Offline evaluation of product search with model-based judgments

Alberto Castelo
Sr. Applied Machine Learning Engineer
MICES 2024

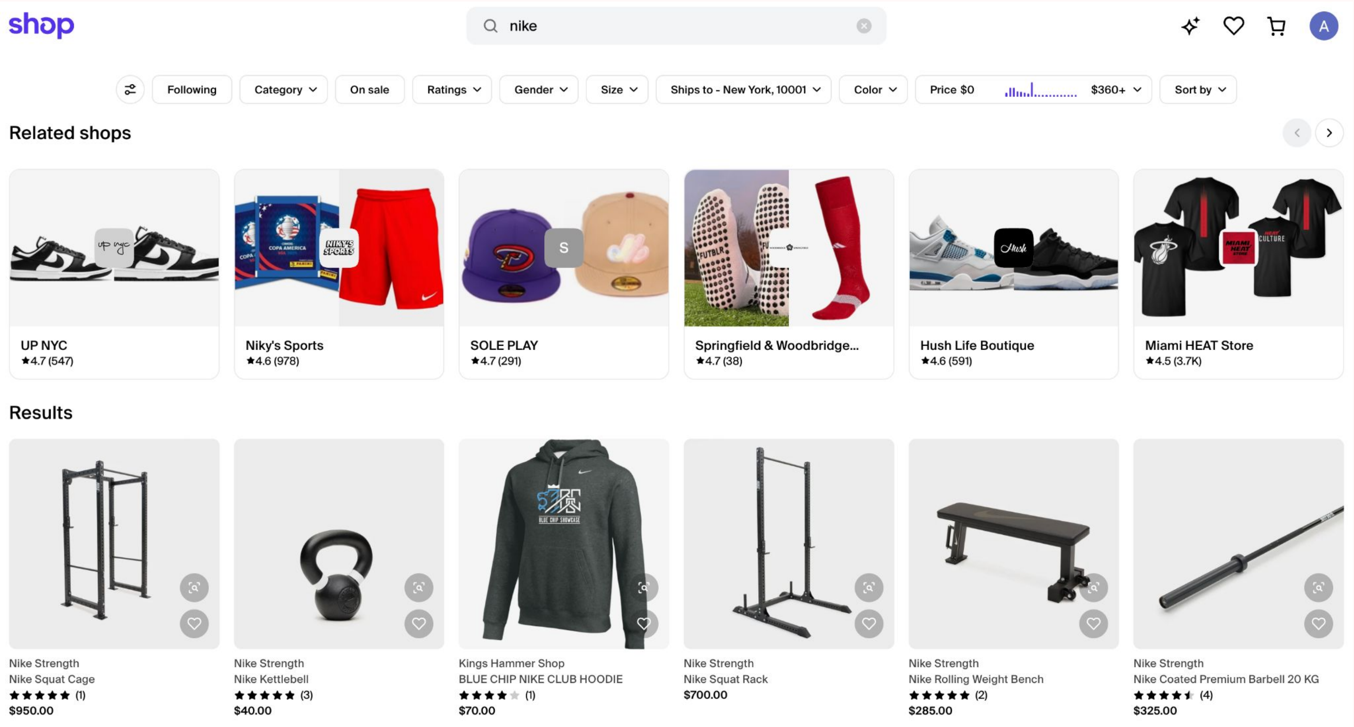
shopify

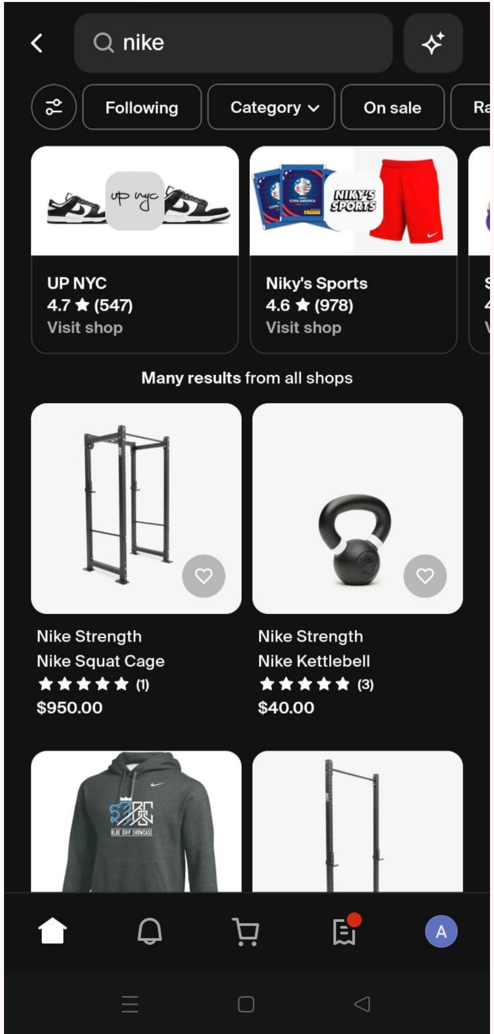
Agenda

- Search@Shopify
- Problem statement
- Solution: Model-based judgments
- Building a binary judgment model
- Conclusions

