E-COMMERCE

DATA ANALYSIS

Improving Shipment Time Prediction

2022

GROUP 9

Clarence Wong

Sociology

Ellen Kwok

Cognitive Science

Kijahre Fikiri

Business Administration

Natalia Nava

Business Administration

2022

TABLE OF CONTENTS

2022

01

PROJECT OVERVIEW

2022

Company Overview

- International e-commerce company
  - *Sells and ships electronic products both domestically and internationally*
  - Sales: Over ***$2M*** *per business cycle*
- Assumption: *Based in United States*
- Analyzed: *Shipping Data*

Industry Overview

- The United States electronics e-commerce market: ***Projected to grow 16%*** *within 3yrs (from $131,491 million in 2022 to $156,779 by 2025)*
- The Global electronics e-commerce market: *Projected to reach* ***$511 billion*** *by 2025.*

Challenges

- Cost function assumptions
- Complexity and Simplification
- Model Accuracy vs. Practicality

Project Goal

- Predict if an item will arrive on time to minimize **money spent on refunds for late deliveries**

SCOPE

THE COMPANY

an international e-commerce company that sells electronic products

DATA

Customer rating and shipping data

SOURCE

Kaggle: https://www.kaggle.com/datasets/prachi13/customer-analytics

2022

02

OPPORTUNITY IDENTIFICATION

2022

Variable	Description	Assumption and Justification
ID Number of Customers	ID Number of Customers	No assumptions made.
Warehouse Block	The company has a warehouse which is divided into blocks: A,B,C,D,F.	No assumptions made.
Mode of Shipment	The company ships via: ship, flight and road.	No assumptions made.
Customer Care Calls	The number of calls made from enquiry for enquiry of the shipment.	We assumed that this occurs after a product is shipped.
Customer Rating.	1: Worst Rating. 5: Best Rating	We assumed that this occurs after a product is shipped.
Cost of the Product	Value in US Dollars	No assumption made.
Prior Purchases	Number of Prior Purchases	No assumption made.
Product Importance	Company categorizes the product in various parameters: Low, medium, or high.	No assumption made.
Gender	Male or Female	No assumption made.
Discounts Offered	Discount offered on that specific product. This value is a percentage. *(The percent discount offered from the cost).*	We assume that the discounts offered are an outcome after product arrival, with the company offering more discounts to compensate for late arrival.
Weight in grams.	Weight in grams.	No assumptions made.
Reached on Time	The Target Variable. Original Dataset: 1: Product has not reached on time. 0: Product has reached on time.	For the notebook we made: 1: Product has reached on time. 0: Product has not reached on time.

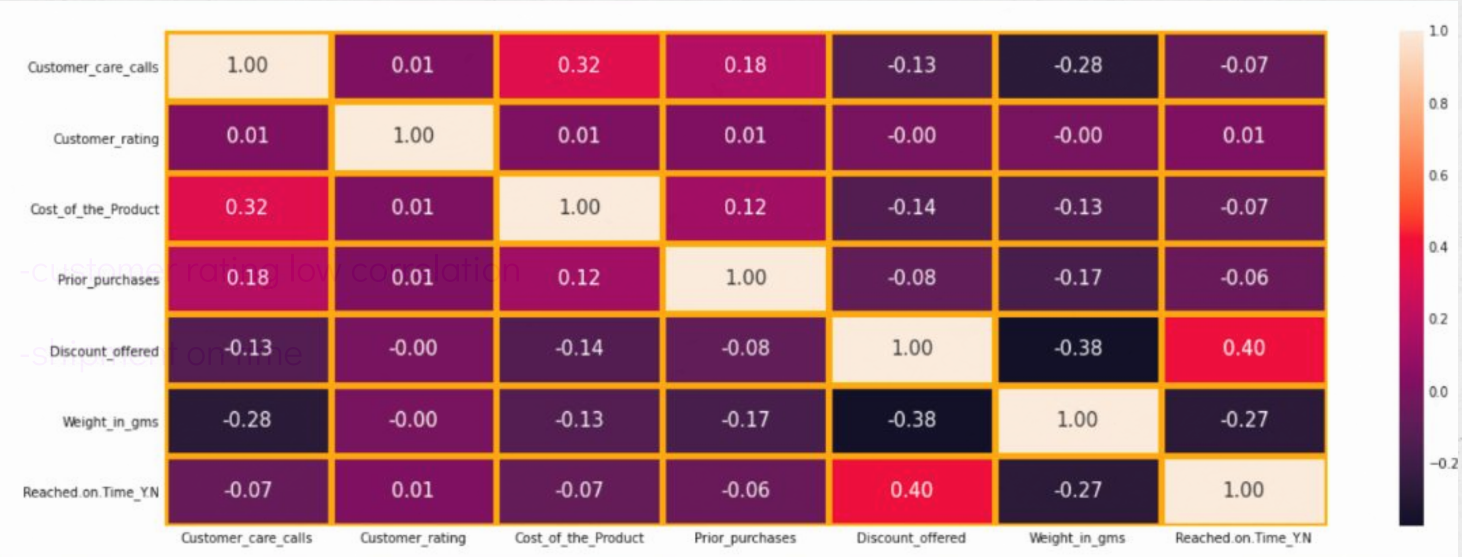
WHAT PROBLEMS IN E-COMMERCE CAN WE SOLVE?

