Understanding Factors That Influence Content Popularity on Netflix

MIT Sloan Analytics Lab | Netflix
Team Members: Joseph Lu, Yen Hann Yoo, Zach Wayne, Rachel Duan

November 2022

1 Introduction

In today's world, there is no lack of entertainment content. Large media companies such as Netflix are consistently producing an endless stream of movies, TV series, and reality TV shows that compete for viewers' screen time. With the rise of streaming services, consumers are facing an unprecedented level of choice, which brings us to the question - what makes content popular?

We collaborate with Netflix in trying to understand the factors that influence content popularity. This analysis aims to help guide Netflix's strategy for producing and releasing new movies and TV shows and answer questions such as:

- What should Netflix look for in a new show before investing to maximize popularity?
- Does having a certain cast and crew lead to greater success?
- How do preferences differ for different regions and countries?

2 Problem statement

To understand content popularity, we use machine learning models to predict whether a certain movie or TV show becomes Top 10 on Netflix. We will explore features in the following categories:

- **Basic Title Information** such as release year, run time, genre, seasons for TV shows, etc.;
- **Popularity of Cast and Crew**, where utilizing the number of votes on their shows on IMDb, we will create a popularity ranking for directors, writers, and actors, detailed methodology below;
- **Plot Line**, which we will explore using NLP techniques on IMDb's "Storyline" of the title;
- **Content Explicitness and Controversy**, measured through the "Parents Guide" section on IMDb;
- **Topicality**, measured through IMDb ratings and number of votes.

We will gather and process data on each of the above aspects and use machine learning models such as decision trees, random forest,s and neural networks to predict which shows would become the top 10. Finally, we will analyze each of the above aspects to understand their contribution to content popularity.

3 Data

Data consisted of a list of 2232 global Netflix originals, with release dates from 2020 onward with a label recording whether or not that original ever made it to any weekly top 10 list, which itself is based on total combined view hours. To obtain basic information for each title, we supplemented this data using the IMDb Datasets. We then matched the set of Netflix titles to their respective IMDb unique identifiers (known as tconst) through web-scraping Google and IMDb search results on the title, with manual spot-checking to

ensure accuracy. With the tconst, we were able to retrieve each title's release date, cast, crew, genre, run time, etc.

Information from the IMDb's "parents guide" of each title was scraped. This section is split into 5 categories: Sex & Nudity, Violence & Gore, Profanity, Alcohol, Drugs & Smoking, Frightening & Intense Scenes. Each category is graded at 4 levels: None, Mild, Moderate, and Severe.

3.1 Feature engineering

In addition to basic cleaning such as imputing missing values (KNN), removing duplicates, and cleaning some rare categories, a significant effort went into engineering informative features to supplement our IMDB title data.

3.1.1 Creating a ranking for directors, writers, and actors

One of the difficulties in measuring cast and crew popularity is time sensitivity. As the popularity of celebrities rapidly changes over time, most rankings on the internet heavily favor actors and crew from the trendiest shows at the moment, which would not be consistent with our dataset (as the release dates span 2+ years). Therefore, we created a ranking for directors, writers, and actors using the following methodology.

- 'title.crew.tsv' within the IMDb Datasets matches each title with its directors and writers. We select directors and writers for movies and entire TV series, excluding ones for singular TV episodes only.
- 'title.principals.tsv' within the IMDb Datasets contains names of the 10 most important principal cast and crew on each title. We use this to obtain the principal actors for each title.
- We select IMDb titles from 1950 to 2020. For each title, we take the number of votes the title has received on its ratings as a proxy for the show's popularity.
- For each person, we sum up the number of votes received across all titles that they are in. 3 rankings on directors, writers, and actors are thus created on the gross number of votes received by a person across all their work.
- We then rank order the people and use the ranking instead of the raw sum of votes, which we found resulted in superior model performance

Note that we are only considering the number of votes from titles from 1950 to 2020 as our Netflix data release dates start from 2020, as we avoid the circularity of using the votes on Netflix shows to predict their own popularity.

To connect this ranking to each title, we average the ranking of the top 3 actors, top 3 writers, and top director to find an average actor, director, and writer ranking for each title. There are some titles with this feature missing, so we will attempt different imputation methods (simple imputation with mean, iterative imputation, KNN) with cross-validation.

3.1.2 Tags and Languages

Next, we incorporate the parental warning tags and original languages. We expected that warning tags such as nudity and violence might be good predictors of popularity, and therefore of a movie reaching top 10 status. We were also expecting that most people like to watch movies in their own language, so being made in a widely-spoken language might make a movie more likely to reach the top 10 as well.

This data is incorporated via a CountVectorizer and OneHotEncoding respectively. To prevent over-fitting, any categories that appear less than X% of the time, with X chosen via cross-validation, was removed.

