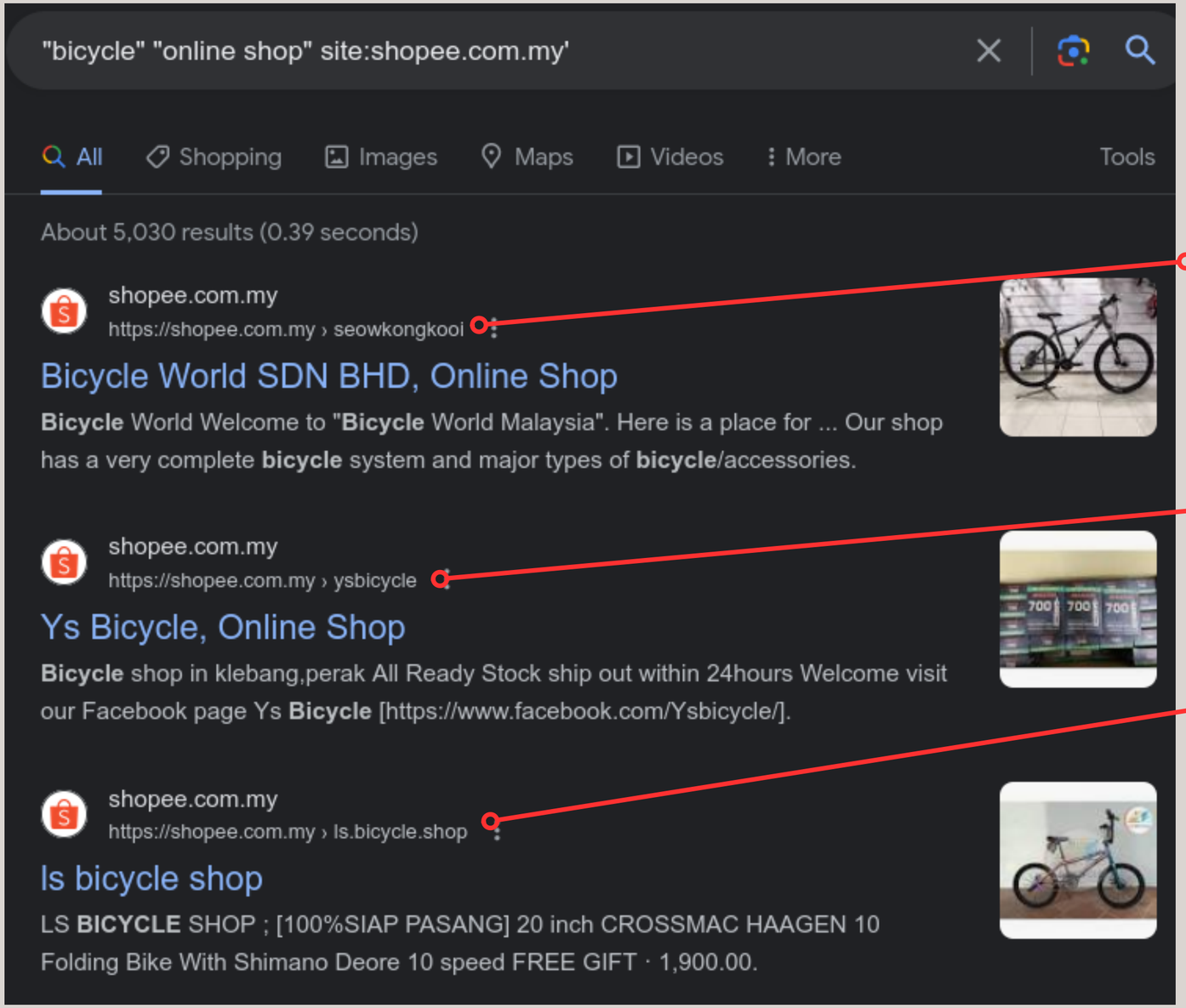
K Youth GA-DSBC-23-003

Shopee Bike Sales Product Classification

Presented by Amir Hamzah Bin Mohd Nasir



shopee.com.my/seowkongkooi

- shopID: 151042127

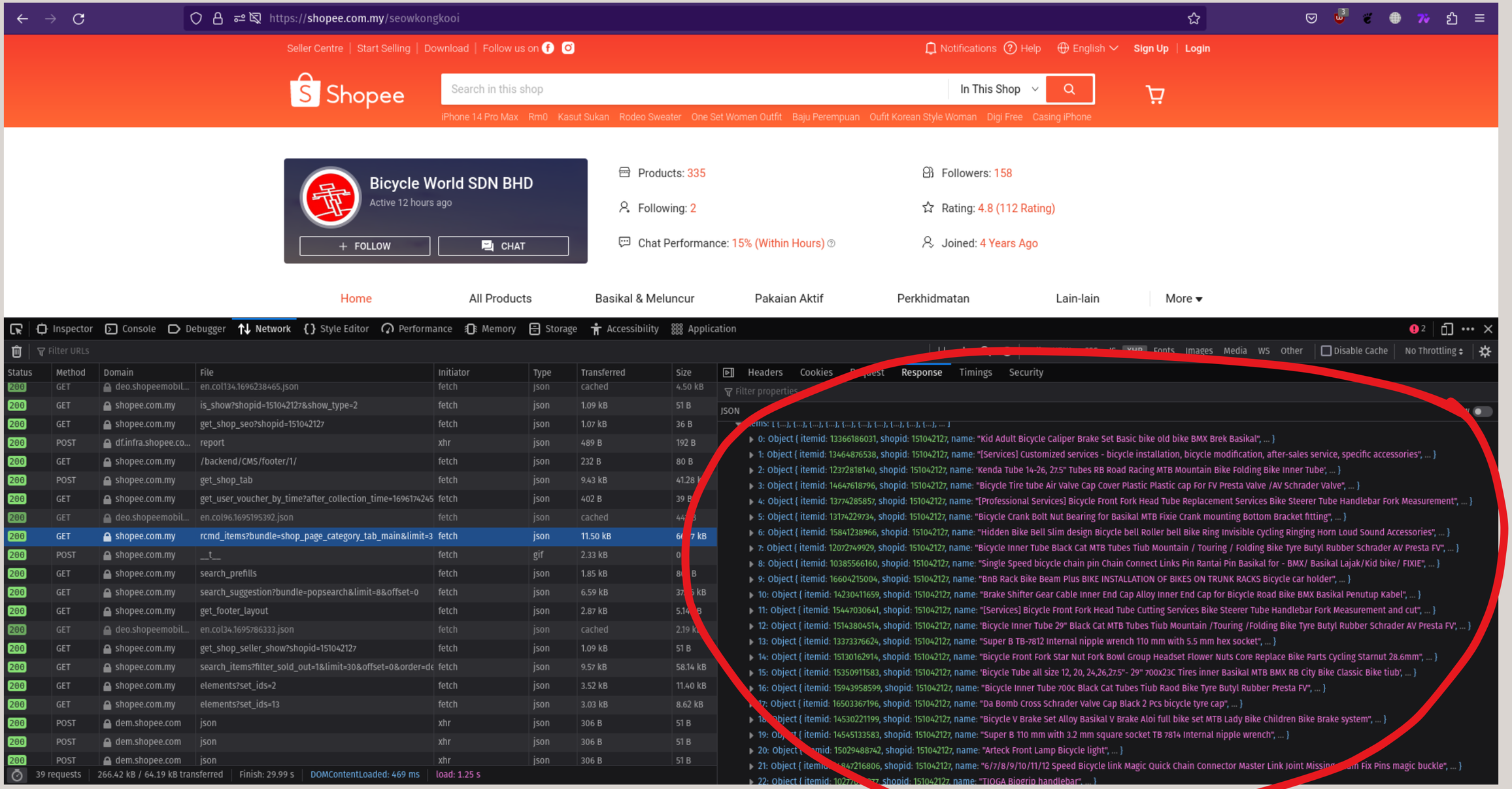
shopee.com.my/ysbicycle

- shopID: 423415991

shopee.com.my/ls.bicycle.shop

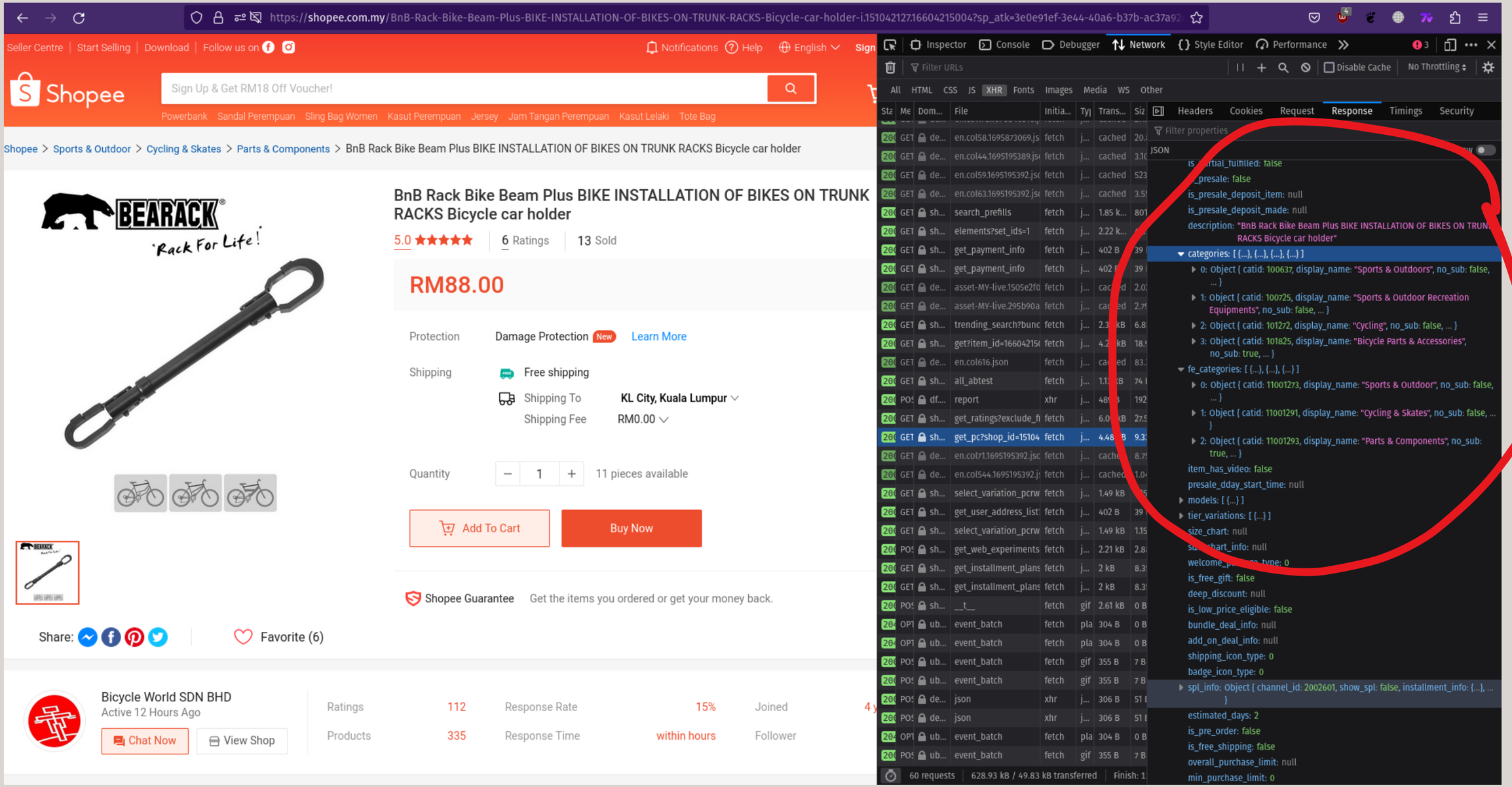
- shopID: 193585755

“https://shopee.com.my/api/v4/shop/rcmd_items” Params = {shop_id : 151042127}



“https://shopee.com.my/api/v4/pdp/get_pc”

Params = {shop_id : 151042127,
item_id : 16604215004}



Google Scraper

- Get each shop selling bicycle related products
- Get each shopID

Shop API

- Get shop details
- Get inventory data
- Get each productID

Product API

- Get details of each individual product

