Predicting Delivery Delays with ML

created by: Novia A. A.

#about-me

Novia Anggita Aprilianti

I am a data analyst with a background in statistics and a strong curiosity about how data can be used to solve real-world problems. I enjoy exploring data, uncovering relevant insights, and simplifying complex matters to make them easier to understand and act upon.

Throughout my career, I’ve believed that good decisions come from data that is truly understood—not just reported. That’s why I’m always open to learning new things and enjoy collaborating across teams and disciplines.

I began exploring the world of data during my university studies and have since strengthened my skills through hands-on experience in various projects. For over three years, I’ve been involved in a wide range of initiatives, from building interactive dashboards for operational and business needs to working on machine learning projects such as regression, classification, and natural language processing (NLP) to support data-driven decision-making.

Let's get started!

Table of Contents

Business Understanding

An e-commerce company aims to analyze its shipping data **to predict whether an upcoming delivery is likely to be delayed or not**. Several variables that are believed to influence shipping performance have been collected. The company also intends **to evaluate the extent to which each variable contributes to the prediction of delivery delays**. The most influential variables will become the focus of relevant teams in order to proactively mitigate delays in the future.

There are two main objectives of this analysis:

- To assess how much influence each variable has on the delivery status.
- To build a predictive model to determine whether a shipment is likely to experience a delay.

Data Overview

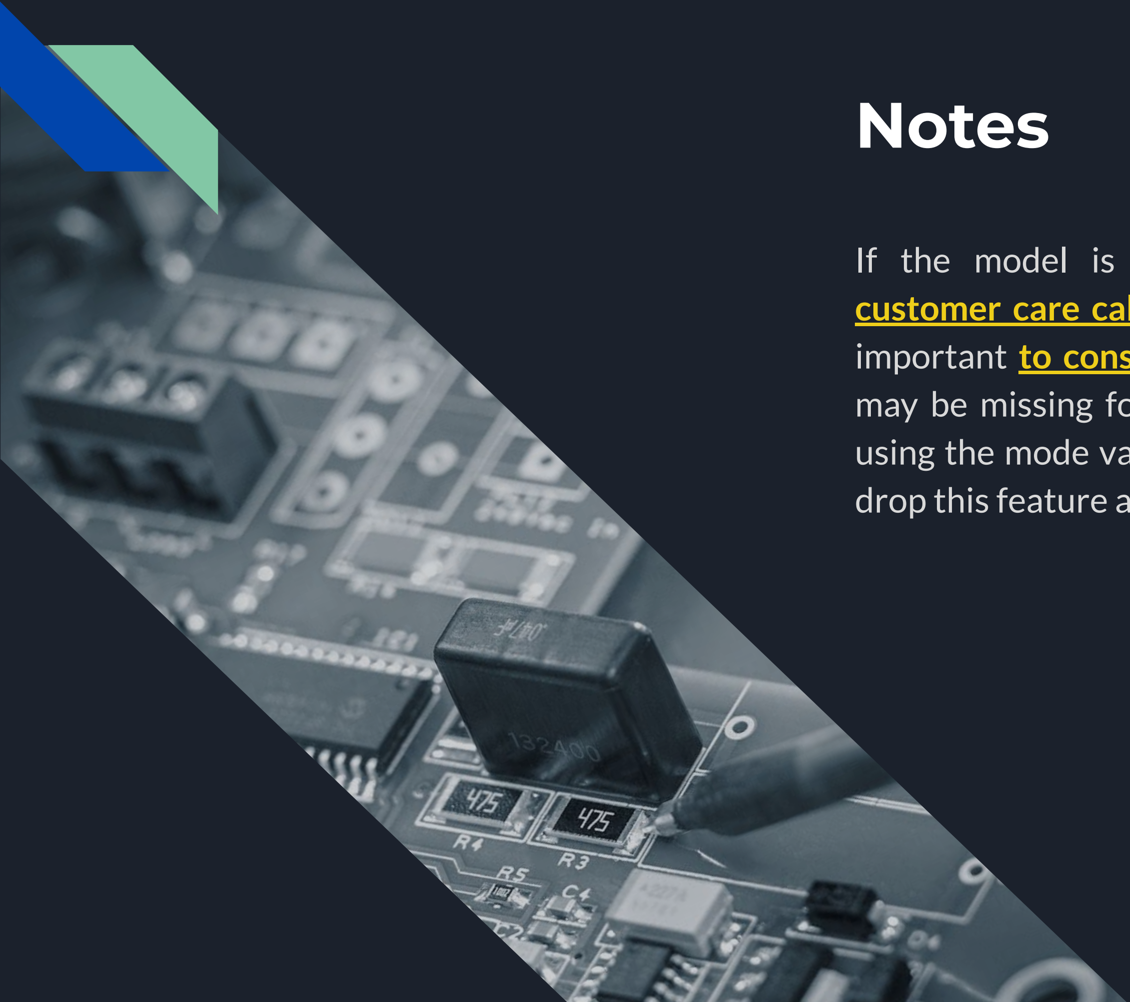
ID	Warehouse_ block	Mode_of_ Shipment	Customer_ care_calls	Customer_ rating	Cost_of_the_ Product	Prior_ purchases	Product_ importance	Gender	Discount_ offered	Weight_in_ gms	Reached.on. Time_Y.N
1	D	Flight	4	2	177	3	low	F	44	1233	0
2	F	Flight	4	5	216	2	low	M	59	3088	0
3	A	Flight	2	2	183	4	low	M	48	3374	0
4	B	Flight	3	3	176	4	medium	M	10	1177	0
5	C	Flight	2	2	184	3	medium	F	46	2484	0

