University of Toronto
Faculty of Applied Science and Engineering
MIE368 - Analytics in Action
Preliminary Report - Team 17

Name	Student Number	Email
H.E.	—	—
I.C.	—	—
J.L.	—	—
A.K.	—	—

Problem Statement

Movie production involves substantial financial investments, such as production costs and marketing expenses. Predicting box office revenue is a measure of people's interest in a particular movie, and governs financial decisions made by producers and investors [1]. The initial motivation is to establish a robust framework that benefits stakeholders in the film industry such as distributors and studios. These factors will aid in making data-driven decisions that will increase the likelihood of financial success. The cast and directors have a direct influence on the revenue since movies with a well-known cast and director tend to help guarantee financial investments and production development, contributing to a greater box office revenue. To showcase the effect of leading cast and directors on box office performance, the original model of predicting box office has been altered to introduce these parameters. The current iteration of the model aims to answer the question: "Given a set of attributes for an upcoming movie, what is the box office produced for the best combination of leading actors and the most suitable director for a movie fitting the given attributes?" Thus, the objective of the model is to determine the optimal set of leading actors and directors, enabling the prediction of box office revenue across various genres.

Data Collection

The initial data collection process obtained various data sources from Kaggle [2] and IMDb's non-commercial database [4]. Upon further research, due to inconsistencies in the Kaggle datasets, and the lack of relevant information provided by IMDb, alternate data sources were researched. "The Numbers" is a non-commercial database website that contains information regarding films dating from the year 1899 to 2031 [4]. To collect the data, Python, pandas, and BeautifulSoup were used to web-scrape the data. Based on data collected from The Numbers and research done to determine the important factors to incorporate in the model, 12 primary decision variables were created. In addition, OpusData, an information services company, provided a sample dataset upon request [5]. The dataset provided consists of approximately 1900 samples from movies ranging from the year 2006 to 2018. Since the dataset does not account for more recent movies, the data will be supplemented with the web-scraped data, which upon processing, will contain approximately 2000 samples, ranging from the year 1995 to 2022. "The Numbers" also contains information regarding actors and directors including the number of movies they participated in, the domestic box office of the movie, and the average domestic box office per movie [6]. From the data selected, the

ultimate factor that influences the revenue of the movie is the population of potential viewers and the quality of the movie. The quality of the movie also highly depends on the cast and director.

Methods

The analysis of the data collected will be conducted through two main methods, which subsequently employ different techniques to arrive at solutions. Via the problem statement, several analysis steps can be extracted, with the major categories being the following:

Genre	Rating	Creative Type	Distributor	Production Method	Source Material

To get both the optimal cast and director, two variations of the knapsack model can define the crux of the problem, where actor and director values representing their star power can be maximized with the decision variables of choosing a particular actor for the category in question. The constraints would be that the number of actors does not exceed the given cast size and that the average salaries do not exceed a certain percentage of the movie budget.

The actor and director values will both be dependent on the average box office earned per movie a, which is the quotient of the total box office gross earned b and the number of movies acted in or directed n. The measure of value for an actor and director, for simplicity, has been defined as the product of a and a parameter λ or γ, denoting the average percentage cut per movie for the actor or director respectively. Since λ and γ are hyperparameters, grid search can be used to determine them with the condition that for an actor, they are not ranked as top 50 and in the top 10 for the director. If they are top-ranked however, then the upper bound of the intervals for percentage cut $0 < \lambda \leq 0.1$ and $0 < \gamma \leq 0.2$ can be assumed due to their popularity and merit [7]. Since the optimization for cast returns a list of actors, the average of the actor values can be taken to get the overall cast value.

Through method stacking, the output of these optimizations would then be the input of a linear regression model predicting an estimated box office, along with the same categories of the movie in question, as well as the release month and the average budget of a movie fitting its likeness. To address the categorical variables, the correlation of each category can be determined via additional EDA and to check for overfitting or bias from the quantifying of the categories, cross-validation and feature engineering can be conducted.

Initial Findings

To conduct an EDA, web scraping techniques were employed to extract data from "The Numbers" website. This process was facilitated by using the BeautifulSoup and Pandas libraries to collect and organize data tables. By web scraping the movies released each year from 1995 to 2023, 52937 rows of movie-related information were collected. The table consists of factors such as movie name, release date, genre, release type, revenue to date, and trailer. By utilizing feature selection to remove columns such as release type, revenue to date, and trailer, and removing null values from the rows, the data set was shrunk to having 39392 rows and three columns. In addition, to highlight the production budget of a movie, a new data set was scraped, containing the production budget of movies from 1995 to 2023. By web scraping this new dataset, important factors such as release date, movie, production budget, domestic gross, and worldwide gross. Via model selection, columns such as domestic gross and worldwide gross were removed, resulting in a dataset with 6322 rows and three columns.

Furthermore, the cast and director of a movie are crucial aspects of the success of a movie. To account for these factors, cast value, and director value variables were created. The data that corresponds to this is also from "The Numbers." The website contains information on the most successful actors for each category as listed in the methods sections. This table contains columns on the total box office gross, the number of movies they took part in, and the average box office per movie, for all categories. This data will be used as inputs for the actor optimization model. The same information is organized by categories for directors as well, which will be extracted for use in the director optimization model.

Completion Plan

The next steps for the project consist of finalizing the EDA and arriving at three main datasets for each model. This entails combining all the scraped data containing all the categories and features of the movies into one, as well as distinct ones for the actor and director information. For these three final datasets to be constructed, additional web scraping must be conducted for the remaining columns of the total domestic box office, the number of movies involved, and the average box office per movie for both actors and directors. To enhance result accuracy, feature scaling can be conducted to reduce deviations in the variables and a new feature of crew value can be engineered through the aggregation of the cast and director values, which helps to mitigate overfitting and bias.

References

[1] Predicting box office revenue for Movies - Stanford University, https://snap.stanford.edu/class/cs224w-2015/projects_2015/Predicting_Box_Office_Revenue_for_Movies.pdf (accessed Oct. 5, 2023).

[2] U. Singh, "Movie dataset: Budgets, genres, insights," Kaggle, https://www.kaggle.com/datasets/utkarshx27/movies-dataset/data (accessed Nov. 3, 2023).

[3] A. Hayes, "What is a franchise, and how does it work?," Investopedia, https://www.investopedia.com/terms/f/franchise.asp (accessed Nov. 3, 2023).

[4] IMDb, https://developer.imdb.com/non-commercial-datasets/ (accessed Nov. 3, 2023).

[5] OpusData, https://www.opusdata.com/about.php (accessed Nov. 3, 2023).

[6] "Box Office Star Records," The Numbers, https://www.the-numbers.com/box-office-star-records/ (accessed Nov. 3, 2023).

[7] R. Provost, Rex Provost, "How much do directors make - DGA rates for Film & TV," StudioBinder, https://www.studiobinder.com/blog/how-much-do-directors-make/#:~:text=This%20 (accessed Nov. 3, 2023).