Machine Learning Final Project Spring 2023

Carter, Camryn
Student
University of Richmond
camryn.carter@richmond.edu

Hu, Penny
Student
University of Richmond
penny.hu@richmond.edu

Turner, Jayana
Student
University of Richmond
jayana.turner@richmond.edu

Sales, Caitlin
Student
University of Richmond
caitlin.sales@richmond.edu

I. Project Aim

For this project, we are aiming to predict the world sales of movies based on features such as release date, genre, runtime, and more. By comparing movies to the top-grossing ones, we hope to learn what makes a movie popular or top-grossing in order to determine whether to make such a film.

II. Overall Plan

A. Introduction

The movie industry is a continuously growing multi-billion dollar industry recovering from damage caused by the COVID-19 pandemic. This global commercial enterprise has a diverse audience and can attract consumers across all target demographics. In 2022, the global box office revenue was $26 billion, a 27% increase compared to 2021. In the United States and Canada, the total earning at the box office was around 7.37 billion U.S. dollars in 2022, which was 4.48 billion dollars more than the previous year.[1] However, revenue earned from the movie industry is still not at the levels it used to be before the COVID-19 pandemic. In 2018, the United States and Canada's box office revenue was 11.89 billion U.S. dollars.[1] The movie industry is still growing despite the economic setbacks attributed to the pandemic.

This lucrative and relatively successful business could be considered an excellent investment opportunity. Producers are major financial supporters of director's film projects or the continuation of popular movie franchises. However, how do they know if they are investing in the right movie? This question is pertinent because investing in an unsuccessful film can lead to a financial loss. In the current study, we aim to predict the world sales of movies, which can further support the popularity and success of a specific movie. Our goal is to use various machine learning models, such as Linear Regression, Polynomial Regression, Regression Decision Trees, etc., to construct an accurate model for predicting the global sales of a movie. The most effective model, as determined by the F1 score, will be able to predict the sucess of future films.

Additionally, using our best-performing model, we aim to predict if a subsequent movie will be successful. Starting a movie franchise can be expensive. With our model, it will be possible to predict success before investing time and money into the project, possibly maximizing profits.

B. Related Work

Previous studies have predicted the success of movies using machine learning models. In a study conducted by Dhir and Raj, the researchers used the Internet Movie Database to predict the IMDb scores of different movies.[2] Dhir and Raj used various machine learning models, including Support Vector Machine, Random Forest, Ada Boost, Gradient Boost, and K-nearest neighbors. Similarly, our current study aims to investigate multiple machine learning models, but to predict the global sales of a movie. This study found that the Random Forest had the highest precision, recall, F1 score, and accuracy compared to all other models. Specifically, the Random Forest model had a predicted accuracy of 61%.

Researchers Lee et al. predicted the success of movies using a classification machine learning algorithm.[3] Movies could be classified into five different classifications, from blockbuster to flop, based on attendance range and revenue range. Lee et al. used various machine learning models, such as gradient tree boosting, adaptive tree boosting, linear discriminant, logistic regression, neural networks, random forests, and support vector classifiers. The researchers used two different types of average percent hit rates (APHR) to measure the performance of their models: Bingo and 1-Away. Based on these two APHRs, gradient tree boosting performed the best. Again, similar to this study, we plan to study various machine learning models. However, we aim to predict the global sales of a movie, not the box-office success of a movie. The machine learning algorithms that the researchers used were classification models, and we plan to use regression models.

C. Dataset Used

The dataset we used was found on Kaggle. It is entitled "Top 1000 Highest Grossing Movies," by Sanjeet Singh Naik.[4] The dataset contains information regarding the top 1000 highest-grossing Hollywood films as of January 10, 2022. This information includes the movie title, movie description, distributor, release date, genre, runtime, domestic sales, international sales, and world sales.

Figure 1 below displays the features present in the dataset with a corresponding example of what the feature is/looks like.

