CMPT 353

FINAL PROJECT REPORT

Netflix's Movies and TV Shows Data Analysis

Instructor: Steven Bergner

Tinh Van Phan - 301384974

Canh Nhat Minh Le – 301384865

April 15, 2022

Table of Contents

I. INTRODUCTION

By offering an enormous cinematic library of various genres, Netflix has inevitably become one of the most popular movie streaming services. However, having too many options to choose from might make subscribers overwhelmed, and raise the question: is Netflix prioritizing quantity over quality? Thus, one of our main goals in this project is to explore the quality of Netflix content by analyzing the different available genres and titles based on audience rating, retrieved from IMDb datasets. In order to do so, we set out a series of questions for further analysis:

1. How accurate can we predict the ranking of a title based on its genres?
2. Is there a difference in quality between all the available genres?
3. The number of Netflix's content in each genre?
4. The IMDB score of Netflix's content over the years ?
5. To help subscribers choose content easily, can we make a title recommendation system?

In order to reach our goal, we have applied a series of statistical analysis methods that we attained throughout this course. We first performed data extraction and data cleaning on the dataset of Netflix titles and IMDb ratings. For the score predicting task, we used the classification Machine Learning techniques and found out the best technique to use for prediction. Additionally, we utilized the ANOVA test to see the difference in quality between genres and find out the genre with the highest average rating score.

II. METHODS

1. Data extraction and cleaning:

Below are the datasets we used for this project:

a. **netflix_titles.csv:** attained from Kaggle. This dataset contains 8807 entries of available titles on Netflix (as of mid-2021), along with other details that are crucial to our project such as directors, cast, and release year.

b. **title.basic.tsv**: attained from IMDb website. This dataset contains over 8841109 entries of available titles on IMDb, including the *'tconst'* column that are crucial for merging with rating dataset. Since this dataset contains information of TV Episodes, we decided to remove all rows that are

marked as 'tvEpisode'. This dataset does not contain the rating for each title.

c. **title.rating.tsv**: attained from IMDb website. This dataset contains 1233157 entries of ratings achieved from taking the weighted average of user votings. Since this dataset also contains information of TV Episodes, we decided to remove all rows that are marked as 'tvEpisode'. We then merged this dataset with title.basic.tsv to create a new dataframe with titles and its rating.

After finishing extracting all data into data frames using the Pandas library, we then merge the Netflix dataframe with the IMDb dataframe to create a new .csv file of Netflix titles with audience rating. This is the file we proceed to utilize for the rest of our project.

2. Data analysis

a. Netflix title distribution between each genre

In order to do this task, we first splitted the genres of each row into a list. Second, we used the `explode()` function on the genres column. We then counted the occurrences of each genre and turned it into a new dataframe. Finally, we used the `barh(...)` function to draw a bar graph displaying the number of titles in each genre available on Netflix.

b. Average IMDB score of Netflix's titles over the years (2000-2020)

We extracted the titles, released years and the average score from the main dataset then calculated the average score of the movies and TV shows separately by years. We also separate them by genre and find the average score of each genre over the year. We chose the period from 2000 to 2020 because we will be able to see the trend from a reasonable amount of titles and the most recent period. Noticing that there are some genres that have fewer titles than the rest, we removed those with less than 500 titles.

c. Analysis of the difference in quality between genres throughout the years using ANOVA

From 27 available genres, we decided to select the top 11 genres with title count larger than 400. Before doing the ANOVA test, we applied the Central Limit Theorem on the 11 selected genres by grouping them by the release year and genre. We proceed to take the average in order to reach closer to normal distribution. To make sure the data is ready for the ANOVA test, we did the normal test and removed 4 genres (Animation, Horror, Mystery, and Thriller).

