Riding the Data Trail

Text-Based Price Classifier for Road Bikes on Carousell

Did you know?

Today ··→ **2030**

525 km of cycling paths

1,300 km of cycling paths

Source: https://www.lta.gov.sg/content/ltagov/en/getting_around/active_mobility/walking_cycling_infrastructure/cycling.html

From 2027: North-South Corridor

- Singapore's first *Integrated Transport Corridor* with **dedicated cycling paths**

Infrastructure enhancement by the government to promote cycling for commuting and leisure purposes.

Source: https://www.lta.gov.sg/content/ltagov/en/upcoming_projects/road_commuter_facilities/north_south_corridor.html

Cycling Popularity Spills Over to Carousell

CYCLING STILL MORE POPULAR NOW THAN BEFORE COVID

Cyclists and retailers said that many people who have picked up the sport have sustained an interest in it even as "normal life" resumed. Decathlon said that it is making 20 per cent more from bicycle sales now compared to pre-pandemic. 11 Mar 2023

todayonline.com
https://www.todayonline.com › singapore › cycling-bicy...

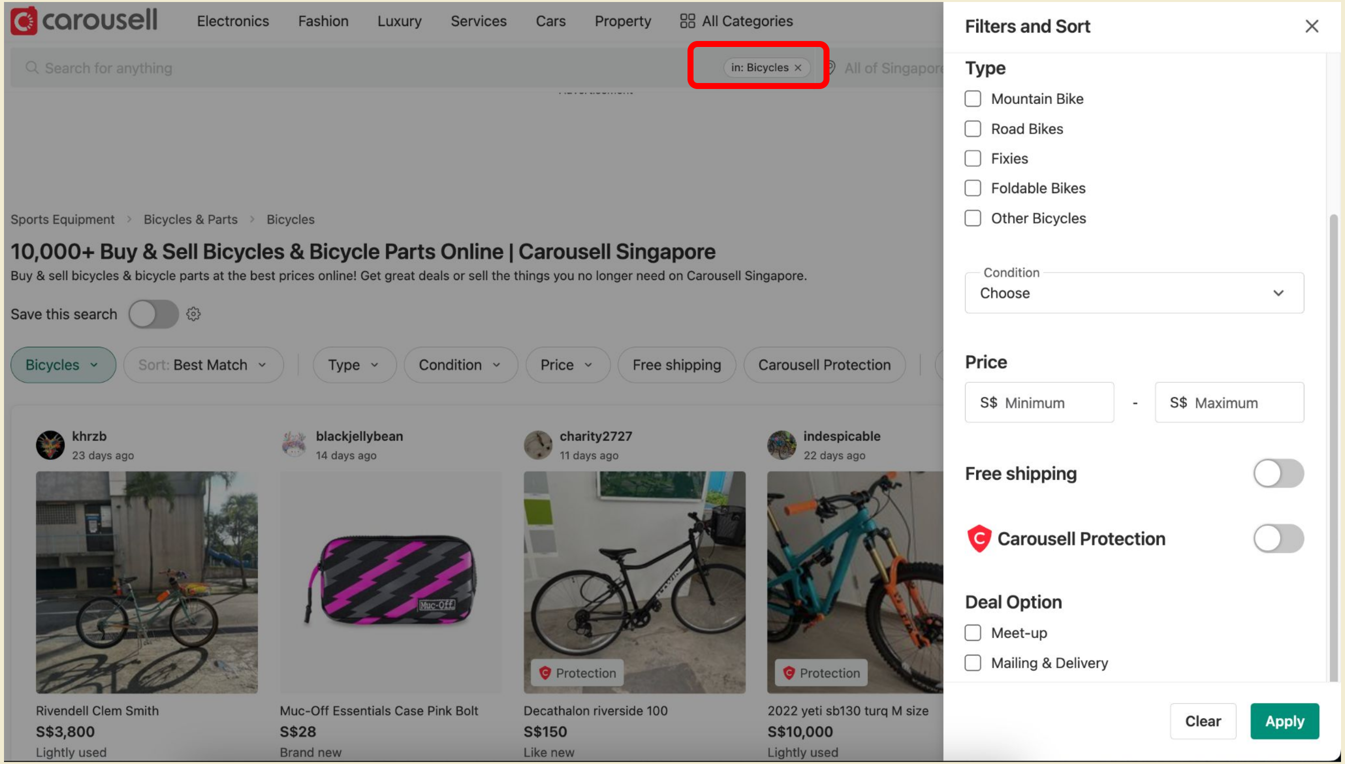
Cycling, bicycle sales on the decline after Covid boom, but ...

carousell

- Founded in Singapore
- For buying / selling new or second-hand goods.. Such as bicycles!

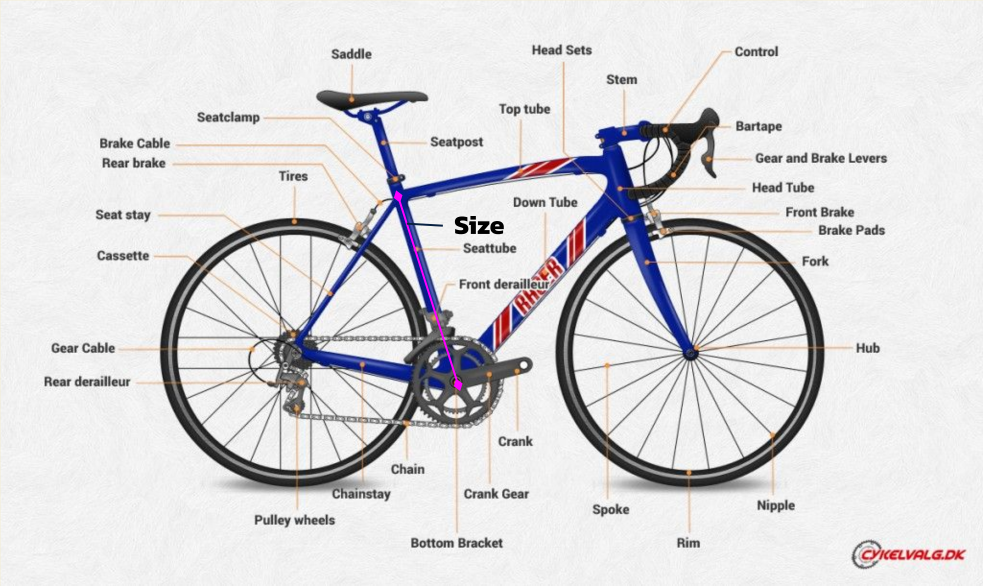
Using Carousell - Pain Points

1. Carousell is not specialized to each product; filters are very generic



Using Carousell - Pain Points

2. There are so many components on a road bike which affects its value!



3. As a buyer / seller, it is difficult to know what's a reasonable price to buy / sell at, and even more challenging as a novice

Image source: https://goodfangsm.life/product_details/62087875.html

Problem Statement → Proposed Solution

1. Carousell's filters are very generic
2. Numerous components affect price
3. Challenging to determine reasonable road bike price, especially as a novice

- Build a tool predicting road bike prices category from user-input values
- Success criteria: buyers / sellers can find out price to buy / sell at **in less time** than scrolling through Carousell to compare prices

Mapping the Journey

Scrape

Listings were scraped from Carousell web, from the 'Bicycles' category

EDA

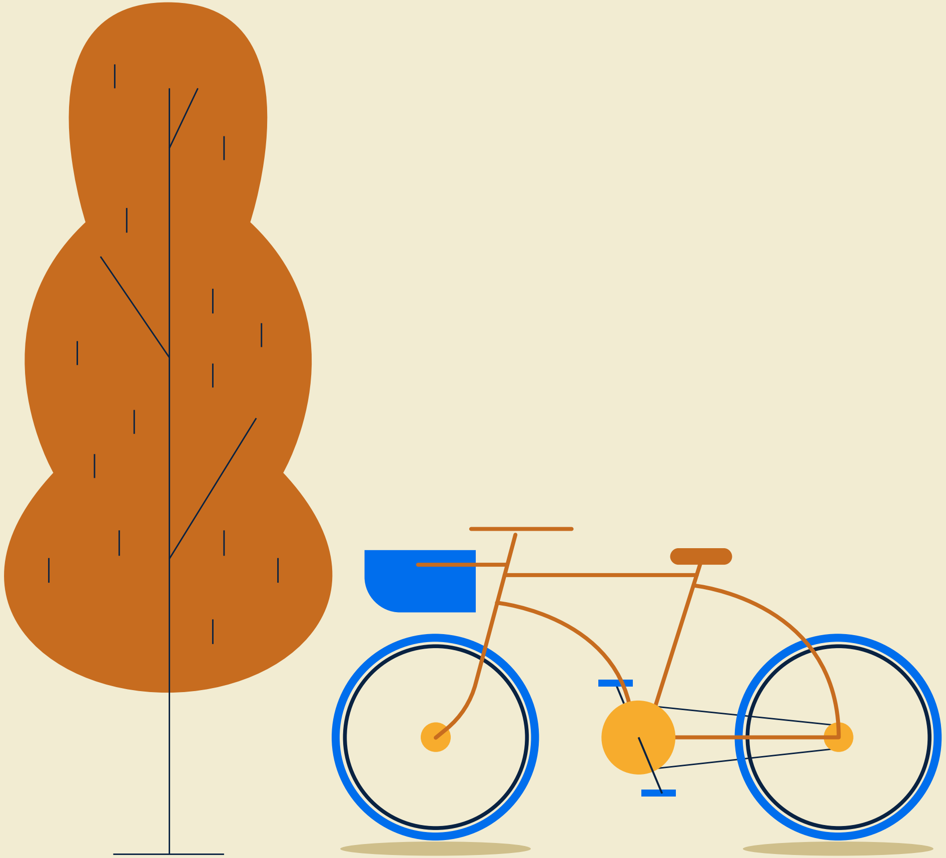
Data cleaning and visualizations; feature engineering and finding patterns

