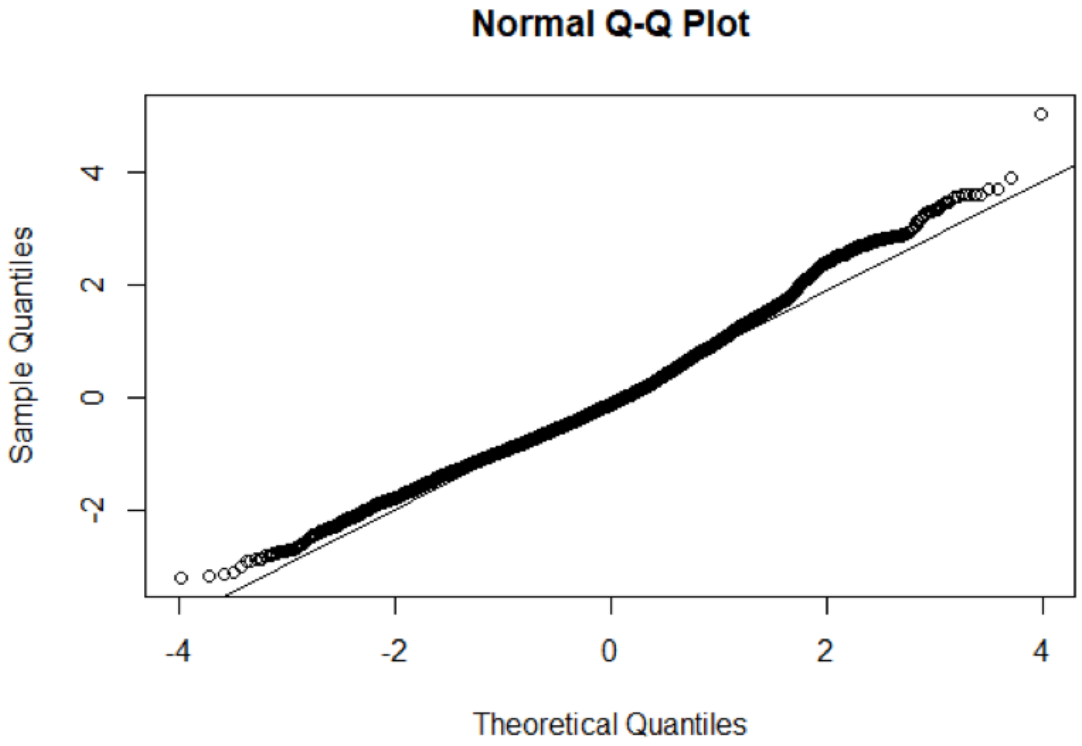


Figure 31. R Output for Model B Normal Q-Q Plot

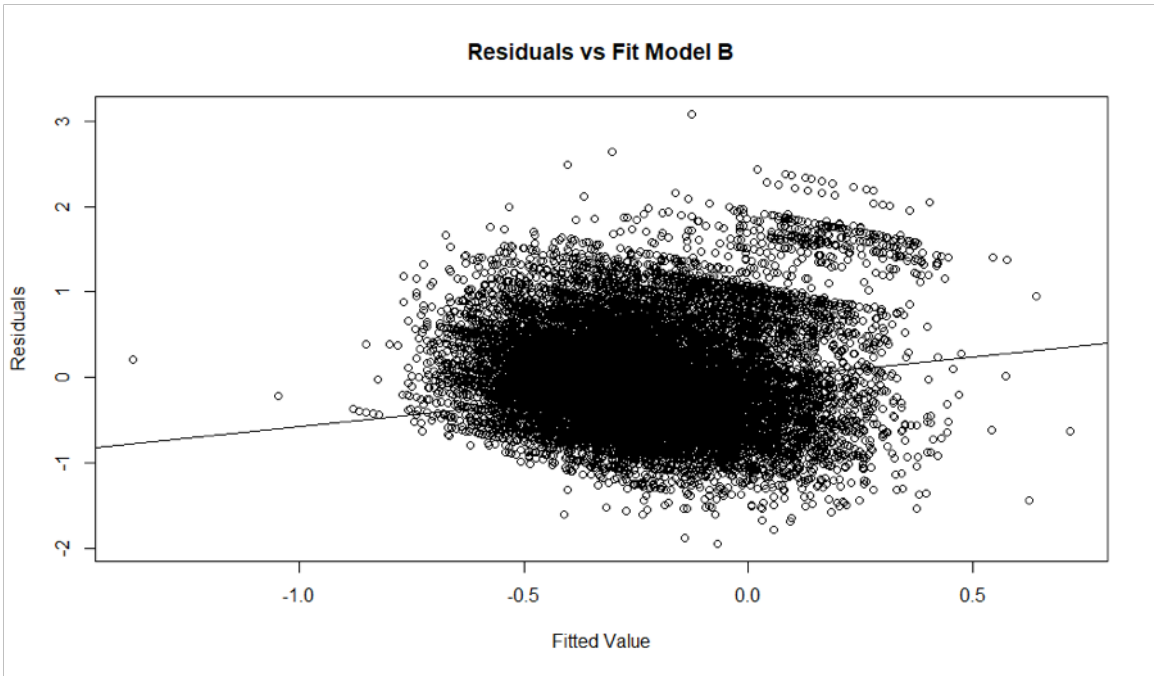


Figure 32. R Output for Model B Residual Plot

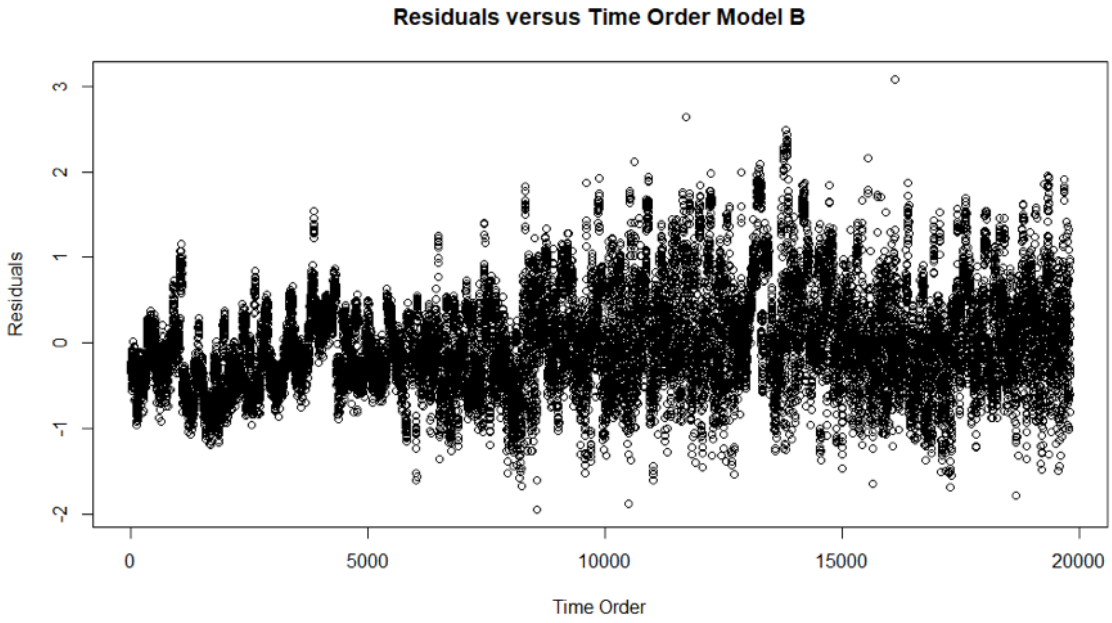


Figure 33. R Output for Model B Residuals vs Time Order

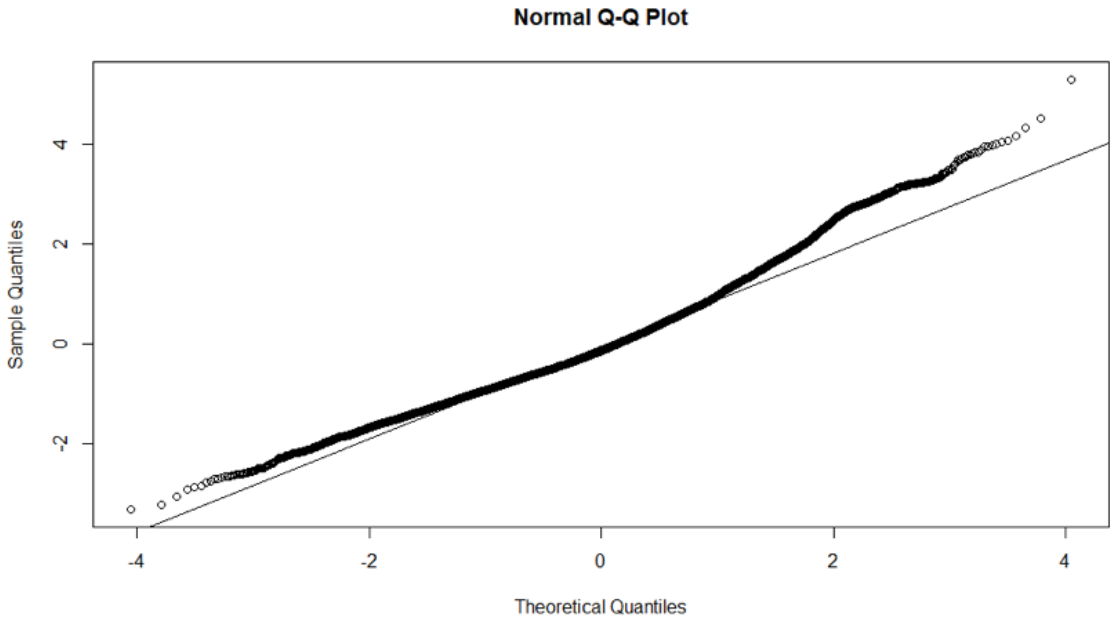


Figure 34. R Output for Model C Normal Q-Q Plot

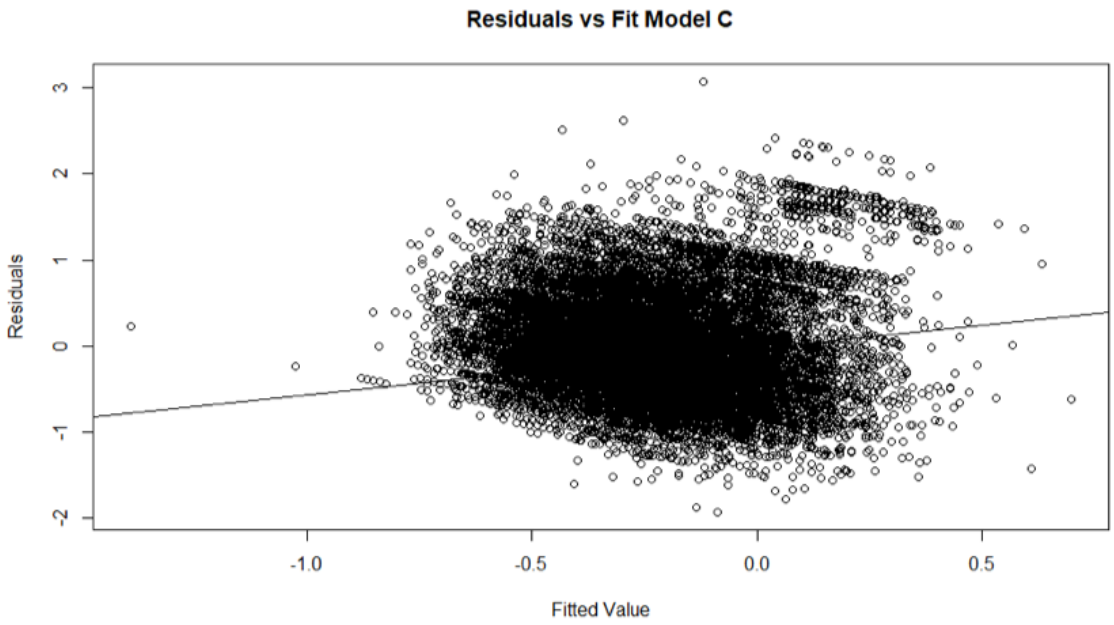


