# Netflix Movie Analysis and Rating Prediction

*Managing Big Data Final Project Report - Team7*

*Wei (Will) Jiang, Xiang (Maggie) Meng, Paakhi Srivastava, Yulin Gai, Joe Ebby Karuthedath*

## Executive Summary

For our final project we collected movie data from IMDB, Wikipedia, and Kaggle to put together a rich dataset and make useful business predictions for Netflix. As a company which profits off of viewers subscriptions, it is essential for Netflix to create a library of attractive movies for its audience. To determine which movies would be good investments for Netflix we collected data on the average movie rating, country of release, duration, genre, and more for a collection of movies from the sources listed above. After the collection of the necessary data, we joined and cleaned the data in SQL. Besides, we also store movie reviews data to AWS DynamoDB for future use. Next, we brought the data into Python for feature engineering, model building, and evaluation. We were able to build a successful KNN model which predicts if a given movie will be “good” which is worth for Netflix to invest in.

## Introduction

Netflix is a subscription based on demand streaming platform, focussing on movies and TV shows. With the advent of other large players like Disney+ and Amazon prime video, it is critical to put quality content on the platform to avert the challenge of eroding market share. Netflix provides its user base both in house productions and content that it acquires from other production houses. Both producing original content and acquiring content licenses are the biggest costs in the Netflix business model.

Optimizing the money spent in this channel is a necessity, as margins in the streaming business are becoming narrower by the day. Also spending the money on good quality content can expand the user base in emerging markets like Asia. This project aims to build a model that can aid Netflix in its decision making when it comes to acquiring content licenses from other

production houses and to see if the decision to produce original content is right or wrong even before the director says “start camera action”. The model does this by making a prediction if the movie can be categorized as a good one or a bad one.

## Data Wrangling - Data Collection and Web Scraping

Based on our goal of the project, we need as much qualitative and quantitative data as possible about Netflix movies. First, we downloaded the “Netflix Movies and TV Shows” dataset from Kaggle, which included over 8000 Netflix movies and TV shows by mid-2021. However, this dataset only included several attributes, such as the duration, country, release year, etc. To get other important attributes like genre, rating of the movie, and popularity, we scraped data from mainly 2 websites -- IMDB and Wikipedia. Besides, for further research, we also scrape the reviews for the movies we have. So the final dataset contains movie and review information for around 4,000 movies on Netflix platform.

## Data Wrangling - Data Cleaning and Storage

In SQL, we joined all the three tables together. Data cleaning part includes dealing with NA, data formatting and part of feature engineering for modeling part later in Python. We create new attributes as follows: whether it is a Netflix original movie, whether the movie is made in the U.S., in which season a film is added to the platform, whether parent guidance is suggested for children to watch a movie. Besides, for the review data, which includes much unstructured information (e.g. text), we choose to use AWS DynamoDB for storage. Such key-value databases are highly partitionable and allow horizontal scaling at scales that other types of databases cannot achieve. With high flexibility and scalability, we can store as many movie reviews as we need for future use.

The challenge and limitation of those datasets is that there is no unique ID for each movie so we have to pair them with the movie title itself, but the format is string and may contain messy symbols and thus cause problems sometimes. Another challenge is the data size itself only contains limited information about movies and the sample size is not big enough as well.

## Insights and Analytics

Exploring the datasets is always the first step before modeling. In fact, there are many potential factors affecting the movie rating, and we can also derive many correlations between these factors and the rating data. Therefore, in order to have a comprehensive understanding of our datasets, we use python to do the exploratory data analysis, and we extract and summarize some of the most relevant points from our findings. One key criteria in classifying movies is the rating class, Figure 3 and Figure 4 show that most movies target mature audiences, the audience class 2 (need parental guidance) also has plenty of market share. We also plot the scatter plot and regression lines for rating vs. duration (Figure 1), the plot suggests there is a stronger positive relationship between movie ratings and duration for non-Netflix-original movies. Another crucial factor that can be used for predicting rating is the release year of a movie, clear patterns are shown in Figure 5 and Figure 6. There is a high-yield period of film production between 2016-2019, but it does not lead to high score movies, instead, movies in the 1960s generally scored high. We also compare average rating between different genres, we notice that Netflix original movies have better rating performance in most of the genres.

