[AI-related FYP] AI for Personalized Content Recommendations

Project ID: 2

Supervisor: Prof Ruohan Zhan

CHENG, Yan Hei

WAN, Nga Chi

CHAN, Tony Yuen Yeung

Department of Industrial Engineering and Decision Analytics

The Hong Kong University of Science & Technology

April 2025

Abstract

This report develops a multi-stage content-recommendation system using the Yelp dataset to tackle data sparsity, new-user cold-start, and long-tail biases. We apply baseline, collaborative-filtering, and deep-learning (DSSM) retrieval models, followed by a Deep FM ranking stage. Evaluation combines simulation metrics, diversity and bias analyses, and GPT-based qualitative assessment. Results show high recall from collaborative filtering, enhanced diversity via DSSM, and effective ranking by Deep FM. We also introduce recent-popular and category-based fallbacks for sparse users. Overall, the system balances precision and diversity, offering a robust, real-world recommendation solution.

Project Team

Name	Roles	Contribution (Project)	Contribution (Write-up)
CHENG, Yan Hei	Leader, Model Optimization, Data Visualization, Video	33%	33%
WAN, Nga Chi	Model Optimization, Data Processing, Video	33%	33%
CHAN, Tony Yuen Yeung	System Design & Integration, Frontend, Report	33%	33%

Preface

This Final Year Project report presents the design, implementation, and evaluation of a multi-stage content-recommendation system built on the publicly available Yelp dataset. Over the academic year, our team investigated key real-world challenges—data sparsity, new-user cold-start, and long-tail biases—and devised a framework that integrates classical collaborative-filtering techniques with deep-learning architectures. Our work encompasses data preparation, model development, and both quantitative and qualitative evaluations to deliver a balanced solution that enhances user experience and broadens business exposure.

Acknowledgments

We gratefully acknowledge the guidance and support of our supervisor, Prof. Ruohan Zhan, whose expertise and thoughtful feedback were invaluable throughout this project. We also thank the Department of Industrial Engineering and Decision Analytics at HKUST for providing the computational resources and collaborative environment that made this work possible.

Table of Contents

1. Introduction

This report describes the development of a content recommendation system built on the Yelp dataset. By addressing challenges such as data sparsity, cold-start scenarios, and long-tail item biases, the system aims to deliver personalized user experiences and improve engagement.

1.1 Motivation

In Hong Kong's dynamic market, recommendation systems enhance the visibility of local businesses—restaurants, retail outlets, and service providers—by tailoring suggestions to a population characterized by high internet penetration and diverse consumer preferences. These systems empower niche enterprises to compete effectively within a densely populated landscape.

Globally, recommendation engines are central to platforms such as Amazon, Netflix, and Google. For example, Amazon's engine contributes about 35 % of its revenue [1]. Personalized recommendations boost sales, increase user retention, and enhance satisfaction across digital ecosystems. More broadly, such systems facilitate online discovery, promote inclusivity, and foster meaningful connections between users and content.

1.2 Industry and Research Landscape

The global recommendation-engine market is projected to grow from USD 9.15 billion in 2025 to USD 38.18 billion by 2030 (CAGR: 33.06 %) [2]. Industry leaders—Google, AWS, and Microsoft—are deploying AI-driven solutions that leverage generative models to deliver more nuanced personalization.

In academia, work continues core challenges such as data sparsity and cold-start. Aggarwal et al. (2022) propose clustering and embedding techniques to mitigate these issues [3], while recent reinforcement-learning approaches have demonstrated improved scalability and accuracy despite limited interaction data [4].

This project implements a multi-stage recommendation framework comprising:

- Baseline models (random, uniform, popular)
- Collaborative filtering (user-based and item-based)
- Deep-learning architectures (Deep Structured Semantic Models and DeepFM)

The evaluation strategy adopts a two-tier retrieval approach—prediction and production simulation—augmented by diversity analysis, ranking metrics (MRR, NDCG), GPT-based qualitative feedback, and popularity-bias assessments.

The report is structured as follows: Section 2 presents background and problem context; Section 3 outlines system design and methodology; Section 4 details evaluation results and analysis; Section 5 summarizes key findings, discusses limitations, and proposes future work; and Section 6 contains appendices.

2. Background and Problem Context

2.1 Introduction to Yelp Dataset

The Yelp dataset, employed as the primary data source in this study, comprises a rich collection of user-generated reviews and ratings [5]. Yelp is a platform that enables users to discover, review, and recommend local businesses—spanning restaurants, retail outlets, service providers, and entertainment venues—across major metropolitan areas in the USA and Canada. Initially released for the annual Yelp Challenge, this dataset has since become a standard benchmark in academic research on recommendation systems.

Unlike high-engagement platforms such as Netflix or Instagram, where users interact continuously with content (e.g., streaming videos or browsing social feeds), Yelp's interactions are more transactional and utility-driven. Users typically consult Yelp with a clear intent—such as choosing a restaurant for a gathering—resulting in fewer overall interactions per user and per item. Consequently, the volume of user–item interaction data is lower, and the urgency for highly aggressive recommendation tactics is reduced.

For more detailed information regarding its properties, please refer to Appendix 1.

2.2 Problem Definition, Impact and Supporting Statistics

Building effective recommendations on the Yelp dataset requires overcoming three principal challenges:

2.2.1 Data Sparsity

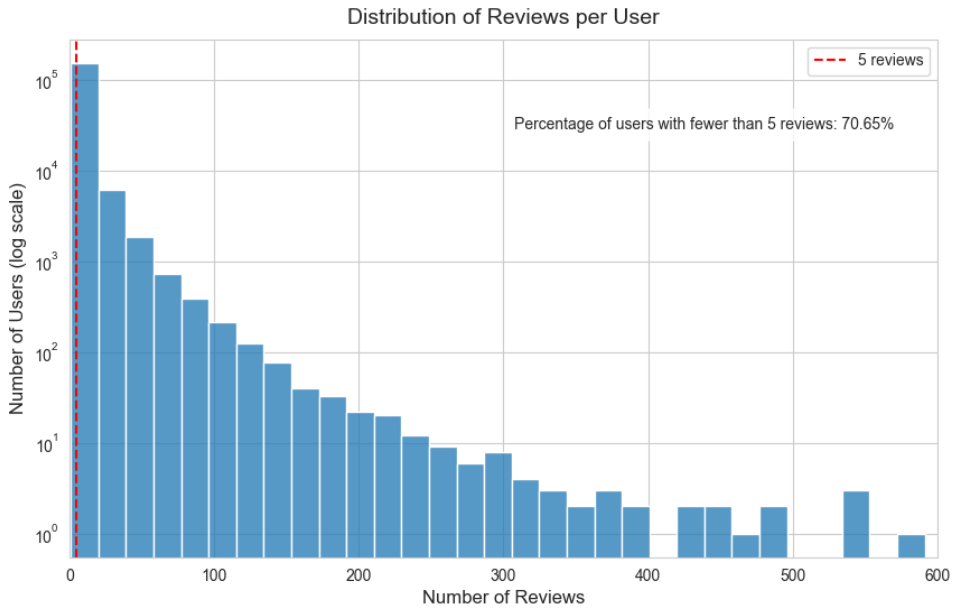
Definition: Data sparsity occurs when only a small fraction of the possible user–item interactions is observed—i.e., most users have reviewed only a handful of businesses, and many businesses have received a minimal number of reviews.

Impacts:

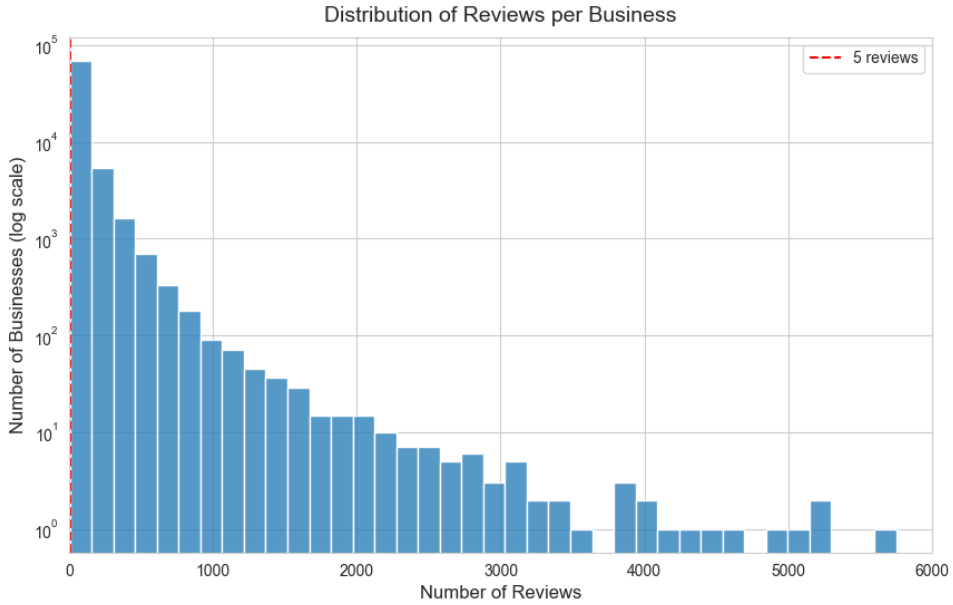
- **Collaborative Filtering**: Models that infer user preferences from patterns of similar users or items fail when there is insufficient overlap in reviews or ratings. Sparse interactions lead to unreliable similarity scores and poorer predictive accuracy.
- **Content-Based Filtering**: These approaches rely on building robust profiles for users and items based on past interactions. When histories are incomplete, profile vectors become partial, weakening the model's ability to match users with relevant items.

- **Risk of Overfitting**: Training on sparse data increases the likelihood that a model will fit noise (e.g., idiosyncratic review behavior) rather than true preference signals. Such models generalize poorly to new users or businesses.

Figure 2.1: Visualization of Data Sparsity in the Yelp Dataset



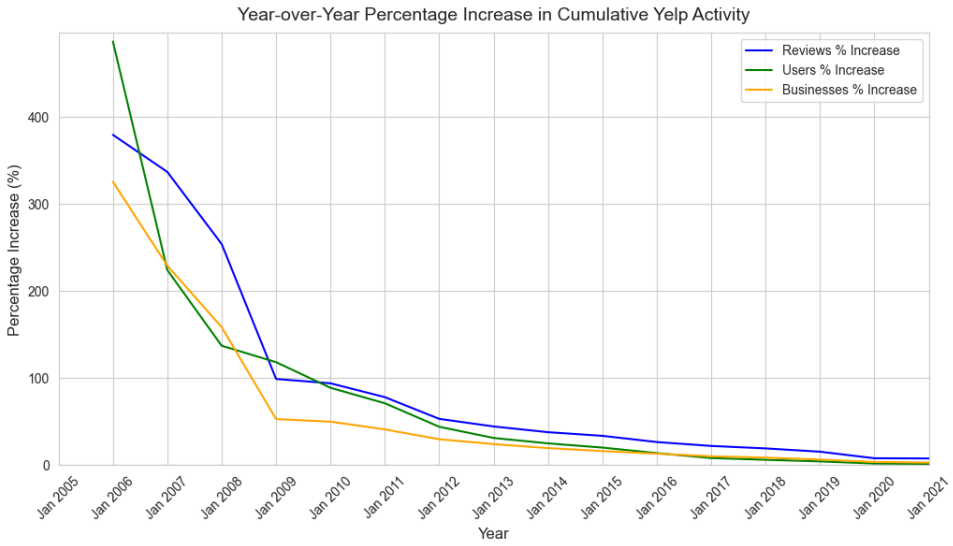
(a) **Histogram of Reviews per User:** Shows that most users contribute only a few reviews, with a long tail of highly active reviewers.



(b) **Histogram of Reviews per Business:** Demonstrates that most businesses receive only a limited number of reviews, leaving many rarely reviewed.

Actual interactions	Possible interactions	Matrix Sparsity	Matrix Density
980,418	11,790,436,855	0.9999	0.0001

(c) **User-Business Interaction Table:** An excerpt of the full interaction matrix, highlighting the large proportion of unobserved entries (empty cells).



(d) **Yearly Growth Line Graph:** Plots cumulative counts of users, businesses, and reviews over time, revealing that while user and business counts grow steadily, review volume increases at a slower rate.

