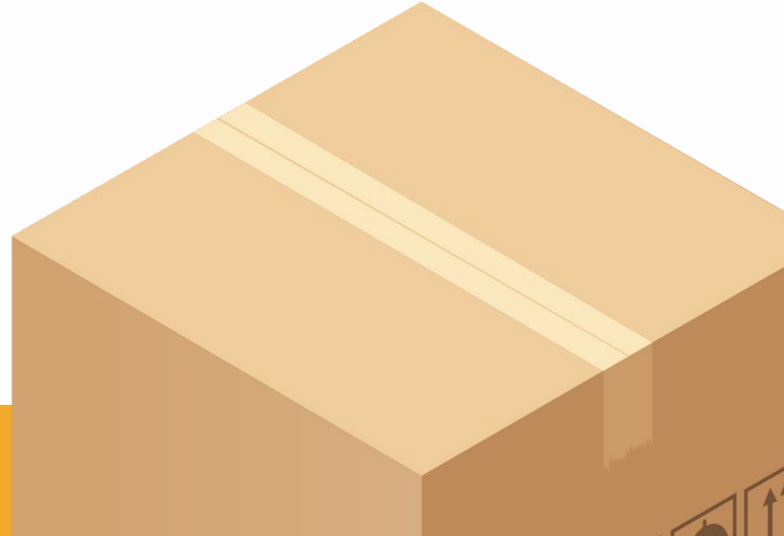
FINAL PROJECT DATA SCIENCE

LATE DELIVERY

PREDICTIVE ANALYTICS

By : Dinda Raraswati

TABLE OF CONTENT



ABOUT ME

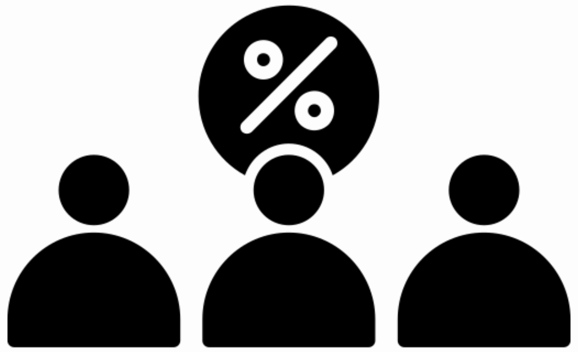
SELF-OVERVIEW
A data enthusiast with a background in Agricultural Engineering who is currently transitioning from academia to industry

EDUCATION
- **Bachelor of Science in Agricultural Engineering (2016 - 2020)**
  Bandung Institute of Technology (ITB)
- **Master of Agricultural Science (2021 - 2023)**
  Kyoto University
- **Data Science Bootcamp (Apr 2025 - present)**
  dibimbing.id

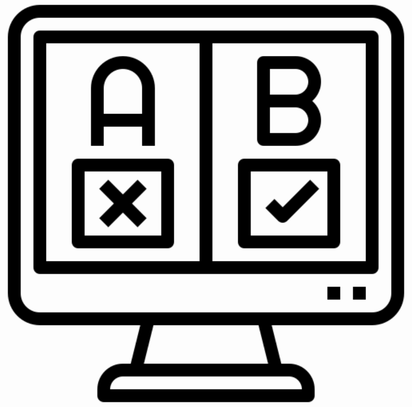
WORKING EXPERIENCE
- **Wageningen Food Safety Research (WFSR) (Nov 2023 - Aug 2025)**
  Researcher
- **Climate Change Center ITB (PPI-ITB) (Dec 2020 - Apr 2021)**
  Project Assistant

PREVIOUS PROJECTS

E-commerce Transaction Analytics
Analyze sales pattern at e-commerce



Bank Customer Churn Prediction
Develop customer churn prediction model using classification algorithms



A/B Testing on Landing Page Designs
Conduct A/B testing to evaluate the effectiveness of different landing page designs on speaker sales

Customer Segmentation of Airline Passengers
Segment airline passengers using K-Means clustering

EXECUTIVE SUMMARY

Problem Statement

DataCo global company has been struggling with late deliveries. Out of 180K transactions over the period of 2015 - 2017, **55% orders were shipped late.** This issue led to **customer dissatisfaction and loss revenue**

Objectives

- Identify key risk factors influencing late delivery risk
- Develop ML-based models to predict delay risk
- Derive Actionable Insights

Methodology

- Used DataCo's transactional data (180K orders)
- Implemented data preprocessing on dataset
- Developed four ML models **(Logistic Regression, Random Forest, Decision Tree, XGBoost)**
- Experimented on different types of data preprocessing (Outlier handling vs original data)
- Tuned chosen model
- Conducted SHAP analysis for model interpretability

Key Findings

- **Standard Class achieved the highest on-time rate** at 61.93% with late deliveries up to 2 days
- On-time rate was quite stable at 44.5% - 46% from January to December
- On average, stores in the e-commerce had an actual lead time of 3.5 days and an expected lead time of 2.9 days, making the **shipping day gap at 0.6 days**
- **XGBoost is the best model with the accuracy of 92%**
- **Shipping schedule, customer city, and shipping mode** are top 3 key drivers of late deliver risk

