Fuentes, Paolo
Pecundo, Allan Magno
Villegas, Mylene

CSCI 271-A

MIDTERM REPORT

I. Introduction & Significance of the Study

Background of the study

The film industry is a major private sector employer. According to a November 2020 study by the Motion Picture Association (2020), 2.5 million jobs are generated by this industry in the USA [1]. In generating jobs, employment opportunities are not only directly provided under film production companies, but this is also done indirectly by supporting local businesses in the filming location. Establishments such as hotels, restaurants, gas stations, dry cleaners, hardware stores and area labor. Local communities reap economic benefits whenever film productions come to town. Thus, the production and distribution of motion pictures and television programs is one of the nation's most valuable cultural and economic resources.

This success at the box office can be attributed to numerous factors. They can be a major actor's star power, genre, runtime, and release date, among others. The abundance of such factors makes manual revenue prediction process difficult. However, due to innovations in technology, volumes of data can be efficiently processed and modeled. Hence the task of predicting gross revenue of a movie can be simplified with the help of modern computing power and the historical data available online in the form of movie datasets.

Specifically, previous studies that used analytics have shown that there is a correlation between production cost or budget and the success of the film in terms of its box office revenue. A study by Pangarker and Smit (2013) attributes production cost as one of the drivers of box office revenue [2]. Similarly, another study conducted by Follows (2019) yielded results showing that there is a Pearson correlation of 0.744 between budget and domestic gross, thereby denoting strong positive correlation [3].

However, this is not always the case as correlation does not immediately mean causation. Vogel (2011) mentioned that " of any ten major theatrical films produced, on the average, six or seven may be broadly characterized as unprofitable and one might break even" [4]. These numbers suggest that the movie industry is one of the riskiest industries. It is because of these high risks in producing movies that makes planning and predicting box office revenues highly essential.

Thus, due to this unpredictability, cinematic success becomes highly contingent on making sound decisions. For instance, should a certain actor be cast over the other? What genre should the movie be? Should the trailers, posters, and advertisements quote some well-known critics, emphasize that its director is an Oscar nominee? Due to this uncertainty,

it is imperative to accurately determine the factors that influence the audiences of a movie, and their impacts assessed so that the risk associated with movie production is minimized.

Significance of the study

The financial success of a film creates an impetus for additional films but a financial failure usually results in an abrupt end to the franchise. Typically called a box-office flop, this phenomenon fails to make as much money at the box office as they cost to produce and release.

The stakeholders involved in this problem are the movie company, the cast, and the crew which includes the director, producer, and screenwriter, among others. They are the primary beneficiaries if their movie tops the box office and exceeds expectations. Likewise, they are also at a loss financially if the movie performs underwhelmingly at the box office, and if the movie franchise is discontinued. Because of this, their target return on investment (ROI) would not be achieved. Therefore, formulating a model that can determine which of the features significantly contribute to the revenue of the movie.

The study aims to address the following questions:

- Can we predict whether a movie will be "successful" before it is released?
- What factors contribute to the success of the movie?

II. Methodology

To address the following questions, the study will use a Machine Learning approach to predicting the movies success, subject to a predefined criteria for success. Factors that contribute to the success of the movie will be determined from the insights derived from the possible parameters of the model coupled with exploratory data analysis as validation.

Pre-processing

First, the dataset was inspected for any missing or atypical observations.. Upon inspecting the dataset, there were a few observations flagged as having atypical values.

	budget	id	popularity	revenue	runtime	vote_average
count	4.803000e+03	4803.000000	4803.000000	4.803000e+03	4801.000000	4803.000000
mean	2.904504e+07	57165.484281	21.492301	8.226064e+07	106.875859	6.092172
std	4.072239e+07	88694.614033	31.816650	1.628571e+08	22.611935	1.194612
min	0.000000e+00	5.000000	0.000000	0.000000e+00	0.000000	0.000000
25%	7.900000e+05	9014.500000	4.668070	0.000000e+00	94.000000	5.600000
50%	1.500000e+07	14629.000000	12.921594	1.917000e+07	103.000000	6.200000
75%	4.000000e+07	58610.500000	28.313505	9.291719e+07	118.000000	6.800000
max	3.800000e+08	459488.000000	875.581305	2.787965e+09	338.000000	10.000000

Figure 1: Summary Statistics of TMDB 5000 Movies Data (Numeric Features)

From the summary statistics of the numeric features seen on Figure 1, there are some movies identified as having 0 dollars in budget or 0 dollars in revenue. These observations were identified to be atypical and were removed from the dataset.

