Facebook Marketing Plan for Movies

MSBA 7012 Class B Group 2

Chen Silu, Ding Heyi, Guan Qianqi, Guo Yijun, Lam Ho Ki, Shen Shulin, Xiao Jinwen, Xu Qingyu

1 Introduction & Executive Summary

1.1 Project Idea

Movie production is undoubtedly one of the most expensive and risky investments in today's world. Producers would make every effort to secure a higher gross revenue. The entertainment industry goes hand in hand with the social media platforms, hence it is vital for companies to understand how to utilize the tools to maintain its presence in the social media platforms e.g. Facebook to promote their movies, engage with their supporters and ultimately boosting the box office revenues. As a marketing company, we assumed a movie producer came to us and requested for an effective digital marketing strategy on Facebook for their multimillion movie productions. Our goal is to analyze the marketing behaviors of previously released films and design an effective Facebook marketing activities for various types of movies to increase user engagement and maximize the box office revenue.

1.2 Data and Methodology

The data we used in this project are the 5 datasets supplied by the client namely *'boxoffice'*, *'fandango_review'*, *'fbcomments'*, *'fbposts'* and *'imdb_movie_overview'*. It contains the characteristics of 203 movies, their promotional behaviors on Facebook and respective responses from audience. To tackle the problem, we performed a 5 step analysis as below: -

Steps	Methods
1. EDA, Data Pre-processing and Feature Engineering	To explore the movie data, summarize key characteristics and patterns and preprocess the data
2. Movie Page Segmentation	Perform a 2-step clustering methods. Group movies according to their features and identify clusters that have similar promotional activities
3. Marketing Performance Analysis	Find out what kinds of Facebook marketing would lead to higher box office revenue. Build machine learning models, use RMSE as evaluation criteria and find out the determinant variables using feature importance
4. User Engagement Analysis	Identify the kinds of user engagement that would result in a higher box office revenue. Conduct text mining on the Facebook comments to see how movie fans react to these promotions
5. Insights and Recommendations	Suggest the suitable marketing plan for different clusters to be adopted by the movie makers

1.3 Key Findings and Results

Our goal is to provide a solution to the film producers by **first identifying the cluster** that their movies belong to and suggest corresponding **marketing plan** for their movies on Facebook and the **user engagement goals** that are optimal in generating a higher gross revenue.

By conducting the clustering analysis, **3 clusters** were identified based on the Genres. We labelled them as **Intense Movies**, **Relaxing Movies** and **Humanities & Art Movies**. We understood that the marketing plan and user engagement might be different for various type of movies. Hence, we further performed a **sub-clustering analysis** to discover **5 important features** (promotion start time, average no. of posts before debut, post type, timing to post and post format) and provided some intuitions for further analysis.

We identified the important features for each category and derived a **recommendation plan** for the 3 clusters of movies respectively. The suggestions are based on **3 dimensions of time** i.e. before movie release, while movie on shows and movie off shows. Meanwhile, we also found out the significant features and provide **recommendations on the user engagements** that the company should put effort to build for respective cluster. Details of our recommendations are listed in section 4.2 and 4.3.

2 Data

To facilitate our analysis onward, we have to preprocess the raw data and perform data cleaning. We extracted useful information and integrated the data into 3 new tables namely *'imdb_clustering.csv'*, *'movie_t123.csv'* and *'audience.csv'*. These tables would be used for subsequent clustering and modeling. The procedures are briefly described as below:

2.1 Data Preprocessing

2.1.1 New Table *'imdb_clustering'*

'imdb_clustering.csv' is derived from the original dataset *'boxoffice'* and *'imdb_movie_overview'*. We selected the columns that are of higher relevancy to our analysis and performed data cleaning before combining them into the new file. Some key transformations are listed below. (For preprocessing details, please refer to our codes with comments added for explanations)