Now that we have a good idea of what our data contains, we conducted feature engineering to make the data more useful for our goals. We created a binary movie class variable which would serve as our target variable by representing if a movie was good by having a rating above 6 or not good by having a rating under 6. Since the distribution of the target variable was uneven, as depicted in Figure 9, we decided to balance the data with oversampling so the proportion was equal. Since Netflix will not know the number of votes for an unreleased movie, we also mapped the number of votes by using 50% percentile as a threshold as a popularity measure, thus Netflix can know whether there is a hot topic about the new movie or not by using online resources. The final part of our feature engineering was transferring categorical variables to dummy variables, such as the season added or the genre of the movie.

Once we were done preparing our dataset, we began the modeling process by splitting the data into test and training sets. We ran a Logistic Regression, KNN, Decision Tree, and a Random Forest with parameter tuning for each. We used the AUC Score to select the best model and as seen in Figure 10, the KNN has the highest score with an AUC of 0.90. To evaluate our models, we created an ROC curve on the test data where we can see the KNN model surpasses the others with a AUC of 0.92, as seen in Figure 11.

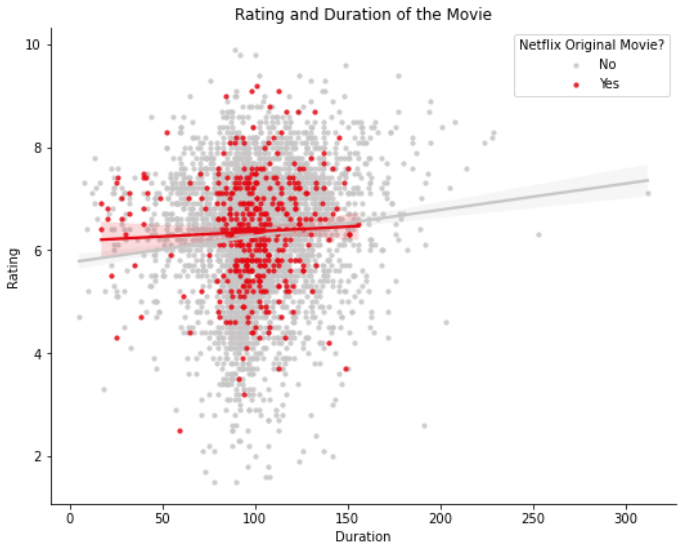
## Discussion

The model uses the data from IMDB to arrive at the target variable, which is widely accepted as a scale to differentiate between good and bad movies. Netflix is currently paying between $100 and $250 million for blockbuster movies, with no guarantee of the movie being a success or not. New releases on the platform are obvious to have a premium acquisition cost. Netflix also has a growing audience for good quality content in foreign languages, which might not have gained wide recognition due the regional nature of the local movie industry. This project can help the Netflix team to identify such high quality foreign cinema. The model can help Netflix executives to make a data driven decision to go ahead with an in house production or to acquire content license from a third party. As a result the platform can spend its valuable resources in acquiring the best quality content for its users, hence retaining the market dominance it currently possesses.

For Opportunities, it can be said a very friendly period for the development of the streaming platform during the pandemic, because people spend more time at home, which means that the streaming platform, such as Netflix, has more potential users. With the increase of user diversity, we can use our model to make predictions more accurately. For Threats, however, Netflix also has many competitors competing with them for loyal users，and there are also some further improvements in the script that can be brought about, such as using natural language processing on the script of the movies to classify them into high performing and low performing movie clusters.

