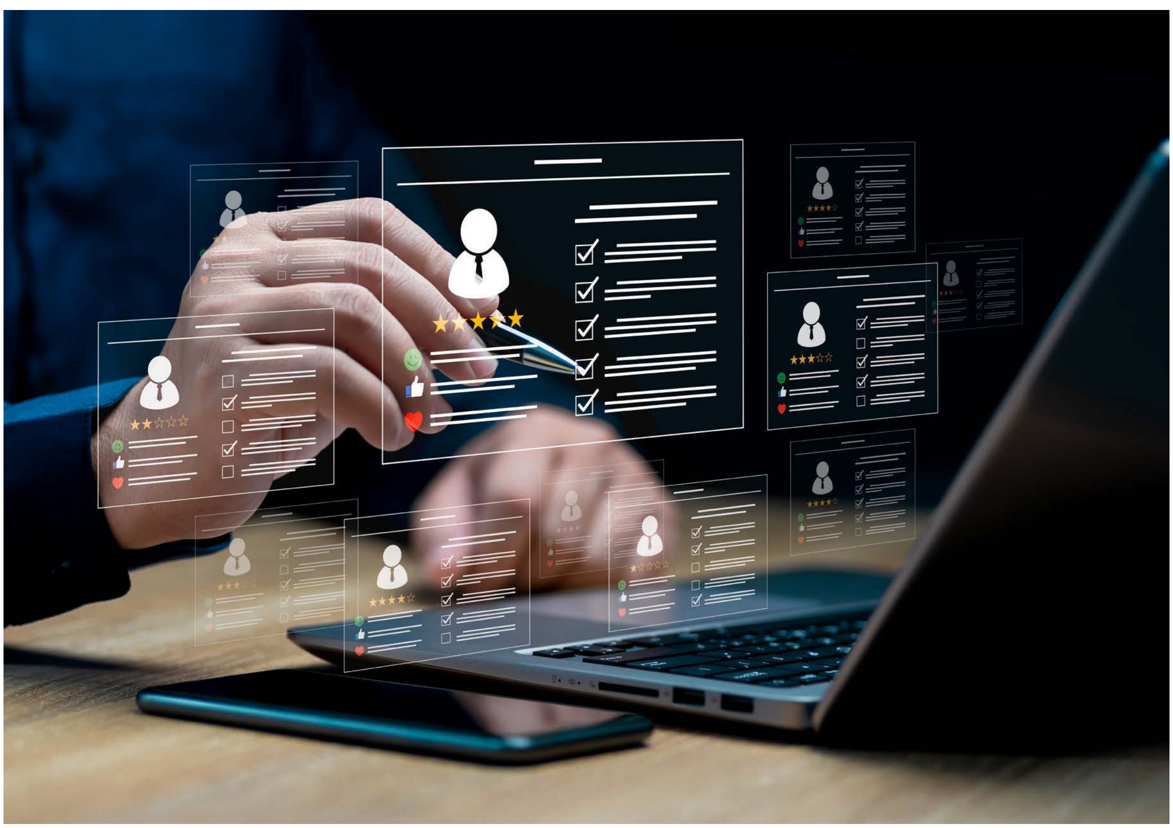
Research Report

KELLY HYDE, NELSON LIM, SARAH NICOLE KOSIC, ANTHONY LAWRENCE, ALICE NGUYEN, STEPHANIE WILLIAMSON, PAUL EMSLIE

Improving the Personnel Records Scoring System for Air Force Active-Duty Officers and Enlisted Personnel



For more information on this publication, visit **www.rand.org/t/RRA2996-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.

Library of Congress Cataloging-in-Publication Data is available for this publication.

ISBN: 978-1-9774-1570-9

Cover: NongAsimo/Adobe Stock

About This Report

The U.S. Air Force (USAF) continually invests in technology to improve its talent management practices. The USAF requested that RAND Project AIR FORCE enhance the talent management technology for active-duty officers and extend its application to include enlisted and civilian personnel. In this report, we describe the approach, methodology, data, and application of the technology, which is a suite of predictive models designed to assess whether personnel have fulfilled expectations for career progression.

The research reported here was commissioned by the Directorate of Plans and Integration, Deputy Chief of Staff for Manpower, Personnel, and Services, and conducted within the Workforce, Development, and Health Program of RAND Project AIR FORCE as part of a fiscal year 2024 project, "Operationalization of a Technology-Enabled Talent Management for USAF Active-Duty and Civilian Personnel."

RAND Project AIR FORCE

RAND Project AIR FORCE (PAF), a division of RAND, is the Department of the Air Force's (DAF's) federally funded research and development center for studies and analyses, supporting the United States Air Force and the United States Space Force. PAF provides the DAF with independent analyses of policy alternatives affecting the development, employment, combat readiness, and support of current and future air, space, and cyber forces. Research is conducted in four programs: Strategy and Doctrine; Force Modernization and Employment; Resource Management; and Workforce, Development, and Health. The research reported here was prepared under contract FA7014-22-D-0001.

Additional information about PAF is available on our website:

www.rand.org/paf/

This report documents work originally shared with the DAF on March 26, 2025. The draft report, dated April 2025, was reviewed by formal peer reviewers and DAF subject-matter experts.

Acknowledgments

We are grateful to Gregory D. Parsons (Deputy Chief of Staff for Manpower, Personnel, and Services, Headquarters USAF [AF/A1X]), for his support throughout the project. We also thank our action officer, Douglas Boerman (chief, Human Resources, Analytics, and Decision Support Division, Plans and Integration, Deputy Chief of Staff for Manpower, Personnel, and Services, Headquarters USAF [AF/A1XD]), for his support and guidance. We also appreciate the feedback and support of Maj Garret Moore, Lang Born, Cara Albright, Eric Primozic, John Sanzone, Lt Col Terrell Tillery, Alex Rosenfeld, and Pedro Riscado in AF/A1XD, whose collective efforts to fulfill our data requests, provide regular feedback on our planned approach, help us address challenges in accessing needed resources, and oversee our Envision workspace were invaluable and indispensable to the project. We also thank Shawna Teague (chief data officer, Headquarters, AF/A1X) for helping us establish the new framework for data-sharing between AF/A1X and RAND on Envision, which allowed us to access and use critical data for this project.

We extend our gratitude to the USAF talent management stakeholders for officer, enlisted, and civilian personnel who spoke with us during this project, including career field managers, development team leaders, headquarters-level force development and personnel policy leaders, functional experts, and human resource data analysts. Their feedback on the nuances of talent management across career fields, key data sources to inform talent management processes, and needs for technology-enabled capabilities was essential to shaping our approach.

This project also received significant contributions from several of our RAND colleagues. We extend a special thanks to Ray Conley, who provided dedicated and thoughtful mentorship and intellectual leadership from the beginning of the project through his retirement in December 2024. We also thank Jennifer Nevius and Daniel Sitterly for sharing their tremendous expertise in DAF personnel policy and talent management processes throughout the project, helping us determine the most critical inputs for the new Personnel Records Scoring System (PReSS), and providing us with a deep contextual understanding of the decision processes it would model. Finally, we are grateful for the contributions of Melissa Bauman, who thoughtfully refined our project briefings; Ojashwi Pathak, who provided useful and innovative methodological ideas as a summer associate; Miriam Matthews and Sean Robson, who provided project support during each of their respective tenures as associate director of the Workforce, Development, and Health Program of Project AIR FORCE; Kirsten Keller and Rob Romer, who led our efforts to explore the feasibility of a future PReSS adaptation or related technology for DAF civilian talent management in a companion report; and David Schulker and Matt Walsh, whose leadership of the original PReSS prototype development set the foundation for this project.

Finally, we thank Maria Lytell for her thoughtful feedback and suggestions throughout the project as our dedicated reviewer.

Summary

Issue

U.S. Air Force (USAF) talent management involves regular large-scale and labor-intensive processes to evaluate personnel. To improve the efficiency of these and other processes relying on human resource data, the USAF has invested substantially in data infrastructure, including the cloud-based analytic platform Envision, which is connected to talent management data sources of record, including the Automated Record Management System—Legacy Conversion (ARMS-LC) and Military Personnel Data System (MilPDS). The availability of these data in an integrated analytic platform introduces the potential for new or updated capabilities with cross-dataset linkages and near–real-time data updates previously difficult to achieve at scale. Prior to USAF-wide adoption of Envision, the Directorate of Plans and Integration, Air Force Directorate of Manpower, Personnel, and Services (AF/A1X) asked RAND Project AIR FORCE to develop a prototype tool, the Personnel Records Scoring System (PReSS), to model career milestones of field grade officers using information from narrative performance reports. For this project, AF/A1X asked RAND Project AIR FORCE to improve and expand on PReSS within Envision, with the goal of developing an operationalizable tool to support active-duty officer and enlisted and civilian talent management throughout the enterprise.

Approach

Using the original prototype as a foundation, we redesigned PReSS to achieve higher accuracy based on more-stringent metrics, adapt more readily to potential changes in data over time, and provide deeper insights into the reasons behind the model's predictions. Our approach began with development on the RAND internal computing resources that hosted the original prototype. In parallel, we collaborated closely with the Human Resources, Analytics, and Decision Support Division, Plans and Integration, Air Force Directorate of Manpower, Personnel, and Services to establish new pipelines for the data required to implement PReSS on Envision, including personnel records from MilPDS and performance report records from ARMS-LC for field grade officers and senior enlisted airmen. Once these data became available, we ported our internal efforts over to Envision, adapting PReSS as required to achieve full compatibility with the platform and navigate emerging limitations. These efforts culminated in a suite of models that measures the extent to which field grade officers and senior enlisted personnel have fulfilled the expectations for their next career development milestones based on their current records.

Although we had intended to develop an analogous tool for the USAF civilian workforce at project outset, discussions with key stakeholders in civilian talent management and our review of available data led to the conclusion that a civilian version of PReSS would lack critical information to

accurately model career development and would require a different approach to meet civilian talent management stakeholders' needs and fit into existing processes. For these reasons, this report focuses on PReSS for field grade officers and enlisted personnel, and our findings related to the civilian workforce are presented in a companion report (Romer, Keller, and Hyde, forthcoming).

Key Findings and Recommendations

Table S.1 summarizes key findings and recommendations derived from developing and testing the new PReSS for field grade officers and enlisted personnel.

Table S.1. Key Insights and Recommendations from the New PReSS for Officer and Enlisted Career Development

Insight	Recommendation
Given enough information, an effective method, and sufficient computing resources, it is possible to quantify whether senior enlisted airmen and field grade officers have fulfilled the expectations for career development milestones, such as promotion, with high accuracy.	Operationalize PReSS for field grade officers and senior enlisted personnel and explore possible extensions of the tool using its outputs as a foundation for related capabilities, such as providing automated feedback to personnel.
The features that correlate with promotion over the course of an airman's career include some that are always relevant and others that become more or less relevant with increased seniority.	Explore ways to analyze and use PReSS outputs to improve understanding and transparency of the key contributors to career outcomes throughout the course of an airman's career, for both talent management stakeholders and personnel themselves.
Although it is possible to predict career outcomes from just MilPDS quantitative features or just performance report qualitative features alone, the most performant models use a combination of both.	When operationalizing PReSS, emphasize use and maintenance of the combined qualitative and quantitative Model 3 and ensure that both types of data remain up to date through robust and regular extraction pipelines.
The USAF already has the key information required for such a capability as PReSS to be operationalized on Envision, but some additional data that could further improve the models or expand PReSS's use cases are not yet usable.	Pursue opportunities to expand the scope of data available to PReSS, such as leveraging existing data connections that have not yet been used (e.g., MyVector) and collaborating with other Department of the Air Force functions to establish new data connections where possible and relevant.

We note that, when porting the new PReSS from RAND's internal server to Envision, we adjusted some features of the tool to overcome limitations of Envision. AF/A1X and other USAF talent management stakeholders should consider these limitations and their implications when using the tool. Additionally, with cutting-edge technologies, such as large language models (LLMs), becoming more broadly available both within Envision and elsewhere, there will be opportunities to build further capabilities into PReSS to meet additional talent management needs without needing to redevelop the tool. In the future, LLM technology could stack on to PReSS by using its outputs to "learn" what matters for career development, providing for more broadly accessible and user-friendly outputs directly to stakeholders and personnel.

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

In recent years, RAND Project AIR FORCE (PAF) has supported the U.S. Air Force's (USAF's) efforts to enhance its talent management practices through technology investments (Schulker et al., 2022; Snyder, 2022; Yeung et al., 2022). These projects demonstrate that leveraging emerging technology can transform the USAF's talent management system to a system in which decisions are frequently refined to meet goals using a wide variety of information, including both quantitative and qualitative data. Beyond conceptual work, PAF developed a prototype tool, Personnel Records Scoring System (PReSS), to assist development teams (DTs) in managing active-duty officers and tested the tool with a select group of DTs. In this report, we describe the work to improve PReSS for active-duty officers and expand the application of the technology to enlisted personnel. Because we discovered irreconcilable data limitations in the development phase of the project, we did not extend PReSS to the civilian workforce; instead, we conducted a review of possibilities, data needs, and exemplar efforts in related domains outside the USAF for an analogous civilian tool. The results of this review will be provided in a companion report (Romer, Keller, and Hyde, forthcoming); our report focuses on the tool's application for active-duty officers and enlisted personnel.

Personnel Records Scoring System

PAF developed PReSS to address a common challenge in USAF human resource management: the difficulty of efficiently and objectively reviewing large volumes of narrative performance information for talent management decisions. Human reviewers face time, energy, and expertise limitations when evaluating officer records for promotion and developmental education opportunities.

PReSS aims to improve the efficiency and objectivity of the nomination process for developmental education by identifying promising candidates and providing an objective benchmark for comparing the decisions of raters and developmental team members. The system is a decision support tool that helps human resource personnel extract meaningful insights from unstructured text in officer performance reports (OPRs).[1]

PReSS is the culmination of two PAF research initiatives: an exploratory study on artificial intelligence in Air Force talent management (Schulker, Lim, et al., 2021) and an implementation of the methodology (Schulker et al., 2024).

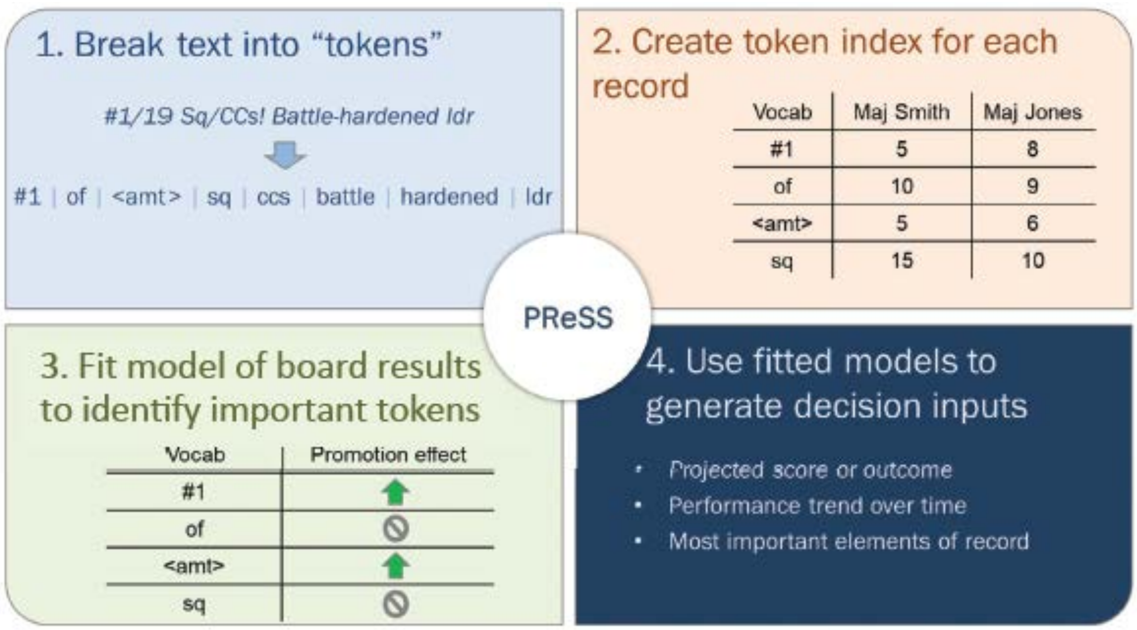
---

[1] Since the development of the PReSS prototype, which used evaluation data from 2012 through 2021, OPRs have been replaced by new-format officer performance briefs (OPBs). In an unpublished 2024 RAND report, researchers Devon Hill, Kirsten M. Keller, Lewis Schneider, Jennifer Nevius, and Zhan Okuda-Lim examined the implications of this transition and a related replacement of enlisted performance reports (EPRs) with enlisted performance briefs (EPBs).

Methodology

PReSS employs a comprehensive four-stage approach to analyze performance narratives (see Figure 1.1). The process begins with text preprocessing (step 1 in the figure), in which the system converts performance text into a standardized format; removes officer names; and normalizes common elements, such as stratifications, to ensure consistency. During this stage, the text is broken down into distinct *tokens*—individual words and phrases that serve as the basic units of analysis.

Figure 1.1. Overview of PReSS Development Steps



SOURCE: Reproduced from Schulker et al., 2024, Figure 1.

Following preprocessing (step 1), step 2 involves feature extraction to transform the tokenized text into measurable data points. The system creates a term-frequency matrix that counts the frequency of each word in a record. Then it applies inverse document frequency (IDF) weighting to prioritize distinctive terms over commonly used ones.[2] To capture meaningful expressions, such as stratifications, PReSS considers multitoken phrases of up to three words in length. The system also normalizes for document length to enable fair comparisons between records with different amounts of text.

Step 3 is model training, which represents the analytical core of PReSS by using supervised machine learning with regularized logistic regression techniques. These models are trained on

---

[2] IDF weighting of a term-frequency (TF) matrix, commonly called *TF-IDF*, divides the term frequencies within each document by a measure of how frequently each term appears across all documents in the sample. This operation results in increased proportional weight placed on terms that appear less frequently, and the most-common terms (e.g., conjunctions; descriptors that apply across documents within a particular group, such as pay grade) receive less weight. TF-IDF is an example of a *stateful* transformation, which depends on the sample on which it is calculated. As discussed later in the report, we designed the new PReSS so it would be less data dependent and less likely to require redevelopment as data and circumstances change; in accordance with this, we did not implement TF-IDF.

historical selection board outcomes to identify which terms and phrases correlate with high performance evaluations. The system's accuracy is ensured through validation methods, including cross-validation and holdout test sets, which confirm that the models can generalize effectively to new records.

In the final stage, prediction generation (step 4 in Figure 1.1), PReSS produces actionable insights from the analyzed data. For promotion potential, the system creates scores on a 0–1 scale indicating the likelihood of selection. For developmental education boards, it produces scores on a 6–10 scale reflecting panel assessments. Beyond these overall metrics, PReSS calculates impact scores for individual OPRs and specific text snippets, enabling the identification of which elements of an officer's record most significantly influence their evaluation. These detailed analytics culminate in comprehensive summaries that clearly show a record's strengths and weaknesses and provide valuable decision support for talent management professionals.

PReSS employs logistic regression rather than a more-complex neural network model because, during model development and testing, the PAF team found that the simpler, more-explainable logistic regression performed better on the accuracy metrics employed. The system achieved 82 percent raw accuracy predicting O-5 promotions and 90 percent raw accuracy for O-6 promotions.[3]

Implementation and Applications

PReSS is designed to complement, not substitute for, human decision processes. It helps identify promising candidates and establishes an objective benchmark for comparing board decisions while leaving final judgments to human experts who can consider contextual factors beyond the scope of algorithmic assessment. This human-centered approach ensures that technology enhances rather than replaces professional judgment. The Air Force Personnel Center and Plans and Integration, Deputy Chief of Staff for Manpower, Personnel, and Services (AF/A1X) are engaged in ongoing work to refine, expand, and operationalize PReSS throughout the service.

Approach

We applied a mixed-methods approach to accomplish the project's objective. As mentioned previously, our current efforts focus on updating models with new training data to maintain accuracy as evaluation practices evolve, expanding PReSS's application to enlisted and civilian personnel, and integrating PReSS seamlessly into operational talent management workflows.

---

[3] *Raw accuracy* refers to the accuracy measure obtained by dividing the number of correct predictions, including both positive and negative, by the total number of predictions (i.e., number of observations in the data). In this report, to describe the updated models, we use *balanced accuracy*, which is a more-robust metric when using data with imbalanced classes. In our case, the extent of class imbalance in the data is determined by the percentage of airmen meeting a particular board who are selected for promotion. Because this percentage is significantly lower than 50 percent in the pay grades in scope for this project, the data on which we trained PReSS overrepresent cases in which an airman was not selected for promotion. This could result in raw accuracy being skewed upward, because an empty model that always predicted 0 (not selected for promotion) would be accurate more than 50 percent of the time. Balanced accuracy corrects for this by taking the average of two accuracy metrics, one for positive cases (i.e., how many airmen who were predicted to be selected for promotion actually were) and the other for negative cases (i.e., how many airmen who were not predicted to be selected for promotion actually were not). In other words, balanced accuracy represents what the raw accuracy *would have been* if the classes had been balanced.

To improve PReSS for officers, we conducted a comprehensive review of the prototype tool by engaging in discussions with DT members and assessing available information. In this initial step, we aimed to gather insights and identify areas for improvement. Subsequently, we expanded the tool with quantitative information based on the findings from the initial step. This expansion enhanced the tool's capabilities and effectiveness. Next, we calibrated the tool using the newly enriched data to ensure its accuracy and relevance.