After these observations were removed from the dataset, checks were done on the remaining categorical features. Under the "status" feature, the remaining observations were classified as either "Released" or "Post Production". Given that the objective is to help stakeholders in the film industry identify whether the movie they release will be successful or not, the observations retained were only movies that were considered "Released". After all the inspection and cleaning of atypical observations, the final dataset was reduced from 4,803 to 3,228 observations; a reduction of 33% from the original dataset).

Aside from the removal of atypical features, some features were engineered and transformed so that they can be utilized by the model. First, the "Year" feature was created from the "Released_Date" feature which consisted of the specific date of when the movie was released represented as a text string. The "Year" feature will represent the numeric year in which the movie was released; a feature that can help predict movie success in relation to the other existing features under the assumption that the context of the year in which the movie was produced is an indicator of success.

Next, the Genres column was unpacked as categorical features with each genre representing a column with 0s and 1s representing whether the movie belonged in that specific genre; into a format that can be used by different models. This was done given that most movies would have more than one genre, hence having one column representing the categorical label of the genre was not possible. Due to movies possibly containing a multiple number of genres, only the more common genres (top 10 in terms of % incidence of observation) that the movies belonged to were maintained while movies that did not belong to any of the common genres were classified under "Others". The final genres represented as separate binary columns were as follows: Action, Adventure, Comedy, Crime, Drama, Family, Horror, Romance, Sci-fi, Thriller, and Others.

Next, the target feature was engineered from the budget and revenue column with the goal of being the measure of whether a movie was successful or not. The "success" made was calculated as a cutoff of the return on investment a movie made. If the movie made more than double its investment, it would be considered a successful movie, otherwise it would not be considered successful.

$$Success = \begin{cases} 1, if\ \dfrac{revenue - budget}{budget} > 1 \\ 0, if\ \dfrac{revenue - budget}{budget} \leq 1 \end{cases}$$

Figure 2. Calculation of "success" feature

Lastly, the final features to be used in building the model were identified to be features that were either:

- Features that could be known before a movie was shown
- Features that were within control of the stakeholders of the movie (e.g., director, producer, writer, etc)
- Features that could be pre-processed for model building within the time constraint of the exercise

With that, the final set of features used were as follows:

Input Features: budget, runtime, year, genre_Action, genre_Adventure, genre_Comedy, genre_Crime, genre_Drama, genre_Family, genre_Horror, genre_Romance, genre_ScienceFiction, genre_Thriller, genre_Others

Target (Output) Feature: success

Model Building

The data was first split into a training set and test set, comprising 80% and 20% of the entire dataset respectively. The model was trained on the training set and the performance metrics were measured against the model's prediction of the test set.

To be able to predict whether a movie would be "successful", there were 3 models that were fitted to the data: (1) Logistic Regression, (2) Decision Tree, and (3) Random Forest.

Each model will have its own distinct way of being able to fit the data to come up with predictions of the target variable based on the provided input variables.

A. Logistic Regression

The Logistic Regression model fits the model using the logistic function.

$$y = \frac{1}{1 + e^{-(mX + b)}}$$

Figure 3: Logistic Function

The model takes the input features vector *X*, and applies a set of weights or weight vector *m*, followed by some constant *b* similar to the linear regression model; which is part of the logistic function. The output of the logistic function, considering the inputs of the weights and input features, will be a value between 0 and 1, representing the probability of the observation belonging to the positive class. To be able to utilize the logistic function for classification, which requires categorical values, there is a threshold of 0.5 applied wherein outputs greater than 0.5 are classified under the positive or 1 class, while the remaining observations are classified under the negative or 0 class [5] .

The model tries to find the best set of weights X and constant b, such that the predicted output classes of the logistic function will be close to the accuracy of the observed classes on the data it was trained on.

