

Data Mining Final Group Project

Minkyung Kim(Monica) / Yunjie Lu(Yena) / Nai-Chieh Yang

Agenda



01

Project Introduction & Overview



02

Project Findings & Model Evaluation



03

Final Recommendation & Conclusion

01

Project Introduction & Overview

What is MUBI?

- MUBI is an art movie streaming platform + community + publication
- 30 movies are handpicked by curators daily (No recommendation system)
- 7-day free trial
- Users can leave ratings and comments regardless of being a paid user or not
- Dataset retrieved from Kaggle (not used on any competition)



Q Search

LIVE

NOW SHOWING

LIBRARY

FEED

NOTEBOOK



NOW SHOWING

TRY 7 DAYS FREE

FILM OF THE DAY



The Problem

Founded in 2007, the same year Netflix platform was launched, MUBI hasn't been able to grow as much as Netflix. In 2015, while Netflix had 75 million subscribers, MUBI had 100k subscribers. Post-Covid19 era is a growth opportunity for movie streaming platforms and MUBI should seize the opportunity through identifying critical attributes for turning non-paying users to subscribers and predicting subscribers to boost growth.

Our Hypothesis

- A user actively rating movies and leaving reviews is more likely to become a subscriber
- Number of movies rated by users during trial period will have high impact on conversion
- Hand-curated films play a big role on turning free or trial users into paying users

02

Project Findings & Model Evaluation

Exploratory Analysis | Overview

List Data

- 23k Users created 80k lists
- 38 movies in list on average
- Max number of lists created by one user is 1263

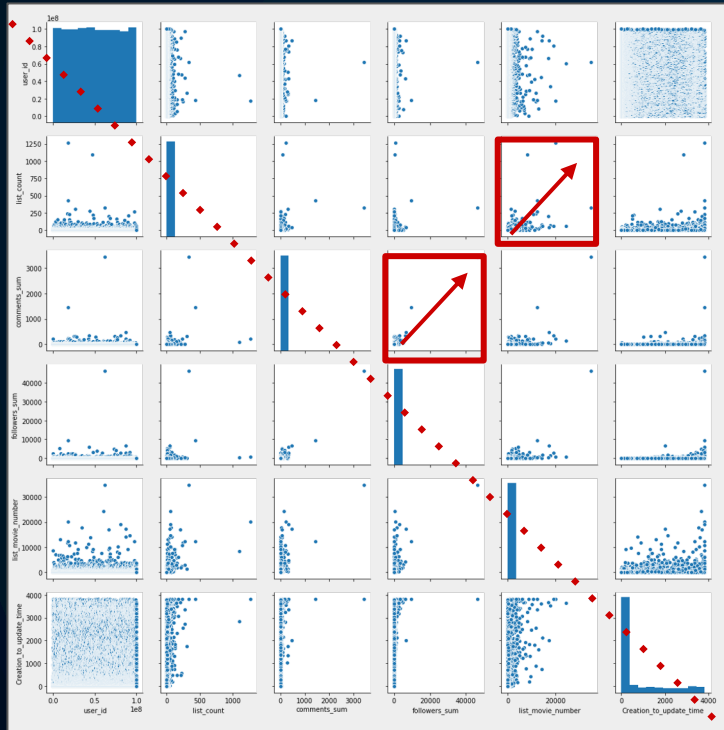
Movie Data

- 226k Movies by 95k Directors
- Popularity('Love' by users) ranges 0 - 13989

Rating Data

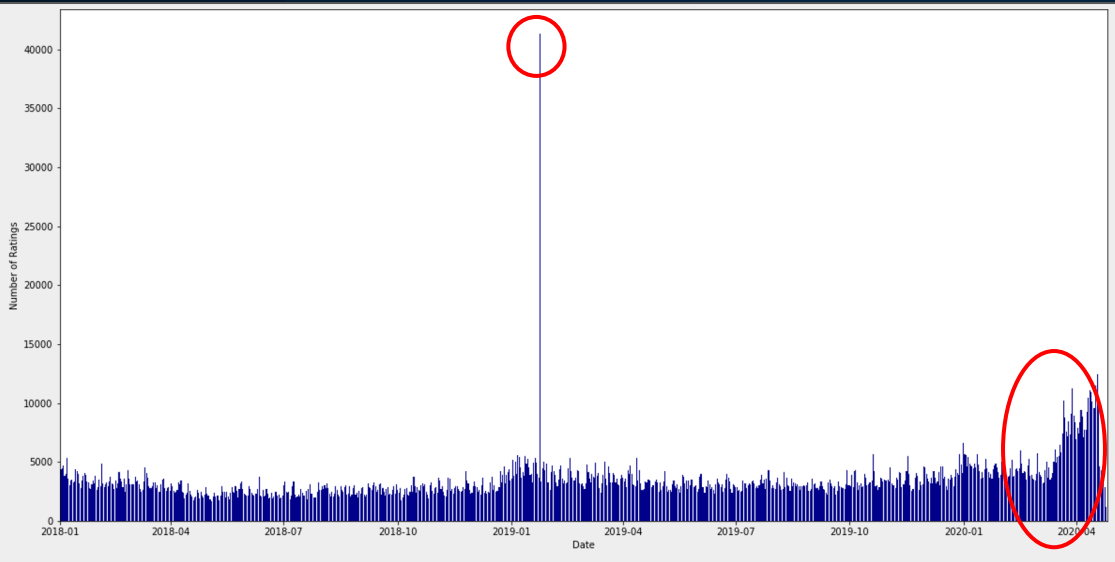
- 15.5million ratings for 142k movies
- Overall mean rating 3.59
- 5.3million ratings were 4