To develop a prototype tool for enlisted personnel, we conducted a thorough review to identify the essential information that should be included. This review involved discussions with DT members and an assessment of available data. Afterward, we processed and assembled the information gathered during the initial task. Subsequently, we calibrated the tool using the assembled data to ensure accuracy and effectiveness. Building on this, we developed a prototype tool for enlisted personnel incorporating the refined information.

We thoroughly reviewed the literature and conducted structured interviews with subject-matter experts from several agencies to explore ways to apply the methodology for civilian personnel. We also reviewed existing data sources to assess their usefulness in developing a prototype tool for civilian personnel. These findings are summarized in the companion report (Romer, Keller, and Hyde, forthcoming).

Structure of This Report

In Chapter 2, we discuss the new PReSS tool for active-duty field grade officers, highlighting key distinctions with the original PReSS prototype. We also present results from the new PReSS, which shed light on the determinants of officer career progression and promotion eligibility. In Chapter 3, we present the new PReSS tool for senior enlisted personnel, which focuses on progression to senior noncommissioned officer pay grades (E-7, E-8, and E-9). Although the foundation of the enlisted PReSS is mostly identical to the officer PReSS, we also describe differences required to accommodate differing data availability and force structure. In Chapter 4, we conclude by summarizing the key takeaways from PReSS development, limitations, and recommendations for operationalizing and maintaining the new PReSS tools moving forward.

Chapter 2

Improving PReSS for Active-Duty Officers

We improved PReSS by refining the original tool's qualitative OPR model and introducing additional quantitative features from an officer's Military Personnel Data System (MilPDS) record. This pipeline created three model categories: Model 1, which uses only performance report data; Model 2, which uses only MilPDS quantitative data; and Model 3, which combines performance report qualitative data and MilPDS quantitative data into a single model.

We explored strategies for combining these features and evaluated each model to identify a final model with optimal performance. Using this final model, we can derive a career development score analogous to the score produced in the original PReSS tool. We can then use various metrics, such as feature importance and feature impact, to interpret the portions of an officer's record that influence their career development score. The improved PReSS tool was implemented on Envision for ease of use by AF/A1X and DTs.

Enhancing the Performance Report Model

We redeveloped every stage of the qualitative performance report model pipeline: text data preprocessing, model input data construction, accuracy and performance measurement, and model output interpretation. We describe each of these components in sequence in the following sections.

Processing the Raw Text Extracted from Performance Reports

To secure the data required for the performance report model, we supported an effort led by Human Resources, Analytics, and Decision Support Division, Plans and Integration, Deputy Chief of Staff for Manpower, Personnel, and Services (AF/A1XD) to update the text extraction method initially developed by RAND researchers for the original PReSS prototype using a combination of methods to accommodate the different formats of OPR and OPBs over time. The text that is ultimately transformed into the input data for the performance report model (Model 1) is initially extracted from Automated Record Management System—Legacy Conversion (ARMS-LC) OPRs and OPBs. The format and file type of these documents has changed over time, introducing the need to standardize across multiple data formats and correct for errors that might be more likely to arise for specific document types. Specifically, the oldest OPRs are often stored as images, requiring an optical character recognition (OCR) method to extract the report text, although the text in newer PDF OPRs and OPBs can be extracted directly. However, even the newer PDFs can vary, with some containing tags (or *metadata*) that make it easier to break up the performance report into individual fields and others not.

However, challenges to using the text from OPRs and OPBs in a predictive model remained after these extraction updates, requiring us to implement a sequence of text preprocessing steps. Specifically, we addressed the following challenges:

- OPR and OPB text might contain the name of the officer being evaluated, which should be excluded from the model input data to ensure that model outputs do not include any personally identifiable information and that any coincidental relationship between names and career development outcomes does not limit model generalizability.
- The text extraction methods, especially for older OPRs stored as images, might sometimes mistake certain characters for other similar-looking characters. For example, the number 8 is sometimes misread as the letter B or vice versa.
- OPR and OPB text is often heavily abbreviated and might contain typos or spelling errors. Additionally, the abbreviations used by different raters are not always consistent, especially for shorthand words (writing "developed" as "dvlpd," "dev'd," etc.). These inconsistencies ultimately reduce the accuracy of the predictive model if they result in a single word being represented by multiple tokens in the input data when a human reader would recognize them as the same word. Although there is an area on the OPR or OPB form for raters to specify the meaning of abbreviations that they have used in their ratings, this field is not always used, does not always contain all relevant abbreviations, and does not follow a standard format across raters.

To address these challenges, we implemented a series of sequential transformations to the raw extracted text that improved readability and consistency while retaining the meaning of the original text. First, to standardize the format of abbreviation definitions, we used a version of GPT-4o mini, an OpenAI-developed large language model (LLM), through an internal RAND service approved for use with controlled unclassified information. We prompted the model to receive an arbitrary list of abbreviations and their definitions and respond with the same list written as "abbreviation 1: definition 1; abbreviation 2: definition 2; . . . ." This allowed us to extract abbreviation-definition pairs from the LLM's response by splitting the text at colons and semicolons; these results could then be used to define find-and-replace rules to replace instances of an abbreviation in an OPR or OPB with its definition. To address the incompleteness of some OPR or OPB abbreviation lists, we created a combined set of abbreviation-definition pairs across OPRs and OPBs in our data, taking the most common definition for an abbreviation in cases in which the same abbreviation was defined in different ways by different raters. The definitions in this combined set were then used as fallback options for each abbreviation in cases in which the abbreviation appeared in an OPR or OPB but was not defined by the rater. In total, the combined set of abbreviation-definition pairs comprised about 10,000 unique acronyms and shorthand words to be replaced in the OPR and OPB data.

Next, we applied spellchecking to the OPR and OPB data using a custom dictionary with select words added to avoid replacing common terms in the USAF context that were not included by default (e.g., "warfighter"). When spellchecking and replacing a nonmatching word with the suggestion generated by the spellchecker, we took steps to ensure that words in all capital letters (which could be either unreplaced abbreviations or titles frequently formatted in all caps, such as exercise or operation names), shorthand words already defined in our dictionary of abbreviation-definition pairs, and other common text features that should not be spellchecked (e.g., plane model names, numbers) were not replaced to avoid undesired changes to the text when applied automatically. Aside from correcting misspellings and typographical errors, a common benefit of the spellchecking approach was that it frequently detected cases in which two words that should have been separated by a space were combined, which is a common issue with OCR text-extraction methods.

Finally, we expanded the dictionary of abbreviation-definition pairs through a semiautomated process that scanned through the OPR or OPB text word by word, checked each word against the spellchecking dictionary and the list of known replacements, and, in cases in which the word did not pass the spellcheck and did not already have a known replacement, prompted the user to enter text with which the word should be replaced or indicate that the text should not be replaced. We applied this computer-assisted, user-driven process to about 10,000 OPRs and OPBs in our data sample, resulting in the addition of about 20,000 find-and-replace rules to our existing list derived from rater-provided abbreviations. These rules addressed cases in which either an abbreviation was never defined by any rater in our data sample or a misspelling or typo was detected by the spellchecker but did not have an automatically suggested replacement.

We developed these find-and-replace rulesets for OPR and OPB text data on RAND's internal server using the OPR data from the original PReSS prototype, which comprised about 50,000 OPRs from O-5 and O-6 promotion boards between 2012 and 2021. When implementing the qualitative model on Envision, we transferred the ruleset over to Envision as a data frame and applied it to the OPR and OPB data that we received there, which included newer OPRs and OPBs from 2022 through 2024. Because the development pipeline involved the use of internal RAND LLM services and a script run in an interactive session in which the user could be prompted for input, the exact process by which this ruleset was built cannot be repeated on Envision. In Chapter 4, we discuss how a similar process could be done entirely within Envision. However, because acronyms and other abbreviations are not likely to change dramatically over time (especially those that are most directly relevant to career development outcomes), the prebuilt set of about 30,000 find-and-replace rules should remain relevant and require only small, case-by-case additions or modifications, which can be performed directly in Envision.

Transforming the Processed Text into Model Input Data

Alongside our improvements to the readability and consistency of the OPR and OPB text, we improved the method by which the text is transformed into the dataset used to model career development outcomes. The original PReSS prototype, as described in Chapter 1, transformed the text into counts of term frequencies, resulting in a table with one column for each token that appears in at least one OPR. A major limitation of this approach is that, after training a predictive model using these data, the vocabulary of any data used in the future—i.e., a new batch of OPRs and OPBs—must be the same, because the columns must match the original training data. This means that, unless the exact same words are used at least once (and no other word is used) in the new data, columns will have to be manually added to or removed from the table of new data to match the original model training data.

To avoid this issue, we used an alternative way of transforming the text into a model training dataset that applies a hashing function to each token. This method is conceptually similar to cryptographic hashing functions used to obfuscate such sensitive information as passwords and personal identifiers, but, unlike these functions, it is not designed to make it difficult to reconstruct the original input from the hash output. The hashing function takes a string of text as input, converts it to binary, applies a sequence of binary operators, and finally converts the result of those operators to an integer. The hashing function also allows the user to specify the maximum integer (usually a power of two) allowable in the output, and, in cases in which the function integer output is greater than this specified maximum, it is replaced by the remainder of dividing it by the specified maximum (i.e., the *modulus*). Using this method, the number of columns desired in the transformed dataset can be prespecified, with a higher number of columns corresponding to a lower risk of *collisions* (i.e., the function outputting the same value for multiple distinct tokens, which can happen if one string maps to an integer greater than the specified maximum, to which the modulo operator is applied, and a different string maps directly to the modulus) and therefore a more-precise mapping. This means that the user can ensure that the transformed dataset always has the same number of columns for any text input without any manual transformations necessary. Although this is the most crucial advantage of the hashing approach for our application, another meaningful advantage is that the hashing approach is orders of magnitude faster and substantially more memory-efficient than the count-based approach used in the original PReSS. Memory efficiency is especially beneficial for the Envision implementation of the qualitative model for reasons discussed in Chapter 4. Finally, because the hashing function we use is reversible, we can recover the original input token from the hashed integer value.[4]

We also extended the model input data with additional features derived from OPRs and OPBs. Stratifications, which are considered one of the strongest predictors of officer career development outcomes, are more accurately represented in numerical form rather than treated as words. Including stratifications as a numerical feature allows the model to account for the fact that, for example, "#2 of 7" and "#4 of 14" represent the same fraction and, therefore, might have similar implications for career outcomes, and "#1 of 5" and "#1 of 20" might both start with "#1," but the latter places the officer above 19 of their peers, while the former places them above only four. If stratifications were treated as words separated by spaces, the model input data would suggest the exact opposite in each of these examples. Instead, we extract stratifications from the OPR and OPB text and convert them to a fractional value (i.e., "#1 of 20" = 1/20 = 0.05). In cases in which an officer received multiple stratifications in the same OPR or OPB, we use the smallest fractional value (i.e., the stratification that places the officer highest among their peers relative to the size of the ranked peer group).

---

[4] The key to this is that the random seed is fixed and known to the user, and the binary operators used in the hashing function are known and invertible. This differs from cryptographic hashing functions, which include several features intended to prevent mapping hashed values back to their original inputs. The hashing function we use, which is provided by both scikit-learn (HashingVectorizer) and PySpark (HashingTF), is trivial to invert because it is designed to be used in fast lookup tables where secrecy of the input is not an objective. While scikit-learn's implementation does not provide a built-in reversing function, another Python package (eli5) offers functions to invert its implementation. Meanwhile, the PySpark implementation that we use on Envision has a built-in method to associate input tokens with their hashed integer values, which is also the index of the column associated with that token in the vectorized output.

Alongside stratifications, we added indicator variables for the ranks of the rater, additional rater, and reviewer, respectively; the officer's duty Air Force Specialty Code (AFSC); and the officer's years in grade. We included these indicators primarily to allow the model to interact them with features from the OPR or OPB text, because the use of a specific word might have different implications depending on the officer's AFSC, how long they have been in the grade at which they are being rated, or the seniority of the rater using the word.

Finally, our approach to constructing model input data from OPR and OPB text differs from the original PReSS in that we looked at each OPR or OPB received by the officer separately and generated a prediction for each rather than combine the history of performance reports received by an officer up to a particular promotion board. This change was partly based on sponsor and stakeholder feedback, which indicated that, although promotion boards can review historical performance reports, it is not typical for them to review anything further than two or three prior reports, and older reports are given significantly less weight than more-recent ones. Additionally, making predictions for each OPR separately avoids issues arising from the history of OPRs and OPBs received in grade being necessarily longer for officers who have been in the grade longer. Varying lengths of input data pose a challenge for predictive modeling because, as the block of text being used gets longer, the number of times each word appears in the text likely increases, especially for words that are commonly used across performance reports. However, this repetition does not necessarily reflect anything about the officer's quality, confounding the relationship between text frequencies and career development outcomes. In the original PReSS prototype, this issue was addressed through a smoothing process, which standardized OPR length by adding placeholder text that resembled a typical OPR. However, as a side effect of this smoothing process, we observed that the original PReSS prototype's predictions were generally skewed toward the average, dampening the distinction between the "top" and "bottom" officers. This effect was especially true when generating predictions from single OPRs, which was the most common use case for talent management stakeholders interested in using the tool. Accordingly, we updated PReSS to generate predictions based on individual OPRs or OPBs so that the score specifically represents the performance report that the officer received for a particular promotion board's consideration.

Determining Whether and When Personnel Were Selected for Promotion

Unlike the original PReSS, which used promotion board files that contained both the performance reports for a particular board and the selections made by the board, our updated PReSS was built on MilPDS personnel records and the performance reports only. Although we were able to reconstruct most of the contents of the promotion board files from these two sources, the MilPDS data do not have a direct analog for the promotion selection indicator itself. However, they do contain sufficient information to infer selection for promotion. The MilPDS data we use for both officers and enlisted personnel contain a *projected pay grade* variable that updates soon after the selections made by a promotion board are released, with selected personnel's grades changing from "[E/O]-X" to "[E/O]-X+1 Select," respectively. This differs from the *actual pay grade* variable in MilPDS, which does not update until the personnel's promotion is pinned on and, therefore, might remain unchanged for several months after their selection for promotion becomes known. Instead, the *projected pay grade* variable typically updates within a couple of months after a promotion board, meaning that promotions inferred from observed changes in the projected pay grade are much closer in time to the actual selection for promotion than those inferred from changes in the actual pay grade. Accordingly, we constructed a promotion selection flag from the MilPDS data that equals 1 in the month immediately preceding an observed change in projected pay grade. For example, if we observe an officer whose projected pay grade was O-4 in May 2022 and O-5 Select in June 2022, the promotion selection flag is equal to 1 for this officer in May 2022, reflecting that, at some point prior to June 2022, the officer was selected for promotion by a board.

Because we cannot directly link performance reports to boards, we also cannot verify whether an officer or enlisted airman was considered eligible for promotion at the time they would have met the board with that performance report. Determining eligibility for promotion is further complicated by the fact that our data contain historical information from periods during which it was possible to be promoted below the promotion zone (BPZ). Accordingly, we do not explicitly restrict our data sample based on eligibility for promotion, and the model might assign predictions of promotion likelihood to years during which the associated airman was not eligible. This could inflate estimated accuracy if the model "learns" to predict eligibility and always assigns a low or zero likelihood to ineligible records. However, our selected model performance metrics (precision, recall, and balanced accuracy) are less likely to be affected; the inclusion of some observations for which promotion is impossible can only decrease precision because false positives would be penalized and true negatives do not enter the formula, would not directly affect recall because neither false positives nor true negatives enter the formula, and would not skew balanced accuracy to the same extent as raw accuracy because the weight that positive cases (and, accordingly, positive predictive values) receive in the balanced accuracy formula is fixed and not reduced by the addition of trivially negative cases. Additionally, we included time in grade as a predictor in both the qualitative and quantitative models to allow for interactions between time in grade and other predictors; this has the side benefit of limiting the extent to which the relative importance of proxies for time in grade above and below the promotion threshold could be inflated by ineligible records.

After inferring promotions from MilPDS data, one challenge remains: To merge the inferred promotion selection flag from MilPDS with the performance report data, it is necessary to also determine when the board that culminated in that selection actually took place. This is necessary because the date in our MilPDS records is the month in which the associated snapshot of the MilPDS data was taken, whereas the date in our performance report records is the date of the promotion board for which each performance report was prepared. Board meeting dates are recorded and typically known in advance, making it theoretically possible to obtain this board date directly rather than inferring it. However, we are not aware of a centralized, analysis-ready database of board meeting dates for officers or enlisted personnel. Instead, this information is usually embedded in board calendar documents from which the dates would need to be extracted, and the lead time with which these board calendar documents are made available is variable and uncertain. As a result, relying on actual board dates would introduce a significant amount of manual work required to maintain PReSS over time and also introduce a dependency on an additional information source that is not (to our knowledge, as of the time of this writing) tracked in Envision.

To avoid external information dependencies and enhance the longevity of the tool, we inferred board dates from the MilPDS data in a similar way to our inference of selection for promotion from the same data. This means that, to run PReSS again in the future, the only strict requirements are updated MilPDS data and updated performance report records. More specifically, for a given calendar year, we can calculate the proportion of observed changes in projected grade within an AFSC (based on the fact that, for officers, board dates differ by career field) that took place in each month. We set the *pseudo board date* for that AFSC, pay grade, and year equal to the earliest month during which at least 30 percent of the observed changes in a projected pay grade took place within that AFSC. For example, suppose that we observe 20 38F O-4s being selected for promotion in calendar year 2022 as implied by changes in projected pay grade. Of these 20 hypothetical promotions, we observe 15 taking place in May and the remaining five taking place at some other point during the year. In this case, our pseudo board date for 38Fs in 2022 would be the date of our MilPDS snapshot in May 2022. This means that, for 38F officers observed in MilPDS to be selected for promotion either on or after May 2022 but before the next pseudo board date, the most-recent performance report that they had received as of May 2022 is the one associated with their selection for promotion. For all other officers in the same AFSC and pay grade for whom we do not observe an increase in projected pay grade by May 2022, the performance report(s) that they received leading up to May 2022 were assigned a 0 for the selection flag (i.e., we infer that they were not selected at the board for which the performance report was prepared).

As a sanity check, we use the end of the reporting period ("report through date") from each performance report to ensure that we do not associate May 2022–derived selections with a performance report dated after May 2022. In cases in which such an assignment would happen, we go back to the previous performance report received by the officer and assign the May 2022–derived selection to this prior performance report until every selection is associated with a performance report for which the evaluation period ended before the change in projected pay grade took place. For example, if we observe an officer's projected pay grade in MilPDS increasing in May 2022 and the officer has two performance reports, the first evaluating their performance from January 2021 through December 2021 and the second evaluating them from January 2022 through December 2022, we associate this selection for promotion in May 2022 with the first performance report ending in December 2021. This is because they cannot have been selected for promotion based on the second performance report because its evaluation period had not yet ended as of May 2022. Instead, the report ending in December 2021 is the most-recent completed performance report they would have received by May 2022 and, therefore, it is the most likely report on which their selection was based.

Using the promotion board data received for the original PReSS, we validated that inferring selections for promotion based on observed changes in MilPDS projected pay grade results is highly accurate at the annual level, with all but one year between 2014 through 2023 having an error rate of less than 2 percent for both officer and enlisted personnel at all pay grades in our scope. The one exception is 2020, during which the coronavirus disease 2019 pandemic disrupted many administrative processes and likely affected data quality. In 2020, we observed an ephemeral increase in error rates for E-6 (about 10 percent) and for E-7 (about 4 percent). However, O-4, O-5, and E-8 error rates for 2020 are in line with other years, suggesting that these data issues were isolated to particular pay grades. After 2020, error rates for both E-6 and E-7 returned to baseline levels of less than 2 percent per year.

For the qualitative model, this pseudo board date inference process affects only the specific performance report that we link to an observed change in projected pay grade, with all other information used in the model coming directly from the performance report. Later in this report, we describe a follow-on model that uses qualitative and quantitative data simultaneously to predict selections. In this combined model, the pseudo board date influences both the specific performance report linked to an observed selection and the specific monthly snapshot of other MilPDS features (e.g., "has ever been a squadron commander") associated with each performance report.

Training a Model to Predict Career Development Outcomes

Since the original PReSS prototype was developed, methodological innovations in data analytics have produced models with significant advantages over standard logistic regression. In particular, gradient-boosted decision tree (GBDT) models have become a first-line method of predictive modeling because they provide many of the same advantages as more-complex methods (e.g., neural networks, which were explored as an option for the original PReSS prototype) while also being more intuitive to visualize through the use of decision trees and providing some built-in functions for interpretability and explainability (specifically, feature importance). For our application, the key advantages of a GBDT model over logistic regression include

- the ability to model complex nonlinear relationships (i.e., when depicted on a graph, the true relationship between a predictor variable and the outcome being modeled can have nearly any shape, not just a straight line, and can be accurately represented by the model)
- the ability to capture any meaningful interactions between the predictor variables (e.g., if the use of a word in an OPR or OPB makes an officer more likely to be promoted in some AFSCs but not others, there is an interaction between the word and AFSC) without having to prespecify them in a formula
- built-in parameters to tailor the model to certain objectives, such as using only the most strongly predictive variables in the final trained model to reduce complexity.

We explored multiple GBDT models, each with unique strengths. The results presented in this report for the qualitative officer model were derived using LightGBM, a GBDT implementation developed by Microsoft Research (Ke et al., 2017). However, when implementing the model on Envision, platform limitations discussed further in Chapter 4 prevented the use of LightGBM. Instead, we use PySpark's GBDT implementation on Envision, which natively supports Envision's distributed computing features and thus avoids these platform limitations. We note that, although direct comparison is not possible because the two implementation models are trained using different datasets, we found generally similar levels of performance across them. We also conducted benchmarking analyses using lasso-logistic regression, the methodology employed by the original PReSS, to assess whether the added complexity of GBDT models was justified by enhanced performance. Across models and pay grades, we found that our GBDT model implementation achieved between 3 and 7 percentage points–higher balanced accuracy and often exhibited a better balance between precision and recall than lasso-logistic regression.

