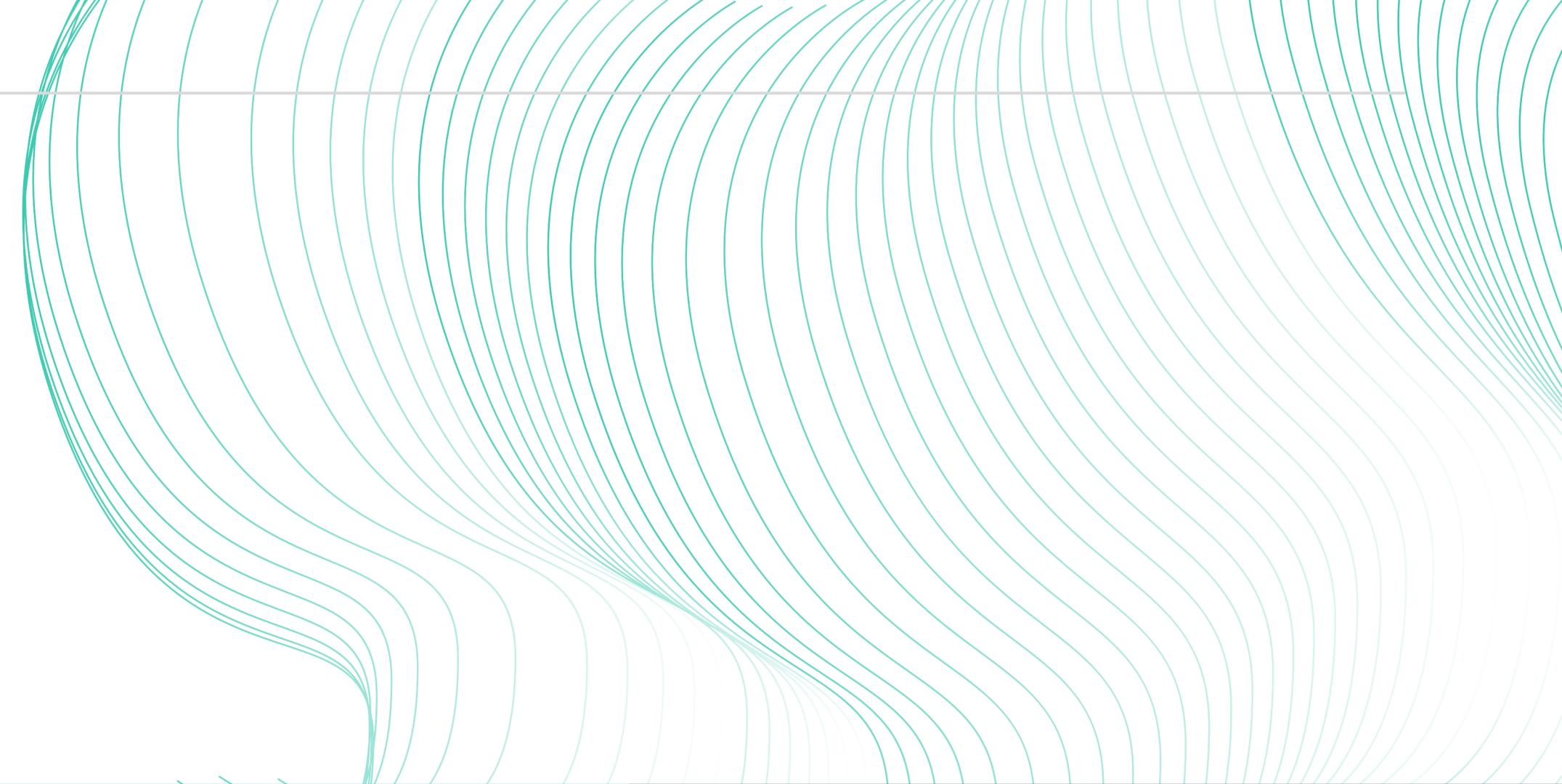


# Supply Chain Analysis

Ge Gao, Yanhuan Huang, Kellly Kao,  
Kojen Wang, Hongyi Zhan, Rebecca Zhang

# Supply Chain Analysis

Make informed strategic decisions for inventory management, order fulfillment, and customer service.



## Demand Prediction

Optimize Inventory Management  
Improve Production Planning  
Reduce Lead Times  
Impact on Sustainability

## Fraud Detection

Prevent Financial Loss  
Preserve Brand Reputation  
Improve Operational Efficiency  
Enhance Customer Confidence

## Fraud Order Clustering

Mitigate Potential Risk  
Adapt Fraud Prevention Strategies  
Stramline investigation Process

# Dataset Overview

We obtained this dataset from Kaggle, and it's originally sourced from the company DataCo Global.

## SupplyChainDataset (180,519 rows, 53 columns)

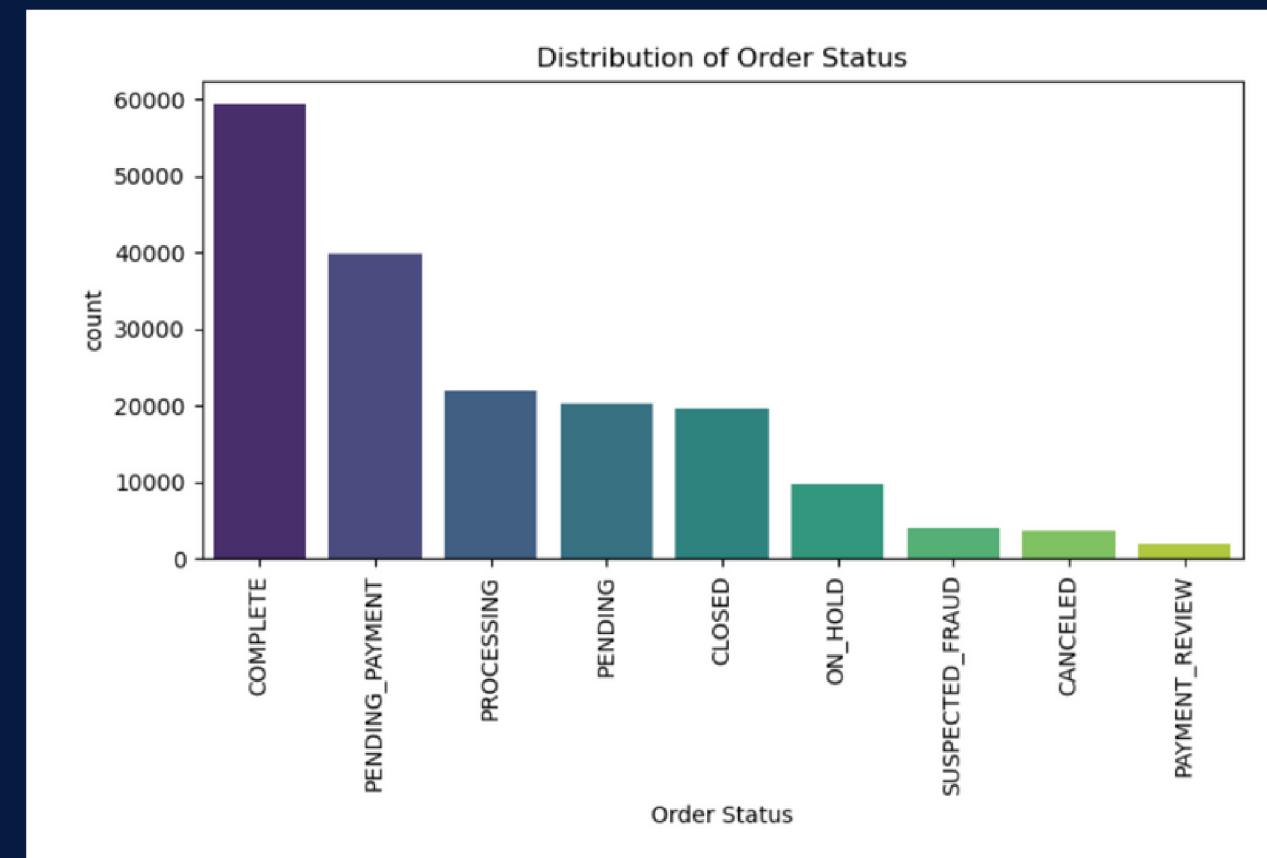
Contains transaction records, customer data, product information, and logistics.

