Team Final Project

Section 101B, Team 11

Team members: Xueyao (Ashley) Gu, Paras Shukla, Rubing (Ruby) Song, Jingnan Wang, Ji (Charlotte) Wu, Madhurima Yella

Business Understanding

Our Client, a VC firm, is interested in entering the media and entertainment market and specifically wishes to invest in movies with a good chance of earning higher than average returns at the box office. However, the impact of COVID-19 has been severe on the Media and Entertainment industry with many production houses and established media giants reporting record losses. Pandemic has given rise to a growing popularity of OTT streaming platforms and has also impacted the box office potential of films. To better guide the investment policy of our client, we have decided to use machine learning to look at the past performance of available films and use this as a reference to try and predict how future films would perform at the box office which will allow our client to make sound investments and pick the right films to produce and back.

Data Understanding

To construct a data analysis in order to solve our business problem dealing with aspects that will lead to a high profit of a movie, we used the internet to find an ideal dataset for our analysis. We chose the 'IMDB 5000 Movie dataset' from the Kaggle Website (https://www.kaggle.com)[1], and our dataset is a four-year-old dataset based on original data scraped from the IMDB Website (https://www.imdb.com), which is recent enough for our analysis. Our dataset describes the information of approximately 5000 movies (28 variables describing the movie information such as aspect_ratio, director_name, gross, budget etc). An instance in our dataset is a unique movie with its traits. The business metric that we would like to target in our problem is the profit of the film. In our dataset, we have two variables - "*gross*" and "*budget*" to calculate the profit of a movie, assuming the entire budget is used during movie

[1] Yueming. (2017, December 16). *IMDB 5000 movie dataset*. Kaggle. Retrieved October 7, 2021, from https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset/version/1.

production, which acts as the cost of the film, given the gross of the movie. Therefore, our target variable to analyze based on our dataset is the *profit* of a movie, and our data analysis is supervised based on this clearly stated variable. We believe that our dataset is an excellent source for analyzing the aspects that influence a movie's profit to provide shareholder insights. At the same time, we acknowledge that certain aspects of the data are exposed to bias for future data preparation and modeling.

Firstly, we have only 5000 rows of data in our dataset, which is considered relatively small for data modeling compared to the total number of movies in human history and could lead to sample bias for modeling. In addition, as we find out, the data is exposed to selection bias. For example, out of 5000 movies in the dataset, 4704 movies have their primary language as English. Of course, our dataset does not represent the real-life setting - such a massive proportion of films from Anglophone countries. We believe this is because IMDB is owned by the American company 'Amazon' which could be causing the bias in data selection.

Secondly, our data contains a considerable amount of missing values and misleading data. After we checked for the 'NA' values for each variable, we noticed substantial amounts of 'NA' values in *'gross'* and *'budget*' - which are the main variables we need to calculate the target variable- *Profit*. We would address them in data cleaning. In addition, after careful investigation, the currency in the gross column is all represented in dollars, but the budget may vary in currency. This discrepancy in currencies posed a problem in the calculation of profit.

