Enhancing E-commerce Delivery Efficiency through Advanced Analytics and Machine Learning

Business Context & Executive Summary

In the highly competitive e-commerce logistics landscape, delivery delays pose a significant threat, impacting not only operational efficiency but also customer satisfaction and brand reputation. A global electronics company initiated this project to address and mitigate these delivery inefficiencies through the strategic application of advanced analytics and machine learning.

Goal:

To identify and prevent delivery delays, optimise packaging and batching logic, and establish a robust routing framework that improves SLA adherence and customer satisfaction.

Solution:

Using a dataset of 10,999 historical deliveries, a multi-faceted approach was adopted, comprising four key analytical pillars:

- Exploratory Data Analysis (EDA) to uncover operational bottlenecks
- Clustering orders to optimise packaging and batching logic based on operational similarities.
- Predictive modelling to score future deliveries by delay risk, enabling proactive intervention.
- Routing logic & courier rules to optimise last-mile performance

Data Dictionary

- ID: Unique identifier for each customer/order.
- Warehouse_block: Warehouse segment where the order was processed (A, B, C, D, E).
- Mode_of_Shipment: Shipment method used: Ship, Flight, or Road.
- Customer_care_calls: Number of calls made by the customer regarding the order.
- Customer_rating: Rating given by the customer (1 = Worst, 5 = Best).
- Cost_of_the_Product: Cost of the product in USD.
- Prior_purchases: Number of purchases made by the customer before this order.
- Product_importance: Importance level of the product (low, medium, high).
- Gender: Gender of the customer (Male or Female).
- Discount_offered: Discount applied to the product (in percentage).
- Weight_in_gms: Weight of the product in grams.
- Reached.on.Time_Y.N: Target variable: 1 = On Time, 0 = Late.
- Courier_ID: Randomly assigned anonymised courier handling the delivery.

Data Overview

The dataset comprises 10,999 entries and 13 columns, providing a comprehensive view of historical e-commerce deliveries. Data quality was high, with no missing values and no duplicate rows. The target variable Reached.on.Time_Y.N was renamed to Reached_on_time for clarity. Approximately 59.7% of deliveries were made on time, while 40.3% experienced delays, highlighting the core problem. Discount_offered showed a wide range, suggesting varied promotional strategies.

	count	mean	std	min	25%	50%	75%	max
ID	10999.0	5500.000000	3175.282140	1.0	2750.5	5500.0	8249.5	10999.0
Customer_care_calls	10999.0	4.054459	1.141490	2.0	3.0	4.0	5.0	7.0
Customer_rating	10999.0	2.990545	1.413603	1.0	2.0	3.0	4.0	5.0
Cost_of_the_Product	10999.0	210.196836	48.063272	96.0	169.0	214.0	251.0	310.0
Prior_purchases	10999.0	3.567597	1.522860	2.0	3.0	3.0	4.0	10.0
Discount_offered	10999.0	13.373216	16.205527	1.0	4.0	7.0	10.0	65.0
Weight_in_gms	10999.0	3634.016729	1635.377251	1001.0	1839.5	4149.0	5050.0	7846.0
Reached_on_time	10999.0	0.596691	0.490584	0.0	0.0	1.0	1.0	1.0

