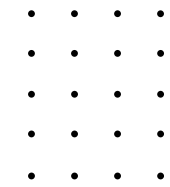


SC1015 Mini-Project: Predicting E-Commerce Order Cancellations

FR1 | Group 8

Gao Anni (N2402461C), Liu Chenyu (N2402421L)

Project Motivation



Order cancellations cause lost revenue, wasted logistics, and inventory issues



Even a small % of cancellations can lead to major operational inefficiencies



Customer trust and satisfaction are negatively impacted by cancellations



Identifying high-risk orders early can lead to cost savings and better planning



A data-driven approach can help predict and reduce cancellations at scale



Motivation

Problem
Statement

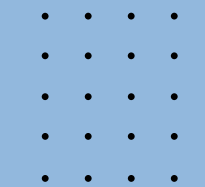
Dataset
Cleaning

EDA

ML Models

Outcomes

Conclusion



Problem

Can we use order-level data to predict order cancellations and reduce revenue loss?



Motivation

Problem
Statement

Dataset
Cleaning

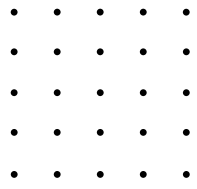
EDA

ML Models

Outcomes

Conclusion

Dataset Overview



Source and Scope

- Real-world dataset from Amazon.in, downloaded from Kaggle
- Contains around 128,975 e-commerce transactions
- Includes product, fulfillment, pricing, and order status details



Fulfilment method
(Amazon vs Merchant)



Ship service level
(Standard vs Expedited)



Amount paid by
the customer



Promotion type
(e.g., No Promo,
Free Shipping,
PLCC, etc.)



Product category

Variables Selected

Target Variable



Order Status



Motivation

Problem
Statement

Dataset
Cleaning

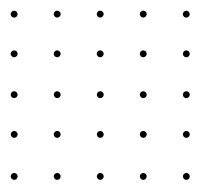
EDA

ML Models

Outcomes

Conclusion

Data Cleaning & Preparation



Removed irrelevant identifiers

Variables to keep

Status
Fulfilment
ship-service-level
Category
Amount
promotion-ids

Variables to remove

Order ID
Date
Sales Channel
Style
SKU
Size
ASIN
Courier Status
Qty
currency
ship-city
ship-state
ship-postal-code
ship-country
B2B
fulfilled-by
index
Unnamed: 22

Variables to add

Category_name



to ensure
interpretability of the
analysis

Handle Missing Values

- Filled missing values in promotion-ids with "no"
- Filled missing values in Amount with 0

Feature Encoding

- Converted Fulfilment, Shipping level, and Promotion type to numeric format
- Encoded Order Status as binary: 1 = Cancelled, 0 = Completed



convert categorical
variables into
numerical format so
that machine learning
models can process
them

Target Filtering

Filtered dataset to only include
Cancelled and Completed orders

Motivation

Problem
Statement

Dataset
Cleaning

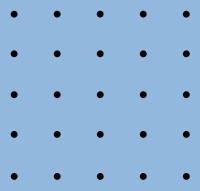
EDA

ML Models

Outcomes

Conclusion

Cleaned Dataframe



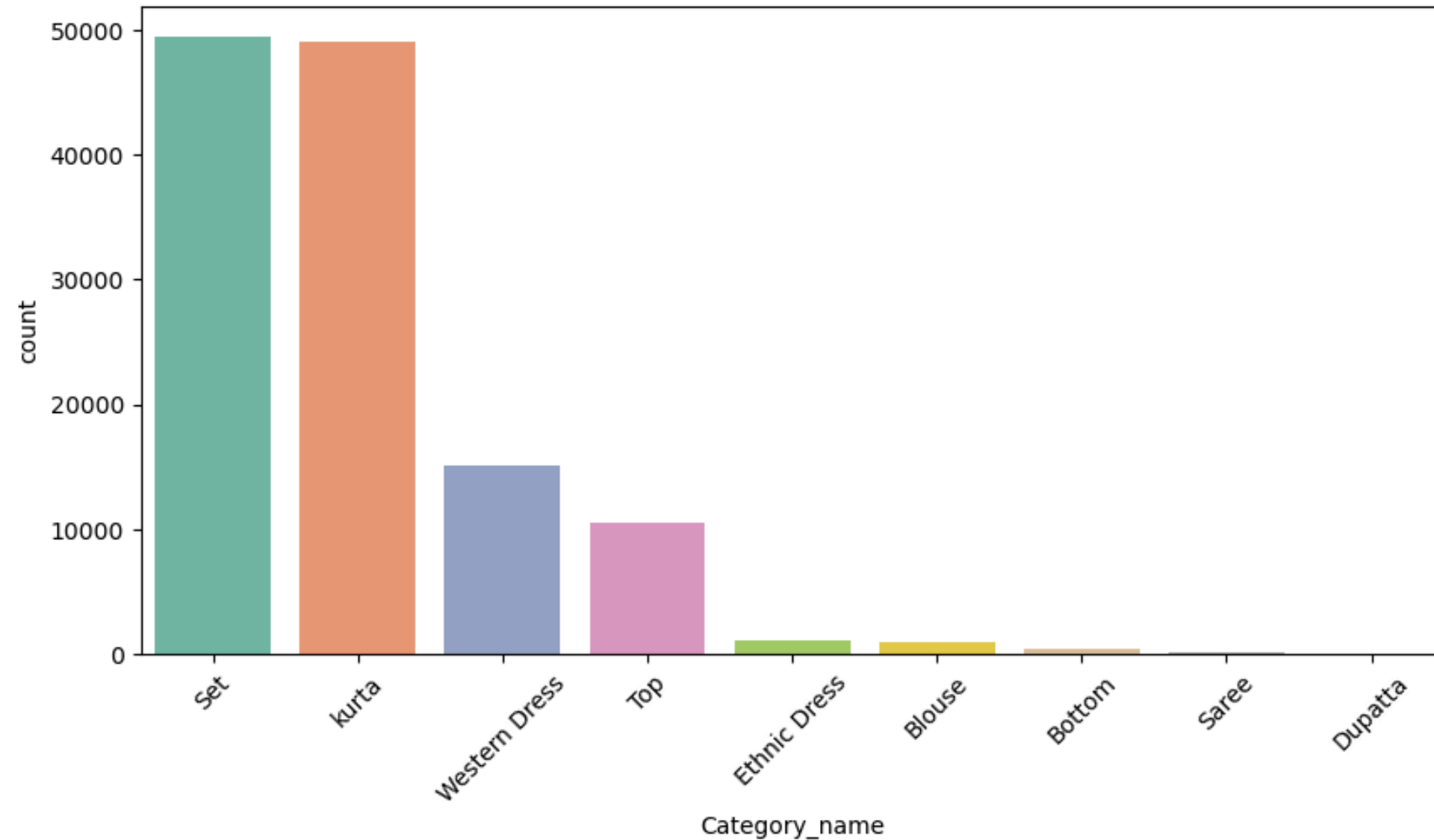
df								
	Status	Fulfilment	ship-service-level	Category	Amount	promotion-ids	Category_name	
0	1	1	0	5	647.62	0	Set	
1	0	1	0	8	406.00	2	kurta	
2	0	0	1	8	329.00	1	kurta	
3	1	1	0	7	753.33	0	Western Dress	
4	0	0	1	6	574.00	0	Top	
...	
128970	0	0	1	8	517.00	0	kurta	
128971	0	0	1	5	999.00	1	Set	
128972	0	0	1	7	690.00	0	Western Dress	
128973	0	0	1	5	1199.00	1	Set	
128974	0	0	1	5	696.00	1	Set	