Model

Building classification models

Deploy

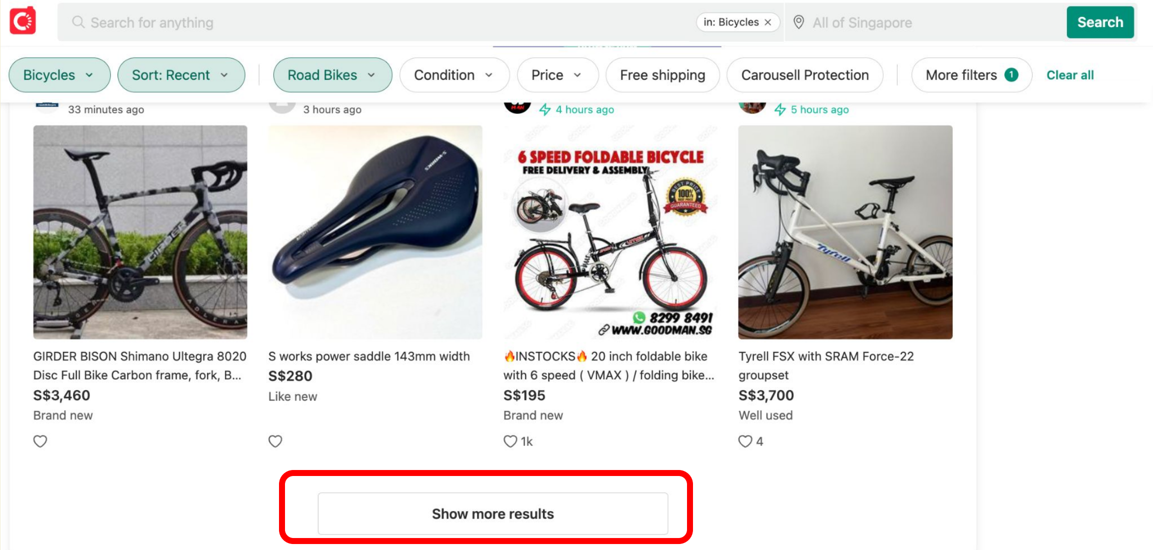
Predicting the class (price level) of a road bike from text input



Data Scraping

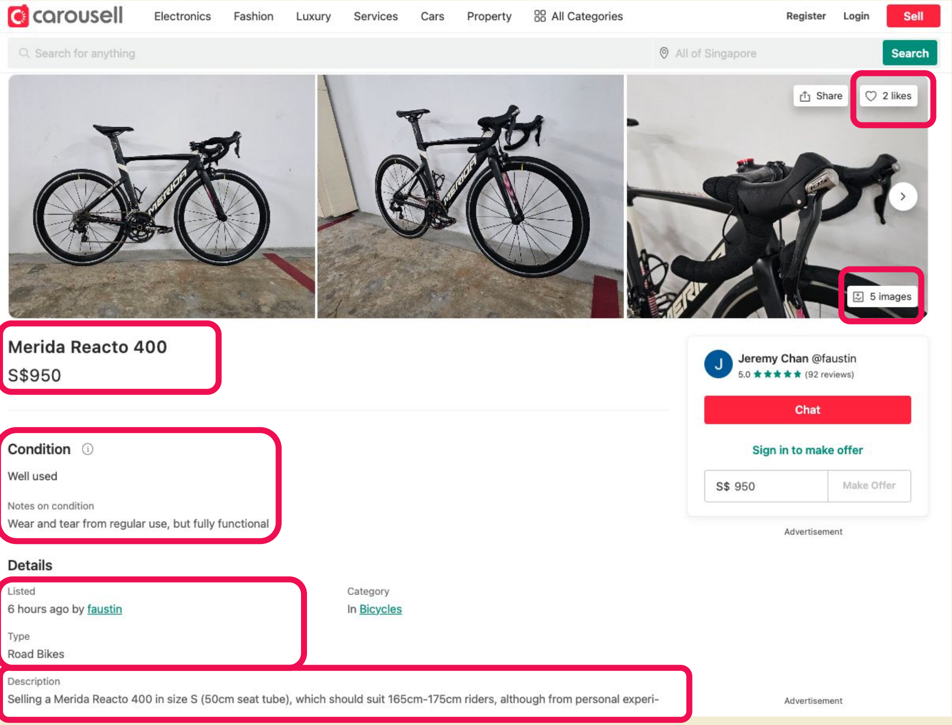
Data scraping from Carousell web

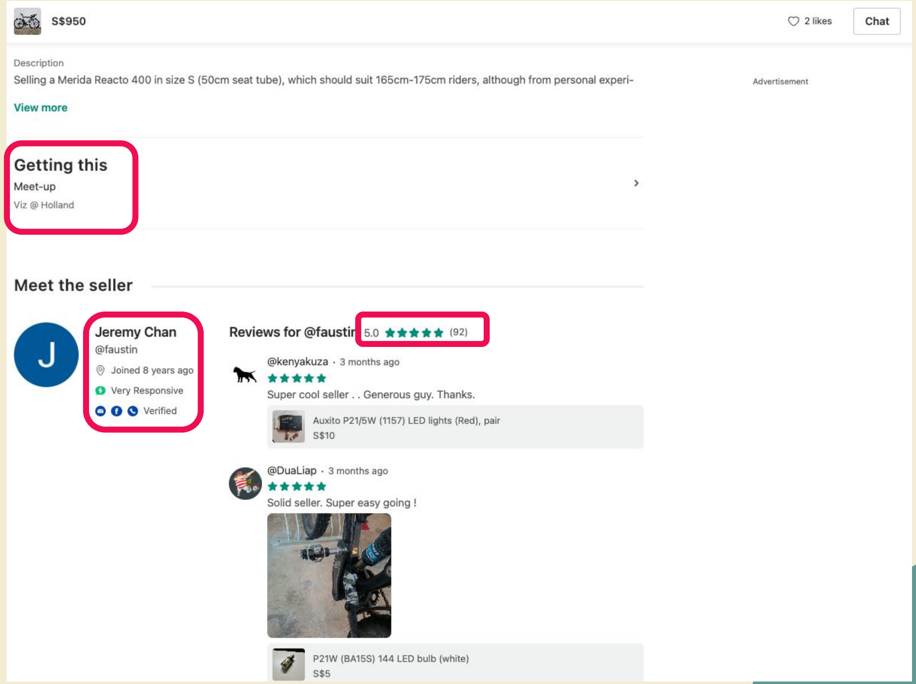
- 'Bicycles' category, with filter for '**Road Bikes**'
- Beautiful Soup, Selenium

