Final Project

Harlan Wittlieff

Data Science, Bellevue University

DSC 550: Data Mining

Dr. Brett Werner

March 5, 2022

Introduction

Many creative individuals dream of pitching their ideas for the next big television show to the major players at Netflix. These creators hope for the success of television series legends such as "Law & Order," "Game of Thrones," "Breaking Bad," "The Office", and the longest running show with thirty-three seasons, "The Simpsons."

However, before green lighting a new show, Netflix must evaluate key factors to attempt to predict whether the tv series will be a smash hit that their audiences will select for viewing out of Netflix's massive catalog. One of these factors that must be considered is the cost of production. When estimating the total costs of producing the television series, Netflix would likely require a solid prediction and understanding of for how many seasons the television series will run.

Through the meticulous analysis of historical television series performance data, Netflix may be able to predict for how many seasons their latest venture is anticipated to run. To achieve this goal, I begin with the Netflix Movies and TV Shows dataset on Kaggle. This dataset contains data describing movies in the Netflix library, their director, cast, release date, date added to Netflix, rating, duration, category, and a description of the movie.

Exploratory Data Analysis

Prior to building a model, the data and relationships amongst variables must first be explored. To begin the analysis, the distribution of the outcome variable will be explored. Figure 1 displays a box plot of the overall distribution of the number of seasons of the television series in the database set. As the chart demonstrates, a large subset of the television shows end after only one season with fewer shows reaching more than one season.

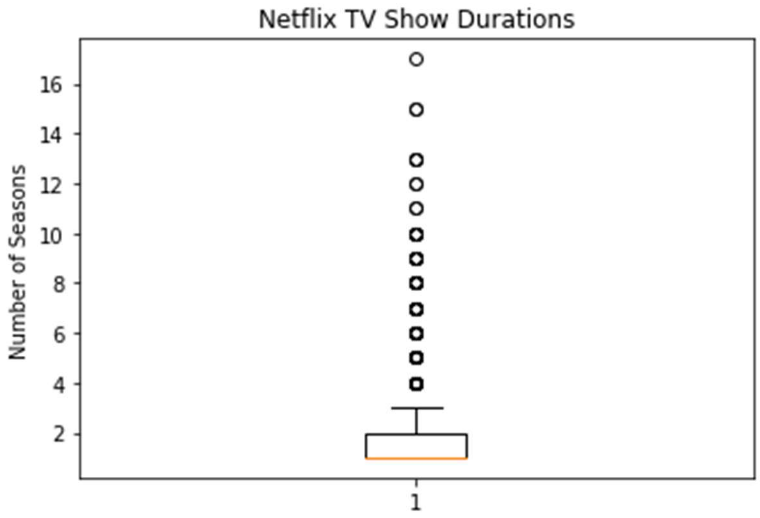

Figure 1: Netflix TV Show Durations

After gaining an understanding of the distribution of Netflix television show durations, the relationship between predictor variables and the number of seasons was explored. One of these relationships was that between the show's rating and duration.

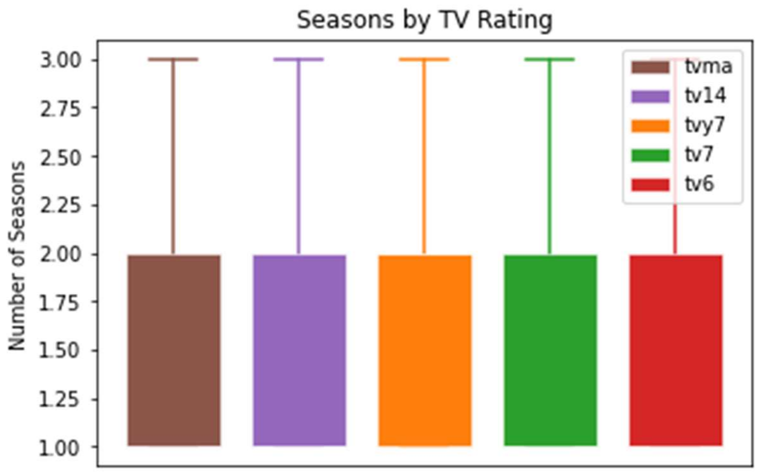

Figure 2: Seasons by TV Rating

As you can see in the figure above, the distribution of durations does not change relative to the tv show's ratings. From this chart, we can assume that the variable tv show ratings does not have a significant impact on the duration of the television series. Therefore, during the data preparation phase the ratings field will be removed.

