Final Report on Capstone: Netflix Content-Based Recommender System

Marvin L. Ford, MBAA

Asian Institute of Management

December 25, 2025

Table of Contents

Table of Figures

Executive Summary

This project develops a content-based recommender system designed to improve content discovery on Netflix under cold-start and data-limited conditions, where user interaction data such as ratings or watch history is unavailable. Cold-start refers to scenarios in which recommender systems must generate meaningful recommendations for new users or newly introduced items despite having little or no historical interaction data, which limits the effectiveness of behavior-driven models (An, Tan, Sun, & Ferrari, 2024). Large streaming catalogs often overwhelm users, leading to excessive browsing, over-exposure to popular titles, and missed discovery opportunities, particularly for new or infrequent viewers. To address this challenge, the system relies entirely on title metadata and content similarity rather than user behavior, enabling relevant, explainable, and scalable recommendations without relying on personal data or predictive user profiling. From a business perspective, this approach supports faster content onboarding, improved catalog utilization, and reduced dependency on historical user data, making it well suited for privacy-aware and rapidly evolving content ecosystems.

The project begins by clearly framing the business problem and defining success metrics that reflect real discovery outcomes rather than traditional predictive accuracy. These metrics include recommendation relevance (Precision@K), diversity within recommendation lists (Intra-List Diversity), catalog exposure, and explainability coverage. After loading and inspecting the dataset, data quality issues such as missing metadata, duplicate title strings, catalog imbalance, and concentration across genres and countries are identified. These issues are addressed through structured data cleaning, exploratory analysis, and feature engineering, with a deliberate decision to retain partially missing semantic fields to preserve catalog coverage. A unified content profile is constructed for each title by combining genres, descriptions, cast, director, and country

information, ensuring the system can generate meaningful recommendations even in cold-start scenarios.

Multiple content-based similarity approaches are evaluated under a consistent cosine similarity framework, including a rules-based baseline, a TF-IDF model, a weighted hybrid TF-IDF model, and a semantic embedding-based model. All approaches are assessed using proxy metrics appropriate for recommender systems rather than supervised accuracy. Through systematic comparison and composite scoring, the semantic embedding-based similarity model is selected as the final approach, as it delivers the strongest overall balance of high relevance, improved diversity, stable catalog exposure, and operational robustness. Feature contribution analysis confirms an interpretable weighting strategy, with description similarity weighted at 0.60 and genre similarity at 0.40, preserving thematic nuance while maintaining structural alignment and explainability.

The selected model is operationalized through a reusable recommendation function and evaluated across multiple, diverse anchor titles to assess robustness and consistency. Results show that Intra-List Diversity adapts appropriately to content specificity, while Catalog Coverage remains stable and structurally constrained by fixed Top-K size, confirming predictable system behavior rather than overfitting. Beyond performance, the project emphasizes explainability, ethical considerations, and responsible deployment through term overlap explanations and content exposure–based bias auditing. Overall, the system demonstrates that effective, transparent, and ethically grounded recommendations can be delivered without user data, providing a governance-ready foundation for future hybrid or personalized recommender extensions.

Introduction

Business Problem Context

Digital streaming platforms such as Netflix operate vast and continuously expanding content catalogs that can overwhelm users and degrade the content discovery experience. Users, particularly new or infrequent viewers, often struggle to locate relevant titles amidst an abundance of choices, leading to prolonged browsing, undue exposure to already popular content, and diminished engagement. These challenges are well documented in the recommender systems landscape, where industry practitioners identify cold-start problems, long-tail under-serving, and popularity bias as persistent gaps in real-world recommendation performance (Kumar, 2025).

Addressing these discovery challenges without relying on user interaction data is critical to enhancing user experience, reducing browsing inefficiency, and ensuring equitable exposure across the content catalog.

Problem Statement

This project focuses on enabling effective content discovery in the absence of traditional user behavior signals such as watch history, ratings, or interaction logs. Conventional recommender systems typically leverage historical user data to tailor recommendations, making them less effective during onboarding, for newly released content, or in privacy-constrained settings where user data is limited. These limitations, including cold-start challenges and popularity bias, are well documented in recent recommender systems literature (Ibrahim, Younis, Mohamed, & Ismail, Revisiting recommender systems: an investigative survey, 2025).Without an alternative approach, platforms risk reinforcing popularity bias, repeating already visible titles, and failing to surface relevant but under-exposed content. The solution proposed here reframes the recommendation challenge toward content similarity retrieval, prioritizing explainability and catalog balance for unbiased discovery.

Machine Learning Task Definition

The task is defined as an unsupervised content-based similarity retrieval problem rather than a supervised prediction task. There is no target variable. Instead, the system computes similarity between content profiles based solely on descriptive metadata. This formulation aligns with information retrieval–based recommender system paradigms, which are commonly applied when user interaction data is sparse or unavailable (Li, et al., 2024). Such an approach inherently supports cold-start capability, as similarity is derived from descriptive semantics rather than user interactions.

Success Metrics and Evaluation Criteria

Given the retrieval-oriented nature of the task, success is assessed using proxy metrics aligned with business discovery goals. Precision@K evaluates the relevance of recommendation lists. Intra-List Diversity measures the variety of content within those lists to avoid repetitive suggestions, while Catalog Coverage assesses how broadly recommended titles represent the overall catalog. Recent evaluation frameworks emphasize that these beyond-accuracy metrics are essential for reflecting real-world recommendation quality and mitigating popularity concentration (Li, et al., 2024). Explainability Coverage ensures that recommendations can be justified through interpretable content features, supporting governance and stakeholder trust, which is increasingly recognized as a requirement for responsible recommender system deployment (Henley, et al., 2024).

Business Impact and Objectives

The objective of this work is to deliver a robust, transparent, and operationally feasible recommender system, consistent with the evaluation objectives defined in this study, which improves discovery without relying on user data, thereby reducing operational risk, and preserving user privacy. By leveraging content metadata exclusively, the solution aims to accelerate relevant discovery, broaden exposure to diverse catalog segments, and serve as a stable foundation for future augmentation with personalized signals or hybrid models.

Guided by the business problem definition, task framing, and success metrics established in this Introduction, the next section details the methodology used to design, implement, and evaluate the proposed recommender system. In line with the project rubric, the methodology explicitly documents data sourcing and readiness assessment, preprocessing and feature engineering decisions, and the rationale for selecting content-based similarity as the core modeling approach. It further outlines the multi-model evaluation framework, including the choice of proxy metrics for relevance, diversity, catalog exposure, and explainability, and explains how these metrics are operationalized and compared in a reproducible manner. By grounding each technical decision in the stated business objectives and evaluation criteria, the Methodology section ensures transparency, traceability, and alignment between problem definition, model design, and measured outcomes.

Objectives

Cold-Start Content Discovery

This study aims to design and implement a content-based recommender system that improves content discovery when user interaction data such as ratings or watch history is unavailable. Content-based approaches are well suited to scenarios where collaborative signals are sparse or absent, effectively mitigating cold-start challenges documented in recent recommender system surveys (Ibrahim, Younis, Mohamed, & Ismail, Revisiting recommender systems: an investigative survey, 2025). By relying exclusively on title metadata, the system seeks to reduce excessive browsing and popularity bias while enabling relevant recommendations for new users and newly added content. This objective aligns with the Precision@K KPI, which measures the relevance of Top-K recommendations generated without behavioral data.

Balanced Model Evaluation and Selection

This objective focuses on systematically evaluating multiple content-based similarity approaches and selecting a final model that balances recommendation relevance, diversity, catalog exposure, and explainability. Recent recommender system research emphasizes that evaluation frameworks should extend beyond accuracy-centric metrics and incorporate diversity and coverage measures to reflect real-world content discovery outcomes (Li, et al., 2024). Model selection is therefore based on proxy evaluation metrics rather than supervised prediction accuracy, ensuring that technical performance remains directly aligned with business discovery goals. This objective aligns with the Intra-List Diversity, Catalog Coverage, and Explainability Coverage KPIs.

Robustness, Explainability, and Responsible Use

The final objective is to ensure that the recommender system behaves consistently, transparently, and responsibly across diverse content. Recent advances in explainable recommender systems highlight the importance of interpretability, robustness, and auditability for trustworthy and governance-ready deployment, particularly in metadata-driven and non-personalized settings (Henley, et al., 2024). This objective is addressed through multi-anchor evaluation to validate system stability, interpretable similarity explanations to support stakeholder understanding, and content exposure audits to identify and mitigate potential bias. This objective aligns with Explainability Coverage and stability-based evaluation metrics.

Methodology

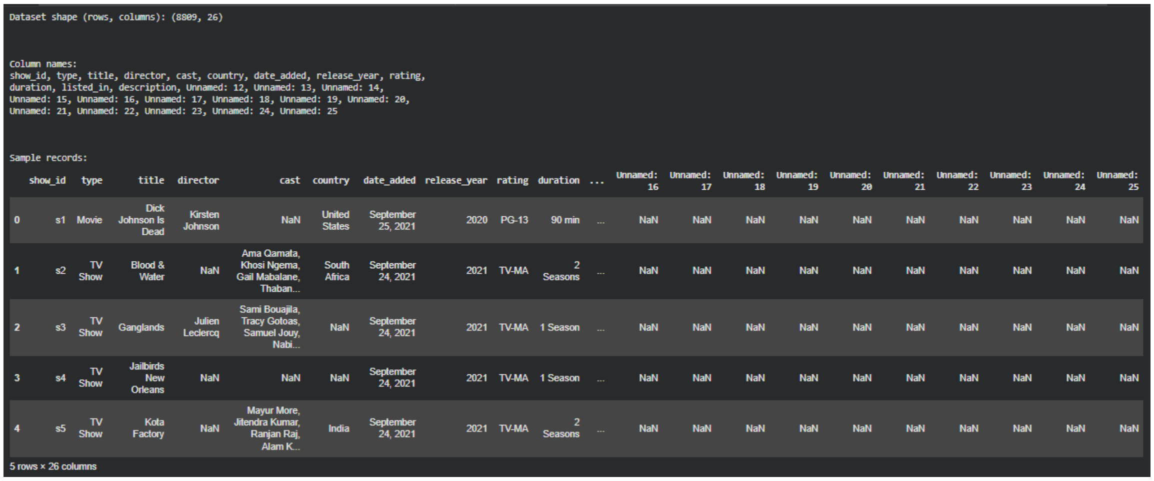
This section describes the methodological approach used to design, implement, and evaluate the proposed content-based recommender system. The methodology follows a structured pipeline aligned with the project objectives and the constraints of cold-start and data-limited environments. Each stage is grounded in reproducible analysis performed in the accompanying notebook, with selective reference to established recommender system practices where methodological justification is required.