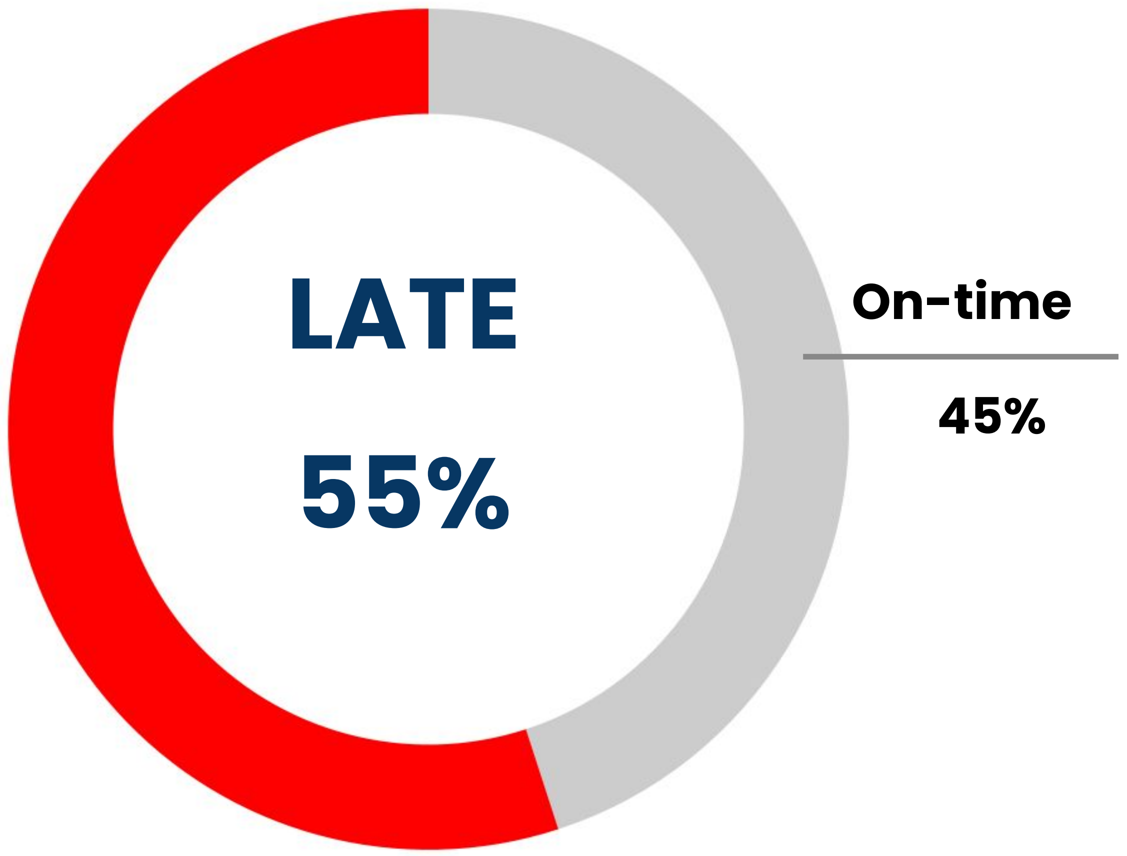
Business Recommendations

- **Adjust shipping schedules :** Develop model to estimate actual shipping days more accurately; Extend shipping days to lower late delivery risk
- **Optimize warehouse locations :** build new warehouses close to regions with the highest number of orders
- **Route optimization**
- **Optimize shipping mode performance** : evaluate and improve shipping mode performances, particularly first class and second class
- **Plan Shipping During Peak Seasons/Hours**
- **Optimize payment process :** Speed up payment confirmation to reduce delays

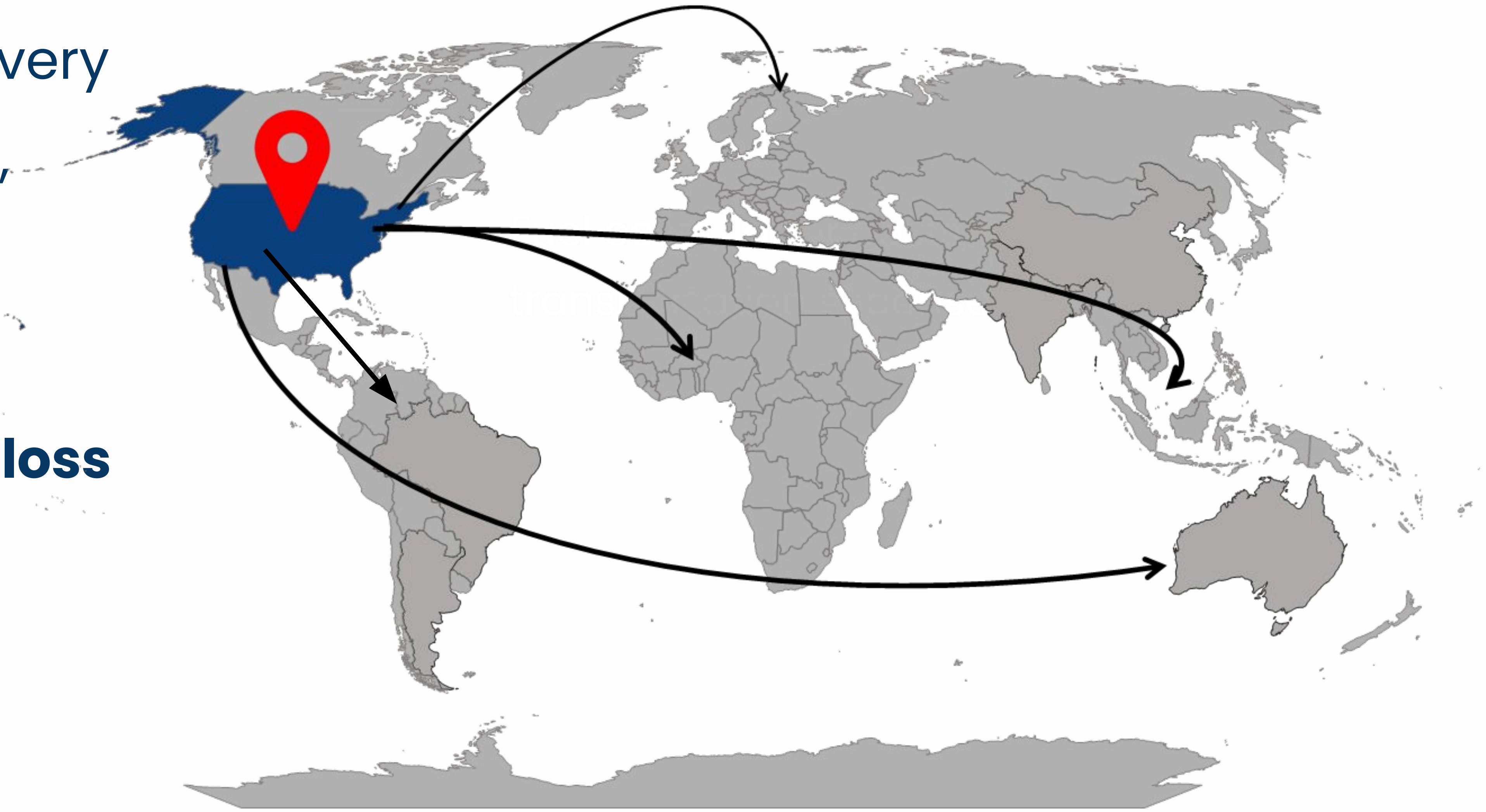
BUSINESS UNDERSTANDING

Problem Statement

- DataCo global company is struggling with late delivery
- Out of 180K transactions in the period of 2015 - 2017, **55% of of total orders were shipped later than expected**
- This issue led to **customer dissatisfaction and loss revenue** [(Medida, 2025)](#)



Possible Causes



 **Severe Weather/Natural Disaster**

 **Transportation Issues**

 **Custom & Regulation**

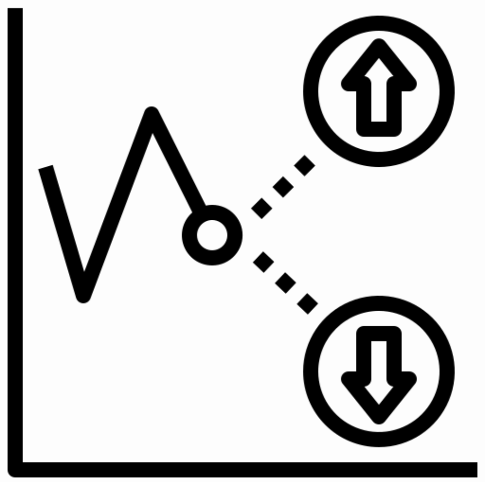
BUSINESS UNDERSTANDING

Key Challenge

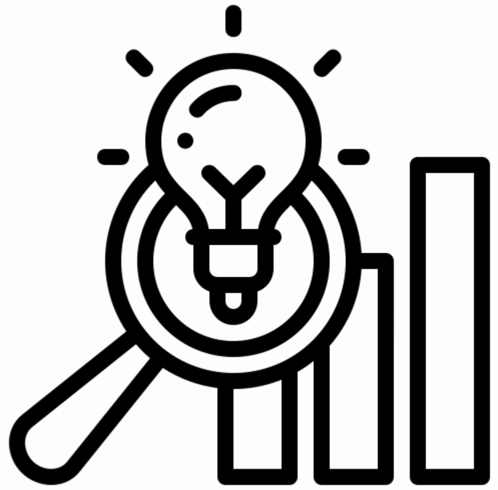
How can we leverage data-driven analysis and prediction model to formulate actionable recommendations for dealing with shipping delays?