Variables	Data Preprocessing
'awards'	There are six types of awards. Referencing to the importance of respective awards in the industry, each award is assigned with different weights to calculate the total score: **(normal)nominations:** 1 point; **(normal)wins:** 2 points; **Nominated for Golden Globes:** 3 points; **Nominated for BAFTA Film Award:** 5 points; **Nominated for Oscar:** 7 points; **Won Oscar:** 10 points
'genres', 'country', 'language'	Use **0/1** to represent whether there is a genre/country/language tag. **For 'genres':** there are 21 types corresponding to 21 columns. For example, *'Action|Comedy'* would be assigned 1 to *'genres_Action'* and *'genres_Comedy'* and be assigned 0 to other genres **For 'country':** Countries with higher frequency, *'United States'*, *'United Kingdom'*, *'France'* and *'Canada'* are extracted, and the rest of the countries is assigned to *'other'* **For 'language':** Languages with higher frequency, *'English'*, *'French'* and *'Hindi'* are extracted and the rest of the countries is assigned to *'other'*
'show_time_week'	Equal to *'release_endtime'* minus *'release_starttime'*
'boxoffice' dataset	Extract *'opening_wkd'*, *'gross_us'*, *'release_starttime'*, *'release_endtime'* and combine to this file

2.1.2 New Tables *'merge.csv'* and *'audience.csv'*

'merge.csv' mainly contains the promotional behaviors of each movie which included the average number of posts per week and the type of posts. It is a table formed from '*movie_t123.csv*' joined to ' *movie_info.csv'* by the unique primary

key imdb_id. *'audience.csv'* mainly contains the audience response behaviors such as average number and total number of likes, comments and shares. *'fbposts'*,*'boxoffice'*, *'fbcomments' and 'fandango_review'* were used to create these new tables. We also included the time information to provide additional insights for the analysis in next step.

Label	Corresponding Time
t1	More than 2 months before the release
t2	Within 2 months before the release
t3	Within 1 month before the release
t4	Movie Release
t5	Within 1 month after release
t6	Within 2 months after release
t7	More than 2 months after release

We extracted the time corresponding to each post and divided it into 3 large dimensions and 7 small dimensions to suit different analytic needs. To look at the promotion time periods in a more holistic picture, we divided into 3 time periods in '*movie_t123*': **(t1) before the release**, **(t2) during the release** and **(t3) after the release**. In order to analyze the average weekly post in more detail, we classified the corresponding time period into 7 dimensions in *'movie_info.csv'*.

2.2 Exploratory Data Analysis

2.2.1 Findings from Movie Features

Two strong positive correlations were found between *'gross_us'* and *'opening_wkd'*, *'release_starttime'* and *'release_endtime'*.

	mpaa_rating	running_time	budget	opening_wkd	gross_us	awards	show_time_week	release_endtime	release_starttime
mpaa_rating	1.000000	-0.050356	0.279857	-0.157528	-0.142674	-0.086780	0.012850	0.016070	0.010983
running_time	-0.050356	1.000000	0.392422	0.279608	0.351822	0.458095	-0.102482	0.153617	0.241323
budget	0.279857	0.392422	1.000000	0.268073	0.262447	0.052803	-0.080132	-0.082321	-0.055039
opening_wkd	-0.157528	0.279608	0.268073	1.000000	0.949269	0.207382	0.162717	0.037656	-0.055134
gross_us	-0.142674	0.351822	0.262447	0.949269	1.000000	0.386731	0.256409	0.161258	0.032481
awards	-0.086780	0.458095	0.052803	0.207382	0.386731	1.000000	0.424129	0.474341	0.296853
show_time_week	0.012850	-0.102482	-0.080132	0.162717	0.256409	0.424129	1.000000	0.520468	-0.000324
release_endtime	0.016070	0.153617	-0.082321	0.037656	0.161258	0.474341	0.520468	1.000000	0.853712
release_starttime	0.010983	0.241323	-0.055039	-0.055134	0.032481	0.296853	-0.000324	0.853712	1.000000

