Business Analytics Project
IEOR 4650

What factors contribute to the success of a movie?

Mainak Pyne
Sandhya Sriramam
Karan Viegas
Kartik Vyas

INDEX

Problem Definition & Business Relevance

Since the advent of the film industry, filmmakers would rely on gut checks and intuition to foretell how well their films would do once released. However, this would oftentimes lead to unforeseen sales and even flops. Now, post the big data revolution, we can leverage an expansive amount of historical data on the performance of films to understand the key factors that contribute towards their box office success. This analysis would subsequently inform the film community how to maximize the probability of success of a film by choosing by choosing these parameters wisely.

We consider a variety of factors that are characteristic to every film such as quality of the cast (in terms of recognized skill, popularity and reputation), popularity of the director, budget, genres and keywords associated with movies, and try to understand how these factors contribute toward the gross revenue of the movie.

An interesting use case would be this: Given a genre, plot, expected duration etc. for a movie to be directed by a specific director with a fixed cast and within a given budget, what gross revenue can the movie expect to make?

Building the Dataset

Data Collection

All data were primarily procured from three sources for the purposes of this project:

> Movie Meta Dataset:
>
> This dataset had plenty of features on almost 5000 movies from the 20th and 21st centuries. The data was limited in that it did not inform us about the attributes of actors and directors of movies but only gave us their names.
>
> Hadley Wickham's Normalized IMDB Movie Data:
>
> We supplemented the initial dataset with data from the IMDB dataset, which additionally contained data on budget, gross, ratings, Facebook likes, movie durations, genres and associated keywords
>
> Web Scraping:
>
> Using Python's BeautifulSoup library, the Twitter sentiment of actors, their collection of awards and nominations were web-scraped to supplement the dataset for the project.

Data Cleaning

Preliminary organization of the data was done on Microsoft Excel with more in-depth cleaning & filtering performed on R. The following filters were applied to the data set after removing null rows:

1. Actor 1 and 2 Facebook likes > 100
2. Actor 3 Facebook likes > 50
3. Director Facebook likes >50

All the above values were decided based on a summary of the whole dataset and the values of variables relative to their median. On Excel, movies were filtered on columns like Languages (only English) and country yielding the final dataset. The cleaned dataset was then used to append with the web scrapped data.

The original dataset also had one column for genres and one for plot keywords in the form of a string separated by "|", this was also separated into multiple binary columns.

Formulation

In order to make the analysis more comprehensive, the awards the actors and directors had won till the year of the movie release was web scrapped and scored according to this formula,

Award Score = 1*(0.3*OAN + 0.7*EAN) + 4*(0.3*OAW + 0.7*EAW)

OA = Ordinary Award Nominations
EA = Extraordinary Award (Oscars, Golden Globes, BAFTA) Nominations
OA = Ordinary Award Wins
EA = Extraordinary Award (Oscars, Golden Globes, BAFTA) Wins
The top 3 awards in the film industry Oscars, Golden Globes and BAFTA were considered and both nominations and wins were weighted relatively.

Actor Publicity Score = 10*B + 5*M + 1*(I + MA + A)

B = Biographies; M = Magazine covers; I = Interviews; MA = Media appearances; A = Articles

Gross:
The square root of gross has been considered as the dependent variable to be predicted. This was done because regressions using gross (as is) was showing significant heteroscedasticity.

Final Dataset

The final dataset had the following features,
Movie title, Director Name, Director FB likes, Actor (1,2,3) Name, Actor (1,2,3) FB likes, Cast Total FB Likes, Number of Critics for Reviews, Number of Faces in Poster, Title Year, Genres, Associated Keywords, Twitter Sentiment of Actor 1 and 2, Award Score for Actor 1, Publicity for actors and directors, age of actors, IMDB score of the movie. For predictive analysis, title year and IMDB scores have been ignored.