Project Objectives

Identify Key Risk Factors
Determine variables such as order location, month, etc that influence late delivery risk



Develop Prediction Model
Develop ML-based models to predict delay probability



Derive Actionable Insights
Gain insights into factors influencing supply chain risks and formulate recommendations for effective strategies

DATA UNDERSTANDING

- Dataset can be downloaded on Kaggle
- Supply chain dataset was used by DataCo Global company for their analysis which include detailed information about customer, shipping, and purchased products
- Dataset contains **180,519 rows** with **53 features**
- Collected from January 2015 to September 2017
- **Dataset has more than one potential target variable** depending on ML problems

DATA PREPROCESSING

Convert Data Types
Convert column timestamp
Object -> Date

Check and Handle Missing Values
Some missing values found

Check and Handle Duplicates
No duplicates found

Check and Handle Outliers
Outliers were transformed

Feature Engineering
Add some columns :
geospatial and temporal
for further analysis

MISSING VALUES

>85%	<1%	0
Missing values in two columns	Missing values in two columns	Remaining columns

DUPLICATES

0

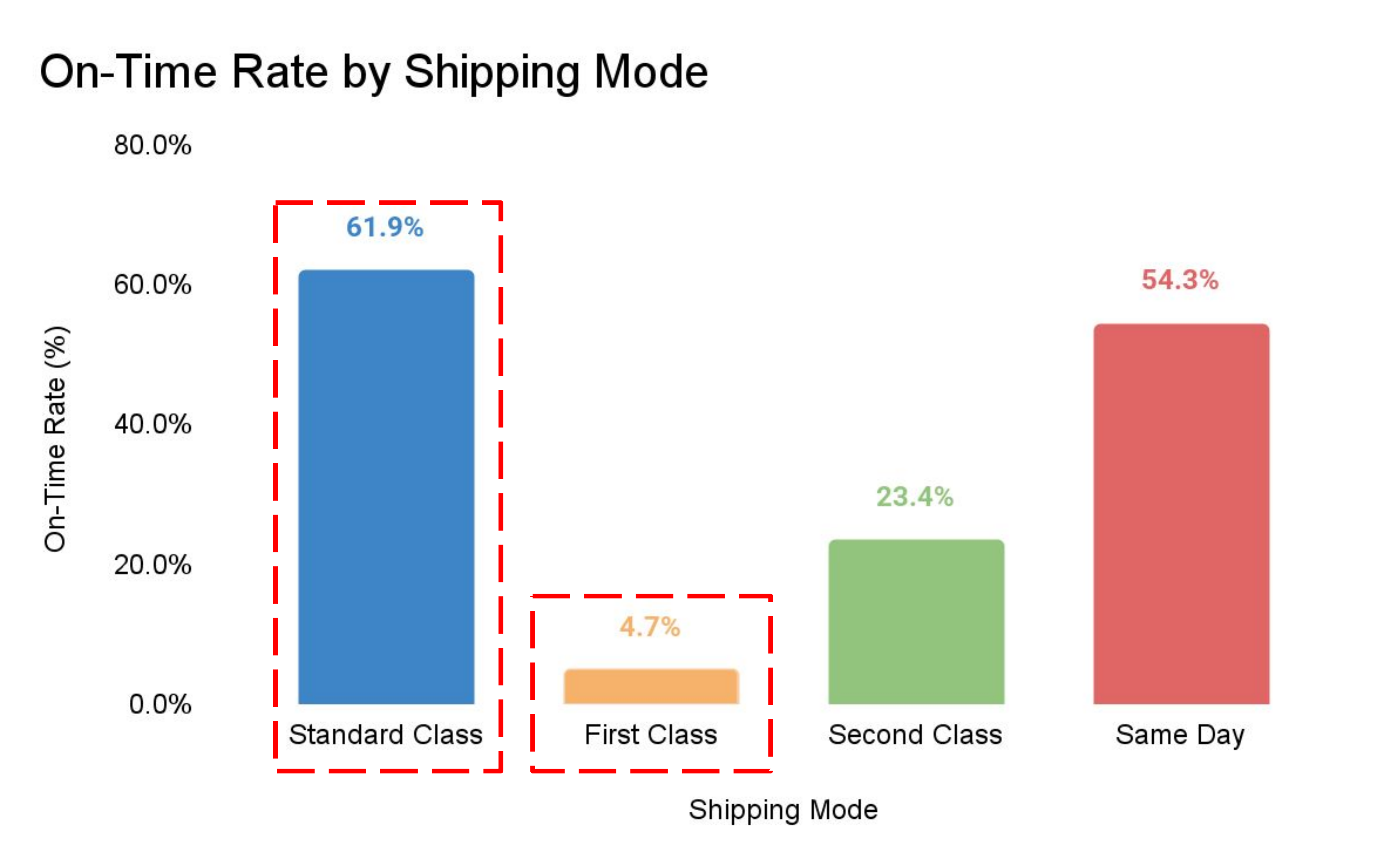
No duplicates

FINAL COLUMNS

56

EXPLORATORY DATA ANALYSIS

Delivery Risk by Shipping Mode

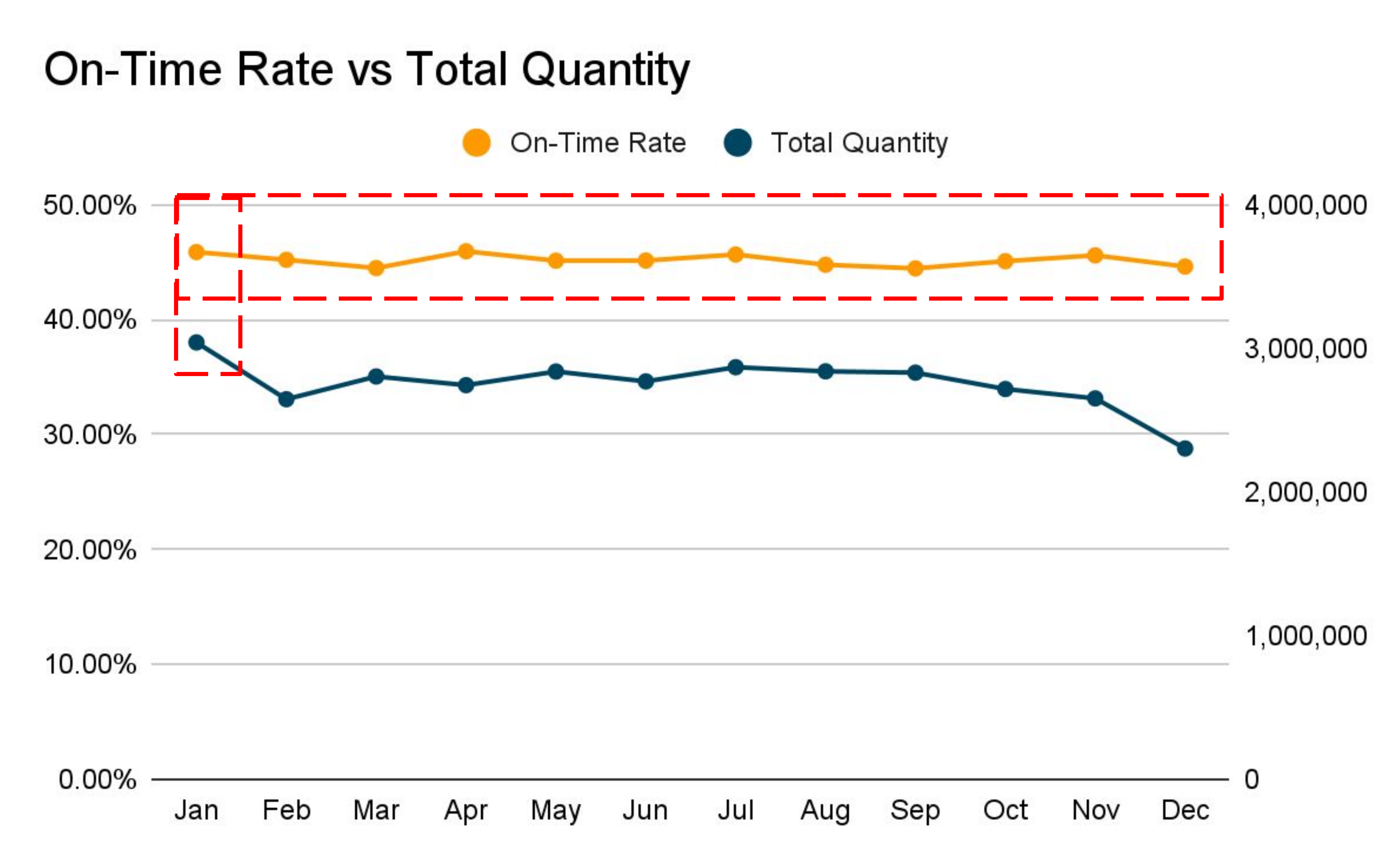


Insights :

1. Surprisingly, **First Class shipping had the lowest on-time rate (4.68%)**, while **Standard Class achieved the highest on-time rate at 61.93%**
2. Second Class shipping had a low on-time rate of 23.37% with deliveries up to 4 days later than scheduled
3. With on-time rate of 54.26%, Same Day delivery were shipped either on schedule or delayed by one day

EXPLORATORY DATA ANALYSIS

