Problem Statement

Design a Machine Learning for QnA Ranking

- FB Community
- Quora
- Reddit

Objectives :

1. Given a Q and answer $\{A_1, A_2 \ldots A_n\}$ predict a score, such that when the answers are sorted by score, the best takes the first place.

In Practice → Maximize the relevance metric. (NDCG)

click / engagement rate.

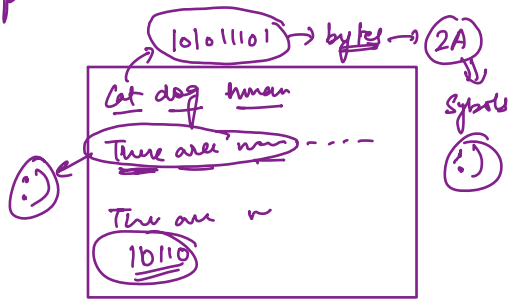
2. Scale → Mn of QnA

Latency < 100 ms

Data Collection & feature engineering layer.

- Historical data on QnA & platform logs

- upvotes, likes, timestamp, edit history, accepted
- User's reputation (warning / bans / removed content)
- Clicks, impressions
- Sub-replies & main replies.
- grading (0-4)
- dwell time



Bits → Bytes → Common sequences → Symbol

Feature Engineering

- Text processing (Tokenisation)
  - lowercase
  - stripped
- Contextual embedding
  - ↳ (transformers Sentence-BERT)
- Metadata

Modeling Strategy

https://www.tensorflow.org/ranking/tutorials/tfr_bert#:~:text=non,the%20list%20is%20either%20truncated

Baseline model

- Simple Supervised model
  - Votes
  - impression
  - dwell

  → logistic Reg, XGBoost.

https://medium.com/acing-ai/quora-ml-platform-for-ranking-answers-baaf00cc97e8

- SBERT

https://sbert.net/examples/sentence_transformer/applications/retrieve_rerank/README.html

LLMs → Prompt (———)

Score
- Explain → { }

https://arxiv.org/html/2411.00142v1#:~:text=Most%20recently%2C%20utilizing%20Large%20Language,approaches%20utilize%20an%20LLM%20either

Evaluation

- Baseline
- F1 score, AUC ...
- NDCG, MRR

Given a query, and a list of answers, the objective of the ranking model is to rank the answers with optimal rank related metrics, such as NDCG

Target / Amazon / Flipkart

Promotions & Campaign

20 years

- P_1
- P_2
- ⋮
- P_{1000}

→ Items $[I_1, I_2, I_3 \dots]$

Promotion → Predict Units

Strategy design

Task → 5 mins

Break → 5 mins

Put answers on chat

Promotions → June & July 2025

⇓

June & July 2024 | 2023 | 2022 ...

Step 1

Most similar promotions from the past year.

Hadoop

↳ HQL (~SQL)

Historical data.

Promotions	Items	Unit Sales	Timestamp. (Fiscal weekend)
P_A	I_1	100	2020-08-01
P_A	I_2	200	2020-08-07
P_A	I_3		2020-08-14
P_B	I_x		
P_n	I_y		

P_C ← I_1

P_{XA} ⇒ I_{XA} [I_1, I_2, I_5]

P_{XA} → P_A ⇒ Metric

nesting

Similarity

(1) Item Similarity → $\frac{A \cap B}{B}$ ⇒ Cont (Common Items / Cont (Items in the current promo)

P_A $[I_1, I_2, I_3, I_4]$

P_{XA} $(I_1, I_2, I_5]$

$\Rightarrow \frac{2}{3}$ → Jaccard

(2) Brand Jaccard

(3) Promo length → 20 days, 10 days (Past promotion)

(4) Location Jaccard

$\Rightarrow \frac{200}{300} \Rightarrow 0.66$ $\Rightarrow$ (2)

Current Promo	Matched Promo	Item Jaccard	Brand Jaccard	Promo length	Location Jaccard	Weighted Score
P_{XA}	P_A	0.66	0.78	2	0.66	
P_{XA}	P_B	—	—	—	—	
P_{XA}	P_C	—	—	—	—	
P_{XA}	P_D	—	—	—	—	
⋮						

$P_{XA} \rightarrow P_A$ $(a \times 0.66 + b \times 0.78 + c \times 2 + d \times 0.66)$

Max(weighted score)

[End of story → Most Similar promotion]

Predict the Sales of promotional Units for the current promotion

1) Target variable → Promotional Units from past promotions.

2) Macro metrics → Count of Promotional Campaigns Once item is used in.

3) Trend based features → Item wise Sales in last Quarter, last 6 months.

Step 1 — Already given

Current promo ⇔	Matched promo	Items	Fx	Last Qtr Sales	Last 6 month Sales	Time stamp	Seasonal Multiplier	Item Jacard
P_{xA}	P_A	I_1	30					
~~P_{xA}~~	P_A	I_2	22					

Brand	Location	Promotion Length	Units
Jacard	Jacard		

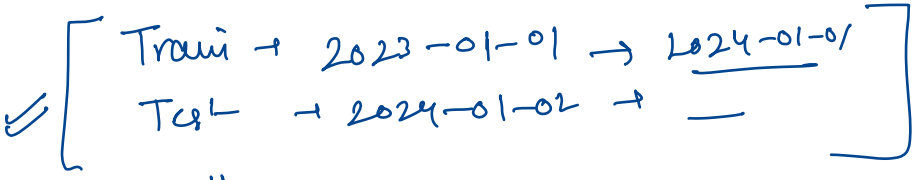
↓

Treatment — Null / Outliers / Negatives

↓

Train Test Split

Time based



Model

XGBoost →

Random Forest

↓

Evaluation | Hyperparameter tuning

↓

Post-Model Analysis → SHAPley Values