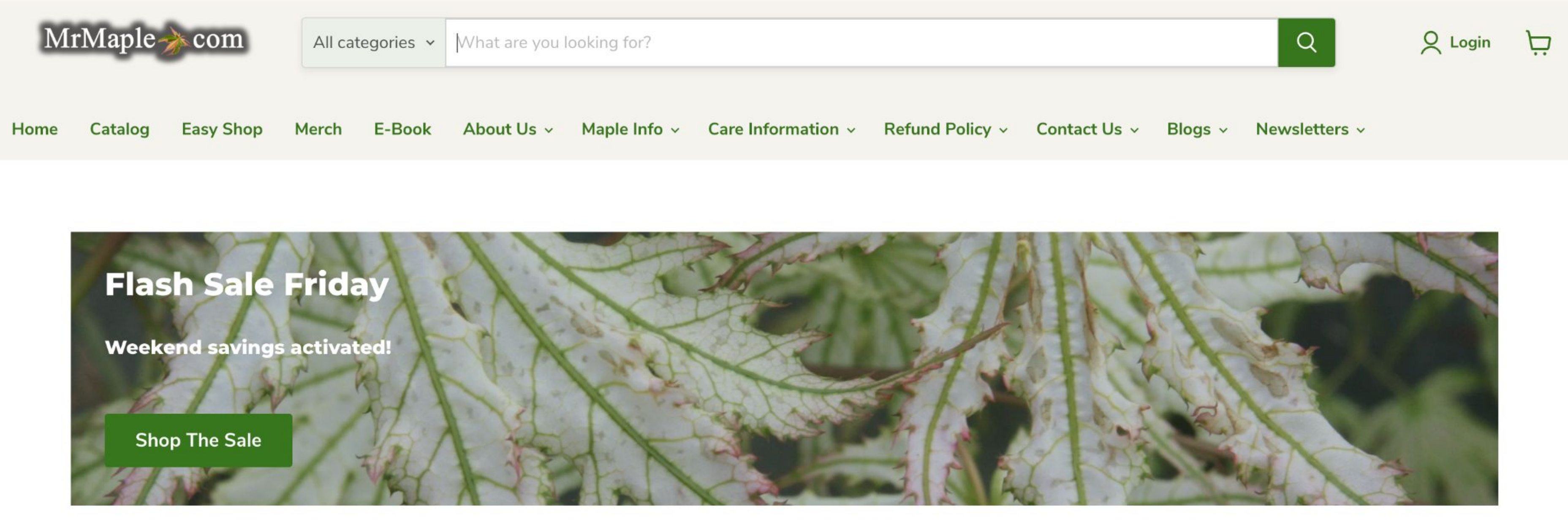
Search@Shopify

Shop

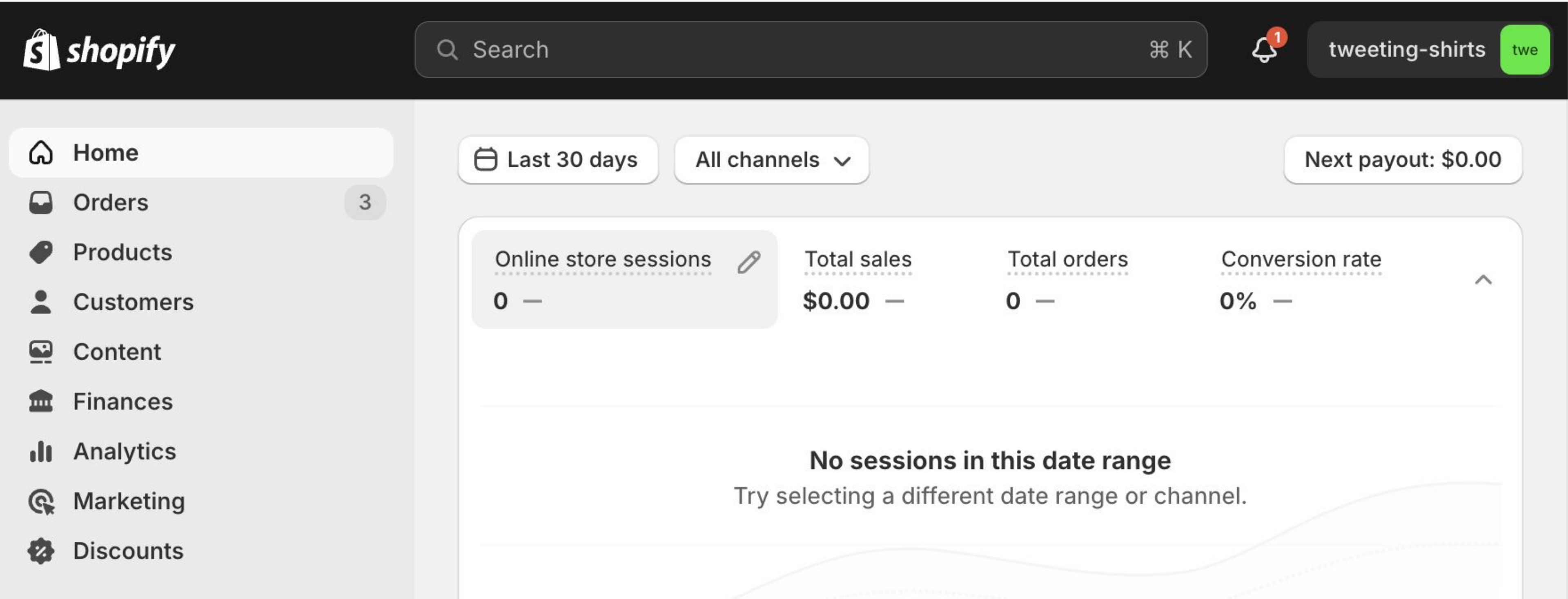

shop
nike
Following
Category
On sale
Ratings
Gender
Size
Ships to - New York, 10001
Color
Price $0
$360+
Sort by
Related shops
UP NYC
4.7 (547)
Niky's Sports
4.6 (978)
SOLE PLAY
4.7 (291)
Springfield & Woodbridge...
4.7 (38)
Hush Life Boutique
4.6 (591)
Miami HEAT Store
4.5 (3.7K)
Results
Nike Strength
Nike Squat Cage
(1)
$950.00
Nike Strength
Nike Kettlebell
(3)
$40.00
Kings Hammer Shop
BLUE CHIP NIKE CLUB HOODIE
(1)
$70.00
Nike Strength
Nike Squat Rack
$700.00
Nike Strength
Nike Rolling Weight Bench
(2)
$285.00
Nike Strength
