Predicting Movie Ratings

Ken Shimokawa, Bingqing Fu, Ting Wu

Abstract

This analysis attempts to answer the question of what the key factors are in making a good film. We analyze movie rating data using machine learning techniques to identify key factors. The most interesting results were that Running Time (the length of the movie), budget and release year were important factors that lead to high rating scores. These factors may reveal interesting biases reviewers may have and indicate to producers the factors making a film.

I. Introduction

There has never been a better time to be a film fanatic. The number of movies that are being released has never been higher and with the rise of streaming they have never been more accessible.

This analysis will seek to answer the question “what factors should a film producer look at in order to decide what film they choose to finance in order to make a highly rated film?” At face value it may seem counterintuitive to assume that the goal of a film producer would be to make a “good” movie instead of a profitable movie at the box office. Once the different revenue streams of a film are considered it makes sense.

Let us take for example the film Iron Man released by Disney and Marvel. Disney and Marvel have a multitude of ways to profit from Iron Man outside of the revenue from box office release. An increase in the number of people who visit Disneyland because of Iron Man, revenue from the soundtrack release, sales of merchandise, revenue from spinoff shows or sequels, home video revenue, streaming revenue, books, comics, parodies etc. From the production companies' standpoint all these streams of revenue are considered. So, while a film may do poorly at the box office, it can turn a net positive through these other means of revenue.

An article by Thomas Caldwell in 2011 explains that: how "Donnie Darko" transformed from a box office disappointment to a cult classic through DVD sales and word-of-mouth. [1]

This is not an isolated incident movies like: Shawshank Redemption, Fight Club, The Big Lebowski, and Blade Runner are all films that initially flopped at the box-office then have gained popularity through reruns, DVD sales and streaming services.

It is important to note that streaming has disrupted these alternative revenue streams dramatically. Matt Damon an accomplished actor and producer who has appeared in films such as Good Will Hunting and recently Oppenheimer explained in a 2011 interview that

“DVD was a huge part of our revenue stream and technology has just made that obsolete and so the movies that we used to make you could afford not to make all of your money when it played in the theater because you knew you had the DVD coming behind the release and six months later you’d get a whole other chunk of it (revenue) it would almost be like reopening the movie…but now, a 25 million dollar (budget) movie with another 25 million for printing and advertising…I have to split everything with exhibitors (movie theaters) I would have to make 100 million dollars...”[2]

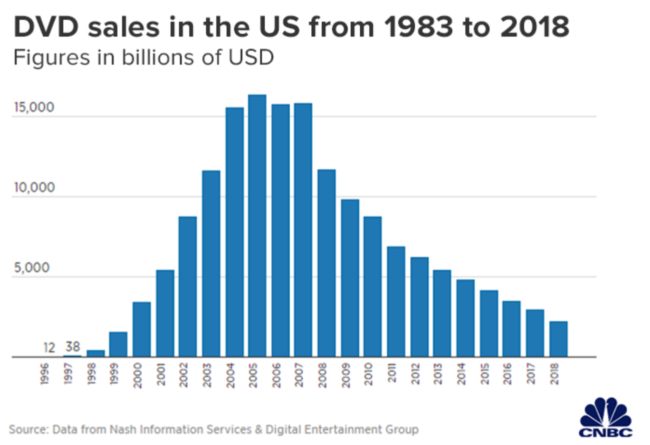


Fig. 1. DVD sales in the US [3]

While this may appear as if it would make our business goal irrelevant. The truth is streaming has also changed since 2011. Streaming services such as Disney Plus, Max and Peacock are all streaming services made by production companies so that instead of having to sell their movie after release to Netflix, Hulu, or Amazon the movie producers can enjoy the entirety of their film's profits via subscriptions.

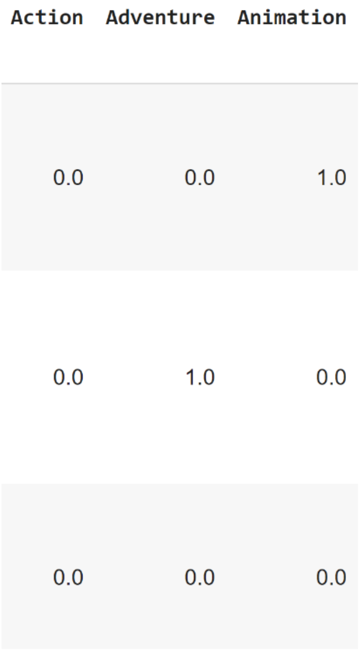
Also, Streaming services produced their own media content and premiered in theaters. Movies such as All Quite on the Western Front, Beasts of No Nation and Marriage Story are all “Netflix originals” that have made money both through box office success and streaming subscription.

It is also simply more difficult to measure the profitability of a film. While Films do disclose box office revenue and production costs, they don’t disclose marketing costs. Many news articles roughly estimate that studios tend to spend an equivalent amount making a movie as they market the movie. This would mean for a film to make a profit purely off box office revenue the film would need to make twice as much as it took to produce it. However, as discussed, this isn’t the only source of revenue for movies. Unfortunately, the revenue from these sources is much harder to track because companies aren’t required to disclose them and in addition it's hard to distribute the profit from these sources to individual movies. For example, if an Ironman toy is sold, we don’t know what share of the income should be attributed to Ironman 1,2 or 3. Due to these characteristics we decided to analyze the ratings of movies instead of tracking revenue from movies.

In this project the “The Movie Dataset” provided by Kaggle is used. [4] The data was first prepared for usage, then an exploratory data analysis was done. This was followed by feature engineering and finally model testing.

First the Data is prepared by creating a table for genre whereas originally genre was represented as a string where a row would list the different genres the film belonged to E.g. “Animation, Comedy, Family”.

However, with a table the data has features for the different genres a film could belong to which was then measured as a Boolean meaning for each row i.e. movie was marked by a 1 for each genre it belonged to and 0 for the genres it didn't.



Action	Adventure	Animation
0.0	0.0	1.0
0.0	1.0	0.0
0.0	0.0	0.0

Fig. 2. One hot encoding

Here (counting from the top) the first movie belongs to the genre "animation" (indicated by the 1) but does not belong to the genre's "action" or "adventure".

