Project - 353 **Summer 2019** **Akshay Nanda**

Objectives/Goal:

- The goal of this project is to find and analyze what variables are important in producing a successful movie. Who affects the profit of a movie is it genre, directors or cast members, critic rating, audience rating?

Data Gathering: Collecting the data.

Data gathering was done from gateways.sfucloud.ca in the directory /home/bigdata/movies.zip. This data extracts into genres.json.gz , omdb-data.json.gz, rotten-tomatoes.json.gz and wikidata-movies.json.gz. All these json files were loaded into the data frames and then further used for data analysis.

Data Exploration and Cleaning:

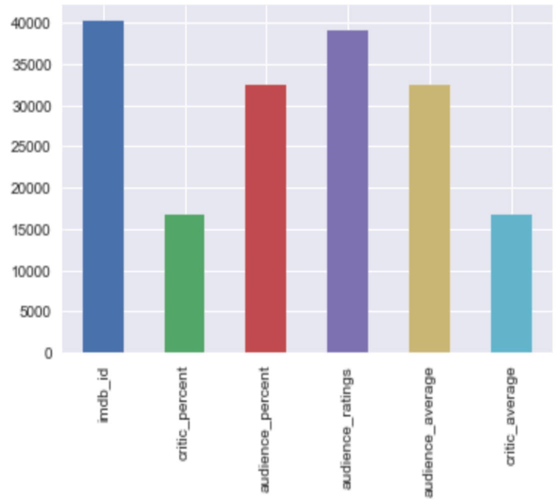
Using visual printing, which data set had the most columns. Wikidata-movies had the most columns. Firstly, we segregated data and checked that which language is the most prominent among our dataset. English language based movies were around 66% and all the rest were below 6%. Interesting column was discovered which old us which movies made profit (it was binary).

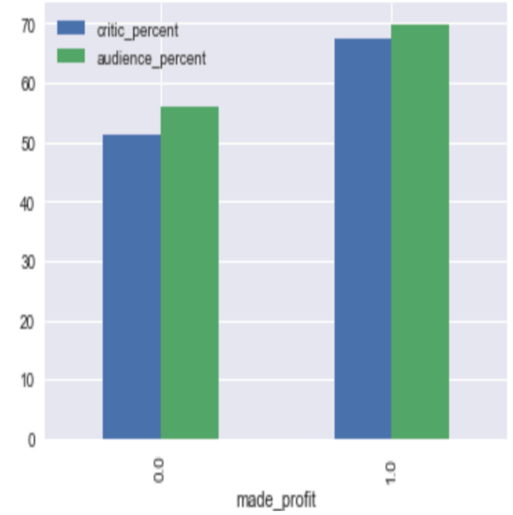
Secondly, we saw the Rotten-tomatoes dataset and counted all the values that given to us and what is missing (refer Fig: 1 in Appendix 1). We found out that critic average and percent were the lowest and had the most data missing (out of 40,000 entries we only had 12,000 entries available). In most data sets there were columns that had very few entries and were removed such as 'based_on', 'cast_member', 'country_of_origin', 'eniki_title','filming_location', 'label', 'main_subject', 'metacritic_id', 'original_language', 'rotten_tomatoes_id', 'series','wikidata_id'.

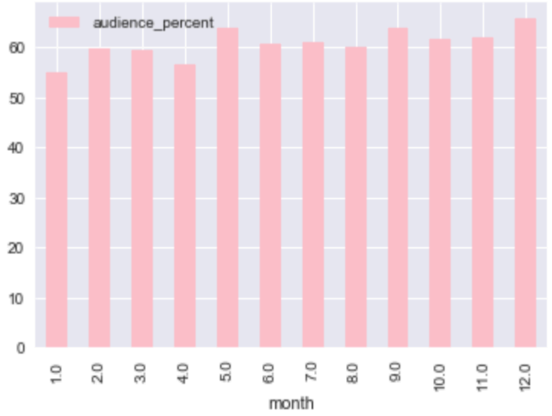
All the data was merged on imdb_ids and rest of the ids were removed. **Outliers** such as directors with profits less than too were inconsistent and did not count towards better profit providers. Plots was also considered as an outlier because it was based on text and would not have been counted as possible predecessor of success of a movie. On the other hand, cast members could have been a good indicator of success but kernel kept on dying every time we tried to Count vectorize and was considered not to be used. Linear regression was used to see correlations between variables but didn’t seem very helpful towards the goal of the project.