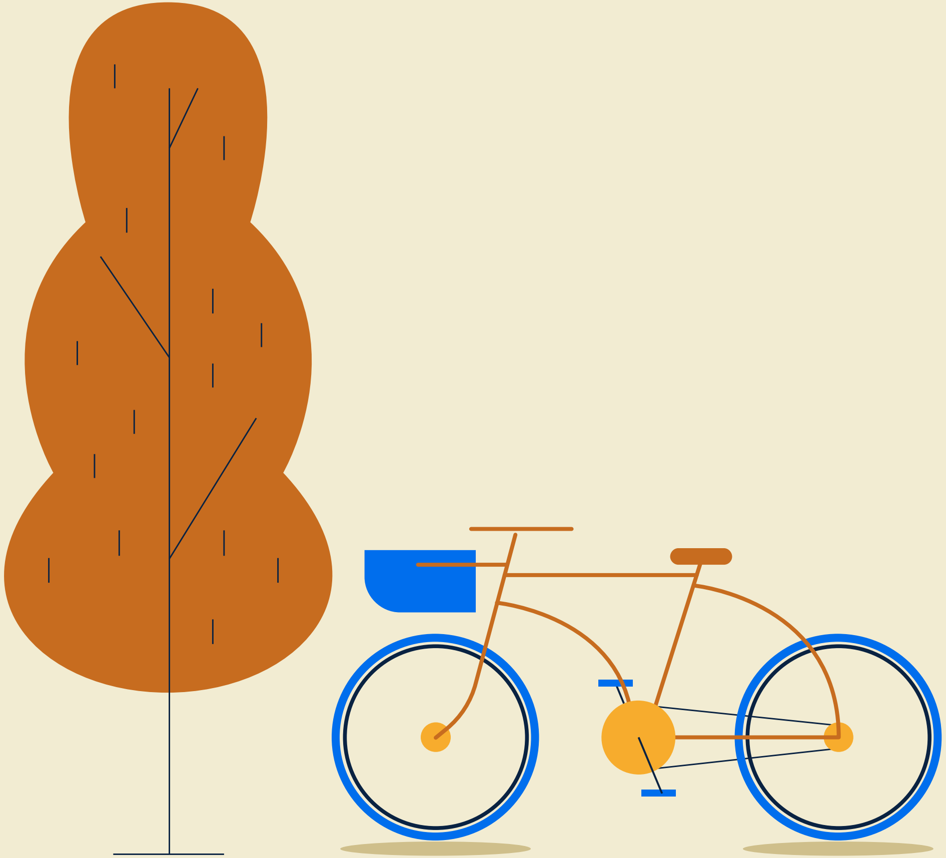


- Dataset obtained: 2896 rows, 29 features

What was Scraped



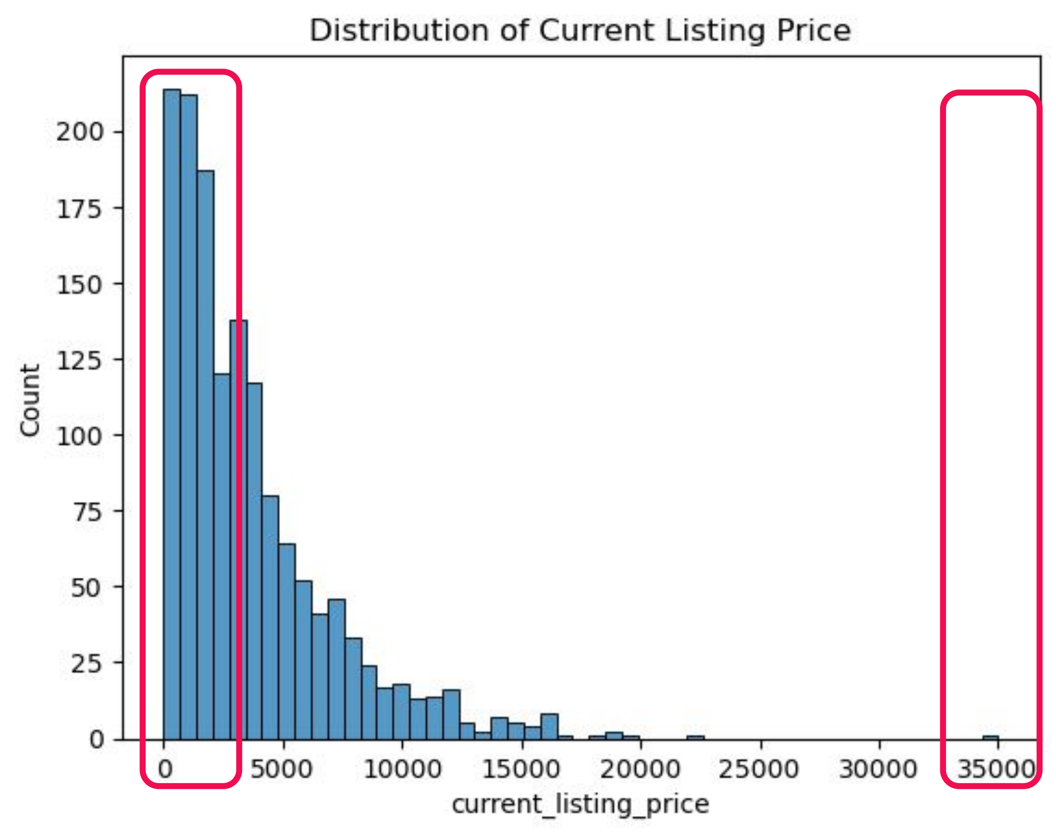




EDA

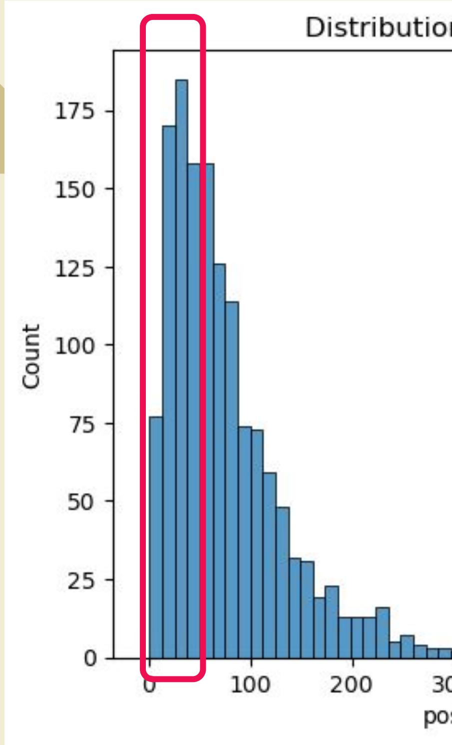
You can enter a subtitle here if you need it

Current Listing Price

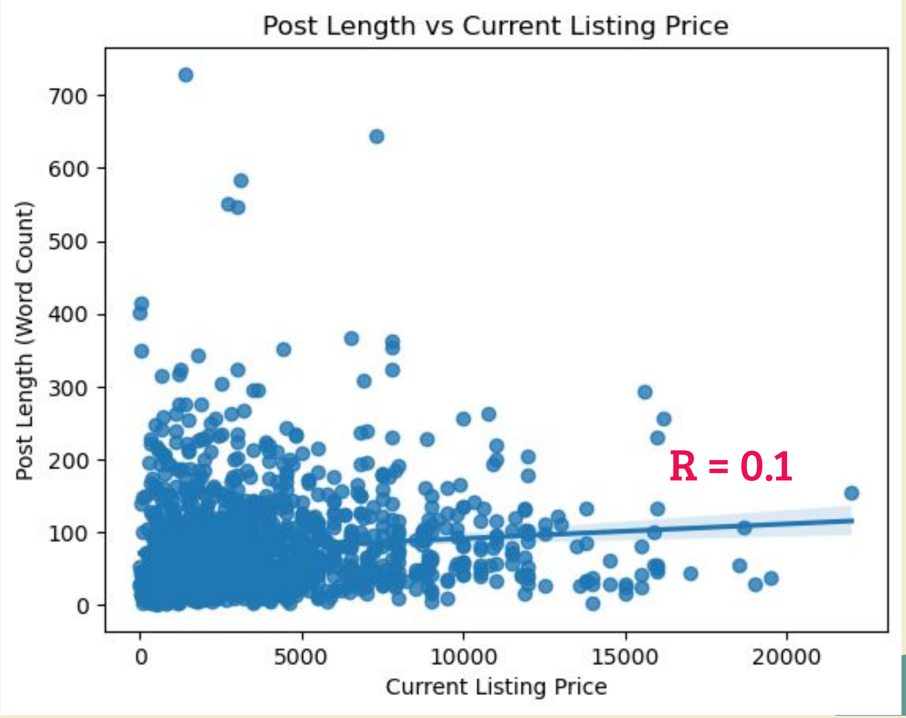


- Listings **up to $2,500** are most common
- The most expensive listing is just **under $35,000 (outlier)**.

Post Word Count



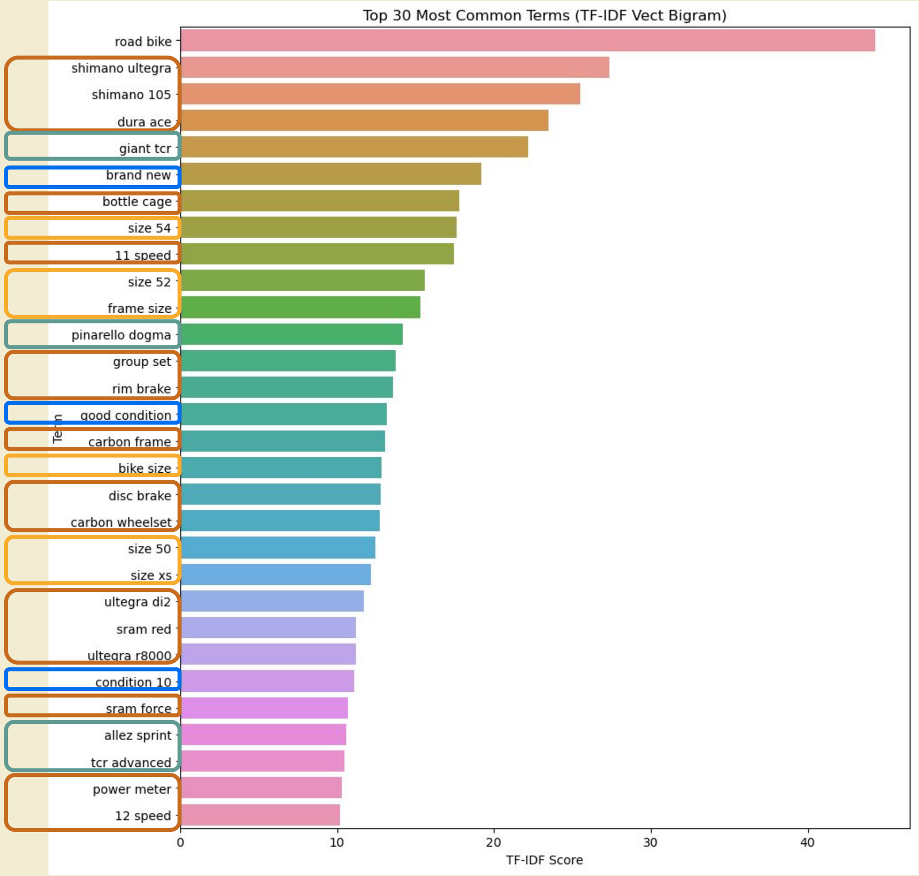
- Most posts have about **10 - 60 words** (from Title and Listing Description)



Number of Images

- **Mean** number of images per post is **5**.
- There are listings which make full use of the **10 image limit**

Top 30 Words (Bigram) - TF-IDF Vectorizer



Legend

Components

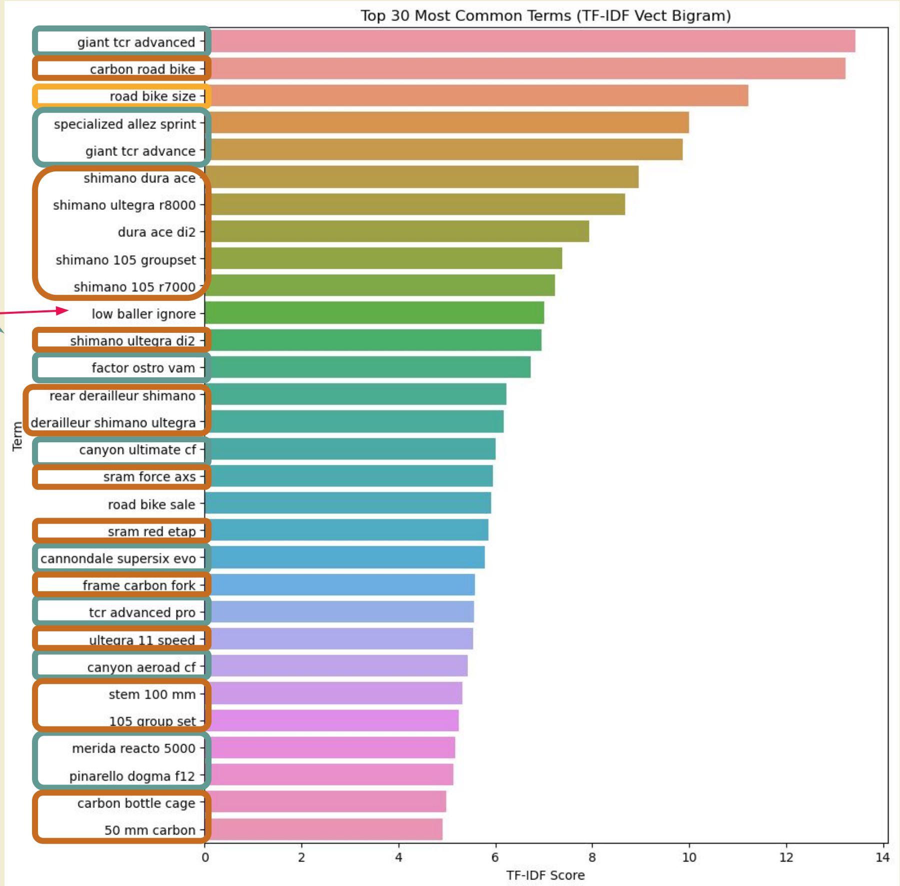
Size

Brand and/or model of bicycles

Condition

- **Condition** bigrams are common despite having a separate input area
- **Size 50, 52, 54** are common

