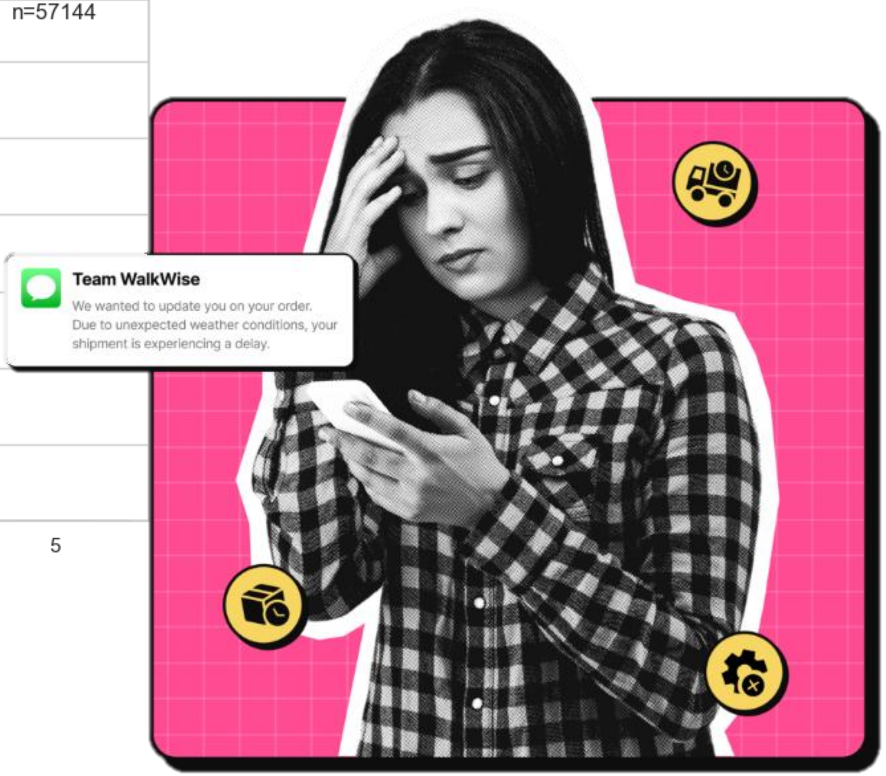
Delivering Trust: Predicting E-Commerce Delivery Delays and Designing Smart Shipping Insurance

Group 1: Wei Xing, Rui Guo

Why Delivery Delays Matter in E-Commerce

- Customers who experienced delivery delays left negative reviews.

- Within 9757 review, 1016 of them (11.07%) mentioned “delay-related” words:



`atraso`, `atrasos`, `atrasada`	atraso = delay
`demora`, `demoras`, `demorado`	delay, delayed
`espera`, `esperando`	wait, waiting
`chegou tarde`	arrived late
`nao chegou`	didn’t arrive
`entrega atrasada` / `entrega atraso`	delayed delivery
`fora do prazo`	out of the deadline

Current Solutions & Our Innovation

- Platforms rely on static SLAs and broad refund policies, often reactive
- Existing delay handling = manual customer service or fixed refund policies

