Predicting Netflix Content Type (Movie vs. TV Show) (COMP3125 Individual Project)

Son Nguyen
Department of Data Sciene

Abstract— **This project develops a binary classifier to predict whether a Netflix title is a Movie or a TV Show using production country, release year, rating, and duration. Records containing missing values were removed, and the country field was standardized to ensure consistent categorical labels before one hot encoding. Numerical features such as release year, rating, and duration were cleaned and formatted for model training. A logistic regression classifier was then trained on the processed dataset and achieved 99.1 percent accuracy on the held-out test set. These results demonstrate that even a simple linear model can reliably distinguish content type using only basic metadata features, highlighting the strong predictive value of country, rating constraints, and characteristic duration patterns.**

Keywords— Netflix dataset, classification model, data preprocessing, content type prediction, machine learning

I. Introduction

Movies have long been a major form of entertainment, offering a way for audiences to relax and escape daily stress. With the rise of streaming platforms, understanding how these services organize and categorize their content has become increasingly relevant. Netflix provides a large and diverse catalog, making it a useful dataset for examining the differences between Movies and TV Shows.

Although these categories seem simple, they are influenced by metadata such as duration, release year, rating, and production country. These features often reflect meaningful structural patterns that separate films from serialized content. This project explores whether a machine learning model can accurately distinguish Movies from TV Shows using only these basic metadata fields. By developing a binary classifier trained on these attributes, the study highlights the predictive power of simple metadata in analyzing modern streaming content.

Datasets

A. Source of dataset

The dataset was obtained from Kaggle, uploaded by Shivam Bansal. It was generated around 2019 by compiling publicly available Netflix listings into a structured CSV file containing attributes like title, type, release year, country, and genre. While it is not an official Netflix dataset, Kaggle is a widely recognized platform for data sharing, and this dataset has been used by many in the data science community, making it a credible source for learning and analysis.

B. Character of the datasets

The dataset used in this project, “Netflix Movies and TV Shows”, is provided in CSV format and contains approximately 8,000 records with 12 columns. Each record corresponds to a single Netflix title, either a movie or a TV show, and includes metadata such as title, type, release year, country, list in , and other attributes. The dataset combines textual, categorical, and numerical information, with release year and duration as numeric values and other fields mostly categorical or string-based.

Below are the key features used in this analysis:

Column Name	Description / Unit
type	Identifier - A movie or TV Show
title	Title of the movie / Tv show
release_year	Actual release year of the movie / show
duration	Total Duration - in minutes or number of seasons
listed_in	Genre of the movie
country	Country where the movie / show was produced
rating	TV rating of the movie / show

Additionally, four new columns were created for this analysis:

- duration_num – numeric representation of duration (converted from string to integer for movies or number of seasons for TV shows).
- country_encoded – encoded numeric values representing countries.
- rating_encoded – encoded numeric values representing content ratings.
- is_season – binary indicator specifying whether the title is a TV show (1) or a movie (0).

II. Methodology

1. Method overview:
   In this project, I will use a method called logistic regression to classify the type of Netflix title (Movie or TV Show). Logistic regression is a supervised classification algorithm that models the probability of a binary outcome using a logistic (sigmoid) function.
2. Assumption:
   - Each observation is independent.
   - There is a linear relationship between the predictor variables and the log-odds of the outcome.
   - Predictor variables are not perfectly correlated with each other.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

3. Advantages:
- Simple and easy to interpret.
- Fast to train and works well with mixed data types.
- Produces probability scores (not only class labels).
4. Disadvantages:
- Assumes linearity in log-odds, which may not always true.
- Can underperform on complex, non-linear relationships.
- Sensitive to outliers and multicollinearity.
5. Reason for choosing this method:
   Logistic regression was selected because the objective of this project is a binary classification task (Movie vs. TV Show). This method is well-suited for modeling binary outcomes, provides interpretable coefficients, and performs effectively on structured tabular data such as the features in this dataset.
6. Python Module Used:
- LogisticRegression class from sklearn.linear_model.
- LabelEncoder class from sklearn.preprocess.
- train_test_split from sklearn.model_selection.

