Movie Revenue Analytics
Realtime and Big Data Analytics - Data Ingestion

Abhishek Verma [av2783]

Overview

As the movie industry grows, the probable profit made by a movie becomes of utmost importance for the stakeholders. Among the movies produced between 2010 to 2020 in the United States, less than 40% of the movies had revenues higher than the production budget. This highlights the importance of knowing the factors contributing to the profitability of a movie to make the right investment decisions and presents us with a bunch of questions like - What can we say about the success of a movie before it is released? Does the release day of the week have anything to do with the popularity and profit of the movie? Does the running time of a movie have an effect on its popularity and profitability of a movie? Can a probable revenue be predicted based on the pre-release data of a movie?

Objective

In this project, we'll be focussing on the following questions:

- Is there a relationship between the release day of the week and the movie's profitability and popularity? If yes, which weekdays as release days turn out to be most lucky for movies in terms of popularity and profit?
- How has the time duration been affecting high Profits, High Voting Average and High Popularity over the years from 2007 to 2017?
- Does the popularity of a director in a particular calendar year affect the chances of them being nominated for an award at the Cannes Film Festival?

In addition to this, if time permits, we will be modeling the movie revenue using machine learning techniques like regression, random forests, etc. using the pre-release data of the movie.

Dataset

TMDB Dataset

This dataset was generated from The Movie Database API. This dataset provides information about different movies, namely, title, cast, crew, budget, revenue, etc. It consists of 2 files- credits.csv and movies.csv, with 4803 movie information records each.

- *credits.csv:* This contains fields like the movie_id(tmdb_id), title, cast and crew of a particular movie.

- *movies.csv:* This has fields including but not limited to movie_id(tmdb_id), the title, budget for the movie production, genre of the movie, the original language, etc.

Data Ingestion

1. Profiling:
    a. *movies.csv:*
        - Total rows: 4806
        - The file has 6 rows which had at least one column missing.
        - 28 rows with Genre missing
        - 1 row with the release date missing
        - 2 rows with runtime missing
        - 86 rows with spoken languages missing
        - 3 duplicate movie titles were present in the file. Upon further examination, it was found that these titles were different movies with different release dates. So, this was ignored.

    b. *credits.csv:*
        - Total rows: 4813
        - 14 rows with at least one column missing.
        - 3 duplicate movie titles were present in the file. Upon further examination, it was found that these titles were different movies with different release dates. So, this was ignored.
        - 42 duplicate cast entries. But these belong to different movies.
        - 27 duplicate crew entries. However, all these entries belong to different movies
2. Cleaning:
    a. movies.csv:
        - Removed the bad rows which had column information missing.
        - Number of rows after cleanup: 4704

    b. *credits.csv:*
        - Removed the bad rows which had column information missing.
        - Number of rows after cleanup: 4799
        - Only the top three cast (IDs) per movie were kept. This reason is that whether a movie will be successful or not depends on the top few casts (the leading actors and the ones with the most screen time). The number three was purely intuitive.
        - From the crew column, the set of directors (IDs) for each movie was extracted. Other information was ignored. This will be used in the hypothesis testing.

3. Extraction:
    a. *movies.csv:*
        - movie_cleaned_data.tsv: A new table was created which contains just a subset of columns, namely, budget, genreIds, movieId, language, popularity, releaseDate, revenue, runTime, spokenLanguages, title, voteAvg. This table will be used in the hypothesis testing and will be used along with other tables for the revenue prediction model.
        - genre_ID-to-Name.tsv: Another table, genreId to genreName mapping was created.
    b. *credits.csv:*
        - cast_ID-to-movie_ID.tsv: A mapping from actors to the movies they have acted in. This will be needed to compute cast_impression_index (as mentioned in the proposal).
        - movie_ID-to-cast_ID.tsv: A mapping from the movie to the top three casts of the movie. Again, this will be used to compute cast_impression_index.
        - cast_ID-to-Name.tsv: Mapping from cast id to cast name
        - director_ID-to-Name.tsv: Mapping from director id to director name

Note: For the instructions to run the code, refer to the README file.