Seasonality Analysis



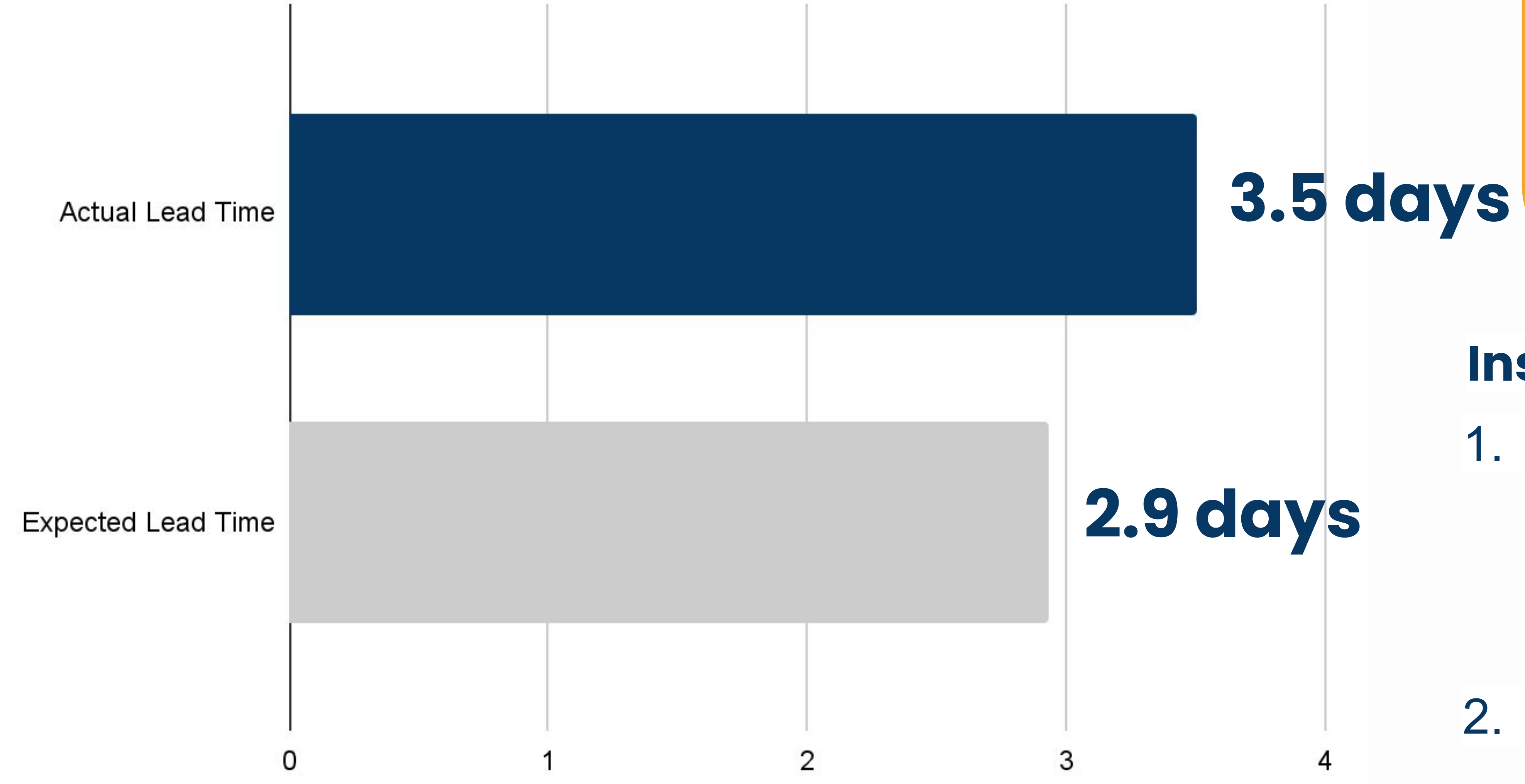
Insights :

1. Overall, **smaller quantity volume led to higher on-time rate**, indicating that delayed delivery is influenced by shipment volume
2. **On-time rate was quite stable at 44.5% - 46%** from January to December with the lowest rate occurring in March and September (44.5%) and the highest rate occurring in April (45.9%)
3. December had both the lowest total quantity sold and relatively low on-time rate, indicating that shipped volume did not influence low on-time rate this month
4. Despite a high volume of shipped products, January still achieved the second-highest on-time rate

EXPLORATORY DATA ANALYSIS

Store Analysis

Gap between actual vs expected delivery performance



0.6 days
Avg. Lead Time Deviation

20.7% slower than expected

Insights :

1. On average, stores/warehouses in the e-commerce had an actual lead time of 3.5 days and an expected lead time of 2.9 days, making the **shipping day gap at 0.6 days**
2. More than half of total stores/warehouses **(54.3%) shipped their products later than expected by more than 0.6 days**, with the worst delay reaching 4 days

DATA PREPARATION FOR MODELING

Train-Test Split

80% Training Data
20% Testing Data

Feature Encoding

- Categorical data with ordered values : **Ordinal Encoding**
- Categorical data with unordered values : **Label Encoding**
- Categorical data with more than 20 values : **Target Encoding**

Feature Scaling

Standard Scaler

Feature Selection

Drop columns that are possibly leakage to target feature (shipping_day_deviation, Shipping Day (real), Delivery Status, Order Status)

Top 3 Highest Correlation with Target Feature

Features	Correlation
Days for shipping (real)	0.4
Days for shipment (scheduled)	-0.37
shipping_day_deviation	0.78