Then, the feature "belongs to a collection" which indicates whether a film is a part of a series E.g. Toy Story I, II and III was converted to the appropriate data type Boolean. The dataset also included the release dates of each movie. This was split into two new features namely its release month and the release year. The release month will be able to account for seasonality and release year will be able to control for changes over time. In addition to these null values, duplicated values and invalid values were all dropped.

Next, an exploratory data analysis was done. By looking at the distribution of the features, some expanded new features are shown to have low variance E.g. Documentary and TV movie. These features will be eliminated. Additionally, a lot of features are not normally distributed, and we have a dataset with a high dimension. We will use a k-means to detect and eliminate outliers.

Using a k-means clustering with a 95% threshold 268 rows are removed. The dataset is left with 5075 rows. Afterwards the rows are classified into three categories: Bad for score from 0-5, Average for score from 5-7 and good for score from 7-10. These categories are based on the following guidelines provided by the website Imdb.

1/10 - 'Do Not Want'
2/10 - 'Awful'
3/10 - 'Bad'
4/10 - 'Nice Try, But No Cigar'
5/10 - 'Meh'
6/10 - 'Not Bad'
7/10 - 'Good'
8/10 - 'Very good'
9/10 - 'Excellent'
10/10 - 'Masterpiece'
[5]

Additionally, the data is split into two. 80% of the rows were used to train our predictive models and the other 20% to test the models we trained.

In this analysis, the random forest model was found to be the most accurate and according to out model run time was the most important feature in determining the rating of each movie.

II. Related Works

While the film industry may seem intuitive there are several things that need to be considered. Most recently as we are all too familiar with the spread of the coronavirus dramatically changed how we lived out day to day lives. In the film industry this meant a halt in production and no film releases for over a year. This led to the postponement of many film releases and a steep decline in box office sales. According to IMDB from 2019 to 2020 total worldwide box office revenue decreased by 77.6%. This has led to a new resurgence in the film industry as movie goers are now able to go see films in theaters.

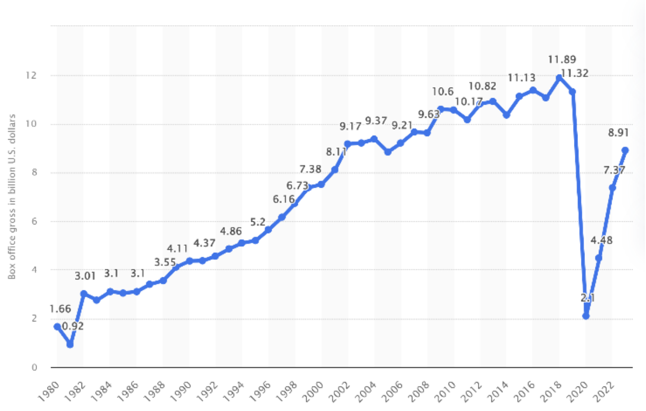


Fig. 3. Box Office [5]

In addition to the coronavirus, the film industry has been in competition with streaming services where consumers can watch movies from home. This has dramatically shaped the film industry. The film: Black Widow produced by Marvel premiered on Disney's streaming site due to the combination of the coronavirus and the sudden rise of streaming.

Others have done similar analyses. An article by Ryan Anderson published in 2019 titled "What makes a successful film? [6] Predicting a film's revenue and user rating with machine learning" The goal of this article was to find out "...whether, knowing only things I could know before a film was released, what the rating and revenue of the film would be. What parameters best predict a good or top grossing film?" The article had three main findings. First, the model accurately predicted film revenue with an r-squared of 0.77. Second, Anderson found that predicting the quality of the film was much harder than revenue (an r-squared of 0.54) even still better than prediction by using the average movie rating. Finally, the author was able to find out that "Film crew turned out to be the difference between a bad and a good film rating prediction, as well as the biggest difference between a well and poorly rated movie. Much more so than actors are."

Another article by Jasmine Holdsworth titles "Predicting IMDB Ratings for Movies" looked to find "how easy it could be to predict the average rating for a movie, and what predictors have the most effect." [7] Ultimately the plan is to develop an app "to allow users to... predict the ratings and revenue a film would generate." Holdsworth used "The Movie Dataset" the same database used in this analysis. Holdsworth first combined the CSV files in the database into a single master table. She then removed features that she felt were irrelevant and cleaned the features she wanted to use. She then created a new feature called Movie Genre. She made sure to measure Movie Genre as a temporary column since movies can belong to more than one genre. Then a pivot table was created for genre, filling all null cells and then was added to the main data frame.

Later, Holdsworth added the cast to the data frame. The roles added were Lead actor, Supporting actor and director. Rows with null for revenue, budget were deleted and rows where the vote count was below 50 and duplicated movies were also deleted. In addition, each column was converted to the appropriate data type.

However, since some films had multiple directors, some duplicated films weren't removed. Holdsworth dummified the director's features and merged them with the original data frame.

Holdsworth then checked for correlations between the predictors using correlation matrix.

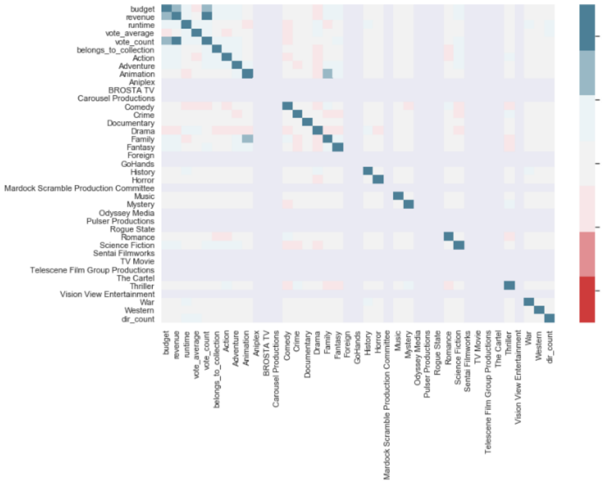


Fig. 4. Heatmap [7]

Since no other variables are highly correlated with vote average Holdsworth concludes that no single features are particularly predictive. However, Holdsworth does find four other strong correlations. Namely: revenue and budget, vote count and budget, revenue and vote count and family and animation.

