## **Machine learning and predective analysis with movie data**

Author: Jinhang Jiang

## Introduction

This machine learning project is the final project for the Business Analytics Capstone Class (BSAN 480, INSTR: Chris Claterbos) at the University of Kansas School of Business. The purpose of the project is to study patterns and derive insights from the Movielens data and IMDB data to predict the performance and the rating of a movie in the future.

## Scope of Work

For the main concentration, the project will focus on performance analysis and prediction. I firstly cleaned, created, and then analyzed the given tables (Appendix A-1 & A-2) to understand patterns and stories that otherwise may have gone untold. Secondly, I created a tag table with 7,075,944 records (Appendix A-3) and did a sentiment analysis with the data. I also created some dashboards to visualize the significant findings. Finally, I tried and compared several different predictive models to forecast profitability performance, as well as the ratings.

I used SQL, R, and Python as the programming languages. I used software, packages and algorithms, such as pandas, NumPy, Node2Vec, Spacyr, WordCloud2, Sentiment Analysis, NetworkX, and K-means for analysis. RStudio, Jupyter Notebook, Oracle Developer were used as IDE. Oracle Analytics Cloud was used to create dashboards.

## Data

* cleandata.csv

This table contains 17,071 English-speaking movie records from the past 114 years, including some important features we are looking for, like **movie ID, revenue, budget, release data, cast information, ratings, genres, and relevance scores.** Full list of attributes can be found at Appendix A-10. Excluding the tag table, all the other tables that focused on some specific feature analysis were generated from this table (Further details in Appendix A-4).

I transformed all the genders to three categories: “M”, “F”, and “N” based on the number 2, 1, and 0.

I did a log transformation for budget and revenue, respectively, since they are too big in numbers and may create outliers when we develop predictive models. I broke the column genre into a dummy matrix for both model development purpose and network analysis.

Since some movies do not have many votes (perhaps because they did not have a social media account or something) while the other of them have a huge number of votes, it would not be reasonable considering it directly as a variable in the model. So, I also did a log transformation to both ML\_RCOUNT and NUMVOTES.

There is a new column called **PERFORM**. I created this one to indicate if this movie is rated above or equal to 7 (since the quantile test shows the top 25% cuts at 7) IMDb score by 1s and 0s. I will use this column to do a logistic regression analysis.

* genre\_matrix.csv

This is the transformed dummy matrix of the genre. It has 17,071 rows and 21 columns. Column 1 is the movie ID, and column two 2 to 21 are all the types of genres that are listed in the original table. The output looks like the following:

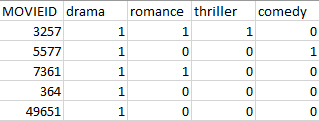


Figure 1 - Partial View of genre\_matrix.csv

In Figure 1, we can see that the table is populated with 1s ("equal to 1" means the movie has this genre listed) and 0s ("equal to 0" means the movie is not in this genre). Take "movieid = 3257" as an example: this film's genres are drama, romance, and thriller.

* tag.csv

It includes movies id, tag, and relevance score, which has 7,075,944 observations. The relevance score ranges from 0 to 1, representing how strongly movies exhibit particular properties represented by tags.

## Analysis Process & Results

#### **Analysis with some sample queries & dashboards:**

At the beginning of the analysis, to study and understand the data better, I used SQL (Appendix A-5) to produce some basic information about the movie data.

###### Movie with $100 million revenue:

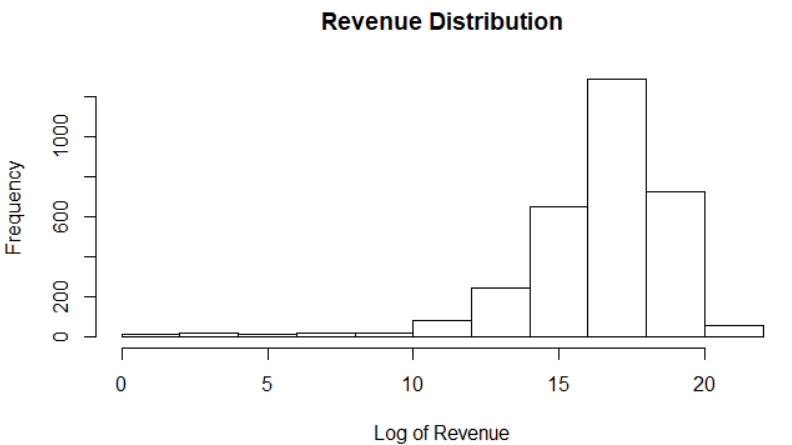


Figure 2 - Log(100 million) ≈ 16.1

Over the most recent 15 years (01/01/2015 – 12/31/2019), the average number of released English-speaking films is **105.64**. According to Figure 2, we can see the sample distribution of the revenue of a fair amount of films cuts at 16, which is $100 million. So, I also looked into the average number of the movies with revenue over $100 million, which is **8.5**, meaning over the last 15 years, approximately eight out of every 100 films annually would have revenue exceeding $100 million.