Out of 55 features**, only 3 features have moderate-strong correlations (>0.3)**

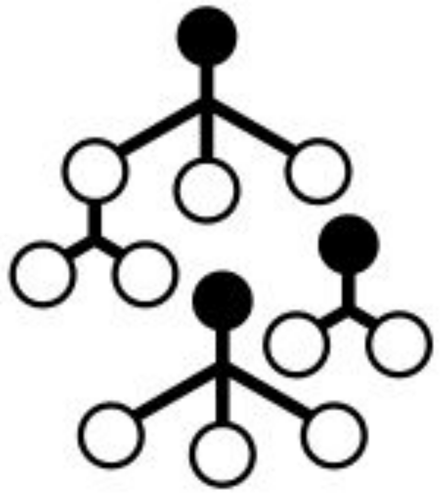
Positive correlation : the **longer** the actual shipping days as well as the **larger** deviations between scheduled and actual shipping days led to **higher probability of late delivery**

Negative correlation : the **longer** expected shipment days led to **lower probability of late delivery**

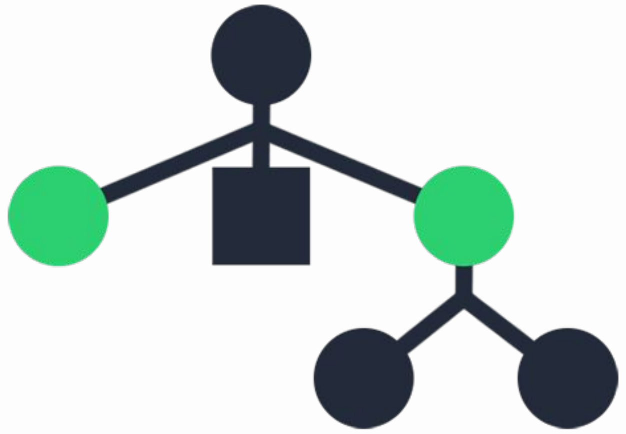
MODEL BUILDING

