Movies' IMDb Rating Prediction

Advait Lonkar
IMT2017002
International Institute of Information
Technology, Bangalore

Gandharv Suri
IMT2017017
International Institute of Information
Technology, Bangalore

Mili Goyal
IMT2017513
International Institute of Information
Technology, Bangalore

Abstract—One simple, but an effective way to determine if a movie is worth watching is to use its IMDb rating. The IMDb top 250 list is, despite the subjective nature of the matter, the list of best movies one can watch. We present a machine learning model to predict the IMDb rating of a movie based on various features provided in the dataset.

Keywords : IMDb rating, numerical features,

I. INTRODUCTION

IMDb (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, fan and critical reviews, and ratings [1].

IMDb registered users can cast a vote (from 1 to 10) on every released title in the database. Individual votes are then aggregated and summarized as a single IMDb rating, visible on the titles main page [2].

The IMDb ratings are accurately calculated using a consistent, unbiased formula, but by no means are these ratings qualitatively *accurate*. An argument can be made that the ratings are too simplistic, which is quite fair : millions of users rate an artistic expression and reduce it into a range of 1 to 10, so obviously some of the nuances of what makes a movie good are lost.

But more often than not, people visit IMDb to rate the latest movie they had watched, check ratings of other movies, make watch lists accordingly, and many other such activities. Though the rating is only a number, it can be very informative and useful in itself. A rating cast by thousands of viewers generally represents a well-aggregated rating of a movie.

The goal of the project is to predict the IMDb rating of a movie provided various features in the dataset.

II. PROBLEM STATEMENT

Given various data-points of a movie, predict the IMDb rating of the movie in the range of 1 to 10.

This problem falls under the domain of **Linear Regression**. The evaluation metrics of the presented model will be the variance, standard deviation, root-mean-square-error of the predicted IMDb ratings v/s the actual IMDb rating of a particular movie.

III. DATA

A. Data Description

We have a dataset of 5,043 samples which will later be split into training and testing sets. The dataset was taken from Kaggle [3].

The dataset consists of the following fields:

- **color** : a string field which identifies the color of the movie, color or black and white.
- **director_name** : a string field which identifies the director of the movie.
- **num_critic_for_reviews** : a numeric field which identifies the number of critic reviewers of a movie.
- **duration** : a numeric field which identifies the duration of the movie.
- **director_facebook_likes** : a numeric field which identifies the number of likes on facebook for the director of the movie.
- **actor_3_facebook_likes** : a numeric field which identifies the number of likes on facebook of the 3rd actor of the movie.
- **actor_2_name** : a string field which identifies the name of the 2nd actor of the movie.
- **actor_1_facebook_likes** : a numeric field which identifies the number of likes on facebook of the 1st actor of the movie.
- **gross** : a numeric field which identifies the gross revenue of the movie.
- **genres** : a string field which identifies the genre of the movie.
- **actor_1_name** : a string field which identifies the name of the 1st actor of the movie.
- **movie_title** : a string field which identifies the title of the movie.
- **num_voted_users** : a numeric field which identifies the number of users who voted for the movie.
- **cast_total_facebook_likes** : a numeric field which identifies the number of likes on facebook of the cast of the movie.
- **actor_3_name** : a string field which identifies the name of the 3rd actor of the movie.
- **facenumber_in_poster** : a numeric field which identifies the number of faces in the poster of the movie.
- **plot_keywords** : a string field which identifies the plot keywords of the movie.

- **movie_imdb_link** : a string field which identifies the imdb link of the movie.
- **num_user_for_reviews** : a numeric field which identifies the number of user for reviews of the movie.
- **language** : a string field which identifies the language of the movie.
- **country** : a string field which identifies the country of the movie.
- **content_rating** : a string field which identifies the conten rating (R or PG-13) of the movie.
- **budget** : a numeric field which identifies the budget of the movie.
- **title_year** : a numeric field which identifies the year of release of the movie.
- **actor_2_facebook_likes** : a numeric field which identifies the number of likes on facebook of the 2nd actor of the movie.
- **imdb_score** : a numeric field which identifies the imdb score of the movie.
- **aspect_ratio** : a numeric field which identifies the aspect ratio of the movie.
- **movie_facebook_likes** : a numeric field which identifies the number of likes on facebook of the movie.

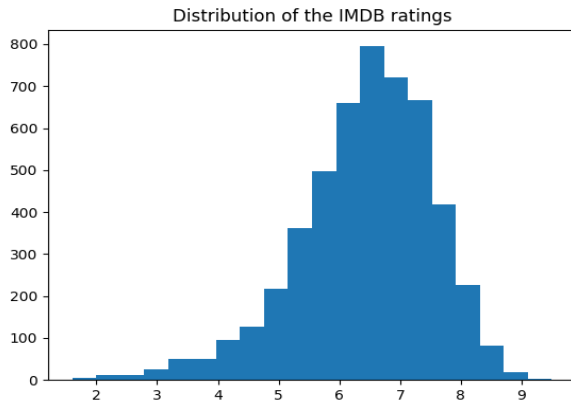## B. Data Exploration and Visualization

Fig. 1: IMDB-Score histogram

IV. PRE-PROCESSING

A. Dropping rows

The columns $gross$ and $budget$ had the maximum number of missing values, 884 and 492 respectively. Imputation of so many values would provide faulty results, so we decided to drop the rows corressponding to the missing fields of columns $gross$ and $budget$.

B. Dropping columns

1) $aspect\_ratio$: The two most frequent values of $aspect\_ratio$ and the rest of the $aspect\_ratio$ values had similar mean IMDb ratings, i.e. the aspect ratio values did not have any significant effect on the IMDb ratings. So we dropped the column $aspect\_ratio$.
2) $language$: Most of the movies (more than 90%) are in English, so we dropped the column $language$.
3) $color$: Most of the movies are colored (more than 90%), so we dropped the column $color$.
4) $plot\_keywords, director\_name, actor\_1\_name, actor\_2\_name, actor\_3\_name, genres, movie\_imdb\_link$: Most of the values in these fields are unique, so we dropped all of these columns.

C. Labelling

The $country$ column had over 79% of values as USA and 8% as UK. So we labelled all the other countries as $Others$.

V. FEATURE EXTRACTION

VI. MODEL BUILDING

REFERENCES

[1] Christian Stocker. "20 jahre internet movie database: Im clubhaus der kinojunkies".

[2] IMDb help. https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/g67y87tfyyp6twav.

[3] Dataset from Kaggle. https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset.