accenture

strategy | consulting | digital | technology | operations

STP 8-A-06

AI DEMAND FORECASTING

FINAL REPORT

Defense Logistics Agency
R&D Weapons System Sustainment Program
September 2019

Contents

GLOSSARY

AAC	Acquisition Advice Code
Absolute Dollar Error	A metric for forecast performance that takes the dollar value of the demand forecast error in a given period
ACE	DLA's Analytics Center for Excellence
ADF	Annual Demand Frequency
AI	Artificial Intelligence
AIDF	Artificial Intelligence Demand Forecasting
APFE	Absolute Percent Forecast Error; one of DLA's current metrics for forecast performance
ATO	Authorization to Operate
CD	Customer Direct; sales channel
CIT	Customer Item Type
CNN-LSTM Network	Convolutional Neural Network—Long Short-Term Network; an algorithmic model used for demand forecasting
COPE	DLA's Center of Planning Excellence
CPU	Central Processing Unit
CSP	Cloud Service Provider
DD	DLA Direct; sales channel
DevSecOps	Development, Security, and Operations; a framework that interlocks security into development and operations
DFU	Demand Forecast Unit; uniquely defines a customer-item relationship for forecasting and is comprised of a demand unit, demand group, location, and forecast model type
DHDT	Demand History Detail Table; data archive that is updated once sales order data is transmitted to JDA (excluding multiple order document types and Non-CLSSA foreign military sales orders)
DME	Demand Month End; concludes the monthly demand planning cycle and therefore, governs major demand planning activities
DMS	Demand Month Start; begins the demand planning cycle and therefore, governs major demand planning activities
DoDAAC	Department of Defense Activity Address Code
DORRA	DLA Office of Operations Research and Resource Analysis
DPA	Demand Plan Accuracy
dSRS	Disproportionate Stratified Random Sampling; sampling technique in which the number of items sampled from each stratum is not proportional to their representation in the total item population
DSVM	Data Science Virtual Machine

EBS	DLA's Enterprise Business Systems
EDW	Enterprise Data Warehouse
Ensemble Modeling	A machine learning method that allows multiple machine learning techniques to be combined into a single resultant forecast
EWMA	Exponentially Weighted Moving Average; an algorithmic model used for demand forecasting
Feature Engineering	A technique used in data preparation that creates new variables that are used by algorithms to generate forecasts
FedRAMP	Federal Risk and Authorization Management Program
FIPS	Federal Information Processing Standards
FISMA	Federal Information Security Management Act
GFE	Government Furnished Equipment
GPU	Graphics Processing Unit
HIST	History Table; data archive that JDA updates during DMS and uses for statistical forecasting (excluding non-forecastable and MTS kit items)
Hyperparameters	A model's unique architecture parameters (e.g., smoothing factors, histororical time period for training)
IL	Impact Level; standardized security levels that are mapped to types of information and information systems across agencies
Input Data Scaling	A technique used in data preparation that adjusts the scales of the input data to speed up the machine learning process
ITCR	IT Change Request
J6T	DLA's Strategic Technology Team
JDA	Vendor for software and supply chain management software currently generating DLA's forecasts
KPI	Key Performance Indicator
LOE	Lines of Effort
LSTM Network	Long Short-Term Memory Network; an algorithmic model used for demand forecasting
MAE	Mean Absolute Error; a metric for forecast error to measure improvement to forecast accuracy
MAPE	Mean Absolute Percent Error; a metric for forecast error to measure improvement to forecast accuracy
MASE	Mean Absolute Scaled Error; a metric for forecast error to measure improvement to forecast accuracy
ML	Machine Learning
MLT	Manufacturing Lead Time
MRO	Maintenance, Repair, and Overhaul; the three main factors that drive the collaborative forecast
MSE	Mean Squared Error; a metric for forecast error to measure improvement to forecast accuracy

MTS	Make-to-stock
NIST	The National Institute of Standards and Technology
Non-CLSSA	Non-Cooperative Logistics Supply Support Arrangement
OA	Obligation Authority
PFE	Percent Forecast Error; one of DLA's current metrics for forecast performance
PLANMO	Planning Analysis Model; a SAS-based predictive analytics engine used by COPE and designed to inform strategic decision-making at DLA by projecting the outcomes of various planning policies
PR	Purchase Request
SAP ECC System	SAP Enterprise Central Component System; where orders are created in EBS
SCPO	JDA's Supply Chain Planning Optimization; tools used by DLA's Enterprise Business Systems to support current Demand and Supply Planning processes
SE	Standard Error; measures the statistical variability over which a sample distribution is representative of a population.
sMAPE	Symmetric Mean Absolute Percent Error; a metric for forecast error to measure improvement to forecast accuracy
SPR	Special Program Requests
SSR	Special Supply Requests
STIG	Security Technical Implementation Guide; used as a verb meaning to make compliant with the STIG
STP	Short Term Project
TCN	Temporal Convolutional Network; algorithmic neural network-based model used for demand forecasting
TSB	Teunter-Syntetos-Babai; an algorithmic model used for demand forecasting that is a variant of Croston's method
VDI	Virtual Desktop Infrastructure
VM	Virtual Machine
WAT	Worldwide Activity Test
What-If	A production-sized JDA environment owned by COPE and currently used for planning process experimentation
Winsorization	A technique used in data preparation that adjusts outliers to limit the impact of extreme history quantities
Winsorized Average	Algorithmic model used for demand forecasting
WSSP	DLA's Weapons System Sustainment R&D Program

EXECUTIVE SUMMARY

DLA's mission to globally support Warfighter demands across nine supply chains is complex. There are no commercial supply chain equivalents that face the dynamic, competing, and urgent needs of its customers. With the current budget climate presenting a challenge to balance weapon system support with fiscal constraints, the need to improve the prediction accuracy of DLA customer demands is critical to effective and efficient service to our Military Services.

Problem

Demand Forecasting is a critical function of DLA's business – the demand forecast is the input to the Agency's downstream supply chain processes that affect customer service.[1] In other words, the accuracy of demand forecasts plays a critical role in DLA's ability to support the Warfighter supply requirements. Under-forecasting can lead to low materiel availability and stockouts, while over-forecasting can lead to cashflow problems and excess inventory. DLA is challenged with both, so the need to explore innovative ways to improve demand forecast accuracy has become a core objective of the DLA Director's Strategic Guidance 2018-2026.

Project Summary

The DLA WSSP R&D office tasked Accenture with the Artificial Intelligence (AI) Demand Forecasting (AIDF) short-term project (STP) to explore the potential of leveraging emergent technology for improving DLA's ability to predict customer demands. The project also assessed DLA's technology environment for AI-based forecasting solutions as a critical component to future production enablement. This STP produced a proof-of-concept in collaboration with the WSSP R&D office, the J34 Center of Planning Excellence (COPE), the Analytics Center of Excellence (ACE), and the Strategic Technology Team (J6T) over a 9-month period, concluding 9/30/2019.

Key Accomplishments & Findings

The key accomplishments and findings of the AIDF STP can be summarized in four main points. Each of these points reflects the results or claims of the R&D study.

1. **Applied AI in Two Ways to Address Challenges:** (1) Developed new AI-based forecast algorithms not in use in the current JDA solution and (2) Applied AI to select and combine up to nine individual forecast models to create an item-specific model.
2. **AI Demand Forecasting Shows Improvement:** This two-fold application of AI demonstrated a **$102M annual reduction in over-forecast error** for the 48k item sample evaluated, without increasing the risk of an under-forecast error.
3. **There is no Universal Solution:** The evaluated AI-powered models were unable to improve accuracy for the sample population of items with extremely sparse demand. This finding supports DLA's current minimum threshold logic for item forecastability.
4. **Scaling AI Needs a Capable Environment:** AI models were developed on offline government laptops with modern data science software. Scalable AI development requires more robust software, hardware, and data pipelines.
5. **There is Value in Alternative Forecast Methods:** Several methods, including simple, non-AI forecasting methods, demonstrated potential improvement relative to current forecasting methods.

[1] Details of forecast impacts on DLA business operations found in the Final Report Introduction

Additional Observations

The AIDF STP demonstrated the potential value to DLA business operations when applying AI to demand forecasting. As part of any R&D effort it's important to highlight observations and challenges discovered during the STP.

1. **AI for Planning can be Targeted:** AI forecasting represents another capability in the Planning toolbox of strategies and can augment demand planner decisions.
2. **Measuring Forecast Accuracy has Limitations:** Forecast metrics are unable to measure sparse demand items accurately. Controlled tests demonstrated the inherent inability of metrics to accurately measure items with less than five days of demand.
3. **Cross-Enterprise Collaboration is Necessary:** AI requires technology, analytic, business function, and change management expertise for successful transformation.
4. **AI Implementation Requires Investment:** For long-term sustainment, additional costs for maintenance and operations plus workforce upskilling must be considered.

Recommendations

The AIDF STP revealed that the application of AI for demand forecasting can yield significant business value to DLA Planning.

Accenture Recommends
1. Develop Prototype to apply AI research to demand forecasting (multiple options) 2. Create infrastructure to support scalable AI Environment 3. Reexamine Enterprise forecast metrics 4. Explore alternative methods for hard-to-forecast items

There are several model options for advancing the proof-of-concept from this STP to a prototype as recommended above. The development and trade-off analysis details are in the **IV. Analysis of Courses of Action (COAs)** section of the Final Report.

Final Report Overview

The AIDF Final Report includes the comprehensive technical details of the research executed during this R&D project. The following table provides a guide for the AIDF final report.

Section	What's Contained in the Section?
I. Introduction	Describes DLA's current environment and project goals
II. Background	Provides initial review of DLA forecasting, challenges, and existing Planning research tools
III. Approach & Findings	Explains methods to address each element of scope (including challenges, assumptions, and procedures), and reviews the analysis performed to determine findings
IV. COA Analysis	For multiple COAs, analyzes business impact, risks, and considerations for next steps to advance the research performed by this STP
V. Conclusions	Summarizes major conclusions and recommends next steps to advance the progress completed by this STP
VI. Appendices	Provides detailed assessments, additional results, and model documentation

I. INTRODUCTION

As the nation's combat logistics support agency, DLA manages a global supply chain that plans, procures, stores, and ships repair parts and supplies for the military services, combatant commands, other partnered federal agencies, and allied nations. As shown in Figure 1, demand planning's primary purpose is to create demand forecasts. These forecasts drive downstream business functions that support the execution of DLA's mission.

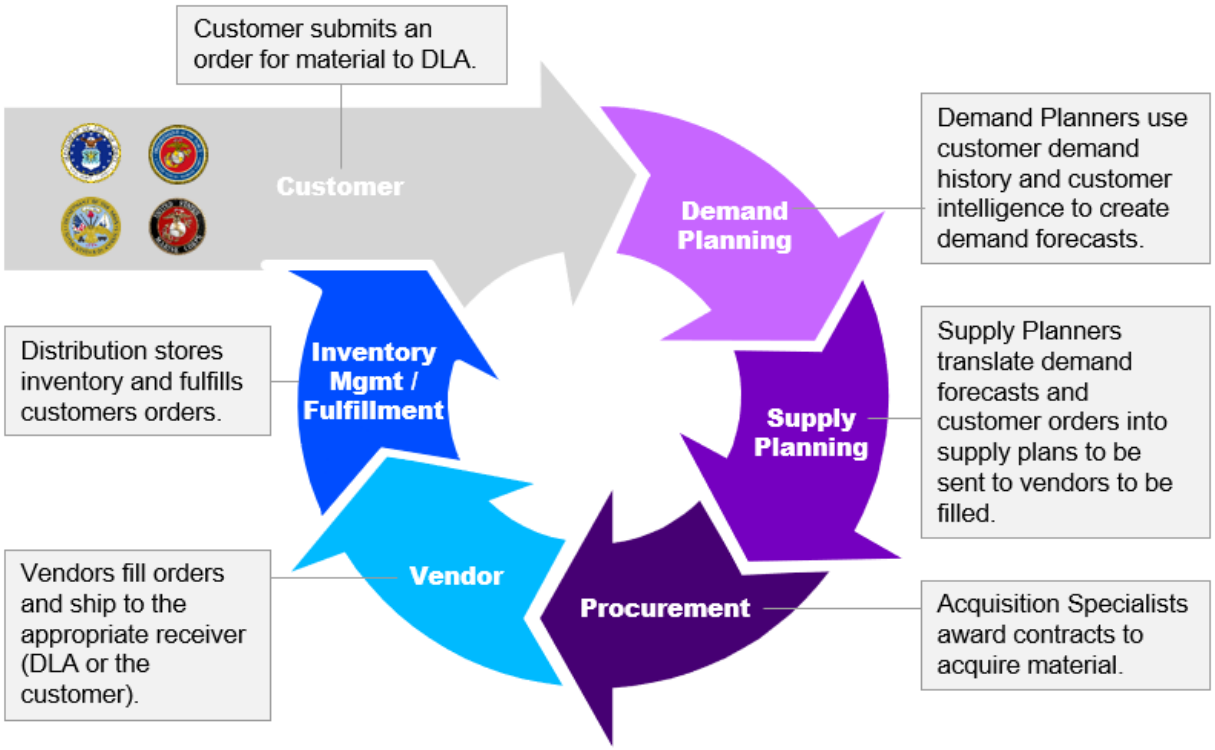

Figure 1: High-Level overview of DLA supply chain processes

The demand forecast drives DLA's core supply chain functions because it is the primary input to the Supply Planning process, which seeks to optimize inventory to meet customer requirements. Supply Planning generates purchase requests (PRs) that are solicited and awarded by Procurement to vendors, who in turn supply the materials. DLA Distribution receipts, stores, and manages the inventory that is aligned to customer orders through order fulfillment. Customer usage data is then fed back to the Demand Planning organization to update future forecasts. DLA's Enterprise Business Systems (EBS) uses JDA Supply Chain Planning Optimization (SCPO) and JDA Collaboration software to support Demand and Supply Planning processes.

Project Goals

The objective of this project was to research the ability of AI/ML to improve DLA's demand forecast accuracy through the development of a proof-of-concept. As the primary input, the demand forecast is critical to the downstream efficiency of DLA's supply chain. Accurate demand forecasts are crucial to providing maximum customer service at the lowest cost to the provider. However, DLA's forecast accuracy often falls short of leadership expectations and can negatively impact the Agency's mission to support Warfighter readiness and weapon system uptime at the expected

cost-to-serve. Given the importance of demand forecast accuracy to DLA's Mission and Strategic Objectives, the R&D Office outlined scope to address two primary questions:

1. Can modern AI/ML algorithms improve DLA's forecast accuracy?
2. How can the Agency efficiently launch AI prototyping projects?

Fundamentally, machine learning techniques leverage advanced algorithms that allow computers to uncover patterns in data, which in turn enable them to make predictions.[2] This a direct analogy to supply chain forecasting based on historical demand, which makes AI solutions built on ML algorithms strong candidates for optimizing forecast accuracy. Indeed, advanced AI/ML techniques for forecasting have recently been growing in commercial use.[3] However, the specialized nature of DLA's business necessitates thorough evaluation of these advanced techniques to determine whether similar accuracy improvements can be achieved.

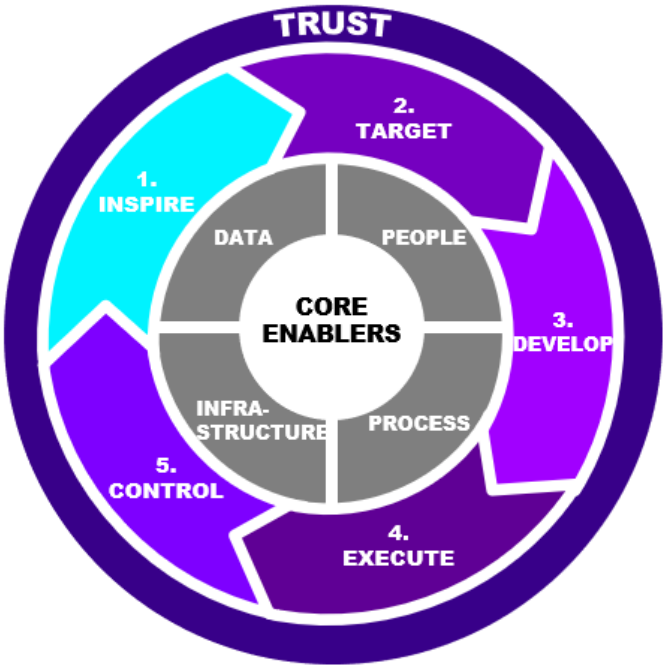

Figure 2: The AI Creation Cycle, as introduced in the R&D AI Groundwork project (STP 8-A-05).

Prototyping specific AI use cases is one of the most efficient methods for an organization to identify and address constraints and/or bottlenecks within AI solution development.[4] As one of the earliest AI use case prototype projects at DLA, AIDF aims to build on the R&D AI Groundwork project (STP 9-A-01) by identifying constraints to the four core enablers as shown in Figure 2:

- **Data**: Identification and classification of security level of data elements to review cleanliness and quality of demand history, JDA forecast, weapon system hierarchy, and item characteristic data for AI Modeling
- **People**: Co-creation and evaluation of AI technology with a cross-functional team (R&D, COPE, ACE, J6T) to understand scientific approach and business impacts
- **Process**: Navigation of unknown processes to create the AI Model Training environment including data classification, Cloud infrastructure build, IT Change Request (ITCR) submission for software, and data transmission
- **Infrastructure**: Development of dedicated AI Model Training Environment to address gaps in the existing hardware and software for development projects

[2] For a primer on AI/ML, see the **AI Awareness Training** developed by the DLA R&D Office in coordination with Accenture, a copy of which can be obtained from Mr. Manuel Vengua (manuel.m.vengua@dla.mil)
[3] https://www.datanami.com/2019/03/22/how-walmart-uses-gpus-for-better-demand-forecasting/
https://www.winsightgrocerybusiness.com/technology/how-grocers-are-reimagining-future-ai
[4] See the **AI Use Case Playbook**, also available via Mr. Manuel Vengua of the DLA R&D Office

II. BACKGROUND

This section defines DLA's current forecasting methods, identifies unique challenges that are a byproduct of the Agency's Mission, and discusses DLA's existing analytics tools and environments. Collectively, a high-level review of these topics provides essential context for understanding the AIDF project team's approach to improving DLA's forecast accuracy.

Forecasting Introduction

Forecasting is the process of predicting future demand based on historical data combined with expert judgement on the current environment and/or external influences. DLA currently relies on traditional demand forecasting techniques. The Agency's approach combines its unique business rules with two sources of intelligence: demand history and customer collaborative input.

Statistical forecasting uses algorithms to analyze demand history for patterns in order to project future requirements. As shown in Figure 3, demand history is captured in monthly buckets. Using this history, a new statistical forecast ("stat forecast") is created each month and is organized into monthly time periods.

Collaborative forecasting leverages requirements submitted by DLA's customers – for example, Navy shipyards and Army depots – to develop the future demand projection. The collaborative forecast ("collab forecast") is generally driven by the customer's maintenance, repair, and overhaul (MRO) schedule. However, stat forecasts are still generated for collaborative customers. Comparisons between the stat and collab forecasts, as well as collab inputs across submission periods for an item facilitate a conversation between DLA Planners and the customer to reach a consensus on the future requirements.

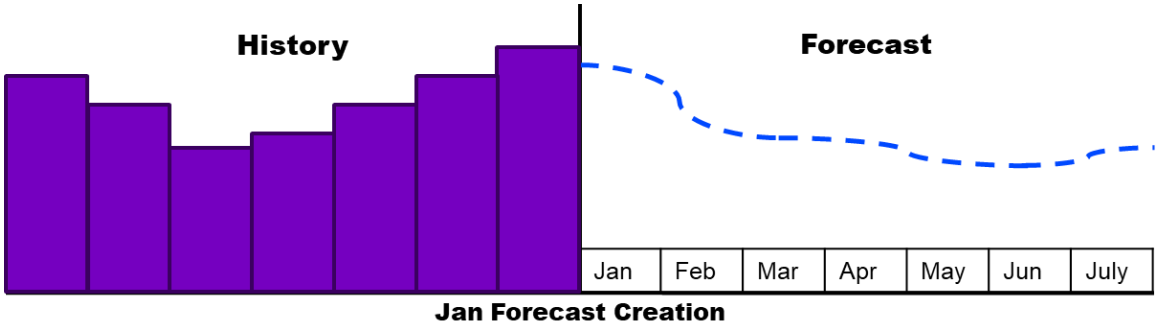

Figure 3: Notional view of statistical forecasting at DLA

DLA-Specific Challenges

DLA manages approximately 5 million items across nine supply chains – supplying 86% of the military's spare parts across more than 2,400 weapon systems. In addition to providing supply chain support for all five branches of the armed services, DLA also supports 10 combatant commands, other federal agencies, and partner and allied nations, including support for disaster response and humanitarian relief efforts both at home and abroad.[5]

The scale of its global supply chain and diversity of item classes create forecasting challenges unique to DLA. Beyond the variety and vastness of its item catalog, long lead times and highly

[5] https://www.dla.mil/AtaGlance/

variable demand combined with a need to maintain inventory critical to Warfighter readiness results in a problem space that is largely unaddressed by commercial supply chains at a comparable scale to DLA.

Diversity of Items

DLA's nine supply chains cover food, clothing, textiles, medical supplies, construction materials, fuel, repair parts, and weapon system consumable end items. The diversity across DLA's item catalog creates a corresponding diversity in customer requirements, demand patterns (seasonality, frequency, etc.), and lead times.

The practical consequence of item diversity for an end-item distributor is the need for many forecast model types, approaches, and expertise to drive consistent forecast accuracy across as many items as possible. For example, the forecast requirements for slow-moving, yet highly important engine parts differ drastically from high-demand, yet prone-to-obsolescence items such as uniforms with operational camouflage changes.

Application of Forecasting Algorithms Options

The JDA planning system currently incorporates six demand forecasting algorithms: Lewandowski, Holt-Winters, Croston, AVS-Graves, Moving Average, and Fourier. While these models are common in commercial organizations, the algorithm selection process is difficult, particularly for DLA's diverse product portfolio.

The as-is forecasting environment uses the heuristically determined rules for model selection (i.e., JDA Demand Classification) that results in top-down groupings of items. Using this approach, items with highly unique demand patterns could be incorrectly classified into poorly fit algorithms available which ultimately harm performance.

The AIDF STP approach uses AI to weight multiple demand forecasts into an item-specialized composite forecast. This is a bottom up-approach, where an item's individual characteristics and demand patterns dictate the forecast model creation.

Sparsity / Variability of Demand

To succeed in its mission, DLA must plan for and stock items whose demand is generally dominated by the Armed services. Shifts in military mission requirements often lead to extremely variable demand and/or sparse demand. Currently, large inventory investments must be made to account for the uncertainty, increasing DLA's risk to never realize demand for those items. Excess inventory constrains the obligation authority available, thereby reducing flexibility for DLA to respond to shifts in demand.

Long Lead Times

Spare parts for military weapon systems are often extremely specialized. As a result, the manufacturing lead times (MLTs) are quite long for a large proportion of the repairable parts managed by DLA – in some cases, years. The item subset population studied in this report had a median MLT of 164 days with the 75% percentile at 235 days.

As an item's lead time increases, so too does the time horizon over which DLA must forecast the item's demand: if an item takes 1 month to award and 6 months to receive from a supplier, then DLA must predict requirements 7 months into the future. Demand that hits within the 6-month window cannot be filled through a new order.

As this forecast horizon increases, the likelihood that demand will increase or decrease within the time period also increases. Historical patterns make future predictions possible, so unexpected spikes or drop-offs in demand within an ordering window can lead to stockouts or excess inventory, respectively. Due to these issues, the analysis reviewed forecast accuracy over a 12-month period, implying that forecast over lead time performance will improve as well.

Specialized Mission & Customer Base

DLA maintains a large, complex supply chain for a diverse array of weapon systems. While the Agency must forecast for a significant number of global locations (DoDAACs), the customer base associated with those locations is comparatively small. A small customer base further reduces agility in the face of changing requirements, because demand that does not materialize where originally expected is not likely to be needed elsewhere. Nevertheless, DLA must order and stock items of highly variable demand to ensure materiel availability in support of Warfighter readiness.

Accuracy in forecasting is ultimately driven by the quality and quantity of information available and the time that predictions are made. The challenges described above all tie back to one or both of these points, and their combination forms a forecasting environment that is completely unique to DLA.

DLA Planning Operations Research

In FY17, DLA established the Center of Planning Excellence (COPE) to oversee DLA Planning processes, drive efficiency through analytics, and share best practices with the supply chains. COPE segmented the demand planning responsibility into two pieces: statistical forecasting, (a COPE responsibility), and collaborative planning (which remains within the individual supply chains). COPE is responsible for innovation and analysis of planning improvement efforts and are project stakeholders for this STP.

COPE owns two distinct tools for the analysis of planning improvements. The first is **What-If**, which is a production-size JDA environment used for experimentation. Using What-If, tactical changes to JDA batch jobs and JDA functionality can be analyzed; however, analysis is slow and requires execution of standard or modified batch jobs. What-If is an excellent capability for final configuration analysis before implementing a new process, but it is not designed for rapid experimentation or AI/ML modelling. For rapid analysis, COPE created the Planning Analysis Model (**PLANMO**), a SAS-based predictive analytics engine designed to inform strategic decision-making at DLA by projecting the outcomes of various planning policies. PLANMO simulates DLA planning and procurement processes, however, is not a forecast evaluation tool.

To build and evaluate new forecast models using AI/ML, DLA needs a dedicated AI model training environment. While initial steps toward the creation of an AI environment were taken during the AI Groundwork STP, the AIDF leveraged a short-term workaround in its research to understand how to use AI to develop new models as well as to select or combine multiple demand forecast models for an item-specific forecast based on the unique demand history and item characteristics.

The outcomes of this STP may lead to another tool or medium for COPE to improve planning performance, powered by AI. Accenture partnered with DLA to form a foundation for AI proof-of-concept research and develop a DLA-tailored approach to demand forecasting. The following sections will describe the approach, results, and conclusions of the exploration of AI for demand forecasting.

III. PROJECT APPROACH & RESULTS

In response to demand forecasting challenges faced by DLA, the AI Demand Forecasting STP established three core principles to guide the project approach:

I. **Partner with DLA to develop a foundation for demand forecasting research.**
   Rather than fit an existing solution to a DLA challenge or build a prototype outside of DLA systems, the AIDF team set out to study DLA-centric approaches. Central to this theme is the proper classification, use, and handling of DLA data, the DLA-focused development of an AI Model Training Environment, and the introduction of modern data science tools to DLA's capabilities.

II. **Challenge and Test simplifying forecasting assumptions (heuristics).**
   The scale of DLA's business required simplifications in the era of constrained computation. With the advent of Cloud and scalable computing power, the AIDF team set out to challenge simplifying assumptions. This approach allows the data and uniqueness of each item to guide the modeling approach for a more robust assessment of best fit forecasting techniques at the individual item level.

III. **Assess modern AI forecast models.**
   DLA's implementation of JDA uses traditional forecast models. Emerging research from the past 20 years have demonstrated forecast improvements using machine learning based approaches.[6] The project set out to research and evaluate modern forecast algorithms' applicability to DLA's business challenges.

Using these three guiding principles, the project developed five workstreams to accomplish the STP scope.

1. **AI Demand Forecasting Environment:** Establishing a technical environment to perform the AI modeling research.
2. **Item Subset Selection:** Targeting items in DLA's portfolio to balance computation needs with statistical significance.
3. **Metrics Evaluation:** Studying forecast metrics to optimize AI models and understand the biases of metrics during interpretation.
4. **Individual Forecast Algorithm Development:** Developing simple to complex forecast algorithms.
5. **AI-Powered Model Selection:** Using AI to select, weight, and combine multiple forecasting models into an Ensemble model.

The following sections will review the methods, techniques, results, and outcomes for each workstream.

[6] Ansen, J.V.H., McDonald, J.B., and Nelson, R.D. (1999). Time Series Predication with Genetic-Algorithm Designed Neural Networks: An Empirical Comparison with Modern Statistical Models, Computational Intelligence, Volume 15, Issue 3, 171-184

AI Demand Forecasting Environment

An AI Model Training Environment is comprised of four critical components: (1) Security, (2) Data Transmission, (3) Computational Power, and (4) Data Science Tools.[7] The security requirements are defined by the type of system – Development/Testing (Dev) vs. Production (Prod) – and the sensitivity of the data. Typically, Dev systems use synthetic or aged data, allowing for a reduction in security requirements. AI models learn patterns in data that are used for future predictions and therefore need to use real data extracts from Prod systems.

Approach

The first step towards building an AI training environment for demand forecasting is to classify the data sensitivity. The AIDF team provided recommendations in the Federal Information Processing Standards (FIPS) 199 form to the DLA CDO for sign-off by applying the guidance from the NIST Special Publication 800-60 to the data required. Additionally, the AIDF team provided recommendations for the DISA Impact Level (IL) for the data requested for AI/ML modeling. The AIDF team worked in conjunction with the R&D, J6T, and ACE offices to obtain approval for the data to reside in an IL-4 environment in order to utilize DLA's development environment systems.

Once security requirements of the to-be system were established, the AIDF team identified multiple options that met minimum computational requirements to act as an AI Model Training Environment, including both on-premise and Cloud environments. The four options are outlined in Figure 4. The team then simultaneously pursued the creation and authorization to use each option as the AIDF training environment.

