Movie Data Analysis

Ryan Ho & Ngawang Kyirong

1. Introduction and Problem Statement

The project described in this report investigates statistical analysis on audience ratings, genres, and how plot summaries can help predict genres. This project is based on the "Wikidata, Movies, and Success" topic: https://coursys.sfu.ca/2019su-cmpt-353-d1/pages/Project. In particular we are interested in several related questions:

1. Are various success criteria correlated with each other?
2. Do some genres have higher audience ratings than others?
3. Can we use plot summaries to predict a particular genre for a movie?

Here we describe the statistical analysis performed to examine the relationship between genre and movie success. Furthermore, we build on this by presenting the results of applying machine learning and NLP techniques to predict the genre of a movie from plot summaries.

The idea to look at genre and plot summary for our project was refined by doing some preliminary analysis on the data provided for the project. Initially, we were interested in looking at multiple factors affecting a genre of a movie such as the actor, director, etc. However we quickly found that for many of those factors (in particular actor and director), there wasn't enough data points for each actor and director. For example, when we filtered for actors that appeared in more than 40 movies to ensure normally distributed data, we were left with 0 data points. With genre however, there was enough data points that even when we filtered for genres having more than 50 movies/records, there was still enough data to do meaningful statistical analysis.

2. Methodology

2.1 Data Extraction, Cleaning, and Transformation

Extraction

The data provided for the project was extracted from WikiData, OMDb, and Rotten Tomatoes as described here: https://coursys.sfu.ca/2019su-cmpt-353-d1/pages/ProjectMoviesData. Our project is mainly concerned with the audience/critic rating data from Rotten Tomatoes and the genre and plot summary data from OMDb.

Cleaning and Transformation
Each of the data sources was read into Pandas DataFrames, and only the columns relevant to the project were kept. Some minor data cleaning was required, for example the rotten tomatoes audience_average values (out of 5) needed to be brought to the same scale as critic_average values (out of 10). Following this, the WikiData, OMDb, and Rotten Tomatoes DataFrames were merged on the "imdb_id" field so that movies could be matched with their corresponding genres, ratings, and plot summaries. This merged DataFrame was then used for subsequent statistical analysis, machine learning, and NLP techniques. One noteworthy transformation that was required was to "explode" the genre column from the OMDb data. Since the value for genre is given as a list, it needed to be split into several rows in order for data analysis techniques using genre to work.

2.2 Data Analysis Techniques

2.2.1 Correlation between success criteria

The project looked at the correlation between the following success criteria: "audience_average", "audience_percent", "critic_average", and "critic_percent" from the Rotten Tomatoes data. The calculations were done using the built in "corr" function from Pandas and the results were visualized using a correlation matrix from matplotlib.

Note that although profit data was available, we noticed that many records were missing this value. When we included profit, we had only ⅛ of the original data (795 records), so we chose to omit profit in this and subsequent analysis.

2.2.2 Analysis of audience average for different genres using ANOVA

ANOVA was performed on the different genres, looking at the audience_average from the Rotten Tomatoes data to see if there was any statistical difference. The stats.f_oneway function from SciPy was used for the ANOVA test. We selected only genres that had more than 50 records/movies to ensure that the audience_average values for each genre were normally distributed. This left us with 22 different genres. To visualize the results, the mean audience_average for each genre was plotting on a bar graph along with the associated standard deviation.

2.2.3 Analysis of genre audience average ratings over time

To look at the change in average audience ratings over time we averaged the audience_average for each genre by year and plotted the results using a line graph. To keep the amount of data on the graph at a reasonable amount, we only plotted the top 6 genres in terms of record count (drama, comedy, action, adventure, crime, and romance).

2.3 Machine Learning

Classification:
We modelled our exploration of predicting genres from plot summaries as a classification problem. However, the data included in the modelling is text. Therefore, our plot data needed to be transformed into a numerical representation of plot summaries. The first step to tackling this program is to transform the text into a representation that will be interpreted correctly by the model. Moreover, our modelling only took into account the 2 most popular genres: Drama and Comedy. We represented Drama with 1 and Comedy with 0.

Vectorization/TF-IDF:
The text analysis was started off with cleaning the plot summaries. This consisted of converting all letters to lowercase and removing punctuation. Moreover, our model can't interpret the text itself so it needed to be transformed. We turned the text into numerical feature vectors using a feature extraction package from scikit-learn. This process included removing meaningless common words within the text (stopwords: "the", "I", "and"...etc), calculating the tf-idf scores for each word, tokenizing the summaries, and transforming this into a sparse matrix which is appropriate for training models.

