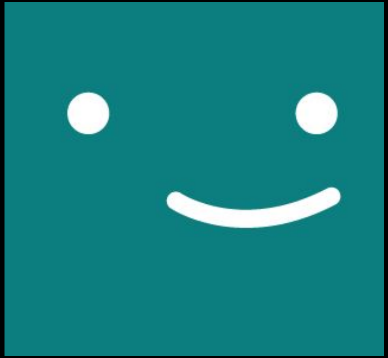
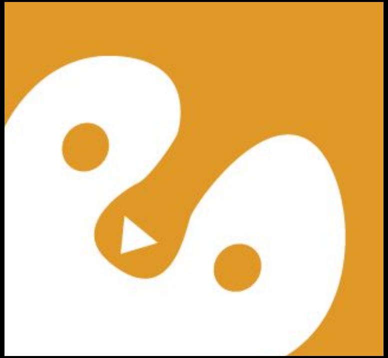
NETFLIX

Lola Lewis, Collin Li, Nico Lopez, Kaya Schubert

AGENDA



Background



Central Question

Hypotheses

Appendix

Background

Company Overview

- Subscription-based streaming service founded in 1997, offering a vast library of movies, TV shows, and original content.
- Began as a DVD rental service, evolved into a global streaming platform available in over 190 countries.

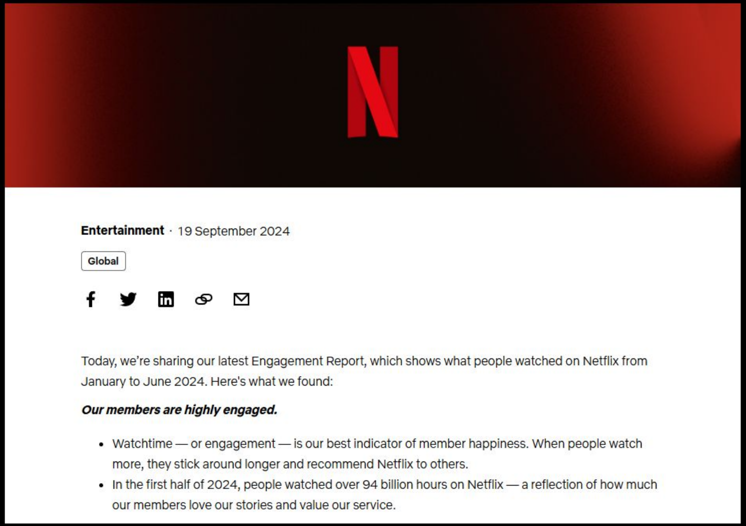
Reed Hastings (left) and Marc Randolph (right)

Home TV Shows Movies Recently Added My List

Data Overview

- 1036 films
- Biannual Netflix “What We Watched” file consisting of data from January to June of 2024
  - Includes **total watch hours** and **number of views** per film.
- Performed a merge with 2 files:
  - Rotten Tomatoes “TomatoMeter” listed as “audienceScore”
  - IMDB file that includes IMDB Score, Cast, and Director information.

What We Watched the First Half of 2024



show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	imdb_score	imdb_votes	hours_viewed	views	audienceScore
s7	Movie	My Little Pony:	Robert Cullen,	Vanessa Hudgens, Kimiko Gl		9/24/2021	2021	PG	91 min	Children & Far	Equestria's div	6.8	3468	9500000	6300000	84
s10	Movie	The Starling	Theodore Melf	Melissa McCa	United States	9/24/2021	2021	PG-13	104 min	Comedies, Dra	A woman adju	6.3	11733	2700000	1600000	71
s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Marshall		9/22/2021	2021	TV-14	94 min	Thrillers	After a deadly	5.3	15464	6900000	4400000	64
s39	Movie	Birth of the Dra	George Nolfi	Billy Magnuss	China, Canada	9/16/2021	2017	PG-13	96 min	Action & Adver	A young Bruce	5.6	8112	500000	300000	67
s42	Movie	Jaws	Steven Spielbe	Roy Scheider,	United States	9/16/2021	1975	PG	124 min	Action & Adver	When an insatiable great white shark terror			6100000	3000000	90
s46	Movie	My Heroes We	Tyler Greco			9/16/2021	2021	PG	23 min	Documentarie	Robin Wiltshire's painful childhood was res			200000	500000	83