Nike Coated Premium Barbell 20 KG
(4)
$325.00


nike
Following
Category
On sale
UP NYC
4.7 ★ (547)
Visit shop
Niky's Sports
4.6 ★ (978)
Visit shop
Many results from all shops
Nike Strength
Nike Squat Cage
(1)
$950.00
Nike Strength
Nike Kettlebell
(3)
$40.00

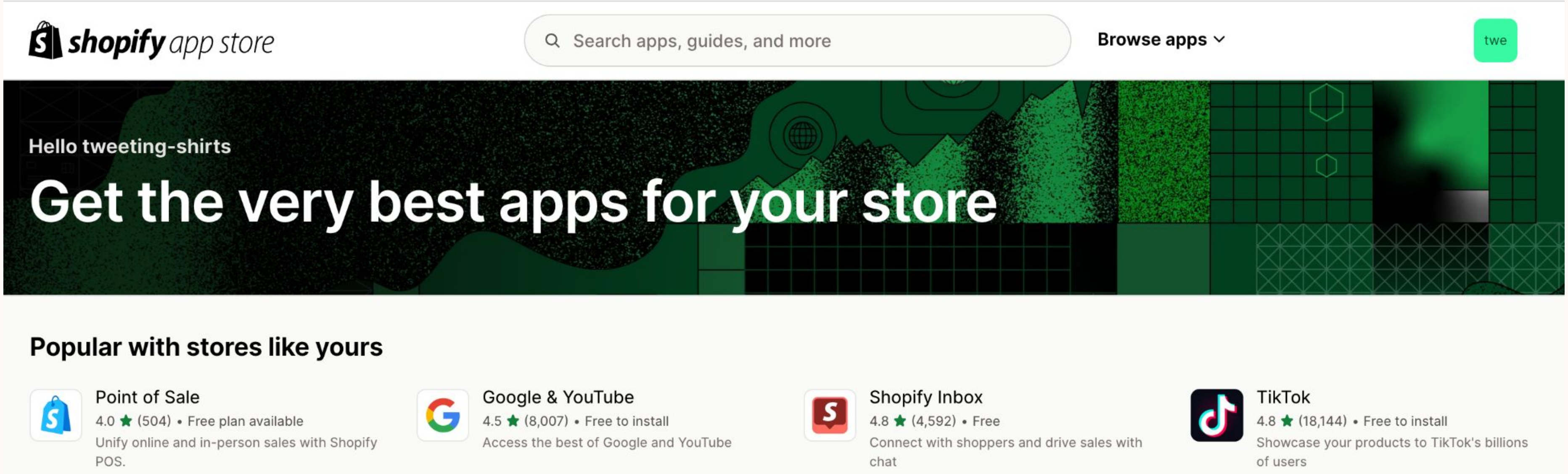
Storefront - MrMaple Example



Admin


shopify
Search
⌘ K
1
tweeting-shirts
twe
Home
Orders
3
Products
Customers
Content
Finances
Analytics
Marketing
Discounts
Last 30 days
All channels
Next payout: $0.00
Online store sessions
0 —
Total sales
$0.00 —
Total orders
0 —
Conversion rate
0% —
No sessions in this date range
Try selecting a different date range or channel.

