Research Report

DAVID SCHULKER, JOSHUA WILLIAMS, CHERYL K. MONTEMAYOR, LI ANG ZHANG, MATTHEW WALSH

The Personnel Records Scoring System

Volume 3, A Methodology for Designing Tools to Support Air Force Human Resources Decisionmaking

The U.S. Department of the Air Force (DAF) has begun to develop and field artificial intelligence (AI) and machine learning (ML) systems for myriad mission areas and support functions, including human resource management (HRM). This report describes an ML decision-support tool to summarize the information in officer performance reports (OPRs) and other narrative-style documents to help the HRM system make personnel decisions more effectively, more efficiently, and in better alignment with the DAF's strategic goals.

KEY FINDINGS

- Department of the Air Force analysts can rapidly develop simple models relating key text in officer records to past decisions. The most-accessible approaches break the text into individual terms, index the records according to which terms they contain, fit a predictive model of the past decisions, and then create decision inputs from the models. We demonstrate these steps through our development process for PReSS.
- The constrained language used in officer performance reports makes them amenable to natural language processing approaches, as shown by the fact that simple models with minimal preprocessing and tuning achieved high levels of accuracy.
- As compared with state-of-the-art machine learning approaches (i.e., deep learning), simple linear models based on the presence or absence of key terms achieve similar levels of predictive performance but have the advantage of being inherently interpretable.
- Key words and phrases that models base predictions on coincide with statements recognizable to expert raters.

Background

The latest data from the *McKinsey Global Survey on Artificial Intelligence* show that private-sector companies have continued the march toward greater adoption of AI, especially for optimizing services or enhancing product offerings (Chui et al., 2021). However, the same survey shows that AI adoption in the field of human resources (HR) is still relatively rare. Further, the percentage of companies that use AI to optimize talent management processes, such as those associated with recruiting or retention, *declined* from 10 percent to 8 percent between the 2020 and

2021 surveys, while those using AI for performance management increased marginally, from 7 percent to 8 percent (Chui et al., 2021; Balakrishnan et al., 2020). Still, those that had adopted AI for HRM continued to report significant cost decreases and revenue increases from adoption.

The relatively low uptake of AI in HRM functions (compared with, say, service operations at 27 percent) could relate to several challenges that are particularly acute in the HRM domain. Work-related attributes and job performance outcomes are complex, and they are difficult to define and objectively measure. AI adoption in HRM can also be stymied by data constraints, ethical or legal unknowns that are challenging to work through, and employee responses to these new systems that have significant impacts on their careers (Tambe, Cappelli, and Yakubovich, 2019).

Despite these challenges, there is a potential opportunity for the DAF to join the minority share of adopters who use AI to improve HRM. This report aims to help demonstrate AI functionality in the HRM arena by proposing a methodology for a class of HRM decision-support tools: the Personnel Records Scoring System (PReSS). PReSS draws on HR records and applies natural language processing (NLP) techniques to produce inputs that support and potentially improve HRM decisions.[1]

This report is one in a series (Table 1) intended to help policymakers address the challenges unique to HRM and move forward with adoption, as part of broader strategies in the U.S. Department of Defense (DoD) and DAF to use "data at speed and scale for operational advantage and increased efficiency" (DoD, 2020, p. 2; see also DAF, 2021). Separate volumes in the series address how policymakers should manage the portfolio of potential use cases and how to evaluate use cases in HRM for safety and equity.

The Untapped Potential of Textual Records for Decisionmaking

The goal of the DAF HRM system is to produce and maintain a workforce that is ready to accomplish the DAF's mission to "fly, fight, and win . . . in air, space, and cyberspace" and that embodies the DAF's core values of integrity, service, and excellence (DAF, 2014; this Air Force Instruction was updated in August 2023, after this report was written). The DAF executes HRM through interdependent functions that determine whom to bring in and how to develop, utilize, advance, and retain them. Layers of policy structure, beginning with Title 10 of the U.S. Code and moving down through subordinate echelons of DoD and DAF, specify the essential form of the

TABLE 1
Outline of Report Series

Volume Number	Report Title	Report Purpose
1	*Leveraging Machine Learning to Improve Human Resource Management: Volume 1, Key Findings and Recommendations for Policymakers* (Schulker, Walsh, et al., 2024)	Overview for senior leaders
2	*Machine Learning in Air Force Human Resource Management: Volume 2, A Framework for Vetting Use Cases with Example Applications* (Walsh et al., 2024)	Framework for how to prioritize ML projects
3	*The Personnel Records Scoring System: Volume 3, A Methodology for Designing Tools to Support Air Force Human Resources Decisionmaking* (Schulker, Williams, et al., 2024)	Technical report on scoring officer records
4	*Safe Use of Machine Learning for Air Force Human Resource Management: Volume 4, Evaluation Framework and Use Cases* (Snoke et al., 2024)	Case study approach to ensure safety of ML systems
5	*Machine Learning–Enabled Recommendations for the Air Force Officer Assignment System: Volume 5* (Calkins et al., 2024)	ML system to inform officer assignments

NOTE: Current report is highlighted.

system and the freedom of action available to DAF HRM policymakers and managers.

At their atomic level, though, HRM outcomes are most often the result of individual decisionmakers, such as those reviewing records of work-related attributes (i.e., knowledge, skills, abilities, and other attributes, or KSAOs) and then making career-altering decisions that best meet strategic HRM goals. Occasionally, these records include quantitative elements, such as test scores. But, because KSAOs and work performance can be difficult to define and quantify, HR records often default to semistructured lists of experiences and narrative descriptions written by supervisors or instructors.

Compared with data that label and track officers for HRM processing (e.g., alpha-numeric codes for career fields and positions), textual records contain rich and specific information on officers. For example, each officer receives an annual OPR with detailed descriptions of his/her duties, most-significant accomplishments, and recommendations for future jobs and developmental opportunities (Schulker et al., 2021). The flexibility of the format enables far richer characterizations of officers and their KSAOs, but this flexibility also makes it difficult to use ML to extract this information.

Thus, many HRM decisions, particularly those affecting officers, call for a subjective process in which a senior officer or a panel of experienced personnel reviews the records and issues a verdict. Panel reviews determine whether officers can reclassify after being eliminated from training and whom to select for developmental education, special assignment, command, and promotion opportunities. Panels determine whom to retain when the DAF must manage force numbers by involuntarily separating members or forcing them to retire early. Panels determine which officers receive prestigious awards and key markers of performance.[2] Panels review officer performance histories and make future assignment recommendations, and assignment teams attempt to match officers with available jobs to facilitate DAF strategic objectives. In sum, statute sets the playing field, but the long-run HR outcomes for any given member are driven primarily by a series of subjective reviews of unstructured HR records at select career milestones.

Anytime a decisionmaker reviews a record and provides input into a decision, they face limitations. The decisionmaker's knowledge and experience are imperfect, they are susceptible to biases, and they have limited time and energy to spend on processing information and making each decision. These limitations create space for ML-based decision inputs to improve the effectiveness or efficiency of the decisionmaking process.

How Decision-Support Tools Can Enhance HRM Decisions

There are many potential ways to incorporate ML into decisionmaking. To illustrate some of these, consider the following implementation designs for leveraging ML in a panel review of records, such as a board process. As discussed in greater detail in another report in this series (Snoke et al., 2024), each of the following implementation designs satisfies different business objectives, and each has different safety and equity implications:

- **Decide.** The ML system automatically generates a decision.
- **Recommend.** The ML system provides recommendations that affect how human raters review the record and reach decisions.
- **Score.** The ML system generates a score representing the degree to which the individual conforms to the decision criteria.
- **Summarize.** The ML system generates a summary for human raters or candidates containing the most-relevant information to the HRM decision.
- **Audit.** The ML system replicates the decisions after the fact to flag unusual or noteworthy cases for further examination.

Implementing each of these designs, regardless of the HRM decision in view, requires a set of rules for converting the records into information that affects the decision process. For example, though there is no NLP involved in the enlisted quarterly assignment process, it is still a prime case of an HRM process that already uses the *decide* implementation of ML decision support. A series of prespecified rules converts information in eligible movers' HR records, such as their preferences, occupation and skill codes, and years of experience, into a final match of personnel to assignments (DAF, 2020). The key contribution of PReSS is that it enables HRM decisionmakers to use unstructured text in the same way they use other personnel attributes in existing HRM processes such as enlisted assignments. The primary novelty of PReSS compared with these other processes is that it arrives at the decision rules through NLP and ML techniques, as we describe in the following section.

Personnel Records Scoring System Conceptual Overview

PReSS draws on NLP techniques because HRM decisions often involve a detailed review of unstructured textual information. In this proof of concept, PReSS focuses specifically on applying NLP to score officer performance narratives on a spectrum from low- to high-performing. Such a system has immediate value for the many decisions that include performance history as a key input. Additionally, the ability to quantify aspects of these narratives could also yield a better understanding of the conditions that lead to higher or lower levels of performance, which would have much wider applicability outside the direct context of a selection board or panel. Though this focus on performance narrows the initial scope of PReSS, the rest of this report will discuss how the same techniques and steps would apply to other HRM decisions with non–performance-related inputs and decision goals.

There are two general classes of methods for creating systems for scoring unstructured text: Human experts can create the scoring rules, or ML methods can derive them from historical board outcomes generated by human experts. Researchers have used human ratings of the positivity or negativity of words to develop high-performing, rule-based NLP systems for scoring text (Hutto and Gilbert, 2014), and these methods are also applicable to scoring officer performance narratives. The primary advantages of handcrafted, rule-based systems are that they are explainable and that they can work in situations where data are limited. However, these systems can be costly to create and maintain, and they depend on human experts articulating a stable set of rules that can be accurately applied to future cases.

The PReSS methodology uses an ML approach known as *supervised ML*, which bypasses the human rating process by using an ML model that learns rules from example pairs of inputs and outputs.[3] The primary advantage of ML approaches is that they automate the creation of the scoring rules, which is useful when the exact set of rules is not known and would be labor-intensive to create and maintain.

Given that both the human and ML approaches have advantages and drawbacks, HRM decisionmakers and analysts should determine the appropriate method on a case-by-case basis. The ML approach is the logical choice for PReSS, because rich historical example data are available, and because the goal of PReSS is to illustrate the general applicability of ML to augment subjective decisions of HRM personnel based on textual records, necessitating a method that is not case-specific.

Table 2 describes some of the ways that a tool like PReSS could be used to augment board processes. Only one of the ML Implementation Designs, *decide*, seeks to fully automate decisions. The remaining four designs provide different inputs to board members and at different points in the scoring process. Thus, some of the implementation designs would have relatively greater influence on board outcomes (e.g., *recommend* and *score*), and others would have relatively less influence on board outcomes (e.g., *summarize* and *audit*).

As described in another report in this series (Snoke et al., 2024), each ML Implementation Design contained in Table 2 meets one or more HRM objectives: reduce workload, improve human decisionmaking, standardize process, advance DAF priorities, increase transparency, and provide feedback. Even so, prior to applying ML in the context of board processes, the U.S. Air Force must ensure that the system meets standards for accuracy, fairness, and explainability.

