Amy Chitnumsab (ac2295)
Tejhan Diallo (td382)
Jennifer Kim (jjk358)

Team 3 – ORIE 4741 Project Proposal

Our project aims to predict the success of future movies, both in terms of commercial and critical success. The dataset used will be *The Movies Dataset*, which has data scraped from the IMDb (Internet Movie Database) website. This dataset has information on 45,000 movies such as budget, revenue, release date, genre, and production company. We hope to use relevant features of this dataset to 1) predict box office revenue and/or ticket sales, which would be a measure of commercial success and 2) IMDb star rating, a potential measure of critical success and/or public reception, which could lean favorable or not. (IMDb ratings are weighted to take into account the number of votes and reviews from the general public and not film critics).

To answer these questions, we will use a combination of data analysis and modeling from the dataset above. Data analysis and visualization can help determine what factors best correlate with the commercial success of a movie, while various models can predict the box office earnings or IMDb ratings of a movie. But first, we may have to first clean, add to, or transform parts of the dataset. We will have to adjust the budget and gross earnings to inflation rates and either add or omit missing data for the budgets and earnings.

Diving into data analysis and visualization, we can use a wide range of graphs to represent any possible correlation between different variables. In general, multi-line charts or scatter plots can be used to compare continuous data and how they change. Bar graphs or pie charts are suitable for comparing discrete or nominal data. Radar charts can also be useful for comparing numerical with nominal data as we can display how different categories rank on a numerical scale.

Once we identify what variables have the strongest correlation with the success of movies, we can use them to create models for predictions. These variables would make up our feature space. An offset may be added if needed. Feature engineering may also be used to transform the raw data into features better for predictions. Linear regression may be used to predict box office earnings. Binary classification can determine if a movie will have a good or bad IMDb rating. This can be done with support vector machines or logistic regression. We can compare multiple models based on their performance.

Given the profitability and prevalence of the movie and entertainment industries, it is of relevance to predict how well a movie might perform in the box office given past data. Thus, the questions analyzed by this project may be of importance to film distributors and studios, who might have financial interest in how well a film performs at the box office given the necessity of film financing and investing. Predicting tickets sold for a movie is still a valid metric for profitability and film success. Concerns that the movie theater industry is dying due to factors like the pandemic and rise of disruptors like Netflix exist, but the future of movie theaters is still bright and evolving. Of course, other factors like continued marketability of films from non-box office revenue sources might also factor in, but predicting movies' box-office performance (and if possible, public reception) would be a reliable tool for studio decision makers and movie investors.