# Appendix

## Figure Collection



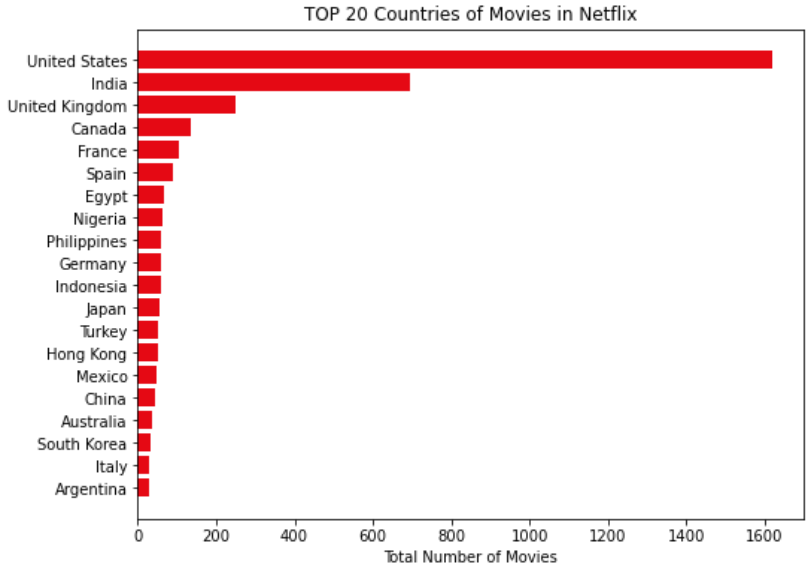
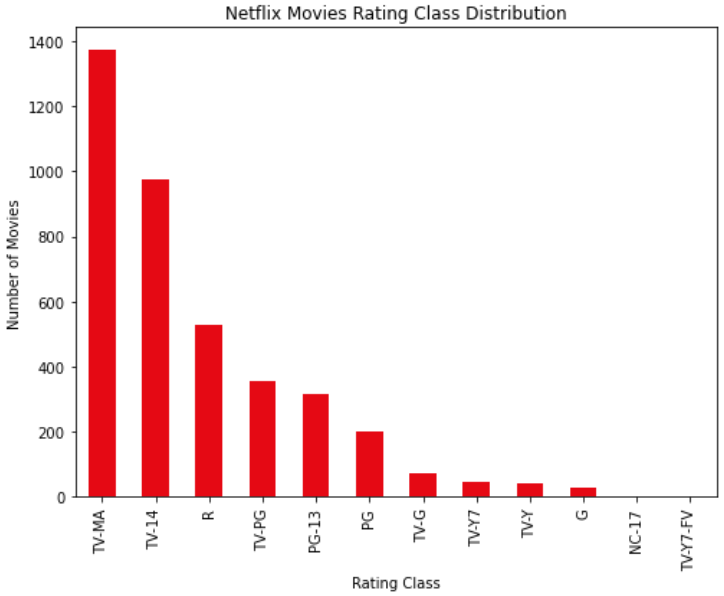