We also used optimization methods to further enhance the performance of our GBDT model implementations. The aforementioned built-in parameters to tailor the model, known as *hyperparameters*, have values assigned prior to training, which can have a significant effect on model performance. There is no one-size-fits-all set of optimal hyperparameter values; instead, the optimal values depend on the dataset and outcome being modeled. For this reason, *hyperparameter optimization*—i.e., a systematic process to identify the set of hyperparameter values that achieve the best model performance—is an important part of GBDT model implementation. Recent developments commonly referred to as AutoML (short for *automatic machine learning*) make hyperparameter optimization faster, more efficient, and less reliant on modeler assumptions. Just as our GBDT model implementations differ between our work on RAND computing resources using the original PReSS data and our work on Envision, we also use different hyperparameter optimization routines across the two platforms, employing Microsoft Research's Fast Library for Automatic Machine Learning (FLAML) (Wang et al., 2021) on RAND servers and Hyperopt (Bergstra et al., 2015) on Envision. However, these packages' respective approaches to optimization are very similar; both use repeated retraining of the GBDT model and a suggestion algorithm to efficiently search for the configuration that achieves the best performance.

Evaluating and Interpreting Model Predictions

Unlike simpler methods, such as logistic regression, GBDT models do not automatically produce intuitive tables of coefficients to describe the respective relationships between the outcome and each predictor variable used in training. Instead, by default, they provide *feature importance*, which is a measure of how much the inclusion of a particular variable in the training data led to the model making more-accurate predictions. Although feature importance is useful to determine which predictors are most strongly associated with the outcome, it does not say anything about how the outcome changes as the predictor value changes (e.g., whether the presence of a particular word in an OPR or OPB is associated with a higher or lower likelihood of being promoted). As a result, without additional tools, GBDT models by default are significantly less interpretable than simpler models.

We addressed this limitation to achieve interpretability comparable with a logistic regression in two ways. For binary variables (e.g., "has ever been a squadron commander," with no = 0 and yes = 1), which make up the majority of our predictors, we simply compared the average prediction for positive cases with the average prediction for negative cases, referring to the difference between these averages as the *impact* of that feature. For example, if among the officers for whom a particular word was used in their OPR or OPB, the average predicted career development score was 0.5, and among all other officers, the average score was 0.3, we assign an impact of +0.2 to this word. Although this is an approximation and might not perfectly represent the impact of a feature, especially for features that strongly interact with other features, it is more simple, scalable, and intuitive than alternatives and likely accurate enough to substantively represent the relationship of interest. For continuous variables (e.g., commissioned years of service), because adapting our approach of comparing averages for binary features would require a potentially infinite number of comparisons between pairs of possible values, we chose to use SHapley Additive exPlanations (SHAP) values (Lundberg and Lee, 2017). In essence, SHAP estimates how much each predictor in the model contributed to its predictions at the individual level. For each observation, a distinct SHAP value is calculated for each predictor in the model, and the sum of these SHAP values equals the model's overall prediction for that observation. Traditionally, SHAP values are depicted on a scatterplot, where each point depicts the value of the feature in a particular observation and its associated SHAP value. To simplify this to a single number that can be interpreted similarly to a regression coefficient, we used the underlying data of this scatterplot in a linear regression, using the coefficient of that regression as the impact. However, because GBDT models can estimate nonlinear relationships, we also note cases in which the impact does not fully represent the estimated relationship because the scatterplot suggests that it is nonlinear. Although it is possible to use SHAP to assess the binary predictors as well, and the comparison-of-averages method is easier to understand, we encountered technical limitations on Envision that required limiting the use of SHAP to cases in which a simple comparison of averages would not suffice. These limitations are discussed in Chapter 4.

Building the MilPDS-Only Model

The previous version of PReSS used only qualitative OPR data to predict promotion. However, prior RAND work has demonstrated the potential for statistical and machine learning models to accurately predict key events in USAF service members' careers using quantitative data (Robson et al., 2022; Schulker, Harrington, et al., 2021; Lim et al., 2025).

We built a quantitative version of PReSS using personnel record data from MilPDS to explore potential advantages of incorporating quantitative data into PReSS. First, we used a feature selection workflow to identify a subset of key career development features. Using the selected features, we built a machine learning classifier (Model 2) to produce a quantitative career development score, analogous to the score from the performance report–only model (Model 1).

Feature Selection

We trained an extreme gradient boosting (XGBoost) classifier to predict an officer's likelihood of promotion. To avoid overfitting, we trained our final models using only a subset of predictors, or features, available within our officer-level dataset. To determine which features to include in our final models, we trained prototype models using different subsets of features and evaluated their performance. Additionally, we considered the following factors:

- relevance to the outcome event (promotion selection)
- lack of redundancy with other features
- lack of relevance to USAF personnel policy
- organizational changes during our project period.

Our feature selection process spanned multiple phases, each prioritizing different sets of these factors. First, to reduce around 130 initial features to a more manageable number, we employed a feature-reduction routine to calculate the variation inflation factors (VIFs) of all variables within our feature set. VIF calculates the extent to which other features can predict a given predictor. For example, if one knows an officer's AFSC, one knows their developmental category (DEVCAT) by extension, so these two features would have a high VIF with respect to one another. Although this example includes two variables, VIF can be extended to sets of many variables, and a variable's VIF score is a measure of how redundant information within that variable is, given other information available in the predictor set. After calculating each predictor's VIF, we pruned variables that exhibited both high VIFs (indicating redundancy) and low Pearson correlation coefficients with the outcome variable of interest (indicating a lack of relevance to the target variable). We iterated until we were left with roughly 80 predictors.

In the next phase of feature selection, we removed features that might be predictive of promotion selection but lack relevance to USAF personnel policy. For example, certain officer variables, such as marital status, might affect retention and career development but are primarily demographic in nature. This limits their usefulness for DTs looking for meaningful policy insights about the career development process. After eliminating features based on this criterion, we were left with a set of roughly 50 features.

In the subsequent phase, we removed variables that seemed attached to organizational processes likely to change in the near future. For example, whether an officer had previously been promoted BPZ was historically an important predictor of promotion board success. However, officer zone promotion board eligibility policies have changed, meaning that features based on BPZ promotions would not be prudent to include in a model intended to be used with current and future data. However, because the MilPDS-only model's historical training data included years during which it was possible to be promoted BPZ and therefore featured several instances of BPZ officers meeting a promotion board, we still included an indicator for whether the officer was BPZ at the time they met the board in the model. This ensured that the lower selection rate among BPZ officers in years when they met the board relative to in-the-promotion-zone (IPZ) officers was not misattributed to other characteristics that might have differed between BPZ and IPZ officers.

This phase of reductions resulted in a set of about 30 variables. We further trimmed this set after considering the feature importance values for each feature within the prototype micromodels we trained. Gain-based feature importance essentially measures how much each input feature (e.g., age, income) improves the accuracy of a decisionmaking model by comparing the accuracy before and after the model leverages that feature. The final set of features used in the final models and their definitions are shown in Table 2.1.

Table 2.1. MilPDS-Only Model Feature Descriptions

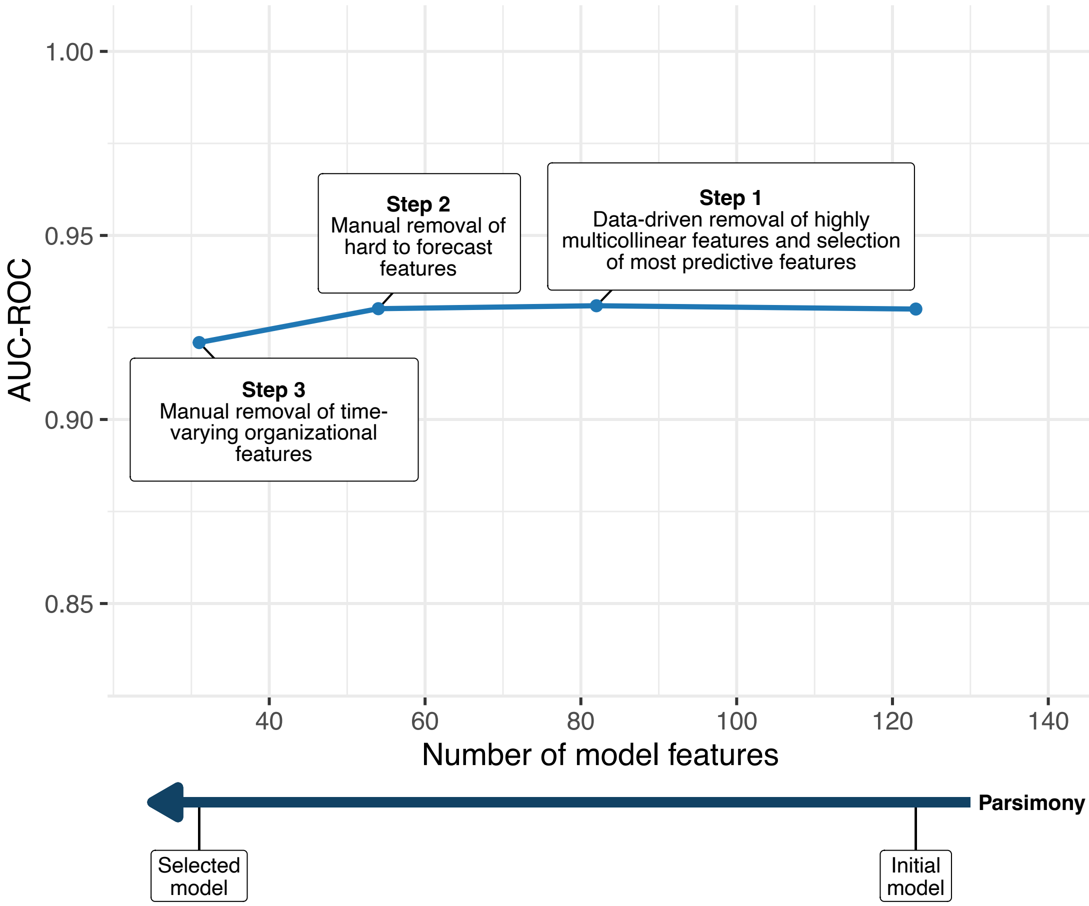
Raw Variable Name	Data Type	Description
CYOS	Numeric	Number of commissioned years of service (CYOS)
MASTER	Numeric	Equal to 1 if officer has a master's degree
PHD	Numeric	Equal to 1 if officer has a doctoral degree
REAL_SQCC_EVER	Numeric	Equal to 1 if officer ever had a squadron commander assignment
WPN_GRAD	Numeric	Equal to 1 if officer is an Air Force Weapons Instructor Course graduate
ML_EVER	Numeric	Equal to 1 if officer has ever been a materiel leader
TIME_IN_GRADE	Numeric	Number of years officer has spent at current pay grade
PMECOMPLETE	Numeric	Equal to 1 if officer completed grade-appropriate professional military education by any method
IDERES	Numeric	Equal to 1 if the officer attended intermediate developmental education (IDE) in-residence
SDERES	Numeric	Equal to 1 if the officer attended senior developmental education (SDE) in-residence
DEVCAT	Character	Core AFSC DEVCAT
PAYGRADE	Character	Current pay grade
SOC2	Character	Source of commission

Across all three phases of feature selection, we evaluated how our feature reductions affected model performance to ensure that we were not degrading the predictive power of the model in appreciable ways. Figure 2.1 displays how the area under the receiver operating characteristic curve (AUC-ROC) of the O-5 MilPDS-only model changed as we progressed through all feature selection phases.[5] As shown by the nearly horizontal line in the plot, minimal performance degradation was observed, even after reducing the number of features by approximately ten times compared with our initial starting point. This indicates that our feature selection process successfully retained the most important features while eliminating redundancy and irrelevant information.

[5] AUC-ROC is one of the most common performance metrics for binary classifiers. It improved understanding on how well a model distinguishes between two classes—in this case, officers who are promoted and those who are not. In more detail,

- the **receiver operating characteristic (ROC) curve** shows a classification model's performance at all classification thresholds. It plots two parameters:
  - **true positive rate (TPR)**, which is the proportion of actual positives correctly identified by the model
  - **false positive rate**, which is the proportion of actual negatives incorrectly identified as positives by the model.
- **area under the curve (AUC)** summarizes the entire ROC curve into a single value. The AUC represents the probability that the model ranks a randomly chosen positive example (an officer who was promoted) higher than a randomly chosen negative example (an officer who was not promoted). An AUC of 1.0 indicates a perfect model, while an AUC of 0.5 indicates a model that performs no better than random chance.

Figure 2.1. The Phases of Feature Selection



For our feature selection prototype models, we used a chronological train-test split for our model training, in which the models were trained on data obtained before 2020 and evaluated on data obtained after 2020. This approach ensures that the models are tested on future data, simulating real-world conditions. As a robustness check, we also explored splitting our train-test split by officer ID instead of time period and noticed similar performance and results. Separate models were trained to predict O-4 and O-5 officer promotions.

Quantitative Model

The officer MilPDS-only model discussed in this section was derived using the XGBoost machine learning algorithm (Chen and Guestrin, 2016). However, as with the qualitative model, we used PySpark's GBDT implementation on Envision and found generally similar levels of performance across the two implementation methods.

We implemented an 85 percent–15 percent training-validation set split on the data. We merged the quantitative data with the qualitative OPR data to ensure that the records used in the training and validation sets were aligned, allowing for a direct comparison of performance on identical validation sets. We preprocessed the data by one-hot encoding the categorical features and scaling continuous features to have a mean of 0 and a standard deviation of 1. We used fivefold cross-validation to tune model hyperparameters before fitting the final model.

We explored the effects of calculating each variable in Table 2.1 in two distinct ways: (1) a binary variable indicating whether an officer had ever achieved a given degree or experience (e.g., master's degree) and (2) a binary variable describing whether the officer had achieved a given degree or experience between their last promotion board meeting and the date when the target outcome was recorded. This lagged approach allows the model to determine whether the recency of a given degree or career development experience is relevant to a given target outcome. However, we did not replicate lagged variables on Envision. Although sometimes the model with lagged variables had the highest accuracy, the difference in performance was often marginal. Additionally, removing lagged variables limits the number of distinct models needing to be run, avoiding excess use of shared Envision resources.

Combining Qualitative and Quantitative Models

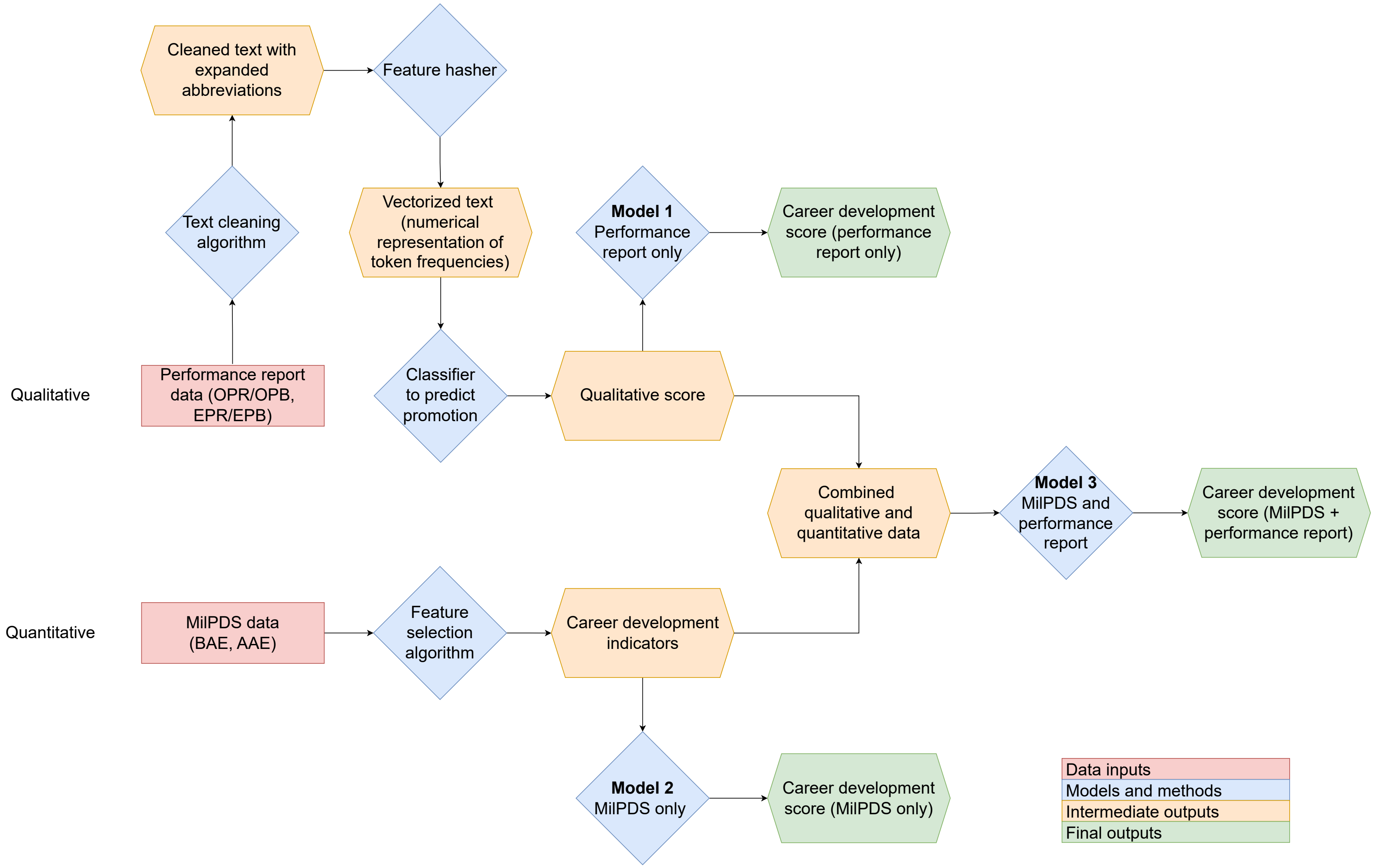
After building the performance report–only model (Model 1) and the MilPDS-only model (Model 2), we explored two methods for combining qualitative and quantitative data to build Model 3:

- **Adding the performance report–only career development score as a feature.** With this method, we used the same methodology as the MilPDS-only model but added a feature with an OPR or OPB score to the training dataset. This OPR or OPB score is the predicted probability of promotion produced by the performance report–only model. In this method, the OPR or OPB score serves as a proxy for qualitative OPR data.
- **Feeding outputs from the two separate models into a stacked classifier.** With this method, we implemented a traditional stacked classifier, which uses an additional model layer to combine the output of a set of individual models and can lead to improved performance (Wolpert, 1992). First, we used the MilPDS-only and performance report–only models to produce distinct probabilities of promotion. Then, we used these two career development scores as input to a logistic regression classifier. This produces a predicted probability of promotion that incorporates information from both individual classifiers.

Taking Stock

Figure 2.2 visually summarizes the updated PReSS pipeline from source data to the three models produced by the tool. As is evident in the figure, our approach to the new PReSS involves several stages that introduce complexity. If one of these layers of complexity does not substantively contribute to model performance, it might be preferable to exclude it in favor of a simpler model, which would be less demanding of computing resources and have fewer layers in which an error or other unanticipated problem could arise. This is why we created and tested the three models. Comparing the performance of each model helps determine whether an approach using only qualitative data or only quantitative data is sufficient or the enhanced performance of the combined model justifies its higher complexity.

Figure 2.2. Improved PReSS Modeling Pipeline



NOTE: AAE = Active Enlisted Master Personnel Extract; BAE = Active Officer Master Personnel Extract.

Model Evaluation and Results

Evaluation of Performance Metrics

We used a variety of metrics to compare the performance of each model. First, we looked at balanced accuracy, which averages the *TPR*, or the percentage of true positives that the model predicts to be positive, and the *true negative rate*, or the percentage of true negatives that the model predicts to be negative. This metric is used because it accounts for false positives and negatives in imbalanced data better than raw accuracy. In addition to balanced accuracy, we assessed precision and recall of the models. *Precision* is the accuracy of positive predictions or the percentage of positive predictions that are true positives. For a model with high precision, when the model predicts that an officer will be promoted, they usually are. *Recall* is the number of true positives that were predicted correctly. In a model with high recall, when an officer is promoted, usually the model predicted they would be. Each of these metrics ranges from 0 to 1, with larger values indicating better performance. We compared the output of the OPR-only and MilPDS-only models, as well as the output of the best-performing combined model. Table 2.2 shows the performance metrics for O-5 models, and Table 2.3 shows the performance metrics for O-6 models. Between the two approaches to combining qualitative and quantitative data that we tested (adding the qualitative OPR score from Model 1 to the quantitative Model 2 as a predictor and using the results of the separate Models 1 and 2 in a stacked classifier, respectively), the best-performing combined model depends on which accuracy metric is considered the most important to maximize. In each table, we have bolded the model that achieves the highest balanced accuracy because this metric combines both possible types of prediction error (false positives and false negatives) into a single summary value. However, prioritizing either maximal precision (i.e., minimizing false positives) or maximal recall (i.e., minimizing false negatives) can change how the models' respective performances stack up.