B. Decision Tree

The Decision Tree algorithm for classification creates a model using a series of successive rules (e.g, similar to if-else statements) in order to predict which class an observation belongs to [5]. The model uses recursive partitioning in coming up with the different rules or conditions to classify the observation. First, the tree will identify a certain predictor which will be used to partition the observations into classes. Second, within that predictor, a specific cut-off value (e.g, condition at a certain feature) will be identified at which the feature can segregate the observations into the two different classes; with the goal of each group having the most homogeneous or similar set of classes within each group. The algorithm repeats this for other features with the algorithm stopping if it meets a certain criteria. Usual criteria that are considered for stopping the tree are (1) the depth of the tree built (i.e., how many successive sets of rules or conditions have been identified for classification), (2) whether adding a new feature rule will improve the homogeneity of the classes in each condition group, or (3) whether the successive groups are already very small in count [6]. The final model will be a tree consisting of a successive set of rules with features and corresponding conditions the observations need to meet for these sets of features.

Summary of Algorithm [5]:

- Step 1: Start at root node (all observations) and identify a single feature *d* among all features D to partition data
- Step 2: Identify split value *s* at feature *d* in which to partition data; optimal split based on selected criteria
- Step 3: Repeat Step 1 (Feature identification) and 2 (Split value identification at feature) until stopping criteria is met;
- Step 4: Predict future observations based on conditions of the final tree

C. Random Forest

The last model that was tested was the Random Forest for classification. The Random Forest model follows a similar algorithm to the decision tree with some specific differences:

- Unlike the Decision Tree, Random Forest will build multiple Decision Trees with each tree coming up with its own prediction for the class

- What differentiates that different trees build are the fact that each tree will only use a randomly subsampled subset of both the input features and observation that are available, whereas the Decision Tree may use all of the features.

These two aspects of the Random Forest make it distinct and usually more powerful than the Decision Tree as it allows the model to generalize better as it takes only a subset of features making the model possibly less complex while scaling up the number of trees that will predict the class. The final class that will be identified for the observation will be the class in which the majority of the trees built will classify it under.

In general, the algorithm behind the Random Forest can be summarized as the following:

Summary of Algorithm [5]:

- Step 1: Take a bootstrap (with replacement) sample of the observations
- Step 2: For the first split, take only sample *p* of the total P input features at random
- Step 3: For each feature, identify the optimal value *s* at which to split the features and group the data; optimal value will be based on criteria such as resulting homogeneity (e.g., how pure are the classes within each group)

- Step 4: Repeat Steps 2 to 3 until the entire tree is grown or fulfills a stopping criteria.
- Step 5: Repeat the entire process of building trees (Steps 2 to 4) until a defined *m* number of trees are built.

After the data was fitted using the 3 types of models mentioned, the accuracy of the model was validated against a test set with observations having the same level of detail that the model has never seen (e.g., trained on). The accuracy of this validation set measures how well the model generalizes its predictions.

Based on the 3 models, the performance metrics were as follows:

Model Type	Precision	Recall	Accuracy
Logistic Regression	0.54	1.00	0.54

Decision Tree (max depth = 5)	0.69	0.55	0.62
Random Forest	0.61	0.75	0.60

Figure 4: Summary of Model Performance

The Logistic Regression model yielded a low accuracy rate and high recall rate because the model predicted almost all the classes to be positive. It might be predict it in this manner if the features individually have no direct linear relationship in determining the likelihood of a movie being successful or not

The Decision Tree model yielded a balanced prediction wherein the accuracy is still higher than the logistic regression model despite it's recall being lower given the relatively lower positive values predicted but a high precision given that there are less false positives predicted in the model compared to Logistic Regression where all cases were predicted as positive, even the actual negative cases.

The Random Forest model also yielded a more balanced prediction than the Logistic Regression although its accuracy is slightly lower than the Decision Tree tuned at max_depth = 5. The precision of the model is also lower which indicates that it did not perform as well as the Decision Tree model in capturing the correct positives as it predicted more false positives than the Decision Tree model. This is also seen in how the Random Forest model's recall is higher than the Decision Tree model's recall.

Both tree models seem to perform better than the logistic regression model as it might be able to better capture complex relationships between the input features and the output (target) variable. Based on these models, the final model used was the Decision Tree model given that it still gave a relatively higher overall accuracy and relatively higher prediction balance than the other two models while still maintaining a level of simplicity wherein insights can be derived from the model via the interpretation of the singular tree.

III. Insights

To be able to come up with actionable insights from the study, we analyzed the feature importance and the tree plot of the trained Decision Tree Model.