3.1.3 Natural language processing on synopses

One major feature of interest found on IMDb are plot synopses, defined by IMDb as summaries that "explain the sequence of events that takes place within a scripted, or non-scripted narrative". These summaries range from 1 to 5 sentences and are crowd-sourced, and are approved by IMDb editors. We attempted 4 different methods to incorporate this unstructured text data:

1. Bag of Words - We try preprocessing by lemmatizing and stemming words, and append the counts into our tabular data set

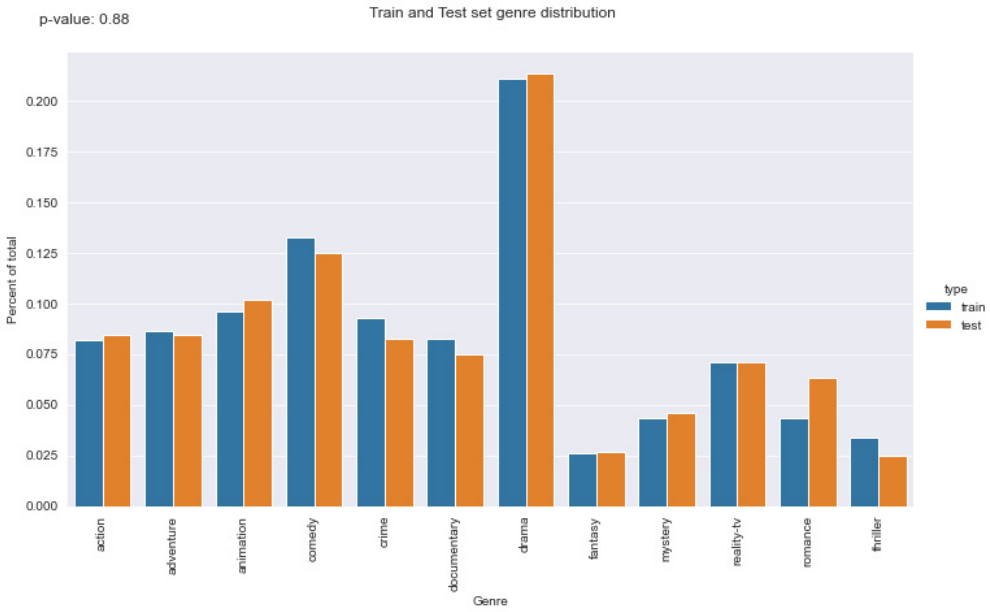

Figure 1: Distribution of genre in train v. test set

2. TF-IDF - Same as above but with weighting

3. BERT (Movies only) - We attempt to fine-tune a pre-trained BERT model to return top-10 predictions from an input string consisting of the tabular data concatenated with the summary data separated by ”.”.

4. DistilBERT (Movies only) - We attempt to utilize a pre-trained encoder which was trained to predict the prevalent emotion for a passage.

Of note is the fact that because we scraped the American IMDb website, all summaries are written in English which may bias the results. This opens the possibility to explore multi-language models in the future.

As a preliminary experiment, we tried each of these above approaches with basic title information and cast & crew rankings on an XGBoost model predicting whether a movie is top 10. The model with plot summary incorporated increased validation AUC by less than 1%. Given this small improvement, we will assess models with and without the summary feature, and evaluate whether adding the improvement is worth adding this layer of complexity.

4 Models

We will use the above features in trying different machine learning models, mainly random forest and XGBoost. We will evaluate each model on validation AUC. In our initial modeling (further explained below), we discovered IMDb number of votes to be the most important feature in predicting the top 10. However, to accommodate for the scenario where Netflix uses this model to evaluate whether a new pitch is worth investing in, as there will be no IMDb ratings or number of votes, we will have 2 final models with and without these features.

We will split the set of 2,232 titles into training and test set. There is a slight class imbalance (64% of titles in Netflix top 10), and the distribution of some independent variables is very imbalanced – for example, 37.9% of the titles are in the drama genre whereas only 3.68% are horror. Therefore, we ensured the balance of both the top 10 ratio as well as the various independent variables in splitting the model.

5 Results

Overall we find that in both the TV Show and Movie domains, the incorporation of IMDB data significantly improves our ability to predict top-10 titles on both a global and regional scale. Interestingly, we also find that removing post-facto data such as ratings and votes does not significantly reduce the predictive power. Therefore we may be able to reasonable predictions about a piece of media's likelihood to reach top-10 prior

to release. Finally, we explore the feature importances and derive some insights to help inform strategic content planning in the future.

5.1 Initial Results with Basic Data

Our initial framework to predict whether a movie will be in the top 10 or not was to use all tabular data available on the front page of an IMDB title such as release date, number of votes, ratings, etc. Data were split into an 80-20 Train-Test set, and all models were 5-fold cross-validated for hyper-parameter tuning. We used 4 basic models consisting of Logistic Regression (LR), Decision Tree (CART), Random Forest (RF), and XGBoost (XGB). The results are as follows:

	Logistic Regression (LR)	Decision Tree (CART)	Random Forest (RF)	XGBoost (XGB)
Validation AUC	0.790	0.780	0.813	0.830
Testing AUC	0.802	0.765	0.828	0.842
Testing F1-Score	0.764	0.739	0.786	0.795
Testing Accuracy	0.745	0.728	0.760	0.783

Table 1: (Movies) Comparison of initial results between Multiple Models

	Decision Tree (CART)	Random Forest (RF)
Validation AUC	0.717	0.794
Testing AUC	0.689	0.757
Testing F1-Score	0.811	0.844
Testing Accuracy	0.712	0.761

Table 2: (TV Shows) Comparison of results between CART and RF

From the results in table 1 and table 2, RF and XGBoost outperformed CART and LR across every metric in both the validation and test sets by a 3 - 8% margin. The significant increase in the validation AUC warrants increased model complexity. These findings also suggest that there are underlying complexities in the relationship between predictors (features) and the response variable, which ensemble models are better at capturing. Therefore, we proceed only with RF and XGBoost as our learning algorithms.