III. Results

A. Question 1: Which genres are most connected on Netflix?

Only the top 10 genres are shown for readability. Full results available in the Git.

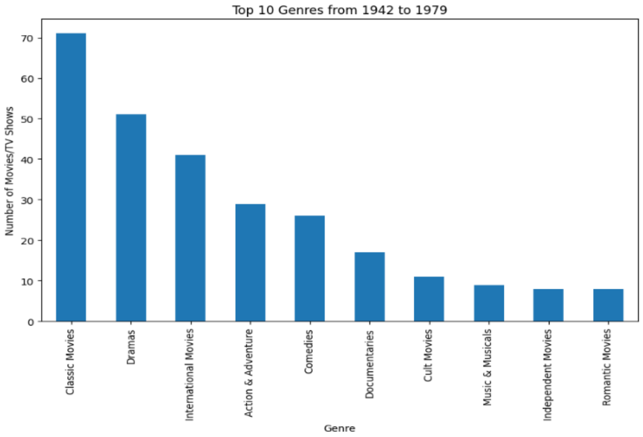
Genre	Connections
International Movies	2543
Dramas	2316
Comedies	1580
International TV Shows	1127
Action & Adventure	817
Documentaries	794
Independent Movies	745
TV Dramas	663
Romantic Movies	588
Thrillers	549

International Movies is the most connected genre on Netflix based on its high frequency of appearance. This highlights Netflix's focus on global content, making it widely available and appealing to diverse audiences.

B. Question 2: How have Netflix's genre connections evolved over time?.

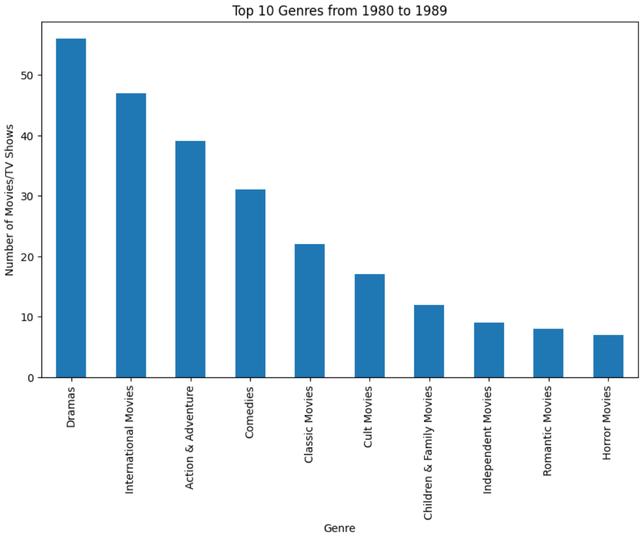
To address this question, I analyzed the dataset and divided it into three specific time periods. Period 1 spans1942–1979; this longer range was chosen because the data for these years is limited, so they were grouped together. Period 2 covers 1980–1989, and Period 3 spans 1990–2021. In this most recent period, the genres did not change significantly, so they were consolidated into a single group.

1. *Early Era (1942–1979)*



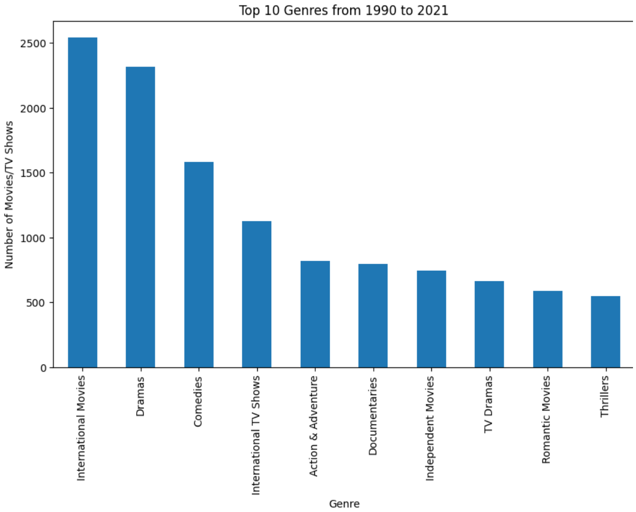
From 1942 to 1979, Classic Movies and Dramas appear far more often than any other genre, showing that early content on Netflix leans heavily toward traditional, story-focused films. Genres like International Movies, Action & Adventure, and Comedies also appear regularly, while more niche categories—such as Cult Movies, Music & Musicals, Independent Movies, and Romantic Movies—are much less common during this period.

