informs
http://pubsonline.informs.org/journal/inte

INFORMS JOURNAL ON APPLIED ANALYTICS
Articles in Advance, pp. 1–12
ISSN 2644-0865 (print), ISSN 2644-0873 (online)

An Air Force Pilot Training Recommendation System Using Advanced Analytical Methods

Nicolas C. Forrest,[a] Raymond R. Hill,[a,*] Phillip R. Jenkins[a]

[a] Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio 45433
* Corresponding author

Contact: nicholas.forrest.4@us.af.mil (NCF); rayrhill@gmail.com, https://orcid.org/0000-0001-8413-8362 (RRH); phillip.jenkins@afit.edu, https://orcid.org/0000-0002-2425-9151 (PRJ)

Received: September 29, 2020
Revised: January 30, 2021; May 7, 2021
Accepted: June 29, 2021
Published Online in Articles in Advance: November 16, 2021

https://doi.org/10.1287/inte.2021.1099

Abstract: The U.S. Air Force has a severe shortage of pilots. The Air Force's Pilot Training Next (PTN) program seeks a more efficient pilot-training environment emphasizing the use of virtual reality flight simulators alongside periodic real aircraft experience. The objective of the PTN program is to accelerate the training pace and progress in undergraduate pilot training. Currently, instructor pilots spend excessive time planning and scheduling flights. This research focuses on methods to autogenerate the planning of in-flight events using hybrid filtering and deep learning techniques. The resulting approach captures temporal trends of user-specific and program-wide student performance to recommend a feasible set of graded flight events for evaluation in students' next training exercise to improve their progress toward fully qualified status.

History: This paper was refereed.

Keywords: recommender system • deep learning • neural networks

Introduction

The U.S. Air Force has a serious operational readiness issue; there are not enough pilots to meet mission demands. At the end of fiscal year 2016, the Air Force's total force structure was 1,555 pilots short of requirements needed to meet national security demands. Lt. Gen. Grosso, then-Chief of Staff for Air Force Personnel, identified the need for pilot production and the Air Force's progressive focus on developing creative, agile solutions to meet pilot demands (McRae 2017).

Pilot-production processes produce qualified pilots. This requires that the Air Force provide sufficient time and realistic training environments for candidates to develop their skills. Currently, undergraduate pilot training (UPT) provides the Air Force's training program for pilots. UPT operates in a three-phase system that spans about one year. The first phase introduces students to basic aircraft control and flying with instruments in an academic environment using mostly aircraft and simulators to help students gain requisite flight experience. Phase 2 begins with a series of basic flight events and transitions into training blocks focused on formation and navigation flying events. Daily evaluations are scheduled, conducted, and reviewed by an instructor pilot (IP), introducing students to legitimate flight hours in a training aircraft. Phase 3 involves more-specified training tracks in either of two other training aircraft. At the end of phase 3, students deemed fully proficient graduate from UPT and move on to their next duty assignment. That next phase, not considered in this work, trains the pilot candidates destined to operate in their newly assigned aircraft in either the mobility air forces or the combat air forces (Korger 2019). See Colbath (2020) for a comprehensive overview of the Air Force UPT system.

The pilot shortage puts pressure on the pilot-training program to produce more pilots at an accelerated pace. However, the required resources are currently not available to increase pilot production. Innovative training methods are needed, which led the Air Force to initiate the Pilot Training Next (PTN) program to provide a more efficient and personalized pilot-training experience. PTN emphasizes the use of virtual reality flight simulators alongside periodic real aircraft experience to progress pilot-training students to qualified status.

PTN differs from traditional UPT in at least four ways: immersive technology; more simulator availability; a personalized syllabus; and experience in a low-risk, high-reward environment. Immersive virtual reality training provides students with training opportunities at reduced costs compared with training in a real aircraft by reducing overall strain and time spent preparing and maintaining aircraft for flight training events. PTN also provides students with more access to virtual reality training with flight simulators. Larger, more-realistic simulators are available in the office, with smaller simulators available at home. More access to flight simulators allows students to continue practicing beyond daily duty hours and gives students more access to training, regardless of external factors such as weather, time of day, or aircraft availability. The use of the simulators to gain proficiency also reduces the dependency of the students having an IP available to guide them. The PTN training schedule focuses on particular pilot-training student competencies and gives students the opportunity to advance in training at their own pace. The first PTN graduating class prepared pilots in approximately 50% less time than UPT (i.e., six months versus 12 months).

A drawback to the PTN individualized training system is that IPs spend hours planning the training events. An automated flight-planning recommender system would alleviate this administrative workload and is a key component of the PTN program. The initiative is called the AutoGradebook. However, unfortunately for the AutoGradebook initiative, there are no recommender systems for pilot-training applications; thus, the research presented in this paper presents a first-ever instance of such a system.

This paper presents the initial recommender system conceptualized, prototyped, tested, and transitioned to the PTN program. The research leading to the recommender system also provided recommendations to zimprove data-collection methods and student-evaluation metrics.

Background and Literature Search

Recommender systems are not new. Most people interact with them regularly, particularly during an internet session. This section introduces recommender systems, some of their applications in research and industry, some common recommendation-generation approaches, and complications associated with implementing more-personalized recommender systems.

Some History on Recommender Systems

Good decision-making approaches help explore all options available. However, exploring all options becomes increasingly difficult as systems grow in size, complexity, and/or influence. Historically, peer and expert recommendations help to simplify these larger decision-making scenarios. As experts or decision makers seek more-personalized recommendations, social methods of acquiring information cannot always provide sufficient advice. Computer-based recommendation systems introduce the ability to obtain more-specified information or advice for a decision maker's interests (Ekstrand et al. 2011).

As computer-based recommendation systems become standard practice in decision support, automated recommender systems will become more common. Early automated recommender systems depended on hard-coded, user-provided specifications to filter through possible options and make suggestions. Today, many online recommender systems do not even require user input to generate recommendations. Instead, modern recommender systems often employ automatically recorded data from user activity to generate effective suggestions (Ekstrand et al. 2011). In fact, effective recommender systems have become essential to the success of major e-commerce companies, such as Amazon and Netflix (Koren et al. 2009).

Common Recommendation-generation Approaches

Two of the most common personalized methods for generating user recommendations are collaborative filtering and content-based filtering.

Collaborative-filtering methods assume that a group of users with highly correlated behavior will have similar preferences. Active users deemed similar to such groups are considered members to provide a reasonable prediction of their preferences (Ekstrand et al. 2011). An early automated collaborative-filtering algorithm is the *k*-nearest-neighbor (k-NN) collaborative-filtering technique by Miller (1995), whereas two of the better-performing methods of collaborative filtering are latent factor models and neighborhood models (Koren 2008).

Item-based collaborative filtering relies on similarities between rating patterns of items rather than user behavior (Ekstrand et al. 2011). Companies that sell consumer goods, such as Amazon, often use item-based collaborative filtering to advertise goods that are predicted to meet customer needs (Linden et al. 2003). Netflix, and other companies looking to provide entertainment services, tends to use hybrid approaches of item-based and k-NN methods to capture personalities of consumers rather than the functionality of a specific item (Koren et al. 2009).

