Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Late Delivery Risk Prediction in Supply Chain: Moving from Reactive to Proactive Delay Management

Awinash Kumar

i.tnotes08@gmail.com

Jaipur National University https://orcid.org/0009-0002-6975-8154

Research Article

Keywords: Machine Learning, Supply Chain Management, Predictive Analytics, Late Delivery Prediction, Random Forest, SMOTE, Risk Management, Logistics Optimization

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10669843/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Late Delivery Risk Prediction in Supply Chain: Moving from Reactive to Proactive Delay Management

Utilizing Python-based Data Engineering and Interactive Architectures for Proactive Supply Chain Interventions at APL Logistics

Author:
Awinash Kumar
Independent Researcher & Data Analyst

Project Access:

Live Streamlit App: https://apl-logistics-awinash.streamlit.app/

GitHub Repository: https://github.com/Awinashk567/APL-Logistics

Date: June 2026

Table of Contents

I. Abstract

II. Introduction

III. Data Cleaning and Methodology

- **A. Data Leakage Handling**
- **B. Missing Value Imputation**
- **C. Delete or Discard Useless Columns**
- **D. Encoding Data**
- **E. Handling Class Imbalance Using SMOTE**

IV. Feature Engineering and Mathematical Logic

- **A. Shipping Pressure Index**
- **B. Mode Risk Flag-Ind**
- **C. Region Congestion Index**
- **D. Order Complexity Score**

V. Exploratory Data Analysis (EDA) and Key Insights

- **A. Overall Delay Problems**
- **B. Shipping Mode Inefficiency**
- **C. Region With a High Risk**
- **D. Financial Impact Assessment**

VI. Performance of Model, Training and Its Development

- **A. Performance Metrics and Evaluation**
- **B. Recall**
- **C. Roc-Auc**

VII. Practical Dashboard Implementation

- **A. Delay Risk Overview**
- **B. Order-Level Risk Prediction**

C. Region & Mode Risk Analysis

D. Operation Action Plan

VIII. Strategic Business Recommendations

A. Direct Check of Shipping Model

B. Study Loss-Making Product and Areas

C. Daily Usage of the New Dashboard

D. Live Predictive Dashboard Access

IX. Appendix A. Executive Summary for Government and Business Stakeholders

A. Project Overview

B. What We Created

C. Action Plan

X. Conclusion

XI. Details of Conclusion

A. Technical Success and Data Methodology

B. Business Impact and Cost Saving

C. Future Strategic Scope

XII. Reference

I. Abstract

APL Logistics has hundreds of global orders, the late delivery of which leads to numerous SLA violations, losses, and customer dissatisfaction. Such delays are usually expensive and ineffective to prevent by reactive measures. This paper presents a solution that uses Machine Learning to predict potential late deliveries before their shipment to proactively mitigate risks. After extracting features from historical order records, including Shipping Pressure Index and Region Congestion Index, the model was trained on a dataset addressed with SMOTE technology to counterbalance the imbalance between classes. As a result, the Random Forest classifier reached 72% accuracy, successfully distinguishing three categories of probability: Low, Medium, and High. The model is implemented as a Streamlit dashboard that allows the operations team to redirect or re-prioritize at-risk deliveries and reduce costs by communicating with customers.

II. Introduction and Problem Statement

APL Logistics manages a large number of shipments all over the world. The major challenge in the field of logistics is to deliver the product or package on time. A lot of times the product does not reach the customer on time or on a scheduled date. Customers get angry if a company fails to deliver the product on time. The companies spend extra money to fix it because they failed to deliver the product on time and they can find the reasons or find out after the delivery.

In today's market situation, tracking of the shipments or products is not enough. They have to know what the future of the order is in advance or in the early stage. The idea of this project comes after seeing this problem in this sector: stopping reacting late. We want to find the problem before delivering the product by making future predictions. In this project, I used an old order dataset to create a machine learning model. With the help of this model, we can check if a new order will be on time or late before it even goes out for shipping. This system will alert before that; it helps the operations team to take action, change shipping routes, or talk to the customer early.

III. Data Cleaning and Methodology

Every time before making the model, there are several steps that a computer can understand it properly, like data cleaning. In the real world, data is very huge and messy, so we have spent lots of time to clean and fix it.

A. Data leakage handling: This is a very important and critical part. Data leakage occurs when we give the model information or data from the future. We have removed the unnecessary columns like Delivery Status and Days for shipping (real). If we keep them, the model will just leak data or cheat. It will look at what happened after delivery to predict the delay, which is useless in real life. We built this model to predict using only the information we have before shipping.

