PREDICTING SUCCESS: BOX OFFICE DECISIONS

Author: Grace Yin, Kevin Asselin Alfaro, Hanjin Ying, Madhu Nathani

It was December 12, 2022, and the manager of Imagine Cinemas in London, Jacob McKeen, sat down to evaluate the box office revenue and resulting profits for Imagine from the past month. In recent years, the film industry has seen a drastic shift away from traditional box office movie theatres to content streaming services. This shift has led to a new set of challenges for movie industry executives in predicting what films will be a success. Lower viewership in the past few months has decreased revenues and subsequently profits for the theatre. McKeen and the rest of the management team needed to better predict which movies audiences would most likely want to see in theatre to maximize profits and make up for these losses.

Given that Imagine Cinemas is a small theatre with only one screen, McKeen is faced with a choice between four movies to license for release in January 2023. McKeen has found a dataset on historical movie metrics from Kaggle, a website containing open-source code and real-world datasets, and will use the knowledge from his university data class to help with this task.

IMAGINE CINEMAS

Located in London, Ontario, Imagine Cinemas is a small movie theatre located in the City Plaza Shopping Centre. While the theatre has eight screens, only one is currently running due to a decrease in viewership since the beginning of the COVID-19 pandemic. Each screen could accommodate approximately 150-200 people prior to the pandemic, and closer to 150 during the pandemic to account for some seats being allocated for social distancing purposes[1]. Therefore, it is highly important for the seats to fill up for the single screen as it is Imagine Cinemas' only stream of revenue.

THE FILM INDUSTRY

The film industry comprises the production, distribution, and exhibition stages[2]. At the distribution level is where movies are made available to the general public through movie theatres, television channels, or at home-viewing through physical media, video-on-demand platforms like Netflix, or download.

The movie licensing or buying process follows a typical timeline. When film production is complete, the studio sends it to a distribution company to come with licensing agreements. The distributors are in control of the marketing for the film and will set up a film screen for prospective buyers like cinemas and film festivals. Buyers will preview the movie and decide on which ones they will leave, and then negotiate with distributors on terms of the lease agreements, including how many weeks and when they will start showing the movie on the screen. The film is sent to the buyer a few days prior to this initial showing date to account for any possible errors with the file.

[1] https://measuringstuff.com/how-many-seats-are-in-a-movie-theater/
[2] https://en.wikipedia.org/wiki/Film_distribution

OPERATING STRUCTURE OF THEATRES

Movie theatres license movies from distributors for a rental fee. This rental fee is around 60% of the ticket prices charged to the end consumer, and profits for the theatre come from the remainder. Movie theatres generate their revenues primarily through ticket sales and concessions where audiences can purchase food and drinks for the viewing. Movie theatre ticket prices vary based on the time of day, day of the week, type of movie, and the movie itself – partially dependent on the rental fee. For example, ticket prices are higher for 3D movies than 2D movies, and higher for evening showtimes over afternoon showtimes. Movie theatres also generate revenue from concessions like popcorn, drinks, and candy. Concession prices are typically higher than at other stores due to the convenience of having it available at the theatre, so profits come from the marked-up prices. Movie theatres also generate revenue from advertising, which can include both pre-movie ads as well as banners and posters throughout the theatre. Having a good sense of movie success allows theatres to better price movie tickets to maximize profits.

FACTORS OF MOVIE SUCCESS

Movie success is inherently unpredictable. There are some factors related to the production of the movie itself that can influence movie success: whether the movie is a stand-alone or part of a series, the popularity and quality of its director and actors, the movie's budget, its genre, and pre-release critic ratings. However, there are additional factors that only emerge post-release like word-of-mouth, review platform ratings, and general audience sentiment that play an equally large if not larger role in a movie's success. This was seen with *Minions: The Rise of Gru*, where a group of young viewers started a trend of watching the movie in full suits, causing incredibly high viewership for the movie from an unintended target market.

CONSUMER SHIFT TO CONTENT STREAMING

With the advent of content streaming services, such as Netflix and Amazon Prime Video, consumers are increasingly turning away from traditional movie theatres. According to the Motion Picture Association of America (MPAA), the number of frequent moviegoers (those who attend at least once a month) has decreased from 25% in 2012 to 17% in 2018[3]. This shift in consumer behaviour has had a significant impact on the film industry. According to the MPAA, the global film industry saw a 2.6% decrease in revenue in 2018, with a majority of the decline being attributed to the shift away from movie theatres to content streaming services[4].

