Fourcasters

Rakamin Academy

E-Commerce Delivery Prediction

Data Science Batch 35

Today's topics

01 **Project background**

02 **What really happened?**

03 **What can we do?**

Fourcasters

Fourcasters was created to “guide” people in the right directions using data-driven decisions.

Get to know the members!

Atthoriq P. P.
Project Lead

Alvida Dwiki C.
BI Analyst

Darell Timothy Tarigan
Data Analyst

Emir Akbar
Data Analyst

Nur Baiti L.
Data Analyst

Reny Rafiqah
BI Analyst

01

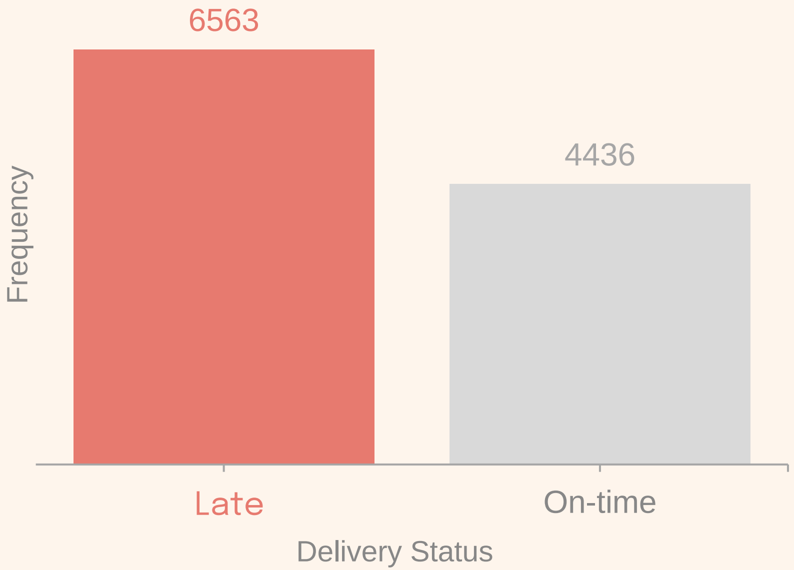
Project Background

Why do we started this project?