As explained in simple terms in an article by Loaiza (2020), Gini Impurity is one of the measures used in a decision tree model to decide the best split. It indicates the probability of incorrectly assigning an observation [7]. When training a decision tree, we can compute how much each feature from the dataset contributes to decreasing the weighted impurity. The greater the contribution, the more important the feature is. The feature importance function as expounded in the documentation from Scikit learn is based on this particular logic [8].

As shown below, here are the variables and their corresponding contributions for our classification model.

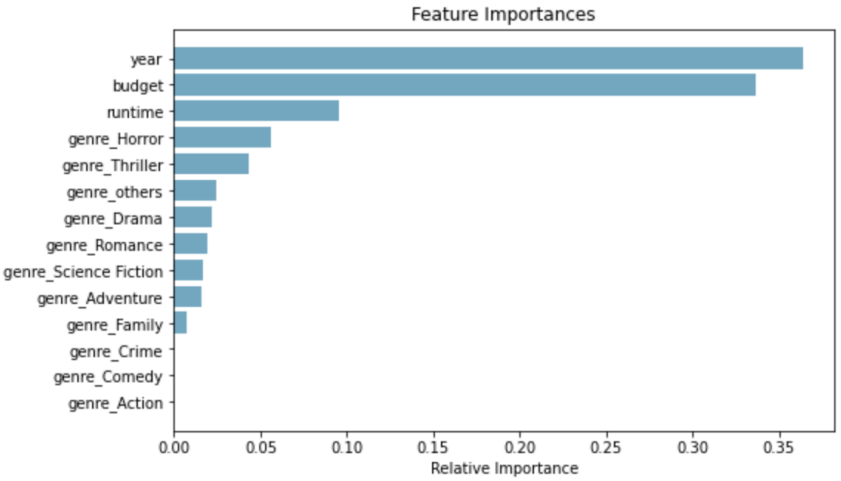

Figure 5: Relative Importance of Features from Decision Tree

In addition to checking the feature importances, we also analyzed the insights that can be derived from checking the decision paths and the end nodes from the tree plot of the trained model. This particular exercise will be helpful in identifying what are the characteristics of the movie that the model will predict as successful or not.

For the initial checking, we analyzed a decision path and the corresponding end node from the model where the genre is drama and budget is greater than $145M dollars. Plotting the end node in a box plot shows that dramas with budgets greater than $145M are tagged successful. This particular insight gives our movie producers an actionable item which is to invest big when it comes to drama films.

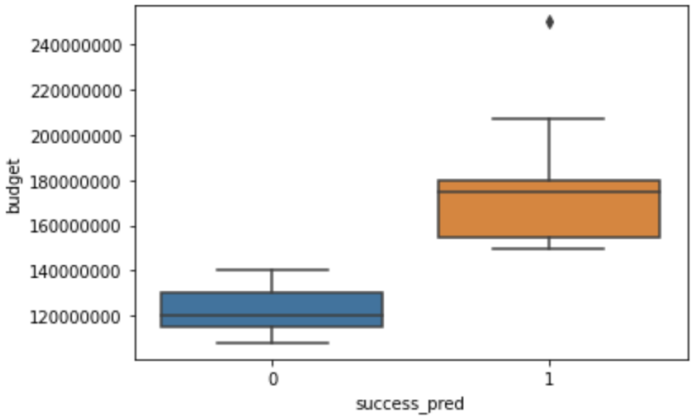

Figure 6: Showing Budget Distributions Among Drama Films By Success Class

For non-drama films, the boxplot of the end node shows that those that are classified as successful are at least 79 minutes long.

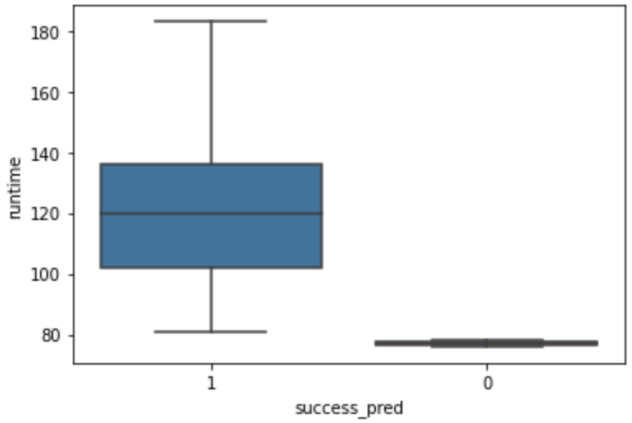

Figure 7: Showing Runtime Distributions Among Non-Drama Films By Success Class

For non-horror films, we were able to see from one of the decision nodes that even those with less budget can still be successful.

