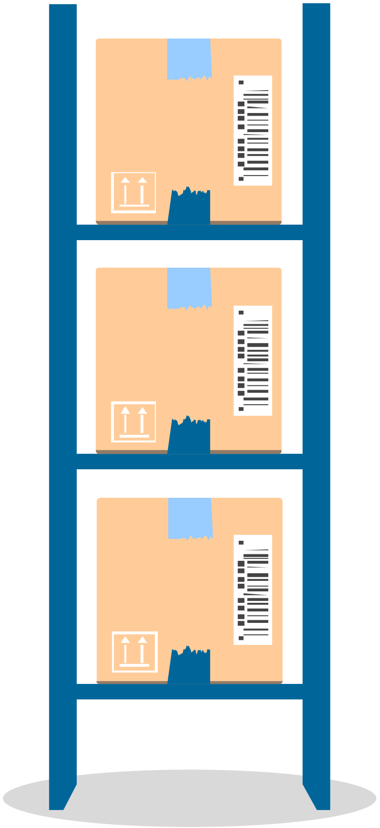
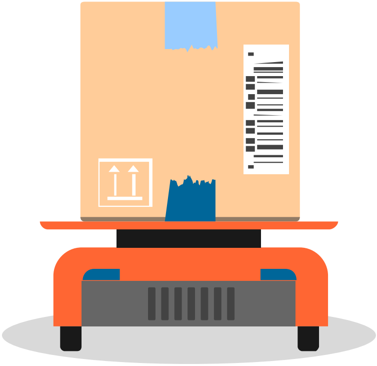


ENHANCING E-COMMERCE EFFICIENCY **WITH MACHINE LEARNING**

NEURAL NET NINJAS

AILEEN ALVAREZ, JENNIFER ALVAREZ, JESSICA ANDRAS, VICKY LASOTA



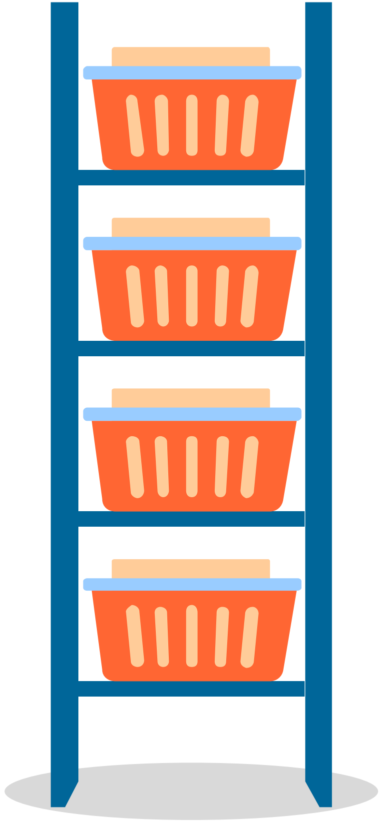
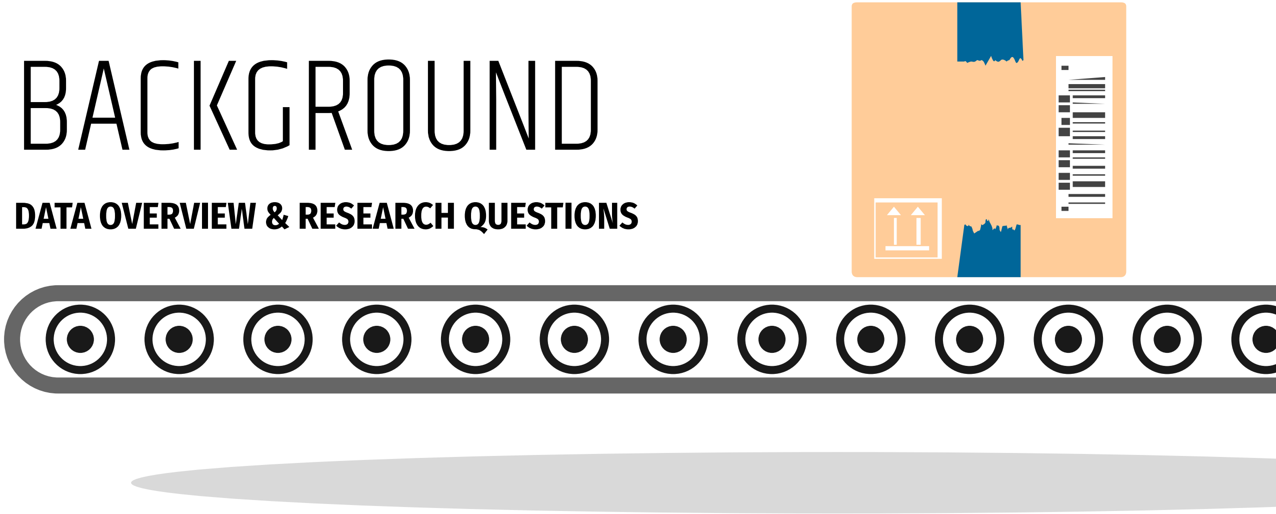
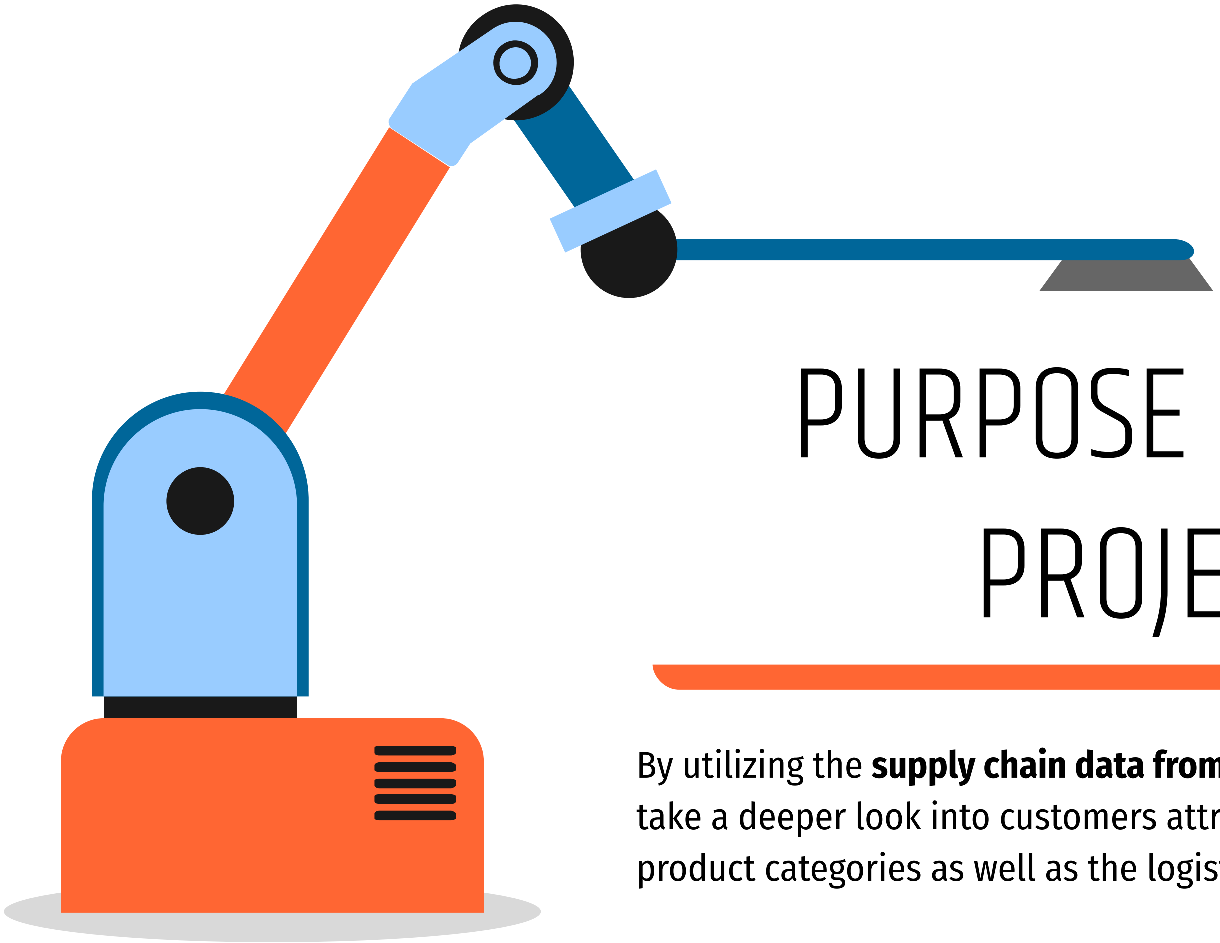


