CAPSTONE PROJECT

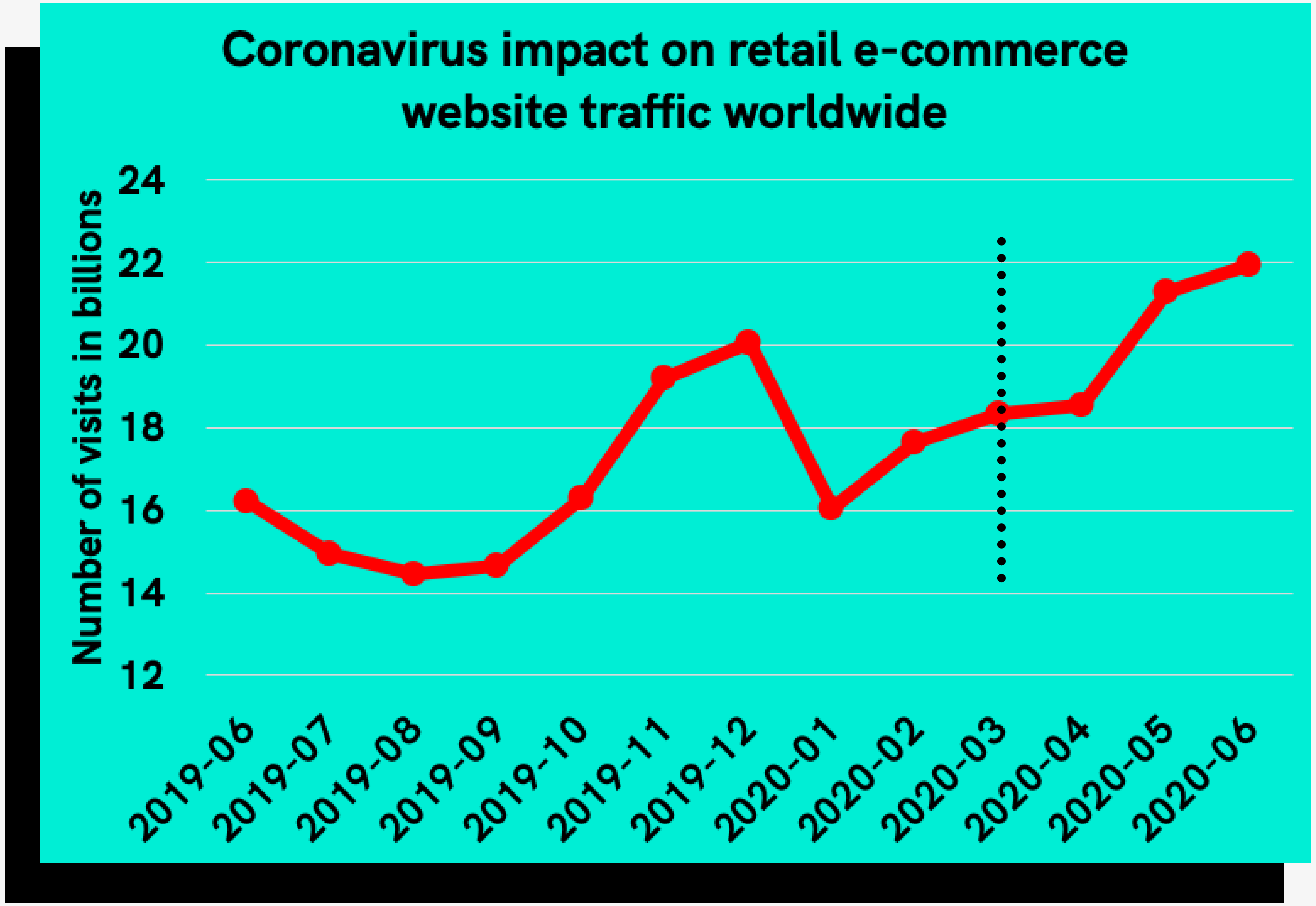
COUNTERFEIT DETECTION ON E-COMMERCE PLATFORMS

Data Science Bootcamp
Analytiks Inc.
October 3, 2020

ONLINE SHOPPING

E-commerce

n. buying and selling of goods or services using the internet



New Normal

By 2040, an estimated 95% of purchases will be made online.

Advantages of Online Shopping

Convenience

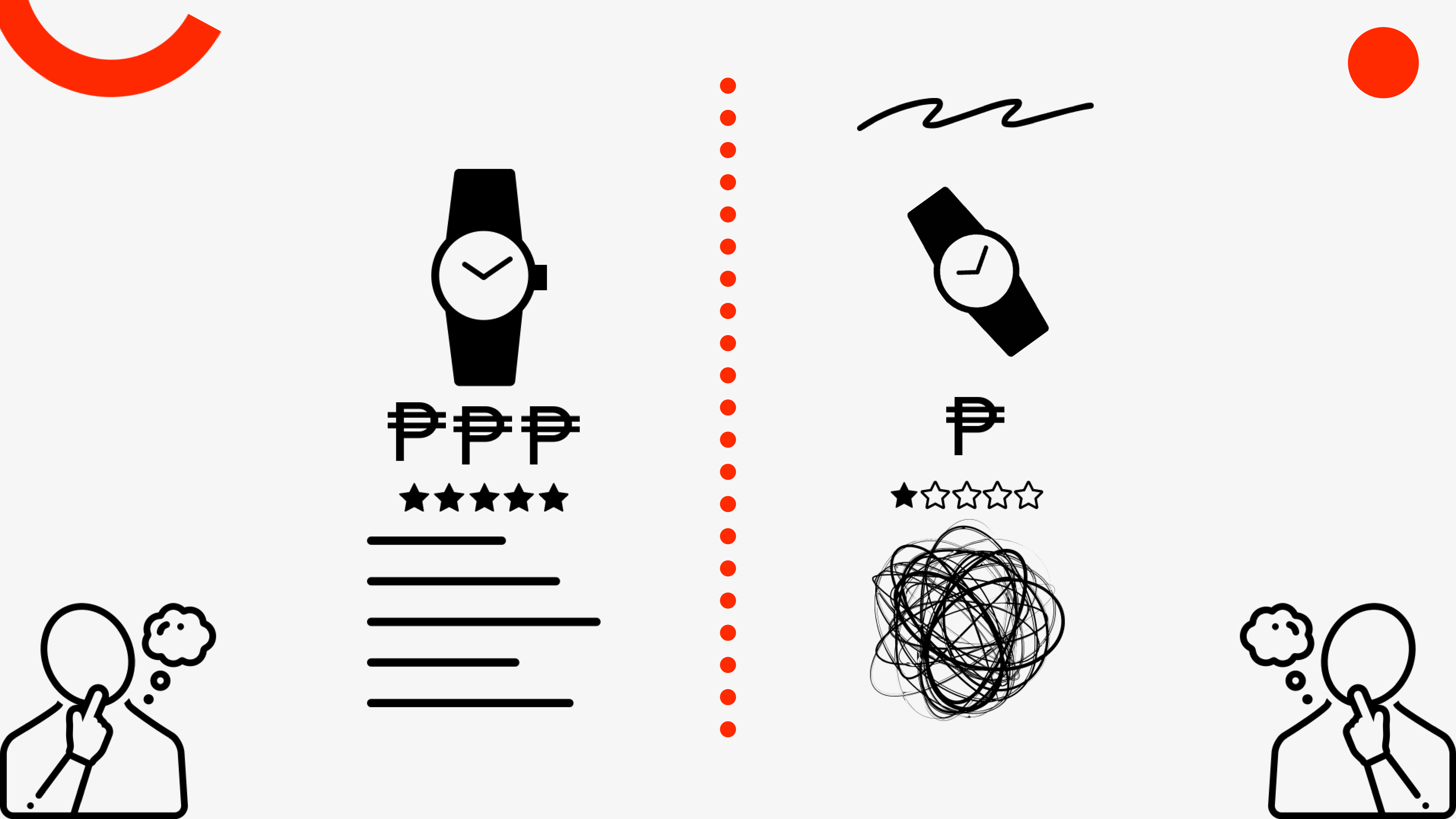
No need to go to stores

Reviews

Testimonials help in decision-making

Diversity

Providers from all over the world



COUNTERFEITING

- fraudulent imitation of a product

Amount of total counterfeiting globally is bound to reach

$ 1.82 Trillion

by the year 2020.

Key Concerns of Counterfeiting:

- Introduce dangerous products into the market
- Weaken environmental, health, and safety regulations
- Diminish tax revenues
- Support organized crime
- Promote child labor

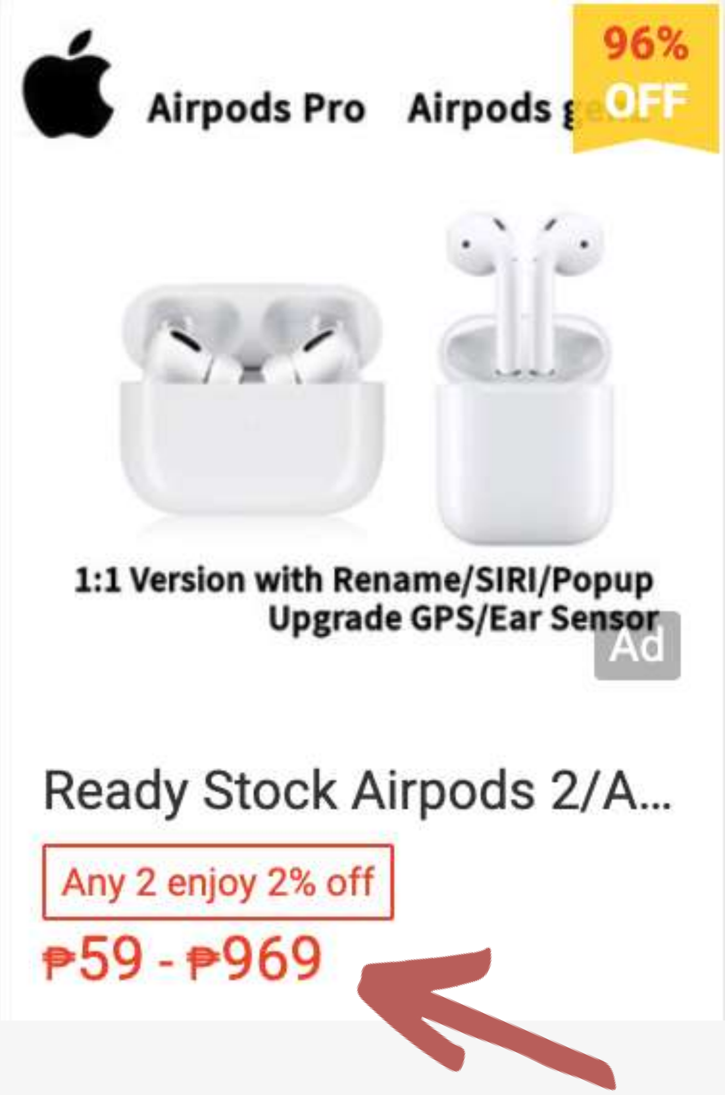
PROBLEM

How can we detect counterfeit products on e-commerce platforms?

Traditional ways to spot counterfeits

Price

Way below SRP



Airpods Pro Airpods g...