Features:

- Warehouse block: The Company have big Warehouse which is divided in to block such as A,B,C,D,E.
- Mode of shipment:The Company Ships the products in multiple way such as Ship, Flight and Road.
- Customer care calls: The number of calls made from enquiry for enquiry of the shipment.
- Customer rating: The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).
- Cost of the product: Cost of the Product in US Dollars.
- Prior purchases: The Number of Prior Purchase.
- Product importance: The company has categorized the product in the various parameter such as low, medium, high.
- Gender: Male and Female.
- Discount offered: Discount offered on that specific product.
- Weight in gms: It is the weight in grams.
- Reached on time: It is the target variable, where 1 Indicates that the product has reached on time and 0 indicates it has NOT reached on time.
- cost_after_discount: (Cost of the product) * (Discount offered)

Notes

If the model is **intended for future prediction, the customer care calls feature should be dropped**. It's also important **to consider the customer rating feature**, as it may be missing for new customers. This can be imputed using the mode value, but it is generally recommended to drop this feature as well.



Workflow

The method used is a classification technique involving several algorithms. The best-performing method will be selected based on predefined evaluation metrics.

- The methods include: Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM.
- The evaluation metric used is the F1 score, as the company considers it important to minimize all types of prediction errors, making the F1 score the most suitable choice.

Step 1

Preprocessing Data:

- Change column names
- Missing/invalid value check
- Data duplicated check
- Outlier check
- Categorical encoding
- Split data (train and test)

Step 2

Data Understanding & EDA

Step 3

Modeling using:

- Logreg
- Decision tree
- Random forest
- XGBoost
- LightGBM

Step 4

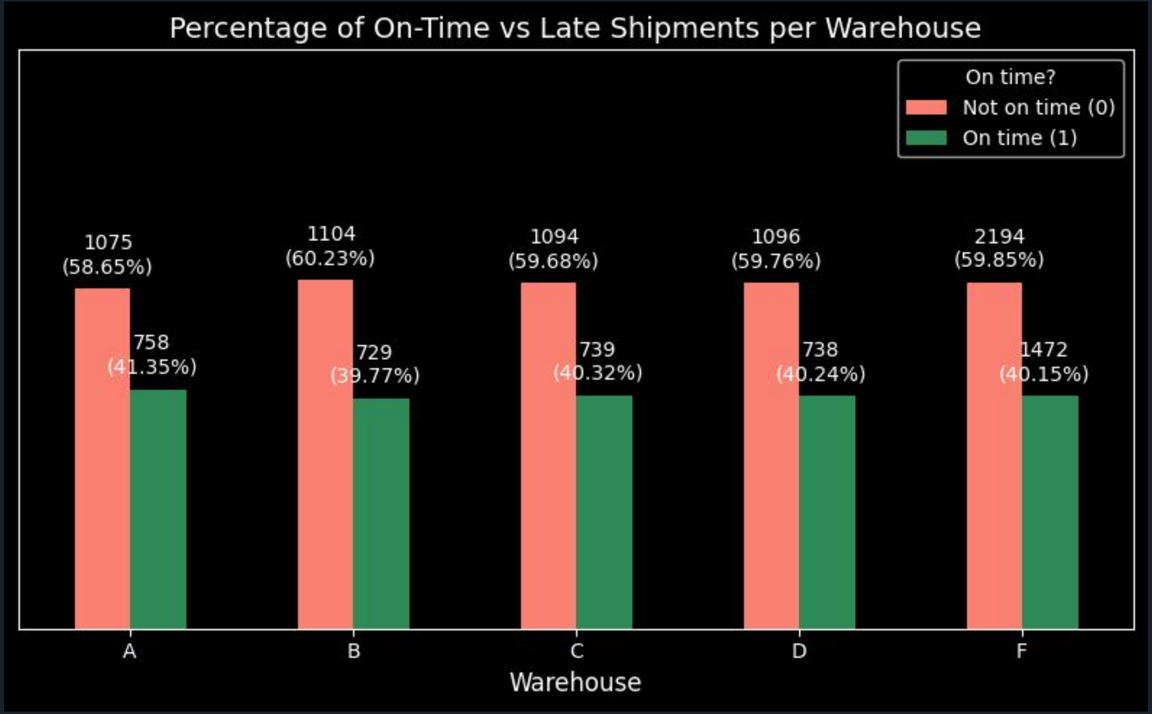
Evaluation model using F1 score. Choosing the best model based on the highest F1 score.

