Final Project Proposal : Predicting Movie Box Office

Fei Xu, Rita Zou, Rus Refati, Xinyu Cao and Yang Li

I. OBJECTIVE

Investing in companies from the film industry on the stock market is harder than investing other companies in traditional industries because these film-related companies' stocks are strongly affected by the financial performance of their upcoming movies in the short term. Hence, accurate and reliable prediction of the Box Office Revenue (BOR) of a movie before releasing is an in pressing need, as it greatly informs the tough investment decision-making process when facing such a high-risk and high-yield opportunity. In addition, the prediction is highly important for advertisement companies that seek to embed their ads in popular movies. Such a prediction can also assist cinemas in scheduling movies and help people choose movies to watch. The objective of this project is to help movie studios and producers to decide whether or not to invest their money and time in production of a movie depending on the factors that are in our model.

II. MOTIVATION

The global movie industry was worth $ 42.2 billion a year in 2019. The impact of the COVID pandemic on worldwide box office revenues has reduced the estimated figure from 44.5 billion U.S. dollars to 16.3 billion for the year 2020.[1] The July estimates indicate a more severe impact than was predicted in March 2020, when revenue was expected to drop to only 32.3 billion U.S. dollars. Across the globe, cinemas were shut down for Q2 2020, and the damage to revenue is projected to last for the next 5 years, although small annual growth is still expected as of 2021. Predicting box office revenue (BOR) of movies before releasing on big screens successfully becomes an emerging need, as it informs investment decisions on the stock market, the design of promotion strategies by advertisement companies, movie scheduling by cinemas, etc. Movie production costs a lot of money and effort. If we could make an accurate prediction on how well a movie will perform based on its budget, cast members, directors, themes, etc., it will greatly help movie producers to optimize their resources for higher return on investment (ROI) movie projects.

III. DATA DESCRIPTION

We will acquire data from The Movie Database (TMDB) through API. TMDB is a community built movie and TV database with every piece of data added by their community dating back to 2008. The reason we choose TMDB over IMDB is that we also want to include the factor of audience's feedback since TMDB has clean well-moderated discussion boards you can access from each movie page. TMDb's strong international focus and breadth of data is largely unmatched. This dataset includes nearly 700,000 movies with key information[2], such as release date, theme, duration, director, cast & crew, budget, revenue, keywords, user score, etc., that covers many different perspectives of a movie. Some of the details are listed below:

- The genre of the film, a categorical variable classifying the film as Action, Children's, Comedy, Documentary, Drama, Horror, Science Fiction, or Thriller.
- The Motion Picture Association of America (MPAA) rating of the film, one of the ratings G (general audiences), PG (parental guidance suggested), PG-13 (possibly unsuitable for children less than 13 years of age), R (children not admitted unless accompanied by an adult), NC-17 (no one under 17 admitted), and U (unrated).
- The origin country of the movie, classified as U.S., English-speaking (but not U.S.), or non-English-speaking.
- The production budget of the film (in millions of dollars).
- The gross revenues (in millions of dollars) for the film's first weekend of general release.
- Ratings: The rating of the movie went from zero to 100% based on users scores.

Besides, various questions can be asked to better understand the data:

- How was the popularity of a movie over the years?
- Considering the five recent years, how is the distribution of revenue in different score rating levels?
- How is the distribution of revenue in different popularity levels?
- What kinds of properties are associated with movies that have high popularity?
- What kind of properties are associated with movies that have a high voting score?
- How many movies are released year by year?
- What are the keywords trends by generation?

We plan to either build a regression model to predict the revenue of a movie or build a classification model to predict the "success" of a movie using ROI that may be easily understood. A successful movie is evaluated by its popularity, vote average score(Ratings) and revenue. There are some keys that can affect the success of a movie. For example, the Budget, Cast, Director, Tagline Keywords, Runtime, Genres, Production Companies, Release Date, Vote Average, etc. In this project, we are going to use Numpy, Pandas, and Matplotlib to investigate the data based on the questions we

came up with. We will also plot the feature importance to find top factors that are related to the revenue or success of a movie.

IV. MEMBER RESPONSIBILITIES

There are five members in our group and all the group members will collaborate on each step of this project and brainstorm appropriate features to include in the model and decide which model to use based on model performance, training time, and interpretability of the model. Yang Li is the project manager and each member will take main responsibility for one step. Member responsibilities are listed as follows:

- Rus Refati: has a computer science background and will be responsible for data gathering from TMDB.
- Rita Zou: has a business analytics background and will clean the queried data and create new features as needed.
- Yang Li: has a business analytics background and will provide help as needed for data engineering. She will perform exploratory data analysis and get insights before model building.
- Fei Xu: has a business analytics background and will be responsible for model building and model comparison.
- Xinyu Cao: has a business analytics background and will use appropriate metrics to evaluate the model and tune the model based on performance.

V. TIMELINE

The milestones of this project are listed below:

- Data gathering: 1-2 weeks, done by Oct 10th
- Data cleaning and pre-processing: 2-3 weeks, done by Oct 24th
- Model building and training: 2 weeks, done by Nov 7th
- Model evaluation and tuning: 2 weeks, done by Nov 21th
- Wrap-up: 1 week, done by Nov 28th

VI. EXPECTED OUTCOME

The expected deliverable is a functioning model that uses the key info of a movie to predict its revenue or success with acceptable accuracy and evaluation metrics (MSE or F1 score). We will also provide recommendations on what factors are more important in movie production. Based on the learned features, we train a mutually-enhanced prediction and ranking model to obtain the box office revenue prediction results. Finally, we apply the framework to the film market and conduct a comprehensive performance evaluation using real-world data. We will use experimental results to demonstrate the superior performance of both extracted knowledge and the prediction results.

REFERENCES

[1] https://www.statista.com/statistics/1170721/impact-coronavirus-global-box-office-revenue/

[2] https://www.themoviedb.org/?language=en-US