96% OFF

1:1 Version with Rename/SIRI/Popup Upgrade GPS/Ear Sensor

Ad

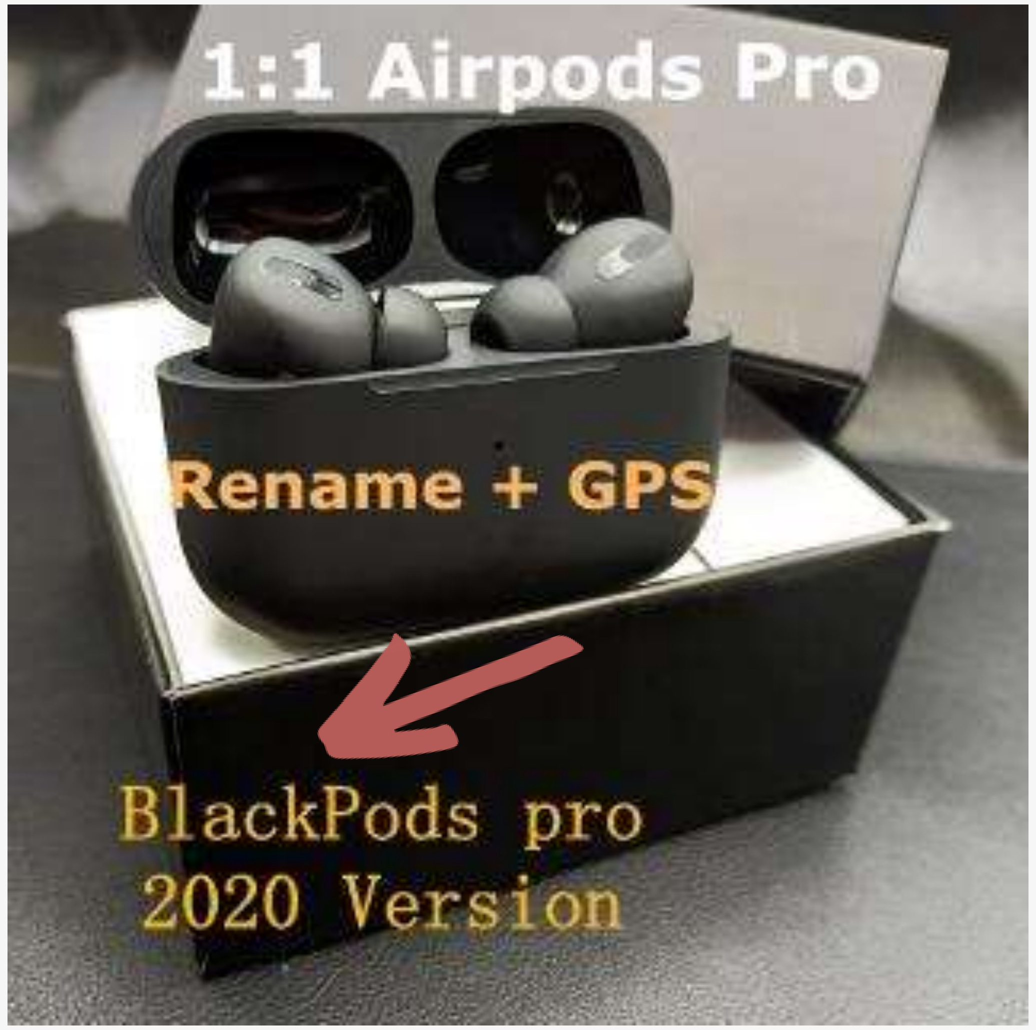
Ready Stock Airpods 2/A...

Any 2 enjoy 2% off

₱59 - ₱969

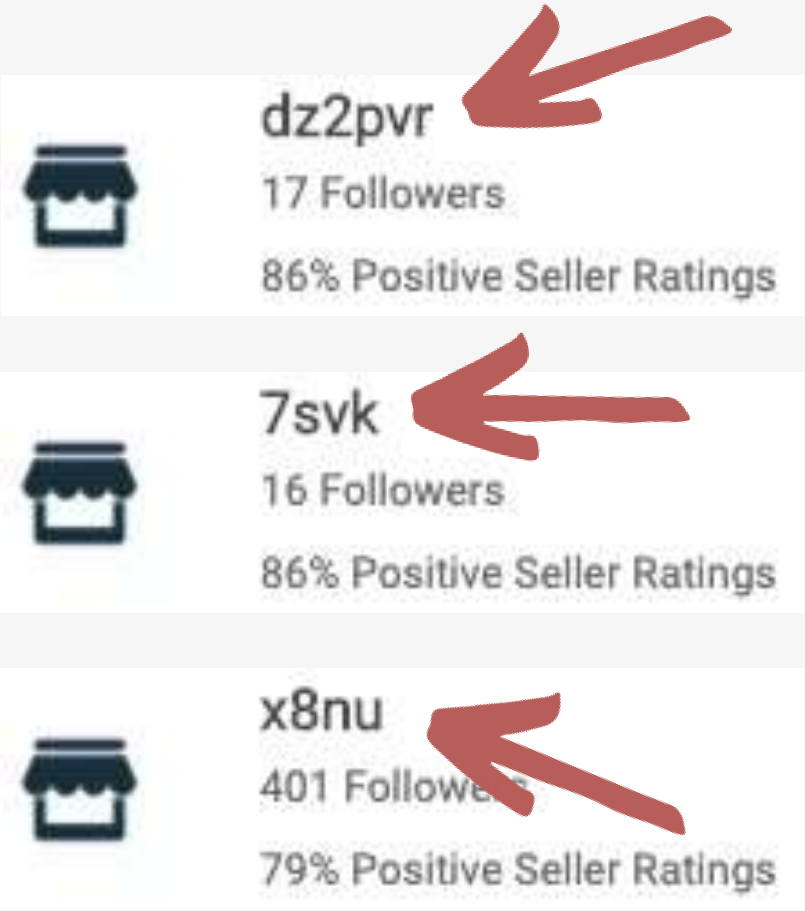
Packaging

Suspicious branding



Place

Sketchy retailers



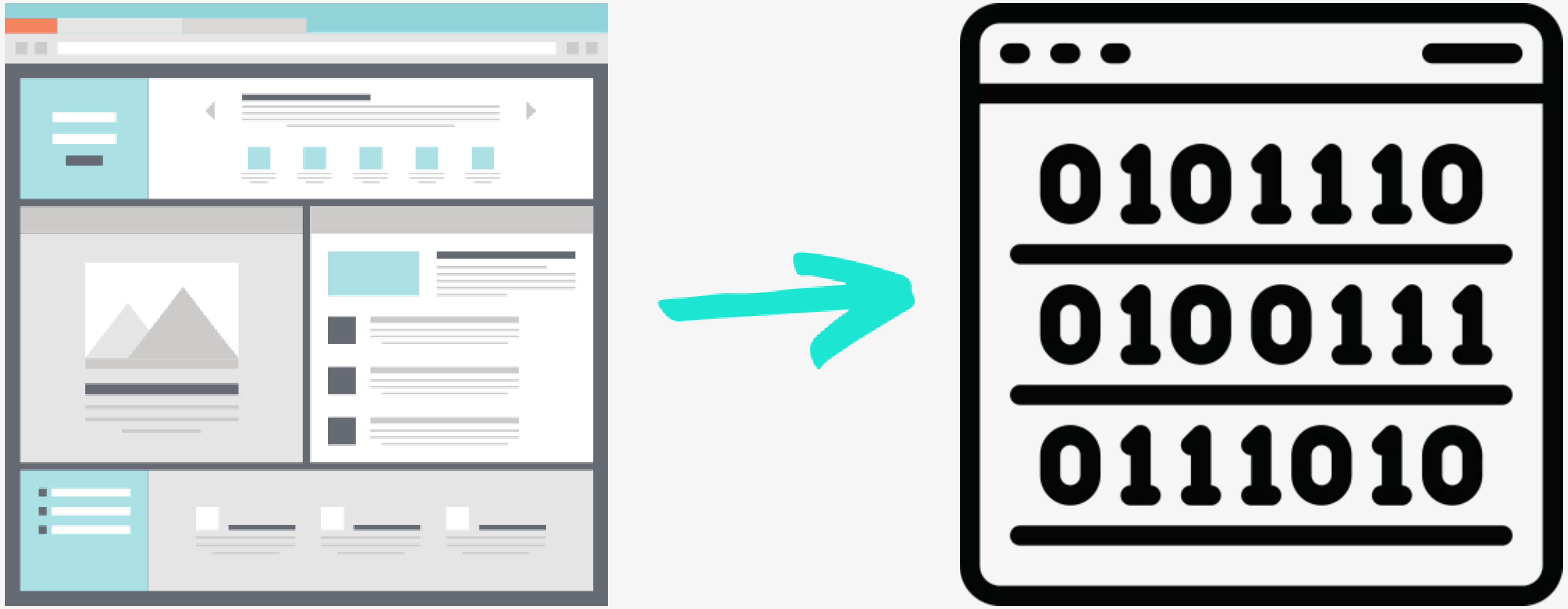
dz2pvr
17 Followers
86% Positive Seller Ratings

7svk
16 Followers
86% Positive Seller Ratings

x8nu
401 Followers
79% Positive Seller Ratings

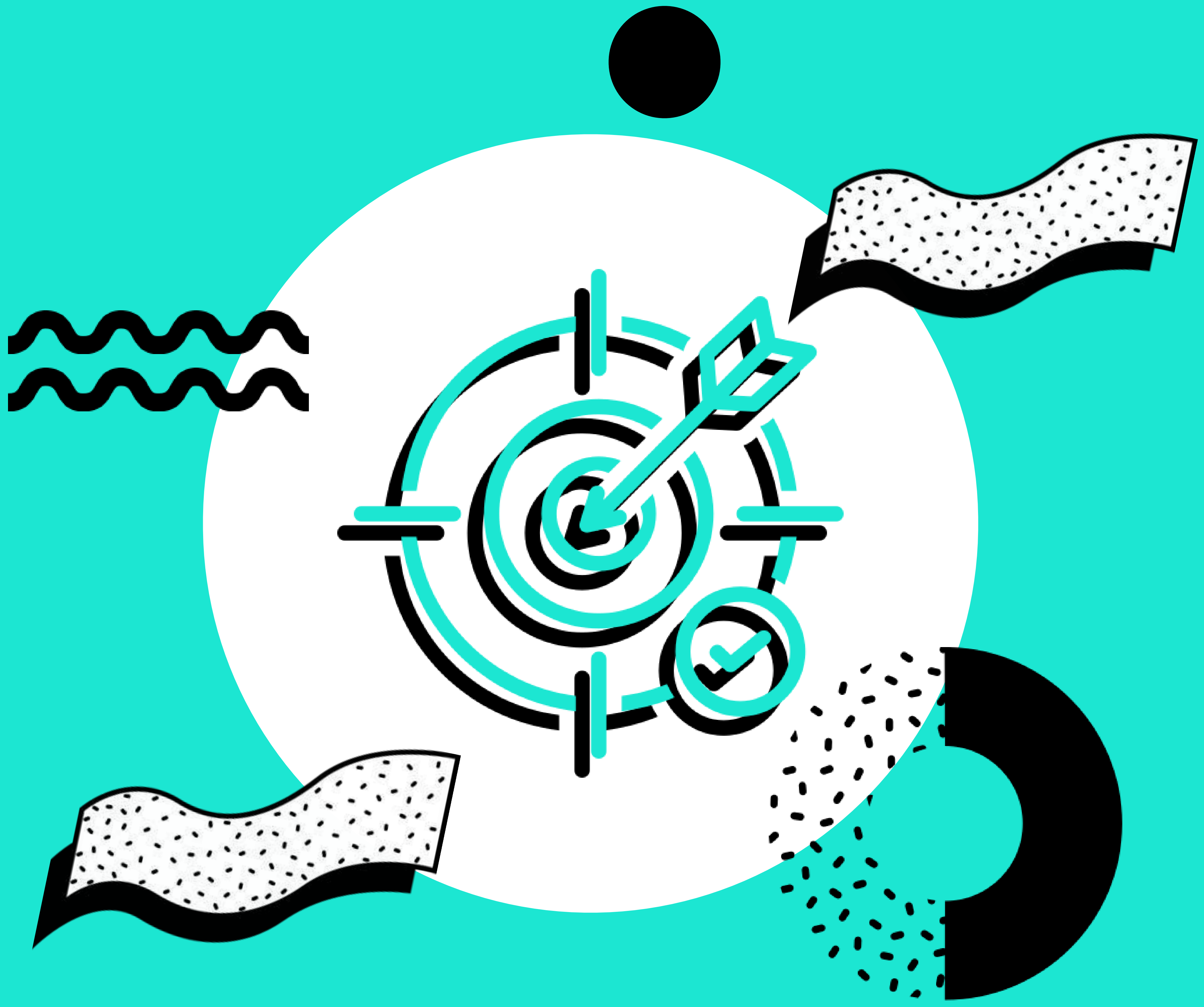
Use machine learning!

Price, Packaging, and Place are features that can be used to classify products using a predictive model.

