An Offering They Can't Refuse: Using Data to Bring Audiences Movies They'll Love

Isaac Gibbs Samir Khan Yu Wang

October 19, 2020

Contents

Executive Summary

The central problem of any business in the movie industry is to show audiences movies they will want to watch. Streaming services like Netflix and Hulu must constantly update their catalogs with attractive content, or else risk losing users who find nothing worth watching; similarly, movie studios must carefully consider which movies will bring in enough of an audience to turn a profit.

In the modern age we live in, more and more data on the relationship between movies and audiences is readily available to all. By leveraging this pervasive data we can steer industries and services towards methods that maximize both user engagement and overall profit. We do this by answering the two following questions:

- **For streaming services,** which movies should be included in their catalogs to maximize consumer engagement?
- **For movie studios,** how do we determine which projects to undertake to maximize profits?

Key Findings

We employed various statistical tests, dimensionality reduction methods, and machine learning models to find that:

(1) People who rate movies online can be classified as **high or low intensity** users. When surveying users to decide what movies to bring onto a service, the best group for predicting user interest is a group that is **90% high-intensity users** and **10% low-intensity users**.

(2) Movies can be classified into two clusters based on their genomic breakdown, which we call "serious-tone" and "lighthearted". We can use the fundamental differences in these clusters to **predict engagement better than any naive model that treats all the data identically**.

(3) Our model identifies genomic tags that are most important in predicting engagement. As a result, we recommend that streaming services focus on a few key features of a movie, such as whether it is **pg-13** and how much **dialogue** it contains, when deciding whether or not to add it to their catalog.

(4) The most **strongly statistically significant** predictors of profitability of a movie pitch are the **average** (not peak) previous success of the lead actor and director and the genre of the film. Importantly, the writer is found to be less important. Industry experts aiming to make profitable films should mostly consider a mixture of information about the genre, lead actor, and director and pay less attention to other factors.

Recommendations

For movie studios, we have devised a statistically significant model to predict profitability of movie pitches based on some basic properties of the pitch.

For streaming services that want to best determine new additions 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.

1 Technical Overview

The remainder of the article is broken down into three sections. Section 2 focuses on how streaming services can utilize feedback from different users to predict the total user engagement garnered by a film. It proceeds by

1. Using **mixture models** to identify distinct classes of users.
2. Fitting a model and then solving a **convex optimization problem** to **minimize variance** of predictions to estimate the optimal method for surveying user preferences under resource constraints.

The end result of this section is a generic procedure that allows a company to input the total number of users that it is capable of surveying, and outputs an allocation that specifies how many users of each type should be recruited. In Section 3 we expand upon this analysis by identifying what specific features of the movie itself drive user engagement. We proceed by

3. Using **principle component analysis** to collapse the space of genome tags to a more tractable set of key features.
4. Using **k-means clustering** in the reduced space to find multiple classes.
5. Using **XGBoost** to fit an **ensemble method** on each class and produce a prediction method that outperforms any naive model that ignores cluster information.

Finally, in section 4 we switch perspectives, and ask how studios can optimize profit. We

6. Use a careful exploratory analysis that combines multiple data sources to construct key features that predict box office success.
7. Combine **linear regression** with a robust **block bootstrap** to form estimates of the effects of our features on a movies profit that are robust to temporal dependencies and heteroskdastic errors.

2 The Users

We begin by examining the data on individual users. Our main goal here is to understand variations in the user base, and then determine how a company like Netflix or Hulu can use this variation to make more informed decisions about what content they should put on their service. Throughout this section and the next, we focus on how many times a movie is rated as a measure of user engagement that companies can optimize.

We believe that this captures the fact that for an online movie or television service having content that is widely discussed and widely rated is even more important than having only high-quality content. Many viral movies are not especially well-reviewed, but nonetheless attract a tremendous amount of attention to the platform on which they are available and thus bring in new users. Since the number of times a movie is rated is a proxy for this kind of success, we take it as our response.