**Key variables:** order IDs, product categories, quantities ordered, customer locations, payment methods, and much more.

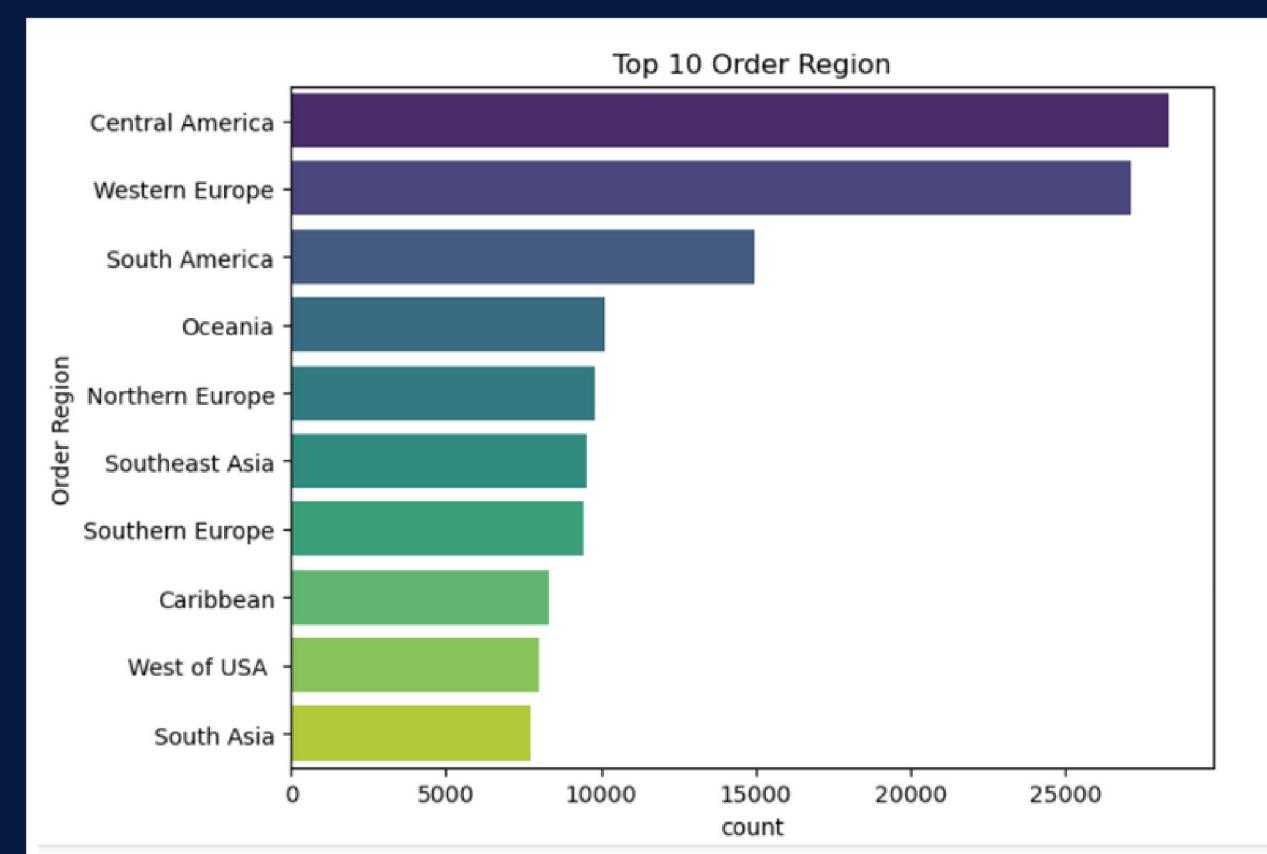
**20,652**  
Customers

**65,752**  
Orders

**118**  
Products



Order status distribution-  
healthy completion rate

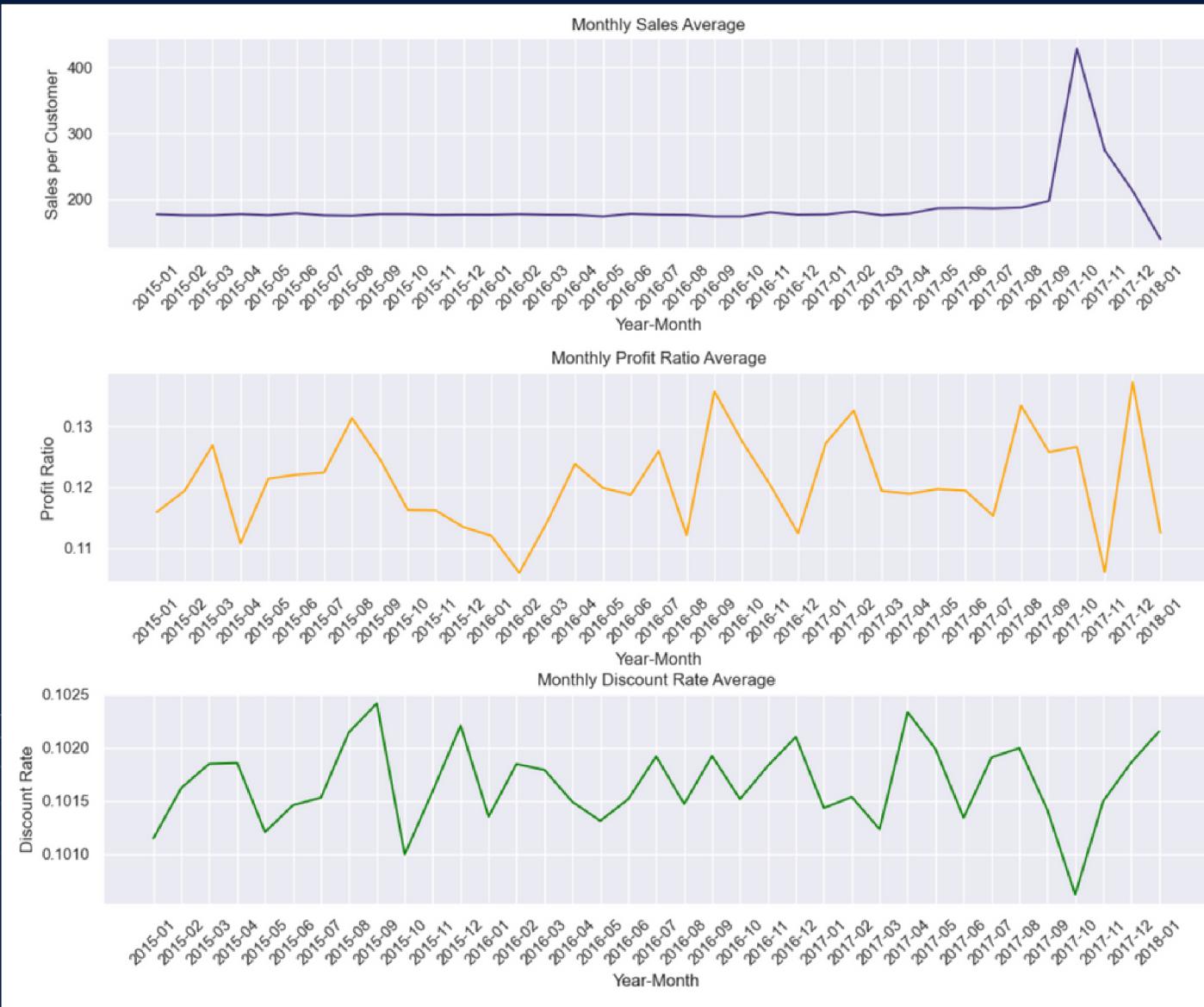


Regional sales volume  
Central America a strong market

For:  
Strategic planning  
Targeting areas for market penetration  
Managing fraud prevention measures