5.2 Incorporation of additional features

We now add in the cast and crew data, parental tags, language, and synopses. We find that in all cases XGB outperforms RF, so for brevity, we only present the best-performing XGB models within each category of synopses treatment. For movies, we also attempted to incorporate more sophisticated deep learning models to varied effects. The DistilBERT model was incorporated by appending the predicted emotion to the rest of the tabular data, while BERT model was trained directly on encodings of the tabular data and summary text.

	BoW	TFIDF	DistilBERT(Emotions)	Fine-Tune BERT	No Summary
Validation AUC	0.830	0.821	0.822	0.605	0.827
Testing AUC	0.841	0.845	0.841	0.643	0.830
Testing F1-Score	0.776	0.769	0.744	0.660	0.761
Testing Accuracy	0.755	0.755	0.712	0.625	0.745

Table 3: (Movies) Comparison of synopses treatments

We find that the BoW summary encoding performs best for movies. A deeper dive into the model parameters shows that the most important predictors related to the BoW encoding are punctuation: particularly commas. We speculate that this is because commas are used in longer, complex sentences, and perhaps this complexity is a proxy for both the effort a fan makes in writing the synopsis and the plot complexity. We find that the incorporation of emotion detection in the synopsis does not add any improvement beyond the text itself. Finally, we find that the attempt at a significantly more powerful model actually results in poor performance. We speculate that this is a prime example of a bias-variance trade-off, as attempting to tune many more parameters than data points available for training leads to over-fitting. This is not to discard deep learning completely, but it appears we require far more data than was made available to us in order to delve deeper and see better performance.

For TV shows we show the performance by choice of preprocessor:

	XGB (with Summary)				XGB
	BoW, S	BoW, L	TFIDF, S	TFIDF, L	(no Summary)
Vali AUC	0.827	0.824	0.825	0.824	0.826
Test AUC	0.879	0.883	0.894	0.868	0.876
Test F1	0.872	0.874	0.888	0.874	0.877
Test Acc	0.805	0.808	0.835	0.812	0.812

Table 4: (TV shows) Comparison of synopses treatments and preprocessing choice

5.3 Final Model

Thus far, we have considered average ratings and the number of votes associated with each title as a feature. However, average ratings and the number of votes are rarely known prior to release. Therefore, in our final models, average ratings and the number of votes are removed and the corresponding models' performance is assessed. Intuitively, we would expect the validation AUC to drop as previous importance measures indicate that average ratings and the number of votes are among the most important features in our models.

	Movies		TV	
	with Ratings and Votes	removed Ratings and Votes	with Ratings and Votes	removed Ratings and Votes
Validation AUC	0.830	0.810	0.827	0.800
Testing AUC	0.841	0.855	0.879	0.857
Testing F1	0.776	0.802	0.872	0.856
Testing Accuracy	0.755	0.783	0.805	0.778

Table 5: (TV Shows) Comparison of chosen model with and without average ratings and votes

Surprisingly we find that removing these two important features results in only a slight degradation in predictive power for both movies and TV shows.

5.4 Regional Results

One question of importance has always been how viewing preferences differ across markets. Netflix provides the top ten movies and TV shows per country going back to June of 2021, so comparing Netflix originals to these top ten lists, we get each title by whether they hit the top ten in a specific market. The markets we choose to look at were U.S., Brazil, France, South Korea, South Africa, and Australia as they represent one country from each continent and should give us a representative view of the differences between various cultures.

We use the model that performs Bag of Words on the summary and does not include the post-facto data. Our findings are as follows:

Country	TV Testing AUC	Movie Testing AUC
Australia	0.850	0.868
Brazil	0.931	0.747
France	0.888	0.741
South Africa	0.821	0.783
South Korea	0.944	0.884
United States	0.797	0.863

Table 6: Testing AUC for market-specific models

We find that in general, these country-specific models outperform the global models. Looking into the most important factors for each model, we see a common theme being having famous directors/actors/writers, suggesting that there is still an importance on a globally popular cast. This is especially true in the US, which is the only market to have the popularity of the actors be the most important feature,

For movies, countries are much more likely to watch romance movies if they're in the country's native language, whereas action movies are the most likely to be watched in another language. Also, the fame of the actors is most important for action movies. Some differences are that the US prefers shorter run times, likes more severe mature content for thrillers and dramas, and finds that the actors are less crucial in thrillers. France also likes severe mature content, for drama, romance, and thrillers, especially, and prefers