Purchase shipping insurance at checkout

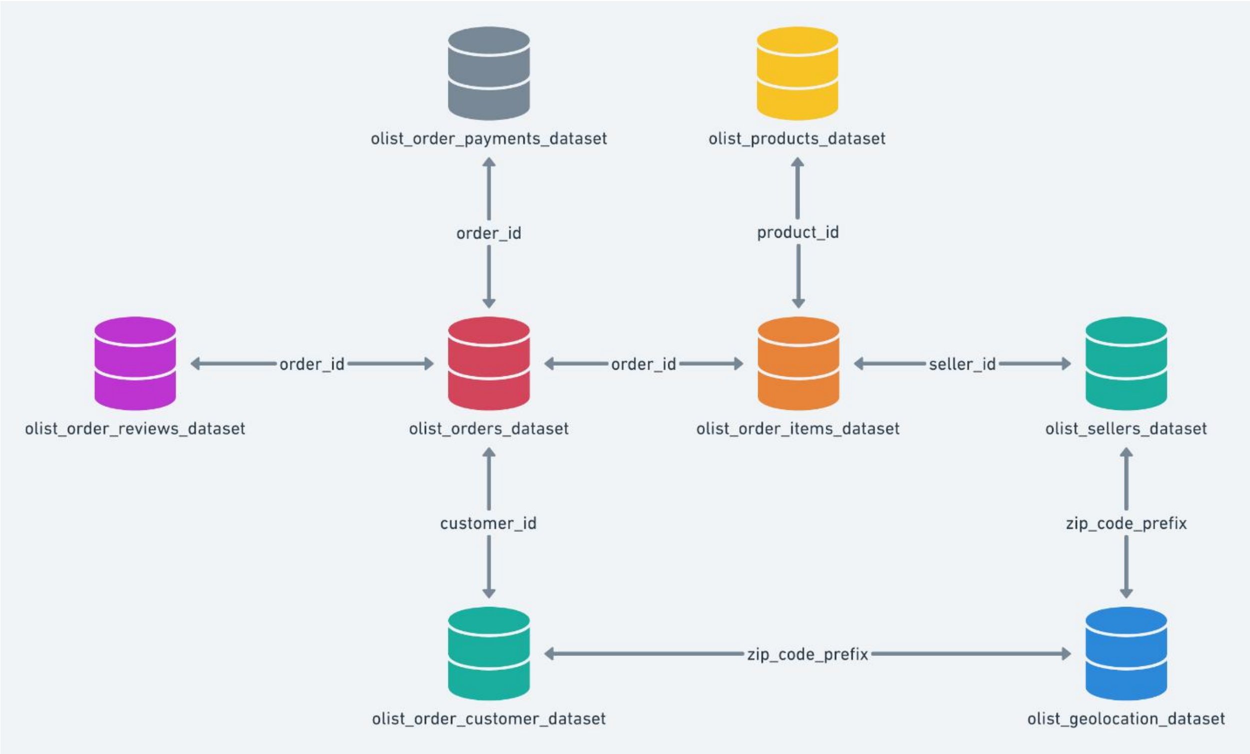
delivery delay

shipping fee instantly refunded

Dynamic Insurance Pricing Powered by Delivery Delay Prediction

Dataset Overview

- Real commercial data, it has been anonymized
- 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil



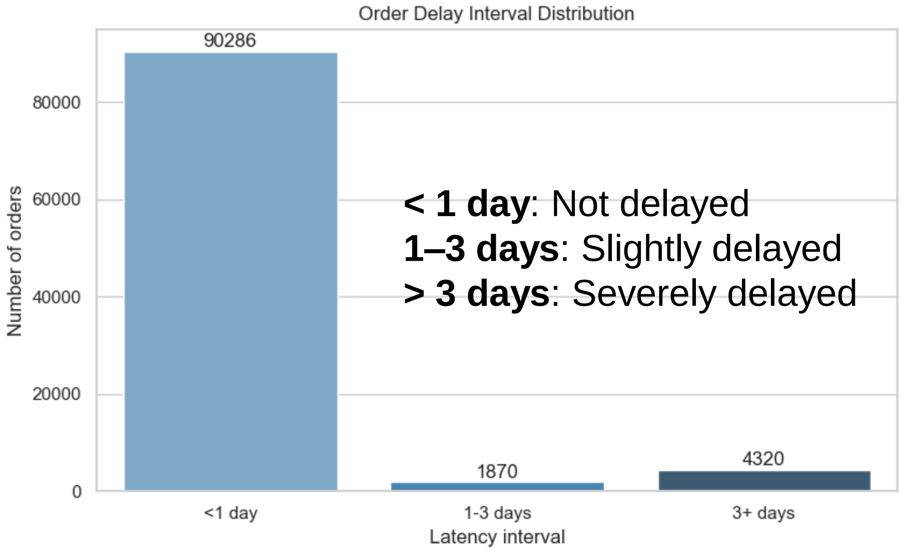
•Order/payment/shipping/delivery timestamps
•Product category, dimensions, weight
•Geolocation of both customers and sellers
•Customer review scores and text feedback

merge

order_full.csv
Size:(115730,21)

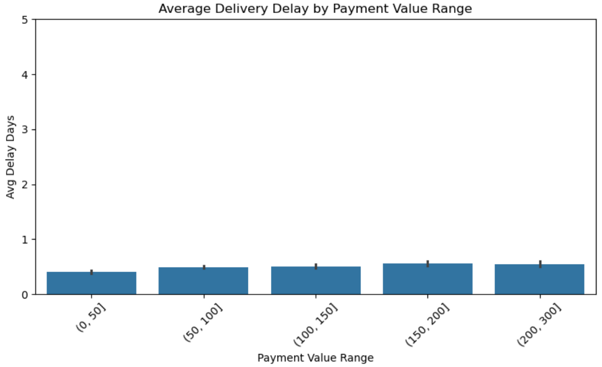
Exploratory Data Analysis (EDA)

- Delays are common: ~6.4% of orders exceed expected delivery date



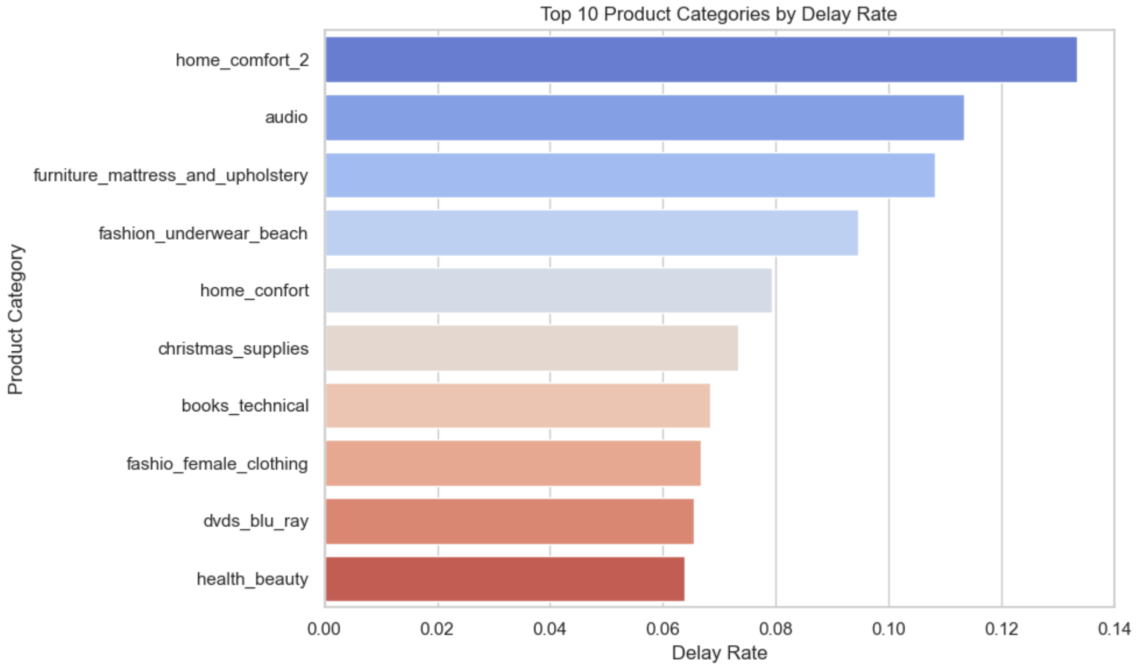
< 1 day: Not delayed
1–3 days: Slightly delayed
> 3 days: Severely delayed

