Group: Unnamed
Project: What Makes A Hit Movie
Member: Qinya Li, Yanghuizi Wang, Xiaohui Li, Xiaoyu Xu, Xiaoxuan Xia

I. Project Overview

1. Introduction

As movies have become important part of our daily lives, people go to see movies for various reasons. Some go for their favorite actors, some go because of other people's good reviews, and some just go for fun. As a group, we think that there is a diversity of potential factors that can significantly contribute to a hit movie, so we made this research, based on some tools we've learned in the course.

2. Data Source

- ***Web Scraping Source:*** Rotten Tomatoes & IMDb
- ***Period:*** 2014 – 1st half of 2018
- ***Total Amount of Movies:*** 2864
- ***Fields:*** Genre, Release Date, Studio, Budget, Open Weekend Revenue, Gross Revenue, Actor, Director, Writer, Meta-Score, Vote, Reviews from Critics, Reviews from Audience

II. Data Analysis

After gathering and filtering the data, we analyzed data to find some interesting patterns and provide useful insights.

1. Genre

Among all 18 genres, "drama" accounts for a large part of the total movies followed by "Comedy", "Thriller", "Action", "Horror" and "Crime". Other 12 kinds of genre only take up approximately 40% of the total movies.

2. Open Week Revenue & Gross Revenue

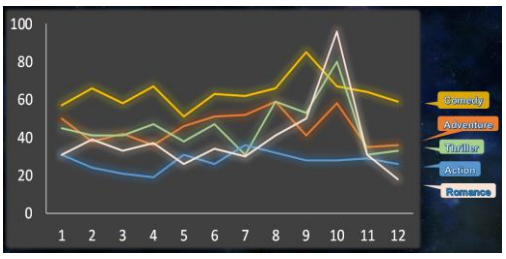
Generally, higher revenues in open week and higher box offices in total usually hand in hand.

3. Actor, Director & Studio

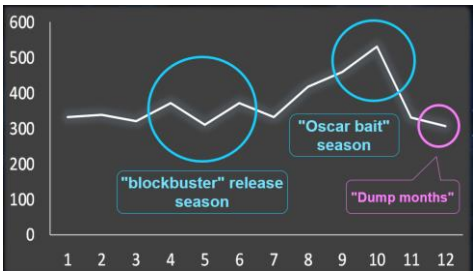
As for movie quality, the boxplot shows that there is no obvious distinction between famous studio and normal studio, as well as famous actors and normal actors. However, the rating of movies directed by famous directors is relatively higher.

4. Release Date

For all genres, the amount of movies released enjoys a surge in "blockbuster" release season, "Oscar bait" season and "Dump months".



Summer is a traditional "blockbuster" release season, when studios release their "Tent-pole movies", because during summer, people can enjoy their vocations watching movies. As for the "Oscar bait" season, studios usually release their "prestige films" to compete for Oscar Prize. Movies released in "Dump months" are usually films that have less prominent stars, cannot be easily marketed.



5. Reviews from critics and audience

Two sides show coherent attitude in general, but they have distinct disagreements on certain types, including comedy, action and entertaining. It makes sense by believing that audience care more about feelings and experiences. They give the film positive reviews and highly recommend it when they're entertained and feel relaxed. On the contrary, critics are much more demanding and consider more about professional criteria, such as plot, character and logic.

6. Sentimental Analysis

We focus on analyzing the general sentiment towards films with text review, specifically along two dimension over Rotten Tomatoes' critic review and IMDb's audience review.

For technical component, we introduce the GloVe[1] word embedding to represent the review sequence as vectors, and further introduce Bidirectional LSTM neural networks to capture the pattern of sequence for sentiment classification. In detail, we use deep learning library Keras and TensorFlow to build the model framework, and applied methods of batch normalization, early stoppage, cross validation, etc. to facilitate training procedure. The best model performance at test set is shown below:

Models / Metrics	Rotten Tomato Critics	IMDb Users
Precision	67%	57%
Recall	64%	46%
F-Score	65%	48%
Accuracy	86%	88%

We also gain insights from analyzing the prediction. In the first place, we implement a real-time sentiment system to monitor mass sentiment over individual movie, which could

