Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 2, 2024
A Regional Conference of the Society for Industrial and Systems Engineering

Assessing the Usefulness of Predictive Logistics Models

Hannah Ball, Jonathan Paynter, Thomas Mussmann, and Dylan Hyde

Department of Mathematical Sciences,
United States Military Academy,
West Point, NY 10996

Corresponding author's Email: hannahlball02@gmail.com

Author Note: I would like to express my gratitude to the Artificial Intelligence Integration Center (AI2C) for the incredible opportunity that they have given me by working with the Department of Mathematical Sciences to offer this project. I would also like to thank my research advisors, LTC Jonathan Paynter, MAJ Thomas Mussmann, and CPT Dylan Hyde for the lessons I have learned from them not only about this specific subject but also about project management and the Army in general. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: The goal of this project is to define a quality threshold for when a future given predictive logistics (PLx) artificial intelligence (AI) model might be most useful. This threshold could inform software developers of a target that they need to meet or exceed in order for their product to be considered for integration into the United States Army's current maintenance process for vehicles. A discrete event simulation is used to build out a model of the current Army maintenance process to establish a baseline and a hypothetical predictive model is incorporated with a time horizon and sensitivity for predictions. This results in a function that maps the quality of the model and time horizon for prediction to readiness rate. This function allows us a way to compare the effectiveness of baseline maintenance versus a model that meets the given quality and time horizon descriptions fed to the simulation.

Keywords: simulation, predictive logistics, predictive maintenance, model quality

1. Introduction

The United States Army has been seeking to integrate artificial intelligence and machine learning into the force. It has begun to recognize the cost-savings, efficiency, and potential for better decision making that comes with implementing such a tool. But there aren't resources to research, develop, and apply artificial intelligence into every niche where it might be applicable. Rather, there are decisions that must be made about where to devote the resources to develop a specific model that solves a specific problem. It has been determined that the world of predictive logistics has great potential for the introduction of AI. Predictive logistics in the context of this paper is the idea of ordering vehicle parts before they are necessary in the right quantities at the right echelons so that they can be used as soon as they are needed. The Army's Artificial Intelligence Integration Center (AI2C) is one group that has been developing a predictive logistics model that takes the user's input of maintenance records and usage data and returns a prediction about failure. This allows maintainers to make better decisions about where to spend their time and money. Such decisions are the essence of predictive logistics - finding ways to make smarter decisions about stocking and ordering parts, with the goal of increasing readiness. Organizations that have well-developed predictive logistics practices might see benefits like reduced vehicle down time, reduced surplus inventory, or reduced costs from unnecessary orders.

Delivering the artificial intelligence solution to this problem is certainly an area the Army is interested in exploring for the Army of 2030. The impacts in the field of vehicle maintenance could be substantial. But how good does the model have to be at positively impacting readiness to be worth the investment of resources towards its development and integration? Our research question is this: Can we determine an appropriate mapping statement that takes model quality as an input and delivers a readiness statement as the output, which allows stakeholders to set a quality requirement for model production?

2. Literature Review

A large component of success for this solution will depend on the proper selection of metrics, both to measure the performance of the model and to measure the effect on readiness. We will discuss a number of measurements for performance of binary classification models, and then we will discuss readiness metrics.

2.1. Model Performance - Binary Classifiers

Metrics for binary classifiers have been analyzed extensively for several decades. Alaa Tharwat's 2020 "Classification assessment methods" gives a thorough overview of them. He explains how a confusion matrix gives the basic four numbers that many of the other useful metrics for classification models are built on. Accuracy, precision, recall, false positive/false negative rate, F-measure, and more can all be found from various calculations based on the confusion matrix. (Tharwat, 2020)