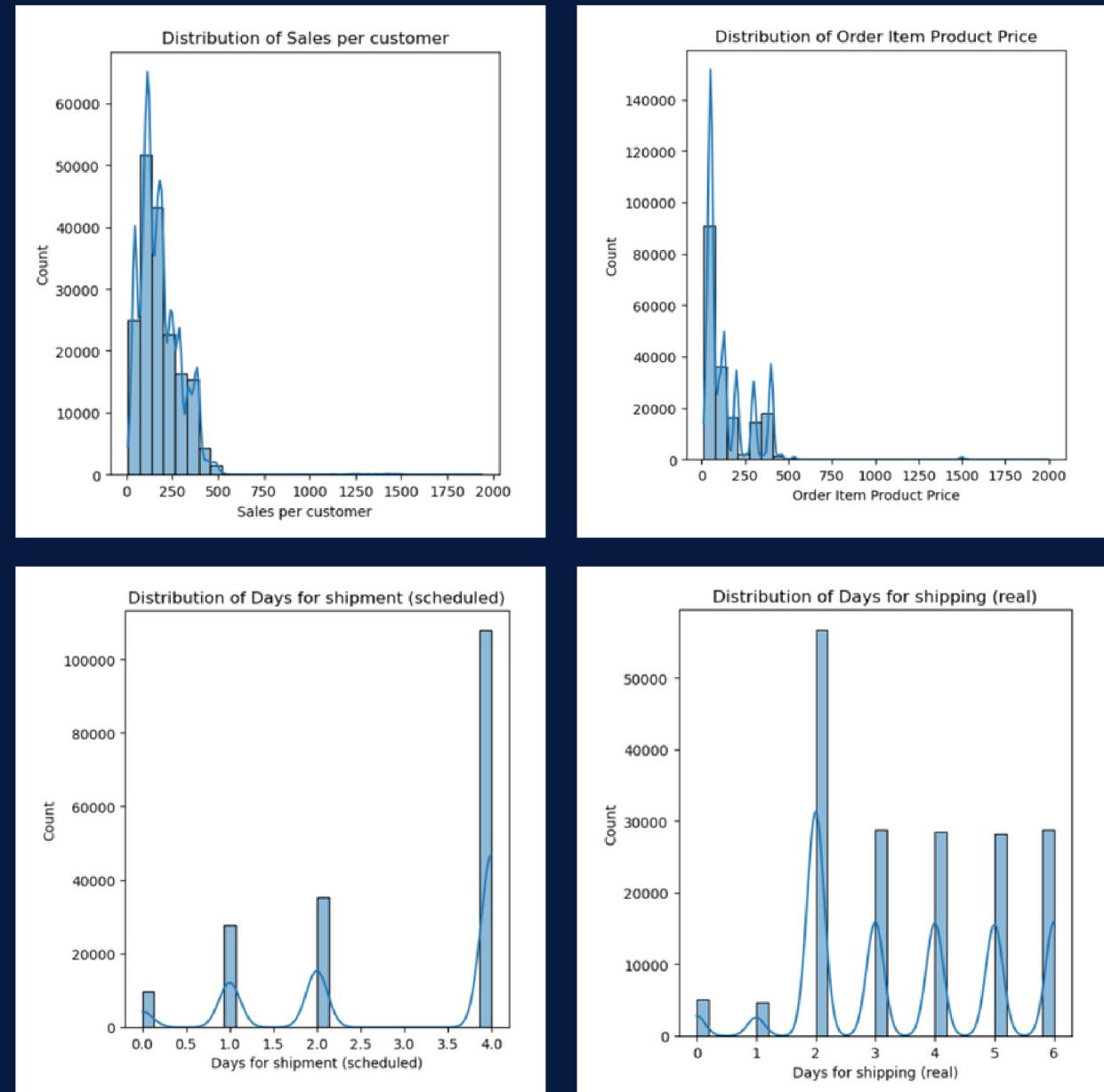
# Exploratory Data Analysis

## Monthly Business Performance Metrics



- Sales per customer are stable, with an exceptional increase in Oct 2017
- The profit ratio follows a seasonal trends, could be the result of business activities like promotions, changes in cost structure.
- Discount rates vary significantly, suggesting active promotional or pricing adjustments.

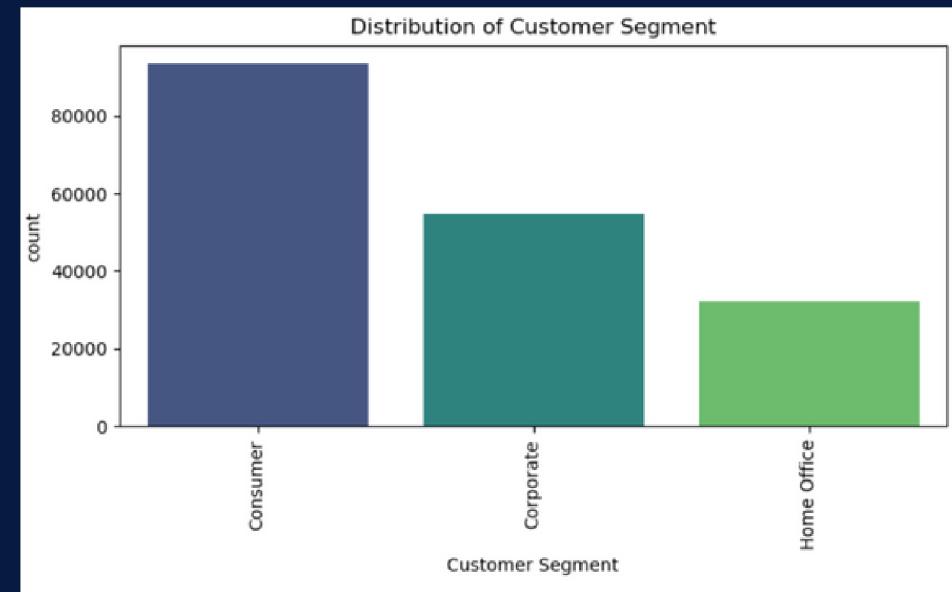
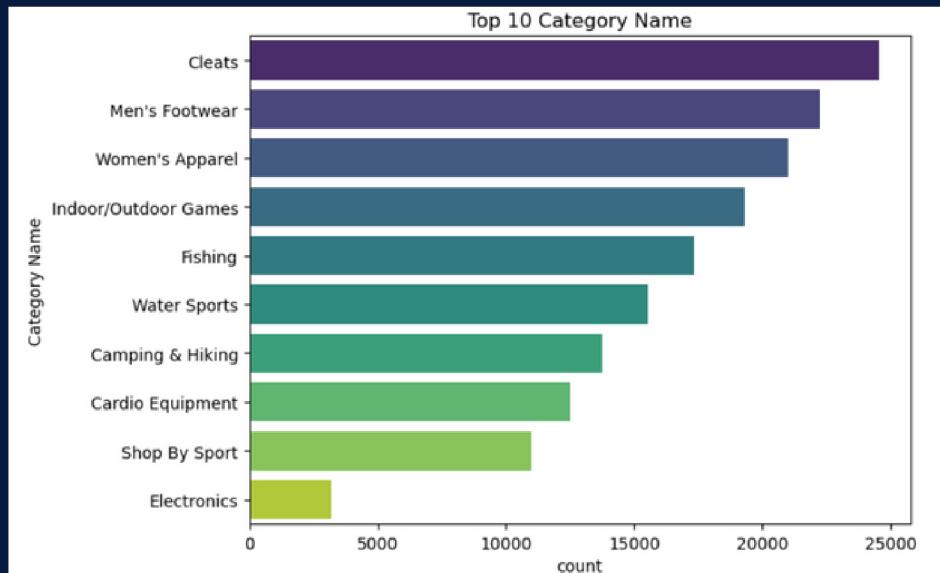
## Overview of Customer Sales and Shipping Dynamics



- Sales per customer and order item prices are skewed towards lower values, business model may focus on high volume sales of lower-priced items.
- The discrepancy between the scheduled and actual shipping times might suggest inefficiencies or unexpected delays in the shipping process.