Option	Description	Status
Air-gapped GPU Laptop	Utilize DLA provision laptops disconnected from the Internet to manually load data for model development and evaluation	• Only approved solution
Big Data and Analytics Stack	Leverage DLA's Big Data (Hortonworks) and Analytics (SAS) environments	• Configuration focused on data storage • Limited AI/ML tools
DLA Azure	Create a new Cloud-based environment in DLA's Development Cloud Environment	• Awaiting approval to build and provision
Accenture Hosted Cloud	Accenture managed environment requiring DLA Authorization To Operate (ATO)	• Not feasible in the timeline for R&D

Figure 4: AI Environment Options Overview

DLA's existing data science tool of choice is SAS 9.4 including SAS/STAT. While SAS can produce many statistical analyses and a sub-selection of forecast algorithms, more sophisticated algorithms require additional licenses. For the speed of R&D STPs, procuring, installing, and configuring Enterprise SAS software was not possible. Thus, the AIDF team sought alternative solutions in the open source community which replicate or exceed the performance of the SAS-based options yet still minimize cost and timeline.

[7] See the **AI Use Case Playbook**, available via Mr. Manuel Vengua of the DLA R&D Office (manuel.m.vengua@dla.mil)

The Accenture team, in conjunction with the DLA Analytics Center for Excellence (ACE) and the R&D office explored the use of open source data science tools – Python and R. Additionally, package and environment managers exist to optimize and streamline the use of data science libraries. Anaconda is the world's most popular data science platform and the foundation of modern machine learning.[8] The AIDF team collaborated with the R&D office and J6T on the submission of four IT Change Requests (ITCRs) through the front door process. The first requested Anaconda, the second requested 35 Python libraries, including dependencies, to enable AI development work to begin in a DADE environment. The third and fourth ITCRs requested Nvidia toolkits and drivers which enable the use of GPUs when performing AI research.

After developing detailed approaches to meet the four components of an AI Model Training Environment, the AIDF team designed a process for forecast algorithm development and evaluation. Rather than building one-off models with custom code for data pipelines, training, and metrics that would have limited reusability, the team standardized the process with modules to allow for more rapid and consistent results.

AI Environment Results

As a result of the ITCR submissions, the Anaconda Distribution has been provisionally approved for air-gapped, stand-alone environments. Additionally, the core Python libraries and Nvidia drivers are approved for development environments with the condition that all libraries are kept up to date.

The AIDF project successfully demonstrated the use of an offline (i.e., air-gapped) GPU laptop with Anaconda software.[9] The environment setup and solution involved the use of two Government Furnished Equipment laptops (GFEs) per developer—one to access data and one to process the data with analysis software in an air-gapped system that adheres to the security requirements of the needed software, as shown in Figure 5. The air-gapped laptop, loaded with both the data and tools, conducted all data science activities related to generating and assessing demand forecasting techniques within DLA. However, this approach does not scale well and should be limited to a small team as an interim solution. Additional technical details for AI development and alternative options are outlined in Appendix B.

[8] anaconda.com/distribution/

[9] Note: The workaround GPU laptops were provisioned to Accenture on June 6, 2019 – six months after the PoP start date, i.e., at the end of the original period of performance – therefore, Accenture agreed to a no-cost project extension to allow for the focus to shift back to its original intent of AI proof-of-concept development. The laptops took approximately two weeks to properly configure and load with the necessary software and data. Modeling was able to begin on June 18, 2019.

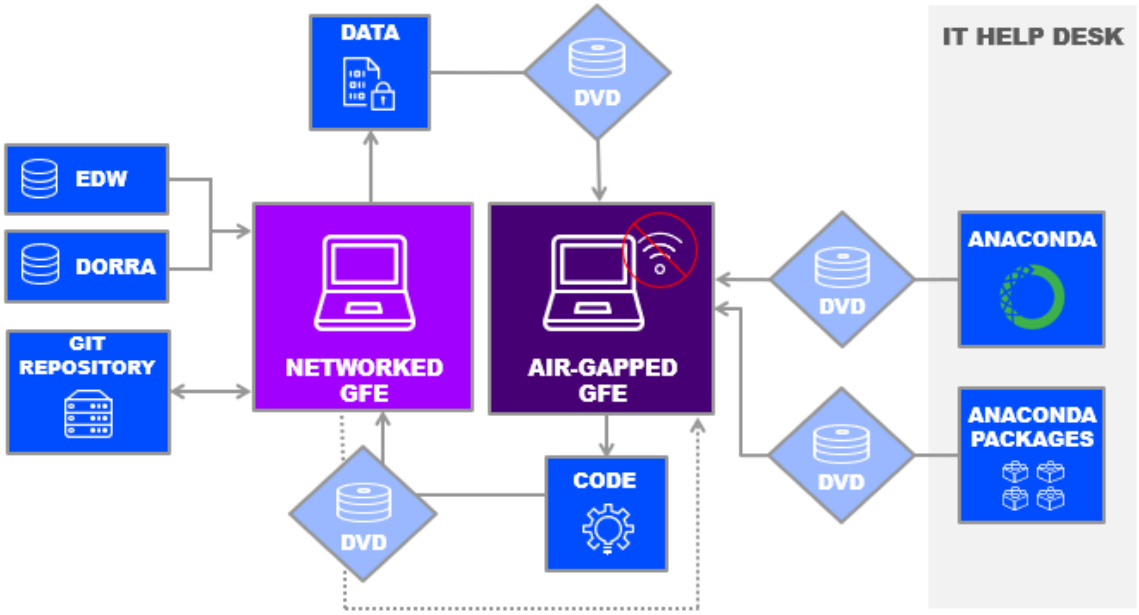

Figure 5: AIDF STP Offline GPU Laptop Planning Operations Research Environment.

Forecast Evaluation Framework

To rapidly develop and consistently evaluate modelling approaches within the AIDF project, a modular, reusable framework was created. All code will be provided to DLA for on-going research and analysis. This system consisted of four primary components:

1. **Data Import:** The data import process is handled through a function that converts raw data into usable model inputs. This function uses explicit definitions of Dev/Test data to ensure that no information from the wrong time period is used, while also forcing modularity at the touchpoints between data import and preprocessing.
2. **Preprocessing:** Preprocessing varies by item and is controlled primarily by consistent functions to enable uniformity and reusability of preprocessing techniques.
3. **Algorithm Implementation:** The implementation of each algorithm is required to have similar input processes and generate a consistent version of the forecast as an output. This enables the output of any given algorithm to be assessed interchangeably.
4. **Results Evaluation:** Algorithmic outputs were designed to have the same format so that forecast models could be rapidly evaluated and compared. This was necessary when performing hyperparameter tuning across many models to enable direct comparison, independent of model type.

Item Subset Selection

DLA has more than 5 million items in its catalog.[10] The WSSP R&D Office had the foresight to recognize that developing an AI Demand Forecasting proof-of-concept at this scale is impractical, which is why Task 6 of the Statement of Work (SOW) stated:

> *"the contractor will select a random sampling of stocked items, to limit the scope of the model development for the proof of concept. The selected stock items will not be limited to those currently forecasted (by JDA or other methods). The contractor will obtain approval from the working group prior to using the samples for the proof of concept."*

Approach

Using a subset of DLA-managed items mitigated the challenge of the computational limitations presented by air-gapped laptops, which inherently constrains the amount of data that can be used for modeling. Additionally, working from a sample of the full item population promotes agility in the solution development: modeling can be conducted faster; and any approaches deemed insufficient or incomplete can be quickly adjusted or thrown out all-together.

The project team began by conducting an analysis of DLA's item population to determine active demand items by Acquisition Advice Code (AAC), supply chain, and material type. For a first phase proof-of-concept, the Working Group agreed on the strategy of limiting the subset to items to AACs D, H, J, and Z within the hardware supply chains (Construction & Equipment, Land, Maritime, Aviation, and Industrial Hardware) with active demand. This included both collaborative and non-collaborative items.

With the subset strategy identified, the project team conducted Stratified Random Sampling (SRS), which extracts a random sample of items across each of the specified the strata, or classes (in this case, the supply chains). Standard SRS creates a random sample that retains proportionality across classes. For example, if a full population is 55% female and 45% male, SRS will select a random sample that is also 55% female and 45% male. The Working Group agreed that this approach was problematic for AIDF, however, because models would be biased to the larger supply chains. Thus, the project team conducted *disproportionate* Stratified Random Sampling (dSRS) to create equal item representation across the five supply chains.

Through collaboration with the Working Group, J34 stakeholders identified that items of high operational significance to DLA must be included in the sample to address items with high business value. Specifically, Aviation, L&M, and Troop Support each have their own item "super groups" – Aviation's "Crown Jewels," L&M's "Super Kids," and Troop Support's "Silver Bullets" (aka "Philly Specials"). In each MSC, these groups are comprised of items that have been flagged by Leadership as top priorities for planning. Because these items make up a comparatively small portion of the population, the AIDF team augmented the random sample produced by dSRS with the top 25k items by Annual Demand Frequency (ADF), independent of supply chain which substantially increased the amount of representation these items received.

After augmentation, the subset was analyzed to ensure sufficient representation across JDA's current demand classes and model types, as well as across the range of historical forecast performance (i.e., an adequate mix of items that have been historically over-forecasted, under-forecasted, and accurately forecasted).

[10] https://www.dla.mil/AtaGlance/

Item Subset Selection Results

Figure 6 demonstrates the initial lack of items with high demand frequencies (left) prior to augmentation to meet the J34 stakeholder's requested representation of high-value item inclusion (right). In the initial subset, only 30 items had 100 days of demand and only 5 items had 200 days of demand. By adding the top 25K items by ADF to the original subset, those quantities rose to approximately 200 items and 25 items, respectively. This augmentation allowed more meaningful statistics to be generated on the important business drivers (50+ days with demand range) without sacrificing the ability to evaluate more sparsely demanded items.

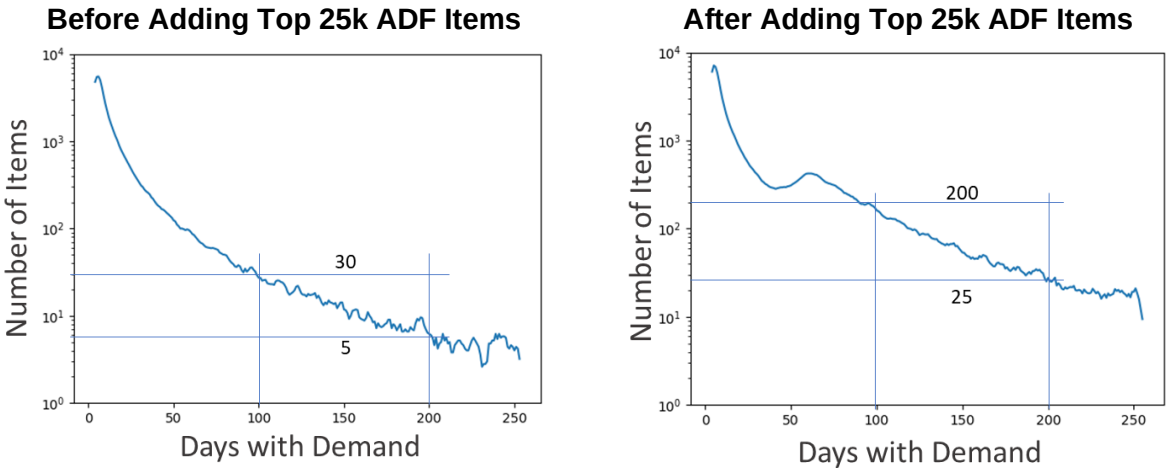

Figure 6: Histogram of Item Days with Demand before augmentation (left) and after augmentation (right)

As a result of the high-ADF item augmentation, the total sample size available for modeling is 145k items. This was the population used for model development, however forecast accuracy comparisons with current JDA performance are limited to only those items with available historical forecasts.

This population is large enough to show statistical significance when evaluating comparative model performance as evidenced in the Appendix D: Model Results section. This population is not meant to provide an exact proportional representation of the entire DLA population, but rather a means to develop a proof-of-concept that accounts for the Working Group input. Further details describing the methodology of selecting the item sample for evaluation is included in Appendix C. The process was an iterative collaboration with the Working Group.

Metrics Evaluation

Model development and tuning requires evaluation metrics to compare forecast performance. Due to the complexity of time-series forecasting and the use of a variety of metrics within academic literature, a study was performed to understand the characteristics, strengths, and weaknesses of various forecast performance metrics as they applied to DLA's data and Mission, then select the best option.

Business Metrics vs. Algorithm Optimization Metrics

Machine learning models for forecasting require metrics for algorithmic optimization. DLA has established *business* metrics to monitor and evaluate forecasts; however, DLA does not have an *optimization* metric suitable for AI development purposes. An ideal optimization metric accurately measures forecast error for all ranges of annual demand frequency. How a metric handles zero forecasts, zero actuals, and other extremes can pose problems when attempting to perform an unbiased evaluation of model accuracy.

Prior to any model development, Accenture evaluated DLA's existing metrics for measuring forecast accuracy and forecast error: Demand Plan Accuracy (DPA) and Absolute Dollar Error, respectively. Absolute Dollar Error uses the dollar value of the demand forecast error to better understand the impact to DLA business outcomes.

In addition, the project team researched and evaluated forecast accuracy metrics commonly found in academic and commercial literature – such as Mean Absolute Percent Error (MAPE), Symmetric Mean Absolute Percent Error (sMAPE), and Mean Absolute Scaled Error (MASE). To determine an optimization metric for model training, a systematic approach to sufficiently weigh the trade-offs between accuracy improvements and interpretability was developed.

Metrics Introduction

The team selected seven metrics to evaluate their performance as optimization metrics. The seven metrics include DPA, MAPE, sMAPE, MASE, Relative Percent Error, Absolute Dollar Error, and Log Error. These metrics represent a variety of approaches to handle zero forecast, zero actuals, and extreme outliers, as well as converting error into a dollarized value instead of a percentage of the actual value.

The project team began by reviewing the basic attributes of the metrics considered for use. For example, Figure 7 compares the range of values for DPA (in the graph it is inverted for comparison) and sMAPE (normalized 0 to 1 for easier comparison), we can see that differences in handling over and under forecast error bias.

Due to error clipping, DPA is a non-continuous metric, making the optimization of ML algorithms difficult. **SMAPE** solves the problem of asymptotically bounding error values, thereby allowing consistently meaningful results over a larger range of error. Additionally, because error is the absolute value, sMAPE does not cancel out error in aggregate views and continues to scale with larger error values, all while still limiting the runaway effect of zero demand periods.

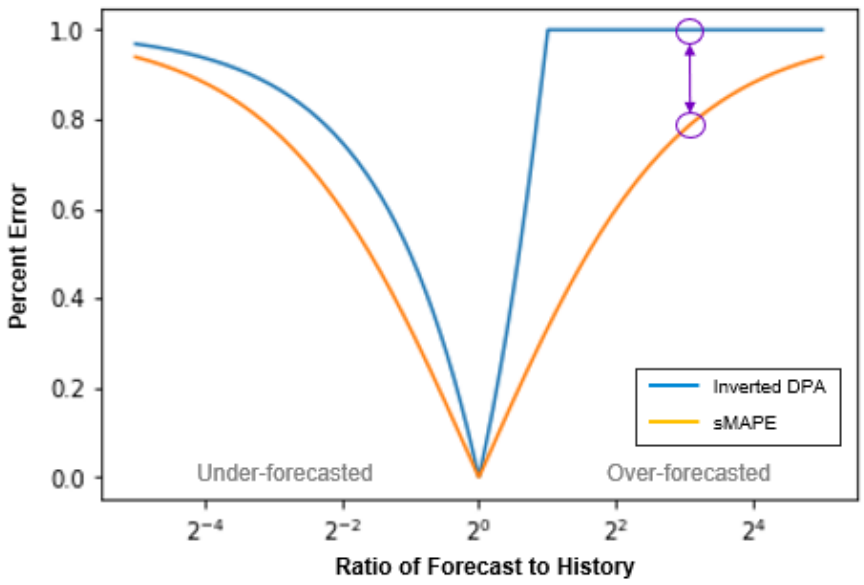

Figure 7: Percent Error Range for DPA (Inverted for Comparative Purposes) and sMAPE

As comparison of example DPA and sMAPE calculations are shown in Figure 8. A zero-forecast is no longer proactively planning and relies on supply planning strategies; thus, metrics should penalize this forecast behavior. Furthermore, an optimization metric ideally increases the penalty for increasing error. SMAPE performance meets these two criteria, while DPA does not. Intuitively, sMAPE provides more information to analyze the forecast error than the DPA, which in the example below stops penalizing over-forecasts beyond 10 (i.e., DPA remains 0%).

	All Zeros	Small Error	Medium Error	Larger Error	Extreme
Forecast	0	8	10	20	50
Actuals	0	5	5	5	5
DPA	100%	40%	0%	0%	0%
sMAPE	0%	46%	67%	120%	164%

Figure 8: Illustrative Example of DPA and sMAPE Differences

Complete details on the metric calculations are included in Appendix F.

Selection of Measurement Periods

To apply the metrics, an aggregation method was selected. The team established goals to evaluate all 12 forecast lags equally, reduce the impact of timing effects lag-to-lag, and aggregate items by how difficult they are to forecast – primarily based on demand frequency. With these goals in mind, the AIDF team selected quarterly measurements of the 12-month forecast that are equally averaged. These selections evaluate all forecast periods equally and allow for the aggregation of metric performance. When aggregated, the metrics are averaged, typically by the days of demand. This approach mimics existing DLA metrics that use multiple measures that are equally averaged.

An example is shown in Figure 9. Beginning from the top, the sMAPE metric is aligned into monthly lags. Moving down the Figure 9, the sMAPE calculation is aggregated into quarterly periods and then averaged across the year. This effectively minimizes the large swings that occur

in forecast accuracy when demand materialization is simply offset by a month, while still penalizing forecasts that do not represent the actual demand pattern.

	Lag0	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7	Lag8	Lag9	Lag10	Lag11
Actuals	8	0	6	7	0	9	2	9	4	20	0	3
Forecast	5	5	5	5	5	5	5	5	5	5	5	5
sMAPE	46.2%	200.0%	18.2%	33.3%	200.0%	57.1%	85.7%	57.1%	22.2%	120.0%	200.0%	50.0%

	Q1	Q2	Q3	Q4
Actuals	14	16	15	23
Forecast	15	15	15	15
sMAPE	6.90%	6.45%	0%	42.10%

Average sMAPE	13.86%

Figure 9: Example Measurement Periods Aggregation

Optimization Metric Evaluation Approach

In order to select an optimization metric, an experiment was designed to evaluate each metric's ability to choose the "best" forecast across ranges of annual days of demand by measuring the impact of intentional bias. Due to individual item's demand history variation, the experiment was performed on the entire Item Subset Population to increase statistical significance.

The AIDF team used the demand history from 2017 to generate four simple forecasts for each item. An illustrative example for this experiment as shown in Figure 10:

1. **True Mean (Green Dashed Line):** The mean value of 2017 demand history values including zeros. This "forecast" is the most accurate flatline forecast possible and should generate the best performance on aggregate.
2. **10% Over Mean (Red Dashed Line):** The True Mean "forecast" increased by 10% to systematically inject over-forecast bias into the values. If this forecast has the best aggregate performance, the metric has an over-forecast bias.
3. **10% Under Mean (Blue Line):** The True Mean "forecast" decreased by 10% to systematically inject under-forecast bias into the values. If this forecast has the best aggregate performance, the metric has an under-forecast bias.
4. **All Zeros (Orange Line):** All zeros are created as the "forecast". Multiple metrics have assumptions to handle zero forecasts, this forecast is designed to test the impact of a zero forecast, especially for very sparse demand items.

Because these "forecasts" are calculated from the known historical usage, they act as controls in an experiment to evaluate the metric's reaction to different forms of bias or error in a forecast. In other words, knowing that the average of the demand history (Forecast 1) should exhibit less error than artificially increased (Forecast 2) and decreased (Forecasts 3 and 4) values, an unbiased optimization metric should choose Forecast 1.

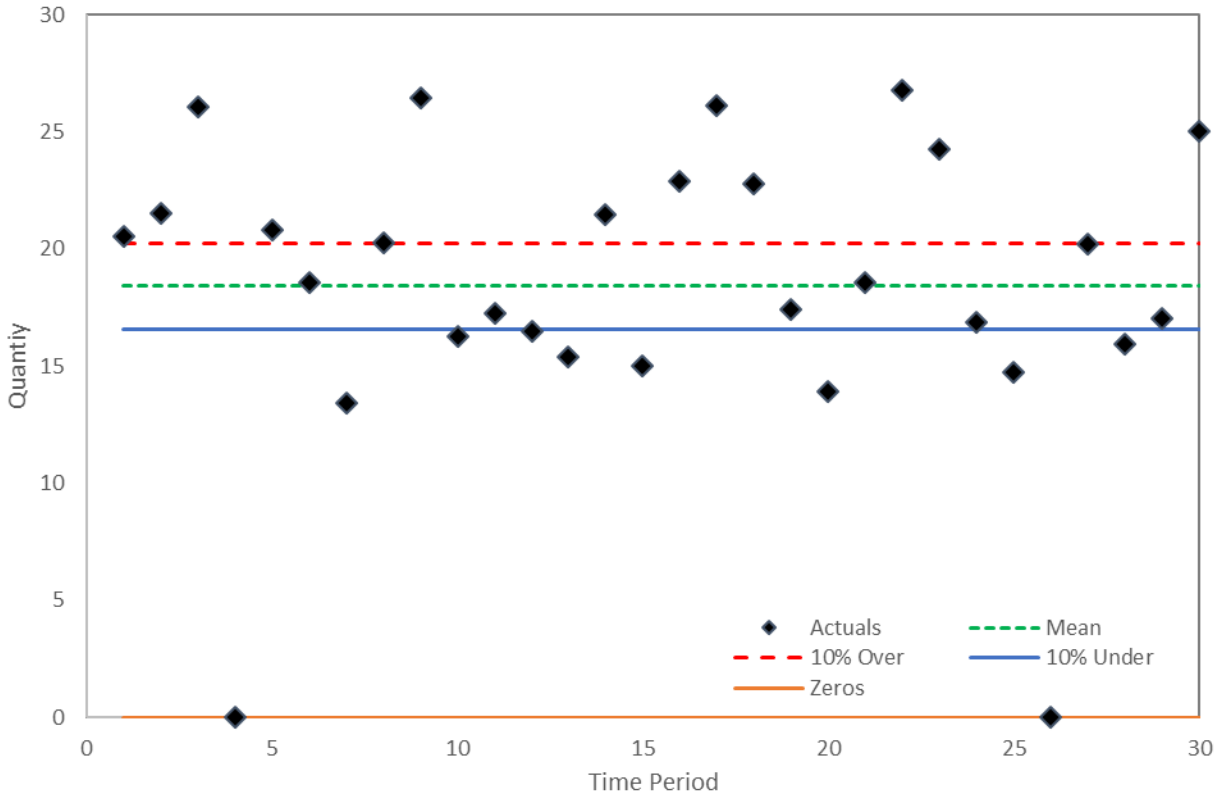

Figure 10: Illustrative example of Optimization Metric Experiment "Forecasts"

After creating the “forecasts” the metrics were calculated for every item. The results were aggregated by days of demand. To view the results smoothing windows were applied, averaging the results (e.g., a 9-day smoothing window is the average of items with 6 to 14 days of demand). This technique is commonly applied to time series problems. Using a smoothing window allows for easier visual interpretation of broader trends by reducing the impact of noise. An illustrative example is shown in Figure 11, of how a smoothing window is applied and the smoothed result.

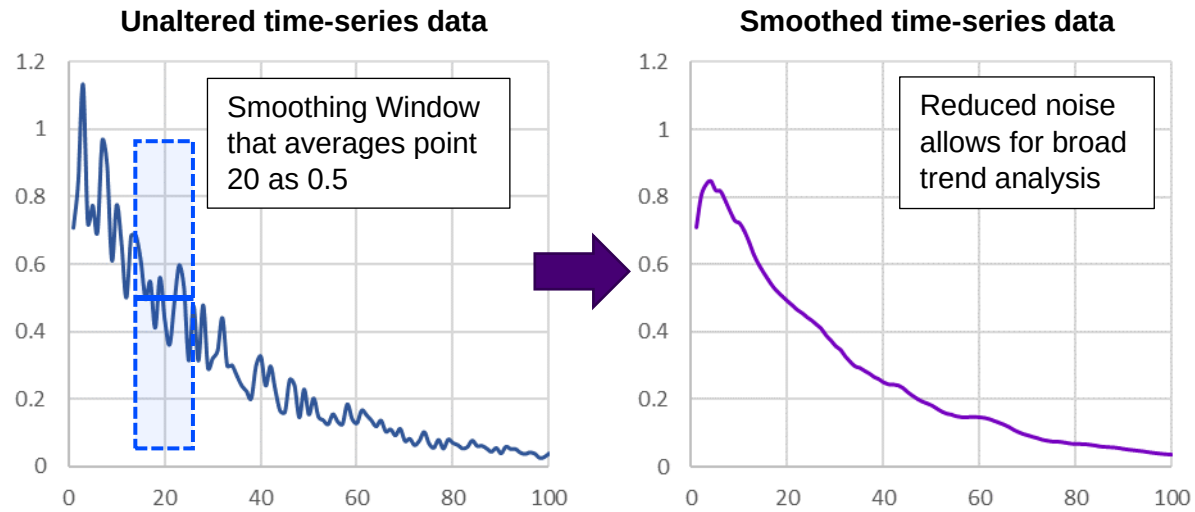

Figure 11: Illustrative Example of a Smoothing Window

Metrics Evaluation Results

The behavior of the metrics changed over the days of demand range quite dramatically. To inspect the results, five ranges were created. A summary of the results are shown in Figure 12 comparing the demand range and which metrics measured the True Mean (ideal) forecast as the best performance (green) vs. which metrics selected an intentionally biased forecast (red).

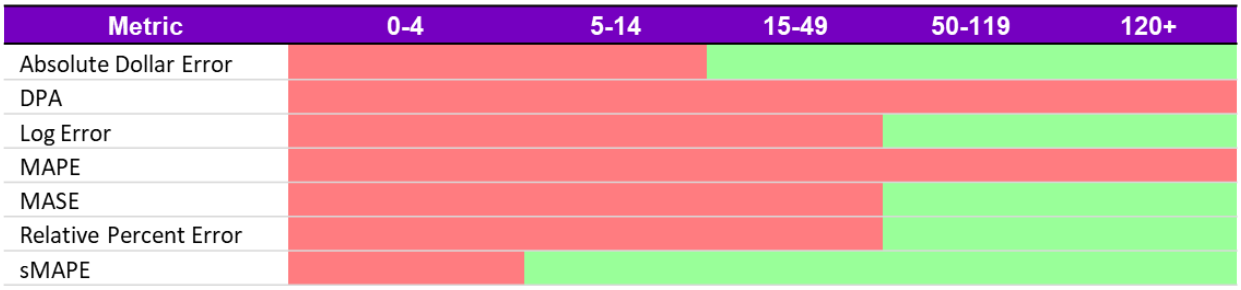

Figure 12: Evaluation of Metric Performance over Demand Frequencies

Further details for each demand range are discussed below.

The 120+ Days with Demand Items

High demand frequency items are expected to have the least issues with metrics assessments since the likelihood of having no demand in a quarter for items with more than 120 days with demand in a year is very low. Additionally, these items are likely to be less erratic than other items with lower demand frequencies as the high amount of demand helps to smooth the demand pattern.

Figure 13 shows the results of this analysis for items with 120 or more days with demand. All metrics except DPA and MAPE selected the true forecast. DPA and MAPE, however, selected the forecast which was purposely under-biased by 10% as the best forecast. This is indicative of known under-forecasting bias in the MAPE metric, on which DPA is heavily based.[11]

Metric	True Forecast	10% Over-Biased Forecast	10% Under-Biased Forecast	Zero Forecast
Absolute Dollar Error	$16,970.21	$19,351.40	$17,763.15	$82,566.06
DPA	75.79%	71.90%	76.73%	0.00%
Log Error	34.20%	37.43%	35.52%	500.00%
MAPE	28.21%	33.28%	26.35%	100.00%
MASE	137.30%	154.96%	146.80%	465.98%
Relative Percent Error	22.61%	24.98%	23.45%	100.00%
sMAPE	23.01%	25.11%	23.96%	200.00%

Figure 13: Comparison of Metric Ability to Identify True Forecast, 120+ Range

The 50-119 Days with Demand Items

The outcome of the metrics analysis in the 50-119 days with demand items is the same as the outcomes for the 120+ days with demand items, as shown in Figure 13 and Figure 14. This indicated that over 50 days of annual demand, the metric behavior is fairly stable and consistent.

[11] https://robjhyndman.com/papers/foresight.pdf

Metric	True Forecast	10% Over-Biased Forecast	10% Under-Biased Forecast	Zero Forecast
Absolute Dollar Error	$6,089.04	$6,560.50	$6,288.88	$21,695.51
DPA	66.31%	63.09%	67.84%	0.03%
Log Error	50.15%	52.79%	50.54%	500.00%
MAPE	47.70%	53.95%	43.58%	99.97%
MASE	136.07%	145.23%	138.58%	415.37%
Relative Percent Error	31.46%	33.40%	31.70%	100.00%
sMAPE	32.70%	34.25%	33.12%	199.95%

Figure 14: Comparison of Metric Ability to Identify True Forecast, 50-119 Range

The 15-49 Days with Demand Items

As shown in Figure 15, as demand becomes less frequent in the 15-49 days with demand range, fewer metrics are able to correctly select the true forecast. The only two metrics that do are Absolute Dollar Error and sMAPE.

