**Data Mining: Movies**

Team Members:

Pierre Casco

Valerie Fowler

Elia Kostyrka

IST 707, Data Analytics

Syracuse University (SU)

July 2019

**Introduction**

The movie industry revolutionized how people found entertainment. The industry as it is today started to take shape in the 1890’s. These basic “films” were less than a minute, with no sound, and in black and white. Many of them consisted of single frame or scene. During what became known as the silent film era, movies were accompanied by an orchestra or individual creating sound effects for the movie. As technology improved, longer films could be made with sound and animation. Actors began to receive screen credit for their roles in the early 20th century. By the 1920’s and 1930’s the cinematic industry was in full swing with actors being paid for roles and technology continuing to improve.

By the 1950’s, a threat to the cinematic industry was gaining popularity, television. The popularity of television lead to numerous bankruptcies and closures of movie theaters. In return, cinema made several adjustments with newer technology, larger screens, and an increase in marketing. Some of the greatest films ever made were created in this time. Something unexpected also took place. Though television was competition, something surprising happened. Television promoted cinematic features instead of threatening it. The next decade or so brought a decline in the traditional studio system. In Individual producers and production companies were now becoming the norm.

A major boost came several decades later with the invention and affordability of VCR’s. This opened up an entirely new revenue stream. Films were now able to be seen in theater and anytime the viewer would like. Studios in return poured more into quality, marketing, and star power-both in front of and behind the cameras. Star power lead to larger box offices and were more in demand to create the next big thing. Technology, again, also aided with this. The transition from actual film to a digital medium and the transition from VHS to DVD had great impact.

The film industry has evolved into a big business. Films now cost millions to make and don’t always get such a return. Movie genres tend to ebb and flow. There is also a rating system based on viewer age in relation to movie content. Gross revenue is calculated within the United States and Internationally. Production companies are a dime a dozen. Numerous actors have their own production companies that allow them to make the films they want, not the other way around. In an ever changing and unreliable industry, one thing remains constant, viewers hold all the power despite the current perception that viewing mediums cannot co-exist (Zeitchik, 2018). Therefore, with a variety of viewing options it is critical that the movie industry identify areas that hold higher entertainment value in order to continue to survive (Buchanan, 2019).

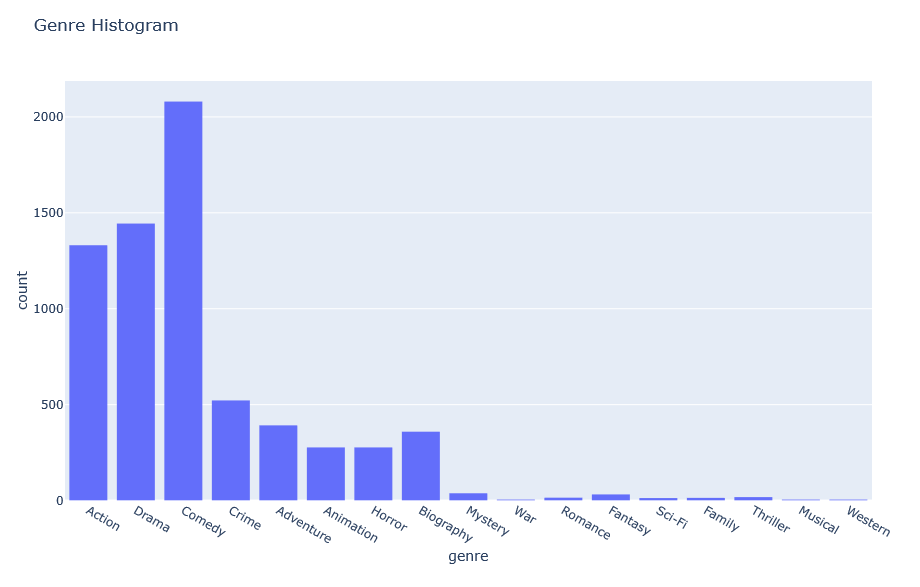
**Analysis**

**The Data.** Two movie data sets pulled from Kaggle were merged to produce one data set. The movies data set contained eighteen (18) attributes and 6,820 observations (Kaggle Open Datasets). There were various data types in the data set, numeric for the budget, gross revenue, and runtime, nominal for the genre, rating, score, success overseas, and country of origin, and text for the movie titles, starts, directors, writers, production company (Appendix A). Limitations for the data included a limited data, later date range for movies (1996-2016), budget, and gross revenue overseas.

Several steps were taken to preprocess the data for various analysis. Five of the eighteen original attributes were used for analysis, rating, genre, IMDb scores, gross revenue, and runtime. In addition, the primary attribute used to predict on whether a movie was successful was created by annotating if a movie had a 7+ IMDb rating and noting it in binary form (yes/no). New attributes were also created that noted whether a writer, director, star, or production company was a top performer by identifying those that had the most (top 97th percentile) movies with a 7+ IMDb user score rating (yes/no binary form). Data was also discretized (IMDb scores, gross revenue, and runtime).