Step 5

Use the model for future shipping forecasting.

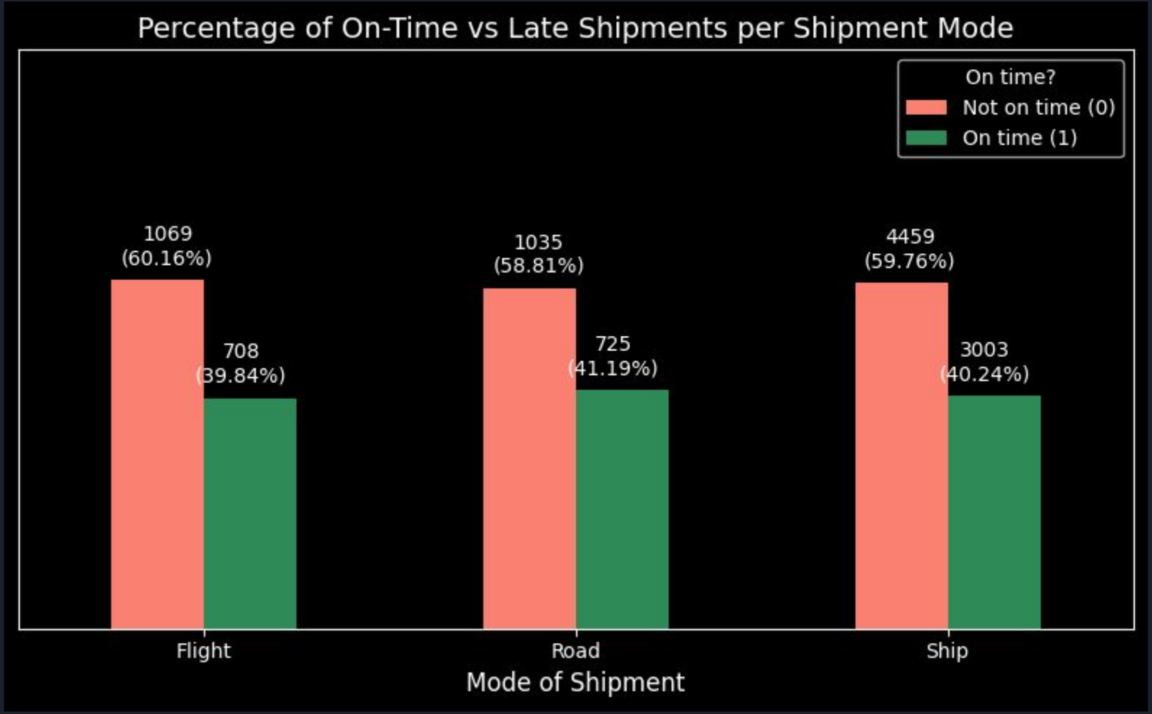
EDA 1:

Do certain warehouses tend to have more delays?



- All warehouses have a delay percentage of over 50%.
- Warehouse B has the highest delay percentage compared to the other warehouses.