Scope of Applications

In this report, we focus on two board processes. The first is officer promotion boards. These boards meet to determine which officers will be promoted to the

TABLE 2

How PReSS Relates to ML Implementation Designs

ML Implementation Design	PReSS Contribution to Design
Decide	HRM personnel could establish a performance cutoff and use predicted performance levels to select a subset of officers for an opportunity (or as a first stage in a multistage selection process).
Recommend	HRM personnel could use predictions to make recommendations to decisionmakers, to recommend either a final decision or a level of scrutiny (e.g., regarding someone in the “gray zone” close to the selection cutoff).
Score	HRM personnel could include the model score alongside the scores of panel members when calculating the total score for a record.
Summarize	Panel members could use the PReSS General Performance Summary, or a tailored version of it, in the scoring process. HRM personnel could use the General Performance Summary when providing feedback to nonselected members.
Audit	Performance-based HRM actions that differ substantially from predicted performance levels could be flagged for further review and refinement of selection processes.

ranks of O-4 and O-5. The purpose of promotion is to “select officers through a fair and competitive selection process that advances the best qualified officers to positions of increased responsibility and authority” (DAF, 2023, p. 12). Outcomes of the promotion process are vital to maintaining the right skill mix and to developing future leaders. And yet, the promotion process is extremely resource-intensive in terms of the time to prepare board members to score records and the time spent scoring records. Additionally, despite instructions and processes to standardize scoring, results are at least somewhat dependent on differences in rater background and experience.

The second board process we focus on is developmental education boards. These boards meet to determine which officers will receive the most-competitive in-residence educational opportunities. Once again, the results of intermediate and senior developmental education (SDE) boards are vital to ensuring the proper future skill mix. As with promotion boards, developmental education boards are resource-intensive and at least somewhat dependent upon the human raters that convene.

By using ML to augment human raters, the DAF could increase the efficiency of promotion and developmental education boards. This could be done, for example, by issuing automated recommendations for the strongest and weakest records and directing human raters’ attention to the most-ambiguous records. Alternatively, this could be accomplished by generating machine summaries of the most-critical information contained in OPRs.

Additionally, by using ML to augment human raters, the DAF could increase the effectiveness of promotion and developmental education boards. Although historical board results are potentially subject to errors and bias of human raters, they reflect the consensus judgment of subject-matter experts. An ML system trained to emulate expert judgment could, at a minimum, detect cases when a future board’s decisions deviate from expectations. Alternatively, an ML system could direct experts’ attention to the most-ambiguous records, allowing the experts to increase resources spent on the most-demanding edge cases. Finally, an ML system that summarizes OPRs could decrease the likelihood that human raters would overlook a key piece of information.

These are just some of the ways that ML could increase the efficiency and effectiveness of human ratings in promotion and developmental education boards.

Purpose of the Report

Most of the methodological techniques that PReSS employs are very common in the field of NLP, and many books describe the details of implementing them in different coding languages. For common techniques, then, we will provide a conceptual summary and point to resources with more-detailed descriptions. However, building an operational decision-support system requires many specifications and tuning decisions that present trade-offs. For example, more-complex models might perform better, but they might also be less traceable and require more computing resources. Therefore, a key purpose of this report is to document and explain the choices that led to the initial version of PReSS so that future DAF efforts can revisit and refine the system.

We illustrate PReSS using the example of developmental education and promotion boards that meet to evaluate and allocate opportunities to mid-career officers at the ranks of O-4 and O-5. However, the models that PReSS uses and the decision-support tools that it supports can be generalized to early-career officers and to the enlisted force.

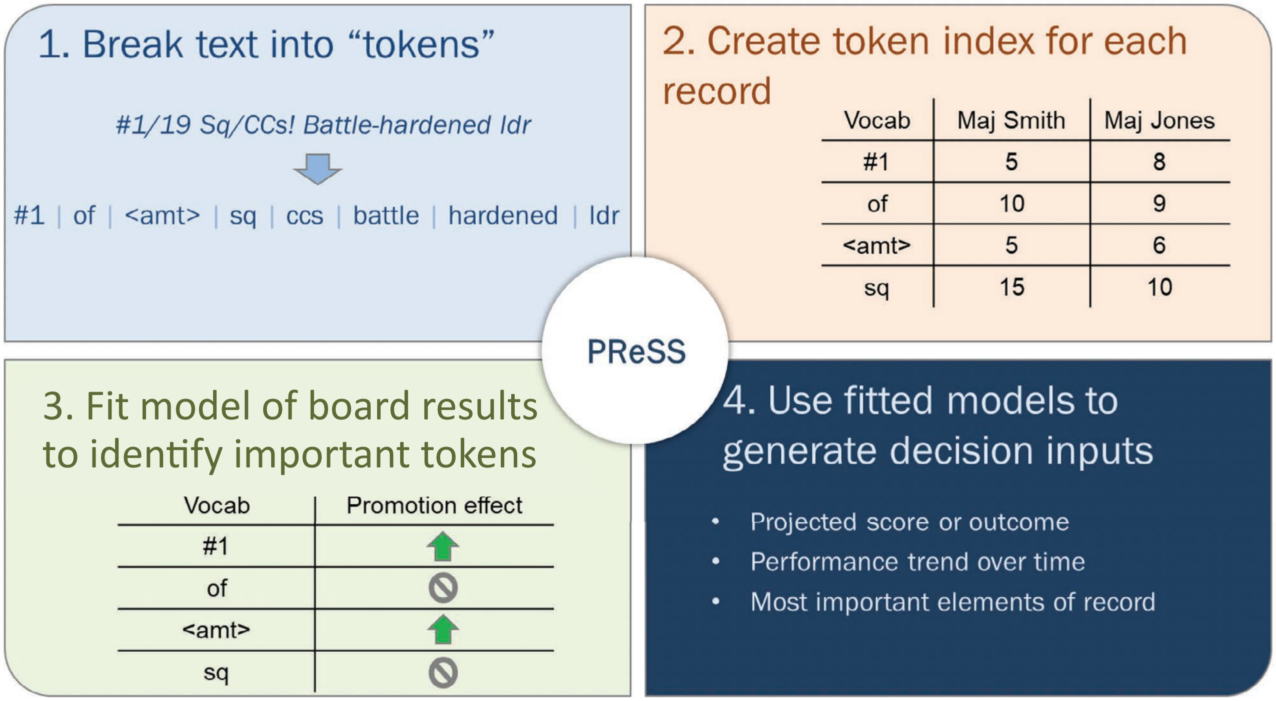
Organization of the Report

Figure 1 shows a high-level overview of the main development steps for PReSS, which is reflected in the structure for this report. The remaining sections are organized as follows:

- The next section describes how the development process begins by breaking the text into distinct tokens and indexing each record according to how many occurrences of each token it contains.
- We then explain how we combined these record-specific indexes with board outcomes or board scores so that an ML model can identify which tokens predict higher levels of performance.
- Next, we describe ways to operationalize a model relating text to performance levels by generating decision inputs. This initial version of PReSS produces a multipurpose decision input, which we refer to as the PReSS General Performance Summary.
- The last section presents our conclusion on how PReSS-type decision-support tools fit in the broader context of adopting ML to improve DAF HRM.

FIGURE 1

Overview of PReSS Development Steps


1. Break text into “tokens”
#1/19 Sq/CCs! Battle-hardened ldr
#1 | of | <amt> | sq | ccs | battle | hardened | ldr
2. Create token index for each record
PReSS
3. Fit model of board results to identify important tokens
4. Use fitted models to generate decision inputs

Vocab	Maj Smith	Maj Jones
#1	5	8
of	10	9
<amt>	5	6
sq	15	10

Vocab	Promotion effect
#1	⬆
of	⊘
<amt>	⬆
sq	⊘

- Projected score or outcome
- Performance trend over time
- Most important elements of record

Numerically Representing the Meaning of the Raw Text

To develop the PReSS ML model for use in HRM decision support, the first step in the process is breaking the text into distinct tokens and indexing each record according to how many occurrences of each token it contains.

The decision inputs in the design implementations (*recommend*, *score*, etc.) all require a method for converting raw text describing performance into a score that accurately represents the level or quality of performance. As Figure 1 shows, the first two steps of this process quantify the raw text so that it can be linked to board results by an ML model.

The right method for quantifying text depends on what information we need to extract to replicate the process that a human panel member might go through when reading the text and assigning a score. For instance, here is how one experienced Air Force Chief Master Sergeant evaluated performance statements:

> When the chief saw a strong bullet, he marked a dash "-" in the margin. If the accomplishment was significant, he crossed the dash with a "+." When the accomplishment had strategic level impact he distinguished the line by drawing a circle "0" around the "+." Consistency was the key in this approach. Scores were then tallied to reveal the strongest package (Jaren, 2017).

There are three main sources of information that NLP methods can attempt to extract from text to mimic the chief's scoring process: (1) information conveyed by the presence or absence of particular words, (2) information conveyed by the meaning of the words in context, and (3) information contained in the order in which the words appear. The simplest NLP approaches focus on (1), but more-complex models that capture elements of (2) and (3) might perform better, depending on how important the context and word order are to interpreting the text.

The example suggests that word order and context are of lesser importance to the chief's scoring process. The human process involves scanning the record for indicators of significant accomplishments (i.e., the presence of key words and phrases), with little attention to the order of the bullets. Further, while the meaning of words in general can vary greatly depending on the context, in the very select world of DAF performance writing, the context is probably less relevant to the meaning. Terms like *Sq/CC* (squadron commander), *MAJCOM* (Major Command), and *officer* have consistent meanings that do not depend on other content.[4] Therefore, it is logical to begin by testing methods that measure the presence or absence of terms in performance report text. ML models can then use this information to predict performance scores (this is discussed later).

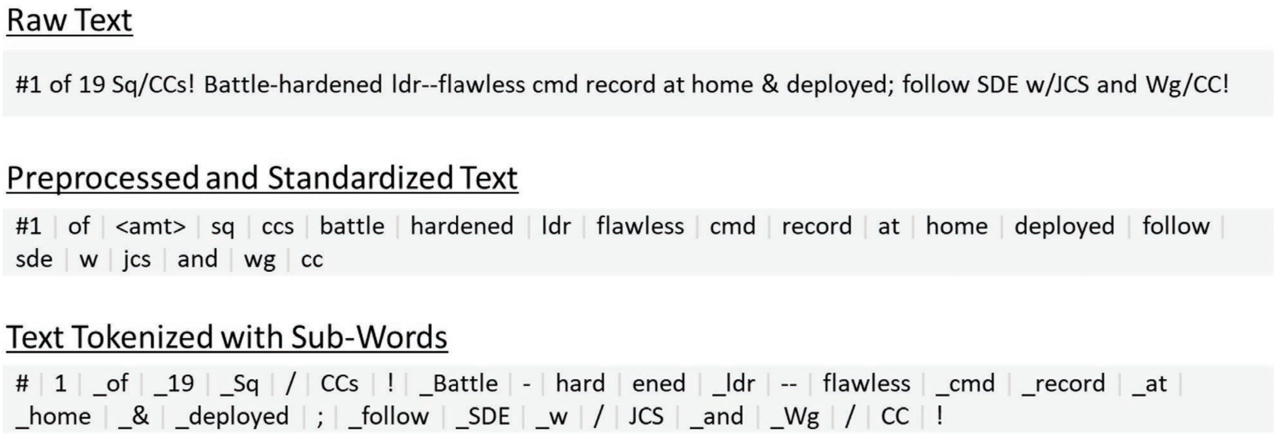
Preprocessing and Standardizing the Text

Most guides acknowledge that some amount of preprocessing can be beneficial when designing an NLP system (Vajjala et al., 2020). Preprocessing removes characters that do not convey significant meaning, and it ensures that the numerical representation of the text is identical for truly equivalent terms.[5] The end product of preprocessing is a final sequence of *tokens* (sets of characters considered by the model as distinct units) that informs the model's vocabulary and provides a summary of the content of each record.

Figure 2 illustrates two preprocessing approaches that we tested. The raw text at the top of the figure shows a typical bullet from an OPR. It contains a stratification statement (i.e., *#1 of 19 Sq/CCs*) that ranks the individual relative to their peers. It also includes school and command push statements (i.e., follow SDE w/JCS [Joint Chiefs of Staff] and Wg/CC) that advocate high-value development opportunities for the individual. The text contains abbreviations such as *ldr* for leader, *SDE* for senior developmental education, and *Sq/CC* and *Wg/CC* for squadron and wing command. Finally, the text contains punctuation such as backslashes, number signs, and exclamation points. All of these features of the raw text convey meaning.

