# SHIPMENT ACTIVITIES IN E-COMMERCE COMPANY

Nguyen Hoang Phuong Ngan
Group 7

# TABLE OF CONTENT


- DATA OVERVIEW
- EXPLORATORY DATA ANALYSIS
- DATA PRE-PROCESSING
- BUILD MODELS AND EVALUATE

# DATA OVERVIEW

## Project goals:

-Find out the factors contribute to the late delivery

-Build models to predict the on time delivered order by using ML and compare the performances between these models

- **ID :** ID Number of Customers.
- **Warehouse block** : The Company have big Warehouse which is divided in to block such as A,B,C,D,E.
- **Mode of shipment** :The Company Ships the products in multiple way such as Ship, Flight and Road.
- **Customer care calls** : The number of calls made from enquiry for enquiry of the shipment.
- **Customer rating** : The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).
- **Cost of the product** : Cost of the Product in US Dollars.
- **Prior purchases** : The Number of Prior Purchase.
- **Product importance** : The company has categorized the product in the various parameter such as low, medium, high.
- **Gender** : Male and Female.
- **Discount offered** : Discount offered on that specific product.
- **Weight in gms** : It is the weight in grams.
- **Reached on time** : It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

***SOURCE:*** *Kaggle*
***LINK: https://www.kaggle.com/datasets/poojakeer/e-commerce-dataset***

# DATA OVERVIEW

**Data info:** df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10999 entries, 0 to 10998 Data
columns (total 12 columns):
#  Column Non-Null Count Dtype
-- ------ -------------- -----
0  ID 10999 non-null int64
1  Warehouse_block 10999 non-null object
2  Mode_of_Shipment 10999 non-null object
3  Customer_care_calls 10999 non-null int64
4  Customer_rating 10999 non-null int64
5  Cost_of_the_Product 10999 non-null int64
6  Prior_purchases 10999 non-null int64
7  Product_importance 10999 non-null object
8  Gender 10999 non-null object
9  Discount_offered 10999 non-null int64
10 Weight_in_gms 10999 non-null int64
11 Reached.on.Time_Y.N 10999 non-null int64
dtypes: int64(8), object(4) memory usage: 1.0+
MB
```

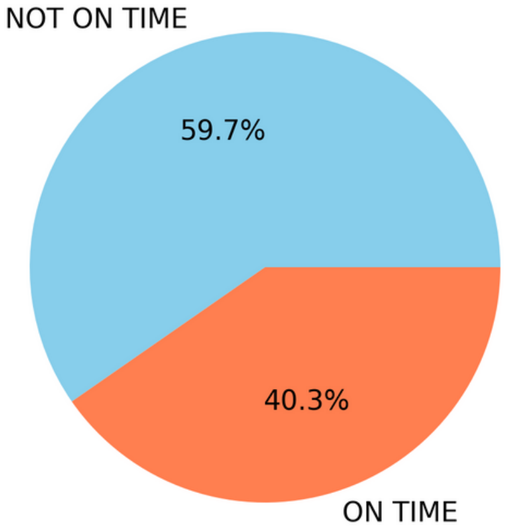
**Drop ID columns:** df= df.drop(['ID'], axis=1)

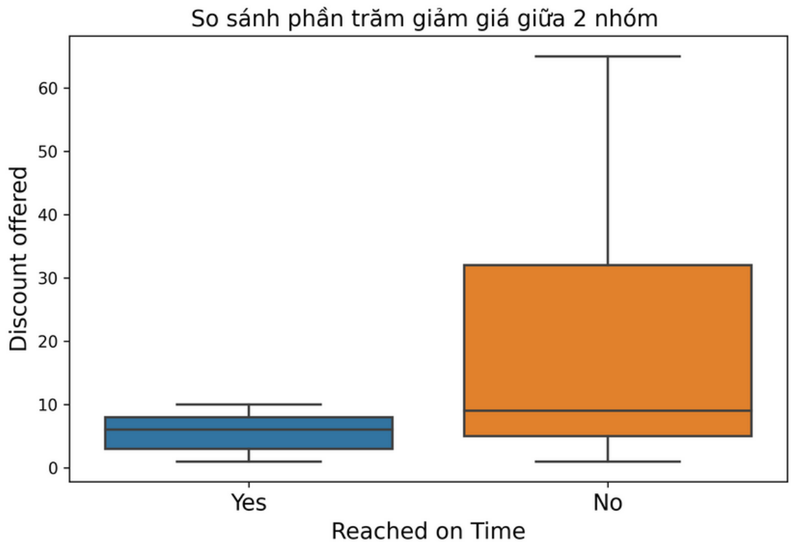
**Check null value:** df.isna().sum()

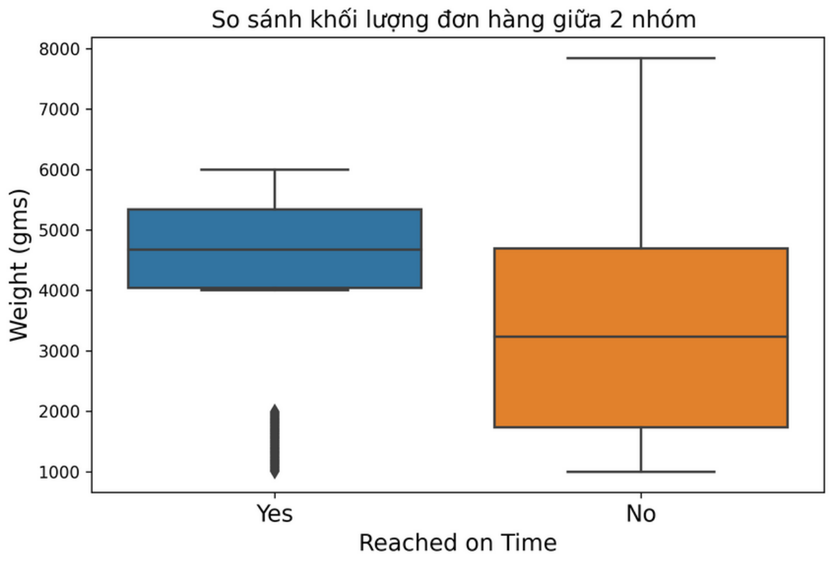
```
Warehouse_block        0
Mode_of_Shipment       0
Customer_care_calls    0
Customer_rating        0
Cost_of_the_Product    0