App store


shopify app store
Search apps, guides, and more
Browse apps
twe
Hello tweeting-shirts
Get the very best apps for your store
Popular with stores like yours
Point of Sale
4.0 (504) • Free plan available
Unify online and in-person sales with Shopify POS.
Google & YouTube
4.5 (8,007) • Free to install
Access the best of Google and YouTube
Shopify Inbox
4.8 (4,592) • Free
Connect with shoppers and drive sales with chat
TikTok
4.8 (18,144) • Free to install
Showcase your products to TikTok's billions of users

Problem statement

Problem statement

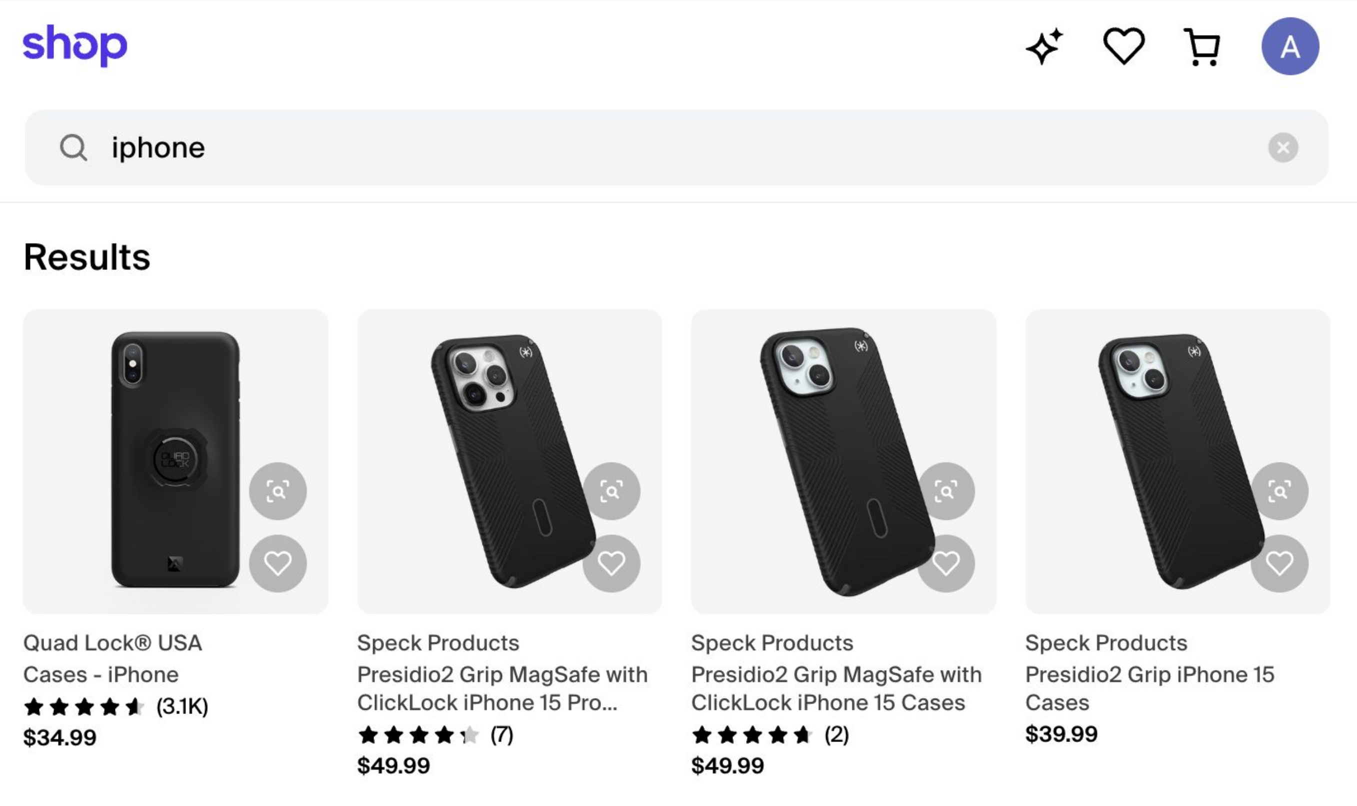
- We need offline evaluation
  - Release with confidence
  - Faster iteration cycles
- Implicit judgments approach is not good enough:
  - Judgment scarcity
  - Bias & noise
  - Not aligned with desired UX