The first preprocessing approach is a minimalist approach that converts the text to lowercase, removes punctuation, and standardizes the appearance of stratifications[6] and dollar amounts.[7] This approach

FIGURE 2
Example Effects of Preprocessing and Standardizing Steps



Raw Text

#1 of 19 Sq/CCs! Battle-hardened ldr--flawless cmd record at home & deployed; follow SDE w/JCS and Wg/CC!

Preprocessed and Standardized Text

#1 | of | <amt> | sq | ccs | battle | hardened | ldr | flawless | cmd | record | at | home | deployed | follow | sde | w | jcs | and | wg | cc

Text Tokenized with Sub-Words

| 1 | _of | _19 | _Sq | / | CCs | ! | _Battle | - | hard | ened | _ldr | -- | flawless | _cmd | _record | _at | _home | _& | _deployed | ; | _follow | _SDE | _w | / | JCS | _and | _Wg | / | CC | !

NOTE: BPE introduces a special character at the beginning of a word to differentiate tokens that begin new words from tokens that occur in the middle of a word. In this figure, we have replaced those special characters with the "_" symbol. The raw text is not from the actual data to protect privacy.

has the advantage of being simple and easy to implement, but it does not allow the model to recognize the similarities between common variants of the same term (e.g., it would not recognize that "mgt," "mngr," "mng'd," and "mgmt" are all variants of the root "manage"). Although humans readily see closely related variants of the same term, the NLP system recognizes only literal matches. The minimalist approach also discards potentially meaningful punctuation, such as exclamation points. The alternative approach uses an algorithm called *byte-pair encoding* (BPE)[8] that breaks the text into individual characters and then reassembles it into subwords according to how frequently combinations occur in the data (Sennrich, Haddow, and Birch, 2016; Gage, 1994). As Figure 2 shows, our use of the BPE method retains symbols, punctuation, and capitalization while occasionally decomposing phrases into components (e.g., "Battle-hardened" becomes "Battle" + "-" + "hard" + "ened"). Thus, BPE captures more information and is more adept at handling word fragments, but it introduces a more complicated vocabulary.

Converting Sequences into Measures of Term Presence and Absence

After the preprocessing cleans and standardizes the text, the next step is to convert the sequence of terms in each record into a matrix with cells that count how many times each word appears in the record. The simplest way to do this is to convert the text strings (such as those in Figure 2) into a term-frequency (TF) matrix, where there is a row for each record (i.e., the concatenation of all OPRs for an officer), a column for each term in the vocabulary, and cells containing the number of times each term appears in each record. Because this step discards the information contained in the word order, it is colloquially known as the "bag-of-words" (BoW) approach. Many popular software packages, such as scikit-learn in Python (Pedregosa et al., 2011), will execute many variants of this transformation very efficiently in a single line of code.

Though the BoW concept is the core of the approach in PReSS, our final set of model inputs requires the following additional steps.

- *Consider multitoken phrases in addition to single tokens.* In addition to single tokens, we enrich the TF matrix by allowing for two- and three-token combinations. For example, converting the minimalist token string from Figure 2 to TFs would include a column for "#1", "#1 of", and "#1 of <amt>" as possible predictors of performance. This will give the ML model additional capacity to capture the impact of key phrases.

- *Limit the size of the vocabulary.* Including all possible tokens would produce a matrix with tens of thousands of columns, most of which would contain words that rarely appear in performance records. Extremely rare terms, by definition, will not strongly relate to any outcome, so we limit the vocabulary to contain only the most-frequent terms. To ensure that this does not limit model performance, we test different vocabulary sizes and examine the effect on model performance.
- *Normalize the TFs.* The potential use cases for PReSS include situations in which the length of the text to score could vary. For example, one might want to use PReSS to compare performance of officers at different points in their careers, when some officers will have accumulated more OPRs than others. To reduce the sensitivity of the model predictions to the overall length, we normalize the TFs by dividing each row by the total number of terms in the row.[9]
- *Multiply by the inverse document frequency (IDF).* Long-standing results in the field of information retrieval show that there is additional benefit to adjusting TFs by a measure of how common each term is across all documents (Robertson, 2004; Sparck Jones, 1972). The intuition for this adjustment is that terms that are common everywhere are likely to be less useful for differentiating records from one another than terms that are less common overall. The IDF adjustment relies on the document frequency of each term, which is the number of records in which the term appears at least once.

The NLP literature refers to the matrix combining TF and IDF values as the *TFIDF matrix*. The final TFIDF matrix contains a row for each record, *r*, and a column for each term, *t*. The following equation defines the values that enter into the matrix:

$$TFIDF_{t,r} = \frac{TF_{t,r}}{\sum_{t'} TF_{t',r}} \left[\log \frac{1+n}{1+DF_t} + 1 \right]$$

where $TF_{t,r}$ is the TF for term *t* and record *r*, DF_t is the document frequency for term *t*, and *n* is the total number of records.

Word Meaning in Context and Sequential Ordering

The NLP methods behind this initial version of PReSS are accurate and interpretable, but they have limited depth. The models contain a vocabulary of common terms, and they can predict how different terms affect selection likelihood when they appear in a record. But they do not "understand" that "#1/22" is a stratification, "F-16CJ" is an aircraft, and "1 FW" is a large unit. Using NLP to build a knowledge base of these types of relations, known as *named entity recognition* within the domain of *information extraction* (Lane, Howard, and Hapke, 2019), could expand the horizon of NLP tools that HRM managers could use to perform other tasks (such as suggesting text or interacting with users as they write). Further, the NLP methods we use have no awareness of the meaning of the words in the vocabulary.

Though the next section of this report shows that the TFIDF approach performs well in the task of accurately predicting a performance score, other PReSS applications might call for alternative approaches that capture the richer meaning of words in context and the effects of how they appear in sequence. These approaches generally fall under the class of deep learning NLP (Vajjala et al., 2020), and they are especially useful in complex NLP tasks like machine translation that require understanding the language deeply enough to accurately generate it.

In contrast to TFIDF, some deep learning approaches represent each word in the text as a vector of values capturing different elements of what the word means (e.g., see Mikolov et al., 2013; Pennington, Socher, and Manning, 2014). Using these word vectors in a model involves replacing the model inputs based on token counts with alternative inputs based on the different dimensions of word meaning. Previous research has developed a set of word vectors specific to the defense context (Schirmer and Léveillé, 2021). However, the research on word vectors shows that they do not necessarily improve the performance of predictive models in defense-related contexts (Schirmer and Léveillé, 2021). More-recent advances in NLP performance have come through extremely large models with whole sections of the architecture dedicated to capturing the meaning of the words in context (Vaswani et al., 2017).

Richer options also exist in the deep learning framework to capture information in the order in which words appear. In the context of scoring OPRs, it is likely that more-recent performance statements carry more weight than earlier statements. Deep learning models that incorporate memory (Hochreiter and Schmidhuber, 1997) or self-attention (Vaswani et al., 2017) could, therefore, perform better, but at a cost of making the model more complex to estimate and less explainable. These techniques should be explored based on theory of the reasoning task and the amount of data available. Then they can be evaluated according to how their performance compares with the far simpler TFIDF models.

Using Labels to Fit a High-Performing Model

Once the TFIDF process quantifies the text in each officer record, the next step is to pair this information with some indicator of the performance level in each record. The ML fitting process then uses these indicators (*labels* in the ML vernacular) to learn which terms correlate with significant accomplishments. The two goals of the fitting process are (1) to produce a set of models capable of predicting the performance level of a new record, and (2) to gain some sense of how accurate each model will be when attempting to generate predictions. Different classes of models require prespecified values that affect how they estimate the scores, so a subgoal of (2) is to find the most accurate version of each model by testing a range of possible values.

Data Overview

The input data for the modeling consisted of 205,782 OPRs that the Air Force Personnel Center extracted from the Automated Records Management System (ARMS). ARMS stores officer records as images or portable document format (PDF) files, so we extracted the text as described in Schulker et al. (2021). The model fitting focused on the performance sections of the OPR rather than the header or job description information, and we combined all available performance text for each officer into a single string in preparation for the TFIDF conversion.[10]

We combined the performance information from the OPRs with two potential sources for quality indicators. First, we used information from O-5 and O-6 selection boards (like Schulker et al., 2021). The only information available from these boards to differentiate record quality is the final selection decision. Second, we used information from Developmental Education Designation Boards (DEDBs) for SDE and intermediate developmental education (IDE). These are based on panel reviews of records that have followed a scoring process like promotion boards since 2014. Unlike the binary decisions from promotion boards, however, the DEDB results include the panel average of numerical scores ranging from 6 to 10, which provides richer information for ML models to discern the relative quality of different records.

Three key aspects of the data are worth highlighting.

1. Though the selection board labels represent a large sample of high-fidelity information on the opinions of board members, initial users must interpret each model's predictions considering the specific board context. For example, O-6 promotion boards seek to differentiate among those in a select sample that consists only of officers who successfully reached the O-5 milestone. Thus, a ruleset predicting these decisions will likely overlook positive and negative statements that affect only officers early on in their careers if these statements are dissimilar to the late-career markers.
2. We provided performance statements only from OPRs to the ML models, but other board-specific information significantly affects the panel scores each record receives. In addition to OPRs, promotion boards receive a promotion recommendation form (PRF), in which supervisors provide comments and a three-tier recommendation to the board. DEDBs receive a separate form with comments and relative rankings from supervi-

sors. PRFs are not generally releasable outside of the board processes, but supervisor comments and rankings would have been available to include in the DEDB models. Even then, we chose not to include the PRF because it would limit the usefulness of the resulting models. If the ML models were to use information that is added to the record only to facilitate a selection board, the resulting model would not be able to accurately predict performance levels in other contexts.[11] The decision to include only general performance information in this initial version ensures that the system will be as broadly applicable as possible.[12]

3. A complete set of records was not available for the study, so the models are limited to a subset of events at which officers met selection boards, and not all records in the data contain officers' complete performance history. Table 3 summarizes the records that we used in the model training process.

Model Estimation Overview

We tested several ML methods for predicting board scores and selection decisions from TFIDF values. In this section, we focus on regression-based methods because they position the system well for evaluation, given that they are simple to use and interpret (Rudin, 2019). The regressions generate a prediction based on a linear combination of TFIDF values. Thus, they enable the analyst to see the direct contribution of any term to the final predicted score.

One challenge with regression stems from the reality that the TFIDF matrix has thousands of possible predictors, and the predictors are likely to be highly correlated with each other. We addressed this challenge with regularization.[13] While a standard regression would allow effects for all possible predictors, regularization pressures the model to focus on only the most-significant predictors (Hastie, Tibshirani, and Friedman, 2009).[14] As a result, any text that does not strongly relate to performance (as signaled by the board results) will not factor into the performance scores at all.

Forming the TFIDF matrix required two decisions, as discussed in the previous section: the overall size of the vocabulary and the allowable limit for multitoken phrases. Incorporating regularization into the regression model added a third decision involving a "penalty" that determines how much to pressure the model to reduce the influence of ever greater numbers of terms. There is no way to know which combination of these values will work best for scoring records, so we designed the model-fitting process to test a range of TFIDF and penalty values and selected the optimal set.