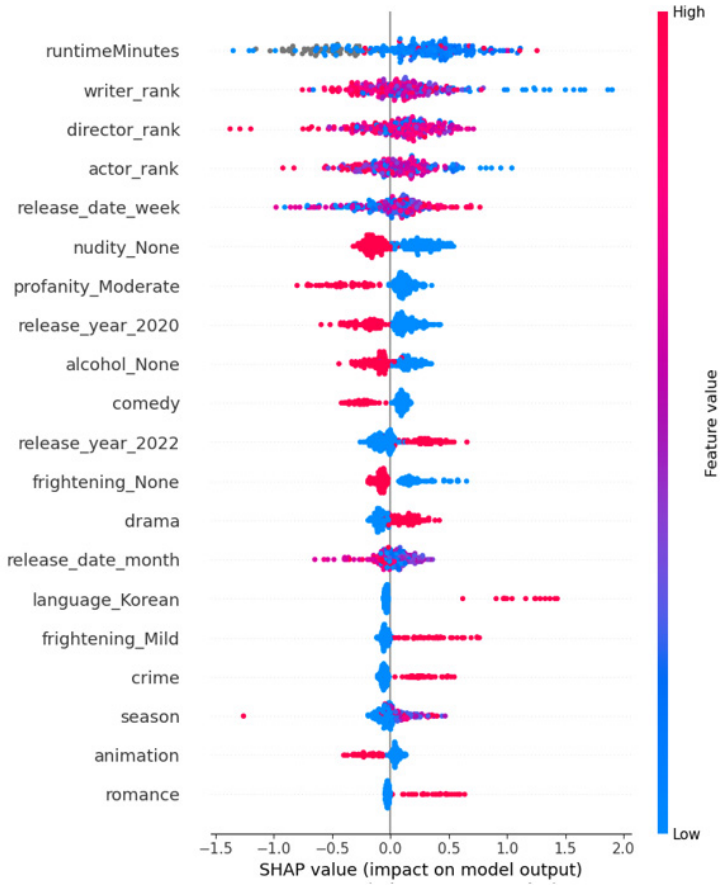

Figure 2: Shap summary plot for the final TV shows model

longer run times. On the other hand, South Korea likes less profanity and finds directors less important for thrillers.

For TV, a theme that is much more prevalent than the movie models is the importance of language. For example, 93% of Korean comedy shows are predicted to reach the top ten in South Korea, compared to only 10% of Korean comedy movies. South Korea also prefers longer run times, with the average run time of a show predicted to hit the top ten being over an hour long. Regarding Brazil, it's found that a higher percentage of Spanish shows are predicted to hit the top ten than Portuguese shows. As for some common TV themes, for animation shows, a good writer is much more consequential than a director or actor, and overall, our models predict fewer TV shows to hit the top ten than movies.

5.5 Feature Importance

5.5.1 Model's Recipe for a Top 10 TV Show

5.5.1.1 Runtime Minutes According to Figure 2, run time is the most important feature in determining whether a TV show will be top 10 or not. Thus, the predicted probabilities for the test set were computed and the run times were binned into four categories: < 30, 30-45, 45-60, and > 60 minutes to ensure an even distribution of observations in the test set within each category. Figure 2 suggests that the shorter the run time (indicated by the cluster of blue dots to the right of the zero-impact line), the more positive the impact on the model output. This is generally in agreement with Figure 7, which shows that the probability of hitting the top 10 is greatest when the run time is between 45-60 minutes; not too long (> 60 minutes) and not too short (< 30 minutes), but generally, the shorter (< 60 minutes) the greater the probability of reaching the top 10.

5.5.1.2 Rankings Unlike movies, Figure 2 suggests that the order of prioritization should be highly ranked writers, then highly ranked directors, and finally highly ranked actors. This finding was verified by scatter plots of writer, director, and actor rankings. Subsequently, each scatter plot was partitioned into four quadrants. In all 3 scatter plots, the top left portion is densely populated, i.e. highly ranked director/writer/actor are more likely to be in the top 10 TV shows. However, note the number of points

outside the top left quadrant. Directors and writers have approximately the same number of points outside the top left quadrant (40 and 50 respectively) indicating that the director and writer rankings influence the probability of reaching the top 10 more than actor rankings, unlike movies. With actor rankings, it has the most points outside the top left quadrant, suggesting that the probability of reaching the top 10 is less sensitive to actor rankings. Hence, the order of prioritization for TV shows would be to employ highly ranked directors and writers, then only consider highly ranked actors, to maximize the likelihood of a TV show hitting the top 10. Intuitively, this makes sense as TV shows, unlike movies, continue with each season and are not "one-offs." Therefore, over time, the quality of the TV show's content is more impactful than the reputation and fame of the actors and actresses themselves.

5.5.1.3 Genres Figure 8 shows that genres such as drama, crime, and romance increase a TV show's likelihood of reaching the top 10. Conversely, both Figures also agree that genres such as comedy and animation are less likely to reach the top 10. In fact, Figure 2 even indicates that these two genres will negatively impact the likelihood of reaching the top 10. Though western, musical, and fantasy are very likely to be in the top 10, as we have limited data points on these genres, they were not vital to the model.

5.5.1.4 Content Explicitness and Controversy Figure 2 suggests that some form of nudity positively impacts a TV show's likelihood of reaching the top 10. Figure 9 also supports this postulation with some nudity (mild, moderate, severe) plays a role in maximizing a TV show's likelihood of reaching the top 10. While Figure 2 suggests that moderate profanity negatively impacts the model output, Figure 9 indicates that severe profanity is better for reaching the top 10. Figure 2 demonstrates that `frightening_None` = 1 negatively impacts the model output whilst `frightening_Mild` = 1 positively impacts the model output. Coupled with the mean predicted probabilities from Figure 9, moderate to severe frightening is desirable in TV shows. Additionally, Figure 2 also implies that `alcohol_None` = 1 negatively impacts the model output, indicating that alcohol consumption (mild, moderate, or severe) in TV shows is desired. Finally, mild or severe violence can play a role in increasing the likelihood of reaching the top 10.