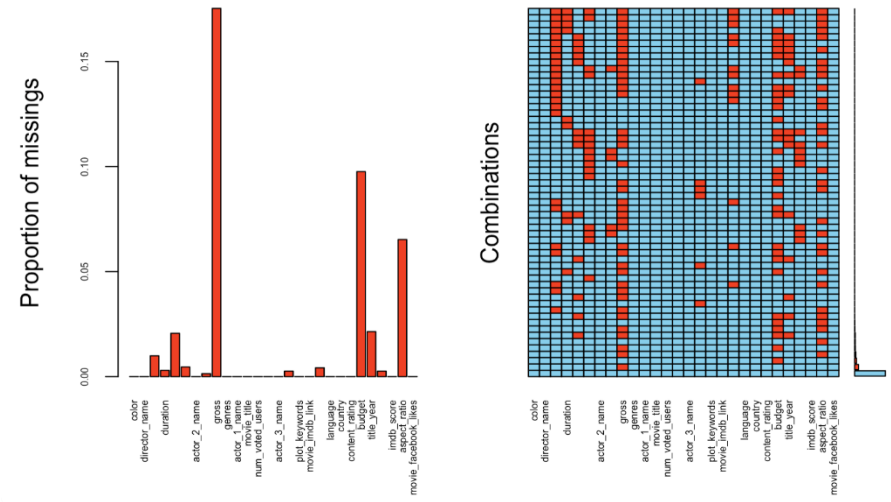

Figure 1: Null Value Visualization

Lastly, we also found out that for certain variables, such as 'director_facebook_likes,' there is a large amount of data showing '0' for this category which is unreasonable in today's media setting. For example, there are 1011 rows in our data that have '0' for director_face_likes. Such a problem will not be helpful for our later analysis and will require further adjustment.

Data Preparation

Apart from the problem of a large amount of missing values in *gross* and *budget*, which we will use to calculate *Profit* as well as the discrepancy between currencies in *budget* and *gross*, another problem we identified is that the gross amount stated in the raw data is domestic gross instead of gross worldwide. We believe gross worldwide is a better and comprehensive metric compared with only the domestic gross to reflect the revenue generated on our investment.

The variable *aspect_ratio* also has a large number of missing values which we could use in our analysis. Based on these concerns, we decided to scrape the data directly from IMDB website for budget and gross worldwide as well as aspect ratio. Because some links for the movie are no longer valid, we managed to scrape 4919 movie records. Subsequently, we cleaned the currency sign as well as the comma for currency display in the *budget* and *gross_world*; converted the string display format of the *aspect_ratio* to float. For the difference in currency translations, we identified those movies whose budget is in foreign currency and utilized the currency exchange rate provided by the Federal Reserve based on each movie's title year and budget currency sign to reach a corrected amount, which we named *budget_corrected*[2]. Because *gross_world* and *budget_corrected* are used to calculate the profit, we dropped all the records which contain NULL values in either gross_world or budget_corrected. We were left with 4188 records and 202 NULL values in aspect_ratio. Instead of dropping these NULL values (since the size of the dataset is small), we filled them with the mean of the existing values in *aspect_ratio* (approximately 2.1176).

For other missing categorical variables like "content_rating", "language", "color", "actor_3_name", "actor_2_name", "actor_1_name", we dropped the NULL values.

In Genres, a movie may have multiple genres. The way genre is formatted is that each genre is separated using "|".

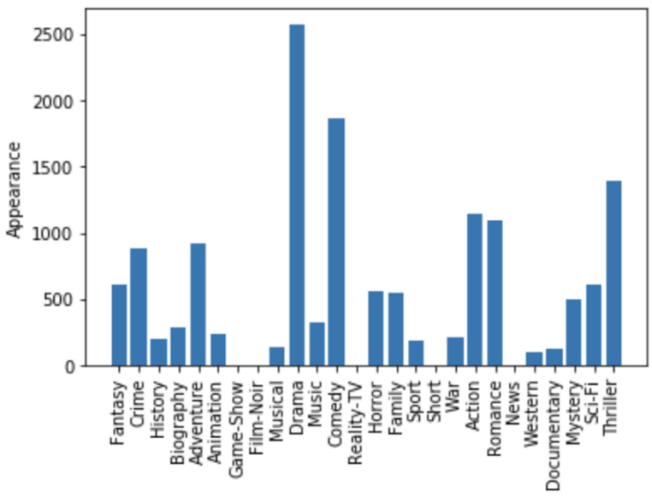

Figure 2: Frequency Table by Genre

[2] *Board of governors of the Federal Reserve System*. The Fed - Foreign Exchange Rates - Country Data - H.10. (n.d.). Retrieved October 9, 2021, from https://www.federalreserve.gov/releases/h10/hist/.