Genre	p-value
Action	0.32721797852539664
Comedy	0.5939379239282525
Drama	0.20523230633865158
Documentary	0.8761377953491402
Crime	0.640618042646904
Romance	0.975610596313673
Thriller	0.023222021340301373
Adventure	0.1321726960675279
Animation	0.005430939095576014
Horror	2.2210223442360937e-05
Mystery	0.03305830610319644

Additionally, we checked for equal variance by evaluating the p-value after using the stats.levene(..)function.

d. Title ranking prediction based on genres

Because predicting the IMDB score of a title requires various high-level data analysis skills, we decided to add another column called "ranking" to our dataset. It contains the ranking of tit based on the system shown below

```
# >8: critical acclaimed
# 6-<8: positive reviewed
# 4-<6: medium
# <4: negative
```

We removed the titles with undefined genres and separated the genres of other titles into lists using split() function. Then we used a MultiLabelBinarizer() object to create a DataFrame containing all the available genres as columns and the rows are the titles with the values 0 or 1 under the genre of that title, see a part of the table below

	Action	Adult	Adventure	Animation	Biography	Comedy
0	0	0	0	0	1	0
1	1	0	0	0	0	0
2	0	0	0	0	0	0

We proceed to 3 different machine learning models on the titles to see which one can produce the most accurate results. The 3 machine learning models we chose are KNeighborClassifier, RandomForestClassifier, GaussianNB, and we printed out the exact match score (EM) and the F-score (F1) of each model.

e. Title recommendation

We constructed a title recommendation system based on the genres, cast, and director(s) of each title. We studied an article from DataCamp in order to learn the proper steps to do this. Since the 'netflix_data.csv' dataset already contains the required information, we only needed to split each cast and director entry into a list. We then converted the name and genre instances into lowercase and stripped spaces between each of them. Next, we created a new column called "soup", in which each entry is a string containing all the director, cast, and the genre(s) of that title. This step is crucial for the `CountVectorizer()` we used afterwards.

	title	director	cast	genres
1	ganglands	[julienleclercq]	[samibouajila, tracygotoas, samueljouy, nabiha...	[action, crime, drama]
3	midnight mass	[mikeflanagan]	[katesiegel, zachgilford, hamishlinklater, hen...	[drama, fantasy, horror]
4	my little pony: a new generation	[robertcullen, joséluisucha]	[vanessahudgens, kimikoglenn, jamesmarsden, so...	[adventure, animation, comedy]
5	sankofa	[hailegerima]	[kofighanaba, oyafunmikeogunlano, alexandradua...	[drama]

Figure 2.2.b.1: Before converting to "soup"

	soup
1	samibouajila tracygotoas samueljouy nabihaakka...
3	katesiegel zachgilford hamishlinklater henryth...
4	vanessahudgens kimikoglenn jamesmarsden sofiac...
5	kofighanaba oyafunmikeogunlano alexandraduah n...

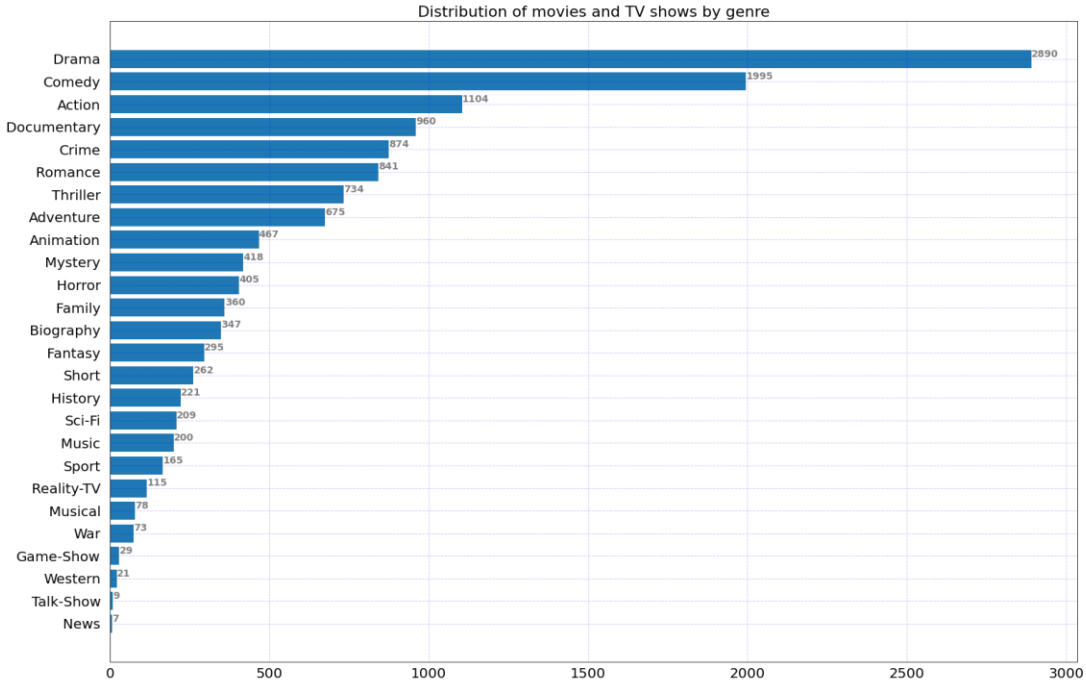
Figure 2.2.b.2: After converting to "soup"