01 Improve customer rating

02 Improve shipment timeliness

Exploratory Data Analysis

Correlation Matrix



Correlation Matrix

	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
Customer_care_calls	1.00	0.01	0.32	0.18	-0.13	-0.28	-0.07
Customer_rating	0.01	1.00	0.01	0.01	-0.00	-0.00	0.01
Cost_of_the_Product	0.32	0.01	1.00	0.12	-0.14	-0.13	-0.07
Prior_purchases	0.18	0.01	0.12	1.00	-0.08	-0.17	-0.06
Discount_offered	-0.13	-0.00	-0.14	-0.08	1.00	-0.38	0.40
Weight_in_gms	-0.28	-0.00	-0.13	-0.17	-0.38	1.00	-0.27
Reached.on.Time_Y.N	-0.07	0.01	-0.07	-0.06	0.40	-0.27	1.00

Customer rating has low correlation with any variables!

2022

03

PROBLEM SPECIFICATION

We additionally found that...

60%

of the product were NOT shipped on time

0.4

of correlation between product arrival late and discount offered

Moreover... discounts are only offered when a product is late!

BY SLIDESGO

2022

$294,710

The amount of money this company spent on discount per business cycle

Goal: Improve product arrival time to lower costs required to compensate for late product arrival

Solution: Create a model to predict what product will be late and suggest a change in carrier

04

MODEL DEVELOPMENT

2022

Our Goal

Using logistic regression and decision tree, build a model that can predict whether a product arrive on time

Cost Function

	Value	Significance
TP	0	0
FP	-$44,90	Cost of discount offered
TN	-$1.51	Cost of changing the carrier
FN	-$1.51	Cost of changing the carrier

Shipping Cost Estimation

What we found in the Data:

- **84%** of products were shipped **internationally**
- **16%** were shipped **domestically**
- **Average weight** of the packages is **8 pound**

Assumption we made:

- **the company uses USPS**

International

The following countries are chosen as they are the most popular countries online companies sell to.

Shipping carrier	US-UK	CA-China	CA-Australia	CA-Canada	CA-Germany	Average
USPS	$88.83	$86.26	$94.34	$58.15	$77.71	N/A
DHL	$81.94	$94.13	$97.57	$57.94	$81.94	N/A
Cost of Changing	$-6.89	$7.87	$3.23	$-0.21	$4.23	$1.65

Domestic

We take the average cost of shipping to the furthest and closest state from California.

Shipping carrier	CA-NY	CA-NEVADA	Average
USPS	$16.1	$16.1	N/A
ShipBob	$17.26	$16.5	N/A
Cost of Changing	$1.16	$0.4	$0.78

84% of products were shipped internationally and 16% were shipped domestically. **As a result the weighted average of changing carriers for late products is $1.5108** ($1.65*0.84 + $0.78 * 0.16).