**Figure 1.** Rating and duration scatter plot

**Figure 2.** Count plot by countries



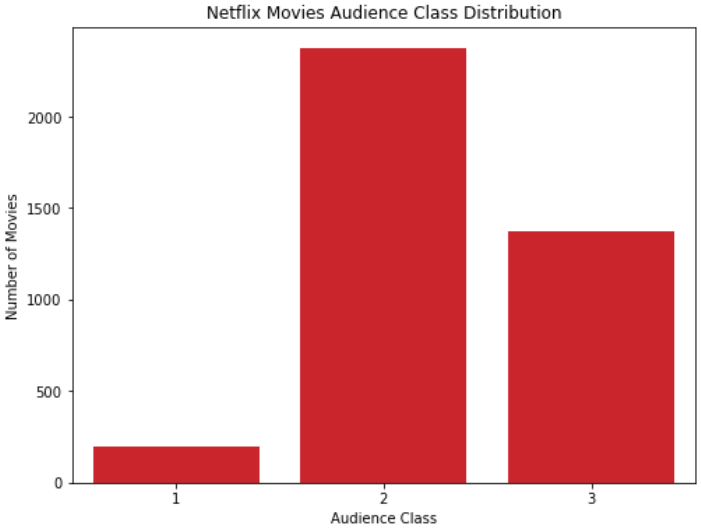
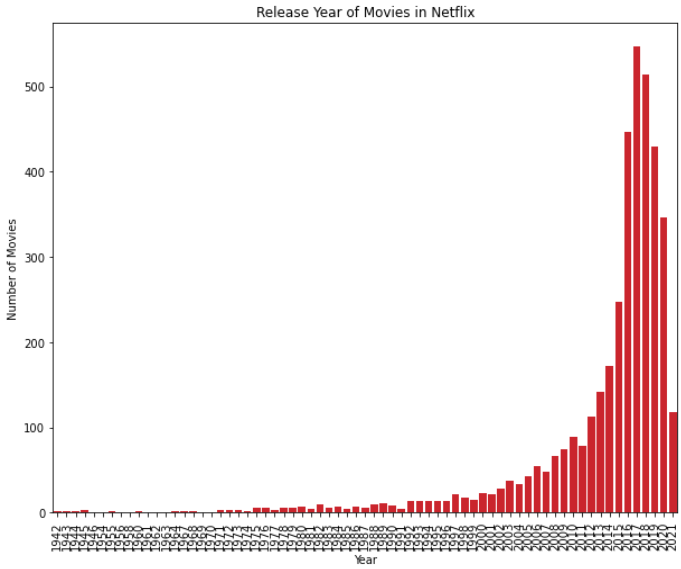



**Figure 3.** Count plot by rating class

**Figure 4.** Count plot by audience class



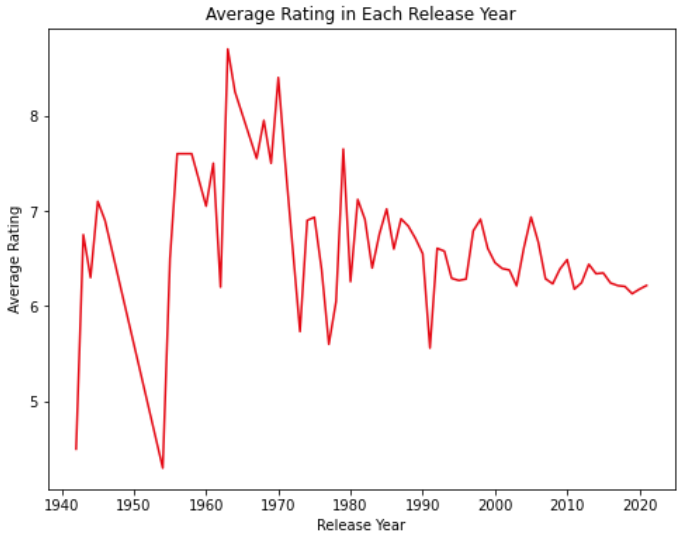
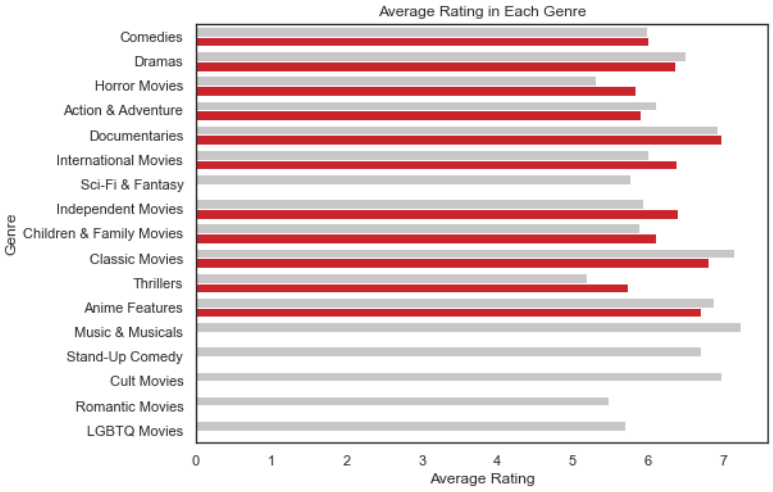




