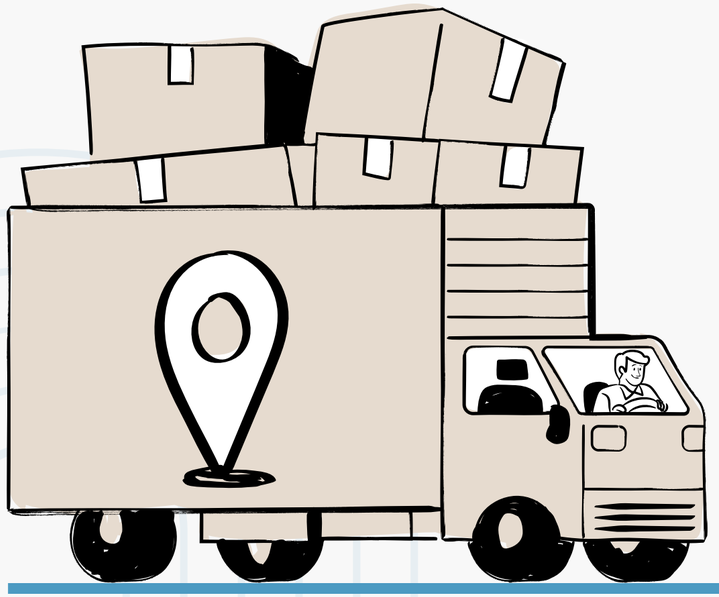
SMARTER ROUTES, FASTER DELIVERIES

DATA MINING FOR SUPPLY CHAIN EFFICIENCY

MRINALI KARTHIK | SAHIL RAWAL | TANISH PATHAK | VAIBHAVI GODSE | UJWAL RAJ



GROUP 2

14 April, 2025

BUSINESS PROBLEM
&
PROJECT OVERVIEW

- In this project, we set out to uncover what truly causes **“shipment delays”** and more importantly, how to predict them before they happen. We analyzed a comprehensive logistics dataset that includes factors like traffic status, temperature, humidity, waiting time, and shipment status. Using this data, we conducted exploratory analysis to identify trends and patterns, and then built predictive models to classify whether a shipment would be delayed or not.

- We conducted **Feature Engineering**, **Exploratory Data Analysis**, and **built “Logistic Regression”, “Decision Tree”,** and **“Random Forest”** models. Focus was placed on improving accuracy and reducing false negatives critical in preventing overlooked delays. Our final model offers data-backed guidance to optimize deliveries and enhance customer satisfaction.

ABOUT CHOSEN DATASET

- We worked with a dataset containing **1,000 observations**.
- The dataset captures a diverse range of features, from environmental conditions (temperature, humidity) to logistics metrics like shipment status, traffic, and waiting time.
- With a binary target variable **“Logistics_Delay”**, we’re able to clearly define and model what constitutes a delay.
- We also get a glimpse into customer behavior, including transaction amounts and purchase frequency adding depth to our analysis.
- Categorical variables like Traffic_Status and Delay_Reason bring in context, though they require encoding before modeling.
- Columns like Asset_ID and Timestamp offer little predictive power and were set aside.
- Some variables, especially those highly correlated, posed a risk of overfitting, and had to be handled with care.