Prior_purchases        0
Product_importance     0
Gender                 0
Discount_offered       0
Weight_in_gms          0
Reached.on.Time_Y.N    0
dtype: int64
```

**The data is in the correct format, without null value**

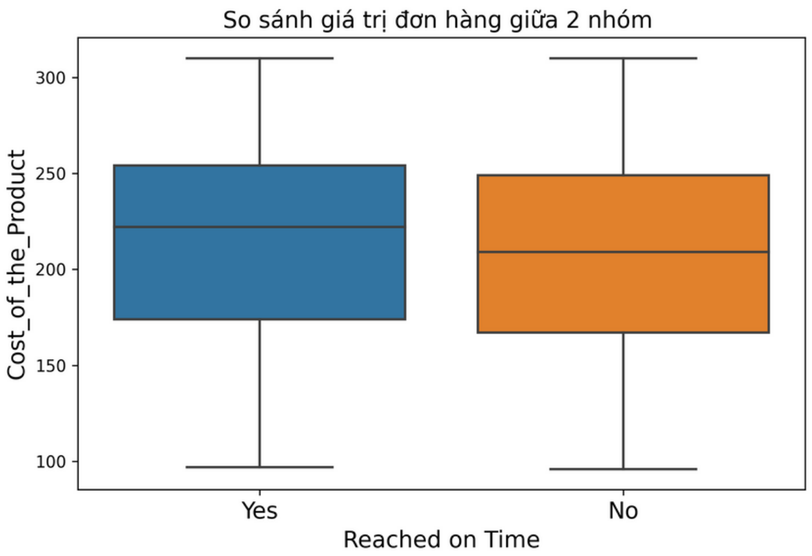
# EXPLORATORY DATA ANALYSIS

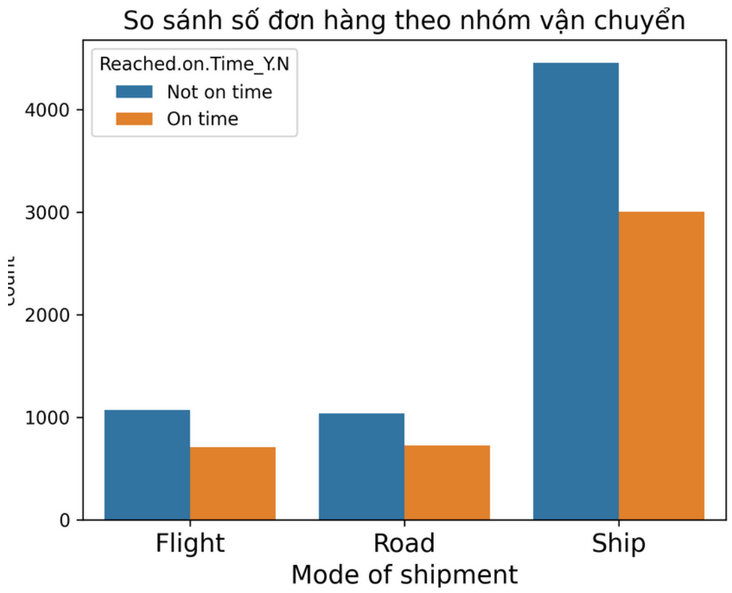


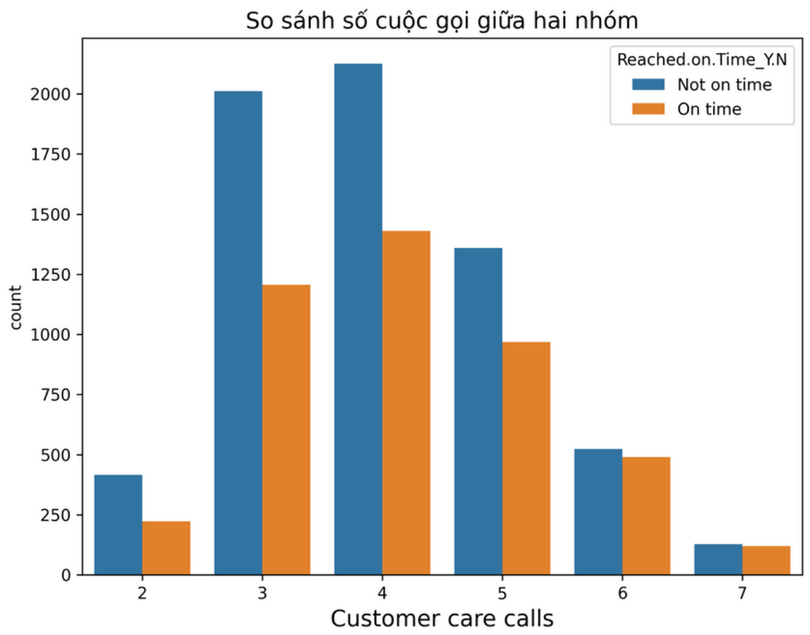



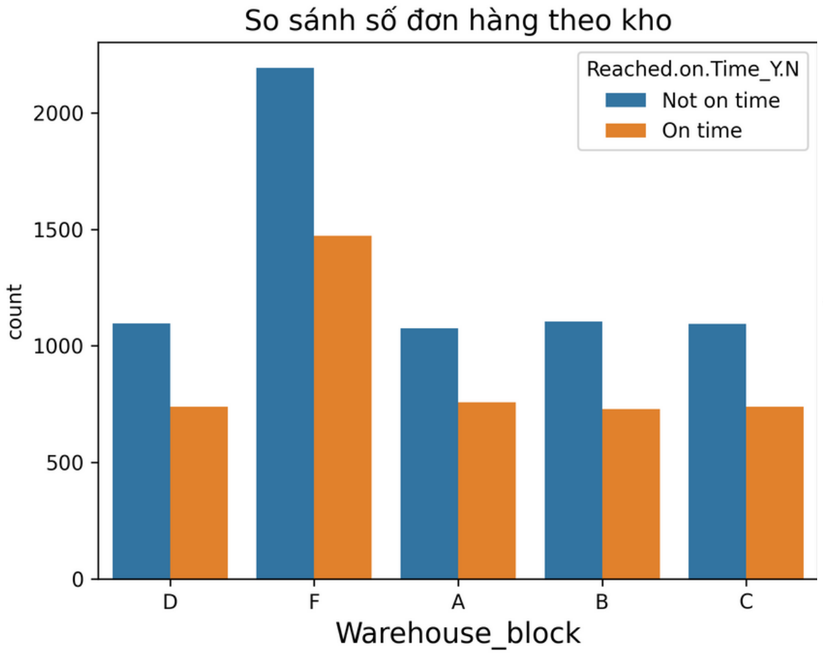

**Build diagrams to illustrate the current performance**

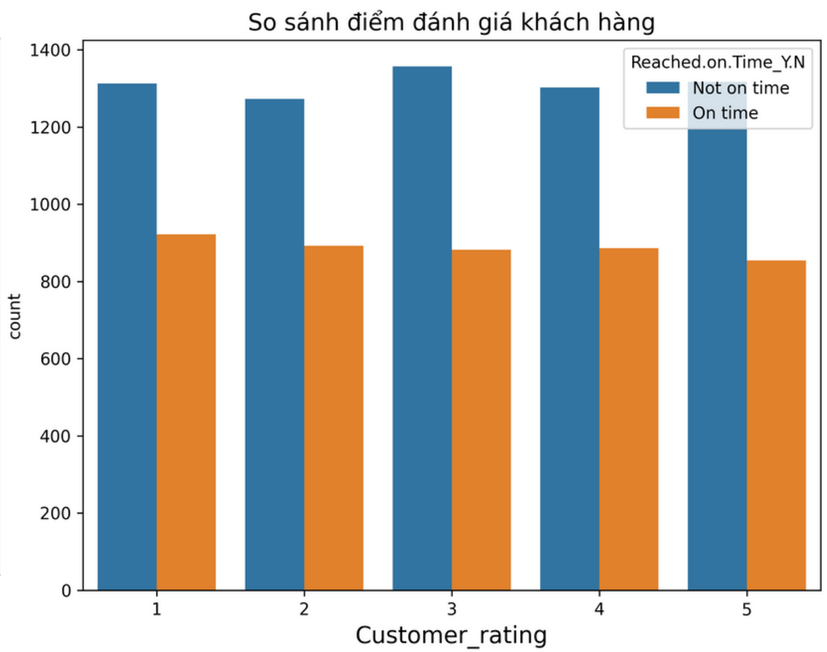


# EXPLORATORY DATA ANALYSIS




## INSIGHTS:

- Nearly 60% of deliveries are late.
- It typically comes on time when the discount is less than 10% and the shipping weighs between 1 and 2 kg and 4 and 6 kg.
- The ratio of late and on-time packages is the same whether using ground or air delivery. However, a large number of the products shipped by ship arrived late.
- Customers who receive deliveries late are more likely to call customer support, regardless of how often they called.
- Wareblock F, which has the most storage space, is where the majority of late packages are placed.
- Customer ratings are the same whether the package arrived on time or not in the review range point.