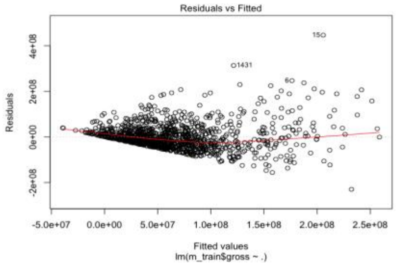
Modeling

Multiple Linear regression

Before proceeding to regress the data, we assessed correlations between independent variables and removed "Cast Total FB Likes" as this was to correlated to actors' FB likes.
We regress gross generated by the movie with the features of the model - cast twitter sentiment, awards, director awards, budget, duration, director's facebook page likes, number of faces in the poster and the genre.

Linear regression of Gross against all independent variables:

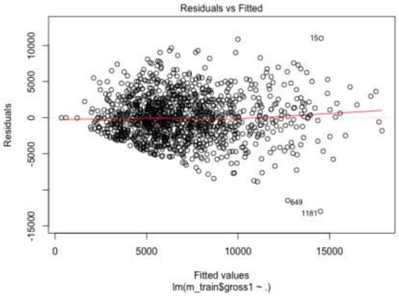
Residual standard error: 55040000 on 1218 degrees of freedom
Multiple R-squared: 0.4451, Adjusted R-squared: 0.4282
F-statistic: 26.41 on 37 and 1218 DF, p-value: < 2.2e-16
MSE = 3.244693e+15



As visible from the plot of residuals, there is significant heteroscedasticity, which is why we decided to transform the dependent variable from Gross to Square root of Gross.

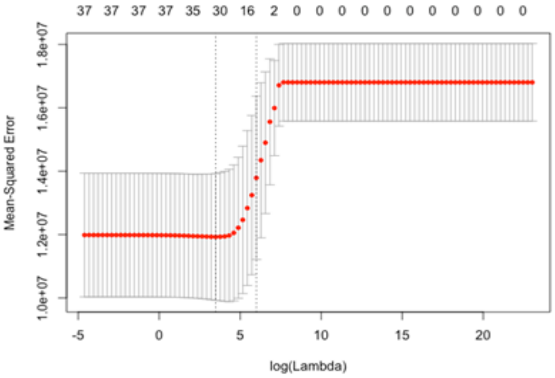
Linear regression of square root of Gross against all independent variables:

Residual standard error: 3091 on 1218 degrees of freedom
Multiple R-squared: 0.4478, Adjusted R-squared: 0.431
F-statistic: 26.69 on 37 and 1218 DF, p-value: < 2.2e-16
MSE = 11033416



As observed in the plot, the residuals are no longer funnel-shaped.

Linear regression of square root of Gross after variable selection using LASSO:



We then conducted a Lasso Regression to determine which features are not significant in the model and can be removed. The results of the Lasso Regression can be viewed in the appendix.

Best lambda = 75.64633

The best value of Lambda was found and the features were shortlisted from the Lasso coefficients.

After removing the insignificant features and keeping only the best subsets of features, we perform another linear regression and obtain the following results,

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.373e+03	7.208e+02	-3.292	0.001024 **
tweet_2	-1.074e+01	8.331e+00	-1.289	0.197702
tweet_1	-9.024e+00	8.182e+00	-1.103	0.270301
num_critic_for_reviews	1.261e+01	8.799e-01	14.330	< 2e-16 ***
duration	3.757e+01	5.021e+00	7.482	1.39e-13 ***

