E-Commerce Shipping Analysis

Shipment of products delivered on time or not?

gizkaolivia@gmail.com
www.linkedin.com/in/gizka-olivia
github.com/gizkaolivia

Curriculum Vitae (CV)

Educational Backgroud

Diponegoro University (2017-2021)
S1 Computer Science/Informatics - GPA 3,82
Dibimbing (Mei 2022 – September 2022)
Data Science Bootcamp
Digital Talent Scholarship - Kementerian Komunikasi dan Informatika (Juli 2022 – September 2022)
Big Data Using Python

Project Experience

- EDA on G-Connect Mobile Ground Sensor
- EDA & Modeling on Health Insurance Cross Sell Prediction:
  Classification
- EDA & Modeling on E-Commerce Shipping Data:
  Classification

Work & Organizational Experience

Dinas Komunikasi dan Informatika Banjarnegara (2020)
IT Support Intern
- Developing a company profile website for the Regional Secretariat

Informatics Student Association (HMIF Undip) - Economics and Finance Division (2018 and 2019)
Treasurer
- Create monthly reporting and submit it with profit results to the HMIF Treasurer at the end of the month *Coordinator Informatics Market*
- Providing the needs of Undip Informatics students both to support lecture activities and other activities
- Create monthly reporting on sales of Infomart at the end of the month

Outline

01 **Problem Statement**
Contains the background of the problem and the purpose of the project

02 **Data Description**
Description of each column in E-Commerce Shipping Data

03 **Data Understanding**
Understand data with Exploratory Data Analysis (EDA)

04 **Data Pre-Processing**
Perform data cleaning, feature selection, feature engineering, imbalanced data handling, and feature scaling

Outline

Problem Statement

Parcel Perform is a software-as-a-service (SaaS) package capable of tracking more than 600 logistics operators globally. Parcel Perform found that there were more than 90 percent of complaints and negative responses from customers related to late delivery. The survey shows 35 percent of customers continue to view shipping as the biggest problem in e-commerce. Optimizing the delivery experience is crucial to increasing the benefits that customers receive because customer satisfaction is the key to customer loyalty. Therefore, the company needs to improve goods delivery services.

Problem:
An international e-commerce company based wants to discover key insights from their customer database. They want to use some of the most advanced machine learning techniques to study their customers. This company sells electronic products.

Objective:
Build a model to predict whether ordered products are delivered on time or not

Data Description

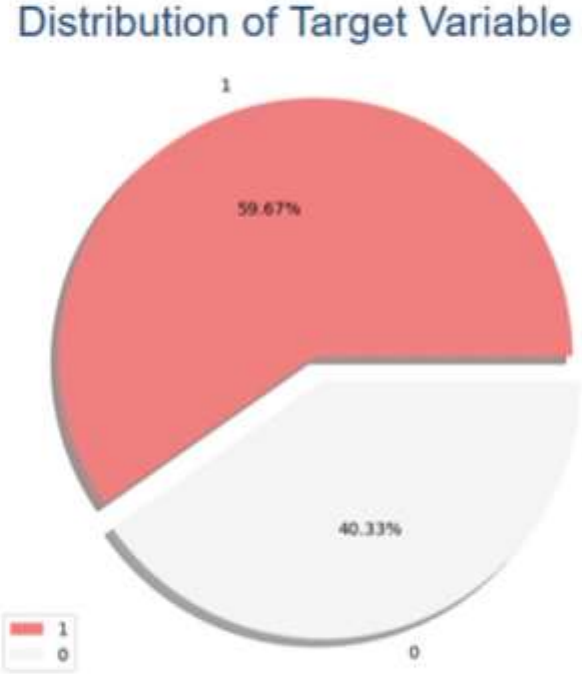
Source: https://www.kaggle.com/datasets/prachi13/customer-analytics
The dataset contained 10999 rows observations of 12 variables

Variable	DESCRIPTION
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 reac hed on time