D. Approach

In our approach, we would first clean the data of any null values and generate new features from the raw data. Next,

Feature	Example
ID	0
Title	Star Wars: Episode VII - The Force Awakens (2015)
Movie Info	As a new threat to the galaxy rises, Rey, a desert scavenger, and Finn, an ex-stormtrooper, must join Han Solo and Chewbacca to search for the one hope of restoring peace.
Distributor	Walt Disney Studios Motion Pictures
Release Date	16-Dec-15
Domestic Sales (in $)	936662225
International Sales (in $)	1132859475
World Sales (in $)	2069521700
Genre	['Action', 'Adventure', 'Sci-Fi']
Movie Runtime	2 hr 18 min
License	PG-13

Fig. 1. Features in "Top 1000 Highest Grossing Movies" dataset

we will graph each feature to the label and identify any relationship between the two. Features will be selcted for use in the model based on this criteria. We will then train multiple types of models with initial thresholds/values. Note that appropriate models will be chosen based on the label we are trying to predict and that any features used in the model will be scaled before training the model. Once initial training is complete, the model will be fine-tuned by adjusting any threshold values and/or adding/removing features from the model. Finally, when we have fine-tuned each model to have the best possible F1 score, we will compare the models to each other, then determine which one performed the best, and choose that model as our final one.

E. Data Analysis Steps

To analyze the data, we first plan to clean the data and calculate for new features. Afterwards, we plan to graph each feature against the label (i.e. world sales) to see any impacts/correlations between the two to determine what features to use in our model.

For cleaning the data, there are some rows where the value for the release date and/or license is "NA". The first idea to handle this is to delete any rows with missing information. While this is a fast solution in order to start testing/training our model, it will lead to a loss in data which will most likely affect the model's performance. Hence, what we would aim to do is to find the missing information from the data through research. Although this would potentially solve the problem of having a loss in data, this research would be time-consuming, and it is not a guarantee that we would find the missing information. Thus, we would do as much research as possible to fill up the missing data and apply the first method to the data set after we have thoroughly researched the information.

In terms of feature extraction, we want to be able to calculate new features from the existing ones in the dataset in order for them to be more usable. For example, one of the existing features is the movie description/information which is a string of text that introduces what the movie is about. However, in its current form, it would be hard to run any sort of model on the description. Hence, we plan to find the most frequently occurring words across all descriptions and create new features that indicate whether a frequently occurring word is in a description or not. Aside from this, we would like to apply one-hot encoding to the genre column of the data set. Since all genres associated with the movie are contained in a list form, we wish to break it up with one-hot encoding to make it more usable for our model. We would also like to make the runtime into a whole number form where the movie runtime is converted into purely minutes. Finally, we wish to break up the release date column into three new features: month, day, and year for the same reasons.

Once these steps are done, we will determine what features seem to have a strong correlation to the label and run our model with those features scaled as our predictors. However, as a note, we will try to avoid using both domestic sales and international sales as features since the sum of those two features would generate world sales.

F. Methodology

Our plan so far is to create and train multiple machine learning models in order to compare the performance of each through the corresponding F1 score and determine which model is most effective. Since we wish to predict world sales which is a continuous value, we decided to have a Linear Regression model. Aside from this, in case the data is non-linear, we would also train a Polynomial Regression model. It must be noted that the Linear and Polynomial Regression models will be trained using cross-validation with k-folds. Aside from this, since we are pursuing Linear and Polynomial Regression models, we would like to also look at using a Support Vector model (SVM) without any kernels and with a Radial Basis Function (RBF) kernel. Finally, given that our data is tabular, we wish to create a Regression Decision Tree model as well for more variety in our models. Additionally, we would like to test and compare the performance of using a single decision tree to a Random Forest model and XGBoost.