Figure 1. Correlation Table Between Movie Features

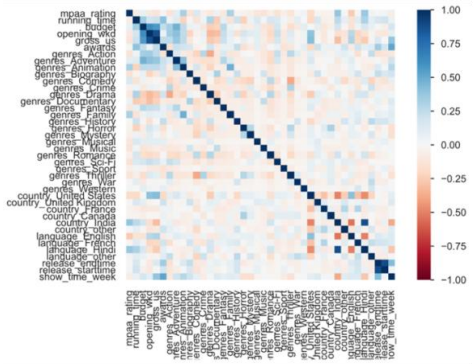


Figure 2. Heatmap Analysis

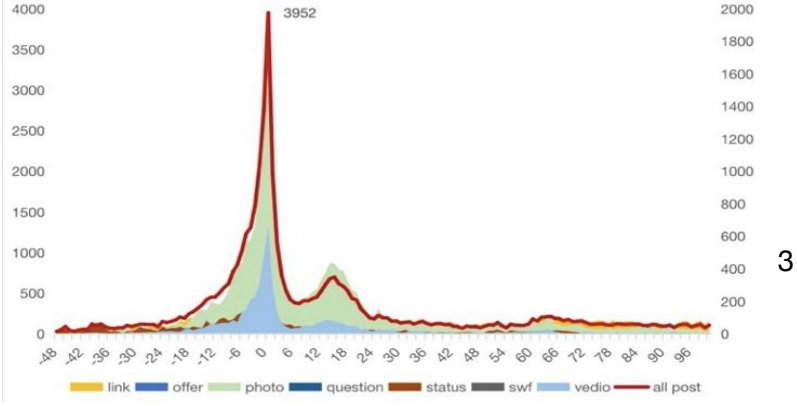
From the genre perspective, 'Adventure' and 'Fantasy' movies generally have **more budgets** while 'Adventure', 'Animation', 'Fantasy', 'Musical' and 'Western' labels have **higher gross revenue**. When movies have 'History', 'Musical' and 'Western' labels, they tend to get **more awards**.

In terms of language, we found that movie with an English version is **13.9 times more box office** than that of a movie without an English version. In the categories with English label, movies that translated to 'French' and 'Hindi' have **more budget and awards**.

From the country perspective, movies released in the United States have **17 times more box office** more than movies not released in the United States. Movies with the 'India' country tag have a **higher budget as much as 11 times** than a movie without the tag.

2.2.2 Findings from Promotional Behaviors

We found that the marketing posts are concentrated in 6 weeks before the release and 24 weeks after the release. There is a high peak and a low peak 3 weeks before release and 15 weeks after release



respectively. Besides, photos and videos are seemingly two most popular types of promotions.

Figure 3. Post Frequency Table

2.2.3 Findings from Movie Fans

We found that the total number of likes, comments, and shares for movie posts reached a peak in the first week of release. This may be due to the high number of posts launched by the marketers in the first week of movie launch. The average number of likes, comments, and shares peaked in weeks 8-10 may be due to the high discussion during this time period.

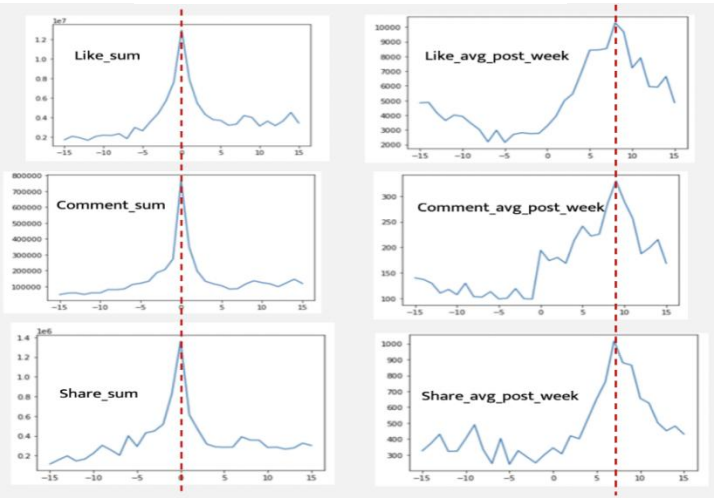


Figure 4. Audience Reaction Analysis

3 Methodology

3.1 Text Mining

3.1.1 Text Preprocessing

From the above new tables, we conducted several steps to clean the text. Firstly, we removed the punctuations and abnormal characters. Meanwhile, Emojis such as '**:)**' , '^^' may indicate some information in the sentiment analysis. We managed to keep these Emojis by creating a dictionary and replaced them with corresponding words. Secondly, we split the words and performed stemming. Lastly, we output the results into 3 new files for downstream analysis.