Data Understanding

- Late = 60%
- On time = 40%

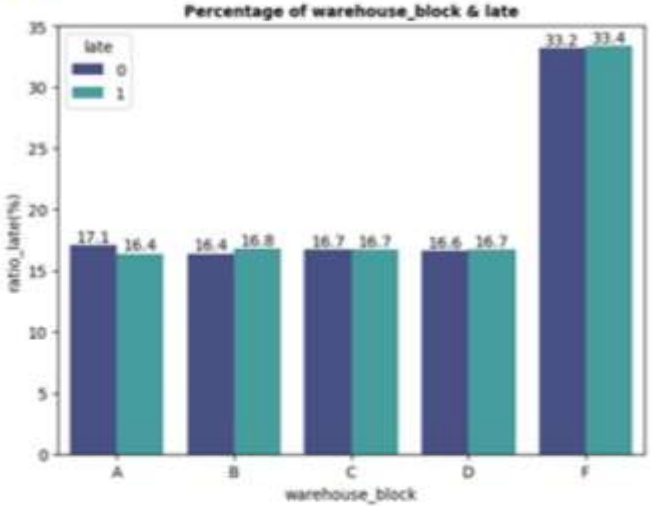
60% of products have delivery delays

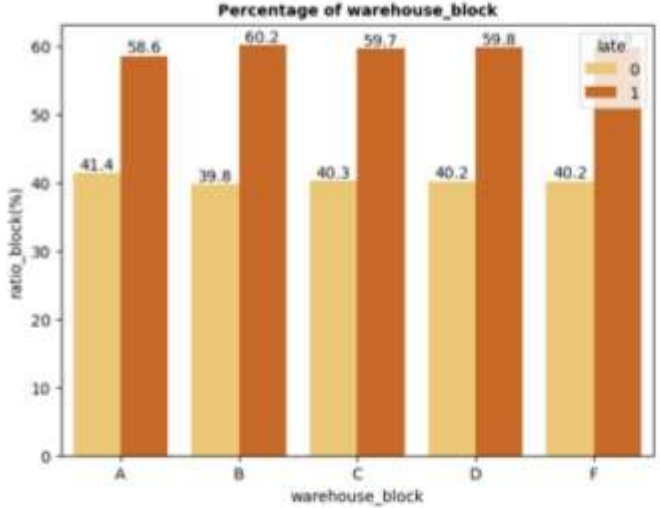


Data Understanding

Exploring of Warehouse_Block

- Shipments from *warehouse_block* F have a higher volume of on time shipments compared to other blocks even though th ey have almost the same difference in the percentage of lates (< 1%). But, *warehouse_block* F can accounts for 33% of all shipment volume.

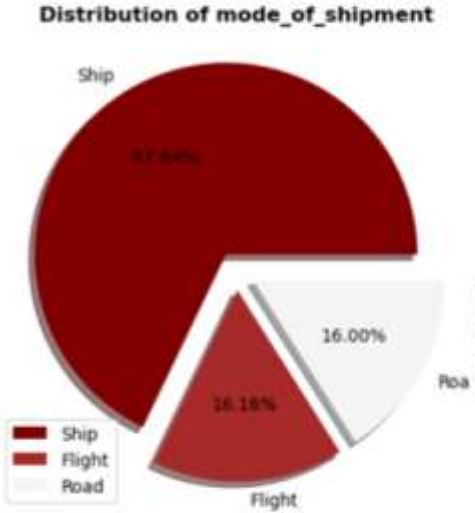




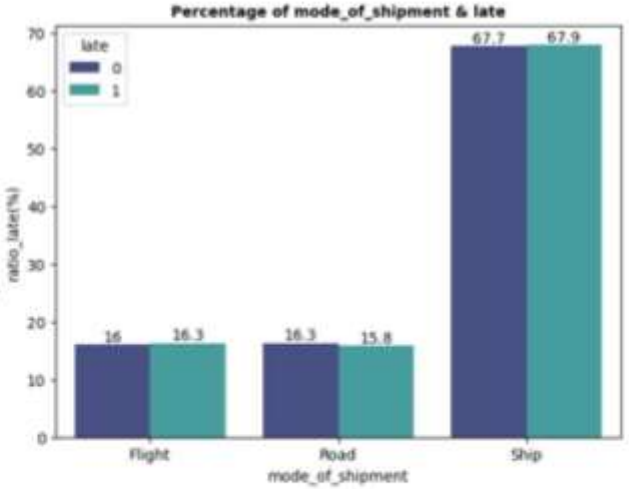
- However, shipments from block A have a better on-time percentage and a smaller late percentage
- Block B can be said to have the worst shipment, this is indicated by the smallest percentage of punctuality and the highest percentage of delays. However, its comparison with other blocks is not too far

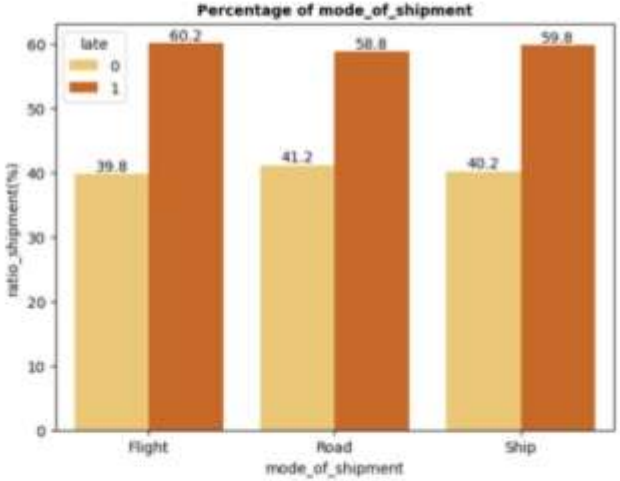
Data Understanding

Exploring of Mode of Shipment



- 68% of all deliveries are made by ship. Shipment delays by Ship tend to be higher due to higher shipping volumes.