DPA, Log Error, MASE, and Relative Percent Error all select the purposely biased 10% under-biased forecast rather than the true forecast which correctly represents the demand over the year. MAPE also selects a misrepresentation of the forecast by selecting a forecast of all zeros. This has potentially dangerous implications to supportability for the Warfighter if models are selected based on the MAPE, as it will tend towards selecting zero-forecasts for sparse items due to inherent flaws in its methodology.

Metric	True Forecast	10% Over-Biased Forecast	10% Under-Biased Forecast	Zero Forecast
Absolute Dollar Error	$4,263.25	$4,406.91	$4,275.76	$10,137.96
DPA	55.33%	53.22%	56.57%	0.79%
Log Error	85.99%	88.38%	85.57%	499.99%
MAPE	101.84%	111.38%	93.47%	99.21%
MASE	140.70%	146.36%	140.38%	323.92%
Relative Percent Error	44.93%	46.69%	44.63%	100.00%
sMAPE	48.52%	49.55%	48.74%	198.42%

Figure 15: Comparison of Metric Ability to Identify True Forecast, 15-49 Range

The 5-14 Days with Demand Items

In the population of items with 5-14 days with demand shown in Figure 16, the metrics assessment becomes more nuanced. Nearly every metric prefers under-forecasting in this range, with the exception of sMAPE. Despite selecting the true forecast, sMAPE is only doing so by 0.08% over the intentional over-forecast. Since items in this range only receive demand between approximately 1 to 4 percent of days within a year from any customer across DLA, the volatility is inherently very high and difficult to forecast.

Notably, MAPE again suggests the best forecast to select is the zero-forecast. This is likely due to the very high error MAPE induces when comparing a positive forecast with an actual demand of zero compared to the other metrics. This would lead to MAPE selecting forecasts comprised entirely of zero values over true representations of the actual forecast encountered.

Metric	True Forecast	10% Over-Biased Forecast	10% Under-Biased Forecast	Zero Forecast
Absolute Dollar Error	$2,155.19	$2,213.25	$2,125.48	$3,268.50
DPA	53.17%	52.31%	53.69%	12.14%
Log Error	229.69%	231.89%	228.54%	499.53%
MAPE	193.44%	205.57%	181.42%	87.78%
MASE	144.49%	148.50%	142.46%	218.55%
Relative Percent Error	67.39%	69.30%	66.36%	100.00%
sMAPE	77.83%	77.91%	78.45%	175.72%

Figure 16: Comparison of Metric Ability to Identify True Forecast, 5-14 Range

The <5 Days with Demand Items

Figure 17 demonstrates that within the range of items with fewer than 5 days with demand in a year, there is no metric that comes close to selecting the true forecast value. Every metric, with the exception of DPA, selects a forecast comprised entirely of zeros in this region.

Notably, DPA does not select the zero forecast, but instead selects the 10% over-biased forecast. This is due to the DPA metric assessing all zero-demand periods as 100% accurate regardless of the forecast. While this avoids issues with selecting the zero-forecast over actual forecasts, it also means that in regions of highly volatile demand, DPA will optimize towards the non-zero demand intervals without accounting for periods without any demand.

Metric	True Forecast	10% Over-Biased Forecast	10% Under-Biased Forecast	Zero Forecast
Absolute Dollar Error	$1,636.10	$1,663.70	$1,620.19	$1,587.60
DPA	75.00%	75.59%	73.82%	53.20%
Log Error	463.16%	463.65%	463.37%	426.62%
MAPE	154.46%	162.20%	146.88%	46.55%
MASE	125.06%	128.33%	122.60%	114.00%
Relative Percent Error	114.40%	117.42%	112.08%	99.96%
sMAPE	140.44%	138.69%	142.98%	93.61%

Figure 17: Comparison of Metric Ability to Identify True Forecast, <5 Range

As an example, Figure 18 shows the sMAPE performance for the range of items with less than five days with demand, which illustratively shows the inability to select the correct forecast for very low demand items. In this case, an all-zero forecast is selected as the top-performing forecast (lowest error). This experimental result is consistent with commercial best practices and existing DLA business rules that limit forecasting to items with sufficient demand by the WAT test.

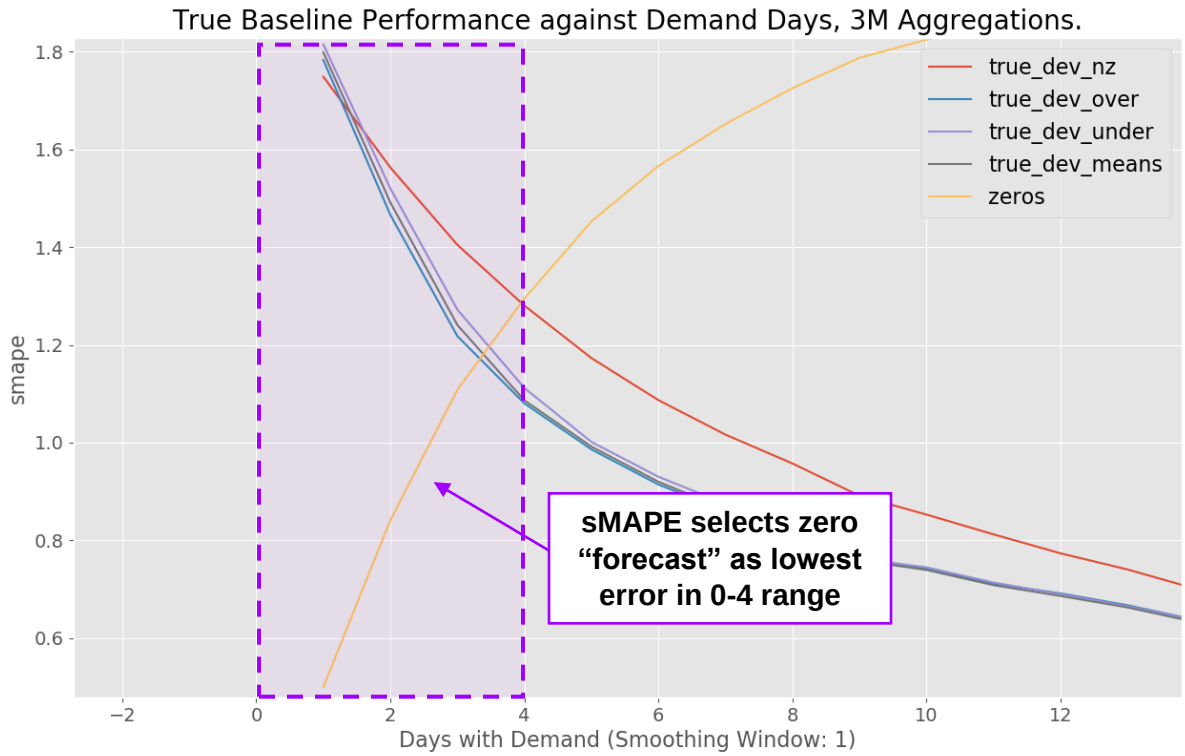

Figure 18: sMAPE Low Demand Baseline Performance

The conclusion that none of the common forecast error metrics perform adequately in the sparse demand region illustrates an underlying issue with the use of discrete, point forecasts for items with extremely low demand frequencies. High uncertainty about whether an item will have any demand in a given period significantly diminishes the business value of a discrete forecast. The AIDF team recommends that ranged forecasts, with the ability to account for variability and probability of zero-demand values be used in these very sparse demand items. Probabilistic ranges would allow for DLA to plan for these items by hedging against high levels of variability while avoiding over-procurement of safety stock.

Separately, the overall range of demand frequencies for DPA is shown in Figure 19. DPA selects the under-forecast bias consistently until nearly continuous demand. The impact of this is that when using DPA as an optimizing metric to choose between models, the models selected will tend towards those with higher under-forecasting errors. As DLA's primary metric, DPA is easy to understand; however, the calculation has inherent limitations that impede the ability to improve the forecast. The AIDF team recommends DLA re-examine the Enterprise forecast metrics to account for this noted bias.

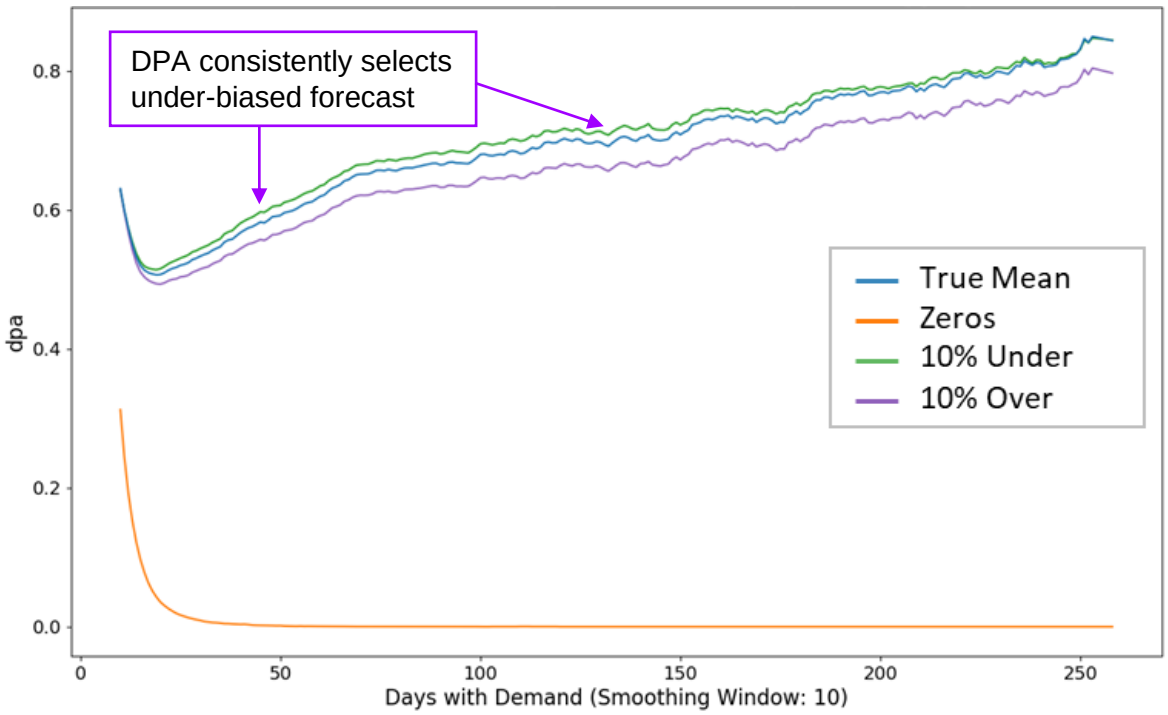

Figure 19: Meta-Evaluation of DPA's Ability to Discern the Correct Forecast

Key Conclusions

- As demand gets sparser, all metrics struggle to correctly select the most accurate forecast.
- DPA consistently exhibits under-forecasting bias in every range of demand (except the <5 range).
- sMAPE is able to select the true forecast across the greatest range of demand density and will therefore be used as the optimization metric. Because of sMAPE's robustness in correctly identifying ideal forecasts, it will also be included as a business metric during model testing and evaluations.
- No metric was able to accurately measure items with fewer than five days of demand, therefore point-forecasts in this region cannot be measured and improved. DLA has existing business logic that requires forecastable items to have at least four periods of demand in the past year which closely resembles the findings of this analysis.

Individual Forecasting Algorithms Modeling

At a high-level, the modeling approach is an **iterative** four-step process that is repeated for each forecast algorithm to prepare models for final testing, as shown in Figure 20. Throughout the model development process, each step can be revisited (often multiple times) to find the best features, architecture, and algorithms as a deeper understanding and intuition of the problem is gained.

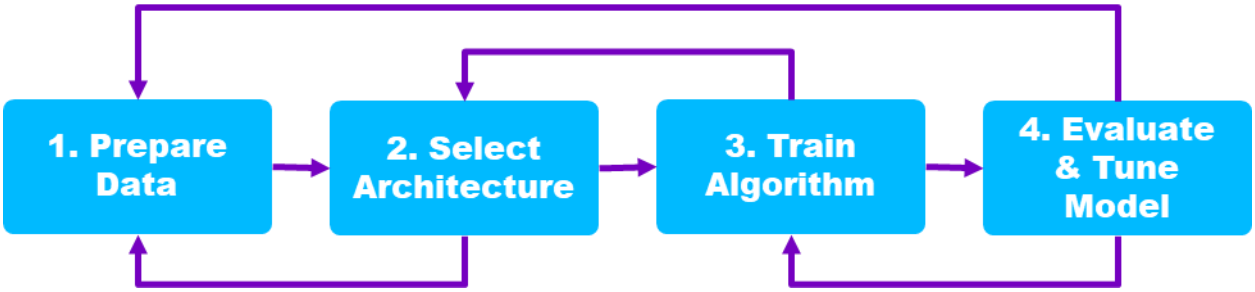

Figure 20: High-Level Overview of Forecast Model Development

Prepare Data

AI/ML modeling begins with data and forecasting is no exception. The AIDF team first identified data elements and sources to extract data. There are three primary types of data: item characteristics, demand history, and historical JDA statistical forecasts. JDA SCPO ("Supply Chain Planning and Optimization") is the production system and often does not contain archived data, thus DLA uses data warehouses to archive key data elements. Data sources are listed in Figure 21.

Data	Source System	Extraction System
Item Characteristics	SAP ECC: MARA, KONP	DORRADW: MTRL_MSTR_HIST_DIMENSION
Demand History	JDA SCPO: ZMDT_Dmd_Hist_Dtl	DORRADW: DP_Hist_Dtl_Fact
JDA Stat Forecast	JDA SCPO: FcstDraft (archived)	EDW: CV_PL_HPL_O20

Figure 21: Key data sources source and extraction systems

After the data was extracted for the item subset population, data validation was performed. First, the data warehouse datasets were cross referenced against the source systems. The JDA Statistical Forecast history was compared against post DMS archives of the FcstDraft table used to populate the What-If system. The FcstDraft immediately after DMS is the unaltered statistical forecast. Additionally, extreme outlier values were manually inspected and validated against source systems. Finally, the data extraction and archiving process was reviewed with the EDW technical teams that support the COPE forecast metrics development and own the data extract, transform, load (ETL) process. ETL errors were found on the CV_PL_HPL_O05 table and removed from further analysis.[12]

After the data was validated the datasets were split into training, development (dev) and test sets. The training set was used to train the AI/ML algorithm and the development set was used to evaluate and tune model hyperparameters. The test set was held out from all analysis until the

[12] These errors were addressed with the EDW team for remediation.

models were ready for final testing and similarly only includes demand and item data prior to 2018. This strict partitioning of data prevents information leakage or bias in the testing of the AI models to provide the most accurate estimate of real-world performance. The test dataset is commonly referred to as the hold-out data.

Seven years of demand history were extracted, and split into train, dev, and test sets based on calendar years as shown in Figure 22. Similarly, historical item characteristics were split and organized by date into train, dev, and test sets.

Figure 22: Item & Demand History data set splits

The components of data preparation (data prep) are preprocessing, transformations, and feature engineering. Data prep is a highly iterative process, especially as models generate results that act as feedback to the data scientists. Multiple options for data scaling and feature engineering were created to evaluate their predictive power. The following techniques were applied with various parameter settings:

- **Winsorization:** adjusts outliers to limit the impact of extreme history quantities. A variety of percentiles were evaluated from 90-99%to adjust outliers.
- **Input Data Scaling:** adjusts the scales of the input data to speed up the machine learning process. AIDF evaluated multiple scaling techniques including standard, mean, min/max, and median scaling
- **Feature generation**: creates new variables that are used by algorithms to generate the forecasts. Features were generated using item details and demand history data.

Pre-Modeling Architecture

Once the master training and development datasets were prepared, the model architecture parameters (also known as hyperparameters) that are relevant to all algorithms were established for evaluation. The combination of hyperparameters create exponentially increasing numbers of potential models for each algorithm. While many of these choices are prescribed within JDA, the AIDF set out to test these choices to determine if algorithmic flexibility could be used to improve performance. Additionally, each algorithm has unique hyperparameters, creating a large potential search area to determine the best design combinations.

The first decision for time-series analysis is the history time period to train. The training dataset includes five years of history. Models were trained on as short as 18 months of history and as long as the full five-year history.

Next, the prediction intervals to generate and evaluate forecasts were tested. Forecast prediction intervals of one week, one month and one quarter were evaluated.

Finally, forecasts were generated at the item-level. Currently, DLA can use CRM cells to subdivide history by customer and develop independent forecasts that are aggregated for supply planning. The AIDF team chose to model at the item-level to provide the AI/ML models the maximum data availability for training.

Individual Forecast Algorithm Training

One of the AIDF project's primary goals is to create a foundation for demand forecast research at DLA. As such, in conjunction with DLA's ACE and COPE, the team chose to start modeling with the simplest algorithms before increasing model complexity. Understanding that additional complexity may not always improve the forecasting process, the AIDF team used simple models as baselines for evaluation against complex, modern forecast algorithms.

Algorithms were divided into three classes – Simple, Classical, and Complex as shown in Figure 23. Additionally, algorithms that are available to JDA were included to evaluate the pre-model architecture choice's impact on the forecast performance.

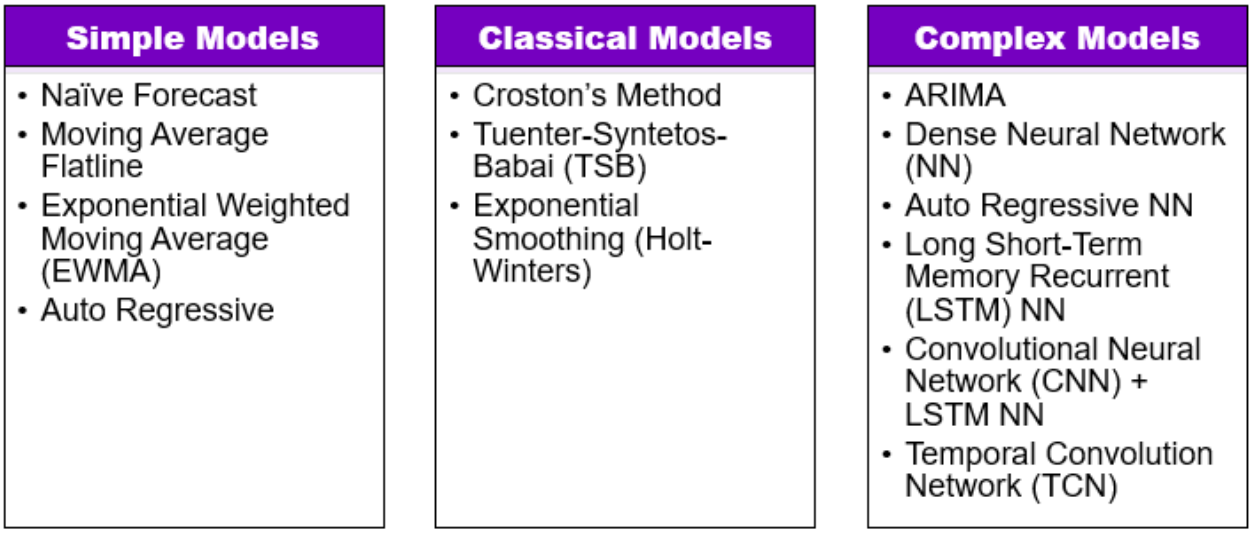

Figure 23: Forecast Algorithms Grouped by Approach

The initial round of modeling used single algorithms trained on only the demand history, also known as univariate models. AI/ML approaches have demonstrated improvements by expanding the aperture of item characteristics or features to train models. The team used weapon system data, item characteristics, and demand history to create new historic features that are included in model training and forecast generation. These models are known as multivariate models. While some models were evaluated using this approach, progress was limited due to time constraints and lack of historical item characteristic data.

Model Tuning and Evaluation

As part of the forecast model training, hyperparameters were tuned to find the best version of the model for the type of data evaluated. The first step was to select the metric used by the training algorithm to optimize the forecast model. Time-series models have a variety of metrics and evaluation periods available for selection. SMAPE was selected as the optimization metric, as discussed in the previous **Metrics Evaluation** section.

To tune AI/ML models, combinations of hyperparameters are traditionally evaluated through trial and error, this is known as a hyperparameter search. Figure 24 outlines the algorithm, hyperparameters, and search technique used to train forecasting models in this STP. Note, the neural network models have similar hyperparameters for all algorithm types utilized. Further details on the neural network models are included in the **Time Series** Model section.

Algorithm	Hyperparameter(s)	Search Method
Moving Average	• Average time period • Winsorization	• Grid
Exponential Weighted Moving Average	• Winsorization • Alpha (smoothing factor)	• Grid
Auto Regressive	• Number of time lags	• Grid
Croston's Method	• Alpha (smoothing factor)	• Grid
TSB	• Alpha (smoothing factor – quantity) • Beta (smoothing factor – interval) • Winsorization • Training history period • Input time periods (daily, weekly, monthly etc.)	• Grid
Exponential Smoothing	• Trend • Damped • Seasonal • Box-Cox • Remove Bias	• Grid
ARIMA	• Number of time lags • Level of differencing • Order of moving average model	• Grid
Neural Networks**	• Network Architecture • Cell / Layer Types • Number of Cells / Layers • Learning Rate • Optimizer • Normalization • Regularization • Auto Regression (Binary)	• Genetic Algorithm • Random (for initial testing)

Figure 24: Overview of Algorithm Hyperparameters and Search Method(s) employed

Individual Model Performance on Dev

Prior to executing the trained models on the test hold-out set, preliminary results were generated on the development dataset. These results gave an indication of performance to inform next steps before running final test procedures.

The models were assessed within four primary ranges of demand frequencies (5-14, 15-49, 50-119, and 120+ days with demand), and are shown with the top-performing hyperparameters for each algorithm type selected to represent that region. For example, the version of Croston's model shown in Figure 25 represents the top-performing hyperparameter combination for Croston's method in the region of items with 5-14 days with demand.

The results of these assessments are shown in Figure 25 through Figure 28.

Figure 25 shows model performance over the 5-14 days with demand region of items. In this region, it appears that high levels of complexity are not able to improve the demand forecast over

simpler methods such as EWMA. This is likely a result of high variability and a lack of clear patterned demand that an AI algorithm can adapt to.

Algorithm	Mean sMAPE	SE sMAPE	Mean DPA	SE DPA
JDA	101.63%	0.37%	42.91%	0.20%
EWMA	94.00%	0.32%	46.72%	0.18%
TSB	94.14%	0.32%	46.38%	0.18%
CNN-LSTM (NN)	94.18%	0.32%	45.59%	0.19%
Croston's	95.04%	0.32%	45.18%	0.19%
Simple Average	95.25%	0.32%	45.58%	0.19%
LSTM (NN)	96.39%	0.32%	42.69%	0.20%
Exponential Smoothing	97.65%	0.34%	41.09%	0.20%
ARIMA	104.65%	0.37%	41.33%	0.20%

Figure 25: Error Measures for the 5-14 Days with Demand Population

Figure 26 details the demand forecast performances over the region of demand with 15-49 days of demand. In this region, the base LSTM showed improved performance over all other methods. In addition to demand history, this LSTM utilizes demand group ordering details to improve its forecast, which may be giving it an advantage in this range.

Algorithm	Mean sMAPE	SE sMAPE	Mean DPA	SE DPA
JDA	69.44%	0.30%	43.01%	0.18%
EWMA	64.31%	0.25%	47.12%	0.15%
TSB	64.44%	0.25%	45.94%	0.17%
CNN-LSTM (NN)	64.80%	0.25%	44.35%	0.17%
Croston's	65.80%	0.25%	44.42%	0.17%
Simple Average	65.64%	0.25%	44.46%	0.17%
LSTM (NN)	63.64%	0.25%	44.97%	0.17%
Exponential Smoothing	66.02%	0.26%	44.14%	0.17%
ARIMA	74.77%	0.34%	43.70%	0.17%

Figure 26: Error Measures for the 15-49 Days with Demand Population

Figure 27 shows the TSB variant of Croston's method outperforming all other forecasting methods for the 50-119 days with demand range. Notably, TSB outperforms the traditional Croston's method in every range of demand frequency, suggesting this modification to Croston's works well for DLA's items.

Algorithm	Mean sMAPE	SE sMAPE	Mean DPA	SE DPA
JDA	45.03%	0.21%	56.89%	0.17%
EWMA	43.22%	0.19%	59.51%	0.15%
TSB	43.17%	0.19%	59.37%	0.15%
CNN-LSTM (NN)	43.33%	0.19%	59.95%	0.15%
Croston's	44.37%	0.21%	57.99%	0.16%
Simple Average	44.26%	0.19%	57.71%	0.16%
LSTM (NN)	43.45%	0.19%	58.85%	0.16%
Exponential Smoothing	44.85%	0.20%	57.04%	0.17%
ARIMA	47.41%	0.22%	55.50%	0.17%

Figure 27: Error Measures for the 50-119 Days with Demand Population

Figure 28 shows the highest-demand items within DLA's catalog. In this region, the more complex CNN-LSTM model appears to show higher performance than all other methods. This outcome suggests that the complexity available from advanced neural networks begins to be realized in higher demand ranges.

Algorithm	Mean sMAPE	SE sMAPE	Mean DPA	SE DPA
JDA	34.61%	0.26%	66.77%	0.24%
EWMA	33.91%	0.26%	68.13%	0.22%
TSB	33.57%	0.26%	68.26%	0.22%
CNN-LSTM (NN)	33.46%	0.26%	68.76%	0.21%
Croston's	33.94%	0.27%	67.96%	0.23%
Simple Average	34.20%	0.27%	67.28%	0.23%
LSTM (NN)	34.83%	0.28%	67.37%	0.24%
Exponential Smoothing	35.96%	0.29%	66.70%	0.24%
ARIMA	38.87%	0.33%	63.78%	0.26%

Figure 28: Error Measures for the 120+ Days with Demand Population

Key Conclusions

- Item demand frequency has a clear and direct impact on model performance.
    - Comparatively simple models tend to perform the best in items with sparse demand. This demonstrates the value of a model selection algorithm to apply models to the items which stand to gain the greatest benefit.
- The TSB variation of Croston's method consistently outperformed traditional Croston's, so the latter was removed from continued analysis.
- Exponential Smoothing and ARIMA methods exhibited poor performance across all demand frequency ranges, leading to them being removed from further analysis.
    - Note that two forms of Exponential Smoothing are currently implemented at DLA: Holt-Winters (a specific type of exponential smoothing) and Lewandowski (a JDA proprietary algorithm is structured similarly to Holt-Winters).
    - Given the poor performance shown by the Exponential Smoothing algorithm when applied to DLA's item population, Accenture recommends evaluating the current performance of these two JDA forecasting models to understand whether their added complexity improves the planning processes' business outcomes.
- Across the four segments, multiple models displayed similar performance.

Additional use of AI for Demand Forecasting

The last point raises a new question: how should models be selected? JDA uses a combination of business rules (known as Demand Classification) and/or human judgement to select a single model from the six models available. Relying on demand planner judgement alone does not scale well to enterprise challenges and always poses risk with potential losses of institutional knowledge, thus an automated system is preferred to augment human judgement.

Accenture formulated two new research questions in attempt to address this problem:

1. Can models be combined to blend the best features from more than one model?
2. Can AI determine which models are useful and how much to weight an individual model in the combined forecast?

Before executing the final test procedures, additional AI models were developed to answer these questions. This decision ensures no information is leaked from the test set to maintain a fair, real-world performance test.

Ensemble Model

For each demand density segment, individual forecasting algorithms show improvements over the existing JDA solution. The dynamic nature of the best model-item combination suggests that performance is dependent on characteristics of the items that need to be aligned to the ideal forecasting method.

AI Ensemble Selection Model Approach

Forecasting algorithms are designed to model observed patterns such as seasonality, trends, and demand intensity. By combining multiple models, the team theorized that performance could be improved by taking the best features of each model, while dampening extreme errors. The AIDF team sought to use AI to automatically select and combine the best models. This approach is known as *ensembling*.

Ensembling is a machine learning method that allows multiple models to be combined to produce a single result. Certain approaches such as decision trees, random forests, and boosting require the generation of base learners to occur concurrently with the generation of the ensemble. These methods, however, do not typically include common time series approaches such as Croston's method or LSTM networks as base learners. Therefore, a unique AI-based approach needed to be developed.

A feed-forward neural network was created to receive the item-characteristics and predict the weights of each individual forecasting algorithm to combine as the Ensemble model. The final forecast is a weighted average of the AI-selected and weighted models. Figure 29 details the process of generating model weights from item-level features.

Figure 29: Model Ensembling Process Visualized

Nine models were selected for the Ensemble Model based on the preliminary performance across a varied range of demand frequencies. These models included:

- 2 LSTMs (Long Short-Term Memory) Network
- 1 CNN-LSTM (Convolutional Neural Network – Long Short-Term Memory) Network
- 2 EWMAs (Exponentially Weighted Moving Average)
- 1 Winsorized Average
- 2 TSB (Teunter-Syntetos-Babai) Variants of Croston's Method
- 1 TCN (Temporal Convolutional Network)

To avoid allowing any information leakage from the test dataset during the training process, the dev dataset was split into a sub-division of train (80% of the dataset) and validate (20% of the dataset). After the Ensemble model was trained in this way, it was evaluated on the test dataset to test performance.

A custom loss function was developed using Python's Keras library to minimize the overall error as well as the possibility of error volatility. This method effectively weights each model based on both the probability of having the lowest error and the risk of extreme errors. As a technical note, the custom loss function can be based on any error/accuracy metric and thereby improve performance on that specific metric.

During training, the neural network learns to select the combination of models with the lowest expected aggregate forecast error. This approach enables flexible combinations of demand forecasting algorithms, in effect creating a custom model for each unique item.

Additionally, this technique allows any number of input forecasting algorithms at any level of complexity. The output results are very interpretable as they are simply the selected models and recommended weight given to each individual model. The flexibility of input models can be selected based on benefit to the business, ease of implementation, maintenance cost, etc. This allows DLA to flexibly adjust what models to include in the demand forecasting process to fit the business needs.

The Ensemble Model development followed the same procedures employed to develop individual models unless otherwise specified.

Test Dataset Model Performance

To test the AI forecast models, 12-month item forecasts were generated starting in January 2018 for all items and models. The forecast accuracy and error metrics were calculated using demand actuals from 2018. SMAPE, DPA, and dollarized forecast error metrics were calculated and summarized for the item subset. Note that all performance reported is on the population of items having 5+ days with demand over the evaluation year due to the systematic inability to assess performance on the population of items with < 5 hits of demand.

The outcomes of this research demonstrated potential business value in multiple models by both sMAPE and DPA metrics. Figure 30 shows average measurements of error for sMAPE and accuracy for DPA for all models researched, including the JDA baseline. When looking at sMAPE and DPA alone, the Ensemble, EWMA v2, LSTM v2, and TSB v2 show strong performance and warrant further inspection.

Model	Mean sMAPE	SE sMAPE
Ensemble	60.34%	0.16%
LSTM v2	61.08%	0.16%
EWMA v2	61.35%	0.16%
LSTM v1	61.77%	0.17%
TSB v2	61.87%	0.17%
TSB v1	61.96%	0.16%
CNN-LSTM	62.41%	0.17%
Simple Average	62.65%	0.17%
TCN	62.89%	0.17%
EWMA v1	63.10%	0.17%
JDA	66.90%	0.19%

Model	Mean DPA	SE DPA
EWMA v2	53.53%	0.10%
Ensemble	53.35%	0.10%
TSB v2	53.26%	0.10%
LSTM v2	53.05%	0.10%
TSB v1	52.78%	0.10%
LSTM v1	52.71%	0.10%
EWMA v1	52.38%	0.10%
Simple Average	52.08%	0.10%
CNN-LSTM	51.01%	0.10%
TCN	50.75%	0.10%
JDA	49.90%	0.10%

Figure 30: Comparison of Model Performance over sMAPE and DPA with Mean and Standard Error (SE)[13]

Figure 31 shows the absolute dollar forecast error and over/under-forecast dollar errors for each model averaged by item. These metrics demonstrate the business value, in dollars, represented by forecast error measurements.

TSB v2 and EWMA v2 have strong absolute dollar error metrics however, the aggregate benefit is a result of improvement in over-forecast error at the expense of under-forecast error. This indicates that the models tend to under-forecast, which can lead to worse material availability. LSTM v2, on the other hand, demonstrates mediocre performance on all dollar error metrics. Finally, the Ensemble model demonstrates the highest performance based on absolute dollar error with a balance between over/under-forecast dollar error. Notably, it outperforms JDA in all five metrics. This result highlights the potential business value that can be derived from the incorporation of AI into the demand forecasting process.

Model	Mean Abs $ Error	SE Abs $ Error
Ensemble	$6,789.56	$315.37
TSB v2	$6,904.51	$323.26
TSB v1	$6,961.43	$321.51
EWMA v2	$6,994.44	$327.96
EWMA v1	$7,002.71	$323.89
LSTM v2	$7,014.38	$320.07
LSTM v1	$7,167.44	$339.83
CNN-LSTM	$7,244.72	$321.12
Simple Avg.	$7,312.84	$339.26
JDA	$7,337.08	$320.99
TCN	$7,347.19	$338.68

Model	Mean Over-Forecast $	SE Over-Forecast $
EWMA v2	$2,515.34	$207.78
TSB v2	$2,530.85	$209.13
TSB v1	$2,790.57	$211.86
Ensemble	$2,827.50	$202.21
EWMA v1	$2,835.01	$215.71
LSTM v2	$2,837.71	$198.76
LSTM v1	$2,996.33	$248.95
Simple Avg	$3,025.75	$217.38
JDA	$3,358.16	$208.57
CNN-LSTM	$3,691.28	$220.31
TCN	$3,915.07	$248.96

Model	Mean Under-Forecast $	SE Under-Forecast $
TCN	$3,432.12	$148.55
CNN-LSTM	$3,553.44	$150.92
Ensemble	$3,962.06	$162.01
JDA	$3,978.93	$156.85
EWMA v1	$4,167.70	$155.97
TSB v1	$4,170.85	$156.97
LSTM v1	$4,171.11	$163.61
LSTM v2	$4,176.67	$180.16
Simple Avg.	$4,287.09	$186.74
TSB v2	$4,373.66	$165.07
EWMA v2	$4,479.10	$176.10

Figure 31: Comparison of Model Performance over Mean Absolute, Under-Forecast, Over-Forecast, and Absolute Dollar Error

The Ensemble Model is the primary focus of further analysis presented because it showed the strongest performance of any new forecasting technique evaluated. Given the strong indication from the performance metrics above, an analysis against the JDA baseline was performed to determine statistically significant business impacts. The goal of this analysis was to further support or refute the value of applying AI to demand forecasting at DLA.

[13] Note: The Standard Error reported for DPA of 0.10% for each model is a product of rounding. Values range from 0.096% to 0.104%

Pairwise Test Methodology

Because all forecasting algorithms have been applied to the same subset of items, a direct comparison of their performance metrics at the item level is possible. This is known as pairwise comparison.

To conduct a pairwise comparison, the Ensemble model metrics were subtracted from the JDA metrics item-by-item. This analysis was done for DPA, sMAPE, Absolute Dollar Error, and Over/Under Dollar error. The analysis was repeated for each item in the measurable subset (48k items). The results from the entire subset were grouped to view the distribution of effects and determine the overall impact. An illustrative example is shown in Figure 32.

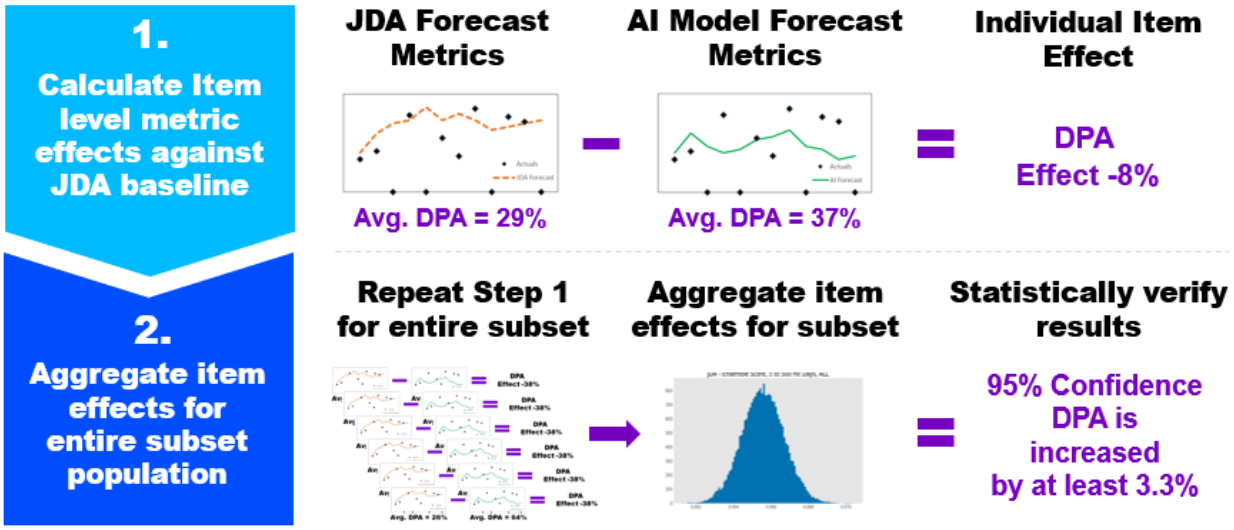

Figure 32: Illustrative Example of Analysis to Determine Effects of AI Model over JDA baseline

Ensemble Model Performance Pairwise Results

Because the Ensemble model demonstrated the best performance using sMAPE (the top evaluation metric, as described in the **Metrics Evaluation** section), the results of its pairwise comparison to JDA are the focus of this section.

The values shown in Figure 33 represent pairwise differences between the Ensemble model and JDA. The Ensemble model shows an average error reduction (sMAPE) of 6.6% when compared to JDA, with 95% confidence of an error reduction of 6.4% or more. Similarly, the Ensemble model's accuracy (DPA) shows an average improvement of 3.5% across the full comparison population.[14] Additional details of sMAPE and DPA performance across demand frequencies and supply chains are available in Appendix D.

The dollarized error metrics listed in Figure 33 (Absolute $, Over-forecast $, and Under-forecast $) refer to the average dollar error the current JDA forecast has over the Ensemble Model per item, each quarter. These results indicate that the primary source of absolute dollarized error avoided comes from a reduction in over-forecast error. Furthermore, this improvement does not come at the expense of under-forecast error. Across the 48k item subset, the error reductions

[14] As a measure of accuracy, the negative numbers associated with DPA comparison indicate JDA had lower accuracy, whereas all other metrics are measurements of error and show JDA having a higher error.

sum to a $105M absolute dollar error reduction annually and $102M over-forecast dollar error reduction annually.

Metric	Average	5th Percentile	95th Percentile
sMAPE	6.6%	6.4%	6.7%
DPA	-3.5%	-3.6%	-3.3%
Absolute $	$545	$453	$648
Overforecast $	$529	$428	$640
Underforecast $	$18	-$60	$91

Figure 33: Average Change and Confidence Intervals for Pairwise Comparison between JDA and the Ensemble Model

These results show that for the items in the measurable 5+ hits of demand range, there was substantial improvement in accuracy of the forecast, and that improvement is primarily driven by decreasing over-forecasting errors. As a proof-of-concept, this shows strong promise for improving the demand forecast, particularly for high-demand items. See the **Model Performance In-Depth Analysis** subsection of Appendix D for more details and example item forecast comparisons.

Inspection of the Ensemble Model Behavior

To understand how the Ensemble Model is behaving, analysis of the individual model weights was performed. Figure 34 shows the average weight of the top five models selected by the Ensemble method across days of demand. While other factors influence the model selection process, the general trend shows that in the lower demand frequency ranges simple models such as EWMA perform well (as indicated by the higher weight percentage assigned to them), while in the higher ranges more complex neural networks are more frequently selected at higher weight values.

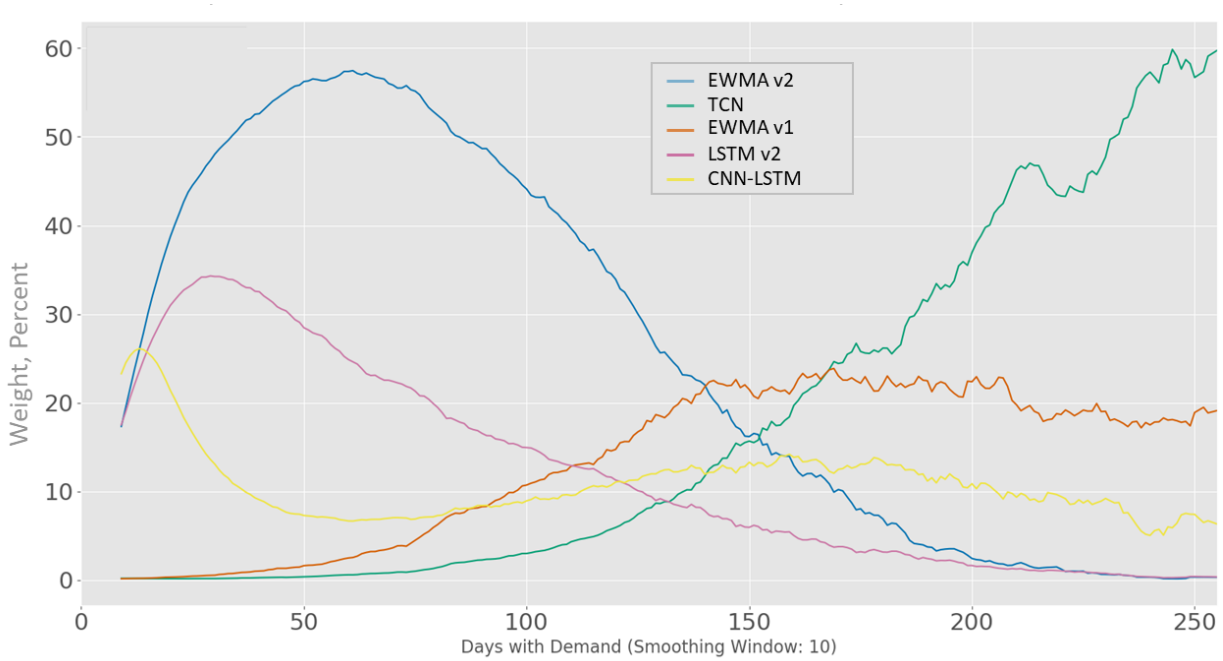

Figure 34: Top Five Model Weights by Demand Frequency Ensemble Model

In addition, histograms of sMAPE by item count for the Ensemble Model and JDA were overlaid to understand where the improvements are likely occurring. Figure 35 illustrates the extreme errors produced by JDA (likely caused by forecasts of all zero quantities) that the Ensemble model tends to avoid.

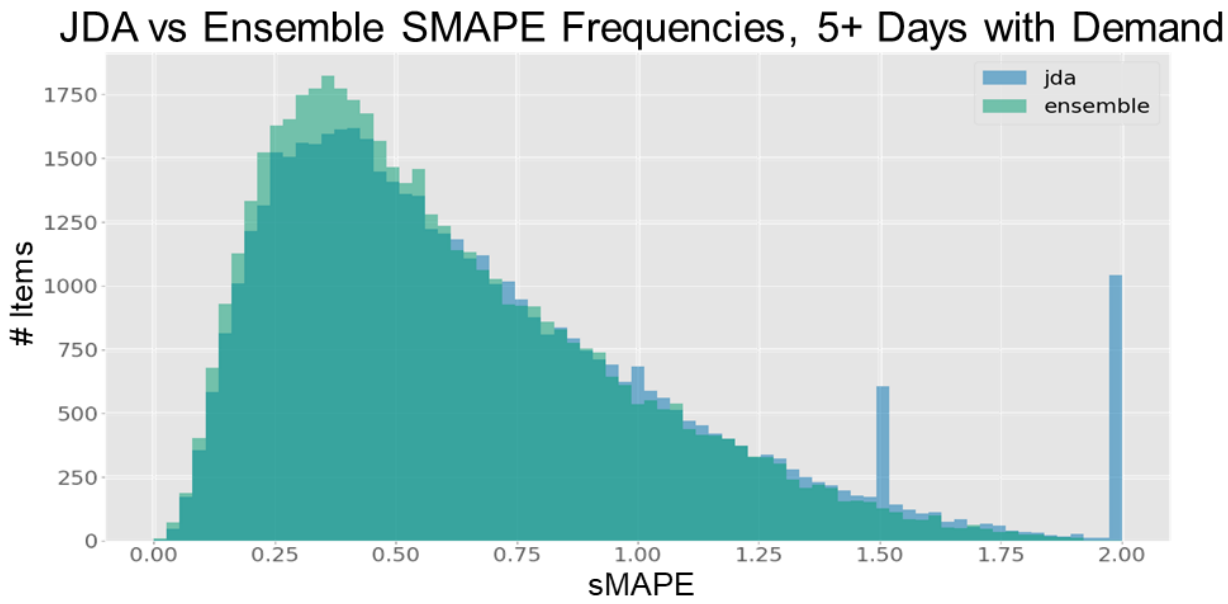

Figure 35: Comparison of JDA vs Ensemble sMAPE Scores

Key Conclusions

The Ensemble Model shows significant improvement to demand forecasting accuracy through AI. It was able to demonstrate:

- Over Forecast Dollar Error reduce by $102M annually over JDA (95% confidence a reduction of at least $83M)
- DPA increased on average by 3.5% over JDA (95% confidence of an average increase of greater than 3.3%).
- sMAPE decreased on average by 6.6% over JDA to Ensemble (95% confidence of a decrease greater than 6.4%)

This research has demonstrated that AI forecasting techniques have potential business value to DLA. Still, the research was not exhaustive. Further experimentation could potentially strengthen the value proposition described in this section.

Collaborative Forecast Considerations

While each DLA supply chain is responsible for its own collaborative planning with the Services, it is the Services that drive collaborative forecast inputs. Each month, the Navy, Army, and Air Force submit their item forecasts directly into JDA through either the Gross Demand Plan (GDP) Demand Data Exchange (DDE). The collaborative forecasts are ingested into the JDA Planning system each month at DMS, and they are transferred to DLA's Supply Planning organization by default unless a DLA Demand Planner overrides them in the system.

As mentioned in the Forecasting Introduction section, these Service-generated forecasts are driven by their specific, depot-level MRO schedule, combined with Bills of Material (BOMs) for the weapon system platforms that they maintain/repair/overhaul. Because DLA does not own the MRO schedule or BOM, it does not own this initial forecast. Thus, in order to make the process collaborative, DLA continues to generate statistical forecasts for all collab items based purely on each item's historical usage.

As indicated by J34 members of the Working Group, the DLA-owned stat forecast is critical to the collaborative process because it provides DLA Planners with a basis for comparison to the Services' demand projections. The current JDA Planning System contains rules-based collaborative forecast exception logic to flag large discrepancies between the collab and stat forecasts. These "collab exceptions" notify DLA Demand Planners where a conversation with their Services counterpart is necessary to determine if the submitted collaborative forecast should be overridden. Thus, the accuracy of the stat forecast is crucial, as it is one of the key components to DLA's exception logic.

Because the Services' own the initial collaborative forecast, the impacts to its accuracy within DLA's control will be inherently process driven, and the stat forecast is the key input to that process. Thus, all of the forecast accuracy improvements demonstrated through this research have the potential to improve DLA's collaborative forecast accuracy by improving the robustness of existing exception logic and arming DLA Planners with the best possible information.

IV. ANALYSIS OF COURSES OF ACTION (COAS)

In order to assist DLA in advancing AI/ML models for demand forecasting it is important to assess the potential business benefits against the cost of exploring the new models further. The AIDF project developed simple to complex forecast models that were compared to the JDA Baseline to determine potential business value, both positive and negative, in order to provide a recommendation that maximizes the benefit to DLA while minimizing the risk.

The identified courses of action below serve as potential options for the path forward to Recommendation 1 from the Executive Summary. Four models were selected as COAs to compare the business benefits of moving from proof-of-concept to working prototype. These models had strong performance and cover a spectrum of techniques ranging from non-AI to a Full AI Ensemble model (which applies AI in two forms):

1. **EWMA:** a non-AI forecast approach, this option serves as a comparison to AI alternatives
2. **LSTM (v2):** a single AI forecast model, this option evaluates the benefits of the top performing individual AI model
3. **Basic AI Ensemble:** AI used to select, weight, and combine five non-AI forecast models, this option isolates the benefit of applying AI to model selection
4. **Full AI Ensemble:** AI used to select, weight, and combine nine forecast models, including four AI forecast models

Potential Benefits

The proof-of-concept model's effectiveness was measured on a set of key performance indicator (KPI) metrics that includes sMAPE, DPA, Absolute Dollar Error, and Over/Under-Forecast Dollar Error on items with at least five annual days of demand. The KPI results for 11 models developed during this STP that improved performance and the JDA baseline are shown in Figure 36.

Model	sMAPE	DPA	Absolute $ Err	Over-Forecast $	Under-Forecast $
Ensemble	60.34%	53.35%	$6,789.56	$2,827.50	$3,962.06
Basic Ensemble	61.05%	53.34%	$6,899.24	$2,720.75	$4,178.49
LSTM v2	61.08%	53.05%	$7,014.38	$2,837.71	$4,176.67
EWMA v2	61.35%	53.53%	$6,994.44	$2,515.34	$4,479.10
LSTM v1	61.77%	52.71%	$7,167.44	$2,996.33	$4,171.11
TSB v2	61.87%	53.26%	$6,904.51	$2,530.85	$4,373.66
TSB v1	61.96%	52.78%	$6,961.43	$2,790.57	$4,170.85
CNN-LSTM	62.41%	51.01%	$7,244.72	$3,691.28	$3,553.44
Simple Avg.	62.65%	52.08%	$7,312.84	$3,025.75	$4,287.09
TCN	62.89%	50.75%	$7,347.19	$3,915.07	$3,432.12
EWMA v1	63.10%	52.38%	$7,002.71	$2,835.01	$4,167.70
JDA	66.90%	49.90%	$7,337.08	$3,358.16	$3,978.93

Figure 36: Summary of Model KPIs

The KPI results illustrate the effectiveness of various models, that aided the creation of COAs for further research and development. On their own, these KPIs do not translate to material business

outcomes (e.g., direct materiel cost reductions or savings). However, the accuracy of a forecast directly influences the quantity of cycle stock and safety stock inventories required to keep an item "healthy" and support DLA customer demands. The KPIs were compared to DLA's current forecasting methods in JDA using the pairwise method (See Approach and Results for details on the methodology). Over/Under Dollar Error KPIs were included to show the relative business value that can potentially be realized. In short, this business case analysis is based on comparing the business benefits for the models used in the STP proof-of-concept with the current forecasting methods employed by DLA on the same item sample population.

The results of comparing each of these forecasting approach COAs to the JDA baseline are represented in Figure 37 below.[15]

	EWMA	LSTM	Basic AI Ensemble	Full AI Ensemble
sMAPE	5.6%	5.8%	5.8%	6.6%
DPA	(3.6%)	(3.2%)	(3.4%)	(3.5%)
Absolute $ Error	$65.7M	$62.2M	$84.4M	$105.1M
Over-Forecast $ Error	$162.2M	$100.0M	$122.8M	$101.9M
Under-Forecast $ Error	($96.1M)	($38.1M)	($38.5M)	$3.5M

Figure 37: Pairwise comparison of business benefits from proof-of-concept models

The results show a trend of increasing business benefit with lower risk of negative outcomes as the application of AI is increased. Note, negative values for DPA mean an improvement, whereas for all other KPIs a positive value is improvement. The comparative results of each of these models is described below in Figure 38.

Rank	Model	Rationale
1	Full AI Ensemble	• Best performance using sMAPE & Absolute Dollar Error • Only option with a positive impact to Under-Forecast Dollar error
2	Basic AI Ensemble	• 2nd best performance using sMAPE and Absolute Dollar Error • Moderate risk of a negative impact to Under-Forecast Dollar Error
3 (tie)	EWMA	• Best Performance using DPA and Over-Forecast Dollar Error • Significant risk of a negative impact to Under-Forecast Dollar Error
3 (tie)	LSTM	• Moderate performance across all metrics • Moderate risk of a negative impact to Under-Forecast Dollar Error

Figure 38: Review of Model Pairwise Results and subsequent Business Benefit Ranking

Risks

There are two types of risk that we considered in the formulation of our recommendation.

1. **Demand Planning Impact –** the potential impacts to DLA operations, both positive and negative. This is calculated by balancing Over/Under forecasting error compared to current methods in JDA. Looking at a solitary metric or KPI to determine a path forward can have adverse effects. Therefore, a holistic assessment of each model's business impacts is critical for minimizing business risk to DLA.

[15] Further details on the statistical validation of these results and additional results for the 0-4 range of items are included in Appendix D in the Course of Action Model Statistical Results section.

2. **Cost-to-Value** – DLA financial commitments before the generation of business value. Costs to advance one, multiple, or all of the models in the proof-of-concept depend on the desire to move incrementally into the next logical development phase of R&D – prototyping – or directly towards a larger-scale production implementation. Following an R&D proof-of-concept, risk can be minimized by selecting a path forward to develop a working prototype, as it would not entail a full-scale technical solution to support the volume and diversity of DLA's product catalogue. As an incremental R&D investment, costs would be controlled and minimized to develop a prototype for one, multiple, or all of these models. In turn, the prototype would provide DLA with the ability to understand real business outcomes via targeted pilot in the near-term.

Our guide for recommending a course of action for DLA accounts for these risks in order to maximize the business impact and minimize cost to value for DLA. As a proof-of-concept, this STP requires more development before an assessment of a full-scale implementation can be made.

Cost Considerations

As a partner committed to helping DLA achieve significant business outcomes, Accenture recommends advancing these models incrementally into a working-prototype as a low-risk investment option for the following reasons:

- Uses controlled R&D investments to develop an AI forecasting prototype providing DLA with a capability to pilot a real-world sample to better understand real improvements in forecast accuracy that can be derived beyond a proof-of-concept.
- Advances the approach without the risk of needing to license proprietary software, data, or technology – the POC demonstrated that this can be accomplished with DLA data, open source software tools, and small infrastructure investments.
- Provides a low-risk opportunity for DLA to understand the relative scale, complexity, resources, and potential change management required for a production implementation without a significant investment.
- Enables Planning to capture real value prior to making production-level investments.

Our estimation of costs does not include full cost of implementation. If DLA chooses to move directly into full-scale implementation, Accenture recommends that DLA perform robust data collection and analysis to identify the full scope of implementation, resource requirements, and associated costs. Long-term implementation costs include cybersecurity, software licenses, data pipelines, infrastructure, and labor for development, integration and maintenance. As demonstrated by the proof-of-concept, DLA has sufficient software and data to move forward and is investing in additional infrastructure. The infrastructure investment is for a dedicated AI Model Training Environment for multiple AI use cases and is not dedicated for this effort alone.

Prototype development costs would be relatively consistent with similar R&D STP investments for any of the models studied during the STP.

Recommended Course of Action

Accenture recommends DLA move forward with both AI selection Ensemble models into a prototype phase to support a real-world pilot. These two options have significant development overlaps that can be synergized to provide DLA with a more holistic assessment of model

performance along with the implementation challenges. We believe this is a positive step for introducing AI into demand forecasting with a business case summarized by the following points:

- ✓ Minimizes investment risk tied to full-scale production implementation
  - Controlled R&D investment significantly reduces the risk of full-scale production implementation and orients DLA toward a "value-first" AI strategy
- ✓ Enables near-term value/benefit for DLA Planning through the use of AI
  - Proof-of-concept results indicate significant decreases in over-forecasting items ($102M) and minimal reduction to under-forecasting items ($3.5M), providing low risk business value to DLA Planning
- ✓ Enables long-term strategy evaluation to potentially transform DLA Demand Forecasting
  - DLA Planning can expand the exploration of AI algorithms used for forecasting, as well as AI for model selection and decision support to augment and/or automate aspects of demand planning
- ✓ Satisfies Objectives 1.4 and 1.8 of the Director's DLA Strategic Plan 2018-2026[16]
- ✓ Demonstrates a valuable contribution for emergent technology from WSSP R&D
  - WSSP strives to provide DLA process areas with emergent technology capabilities that improve process performance; this would add an AI capability to the Planning toolbox in support of better business outcomes

[16] DLA Strategic Plan 2018-2026 found at https://www.dla.mil/Info/strategicplan/LinesOfEffort/#WarfighterFirst

V. CONCLUSIONS & RECOMMENDATIONS

The AIDF project was the first R&D STP to attempt the development of a custom-built AI proof-of-concept using DLA's existing IT infrastructure. As such, the AIDF team encountered multiple process and technology obstacles to model development. Nevertheless, the team generated a proof-of-concept that demonstrated the business value of modern AI/ML algorithms for improved forecast accuracy.

In addition, the AIDF team initiated the processes necessary to establish a Cloud-based AI Model Training Environment to facilitate future AI use cases. This section highlights key conclusions drawn from the accomplishments above and provides recommendations on next steps for the Agency to facilitate scalable and sustainable AI model development.

Conclusions & Findings

AI Techniques Improve Demand Forecasting

The AIDF STP was able to generate significant improvement in the DPA, sMAPE, and Absolute Dollar error metrics. Key takeaways from the AIDF STP's evaluation of modern models in the demand forecasting process showed that:

- An AI-based "bottom-up" approach to model selection can create item-specific models that improve the forecast when compared to the existing JDA "top-down" rules-based model selection system.
- The AI-powered Ensemble Model helped to dampen the effects of extreme errors. It demonstrated an Over-Forecast Dollar Error decrease of $102M annually, an average DPA increase of 3.5%, and an average decrease in sMAPE of 6.6% when compared to JDA.
- Simple models can outperform complex models when the item's history is truly erratic and unpredictable. In these situations, the AI-powered approach to model selection is beneficial.

Accuracy/Error Metric(s) for Model Evaluation Can Be Biased

In order to verify the impact of the AIDF project, an unbiased metric was needed. New metrics, in addition to existing DLA metrics (DPA and Dollarized Forecast Errors) were analyzed for inherent bias. This analysis led to the discovery of an under-forecasting bias in DPA when applied to DLA's item population. Additionally, the study showed that no metric was able to assess items having fewer than five days of annual demand. This finding supports DLA's existing business logic for identifying forecastability.

Furthermore, extremely sparse demand items are not good candidates for traditional point forecasting. These items may benefit from a range-based, probabilistic approach. This implies that applying the AI-techniques developed during this STP do not demonstrate universal applicability for demand forecasting.

An AI Model Training Environment is Needed for Scaling

As outlined in the AI Groundwork STP, scalable, long-term AI solution development requires a dedicated AI Model Training Environment with the requisite hardware, software, and data access. Such an environment does not currently exist within DLA. To this end, the project team coordinated with key stakeholders across the Agency to initiate the establishment of an AI Model

Training Environment within the DLA Azure Cloud. The team also submitted multiple AI technologies through the DLA Front Door to make basic AI development toolkits available to DLA end users. Finally, the team coordinated with DLA stakeholders to establish data impact level assessments to enable data to be transferred into the Cloud development environments.

All of the above processes remain ongoing, so the project team worked with stakeholders to establish short-term workarounds in order to proceed with modeling. The environment obstacles encountered during the AIDF STP are illustrated in Figure 55 and details on the required workarounds can be found in Appendix B. This Technology Assessment is directly relevant to further research and development.

The Value of a Forecast Model Analysis Framework

The AIDF team created a modular forecasting analysis framework to enable the rapid evaluation of different models. The analysis framework is flexible but specifically focused on demand forecasting, allowing analysis to occur across customizable subsets of items or to experiment with new forecasting approaches. This framework can be used to develop and evaluate new strategies for under-performing items.

Recommendations

Develop an AI Prototype

The AIDF STP demonstrated that AI has potential to improve demand forecasting at DLA. Accenture recommends DLA continue this progress through the creation of an AI forecasting prototype using controlled R&D investment for further development. The AIDF team recommends prototyping the Basic AI Ensemble and Full AI Ensemble models. This approach would allow DLA to continue valuable research and to determine the relative business value versus the management complexity before larger investments are made.

Continue Developing a Scalable AI Model Training Environment

In order to advance AI at DLA, development of an AI Model Training Environment should continue. Additional improvements for long-term scalability include:

- Flexible, cloud-based infrastructure capable of adjusting capacity based on business need in an IL-5 security region.
- The continued introduction and full software approval of modern, open-source AI tools including Anaconda, RStudio, and Python packages for data science. These tools will need to be available to production-level environments to enable efficient integration of AI solutions into production settings.
- Improving the data pipeline to AI development environments to enable the rapid addition of new data elements into the demand forecasting process.

Examine Enterprise Forecast Metrics

DPA consistently exhibits under-forecasting bias for items with more than five days of annual demand and exhibits an over-forecasting bias for items in the 0-4 days of demand range. As DLA's primary metric, DPA introduces supply chain risk that must be managed through manual overrides or supply-side strategies. Based on these results, Accenture recommends that the DLA Planning Process Owner use the outcomes of this STP to examine Enterprise forecast metrics.

Explore Alternative Demand Forecasting Methods

The AIDF STP made promising first steps for implementing AI into the demand planning process. To build upon this progress and further improve the demand forecasting solution, Accenture recommends the following steps:

- Evaluate alternative range-based forecasting methods for sparsely demanded items to improve their materiel availability and reduce inventory levels.
- Continue to investigate the introduction of additional data elements into multivariate AI forecasting models.
- Explore potential processes for systematically introducing new demand forecasts into the existing JDA system to evaluate the impact of CRM cells, and explore methods to transfer forecasts to Supply Planning.

APPENDIX A: DLA DEMAND PLANNING OVERVIEW

DLA's current demand planning process exists as an evolution of ideas, both commercial and federal, to improve the prediction of materiel requirements. The following section review key aspects of existing processes that the AIDF project sought to test and evaluate to improve the forecast accuracy using AI/ML approaches.

Forecast Terminology

Analyzing and measuring forecast accuracy is a complex process and requires a common terminology for discussion. A forecast at DLA is generated each month and in monthly increments. Following the example illustrated in Figure 39, in January 2019 quantities are forecasted for each of the next 12 months (including the current month). In February 2019, another 12 months of forecast is created, which results in multiple forecast values for every month but one. In order to differentiate between these forecast values, the term "lag" is used. Lag represents the relationship between the month being forecasted and the month that that forecast was created. For example, the lag0 forecast for Feb. '19 was generated in February 2019. The lag1 forecast for Feb. '19 was generated in January 2019.

When new forecasts are created the forecast horizon shifts, but the lag position remains the same: e.g., the lag6 forecast is always the 7th month of forecast, however the lag6 time period shifts to the right on the calendar as each subsequent forecast is generated. Similarly, the lag0 forecast always refers to the same month that the forecast was created.

Forecast Created	Jan '19	Feb '19	Mar '19	Apr '19	May '19	Jun '19	July '19	Aug '19	Sep '19	Oct '19	Nov '19	Dec '19	Jan '20	Feb '20
Jan '19	lag0	lag1	lag2	lag3	lag4	lag5	lag6	lag7	lag8	lag9	lag10	lag11		
Feb '19		lag0	lag1	lag2	lag3	lag4	lag5	lag6	lag7	lag8	lag9	lag10	lag11	
March '19			lag0	lag1	lag2	lag3	lag4	lag5	lag6	lag7	lag8	lag9	lag10	lag11

Figure 39: Description of Forecast Lags

By using forecast lag terminology to refer to the position of the forecast rather than the calendar month, we can measure forecasts in the same manner throughout time. The use of lags is particularly important to articulate time periods for forecast accuracy measurements and the inspection of forecast quantities.

Batch Cycles

DLA planning systems run on batch cycles – processes executed at set intervals rather than real-time – on monthly, weekly and daily cadences. The monthly demand month start (DMS) and demand month end (DME) processes govern major demand planning activities. A simplified timeline is shown in Figure 40 including key batch jobs for demand planning.

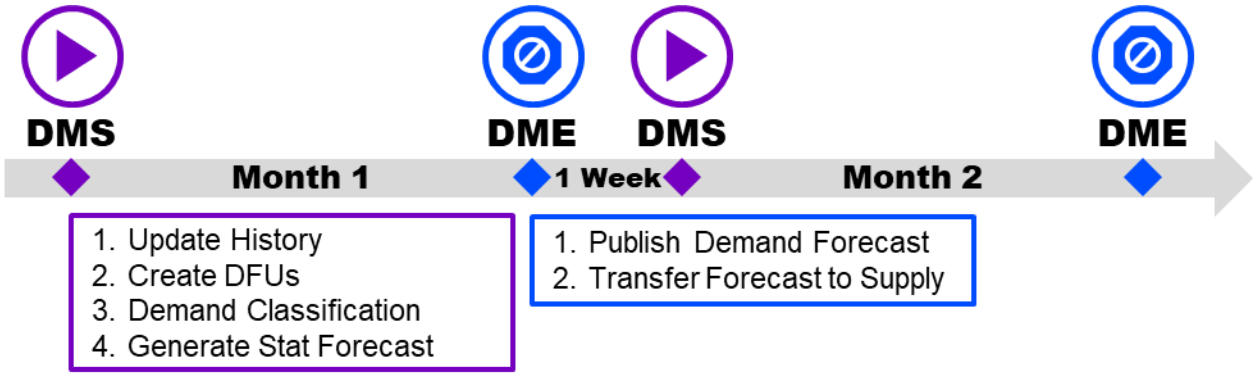

Figure 40: Overview of DLA DMS / DME Planning Cycles

DMS begins the demand planning cycle; starting with updating demand history, creating demand forecast units (DFUs), running demand classification, and generating statistical forecasts. Throughout the month, demand planners can adjust and tune the draft forecast until DME. During DME, the forecast is published and then transferred to supply planning. DME concludes the monthly demand planning cycle. There is a one-week gap between DME and the next DMS execution. Additionally, the collaboration process to obtain customer requirements automatically also executes on a monthly process. Each month, customers submit requirements which DLA review before publishing as part of the overall forecast generation process.

History Processing

Each month during DMS, the demand history is updated in JDA. When DLA's customers submit orders to DLA, sales orders are created in EBS in the SAP Enterprise Central Component (ECC) system. During Nightly Fulfillment, a daily batch job, sales order data is transmitted to JDA, updating the demand history detail table (DHDT). The data is stored at the order level, including customer details. The JDA DHDT contains 60 months of history. The demand history detail table excludes multiple order document types and Non-Cooperative Logistics Supply Support Arrangement (Non-CLSSA) foreign military sales orders.

During DMS, JDA updates the history table (HIST), which is used for statistical forecasting. The HIST contains 36 months of history, aggregated to the monthly level (e.g., multiple sales orders in one month will be aggregated to a total quantity for the month). During aggregation, only active sales orders are considered. Sales orders can be inactive for multiple reasons such as terminal item status, channel switch – customer direct to DLA direct ("CD to DD"), or service requests. The HIST table does not store data for non-forecastable items and make-to-stock (MTS) kit items.

By default, 36 months of HIST and monthly aggregates of order quantity are business rules imposed on the JDA planning system that limit the flexibility of the statistical forecasting algorithms. JDA has capabilities to adjust history start date and weight periods more heavily however it is not used commonly across DLA.

DFU Creation / Forecastability

A DFU uniquely defines a customer-item relationship for forecasting. Specifically, a DFU is comprised of a demand unit, demand group, location, and forecast model type. History processing and DFU creation / forecastability are two components of demand planning processes which run in concert. First, the DHDT is prepared for monthly processing, then DFUs are created, and finally the HIST is updated. An item that is forecastable has corresponding DFUs that are used to

generate its forecasts for each customer. DLA has four distinct methods to create DFUs and thereby forecast requirements, outlined in Figure 41. If an item meets the criteria of any of the four DFU creation methods, it is considered forecastable.

Adequate Demand	Dominant Customer	Collaborative Customer	Open SSR / SPR
Sufficient demand history to meet worldwide activity test (WAT): • Individual rules by profit center • E.g. 4 months of demand in last 13 months with a quantity of at least 10 • Less strict requirements to remain forecastable	Manually activated DFU using a Customer Item Type (CIT) Record: • Forces the creation of a DFU without history or customer collaboration • CIT Type D	Formal collaboration for customers to submit projected constrained supply plans • Customers submit through Demand Data Exchange • A DFU is created to store collaborative requirements for evaluation and addition into the DLA Forecast	Method for customer to submit requirements who are not part of formal collaboration: • DFU is created for Special Supply Requests (SSR) and Special Program Requests (SPR)

Figure 41: Methods to create DFUs used to generate demand forecasts.

It is important to note that while stat forecasts are generated for collaborative items, collaborative forecasting is a separate monthly process. The primary focus of this R&D project was to assess new statistical forecast algorithms, that have potential to improve the forecasting power of non-collaborative, purely statistically-forecasted items, and can also be used to generate exceptions for collaborative forecast inputs through comparison, allowing demand planners to research customer inputs to verification.

Demand Classification and Model Tuning

Forecast algorithms are mathematical equations used to forecast future requirements. Demand classes are segments of the item population that use different forecasting methods to generate the forecast DLA has six distinct forecast algorithms available in JDA used to create demand forecasts. Each model has tunable parameters, making manual evaluation of potential models for each DFU a herculean undertaking. Instead, DLA uses a process called Demand Classification to align forecast algorithms to DFU demand history. Each classification has one to four forecast algorithms to evaluate, reducing the manual effort to select a statistical forecast algorithm.

There are three sub-processes in demand classification:

I. **Classify** – Performs analysis on history to categorize DFUs into one of ten demand classes based on history characteristics.
II. **Optimization** – Recommends the best forecast algorithm and appropriate parameters based on the classification
III. **Add/Update** – Allows recommendations to be committed to production

DLA uses eight of the ten JDA demand classifications. Details of the eight classes, criterion, and potential algorithms are outlined in Figure 42. Note, the two additional JDA classes that DLA does not use are assigned into management control.

Class	Criterion	Algorithm(s)
Continuous Seasonal	• Maximum 15% of history periods are zero • Seasonal pattern found	• Fourier • Holt-Winters • Lewandowski
Continuous Non-Seasonal	• Maximum 15% of history periods are zero • No seasonal pattern found	• Fourier • Lewandowski
Erratic Seasonal	• Maximum 35% of history periods are zero • Seasonal pattern found	• Fourier • Holt-Winters • Lewandowski
Erratic Non-Seasonal	• Maximum 35% of history periods are zero • No seasonal pattern found	• AVS Graves • Croston • Fourier • Lewandowski
Lumpy Seasonal	• Maximum 95% of history periods are zero • Coefficient of Variation > 2 • Seasonal pattern found	• Fourier • Holt-Winters
Lumpy Non-Seasonal	• Maximum 95% of history periods are zero • Coefficient of Variation > 2 • No seasonal pattern found	• Croston
Obsolete	• Trailing 18 periods are zero demand	• Lewandowski
Management Control	• 12 or fewer periods of history • All other conditions not covered	• Fourier

Figure 42: Overview of demand classes used at DLA to select forecast algorithms

After the classification process is complete, a tuning process is run in order to generate algorithm and optimal parameter values recommendations. Using the DFU class and demand history, tuning recommends up to three algorithms and the corresponding parameters. During the tuning process, DLA demand planners can adjust outliers by replacing them with the outlier limit or mean, adjust the threshold to determine outliers, choose the forecast algorithm for items in management control, determine the smoothing algorithm and determine the regression algorithm.

After the classification and tuning process are complete, recommendations can be committed to demand planning using the Add/Update model process.

DLA Performance Metrics

The goal of forecast metrics is to measure and monitor the accuracy of the forecast. Because the demand forecast projects requirements in the future, we must wait until actuals accrue to measure accuracy. Forecast metrics can evaluate one lag or a range of lags, additionally, metrics can be aggregated across multiple forecast generations. For example, a three-month lag0 (3ML0) metric evaluates the forecast accuracy for the lag0 period across three consecutive forecast generations.

To evaluate a forecast generated in January 2019, we must select the measurement period (forecast lags) and metric. In Figure 43, we have actual demand for lag0 to lag5 (effectively

demand through June of 2019). When calculating a forecast metric, we must first identify the period's error, which is the difference between the forecast quantity and the actual demand, as shown in the equation below.

$$Forecast\ Error_{lagi} = Forecast_{lagi} - History$$

In Figure 43, lag0, lag1, lag4, and lag5 are over-forecasted for that individual period. Lag2 and lag3 are under-forecasted for that individual periods. However, if we aggregate across lag0 to lag4, the total error is nearly zero.

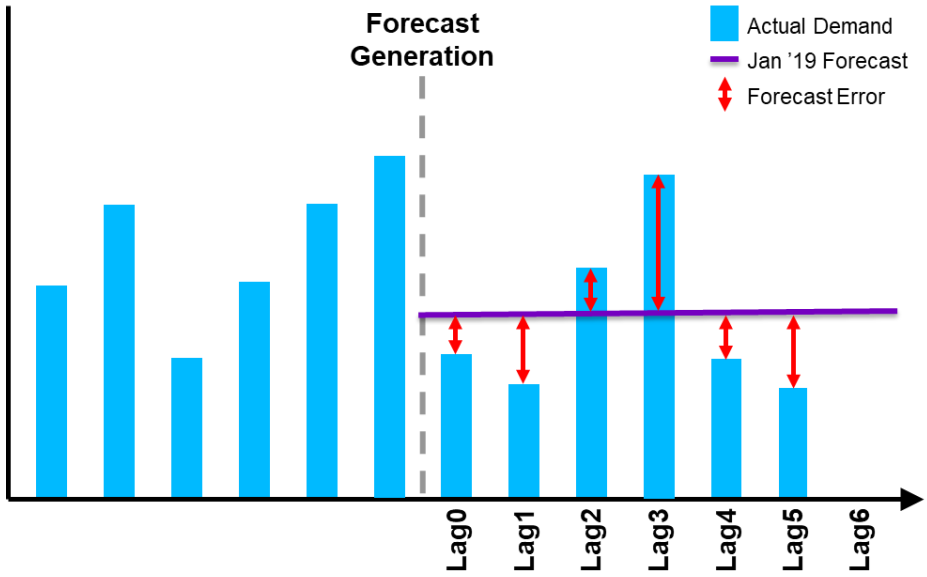

Figure 43: Example forecast vs. demand for a forecast generated in January of 2019

As demonstrated in the example, there are a wide range of options to assess forecast accuracy. For an agency such as DLA, with a very diverse product mix, selecting metrics that evaluate the proper timeframes is challenging. DLA uses several forecast metrics and aggregation periods to measure forecast accuracy Agency-wide. DLA uses three primary metrics, Demand Plan Accuracy (DPA), Net Percent Forecast Error (PFE), and Absolute Percent Forecast Error (APFE).

- DPA represents the percentage accuracy of the forecast and is bound between 0 and 100%. The goal is to maximize DPA in order to generate the best forecasts.
- PFE provides the ability to understand over or under forecast bias and the magnitude of the error. By measuring error, the goal is to have a 0% PFE. PFE is useful to compare individual DFUs however, aggregating PFE offsets positive and negative values therefore it can mask error across the population.
- APFE is similar to PFE however it uses the magnitude value of the error. APFE can be used to aggregate results when the direction of the error is not important, just the size of the error. Similar to PFE, the goal is to measure a 0% APFE

When aggregating PFE and APFE forecast metrics across multiple DFUs, DLA monetizes forecast and history in the calculations. DLA uses multiple metrics and measurement periods in order to identify different issues.

All metric equations can be found in Appendix F.

APPENDIX B: TECHNOLOGY ASSESSMENT

Accenture identified that the efficacy of AI within DLA's landscape is dependent upon the intertwined relationships between four critical enablers: Cybersecurity, Data, Software Tools, and Compute Power. These enablers in unison create a robust AI Development Platform (Figure 44). The three technical enablers (data, software tools, and compute power) drive AI development but must also comply with the existing cybersecurity measures that underlie and enable all operations at DLA.

Although the specifications of each enabler will differ on a case-by-case basis, this Technology Assessment aims to determine the minimum viable solutions for the AI Development Platform for both the AIDF project and forthcoming AI use cases. Accenture evaluated DLA's current technological architecture and IT procedures against the prerequisites for successful AI development by exploring the following three key areas:

1. **The AI Development Platform**
2. **AI Development Platform Options**
3. **Environment Process Standardization**

Accenture recommends DLA continue to leverage and expand existing data science tools and resources that promote the AI Development Platform. However, in order to allow the technical enablers to work in tandem with current cybersecurity requirements, DLA must establish streamlined processes that support the flexibility, scalability, and speed that differentiates AI from traditional IT. The success of DLA's AI journey is, therefore, contingent upon both the establishment of a secure AI Development Platform and a process realignment to support the configuration of that Platform.

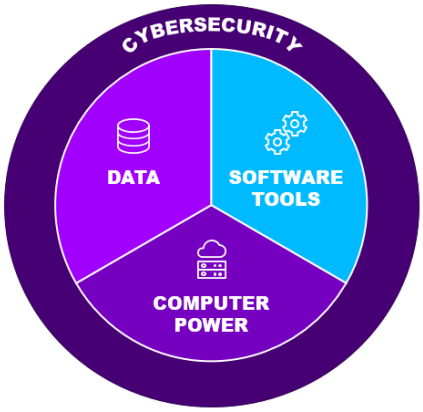

Figure 44: The AI Development Platform

AI Development Platform

AI development relies on rapid learning, fast failure, effective design iterations, and consistent adaptation. Cybersecurity acts as an underlying control to protect DLA's assets, while enabling the other three technical components to securely drive development. Production data is needed to train, evaluate and test the AI model to better predict demand by intaking both time-series and attribute data. The software tools interface with this data to develop the model, while the compute power provides the processing to train and evaluate the model. In order to assess the most viable development approach for the AIDF project, it is important to first understand the relationship of these critical enablers and how they work conjointly to create a robust AI Development Platform.

Cybersecurity

Cybersecurity defines all operations at DLA and is the foundational enabler that must be considered at each step of the development process. Therefore, the usage of data, software tools, and compute power must all be considered within the scope of existing security standards.

The AI model development process requirements differ from traditional IT in because both greater processing power and real historical data from the production environment at the development stage (previous application development used stand-in data). The security requirements to

manage and use production data, however, remain the same, despite AI now requiring use of production data earlier in a model's lifecycle. Both Cloud services and on-premise IT are viable solutions for infrastructure and have established standards for secure application at DLA. The security requirements for on-premise IT are defined by the Federal Information Security Management Act (FISMA), whereas the controls for Cloud services are outlined by the Federal Risk and Authorization Management Program (FedRAMP). Both hold each respective infrastructure service to the same security requirements and controls defined by the NIST 800-53 document and Federal Information Procession Standard (FIPS) 199.

The Department of Defense supplemented FedRAMP's security measures for Cloud services by creating additional security requirements for Cloud Service Providers (CSPs) specifically for within the DoD. The program, FedRAMP+, outlines qualifications for four different data Impact Levels (ILs) based on the sensitivity and confidentiality level of the information (see Figure 45 for further details on the FedRAMP+ Impact Level metrics). The DLA EBS Application Owner has classified all DLA data as Impact Level 5. IL5 is for Controlled Unclassified Information related to National Security Systems and is the highest level of security for unclassified data. Although this is the standardized classification for DLA data, it is important to note that within the scope of the AIDF project the needed data subset underwent an approval process to allow data to be put into DADE for test and development. Microsoft Azure Cloud, a CSP that DLA has already begun investing in, has dedicated Defense regions (DLA Azure Cloud) that meet FedRAMP+ Impact Levels 2, 4, and 5. Infrastructure as a Service (IaaS) is currently the only service available at IL5 though and consequently, the only service permissible by DLA. DLA is primarily focused on investment in environments that meet IL4 and IL5, as those environments would automatically encompass any requirements for information that falls under IL2.

Impact Level	Information Sensitivity	Security Controls	Location	Off-premises Connectivity	Separation	Personal Requirements
2	Public or Non-critical Mission Information	FedRAMP v4 Moderate	US/US outlying areas or DoD on premises	Internet	Virtual/Logical PUBLIC COMMUNITY	National Agency Check and Inquiries (NACI)
4	CUI or Non-CUI, Non-critical Mission Information, Non-National Security Systems	Level 2 + CUI Specific Tailored Set	US/US outlying areas or DoD on premises	NIPRNet via CAP	Virtual/Logical Limited "Public" Community, Strong Virtua: Separation between Tenant Systems & Information	US Persons ADP-1 Single Scope Background Investigation (SSBI), ADP-2 National Agency Check with Law and Credit (NACLC), Non-Disclosed Agreement (NDA)
5	High Sensitivity CUI, Mission Critical Information, National Security Systems	Level 4 + NSS & CUI Specific Tailored Set	US/US outlying areas or DoD on premises	NIPRNet via CAP	Virtual/Logical FEDERAL GOV. COMMUNITY, Dedicated Multi-Tenant Infrastructure Physically Separate from Non-Federal Systems, Strong Virtual Separation between Tenant Systems & Information	
6	Classified SECRET, National Security Systems	Level 5 + Classified Overlay	US/US outlying areas or DoD on premises, CLEARED/ CLASSIFIED FACILITIES	SIPRNet DIRECT with DoD SIPRNet Enclave Connection Approved	Virtual/Logical FEDERAL GOV. COMMUNITY, Dedicated Multi-Tenant Infrastructure Physically Separate from Non-Federal AND Unclassified Systems, Strong Virtual Separation between Tenant Systems & Information	US Citizens with Favorably Adjudicated SSBI & SECRET Clearance NDA

Figure 45: Overview of FedRAMP+ Impact Levels

Cloud services and on-premise IT differ in responsibility for security requirements. Cloud service

models share the responsibility between DLA and the CSP to procure, provision, operate and secure; whereas with an on-premise solution, DLA is responsible for all aspects of infrastructure management and cybersecurity. As seen in Figure 46, DLA is responsible for what is above the security boundary line, and the CSP is responsible for what is below. The security boundary line can also shift up and down based on which Cloud service model (IaaS, PaaS or SaaS) is selected.

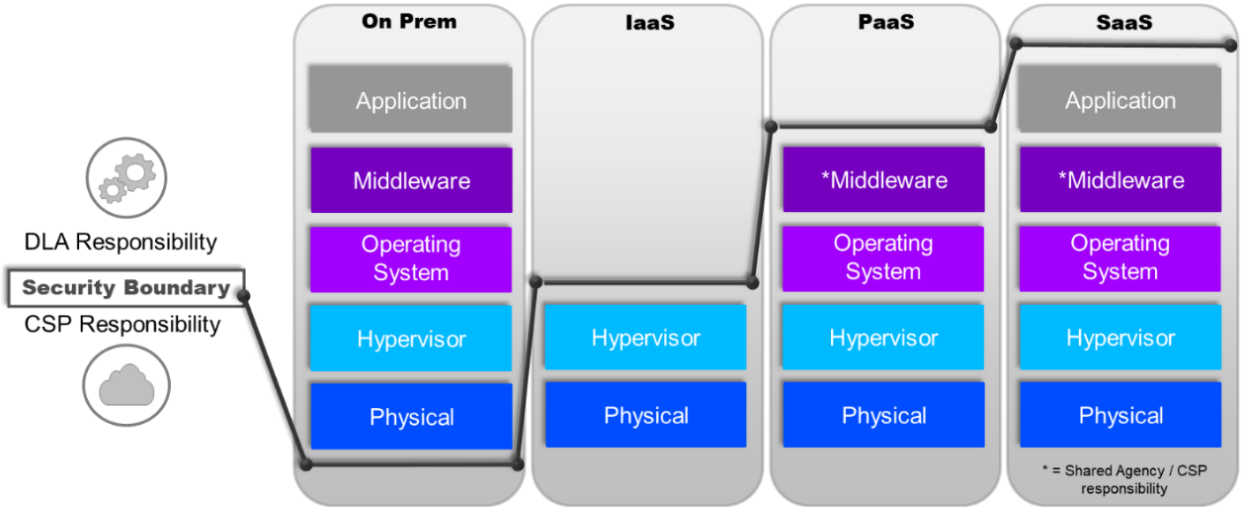

Figure 46: Division of Responsibility for On-Premise and Cloud Service Solutions

Data

AI models learn from historical data and, therefore, access to production data is critical to ensuring that the model makes accurate predictions. Because of this data need, IL security controls for data transfer, storage and usage must be applied to all AI Model Training Environments. Data ILs are determined in accordance with the NIST Special Publication 800-60 Version 2 Revision 1's guide on mapping information to Impact Levels. It examines the potential impact that the loss of the following objectives would have "on organizational operations, organizational assets, individuals, other organizations, or the Nation:"

1. **Confidentiality**: Preservation of "authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information"
2. **Integrity**: Protection "against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity"
3. **Availability**: Assurance of "timely and reliable access to and use of information"[17]

Each data category has a recommendation within NIST's guidelines for its level across the three qualifications, along with various caveats that if met exclude the data from the originally recommended categorization.

The AIDF project extracted a portion of DLA's data primarily from the DLA Office of Operations Research and Resource Analysis (DORRA) Date Warehouse and the Enterprise Data Warehouse (EDW). Only historical data was utilized for the model with exception of supplemental contextual data, which does not have historically archived values (e.g., item attributes, weapon

[17] https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v2r1.pdf

system classification). DLA's data in its entirety is classified as IL5. However, using the principles outlined by NIST, the categories of data that are needed for the AIDF map to low Impact Levels (Figure 47). Although NIST's impact level guide was created separately from Cloud Services (and FedRAMP+ Impact Level metrics are specifically for CSPs), it was determined through coordination with the R&D Office and the CDO that the low impact level ranking from NIST was sufficient for use within the FedRAMP+ DADE IL4 environment.

Category			Security Objective		
			Confidentiality	Integrity	Availability
3.1		Administrative Management			
	3.1.1	Facilities, Fleet, and Equipment Management Information Type	Low	Low	Low
3.4		Supply Chain Management			
	3.4.1	Goods Acquisition Information Type*	Low	Low	Low
	3.4.3	Logistics Management Information Type	Low	Low	Low

Figure 47: Data extracted for the AIDF project and its corresponding potential Impact Level

Because the IL5 STIGed image has not yet been implemented into the DADE environment, data usage within Microsoft Azure's Test and Development (T&D) environment is restricted to IL4 and below. Therefore, recategorizing the extracted data to IL4 allows the data to be batch transferred and used within the scope of the AIDF project. This process of mapping and recategorizing the data is only relevant for Cloud service solutions and would not be applicable for on-premise IT (as FedRAMP+ was designed strictly for CSPs).

Through the steps taken to recategorize AIDF's extract of data, Accenture identified a pressing need for further data governance at DLA. Policies are needed to define requirements and accountability by bridging DLA's operational processes with its data requests to more thoroughly include data management. Throughout the span of the project, there was no prior approval processes in place for data recategorization, nor was there clear ownership of overall data quality.

An AI model's reliability mirrors that of the data it learns from. If the data is unclear, inaccurate or mislabeled, the model will project the same errors into its own system. Therefore, the quality and clarity of the data is critical to the performance of the model. Data recategorization requests will increase with the rise of data science and the simultaneous larger movement towards the Cloud at DLA. Completing manual approval processes for each project impedes consistent data visibility and risks potential missteps from human error. The institution of a streamlined process that defines data access management roles would not only encourage AI development, but it would ultimately benefit security by maintaining consistent protocol for data usage and maintenance.

Software Tools

Cutting edge resources are needed in order to support execution of rapid AI solutions. The data science tools that are required for model development include software, programming languages and code libraries. These tools predominantly fall into the following three distribution categories:

1. **Enterprise Software:** commercially developed software where users purchase a license for use but are restricted from modification or further distribution (i.e. SAS, Revolution, or Anaconda Enterprise)

2. **Open Source:** publicly developed software that is maintained by a global community of data scientists and can be copied, modified or redistributed without associated fees (i.e. Python, RStudio, Anaconda)
3. **Managed Services:** enterprise software that is purchased but owned and maintained by the vendor (i.e. DSVM)

The need of new software necessitates an approval process within DLA to validate that the identified software is secure. An IT Capability Request (ITCR) is submitted through the Front Door, which starts a Risk Assessment for use of the submitted product within the DLA network. If approved by the CTO, the software tool is added into IT Solutions Document (ITSD) Accenture assessed leading data science tools.

Python

Python, an open-source programming language, is fundamental for model development. Python is approved software on the ITSD. The development process is specifically reliant on Python's open-source data science libraries. These libraries comprise of optimized code that perform core tasks and have been maintained globally by top data scientist groups, such as Google. Utilization of the libraries increases efficiency and efficacy of model development by allowing developers to focus on testing and model evaluation instead of rebuilding basic programming structures and algorithms.

