**Netflix Data Exploration**

**A blue and black logo

Description automatically generated**

**CS 661 – Python Programming**

**Spring 2024**

|  |  |
| --- | --- |
| Under the supervision of: | Submitted by: |
| **Dr. Kshitij Sharma** | **Harsh Modi**  **Dhruv Chamaria**  **Nitant Bhagat** |

**Abstract**

In this project, we conducted a comprehensive analysis of Netflix's extensive library of movies and TV shows, leveraging a tabular dataset containing detailed listings of content along with various attributes such as cast, directors, ratings, release years, durations, and more. Netflix requires to understand consumer preferences and content trends because of its big global subscription services base. Through analysis, we are able to gain insights into a variety of topics, including the evolution of material over time, genre preferences, the impact of directors and cast, rating distributions, duration patterns, distribution by region, and date-added trends. The results provide insightful information for subscribers, platform managers, and content providers, facilitating informed choices and improving our knowledge of the patterns in digital media consumption.

**Introduction**

Netflix is a well-known brand in the current digital era, providing a wide range of films and TV series to audiences worldwide. With over 8600 titles available, Netflix has transformed how we consume entertainment. For both producers and spectators, it is crucial to comprehend the motivations behind this enormous collection. We'll go deeply into Netflix's content environment in this project, looking at popular genres, release trends, director influence, audience ratings, and worldwide distribution. By analyzing this data, we hope to uncover insights that provide valuable perspectives on how we engage with digital media.

**Dataset**

The dataset provides a comprehensive listing of movies and TV shows available on Netflix, encompassing various attributes such as type (movie or TV show), title, director, cast, country of production, date added to Netflix, release year, content rating, duration (in minutes or seasons), genre, and a brief description. Since Show ID is not necessary for the analysis, it will be deleted. This dataset, which includes approximately 8600 entries from 2008 to 2021, provides insights into the wide range of material offerings from one of the top streaming services in the globe.

**Preprocessing**

Data preprocessing is necessary to prepare the data before we can begin analyzing it. We cleaned up missing information, gave columns more understandable names, and got rid of unnecessary data like show IDs. Additionally, we ensured that no duplicate entries were causing issues. We transformed dates into a more readable format and removed unnecessary data like the day, month, and year for better understanding. Overall, we organized the data to make it simpler to work with and evaluate.

**Architecture/Methodology**

1. **Data Collection and Preprocessing:**
   * Gathered data on Netflix movies and TV shows.
   * Cleaned and preprocessed the dataset.
   * Handled missing values, encoded categorical variables, and extracted relevant features.
   * Selected, transformed, and created new features to improve model performance.
   * Extracted features such as release year, type (movie or TV show), and rating category.
   * Encoded categorical variables for modeling purposes.
2. **Exploratory Data Analysis (EDA):**
   * Conducted exploratory data analysis to understand dataset characteristics.
   * Examined summary statistics and visualized data distributions.
   * Explored relationships between variables using pie charts, count plots, and scatter plots.
3. **Modeling:**
   * Applied various machine learning algorithms for predictive modeling.
   * Utilized techniques like linear regression, logistic regression, decision trees, and random forests.
   * Trained models on subsets of the data and evaluated their performance using appropriate metrics.
4. **Model Evaluation and Interpretation:**
   * Evaluated trained models using cross-validation techniques and performance metrics.
   * Analyzed model interpretations, including feature importance and decision rules.
   * Assessed effectiveness in predicting target variables such as content duration, type, or audience rating.
5. **Visualization:**
   * Visualized analysis results using plots, charts, and tables.

**Results**

* Movies vs TV Shows Distribution:
  + Approximately 70% of the content on Netflix consists of movies, while the remaining 30% are TV shows.
* Distribution of Content based on Year Released:
  + The highest number of movies were released in 2018.
* Content Released per Year:
  + The year 2018 had the highest number of movie releases. TV show production surged after 2015, surpassing movie production in 2021.
* Content across Countries:
  + The US has the most content for movies, followed by India and the UK. For TV shows, the US and UK are leading producers, with Japan and South Korea also notable.
* Top 10 Directors for Movies and TV Shows:
  + Rajiv Chilaka, Jan Suter, and Raul Campos are the most active directors for movies. Each directed around 20 movies. For TV shows, directors typically have fewer productions, ranging from 2 to 3 shows.
* Top 10 Genres for Movies and TV Shows:
  + The International genre is the most common for both movies and TV shows, followed by Drama and Comedy.
* Count plot on Movies and TV shows by country:
  + Japan and South Korea has only Tv shows released in their country ,while Indonesia ,Hongkong and Philippines has only movies.
* Classification of Content on Netflix by Age Group:
  + Most content on Netflix is produced for adults, followed by teens and kids. This indicates a skew towards mature audiences.
* Movies Added on Netflix by Year, Month, and Weekday:
  + Most movies were released in 2019 and 2020, with July being the most popular month. Fridays see the highest number of releases.
* TV Shows Added on Netflix by Year, Month, and Weekday:
  + Similar to movies, most TV shows were released in 2019 and 2020, with December being the most popular month. Fridays also see the highest number of releases.
* Average Duration of Movies and TV Shows:
  + Most movies have durations between 65 and 125 minutes, while TV shows typically consist of around 2 seasons.
* Top 20 TV Shows by number of seasons
  + Grey’s anatomy Has more than 15 seasons and top 14 TV shows has minimum 10 seasons and some of them has more than that.
* Top 10 Actors in Movies and TV shows:
  + Anupam Kher leads with the highest number of movies, followed by Shah Rukh Khan and Naseeruddin Shah.
  + Unlike movies, TV shows feature actors from various countries in the top 10 list.
* Distribution of ratings among genres
  + International and Drama is mostly rated by adults. Children movies and Kids TV is only rated by kids.
* Distribution of Genres Across Countries:
  + Most famous type of genre is international and drama among countries.
* Distribution of ratings among countries
  + Only adults and children have rated on Netflix in the US. Only India has ratings from adults, children, and teens all three of them.

**Conclusion**

Many important insights into the platform's content landscape and viewer preferences are obtained from this analysis of the Netflix dataset. One thing that stands out is Netflix's commitment to providing a wide range of entertainment options, including a plethora of movies spanning different genres. However, there's a noticeable trend towards producing more TV shows in recent years, indicating a shift in audience preferences towards serialized storytelling. Comedy, drama, and international genres tend to appeal to viewers everywhere. Moreover, our analysis sheds light on release patterns, average content duration, and audience demographics, providing valuable insights that can inform content creation and platform strategy decisions.