director_facebook_likes	6.448e-02	2.856e-02	2.258	0.024145 *
actor_1_facebook_likes	6.554e-03	4.479e-03	1.463	0.143641
facenumber_in_poster	-1.192e+02	5.664e+01	-2.104	0.035597 *
budget	1.997e-06	1.185e-06	1.685	0.092299 .
actor_2_facebook_likes	4.725e-02	1.674e-02	2.823	0.004832 **
Action	9.142e+02	2.508e+02	3.646	0.000278 ***
Adventure	9.736e+02	2.733e+02	3.562	0.000382 ***
Fantasy	6.422e+02	2.962e+02	2.168	0.030360 *
Sci.Fi	1.100e+02	2.909e+02	0.378	0.705484
Thriller	2.144e+02	2.392e+02	0.896	0.370184
Romance	2.388e+02	2.331e+02	1.024	0.305864
Animation	1.066e+03	5.897e+02	1.808	0.070845 .
Comedy	9.879e+02	2.444e+02	4.043	5.61e-05 ***
Family	2.408e+03	3.901e+02	6.173	9.12e-10 ***
Musical	-4.594e+02	6.330e+02	-0.726	0.468188
Western	-8.116e+02	8.265e+02	-0.982	0.326311
Drama	-1.301e+03	2.282e+02	-5.700	1.50e-08 ***
History	-9.291e+02	5.081e+02	-1.829	0.067683 .
Sport	5.283e+02	4.416e+02	1.196	0.231795
Crime	-2.706e+02	2.530e+02	-1.070	0.284899
Horror	-7.804e+02	3.658e+02	-2.134	0.033060 *
War	6.316e+02	4.748e+02	1.330	0.183646
Music	5.409e+02	4.726e+02	1.144	0.252714
Documentary	7.102e+02	1.190e+03	0.597	0.550764
actor1_age	3.929e+01	6.847e+00	5.738	1.21e-08 ***
actor1_awards	5.376e-01	8.281e-01	0.649	0.516326
director_awards	-4.525e+00	9.020e-01	-5.016	6.04e-07 ***
actor1_publicity	8.047e-01	3.468e-01	2.320	0.020499 *
director_publicity	1.252e+00	1.355e+00	0.924	0.355601

Residual standard error: 3086 on 1222 degrees of freedom
Multiple R-squared: 0.4477, Adjusted R-squared: 0.4328
F-statistic: 30.01 on 33 and 1222 DF, p-value: < 2.2e-16
MSE = 11007985

The results indicate an improved Adjusted R-squared

Insights:

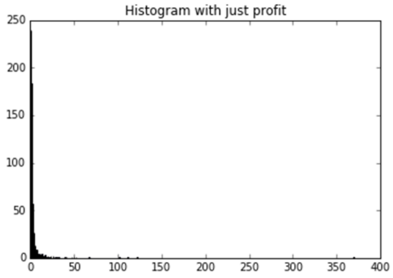
1. Genres:
    a. Action, Adventure, Comedy, Fantasy, Family genres have strong positive correlation with gross revenue
    b. Horror & Drama have a negative correlation with box office success
2. Publicity and age of the lead actor has high correlation to gross
3. Movies with longer duration tend to significantly have higher gross
4. Movies that have more critics writing reviews about the film have high gross
5. Budget has been accounted for in this model. Although the budget is positively correlated with gross, the p-value indicates that it is not significant. Hence, having a high or low budget does not necessarily convert to high gross
6. Having too many faces in the posters also negatively impacts gross

We then moved on to classification algorithms that give more insights into different factors.

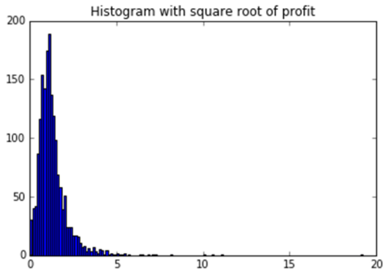
K-means clustering

K-means clustering was done to identify clusters that exist to classify movies as success or failure. K-means was done using imdb score and profits.

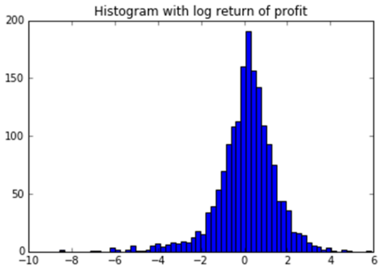
To cluster the profits generated by movies, we first normalize the data. The charts below show the transition on using square root and logarithmic function on profits. Transforming using logarithmic function yields the closest to normalized data, as shown below.



