Kurt Lanting – 2021-09489 (THV)
Raphael Tan – 2021-21335 (THU)
Gabbie Purisima – 2021-10796 (THV)

Movies and TV Shows Age Rating Predictions Across Multiple Streaming Platforms via Naive Bayes

Project Summary

The project sets out to pioneer a Machine Learning Model using Naive Bayes solution, a novel approach, to accurately predict age ratings for movies and TV shows on major streaming platforms such as Netflix, Disney+, Amazon, Prime Video, and Hulu.

With the wide variety of content available, the major challenge is the need for precise age ratings or restrictions for movies and TV shows on the abovementioned streaming platforms. It becomes difficult for viewers to select age-appropriate material, resulting in the need for an automated solution to predict age ratings accurately.

This problem is of utmost importance as it directly affects the viewer's ability to make informed decisions about the content they are about to watch, especially concerning age-appropriateness. This project will benefit viewers by enhancing their viewing experience and promoting responsible media consumption habits by providing an accurate age rating prediction. Moreover, by implementing improved content moderation, streaming platforms cannot only gain popularity and attract a larger audience but also ensure a safer and more tailored entertainment experience for their users. This enhancement could increase user satisfaction, thereby boosting their position in the competitive streaming industry.

Background

We cannot deny that movies and TV shows greatly influence our lives in various ways. As mentioned by Rothwell, movies and TV shows serve as a mirror to society, reflecting its joy, struggles, and complexities. With its increasing popularity and the coming of the digital age, the entertainment industry has seen a significant shift towards digital streaming platforms, providing viewers with unlimited access to a vast amount of movies and TV shows. However, one prevailing problem is that viewers are presented with many films and TV shows, and many still need to be labeled or with a rating (Powell, 2021). This lack of precise age restrictions or ratings can lead to the viewer being poorly directed towards a movie or TV show content suitable for their age group.

The proposed project aligns with the theme ‘AI for Everyday Life,’ which addresses a familiar real-world problem millions of viewers worldwide face. As streaming platforms continue to offer an ever-expanding array of content, the need for an accurate age rating becomes increasingly appropriate. By utilizing AI techniques, we can automate the process of age-rating predictions for movies and TV shows across multiple streaming platforms, making entertainment consumption safer and more tailored to individual preferences. This project can enhance the viewing experience and ensure the appropriateness of content on various streaming platforms for viewers.

A successful implementation of the project can make a significant impact on both viewers and entertainment streaming platforms. By providing accurate and standardized age rating predictions, we can empower viewers to make more informed decisions on the content they consume and promote responsible viewing habits. This, in turn, can lead to a more enjoyable and ethically responsible consumption of movies and TV shows. In addition, streaming platforms can increase user satisfaction by improving content moderation practices, thereby contributing to the improvement of the entertainment industry as a whole.

Materials and Methods

Data Collection

Data Source

Our data source will be Kaggle, specifically from datasets containing information on movies and TV shows on Netflix, Disney+, Amazon Prime Video, and Hulu. Each dataset includes features such as show ID, type (movie or TV show), title, director, cast, country of production, date added, release year, rating (which will be our target feature), duration, genre, and description. These datasets provide comprehensive content listings on each streaming platform, allowing in-depth analysis and modeling. Altogether, they comprise 23,863 records.

Justification for the Chosen Datasets

- **Relevance**: These datasets are highly relevant to the theme "AI for Everyday Life," as streaming services are integral to daily entertainment for millions worldwide.

- **Diversity**: The datasets span four major streaming platforms, providing a diverse range of content and subscriber demographics. This diversity allows for more comprehensive analysis and model building.
- **Data Richness**: Each dataset contains detailed information about movies and TV shows, including metadata such as cast, directors, ratings, release year, duration, and description. This richness enables a variety of potential analyses and AI applications.
- **Popularity**: With millions of subscribers, these platforms represent a significant portion of the global streaming market. Analyzing data from these platforms can yield insights with broad applicability.