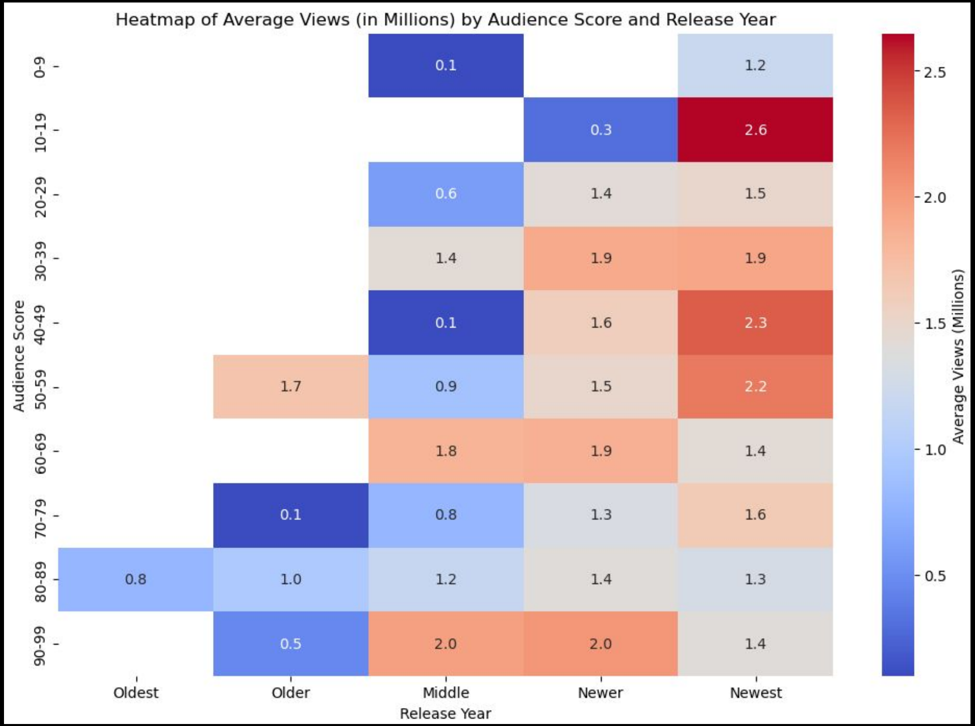
Central Question

What content characteristics influence viewer engagement, performance, and audience satisfaction on Netflix?

Hypotheses

Problem 1 - Relationship between Audience Score and Viewership

Hypothesis: **High audience scores do not always correspond to high viewership**, suggesting that external factors, such as marketing, release timing, or genre appeal, may outweigh viewer satisfaction in driving Netflix popularity.



Oldest: 1970 - 1987
Older: 1988 - 2005
Middle: 2006 - 2019
Newer: 2020 - 2023
Newest: 2024

Correlation Analysis Between audienceScore and views:

	audienceScore	views
audienceScore	1.000000	-0.064324
views	-0.064324	1.000000

Finding:

- Titles with **moderate audience scores** achieve **higher views**
- **Newer titles consistently** perform better
- Inverse correlation between audience score and views

Managerial Insights:

- Promotion Strategy
- Recency Effect

Problem 2 - Predicting Movie Performance

Hypothesis: The performance of a movie, as measured by IMDb scores and audience scores, is most significantly influenced by the number of stars in the cast, viewership hours, and duration. For Logistic Regression, we predict the number of stars in the cast are positively correlated with the performance, while older release years, viewership hours, higher duration, and more directors have a negative correlation with performance.

Applications:

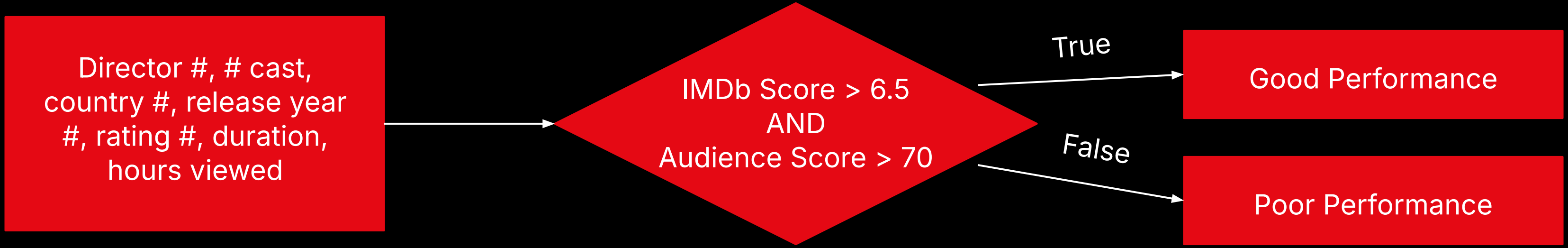
Content Curation and Investment Decisions

