# Predicting Pilot Success Using Machine Learning

THESIS

Aaron C. Giddings, Captain, USAF

AFIT-ENS-MS-20-M-150

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**

***AIR FORCE INSTITUTE OF TECHNOLOGY***

**Wright-Patterson Air Force Base, Ohio**

The views expressed in this document are those of the author and do not reflect the official policy or position of the United States Air Force, the United States Department of Defense or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENS-MS-20-M-150

# PREDICTING PILOT SUCCESS USING MACHINE LEARNING

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 Science in Operations Research

Aaron C. Giddings, BS
Captain, USAF

26 March 2020

AFIT-ENS-MS-20-M-150

# PREDICTING PILOT SUCCESS USING MACHINE LEARNING

THESIS

Aaron C. Giddings, BS
Captain, USAF

Committee Membership:

Dr. Raymond R. Hill, Ph.D.
Chair

Lt Col Jason R. Anderson, Ph.D.
Reader

AFIT-ENS-MS-20-M-150

# Abstract

The United States Air Force has a pilot shortage. Unfortunately, training an Air Force pilot requires significant time and resources. Thus, diligence and expediency are critical in selecting those pilot candidates with a strong possibility of success. This research applies multivariate and statistical machine learning techniques to pilot candidates pre-qualification test data and undergraduate pilot training results to determine whether there are selected pre-qualification tests or specific training evaluations that do a “best” job of screening for successful pilot training candidates and distinguished graduates. Flight experience, both during training and otherwise, indicates pilot training completion and performance.

# Acknowledgements

*Anything begun in vanity ends in humility.*

Much thanks to my fellow students for the support they showed me throughout my studies, my family for their relentless encouragement, and my advisor for keeping me focused on the end game while at AFIT.

Aaron C. Giddings

# Table of Contents

Page

# List of Figures

# List of Tables

Table Page

PREDICTING PILOT SUCCESS USING MACHINE LEARNING

# I. Introduction

## 1.1 Problem Statement

The United States Air Force has a pilot shortage, especially fighter and special operations pilots. According to the United States Government Accountability Office (GAO) there are a combined 445 unfilled operational positions from those two communities which results in 87 percent fill for fighter pilot positions and 97 percent fill for special operations pilot positions [1]. The United States Air Force has explored multiple ways to expand the pilot pool from providing different retention bonuses to changing the medical requirements to expand the candidate pool. Other analyses have examined personnel demographics studies to examine candidate retention by different categorical factors such as age, gender, and race. Meyer [2] conducted this study and found that women and minorities were more likely to attrite UPT, but found no trend in age. Pilot training also takes a long time; over a year in most cases. Air Force Education and Training Command (AETC) is attempting to shorten Undergraduate Pilot Training (UPT) using different techniques. One of the techniques that AETC is evaluating is Pilot Training Next (PTN) which utilizes virtual reality technology to help instruct pilot candidates.

Another avenue explored by AETC is saving money by streamlining the number of tests that candidates need to take, while not undermining the value these tests provide as a screening mechanism for pilot candidates. Historically, UPT for the United States Air Force provides numerous pre-evaluation tests to evaluate the potential success of pilot candidates and skills assessment tests once enrolled. Pre-evaluation tests include the Air Force Officer Qualifications Test (AFOQT), the Pilot Candidate Selection Method (PCSM), and the Test of Basic Aviation Skills (TBAS). While in UPT, pilot candidates are assessed using written tests for instrument knowledge and orientation, daily rides and check rides for situational proficiencies, and flight commander rankings to assess the candidates understanding of group dynamics. This thesis explores the plethora of tests to find out whether any specific test or combination of tests may more accurately predict the success of pilot candidates in their UPT experience. Of the many possible factors that a pilot candidate has, the next section focuses the study on specific areas of interest.

## 1.2 Identifying the Research Lens

When applying to UPT, potential candidates need to complete certain tests that assess verbal and scientific reasoning, as well as spatial reasoning and baseline reaction timing crucial for pilot success. Through the previous studies conducted and the data collected from the Air Force Personnel Center, this study hopes to define any relationships between the pre-aptitude and proficiency tests and success in pilot training. With the scope of the research established, the next section addresses the purpose of the research.

## 1.3 Research Purpose

This study examines the relationships between the pre-aptitude and proficiency tests and eventual success in pilot training. This work builds on previous studies conducted and uses the data collected from the Air Force Personnel Center [2].

Applying to UPT, cadets need to complete certain tests that assess verbal and scientific reasoning, as well as spatial reasoning and baseline reaction timing crucial for pilot success. It is the result of these tests that are used in the analyses provided in this work. The next section takes the purpose of the research and formulates questions that this research answers.

## 1.4 Research Questions and Hypothesis

The primary research question addresses the measures of pilot success: what tests seem to best measure pilot training success? Since the methodology employed uses multivariate (or machine learning) methods, a sub-question is which multivariate technique preforms the best for this data? Which of the multivariate techniques provide the most useful insight into the test data examined to help answer the driving research question. Ultimately, the study examines whether candidates with more experience flying and higher PCSM and TBAS scores will perform better in UPT.

## 1.5 Overview

Chapter II describes the previous research and studies done on pilot evaluations as well as an in-depth description of the pilot parameters and assessments and an overview of the machine learning techniques used in this study. Chapter III explains the process used for evaluating the data to include data cleaning performed, pre-processing techniques used, prediction algorithms utilized, and verification procedures executed. Chapter IV goes through the results of the study evaluating the accuracy of the predictions, different characteristics deemed important, and overall trends. Lastly, Chapter V provides conclusions as well as further avenues of research.

# II. Literature Review

## 2.1 Background

Since the United States Air Force officially started in 1947, leaders have attempted to find pilots that have “the right stuff.” Hunter and Burke [3] identified 68 published studies from 1940 to 1990 addressing which predictor variables to include in pilot selection that accurately quantifies “the right stuff”.

The most significant predictive factors they found for graduating pilots were cognitive ability, including verbal, quantitative, and undergraduate grade point average (GPA). This is referenced in several other studies, from fighter pilot performance analysis during the Korean War in 1957, to validation of the USAFs current Pilot Candidate Selection Method (PCSM) and use of the AFOQT, 2008 through 2018 [4][5] [6] [7][8]. Hunter and Burke also confirmed that job skills, dexterity, and reaction time were also good predictive factors of pilot training success [3].

When it comes to addressing the retention problem of Air Force pilots, leadership wants to recruit the people more likely to embrace the pilot culture in addition to being technically proficient and able to handle the stresses of flying. Wilson conducted a study that measured the stresses a pilot endures during a flight and found that during different stages of a flight, the pilot’s heart rate varied by a rate of 6 hertz (Hz) during takeoff and landing, rising from 80 beats per minute (bpm) to 98 bpm during the takeoff phase and decelerating from 113 bpm to 98 bpm during landing [9]. To identify prior to UPT whether a candidate will handle those stresses, the Air Force attempts to find candidates that possess the right resiliency characteristics necessary for UPT. This is done using the personality test conducted as a component of the AFOQT called the Self-Description Inventory (SDI+). The SDI+ is a 220-item, trait-based measure that assesses the Big 5 domains of Neuroticism (the ability to feel distressed or anxious), Extraversion, Openness, Agreeableness, and Conscientiousness (the ability to be careful or diligent), with a measure of Machiavellianism (the ability to manipulate others) included in the test [7].

## 2.2 Prior Studies

This thesis uses the same data most recently used for demographic analysis of both attrition and performance. In that recent study, Meyer [2] used statistical hypothesis testing to find categorical biases between age, gender, and race. Using linear regression techniques, she found that women and minorities were more likely to leave pilot training while no such bias existed for age. This data contains various test results ranging from personality tests to prior entry tests (like the AFOQT and PCSM) to check ride scores and flight commander evaluations from all pilot candidates between 2011 and 2018 which includes over 27,000 data points spread out between 12,000 unique manned and unmanned pilot candidates. Since the pre-selection process is similar between the Remotely Piloted Aircraft (RPA) and the manned aircraft pilots, and the two training programs contain comparable measures, the studies and analysis conducted are combined. The AFOQT and PCSM tests attempt to characterize the traits needed for pilot success [10]. These next sections address specific aspects of prior studies conducted on UPT students and applicants in different eras, but none of the subsequent studies conducted utilized the same data as the study Meyer conducted analyzing UPT attrition on the demographic areas of age, gender, and race [2].

## 2.3 USAF Pilot Selection Process

The United States Air Force chose pilot candidates from the pool of commissioned USAF officers by selection boards from 1965 through 1995. They evaluated these candidates based on pilot suitability, including medical and physical fitness factors, academic performance, aptitude test scores, commanders recommendations, and previous flying experience [5]. While a majority of that data can still be attained, everything except for academic performance and aptitude test scores is based solely on the results of the pre-entry tests conducted on pilot candidates.

Candidates progressing through UPT are assessed primarily on their academic performance and aptitude test scores [11]. The way the Air Force quantifies aptitude prior to acceptance into UPT is through the AFOQT, the PCSM, and the TBAS. The PCSM also assesses a candidate's previous flying experience by adding the flight hours to the overall score. Pilot aptitude is tested in two different aspects: operational and technical proficiency. To assess a candidate's operational proficiency while in pilot training, instructors ask them to perform progress checks and elimination checks [11]. Meanwhile, a candidate's technical proficiency is measured by academic performance using classroom tests [11].

## 2.4 Pre-Aptitude Tests

Pilot candidates in the United States Air Force have many tests they must take prior to entry into pilot training. Similar to the Scholastic Assessment Test (SAT) or the American College Test (ACT) that high school seniors take prior to entry into college, the aptitude tests required for pilot candidates assess their communicative and analytical competences. The difference with these tests are the assessments for spatial awareness, reactive capability, and mental acuity required for UPT. The aptitude tests pilot candidates take include the AFOQT, TBAS, and the PCSM.

### 2.4.1 Air Force Officer Qualifying Test

The AFOQT is comprised of five different subtests designed to assess the different relevant skills necessary to become an Air Force Officer. The first subtest evaluates a candidate's verbal skills by testing their verbal analogies, which tests the candidate's ability to reason and determine relationships between words and their word knowledge, which assesses verbal comprehension including the ability to understand written languages through the use of synonyms [12]. The second subtest assesses a candidate's quantitative skills by measuring their arithmetic reasoning skills, which evaluates their ability to understand arithmetic relations expressed as word problems and their math knowledge, which provides a measure of the ability to use mathematical formulas, relations, and terms [12]. The third subtest evaluates a candidate's spatial reasoning skills using block counting, rotated blocks, and hidden figures. Block counting analyzes three-dimensional sets of blocks for spatial awareness. Rotated blocks allow for the candidate to visualize and mentally manipulate objects. Hidden figures assesses the candidate's ability to breakdown complex figures into simple components [12]. The fourth subtest is specifically designed for the focus of this research: the aircrew evaluations contain instrument comprehension, aviation information, and general science. Instrument comprehension measures the ability for the student to determine altitude based on instrument readings, aviation information assesses knowledge of general aviation concepts, principles, and terms, and general science provides a measure of knowledge and understanding of scientific concepts, instruments, principles, and terms [12]. The last subtest utilizes table reading to measure the candidate's ability to quickly and accurately extract information from tables [12].

Since the adoption of the AFOQT, there have been many studies done on its validity. The Air Force Research Laboratory (AFRL) has conducted studies comparing the results from different versions of the AFOQT and adjusting the score qualifications. With regards to adjusting the score qualifications, AFRL conducted a study to observe the adverse impact of making the AFOQT qualifications more restrictive on gender and race admissions in 2006. Comparing the selection rates, AFRL found an increased adverse impact on women and minorities for officer qualifications [13].

In 1998, AFRL conducted another study attempting to quantify the barriers for entry into UPT and Navigator Training divided by the different sources (direct accessions via the Air Force Academy, Officer Training School, and Reserved Officer Training Corps, and active duty cross training from different career fields) [14]. While they did not evaluate the effectiveness of the AFOQT in this study, they used it as one of the barriers for UPT, but not for navigator training because at that time navigators cadets from the Air Force Academy were not required to take the AFOQT to become a navigator [14]. The 1998 study utilized UPT and navigator training applicants from 1992-96 and evaluated the barriers comparing gender and racial categories as well as compare the relationships between the different requirements for entry into the two trainings - Physical Training Test scores, Relative Standing Scores, Grade Point Average (GPA), the AFOQT Verbal and Quantitative scores, and a Categorized Order of Merit [14]. The Categorized Order of Merit (COM) combines the previous measures while assigning weights based on priority, assigning higher weight to GPA and Standing [14]. The results from [14], only taking into account the Air Force Academy applicants from the class of 1996, indicate a direct correlation between Standing and Board Rating, a positive correlation between GPA and selection relative to class rank, positive correlations between Military Performance Averages and Flight Screening performance relative to selection in class rank, and no trend in AFOQT Pilot scores and selection relative to class rank [14].

While these past studies accounted for the effects of the barriers on demographic applicants, the focus of the current study is similar to components of the 1996 AFRL study, the relationships between the different prerequisite tests and success in pilot training without looking at divisions for race, gender, or where the candidate is coming from.

### 2.4.2 Test of Basic Aviation Skills

The TBAS is a computer-administered cognitive and perceptual-motor-based test specifically designed to assess pilot skills directly [15]. The TBAS contains nine different subtests: the first four subtests test basic components, the next three tests combine the first four tests, one of the tests evaluates emergency procedures, and the last test assesses unmanned-aerial vehicles. Three and five digit listening tests (3DIG, 5DIG) utilize the candidate's listening skills by having them wear headphones and listen for three or five specified digits or targets when the test lists out a series of letters and numbers [15].

For example, in the three digit listening test, the candidate may have to listen for 1, 4, and 8. If they hear "A H B T U K W N Q 6 L 8 P M 9 4 X" they should click the trigger immediately after hearing the number 8 and the number 4. After the two listening tests, the candidate takes an airplane tracking test (ATT) section that measures the ability to track a moving target horizontally and vertically by keeping a set of crosshairs centered on an airplane that appears on the computer screen [15]. Next the candidate performs a horizontal tracking test (HTT) by using rudder pedals to adjust the aircraft's speed and keep the aircraft inside a box [15]. The next subtest, airplane tracking and horizontal tracking test (AHTT), combines the previous two test by having the candidate keep the plane in the box using rudder pedals while tracking another plane using crosshairs [15]. The two listening tests are then added to the testing tasks for the next two subtests, the airplane tracking, horizontal tracking, and three digit listening test (AHTT3) and the airplane tracking, horizontal tracking, and five digit listening test (AHTT5) [15]. These two tests have the candidate perform the three assigned tasks simultaneously and are graded on accuracy [15]. In the penultimate subtest called the emergency scenario test (EST), the candidate responds to audio warnings indicating an emergency situation while performing the airplane tracking and horizontal tracking tests [15]. The last subtest, the UAV Test, requires the candidate is given a UAV with its heading and a map of the ground view. The candidate must identify map locations with this information [15].

In 2005, AFRL conducted a study to validate the creation of the TBAS to replace the Basic Attributes Test (BAT) as a supplement to the PCSM [15]. This replacement was deemed necessary since the BAT hardware and software had not been updated since 1993 [15]. While testing the validity of the TBAS components, they evaluated the subtests individually combined with the AFOQT and compared the regression results with the Specialized Undergraduate Pilot Training (SUPT) T-37 outcome (whether the student passed or failed) and the T-36 total score that combines the various checks that they do. Using a sample size of 994 SUPT candidates, their results when comparing the SUPT total scores indicate a 0.0027 incremental improvement when components are added to the AFOQT [15]. Comparing the individual components of the TBAS added to the AFOQT comparing the SUPT total scores, the highest individual incremental improvement is shown when the UAV test is added to the AFOQT [15]. They also recommended prior to implementation of the TBAS a subgroup analyses and supporting documentation that conveys testing policies. Since this study is not testing the validity of the tests required for UPT, TBAS is only used as a feature for analysis being tested against the success of pilot candidates.