TABLE OF CONTENTS

BACKGROUND

DATA OVERVIEW & RESEARCH QUESTIONS





PURPOSE OF THE PROJECT

By utilizing the **supply chain data from DataCo Global**, we aim to take a deeper look into customers attributes/preferences in product categories as well as the logistics of the delivery process.

DATA OVERVIEW

This dataset contains **structured data** related to orders, including payment information, shipping details, and product information.

- Total of 53 columns and 180,519 rows of data.
- Geographical Scope -
  - Customers Based in the contiguous United States and Puerto Rico.
  - Orders shipped out to 164 different countries across the globe.
- Time -
  - Dataset released in 2019, has not been updated since.

DATA OVERVIEW *(Contd.)*

Data at a glance:

- **180,519 orders** in the dataset.

Total Count of Delivery Statuses

Delivery Status	
Advance shipping	41,592
Late delivery	98,977
Shipping canceled	7,754
Shipping on time	32,196

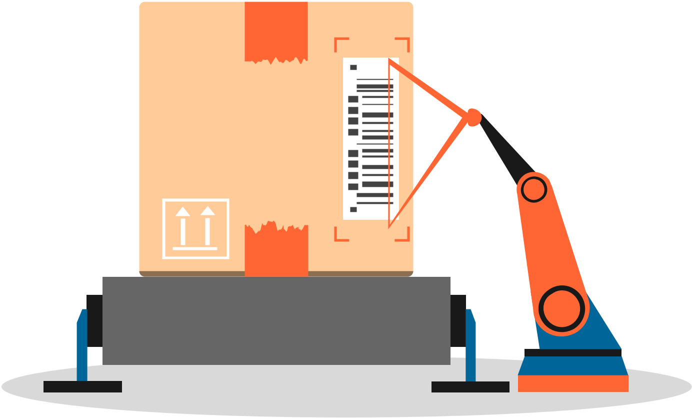
Count of Order Id

7,754 98,977

- Top 5 cities where **Customers** are based:
  - Caguas, PR
  - Chicago, IL
  - Los Angeles, CA
  - Brooklyn, NY
  - New York, NY