A. Revenue and Budget

Since in the initial linear model data points were clustered together Holdsworth opted to use a double log scale to understand the relationship better.

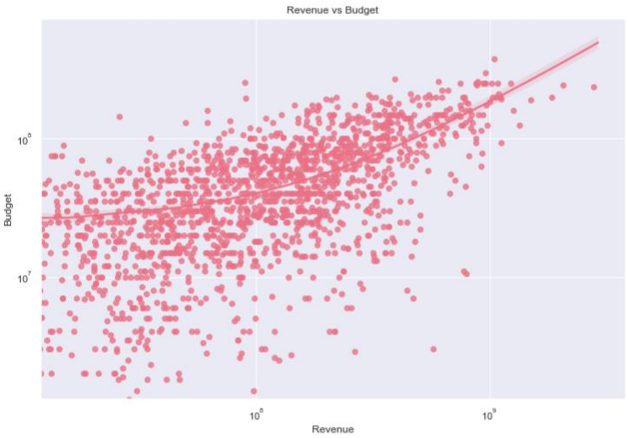


Fig. 5. Revenue vs Budget [7]

"The relationship is strong at a r-squared of a 0.69, although it is not represented perfectly by a straight line. The correlation coefficient between revenue and budget. It also makes intuitive sense that movies that have larger budgets will attract larger audiences." [7]

B. Revenue and Vote Count

"The correlation coefficient between revenue and vote count is even more significant at 0.74. As with the relationship between revenue and budget, the larger an audience the higher the number of votes can be expected." [7]

C. Vote Count and Budget

A double log model of vote count and Budget has a correlation coefficient of 0.52 and while the relationship is nonlinear, "It appears films with a larger budget spent producing them tend to generate more votes, which makes sense."

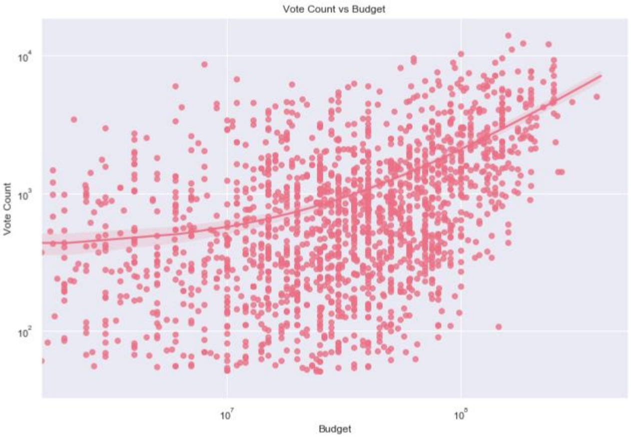


Fig. 6. Vote count vs Budget [7]

D. Family and animation

In the case of family and animation while there is a total of "78 films of genre 'Animation' and 186 films of genre 'Family'. While most Animation films are also classified as Family, the same cannot be said for the other way around." For these reasons Holdsworth keeps both predictors in the model.

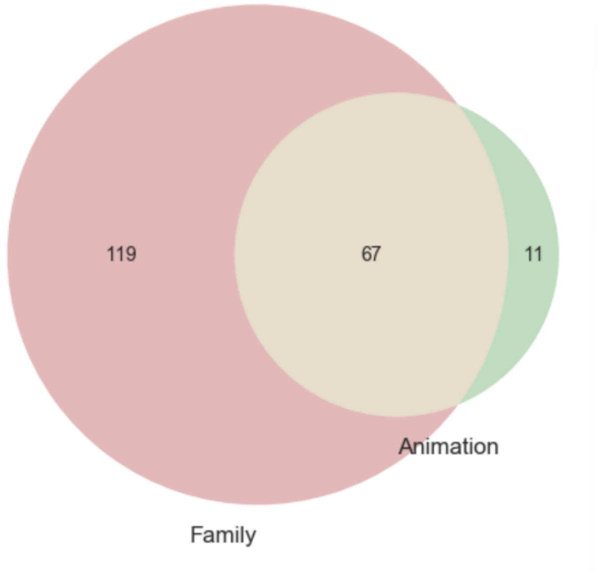


Fig. 7. Family and animation [7]

Holdsworth then splits the data into the training and testing datasets. Holdsworth also creates a “Baseline score with which to compare the scores for all models moving forward. This score will represent the score one would get if they were just to predict the mean value of y. If my model outperforms this score, I know it is doing well.” The score is 0.83 meaning “that the predicted score is within 0.83 of a mark out of 10.” [7] This is the benchmark to measure the performance of models. The models tested include Linear regression, Regularized-Ridge, regularized–Lasso, Regularized-Elastic Net, Simple Decision Tree, Random Forest, Grid searched Random Forest, Bagged Decision trees, Linear SVR, RBF and Poly. Finally, Holdsworth exports the finalized dataset and pickles the SVR model both to use in the application.

An article published by Zhuang, Jing, and Zhu in 2006 discusses how the rise of the internet has made online reviews increasingly valuable for consumers. Researchers are now focusing on automatic review mining and summarization, particularly in specific domains like movie reviews. This paper proposes a multi-knowledge-based approach that combines WordNet, statistical analysis, and movie knowledge to extract features from reviews and determine their sentiment (positive or negative). The experiments demonstrate the effectiveness of this approach in mining and summarizing movie reviews. [10]

III. Proposed Methodologies and Machien Learning Techniques

During this section, we’ll describe in reasonable detail the ML algorithm(s) and pipeline we used to address our movie rating prediction problem.

Based on the characteristics of our dataset, we initially considered five algorithms for our prediction task, along with one algorithm dedicated to outlier elimination. Here are the reasons for our selection:

Logistic Regression and KNN serve as strong starting points due to their simplicity, efficiency, and ability to handle multiclass classification tasks effectively. KNN, in particular, demonstrates effectiveness with moderately sized datasets. Additionally, SVMs offer the capability to address imbalanced datasets through adjustments in class weights during training.