###### The annual number of each genre for most recent 15 years:

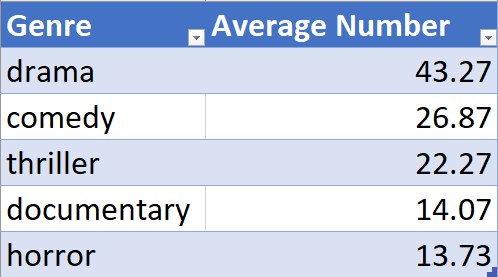


Figure 3 - Top five genres released annually

The five most popular genres (Figure 3) are drama, comedy, thriller, documentary, and horror. According to the analysis of the genre table, the documentary is the 11th in the total numbers of the dataset, meaning documentary must have become increasingly popular over the past 15 years.

###### 90s, Cinema’s fairytale decade:

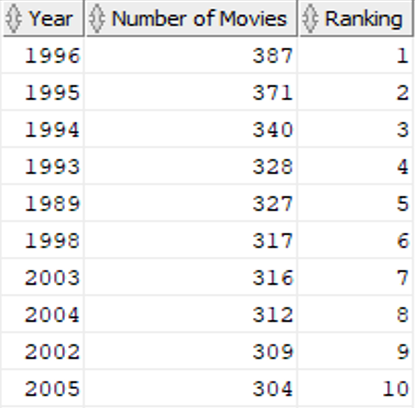


Figure 4 – 90s, Cinema’s fairytale decade

Figure 4 shows that the 90s take up half of the top ten (four of the top five) most productive years in the past 114 years of cinema history. The reason why the cinema industry was booming in the 90s could be that the low-budget independent films unceasingly rose and maintained their popularity in the industry within the decade. Also, in the late 90s, Netflix started to offer rental DVDs service, which could be another reason. Moreover, with the introduction of the Internet and PCs, it may had shaped the consumers' movie-watching practices and stimulated the industry as well. More information can be found in Appendix B or [here](https://www.filmsite.org/90sintro.html).

###### Correlation coefficient between ratings and revenue:

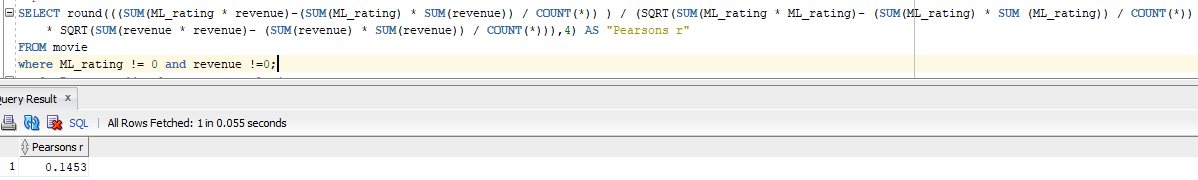


Figure 5 – Rating and Revenue aren’t strongly correlated

According to the definition of Pearson’s Product-moment Correlation Coefficient, the test gives a measurement from -1 for a perfect negative correlation to 1 for a perfect correlation. And a correlation of 0 means that there is no relationship between the two. Given the result is **0.1453** (Figure 5), I do not think there is a strong link between rating and revenue. It could be due to the range of the time in this dataset is too large, and the amount of revenue for an individual movie has significantly increased over the years.

###### Ranking of profitability:



Figure 6 – Top ten most profitable movie in history

I also did a quick analysis (Figure 6) of profitability by looking at the profit margin. Some very famous movies are on the list, like Modern Times. These movies are not necessary to be the best movies in the history, but they are definitely worth some further study if you are interested.

###### Box-Office analysis for the past 30 years:

Since our dataset contains the movies from the past 114 years, I decided only to analyze the box office of the most recent three decades to make the analysis more relevant.