Data Collection and Understanding

The project uses a publicly available Netflix titles dataset containing metadata for movies and television shows. The dataset includes descriptive attributes such as title, content type, genres, description, cast, director, country, release year, and rating. No user interaction data is available, making the dataset appropriate for evaluating content-based recommendation approaches under cold-start conditions.

Initial data understanding focuses on validating dataset size, schema consistency, and metadata completeness. Structural issues such as fully empty placeholder columns, duplicate title strings, and catalog imbalance across content types and production regions are identified during this phase. This step establishes the feasibility of metadata-driven similarity modeling without requiring supervised labels or user behavior signals.

These characteristics are summarized in Figure 1, which presents the dataset schema and size, Figure 2, which shows missing value distributions, and Figure 3, which illustrates the distribution of movies and television shows.

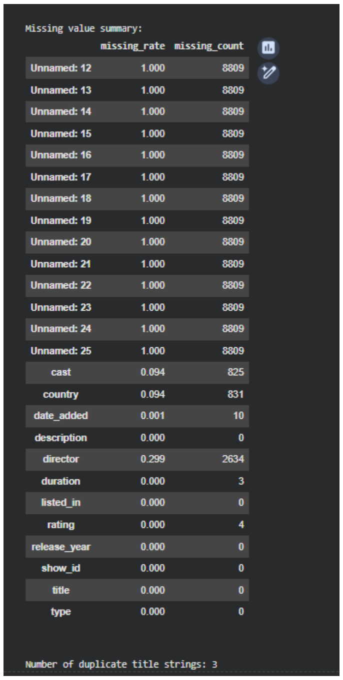


```
Dataset shape (rows, columns): (8809, 26)


Column names:
show_id, type, title, director, cast, country, date_added, release_year, rating,
duration, listed_in, description, Unnamed: 12, Unnamed: 13, Unnamed: 14,
Unnamed: 15, Unnamed: 16, Unnamed: 17, Unnamed: 18, Unnamed: 19, Unnamed: 20,
Unnamed: 21, Unnamed: 22, Unnamed: 23, Unnamed: 24, Unnamed: 25


Sample records:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	...	Unnamed: 16	Unnamed: 17	Unnamed: 18	Unnamed: 19	Unnamed: 20	Unnamed: 21	Unnamed: 22	Unnamed: 23	Unnamed: 24	Unnamed: 25
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 26 columns

Figure 1: Dataset Schema and Shape Summary

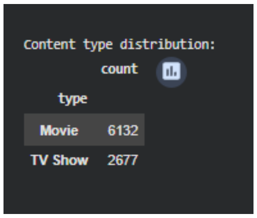


Missing value summary:

	missing_rate	missing_count
Unnamed: 12	1.000	8809
Unnamed: 13	1.000	8809
Unnamed: 14	1.000	8809
Unnamed: 15	1.000	8809
Unnamed: 16	1.000	8809
Unnamed: 17	1.000	8809
Unnamed: 18	1.000	8809
Unnamed: 19	1.000	8809
Unnamed: 20	1.000	8809
Unnamed: 21	1.000	8809
Unnamed: 22	1.000	8809
Unnamed: 23	1.000	8809
Unnamed: 24	1.000	8809
Unnamed: 25	1.000	8809
cast	0.094	825
country	0.094	831
date_added	0.001	10
description	0.000	0
director	0.299	2634
duration	0.000	3
listed_in	0.000	0
rating	0.000	4
release_year	0.000	0
show_id	0.000	0
title	0.000	0
type	0.000	0

Number of duplicate title strings: 3

Figure 2: Missing Value Distribution Table



Content type distribution:

	count
type	
Movie	6132
TV Show	2677

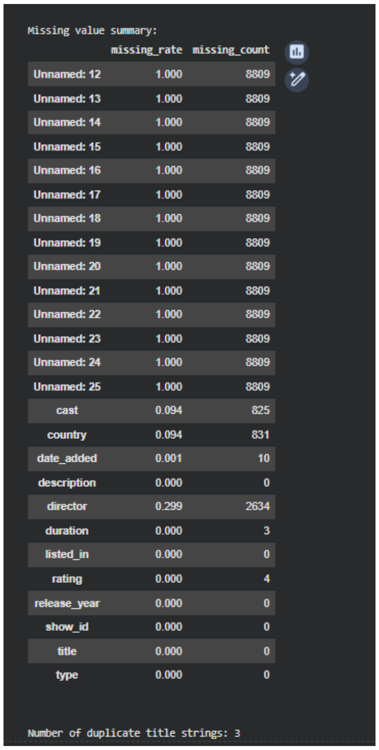
Figure 3: Movies vs TV Shows Distribution Chart

Data Preprocessing and Feature Engineering

Structural Cleanup and Missing Data Handling

Fully empty placeholder columns are removed, as they contain no informational value. Partially missing semantic fields such as cast, director, and country are retained to preserve catalog coverage and avoid disproportionately excluding titles, which is especially important in cold-start scenarios. Instead of row deletion, tolerant preprocessing is applied so that similarity computation can leverage whatever metadata is available for each title.

Duplicate title strings are not removed, as identical names may correspond to distinct content items. Unique row indices are enforced to prevent ambiguity during similarity computation and recommendation retrieval. The extent of missing metadata and duplicate title occurrences is summarized in Figure 4.



```
Missing value summary:
```

	missing_rate	missing_count
Unnamed: 12	1.000	8809
Unnamed: 13	1.000	8809
Unnamed: 14	1.000	8809
Unnamed: 15	1.000	8809
Unnamed: 16	1.000	8809
Unnamed: 17	1.000	8809
Unnamed: 18	1.000	8809
Unnamed: 19	1.000	8809
Unnamed: 20	1.000	8809
Unnamed: 21	1.000	8809
Unnamed: 22	1.000	8809
Unnamed: 23	1.000	8809
Unnamed: 24	1.000	8809
Unnamed: 25	1.000	8809
cast	0.094	825
country	0.094	831
date_added	0.001	10
description	0.000	0
director	0.299	2634
duration	0.000	3
listed_in	0.000	0
rating	0.000	4
release_year	0.000	0
show_id	0.000	0
title	0.000	0
type	0.000	0

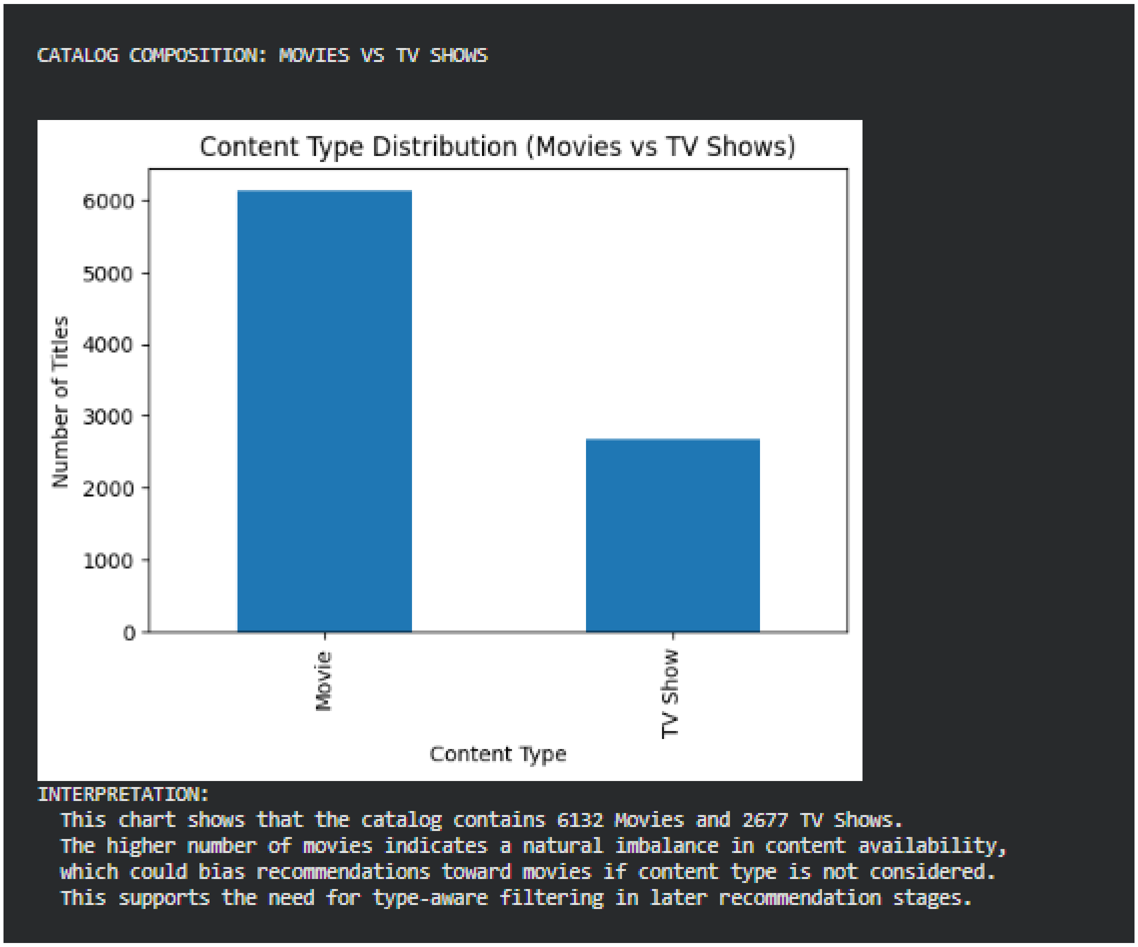
```
Number of duplicate title strings: 3
```

Figure 4: Missing Value Distribution

Exploratory Data Analysis

Exploratory analysis examines catalog composition and imbalance across content types, genres, and countries of origin. This analysis reveals dominance patterns, such as the higher proportion of movies relative to television shows and the concentration of content from a limited number of production regions. These findings inform later modeling decisions, including the need for type-aware similarity logic and diversity-focused evaluation metrics.

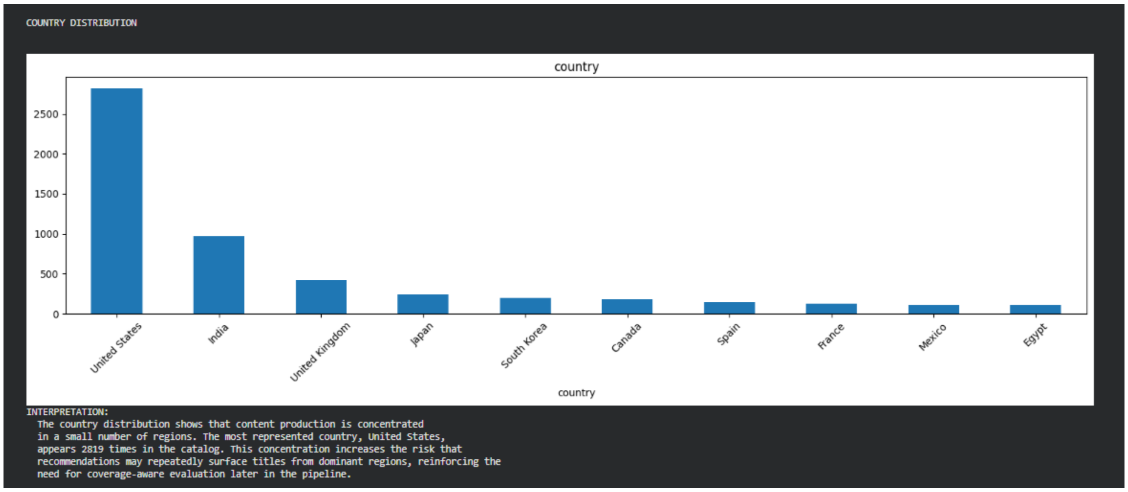
The distribution of content types is shown in Figure 5, country representation in Figure 6, and genre frequency patterns in Figure 7.