Initial Problem

Late Deliveries: 60% across all orders

From a total of 10.9K orders

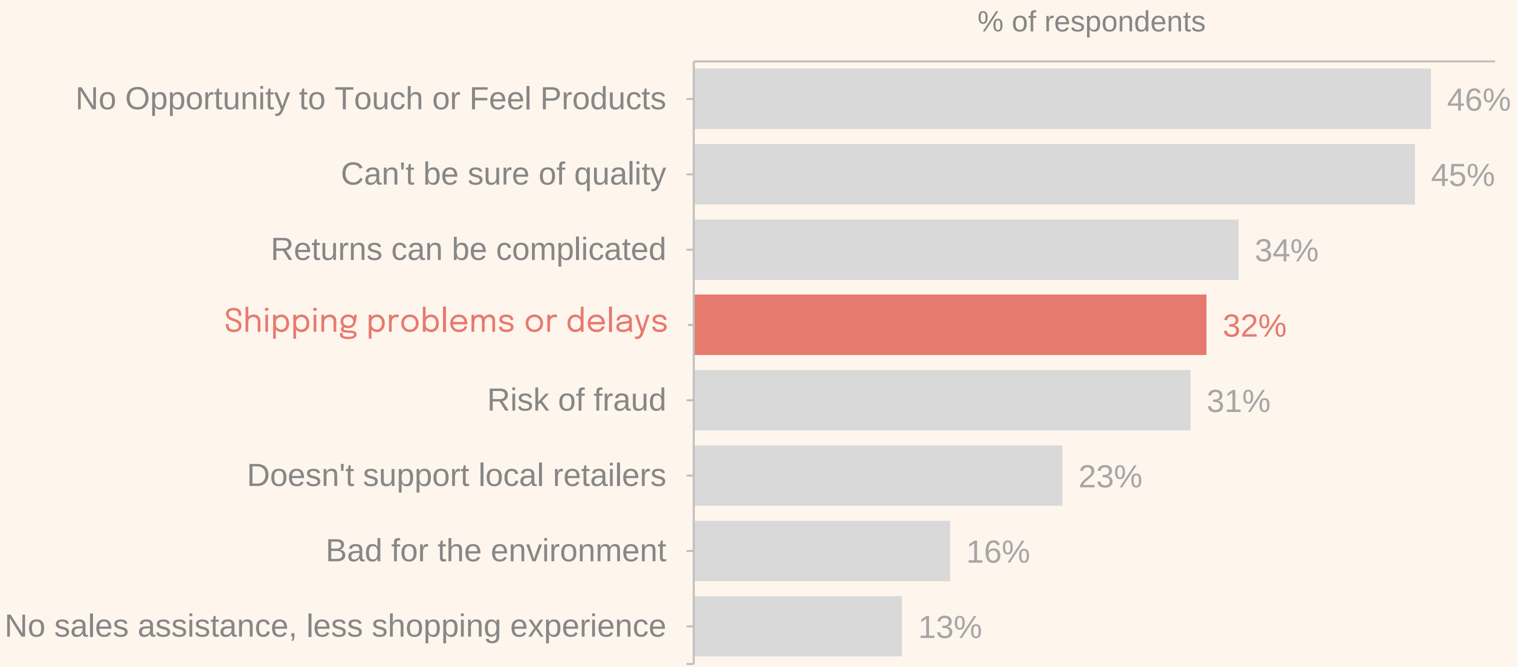


Why do we care about these late deliveries, anyway?

Let's see a worldwide survey of e-commerce biggest drawbacks.

Late Deliveries: Top 4th biggest concerns of e-commerce

From a total of 11.2K respondents



According to a survey titled “Main disadvantages of online shopping worldwide 2022” by Koen Van Gelder
Published via Statista on September 29, 2023

How will late deliveries handled in this project?

Meet the Project Body

Brief details

Background	Concerning high rate of late deliveries that could lead to customer loss
Goal	Increase On-time Delivery Rate by implementing solutions related to the factors causing delays.
Objective(s)	Create and use a machine learning model to predict delivery status and identify the root causes. Then, recommend solutions and business actions to address those root causes.
Business Metrics	On-time Delivery Rate

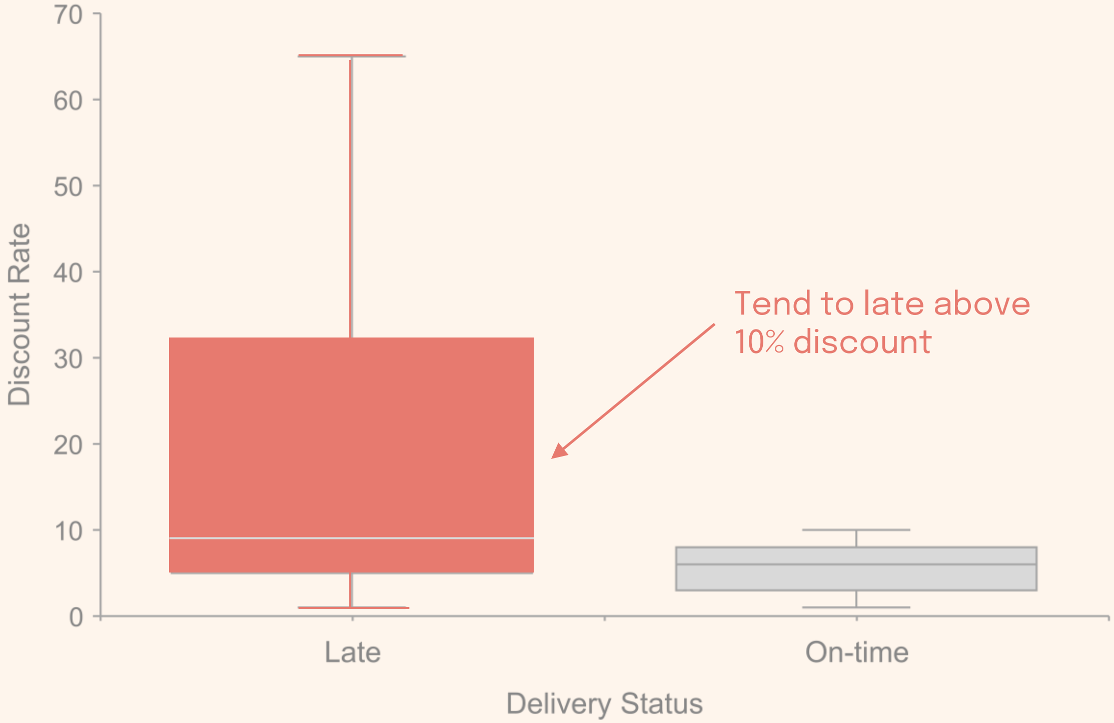
02

What Really Happened?

