Examining the Netflix Top 10 Feature (Team 166)

Roshaun Brady *Gabriel Carpio* *Siva Nagarajan* *Jessica Ruiz*

Introduction and Motivation

Netflix's Top 10 list has become a go-to resource for viewers to discover the most popular shows and movies on the platform. However, this list alone lacks deeper insights into the factors that contribute to a program's success. Our project aims to uncover the key drivers behind a movie's Top 10 ranking, which can empower viewers to make more informed viewing choices and help content creators and business executives make strategic decisions.

The rise of streaming platforms like Netflix has revolutionized the way we consume entertainment. These platforms offer an abundance of content, personalized recommendations, and the convenience of on-demand viewing. While this has given viewers a sense of control, the underlying algorithms that power these platforms can sometimes create an illusion of choice (Jenner 2020). Even in the early days of Netflix, algorithms were used to encourage binge-watching behavior (Matrix 2014).

As streaming platforms continue to evolve, there is an increasing concern for incorporating a "human touch" in their products and services (Van Es 2022). The addition of the Top 10 feature on Netflix is a step in this direction, as it provides viewers with insights into what other users are watching. However, the reasons behind a title's inclusion in the Top 10 list remain largely opaque.

Problem Definition

The primary problem we aim to address is the lack of deeper insights into the factors that contribute to a movie's Top 10 ranking on Netflix. Specifically, we want to answer the following question:

> Given a set of movie attributes (e.g., genre, release year, runtime, director, cast, etc.) and the corresponding Top 10 rankings on Netflix, can we develop a predictive model to accurately forecast a movie's Top 10 ranking?

Understanding recommendation systems and the drivers of content popularity is critical for both end-users and businesses (Brookey et al. 45; Ferchaud and Proffitt 28; Roy and Dutta 2; Wang et al. 8). By developing a Tableau dashboard that visualizes these insights, we aim to empower viewers to make informed consumption choices and enable content creators and business executives to make strategic production and distribution decisions (Islam 2019).

Proposed Method

This project's main goal is to develop a predictive framework for assessing the likelihood of media programs, including movies and TV shows, reaching the Top 10 category on Netflix. Now that the Internet is being used to deliver us on-demand content, Netflix has an incentive to keep viewers on the platform, and displaying the Top 10, as a way to increase the sense of community, is one way to do that (Chalaby 2023). Utilizing our cleaned and reformatted dataset, the integration of the SMOTE technique with a logistic regression model enables the creation of dynamic visualizations within Tableau. These visualizations enable users to explore historical trends from 2020 to 2022, revealing highly ranked programs across various attributes such as genre and country.

Data

The team acquired a publicly available dataset on the Netflix top 10 feature to use for their analysis (link). To further enrich the original dataset, an additional Netflix dataset was obtained to provide additional attributes that could be leveraged for predictive capabilities (link). The datasets were mostly clean, but the team did encounter some issues with dirty data, particularly in fields like title, country, genre, and duration. For example, the genre field simply listed all genres of a program separated by commas, all in the same column. This was problematic because most programs are in multiple genres, and the number of genres varies per program. The duration field also had several types of values depending on whether it was a TV show or a movie. In the cases of shows, it was the number of seasons, while for movies, it was the number of minutes.

To address these data quality issues, the team employed various null handling techniques to clean the data and widen the dataset to create features for genres. This involved parsing the genre field to extract individual genres and creating separate columns for each genre, as well as normalizing the duration field to have consistent units (e.g., minutes for both movies and TV shows). One key challenge the team faced was finding Netflix data that could effectively enhance the top 10 list dataset. The team tried to use data from IMDB but found it to be too dirty to incorporate due to the overarching inconsistency of data accuracy. This data source was created by users and therefore includes errors that need to be corrected. Given the scale and nature of this project, the team determined that addressing these data quality issues was beyond the scope of the current effort. Despite the data quality challenges, the team successfully enhanced the original Netflix top 10 dataset by incorporating additional attributes from a separate Netflix dataset. This expanded dataset provided more comprehensive information that could be leveraged for predictive analytics and other data-driven insights. The team's diligence in cleaning and transforming the data was crucial in ensuring the reliability and usefulness of the final data set for our analysis.