5.5.1.5 Release Timing Releases should be avoided at the beginning and end of summer. This is likely due to work-related circumstances as people have less time just before and after the summer holidays but have more time to expend during the summer holidays. While April and May have very high probabilities, only TV shows with Drama and Action genres should be released in these months. This is because the number of Drama and Action related TV shows released in April and May far outnumber other genres. These two genres also fare very well throughout the year in reaching the top 10 and hence the probabilities in April and May are skewed towards these two genres. Nonetheless, there are some seasonality patterns to be expected as well. For example, Romance fares very well in February (Valentine's Day) and festive seasons in December. Conversely, Horror also does very well in the last weeks of October (Halloween). Overall, from Figure 10, the model predicts the probability of reaching the top 10 across all genres is most stable and greatest for releases during the last 3 months of the year: October, November, and December. This is in agreement with the Shap summary plot that shows the model output is positively impacted by later (higher) release week values.

5.5.2 Model's Recipe for a Top 10 Movie

5.5.2.1 Runtime Minutes The most important feature is `runtimeMinutes`. A longer run time is preferred to a shorter run time. From figure 11, it is clear that movies with >100 minutes run time significantly outperform shorter movies, and each bucket longer than 100 minutes has higher predicted average probabilities of making it to the top 10.

5.5.2.2 Actor, director ranking Actor and director rankings are identified as important variables in figure 3, whereas writer ranking is not. From figure 12, we can a high concentration of titles in the top left quadrant in all three plots, as better rankings in cast & crew lead to a higher likelihood of content popularity. However, it is visually clear that for actors and directors, there are more points in the top left quadrant, as the Shap plot also suggests. This could potentially imply audience especially value big-names in actors and directors in movies, therefore there may be an area that's worth Netflix's investment.

5.5.2.3 Nudity, profanity We are seeing similar patterns as `nudity_None` and `profanity_Moderate` negatively impact the probability of the top 10. This potentially shows Netflix has predominantly adult clients, therefore content geared towards children, teens perform worse than those geared towards a more mature market.