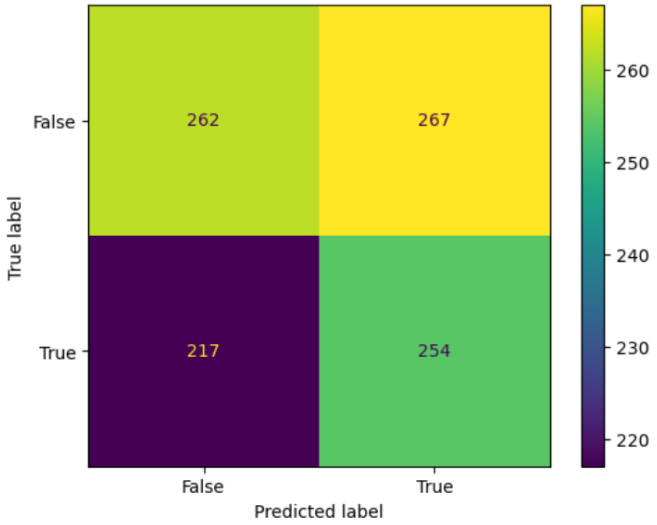


Figure 1: Confusion Matrix: Many of the metrics from the model can be derived from a confusion matrix, such as the one in this figure, derived from python code. The vertical axis is the actual value and the horizontal axis is the label predicted by a model. For example, in the top left square, the actual item is false, but the model predicted that it was true. This happened 262 times before the confusion matrix was generated. All 262 of these values are false positives. Values from each of the sections of a confusion matrix are often the only elements needed to compute performance metrics for a model.

Preliminary stages of the model will rely on binary classification metrics, which lends to easy comparison. As will further be explained later, the human decision phase can be simplified into a binary decision which will allow us to build a simple model to search for any initial trends. The initial model relies on sensitivity because of the way each section of the confusion matrix was incorporated into the simulation, and specificity will be included in future work. (*An Introduction to Statistical Learning*, n.d.)

2.2. Readiness

Operational readiness is defined as "The Army's ability to provide and support CCDRs [Combatant Command Commanders] with trained and ready forces in the quantity and with the capabilities required within needed timelines to meet operational plan requirements of the Defense Planning Guidance" - in other words, does the Army have the people and equipment it needs whenever it is demanded? (*Army Strategic and Operational Readiness*, 2020) Readiness is reported in the Commander unit-status report, a periodic assessment of a unit's readiness across four domains and includes an assessment of equipment. (*Army Unit Status Reporting and Force Registration-Consolidated Policies*, 2022) Many requirements for equipment readiness include reporting the ratio of fully mission-capable equipment versus equipment in other stages of readiness over time, a standard which is governed at least in part by DA PAM 750-8, which describes the procedures for the documentation of equipment failures. (*The Army Maintenance Management System (TAMMS) Users Manual*, 2005) Based on these ideas, our simulation defines operational readiness as the average percentage of vehicles which had no critical part failures over some amount of time.

3. Methodology

We will describe modeling the relationship between model quality and unit readiness in stages. The first stage is to model the current or baseline state of Army maintenance. The second stage adds in a perfectly predictive element. The third

stage is creating imperfect predictions. Finally, we iterate the simulation to create a mapping between the quality of the model as described in the third phase and readiness, then compare it to the baseline. The following subsections will describe these phases in more detail, beginning with the initial simulation.

3.1. Baseline simulation

The initial simulation is meant to model the current Army maintenance process. It models a fleet of vehicles, each vehicle composed of several parts of varying importance to the function of the vehicle. Parts break and are fixed according to a distribution described upon initialization. Each vehicle may also be either functioning and operating or not functioning and in maintenance. At each time step, the fleet's percentage of functioning vehicles can be calculated, visualized in Figure 5.

The simulation follows the typical practice for discrete-event style simulations of using a future event list (FEL) to advance time in the simulation. The current modeled events are a break or a fix. Every part is initialized as working, so the simulation initializes the first break for each part. When the simulation encounters a break in the FEL, it creates the next fix event for that part. If the part is considered a critical part, the broken part deadlines the vehicle that it belongs to and the fleet's readiness drops. The simulation does not model vehicle use time - it could be thought of as runtime per vehicle over arbitrary time units rather than each time step representing an amount of time past the start of the simulation.

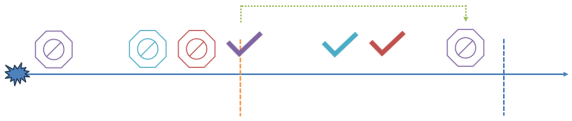


Figure 2: Current Simulation: The simulation modeling the current state of reality employs a typical future event list. The horizontal line represents a timeline, with a starburst image at the moment of initializing the simulation and a dotted line representing the end of the simulation. The crossed-out bolt icon represents the moment a part breaks, and the check marks represent the moment the part is fixed. The color of each icon represents a part - to follow a certain part, follow the icons of the same color. The green line shows a fix of a purple part with the associated next break of a purple part.