Comparing Models:

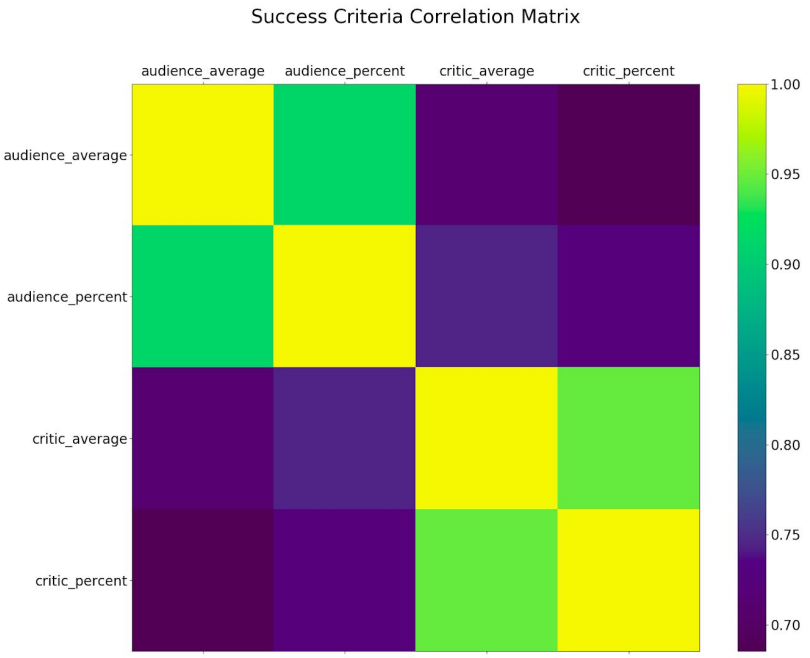
Using TF-IDF:

- **Multinomial Naive Bayes**
    - Prediction score: 0.6
- **Logistic Regression:**
    - Prediction Score: 0.6047
- **Support Vector Machine** (C=1.0, kernel='linear', degree=2, gamma='auto')
    - Prediction score: 0.5933
- **Nearest Neighbours** (k=100)
    - Prediction score: 0.6695
- **Neural Network**
    - Prediction score: 0.5743

3. Results

3.1 Correlation of Success Criteria

Looking at the correlation of success criteria found that they were generally correlated. Not surprisingly, the audience criteria was more highly correlated with the same holding true for the critic criteria. Results from the correlation matrix are shown below:



3.2 Effect of Genre on Movie Success

In answering the question "Do some genres have higher audience ratings than others?", we first performed an ANOVA test to see if there was statistical difference between the audience averages for movies from different genres. The ANOVA for 22 different genres is shown below:

F-value = 138.59422418659716
p-value = 0.0

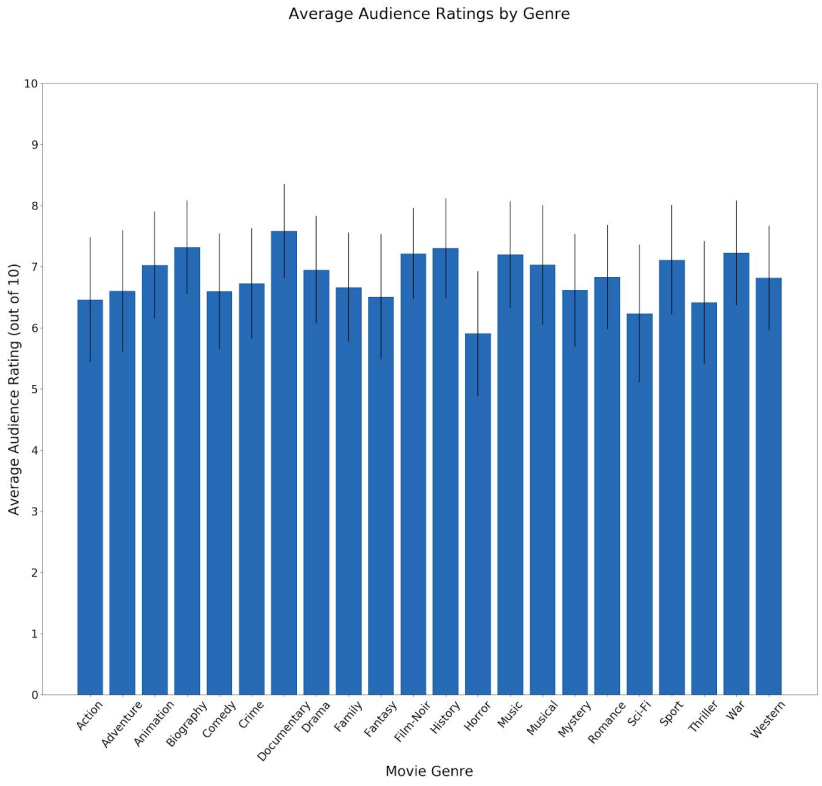
Unfortunately because of underflow, the p-value from the ANOVA test was 0.0. However since it is below the smallest representable normal number in Python (2.2250738585072014e-308), we can conclude there is a statistical difference between the audience average ratings for different genres. The F-value confirms this as well, since it is well above 1.0.