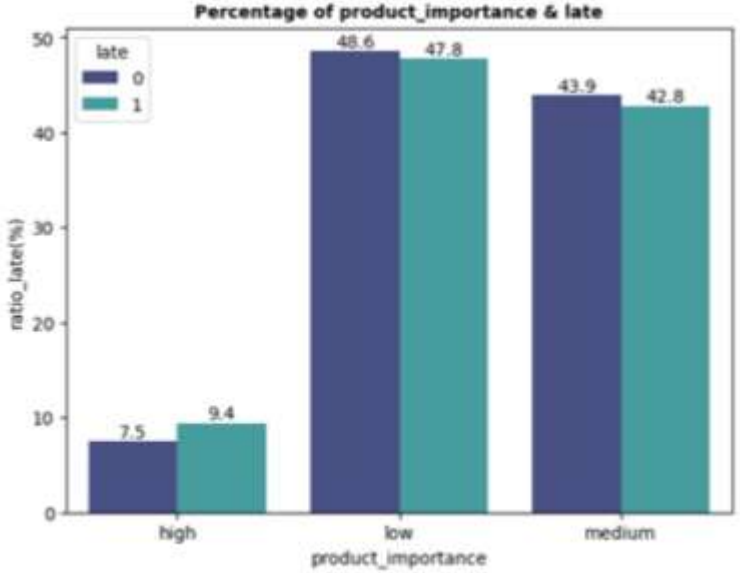


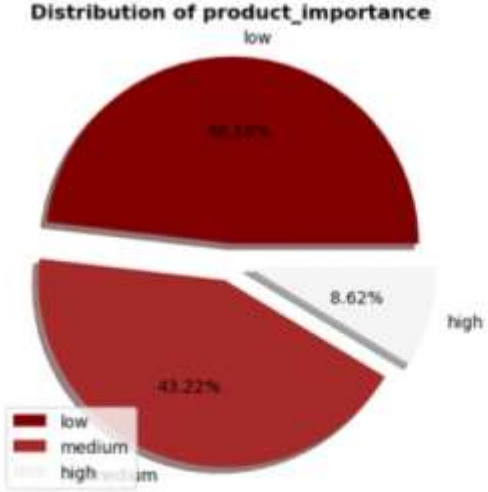
- Each mode_of_shipment has almost the same performance in shipping products.

Data Understanding

Exploring of Product of Importance

- Products with medium and low importance show a larger total late shipments due to higher shipping volume, but actually on time shipments have a higher percentage than late shipments.
- In contrast to product_importance high which tends to be late shipment with small shipping volume, it's only 9% of the total shipping volume.