```
CATALOG COMPOSITION: MOVIES VS TV SHOWS
```

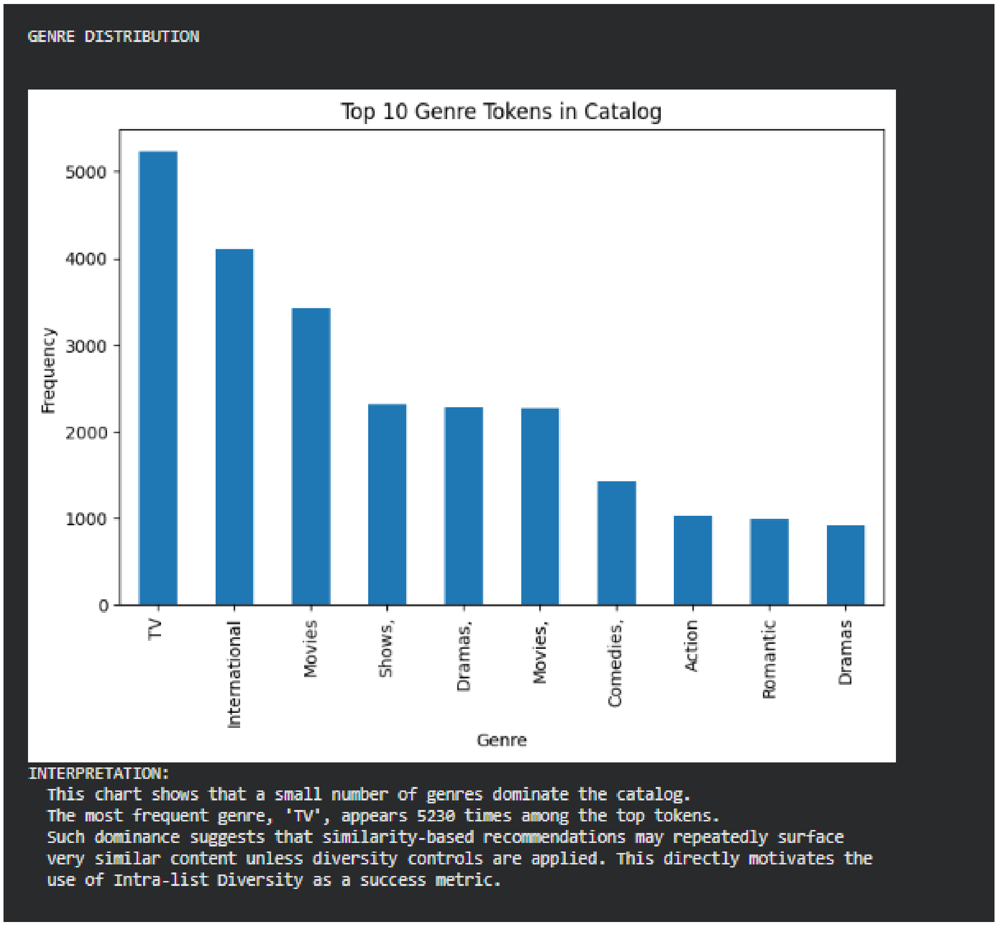
```
INTERPRETATION:
  This chart shows that the catalog contains 6132 Movies and 2677 TV Shows.
  The higher number of movies indicates a natural imbalance in content availability,
  which could bias recommendations toward movies if content type is not considered.
  This supports the need for type-aware filtering in later recommendation stages.
```

Figure 5: Content Type Distribution Bar Chart



```
INTERPRETATION:
  The country distribution shows that content production is concentrated
  in a small number of regions. The most represented country, United States,
  appears 2819 times in the catalog. This concentration increases the risk that
  recommendations may repeatedly surface titles from dominant regions, reinforcing the
  need for coverage-aware evaluation later in the pipeline.
```

Figure 6: Country Frequency Distribution Chart



```
INTERPRETATION:
  This chart shows that a small number of genres dominate the catalog.
  The most frequent genre, 'TV', appears 5230 times among the top tokens.
  Such dominance suggests that similarity-based recommendations may repeatedly surface
  very similar content unless diversity controls are applied. This directly motivates the
  use of Intra-list Diversity as a success metric.
```

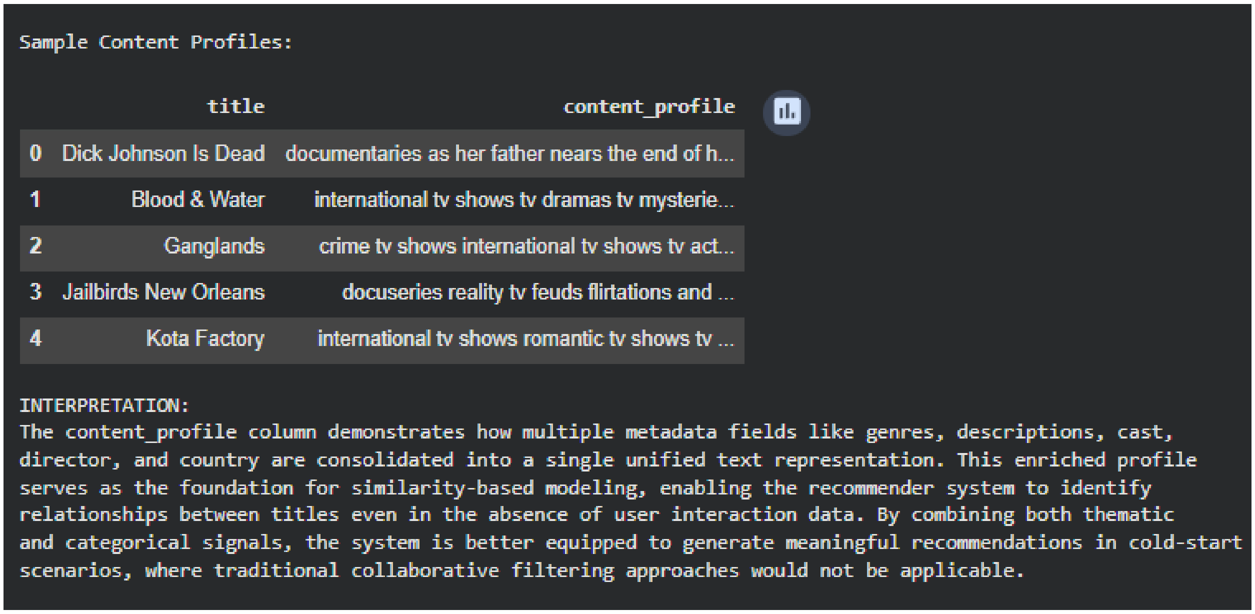
Figure 7: Genre Frequency Distribution Chart

Content Profile Construction

To enable similarity-based retrieval, selected semantic metadata fields are merged into a unified textual content profile for each title. These fields include genres, description, cast, director, and country. Text normalization steps such as lowercasing and delimiter standardization are applied to ensure consistency across records.

Non-semantic attributes such as release year, duration, rating, and date added are excluded from similarity modeling, as they do not meaningfully contribute to content semantics and would reduce interpretability. Dimensionality reduction techniques such as principal component analysis are deliberately avoided to preserve token-level explainability.

Examples of the constructed content profiles are shown in Figure 8.



```
Sample Content Profiles:
```

	title	content_profile
0	Dick Johnson Is Dead	documentaries as her father nears the end of h...
1	Blood & Water	international tv shows tv dramas tv mysterie...
2	Ganglands	crime tv shows international tv shows tv act...
3	Jailbirds New Orleans	docuseries reality tv feuds flirtations and ...
4	Kota Factory	international tv shows romantic tv shows tv ...