126825 rows x 7 columns



Exploratory Data Analysis

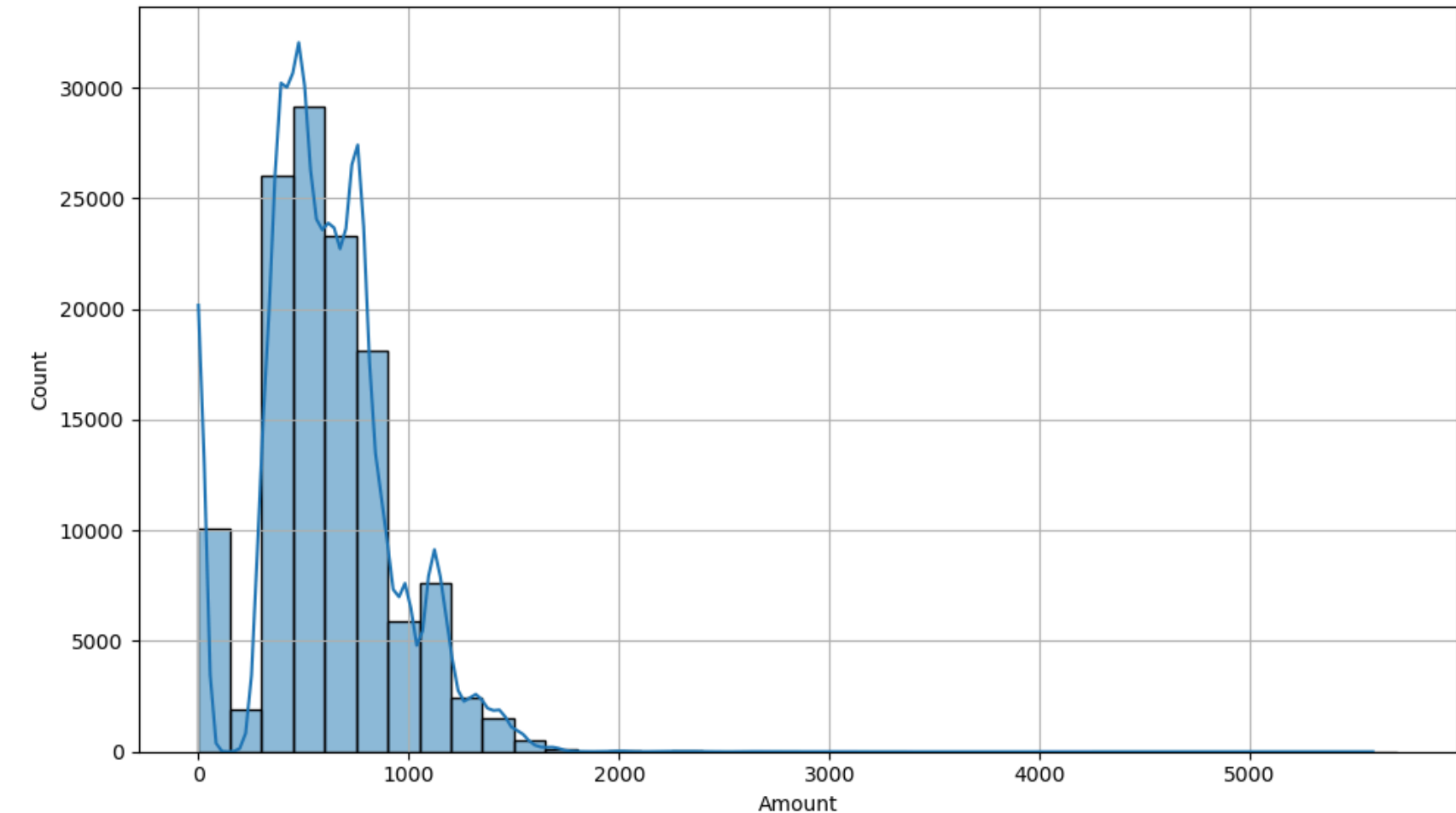
Number of orders for each product category



- The majority of orders come from just a few product categories, indicating a highly imbalanced category distribution.
- To ensure model stability and avoid noise from low-frequency categories, we decided to focus on the top 4 product categories for modeling.



Distribution of Sales Amount across Count of Orders



- Most orders are between ₹300–₹1000.
- The distribution is right-skewed with a long tail of high-value orders.