Because ML models are so flexible, they can always continue to improve predictions on the available data by making the models more tailored and

TABLE 3
Summary of Selection Board Records Included in Analysis

Selection Board	Decision Years	Total Number of Decisions	Number of Decisions Matched to Records	Average Number of OPRs per Matched Record
O-6 promotion[a]	2012, 2014–2016, 2019	35,030	5,017	12.78
O-5 promotion[a]	2006–2011, 2013–2017	64,900	17,790	10.70
SDE[b]	2013–2021	9,228	7,054	11.86
IDE[b]	2013–2021	12,104	6,163	7.91

[a] The starting number of decisions for promotion boards includes cases in which members met the board in the promotion zone (IPZ) and above the promotion zone (APZ) and members who were selected below the promotion zone (BPZ). We did not include BPZ nonselects because their quality level is not directly comparable to IPZ and APZ nonselects. In other words, high-quality BPZ nonselects are not selected because BPZ receives a small share of the opportunity, not because of the quality of the records. However, board processes confirm that BPZ selects are at or above the quality level of the lowest IPZ select, so these records are valid examples of records of sufficient quality to merit selection for the model to consider.

[b] Officers receive multiple considerations for developmental education, or multiple "looks." In our analysis, we included all available "looks" for each officer.

complex, but such improvements reach a point where they know the available data too well and perform worse on never-before-seen examples. Procedures to prevent such *overfitting* involve withholding data from the model and then using the withheld data as a proxy for new examples to evaluate how models perform. By testing different values for the model parameters on data that the model has not yet seen, the model-fitting process arrives at an optimal configuration and provides quantitative measures for how well it is expected to perform on future data, given that the process for generating the future data resembles the one that generated the available data. The text box on the opposite page, entitled "PReSS Model Selection Process," describes our exact process in more detail.

Results of Model Selection

Tables 4 and 5 illustrate the main product of the model selection process by comparing the performance statistics for the best-performing regularized logistic/linear regression models with two alternatives, which are described in more detail in the appendix. The first alternative, referred to as *Naïve Bayes*, learns a model of how likely one word is to come after another word within the set of promote records or do-not-promote records. Using these models, Naïve Bayes generates a prediction by estimating how much more (or less) likely a word or phrase is to come from the promote corpus versus the do-not-promote corpus and gives a label based on which corpus that the word or phrase is most likely to stem from. The second alternative, *AttentionNet*, is a neural-network–based approach that replaces the TFIDF values with "attention" values that the model estimates during the fitting process. This alternative has a much greater ability to account for interactions or nonlinear effects of words and phrases compared with the other methods. We selected these alternative approaches as bookends on the complexity spectrum, with the primary method, logistic regression, falling between Naïve Bayes and AttentionNet.

The best-performing methods for each measure are highlighted in Tables 4 and 5. All methods perform comparably for promotion outcomes, as seen in the similar values for accuracy and AUC. The DEDB results show that the best models explained 33 to 34 percent of the variation in scores, and the average prediction error was between 0.40 and 0.51 points on the 6-to-10-point scale.

Figures 3 and 4 further illustrate the model fit of the regression approaches for promotion boards and DEDBs by comparing predictions of the test data with the actual results. Figure 3 compares the model's assessment of each record, in the form of a predicted

TABLE 4

Performance Statistics for Predicting Promotion Board Results

	Accuracy (%)	Precision	Recall	AUC
O-6 promotion				
Logistic regression (minimalist)	81.8	0.805	0.755	0.895
Logistic regression (BPE)	82.5	0.800	0.781	0.899
Naïve Bayes	76.3	0.718	0.725	0.864
AttentionNet	73.5	0.718	0.929	0.827
O-5 promotion				
Logistic regression (minimalist)	88.9	0.914	0.869	0.952
Logistic regression (BPE)	90.2	0.929	0.879	0.963
Naïve Bayes	94.0	0.913	0.978	0.950
AttentionNet	85.4	0.898	0.926	0.875

NOTE: AUC = area under the (receiver operating characteristic) curve (see text box).

PReSS Model Selection Overview

Possible Ranges for Parameters

We considered possible ranges for tuning parameters that affect the TFIDF matrix and the regularization strength. We then sampled randomly from these ranges to generate 99 candidate configurations to explore. We expanded the ranges anytime we saw that best model called for a parameter near the boundary of the range. The ranges we considered were as follows:

- Vocabulary size can take seven possible values ranging from 5K to 40K.
- Tokens can consist of a single word or can include combinations of up to five words.
- Penalty parameters were selected from a uniform distribution ranging from zero to 200 for binary outcome models and zero to 0.0001 for continuous outcome models.[a]

Procedure for Measuring Performance on New Records

Our procedure for splitting the data to prevent overfitting included the following steps:[b]

- We randomly split the data into a training set containing 80 percent of the records and a test set with the remaining 20 percent. The test set was not available to models at any point until all final decisions had been made.
- With the training data, we applied cross-validation (which also uses random splits in the data) to measure the performance of a given set of parameters.
- We selected the model that maximized performance and fitted this model to the entire set of training data.
- We used the best model in each class to predict outcomes for the test data to measure performance of each class on new records.

Model Performance Metrics and Definitions

We relied on well-known metrics to compare the performance of models for promotion board outcomes, where the only available outcome was a binary indicator for whether an officer was selected. For accuracy, precision, and recall, we generated predictions according to whether select or nonselect was more likely for a given officer. The metrics we used were as follows:

- *Accuracy*: the fraction of records the model predicts correctly
- *Precision*: the fraction of predicted selections that are actually selected
- *Recall*: the fraction of actual selects that the model predicts will be selected
- *Area under the (receiver operating characteristic) curve (AUC)*: a continuous statistic, ranging from zero to one with higher values indicating better fit, that measures the ability of a model to discriminate between positive and negative cases and has several different interpretations (Hastie, Tibshirani, and Friedman, 2009).

For DEDB scores, which are continuous, we used the following metrics for model performance:

- *R^2*: the proportion of the total amount of error that is accounted for by the model
- *Root mean squared error (RMSE)*: the square root of the average squared difference between the model's predicted value and the true value
- *Mean absolute deviation (MAD)*: the average difference between the model's predicted value and the true value.

[a] In scikit-learn's implementation of L1 regularization for logistic regression, larger values indicate weaker regularization, whereas in the implementation of L1 for linear regression, smaller values indicate weaker regularization. In general, models with weaker regularization performed better in both cases.

[b] Our splitting procedure did not factor in the time of the boards. For example, an alternative procedure might further structure the withholding so that past records are used for training and new board records are used for testing. Thus, it is possible that the process overestimates how well the models will perform on new records, depending on how much the data and scoring process vary over time.

probability of selection, with the actual board decisions. The figure shows that the models differentiate well between selects and nonselects, especially for the O-5 board, where very few nonselects received a high probability from the model and very few selects received a low probability.

TABLE 5

Performance Statistics for Predicting DEDB Scores

	R^2	RMSE	MAD
SDE			
Linear regression (minimalist)	0.361	0.501	0.394
Linear regression (BPE)	0.359	0.502	0.397
Naïve Bayes	0.055	0.609	0.483
IDE			
Linear regression (minimalist)	0.369	0.393	0.309
Linear regression (BPE)	0.375	0.391	0.308
Naïve Bayes	0.023	0.487	0.398

NOTE: Given the performance of AttentionNet in the initial testing of promotion board models, we chose to focus on other models when the DEDB outcomes became available. To apply the Naïve Bayes model to average board scores, we rounded the board scores to the nearest half-point and treated the score as a categorical outcome.

