Performance prediction of movies in India based on quantitative parameters

Adittya Gupta [1] **Gautam Sharma** [1] **Sparsh Mittal** [1] **Yash Raj Singh** [1]

Abstract

The current Cinema industry has about a ₹ 15,000 crore market size in India, with about 3,000 films being produced each year (Wikipedia contributors, 2024). In light of the above, predicting a movie's performance in the Indian audience becomes a helpful measure for all the stakeholders of the industry to improve and modify their cinema to closely cater to the needs of the Indian audience, while also booking a profit in the process. This project attempts to solve the problem of predicting any movie's performance in the Indian market by using machine learning.

1. Introduction

The current Cinema industry is not able to efficiently cater to the needs of its audience as while creating a movie, the creativity of the makers doesn't take into account the market trends, which are reflective of how a movie would perform at the box office.

The entire industry's success is based on hit and trial, or luck, which should not be ideal, given the industry's size and influence over the Indian GDP. If the makers of a film could see how their cinema would perform at the box office at the current time, then it would be a win-win situation for both the makers and the audience, as the audience would get a good (intended) experience in the theaters and the producers would be able to book good profits off the movie.

We plan to make a predictive model trained on years of historical data of movies, to be able to predict a movie's performance at a certain time.

2. Literature Review

Apala et al., had previously applied K-means on a dataset generated using social media for the given task (Apala et al., 2013), but they employed a pre-trained model, in the form of Weka to apply K-means. Abidi et al., had applied various models using a relatively small training set with very less predictors to solve the above problem (Abidi et al., 2020). The former group found the **Generalized Linear Model (GLM)** to be the best performer among all the methods used. Sharda and Delen applied neural networks with almost exhaustive inputs to solve the problem (Sharda & Delen, 2006). Earlier Sawhney and Eliashberg (Sawhney & Eliashberg, 1996) used the early box office collections of a movie to predict its lifetime performance at the box office.

2.1. Proposed Novelties

Our work extends on the previously done works, as we employ a bigger dataset, sensitive to Indian lingual cinema. We also plan to apply NLP on the movie's storyline, fetched from the dataset to further classify the movies.

3. Dataset Generation

We generate the dataset through web scrapping the IMDb website. We have included the following features in our dataset: Time of release, Genre, Cast, Directors, Production Company, Storyline, Original Language, Country of Origin, and Duration.

The code for dataset generation is provided here.

A histogram of the dataset for the IMDb rating is given in Figure 1.

References

Abidi, S. M. R., Xu, Y., Ni, J., Wang, X., and Zhang, W. Popularity prediction of movies: from statistical modeling to machine learning techniques. *Multimedia Tools and Applications*, 79(47):35583–35617, Dec 2020. ISSN 1573-7721. doi: 10.1007/s11042-019-08546-5. URL https://doi.org/10.1007/s11042-019-08546-5.

Apala, K. R., Jose, M., Motnam, S., Chan, C.-C., Liszka, K. J., and de Gregorio, F. Prediction of movies box office performance using social media. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 1209–1214, 2013. doi: 10.1145/2492517.2500232.

[1]Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati, India. Correspondence to: Adittya Gupta <g.adittya@iitg.ac.in>, Gautam Sharma <gautam.sharma@iitg.ac.in>, Sparsh Mittal <m.sparsh@iitg.ac.in>, Yash Raj Singh <yash.singh@iitg.ac.in>.

Course Project, Instructor: Prof. Ashish Anand

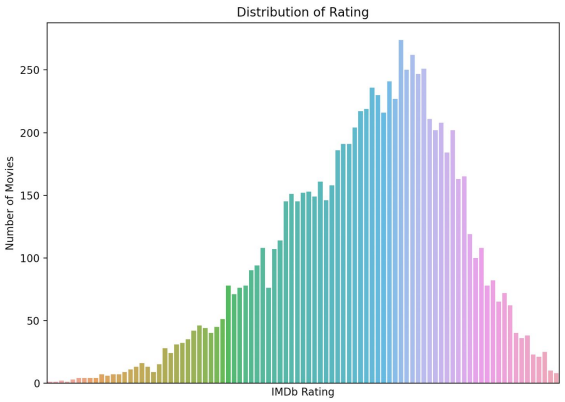

Figure 1. Histogram of the custom dataset for the IMDb rating

Sawhney, M. S. and Eliashberg, J. A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, 15(2):113–131, 2024/02/22/ 1996. URL http://www.jstor.org/stable/184189. Full publication date: 1996.

Sharda, R. and Delen, D. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2):243–254, 2006. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2005.07.018. URL https://www.sciencedirect.com/science/article/pii/S0957417405001399.

Wikipedia contributors. Cinema of india — Wikipedia, the free encyclopedia, 2024. URL https://en.wikipedia.org/w/index.php?title=Cinema_of_India&oldid=1209395942. [Online; accessed 22-February-2024].