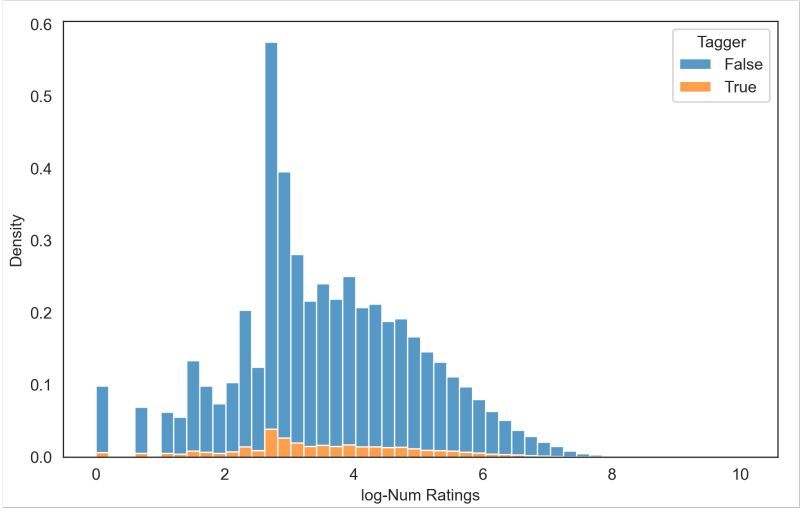

Figure 1: Histogram of log(# ratings) among users who tag movies and users who do not tag movies

2.1 Exploratory Analysis

In this section we will study the rating and tagging habits of the 283227 users in the data set. We focus on the number of movies a user has rated, partially because it is a natural measure of enthusiasm for movie watching and engagement with the platform, and partially because it shows significant variation: fifty percent of these users have rated fewer than 30 movies, while one of them has rated an impressive 23715 movies. The histogram in Figure 1 shows the exact distribution; the goal of this section is to differentiate users based on this data.

2.1.1 Users who tag movies are no different from those who do not

One obvious variation in how users interact with a movie service is whether or not they assign custom tags to movies. We might expect users who tag movies to be the same users who rate many movies, since both of these stem from higher interest in watching movies. However, the histogram in Figure 1 does not show evidence of this, and we can use the Mann-Whitney U-test to evaluate

H_0 : the distribution of number of ratings is the same for users who tag and users who do not

against

H_1 : the distribution of number of ratings is different for users who tag and users who do not.

This test gives a p-value of 0.39, so we fail to reject the null hypothesis. This means that users who tag movies are not engaging with the service more, in the sense of rating more movies, than those who do not.

2.1.2 The distribution of number of ratings can modeled as a mixture distribution

The distribution in Figure 1 strongly resembles a mixture distribution: the sharp spike around $e^3 \approx 20$ ratings is much larger than any other feature of the histogram, and suggests that there

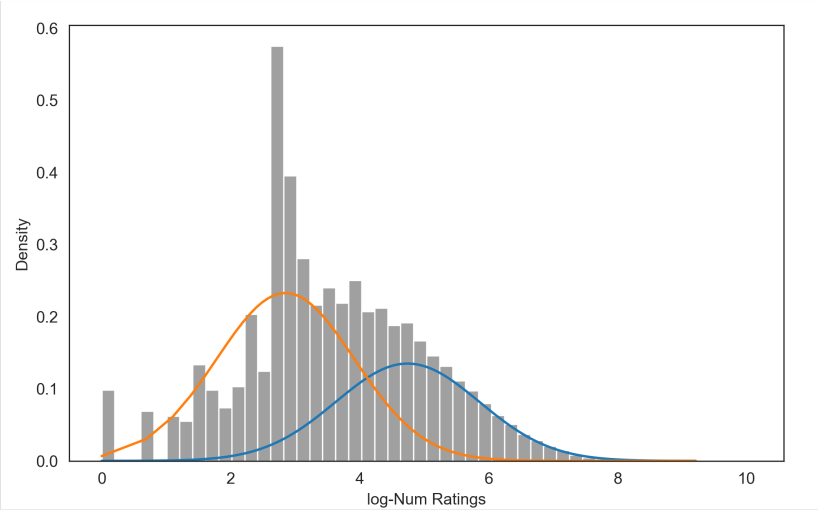

Figure 2: Fitted Gaussian mixture model for distribution of number of ratings, showing high-intensity and low-intensity users

are actually two different kinds of users in this picture. There are many low-intensity users, who contribute to this peak, and a smaller number of high-intensity users, who contribute to the heavy right tail.