As evidenced, audiences were already shifting to content streaming prior to the COVID-19 pandemic. However, COVID-19 has accelerated this shift, with 41% of consumers rarely going to see movies in theatres anymore and 18% not going to theatres at all[5].

With the rise of streaming, decisions have to be made on whether a movie should be primarily released on streaming platforms or in theatre. Some of the movies that are going to theatres are not commercial enough, but some movies sent directly to streaming might be better suited to

[3] https://www.motionpictures.org/wp-content/uploads/2017/03/MPAA-Theatrical-Market-Statistics-2016_Final-1.pdf
[4] Ibid.
[5] https://www.theguardian.com/film/2022/dec/02/box-office-blues-films-bombing-hollywood

theatres. Some movies also do much better than others. For example, Knives Out's sequel, Glass Onion, did extremely well in theatres over Thanksgiving, while other movies did poorly.

THE MOVIE CHOICES

A Man Called Otto
Otto is a grump who's given up on life following the loss of his wife and wants to end it all. The movie's runtime is 2h 6m, censorship rating is PG, and main genre is comedy. The director is Marc Forster, who is labelled with a C in the director.csv dataset. The key leading actor is Tom Hanks, who is labelled with an A in the actor.csv dataset. The film had its first viewing at the Academy Museum in early December and received high ratings from voters and critics[6]. Many believed this movie will be Tom Hanks' golden ticket to being nominated for the Best Male Actor Oscar awards next year.

Alice, Darling
A young woman trapped in an abusive relationship becomes the unwitting participant in an intervention staged by her two closest friends. The movie's runtime is 1h 30m, censorship rating is 18, and main genre is drama. The director is Mary Nighy, who is labelled with a C in the director.csv dataset. The key leading actor is Anna Kendrick, who is labelled with a B in the actor.csv dataset. The movie released its first trailer in early December and is expected to play at the Palm Springs Film Festival in January. Kendrick stated that she drew on her personal toxic relationship experiences to channel out her character in the drama[7]. She believes the movie will be emotionally striking for the audience base.

Fear
A much needed getaway and a celebration weekend turns into a nightmare due to the contagious airborne threat. The movie's runtime is 1h 40m, censorship rating is A, and main genre is horror. The director is Deon Taylor, who is labelled with an A in the director.csv dataset. The key leading actor is Joseph Sikora, who is labelled with an A in the actor.csv dataset. The movie is expected to be released in late January 2023[8]. However, due to the main genre being horror, it is important to consider that this genre is not for everyone and could only attract a smaller and niche audience base.

M3GAN
A robotics engineer at a toy company builds a life-like doll that begins to take on a life of its own. The movie's runtime is 1h 50m, censorship rating is PG, and main genre is horror. The director is Gerard Johnstone, who is labelled with a B in the director.csv dataset. The key leading actor is Allison Williams, who is labelled with a B in the actor.csv dataset. The movie is expecting strong performance in 2023. Additionally, the Titular AI doll in the film has attracted a lot of attention through her viral dance scene in trailers[9]. This drove high social media

[6] https://variety.com/2022/awards/awards/a-man-called-otto-tom-hanks-oscars-best-actor-1235453568/
[7] https://variety.com/2022/awards/news/alice-darling-trailer-anna-kendrick-1235451636/
[8] https://moviesandmania.com/2022/11/23/fear-movie-film-psychological-horror-2023-trailer-release-news/
[9]https://www.boxofficepro.com/long-range-box-office-forecast-blumhouses-m3gan-tracking-for-a-solid-2023-lead-off-plus-the-latest-avatar-the-way-of-water-projections/

interactions on platforms such as Tik Tok. However, many critics say the viewership of M3GAN will be negatively impacted by the release of Avatar: The Way of Water.

CONCLUSION

As the movie industry continues to shift away from traditional movie theatres to content streaming services, movie industry executives need to be able to better predict which movies will be a success in theatre specifically. Additionally, movies make a third of their entire domestic box office during the opening weekend. Opening week box office is increasingly vital due to its power in creating buzz and drawing more movie-goers to theatres in the coming weeks. By analyzing data on historical movie revenues with stated factors like genre, director, actors, and more, models can be created to predict the success of upcoming movies to some extent. McKeen and his team can use this to help inform their decision-making about which movie to license. With January soon approaching, it is vital for McKeen and his team to make a decision quickly on which movie to show in order to allocate enough time to negotiate licensing agreements with distributors.

SOLUTION STATEMENT

THE DATASETS

McKeen found a dataset titled “IMDB 5000 Movie Dataset” on Kaggle, which comprises data from 2000 historically released movies. It had columns ‘Movie_Title’, ‘Year’, ‘Director’, ‘Actors’, ‘Rating’, ‘Runtime’, ‘Censor’, ‘main_genre’ and ‘Total_Gross’, which was the variable McKeen wanted to predict.