Google Scraper

- Get each shop selling bicycle related products
- Get each shopID

Shop API

- Get shop details
- Get inventory data
- Get each productID

Product API

- Get details of each individual product

Privatized!

Shop Scraper vs Product Scraper

Inventory Data (Shop Scraper)

- Product Name
- Price
- Stock
- Main Category

Product Data (Product Scraper)

- Product Name
- Product Description
- Price
- Stock
- Main Category
- Sub Category 1
- Sub Category 2

Among others..

Problem Statement

- Not having subcategories provide little to no insight in data analysis
- Not being able to derive meaningful insights from scraped future bikes sales data

Inventory Data (Shop Scraper)

- Product Name
- Price
- Stock
- Main Category

Product Data (Product Scraper)

- Product Name
- Product Description
- Price
- Stock
- Main Category
- Sub Category 1
- Sub Category 2

SHOPEE

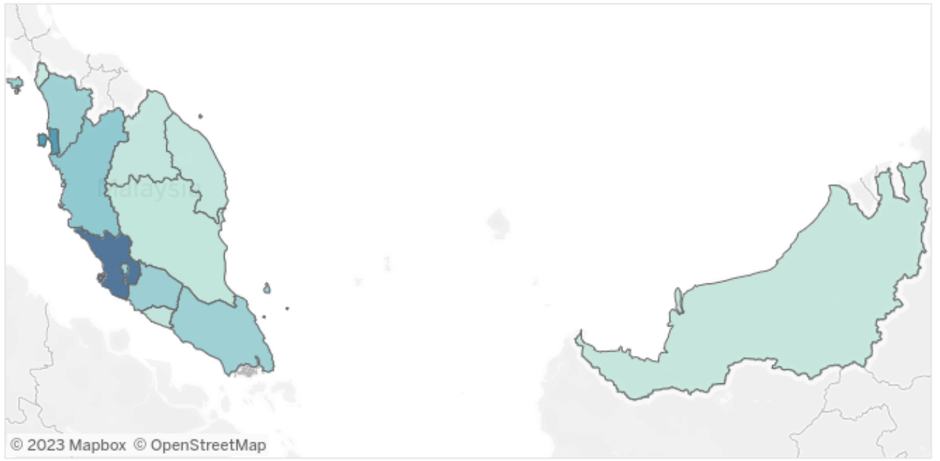
Category: Bicycles

Overview | Store Analysis

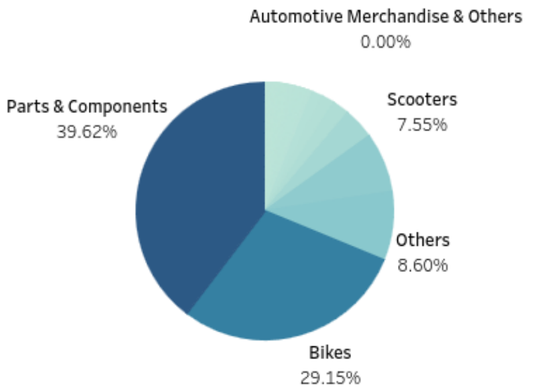
Total Revenues	Number of Stores	Number of Unique Products	Total Products Sold
RM2,706,927.66	208	8,139	82,529

Country	State/..	
Malaysia	Selangor	RM885,034.23
	Penang	RM565,242.67
	Perak	RM252,005.82
	Negeri Sem..	RM200,890.85
	Kuala Lump..	RM191,833.05
	Kedah	RM188,681.42
	Johor	RM187,803.43
	Melaka	RM20,580.85
	Pahang	RM19,546.01
	Pulau Pinang	RM18,406.07
	Kelantan	RM12,601.06
	Terengganu	RM9,352.00
	Sarawak	RM822.00
	Perlis	RM358.00
Mainland China	Mainland C..	RM146,981.38
Taiwan	Taiwan	RM6,685.52
Indonesia	Indonesia	RM103.30