**Figure 5.** Count plot by release year

**Figure 6.** Line plot of average rating per year



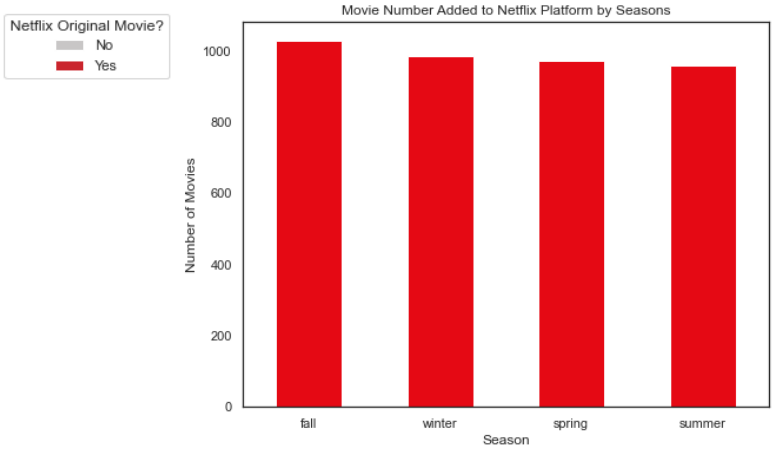




**Figure 7.** Bar plot of average rating per genre

**Figure 8.** Count plot by season

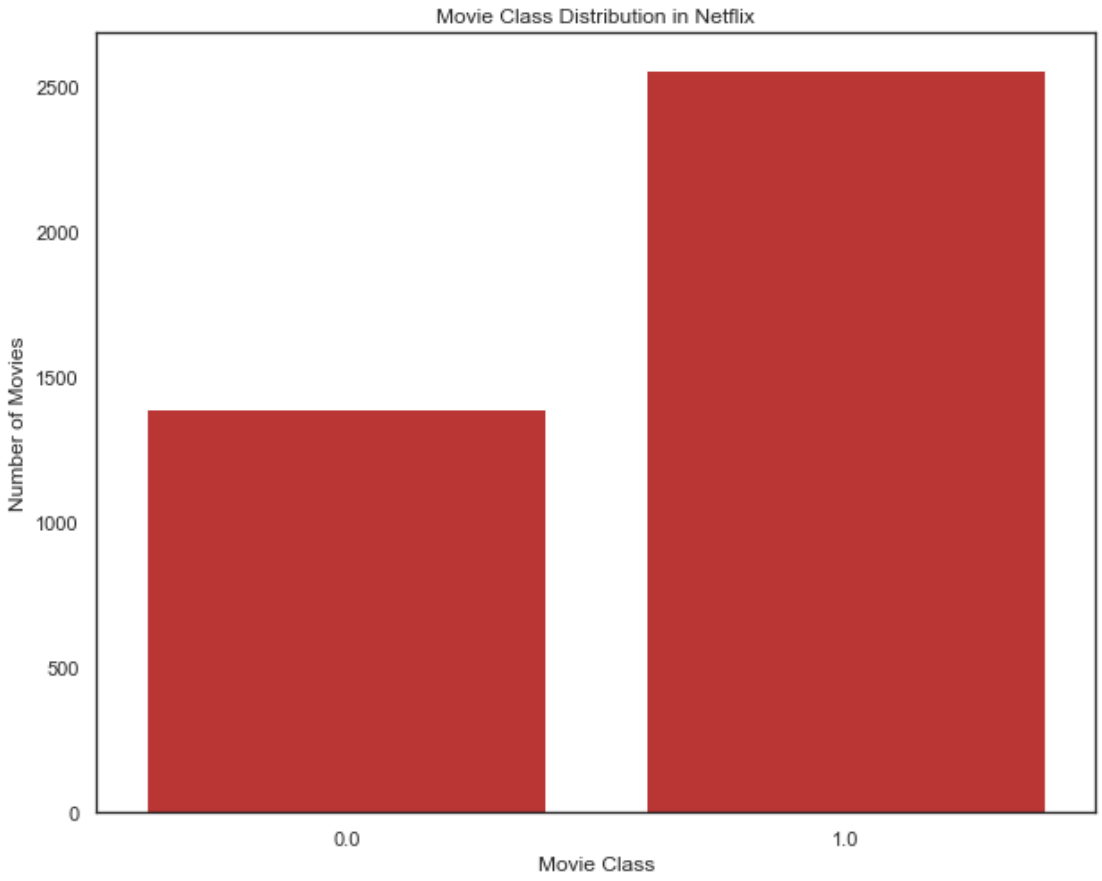



**Figure 9.** Distribution of target variable before oversampling

```
5-fold cross validation:

ROC AUC: 0.71 (+/- 0.01) [Logistic Regression]