Additionally, McKeen also found two datasets: ‘actor.csv’ and ‘director.csv’. Both datasets contained the name of directors and actors, as well as their corresponding ratings. McKeen chose them because ‘IMDB.csv’ did not contain metrics such as ratings that could access directors’ or actors’ performance.

Finally, McKeen compiled four movies he wanted to predict in ‘Movies.csv’. This contained four movies to be shown in 2023, and he wanted to apply them to the three models to see their performance and decide which to license.

DATA CLEANING PROCESS

After importing the datasets into the R workplace, McKeen took a closer look and found that there were nine columns in the main dataset ‘IMDB.csv’. The first step was to decide which columns to use as independent variables. Since ‘Movie_Title’ and ‘Year’ had no clear causal relationships with a movie’s gross revenue, he removed these two columns. For ‘Total_Gross’, which was the dependent variable, McKeen deleted the N/A rows, and for the rest of the valid rows, he converted every character ‘M’ into zeros so that all data were numerical. He did similar work on columns ‘Censor’ and ‘Directors’. For column ‘Actors’, McKeen realized there were multiple names within each cell, disabling the model training process. Therefore, McKeen split the ‘Actors’ column into four separate columns, each representing one actor. Out of simplicity, he removed the last three actors and left only the first leading actor.

Lastly, McKeen realized a potential flaw in his current dataset: director and actor names were meaningless in determining a movie’s gross. Therefore, he downloaded two more datasets containing director/actor names and corresponding ratings in three levels: A, B and C. He then linked the names with the ratings and added ‘DirectorRating’ and ‘ActorRating’ to the model.
In summary, after data cleaning, the columns that will be used for predicting ‘Total_Gross’ were: ‘DirectorRating’, ‘ActorRating’, ‘Rating’, ‘Runtime’, ‘Censor’ and ‘main_genre’.

MODEL SELECTION DECISION

Now McKeen had a clear view of what his cleaned dataset looked like, he wanted to decide what models to adopt for movie predictions, given that the types of independent variables were mixed: ‘Rating’ and ‘Runtime’ were numerical, while ‘Censor’, ‘main_genre’, ‘DirectorRating’ and ‘ActorRating’ were categorical. After preliminary research, McKeen found that the decision tree was a good match, since every split in a tree is based on a feature: If the feature is categorical, the split is done with the elements belonging to a particular class; If the feature is continuous, the split is done with the elements higher than a threshold.

To avoid the decision tree's tendency for overfitting, McKeen also wanted to try the random forest. By using the voting approach, this model is applicable to more generalized data, and pruning is not required.

As McKeen expected, the decision tree and random forest would return him numerical predictions for movie gross revenue. McKeen, therefore, wondered if there was a supplementary model that could give him categorical outputs, namely A, B, C and D. Although trees and forests could handle such classification problem, McKeen decided to use KNN, since KNN offered a non-parametric method that doesn't learn an explicit mapping function during training.

In summary, McKeen decided to try decision tree, random forest and KNN. All of these three methods were supervised learning, since McKeen had access to true labels. In this way, for any future data, he would first feed that data into KNN to see what bucket it falls in. If the movie obtained A or B, or any performance threshold McKeen desired to set, he would then take a closer look by testing it on a decision tree and random forest.

Model 1: Decision Tree

To construct a decision tree, McKeen selected 'DirectorRating', 'ActorRating', 'Rating', 'Runtime', 'Censor' and 'main_genre'. He wanted to see how those multiple independent variables impact 'Total_Gross'. McKeen recalled that leaving the tree depth free might lead to overfitting. Therefore, he visualized the relative error as a function of tree size and chose a cp. As shown in Exhibit 1, a cp threshold at around 0.024 would be the optimal number: if cp was larger than this number, a decrease in relative error was marginal.

Now McKeen looked at the tree and reached the following findings: as shown in Exhibit 2, 'Runtime' was the root node, representing the starting point. 'Runtime', therefore, had the highest hierarchy in the tree: for a movie to have a high box office, the runtime first needs to be reasonable and fall into the average interval that viewers accept. The second determinant factor was 'Rating': before a movie was shown to the public, it usually had to go through an inside viewing event. This was where professional movie critics came and rated, given their first impressions. Additionally, 'main_genre' and 'ActorRating' were the other two important factors in determining the box office. People only choose to see a movie if its genre fits their preference or if it has their favourite actors.