FROM RAW TO REFINED

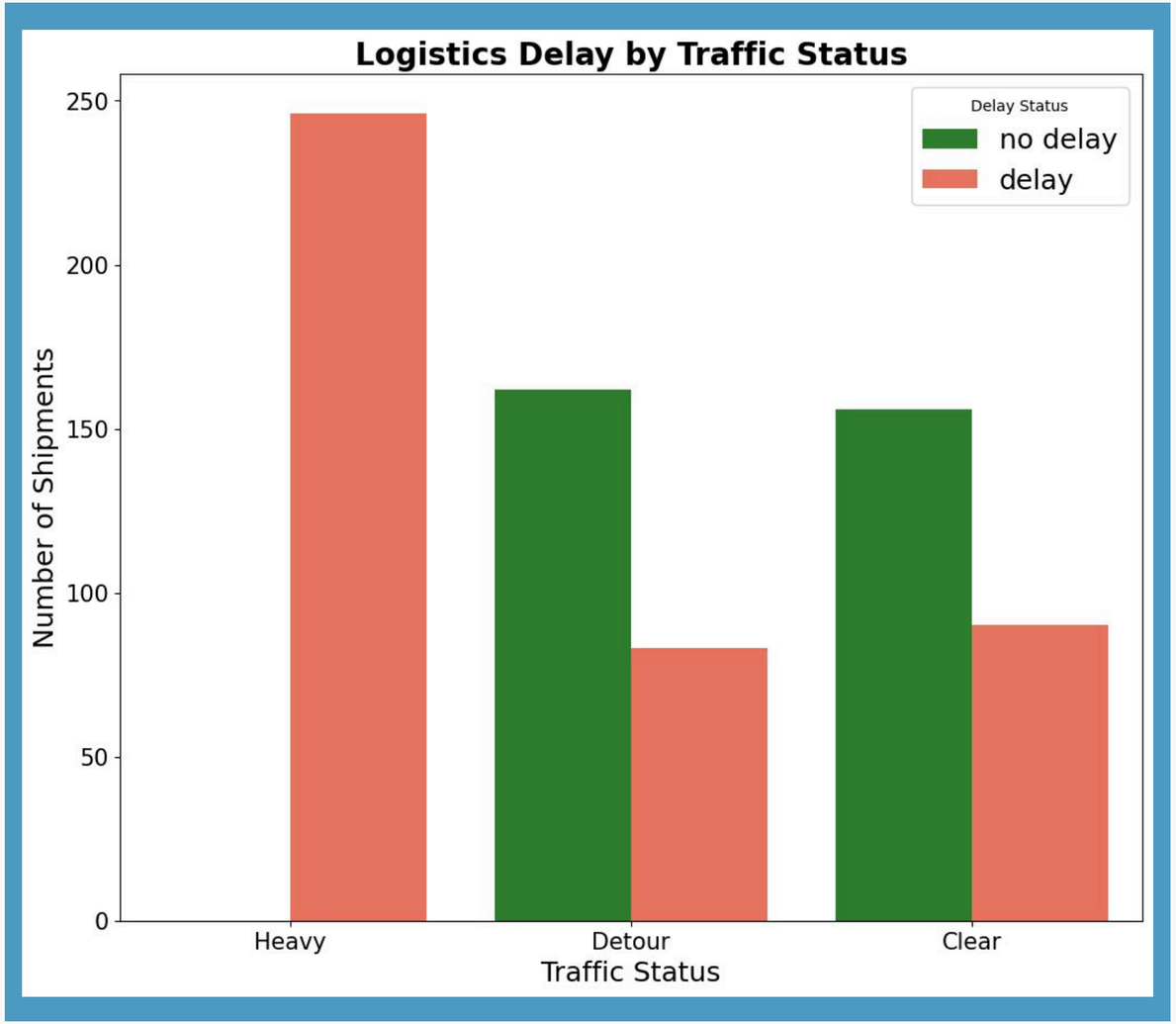
PRE-PROCESSING THE DATA

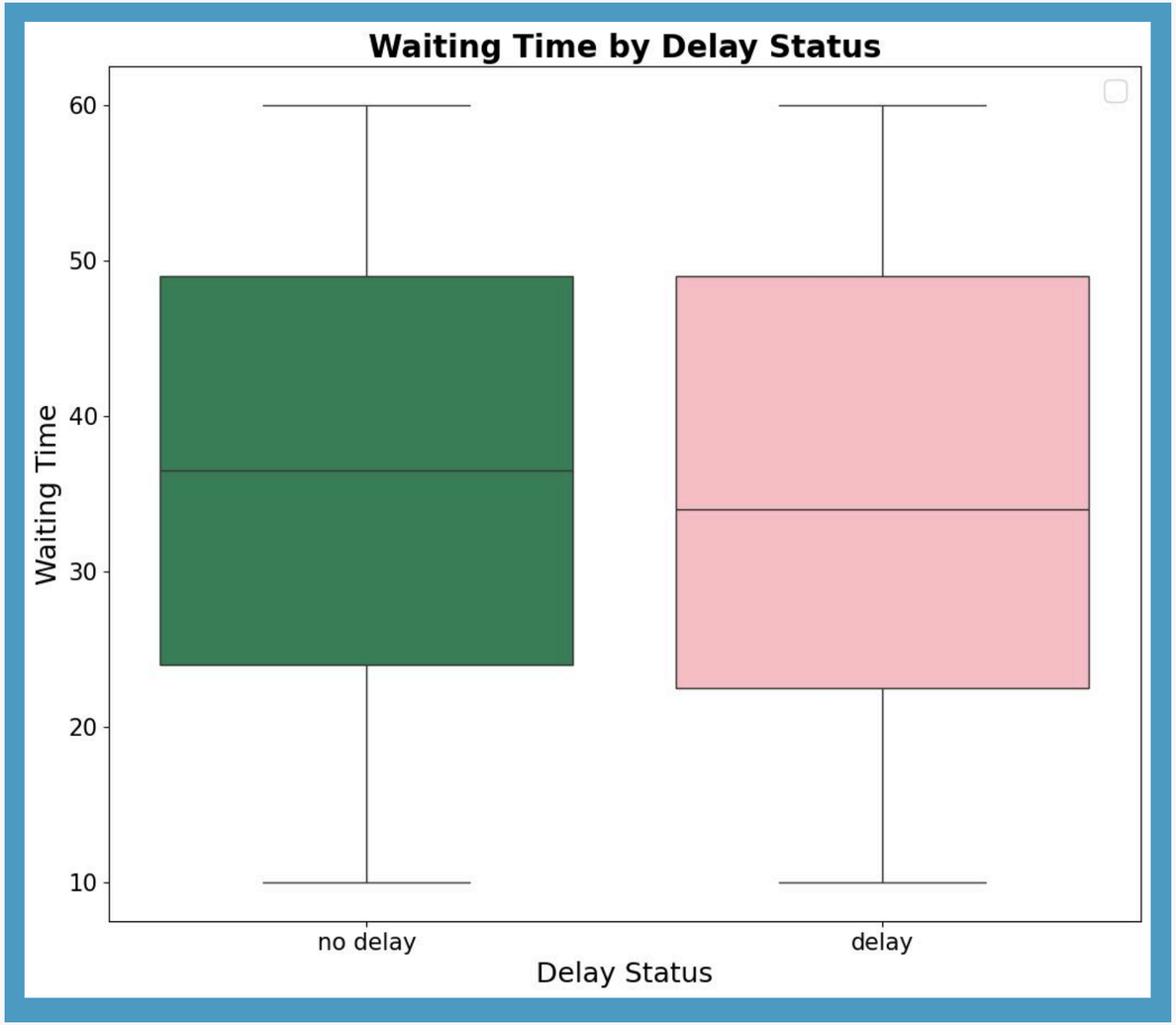
- We began by cleaning the dataset; null values were removed to keep only complete records.
- From the Timestamp column, we extracted **“date”** and **“month”** information to enable seasonal trend analysis.
- Dropped unnecessary columns like **“Asset_ID”** and **“Timestamp”** to reduce noise and improve model focus.
- We created a new **“MonthName”** column for clearer visualisation's during Exploratory Data Analysis.
- Categorical features like Traffic_Status, Shipment_Status, and Logistics_Delay_Reason were converted to **“Dummy Variables”** to make them model-ready.
- We also created a **“binary DelayStatus”** variable for easier interpretation during visualization.

