Forecasting Box Office Earnings: A Machine Learning Approach to Assess pre-release Movie metrics and Online Sentiment

Noor Assi

Level of Interest 7/10

Description:

As an avid movies' enthusiast, I spend a lot of my free time watching movies of all genres. I have seen some movies that I enjoyed a lot but ended up not doing well in the Box Office. I am curious in trying to predict and study the factors that could cause that. The objective of this project is to predict the financial success of a movie before its release. The entertainment industry invests heavily In movies, and predicting their success can guide marketing strategies and investment decisions. This project would explore how factors like marketing spend, star power, genre, and pre-release social media sentiment can predict a movie's box-office earnings.

Data:

The dataset would include both numerical and categorical features. Numerical features could include marketing spend, IMDB rating of the lead actors, and social media mentions (number of hashtags about the movie.) Categorical data includes the genre, MPAA rating, whether the movie is a sequel, and if the movie has a similar possible competitor.

Example Dataset Instances:

Marketing Spend	Lead actor Rating - IMDb	Genre	MPAA rating	Is-Sequel	Social Media mentions	Target: Success
12 Million	7.5	Action	PG-13	Yes	50000	Yes

Gathering Data:

Gathering the data requires multiple resources. Industry reports provide marketing spends. Ratings from IMDb could be accessed by anyone through the website. Genre and MPAA ratings are also provided through the movies databases publicly. Social media mentions could be calculated across different social media platforms like Twitter and Tiktok using the NLP technique. As for the target variable, it could be determined based on the box-office earnings relative to the budget. Box Office Mojo is a website that provides these details.

The data will require cleaning and normalization intensively to be ready for a machine learning model, but there are a lot of other factors that could be considered as well. I also think that the project has real-world capabilities since the entertainment industry is in need of reliable predictive models for movie success.