- Payment amount has little correlation on delivery delay



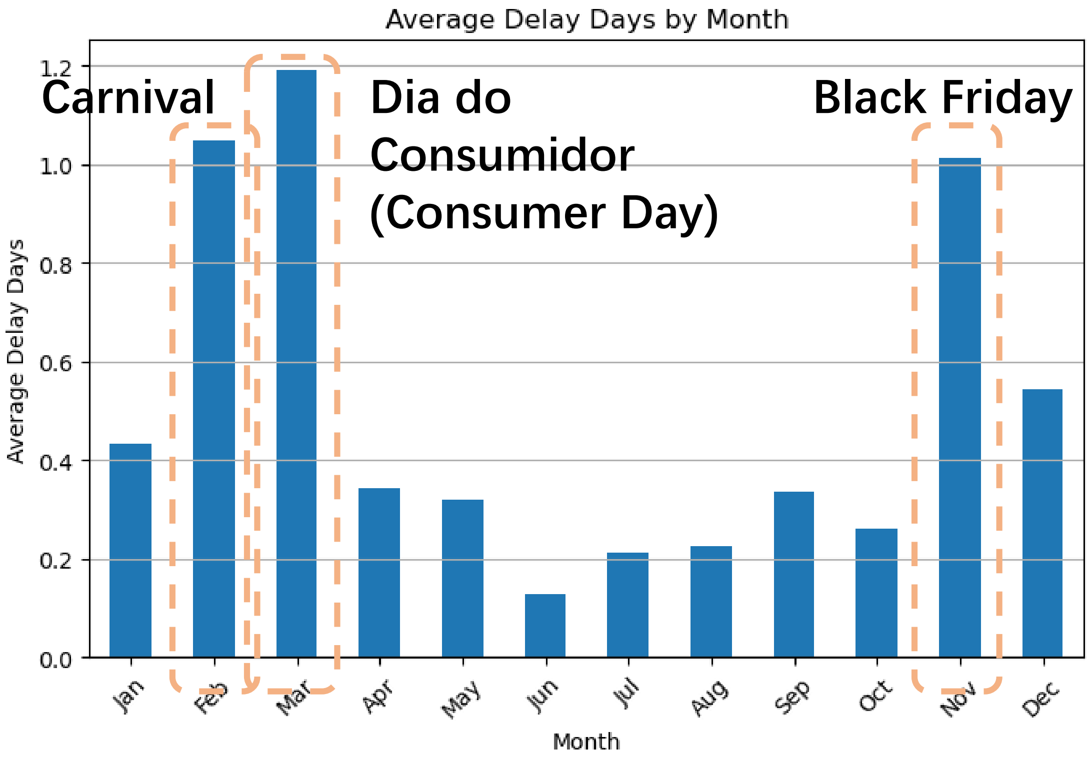
- Top 5 Delay-Prone Product Categories

- ☐ Seasonal home comfort appliances
- ☐ Audio devices and equipment
- ☐ Furniture and home textiles
- ☐ Intimate wear and beachwear
- ☐ General home comfort goods

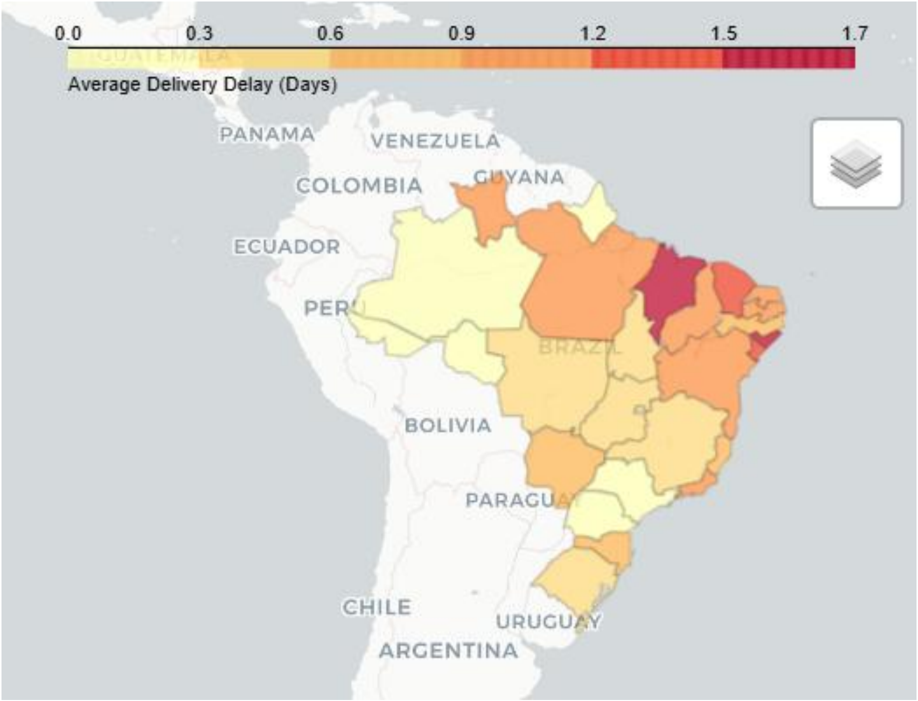


Exploratory Data Analysis (EDA)

- Strong seasonality effects (holidays, sales events)



- Geographic patterns: Rural destinations have higher delay rates



What Are We Trying to Discover?

RQ1: What factors most strongly predict delivery delays?

RQ2: How can we use existing order information to provide platforms with a fair and data-driven shipping insurance pricing strategy?