FIGURE 3

Predicted Promotion Probability Versus Actual Board Decisions for Test Data

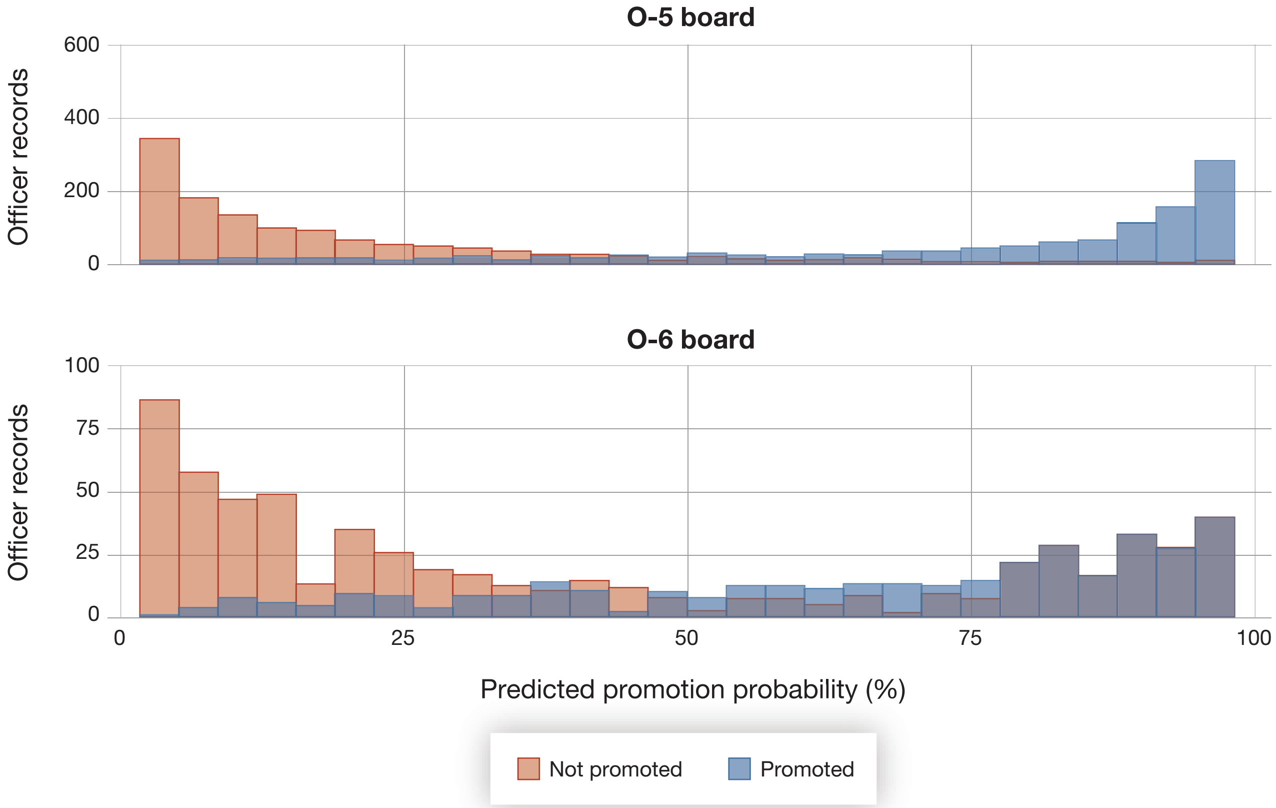


FIGURE 4

Predicted DEDB Score Versus Average Panel Score for Test Data

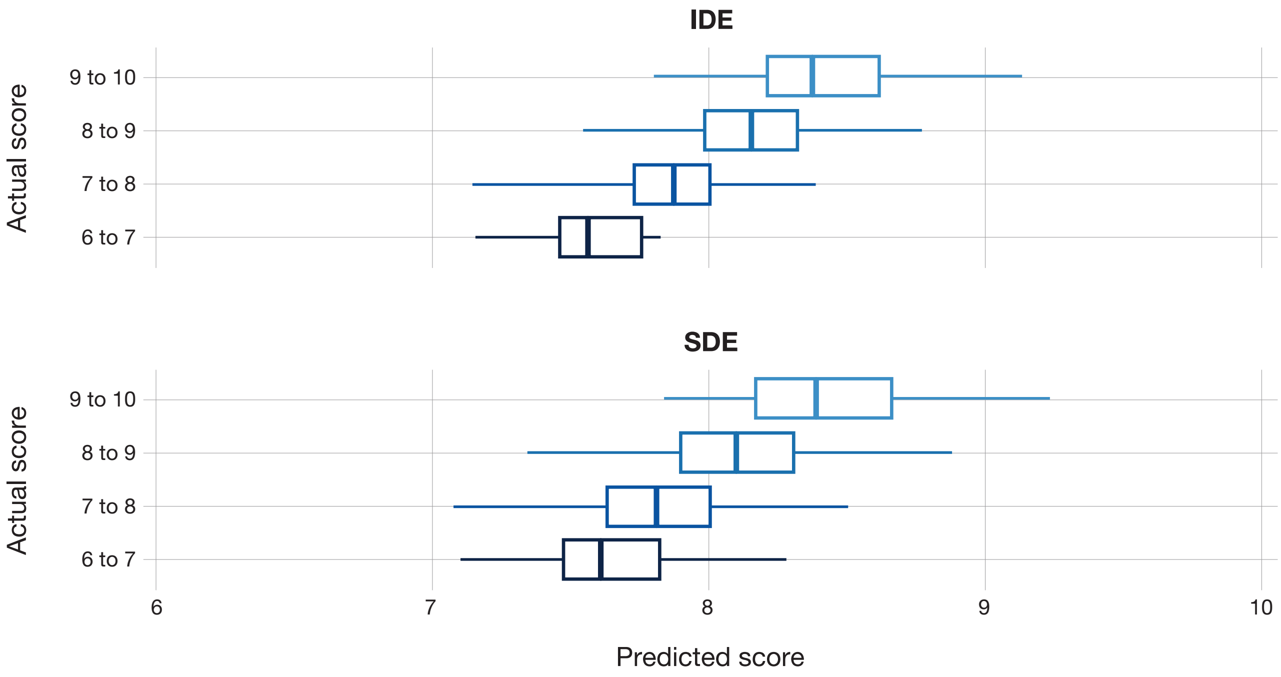


Figure 4 compares predicted DEDB scores on the test data to the actual record scores. The model generally predicts higher scores for records that received higher scores, as shown by the fact that the box plots shift further to the right with each successive category on the vertical axis. However, all the boxes are compressed toward the median score (approximately 8.0), compared with their actual scores, which means that the model overpredicts for the lowest-scoring records and underpredicts for the highest-scoring records. In other words, the model fails to find enough information in OPRs to accurately identify an outstanding record scoring above a 9.0 compared with a strong record that received an 8.5. The best explanation for this pattern is that factors that differentiate quality in the extremes relate to supervisor comments and rankings not contained in the OPRs.

If the more-complex models had performed significantly better than the regression-based approaches, this would have presented a possible trade-off between model accuracy and its ease of use and interpretability. However, as is often the case in applications with clear structure and relevant, high-quality data (Rudin, 2019), performance did not meaningfully differ between model classes. Therefore, it is pragmatic to default to the simplest and most interpretable method: regression techniques with minimal text preprocessing. The remaining sections in this report will reference these models in all results, but the performance targets in Tables 4 and 5 serve as benchmarks that future modeling efforts can improve upon as additional data become available.

Using Models to Generate Decision Inputs

The model-fitting process produces a set of scoring rules, in the form of a regression model, that can convert performance statements in an officer's record to a prediction of the board score or result. These predictions would be the only necessary ingredient if decisionmakers were to adopt the *decide* implementation for these specific boards. However, we designed PReSS to be a general-purpose scoring tool that is also useful outside these boards in the *recommend*, *score*, or *summarize* ML implementations. This section describes PReSS's novel method for generating predictions that can generalize outside a particular board context. It then presents an overview of the initial design of PReSS's main output: a general summary of the performance level, strengths, and weaknesses in an officer's record.

Predicting the Quality of a Performance Record

Using PReSS Models to Evaluate Snippets of Text

Each model learned its scoring function using past examples, in which a board had reviewed officer records and scored them as part of the process for determining whom to select for promotion or developmental education opportunities. In this case, each officer's record is made up of a concatenated stack of OPRs containing dozens of OPR bullets. This means that the model is *expecting* new records to take the same form as ones used for training, and it is unclear how well it will perform in novel applications. For example, using PReSS to score individual OPRs or text snippets, which are key components of PReSS's general performance summary, will likely yield unstable results that are difficult to interpret.

To illustrate this instability, Table 6 applies the PReSS models for O-6 promotion and SDE board scores to predict the quality of five example performance statements. The first row offers the model an empty record with no recognizable terms, while the second statement is badly written and intentionally runs counter to the model's expectations. We crafted the final three statements to realistically indicate increasingly high levels of performance. The first column for each board ("raw score") shows model predictions for each individual statement. These raw scores are often extreme and difficult to interpret. The extreme results stem from the fact that the TFIDF process normalizes the text for length. Thus, the model assumes that it has an entire record of text with only the terms included in each statement, which would be a highly irregular record. For the blank record and the second statement, the model expects no chance of selection for O-6, whereas it predicts certain selection for the fourth and fifth statements. The DEDB model also predicts scores at the extremes that are sometimes well outside the possible scoring range. While it might be possible to trace the model's logic (e.g., a hypothetical record where every bullet contained a MAJCOM stratification would seem to have a high chance of selection), these outputs would not earn the trust of decision-makers or be useful for comparing the statements with each other.

Instead of asking the models to consider each statement as if it were representative of the entire record, a better-designed question for the models would be: What would happen if this statement were added to an otherwise average record? This intuition forms the basis of the predictions in the "smoothed score" columns. To generate the smoothed predictions, we take the TF values from the respective state-

TABLE 6

Raw and Smoothed Performance Predictions for Example Statements

	O-6 Promotion Probability		SDE Score	
Performance Statement	Raw Score	Smoothed Score	Raw Score	Smoothed Score
Blank	0.00	0.54	7.06	7.95
Not war ready, very inexperienced, will fail in battle if deployed—likely to crash acft	0.00	0.53	6.50	7.95
My CAF expert! From operations to staff a full-spectrum leader; SDE/Jt Staff absolutely; future Ops Sq/CC!	0.47	0.54	9.38	7.97
#1/202 HQ AMC O-4s! Charismatic ldr--raised standard for my CAG; JCS & Sq/CC after IDE; fast track to SDE!	1.00	0.61	10.82	7.99
#1 of 19 Sq/CCs! Battle-hardened ldr—flawless cmd record at home & deployed; follow SDE w/JCS and Wg/CC!	1.00	0.70	18.12	8.06

NOTE: The smoothed scores for the blank records in the first row contain only average text, so these values (0.54 for O-6 promotion and 7.95 for SDE score) are approximately the average outcome for officers facing each board.

ments and round them out by adding new terms that represent the average use of each term in the training data until we reach the average number of terms for the records that met each board.[15] The results accord much more with what a decisionmaker might expect to see when comparing the statements. The second statement, though it reads quite negatively to a human, does not contain any terms that dramatically alter the score in the context of a larger record of performance. The final three statements begin to elevate the record, particularly in the eyes of the O-6 promotion board. The last statement, which contains arguably the most important performance marker one could have—a #1 stratification as Squadron Commander—raises the probability of selection to O-6 by 16 percentage points.

Applying the Snippet Method to Break Down and Summarize a Record

Beyond allowing decisionmakers to compare the value of individual performance statements, this smoothing concept enables any PReSS model to assess the impact of different portions of an officer's performance history, such as individual OPRs or even individual snippets of text within an OPR. We can generate an impact score of a single OPR or a text snippet by replacing the selected text with average-looking text and recomputing the prediction. If replacing the selection causes the performance score to decline significantly, this would indicate that the selection contains important markers of performance. By contrast, if replacing the selection causes the performance score to increase, then the text could signify below-average performance. High-impact snippets (both positive and negative) form the basis of PReSS's visualizations and textual summaries, which we describe next.

Overview of the PReSS General Performance Summary

In this early stage of the development process, it would be premature to build out the system into an interactive tool supporting a single use case, especially considering that each use case faces different types of complexity that affect its prospects and that the design process for a use case will need to include testing and evaluation of multiple implementation options for safety considerations (Walsh et al., 2024; Snoke et al., 2024). Instead, we provided code and models to generate general performance summaries. These general performance summaries and the functionality underlying them can be folded into specific USAF applications in the near future.

To help decisionmakers understand PReSS's potential uses, we also created a general performance summary that outlines the performance information in each officer's record with an eye toward the *recommend, score,* and *summarize* implementations. This summary report might be directly useful, but, if not, it could also serve to spur ideas for adapting the PReSS concept to different decision-support applications. The PReSS general performance summary contains three main sections: (1) an overview of the performance scores according to each of the four models, (2) a visualization of an officer's performance over time, and (3) a list of the most-significant text contained in the report.

Because parts of the PReSS general performance summary display text from the officer's record, we limit our use of examples from the actual data because of privacy concerns. Instead, we created a pseudo-record, which contains notional performance statements (like those in Table 6). We combined these simulated performance statements with textual "noise" to illustrate how the final section of the general performance summary extracts the most-meaningful snippets from the record. We used selections from the first chapter of Air Force Instruction (AFI) 1-1, *Air Force Standards* (DAF, 2014) as the noise text, because the AFI text contains terms from the model's vocabulary (e.g., "mission") without conveying any information about performance. The following text box contains an example of a performance narrative from one of the five OPRs in the pseudo-record.

Performance Overview Section

The first section of the PReSS general summary provides a simple table and visual of the overall performance level in the officer's record (Figure 5). The left-hand side lists the predicted probability that the

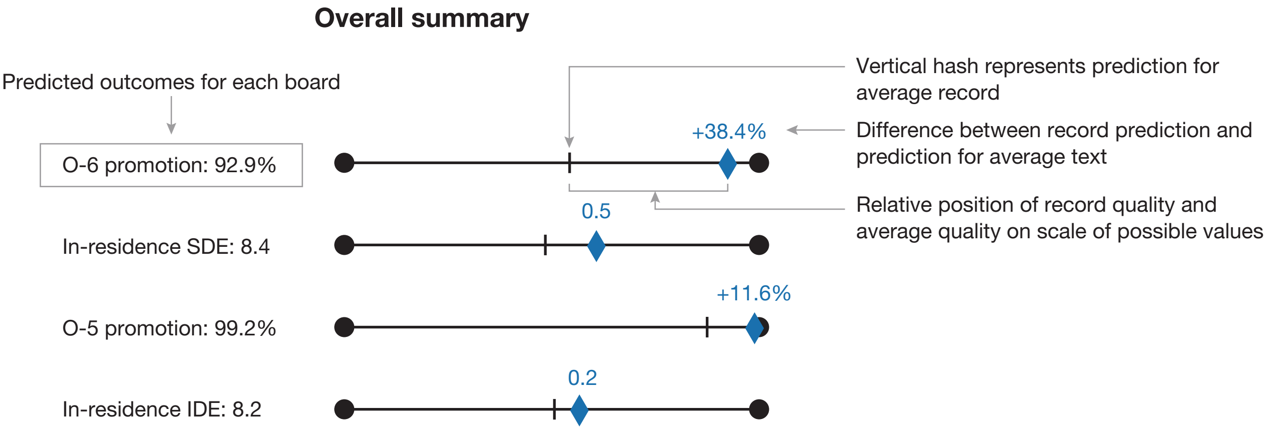
Sample OPR from the Illustrative Pseudo-Record

Rank: 2nd Lt
Date range: January 1, 2016–December 31, 2017
Performance text: Overview. The Air Force environment, whether at home station or forward deployed, encompasses the actions, values, and standards we live by each and every day, whether on- or off-duty. From defined missions to force structure, each of us must understand not only where we fit, but why. Mission. The mission of the United States Air Force is to fly, fight, and win . . . in air, space, and cyberspace. To achieve that mission, the Air Force has a vision: The United States Air Force will be a trusted and reliable Joint partner with our sister Services known for integrity in all of our activities, including supporting the Joint mission first and foremost. We will provide compelling air, space, and cyber capabilities for use by the combatant commanders. We will excel as stewards of all Air Force resources in service to the American people, while providing precise and reliable Global Vigilance, Reach, and Power for the Nation. 80 FTW CGOQ...#1/214 CGOs! Crucial bearer of Wg's primary msn--#1 of 154 Sq IPs in CY11 student sorties flown; 746 EAS Pilot of Month...#1/24 pilots; cmbt-proven ldr--groom for Asst Flt/CC; AC upgrade next; PDE for sure.