The Python libraries required approval through DLA's Risk Assessment process. Accenture identified 35 libraries as valuable and necessary for the AIDF project (Figure 48). While the majority of the libraries needed for future AI development fall within the 35 libraries identified for AIDF, new libraries will be required depending on the scope of future AI prototypes.

Library Name	Version
absl-py	0.7.1
astor	0.7.1
cycler	0.10.0
Cython	0.29.7
gast	0.2.2
grpcio	1.20.1
h5py	2.9.0
keras	2.2.4
keras-applications	1.0.7
keras-preprocessing	1.0.9
kiwisolver	1.1.0
Markdown	3.1
matplotlib	3.0.3
mock	3.0.4
numpy	1.16.3
pandas	0.24.2
pip	19.1.1
prophet	0.1.1
protobuf	3.7.1
Pyparsing	2.4.0
python-dateutil	2.8.0
pytz	2019.1
PyYAML	5.1
scikit-learn	0.20.3
scipy	1.2.1
setuptools	41.0.1
six	1.12.0
Statsmodels	0.9.0
tensorboard	1.13.1
tensorflow	1.13.1
tensorflow-estimator	1.13.0
tensorflow-gpu	1.13.1
termcolor	1.1.0
Werkzeug	0.15.2
wheel	0.33.1

Figure 48: Identified Python libraries requested for approval for AIDF

Anaconda

Anaconda is the "world's most popular Python data science platform" and environment manager, containing extensive data science libraries for both Python and R.[18] Many of the Python libraries have dependencies that need to be thoroughly tracked and managed; Anaconda controls these dependencies and eliminates added time spent by resources to manage manually. Additionally, Anaconda curates packages and maintains environments to ensure compatibility and improve performance on both GPU and CPU-based algorithms. For example, TensorFlow is an open source software library that is based on data flow graphs and supports machine learning methods. Its implementation has substantially faster processing speeds with Anaconda than manual installation.[19] Accenture's AI Technology Radar has identified TensorFlow as a valuable product within the Artificial Intelligence Ecosystem and recommends its use in full-scale production

[18] https://www.anaconda.com/wp-content/uploads/2019/03/2018-06-Anaconda-State-of-Data-Science.pdf
[19] https://www.anaconda.com/tensorflow-cpu-optimizations-in-anaconda/

deployments. Anaconda effectively manages the TensorFlow package to allow developers to focus on algorithmic development instead.

Security concerns prolonged the approval of Anaconda usage at DLA. Reciprocity was initially requested, because other DoD agencies have already authorized Anaconda and actively utilize it. Reciprocity was unsuccessful, however, and Anaconda only received approval for usage on "air-gapped" laptops for the AIDF project. Due to the importance of Anaconda for development and a lack of alternative software options, this restricted approval ultimately determined which AI Development Platform option was pursued for completion of the project.

RStudio

While Python accomplishes the needs of the AIDF effort and future data science projects, R, and its accompanying environment RStudio, is an effective machine learning system and valuable to consider for future development. Accenture's AI Technology Radar lists it as a prevalent part of the data science ecosystem. However, Accenture could not test RStudio within this period of performance. DLA should continue to explore and evaluate the potential of RStudio as an AI enabler for future AI prototypes.

DSVM

Azure Data Science Virtual Machines (DSVMs) are customized images that are pre-installed with a broad range of data analytics tools that are pre-configured for use in the Azure environment. Azure DSVMs can allow data scientists access to popular analytics tools without overhead costs. DSVMs are IaaS offerings, but the DSVM image provided in Azure Government is the same image provided for commercial use. In order for the DSVM to meet IL5 requirements, each software tool within DSVM not already on DLA's ITSD require an individual Risk Assessment from J61 and STIG implementation. These types of solutions will be more valuable once Microsoft Azure provides them as PaaS.

Compute Power

AI development requires infrastructure to power models to learn and perform. However, infrastructure specifications differ with each AI use case. This creates a unique need for compute power to be flexible and scalable in order to tailor the compute power to the specific prototype. Both on-premise and Cloud services are feasible options for AI infrastructure. On-premise solutions are reliable and allow DLA to maintain a level of control that the Cloud generally does not permit. The cost and time investment of setup, management, and maintenance though is exponentially greater than the Cloud. Additionally, On-Prem agility is limited to the constraints of what can be procured and consequently limits achieving the true value of AI. Alternatively, Cloud services utilize virtual machines (VMs) with a variety of storage and processor types that can be easily scaled up or down as needed. This grants DLA the opportunity for flexibility to adjust data size throughout the prototype phase, along with long-term scalability of AI models across DLA.

DLA's Strategic Plan 2018-2026 notes exploiting "an innovative culture" by driving "business improvements with emerging technologies."[20] Cloud services represent a technical cultural shift that should be embraced in order to promote AI capabilities at DLA. DLA Leadership has already established an existing subscription with Microsoft Azure Cloud with specific regions for the DoD that meet FedRAMP+ controls for IL 2, 4, and 5. MS Azure includes software tools in pre-imaged

[20] https://www.dla.mil/Portals/104/Documents/Headquarters/StrategicPlan/DLAStrategicPlan2018to2026AP.pdf

VMs with supporting machine learning capabilities. Although the benefits of Cloud services exceed its limitations, it's important to emphasize that Cloud services must share the security responsibility with the vendor (as addressed and discussed further within the Cybersecurity section). Further, the success of the Cloud is dependent upon having a standardized configuration process that promotes repeatability and delivers upon the process speeds that differentiate the Cloud from on-premise IT.

Cybersecurity, Data, Software tools, and Compute power work cohesively to enable the AI Development Platform. AI introduces specific data and compute power requirements at the development stage that historically weren't needed for projects until production. In order to adapt DLA's configuration processes to reflect the new needs, it is critical to understand the four enablers and their interconnected relationships (Figure 49). Although each component's needs will vary depending on the use case, they must mesh together to support the overarching requirements of the prototype and DLA's preexisting standards.

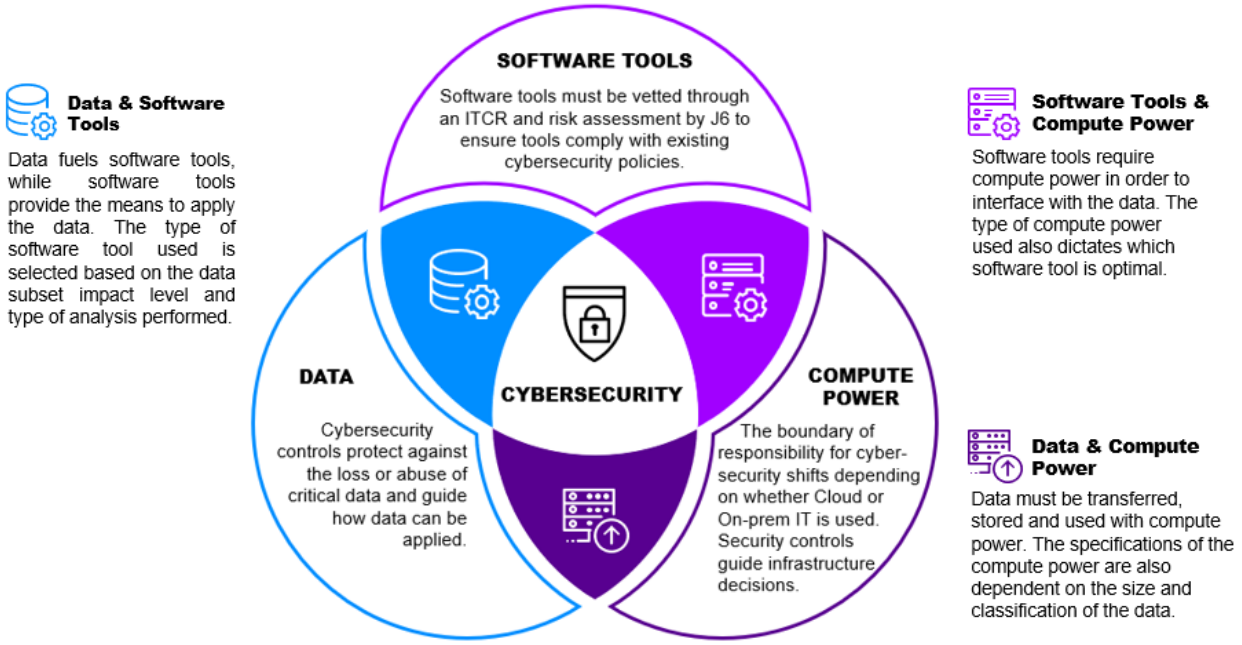

Figure 49: Overview of the relationship dependencies between the four core enablers of the AI Development Platform

AIDF AI Development Platform Options

With a focus on the four key enablers of the AI Development Platform, Accenture assessed the benefits and drawbacks to determine viability and long-term success of potential development solutions. Enterprise SAS 9.4 and Big Data Stack were deemed infeasible after initial consideration and removed from further assessment within the scope of the AIDF project. DLA has set up SAS to develop analytical products however, additional licenses would be needed to cover the more advanced methods that are essential for AI development. Procurement of these licenses would take longer than the time allotted for the AIDF's period of performance and therefore SAS was concluded unviable. However, even with additional time for procurement, it is

recommended that specialized, open-source tools within the data science ecosystem be prioritized over SAS for AI development.

The Big Data Stack was also not pursued because it's still in the early stages of development and was designed as a data lake, not an AI/ML development environment. The Big Data Environment includes Spark, which can be used for AI/ML development however further configuration is required. The three solutions outlined below were identified as feasible and evaluated further by Accenture for the AIDF project.

Interim "Air-gapped" Laptop Solution

The Interim "Air-gapped" Laptop Solution involves the use of two GFEs per developer—one to access data and one to process the data with analysis software in an air-gapped system. The "air-gapped" laptop equipped with both the data and tools conducts all data science activities related to generating and assessing demand forecasting techniques within DLA (Figure 50). The "air-gapped" laptop functions as on-premise IT, which allows developers to access and transfer the needed data via DVDs. Testing in the AIDF project environment is limited to a small team of developers working on algorithm development. Team size will generally vary based on the scope of each AI use case; however, for the AIDF project, a small team is only suitable as an interim solution, and more resources would be needed to pursue long-term implementation.

The most significant benefit of this solution is its immediate viability in large part because it's on-premise IT that permits Anaconda. With the elimination of sharing the security responsibility with a third-party CSP, the enablement process timelines for data and software tools are expedited. However, there are notable drawbacks to this solution. The inefficiency caused by DVD data transfer does not allow the development team to take full advantage of AI. Additionally, the Git repository tracks changes and typically is available for developers to simultaneously access. Shared access to version control and development of the same code allows the team to overlap modules instantly. Without the ability to do so, collaboration between developers and identification/resolution of issues is confined to scheduled overlaps. Lastly, this solution does not enable long-term adoption, as scalability is restricted to the additional procurement of hardware.

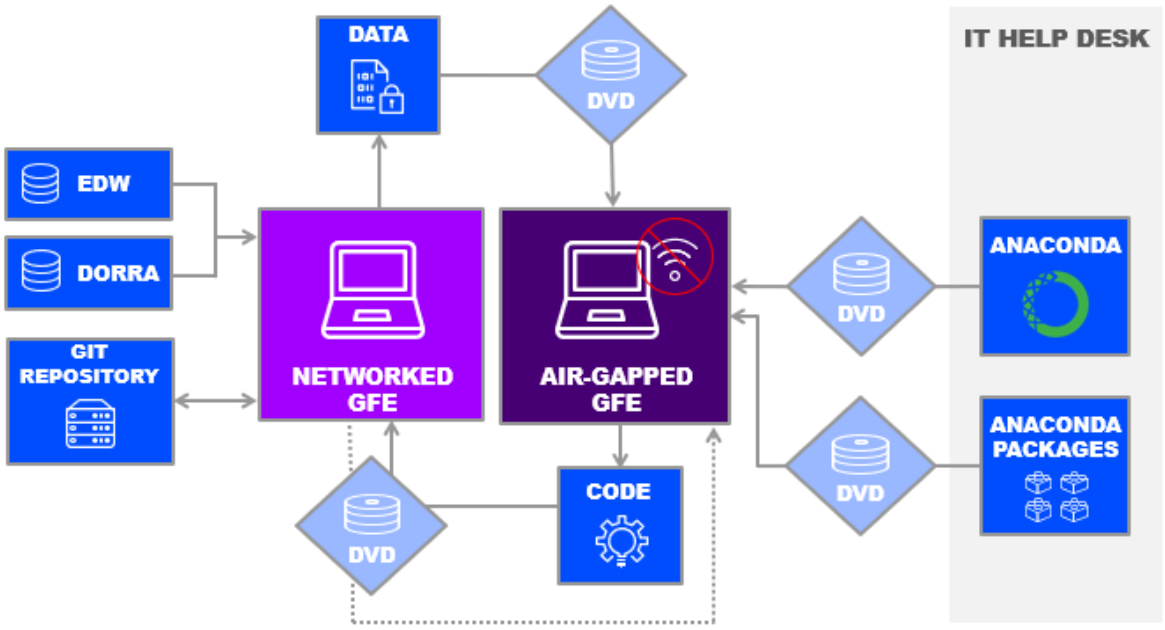

Figure 50: The Interim "Air-gapped" Laptop Solution

Microsoft Azure Solution

The Microsoft Azure Solution leverages the use of Cloud services instead of On-premise IT. A Virtual Machine would be configured and scaled to fit the specific infrastructure needs of the AIDF project (Figure 51). Because DLA currently only has a T&D environment fit to meet IL4 requirements, the extracted data needs to be mapped and recategorized from its current IL5 rank to IL4 using NIST's Impact Levels guide. Once recategorization is approved, the production data is allowed in T&D in Microsoft Azure, which could be maintained by central authority (ACE or R&D, for example).

This solution allows teams to use production data and Cloud services—the optimal compute power for AI development—without breaching the security restrictions of the T&D environment. Further, Cloud removes the constraints on collaboration from restricted access to the Git repository that would be experienced with the Interim "Air-gapped" Laptop Solution. The development efficiency and scalability also improve with this solution, which thereby enhances the developers' abilities to rapidly innovate and iterate designs. Upon completion of the AIDF project, project items can be stored within the Cloud or services can be turned off until needed for future AI prototypes. However, this solution is limited to the timeline of the recategorization process. Future AI projects will continually experience delays if the approval process of recategorization is individually determined on a case-by-case basis. Therefore, in order for this solution to be effective, DLA should define key data requirements and further standardize steps for recategorization and approval.

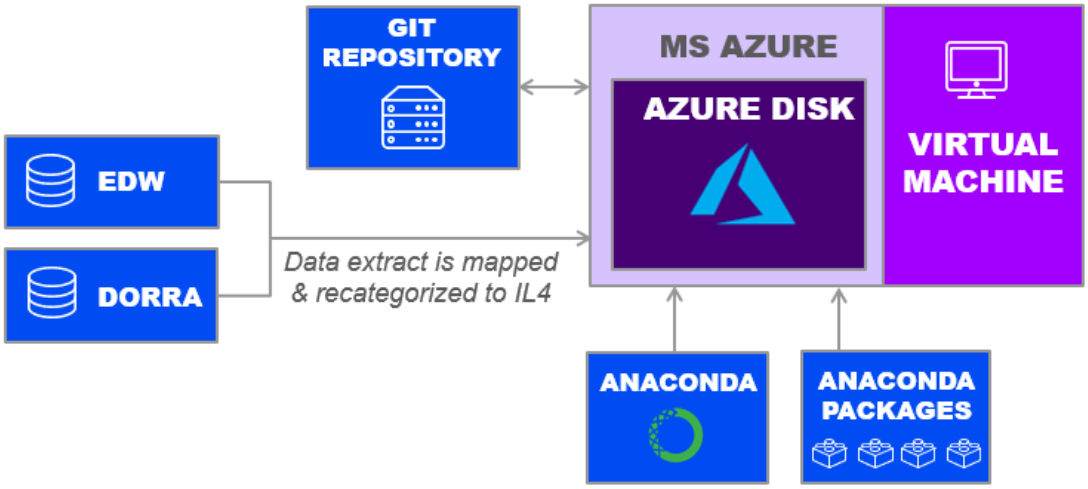

Figure 51: The Microsoft Azure Solution

Microsoft Azure with VDI Enablement Solution

In the Microsoft Azure with VDI Enablement solution, the Microsoft Azure Solution is enhanced to simplify and streamline tool and data access. DLA would use multiple Virtual Machines (VMs) with ranging hardware options for processor type (CPU and GPU) and size. The VMs could be used for AIDF-like projects, along with future prototype projects across the agency (Figure 52). Although the Cloud service would encounter the aforementioned security restrictions, DLA would remain in compliance by establishing an IL5 environment in T&D. This would be completed by putting an IL5 image into T&D and STIGing the OS and standard data science tools to IL5 requirements. The service would then maintain a core image that is STIGed and attached to the VMs of differing sizes.

There are numerous benefits to this solution. The Cloud services that are needed to support a specific AI use case could be turned off once the prototype is completed, ending excess cost accrual. Services could thereafter be reactivated and deactivated based on tailored needs for future AI use cases. Further it would offer increased scalability and encourage collaboration by conjointly using a single environment. The increased initial setup timeline and investment would be an unavoidable limitation, but once completed, there would be improved efficiency for future VM setup and attachment of the STIGed image.

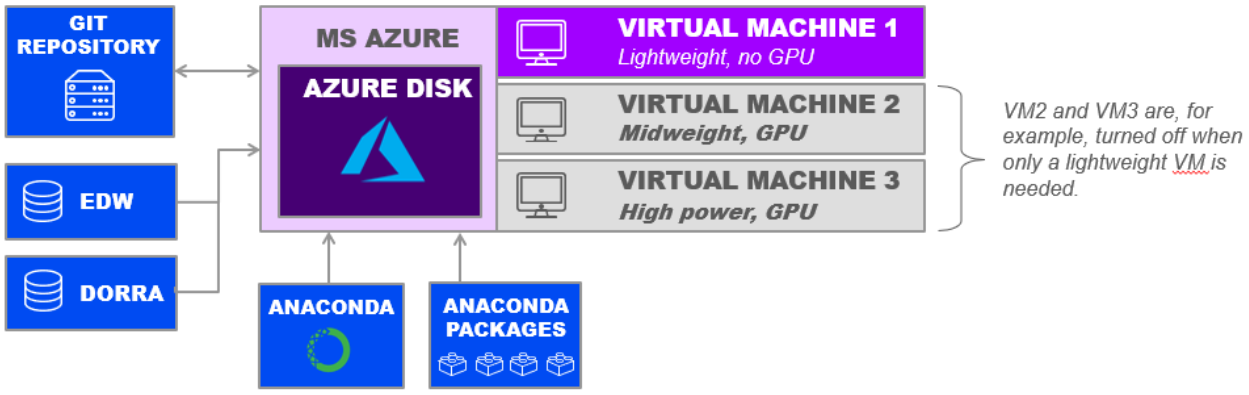

Figure 52: The Microsoft Azure with VDI Enablement Solution

The Interim Laptop, Microsoft Azure, and Microsoft Azure with VDI Enablement solutions offer incremental methodologies for AI model development (Figure 53). The Interim Laptop solution is an effective workaround for the FedRAMP+ security obstacles and ultimately is the only solution currently approved to use Anaconda. Although it does not have the capacity to support long-term AI implementation at DLA, it is a concrete first step in enabling AI development and transitioning DLA's IT landscape to fit data science needs. The Azure solution scales the interim laptop solution by eliminating both inefficiencies caused by DVD transfer and constraints on collaboration through use of MS Azure. Lastly, the Azure with VDI enablement solution is the ideal state for AI development at DLA, as it builds on the Azure solution by establishing capabilities to drive AI activities long-term.

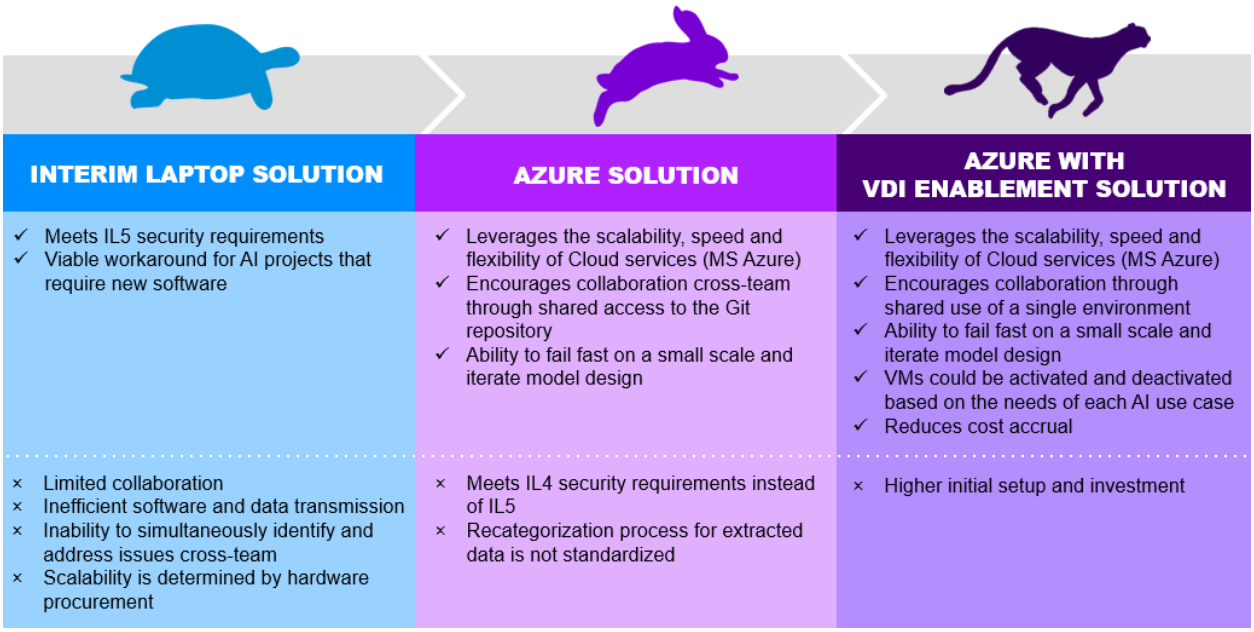

Figure 53: Overview of AI Development Platform options for the AIDF Project

Environment Process Standardization

Standardization promotes innovation and the dissemination of forward-looking ideas through structured methods and procedures. The technical obstacles that were met throughout the AIDF project were largely due to a lack of foundational processes to approve and enable the technology needed for AI development. The data science ecosystem is equipped with robust tools and technology, and DLA has business processes for configuration of traditional IT needed in the past. However, there is currently a missing link between the new resources necessary for AI and DLA's existing configuration processes. Each approval procedure—such as the data recategorization or the Python libraries approval—had a lack of guidelines on the steps that needed to be taken and instead was created ad-hoc. As seen in Figure 55, this resulted in cumbersome processes with pain points and delayed timelines that impeded the project's progress.

As DLA continues to integrate AI into operations, enablement will need to be repeated for future AI use cases. The success of DLA's AI journey and long-term growth, therefore, depends on shifting the workflow to include guides and standards for these processes instead of creating interim solutions on a case-by-case basis. Standardizing the enablement processes would encourage further AI development, while also preventing issues with consistency, efficiency and

productivity. DLA should work with and align to the larger DoD, who already leverage open-source software tools (i.e. Anaconda, Containerization) for development. Additionally, as cybersecurity is a main focal point within the configuration process, Accenture recommends that DLA adopt the Development Security Operations (DevSecOps) framework to interlock security into project operations by incorporating security into the design phase. Traditionally, security and operations are determined afterwards (Figure 54); however, DevSecOps instead allows developers to create secure solutions by meshing security into the existing tools and procedures without hindering development speed. Aligning development, security, and operations addresses and mitigates procedural obstacles by streamlining all aspects of the model's lifecycle throughout project delivery. The establishment of procedures and a solidified framework for the enablement of AI is the critical link to effectively bridge the needs of the AI Development Platform with DLA's IT landscape.

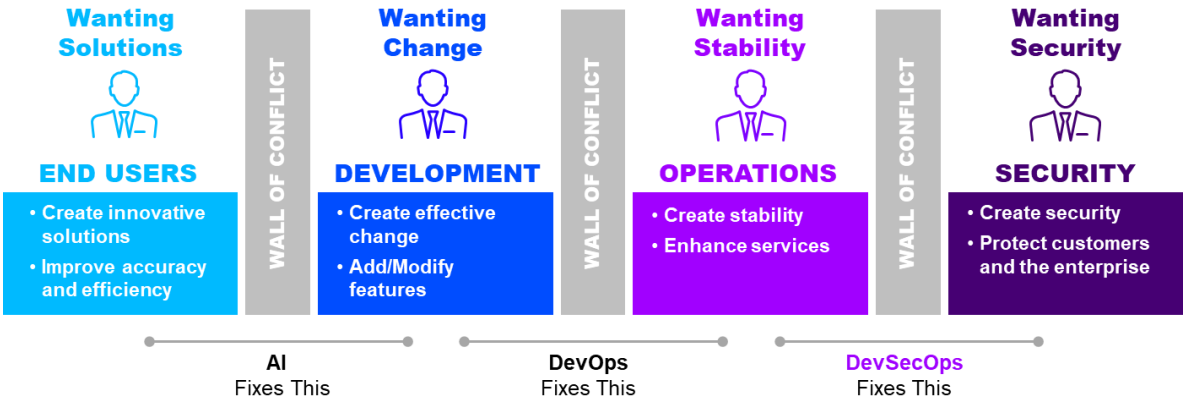

Figure 54: DevSecOps Breaks Down Walls of Conflict

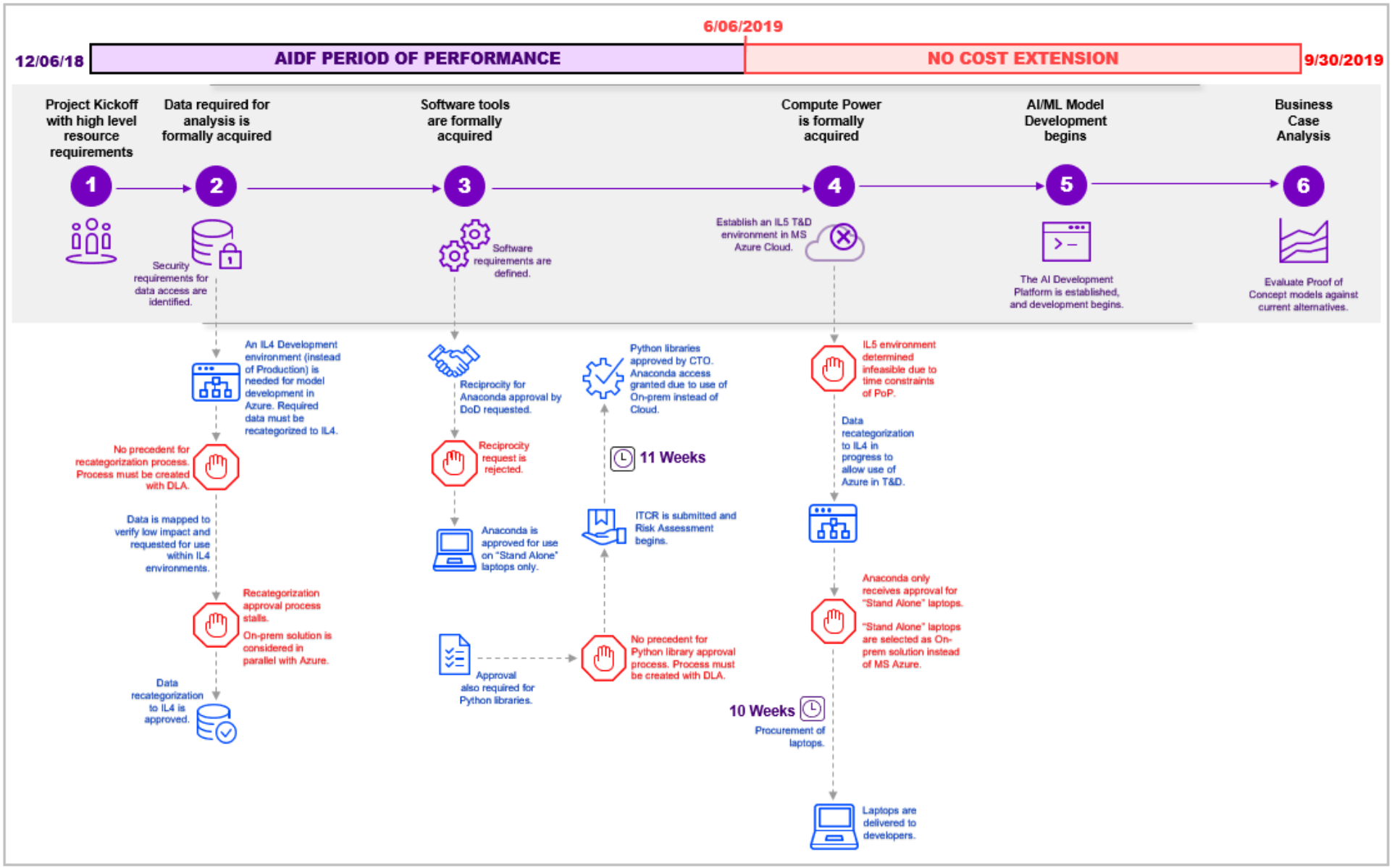

Figure 55: Overview of AIDF Project process obstacles

APPENDIX C: ITEM SUBSET SELECTION METHODOLOGY

This Appendix details the methodology for selecting the item subset that was used for all AI/ML modeling. The team used the six steps outlined below to determine an item subset that is both representative of the items with active demand and large enough to produce statistically significant results.

