Ben Dieck & Tim Tieng

Team SOF Science

Syracuse University

IST 707 – Data Analytics

Dr. Jeremy Bolten

Final Presentation

# Aggregate Compilation of Netflix Data

In the modern digital era, the entertainment landscape has undergone a profound transformation, with streaming services emerging as the predominant medium for content consumption and at-home entertainment. These platforms, boasting millions of global subscribers, offer an extensive repository of multimedia content, accessible at users' convenience. However, the triumph of a streaming service is contingent not merely on content volume but, critically, on its adeptness in personalizing content recommendations, elevating user experiences, and retaining a loyal subscriber base. This juxtaposition of data understanding and machine learning, central to the streaming industry, forms the nucleus of our investigation.

At the heart of effective machine learning lies an intricate comprehension of data. In the context of streaming services, this entails a holistic understanding of user behaviors, content preferences, and consumption patterns. Through meticulous scrutiny of the vast expanse of user-generated data, streaming platforms can unearth invaluable insights, empowering them to refine content curation, amplify user engagement, and optimize core business functions. Therefore, the fusion of data understanding and machine learning stands as an unequivocal prerequisite for attaining a competitive advantage in the fiercely competitive streaming domain.

Machine learning algorithms, driven by profound data understanding, equip streaming services with the capacity to furnish personalized content recommendations that captivate users. These algorithms harness historical user interactions, demographic information, and even real-time data to serve tailored content suggestions that resonate with individual tastes. This not only augments user satisfaction but also bolsters user retention—a pivotal metric underpinning the profitability and durability of streaming platforms. Additionally, machine learning techniques have the potential to fine-tune content delivery, ensuring judicious bandwidth allocation and mitigating user frustration arising from buffering or sluggish loading times.

As we plunge deeper into the intricacies of the interconnected world of streaming services, it becomes obvious that an adept understanding of data coupled with the application of machine learning is quintessential for amplifying business performance. In this report, we address the approach we took, models we used, and our insights we derived to help highlight how streaming services, can leverage their data to optimize their business.

**About the Data – Netflix Userbase Dataset & Netflix UK Audience Behavior**

The Netflix Userbase Dataset offers a comprehensive glimpse into a sample of Netflix users, encompassing a wide array of data facets pertaining to user subscriptions, financial metrics, account particulars, and user activity. Each row within the dataset corresponds to a distinct user and is identified by a unique User ID. Vital information contained within the dataset comprises the user's subscription tier (Basic, Standard, or Premium), the monthly revenue attributed to their subscription, their Netflix registration date (Join Date), the date of their most recent payment (Last Payment Date), and their geographical location. The initial columns of the two dataset provides interesting insight to user composition and behaviors, while the aggregation of the two datasets develops predictive insights into how users will interactive with content, and what attributes the content has that might impact that interaction.

In addition to these fundamental attributes, the dataset incorporates supplementary columns designed to provide insights into user behaviors and preferences. These supplementary columns encompass Device Type, encompassing device categories such as Smart TV, Mobile, Desktop, and Tablet, as well as Account Status, which signifies whether an account is currently active or inactive. It is essential to note that the dataset is entirely synthetic, thus bearing no resemblance to real Netflix user data. However, it serves as a valuable resource for analytical pursuits and model development, facilitating a deep exploration of hypothetical user trends, preferences, and revenue generation patterns within an artificial Netflix userbase.

In addition to the core attributes that were included with the dataset, the aggregate dataset developed to create modeling incorporates several derived columns to assist in analysis. These were generated utilizing measures within the dataset. The most important of these is the generated “behavior” metric which is a discrete variable that categorizes user behavior 0-3. “0” classifies programming that was skipped over without watching. “1” indicates programming that was viewed for a small period of time, and then the user moved away from the content. “2” classifies content that was watched for the full duration, and then turned off. “3” is the target variable, content that was watched for a duration greater than 100% of its runtime. This is binge watched content that kept users on the platform beyond the scope of other iterations.