Methodology：Predicting Delivery Delays

Feature engineering	Feature Creation •delay_days = actual_delivery_date - estimated_delivery_date •actual_shipping_days = delivery_date - purchase_timestamp •is_delayed = 1 if delay_days > 1 else 0 Missing Value Handling: remove missing values One-Hot Encoding on categorical variables	
Problem	on-time vs. delay	number of delay days
Method	Classification	Regression
Model	Logistic Regression, Random Forest Classifier	Linear Regression, Random Forest Regressor
Evaluation	Accuracy, Precision, Recall	RMSE, R^2

Variable Selection

Feature Name	Reason for Not Selection
customer_id, order_id	These are ID fields that carry no informational value and cannot be vectorized.
product_category_name, seller_id	Too high-dimensional (too many categories), not suitable for direct input into the model; would require further embedding or clustering to use.
review_score	This is an "outcome" variable, likely influenced by delays, and therefore cannot be used to predict them.
geolocation, city, state	Geographical data is important, but requires preprocessing into numeric or distance-based features; processing is more complex.
order_purchase_timestamp, order_delivered_customer_date	These time-based features have already been converted into month and shipping_days for use.

Feature Name	Meaning	Reason for Selection
Estimated_shipping_days	Estimated number of days from purchase to delivery	Important factor about whether a delay occurred; serves as the key label feature for learning delay patterns.
freight_value	Shipping cost	Shipping cost may be related to distance, weight, or priority, indirectly affecting delay probability.
price	Price of items in the order (product value)	Product price may influence the shipping method (e.g., expensive items may use faster logistics).
payment_value	Total payment amount by customer	May include multiple products and reflect the "importance" of the order; also potentially overlaps with freight, so retained for comparison.
month	Month of purchase	Seasonality and holidays may increase delivery risk; provides essential temporal signals.

Results & Insights

Binary Classification

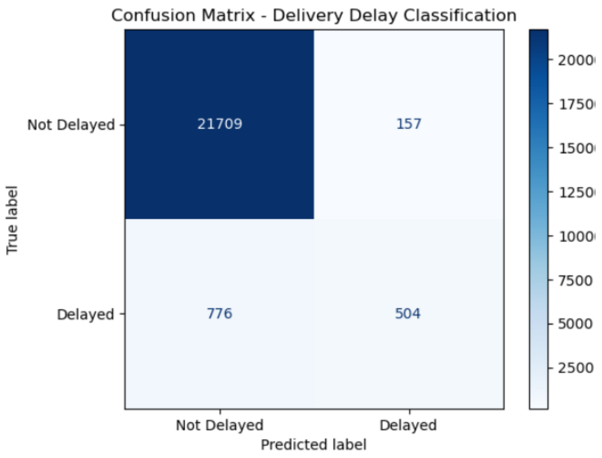
Classification Report: **Logistic Regression** → **Random Forest Classifier**

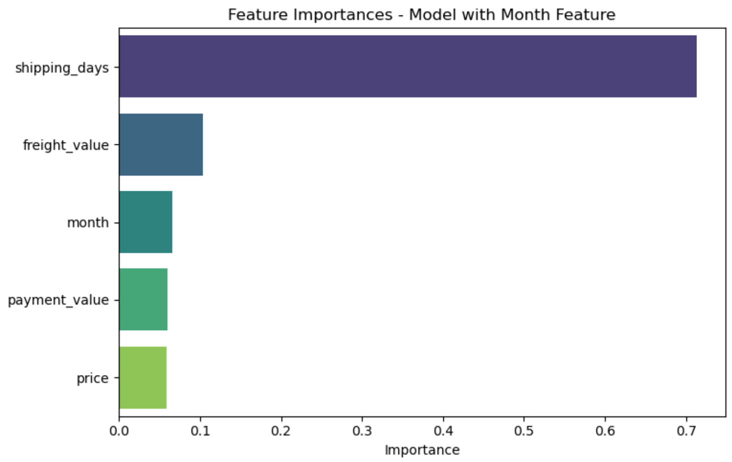
Logistic Regression

	precision	recall	f1-score	support
0	0.97	0.99	0.98	21866
1	0.76	$\boxed{0.39}$	0.52	1280
accuracy			0.96	23146
macro avg	0.86	0.69	0.75	23146
weighted avg	0.95	0.96	0.95	23146