Implicit judgments - CTR Scarcity

- Scarcity tends towards reinforcing old good products

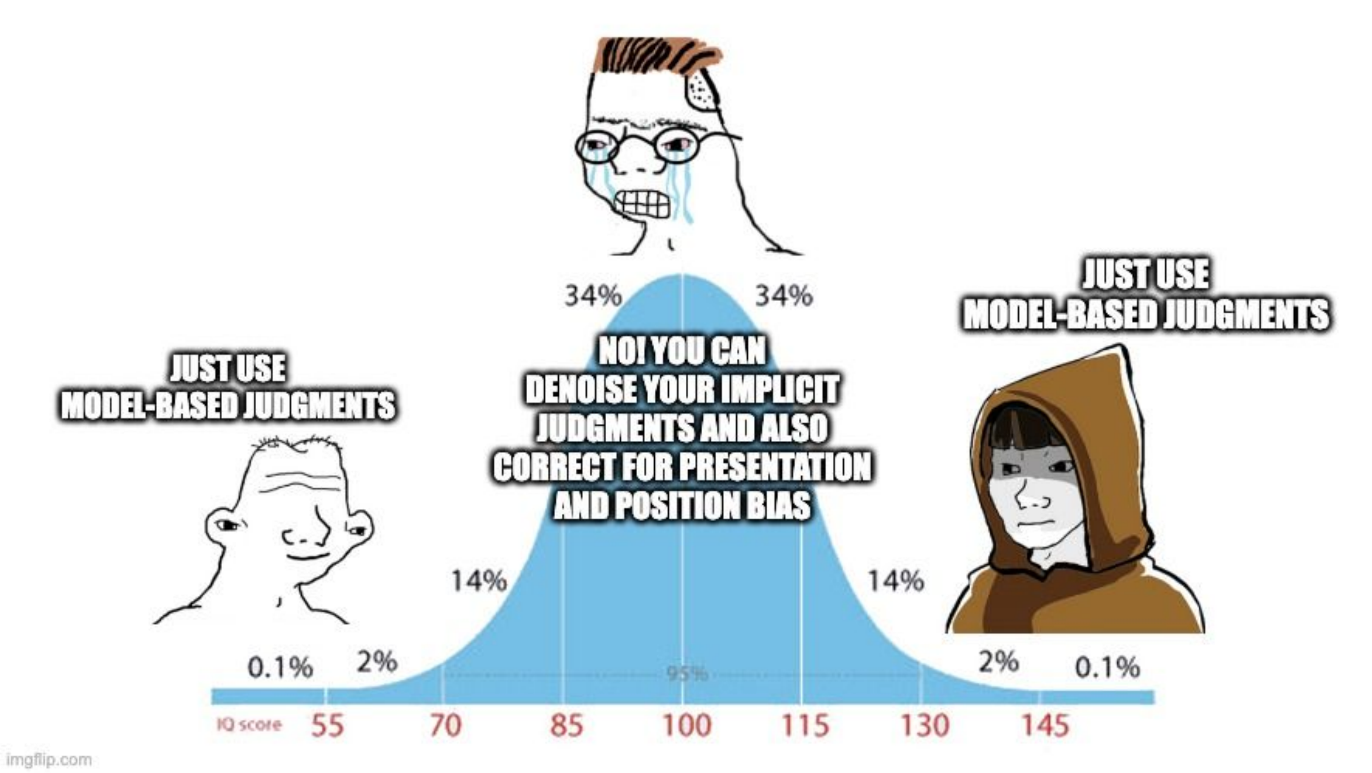
Product Id	Judgment
4634	0.23
2156	0.25
1234	0 or Mean
7891	0 or Mean
12945	0.12

Implicit judgments - UX impact



Solution: Model-based Judgments

Solution



JUST USE MODEL-BASED JUDGMENTS
NO! YOU CAN DENOISE YOUR IMPLICIT JUDGMENTS AND ALSO CORRECT FOR PRESENTATION AND POSITION BIAS
JUST USE MODEL-BASED JUDGMENTS
34%
34%
14%
14%
0.1%
2%
2%
0.1%
IQ score
55
70
85
100
115
130
145
imgflip.com