### 2.4.3 Pilot Candidate Selection Method

The PCSM is a composite test that previously combined the scores of the AFOQT pilot composite, several subtests from the BAT, and any previous flying experience, prior to the implementation of the TBAS [13]. AFPC processes the component results into a constantly refined model that ranks an applicant's probability to succeed in pilot training. The PCSM combines the AFOQT pilot composite, the TBAS, and any previous flying experience. In previous studies, PCSM demonstrated a direct correlation to pilot training success; the higher the PCSM score, the greater probability of completing pilot training, higher class ranking, and increased likelihood of being qualified to fly a fighter platform [13]. While there are no minimum qualification scores for the PCSM to qualify for pilot training, AFRL has suggested using a 25th percentile cutoff prior to operational implementation and a 50th percentile cutoff four years after implementation [13]. However, a 2006 study indicates that implementing these qualification standards would adversely affect female selection for pilot training and would slightly affect the minority selection into pilot training (in the latter case, the affect is only visible implementing a 50th percentile cutoff) [13].

Armstrong Laboratory conducted a study on the PCSM prior to the adoption of the TBAS in 1992 [16]. This study included the BAT and the different components associated with the test. According to their descriptions of those component subtests, the BAT included a psychomotor test that required the candidate to complete similar tasks to the ATT and the HTT subtests of the TBAS as well as tests that evaluate a candidate's willingness to take risks and process information quickly. In performing regression analysis to compare all of the tests (AFOQT, the three BAT components, and flying experience) as well as explore models utilizing combinations of the five test variables and comparing the passing or failing of UPT as well as class rank, the results indicated that individually, the AFOQT had the most impact on passing UPT and class rank with a correlation factor of 0.308 for passing UPT and 0.347 for class rank [16]. When observing the combinations of tests, the largest contributor to a change in success, when combined with the AFOQT, was flying experience observing a 3.6 percent increase in passing UPT and a 4.1 percent increase in improved class rank [16]. The improvement in correlation when all of the tests were added to the model was 7.1 percent to passing UPT and 7.9 percent to improved class rank. This current study does not utilize the TBAS, so these results are used anecdotally when determining the model combination that best predicts pilot success [16].

## 2.5 Personality Tests

Unlike the easily quantifiable qualities of pilots outlined in the other tests, personality, while testable and important, is not as easy to quantify. Many studies attempted to quantify the importance of certain personalities in certain professions. Tupes and Christal [7] were the first to furnish evidence for the five factor personality model based on their work for the United States Air Force. Much of their initial work remained unknown and kept internally until their later publication in 1992, but the work led to the development of the Self-Description Inventory (SDI) for the AFOQT. Through the collaborative efforts of The Technical Cooperation Program (TTCP), the SDI has been adopted by the militaries of the United Kingdom, Canada, New Zealand, and Australia for research purposes. For the United States Air Force, the SDI is a 220-item (changed from 163 at it's inception), trait-based measure that assesses the Big 5 domains of Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, with a measure of Machiavellianism included in the test [7].

Darr [17] performed a meta-analysis of the SDI evaluating the SDI characteristics to job performance among enlisted personnel. She examined three different lengths of the SDI: a 163-question SDI used by the USAF, a 75-question SDI used by the

Canadian Armed Forces, and a 172-question SDI used by the Royal British Navy and compared the results to data on the 34,217 military members in her study [17]. She found that the trait most positively correlated to job performance was conscientiousness, the trait that correlated the most with training performance was extraversion, and the trait that correlated the most with counter-productive work behavior was neuroticism. Darr also points out that some of the negative correlations with personality align with personalities desired for certain military situations (e.g. in the case of a soldier working with chemical or nuclear material, thinking outside the box or experimenting with new techniques could have adverse consequences) [17]. Her study also recognizes the limitation of only testing the correlations of those certain personalities to military success and the validity of performing a study within a civilian work structure [17]. Similar to Darr's study, this current study quantitatively analyzes the SDI testing results to model success in pilot training.

Prior behavioral studies were conducted by Armstrong Laboratory. One of those studies in 1992 evaluated a generic personality inventory on pilot candidates to search out desirable and undesirable characteristics with the goal of eventually implementing a personality test that searches for positive personality traits such as resiliency and filters out negative personality traits such as anxiety [18]. Siem [18] tested the Automated Aircrew Personality Profiler (AAPP) which consists of 202 items representing 16 scales from several instruments: the Minnesota Multiphasic Personality Inventory, one of the more commonly used diagnostic tools in clinical practice; the State-Trait Anxiety Inventory; the Personal Orientation Inventory, an instrument designed to assess an individual's aptitude for self-actualization; the Interpersonal Behavior Scale, which measures assertive and aggressive tendencies; and the Jenkins Activity Survey, designed to measure personality factors associated with chronic heart disease [18]. Using principal factoring with oblique rotation, five factors emerged with eigenvalues greater than 1.0 - Hostility, Self-Confidence, Values Flexibility, Depression, and Mania - on which all but one scale (Amorality) manifested factor loadings with an absolute value greater than 0.30 [18].

## 2.6 Data and Sample

The data used in this study contains 216 different factors and over 27,000 data points spread out over 12,000 UPT, Undergraduate RPA Training (URT), and Undergraduate Combat Systems Officer (CSO) Training (UCT) applicants. All three different trainings require the same components to determine a final score: Category Check T-Score, Daily Maneuver T-Score, Academic T-Score, and Flight Commander Ranking T-Score.

One difference between the three training avenues is the weighting between the Academic T-Score and the Flight Commander Ranking T-Score between the UPT/URT and UCT. For UPT and URT, the Category Check T-Score is 40 percent of the total score, the Daily Manuever T-Score is 20 percent of the total score, the Academic T-Score is 10 percent of the total score, and the Flight Commander Ranking T-Score is 30 percent of the total score. For UCT, the first two categories have the same weight, but the Academic T-Score is 30 percent of the total score and the Flight Commander Ranking T-Score is 10 percent of the total score.

The applicants in the data span the entirety of the three commissioning sources of the United States Air Force Academy (USAFA), Air Force Reserve Officer Training Corps (AFROTC), and Officer Training School (OTS). Each applicant should have at least three data points associated within the data set assuming the applicant passes every step of their pilot training. Assuming the applicant's acceptance into pilot training, they have at least one data point within the data set even if they fail the Initial Flight School (IFS). With completion of the initial flight school, the data contains all of their pre-UPT assessment scores (including AFOQT, PCSM, and TBAS) as well as all of their flight check score and their class ranking (including their class size).

## 2.7 Machine Learning Techniques

This study examines machine learning classification techniques as well as a regression technique with the data provided for the study. These techniques are reasonable given the output of this study is a binary outcome predicting whether a pilot candidate completes UPT. Classification techniques used in this study include Neural Nets, Naive Bayes, Random Forests (in the form of Bootstrap Forest), Decision Trees (in the form of Boosted Tree), and $K$-Nearest Neighbors while the regression technique examined is Logistic Regression.

### 2.7.1 Logistic Regression

The first technique applied in this study is Logistic Regression. Logistic Regression is a type of linear regression where the response variable has only two possible outcomes denoted by 0 and 1. The models in this study follow a formula similar to

$$y_i = x_i'\beta + \epsilon_i. \tag{1}$$

Since the response is binary, the error term, $\epsilon_i$ can only take on two values, namely,

$$\begin{aligned} \epsilon_i &= 1 - x_i'\beta, \quad \text{when } y_i = 1 \\ \epsilon_i &= -x_i'\beta, \quad \text{when } y_i = 0\ . \end{aligned}$$

Each variable $x_i$ is a factor in the model while the $\beta$ values are the applicable weights for the factors that help best classify each candidate data point.

Glonek and McCullagh [19] described the many ways to implement logistic regression when there are multiple factors to consider. They include two real-world applications to data sets, one of which is called ‘Six Cities Data’. The ‘Six Cities Data’ tested the health effects of air pollution using an annual binary response indicating the presence or absence of wheeze between the ages of 7 and 10 for each of 537 children from Stuebenville, Ohio [19]. In the study, the factors tested were various fits for the marginal odds for wheeze (log, bivariate, trivariate, and fourth-order), the mother’s smoking habits, the child’s age, and the interactions between the marginal odds for wheeze, the mother’s smoking habits, and the child’s age. Glonek and McCullagh compared all the different deviations and combinations of the factors and the variable fits with varying results. However, only one of their models involves the same set of factors (five models in total) with no second-order effects, interactions, or beyond tested to maintain simplicity. There is no previous literature on the UPT data to indicate necessary interactions or the need to explore second-order effects.

### 2.7.2 Decision Trees

Some machine learning classification techniques utilize variations of Decision Trees. Decision Trees are a type of directed, acyclic graph where the nodes represent decisions and the branches have two or more descendent nodes representing possible paths from one node to another defined by the algorithm of the decision tree.

Friedl and Brodley used multiple types of Decision Trees in their study on the “Classification of Land Cover from Remotely Sensed Data” [20]. The three types they examined were Univariate Decision Trees, Multivariate Decision Trees, and Hybrid Decision Trees. They cite several advantages for decision trees over traditional classification procedures used in the remote sensing realm: trees are strictly non-parametric, trees do not require assumptions regarding the distributions of input data, and trees handles nonlinear relations between features and classes, allow for missing values, are capable of handling both numeric and categorical inputs, and their classification structure is explicit making it easily interpretable [20]. To assess the performance of the Decision Tree technique, they used cross-validation by splitting the data into three parts: 70 percent training, 20 percent pruning, and 10 percent testing [20]. The pruning step removes inaccurate classifications. These steps are necessary in their implementation of Decision Trees because the data set they used contains 11 different classifications of their dependent variable for each of the different terrain features in the land data they considered. Since the dependent variable in this study only classifies between two options, pruning is not necessary.

This study uses the Boosted Tree classification in JMP-13 Pro. Boosted tree takes a large additive decision tree and builds upon it using layers. Each layer is built using the residuals of the previous layer which corrects for any poor model fit. The final prediction of an observation is the sum of all the predictions from the previous layers.

### 2.7.3 Random Forests

Another Decision Tree variation used in this study is Random Forest classification. Random Forest classification utilizes multiple decision tree classifications that are combined using bootstrap aggregation, or bagging. Bootstrap aggregation is a technique that reduces overfitting and improves the outcome of learning on limited sample size. Bagging works by creating some $M$ specified subsets of the data with $n$ samples per subset. The $n$ samples are uniformly sampled with replacement from the original data set. For each bootstrapped sample, the labels are preserved. Next, $k$ individual learning modules are created for each bootstrapped sample. The outputs of these modules are then combined. JMP-13 Pro uses Bootstrap Forest classification which is an application of Random Forest classification. This utilization of Bootstrap Forest and Boosted Tree assesses the accuracy of the models and predictors.

### 2.7.4 Naive Bayes

Naive Bayes is a classification technique that uses Bayes' theorem (below) to categorize data.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

The Naive Bayes algorithm is called naive because of the assumptions made about the data. One assumption is that all of the features of the data set are of equal importance and are independent [21]. However even if these assumptions are violated, this algorithm still performs well [21]. Depending on the number of variables taken into account in the model, Naive Bayes could take longer to run as model complexity increases [21].

Rish [22] studied Naive Bayes classification to find the scenarios that best utilized the strengths of the technique. He found that the accuracy of Naive Bayes is not directly correlated with the amount of feature dependencies measured [22]. However, he did find that a better predictor of accuracy is the loss of mutual information between features [22]. This study uses Naive Bayes to assess the accuracy of the models and predictors.

### 2.7.5 Neural Nets

Neural nets is short for neural networks. This technique emulates neurons in the human brain respond to stimuli with sensory inputs. Neural Nets use hand-labeled data that feed into a system of input nodes. Those input nodes feed forward in one direction into other layers of nodes. Each connection between layers in the neural net have certain weights associated with them that are multiplied by the value from the previous node. As the data moves forward through the neural net, the products in every node are added together and compared to a threshold value within the layer as depicted in equation 3.

$$y(x) = f(\sum_{i=1}^{n} w_i x_i) \tag{3}$$

If the sum is below the threshold value, no information is passed. However, if the sum exceeds the threshold value, the node "fires" like in the brain, sending the sum to the next layer until the data are correctly classified. Initially, the weights assigned between the nodes are random values. As the data flows through the neural net, those random weights are adjusted until the training data with the same labels consistently yields similar results. Figure 1 depicts an example of a Neural Net.

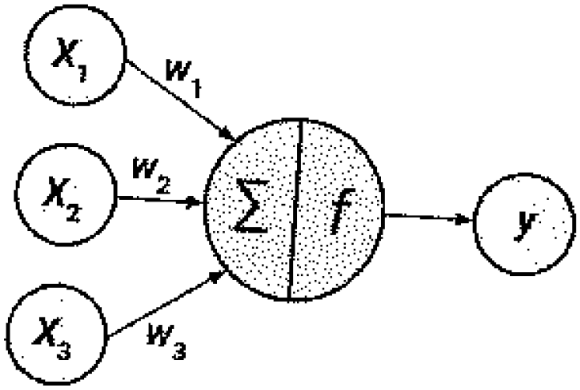

**Figure 1. Neural Net Example [21]**

Many studies use neural nets for pattern recognition. Ciresan used neural nets in handwriting recognition [23]. He trained five models with two to nine layers containing a different amount of nodes in each layer utilizing a test-to-training set ratio of 1:6 with 70,000 total data points [23]. He cites an error percentage of less than 1 percent for all of the models. However, each of the input layers has between 1,000 and 9,000 neurons, has multiple layers, and run times between 23 and 115 hours [23].

Contrary to the Ciresan study, the current use of Neural Nets uses a larger portion for testing since this data set is smaller than Ciresan's used for handwriting recognition. This study uses Neural Nets to assess the accuracy of the models and predictors. JMP-13 Pro can create one-layer or two-layer Neural Nets that utilize hyperbolic tangent nodes, linear nodes, and Gaussian nodes. Those one-layer or two-layer Neural Nets can then be boosted by a specified number of models. The number of models specified by boosting takes the predicted values from the first model, scales them using the learning rate, them subtracts them from the actual values to produce a scaled residual. These residuals are then fit to a new model where the responses are scaled residuals of the previous model. This process continues until the number of models are completed or the addition of a new model fails to improve the validation statistic. The hyperbolic tangent transforms values to be between -1 and 1, and is centered on the scaled version of the logistic function. The hyperbolic tangent function follows equation 4

$$\frac{e^{2x}-1}{e^{2x}+1}. \tag{4}$$

The linear transformation is the identity function. The linear combination of factors is not transformed. This transformation is most often used in conjunction with one of the nonlinear activation functions. In this case, the linear activation function is in the second layer while the nonlinear activation functions are in the first layer. For a continuous target variable, if only linear activation functions are used, the model for the target variable reduces to a linear combination of the factors. For a nominal target variable, the model reduces to a logistic regression. The Gaussian function is useful for radial basis function behavior or when the response surface is Normal in shape. This function follows equation 5, where $x$ is a linear combination of the factors

$$e^{-x^2}. \tag{5}$$

### 2.7.6 K-Nearest Neighbors

$K$-Nearest Neighbors is both a classification technique and an imputation technique that classifies and imputes missing data. The technique works by labeling feature data or blank data based on feature data that is in close proximity. For each example in the data, the algorithm calculates the distance between that example and the neighbor examples already added. The distance and the index are then added to an ordered collection that are sorted from smallest to largest. The algorithm then picks the first $k$ entries and obtains their labels, yielding the mode of the classification. For the implementation of $K$-Nearest Neighbors, the optimal $k$ varies based on the number of factors in the models.

Horton and Nakai used $K$-Nearest Neighbors as a way to predict the cellular localizations sites of proteins in yeast and E.coli [24]. They compared KNN with Decision Trees, and Naive Bayes, as well as their own probabilistic model, and found that KNN performs consistently among the best in both data sets [24]. They determined unique $k$ values for each data set using leave-one-out cross validation [24]. Like the Horton and Nakai study, this study utilizes $K$-Nearest Neighbors to assess the accuracy of the models and predictors. However, KNN will also be used to impute missing values to enhance the accuracy of the unknown data.

This chapter addressed the data collected in the study and previous studies done on subsets of the data. Some of the studies conducted were more recent but not conducted on potential pilots or actual aviators, while other studies were conducted on pilot candidates but are not reflective of the current tests. With the previous studies on the data observed, the next section of this chapter identified machine learning classification techniques that can be applied to the data, how those techniques work, and previous studies conducted using those techniques. The next chapter addresses the specific applications of the machine learning techniques on the data used for this study.