- Understanding which features drive movie performance enables Netflix to invest in content that aligns with audience preferences, maximizing returns on investments

Audience Engagement and Recommendations

- By identifying factors influencing a movie's success, Netflix can enhance their recommendation algorithms to better match content with viewer interests

Problem 2 - Predicting Movie Performance



Logistic Regression

	precision	recall	f1-score	support
0	0.64	0.61	0.63	140
1	0.63	0.66	0.64	140
Accuracy			0.64	280

- Model is correct in approximately 64% predictions across all classes
- Precision, recall, and f1-score all similar due to SMOTE
- Balanced performance with moderate accuracy

Random Forest Classifier

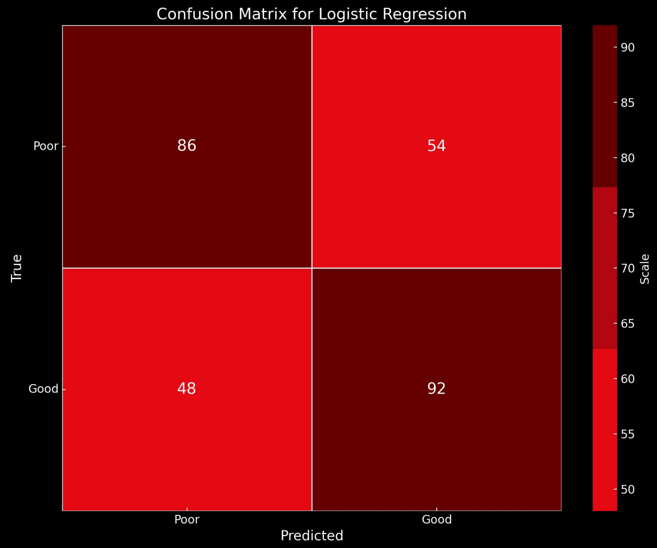
	precision	recall	f1-score	support
0	0.78	0.79	0.78	140
1	0.78	0.78	0.78	140
Accuracy			0.78	280

- Model is correct in approximately 78% predictions across all classes
- Precision, recall, and f1-score all similar due to SMOTE
- Better than Logistic Regression at capturing non-linear relationship and interactions in data

Home TV Shows Movies Recently Added My List

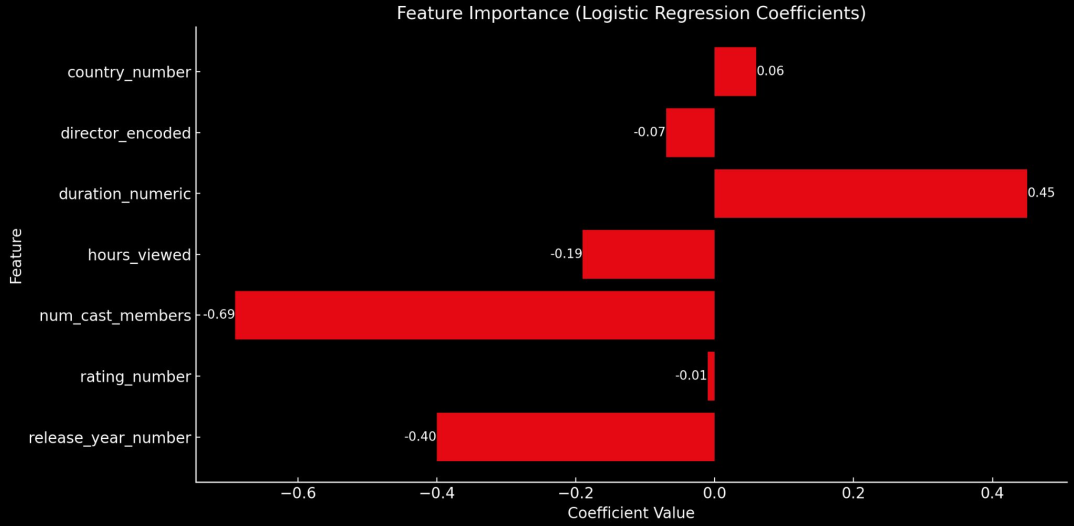
Problem 2 - Predicting Movie Performance

Logistic Regression



- Predicted 86 ‘poor performance’ movies correctly, 48 incorrectly
- Predicted 92 ‘good performance’ movies correctly, 54 incorrectly