Top 30 Words (Trigram) - TF-IDF Vectorizer



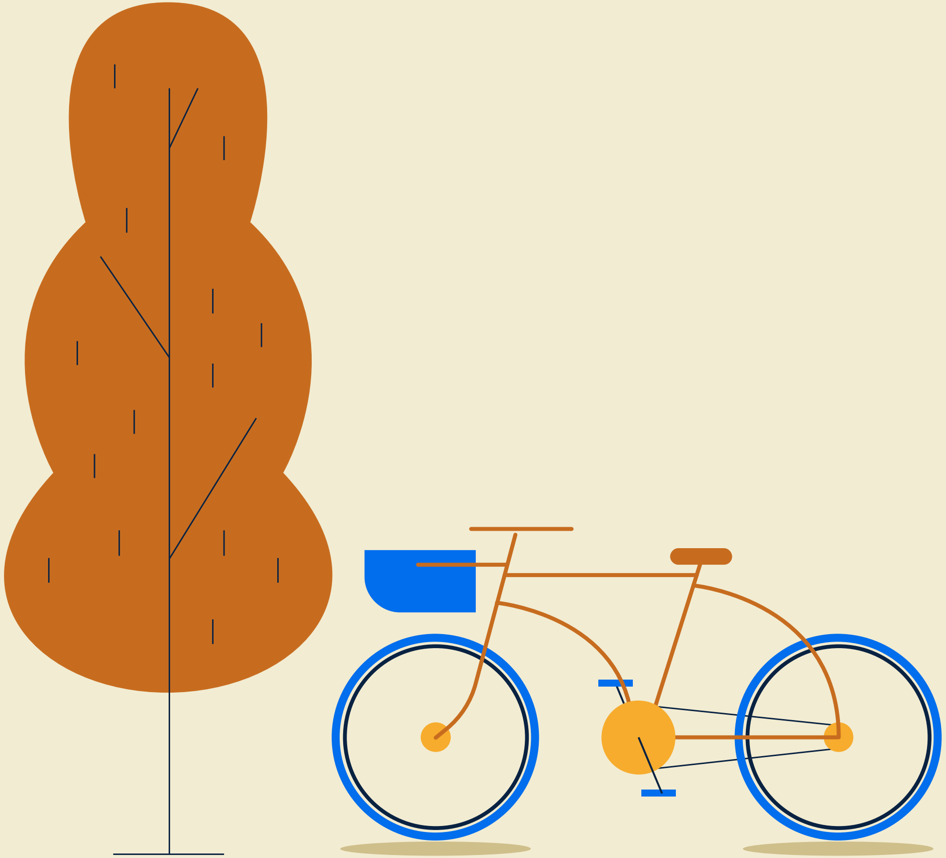
Legend

Components

Brand and/or model of bicycles

Size

- **Components, brand and/or model of bicycles** are most frequently mentioned
- "Low ballers will be ignored!"



Modeling

Pre-Modeling Preparation

1. **Identify** columns to send for modeling

2. Use **BERTopic** for topic identification:
   - Obtain **top 10 words** for each topic and append to dataframe

3. Narrow down dataframe to listings **up to $2000**

4. **Feature engineer** price:
   - Class 0 ($0 – $600)
   - Class 1 ($601 – $1200)
   - Class 2 ($1201 – $2000)

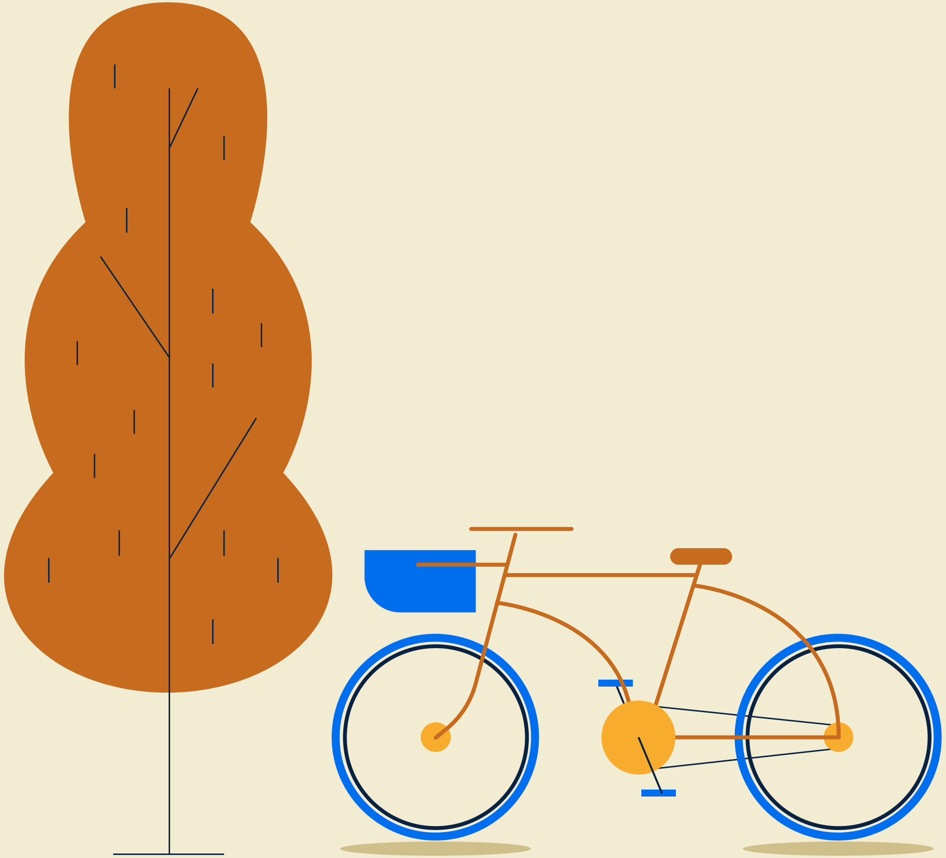
Topics from BERTopic

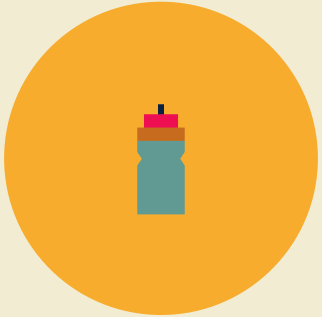
```
Topic 0: bike, size, carbon, shimano, frame, mm, road, new, cm, ultegra
Topic 1: specialized, bike, mm, carbon, tarmac, shimano, size, saddle, sram, ultegra
Topic 2: giant, tcr, shimano, advanced, size, pro, disc, carbon, wheelset, advance
Topic 3: merida, reacto, carbon, bike, shimano, ultegra, scultura, size, frame, road
Topic 4: trek, madone, bontrager, oclv, bike, slr, size, aeolus, emonda, carbon
Topic 5: canyon, cf, bike, ultimate, endurace, aeroad, size, sl, shimano, new
Topic 6: pinarello, dogma, size, di2, f12, dura, ace, bike, wheelset, f10
```

Each topic has a mixture of **brand** and **model** name, **components**, **properties** of the bike

Modeling Summary

	Accuracy (Train)	Accuracy (Test)	AUC	Recall	Precision	F1
Gradient Boosting Classifier (baseline)	0.8365	0.8276	1.0000	0.8276	0.8918	0.8348
Logistic Regression	0.9406	0.9212	0.9869	0.9212	0.9211	0.9210
k-nearest Neighbours	0.9682	0.9655 ↑ **17%**	0.9981	0.9655	0.9675	0.9658