3.2. Simulation with Perfect Prediction

This stage of modeling will incorporate perfect prediction. The scenario in consideration is modeled in Figure 3. The idea is that while a normal discrete event simulation only "sees" and handles the first event in the list, a predictive model can "see" for some time horizon in the future. It may "see" that a part is to break in the future, resulting in an earlier fix. For our preliminary predictive simulation, all possible true positives within the given time window were added. We refer to the time window that the simulation can "see" into the future as the time horizon and the model predicting a break within the time horizon as a "lightbulb moment".

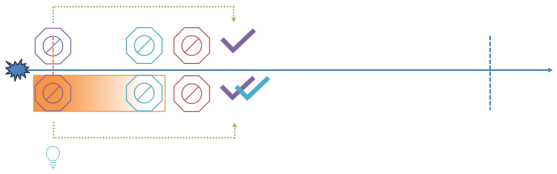


Figure 3: Lightbulb Moment: The second stage of the simulation is to add in the true positive aspect of prediction by adding all possible lightbulb moments within the time threshold and adjusting the simulation to act accordingly afterwards. Most of the symbols are the same as in Figure 2. The parts on the top of the figure are modeling baseline maintenance. The bottom of the figure models what happens when we incorporate perfect true positives. The orange dotted line represents the current moment that the simulation "sees" without prediction, while the orange box represents the time horizon for prediction. The shading represents a variable amount of certainty about predictions, generally decreasing the further in the future the model looks. The lightbulb icon represents the model's recognition or prediction that the blue part will break within the time horizon - the "lightbulb moment".

ISBN: 97819384962-4-0

3.3. Simulation with Imperfect Prediction

After implementing true positives in the model, we were able to easily reduce the chances that the simulation generated a lightbulb moment for every break within the time window. In terms of how the simulation behaves, that meant that sometimes a lightbulb would be generated, and sometimes the simulation would behave the same as the baseline due to a false negative. False negatives could also be thought of as a missing lightbulb moment. This implementation meant that the parameter adjusted to produce a higher or lower percentage of lightbulb moments is actually sensitivity. Recall that the goal of this project is to determine how good an arbitrary predictive model has to be to be worthwhile - while sensitivity is usually a quality metric calculated after the outputs of a parameter, it is actually a parameter on the input side of this simulation.

3.4. Future Simulation Work

The next stage of implementation would be to incorporate false positives into the model and measure their effects alongside the other two concerns of time horizon and sensitivity. However, the effect of a false positive has much more to do with cost to the unit in terms of storage, rather than operational readiness. The current simulation has no way to track fleet level resources like mechanics, scheduling, or inventory, so there is no way to ascribe cost to a false positive in a meaningful way. To deliver a complete picture of the problem, specificity should be incorporated into the simulation next.

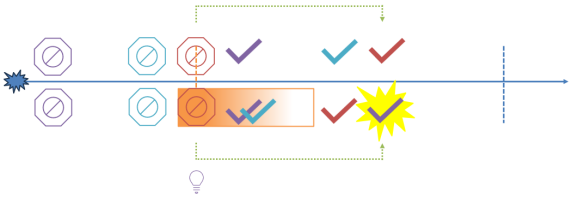


Figure 4: Imperfect Lightbulb Moment: The next stage of simulation would consider false positives, but we do not incorporate them yet in the results.

Varying the time horizon and balance between false positives and false negatives allows us to start asking the main research question again - how good do these models need to be in order for them to still contribute enough value to the Army maintenance system to be worthwhile? As we add more elements of prediction into the model, we begin to form a clearer picture of requirements for what a truly useful predictive model would have to be. This would also be the stage of modeling where we might begin to encounter conditions where a predictive model is actually worse than leaving the current maintenance process alone after considering other types of cost.

4. Results

4.1. Per-Iteration Outputs

Preliminary output metrics include a graph of the percentage of working vehicles for a single simulation. Since deadlining a vehicle is an instantaneous event, the graph appears step-like in nature. Figure 5 simulates only five vehicles with one critical part each, but the code is structured so that it is very easy to scale how many vehicles and critical parts are modeled.

4.2. Aggregate outputs

Iterating the simulation for various sensitivities and time horizons is the final step. This results in a 3D surface - the connection between the previous figure and the 3D graph is that the 2D version is an instantaneous OR rate for a specific sensitivity and time horizon. The 3D surface takes the average of the 2D surface over time and plots that average over time horizons and sensitivities, resulting in Figure 6a.