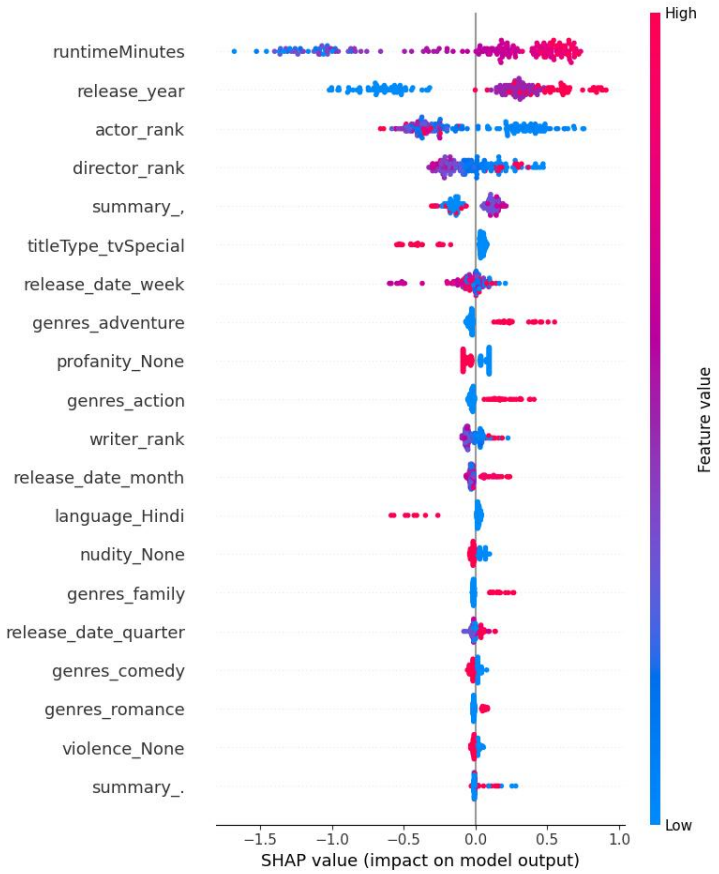

Figure 3: Shap summary plot for the final movies model

6 Appendix

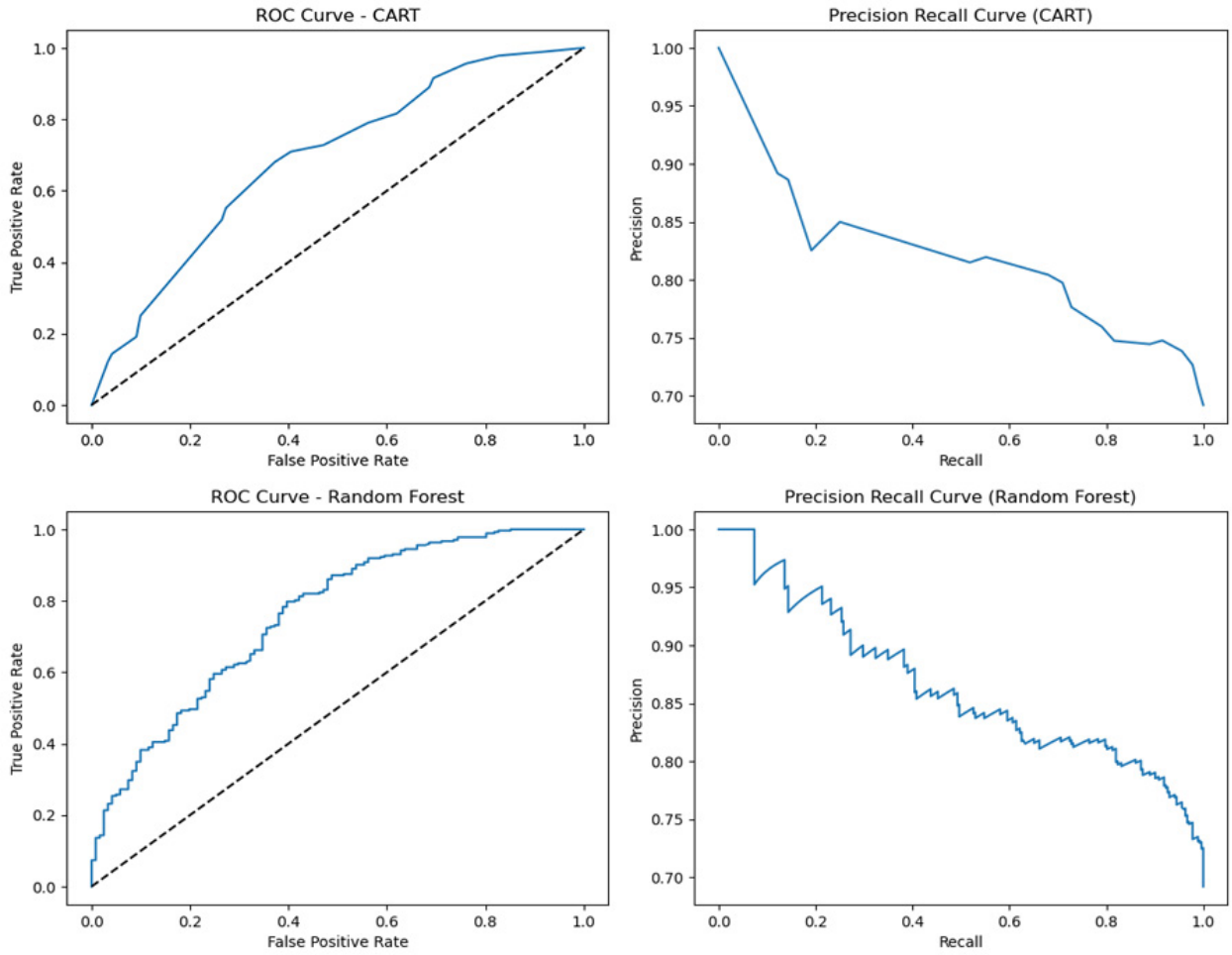

Figure 4: TV Shows - ROC-AUC and P-R curves for CART and RF models

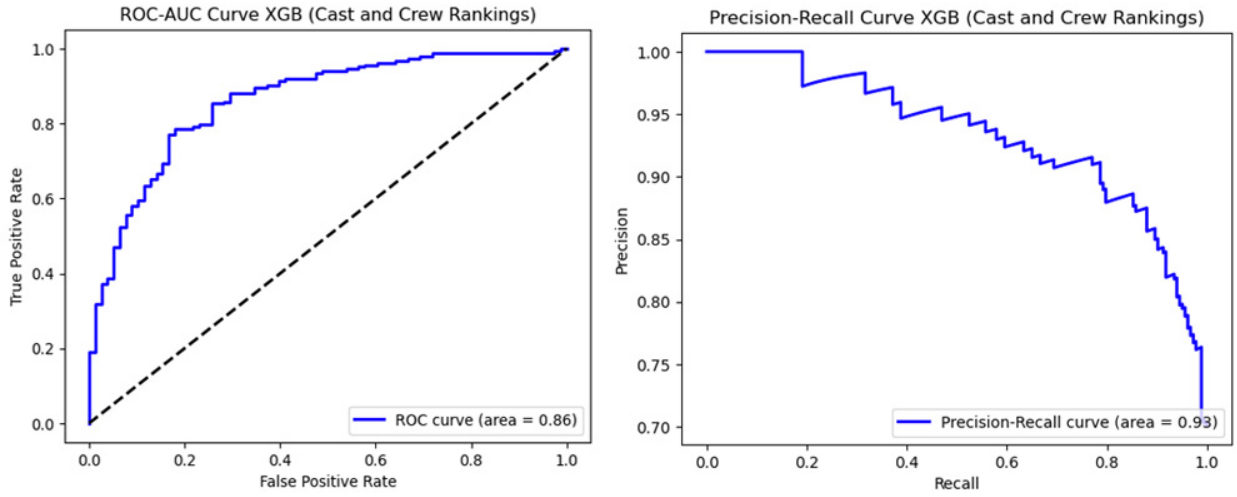