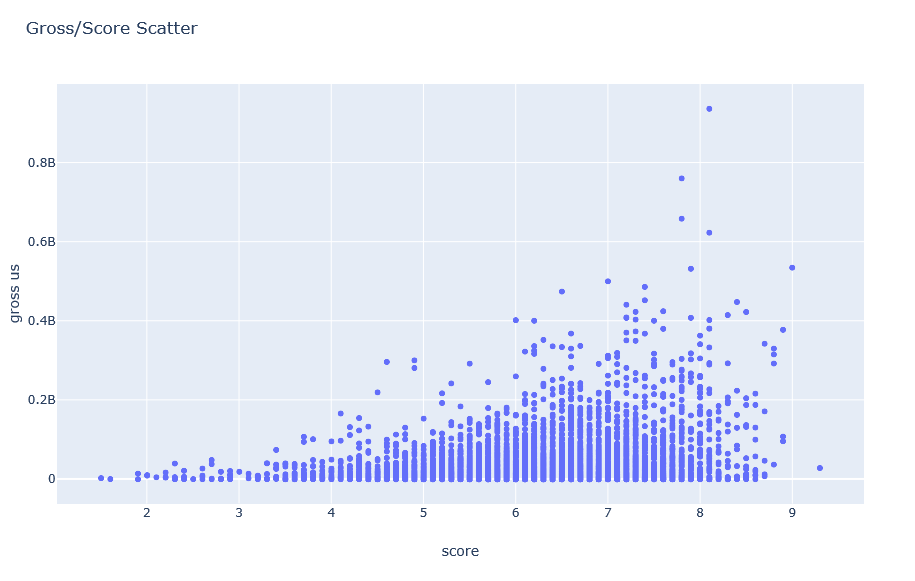
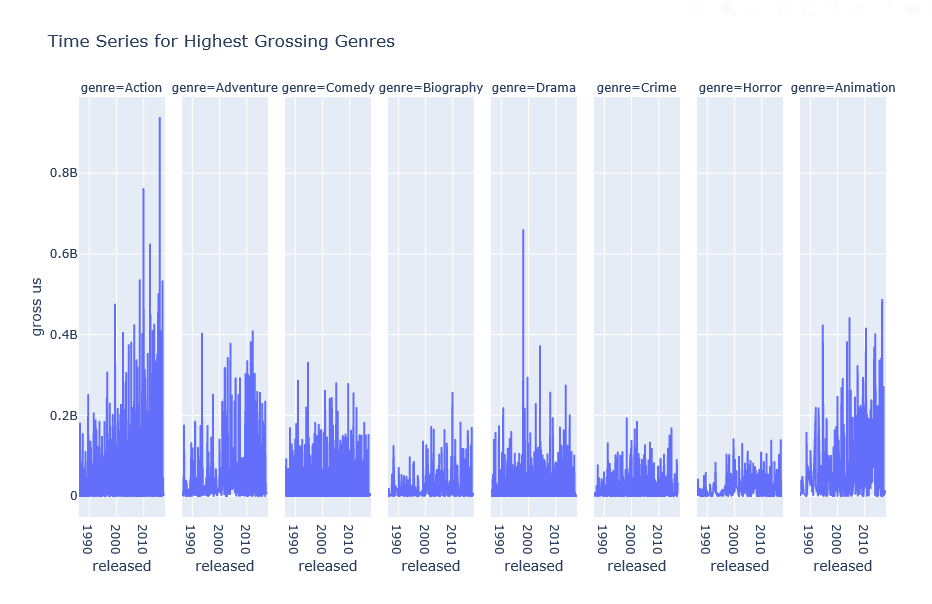
**Exploratory Data Analysis.** Various areas were explored for initial insights. Movie availability by genre was explored using a histogram to count the different genres in the dataset (Figure 1). The histogram showed that comedies made up a large chunk of the dataset, followed by drama and then action movies. The other genres did not contribute as much as the other three genres.

Figure 1. Movie Availability by Genre



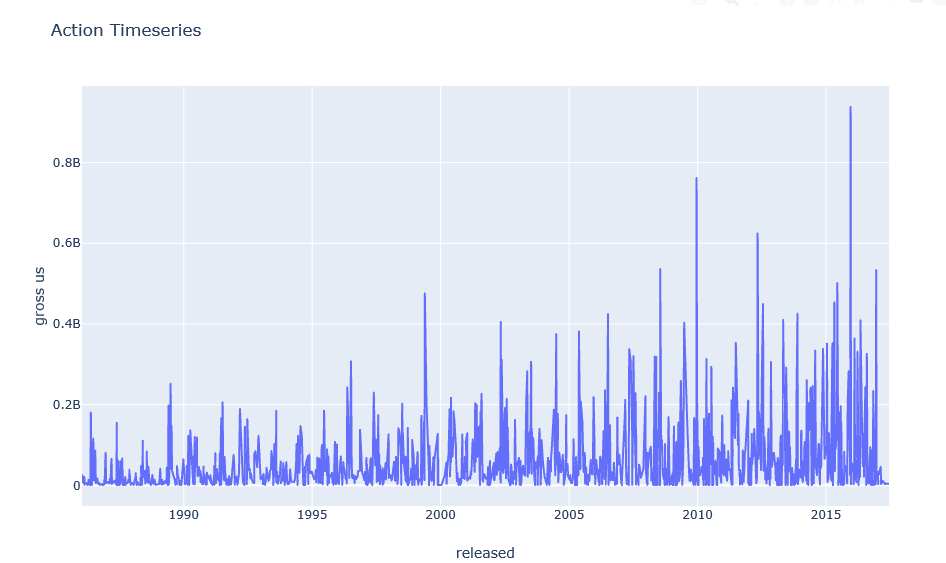
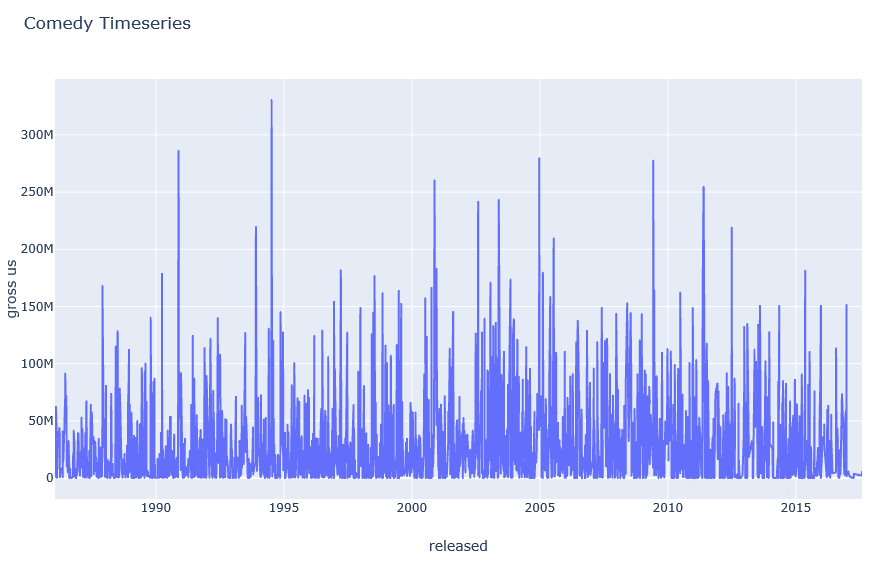
Secondly, IMDb user rating and gross revenue was explored using a scatter plot to see if there was a correlation between money made and user rating (Figure 2). As expected, the higher rated movies made more money than the lower rated movies. Understandably higher rating can create buzz, generating interest and driving box office sales up. Thirdly, yearly gross revenue by genre was explored using a time series plots for the different genres to see if there was a cyclical nature for when movies are released (Figure 3). There is a bit of seasonality. The term “summer blockbuster” comes to mind, meaning a movie with a big budget is released during peak summer months.