Random Forest and Gradient Boosting emerge as robust choices for handling high-dimensional datasets with intricate relationships. Known for their resilience to noise and overfitting, they are well-suited for multiclass classification tasks featuring a moderate number of features.

A. Algorithms Description

1) LogisticRegression with ovo: logistic regression is a supervised machine learning algorithm widely used for binary classification tasks. This approach utilizes the logistic (or sigmoid) function to transform a linear combination of input features into a probability value ranging between 0 and 1. With OvO, we train a binary classifier for every pair of classes in the dataset. During prediction, each classifier makes a binary decision between two classes. The final class assignment is determined by aggregating the results of these binary decisions, often through voting.

2) Support Vector Machine: SVM is a powerful supervised learning algorithm that can handle both linear and non-linear classification tasks by finding the optimal decision boundary in the feature space. It is widely used in various applications, including classification, regression, and outlier detection. SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Being max-margin models, SVMs are resilient to noisy data (for example, mis-classified examples).

3) Random Forest: Random forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

4) XGBoost: XGBoost (Extreme Gradient Boosting) is an efficient and scalable machine learning algorithm that belongs to the ensemble learning family, specifically boosting algorithms. It works by building a series of decision trees sequentially, with each subsequent tree aiming to correct the errors of the previous one.

5) k-nearest neighbors algorithm (K-NN): k-nearest neighbors algorithm (K-NN) is a non-parametric supervised learning method. It is used for classification and regression. Input consists of the k closest training examples in a data set. In K-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

6) K-means clustering: k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. Given a set of observations (x1, x2, ..., xn), where each observation is a *d*-dimensional real vector, k-means clustering aims to partition the n observations into k (⩽ n) sets S = {S1, S2, ..., Sk} so as to minimize the within-cluster sum of squares (WCSS)

B. *Workflow and Pipiline*

In this section, we present a high-level overview of our entire machine learning process, delineated into two crucial pipelines. The first pipeline outlines the process of determining the optimal feature engineering techniques and subsequently applying them to our dataset. The second pipeline focuses on identifying the best-performing model with default parameter settings and further refining it through parameter tuning to enhance its classification performance metrics.

1) Overall Workflow

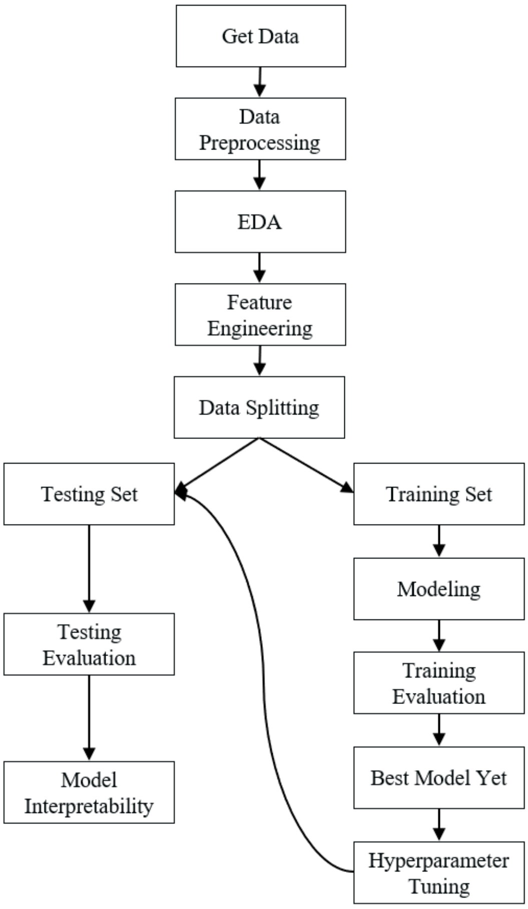


Fig. 8. High-Level Workflow

2) Sub Pipeline

We initially explored the application of classic feature engineering techniques to optimize our dataset. This included the standardization of features through scaling, detection and removal of outliers, oversampling to address heavily imbalanced classes, and dimensionality reduction methods.

Given the inherent complexity of our dataset, characterized by heavily imbalanced classes and a non-normal distribution of most features, determining the most effective feature engineering techniques posed a challenge. Furthermore, our exploratory data analysis (EDA) revealed the presence of outliers, and the expansion of features through the creation of dummy variables for categorical multivalue feature 'genres' added to the complexity.

To identify the most suitable feature engineering approach, we employed the GridSearchCV method. This method enabled us to systematically search through a range of parameter values for each feature engineering technique and evaluate their performance. Ultimately, StandardScaler emerged as the most effective technique, allowing us to standardize our dataset. Although StandardScaler was selected as the final winner, the experimentation process was essential in guiding our decision-making and ensuring the robustness of our feature engineering pipeline. Below is the pipeline of our determination of feature engineering techniques.

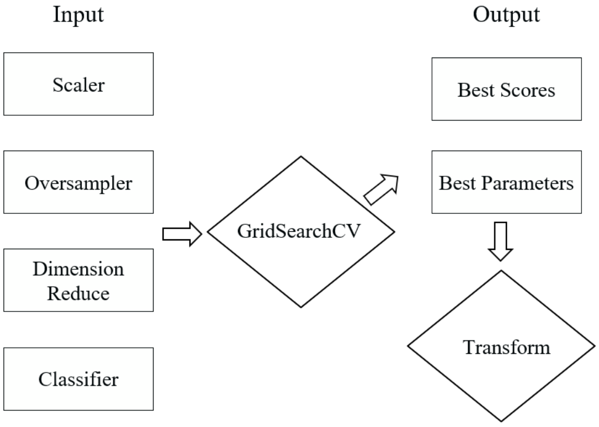


Fig. 9. Feature Engineering Pipeline

As previously discussed, we have selected five classification algorithms suitable for our task. Below is the pipeline outlining the process of selecting and fine-tuning classifiers to train a robust model for our predictions.