Content-based filtering systems produce recommendations based on items previously preferred by the user (Ekstrand et al. 2011). Multivariate techniques, such as Bayesian classifiers, cluster analysis, decision trees, and artificial neural networks, are examples of methods used for such user profiling (Sanghavi et al. 2014). Pandora, an online music-streaming company, has had success applying a content-based filtering algorithm to recommend new music to users (Koren et al. 2009).

Ensemble modeling employs multiple modeling approaches applied to a common problem to produce a solution to that problem. Ensemble modeling overcomes limitations associated with any particular modeling approach when applied to problems of interest. Not surprisingly, in machine learning, an ensemble of models generally outperforms individual models. Hybrid filtering is an approach that employs an ensemble approach via preliminary cascading techniques to combine model results (Sanghavi et al. 2014). See Ricci et al. (2011) for an in-depth overview of such recommendation systems and methods.

Deep learning models (i.e., multilayer artificial neural networks) are widely discussed because of their ability to learn and exploit the unknown structures within data. The various types of neural network model architectures that are suitable for different recommendation tasks are viewed as neural building blocks for complex models. Deep neural networks are composed of multiple neural building blocks that combine to form one functioning model. These deep learning models can model vast amounts of complex data, providing an additional advantage for content-based recommendation tasks.

The PTN recommender defined and prototyped in this work uses the temporally ordered sequences of flight events found in the exercise plans for each student produced throughout pilot training to produce legitimate recommendations for the flight events evaluated in upcoming exercises. A deep learning model structure that has proven successful in capturing temporal data trends for prediction purposes is the long short-term memory recurrent neural network (LSTM RNN) (Xie and Wang 2018). The LSTM RNN algorithm is used in this work to generate an initial set of events for the next training exercise for an individual pilot candidate.

Potential Complications of Personalized Recommendations

Personalizing recommender systems introduces complications. A highly personalized recommendation system can become inconsistent when placed in general use. In a pilot-training recommender system, such inconsistencies can introduce overall inefficiencies when student performance varies from average student performance. Students whose performance differs from the average performance are labeled “black sheep,” whereas an average performer is the “white sheep.” Further complicating the pilot-training system is that both types of performers can interchange roles, owing to training-performance lapses or advances. Recommender systems struggle with the black-sheep users because of their differences (Srivastava et al. 2019). A PTN recommender system should focus on moving the poor-performing black sheep toward the better performer. This movement of performance toward the better performer is a novel aspect of the recommender system described in this paper and implemented in the collaborative filtering component of the hybrid algorithm developed.

Two final complications are briefly noted. First, data formats and content can change, thus degrading recommender-system performance. Second, measuring recommender-system performance is hard when the number of possibly correct recommendations is not unique (Ekstrand et al. 2011). We address both of these final complications in our model development and testing effort.

Table 1. Point Allocation per Event Graded Evaluation

Recorded grade	Definition	Point equivalent
E	Excellent	4
G	Good	3
F	Fair	2
U	Unsatisfactory	1
NG	No grade	0
N/A	No recorded data	0

Data Description, Collection, and Preparation

The data used for this work came from the first PTN class. The raw data set consists of the scores received on every graded event performed during each training exercise for the 19 students in the original PTN class. Those scores are well defined and summarized in Table 1. Ultimately, 18 students were used, as one trainee left the program. The data include information on 128 individual flight events executed during pilot training. Training events fall into 10 different event categories: basic, patterns, contact, instruments, basic formations, tactical formations, low-level, four-ship formations, combat air forces introduction, and mobility air forces introduction, as listed in Table 2. Only a subset of all possible flight events are performed during each training exercise. Each student has an identification number, and each record in the data represents the information for a single training exercise for a given student.

The data naturally required some cleaning, which is not detailed here. The data are also limited in some respects, such as a lack of weather information, start times, and individual event-ordering details. These limitations were noted in data-engineering recommendations provided to the research customer, but these did not impact the recommender-system development effort.

Exploratory Data Analysis

PTN features tailored training programs. The first PTN class of 18 graduated students required a minimum of 60 days, a maximum of 100 days, and an average of 83 days to complete the training. This range in the training length shows individualization of the training programs, but program individualization impacts IP efforts. The IPs spend a significant amount of time creating the tailored sequence of events for each pilot trainee’s next exercise, choosing among the 128 possible training events. IPs must consider trainee

Table 2. List of Graded Flight Events by Event Category Provided by the PTN Program

Event categories	Graded flight events
Basic	Mission Analysis/Products, Ground Ops, Takeoff, Departure, Basic Aircraft Control, Cross-Check, Enroute Descent/Recovery, Inflight Checks, Inflight Planning, Clearing/Visual Lookout, Communication, Risk Mgmt/Decision Making, Situational Awareness, Task Management, Emergency Procedures, General Knowledge
Patterns	Overhead/Closed Pattern, Visual St-In, Landing, No-Flap Landing, Go-Around, Emergency Landing Pattern
Contact	G-Awareness, TP Stalls, Slow Flight, Power On Stalls, Contact Recoveries, Spin Recovery, Aileron Roll, Barrel Roll, Pitchback/Sliceback, Cloverleaf, Cuban Eight, Immelmann,Lazy Eight, Loop, Split S
Instrument	Vertical S, Unusual Attitudes, Steep Turns, Intercept/Maintain Arc, Fix to Fix, Holding, Full Procedure Approach, Non-Precision Final, Precision Final, Circling Approach, Missed Approach, Night Landing
Basic formation	Wing Takeoff, Interval Takeoff, Instrument Trail, G-Warmup/Awareness, Lead Platform, Pitchout (Both), Fingertip (Wing), Route (Wing), Fighting Wing (Wing), Straight Ahead Rejoin, Turning Rejoin, Overshoot, Echelon (Wing), Breakout (Wing), Lost Wingman (Both), Extended Trail (Wing), Position Change, Formation Approach (Both), Formation Landing (Both), Battle Damage Check, Flt Integrity/Wingman Consideration
Tactical formation	Delay 90, Delay 45, Hook Turn, Shackle, Cross Turn, Fluid Turn, Tactical Rejoins, Fluid Maneuvering, Tac Initial
Low-level	Course Mx, Course Entry, Time Control, Altitude Control, Checkpoint ID, LL GPS Integration, Tactical Maneuvering, LL Lead Change
4-Ship formation	Four Ship Admin, Fluid 4, Box Formation, Offset Box, Wall, 4-Ship Fingertip, Straight Ahead Rejoin, Turning Rejoin
CAF introduction	Heat to Guns Setup, Heat to Guns Maneuvering, Fuel Awareness/Management, Advanced Handling, Perch Setups, Maneuver Selection, Offensive Fighter Mnvr Exec, Defensive Fighter Mnvr Exec, CZ Recognition, Air to Air Weapons Employ, HA Lead Turn Exercise, HA Butterfly Setups, HA BFM Flt Analysis, SA Conventional Range, SA Tactical Range Proc, SA Safe-Escape Maneuver, SA Threat Reaction, SA Weapons Employment, Air to Ground Error Analysis, TACS/JFIRE Procedures, Air to Ground 2-Ship Mutual Supt
MAF introduction	Mission Management, VFR Arrival, Tanker Procedures, Reciever Procedures, Airdrop Procedures, Crew Coordination, Single Engine Approach, Single Engine GA/Missed Appch, A/R Overrun, A/R Breakaway, FD/AP Operations, FMS Operations

Notes. These 128 events fall into 10 categories. Subsets of each are selected for each pilot-training event. CAF, combat air force; MAF, mobility air force.

event proficiencies and the introduction of new events to build the trainee's breadth of experience. An initial data analysis provides useful insights for defining the PTN recommender system. In particular, occurrence distributions of each event provide insight into when events are typically scheduled during the training campaign for each student and are used to define the event scores used in the collaborative filtering component of the algorithm described below. Details of the analysis are available in Forrest (2020).