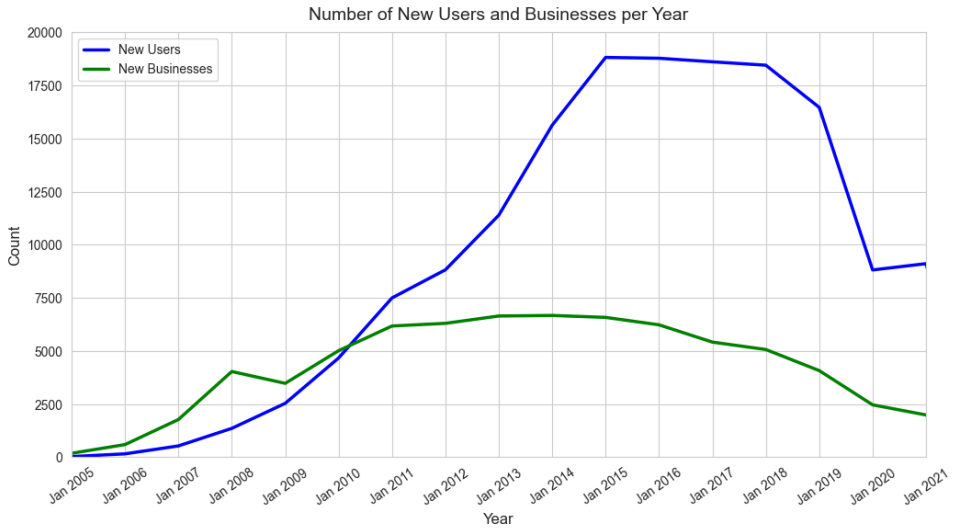
2.2.2 Cold-Start Problem (New Users)

Definition: The cold-start problem arises when a recommendation system has little or no historical data about new users or new items. Here, we focus on new users.

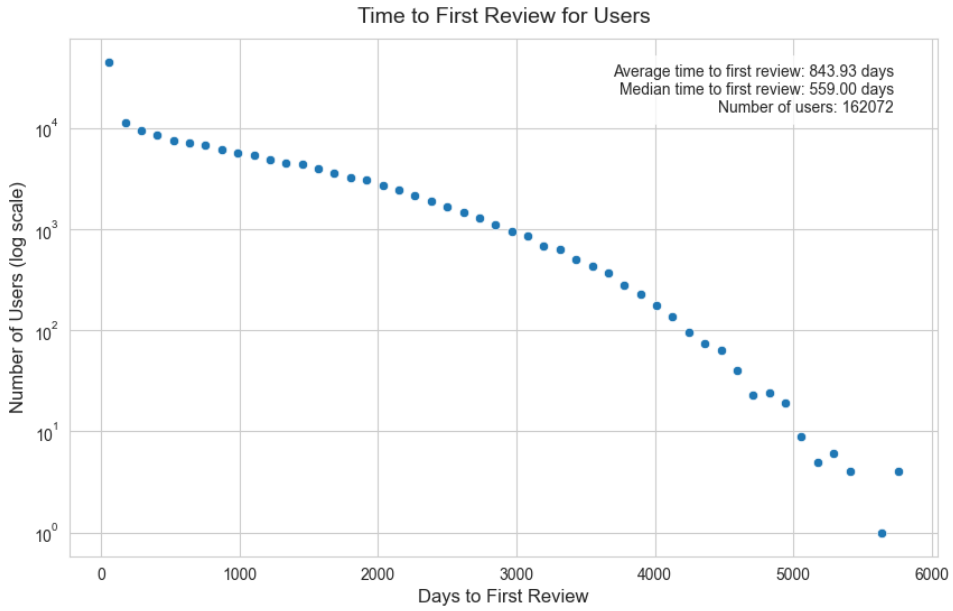
Impacts:

- **Lack of Personalization**: Without past reviews or ratings, the system defaults to generic strategies—commonly recommending the most popular or highest-rated businesses.
- **User Engagement Risk**: If initial suggestions feel irrelevant, new users may disengage before providing any feedback, perpetuating the lack of data.
- **Delayed Feedback Loop**: The time between a user joining and contributing their first review can be substantial, stalling personalized recommendations.

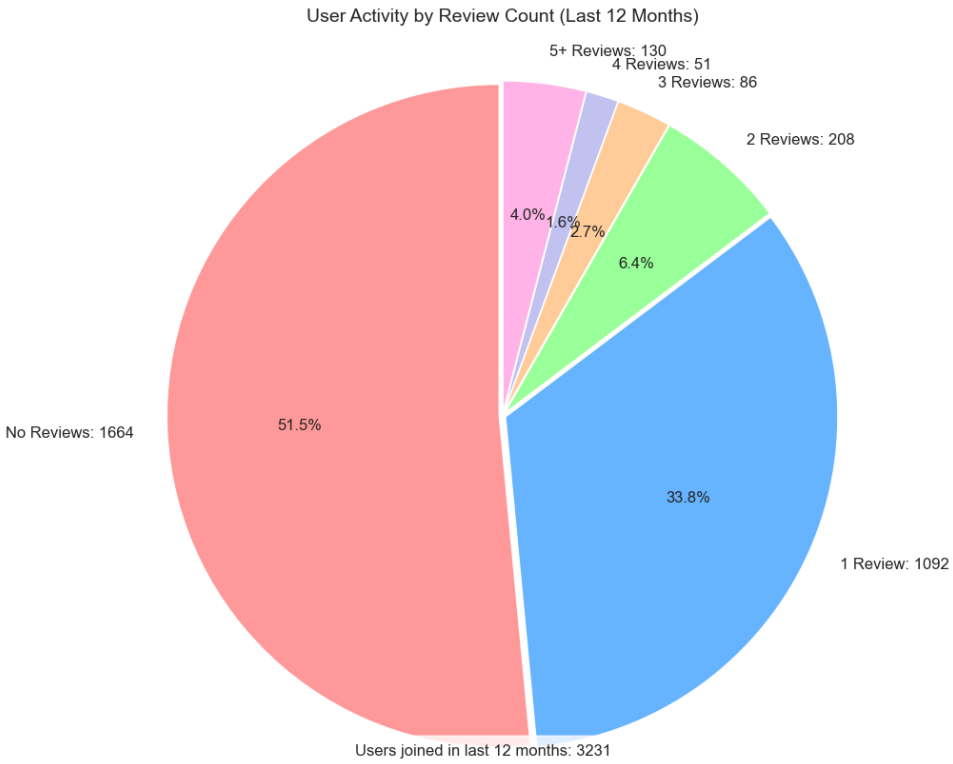
Figure 2.2: Visualizing the Cold-Start Problem in the Yelp Dataset



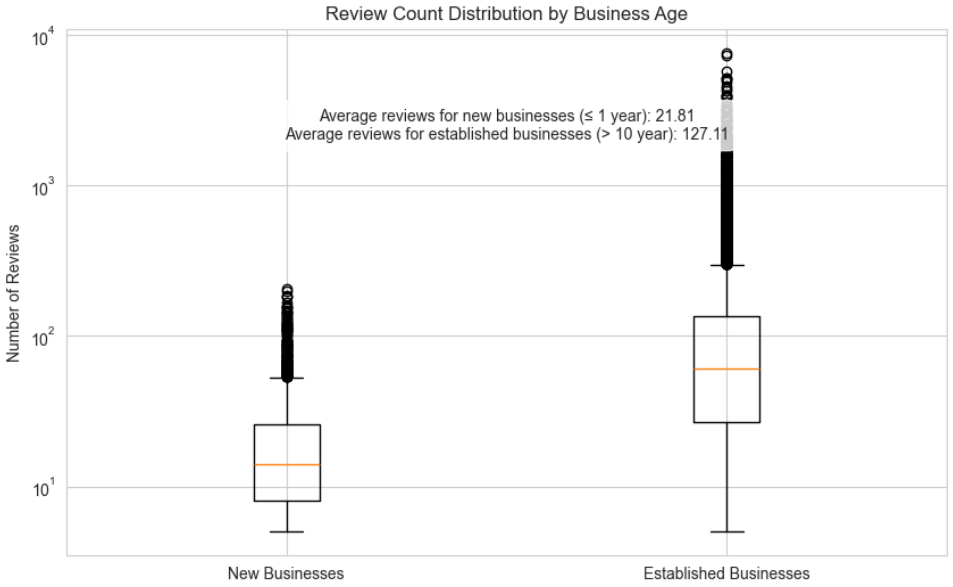
(a) **New Users and Businesses Over Time**: A line graph showing the steady influx of new users and new businesses joining the platform.



(b) **Time to First Review Plot:** Illustrates that many new users wait days or weeks before posting their first review.



(c) **Pie Chart of Review Activity for New Users:** Indicates that fewer than 50 % of new users write more than one review in their first year.



(d) **Box Plot Comparing Review Counts**: Contrasts review counts for businesses established within the past year against those active for longer, highlighting how new businesses receive significantly fewer reviews.

2.2.3 Long-Tail Items

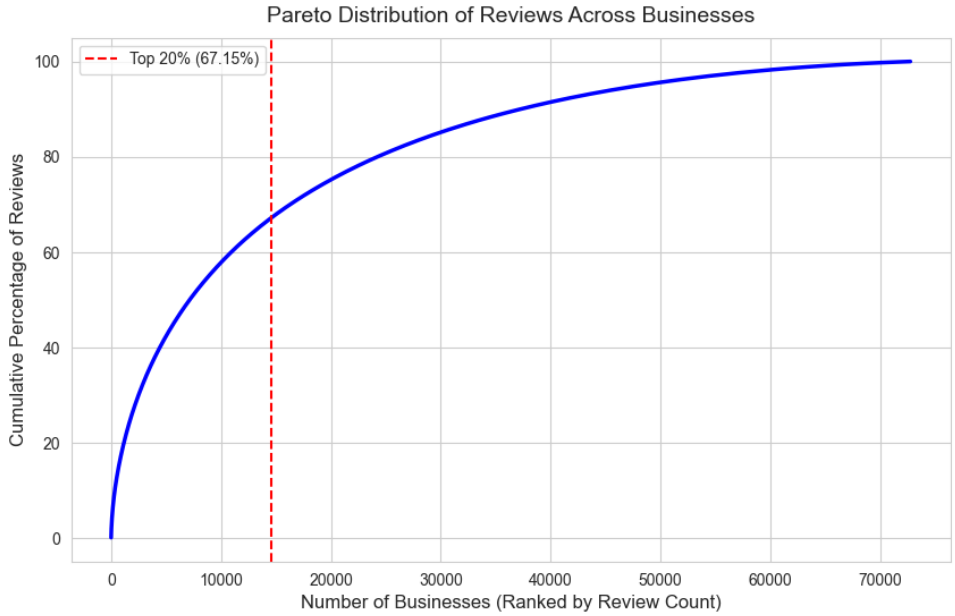
Definition: Long-tail items are businesses with relatively few reviews compared to the small set of extremely popular businesses. This imbalance creates a skewed distribution in user–item interactions.

Impacts:

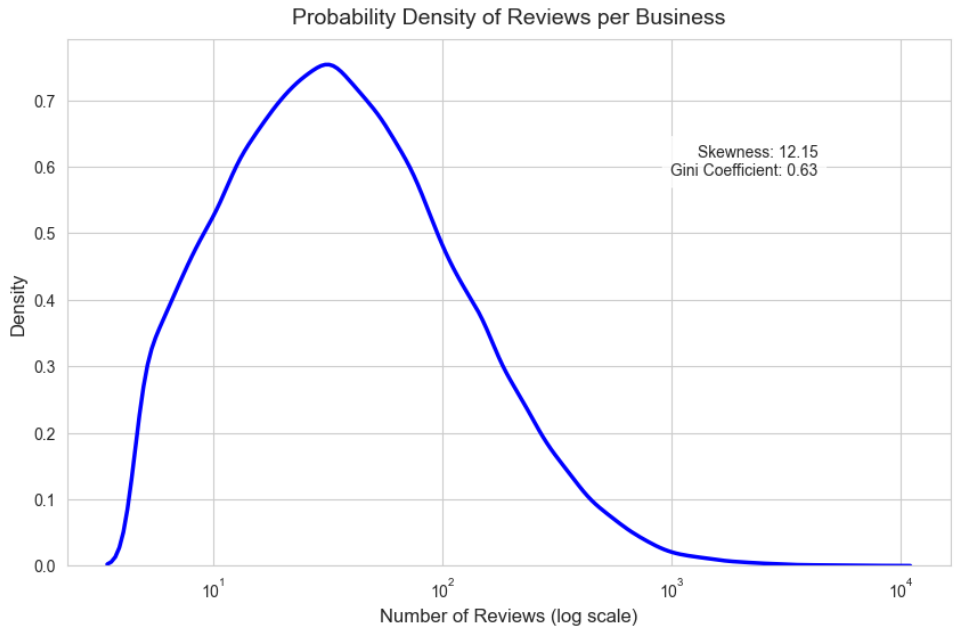
- Limited Content Diversity: Recommendation algorithms optimized for accuracy tend to favor high-interaction ("head") items, reducing exposure of niche ("tail") businesses that may match unique user interests.

