Movie Revenue Analytics
Realtime and Big Data Analytics

Ishita Jaisia [ij2056]
FNU shivanshi [ss14396]
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.

Datasets

We will be using two public datasets - both of which are downloadable from the Kaggle website. The data is static in nature.

1. 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.

2. ThemovieDB
   The primary source of this dataset is www.themoviedb.org. This dataset contains information on various directors, awards for which they were nominated, the number of total awards won, etc. This dataset consists of 8 files with varying numbers of records. The files we will be using from this dataset are 900_acclaimed_directors_awards.csv and 220k_awards_by_directors.csv.
   - *900_acclaimed_directors_awards.csv:* This presents us with details about different directors and the number of awards they received.
   - *220k_awards_by_directors.csv:* This consists of details about the award nominations that different directors received in various award ceremonies.

3. The Movies Dataset
   This dataset is an ensemble of data collected from TMDB and GroupLens. This is a database of 45000 movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. The files in this dataset that are useful to us are
   - *movies_metadata.csv:* This has details about the genre, budget, language, popularity, etc.
   - *links.csv:* This is a file to link the movie IDs in this dataset to their respective tmdb_ids.
   - *credits.csv:* This contains the crew and cast information of the movies.

Note: We plan to combine the relevant data from the TMDB dataset and The Movies Dataset for a better understanding of the problem statement at hand which will lead to better analysis.

Software Architecture/Tools

1. Java
2. Hive / HBase
3. HDFS

4. Hadoop

Workflow

1. **Data ETL and Profiling:** There are a total of 5 ETL Jobs to get the required data for our analysis.

Team Member	Dataset Assigned	Tasks	Usecase
Ishita Jaisia	ThemovieDB	ETL Task 1	→ Hypothesis testing
Shivanshi	The Movies Dataset	ETL Task 2	→ Hypothesis testing → Revenue prediction model
Abhishek Verma	TMDB Dataset	ETL Task 3	→ Hypothesis testing → Revenue prediction model

ETL Task 1

I. From the file *900_acclaimed_directors_awards.csv,* we extract the columns: director_name, tmdbId, and the total_number_of_awards. Badly formatted rows will be filtered out.
II. From the file *220k_awards_by_directors.csv*, we extract the data pertaining to the Cannes Film Festival of each director and a list of years when this particular director was nominated for an award. Again, any invalid rows will be removed.
III. We combine the data from the above two subtasks and put it into a single table with columns: TmdbId, Director_name, Total_number_of_awards and Nomination_years.

ETL Task 2

I. We extract the data from *movies_metadata.csv*. The only columns useful for our analysis are Movie_id, Title, Genre, Budget, Revenue, Language, Year, Popularity, Spoken_Languages, Vote_Average. Rows with missing values will be filtered out.
II. Map the movie_id in the above table to the corresponding tmdb_id using the file *links.csv*.
III. We create a mapping of a cast to a list of movies in which they are featured using *credits.csv*.
IV. We create a mapping of a movie to the cast list in the movie using *credits.csv*.

ETL Task 3

I. We extract the data from *movies.csv*. The only columns useful for our analysis are Movie_id, Title, Genre, Budget, Revenue, Language, Year, Popularity, Spoken_Languages, Vote_Average. Rows with missing values will be filtered out.
II. We create a mapping of a cast to a list of movies in which they are featured using *credits.csv*.
III. We create a mapping of a movie to the cast list in the movie using *credits.csv*.

Note: ETL tasks 2 ad 3 may seem very similar but they are not. This is because these tasks have to be performed on two different datasets, both of which have different columns. This will require separate map-reduce tasks for each of these.

2. **Hive/HBase processing:** These technologies will be used for the following tasks:

2.1. The data from all the ETL tasks will be aggregated in a single table for analysis and to be used in the modeling of revenue.
2.2. For the Revenue Prediction Model, data from ETL tasks 2 and 3 will be used.
2.3. As we can notice that we will get few duplicate rows from ETL tasks 2 and 3 as some movies will be present in both the datasets. We can do the union of the data from ETL tasks 2 and 3 before doing Revenue Prediction to avoid any redundant calculations.
2.4. We calculate the **cast_impression_index** of the whole cast which suggests the box office appeal of the cast for each movie. This is done using a weighted average of historical revenue of all the movies of the cast members individually in the list. We will use Hive/HBase to store the movie and cast information because a lot of highly complex join operations are required from both the TMDB tables and these technologies are known for their high scalability when it comes to table-like querying. This index will be used to train the Revenue Prediction model.

3. **Revenue Prediction model**

Once the Hive/HBase processing is complete, we will use the processed data and machine learning techniques based on regression and random forest to predict the revenue a movie will generate based on its cast impression index,

budget, genre, language, popularity, and the number of awards won by the director till date.

References

1. TMDB 5000 Movie Dataset | Kaggle
2. 350 000+ movies from themoviedb.org | Kaggle
3. The Movies Dataset | Kaggle