Four ML models were developed to compare their performances

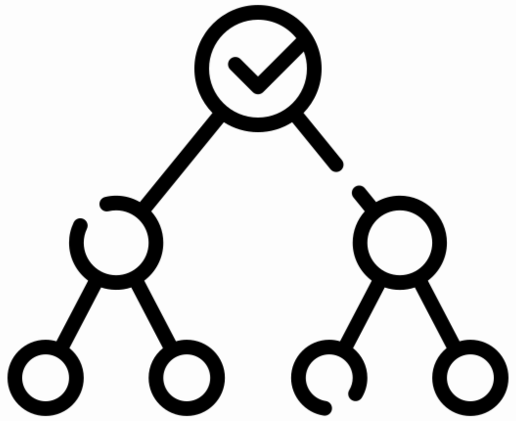
Logistic Regression



Random Forest



XGBoost

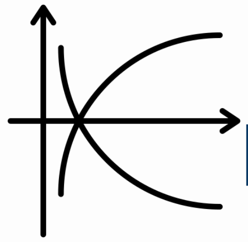


Decision Tree

EXPERIMENT 1

Baseline Dataset

EXPERIMENT 2

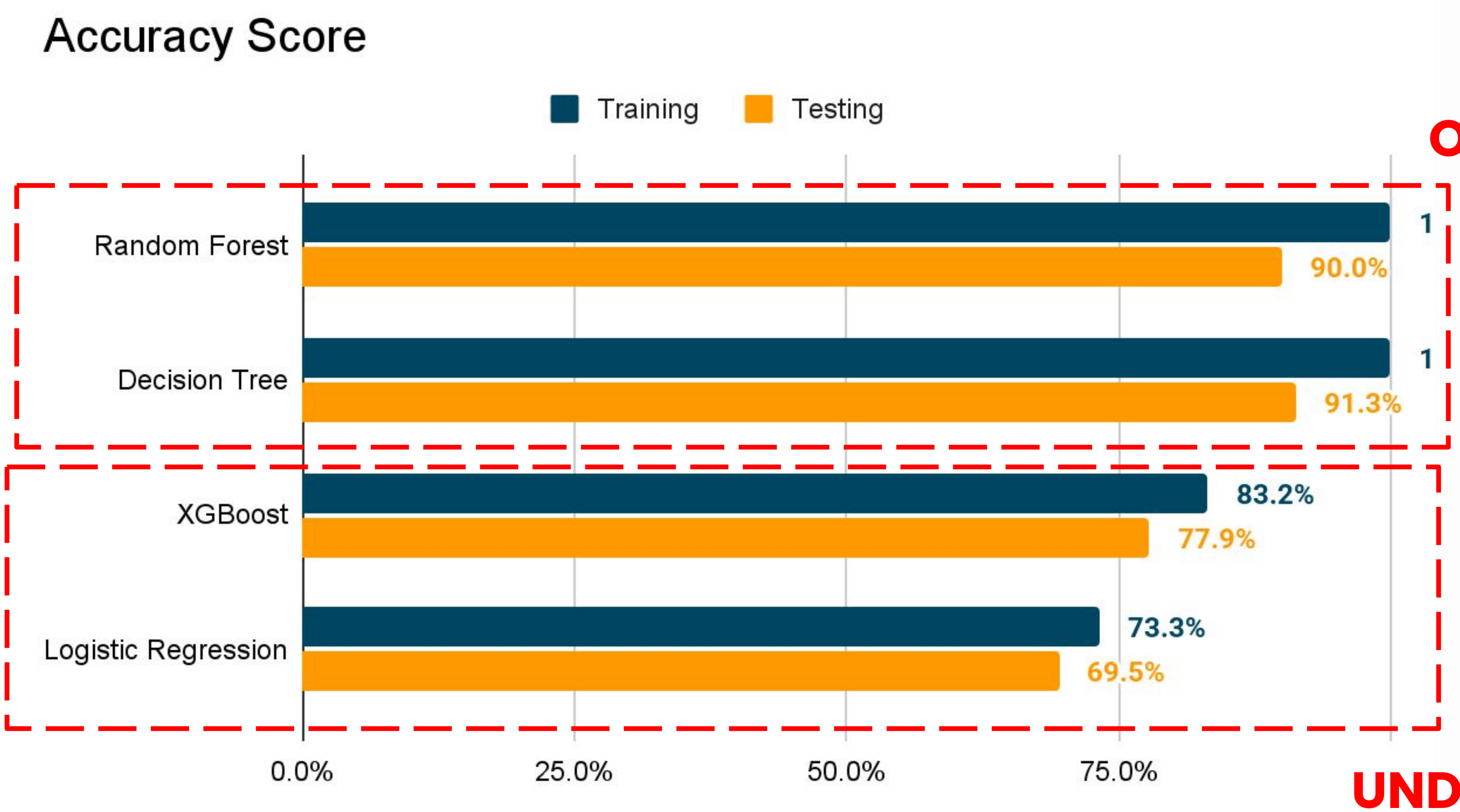


Outlier Handling

Log Transformation

MODEL EVALUATION

Original Dataset

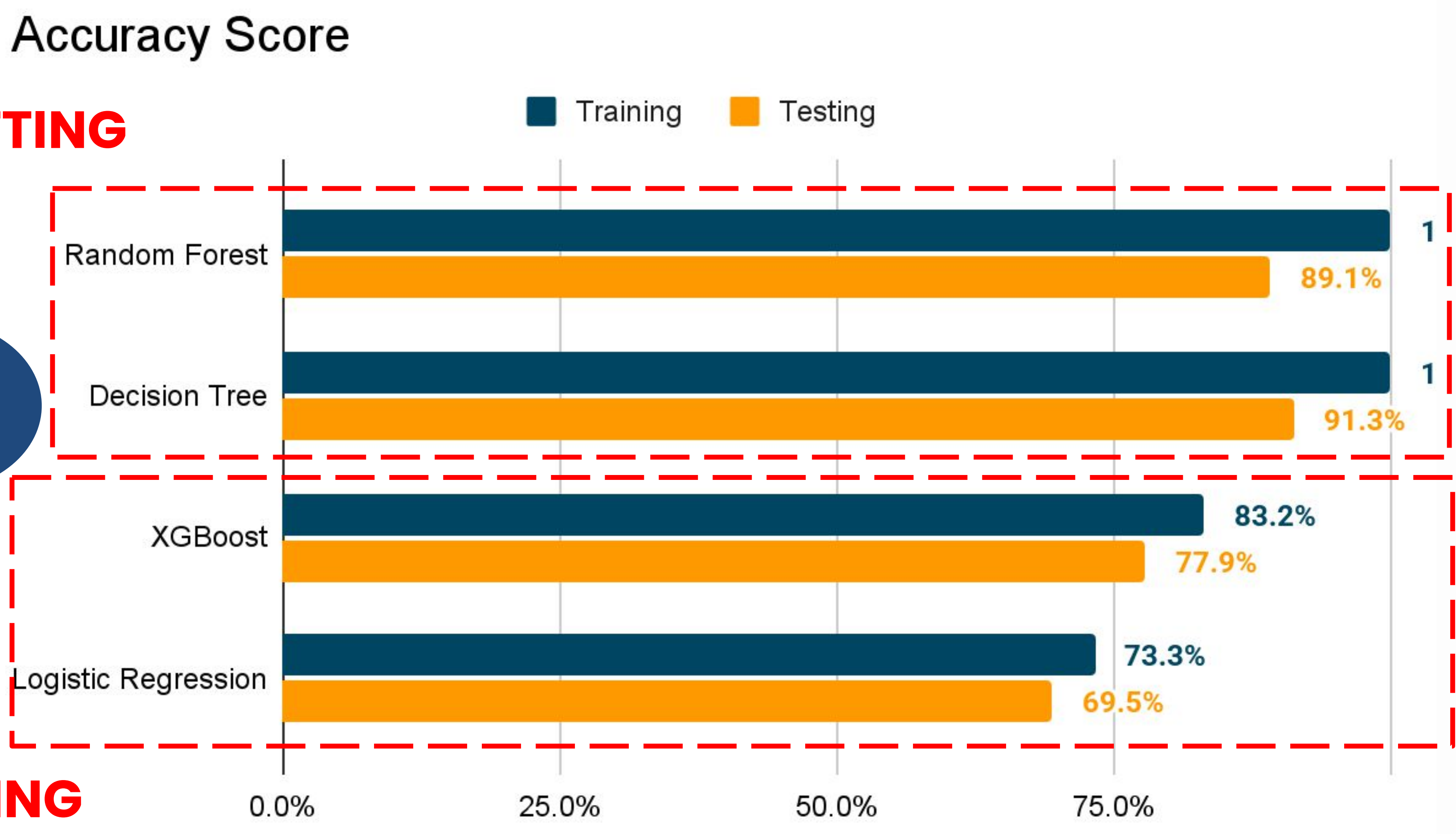


OVERFITTING

VS

UNDERFITTING

Outlier Handling



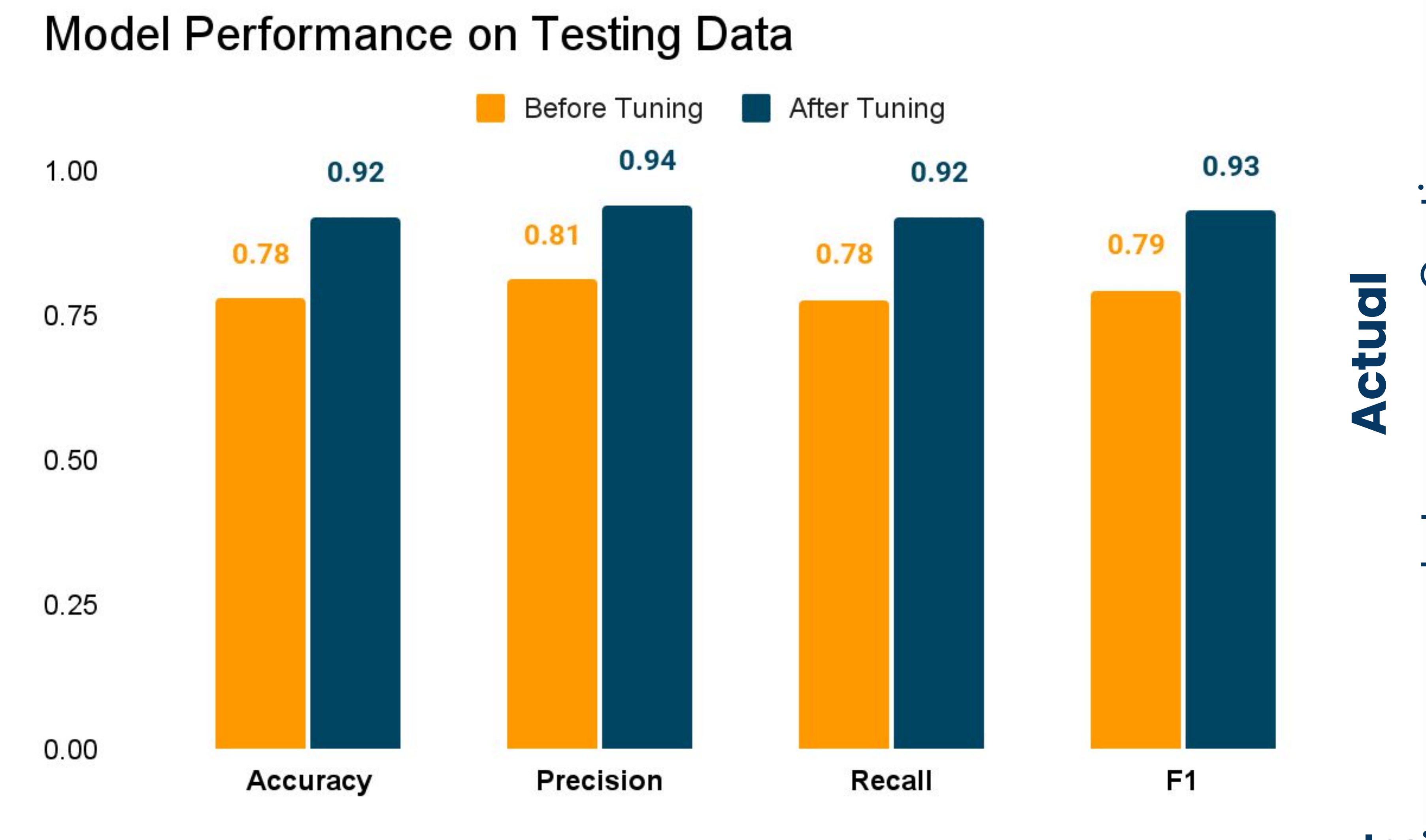
Insights :