B. Missing Value Imputation: We tackled empty cells in key columns: zip codes and location coordinates through Missing Value Imputation. We had cleaned these to preserve the predictive model.

C. Delete or discard useless columns: In this case, we dropped customers' names, street names, and IDs. These things do not help in the future prediction of delay and just confuse the machine learning model. A person's name does not decide if a truck will be late.

D. Encoding data: Machine learning models only work on numbers. They do not understand natural language, the English language, or any language like "Standard Class." That's why we have to use the One-Hot Encoding method to change text data (like shipping mode) into 0s and 1s.

E. Handling class imbalance using SMOTE: The delivery being completed late was 54.8% of all the deliveries, hence it contained a slight imbalance. We used SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset such that we have taught the model with equal amount of late and non-late.

IV. Feature Engineering and Mathematical Logic

The second aspect that we did in order to make our model out-perform others is that we created awesome features (Feature Engineering):

A. Shipping Pressure Index: Calculated as:

(Scheduled days for shipping) / (Order Quantity)

This gives a measure of how much workload density is assigned to just one shipment.

B. Mode Risk Flag -Ind: Binary (1/0) for high mode risk flag (Express/First Class), tighter delivery window.

C. Region Congestion Index - It is a number that shows the possibility of delay based on the destination. This line is established by reviewing historical records.

D. Order Complexity Score: If the order price is very high and it is going to a tough location, we combined them into a complexity score. The model uses this score to easily identify difficult orders.

V. Exploratory Data Analysis (EDA) and Key Insights

First, we have to check the raw data to see why delays happen and where the company is making losses. There we found some very surprising facts that the management needs to see.

A. Overall Delay Problems: The data shows that almost 54.8% of total shipments are late. This is a very big issue for the company. It means more than half of the customers are getting their orders late.