# Exploratory Data Analysis

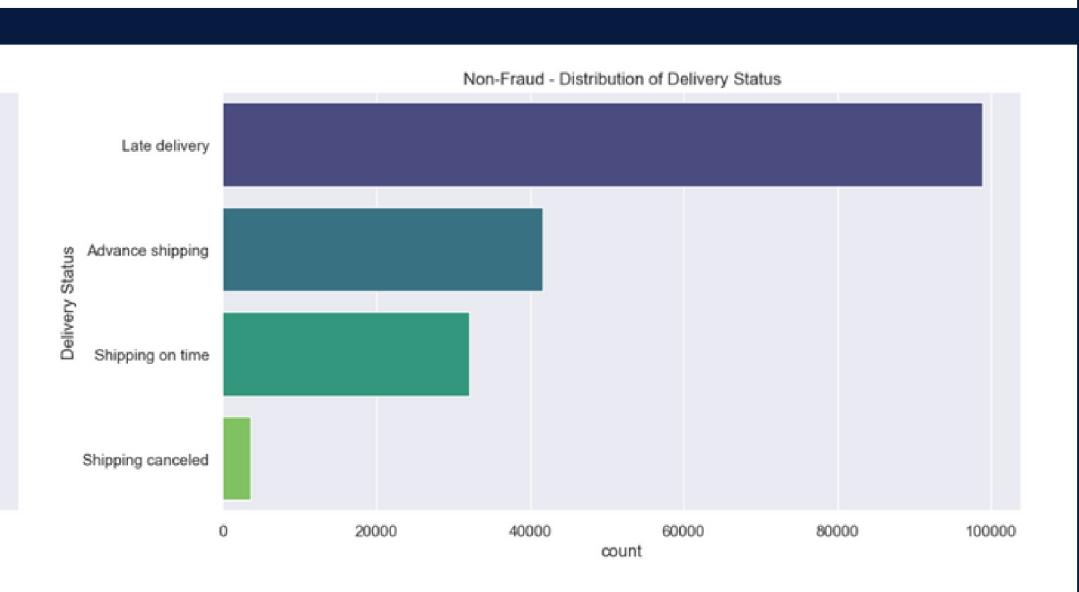
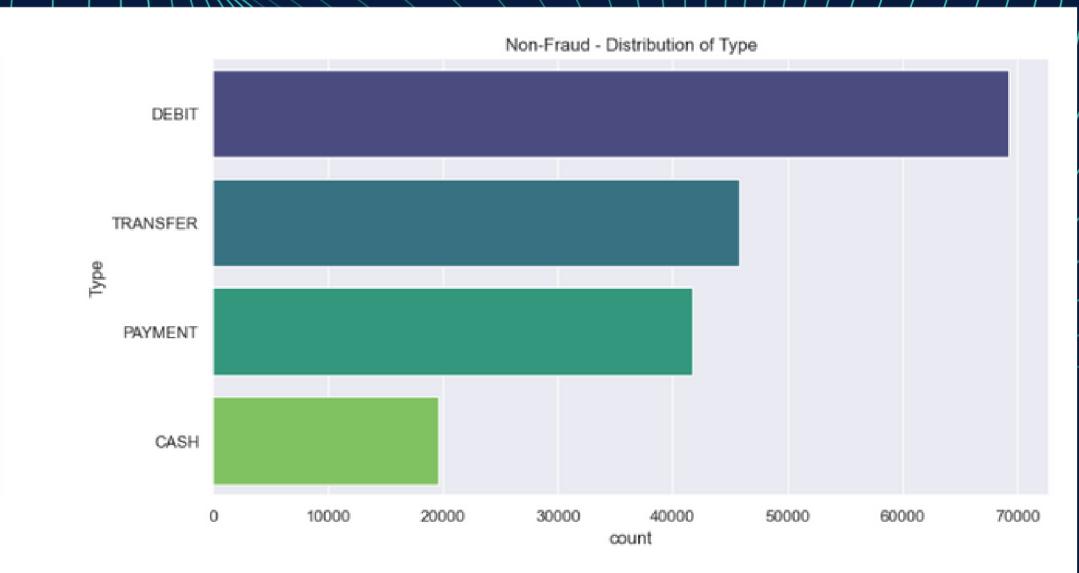
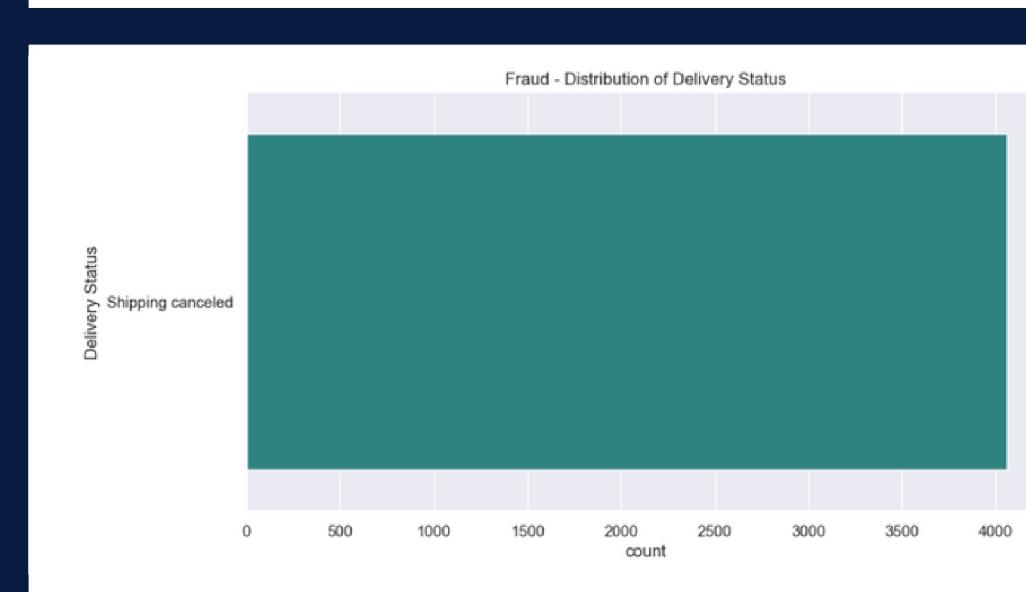
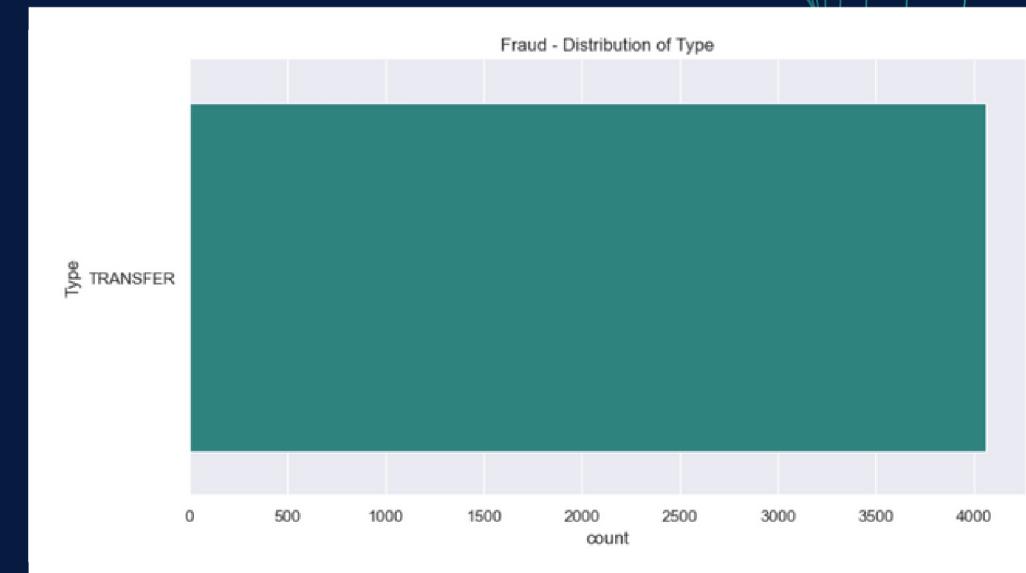
## Product Category & Customer Segment



- Popular categories: sports apparel and equipment, less popular: electronics
- The 'Consumer' segment is predominant, which may influence marketing strategies, product offerings, and sales efforts.

## Fraud Order v.s. Non-Fraud Order

Please visit our [github](#) for more detailed analysis!