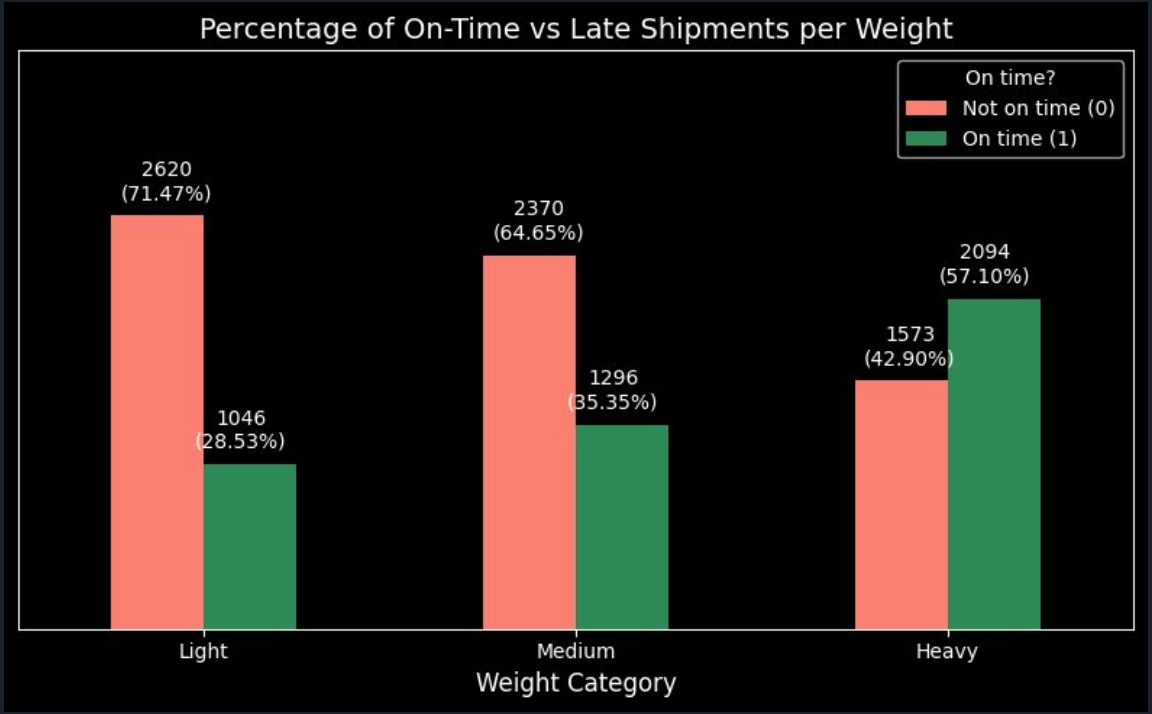
EDA 2:
Which shipment mode is most often on time?



- All shipment modes have a delay percentage of over 50%.
- Flight has the highest delay percentage compared to other shipment modes.
- Road has the highest on-time delivery percentage among all shipment modes.

EDA 3:

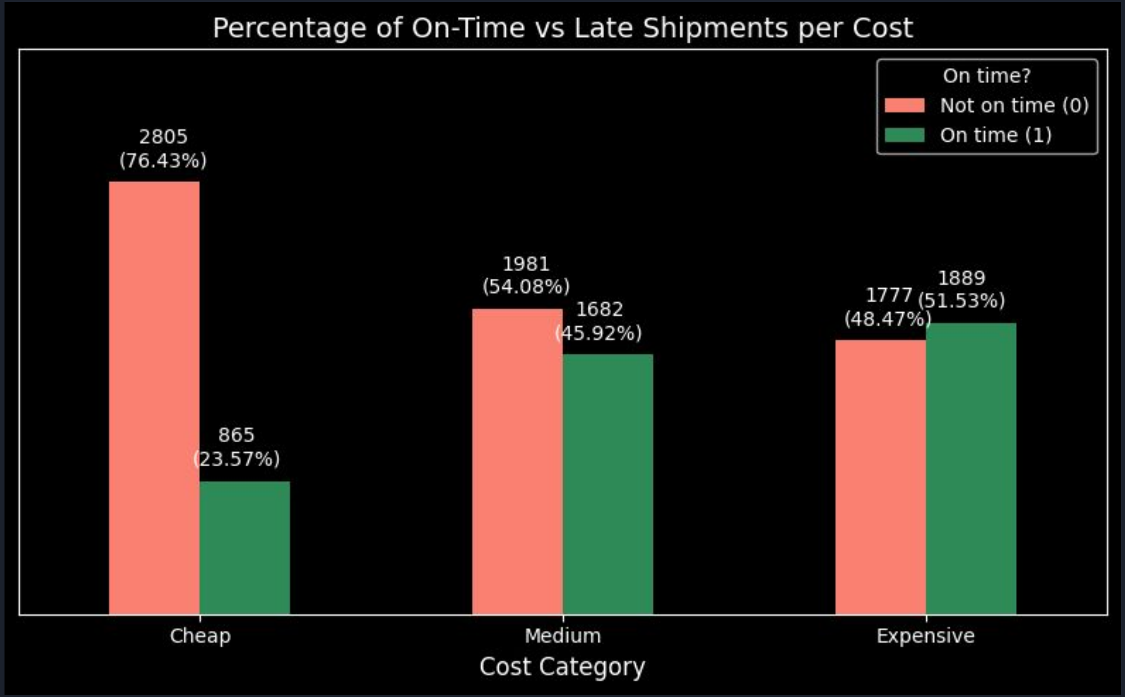
Do heavier products tend to be delayed?