Data Analysis

The team achieved building a classification machine learning model to predict whether a show would land in Netflix's top 10. We undertook several preprocessing steps, including using regex to parse the duration column into separate features for the number of seasons and the duration for movies, as it encompassed metrics for both TV shows and movies such as ‘2 seasons or 60 min’. Additionally, we addressed the mixed-genre listings in the 'listed_in' column by deduplicating and creating binary columns for each genre as a row could contain a mix of genres such as ‘Horror, Thriller, Drama’. To tackle the imbalance between titles in the top 10 and the entire Netflix catalog, we employed the SMOTE technique. By using one-hot encoding on categorical data, our model could effectively make predictions and our initial logistic regression model. After cross-validation, we achieved an accuracy of 0.85 and showcases manipulating the data into numerical and tangible data allows algorithms to understand is essential to allow big data to turn into actionable tools (Mackenzie 2015).This approach is novel as it offers a predictive framework for determining top 10 titles, enabling data-driven decisions for producers and filmmakers to comprehend what users prefer; it is critical for creators to be more targeted and understand their audience better as social media, short-form content, and alternative content are all competing in the attention economy for engagement. (Liang 2022). After constructing a logistic regression model, the team delved into the exploration of other advanced algorithms, including XG Boost, Random Forest, and SVM, to ensure robustness and effectiveness. To assess these models, the team used Area Under the Curve (AUC) metric to determine performance as accuracy can be misleading with imbalanced datasets and summarizes the model's overall efficacy across different thresholds, making it possible to compare models and assess their practical utility in real-world scenarios

Comparatively, XG Boost and Random Forest exhibited a notable AUC of 0.80, outperforming logistic regression, which achieved a score of 0.71. Despite this, the team opted to retain the logistic regression model within the Tableau platform for its simplicity. Nonetheless, they acknowledged the potential for enhancement and viewed the superior performance of XG Boost and Random Forest as an avenue for future iteration.

Data Visualization

The Tableau visualizations were mostly developed using the pre-processed data, except our prediction visualization which utilizes the output from our logistic regression model. These visualizations allowed us to dynamically play with the data to discover patterns and trends that support the case for a program making it to the Top 10. They also allow the end user to visually interact with Top 10 trends in an easily digestible way.

One of our visualizations is a stacked bar chart that can be dynamically changed by the user to show how the distribution of programs in the Top 10 changes over time across various attributes. The stacked bar always adds up to 100%, and the color split changes to different attributes, such as TV show vs. movie, or release year, based on the user's choice. The x-axis defaults to one bar per year but can be adjusted for greater granularity using plus and minus indicators. A filter enables selection of specific ranks or rank ranges, allowing users to focus on, for example, only the Top 5 programs.

Here are some key takeaways from the stacked bar chart, which truly lets us begin to see how Netflix's Top 10 feature may be influencing viewers:

- TV shows make up more than half the Top 10, and this percentage is only increasing over time.
- Netflix exclusives also make up more than half the Top 10, and this percentage rises all the way to 75% for 2022.
- More recent programs dominate the Top 10 – as soon as a new year begins, a plurality of the Top 10 are programs from that year.

Because these are all reasonable expectations of the viewer experience, this does not immediately give any reason to assume the Top 10 data are not reliable, as the Critical Studies in Television article questions (Scarlata 2022).

Another tab displays bar charts for genres and countries where productions were filmed. Initially excluded from the stacked bar chart due to multiple genres or filming locations per program, each country or genre now has its own bar. This decision makes it easier to fit all elements on the chart. Programs spanning multiple countries or genres are counted multiple times, one in each category. While this may result in bars exceeding the number of programs, it effectively showcases dominant countries and genres in the Top 10. Users can also exclude the United States for clearer visualization, as its inclusion often dominates the view.

The next three tabs provide different ways of viewing the top programs in the Top 10 database. The third tab shows a temporal view of the number one ranked program for that month. The user can filter based on month, year in Top 10, and release date. Because the overall Top 10 dataset is daily, the large size of it can make it cumbersome to work with and understand. This monthly view lets users of the dashboard see which all programs made it to #1 at any point in the month, which is easier to see in this list format for each month, rather than having to scroll through day-by-day, and provides enough information regardless, because it's unlikely that it would matter to a future creator specifically which day of the month a program reached #1.

The fourth tab is an animated view from the inception of the Top 10 feature until the end of the 2-year span for which we have the data, showing the total number of days a program has been in the Top 10 over that span, and the animation can be played, paused, and rewound to show change over time. There are options to change the color-coding of the dots and to filter by genre by manually typing in a genre. The list of valid genres can be found on the Prediction tab. The main interesting takeaway from the final result of this one is how much Cocomelon dominates, not only in terms of total days, but how its curve is always rising – which means that it remains in the Top 10 at all times of year.