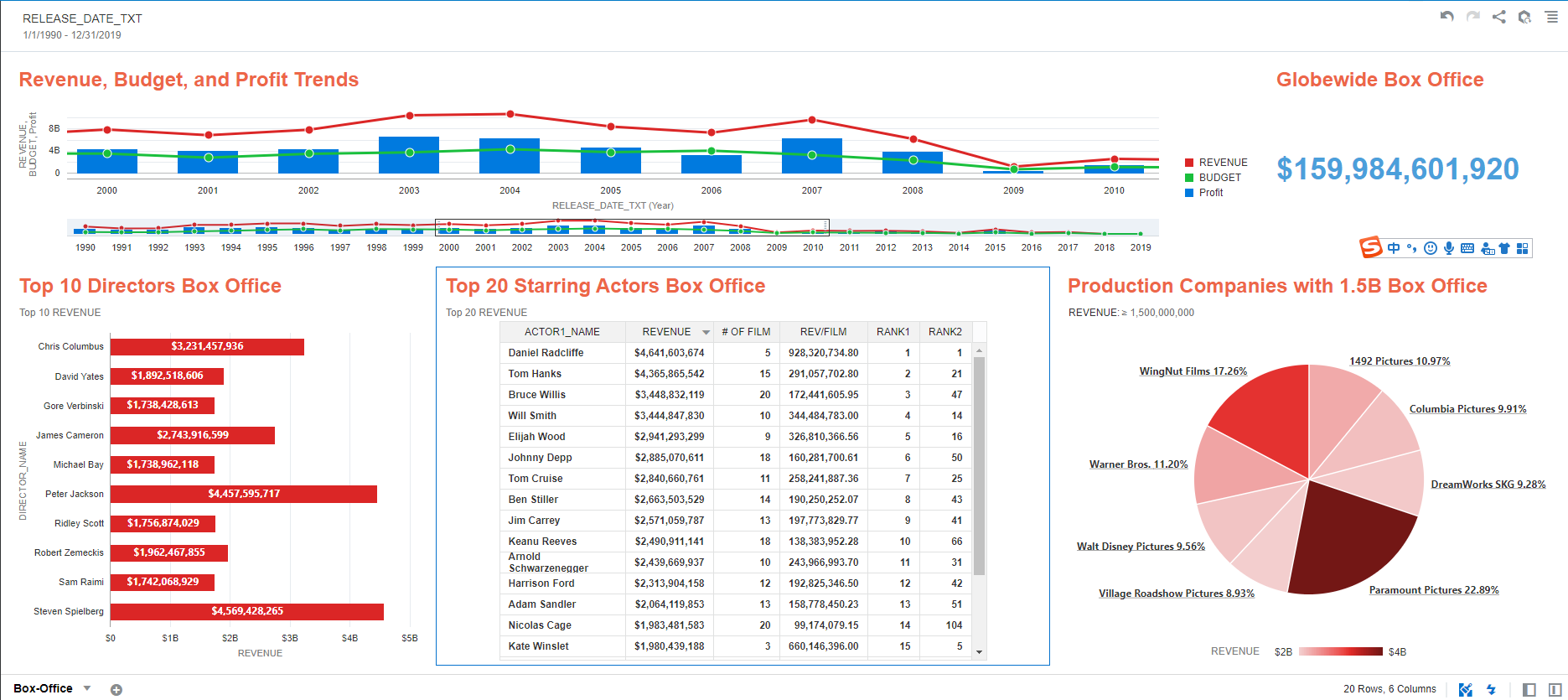


Figure 7 – Box office analysis (1990-2019)

According to the dashboard (Figure 7), the total box office for the English-speaking films in the past 30 years is approximately ***$160B* globe wide**. And **Paramount Pictures took 22.89%** of the total revenue, followed by **WingNut Films 17.26%**, **Warner Bros. 11.20%**, and **1492 Pictures 10.97%**.

I displayed the top 10 directors, and it shows ***Steven Spielberg*** and ***Peter Jackson*** had the most box office since 1990.

In the table above, I displayed the top 20 starring actors based on the total box office first. Soon I realized that it should not be the only factor to measure the box office drawing power of the top actors since some of them may have made many films in the past three decades while some of them may only have made a few movies due to the late debuting or something similar. So, I also analyzed the revenue per film to understand the box office better. "Rank 1" is the ranking of the total box office. "Rank 2" is the ranking of the box office per film. From the two rankings, we can see that **Daniel Radcliffe** fully deserves the most powerful box office draw by being ranked as number one on both lists. Interestingly, **Tom Hanks** is *number two* for the total box office list but only *ranked 21* for the box office per film.

For the line chart named "Revenue, Budget, and Profits Trends," I zoomed in to focus on the decade between 2000 and 2010 since there is a sharp drop around 2007-2009, very likely due to the great recession. Looking at the whole picture of the past thirty years, I also found out that the revenue and profit have been decreasing while the budget gradually increased.

###### The analysis of gender in cinema industry:

In this part of analysis, I focused on how the gender could affect the performance, revenue, and budget. “M” is male, “F” is female, “N” is untold. Since “N” is only a small portion of the whole dataset, I decided to leave it alone instead of doing any transformation. This analysis only focuses on the most recent three decades as above.