- Top 5 countries receiving **Orders:**
  - United States
  - France
  - Mexico
  - Germany
  - Australia

RESEARCH QUESTIONS



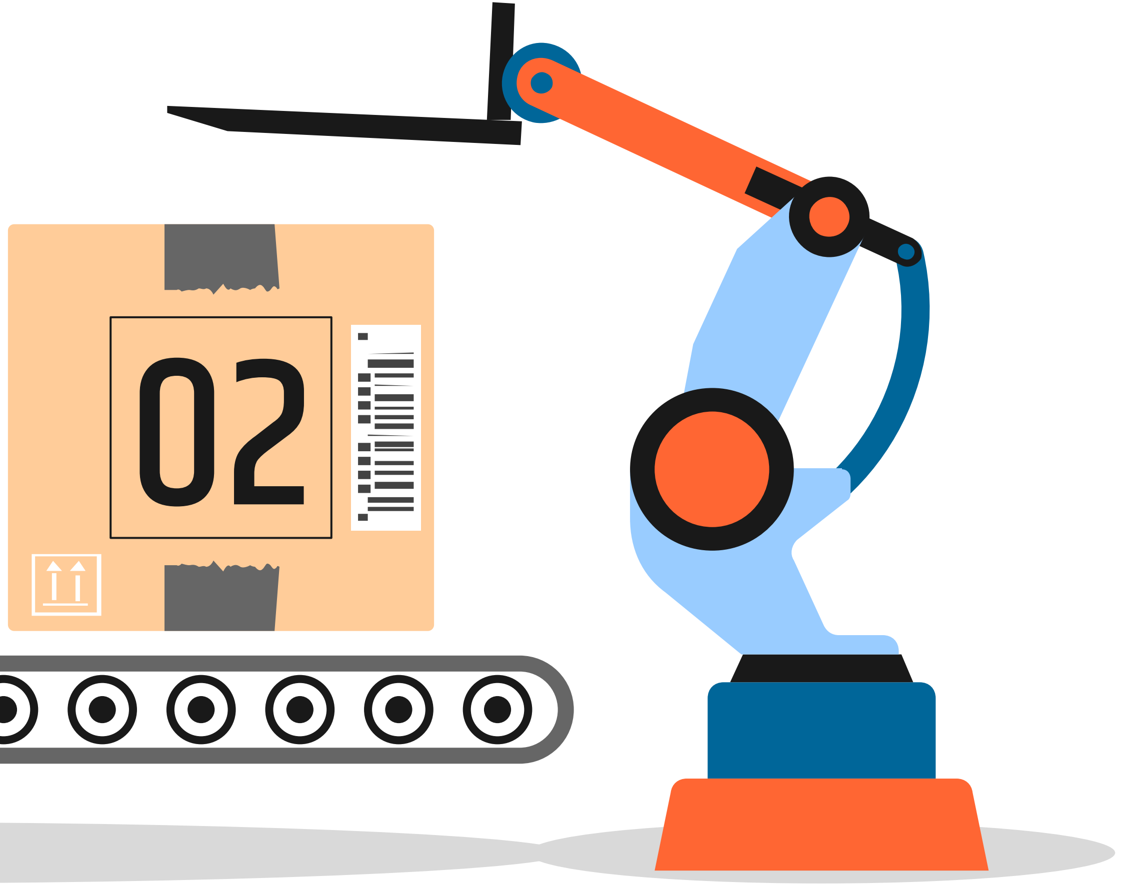
What are the predicted sales based on order volume?

What is the risk of a delivery being late?