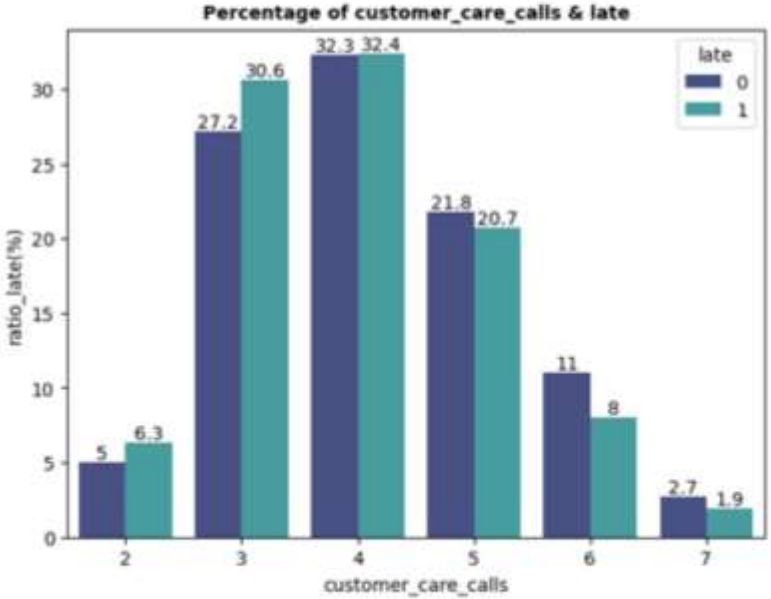


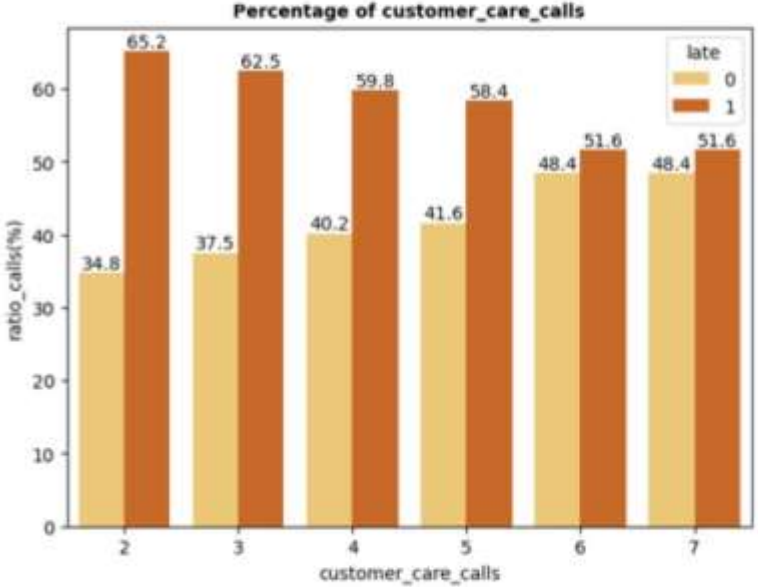


Data Understanding

Exploring of Customer Care Calls

- More than 80% of customers make 3-5 calls during the shipment process

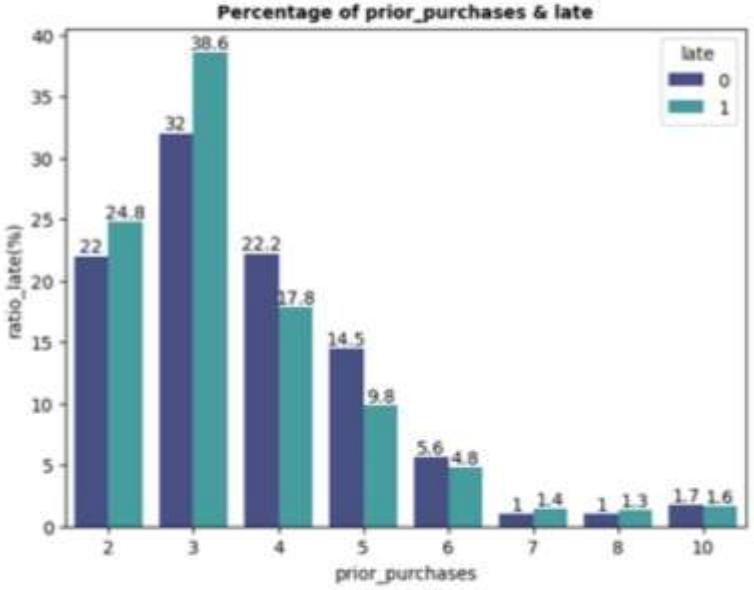




- The more often the customer calls, the higher the chance that the shipment will be made on time

Data Understanding

Exploring of Prior Purchases



- The highest shipment delay occurs in customers who previously made 2-3 purchases. This is also influenced by the high volume of shipments, which is 60%
- prior_purchases above 5 times tends to experience on time shipment

Data Understanding

Exploring of Cost of the Product

cost_of_the_product	delayed_shipments	value_count
(95.786, 117.4]	66.216216%	2.018365%
(117.4, 138.8]	66.353383%	4.836803%
(138.8, 160.2]	63.161609%	12.882969%
(160.2, 181.6]	62.960180%	12.101100%
(181.6, 203.0]	62.197802%	12.410219%
(203.0, 224.4]	58.882083%	11.873807%
(224.4, 245.8]	57.546012%	14.819529%
(245.8, 267.2]	56.097561%	17.519775%
(267.2, 288.6]	57.187828%	8.664424%
(288.6, 310.0]	46.835443%	2.872988%