For better analysis, we separated each genre and used each distinct value as a variable. If a movie belongs to a certain genre, we denote the value under that genre to 1, if not, 0. Based on the bar plot shown above, *Game-show, Film-Noir*, *Reality-TV*, *Short* and *News* appear most infrequently, we then categorize these genres as *Genre_Others*.

For *plot_keywords*, which is a lot more complicated and diverse compared to *genre*, we decided to add up all the frequencies of a certain keyword within the dataset and use that to gauge how niche/mainstream a movie is. The higher the score, the more main-steam a movie is. We denote that variable as *main_stream.* This treatment might be questionable. A better treatment can be to investigate the frequency of the plot_keywords from all the movies on IMDB website, but for now we assume that our sample is representative.

Columns which are difficult to measure simply using text- *director_name, actor_1_name, actor_2_name, actor_3_name*, we referred to the IMDB official website to find the top 100 directors and actors to measure their influence. If the name from all four variables appears in the list, we denote 1, else we will denote 0. The respective names for the columns are *director_name_100, actor_1_name_100, actor_2_name_100, actor_3_name_100*.

We dropped *director_facebook_likes* because of a lot of '0's in the data column which had little to no use for conveying the true popularity of directors, and we endorsed that the newly created *director_name_100* would explain their popularity.

For *language*, due to the disproportionate number of English films and other languages, we created *language_English* to denote whether a language is English or not.

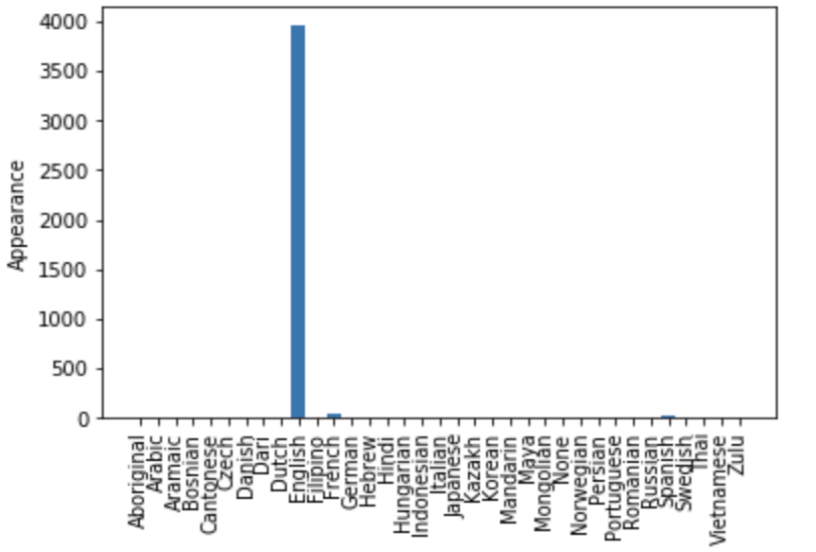

Figure 3: Frequency Table by Language

We applied the same treatment to the country, using country_USA as a dummy variable.

For other variables, which cannot be obtained until the movie has been released, we decided to not use them. Those variables won't be available when we make our investment decisions.