Methodology

Step 1 – Limit by Acquisition Advice Code & Active Demand

DLA's total item inventory contains approximately 5 million items; however, many of these items are Service/locally managed, restricted, terminal, or condemned, as specified by each item's Acquisition Advice Code (AAC). The project team analyzed DLA's item population by AAC and determined that only D, H, J, and Z items are most relevant to the development of AI/ML forecasting models (see Figure 56 for DLA Handbook definitions). Combined, they make up 80% DLA's item catalog. Thus, limiting the item subset to the four AACs below reduced the total count from over 6 million to approximately 5 million.

AAC	Definition	Justification
D	Issue, Transfer, or Shipment is not subject to specialized controls other than those impacted by the Integrated Materiel Manager/Service supply policy. **The item is centrally managed, stocked, and issued**.	Items within AAC D, H, J, and Z are all centrally managed, non-obsolete items which can benefit from DLA-centered demand forecasting models.
H	Direct Delivery Under a Central Contract. **The item is centrally managed and procured**. Normal issue is by direct shipment from vendor to the user.	
J	Not stocked. **Procurement initiated only after receipt of order**.	
Z	Insurance/Numeric Stockage Objective Item. Items that are generally not subject to periodic replacement or wear-out, but **subject to infrequent replacement** as the result of accidents and other unexpected occurrences.	

Figure 56: DLA definitions for item Acquisition Advice Codes included in the AIDF Proof-of-Concept.

The AIDF team further limited the subset within the four AACs above by focusing only on items with active demand, which is defined as demand within the final year of the training dataset (2016). As shown in Figure 57 and Figure 58, introducing this restriction reduced the total eligible item population from 5 million to less than 1 million.

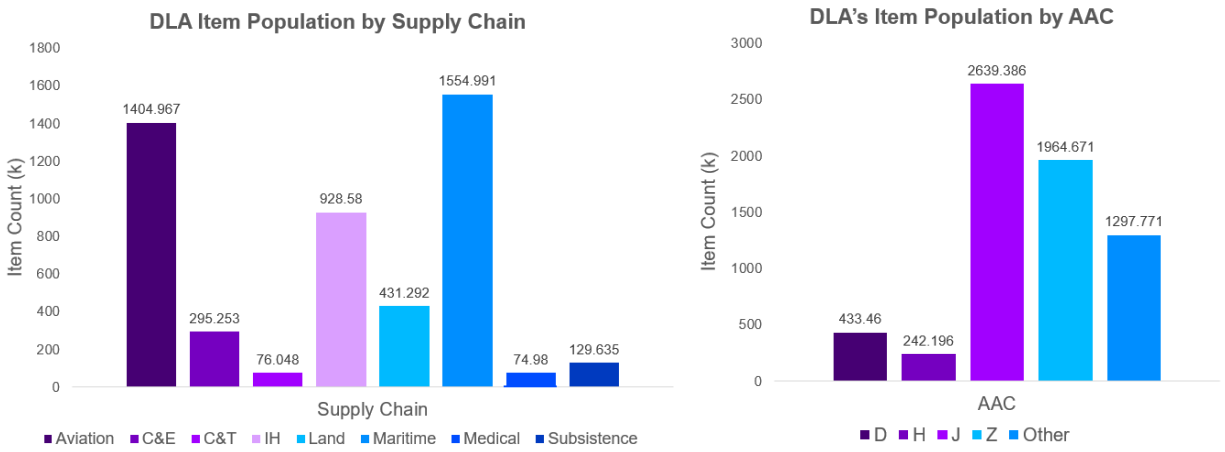

Figure 57: Overview of DLA's Total Item Population as of December 2017

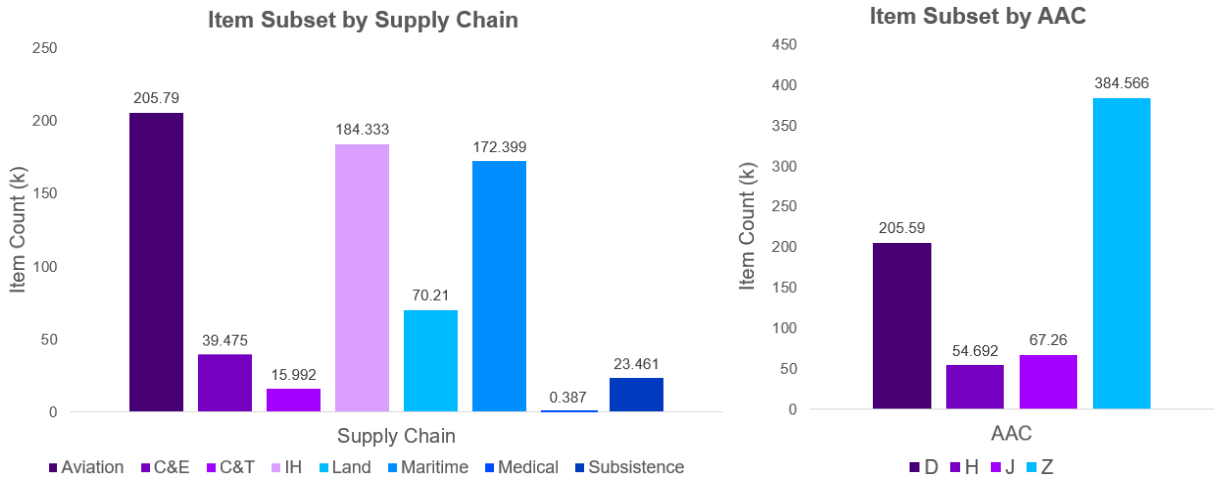

Figure 58: Item Population within Relevant AACs and Demand in the Training Years

Step 2 – Stratify by Material Type, then Supply Chain

The approximately 1 million items remaining after Step 1 were categorized into four primary materiel types: **Subsistence**, **Medical**, **Clothing & Textiles**, and **Hardware**. As shown in Figure 59, the item counts for the first three supply chains make up only 3% of the subset. In addition, the supply chains for these materials exhibit properties that make their items non-ideal for a Phase 1 forecasting proof-of-concept:

- The **Subsistence** and **Medical** supply chains are primarily outsourced: while DLA manages procurement, the planning and inventory management for these items is largely overseen by vendors.
- The **Clothing & Textiles** supply chain uses a forecasting hierarchy (PGC-level) that differs from the rest of the Agency.

The uniqueness of the supply chains above, combined with their low item counts, led the Accenture team to focus its modeling efforts on items within the **Hardware** supply chains (Figure 59). However, this focus did not create a substantial reduction in the overall item subset.

Supply Chain	Item Count (k)	Material Type
Aviation	205.8	Hardware
C&E	39.5	Hardware
C&T	15.9	Clothing & Textiles
IH	184.3	Hardware
Land	70.2	Hardware
Maritime	172.4	Hardware
Medical	0.4	Medical
Subsistence	23.5	Subsistence

Figure 59: Accenture focused its forecast modeling efforts to items within the Hardware supply chains.

Step 3 – Consider Computational Limitations

Refining the data to AAC D, H, J, and Z items with active demand within the Hardware supply chains left the total item subset count at approximately 672k. The team then considered the computational limitations imposed by the lack of an established AI Model Training Environment at DLA:

- **Data Extraction:** Seven years of historical usage and forecast data needed to be extracted from DLA systems[21] for each item, so the time the data took to download was directly proportional to the total number of items selected for the subset.
- **Data Transfer:** DLA required that the team conduct all modeling activities on air-gapped laptops (i.e., no internet connectivity). Therefore, all downloaded data had to be burned and then transferred (as directed by DLA) via DVD-Rs– an extremely time intensive process that also scales proportional to the size of the subset.
- **Compute Power and Data Storage:** Using laptops to develop all ML algorithms introduced hard limits on data storage and compute power. Details on model performance at the item level also needed to be stored, which significantly increased the data footprint required.

The limitations above meant that project progress would be hindered by selecting more items than necessary for a successful prototype. Noting that the smallest Hardware supply chain (C&E) had 39.5k items with active demand, the sample size per supply chain needed to be at or below this value to create equal item representation. Through experimentation with the extraction, transfer, and preliminary modeling processes, Accenture determined that 25k items per Hardware supply chain was an optimal sample size (i.e., 125k total items).

[21] Demand history and item attributes were extracted from DORRADW. Statistical forecast history was pulled from EDW. See section Forecasting Algorithms Modeling for details.

Step 4 – Leverage Disproportionate Stratified Random Sampling

The traditional approach to Stratified Random Sampling (SRS) retains proportionality across the strata (in this case, the five supply chains); however, this would create much lower representation for items within the C&E and Land supply chains. To mitigate the risk of biasing toward any one supply chain, the project team used disproportionate Stratified Random Sampling (dSRS) to extract the initial 125k item subset. As shown in Figure 60, dSRS randomly selects the same number of items from each supply chain to enforce equal representation.

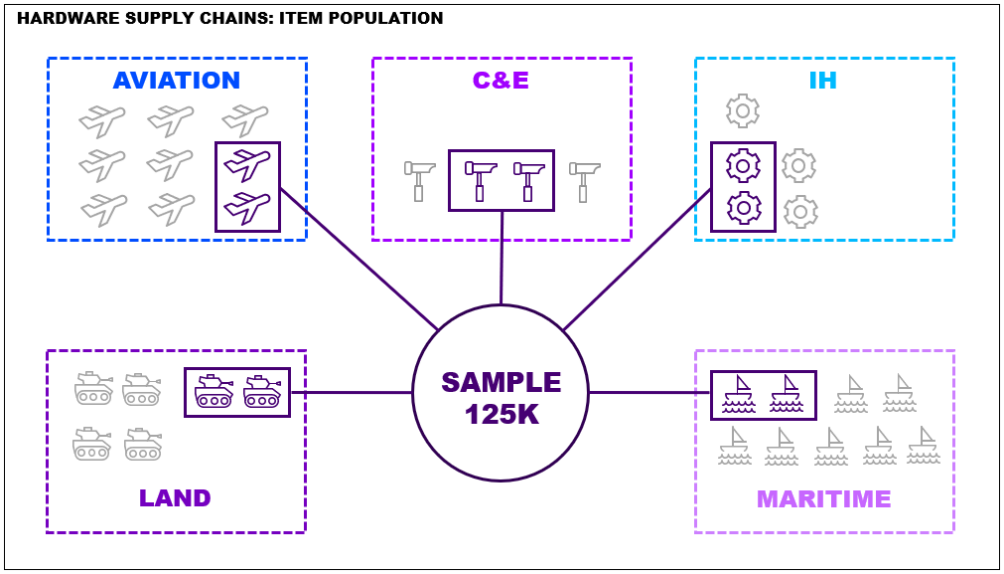

Figure 60: Depiction of disproportionate Stratified Random Sampling

Step 5 – Evaluate Operational Significance, Current Model Diversity, and Historical Forecast Bias

Operational Significance

After running dSRS, Accenture evaluated the operational significance of the resulting item subset. Specifically, the team analyzed the 125k items for overlap with DLA's Super Groups to ensure sufficient representation. The Super Groups include Aviation's "Crown Jewels," L&M's "Super Kids," Troops Support's (i.e., C&E, and IH's) "Silver Bullets," and other DFUs with extremely high-value forecasts.

During this validation process, Accenture determined that the 25K sample size per supply chain did not capture an adequate number of high annual demand frequency (ADF) items for analysis. Excluding these items would constrain the team's ability to predictively model high-ADF items and would consequently reduce the overall robustness of the proof-of-concept. Therefore, the item subset was augmented with the top 25k NIINs by ADF, independent of supply chain, as shown in Figure 61 by request from J34. This data augmentation improved Accenture's ability to model high-ADF items, thereby maximizing the efficacy of the proof-of-concept.

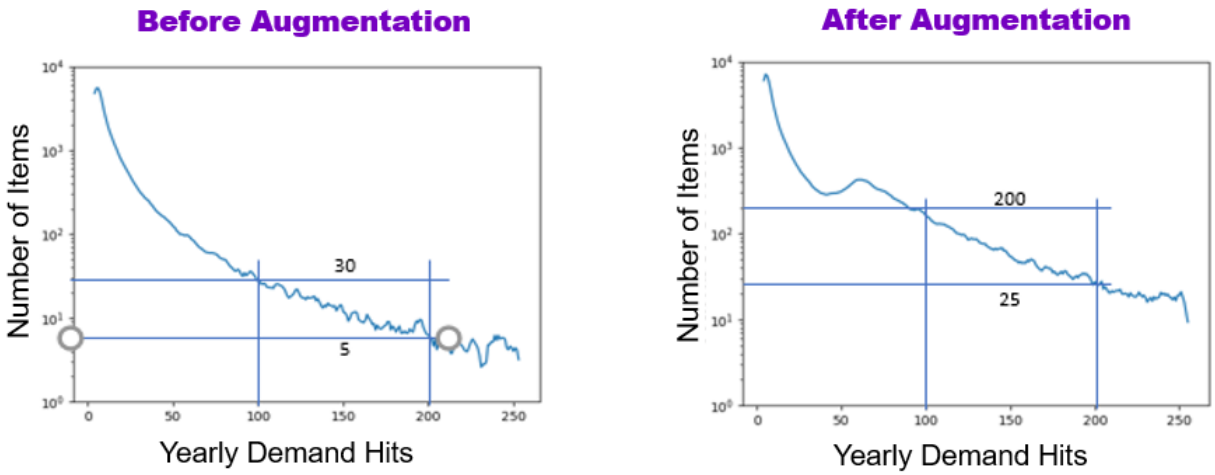

Figure 61: Graphs of Before and After the Item Subset was Augmented to Include High-ADF Items

Supply Chain	Eligible Items	dSRS	dSRS + Top ADF
Aviation	201.3	25	31.6
C&E	40.7	25	25.7
IH	191.1	25	31.7
Land	71.4	25	27.4
Maritime	172.4	25	28.6

Figure 62: Item Subset Augmented for Top ADF. Note that total augmented item count is only 145k because of overlap between the dSRS and Top ADF item populations.

Current Model Diversity

The item subset was also evaluated for diversity across existing JDA statistical forecasting models in order to verify adequate representation of different demand classes (e.g., continuous non-seasonal, lumpy seasonal, etc.). Accenture captured the current statistical forecast models for all DFUs corresponding to the items in the augmented subset. As seen in Figure 63, the subset contains DFUs that span all six of the existing models. Because Lewandowski is a proprietary forecasting algorithm similar to Holt-Winters, the seeming underrepresentation of Holt-Winters can be ignored. The low count of Moving Average DFUs is also mitigated by the project team's inclusion of the Moving Average model type in its research.

Model Type	Number of DFUs (k)
AVS-Graves	48.5
Croston's	16.0
Fourier	42.0
Holt-Winters	2.9
Lewandowski	35.6
Moving Average	2.0

Figure 63: Distribution of item subset DFUs across existing statistical forecast models.

Historical Forecast Bias

In addition to operational significance and current model diversity, the team evaluated the item subset for historical forecast bias. The data should be inclusive of both accurate and inaccurate historical forecasts. Therefore, the subset should include sufficient representation of items that were under, over, and accurately forecasted. As shown in Figure 64, Accenture determined that the items in the augmented subset exhibited an adequate forecast accuracy distribution.

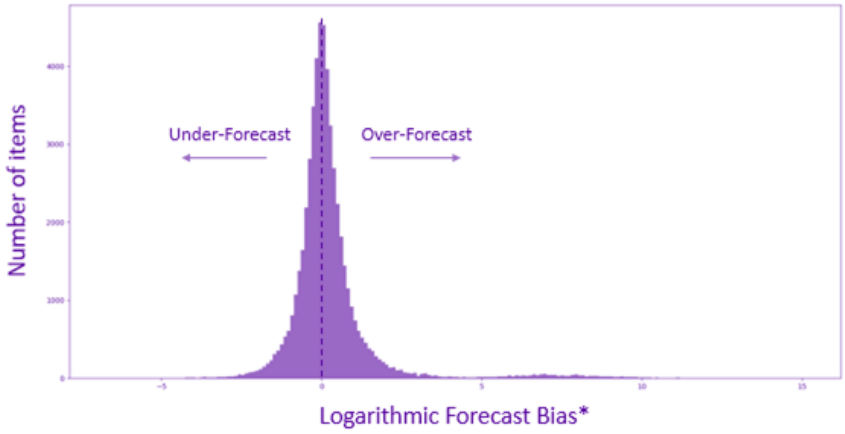

Figure 64: Distribution of historical forecast accuracy within the item subset.

Step 6 – Calculate the Statistical Power to Verify the Item Subset's Size

Statistical power, which ranges between 0 and 1, can be used to determine whether a sample is large enough to produce statistically significant results. The larger the sample, the higher the power. Thus, the closer the power is to 1, the greater the likelihood that conclusions drawn on the item subset will be statistically meaningful.

As a final step, the AIDF team calculated the power of the item subset to ensure that the sample size was large enough to draw statistically significant conclusions. The subset was segmented into bins based on days with demand over the evaluation period 5-14, 15-49, 50-119, and 120+. A separate power calculation was completed for each range because initial modeling research revealed that the best-performing algorithms varied by demand frequency (Figure 65). The power of the item subset is 1 (indicating 100% ability to define differences when differences exist to be found) for the 5-14 and 15-49 ranges, 0.90 for the 50-119 range, and 0.20 for the 120+ range.

This result concludes that a statistically significant difference would be achieved when there is one to be found 90% to 100% of the time between the frequency range of 5-119 and 20% of the time in the 120+ range. This includes a data augmentation in the high-demand frequency range of items to improve statistical power. As an additional mitigation to this lower power in the high-frequency range of items, it is likely that more complex algorithms will increase the effect size in high-frequency range items over the very conservative estimation created by the simple average approach. Additionally, if needed, a more precise pairwise statistical test can be performed to increase the effect size which would likely limit the population standard deviation.

These results are generated with the alpha value set to 0.01, which sets an allowable 1% chance the results are due to chance rather than the test performed. This statistical power calculation is a good indication that the item subset is adequate in size and will likely result in significant findings if there is indeed significant improvement from the forecasting approaches implemented.

The following formulas were used to calculate the statistical power.

Standard Error

$$SE = \frac{\sigma}{\sqrt{Sample\ size}}$$

Critical Value

$$Z_{critical} = \frac{\bar{X}_{critical} - \mu_0}{\sigma}$$

Inverse the normal cumulative distribution for the hypothetic mean ($\bar{X}_{critical}$) and the standard deviation to find the critical value.

Effect Size

$$Effect\ Size = \frac{(JDA\ Average\ sMAPE - Simple\ Average\ Average\ sMAPE)}{\sigma}$$

Take the difference between the two groups (JDA and Simple Average) and divide by the standard deviation of the Simple Average group.

Actual Mean

$$\bar{X} = (\bar{X}_{critical} + Effect\ Size)\sigma$$

Beta

$$\beta = P(\bar{X} \geq \bar{X}_{critical} | \mu)$$

Calculate the cumulative distribution function for the critical value of the actual mean and standard deviation to find beta (the probability of making a Type II error).

Power

$$Power = 1 - \beta$$

RANGE 1: 5-14	
Algorithm	Simple Average
Hypothetic Mean	0
Standard Deviation	0.363
Alpha	0.010
Sample Size	12589
Standard Error	0.003
Critical Value	0.008
Effect Size	0.173
Actual Mean	0.063
Beta	5E-66
Power	**1.000**

RANGE 2: 15-49	
Algorithm	Simple Average
Hypothetic Mean	0
Standard Deviation	0.303
Alpha	0.010
Sample Size	14551
Standard Error	0.003
Critical Value	0.006
Effect Size	0.125
Actual Mean	0.038
Beta	8E-38
Power	**1.000**

RANGE 3: 50-119	
Algorithm	Simple Average
Hypothetic Mean	0
Standard Deviation	0.235
Alpha	0.010
Sample Size	14786
Standard Error	0.002
Critical Value	0.004
Effect Size	0.030
Actual Mean	0.007
Beta	0.098
Power	**0.902**

RANGE 4: 120+	
Algorithm	Simple Average
Hypothetic Mean	0
Standard Deviation	0.203
Alpha	0.010
Sample Size	5785
Standard Error	0.003
Critical Value	0.006
Effect Size	0.020
Actual Mean	0.004
Beta	0.796
Power	**0.204**

Figure 65: Power Calculation of Item Subset Per Identified Range

Summary

IDENTIFY & STRATIFY			EXTRACT	EVALUATE	VERIFY
1	2	3	4	5	6
Determine data requirements and limit by Acquisition Advice Code & Active Demand.	Stratify by Materiel Type, then Supply Chain.	Consider Computational Limitations	Leverage Disproportionate Stratified Random Sampling.	Evaluate Operational Significance, Current Model Diversity, and Historical Forecast Bias.	Calculate the statistical power of the sample to verify the size of the item subset.

Figure 66: Overview of the Item Selection Process

APPENDIX D: MODEL RESULTS

Model Performance In-Depth Analysis

Statistical Testing Methodology

To evaluate the performance of the ensemble technique against the existing JDA solution, the results were compared pairwise by items to directly assess the statistical significance of the population performance divergences. To accomplish this, items in both the item subset and the existing JDA solution were evaluated for performance on sMAPE, DPA, and dollarized error metrics to determine comparative performance.

To achieve this, prediction quantities for the Ensemble method, JDA output, and actual demand were aggregated into 3-month buckets over the Test dataset year (2018). The results of the Ensemble method and the JDA outputs were then compared with the actual quarterly demands. These comparisons were then analyzed using sMAPE, DPA, and dollarized errors and were done by subtracting the Ensemble score from the JDA score.

The result of this comparison yields JDA and Ensemble scores for each metric which can then be compared pairwise, item-by-item. This enables an explicit evaluation of performance improvement by taking the same items under the same circumstances and evaluating outcomes based on separate forecasting techniques.

Two separate methodologies were employed to evaluate the comparative performance of the two forecasting techniques (JDA and Ensembling). First, a parametric approach which determines the population mean and confidence intervals was used. This approach requires the assumption of normality and does not fit non-normal data well. Secondly, a technique known as bootstrapping was used wherein the population of items and their resultant scores are resampled with replacement to generate an empirical evaluation of performance.

The bootstrapping methodology is shown in Figure 67 where the histogram of a skewed population of score differences between the JDA and Ensemble models is resampled with replacement resulting in an un-skewed, normal range of population means from which a confidence interval of the true mean can be derived without any assumption about the original distribution’s normality.

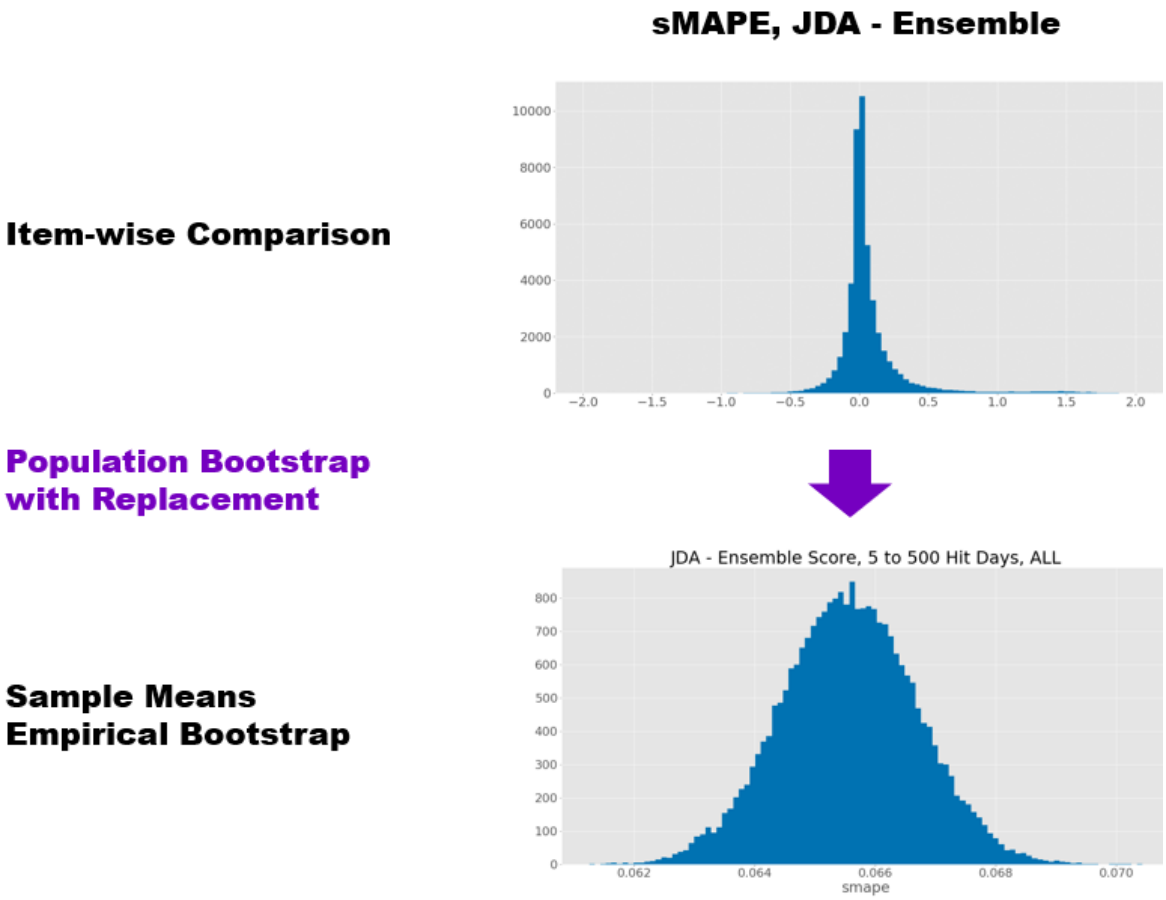

Figure 67: Example of Bootstrapping Methodology to Empirically Approximate Population Means

Ensemble Model Results – Breakouts

In-depth, subgroup analysis of item performance was performed using the two previously detailed statistical tests. The definitions used within these breakouts are outlined below:

- **hits_range**: The range of days with demand for items evaluated
- **med_sim**: The median of the simulated bootstrap sample means
- **5th_pct_sim**: The empirical 5th percentile of bootstrapped sample mean simulations
- **95th_pct_sim**: The empirical 95th percentile of bootstrapped sample mean simulations
- **para_mean**: The parametric mean of the item population
- **para_5th**: The parametric 5th percentile, calculated on the original pairwise comparison
- **para_95th**: The parametric 95th percentile, calculated on the original pairwise comparison
- **para_stderr**: The standard error of the original pairwise comparison
- **num_obs**: The number of items within the range specified

For all comparative evaluations performed, the Ensemble score was subtracted from the baseline JDA score. The result of this is that for error metrics such as **sMAPE and Dollar Error, a positive**

value represents an improvement, whereas for accuracy scores such as **DPA, a negative score represents an improvement**.

sMAPE and DPA scores for items within both the item subset and JDA having more than five days with demand are show in Figure 68 and Figure 69. These tables show meaningful improvement in each of the demand frequency regions assessed.

hits_range	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
(5, 15)	9.12%	8.64%	9.60%	9.11%	8.63%	9.60%	0.29%	12941
(15, 50)	7.76%	7.43%	8.09%	7.75%	7.43%	8.08%	0.20%	15055
(50, 120)	4.37%	4.12%	4.61%	4.37%	4.12%	4.61%	0.15%	14444
(120, 500)	3.16%	2.85%	3.48%	3.16%	2.84%	3.47%	0.19%	5740
(5, 500)	6.56%	6.37%	6.74%	6.56%	6.37%	6.74%	0.11%	48180

Figure 68: sMAPE for Items within the Item Subset compared with JDA

hits_range	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
(5, 15)	-2.45%	-2.69%	-2.21%	-2.45%	-2.69%	-2.20%	0.15%	12941
(15, 50)	-3.67%	-3.85%	-3.48%	-3.67%	-3.85%	-3.48%	0.11%	15055
(50, 120)	-4.11%	-4.29%	-3.94%	-4.11%	-4.29%	-3.93%	0.11%	14444
(120, 500)	-3.49%	-3.76%	-3.22%	-3.49%	-3.75%	-3.22%	0.16%	5740
(5, 500)	-3.45%	-3.56%	-3.34%	-3.45%	-3.56%	-3.34%	0.06%	48180

Figure 69: DPA for Items within the Item Subset compared with JDA

When the error is dollarized as shown in Figure 70, the results show that for most regions of item demand frequency, performance is improved. This applies to the overall range of items having 5+ days with demand, showing an average of $545/item error avoided across 48k items. The impact of this is a net avoided forecast error of ~$26M for items within the subset for each quarter. Aggregated yearly, this amounts to ~$105M. Of this, the extremely high demand items with 120+ days with demand account for ~$9M quarterly, or ~$36M yearly despite being the smallest population of items. This suggests the Ensemble model provides the greatest benefit to DLA's high demand items.

hits_range	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
(5, 15)	-$98.06	-$174.15	-$31.28	-$99.51	-$171.33	-$27.68	$43.67	12941
(15, 50)	$557.69	$434.27	$682.39	$557.79	$434.03	$681.54	$75.23	15055
(50, 120)	$689.15	$546.69	$876.81	$696.80	$531.41	$862.20	$100.54	14444
(120, 500)	$1,579.43	$1,046.01	$2,235.17	$1,603.70	$1,002.43	$2,204.96	$365.51	5740
(5, 500)	$545.27	$453.48	$648.07	$547.52	$450.20	$644.84	$59.16	48180

Figure 70: Absolute Dollar Error for Items within the Item Subset compared with JDA

Looking deeper into the breakout of the absolute dollar error, the forecast difference is split into over and under-forecasts, as shown in Figure 71 and Figure 72. The primary source of improvement appears to come from limiting the impacts of over-forecasts ($529/item) while making negligible improvements to the under-forecast ($17.56/item). The total population over-forecast avoided by moving from the JDA to Ensemble model for the ~48k items then sums to ~$25M each quarter, ~$102M yearly.

hits_range	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
(5, 15)	-$417.73	-$501.33	-$344.41	-$419.27	-$497.99	-$340.56	$47.85	12941
(15, 50)	$390.79	$246.29	$557.06	$394.64	$239.95	$549.32	$94.03	15055
(50, 120)	$979.81	$816.60	$1,173.65	$986.00	$807.01	$1,164.98	$108.81	14444
(120, 500)	$1,857.08	$1,308.15	$2,538.60	$1,883.21	$1,263.74	$2,502.68	$376.58	5740
(5, 500)	$529.18	$428.63	$640.65	$530.65	$425.11	$636.19	$64.16	48180

Figure 71: Over-Forecast Dollar Error for Items within the Item Subset compared with JDA

hits_range	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
(5, 15)	$318.81	$278.70	$365.31	$319.77	$276.20	$363.33	$26.49	12941
(15, 50)	$173.09	-$13.55	$312.70	$163.15	-$2.31	$328.60	$100.58	15055
(50, 120)	-$291.47	-$410.06	-$161.07	-$289.19	-$413.91	-$164.48	$75.81	14444
(120, 500)	-$280.36	-$612.02	$50.27	-$279.51	-$607.09	$48.06	$199.14	5740
(5, 500)	$17.56	-$60.33	$91.49	$16.87	-$58.86	$92.60	$46.04	48180

Figure 72: Under-Forecast Dollar Error for Items within the Item Subset compared with JDA

Proportional Supply Chain Breakouts

To compare the performance of the Ensemble method against JDA for the measurable 5+ days with demand items in a manner that represented an un-augmented and proportional sample within each supply chain, the stratified random sample of items prior to augmentation with high-ADF items was used. Far fewer items were evaluated in this comparison due to the fact that the data augmentation substantially increased the item subset's overlap with JDA's forecastable item list.