NOTE: Distracting text from AFI 1-1 (DAF, 2014; AFI text has since been revised) shown in black. Performance text shown in red.

FIGURE 5

Performance Overview Section of General Performance Summary



record would be selected by an O-5 and O-6 promotion board, along with the predicted record score from the intermediate and senior DEDBs. Promotion probabilities range from 0 to 100 percent, and DEDB scores range from 6 to 10. To place these values in context, the line graphs on the right-hand side show the difference between the predicted values (marked by blue diamonds) and the prediction for a record containing the average level of each term in the vocabulary (marked by the vertical line segment). The models appear to have high regard for the pseudo-record, though this is likely because the pseudo-record is shorter than a normal record for O-5 or O-6 promotion. Smoothing would further improve the interpretability of the predictions.

In addition to providing a top-line review of the performance in the record, the performance overview section illustrates possible inputs that HRM personnel could use for the *score* or *recommend* implementations. The figure provides an estimated score, which could be used as an input into the human scoring process. A decision-support tool could also use the promotion predictions to form recommendations that influence the decisions of panel members (see Snoke et al., 2024, for a more detailed discussion of possible ML design implementations in support of selection boards).

Performance over Time Visualization

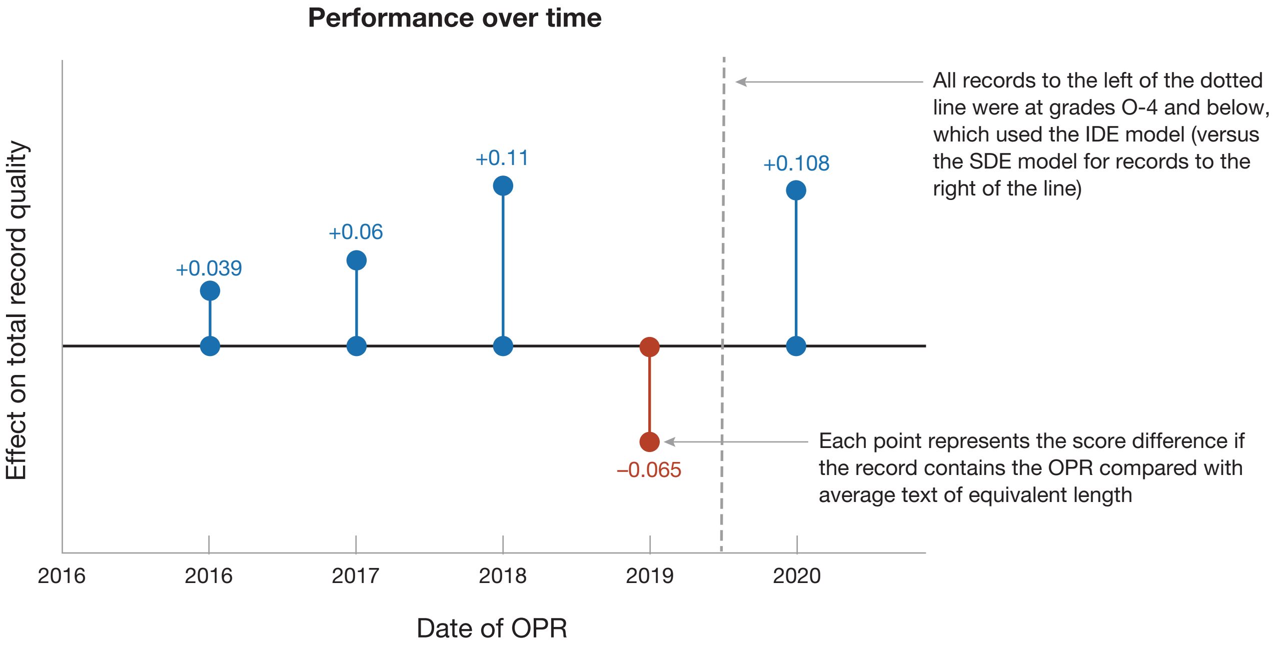
Directly below the performance overview section, the PReSS general performance summary includes a plot of the officer's performance trend over time (Figure 6). PReSS calculates the value for each OPR in Figure 6 by removing the OPR text from the record, replacing it with average-looking text, and recomputing the overall score. This simple-yet-powerful visual helps reveal different facets of an officer's performance history that are well known to experienced board members. The time trends could help an HRM decisionmaker understand how an officer's recent performance compares with the overall record, or to determine whether a particularly weak OPR is an aberration or characteristic of a broader trend. The trends also show which officers go through periods of regression, when their performance appears to decline, and if the officer is a "late bloomer," whose recent performance indicates much more potential than the cumulative record. In the case of the pseudo-record, all OPRs except the fourth in the series contain performance statements with key markers, such as stratifications. The negative value attached to this OPR reflects the fact that the overall record score improves when we replace it with average-looking text of equal length.

Comparing Records with Performance Summary Visualizations

The performance overview and performance over time visualizations would enable panel members to rapidly acquire a general awareness of the performance in a set of competing records, as shown in Figure 7. The figure compares the outputs for an actual record that received a high score from the SDE board (top panel) with a record that received an average score (bottom panel). The model predictions on the left side show that the officer with the strong record is a shoo-in for promotion (99.8-percent chance) and should receive above-average scores from the DEDBs. The models predict that the officer with the average record is likely to make O-5, but there is about a 27-percent chance that they will not make O-6. The performance over time visuals are

FIGURE 6

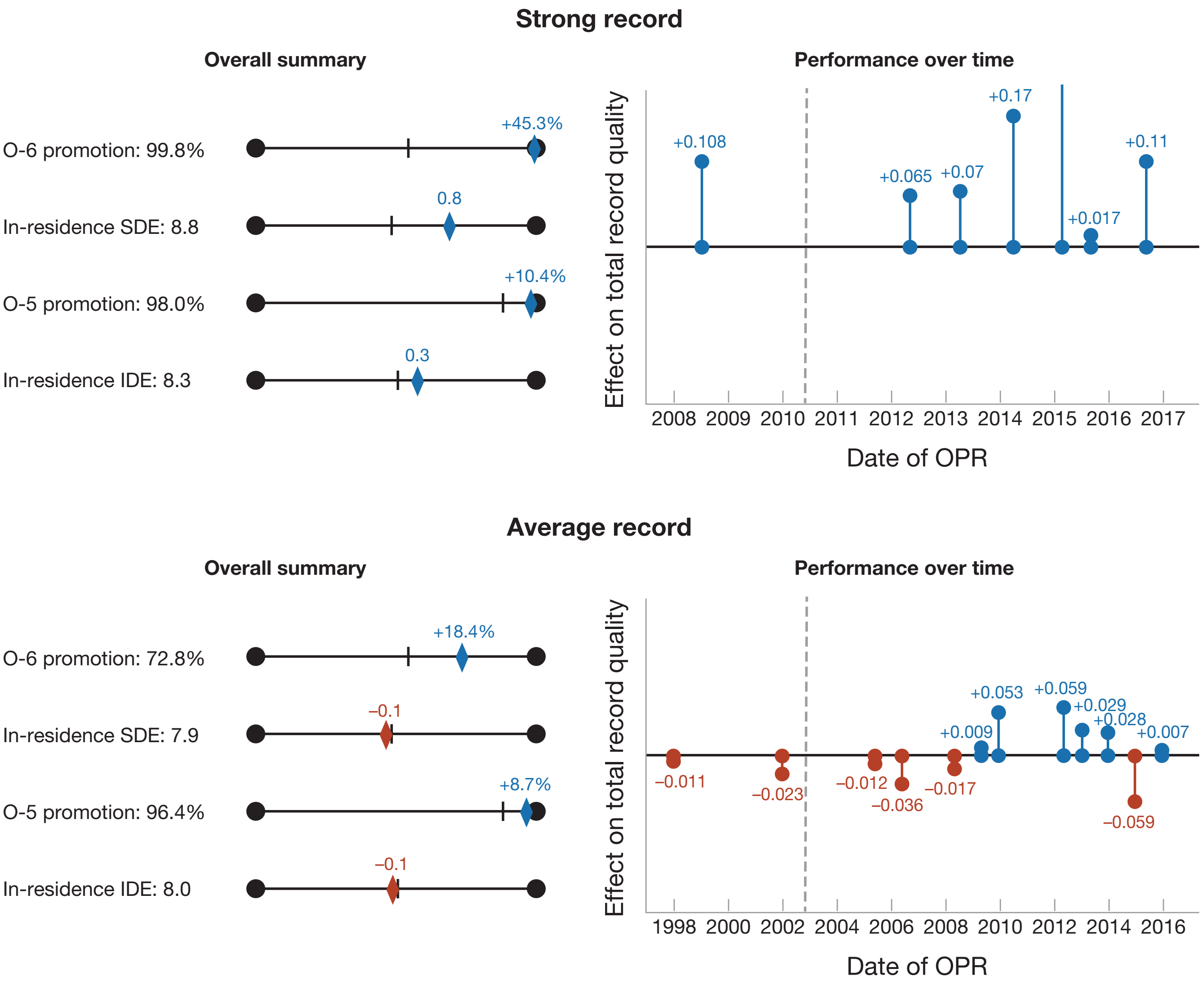
Performance Trend Section of General Performance Summary



NOTE: Though any of the PReSS models could generate this visual in its entirety, the initial version applies the IDE model to evaluate OPRs at grades O-4 and below (all OPRs to the left of the dotted line) and the SDE model for OPRs at grades O-5 and above (the one OPR to the right of the dotted line). Thus, the values in the figure are on the standard 6-to-10-point quality scale.

FIGURE 7

PReSS General Performance Summary Visualizations



also quite informative. The officer with the strong record appears to get better with each successive OPR as an O-5 (i.e., all points to the right of the vertical dotted line), peaking at around the time that the officer would be approaching leadership opportunities, such as squadron command. By contrast, very few of the OPRs for the officer with the average record stand out, and the performance level trends downward after this officer was promoted to O-5.

List of High-Impact Text

The final section of the PReSS general summary provides a grade-by-grade list of the most-significant text snippets in a record. PReSS creates this summary by breaking the record into snippets using common punctuation, such as semicolons and dashes, as delimiters. Then, the system calculates the impact of each snippet by replacing it with an equal amount of average-looking text and recomputing the record score. All snippets with impacts above a user-specified threshold appear in a table at the bottom of the general performance summary.

Figure 8 shows part of this table for the pseudo-record. Each snippet includes shading with progressively darker shades indicating higher impact, allowing the HRM decisionmaker to quickly review and compare performance statements at each grade. The table separates the list into two columns: one for “record-enhancing” text and one for “record-moderating” text.

FIGURE 8

Summary of High-Impact Text Portion of General Performance Summary

Grade	Record-enhancing text	Record-moderating text
O-3 (cont'd)	Instructor next/groom for WIC	including weekends
	PDE a must	Dlvrd 3.1m lbs of fuel to spt 45 TICs
	The mission must be accomplished	
O-2	My #1 of 10 schedulers!	Service Before Self tells us that professional duties take precedence over personal desires
	My #2 of 10 Fit/CCs!	
	It is the moral compass of the inner voice	
Q1	#1/214 GCOs!	including supporting the Joint mission first and foremost
	#1/24 pilots	whether on- or off-duty
	#1 of 154 Sq IPs in CY11 student sorties flown	whether at home station or forward deployed
	Crucial bearer of Wgs primary msn	groom for Asst Flt/CC

NOTE: Though any of the PReSS models could generate this visual in its entirety, the initial version applies the IDE model to evaluate OPRs at grades O-4 and below and the SDE model for OPRs at grades O-5 and above.

