PREDICTING THE MAIN GENRE ON

NETFLIX

Play

JOUCLARD Lucie, MATRAND DE LA BOURDONNAYE Daphné, HUANG Silin, JIA Xixu

INTRODUCTION & OBJECTIVE

NETFLIX

CONTEXT

- Personalising content is key to building user loyalty.
- Netflix, with its constantly growing catalogue, needs to better classify its titles by genre.

ISSUES

- Manual classification is often imprecise or incomplete.
- Under-exploited metadata (title, description, type, year).

OBJECTIVE

1. Develop a machine learning model to automatically predict the main genre based on textual metadata.
2. Compare several models (logistic regression, random forests, XGBoost) with TF-IDF.

ADDED VALUE

1. Automated and reliable content indexing.
2. Improved personalised recommendations.
3. Support for marketing and editorial decisions.

BUSINESS UNDERSTANDING

NETFLIX

STRATEGIC CONTEXT

Netflix, the world leader in streaming, offers thousands of different types of content.
Precise classification by genre is key to :

1. Improve personalised recommendations
2. Optimise catalogue navigation
3. Guide marketing, purchasing and production decisions

PROBLEM

- Categorisation based on declarative and sometimes subjective information.
- Multiple genres per title or incorrect/absent labelling.
- Negative impact on user experience and performance analysis.

NEED

An automatic system to predict the main genre based on accessible data (description, year, type).

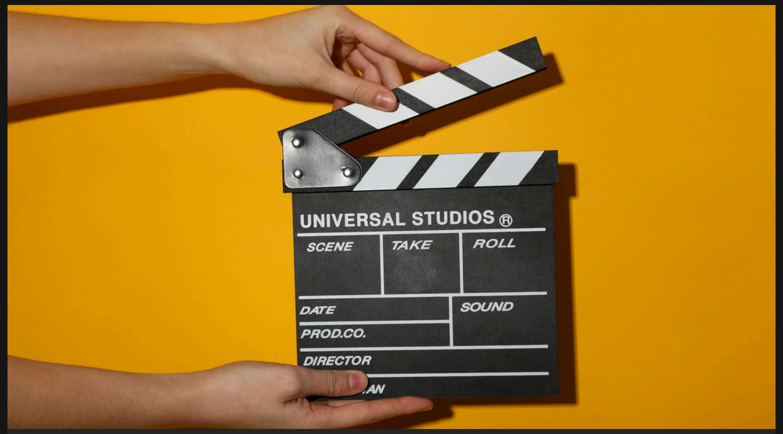
Objective: make classification more reliable, speed up the integration of new content, increase subscriber satisfaction.

ML

- Multiclass supervised classification task: predict a single main genre per title.
- Training of a robust model on a public, reproducible and relevant dataset.

DATA UNDERSTANDING

NETFLIX



Data Source

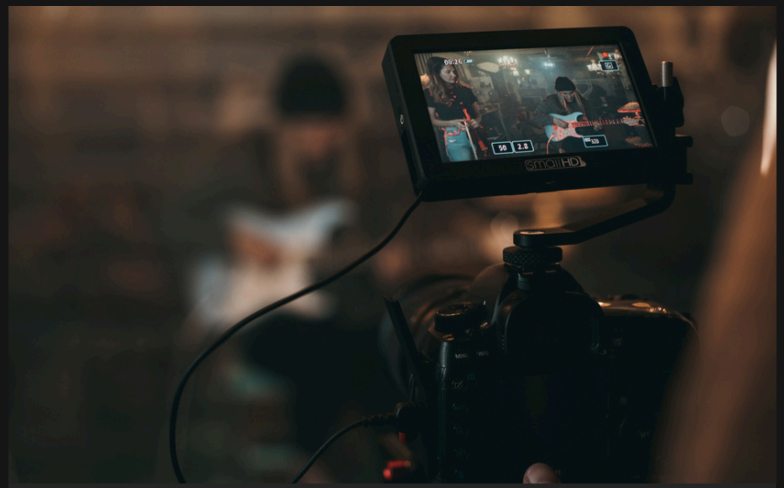
- 📍 **Origin**: Public dataset from Kaggle (≈8,000 Netflix titles)
- 🎬 **Content**: Metadata for shows and movies:
- Title, Description
- Release Year, Type (Movie or TV Show)
- Genre(s), Age Rating, Country, Director, etc.