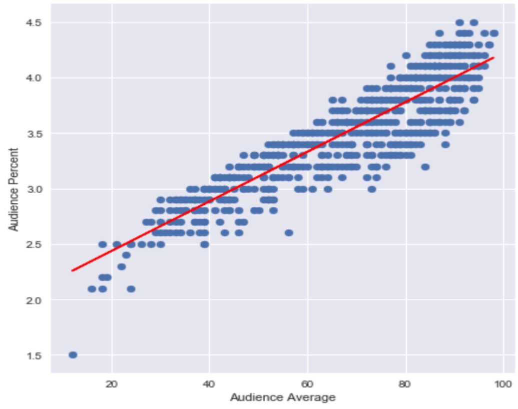
Workflow and Pre - Results:

Next, using made profit and merging wiki-data with rotten tomatoes we wanted to see if people rate profitable movies higher or not. So we plotted a bar plot and saw that profitable movies are rated higher. This showed us a way towards using audience percent and critic percent as good indicator of profit making and can be used for machine learning. Similarly, we saw director ratings by implying ratings a movie is directly proportional to critic percent and audience percent. Also explored, does time of the year or month of the year produce higher ratings.









Thirdly, we saw in the omdb-dataset, that information regarding awards and winings could be a good factor to measure success as these Oscars, nominations, other awards, bafta and golden globe are the movies that end up making profit because of their quality. Regular expression was used to gather all these values and placed in columns and placed against their respective imdb_ids.

We also though about genres in our data set and wanted to see if they can make a difference in so used sklearn library (feature extraction.text) and used CountVectorizer to restructure the dataset and placed 400 more columns as some movies had multiple genres. We thought it would be a good indicator of profit. Some genres can be more successful and produce more profitable movies.

Next up, checking the dependency of each variable on made profit or not. Mutual info classification was used to produce between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. We plotted the mutual dependency of each variable in predicting profit (Fig 4).

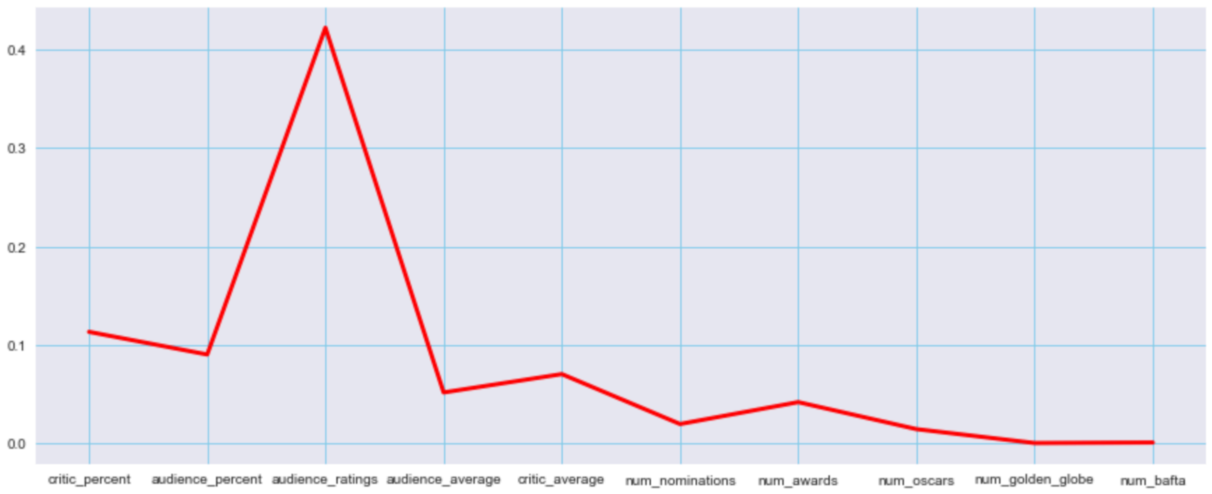


Fig 4: Mutual classification info plotted against respective variables. Y-axis denotes the mean of mutual dependency and X-axis denotes the variables.

Now we see in Figure 4, audience ratings have the most dependency and rest of the variables like Oscars and awards are important contributors in comparison to golden globe awards and British awards (Bafta). All these variables will be important contributors while preparing the different models for predicting a profitable movie.