Profit distribution

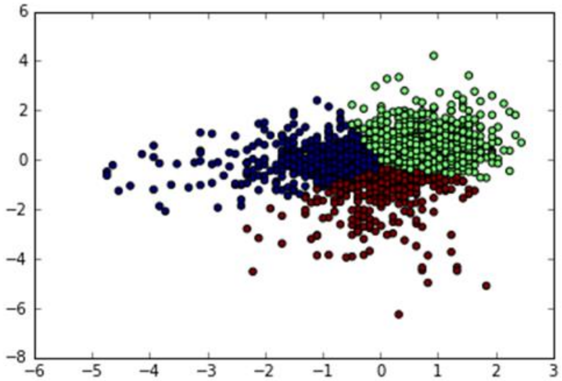


Square root of profits

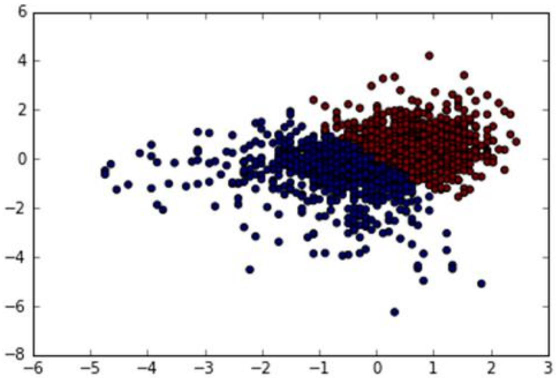


Logarithmic function of profit

In order to find the optimal value of "k", a graph of within cluster variation for each "k" was plotted. We then identified that the highest change or least within cluster variation occurred for "k"= 2 to 3. So we decided to plot the confusion matrix for the "k" values to arrive at optimal "k".