Exploratory Data Analysis (EDA) Report

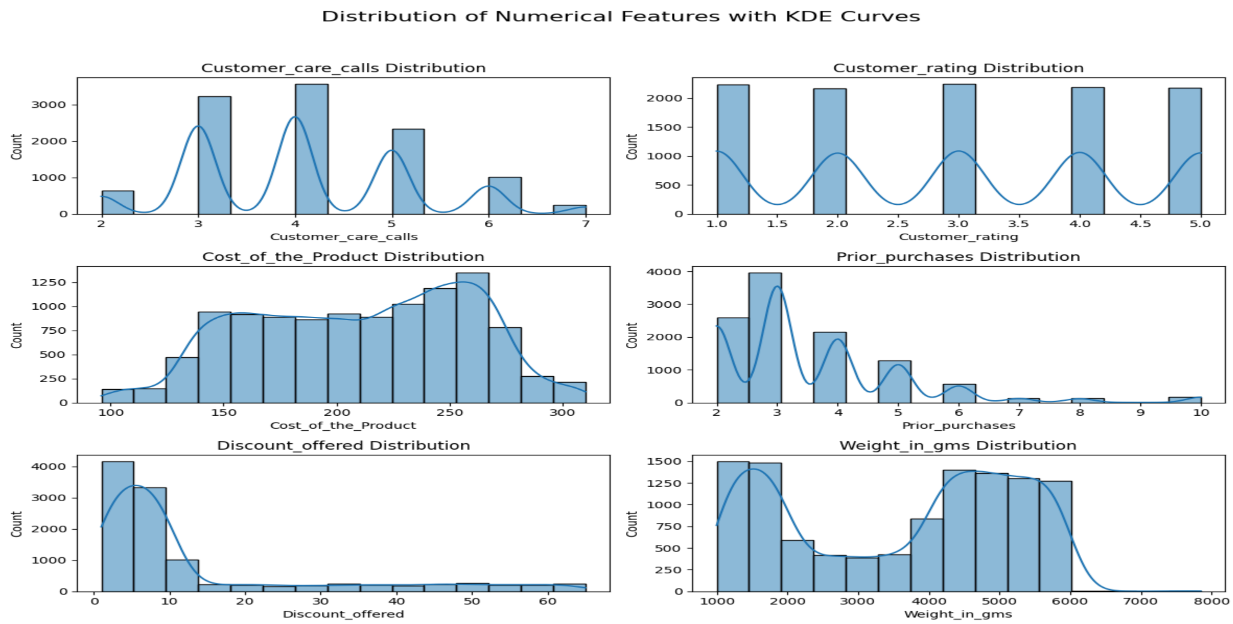
Univariate Analysis

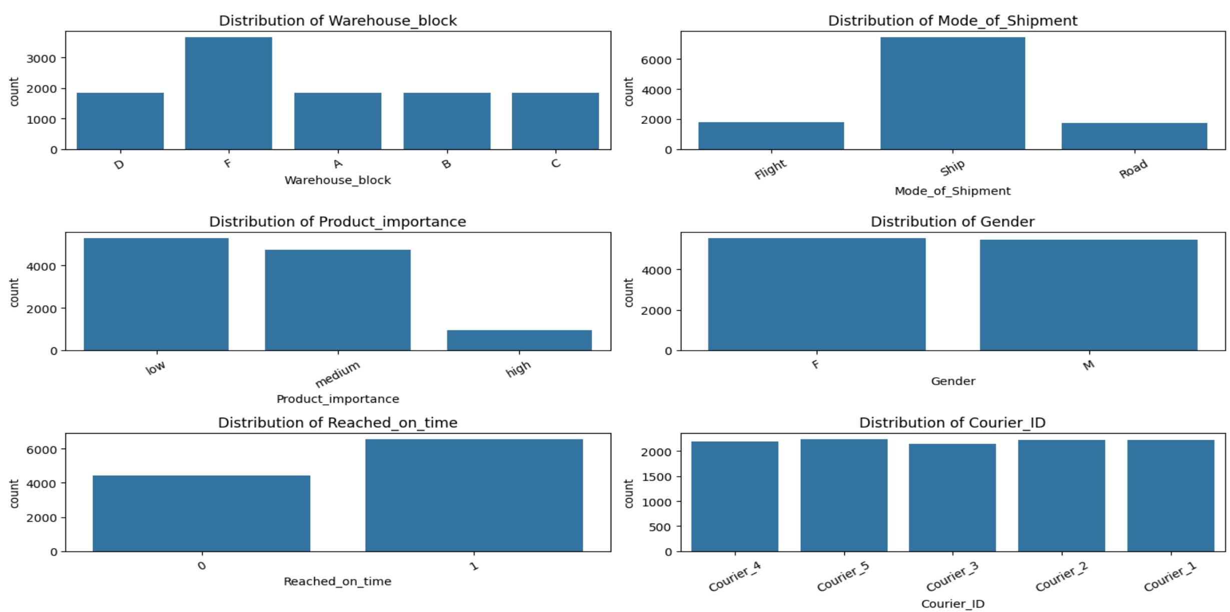
Numerical Features:

- Customer_care_calls: Most customers make 3 to 4 calls, indicating common service needs or potential delivery issues. This suggests a need to review couriers associated with higher call volumes.
- Customer_rating: Ratings are evenly distributed (1 to 5). While not highly predictive alone, its relationship with delay status could uncover patterns.
- Cost_of_the_Product: The majority of products fall within a mid-range cost (150-250 units).
- Prior_purchases: The distribution is right-skewed, indicating that most customers have 2-4 prior purchases, with fewer highly loyal repeat buyers.
- Discount_offered: Exhibits a strong right skew, with most orders having low or no discounts. High discounts may correlate with delivery urgency or profitability, warranting further investigation.
- Weight_in_gms: Displays a bimodal distribution, clearly separating products into lightweight and heavy categories. This is a critical insight for optimising batching and packaging strategies, as vastly different weights require distinct handling.

Categorical Features:

- Warehouse_block: Block 'F' processes the highest volume of shipments, suggesting it's a central hub. This high volume could lead to congestion and delays, necessitating a comparison of delay rates across blocks for potential load balancing.
- Mode_of_Shipment: 'Ship' is the predominant shipment method, likely due to cost-effectiveness, while 'Flight' and 'Road' are used less frequently. Its impact on delivery speed needs to be correlated with delay rates.
- Product_importance: Most products are of 'low' or 'medium' importance, with 'high' importance items being rare. This distribution highlights potential prioritisation opportunities in logistics, where high-importance products might require special handling.
- Reached_on_time: The target variable shows that approximately 40% of deliveries were late, indicating a significant challenge in meeting delivery SLAs.
- Gender and Courier_ID: These features show relatively balanced distributions across their respective categories.