Data Description

The datasets from Netflix, Disney+, Amazon Prime Video, and Hulu contain comprehensive information about movies and TV shows available on these platforms. Each dataset includes the following features:

- **Show ID:** A unique identifier for each movie or TV show.
- **Type**: Indicates whether the entry is a movie or a TV show.
- **Title**: The title or name of the movie or TV show.
- **Director**: The director(s) of the movie or TV show.
- **Cast**: The cast members of the movie or TV show.
- **Country**: The country of production for the movie or TV show.
- **Date Added**: The date when the movie or TV show was added.
- **Release Year**: The year when the movie or TV show was released.
- **Rating**: The rating or film rating assigned to the movie or TV show (this will be our target feature).
- **Duration**: The total duration of the movie or TV show.
- **Listed In:** The genre(s) of the movie or TV show.
- **Description**: A description of the movie or TV show, providing insight into its content and themes.

The dataset consists of both categorical and numerical data types. Categorical features include type (movie or TV show), country, and listed (genre), while numerical features include release year and duration. The target feature, rating, is also definite, representing different levels of viewer ratings assigned to the movies or TV shows.

Methodology

Data Preprocessing

The dataset will undergo preprocessing steps to ensure its suitability for the model. This includes cleaning the data to handle missing values and outliers. For the description feature, text preprocessing techniques such as tokenization, removing stop words, and stemming or lemmatization will be applied to standardize the text data and prepare it for modeling.

Feature Extraction

For predicting the target feature, which is the age rating assigned to the movie or TV show, the following features from the dataset will be used:

- **Description**: A detailed description of the movie or TV show provides direct information about its content and themes, which can significantly impact the assigned age rating.
- **Rating**: The age rating or film rating assigned to the movie or TV show, such as PG, G, R, TV-MA, etc. This will serve as our target feature, representing the desired outcome of the prediction task.

These features collectively contribute to determining the appropriate age rating for movies and TV shows and are relevant for predicting the target feature.

Model Selection

The chosen model for this project is Multinomial Naive Bayes. This model is suitable for text classification tasks and is particularly effective when dealing with textual data such as movie or TV show descriptions.

Model Implementation

The implementation of the Multinomial Naive Bayes model involves several steps. First, the description feature will be preprocessed to convert the text into numerical representations. CountVectorizer or TF-IDF will be tested for best suitability. This process will generate a document-term matrix, where each row represents a movie or TV show description, and each column represents a tokenized word from the descriptions. The model will be trained using the transformed data, with the rating as the target variable. Predictions on new data will be made using the trained model.

Evaluation Metrics

Our evaluation metrics will focus on the following metrics to assess the model's performance in predicting age ratings:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision**: Indicates the proportion of correctly predicted age ratings out of all expected age ratings.
- **Recall**: Measures the proportion of correctly predicted age ratings out of all actual age ratings.
- **F1 Score:** Harmonic mean of precision and recall, providing a balanced measure of the model's performance.
- **Confusion Matrix**: Provides insights into the types of errors made by the model, including true positive, false positive, true negative, and false negative predictions.

By leveraging these evaluation metrics, we aim to gauge the Multinomial Naive Bayes classifier's effectiveness in accurately predicting age ratings assigned to movies or TV shows based on their features.

References and Data Sources

Rothwell, J. (2019, July 25). You are What You Watch? The Social Effects of TV. The New York Times. Retrieved May 15, 2024, from https://www.nytimes.com/2019/07/25/upshot/social-effects-television.html

Powell, R. (2021, November 15). What are TV and Streaming Platform Age Ratings? Spherex. Retrieved May 15, 2024, from https://www.spherex.com/parental-guidelines-for-tv-and-streamers

Netflix Movies and TV Shows. Retrieved May 15, 2024, from https://www.kaggle.com/datasets/shivamb/netflix-shows

Amazon Prime Movies and TV Shows. Retrieved May 15, 2024, from https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows

Disney+ Movies and TV Shows. Retrieved May 15, 2024, from https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows

Hulu Movies and TV Shows. Retrieved May 15, 2024, from https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows

Tools and Technologies

- **Python**: The primary programming language used for data preprocessing, model training, and evaluation.
- **Pandas**: Utilized for data manipulation and analysis, including reading the dataset from a CSV file, handling missing values, and structuring the data for model training.
- **scikit-learn**: A machine learning Library in Python that provides efficient data preprocessing, modeling, and evaluation tools. Specifically, sci-kit-learn's CountVectorizer and TF-IDF are employed to convert text data into numerical representations, and MultinomialNB is used to implement the Multinomial Naive Bayes model.
- **Jupyter Notebook**: The interactive development environment used for writing and executing Python code, allowing for easy experimentation and documentation of the modeling process.