To formalize this intuition, we fit a Gaussian mixture model with two components to the data. The resulting fit is shown in Figure 2, with one component in orange and one component in blue. This model is visually a very good fit to the data, and naturally leads to a method for separating the users into two groups: if a user is more likely to have come from the orange component, we label them as a low-intensity user (LIU), and if a user is more likely to have come from the blue component, we label them as a high-intensity user (HIU).

2.2 Optimal user surveying practices

In the previous section, we identified two different groups of users. Now, if a company is seeking feedback from users on whether or not a movie should be added to their service, they must find a way to balance the inputs of these two groups. Is it necessary to consider feedback from both groups? If so, how should the amount of feedback collected from each group vary? In this section, we address these questions.

2.2.1 Both kinds of users are predictive of engagement

To determine whether or not each kind of user has ratings that are predictive of engagement, we fit the regression

$$\log(\text{number of ratings}) \sim \beta_0 + \beta_1 \cdot \text{median rating among LIUs} + \beta_2 \cdot \text{median rating among HIUs}.$$

The fitted model is shown in Table 1

None of the 95% confidence intervals include 0, so we conclude that median rating among both high-intensity and low-intensity users is predictive of the engagement a movie receives. Thus, if

Variable	*Coef.*	*Std. Err.*	*95% Conf. Int.*
β_0	2.0614	0.083	(1.898, 2.225)
β_1	0.0897	0.016	(0.058, 0.122)
β_2	0.7696	0.025	(0.720, 0.819)

Table 1: Summary of fitted model predicting number of ratings from median rating among each group of users

we are surveying users to decide whether or not to add a movie to the service, it is in fact important to survey both kinds of users.

2.2.2 Optimal survey design

However, looking at the fitted coefficients, the coefficient for high-intensity users is larger; their median rating of a movie is more predictive of how much engagement a movie will get. This raises the question of how should we balance the number of users of each kind we survey to best estimate engagement.

To address this question, suppose we are going to survey n total users, of which n_0 will be low-intensity users and n_1 will be high-intensity users. Then, our estimate of engagement will be

$$2.0614 + 0.0897 \cdot \text{ sample median of } n_0 \text{ users} + 0.7696 \cdot \text{ sample median of } n_1 \text{ users.}$$

The variance of a sample median from a population of size n is proportional to $1/n$, so the variance of the resulting estimate is roughly $\frac{0.0897^2}{n_0} + \frac{0.7696^2}{n_1}$.

The optimal choices of n_0 and n_1 then, are the solutions to the optimization problem

$$\text{minimize } \frac{0.0897^2}{n_0} + \frac{0.7696^2}{n_1} \text{ subject to } n_0 + n_1 = n, \quad n_0 \geq, \quad n_1 \geq 0$$

which is convex and thus can easily be solved numerically or analytically. For example, when $n = 50$, we find that we should include 5 low-intensity users and 45 high-intensity users, reflecting the fact that although both kinds of users have predictive ratings, the high-intensity users have significantly more predictive ratings. In general, the optimal allocation is to have 89.6% of the group be high-intensity users and the remaining 10.4% be low-intensity users.

3 The Movies

In this section we investigate which properties of a movie are most indicative of success. Following the theme of the previous section, we aim to determine which movies generate high engagement across users. Our goal is to help movie directors predict user engagement via the genomic breakdown of their movie - that is, we aim to utilize the various scores and tags a movie could have to both generate a "recipe for success" and predict overall engagement based on a descriptive pitch.

3.1 Exploratory Analysis

We work primarily with data from the genome-scores.csv file in which each line contains a movie ID, a tag ID, and a [0,1]-valued score indicating how well the tag describes the movie. By pivoting this file, we are able to generate a dataframe with 13176 movies and 1128 tags.

3.1.1 High Dimensionality of Data

Due to the high dimensionality of the data and the intuition that a lot of the tags are highly correlated (e.g. ”animated”, ”animation”, and ”anime” are all separate tags), we begin our analysis by projecting the data into a lower dimensional space. We do this by first centering the data and then applying Principal Component Analysis (PCA). Figure 3 is a scree plot showing the percentage of variance explained by each of the principle components. By using the elbow method we determine that 9 principal components seems like a sensible choice. These 9 components explain 45% of the variance in the data and thus using just these components we are able to recover most of the signal from the genome tags without over fitting to the residual noise.