Table 3 depicts the two basic training groupings uncovered. Events in group A are introduced early; events in group B begin appearing after about training exercise 40. Table 4 breaks out the occurrence distribution of each event. The key takeaways from these data are as follows:

- The training features depth of event coverage to achieve proficiency.
- The training features breadth to cover all the events.
- There are clearly timing issues regarding the repetition of trained events.
- There is the introduction of new, as-yet-untrained events.

Not detailed here, but available in Forrest (2020), is the distributional information pertaining to the frequency of event selection based on the particular time in the training program (i.e., the likelihood of the event occurring in a particular exercise), as implied by the data in Table 4.

Legacy and Proposed Progress Score

All PTN students progress at their own pace. Training focuses on proficiency depth as well as skills breadth. The current overall progress metric is the maneuver item file (MIF). Events are graded on the zero-through-four ordinal scoring scale, corresponding to "No Grade" through "Excellent," highlighted in Table 1. The MIF is the cumulation of the maximum recorded scores over all the events available.

Figure 1 shows the cumulative MIF for the first PTN class (excluding combat air force and mobility air force events, which may not be common to all completing students). The standard MIF threshold, represented by the thick dashed line, is the maximum cumulative MIF score a student can achieve. The level area early in the training campaign seems to indicate progress stagnation, but is actually a period of increasing breadth of skills. The MIF does not adequately account for breadth and only focuses on proficiency levels. Such a measure is not particularly conducive to building a recommender system.

Table 3. Grouped Basic Formation Events Based on Event Occurrence Frequencies Shown Here as Group A and Group B

Group A	Group B
Wing Takeoff	Interval Takeoff
G-Warmup/Awareness	Fighting Wing (Wing)
Lead Platform	Instrument Trail
Pitchout (Both)	Turning Rejoin
Fingertip (Wing)	
Route (Wing)	
Straight Ahead Rejoin	
Overshoot	
Echelon	
Breakout (Wing)	
Lost Wingman (Both)	
Extended Trail (Wing)	
Position Change	
Formation Approach (Both)	
Formation Landing (Both)	
Battle Damage Check	
Flt Integrity/Wingman Consideration	

Note. Events in group A occurred early in a pilot trainee program, whereas events in group B occurred later in the pilot trainee program.

The Forward Progress Score (FPS) was designed to better model student progress by incorporating both depth and breadth aspects of training into a single metric. Achieving proficiency in each event is assumed as the primary goal for each student. The FPS uses a percent value of the individual maximum MIF scores as a variable representing student progress toward proficiency in each event. Applying percentages of total progress toward a set goal addresses skill depth in the training campaign toward overall proficiency more appropriately than simply considering recorded grades. Additional details on the calculation of the FPS are contained in Appendix A.

Visuals representing student FPS score over time, again excluding combat air force and mobility air force event evaluations, are represented in Figure 2. Unlike Figures 1, 2 shows a consistent progression of student performance throughout training, a metric more useful to building a recommender system. This alternate progress measure was positively received by the PTN program and is being examined for employment.

Defining the PTN Recommender System

The flight-training planning process involves many aspects. Among them are repetition of tasks to gain and improve proficiency, the introduction of new tasks to gain requisite breadth of skills, and the timely progression toward full qualification. Each training exercise involves a set of events accommodating these key aspects of the training. The IP will spend an inordinate amount of time creating each trainee's sequence of events for each exercise.

A guiding principle in the building of the recommender system was that the IP-produced exercise sequence of events embedded in the PTN data were correct. Thus, an accurate recommender system should closely match the IP suggestions. The overall recommender-system approach is graphically defined in Figure 3 and described next.

An IP plans a student's training flight as a sequence of events. The recommender system needs to generate such sequences. However, the PTN data set was not large enough to train the sequence-recommendation component. Thus, the data for the 18 students, considering the 128 possible events, over the 60–100 training exercises available, were presented as sets of events (versus ordered vectors of events). This expansion of the data presentation proved sufficient to actually train the event-sequence-generation component of the recommender system.

The first component of the hybrid recommender system produces the initial event set recommended for the next training exercise. An LSTM RNN is trained to output an unordered set of events for the next exercise combined to create the initial exercise plan. Of the 18 student records available, the first four were held out as test data, with the remaining used as training data. The training involved leave-one-out cross-validation to ensure the quality of the model. The data provided for each student involved the event scores for that student over past training, with the scores normalized to a percentage of the required proficiency level for that event. The training metric employed used the agreement between the algorithm-recommended set of events with those suggested by the IP, because the initial event set suggested should resemble that coming from the IP. Model training involved 100 iterations, unless there was a lack of improvement, at which point the training was terminated. This overall approach kept with the objective of building a recommender that mimics the work of the IP.

Just using the initial recommended set of events causes stagnation in pilot student progression. In practice, training must focus on improving every students' performance—those weak students who need to improve toward the overall average performance and stronger students who must continue to improve. A novel collaborative filtering extension is employed to achieve improved recommender-system performance.

The collaborative filtering component is a swap heuristic that replaces events selected by the LSTM RNN with nonselected events if the swap is deemed beneficial. This is accomplished by using a measure associated with each event that helps determine whether that event will help to improve student performance.

At each stage of the training campaign, a gold standard of performance is defined. This is the performance

Table 4. Statistics for Reaching Proficiency Across Original PTN Class Measured by Training Exercise