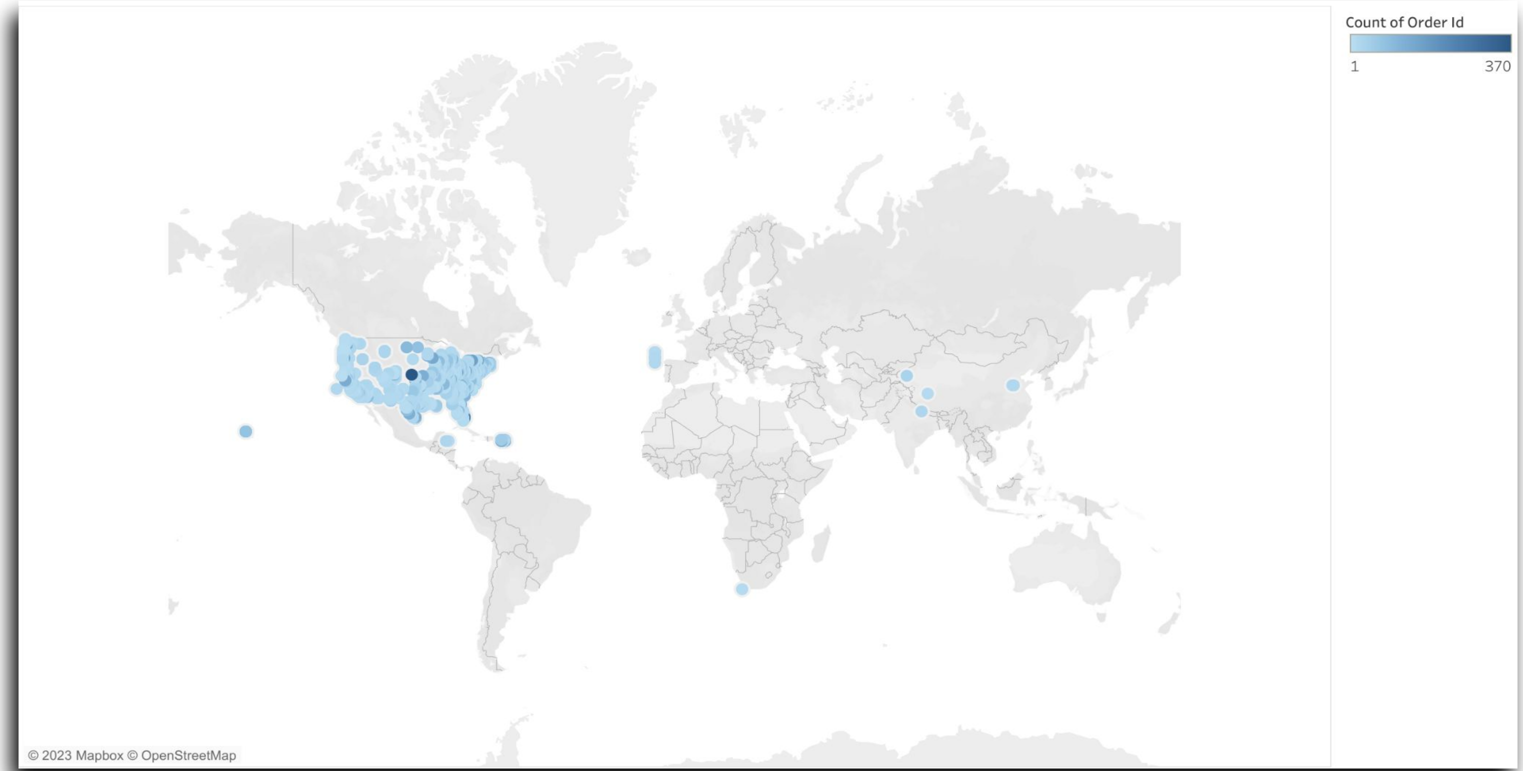
How many orders can we anticipate to be fraudulent?