Figure 2: IMDb Rating and Gross Revenue Figure 3. Yearly Gross Revenue by Genre

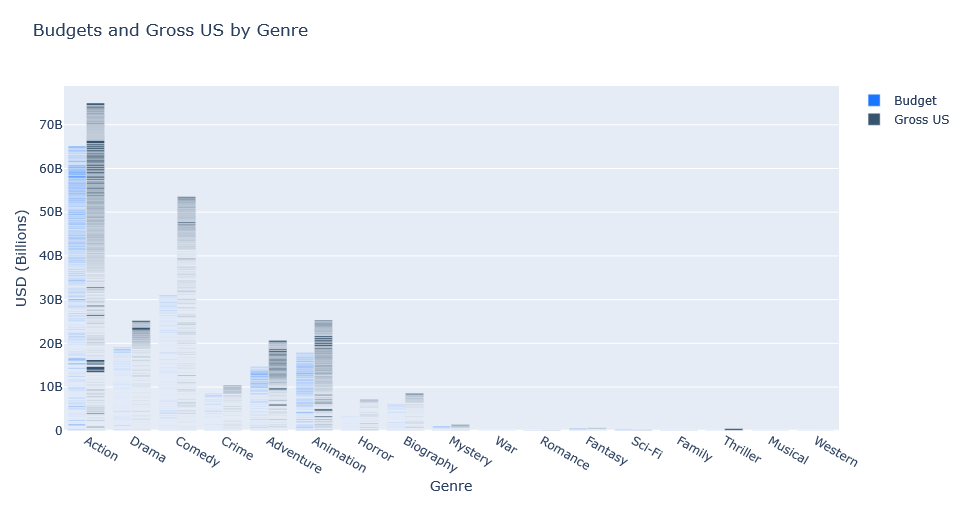
Fourthly, action and comedy time series plots were further explored to see the seasonality in the movies by studying the different peaks (Figures 4 and 5). Of note, as time moves forward, action movies made more money while comedies are relatively steady throughout the peaks on the plot.

Figure 4: Action Movie Time Series Figure 5: Comedy Movie Time Series

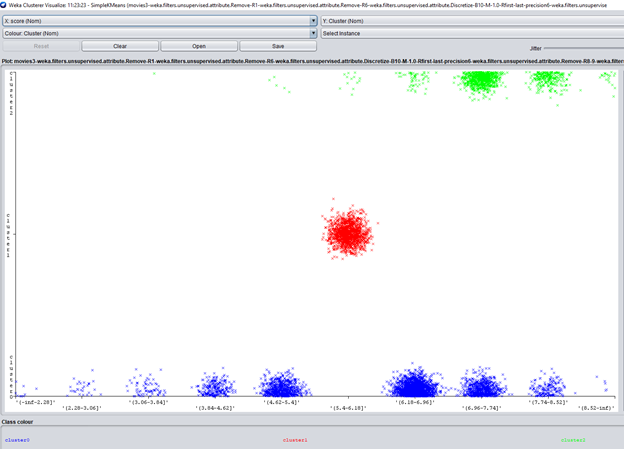
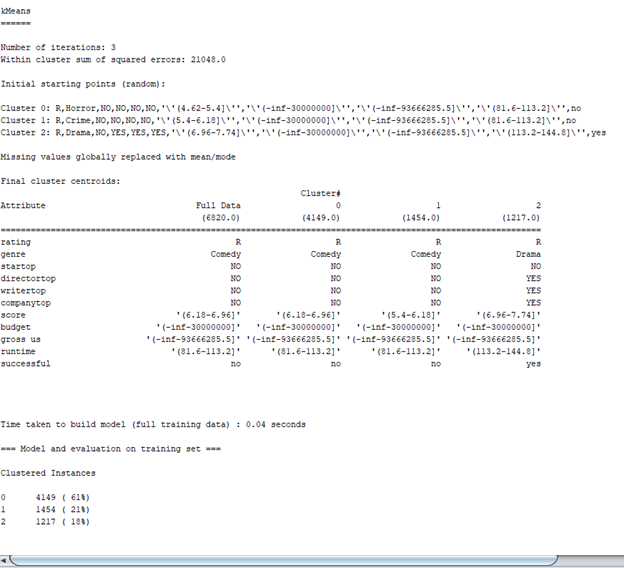
Lastly, a budget and gross revenue plot was created to explore budget versus revenue for different genres. Nearly all movies generated a profit; however, comedies made the most money since the budgets were lower and the profits were higher. An ideal mix would likely be an action comedy that leans more toward the comedy side which keeps the budget down.