KEY VARIABLES

Column Name	Description	Type
show_id	Unique ID	text
type	Movie or TV Show	categorical
title	Title of the content	text
listed_in	Genre list (e.g. Dramas, Comedy...)	categorical (multi-label)
description	Text description of the content	text
release_year	Year of release	numerical

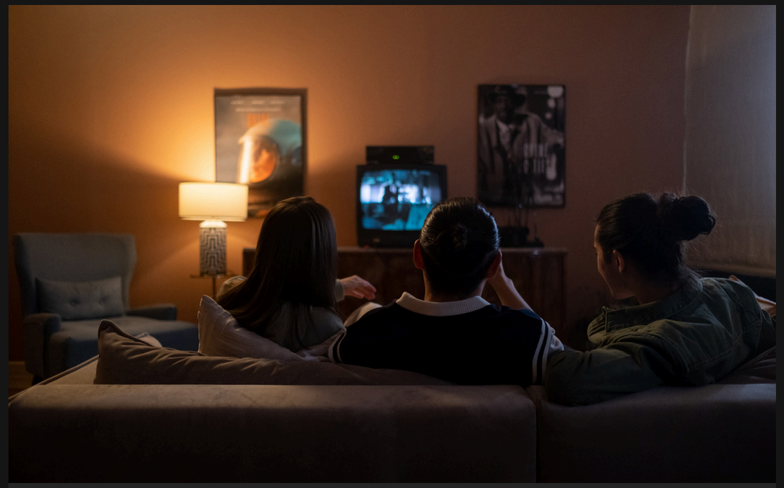
Other columns such as rating, duration, country, cast, director, and date_added are mostly informative and were considered less relevant for our predictive modeling.

DATA PREPROCESSING



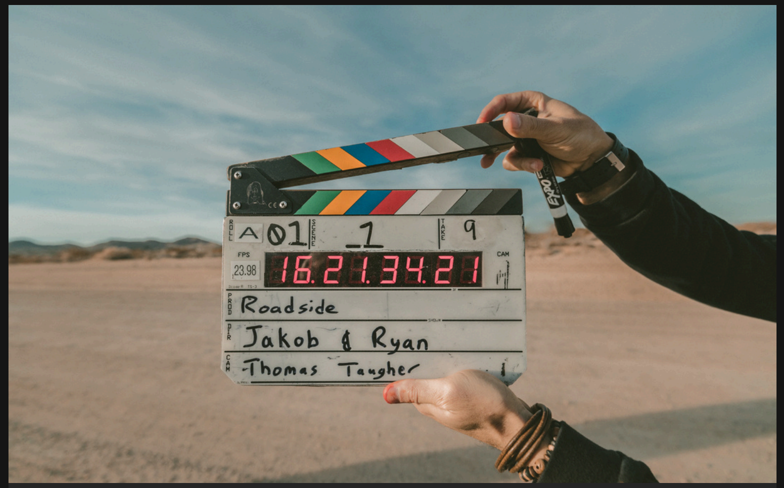
1. Cleaning and Filtering

- Removed duplicates and rows missing key columns (description, listed_in)
- Excluded genres with fewer than 5 occurrences to reduce class imbalance



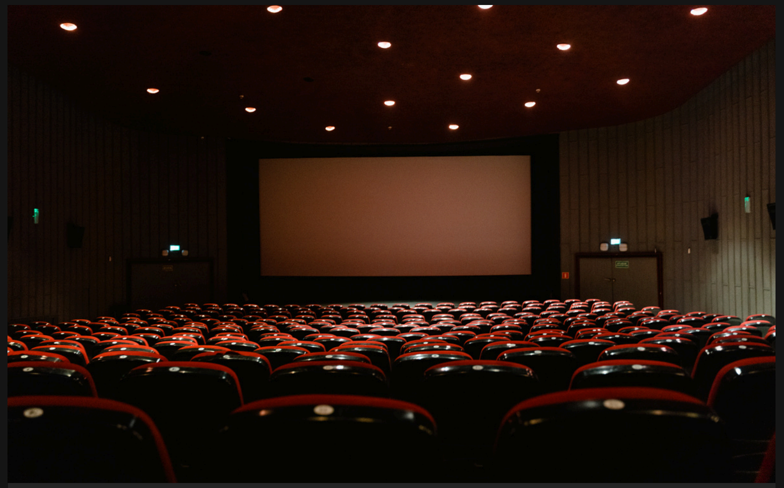
2. Genre Simplification

- Extracted the main genre by taking only the first genre listed in listed_in
- Example: "Dramas, International Movies" → "Dramas"



3. Text Preprocessing

- Lowercased descriptions
- Removed numbers, punctuation, and stopwords (custom list)
- Result: clean, standardized text for vectorization



4. Feature Engineering

- description → TF-IDF vector (top 1000 terms)
- main_genre → label encoded (target variable)
- type (Movie/TV Show) → one-hot encoded
- release_year kept as numeric
- Combined all features into a single sparse matrix
- Applied 80/20 train-test split, stratified by genre

MODELISATION

NETFLIX

Chosen Models

- **Logistic Regression:** A simple, interpretable linear model, effective for TF-IDF text features.
- **Random Forest:** A tree-based ensemble that captures complex patterns and handles mixed data types.
- **XGBoost:** A powerful boosting algorithm known for strong performance, especially on imbalanced data.