Modelling the Data (Results)

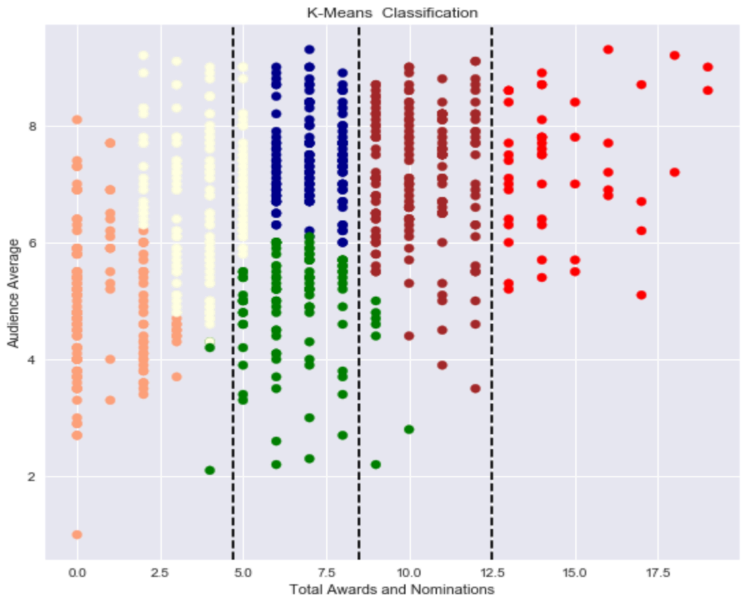


Figure: K-means clustering of audience average based on all awards and honorable mentions.

Round 1: K-means clustering was done using 6 clusters to see regions of success of the movie. Highly rated movies produce better results in terms of awards.

Round 2: Using the variable selected after mutual classification and count vectorizing genres with their respective movies. We used different models to predict the if a movie made profit or not. Made_profit was binary. Making profit means a value of one and not making profit a value of zero. Three different models were used.

Model	**Accuracy Score**
Gaussian NB	0.294 (t = 2sec)
KNN	0.864(t = 4sec)
SVM	0.793(t = 35sec)

Table 1: Predicting made_profit using different models. Time = t to show the result.

Round 3: Secondly, running a test on Knn model to see how many neighbors provide the best accuracy score while predicting made_profit. Find the best parametres for predicting profit using KNN model.

Feature	**Best Result**
Max score	0.8569
@variables	3
Best Variables To predict profit	'critic_percent', 'audience_percent' 'audience_ratings'
Number of Clusters	15

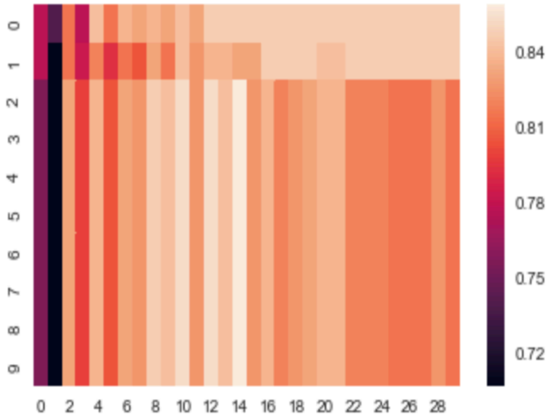


Figure 5: Heat-map shows how accuracy score changes with adding one extra-variable column at each iteration.

Heat map shows that number of Oscars, awards and rest after 3 variables ('critic_percent', 'audience_percent','audience_ratings') do not improve the accuracy score and can be discarded.

Round 4: After last two rounds of modeling we realized that all the three variables discussed above not pre-variables while making movies they come out after you have invested time and money
.
Pre-variables that indicate profit are also very important to note. Such as directors and cast members which are decided before a movie is made. Can directors predict the profit? was the question we wanted to answer. **Note:** cast members were not taken into account as kernel when we tried to count vectorize and adding respective columns to the respective movies.

Knn_model being the best model. We used that produced an estimated accuracy score for predicting profit to be 0.84 and SVC also produced a similar accuracy score of 0.83. It was a great hit for us as it shows that good directors bring good profits to the producers.