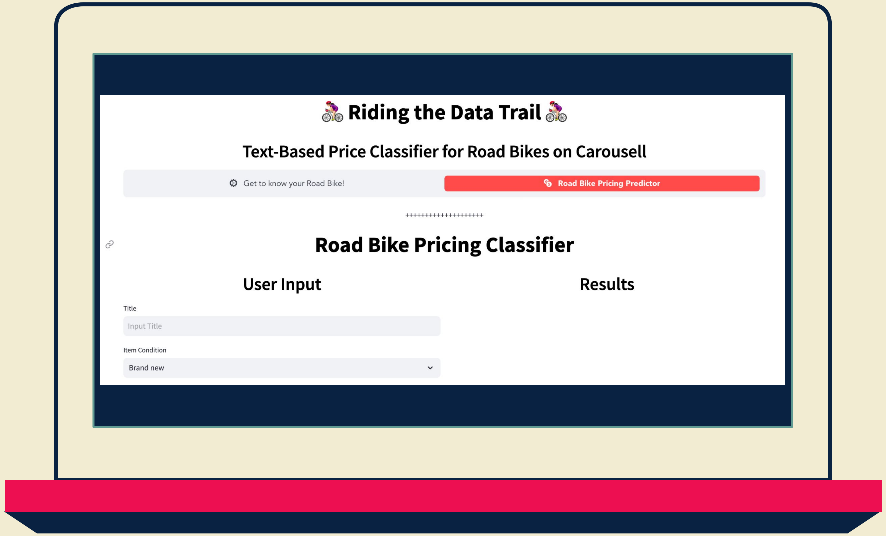


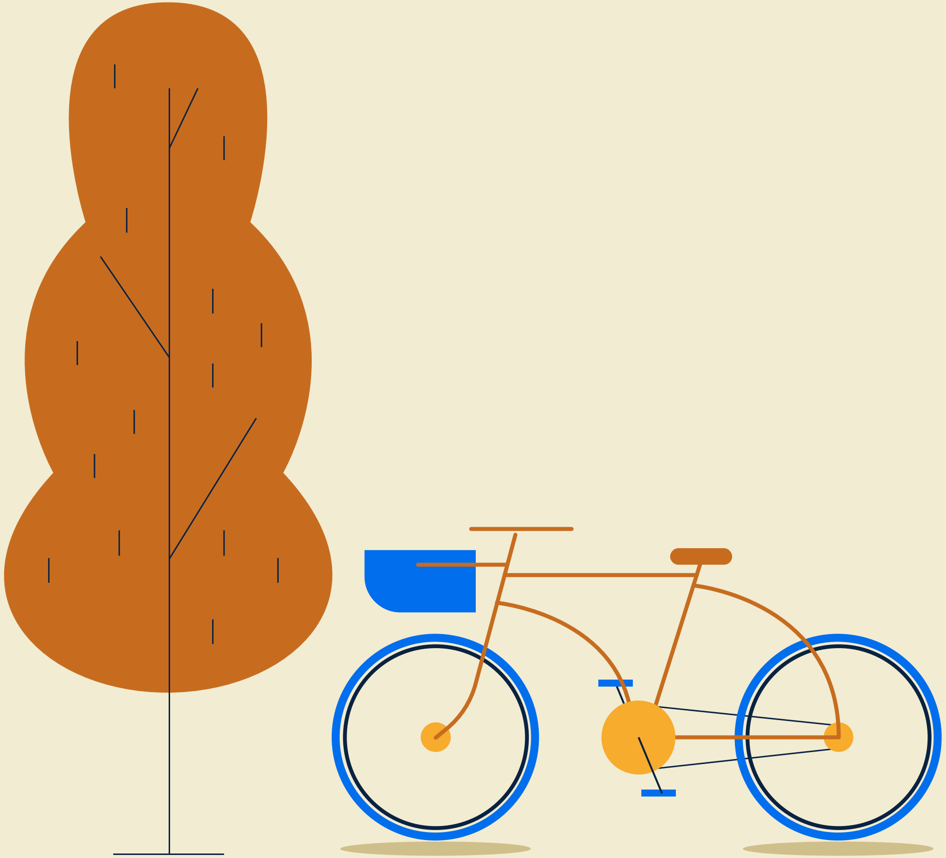


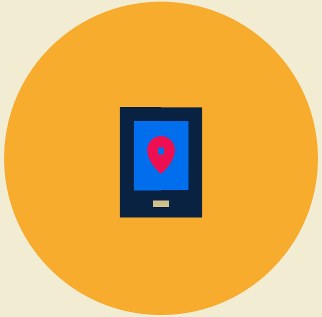
Deployment

You can enter a subtitle
here if you need it

Streamlit demo







Key Insights & Impact

Key Insights

Components maketh the bike

Sellers: list your bike components, ensuring to use keywords

Buyers: know what components you want/need, and search by keywords

Bike sizes 50, 52, 54 are listed frequently

Sellers: if you're selling a bike of this size make sure to have a USP / good price

Buyers: many choices if you are looking for bicycles of these sizes

Post quality

Sellers: aim to use up to 100 words and upload 5 images

Buyers: avoid posts which are too brief; it will be difficult to ascertain pricing

Recap

Problem Statement

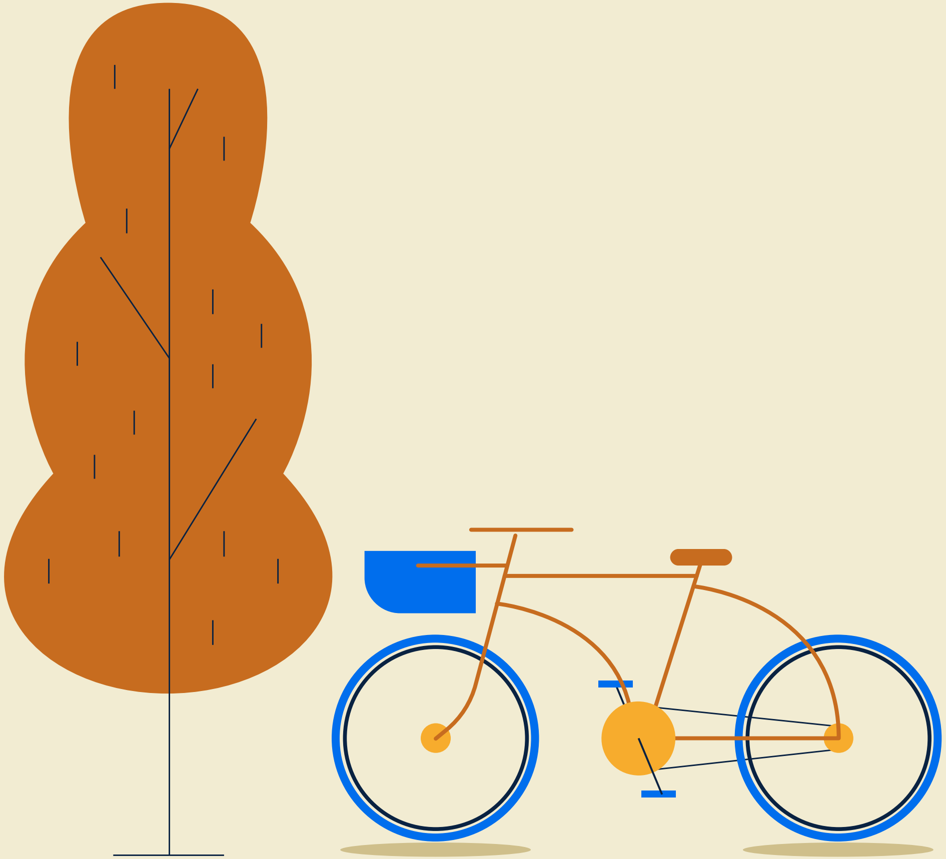
1. Carousell's filters are very generic
2. Numerous components affect price
3. Challenging to determine reasonable road bike price, especially as a novice

Proposed Solution

- Build a tool predicting road bike prices category from user-input values

Impact

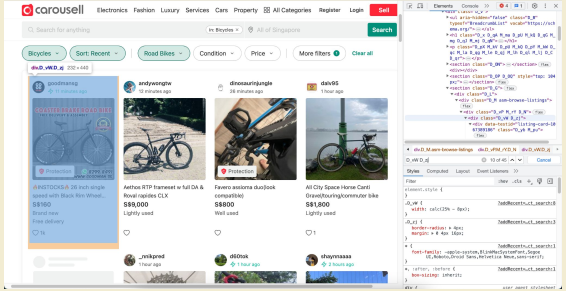
- Users are able to **save 10 – 15 min** of 'research' time
- Takes **less than 2 min** to input keywords and get a recommended price range



Challenges & Solutions

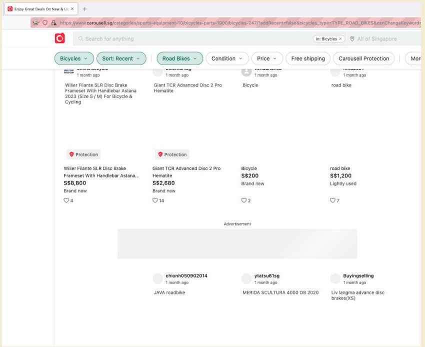
You can enter a subtitle here if you need it