Figure 6: Movie Budget Revenue by Genre



**Cluster Analysis.** Cluster analysis was conducted using Weka (parameters – SimpleKMeans clustering, Euclidean distance, Manhattan distance, 3 clusters, 500 max iterations, and 10 random seeds) to identify clusters in the data associated with successful movies. For both the Euclidean and Manhattan distances three clusters were formed, two of which had unsuccessful movies, and one with successful movies as the mode. Data preprocess include discretizing all numeric values. Attributes used for analysis included genre, top performing stars, directors, writers, and production companies, gross us revenue, runtime, and the successful category. Cluster 2 showed the mode successful attribute was “yes”, so this is the cluster of focus. The mode for this cluster was rated R movies, dramas, no top stars, but included top directors, writers, and production companies. The Average IMDb scores were between 6.96-7.74 and had longer runtimes between 113-144 minutes (Figures 7-12).

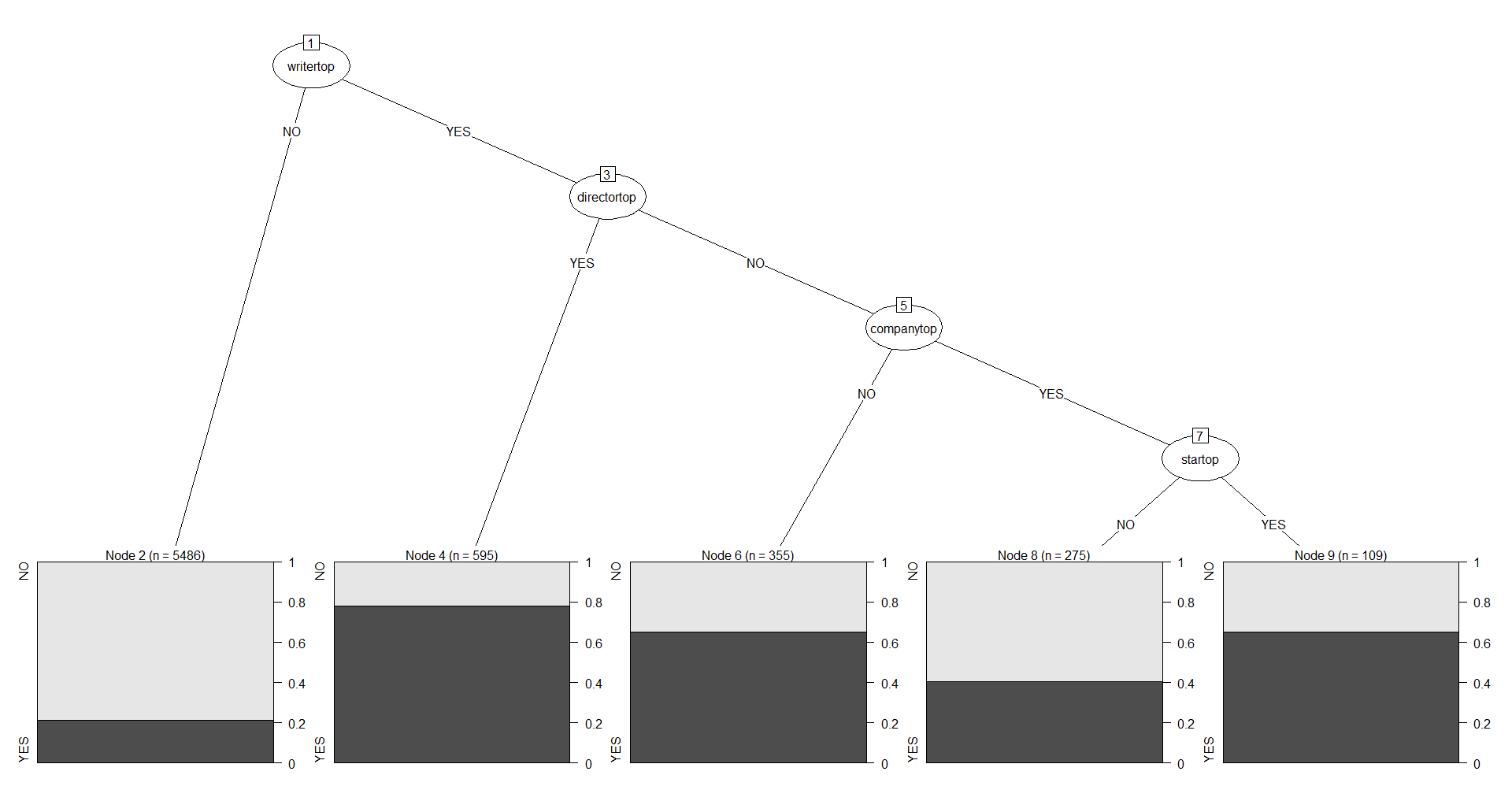
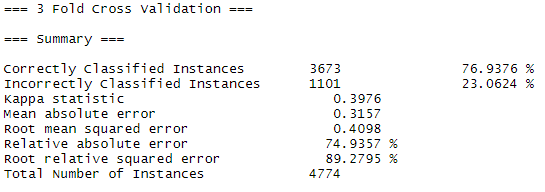
Figure 7. Cluster Analysis Results Figure 8. Clusters and IMDb scores