DATA VISUALIZATIONS

EXPLORATORY ANALYSIS

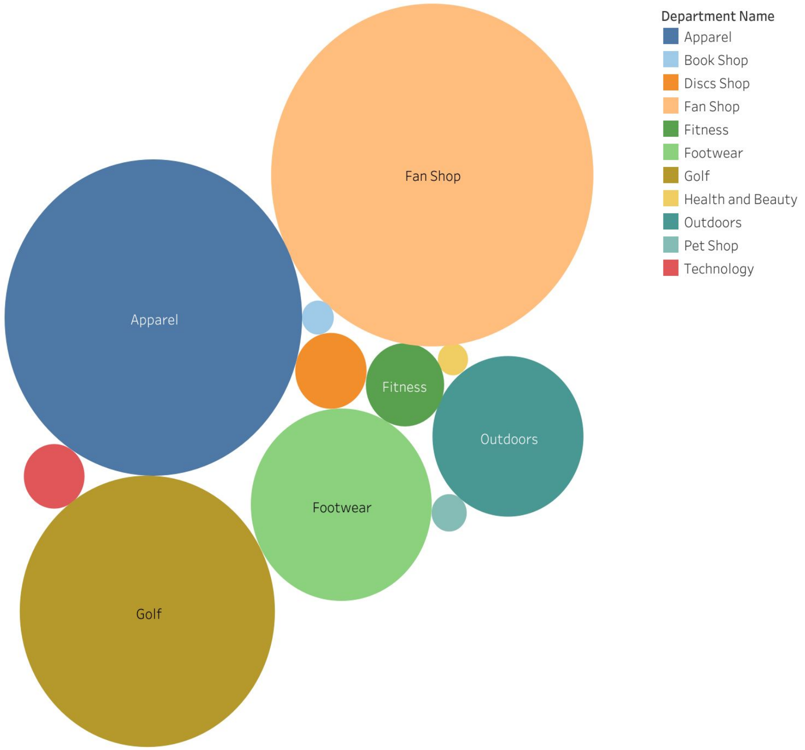


MAP OF CUSTOMER LOCATION BY LATITUDE & LONGITUDE

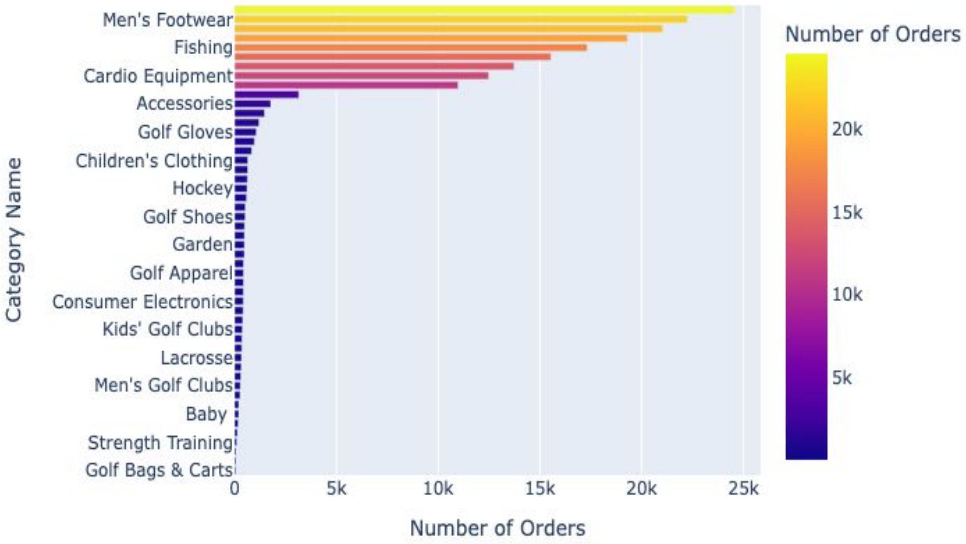


CUSTOMER PREFERENCES

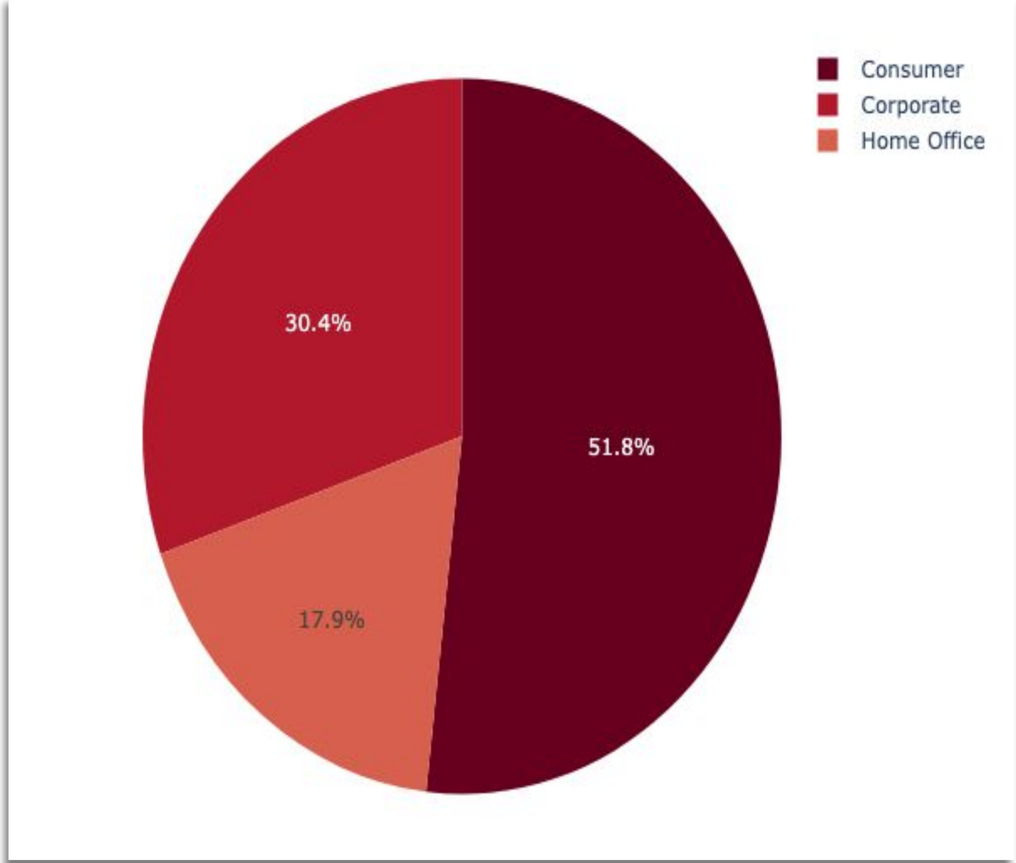
Most Popular Products by Order Count

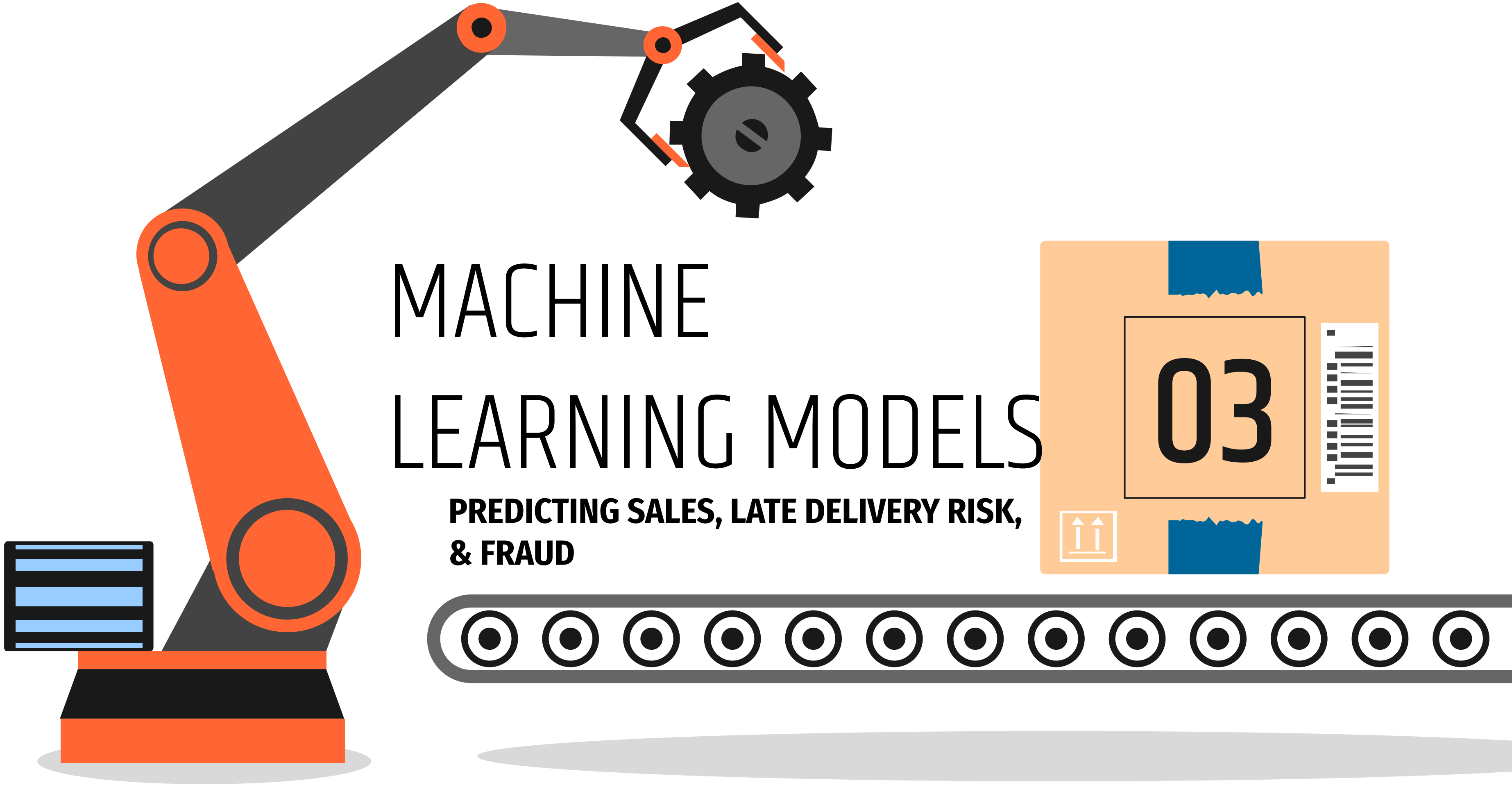


Most Popular Categories by Order Count