Random Forest Classifier

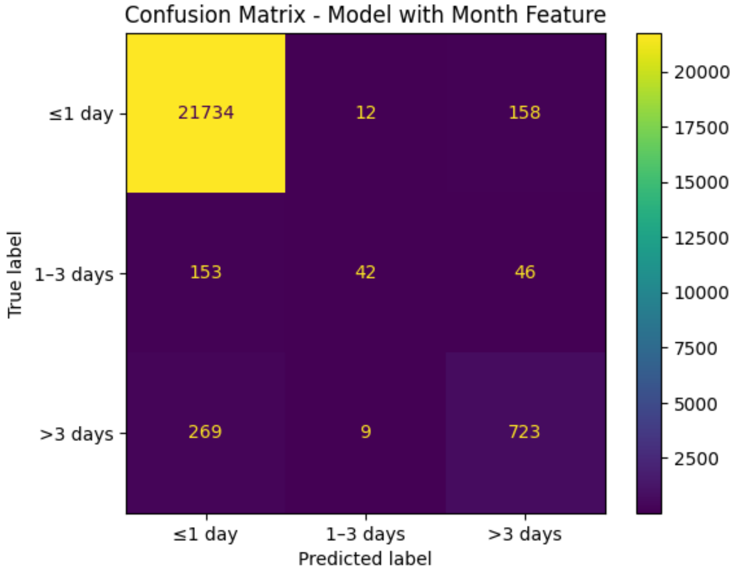
	precision	recall	f1-score	support
0	0.98	0.99	0.99	21904
1	0.81	$\boxed{0.68}$	0.74	1242
accuracy			0.97	23146
macro avg	0.90	0.83	0.86	23146
weighted avg	0.97	0.97	0.97	23146





Multi-class Classification

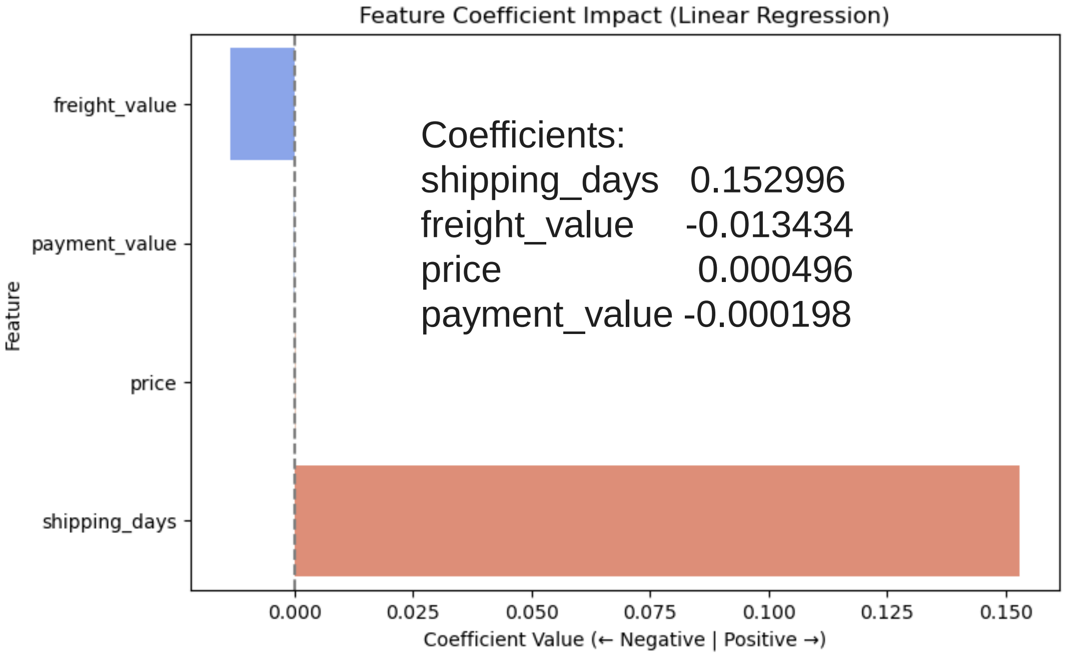
	precision	recall	f1-score	support
0	0.98	0.99	0.99	21904
1	0.67	0.17	0.28	241
2	0.78	0.72	0.75	1001
accuracy			0.97	23146
macro avg	0.81	0.63	0.67	23146
weighted avg	0.97	0.97	0.97	23146