Motivation

Problem
Statement

Dataset
Cleaning

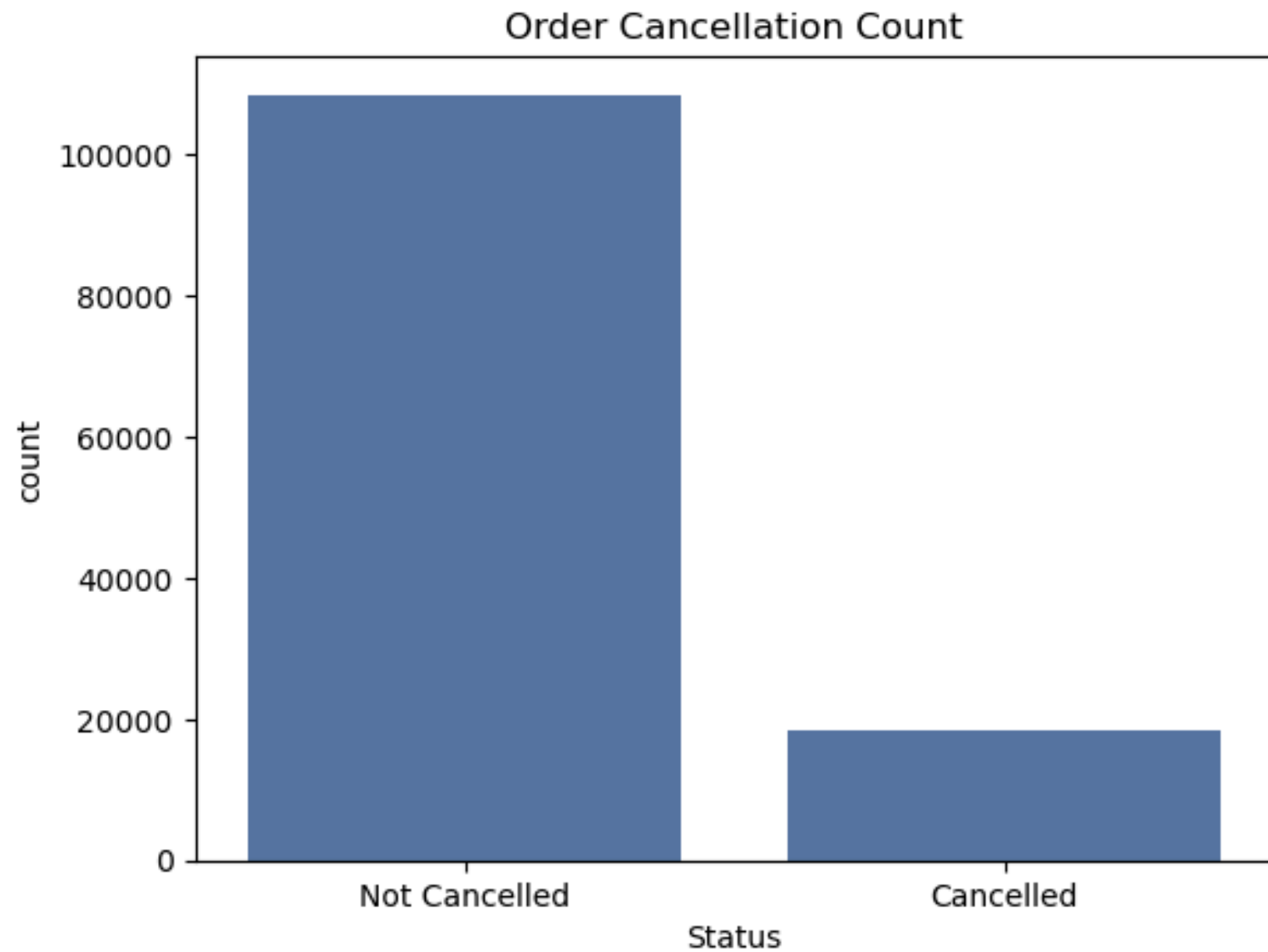
EDA

ML Models

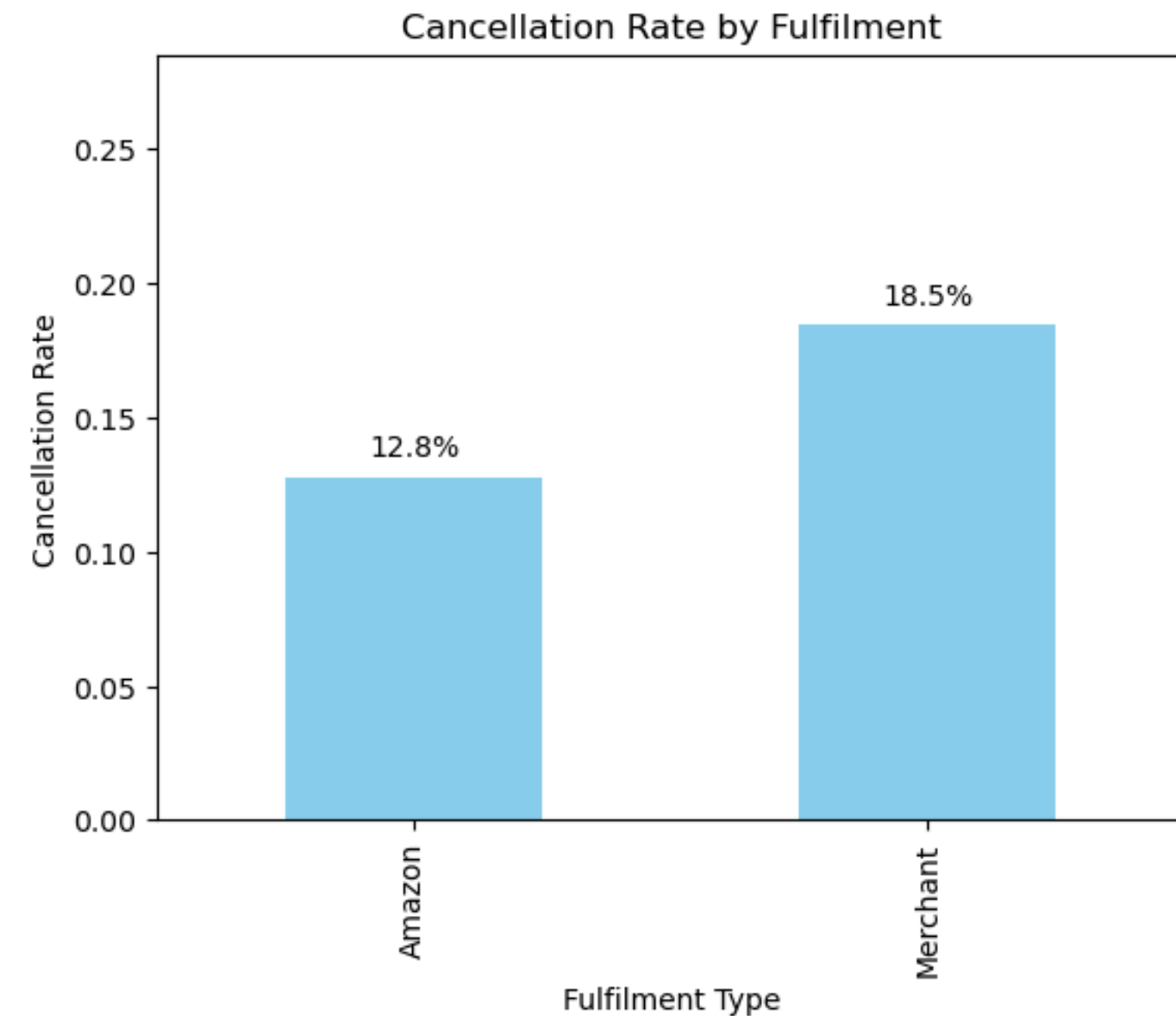
Outcomes

Conclusion

Exploratory Data Analysis



- 85.55% of orders are Not Cancelled
- 14.45% of orders are Cancelled
- Significant class imbalance – needs to be addressed in modeling (e.g., using class weights)

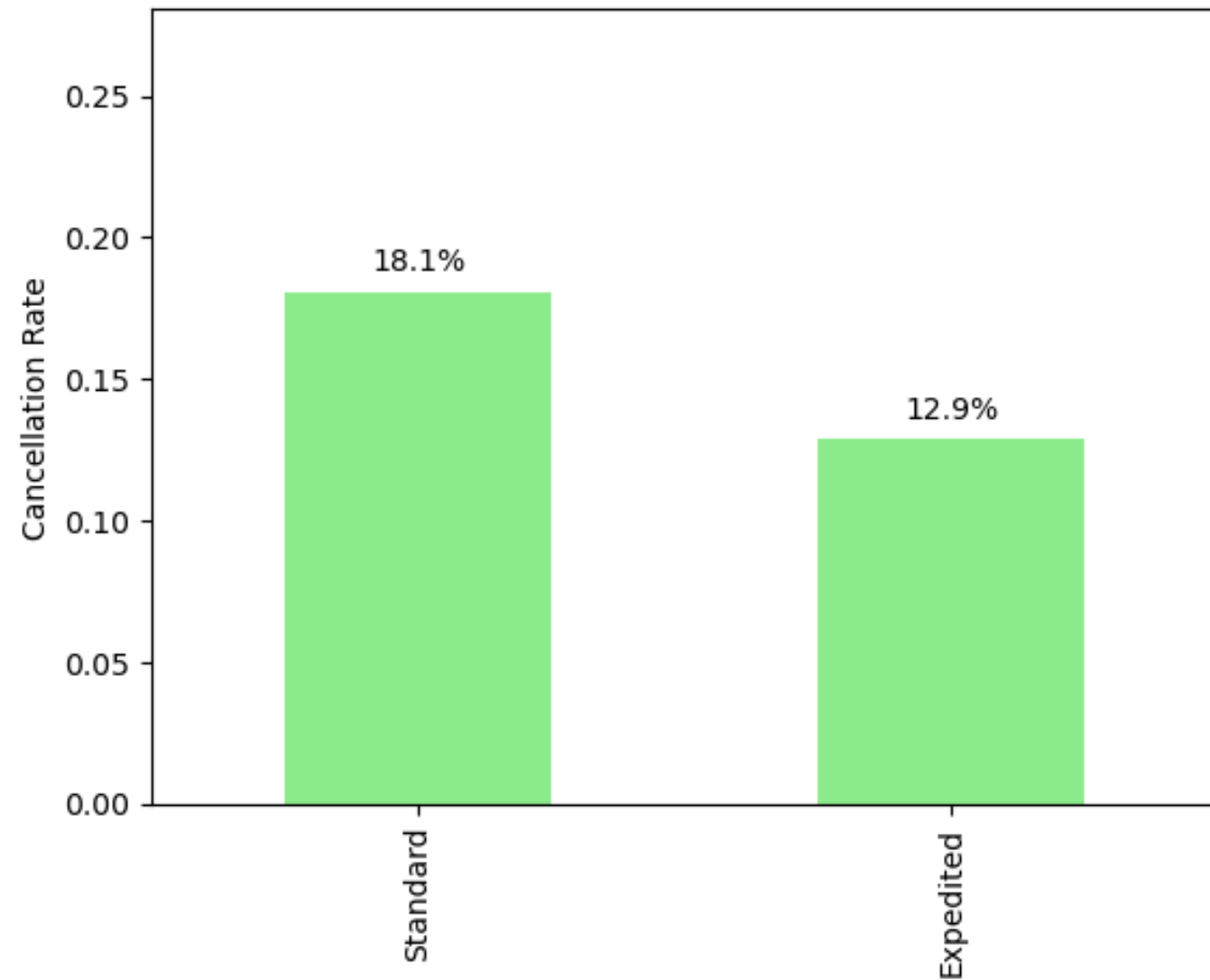


- Merchant orders are ~45% more likely to be cancelled
- Fulfilment method is a key factor influencing cancellations



Exploratory Data Analysis

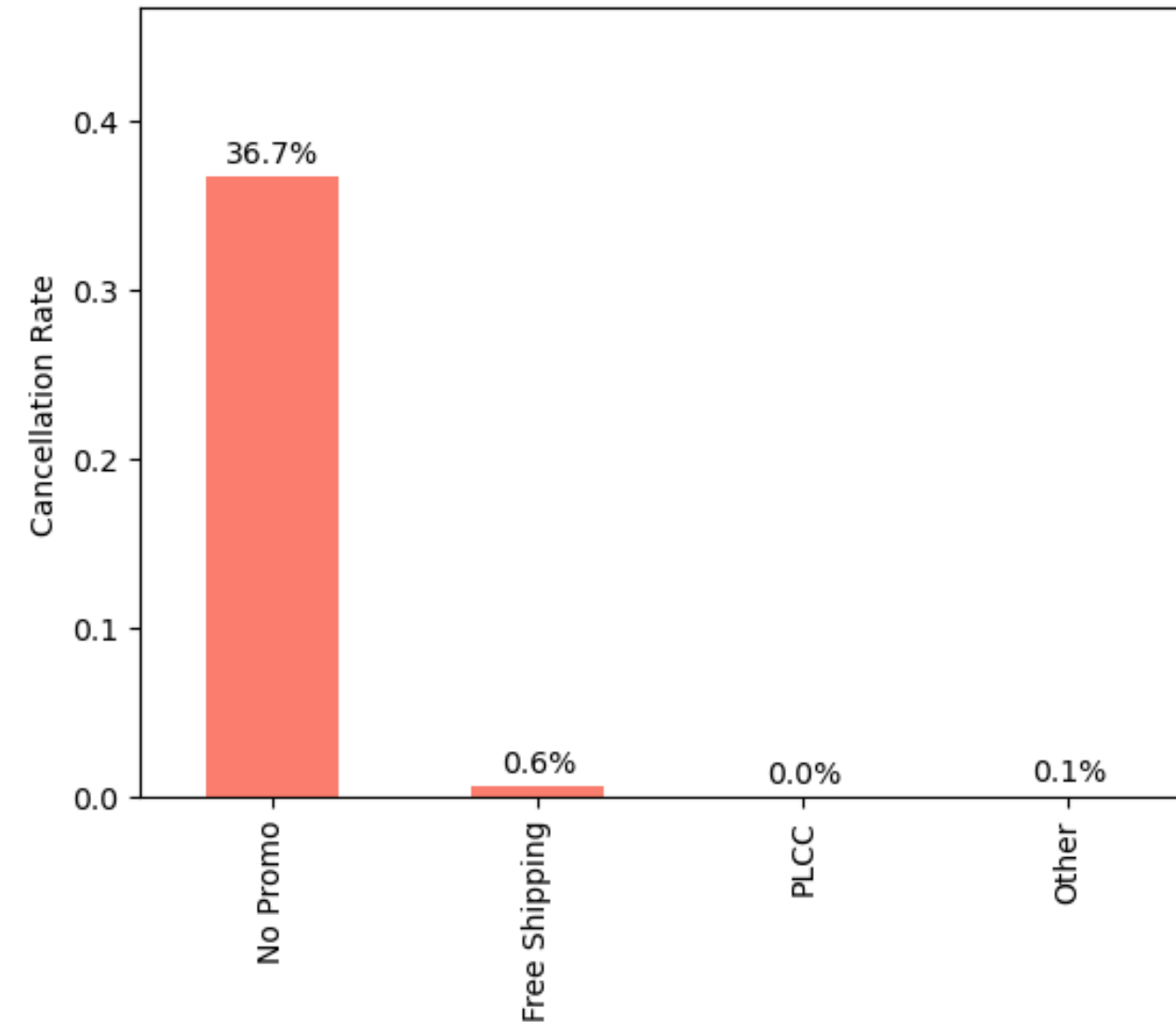
Cancellation Rate by Shipping Service Level