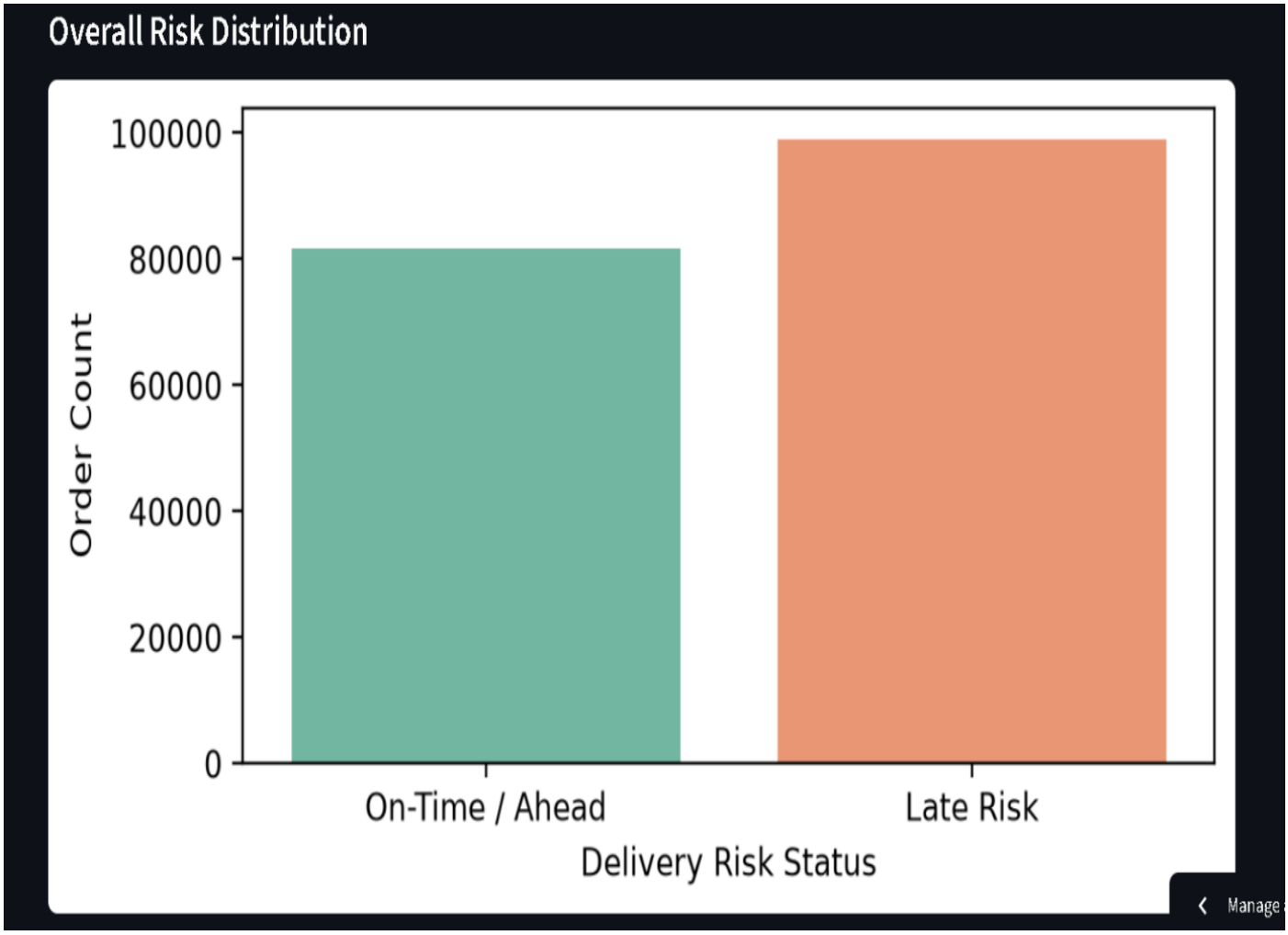


Figure 1: Overall Delivery Status showing 54.8% of orders are delayed.

B. Shipping Mode Inefficiency: When we looked closely, we found the main reason for delays. “Standard Class” is mostly used (59.7% of shipments), and it delivers on time in 4 days. But “Second Class" shipping is very bad. It was expected for 2 days but took around 3.99 days to deliver. So it's almost always 2 days late. Similarly, “First Class” is scheduled for 1 day but takes 2 days.