ROC AUC: 0.90 (+/- 0.01) [kNN]
ROC AUC: 0.76 (+/- 0.01) [Decision Tree]
ROC AUC: 0.87 (+/- 0.01) [Random Forest]
```

**Figure 10.** Final AUC scores for the different models

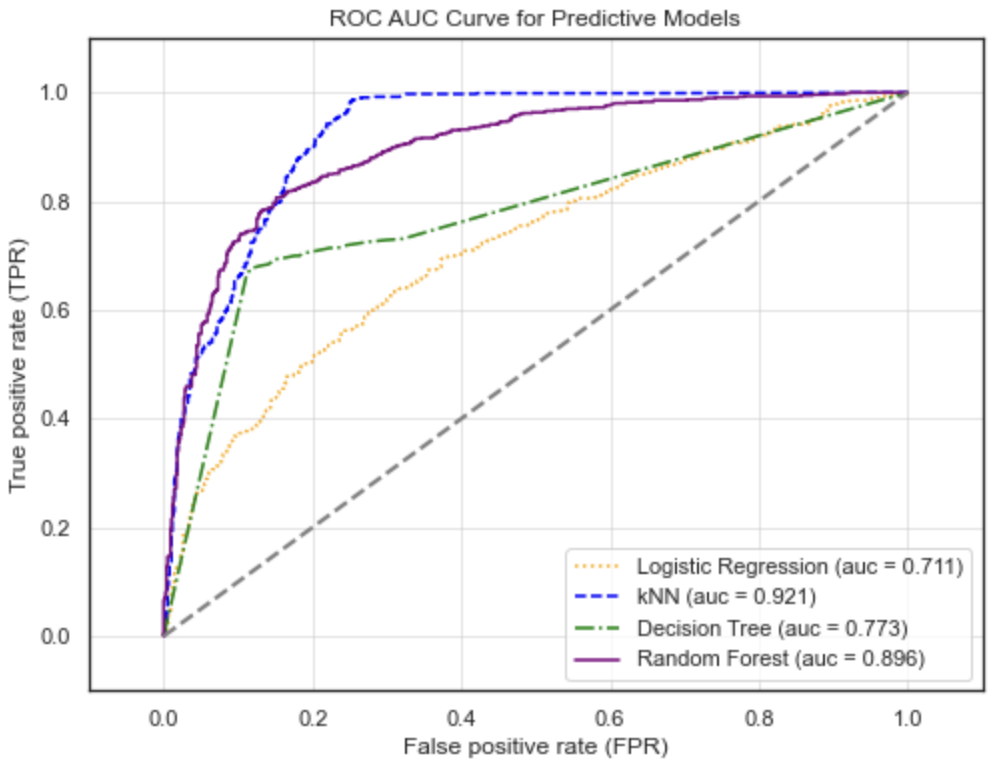



**Figure 11.** Final ROC curve for all models used

## Data Source

https://www.kaggle.com/shivamb/netflix-shows

https://en.wikipedia.org/wiki/List_of_Netflix_original_films

https://www.imdb.com