Graded event	Min	Median	Mean	Max	Graded event	Min	Median	Mean	Max
Mission Analysis/Products	13	31.5	33	56	Extended Trail (Wing)	56	66	66	77
Ground Ops	14	28	32	64	Position Change	28	51	49	68
Takeoff	1	28.5	28	56	Formation Approach (Both)	54	57.5	58	61
Departure	3	28	29	57	Formation Landing (Both)	49	49	49	49
Basic Aircraft Control	13	36	38	79	Battle Damage Check	45	55	57	72
Cross-Check	19	33	39	86	Flt Integrity/Wingman Consider	31	47.5	49	73
Enroute Descent/Recovery	14	31	33	74	Delay 90	41	52	56	77
Inflight Checks	14	31	32	57	Delay 45	52	61	63	77
Inflight Planning	19	34.5	38	74	Hook Turn	52	57	61	77
Clearing/Visual Lookout	14	30	33	74	Shackle	49	61.5	62	77
Communication	13	29	32	67	Cross Turn	54	59	61	75
Risk Mgmt/Decision Making	12	29	30	53	Fluid Turn	77	77	77	77
Situational Awareness	14	32.5	36	86	Tactical Rejoins	44	61	56	62
Task Management	19	31.5	35	76	Fluid Maneuvering	54	61	61	68
Emergency Procedures	19	31	38	86	Tac Initial	47	53.5	56	69
General Knowledge	19	29	37	79	Course Mx	47	66	65	88
Overhead/Closed Pattern	14	30	30	52	Course Entry	38	55	60	88
Visual St-In	7	32	26	33	Time Control	41	51	54	70
Landing	2	28	27	45	Altitude Control	50	52	60	88
No-Flap Landing	48	58	64	88	Checkpoint ID	51	63	64	88
Go-Around	19	34	36	60	LL GPS Integration	49	58	62	88
Emergency Landing Pattern	11	17	18	31	Tactical Maneuvering	51	66	63	76
G-Awareness	28	32.5	36	57	LL Lead Change	52	67	63	76
TP Stalls	28	30.5	33	45	Four Ship Admin	62	62	62	62
Slow Flight	—	—	—	—	Fluid 4	56	56	56	56
Power On Stalls	22	33	33	45	Box Formation	—	—	—	—
Contact Recoveries	21	30	33	57	Offset Box	62	62	62	62
Spin Recovery	2	7.5	10	36	Wall	70	70	70	70
Aileron Roll	16	28	28	40	4-Ship Fingertip	—	—	—	—
Barrel Roll	29	39.5	40	53	4-Ship Straight Ahead Rejoin	—	—	—	—
Pitchback/Sliceback	28	28	28	28	4-Ship Turning Rejoin	62	62	62	62
Cloverleaf	13	29.5	30	42	Heat to Guns Setup	65	69	69	73
Cuban Eight	3	32	28	38	Heat to Guns Maneuvering	59	65	66	72
Immelmann	11	32	32	43	Fuel Awareness/Management	59	69	70	81
Lazy Eight	28	40	40	56	Advanced Handling	—	—	—	—
Loop	27	31	32	40	Perch Setups	59	65	66	73
Split S	4	37	36	57	Maneuver Selection	61	65	67	73
Vertical S	—	—	—	—	Offensive Fighter Mnvr Exec	62	66	67	73
Unusual Attitudes	16	19	19	22	Defensive Fighter Mnvr Exec	64	73	73	81
Steep Turns	71	75	75	79	CZ Recognition	65	70	71	78
Intercept/Maintain Arc	31	53	54	84	Air to Air Weapons Employ	—	—	—	—
Fix to Fix	4	19	22	61	HA Lead Turn Exercise	—	—	—	—
Holding	29	48.5	46	62	HA Butterfly Setups	—	—	—	—
Full Procedure Approach	27	37	38	56	HA BFM Flt Analysis	72	74.5	75	79
Non-Precision Final	26	31	36	78	SA Conventional Range	—	—	—	—
Precision Final	1	29.5	30	55	SA Tactical Range Proc	—	—	—	—
Circling Approach	16	58.5	55	86	SA Safe-Escape Maneuver	—	—	—	—
Missed Approach	11	45	45	63	SA Threat Reaction	77	77	80	87
Night Landing	17	37	34	48	SA Weapons Employment	76	85	83	89
Wing Takeoff	14	43	46	75	Air to Ground Error Analysis	—	—	—	—
Interval Takeoff	48	59	58	67	TACS/JFIRE Procedures	—	—	—	—
Instrument Trail	64	72	72	80	Air to Gnd 2-Ship Mutual Supt	—	—	—	—
G-Warmup/Awareness	14	59	58	77	Mission Management	71	85	84	97
Lead Platform	33	48	46	58	VFR Arrival	68	86	80	86
Pitchout (Both)	14	48	42	57	Tanker Procedures	—	—	—	—
Fingertip (Wing)	14	54	50	68	Reciever Procedures	—	—	—	—
Route (Wing)	42	54	53	72	Airdrop Procedures	—	—	—	—
Fighting Wing (Wing)	52	57.5	58	72	Crew Coordination	63	85	80	87
Straight Ahead Rejoin	40	43	50	67	Single Engine Approach	—	—	—	—
Turning Rejoin	51	55	57	66	Single Engine GA/Missed Appch	—	—	—	—
Overshoot	44	49	49	54	A/R Overrun	—	—	—	—
Echelon (Wing)	50	50	50	50	A/R Breakaway	—	—	—	—

Table 4. (Continued)

Graded event	Min	Median	Mean	Max	Graded event	Min	Median	Mean	Max
Breakout (Wing)	33	53	53	64	FD/AP Operations	—	—	—	—
Lost Wingman (Both)	43	62.5	59	77	FMS Operations	72	72	78	90

Notes. Each exercise entry provides the minimum, median, mean, and maximum values. These distributional data are used in the recommender system to produce event utilities, key to the collaborative filtering component of the recommender system.

level toward which all underperforming students should gravitate. The gold standard is defined as the average event scores of the top 10% of performers in the class. Underperforming pilot trainees need training plans designed to push them toward the gold-standard performer. The collaborative filter does this by considering all possible events and swapping LSTM RNN-selected events with nonselected events when prudent, as determined by the calculated event-utility measure.

The individual event fraction (IEF) is the difference in proficiency between a student's event score and that of the gold-standard performer. This score is then weighted by the likelihood of that event being included in the training plan at this point in the training program. These likelihoods are based on Table 4 data. This weighted score is called the event utility. The collaborative filter varies in how it handles weak students (below the standard) and strong students (above the standard).

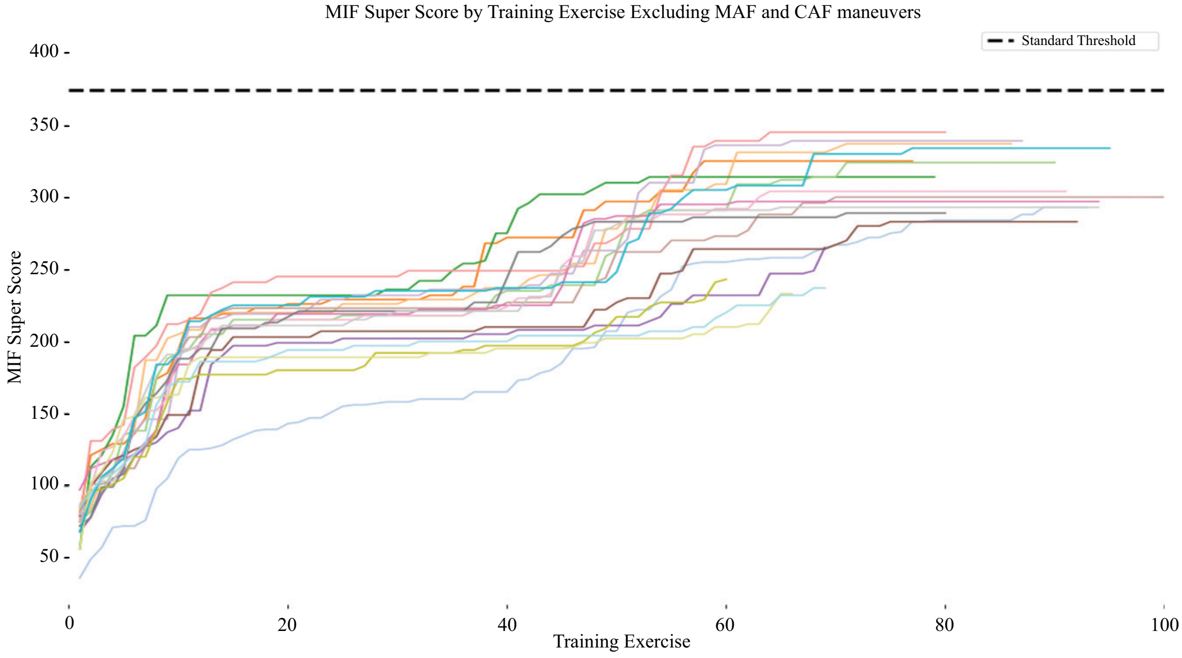
For the weaker students, the collaborative filter swaps low-utility events in the recommended event set with higher-scoring events not selected. The number of swaps is kept to a fraction of the total low-utility events for that student to maintain an element of consistency in the final training-exercise event set while still pushing the training toward desired proficiency.