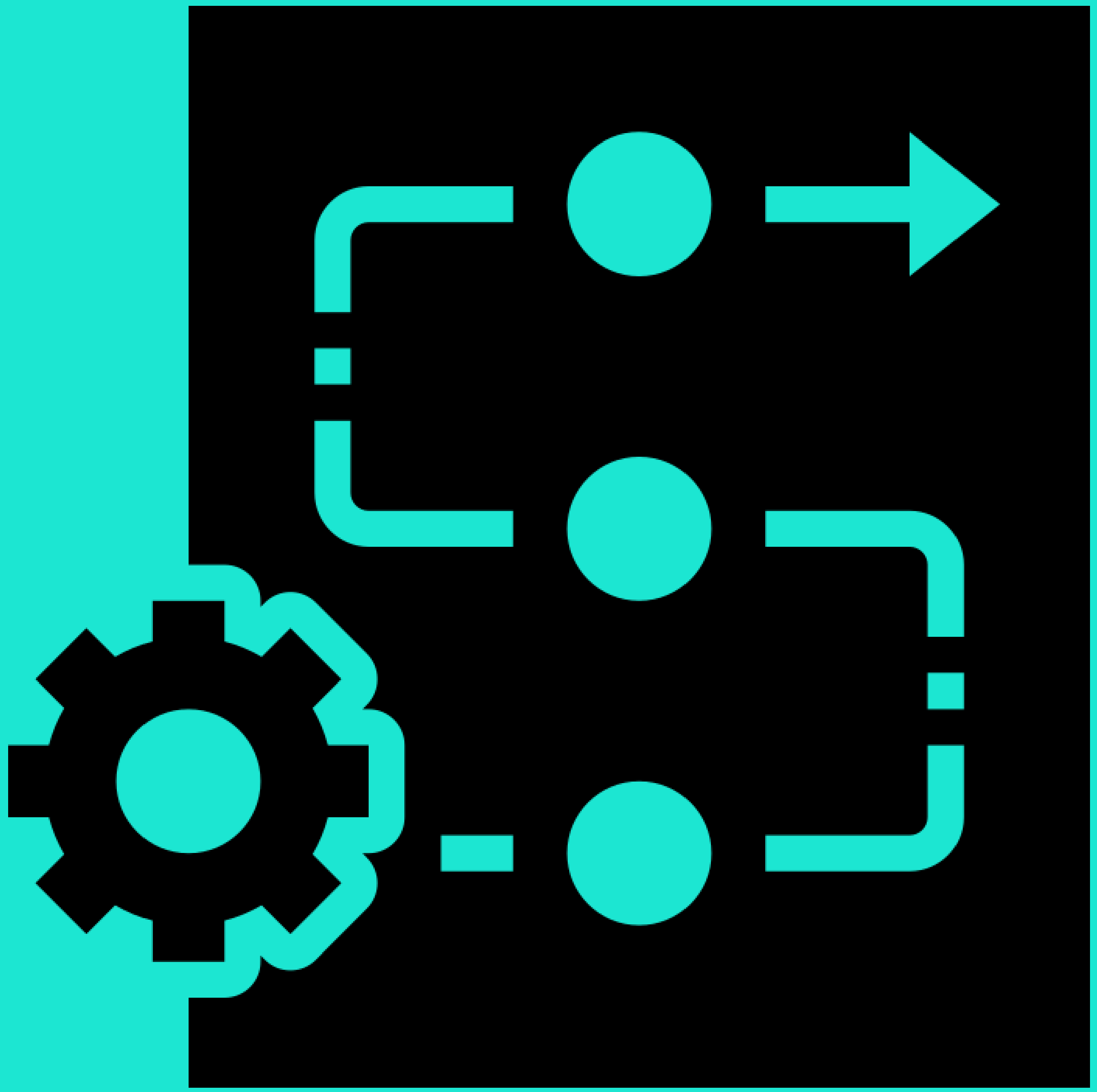


OBJECTIVES

1. Use machine learning to **predict** whether or not an e-commerce listing is authentic.
2. **Determine** how the features affect the decision-making of the model.



PROJECT FLOW



Step 1
Assessment

Step 2
Collect data

Step 3
Prepare data

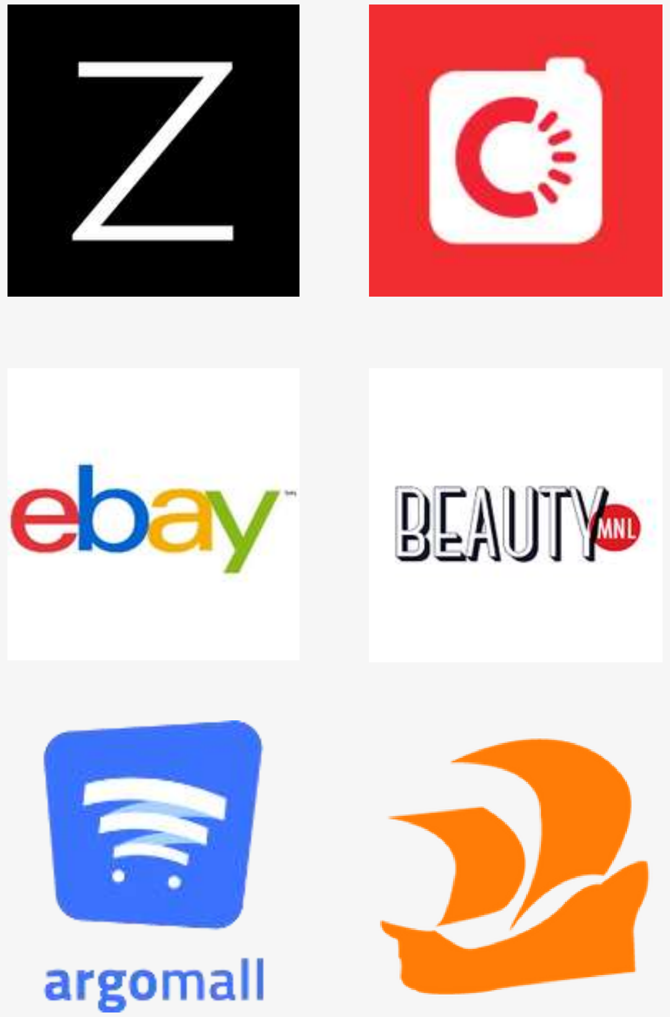
Step 4
Build model

Step 5
Evaluate model

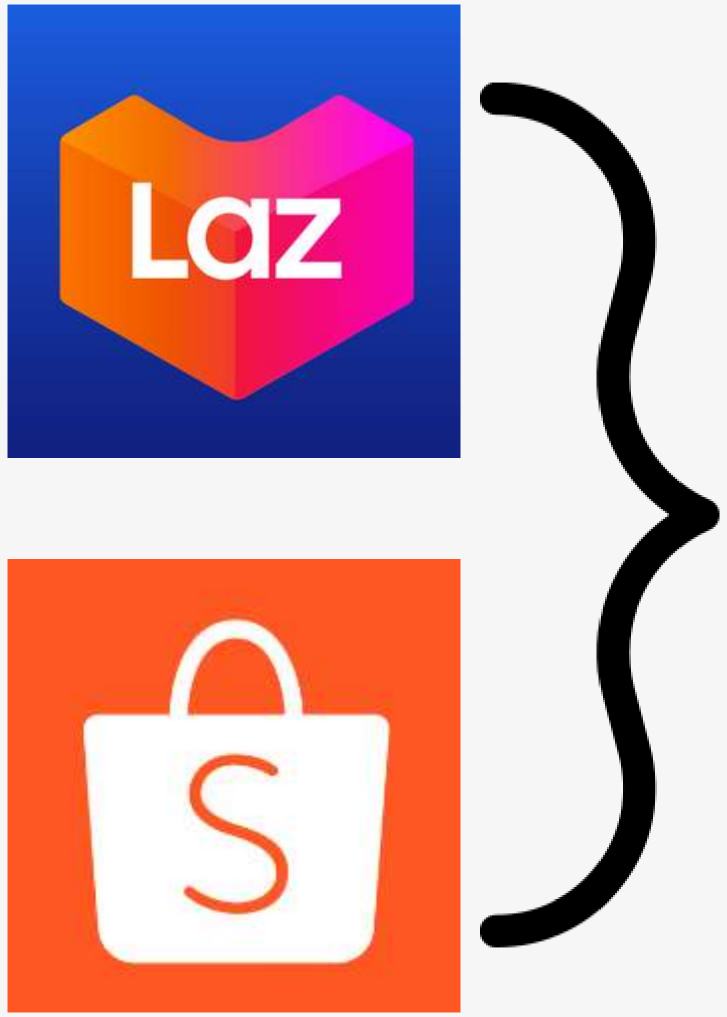
Step 1
Assessment

Evaluate options

1. Platforms



Lazada



Shopee

- Top 2 e-commerce sites based on traffic
- Diverse selection of products
- Massive user base

Step 1
Assessment

Evaluate options

2. Products

- In the Philippines, the e-market's largest segment is Electronics & Media.
- In 2016, electronics accounted for 35% in the global trade of fake goods.

airpods pro

apple watch

ps4 controller

Step 2
Collect data

How to collect?

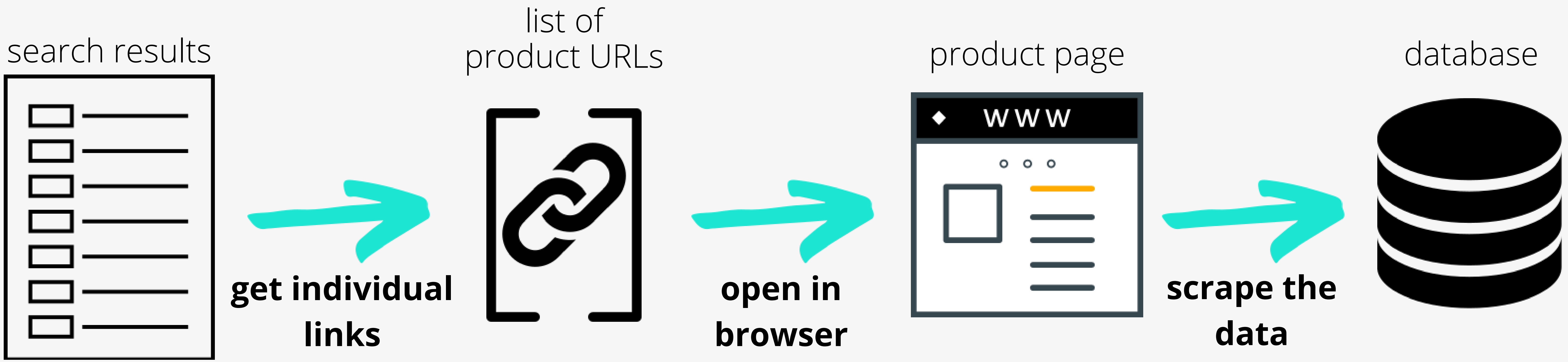
Dynamic web scraping

Use Selenium

- Automated
- Simulate “human-like” behavior

Page-by-page collection

- Lazada: 40 listings/page
- Shopee: 50 listings/page



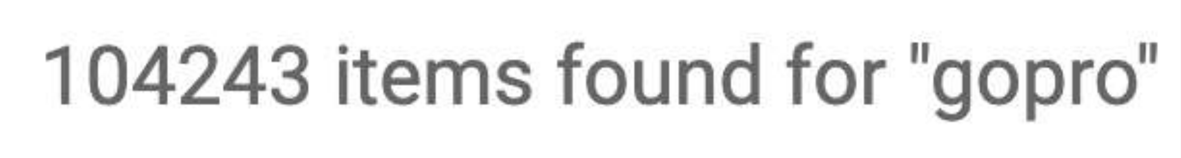
Step 2
Collect data

How much data?

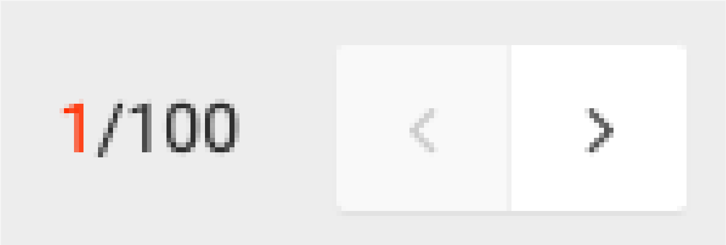
All of it, of course!