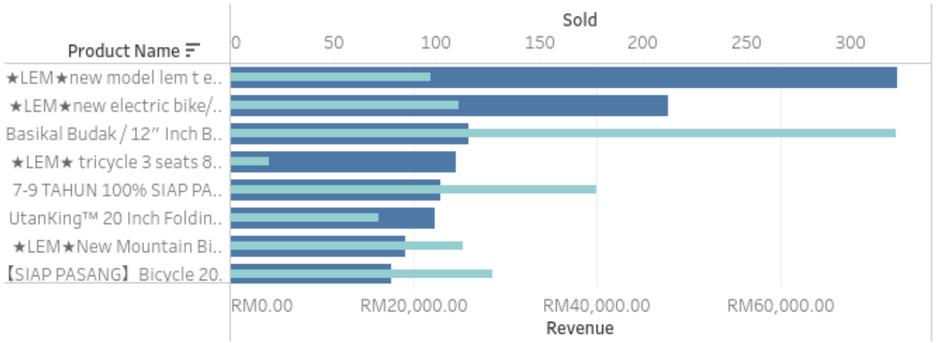
Total Revenue Per Region



Market Share by Category



Best sellers by Revenue/Sold



Market Share by Shops Revenue

EDA

Some Considerations

01
Data can’t be filtered further much to help in modeling

02
Only able to train on product name not product description

03
Main category is filterable

04
F1-score to be target metric due to likely class imbalance

05
There will be a lot of noise, nothing we can do about it

06
Flat approach

Present Data

- Product Name
- Product Description
- Main Category
- Sub Category 1
- Sub Category 2

Future Data

- Product Name
- Price
- Stock
- Main Category

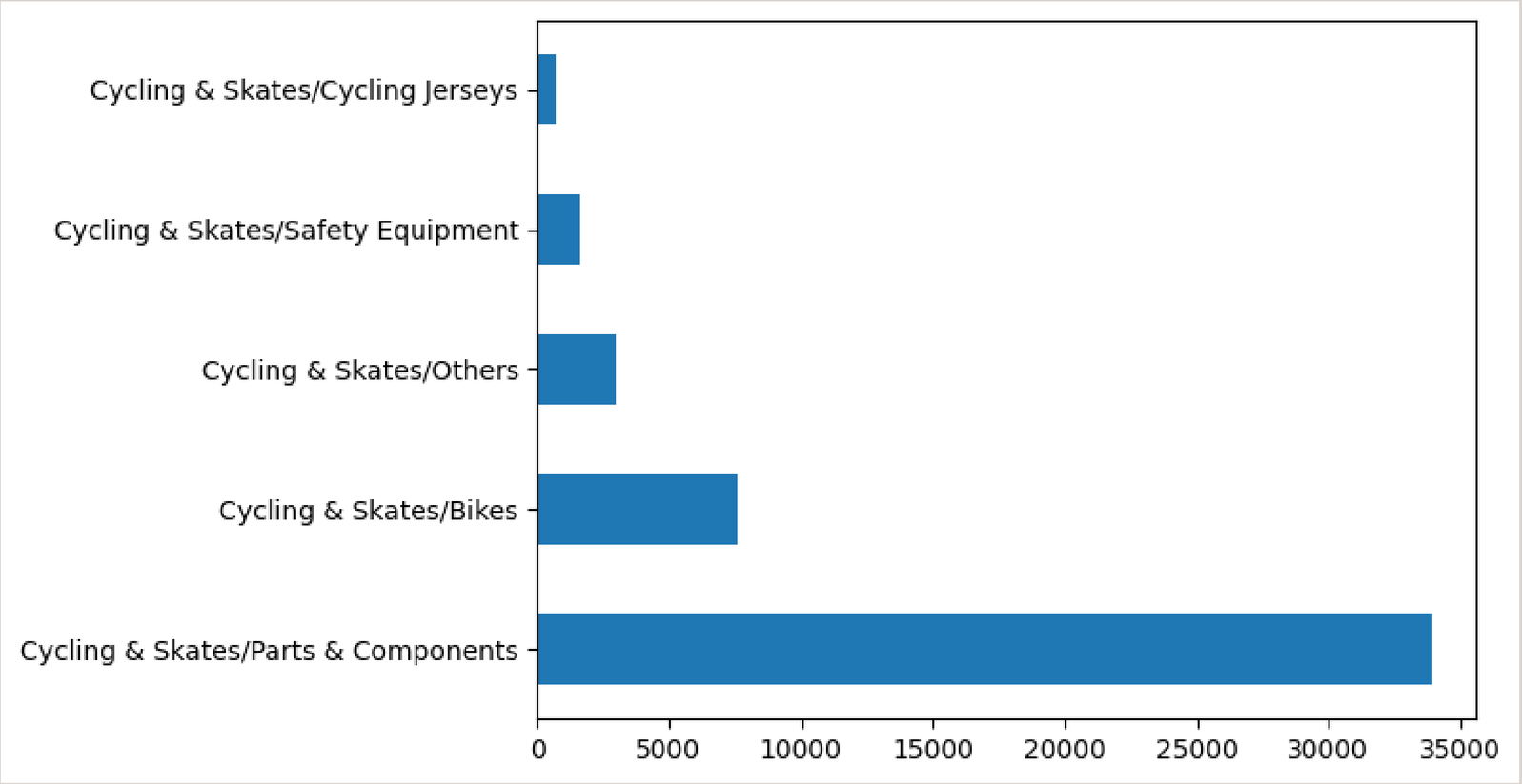
Flat Approach

Category	Shopee > Sports & Outdoor > Cycling & Skates > Safety Equipment
Brand	Bell Helmets

