Netflix Analysis Using ElasticSearch, Kibana, AzureML

Hannah Hae-In Kim, David Montes, Vanessa Munoz, Shawn Tran
Department of Information Systems, California State University Los Angeles
CIS 3200-01 Data Processing & Analytics
hkim134@calstatela.edu dmonte29@calstatela.edu vmunoz28@calstatela.edu stran13@calstatela.edu

Abstract: Our group decided to conduct analysis from a streaming app called Netflix Movies and TV Shows. The overall main goal is to not only gain more knowledge about Netflix's database but to also understand when Netflix users are watching either TV Shows or Movies at a certain year. Every year, the Netflix company has been collecting data but they are unsure of how to utilize the information they have gathered. Our team had produced an analytical dashboard to show the variety of metrics and visualizations.

1. Introduction

Data processing and analysis is an incredibly important aspect of business in today's landscape. Being able to process and analyze data about the public can help businesses make informed decisions that can create an incredible competitive advantage. In this case, having knowledge of what kind of shows are very popular at the time can help a streaming service create the next blockbuster show.

Our group decided on this topic because streaming television shows and movies have become a large part of our entertainment industry as it provides a way for consumers to enjoy films and shows in the comfort of their own home. Online streaming services have taken the entertainment industry by storm as many of the young generation have turned to Netflix, Hulu, Youtube, and other streaming platforms as their main source of entertainment compared to television. This topic is significant because streaming platforms will be the future of entertainment in the coming years as we have seen during the pandemic. Many theatre release movies have instead turned to streaming service releases during the pandemic and have had great success, Disney+'s Mulan for example. Data analytics can be used by these streaming services to discover viewer trends and find what is popular, which can lead them to release shows that will become incredibly popular.

As technology advances, there are many different streaming services battling for our money. And instead of going to a concert or watching your favorite movie in a theater, people would rather stay at home and watch a movie or tv show through a streaming service. As streaming television shows and movies have grown globally these days, Netflix streaming application has become a tremendous hit to millions of home subscribers.

In this paper, section 2 will be based on related work. Section 3 will be based on existing work on Netflix services and why it is needed, especially playing a significant role on their success. Section 4 describes.. And finally, Section 5 is our Conclusion.

2. Related Work

Our project is specifically looking at over 7000 Netflix shows data, for example director, cast, genre, etc. We can observe this data and separate shows into categories based on the year they were released to find trends among the recent years of release rates of shows. We can also categorize this data based on the genre of shows or movies to find which genres are popular during specific time frames as well. The Netflix production company can also use this data to plan future releases that they believe will be extremely popular by using past data to predict future trends and patterns of popular genres.

Related research done about Netflix and its effects on digital media and our society states similar things about how streaming will take over the entertainment industry compared to normal television within the younger generation. "huge percentages of Netflix subscribers watched back-to-back episodes, devouring a season of content in just days".[1] They also state how a large portion of Netflix's viewer base belong to the millennial generation showing how television as a form of entertainment is slowly losing its grasp on the industry and streaming will be the new king. This strategy of releasing shows, all the episodes of a season at once, has created a binge-watching culture within our society where viewers will watch an entire season of a show in one sitting. This type of viewing has taken the world by storm as many have chosen to consume their

media in this way. This is not confined to Netflix either, other streaming services such as Hulu or Amazon Prime also have subscribers that binge watch series as well.

Another case study about Netflix is regarding their users, comparing their expectations using Netflix's streaming service and using on-demand movie and show services. It was discovered "that users have different levels of expectation based on the method used to deliver the video content".[2] Users of online streaming services such as Netflix or Hulu have lower levels of expectations compared to when they are viewing shows or movies through an on-demand service. This lowered expectation impacted how viewers perceive the video quality as well as the amount they are willing to spend on these entertainment services. This supports the previously stated comment about how online streaming services will soon take over the entertainment industry as users do not expect as much from online streaming services like Netflix or Hulu compared to traditional on-demand services.

3. Background/Existing Work

In the beginning, Netflix started in 1998 where customers would rent out and sell DVD copies through the mail. Renters would filter through watch options and return the DVD's through mail after they have watched them. Now as technology advances, the need to rent through the site and request a DVD to be sent via mail is no longer needed. Netflix has transformed into a subscription-based site where customers can stream their favorite movies in less than seconds. Apart from becoming a major streaming service, Netflix has successfully become their own production company. Because of the amount of movies and customers that Netflix has accumulated throughout the years, a crucial member of their team is now Data Scientists. They are needed to see and keep track of all of Netflix's datasets. When you're a data scientist, it is important to gather all different necessary types of data and analyze it. This type of data can be based on anything such as seeing the Netflix subscriber's gender, date of birth, and their taste in movies or tv shows. Knowing and analyzing this information updates each subscriber's algorithm to specifically fit what they want to watch. By incorporating concepts like data analysis, machine learning, and statistics, Data Science can help not just Netflix but any business to grow exponentially regardless of sector. [3]

With Netflix bringing a tremendous amount of different users across the world, each subscriber generates hundreds of ratings per day based on their search and adding tv shows and movies to their watch-list, this data ultimately becomes part of big data. [3] Netflix stores all of their Big Data by using key machine learning algorithms, it builds specific patterns to see the subscriber's taste in movies and tv shows.