EXPLORATORY DATA ANALYSIS

Index	Latitude	Longitude	nventory_Leve	Temperature	Humidity	Waiting_Time	Transaction_A	Purchase_Freq	sset_Utilizatio	emand_Foreca	.ogistics_Delay	Month
Latitude	1	-0.00188656	0.00149757	-0.00592858	-0.0504702	-0.00410512	-0.0390312	-0.00317731	-0.0193202	0.00015896	0.0583313	0.0547979
Longitude	-0.00188656	1	-0.027369	-0.0154026	0.0342985	-0.00025428	-0.0207148	-0.0324096	0.0233405	0.121546	0.0406014	-0.0510273
Inventory_Level	0.00149757	-0.027369	1	-0.047892	0.0479632	-0.0054145	-0.0437	-0.0222163	0.0199633	-0.0341177	-0.017599	0.0324092
Temperature	-0.00592858	-0.0154026	-0.047892	1	-0.0140601	0.0277914	-0.0108137	-0.00101826	0.0619203	0.00462642	-0.0391206	-0.00844859
Humidity	-0.0504702	0.0342985	0.0479632	-0.0140601	1	0.0350077	-0.00987434	0.0449618	-0.0333077	-0.00495641	-0.0056439	-0.070221
Waiting_Time	-0.00410512	-0.00025428	-0.0054145	0.0277914	0.0350077	1	0.0336076	0.0551975	-0.0264179	-0.0171432	-0.0592418	0.0309371
User_Transaction_Amount	-0.0390312	-0.0207148	-0.0437	-0.0108137	-0.00987434	0.0336076	1	0.0649201	0.0370841	-0.0311888	0.0256256	0.0421021
User_Purchase_Frequency	-0.00317731	-0.0324096	-0.0222163	-0.00101826	0.0449618	0.0551975	0.0649201	1	0.0272649	-0.0933327	-0.0255592	-0.00707415
Asset_Utilization	-0.0193202	0.0233405	0.0199633	0.0619203	-0.0333077	-0.0264179	0.0370841	0.0272649	1	-0.113889	0.0339304	-0.00816171
Demand_Forecast	0.00015896	0.121546	-0.0341177	0.00462642	-0.00495641	-0.0171432	-0.0311888	-0.0933327	-0.113889	1	-0.00405289	0.0266973
Logistics_Delay	0.0583313	0.0406014	-0.017599	-0.0391206	-0.0056439	-0.0592418	0.0256256	-0.0255592	0.0339304	-0.00405289	1	-0.0245346
Month	0.0547979	-0.0510273	0.0324092	-0.00844859	-0.070221	0.0309371	0.0421021	-0.00707415	-0.00816171	0.0266973	-0.0245346	1
Shipment_Status_Delivered	-0.0693691	0.00985331	-0.00609321	0.0235102	0.0351145	0.0252027	-0.0341497	-0.0470554	-0.0511794	0.0208946	-0.319574	-0.0427542
Shipment_Status_In Transit	0.00574285	-0.028812	0.0814216	0.0115724	0.0251939	0.0116232	0.00052937	-0.00411283	0.0242826	-0.0197183	-0.329664	0.0330766
Traffic_Status_Detour	0.0564949	-0.0448425	-0.0281666	0.0433133	-0.0200213	-0.00470619	1.85243e-05	0.0564188	-0.0262526	-0.00895763	-0.327337	0.0607468
Traffic_Status_Heavy	0.039602	0.0546406	0.0318932	-0.0158424	0.0445276	-0.0262246	-0.00643953	-0.0793886	0.0551657	-0.0180285	0.616642	-0.0320199
Logistics_Delay_Reason_Traffic	-0.00904113	0.0321637	-0.039486	0.0405572	0.0534054	0.0270672	0.030384	0.0467181	-0.0646714	0.0079652	0.00486795	0.024682
Logistics_Delay_Reason_Weather	0.0155816	-0.0104321	0.0472546	-0.00134208	-0.022932	-0.0532674	0.00280588	-0.00728099	-0.0130609	0.00507837	-0.00453182	-0.0611299