The fifth tab lets the user view the top programs within the 2-year time span based on the number of days spent with a Top 10 ranking. Users can filter on rank, year, and type of program. Many reoccurring top 10 shows are Netflix exclusives with similar attributes, showing an indication of recycled content being released by Netflix (Cuelenare 2024).

Lastly, there is a tab labeled "Make Your Own Program," allowing users to select different attributes for their ideal program. Leveraging our logistic regression model, we estimate the probability of this program reaching the Top 10. This is the main product that is the output of our project and research, which we would want our end users to utilize when attempting to develop their own program for Netflix. All of the next section about Experiments and Evaluations is based on this tab.

Our Tableau Dashboard can be viewed here:
https://public.tableau.com/app/profile/siva.nagarajan/viz/Team166ProjectDVASpring2024/ofTop10byAttribute

Experiments and Evaluations

The primary inquiries guiding our experiments and evaluations were twofold:

1. What attributes of Netflix programs increase their likelihood of reaching the top 10?
2. Does the Tableau tool we developed accurately predict a program's success and effectively assist creators?

The prediction tab was created by taking the logistic regression equation that was the output of the aforementioned analytics, and allowing the user to "plug and play" choosing different values for attributes such as the duration of the program, the type of program, and its rating and genres, to see the likelihood of this program reaching the Top 10 based on our model.

Due to the dataset's coverage of all U.S. Netflix programs over a 2-year span, where only a fraction reaches the top 10, it's logical that most of the coefficients in the model are negative. This suggests a low success rate for random programs. However, focusing on the few positive coefficients offers insight into the primary drivers of success, defined as reaching the Top 10.

Most of the attributes with positive coefficients are the binary attributes we made for each genre, where the value is 1 if the program is in that genre, and 0 otherwise. Having a value of "1" for any of these genres had a positive coefficient: Crime TV, Reality TV, Action TV, Drama TV, Mystery TV, and Sci-Fi and Fantasy TV. In addition, for the binary attribute indicating whether a program is a TV show or Movie, the TV Show value had a positive coefficient. When a program was a movie, the duration attribute (length in minutes) of the program also had a positive coefficient. On the other hand, when a program is a TV show, the duration attribute (number of seasons) has a negative coefficient, indicating that having fewer seasons is better. While this might at first seem counterintuitive, it makes sense due to the rise of miniseries, and that many TV shows that end up in the top 10 are miniseries, which get coded as a 1-season TV show in the data.

To test out the basic functionality of the prediction tool in Tableau, one simple experiment that was done was to start with a baseline of a 1-season TV show that is rated TV-MA (while all the ratings have negative coefficients, TV-MA is the most frequent rating to appear in the Top 10 as discovered using the stacked bar chart) and then manipulate the different genre attributes mentioned above to make sure the percentage chance increases for those 6 genres. Indeed it does, and in fact, if all 6 genres are chosen, it increases the chance of the program reaching the Top 10 to 93%! However, this might indicate a small flaw in the tool, that being that the tool allows for the creation of unrealistic programs – it's unlikely, perhaps impossible for a program to fit all 6 of those genres. Even if it were possible, we are trying to quantify something that is partially based on emotion (viewer sentiment) and shoving all 6 genres into the same program may reduce the actual positive impacts that those genres have on the viewer experience. Nevertheless, for more realistic programs with 2-4 genres that go along with each other, the tool accurately reflects the likelihood.

To test this, another experiment was to take the list of programs that made the "Top 10 of the Top 10" on the similarly named Tableau tab, and test out each of Top 10 programs, with their respective attributes, on our tool to see what the readout would be.

Top 10 of the Top 10	**Rating**	**# of Seasons**	**Genres**	**Tool Output**
Cocomelon	TV-Y	3	Kids' TV	3%
Ozark	TV-MA	3	Crime TV Shows, TV Dramas, TV Thrillers	80%
Cobra Kai	TV-14	3	TV Action & Adventure, TV Comedies, TV Dramas	72%
Manifest	TV-14	3	TV Dramas, TV Mysteries, TV Sci-Fi & Fantasy	69%
Outer Banks	TV-MA	2	TV Action & Adventure, TV Dramas, Teen TV Shows	88%
Squid Game	TV-MA	1	International TV Shows, TV Dramas, TV Thrillers	19%
All-American	TV-14	3	TV Dramas, Teen TV Shows	62%
Bridgerton	TV-MA	1	Romantic TV Shows, TV Dramas	66%
Lucifer	TV-14	6	Crime TV Shows, TV Comedies, TV Dramas	63%
Virgin River	TV-14	3	Romantic TV Shows, TV Dramas	42%