Figure 5: TV Shows - ROC-AUC and P-R curves for XGB (Rankings) model

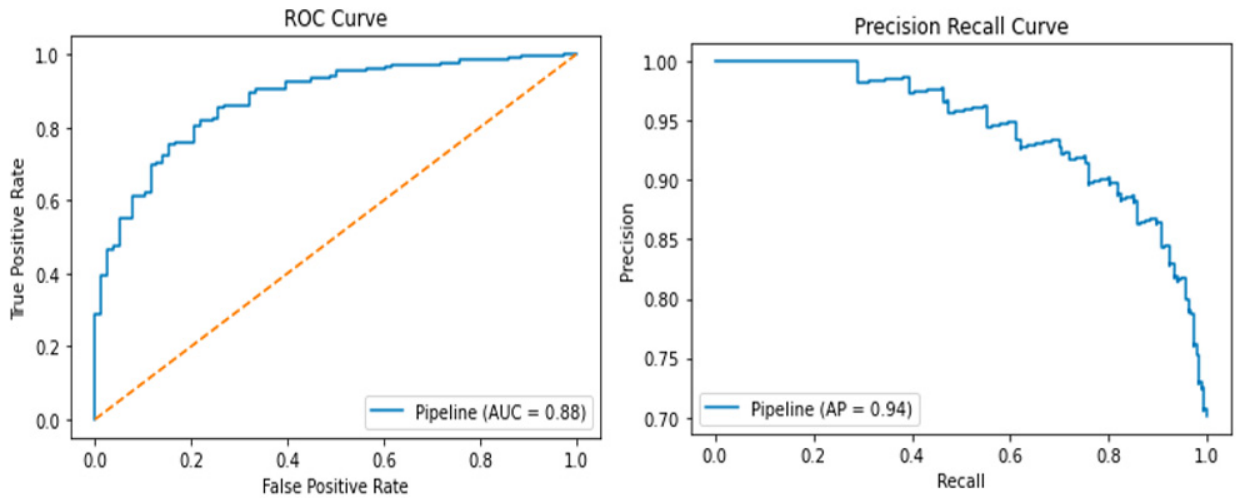

Figure 6: TV Shows - Final model ROC-AUC and P-R curves

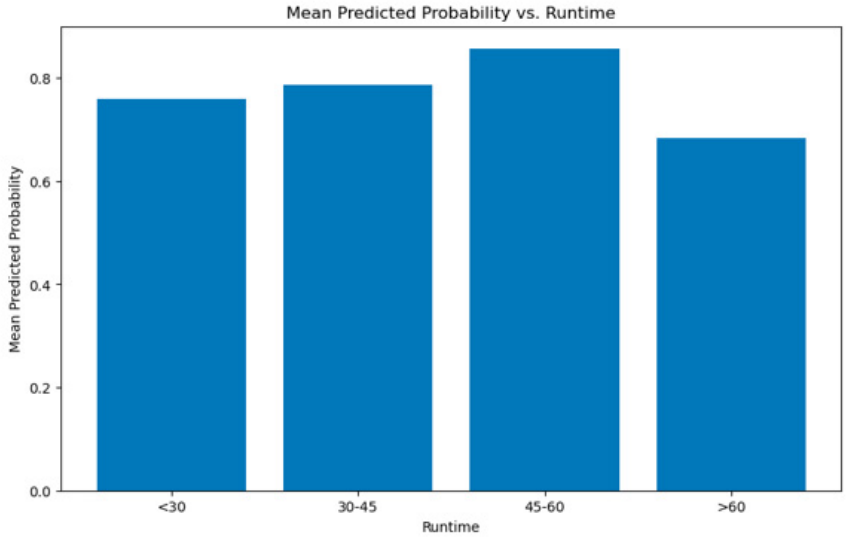

Figure 7: TV Shows - Test set mean predicted probabilities vs. runtime (minutes)