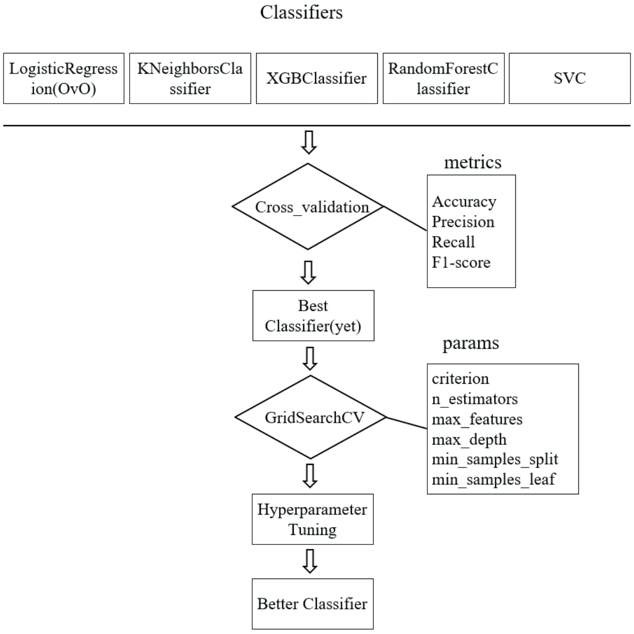


Fig. 10. Model Development Pipeline

TABLE I. SPECIFIC STEPS

```
Movie Rating Using Various Classification Algorithms
Input       Original movie dataset
Output      Rating label
Preprocessing
1.   dataset load
2.   dataset preprocessing: clean genres, belong_to_colletion and budget
3.   Encoding categorical feature by MultiLabelBinarizer()
4.   Filter nulls, budget=0 and vote_count<50
Feature Engineering
5.   Detect and eliminate outliers using k-means
6.   Remove low variance features by 95% threshold
7.   Classify the vote_average to 3 classes ['bad movie' rating from 0-5: label as 0, 'average' movie rating from 5-7: label as 1, 'good movie' rating from 7-10: label as 2].
8.   Find the best feature engineering techniques by GridSearchCV using precision metric
9.   Scaling independent features with StandardScaler(), skip the oversampling and PCA steps.
Modelling and Evaluation
10.  Split the dataset into train and test set with test_size as 20%
11.  Train and Evaluate each model in their default settings on train set using 4 metrics[accuracy, precision, recall, f1_score] to find the best performance model with cross_validate and pipeline techniques
12.  Hyperparameter tuning the best model yet with GridSearchCV with f1-score metric
13.  Train the best yet model with best parameters
14.  Evaluate train set after hyperparameter tuning
15.  Predict and evaluate tuned model on test set
16.  Check the feature importance with Random Forest
```

C. *Case Study*

To demonstrate our methodology, let's consider an example to provide a succinct overview of how our approach functions and how rating labels are assigned. Additionally, we will compare our predicted labels with the labels in the IMDB dataset to evaluate the effectiveness of our algorithm.

Let us give an example with below feature values(not display genre features with 0 value):

TABLE II. FEATURES

vote_count	runtime	budget	month	collection	crime	drama	label
3418	**200**	**13000000**	**12**	**1**	**1**	**1**	**2**

Our algorithm has assigned a label of 2 to this movie, which, according to our rating class, indicates a good movie. The actual rating and title of this movie are revealed to be 'The Godfather: Part II,' with an original vote_average of 8.3. It aligns with our classification as a good movie with label 2.

IV. Experimental Evaluation

A. *Dataset Description*

In the analysis, the data set was first cleaned of useless information (Ex. the link for the imdb webpage, imdb ID which is redundant information when considering that there is already an ID feature, taglines etc.). By removing redundant information and features the data was made to be more manageable and easier to manipulate and because these features were basically irrelevant will not be expanded on any further.

Title is the title of the film as listed on Imdb. ID is a primary key for each film. The feature "belongs to a collection" is an indication of whether a movie belongs to a film franchise (think Avengers and Star Wars movies). Budget is a numerical feature that measures the production budget for each movie. As discussed previously this number does not include the promotional budget or any other costs that aren't directly used to produce the film. Runtime is how long the movie is measured in minutes and the vote average is the rating the film received from viewers on Imdb. Vote count is how many people reviewed the movie on a one to ten scale on Imdb and the following features such as: animation, action and adventure are a tmp table measured as a Boolean data type that measures which genre the film belongs to.

B. *Experimental Settings*

We evaluate our model based on several criteria including accuracy, precision, recall, and F1-score. Our goal is to ensure a comprehensive assessment of model performance across different models in various aspects and confusion matrix are used to define potential results as four categories in Table []. The ranking evaluation criteria is as below:

TABLE III. CONFUSION MATRIX

	Predict Good	Predict Average	Predict Bad
Actual Good			FP-B
Actual Average			FP-B
Actual Bad	FN-B	FN-B	TP-B

C. *Accuracy*

This criteria measures of what percentage movie is correctly classified into bad compared with the total predictions made. Accuracy can be computed by equation().

$$Accuracy_B = \frac{TP_B}{TP_B + FP_B + FN_B}$$

D. *Precision*

This criteria measures of what percentage movie is correctly classified into good, average or bad compared with the total predictions made. Precision can be computed by equation().

$$Precision_B = \frac{TP_B}{TP_B + FP_B}$$

E. *Recall*

This criteria measures of what percentage actual bad movie is correctly identified with the total predictions made. Recall can be computed by equation().

$$Recall_B = \frac{TP_B}{TP_B + FN_B}$$

*1) F1-Score:*We use F1-score as harmonic mean of precision and recall, given a balance between them. F1-score can be coputed by equation().

$$F1 - Score_B = 2 \times \frac{Precision_B \times Recall_B}{Precision_B + Recall_B}$$

And in same way we can have confusion matrix Good and confusion matrix average. Then we compute the mean as Macro-average score.

