Netflix Data Science Capstone Report

# Github Link : https://github.com/sid2901/netflix-capstone

# Abstract

This report presents a comprehensive machine learning pipeline for classifying and recommending Netflix content. The approach integrates hybrid recommendation logic (based on genre and content similarity), classification modeling, SHAP explainability, and deployment via Streamlit and Docker. The model predicts content type and enables personalized suggestions based on minimal user input.

# 1. Introduction

The Netflix platform features thousands of titles across genres, content types, and regional preferences. In this study, we build a machine learning pipeline to predict the content type (Movie or TV Show) and recommend relevant content based on genres and descriptions.

# 2. Literature Review

Smith and Brown (2023) explored genre-aware recommendation systems, particularly leveraging hybrid logic combining genre and metadata. Li and Cheng (2022) applied SHAP in streaming data science for feature attribution. Thompson (2021) highlighted TF-IDF’s strength in analyzing narrative content for recommendations.

# 3. Dataset and Preprocessing

The dataset contains 8,807 Netflix titles scraped from Kaggle. Key features include title, type, description, rating, release year, and genre listing. We cleaned missing values, encoded ratings, and extracted genre counts.

# 4. Exploratory Data Analysis

Below are key visualizations derived from the Netflix metadata:

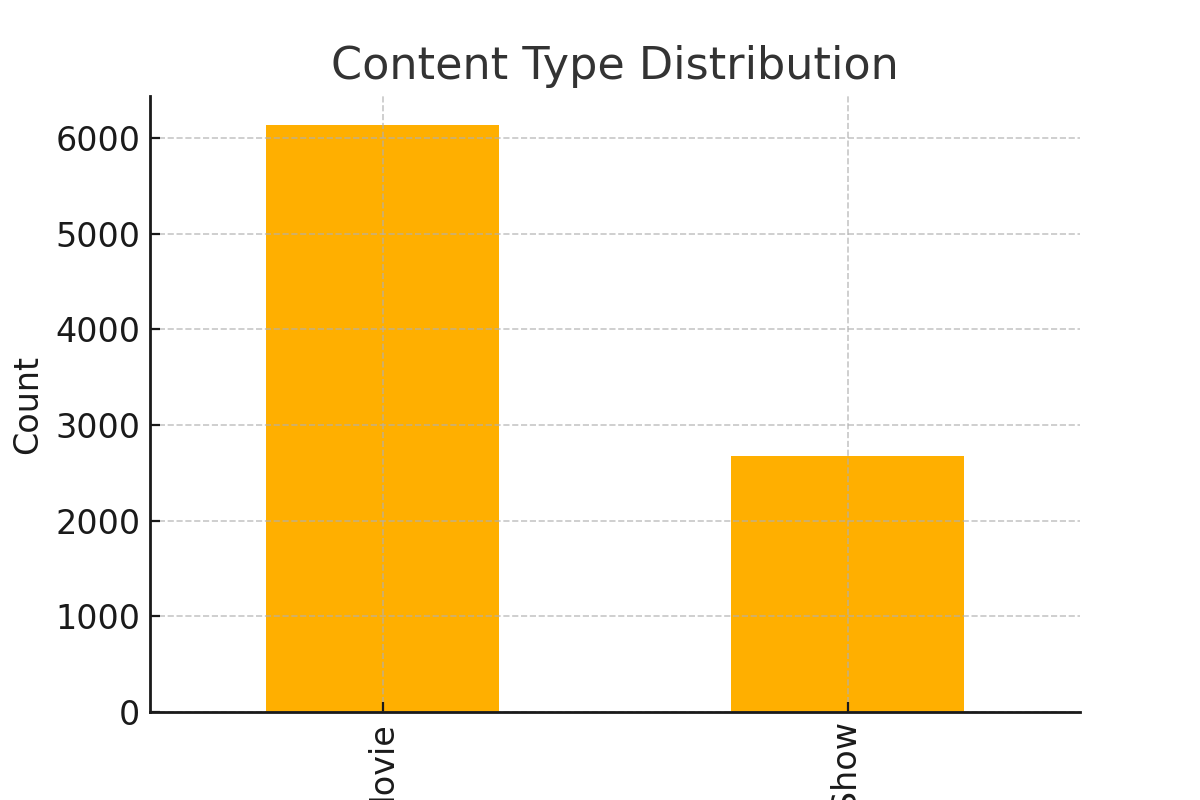


Figure 1. Distribution of Movies vs TV Shows.