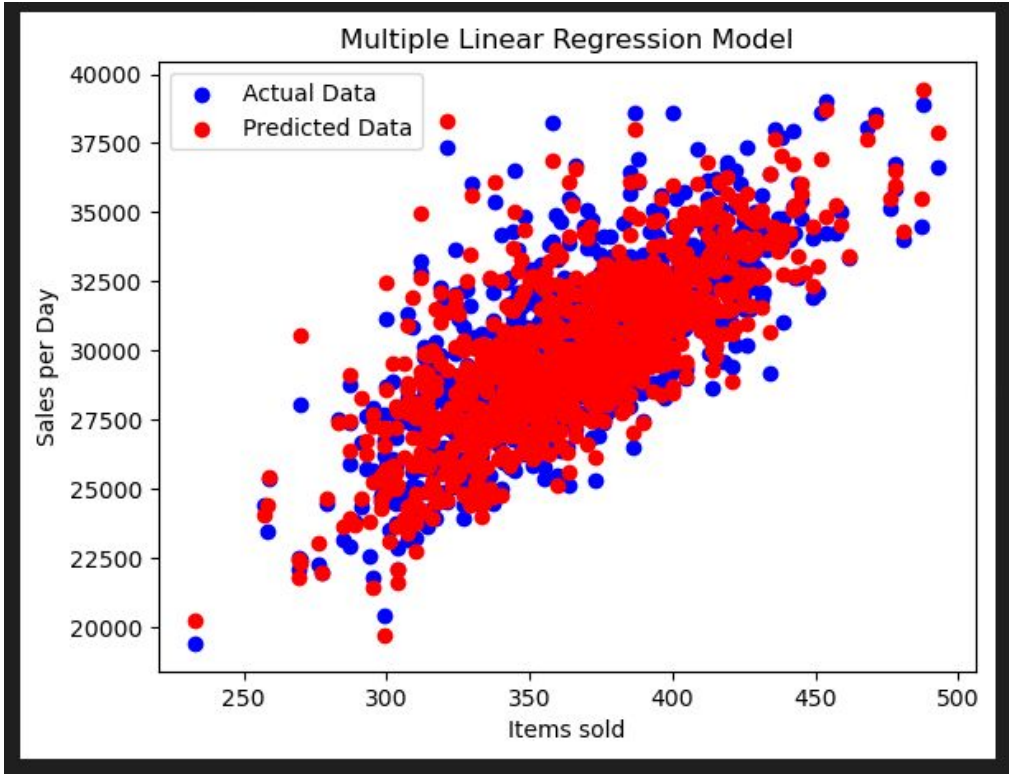


Number of Orders of Different Customer Segments




MACHINE
LEARNING MODELS
PREDICTING SALES, LATE DELIVERY RISK,
& FRAUD
03

Sales Predictions



Multiple Linear Regression Model

Prediction based on the combination of:

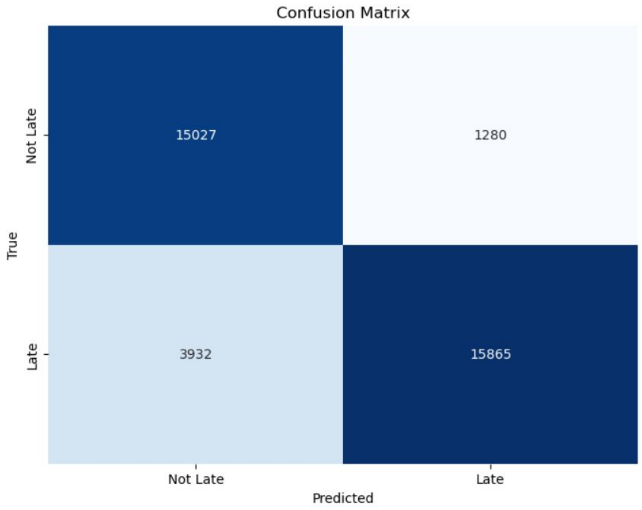
- item sales volume
- average product price & discount