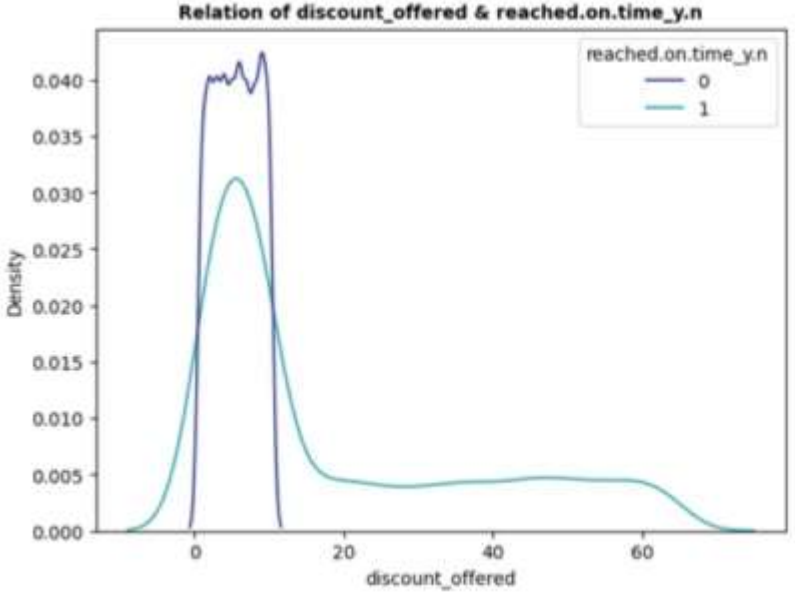
The higher the cost_of_the_product, the greater the possibility of on time shipment

Data Understanding

Exploring of Discount Offered

- Shipments that have a discount of less than 13.8 tend to experience on time shipments, these shipments account for 77 % of the total volume
- All shipments that have a discount offer greater than 13.8 experience delays

discount_offered	delayed_shipments	value_count
(0.936, 7.4]	47.297297%	53.150286%
(7.4, 13.8]	49.212894%	24.256751%
(13.8, 20.2]	100.000000%	3.036640%
(20.2, 26.6]	100.000000%	2.418402%
(26.6, 33.0]	100.000000%	2.818438%
(33.0, 39.4]	100.000000%	2.682062%
(39.4, 45.8]	100.000000%	2.554778%
(45.8, 52.2]	100.000000%	3.391217%
(52.2, 58.6]	100.000000%	2.627512%
(58.6, 65.0]	100.000000%	3.063915%

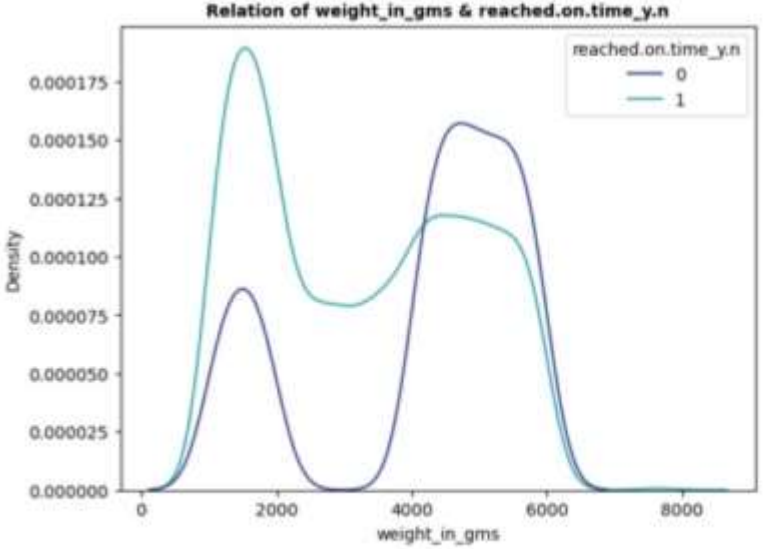


Data Understanding

Exploring of Weight in Gms

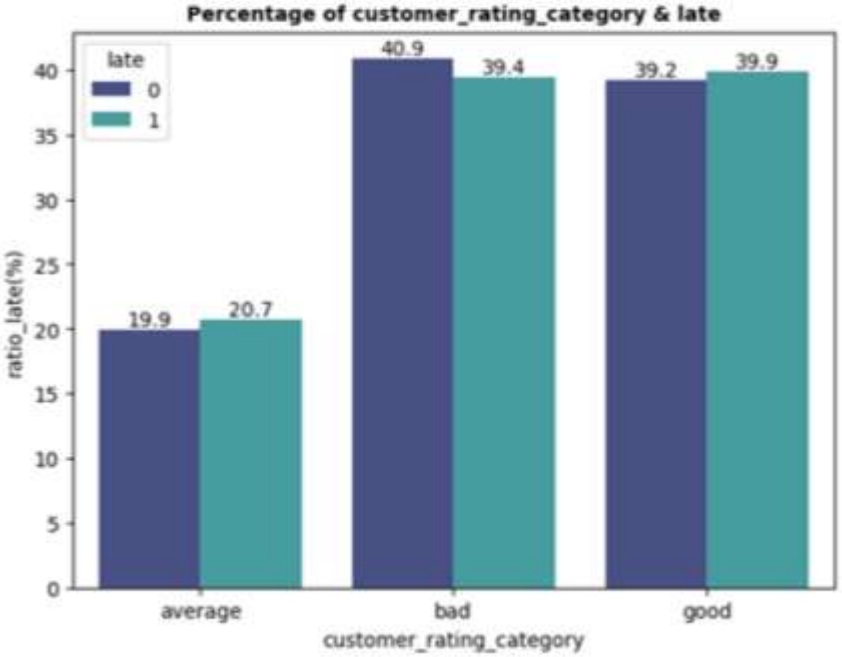
- Shipment of products weighing less than 4000 grams (4kg) tends to be late, while those more than 4kg tend to be on time. Shipments more than 4kg account for 56% of the total volume
- All shipment of products weighing 2370-3739 grams and more than 6477 grams are delayed