Exploratory Analysis | List Data



- Number of lists has a positive relationship with the number of movies in the lists.
- The number of followers has a positive relationship with the number of comments.

Exploratory Analysis | Rating Data



- Number of ratings and unique users rating daily increased post-Covid19
- Some kind of event happened on 01/24/19 that attracted a lot of users to rate movies(highest of all time) but also suspicious users exists

Exploratory Analysis | Rating Data

Year	Movie Count	Movie popularity	Average Rating
1878~1950	10,980	945,853	3.93
1950~2000	55,305	7,185,597	3.77
2000~2020	76,358	7,388,457	3.37


Users love old movies but there are a smaller number of old movies.

Exploratory Analysis | User Data

- Among 451k users(who rated a movie at least once), 23% are subscribers, 8% are trial users.
- One user rated 34 movies on average
- 5.1% of the users created a movie list
- All the trialist will become a subscriber

	Trialist = Yes	Trialist = No
Subscriber = Yes	36,575	67,661
Subscriber = No	0	347,521

→ Potential customers

- User ID 19983 is probably an imposter! 
 - Not a subscriber, rated 20,000 movies with 1.39 average rating score, rated 14 movies per day for 1428 days

Model Evaluation

2 Binary classification problems (predicting subscribers, predicting trialists) with a dataset of 21 variables and 451k rows

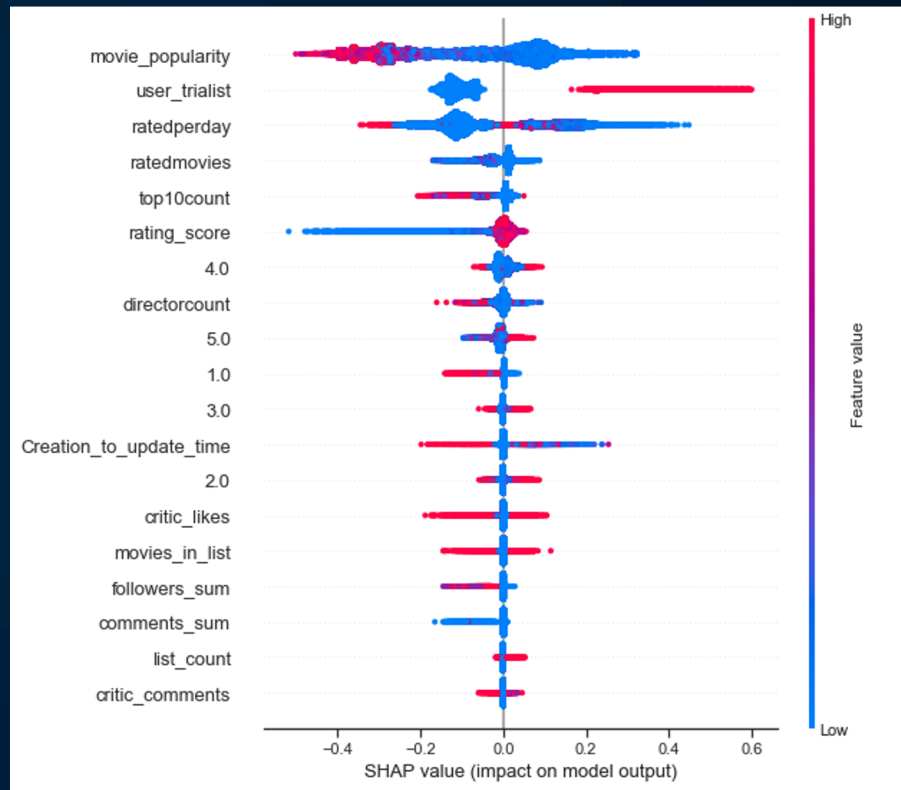
- Logistic Regression(+PCA for selecting features)
- Random Forest
- XGBoost (+Hyperopt)
- LightGBM (+Hyperopt)