Deep dive into the data itself

Late deliveries are happening at a higher discount rate

More frequently above the 10% discount area

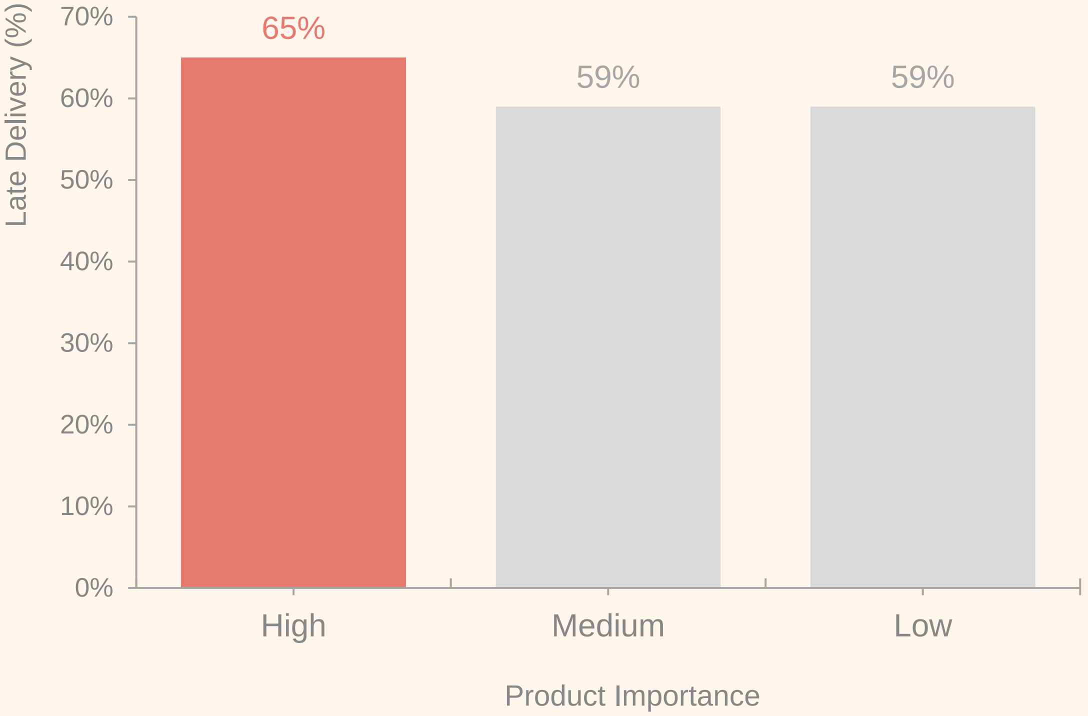


Assumption:

The company may prioritize orders with lower discounts, as they generate higher profit margins.

High-importance orders face delivery delays more

There may be difficulties or inefficiencies that are happening in the shipment process for high importance products.