# III. Methodology

This study tests machine learning classification techniques as well as a regression technique using the data provided described in the previous section as well as demographic data. The demographic data assessed includes race, gender, commissioning source, college, major, and whether the candidate was enlisted or an officer. Classification techniques used in this study include Neural Nets, Naive Bayes, Bootstrap Forest, Boosted Tree, and $K$-Nearest Neighbors while the regression technique tested is Logistic Regression.

## 3.1 Assumptions/Delimitations

This study assumes that the UPT student data were entered accurately and that the student records maintained their integrity during the data transfer from AFPC/DSYX to AFIT/ENS. The delimitation of this analysis is that causes for statistically significant differences are not proposed. This study aims only to analyze pilot training applicants and students; it does not address the process prior to pilot candidate selection nor does this study directly address how to increase USAF pilot recruitment.

## 3.2 Implications

This study identifies areas for further research regarding which tests, personalities, or other performance measures indicate varying levels of success in UPT for both course completions and distinguished graduates (DGs). If a coherent model is developed, it is possible to tailor training and recruitment based on the applicable model factors. Various other pilot-affiliated organizations and commissioning sources could use these results to determine candidates worthy of scholarships and which candidates would be more competitive applicants for UPT.

## 3.3 Analysis Process

Prior to a data analysis, the data must be cleaned and verified, and missing values imputed before performing the analysis. The previous study conducted by Meyer [2] did not implement analytical tools, meaning there are no definitive models defined for predicting UPT completions and DGs. This suggests the need for using multiple exploratory multivariate techniques before attempting to create a model that utilizes only the significant factors. Multivariate analysis techniques are implemented in Python and JMP-13 Pro to handle the computations.

### 3.3.1 Data Cleaning & Imputation

The data were cleaned to get a more accurate representation of UPT candidates. Of the 27,894 data points and the 232 columns of data, 34 columns had less than 11 percent entries filled and only 10 columns had 100 percent entries with non-blank values. Some columns are easily filled such as the “Age at Class Start” which is calculated using the difference between the “Start Date” and “Date of Birth” columns. Other columns received imputed values based on whether the student completed the course. Because of those different imputation considerations, the data split into the students that completed courses and those that did not. Once the split was made, logical imputations were made. For the students that completed the course, their “Attrition Reason” became “N/A” while the ones that did not complete the course had an “Unknown” attrition reason. The students that did not have policy waivers had “N/A”. The students that did not have prior enlisted experience had “0” put in the appropriate columns. The Python code utilized for data cleaning and imputation can be found in Appendix A.

Using the splits of passing student data and failing student data, the data set is first imputed exhausting all the implications from the literature review and previous thesis conducted. After completing the logical data imputations, the columns are then evaluated using a percent fill criteria. Since a majority of the important test data and UPT check and composite score columns had at least 60 percent fill, that was the criteria for the passing student data leaving 198 columns. That data entries were then filtered for any blank entries within the row, downsizing a subset that originally contained 25,832 entries to 9,995 complete entries. To maintain that balance with the failed student subset, those same 198 columns from the passing student subset were used for the failed student subset. The filtering for any blank entries within the rows downsized the failed student subset from 2,062 entries to 55 complete entries. The two subsets were then recombined to make a working data set with 10,050 total data points. After being combined again, the columns were evaluated to ensure a balance of categorical feature data.

After the data were combined and filtered, 26 of the 199 features contained categorical data. To use that categorical data, those features are One-hot encoded which involves giving each category within a feature their own binary feature which reveals whether that category is used or not. The categorical features were then evaluated for the number of categorical bins and then filtered to minimize the number of bins per feature. Once complete, the categorical features that contained more than 20 identifiable categories were removed bringing the number of features down from 199 to 188. Applying one-hot encoding, the number of features then increased from 188 to 292 features that contained more than one entry.

Running the data through each classification technique and logistic regression chose to classify every prediction as a pass while accepting the losses. This was attributed to the lack of balance between UPT successes and failures (0.55 percent of the total remaining data is failures compared to 99.45 percent successes). This imbalance led to restarting the data cleaning and imputation process.

From the initial 232 features, features with dates, selected categorical features, and features with less than 60 percent fill were eliminated. This time, after one-hot encoding the categorical features, the remaining columns with missing data were imputed using $k$-nearest neighbors (with $k$=1 and $k$=7 to test accuracy results once the machine learning techniques were implemented). This maintained a 27,894-point data set, but increased the number of features from 179 to 283.

For the distinguished graduate (DG) study, a new feature was created, 'Class Percentile'. 'Class Percentile' was calculated by dividing the 'Class Rank' feature by the 'Class Size' feature. 'DG Indicator' was then made by taking the candidates that graduate in the top 10 percent or better and assigning them a '1' while all other candidates were assigned a '0'. The new binary variable is used as a target variable for another study and not as a feature for the UPT Completion study. The addition of the new feature increased the total number of features to 284.

### 3.3.2 Predictive Screening

Once the data set was finalized, the JMP 'Predictive Screening' tool was used on the complete data set containing 284 factors. JMP-13 Pro uses Bootstrap Forest to rank-order the given factors in order from most applicable to least applicable to the target variable. Bootstrap Forest is a combination of bootstrapping and random forest classification. JMP-13 Pro bootstraps the specified training data set by extrapolating data points. It then creates many decision trees off of that data set, and then takes an average of the results of the decision trees to complete the Bootstrap Forest.

Through the 'Predictive Screening' tool in JMP, the overall best factor in predicting who will complete pilot training is the 'Flight Experience Hours' factor. To obtain improved fidelity in predicting UPT completion, three different subsets of data were created depicting different stages of the UPT process: the first subset included all of the factors a candidate has prior to acceptance into UPT to include demographic information (i.e. race, age, flight experience, etc.) called the preemptive subset, the second subset includes only the test scores of the tests that the candidates take prior to UPT acceptance (i.e. AFOQT, PCSM, personality tests, etc.) called the test subset, and the third subset includes all of the factors that a candidate is tested on while at UPT (i.e. check scores, online hours, offline hours, etc.) called the post-acceptance subset.

After making the three subsets of data, predictive screening was performed again to provide a starting point of relevant factors for making models on the three subsets. The factors included in the model for the preemptive subset are ‘Flight Experience hours’, ‘Age at Class Start’, and ‘Dominance/Leader’ which is a result of a personality test conducted by the AFOQT. The factors included in the model for the test subset, ‘TCO’, ‘Reading Comprehension’, ‘Dominance/Leader’, ‘Well-Adjusted’, ‘Interpersonal Tactics’, ‘Team Player’, ‘AFOQT-Pilot’, and ‘PCSM Score’. ‘Well-Adjusted’, ‘Interpersonal Tactics’, and ‘Team Player’, are also personalities evaluated on the AFOQT. Lastly, the factors included in the model for the post-acceptance subset are ‘Flight Count’, ‘Offline Aircraft hours’, ‘Device Experience hours’, ‘Device Aircraft hours’, ‘Online Count’, ‘Offline Count’, ‘Require Flight Count’, ‘Check Score’, and ‘Flight Aircraft hours’.

Through the ‘Predictive Screening’ tool in JMP, the top eight best factors in predicting DGs came from the test scores students obtain while attending UPT, which are: ‘Composite Score’, ‘Daily T-Score’, ‘Flight Weighted Score’, ‘Check T-Score’, ‘Flt/CC [Flight Commander] T-Score’, ‘Flt/CC Raw Score’, ‘Daily Score’, and ‘Academic T-Score’. Like the analysis observing students completing UPT, the same three subsets were used to test DGs. But when ‘Predictive Screening’ was used on the Preemptive and Test subsets, the relevant factors between the two subsets were nearly identical. This prompted the consolidation of the ‘Preemptive’ subset and the ‘Test’ subset into a singular ‘Pre-Acceptance’ subset for the DG study. For the DG study, two subsets were observed: a ‘Pre-Acceptance’ subset that includes all of the entry tests required, demographic information, and prior flying experience; and a ‘Post-Acceptance’ subset that includes all of the tests that a student performs at UPT.

After distinguishing the two subsets for the DG study, the ‘Predictive Screening’ tool in JMP was used to provide a starting point of relevant factors in predicting DGs. For the ‘Pre-Acceptance’ subset, the relevant factors included in the model were ‘Flight Experience (in hours)’, ‘FAA Flight Hours’, ‘PCSM Score’, and ‘AFOQT-Pilot Score’. The relevant factors included in the ‘Post-Acceptance’ model for DGs were ‘Composite Score’, ‘Daily T-Score’, ‘Flight Weighted Score’, and ‘Check T-Score’.

### 3.3.3 Factor Building

Once the factors were determined, a factor analysis was used on each set of factors. If there are any correlations between the factors, then factor analysis will group similar factors together so that a weighted combination of the grouped factors can be created for the final model based on the importance of each factor within the group. Prior to creating the formula for each new aggregate factor, the sub factors are normalized using the Min-Max normalization technique which divides the difference between each data point and the minimum by the difference between the maximum and the minimum of each factor as displayed 6

$$\frac{v - v_{min}}{v_{max} - v_{min}}. \tag{6}$$

For the preemptive subset, no such aggregate factors are made. For the test subset, the AFOQT-Pilot score and the PCSM scores are combined together based on equation 7, now called the 'AFOQT-Pilot & PCSM Score'.

$$0.6 * AFOQT\ Pilot\ Score\ +\ 0.4 * PCSM\ Score. \tag{7}$$

Also in the test subset, 'Dominance/Leader', 'Well-Adjusted', 'Interpersonal Tactics', and 'Reading Comprehension' scores are combined based on equation 8 hereby called the 'Leadership Personalities & Reading Comprehension Score'.

$$\begin{split}0.3 * Well\ Adjusted\ Score\ +\ 0.3 * Interpersonal\ Tactics\ Score\\ +\ 0.25 * Dominance/Leader\ Score\ +\ 0.15 * Reading\ Comprehension\ Score.\end{split} \tag{8}$$

For the post-acceptance subset, 'Flight Count', 'Required Flight Count', and 'Flight Aircraft hours' are combined together based on equation 9 hereby called the 'Combined Flight Score'.

$$\frac{1}{3}Flight\ Count\ +\ \frac{1}{3}Required\ Flight\ Count\ +\ \frac{1}{3}Flight\ Aircraft\ Hours. \tag{9}$$

Also in the post-acceptance subset, 'Offline Aircraft hours', 'Online Count', and 'Offline Count' are combined based on equation 10 and hereby called the 'Combined Online-Offline Score'.

$$0.4 * Offline\ Aircraft\ Hours\ +\ 0.4 * Offline\ Count\ +\ 0.2 * Online\ Count. \tag{10}$$

Finally, in the post-acceptance subset, ‘Device Experience hours’ and ‘Device Aircraft hours’ are combined together based on 11 below and hereby called the ‘Combined Device Score’.

$$0.5 * Device\ Aircraft\ Hours\ +\ 0.5 * Device\ Experience\ Hours. \tag{11}$$

Another factor analysis was performed on the two DG models, but no significant relationships were observed between the two sets of factors.

## 3.4 Technique Parameters

After building the necessary factors, each model was evaluated using five machine learning techniques and a regression technique. These techniques used accommodate the binary response outcome of whether a pilot candidate completes UPT. The classification techniques used are: Neural Nets, Naive Bayes, Bootstrap Forest, Boosted Tree, and $K$-Nearest Neighbors while the regression technique tested is Logistic Regression. To measure the effect of ‘Flight Experience (in hours)’, Boosted Tree, Bootstrap Forest, and Neural Nets were run solely on that one factor and compared to the preemptive model. For these techniques, 33 percent of the data are utilized for training while the remaining 67 percent of the data are utilized for testing.

### 3.4.1 Logistic Regression

Logistic Regression is linear regression where the response variable has only two possible outcomes denoted by ‘0’ and ‘1’. In the first study, successfully passing UPT is a ‘1’ and failing UPT is a ‘0’. In the second study, graduating UPT in the top 10 percent is a ‘1’ while the remaining 90 percent is a ‘0’.

For all of the logistic regression models, the target level is 1. Interactions were considered, but rejected. The Preemptive model follows the form of

$$y = 1.38917363 \ + \ 0.00140082 * Flight\ Experience\ Hours\ + \\ 0.04420651 * Age\ at\ Class\ Start \ + \ (-0.2343582) * Dominance/Leader\ Score. \tag{12}$$

The Test model follows the form of

$$y = 2.69867331 \ + \ (-0.0001581) * Team\ Player\ Score \ + \ (-0.000031721) * TCO \\ + \ 1.39851123 * AFOQT\ \&\ PCSM\ Score \\ + \ (-2.2995235) * Leadership\ Personalities\ \&\ Reading\ Comprehension\ Score. \tag{13}$$

The Post-Acceptance model follows the form of

$$y = (-0.0408055) \ + \ 0.03417692 * Check \ + \ (-3.4968712) * Combined\ Flight\ Score \\ + \ (-3.4561414) * Combined\ Online/Offline\ Score \\ + \ 9.39623253 * Combined\ Device\ Score. \tag{14}$$

The Pre-Acceptance Model for the Distinguished Graduate analysis follows the form of

$$y = (-2.768466) \ + \ (-0.0103508) * Flight\ Experience\ hours \\ + \ (-0.0137626) * AFOQT - Pilot\ Score \ + \ 0.00071588 * FAA\ Flight\ Hours \\ + \ 0.02639282 * PCSM\ Score. \tag{15}$$

The Post-Acceptance Model for the Distinguished Graduate analysis follows the form of

$$\begin{aligned} y = (-39.96381) \ + \ 0.53020677 * Composite\ Score \\ + \ (-0.0966235) * Flight\ Weighted\ Score \ + \ 0.12023264 * Daily\ T - Score \\ + \ 0.05206896 * Check\ T - Score. \end{aligned} \tag{16}$$

### 3.4.2 Machine Learning Techniques

Some of the machine learning classification techniques utilize variations of Decision Trees. One of the Decision Tree variations utilized in this study is Boosted Tree classification. Boosted tree takes a large additive decision tree and adds layers. Each layer is built using the residuals of the previous layer which corrects any poor fitting. The final prediction of an observation is the sum of all the predictions of the previous layers.

The settings employed for Boosted Tree were as follows: 50 layers with a minimum five splits and three splits per tree, a 10 percent learning rate and an overfitting penalty of 0.01 percent. All instances of Boosted Tree fully examine all of the rows and columns specified for the models. The layers of Boosted Tree for each model follows a similar format to Figure 2.

The settings employed for Bootstrap Forest were as follows: 100 trees in the forest with one term sampled per split, a minimum of 10 splits per tree, a maximum of 2,000 splits per tree, and a minimum split size of 27. Additionally, a 1 percent bootstrap sample rate was applied in this technique. No other specific settings were needed to implement Naive Bayes. Naive Bayes had no other specified settings outside of the specified partition for testing and training.

Each implementation of Neural Nets uses a 'Holdback' validation method for split-

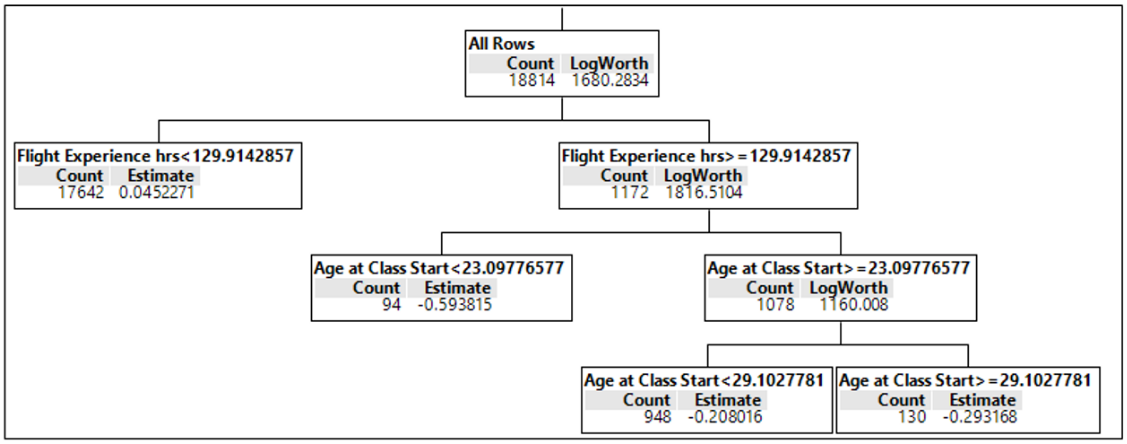

**Figure 2. Boosted Tree Layer 1 for the Preemptive Model**

ting the data set into test and train with the same 33 percent-67 percent ratio split. For the structures, each net has two layers composed of three hyperbolic tangent sigmoid nodes, three linear nodes, and three Gaussian nodes with five model iterations for boosting. This means that each implementation of Neural Nets contains 90 nodes. Each Neural Net utilizes a learning rate of 10 percent and squared penalty method that fits one tour. The implementations of Neural Nets for each model follows a similar format to Figure 3.

