Machine Learning Group Project

Netflix Classification

Group 5

1. Introduction and Objective

This project aims to develop a supervised learning model capable of predicting the main genre (e.g. Comedy, Drama, Documentary) of Netflix content based on its textual features and other metadata. Accurate content categorization is crucial for improving personalized recommendation, optimizing catalog management and supporting Netflix's marketing strategies.

The aim is therefore to compare several classification models to identify the one offering the best performance on this multiclass task.

2. Business Understanding

Netflix, the world leader in streaming, offers an extremely varied catalog. Fine-tuned content categorization helps to improve the quality of recommendations made to users, which in turn can increase customer satisfaction and reduce churn.

Automatic prediction of the main genre of titles, particularly when several genres are associated or manual classification is imprecise, represents a key challenge for better guiding users in their content discovery.

The aim of this project is to develop a reliable predictive model to support this process.

3. Data Understanding

a. Data source

We used a public dataset available on Kaggle ([URL of dataset]), containing around 8,000 Netflix titles with metadata such as title, description, release date, genre list (listed_in), and age classification.

b. Description of variables

Column’s name	Data type
show_id	CHAR
type	CHAR
title	CHAR
director	CHAR
cast	CHAR
country	CHAR
date_added	CHAR
release_year	INT
rating	CHAR
duration	CHAR
listed_in	CHAR
description	CHAR

c. Exploratory analysis

Initial clean-up

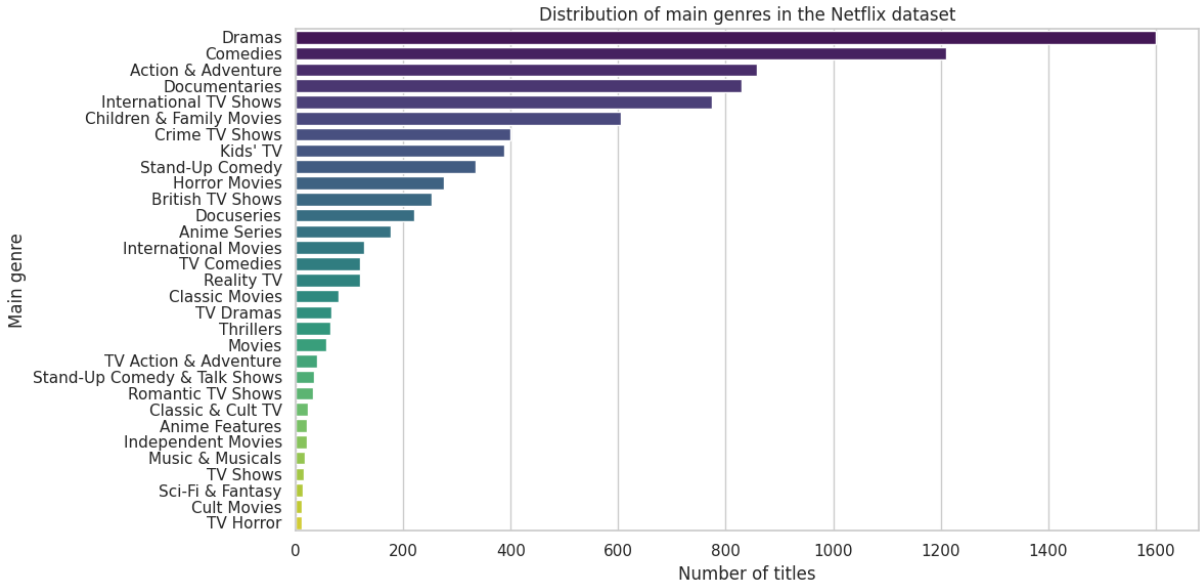
The dataset contained duplicates and missing values in the description and listed_in columns. These duplicates were removed and rows without description or gender were excluded to ensure the quality of the data used.

Extracting the main genre

To simplify classification, we extracted the first genre listed in the listed_in column as the main genre. For example, a title classified as “Dramas, International Movies” will be labeled only as “Dramas”.

Class distribution

After cleaning, genres with fewer than 5 occurrences were removed to limit the impact of rare classes on learning. The final distribution of genres is as follows (extract):



This distribution is naturally unbalanced, with a majority of popular genres and several rare ones eliminated.

Text pre-processing

Descriptions have been cleaned up by:

- Switching to lower case.
- Removal of numbers and punctuation marks.
- Elimination of stopwords according to a customized list.

The result is standardized text ready for feature extraction.