- Faster delivery is associated with lower cancellation likelihood
- Shipping speed is a relevant feature for cancellation prediction

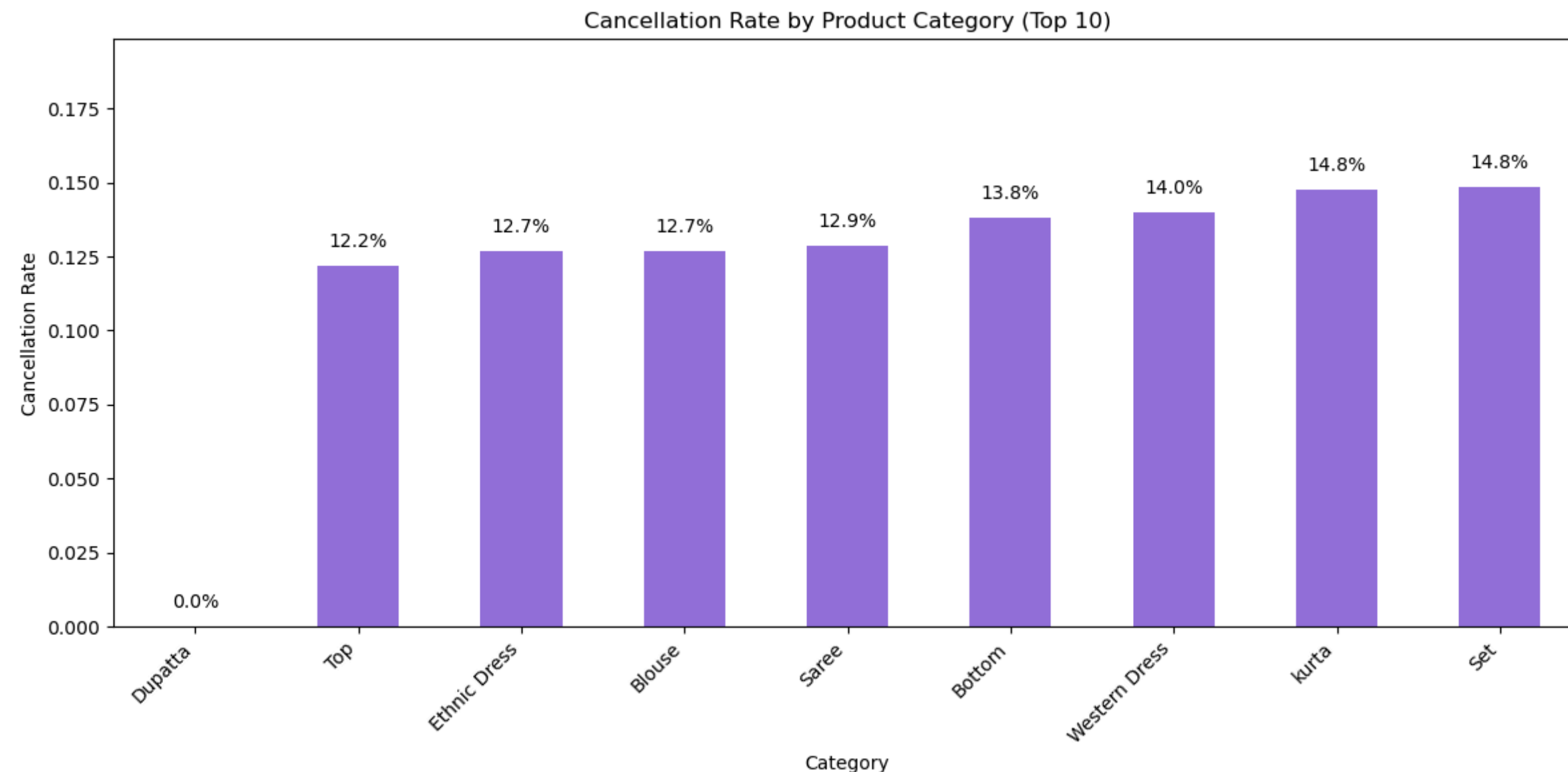


Cancellation Rate by Promotion Type



- Promotions significantly reduce the risk of order cancellation
- Promotion Type is a highly predictive feature