We also wanted to take a look at the top 6 genres in terms of record count in the data (drama, comedy, action, adventure, crime, and romance). This was done to get a more manageable

amount of genres to do further analysis. For example, as described in section 3.3, we looked at the change in the genre ratings over time, which would be too much to do with 22 genres. As shown below, results of the ANOVA indicate that with the 6 genres, there were statistically significant differences in the audience averages:

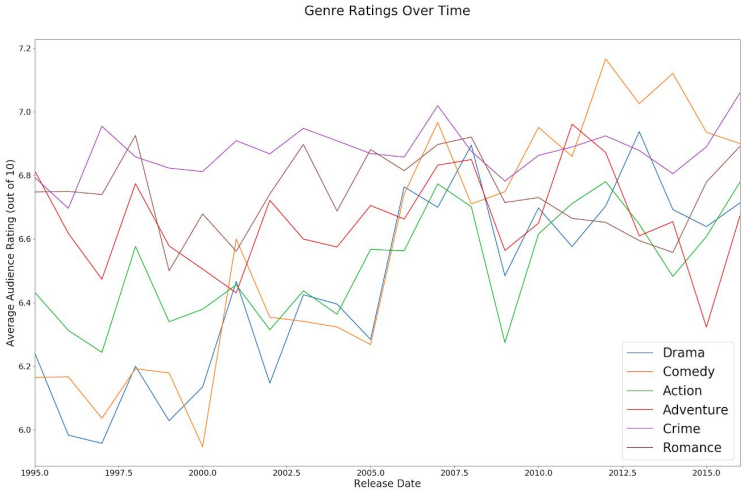
F-value = 100.42221368144865
p-value = 2.6630253898797657e-104

To visualize the differences in audience averages, we graphed the mean audience average rating for each genre. As shown in the graph below, the audience average ratings for each genre were in the range of 6-8 (out of 10).



3.3 Change in Success of Genres over Time (1995 - 2017)

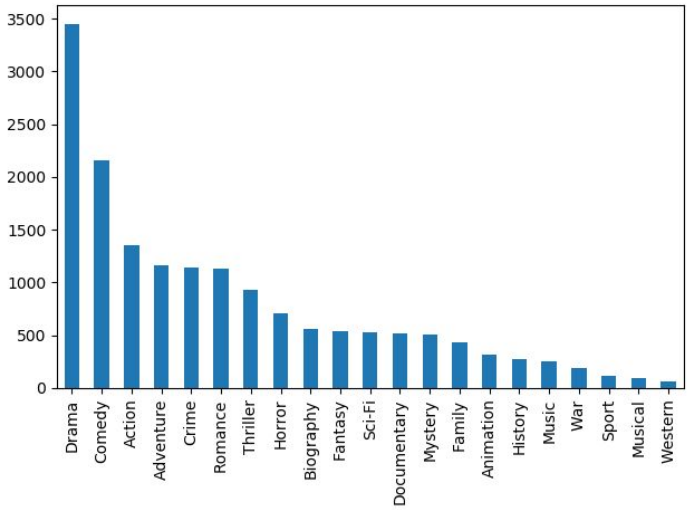
Building on the results from 3.2, we wanted to look at whether there were any changes in genre success over time. We chose to look at the top 6 genre by record/movie count so that the amount of data would be manageable. A plot of the mean audience average by year is shown below. Interestingly, some genres such Drama and Comedy showed a clear upward trend in the audience average ratings over the time period examined.