- Feedback-Loop Reinforcement: Recommending popular items increases their visibility and future interactions, further marginalizing long-tail items in a self-reinforcing cycle.

Figure 2.3: Visualizing the Distribution of Reviews in the Yelp Dataset



(a) **CDF of Review Distribution:** Demonstrates that the top 20 % of businesses account for approximately 67.15 % of all reviews, underscoring the head-heavy nature of engagement.

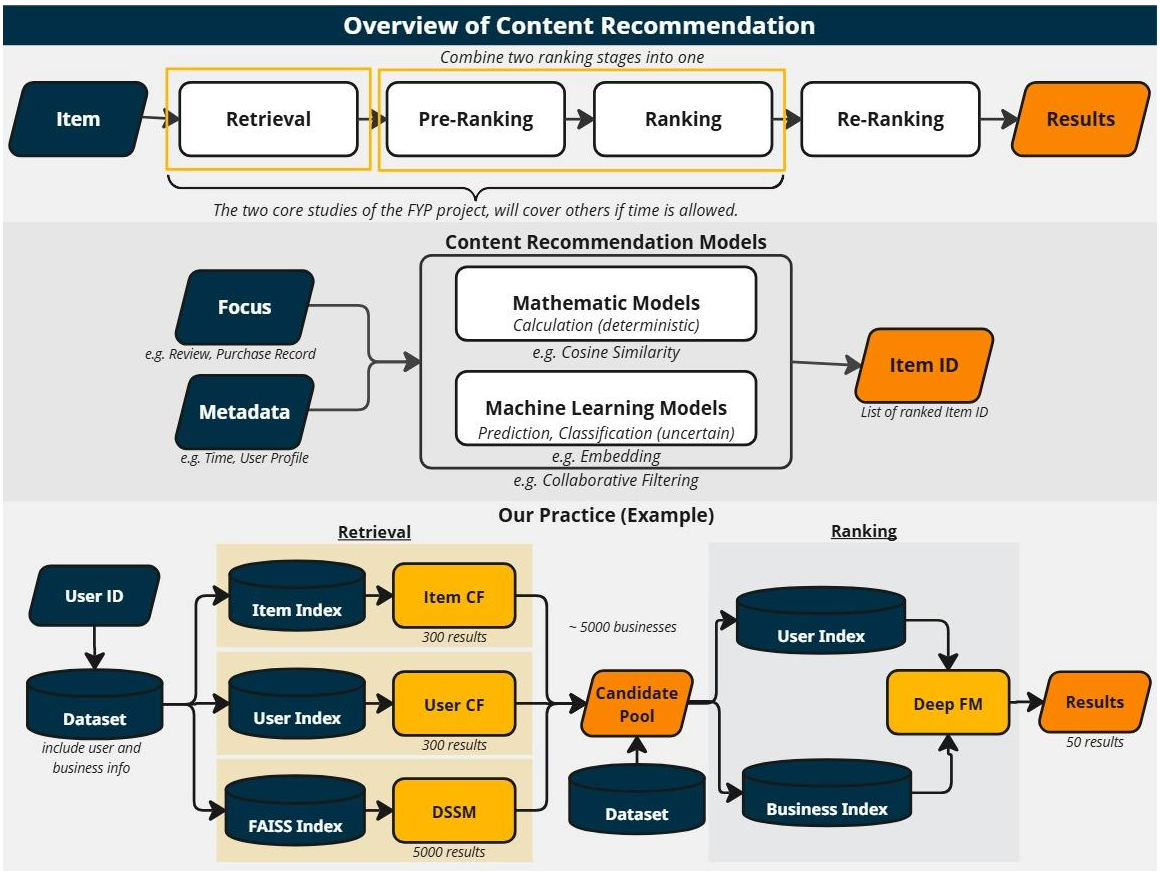


(b) **PDF of Review Counts per Business:** Depicts the skewness of the distribution and reports the Gini Coefficient as a quantitative measure of inequality among businesses.

Together, these challenges—data sparsity, the cold-start of new users, and the long-tail bias—highlight the complex landscape in which a Yelp-based recommendation system must operate. Addressing them requires a combination of model architectures and evaluation strategies, which we discuss in Section 3.

3. System Design and Methodology

3.1 Overview of Content Recommendation Systems



Content recommendation systems typically employ a four-stage pipeline to produce a final ranked list of items:

1. **Retrieval**: Identify a broad set of candidate items from the full catalog.
2. **Pre-Ranking**: Apply lightweight scoring (e.g., simple heuristics) to filter out low-relevance candidates.
3. **Ranking**: Use more complex models and richer features to score and order the surviving candidates.
4. **Re-Ranking**: Adjust the top results (e.g., to boost diversity or fairness) before presenting them to the user.

By escalating model complexity at each stage, the system can efficiently process millions of items and deliver a tailored recommendation list. In our implementation:

- **Retrieval Stage**: Three distinct models generate initial candidates.
- **Combined Pre-Ranking & Ranking Stages**: Advanced learning-based models refine and order those candidates.

3.2 Model Design and Architecture

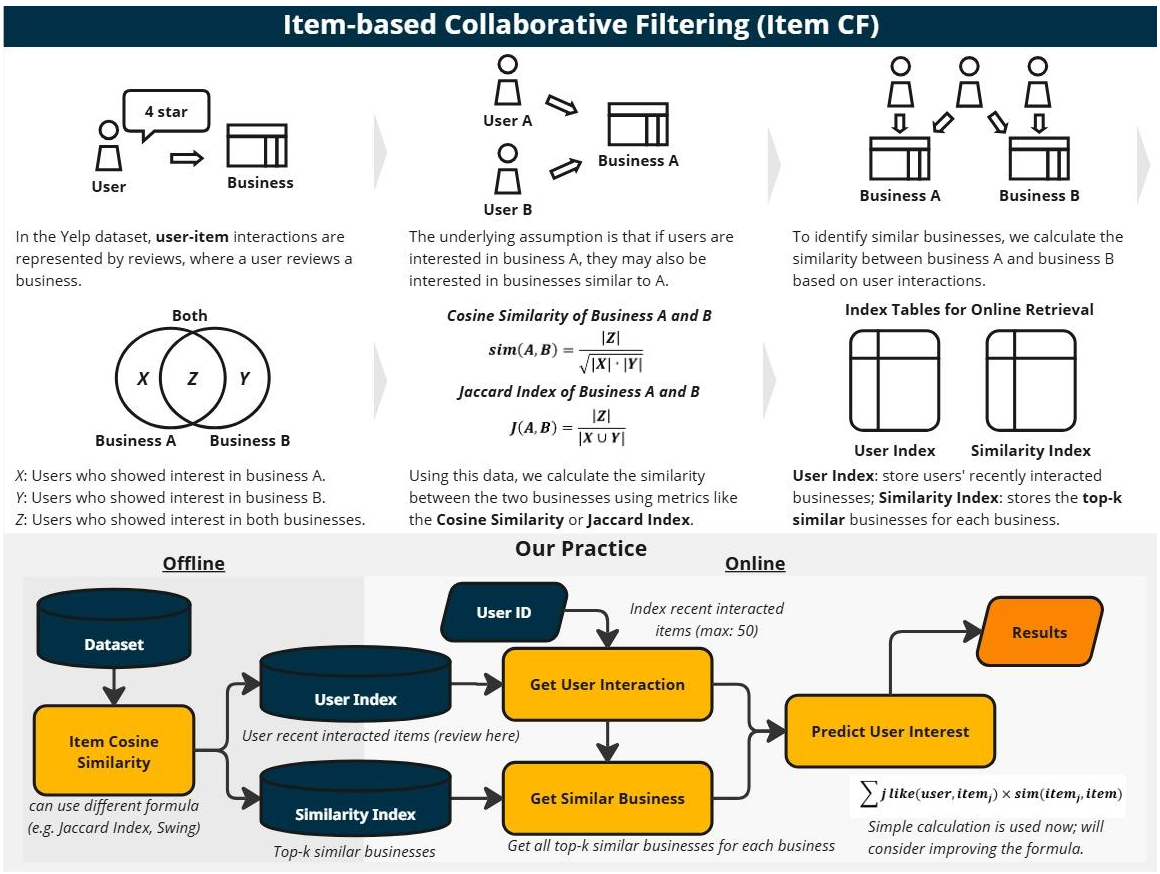
We employ a suite of models spanning baseline heuristics to deep-learning architectures. Diagrams (see Section 3.1) illustrate the end-to-end flow.

3.2.1 Baseline Models (Random, Uniform, Popular)

These simple approaches serve as benchmarks in the Retrieval Stage:

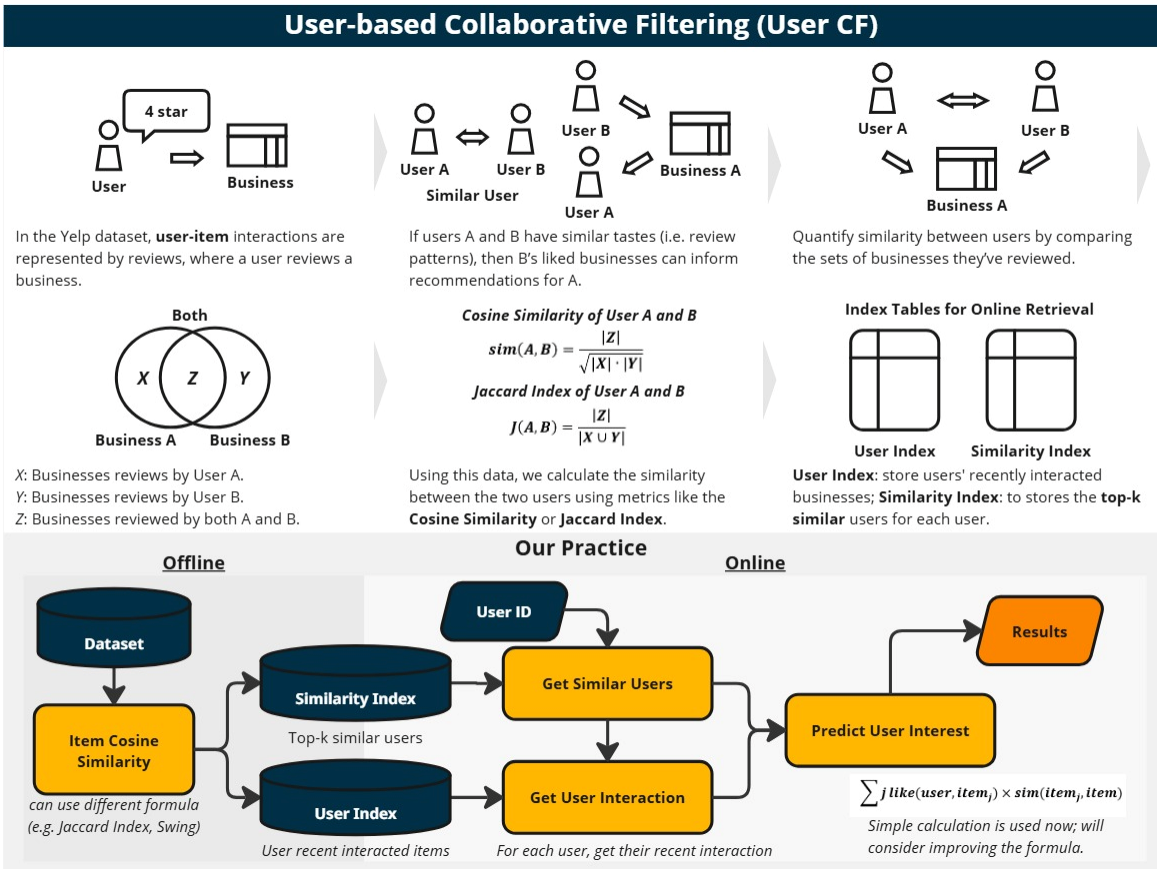
- **Random**: Select k businesses uniformly at random for each user.
- **Uniform**: Assign the same k randomly selected businesses to every user.
- **Popular**: Choose the top k businesses by combining recent review counts and average rating (e.g., past month).