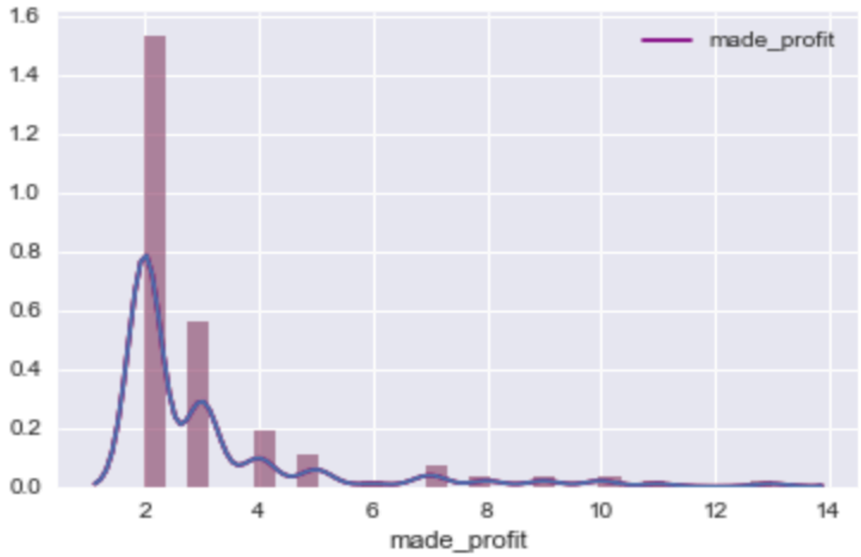


Figure 6: Show the plot of Mean of Directors count versus their profit count. So therefore, there a lot of directors with less consistency of bringing profit, but there less directors who almost always make a profitable movie.

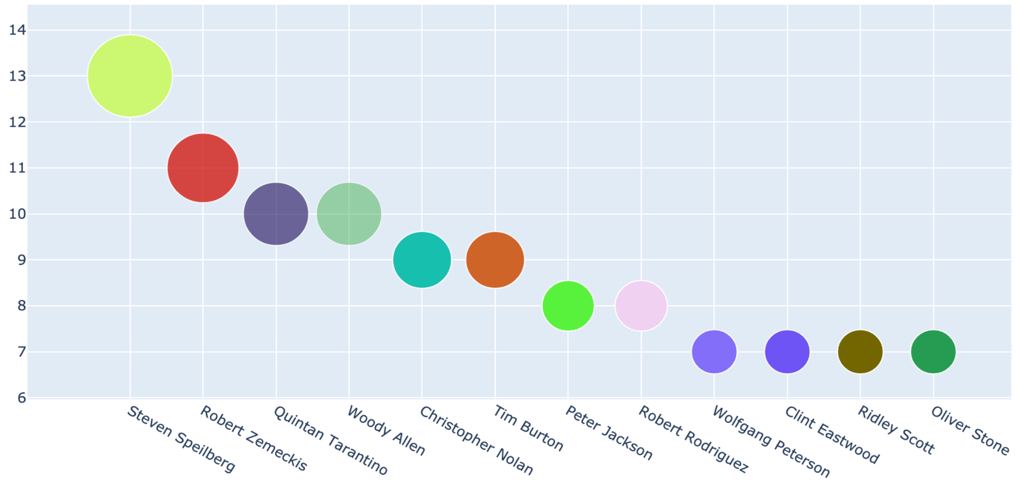


Figure 7: Bubble plot of most successful directors based on made_profit from their different movies. Bubble size denotes their higher chances of bringing profit in. (Note : wiki-data-ids were replaced by actual names of directors.)

Figure 6 and 7 are really good results that show that directors are the most important factor in producing a good movie.

Conclusion

- Seasons and months don't matter while releasing a movie as audience ratings don't change during the course of the year. They are always very genuine in giving their review.

- Better ratings from audience and critic brings success and more awards hence more profit. If they don't rate your movie too well in the opening weeks. Movie most probably be not very successful.

- Three variable Critic_percent , audience_percent and audience_ratings account most towards predicting a successful movies and hence will bring more success and profit.

- Choosing a consistent director with good past history of profit getting is the key to getting a profit out of the money and time invested in a movie.

- Based on our data, hire these Directors and producers will most likely make profit Steven Spielberg, Robert Zemeckis , Quintan Tarantino, Woody Allen.

Limitations and what can be improved in the future?

- A lot of data is missing from some datasets. That could bring more meaning to our goal.
- Count vectorizing and creating table such as TF-IDF were taking long time and kernel dies most of the time. Spark could be used to reproduce all the results.
- More outliers can be removed such as we could have only worked with English movies rather than generating a global perspective of success.
- While using regular expression we might have missed useful info as there were many different ways the data was used. It was hard to fetch everything in the right manner.

In the future, we would like to continue the project and keep working on filling the missing data and using other parameters and produce better results.

Accomplishment Summary

- Cleaned using Regular expression and merged important data for data analysis.
- Extracted genres and directors for each movie using CountVectorizer.
- Produced various accuracy scores using multiple Machine Learning Algorithms like KNN, SVC and Gaussian Naïve Bayes.
- Used K-means Clustering technique to cluster movies with high and low total awards vs their audience ratings.
- Visualized data using plotly and matplotlib & generated Heatmaps, Linear regression graphs, distplots, bubble plots to better understand the data.
- Was able to tell a story about my work and produced formal report with robust conclusions.