**Classification Analysis**. Decision tree and Naive Bayes were used for classification to train the models to predict the success of a movie based on various attributes. In addition, three algorithms were compared for classification analysis to identify a model that could provide higher accuracy, SVM, k-NN, and Random Forest.

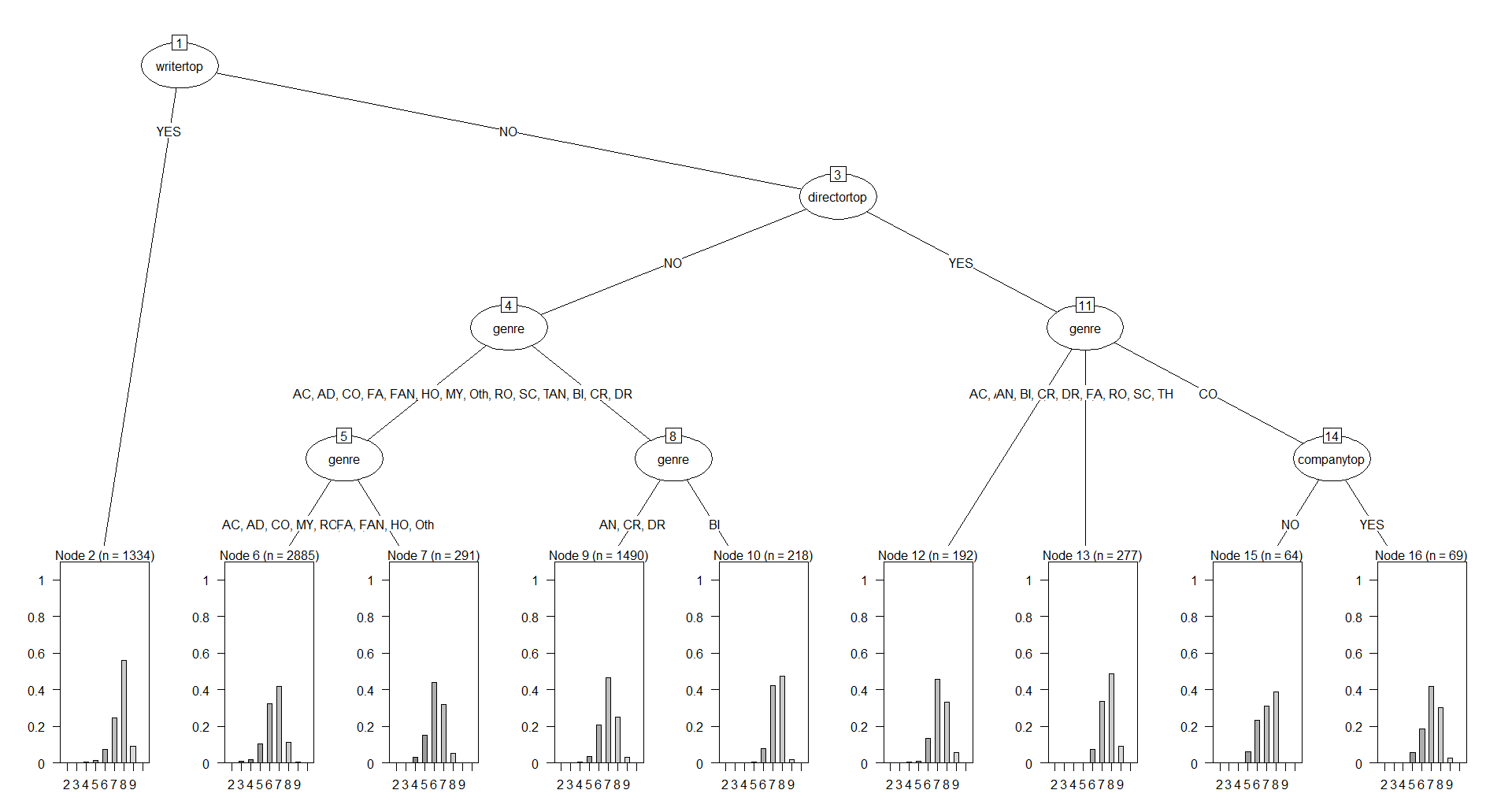
**Decision Tree Model.** Two additional preprocessing steps were taken to run a decision tree. Category names of the genre category were listed as abbreviations for easy reference in the results (e.g., Drama became DR) and an additional “score level” column was added denoting the score level for a movie. Ratings up to a rating were given the higher rating, e.g., rating 0-1 were given a level 1, rating 1.1-2 were given a level 2, rating 2.1-3 were give a score level 3, etc. A decision tree was built in R using J48 (parameters – unpruned tree, 2 number of instances per tree, .5 pruning confidence, 3 numfolds, and 1 seed) to determine the percentage of correctly classified instances. Given the attributes rating (e.g., PG, R, etc.), genre, top performing starts, writers, directors, and production companies, and the successful category, a decision tree was built that provided 76.94% accuracy on the test data (Figure 13).