For the stronger students, the collaborative filter considers event swaps that either further improve student performance (achieve a depth of experience) or add new events (to achieve a breadth of experience). Once the collaborative filter processing is complete, the event set for that student is provided as the next recommended training exercise.

Model Results and Testing

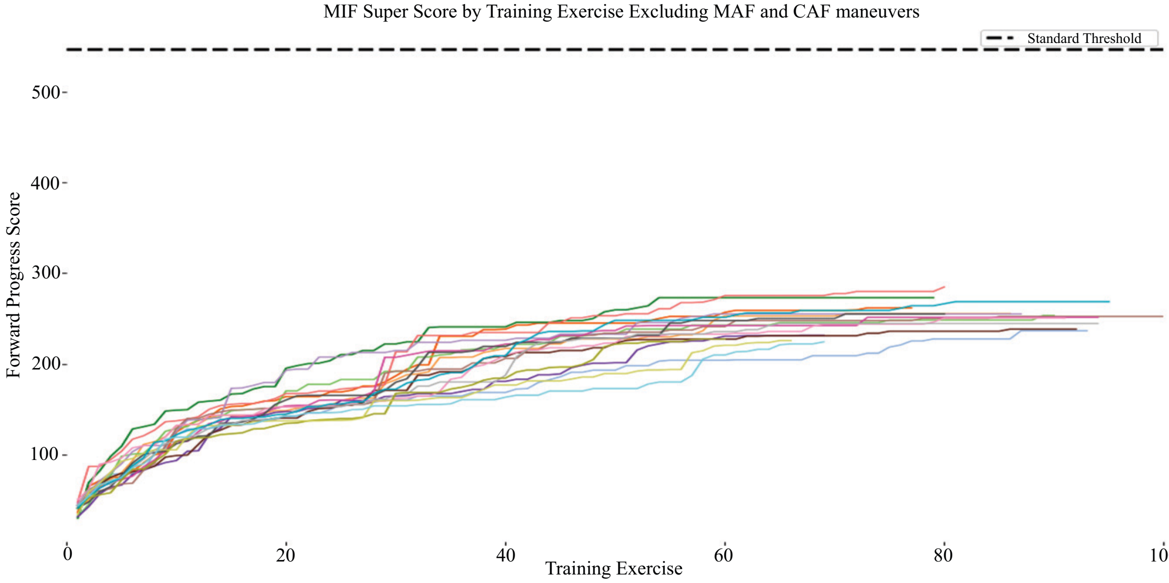
Figures 4 and 5 provide examples of what the recommender system accomplishes. These represent a particular student with an IP-generated plan, alongside

Figure 1. (Color online) Cumulative Student Performance over Time Using Cumulative MIF Metric for First PTN Class, Excluding Combat Air Force and Mobility Air Force Tracks



Note. The middle area seems to indicate progress stagnation, but is really an area where the training is focused on breadth of training.

Figure 2. (Color online) Cumulative Student Performance over Time Using FPS Metric for First PTN Class, Excluding Combat Air Force and Mobility Air Force Tracks

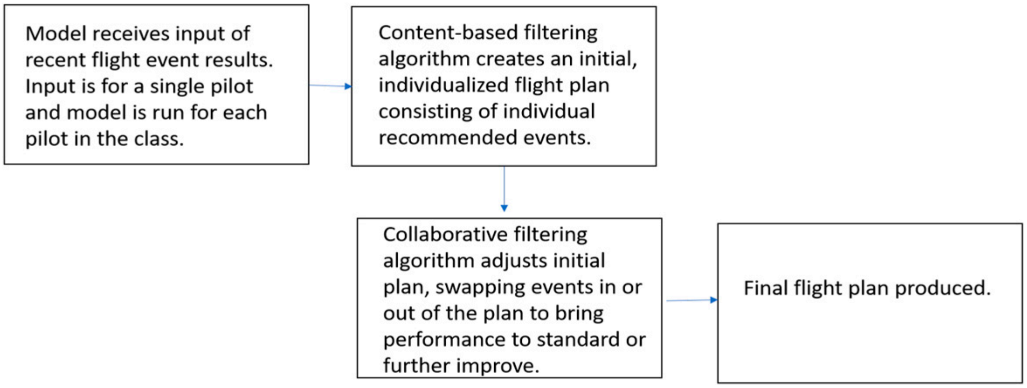


Note. The key difference here is the lack of an area of apparent progress stagnation because the FPS metric specifically considers training breadth.

the recommender-system-generated plan for exercises 3 and 21, respectively. For exercise 3, there are two preceding exercises, and for exercise 21, there are 20 preceding exercises, all of which were manually IP-generated. These preceding exercises provide the input event lists and scores used by the recommender system to generate the initial recommended plan. On the right of each figure are the IP-generated and recommender-system-generated listings for exercises 3 and 21. The agreement appears quite adequate. This is the agreement level examined in the testing of the recommender system.

The recommender-system evaluation uses two metrics, each based on a binary variable, set to zero when the IP and recommender system disagree on an event inclusion and one when there is agreement. Testing loss is the binary cross-entropy measure, whereas testing accuracy is the average value of the binary variable. Note that these are measures compiled by using the 128 possible events.

Figure 3. Overall Sequence of Events for Each Student to Develop the Final Recommended Training Plan


Model receives input of recent flight event results. Input is for a single pilot and model is run for each pilot in the class.