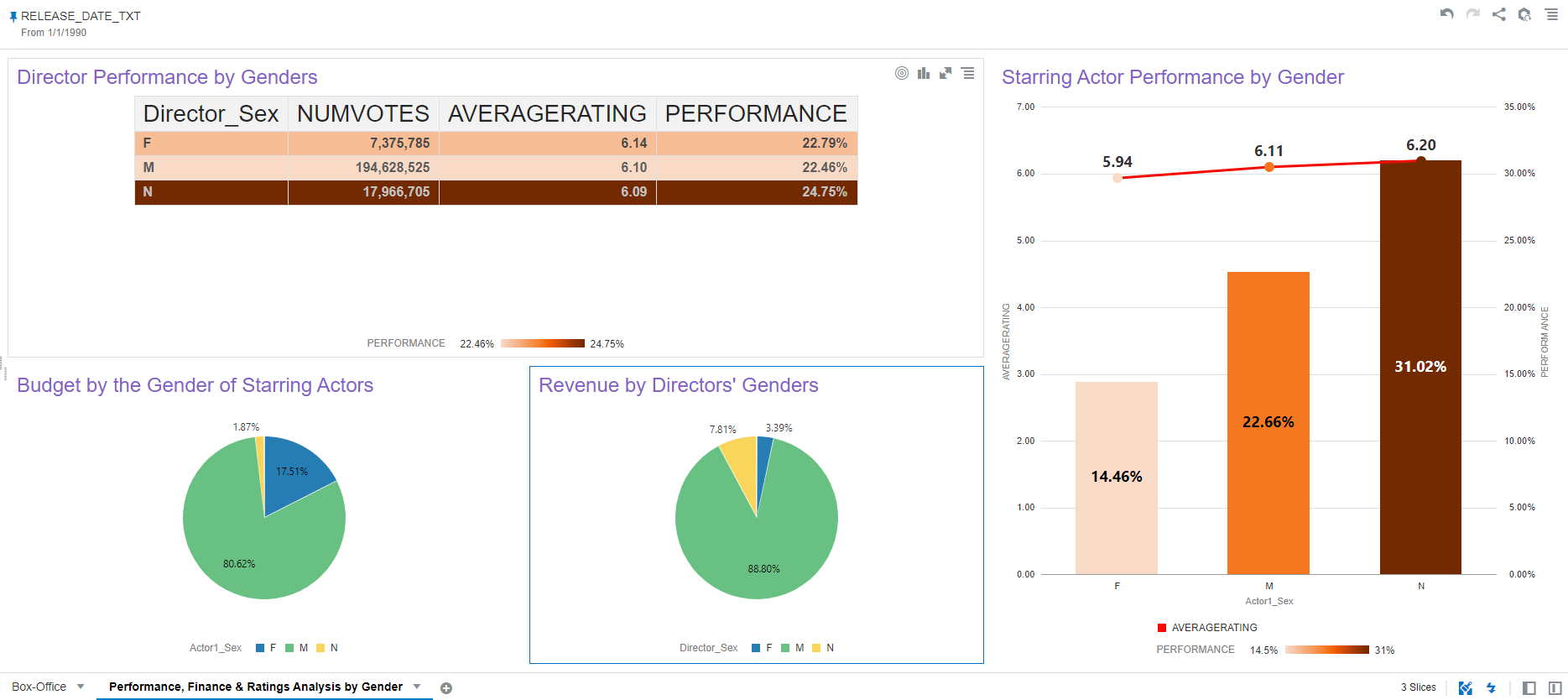


Figure 8 – Analysis of performance & finance by genders

In figure 8, we can see the first chart called “Director Performance by Genders.” For the whole dataset, the possibility for a male director's film to be rated above *7* was **24.82%**, while **22.83%** for a female director. However, for the past thirty years, the average performance for female directors surpassed male directors by ***0.33%***. The performance of female directors was very consistent in general. On the other hand, male directors underperformed over the past three decades.

In contrast, the movies with male starring actors outperformed female starring actors with **8.2%**. And the average rating was also **0.17** points higher.

In the charts regarding budget and revenue, we can tell that male directors and male actors took the most resources. A relatively large amount of budget for a movie will go to actors, and the starring actors tend to take more than others. Moreover, I believe the vast difference is a sign of inequality in the cinema industry since male directors and actors are getting more good opportunities as well as potential income. This hypothesis is consistent with an analysis article of Forbes in 2017 (Article can be found in Appendix B or [here](https://www.forbes.com/sites/natalierobehmed/2017/08/22/full-list-the-worlds-highest-paid-actors-and-actresses-2017/#67907d793751)).

#### **Network Analysis of Genres:**

To study and predict the genre combinations of top-rated movies, I decided to do a deep learning network analysis, using the algorithms of Node2Vec and K-means. [Node2Vec](https://cs.stanford.edu/~jure/pubs/node2vec-kdd16.pdf) is an algorithmic framework for representational learning on graphs developed by Stanford (Appendix B) and the [usage demonstration](https://medium.com/analytics-vidhya/analyzing-disease-co-occurrence-using-networkx-gephi-and-node2vec-53941da35a0f) of this algorithm can be found in Appendix B.

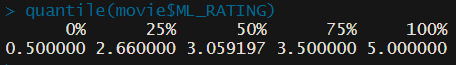


Figure 9 – Top rating cuts off at 3.5

To find out top rated movies, I firstly did a quick check of the quantile of the ML\_Rating (Figure 9) and decided to cut at 3.5, where is top 25% (4335 records). Then I created a “top rating adjacency matrix of genre” (Appendix A-6) that looks like the following:

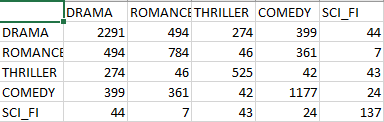


Figure 10 – Adjacency Matrix

Then I use the Node2Vec package in python (Appendix A-7) to generate a set of k-means features for the selected genre data. Then I used R (Appendix A-8) to do the rest k-means analysis and plotting.