Challenge #1

- Unrealistic! Customer cannot possibly view all the listings.



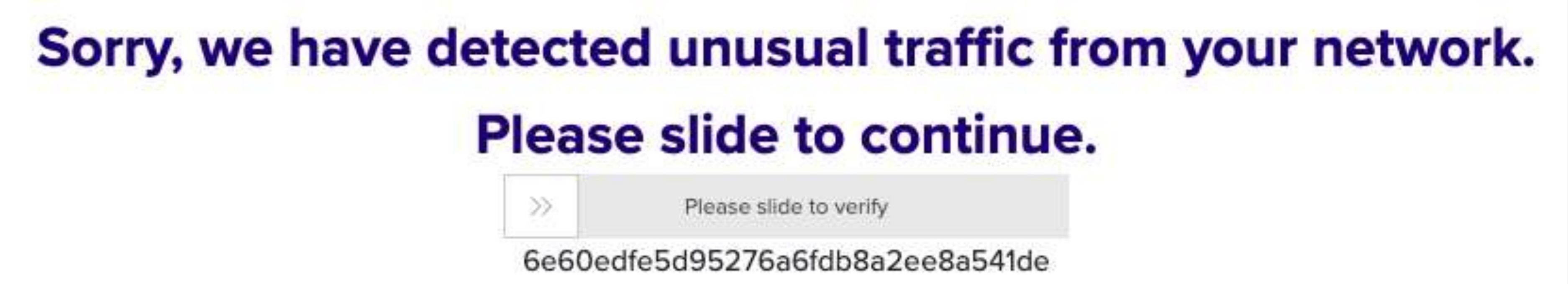
104243 items found for "gopro"



1/100

Challenge #2

- Need to limit request rate.



Sorry, we have detected unusual traffic from your network.

Please slide to continue.

Please slide to verify

6e60edfe5d95276a6fdb8a2ee8a541de

Step 2
Collect data

How much data?

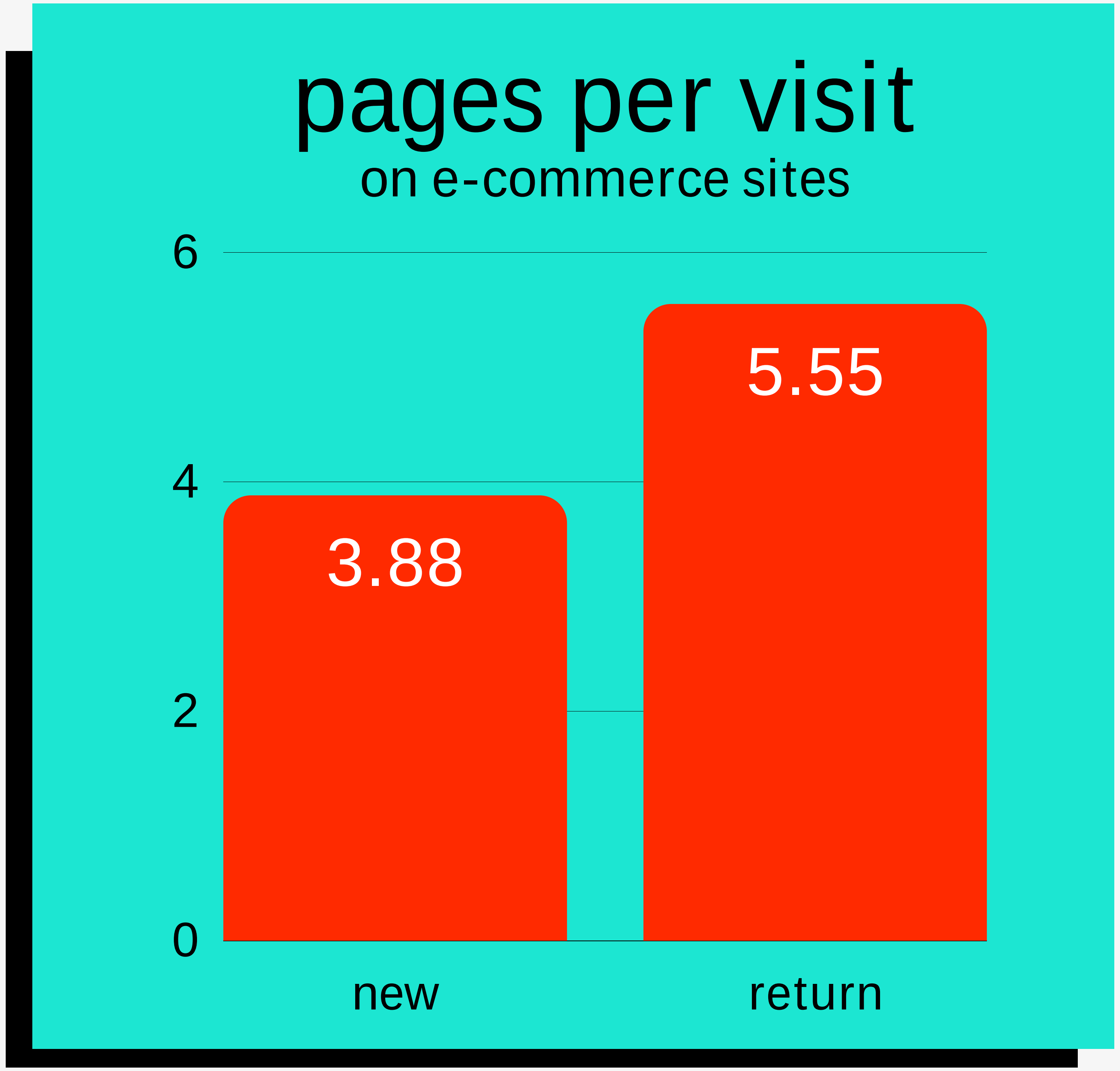
Simulate real-life scenario

Page-by-page collection

- Lazada: 40 listings/page
- Shopee: 50 listings/page

Target:

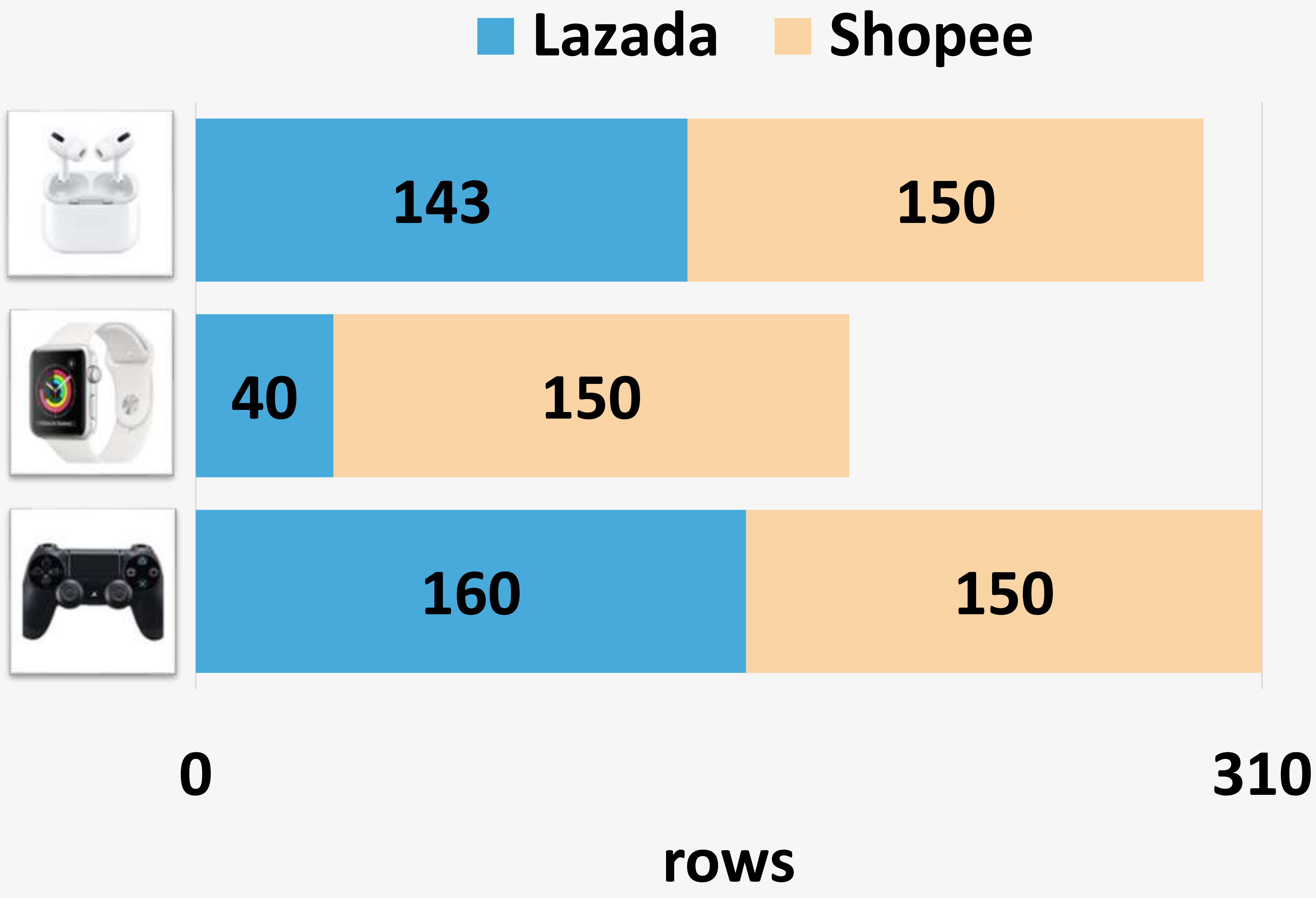
- Lazada: 4 pages = 160 listings
- Shopee: 3 pages = 150 listings
- **Total: 310 per product**



Step 2
Collect data

We now have data!

We collected six datasets.



Why are there missing data?

- Listing is displayed but product page is unavailable.
- Not enough listings to display.

Step 2
Collect data

We now have data!

We collected twelve features.

Categorical (4)

'ProductURL'
'ProductTitle'
'SellerName'
'ChatResponseTime'

Numerical (7)

'IsMall'
'Price'
'ProductRatingAverage'
'ProductRatingCount'
'SellerRating'
'ShipOnTime'
'ChatResponseRate'

Target

'Class'

Step 2
Collect data

Target variable

'Class', dtype: int

We manually labeled each row of data, as follows:

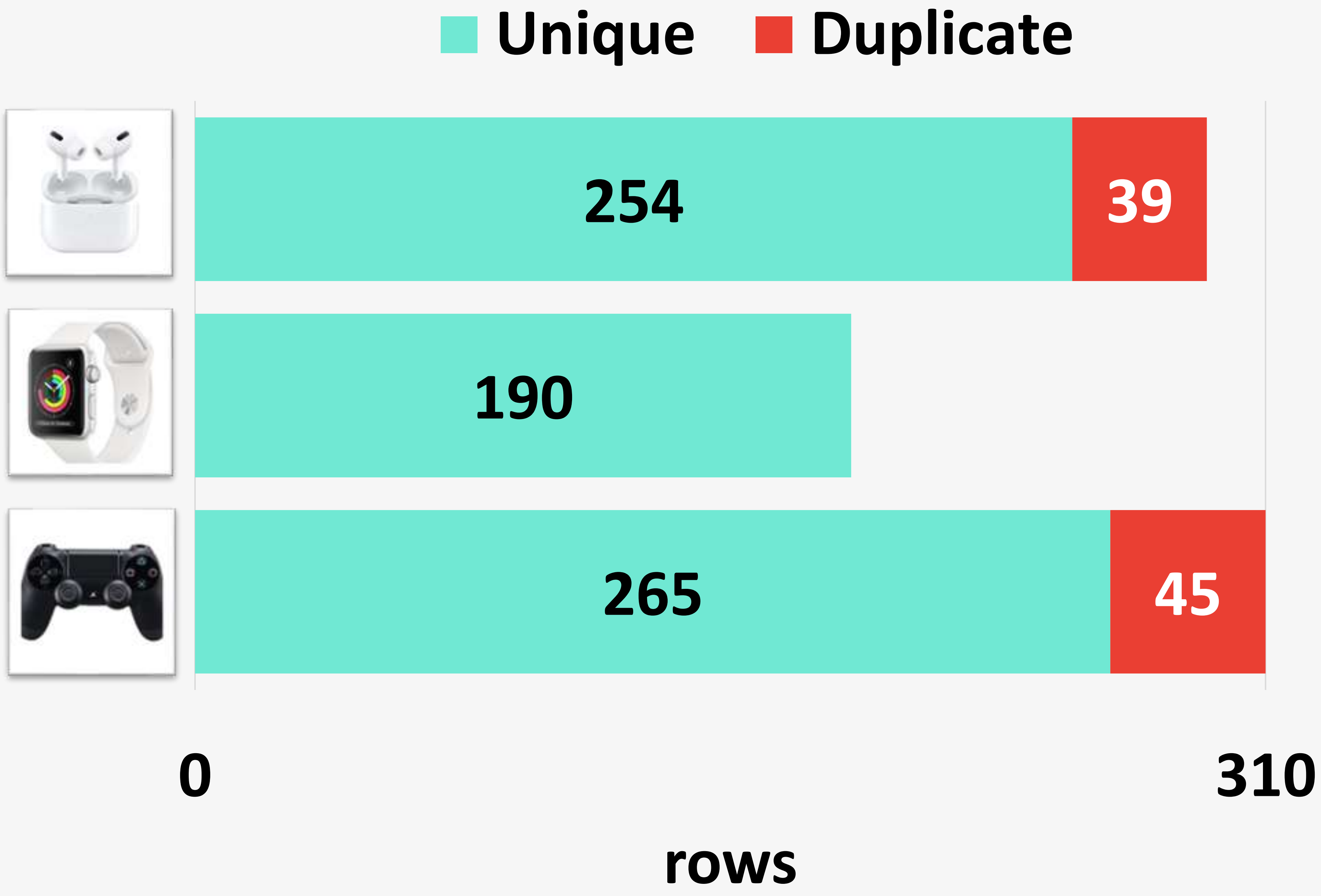
- 0 - Counterfeit
- 1 - Authentic
- 2 - Unsure
- 3 - Different Product

Each entry was reviewed by another member at least once.

Step 3
Prepare data

Data Preparation

Drop duplicates



Step 3
Prepare data

Data Preparation

Handle missing values

ChatResponseTime : unavailable on Lazada product pages; difficult to scale

within minutes
within hours
Active in: 1 hours
Active in: 10 mins
Active in: hours

ShipOnTime : not provided by Shopee

Resolution : drop both features

(1) high missing value count
(2) arguably irrelevant

Step 3
Prepare data

Data Preparation

Handle incorrect column dtype

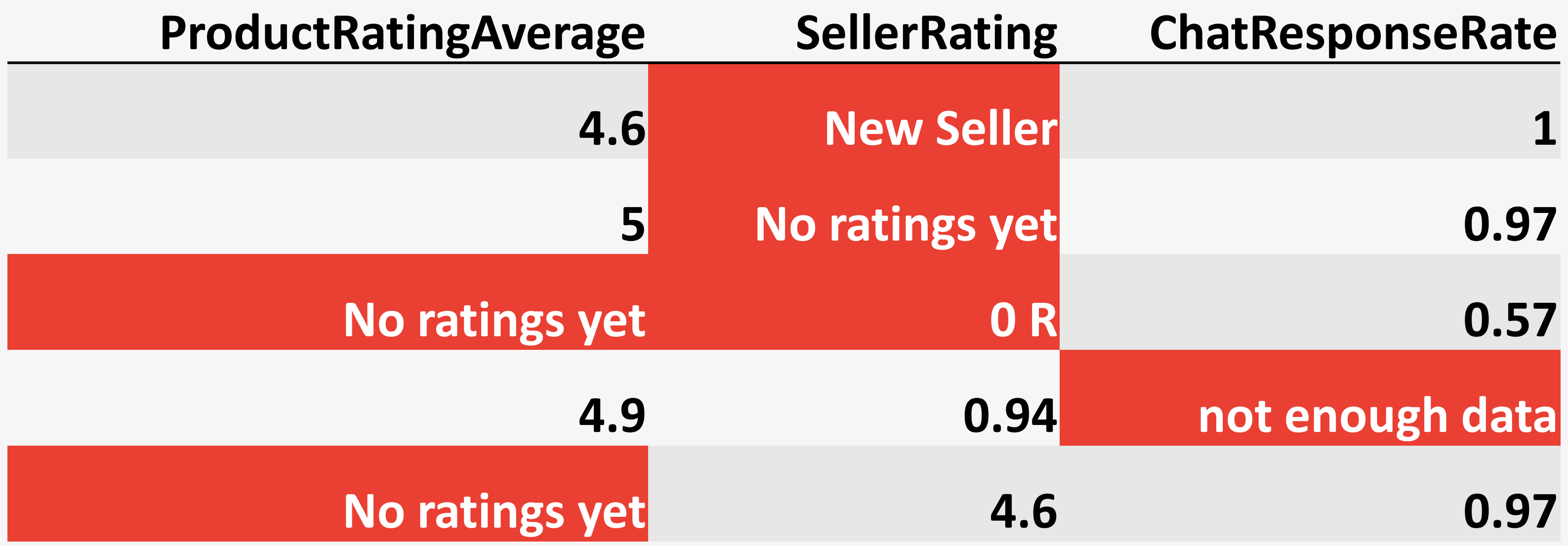
These features should be float-type, but contain object-type data.

ChatResponseRate	: ‘not enough data’
ProductRatingAverage	: ‘No ratings yet’
SellerRating	: ‘New Seller’, ‘No ratings yet’, ‘0 R’

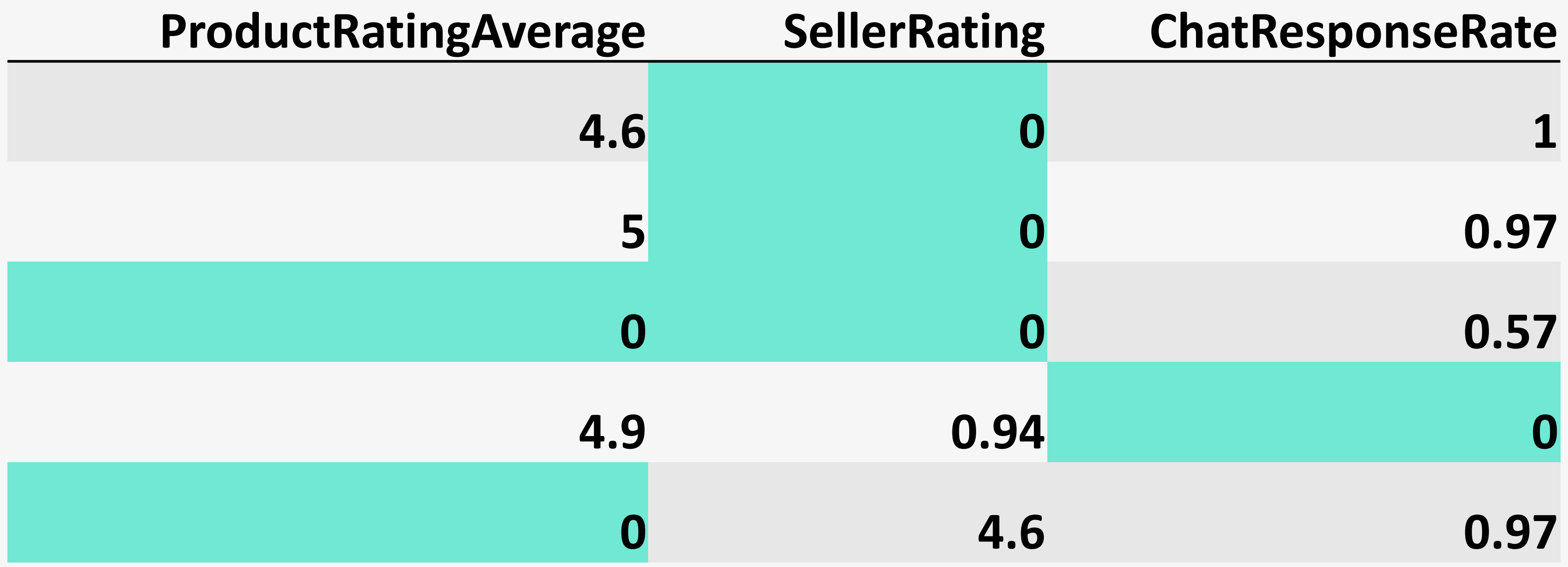
Data Preparation

Handle incorrect column dtype

Resolution : replace with 0



ProductRatingAverage	SellerRating	ChatResponseRate
4.6	New Seller	1
5	No ratings yet	0.97
No ratings yet	0 R	0.57
4.9	0.94	not enough data
No ratings yet	4.6	0.97

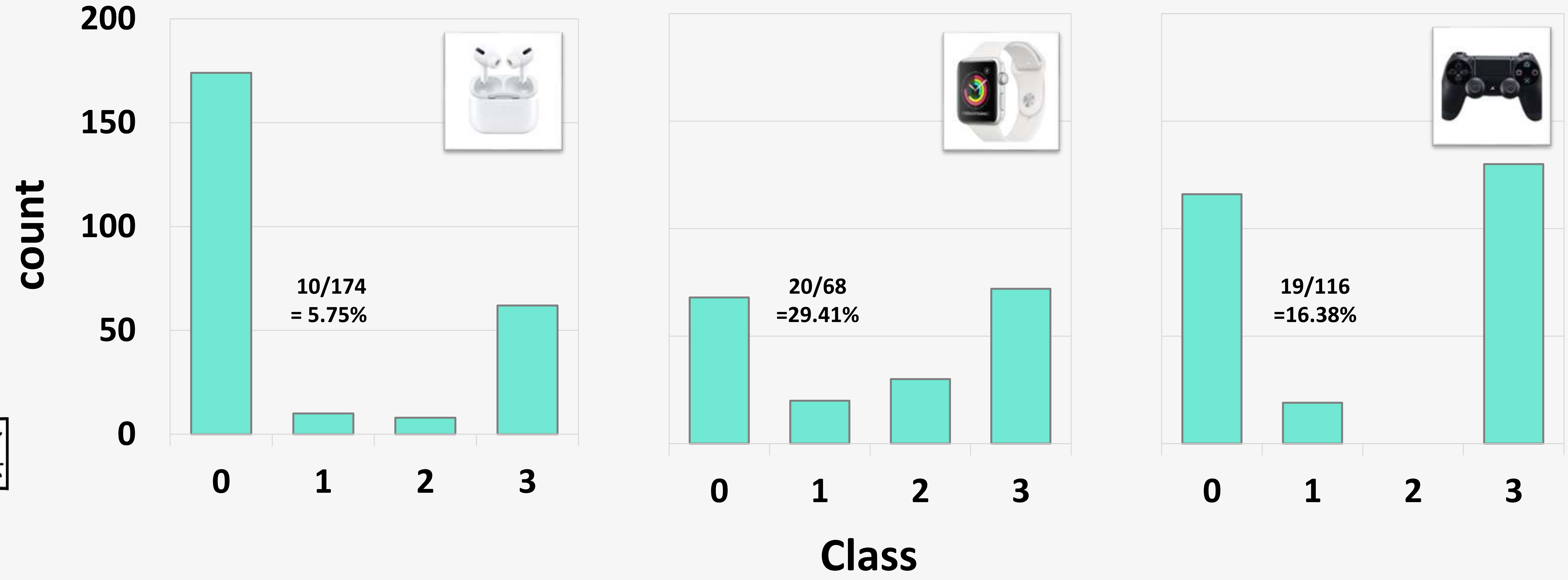


ProductRatingAverage	SellerRating	ChatResponseRate
4.6	0	1
5	0	0.97
0	0	0.57
4.9	0.94	0
0	4.6	0.97