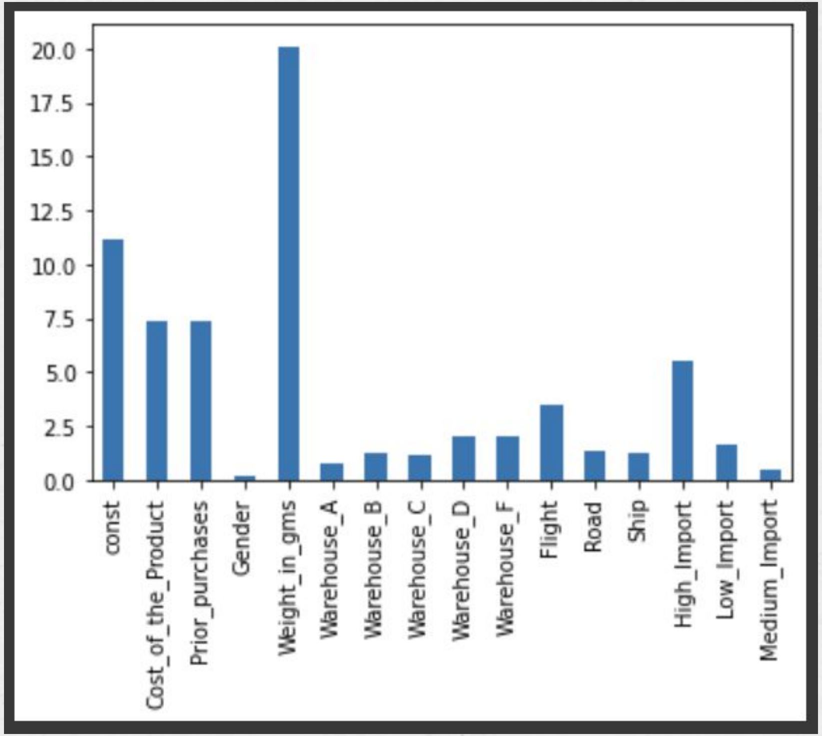
Cost Function

	Value	Significance
TP	0	0
FP	-$44,90	Cost of discounts offered
TN	-$1.51	Cost of changing the carrier
FN	-$1.51	Cost of changing the carrier

False positive is costly - so we would like to reduce false positive

LOGISTIC REGRESSION
TRAINING THE MODEL

LOGISTIC FIT SUMMARY

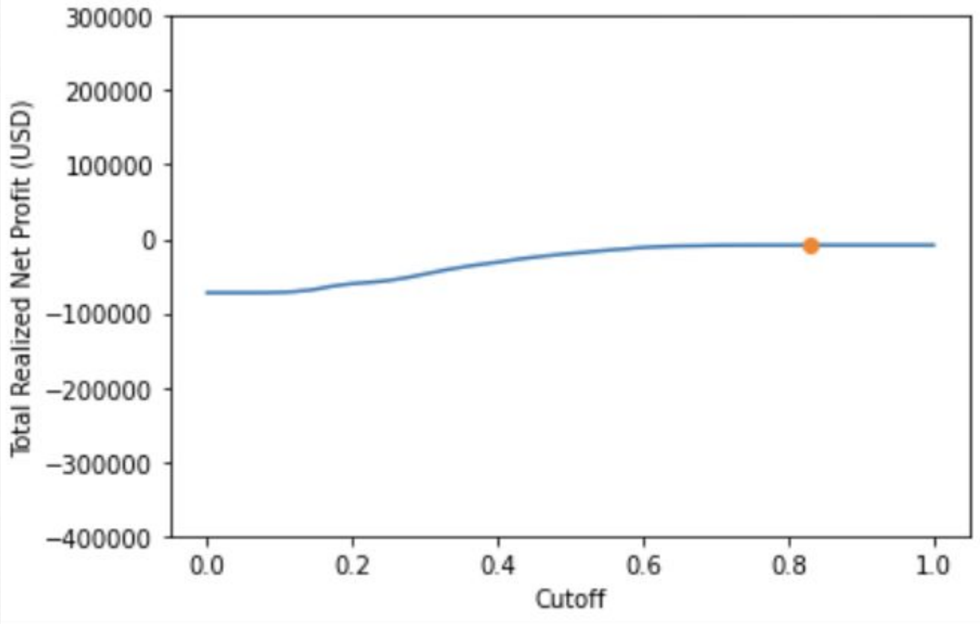


We find that some of these variables are more predictive than the others.