weight_in_gms	delayed_shipments	value_count
(994.155, 1685.5]	67.610063%	20.238203%
(1685.5, 2370.0]	75.979305%	12.301118%
(2370.0, 3054.5]	100.000000%	5.473225%
(3054.5, 3739.0]	100.000000%	5.673243%
(3739.0, 4423.5]	53.741054%	13.973998%
(4423.5, 5108.0]	42.153549%	18.828984%
(5108.0, 5792.5]	43.028486%	18.192563%
(5792.5, 6477.0]	41.105354%	5.264115%
(6477.0, 7161.5]	100.000000%	0.009092%
(7161.5, 7846.0]	100.000000%	0.045459%



Data Understanding

Which rating has an influence on delivery delays, good or bad rating?

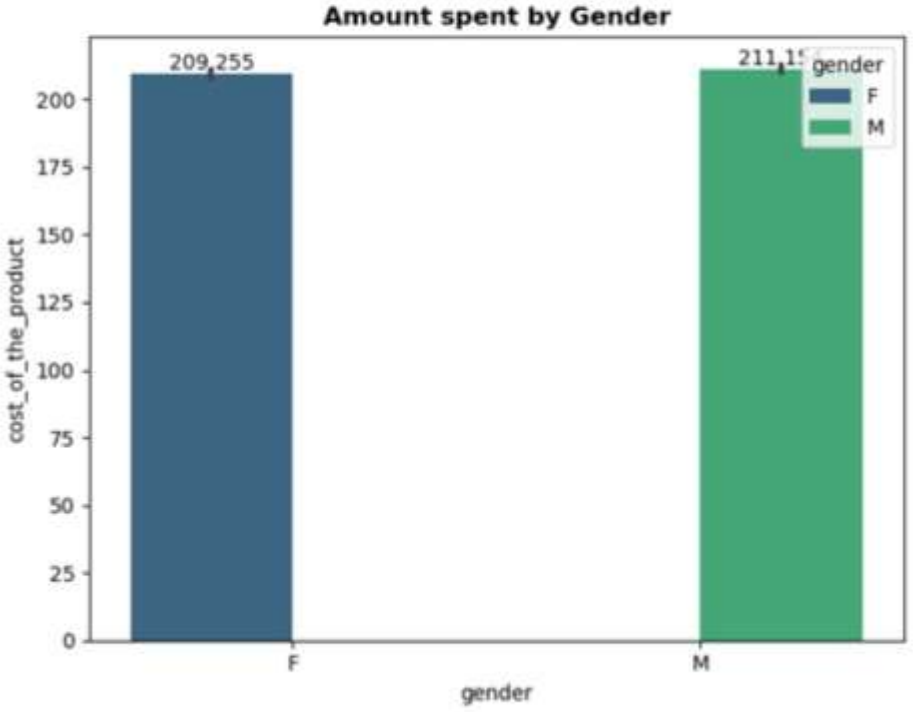


Categorize customer ratings into 3, namely

- Rating 1-2 = Bad
- Rating 3 = Average
- Rating 4-5 = Good

Surprisingly, bad ratings are given to shipments that tend to be on time compared to good ratings