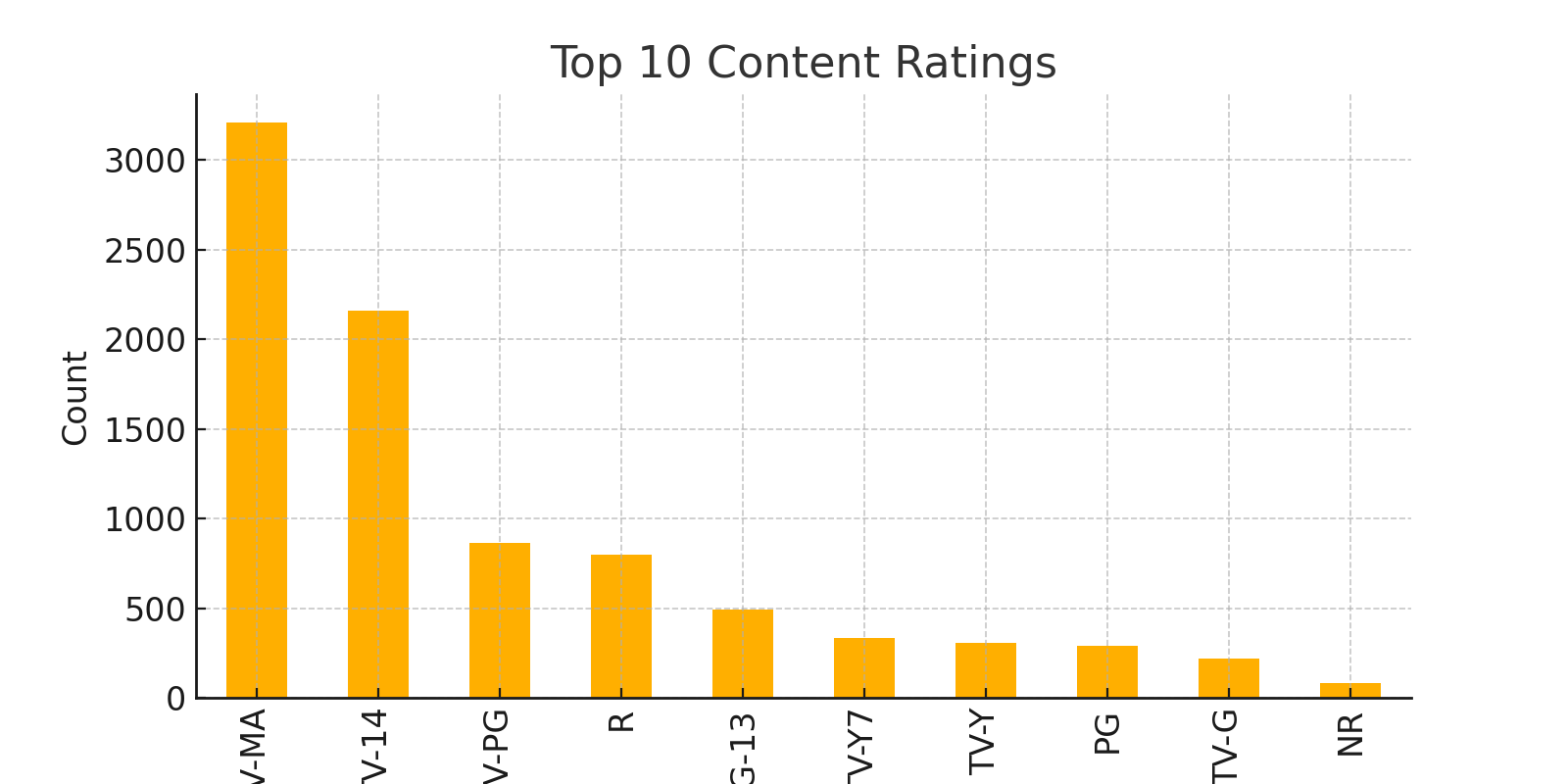


Figure 2. Top 10 Most Common Content Ratings.