The Netflix Audience Behavior - UK Movies dataset also offers a unique window into other user preferences on the Netflix platform. It focuses on users in the United Kingdom who willingly consented to have their anonymous browsing activities tracked. This dataset exclusively encompasses desktop and laptop interactions, accounting for approximately 25% of global Netflix traffic, and spans a fixed timeframe from January 2017 to June 2019. It meticulously records each instance when a member of the tracked user panel in the UK accessed a Netflix.com/watch URL associated with a movie, providing valuable insights into audience engagement.

The 'Duration' attribute within the dataset quantifies the time (in seconds) until a user interacted with another URL after visiting a movie page. A watch time of zero seconds signifies that the user briefly visited the page without further engagement. This data holds profound significance as it addresses a critical need in the media industry - the ability to bridge the gap between content creators and audience preferences. In an increasingly privatized media landscape, where streaming dominates content distribution, this dataset becomes a vital resource for filmmakers and businesses. It sheds light on audience behavior and consumption patterns, facilitating the creation of commercially viable projects and informed business strategies in an era where data is the bedrock of decision-making, especially in the realm of streaming media. Despite its imperfections, this dataset remains a crucial global measure of VOD (Video on Demand) activity, serving as a beacon of insights in the face of a data scarcity that surrounds this evolving industry. With increased competition in market share comes the requirement of tools that create predictive insight to achieve market advantage.

*Figure 1 – First 10 Rows of the netflixUsersData Dataframe*

*A screenshot of a computer

Description automatically generated*

*Figure 2 – Data Structure: NetflixUsersData*

*A black and white text

Description automatically generated*

**Exploratory Data Analysis / Macro – Level Understanding**

Exploratory data analysis (EDA) plays a pivotal role in unraveling the underlying insights within the Netflix user dataset, enabling the formulation of actionable business recommendations. The dataset contains three key categorical attributes: Subscription Types (Basic, Standard, Premium), Country (providing a global view of users), and Device (comprising Tablet, Laptop, Smart TV, and Device). Each of these attributes holds immense significance in the context of the analysis.

Firstly, Subscription Type offers a lens through which we can discern the revenue contribution of different regions. By understanding which subscription tiers are prevalent in specific geographic areas, Netflix can strategically tailor its content offerings and pricing structures to maximize revenue.

Second, the Country attribute is instrumental in comprehending the geographical distribution of the user base. This insight not only aids in infrastructure planning and capacity forecasting but also informs Netflix's global expansion strategies, ensuring that it can cater to its diverse user base without compromising service quality.

Finally, the Device attribute is invaluable for developers seeking to optimize the streaming experience. Analyzing how users access Netflix on various devices provides essential insights into device-specific streaming quality and allows developers to fine-tune content formats for enhanced user satisfaction, aligning with Netflix's commitment to delivering top-notch viewing experiences across devices.

EDA of these attributes empowers Netflix to make data-driven decisions, ranging from revenue optimization and global scalability to tailored content formatting, ultimately enhancing the user experience and solidifying Netflix's position as a leading streaming service provider. This was the approach our data analytics team took to understand the macro-level implications of our data.

**User Subscription Breakdown**

Our team has adopted a multifaceted approach to comprehending the breakdown of user subscriptions on Netflix, with a keen focus on categorizing users into three distinct subscription types: Basic, Standard, and Premium. Leveraging data analytics techniques, we are meticulously analyzing the dataset to determine the distribution of users across these subscription tiers. This involves creating descriptive statistics, such as frequency counts and percentages, to quantify the proportion of users subscribed to each category. Additionally, we are employing data visualization tools to craft intuitive graphs and charts that vividly illustrate the subscription breakdown to empower executive decision making.