3.2.2 Item-based Collaborative Filtering (Item CF)



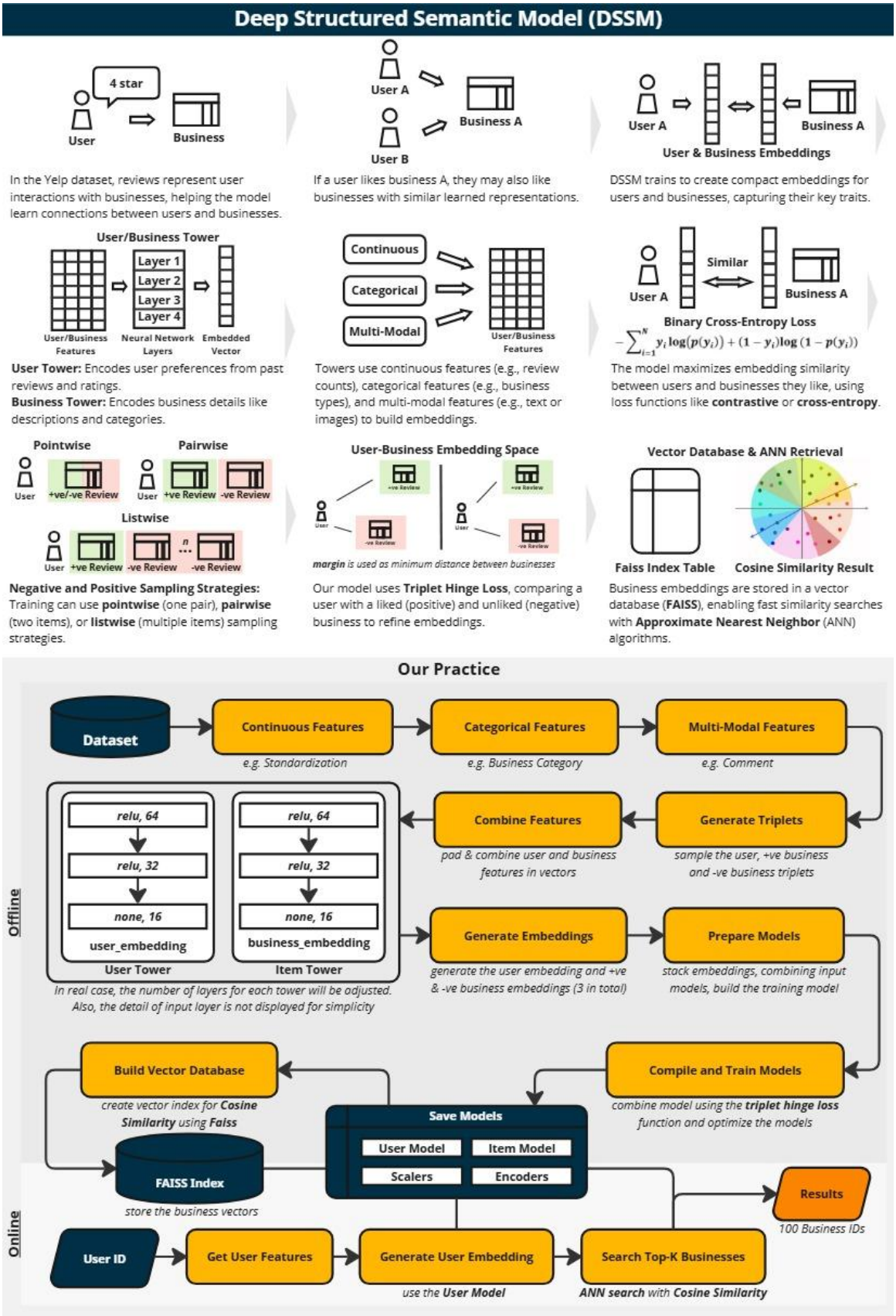
Item CF computes cosine similarity between business review vectors [6]. Businesses frequently co-reviewed with positive feedback are deemed similar, enabling recommendations of items akin to those a user has already liked. This method excels when historical item preferences are stable and well-documented.

3.2.3 User-based Collaborative Filtering (User CF)



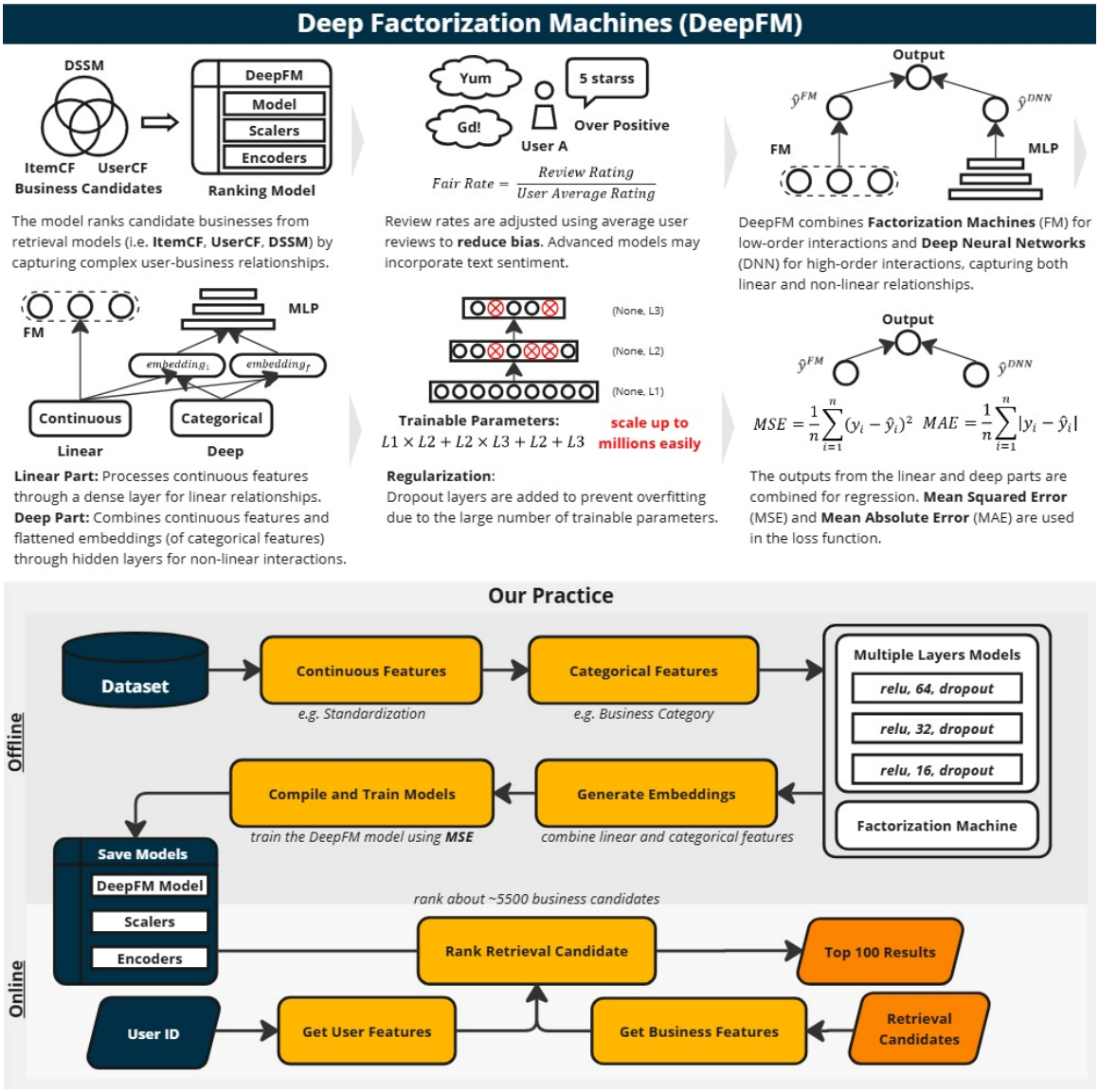
User CF finds peers whose review histories mirror the target user's patterns. By applying cosine similarity to users' rating vectors, it infers preferences from the community's collective behavior, producing personalized suggestions when interaction volume is ample.

3.2.4 Deep Structured Semantic Model (DSSM)



Adopting a dual-tower architecture [7], DSSM embeds users and businesses into a shared high-dimensional space. Continuous and categorical features are transformed into dense vectors. A triplet hinge loss trains the model to pull positive interactions closer and push negatives farther apart [8]. During retrieval, we use FAISS for Approximate Nearest Neighbor searches, ensuring fast, scalable look-ups.

3.2.5 Deep Factorized Machine (Deep FM)



In the Ranking Stage, Deep FM combines a factorization-machine component (for low-order feature interactions) with a deep neural network (for high-order patterns). Trained with a regression loss to predict engagement likelihood (e.g., probability of a positive review), Deep FM scores and orders the retrieval candidates to surface the most relevant items.

3.3 Evaluation Methodologies

This section outlines the diverse evaluation approaches used to assess the performance and real-world applicability of our recommendation models. Our evaluation is structured into distinct parts, addressing both offline prediction and production-like retrieval, as well as candidate analysis and ranking.

3.3.1 Retrieval Evaluation Strategy: Two-Tier Evaluation

The Two-Tier Evaluation framework is designed to bridge the gap between offline model performance and real-world system behaviour. It is divided into two tiers:

Tier 1: Prediction Evaluation

To measure the accuracy of the model's predicted scores or probabilities for user–item interactions using a controlled test set with ground-truth labels.

Input	A set of user–item pairs sampled from historical interactions.
Process	1. Use the model to predict a rating or probability score for each pair. 2. Compare the predictions against actual (positive/negative) labels.
Metrics	Accuracy, Precision, **Recall**, F1-score, **F-Beta** (B=2), **Unrated Percentage**.
Benefits	1. Provides a clear assessment of the model's ability to differentiate between positive and negative interactions. 2. Aids in diagnosing issues related to model training and feature representation. 3. Enables rapid iteration due to the smaller candidate set and straightforward evaluation process.

Tier 2: Retrieval Evaluation (Production Simulation)

To simulate a production scenario by retrieving the top-k candidate items for a user from a large catalog, thus evaluating the model's practical performance in identifying items that match historical positive interactions.

Input	A user identifier and the entire (or a large subset of the) item catalogue.
Process	1. Retrieve the top-k nearest neighbour items based on the model's embedding or similarity score. 2. Determine if the retrieved items include those with which the user has positively interacted.
Metrics	Accuracy, Precision@k (fraction of items in the top-k list that are truly positive), **Recall@k** (fraction of truly positive items that appear within the top-k list), F1 Score, **F-Beta** (B=2).
Benefits	1. Directly mimics a production scenario, thereby revealing the model's performance under real-world constraints such as runtime efficiency. 2. Highlights challenges like the "needle in a haystack" phenomenon, wherein even a strong model might exhibit low recall if the relevant items are scarce within a large candidate pool.

	3. Verifies the alignment of the model's training objectives with the demands of the retrieval task; models solely optimized for prediction may underperform in ranking.

The prediction simulation uses controlled, ground-truth data to reveal the model's performance when interactions are limited. The production simulation aspect is particularly well-suited for addressing **Data Sparsity** by testing the model on realistic, large-scale candidate pools where sparse interactions exist.

3.3.2 Retrieval Candidate Pool Analysis

This analysis focuses on the overall candidate set generated by the various retrieval models, with an emphasis on the diversity, uniqueness, and contribution of each model.