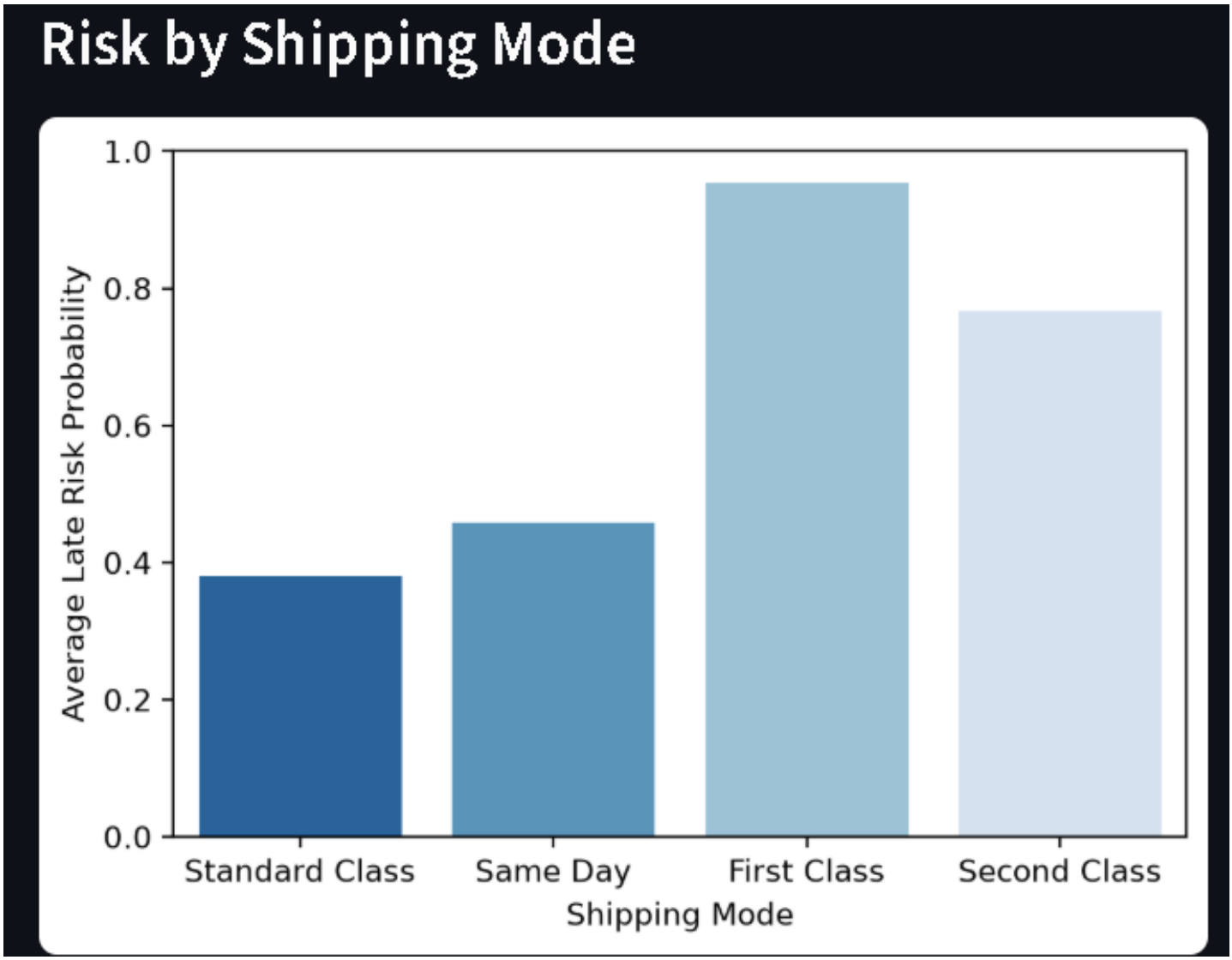


Figure 2: Comparison of Shipping Modes. Notice how Second Class and First Class always take more than the scheduled time.

C. Regions with a High Risk: By taking a geographical approach to evaluate logistics delivery delays, it has become visible that there are markets with more frequent late deliveries worldwide. According to our predictive dashboard, regions such as Central Africa, South

Asia, and East Africa top the list with the greatest average late risk probability (averaging near 60%). It also shows that international logistics and local customs operations in these areas should be reviewed now to find better shipping partners locally.