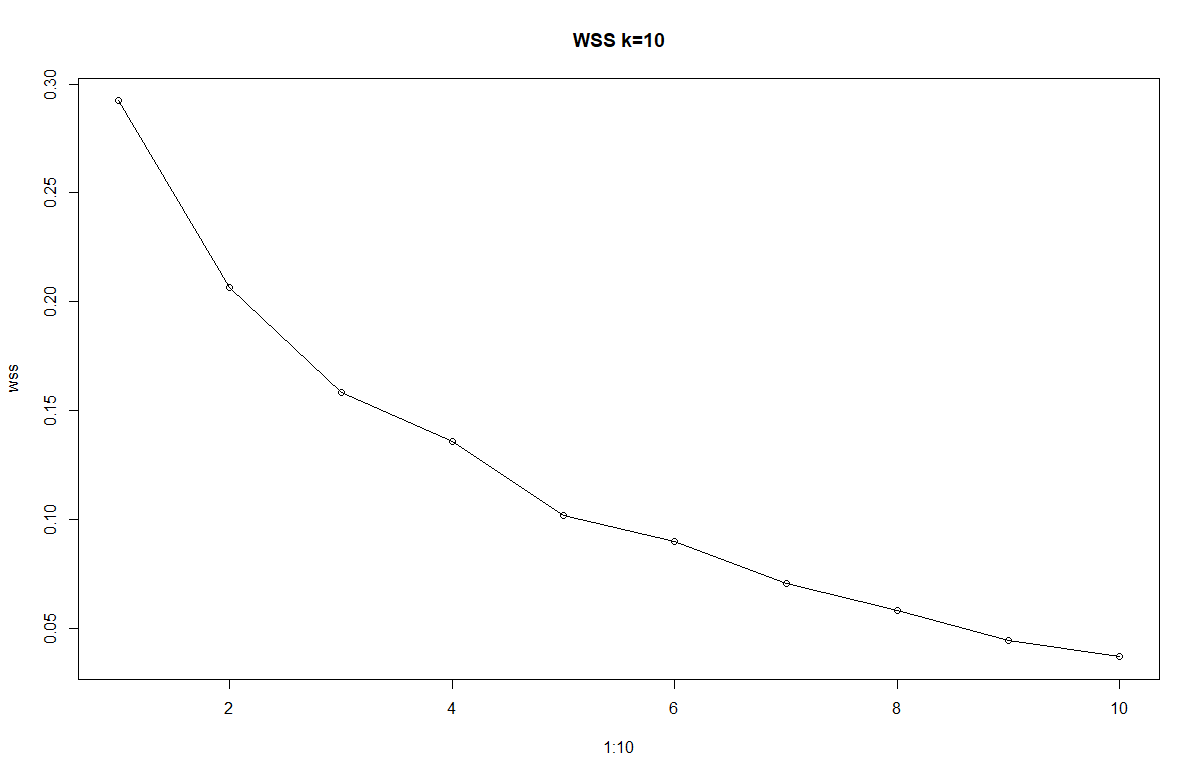
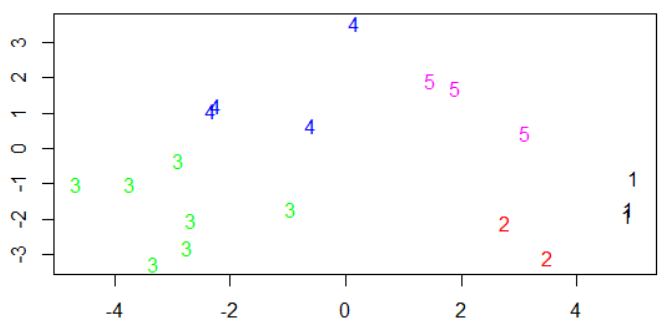
 

Figure 11 - Elbow curve shows at 5 Figure 12 – K-means plot

Based on Figure 11, I decided to use k=5 since it is where the elbow curve shows. And the final plot looks like Figure 12, which is pretty neat and clear. I ran a subsample afterward to gain insights about the popular genre combinations for top-rated movies.

The genre combination of the first subsample group of top-rated movies is composed of “**drama, comedy, animation, and children**.”

The genres of the second group are “**drama, thriller, and IMAX**.”

The genres of the third group are “**drama, adventure, action, mystery, war, documentary, and film-noir**.”

The genres of the fourth group are “**drama, romance, musical, crime, and fantasy**.”

The genres of the fifth group are “**drama, sci-fi, western, and horror**.”

By studying the popular genre combinations above, movie producers will be able to make a better decision on what kind of film scripts they should pick. And this model also can be applied to an advanced recommendation algorithm for streaming platforms, like Netflix. For example, if a customer watched a lot of movies with genres in drama and comedy, it will be reasonable to recommend the customer some animation or children movie.

#### **Natural Language Processing Analysis:**

###### Wordcloud2 & sentiment analysis:

In this step, I was very interested in studying the relationship between the TAGs and the high Relevance scores. A quantile test shows that **the top 10%** relevance scores cut off at ***0.2905***. So, I subsampled the TAG table, and get a new table with **707,734** records.