# DATA PRE-PROCESSING

## Data Encoding

```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['Warehouse_block'] = le.fit_transform(df['Warehouse_block'])
df['Mode_of_Shipment']= le.fit_transform(df['Mode_of_Shipment'])
df['Product_importance']= le.fit_transform(df['Product_importance'])
df['Gender']= le.fit_transform(df['Gender'])
df
```

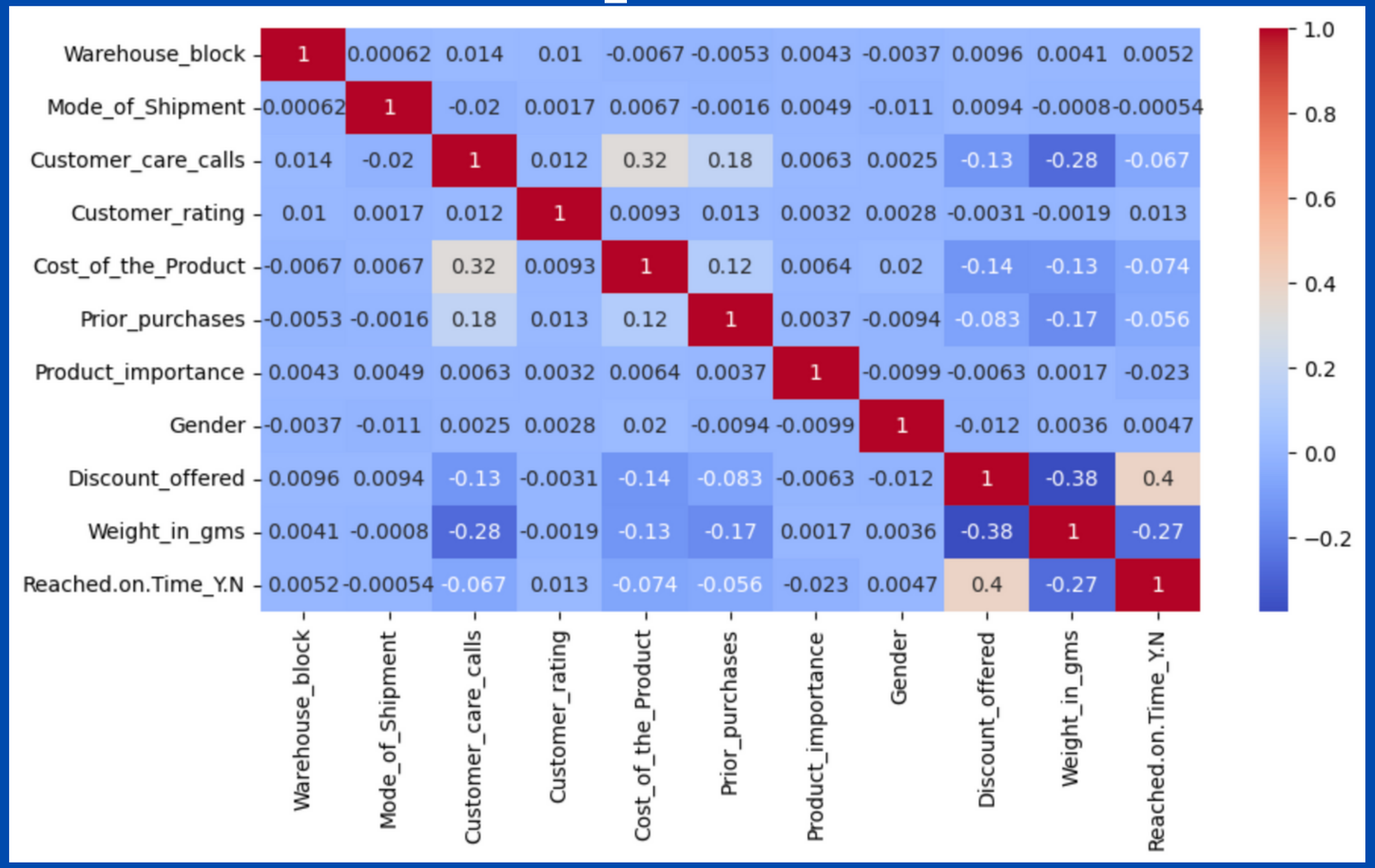
**Before**

| Warehouse_block | Mode_of_Shipment | Product_importance | Gender |
|---|---|---|---|
| D | Flight | low | F |
| F | Flight | low | M |
| A | Flight | low | M |
| C | Flight | medium | F |
| ... | ... | ... | ... |
| A | Ship | medium | F |
| B | Ship | medium | F |
| C | Ship | low | F |
| F | Ship | medium | M |
| D | Ship | low | F |

**After**

| Warehouse_block | Mode_of_Shipment | Product_importance | Gender |