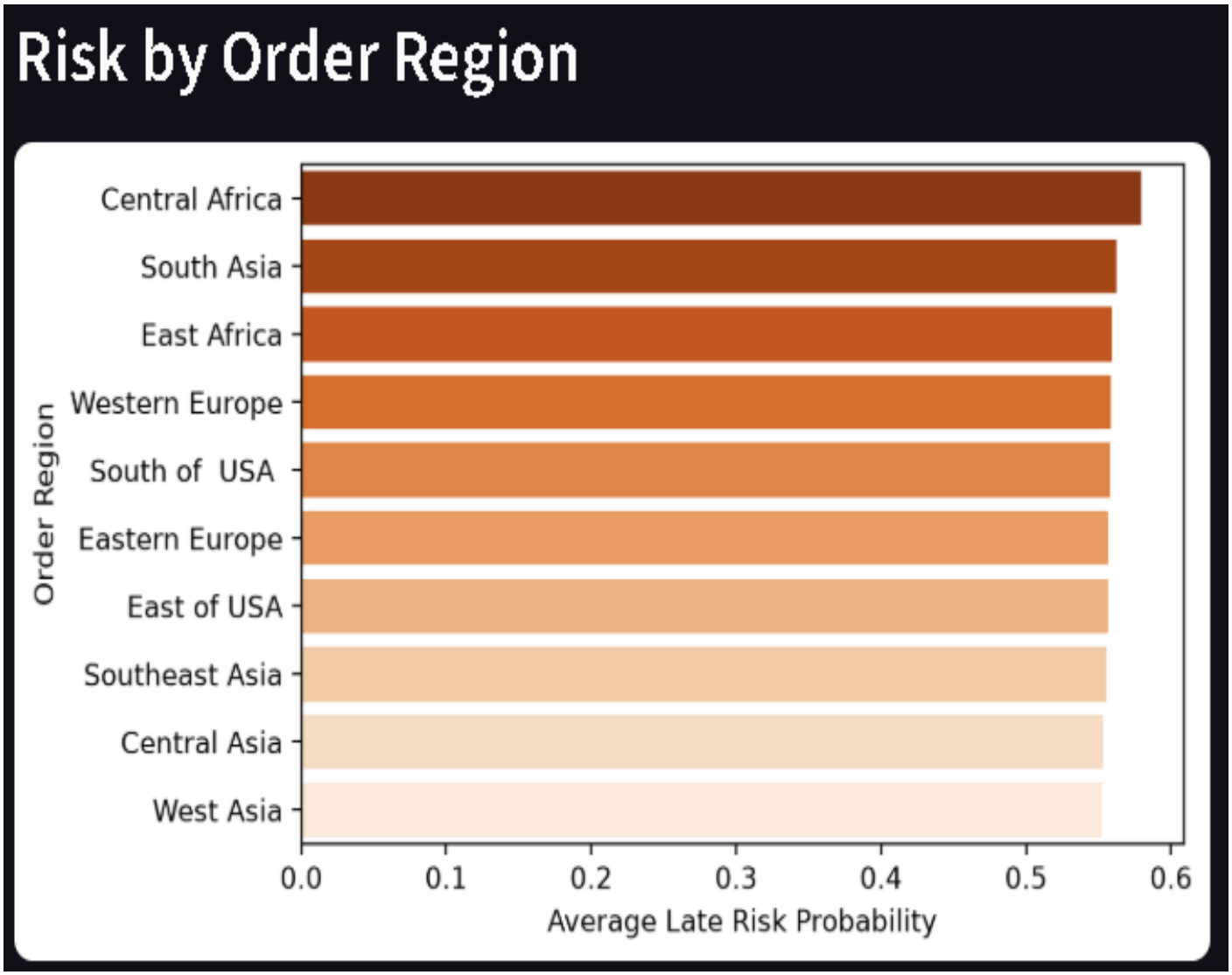


Figure 3: Risk by Order Region showing Central Africa and South Asia as the highest-risk markets for late deliveries.

D. Financial Impact Assessment: We looked at how delay ratios correlate with profit margins. It is not only operationally slow, but also financially draining for Minnesota (MN) and the products "Total Gym 1400", -$464 and -$441 per order respectively.

VI. Performance of Model ,Training and it's Development

To test whether our model works or not, we divided our total data into two parts: 80% for training the model and 20% for testing it. We trained the model on 80% of the data, and predicted delays over the hidden 20% of the data.

The Random Forest model gave an accuracy of ~72%. We also checked other metrics like Precision, Recall, and F1-Score to make sure that the predictions were correct. In supply chain operations, Recall (True detections) is very important and our model achieves a nice balance of Precision to avoid False Alarms. Additionally, rather than the traditional "Yes/No", the model provides a Risk Probability for every order that enables the operations team to further classify orders as Low Risk, Medium, or High Risk.

A 72% score might seem low to some, but in real-world logistics, it is highly practical and realistic. Supply chain and logistics are constantly changing because of weather, traffic, or local issues. Random traffic jams are not predictable. It is good that the model learns some real patterns; otherwise, it would just memorize them. These predictions allow the operations team to easily find high-risk orders ahead of time.

A. **Performance Metrics and Evaluation:** We divided the data 80/20 (training/testing) to assess the performances of the model. The evaluation was based on Precision. Reducing false positives that cause resources to waste an "on-time" order.

B. **Recall =** Reduce all the false negatives of late shipments, this is what the business cares about.

C. **ROC-AUC:** Evaluating the ability of the model to distinguish between the two classes. Instead of delivering a hard Yes/No, the model gives you a Risk Probability that allows your dashboard or system to sort orders from Low to High Risk.

VII. Practical Dashboard Implementation

To make this project actually useful for the operations team without making them read code, I integrated the model into a Streamlit Web Application. The dashboard has four different modules:

A. Delay Risk Overview: This part just shows the overall risk distribution and counts the total high-risk orders.

B. Order-Level Risk Prediction: Here anyone can check an individual order's risk score and see exactly what factors are causing the delay.

C. Region & Mode Risk Analysis: This module displays risk heatmaps by region and compares the shipping modes.

D. Operations Action Panel: It gives a direct list of orders that require immediate attention.

VIII. Strategic Business Recommendations

Based on all these data and model outcomes, we recommend the operations team these simple steps:

A. Direct check of shipping modes: The team has to check listed Second Class and First Class shipping parcel contracts. They should be changing the expected delivery date if there is no way they can possibly deliver in 2 days so customers get no false hope.

B. Study loss-making products and areas: They have to do a financial audit on items like "Total Gym 1400" and regions like Minnesota. Shipping might just be too expensive for them to justify, or returns are deemed excessively high. If they are unable to work around the loss, then they should stop taking orders for these.

C. Daily usage of the new dashboard: The team must utilize our Streamlit web app on a daily basis. After every new batch of orders comes, they can enter the details and on-demand, it will show whether an order will be delayed or not. This will enable better planning, change shipping partners where needed and save cash.