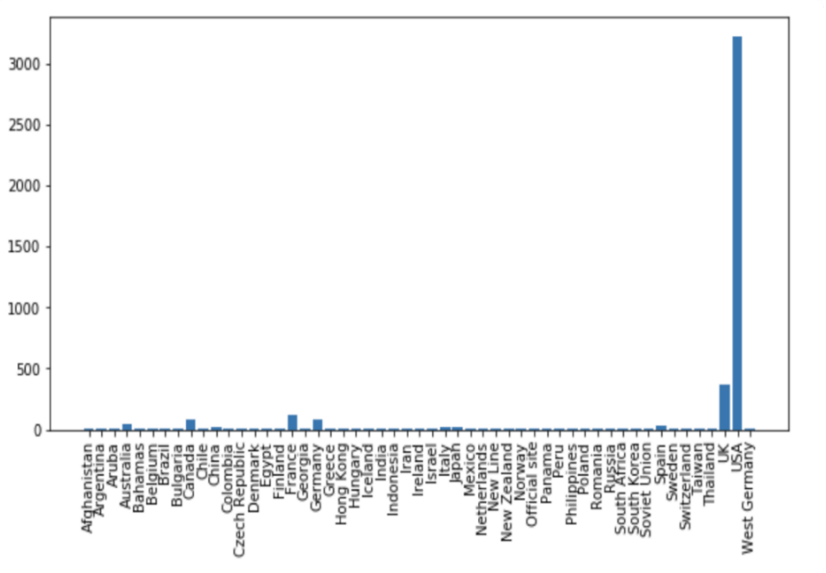

Figure 4: Frequency Table by Country

After the cleaning, we are left with features *title_year, country_USA, Language_English, color, duration, aspect_ratio, director_100, actor_1_name_100, actor_1_facebook_likes, actor_2_name_100, actor_2_facebook_likes, actor_3_name_100, actor_3_facebook_likes, cast_total_facebook_likes, main_stream, Horror, Fantasy, Crime, History, Biography, Family, Adventure, Animation, Sport, War, Musical, Drama, Music, Action, Romance, Western, Documentary, Mystery, Sci-Fi, Thriller, Comedy, Genre_Others, budget_corrected*, and target variable *Profit,* which is calculated using *gross_world - budget_corrected*. For better communication, we used million to describe *Profit* in future model building processes.

Exploratory Data Analysis

Multicollinearity in our dataset is a potential problem. As we constructed a correlation matrix for variables in our dataset, we noticed certain independent variables have huge correlation between each other which may suggest adjustment in further modeling. For example,

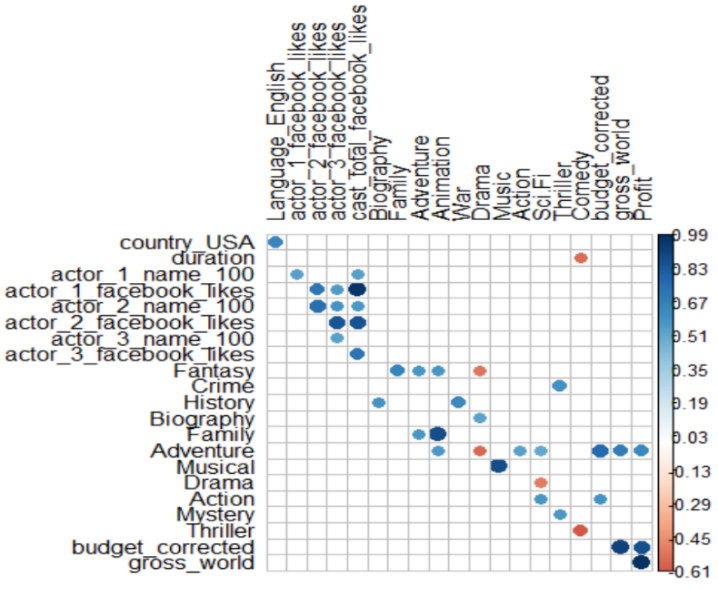

Figure 5: Correlation Matrix

actor_1_facebook_likes and cast_total_facebook_likes have high correlation between each other (above 90%). With these problems and bias in mind, we will try to best address them in further modeling.

As we created over 20 dummy variables for the genres of movies, we are interested to see the relationship between a movie's genre and its profit. To accomplish this goal, we created the bar graph to analyze this point. We calculated the average profit for movies in each genre and aimed to compare the average profit in a barplot. From Figure 6, we can see that animation, adventure movies are among the genres which generate the highest profit, while documentary and sport movies generally generate lower profit. The visualization helped us gain insights for further modeling.