X=Categories
y=Product Name

- Imbalanced Dataset
- Has 6:1 ratio from the 1st and 2nd class.
- Data required a lot of converting.
- Cleaning was kept to a minimum to retain info
- Total no. of categories: 54
- Total no. of samples: 54041

Top 5 categories




```
fe_categories
Cycling & Skates/Parts & Components    0.627450
Cycling & Skates/Bikes                 0.140689
Cycling & Skates/Others                0.056124
Cycling & Skates/Safety Equipment      0.029718
Cycling & Skates/Cycling Jerseys       0.012491
Name: proportion, dtype: float64
```

Modelling

Base model evaluation

Various algorithms trained on data with base parameters.

Algorithms used:

1. Logistic Regression
2. LinearSVC
3. Random Forest
4. MultinomialDB
5. XGBoost

Hyperparameter Tuning

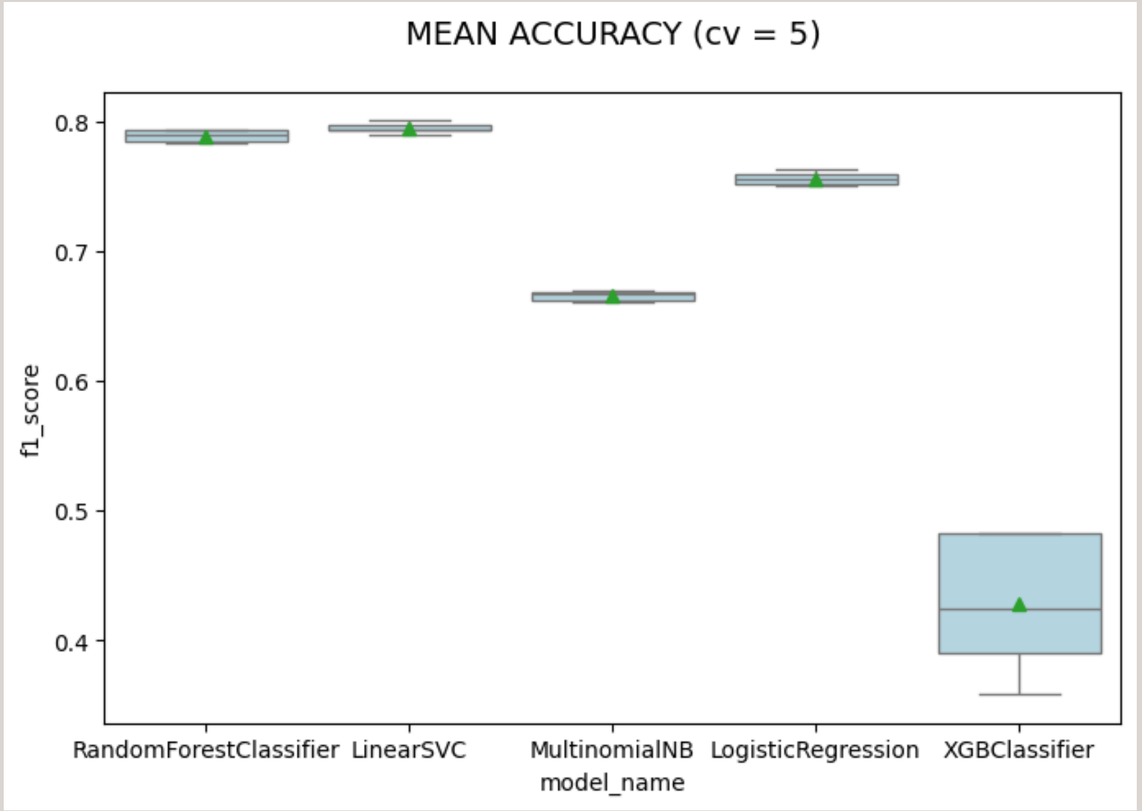
Hyperparameter tuning is done on the best base models to improve scores.

Well.. supposed to :)

```
Best Hyperparameters: {'C': 1.2989993092595722, 'class_weight': None}
Best Score: 0.7936280309801937
CPU times: user 48 s, sys: 45.7 s, total: 1min 33s
Wall time: 36min 25s
```

Base Model Evaluation

model_name	Mean F1-Score	Standard deviation
LinearSVC	0.794682 ✓	0.004338
LogisticRegression	0.755511 ✓	0.005381
MultinomialNB	0.665578	0.004049
RandomForestClassifier	0.788467 ✓	0.004751
XGBClassifier	0.427763	0.055041



Hyperparameter “Tuning”

model_name	Mean F1-Score	Standard deviation
LinearSVC	0.794682	0.004338
LogisticRegression	0.755511	0.005381
MultinomialNB	0.665578	0.004049
RandomForestClassifier	0.788467	0.004751
XGBClassifier	0.427763	0.055041