3.1.2 Sentiment Analysis

Based on the cleaned text data we got, we used 'VADER' package to analyze the sentiment of the text. 'VADER' labels texts according to their semantic orientation as either positive or negative and tells us about how positive or negative a sentiment is. The compound scores of the 'VADER' output tell us how positive or negative a sentiment is. When compound score < 0, it means negative, otherwise it's positive. For example, "looking amazing", its sentiment score is 0.5859 and its sentiment is positive. We assigned a sentiment score to every comment in the dataset accordingly.

	imdb_id	comment_text	sentiment_score	sentiment
0	tt1409024	looks amazing	0.5859	Positive
1	tt1409024	haha smith always main guy riddick comes	0.4588	Positive
2	tt1409024	looks mint n funny	0.4404	Positive
3	tt1409024	whooo smith	0.0	Positive
4	tt1409024	NaN	None	None
...	...	...	...	...

Figure 5. Sentiment Analysis (Positive)

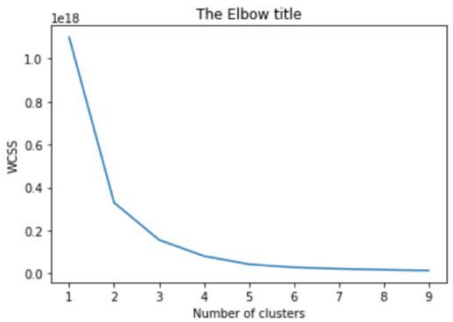
	imdb_id	posttext	sentiment_score	sentiment
0	tt0485985	neyo dont want play singers irish independent ...	-0.6312	Negative
1	tt0485985	neyo dont want play singers belfast telegraph ...	-0.6312	Negative
2	tt0485985	neyo like roles little nothing regular life me...	0.7003	Positive
3	tt0485985	film sheds ligh: plight lakotas san francisco ...	0.9618	Positive
4	tt0485985	disney buying lucasfilm b wdth wdtndisney buyi...	-0.34	Negative
...	...	...	...	...

Figure 6. Sentiment Analysis (Negative)

3.2 Clustering

We conducted a two-step clustering to find out the best promotional behaviors for a specific group of movies with common characteristics.

3.2.1 Step 1: Movie Features Clustering



This clustering step is based on the idea that movies with different characteristics may have different emphases when conducting promotion strategies. We tried two clustering methods, K-means Clustering and Hierarchical Clustering. For the K-means clustering, we used the within-cluster sum of squares (i.e. variance) as an evaluation metric to choose the best number of clusters (i.e. k) and eventually chose k=3. For the hierarchical clustering, we tried 6 methods: 'centroid', 'single', 'complete', 'weighted', 'median' and 'ward'.

We found that K-means clustering gave us better results, hence we would continue to use the output from this method for the analysis onwards. The variables we included in clustering are *'Budget', 'Running time', 'Genres', 'Show time', 'Country', 'Language'* and *'Mapp rating'* from the *'imdb_clustering'* table.

Figure 7. Clustering Elbow Method

3.2.2 Step 2: Movie Promotional Behaviors Clustering

This step is conducted in each of the clusters we obtained from the first step. We would like to compare and find out the best performing sub-clusters in respective clusters by using the 'gross revenue' as the evaluation metrics. Further analysis of the promotion behaviors in the selected best performing sub-cluster would be carried out to get an initial insight for the features importance of different clusters of movies.

We used K-means clustering again to identify the sub-clusters. Within-cluster sum of squares (i.e. variance) is used to choose the best number of clusters (i.e. k) in each cluster. We eventually used the result of k=2 in the first cluster, k=3 in the second cluster and k=2 in the third cluster. The variables we used are the publish time, number of post and various post types in all the 3 time periods in the *'merge'* table.

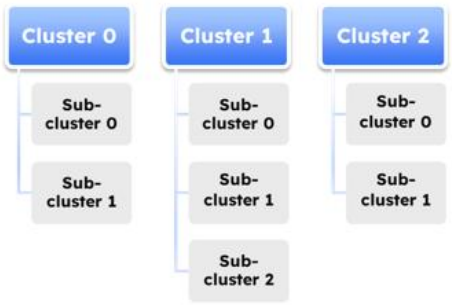


Figure 8. Clustering Logic Flow

3.3 Marketing Behaviour Analysis

To analyze the promotional behaviors of the movies, we would first summarize and select the key features for each cluster and interpret the impact of these features on the marketing behaviors of the movies. Inspired from the sub-cluster analysis, we extracted features that are important to drive a higher gross revenue such as the post publish time, number of posts published at different time period, post types such as video, status, photo, link, and question for further analysis. To better understand the business meaning of the post contents, we applied text mining techniques to analyze the text in the posts.