Coverage of Recommended Items: The proportion of unique businesses featured across all candidate lists.

$$Coverage = \frac{Number\ of\ unique\ businesses\ recommended}{Total\ businesses\ in\ dataset}$$

Contribution Analysis: Examines the overlap and unique contribution of each retrieval model, providing insights into how different algorithms complement or duplicate each other's recommendations.

Entropy of Recommendations: Measures the diversity among the recommended items; a higher entropy indicates a more varied recommendation set.

$$H = -\sum_{i} P(i) log P(i)$$

where $P(i)$ is the probability of a business appearing in recommendations.

This analysis helps evaluate the model's robustness in **Data Sparsity** scenarios by examining how well it retrieves a broad and varied set of candidates despite limited interaction history. It also reflects the model's ability to generalize beyond popular or frequently seen items.

3.3.3 Ranking Stage

For the ranking stage, performance is evaluated using a refined candidate list. The testing data primarily consists of historical user–item pairs with positive interactions.

Input	A set of user–item pairs sampled from historical interactions.
Process	1. Filter the test set to include only positive interactions.

	2. For each user, combine their n positive interactions with randomly selected businesses to form a candidate list of k items. 3. Rank the k candidates using the ranking model.
Metrics	**Mean Reciprocal Rank (MRR)**, First Relevant Rank (average rank of the first item for user), Normalized Discounted Cumulative Gain (NDCG), **NDCG@10 (NDCG for the top 10 recommendations)**.

In addition to quantitative evaluation, GPT is employed to simulate user behaviour and provide qualitative feedback on the recommendations:

Input	User profiles that include metadata such as recent interactions and preferred categories.
Process	1. GPT simulates user interests based on the provided profile. 2. GPT compares the recommended businesses (based on business categories, ratings, and previous reviews) against simulated preferences.
Metrics	Recommendation score (qualitative rating) and comments providing feedback on the relevance and suitability of recommendations.

By ranking a mix of known positives and randomly sampled negatives, the evaluation directly tests the model's ability to distinguish relevant content under limited interaction settings, helping to tackle the **Data Sparsity** challenge.

3.3.4 Popularity Analysis

This analysis aims to identify and quantify any popularity bias in the recommendations. Businesses are categorized into popular and long-tail items based on a percentile cutoff following the Pareto Principle.

Long-Tail Coverage: Proportion of recommended businesses that fall into the long-tail category.

$$Long\text{-}Tail\ Coverage = \frac{Number\ of\ long\text{-}tail\ businesses\ recommended}{Total\ businesses\ recommended}$$

Gini Coefficient: Quantifies the inequality in the distribution of recommendations across businesses. A higher Gini indicates a greater concentration on popular items.

$$G = 1 - 2\sum_{i=1}^{N} P(i)(N - i + 0.5)$$

where $P(i)$ represents the fraction of recommendations received by a business.

Item Popularity Bias: Measures the fraction of recommendations that are allocated to popular items as opposed to long-tail items.

$$Popularity\ Bias = \frac{\sum_{popular\ items} recommendations}{\sum_{all\ items} recommendations}$$

This analysis is targeted at identifying and quantifying popularity bias, making it a direct approach for addressing the **Long-Tail Items** problem by encouraging more equitable exposure across the item spectrum.

4. Evaluation and Results

In this study, reviews rated 4 or 5 are treated as positive interactions, while those rated 1–3 are considered negative. It is important to acknowledge that rating behavior carries inherent bias—for example, a 2- or 3-star review does not necessarily imply disinterest. Our primary focus is on the presence of user–business interactions, which reflect a degree of intent or engagement.

Detailed training logs are available in the accompanying Excel.

4.1 Evaluation Results and Problem Resolution

4.1.1 Retrieval Evaluation Strategy: Two-Tier Evaluation

We assess retrieval performance in two phases: a **prediction** perspective (how well models predict known interactions) and a **production simulation** perspective (how well models would perform in a live setting).

Retrieval Models Overview:

		Item CF	User CF	DSSM	Baseline
Retrieval	**Accuracy**	0.53	0.54	0.55	0.32
	Precision	0.55	0.56	0.63	0.74
	Recall	0.25	0.33	0.15	0.02
	F1 Score	0.35	0.42	0.24	0.05
	Fβ (B=2)	0.29	0.36	0.17	0.03
Prediction	**Accuracy**	0.60	0.55	0.48	
	Precision	0.58	0.62	0.48	
	Recall	0.36	0.01	0.94	
	F1 Score	0.45	0.02	0.63	
	Fβ (B=2)	0.39	0.01	0.79	
	Unrated	0.43	0.91	0.00	

Among the three retrieval approaches, **User-based Collaborative Filtering (User CF)** achieves the strongest overall results. Notably, **the Deep Structured Semantic Model (DSSM)** demonstrates robust generalization—especially in the "Unrated" category of the prediction phase.

Baseline Models Overview:

	Metric/Model	Random	Popular	Uniform
Retrieval	Accuracy	0.31	0.32	0.31
	Precision	0.75	0.74	0.83
	Recall	0.00	0.02	0.01
	F1 Score	0.01	0.05	0.02
	Fβ (B=2)	0.00	0.03	0.01

All baseline models (random, uniform, popular) perform poorly in the retrieval simulation, underscoring the need for more sophisticated techniques.

Item-based Collaborative Filtering (Item CF):

	Metric/Model	1001	1002	1003	1004
Retrieval	Accuracy	0.53	0.53	0.51	0.56
	Precision	0.55	0.55	0.48	0.61
	Recall	0.25	0.25	0.04	0.27
	F1 Score	0.35	0.35	0.07	0.37
	Fβ (B=2)	0.29	0.28	0.04	0.30
Prediction	Accuracy	0.60	0.56	0.60	0.56
	Precision	0.58	0.60	0.61	0.59
	Recall	0.36	0.17	0.05	0.17
	F1 Score	0.45	0.27	0.10	0.26
	Fβ (B=2)	0.39	0.20	0.06	0.20
	Unrated	0.43	0.44	0.76	0.44

Model	Objective
1001	Baseline Model
1002	Add Time-Decay Feature
1003	Apply Jaccard Similarity
1004	User Cluster as Interaction Data

Incorporating user clusters to enrich business interactions yields the best Item CF results, although gains over the baseline remain modest.

User-based Collaborative Filtering (User CF):

	Metric/Model	1001	1002	1003	1004
Retrieval	Accuracy	0.54	0.53	0.51	0.55
	Precision	0.56	0.55	0.53	0.56
	Recall	0.33	0.25	0.10	0.35
	F1 Score	0.42	0.34	0.17	0.43
	Fβ (B=2)	0.36	0.28	0.12	0.38
Prediction	Accuracy	0.55	0.56	0.58	0.57
	Precision	0.62	0.57	0.58	0.57
	Recall	0.01	0.37	0.11	0.37
	F1 Score	0.02	0.45	0.18	0.45
	Fβ (B=2)	0.01	0.40	0.13	0.40
	Unrated	0.91	0.43	0.78	0.42

Model	Objective
1001	Baseline Model
1002	Add Time-Decay Feature
1003	Apply Jaccard Similarity
1004	User Cluster as Similar User Data

Similarly, expanding the neighbor pool via clustering improves User CF retrieval metrics.

Deep Structured Semantic Model (DSSM):

	Metric/Model	1001	1002	1003	1004
Retrieval	Accuracy	0.55	0.53	0.52	0.54
	Precision	0.63	0.66	0.50	0.61
	Recall	0.15	0.02	0.06	0.09
	F1 Score	0.24	0.04	0.11	0.16
	Fβ (B=2)	0.17	0.02	0.08	0.11
Prediction	Accuracy	0.48	0.51	0.50	0.57
	Precision	0.48	0.44	0.49	0.55
	Recall	0.94	0.10	0.88	0.56
	F1 Score	0.63	0.17	0.63	0.55
	Fβ (B=2)	0.79	0.12	0.76	0.56

Model	Objective
1001	Baseline Model
1002	Binary Cross Entropy Loss Function
1003	GeoHash Feature
1004	Business Categories Feature

The DSSM baseline attains the highest retrieval scores overall but exhibits an overly optimistic bias in prediction—classifying many user–business pairs as positive.

Importantly, DSSM tends to recommend businesses distinct from those selected by CF models, introducing broader yet still relevant suggestions. We analyze this diversity effect further in Section 4.1.2.

4.1.2 Retrieval Candidate Pool Analysis

Background Statistic:

	Unique Users	Num of Businesses/User	Unique Businesses	Entropy
Item CF	848 (64.19%)	201.77	29083 (48.15%)	9.40
DSSM	1000 (75.70%)	5000.00	43304 (71.70%)	9.56
User CF	848 (64.19%)	458.57	43921 (72.72%)	9.99
Total	1321	2062.30	60395	9.74

By combining candidates from all retrieval models, the pooled set covers about 77 % of all businesses and achieves an entropy score of 9.74.

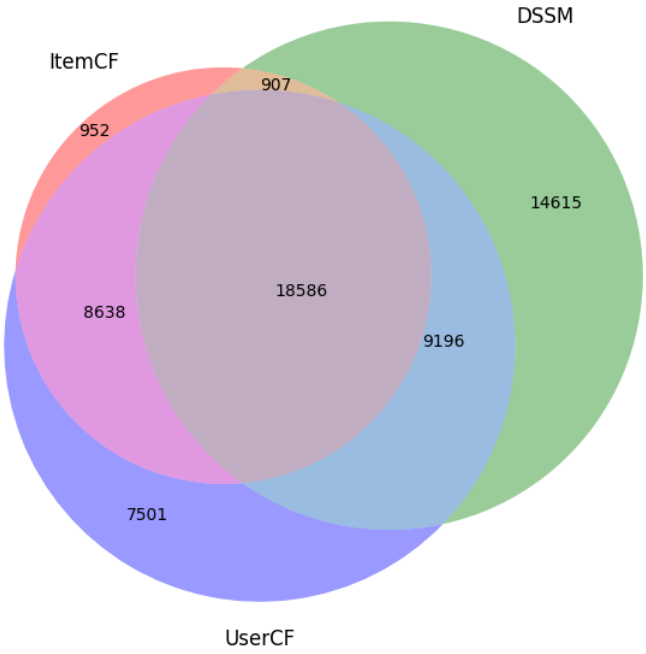
Overlap of User and Business Pairs:

	Frequency	Percentage
Item CF & DSSM	20825	0.39%
Item CF & User CF	110614	2.05%
DSSM & User CF	39971	0.74%
All Models	13631	0.25%

Retrieval outputs from different models rarely coincide on the same user–business pairs; the highest overlap occurs between Item CF and User CF.

Overlap of Business Recommended:

Figure 4.1.1: Venn Diagram of Item Overlap Across Models



	Frequency	Percentage
ItemCF & DSSM	19493	32.28%
ItemCF & UserCF	27224	45.08%
DSSM & UserCF	27782	46.00%
All Models	18586	30.77%

The most frequently co-recommended businesses are typically the platform's most popular. Crucially, DSSM contributes unique, less-surfaced businesses, increasing overall recommendation diversity.

4.1.3 Ranking Stage

Deep Factorization Machines (Deep FM):

	Metric/Model	1001	1002	1003	1004
Ranking	**MRR**	0.08	0.06	0.06	0.05
	FRR@100	12.00	17.00	17.00	20.00
	NDCG@10	0.05	0.03	0.03	0.02
	NDCG@100	0.26	0.24	0.24	0.23

Model	Objective
1001	Baseline Model
1002	GeoHash Feature
1003	Business Categories Feature
1004	GeoHash & Business Categories Feature

In the ranking phase, the baseline ranking model delivers the best performance. On average, the first relevant item appears within the top 12 candidates out of 100, indicating satisfactory ranking effectiveness.

GPT Ranking Evaluation:

GPT was employed as a proxy user to assess whether the top-ranked recommendations align with inferred preferences derived from user profiles.