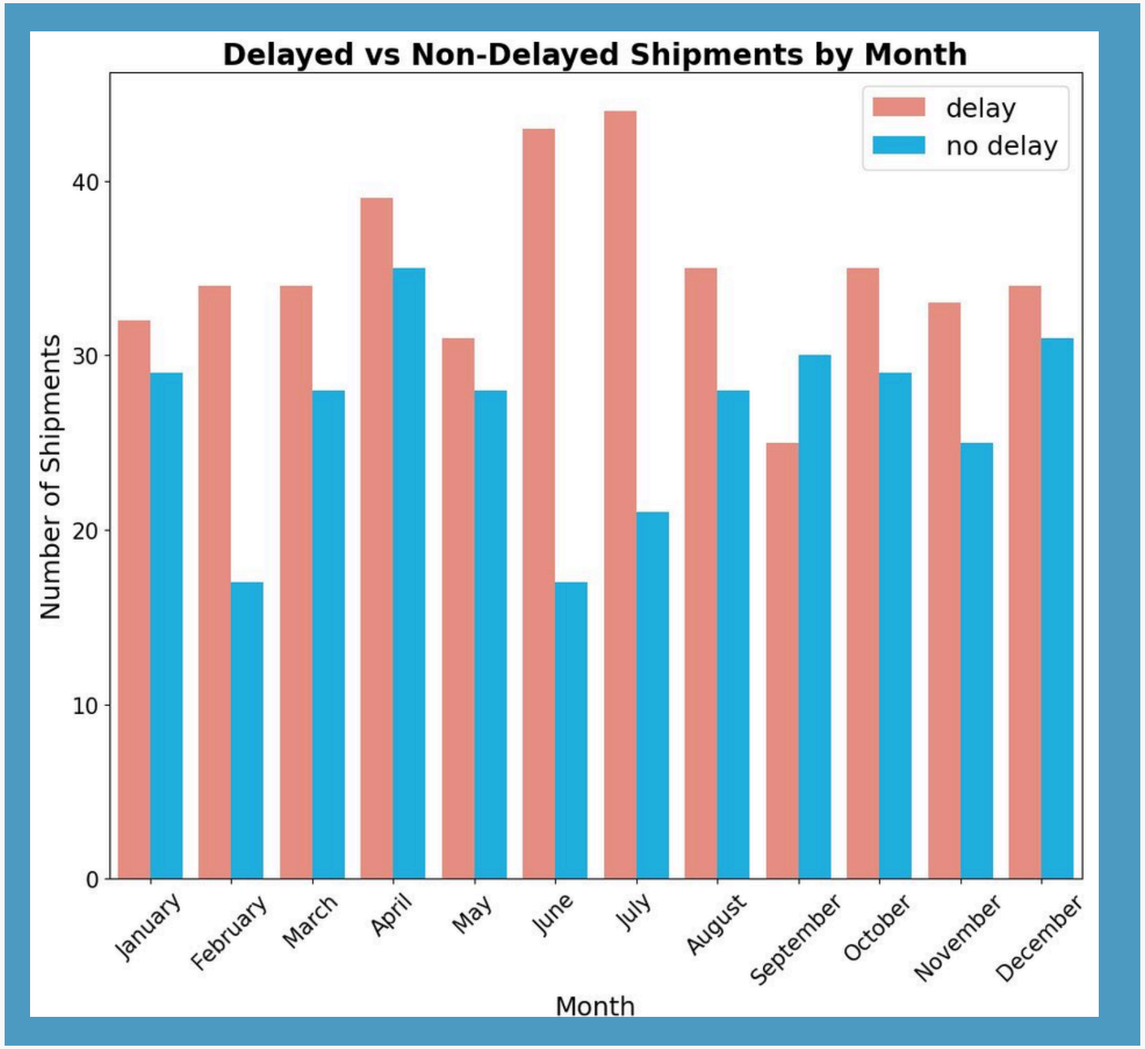
The correlation table uncovers relationships between variables, helping us identify which features may influence delays the most.