- 'TRANSFER' is a common transaction type in fraudulent activities, whereas 'DEBIT' is the most common type for non-fraudulent transactions.
- Fraudulent transactions predominantly result in 'Shipping canceled', which is a logical step to mitigate loss due to fraud; Non-fraudulent transactions are mostly 'Late delivery'

# Scenario Overview

## Demand Prediction

### Data Preprocessing

Missing values, Feature Engineering, One hot encoding, Standardization

### Aggregation to get predictors and target

X: sales, shipping days, product price...  
y: monthly order quantity

### Feature Selection

By Random Forest

### Model and Hyperparameter Tuning

LSTM, RF, GB, LR

## Fraud Detection

### Data Preprocessing

Feature Engineering, Label encoding, Standardization

### Handle Imbalanced Dataset

Oversampling fraudulent orders

### Feature Selection

By Recursive Feature Elimination (RFE)  
Selected features: type, order hour, latitude..

### Model and Hyperparameter Tuning

Ensemble methods-voting (LR, RF, SVM)  
XGBoost, LightGBM

## Fraud Order Clustering

### Data Preprocessing

Filter fraud, Drop unary variables and correlated features, one hot, Standardization

### Decide the number of clusters

Calculate inertia

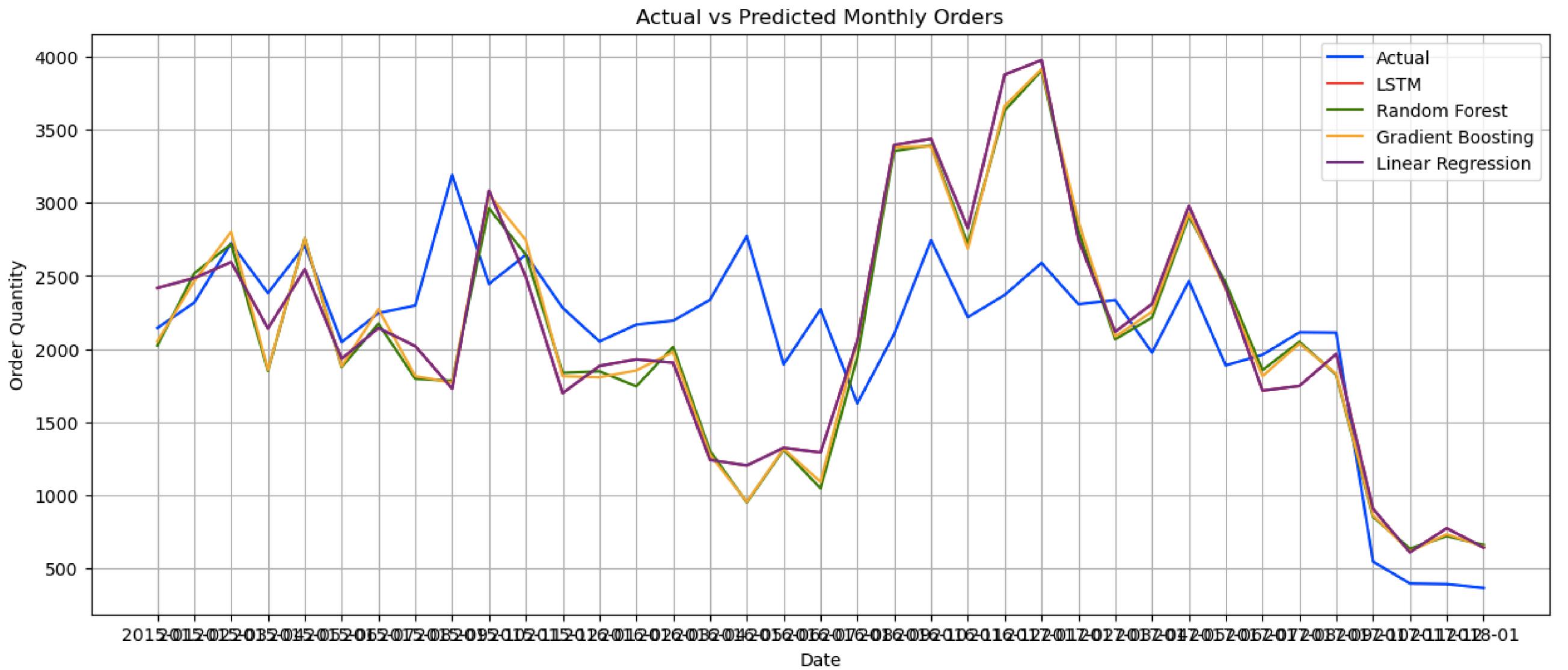
### KMeans Clustering

Inverse transform the cluster centroids and visualize features of each cluster

### DBSCAN

Obtain 6 clusters and identify the noise

# Demand Prediction - Result



Model	MSE	MAE
LSTM	132.55	4.7
RF	19.42	<b>0.48</b>
GB	<b>9.05</b>	1.08
LR	132.55	4.7

## Random Forest and Gradient Boosting Hyperparameters

- the number of trees: 30, 50, 70, **90**
- maximum number of features when splitting a node: 5, 10, 15, **20**, square root, log2
- minimum number of samples to split: 1, **50**, 100, 150, 200

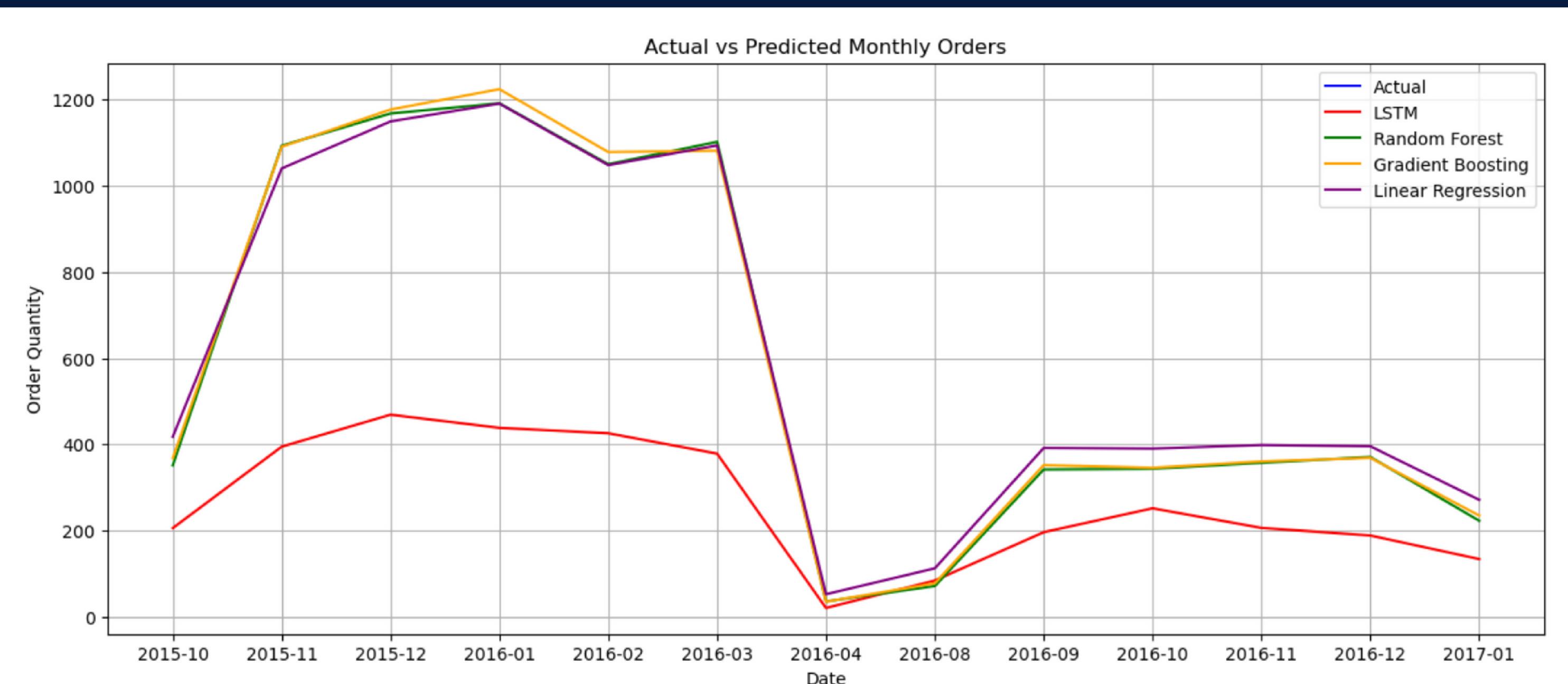
- **Random Forest** has the smallest **MAE**
- **Gradient Boosting** has the smallest **MSE**