As evident from the above analysis, most of the "Top 10 of the Top 10" are accurately predicted by our tool, with a probability exceeding 60% of reaching the Top 10. Even the 42% likelihood for Virgin River should be regarded positively, suggesting a significant chance of success solely based on the identified attributes. However, two outliers, Cocomelon and Squid Game, were predicted with low success rates of 3% and 19%, respectively. Despite this, they represent exceptional cases within their genres—international TV and kids' TV—highlighting the limited predictive power of our tool in capturing such unicorns. These instances underscore the distinction between correlation and causation, emphasizing that our tool should not be perceived as a guarantee of success or failure. Human emotions, such as viewer sentiment, cannot be fully captured through quantitative analysis alone.

An intriguing observation from the "Top 10 of the Top 10" is that when filtering to show only ranks 1-3, Cocomelon disappears entirely. This suggests that while Cocomelon consistently makes it to the Top 10, it rarely achieves the highest ranks. This phenomenon may indicate that our model is more adept at predicting programs that reach the pinnacle of success rather merely making it into the overall Top 10.

Conclusions

In conclusion, our model has pinpointed key attributes for Netflix Top 10 success, including being a TV show, having fewer seasons (or longer runtime for movies), and specific genres like crime, reality, drama, action, mystery, sci-fi, and fantasy. Our Tableau tool enables creators to experiment with attribute combinations to predict success, achieving an 85% accuracy rate in cross-validation. However, it's crucial to recognize that predicted success rates are based on correlation, not causation. Future research and improvement areas include assessing the tool's validity beyond the U.S. Top 10, addressing unrealistic scenarios allowed by the tool (such as mixing TV and movie genres), and refining predictions for programs with no genre or excessive genres. While our tool aids in estimating success, it should supplement rather than dictate creative decisions, serving as one of many factors in the creative process.

Statement of Effort

All team members contributed a similar amount of effort.

Software

The packaged repository contains two datasets in csv files, two Jupyter notebooks containing data preprocessing and model building. All code is executable in these notebooks. The data visualization is published on the Tableau Public site for viewers.

APPENDIX

Literature Survey

Brookey, Robert Alan, et al. Triaging the Streaming Wars. 1st ed., Routledge, 2023.

Cuelenaere, Eduard. "How 'original' are Netflix original films? mapping and understanding the recycling of content in the age of Streaming Cinema." Media, Culture &amp; Society, 31 Jan. 2024, https://doi.org/10.1177/01634437231224081.

Chalaby, J. K. (2023). The streaming industry and the platform economy: An analysis. Media, Culture & Society, 0(0). https://doi.org/10.1177/01634437231210439

Ferchaud, Arienne, and Jennifer M. Proffitt. Television's Streaming Wars. Routledge, 2023.

Islam, Mohaiminul, and Shangzhu Jin. "An overview of data visualization." *2019 International Conference on Information Science and Communications Technologies (ICISCT)*, 4 Nov. 2019, https://doi.org/10.1109/icisct47635.2019.9012031

Jenner, Mareike. "Researching binge-watching." Critical Studies in Television: The International Journal of Television Studies, vol. 15, no. 3, Sept. 2020, pp. 267-279 https://doi.org/10.1177/1749602020935012.

Liang, M. (2022). The end of social media? How data attraction model in the algorithmic media reshapes the attention economy. Media, Culture & Society, 44(6), 1110-1131. https://doi.org/10.1177/01634437221077168

Mackenzie, A. (2015). The production of prediction: What does machine learning want? European Jouranal of Cultural Studies, 18(4-5), 429-445. https://doi.org/10.1177/1367549415577384

Matrix, Sidney Eve. "The Netflix Effect: Teens, Binge Watching, and On-Demand Digital Media Trends." Jeunesse Young Peoples Texts Cultures, vol. 6, no. 1, 2014. ResearchGate, https://www.researchgate.net/publication/270665559_The_Netflix_Effect_Teens_Binge_Watching_and_On-Demand_Digital_Media_Trends

Roy, D., Dutta, M. A systematic review and research perspective on recommender systems. J Big Data 9, 59 (2022). https://doi.org/10.1186/s40537-022-00592-5

Scarlata, Alexa. "'What are people watching in your area?': Interrogating the role and reliability of the Netflix top 10 feature." Critical Studies in Television: The International Journal of Television Studies, vol. 18, issue 1, 2022. SageJournals, https://journals.sagepub.com/doi/10.1177/17496020221127183.