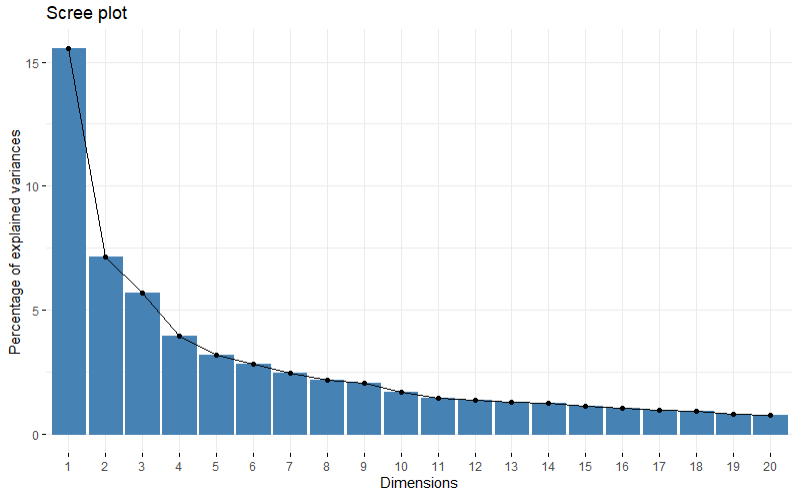

Figure 3: Percentage of variance explained per principal component

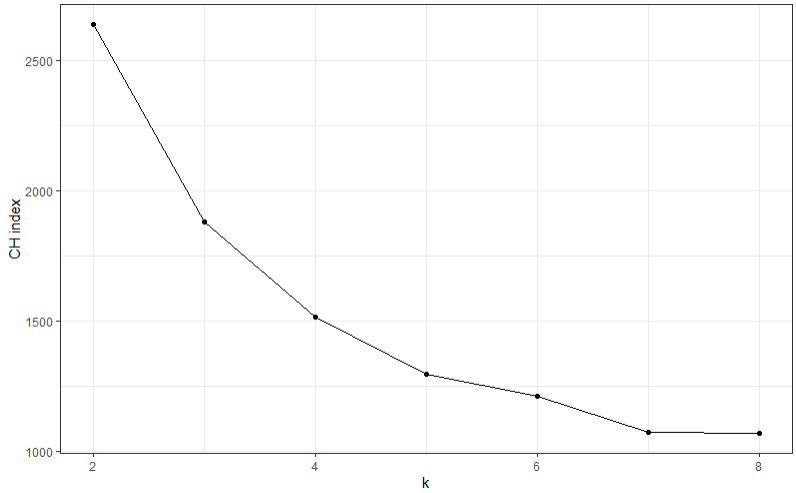

Figure 4: Calinski-Harabasz index, calculated via within sum of squares and between sum of squares for each cluster. Higher values are more desirable.

3.1.2 Clustering data aids in prediction

It is intuitively obvious that users engage with different types of movies in different ways. A lighthearted comedy might be thrown on in the background on a rainy day, while Oscar nominated dramas are usually viewed only when we can give them our full attention. Thus, when we predict engagement from the genomic data, we expect that any good model should predict differently for different types of movies. Additionally, it is also reasonable to expect that the genomic pattern underlying the different types of films should be both different and discernible. Although some genre information was provided for all the movies, we believe that due to the high granularity of the genomic data we can cluster the movies in a more sophisticated way than what can be obtained by using typical labels like "action" or "comedy" that might not fully capture the mood or tone of a film.

Motivated by these ideas we run k-means classification on our 9 principal components. To choose the appropriate k, we use the Calinski-Harabasz index, an index that balances a trade-off between minimizing the within cluster variances and maximizing separation between clusters. In general, the goal is to maximize this index. The values of this index across $k \in \{1, \ldots, 8\}$ are shown in Figure 4 and based on this information we choose 2 clusters. As a sanity check we investigate how the provided genre labels vary across the 2 clusters. From Figure 5, we identify the clusters as "serious-tone" or "lighthearted".

To visualize the clusters in two dimensions, we employ Uniform Manifold Approximation and Projection (UMAP). UMAP is expected to perform better than simply viewing the data in principal component space because the PC projection performs well only if the shape of the data is largely Gaussian and most of the variance can be explained in the first 2 principal components. For our data, we have no reason to believe the former assumption and the latter assumption was already found to be false from observing the scree plot in Figure 3.

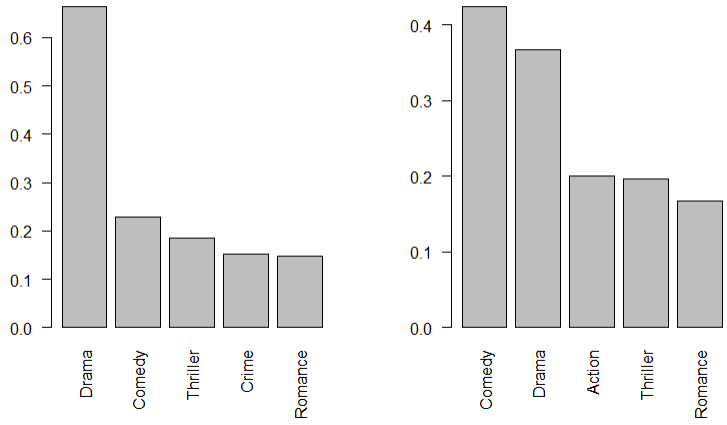

Figure 5: Top 5 genres appearing in each cluster. The clusters can be largely summarized as "serious-tone" vs "lighthearted"

UMAP helps us bypass both these assumptions by identifying a lower-dimensional manifold that the data is "close" to. Applying UMAP on a random sample of 500 data points allow us to visualize the separation of the clusters (Figure 6). As you can see there is a good separation of the clusters in the lower dimensional space.

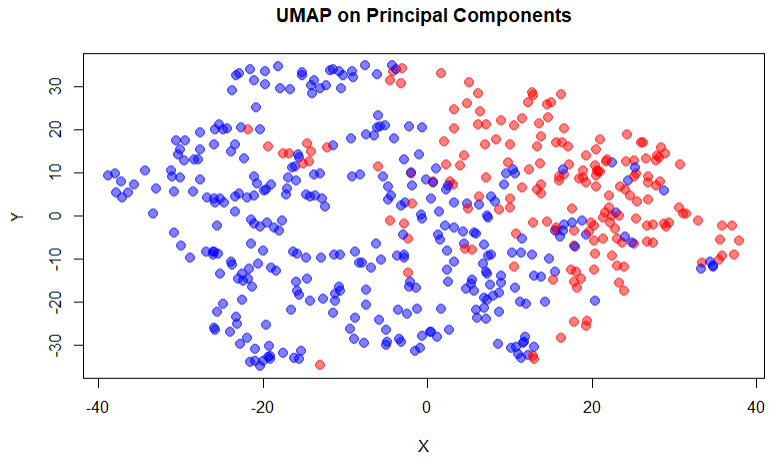

Figure 6: Projection of principal component data onto a lower-dimensional manifold via UMAP, which visualizes the distinction between red ("serious-tone") and blue ("lighthearted") films.

3.2 Predicting Engagement

As in Section 2, we predict log(engagement) of the movies, but now we use the genome scores as the predictors. Given the high number of predictors ($p = 1128$), we decided that a gradient boosting algorithm such as XGBoost will provide us with a good model.

We also propose a new model (described in detail below) that we will call XGCluster. This is an ensemble method that trains XGBoost on each cluster separately and thus incorporates our prior belief that movies with lighter or darker tones are treated differently by viewers. Because we do not have oracle information on the clusters, the training and testing algorithms are constructed as summarized below:

Algorithm 1: Training XGCluster

Input: data
clusters $\leftarrow$ train k-means(data);
for $i \leftarrow 1$ **to** k **do**
  XGCluster [i] $\leftarrow$ train `XGBoost` (clusters [i]);
Output: XGCluster, clusters

Algorithm 2: Predicting XGCluster

Input: data, XGCluster, clusters
for $i \leftarrow 1$ **to** *len*(data) **do**
  j $\leftarrow$ classify(clusters, data [i]) ;
  pred [i] $\leftarrow$ predict(XGCluster [j], data [i]) ;
Output: pred

As a baseline, we will compare these models to a simple linear regression and a ridge regression model. As for XGBoost we also fit these models in two ways by either fitting them once on the entire dataset (hereby referred to as lm and Ridge) or separately on each cluster (hereby referred to as lmCluster and RidgeCluster).

3.2.1 Methodology

We perform a 75%/25% split of our data, giving us 9882 movies on which to train our models and 3294 movies to test them on. We run a simple linear model, Ridge Regression, XGBoost, and XGCluster on our training data. For all models outside of the simple linear model we use cross validation to tune the hyperparameters. As is typical, models are fit on the training data and their perfomance is then evaluated on the test data.

3.2.2 Results

We display the out-of-sample MSEs in Table 2. As expected, the linear model performs the worst as it is unable to deal with the high dimensionality of the covariates. Ridge regression, which accounts for the dimensionality by regularizing, does not perform much better. XGBoost and XGCluster both outperform the linear models considerably and XGCluster is able to outperform the naive XGBoost by taking the clustering into account.

Model	lm	lmCluster	Ridge	RidgeCluster	XGBoost	XGCluster
OOS MSE	0.6717	0.6749	0.6492	0.6211	0.4886	0.4754

Table 2: Out-of-sample MSE for log(engagement) calculated on 3294 movies

Remarkably, the variance of log(engagement) in the test set is $\hat{\sigma}^2 = 2.687$. Thus, taking

$$\% \text{ variance explained} = 1 - \frac{\text{MSE}}{\hat{\sigma}^2}$$

we see that XGCluster using genomic scores as predictors was able to account for an astounding 82.3% of out-of-sample variance in log(engagement).

One benefit of XGBoost is that it provides feature importance. This allows us to create a "recipe for success" as mentioned earlier. For dramas, important genomic attributes include "pg-13", "awesome", "nudity", "exciting", and "documentary". For comedies, important attributes include "unfunny", "stand-up comedy", "dialogue", "quotable", "crude humor".

Intuitively, these tags make sense - audiences enjoy dramas that are intense but accessible ("pg-13") and nudity in movies is frequently discussed among viewers. For comedies, memorable scenes drive engagement. Unfunny could also drive engagement in that although the ratings may overall be bad, the proverbial statement that "any publicity is good publicity" can drive audiences to continue to watch the film.

4 The Industry

In this section we investigate a fundamental question for movie studios: what movies are worth funding? We focus on answering this question using only information that is available when the movie is initially pitched to producers. More precisely, we assume the only information available is the director, writer, proposed star, budget, rating (e.g. R, PG-13), and release year of the movie. We are interested in using only this information to predict

$$\text{percent return} = \frac{\text{gross profit}}{\text{budget}} \cdot 100.$$

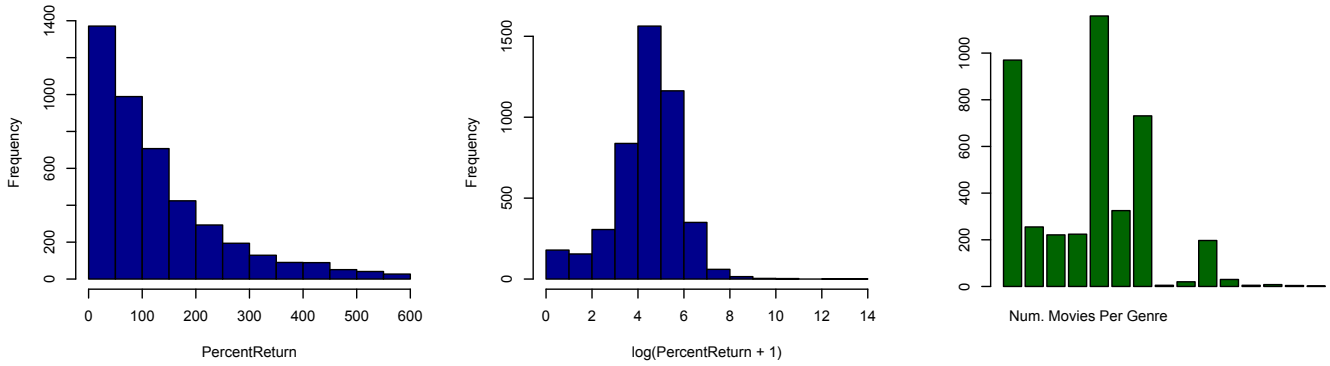

Figure 7: From left to right we have histograms of the bottom 95% of the percent returns, the log(percent returns + 1), and the number of movies in each genre.