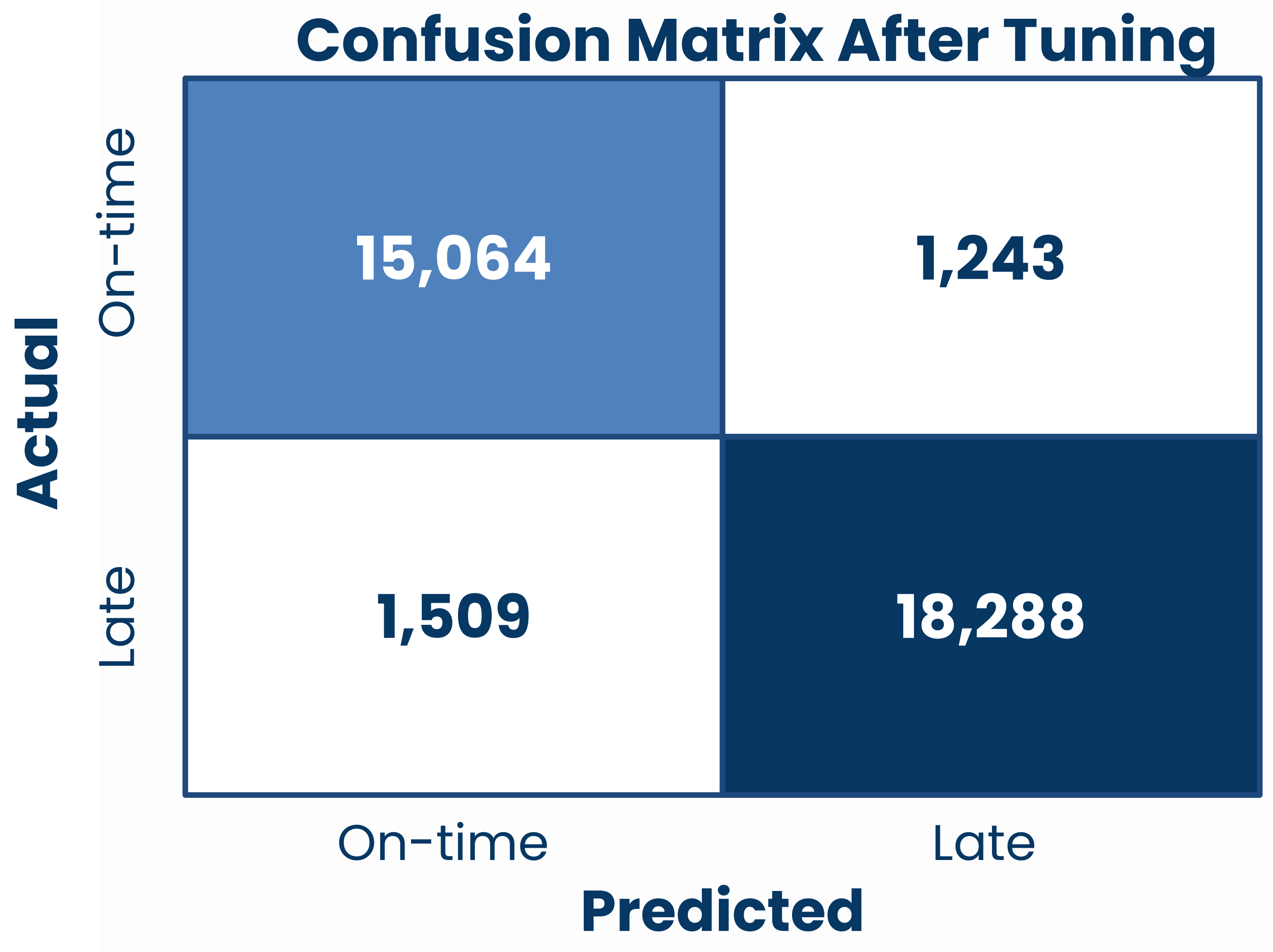
1. Random Forest and Decision Tree are overfitting to the training data, shown by all the metric scores of 1
2. Although Logistic Regression and XGBoost are good for data generalization, these models are possibly underfitting
3. **There is no difference in model performance between original dan outlier handling**

MODEL EVALUATION

Hyperparameter Tuning

Optuna optimization was implemented to choose the best parameters for XGBoost model





FP Rate: 0.08

FN Rate: 0.08

TP Rate: 0.92

TN Rate: 0.92

avg of 17.42% increase in evaluation metrics

Insights:

- **Hyperparameter tuning has improved XGBoost performance,** despite making it overfitting (all the metric scores on training data =1)
- **False Positive Rate (0.08) :** About 8% of on-time delivery were incorrectly predicted as late delivery risk
- **False Negative Rate (0.08) :** About 8% of late delivery were incorrectly predicted as on-time

MODEL INTERPRETATIONS

Top 10 Key Drivers of Late Delivery Risk

1. **Expected Shipping Schedule** : **shorter shipping scheduled** days led to **higher delay risk**
2. **Customer Street :** certain customer's location prone to late delivery
3. **Shipping Mode** : **first class** shipping mode has **higher risks** of late delivery
4. **Payment Type :** payment process influence late delivery risk since it is related to duration of payment confirmation
5. **Order City:** certain customer's location prone to late delivery
6. **Order Item ID :** certain products might influence late delivery risk
7. **Shipping Month :** some months with **high shipping volumes** led to **higher late delivery risk**
8. **Shipping Hour :** some hours with **high shipping volumes** led to **higher late delivery risk**
9. **Store ID :** some stores may have higher late delivery risk
10. **Store's Latitude :** store's location may influence delay since it is related to distance

POTENTIAL BUSINESS IMPACT

Current (Without Model)

55%

Late deliveries

- No early warning for high-risk shipments
- Too late to take measures on mitigating late delivery risk

With Prediction Model

92% Accuracy

Early Identification

- Proactive steps to avoid delays
- Smarter resource allocation
- Inform customers in advance or set slightly longer expected shipping days

Potential impacts : lower actual late rate, improved customer satisfaction, probably increased customer lifetime value (CLV)

A model with 92% accuracy does not directly fix shipping delays, but empowers the company to take actions early to mitigate late delivery risk

BUSINESS RECOMMENDATIONS

Adjust Shipping Schedules

- Develop model to estimate actual shipping days more accurately
- Extend shipping days to lower late delivery risk

Optimize Warehouse Locations

- Establish warehouses near regions with the highest number of orders
- **Cons :** need high cost to build new warehouses

Route Optimization

- Develop routing algorithm to make shipping efficient
- Cluster nearby regions, so deliveries can be completed faster and more efficiently

Optimize Shipping Mode Performance

- Evaluate First Class shipping mode and improve its performance
- Remove shipping modes with low on-time rate and high costs

Plan Shipping During Peak Seasons/Hours

- Allocate extra resources during peak months
- Prioritize early day shipping for high-risk orders

Optimize Payment Process

- Speed up payment confirmation to reduce delays

ROOM FOR IMPROVEMENTS

Address Overfitting

Feature engineering, collect data on new features (weather, distance, etc), simplify the model