Model Evaluation | Predicting Subscribers

	XGBoost	LightGBM	Random Forest	Logistic Regression
Test AUC	0.9007	0.8657	0.8756	0.7780
Train AUC	0.9145	0.8685	0.8725	0.7790

XGBoost-SHAP Insights

- Low movie popularity tend to have a positive impact on becoming a subscriber
- Trial users have a higher tendency of becoming a subscriber
- High number of rated movies per day can either have negative impact or no impact on the model output

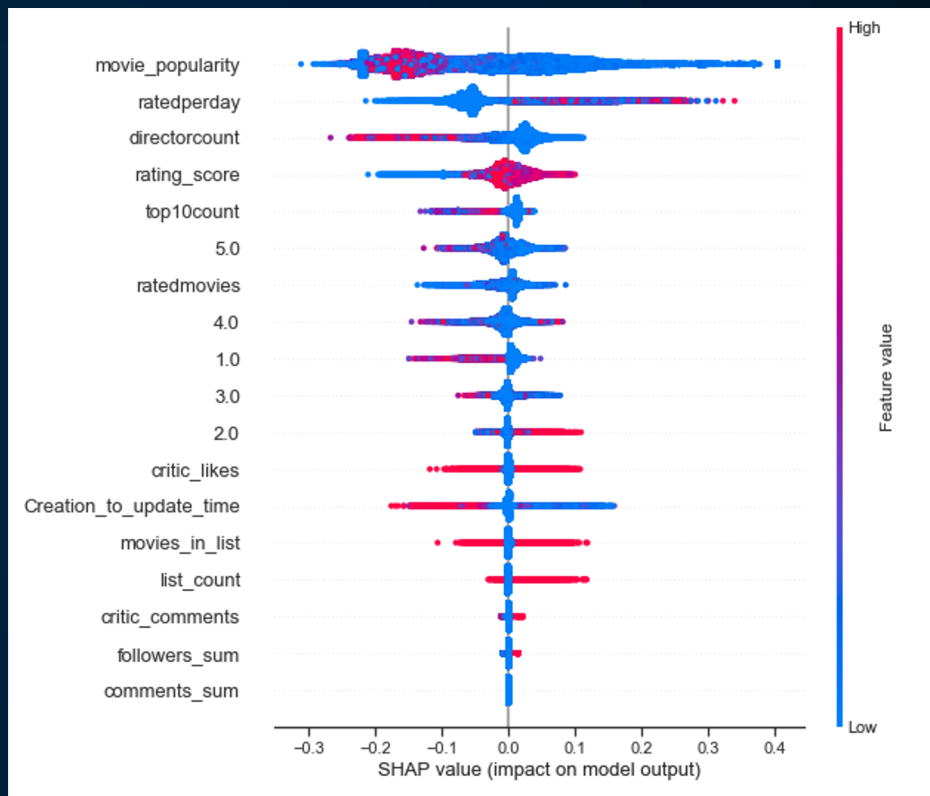


Model Evaluation | Predicting Trialists

	XGBoost	LightGBM	Random Forest	Logistic Regression
Test AUC	0.7795	0.7515	0.6803	0.6369
Train AUC	0.8061	0.7516	0.6828	0.6371

XGBoost-SHAP Insights

- High movie popularity tend to have a negative impact on becoming a trialist
- Users are tend to become trialists when they rated higher number of movies per day
- Small number of directors of rated movies have positive impact on model output



03

Final Conclusion & Recommendations

Conclusion

1. Best performing model : XGBoost

2. Hypothesis Review and Verification



Users actively rate movies and leave reviews \Rightarrow more likely to be a subscriber.

- False. Actively rating has no impact or negative impact to be a subscriber. Leaving reviews is less likely to influence a user to be a subscriber.



Number of movies rated by users during trial period \Rightarrow have high impact on conversion

- True. Plus, being a trialist have positive impact on becoming a subscriber.



Hand-curated films plays a big role on turning free or trial users into paying users

- “movie_popularity” is the variable with the most impact. Low average popularity of movies rated have positive impact.

Recommendation

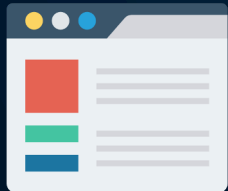
Data Collection Recommendation:

1. Try to avoid or get rid of outlier data
 - a. The sharp peak on 01/24/2019
 - b. The person that rated 20k movies⇒ Cyber security system should block the users that are acting abnormal.
1. Take more variables into consideration
 - a. Add variables: Viewed movie list, average stay time, movie category, etc.
 - b. Align with marketing tools such as Google Analytic and combine the information with existing datasets, labeling users and better understanding their customer journey.