Based on this result, we can remove features that do not provide much predictive power and re-run our model

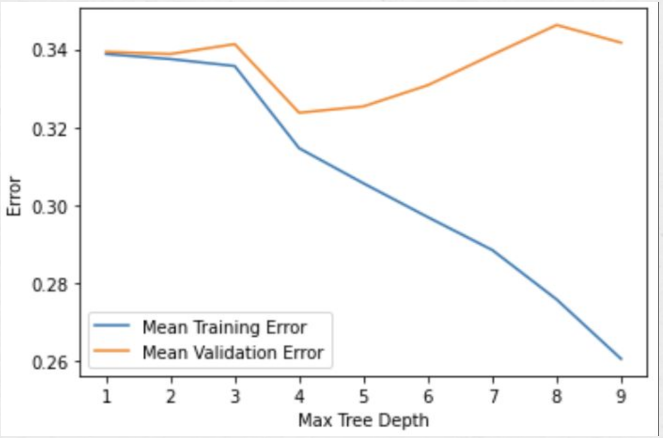
In equation form, this is

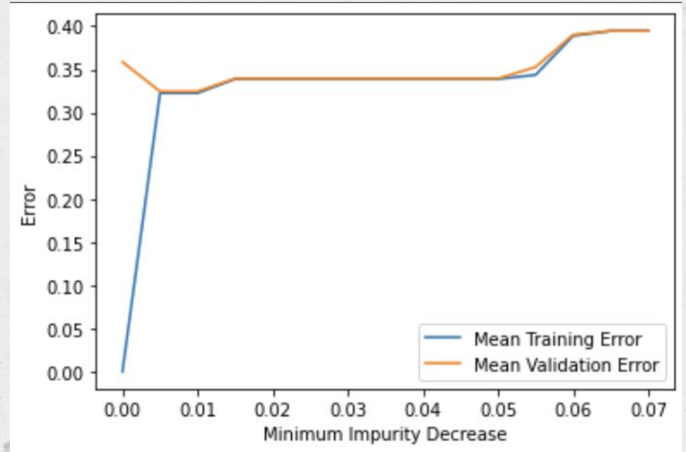
$$\begin{aligned}\text{Logit equation}=(-0.3751)+\\ &0.2247\times \text{Cost\_of\_the\_Product}\\ &0.2190\times \text{Prior\_purchases}\\ &0.6357\times \text{Weight\_in\_gms}\\ &(-0.0957)\times \text{Warehouse\_D}\\ &(-0.0670)\times \text{Warehouse\_F}\\ &(-0.4646)\times \text{High\_Import}\\ &(-0.1458)\times \text{Flight}\end{aligned}$$