Applications:

- revenue & volume benchmarking
- setting optimal discount rates
- item pricing

R-squared: 0.94
RMSE: 741.45
MAE: 575.72

Late Delivery Risk




```
True Negatives (TN): Model correctly predicts a delivery as 'Not Late'.
False Positives (FP): Predicted as 'Late', but are actually 'Not Late' .
False Negatives (FN): Predicted as 'Not Late', but are actually 'Late'.
True Positives (TP): Model correctly predicts a delivery as 'Late'.
-------------------------
TN: n= 15027 (41.62%)
FP: n= 1280 (3.55%)
FN: n= 3932 (10.89%)
TP: n= 15865 (43.94%)
```

- Using a RandomForestClassifier to predict the risk of late deliveries based on input features from the dataset.
- After training the model and making predictions on a test set, it evaluates the model's performance and visualizes the results using a confusion matrix and a classification report.

```
Accuracy and Classification Report
Key: (0) = Not Late, (1) = Late
------------------------------
Accuracy: 0.8556392643474408
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.92      0.85     16307
           1       0.93      0.80      0.86     19797

    accuracy                           0.86     36104
   macro avg       0.86      0.86      0.86     36104
weighted avg       0.87      0.86      0.86     36104
```

Fraud Predictions Model

Random Forest Model &
Extreme Gradient Boosting

	Legitemate [0]	Fraud [1]
Legitemate [0]	12616	3142
Fraud [1]	61	278

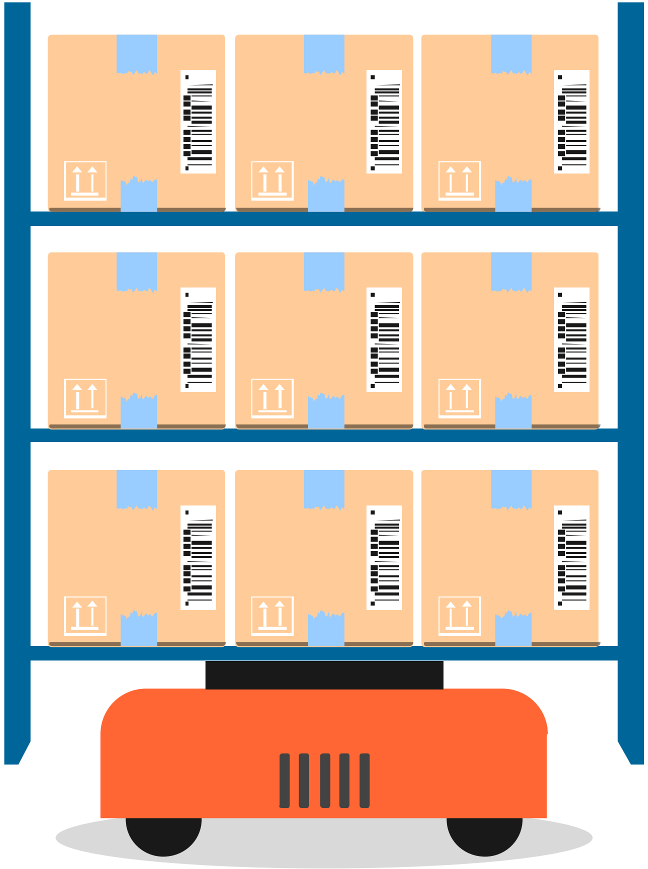
```
Accuracy Score : 0.8010188233832392
Classification Report
              precision    recall  f1-score   support

           0       1.00      0.80      0.89     15758
           1       0.08      0.82      0.15       339

    accuracy                           0.80     16097
   macro avg       0.54      0.81      0.52     16097
weighted avg       0.98      0.80      0.87     16097
```

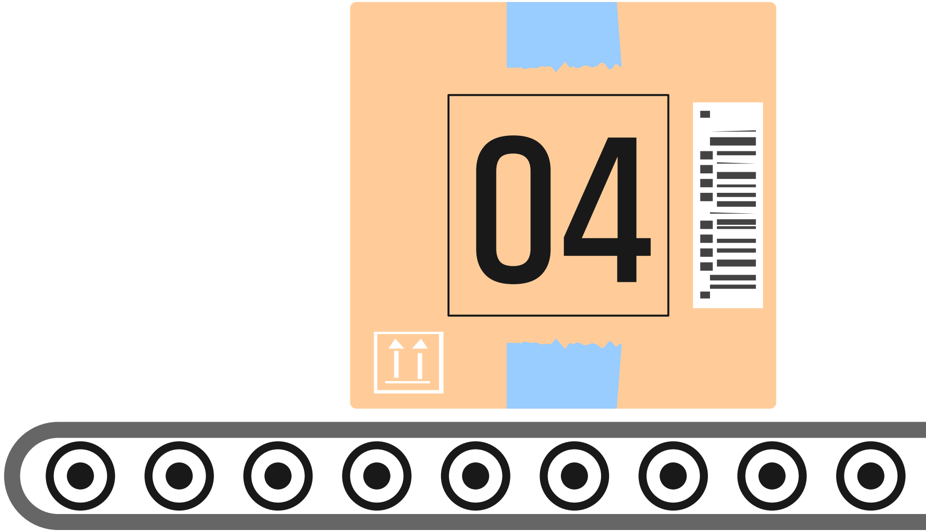
Struggling with applicability of the model due to accuracy/recall/precision trade off.