Moving forward, we utilized the `cosine_similarity` function to calculate the distance between the embeddings. Finally, we created a `get_recommendation(...)` function that takes two parameters: the title of a movie/tv show that we want to get similar recommendations from and the list of cosine similarity scores between that title and other titles in the dataset. This function will later return the top 10 recommendations.

III. RESULTS

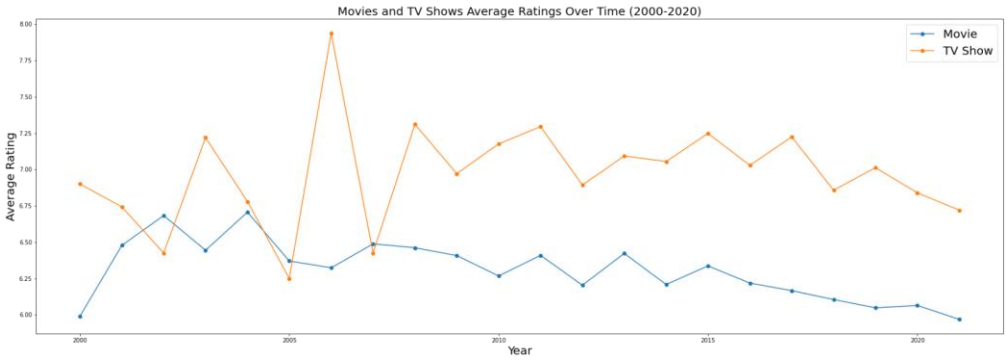
1. Distribution of titles between each genre

The bar graph below displays the distribution of Netflix titles in each genre.

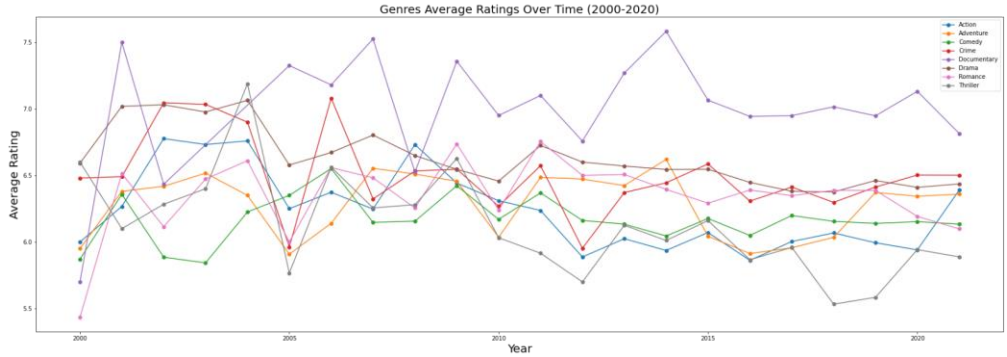


We can see that as of mid-2021, Netflix has the most titles that are classified as "Drama", while the genre "News" has the least titles. The top 3 most genres with the most content on Netflix are "Drama", "Comedy" and "Action", the least ones are "News", "Talk-show" and "Western". We can see the audience preference of entertainment very clearly from this graph.