Case Study: User Molly

Business	Categories	Alignment Assessment	Relevance Score
Tumerico	Mexican, Gluten-Free, Vegetarian, Restaurants, Vegan	Offers a restaurant experience but focuses on vegetarian and vegan options, misaligned with the user's preference for American cuisine. May appeal to health-conscious users.	7/10
Czerw's Kielbasy	Food, Specialty Food, Meat Shops	Focuses on meat products, which does not align with preferences for bars, nightlife, or coffee. Lacks relevance to the user's preferred categories.	5/10
ICI Macarons & Cafe	Desserts, Bakeries, Food, Coffee & Tea	Aligns well with the user's interest in coffee and tea, offering café experience. Lacks the nightlife aspect but remains highly relevant.	8/10
Cathedral Basilica of Saint Louis	Churches, Landmarks & Historical Buildings	Does not align with preferences for nightlife, bars, coffee, or American cuisine. Unlikely to interest the user due to its religious and historical nature.	4/10
Latteria	Food, Ice Cream & Frozen Yogurt, Coffee & Tea	Aligns with the user's interest in coffee and tea, though it does not cater to nightlife or bars. Offers a relevant and appealing option.	7/10

The GPT-based evaluation effectively overcomes sparse user data challenges by generalizing preferences from limited information. The results balance highly relevant recommendations with suggestions that introduce new experiences.

For more evaluation result, please refer to Appendix 2.

4.1.4 Popularity Analysis

Figure 4.1.2: Long-Tail Coverage on Recommendations (Unique Businesses)

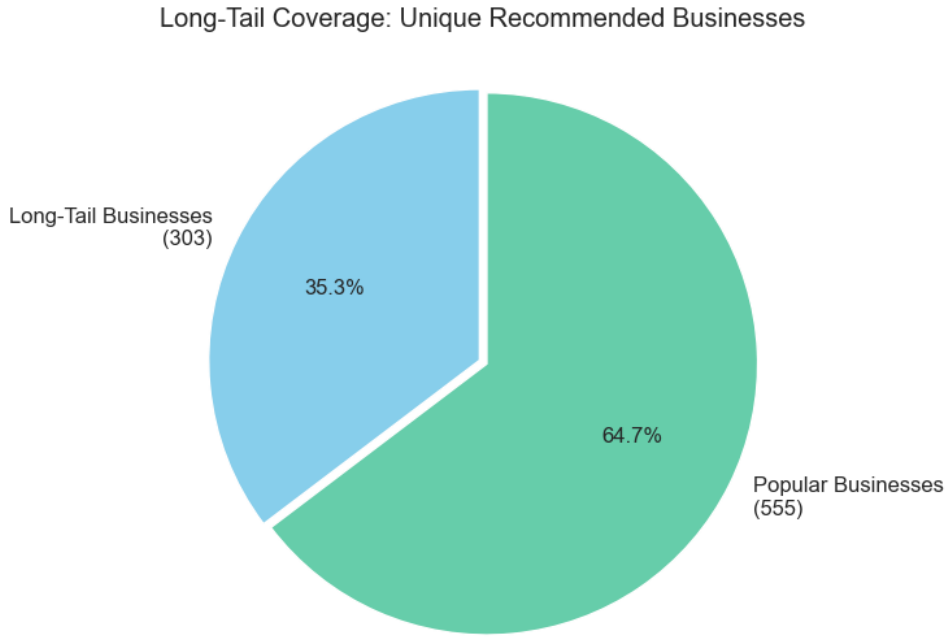


Figure 4.1.3: PDF and Gini Coefficient in Recommendations

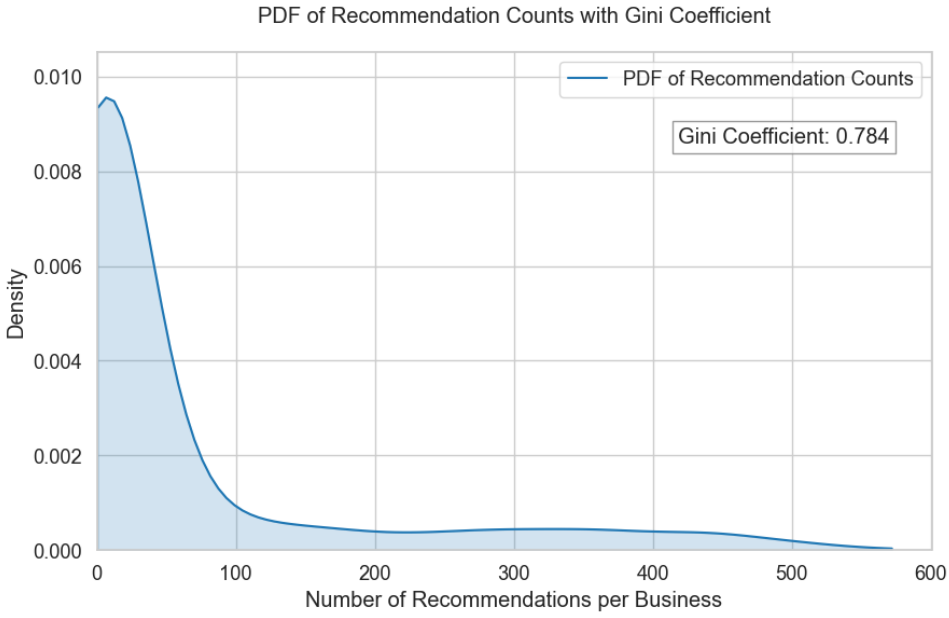
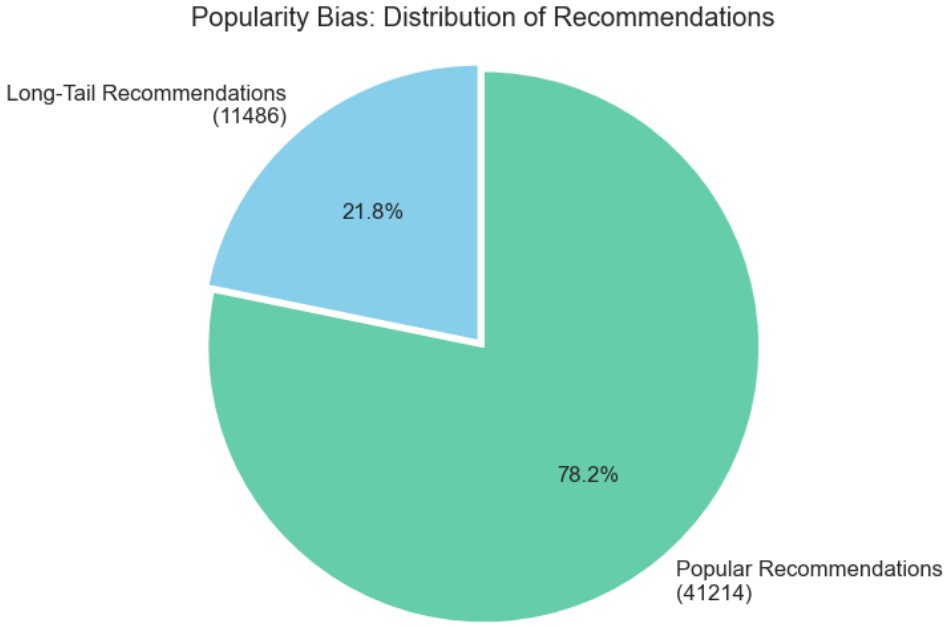


Figure 4.1.4: Popularity Bias on Recommendations (Distribution of Recommendations)



As shown, the system achieves a **long-tail coverage** of 0.35, a **Gini coefficient** of 0.78, and a **popularity bias** of 0.78. These metrics indicate a balanced trade-off between recommending popular businesses and maintaining diversity.

4.2 Limitations of Personalized Recommendations and Solutions

4.2.1 Diminishing Effect of Number of Reviews

Personalization quality is strongly correlated with the number of reviews a user has submitted. Users with sparse review histories experience lower retrieval performance across all models:

Figure 4.2.1: Performance Comparison of Item CF Model

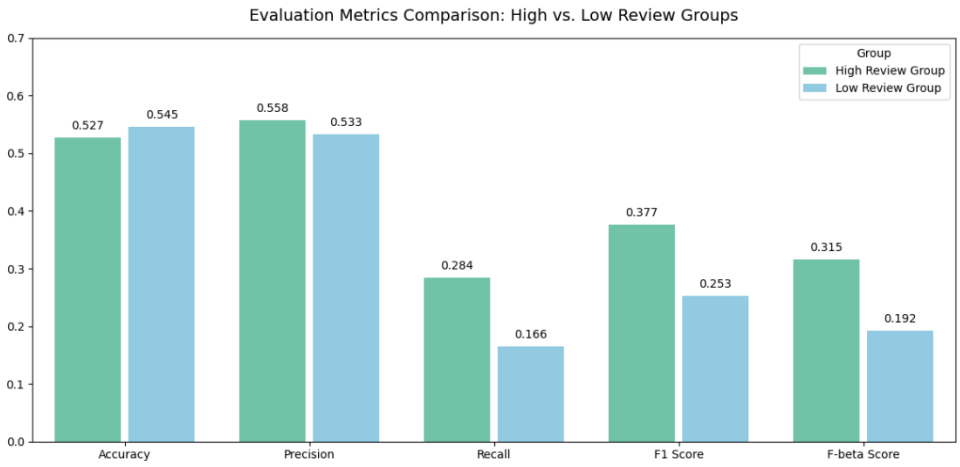


Figure 4.2.2: Performance Comparison of User CF Model

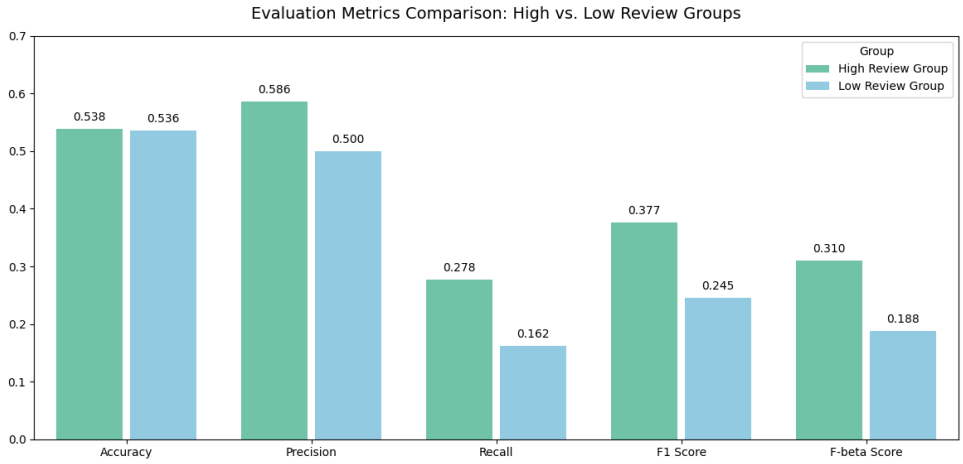
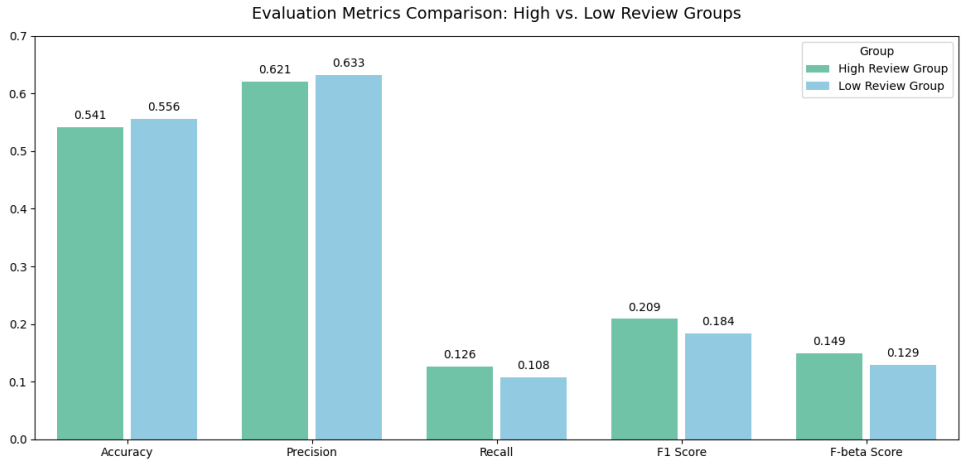


Figure 4.2.3: Performance Comparison of DSSM Model



To address this, we recommend strategies that encourage early review contributions, thereby accelerating the data collection needed for effective personalization.

4.2.2 Generalized Recommendation Strategies

For users lacking sufficient personal data, we explored two fallback approaches:

1. Recent Popular Recommendations

Suggest the highest-rated and most-reviewed businesses over the past two months.