Figure 13. R J48 Decision Tree Results



A scaled down decision tree using c50 in R with only the genre, top performing stars, directors, writers, and production companies showed that using a top writer produced more movies rated at higher scores. The decision tree also showed that only the genre “comedy” was segmented out to predict score level. Comedies that used a top director produced more movies with score level 8. Action movies with a top director could predict score level as well (Figure 14).

Figure 14. R c50 Decision Tree Results



**Naive Bayes Model.** A Naïve Bayes model was built in R using the parameter laplace 1 to determine the percentage of correctly classified instances. Given the attributes rating (e.g., PG, R, etc.), genre, top performing stars, directors, writers, production companies, and the successful category, the model had the ability to predict unsuccessful movies by 88.8% and decreased to 54.5% success of predicting the yes cases on the test data (Figure 15 and 15a).

Figure 15: Naïve Bayes Confusion Matrix 1

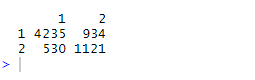
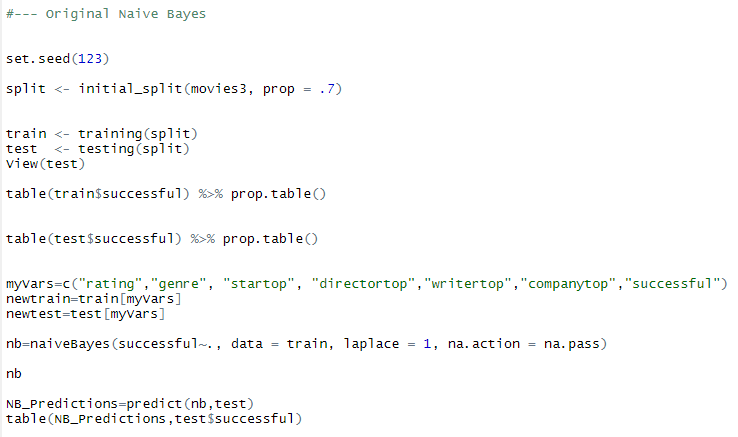


Figure 15a: Sample R J48 code



A second Naïve Bayes model was built in R using e1071 (parameters - 3-fold CV) to determine the percentage of correctly classified instances and was able to predict unsuccessful movies by 89.8% and decreased to 48.8% success of predicting the yes cases on the test data (Figure 16 and 16a).

Figure 16: Naïve Bayes Confusion Matrix 2

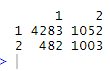
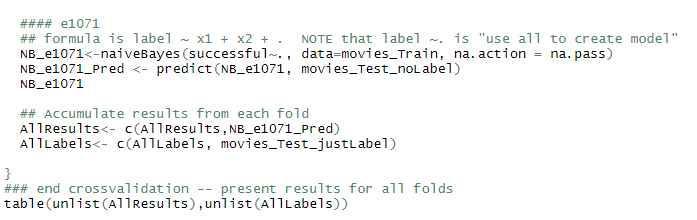
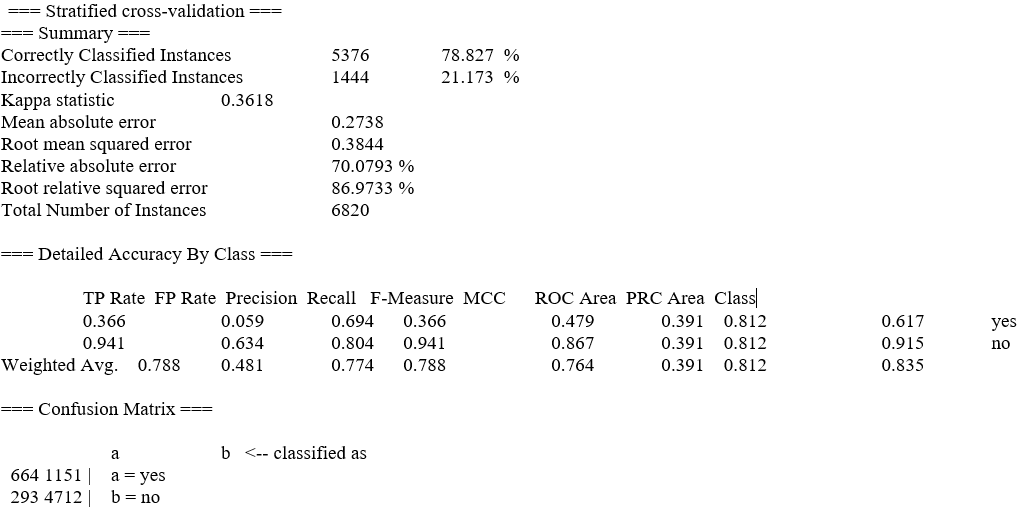