Figure 35. R Output for Model C Residual Plot

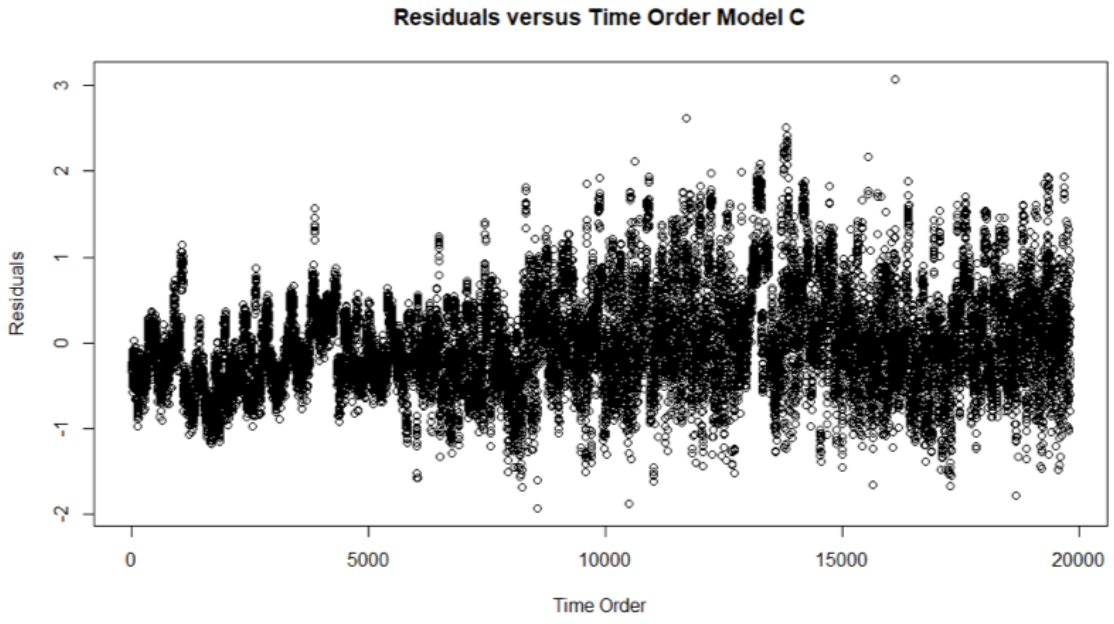


Figure 36. R Output for Model C Residuals vs Time Order

Bibliography

1. Abel, A. and Bernanke, B. [2005], *Macroeconomics*, 5 edn, Pearson Education, Inc., New York, NY.

2. Apltekin, A. and Levine, P. [2012], 'Military expenditure and economic growth: A meta-analysis', *European Journal of Political Economy* **28**, 636, 650.

3. Bruce, P. and Bruce, A. [2017], *Practical Statistics for Data Scientists*, 1 edn, OReilly Media.

4. Bureau, U. C. [2020*a*], 'Business formation statistics', *U.S. Department of Commerce* . Accessed December 2020.
URL: *https://www.census.gov/econ/bfs/index.html*

5. Bureau, U. C. [2020*b*], 'Manufacturers shipments, inventories, & orders', *U.S. Department of Commerce* . Last Accessed on 1 July 2020.
URL: *https://www.census.gov/manufacturing/m3/historical_data/index.html*

6. D. Montgomery, E. Peck, G. V. [2012], *Introduction to Linear Regression Analysis*, 5 edn, Wiley & Sons, Inc., Hoboken, NJ.

7. Elliott, J. T. [2018], Air force officer attrition: An econometric analysis, Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

8. for Economic Co-operation, O. and Development [2016], 'Consumer confidence index (cci)'. Accessed on 06 December 2020.