|---|---|---|---|
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 |
| 2 | 0 | 2 | 0 |
| ... | ... | ... | ... |
| 0 | 2 | 2 | 0 |
| 1 | 2 | 2 | 0 |
| 2 | 2 | 1 | 0 |
| 4 | 2 | 2 | 1 |
| 3 | 2 | 1 | 0 |

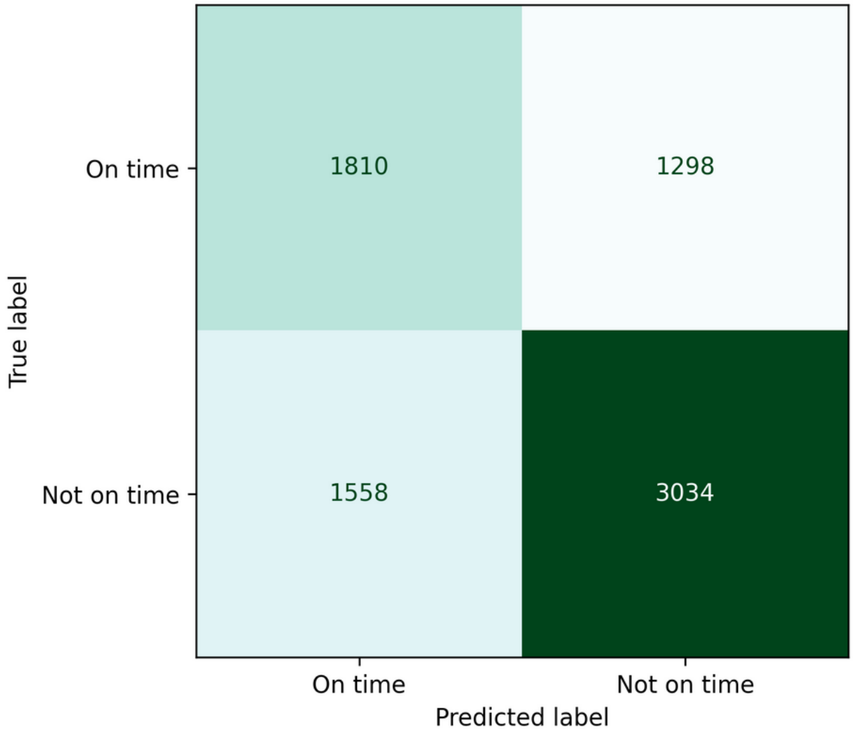
# Correlation matrix between Reached.on.Time_Y.N and other variables




- Discount_offered is the feature that most highly correlates to whether a product is delivered on time

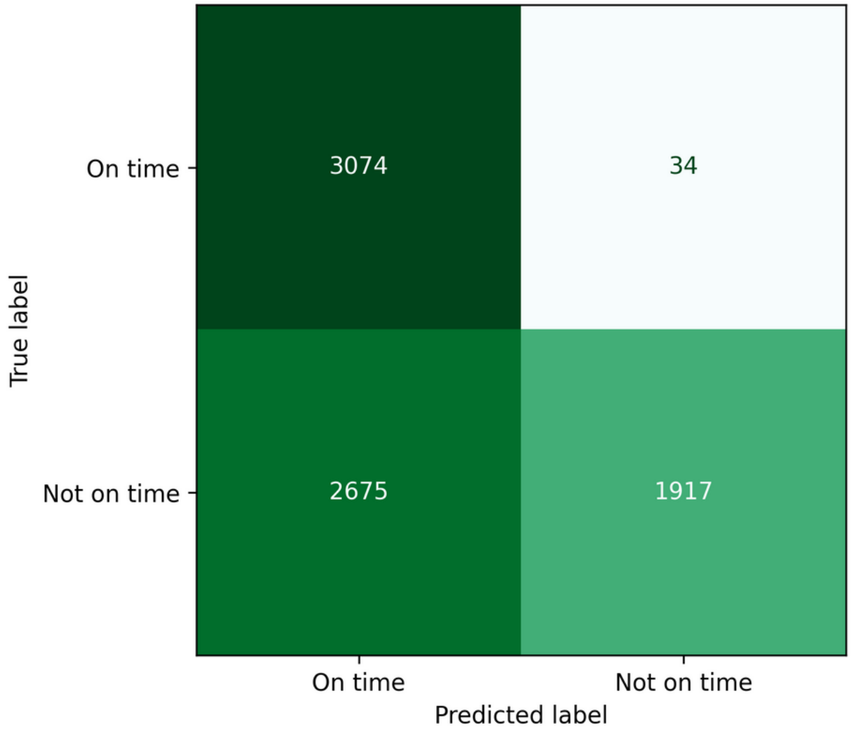
# BUILD MODELS AND EVALUATE

## Logistic Regression




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.54 | 0.56 | 0.55 | 3107 |
| 1 | 0.69 | 0.67 | 0.68 | 4593 |
| accuracy | | | 0.63 | 7700 |

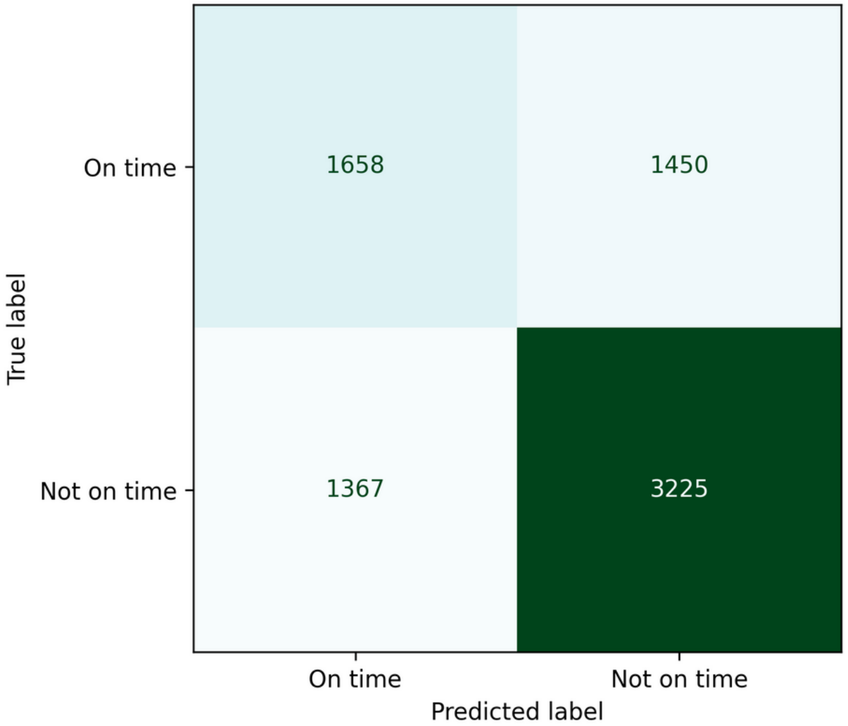
## Gaussian Naive Bayes




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.53 | 0.99 | 0.69 | 3108 |