2) Macro-average: And in same way we can have confusion matrix good and confusion matrix average. Then we compute the mean as Macro-average score. To make the equation concise, we abbreviate the output $Accuracy_B$ as A_B in the equation $Macro - accuracy$ and others are treated in same way. *Macro-average* can be computed by equation().

$$Macro - Accuracy = \frac{A_G + A_A + A_B}{3}$$

$$Macro - Recall = \frac{R_G + R_A + R_B}{3}$$

$$Macro - F1 - Score = \frac{F_G + F_A + F_B}{3}$$

For the model, we also make hypotheses as below:

H1: "Ensemble methods, particularly XGBoost and Random Forest, will perform better than logistic regression due to their ability to interactions between features."

H2: "Genres, will significantly enhance model performance by allowing models to capture more complex patterns in the data."

To gain more understanding for the movie prediction problem, we summarize the related work and take budget, genres, release time as independent variables and the movie grade (good, average or bad) as independent variables.

When finished data cleaning, we maintain a dataset of about 4,000 samples and we split the training and test data in a 80:20 percentage as the dataset is not a huge data and we need to have enough test data to evaluation our models. The analysis includes cross-validation to validate the models' performances robustly. We have visualized the feature distributions and correlations to understand the data better and to refine our feature selection process.

F. Results

This section describes results from our experiments with various machine learning models on the movie metadata dataset. We used graphs and histograms to visually represent the data, which provide clearer insights than tabular data, particularly in highlighting the basic differences and statistical significance of the results.

G. Correlation analysis:

THROUGHVISUALIZATION, CORRELATION ANALYSIS CAN MAKE THE ORIGINAL DATA PRESENT A MORE INTUITIVE CORRELATION BETWEEN THE RELATIONSHIP BETWEEN THE DATA, FOR FURTHER DATA PROCESSING AS WELL AS MODELING DATA SCREENING TO PROVIDE A RELIABLE IDEA.

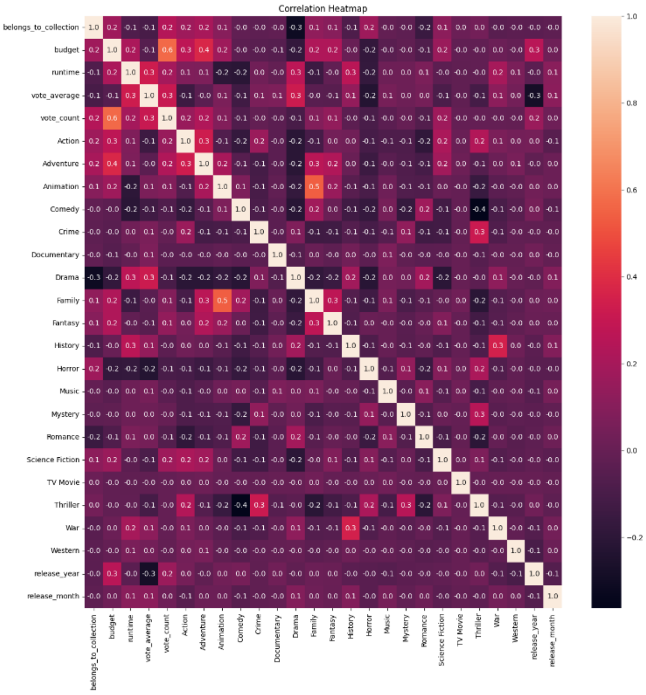


Fig. 11. Heatmap

From the heatmap, we can find there are positive correlations between budget and vote_count by 0.6, and another significant correlation is found in genres which the Family and Animation have a correlation by 0.5.

H. Classifier results:

In the initial stage of modeling, for the classification problem, we tried to use five algorithms OneVsOne, KNeighbors, SVC , RandomForest and GradientBoosting for predicting the movie ratings. Logistic Regression and KNN are good starting points, they are simple and efficient. And can handle multiclass classification tasks well. KNN can be effective with a moderately sized dataset. Random Forest and Gradient Boosting are effective for handling high-dimensional datasets with complex relationships. They are robust to noise and overfitting, making them suitable for multiclass classification tasks with a moderate number of features. For the output, we based on the Accuracy, Precision, Recall and F1 scores to measure the performance of the model.

TABLE IV. TRAINING SET PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score
OneVsOne	0.99	0.99	0.98	0.99
KNeighbors	1.00	1.00	1.00	1.00
SVC	0.79	0.73	0.47	0.48
RandomForest	0.80	0.72	0.54	0.58
XGB	0.78	0.68	0.48	0.50

TABLE V. TEST SET PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score
OneVsOne	0.99	0.99	0.98	0.99
KNeighbors	1.00	1.00	1.00	1.00
SVC	0.79	0.73	0.47	0.48
RandomForest	0.80	0.72	0.54	0.58
XGB	0.78	0.68	0.48	0.50

Random forest saw a perfect score of 1, this may indicate some parameters need more training. KNN also have good performance in training set however in the test set it have weaker performance. One vs one classifier as a basic algorithm is considered as a benchmark, we are happy to see all other algorithms show a better performance.

1) Hyperpamameter tuning

From the initial model attempt, we found that random forest may have better performance for this kind of problem, we use GridSearchCV to adjust the relevant hyperparameters to optimize the performance of the model in our further work, and the adjusted random forest model performs as follows.

TABLE VI. CONFUSION MATRIX

Model	Random Forest
Accuracy	0.75
Precision	0.60
Recall	0.52
F1 Score	0.55

I. Confusion matrix for each class:

To gain more detailed understanding with our model across different classifications of movie quality, we made a detailed confusion matrix by class to find if some classes are significantly have better or worse performance than others. The model performs well in identifying Good and Average movies with high overall accuracy and reasonable balance between precision and recall.

However, the model struggles with the Bad category, where it shows good accuracy but suffers from low precision, indicating many movies are incorrectly labeled as Bad.