Results & Insights

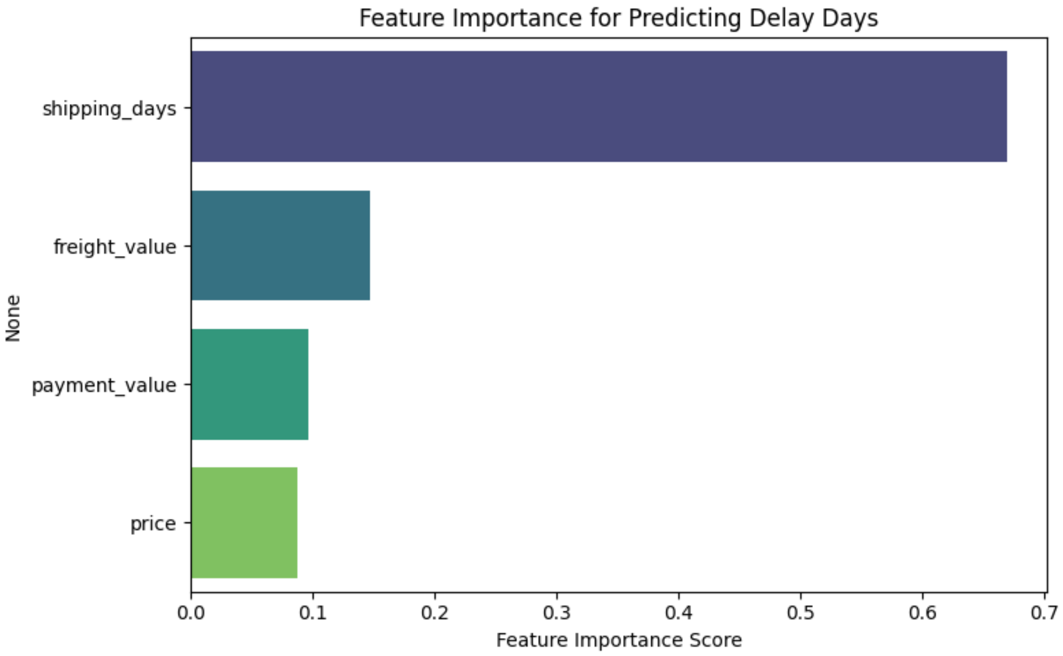
Linear Regression

```
Linear Regression RMSE: 2.27
Linear Regression R^2 Score: 0.25
```



Random Forest Regression

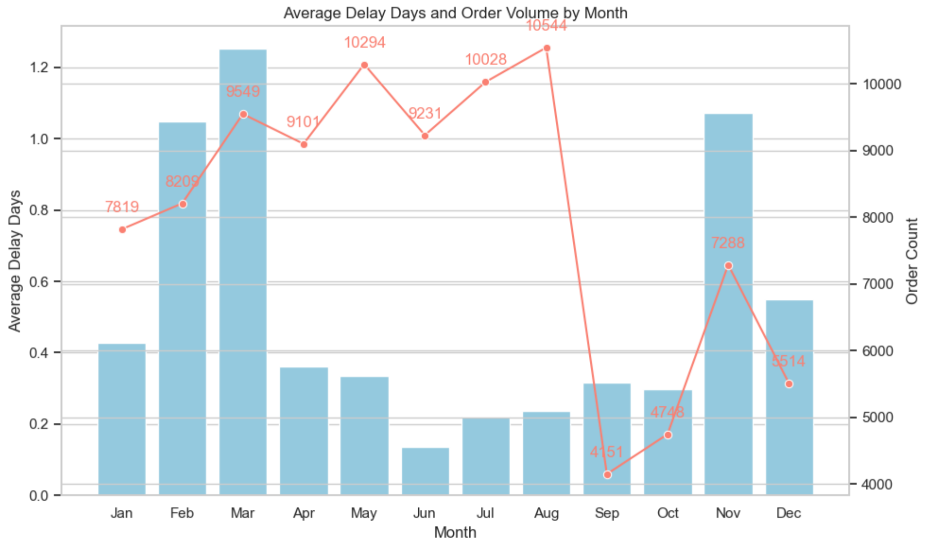
```
Random Forest Regression
RMSE: 2.21   R²: 0.67
```



Results & Insights

Further data mining

Considering the significant variation in delays across months, we trained separate models for each month's orders.



