An Offering They Can’t Refuse: Using Data to Bring Audiences Movie They’ll Love

Team 7: Isaac Gibbs, Samir Khan, and Yu Wang

THE PROBLEM

Opening Netflix, scrolling through the catalog, seeing nothing interesting, and then navigating away is an all too common experience. As a result, Netflix and other streaming services must constantly update their catalog with new and attractive content to keep users watching. We used data on user ratings and movie tags to break down the best ways for streaming companies to find out which titles will keep users watching and which ones aren’t worth their licensing fees.

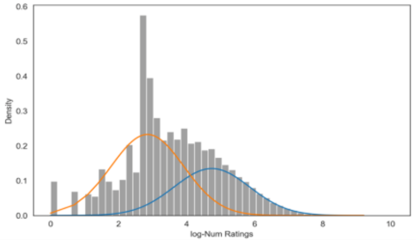
KEY TAKEAWAYS

- There are different kinds of users: movie buffs, and more casual viewers. When deciding what movies to add to a service, it is important to consider the opinions of both of these kinds of users, but for the best predictions you should survey mostly high-intensity users.
- How much engagement a movie receives on the platform can be predicted from certain tags. Especially important tags include “stand-up comedy,” “dialogue,” and “nudity.” Most of the success or failure of a movie will be driven by how much of these it features.

MODELING USER BEHAVIOR

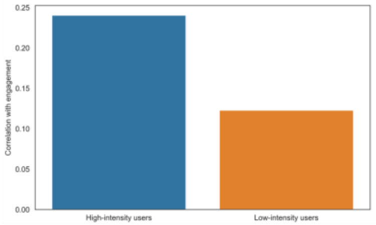
Two kinds of users

For every user, we have access to the number of movies they rated. Analyzing this distribution reveals two different kinds of users: those that rate only a small number of movies, and thus use the service less, and those that rate a large number of movies, and thus use the service more. For each user, we determine which group they are more likely to have come from, and assign them to that group. We call these “low-intensity” and “high-intensity” users.



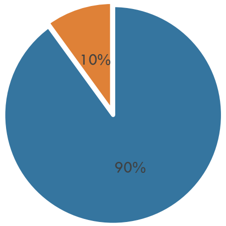
Both kinds matter…

We fit a regression model to predict the total number of times that a movie is rated (which is a measure of its popularity) based on its median rating among high-intensity users and its median rating among low-intensity users. In the resulting regression, the ratings of both kinds of users were statistically significant predictors of a movie’s engagement.



…but one kind matters more

Based on the regression model, we can form an estimate of the popularity of a new movie by asking a group of users to rate it. We solved an optimization problem to determine that, for the most accurate predictions, the survey group should consist of 10% low-intensity users and 90% high-intensity users.



MODELING MOVIE ATTRIBUTES

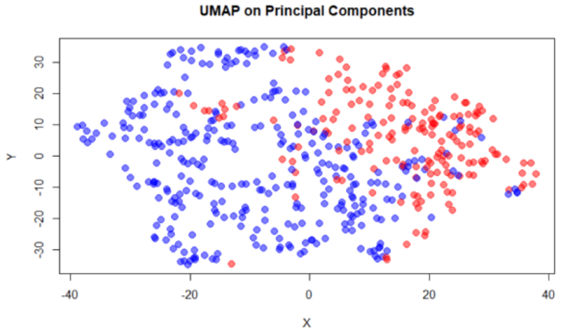
Visualizing the space of movies

Our data consist of several thousand movies and a list of tags such as “boston” and “powerful ending.” We have numerical values of how much each tag applies to each movie, and we can use these to form a graphic representation of the set of movies, shown below. Then, using statistical methods, we identify two clusters within this space.

Powerfully predicting engagement…

We can leverage the clusters found above to build a model for predicting engagement from a movies tag data. By fitting a different model for each cluster, we are able to improve performance, and make the most accurate predictions possible.

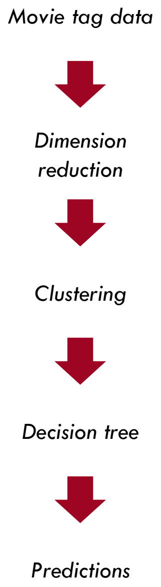
82.3% variance explained



… while finding the most important features

Our model of choice for each cluster is XGBoost, a decision tree model. As such, we can use it to identify several especially important tags which streaming services should pay attention to when trying to model engagement. The tags that are important vary between clusters. Some important tags include:

documentary ***pg-13*** ***quotable***


Movie tag data

Dimension reduction

Clustering

Decision tree

Predictions

FINAL RECOMMENDATIONS

For streaming services like Netflix or Hulu that want to determine which movies to add to the collection, we recommend creating focus groups based on a mixture of high and low intensity users. From this, we recommend surveying these focus groups on key movie attributes and for their overall impressions in order to best forecast engagement.