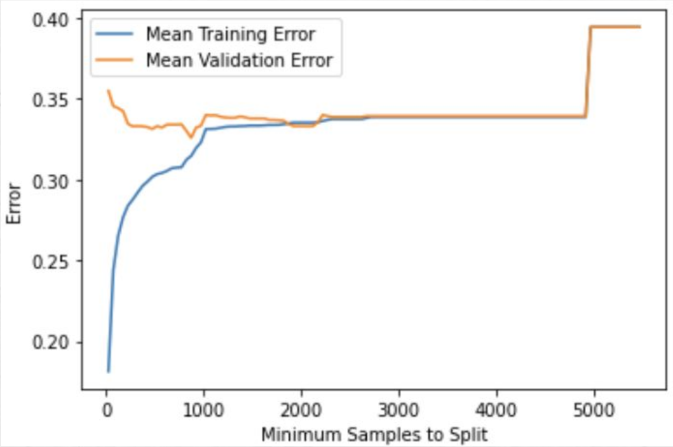


0.83 appears to be the best cutoff value

DECISION TREE
TRAINING THE MODEL

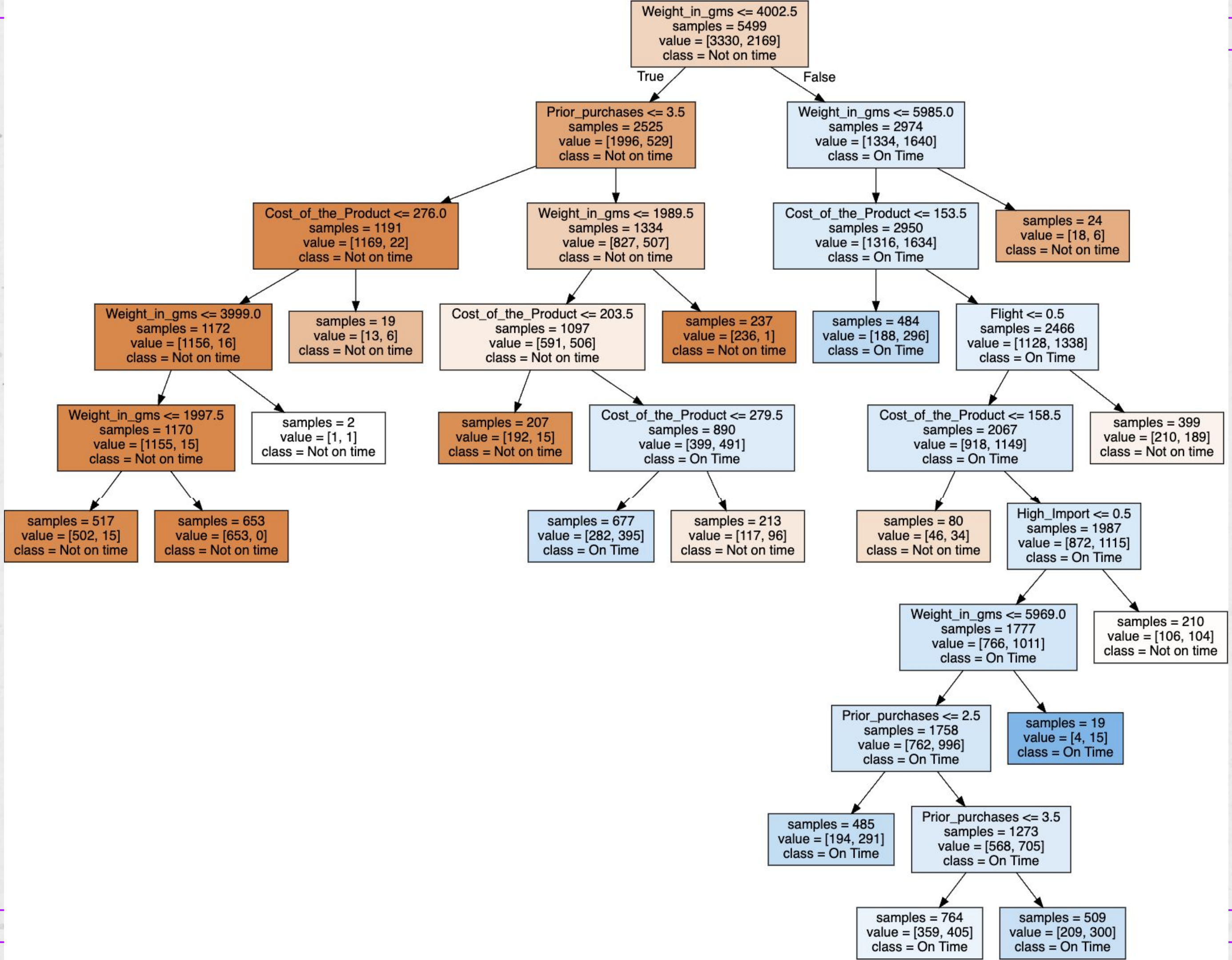




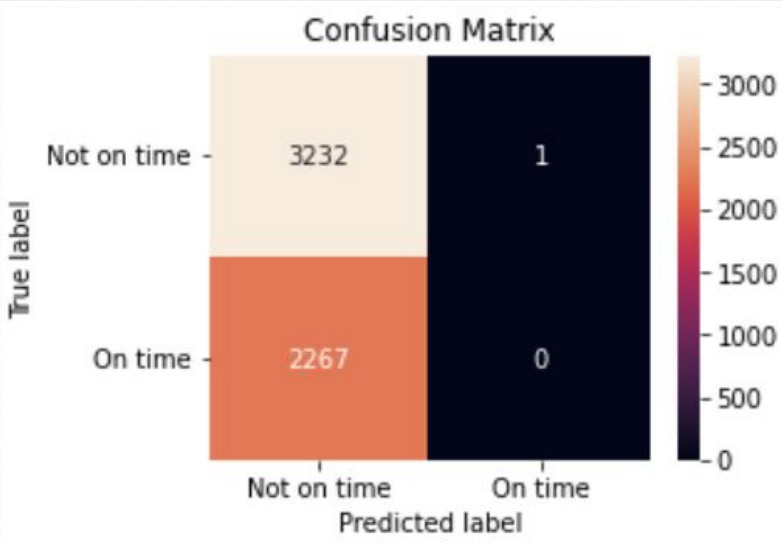


We created three decision trees of different depth, each based on the best values of **depth(4)**, **sample split (875)**, and **impurity(0.005)** found through **cross-validation**

Ultimately, we found that the decision tree with a sample split of 875 gives the best decision tree, as in, the lowest false positive rate.