- Products in the lightweight category tend to experience more delays.
- Products in the heavyweight category are generally delivered faster or more often on time.

EDA 4:

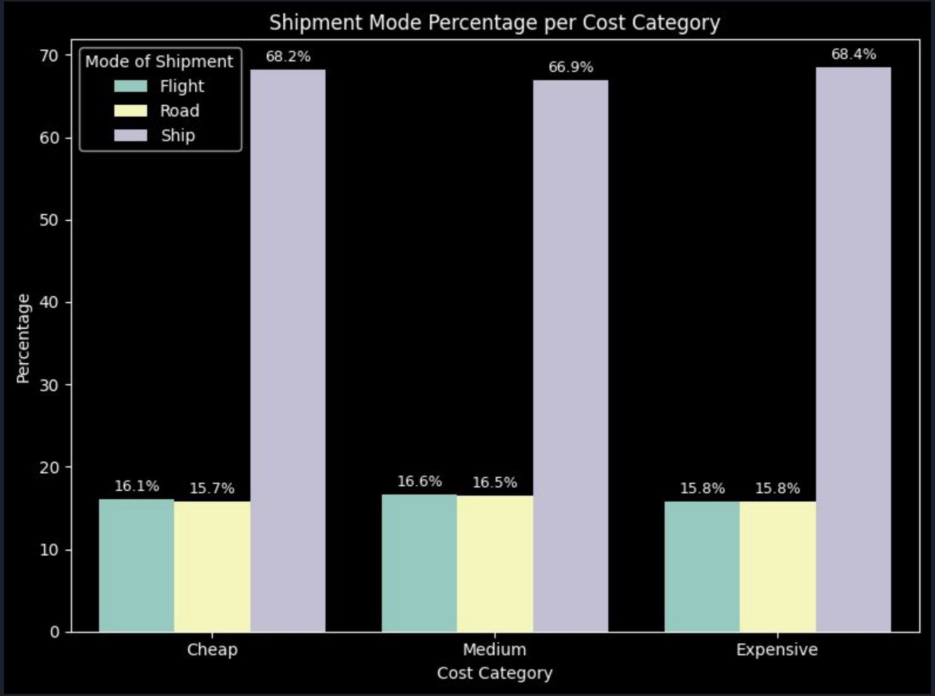
Do lower-priced products result in more shipments and more frequent delays?



- Low-priced products tend to experience more delays.
- What's important to highlight is that, although low-cost products do not lead to a significant increase in the number of shipments (all three cost categories have around 3,660 shipments), the delay percentage for low-cost products is relatively high compared to the other cost categories.
- Products in the medium-to-high price range tend to be delivered faster or more often on time.