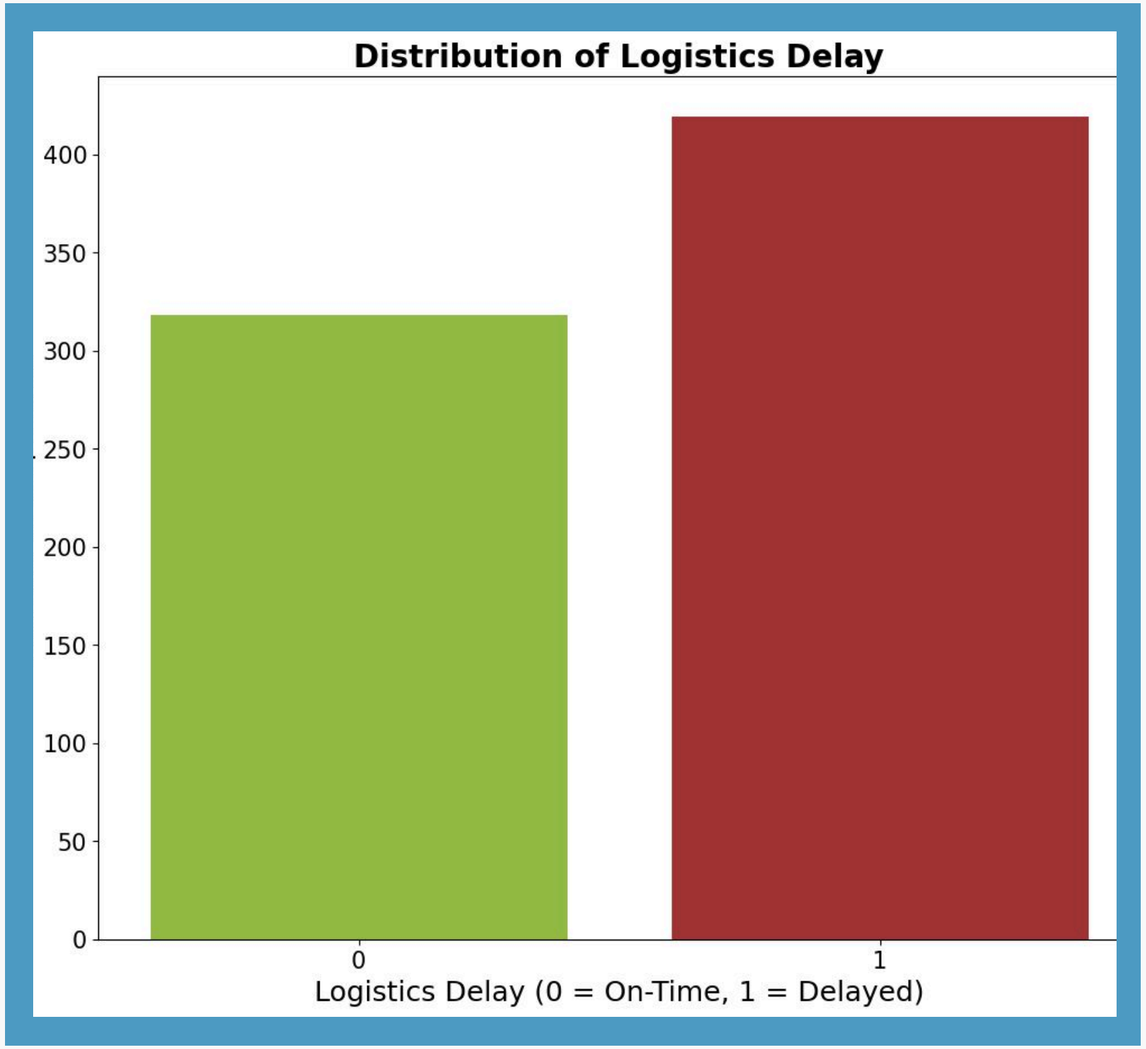




Heavy traffic strongly correlates with shipment delays, while delayed shipments also show slightly longer and more variable waiting times.

VISUAL INSIGHTS





Delayed shipments are more frequent overall, with noticeable peaks during mid-year months, indicating both general and seasonal delay patterns.

PREDICTIVE MODELS TESTED

Logistic Regression:

- It helped us understand feature importance and served as a benchmark for performance.
- It showed strong results in precision, F1 score, and lowest false negatives.

Decision Tree Classifier:

- It allowed us to capture non-linear relationships between features.
- It achieved high F1 score in training but slightly overfit the data.