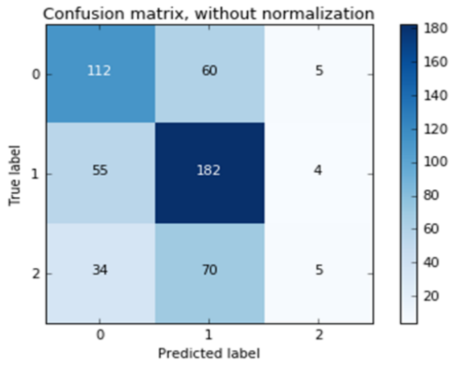


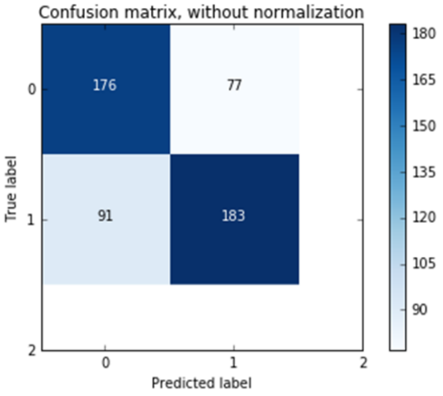
2 clusters



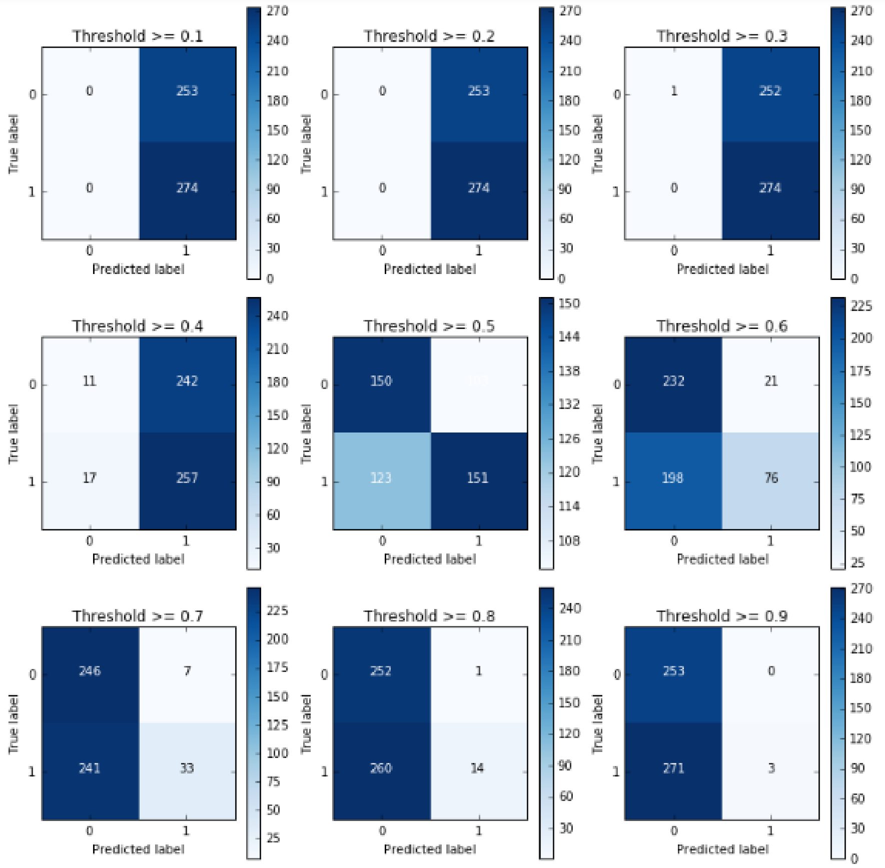
3 clusters

When we set up the confusion matrices we saw a clear difference in the true positives which were much higher for 2 clusters. We chose to continue our analysis with 2 clusters.

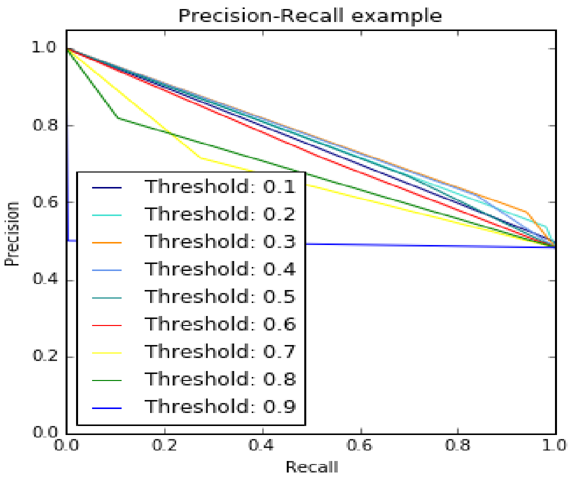


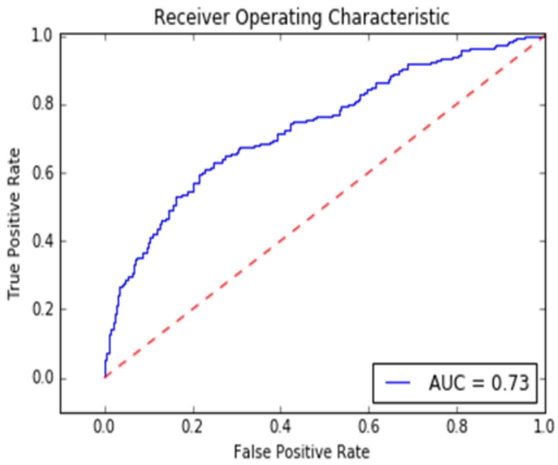


For every parameter, we found the mean recall volume and score using K-fold cross validation, after which we plot the confusion matrices with different thresholds. The images below show the True positive and true negatives for every threshold for the classifier. While in this case, classifying a movie that would be a success to be a failure is accepted, increasing the threshold also reduced the accuracy of predictions. Hence, we made a trade-off and the best thresholds are around 0.6-0.7.