Figure 16a: Sample R e1071 code



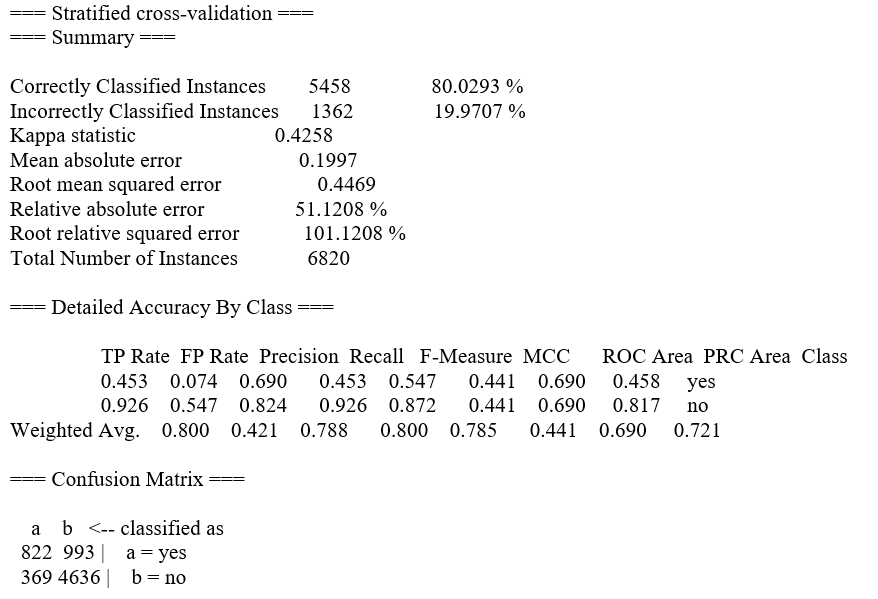
**k-NN Model.** Three algorithms were compared for classification analysis to identify a model that could provide higher accuracy, k-NN, SVM, and Random Forest. The scores were removed, and the successful attribute was kept in order not to skew the results. The first being the k-NN model. A k-NN model was built in Weka using cross-validation of 3 folds, different k nearest neighbors (0.5, 1, 5, 10), and both distance weighting of weight by 1-distance and weight by 1/distance. Using the successful attribute as the target, the best results included the one with a k-NN value of 10 and weighted by 1-distance. The k-NN model had the ability to predict unsuccessful movies by 80% and decreased to 69.4% success of predicting the yes cases on the data (Figure 17).

Figure 17. k-NN Model Results



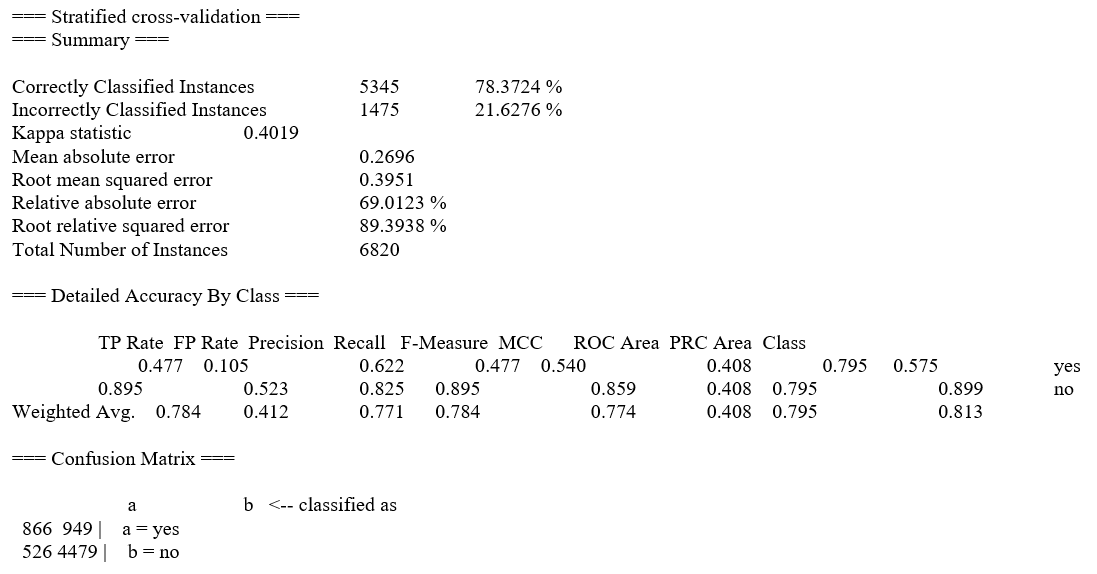
**SVM Model.** An SVM model was built in Weka using cross-validation of 3 folds, different cost amounts (0.3, 0.5, 0.8 1.0), and three kernels were used, the Poly kernel, RBF kernel, and normalized kernel. Using the successful attribute to predict on, the best results included the one with a cost value of 0.5. and the poly kernel. The SVM model had the ability to predict unsuccessful movies by 82% and decreased to 69% success of predicting the yes cases on the data (Figure 18).

Figure 18. SVM Model Results



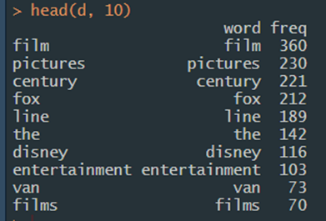
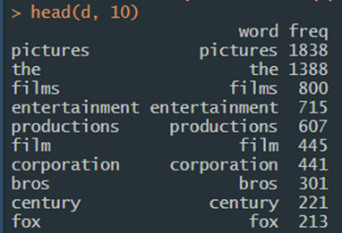
**Random Forest Model.** A Random Forest model was built in Weka using cross-validation of 3 folds and different number of trees (2, 5, 10, 15). Using the successful attribute as the target, the best results included the one with 15 trees. The Random Forrest model had the ability to predict unsuccessful movies by 82.5% and decreased to 62.2% success of predicting the yes cases on the data (Figure 19).