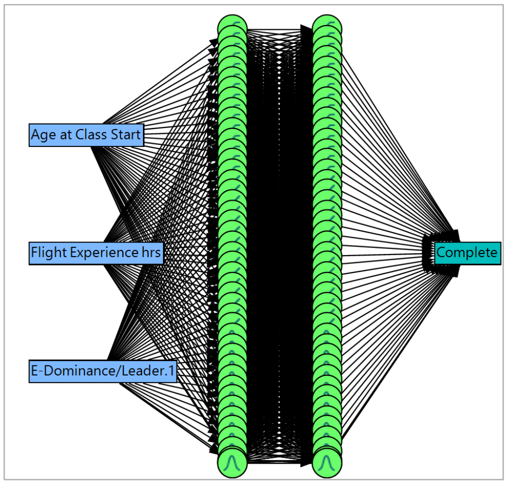

**Figure 3. Neural Net for the Preemptive Model**

For the implementation of $K$-Nearest Neighbors, the optimal $k$ varies based on the number of factors in the models. The more variables included in the model, the greater value of $k$ required to maximize predictive power. As the $k$-value approaches the number of variables in the model, the model approaches the maximum value of accurate classifications. The way JMP implements $K$-Nearest Neighbors to a given data point is by choosing the $k$ smallest Euclidean distance between predictor values of that data point and the predictor values of the surrounding data points. This process is completed for all data points and then repeated for a specified $k$ iterations. The $k$-values in the confusion matrices in Appendix B for the completion analysis and Appendix C for the Distinguished Graduate Analysis are the values that maximize prediction accuracy and reflect the best performing iteration.

# IV. Analysis

## 4.1 Pilot Training Completion Analysis

The machine learning techniques and models are compared using the type I error rates (false positives) and accuracy of predicting successes and failures which are depicted in the histograms below. Each technique applied to each model is assessed by maximizing the accuracy of predicting both failures and successes while minimizing the amount of type I errors (false positives). In this study, there are 25,832 students that complete UPT and 2,062 students that fail to complete UPT. The confusion matrices for each technique applied in each model can be found in Appendix B. Corresponding Receiver Operator Characteristic (ROC) Curves for Neural Nets, Bootstrap Forest, Boosted Tree, and Logistic Regression are found in Appendix D, but do not add anything to the actual results of the technique comparisons.

### 4.1.1 Preemptive Model

The Preemptive model utilizes a candidate's 'Flight Experience (in hours)', age at the start of UPT, and their 'Dominance/Leader' score. Since 'Flight Experience (in hours)' is the best predictive indicator of pilot success, that individual factor used in Neural Nets, Bootstrap Forest, and Boosted Tree is also compared to the created model.

For the Preemptive model, Logistic Regression and the Neural Net that only utilizes a candidate's 'Flight Experience (in hours)' are ineffective since they predict all candidates as successful completions. Evaluating the techniques in the preemptive model, Boosted Tree, $K$-Nearest Neighbors, and Bootstrap Forest utilizing 'Flight Experience' are the best techniques in predicting failures in UPT. By a narrow margin, $K$-Nearest Neighbors outperforms all other machine learning techniques tested because it more accurately predicts failures (see Figure 4) and it has one of the smallest false positive rates (see Figure 5). The model itself seems to have large variance of predicting UPT failures, which works better with a Neural Net, Boosted Tree, and $K$-Nearest Neighbors.

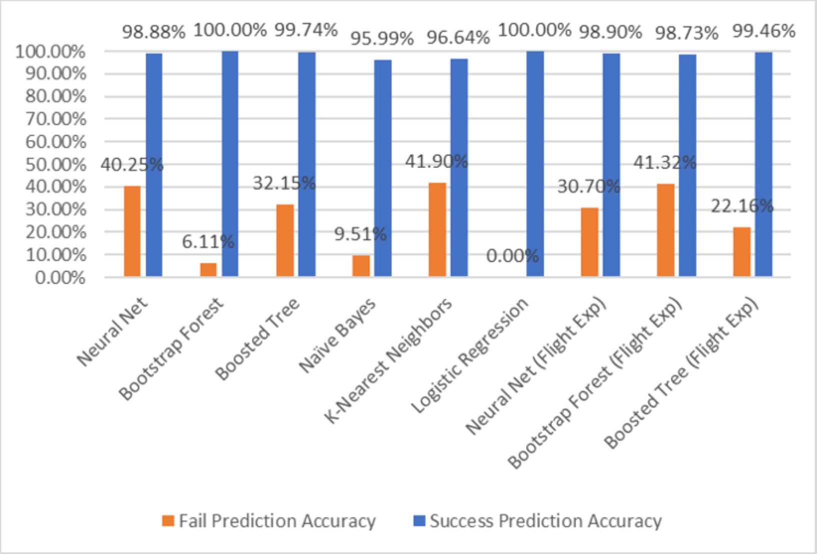

**Figure 4. Preemptive Model Accuracy**

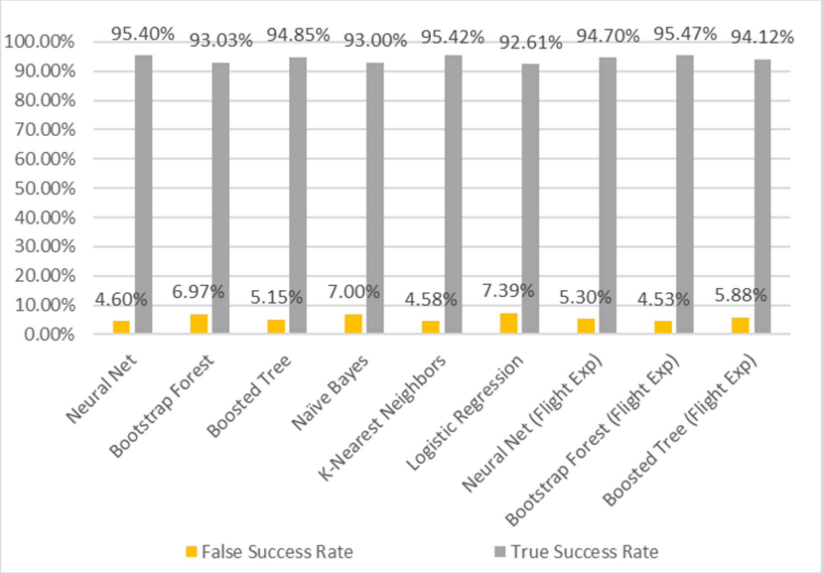

**Figure 5. Preemptive Model Success Rates**

### 4.1.2 Test Model

The Test model utilizes a candidate's TCO score, their 'Team Player' score, their 'AFOQT-Pilot & PCSM score', and their 'Leadership Personalities & Reading Comprehension score'.

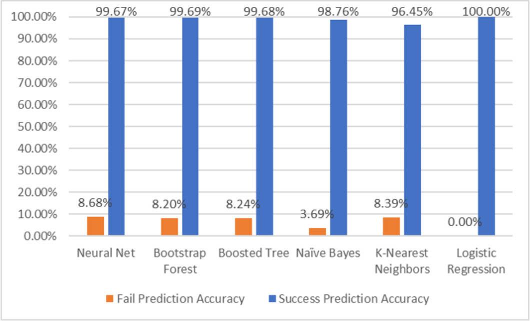

**Figure 6. Test Model Accuracy**

For the Test model, Logistic Regression is ineffective since it predicted all points as successful completions. Besides the Logistic Regression model, all other machine learning techniques perform similarly. Naive Bayes, however, performs the second worst with around 4 percent of successful failures predicted. By a narrow margin, Neural Net outperforms the other machine learning techniques because it predicts the most failures (see Figure 6) and has the smallest type I error rate (see Figure 7). Comparing the Test model to the Preemptive Model, the Preemptive model predicts failures better than the Test model.

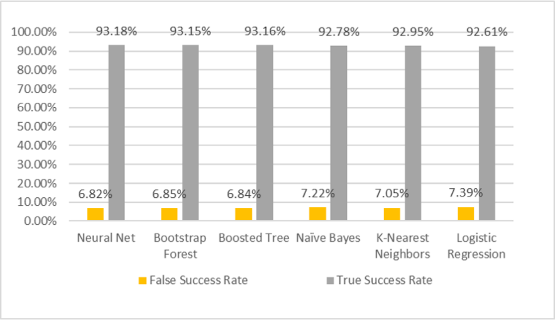

**Figure 7. Test Model Success Rates**

### 4.1.3 Post-Acceptance Model

The Post-Acceptance model utilizes a candidate's 'Flight Check' score, their 'Combined Online-Offline score', their 'Combined Flight Score', and their 'Combined Device Score'.

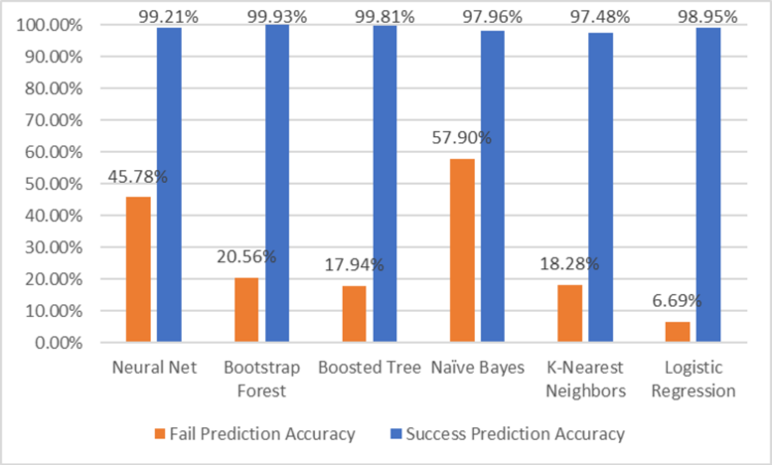

**Figure 8. Post-Acceptance Model Accuracy**

For the Post-Acceptance model, Naive Bayes substantially outperforms all other machine learning techniques by predicting more than twice as many failures as the other techniques (see Figure 8) and almost half as many type I errors (see Figure 9).

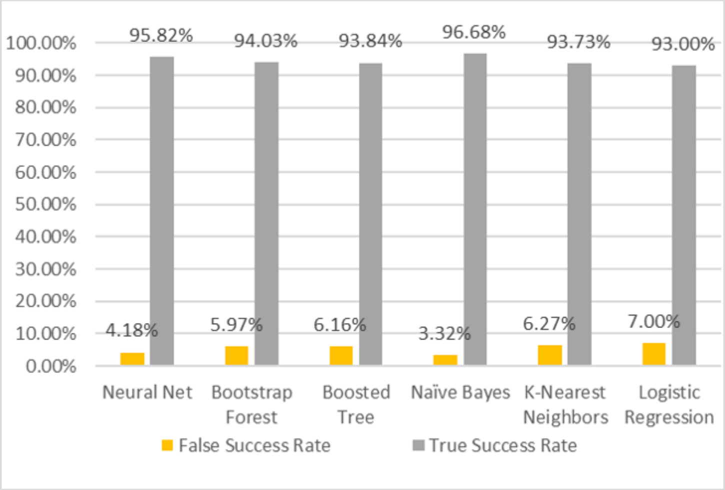

**Figure 9. Post-Acceptance Model Success Rates**

Comparing the Post-Acceptance model to the other two models, there are some techniques where the Preemptive model outperforms the Post-Acceptance model. Figure 10 depicts the best performing model for each technique in terms of predicting UPT failures and false success rates. Based on the results in Figure 10, half of the best performers utilize the Preemptive model (excluding the best Bootstrap Forest output that solely utilizes 'Flight Experience in hours'), but the best performing model is when Naive Bayes utilizes the Post-Acceptance model.

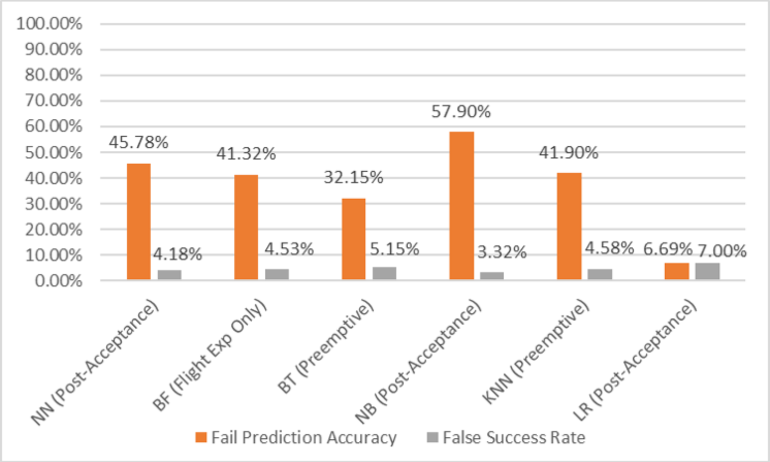

**Figure 10. Best Model Comparisons**

## 4.2 Distinguished Graduate Analysis

This next analysis assesses the same factors but changes the criteria from whether a pilot candidate succeeds or fails to whether a candidate is eligible to finish as a distinguished graduate (DG). The purpose is to determine what factors in certain points of the UPT process could best predict whether a pilot candidate could graduate as a DG. In this analysis, a DG is a binary variable with a success defined as a candidate that graduates UPT in the top 10% of their class. Redefining success as being a DG shifts the balance between success and failure even further than the completion criteria. In this data set, there are 1,338 pilot candidates that fulfill the defined DG criteria while there are 26,556 that do not fulfill that criteria.

The techniques compared in this study are the same techniques used in the UPT completion study, however instead of evaluating the techniques for false positives and accurate predictions of both successes and failures, they are evaluated for false predictions and accurate DG predictions. The full confusion matrices for each Machine Learning technique applied to each model are in Appendix C. The ROC curves for each technique applied to each model are in Appendix E but do not add much to the analysis.

### 4.2.1 Pre-Acceptance Model

The Pre-Acceptance Model is composed of four factors: 'Flight Experience (in hours)', 'FAA Flight Hours', 'AFOQT - Pilot' Score, and 'PCSM Score'. Using the same techniques in the completion analysis, the Pre-Acceptance model does a poor job of predicting DGs. Based on the results found in Figure 11, all of the techniques applied inaccurately predicted DGs with the two best techniques predicting roughly 12 percent (Naive Bayes) and 10 percent ($K$-Nearest Neighbors) of the DGs. On a positive note, all of the machine learning techniques applied to this model yield low incorrect prediction rates with all lower than 6 percent.