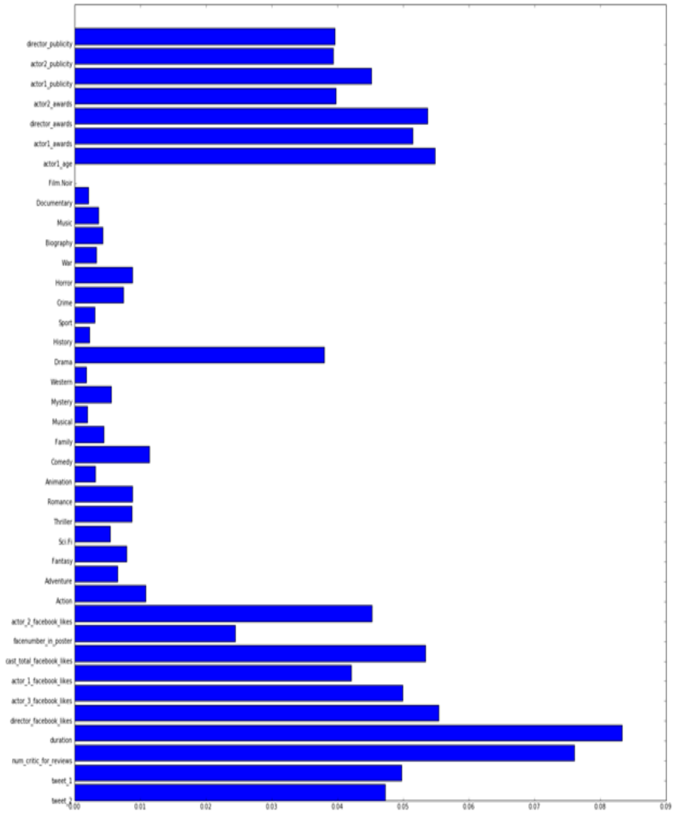
Below is the Receiver Operating characteristic curve,





Thus, using this clustering technique it is possible to classify a movie based on all other features except IMDB score as a success or a failure. Additionally, increasing number of clusters to 4 also gives us an accuracy of 56%. This indicates that the model is does a better job of classifying movies as compared to randomly choosing a cluster (probability of which would 25% for each cluster.).

Random Forest



Feature importance

A random forest was built to gain insight into the feature importances in the dataset. The chart below shows the important features and their relative measures, that contribute to the prediction, with an accuracy of 67.5%.

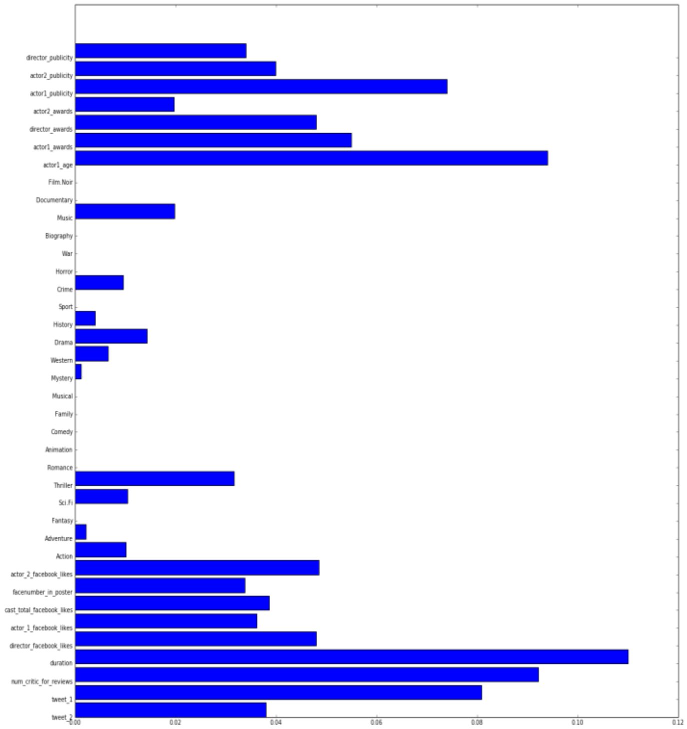
Insights:

1. It can be seen that two features- duration and number of critic for reviews are relatively more important features. This suggests that getting more reviews from critics or inviting more critics to premieres can have a positive impact on the success of a movie
2. This also shows relative importance of genres, which can be useful for production houses to make decisions based on relative importance of genres.