As Data Scientists' gathers and analyzes Big Data, it has helped improve the ratings of all Netflix subscribers. Especially today, Data Scientists have such a unique way in attracting all of their new and old subscribers. Eventually, Netflix will know how long a subscriber is using their streaming service and see what time a subscriber is on Netflix. With this information, Netflix can see the behavior of a subscriber and the time of day when the user is going to come back, just by looking at their history. The algorithm also boosts different genres of shows that may be new to the subscriber. This is to slowly introduce watchers to try a different type of show. Netflix's goal is to get their audience to stream their shows and movies for as long as possible. And with the big data collected through Netflix's subscribers account, Netflix will regularly change their cover art and taste in genres which will attract their audiences.

With Netflix being available to stream right onto your laptop, Mac, mobile, tablet, and other devices, it is obvious to see why their streaming service has become a successful growth for their company. With the Big Data gathered through the Data Scientist collection, it has brought a big advantage to attract their Netflix subscribers. As a result, people that are not Data Scientist can also look through Netflix's dataset and create a glance of what the Netflix universe has created for all of their subscribers. According to Jack Kerschner's report, he has created an expansive Netflix catalogue using a dataset obtained on Kaggle that was manipulated to allow for new groupings. [4] At first, Jack used Microsoft Excel to separate all of the names such as cast, genres, countries, directors, and etc. After he created all of his data columns, he was ready to put them into Tableau. At the end of his result, he generated 55,953 names and 41 sheets to see Netflix's catalog. With Jack's Tableau visualization, you're able to see many different types of datasets that Netflix has collected throughout the years. Our work will be very similar to Jack's Tableau visual. And it will be based on the datasets used on Kaggle but it will have a simpler and cleaner form of different visuals such as bar charts, pie charts, and maps shown in Kibana.

4. Our work

Table 1. Quick View of the Netflix_Titles Dataset.

show_id	type	title	director	cast	country	date_adde	release_ye	rating	duration	listed_in	description
s1	TV Show	3%		JoÃ£o Mig	Brazil	########	2020	TV-MA	4 Seasons	Internatio	In a future where the e
s2	Movie	7:19	Jorge Mich	DemiÃ¡n B	Mexico	########	2016	TV-MA	93 min	Dramas, In	After a devastating ear
s3	Movie	23:59	Gilbert Chi	Tedd Chan	Singapore	########	2011	R	78 min	Horror Mo	When an army recruit i
s4	Movie	9	Shane Ack	Elijah Woc	United Sta	########	2009	PG-13	80 min	Action & A	In a postapocalyptic wo
s5	Movie	21	Robert Luk	Jim Sturge	United Sta	1-Jan-20	2008	PG-13	123 min	Dramas	A brilliant group of stud
s6	TV Show	46	Serdar Aka	Erdal BeÅ	Turkey	1-Jul-17	2016	TV-MA	1 Season	Internatio	A genetics professor ex
s7	Movie	122	Yasir Al Ya	Amina Kha	Egypt	1-Jun-20	2019	TV-MA	95 min	Horror Mo	After an awful accident
s8	Movie	187	Kevin Reyr	Samuel L. .	United Sta	1-Nov-19	1997	R	119 min	Dramas	After one of his high sc
s9	Movie	706	Shravan Ki	Divya Dutt	India	1-Apr-19	2019	TV-14	118 min	Horror Mo	When a doctor goes mi
s10	Movie	1920	Vikram Bh	Rajneesh D	India	########	2008	TV-MA	143 min	Horror Mo	An architect and his wif
s11	Movie	1922	Zak Hilditc	Thomas Ja	United Sta	########	2017	TV-MA	103 min	Dramas, TI	A farmer pens a confes
s12	TV Show	1983		Robert Wa	Poland, Ur	########	2018	TV-MA	1 Season	Crime TV S	In this dark alt-history t
s13	TV Show	1994	Diego Enrique	Osorno	Mexico	########	2019	TV-MA	1 Season	Crime TV S	Archival video and new
s14	Movie	2,215	Nottapon	Artiwara K	Thailand	1-Mar-19	2018	TV-MA	89 min	Document	This intimate documen
s15	Movie	3022	John Suits	Omar Epps	United Sta	########	2019	R	91 min	Independe	Stranded when the Eart
s16	Movie	1-Oct	Kunle Afol	Sadiq Dabo	Nigeria	1-Sep-19	2014	TV-14	149 min	Dramas, In	Against the backdrop o
s17	TV Show	9-Feb		Shahd El	Yaseen, Shai	########	2018	TV-14	1 Season	Internatio	As a psychology profes
s18	Movie	22-Jul	Paul Greer	Anders Da	Norway, Ic	########	2018	R	144 min	Dramas, TI	After devastating terro
s19	Movie	15-Aug	Swapnane	Rahul Peth	India	########	2019	TV-14	124 min	Comedies,	On India's Independenc
s20	Movie	'89		Lee Dixon,	United Kin	########	2017	TV-PG	87 min	Sports Mo	Mixing old footage with
s21	Movie	â€¢â€¢Kuch Bheege	Onir	Geetanjali	India	1-Sep-18	2018	TV-14	110 min	Dramas, In	After accidentally conn
s22	Movie	â€¢Goli Soda 2	Vijay Milto	Samuthira	India	########	2018	TV-14	128 min	Action & A	A taxi driver, a gangster