After “Tuning”

	Model_Name	Accuracy	Precision	Recall	F1-Score	
0	LinearSVC	81.019401	79.664053	81.019401	79.480852	+0.02
1	Random Forest	81.182012	80.120763	81.182012	79.311302	+0.47
2	Logistic Regression	79.140967	77.256282	79.140967	76.487908	+0.93

Reducing Dataset

- “Tuning” didn’t provide much improvement.
- Reduced dataset to minimize noise
- Removed all categories with < 450 samples

Base Dataset

fe_categories	
Cycling & Skates/Parts & Components	33908
Cycling & Skates/Bikes	7603
Cycling & Skates/Others	3033
Cycling & Skates/Safety Equipment	1606
Cycling & Skates/Cycling Jerseys	675
Performance Wear/Sets	520
Water Sports/Others	520
Sports Wear/Jerseys	486
Men's Sports Shoes/Others	468
Cycling & Skates/Scooters	454
Performance Wear/Pants & Shorts	399
Performance Wear/Others	371
Outdoor & Adventure/Camping & Hiking	353
Kids Sports & Outdoor Play/Sport Shoes	313
Outdoor & Adventure/Others	302
Racket Sports/Others	224
Fishing/Others	178
Outdoor & Adventure/Outdoor Bags	177
Exercise & Fitness Equipment/Exercise Bikes & Treamills	173
Performance Wear/Hats	166
Fishing/Hooks	152
Outdoor & Adventure/Lights & Lanterns	145
Fishing/Baits, Lures & Flies	141
Exercise & Fitness Equipment/Others	116
Fishing/Rod & Reels	115
Stick & Ball Games/Golf	100
Running/Running Accessories	98
Exercise & Fitness Equipment/Weight Training Equipment	95
Sports Wear/Training Wear	94
Exercise & Fitness Equipment/Supports, Tapes & Braces	92
Acrobatic & Martial Arts/Training Equipment	85
Exercise & Fitness Equipment/Yoga & Pilates	76
Exercise & Fitness Equipment/Accessories	75
Performance Wear/Compression Tops	75

Reduced Dataset

fe_categories	
Cycling & Skates/Parts & Components	33908
Cycling & Skates/Bikes	7603
Cycling & Skates/Others	3033
Cycling & Skates/Safety Equipment	1606
Cycling & Skates/Cycling Jerseys	675
Performance Wear/Sets	520
Water Sports/Others	520
Sports Wear/Jerseys	486
Men's Sports Shoes/Others	468
Cycling & Skates/Scooters	454

Name: count, dtype: int64

	precision	recall	f1-score	support
Cycling & Skates/Bikes	0.82	0.62	0.71	2535
Cycling & Skates/Cycling Jerseys	0.71	0.78	0.75	213
Cycling & Skates/Others	0.56	0.27	0.36	999
Cycling & Skates/Parts & Components	0.85	0.95	0.90	11184
Cycling & Skates/Safety Equipment	0.91	0.85	0.88	538
Cycling & Skates/Scooters	0.84	0.72	0.77	134
Men's Sports Shoes/Others	0.73	0.77	0.75	146
Performance Wear/Sets	0.73	0.70	0.71	179
Sports Wear/Jerseys	0.84	0.64	0.73	168
Water Sports/Others	0.99	0.86	0.92	165
accuracy			0.84	16261
macro avg	0.80	0.71	0.75	16261
weighted avg	0.83	0.84	0.83	16261

	precision	recall	f1-score	support
Bikes	0.82	0.62	0.71	2535
Cycling Jerseys	0.71	0.78	0.75	213
Jerseys	0.84	0.64	0.73	168
Others	0.70	0.42	0.52	1310
Parts & Components	0.85	0.95	0.90	11184
Safety Equipment	0.91	0.85	0.88	538
Scooters	0.84	0.72	0.77	134
Sets	0.73	0.70	0.71	179
accuracy			0.84	16261
macro avg	0.80	0.71	0.75	16261
weighted avg	0.84	0.84	0.83	16261

Classification Report

- Diminishing returns the more the dataset is reduced
- Subcategory2 - Others have significantly lower scores than other subcategories
- Others is too vague of a subcategory.

```
reduced_df_noothers = df[df['fe_category2'] != 'Others']
```

= Bye Others!

	Model_Name	Accuracy	Precision	Recall	F1-Score
0	LinearSVC	81.019401	79.664053	81.019401	79.480852
1	Random Forest	81.182012	80.120763	81.182012	79.311302
2	Logistic Regression	79.140967	77.256282	79.140967	76.487908
3	LinearSVC Reduced	84.023123	82.984387	84.023123	82.776058
4	Random Forest Reduced	84.853330	84.212341	84.853330	83.379299
5	Logistic Regression Reduced	83.481951	82.575642	83.481951	81.684274
6	LinearSVC No Others	87.110420	86.721738	87.110420	86.371617
7	Random Forest No Others	87.110420	87.018846	87.110420	86.116059
8	Logistic Regression No Others	85.256610	84.595317	85.256610	83.767911
9	LinearSVC No Others Reduced	89.179055	88.913741	89.179055	88.621916
10	Random Forest No Others Reduced	89.748226	89.634477	89.748226	89.116105
11	Logistic Regression No Others Reduced	88.489353	88.365208	88.489353	87.668393