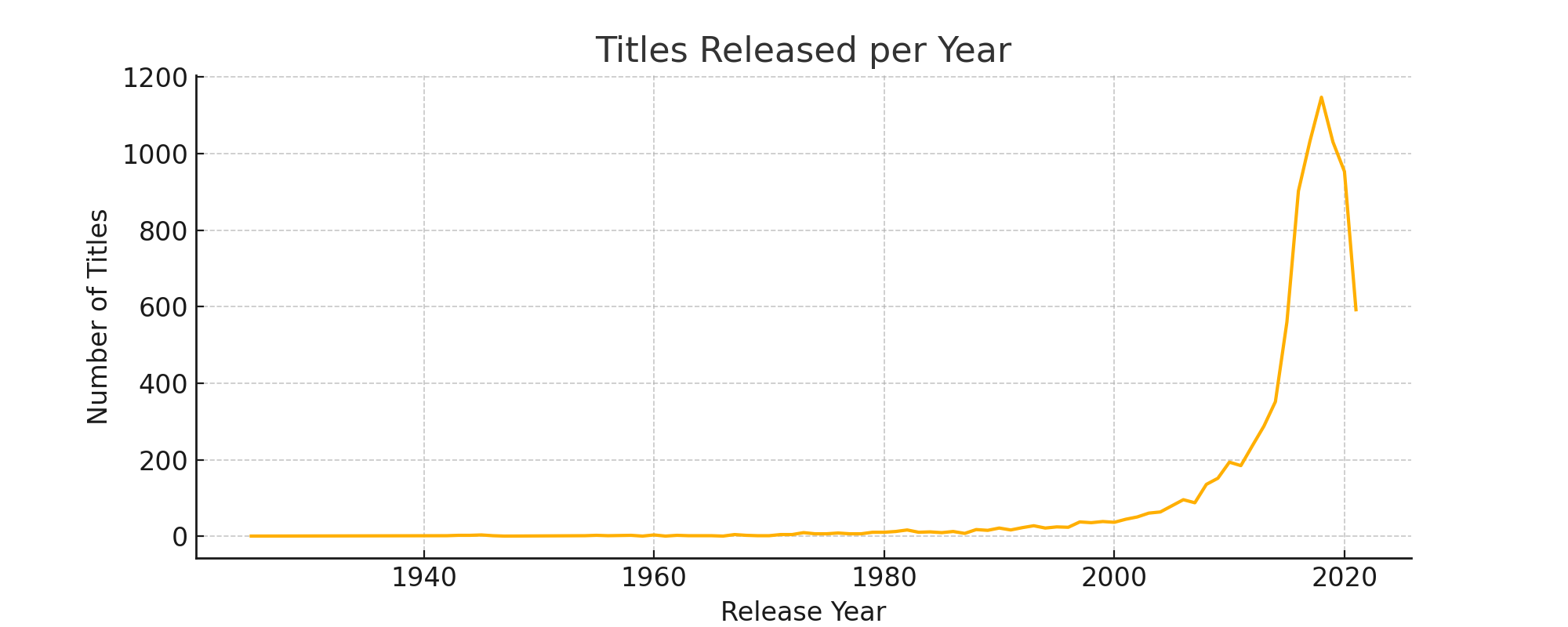


Figure 3. Number of Titles Released Per Year.