Recommendations(Continued)

Marketing/Operation Recommendation:

1. Subscriber and Trialist don't like to go with the flow
⇒ Launch a discovery section on website and promote older or rare movies for those art movie lovers
⇒ Target Boomers
1. Some people who left comments and ratings are not a subscribed user
⇒ Seperate the movie discussion platform and movie streaming platform.

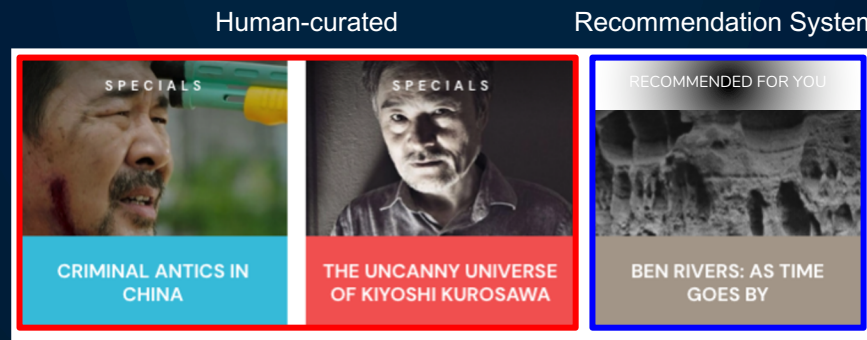


Art Movies Rating Platform

Art Movies Streaming Platform

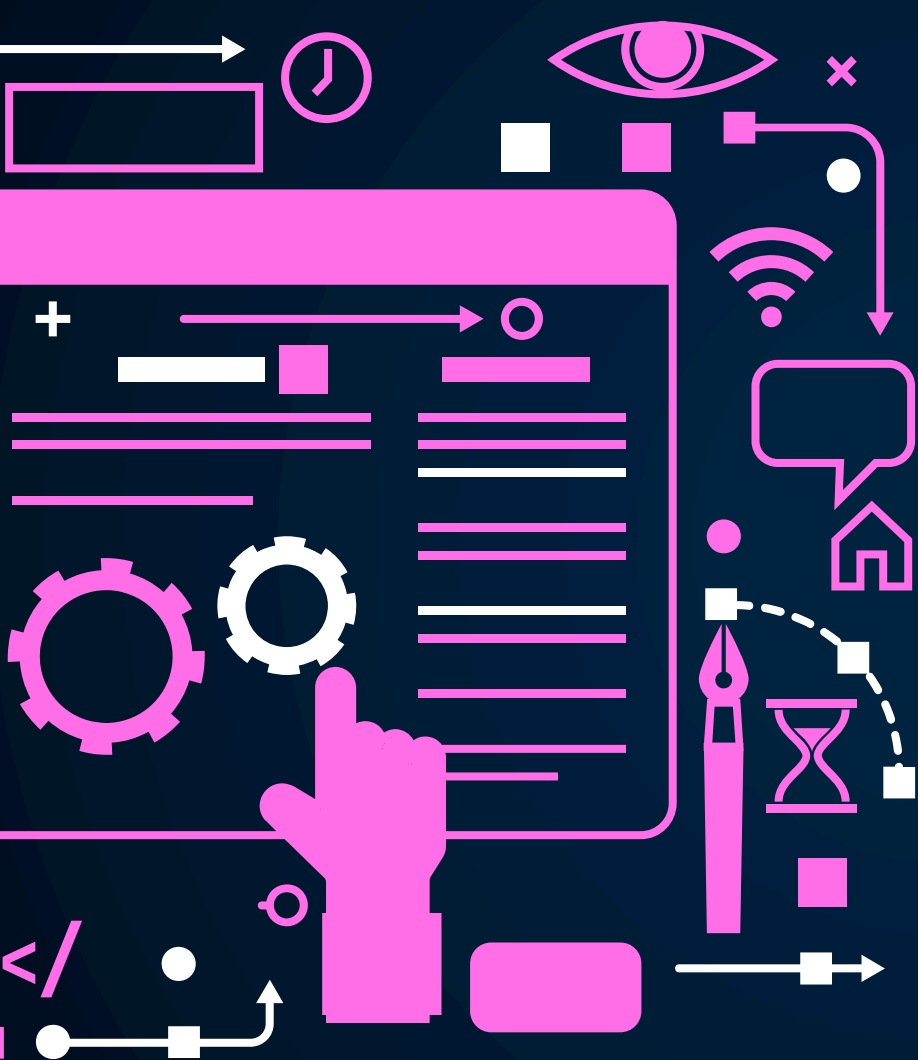
Recommendation System

- Boosting films with low popularity/no ratings for discovery
- Human curated section + recommendation system section
- For non-subscribers, recommend movies that a user is predicted to give high ratings



Combination of human-curated and system recommended movies on MUBI (for placement only)

Any Questions?



Thank You

Minkyung Kim(Monica) / Yunjie Lu(Yena) / Nai-Chieh Yang