The example snippets from the pseudo-record highlight two key points for how HRM personnel interpret the list of high-impact text. First, the fact that a snippet appears in the table does not guarantee that board members view it positively (or negatively). Rather, the precise interpretation is that the terms in the statement, on net, are more strongly correlated with board scores than average text.[16] Second, it is not always obvious why a phrase affects the estimated quality of the record in either direction. In particular, the record-moderating text is difficult to interpret. We have observed that snippets that are long and vague tend to moderate the strength of a record. This is possibly useful, but not as compelling as examples of record-enhancing text.

This initial PReSS static report presents some options to decisionmakers that they could adapt in several ML design implementations. Further, the prediction method that forms the core of the PReSS general performance summary would work just as easily for any model that uses TFIDF as an input, should DAF analysts continue to improve upon the performance of the models with new alternatives.

Limitations of PReSS

The models used by PReSS accurately predict promotion outcomes and DEDB scores. Additionally, the predicted outcomes and summaries of high-impact text open the door for several ML design implementations. Nevertheless, PReSS is limited in at least three significant ways.

First, PReSS is trained to emulate historical decisions of human board members. One goal of DEDB and promotion boards is to identify officers with the greatest leadership potential and to prepare them to serve in positions of greater responsibility. Past decisions of human board members are, at best, an indirect measure of leadership potential. Thus, PReSS is trained using an imperfect measure of ground truth. To overcome this limitation, the U.S. Air Force could define and collect more-direct measures of the primary outcome—performance in leadership positions.

Second, the criteria used by DEDB and promotion boards to score records evolve over time. As of this writing, PReSS gives equal weight to all records, regardless of age. To overcome this limitation, PReSS could be trained using an objective function that gives greater weight to more-recent records.

Third, the content of records may also evolve over time. For example, phrases like *joint all domain command and control*, or *JADC2*, may appear for the

PReSS takes the first step toward helping supervisors and HRM personnel understand and use this rich performance information more effectively.

first time in new records that must be scored. As of this writing, PReSS ignores terms that are not contained in data used to train the models. However, the models can be retrained after each board, after which new terms take on significance.

Conclusions and Recommendations

Large-scale data management and ML tools have become increasingly available to analytic organizations, enabling analysts to build systems that transform unstructured text into decision inputs, such as scores or predictions. The HRM domain stands to disproportionately benefit from these advances because it relies on the “rich medium” of textual narratives (Brutus, 2010). Despite this potential, surveys have documented that firms have been slow to adopt AI-based decision-support systems for HRM processes, relative to other areas (Chui et al., 2021). If the DAF is to capitalize on its infrastructure investments and become more “data-centric” in its approach to HRM, it will need to become very effective and efficient at spotting, developing, and safely deploying decision-support systems.

Among the many ways HRM personnel use unstructured text, a ubiquitous class of decisions involves cases in which individuals review officer performance narratives to select a subset of officers for an HRM action. These decisions are almost always constrained in the amount of time and effort the individual can spend reviewing the records and in the amount of expertise possessed by any individual. Even the most-seasoned officers are more knowledgeable about their own functional areas than others, which could create blind spots in their judgment.

We developed PReSS to address this common challenge. Because the volume of performance information presents a difficulty in many HRM actions, a well-functioning ML decision-support system has the potential to add significant value if it reaches maturity. For example, we presented earlier a narrative description of one Air Force Chief Master Sergeant’s process for systematically scoring a set of performance statements, one by one. Without fundamentally changing the level of human control in the decision, the PReSS general performance summary would enable the following alternative scoring process:

> The chief started by taking the highest and lowest prescored records and saving them for the end of the day, when he would be running low on energy. In later reviewing these records, the chief focused on confirming the presence or absence of obvious significant accomplishments and/or red flags. He quickly looked through the summaries of midrange records for those who had earned significant stratifications or other markers and tentatively placed them in the “select” pile. All panel members spent extra time discussing performance trends, strengths, and weaknesses of records close to the cutoff, in light of the strategic HR goals described in the board memorandum.

PReSS takes the first step toward helping supervisors and HRM personnel understand and use this rich performance information more effectively. The system can further support any of the ML implementations (Table 1) that we discuss in this report and other reports in the series. Nonetheless, the findings that ML can technically be applied to a particular HRM process, and that it can potentially yield decisions that are more efficient and effective than is currently possible, does not mean that the DAF should immediately adopt the system. Before doing so, the DAF must take steps to ensure that the system is safe.

(Another report in this series presents a framework for evaluating the safety of ML systems for HRM, and we apply the framework with PReSS as the use case.) In the meantime, this research offers the following conclusions and recommendations for related DAF efforts to design and implement PReSS-like decision-support tools.

Conclusion 1: Several Noteworthy Attributes of the PReSS Use Case Cause It to Work Particularly Well

Several attributes of the PReSS use case cause it to work particularly well. First, PReSS seeks to support a task that human judges can perform with precision. This means that there are clear patterns for the model to find that help it accurately predict performance. In fact, board processes build in checks to ensure that human judges evaluate records according to standardized criteria (as the earlier chief's quote noted: "Consistency was the key in this approach"). ML approaches will tend to be less accurate if they are attempting to support decisions in which human judges disagree—in that case, outcomes used to train models are less reliable. Second, with PReSS, a sufficient sample of examples wherein human judges scored records is available, thanks to the high-quality historical records on selection boards. In other use cases, if sufficient data are not available, the DAF would have to invest in additional labor to either create examples for training ML models or to create detailed scoring rules that reach a sufficient level of accuracy. Third and finally, supervisors use a constrained language when writing performance evaluations, and policies and regulations limit how supervisors can use language when writing evaluations. The consistent way language is used somewhat reduces the complexity of the NLP task.

Recommendation 1a: Use Organic Analytic Resources to Test the Viability of Use Cases with These Attributes

In hindsight, because of the rich data available and the high precision with which human judges perform the task, an NLP-based performance-scoring tool is low-hanging fruit. Still, there are likely many other HRM use cases that share these attributes. Assignment matching, occupational classification, and training curation are examples of other areas in which analysts could apply the same steps and, at a minimum, test the viability of a decision-support tool. DAF analytic organizations could take PReSS as a "worked example" to do rapid business case analyses of other applications, which would involve confirming the available data and testing the performance of basic supervised ML modeling approaches.

Conclusion 2: Well-Established Computational Methods Are Competitive with State of the Art

In recent years, there have been breakthroughs in state-of-the-art large language models for interpreting and generating text and neural networks for predicting outcomes based on complex inputs. In two use cases—promotion boards and DEDB boards—we found that relatively simple approaches such as linear regression and Naïve Bayes performed as well as or better than AttentionNet. The implication is that although state-of-the-art approaches may be justified in some cases, other long-standing techniques may be adequate or even preferable.

Recommendation 2a: Begin New Exploratory Efforts with Simple, Explainable Approaches

PReSS is an example in which "small data" rather than "big data" can still prove useful for identifying predictive relationships between text patterns and HRM decisions, which is common in HRM (Tambe, Cappelli, and Yakubovich, 2019). Simple approaches also have the advantage of being traceable and explainable. In seeking out PReSS-like applications, we recommend analysts begin by exploring the data and optimizing the performance of simple, regression-based approaches before moving on to test more-complex (but potentially higher-performing) methods.

Conclusion 3: Data Refinements Could Further Improve the Functionality of PReSS and Other Decision-Support Tools

Building on the previous conclusions, this research shows that DAF analysts can develop simple, high-performing systems to improve HRM decisions, even in the presence of imperfect data. In the process of designing PReSS, we identified several recommendations for the DAF to consider in the data management realm.

Recommendation 3a: Efforts to Digitize and Centralize HR Records, Such As myEval, Could Further Streamline the Development of Models and Improve Performance

Though we use a rich dataset of several thousand HRM decisions to develop the PReSS models, we also show that our dataset is incomplete. This limitation stems from the challenges inherent in transferring and processing records using current data management systems. Initial user-facing challenges aside (Cohen, 2022), moving to a system that accepts and stores all performance information digitally and in a format that analysts can more easily access should improve model quality by making more-complete inputs available.

Recommendation 3b: Tailor Data Inputs for Particular Use Cases

Our goal for this analysis was to create a general-purpose performance-scoring tool, which means that we intentionally limited the data inputs to include only information that would be available for each officer at any point in his/her career. To create a PReSS-like capability to support a particular HR decision, analysts should consider tailoring the data inputs to improve performance. This analysis focused solely on officer evaluations, but tool designers could tailor the method to particular processes by including inputs specific to those decisions. For example, with the use case of DEDBs, rater recommendations and assignment histories are promising sources for information that could improve model performance.

A second, less obvious way to tailor data inputs for a particular use case is to create training examples for ML that exemplify alternate rating schemes that advance U.S. Air Force objectives. For example, in performance scoring, the ML models (like the board members they emulate) tend to focus on stratifications and push statements as the most important factors that drive decisions (Schulker et al., 2021). To steer the models in a different direction, developers could create new training examples that reflect the effects of scoring records using alternative rules.

APPENDIX

Alternative Methods for Predicting Performance Levels

The modeling results discussed in the main body of the report focus on regularized linear or logistic regression. To understand the performance trade-offs involved in selecting a method, we tested two alternatives in the development process. Here, we describe these methods in more detail.

Naïve Bayes

Naïve Bayes is a simple and popular method of classifying data based on Bayes Rule with the assumption that all input features are independent of one another (Hastie, Tibshirani, and Friedman, 2009). This means we assume that OPRs consist of an unspecified number of statements about each candidate, where each statement (e.g., “Great future officer–promote today”) is independent of all other statements (i.e., evaluators are not basing their evaluation on any other evaluation). Naïve Bayes then calculates the probability of promotion as the likelihood that the language used in a candidate’s OPR appeared in the OPRs of other candidates who were promoted or not promoted.

Computing this likelihood relies on a simple N-gram model that learns the probability of a sequence of words and/or tokens appearing within the OPRs of officers who were promoted or not promoted (Hovold, 2005). From these learned probabilities, we can directly calculate the likelihood of a statement appearing in the OPRs of other officers.

For these evaluations, we allowed fourth-order N-grams (i.e., we consider at most four words or tokens at a time to capture stratification statements in a single N-gram—e.g., "# 1 of 25"). To compute DEDB scores, we treated DEDB as a classification task with classes spaced in half-point intervals from six to ten.

AttentionNet

AttentionNet (Yoo et al., 2015) takes as input a tokenized sequence of text and uses a series of neural network layers to output a promotion probability between 0 and 1. The tokenized text is passed through an embedding layer where each tokenized word is assigned 256 embeddings. Next, the embeddings are passed through an *Attention* layer, which assigns a weight to each embedded token. The sum of all weights in a sequence equals 1. The purpose of the Attention layer is to learn the importance of certain words in a sequence of text, relative to the outcome variable (promotion outcome). The Attention layer feeds into several fully connected layers, which yield the final predicted promotion probability.

The network is fit using Adam stochastic gradient descent on a binary cross entropy loss function with respect to the promotion variable or, in the case of DEDB, continuous scores. We trained the network using a train-test-validation split. The test set was used to tune the following hyperparameters: embedding dimension, training batch-size, and dropout-rates for the fully connected layers.

Notes

[1] The research conducted on language models was completed in April 2023, prior to the emergence and widespread adoption of large language models (LLMs), such as ChatGPT. Therefore, the findings and conclusions presented in this study may not reflect the current state of the field and should be interpreted with this context in mind. It is important to note that the rapid development and evolution of LLMs may have significant implications for the study of language models, and future research should take these advancements into consideration.

[2] For example, management-level review boards award "definitely promote" designations prior to a promotion board.

[3] For a more detailed discussion of the different classes of AI and ML techniques, see Walsh et al. (2024).