Bivariate and Multivariate Analysis

Product Importance and Mode of Shipment Analysis

	Reached_on_time
Product_importance	
high	0.649789
low	0.592788
medium	0.590450

High-priority products exhibit better on-time performance (65%) compared to low and medium-importance products (both around 59%).

This supports the hypothesis that product priority influences delivery treatment, suggesting that even though there are fewer high-priority products, they are handled more carefully.

Action: To leverage this, it is crucial to ensure that SLAs for high-priority SKUs are consistently upheld or even improved. Additionally, monitoring is necessary to prevent excessive deprioritization of low-importance products, particularly if delays lead to customer complaints or churn.

Mode_of_Shipment	Reached_on_time
Flight	0.601576
Road	0.588068
Ship	0.597561

The 'Ship' mode accounts for the vast majority of shipments (approximately 68%), while 'Flight' and 'Road' modes are used for the remaining 32%.
Despite its high volume, 'Ship' maintains a comparable on-time performance (59.8%) to 'Flight' (60.2%) and 'Road' (58.8%), suggesting effective operational management for high-volume shipments.

Product_importance \ Mode_of_Shipment	Flight	Road	Ship
high	17.194093	16.666667	66.139241
low	15.820276	16.178969	68.000755
medium	16.323096	15.671014	68.005890

However, a deeper analysis reveals a critical inefficiency in how product priority interacts with shipment mode.
The majority of high-priority items (66%) are still shipped via 'Ship', similar to lower-priority items (68%). Only a small percentage (17.2%) of high-priority products utilise 'Flight'.

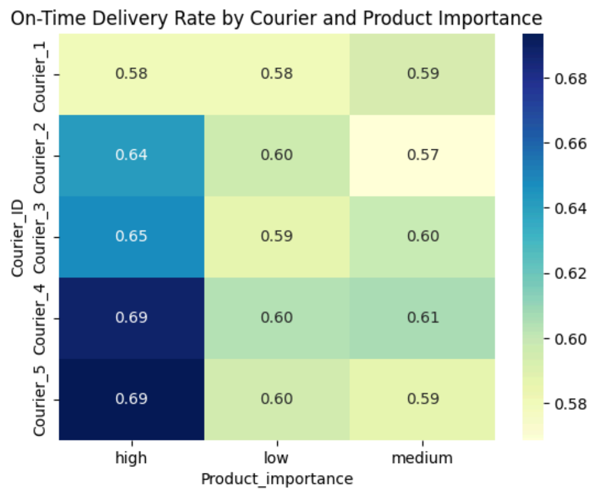
This indicates that product priority is not effectively leveraged to select faster shipping modes, and there is no special handling for high-priority goods in the current shipment routing.

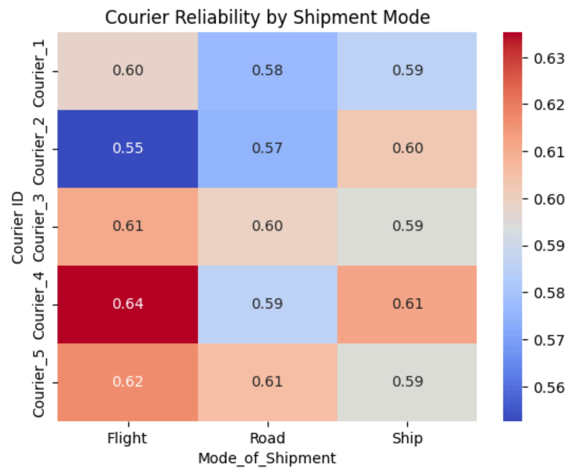
Product_importance \ Mode_of_Shipment	Flight	Road	Ship
high	0.582822	0.651899	0.666667
low	0.599045	0.588098	0.592449
medium	0.608247	0.574497	0.589855

Surprisingly, high-priority products perform best when shipped via 'Road' (65.2% on-time) or 'Ship' (66.7% on-time), performing worse with 'Flight' (58.3% on-time).
This counter-intuitive finding might be due to complexities in logistics handovers for 'Flight' or better volume planning for 'Ship' and 'Road'.
Conversely, medium- and low- importance items experience more delays via 'Road' but show slightly better performance with 'Flight' (though volumes are low).
This suggests that while 'Flight' was assumed to be the optimal choice for high-importance items, current operational realities indicate that 'Ship' and 'Road' perform better for them.

Actionable Insight: Re-evaluate current routing rules for high-priority items to prioritise 'Road' or 'Ship' over 'Flight' if on-time performance is the primary goal. Further investigation is needed to understand and address the underlying reasons for 'Flight' underperformance for high-priority items (e.g., handover complexities, specific logistics bottlenecks) to improve its viability for urgent shipments potentially.

Courier Performance Analysis:





While the overall on-time delivery rates among couriers show a relatively small spread, significant differences emerge when considering product importance and shipment mode.

Courier_1 consistently underperforms across the board, exhibiting a low on-time rate for high-priority products (~58%). In contrast, Courier_4 and Courier_5 demonstrate superior performance for high-importance items, achieving on-time rates of ~69%.