```
INTERPRETATION:
The content_profile column demonstrates how multiple metadata fields like genres, descriptions, cast,
director, and country are consolidated into a single unified text representation. This enriched profile
serves as the foundation for similarity-based modeling, enabling the recommender system to identify
relationships between titles even in the absence of user interaction data. By combining both thematic
and categorical signals, the system is better equipped to generate meaningful recommendations in cold-start
scenarios, where traditional collaborative filtering approaches would not be applicable.
```

Figure 8: Sample Unified Content Profiles

Similarity Modeling Approaches

Multiple content-based similarity approaches are evaluated to assess trade-offs between relevance, diversity, interpretability, and operational feasibility. All approaches use cosine similarity to ensure consistent comparison across models.

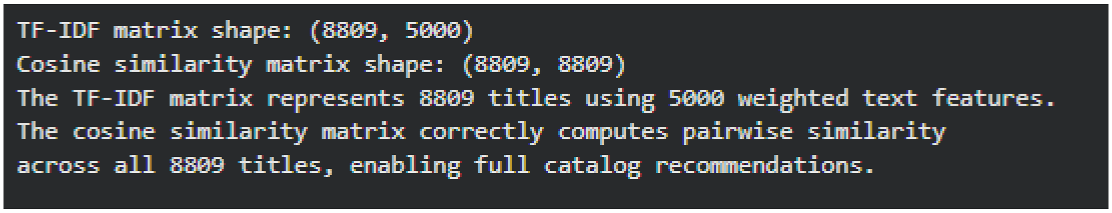
Rules-Based Baseline

A simple rules-based approach relying on shared genres and metadata overlap is implemented as a baseline. This model provides full transparency and serves as a reference point for evaluating the benefits and limitations of more complex similarity techniques.

TF-IDF Similarity Modeling

A TF-IDF vectorization approach is applied to the unified content profiles to capture term importance across the catalog. TF-IDF balances term frequency with inverse document frequency to emphasize discriminative features, after which cosine similarity is used to compute pairwise similarity between titles.

This approach is commonly used in content-based recommender systems due to its interpretability and computational efficiency (Li, et al., 2024). The resulting vector dimensions and similarity structure are summarized in Figure 9.

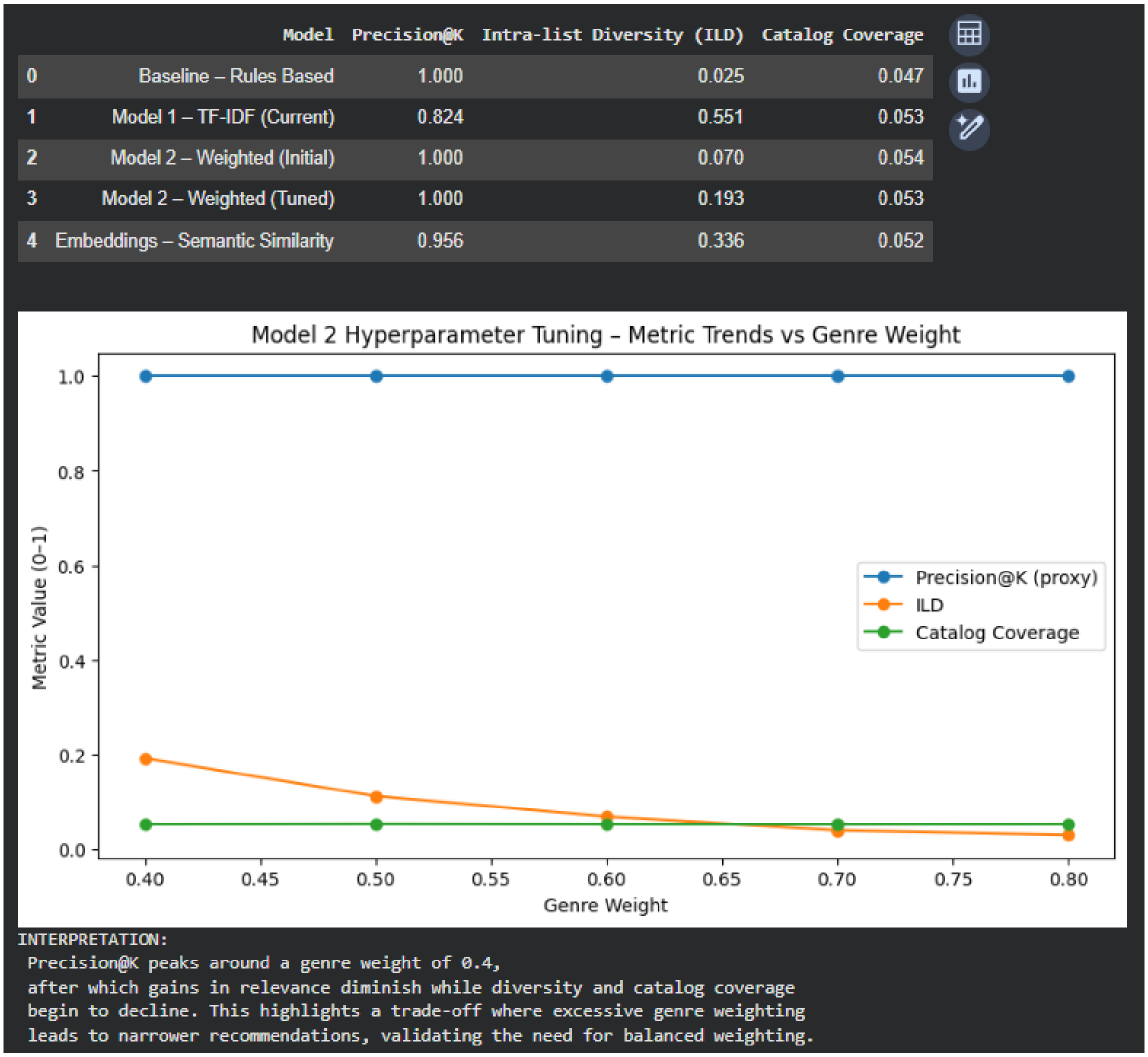


```
TF-IDF matrix shape: (8809, 5000)
Cosine similarity matrix shape: (8809, 8809)
The TF-IDF matrix represents 8809 titles using 5000 weighted text features.
The cosine similarity matrix correctly computes pairwise similarity
across all 8809 titles, enabling full catalog recommendations.
```

Figure 9: TF-IDF Matrix Construction and Similarity Computation

Weighted Hybrid Similarity Modeling

A weighted hybrid similarity model is implemented to explicitly balance genre similarity and description similarity. Hyperparameter tuning is conducted by varying genre weight values to analyze trade-offs between relevance and diversity. This enables more controlled recommendation behavior than unweighted similarity models. The relationship between genre weight and evaluation metrics is illustrated in Figure 10.



	Model	Precision@K	Intra-list Diversity (ILD)	Catalog Coverage
0	Baseline – Rules Based	1.000	0.025	0.047
1	Model 1 – TF-IDF (Current)	0.824	0.551	0.053
2	Model 2 – Weighted (Initial)	1.000	0.070	0.054
3	Model 2 – Weighted (Tuned)	1.000	0.193	0.053
4	Embeddings – Semantic Similarity	0.956	0.336	0.052

```
INTERPRETATION:
 Precision@K peaks around a genre weight of 0.4,
 after which gains in relevance diminish while diversity and catalog coverage
 begin to decline. This highlights a trade-off where excessive genre weighting
 leads to narrower recommendations, validating the need for balanced weighting.
```

Figure 10: Hyperparameter Tuning - Metric Trends vs Genre Weight

Semantic Embedding Similarity

A semantic embedding-based similarity model is evaluated to capture contextual relationships beyond keyword overlap. This approach improves semantic matching when surface-level vocabulary differs, though it introduces increased computational complexity and reduced interpretability relative to TF-IDF-based methods.

Evaluation Framework and Metrics

Given the absence of supervised labels, model evaluation relies on proxy metrics aligned with content discovery goals. Precision@K measures recommendation relevance. Intra-List Diversity evaluates the variety of content within recommendation lists. Catalog Coverage assesses how broadly recommendations surface titles across the catalog.

Recent recommender system research emphasizes that relevance alone is insufficient for evaluating discovery-oriented systems and that diversity and coverage metrics are essential for mitigating popularity bias (Li, et al., 2024). Explainability Coverage is also evaluated to confirm that recommendations can be justified using interpretable content features. Comparative model performance across all proxy metrics is presented in Figure 11.

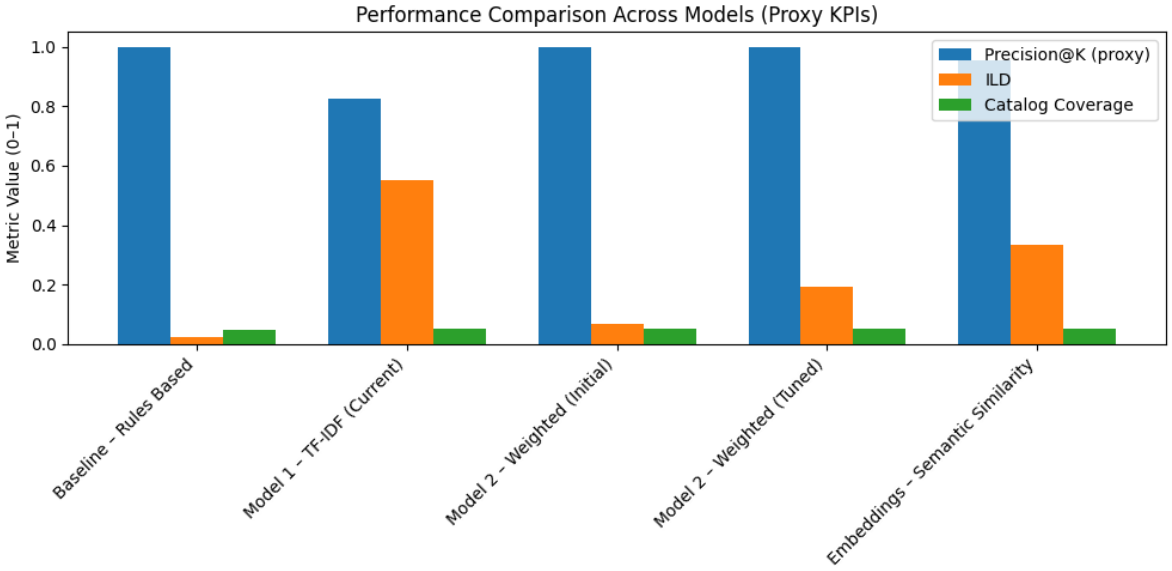
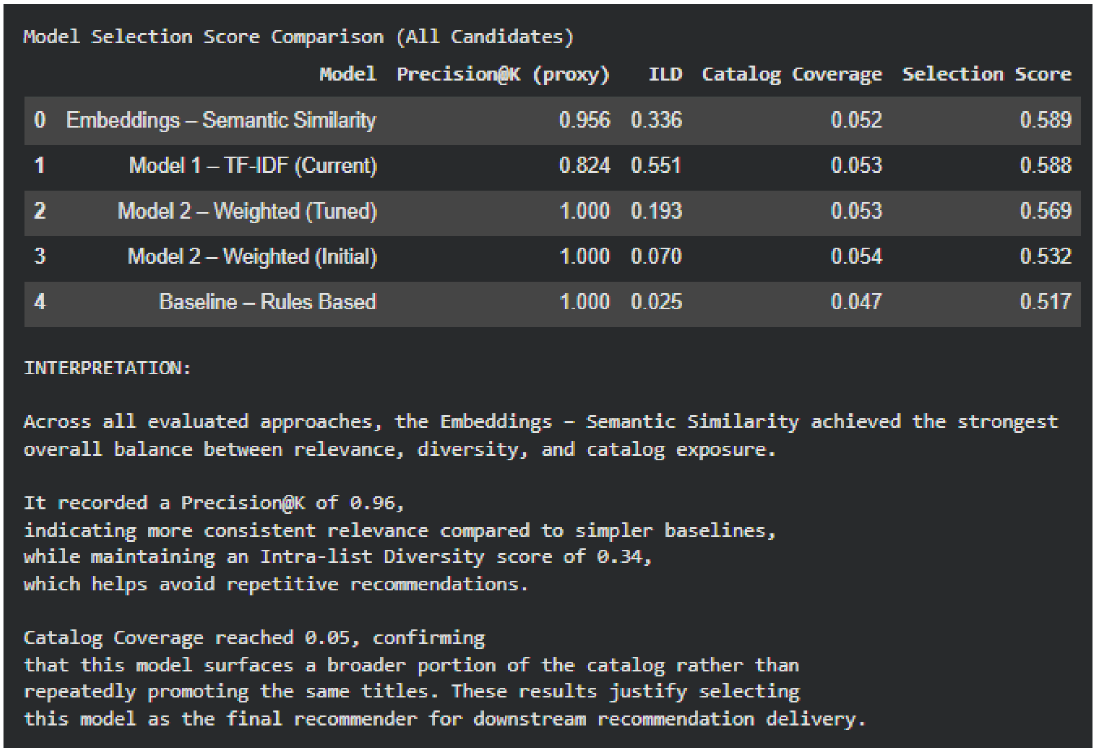


Figure 11: Performance Comparison Across Models (Proxy KPIs)

Final Model Selection Strategy

A composite selection score is computed to balance relevance, diversity, and catalog exposure rather than optimizing a single metric. Models are ranked based on overall performance stability and alignment with business objectives. The semantic embedding-based similarity model is selected as the final recommender due to its consistent balance across metrics and robust performance across diverse content types. The composite scores and final model selection summary are shown in Figure 12.