[4] One additional advantage of using an approach that is not sensitive to word order and syntax is that, if the structure of the OPRs changes (for instance, by shifting from bulleted statements to paragraphs) but the words and phrases conveying impact remain, the models could still be useful.

[5] Common preprocessing steps involve changing all words to lowercase, removing stop words that contain limited meaning (e.g., "the," "and," "of"), and applying *stemming* or *lemmatization* rules that strip endings and suffixes from words to reduce them to their root (Lane, Howard, and Hapke, 2019). We explored these techniques and did not see any improvement in model performance.

[6] The stratification standardization replaces the denominator with the token "<amt>," which assumes that all denominators are equivalent. Without this step, different denominators (e.g., 5, 8, 10) would be treated as distinct words in the vocabulary. An alternative would be to strip the numerator and denominators of stratifications and add them in as quantitative variables (rather than tokens). We explored this alternative approach and did not see significant gains in performance, so we defaulted to the simpler approach.

[7] It is not shown in the example, but this step replaced any dollar amount, such as $1.5M, with the token "<mny>." As in the discussion about stratifications, this step discards any information contained in the quantitative value of the dollar amount.

[8] BPE was initially developed to solve the problem of encountering words that are not part of the original model's vocabulary, as it allowed the model to reconstruct the novel word using subwords that it recognized. Subword tokenization is used in most advanced NLP applications (Wolf et al., 2020).

[9] The default setting in scikit-learn's implementation of term frequency inverse document frequency (TFIDF) normalizes the final TFIDF value by dividing the row by the L-2 norm (i.e., dividing by the sum of squared values). We found that this normalization introduced unusual behavior in the prediction and summary functions when we wanted to introduce a variable amount of "average" text into a prediction. We discuss this more later.

[10] During the process of linking OPRs to board scores and outcomes, we omitted any text that entered the record too late to be considered.

[11] For example, the model could not readily provide accurate performance scores for more-junior officers who had not yet been nominated to meet a DEDB.

[12] If decisionmakers desire a model specifically to support the selection board process, incorporating these other elements could further improve predictive accuracy.

[13] Specifically, we used L1 regularization in all models, which penalizes the absolute size of the model coefficients. L2 regularization, which penalizes the squared coefficient, is also a common alternative, as is the elastic-net penalty, which is a compromise between the two. A key benefit of L1 is that it pressures irrelevant coefficients to be exactly zero (rather than just small), making it easy to view just the subset of the vocabulary that is relevant to performance.

[14] Regularization constrains the sum of all coefficients at a prespecified value. The smaller that value, the simpler the final model will be. Because it tends to reduce the amount of influence

that variables can have in the model, regularization falls under a broader class referred to as *shrinkage* methods.

[15] For promotion boards, we use the average frequency of text from IPZ records to prevent above-the-zone records from bringing down the average. This ensures that the prediction for the average record is close to the promotion rate for officers in the zone. For DEDBs, we used the average frequency from across all "looks."

[16] In fact, very few records contain statements describing negative performance. Instead, raters differentiate performance by replacing highly recognizable signals with vaguely positive placeholders. Thus, a statement along the lines of "groom for Asst Flt/CC" could indicate average or below-average performance.

References

Balakrishnan, Tara, Michael Chui, Bryce Hall, and Nicolaus Henke, "The State of AI in 2020," McKinsey & Company, November 17, 2020. As of July 6, 2022: https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2020

Brutus, Stephane, "Words Versus Numbers: A Theoretical Exploration of Giving and Receiving Narrative Comments in Performance Appraisal," *Human Resource Management Review*, Vol. 20, No. 2, 2010.

Cohen, Rachel S., "Air Force Leaders Pledge to Fix Hated myEval Software," *Air Force Times*, September 2, 2022. As of September 12, 2022: https://www.airforcetimes.com/news/your-air-force/2022/09/02/air-force-leaders-pledge-to-fix-hated-myeval-software/

Chui, Michael, Bryce Hall, Alex Singla, and Alex Sukharevsky, "The State of AI in 2021," McKinsey & Company, December 8, 2021. As of July 6, 2022: https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

DAF—*See* U.S. Department of the Air Force.

DoD—*See* U.S. Department of Defense.

Gage, Philip, "A New Algorithm for Data Compression," *C Users Journal*, Vol. 12, No. 2, February 1994.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Science and Business Media, 2009.

Hochreiter, Sepp, and Jurgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, Vol. 9, No. 8, 1997.

Hovold, J., "Naive Bayes Spam Filtering Using Word-Position-Based Attributes," conference paper, in CEAS, Conference on Email and Anti-Spam, July 2005. As of July 18, 2023: https://ceas.cc/2005/papers/144.pdf

Hutto, Clayton, and Eric Gilbert, "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the International AAAI Conference on Web and Social Media*, Association for the Advancement of Artificial Intelligence, Vol. 8, No. 1, 2014.

Jaren, Eric R., *Brown Bag Lessons: The Magic of Bullet Writing*, Air University Press, December 2017.

Lane, Hobson, Cole Howard, and Hannes Max Hapke, *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*, Manning Publications, 2019.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," in *ICLR Workshop Papers*, International Conference on Learning Representations, 2013.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Edouard Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, No. 85, 2011.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning, "GLoVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Robertson, Stephen, "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF," *Journal of Documentation*, Vol. 60, No. 5, 2004.

Rudin, Cynthia, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, Vol. 1, No. 5, May 2019.

Schirmer, Peter, and Jasmin Léveillé, *AI Tools for Military Readiness*, RAND Corporation, RR-A449-1, 2021. As of July 6, 2023: https://www.rand.org/pubs/research_reports/RRA449-1.html

Schulker, David, Nelson Lim, Luke J. Mathews, 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 July 6, 2023: https://www.rand.org/pubs/research_reports/RRA812-1.html

Schulker, David, Matthew Walsh, Avery Calkins, Monique Graham, Cheryl K. Montemayor, Albert A. Robbert, Sean Robson, Claude Messan Setodji, Joshua Snoke, Joshua Williams, and Li Ang Zhang, *Leveraging Machine Learning to Improve Human Resource Management*: Vol. 1, *Key Findings and Recommendations for Policymakers*, RAND Corporation, RR-A1745-1, 2024.

Sennrich, Rico, Barry Haddow, and Alexandra Birch, "Neural Machine Translation of Rare Words with Subword Units," in *54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2016.

Snoke, Joshua, Matthew Walsh, Joshua Williams, and David Schulker, *Safe Use of Machine Learning for Air Force Human Resource Management: Evaluation Framework and Use Cases*, RAND Corporation, RR-A1745-4, 2024.

Sparck Jones, K., "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, Vol. 28, 1972.

Tambe, Prasanna, Peter Cappelli, and Valery Yakubovich, "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward," *California Management Review*, Vol. 61, No. 4, 2019.

U.S. Department of the Air Force, Air Force Instruction 1-1, *Air Force Culture: Air Force Standards*, August 7, 2012, incorporating Change 1, November 12, 2014.

U.S. Department of the Air Force, Air Force Instruction 36-2110, *Total Force Assignments, Guidance Memorandum 2020-01*, July 28, 2020.

U.S. Department of the Air Force, *Department of the Air Force Implementation Plan of the DoD Data Strategy*, February 2021.

U.S. Department of the Air Force, Department of the Air Force Guidance Memorandum to DAFI 36-2501, *Officer Promotions and Selective Continuation*, January 20, 2023. As of July 17, 2023: https://static.e-publishing.af.mil/production/1/af_a1/publication/dafi36-2501/dafi36-2501.pdf

U.S. Department of Defense, *DoD Data Strategy*, Washington, D.C., 2020. As of May 5, 2021: https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF

Vajjala, Sowmya, Bodhisattwa Majumder, Anuj Gupta, and Harshit Surana, *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*, O'Reilly Media, 2020.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems*, Long Beach, Calif., 2017.

Walsh, Matthew, Sean Robson, Albert A. Robbert, and David Schulker, *Machine Learning in Air Force Human Resource Management:* Vol. 2, *A Framework for Vetting Use Cases with Example Applications*, RAND Corporation, RR-A1745-2, 2024.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, et al., "Transformers: State of the Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, 2020.

Yoo, Donggeun, Sunggyun Park, Joon-Young Lee, Anthony S. Paek, and In So Kweon, "AttentionNet: Aggregating Weak Directions for Accurate Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers, 2015.

Acknowledgments

We thank Gregory Parsons, Director of Plans and Integration, Deputy Chief of Staff for Manpower and Personnel, Headquarters U.S. Air Force (AF/A1X), for his support throughout the project. We thank Col Laura King (AF/A1H) and Doug Boerman (AF/A1X). This research benefited greatly from their input and support. We are deeply appreciative of the assistance we received from many personnel, including Lt Col Monique Graham (2022 RAND Air Force Fellows Program). Finally, we thank the many RAND colleagues who helped with this work: principally, but not exclusively, Melissa Baumann, Michael Ryan, Ignacio Lara, Nelson Lim, and Miriam Matthews.

Abbreviations

AFI	Air Force Instruction
AI	artificial intelligence
APZ	above the promotion zone
ARMS	Automated Records Management System
AUC	area under the curve
BPE	byte-pair encoding
BPZ	below the promotion zone
DAF	U.S. Department of the Air Force
DEDB	Developmental Education Designation Board
DoD	U.S. Department of Defense
HR	human resources
HRM	human resource management
IDE	intermediate developmental education
IDF	inverse document frequency
IPZ	in the promotion zone
KSAOs	knowledge, skills, abilities, and other attributes
MAD	mean absolute deviation
MAJCOM	major command
ML	machine learning
NLP	natural language processing
OPR	officer performance report
PReSS	Personnel Records Scoring System
PRF	promotion recommendation form
RMSE	root mean squared error
SDE	senior developmental education
TF	term frequency
TFIDF	term frequency inverse document frequency

About This Report

The Department of the Air Force (DAF) has begun to develop and field artificial intelligence and machine learning (ML) systems for myriad mission areas and support functions, including human resource management (HRM). ML systems could accelerate existing decision processes and enhance decision quality by leveraging data. Further, by allowing the DAF to make decisions at greater speed and scale, ML systems have the potential to enable entirely new decision processes.

One of the richest sources of personnel data available to the DAF is performance reports. Information about service members' knowledge, skills, abilities, and other attributes is contained in these reports, along with supervisor assessments of their performance and leadership potential. This information could be used to directly inform selection boards and to indirectly inform a range of other HRM processes that consider service member experiences and performance. However, it is difficult to access and use this information because it is embedded in unstructured, narrative text.

We propose a system that uses natural language processing to extract meaning from narrative text and make it available as inputs to a range of HRM processes. This report describes the methodology behind the system, known as the Personnel Records Scoring System (PReSS), and illustrates the primary output of the initial release of the system—a general summary of performance information contained in an officer's record.

The research reported here was commissioned by the Director of Plans and Integration, Deputy Chief of Staff for Manpower and Personnel, Headquarters U.S. Air Force (AF/A1X) and conducted within the Workforce, Development, and Health Program of RAND Project AIR FORCE as part of a fiscal year 2022 project, "Machine Learning Decision-Support Tools for Talent Management Processes." This is one of five related reports originating from the project: The others in the series describe (1) strategic considerations for the DAF as it pursues applications of ML to HRM, (2) a framework for selecting a portfolio of ML projects, (3) a test and evaluation strategy to ensure that ML models meet safety standards, and (4) a conceptual implementation of an ML system for informing officer assignments.

RAND Project AIR FORCE

RAND Project AIR FORCE (PAF), a division of the RAND Corporation, is the Department of the Air Force's (DAF's) federally funded research and development center for studies and analyses, supporting both 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 September 13, 2022. The draft report, dated April 2023, was reviewed by formal peer reviewers and DAF subject-matter experts.

For more information on this publication, visit www.rand.org/t/RRA1745-3.

www.rand.org

RR-A1745-3