Month	RMSE	R^2	Month	RMSE	R^2
Jan	1.56	0.53	July	0.89	0.73
Feb	1.95	0.76	Aug	1.0	0.3
Mar	1.97	0.75	Sep	1.07	0.71
Apr	1.11	0.64	Oct	1.62	0.53
May	0.94	0.61	Nov	1.82	0.79
Jun	0.73	0.73	Dec	1.36	0.75

Use the month as an independent variable, 'order_month'

RMSE: 1.82 R^2 Score: 0.73

Methodology： Insurance Pricing Design

Predict whether delay with our classification model

Estimate the "delay probability" based on the predicted value for each month

pred_delay_prob = (TP + FP) / total_samples

Design a risk-based pricing mechanism

insurance_price = (delay_prob × average_shipping_cost × safety_factor) / coverage_rate

- **This price represents an "expected cost plus profit" model, designed to cover the potential risk borne by the platform while ensuring a reasonable return.**

Results & Insights

order_id	customer_id	order_purchase_timestamp	order_estimated_delivery_date	shipping_days	order_item_id	price	freight_value	product_category_name	payment_value	customer_state	Insurance price
6514b8ad8028c9f2cc2374ded245783f	9bdf08b4b3b52b5526ff42d37d47f222	2017/5/16 13:10	2017/6/7	9	1	59.99	15.17	automotivo	75.16	RS	1.22
76c6e866289321a7c93b82b54852dc33	f54a9f0e6b351c431402b8461ea51999	2017/1/23 18:29	2017/3/6	9	1	19.9	16.05	moveis_decoracao	35.95	SP	1.54
e69bfb5eb88e0ed6a785585b27e16dbf	31ad1d1b63eb9962463f764d4e6e0c9d	2017/7/29 11:55	2017/8/23	18	1	149.99	19.77	moveis_escritorio	161.42	SP	0.97
e6ce16cb79ec1d90b1da9085a6118aeb	494dded5b201313c64ed7f100595b95c	2017/5/16 19:41	2017/6/7	12	1	99	30.53	ferramentas_jardim	259.06	RJ	2.45
34513ce0c4fab462a55830c0989c7edb	7711cf624183d843aafe81855097bc37	2017/7/13 19:58	2017/8/8	5	1	98	16.13	informatica_acessorios	114.13	RJ	0.79
82566a660a982b15fb86e904c8d32918	d3e3b74c766bc6214e0c830b17ee2341	2018/6/7 10:06	2018/7/18	12	1	31.9	18.23	perfumaria	50.13	SP	0.44
5ff96c15d0b717ac6ad1f3d77225a350	19402a48fe860416adf93348aba37740	2018/7/25 17:44	2018/8/8	4	1	19.9	12.8	cama_mesa_banho	32.7	MG	0.63
432aaf21d85167c2c86ec9448c4e42cc	3df704f53d3f1d4818840b34ec672a9f	2018/3/1 14:14	2018/3/21	11	1	38.25	16.11	brinquedos	54.36	SP	4.63
dcb36b511fcac050b97cd5c05de84dc3	3b6828a50ffe546942b7a473d70ac0fc	2018/6/7 19:03	2018/7/4	13	1	132.4	14.05	perfumaria	146.45	SP	0.34
403b97836b0c04a622354cf531062e5f	738b086814c6fcc74b8cc583f8516ee3	2018/1/2 19:00	2018/2/6	17	1	1299	77.45	construcao_ferramentas_construcao	1376.45	GO	7.42

- If freight_value = 10

Month	Delay Rate	Insurance Price
Jan	3.99%	$0.96
Feb	10.67%	$2.56
Mar	11.97%	$2.87
Apr	2.40%	$0.58
May	3.34%	$0.80
Jun	1.00%	$0.24
July	2.04%	$0.49
Aug	1.85%	$0.44
Sep	2.58%	$0.62
Oct	1.97%	$0.47
Nov	10.18%	$2.44
Dec	5.02%	$1.20

Conclusions & Reflection

Data mining on E-commerce delay

In our dataset, delivery delays are not strongly correlated with order value, but show significant associations with product categories, seasonal factors, and geographic distribution.

- The product categories with the highest average delays are: Seasonal home comfort appliances, Audio devices and equipment and Furniture and home textiles
- February, March, and November are peak delivery months, likely due to national events and sales campaigns.
- In Brazil's Northeastern region, where logistics infrastructure is less developed, the average delivery time is notably longer.

Prediction model

We demonstrated the feasibility of predictive delay modeling in E-commerce. We reach 68% recall on classification model and 0.73 R-square on regression model.

Insurance Pricing

We built a smart shipping insurance pricing mechanism that both improve customer satisfaction and earn 20% extra money.

Limitations : single-region dataset, lack of logistics info of manufacturer, unbalanced data