# Demand Prediction Application

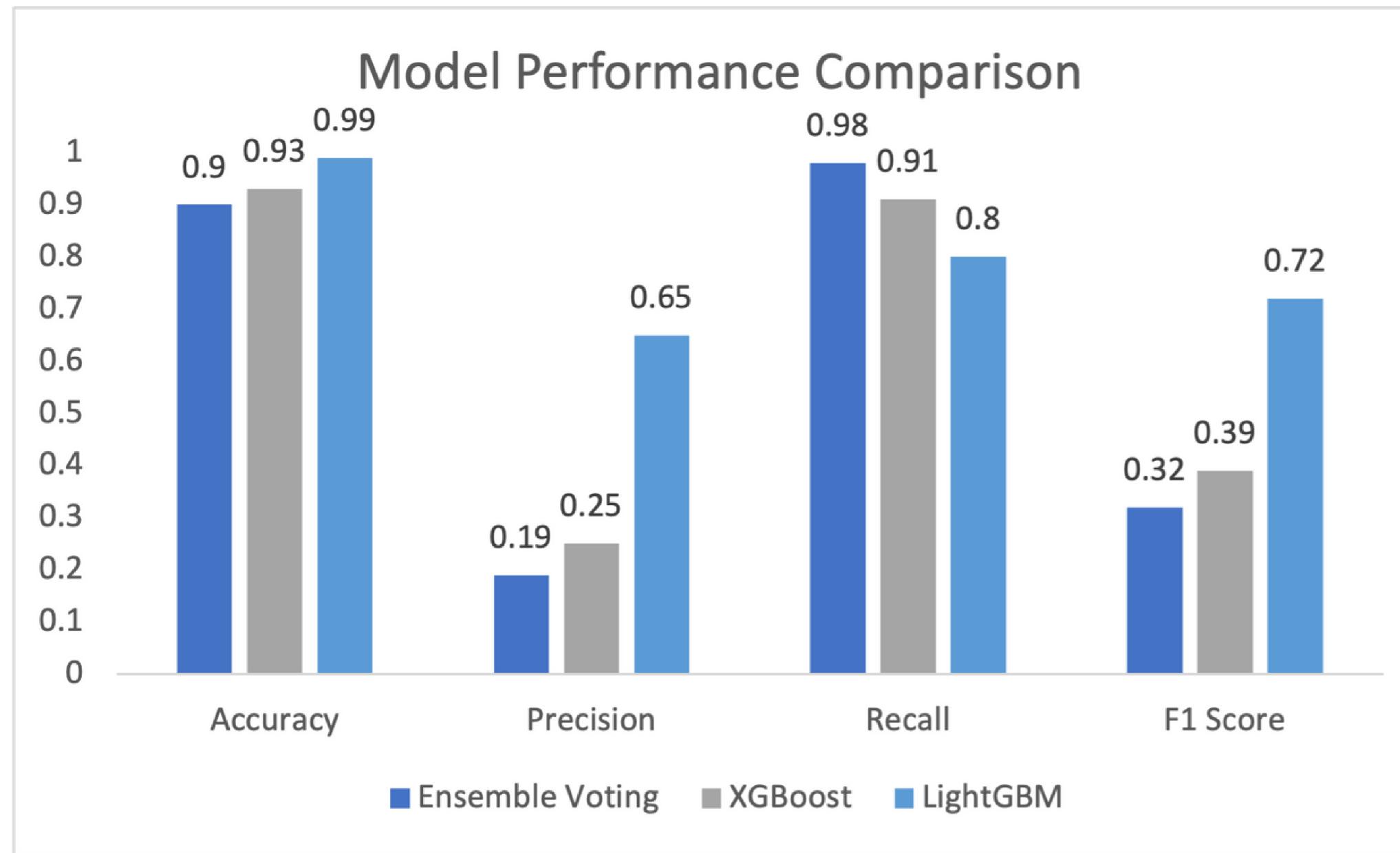
The models can be used to predict demand for specific product category and region

Product: Cardio Equipment

Region: Pacific Asia



# Fraud Detection - Result



### XGBoost Hyperparameters

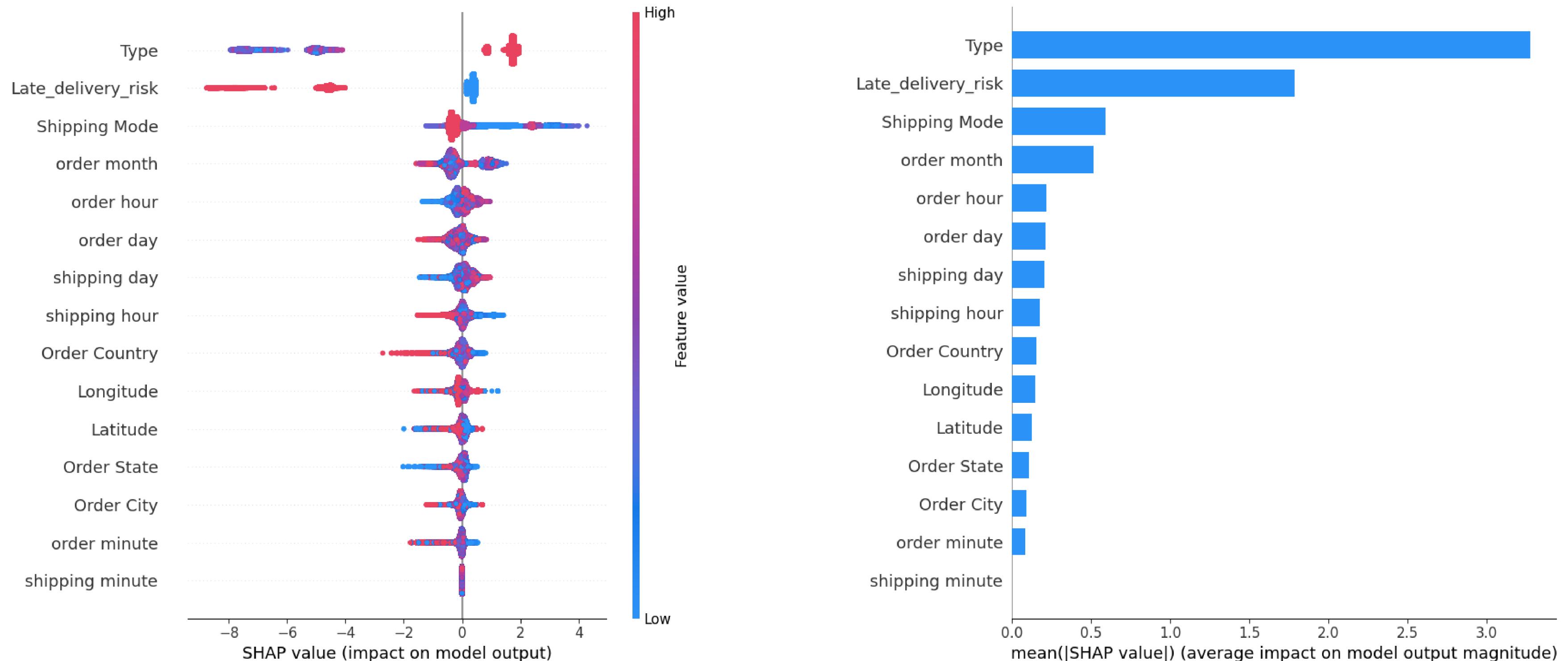
- maximum depth of a tree: 3, 4, **5**
- learning rate: **0.1**, 0.01
- the number of trees: 100, **200**

### LightGBM Hyperparameters

- maximum number of leaves: 31, **50**
- learning rate: **0.1**, 0.01
- the number of trees: 100, **200**