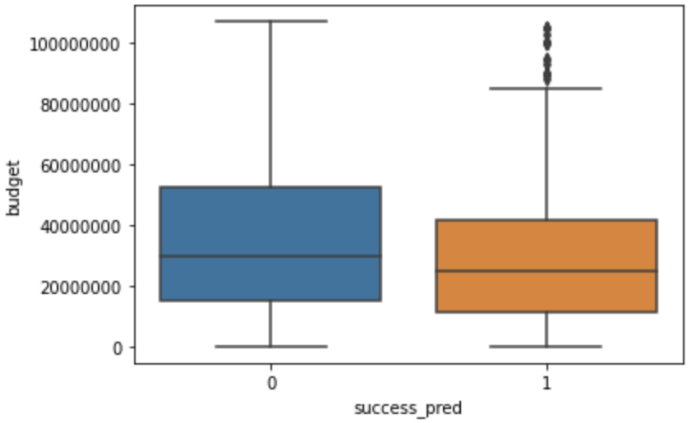

Figure 8: Showing Budget Distributions Among Non-Horror Films By Success Class

Lastly, upon checking an end node where the genre is horror and non-romance, most of the successful movies falling in this category have shorter run times. This can be an actionable item for movie producers to keep non-horror films short.

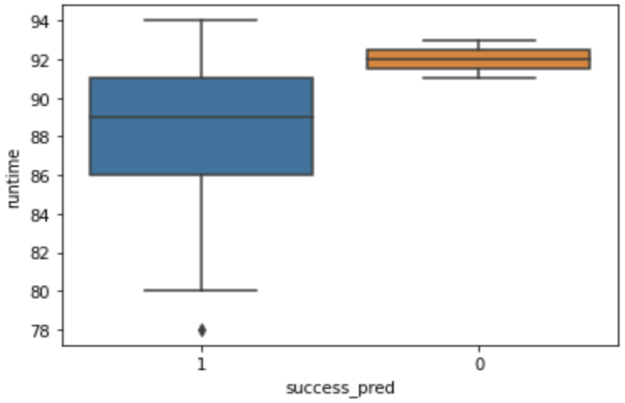

Figure 9: Showing Runtime Distributions Among Horror + Non-Romance Films By Success Class

Analyzing the decision paths and the corresponding end nodes of the trained model helped us in understanding what are the characteristics of the movie that will most likely be tagged as successful by the model. From this, we were able to produce actionable insights that could help the movie producers in their decision making.

IV. Conclusion

Predicting the success of a movie is a significant task in managing the risks when it comes to filmmaking. After conducting the study, we were able to come up with a model that could predict if a movie will be successful or not using the limited variables known before the release of the movie. Some of the important variables from the available data are year, budget, and runtime. From the results of the splitting done by the trained Decision Tree model, we were able to identify the characteristics of the movie that are most likely to be predicted as successful. These results and insights can be used by the filmmakers for their decision making before the release of the movie.

V. References

[1] "The American Motion Picture and Television Industry: Creating Jobs, Trading around the World." Motion Picture Association, 2 June 2021, https://www.motionpictures.org/research-docs/the-american-motion-picture-and-television-industry-creating-jobs-trading-around-the-world-2/

[2] Pangarker, N.A. & Smit, Eon. (2013). The determinants of box office performance in the

film industry revisited. South African Journal of Business Management. 44. 47-58. 10.4102/sajbm.v44i3.162.

[3] Follows, Stephen. “Is a Movie's Box Office Gross Connected to Its Budget?” Stephen Follows Film Data and Education, 6 July 2021, https://stephenfollows.com/is-a-movies-box-office-gross-connected-to-its-budget/.

[4] Vogel, H. (2011). Entertainment Industry Economics (8th ed.). New York: Cambridge University Press.

[5] Bruce, Peter C., and Andrew Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. O'Reilly Media, 2018.

[6] “Sklearn.tree.decisiontreeclassifier.” *Scikit-Learn: Machine Learning in Python*, https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[7]Loaiza, Steven. “Gini Impurity Measure.” *Medium*, Towards Data Science, 23 Mar. 2020, https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33

[8]“Feature Importances with a Forest of Trees.” *Scikit*, https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html