2. Average IMDB score of Netflix's titles over the years (2000-2020)

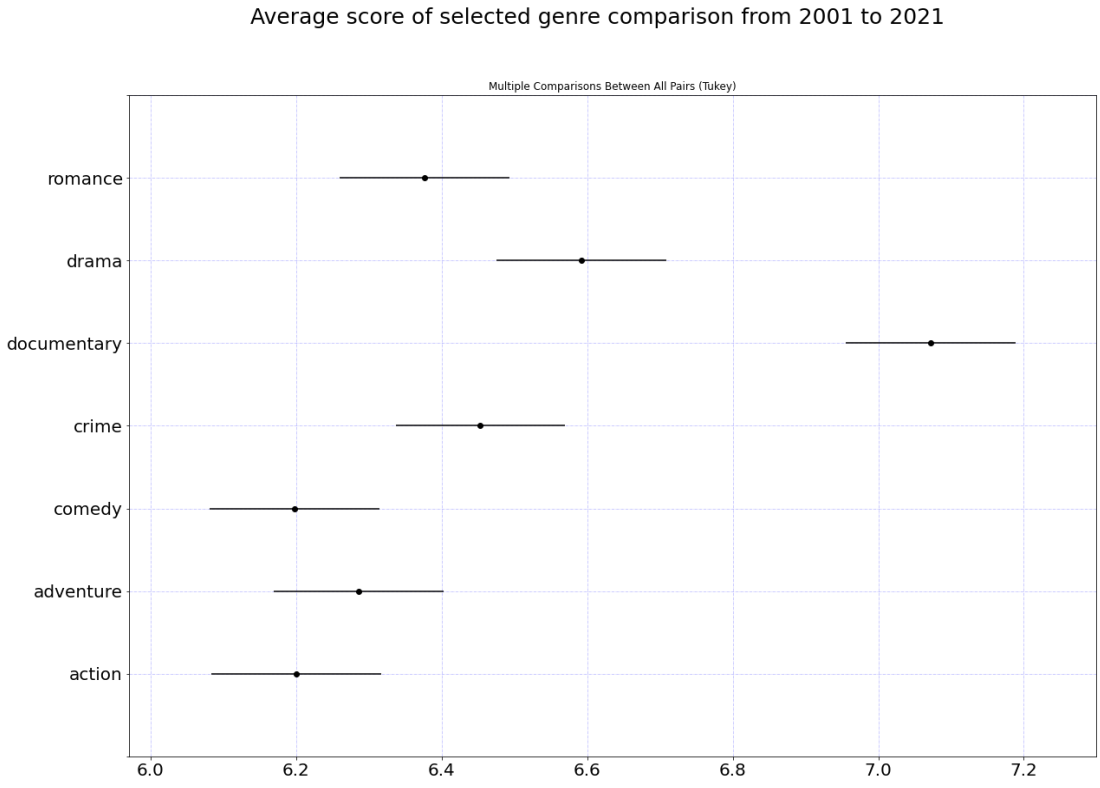


Throughout the period of 2000 to 2020 we can see a significant difference in the average IMDB scores between Netflix's movies and TV shows. In general, TV shows have higher ratings than movies.In 2001, 2005 and 2007 movies' ratings dropped lower than movies but they always rose up again in the next year; especially in 2006, the IMDB rating of TV shows peaked at 7.99.



Documentaries started second lowest in the chart but they peaked in the next year at 7.5 and continued to have a significantly higher level of rating than the other genres. While the drama genre is pretty stable throughout the years, action, adventure, crime and drama fluctuate a lot during this period. While comedy and romance started and grew gradually in the following years, thriller started in a high place but their rating decreased and became the lowest rated genre in 2020.

3. Analysis of the difference in quality between genres throughout the years using ANOVA



The ANOVA test resulted in a p-value of 9.656284532297942e-23, indicating that there is a statistical difference in the yearly average rating score between each genre. Based on the graph above, it is apparent that the 'Documentary' genre has the highest average rating score throughout the selected time period.

4. Tile ranking prediction based on genres

From the models created, we collected the score for each models:

Models	EM	F1
KNN	0.521	0.530
Random Forest	0.655	0.552
Naive Bayes	0.127	0.169

Random Forest model has the highest score, which shows that it is the most suitable model for predicting the ranking of Netflix's titles.

5. Title recommendation system

We input the movie "Before I Wake", which was classified as a Horror/Drama/Fantasy movie and the system output 10 titles with similar genres and directors.