Challenges & Solutions



Data scraping - html class changes every day

- Re-inspect soup and update code before scraping
- Use div-id where possible

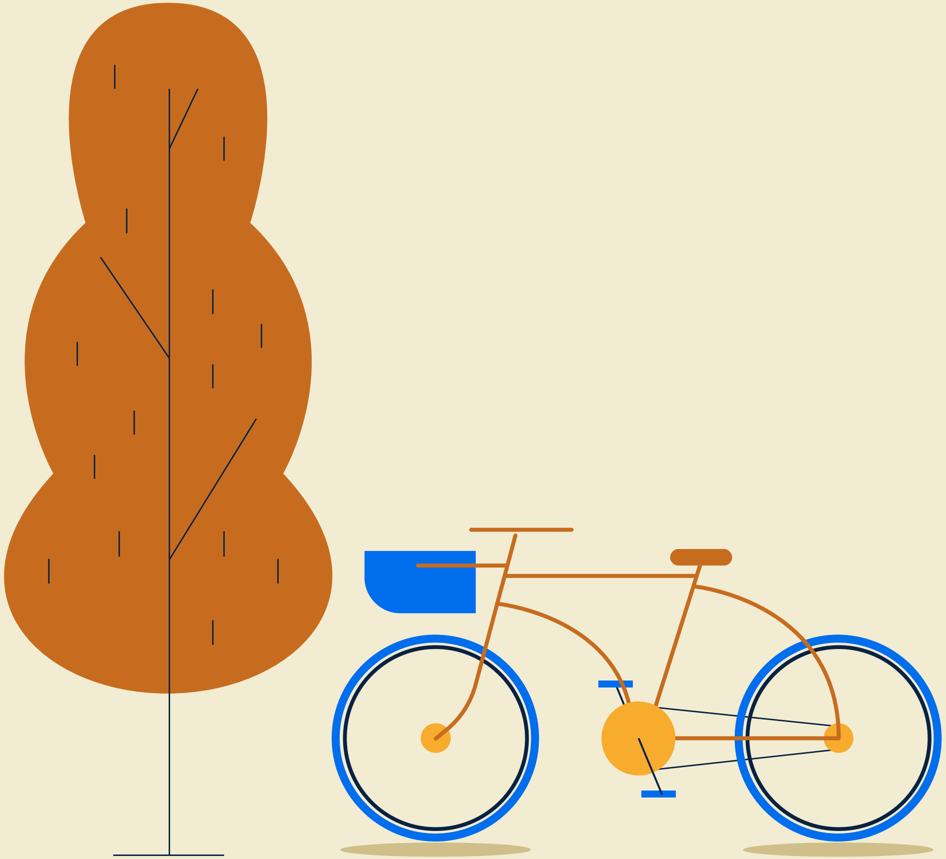


Limited to about 2,500 listings scraped per run

Scrape data every few days to get newer listings

Noise - incorrectly categorised items

Feature engineered a 'brands' dictionary containing bicycle brands



Future Work

You can enter a subtitle here if you need it

Future Work

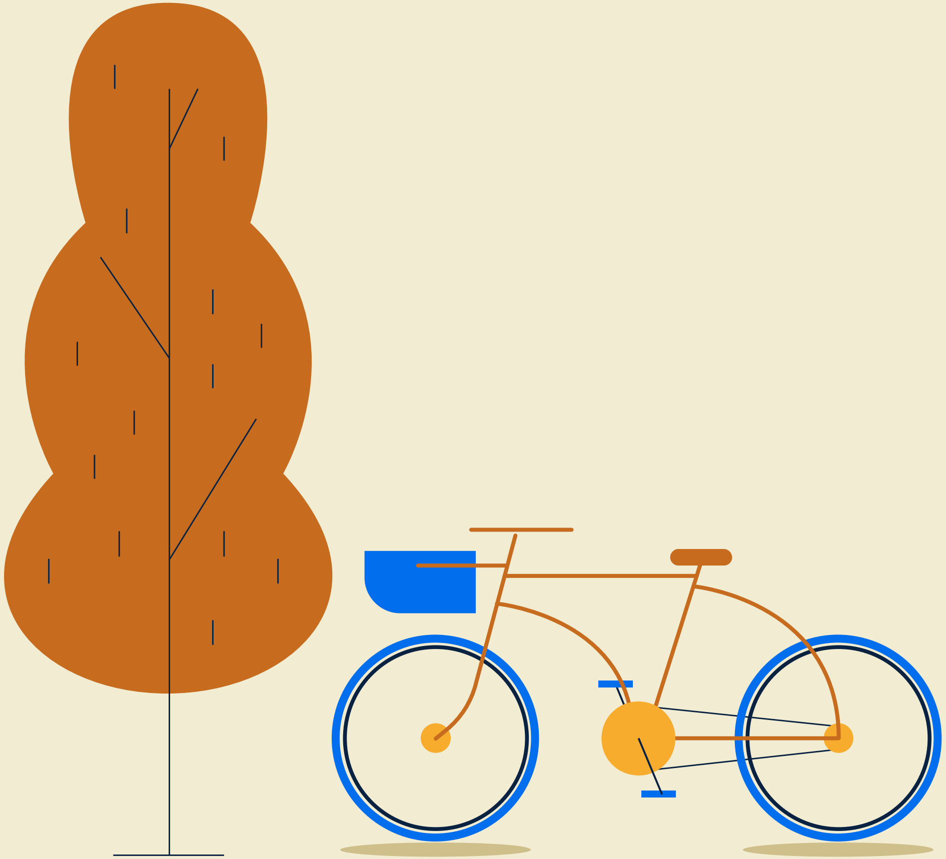
1. **Increase size** of **dataset** for Road Bikes, with **higher price ranges**

2. **Decrease price window** of **predicted class** for more actionability

3. Expand analysis to **other types of bicycles** as well (Mountain Bikes, Foldies, Hybrid)

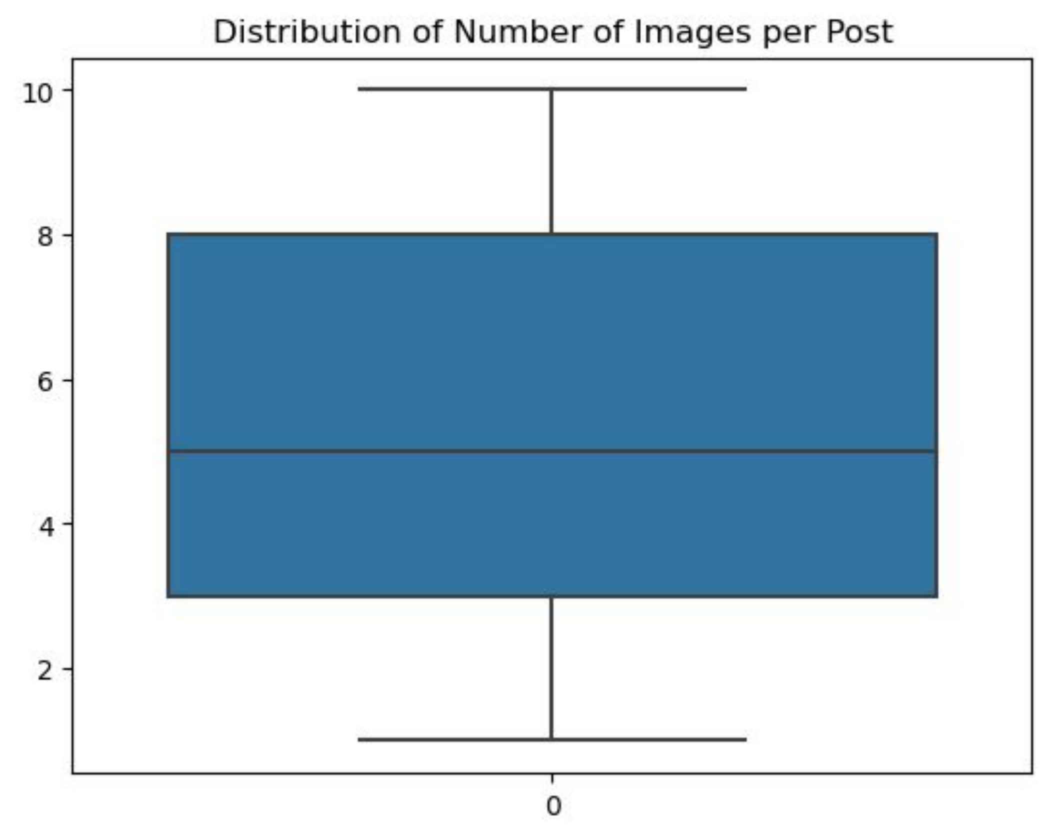
Thank you!

Time to buy / sell that bike on Carousell!