Much better metrics

Model winner:
RF No Others Reduced

Base + No others
+ 4%~

Reduced + No others
+ 7%~

Model Evaluation

ROC Curve

- Brawling between classes
- Model's ability to distinguish between predictions

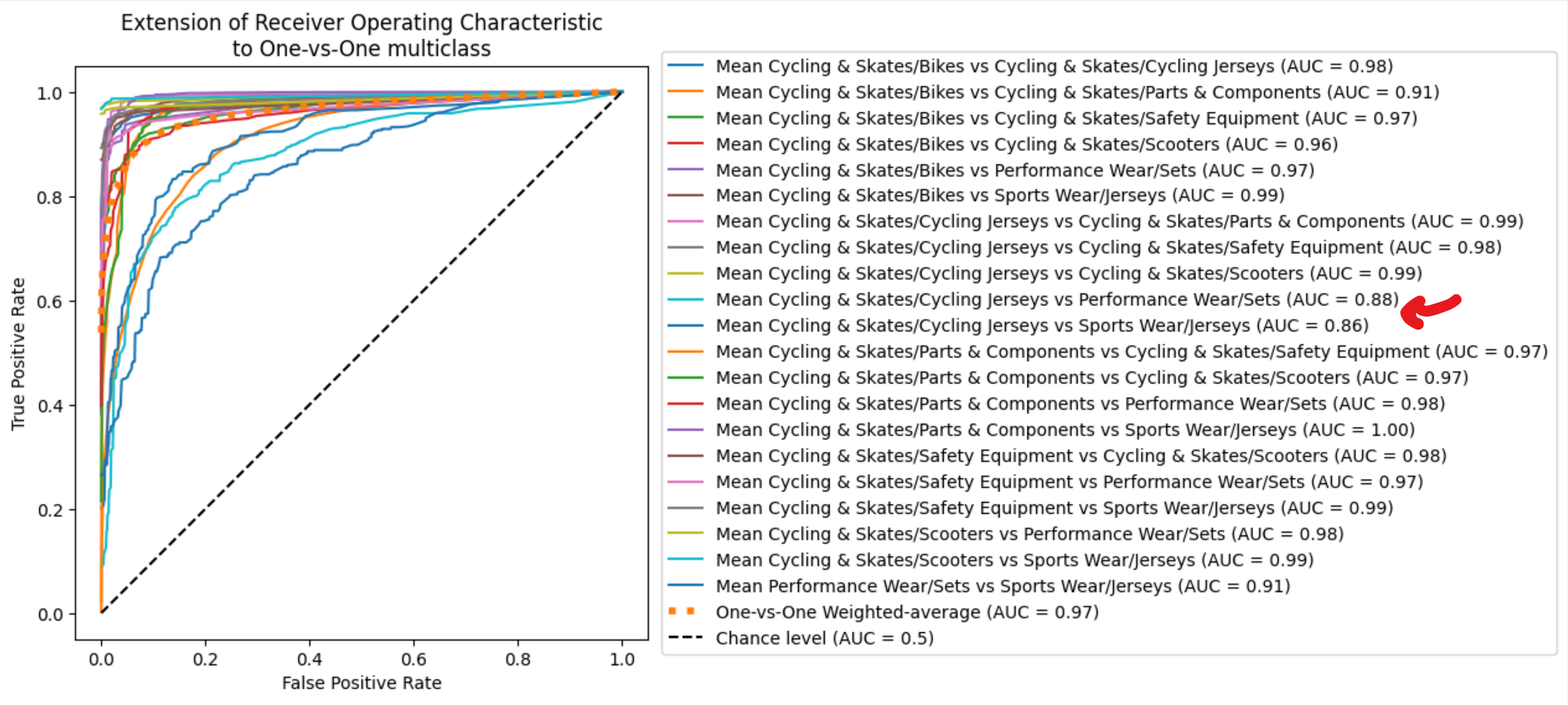
Coefficient Analysis

- See what features are important in classifying each class

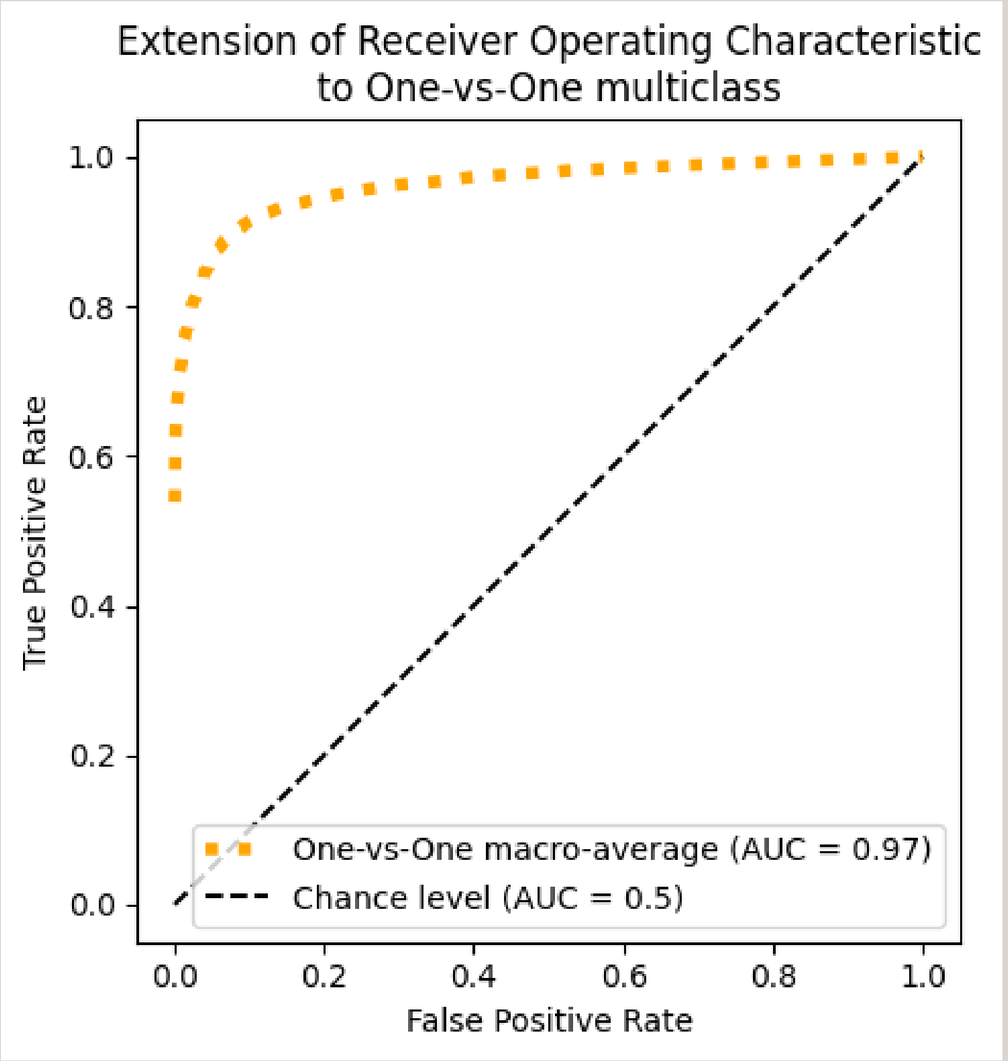
Predict New Data

- Predict with fresh new data and see firsthand if predictions are reasonable

Random Forest Classifier Reduced No Others



AUC not factored by class imbalance



Coefficient Analysis

```
Top features for class Cycling & Skates/Parts & Components (Weight: 33908):
bicycle: 3.8438
fender: 3.5730
handlebar: 3.1309
saddle: 3.0796
chain: 2.8801
brake: 2.7803
seat: 2.6406


Top features for class Cycling & Skates/Bikes (Weight: 7603):
bike: 4.1272
speed: 2.9157
20: 2.8534 ✗
shimano: 2.6000
bicycle: 2.5065
gomax: 2.3541
camp: 2.3046 ✗
```

```
Top features for class Cycling & Skates/Safety Equipment (Weight: 1606):
helmet: 10.3053
helmets: 4.4488
knee: 2.4509
mips: 2.4477
abus: 2.3603 ✗
park: 2.3076 ✗
frontier: 2.2981


Top features for class Cycling & Skates/Cycling Jerseys (Weight: 675):
jersey: 7.8840
cycling: 3.5513
cyclepro: 1.7784
shirts: 1.6969
clothing: 1.6627
downhill: 1.6165 ✗
ykywbike: 1.5804 ✗ Huh?
```

New predictions

	product_name	categories_decoded
51792	YKYWBIKE Pro Cool Mesh Superlight Bike Cycling Base Layers Bicycle Short Sleeve Shirt Breathable Cycling Jerseys Cycling Clothin	Cycling & Skates/Cycling Jerseys