Firstly, we used an automated web crawler 'Octoparse' to crawl data from IMDB official website to complement our data. We got the movies' overviews, casts' names, directors' names and writers' names from the website. Secondly, we conducted text cleaning which included removing stop words and name identification using bigram. Thirdly, we counted the number of times that the casts, directors, and writers' names were mentioned in each post. We calcultated the star-rate, director-rate, and writer-rate for each post by dividing the count by the number of posts accordingly. Lastly, we calculated the cosine similarity scores of the posts and movie overview based on the tf-idf.

3 tree-based models were built to identify the key features and explored the correlation between the features and movie gross revenues in respective clusters. Linear regression was used to provide further insights on our conclusions. We tuned the models and selected the best hyper-parameters by minimizing the RMSE. We selected the features that were of significant importance by integrating the results of the 3 tree-based models. The way we select the features is to list out the top 10 important features output by the 3 models and look at the occurrence of the features in these 3 lists.

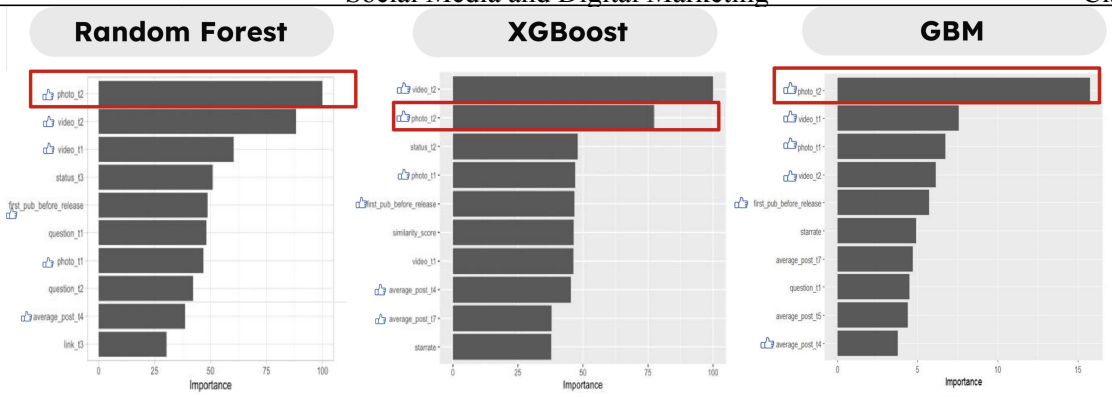


Figure 9. Feature Importance Outputs

Take cluster 2 as an illustration. *'photo_t2'* was considered as an important feature as it ranked in the top 2 of all 3 models. We would then focus on analyzing the correlation between the extracted features and the movie gross revenue. The same logic applied to the other two clusters. After we identified the key features, we would output and interpret their pdp plots.

3.4 Users Engagements on Facebook

To analyze the audience engagements, we built 3 tree-based models which included Random Forest, GBM and XGBoost for each cluster and found out the most important feature that would affect the gross revenue. We tuned the hyperparameters for each model by minimizing the RMSE. Then, we would focus on the top 4 important features of each model. The total score of the features would be calculated by referencing the ranking of the features in each model; rank 1 in each model would be assigned with a weight of 4, rank 2 would be assigned with a wight of 3 and so on. Features with a higher score are considered as more important. After choosing the important features, we would plot pdp to visualize the relationship between the box office revenue and the selected features.