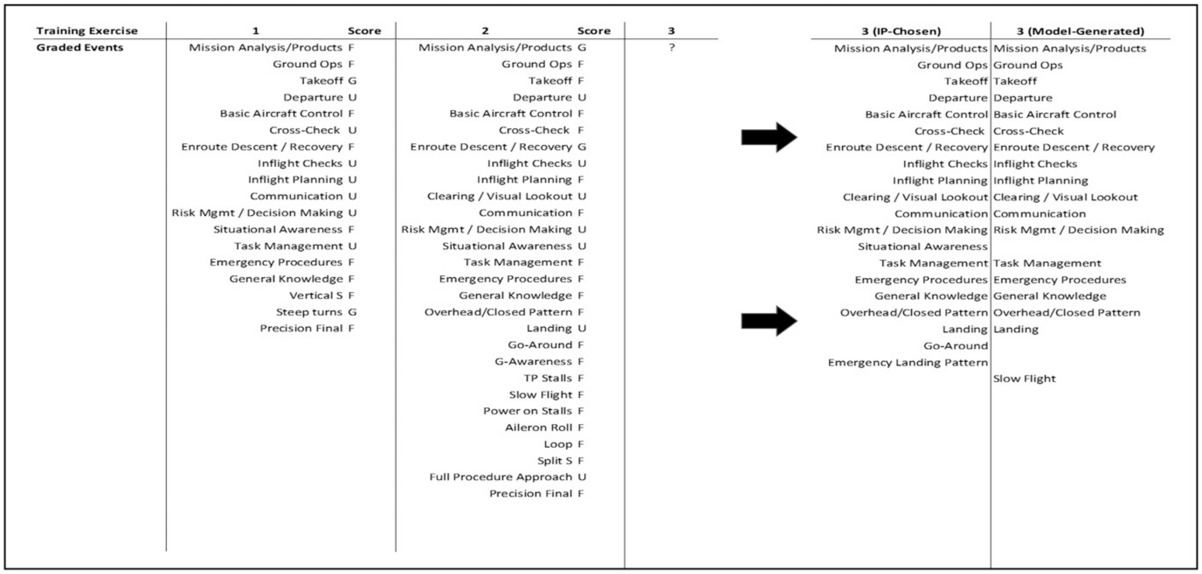
Content-based filtering algorithm creates an initial, individualized flight plan consisting of individual recommended events.

Collaborative filtering algorithm adjusts initial plan, swapping events in or out of the plan to bring performance to standard or further improve.

Final flight plan produced.

Notes. An initial individualized plan is generated by using a content-based filtering algorithm applied to the student’s past scored events. The collaborative filtering algorithm then adjusts the initial plan to promote improvement in pilot trainee efficiency and breadth.

Figure 4. IP-Chosen Event Recommendations vs. Model-generated Event Recommendations on Training Exercise 3

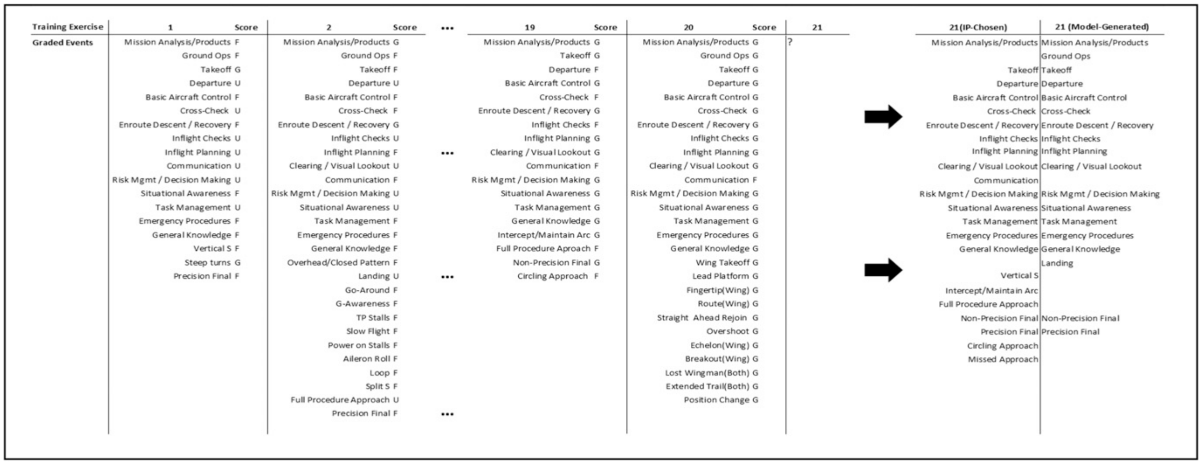


Training Exercise	1	Score	2	Score	3	3 (IP-Chosen)	3 (Model-Generated)
Graded Events	Mission Analysis/Products	F	Mission Analysis/Products	G	?	Mission Analysis/Products	Mission Analysis/Products
	Ground Ops	F	Ground Ops	F		Ground Ops	Ground Ops
	Takeoff	G	Takeoff	F		Takeoff	Takeoff
	Departure	U	Departure	U		Departure	Departure
	Basic Aircraft Control	F	Basic Aircraft Control	F		Basic Aircraft Control	Basic Aircraft Control
	Cross-Check	U	Cross-Check	F		Cross-Check	Cross-Check
	Enroute Descent / Recovery	F	Enroute Descent / Recovery	G		Enroute Descent / Recovery	Enroute Descent / Recovery
	Inflight Checks	U	Inflight Checks	U		Inflight Checks	Inflight Checks
	Inflight Planning	U	Inflight Planning	F		Inflight Planning	Inflight Planning
	Communication	U	Clearing / Visual Lookout	U		Clearing / Visual Lookout	Clearing / Visual Lookout
	Risk Mgmt / Decision Making	U	Communication	F		Communication	Communication
	Situational Awareness	F	Risk Mgmt / Decision Making	U		Risk Mgmt / Decision Making	Risk Mgmt / Decision Making
	Task Management	U	Situational Awareness	U		Situational Awareness	
	Emergency Procedures	F	Task Management	F		Task Management	Task Management
	General Knowledge	F	Emergency Procedures	F		Emergency Procedures	Emergency Procedures
	Vertical S	F	General Knowledge	F		General Knowledge	General Knowledge
	Steep turns	G	Overhead/Closed Pattern	F		Overhead/Closed Pattern	Overhead/Closed Pattern
	Precision Final	F	Landing	U		Landing	Landing
			Go-Around	F		Go-Around	
			G-Awareness	F		Emergency Landing Pattern	
			TP Stalls	F			Slow Flight
			Slow Flight	F			
			Power on Stalls	F			
			Aileron Roll	F			
			Loop	F			
			Split S	F			
			Full Procedure Approach	U			
			Precision Final	F			

Notes. The first two columns are training exercises 1 and 2. These lead to the IP-chosen set of events and the model-generated set of events. Notice Go-Around is selected by the IP, but not by the model, whereas Slow Flight is selected by the model, but not by the IP.