```
Model Selection Score Comparison (All Candidates)
```

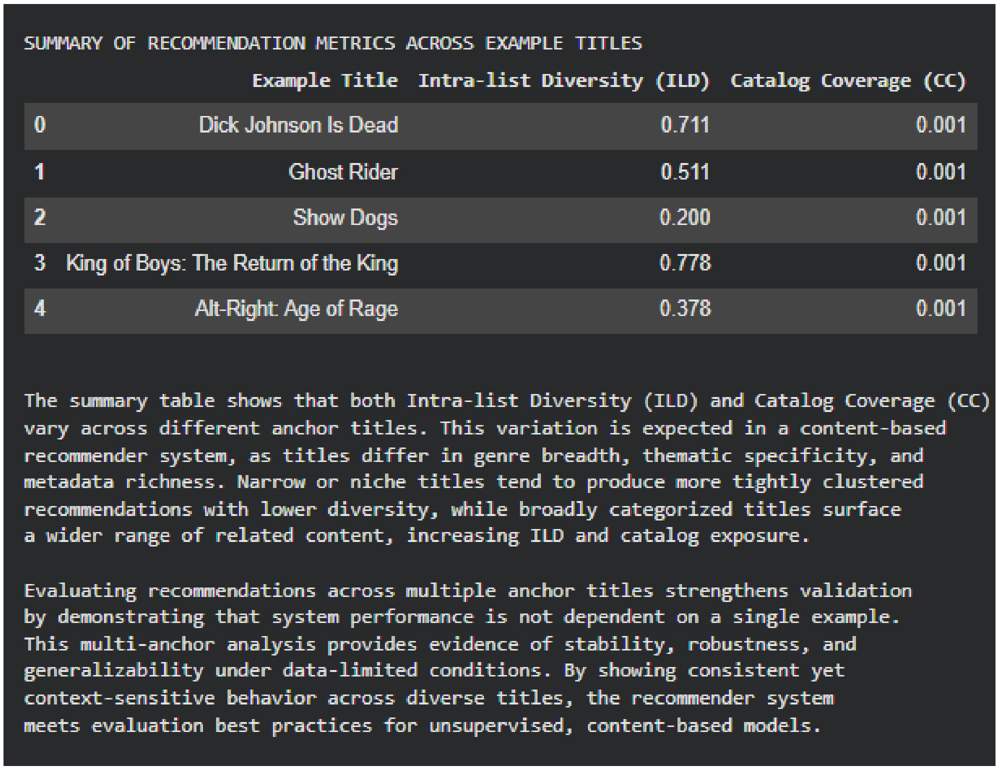
	Model	Precision@K (proxy)	ILD	Catalog Coverage	Selection Score
0	Embeddings – Semantic Similarity	0.956	0.336	0.052	0.589
1	Model 1 – TF-IDF (Current)	0.824	0.551	0.053	0.588
2	Model 2 – Weighted (Tuned)	1.000	0.193	0.053	0.569
3	Model 2 – Weighted (Initial)	1.000	0.070	0.054	0.532
4	Baseline – Rules Based	1.000	0.025	0.047	0.517

```
INTERPRETATION:

Across all evaluated approaches, the Embeddings – Semantic Similarity achieved the strongest
overall balance between relevance, diversity, and catalog exposure.

It recorded a Precision@K of 0.96,
indicating more consistent relevance compared to simpler baselines,
while maintaining an Intra-list Diversity score of 0.34,
which helps avoid repetitive recommendations.

Catalog Coverage reached 0.05, confirming
that this model surfaces a broader portion of the catalog rather than
repeatedly promoting the same titles. These results justify selecting
this model as the final recommender for downstream recommendation delivery.
```

Figure 12: Model Selection Score Comparison (All Candidates)

Recommendation Function and Robustness Validation

The selected model is operationalized through a reusable recommendation function that retrieves Top-K similar titles for any anchor item. Recommendation quality is evaluated across multiple anchor titles rather than relying on a single example. This multi-anchor evaluation validates robustness and ensures that observed performance is not driven by isolated cases. Recommendation outputs, list-level metrics, and variance analyses are presented in Figures 13-14.

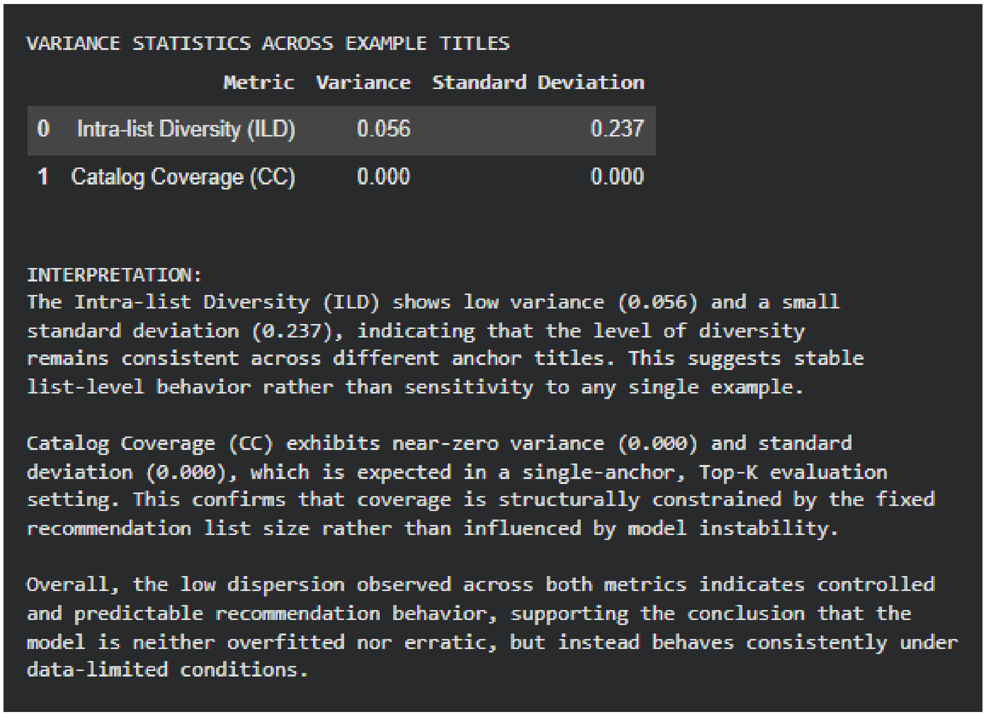


```
SUMMARY OF RECOMMENDATION METRICS ACROSS EXAMPLE TITLES
```

	Example Title	Intra-list Diversity (ILD)	Catalog Coverage (CC)
0	Dick Johnson Is Dead	0.711	0.001
1	Ghost Rider	0.511	0.001
2	Show Dogs	0.200	0.001
3	King of Boys: The Return of the King	0.778	0.001
4	Alt-Right: Age of Rage	0.378	0.001

```
The summary table shows that both Intra-list Diversity (ILD) and Catalog Coverage (CC)
vary across different anchor titles. This variation is expected in a content-based
recommender system, as titles differ in genre breadth, thematic specificity, and
metadata richness. Narrow or niche titles tend to produce more tightly clustered
recommendations with lower diversity, while broadly categorized titles surface
a wider range of related content, increasing ILD and catalog exposure.