Data Understanding

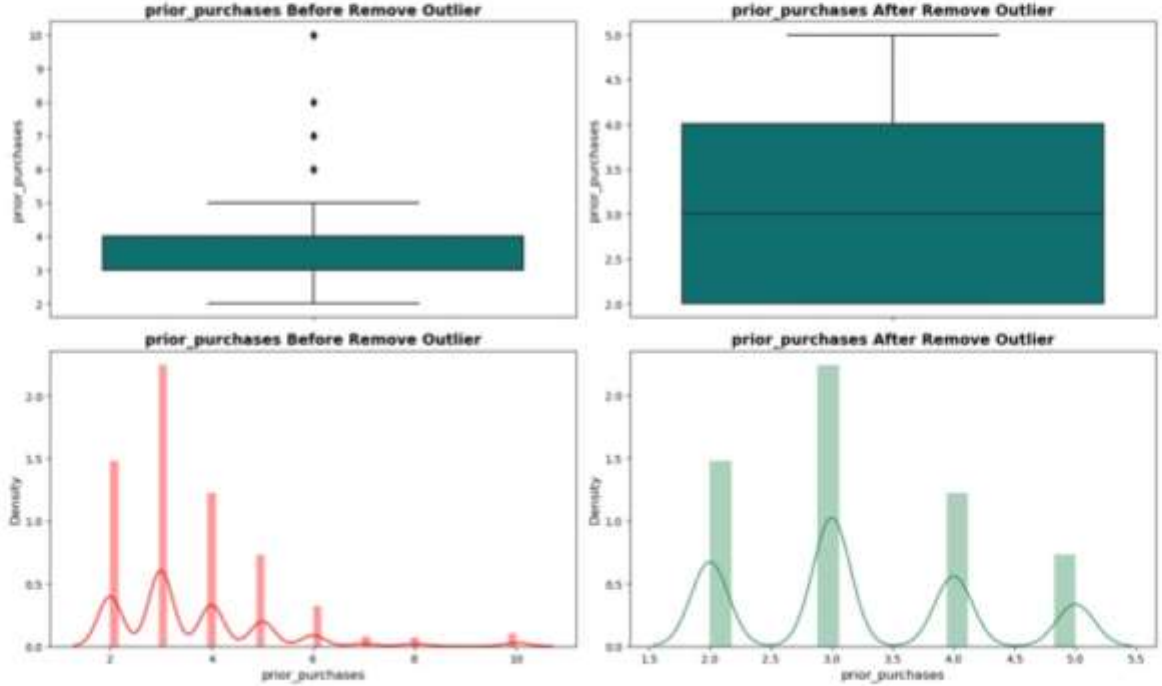


How do customers spend money to buy products?

Both Female and Male look the same in spending money when buying a product

Data Pre-Processing

Remove Outlier



Data Pre-Processing

Feature Engineering

Label Encoding → mode_of_shipment
Ordinal Encoding → warehouse_block, product_importance, customer_rating_category

Feature Selection

Drop the grouped numeric column → customer_rating
Drop unnecessary columns → id and gender

Imbalanced Data

Because the ratio of target variables is 40% and 60%, there is no need to handle imbalance data

Feature Scaling

We normalize the data on features that have a skewed distribution, the rest I do standardization.

Machine Learning Model

Evaluation Model

Model	Precision		Recall		F1-Score		ROC-AUC		Accuracy	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
LogisticRegression (Baseline)	0.72601	0.72957	0.64857	0.68429	0.68511	0.70620	0.64330	0.65722	0.64431	0.66227
LogisticRegression	0.71256	0.70892	0.73258	0.69466	0.72243	0.70172	0.63477	0.60762	0.65637	0.62973
DecisionTreeClassifier	0.71599	0.73248	0.73932	0.70428	0.72747	0.71810	0.63990	0.63596	0.66186	0.65331
RandomForestClassifier	0.74906	0.77322	0.67415	0.67716	0.70963	0.72201	0.66014	0.67167	0.66323	0.67306
KNeighborsClassifier	0.70462	0.77322	0.70224	0.67716	0.70343	0.72201	0.62049	0.67167	0.63854	0.67306
GaussianNB	0.99469	0.98547	0.42134	0.41557	0.59194	0.58461	0.70891	0.70263	0.64540	0.62973
SVC	0.78116	0.77437	0.66179	0.61854	0.71654	0.68774	0.68565	0.65780	0.68038	0.64783
XGBClassifier	0.72944	0.75825	0.68764	0.68329	0.70792	0.71882	0.64399	0.65855	0.65363	0.66483

Decision Tree has the highest recall score

Recall is used to suppress the number of False Negatives, namely the number of deliveries that are actually delayed but are predicted to be on time

Machine Learning Model

Evaluation Model of Decision Tree

Based on the evaluation using Recall, it is can assumed that 728 (TN : 386 + FN : 338) orders are predicted to be delivered on time. Where 338 orders were predicted to be wrong, actually experienced delays in delivery. This model can correctly predict 386 orders delivered on time.