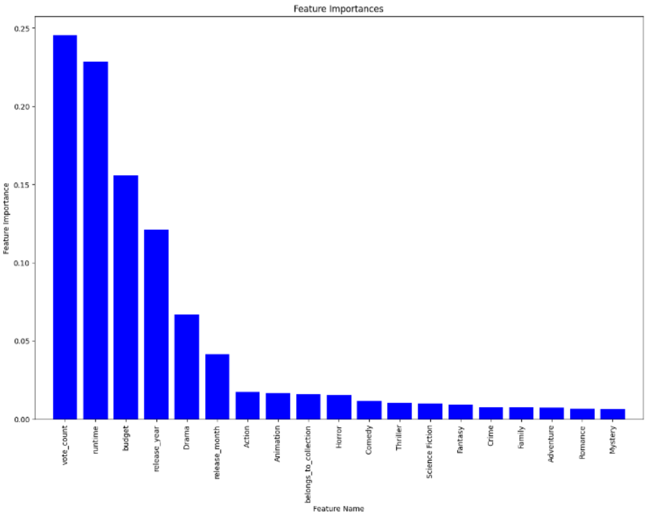
TABLE VII. CONFUSION MATRIX

	Accuracy	Precision	Recall	F1 Score
Good	0.91	0.37	0.25	0.30
Average	0.75	0.79	0.89	0.84
Bad	0.91	0.36	0.25	0.29

J. Feature Importance:

Understanding which feature significantly impact the model also take a important part in our work. The bar chart below illustrates the relative importance of various features used in the model.

.



The bar chart shows vote count, runtime, budget, release time and drama takes higher importance than other features while mytery, romance and adventure may not be be as decisive for predicting a movie's success in the dataset used.

K. Discussion

1) Performance of classifer on different classes: The results from Table III support several key insights into the strengths and weaknesses of the predictive model:

- Strengths: The model is highly accurate in identifying movies as good or bad, which is critical for applications such as investing decision making. The high precision for average movies is particularly advantageous for avoiding misclassification of moderately received movies as extremes.

- Weaknesses: The low precision in the bad category indicates possibility for false positives in this category, which could lead to financial loss when missing some potential good movies. Additionally, the moderate recall rates across categories suggest there is room for improvement in capturing all relevant instances within each category. This may be caused by the imbalanced raw data as the bad movie only accounts for 10% of the average one, we used oversampling to improve the model performance and get a satisfied overall model, sample rebalancing for bad movie could be considered in follow-up work to improve the model performance by class.

Compared with other research models made on movie success predicting, our model show a high accuracy as we only need to decide a general category of movies which is enough for investing decision making.

2) Performance of classifer accuracy: The Random Forest model was compared against other machine learning models, such as Support Vector Machines (SVM) and Logistic Regression, which were also considered during the initial phases of the project. The strengths of the Random Forest model lie in its ability to handle high-dimensional data and its robustness to overfitting compared to simpler models like Logistic Regression. Unlike SVM, the Random Forest is more interpretable, especially with regards to understanding feature importance.

The Random Forest model was compared against other machine learning models, such as Support Vector Machines (SVM) and Logistic Regression, which were also considered during the initial phases of the project. The strengths of the Random Forest model lie in its ability to handle high-dimensional data and its robustness to overfitting compared to simpler models like Logistic Regression. Unlike SVM, the Random Forest is more interpretable, especially with regards to understanding feature importance.

The performance of the Random Forest can be attributed to its ensemble method, where multiple decision trees are combined to improve accuracy and control overfitting. The diversity among the trees, achieved through bootstrapping samples and feature randomness, ensures that the model captures a wide range of data characteristics. Moreover, the dataset's features, such as runtime and budget, likely have inherent predictive power which the Random Forest can exploit effectively due to its comprehensive approach to feature handling.In conclusion, the Random Forest model's application to predicting movie ratings demonstrates its capability to manage complex datasets and provide substantial accuracy, making it a preferred choice over simpler or less robust models. Future work might explore the integration of more granular temporal data or the use of advanced ensemble techniques like boosting to further enhance predictive performance.

V. Conclusion

In our analysis we found that the Random Forest Model to be the most well-suited model. While we did need to marginally prune the model, we found that Train Accuracy was 0.94, Train Precision was 0.942, Test Accuracy was 0.88 and Test Precision to be 0.89. These results are much better than random chance. Out of the many features that were predictors of rating it was interesting to see that Runtime was the second most likely predictor of ratings. In an article titled: Longer Movies are better, according to the Tomato Meter by Mark Hofmeyer discusses the effects of movie duration on rotten tomato ratings [9]. The article delves into the relationship between a film's length and its critical reception, analyzing data from 1,431 wide-release movies since 2010. Films were categorized based on their duration: less than 100 minutes (about 1 and a half hours), 100-120 minutes (about 2 hours), 120-140 minutes (about 2 and a half hours), and over 140 minutes (about 2 and a half hours).

Movies under 100 minutes (about 1 and a half hours) were found to be the most likely to receive negative reviews, with only 34% being rated Fresh on Rotten Tomatoes. While exceptions like "Lady Bird" existed, most shorter films were criticized for lack of depth or substance. Conversely, films in the 100–120 minute range were more likely to receive positive reviews, particularly notable for genre hits like "Spider-Man: Into the Spider-Verse" and "Get Out. " The 120–140 minute category emerged as a sweet spot, with a 64% Tomatometer average and a higher proportion of critically acclaimed films, including Best Picture winners like "The Shape of Water." Longer films, over 140 minutes (about 2 and a half hours), boasted the highest Tomatometer average at 70%, featuring acclaimed titles like "The Wolf of Wall Street" and "Once Upon a Time in Hollywood." However, the article notes a decline in Best Picture wins for lengthy films since the 1990s.

Analysis suggests that longer runtimes often allow for more depth and complexity in storytelling, leading to higher critical acclaim. However, it also acknowledges that a film's success depends on how effectively it utilizes its runtime to engage audiences. Personal preference for ideal movie length varies; some viewers may prefer concise narratives, while others appreciate the depth offered by longer films.

In summary, while longer films tend to receive higher critical ratings, the ideal movie length is subjective and dependent on factors such as storytelling, genre, and audience engagement. Ultimately, the quality of a film is determined by how well it utilizes its runtime to deliver a compelling and resonant experience.