*Figure 3 – User Subscription Breakdown – Pie Chart*

*A pie chart with different colors

Description automatically generated*

*Figure 4 – User Subscription Type Percentage Breakdown*

*A screenshot of a cell phone

Description automatically generated*

Understanding the breakdown of user subscriptions is of paramount importance to Netflix for several compelling reasons. Firstly, it provides critical insights into the revenue landscape of the streaming service. By discerning which subscription types are the most prevalent among their user base, Netflix can strategically adjust pricing strategies and promotional campaigns to maximize revenue generation. Secondly, subscription breakdown analysis aids in tailoring content and feature offerings that might be unique to different types of subscriptions. This personalization not only bolsters user satisfaction but also contributes to user retention, a cornerstone of Netflix's long-term profitability. Furthermore, this granular understanding empowers Netflix to make informed decisions regarding content investments and global expansion strategies, ensuring they can allocate resources efficiently and effectively to meet the diverse needs of their user base across different subscription tiers. In essence, a nuanced grasp of user subscription breakdown equips Netflix with a competitive edge and a data-driven foundation for strategic decision-making in an increasingly dynamic streaming landscape.

**Users By Country Analysis**

Understanding the user base per country is of paramount significance for a streaming service like Netflix due to its global reach and diverse audience. This insight provides Netflix with a nuanced view of its user demographics, preferences, and behaviors, enabling the customization of content libraries to cater to the specific tastes and cultural nuances of each region. Moreover, it plays a pivotal role in infrastructure planning, ensuring that Netflix can allocate resources efficiently to maintain seamless service quality as it expands into new markets and scales existing ones. Additionally, understanding the geographical distribution of users is instrumental in crafting targeted marketing campaigns, optimizing pricing strategies, and tailoring customer support services to enhance user satisfaction and foster long-term loyalty. Ultimately, a profound grasp of the user base per country empowers Netflix to not only adapt its content offerings but also make informed decisions that contribute to its sustained growth and competitive advantage in the global streaming landscape.

*Figure 5 – Number of Netflix Users by Country*

*A graph with colorful bars

Description automatically generated*

The observations regarding user counts by country in the dataset hold significant implications for understanding Netflix's global reach and potential limitations. It's evident that Spain, the United States, and Canada boast the highest user counts, underscoring Netflix's stronghold in these regions. However, the uniformity of user counts across the "Rest of World" category raises questions about the dataset's completeness, particularly given the absence of India, one of the world's most populous nations. This observation highlights a potential data gap, emphasizing that the dataset might not fully represent Netflix's user base worldwide.

Recognizing this limitation is crucial for any analysis or business decision-making, as it suggests that insights drawn from the dataset may not summarize the entire field of Netflix's global user demographics and behaviors.

*Figure 6 – Global Heatmap of Users*

A map of the world

Description automatically generated

This heatmap was another means for the team to understand the user breakdown of the dataset. This Heatmap shows clusters of users in each country to potentially highlight which areas of the world has the most active users.

This is imperative when it comes to ensuring your service is highly available and fault tolerant. This type of study can help Netflix understand where to have servers to ensure there is capacity to accommodate all its users.

**User Subscription Type by Country – Revenue Implications**

Next, it is important how different users are arrayed across the globe, and to see if there is a specific region where one specific subscription type is dominant. To accomplish this, we created a simple bar chart to highlight this information requirement.

*Figure 7 – Subscription Type breakdown per Country*

**A graph of different colored bars

Description automatically generated**

*Figure – 8 – Monthly Revenue Analysis – Country View*

*A graph of a bar graph

Description automatically generated with medium confidence*

During this phase of analysis, we were made two distinct observations:

1. The top three countries are the same, but in different order. We can see that U.S generates the most revenue where it was ranked as the 2nd highest country with the most users. This can be attributed to the subscription breakdown per country.
2. We can also see more variance between revenue figures compared to user base.