Figure 6a shows the current progress in the simulation which incorporates both the time horizon and sensitivity aspects of model quality. We can map the model's sensitivity at some time horizon and see its expected readiness rate. Consider first the line where time horizon equals zero. This represents any model which does not have the capacity to "look ahead" into the future. Its sensitivity becomes irrelevant, and surely enough we see on this graph a fairly level readiness rate. Similarly, there is a relatively level readiness rate where sensitivity is zero and no predictions are made. As sensitivity and time horizon generally increase, there is a positive impact on readiness rate, until the effect generally plateaus when time horizon is far enough in advance at a high average percentage of working vehicles per 1000 events.

ISBN: 97819384962-4-0

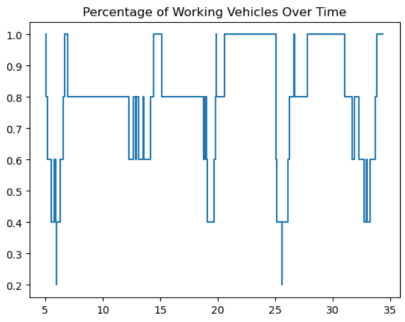
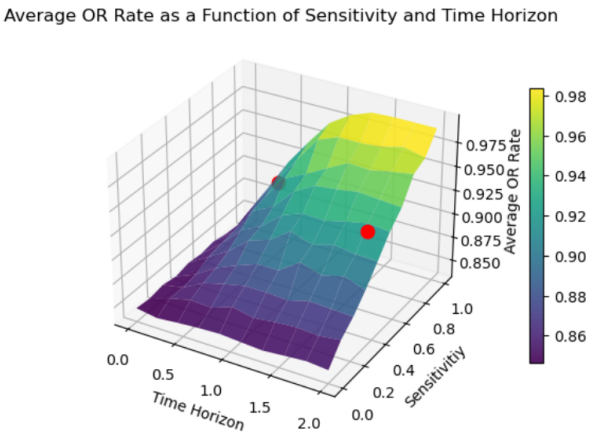
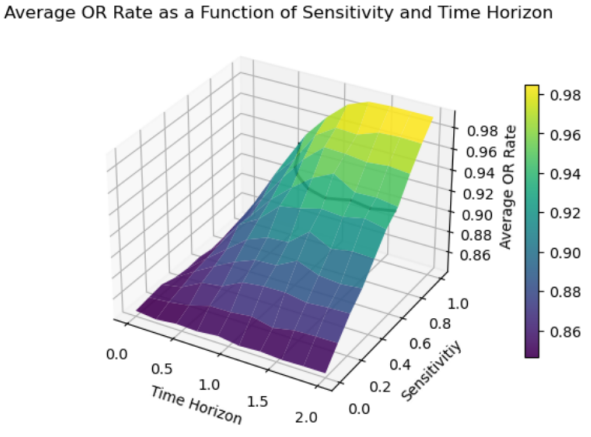


Figure 5: Readiness metrics: At the fleet level, we can access metrics like those incorporated in this graph that represents the percentage of working vehicles (y-axis) over time (x-axis, units arbitrary). This particular fleet fleet models only five vehicles with one critical and one non-critical part each to better demonstrate how some of the changes from step to step might appear.



(a) 20:1 shipping time to fix time.



(b) 2:1 working time to fix time.

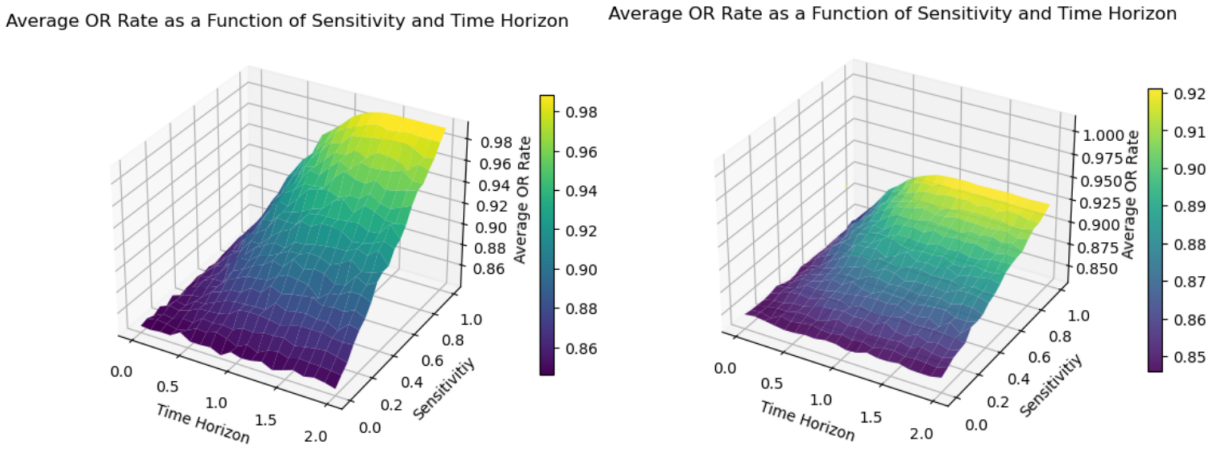
Figure 6: Two graphs generated with different ratios between time that a part is expected to work and time that it takes to fix the part.