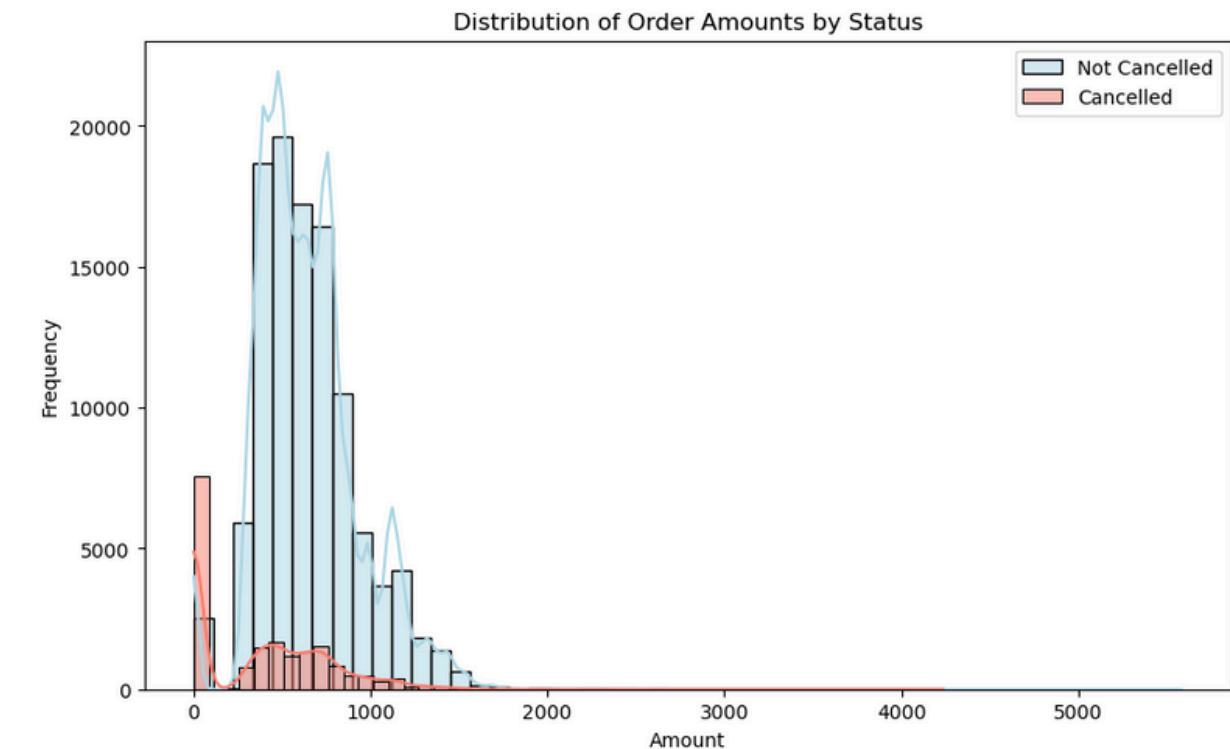
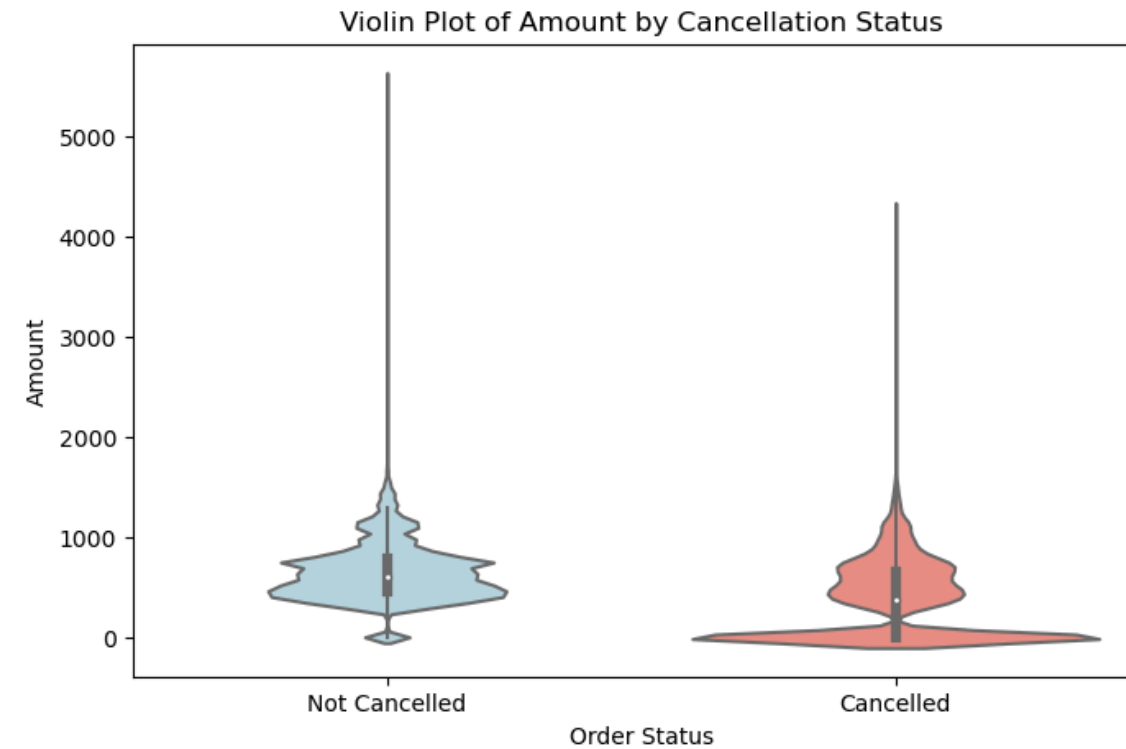
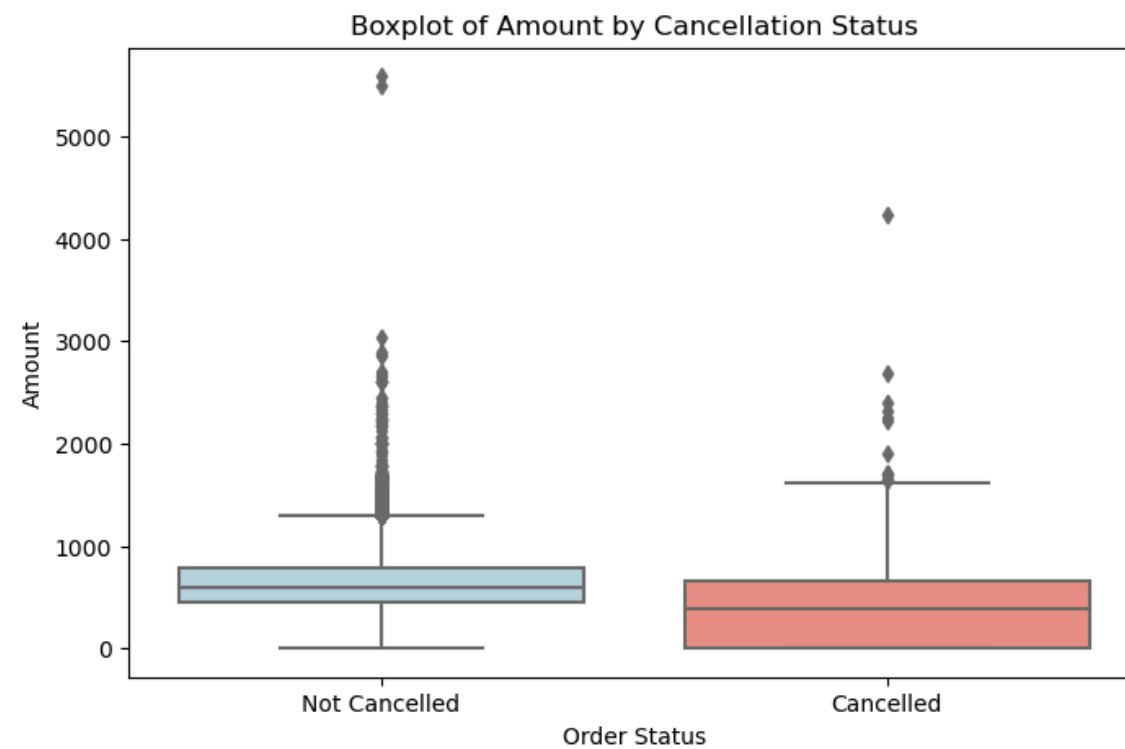
Exploratory Data Analysis



- Highest cancellation in Set and Kurta (14.8%) – also top-selling categories
- Suggests product type may influence cancellation behavior



Exploratory Data Analysis



- Cancelled orders are typically lower in value.
- Higher-value orders are less likely to be cancelled.
- While order amount alone may not fully explain cancellations, it helps enhance predictions when used alongside fulfilment type, promotion, and category.



Motivation

Problem
Statement

Dataset
Cleaning

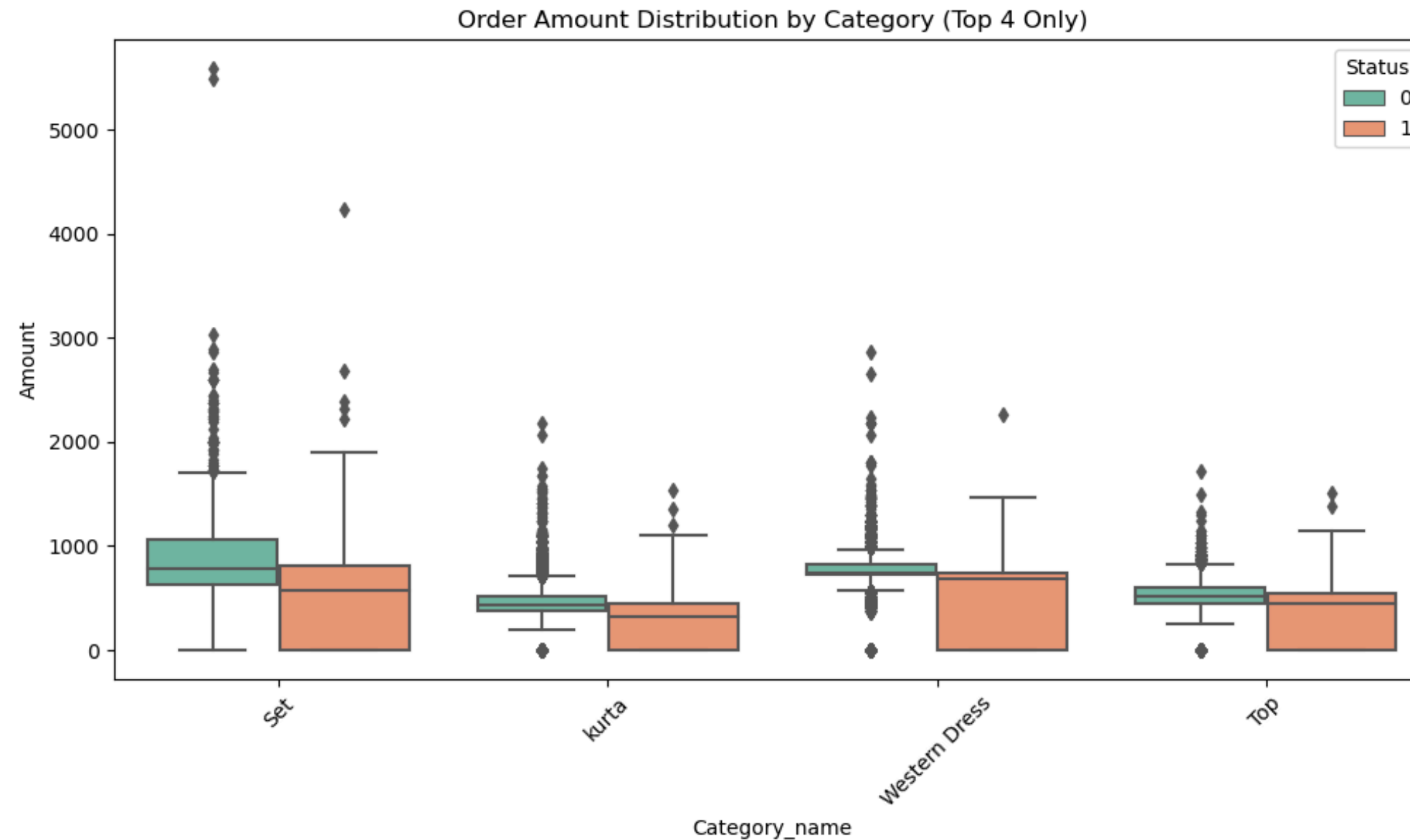
EDA

ML Models

Outcomes

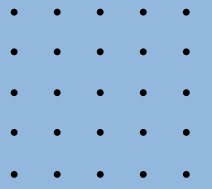
Conclusion

Exploratory Data Analysis



- In the top 4 categories, cancelled orders tend to be lower in value
- Suggests order amount + category interaction may be predictive
- These patterns are important for modeling cancellation risk