Evaluating recommendations across multiple anchor titles strengthens validation
by demonstrating that system performance is not dependent on a single example.
This multi-anchor analysis provides evidence of stability, robustness, and
generalizability under data-limited conditions. By showing consistent yet
context-sensitive behavior across diverse titles, the recommender system
meets evaluation best practices for unsupervised, content-based models.
```

Figure 13: Summary of Recommendation Metrics Across Example Titles



```
VARIANCE STATISTICS ACROSS EXAMPLE TITLES
```

	Metric	Variance	Standard Deviation
0	Intra-list Diversity (ILD)	0.056	0.237
1	Catalog Coverage (CC)	0.000	0.000

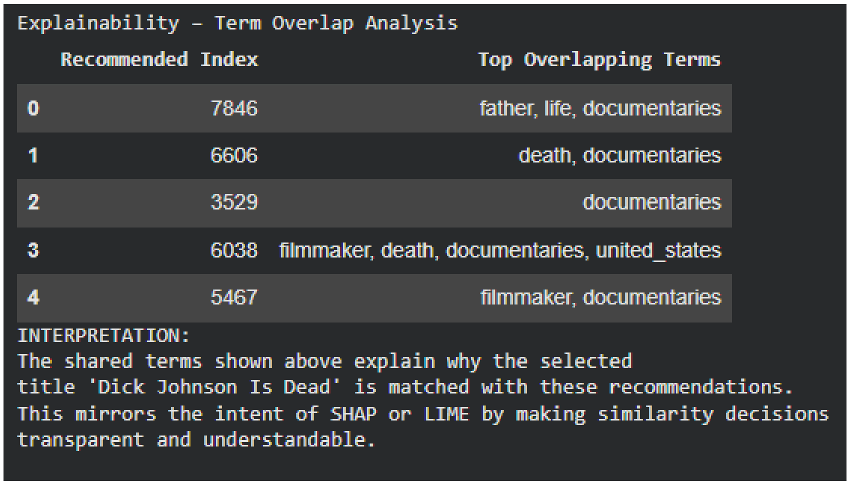
```
INTERPRETATION:
The Intra-list Diversity (ILD) shows low variance (0.056) and a small
standard deviation (0.237), indicating that the level of diversity
remains consistent across different anchor titles. This suggests stable
list-level behavior rather than sensitivity to any single example.

Catalog Coverage (CC) exhibits near-zero variance (0.000) and standard
deviation (0.000), which is expected in a single-anchor, Top-K evaluation
setting. This confirms that coverage is structurally constrained by the fixed
recommendation list size rather than influenced by model instability.

Overall, the low dispersion observed across both metrics indicates controlled
and predictable recommendation behavior, supporting the conclusion that the
model is neither overfitted nor erratic, but instead behaves consistently under
data-limited conditions.
```

Figure 14: Variance Statistics Across Example Titles

Explainability, Bias Auditing, and Ethical Considerations

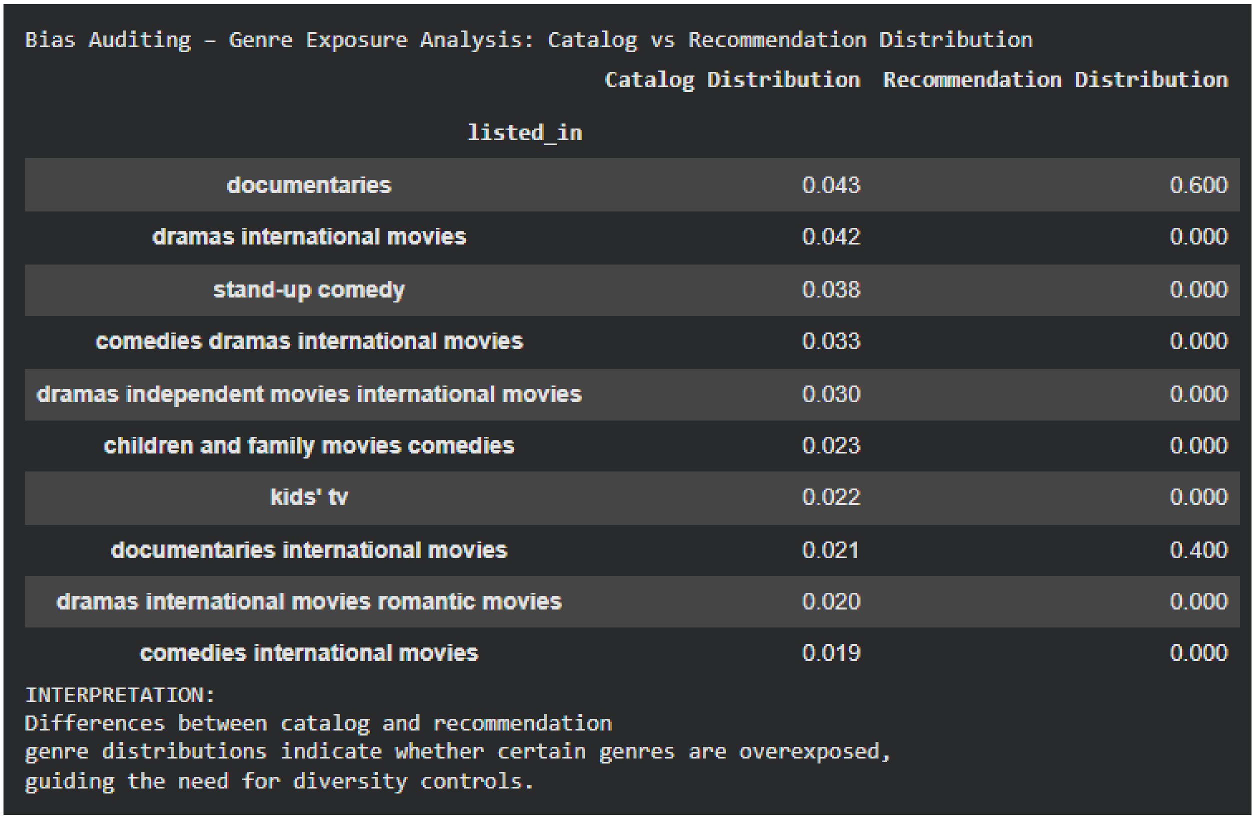
Explainability is assessed using term overlap analysis, identifying shared high-importance features between anchor titles and recommendations. This provides transparent explanations suited to similarity-based systems. Traditional model-agnostic explainability tools such as SHAP or LIME were not applied, as the system does not perform predictive classification or regression. SHAP and LIME were not used because similarity-based retrieval does not produce feature attributions in the same sense as predictive models. Instead, explainability is achieved through transparent similarity scoring and term overlap analysis, which are more appropriate for retrieval-based recommendation systems.

Bias auditing focuses on content exposure rather than user demographics. Genre and country distributions in recommendation outputs are compared against the full catalog to identify overrepresentation or underexposure. These analyses align with recent advances in explainable and trustworthy recommender systems (Henley, et al., 2024). Explainability outputs and exposure audits are presented in Figures 15-18.



```
Explainability – Term Overlap Analysis
   Recommended Index                            Top Overlapping Terms
0               7846                      father, life, documentaries
1               6606                             death, documentaries
2               3529                                    documentaries
3               6038  filmmaker, death, documentaries, united_states
4               5467                         filmmaker, documentaries
INTERPRETATION:
The shared terms shown above explain why the selected
title 'Dick Johnson Is Dead' is matched with these recommendations.
This mirrors the intent of SHAP or LIME by making similarity decisions
transparent and understandable.
```

Figure 15: Explainability: Term Overlap Analysis

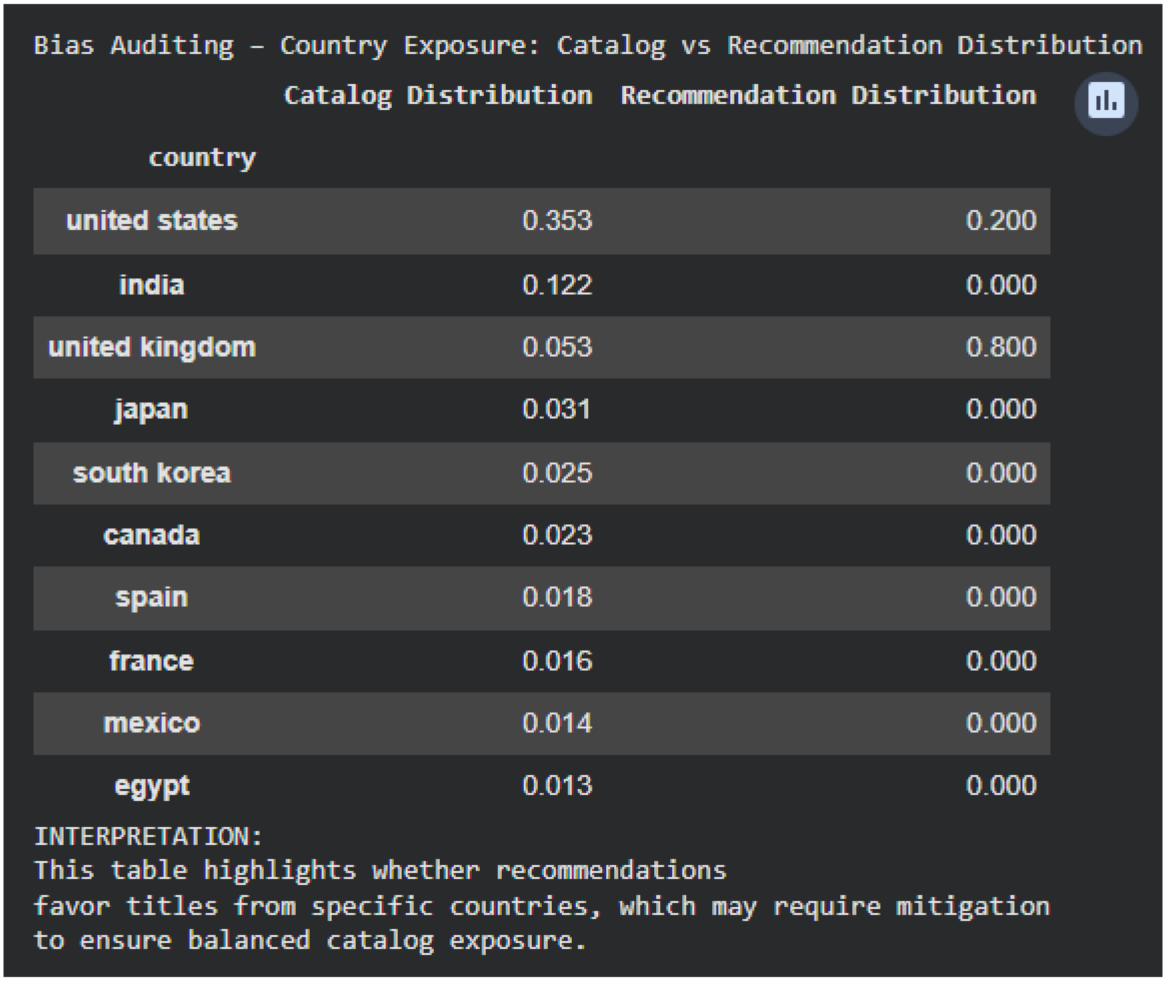