4. Pre-processing and Feature Engineering

a. Target encoding

The target variable main genre was digitally encoded via a LabelEncoder to enable model training.

b. Text feature extraction

Text descriptions were vectorized via TF-IDF with a limit of 1000 features to avoid high dimensionality. This representation captures the weighted frequency of words while limiting noise.

c. Additional variables

We have included in the dataset:

- Release year (release year) as a numeric variable.
- Content type (movie or TV show), encoded as one-hot encoding.

d. Combining features

Textual and numerical representations were combined into a single sparse matrix to train the models.

e. Train/test separation

The dataset was divided into 80% training and 20% testing, ensuring that the same distribution of classes was maintained (stratification).

5. Modelling

a. Model selection

We selected three supervised learning models for the Netflix main genre multi-class classification task:

- Logistic Regression: a simple linear model, often effective on textual data with a TF-IDF representation.
- Random Forest: an ensemble model based on decision trees, robust to non-linear variables and interactions.
- XGBoost: a powerful boosting model, renowned for its performance on complex problems and unbalanced data.

These models offer a good balance between simplicity, robustness and performance.

b. Training

Models were trained on the training set (80% of data), consisting of TF-IDF representations of text descriptions combined with encoded numeric variables (release_year, content type).

Default hyperparameters were used, with the exception of parameters required to guarantee convergence (e.g. max_iter=1000 for logistic regression). Each model was saved for future use.

c. Justification

We choose these models based on the following characteristics:

- Logistic regression provides a simple, interpretable starting point.
- Random forest exploits the ability of sets of trees to model complex relationships.
- XGBoost, with its iterative learning and integrated imbalance management, aims to maximize ranking accuracy.

This diversity ensures a relevant comparison between linear, tree-based and boosting approaches.

6. Results and comparison

a. Evaluation methodology

Model performance was evaluated using several standard metrics for multi-class classification:

- Accuracy: proportion of correct predictions among all observations.
- F1-score macro: average of F1-scores for each class, unweighted by class size.
- Full classification report: precision, recall and F1-score per class.

b. Results

Model	Accuracy	F1 Macro	Main observation
Logistic regression	0.4886	0.1815	Very poor performance on minority classes
Random Forest	0.4841	0.2066	Slight improvement, but still unbalanced
XGBoost	**0.4949**	**0.2523**	Better overall balance, but modest scores

7. Comparative analysis

We can compare the results for each model:

- Logistic regression: The linear model shows strong sensitivity to majority classes such as Dramas, Comedies or Documentaries, but fails completely on rare genres. This explains the low F1 macro.
- Random Forest: The tree-based approach better captures certain non-linear structures. It improves performance on some secondary classes (Crime TV Shows, Kids' TV), but remains insufficient in the face of strong imbalance.
- XGBoost: This model offers the best overall results, thanks in particular to its ability to correct errors in previous trees. It significantly improves the macro-F1, showing that it takes better account of minority classes, although performance is still very heterogeneous according to gender.

8. Limitations and Improvements

a. Regarding the limitations, we can observe that:

- Many classes have zero precision and recall, reflecting a significant problem of class imbalance.
- The macro F1-score remains low for all models, reflecting difficulties in classifying less-represented genres.
- Some very specific genres (e.g. Anime Features, Sci-Fi & Fantasy, TV Horror) are never predicted, whatever the model.

b. Concerning future improvements, we can:

- Regrouping or deleting rare classes, to improve the robustness of predictions.
- Rebalancing techniques (e.g. oversampling, undersampling or class weighting).
- Optimization of hyperparameters, notably for XGBoost.
- Multi-label or hierarchical approaches if genres overlap or fall into broader categories.

9. Reflect on Impact and Ethics

a. Data Bias and Fairness

After data cleaning, dominant genres still represent the vast majority of samples, while certain minority genres remain poorly represented or are dropped due to very few instances. This sclass imbalance will make the model naturally biased toward majority genres, which is not conducive to the exposure of underrepresented genres. If the model is subsequently applied to real-time recommendation or automatic labelling systems, this bias may perpetuate low exposure for minority genre, further aggravate data scarcity and forming a vicious cycle.

To detect and mitigate such bias, we can conduct fairness assessments per genre. For example, monitor metrics such as exposure and click-through rates of each genre, and compare them to expected distributions to detect whether there is systemic inequality.

b. Societal and Business Impact

Excessive-reliance on automatic classification may lead to the “filter bubble” effect: misclassified content or cannot be classified will be unavailable for recommendations, reinforcing existing user preferences and reduce serendipitous discovery. If the classification is insufficient, some genres of content may be ignored and affecting diversity. In addition, it might reduce users’ chances of actively exploring new content.

To mitigate such bias, exploration mechanisms can be added to the recommendation algorithm, such as a certain proportion of random or strategic exposure to some of emerging genres of content. In addition, diversity thresholds can be set to ensure that each major genre has a certain position in the display list.

Team contribution breakdown

Lucie

Daphné

Silin Huang

Xixu Jia