EDA 4:

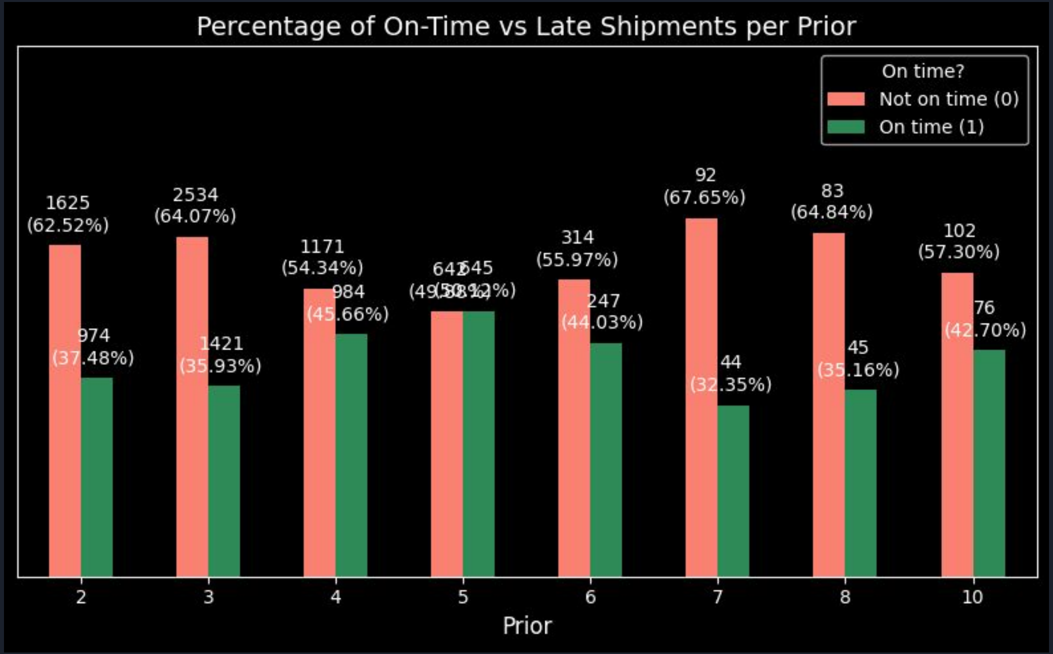
Do lower-priced products result in more shipments and more frequent delays?



- Another interesting finding is that all types of products—from cheap to expensive—follow the same shipment mode pattern. This means that low-priced items are not necessarily shipped via road or ship, which are typically slower.
- This indicates that shipment mode is not a key factor contributing to the higher delay rate observed in low-cost products.

EDA 5:

How is the pattern of delivery status between customers?



- In general, customers who frequently make shipments tend to not experience delivery delays.
- However, there is an anomaly where customers who have previously made shipments 7 to 8 times also often experience delays.

Model Selection

Model	F1 Score	
	Train set	Test set
Logistic regression	49.01%	49.50%
Decision tree	71.87%	71.35%
Random forest	73.79%	68.24%
XGBoost	71.14%	70.48%
LightGBM	70.97%	70.52%

- Although Random Forest achieved the highest F1 score on the training set, it experienced a significant drop in performance on the test set. This indicates a tendency toward overfitting, where the model performs well on training data but poorly on unseen data.
- Therefore, Decision Tree is considered the best model, as it demonstrates more stable performance and achieves the highest F1 score on both the training and test sets compared to other stable models.