Machine Learning

- *Logistic Regression*
- *Random Forest*
- *eXtreme Gradient Boosting*



Motivation

Problem
Statement

Dataset
Cleaning

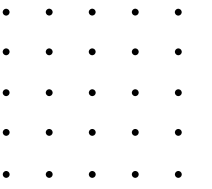
EDA

ML Models

Outcomes

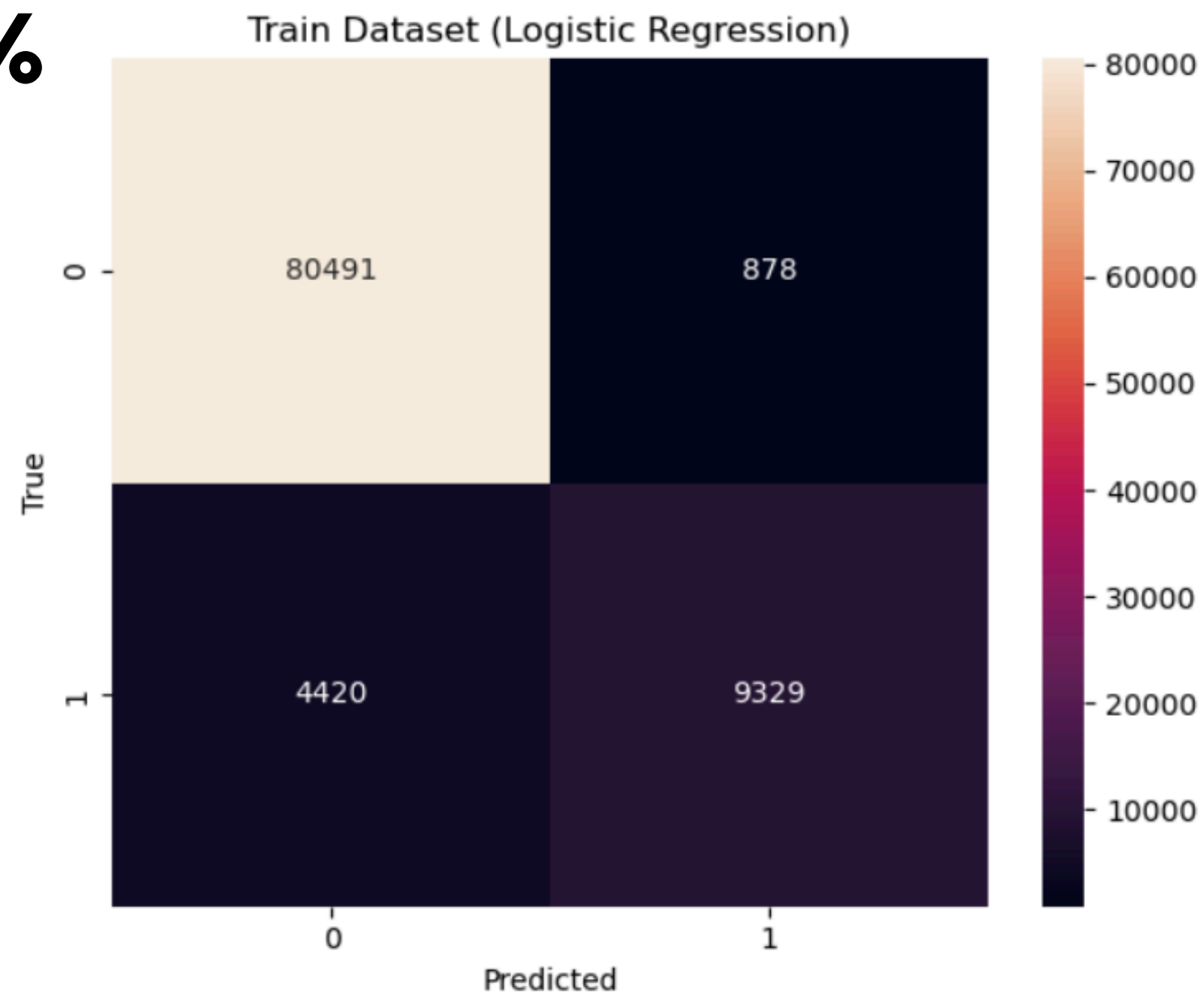
Conclusion

Logistic Regression

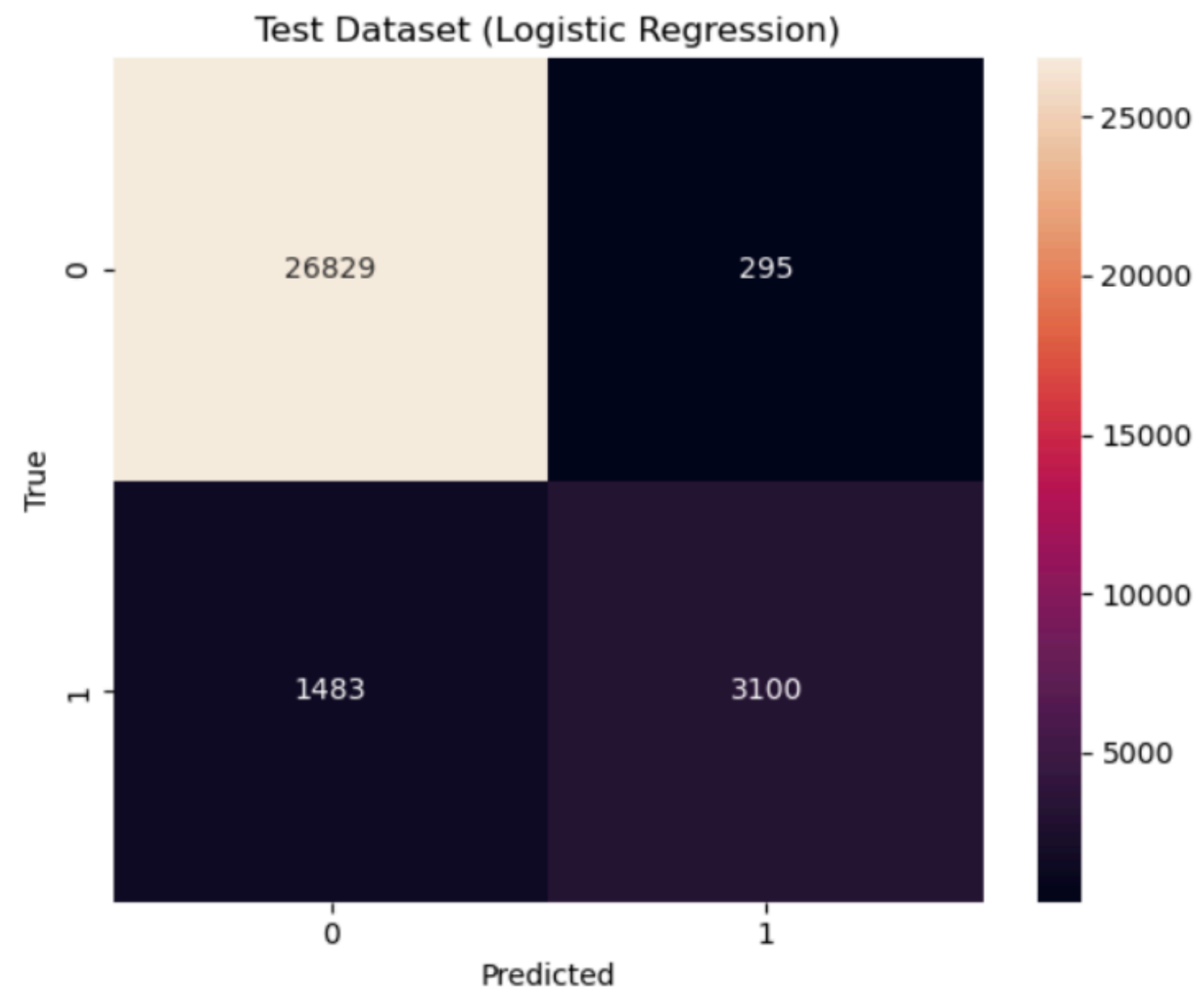


- Used as a **baseline model** to test if order features are predictive of cancellations
- Simple and interpretable: easy to understand the relationship between features and target
- Fast and efficient to train, especially on large datasets or early-stage analysis