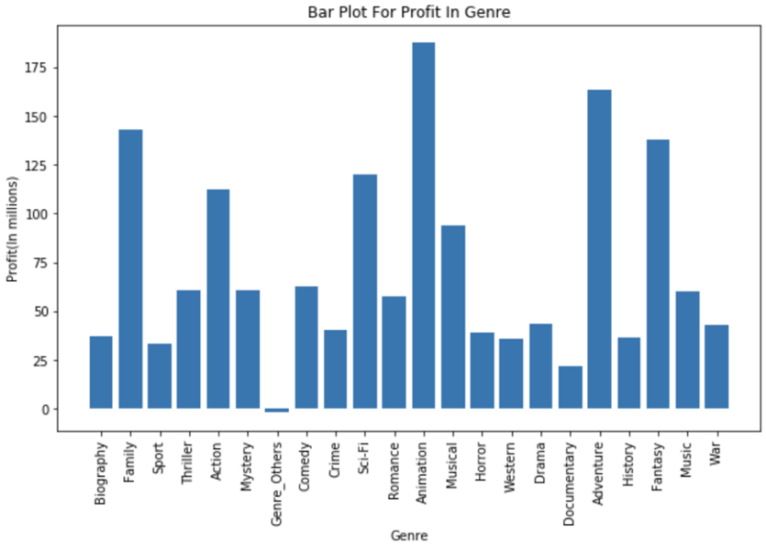

Figure 6: Profit by Genre

In addition, when we choose the profit for a movie as the business metric to analyze, we use the budget of the movie and its gross. We wanted to find whether a highly budgeted movie will ensure a high return in profit. To analyze this question, we fit a scatter plot for the profit of the movie against it budget. The results showed that a movie's profit is positively related to its budget, suggesting that higher the cost of the movie will generate a higher profit.

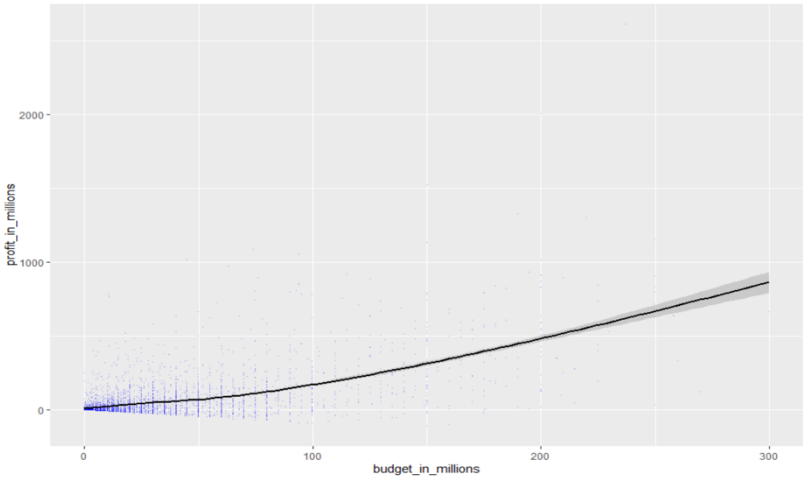

Figure 7: Scatter Plot Between Budget and Profit

Modeling

Our core task is to build a regression model to forecast profit based on given features. Models we chose are Linear Regression, Lasso Regression and Random Forest.

Linear Regression is the most used method used in regression. The upside of using Linear Regression, in our case, is that we could identify which elements of a film could potentially affect profit and by what amount. However, downside is that it assumes a linear relationship between features and target, which might not be a great fit if the underlying connection is complex. In addition, human judgment must be used to select features based on P-values.

Lasso is another way to build a regression model. It is for regularization purposes to avoid overfitting. The upside of it is that it can select features for us to prevent the overfitting problem. The downside is that we need to decide on the "Lambda" to penalize based on the number of features we selected. It might also overlook some features that may be insignificant but interesting.