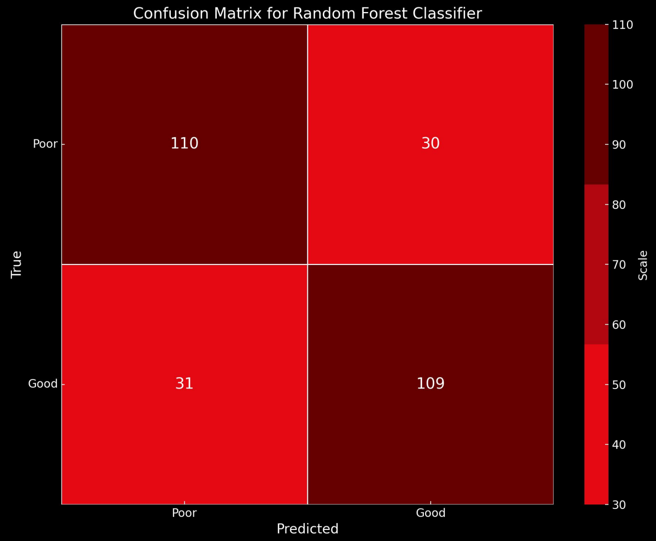
- num_cast_members : -0.69
- release_year_number: -0.40
- hours_viewed : -0.19
- duration_numeric: 0.45



Home TV Shows Movies Recently Added My List

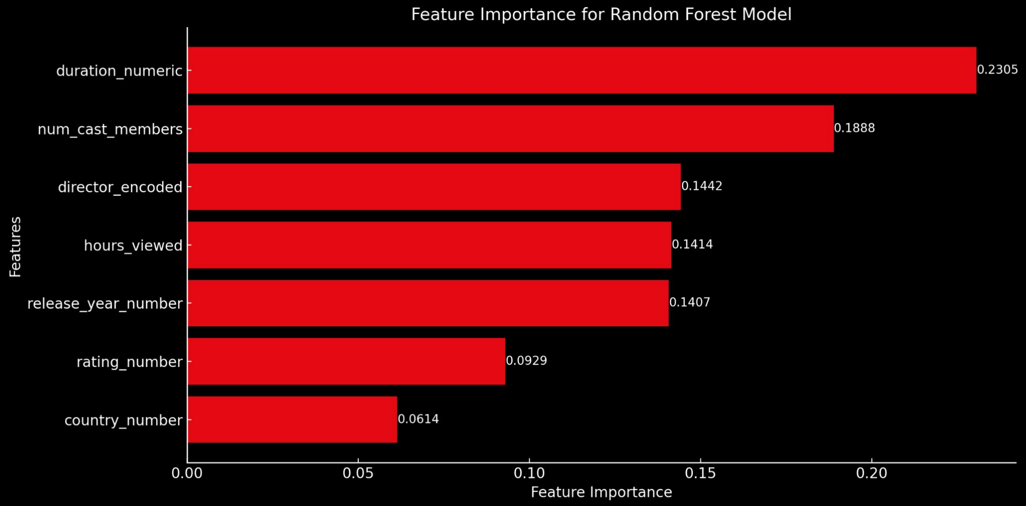
Problem 2 - Predicting Movie Performance

Random Forest Classifier



- Predicted 110 'poor performance' movies correctly, 31 incorrectly
- Predicted 109 'good performance' movies correctly, 30 incorrectly

- duration_numeric : 0.2305
- num_cast_members: 0.1888
- director_encoded: 0.1442
- hours_viewed: 0.1414



Problem 2 - Predicting Movie Performance

Observations:

- The Random Forest model outperformed Logistic Regression with an accuracy of 78% compared to 64%, highlighting its ability to capture non-linear relationships between features and movie performance

- The models validated our thesis that duration, number of cast members, and viewership hours are significant drivers of movie performance, with duration emerging as one of the most impactful feature in both Logistic Regression and Random Forest

Managerial Insights:

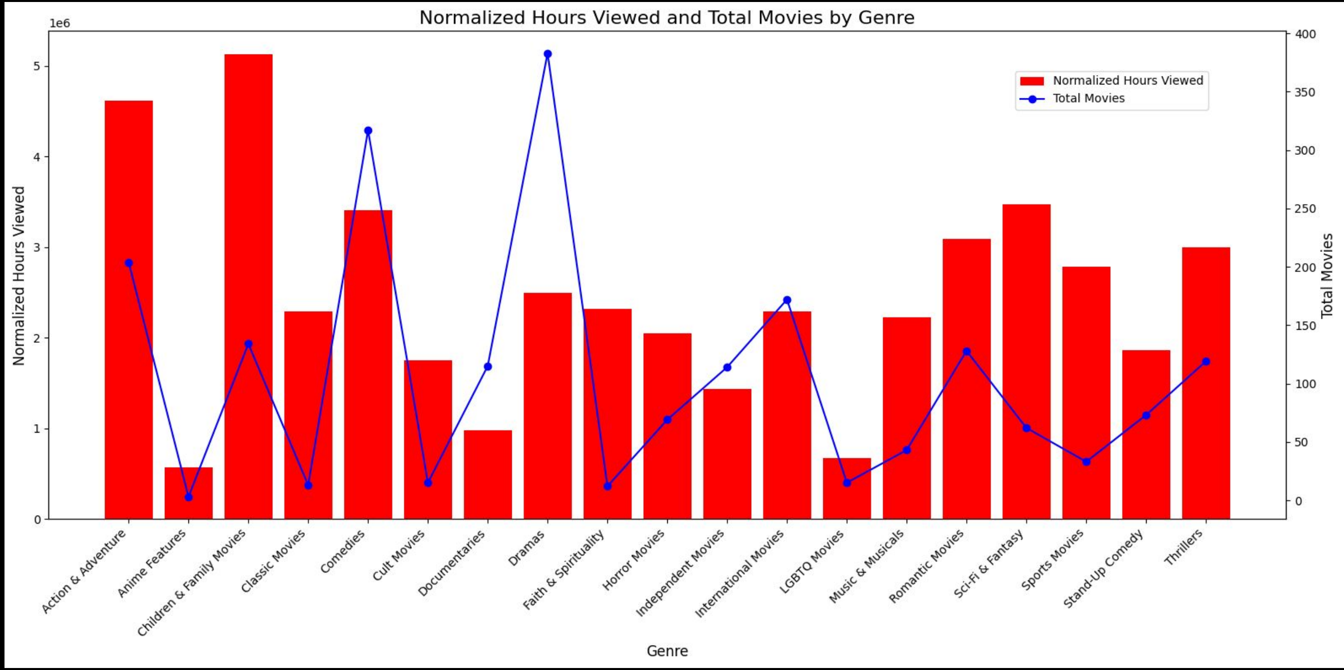
- Quantity does not equal quality. The negative relationship between cast size and performance suggests that overcrowding a film with stars can take away focus. Additionally, higher viewership does not equate to a better performing movie. A balanced cast and effective use of duration can maximize performance

- Focus marketing, promotion, and recommendation efforts on newer titles, as they are more likely to appeal to audiences and achieve higher performance metrics. Netflix can strategically position older movies in categories like "classics" or as part of limited-time promotions, rather than treating them as flagship content

Home TV Shows Movies Recently Added My List

Problem 3 - Movie Hosting

Genres with high normalized hours viewed but fewer total movies indicate strong audience engagement and potential for strategic content expansion.



THANKS

Appendix

Data Dictionary

- **title** - Name of the Netflix title
- **audienceScore** - Audience ratings on a scale of 0 - 100 (Rotten Tomatoes)
- **views** - Total number of viewers for the title
- **hours_viewed** - Total hours watched for the title
- **release_year** - Year the title was released
- **rating** - Age-appropriate rating (e.g., PG-13, TV-MA)
- **listed_in** - Genres or categories associated with the title (e.g., Action, Drama)
- **imdb_score** - IMDb rating for the title
- **imdb_votes** - Number of votes the title received on IMDb
- **date_added** - Date the title was added to Netflix's catalog
- **duration** - Length of the title (in minutes for movies)
- **director** - Name(s) of the director(s)
- **cast** - Names of the cast members
- **country** - Country or countries where the title was produced
- **director_encoded:** The numerical value assigned to the director(s) of a movie. Movies with multiple directors are assigned a fixed value, while unique directors are mapped to unique integers
- **num_cast_members:** The numerical count of the number of stars in a movie. Calculated by splitting the cast column (a comma-separated list of actors) and counting the entries
- **country_number:** The numerical value assigned to a specific country associated with a movie. Movies associated with multiple countries are assigned a fixed value.
- **release_year_number:** Converts the release year of the movie into a sequential numeric format (e.g., the earliest year is 0, the next is 1, etc.).
- **rating_number:** Converts movie ratings (e.g., G, PG, PG-13) into numeric values based on a predefined mapping.
- **duration_numeric:** Represents the runtime of the movie in minutes, extracted as a numeric value from the duration column (e.g., "120 min" becomes 120).
- **hours_viewed:** Indicates the total hours a movie has been viewed, providing a direct measure of its popularity. Missing values are filled with the mean.