Predicting The Most Profitable Movie Genre

Prepared by: Khalid Bawazer

Introduction

Film industry considered one of the most profitable and risky businesses in the entertainment and media fields, has a massive audience and its profits are estimated in millions/billons. It's extremely important for the filmmakers and producers to know and understand their audience preferences regarding movie genres.

Objectives

Examine and predict which genre of movies will win the audience reception and acceptance within upcoming years based on several factors such as profit reception, Ratings, Views, and others. This model will benefit the filmmakers, and producers in the film industry to know which kind of movies the audience prefers to watch to grant the financial return.

Design

The dataset is obtained from IMDb website (an acronym for Internet Movie Database), is an online database of information related to films, television programs, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

Data

The dataset consists of two datasets combined into one, include 77 features and 85,855 observations. There are 17 categorical features such as: Movie ID, Movie Name, Genre, Language, Director, Budget, Gross, etc. And 54 numerical such as: Duration, Average Votes, Total Votes, Number of reviews from users and critics, etc.

Algorithms

Feature Engineering:

- Merging two datasets into one (Movie details and ratings).
- Clean the dataset which contain null values.
- Omit non-dollars movie profit records.
- Converting categorical features into numerical.

Models:

The dataset is considered as unlabeled data which need models to determine the hidden pattern. The chosen models are:

- K-Means clustering.
- Logistic regression.

Tools

Data processing: Numpy and Pandas.

Modeling: Scikit-learn.

Visualizations: Matplotlib and Seaborn.

Communication