72185	Swivels Connectors Copper And Pvc Feeder Hook Small About 13mm/Large About 15mm	Cycling & Skates/Parts & Components
30176	[Speed Park] Fixed Line Tube For Frame C-Shaped Buckle Outer Fixing Set Price Is One C	Cycling & Skates/Parts & Components
46783	BSC Women Essentials TOUR Jersey 2.0 Sand	Sports Wear/Jerseys
31222	FKR 16X1.75/1.95 AV BICYCLE BUTYL TUBE	Cycling & Skates/Bikes
41462	3T Full Carbon Road Bike Handlebar Matte/Gloss Black Top Carbon Fiber Road Bicycle Handlebar 31.8mm*400/420/440mm	Cycling & Skates/Parts & Components
19253	BikeDIY [LOCAL] MTB Bike Hydraulic Brake Needle Compressing Bushing Brake Hose Bicycle Parts SLX XT XTR Deore 446 T01	Cycling & Skates/Parts & Components
17731	gofondo® Minpoc MAGNETIC clip-on small bag for your Brompton	Cycling & Skates/Bikes
26194	Time Road Pedal Xpro 10 (Black/Grey)	Cycling & Skates/Parts & Components
34337	[MARIN] FAIRFAX 3 700C V URBAN BIKE (2022)	Cycling & Skates/Bikes

It works! Kinda

Recommendation

1

Training Data from one month only won't be able to reliably predict future unseen categories.

Ideally, obtain a larger dataset that spans multiple months for reliable predicting.

2

Future data can't reliably validate model. Only Current data can be validated as they have real categories.

3

Utilize pre-trained models such as BERT or further tuning once a better machine is obtained.

K Youth GA-DSBC-23-003

Thank You

Presented by Amir Hamzah Bin Mohd Nasir