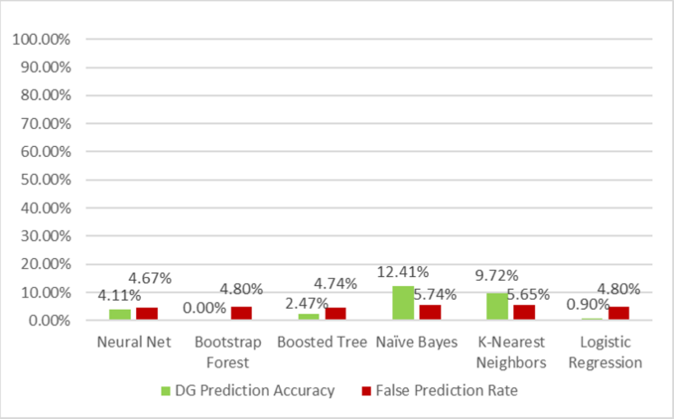

**Figure 11. Pre-Acceptance Distinguished Graduate Model Accuracy**

### 4.2.2 Post-Acceptance Model

The Pre-Acceptance Model is composed of four factors: 'Composite Score', 'Daily T Score', 'Flight Weighted Score', and 'Check T Score'. Comparing this model to the Pre-Acceptance model, this Post-Acceptance model does a much better job of predicting DGs. Based on the results in Figure 12, all techniques accurately predicted at least half of the DGs with most techniques hovering between 60 percent and 67 percent roughly five times more than the Pre-Acceptance model. Naive Bayes performs the best by predicting roughly 87 percent of the 1,338 DGs more than seven times the performance of the Naive Bayes in the Pre-Acceptance model. All of the techniques applied have low incorrect prediction rate with most techniques maintaining below a 3 percent false prediction rate. The Naive Bayes is the only exception maintaining a false prediction rate of roughly 5 percent.

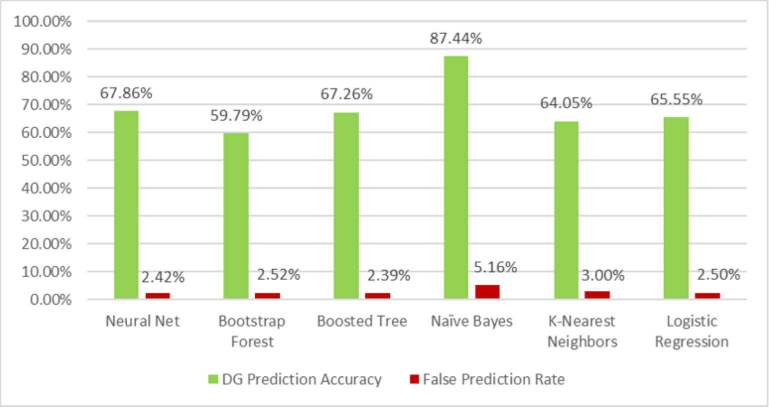

**Figure 12. Post-Acceptance Distinguished Graduate Model Accuracy**

# V. Conclusion

## 5.1 Insights on Pilot Training Completion

This study has confirmed the USAF selection criteria for acceptance into the UPT program. Above all other factors in the application process, the factor that best predicts that a candidate will complete UPT is, not surprisingly, how well a candidate performs at UPT. The amount of flight experience a candidate shows enough promise to warrant further investigation. Of the different tests that a candidate takes prior to acceptance into UPT, some are more applicable to UPT, such as the Pilot Score of the AFOQT and the PCSM Score as these appear to be the best indicators of UPT completion. Overall, flight experience by itself and UPT performance are better indicators of pilot success than the AFOQT-Pilot Score and the PCSM Score.

## 5.2 Insights on Distinguished Graduates

This study confirms the USAF selection process for DG are the best performers at UPT. While good pilot-related AFOQT scores, PCSM scores, and prior flight experience are beneficial for successfully completing UPT, those factors do not accurately predict whether a pilot candidate will graduate in the top 10 percent of their class. There is some underlying factor besides the residual error that prevents some of those top performers from earning DG. That could be explained by some sort of disciplinary or other factor unable to be collected.

## 5.3 Areas of Further Research

While the data included in this study uses a lot of demographic information to include commissioning source, race, gender, and age, there is some demographic information that is not included such as university or undergraduate major that had so many unique categorical factors that could not be reasonably simplified. Some information that would capture similar information that is not included in this study would be whether the degrees the candidate has earned are technical or non-technical degrees as well as what region or state the candidate is from.

The data used was sparse, at least 60 percent fill was used for this study. The features that were not completely filled were imputed using $K$-Nearest Neighbors. Imputation can lead to the inaccuracy of some of the models and factors. A clearer data picture might have made Naive Bayes more accurate in some of the models. Previous research also indicates that Random Forests would work better if the dependent variable values are balanced between the classes. Including all applications might give more balance to the data. While the start date of a UPT class was included in this study, it's primary use was to calculate the age of a student at the start of a class. The class start date could also be used to perform a time-series analysis to explore any possible cyclical trends in UPT data over time.

As more pilot candidates go through the Pilot Training Next (PTN) program, there could be a comparison study to the traditional UPT tracks. This study could also expand into a joint-service study to compare the quality of pilot training between the military services. Instead of having to estimate whether a candidate earns DG, the accuracy of the study would be improved if that information is added to the data set from the candidate's records.

# Appendix A. Python Code for Data Cleaning and Imputation

```python
#Necessary Packages
import pandas as pd
import numpy as np
import datetime as dt
import time
from dateutil.relativedelta import relativedelta
from datetime import date
from sklearn import preprocessing

#upt_data2 = pd.read_excel(r’D:\Thesis Work\Merged data - cleaned for Capt Uber - 24 Aug.xlsx’)
filename = ’Merged data - cleaned for Capt Uber - 24 Aug.xlsx’
upt_data = pd.read_excel(filename)
upt_guide = pd.read_csv(r’D:\Thesis Backup\Thesis Work\Data for Modeling\UPT Test Data3.csv’)

#Observing column fills
upt_pct = upt_data.count()/27894

#Adding a class percentile
upt_class_pct = upt_data["Class Rank"] / upt_data["Class Size"]
upt_data.insert(12, "Class Percentile", upt_class_pct, True)

#Changing the format of the Date columns.
upt_data["Date of Birth"] = pd.to_datetime(upt_data["Date of Birth"],format=’%Y%m%d’)
upt_data["Class Start"] = pd.to_datetime(upt_data["Class Start"], format=’%Y%m%d’)
upt_data["Completion Date"] = pd.to_datetime(upt_data["Completion
Date"],format='%Y%m%d')
upt_data["Attrition Date"] = pd.to_datetime(upt_data["Attrition Date"],format='%Y%m%d')
upt_data["Entrance Active Duty"] = pd.to_datetime(upt_data["Entrance Active Duty"],format='%Y%m%d')

#Removing unusable columns.
upt_data = upt_data.drop(['Platform earned', 'Acft Type', 'Major', 'Undergraduate', 'Years Prior Service'], axis=1)

#Other Imputations & Fixes
upt_data["Time of Enlisted Yrs"].fillna(0,inplace=True)
upt_data["Time of Enlistment Months"].fillna(0,inplace=True)
upt_data["Policy Waiver"].fillna("N/A",inplace=True)
upt_data['Complete'] = upt_data['Complete'].replace(2,1)
#Fixes for Course column
upt_data.loc[upt_data['Course'].str.contains('P-V4A-G',case=False), 'Course'] = 'P-V4A-G'
upt_data.loc[upt_data['Course'].str.contains('3 (T-38C)',case=False), 'Course'] = 'P-V4A-N-3'
upt_data.loc[upt_data['Course'].str.contains('(T-6)',case=False), 'Course'] = 'P-V4A-N'
upt_data.loc[upt_data['Course'].str.contains('CV4PA',case=False), 'Course'] = 'C-V4P-A'
upt_data.loc[upt_data['Course'].str.contains('PV4C-E SUPTH',case=False), 'Course'] = 'P-V4C-E SUPTH'
#Fixes for Track column
upt_data['Track'].fillna('missing',inplace=True)
upt_data.loc[upt_data['Track'].str.contains('3B',case=False), 'Track'] = '3B'
upt_data.loc[upt_data['Track'].str.contains('3A',case=False), 'Track'] = '3A'
upt_data.loc[upt_data['Track'].str.contains('INTERNATIONAL BF',case=False), 'Track'] = 'INTL'
upt_data.loc[upt_data['Track'].str.contains('missing',case=False), 'Track'] = np.nan
upt_data['Track'].unique()
#Fixes for the Education column
upt_data['Education'].fillna('missing',inplace=True)
upt_data.loc[upt_data['Education'].str.contains('e.g.',case=False), 'Education'] = 'Undergraduate Degree'
upt_data.loc[upt_data['Education'].str.contains('missing',case=False), 'Education'] = np.nan
upt_data['Education'].unique()

#Data Checks
upt_data['Policy Waiver'].unique()
upt_data["Attrition Reason"].unique()
pd.value_counts(upt_data['Course'].values, sort=True)

#Trying to find a pattern of data fill to find imputation techniques.
upt_pass_data = upt_data[upt_data['Complete'] != 0]
upt_fail_data = upt_data[upt_data.Complete == 0]
upt_pass_pct = upt_pass_data.count()/25832
upt_fail_pct = upt_fail_data.count()/2062

#More Data Checks
upt_pass_data['Attrition Reason'].unique()
upt_fail_data['Attrition Reason'].unique()
upt_data['Composite Score'].describe()
upt_fail_data['Composite Score'].describe()

#Fail Data Imputation
```

```python
upt_fail_data['Attrition Reason'].fillna('UNKNOWN',inplace=True)
upt_fail_data['Completion Date'].fillna('N/A',inplace=True)
upt_fail_data['Age at Class Start'] = upt_fail_data['Attrition Date'] - upt_fail_data['Date of Birth']
upt_fail_data['Age at Class Start'] = upt_fail_data['Age at Class Start']/np.timedelta64(1,'Y')

#Pass Data Imputation
upt_pass_data['Attrition Reason'].fillna('None',inplace=True)
upt_tl = pd.Timedelta(pd.offsets.Day(730))
grad_date = upt_pass_data['Class Start'] + upt_tl
upt_pass_data['Attrition Date'].fillna(grad_date,inplace=True)
upt_pass_data['Age at Class Start'] = upt_pass_data['Class Start'] - upt_pass_data['Date of Birth']
upt_pass_data['Age at Class Start'] = upt_pass_data['Age at Class Start']/np.timedelta64(1,'Y')

#Imputation Exploration Data Checks
upt_pass_data['Class Size'].describe()
upt_pass_data['Class Size'].median()
upt_data['Age at Class Start'].describe()

#After Logical Imputations, append Pass & Fail Data.
upt_data2 = upt_pass_data.append(upt_fail_data)
upt_pct2 = upt_data2.count()/27894

#Removing date formatted columns.
upt_data2 = upt_data2.drop(['Date of Birth', 'Completion Date', 'Class Start', 'Attrition Date', 'Entrance Active Duty', 'AFOQT Date Tested', 'DATE of Original Scan', 'Date Scanned'], axis=1)

#Trying to find a pattern of data fill to find imputation techniques
```

```
    .
upt_pass_data2 = upt_data2[upt_data2['Complete'] != 0]
upt_fail_data2 = upt_data2[upt_data2.Complete == 0]
upt_pass_pct2 = upt_pass_data2.count()/25832
upt_fail_pct2 = upt_fail_data2.count()/2062

#Obtaining Testable Copies of the Datasets
upt_test = upt_data2.copy()
upt_ftest = upt_fail_data2.copy()
upt_ptest = upt_pass_data2.copy()
upt_ptest_pct = upt_ptest.count()/25832
upt_test_pct = upt_test.count()/27894
upt_ftest_pct = upt_ftest.count()/2062

#Remove columns with <60% fill
#Imputation Column Fill Tolerance
tol = 0.6

#Step 0: Remove columns with <60% fill.
upt_data2.dropna(axis=1,thresh=27894*tol,inplace=True)

#Test 1: No further imputation
upt_test.dropna(axis=1,thresh=27894*tol,inplace=True)
upt_test.dropna(axis=0, how='any', inplace=True)

upt_data2.dropna(axis=1,thresh=27894*tol,inplace=True)
upt_data2.dropna(axis=0, how='any', inplace=True)

pd.value_counts(upt_data2['Complete'].values, sort=True)

#Test 2: w/ Separate Filtering Tolerance Criteria
#Column Filtering for Pass Data.
```

```
upt_ptest.dropna(axis=1, thresh=25832*tol, inplace=True)
upt_ptest.dropna(axis=0, how=’any’, inplace=True)

#Column Filtering for Fail Data.
tol2 = 0.08
upt_ftest.dropna(axis=1, thresh=2062*tol2, inplace=True)
upt_ftest.dropna(axis=0, how=’any’, inplace=True)

#Testing Age Imputation
upt_ftest_ac[’Date of Birth’].describe()
upt_f_age = upt_ftest_ac[’Attrition Date’] - upt_ftest_ac[’Date of Birth’]
upt_f_age = upt_f_age/np.timedelta64(1,’Y’)
upt_ftest_ac[’Age at Class Start’].fillna(upt_f_age,inplace=True)
#Testing Class Info Imputations
upt_ftest[’Class Start’].fillna(upt_ftest[’Attrition Date’],inplace=True)
upt_ftest[’Class Size’].fillna(50,inplace=True)
upt_ftest[’Class Rank’].fillna(50,inplace=True)
upt_ftest[’Class Percentile’] = upt_ftest[’Class Rank’]/upt_ftest[’Class Size’]

#Categorical Imputation
upt_test[’Track’].fillna(’missing’,inplace=True)
upt_test[’Officer or Enlisted’].fillna(’missing’,inplace=True)
upt_test[’Gender’].fillna(’missing’,inplace=True)
upt_test[’Reserve or Guard’].fillna(’missing’,inplace=True)
upt_test[’TBAS Version’].fillna(’missing’,inplace=True)
upt_test[’Education’].fillna(’missing’,inplace=True)
upt_test[’Status’].fillna(’missing’,inplace=True)
upt_test[’Aero Rating’].fillna(’missing’,inplace=True)
upt_test[’Source’].fillna(’missing’,inplace=True)
```

```
upt_test['Race'].fillna('missing',inplace=True)
upt_test['EDLEV'].fillna('missing',inplace=True)
upt_test['DEGREE'].fillna('missing',inplace=True)

upt_test = upt_test.drop(['Class'], axis=1)
upt_test = upt_test.drop(['Key # to match'], axis=1)
upt_test = upt_test.drop(['EDLEV'], axis=1)
upt_test = upt_test.drop(['DEGREE'], axis=1)
upt_test = upt_test.drop(['TestDate'], axis=1)
upt_test = upt_test.drop(['ProcDate'], axis=1)
upt_test = upt_test.drop(['Status'], axis=1)
upt_test = upt_test.drop(['Age at Exam'], axis=1)
upt_test = upt_test.drop(['Required Flight Check'], axis=1)


upt_test_ohe = pd.get_dummies(upt_test)
upt_test_ohe = upt_test_ohe.loc[:, (upt_test_ohe != 0).any(axis=0)]
upt_test_ohe_labels = list(upt_test_ohe.columns.values)
upt_test_ohe_pct = upt_test_ohe.count()/27894

#K-Nearest Neighbors Imputation
pip install missingpy
from missingpy import KNNImputer
imputer = KNNImputer(n_neighbors=7, weights="uniform")
upt_test_knni7 = imputer.fit_transform(upt_test_ohe)
upt_knni7_pct = upt_knni7.count()/27894

upt_knni7 = pd.DataFrame(upt_test_knni7, columns=upt_test_ohe_labels)


#Fail Test Data Checks
```

```
upt_ftest_ac['Daily Score'].describe()
upt_ftest_ac['Daily T Score'].describe()
upt_ftest_ac['Check'].describe()
upt_ftest_ac['Check T Score'].describe()
upt_ftest_ac['Flight Weighted Score'].describe()
upt_ftest_ac['Academic'].describe()

upt_ptest_ac['Completion Date'].describe()
pd.value_counts(upt_ptest_ac['Completion Date'].values, sort=True)
upt_ptest_ac = upt_ptest_ac.drop(['Attrition Date'], axis=1)
upt_ftest_ac = upt_ftest_ac.drop(['Attrition Date'], axis=1)
upt_ptest_ac = upt_ptest_ac.drop(['Completion Date'], axis=1)
upt_ftest_ac = upt_ftest_ac.drop(['Completion Date'], axis=1)

#Column Filtering for Pass Data.
upt_ptest_ac.dropna(axis=1, thresh=25832*tol, inplace=True)
upt_ptest_ac.dropna(axis=0, how='any', inplace=True)
upt_ptest_ac_pct2 = upt_ptest_ac.count()/9995

#Column Filtering for Fail Data.
tol2 = 0.08
hdrs_upt_ptest_ac = list(upt_ptest_ac.columns.values)
upt_ftest_ac = upt_ftest_ac[hdrs_upt_ptest_ac]
upt_ftest_ac2 = upt_ftest_ac.copy()
upt_ftest_ac.dropna(axis=1, thresh=2062*tol2, inplace=True)
upt_ftest_ac.dropna(axis=0, how='any', inplace=True)

#After Logical Imputations, append Pass & Fail Data.
upt_data3 = upt_ptest_ac.append(upt_ftest_ac)
upt_pct2 = upt_data2.count()/27894

#Sanity Checks
```

```
pd.value_counts(upt_test[’Officer or Enlisted’].values, sort=True)
pd.value_counts(upt_test[’Attrition Reason’].values, sort=True)
pd.value_counts(upt_test[’Training Location’].values, sort=True)
pd.value_counts(upt_test[’Course’].values, sort=True)
pd.value_counts(upt_test[’Gender’].values, sort=True)
pd.value_counts(upt_test[’Race’].values, sort=True)
pd.value_counts(upt_test[’Source’].values, sort=True)
upt_test[’Age at Class Start’].describe()
upt_test[’Required Flight Check’].unique()


#Exporting to a CSV
upt_knni7.to_csv(’UPT Data KNN7 Imputed.csv’)
```

# Appendix B. Pilot Training Completion Analysis Confusion Matrices

Confusion matrices are used on the validation sets for each technique for comparison, with the exception of logistic regression which uses the entire data set.