Input:

title	director	cast	rating	duration	originalTitle	year	genres
before i wake	Mike Flanagan	Kate Bosworth, Thomas Jane, Jacob Tremblay, An...	PG-13	97 min	Before I Wake	2016	Drama,Fantasy,Horror

Output:

title	director	cast	rating	duration	originalTitle	year	genres
midnight mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	TV-MA	1 Season	Midnight Mass	2021	Drama,Fantasy,Horror
the haunting in connecticut 2: ghosts of georgia	Tom Elkins	Abigail Spencer, Chad Michael Murray, Katee Sa...	R	101 min	The Haunting in Connecticut 2: Ghosts of Georgia	2013	Drama,Horror,Mystery
the old ways	Christopher Alender	Brigitte Kali Canales, Andrea Cortes, Julia Ve...	TV-MA	90 min	The Old Ways	2020	Drama,Fantasy,Horror
ghoul	Patrick Graham	Radhika Apte, Manav Kaul, Ratnabali Bhattachar...	TV-MA	1 Season	Ghoul	2018	Drama,Fantasy,Horror
the 3rd eye 2	Rocky Soraya	Jessica Mila, Bianca Hello, Nabilah Ayu, Sophi...	TV-MA	117 min	Mata Batin 2	2019	Drama,Fantasy,Horror
gerald's game	Mike Flanagan	Carla Gugino, Bruce Greenwood, Henry Thomas, C...	TV-MA	103 min	Gerald's Game	2017	Drama,Horror,Thriller
wildling	Fritz Böhm	Bel Powley, Brad Dourif, Liv Tyler, Collin Kel...	R	93 min	Wildling	2018	Drama,Fantasy,Horror
the maus	Yayo Herrero	Alma Terzic, August Wittgenstein, Aleksandar S...	TV-MA	90 min	The Maus	2017	Drama,Fantasy,Horror

IV. LIMITATIONS

We experienced some difficulties dealing with duplicate titles; in particular, to get the imdb score of the titles in the netflix_titles.csv file, we have to perform merging twice to get the IMDB scores of all titles and then the titles available on Netflix. We encountered issues with duplicates because of the large number of similar titles. We decided to remove the titles with the type "TvEpisode" and perform inner join on the primary title and the released year of the titles. It took us more than 10 hours to figure it out.

We also found it difficult trying to predict the IMDB score of the titles. Since the IMDB score is numerical, it required more knowledge than what we learned in the course. Instead of predicting the IMDB scores, we changed to prediction of the ranking of Netflix's title.

V. PROJECT EXPERIENCE

Canh Nhat Minh Le

- Collaborated with partner to come up with ideas and plan tasks
- Fluently applied acquired course knowledge in data cleaning and preparation using Pandas library
- Learned and built a title recommendation system by calculating the cosine similarity score
- Utilized the matplotlib library to construct necessary graphs for data visualization
- Conducted the ANOVA test to analyze average rating score between different genres

Tinh Van Phan

- Performed research on chosen topic and efficiently discussed with partner on the problems statements
- Worked closely with partner in data preparation and problem solving
- Fluently applied machine learning concepts and sufficiently undertook classification problems
- Constructed and modified multiple lines plot using matplotlib
- Learned Multi Label Classification using MultiLabelBinarizer in sklearn.preprocessing library and successfully applied to title ranking prediction

VI. REFERENCE:

https://www.geeksforgeeks.org/bar-plot-in-matplotlib/

https://www.datacamp.com/community/tutorials/recommender-systems-python

https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification

https://www.kaggle.com/discussions/questions-and-answers

https://jovian.ai/jandrewtomich/new-netflix-and-imdb-analysis?fbclid=IwAR38qlgBMhqutyv6BsGuN1YpVNRiZIvnGZcVahdZI__KTunCJC5o7Yg0f3Q

https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/?fbclid=IwAR1s91P_3vl54K8mxpCdvZ-6Gnj7FdLl9USvrMxcIUG9Bx1pXarlQoTYqfs

https://www.imdb.com/interfaces/?fbclid=IwAR204w_cCrqOLFx6_6hXLIo_kPbWc9LbOhFTZjvdzJkbK-Ym3-YsxXkKNlM

https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset?resource=download&select=title.basics.tsv.gz&fbclid=IwAR214H5hQR_dx5LyencR2ib2e1fz-3SFjwVHDc5Hbb2o0SNj0nXRa6lCJFc

https://www.kaggle.com/datasets/sankha1998/tmdb-top-10000-popular-movies-dataset?fbclid=IwAR2Y_3j3h9IrJN2HDfRZyHuCLNLgIdCU5NLZdQSQl3sktX59LLMu7jhZjZQ