Table 5 depicts the full testing accomplished, examining results generated for exercises 11, 26, and 51, as based on 10, 25, and 50 preceding exercises, respectively. Overall, the results are promising, as the recommender system does well at building each exercise set, and there does not appear to be much sensitivity with respect to the number of preceding exercises required to achieve the results. However, Table 5 does not provide the full temporal picture. Figure 6 plots the testing accuracy metric as a function of training-exercise number, which represents elapsed time in PTN. These recommendations involve a preceding exercise cap of 50, meaning that the algorithm does not take into account any training that occurred outside of the 50 most recent exercises. The recommender system performs well early and late in the training programs, but appears to struggle throughout the middle of the training programs. An exact rationale for the behavior, and algorithmic corrections, is part of ongoing work and extensions as more PTN

Figure 5. IP-Chosen Event Recommendations vs. Model-generated Event Recommendations on Training Exercise 21



Training Exercise	1	Score	2	Score	•••	19	Score	20	Score	21	21(IP-Chosen)	21 (Model-Generated)
Graded Events	Mission Analysis/Products	F	Mission Analysis/Products	G		Mission Analysis/Products	G	Mission Analysis/Products	G	?	Mission Analysis/Products	Mission Analysis/Products
	Ground Ops	F	Ground Ops	F		Takeoff	G	Ground Ops	G			Ground Ops
	Takeoff	G	Takeoff	F		Departure	F	Takeoff	G		Takeoff	Takeoff
	Departure	U	Departure	U		Basic Aircraft Control	G	Departure	G		Departure	Departure
	Basic Aircraft Control	F	Basic Aircraft Control	F		Cross-Check	F	Basic Aircraft Control	G		Basic Aircraft Control	Basic Aircraft Control
	Cross-Check	U	Cross-Check	F		Enroute Descent / Recovery	G	Cross-Check	G		Cross-Check	Cross-Check
	Enroute Descent / Recovery	F	Enroute Descent / Recovery	G		Inflight Checks	F	Enroute Descent / Recovery	G		Enroute Descent / Recovery	Enroute Descent / Recovery
	Inflight Checks	U	Inflight Checks	U		Inflight Planning	G	Inflight Checks	G		Inflight Checks	Inflight Checks
	Inflight Planning	U	Inflight Planning	F	•••	Clearing / Visual Lookout	G	Inflight Planning	G		Inflight Planning	Inflight Planning
	Communication	U	Clearing / Visual Lookout	U		Communication	F	Clearing / Visual Lookout	G		Clearing / Visual Lookout	Clearing / Visual Lookout
	Risk Mgmt / Decision Making	U	Communication	F		Risk Mgmt / Decision Making	G	Communication	F		Communication	
	Situational Awareness	F	Risk Mgmt / Decision Making	U		Situational Awareness	G	Risk Mgmt / Decision Making	G		Risk Mgmt / Decision Making	Risk Mgmt / Decision Making
	Task Management	U	Situational Awareness	U		Task Management	G	Situational Awareness	G		Situational Awareness	Situational Awareness
	Emergency Procedures	F	Task Management	F		General Knowledge	G	Task Management	G		Task Management	Task Management
	General Knowledge	F	Emergency Procedures	F		Intercept/Maintain Arc	G	Emergency Procedures	G		Emergency Procedures	Emergency Procedures
	Vertical S	F	General Knowledge	F		Full Procedure Aproach	F	General Knowledge	G		General Knowledge	General Knowledge
	Steep turns	G	Overhead/Closed Pattern	F		Non-Precision Final	G	Wing Takeoff	G			Landing
	Precision Final	F	Landing	U	•••	Circling Approach	F	Lead Platform	G		Vertical S	
			Go-Around	F				Fingertip(Wing)	G		Intercept/Maintain Arc	
			G-Awareness	F				Route(Wing)	G		Full Procedure Approach	
			TP Stalls	F				Straight Ahead Rejoin	G		Non-Precision Final	Non-Precision Final
			Slow Flight	F				Overshoot	G		Precision Final	Precision Final
			Power on Stalls	F				Echelon(Wing)	G		Circling Approach	
			Aileron Roll	F				Breakout(Wing)	G		Missed Approach	
			Loop	F				Lost Wingman(Both)	G			
			Split S	F				Extended Trail(Both)	G			
			Full Procedure Approach	U				Position Change	G			
			Precision Final	F	•••							

Notes. The first four columns are training exercises 1–20. These lead to the IP-chosen set of events and the model-generated set of events. Notice Vertical S is selected by the IP, but not by the model, whereas Landing is selected by the model, but not by the IP.

Table 5. Model Testing Results

Sequence length	Testing loss	Testing accuracy
50	0.2146	0.9145
25	0.2113	0.9170
10	0.2151	0.9136

data are received. However, one conjecture is that the dip in performance is due to the introduction of breadth into the training-event selection process, meaning that additional work on the algorithm might focus on emphasizing the breadth of event selection at some point estimated as midway through a training program.

Overall, the recommender system generated more than 300 exercises based on this single PTN data set. The recommender system produced each recommendation at an average of 8.14 seconds (standard deviation of 0.10) compared with the hours that are expected for an IP to accomplish a similar task. These time savings are impressive, particularly when coupled with the demonstrated accuracy of this initial recommender system.

Conclusions and Future Work

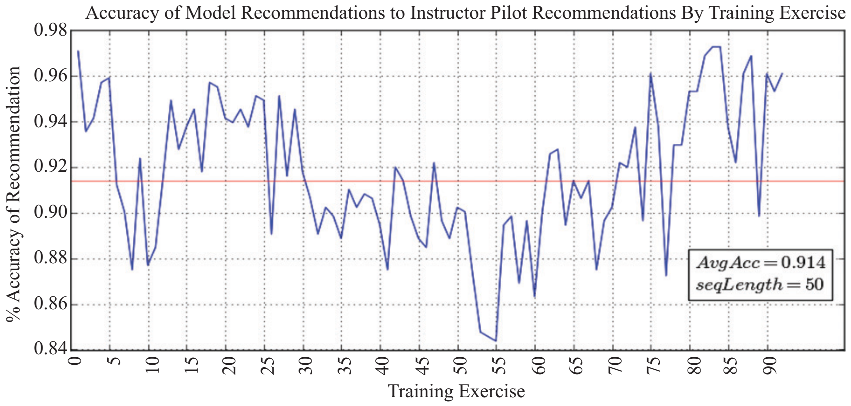
The Air Force's PTN program has had success in effectively shortening the length of UPT campaigns for pilot-training students. Operational differences from traditional undergraduate pilot training with the emphasis of virtual reality flight simulation alongside periodic real aircraft experience appear to provide a more efficient pilot-training process. Automating tedious tasks performed by IPs by using a recommender system provides an opportunity to make the pilot-training process even more efficient.

A new metric, called the FPS, was developed to better track student progress throughout UPT. Unlike current methods, FPS incorporates both breadth and depth of student skill advancement. The FPS uses progress-tracking metrics, such as proficiency of individual graded events, quantity of introduced events and event categories, and proficiency of entire event categories, to capture the multidimensional aspect of student advancement through a training campaign.

An initial recommender system employing a hybrid filtering approach using both content-based and collaborative models was shown to accurately generate a set of appropriate flight events for evaluation in a student's next training exercise. An LSTM RNN was trained using actual IP recommendations to produce initial student-specific recommendations. A collaborative filtering component improved this recommended set. Testing demonstrated the model's ability to produce flight-event recommendations, averaging 92% similarity to actual IP recommendations at a fraction of the time currently required.

The inherent uniqueness of PTN and the AutoGradebook concept open a variety of focus areas for future research. Minor parameter tuning was conducted in this study. As additional classes complete PTN and the improvements to the data-collection process that were recommended by this initial study are put into place, additional model parameters and finer levels of parameter tuning can be examined. The limited data available for this initial development effort limited algorithm choices. With more data, set-generation methods can be examined. The current approach considers events independently. Future work can define and incorporate event linkages and various IP "rules of thumb" to further improve the recommendations produced. Trainees can experience proficiency regression, requiring remedial action in the form of reintroducing previously trained events. Current research is focused on methods to identify and recommend rectifying actions for such training regression. Finally, the current algorithms focus on pushing event selection toward top

Figure 6. (Color online) Model Accuracy with Regard to Training Exercise



Note. *AvgAcc*, average accuracy; *seqLength*, sequence length used for training.

performers. This approach can be reassessed to possibly employ multiple student standards in the collaborative filtering component of the recommender system.