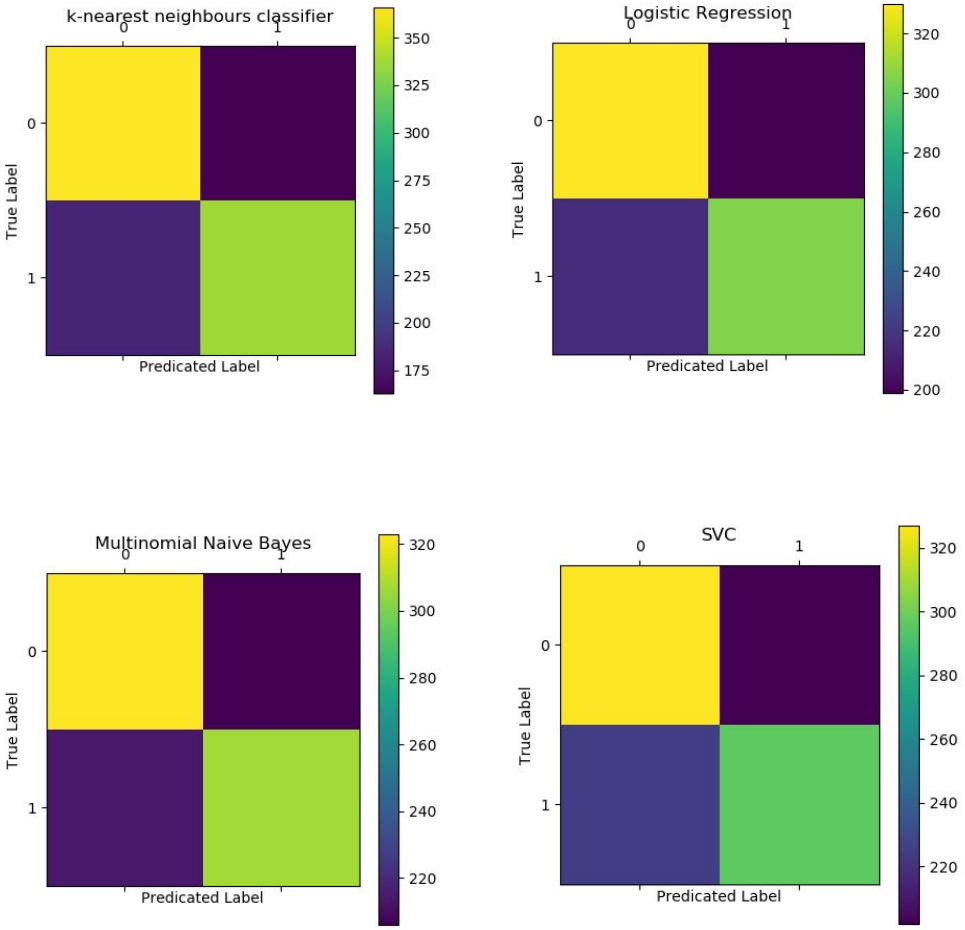
3.4 Building predictive models

The models used to try and predict genres from plot summaries are: Multinomial Naive Bayes, Logistic Regression, SVC, k-nearest neighbour classifier, and a neural network. After training and testing the models, the k-nearest neighbour classifier outperformed the rest of the models. Although it's prediction score still wasn't great (0.6695), it did much better than the rest.

When trying to train a Multinomial Naive Bayes model, the predictions were only of the genre that had the most training data. Therefore, the data was trimmed down to 2100 data points each for Drama and Comedy. This was because the dataset was unbalanced as shown below which influences that outcome from the models.



Each predictive model except for the neural network is represented as a confusion matrix below. The matrix describes which predictions were correct and incorrect for each model. This allows us to see the frequency of which predictions were correct or incorrect which may suggest some tinkering in the models.



4. Limitations

4.1. Data Limitations

One limitation we experienced was with the profit data from Wikidata. This would have been a great metric to measure movie success, however since so few records actually had this value, we couldn't include it in our analysis. An extension to this project could look at other sources of data that has more complete values for profit and see if the movie genre is associated with higher earnings.

Another limitation we experienced was the unbalanced data. This caused some models to be bias towards the data points that dominated in sample size. Therefore, we balanced out the dataset and trained on an equal amount of data. Moreover, I think that a more in depth summary for the movies would've helped differentiate each genre.

Project Experience Summaries

Ryan Ho

My main contribution to the project was the data extraction, cleaning, and transformation step. I took a preliminary look at the data to see what was available and what sort of data cleaning and transformation needed to be done. To accomplish this task, I used what I had learned about Pandas to get the data in a usable form. As a result of my efforts, the project had the data required to carry out the statistical analysis, machine learning, and NLP. I was also involved in the statistical analysis to answer the first two questions presented in the introduction. I used various statistical tools learned throughout the course such as ANOVA, to find differences in audience ratings for different genres of movies. I also used plotting software (matplotlib) to visualize the results and get an idea of what the data trends were. This contribution to the project allowed us to answer whether or not genre plays a role in the success of a movie.

Ngawang (Choenden) Kyirong

My main contribution to the project was forming ideas for the project, analysis of ratings over time, choosing and training the predictive models, and analyzing the results. In order to accomplish the tasks I had to study and use the statistical methods described in class, understand their underlying mechanisms, and test appropriately. Moreover, in order to execute the analysis in python I had to be comfortable using the libraries learnt in class such as: pandas, scikit-learn, numpy, and matplot. Furthermore, in order to display the results in a meaningful way I had used maplot and choose appropriate figures that would describe important information.