Netflix has become one of the biggest streaming services in the entertainment industry in the past decade. Data Analysis will help them overcome their rivals and take over the entertainment industry for years to come. The dataset size is 3MB and holds 12 columns with 7,787 records.

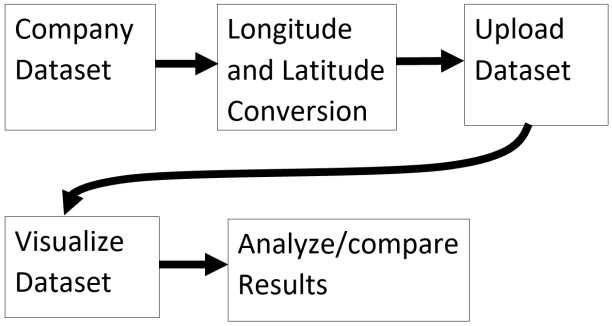

Figure 1. Data Implementation Process.

Our workflow diagram can be seen above in figure 1. We obtained a dataset on Netflix from kaggle and imported it into kibana after converting the countries to coordinates to be able to map the data. We then created index patterns and visualized the dataset through a pie chart, bar chart, and geo map.

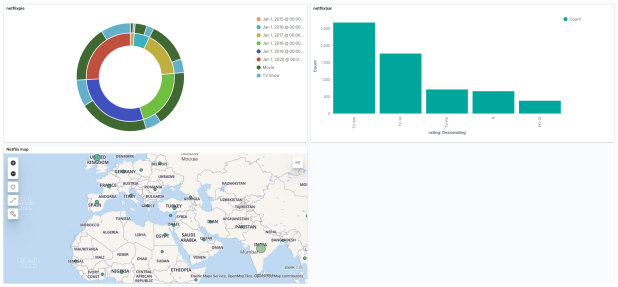

Figure 2. Data Visualizations

Our dashboard containing all three visualizations is displayed above in figure 2. It contains our pie chart, bar chart, and geo map.

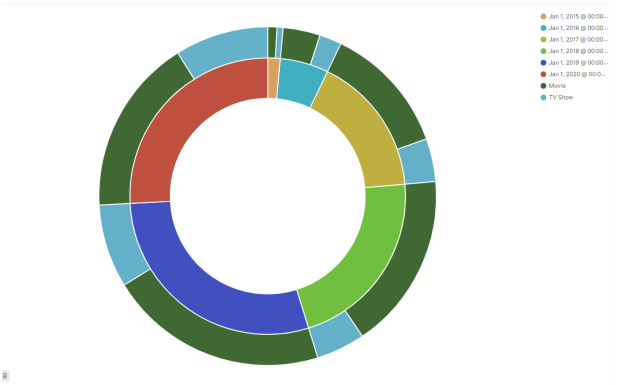

Figure 3. Types of films added since 2015

Next we will go more in depth with each visualization starting with the pie chart. The inner ring is divided into sections for each year from 2015-2020 starting with the small sliver going clockwise per year. The outer pie refers to the types of films released during that specific year. The blue refers to tv shows and the green refers to movies. Here we see the number of movies far outweigh the number of tv shows in all years, but recently tv shows have started to grow each year.

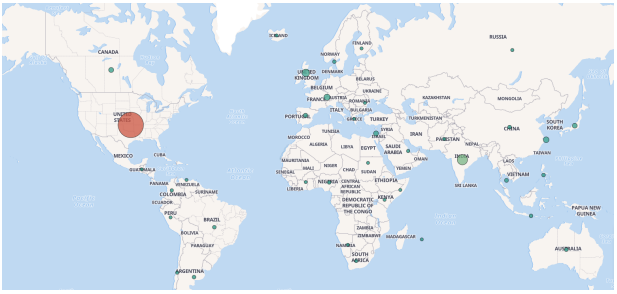

Figure 4. Geo-Spatial Map

The figure above displays our Geo-spatial map which displays counts of films from netflix produced in a specific country. As we can see from the map, the United States is by far the leader in producing Netflix films while India is not too far behind showing that Bollywood may rival Hollywood in the future.