This is of course a limited set of features and so we cannot expect to make optimal predictions. Instead, we focus on understanding which features are most important in determining a movies success, so this information can guide decision making.

4.1 Exploratory Data Analysis

4.1.1 Transforming percent return data

The first panel of Figure 7 above shows a histogram of the bottom 95% of percent returns for the movies. The heavy right tail violates the assumptions of the linear model, so we log-transform the percent revenue. The second panel of Figure 7 shows that this transformation improves normality.

Finally, the third panel shows the distribution of genres. We see that some genres have very few movies, so we merge the smaller categories. Specifically, we treat all thrillers as action movies, all westerns and sci-fi movies as adventures, and all romances as dramas.

4.1.2 Feature construction

We built features from the industry.csv and oscar.csv data files. For the industry data, there are 2504 unique star actors. This is too many to capture the influence of each actor individually, so we summarize an actors influence in three ways:

Quality score an indicator of whether or not an actor has been previously nominated for an Oscar in an acting-related category

Average profit score average log percent return of all films that the actor starred in, in the 5 years proceeding the release data of the current movie

Max profit score same as above, but maximum percent return instead of average

(For first-time actors, we replace the latter two by the average percent return of all movies from the last 5 years for which data is available.)

Similarly, there are too many directors and writers to capture individual effects, but we can define analogous quality and profit scores for them. Figure 8 visualizes the distributions of some of these features across the movies.

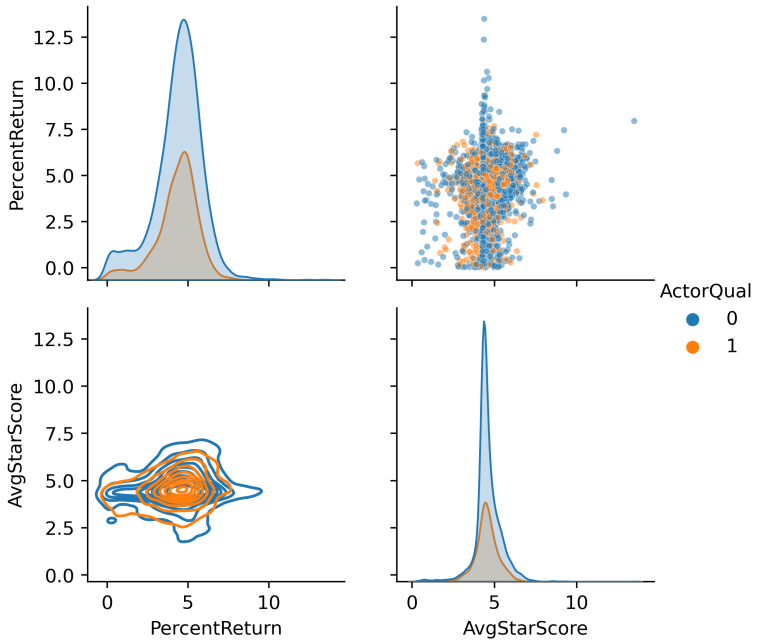

Figure 8: These plots show the relationships between percent return, average star score, and actor quality. The density estimates show that the marginal distributions of star score and percent return are similar for actors of either quality, although there are significantly more low-quality actors. The contour plot shows that the joint distribution is also similar for actors of either quality.

4.1.3 Data cleaning

In addition to the above, we perform three important data-cleaning steps: first, we correct several discrepancies between the release year and release date; second, we remove 2192 movies from the data set for which some features are missing; third, we drop movies from 1986 to 1990 because we cannot form the necessary features. After finishing these 3 data cleaning steps we are left with 4156 movies from which to build our model.

4.2 Model Fitting and Hypothesis Tests

We fit a linear model

$$\log(\text{percent return} + 1) \sim \text{ all other features.}$$

Note that because of how we constructed profit scores, our features are actually functions of the response variables at earlier time points. Formally, the data (X_t, Y_t) for time points $1 \leq t \leq T$ have the property that X_t depends on Y_s for $s < t$.