Back up slides

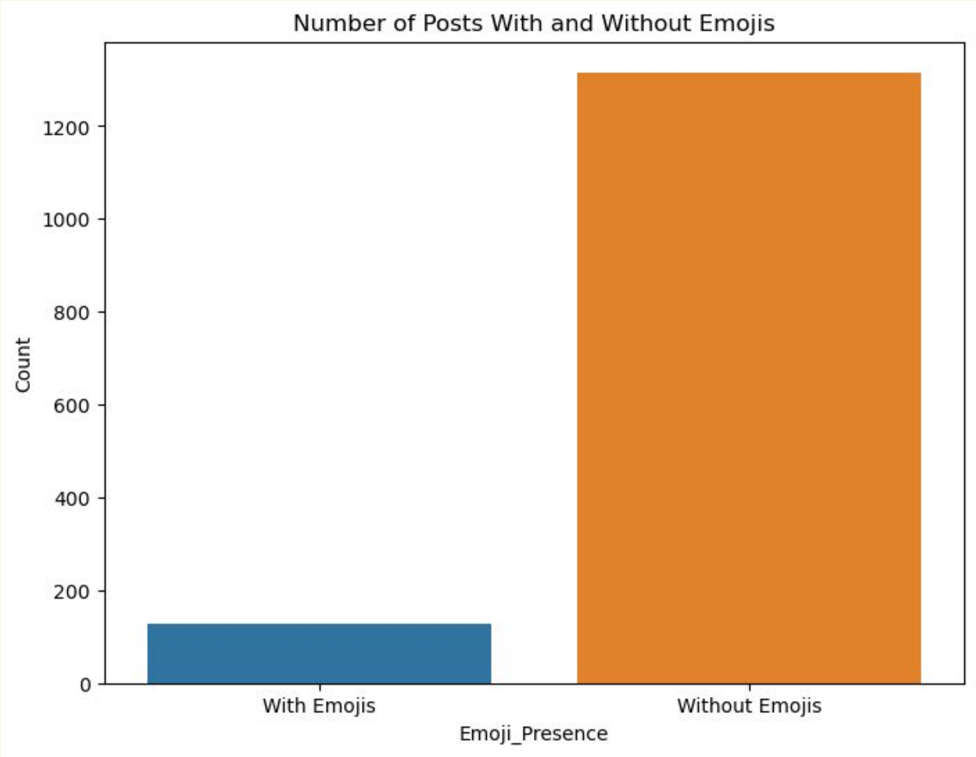
Number of Images



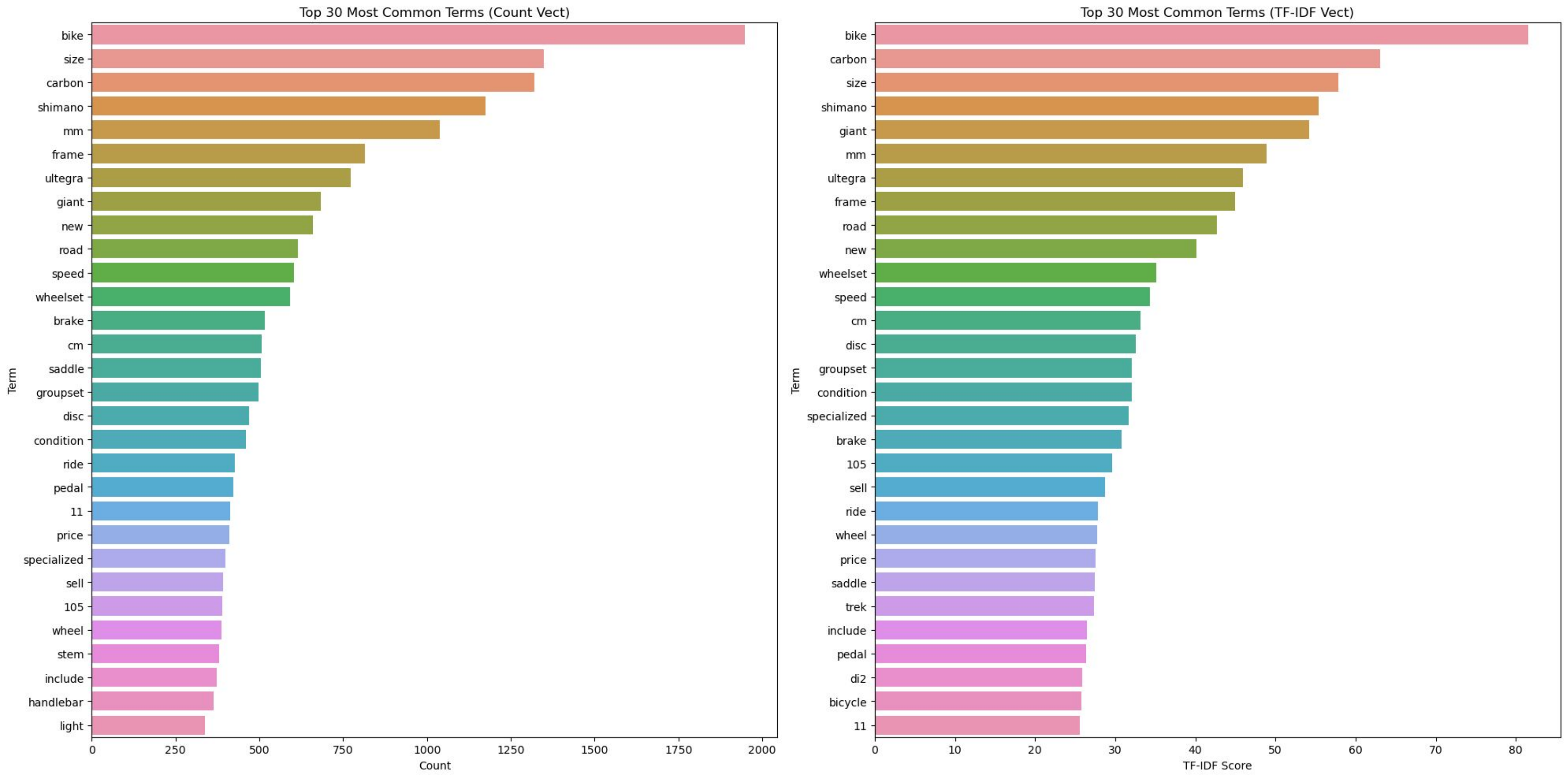
Features of scraped dataset

---	-------	--------------	-----
0	no_of_likes	2885 non-null	object
1	no_of_images	2885 non-null	object
2	title	2885 non-null	object
3	listing_price	2885 non-null	object
4	item_condition	2885 non-null	object
5	deal_method	2885 non-null	object
6	post_date	2885 non-null	object
7	category_type	2885 non-null	object
8	post_type	2885 non-null	object
9	condition_subtext	2884 non-null	object
10	listing_description	2885 non-null	object
11	mailing_option	2885 non-null	object
12	delivery_options	435 non-null	object
13	mail_speed	435 non-null	object
14	meetup_option	2885 non-null	object
15	meetup_location	2566 non-null	object
16	seller_id	2885 non-null	object
17	seller_join_date	2885 non-null	object
18	seller_response	2796 non-null	object
19	seller_verif	2885 non-null	object
20	verified_by_email	2189 non-null	float64
21	verified_by_facebook	2189 non-null	float64
22	verified_by_mobile	2189 non-null	float64
23	seller_stars_rating	2608 non-null	float64
24	reviews_of_seller	2608 non-null	object
25	url	2885 non-null	object
26	deal_location_lat	647 non-null	float64
27	deal_location_lon	647 non-null	float64
28	posts	2885 non-null	object

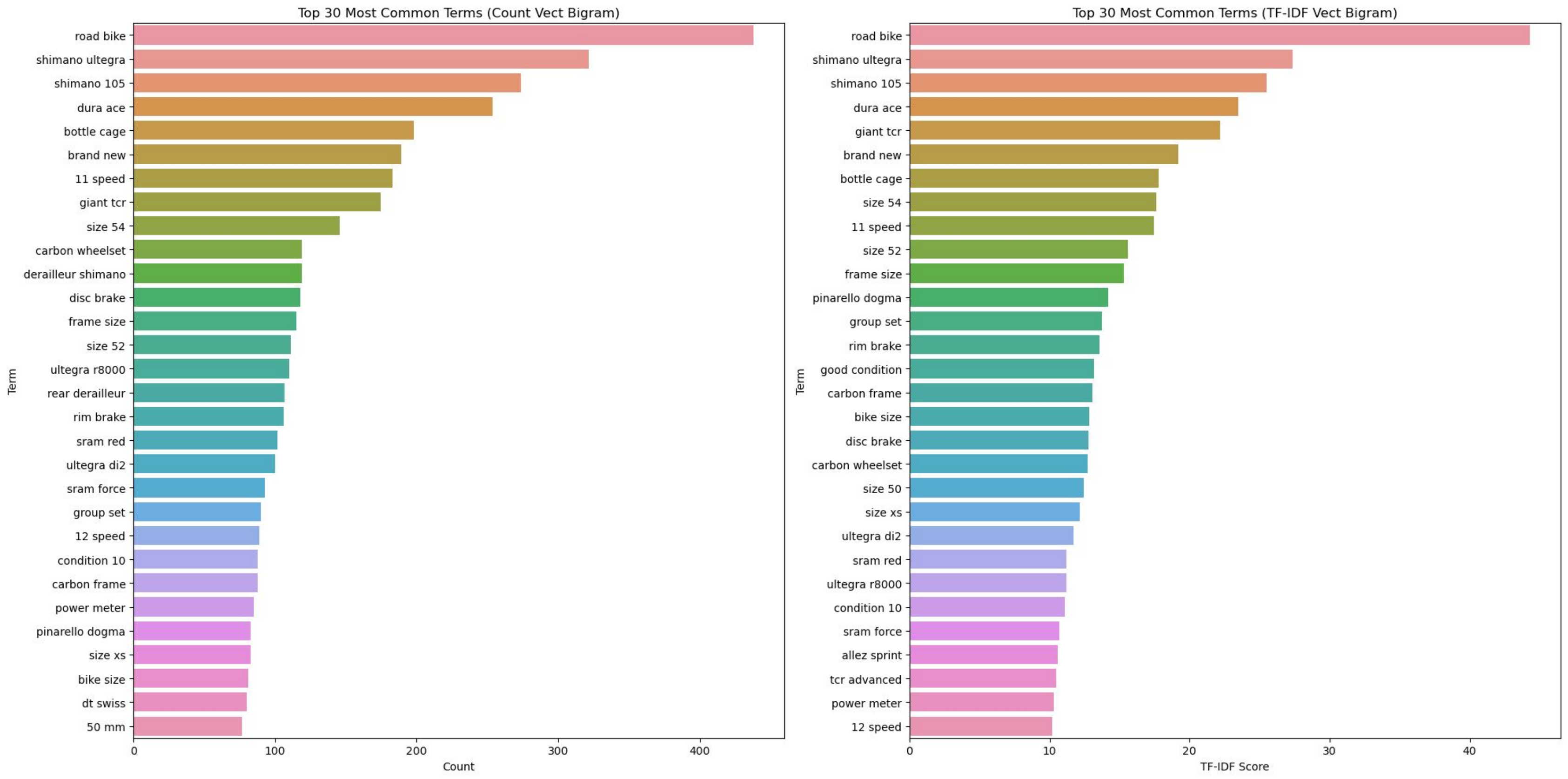
Emoji Use


Number of Posts With and Without Emojis
Count
0
200
400
600
800
1000
1200
With Emojis
Without Emojis
Emoji_Presence