Decision Trees

Another classification algorithm, a decision tree was built to classify movies based on profits. The accuracy of the tree was around 63%. The depth of the tree was decided based on mean accuracy scores of training and test data. The depth of the tree is 9 and the tree considers the following features at a depth of 9. Python's Decision Tree Classifier helps us generate the probability score of all the features. The array of probabilities looks something like this:

```
array([ 0.04,  0.08,  0.09,  0.11,  0.05,  0.04,  0.04,  0.03,
0.05, 0.01,  0.  ,  0.  ,  0.01,  0.03,  0.  ,  0.  ,  0.  ,  0.  ,0.  ,
0.  ,  0.01,  0.01,  0.  ,  0.  ,  0.01,  0.  ,  0.  ,0.  ,  0.02,  0.  ,
0.  ,  0.09,  0.05,  0.05,  0.02,  0.07, 0.04,  0.03])
```



Features in the tree

Conclusion

Based on our model, we recommend producers to concentrate on getting more critics for reviews and on the duration of a movie.
In terms of actors, it helps to have actors who are older (viz. More experienced) and are in general more popular in the entertainment industry.
In terms of genres, family, comedy and action movies generate more revenues and are more successful as opposed to thrillers or horror movies.

Finally**,** this model is only a heuristic for prediction and should not be considered too literally.

Scope for future work

Using the Twitter sentiment of an actor at the time of movie release would help in improving the accuracy of the data drastically.

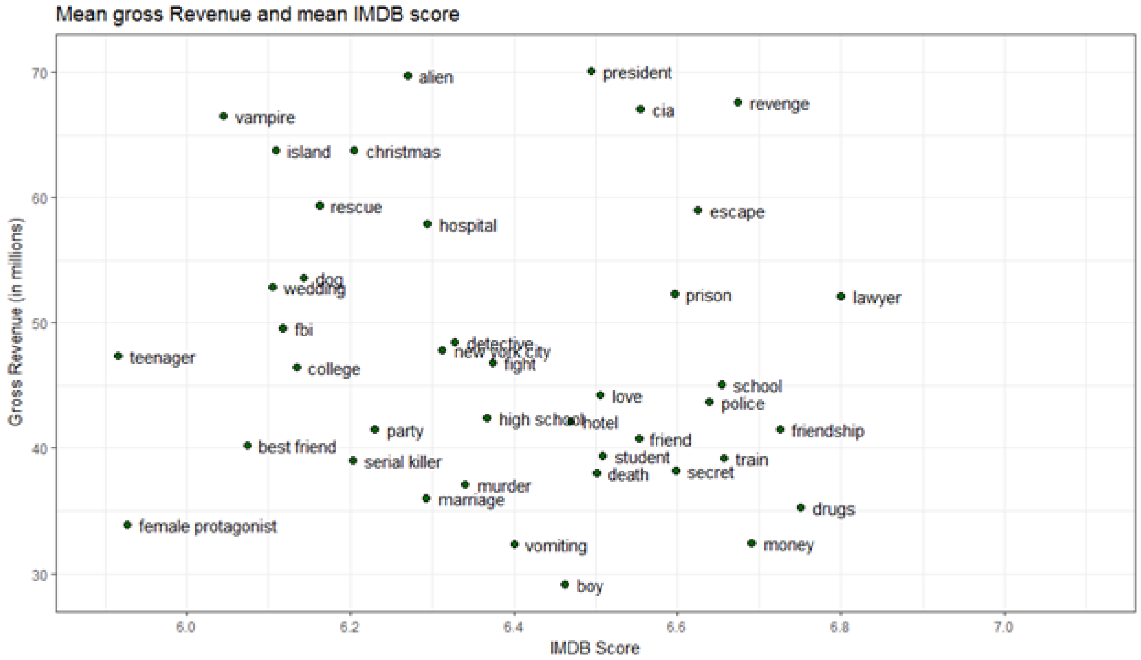
References

1. http://www.imdb.com/
2. www.kaggle.com
3. www.stackoverflow.com

Appendix

In addition to that, we did multiple analysis with individual feature to understand the dependencies better.

Revenue vs Plot Keywords



The chart above shows genres plotted with respect to IMDB scores and Gross Revenues they generate. We can clearly see that a good IMDB score does not necessarily define success and hence has other factors contributing to it.