Each point on Figure 6a represents a model that could be described by the sensitivity and time horizon axes. The dot in the foreground of the figure would represent a model with a sensitivity of 0.5 and a time horizon of 1.75. We can visually see the benefits of a model that performs with sensitivity of 0.5 at a time horizon of 1.75 time units, because we can see the elevation above the level curves that represent baseline maintenance. However, the dot in the background has a very similar OR rate, but at a time horizon of 0.3 time units and a sensitivity of 0.9. This demonstrates how multiple models could perform very differently regardless of underlying mathematics but still provide the same effect.

That fact brings us back to the research question at hand - how good does a model need to perform to be worthwhile? This simulation cues up a future decision about the actual number, but we are still equipped to visualize what the requirement that we would give to a developer could look like. As long as a developer‘s model performs at or above the contour line in Figure 6b, it meets the requirement. The surface shows the mapping between several acceptable ways to achieve the performance requirement that a developer could aim for.

One major factor that influences this function is the ratios between expected shipping time and the expected time it takes to fix a part. Figure 7a and 7b are visually similar since the time units are arbitrary, but could be thought of as different scales, like 7a is on a scale of weeks versus 7b is on a scale of days. This effect could be less pronounced if the part hardly has any shipping time - there‘s not as much room for any benefits to be realized. The simulation does bring to light how much of an effect something like reducing shipping time might have for parts of drastically different shipping times.

Figure 6b provides us is some insight about mapping combinations of time horizon and sensitivity to an OR rates. Comparing these numbers to the simulated baseline of Army maintenance with no prediction will allow us to generate a threshold that, if met, lands a predictive model’s performance in a space that improves the current maintenance process by an amount



(a) 20:1 shipping time to fix time.

(b) 2:1 working time to fix time.

Figure 7: Two graphs generated with different ratios between time that a part is expected to work and time that it takes to fix the part.

significant enough to warrant its development.

5. Conclusion

In the predictive logistics space, it is necessary to determine how good a model has to be in order for it to add enough value to the current maintenance process to be worth developing and integrating. We developed a simulation that allows us insight into what time horizon and sensitivity the model should reach in order for it to be better than baseline. Modifications that we will incorporate include the idea of limited resources to the simulation, scheduling for vehicles and mechanics, time-triggered maintenance, and modeling based on real days, not just run hours. We anticipate that some of these modifications will directly impact the way false positives might change the output of the simulation while others will have more to do with the interpretation of the results. We can only imagine that a model that incorporates more aspects of reality becomes increasingly trustworthy and useful.

As it stands, this project is useful for developers to think about during their work, but the long-term goal of the project is to deliver a reliable way to determine a performance requirement that developers must meet in order to be considered as a contender for a contract. The work described in this paper paves the way for future researchers to deliver an incredibly powerful statement that guides a piece of the future direction of AI involvement in the Army's maintenance process.

6. References

The Army Maintenance Management System (TAMMS) Users Manual. (2005, August). Department of the Army. Retrieved 2024-03-22, from https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB$_ID = 81693$

Army Strategic and Operational Readiness. (2020, April). Department of the Army. Retrieved 2024-03-22, from https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB$_ID = 1006430$

Army Unit Status Reporting and Force Registration-Consolidated Policies. (2022, August). Department of the Army. Retrieved 2024-03-22, from https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB$_ID = 1021857$

An Introduction to Statistical Learning. (n.d.). Retrieved 2023-09-23, from https://www.statlearning.com

Tharwat, A. (2020, January). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. Retrieved 2023-09-22, from https://doi.org/10.1016/j.aci.2018.08.003 (Publisher: Emerald Publishing Limited) doi: 10.1016/j.aci.2018.08.003