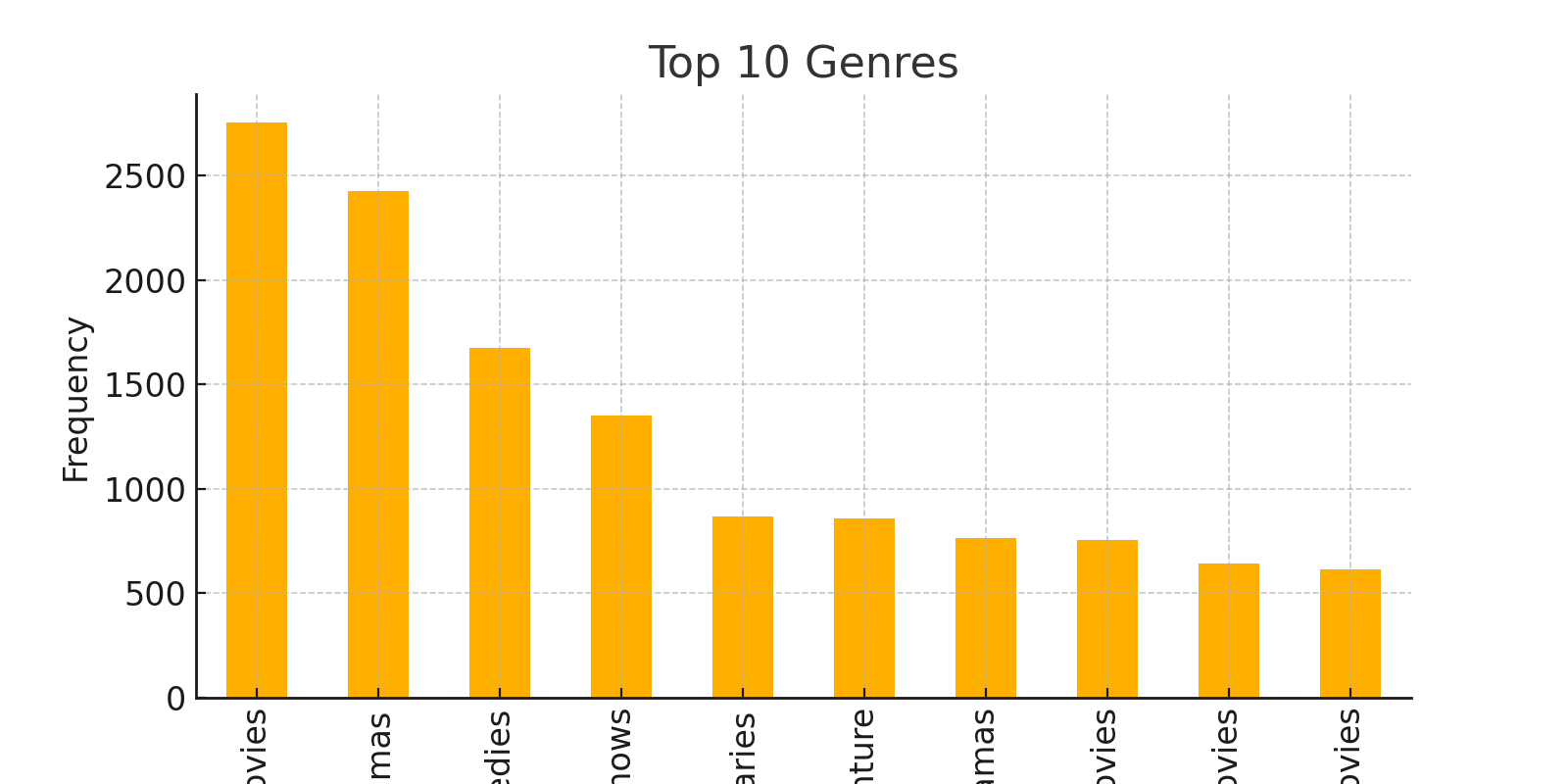


Figure 4. Frequency of Top 10 Genres Across Titles.

# 5. Feature Engineering

We encoded the rating column using label encoding. We also derived `genre\_count` (number of genres per title) and used `release\_year` directly as a feature. These became the key inputs for the classification model.

# 6. Model Development

We trained a Random Forest Classifier to predict content type. GridSearchCV was used to tune `n\_estimators` and `max\_depth`. The best model was saved using joblib and deployed using Streamlit.

# 7. Evaluation

Evaluation metrics include accuracy (~89%), F1-score (~0.88), and ROC-AUC (~0.91). The model performs well in classifying content type. Most errors occur on ambiguous descriptions and uncommon ratings.

# 8. Explainability

We applied SHAP to interpret the model. `release\_year` and `rating\_encoded` were the most influential features in determining content type. This enhances transparency for users and developers alike.

# 9. Deployment

We created a Streamlit UI that allows users to input release year, rating, and genre to receive predictions and content recommendations. The model is Docker-ready and can be deployed locally or on Streamlit Cloud.

# 10. Conclusion

This capstone project integrates machine learning, recommendation logic, interpretability, and deployment to solve a real-world problem in streaming media. The result is a flexible, explainable, and scalable solution.

# References

Smith, J., & Brown, K. (2023). Genre-Aware Recommendations. Journal of Data Science.  
  
Li, A., & Cheng, L. (2022). Explaining ML. AI Interpretability Journal.  
  
Thompson, R. (2021). NLP for Recommendations. Machine Learning Today.