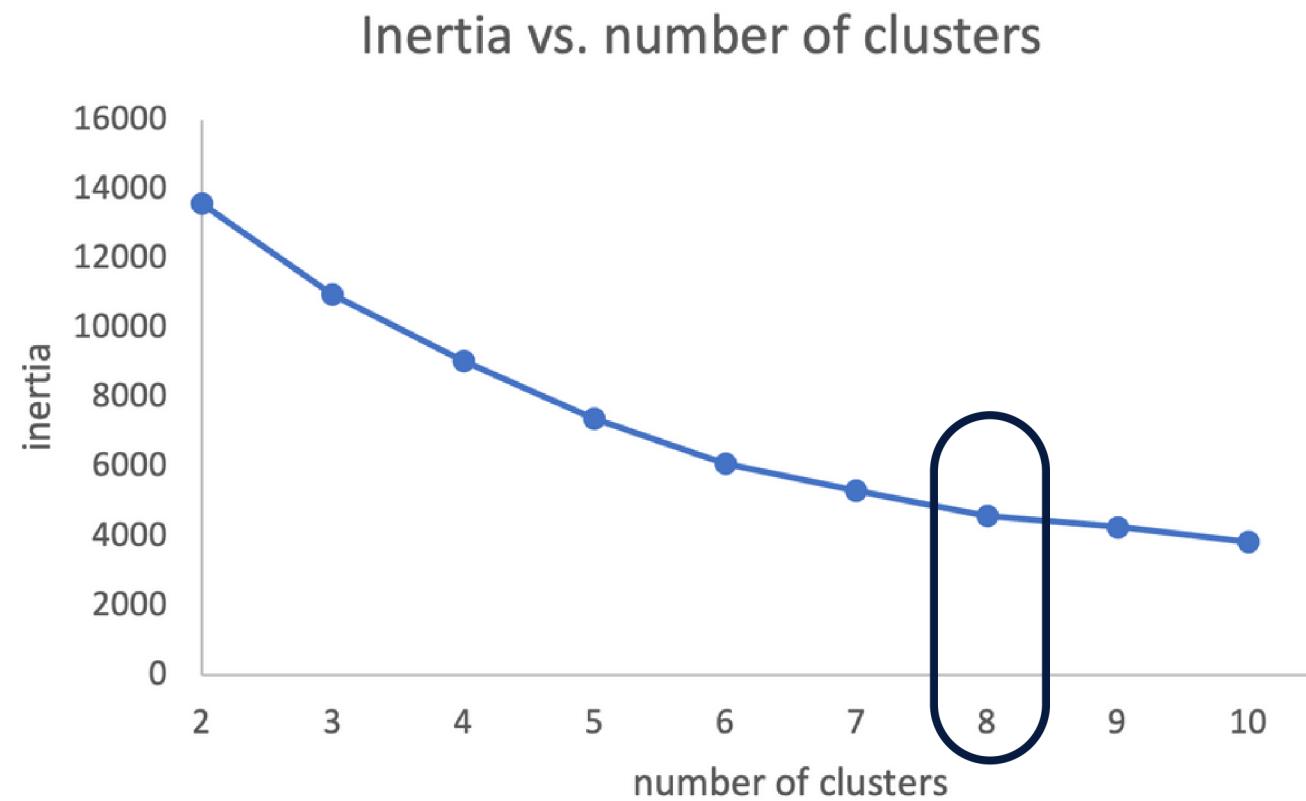
- **LightGBM has the highest accuracy (0.99), precision (0.65) and f1 score (0.72)**
- Ensemble voting (from LR, RF, SVM) has the highest recall (0.98)
- Flagging real customer orders as fraudulent orders (false positive) damages the customer-business relationship, **prioritize precision!**

# Fraud Detection - Feature Importance



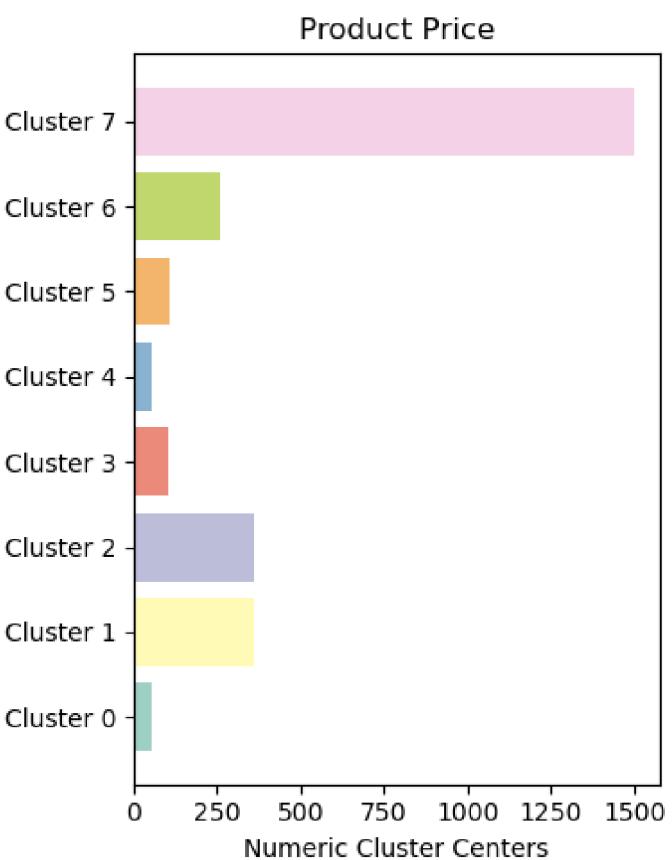
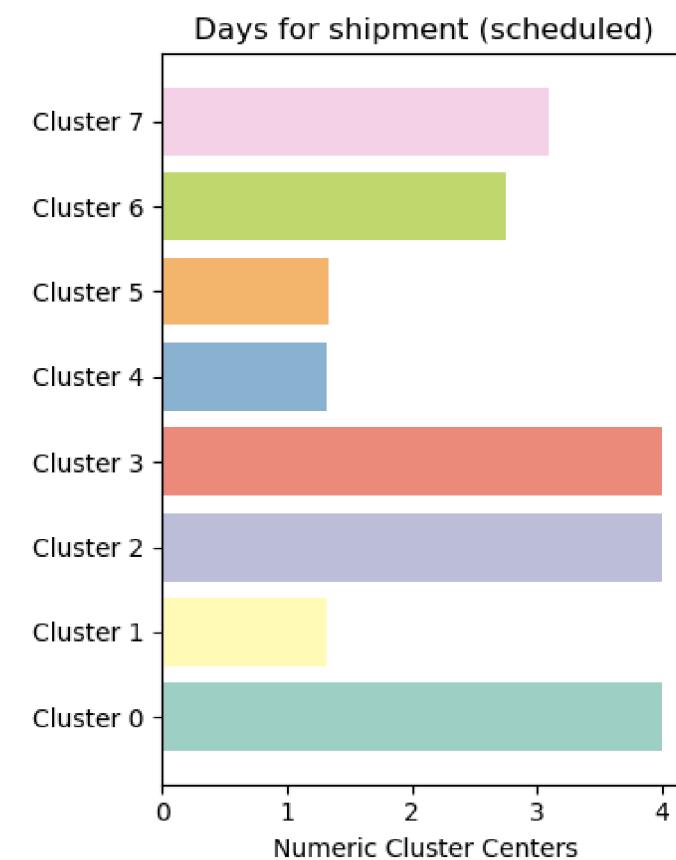
**Type of transaction** (debit, transfer, payment, cash), **late delivery risk, shipping mode** (standard, second, first, same day), **order month** are important features detecting fraudulent orders