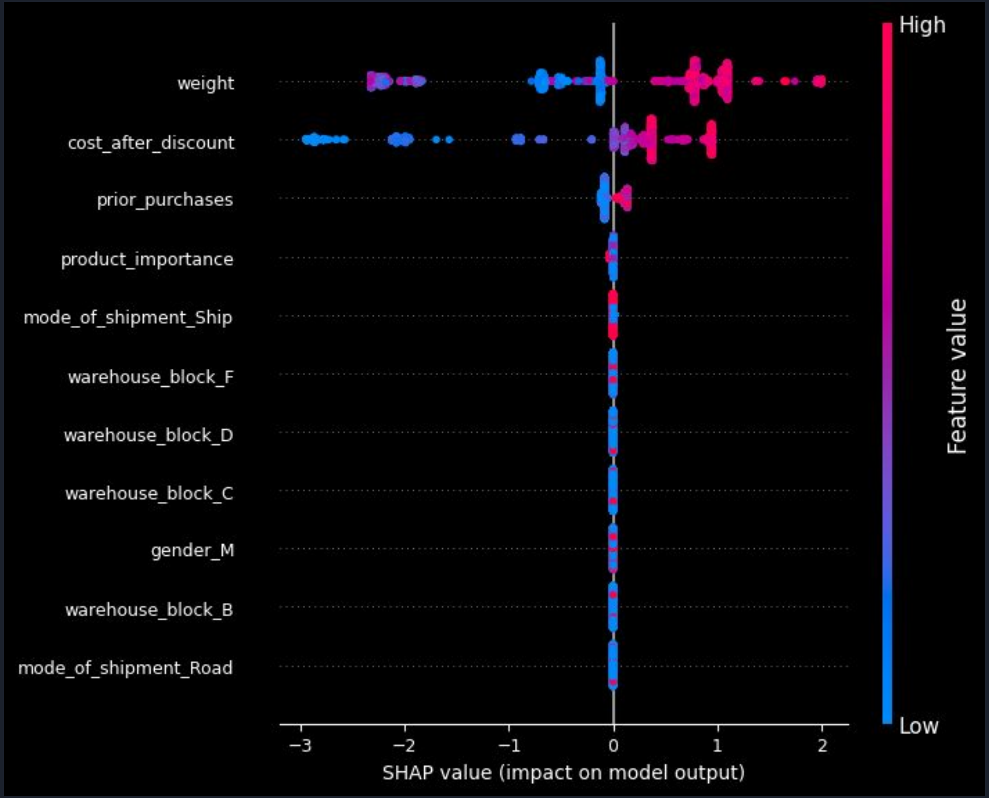
Features Contribution

Explanation:

- The more positive the SHAP value, the greater its contribution to the product arriving on time, and vice versa.
- The color indicates the magnitude of the value: the redder the color, the higher the value; the bluer the color, the lower the value.

There are three features that contribute most significantly to the delivery status (on-time or delayed):

1. Weight: the heavier the item, the faster it arrives.
2. Cost after discount: the higher the price of the item, the faster it arrives.
3. Prior purchases: the more frequently a customer shops, the faster their ordered items arrive.



Features Contribution

Feature	Contribution	Justification
Weight	The heavier the item, the faster it arrives	● Lightweight items tend **to be consolidated first** with other lightweight goods before shipping, so they require a longer waiting time. **Shipping them individually incurs a relatively low cost**, so they are often delayed. Even if demand is high, the items need to "accumulate first" before being shipped. ● Lightweight items are **typically scheduled for shipping every 2–3 days** when not urgent, often waiting for remaining volume (included in the last/non-priority shipments).
Cost	The more expensive the item, the faster it arrives	● **Company regulations** may prioritize the shipping of products based on price, where **higher-value items are considered more critical and thus prioritized**. ● Lower-priced items are shipped through **cheaper channels**, such as non-express land transport, combined with other orders (consolidated shipping). ● Lower-priced items are **often packaged with many other orders** for efficiency, which can result in waiting for a full batch and potentially cause delays.
Prior Purchase	The more frequent the purchases, the faster it arrives	● **Loyal customers may receive priority** in the picking, packing, and warehouse shipping process; they may also frequently receive discount vouchers for shipping. ● Due to regular transactions, **loyal customers may better understand shipping patterns**, such as avoiding known delay periods and knowing the ideal time to place an order, as well as which vendor or shipping method is the fastest.

Conclusion:

1. Based on the charts showing an overview of the data, several key points can be highlighted:
   - The data shows a **fairly even distribution across different modes of shipment**. This indicates that regardless of the shipment mode, the distribution of on-time and late deliveries is relatively similar.
   - **Products with higher weight or cost are actually more likely to arrive on time**, and the graph indicates that the heavier and more expensive the item, the higher the likelihood it will be delivered on time.
2. The **Decision Tree model was selected** as the best-performing model based on model stability and the F1 score metric.
3. There are **three features with the highest contribution** in predicting whether a delivery will be late or on time, namely: weight of product, cost of product (after discount), and prior purchase.

Recommendations:

1. **Add other relevant features** such as distance/travel time, destination location or store location, order time (hour/day/date/month), number of products, product category (electronics, clothing, food, etc.).
2. **Pay attention to smaller-sized and/or lower-priced products** to reduce delivery delays.
3. **Offer discounts only during specific events or occasions**, ensuring they do not coincide with periods of high shipping demand (such as weekends, holidays, etc.). If avoiding such periods is not possible, anticipate the surge in shipments by **preparing a larger delivery fleet** than on regular days.

Attachments

Google slide presentation

Google colab

Dataset

Thank You

created by: Novia A. A.