Table 2.2. O-5 Model Performance Metrics

Model	Balanced Accuracy	Precision	Recall
OPR-only	0.77	0.52	0.78
MilPDS-only	0.74	0.83	0.50
MilPDS + OPR feature (combined model)	**0.83**	**0.69**	**0.70**
Stacked classifier (combined model)	0.82	0.76	0.67

Table 2.3. O-6 Model Performance Metrics

Model	Balanced Accuracy	Precision	Recall
OPR-only	0.79	0.68	0.74
MilPDS-only	0.81	0.74	0.66
MilPDS + OPR feature (combined model)	**0.83**	**0.86**	**0.67**
Stacked classifier (combined model)	0.78	0.77	0.59

The O-5 combined models outperformed the models that use either only qualitative or only quantitative data. Although the stacked classifier model's precision is 0.07 lower than the MilPDS-only model, its recall is 0.17 higher than the MilPDS-only model. A similar difference in score can be found between the OPR-only model and the combined model that adds the OPR score as a feature. For the Deputy Chief of Staff for Manpower, Personnel, and Services's (AF/A1's) use case, the best model should be balanced between precision and recall. If a model has low precision, we will think that more officers have fulfilled expectations for the next career development milestone than actually have. Conversely, if a model has low recall, we will fail to identify many of the officers who have actually fulfilled expectations. It is better to have a balance between these two metrics to not over or underpredict career development scores, which is why the combined models are best, despite not having the highest scores in every metric.

The best model for O-6 is the one that adds the OPR score as a feature to the MilPDS model. This model has the highest accuracy and precision, although its recall is slightly lower than the OPR-only model. For both O-5 and O-6 models, models that combine MilPDS and OPR information perform the best.

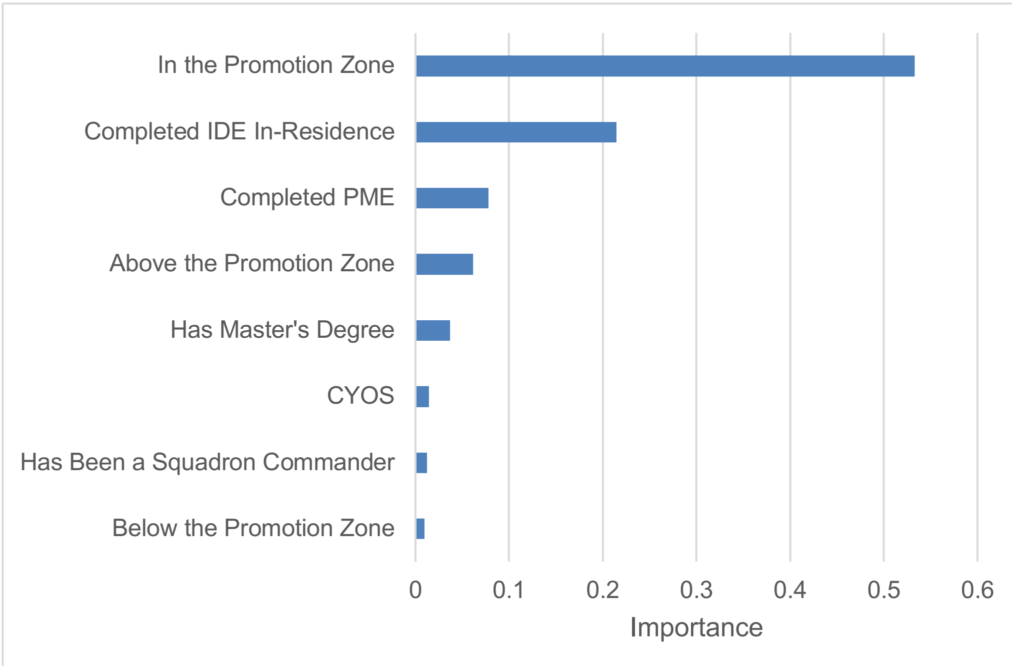
Feature Importance

For the best-performing models, we can further investigate the contribution of individual features to the prediction by using feature importance. More specifically, we use a measure of feature importance called *gain*, which measures how much each input feature improves the accuracy of a decisionmaking model by comparing the accuracy before and after the model leverages that feature.[6] We can assess the importance of each feature in the model to determine which features are the strongest predictors of promotion. In Figure 2.3, we can see that being IPZ and completing IDE in-residence are the strongest quantitative predictors of promotion to O-5. Additionally, in Figure 2.4,

[6] More specifically, gain approximates the average increase in predictive accuracy achieved by the model across all decision tree splits using a specific variable. Unlike other feature importance measures, such as *frequency* (which simply counts the number of times a variable is used in a tree split), gain accounts for both how often a variable is used in a GBDT *and* the impact of those uses on a performance metric. This means that variables with low variance (e.g., a dummy variable that equals 1 for only a single observation and 0 for all others) are unlikely to have high gain because they have a limited impact on overall model performance, even if they are frequently used in splits.

we can see that an officer's promotion zone, whether IPZ, above the promotion zone (APZ), or BPZ are the top three most important features to predict promotion to O-6.[7]

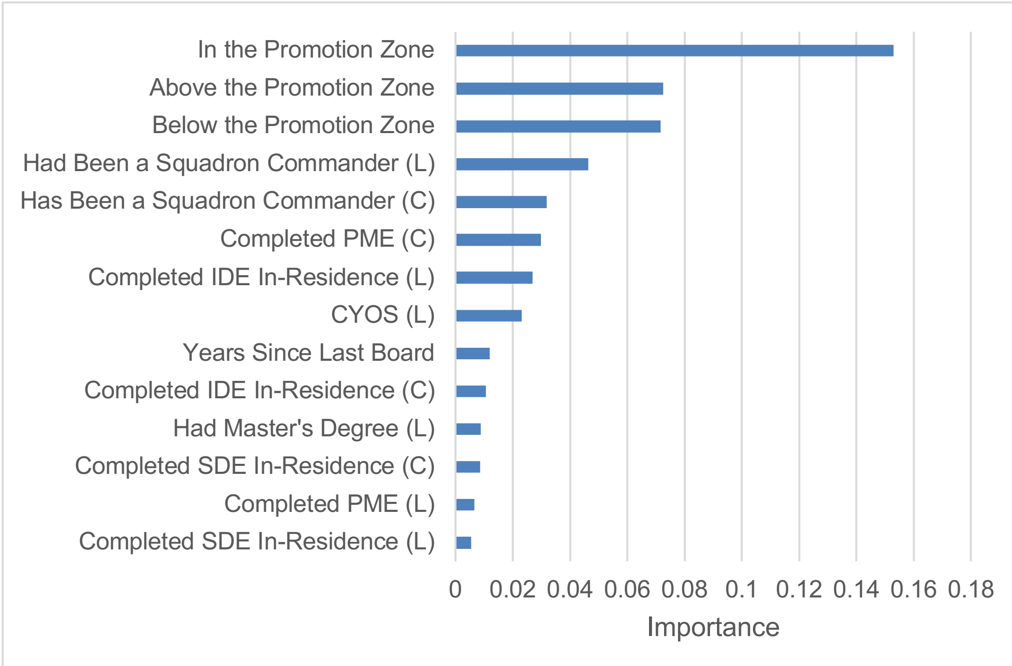
Figure 2.3. O-5 Quantitative Feature Importance



NOTE: PME = professional military education.

[7] Because it is no longer possible for officers to be promoted BPZ, BPZ officers will not appear in new data used to retrain PReSS in the future. Given enough years of new data, it will eventually become possible to retrain PReSS on an amount of data similar to what we used to train the new officer PReSS without including any data from years in which it was possible to be promoted BPZ. At that time, the indicator for the officer being BPZ should be removed from the model. However, the indicator for the officer being APZ should be retained, because that will still apply to some officers meeting future boards, and evidence suggests that it is a significant predictor.

Figure 2.4. O-6 Quantitative Feature Importance



NOTE: (C) = at current board; (L) = at last board. In other words, (C) features are contemporaneous with the board being predicted, but (L) features are lagged by one year.

Similarly, we can extract the feature importance of qualitative predictors. The qualitative data have one feature for each unique token found in the OPR text. Therefore, the feature importance of the qualitative model shows how strongly each token can predict promotion. For example, in Figure 2.5, we can see that the words "senior" and "intermediate" are the most influential token predictors for promotion to O-5. In Figure 2.6, we can see that stratification and the word "group" are the most influential token predictors for promotion to O-6.

Figure 2.5. O-5 Qualitative Feature Importance

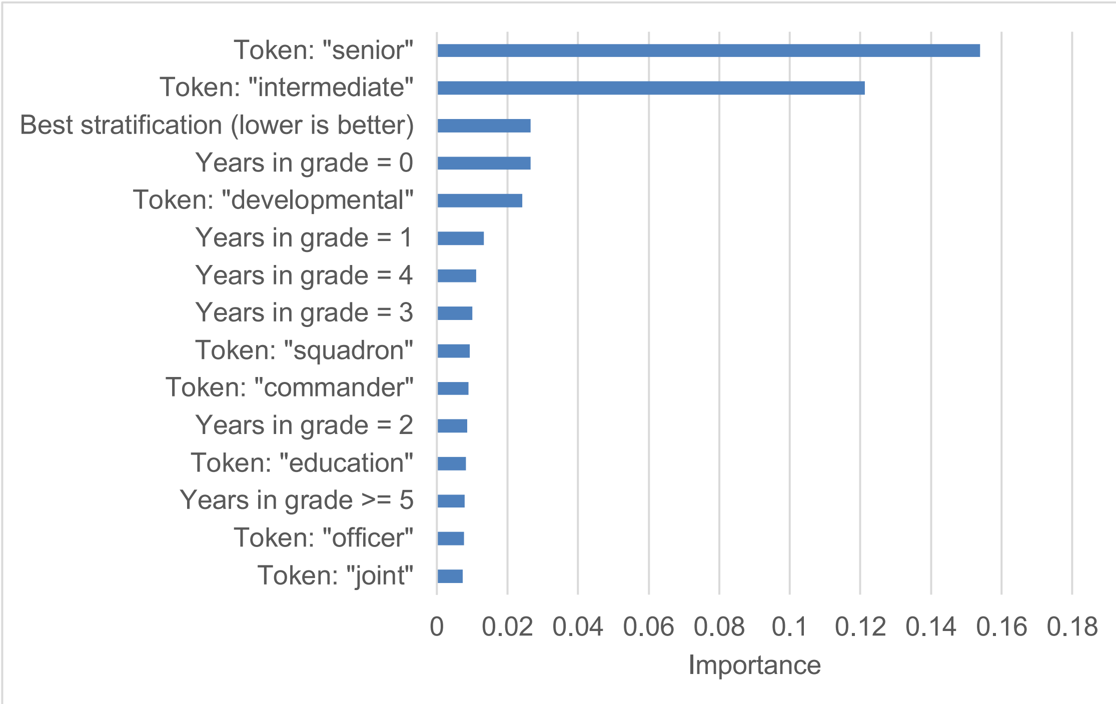
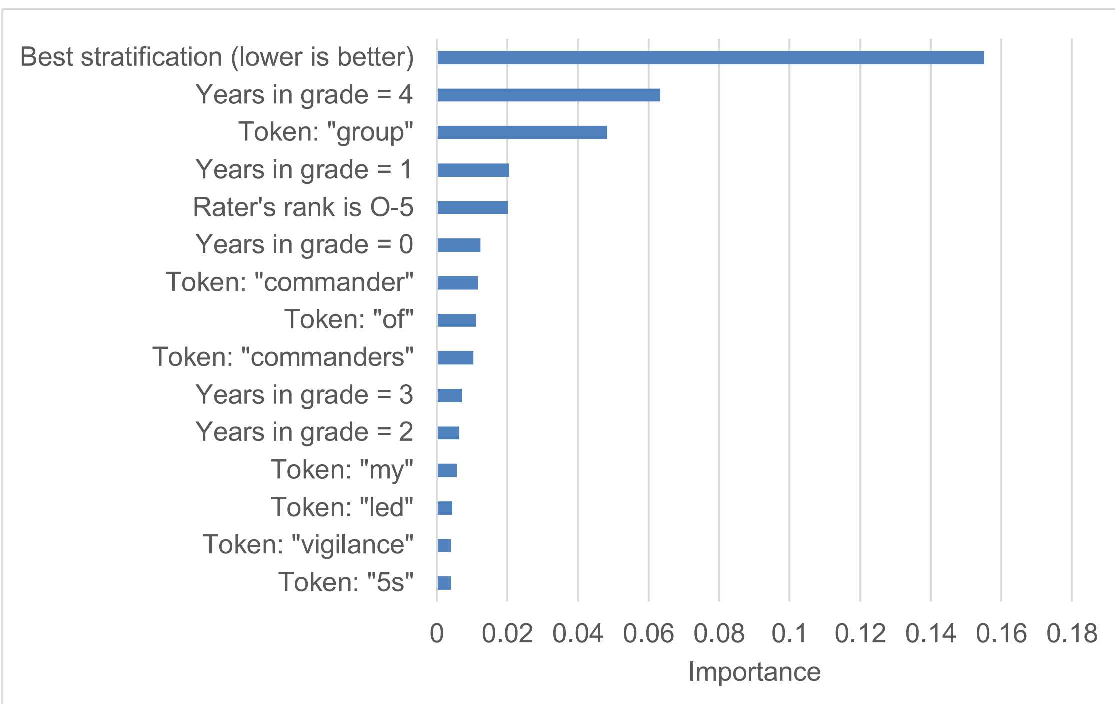


Figure 2.6. O-6 Qualitative Feature Importance



Feature Impact

Although feature importance can tell us which features are strong predictors of our outcome variable, it does not reveal whether that feature is a positive or negative predictor. For example, IPZ is the feature with the highest importance for the O-5 quantitative model. However, we do not know whether this is because officers with IPZ tend not to be promoted or officers with IPZ are more likely to be promoted.

In addition to feature importance, we can assess *feature impact*, which represents how the predicted value (likelihood of promotion) tends to change as the predictor's value increases or changes. For binary and categorical variables, we compute the impact by taking the difference in expected career development score for each value of the feature. For continuous variables, we compute the impact based on SHAP values (SHAP, undated). SHAP values measure the contribution of each predictor toward the outcome for each observation. We convert individual SHAP values to a single measure of impact by fitting a linear regression on the feature and its SHAP values and taking the coefficient.

In Figure 2.7, we see that IPZ and IDE in-residence, the two most important features in the O-5 model, are both positive indicators of promotion. Conversely, in Figure 2.8, we see that IPZ is a positive indicator of promotion to O-6, and APZ and BPZ are negative indicators of promotion to O-6. Similarly for the qualitative models, "senior" is a positive indicator, and "intermediate" is a negative indicator in the O-5 model (Figure 2.9). For the O-6 model, the fractional value of a stratification is a negative predictor (which makes sense, because a lower stratification value means a higher ranking relative to peers), and "group" is a positive predictor (Figure 2.10).

Figure 2.7. O-5 Quantitative Feature Impact

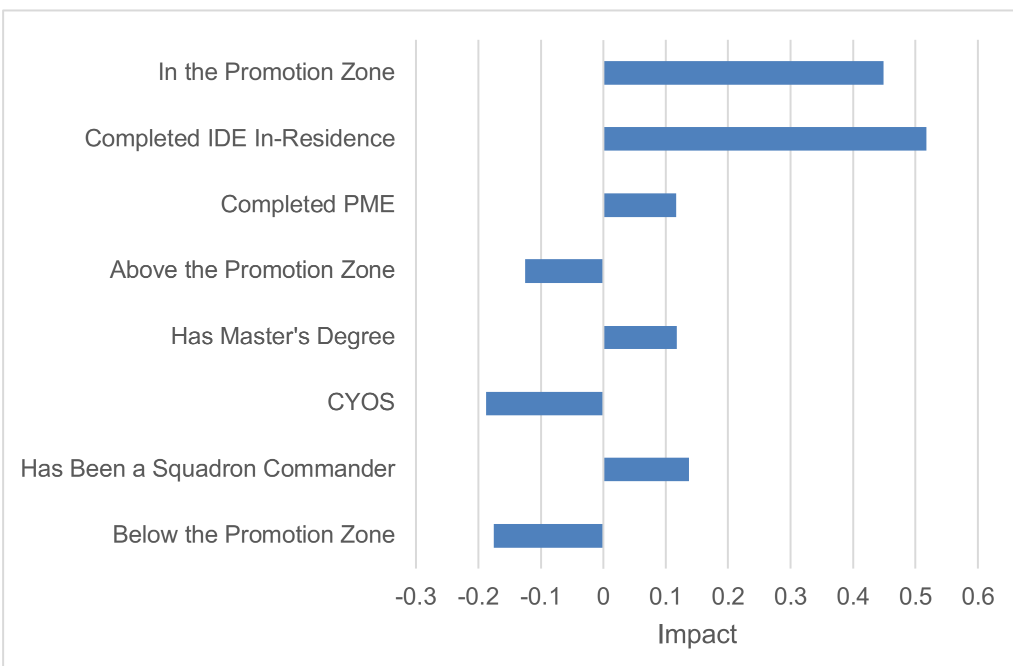
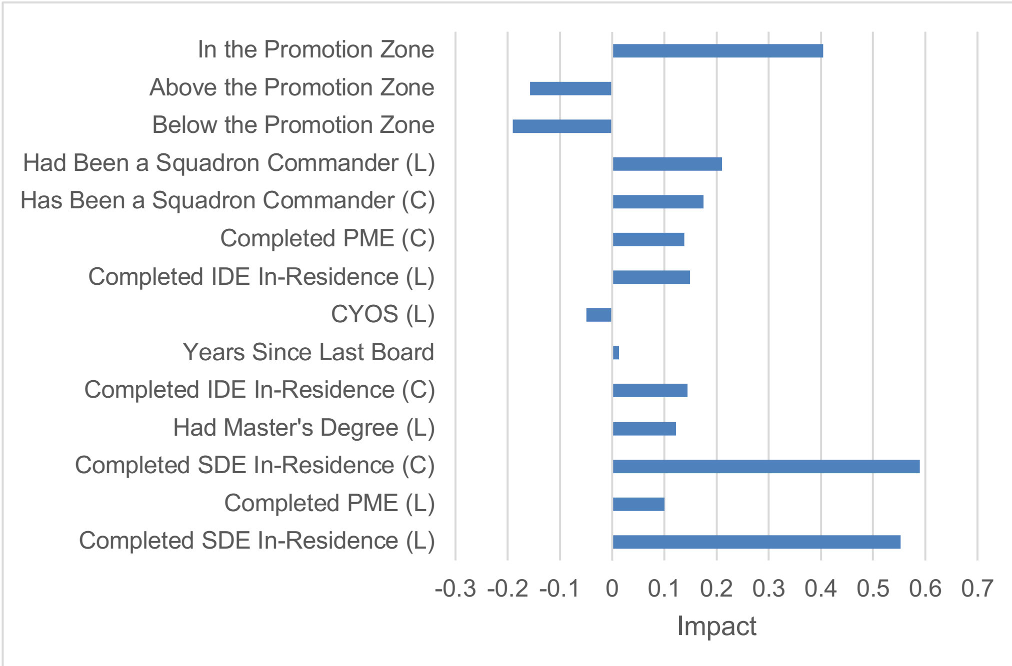


Figure 2.8. O-6 Quantitative Feature Impact



NOTE: (C) = at current board; (L) = at last board. In other words, (C) features are contemporaneous with the board being predicted, but (L) features are lagged by one year.