**Table 1. Confusion Matrix of Logistic Regression for the Preemptive Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 0 | 2,062 |
| **1** | 0 | 25,832 |
| $n$ = 27,894 | | |

**Table 2. Confusion Matrix of Neural Net for the Preemptive Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 267 | 421 |
| **1** | 81 | 8,529 |
| $n$ = 9,298 | | |

**Table 3. Confusion Matrix of Bootstrap Forest for the Preemptive Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 14 | 642 |
| **1** | 0 | 8,593 |
| $n$ = 9,249 | | |

**Table 4. Confusion Matrix of Boosted Tree for the Preemptive Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 260 | 428 |
| **1** | 51 | 8,503 |
| $n$ = 9,242 | | |

**Table 5. Confusion Matrix of K-Nearest Neighbors for the Preemptive Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 308 | 404 |
| **1** | 286 | 8,307 |
| $n$ = 9,305; $k$ = 2 | | |

**Table 6. Confusion Matrix of Naive Bayes for the Preemptive Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 60 | 635 |
| **1** | 268 | 8,300 |
| $n$ = 9,263 | | |

**Table 7. Confusion Matrix of Neural Net for Flight Experience Only Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 276 | 412 |
| **1** | 173 | 8,437 |
| $n$ = 9,298 | | |

**Table 8. Confusion Matrix of Bootstrap Forest for Flight Experience Only Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 245 | 454 |
| **1** | 85 | 8,472 |
| $n$ = 9,256 | | |

**Table 9. Confusion Matrix of Boosted Tree for Flight Experience Only Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 128 | 565 |
| **1** | 35 | 8,560 |
| $n$ = 9,242 | | |

**Table 10. Confusion Matrix of Logistic Regression for the Test Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 0 | 2,062 |
| **1** | 0 | 25,832 |
| $n$ = 27,894 | | |

**Table 11. Confusion Matrix of Neural Net for the Test Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 60 | 628 |
| **1** | 33 | 8,577 |
| $n$ = 9,298 | | |

**Table 12. Confusion Matrix of Bootstrap Forest for the Test Model**

| Actual | Predicted Count | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 55 | 626 |
| **1** | 23 | 8,521 |
| $n$ = 9,225 | | |

**Table 13. Confusion Matrix of Boosted Tree for the Test Model**

| Actual | Predicted Count | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 63 | 600 |
| **1** | 28 | 8,536 |
| $n$ = 9,227 | | |

**Table 14. Confusion Matrix of K-Nearest Neighbors for the Test Model**

| Actual | Predicted Count | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 32 | 665 |
| **1** | 148 | 8,386 |
| $n$ = 9,231; $k$ = 4 | | |

**Table 15. Confusion Matrix of Naive Bayes for the Test Model**

| Actual | Predicted Count | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 82 | 600 |
| **1** | 457 | 8,062 |
| $n$ = 9,201 | | |

**Table 16. Confusion Matrix of Logistic Regression for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 138 | 1,924 |
| **1** | 272 | 25,560 |
| $n$ = 27,894 | | |

**Table 17. Confusion Matrix of Neural Net for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 320 | 368 |
| **1** | 65 | 8,545 |
| $n$ = 9,298 | | |

**Table 18. Confusion Matrix of Bootstrap Forest for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 197 | 520 |
| **1** | 26 | 8,549 |
| $n$ = 9,292 | | |

**Table 19. Confusion Matrix of Boosted Tree for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 143 | 529 |
| **1** | 13 | 8,537 |
| $n$ = 9,222 | | |

**Table 20. Confusion Matrix of K-Nearest Neighbors for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 414 | 283 |
| **1** | 170 | 8,432 |
| $n$ = 9,299; $k$ = 4 | | |

**Table 21. Confusion Matrix of Naive Bayes for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 125 | 537 |
| **1** | 231 | 8,323 |
| $n$ = 9,216 | | |

# Appendix C. Distinguished Graduate Analysis Confusion Matrices

Confusion matrices are used on the validation sets for each technique for comparison, with the exception of logistic regression which uses the entire data set.

**Table 22. Confusion Matrix of Logistic Regression for the Pre-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 26,542 | 1,326 |
| **1** | 14 | 12 |
| $n$ = 27,894 | | |

**Table 23. Confusion Matrix of Neural Net for the Pre-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,849 | 3 |
| **1** | 429 | 17 |
| $n$ = 9,298 | | |

**Table 24. Confusion Matrix of Bootstrap Forest for the Pre-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,829 | 0 |
| **1** | 432 | 0 |
| $n$ = 9,261 | | |

**Table 25. Confusion Matrix of Boosted Tree for the Pre-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,761 | 15 |
| **1** | 416 | 15 |
| $n$ = 9,207 | | |

**Table 26. Confusion Matrix of K-Nearest Neighbors for the Pre-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,749 | 112 |
| **1** | 385 | 42 |
| $n$ = 9,288; $k$ = 4 | | |

**Table 27. Confusion Matrix of Naive Bayes for the Pre-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,666 | 143 |
| **1** | 383 | 59 |
| $n$ = 9,251 | | |

**Table 28. Confusion Matrix of Logistic Regression for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 26,321 | 235 |
| **1** | 461 | 877 |
| $n$ = 27,894 | | |

Table 29. Confusion Matrix of Neural Net for the Post-Acceptance Model

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,765 | 87 |
| **1** | 148 | 298 |
| $n$ = 9,298 | | |

Table 30. Confusion Matrix of Bootstrap Forest for the Post-Acceptance Model

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,708 | 64 |
| **1** | 178 | 257 |
| $n$ = 9,207 | | |

Table 31. Confusion Matrix of Boosted Tree for the Post-Acceptance Model

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,743 | 83 |
| **1** | 147 | 303 |
| $n$ = 9,276 | | |

Table 32. Confusion Matrix of K-Nearest Neighbors for the Post-Acceptance Model

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,738 | 115 |
| **1** | 138 | 283 |
| $n$ = 9,274; $k$ = 4 | | |

**Table 33. Confusion Matrix of Naive Bayes for the Post-Acceptance Model**

| **Actual** | **Predicted Count** | |
|---|---|---|
| **Complete** | **0** | **1** |
| **0** | 8,358 | 443 |
| **1** | 56 | 378 |
| $n$ = 9,235 | | |

# Appendix D. Pilot Training Completion Analysis ROC Curves

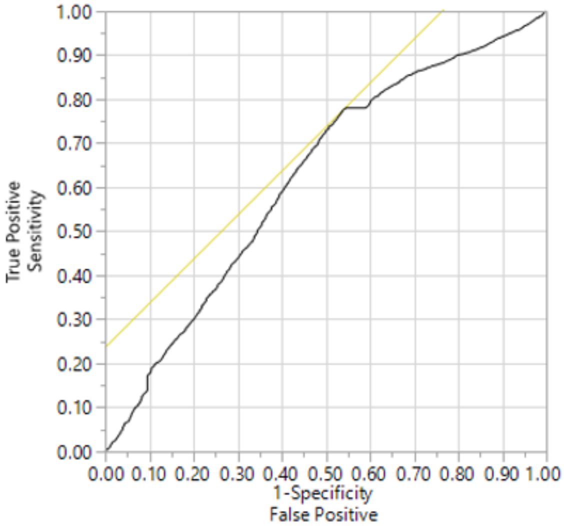

**Figure 13. Preemptive Model ROC Curve - Logistic Regression**

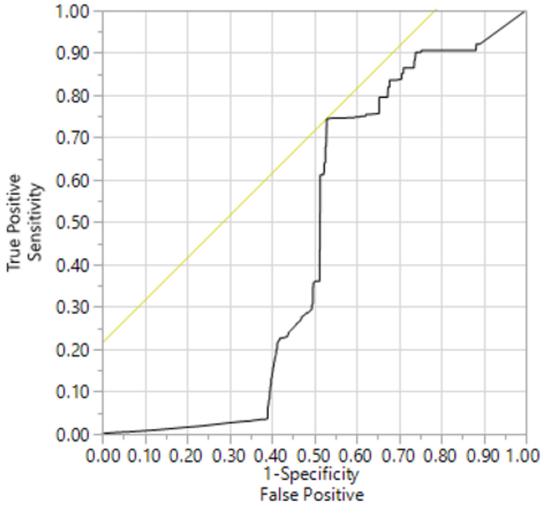

**Figure 14. Preemptive Model ROC Curve - Logistic Regression (Flight Experience Only)**

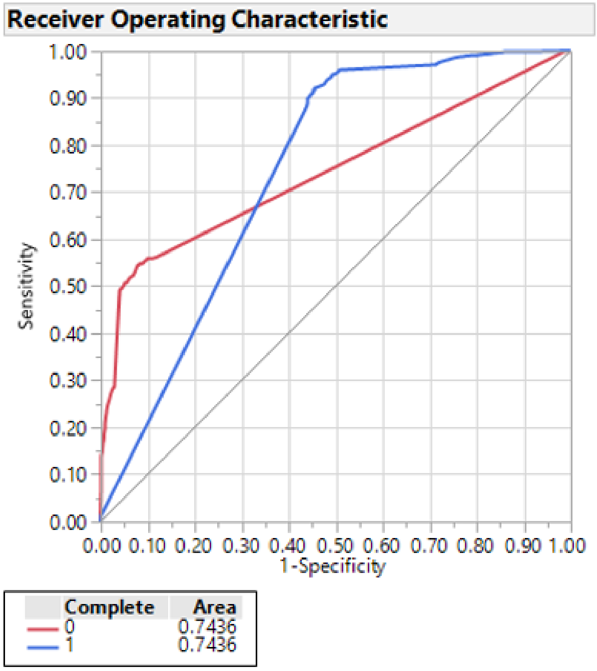

Figure 15. Preemptive Model ROC Curve - Bootstrap Forest

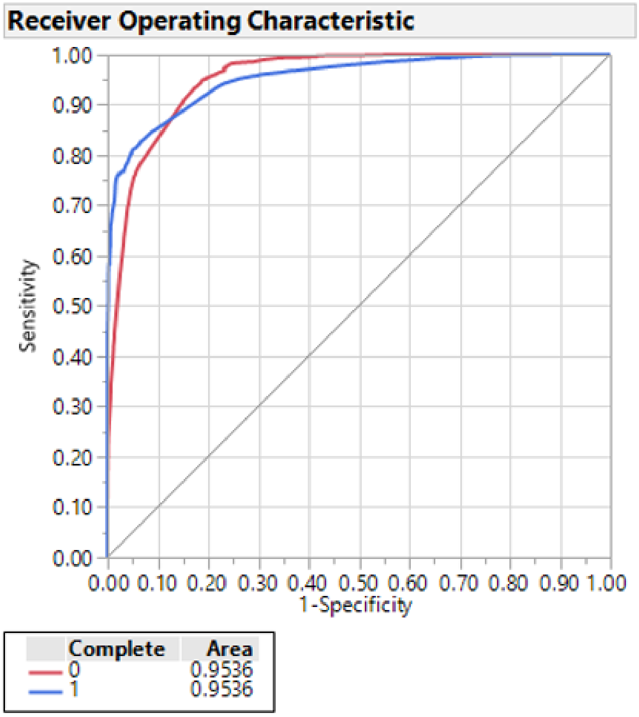

Figure 16. Preemptive Model ROC Curve - Bootstrap Forest (Flight Experience Only)

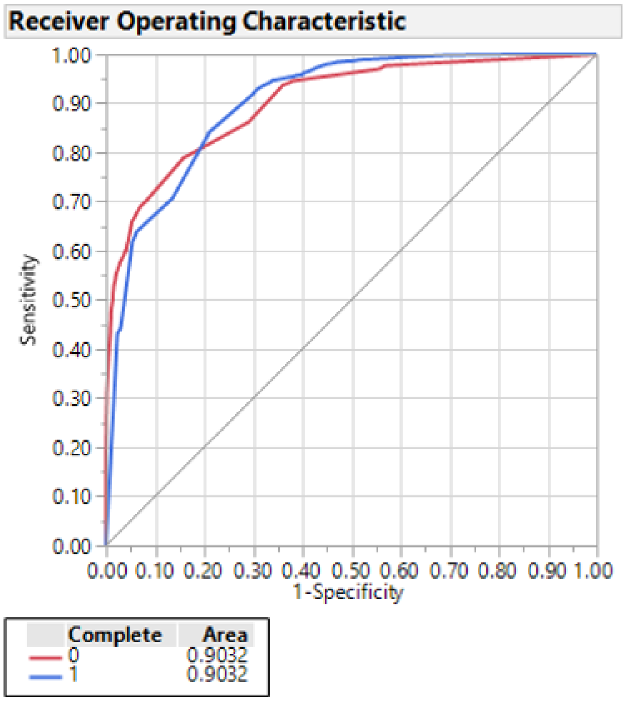

Figure 17. Preemptive Model ROC Curve - Boosted Tree

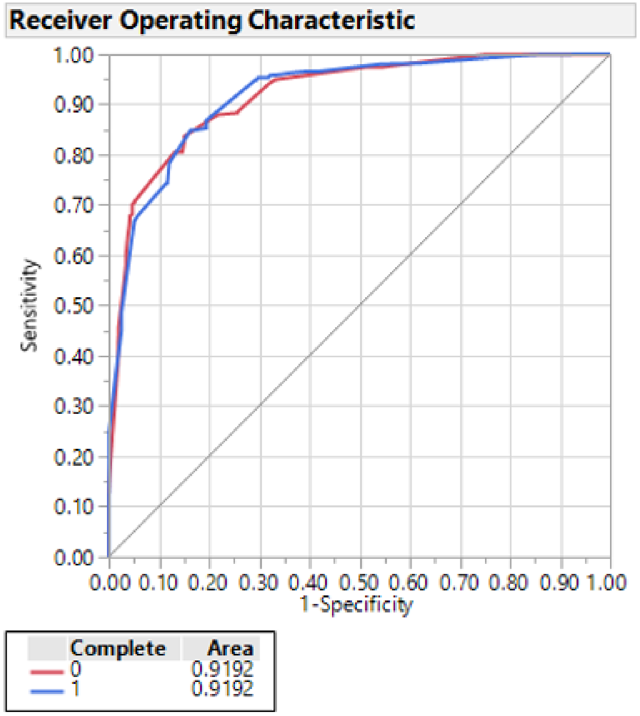

Figure 18. Preemptive Model ROC Curve - Boosted Tree (Flight Experience Only)