In summary, the decision tree helped McKeen filter out those four leading independent variables that had the most weight in determining a movie's gross revenue. Using this model, he would input any future movie he wanted to predict and receive a numerical gross revenue.

Model 2: Random Forest

McKeen also wanted to use a random forest, which was more accurate and robust than a single tree, since all trees within the forest did not learn the same. McKeen chose 'DirectorRating', 'ActorRating', 'Rating', 'Runtime', 'Censor' and 'main_genre' to build the forest. To avoid a large training time when having big data, McKeen first plotted the error (MSE) as a function of the number of trees, as shown in Exhibit 3, and limited the number of trees to 100. Different from the decision tree, McKeen was unable to visualize and gain intuition simply from looking at the graph. This was a tradeoff McKeen made when building a random forest.

Model 3: KNN
Mckeen could also take a different approach in predicting what makes movies successful, so he decided to try a classification method - KNN. He started by, splitting movies by percentile of their gross revenue into four classes: A, B, C, and D: with A being the highest grossing and D being the lowest grossing in order to create classes. Moreover, since Mckeen had previously used A, B, and C systems to rank directors and actors in films, he also needed to transform the data into numerical 3,2,1 rankings respectively in order to utilize a classification model.

Mckeen then normalized his data set to make sure the scale of all features was the same. He then created a testing and training set at a sampling rate of 0.8, followed by creating the training features, training labels, and testing features. After all of these steps were taken, Mckeen applied his KNN model to the testing set to compare to the actual values.

COMPARISON OF ERROR METRICS

After all 3 models had been built, it was time for Mckeen to compare their accuracies to determine which one(s) he should use to make his decision.

Model 1: Decision Tree
The decision tree had an RMSE of 51,972,790 but will vary depending on the random sample that is taken.

Model 2: Random Forest
The random forest had an RMSE of 55,442,910 but will also vary depending on the random sample that is taken.

Model 3: KNN
The KNN error measure is different as it is a classification method but it had an average misclassification rate of 44.39% percent after 5-fold cross-validation.

According to the RMSE measures, the decision tree better fits the data than the random forest. However, these values are very close and depending on the random sample that is taken, the RMSE could be lower for either model. Therefore when comparing these models, Mckeen can use both to make predictions are they provide similar error metrics. One is not materially better than the other.

When interpreting the KNN, the 44.39% misclassification rate initially seems very high, however as we have four classes, the random misclassification rate would be 75% (¾). Therefore, the misclassification rate is significantly better than a random classification. Additionally, since our classes are based on continuous data, the magnitude of misclassification also matters - not all misclassifications are equal, some are more incorrect than others. As we can see by the confusion matrix (Exhibit 4), the misclassifications are usually close to the correct class, this means that most misclassifications are small misclassifications.

PREDICTION RESULTS FOR FOUR MOVIES

McKeen collected four movies on IMDB that will be launched in 2023 and stored those values in the dataset ‘Movies.csv’. There are seven columns, namely ‘Movie_Title’, ‘Rating’, ‘Runtime’, ‘DirectorRating’, ‘ActorRating’, ‘Censor’ and ‘main_genre’.

For the decision tree and random forest, he simply input the ‘Movies.csv’ and received the predicted gross revenue. Since numerical inputs are required for KNN, McKeen first excluded ‘Censor’ and ‘main_genre’ from the dataset (since they cannot change to numerical values), then he transformed 'DirectorRating' and 'ActorRating' from character to numerical (based on the assumption that 'A'=3, 'B'=2, 'C'=1).

As shown in Exhibit 5, KNN assigned ‘A’, ‘D’, ‘D’ and ‘A’ to the four movies. This indicated that movies ‘A Man Called Otto’ and ‘M3GAN’ were the most promising to invest in. On the other hand, the random forest output 62970571, 57890649, 59534280 and 41199848, indicating that ‘A Man Called Otto’ and ‘Fear’ ranked at the top (Exhibit 6).

Except for the fact that ‘A Man Called Otto’ was consistent among the three models (always ranking at the top), McKeen found an inconsistency in the predicted output among the three models, especially between KNN and random forest. This was because different columns were selected for building different models, so McKeen decided not to overly depend on one single model as his forecasting tool, but rather a combination of all three, as well as the qualitative factors which were not reflected in the models.

