Data Preparation, Cleaning & Feature Generation

Farida Bey
900212071

&

Youssef Ghaleb
900211976

DATASET OVERVIEW AND RATIONALE

Name: TMDb Movies Dataset 2023 (930k Movies)
Source: Kaggle - TMDb Movies Dataset

This dataset provides detailed information about movies, including features such as title, runtime, genres, overview, and keywords. It is highly suitable for genre classification tasks, aligning with the goals of our project. The dataset includes approximately *930,000 movie entries* and contains *24 features*, but we have selected a subset that best fits our needs for predictive analysis.

Features:
id: Unique identifier for each movie.
title: Movie title.
vote_average: Average user rating (numeric).
vote_count: Number of votes (numeric).
status: Movie status (e.g., Released, Post Production).
release_date: Date of release (date format).
revenue: Revenue generated (numeric).
runtime: Duration of the movie in minutes (numeric).
adult: Whether the movie is for adults (boolean).
backdrop_path: URL path for the backdrop image.
budget: Budget of the movie (numeric).
homepage: Website URL.
imdb_id: IMDB identifier (string).
original_language: The original language of the movie (nominal).
overview: A brief synopsis of the movie (text).
popularity: Popularity score (numeric).
poster_path: URL path for the poster.
tagline: Movie tagline (text).
genres: Movie genres (multi-label, nominal).
production_companies: Companies involved in production (multi-label, nominal).
production_countries: Countries where the movie was produced (multi-label, nominal).
spoken_languages: Languages spoken in the movie (multi-label, nominal).
keywords: Keywords related to the movie (text).

We found this dataset particularly relevant due to its large number of instances, allowing for better generalization of machine learning models. The diversity of features, such as *overview* and *genres*, offers rich semantic content for the model to understand the relationship between movie descriptions and genre classification. Additionally, *keywords* provide further context about the movie, enhancing the depth of classification.

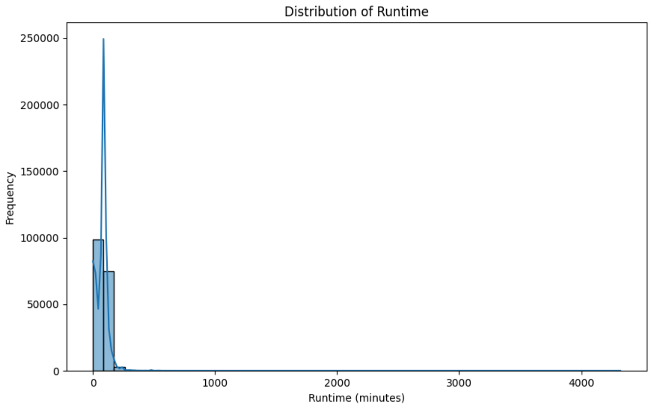
FEATURES ANALYSIS

1. Statistical Distribution:

The runtime feature represents the duration of each movie in minutes. The following summary statistics provide insights into its distribution (after cleaning):

Count: 177,474
Mean: 75.32 minutes
Standard Deviation (Std): 49.08 minutes
Minimum (Min): 1 minute
25th Percentile: 44 minutes
Median (50th Percentile): 84 minutes
75th Percentile: 99 minutes
Maximum (Max): 4,320 minutes

The histogram below demonstrates the distribution of the runtime values in the dataset (after cleaning):



2. Missing Values:

Using a Python script to analyze the percentage of missing values in our dataset, we categorized the features into two groups: non-numeric (nominal, string, boolean, date, etc.) and numeric features. The results of this analysis are presented in the following tables:

2.1. Non-Numeric Features:

Feature	*Empty Cells (%)*
title	*0.001171*
status	*0.000000*
release_date	*15.784377*
adult	*0.000000*
backdrop_path	*73.228583*
homepage	*89.361006*
imdb_id	*46.322776*
original_language	*0.000000*
original_title	*0.001171*
overview	*20.008845*
poster_path	*31.014530*
tagline	*85.939097*
genres	*39.663250*
production_companies	*54.578233*
production_countries	*43.816649*
spoken_languages	*42.200362*
keywords	*72.608244*

2.2. Numeric Features:

Feature	*Empty Cells (%)*	*Percentage of Zeros (%)*
id	*0.0*	*0.000000*
vote_average	*0.0*	*68.523841*
vote_count	*0.0*	*68.502944*
revenue	*0.0*	*98.150242*
runtime	*0.0*	*27.179542*
budget	*0.0*	*94.909021*
popularity	*0.0*	*10.828239*

3. Unique Values:

The following table presents the frequency (how many times that genre appears) of each genre in the dataset:

Genre	Frequency
Action	16,492
Adventure	9,646
Animation	14,083
Comedy	41,431
Crime	11,553
Documentary	39,933
Drama	60,208
Family	9,025
Fantasy	8,154
History	5,932
Horror	16,658
Music	12,768
Mystery	6,746
Romance	18,261
Science Fiction	8,349
TV Movie	7,242
Thriller	15,394
War	3,816
Western	4,017

The distribution of values in the adult column is as follows:

Adult	Count
False	171,641
True	5,833

4. Correlation with Label:

While a formal correlation analysis was not conducted, insights from industry experts (assistant directors, and actors) indicate that certain genres, such as *Comedy* and *Drama*, are often associated with longer runtimes, whereas genres like *Documentary* and *Animation* may have varied runtime distributions. In addition some genres are more likely to be labeled as *adult* and others are more likely to be labeled as *not adult*, indicating a correlation with audience demographics. This understanding implies that runtime and the adult feature could be relevant for genre classification, guiding the model to enhance its predictive accuracy based on industry knowledge.

5. Semantic Importance:

Overview:

The overview feature encapsulates great information about the movie from key themes, character arcs, and major events, which are all often indicative of a film's genre. By analyzing the language used in the overview, the model can identify keywords or phrases that are strongly associated with specific genres (e.g., "detective," "love," "sci-fi," etc.).

Natural Language Processing (NLP) techniques, such as TF-IDF (Term Frequency-Inverse Document Frequency), can be applied to capture the semantic meaning of the overview and reveal how certain plots align with particular genres.

Keywords:

This feature offers a list of important terms related to the movie's themes, concepts, or notable elements. This feature can act as a supplementary descriptor for the movie, further enhancing genre classification. They can reflect genre-specific tropes or motifs (e.g., "vampire," "space travel," "superhero," etc.), which may not always be explicitly stated in the overview but are essential for classification. When used in conjunction with the overview, the keywords can help the model differentiate between similar genres that may share overlapping plot elements but have distinct thematic differences.

Integrating these textual features allows the model to develop a more nuanced understanding of the relationships between the content of the movies and their genres. The ability to interpret and analyze these features effectively can lead to improved accuracy in genre classification.

DATA CLEANING AND PREPROCESSING

1. Data Cleaning

Dropping Unimportant Features:
First, we removed features that were deemed irrelevant to our task (e.g., backdrop_path, poster_path, and homepage, e.t.c.).

Handling Missing Values:
Features with high percentages of missing data or zeros (in case of numerical features) were dropped. Additionally, duplicate rows were also removed to avoid redundancy.

Although keywords had an empty cell percentage exceeding 72% , it supports overview by providing more specific words to describe the movie, thus we didn’t remove it. Also we removed all the rows containing the empty cells.

2. Data Preprocessing

One-Hot Encoding:
The genres feature, being multi-labeled, was converted into separate binary columns for each genre using one-hot encoding.

Scaling:
We scaled the runtime feature using two methods:
Min-Max Scaling and ***Standard Scaling***. We will evaluate both scaling methods in later stages to determine which performs better.

Binary Encoding:
The adult feature was encoded into binary values, where True became 1 and False became 0.

3. Text Processing

For the text features (overview and keywords):

Stop Word and Punctuation Removal:
We removed common stop words and punctuation marks to focus on meaningful content.

Stemming:
Stemming was used to reduce words to their root form.

However, we found that this process distorted some words, and we are researching alternative methods that do not distort words as much.

TF-IDF Transformation:
Finally, we applied TF-IDF to convert the processed text into numerical vectors, emphasizing important terms in the dataset.

FINAL DATASET OVERVIEW

The final dataset after thorough cleaning and preprocessing includes *177,475 movie entries* and contains *6 features*.

The selected features:

Title: The title of the movie.

Adult_encoded: Indicates whether the movie is intended for an adult audience, encoded (with 1 for True and 0 for False).

Runtime_minmax_scaled: Min-Max scaled values of the runtime.

Runtime_standard_scaled: Standard scaled values of the runtime.

Overview_cleaned: Cleaned movie descriptions, after preprocessing (stop word removal, punctuation removal, stemming).

Keywords_cleaned: Cleaned keywords associated with the movie, also preprocessed similarly.

One-Hot Encoded Genres: Each genre (e.g., Action, Adventure, Animation, etc.) is represented by a separate binary column (1 for presence and 0 for absence). ***[label]***

The final features include *title, runtime, adult, one-hot encoded genres, cleaned overview, and cleaned keywords.* Title is kept for reference, while runtime is important for genre differentiation and is scaled using both min-max and standard scaling. The adult feature is encoded as a binary variable to reflect content type. One-hot encoding for genres allows us to capture multi-label information effectively. Lastly, the overview and keywords were preprocessed and transformed using TF-IDF to enhance model performance. This selection balances numerical, categorical, and textual data for genre prediction.