Step 3
Prepare data

EDA

‘Class’ distribution of each product

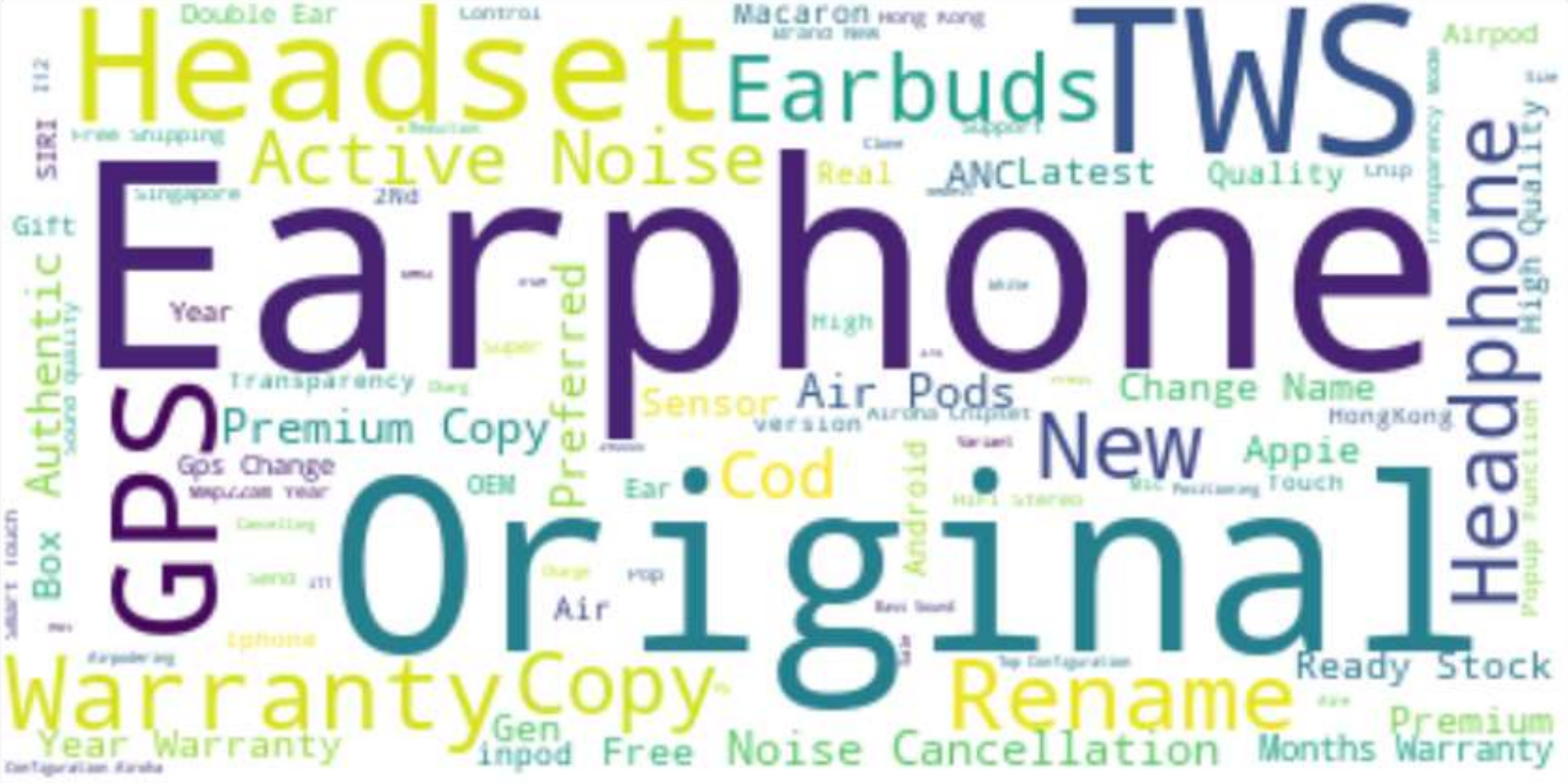


0 – Counterfeit | 1 – Authentic | 2 – Unsure | 3 - Different Product

Step 3
Prepare data

EDA

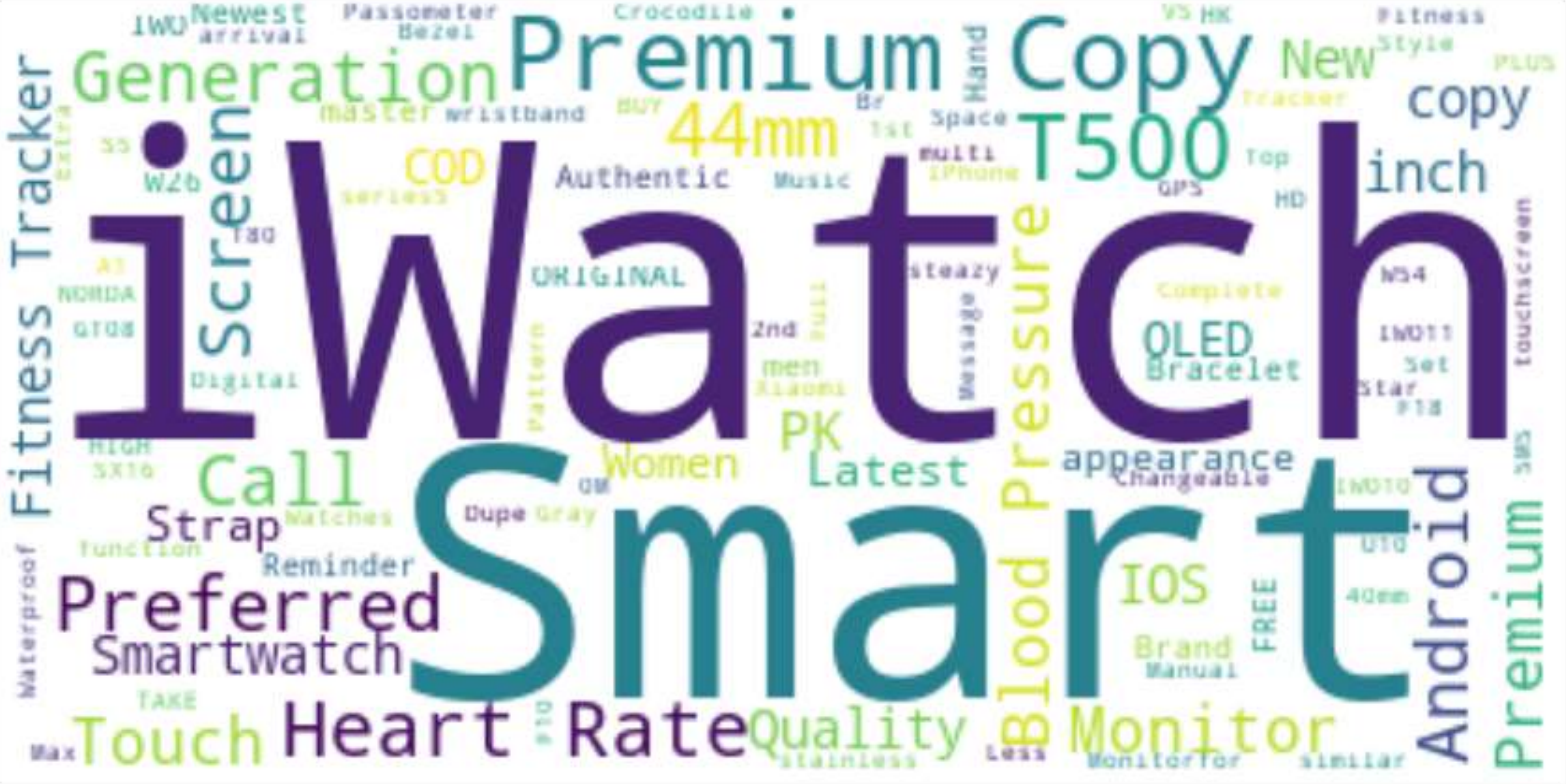
Word cloud of common terms in counterfeit product labels



Step 3
Prepare data

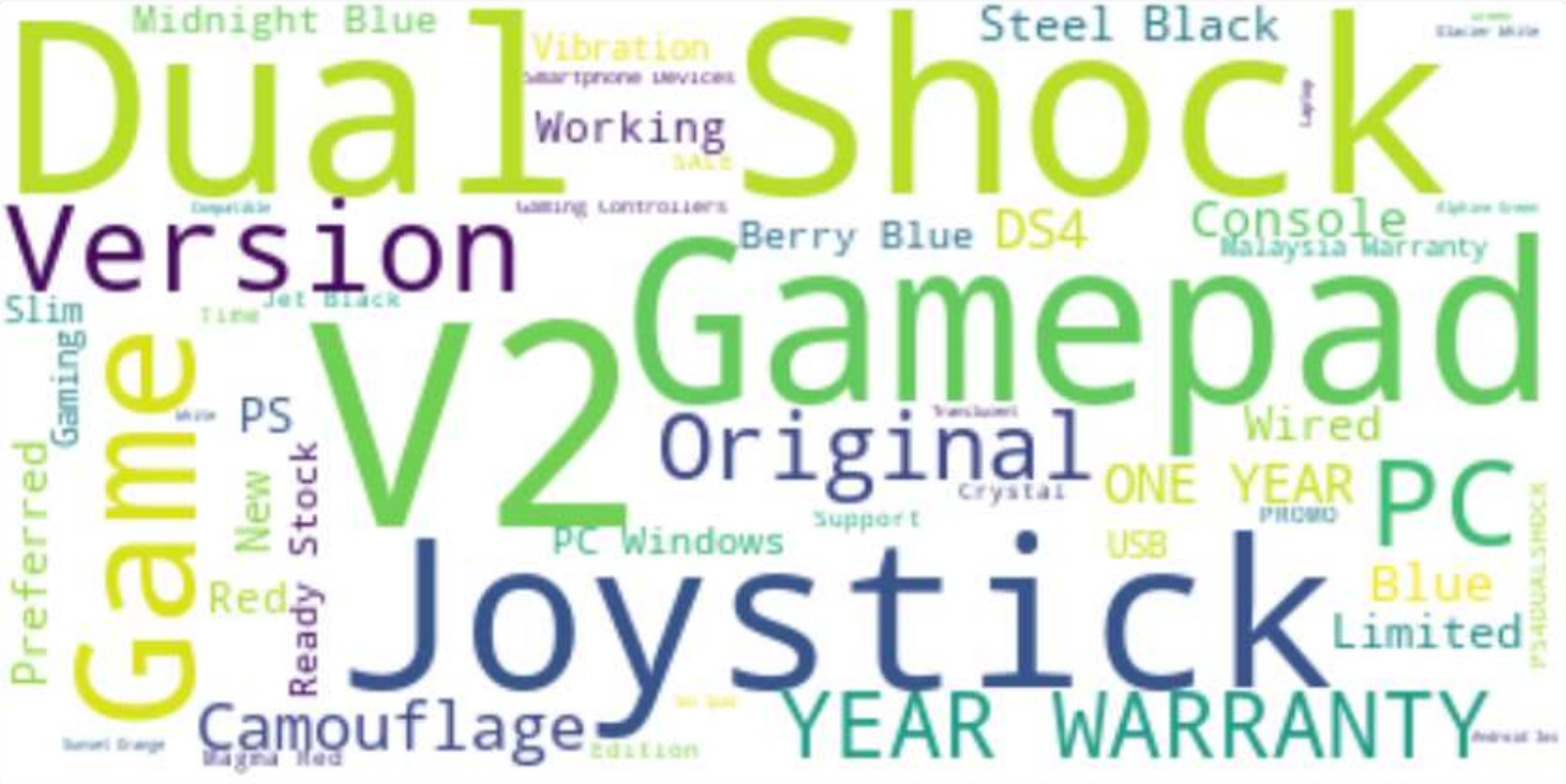
EDA

Word cloud of common terms in counterfeit product labels



EDA

Word cloud of common terms in counterfeit product labels



Step 4
Train model

Modelling

Getting Started

Data:

- 6 datasets, 6 (numerical) features, 1 target

Model:

- Supervised
- Classification
  - ➢ Voting: Hard, Soft (Uniform), Soft with GridSearchCV

➢ Logistic Regression (scaled)
➢ K-Nearest Neighbors
➢ Gaussian Naïve-Bayes (scaled)
➢ Decision Tree
➢ Random Forest