Figure 13 – Word Clouds I

Figure 13 shows that there are some common and meaningless words, like “good”, “great”, and “best”. I removed some of them to help gain a better understanding.

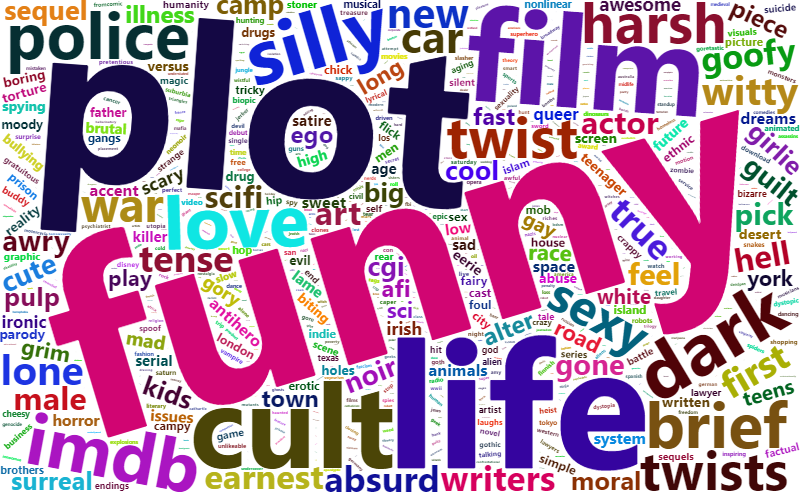


Figure 14 – Word Clouds II

Figure 14 looks like there are plenty of emotional words and expressions, so I believe it is worth a sentiment analysis. It is not necessary that a top-rated movie is linked to positive tags. “Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material ([Gupta 2018](https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17); Appendix B).” A sentiment analysis will help the movie business understand the relationship between the [tag genome](http://files.grouplens.org/papers/tag_genome.pdf), a data structure that extends the traditional tagging model to provide enhanced forms of user interaction (Appendix B), and top-rated movies, as well as the social sentiment to the top-rated movies.

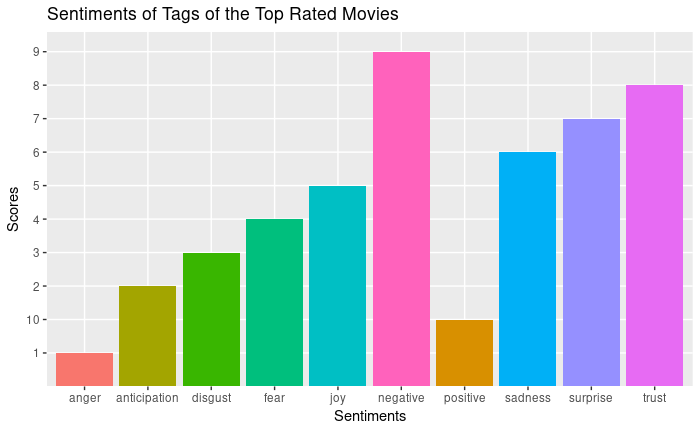


Figure 15 – Sentimental Analysis Plot

Surprisingly, the final output for the sentimental analysis (Figure 15) of the movies with high ratings suggests that the sentiment of “negative” has the highest score, and sentiment of “joy” is in fifth place. I came up with three hypotheses regarding this output.

* First, in this case, perhaps the contents of a quite portion of the top-rated movies are associated with a negative attitude to attract customers’ attention since a lot of Hollywood blockbusters do not have happy endings.
* Second, the sentiment analysis mistook the “intensive” or something similar for “negative” due to the deficiency of the algorithm.
* Third, a commercial movie tends to create buzz on social media, and a lot of arguments and conversations could eventually contain negative sentiments.

***The insights of this part of analysis will help make better marketing decisions for the movie production companies. Also, the media-service companies will also be able to develop interfaces that combine tagging and recommendations, such as Movie Tuner and Music Explaura system*** ***[***[***Green et al. 2009***](https://dl.acm.org/doi/abs/10.1145/1639714.1639768)***] (Appendix B).***

###### Specificity Analysis

I collected all the tags into one document and then conducted a *Specificity* analysis over all the tags with a relevance score above 0.2905 (Top 10%) as well as for the group of tags with a relevance score below 0.012 (Bottom 10%). *Specificity*, as the number of specific words or phrases conveying specific information (Hope, 2016) relevant to the movie, divided by the number of total words in the tags. To extract specific entity names, I used the Named Entity Recognition (NER) technique with Spacyr software. The higher value of *Specificity* is, the more specifically the tags described the movies. For information regarding *Specificity* can be found in Appendix B.

The whole dataset of tags with top 10% relevance scores contains *707,734* records, and *1,030,029* of them are identified as words by Spacyr. After parsing the data, I was able to extract ***42,682 Named Entities***. So, the average Specificity of the tags with the top 10% relevance score is **0.0414** (Appendix A-12).