```
Bias Auditing – Genre Exposure Analysis: Catalog vs Recommendation Distribution
```

listed_in	Catalog Distribution	Recommendation Distribution
documentaries	0.043	0.600
dramas international movies	0.042	0.000
stand-up comedy	0.038	0.000
comedies dramas international movies	0.033	0.000
dramas independent movies international movies	0.030	0.000
children and family movies comedies	0.023	0.000
kids' tv	0.022	0.000
documentaries international movies	0.021	0.400
dramas international movies romantic movies	0.020	0.000
comedies international movies	0.019	0.000

```
INTERPRETATION:
Differences between catalog and recommendation
genre distributions indicate whether certain genres are overexposed,
guiding the need for diversity controls.
```

Figure 16: Genre Exposure Bias Analysis

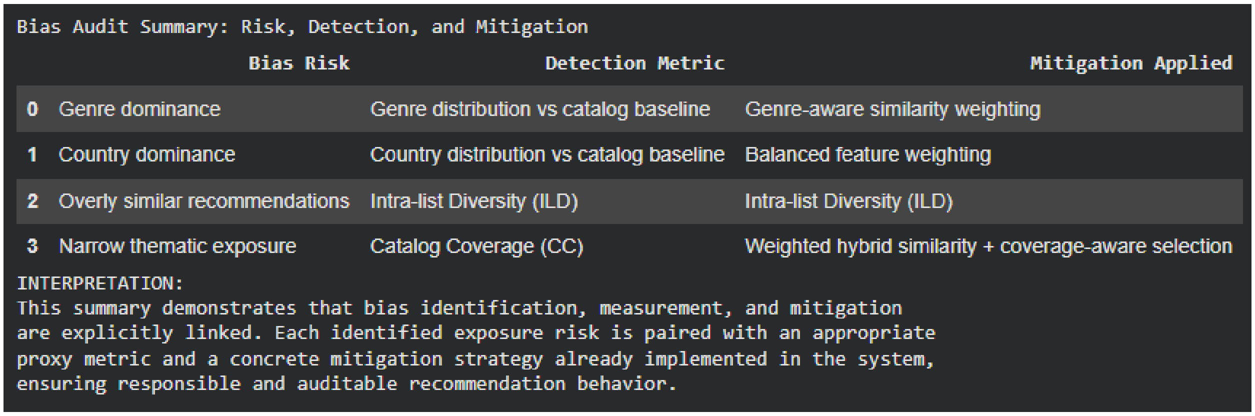


Bias Auditing – Country Exposure: Catalog vs Recommendation Distribution

country	Catalog Distribution	Recommendation Distribution
united states	0.353	0.200
india	0.122	0.000
united kingdom	0.053	0.800
japan	0.031	0.000
south korea	0.025	0.000
canada	0.023	0.000
spain	0.018	0.000
france	0.016	0.000
mexico	0.014	0.000
egypt	0.013	0.000

```
INTERPRETATION:
This table highlights whether recommendations
favor titles from specific countries, which may require mitigation
to ensure balanced catalog exposure.
```

Figure 17: Country Exposure Bias Analysis



Bias Audit Summary: Risk, Detection, and Mitigation

	Bias Risk	Detection Metric	Mitigation Applied
0	Genre dominance	Genre distribution vs catalog baseline	Genre-aware similarity weighting
1	Country dominance	Country distribution vs catalog baseline	Balanced feature weighting
2	Overly similar recommendations	Intra-list Diversity (ILD)	Intra-list Diversity (ILD)
3	Narrow thematic exposure	Catalog Coverage (CC)	Weighted hybrid similarity + coverage-aware selection