Random Forest Classifier:

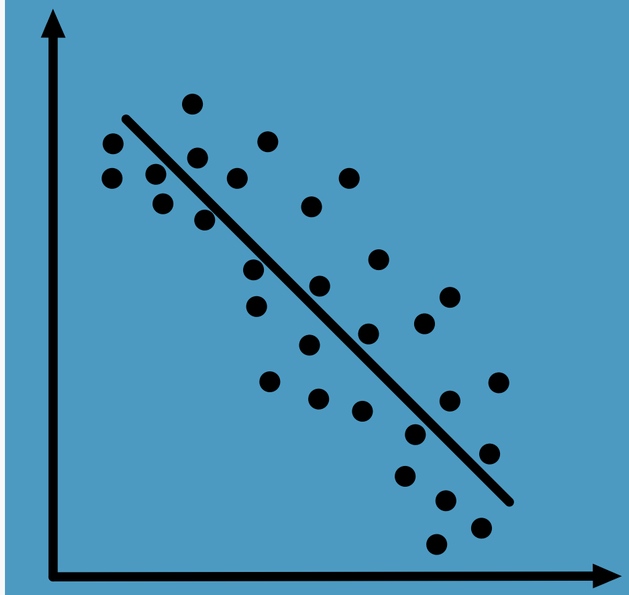
- It improved generalization and reduced overfitting.
- It gave high training accuracy, but false negatives were slightly higher than logistic regression.

Performance Metrics of Tested Models

Model	F1 Score	Recall	Precision	False Negatives
Logistic Regression (Default)	0.8	0.72	0.91	38
Logistic Regression (Tuned)	0.87	0.95	0.8	6
Decision Tree	0.8	0.74	0.91	34
Random Forest	0.82	0.78	0.87	32

And the Best Model Is...

- **“Logistic Regression”** delivered consistently high F1 and recall scores.
- After adjusting the classification **threshold** to **0.3**, it minimized **false negatives (6)** significantly that are critical in logistics where missing a delay can lead to operational risks.
- While Random Forest and Decision Tree performed well, they couldn’t match the reliability and precision of Logistic Regression in flagging delayed shipments correctly.
- We selected Logistic Regression not just for its solid performance, but because it aligned best with our business goal: **“reducing undetected delays”**.



CONFUSION MATRIX

Index	Pred:0	Pred:1
Actual:0	56	32
Actual:1	6	128

KEY INSIGHTS FROM THE DATA

Key Factors Contributing to Delays

Waiting Time, Traffic Status, and Inventory Level are the top predictors of shipment delays. Delayed shipments typically have higher waiting times and are more frequently associated with heavy or detour traffic conditions.

Impact of Traffic Conditions

Shipments that encounter 'Heavy' or 'Detour' traffic are significantly more likely to be delayed compared to those with 'Clear' status. This suggests traffic-based route optimization could reduce delays.

Role of Inventory Level

Low inventory levels are often associated with higher delay rates, possibly due to restocking delays or under-optimized asset utilization.

KEY INSIGHTS FROM THE DATA

Environmental Conditions May Have Limited Influence

Humidity and temperature show weaker correlations with delays, suggesting that while weather matters, it's not a dominant factor in this dataset.

Predictive Model Can Accurately Flag Delays

Using logistic regression and feature selection, we achieved an F1 score of **~0.80**, showing that even basic models can provide strong predictive performance for shipment delays.

Operational Recommendations

We recommend prioritizing high-waiting-time shipments for proactive management, integrating real-time traffic insights into dispatching decisions, and ensuring buffer inventory during peak months like July and August.

CONCLUSION

- In the complex world of logistics, even small delays can have a ripple effect on efficiency, cost, and customer satisfaction. Through this project, we transformed raw operational data into meaningful insights; uncovering patterns hidden in traffic, time, and seasonality.

- By testing multiple models and evaluating them through important metrics, we found that Logistic Regression when thoughtfully tuned offered the best balance of accuracy and risk mitigation, particularly by minimizing false negatives.

- Our approach not only predicts delays but also empowers decision-makers to act early, optimize operations, and move one step closer to a smarter, more resilient supply chain.

Thank you