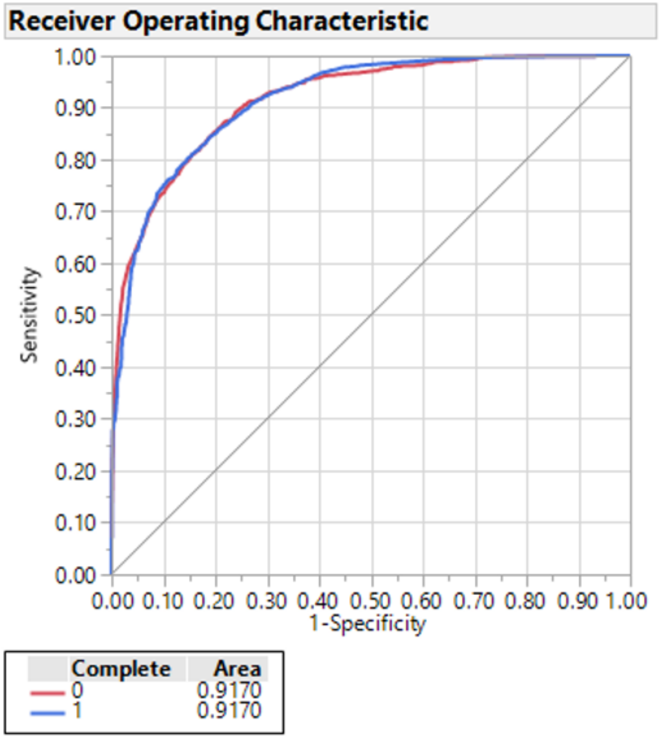

Figure 19. Preemptive Model ROC Curve - Neural Nets

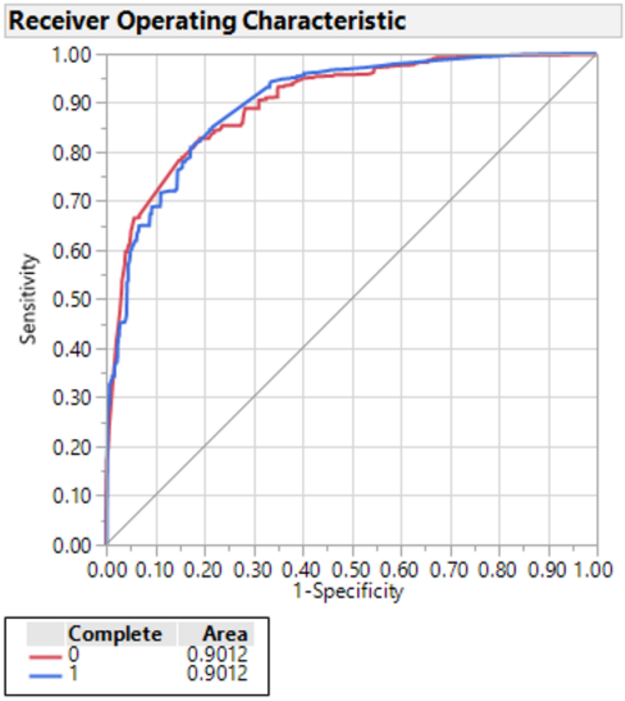

Figure 20. Preemptive Model ROC Curve - Neural Nets (Flight Experience Only)