Conclusion

Question 1 : Find which features has high contribution

- Discount_offered with positive correlation

Question 2 : Find the best model to predict on-time shipment reach

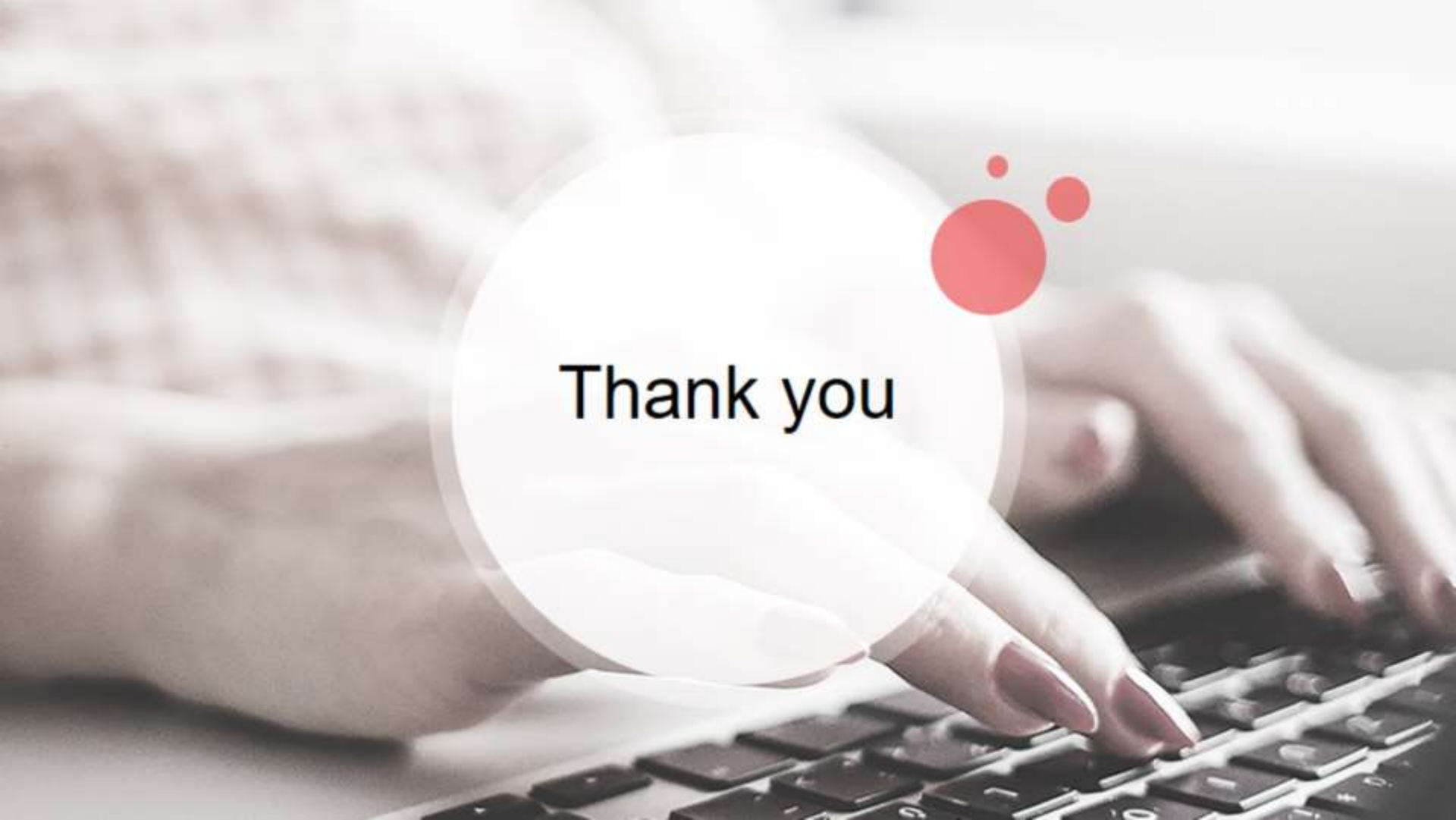
- Decision tree has the highest recall score compared to other models, which is 70%

Question 3: Find the factors that make product shipment on time

- A large discount (>14%) can cause the number of sales to increase, so that the shipment volume increases and causes delay shipments
- Likewise with cost_of_the_product, the more expensive the product sold, the higher the possibility of delivery on time
- Low and medium importance products actually experience on-time delivery.
- Shipments by road and block F tend to be on time
- Make a call to the seller at least 2 times
- Shipment of products weighing > 4kg tends to be on time

Recommedation

- Anticipating shipping overload by choosing the right expedition. One of the causes of overload is during big discounts so that cost of product become cheaper, this can happen during major holidays, year-ends, etc. Overload can also occur in products that have high importance.
- In addition to choosing the right expedition, during peak season the company can provide information that there is a possibility of delivery delays with the delivery range being longer than usual days.
- The company can distribute products well, so that there is no accumulation of goods in one block, especially in block F.


Thank you