2. Category-based Recommendations

Identify each user's top five business categories from any existing reviews (or default to popular categories). Users are then grouped via the nearest-neighbor matching on category preferences, and CF models are trained within each cluster.

These generalized strategies not only improve engagement for new or sparsely active users but also generate additional interaction data to enable a seamless transition to fully personalized recommendations over time.

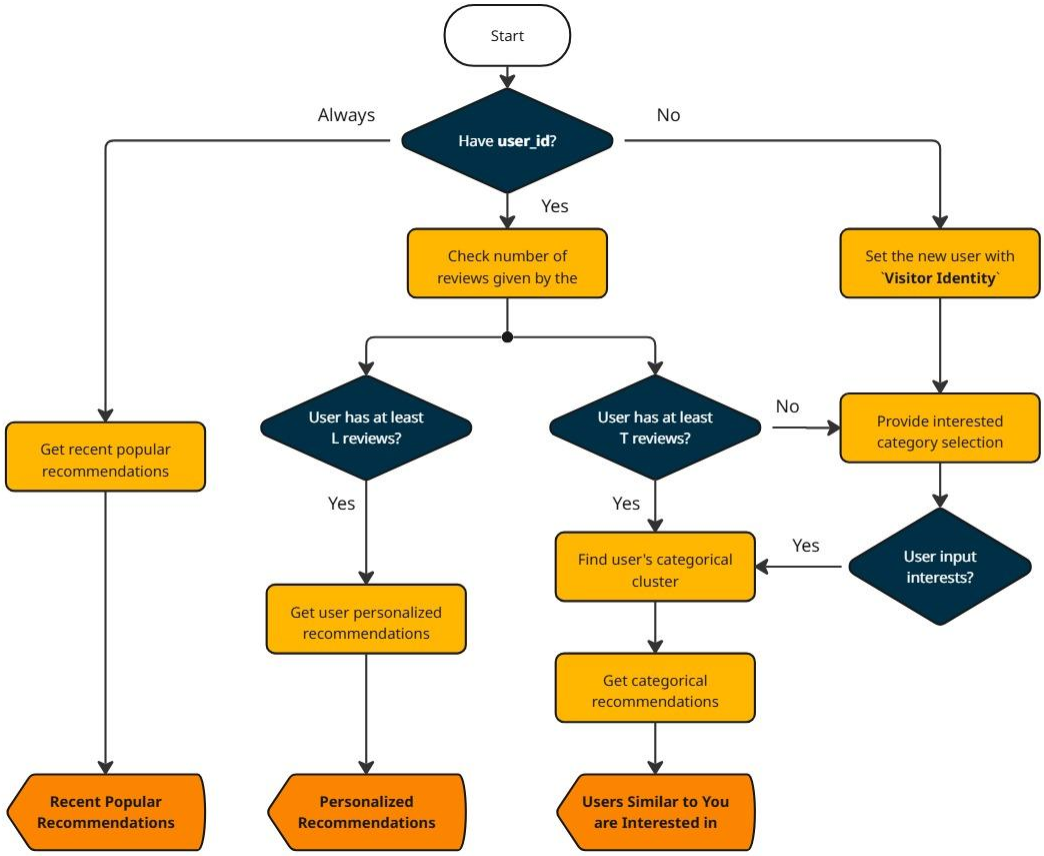
4.3 Content Recommendations System

Building on the evaluation insights, we developed a full-stack content recommendation system with both frontend and backend components.

see Appendix 3 for software details

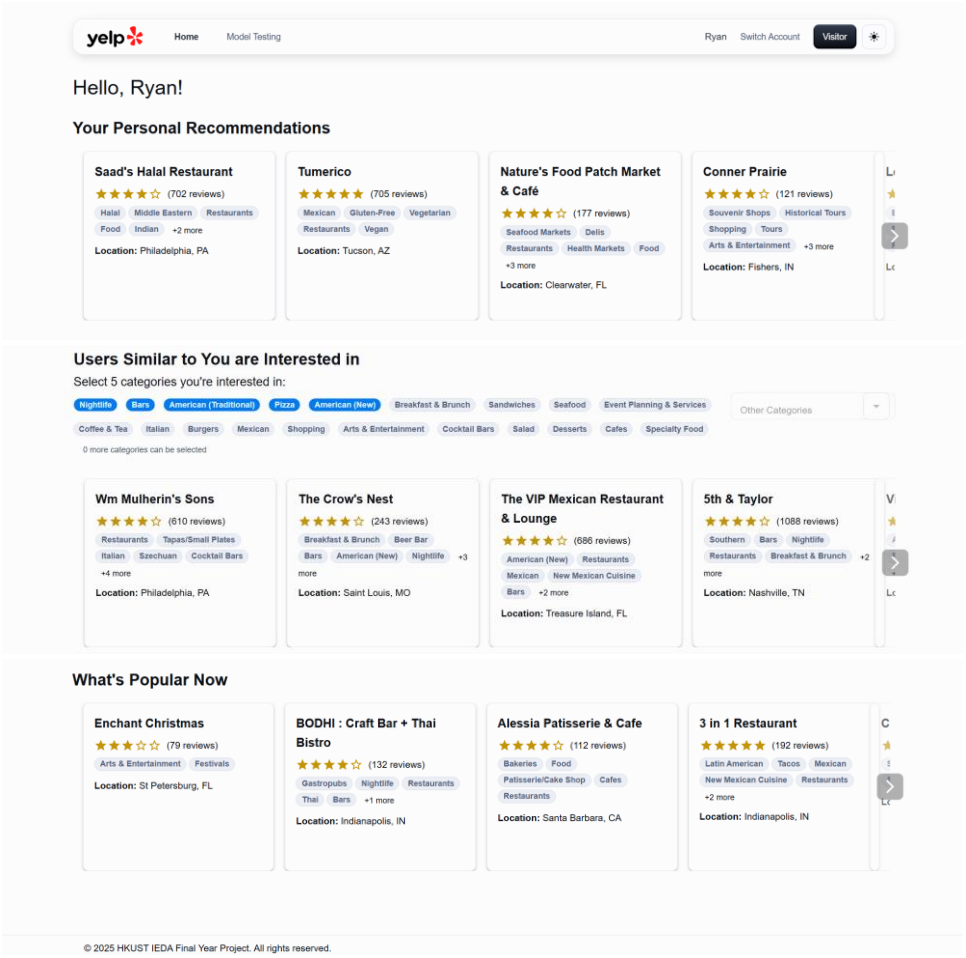
4.3.1 User Journey Map

Figure 4.3: User Flow Chart in the Content Recommendation System

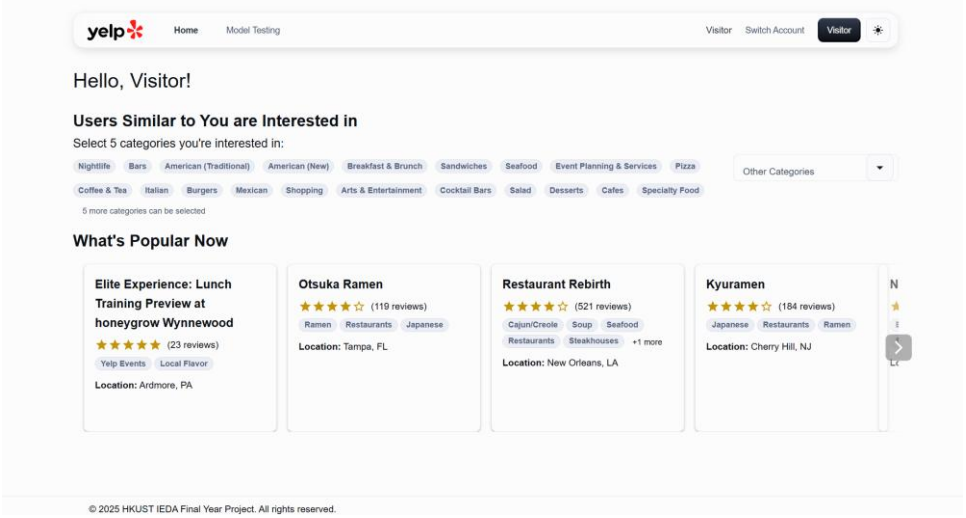


- Users with ≥ L reviews (e.g., 25) receive personalized recommendations.
- All users—regardless of history—see both Recent Popular and Category-Based recommendations (using default categories if sufficient data exists).

4.3.2 Content Recommendations System Web App



Homepage for Logged-In Users ("Ryan"): Personalized suggestions based on review history.



Homepage for Visitors: Generalized recommendations.

For a detailed demonstration, please watch the video provided on GitHub page [9].

5. Conclusions

5.1 Summary of Key Findings

In the retrieval stage, collaborative-filtering models (Item CF and User CF) achieve the best overall performance, although both are constrained by data sparsity. To alleviate this limitation, we apply clustering techniques to expand the interaction data available for similarity calculations. From a different perspective, the deep-learning model (DSSM) can generate diverse recommendations even when user interactions are limited. Although the use of approximate nearest-neighbor (ANN) search slightly reduces DSSM's prediction accuracy, the retrieval candidate–pool analysis demonstrates that DSSM still contributes valuable alternative suggestions.

In the ranking stage, the Deep Factorization Machine (Deep FM) model effectively captures both low-order and high-order feature interactions, modeling linear and non-linear relationships simultaneously. Its performance is promising: on average, a relevant item appears between positions 15 and 20 in the ranked list of 100 candidates, demonstrating that Deep FM reliably selects a high-quality top 100 from a pool of roughly 6,000 retrieval candidates. In addition, GPT-based qualitative evaluation provides deeper insights into how well these ranked recommendations align with user preferences.

However, the limited number of reviews per user remains a major bottleneck for personalized recommendations. To bridge the gap between cold-start and fully personalized experiences, we propose generalized strategies—Recent Popular and Category-Based recommendations—that adapt to users with sparse histories and incrementally improve personalization as more data are collected.

5.2 Limitations and Challenges

This project encountered two primary challenges:

1. **Lack of Implicit Interaction Data:**
Without access to clicks, views, or other implicit signals, our models rely solely on explicit feedback (reviews), which restricts the depth of user-behavior modeling. To make the most of available data, we incorporated two collaborative-filtering variants that leverage different aspects of the review matrix.
2. **Absence of A/B Testing:**
In the absence of real users, we could not conduct live A/B experiments. Instead, we employed retrieval-simulation metrics and GPT-based qualitative assessments as proxies for online performance, though these remain indirect measures of true user satisfaction.

Additionally, we down-sampled the dataset to meet computational constraints, which may have altered some inherent data relationships.

Further details regarding the sampling process can be found in Appendix 4.

5.3 Future Work and Potential Applications

5.3.1 Deep Learning Models

Deep-learning architectures hold significant room for enhancement. Future research could investigate advanced network designs, improved sampling methods to reduce bias, and richer feature-engineering pipelines. Given the scope and complexity, this effort may warrant a dedicated study on deep-learning approaches in recommendation systems.

5.3.2 Ranking Stage

In the current implementation, only Deep FM is employed in ranking. Future work could experiment with multiple, diverse ranking models—effectively approximating a mixture-of-experts approach—so that each model contributes its unique strengths. To support this experimentation, we should collect additional online metrics such as click-through rates, "like" rates, and review rates. Furthermore, enhancing the quality and relevance of the retrieval candidate pool will be essential to maximize any ensemble or multi-model ranking strategy.

5.3.3 Potential Applications

One promising application is a course-recommendation system for HKUST. Although initial data-preprocessing challenges—standardizing Student Feedback Questionnaires (SFQs), mapping courses to consistent categories, and anonymizing student records—prevented its realization here, the underlying methodology directly parallels our Yelp-based framework. In such a system:

- **SFQs as Review Data**: Students' course evaluations (ratings and written feedback) would serve as explicit interaction signals, analogous to Yelp reviews.
- **Rich Student Metadata**: Profiles including faculty affiliation, major, year of study, and elective versus core status could form additional features in both collaborative-filtering and deep models.
- **Category-Based Clustering**: By grouping students according to their top-rated course categories (e.g., "Data Science," "Humanities," "Engineering Foundations"), we could address cold-start users and refine recommendations for niche subject areas.