Figure 2.9. O-5 Qualitative Feature Impact

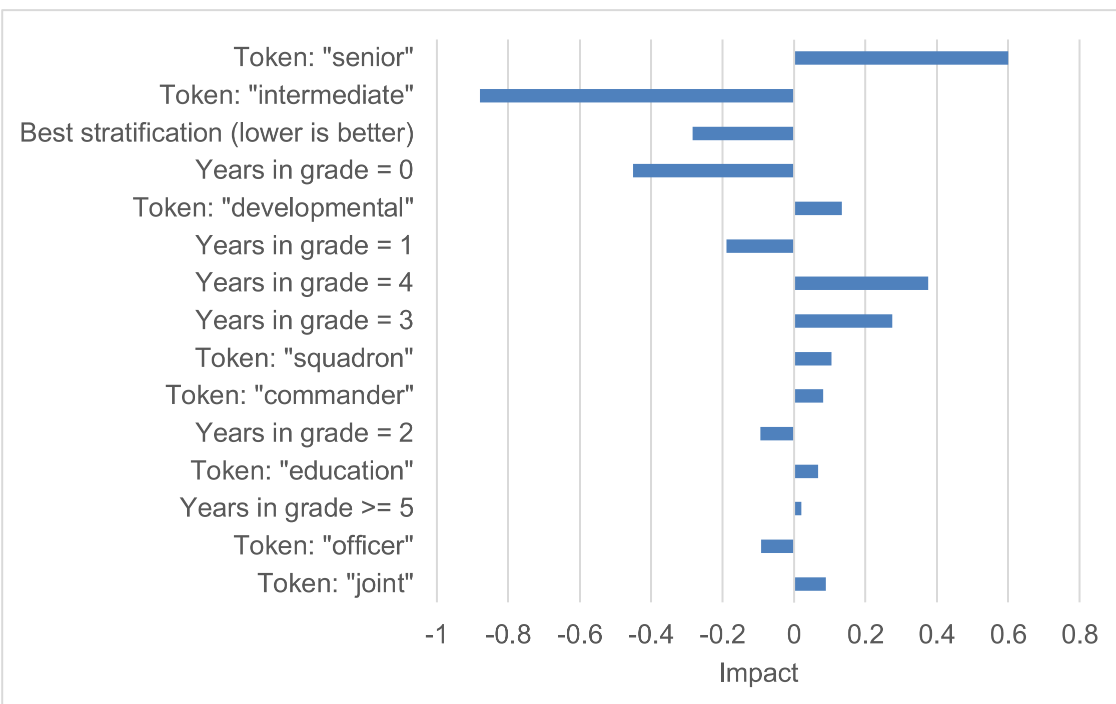
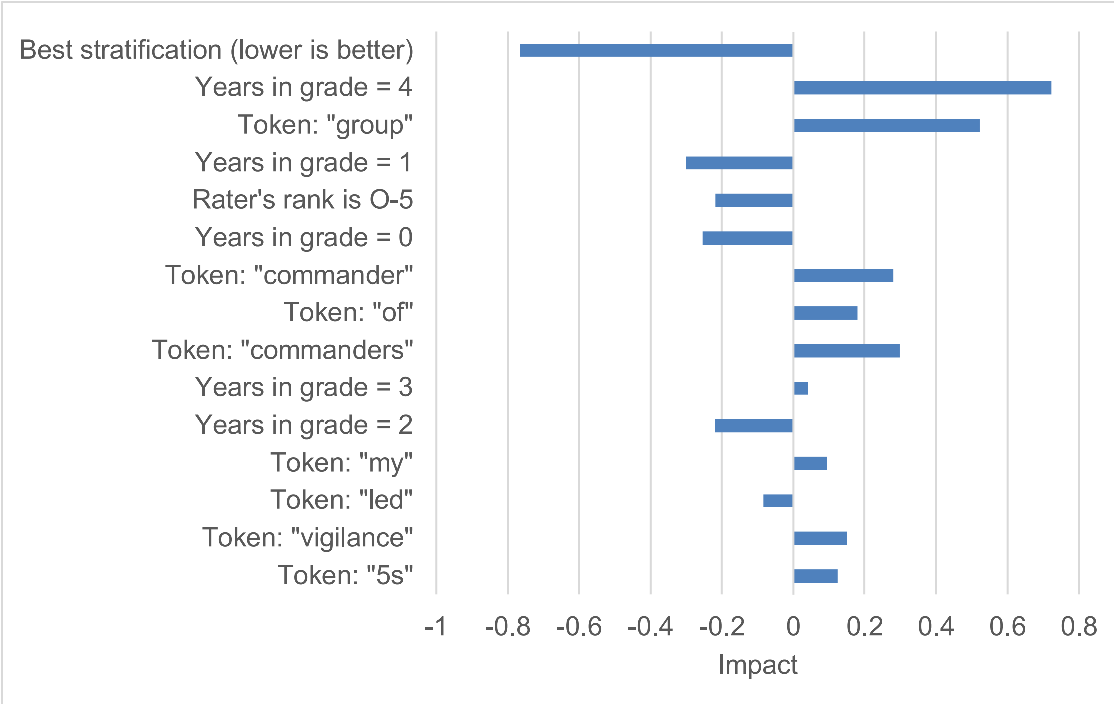


Figure 2.10. O-6 Qualitative Feature Impact



Interpretation of Individual Records and OPRs with Respect to Career Development Score

We can inspect each individual's MilPDS and OPR data to see how each portion of their record contributes to their career development score. As described in the previous section, feature impacts can tell us whether a feature is a positive or negative indicator for promotion. Then, we can look at the quantitative record and OPR score and see which positive indicators are present or missing and vice versa for negative indicators. This will give an idea of why an officer received the career development score they did.

In Table 2.4, we see four example records for majors (MAJ; O-4s) who were considered for promotion to O-5. MAJ-1 has a strong quantitative record and a strong OPR score, leading to a very high career development score. MAJ-2 and MAJ-3 have similar quantitative records, but MAJ-3 has a much stronger OPR score, leading to a higher career development score. MAJ-4 has a weak OPR score, but their strong quantitative record still leads to a high career development score.

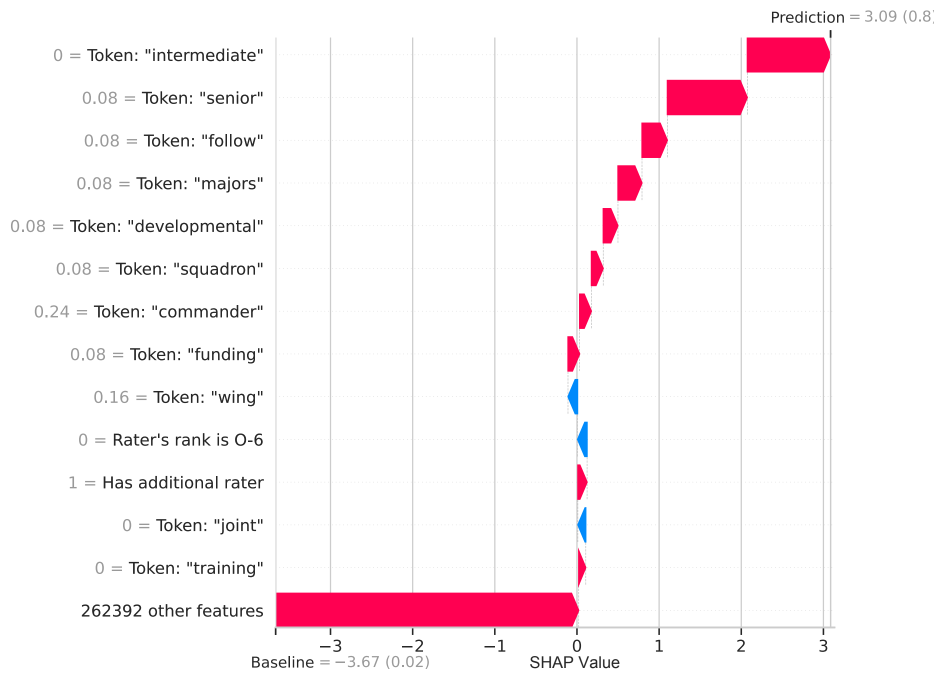
Table 2.4. Example O-5 Candidate Records

Example Record	Promotion Zone	IDE In-Residence	PME Completed	Has Master's Degree	Has Been a Squadron Commander	OPR Score	Career Development Score	Was Promoted?
MAJ-1	IPZ	Yes	Yes	Yes	Yes	80	100	Yes
MAJ-2	APZ	No	Yes	Yes	No	4	0	No
MAJ-3	IPZ	No	Yes	Yes	No	80	98	Yes
MAJ-4	IPZ	Yes	Yes	Yes	Yes	54	96	Yes

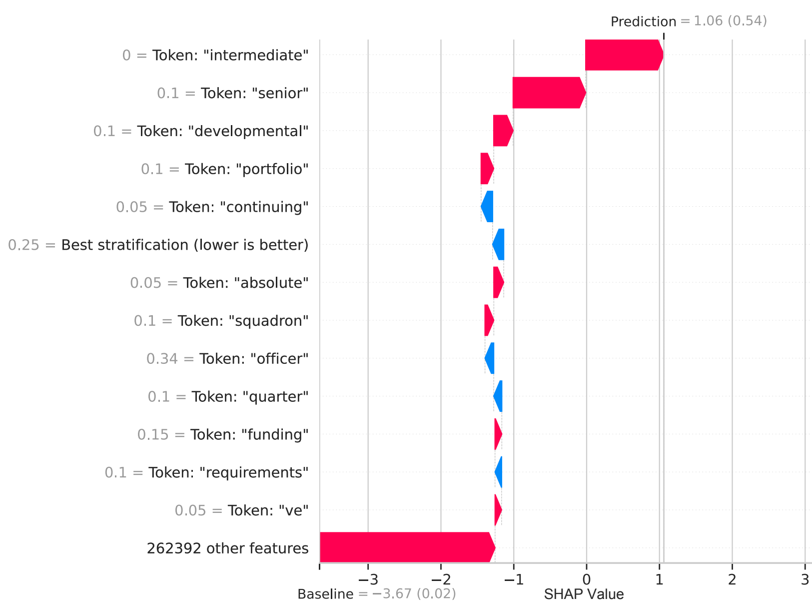
We can inspect the SHAP values of OPR tokens for specific individuals to dig deeper into the OPR score. In Figure 2.11, we see the SHAP values for MAJ-1, who has a relatively strong OPR score, and MAJ-4, who has a relatively weak OPR score. In the figure, red bars (pointing right) indicate features that are positive indicators of promotion, and blue bars (pointing left) indicate features that are negative indicators of promotion. The lengths of the bars indicate the magnitude of the impact that the feature had on the model's prediction. Of the top 13 most impactful predictors, only three have a negative impact on the prediction for MAJ-1, and five have a negative impact for MAJ-4. Additionally, the other 250,000+ features have a stronger cumulative positive impact for MAJ-1 than for MAJ-4.

Figure 2.11. Explaining MAJ-1 and MAJ-4 OPR Scores

Panel A. MAJ-1



Panel B. MAJ-4



NOTE: In the figure, red bars (pointing right) indicate features that are positive indicators of promotion, and blue bars (pointing left) indicate features that are negative indicators of promotion.

Additionally, in Table 2.5, we inspect four example lieutenant colonels (LTC; O-5) to see which parts of their records contribute to their career development scores. LTC-1 has both a strong quantitative and qualitative record, leading to a high career development score. LTC-2 and LTC-3 have similar quantitative records, but LTC-3 has a much stronger OPR score, leading to a higher career development score. Finally, LTC-4 has a strong OPR, but a weak quantitative record prevents LTC-4 from obtaining a high career development score.

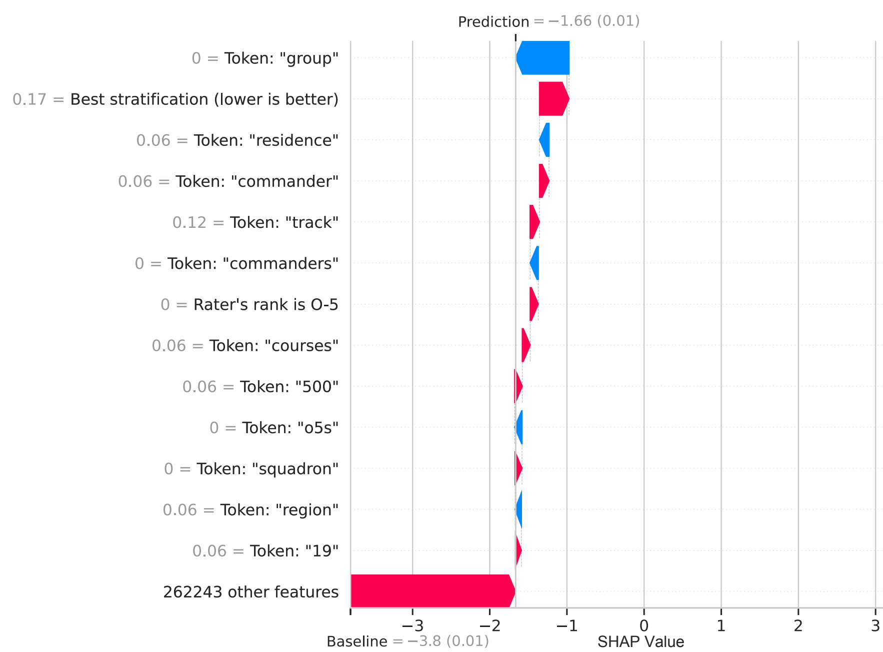
Table 2.5. Example O-6 Candidate Records

Example Record	Promotion Zone	SDE In-Residence	PME Completed	Has Master's Degree	Has Been a Squadron Commander	OPR Score	Career Development Score	Was Promoted?
LTC-1	IPZ	Yes	Yes	Yes	Yes	83	99	Yes
LTC-2	APZ	No	Yes	Yes	No	1	1	No
LTC-3	IPZ	No	Yes	Yes	Yes	62	79	Yes
LTC-4	IPZ	No	No	Yes	No	77	22	No

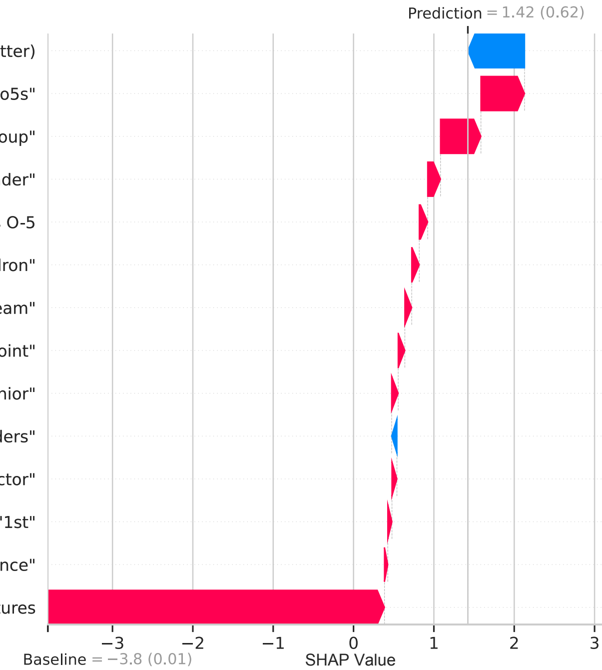
In Figure 2.12, we look at the SHAP values for LTC-2's and LTC-3's OPR scores. LTC-2 has a weak OPR score, and LTC-3 has a moderate OPR score. LTC-2's top predictor, "group," has a negative impact, and the next highest predictor, stratification, has a positive impact. LTC-3's top predictor, stratification, has a negative impact, but the next two predictors, "O-5s" and "group," have positive impacts. Additionally, LTC-3's other features have a stronger positive impact than LTC-2's.

Figure 2.12. Explaining LTC-2 and LTC-3 OPR Scores

Panel A. LTC-2



Panel B. LTC-3



Implementation on Envision

The model training and evaluation pipeline we described was implemented on internal RAND servers and later adapted for and uploaded to Envision for use by AF/A1X and DTs. The distributed computing environment on Envision required us to translate the Python implementations of the performance report–only and MilPDS-only models to PySpark. This included changing from using LightGBM and XGBoost models to PySpark ML's GBTClassifier. Additionally, because of a lack of compatibility between the SHAP package and PySpark, we used an alternative to SHAP values to estimate feature impact on Envision, which we describe in Chapter 3.

Chapter 3

Development of PReSS for Enlisted Personnel

We replicated the improved officer PReSS tool described in Chapter 2 for enlisted personnel. Although the enlisted PReSS tool follows the same general methodology as the officer PReSS tool, there are a few key differences in their implementation because of technical limitations and policy differences.

Methodological Differences with Officer Models

Unlike the officer PReSS, which was developed on internal RAND servers and modified for use on Envision, the enlisted PReSS tool was implemented solely on Envision. This difference in implementation was because EPR and EPB data are solely available for use on Envision. As a result, we were not able to replicate some packages and methods used in training and evaluating the officer PReSS models for the enlisted PReSS models.

Notably, the SHAP package is limited in its compatibility with PySpark models for reasons that we discuss later in this chapter. Therefore, our SHAP value analysis used to interpret feature impacts and individual officer records in the officer PReSS could not be recreated identically for the enlisted PReSS. Instead, for each of the top 20 most important features in the enlisted qualitative model, we estimate a linear regression with the EPR promotion score output (a value between 0 and 1) as the dependent variable and the input feature as the independent variable.[8] The coefficient of this linear regression is our estimate of the impact of the associated variable. In essence, this replicates the last step of our feature impact estimation process for the officer models, but, instead of using calculated SHAP values as the independent variable in the linear regression component, we use the predictions of the model directly. Consequently, unlike those in the officer models, our impact estimates for the enlisted models might not account for possible interactions between the input features and do not directly leverage the tree structure of the predictive model, making them less robust.

[8] Although we had access to raw data from both EPRs and EPBs on Envision, the routine to separate the extracted text from EPBs into distinct features (e.g., rater's pay grade, rater's assessment by line, ratee's core or duty AFSC) under development by AF/A1XD was not complete at the time that the presented analyses were run. EPRs had the same embedded metadata to help with feature extraction as OPRs, meaning that we were able to adapt our existing OPR extraction pipeline from the original PReSS prototype to EPRs. However, unexpectedly, even though the last few years of EPRs almost always featured these metadata, the newer EPBs do not. For this reason, the results presented in this report do not include EPBs (2022–present), and we henceforth refer only to EPRs. However, the Envision capability is set up so that, once the separate features have been extracted from the raw EPB data (provided that this extraction conforms to the same structure as the extracted EPR data), rerunning the enlisted model pipeline will automatically incorporate the new EPB data.

Adapting the Officer Qualitative Model to Enlisted Personnel

Aside from the technical limitations noted previously related to feature impact estimation, our implementation of the qualitative model for enlisted personnel is functionally identical to the implementation for officers. Although the formats of performance reports for officers and enlisted personnel differ, for the subset of enlisted pay grades in scope for our project (E-6 through E-9), the key information that we use in our qualitative models from the narrative assessments is structured near-identically across the two personnel groups. As a result, our process of combining the narrative ratings into a single text column, expanding and standardizing abbreviations used therein, constructing a numerical representation of the reformatted text, and feeding this numerical representation into a GBDT predictive model from the officer PReSS was directly transferable to the enlisted PReSS. The additional fields that we used from officer performance reports alongside the text data (the airman's number of years in grade, their core AFSC, and the respective pay grades of their raters for a particular board) were also available from enlisted performance reports and, therefore, were also able to be derived and used in the predictive model in the same way. Finally, we applied the same process of inferring both selections for promotion and the approximate timing of promotion boards from the officer PReSS to the enlisted PReSS. However, unlike officers, enlisted personnel across AFSCs meet promotion boards at the same time; as a result, although officers' pseudo board dates are assigned by pay grade, year, and AFSC, enlisted pseudo board dates are assigned by pay grade and year only, making them simpler to derive and less likely to be affected by small sample sizes.

Adapting the Officer Quantitative Model to Enlisted Personnel

Differences between officer and enlisted promotion policies led to changes in how the quantitative component of the enlisted PReSS was implemented. In the officer PReSS, we use a robust feature selection routine to identify a small subset of relevant and influential predictors of promotion. This strategy was informed by the fact that the officer promotion process is determined by a central selection board that considers a wide variety of factors to make promotion decisions.

Conversely, the enlisted promotion process is much more deterministic than officer promotions. Historically, enlisted promotions were determined by Weighted Airman Promotion System (WAPS) factors, which include standardized test scores, promotion recommendations, annual performance evaluations, decorations earned, time in service, and time in grade. Promotions to E-5 and E-6 were solely determined by WAPS factors, and promotions to E-7 through E-9 used both WAPS factors and a selection board. We choose to model only promotions to E-7 through E-9 because earlier grades are entirely deterministic.

More recently, promotions to E-7 through E-9 are solely conducted via selection board. However, the historical importance of these limited WAPS factors remains. As a result, instead of using a feature selection routine, we manually selected a set of features that would be relevant to enlisted promotions based on current organizational policy and historical factors. The features included in the enlisted MilPDS-only model are listed in Table 3.1.

Table 3.1. Enlisted MilPDS-Only Model Feature Descriptions

Raw Variable Name	Data Type	Description
CONTROL_AFSC_3DIGIT	Character	First three digits of control AFSC
E4BPZ	Numeric	Equal to 1 if airman was promoted to E-4 BPZ
AFQT_cat	Character	Category of Armed Forces Qualification Test (AFQT) score (I, II, IIIa, IIIb, IV, V)
MECH_cat	Character	Category of Armed Services Vocational Aptitude Battery (ASVAB) Mechanical score (I, II, IIIa, IIIb, IV, V)
ADMIN_cat	Character	Category of ASVAB Administrative score (I, II, IIIa, IIIb, IV, V)
GEN_cat	Character	Category of ASVAB General score (I, II, IIIa, IIIb, IV, V)
ELEC_cat	Character	Category of ASVAB Electronics score (I, II, IIIa, IIIb, IV, V)
TISSC	Numeric	Time-in-service score
TIGSC	Numeric	Time-in-grade score
PHS_PT	Numeric	Phase point (time in service at the time of promotion to current grade)
YRS_TO_ETS	Numeric	Number of years until Expiration Term of Service (ETS)
ADSC	Numeric	Years left in active-duty service commitment (ADSC)
DECSC	Numeric	Decoration score; calculated based on point values attributed to different medals earned; maximum score of 25
BRS	Numeric	Equal to 1 if airman is enrolled in the Blended Retirement System (BRS)
HYT	Numeric	Equal to 1 if airman's High Year of Tenure (HYT) date is before the next promotion cycle
EPR_CURR, EPR_PREV, EPR_PREV2	Character	The past three annual EPR evaluation scores received
VISIBLE_A	Numeric	Equal to 1 if airman received an A on any EPR at the current grade
VISIBLE_V	Numeric	Equal to 1 if airman received a V on any EPR at the current grade
PROMREC_CURR, PROMREC_PREV, PROMREC_PREV2	Character	The past three promotion recommendations received; promotion recommendations are received with every eligible promotion cycle (E-7 cycle and below)
STRAT_CURR, STRAT_PREV	Character	The past two stratifications received; stratifications are received with every eligible promotion cycle (E-8 and E-9 cycles)

Model Evaluation

Evaluation of Performance Metrics

As we did with the officer PReSS tool, we evaluated the performance of the enlisted PReSS using balanced accuracy, precision, and recall. In Tables 3.2–3.4, we bold the model(s) in each table with the highest balanced accuracy (putting more than one in bold in the case of a tie). However, unlike our findings in the officer models, the best model in each enlisted case does not depend on which accuracy metric is prioritized, because the model with the highest balanced accuracy also achieves the highest precision and the highest recall across all three cases.