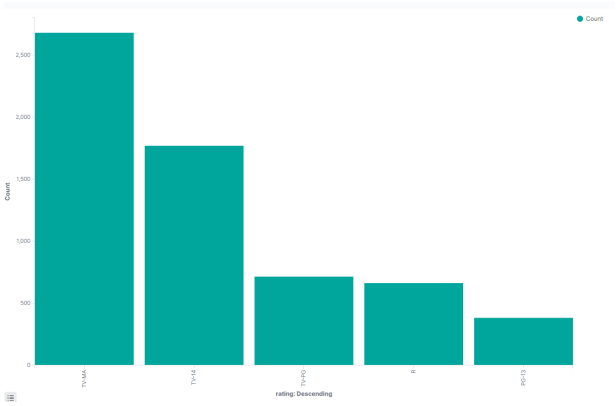

Figure 5. Ratings of films by count

This vertical bar chart depicts counts of ratings of Netflix movies and tv shows. From highest to lowest, the ratings rank TV-MA, TV-14, TV-PG, R, and PG-13. Through this visualization we can see that Netflix mainly carries mature content and usually does not carry children focused content such as their competitors such as Disney's streaming service Disney+.

Feature	Score
type	10.275009
country	5.646401
listed_in	4.355015
rating	1.012322
date_added	0.868689
title	0.019563
release_year	-0.01811
cast	-0.627377
director	-1.776813

Figure 6. Permutation Feature Importance

Next we will be covering figures that pertain to our AzureML experiments that aim to predict the duration based on other data in the dataset.. Here in figure 6 we can see the permutation feature importance of the columns included in the dataset with type being the most significant and director being the least. The higher the score the more impactful the column is at predicting the column in question.

Metrics

Mean Absolute Error	56.666623
Root Mean Squared Error	68.224924
Relative Absolute Error	0.777588
Relative Squared Error	0.779284
Coefficient of Determination	0.220716

Metrics

Mean Absolute Error	60.024264
Root Mean Squared Error	71.862782
Relative Absolute Error	0.793361
Relative Squared Error	0.842261
Coefficient of Determination	0.157739

Figure 7. AzureML Evaluate Model

Lastly, in the figure above we see evaluation results of the two models we incorporated into our experiments. Both models contain a relative squared error of 0.77 and 0.84 as well as a coefficient of determination of 0.22 and 0.15 respectively. Through these statistics we see that these models are performing poorly at predicting duration. We have tried utilizing different regression algorithms as we had used linear regression for this dataset. However, the relative squared errors and coefficient of determinations remained very similar.

5. Conclusion

In conclusion, our group successfully downloaded the dataset, uploaded it into Elasticsearch and Kibana. We then visualized the dataset by creating charts such as a bar chart, pie chart, and geo-spatial map through Kibana. We also created a dashboard to display all three charts in the same interface. Utilizing these charts we were able to find trends regarding Netflix's streaming database which can be utilized by Netflix or other streaming services in the entertainment industry to attain a competitive advantage amongst one another. Data analysis is incredibly important in business and can propel a company into the limelight if they are able to act upon information and data accordingly to be put in the best position possible.

References

[1] Matrix, S. (2014). The Netflix effect: Teens, binge watching, and on-demand digital media trends. Jeunesse: Young People, Texts, Cultures, 6(1), 119-138.https://muse-jhu-edu.mimas.calstatela.edu/article/553418/summary

[2] Jackson F., Amin R., Fu Y., Gilbert J.E., Martin J. (2015) A User Study of Netflix Streaming. In: Marcus A. (eds) Design, User Experience, and Usability: Design Discourse. Lecture Notes in Computer Science, vol 9186. Springer, Cham. https://doi-org.mimas.calstatela.edu/10.1007/978-3-319-20886-2_45

[3] Costa, C. D. (2020, April 19). *How Data Science is Boosting Netflix*. Medium. https://towardsdatascience.com/how-data-science-is-boosting-netflix-785a1cba7e45
[4] Kerschner, J. (2021, March 10). Tableau Public. https://public.tableau.com/profile/jack.kerschner#!/vizhome/Netflix_Catalogue/Story1.
[5] https://github.com/stran13/CIS3200Netflix
[6]https://gallery.cortanaintelligence.com/Experiment/Netflix-titles-2
[7]https://www.kaggle.com/shivamb/netflix-shows