Acknowledgments

The authors acknowledge the tremendous support from AFWERX and the PTN program. The authors also thank the editorial staff and anonymous reviewers for their constructive comments that helped improve the content and presentation of this paper. The views expressed in this article are those of the authors and do not reflect the official policy or position of the U.S. Air Force, the U.S. Department of Defense, or the U.S. Government.

Appendix A. Mathematical Development of Forward Progress Score

Understanding FPS requires some mathematical structure. Define the sets *STUDENTS* of size S as the pilot candidates in the program, *EVENTS* of size E as the potential training events, and *CAT* of size C as the categorical grouping of events. See Table 2 for the list of sets E and C. Each student, $s \in STUDENTS$, receives an IEF score:

$$IEF_i^s = \frac{MaxScore_i^s}{MaxMIF_i} \qquad s \in STUDENTS; i \in EVENTS, \qquad \text{(A.1)}$$

where the $MaxScore_i^s$ is the highest score for student s on event i, and $MaxMIF_i$ is the highest possible score for event i. These scores are used to obtain the cumulative event points (CEPs) score based on Table A.1. The CEP is a summation of the points earned by the student for each event. These values are retained for each student as CEP_i^s.

The FPS for each student, $s \in STUDENTS$, at the end of training exercise ℓ, is then defined as:

$$FPS_{s,\ell} = \sum_{i=1}^{E} CEP_i^s + \sum_{j=1}^{C} \left[SW_j I_j^s + MW_j H_j^s \right] \qquad \forall\, s \in STUDENTS, \qquad \text{(A.2)}$$

where SW_j and MW_j are the weights of having been introduced to all events and having achieved proficiency in all events, respectively. The I_j^s and H_j^s are indicator functions pertaining to whether student s has been introduced to all events and has achieved proficiency in all events, respectively.

Table A.1. Point Allocation per Event Graded Evaluation Used in the Recommender System

IEF value	Additional points allotted	Cumulative points allotted
$IEF \geq 1$	2.25	5
$1 > IEF \geq \frac{2}{3}$	0.50	2.75
$\frac{2}{3} > IEF > \frac{1}{3}$	0.75	2.25
$\frac{1}{3} \geq IEF > 0$	1.5	1.5
$IEF = 0$	0	0

Appendix B. Metrics Used in Evaluation

Let e represent a particular event of interest (of the 128 available). Define t_e as the binary variable representing IP selection of the event (value of one for selection), p_e as the binary variable representing the recommender-system selection of the event, and b_e as the binary variable representing agreement of the IP and the recommender system (a value of one when $te = p_e$). Then,

$$\text{Testing Loss} = \frac{-1}{128} \sum_{e=1}^{128} \left[t_e \log(p_e) + (1 - t_e) \log(1 - p_e) \right], \qquad \text{(B.1)}$$

$$\text{Testing Accuracy} = \frac{-1}{128} \sum_{e=1}^{128} b_e, \qquad \text{(B.2)}$$

where the testing loss metric is a commonly used metric known as binary cross-entropy.

Appendix C. Acronyms Used

Acronym	Meaning
IP	Instructor pilot
UPT	Undergraduate pilot training
PTN	Pilot training next
k-NN	*k*-nearest neighbors
LSTM	Long short-term memory
RNN	Recurrent neural network
MIF	Maneuver item file
FPS	Forward Progress Score
CEP	Cumulative event points
IEF	Individual event fraction

References

Colbath DJ (2020) A model to analyze the capacity of pilot training production. Unpublished master's thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH.

Ekstrand MD, Riedl JT, Konstan JA (2011) Collaborative filtering recommender systems. *Foundations Trends Human-Comput. Interaction* 4(2):81–173.

Forrest NC (2020) Conceptualization and application of deep learning and applied statistics for flight plan recommendation. Unpublished master's thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH.

Koren Y (2008) Factorization meets the neighborhood: A multifaceted collaborative filtering model. *KDD'08 Proc. 14th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery), 426–434.

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Technical report, IEEE Computer Society, Washington, DC. Accessed September 15, 2020, https://data-science/data-science-repo/Recommender-Systems-[Netflix].pdf.

Korger C (2019) The road to wings: How to become a US Air Force pilot. Accessed September 15, 2020, https://sofrep.com/fightersweep/how-to-become-a-us-air-force-pilot/.

Linden G, Smith B, York J (2003) Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* 7(1):76–80.

McRae J (2017) AF addresses pilot shortage. Accessed September 15, 2020, https://www.af.mil/News/Article-Display/Article/1135435/af-addresses-pilot-shortage/.

Miller BN (1995) GroupLens: An open architecture for collaborative filtering. Technical report, Pennsylvania State College of Information, Science, and Technology, University Park, PA. Accessed September 15, 2020, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.3784&rep=rep1&type=pdf.

Ricci F, Rokach L, Shapira B, Kantor PB, eds. (2011) *Recommender Systems Handbook* (Springer Science+Business Media, LLC, New York).

Sanghavi B, Rathod R, Mistry D (2014) Recommender systems—comparison of content-based filtering and collaborative filtering. *Internat. J. Curr. Engrg. Tech.*. 4(5):3131–3133.

Srivastava A, Bala PK, Kumar B (2019) New perspectives on gray sheep behavior in E-commerce recommendations. *J. Retailing Consumer Services* 53:1–11.

Xie J, Wang Q (2018) Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge. *CEUR Workshop Proc.* 2148:97–102.

Verification Letter

Lt. Col. Robert M. Knapp, Director of Operations, Detachment 24, Department of the Air Force, Nineteenth Air Force, Joint Base San Antonio-Randolph, San Antonio, Texas 78234, writes:

"This letter serves to verify that An Air Force Pilot Training Recommendation System using Advanced Analytical Methods and its associated scripts and models were applied to the next generation of Pilot Training Next (PTN) data with some code refactoring. The resulting models were found to be accurate using the next iteration of student data, with some degradation related to novel tasks. The approach demonstrated in An Air Force Pilot Training Recommendation System Using Advanced Analytical Methods resulted in a requirement for a similar recommendation system to be built in the follow-on pilot training program through the Defense Innovation Unit."

Nicholas C. Forrest is an active-duty First Lieutenant in the U.S. Air Force. He completed his MS in operations research in March 2020 and continues to explore aspects of machine learning and artificial intelligence in his daily Air Force analytical duties.

Raymond R. Hill is a professor of operations research in the Department of Operational Sciences at the Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio. He holds a PhD in industrial and systems engineering and has research interests in applied statistics, simulation, and defense analytics.

Phillip R. Jenkins is an assistant professor of operations research in the Department of Operational Sciences at the Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio. He holds a PhD in operations research and has research interests in all aspects of probabilistic modeling, machine learning, and artificial intelligence.