Table 3.2. E-7 Model Performance Metrics

Model	Balanced Accuracy	Precision	Recall
EPR-only	0.72	0.60	0.62
MilPDS-only	0.65	0.62	0.41
MilPDS + EPR feature	**0.96**	**0.94**	**0.95**
Stacked classifier	0.92	0.93	0.88

Table 3.3. E-8 Model Performance Metrics

Model	Balanced Accuracy	Precision	Recall
EPR-only	0.85	0.57	0.74
MilPDS-only	0.83	0.62	0.70
MilPDS + EPR feature	0.86	0.68	0.75
Stacked classifier	**0.87**	**0.79**	**0.75**

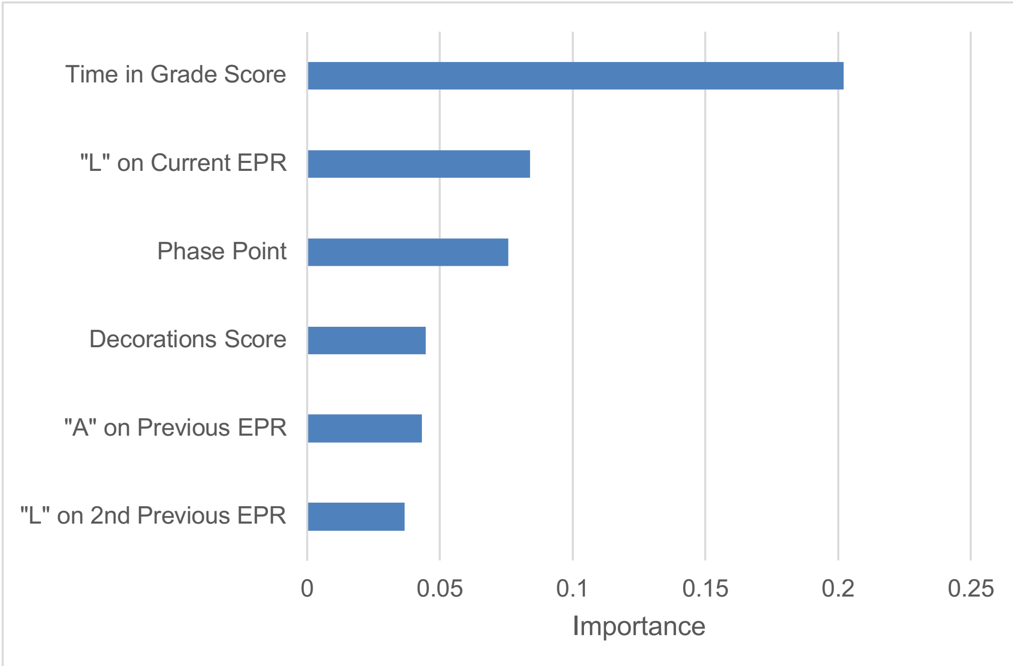
Table 3.4. E-9 Model Performance Metrics

Model	Balanced Accuracy	Precision	Recall
EPR-only	0.84	0.71	0.75
MilPDS-only	0.97	0.76	1.00
MilPDS + EPR feature	**0.99**	**0.86**	**1.00**
Stacked classifier	**0.99**	**0.86**	**1.00**

Feature Importance

As we did with the officer models, we assessed feature importance for the predictors included in the quantitative models for E-7, E-8, and E-9 promotions. In Figure 3.1, we see the feature importance values for the E-7 quantitative model. Time-in-grade score is the most important predictor, indicating that an airman's experience in their current grade is a key consideration for promotion. Receiving an L on the current EPR is the second most important predictor for promotion to E-7. An L represents "exceeds most if not all expectations" and is the highest evaluation an airman can receive on their annual EPR. This feature being important for the E-7 quantitative model indicates that high scores on EPR evaluations are important for promotion consideration.

Figure 3.1. E-7 Quantitative Feature Importance



NOTE: EPRs used two different overall scoring systems over our data sample period. The first, used before 2015, was a numeric 1–5 scale in which higher numbers represented better overall performance. The second, used from approximately 2015 up through the replacement of EPRs with EPBs, was a letter scale with levels “E,” “V,” “A,” and “L” in ascending order. In other words, a 5 under the numeric scale is roughly equivalent to an L under the E-V-A-L letter scale. Because our data sample spans both scoring systems and the scales are not exactly equivalent, we include separate predictors representing 1–5 and E, V, A, and L in the enlisted PReSS models.

In Figure 3.2, we see the feature importance values for the E-8 quantitative model; these reveal that receiving a stratification for the current promotion board is the most important predictor for promotion. This finding aligns with enlisted promotion policy and our findings from the officer models. Stratifications are used only in enlisted promotion beginning at E-8. Once stratifications are available, they are a key indicator of promotion for airmen, similar to how stratification was a key predictor in the officer qualitative models. However, in Figure 3.3, we see that receiving a stratification is not as important of a predictor in the E-9 model. Instead, similar to the E-7 model, time in grade is the most important predictor. This might indicate that an airman’s experience is more important to the selection board than stratifications for promotion to E-9.

Figure 3.2. E-8 Quantitative Feature Importance

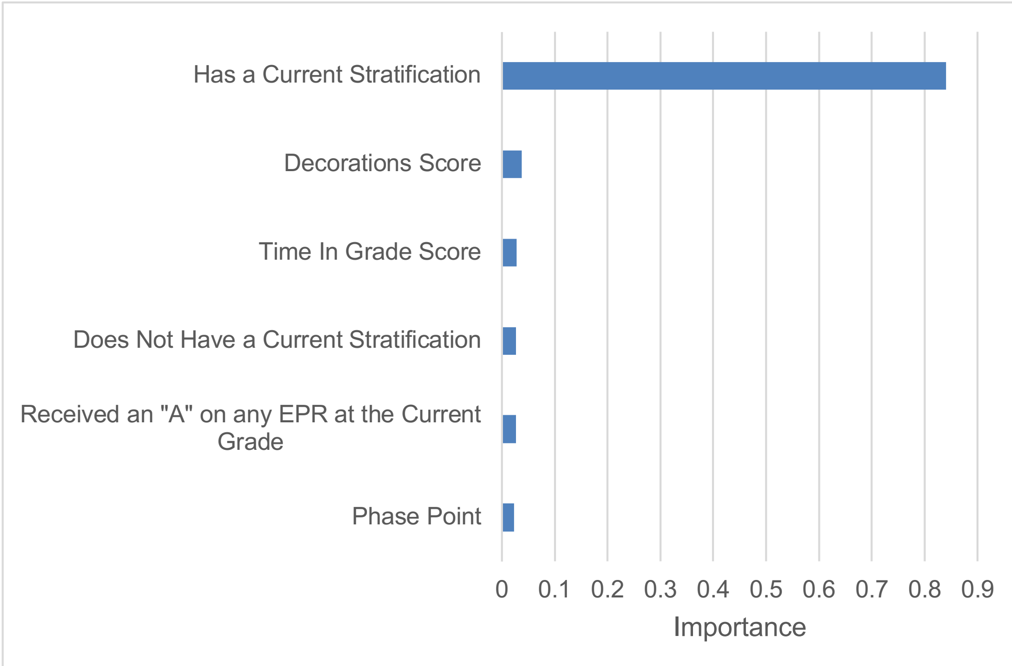
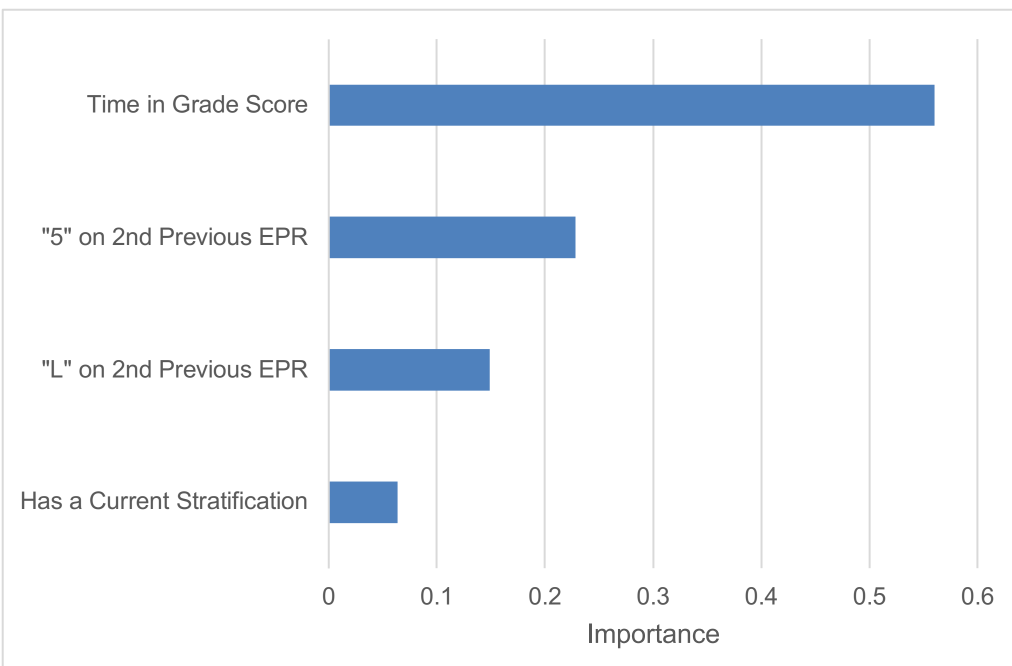
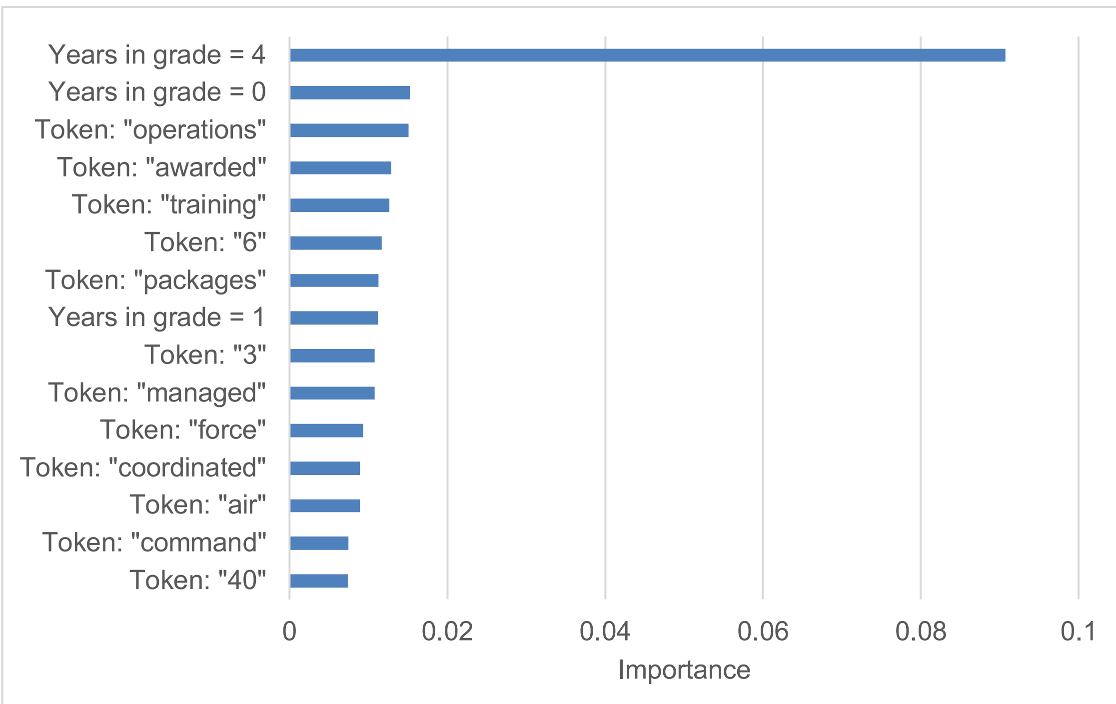


Figure 3.3. E-9 Quantitative Feature Importance



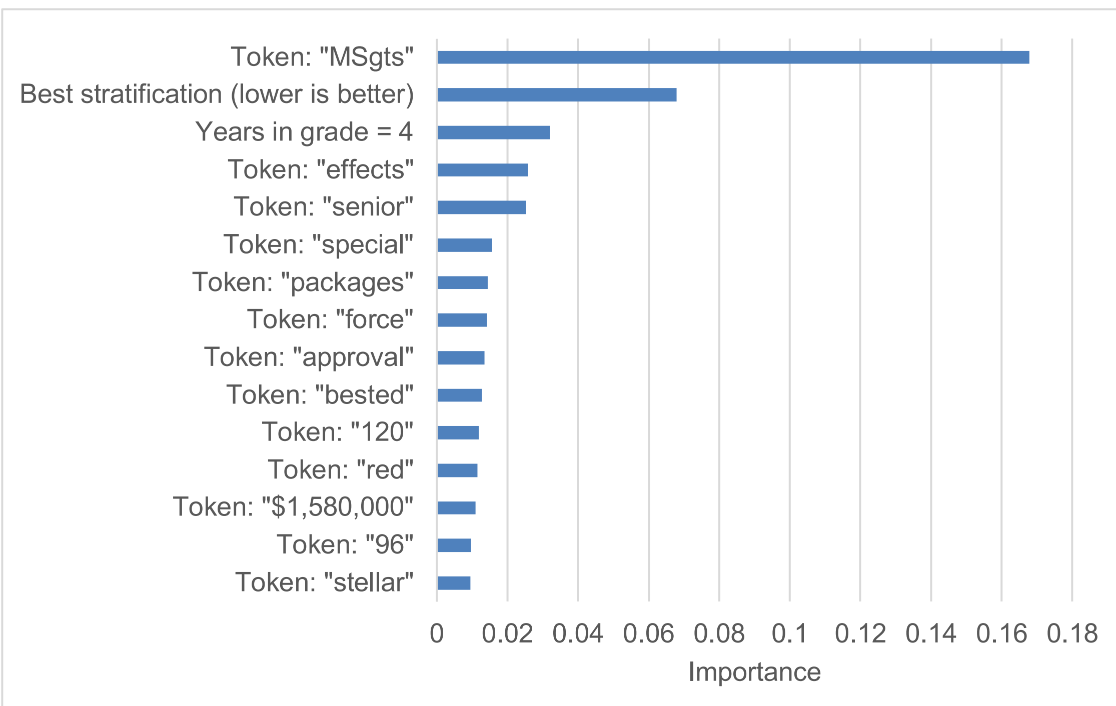
Next, we present the feature importance values for predictor variables included in the E-7, E-8, and E-9 qualitative models. Across the three qualitative models, we see patterns in relative feature importance that generally follow the patterns we saw in the officer qualitative models. For promotion from E-6 to E-7, the two most important predictors are both related to the airman's time in grade. Other top predictors are numbers: "6," "3," and "40." However, the relevance of these specific numbers does not always have an obvious interpretation, because there are several reasons that a number might appear in a performance report. We note that mentions of quantities are extremely common in performance reports, because raters quantify the ratee's accomplishments (e.g., "briefed 6 General Officers," "developed process for 3 Areas of Responsibility," "taught 40 students"). However, a number might also be used to refer to a particular squadron, group, or wing, meaning that the presence of a number as an important predictor might reflect differences in career trajectories across these organizational units. Other high-importance tokens for E-7 include descriptors of professional responsibilities or accolades: "operations," "awarded," "training," "managed," and "coordinated."

Figure 3.4. E-7 Qualitative Feature Importance



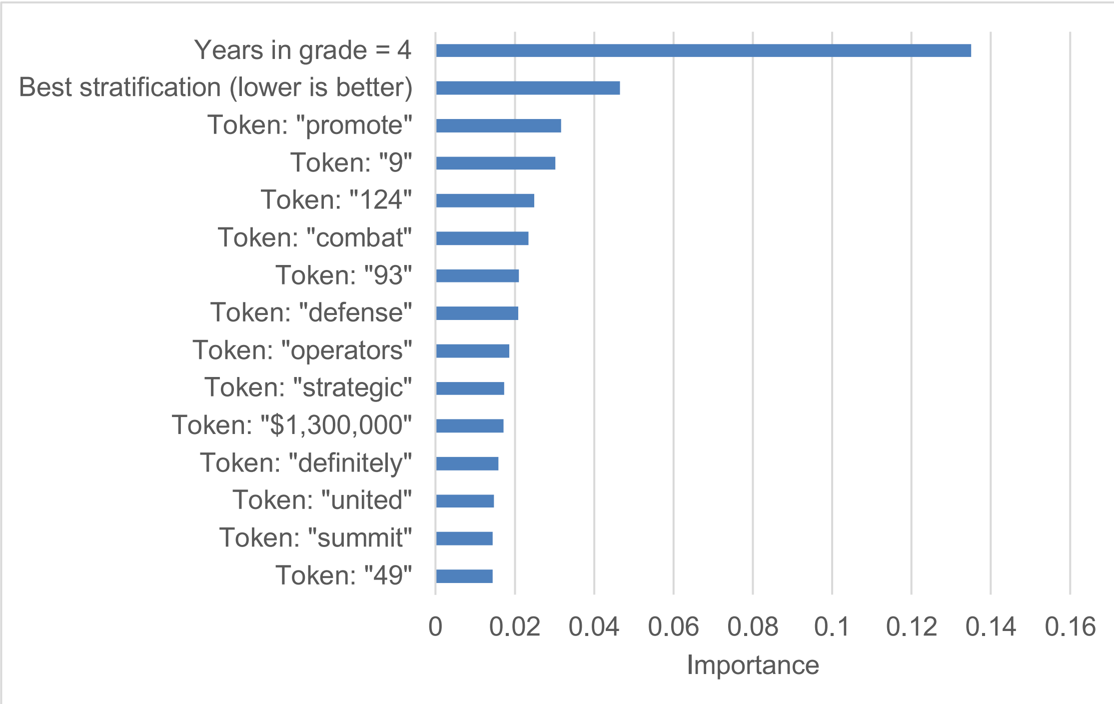
For the E-7 to E-8 and E-8 to E-9 qualitative models, we see several similarities with the E-6 to E-7 model and one key difference: stratifications. Stratifications do not rank highly in importance for promotion from E-6 to E-7 (likely because stratifications are a norm only for E-7 and above; there is no stratification box on the EPR form used for E-6s) but are among the most important predictors for promotion to the two highest enlisted pay grades. This tracks with our findings for officers, where, although stratifications had high importance for promotion to both O-5 and O-6, the magnitude of this importance was significantly larger for promotion to O-6, suggesting that receiving high stratifications is increasingly critical to career progression as one's pay grade rises. Interestingly, "MSgts," a proxy for the presence of a stratification in an E-7's performance report, is the most important predictor for promotion to E-8, although "MSgts" is not among the top predictors for promotion to E-9. In sum, these findings suggest that stratifications are not among the key predictors of promotion to E-7, the presence of *any* stratification is a key predictor for promotion to E-8 and the quality of that stratification has an additive effect, and the presence of a *high* stratification is a key predictor for promotion to E-9. We also see the word "promote" among the most important predictors of promotion to E-9, suggesting similarities with the officer promotion to O-5 model in which push lines were found to be critical determinants of selection.[9] Also, as in the E-7 model, we find some numerical predictors ("120," "96") among the most important predictors, along with a monetary amount ("$1,580,000"). Although $1,580,000 might appear too specific to be a useful predictor, it appears in 27 unique EPRs, making a predictive association plausible. In this case and elsewhere, monetary amounts in EPRs represent the dollar value of an effort to which the airman contributed or resources that the airman oversaw or procured. Although the specific amount varies, such values appear to be quite common across EPRs (and, to a lesser extent, OPRs) and might be used to signal the airman's capability to handle high-value, high-stakes responsibilities.

Figure 3.5. E-8 Qualitative Feature Importance



[9] A *push line* is the very last line of an OPR or EPR, in which raters directly suggest a course of action for the ratee in the near future. For example, a rater might say "send to IDE soonest" to indicate that they believe the airman is qualified to attend IDE at the next possible opportunity. Subtle differences in push line word choices could signal differing levels of enthusiasm or immediacy: "IDE soonest" might indicate that the rater believes that the airman is ready for IDE right now, and "IDE next" could mean that the the the rater thinks that airman is tracking toward IDE but needs to do something else first.