This technique creates a view by supply chain of the proportional expected improvement that most directly mimics scaling the ensemble technique to the entire supply chain. Figure 73 details the sMAPE improvements across supply chains using this proportional item sample while Figure 74 details the DPA comparisons. In both of these measures, there are significant performance improvements, with Land showing the greatest overall accuracy gain.

sup_chain	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
MRTM	9.01%	8.31%	9.73%	9.01%	8.30%	9.73%	0.43%	4781
LAND	10.39%	9.66%	11.16%	10.40%	9.65%	11.15%	0.46%	5125
IH	8.99%	8.36%	9.63%	9.00%	8.36%	9.63%	0.38%	6262
AVTN	7.40%	6.89%	7.93%	7.41%	6.89%	7.92%	0.31%	6976
C&E	6.40%	5.87%	6.93%	6.40%	5.87%	6.93%	0.32%	5971

Figure 73: sMAPE Comparison for a Representative Item Sample across Supply Chains

sup_chain	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
MRTM	-3.67%	-4.04%	-3.32%	-3.68%	-4.03%	-3.32%	0.21%	4781
LAND	-4.72%	-5.09%	-4.34%	-4.72%	-5.09%	-4.34%	0.23%	5125
IH	-3.48%	-3.80%	-3.16%	-3.48%	-3.80%	-3.16%	0.20%	6262
AVTN	-2.76%	-3.05%	-2.47%	-2.76%	-3.04%	-2.47%	0.17%	6976
C&E	-2.90%	-3.21%	-2.61%	-2.91%	-3.21%	-2.60%	0.18%	5971

Figure 74: DPA Comparison for a Representative Item Sample across Supply Chains

Absolute Dollar Error, Over-Forecast Dollar Error, and Under-Forecast Dollar Error are shown for the representative supply chain samples in Figure 75, Figure 76, and Figure 77, respectively.

Again, the Land supply chain shows the highest average error reduction using the Absolute Dollar Error metric.

sup_chain	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
MRTM	$303.80	$191.35	$463.99	$313.14	$176.09	$450.18	$83.31	4781
LAND	$622.95	$394.43	$848.03	$621.47	$392.17	$850.77	$139.39	5125
IH	$85.74	$47.75	$149.31	$89.72	$37.71	$141.73	$31.62	6262
AVTN	$549.43	$299.84	$916.42	$572.19	$256.88	$887.49	$191.67	6976
C&E	$512.06	$247.78	$876.08	$527.29	$214.89	$839.70	$189.91	5971

Figure 75: Absolute Dollar Error Comparison for a Representative Item Sample across Supply Chains

sup_chain	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
MRTM	$139.48	$3.81	$349.88	$153.62	-$23.43	$330.67	$107.63	4781
LAND	$364.80	$117.60	$600.13	$361.55	$119.91	$603.19	$146.89	5125
IH	$49.83	$27.67	$77.41	$50.69	$25.85	$75.53	$15.10	6262
AVTN	$22.56	-$233.80	$417.63	$49.09	-$284.27	$382.45	$202.65	6976
C&E	$388.05	$116.41	$756.29	$405.09	$82.93	$727.25	$195.84	5971

Figure 76: Over-Forecast Dollar Error Comparison for a Representative Item Sample across Supply Chains

sup_chain	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
MRTM	$160.94	$78.20	$237.72	$159.52	$79.52	$239.52	$48.63	4781
LAND	$257.23	$115.26	$417.14	$259.92	$109.83	$410.01	$91.24	5125
IH	$35.92	-$1.57	$99.91	$39.04	-$13.19	$91.26	$31.75	6262
AVTN	$515.18	$345.76	$728.19	$523.09	$331.01	$715.17	$116.76	6976
C&E	$123.05	-$47.50	$294.93	$122.20	-$48.41	$292.82	$103.72	5971

Figure 77: Under-Forecast Dollar Error Comparison for a Representative Item Sample across Supply Chains

Ensemble Item Performance Comparisons

To understand where the Ensemble method was improving performance, individual item analysis was done on the extreme errors avoided. One example shown in Figure 78 demonstrates JDA's extreme overreaction to certain patterns which can lead to costly errors in the forecast.

Two benefits of the method are shown here. The first is the item-centric model selection that allows the Ensemble to leave out those models which are unlikely to fit the data. The second is the effect of the Ensemble model dampening the input of outlying models by mediating their results with the results with the consensus of the other models.

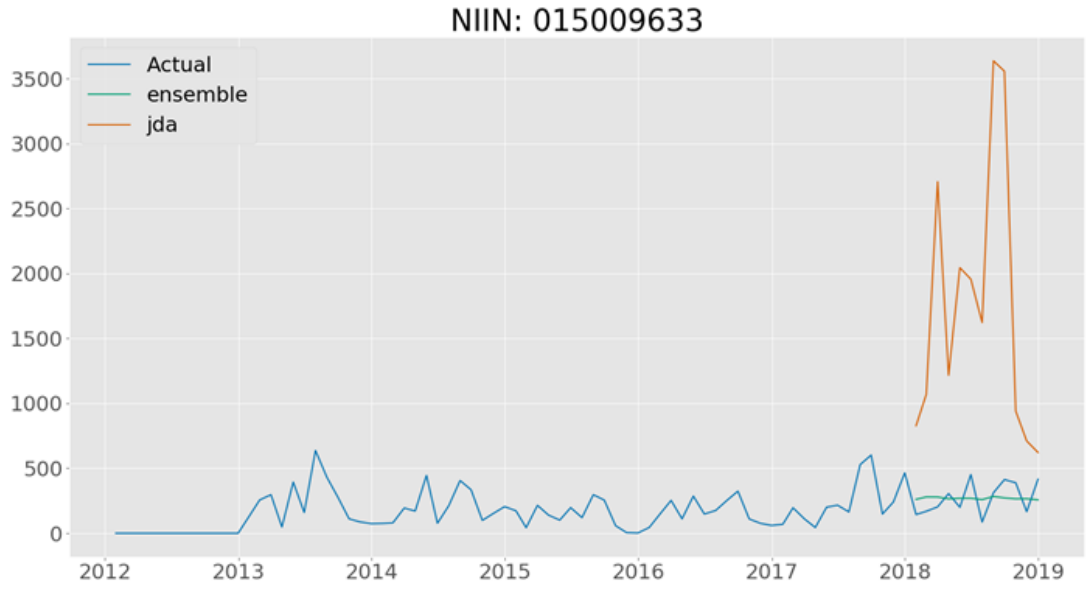

Figure 78: Illustrative Example of a Large Error Avoided by the Ensemble Method

Figure 79 similarly shows an example of a JDA algorithm selecting drastically under-forecasting an item's future demand based on a pattern that did not materialize.

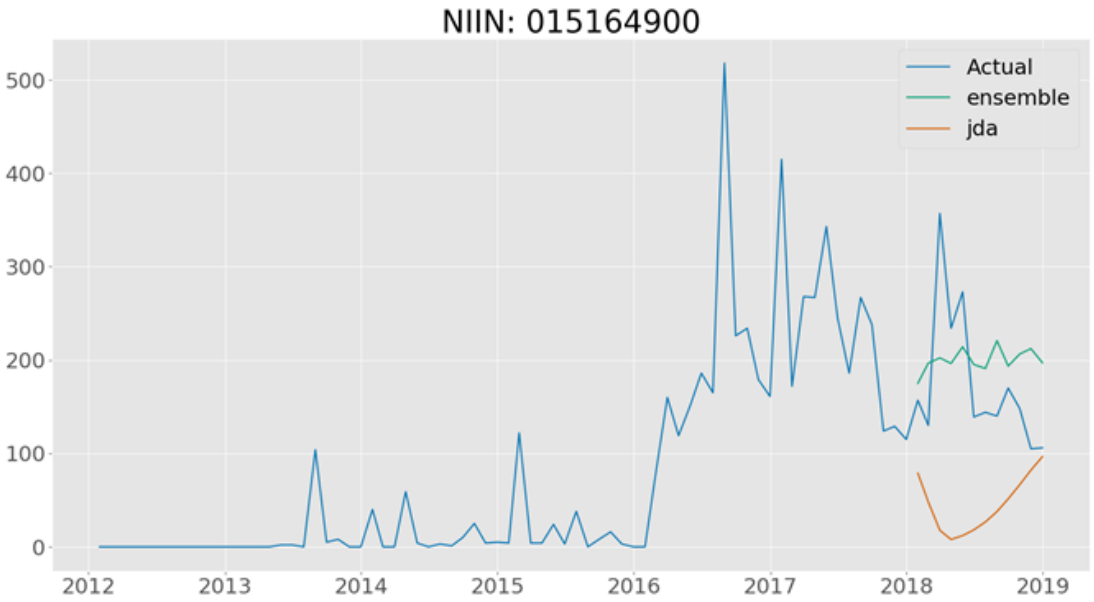

Figure 79: Illustrative Example of a Large Error Under-Forecasting Error Avoided by the Ensemble Method

JDA's algorithms, by showing some tendencies to overfit historical data into trends, do also sometimes correctly identify hard-to-predict demand patterns when the Ensemble technique does not, as shown in Figure 80. However, this approach can come at the cost of extreme errors which generate more error than are ultimately reduced when aggregated, leading to the results of comparatively high errors shown in Figure 78.

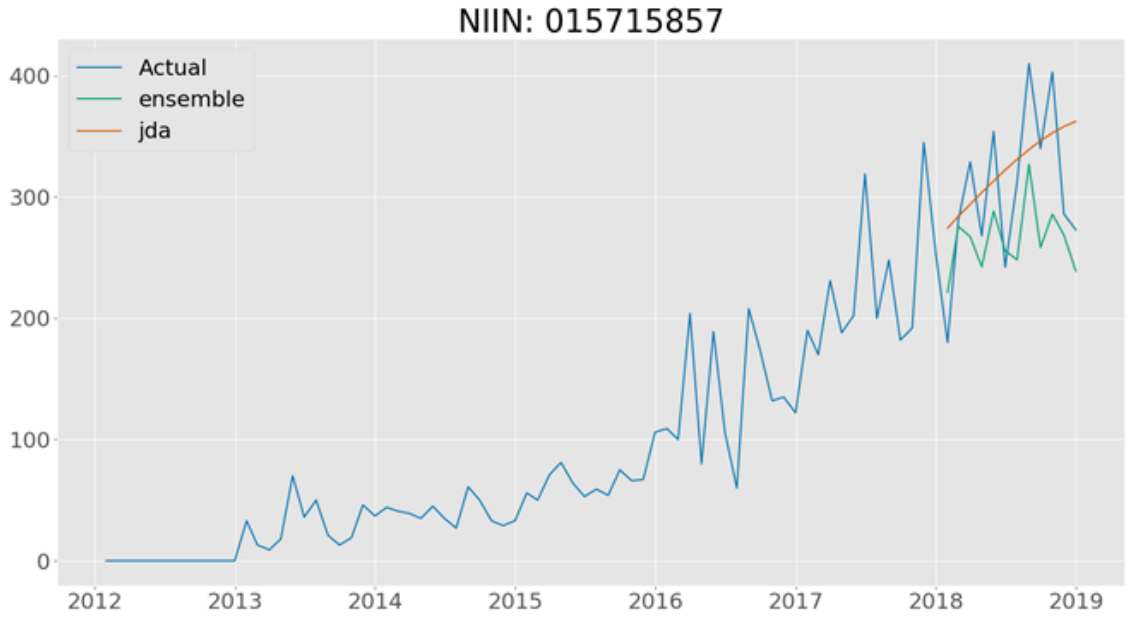

Figure 80: Illustrative Example of JDA's Algorithm Aggressively Selecting the Correct Trend

Course of Action Model Statistical Results

For completeness and statistical verification of data presented in the Analysis of Courses of Action (COAS) section, detailed breakouts of value by model type and metric over the 0-4 and 5+ regions of days with demand have been included here. All comparisons are performed using a pairwise comparison against the JDA baseline wherein a higher number represents a higher metrics score by JDA compared against the alternative.

Note that given the severe metrics issues associated with items having fewer than 5 days with demand, the **results shown for this range of items do not have clear, consistent, or business-relevant interpretations and the AIDF team's recommendation for this group of items is to adopt management strategies outside of point forecasting altogether.** For completeness, these metrics are included in Figure 81, Figure 83, Figure 85, and Figure 87.

For the items having 5 or more days with demand, the analyses included below in Figure 82, Figure 84, Figure 86, and Figure 88 provide a representation of the statistical confidence intervals for each metric over the item subset.

EWMA

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-2.67%	-2.88%	-2.47%	-2.67%	-2.88%	-2.47%	0.13%	9206
sMAPE	-19.48%	-20.41%	-18.56%	-19.49%	-20.41%	-18.56%	0.56%	9206
Over-Forecast $	$223.02	-$98.42	$793.71	$254.78	-$208.21	$717.77	$281.45	9206
Under-Forecast $	$53.28	$34.93	$71.35	$53.24	$35.10	$71.38	$11.03	9206
Absolute $ Err	$281.04	-$37.38	$857.61	$308.02	-$153.68	$769.73	$280.67	9206

Figure 81: EWMA Scores for Items with 0-4 Days with Demand

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-3.63%	-3.74%	-3.53%	-3.63%	-3.74%	-3.53%	0.06%	48180
sMAPE	5.55%	5.37%	5.73%	5.55%	5.37%	5.73%	0.11%	48180
Over-Forecast $	$841.70	$737.75	$956.69	$842.82	$733.66	$951.98	$66.36	48180
Under-Forecast $	-$498.54	-$589.17	-$417.30	-$500.18	-$586.58	-$413.77	$52.53	48180
Absolute $ Err	$341.09	$247.57	$444.05	$342.64	$244.08	$441.21	$59.92	48180

Figure 82: EWMA Scores for Items with 5+ Days with Demand

LSTM

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	0.46%	0.23%	0.70%	0.46%	0.22%	0.70%	0.15%	9206
sMAPE	-22.57%	-23.51%	-21.64%	-22.57%	-23.50%	-21.64%	0.56%	9206
Over-Forecast $	-$1,772.81	-$2,414.16	-$1,392.34	-$1,815.43	-$2,339.69	-$1,291.18	$318.70	9206
Under-Forecast $	$196.79	$172.08	$224.00	$197.45	$171.15	$223.76	$15.99	9206
Absolute $ Err	-$1,578.03	-$2,209.04	-$1,204.17	-$1,617.98	-$2,137.71	-$1,098.25	$315.95	9206

Figure 83: LSTM Scores for Items with 0-4 Days with Demand

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-3.15%	-3.27%	-3.03%	-3.15%	-3.27%	-3.03%	0.07%	48180
sMAPE	5.81%	5.61%	6.01%	5.81%	5.61%	6.01%	0.12%	48180
Over-Forecast $	$518.94	$392.98	$655.48	$520.45	$388.93	$651.96	$79.95	48180
Under-Forecast $	-$194.30	-$315.44	-$92.69	-$197.74	-$309.37	-$86.11	$67.86	48180
Absolute $ Err	$321.59	$208.41	$439.65	$322.71	$206.40	$439.01	$70.70	48180

Figure 84: LSTM Scores for Items with 5+ Days with Demand

Ensemble

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-0.76%	-0.99%	-0.53%	-0.76%	-0.99%	-0.54%	0.14%	9206
sMAPE	-20.30%	-21.23%	-19.38%	-20.30%	-21.22%	-19.37%	0.56%	9206
Over-Forecast $	-$812.43	-$986.93	-$585.67	-$802.63	-$1,004.49	-$600.76	$122.71	9206
Under-Forecast $	$165.69	$145.13	$187.64	$165.95	$144.80	$187.10	$12.86	9206
Absolute $ Err	-$647.39	-$813.55	-$422.97	-$636.68	-$834.09	-$439.26	$120.01	9206

Figure 85: Ensemble Scores for Items with 0-4 Days with Demand

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-3.45%	-3.56%	-3.34%	-3.45%	-3.56%	-3.34%	0.06%	48180
sMAPE	6.56%	6.37%	6.74%	6.56%	6.37%	6.74%	0.11%	48180
Over-Forecast $	$528.52	$428.37	$638.46	$530.65	$425.11	$636.19	$64.16	48180
Under-Forecast $	$18.26	-$60.89	$91.47	$16.87	-$58.86	$92.60	$46.04	48180
Absolute $ Err	$545.51	$455.32	$647.05	$547.52	$450.20	$644.84	$59.16	48180

Figure 86: Ensemble Scores for Items with 5+ Days with Demand

Basic Ensemble

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-2.39%	-2.60%	-2.19%	-2.40%	-2.60%	-2.19%	0.13%	9206
sMAPE	-19.25%	-20.16%	-18.33%	-19.25%	-20.17%	-18.32%	0.56%	9206
Over-Forecast $	-$303.83	-$441.74	-$168.74	-$305.07	-$441.09	-$169.06	$82.68	9206
Under-Forecast $	$78.26	$59.38	$97.04	$78.22	$59.39	$97.06	$11.45	9206
Absolute $ Err	-$225.28	-$358.88	-$97.18	-$226.85	-$358.13	-$95.56	$79.81	9206

Figure 87: Basic Ensemble Scores for Items with 0-4 Days with Demand

score	med_sim	5th_pct_sim	95th_pct_sim	para_mean	para_5th	para_95th	para_stderr	num_obs
DPA	-3.45%	-3.55%	-3.34%	-3.45%	-3.55%	-3.34%	0.06%	48180
sMAPE	5.85%	5.67%	6.03%	5.85%	5.67%	6.03%	0.11%	48180
Over-Forecast $	$635.94	$533.03	$748.31	$637.41	$530.23	$744.58	$65.15	48180
Under-Forecast $	-$198.47	-$285.68	-$118.83	-$199.56	-$283.48	-$115.64	$51.01	48180
Absolute $ Err	$436.46	$341.11	$541.38	$437.85	$338.34	$537.35	$60.49	48180

Figure 88: Basic Ensemble Scores for Items with 5+ Days with Demand

APPENDIX E: MODEL DETAILS

Time Series Model Overview

Many model architectures were evaluated over the course of the AIDF project. The primary models and their hyperparameters are discussed below.

Croston's Method

Croston's method for demand forecasting has been heavily used to forecast intermittent demand items due to its ability to blend the probability of incoming demand intervals with the probability of demand quantities.[22] In evaluating this method, the alpha parameter, which controls the smoothing of both demand interval probability and demand quantities, was iterated and minimized across demand frequency regions.

Teunter-Syntetos-Babai Variant of Croston's Method

Since the original introduction of Croston's method in 1972, multiple modifications have been proposed. One particular improvement has been to separate the smoothing of demand probability within an interval and the demand quantity.[23] The Teunter-Syntetos-Babai (TSB) variant of Croston's method takes advantage of this and was included in the analysis process.

Hyperparameter optimization for this methodology was performed by a grid search over the alpha and beta hyperparameters, controlling the demand probability and demand quantity smoothing factors. Next, the top performing models were assessed across different lengths of demand history, showing intuitively that more demand (the 5-year period) promoted better forecasts. Monthly demand inputs and a 5-year window were shown to have the best performance and were used as the base selections as model inputs.

Once the region of best performance was refined, Winsorization preprocessing was added to the limited grid search of top performing alpha and beta hyperparameter ranges. Notably high Winsorization options were able to improve both DPA and sMAPE but at a large cost to under-forecasting error. An area of further analysis could be to find a balance between using Winsorization and avoiding forecast bias for sparse-demand region items.

Simple Average

Average forecasts were included in the AIDF project's analysis in order to form a baseline of comparison and provide a forecast to the ensembling technique for items unlikely to benefit from complex models. The averages were tuned by grid searching the amount of history to include and Winsorization levels.

The special case of the 1-month simple average is also the naïve forecast. This forecast was rolled into the simple average results in all analyses.

---

22 https://www.researchgate.net/publication/254044245_A_Review_of_Croston's_method_for_intermittent_demand_forecasting

23 https://kourentzes.com/forecasting/wp-content/uploads/2014/06/Kourentzes-2014-Intermittent-Optimisation.pdf

Exponentially Weighted Moving Average

The exponentially weighted moving average approach was included as a simple approach that effectively smooths demand history. This approach's hyperparameters were optimized via grid search on its smoothing parameter, alpha, and Winsorization levels.

Autoregressive Integrated Moving Average

The Autoregressive Integrated Moving Average (ARIMA) model combines Autoregressive and Moving Average approaches with a "differencing" component to create statistical stationarity within the time-series. This approach has been used in multiple time-series forecasting applications and is able to fit complicated patterns of demand.[24] The ARIMA algorithm was optimized over its p, d, and q hyperparameters, indicating the order of the autoregressive model, the degree of differencing, and the order of the moving average model, respectively, through a grid search. This model was not deeply evaluated due to very poor performance in the initial grid search leading to negative and unpredictable errors.

Exponential Smoothing

Exponential smoothing is a common technique for forecasting that at its simplest single form provides a means to weight historical data with recency bias and at its most complex formulation can account for trends and seasonality in the form of triple exponential smoothing.[25] The hyperparameters evaluated the inclusion of trend, dampening, seasonality, box-cox estimation, and automatic bias removal. Like ARIMA, this algorithm produced poorly performing results and was not heavily evaluated after initial grid searches.

Neural Networks

Neural networks for time series have become more prevalent due to increased academic interest in the advancement of recurrent neural networks since 2014 with the Gated Recurrent Unit (GRU), a network initially used in speech recognition.[26] Multiple neural networks model architectures were evaluated within the AIDF project, including Temporal Convolutional Network (TCN), Convolutional Neural Network to Long Short-Term Memory (CNN-LSTM) Network, Long Short-Term Memory (LSTM) Network, and Feedforward Neural Networks. These networks were evaluated by modifying loss functions, architectures, regularizers, normalization, auto-regression, learning rate, and resampling rate. These were optimized via a genetic algorithm and manual modifications.

[24] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889
[25] https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm
[26] https://arxiv.org/pdf/1406.1078.pdf

APPENDIX F: METRICS CALCULATIONS

The following are the error metrics used in the metrics assessment to determine viable metrics options. Note that $\boldsymbol{\epsilon}$ refers to a very small number meant to avoid divide-by-zero type errors. Additionally, MAPE, MASE, Relative Percent Error, and Log Error were limited to a maximum individual score of +/- 5 to avoid the influence on single outliers drastically shifting the net score.

For all metrics calculations: $\boldsymbol{y}$ refers to the actual demand value and $\boldsymbol{\hat{y}}$ refers to the forecast value.

MAPE

$$= \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{|\boldsymbol{\hat{y}_i} - \boldsymbol{y_i}|}{|\boldsymbol{y_i} + \boldsymbol{\epsilon}|}$$

DPA

$$= \begin{cases} \boldsymbol{if\ y_i = 0, then\ 1} \\ \boldsymbol{if\ MAPE > 1, then\ 0} \\ \boldsymbol{if\ MAPE \leq 1, then\ 1 - MAPE} \end{cases}$$

sMAPE

$$= \frac{\mathbf{2}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{|\boldsymbol{\hat{y}_i} - \boldsymbol{y_i}|}{|\boldsymbol{y_i}| + |\boldsymbol{\hat{y}_i}| + \boldsymbol{\epsilon}}$$

Absolute Dollar Error

$$= \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} |\boldsymbol{\hat{y}_i} - \boldsymbol{y_i}| * \boldsymbol{Unit\_Cost}$$

Over-Forecast Dollar Error

$$= \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{Max}(\boldsymbol{\hat{y}_i} - \boldsymbol{y_i}, \mathbf{0}) * \boldsymbol{Unit\_Cost}$$

Under-Forecast Dollar Error

$$= \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{Max}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}, \mathbf{0}) * \boldsymbol{Unit\_Cost}$$

Relative Percent Error

$$\boldsymbol{\bar{y}} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{y_i}$$

$$\boldsymbol{Relative\ Percent\ Error} = \frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \frac{|\widehat{\boldsymbol{y}}_{\boldsymbol{\iota}} - \boldsymbol{y}_{\boldsymbol{i}}|}{\overline{\boldsymbol{y}} + \boldsymbol{\epsilon}}$$

MASE

$$\boldsymbol{MASE\ Denominator} = \frac{\mathbf{1}}{\boldsymbol{m}-\mathbf{1}}\sum_{\boldsymbol{j=2}}^{\boldsymbol{m}} |\boldsymbol{y}_{\boldsymbol{j}} - \boldsymbol{y}_{\boldsymbol{j-1}}|$$

$$\boldsymbol{MASE} = \frac{\frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} |\widehat{\boldsymbol{y}}_{\boldsymbol{\iota}} - \boldsymbol{y}_{\boldsymbol{i}}|}{\boldsymbol{Mase\ Denominator} + \boldsymbol{\epsilon}}$$

APPENDIX G: MODEL RUN DOCUMENTATION

Environment Setup

Due to the lack of a centralized AI environment, models will need to be run in an environment with the same capabilities of the interim Anaconda-enabled, air-gapped laptop solution. Prerequisites to running the models are defined below:

- An Anaconda environment with a GPU and updated Nvidia drivers
- The ability to install the environment packages listed in the conda_build.txt file
- The provided project folder with source code and input data

With the necessary prerequisites, the following steps outline the environment setup:

1. Add the project folder provided to the "C:\Users\AIDF" folder
2. Create an anaconda environment based on the conda_build.txt file using the anaconda prompt command "conda create --name AIDF --file ...\project\conda_build.txt"
3. Move the custom_paths.pth file from the "...\project\ folder" to the site-packages folder in your python environment.
    a. This enables the project subfolder to be treated as a library.
4. Open Anaconda Prompt and change directory to "C:\Users\AIDF\project"
5. Run the command "conda activate AIDF"

Generate Ensemble and Base models

1. Modify the parameters in the ensemble_predict.py file located in the …\project\analysis folder under the "Change these parameters" section, as needed
2. Change directory to the "…\project\analysis" folder
3. Run the command "python ensemble_predict.py"
4. All base model results will be stored in the defined save_dir folder, defaulted to "…\project\data\processed\test_model_predictions\ens_in\"
5. The Ensemble model will generate output to the ensemble_destination folder, defaulted to "…\project\data\processed\test_model_predictions\ens_out\"

There are multiple possible ways to run this analysis; however, this method most similarly mirrors the method used to generate the base and Ensemble models evaluated during the AIDF project.

Using the Template File

To enable further research using the environment provided, the template.py file is provided within the "…\project\analysis\" folder. This template provides a basis of comparison for alternative model implementations to be considered.

To use this file, follow the instructions within the Environment Setup section and additionally modify the template.py file to perform your specific analysis. The template accomplishes the following tasks:

1. Loads either dev or test data with a specified lookback period.
2. Splits out train and dev data into separate DataFrames.
3. Preprocesses the data in a modular way that enables a replaceable touchpoint for alternative preprocessing comparisons.

4. Iterates over model parameter options for a given model and saves the model outputs to a specified folder.
5. Aggregates the results of all the models in the specified save folder and generates a table with comparative results.
    a. This section enables selection of multiple different metrics, including sMAPE, DPA, Over-Forecast Dollar Error, Under-Forecast Dollar Error, and Absolute Dollar Error.
6. Graphically compares performance of the top performing models by specified ranges against days with demand.

This template is meant to serve as a basis for further evaluation and a demonstration of the existing code framework. Modularity is critical to the success of this process, so each step can be replaced with an alternative input or approach to enable further analysis.

Note, this not a production environment, but rather a setup that enables quick and flexible implementations of model comparisons and tuning within the context of a proof-of-concept.