9. Frank E. Harrell, J. [2015], *Regression Modeling Strategies*, 2 edn, Springer International Publishing, Switzerland.

10. Franzen, C. N. [2017], Survival analysis of us air force rated officer retention, Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

11. Gaupp, M. P. [1999], Pilot inventory complex adaptive system (picas): An artificial life approach to managing pilot retention, Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

12. Hanson, M. L. and Nataraj, S. [2011], 'Expectations about civilian labor markets and army officer retention'.

13. Help, J. P. . [n.d.*a*], 'Bagging'. Accessed December 2020.
URL: *https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/boosted-tree.shtml#*

14. Help, J. P. . [n.d.*b*], 'Boosted tree'. Accessed December 2020.
URL: *https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/boosted-tree.shtml#*

15. Help, J. P. . [n.d.*c*], 'Bootstrap forest'. Accessed December 2020.
URL: *https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/bootstrap-forest.shtml#*

16. Help, J. P. . [n.d.*d*], 'Neural networks'. Accessed December 2020.
URL: *https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/neural-networks.shtml#*

17. Help, J. P. . [n.d.*e*], 'Partition models'. Accessed December 2020.
URL: *https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/partition-models.shtml*

18. Help, J. P. . [n.d.*f*], 'Predictor screening'. Accessed December 2020.
URL: *https://www.jmp.com/support/help/en/15.2/index.shtml#page/jmp/predictor-screening.shtml#*

19. Hill, R. R. and Gaupp, M. P. [2006], 'Visualization tools for prescriptive analysis of a complex adaptive system model of pilot retention', *International Journal of Modeling and Simulation* **6**(4), 378–384.

20. Jantscher, H. L. [2016], An examination of economic metrics as indicators of air force retention, Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

21. of Commerce, U. D. [2004-2019], 'Manufacturers' shipments, inventories, and orders', *U.S. Census Bureau* .

22. Patterson, C., Petkun, J. and Skimmyhorn, W. [2020], 'I want you!(but not you): Selection in military retention'.
URL: *https://jbpetkun.github.io/pages/working_papers/Manuscript_PPS.pdf*

23. Pujats, T. S. [2020], Forecasting attrition by afsc for the united states air force, Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

24. RESEARCH, N. B. O. E. [n.d.*a*], 'About the nber'. Accessed December 2020.
URL: *https://www.nber.org/about-nber*

25. RESEARCH, N. B. O. E. [n.d.*b*], 'Business cycle dating'. Accessed December 2020.
URL: *https://www.nber.org/research/business-cycle-dating*

26. Schofield, J. A. [2015], Non-rated air force line officer attrition rates using survival analysis, Master's thesis, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
27–03–2021	Master's Thesis	August 2019 — March 2021

4. TITLE AND SUBTITLE

Modeling Air Force Retention with Macroeconomic Indicators

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

McGee, Michelle K., BS, 2Lt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-M-176

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

HAF/A1XD
Douglas A. Boerman
1550 W. Perimeter Rd,, Rm 4710
Joint Base Andrews NAF Washington, MD 20762-5000
Email: douglas.a.boerman.civ@mail.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A:
Approved for Public Release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Officer retention has been a longstanding problem for Air Force leadership. Both intuition and previous research suggest economic and demographic factors play important roles in an officers decision to separate from service. Leading economic indicators are nationally reported statistics that tend to be predictive of where the economy is heading. This work targets the research gap of how leading economic indicators explain and predict attrition. Due to the noisy, complex nature of the data, the model had varied success in accurately predicting future attrition rates. As a result of this research, the current models can incorporate these findings to become more targeted and precise. Leadership who make key decisions regarding critically-manned career fields, career-specific retention bonuses, force shaping, consolidation measures, etc. will have higher-quality, better scoped results.

15. SUBJECT TERMS

statistical modeling, retention, personnel models, macroeconomics

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Raymond R. Hill, AFIT/ENS
U	U	U	UU	83	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x7469; rhill@afit.edu