Figure 3.6. E-9 Qualitative Feature Importance



Feature Impact

To estimate the *impact* of the most important predictive features in the enlisted quantitative and qualitative models, we conducted a calculation similar but not identical to the one we used for the officer models. As we mentioned at the beginning of the chapter, the officer models were initially trained on RAND's internal computing resources and then ported over to Envision, which meant that we were able to use the SHAP package with RAND's internal data to estimate SHAP values for each predictor. We then summarized these values across all observations with a linear regression to obtain a single impact estimate for each officer model. However, the SHAP package could not be used with our enlisted models on Envision because it is not fully compatible with the PySpark MLLib classifier that we use.[10] Consequently, instead of a SHAP value, we associated the predicted probability output of the model with each predictor's values through a similar linear regression. Although this has a similar interpretation (i.e., as this predictor's value increases by one unit, the average prediction increases or decreases by X units), it is less robust than our SHAP-based approach for officers because it does not account for potential interactions between predictors or isolate the contribution of a specific feature. If the usability of SHAP on Envision changes in the future, we recommend updating

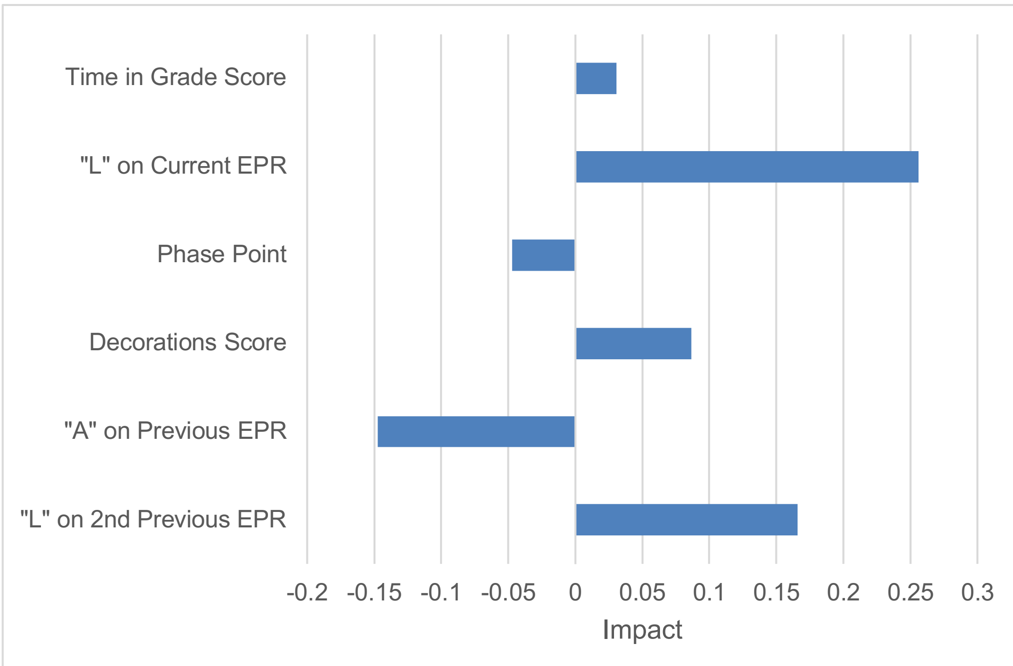
[10] More specifically, the SHAP package can work on Spark with the PySpark MLLib GBTClassifier function, but not if the model input data have predictors with categorical splits. For example, the core AFSC variable we include in the qualitative models is categorical. To include this in our models, it is necessary to reshape the variable from a column of text strings to an array of indicators with one for each unique core AFSC (a process known as *one-hot encoding*). In PySpark, this is called a *categorical split*. If the training data for a PySpark model included any variables with categorical splits, the functions of the SHAP package that we used for officers cannot be used, and alternative functions provided by the package (e.g., the model-agnostic KernelExplainer) either take prohibitively long to compute on Envision or do not support tree-based models at all.

the code that generates these impact figures to use SHAP values instead of the raw prediction output of the model.

Nonetheless, our impact estimates have intuitive implications that help contextualize the feature importances in Figures 3.1–3.6. In Figures 3.7–3.12, we present the impact estimates for the most important features by order of importance.

Figure 3.7 shows the impacts of the most important features for the E-7 quantitative model. Receiving L ratings on an EPR have the most positive impact on an officer's promotability, and receiving an A, a lower score than L, is the most negative indicator of promotion.

Figure 3.7. E-7 Quantitative Feature Impact



In both the E-8 and E-9 quantitative models, receiving a stratification is the most positive indicator of promotion, as seen in Figures 3.8 and 3.9. This finding supports stratification as a key accomplishment that improves an airman's probability of promotion. Although nonstratification features have very little impact in the E-8 model, time in grade has a positive impact in the E-9 model, indicating that having more experience at E-8 leads to an increased likelihood of being promoted to E-9.

Figure 3.8. E-8 Quantitative Feature Impact

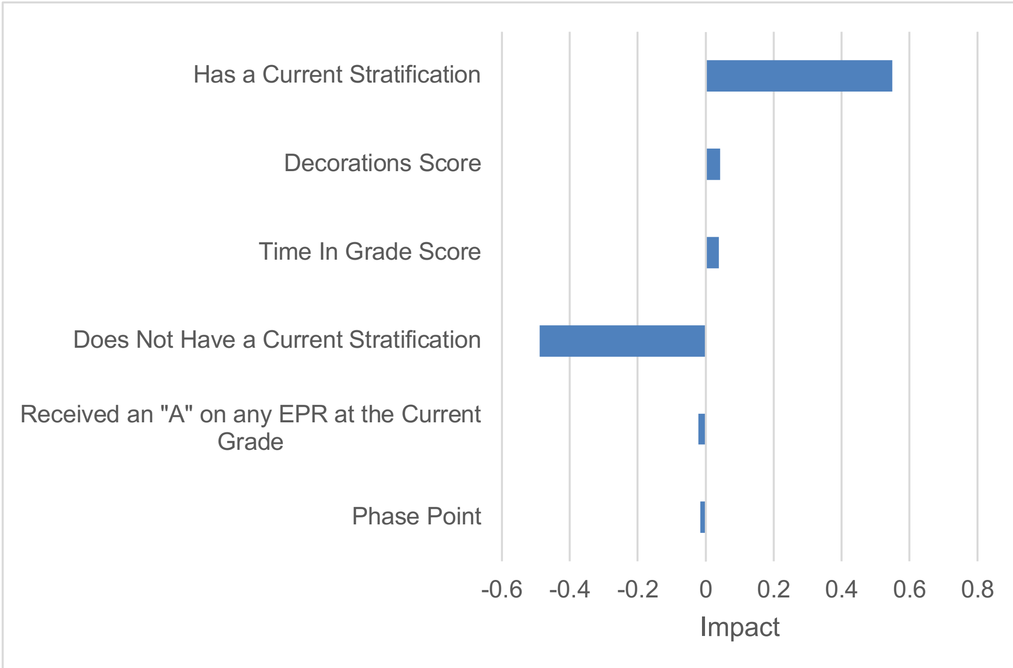
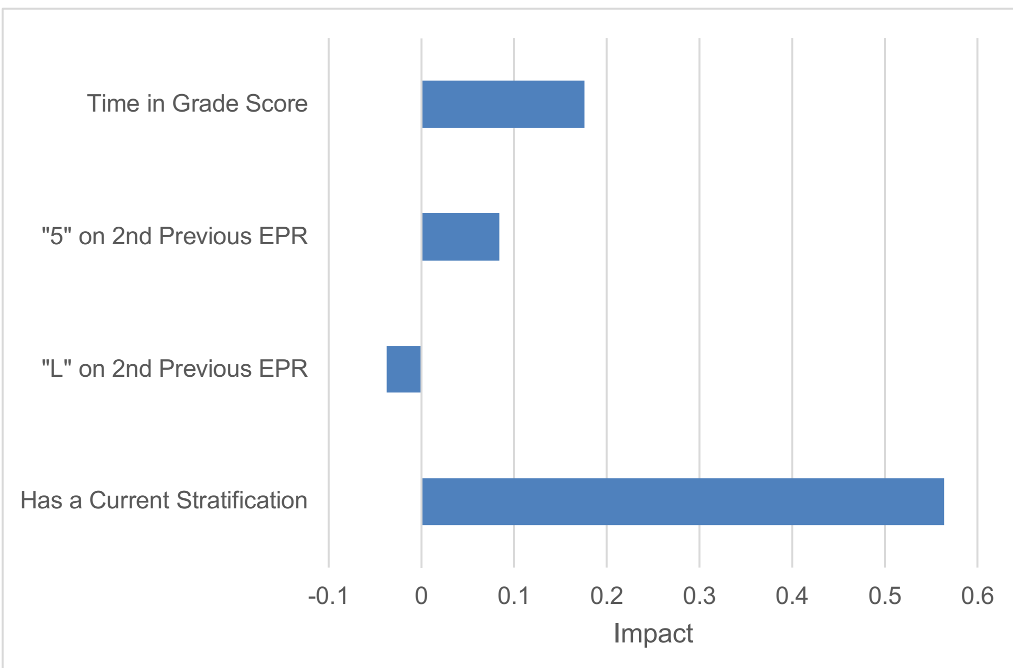


Figure 3.9. E-9 Quantitative Feature Impact



For promotion to E-7 (Figure 3.10), we see a similar pattern with regard to years in grade as we saw with field grade officers, with too much (four years) or too little (less than one year) time in grade making promotion less likely. In the absence of stratification, we also see that "awarded," which might serve a similar function by signaling the airman's position among their peers, has a moderate positive impact. We also see small positive impacts of the airman having contributions to certain functions ("operations," "training") highlighted in their EPR. For promotion to E-8 and E-9 (Figures 3.11 and 3.12), when stratifications become policy, we see an outsized impact for stratification-related features relative to others. Although the feature importance of "promote" was close behind the best stratification value in the E-9 model, the impact that we estimate for "promote" is substantially smaller than the impact of the stratification value. This indicates that, although the best stratification and the presence of a "promote" push line are both statistically predictive of selection for E-9, a high stratification has a stronger effect.

Figure 3.10. E-7 Qualitative Feature Impact

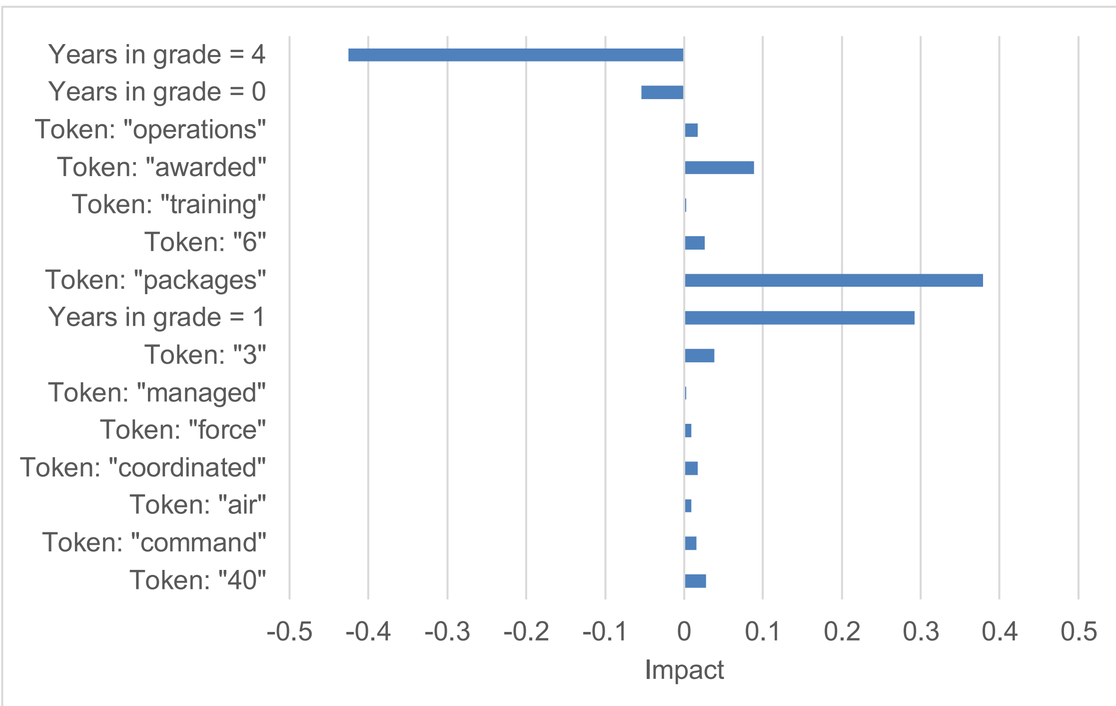


Figure 3.11. E-8 Qualitative Feature Impact

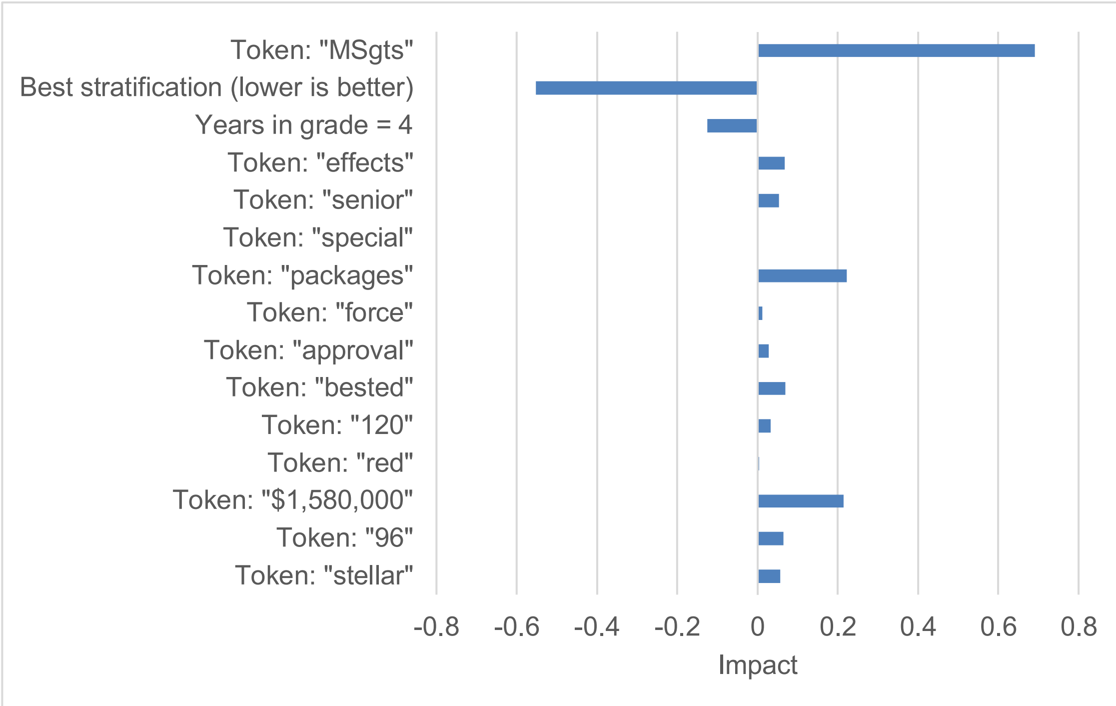
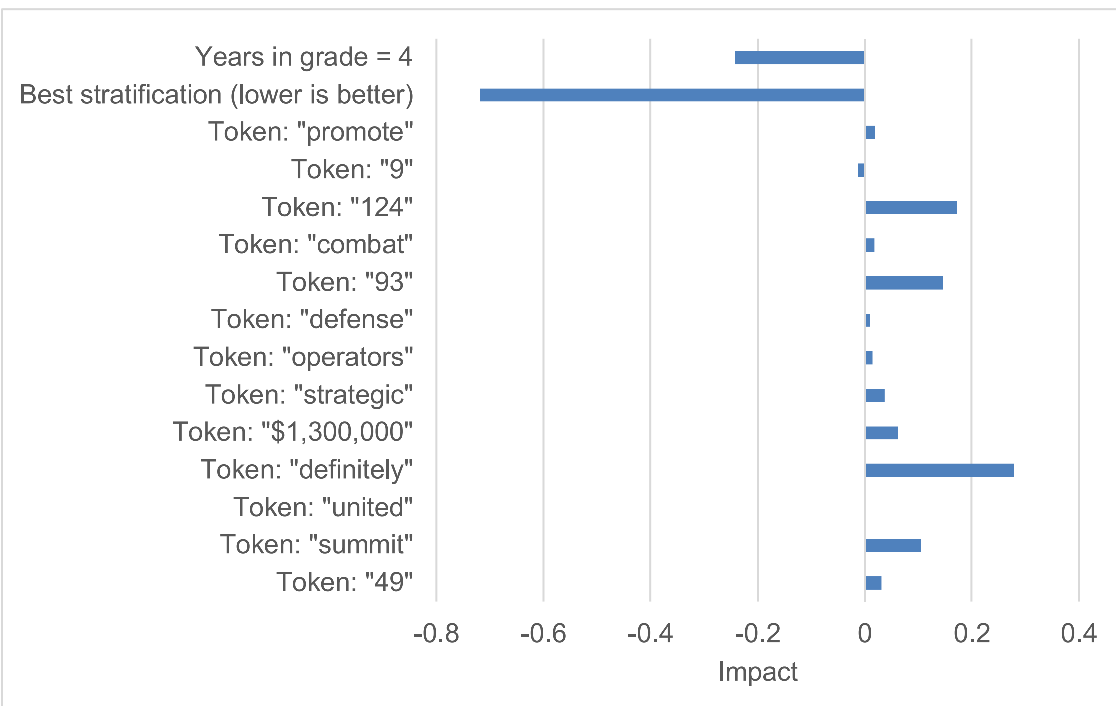


Figure 3.12. E-9 Qualitative Feature Impact



Chapter 4

Conclusion

Building on the prototype PReSS tool that PAF developed for AF/A1X in prior projects as a foundation, we built a suite of tools to help USAF talent management stakeholders analyze the career development trajectories of both officer and enlisted personnel. We summarize the key improvements in Table 4.1.

Table 4.1. Key Improvements in New PReSS from Original Prototype

Feature	Original Prototype	New PReSS
Predictors used	Qualitative OPR data only	Key features from MilPDS and qualitative OPR and OPB data
Prediction method	Logistic regression	GBDT classifiers
Evaluation metrics	Accuracy, AUC-ROC	Balanced accuracy, precision, and recall
Explanation methods	Extraction of snippets from text with largest impact on prediction	Feature importance, SHAP
Text representation (vectorizer)	Weighted counts of tokens	Feature hashing and numerical representation of stratifications
Text preprocessing	Some standardization of common features (stratifications, monetary amounts)	Standardization of common features and abbreviation expansion, spellchecking
Text source	All OPRs received in grade combined	Annual OPRs and OPBs processed individually
Platform	RAND internal server	Envision with initial research and development on RAND internal server

Unlike the prototype, which was implemented locally on RAND's computing infrastructure and therefore required the USAF to ask PAF analysts to rerun the model with new data, **our updated PReSS is implemented entirely on the Department of the Air Force's cloud-based analytic computing platform Envision.** As a result, PReSS is purpose-built for operationalization into USAF talent management processes because it already lives on, and therefore is assured to be compatible with, the primary analytic platform that the USAF uses for those processes internally.

Additionally, over the course of the project, **we worked closely with AF/A1XD data experts to establish new, repeatable pipelines to build the input data required for PReSS**, meaning that the entire stack from initial retrieval from MilPDS and ARMS-LC data connections through the final prediction output of the model is replicable. This improvement addresses a primary limitation of the original PReSS prototype, which was trained on a fixed snapshot of data and would need to be rebuilt from the ground up with new data inputs, especially if the structure of those inputs changed, as occurred recently with the transition from the OPR and EPR format to the OPB and EPB format. Although certain elements of the new PReSS are tailored specifically to the Envision platform, the core functions of the tool are written in PySpark and do not rely on any functionality exclusive to Envision. If either changes to the Envision platform or the adoption of a different cloud-based analytic platform require it in the future, PReSS can be transferred to any other platform that supports Python (nearly universal) and that is either built on a Spark cluster or offers the capability to create one locally.

The new PReSS was built to automate as much of the pipeline from initial data retrieval to final outputs as possible, but some elements might still need to be updated over time. Most notably, the dictionary that we built to replace abbreviations in performance reports with their full meaning, standardize their format, and correct typos and other issues that would introduce inconsistencies in the input data was developed using only the OPR data that RAND received for the original PReSS prototype project. Because we did not have analogous EPR data, this means that abbreviations and other potential sources of inconsistency specific to enlisted personnel data would not be represented in this dictionary and would therefore not be replaced. To limit the burden of updating it over time, **we saved the dictionary used to replace abbreviations and correct common errors as a simple data frame within Envision, and updating the dictionary requires only adding or modifying rows of this data frame**. The replacements made based on the dictionary use relatively complex regular expressions (RegEx), but the data frame backing the dictionary is simple text and does not require RegEx expertise to update. However, the issue of EPB data not having metadata for straightforward entity extraction will require more manual effort and either RegEx expertise or knowledge of PDF file structure to address. At the time of writing, we are aware of ongoing efforts within AF/A1XD to address this issue. When these efforts are complete, running the enlisted PReSS with EPB data will be achievable by swapping out the EPR data input at the start of the tool pipeline (for which EPB data are currently given blank values for the feature columns used by PReSS) with the new, more-complete input data.

Limitations of Envision as a Platform for PReSS

Although building the new PReSS in Envision has the crucial advantage of reducing barriers to operationalizing the tool in USAF management processes, we did face several limitations associated with the platform. We list some of the main limitations we faced in Table 4.2 and discuss in more detail the implications for the USAF's ongoing use of the tool in this section.

Table 4.2. Limitations Faced on Envision and Related Adaptations to PReSS

Limitation	Adaptation to PReSS
Limited ability to use supplemental packages to extend Python and PySpark functionality	Where possible, we manually built required packages from source on Envision; otherwise, we used less preferred alternatives where possible and omitted some features with no suitable alternatives.
Blocking resource limitations, particularly for functions that do not natively support Envision's Spark cluster	We replaced some functions with less resource-intensive alternatives to prevent out-of-memory (OOM) and other errors preventing the tool from running.
Deprecation of some Envision functions and lack of support for more-recent features of the underlying Palantir Foundry system	We worked exclusively in Code Repositories instead of unavailable newer functions better suited for rapid iteration in early phases of development.
Limited user permissions to customize environment to address incompatibilities or improve efficiency	We rewrote code where possible to work around incompatibilities, resulting in longer run time and/or reduced model performance.

The ability to use supplemental packages to extend Python and PySpark functionality is limited. Although Envision does offer a built-in repository that allows users to install commonly used packages directly into their Code Repositories without needing elevated user account privileges or administrator support, the universe of packages offered is a small subset of what is generally available in popular public repositories (e.g., conda-forge), and those that are available are frequently significantly out of date. The staleness of the available packages is particularly problematic in cases in which the versions available to install from Envision's built-in repository are no longer compatible with the latest version of Python. In our case, the most relevant limitation that this imposed on the project was that the packages that we used for predictive modeling with RAND's internal computing resources, XGBoost and LightGBM, were available only in outdated versions that did not support critical functionality required by PReSS. However, we faced a similar issue with several other less critical but still desired packages. In some cases, we were able to work around this limitation by installing a package in a Python environment outside Envision, uploading the source code from this installation to Envision, and repackaging that source code in a local repository in our Envision project's scope. This worked well for simple, self-contained "pure Python" packages, but it did not work for more-complex packages with additional dependencies (e.g., both XGBoost and LightGBM, for which the Python package is essentially just a wrapper around a C library that needs to be compiled separately). Early in the project, we engaged with Envision support to request the addition of a PDF-parsing package required by the original PReSS prototype, but they indicated that the addition of a custom package beyond those already available in the platform was not supported.