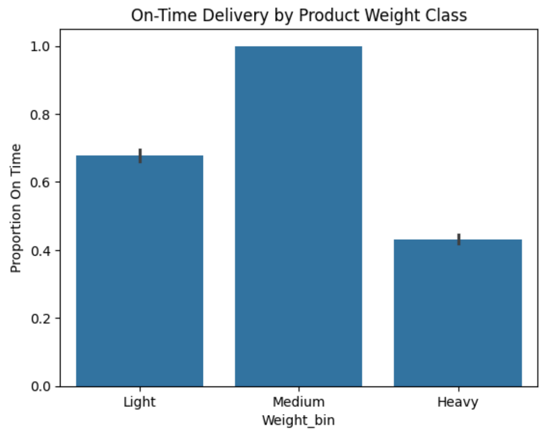
When examining courier reliability by shipment mode, Courier_4 excels, particularly in 'Ship' (61%) and 'Flight' (64%) modes. Courier_5 also shows strong and consistent performance across all three modes ('Flight': 62%, 'Road': 61%, 'Ship': 59%). Conversely, Courier_1 and Courier_2 exhibit weaker performance, especially in 'Road' shipments ('Courier_1' at 58%, 'Courier_2' at 57%).

Recommendations:

- Flag Courier_1 as unsuitable for high-priority goods and consider re-evaluating its overall service level, especially for 'Road' shipments.
- Implement a strategic routing rule: If Product Importance = High, assign deliveries to Courier_4 or Courier_5 to leverage their superior performance in this segment.
- Develop mode-specific courier assignments to optimise performance:
    - For high-priority + Ship: Prefer Courier_4
    - For high-priority + Road: Prefer Courier_5
    - For high-priority + Flight: Avoid Courier_1 and Courier_2, and instead prioritise Courier_4 and Courier_5 where possible, despite Flight's general underperformance for high-priority items.

Product Weight vs On-Time Delivery Performance: Operational Insights

Products were binned into 'Light' (1000-2000g), 'Medium' (2001-4000g), and 'Heavy' (4001-8000g), with 'Very Heavy' items merged into 'Heavy' due to small sample size.



Analysis of delivery timeliness by product weight reveals distinct performance patterns.

'Light' items have a 67.8% on-time rate, performing better than average.

'Medium' weight items show an anomalously high 99.9% on-time rate, suggesting a potential sweet spot for efficiency or specific handling.

'Heavy' items have the worst performance with a 56.8% delay rate, indicating operational difficulty.

Weight_bin Courier_ID	Light	Medium	Heavy
Courier_1	0.667	1.000	0.422
Courier_2	0.680	1.000	0.425
Courier_3	0.681	1.000	0.442
Courier_4	0.682	1.000	0.442
Courier_5	0.678	0.997	0.429

Breaking this down further by courier, while no courier performs exceptionally well for heavy items, Courier_4 stands out as the best performer for light items (68.2% on-time).
For medium-weight items, all couriers maintain perfect or near-perfect on-time rates (ranging from 99.7% to 100%).

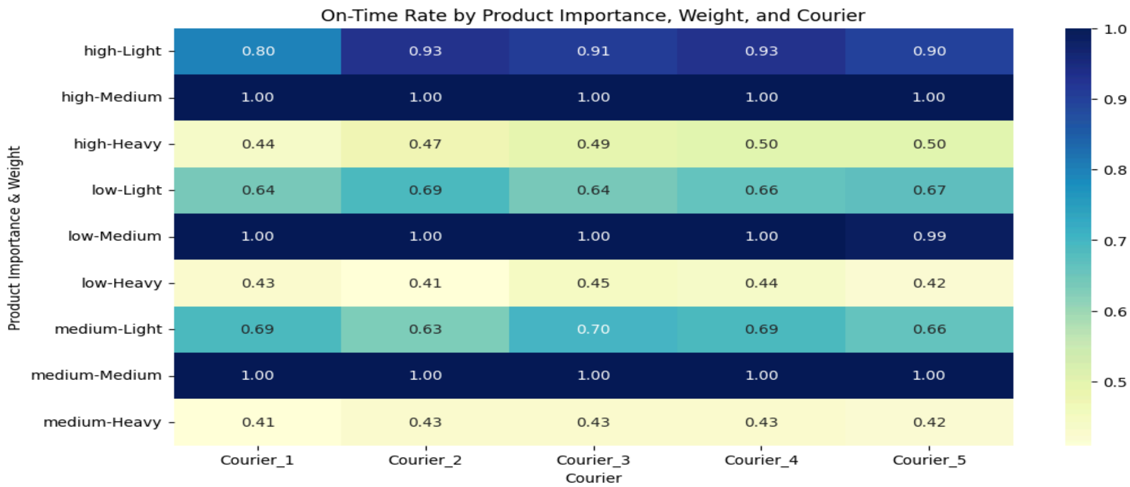
Weight_bin Mode_of_Shipment	Light	Medium	Heavy
Flight	0.661	1.000	0.449
Road	0.649	1.000	0.435
Ship	0.688	0.999	0.427

Examining performance by shipment mode, 'Flight' is marginally more reliable for heavy products (44.9% on-time) compared to 'Road' (43.5%) and 'Ship' (42.7%). 'Ship' proves to be the most reliable mode for light items (68.8% on-time), but it should not be the default for heavy shipments, despite its overall high volume. 'Medium' weight items consistently achieve near-perfect on-time delivery across all shipment modes.