Tasks

- Binary relevance

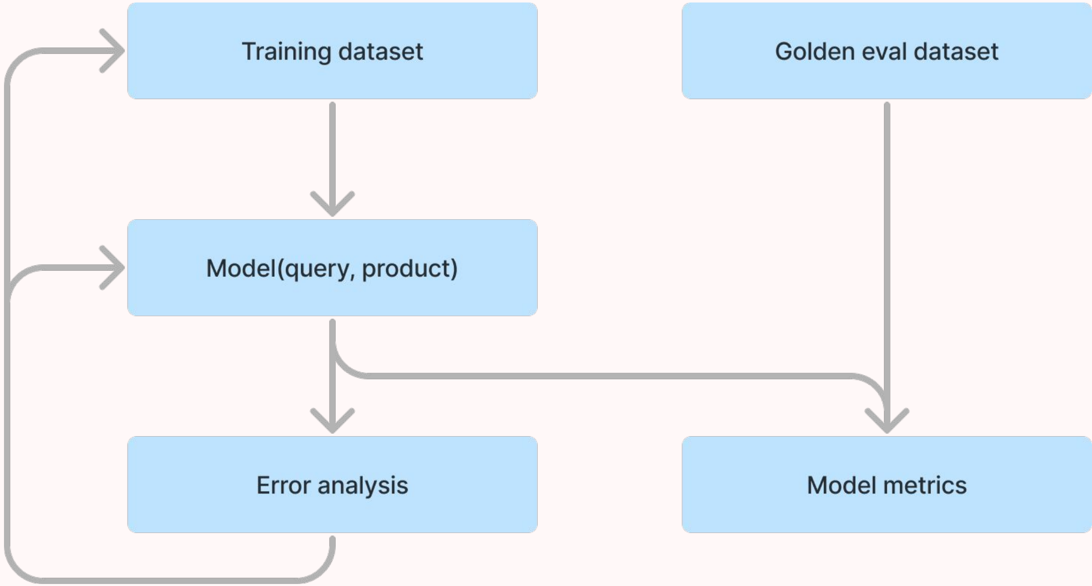
- Ranking

Ranking vs. relevance

Building a binary judgment model

Model flywheel

- Build your golden eval dataset
  - **Manual annotation**
  - Representative of business objectives
- Set a training loop:
  - Train dataset
  - Model
  - Error analysis



Building a training dataset

- Leverage public datasets
    - [ESCI (Amazon search)]()
- Synthetic data
    - Distilling from GPT4 labels
    - Real data + synthetic
- Manual annotation
    - Clear guidelines

```
SYSTEM_PROMPT = """
You are an expert on ecommerce working on a product search engine. Your job is to:
1. Understand the product provided.
2. Generate an inexact query that:
 * Could happening during a ecommerce product search session.
 * where the product does not satisfy all the conditions or characteristics of the query.
 * has a similar structure than the exact query provided.
3. Answer in JSON format.

Use these definitions:
* Exact query: a query that matches the product characteristics.
* Inexact query: query that do not match the product (even if the product is close to match).

Examples:
1. Product: an Iphone 14 phone.
 a. Exact query: "iphone 14",
 b. Inexact query: "iphone 15"
2. Product: jordan's sneakers
 a. Exact queries: "nike"
 b. Inexact queries: "adidas"
3. Product: A bed for dogs
 a. Exact queries: "dog bed"
 b. Inexact queries: "cat bed"
"""


USER_PROMPT = """
Take a look at the product:
{product_dict}

In JSON format, generate a close yet inexact query with a similar structure as the exact query "{query_string}":
"""
```