Functions that do not natively support Spark very frequently run into blocking resource limitations, requiring the use of alternatives that are not always equivalent. As a distributed computing platform, Envision is designed to run computing tasks that can be divided into smaller pieces, run on separate hardware nodes (Spark *executors*), and then have the results sent back to the main device (the Spark *driver*). Because of this design principle, the resources available to the main device are purposefully limited, with the default configuration on Envision providing only 3 gigabytes

(GB) of memory to the Spark driver. Consequently, computing tasks that require a larger amount of memory (especially with large datasets, such as the MilPDS and ARMS-LC extracts that PReSS uses) can be performed only if they support distribution over a Spark cluster. Attempting to perform any task that does not support this has a very high likelihood of encountering a fatal OOM error. This limits the universe of packages that one can use on Envision, particularly for memory-intensive applications, such as machine learning. Additionally, even some tasks that do support Spark still culminate in an OOM error if any component of the task tries to pull too much data from the executors to the driver, which is sometimes difficult to predict. In general, we found that it was best to rely on functions in the PySpark package wherever possible, even when functions from third-party packages claim to support Spark. The PySpark library is fortunately quite extensive and builds in versions of several functions from commonly used packages (e.g., scikit-learn) that are explicitly designed to run on Spark. These include a suite of machine learning functions called MLLib, which contains several of the functions that we use in PReSS, including the GBDT classifier, the hashing function used to generate numerical representations of performance report text, and other model input data processing functions. However, the MLLib implementation of some functions has meaningful disadvantages over the third-party reference function that it replaces. For example, the GBDT classifier that we use for PReSS on Envision uses an older, less performant gradient boosting method than the method used in XGBoost and LightGBM and lacks some other features of those packages that would have simplified certain components of PReSS, particularly the estimation of feature impacts.

At the time of writing this report, Envision does not provide all of the functionality in the latest versions of the Palantir Foundry platform on which it is based, and some included functions are no longer under active development. Most notably, Foundry features intended to simplify the integration of analyst workflows from other platforms, such as Code Workspaces, and integration with the popular VSCode-integrated development environment are not implemented in Envision. Additionally, some functions that are available in Envision, such as Code Workbooks (the predecessor to the more recent Workspaces feature), are marked "legacy" or "sunset" in the Foundry documentation and therefore will not receive future updates. For Code Workbooks, the most relevant limitation is that the latest version of Python is not supported, which, in turn, means that the latest version of Python packages that depend on a newer version of Python cannot be used. Because Code Workspaces are not implemented on Envision, this means that there is no up-to-date, actively developed capability to run an interactive Python session. Instead, work on Envision to implement PReSS and other analytic tools like it must be done within a Code Repository, which is significantly less flexible and responsive than a Code Workspace would be. Because testing and iteration in a Code Repository requires building complete Python scripts, development of such capabilities as PReSS likely will take longer and might require more computing resources in Envision than in more-flexible platforms that allow for real-time iteration and piecemeal code execution. Additionally, although the underlying PySpark code of such a capability as PReSS is readily portable to other platforms, some Foundry capabilities that could be built on top of PReSS, such as model logging, are specific to Foundry systems and, therefore, would not transfer from Envision to another platform.

Although Envision offers the ability to import custom Spark configuration profiles, many profiles can be applied only by platform administrators. In some sense, this is a feature and not a bug of the platform, because preventing users from applying certain custom Spark profiles prevents excess resource usage and limits the cost of the platform. However, the defaults available to Envision users can be restrictive, particularly in terms of compute resources. By default, the custom Spark profiles available to Envision users without administrator intervention allow only up to eight executors, two central processing unit (CPU) cores per executor, 13 GB of memory for the driver, and 27 GB of memory per executor. Although this is more than enough for simpler data processing operations, such as defining new variables based on source data or calculating summary statistics, compute-hungry machine learning processes and more-complex routines to process large datasets might run into resource constraints.

Working around resource constraints for those without administrator privileges in Envision detracts from performance, making routines take longer to run and, in some cases, impossible to run. For PReSS on Envision, the main consequence of this limitation is that some steps of the pipeline from raw input data to final predictions take significantly longer to run on Envision than they did on RAND's internal servers, where higher amounts of available memory and CPU cores allowed for greater leverage of parallel processing to speed up complex calculations. As of this writing, with between 10,000 and 50,000 rows of data being used to train and validate the models, the entire PReSS pipeline takes about 12 hours to run from start to finish, not including the initial sourcing of the data from MilPDS and ARMS-LC databases that was done for us by AF/A1XD.

Although the lengthy run time for PReSS on Envision might not be a challenge if PReSS is run only every few months (because it takes time for a meaningful number of new performance reports, personnel records, and board selections to become available), slower run times do limit the scalability of the tool. Currently, our data for officers on Envision include only a selection of seven core AFSCs for which we were able to speak with USAF talent management stakeholders early in the project to assess their needs for PReSS capabilities and inform the design of the tool. Expanding this dataset to the full universe of AFSCs will increase the time required to run the models, as will natural growth in the data over time for the AFSCs already included. The two key components of the PReSS pipeline that contribute the most to that run time—the replacement of abbreviations with their full definitions and the actual fitting of the models—both leverage parallel processing and, therefore, would likely complete much faster if more executors and/or CPU cores were available to run the models. Because of time limitations, we did not pursue permission to import more-expansive Spark profiles from Envision support. However, AF/A1X could pursue such permission if the run time of the tool becomes a concern in the future.

Recommendations Based on Insights from PReSS

Although the primary objectives of this project were to update and expand the PReSS tool for multiple levels of USAF talent management and help facilitate AF/A1X's operationalization of the updated tool, our findings also contribute to a broader understanding of officer and enlisted personnel career development dynamics at the grades in our scope. We summarize the key insights and associated recommendations in Table 4.3 and follow with a more detailed discussion.

Table 4.3. Key Insights and Recommendations from the New PReSS for Officer and Enlisted Personnel Career Development

Insight	Recommendation
Given enough information, an effective method, and sufficient computing resources, it is possible to quantify whether senior enlisted airmen and field grade officers have fulfilled the expectations for career development milestones, such as promotion, with high accuracy.	Operationalize PReSS for field grade officers and senior enlisted personnel and explore possible extensions of the tool using its outputs as a foundation for related capabilities, such as providing automated feedback to personnel.
The features that correlate with promotion over the course of an airman's career include some that are always relevant and others that become more or less relevant with increased seniority.	Explore ways to analyze and use PReSS outputs to improve understanding and transparency of the key contributors to career outcomes throughout the course of an airman's career, for both talent management stakeholders and personnel.
Although it is possible to predict career outcomes from just MilPDS quantitative features or just performance report qualitative features, the most-performant models use a combination of both.	When operationalizing PReSS, emphasize the use and maintenance of the combined qualitative and quantitative Model 3, and ensure that both types of data remain up to date through robust and regular extraction pipelines.
The USAF already has the key information required for such a capability as PReSS to be operationalized on Envision, but some additional data that could further improve the models or expand PReSS's use cases are not yet usable.	Pursue opportunities to expand the scope of data available to PReSS, such as leveraging existing data connections that have not yet been used (e.g., MyVector) and collaborating with other USAF functions to establish new data connections where possible and relevant.

Given enough information, an effective method, and sufficient computing resources, it is possible to quantify whether senior enlisted airmen and field grade officers have fulfilled the expectations for career development milestones, such as promotion, with high accuracy. The original PReSS prototype was an initial demonstration of this capability for O-4 and O-5 officers; it showed a relatively high degree of accuracy in predicting selections for promotion among officers at these pay grades using information from only OPRs. We improved this capability by overhauling the OPR processing pipeline to further standardize input data; correct (to the extent possible) typos and misspellings; and incorporate contextual features accompanying the performance report narratives, such as rater pay grade(s), ratee AFSC, and a continuous numerical representation of stratifications embedded in the report. Through these enhancements, we were able to improve model performance and achieve a numerically similar level of accuracy with both a more-challenging prediction task (using a single OPR to predict rather than the officer's full history, limiting the amount of data available to the classifier) and more-stringent accuracy metrics (balanced accuracy, precision, and recall to ensure that the model's measured accuracy is not artificially inflated when predicting a binary outcome with significantly fewer than one-half of positive cases). We further built on these improvements by incorporating personnel record data from MilPDS to obtain a more-complete picture of an officer's duty history and other career elements known to be relevant but that might not be represented in a performance report. Finally, we expanded this methodology to analyze promotion to three senior enlisted pay grades (E-7, E-8, and E-9), demonstrating that the method that we developed for officers can also achieve a similar level of accuracy and actionable information when applied to senior enlisted personnel.

The features that correlate with promotion over the course of an airman's career include some that are always relevant and others that become more or less relevant with increased seniority. Across all of our qualitative models, we consistently found that certain descriptors reflective of the ratee's leadership skills, accomplishments, scope of experience and responsibilities, and value added were associated with higher likelihoods of promotion. We also found consistent patterns with respect to time in grade, with each grade level having a sweet spot where promotion is most likely, after which further time in the same grade begins to associate negatively with the likelihood of promotion. However, we also found intuitive nuances across pay grades. For enlisted personnel, because stratifications are not a norm in E-6 evaluation, we found that promotion to E-7 is associated with other qualitative factors. However, for promotion to E-8 and E-9, we found that stratifications become immediately relevant to a similar level for senior enlisted personnel as for officers as soon as they are introduced into the enlisted evaluation process. We also found that, although the push lines that raters use to tell the board exactly what they think should happen to their ratees are relevant at all career stages, the magnitude of their relevance does appear to vary by level: Promotion to O-5 and E-9 are strongly associated with push line components (e.g., "promote," "send to SDE"), but promotion to O-6, E-7, and E-8 does not appear to emphasize these features to the same degree. Particularly for elements that become more or less relevant as careers progress, PReSS offers an opportunity to increase transparency of the key factors contributing to career outcomes at all levels with information backed by data. In turn, future analyses of the underlying reasons for these differences and the extent to which they generalize across career fields could be built on PReSS outputs.

Although it is possible to predict career outcomes from just MilPDS quantitative features or just performance report qualitative features, the most-performant models use a combination of both. The extent to which components of an airman's record actually influence their likelihood of promotion was an open question at the outset of this project. Conversations early in the project with key USAF talent management stakeholders suggested a variety of perceptions about the relevance of narrative performance report information, with some seeing it as critical and others seeing the narrative descriptions as significantly less relevant than the duty history from MilPDS and the stratifications on the performance report alone. The original PReSS prototype demonstrated that OPR information could be used to accurately predict career milestones. However, the question of whether its predictive accuracy reflected the unique value of OPR information relative to quantitative personnel records or instead reflected only the OPR's ability to proxy for duty history and other record features remained open. In the new PReSS, we explicitly tested three approaches to predicting career outcomes: one using only personnel records, another using only performance reports, and a third using both. Our findings suggest that, although models using one or the other are still quite accurate, the highest performance is typically achieved by the models using both types of information, suggesting that each information source provides at least some unique predictive information. This is particularly true if the objective is to balance precision and recall, because we found that performance report–only models tend to have better recall (i.e., using information from performance reports can enable the model to identify the top prospects, but it might struggle to discriminate between true top prospects and others with ostensibly similar records that do not culminate in promotion as often), and personnel record–only models tend have better precision (i.e., a top prospect identified by this model is very likely to actually be a top prospect, but some others will likely be missed). The models combining the two types of information into a single model achieve the greatest balance between these two dimensions of accuracy, and, therefore, we recommend focusing on these models when operationalizing and maintaining PReSS.

The USAF already has the key information required for such a capability as PReSS to be operationalized on Envision, but some additional data that could further improve the models or expand PReSS's use cases are not yet usable. Although our training data included most of the highest-priority measures for predicting career outcomes (e.g., IDE or SDE completion, squadron commander experience, time in grade), there are some other data sources that we were not able to use for the new PReSS that might further improve model performance. In particular, we offer three insights:

- The original PReSS prototype used promotion board files, which recorded the exact board date on which each officer was selected or not selected for promotion. With MilPDS data alone, we cannot determine the extent to which measurement error in our inferred selections reduces the accuracy of our models.
- There are other pieces of information highly likely to be relevant to promotion outcomes that we were not able to obtain on Envision, including the contents of promotion recommendation forms and promotion line numbers.
- Other qualitative information outside a performance report might be available through such platforms as MyVector, which we were not able to access. It is possible that information in MyVector that reflects voluntary participation in certain career development activities or signaling opportunities (e.g., statements of interest, talent marketplace engagement) would help distinguish between cases in which an individual is not promoted because they are not qualified and cases in which an individual is voluntarily not seeking a promotion.[11]

Additional Opportunities for Future Work and Further Development

The new PReSS for officer and enlisted personnel is highly accurate and capable of identifying intuitive key predictors of career outcomes and, therefore, can be operationalized for that purpose. However, other use cases that subject-matter experts identified during discussions early in the project would require further development of PReSS and complementary tools to support them. For example, subject-matter experts expressed interest in the use of LLMs to simplify certain labor-intensive talent management processes by summarizing the key components of candidate records, extracting certain

---

[11] The ability to distinguish between these cases is not necessarily critical for officer and enlisted personnel because of the mandatory, regularly scheduled board structure that ensures that all personnel eligible for promotion receive an evaluation regularly. However, when exploring the possibility of expanding the PReSS methodology to the civilian workforce, we determined that there was not sufficient information on this latent motivation factor to accurately predict promotions because the mechanisms that lead to civilian promotion are voluntary. Information in MyVector for civilian employees might be especially valuable to help distinguish between those employees who are actively seeking advancement opportunities and those who are not by choice.

features of interest not readily available in an existing spreadsheet, or conducting targeted searches of candidate records to identify personnel who have certain needed qualifications or qualities. To the extent that these components are reflected in either MilPDS or performance report data, the new PReSS can contribute to these use cases, but it is not directly designed to support them. However, the outputs created by PReSS could be used in downstream applications to support these use cases more directly. For example, PReSS predictions could be used to fine-tune an LLM to understand the determinants of promotion based on the classifier's predictions, and then that LLM could subsequently be used by talent management stakeholders or even personnel themselves to extract information in a more-interactive and less code-dependent way.

In addition to fine-tuning an LLM, PReSS could be used to formulate direct feedback mechanisms across personnel groups. Talent management stakeholders whom we engaged with throughout the project indicated that some career fields have established processes to provide direct feedback to officers, including letters from DTs that provide the officer's developmental education panel score relative to the distribution of their peers and bubble charts that visually summarize key attributes of an officer's records, particularly for different types of experience. However, to our knowledge, such an effort has not been scaled up across the enterprise and remains limited to a select few AFSCs. By providing a way to efficiently evaluate officer and enlisted performance records in bulk and calculate simple career development score statistics that could help personnel understand whether they are on track to meet their career goals, PReSS could enable future capabilities to increase transparency of talent management and performance evaluation processes.

Abbreviations

AF/A1	Deputy Chief of Staff for Manpower, Personnel, and Services
AF/A1X	Plans and Integration, Deputy Chief of Staff for Manpower, Personnel, and Services
AF/A1XD	Human Resources, Analytics, and Decision Support Division, Plans and Integration, Deputy Chief of Staff for Manpower, Personnel, and Services
AFSC	Air Force Specialty Code
APZ	above the promotion zone
ARMS-LC	Automated Record Management System—Legacy Conversion
AUC-ROC	area under the receiver operating characteristic curve
BPZ	below the promotion zone
CPU	central processing unit
CYOS	commissioned years of service
DAF	Department of the Air Force
DEVCAT	developmental category
DT	development team
EPB	enlisted performance brief
EPR	enlisted performance report
GBDT	gradient-boosted decision tree
IDE	intermediate developmental education
IPZ	in the promotion zone
LLM	large language model
LTC	lieutenant colonel (O-5)
MAJ	major (O-4)
MilPDS	Military Personnel Data System
OCR	optical character recognition
OOM	out of memory
OPB	officer performance brief
OPR	officer performance report
PAF	RAND Project AIR FORCE
PME	professional military education
PReSS	Personnel Records Scoring System
SDE	senior developmental education
SHAP	SHapley Additive exPlanations
TPR	true positive rate
USAF	U.S. Air Force

WAPS	Weighted Airman Promotion System
XGBoost	extreme gradient boosting

References

Bergstra, James, Brent Komer, Chris Eliasmith, Dan Yamins, and David D. Cox, "Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization," *Computational Science & Discovery*, Vol. 8, No. 1, January 2015.

Chen, Tianqi, and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, August 2016.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates, Inc., December 2017.

Lim, Nelson, Sarah Nicole Kosic, Eric I. Salehi, and Helin Hernandez, *A Brief Guide to Workforce Statistics: For Airmen and Guardians*, RAND Corporation, TL-A3179-3, 2025. As of July 25, 2025: https://www.rand.org/pubs/tools/TLA3179-3.html

Lundberg, Scott M., and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates, Inc., December 2017.

Robson, Sean, Maria C. Lytell, Matthew Walsh, Kimberly Curry Hall, Kirsten M. Keller, Vikram Kilambi, Joshua Snoke, Jonathan W. Welburn, Patrick S. Roberts, Owen Hall, and Louis T. Mariano, *U.S. Air Force Enlisted Classification and Reclassification: Potential Improvements Using Machine Learning and Optimization Models*, RAND Corporation, RR-A284-1, 2022. As of April 27, 2025: https://www.rand.org/pubs/research_reports/RRA284-1.html

Romer, Robert, Kirsten M. Keller, and Kelly Hyde, *An Initial Assessment of Opportunities for Tech-Enabled Civilian Talent Management in the Department of the Air Force*, RAND Corporation, RR-A2996-2, forthcoming.

Schulker, David, Lisa M. Harrington, Matthew Walsh, Sandra Kay Evans, Irineo Cabreros, Dana Udwin, Anthony Lawrence, Christopher E. Maerzluft, and Claude Messan Setodji, *Developing an Air Force Retention Early Warning System: Concept and Initial Prototype*, RAND Corporation, RR-A545-1, 2021. As of April 27, 2025: https://www.rand.org/pubs/research_reports/RRA545-1.html

Schulker, David, Nelson Lim, Luke J. Matthews, Geoffrey E. Grimm, Anthony Lawrence, and Perry Shameem Firoz, *Can Artificial Intelligence Help Improve Air Force Talent Management? An Exploratory Application*, RAND Corporation, RR-A812-1, 2021. As of April 27, 2025: https://www.rand.org/pubs/research_reports/RRA812-1.html

Schulker, David, Matthew Walsh, Nelson Lim, and Ajay K. Kochhar, *How the U.S. Air Force Can Incorporate New Data Technologies into Its Talent Management System: Framework and Use Cases for Technology-Enabled Talent Management*, RAND Corporation, RR-A1198-2, 2022. As of April 27, 2025: https://www.rand.org/pubs/research_reports/RRA1198-2.html

Schulker, David, Joshua Williams, Cheryl K. Montemayor, Li Ang Zhang, and Matthew Walsh, *The Personnel Records Scoring System*: Volume 3, *A Methodology for Designing Tools to Support Air Force Human Resources Decisionmaking*, RAND Corporation, RR-A1745-3, 2024. As of March 28, 2025: https://www.rand.org/pubs/research_reports/RRA1745-3.html

SHAP, "An Introduction to Explainable AI with Shapley Values," webpage, undated. As of July 28, 2025: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

Snyder, Don, *Funding Technology-Related Business Initiatives in the Department of the Air Force*, RAND Corporation, RR-A1198-3, 2022. As of August 6, 2025: https://www.rand.org/pubs/research_reports/RRA1198-3.html

Wang, Chi, Qingyun Wu, Markus Weimer, and Erkang Zhu, "FLAML: A Fast and Lightweight AutoML Library," *Proceedings of Machine Learning and Systems*, Vol. 3, 2021.

Wolpert, David H., "Stacked Generalization," *Neural Networks*, Vol. 5, No. 2, 1992.

Yeung, Douglas, Elicia M. John, Jeannette Gaudry Haynie, James Ryseff, Bonnie L. Triezenberg, and Nelson Lim, *Implementing Technology-Enabled Human Resources Capabilities in the U.S. Air Force: Insight from the Private Sector and Military Services*, RAND Corporation, RR-A1198-1, 2022. As of April 27, 2025: https://www.rand.org/pubs/research_reports/RRA1198-1.html

RAND PROJECT AIR FORCE

The authors expanded the Personnel Records Scoring System (PReSS) prototype, developed by RAND, to build a suite of tools to help U.S. Air Force (USAF) talent management stakeholders analyze the career development trajectories of both officer and enlisted personnel.

PReSS is designed to improve the efficiency and objectivity of assessing whether personnel have fulfilled expectations for career progression by identifying promising candidates and providing an objective benchmark for comparing the actual decisions of raters and developmental team members. The system is a decision support tool that helps human resource personnel extract meaningful insights from unstructured text in officer and enlisted performance reports and structured administrative recordkeeping data.

The updated PReSS tool achieves higher accuracy based on more-stringent metrics, adapts more readily to potential changes in data over time, and provides deeper insights into the reasons behind the model's predictions.

$39.00

ISBN-10 1-9774-1570-9
ISBN-13 978-1-9774-1570-7

53900

9 781977 415707

www.rand.org

RR-A2996-1