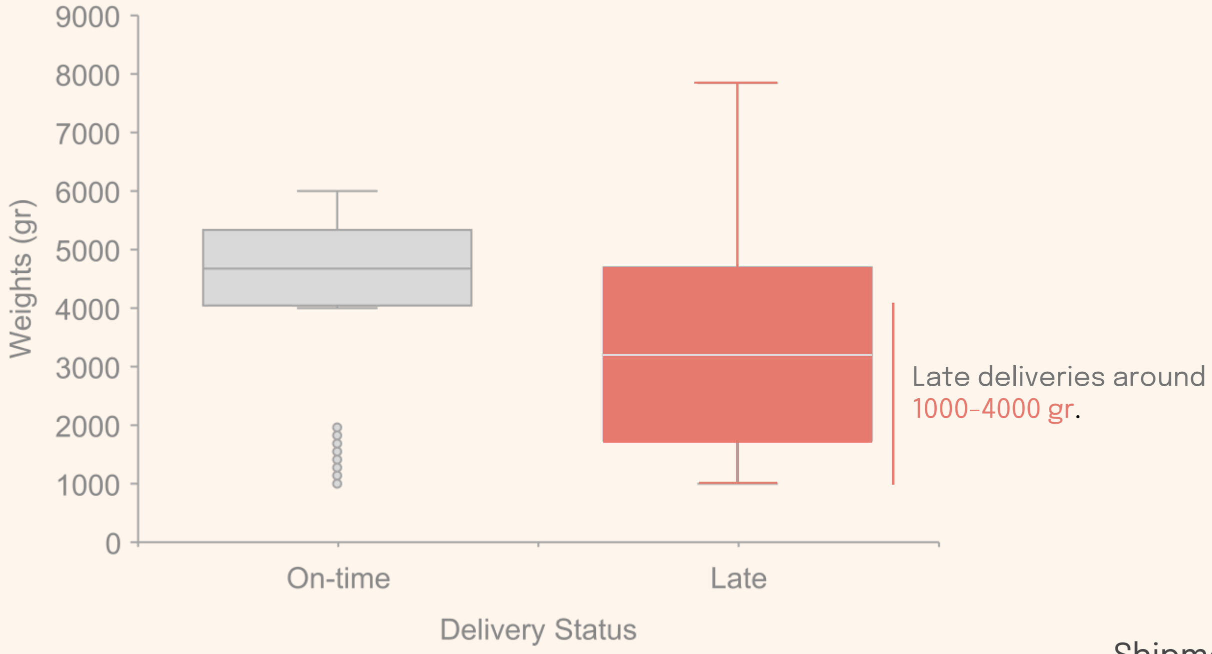
Assumption:

High importance orders may require more complex handling, like specialized packaging or stricter quality checks

Up next: Product Weights

Lighter weight ranges are experiencing late deliveries

Products around the highlighted ranges always running late



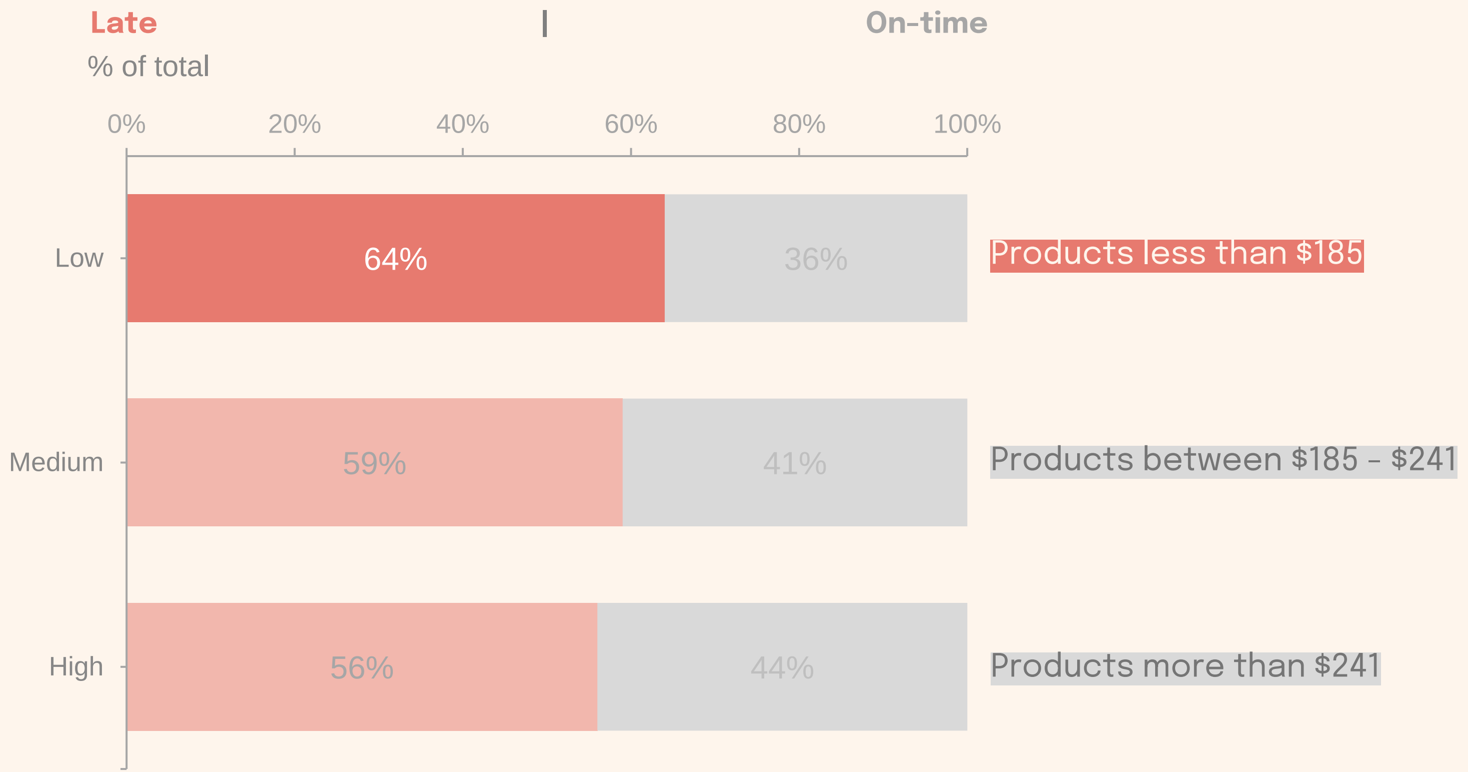
Assumption:

Shipments weighing more than 4000 gr are more prioritized by the company.

Up next: Product Cost Class

Low cost products suffer from more late deliveries

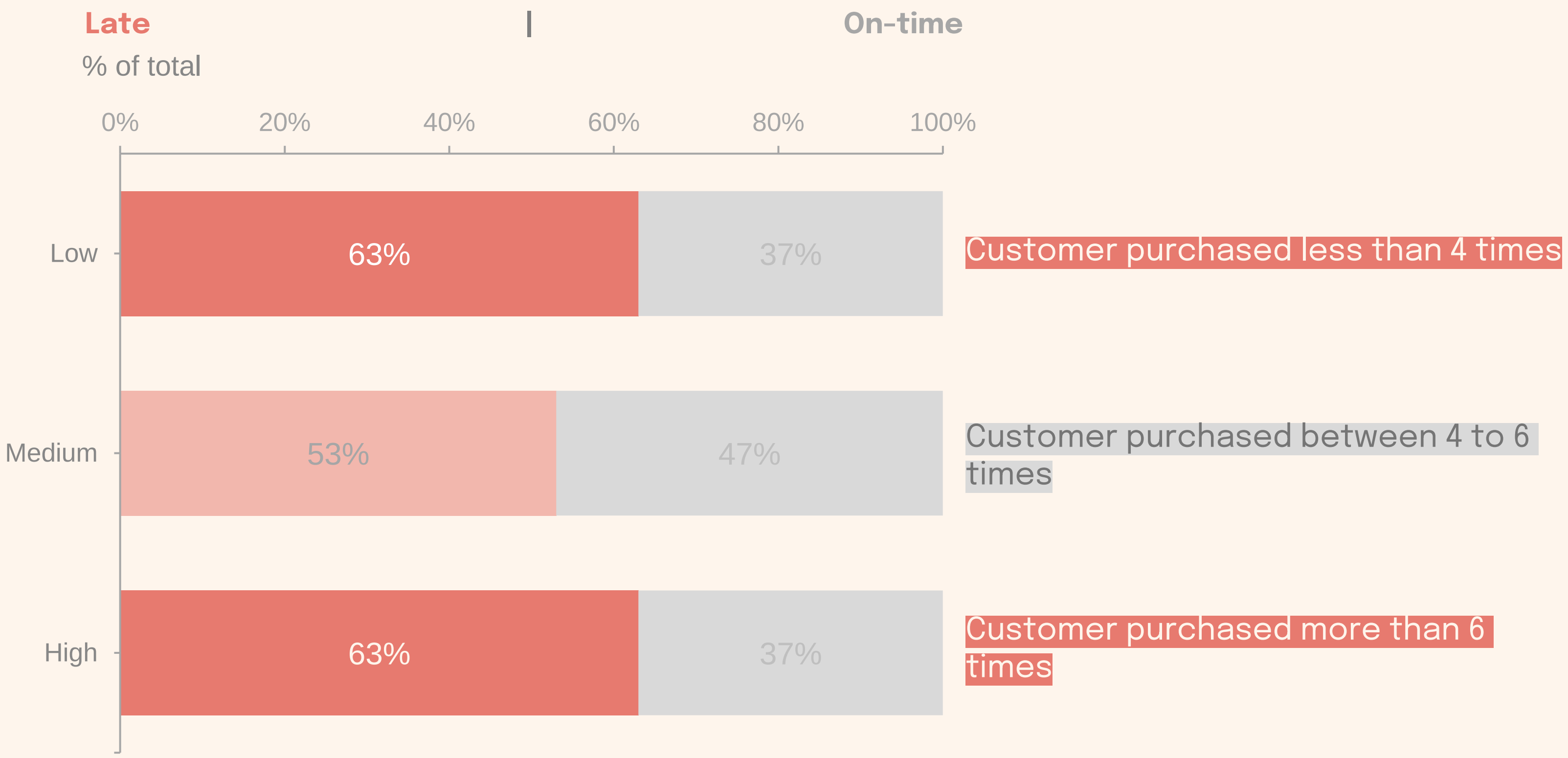
Even though the lateness is higher than the on-time deliveries across all cost class