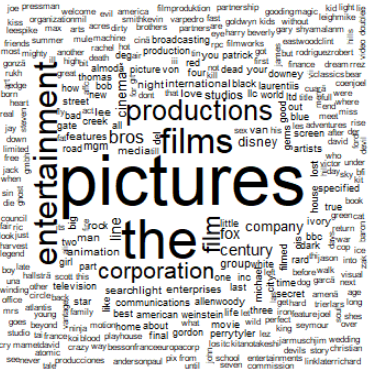
Figure 19. Random Forest Results



**Text Mining.** Initial text mining steps were taken, including the development of two word clouds to show higher word frequencies. The first contained all attributes from the original data set (Figure 20). The second removed all numbers and years (Figure 21). This was done to determine if a particular year had more projects made or were more successful.

Figure 20. High Word Frequencies, All Figure 21: Number and Years Negated

**Results**

The k-NN training model was able to more accurately predict successful movies given the attributes rating, genre, top performing stars, directors, writers, and companies, and the target successful category (IMDb rating of 7 or more), 69.4% compared to the SVM model (66.7%) and the Random Forrest (62.7%) models. Issues included preprocessing text data for text mining due to the large variety of different actors, writers, directors, and production companies. One workaround was to classify those categories by identifying high performers. Not having access to the budget data for all movies (less than 10%) prevented deeper analysis into return on investment versus solely gross revenue. Decision trees and clustering worked well to display the results; however, the supervised learning techniques would be most beneficial to train on new movie data not part of the data set used (movies produced 2017-current).

The word clouds confirmed what was found in the previous analysis. Production companies, writers, and directors are the common denominator to having a successful movie. The larger, more established production companies like Searchlight, Fox, Century, and Warner Brothers are among the most frequently found in the data set. Writers and Directors such as, Tyler Perry, Woody Allen, and Clint Eastwood were among the top performers. No particular year stood out to drastically different from any other.

**Conclusion**

Data mining the movie industry data revealed that not one sole attribute could predict the success of a movie. Specific findings stemming from the analysis of the movie industry data include 1) areas that could provide entertain value included top directors, writers, and production companies, 2) profitability of specific categories included comedy movies which had the best return on investment (e.g., budget versus gross revenue), and 3) models can be used to predict the future success of movies to a degree but more information would provide better predictions.

While some results were surprising, others were not. Comedies having the largest return on investment makes sense. Comedies make people laugh and allow them to forget any problems or struggles they are facing. A comedy may be seen as more of an escape. Comedies are also cheaper (mostly) to make versus an action, animation, or war movie. A cheaper budget in combination with a large gross revenue will always lead to a greater return.

Surprisingly, specific actors do not guarantee success. Top directors and writers are more influential than anything else. In some cases, the writer and director were the same individual. With more data, it would be interesting to see if this trend applied in the overseas market. Production companies also play a large role in the success of the movie. This, again, makes sense. Larger companies are able to support numerous projects at once. They have more access to funding and can pull in the best writers and directors for projects.

Obtaining this information can help the movie industry as a whole. The results can be further analyzed by the top production companies. This will show specifics on what genres are more successful for the company and therefore what to focus on in the future. It will also show where they are lacking in comparison to others. Having the strengths and weaknesses of your company and your competition can shape decisions and projects. Smaller production companies will also benefit. They will have the potential to see where the larger companies are lagging and attempt to find their niche there; or know what to stay away from.

Overall, as the movie and entertainment industry continue to change, so will the trends. For example, the market will become oversaturated with comedies and people will grow tired of them or the movie plot will start to become familiar. This will lead to a new trend, say, action/ superhero movies. They will then control the box office until they don’t, and the next trend will start. The entertainment industry is all about what is coming next, not what is happening now. The most popular concern moving forward is the decline of the movie industry due to streaming technology. The movie industry has dealt with technological advancement before and came out just fine. Will they be able to do it again?

References

Buchanan, K. (2019). How Will the Movies (As We Know Them) Survive the Next 10

Years? Retrieved from https://www.nytimes.com/interactive/2019/06/20/ movies/movie-industry-future.html

Kaggle Open Datasets (2019). Movies Industry, Three decades of movies. Retrieved

from https://www.kaggle.com/danielgrijalvas/movies

Kaggle Open Datasets (2019). Movies Industry, Three decades of movies. Retrieved

from https://www.kaggle.com/prathyushavudatha/movie-analysis

Zeitchik, S. (2018). Is Netflix Killing the Movie Theater? Not So Fast. Retrieved from

https://www.washingtonpost.com/business/is-netflix-killing-the-movie-theater-not-so-fast/2018/12/24/7a16dbf8-037a-11e9-8186-4ec26a485713\_story.html

Appendix A

Data Dictionary

|  |  |
| --- | --- |
| **Term** | **Description** |
| budget | the budget of a movie. some movies don't have this, so it appears as 0 |
| company | the production company |
| country | country of origin |
| director | the director |
| genre | main genre of the movie |
| gross | total revenue of the movie |
| gross overseas | revenue of the movie overseas |
| gross us | revenue of the movie in the us |
| more success overseas | higher revenue overseas than in the us |
| movie title | name of the movie |
| rating | rating of the movie (R, PG, etc.) |
| released | release date (MM-DD-YYYY) |
| runtime | duration of the movie |
| score | IMDb user rating |
| star | main actor/actress |
| votes | number of user votes |
| writer | writer of the movie |
| year | year of release |