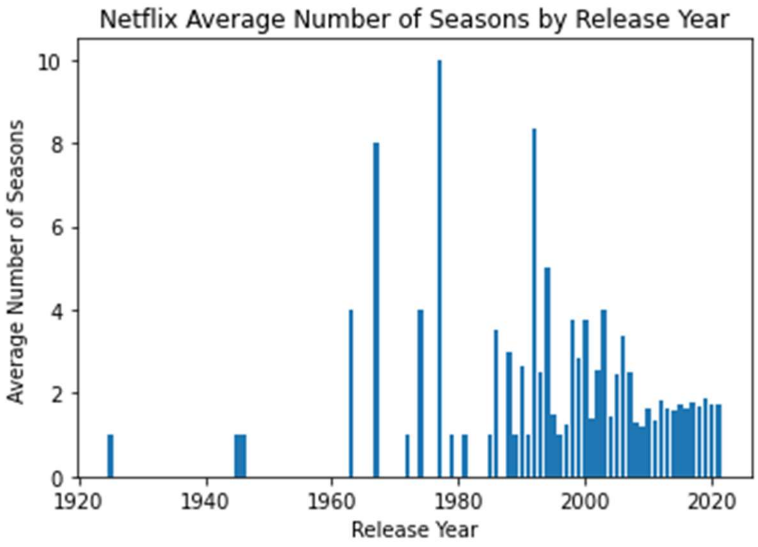

Figure 3: Netflix Average Number of Season by Release Year

Lastly, the relationship between release year and number of seasons was explored. The chart above displays the average number of seasons by television series released each year. The chart highlighted a few outliers, especially in the 1980s and 1990s. This is likely due to the low number of movies released during each of the years on Netflix. For the purpose of this data analysis, these values will be left in the dataset.

Data Preparation

During the data preparation phase, unneeded columns were removed. These unnecessary columns included show id, type, listed in, and rating. The original dataset also contained information about Netflix's library of movies. Since movie data was not required for the analysis, all rows where type equals movie were removed. Shows released in the current year that may still be running were excluded as their total duration is still unknown.

To allow for effective data analysis, the format of multiple fields needed to be transformed. For example, I translated the duration field to feature numerical seasons values. I then created a binary field based on the show duration with any shows lasting longer than one season were assigned a value of "1", while shows the lasted only one season were assigned a value of "0."

After these initial data preparations were completed, the descriptions needed to be cleaned. The first step in this description cleaning process was to remove punctuation and excess whitespace from the text. The text was then tokenized and any stop words were removed. Finally, I applied NLTK's PorterStemmer.

In final preparation for the model, the data was split into test and training sets. Eighty percent of the data was utilized to train the model. While the remaining twenty percent of the data was reserved for testing.

Model Building and Evaluation

I selected KNN to be the first model to be evaluated because the dataset contained descriptions and release dates. The assumption with the first model is that what is popular (more likely to have multiple seasons) will change over time and grouping like shows together may allow for a better prediction.

Initially four neighbors were chosen. However, through the utilization of a grid search it was discovered that sixteen neighbors provided the highest accuracy for the model. The accuracy for the first model is 67.4%. Next, a grid search was performed to compare the KNN model to a logistic regression model and a random forest model. The result of the grid search found that the random forest provided an equivalent accuracy of 67.4%.

To evaluate the model, a confusion matrix was created. This confusion matrix uncovered the flaw with the model. Due to the large number of Netflix shows that only last one season, in all but four cases the model predicted that the series would only last for one season.

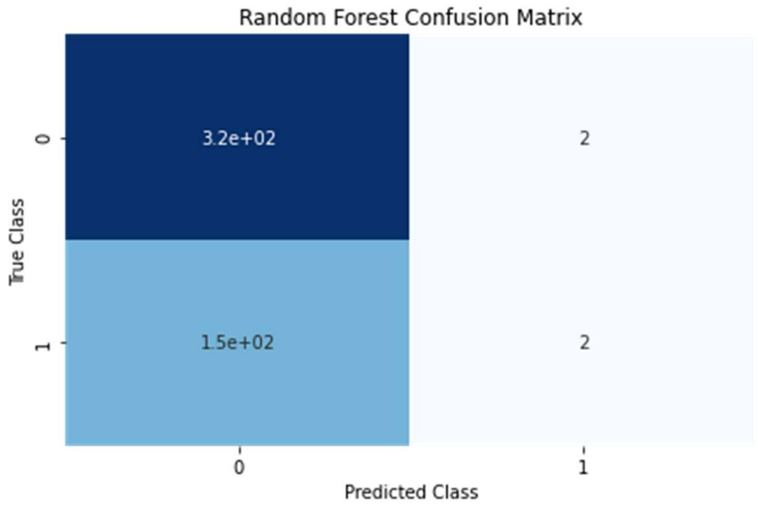

Figure 4: Random Forest Confusion Matrix

Conclusion

After analyzing the model, I would not recommend implementation because the model does not perform better than simply guessing that every series will last one season. Expanding the dataset to include television shows outside of the Netflix library may provide more opportunities for the model to learn from shows that last for longer durations. Another suggestion would be to build a model that takes into consideration the performance of television shows' first season with potential metrics such as IMDB ratings, meta scores, reviews, etc. to predict the series duration.

References

Bansal, S. (2021, September 27). *Netflix movies and TV shows*. Kaggle. Retrieved March 5, 2022, from https://www.kaggle.com/shivamb/netflix-shows