Our observations regarding monthly revenue analysis for Netflix by country reveal an intriguing interplay between user counts and revenue generation. While the top three countries remain consistent, albeit in varying order, the data underscores the critical role of subscription breakdown per country in shaping revenue patterns. Notably, the United States emerges as the top revenue generator, despite being the second-highest in terms of user count. This highlights the influence of subscription types and pricing strategies on revenue dynamics. Moreover, the greater variance observed in revenue figures, compared to user base statistics, emphasizes the intricate web of factors—ranging from subscription choices to regional economic disparities—that contribute to Netflix's financial landscape. These insights underscore the complexity of the streaming industry and underscore the importance of data-driven decision-making in optimizing revenue streams for a global service like Netflix.

**Device Breakdown – Classification Modelling**

The team attempted to run a classification model on our dataset to see if we could predict the type of device a user uses, Smartphone, Tablet, laptop, or Smart TV, based on the available attributes in the dataset.

Device prediction in the context of a streaming service like Netflix holds paramount importance due to its multifaceted implications for user experience enhancement and strategic decision-making. Firstly, predicting the type of device a user employs to access the platform offers a pivotal opportunity to optimize the streaming quality. Different devices possess distinct capabilities, screen sizes, and processing power. Accurate device predictions empower Netflix to tailor content formats and streaming resolutions, ensuring that users receive an optimal viewing experience irrespective of their chosen device. This personalized approach not only enhances user satisfaction but also mitigates issues like buffering, ensuring that users can seamlessly enjoy their favorite content on any device, from smartphones to smart TVs.

Secondly, device prediction serves as a vital component of Netflix's data-driven strategy. By understanding how users interact with their service across various devices, Netflix can formulate targeted development plans and resource allocation strategies. This insight aids in the prioritization of device-specific feature enhancements and app optimizations, aligning with evolving user preferences. Additionally, it assists in content recommendations, as it enables Netflix to fine-tune suggestions based on the user's device capabilities and viewing habits. Ultimately, device prediction is an instrumental tool in Netflix's quest to continually innovate and deliver top-tier streaming experiences, making it a essential to their mission to maintain their competitive edge in the ever-evolving streaming landscape.

*Figure 9 – Classification Model Output*

***A screenshot of a computer

Description automatically generated***

Accuracy: This is the overall accuracy of the model, which measures the proportion of correctly classified instances. In this case, the accuracy is 0.2485, which is quite low.

95% CI: The 95% confidence interval for accuracy.

No Information Rate (NIR): The accuracy that would be achieved by predicting the majority class for all instances. In this case, it's 0.2545.

P-Value [Acc > NIR]: The p-value associated with comparing the model's accuracy to the No Information Rate. It suggests whether the model performs significantly better than random guessing.

Overall, the confusion matrix and associated statistics indicate how well your model is performing for each class. In this case, it appears that the model is struggling to correctly classify instances into most classes, which may require further model tuning or data preprocessing.

**Derived Insight from Classification Modeling**

The information set was overly dimensional in its original form to allow for any valuable insight utilizing ARM, as the rules tended towards reflecting the attributes of the majority of transactions within the set, and not the overall trend of “What makes people watch content?”. This was resolved by reducing the dataset to a random sampling of 1000 observations of each type of user behavior, 0-3. The simplified dataset allowed for a better understanding of associative rules that encouraged increased content usage.

A diagram of a number of words

Description automatically generated with medium confidence

A diagram of a network

Description automatically generated with medium confidence The highest support for behavior 3 indicates that users do not binge watch modern programming, but more interesting older content produced between 2010-2020. In general they are, specifically, movies of average popularity and US production. The popularity of US programming within the global market is predictable and consistent with observed global trends in the entertainment industry, even as Netflix attempts to increase market capture by incorporated programming from other countries including China, Japan, Korea, and India. A caveat of this analysis is that the programming described by these associative rules also describes the majority of the content in the set. So there isn’t necessarily anything startling about the results of the model.