Up next: Prior Purchases Class

Low and high prior purchases experience more lateness

Why do we neglect these new and repeat customers' deliveries?

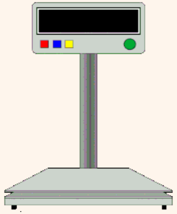


Some of these associations are backed up by statistical measures such as correlation

To sum it up

5 Factors that affect deliveries

Discount offered



Product's Weight

Product Importance

Product Cost

Prior Purchases

03

What can we do?

Model Development &
Business Actions

Before the model development

We cleaned and preprocess the data

To make the learning process efficient and appropriate

Cleaning Checklist

Data State	Are there any?	Actions
Missing Values	✗	Not required
Duplicates	✗	Not required
Invalid Values	✗	Not required
Outliers	✓	Transformation and outlier-robust model
Skewed	✓	Transformation

The values are still valid (not input error)

Discount feature (right-skewed)

Preprocessing Checklist

Task	Necessary?	Actions
Encoding	✓	Ordinal Encoding
Scaling	✓	Standardization

Then,

We developed a model

That could predict whether a delivery will be late or on-time

Let's see the initial evaluation of the model

Initial Evaluation of Models

Model Algorithm

Metrics	Logistic Regression	Decision Tree	Random Forest	kNN	SVM	Ada Boost	XGBoost
Accuracy	0.63	0.64	0.65	0.64	0.67	0.67	0.65
Precision	0.67	0.70	0.71	0.71	0.90	0.81	0.73
Recall	0.75	0.70	0.69	0.67	0.51	0.58	0.65
F1	0.71	0.70	0.70	0.70	0.65	0.68	0.69
AUC	0.72	0.63	0.73	0.72	0.74	0.74	0.73

Recall is used to reduce the number of orders that are actually late but predicted to be on time

We selected these highlighted models to be tuned further prioritizing the recall and AUC scores but keep it from going overoptimistic

Evaluation of Tuned Models

Model Algorithm

Metrics	Logistic Regression	Random Forest	XGBoost
Accuracy	0.63	0.63	0.62
Precision	0.66	0.65	0.63
Recall	0.77	0.83	0.88
F1	0.71	0.73	0.73
AUC	0.72	0.75	0.74

Random forest has good recall, highest Accuracy and AUC scores

Despite XGBoost having the highest recall, the number is too overoptimistic because it would produce more False Positive too
That is, orders that are actually on-time but predicted to be late