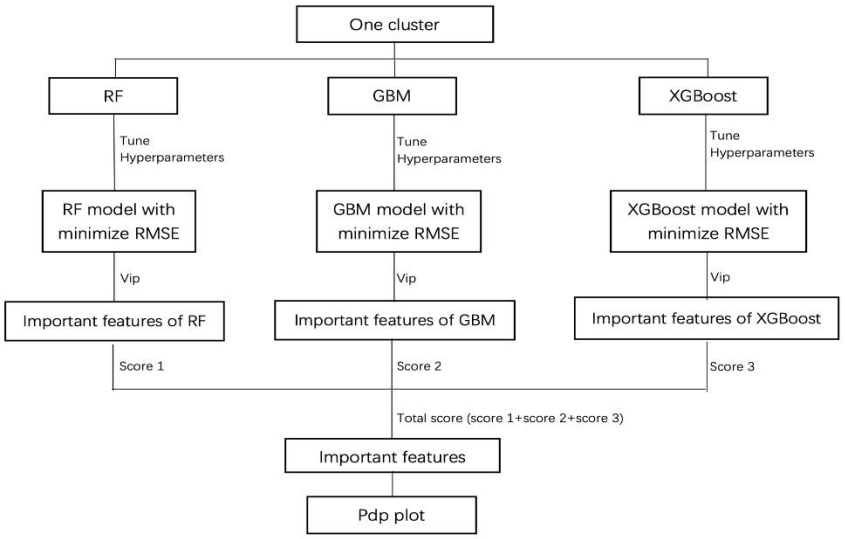


Figure 10. User Engagement Analysis Logic Flow

4 Results

4.1 Clustering Analysis

4.1.1 Step 1: Movie Features Clustering

We obtained 3 clusters from this step and labelled them based on the most obvious features 'Genres'. Intense, Relax and Humanities & Art movies were identified. The distinctiveness and the important characteristics of the 3 clusters are summarized and presented in the following table: -

	Intense Movies	Relaxing Movies	Humanities & Art Movies
Total no. of Movies	49	78	73
Genres	Action, Crime, Horror, Mystery, Sci-Fi, Thriller	Comedy, Family, Romance	Adventure, Biography, Drama, Documentary, Fantasy, History, Music
Mean Budget	Highest: $75,022,694	Middle: $60,264,426	Lowest: $41,159,312
Running Time	Middle: 106 min	Shortest: 103 min	Longest: 113 min
Show Time	Shortest: 10 weeks	Middle: 14 weeks	Longest: 15 weeks
Countries (besides USA)	France, Canada	India	UK and other countries
Language (besides English)	Other languages	Hindi	French
MPAA Rating*	Mainly 1 & 2 but not include 5	Mainly 1, 2, 3, 4 but not include 5	All ratings

*MPAA Rating: 1 = 'PG-13', 2 = 'R', 3 = 'PG', 4 = 'Not Rated', 5 = 'G'

4.1.2 Step 2: Movie Promotional Behaviors Clustering

From the result of the second step clustering, we further divide the intense movies into 2 subclusters, relaxing movies into 3 subclusters and Humanities & Art movies into 2 subclusters. To find out the best promotional behaviors in each cluster, we compared the gross revenue of the subclusters in their respective clusters. Since subcluster 1 in all the 3 clusters has a higher average gross revenue, we would analyze the subclusters in detail.

There are some common characteristics across these subclusters. Firstly, the marketing efforts would be focused in 8 weeks before the movie launch. Secondly, the most popular post type are photos, link, status, video and lastly questions. Thirdly, marketers preferred creating notes more than launching an event on Facebook wall post. To deduce the difference in these clusters, we first calculated the mean for all the variables and explored the dissimilarity in their respective best performing sub-clusters. A few initial observations were spotted in this stage and summarized in the following table:

	Intense Movies (Sub-Cluster 1)		Relaxing Movies (Sub-Cluster 1)		Humanities & Art Movies (Sub-Cluster 1)	
Avg. gross revenue	$153,500,000		$60,294,740		$43,694,940	
Promotion start time	85 weeks		48 weeks		33 weeks	
Average no. of posts before debut	45 posts		33 posts		23 posts	
Timing for post type	Video, Status, Link, Question	Before	Video, Status, Link, Question	Before	Video, Status, Link, Question	Before
	Photo	**After**	Photo	**Before**	Photo	**On Show**
Post type	Photo → Link → Status → Video → Question					
Post Format	Story > Post	**Before**	Story > Post	**Before**	Story > Post	**Before + On Show**

By conducting the above preliminary analysis, we could generate some initial insights on each cluster and acted as an indicator for the key features for the next section.