75%



25%



Motivation

Problem
Statement

Dataset
Cleaning

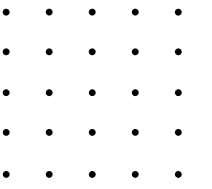
EDA

ML Models

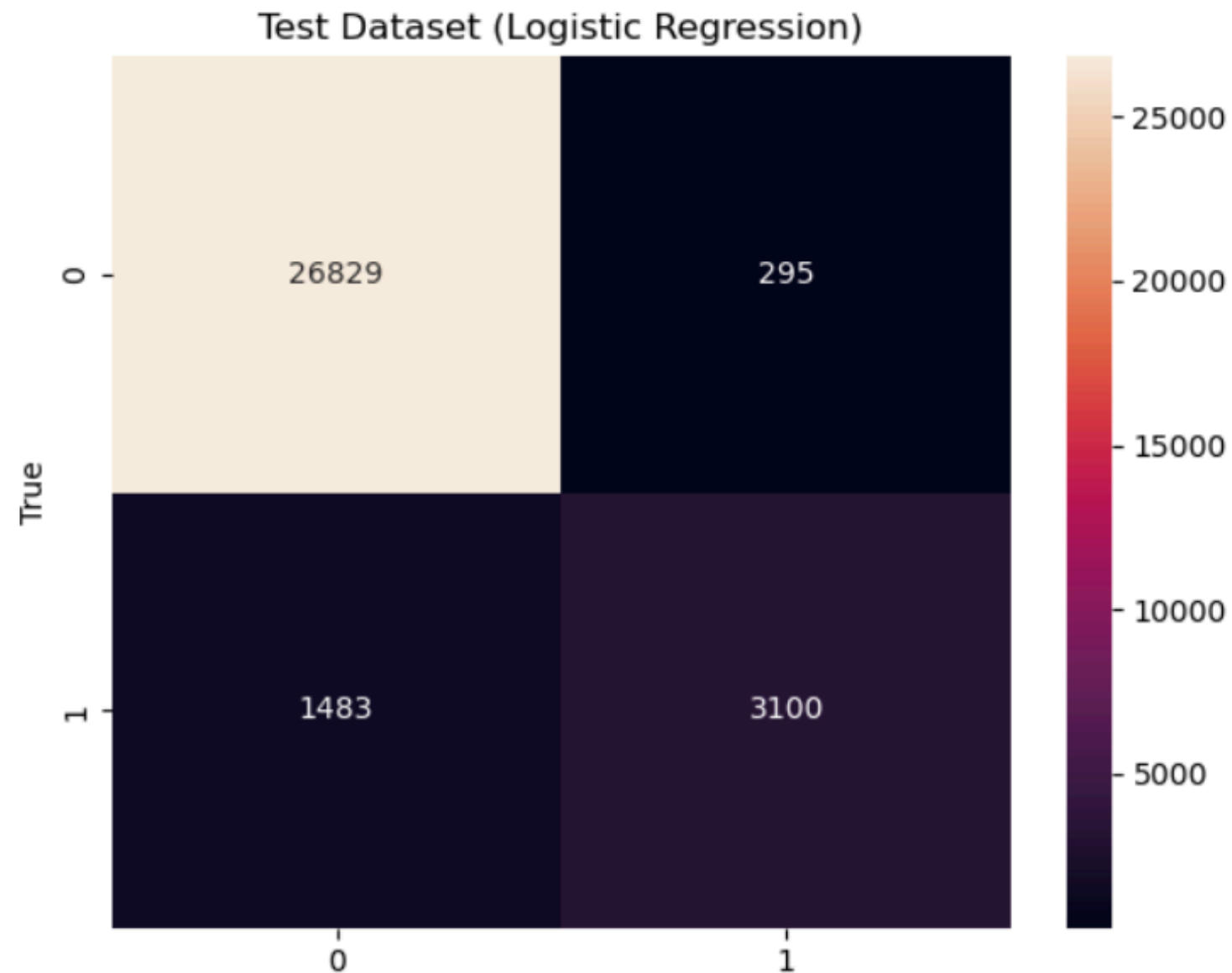
Outcomes

Conclusion

Logistic Regression



Goodness of Fit of Model: Test Dataset
Classification Accuracy : 0.9439
True Negative Rate : 0.9891
True Positive Rate : 0.6764
False Negative Rate : 0.3236
False Positive Rate : 0.0109



Accuracy: 94.39%

→ Overall prediction is highly accurate; the model performs well on general order classification.

True Negative Rate: 98.91%

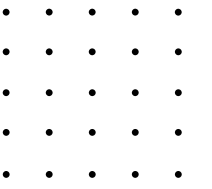
→ Excellent at identifying non-cancelled orders; very few "good orders" are mistakenly flagged.

True Positive Rate (Recall): 67.64%

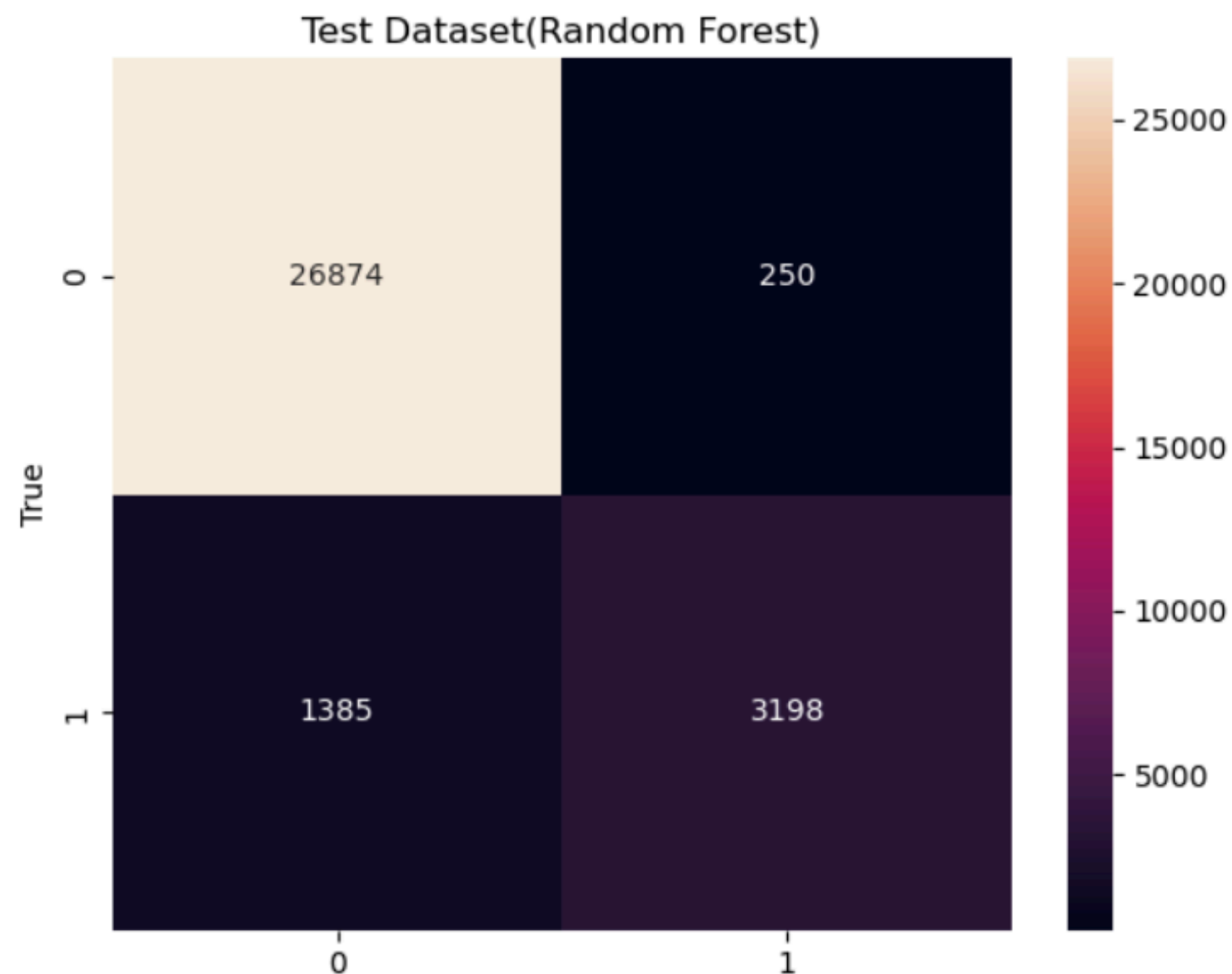
→ Correctly catches about two-thirds of cancelled orders, but misses ~32% of high-risk cases.