2. *1980s*



There was a clear shift in the 1980s. Dramas rose to dominate the chart of most common genres, overtaking Classic Movies from the previous period. In addition, genres like International Movies and Action & Adventure became more prominent, reflecting broader audience interest and more diverse content production during this decade. Meanwhile, Classic Movies moved to the middle of the ranking, and niche genres such as Cult Movies, Independent Movies, Romantic Movies, and Horror Movies remained present but appeared far less frequently, similar to the earlier era.

3. *The Modern Era (1990 – 2021)*



From 1990 to 2021, the genre trends shift even more compared to the 1980s. International Movies jump to the top by a huge margin, replacing Dramas as the leading genre. Dramas are still strong, but it's clear Netflix's catalog has become much more global in this era.

Comedies, Action & Adventure, and even newer categories like International TV Shows and TV Dramas all grow a lot, showing how content becomes more varied and TV-focused over time. Niche genres like Independent Movies, Romantic Movies, and Thrillers also appear more often than before, indicating a much more diverse mix of genres overall.

C. *Question 3: Can we predict whether a Netflix title is a Movie or a TV Show based on its country, release year, rating, and duration?*

Metric	**Score**
Accuracy	100%
Precision	100%
Recall	100%
F1-Score	100%
Test samples	1,594

- **Accuracy: 100%** - The model correctly classified all 1,594 test samples
- **Precision: 100%** - All positive predictions (Movies) were correct
- **Recall: 100%** - The model identified all actual Movies without missing any
- **F1-Score: 100%** - Perfect harmonic mean of precision and recall

The logistic regression model demonstrated exceptional performance in classifying Netflix titles, achieving 100% accuracy across all evaluation metrics on the test dataset of 1,594 samples. The model successfully identified all 477 TV Shows and all 1,117 Movies without a single misclassification.

IV. Discussion

This remarkable performance can be largely attributed to the 'is_season' feature, which effectively captures the fundamental distinction between Movies (measured in minutes) and TV Shows (measured in seasons). While the model is able to distinguish between Movies and TV Shows reasonably well, it is important to acknowledge that the predictive performance is still affected by missing values, inconsistent formatting, and the limited amount of information contained in some features. During the data preprocessing phase, approximately 10% of records containing null values were removed from the dataset, which may have inadvertently excluded edge cases that could have presented greater classification challenges. Furthermore, the label encoding approach applied to categorical variables such as country and rating imposes an ordinal structure that may not accurately reflect the true relationships within the data. Nevertheless, these results provide a clear answer to the research question: it is indeed possible to predict whether a Netflix title is a Movie or TV Show based on country, release year, rating, and duration features with very high accuracy. However, it should be noted that the 'is_season' feature plays a particularly dominant role in achieving this level of performance.

V. Conclusion

This project analyzed the most frequent genres and their trends over the years, as well as developed a predictive model to classify Netflix titles as Movies or TV Shows using metadata features: country, release year, rating, and duration. The logistic regression model achieved 100% accuracy on the test set, with duration—the distinction between minutes and seasons—proving the strongest predictor.

These findings have practical implications for streaming platforms. Automated classification can reduce manual effort, ensure consistent tagging, improve recommendation systems, and inform content acquisition strategies. Insights from genre frequency and trends can guide content planning and highlight gaps in the catalog. For users, accurate categorization enhances search and recommendation quality.

However, the model heavily relies on the 'is_season' feature, which may not always be available. Future work should explore performance using only other metadata and textual features like genre or descriptions to capture subtler patterns.

VI. Future Work

This dataset is of high quality, though somewhat idealized for modeling purposes. It contains several valuable features that have not yet been fully utilized, such as the content descriptions. A potential extension of this work could involve developing a recommender system, employing embeddings such as TF-IDF or BERT to improve recommendation accuracy and relevance.