But,

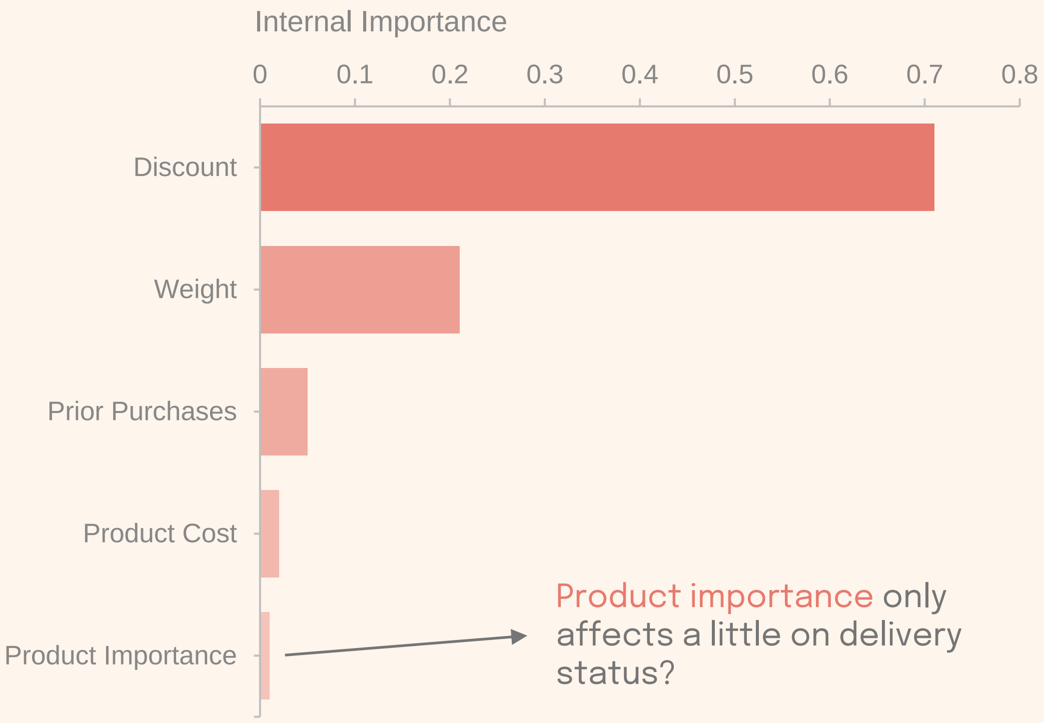
Which features are important?

According to our
Random Forest model

Let's see the feature importance of the model

Discount and weight are the main factors that contribute to the delivery status

Interestingly, product importance doesn't play any role in this model



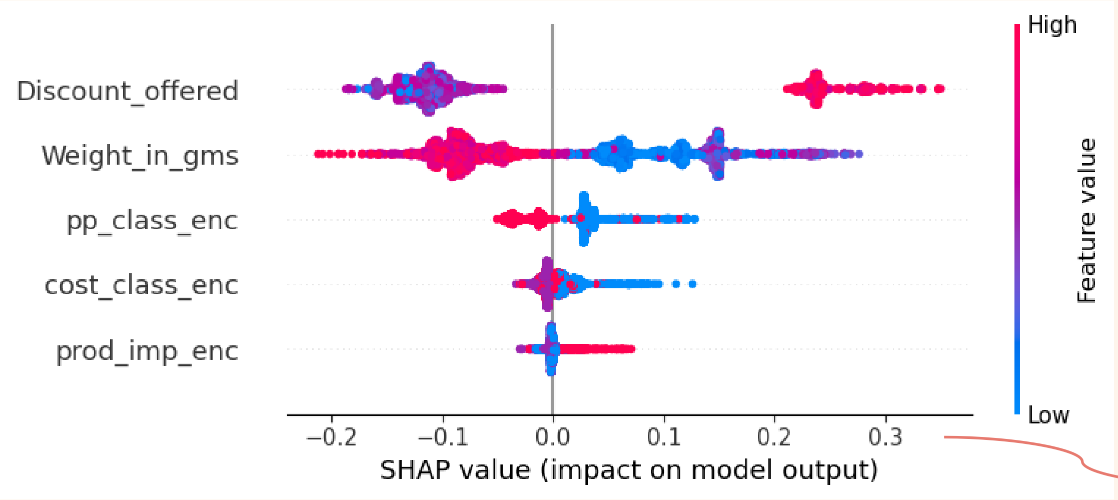
But still,

Discount is the most prominent feature that is associated with deliveries

We don't know the direction of these features towards the delivery status.

For that, we're gonna use SHAP values.

Which direction?



More lateness!

Higher discounts → late deliveries

Lighter weights → late deliveries

Lower prior purchases → late deliveries

These are associations only, not causal relationship.

Finally,

What should we really do?

After all the data exploration and model building

Business actions and simulations

We recommend to...

Implement a Discount Threshold Policy

Bundled Products Shipment

Discount Threshold Policy

++ profit margin!! (assumption)

Balancing customer interest and on-time delivery

Challenges

Need to investigate further why higher discounts lead to late deliveries.

Risk that some customers may lose interest if we limit discounts

Mitigations

Relevant data, such as detailed warehouse and logistics data

Loyalty rewards, points, or early access to new products

Bundled Products Delivery

A simple and invisible solution

- Lighter products are more likely to be late than medium weight products.
- Bundle lighter products together into heavier packages to reduce the likelihood of late deliveries.
- This process would be invisible to customers.