Training Pipeline

- **Inputs:** TF-IDF vectors (from descriptions) + numerical/categorical features (release_year, type)
- **Split:** 80% training / 20% testing (stratified to preserve genre balance)
- **Preprocessing:** Models trained on cleaned, vectorized, and combined feature matrix
- **Hyperparameters:** Defaults used; max_iter increased for logistic regression to ensure convergence

EVALUATION & METRICS

NETFLIX

EVALUATION

- multi-class supervised classification
- strong gender imbalance
- essential to use appropriate metrics
- evaluate overall performance and class by class

METRICS USED

Overall accuracy
- percentage of correct predictions
- limited in the case of unbalanced classes

F1-score macro
- unweighted average of F1-scores
- ability to correctly predict all classes, including minority classes

METHOD

1. Evaluation carried out on the test set (20%), never seen bu the models during training
2. No artificial balancing applied before this firstevaluation, to obtain a raw and objective performance

WHY THESE CHOICES?

F1-score reflects the model's ability to handle all the classes well
even poorly represented
key issue in our business problem

RESULTS

NETFLIX

Model	Accuracy	F1-score macro	Main observation
Logistic regression	0.4886	0.1815	High sensitivity to majority classes, total failure on rare genre
Random Forest	0.4841	0.2066	Slight improvement in F1, but persistent difficulty on minority classes
XGBoost	0.4949	0.2523	Better overall performance, relative balance between classes

Logistic regression: basic model, fast but too limited to manage class imbalances
Random Forest: better at capturing certain non-linear structures, but still penalised by the absence of an adjustment mechanism for rare classes
XGBoost: performs well on unbalanced data thanks to iterative optimisation; it obtains the best results

LIMITATIONS & FUTURS IMPROVEMENTS

NETFLIX

LIMITATION

- Severe **class imbalance**: many genres show zero precision/recall
- **Low macro F1** across all models: poor performance on under-represented genres
- Certain niche genres (e.g., Anime Features, Sci-Fi & Fantasy, TV Horror) **never get predicted**

IMPROVEMENTS

- Merge or remove very classes to strengthen robustness for main classes
- Resampling strategies (oversampling, undersampling, weighted loss)
- Hyperparameter optimization (e.g., grid/Bayesian search for XGBoost)
- Consider multi-label or hierarchical classification if genres overlap or form broader categories

ETHICAL ASPECT

NETFLIX

THE ETHICS OF AUTOMATED CLASSIFICATION

- Based on a model trained on descriptions ==> often written **subjectively**
- **Risk of bias:** systematically associated with stereotyped genres
- System: reinforce **normative representations** of content, without editorial contextualisation

TRANSPARENCY & EXPLICABILITY

- models offer a **degree of interpretability** (logistic regression or random forest)
- XGBoost, more complex to explain
- if a solution is integrated ==> essential to guarantee **traceability of algorithmic decisions**

RESPECT FOR PRIVACY

- public and contains no sensitive info or user data
- no direct risk to the privacy or security of individuals

LIMITS OF USE IN PRODUCTION

- a decision-making tool, not a complete substitute for human judgement
- **real context:** editorial validation of the predicted genres would be essential

CONCLUSION

NETFLIX

PROJECT SUMMARY

Implementation of a complete processing chain:

- textual pre-processing (TF-IDF)
- integration of contextual variables
- training of three supervised models
  - Logistic regression
  - Random forest
  - XGBoost
- rigorous evaluation on a stratified test set

Result:

- XGBoost performed best
- performance limited

BUSINESS ISSUES

- Such a model can automate the classification of new or poorly labelled content
- A decision-making tool for editorial and product teams
- Can enhance recommendation systems and improve catalogue navigation

FUTURE DEVELOPMENTS

- Rebalancing classes (SMOTE, class weighting)
- Tuning hyperparameters
- Multi-label approach to reflect the reality of the Netflix catalogue
- Moving to advanced models (BERT, LLMs) for better semantic understanding

NETFLIX

THANKS FOR YOUR ATTENTION

End