Monitor Performance

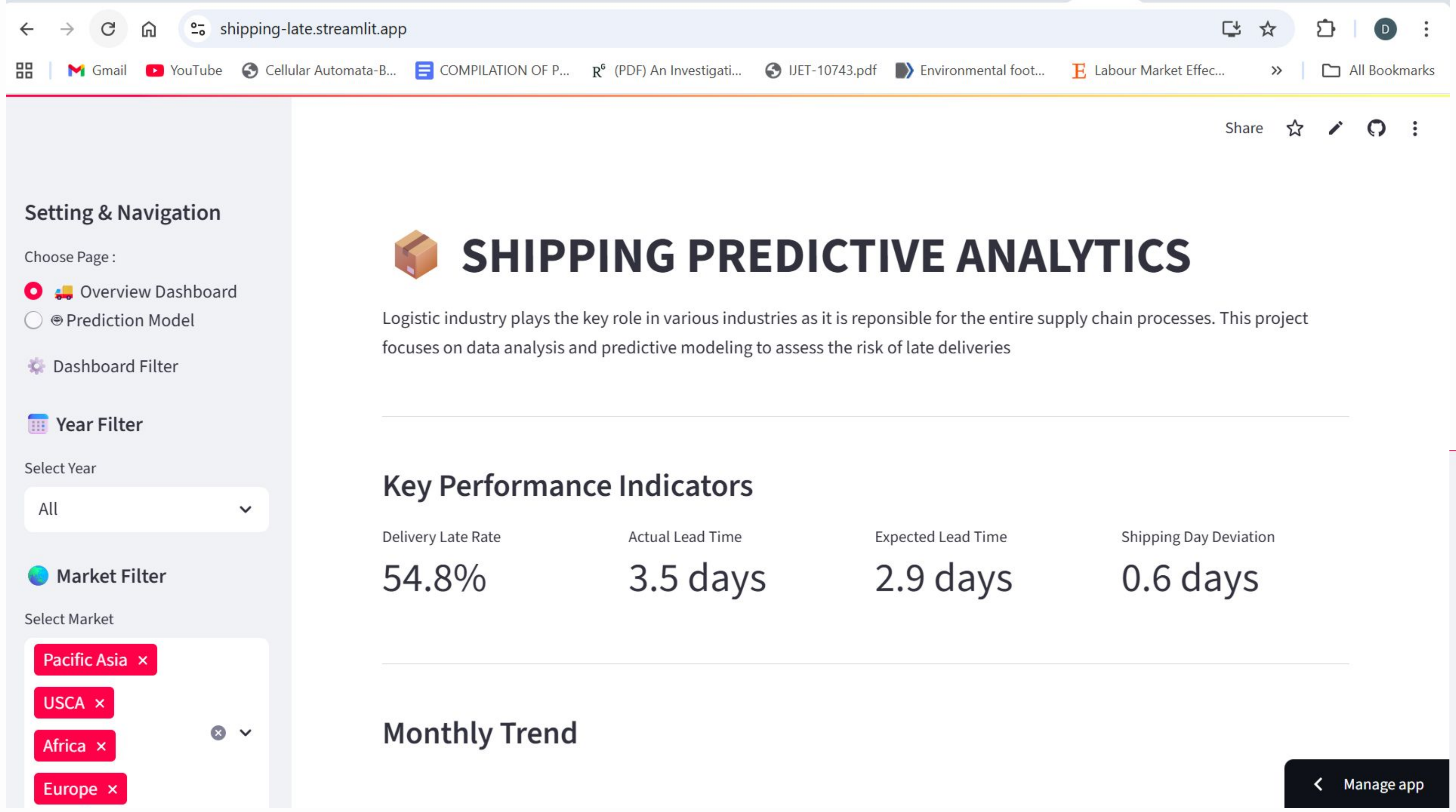
Track accuracy regularly and refine the model if the accuracy worsened

Deploy Model

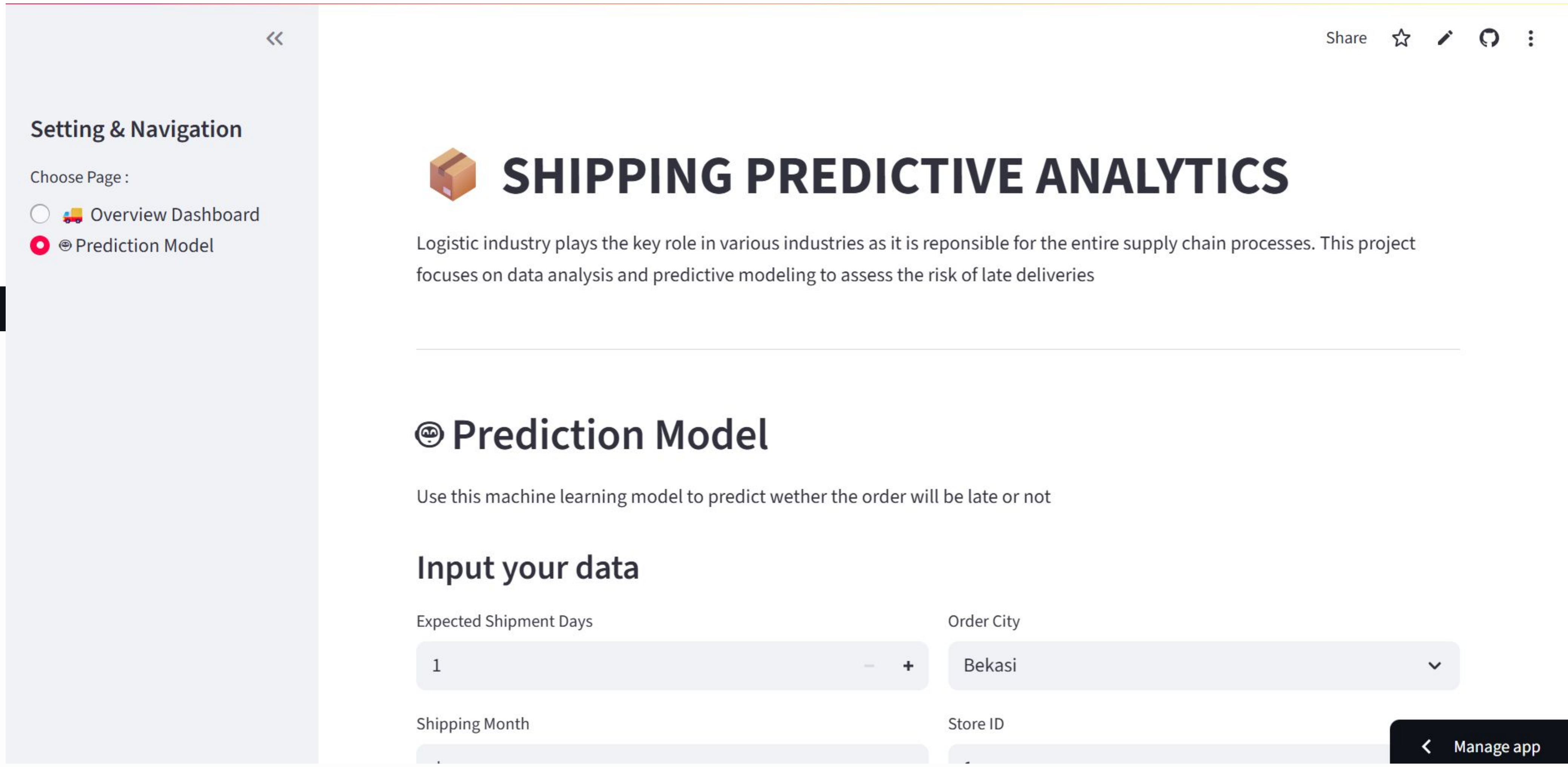
Implement model in the company system either via web or app

STREAMLIT DEPLOYMENT

Analysis Dashboard



Prediction Model



[LINK STREAMLIT]()

[LINK GOOGLE COLAB]()

Thank you!

dindararas

Dinda Raraswati

Dinda Raraswati