# Fraud Order Clustering

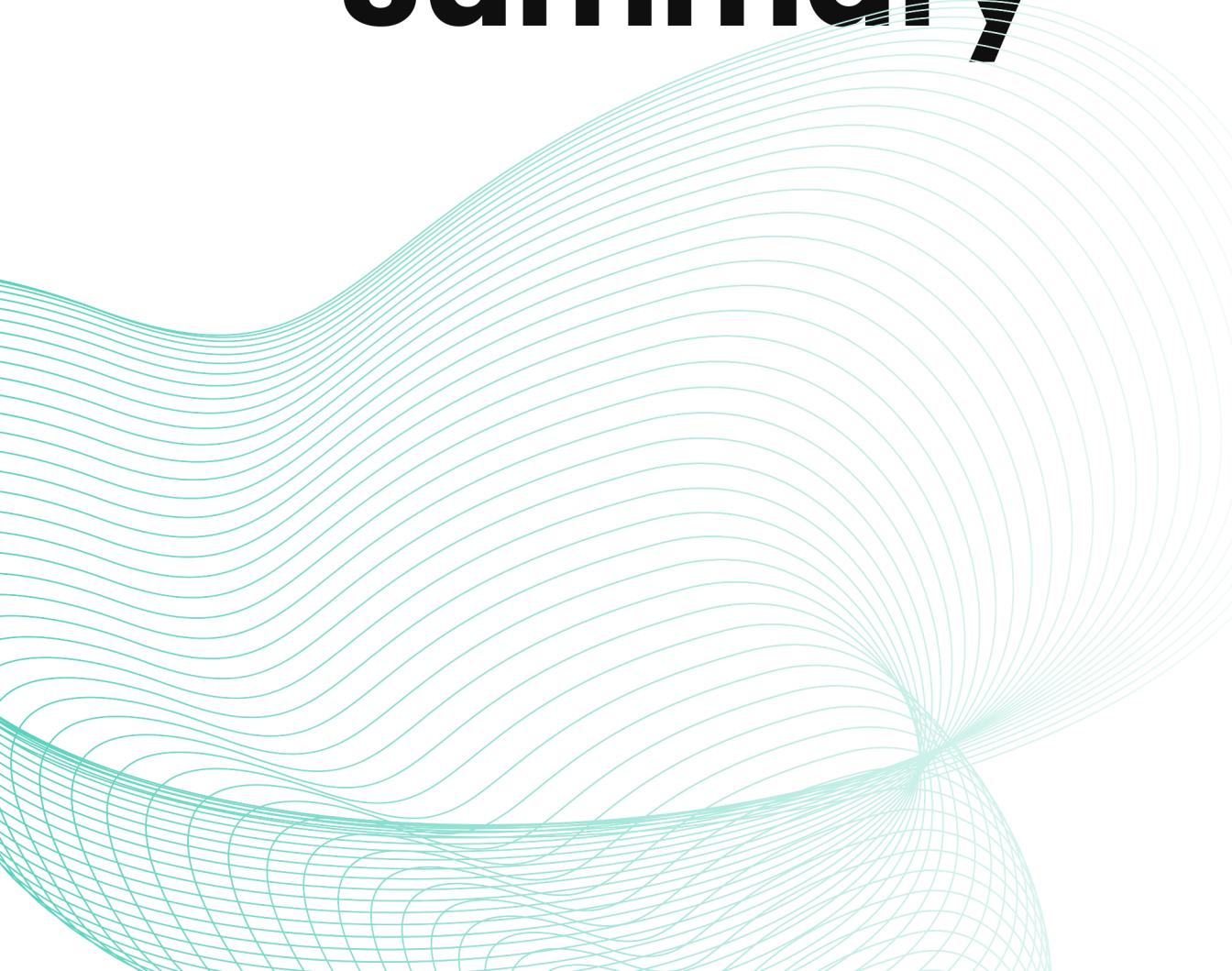


Cluster 0: High Volume,  
 Moderate Profit  
 Cluster 1: Low Volume, Quick  
 Delivery  
 Cluster 2: Low Volume, First  
 Class Shipment  
 Cluster 3: Low Volume, Low  
 Profit

Cluster 4: High Volume,  
 Moderate Profit, Quick Delivery  
 Cluster 5: Low Volume, Low  
 Profit, Quick Delivery  
 Cluster 6: Moderate Volume,  
**Negative Profit**  
 Cluster 7: Low Volume, High  
 Profit and Price



# Summary



## Exploratory Data Analysis

- Discrepancy in scheduled vs. actual shipping times suggests process inefficiencies
- Type of transaction - transfer is common in fraud, debit is common in non-fraud
- Fraudulent transactions often result in shipping canceled while non-fraud transactions frequently face late delivery, indicating logistics or inventory issues.

## Fraud Detection

- LightGBM had an accuracy of 0.99 and a precision of 0.65, which performed best among 3 models
- Type of transaction, late delivery risk, shipping mode, order month are important features detecting fraud
- Fraud detection prevents financial loss, and further improve the prediction of demand prediction

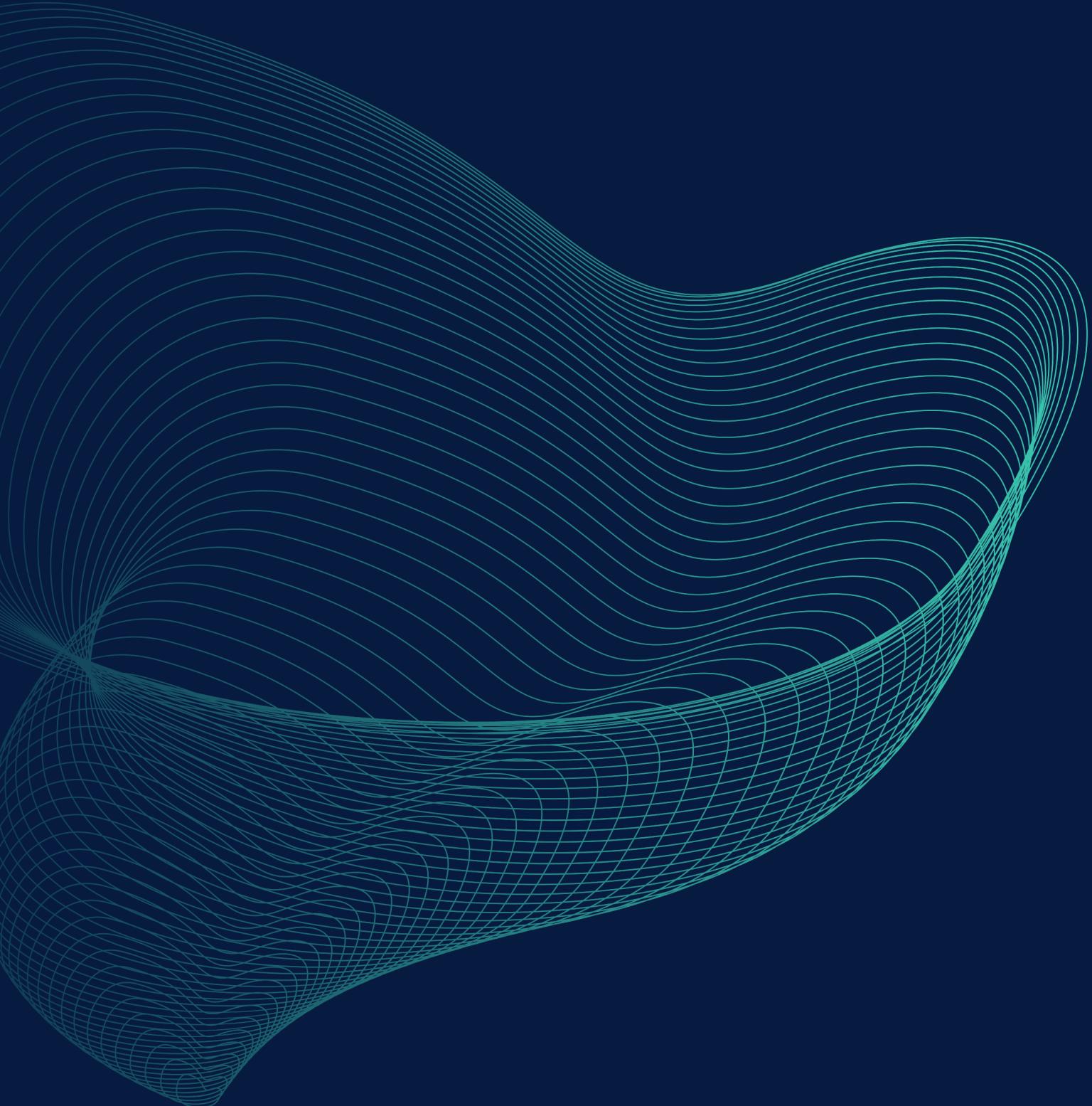
## Demand Prediction

- Gradient Boosting performed the best among 4 models with MSE equal 9.05
- The prediction model can be used to predict demand for specific product category and region
- Accurate prediction can help product planning and improve inventory management

## Fraud Order Clustering

- KMeans clustering divided fraudulent orders to 8 clusters while DBSCAN to 6 clusters
- By clustering, we can analyze the patterns of fraud orders. While fraudsters constantly evolve their strategies to bypass detection mechanisms, the model can adapt to new patterns by updating the model with new data

---



Thank you!