- **Hybrid Modeling**: Combining collaborative filtering on SFQs with content-based embeddings of course descriptions and metadata would help surface relevant electives and core courses tailored to each student's background and interests.

Implementing this system could enhance students' course-selection process, aid academic advising, and inform curriculum design—delivering significant value to the HKUST community.

6. Appendixes

1. *Yelp Dataset Information*

business_details (business)

Field Name	Data Type	Description
business_id	TEXT	Unique identifier for each business
name	TEXT	Name of the business
address	TEXT	Street address of the business
city	TEXT	City where the business is located
state	TEXT	State where the business is located
postal_code	TEXT	Postal code of the business
business_id	TEXT	Foreign key referencing business_details
category	TEXT	Business category

business_categories (business)

Field Name	Data Type	Description
business_id	TEXT	Foreign key referencing business_details
category	TEXT	Business category

checkin_data (business)

Field Name	Data Type	Description
business_id	TEXT	Foreign key referencing business_details
checkin_date	TEXT	Date of check-in (format: YYYY-MM-DD HH:MM:SS)

tip_data (tip)

Field Name	Data Type	Description
user_id	TEXT	Foreign key referencing user_data
business_id	TEXT	Foreign key referencing business_details
text	TEXT	Tip content
date	TEXT	Date of the tip (YYYY-MM-DD HH:MM:SS)
compliment_count	INTEGER	Number of compliments received for the tip

review_data (review)

Field Name	Data Type	Description
review_id	TEXT	Unique identifier for each review
user_id	TEXT	Foreign key referencing user_data
business_id	TEXT	Foreign key referencing business_details
stars	REAL	Star rating given in the review
date	TEXT	Date of the review (YYYY-MM-DD HH:MM:SS)
text	TEXT	Review content
useful	INTEGER	Useful votes received
funny	INTEGER	Funny votes received
cool	INTEGER	Cool votes received

user_data (user)

Field Name	Data Type	Description
user_id	TEXT	Unique identifier for each user
name	TEXT	Name of the user
review_count	INTEGER	Number of reviews written by the user
yelping_since	TEXT	Date the user joined Yelp (YYYY-MM)
useful	INTEGER	Number of useful votes received
funny	INTEGER	Number of funny votes received
cool	INTEGER	Number of cool votes received
fans	INTEGER	Number of fans
average_stars	REAL	Average star rating given by the user
friends	TEXT	List of friends stored as a string
elite	TEXT	Years user was elite stored as a string
compliment_*	INTEGER	Counts of specific compliments received

2. *GPT Results*

User: Ryan

Business	**Categories**	**Alignment Assessment**	**Relevance Score**
Nathaniel Reid Baker	Sandwiches, Food, Desserts, Restaurants, Bakeries	This business focuses on bakery items and sandwiches, which do not align with the user's primary interests in nightlife, bars, and American cuisine. While it may be a good place for a casual meal, it lacks the nightlife aspect that the user is looking for.	5/10
Cathedral Basilica of Saint Louis	Churches, Landmarks & Historical Buildings	This business is entirely unrelated to the user's interests in nightlife and dining. It is a historical site and does not offer any bar or restaurant services.	3/10
Czerw's Kielbasy	Food, Specialty Food, Meat Shops	This business focuses on meat products and specialty foods, which may appeal to some users but does not cater to the user's specific	6/10

		interests in bars and American cuisine. It lacks the social aspect of nightlife.	
Insta Quick Oil Change	Automotive, Oil Change Stations	This business is completely unrelated to the user's profile, which focuses on dining and nightlife. It does not meet any of the user's preferences.	2/10
Rawk Star Cafe	Vegan, Juice Bars & Smoothies, Food, Gluten-Free	This business focuses on vegan and health-oriented food, which does not align with the user's preference for American (Traditional) and pizza. It also lacks the nightlife aspect.	5/10

User: Richard

Business	**Categories**	**Alignment Assessment**	**Relevance Score**
Insta Quick Oil Chang	Automotive, Oil Change Stations, Auto Parts & Supplies, Auto Repair	This business is completely unrelated to the user's interests in food categories such as breakfast, brunch, pizza, Mexican, and desserts. It does not meet any of the user's preferences.	2/10
Artisan Boulanger Patissier	Food, Restaurants, Bakeries	This bakery specializes in pastries and baked goods, which could appeal to the user's interest in desserts. However, it does not align with the user's preferences for breakfast, brunch, pizza, or Mexican cuisine. Additionally, it is currently closed, which makes it less relevant.	6/10
ICI Macarons & Cafe	Desserts, Bakeries, Food, Coffee & Tea, Specialty Food, Macarons	This business aligns well with the user's interest in desserts and offers a café experience. However, it does not cater to the user's preferences for breakfast, brunch, pizza, or Mexican food. It is a good option for dessert lovers but lacks the broader appeal the user might be looking for.	8/10
Epi's A Basque Restaurant	French, Basque, Restaurants	This restaurant offers a different cuisine (Basque) that does not align with the user's preferences for American (New), pizza, or Mexican food. While it may provide a good	5/10

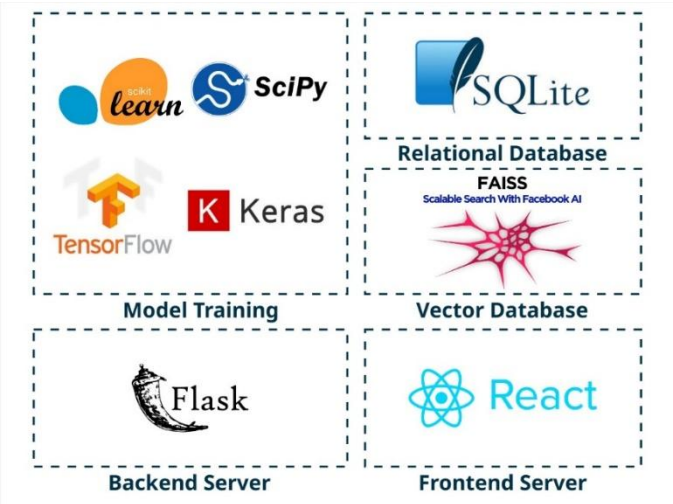
		dining experience, it does not meet the user's specific interests.	
Freakshakes at TeeJay's Sweet Tooth	Yelp Events, Local Flavor	This event focuses on ice cream treats, which may appeal to the user's interest in desserts. However, it does not align with the user's preferences for breakfast, brunch, pizza, or Mexican food. Additionally, it is an event rather than a permanent business, which may limit its relevance.	7/10

User: Elizabeth

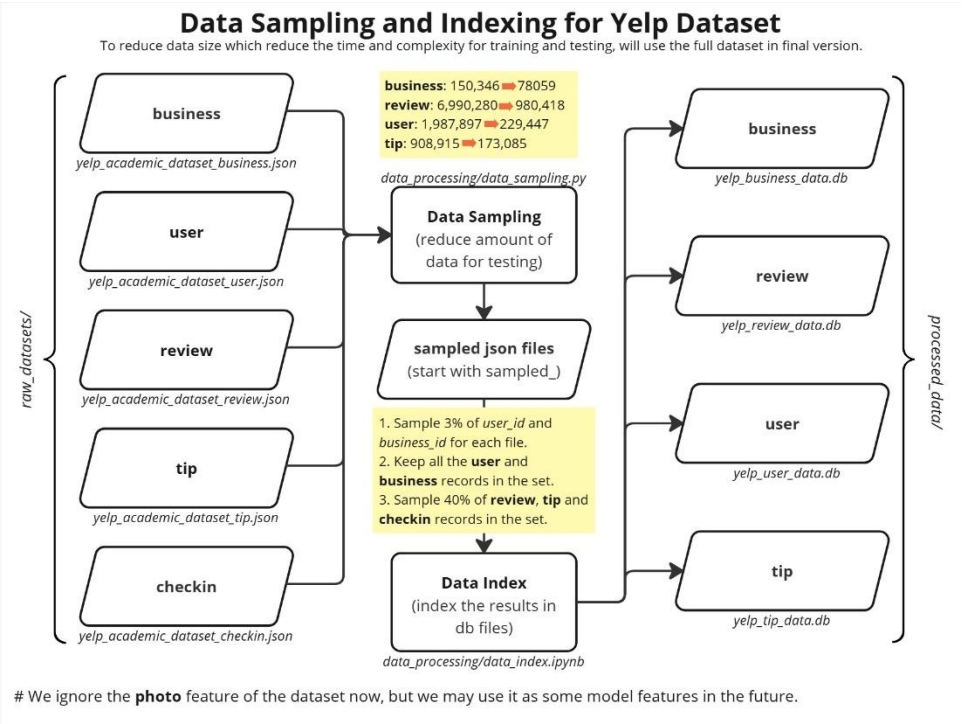
Business	**Categories**	**Alignment Assessment**	**Relevance Score**
Poke' Cafe'	Food, Restaurants, Seafood, Poke, Hawaiian, Sushi Bars	This business focuses on poke and seafood, which does not align with the user's preferences for Vietnamese, sandwiches, burgers, American (Traditional), or pizza. While it may offer fresh ingredients, it does not meet the user's specific dining interests.	5/10
Ice Cave	Food, Shaved Ice, Desserts, Bubble Tea	This business specializes in shaved ice and desserts, which could appeal to the user but does not align with their primary interests in Vietnamese cuisine, burgers, or pizza. It is a dessert-focused establishment and lacks the savory options the user prefers.	6/10
Mayo Ketchup By Plantain Girl	Street Vendors, Food, Dominican, Event Planning & Services, Puerto Rican, Restaurants, Caterers, Caribbean, Cuban	This restaurant offers Caribbean and Dominican cuisine, which may appeal to the user as it includes sandwiches (like the Jibarito) and other savory options. However, it does not directly align with the user's preferences for Vietnamese, burgers, or pizza.	8/10
Pass-A-Grille Beach	Active Life, Event Planning & Services, Hotels & Travel,	This is a beach destination and does not align with the user's dining preferences. While it may be a nice place to visit, it does not cater to the	4/10

	Hotels, Restaurants, Beaches	user's interests in specific food categories.	
Brigtsen's Restaurant	Restaurants, Cajun/Creole, Active Life, Southern	This restaurant offers Cajun and Creole cuisine, which may appeal to the user as it includes hearty, flavorful dishes. However, it does not directly align with the user's preferences for Vietnamese, burgers, or pizza.	7/10

3. *Software Listing*



4. *Sample Down Dataset*



7. References

[1] The Business Research Company, “Content Recommendation Engine Global Market Report 2024,” The Business Research Company, 2024. [Online]. Available: https://www.thebusinessresearchcompany.com/report/content-recommendation-engine-global-market-report. Accessed: Apr. 29, 2025.

[2] Mordor Intelligence, “Recommendation Engine Market Size, Share (2022-2027) | Industry Forecast,” Mordor Intelligence. [Online]. Available: https://www.mordorintelligence.com/industry-reports/recommendation-engine-market. Accessed: Apr. 29, 2025.

[3] D. Roy and M. Dutta, “A systematic review and research perspective on recommender systems,” *Journal of Big Data*, vol. 9, no. 1, May 2022, [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00592-5.

[4] V. AI, “Trends in Emerging Recommender Systems,” *Medium*, Mar. 1, 2023. [Online]. Available: https://medium.com/@ValkyrieAI/trends-in-emerging-recommender-systems-3c779642e364.

[5] Yelp, “Data Licensing – Yelp Open Dataset,” *Yelp for Business*, Jan. 15, 2025. [Online]. Available: https://business.yelp.com/data/resources/open-dataset/.

[6] Google Developers, “Collaborative Filtering | Recommendation Systems,” *Google Developers*. [Online]. Available: https://developers.google.com/machine-learning/recommendation/collaborative/basics.

[7] W. Shusen, “RecommenderSystem,” GitHub repository, 2022. [Online]. Available: https://github.com/wangshusen/RecommenderSystem.

[8] J.-T. Huang *et al*., “Embedding-Based Retrieval in Facebook Search,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2020, [Online]. Available: https://arxiv.org/abs/2006.11632.

[9] tonyctyy, “content-recommendation: This is a project studying about the content recommendation system,” GitHub repository, 2024. [Online]. Available: https://github.com/tonyctyy/content-recommendation.