| 1 | 0.98 | 0.42 | 0.59 | 4592 |
| accuracy | | | 0.65 | 7700 |

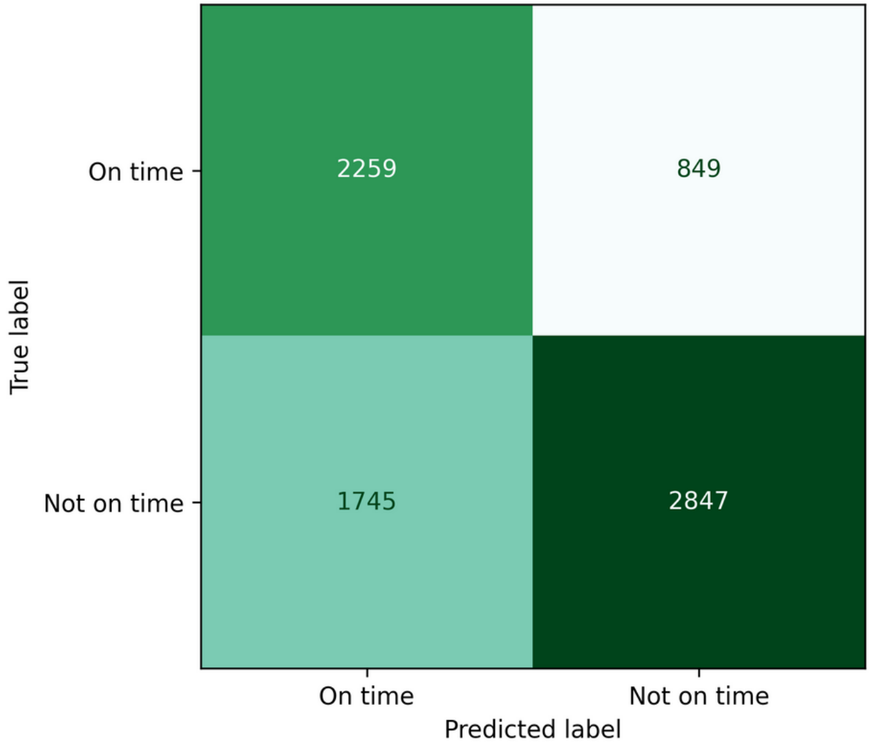
# BUILD MODELS AND EVALUATE

## Decision tree




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.55 | 0.53 | 0.54 | 3108 |
| 1 | 0.69 | 0.70 | 0.70 | 4592 |
| accuracy | | | 0.63 | 7700 |

## Random forest

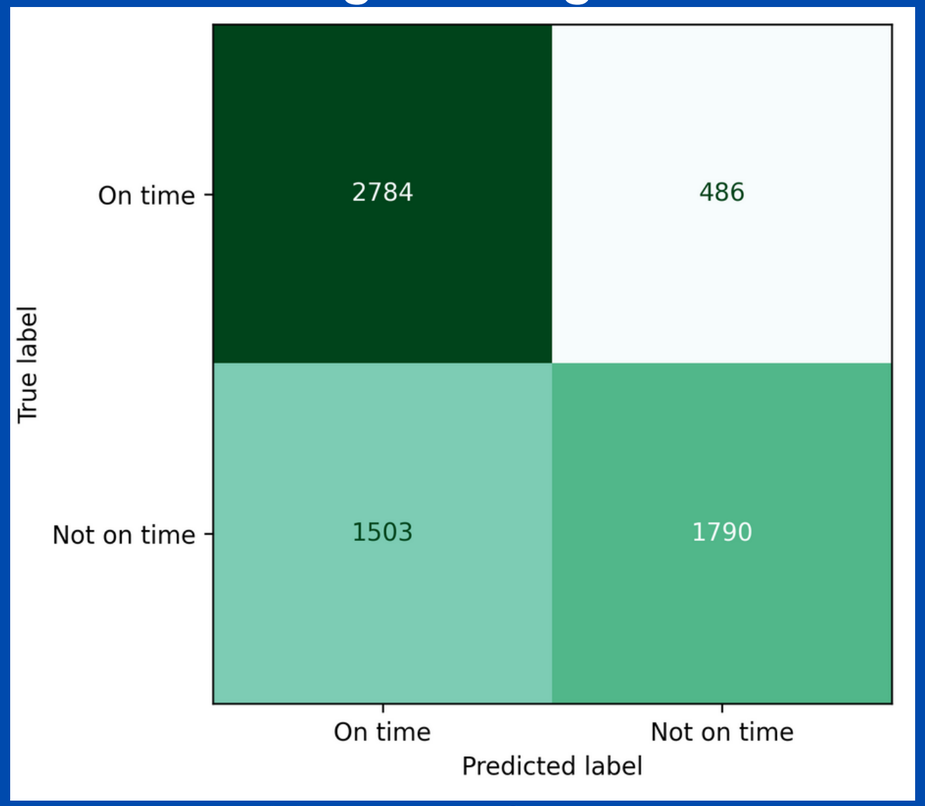



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.56 | 0.73 | 0.64 | 3108 |
| 1 | 0.77 | 0.62 | 0.69 | 4592 |
| accuracy | | | 0.65 | 7700 |

# BUILD MODELS AND EVALUATE

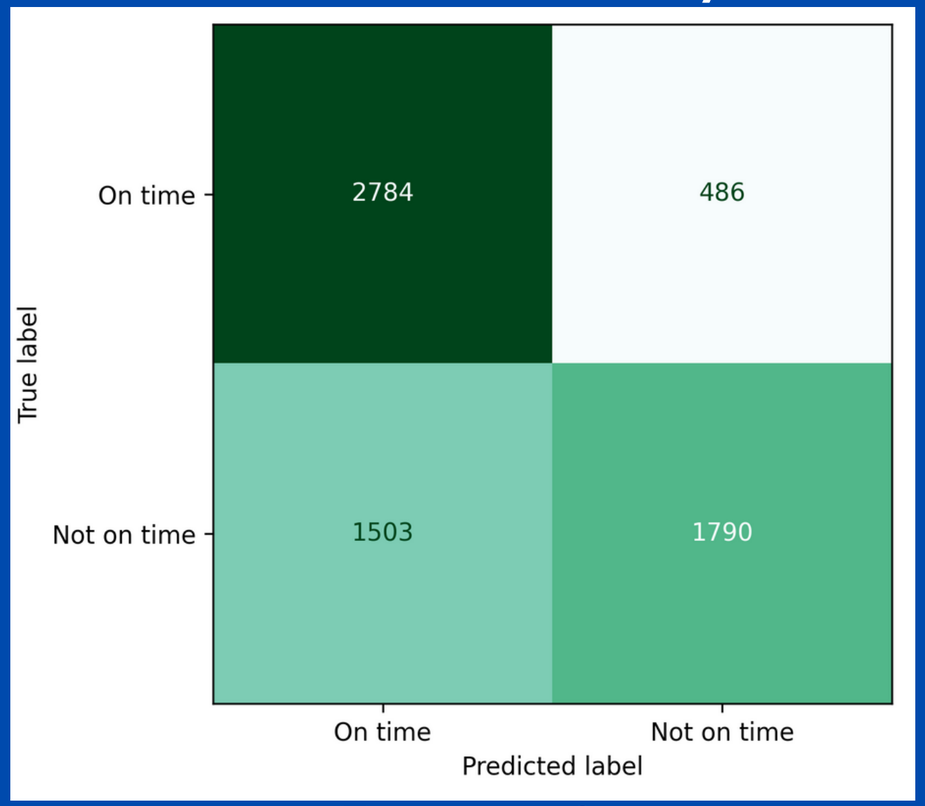
## SMOTE

### Logistic Regression




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.65 | 0.84 | 0.73 | 3270 |
| 1 | 0.78 | 0.55 | 0.64 | 3293 |
| accuracy | | | 0.69 | 6563 |

### Gaussian Naive Bayes

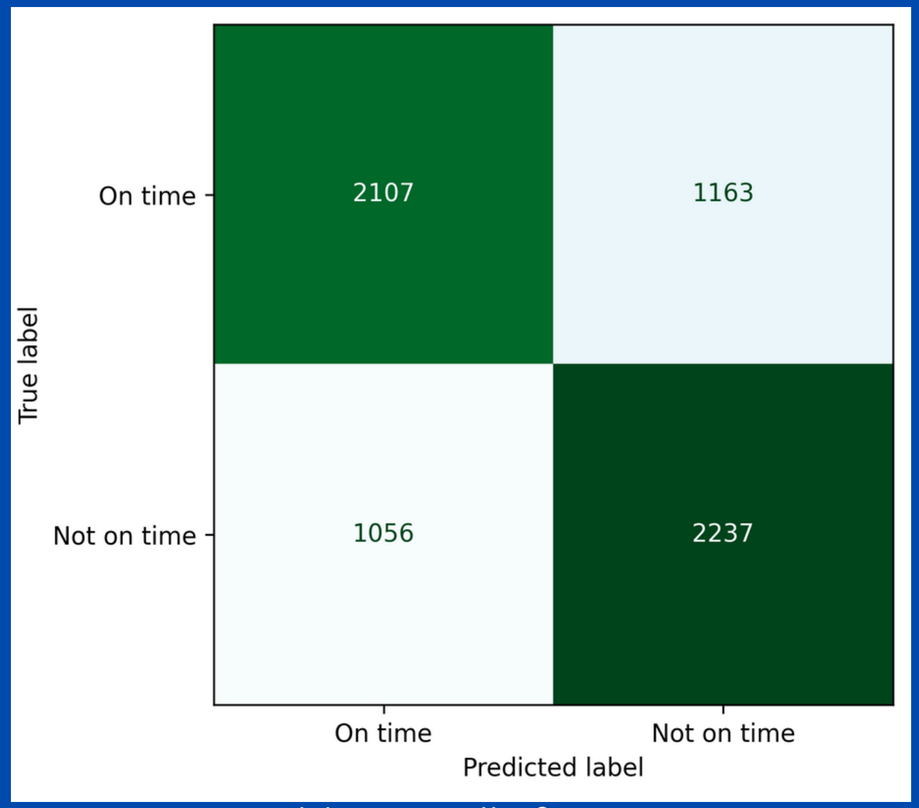