In our initial training for each model, we would start with initial thresholds/values. For the Linear Regression and Polynomial Regression models, we would start with cross-validation using a 3-fold split. Afterward, we would increase the number of splits and determine which k-fold produced the highest F1 score. As for the SVM with and without a kernel, we will determine our initial threshold values through research and fine tune the parameters from there. Then, for the decision trees, while there are no parameters that we would adjust for the single decision tree, we would adjust

the different parameters for the Random Forest model and XGBoost including but not limited to the number of decision trees, the maximum number of features to consider and the maximum depth of the tree. Note that the starting threshold values for the Random Forest model will be determined through research and consultation with our professor while the starting values for the XGBoost will simply be the default ones from the package.

Once each model is fine-tuned, we will compare the final F1 scores for each model and choose the model with highest score to be our final chosen model. However, it must be noted that each model trained and created will be discussed and presented on.

G. Expected Results

The experiment is used to predict the world sales of movies based on features such as release date, genre, runtime, and more. The expected results for the experiment are that if a movie is classified under the genre of action, adventure or sci-fi, they are more likely to be a popular and top-gross movie. Additionally, by comparing movies to the top-grossing ones, we also predict that the popular and top-gross movie has a run time that is more than 1 hour and 30 mins (i.e. 90 minutes) but less than 3 hours (i.e. 180 minutes) with no linear correlation between the popularity and the run time of the movie. Finally, based on the previous study done by Dhir and Raj, we believe that the Random Forest model would have the best performance.

III. Tentative Timeline and Responsibilities

Figure 2 represents our tentative timeline. For each task, there are dedicated person/s assigned to accomplish the task as well as the amount of time dedicated for that task.

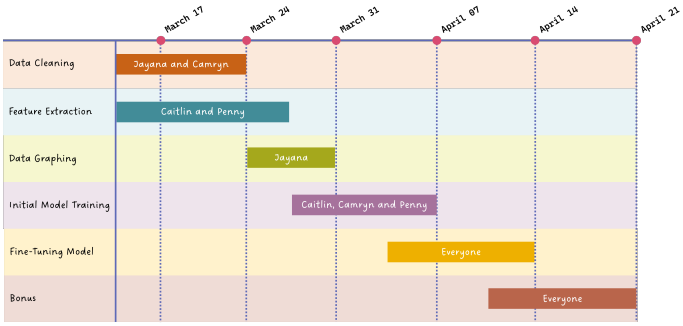

Fig. 2. Timeline of Tasks and Responsibilities

IV. Optional-Above and Beyond

After we predict the possible success of a movie (i.e. top-gross), for our bonus, we will add a suggestion component to the project. When a movie's predicted sales are under a threshold, the model will recommend additional features based on the training data for the movie makers to add to or change their film so that they can achieve higher gross and popularity. For example, a romantic movie that has a run time of 80 minutes is being planned and when the movie is predicted to have low sales, then the bonus portion of the project will recommend additional features like adding drama and fantasy to the genre and extend the run time of the movie to be 100 minutes for the movie to have a higher predicted sales.

V. References

1. Box Office Mojo. "Box Office Revenue in The United States and Canada from 1980 to 2022 (in Billion U.S. Dollars)." Statista, Statista Inc., 10 Feb 2023, https://www.statista.com/statistics/187069/north-american-box-office-gross-revenue-since-1980/

2. Dhir, Rijul, and Anand Raj. "Movie success prediction using machine learning algorithms and their comparison." 2018 first international conference on secure cyber computing and communication (ICSCCC). IEEE, 2018.

3. Lee, Kyuhan, et al. "Predicting movie success with machine learning techniques: ways to improve accuracy." Information Systems Frontiers 20 (2018): 577-588.

4. Naik, Sanjeet Singh. "Top 1000 Highest Grossing Movies." Kaggle, 15 Jan. 2022, https://www.kaggle.com/datasets/sanjeetsinghnaik/top-1000-highest-grossing-movies?select=Highest%2BHolywood%2BGrossing%2BMovies.csv.