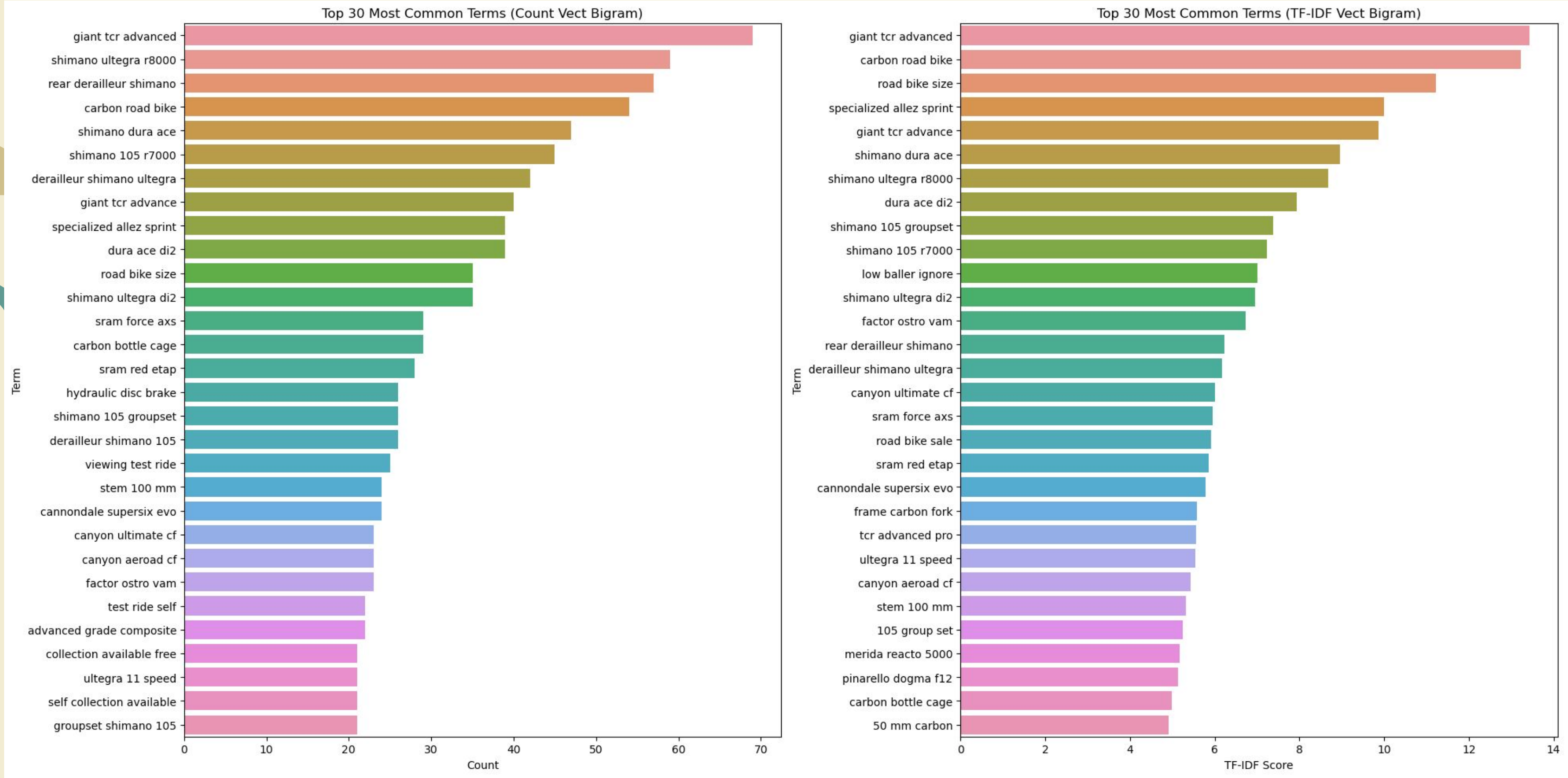
Top 30 Words - Count & TF-IDF Vectorizer



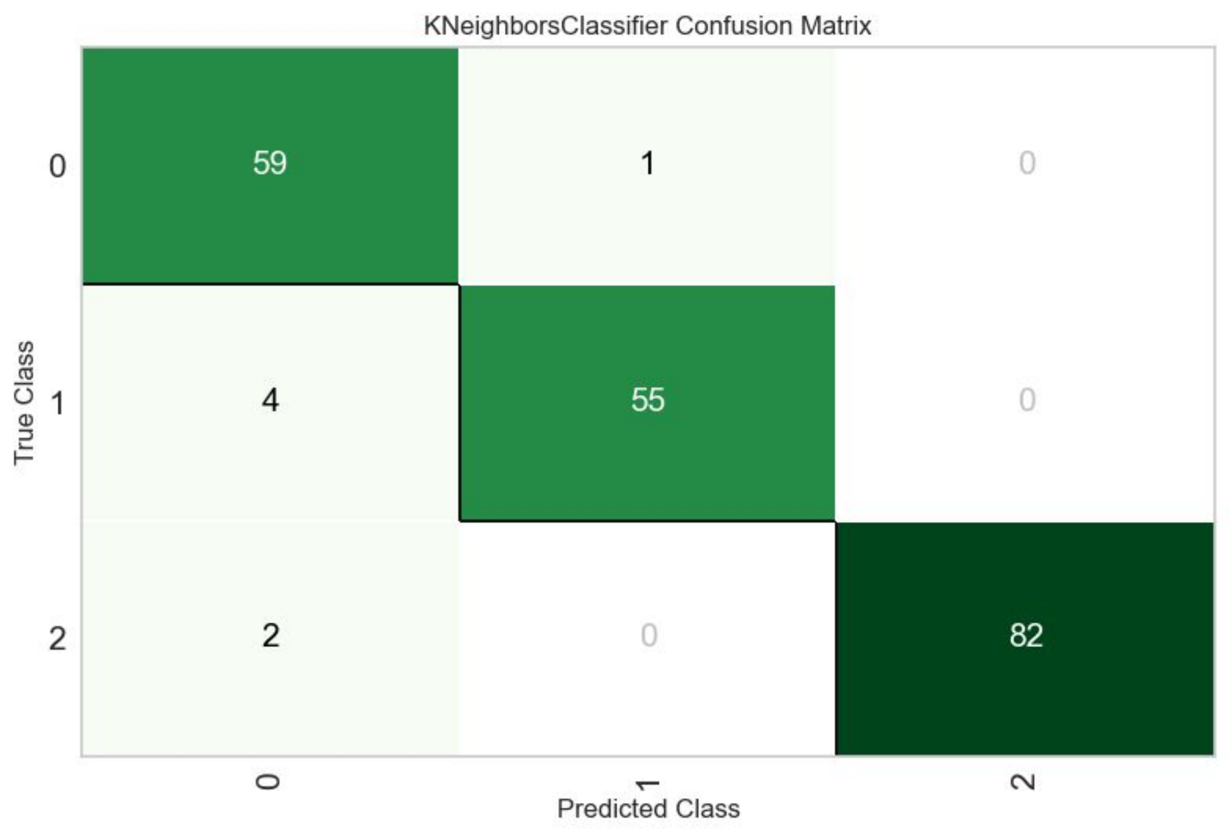
Top 30 Words (Bigram) - TF-IDF Vectorizer



Top 30 Words (Trigram) - TF-IDF Vectorizer

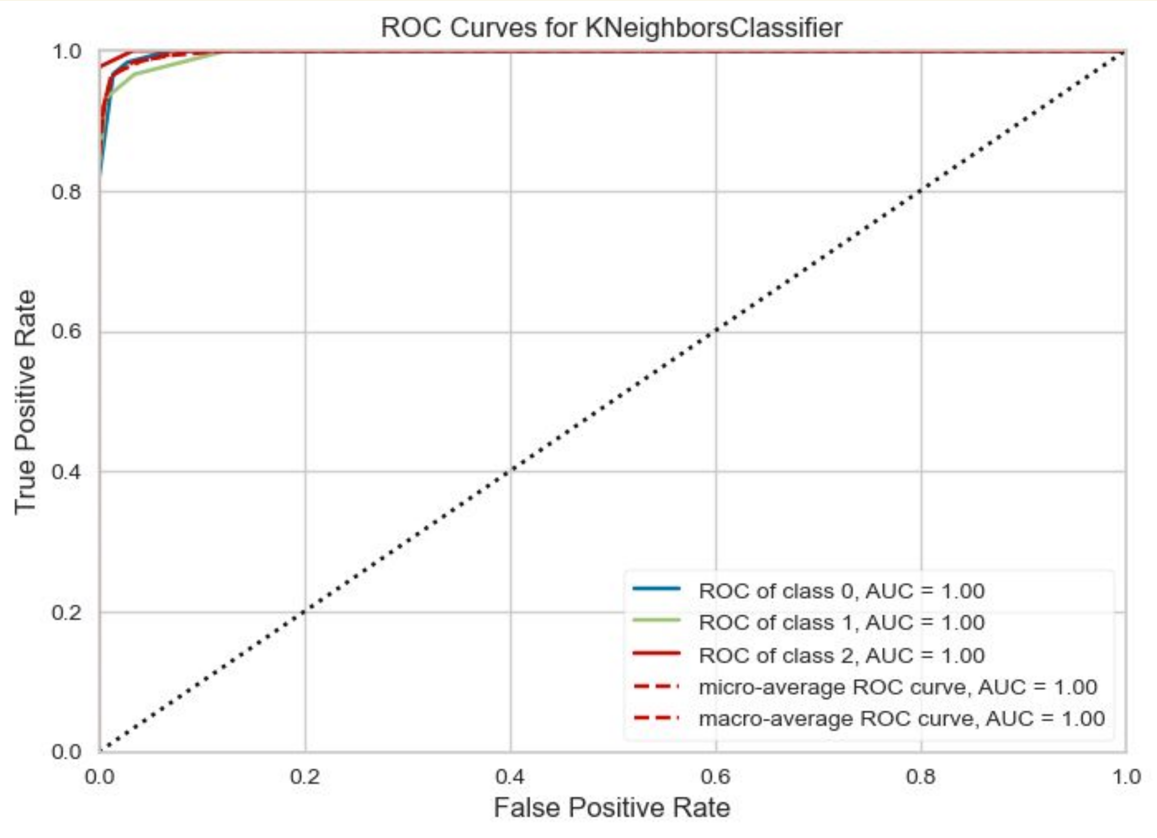


Confusion Matrix



- Class 1: Low (\$0 – \$600)
- Class 2: Med (\$601 – \$1200)
- Class 0: High (\$1201 – \$2000)

ROC AUC



- Class 1: Low ($0 – $600)
- Class 2: Med ($601 – $1200)
- Class 0: High ($1201 – $2000)

Brands Dictionary

```
brands = [
r'\bArgon(18)?\b',
r'\b(Giant|Qiant)\b',
r'\bScott\b',
r'\bBianchi\b',
r'\bTrek\b',
r'\bStandert\b',
r'\bOrbea\b',
r'\bFactor\b',
r'\bCanyon\b',
r'\b(Pinarello|Pina|Dogma)\b',
r'\bEddy Merckx\b',
r'\bColnago\b',
r'\b(Specialized|Specilized|Specialize|Specialised)\b',
r'\bS[-\s]?[Ww]ork\b',
r'\bMerida\b',
r'\bDean\b',
```