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.63 | 0.99 | 0.77 | 3270 |
| 1 | 0.98 | 0.43 | 0.60 | 3293 |
| accuracy | | | 0.71 | 6563 |

# BUILD MODELS AND EVALUATE

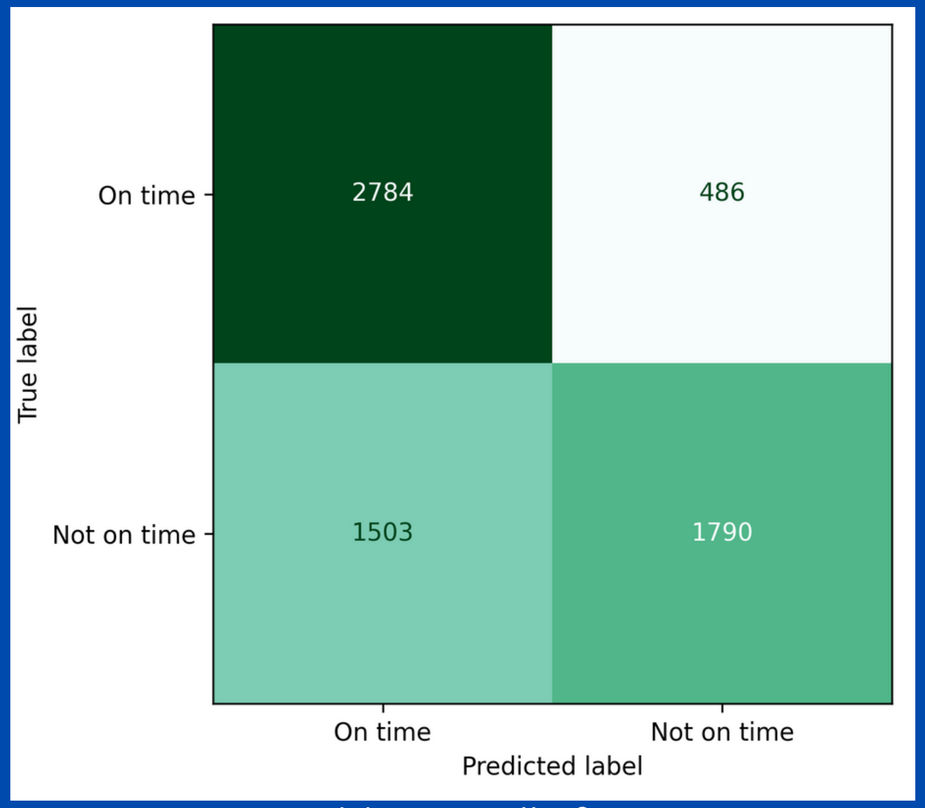
## SMOTE

### Decision tree




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.65 | 0.84 | 0.73 | 3270 |
| 1 | 0.78 | 0.55 | 0.64 | 3293 |
| accuracy | | | 0.69 | 6563 |

### Random forest



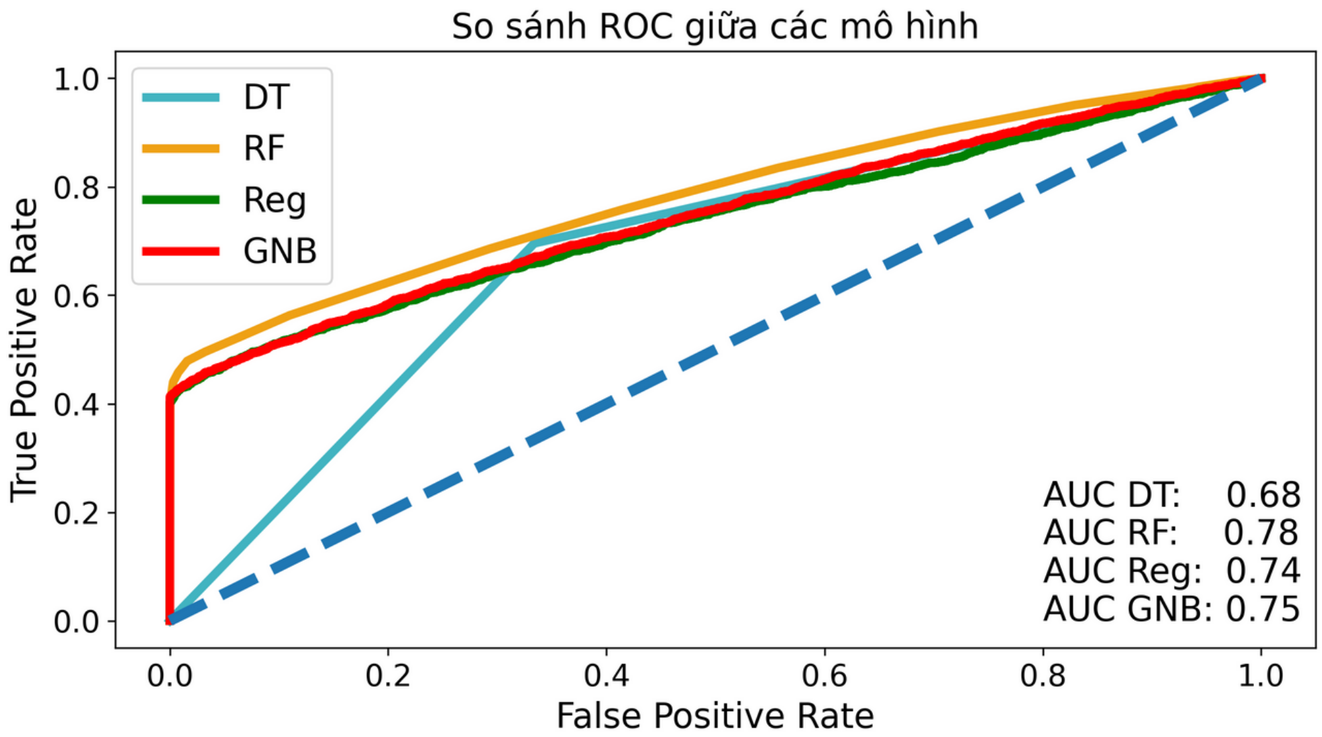

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.81 | 0.74 | 3270 |
| 1 | 0.77 | 0.62 | 0.69 | 3293 |
| accuracy | | | 0.72 | 6563 |

## Before

| Model | | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | On time | 0.54 | 0.56 | 0.55 | 0.63 |
| | Late | 0.69 | 0.67 | 0.68 | |
| Gaussian NB | On time | 0.53 | 0.99 | 0.69 | 0.65 |
| | Late | 0.98 | 0.42 | 0.59 | |
| Decision Tree | On time | 0.55 | 0.53 | 0.54 | 0.63 |
| | Late | 0.69 | 0.70 | 0.70 | |
| Random Forest | On time | 0.56 | 0.73 | 0.64 | 0.65 |
| | Late | 0.77 | 0.62 | 0.69 | |

## After

| Model | | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | On time | 0.65 | 0.84 | 0.73 | 0.69 |
| | Late | 0.78 | 0.55 | 0.64 | |
| Gaussian NB | On time | 0.63 | 0.99 | 0.77 | 0.71 |
| | Late | 0.98 | 0.43 | 0.60 | |
| Decision Tree | On time | 0.65 | 0.84 | 0.73 | 0.69 |
| | Late | 0.78 | 0.55 | 0.64 | |
| Random Forest | On time | 0.68 | 0.81 | 0.74 | 0.72 |
| | Late | 0.77 | 0.62 | 0.69 | |




Random forest Classification model has the highest predicting rate of on-time delivery in F1-score and Accuracy.

# THANK YOU

## DO YOU HAVE ANY QUESTION?

End of Slide