This means that the usual linear model inference will not be valid, so we instead construct bootstrap confidence intervals. We use the block bootstrap, which re-samples blocks of 26 time-adjacent data points with replacement to form new data sets $\{X_t^b, Y_t^b\}_{1 \leq t \leq T}$, and then re-computes both the features (i.e. the profit scores) and the regression estimates on these bootstrap datasets. The quantiles of the t-statistics from the resampled data give us confidence intervals, some of which are shown in Figure 9.

We see that only some genres have significant effects. The baseline genre is action so the confidence intervals in Figure 9 indicate that comedy movies behave significantly different from action

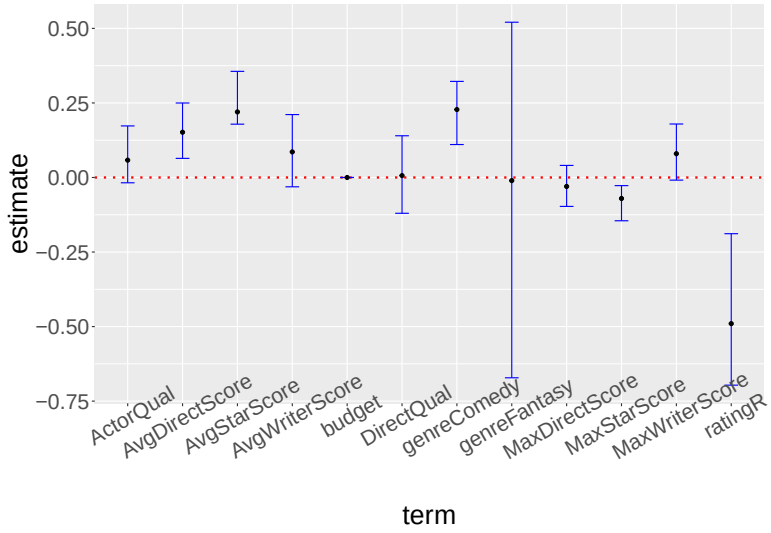

Figure 9: Bootstrap confidence intervals of selected coefficients.

movies, while fantasy movies are more similar. Additionally, we find that previous monetary success of the star actor and director is significantly correlated with future success, whereas the same cannot be said for writers. However, it should be noted that it is the average previous performance of the actor and director, not their peak performance, that is most important. We also find that previous Oscar nominations are not significant predictors of future profits. This is perhaps not surprising since Oscar nominated movies tend to have a more niche appeal that precludes them from garnering massive success at the box office. Finally, we find that R-ratings deflate profits most likely because they exclude many potential viewers.

5 Conclusion

By examining all of the available data, we have identified several key insights that can inform the decisions of streaming companies and movie studios. We found two distinct groups of users, both of whom can be surveyed to predict the amount of engagement a film will receive. Additionally, we found that it is crucial to survey a larger number of serious movie watchers. Then, we examined the space of all movies, and used a division into "serious-tone" and "lighthearted" films to improve the performance of a predictive algorithm. This algorithm identified key attributes for predicting success, such as "dialogue" and "nudity." Taken together, these give a recipe for determining whether or not to bring a title to a streaming service: construct a group of users that has an appropriate balance of high and low intensity users, ask them for ratings and impressions, focusing on key tags, and then use these to predict engagement. We also considered the perspective of studios, who must predict movie profits from the limited data available before a movie is made, and found that they should be looking mainly at the average previous success of the star actor and director, not their peak performance.

Our work enables movie industry players to make more informed decisions, but there is still room to do more. Our main measure of engagement is the number of times a movie was rated, which we believe can be even more important than how highly a movie is rated, since a large number of ratings shows a high level of user interest. However, if we could directly ask users how interested they are in a movie, this could be an even better metric. Additionally, our analysis

of industry data was largely separate from our analysis of the ratings data: if we could connect these two sets of data, and determine how directorial choices and casting decisions affect user preferences, we could form an even more complete picture of what movies users want to see. Finally, our findings are not causal. This is not easy to remedy with the data available, but it should be taken into account when interpreting our results.