Binary relevance model-based judgments

- LLM as a judge
- Classifier as a judge

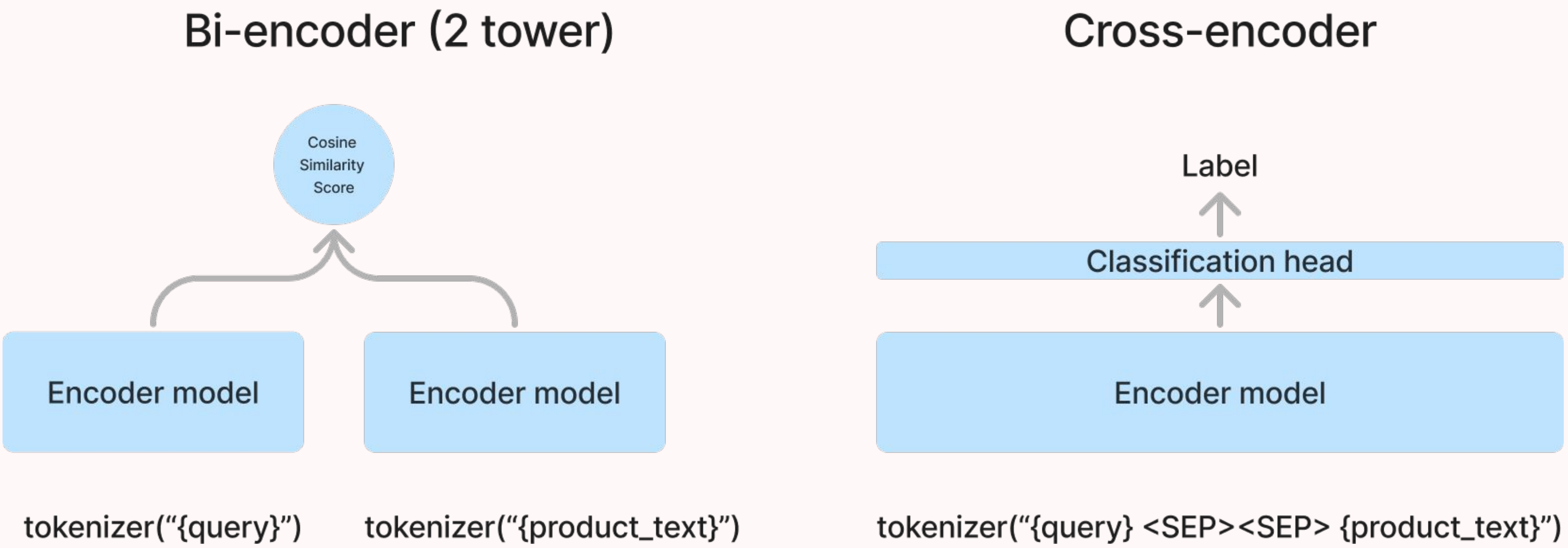
LLM-Llama 3

- 8B-Instruct
  - A couple of few-shot examples.
- 8B (un-instruct) LoRA finetune
  - O(100) samples

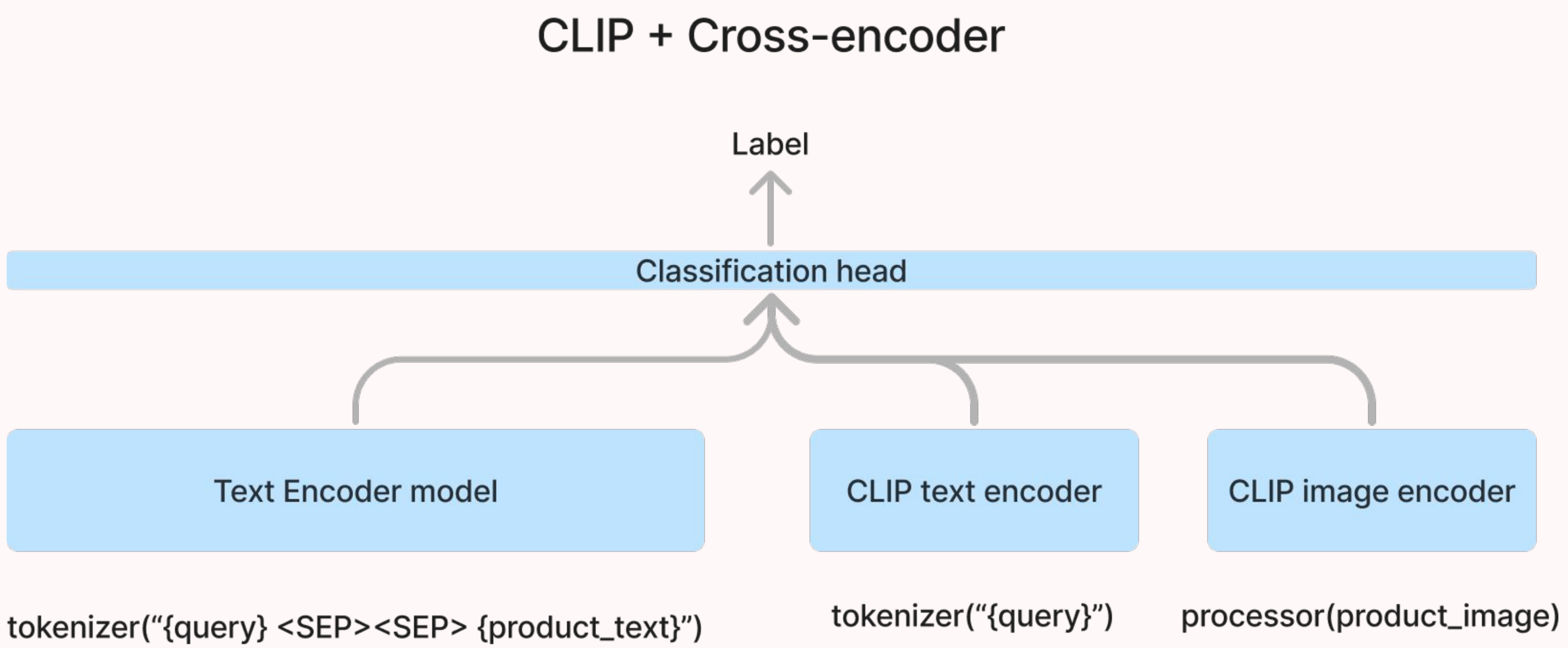
```
You are a relevance engineer working on ecommerce search.
Your job is to tag a pair of query and product as 'E' (exact match) or 'I'
(inexact match) depending on whether the product fully satisfies/matches the
intent of the query
### Query:
nike
### Product:
# title
jordan's
# vendor
nike
# shop name
shoes retailer
# product category
shoes
# product attributes
black
# image description
A man wearing a Nike shoe
# price
120
# description
Great Jordan's

### Response:
```