The number of tags with bottom 10% relevance scores is *685,361*. It has a total of *1,020,717* words. **234,835** of them are identified as Named Entities, which means, interestingly, the average Specificity of the tags with the bottom 10% relevance score is **0.23**.

The correlation coefficient between the Specificity and Relevance is approximately **-1**, which means they are strongly negatively correlated. In this case, I infer that if the tags are more relevant to the movie, the content of the tags will be less specific, which means a lot of emotions will be included in the tags.

#### **Model Development and Comparison:**

###### Model 1 – IMDb Average Rating Prediction

I firstly tried a linear regression model with a dataset containing 32 variables. I split the dataset into 80% training and 20% testing. I ran a stepwise check for both direction and here is the final list of variables for model 1 and output for summary (Figure 16 & 17):

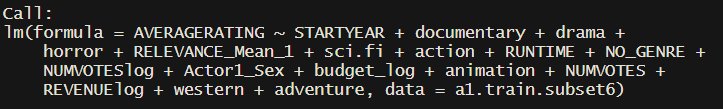


Figure 16 – Model 1

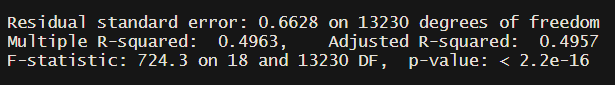


Figure 17 – Summary output for model 1

All the variables are statistically significant in this model; however, the Adjusted R-square is only 0.4957 which means only **49.5%** data can be explained by this model. Plus, the residual plots suggest a Y-transformation needed. Model follows normal distribution. The Mean Squared Prediction Error (MSPE) is **0.6627**. The AIC is **26721.63**.

###### Model 2: Use same data to predict “PERFORMANCE”

In this model, I manually set parameter for the averagerating at 7. So, if the movie is rated greater or equal to 7, performance will be 1, otherwise performance will be 0. I fitted the data with a logistic regression model, and the final outputs are the followings:

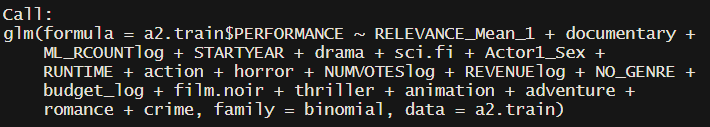


Figure 18 – Model 2

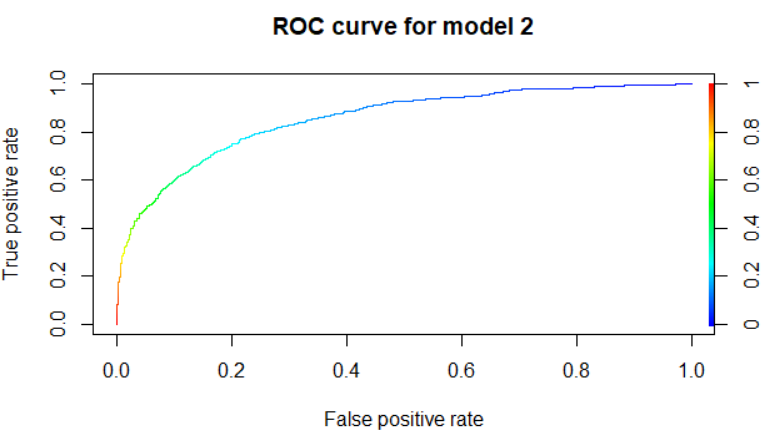


Figure 19 – ROC curve for model 2

**The value of AUC: 0.8536802**

**The AIC: 10573**

**The Misclassification Rate (MR): 0.2164616**

**The False Negative Rate (FNR): 0.06239016**

**The False Positive Rate (FPR): 0.1540715**

The FPR is too high compared to FNR, which is not ideal as a movie production company would rather not to produce a movie with the potential to succeed than create a film that is very likely not to perform well. So, in the future analysis, we may have to change the model to lower FPR. The AUC looks good, and the AIC of model 3 is also much lower than the AIC of model 2.

###### Model 3: Prediction based on ROI

Because the correlation coefficient between revenues and ROIs for the past 30 years in my dataset is only **-0.00827**, I realized that a movie with a high box office is not necessarily successful from a perspective of investment. Dr. Michael T. Lash (2017) decided to use return on investment (ROI) as a measurement to predict if a movie is "truly profitable" (Appendix B). ROI is defined as "**profit divided by budget**." I used the same method to build a model here.

