IMDB Dataframe Explanation

November 17, 2021

We use the dataset IMDB movies which contains data from 85,856 films in the following features taken from IMDB.com: IMDB title ID (a unique ID number assigned to each film by IMDB), title name, original title name, year of release, date of release, movie genre, duration in minutes, country of origin, movie language, director name, writer name, production company, actor names, plot description, average vote (the IMDB rating), number of votes, budget, US gross income, worldwide gross income, metascore rating, number of reviews from users, and number of reviews from critics.

Our goal is to predict what specific features a film should have in order to maximize (1) critic reception and (2) popular reception. The way we will determine this is by running regressions on the data against the dependent variables **metascore** (to determine critical reception) and **worldwide gross income** (to determine popular reception). The independent variables we will be analysing are: **movie genre, duration, director, writer, production company, actors, average vote, number of votes,** and **budget**. The IMDB title ID will be used as the indexer for our dataframe, and average vote and number of votes will only be used to predict the metascore for films that don't initially have a metascore (which will be discussed more in depth later).

Because our project is so heavily founded on the independent variables, we remove all films that have any data missing from the independent variables. This leaves us with a dataset of 22,912 films. Duration, average vote, number of votes, budget, worldwide gross income, and metascore are all numerical values and are ready to go, but genre, director, writer, production company, and actors are categorical features that need to be one-hot encoded.

We one-hot encode the genre feature so that a new column is created for each genre type, and a film of a certain genre will have a value of 1 in the column representing its genre and a 0 in every other genre column. Keep in mind that most films are categorized under more than one genre, so most films have a 1 in more than one genre column.

We proceed in like manner with the other categorical features, but we also implement a method to sift out individual cases that appear less than a specified threshold. A person is considered a serial killer if they have killed 3 or more people. Similarly, we define a serial director, writer, or actor as a person who has directed, written, or acted in a film at least 3 times. Since we are looking for consistent information to predict a successful movie, we only consider serial

directors, writers, and actors in our dataset. So, we one-hot encode the director, writer, and actors features but only keep the columns that sum to at least 3. We one-hot encode the production company feature in a similar manner, but we set the threshold to 10 or more since a company that has produced less than 10 films is probably out of business.

*** Note to Group: This needs to be changed if we decide to use an actors threshold of 5 ***

We one-hot encode each of these features separately in their own dataset and then join them together in our original dataset after removing the initial genre, director, writer, production company, and actors columns. We are left with 24 unique genres, 2,172 unique serial directors, 3,095 unique serial writers, 161 unique production companies that have produced at least 10 films, and 24,945 unique serial actors. This generates a dataset (after being joined together with duration, average vote, number of votes, budget, worldwide gross income, and metascore) of 22,912 films with 30,403 columns.

*** Exchange Rate and Inflation info should go here ***

After dropping the films that have foreign currencies that cannot be converted to dollars, we are left with a dataset of 21,028 films with 30,403 columns.

*** If we decide to use an actors threshold of 5, this dataset changes to 21,028 by 17,319 after dropping votes and avg_vote ***

IMDB assigns a film's metascore by taking a weighted average of reviews made by professional film critics from around the world. The scale is from 0 to 100 where a higher score indicates that the film was better received by critics. Because of this, we use the metascore, and not the average vote, to determine and predict the critical reception of films in this project. However, not every film has a metascore on IMDB, and if we dropped every film in our dataset that didn't have a metascore, our dataset would drop from 22,912 films to only 7,685 films. On the other hand, every film on IMDB does have an average vote and number of votes, so rather than simply drop every film that doesn't have a metascore, we decide to predict metascores for those films using their average votes and number of votes.

*** Daniel's info about Metascore should go here ***

All Datasets should have the same column size of 30,403 (but this might include votes and avg_vote?) OR 17,319 after dropping votes and avg_vote.

Dataset Row Sizes:

Original Dataset = 22,912
Exchange Rate Dataset = 21,028
Filled_Metascores Dataset = 21,028
Original_Metascores Dataset = 7,551
Gross_Income Dataset = 12,196

Citations

kaggle (2020), IMDB movies (dataset). Retreived from www.kaggle.com/stefanoleone992/imdb-extensive-dataset