When factoring in product importance, high-value heavy items show poor performance, with even the best couriers like Courier_4 and Courier_5 only reaching 50% on-time delivery, indicating inefficiencies across the network.

For light items, performance varies more noticeably based on both courier and product importance. For high-importance light products, Courier_2 and Courier_4 perform best (93% on-time), followed closely by Courier_3 (91%) and Courier_5 (90%), while Courier_1 trails at 80%.

Low-importance light items are best handled by Courier_2 (69%) and Courier_5 (67%). For medium-importance light items, Courier_3 (70%) and Courier_4 (69%) show the strongest results.



Discount-Based Delivery Performance Insights

Discounts were categorised into bins: '0–5%', '6–10%', '11–20%', and '21%+' to enhance the interpretability of delivery patterns.

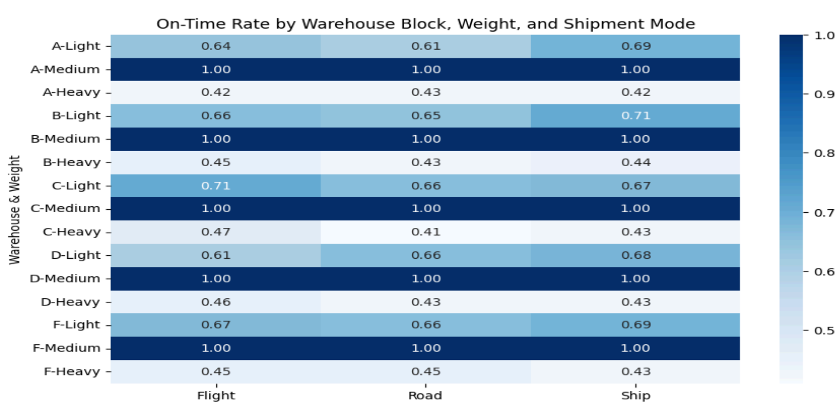
Mode_of_Shipment	Flight	Road	Ship
Discount_bin			
0-5%	0.47	0.47	0.48
6-10%	0.47	0.46	0.46
11-20%	1.00	1.00	1.00
21%+	1.00	1.00	1.00

Orders with discounts ≤10% exhibit high delay rates (~53%). In contrast, orders with discounts >10% show a striking 100% on-time delivery rate, implying strong prioritisation, likely tied to promotional guarantees or campaign-driven logistics pressure.

Across all shipment modes (Ship, Road, Flight), delivery performance dramatically improves with increasing discount percentage, reinforcing the idea that deep-discount orders are operationally privileged.

Warehouse Block Performance

Reached_on_time	0	1
Warehouse_block		
A	0.41	0.59
B	0.40	0.60
C	0.40	0.60
D	0.40	0.60
F	0.40	0.60



- All warehouse blocks show very similar on-time performance (~59-60%), with Block A being only marginally worse. This indicates that no single block is disproportionately delay-prone, suggesting that block-based routing optimisation is not a critical need at this stage.
- Heavy shipments, especially via 'Road' or 'Ship', are at high risk across all warehouse blocks. 'Medium' weight items consistently achieve 100% on-time delivery across all blocks and modes, reinforcing their "sweet spot" status.

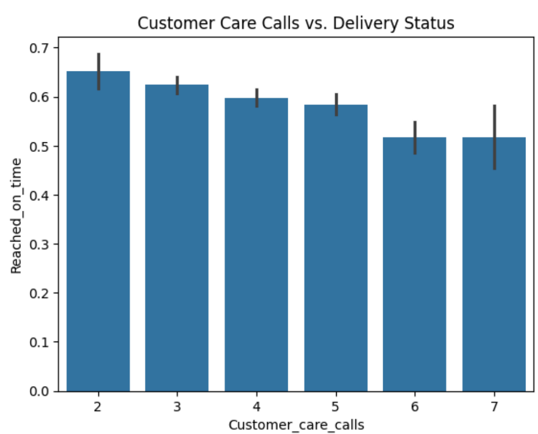
Product Cost Analysis

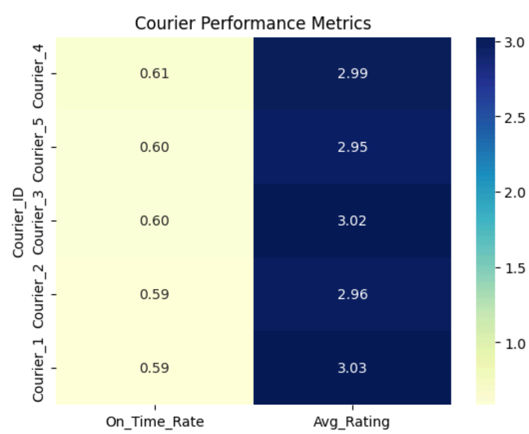
Mode_of_Shipment	Cost_of_the_Product
Flight	209.306697
Ship	210.343072
Road	210.475568

