DPA Project – Proposal & Outline (CSP 571)

Group Members:

1) Mohana uma sushmanth Penumarthi - A20525576
2) Vinnapala Sai Ramya - A20526253
3) Harini Vaidya - A20525926

Project Title – Movie recommendation system

Project Proposal

Description of Project & Research goal

Now-a-days, the crucial task for the online streaming platforms is to attract users by recommending the relevant content what they like to watch. Recommendation systems play a crucial role in sectors like e-commerce and online streaming services, including platforms like Netflix, YouTube, and Amazon. Accurate recommendations for the next product, music, or movie enhance user satisfaction, prolong user engagement, and contribute to sales and profit growth.

The main objective of this project is to develop a recommendation system that assists users in discovering the best movie content based on ratings and genre. For this project, we take IMDB movie dataset from the official IMDB website and would like to analyze what kind of movies are more successful or got a higher reach to audiences. The results from this project can also help the streaming platform companies to understand the factors of successful movie and make a decision regarding future movie acquisitions.

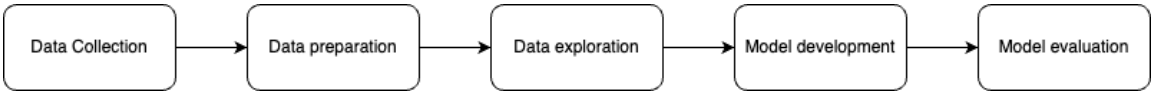
Specific Questions

The following questions would be addressed in this project:

- What are the average ratings for different genres?
- Are there certain genres that tend to have longer or shorter runtimes?
- Which genres receive the highest number of votes on average?
- Do these top-rated movies belong to specific genres or time periods?
- What kind of movies are most produced in yearly based?
- What is the distribution of the number of votes received by movies?
- What sort of movies in the dataset are classified as adult content?

Proposed methodology

The proposed methodology includes the following steps:



- **Data Collection:** we gathered the data from official IMDB website. We have collected two datasets. (Movies dataset and ratings dataset)
- **Data preparation:** In this phase, we process the data by cleaning and removing null values and transforming it into a suitable format for subsequent analysis.
- **Data exploration:** We perform data analysis in this step. we will use exploratory data analysis (EDA) techniques to identify trends, patterns in the data.
- **Model development:** This stage involves the development of various machine learning and statistical models, such as linear regression, KNN, also employing model selection techniques to determine the most suitable model for our dataset.
- **Model evaluation:** In this step, we will evaluate the performance of our models. We will use a variety of evaluation metrics, such as accuracy, precision, recall, F1 score, Mean Absolute Error, Mean squared error (loss function) to assess the performance of our models

Metrics for Measuring analysis results

We will use the following metrics to measure the results obtained from this project:

- Precision: This metrics measures the proportion of true positive predictions among all positive predictions
- Recall: This metric measures the proportion of actual positive cases that are predicted correctly.
- F1 Score: This is a harmonic mean of precision and recall.
- Mean absolute error: This metric is used to measure the average of the sum of absolute differences between predictions and actual observations.
- Mean squared error: This metric measures the average of the squared differences between the predicted ratings and the actual ratings.

Project Outline

Literature review and related work

We have reviewed the existing projects and relevant papers based on movie recommendation system. Here are some,

Relevant projects:

- https://susanli2016.github.io/Modeling-Prediction-Movies/
- https://rpubs.com/susan_li/movie-time

Relevant papers:

- https://ddd.uab.cat/pub/elcvia/elcvia_a2020v19n3/elcvia_a2020v19n3p18.pdf
- https://link.springer.com/chapter/10.1007/978-981-13-1927-3_42

Related Datasets:

- https://data.world/datasets/movies
- https://grouplens.org/datasets/movielens/

Data sources and reference data with descriptions

we have collected the datasets from official IMDB website. In this project we are considering two datasets - movie dataset and rating dataset.

Dataset link: https://developer.imdb.com/non-commercial-datasets/

Movie dataset (title.basics.tsv.gz) consist of 10,48,575 rows and 9 features. The features of movie dataset are:

- **tconst** (string) - alphanumeric unique identifier of the title
- **titleType** (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- **primaryTitle** (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- **originalTitle** (string) - original title, in the original language
- **isAdult** (boolean) - 0: non-adult title; 1: adult title
- **startYear** (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
- **endYear** (YYYY) – TV Series end year. '\N' for all other title types
- **runtimeMinutes** (integer)– primary runtime of the title, in minutes
- **genres** (string array) – includes up to three genres associated with the title

Rating dataset (title.ratings.tsv.gz) consist of 10,48,575 rows and 3 features. The features of rating dataset are:

- **tconst** (string) - alphanumeric unique identifier of the title
- **averageRating** (integer)– weighted average of all the individual user ratings
- **numVotes** (integer) - number of votes the title has received

Since there are two distinct datasets, the crucial step in the data preprocessing phase involves merging the datasets using the "tconst" attribute as a key for the join operation. Furthermore, we have identified the null values within specific attributes, like "endYear", "runtimeMinutes," etc, will be handled and resolved during the data preprocessing step.

Data processing and pipeline

We will perform the following steps in data processing,

- Data collection & merging: We have collected separate datasets for movies and ratings. Now, we need to merge them using the appropriate key (tconst).
- Data cleaning: The data obtained from IMDB is in raw form, needs to be cleaned for accurate results. In this step, we will remove unnecessary attributes and replace missing values (NA/ null values) with suitable values.
- Data transformation: We will transform the data into a format that is suitable for our analysis.
- Data analysis: We will analyze the data based on the data description.

Data stylized facts

We'll analyze the data distribution and perform grouping operations on attributes that are correlated and exhibit similar observations to identify the patterns or changes in trends. Additionally, we'll use data visualization libraries such as ggplot2 to enhance clarity in understanding the data.

Model selection

The primary objective of this project is to recommend the top-performing movies by predicting ratings based on attributes such as movie type, number of votes, genres, isAdult, etc. To achieve precise results, we will train the data using several machine learning and statistical models, including linear regression, KNN, etc. Additionally, we will utilize model selection techniques to identify the most appropriate model for our dataset.

Software packages and applications

Software: R studio, R
Libraries: tibble, ggplot2, data.table, plotly, dplyr, caret, corrplot, Metrics, MASS, lubridate.

References

1) https://link.springer.com/chapter/10.1007/978-981-13-1927-3_42
2) https://susanli2016.github.io/Modeling-Prediction-Movies/
3) https://data.world/datasets/movies
4) https://rpubs.com/vsi/movielens
5) https://ieeexplore.ieee.org/abstract/document/8663822