Initial very high Accuracy of 98% resulted in 0 recall for fraud delivery due to imbalanced data.

Improved fraud recall by applying multiple strategies to a random forest model:
- scaling [no impact],
- SMOTE & RandomUnderSampler [minor impact],
- adjusting classifier to balance class_weight [minor impact],

Applied XBoost which improved recall BUT at a price of overall accuracy & misclassification of legitimate transactions





KEY TAKEAWAYS

DATA LIMITATIONS & ADDRESSING MODEL COMPLEXITIES

DATA LIMITATIONS

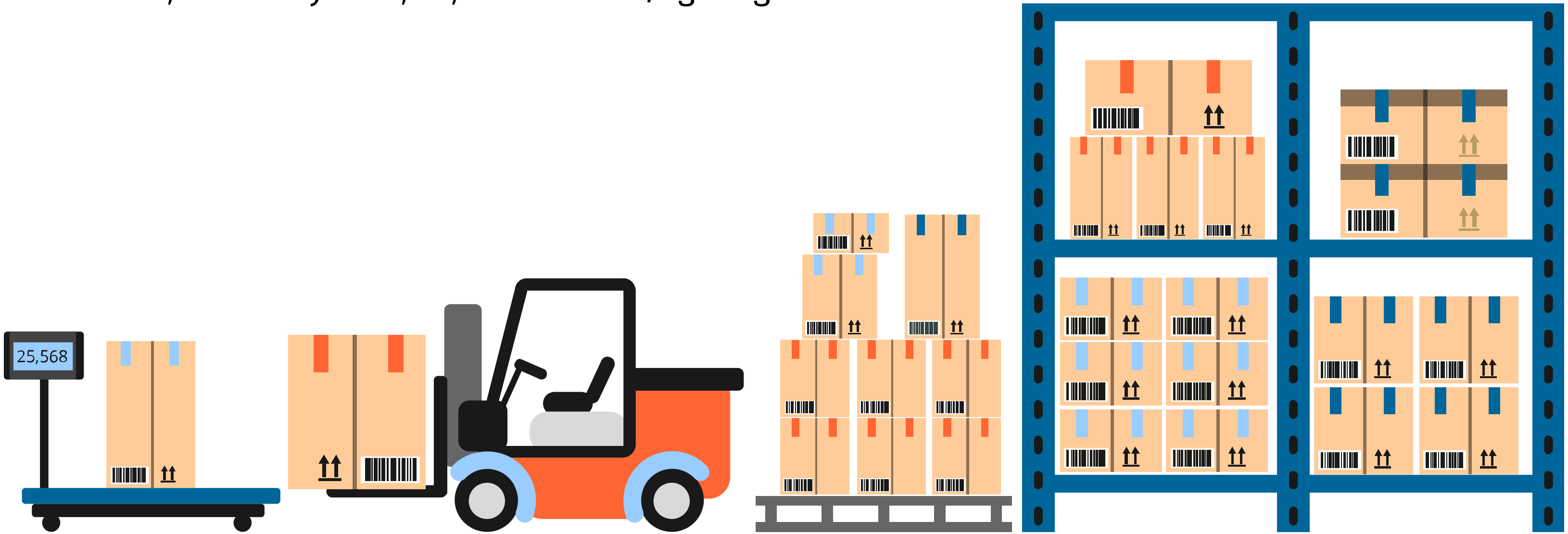
- The data is **synthetic**, meaning it may not fully capture the complexity and nuances of real data
- **High-dimensional data** can make it **challenging** for the model to identify relevant features and may lead to overfitting
- **Further analysis** could include the utilization of the **Principal Cost Analysis**, reducing the number of features

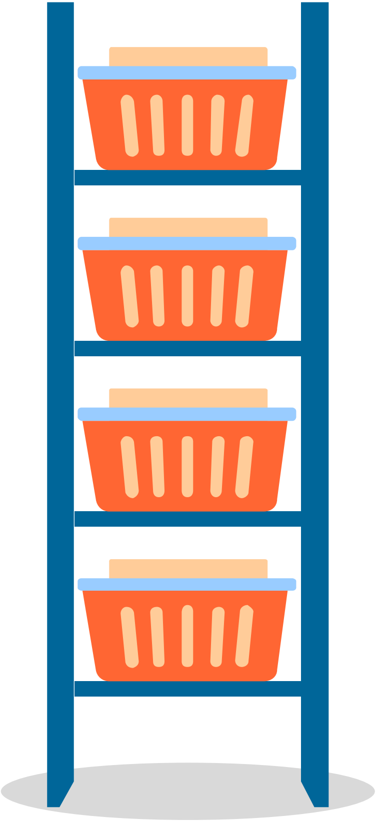
MODEL COMPLEXITIES

- **SMOTE** and **RandomUnderSampler** can address class imbalance but also may introduce noise or reduce the amount of training data
- Understanding of the business context is critical when **balancing the trade-offs between recall and overall accuracy**

DATA REFERENCE:

Constante, Fabian; Silva, Fernando; Pereira, António (2019), “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS”, Mendeley Data, V5, doi: 10.17632/8gx2fvg2k6.5





THANKS!

ANY QUESTIONS?