- all default except for max_depth = 10 for DT
- 5-fold cross validation

Metrics

Lazada ‘airpods pro’: best parameters

Step 5
Evaluate model

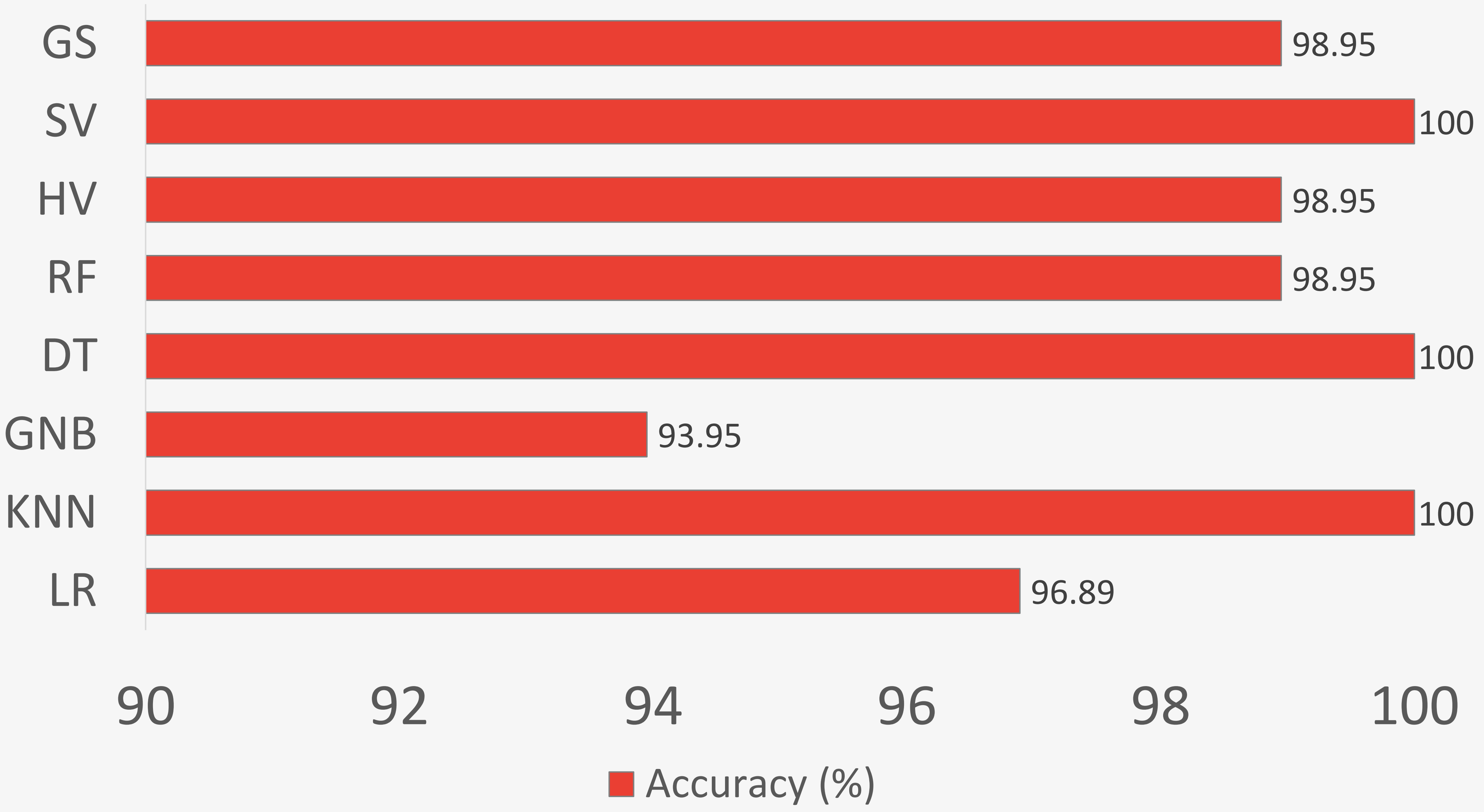
```
param_grid = {
        'knn__n_neighbors' : [1,3],
        'rf__max_depth': [1, 4, 7],
        'rf__criterion' : ['gini', 'entropy’]
        }
```

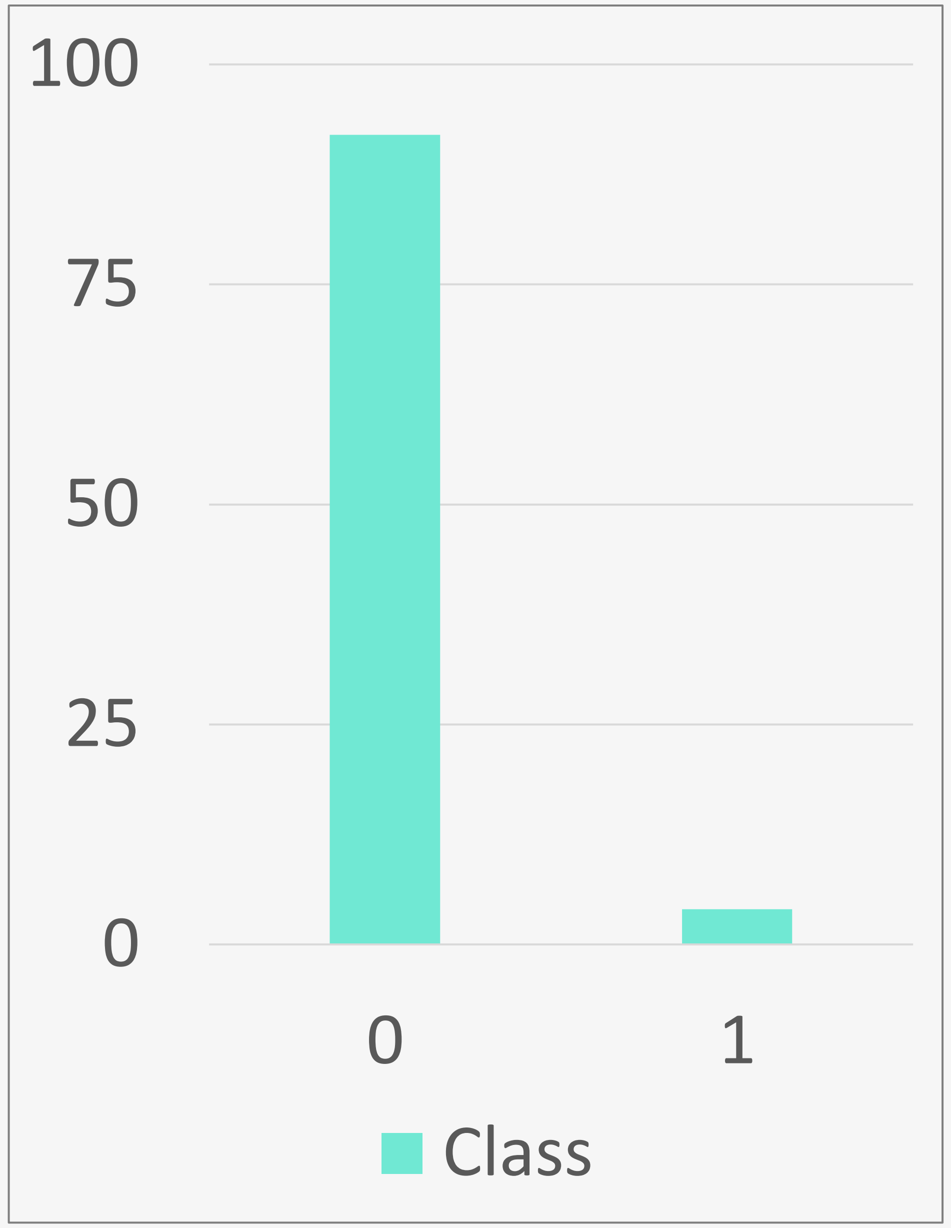
```
Best parameters = {
        'knn__n_neighbors': 1,
        'rf__max_depth’: 4
        'rf__criterion': 'gini’,
        }
```