Our dataset contains 17,071 records over the past 114 years, but the revenues of movies have increased significantly. So, I will only use the data from the most recent 30 years with 1,495 valid records, which is more relevant for prediction of profitability. The following is the output of the model 3:

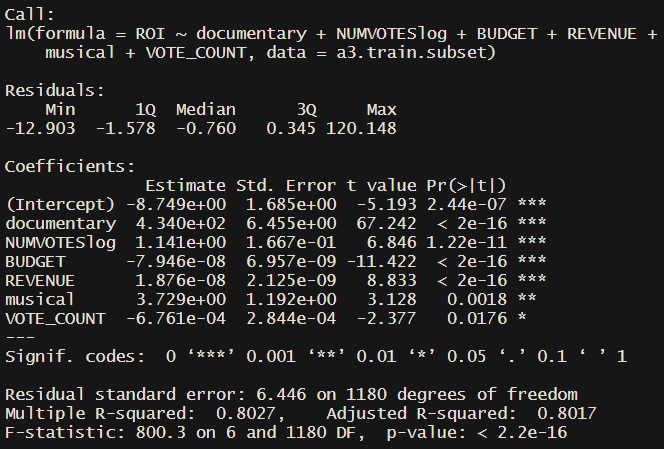


Figure 20 – Output for model 3

It looks like the **genre of documentary and musical** will have a positive impact on the ROI, and if the movies get votes on IMDb will also increase their ROIs. The adjusted **R-squared is 0.8017**. The **AIC is 7801.35**. And the **MSPE is 1966.634**. Overall, the model has its potential, but it does not perform as well as I expected.

I think both Model 2 and Model 1 have better performance than Model 3. Moreover, further development has the potential to turn Model 2 to an optimization algorithm for movie production. All the code for fitting the models can be found at Appendix A-11.

## Conclusion

In this machine learning project with the given movie datasets, I firstly joined multiple tables in one table and did some necessary data preparation for analysis. Then, I did some SQL query analysis and created two dashboards to reveal many significant and meaningful insights about the data. I did a deep learning network analysis to study the relationship among genres for the top-rated movies and find five popular combinations. Moreover, I used some contextual text mining techniques to explore the relationship between tags and top-rated films. Finally, I built three models to predict if the movie can be rated above 7 on IMDb and ROIs. The prediction results are acceptable and reasonable. ***Movie production business should find my findings useful and will be able to develop a better recommender system, optimize the selection decisions of future movies, and make better decisions on marketing campaigns.***

## APPENDIX A – Supplementary Materials

You may check the partial codes and relevant tables on my [GitHub](https://github.com/jinhangjiang/BSAN480-FinalProject).

1. [SQL code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/MovieTable_sqlcommand.txt) for cleaning and creating the original table
2. [R code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Model_analysis.R) for modifying the movie table
3. [SQL code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Tag_creation.sql) for creating the tag table
4. [Details and references](https://github.com/jinhangjiang/BSAN480-FinalProject/tree/master/Reference) for the tables
5. [SQL code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/FinalProject.sql) for the basic analysis and queries
6. [SQL code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Top_Rating_Genre_Matrix.txt) for creating the top\_rating\_genre\_matrix
7. [Python code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Movie_genre_network.py) for generating Node2Vec features of genres
8. [Code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Genre_networks.R) for creating genre networks in R
9. [Sentiment Analysis Code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/NLP_Analysis_Report.pdf) of the tags in R
10. [Dictionary](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Reference/Dictionary%20for%20cleandata%20table.png) of attributes in cleandata.csv
11. [Code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Fit_Model.R) for fitting models in R
12. [Code](https://github.com/jinhangjiang/BSAN480-FinalProject/blob/master/Scripts/Specificity%20Analysis.R) for conducting Specificity analysis

## Appendix B – Reference

Dirks, T. (n.d.). The History of Film The 1990s. Retrieved May 13, 2020, from <https://www.filmsite.org/90sintro.html>

Green, S. J., Lamere, P., Alexander, J., Maillet, F., Kirk, S., Holt, J., … Mak, X.-W. (2009). Generating transparent, steerable recommendations from textual descriptions of items. *Proceedings of the Third ACM Conference on Recommender Systems - RecSys 09*. doi: 10.1145/1639714.1639768

Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi: 10.1145/2939672.2939754

Gupta, S. (2018, January 19). Sentiment Analysis: Concept, Analysis and Applications. Retrieved May 13, 2020, from <https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17>

Hope, O.-K., Hu, D., & Lu, H. (2014). The Benefits of Specific Risk-Factor Disclosures. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2457045

Jiang, J. (2020, May 7). Analyzing disease Co-occurrence Using NetworkX, Gephi, and Node2Vec. Retrieved May 13, 2020, from <https://medium.com/analytics-vidhya/analyzing-disease-co-occurrence-using-networkx-gephi-and-node2vec-53941da35a0f>

Lash, M., Fu, S., Wang, S., & Zhao, K. (2015). Early Prediction of Movie Success — What, Who, and When. *Social Computing, Behavioral-Cultural Modeling, and Prediction Lecture Notes in Computer Science*, 345–349. doi: 10.1007/978-3-319-16268-3\_41

Robehmed, N. (2017, November 10). Full List: The World's Highest-Paid Actors And Actresses 2017. Retrieved May 13, 2020, from <https://www.forbes.com/sites/natalierobehmed/2017/08/22/full-list-the-worlds-highest-paid-actors-and-actresses-2017/#284ed9323751>

Vig, J., Sen, S., & Riedl, J. (2012). The Tag Genome. *ACM Transactions on Interactive Intelligent Systems*, *2*(3), 1–44. doi: 10.1145/2362394.2362395