Cross-encoders



CLIP + crossencoder

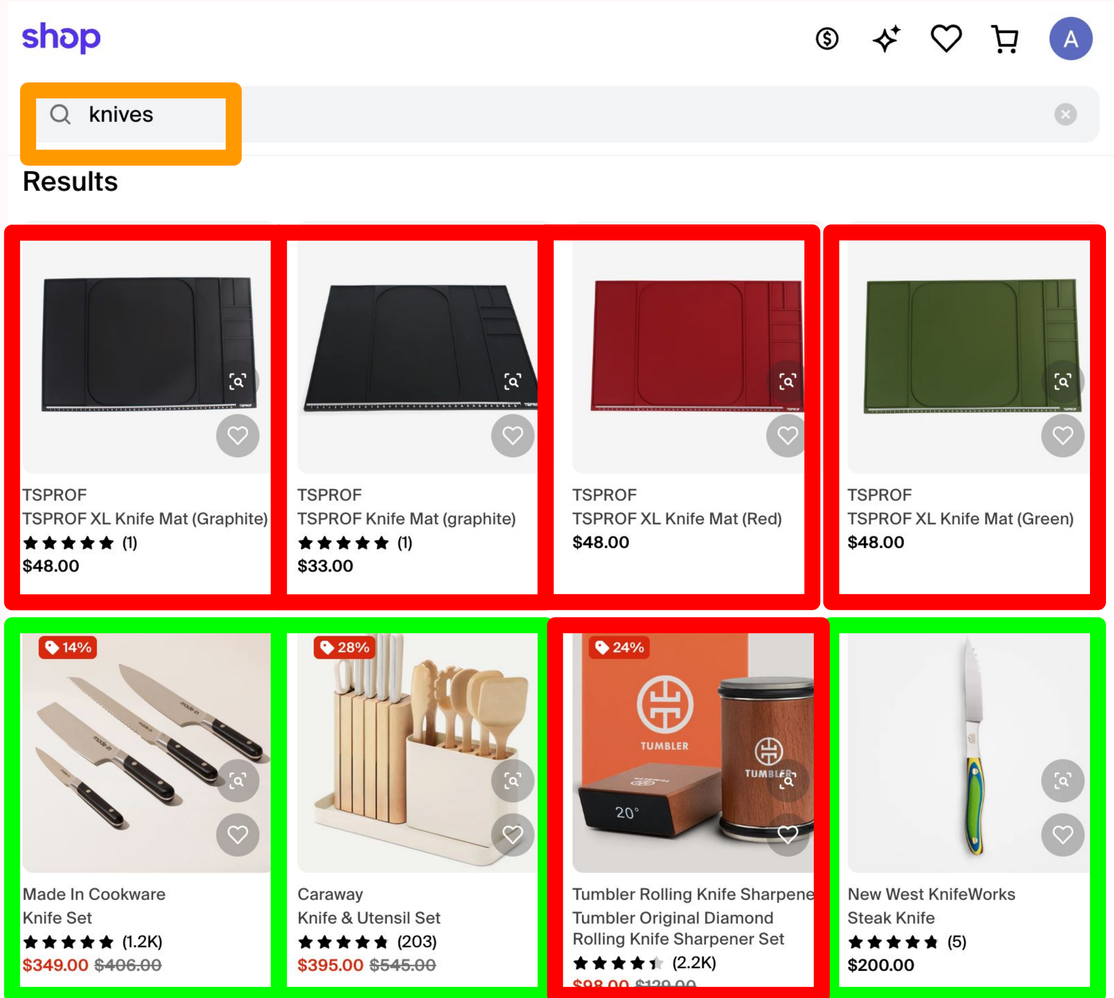


Performance Summary

Model	Data required	Inference Cost (1M pairs)	Classification Performance
Llama 3 8B-Instruct	O(1)	O($100)	Low
Llama 3 8B LoRA FT	O(100)	O($100)	High
Crossencoder	O(1k)	O($10)	High

SERP-Precision

- **Precision@1:** 0/1
- **Precision@4:** 0/4
- **Precision@8:** 3/8



Winner/Loser comparison

- Comparing 2 search strategies: Precision@1

Query: ebike

2024_04_10_elasticsearch_multi_match

2024_04_11_elasticsearch_basic

Position: "0"
Title: "Dubbel Ebike"
Shop Name: "Blix Electric Bikes"
Predicted Category: "Sporting Goods > Outdoor Recreation > Cycling > Bicycles"
Relevance score: "0.98"

Position: "0"
Title: "eBike Cranks"
Shop Name: "20TwentyStore"
Predicted Category: "Sporting Goods > Outdoor Recreation > Cycling > Bicycle Parts > Bicycle Drivetrain Parts > Bicycle Cranks"
Relevance score: "0.17"

Conclusions

Conclusions

- Model-based judgments for the win
- Specialized models >> LLM
  - For now...
- Data is key

Thank you!

acaste10

alberto-castelo-becerra

shopify

Offline evaluation of product search with model-based judgments

Alberto Castelo
Sr. Applied Machine Learning Engineer
MICES 2024

shopify