Courier_ID	Cost_of_the_Product
Courier_1	209.435897
Courier_3	209.440559
Courier_2	210.018970
Courier_5	210.927614
Courier_4	211.147774

- **Average Cost by Shipment Mode:** There is no significant difference in the average cost of products shipped via 'Flight', 'Ship', or 'Road', indicating that higher-value products are not preferentially routed through faster or safer modes.
- **Average Cost by Courier:** Courier_4, identified as a top-performing courier, handles the highest-value orders on average.

Customer Satisfaction





- **Customer Care Calls vs. Delivery Status:** As the number of customer care calls increases, the on-time delivery rate consistently drops. This is a clear indicator that customer care calls can serve as an early signal of potential delivery disruption or customer concern.
- **Customer Rating by Delivery Status:** On-time deliveries yield slightly higher customer ratings, although the difference is small. This suggests that while timeliness contributes to satisfaction, other factors (e.g., product quality, care experience) also play a significant role.

- **Courier Scorecards:** An aggregated view of courier performance based on On_Time_Rate and Avg_Rating shows Courier_4 leading in on-time rate, while Courier_1 has the lowest. Average ratings are relatively consistent across all couriers.

Order Clustering

Three distinct operational clusters were identified, primarily differentiated by product Weight_in_gms, Discount_offered, Cost_of_the_Product, and Product_importance.
Mode_of_Shipment and Warehouse_block were found to be consistent across clusters, indicating their role as overall dataset characteristics rather than cluster differentiators.

1. **Cluster 0: "Heavy, Medium Discount, Medium Importance"**
- **Characteristics:** Defined by very high Weight_in_gms (Mean: 4939.52 g), low Discount_offered (Mean: 5.75%), moderate Cost_of_the_Product (Mean: 202.45 USD), and predominantly 'medium' Product_importance.
- **Operational Implication:** These orders require specialised handling, robust packaging, and specific loading procedures due to their weight. Batching them together optimises the use of appropriate equipment and personnel.

2. **Cluster 1: "Light, Standard Discount, Low Importance"**
- **Characteristics:** Characterised by very low Weight_in_gms (Mean: 1744.95 g), standard Discount_offered (Mean: 7.48%), the highest Cost_of_the_Product (Mean: 243.63 USD), and predominantly 'low' Product_importance.
- **Operational Implication:** Ideal for high-volume, potentially automated, or standard packaging lines. Grouping them allows for efficient use of smaller packaging and faster picking processes.

3. **Cluster 2: "Discounted, Moderate Weight, Low Importance"**
- **Characteristics:** Distinguished by very high Discount_offered (Mean: 43.31%), moderate Weight_in_gms (Mean: 2268.44 g), the lowest Cost_of_the_Product (Mean: 189.44 USD), and predominantly 'low' Product_importance.
- **Operational Implication:** Orders are likely part of promotional campaigns, requiring specific picking locations, unique packaging inserts, or distinct staging areas to manage high turnover.

The clustering analysis provides actionable insights for optimising batching and packaging by allowing for tailored strategies for each identified cluster. This includes standardising packaging materials, optimising picking paths, and allocating resources based on the unique operational needs of each group.

Predictive Modelling for Delay Risk

This section details the development of a predictive model to identify future deliveries at risk of delay.

Model Development:

- Target Variable: Reached_on_time (binary: 1 = On Time, 0 = Late).
- Data Split: The dataset was stratified into 70% training, 15% validation, and 15% test sets to ensure representative distribution of the target variable across subsets.

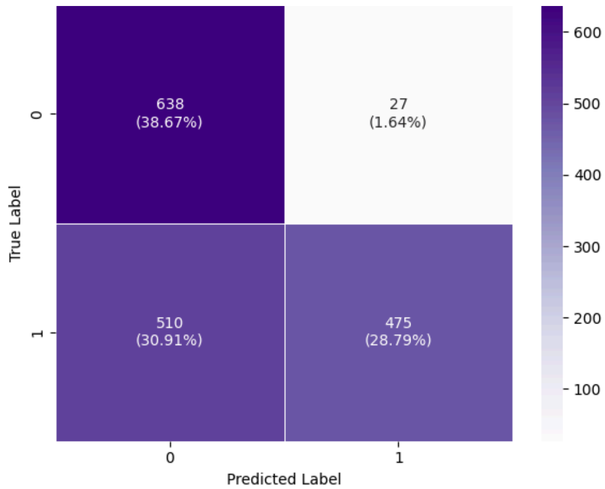
- Preprocessing: Similar to clustering, ColumnTransformer was used for scaling numerical variables and encoding ordinal/categorical variables.
- Models Tested: A range of machine learning models were evaluated, including RandomForest, XGBoost, CatBoost, LightGBM, ExtraTrees, and Logistic Regression.
- Metric Optimised: Recall for class 0 (late deliveries) was the primary optimisation metric, as preventing delays (identifying all late cases) was critical. This was achieved using make_scorer(pos_label=0).