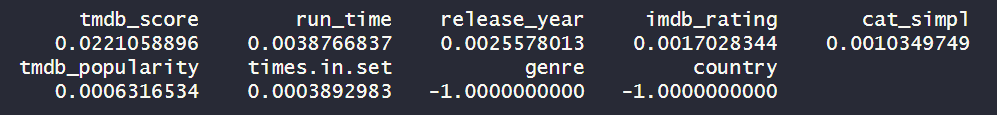
The inverse model, displaying “0” classified behavior, also includes those films that are considered Average scoring, released between 2010-2020 (Pre-Pandemic). So it is difficult to rule the behavior “3” model as providing any kind of predictive insight. We can see that the majority of the rules for the entire shortened dataset involve those attributes.

A graph of a bar graph

Description automatically generated with medium confidence

But the 0-model also includes high lift for longer programming, 2-3 hours in length, and indicates significant support from the lower end of the rotten tomatoes database score. This indicates the the tmdb\_score may provide a more significant degree of predictive insight into if a user will skip content.

Exploring this possibility, our team utilized CORElearn attribute evaluation to measure the information gain ratio of the original dataset of 175k observations. The tmdb score was by far the most influential attribute in decision making within the dataset.



The plethora of discreet data within the set made usage of other models unviable. Decision trees on their own could not handle the dimensionality of even the truncated data set. SVM, kNN, and Random Forest could not process the information, so the dataset was transformed, again, to attempt to resolve the problem. Converting the categorical variables into numeric variables involved utilizing a sequence of dummy variables using binary values. The total attributes, after discretization of attributes like imdb\_rating, tmdb\_rating, run\_time, etc… developed another highly dimensional dataset of 4000 observations x 399 attributes.

A black background with yellow squares

Description automatically generated

Attempts to use PCA to reduce dimensionality created variables whose original names were obfuscated, which didn’t produce any actionable insight.

This new data structure allowed for more accurate classification utilizing kNN. After several runs, kNN was optimized at k = 17 associated datapoints, using 5 fold cross verification. It was run using all available datapoints within the set which produced a predictive model that was capable of an average of 62.2 % accuracy in predictive ability. Ironically, however, the model does even better at extreme behaviors, skip and binge watching. Behavior 0, and 3, produced accuracy as high as 68.5% accuracy. This is significantly better than random guessing, but not necessarily enough accuracy to pass a composite risk management plan for business strategy.

A number on a black background

Description automatically generated

A screenshot of a computer

Description automatically generated

**Conclusions**

In conclusion, the transformative impact of streaming services on the contemporary entertainment landscape is indisputable, with millions of subscribers worldwide seeking convenient access to a vast array of multimedia content. However, the key to success in this highly competitive arena lies not only in content quantity but, crucially, in the art of personalization, user satisfaction, and subscriber retention. Our investigation has illuminated the central role of data understanding and machine learning in achieving excellence within the streaming industry.

Effective machine learning hinges on a deep understanding of data, encompassing user behaviors, content preferences, and consumption patterns. By meticulously dissecting user-generated data, streaming platforms gain invaluable insights that empower them to refine content curation, boost user engagement, and optimize core business operations. This fusion of data understanding and machine learning is undeniably pivotal for securing a competitive edge in the dynamic streaming domain.

Machine learning algorithms, driven by comprehensive data understanding, enable streaming services to offer personalized content recommendations that captivate users, bolstering both satisfaction and retention rates. Moreover, these algorithms optimize content delivery, mitigating issues such as buffering or slow loading times. As we delve deeper into the intricacies of streaming services, it becomes apparent that data comprehension coupled with machine learning is the cornerstone of accelerated business performance. Our report has explored our approach, models, and the insights derived to underscore how streaming services can harness their data to optimize their operations and further elevate the streaming experience for their subscribers.

# References

Soeiro, V. (2022, May 10). *Netflix TV Shows and Movies.* Retrieved from kaggle.com: https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

VOD Clickstream, Eliel CM, & Follows, S. (2019, June 28). *Netflix Audience Behavior - UK Movies.* Retrieved from kaggle.com: https://www.kaggle.com/datasets/vodclickstream/netflix-audience-behaviour-uk-movies