4.2 Marketing Behaviour Analysis

Below are the key features of the three clusters we selected from the feature importance.

Intense: *'first_pub_before_release', 'average_post_t2', 'average_post_t4', 'video_t2', 'photo_t2', 'similarity_score', 'starrate'*

Relaxing: *'average_post_t3', 'average_post_t4', 'photo_t3' , 'similarity_score', 'photo_t1', 'photo_t2'*

Humanities & Art: *'first_pub_before_release', 'average_post_t4', 'video_t1', 'video_t2', 'photo_t1', 'photo_t2'*

Selected features were further analyzed by using the partial dependence plot and we summarized the best practice for marketers. Take *'photo_t1'* as an example, we can see that there are several turning points on the line. As indicated by the dotted line, when *'photo_t1'* ranges between 0.2 and 0.5, the gross revenue reaches the highest value. After de-standardized the x-axis, we found that the best range of *'photo_t1'* is between 45 to 65. In other words, marketers should post 45-65 photo posts before the movie release to maximize the revenue.

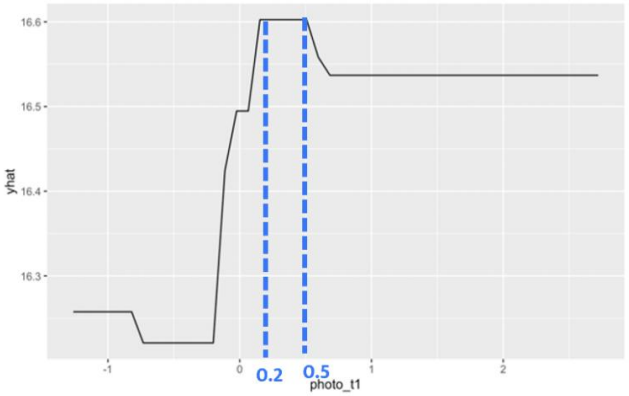


Figure 11. PDP Illustration

➢ **Marketing Recommendations For Intense Movies**

	Before Release	On Shows	Off Shows
No. of posts per week	> 11 posts before 1 month of release	3 – 4 posts	/
Type of posts	> 48 photo posts	45 – 65 photo posts	33 – 43 photo posts
Content relevancy with movie overview	Cosine similarity score > 9		

➢ **Marketing Recommendations For Relaxing Movies**

	Before Release	On Shows	Off Shows
No. of posts per week	> 2 posts before 1 – 2 month of release	< 8 posts	/
Type of posts	/	> 3 video posts > 65 photo posts	/
Content relevancy with movie overview	Cosine similarity from 8 - 18		
Frequency of star names	> 0.6 in each post		

➢ **Marketing Recommendations For Humanities & Art Movies**

	Before Release	On Shows	Off Shows
Publish Time	> 5 weeks before release	/	/
No. of posts per week	> 48 photo posts	< 6 posts	/
Type of posts	> 20 photo posts > 7 video posts	< 48 photo posts > 3 video posts	/

4.3 User Engagement Analysis

Intense Movies: the average number of comments per post during on show and the positive rate of audience reviews stood out from other features. As indicated by pdp plots, while movies are on show, film producers should take measures to ensure that there are more than 170 comments on each post and to acquire as high positive rate in the comment as possible.

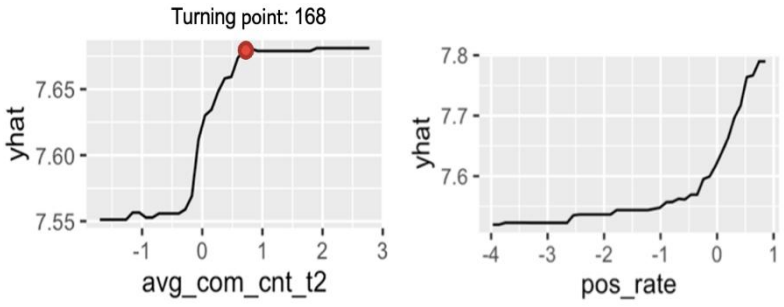


Figure 12. PDP Plot Illustration (Intense Movies)

Relaxing Movies: the average star rating and positive rate are found to be the important features in this category. Companies should try their best to keep a high rating and high positive rate. However, companies do not need to pursue high rating and high positive rates endlessly since overly high reviews may create an opposite effect to user engagement. It would be better to have some negative reviews in this category of movies.