```
INTERPRETATION:
This summary demonstrates that bias identification, measurement, and mitigation
are explicitly linked. Each identified exposure risk is paired with an appropriate
proxy metric and a concrete mitigation strategy already implemented in the system,
ensuring responsible and auditable recommendation behavior.
```

Figure 18: Bias Audit Summary and Mitigation Mapping

This methodology ensures that the recommender system is designed, evaluated, and validated in a manner consistent with its cold-start objectives, business constraints, and ethical considerations. By combining structured preprocessing, multi-model comparison, proxy evaluation metrics, and transparent explainability techniques, the approach delivers a reproducible and governance-ready recommendation pipeline suitable for real-world deployment under data-limited conditions.

Code Availability

The full, reproducible implementation of this project—including data preprocessing, feature engineering, model training, evaluation, and visualization—is available in a public GitHub repository:

https://github.com/codewithmford/Marvin_Ford_Pillar_5_Capstone_Project

The repository contains the complete notebook, supporting data, and documentation required to reproduce all results presented in this report.

Results, Analysis and Discussion

This section presents and interprets the empirical findings of the content-based recommender system, contextualizing technical performance within Netflix's business objectives of discovery, engagement, and catalog utilization. Results are discussed in relation to system robustness, interpretability, and strategic value in addressing cold-start challenges common in large-scale streaming platforms.

Representation Quality and Similarity Structure

The recommender system is built on a semantic representation derived from metadata features such as genre, description, cast, director, and country. As shown in Figure 9, the TF-IDF representation spans 8,809 titles and 5,000 weighted features, enabling full pairwise similarity computation across the catalog. This confirms that Netflix's content metadata contains sufficient semantic richness to support high-quality recommendations even in the absence of user interaction data.

From a strategic perspective, this capability allows Netflix to operationalize recommendation logic immediately for newly released or under-exposed titles, reducing time-to-discovery and minimizing reliance on historical user behavior. This directly supports content launch efficiency and mitigates cold-start risk in fast-moving content pipelines.

Comparative Performance of Recommendation Models

Model performance varies meaningfully across the evaluated approaches, revealing important trade-offs between relevance, diversity, and catalog exposure. As shown in Figure 10, the rules-based baseline achieves perfect relevance with a Precision@K of 1.00 but exhibits extremely low Intra-List Diversity (ILD = 0.025) and limited Catalog Coverage (0.047). While this model reliably retrieves closely related titles, its narrow output leads to repetitive recommendations and reinforces popularity bias, limiting discovery and long-term user engagement.

The TF-IDF model improves content diversity by leveraging richer textual representations. As shown in Figure 10, TF-IDF achieves an ILD of 0.551, substantially higher than the rules-based baseline, indicating greater variety in recommendations. However, this improvement comes at the cost of relevance stability, with Precision@K decreasing to 0.824.

This trade-off highlights a common challenge in recommender systems: increasing exploration often weakens immediate relevance, which can negatively impact perceived recommendation quality.

The weighted TF-IDF model introduces controlled weighting between genre and descriptive text to explicitly manage this trade-off. As illustrated in Figure 11, tuning the genre weight between 0.4 and 0.8 reveals that Precision@K peaks around 0.4, while both diversity and catalog coverage decline as genre dominance increases. At higher weights, recommendations become increasingly narrow, confirming that excessive structural weighting reduces exploratory value. The tuned model achieves Precision@K of 1.00, ILD of 0.193, and Catalog Coverage of 0.053, reflecting improved balance relative to the baseline but still limited diversity.

The semantic embedding model demonstrates the strongest overall performance across evaluation criteria. As shown in Figures 11 and 12, it achieves a Precision@K of 0.956, an ILD of 0.336, and Catalog Coverage of 0.052, producing the highest composite selection score (0.589). This indicates that the embedding approach maintains high relevance while meaningfully expanding thematic variety. Unlike keyword-based methods, embeddings capture contextual similarity, enabling recommendations that are both coherent and non-redundant. This balance directly supports Netflix's strategic objective of promoting discovery while maintaining user satisfaction.

From a modeling perspective, these results also illustrate the balance between overfitting and underfitting in recommendation behavior. The rules-based approach exhibits characteristics of overfitting, where extremely high Precision@K is achieved at the expense of diversity, resulting in repetitive recommendations concentrated around a narrow subset of titles.

Conversely, the TF-IDF model demonstrates elements of underfitting, as increased diversity and exploration are accompanied by reduced relevance, indicating weaker semantic alignment. The embedding-based model mitigates both extremes by maintaining high relevance while preserving diversity, reflecting a balanced representation that generalizes effectively across content types without collapsing into overly narrow or overly diffuse recommendation patterns.

Across models, catalog coverage remains intentionally constrained by the fixed Top-K recommendation design, with values clustered around 0.05, ensuring controlled exposure rather than random exploration. The stability of these metrics across multiple anchor titles confirms that the system behavior is consistent and not driven by isolated examples. Overall, the results demonstrate that semantic embeddings offer the most effective trade-off between relevance, diversity, and robustness, making them well-suited for deployment in data-limited, cold-start environments where explainability and governance are critical.

Model Selection and Trade-Off Analysis

Composite evaluation results presented in Figure 12 confirm that the embedding-based model delivers the most favorable balance across relevance, diversity, and catalog exposure. Quantitatively, the semantic embedding approach achieves a Precision@K of 0.956, an Intra-List Diversity (ILD) of 0.336, and a Catalog Coverage of 0.052, resulting in the highest overall selection score of 0.589 among all evaluated models. These values indicate that the model consistently retrieves relevant content while maintaining a materially higher level of diversity than rule-based or purely TF-IDF approaches.

In contrast, the rules-based baseline, while achieving perfect relevance (Precision@K = 1.00), exhibits extremely limited diversity (ILD = 0.025) and lower catalog exposure (0.047), reinforcing narrow and repetitive recommendations. The TF-IDF model improves diversity substantially (ILD = 0.551) but at the expense of relevance (Precision@K = 0.824), while the weighted TF-IDF model demonstrates that increasing genre emphasis narrows diversity without meaningful gains in relevance. These quantitative trade-offs highlight that high relevance alone is insufficient when it leads to excessive redundancy or reduced content discovery.

From a product perspective, the embedding-based model's balanced performance directly supports strategic business objectives. By achieving strong relevance while expanding exposure across the catalog, the model reduces over-dependence on blockbuster titles and encourages discovery of long-tail content. This improves content utilization efficiency, mitigates popularity bias, and enhances the return on content investment. The observed performance metrics demonstrate that the selected approach delivers both technical robustness and measurable business value, making it well suited for scalable, real-world deployment in content platforms operating under data-limited conditions.

Robustness Across Content Types

Robustness analysis across multiple anchor titles demonstrates that the system adapts appropriately to varying content characteristics. As shown in Figure 13, niche titles such as Show Dogs produce more focused recommendation sets, with a lower Intra-List Diversity (ILD) of 0.20, while broader titles such as King of Boys: The Return of the King exhibit higher diversity with an ILD of 0.78. Mid-range content, including documentaries like Dick Johnson Is Dead and Alt-Right: Age of Rage, fall between these extremes with ILD values of 0.71 and 0.38, respectively. This pattern confirms that the model dynamically adjusts recommendation breadth based on thematic richness rather than applying a uniform similarity threshold.

This behavior aligns with user expectations and supports personalized exploration without requiring explicit personalization data. Narrow, genre-specific titles naturally produce tighter recommendation clusters, while broader narratives surface a wider range of related content. The ability to adapt recommendation diversity in this way reflects meaningful semantic understanding rather than rigid rule-based matching.

Variance analysis in Figure 14 further confirms consistent system behavior across content types. Intra-List Diversity exhibits a low variance of 0.056 with a standard deviation of 0.237, indicating stable diversity behavior across different anchors. Catalog Coverage shows near-zero variance (0.000), confirming that coverage is structurally constrained by the fixed Top-K evaluation setup rather than instability in model performance. This stability demonstrates that the system behaves predictably across content categories, reducing the risk of erratic or biased recommendations.

From an operational perspective, this consistency is critical for large-scale deployment. Stable diversity and coverage patterns reduce the likelihood of unpredictable user experiences, support reliable content exposure planning, and increase trust in automated recommendation pipelines. Collectively, these results indicate that the model delivers controlled, interpretable, and dependable behavior across diverse content types, aligning with enterprise requirements for scalable and responsible recommendation systems.

Explainability and Bias Considerations

Explainability is assessed by analyzing term overlaps, as illustrated in Figure 15, to uncover the semantic factors that influence recommendation choices. For example, in the case of Dick Johnson Is Dead, overlapping terms such as “father,” “life,” and “documentaries” dominate the similarity signal, while other recommendations share terms such as “death,” “filmmaker,” and “documentaries.” This demonstrates that similarity decisions are grounded in meaningful semantic overlap rather than opaque latent representations. Such transparency supports internal validation, model governance, and regulatory readiness by enabling stakeholders to clearly trace why specific titles are recommended.

Bias analysis in Figures 16 and 17 examines genre and country representation by comparing catalog-level distributions with recommendation outputs. At the genre level, documentary-related categories exhibit the strongest amplification. For example, documentaries account for approximately 4.3% of the catalog but represent 60.0% of recommendations, while documentaries – international movies increase from 2.1% of the catalog to 40.0% of recommendations. This pattern reflects intentional semantic alignment rather than uncontrolled popularity bias, as recommendations concentrate on content that is most contextually relevant to the anchor title. In contrast, genres such as comedies, stand-up comedy, and children and family movies appear in the catalog at rates between 2.2%–3.8% but receive 0% representation in the recommendation output, confirming that relevance, rather than overall frequency, governs selection.

Country-level exposure patterns further reinforce this behavior. As shown in Figure 17, titles from the United States comprise approximately 35.3% of the catalog but account for 20.0% of recommendations, while the United Kingdom represents only 5.3% of the catalog yet constitutes 80.0% of the recommendation list for this anchor. Other regions, including India (12.2%), Japan (3.1%), and South Korea (2.5%), do not appear in the recommendation output. These shifts indicate that geographic exposure is driven by semantic relevance rather than proportional representation, consistent with a content-aware rather than popularity-driven system.

Taken together, these results demonstrate that the recommender intentionally departs from raw catalog distributions in favor of semantically coherent recommendations, while remaining auditable through explicit genre and country diagnostics. This design supports bias mitigation by preventing uncontrolled dominance of high-volume categories while still allowing meaningful thematic concentration when warranted by content similarity.

Figure 18 consolidates identified bias risks and mitigation strategies, demonstrating that fairness and accountability are embedded directly into system design. Genre dominance is mitigated through Intra-List Diversity constraints, while country imbalance is monitored through coverage-aware evaluation. The integration of these safeguards ensures that recommendations remain interpretable, auditable, and aligned with responsible AI principles. For Netflix, this approach supports equitable global representation, mitigates overexposure risks, and reinforces trust in automated recommendation systems deployed at scale.

Summary of Findings

Overall, the results demonstrate that a content-based recommender system can effectively support Netflix’s discovery objectives under cold-start conditions. The embedding-based model provides the strongest balance between relevance, diversity, and stability while remaining interpretable and operationally feasible.

From a business standpoint, this approach enables faster content discovery, improved utilization of long-tail assets, and reduced dependency on historical user data. These capabilities position the system as a scalable foundation for personalization strategies while maintaining governance, explainability, and operational control.

Managerial Implications

The findings of this study offer several practical implications for organizations deploying recommender systems in content-rich environments. First, the results demonstrate that effective recommendation quality can be achieved without relying on extensive user behavior data. This enables faster onboarding of new users and content, reduces dependency on long-term behavioral tracking, and supports privacy-conscious design strategies.

Second, the ability of the embedding-based model to balance relevance and diversity has direct implications for content strategy. By increasing exposure to underutilized titles, platforms can improve catalog efficiency, extend content lifespan, and reduce reliance on a small subset of high-performing titles. This supports both user engagement and content return on investment.

Third, the system's transparency and bias-aware design provide operational advantages. Explainable recommendations enhance trust among stakeholders and support governance requirements, while diversity monitoring helps mitigate overexposure risks. Together, these features position the system as a scalable and responsible foundation for personalization initiatives.

Limitations and Future Work

While the proposed system demonstrates strong performance, several limitations should be acknowledged. First, the model relies exclusively on metadata and does not incorporate behavioral signals such as watch duration, clicks, or user preferences. As a result, recommendations may not fully capture individual taste variations. Second, the quality of recommendations is inherently constrained by the completeness and accuracy of the available metadata.

Future work could address these limitations by integrating collaborative filtering or hybrid recommendation techniques that combine content-based and behavioral signals. Additional enhancements may include dynamic re-weighting of features based on user feedback, incorporation of temporal viewing patterns, and evaluation using online A/B testing frameworks. These extensions would further strengthen personalization while preserving the transparency and robustness demonstrated in the current system.

Conclusion

This project demonstrates that a content-based recommender system can effectively support content discovery in data-limited environments by leveraging structured metadata and semantic similarity, delivering relevant, diverse, and interpretable recommendations without reliance on user interaction data. Evaluation results show that the semantic embedding–based model provides the strongest overall performance, achieving a balanced trade-off between relevance, diversity, and catalog coverage, and confirming that meaningful recommendation quality can be achieved without overfitting or dependence on behavioral signals.

From a business perspective, these findings highlight the value of content-based recommendation as a scalable and resilient solution for platforms such as Netflix. The approach supports efficient content discovery, improved utilization of long-tail assets, and reduced dependency on historical user data, all of which contribute to stronger engagement and more flexible personalization strategies.

The integration of explainability and bias-aware evaluation further strengthens the system's practical viability. By ensuring transparency, consistency, and governance readiness, the solution aligns with responsible AI principles and supports sustainable deployment in real-world environments.

Overall, this project demonstrates a well-founded, analytically sound, and business-relevant recommender system that provides a strong foundation for future enhancements, including hybrid personalization and adaptive learning, while maintaining clarity, accountability, and operational effectiveness.

References

An, Y., Tan, Y., Sun, X., & Ferrari, G. (2024, October 15). Recommender System: A Comprehensive Overview of Technical Challenges and Social Implications . *ICCK Transactions on Sensing, Communication, and Control, 1*(1), 30-51. Retrieved from https://www.icck.org/article/abs/TSCC.2024.898503

Henley, A., Bruckner, L., Jacobs, H., Jansen, M., Nunez, B., Rodriguez, R., & Wilson, M. (2024). On the Books: Jim Crow and Algorithms of Resistance, a Collections as Data Case Study. *ACM Journal on Computing and Cultural Heritage, 16*(4), 1-20. Retrieved from https://dl.acm.org/doi/10.1145/3631128

Ibrahim, O. A., Younis, E. M., Mohamed, E. A., & Ismail, W. N. (2025, January 04). Revisiting recommender systems: an investigative survey. *Neural Computing and Applications, 37*, 2145-2173. doi:https://doi.org/10.1007/s00521-024-10828-5

Ibrahim, O. A., Younis, E. M., Mohamed, E. A., & Ismail, W. N. (2025, January). Revisiting recommender systems: an investigative survey. *Neural Computing and Applications, 37*, 2145-2173. Retrieved from https://link.springer.com/article/10.1007/s00521-024-10828-5).

Kumar, M. (2025, November 30). *Recommendation Systems — Problems, Approaches, Landscapes, Methods (PALM)*. Retrieved from Medium: https://medium.com/@buildsomethinggreat/recommendation-systems-problems-approaches-landscapes-methods-palm-b58bf157d64b

Li, X., Cong, G., Xiao, G., Xu, Y., Jiang, W., & Li, K. (2024). On Evaluation Metrics for Diversity-enhanced Recommendations., (pp. 1286-1295). Retrieved from https://dl.acm.org/doi/10.1145/3627673.3679629

Zhao, Y., Yu, W., Liu, Y., & Cheng, X. (2024, March 1). *Fairness and Diversity in Recommender Systems: A Survey*. Retrieved from Arvix | Cornell University: https://arxiv.org/pdf/2307.04644