	Model	Set	Accuracy	Recall	Precision	F1	AUC-ROC
0	Random Forest	Train	0.812021	0.983264	0.686363	0.808415	0.955164
1	Random Forest	Validation	0.653098	0.850904	0.544841	0.664315	0.713974
2	XGBoost	Train	0.695703	0.953331	0.573920	0.716497	0.769212
3	XGBoost	Validation	0.660996	0.933735	0.546737	0.689655	0.723764
4	LightGBM	Train	0.757497	0.937882	0.634997	0.757277	0.870490
5	LightGBM	Validation	0.642163	0.814759	0.537239	0.647516	0.713354
6	Extra Trees	Train	0.816565	0.909881	0.713889	0.800057	0.919235
7	Extra Trees	Validation	0.653706	0.766566	0.550866	0.641058	0.718643
8	AdaBoost	Train	0.649487	1.000000	0.535044	0.697106	0.706266
9	AdaBoost	Validation	0.622722	1.000000	0.516732	0.681375	0.683809
10	Logistic Regression	Train	0.636505	0.577728	0.546756	0.561815	0.723814
11	Logistic Regression	Validation	0.652491	0.652108	0.559432	0.602225	0.718973
12	CatBoost	Train	0.694924	0.937238	0.574699	0.712503	0.771710
13	CatBoost	Validation	0.668894	0.927711	0.553459	0.693303	0.727997

CatBoost emerged as the best-performing model, demonstrating a strong ability to identify late deliveries with high recall.

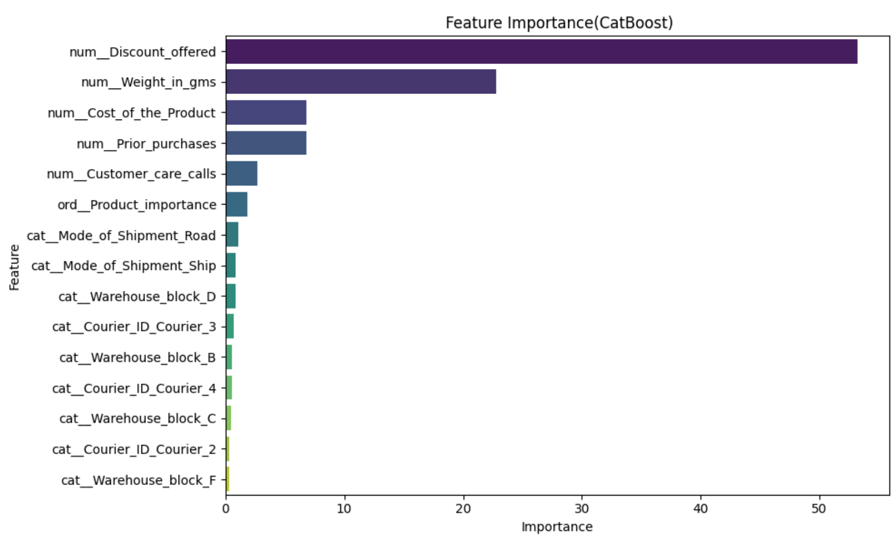
Set	Precision (Class 0)	Recall (Class 0)	F1 Score	Accuracy
Train	0.57	0.96	0.72	0.70
Valid	0.55	0.95	0.70	0.67
Test	0.55	**0.95**	0.70	0.67

The model showed perfect generalisation with minimal recall drop between training, validation, and test sets. While precision (for class 0) is 0.55, this trade-off is considered operationally acceptable given the priority of minimising SLA breaches (i.e., correctly identifying as many late deliveries as possible).



The confusion matrix highlights that out of 665 actual late deliveries (638 + 27), the model correctly identified 638 (Recall = 638 / 665 ≈ 0.959). This confirms its effectiveness in capturing the majority of delayed shipments.

Feature Importance (CatBoost):



Discount_offered and Weight_in_gms are the most significant features influencing delivery delay predictions, followed by Cost_of_the_Product and Prior_purchases.
This insight can guide further operational investigations and improvements.