INSIGHTS GAINED

After the analysis, students are expected to choose ‘A Man Called Auto’ to screen, however, students may also choose ‘Alice, Darling’ based on qualitative factors inside the case. This emphasizes the 3 main learning objectives for this case:

1. How to manipulate dirty data into something usable
2. How to choose, build, and asses models
3. How models impact real-world decisions

Students must navigate the dirty data through many methods before making models as in real situations, data is never perfect. Choosing models is also key as students may or may not opt for a classification or predictive type, there is no one correct answer, and students can reach similar conclusions in many ways. Lastly, no model will ever be perfect thus when making decisions students must also consider qualitative factors and understand that models are supplementary to large decisions. This further emphasizes the idea that there is no one correct answer as long as students justify themselves.

Overall, Students are learning how to leverage data from start to finish in real-life situations to allow them to make better (not perfect) decisions.

Exhibits:

Exhibit 1: Decision Tree - relative error as a function of cp.

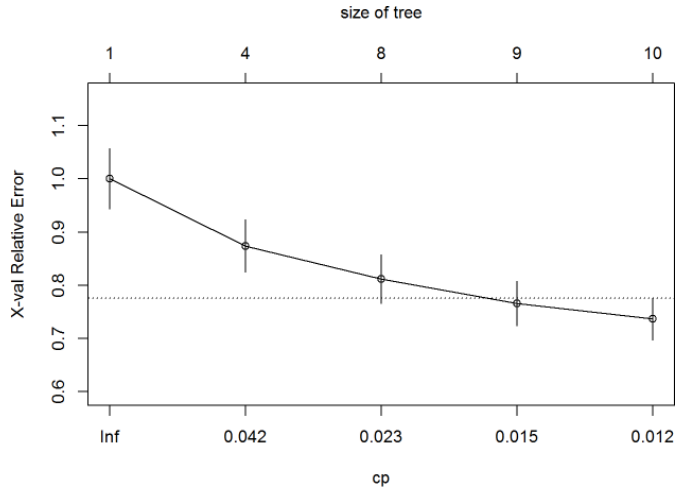

Exhibit 2: Decision Tree.

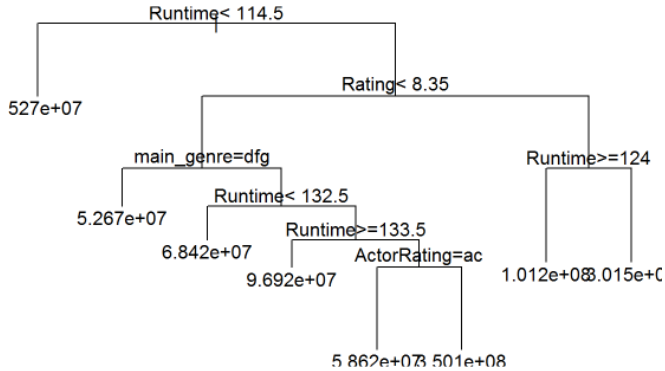

Exhibit 3: Random Forest - error as a function of the number of trees.

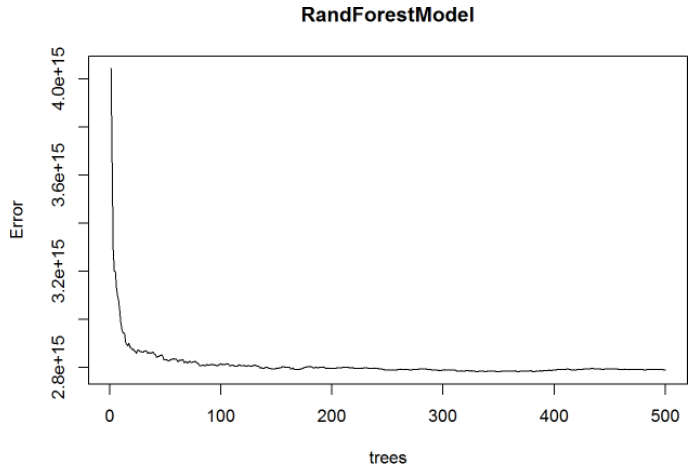

Exhibit 4: KNN - confusion matrix.

```
table(testingSet$GrossRating,predictedLabels)
```

```
##    predictedLabels
##       A   B   C   D
##   A 107  26  24  11
##   B  23  84  30  18
##   C  29  25  74  27
##   D  30  27  34  91
```

Exhibit 5: Prediction for four movies - KNN.

```
prediction_KNN = knn(trainingfeatures,movie_KNN,traininglabels,k=3)
prediction_KNN
```

```
## [1] A D D A
## Levels: A B C D
```

Exhibit 6: Prediction for four movies - Random Forest.

```
prediction_forest <- predict(RandForestModel, movie_predict)
prediction_forest
```

```
##        1        2        3        4
## 63940702 68552429 61226339 40172513
```