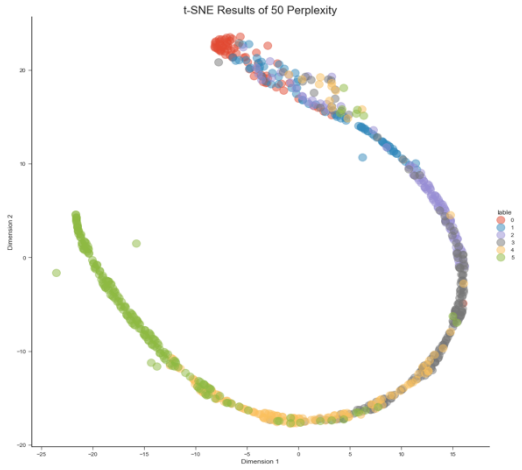
[1] Pennington, et al., 2014. GloVe: Global Vectors for Word Representation

scrape all the IMDb reviews whether rated or not at daily basis and output a pool averaged sentiment index. Moreover, analysis on the predicted sentiment show up a divergence case between authorized critics sentiment and general audience sentiment, reinforcing the perspective difference over critics and audience sentiment. For detailed reference, we created an online report webpage[2].

III. Model Construction

After cleaning data, analyzing data, it's time for us to construct model to predict the Box Office.

From the statistical distribution and also consideration of pratical value, we classfy the Box Office into six type as our predicted task. And we start with an unsupervised learning algorithm t-SNE[3] to get an intuitive image of the outcome. There is a pattern with perplexity of 50, especially for the extreme large and small revenue (label 0 and 5), shedding light on the prediction.



We in total built four machine learning models: Naïve Bayesian Classfier, Artificial Neural Network, SVM[4] and XGBoost. XGBoost stands out when we used four critical metrics to evaluate our models.

1. Choosing Model

Considering the large amount of inputs and potential complexity of their relationship with the target, we select four models that is relative efficient in dealing with the situation.

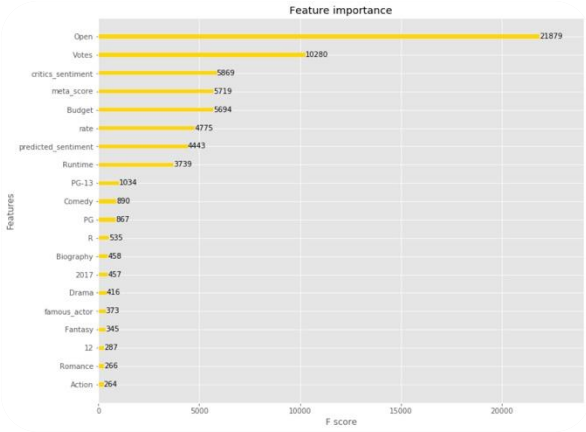
We evaluate our models by four critical metrics on the test set, the result of which is illustrated below.

Models / Metrics	Naïve Bayesian	Neural Network	SVM	XGBoost
Precision	67%	57%	84%	77%
Recall	64%	46%	43%	76%
F-Score	65%	48%	51%	76%
Accuracy	86%	88%	82%	75%

2. Evaluation

XGBoost[5] is an efficient optimized distributed gradient boosting tree model with efficient, flexible and portable feature. In our dataset setting, it validated its robust generalization under test set. And we therefore infer our final model could predict the trend of future box office with opening week feature.

By studying the feature importance of our XGBoost tree model, opening week box office lies as a significant contributor, while our predicted sentiments also play roles in determining the final revenue.



IV. Conclusion

Big-deal sci-fiction movies flooding in during the summer vacation bring a feast for eyes to the audience, though not always acquiring corresponding praises. Instead, during October and November, the grand exhibition of Oscar-level hit movies is likely to make you a surprise.

What matters for a movie investor is the combo of famous director with all-star actors which contributes to a guarantee of considerable box office revenue. Nevertheless, open-week marketing is never less important as for promoting the earnings. At length, what equips a movie with the sustainability to survive the time is of course, its quality.

[2] https://xiaohui-victor-li.github.io/projects/film-sentimental-analysis/

[3] L.Maaten, G.E.Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE

[4] implement with scikit-learn library

[5] https://xgboost.readthedocs.io/en/latest/