Other categories were more to be expected. We found earlier in our data exploration stage that Vote count was correlated to rating and that budget was related to vote count etc. It is surprising that in our model that release year was a predictor of rating. This would imply that the average rating of movies overall is increasing over time. Are movies getting better with time? It's something to think about.

Some genres had more of a rating effect than others. Drama and Animation were the 5th and 7th strongest predictors of movie ratings respectively. The general public may feel someways about certain genres overall. For example, Animation and Family have a huge overlap, is it possible that the criteria for what makes a good “family movie” is different from the criteria for personal favorites. If this were true, how would it impact the ratings? Would it imply that the ratings are skewed towards factors other than individual enjoyment of the film, but rather influenced by how suitable viewers perceive the movie to be for children?

In addition to this it could be possible that Drama is considered a more prestigious genre of movie and thus tends to receive higher scores than it deserves because of how “important” the subject is etc.

Another intriguing revelation emerges when examining the predictive power of the feature "belongs to a collection" on movie ratings. One might naturally presume that sequels are produced due to the positive reception of the original film, thereby both the original and its sequel falling under the "belongs to a collection" category. Surprisingly, this assumption does not hold true in the analysis.

It prompts one to ponder: could viewers' expectations set by the excellence of the original film inadvertently lead to disappointment with its sequel? The phenomenon of sequels failing to live up to the acclaim of their predecessors is not uncommon. Perhaps viewers, having been enamored by the first installment, approach subsequent entries with heightened scrutiny, causing them to be more critical in their assessments.

This raises the question of whether the weakness of "belongs to a collection" as a predictor of movie ratings stems from a tendency for audiences to judge sequels more harshly. If the initial film sets a high bar in terms of storytelling, character development, and overall entertainment value, its sequel may struggle to meet these heightened expectations. Consequently, even if the sequel is well-crafted, it may still fall short in the eyes of viewers who hold it up against the standard set by its predecessor.

Furthermore, factors such as changes in cast, creative direction, or narrative focus between installments could also contribute to the disparity in ratings between original films and their sequels. Viewers may perceive these alterations as deviations from the essence of the original story or as attempts to capitalize on the success of the franchise without offering substantive contributions to the overarching narrative.

In essence, while one might expect films belonging to a collection to benefit from the goodwill generated by their predecessors, the reality appears to be more nuanced. The phenomenon of sequel disappointment and the complex interplay of audience expectations and cinematic execution underscore the multifaceted nature of predicting movie ratings in the context of franchise films.

VI. Future Work

A study focusing on the profitability of films would be an intriguing avenue for exploration, particularly given the inherent complexities involved in assessing a film's financial success and the more direct business application. Traditional metrics such as box office revenue and production costs provide only a partial picture, as they overlook significant factors such as marketing expenses and revenue from ancillary sources such as merchandise.

The challenge of accurately gauging a film's profitability is underscored by the opaque nature of the film industry, where studios often refrain from disclosing comprehensive financial data. While box office receipts and production budgets are typically made public, marketing expenditures are often undisclosed, leaving analysts to speculate based on rough estimates.

However, as mentioned this narrow focus on box office revenue overlooks the myriad revenue streams that contribute to a film's overall profitability. Revenue from sources such as merchandise sales, licensing agreements, streaming platforms, and ancillary markets like DVD and Blu-ray sales can significantly augment a film's earnings. Yet, tracking and attributing these revenues to individual films poses considerable challenges. While the profitability of films remains a complex and multifaceted subject, it offers a promising avenue for research.

Furthermore, delving into the realm of personalized film recommendations based on user data held by streaming services presents an intriguing frontier for exploration. With the vast amount of information that streaming platforms collect about their users' viewing habits, preferences, and demographic profiles, there exists a rich trove of data ripe for analysis and application in the realm of content recommendation systems.

These systems can consider various factors such as genre preferences, past viewing history, ratings given to previous films, time of day, device usage patterns, and even contextual information such as weather or holidays.

Moreover, personalized recommendations can also benefit content creators and distributors by facilitating the discovery and promotion of niche or lesser-known films to audiences who are most likely to appreciate them. This can help to diversify the range of content available on streaming platforms and provide greater visibility to independent filmmakers and underrepresented voices in the industry.

However, the implementation of personalized recommendation systems is not without its challenges. Issues such as privacy concerns, algorithmic bias, and the risk of creating "filter bubbles" where users are only exposed to content that reinforces their existing preferences, must be carefully navigated.

VII. Works Cited

[1] T.Cladwell"The cult of Donnie Darko," *The Independent*, Aug. 02, 2005.https://www.independent.co.uk/arts-entertainment/films/features/the-cult-of-donnie-darko-303185.html

[2] "Matt Damon Sweats From His Scalp While Eating Spicy Wings | Hot Ones," *www.youtube.com*. https://www.youtube.com/watch?v=yaXma6K9mzo

[3] S. Whitten, "The death of the DVD: Why sales dropped more than 86% in 13 years," *CNBC*, Nov. 08, 2019. https://www.cnbc.com/2019/11/08/the-death-of-the-dvd-why-sales-dropped-more-than-86percent-in-13-years.html

[4] "TheMoviesDataset,"*www.kaggle.com*.https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data?select=ratings_small.csv (accessed May 04, 2024).

[5] "How to rate movies on IMDB.," *IMDb*. https://www.imdb.com/list/ls076459507/

[6] "North American box office revenue 1980-2019,"*Statista*. https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/

[7] R. Anderson, "What makes a successful film? Predicting a film's revenue and user rating with machine learning," *Medium*, Aug. 06, 2019. https://ryan-anderson-ds.medium.com/what-makes-a-successful-film-predicting-a-films-revenue-and-user-rating-with-machine-learning-e2d1b42365e7

[8] "Jasmine Holdsworth," *jazpeng.github.io*. https://jazpeng.github.io/predict_movie_ratings/ (accessed May 04, 2024).

[9] "Longer Movies Are Better, According to the Tomatometer (And Really Long Movies Are Even Better Than That)." https://editorial.rottentomatoes.com/article/longer-movies-are-better-according-to-the-tomatometer/

[10] CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management