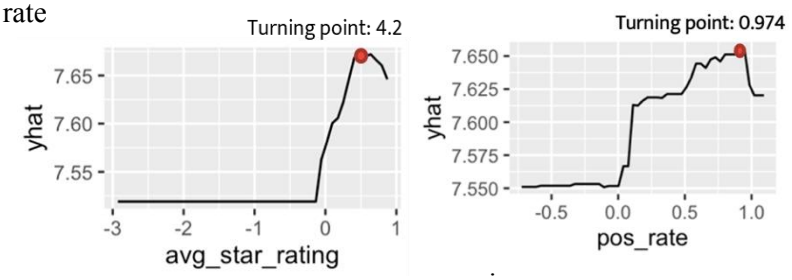


Figure 13. PDP Plot Illustration (Relaxing Movies)

Humanities & Art Movies: the average number of comments per post during on show and the positive rate of audience reviews are found to be important. While on shows, film producer should ensure that there are at least 125 comments in each post. For positive rate, 92% would be an optimal goal for company to pursue. We would suggest marketers to allow the existence of different voices and never pursue a very high positive rate in this cluster.

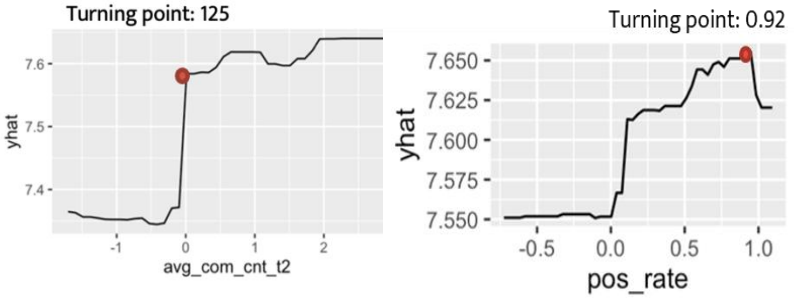


Figure 14. PDP Plot Illustration (H&A Movies)

5 Limitations and Future Work

Limited Sample Size: There are only 203 movies in the entire dataset, of which only 180 contained promotional behaviors. Hence we have limited sample size for tree-based model training. Therefore, the partial dependence plot curve may not reflect a very accurate evaluation of the promotion behaviors. If more films could be included and analyzed, we could get a more convincing relationship between the promotional behaviors and the box office revenue. The credibility of our recommendations would be higher.

Text Mining: We have concluded that more likes, comments and shares have a positive impact on the box office. Further analysis on the kind of contents that leads to higher number of these interactions could be carried out. We have conducted word embedding and topic modelling analysis, yet the results were not very significant and not optimal to be included in this report. In future, we hope to further deepen the analysis between the text content and the corresponding audience interactions.

Evaluation on our recommendations: Due to limited sample size, we were not able to test our recommendation on a testing dataset. We tried to use movies without promotional behavior as a benchmark model to see if our recommendations could help them to generate a higher revenue. However, there were not enough data, and the results were not representable. For future work, we hope to gather more data so that we could separate the dataset into training and testing to see if our recommendations would be helpful to provide some indications for the movie makers to adopt in the social media platform.

6 Work Allocation

Name	Student ID	PIC
Chen Silu	3035882873	Data preprocessing, EDA, Marketing behavior analysis
Ding Heyi	3035883736	Data preprocessing, EDA, User engagement analysis
Guan Qianqi	3035879553	Data preprocessing, EDA, User engagement analysis
Guo Yijun	3035884510	Text Mining (Sentiment Analysis), Clustering Analysis
Lam Ho Ki	3035880540	Text Mining (Word Embedding), Clustering Analysis
Shen Shulin	3035883293	Text Mining (Sentiment Analysis, Topic mining), Marketing behavior Analysis
Xiao Jinwen	3035876771	Text Mining (Word Embedding), User engagement analysis
Xu Qingyu	3035877373	Text Mining (Preprocessing), Marketing behavior analysis

* PIC were responsible for the code, analysis, interpretation, presentation and reporting for each item