Van Es, Karin. “Netflix & Big Data: The Strategic Ambivalence of an Entertainment Company” Television & New Media, vol. 24, issue 6, 2022. SageJournals. https://journals.sagepub.com/doi/10.1177/15274764221125745

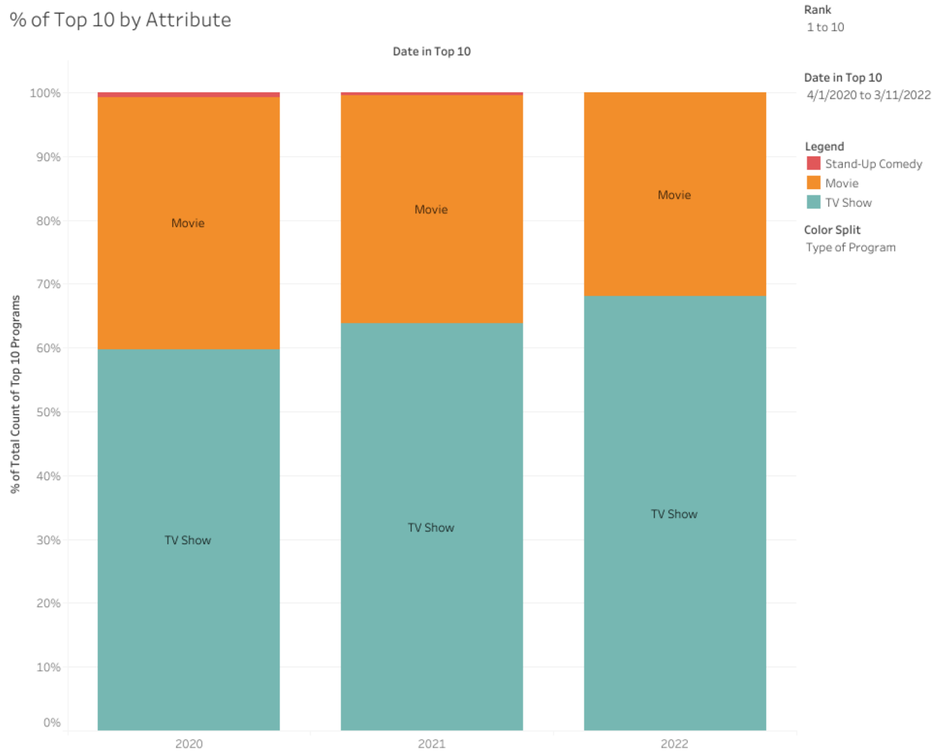
Wang, Zan, et al. “An Improved Collaborative Movie Recommendation System Using Computational Intelligence.” Journal of Visual Languages and Computing, vol. 25, no. 6, Dec. 2014, pp. 667–75.

Final Data Sources

Netflix daily top 10
Netflix Daily Top 10 Movie/TV Show in the United States from 2020 - Mar 2022.
https://www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us

Netflix Movies and TV Shows
Listings of movies and tv shows on Netflix - Regularly Updated
https://www.kaggle.com/datasets/shivamb/netflix-shows

Tableau Images



Duration (minutes if movie, number of seasons if TV show)
6

Type of Program
TV Show

Rating
TV-MA

Probability of Your Program Ending Up in the Top 10: 54.99%

Independent Movies Genre	International TV Shows Genre	Romantic Movies Genre	Romantic TV Shows Genre	Sci-Fi & Fantasy Genre	Science & Nature TV Genre	Spanish-Language TV Shows Genre
No	No	No	No	No	No	No