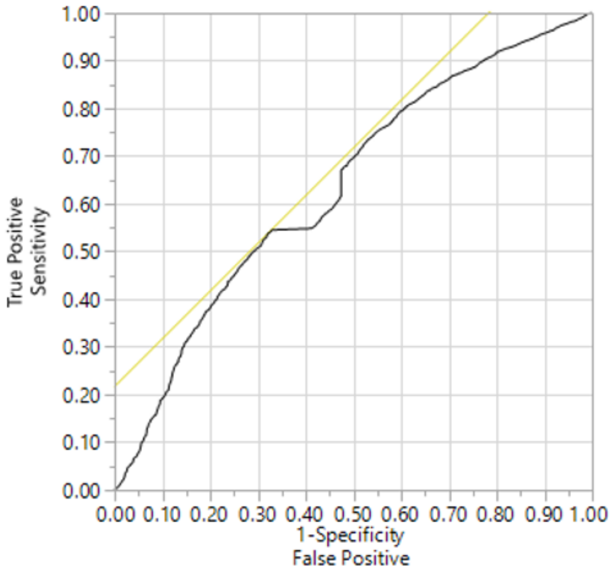

**Figure 21. Test Model ROC Curve - Logistic Regression**

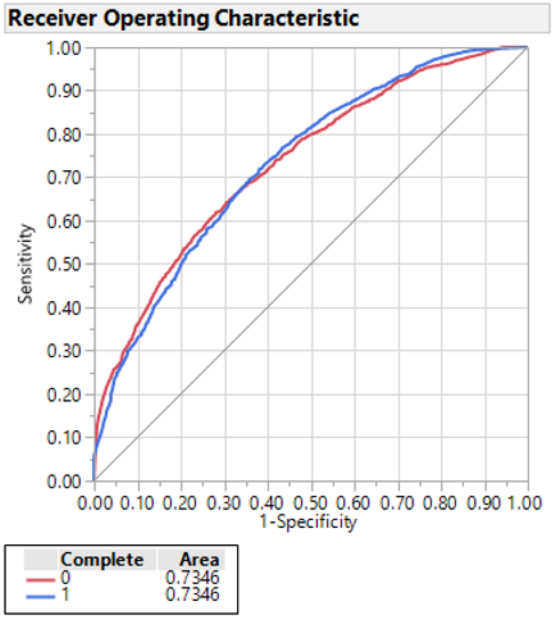

**Figure 22. Test Model ROC Curve - Bootstrap Forest**

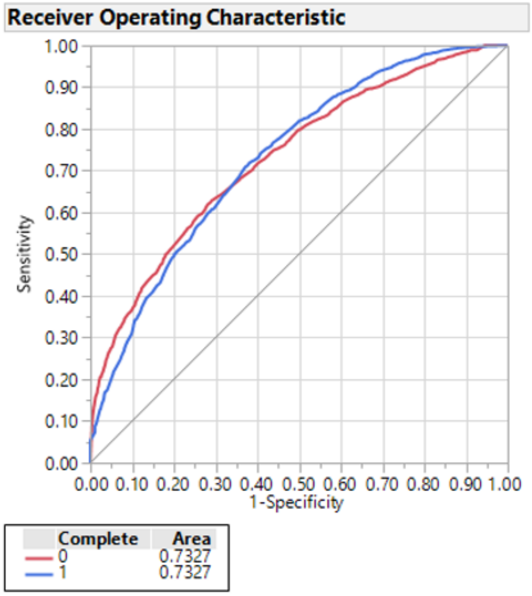

**Figure 23. Test Model ROC Curve - Boosted Tree**

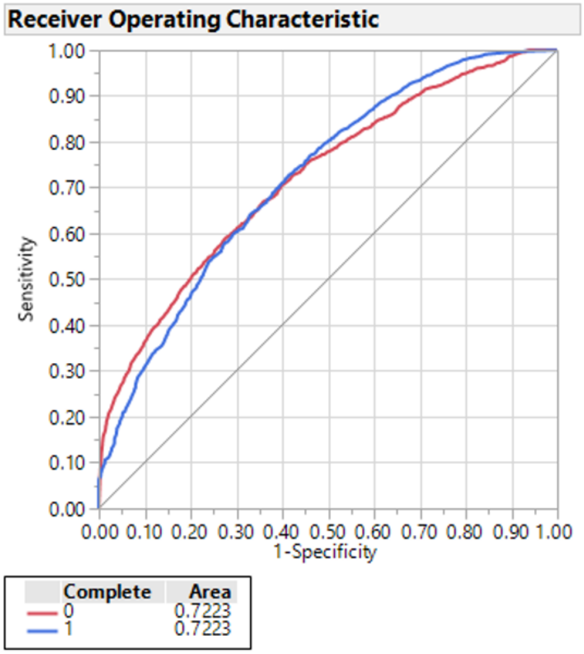

**Figure 24. Test Model ROC Curve - Neural Nets**

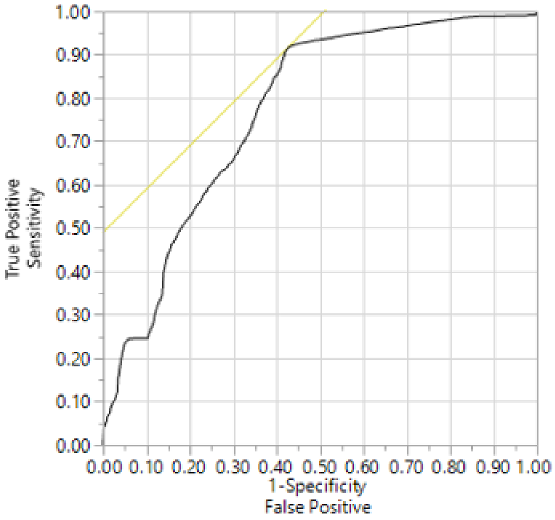

**Figure 25. Post-Acceptance Model ROC Curve - Logistic Regression**

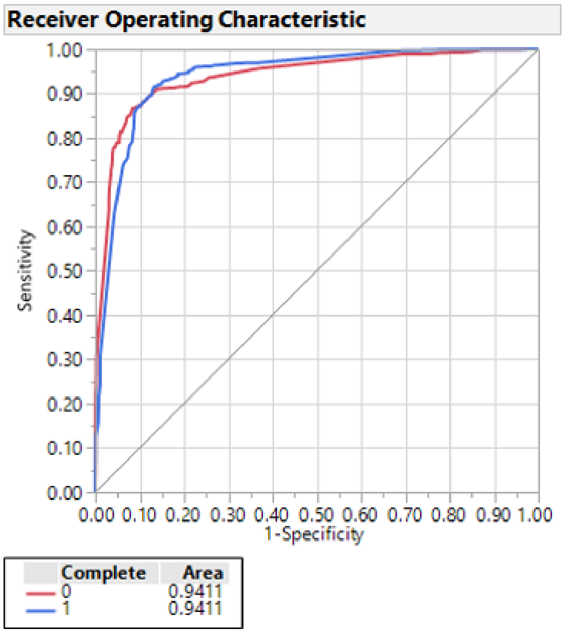

**Figure 26. Post-Acceptance Model ROC Curve - Bootstrap Forest**

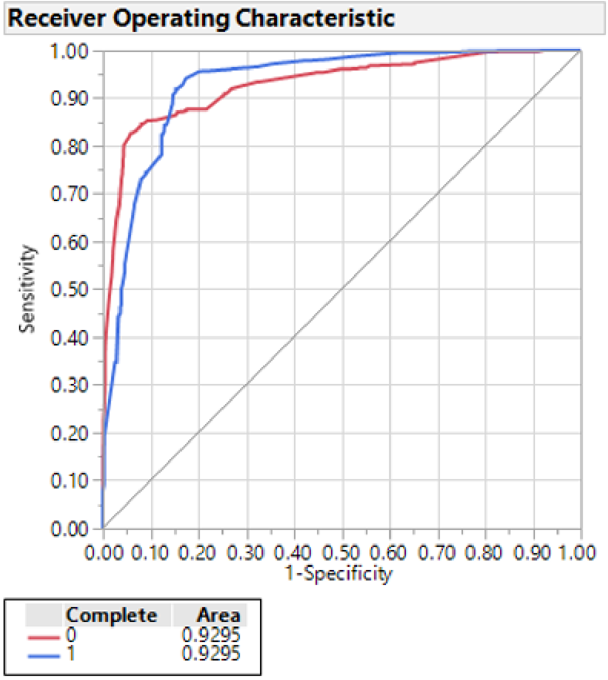

Figure 27. Post-Acceptance Model ROC Curve - Boosted Tree

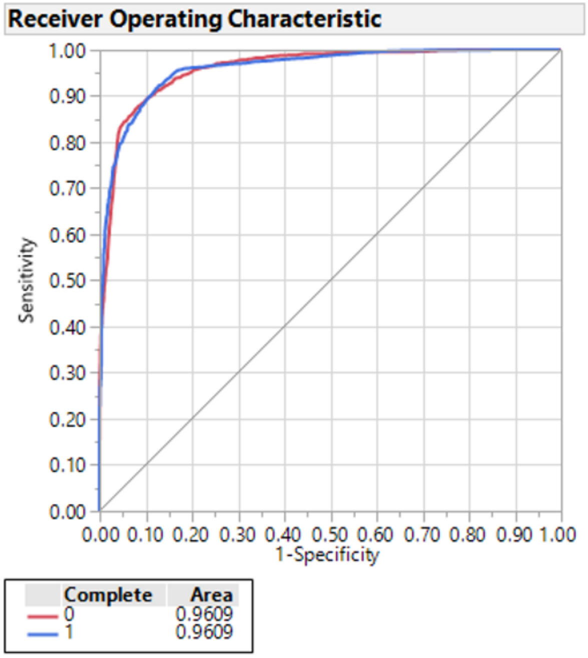

Figure 28. Post-Acceptance Model ROC Curve - Neural Nets

# Appendix E. Distinguished Graduate Analysis ROC Curves

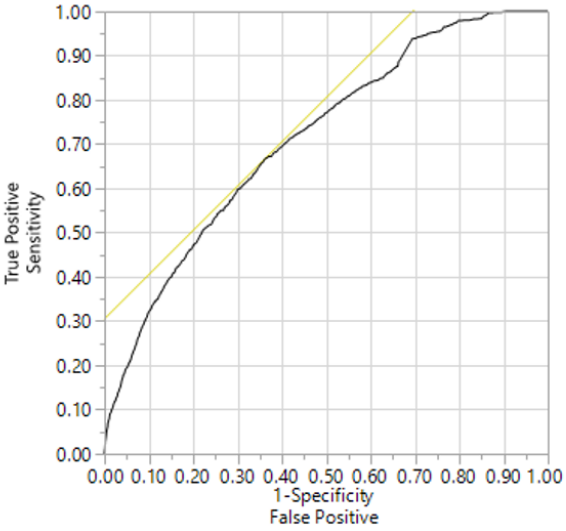

Figure 29. Pre-Acceptance DG Model ROC Curve - Logistic Regression

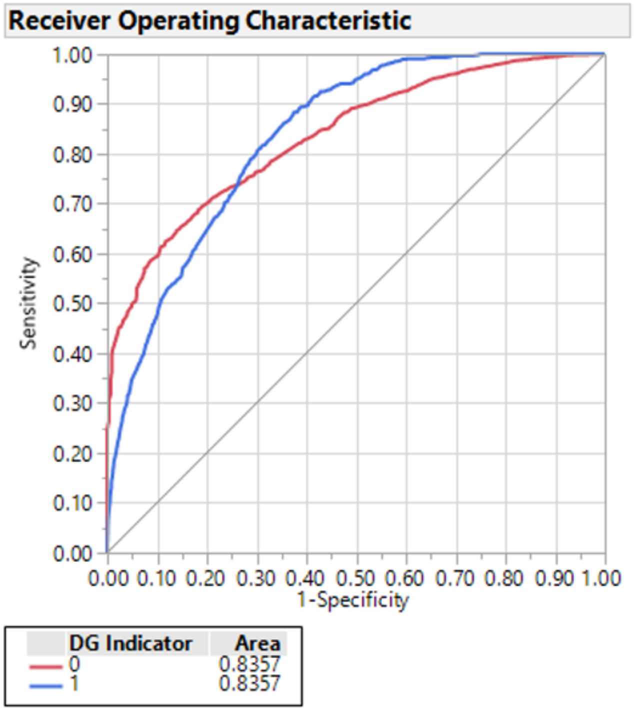

Figure 30. Pre-Acceptance DG Model ROC Curve - Bootstrap Forest

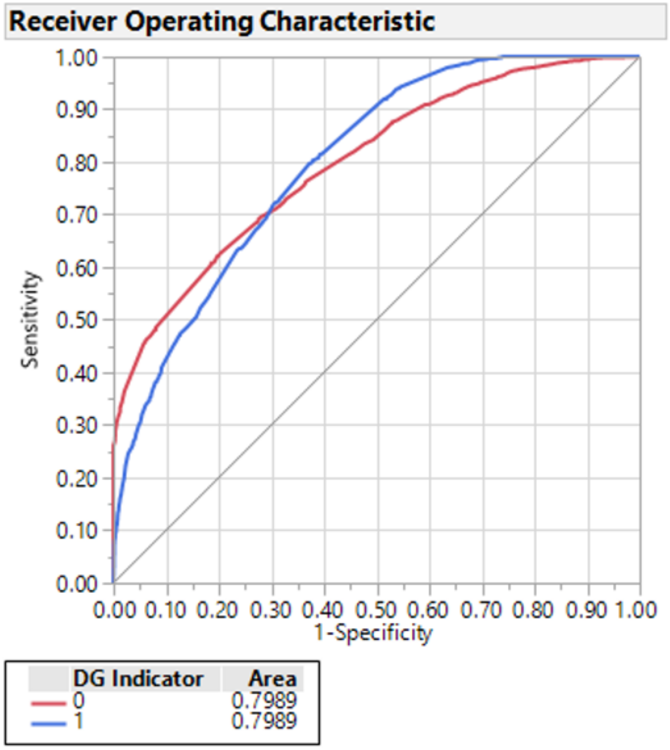

**Figure 31. Pre-Acceptance DG Model ROC Curve - Boosted Tree**

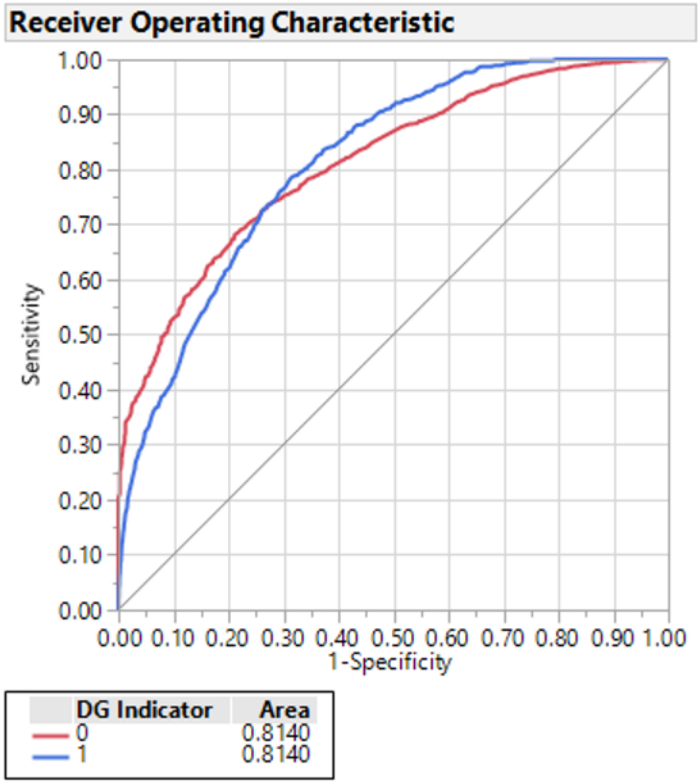

**Figure 32. Pre-Acceptance DG Model ROC Curve - Neural Nets**

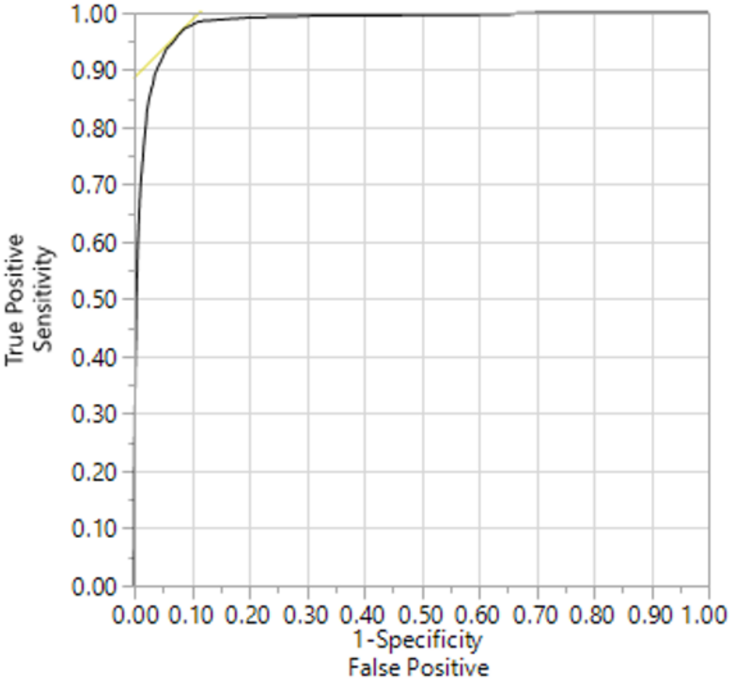

**Figure 33. Post-Acceptance DG Model ROC Curve - Logistic Regression**

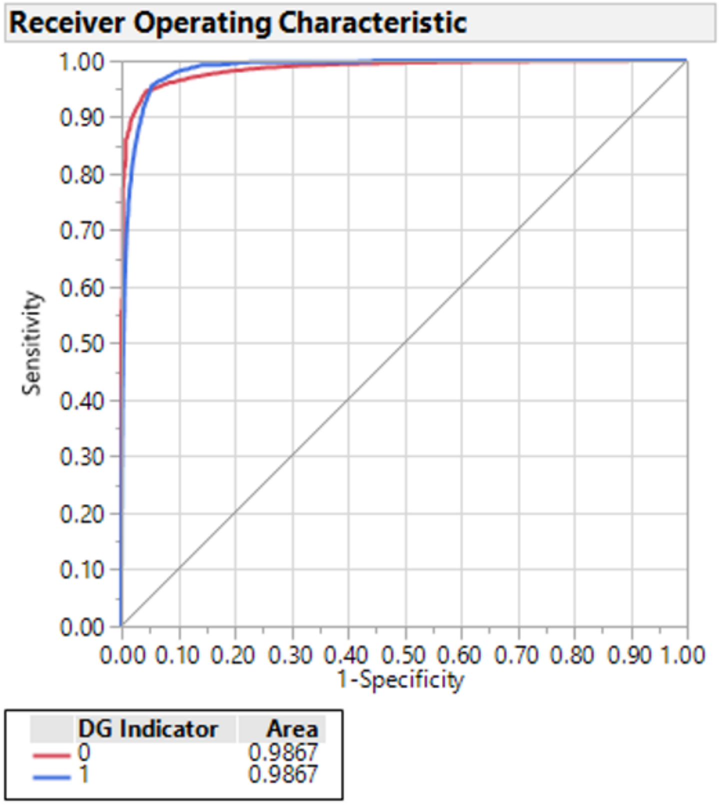

**Figure 34. Post-Acceptance DG Model ROC Curve - Bootstrap Forest**

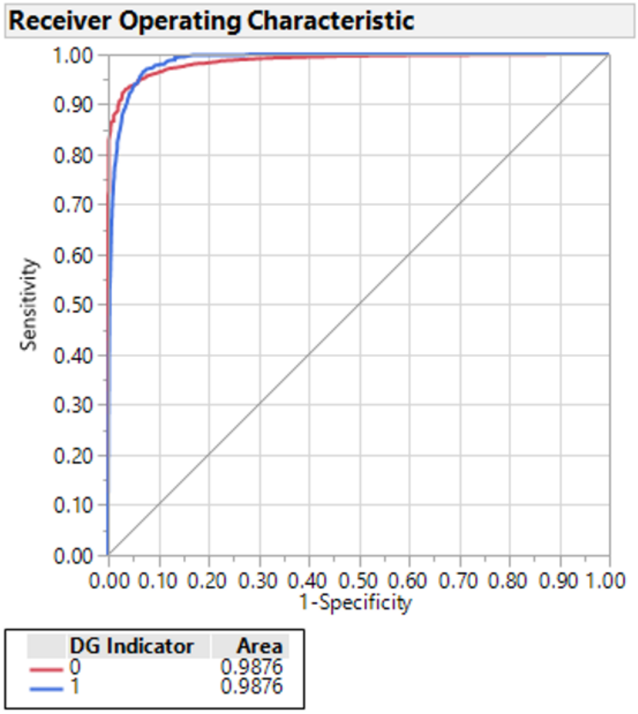

**Figure 35. Post-Acceptance DG Model ROC Curve - Boosted Tree**

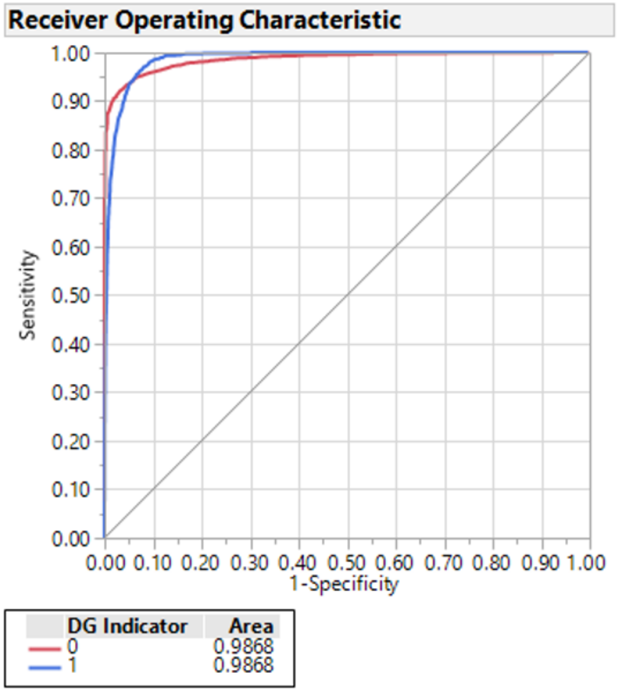

**Figure 36. Post-Acceptance DG Model ROC Curve - Neural Nets**

# Bibliography

1. US Government Accountability Office, "DoD Needs to Reevaluate Fighter Pilot Workforce Requirements," Tech. Rep., Washington D.C., USA, 2018.

2. Keisha A. Meyer, "Does Age, Gender, or Race Affect Undergraduate Pilot Training Attrition Rates and Composite Scores?," M.S. thesis, Air Force Institute of Technology, Wright-Patterson AFB, OH, 2019.

3. D.R. Hunter and E.F. Burke, "Predicting Aircraft Pilot-Training Success: A Meta-Analysis of Published Research," *The International Journal of Aviation Psychology*, vol. 4, no. 4, pp. 297–313, 1994.

4. E. Torrance, C.J. Rush, H. Koh, and J. Doughty, "Factors in Fighter-Interceptor Pilot Combat Effectiveness," Tech. Rep., Air Force Personnel and Training Research Center, Lackland Air Force Base, TX, USA, 1957.

5. T.R. Carretta and M.J. Ree, "Air Force Officer Qualifying Test Validity for Predicting Pilot Training Performance," *Journal of Business and Psychology*, vol. 9, no. 4, pp. 379–388, 1995.

6. T.R. Carretta, M.S. Teachout, M.J. Ree, E.L. Barto, R.E. King, and C.F. Michaels, "Consistency of Relations of Cognitive Ability and Personality Traits to Pilot Performance," *The International Journal of Aviation Psychology*, vol. 24, no. 4, pp. 247–264, 2014.

7. L.G. Barron, T.R. Carretta, and M.R. Rose, "Aptitude and Trait Predictors of Manned and Unmanned Aircraft Pilot Job Performance," *Military Psychology*, vol. 28, no. 2, pp. 65–77, 2016.

8. AFPC/DSYX, "Pilot Candidate Selection Method (PCSM)," Tech. Rep., Air Force Personnel Center, San Antonio, TX, USA, 2018.

9. Erland A. I. Svensson and Glenn F. Wilson, "Psychological and Psychophysiological Models of Pilot Performance for Systems Development and Mission Evaluation," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 95 – 110, 2002.

10. P.D. Retzlaff and M. Gibertini, "Air Force Pilot Personality: Hard Data on the "Right Stuff"," *Multivariate Behavioral Research*, vol. 22, no. 4, pp. 383–399, 1987.

11. AETC, *AETC Instruction 36-2205, Volume 4*.

12. Thomas R Carretta, "Predictive Validity of the Air Force Officer Qualifying Test for USAF Air Battle Manager Training Performance Human Effectiveness Directorate Supervisory Control Interfaces Branch," Tech. Rep., Wright-Patterson AFB, OH, 2009.

13. T. R Carretta, “Evaluation of Adverse Impact for US Air Force Officer and Aircrew Selection Tests,” Tech. Rep., Wright-Patterson AFB, OH, 2006.

14. Joseph L. Weeks and Warren E. Zelenski, “UNITED STATES AIR FORCE ENTRY TO USAF UNDERGRADUATE FLYING TRAINING,” Tech. Rep., Brooks AFB, TX, 1998.

15. Thomas R Carretta, “Development and Validation of the Test of Basic Aviation Skills (TBAS),” Tech. Rep., Wright-Patterson AFB, OH, 2005.

16. T.R. Carretta and M.J. Ree, “Pilot Candidate Selection Method (PCSM): What Makes It Work?,” Tech. Rep., Brooks AFB, TX, 1993.

17. Wendy Darr, “Military Personality Research: A Meta-Analysis of the Self Description Inventory,” *Military Psychology*, vol. 23, no. 3, pp. 272–296, 2011.

18. Frederick M. Siem, “Predictive Validity of an Automated Personality Inventory for Air Force Pilot Selection,” *The International Journal of Aviation Psychology*, vol. 2, no. 4, pp. 261–270, 1992.

19. G.F.V. Glonek and P. McCullagh, “Multivariate Logistic Models,” *Journal of the Royal Statistical Society*, vol. 57, no. 3, pp. 533–546, 1994.

20. C. E. Brodley and Mark A. Friedl, “Decision Tree Classification of Land Cover from Remotely Sensed Data,” *Remote Sensing of Environment*, vol. 61, no. 3, pp. 399–409, 1997.

21. Brett Lantz, *Machine Learning with R*, Packt Publishing, Birmingham, UK, second edition, 2015.

22. I. Rish, “An Empirical Study of the Naive Bayes Classifier,” in *IJCAI Workshop on Empirical Methods in Artificial Intelligence.* Georgia Tech, 2001, pp. 41–46.

23. Dan Claudiu Ciresan, “Deep, Big, Simple Neural Nets for Handwritten Digit Recognition,” *Neural Computation*, vol. 22, pp. 3207–3220, 2010.

24. Paul Horton, “Better Prediction of Protein Cellular Localization Sites with the K-Nearest Neighbors Classifier,” in *International Conference on Intelligent Systems for Molecular Biology.* 1997, pp. 147–152, American Association for Artificial Intelligence.

| | |
|---|---|
| **REPORT DOCUMENTATION PAGE** | *Form Approved*<br>*OMB No. 0704–0188* |

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704–0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

| | | |
|---|---|---|
| **1. REPORT DATE** *(DD–MM–YYYY)*<br>26-03-2020 | **2. REPORT TYPE**<br>Master's Thesis | **3. DATES COVERED** *(From — To)*<br>Sept 2018 — Mar 2020 |

| | |
|---|---|
| **4. TITLE AND SUBTITLE**<br>PREDICTING PILOT SUCCESS USING MACHINE LEARNING | **5a. CONTRACT NUMBER** |
| | **5b. GRANT NUMBER** |
| | **5c. PROGRAM ELEMENT NUMBER** |
| **6. AUTHOR(S)**<br>Aaron C. Giddings | **5d. PROJECT NUMBER** |
| | **5e. TASK NUMBER** |
| | **5f. WORK UNIT NUMBER** |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**<br>Air Force Institute of Technology<br>Graduate School of Engineering and Management (AFIT/EN)<br>2950 Hobson Way<br>WPAFB OH 45433-7765 | **8. PERFORMING ORGANIZATION REPORT NUMBER**<br>AFIT-ENS-MS-20-M-150 |
| **9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**<br>Air Force Air Education and Training<br>Commander, Detachment 21<br>Randolph AFB TX 78150<br>Email: jason.colborn@us.af.mil | **10. SPONSOR/MONITOR'S ACRONYM(S)**<br>AETC/Det 21 |
| | **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**

The United States Air Force has a pilot shortage. Unfortunately, training an Air Force pilot requires significant time and resources. Thus, diligence and expediency are critical in selecting those pilot candidates with a strong possibility of success. This research applies multivariate and statistical machine learning techniques to pilot candidates pre-qualification test data and undergraduate pilot training results to determine whether there are selected pre-qualification tests or specific training evaluations that do a "best" job of screening for successful pilot training candidates and distinguished graduates. Flight experience, both during training and otherwise, indicates pilot training completion and performance.

**15. SUBJECT TERMS**

Pilot training, multivariate analysis

| **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 | U | 90 | **19b. TELEPHONE NUMBER** *(include area code)*<br>(937) 255-3636, x7469; rhill@afit.edu |