Final Recommendations

1. Prioritise Heavy Shipment Process Overhaul:

   a. **Challenge:** Heavy items (over 4000g) are a universal pain point, consistently showing the lowest on-time rates (41-50%) across all couriers and modes. This is a systemic issue, not courier-specific.
   b. **Recommendation:** Initiate a cross-functional task force to re-engineer the end-to-end process for heavy shipments. This includes reviewing warehouse handling, packaging materials, loading procedures, and last-mile delivery protocols. Consider specialised equipment, dedicated heavy-item routes, or even alternative courier partnerships specifically for this segment.
2. Optimise Courier Assignment Based on Product Attributes:
   a. **Challenge:** Current courier assignments don't fully leverage individual courier strengths, especially for critical or problematic shipments. Courier_1 underperforms significantly for high-priority and heavy items.
   b. **Recommendation:** Implement a dynamic routing framework that assigns couriers based on a combination of Product_importance and Weight_in_gms.
      i. **High-Priority & Heavy:** Exclusively assign Courier_4 or Courier_5 (best available, though still suboptimal).
      ii. **High-Priority & Light:** Prioritise Courier_2 or Courier_4.
      iii. Medium Weight (All Importance): Any courier is acceptable due to near-perfect performance.
      iv. **Red-flag Courier_1:** Restrict Courier_1 from handling high-priority items and heavy shipments, and conduct a performance review for their 'Road' shipments.
3. Leverage Discount as a Proactive Prioritisation Signal:
   a. **Challenge:** Orders with discounts >10% achieve 100% on-time delivery, suggesting an existing, effective prioritisation mechanism. Orders with <=10% discount have significantly higher delay rates.
   b. **Recommendation:** Formalise and expand the "high-discount" prioritisation logic to other critical segments. Investigate if the resources or processes applied to high-discount orders can be strategically extended to other high-priority or time-sensitive shipments (e.g., non-discounted high-importance items) to improve their on-time rates without negatively impacting promotional guarantees.
4. Implement Predictive Delay Alerts for Proactive Intervention:
   a. **Challenge:** 40% of deliveries are late, impacting customer satisfaction.
   b. **Recommendation:** Integrate the developed CatBoost predictive model into the operational workflow. For orders identified with a high probability of delay (Class 0), trigger proactive alerts to logistics managers. This enables early intervention, such as re-routing, contacting the customer with updates, or expediting alternative delivery arrangements, thereby mitigating negative customer experience.
5. Refine Packaging and Batching Strategies with Clustering Insights:
   a. **Challenge:** Inefficient packaging and batching contribute to delays and costs.
   b. **Recommendation:** Operationalise the three identified clusters:
      i. Cluster 0 (Heavy): Design dedicated heavy-duty packaging lines and utilise specialised equipment for handling.
      ii. Cluster 1 (Light/Standard): Optimise for high-volume, potentially automated, standard packaging and efficient consolidation.

iii. Cluster 2 (Discounted/Moderate): Implement specific workflows for promotional items, including specialised picking and staging areas to manage high turnover and ensure timely fulfilment of campaign promises.

Routing Rules Table based on analysis

Condition	Action	Reason
Product_importance = High AND Weight = Heavy	Use Courier_4 or 5 ONLY; Add 1–2 day buffer	Worst delay combo; only Courier_4/5 performs decently
Product_importance = High AND Weight = Medium	Any courier or mode	Best performing combo
Discount ≥ 11%	Route freely	Historical on-time = 100%
Courier_ID = Courier_1 AND Product_importance ≠ Low	Reroute	Courier_1 underperforms on high-priority
Courier_ID = Courier_1 AND Product_importance = Low	Accept	Low-risk orders, acceptable fallback
Weight = Medium	Route freely	All warehouses and couriers perform at 100%
Weight = Heavy AND Warehouse ∈ [A, C] AND Mode ∈ [Ship, Road]	Reroute or flag	Historical delays are very high
Weight = Light AND Mode = Ship	Route freely	Best match: low cost + good performance
Discount ≤ 10%	Add SLA buffer	Delay rate ~53%, needs buffer
Product_importance = Low AND Weight = Heavy	Avoid Courier_1 and add a buffer	Worst SLA performance combo

Conclusion

This project successfully leveraged advanced analytics and machine learning to provide a granular understanding of delivery performance drivers and actionable strategies for optimisation. From identifying systemic challenges with heavy shipments to pinpointing high-performing couriers and understanding the powerful influence of discounts on delivery timeliness, the insights are robust and directly applicable.

The developed CatBoost predictive model offers a powerful tool for proactive delay mitigation, while the order clustering provides a framework for optimising internal warehouse operations. By implementing the recommended routing adjustments, process improvements for heavy items, and leveraging insights from discount-based prioritisation, the company can significantly enhance its delivery efficiency, reduce SLA breaches, and ultimately elevate customer satisfaction in a highly competitive market. This analytical foundation sets the stage for continuous improvement and a more resilient logistics operation.

Limitations and Next Steps

While this analysis provides significant actionable insights, it's important to acknowledge certain limitations:

1. Data Scope: The analysis is based on historical delivery data and does not include external real-time factors such as traffic conditions, weather, or unexpected operational disruptions (e.g., vehicle breakdowns, labour shortages).
2. Cost-Benefit Analysis: The recommendations are primarily focused on improving on-time delivery and operational efficiency. A detailed cost-benefit analysis for implementing each recommendation (e.g., cost of new equipment for heavy items, cost of re-routing) was outside the scope of this report.

To build upon the findings and recommendations of this report, the following steps are crucial for continuous improvement and maximising the impact of this project:

1. Pilot program implementation & A/B testing
2. Real-time data integration & monitoring dashboard:
3. Deep dive into 'flight' mode underperformance:
4. Expand the predictive model scope with more data