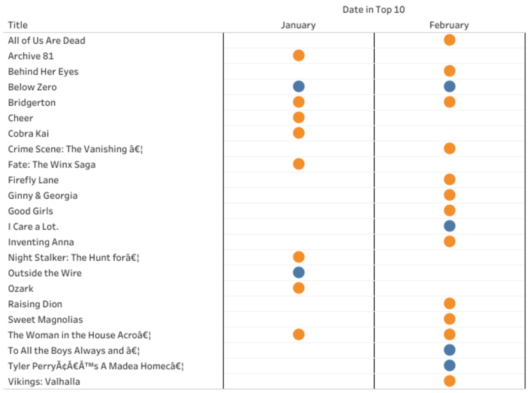
Reaity TV Genre	Sports Movies Genre	Stand-Up Comedy & Talk Shows Genre	Stand-Up Comedy Genre	Teen TV Shows Genre	Thrillers Genre	TV Action & Adventure Genre	TV Comedies Genre
No	No	No	No	No	No	No	No

Music & Musicals Genre	Cult Movies Genre	Documentaries Genre	Docuseries Genre	Dramas Genre	Faith & Spirituality Genre	Horror Movies Genre	TV Dramas Genre	TV Horror Genre
No	No	No	No	No	No	No	No	No

LGBTQ Movies Genre	TV Mysteries Genre	TV Sci-Fi & Fantasy Genre	TV Thrillers Genre	Action & Adventure Genre	Anime Features Genre	Anime Series Genre	British TV Shows Genre	Classic & Cult TV Genre
No	No	No	No	No	No	No	No	No

International Movies Genre	Korean TV Shows Genre	Kids' TV Genre	Classic Movies Genre	Children & Family Movies Genre	Comedies Genre	Crime TV Shows Genre
No	No	No	No	No	No	No

Highest Ranked Program by Month



	Date in Top 10	
Title	January	February
All of Us Are Dead		🟠
Archive 81	🟠	
Behind Her Eyes		🟠
Below Zero	🔵	🔵
Bridgerton	🟠	🟠
Cheer	🟠	
Cobra Kai	🟠	
Crime Scene: The Vanishing â€¦		🟠
Fate: The Winx Saga	🟠	
Firefly Lane		🟠
Ginny & Georgia		🟠
Good Girls		🟠
I Care a Lot.		🔵
Inventing Anna		🟠
Night Stalker: The Hunt forâ€¦	🟠	
Outside the Wire	🔵	
Ozark	🟠	
Raising Dion		🟠
Sweet Magnolias		🟠
The Woman in the House Acroâ€¦	🟠	🟠
To All the Boys Always and â€¦		🔵
Tyler PerryÃ¢â‚¬â„¢s A Madea Homeâ€¦		🔵
Vikings: Valhalla		🟠

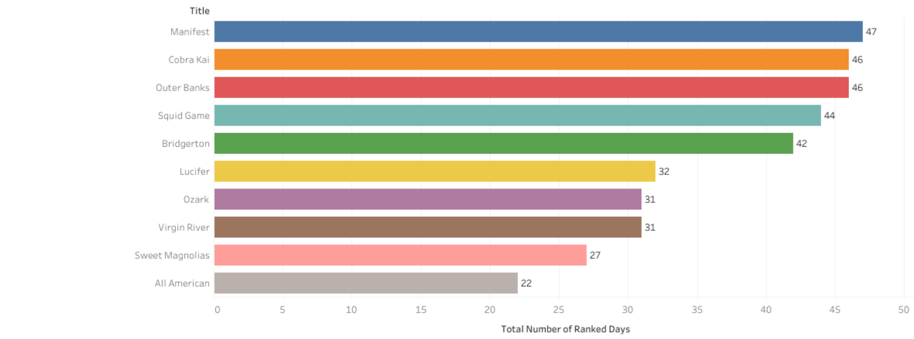
Month in Top 10
☑ January
☑ February
☐ March
☐ April
☐ May
☐ June
☐ July
☐ August
☐ September
☐ October
☐ November
☐ December

Year in Top 10
☑ 2020
☑ 2021
☑ 2022

Netflix Release Year
☑ 2013
☑ 2014
☑ 2015
☑ 2016
☑ 2017
☑ 2018
☑ 2019
☑ 2020
☑ 2021
☑ 2022

Type
🟦 Movie
🟧 TV Show

Top 10 of the Top 10



Type
☑ Movie
☑ Stand-Up Comedy
☑ TV Show

Rank
☑ 1
☑ 2
☑ 3

Year of Date Added
☑ 2018
☑ 2019
☑ 2020
☑ 2021

Title
🟦 Manifest
🟧 Cobra Kai
🟥 Outer Banks
🟩 Squid Game
🟩 Bridgerton
🟨 Lucifer
🟪 Ozark
🟫 Virgin River
🟥 Sweet Magnolias
⬜ All American