Challenge

Inventory systems must accurately account for bundled products. Failure to do so can lead to mistakes in the shipment process.

Mitigation

Use additional barcode or radio-frequency identification (RFID) technology to track bundled products.

Up next: Discount and Weight Simulation

Currently, we have

40%

On-Time Delivery Rate

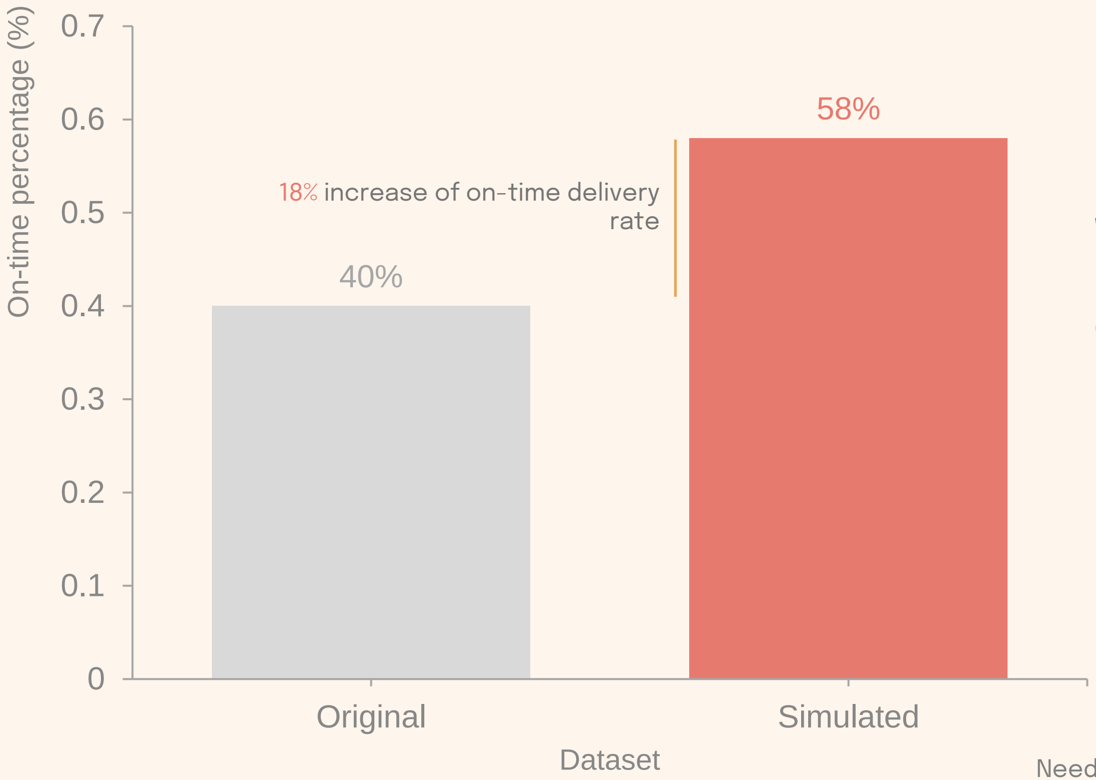
Finding the optimal discount threshold

Finding the optimal weight ranges

Assessing the on-time delivery rate

Original vs. simulated on-time delivery rates

Discount and weight are the main factors in increasing the on-time delivery rate



What did we simulate?

4% maximum discount threshold

5000gr minimum weights (bundled)

Need additional shipment identifier for the shipment weight bundle

*These are only temporary solutions

But wait,

There's one extra catch in the simulation process

We play around with the prior purchases and turns out:
Higher prior purchases = more on-time deliveries

We're **neglecting** these new customers' delivery commitment !!!

Use more reliable shipping carriers for orders from new customers.

Offer incentives to new customers who experience late deliveries

Additional Actions

In the case of late deliveries

Inform customers of delivery delays

- Contact the customer to inform them of the delay.
- Extend the SLA (Service Level Agreement) by x days.

Offer incentives for late deliveries

- Loyalty points, free delivery, or exclusive access to new products.
- Give customers the option to choose the incentive

We get the 'x' days if we develop a new model that can predict the order delivery date.

Fourcasters

Thank you!