Weight_in_gms <= 4002.5
samples = 5499
value = [3330, 2169]
class = Not on time
True
False
Prior_purchases <= 3.5
samples = 2525
value = [1996, 529]
class = Not on time
Weight_in_gms <= 5985.0
samples = 2974
value = [1334, 1640]
class = On Time
Cost_of_the_Product <= 276.0
samples = 1191
value = [1169, 22]
class = Not on time
Weight_in_gms <= 1989.5
samples = 1334
value = [827, 507]
class = Not on time
Cost_of_the_Product <= 153.5
samples = 2950
value = [1316, 1634]
class = On Time
samples = 24
value = [18, 6]
class = Not on time
Weight_in_gms <= 3999.0
samples = 1172
value = [1156, 16]
class = Not on time
samples = 19
value = [13, 6]
class = Not on time
Cost_of_the_Product <= 203.5
samples = 1097
value = [591, 506]
class = Not on time
samples = 237
value = [236, 1]
class = Not on time
samples = 484
value = [188, 296]
class = On Time
Flight <= 0.5
samples = 2466
value = [1128, 1338]
class = On Time
Weight_in_gms <= 1997.5
samples = 1170
value = [1155, 15]
class = Not on time
samples = 2
value = [1, 1]
class = Not on time
samples = 207
value = [192, 15]
class = Not on time
Cost_of_the_Product <= 279.5
samples = 890
value = [399, 491]
class = On Time
Cost_of_the_Product <= 158.5
samples = 2067
value = [918, 1149]
class = On Time
samples = 399
value = [210, 189]
class = Not on time
samples = 517
value = [502, 15]
class = Not on time
samples = 653
value = [653, 0]
class = Not on time
samples = 677
value = [282, 395]
class = On Time
samples = 213
value = [117, 96]
class = Not on time
samples = 80
value = [46, 34]
class = Not on time
High_Import <= 0.5
samples = 1987
value = [872, 1115]
class = On Time
Weight_in_gms <= 5969.0
samples = 1777
value = [766, 1011]
class = On Time
samples = 210
value = [106, 104]
class = Not on time
Prior_purchases <= 2.5
samples = 1758
value = [762, 996]
class = On Time
samples = 19
value = [4, 15]
class = On Time
samples = 485
value = [194, 291]
class = On Time
Prior_purchases <= 3.5
samples = 1273
value = [568, 705]
class = On Time
samples = 764
value = [359, 405]
class = On Time
samples = 509
value = [209, 300]
class = On Time

EVALUATION



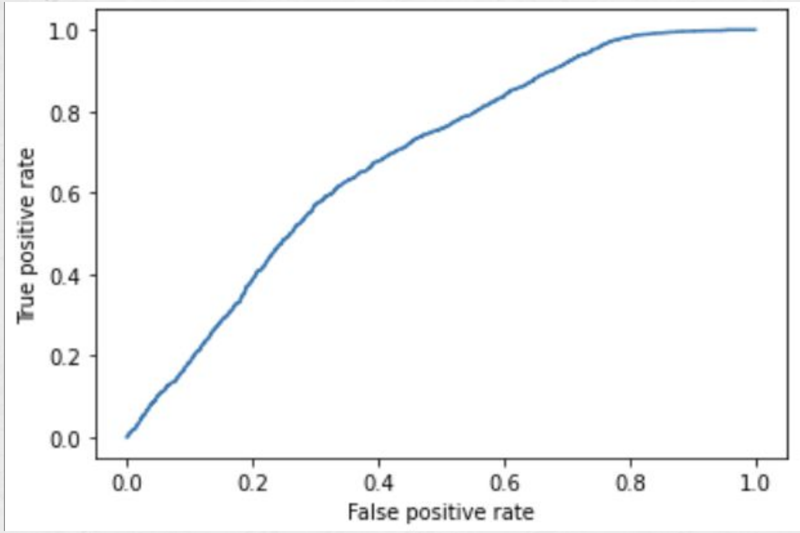
Under the cutoff of 0.83:

Specificity = 0.997

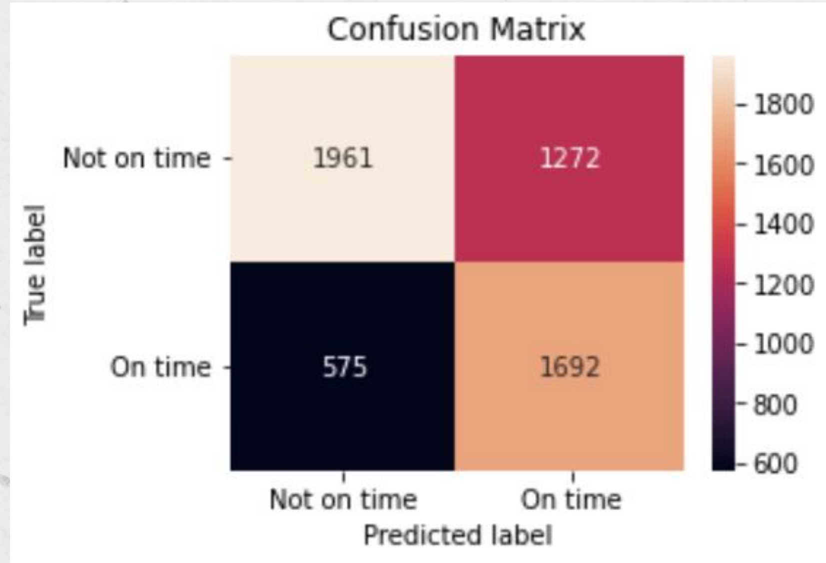
The model correctly identifies 99.7% of all products that did not reach on time

Calculating the estimated profit with the previous generated cost function, **the expected cost in refund with this model is -$8,325.9 USD, , which will save $ (294,710-8,235.9)= $286,474.1**

ROC Curve



Our classifier has a 68.3 AUC score. We will consult the company for other available data like shipping carrier, weather in the future to improve the model performance.



With the decision tree's prediction

Sensitivity: 0.7464

Specificity = 0.606

The model correctly identifies 60.6% of all products that did not reach on time

```python
# True positives are in the lower-right (row 1, column 1)
TP = cm_test[1, 1]
# True negatives are in the upper-left (row 0, column 0)
TN = cm_test[0, 0]
# False positives are in the upper-right (row 0, columns 1)
FP = cm_test[0, 1]
# False negatives are in the lower-left (row 1, column 0)
FN = cm_test[1, 0]
# Profit as computed before
profit = \
    0 * TP + \
    -1.51 * TN + \
    -1.51 * FN + \
    -44.90* FP
profit
```

Calculating the estimated profit with the previous generated cost function, **the expected cost in refund with this model is -$60,942 USD, which will save $ (294,710-60,942)= $233,768**

2022

05

LIMITATION

Low ROC score

We can consult the company for other available data like shipping carrier, weather in the future to improve the model performance.

Assumptions

We have made assumptions about shipping carriers and the shipping cost. This is a key part of our model and insights in the actual shipping cost of the company can greatly change the model.

2022

06

RECOMMENDATION

Shipment Carrier Change

Our model will predict whether or not a product will be late. If the model predicts that the item will be late, then the company should **change shipping delivery methods in order to ensure that the product arrives on time**. By making sure that the product arrives on time, the company is saving money that would otherwise be spent on giving discounts to late products.

After the Model

Clustering

R programming that if we use k-means scatter analysis, the best number of clusters to use for this data would be three.

Splitting the Data

Splits the dataset into 3 sets (Product importance: low, medium, high) for the model

- yield better predictions
- allow the company to make different business decisions based on product importance

Special Shoutout - Our unsupervised learning model created in R
https://drive.google.com/file/d/1fw9Q9AXnBrYXa6rk6ulrBsZdQrUuHiAO/view?usp=sharing

Thank You!