Random Forest



```
Goodness of Fit of Model: Test Dataset(Random Forest)
Classification Accuracy : 0.9484
True Negative Rate      : 0.9908
True Positive Rate      : 0.6978
False Negative Rate     : 0.3022
False Positive Rate     : 0.0092
```



Why Are the Results Similar?

Class Imbalance

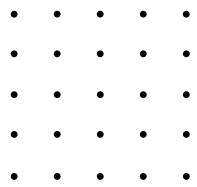
Our target variable is imbalanced:
most orders are not cancelled (Status = 0)

Default Classifiers Are Biased

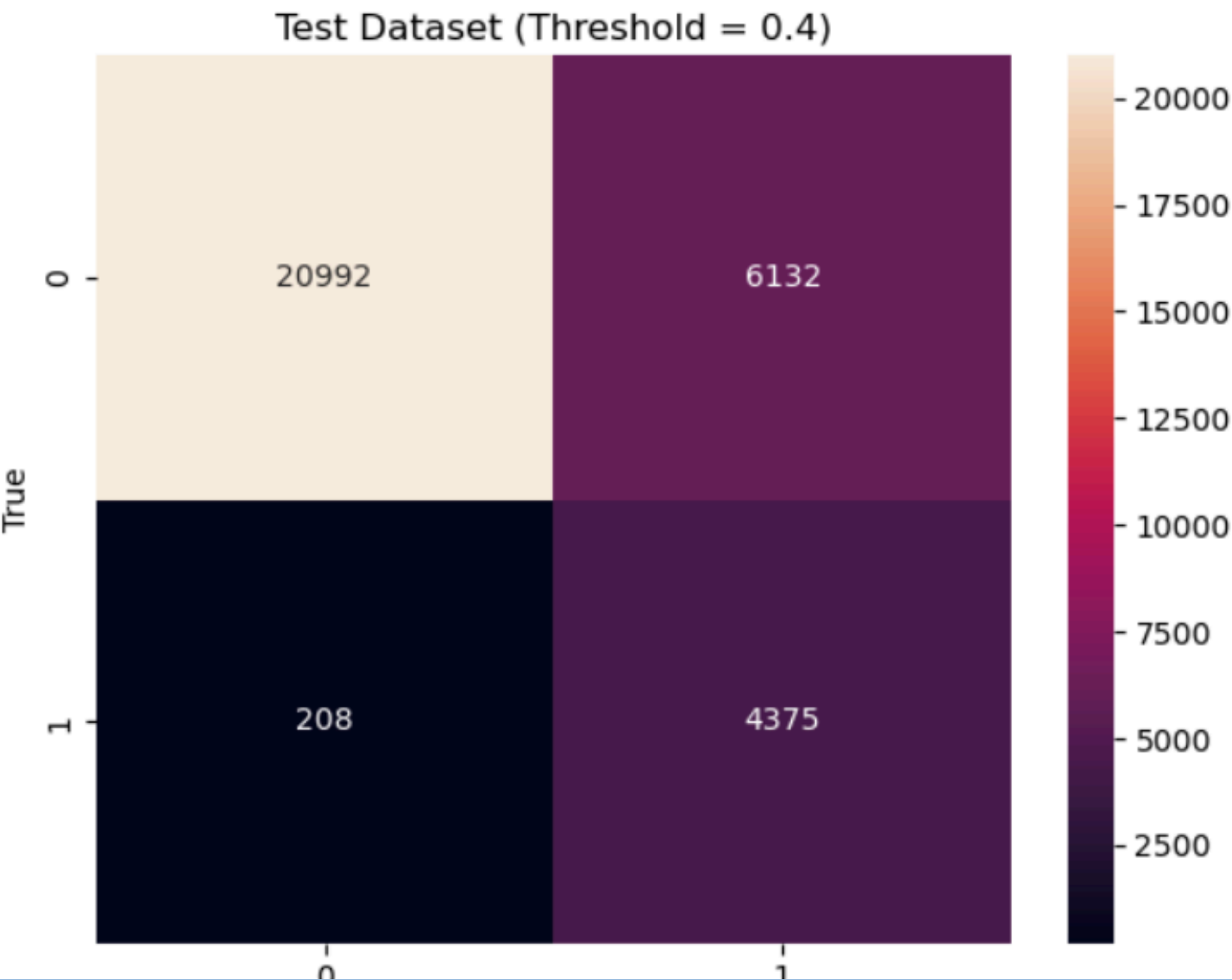
Both Logistic Regression and Random Forest focus on predicting the majority class to maximize accuracy.



Our Solution: Make the Model Sensitive



Goodness of Fit of Model: Test Dataset (Balanced, Threshold = 0.4)
Classification Accuracy : 0.8000
True Negative Rate : 0.7739
True Positive Rate : 0.9546
False Negative Rate : 0.0454
False Positive Rate : 0.2261



What Changed After Tuning?

Much Higher Recall

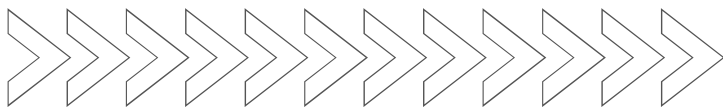
True Positive Rate improved from 69.8% → 95.5%
Model now catches almost all cancelled orders.

But at a Cost

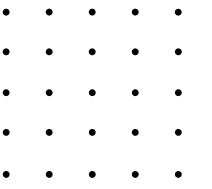
False Positive Rate increased from 0.9% → 22.6%
Meaning: More non-cancelled orders are mistakenly flagged.

Good for Risk-Averse Businesses

This setting is ideal if the cost of a missed cancellation is higher than the cost of a false alarm.



eXtreme Gradient Boosting



Why we tried XGBoost?

- Known for high performance in many real-world prediction tasks
- Handles complex, non-linear feature interactions

What we found?

Models	Random Forest	XGBoost
Accuracy	94.84%	94.77%
Recall	69.78%	68.89%
True Negative Rate	99.08%	99.14%

WHY?

- Data is structured & not very nonlinear
- Feature signals may already be captured by simpler models
- Random Forest is already a strong baseline

Model complexity ≠ better performance

Data matters more

XGBoost didn't lead to better results.



Motivation

Problem
Statement

Dataset
Cleaning

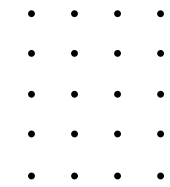
EDA

ML Models

Outcomes

Conclusion

What we learned?



Key Model Observations

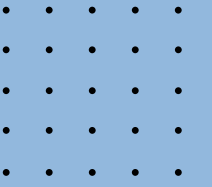
- Logistic Regression is simple but effective → good baseline
- Random Forest performed best even before tuning
- XGBoost didn't outperform → complexity ≠ better

Optimization Insights

- Tuning (threshold + class_weight) helps maximize recall
- Best model choice depends on business goals
- Imbalanced target affects all models → accuracy ≠ everything



Project Outcomes



Using Our ML Model, Sellers Can:

For risk-averse platforms

Use the **balanced model** with a lower threshold.

- Catches almost all cancelled orders
- Accepts some false alarms

where cancellation losses outweigh the cost of intervention

Example: **luxury** e-commerce platforms

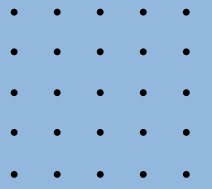
For cost-sensitive platforms

Use the **unweighted (default)** model

- Fewer false positives, more accurate overall
- May miss some cancellations

Ideal for cost-sensitive platforms, with limited resources for manual intervention

Example: **fast-moving** consumer goods



Conclusion

- *Data-Driven Insights*
- *Recommendations*



Motivation

Problem
Statement

Dataset
Cleaning

EDA

ML Models

Outcomes

Conclusion

Conclusion

Data-Driven Insights

High Order Amount → Lower Cancellation

Customers placing larger orders tend to cancel less often, indicating stronger purchase intent.

Fulfilment Method Matters

Merchant fulfilment shows higher cancellation rates—likely due to delays or service issues.

Promotion Usage Helps

Orders with promotions show lower cancellation rates, possibly driven by stronger incentives or urgency.

Recommendations

Predict and Prevent with RF-Model

Use a trained Random forest ML model to flag high-risk orders early and trigger proactive follow-up by customer service.

Improve Fulfilment Reliability

Optimize logistics for high-risk fulfilment types to reduce delays and cancellations.

Targeted Promotions for High-Risk Categories

Offer time-limited discounts or free shipping to reduce cancellations in vulnerable product categories.

Thank you

FR1 | Group 8

Gao Anni (N2402461C), Liu Chenyu (N2402421L)