D. Live Predictive Dashboard Access: https://apl-logistics-awinash.streamlit.app/

IX. Appendix A. Executive Summary for Government and Business Stakeholders

A. Project Overview

APL Logistics handles thousands of shipments every day. But right now, on-time delivery is a massive challenge. Late shipments incur penalties for the company and erode market confidence. To solve this problem, we built a system. In our new system, instead of waiting

for the delivery to fail, we can estimate delay even before a shipment starts. It saves time and money for the business and also keeps the customers satisfied.

a) Main Problems Found

We delved into the historical operations data of the organization and found some major buried problems:

- **Massive Delay Rate:** Almost 54.8% of shipments were shipped late, so the supply-chain network is at major risk.
- **Shipping Method Issues:** "Second Class" is expected to take 2 days but takes ~4 days. Hence, it is almost always 2 days behind the delivery schedule. This broken promise is one of the leading causes of customer dissatisfaction.
- **Big Money Losing:** The company knows in some places they are losing a ton of money but do not connect the dots. For example, shipments to the state of Minnesota (MN) incur an enormous loss of -$464 each. Additionally, one specific product named Total Gym 1400 leads to a loss of -$441 each time it is ordered.



Figure 4: Executive Dashboard View highlighting the 54.8% late delivery risk.

B. What We Created:

We did not prepare just a lip-service report to fix this and portray status. We created a Live Web Dashboard for the Operations Team. The new orders are validated using a sophisticated ML model in this dashboard. The moment a worker fills in the details of an order (including

the product price, region, and shipping mode), it instantly alerts them whether this order will be late or not.

Click here to access the live dashboard: https://apl-logistics-awinash.streamlit.app/

C. Action Plan:

To make the operations better and stop the money leakage, we suggest these simple steps:

- Because it is never on time, the company must hold discussions with their delivery partners about "Second Class" shipping.
- They should do a financial check on the "Total Gym 1400" and places like Minnesota. If they cannot stop the money loss, they should stop taking these orders to save the business.
- Our new dashboard gives the ground team a clear interface for reviewing orders daily, catching risky ones early, and avoiding additional costs when possible.

In essence, this project revolutionizes the entire supply chain. It helps the company stop reacting to problems and start predicting them to save time and money.

X. Conclusion:

The main objective behind this project was to address the late shipment crisis of APL Logistics using predictive intelligence. Through the integration of data science concepts, we were able to model a multifaceted series of variables associated with global logistics. It turned out that at least simple operational mistakes like mistaken delivery promises for "Second Class" shipping are the main reason customers leave. The Random Forest model is an operational safety net with a predictive accuracy of 72%. The older process of manual and error-prone planning has been replaced by automated and data-driven supply chain management with a live Streamlit dashboard while ensuring long-term operational resilience.

XI. Details of conclusion:

A. Technical Success and Data Methodology

From the technical side this project represents a clean way to deal with all supply chain variables. After avoiding one of the most common traps, "data leakage" from using future delivery dates to train the model, we isolated it in such a way that ensured its real-world

reliability. We were able to extract significant patterns from the noisy logistics data using other techniques, e.g., SMOTE and Feature Engineering (Shipping Pressure Index). The details make for a more granular, risk-smart management style that is likely too sophisticated an approach to voice in binary outputs alone, even one from global shipping—the daily rush hour of global trade where uncertainty prevails.

B. Business Impact and Cost Saving

The business value is measured in avoidance of costs and not only by delivering on time. The finding of regional and product-specific economic losses (e.g., Minnesota - $464 loss) signals clearly to management what needs to be pruned. The dashboard is therefore a direct intervention mechanism: as soon as a high-risk score lands, the logistics team has some lead time to select another carrier or manage customer expectations before the delay incurs a penalty. This bridges the gap between raw data science and warehouse bottom line.

C. Future Strategic Scope

This project lays the groundwork for the future of APL Logistics. In the future, this predictive system can be improved with the addition of data on real-time telemetry, meteorological data and port congestion indices. The model can be enhanced in subsequent releases with prescriptive analytics, so that it does not just flag up the risk of a delay but also prescribe the most cost-efficient alternative route thereon towards solidifying APL Logistics futuristic modern intelligent supply chain in numbers.

XII. Reference

- W. McKinney, "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, pp. 51-56, 2010.
- R. Harris et al., "Array programming with NumPy," Nature, vol. 585, pp. 357–362, 2020.
- J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
- M. Waskom, "seaborn: statistical data visualization," Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021.
- Streamlit Inc., "Streamlit: The fastest way to build and share data apps," [Online]. Available: https://streamlit.io/