Random forest is an ensemble learning that randomly selects N features on a bootstrapped dataset, iterating multiple times to create several individual decision trees and splits out the average prediction. The upside of it is that it avoids overfitting by averaging the performance of individual trees. It also performs well on large datasets with a wide range of variables. The downside of it is that, we won't know which features might affect the profit and how.

All three models can perform regression analysis. Given unique features of specific films and budget estimation, we can predict profit and decide on the highest one to invest in or the ones with the highest total profit, given that total estimated budget does not exceed our expected amount.

For linear regression, we removed insignificant variables such as country_USA, Language_English, title_year, color, actor_1_name_100, actor_2_name_100, actor_3_name_100 as well as main_stream, we also removed cast_total_face book_likes given the high correlation it shares with actor_1_facebook_likes.

We used a 10-fold analysis to decide on the lambda that produces the minimum cross-validation error for lasso regression. We also tested 1 to 38 features selection for random forest and decided on the one with the least MAE.

Evaluation

To test the predictive power of the three models, we conducted a 10-fold cross-validation test. We split the 4146 datasets into 10-fold. For each fold, 90% of the dataset would be the training set, and the rest 10% would be used as the validation set. The metric we used to decide which model to choose is mean absolute error (MAE). The final results are as follows:

	mtry	validate_mean_mae	validate_mean_r2	train_mean_r2
Linear Regression		67.51	0.36	0.37
Lasso Regression		67.76	0.37	0.39
Random Forest	9	60.65	0.44	0.45

Figure 8: Model Performance

It turns out that a random forest with nine randomly selected features produces the least average MAE among the validate set and the highest correlation within both the train and the validate set. Linear regression slightly outperforms lasso regression in terms of average MAE and the average R^2 within the validate set. In light of that, random forest generated a minor cross-validation error; we will choose it as our model.

In terms of movie investment choice, we need to collect our model's features. Since Random Forest is a black-box algorithm, we won't know what features come into play, but all those variables are relatively easy and cheap to collect. Based on features and estimated budget, we can predict profit.

Deployment

For deployment, we need to collect all the required variables in the random forest and get the predicted profit to determine the most commercially successful movies for our client to invest in.

There are several problems and risks regarding our deployment:

1. The inaccuracy and bias of our model

We have relatively limited data, and our model's MAE reaches 60.65 million. The model and the outcome that we generate may not be that accurate. If we use it as guidance to make investment decisions, it could let us suffer from loss. Also, our data is biased in several aspects. For example, it mainly contains English movies, so its usefulness to predict non-English movies' profit isn't going to be accurate and probably blatantly wrong.

2. The incomplete selection of variables

Being restricted by the dataset, we could only consider limited variables that may impact the profit. But there could be more factors in the actual film market, such as the movie release dates (in holidays or not), the quality of the trailers, etc. We can omit some significant factors.

3. The possibility of missing some unusual but successful movies

As we use the historical data to build the models, we could only find the movies that have the similar success as the old ones. Some independent movies may not have famous directors and actors, and not belong to a welcomed genre, but they have their unique successful factors, which our model may fail to identify them.

4. Ethical Considerations

Our client seeks to invest in the most profitable films, however while these films will likely to be successful in the box office, they might have messages that don't resonate well with everyone, some films might have violent and provocative themes that aren't appropriate themes or might have social constructs and movements that fester through social media that could impact mental health of teenagers and young adults. As a result, our clients' investment might have far reaching implications which will be important to keep in mind.

Possible ways to mitigate the problems and risks:

We could continuously add more data and variables to training our random forest model better to increase its accuracy. Also, as models may suffer from the inability to identify some unique successful movies, we could use our model as a reference, combining with some expert's views in the media and entertainment industry, to better guide our investment. As for the ethical problems, we might need to balance our profitability with our possible social influence.