Metrics

Lazada ‘airpods pro’: accuracy

Step 5
Evaluate model

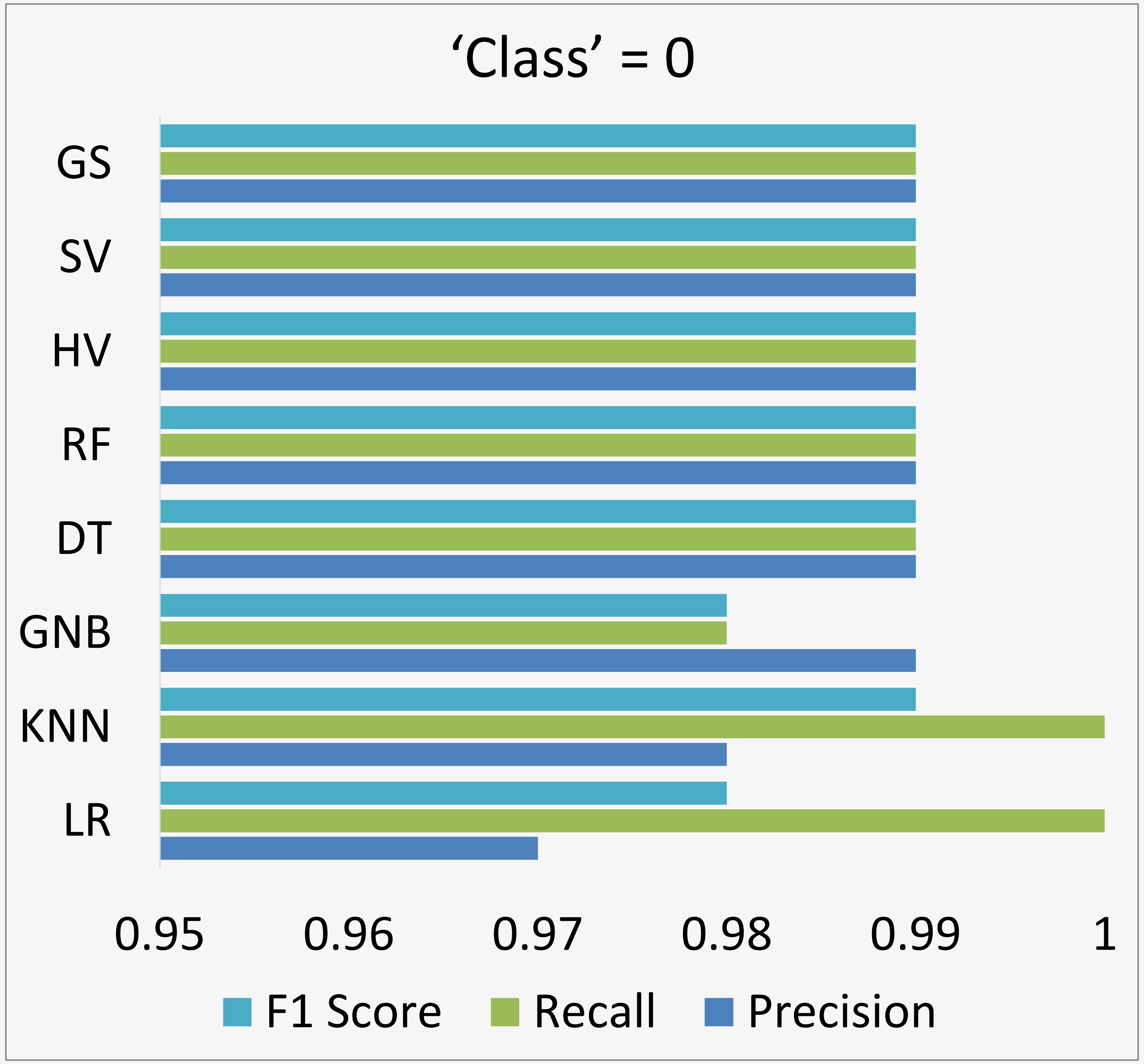


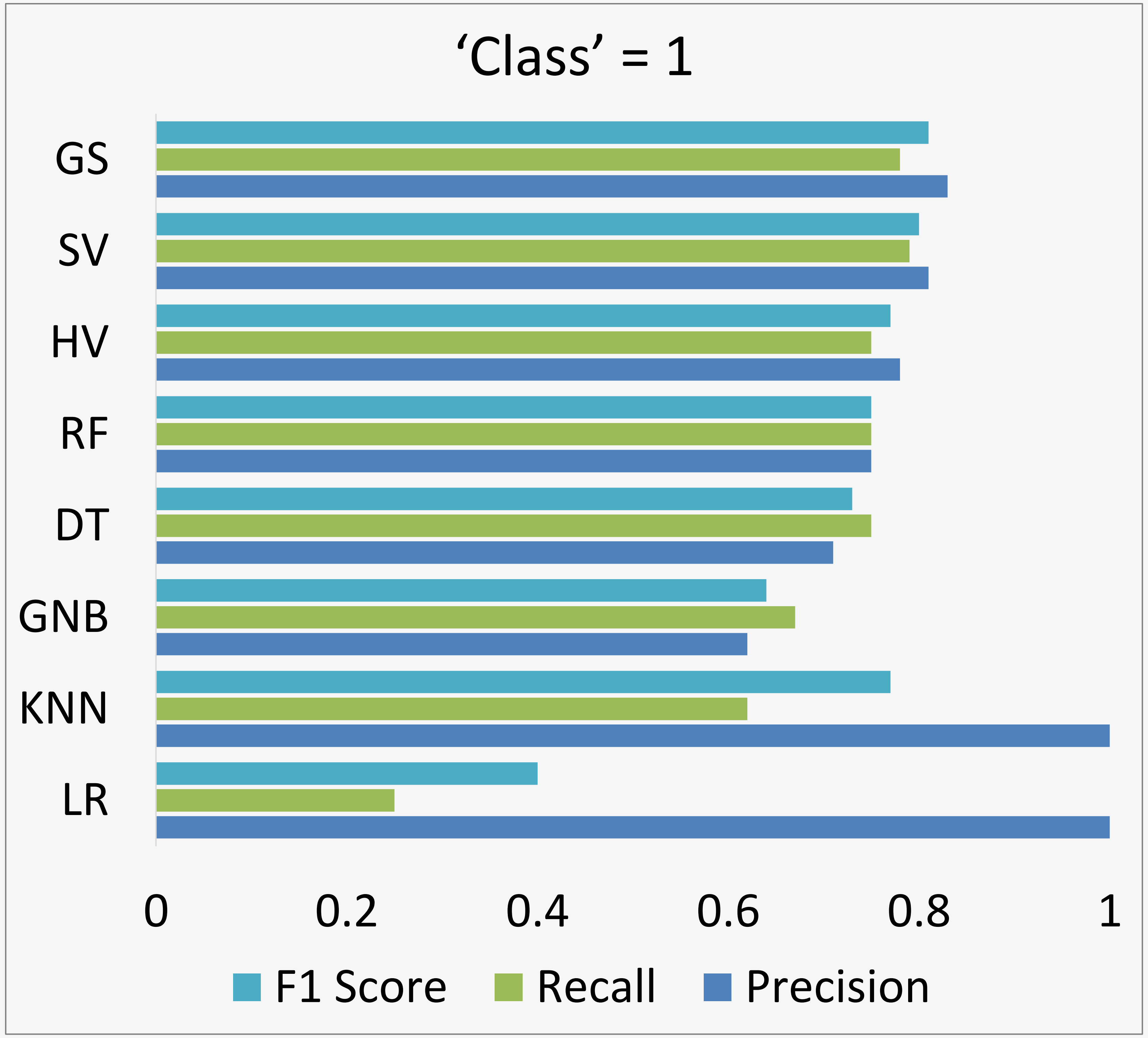


Metrics

Lazada ‘airpods pro’: other scores

Step 5
Evaluate model

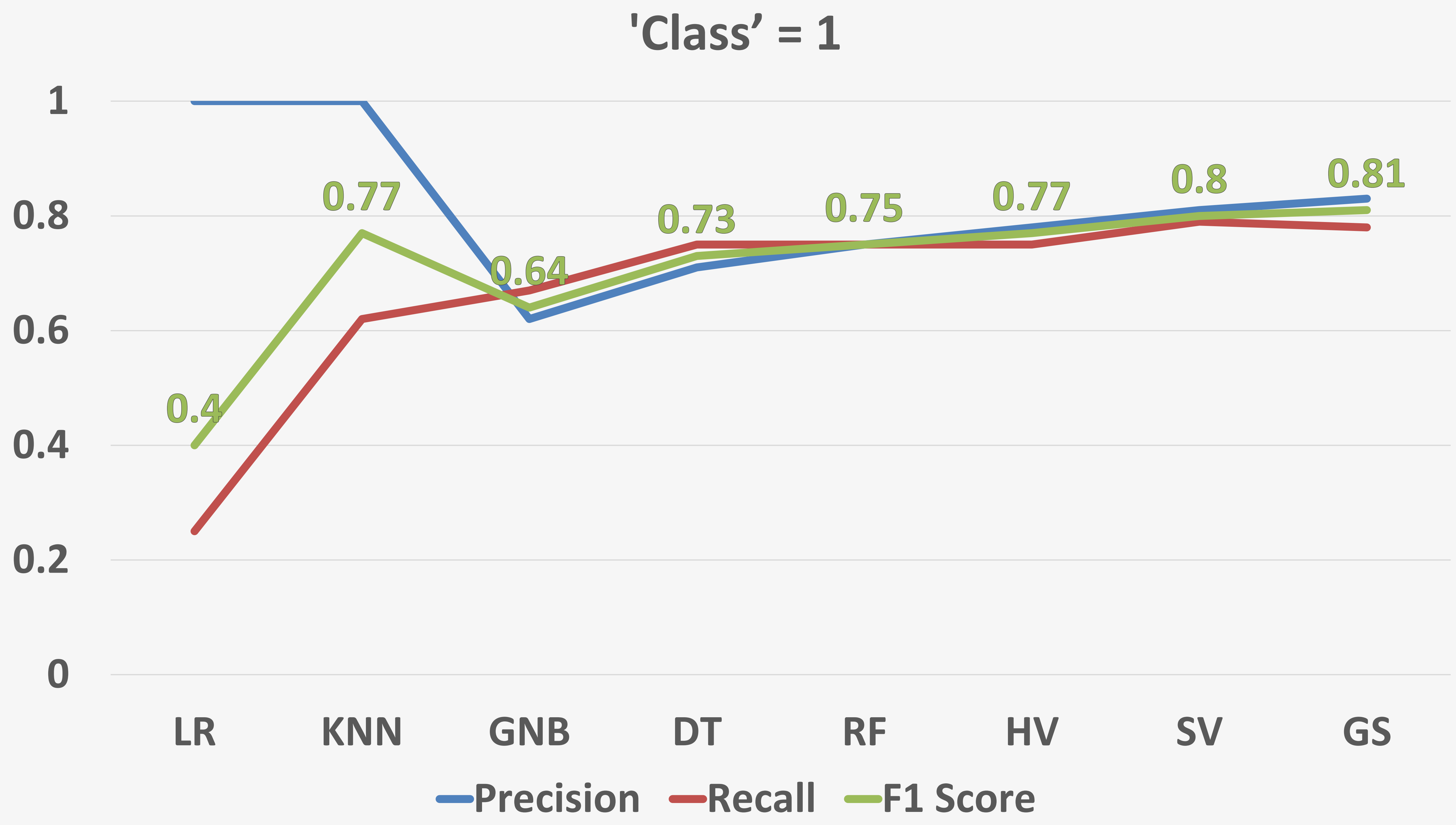




Metrics

Lazada ‘airpods pro’: best estimator

Step 5
Evaluate model



Metrics

All products: best estimator

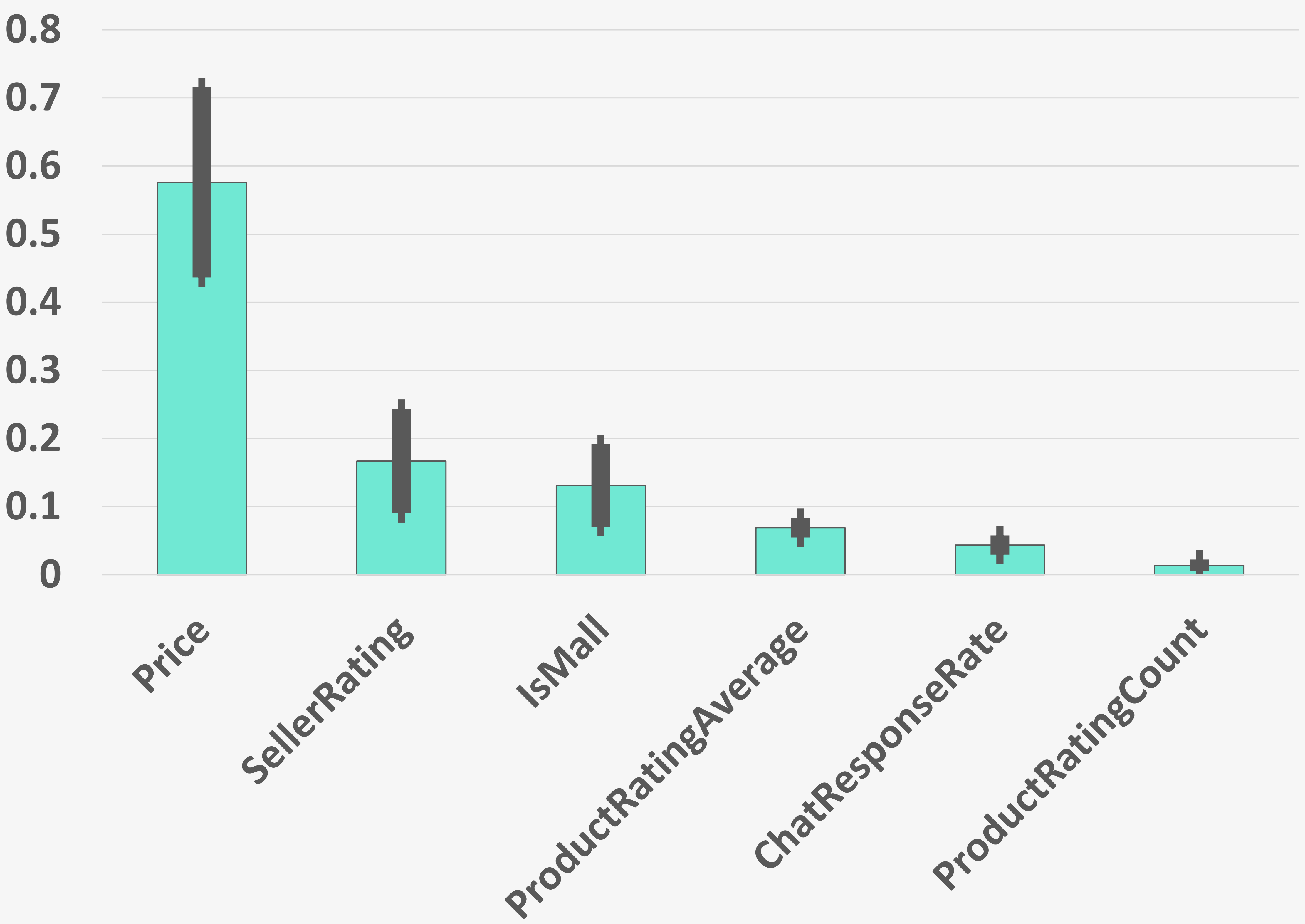
Step 5
Evaluate model

Dataset	Best estimator	F1 score
Lazada ‘airpods pro’	Soft Voting with GridSearchCV	0.81
Lazada ‘apple watch’	LR	1
Lazada ‘ps4 controller’	Soft Voting with GridSearchCV	0.77
Shopee ‘airpods pro’	Soft Voting with GridSearchCV	0.78
Shopee ‘apple watch’	Soft Voting with GridSearchCV	0.85
Shopee‘ps4 controller’	Soft Voting with GridSearchCV	0.86

Metrics

All products: Feature Importance

Step 5
Evaluate model



Summary

and Conclusions

In summary,

- **Detection of counterfeit items** on e-commerce platforms can be achieved through machine learning
- **Price, Seller Rating, and authentic certification** (e.g., LazMall and ShopeeMall) are key indicators of authenticity

1. How will we scale this in the real world?

2. Points of improvement

Thank you for listening!

Presented by:

Romar Acejo

Chris Dion Bautista

Ladie Guevarra