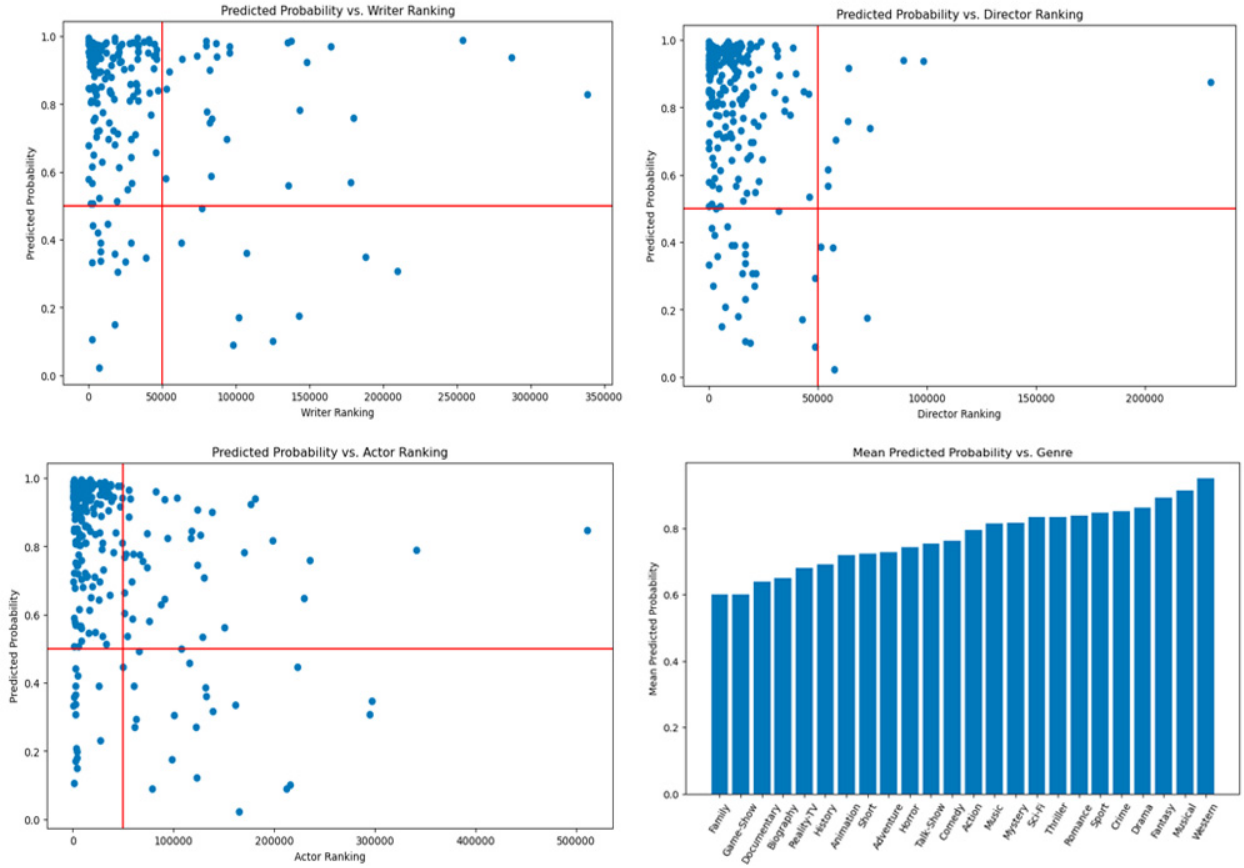

Figure 8: TV Shows - Test set mean predicted probability vs. writer / director / actor ranking and genre

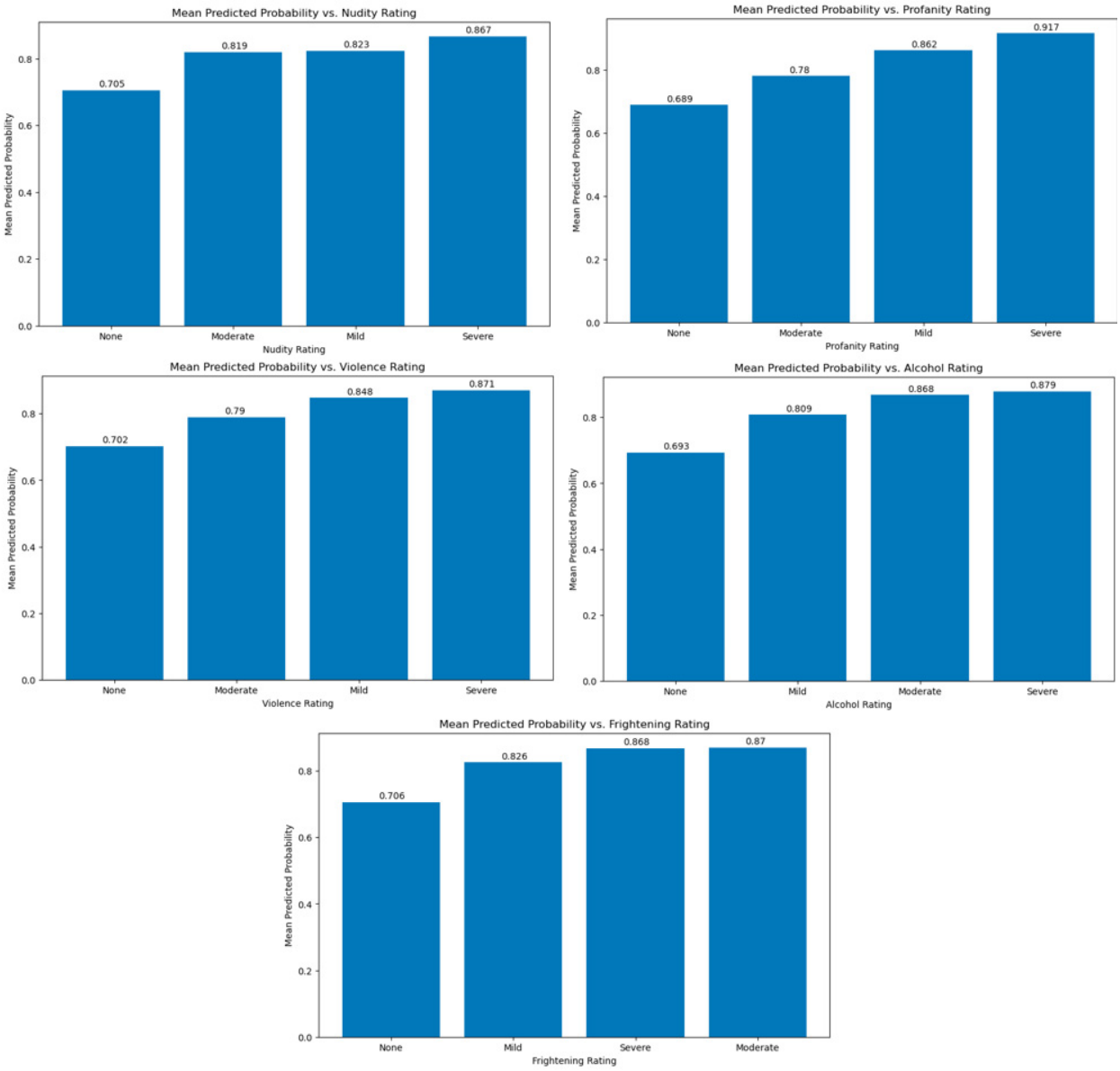

Figure 9: TV Shows - Test set mean predicted probability of hitting the top 10 vs. parents guide rating

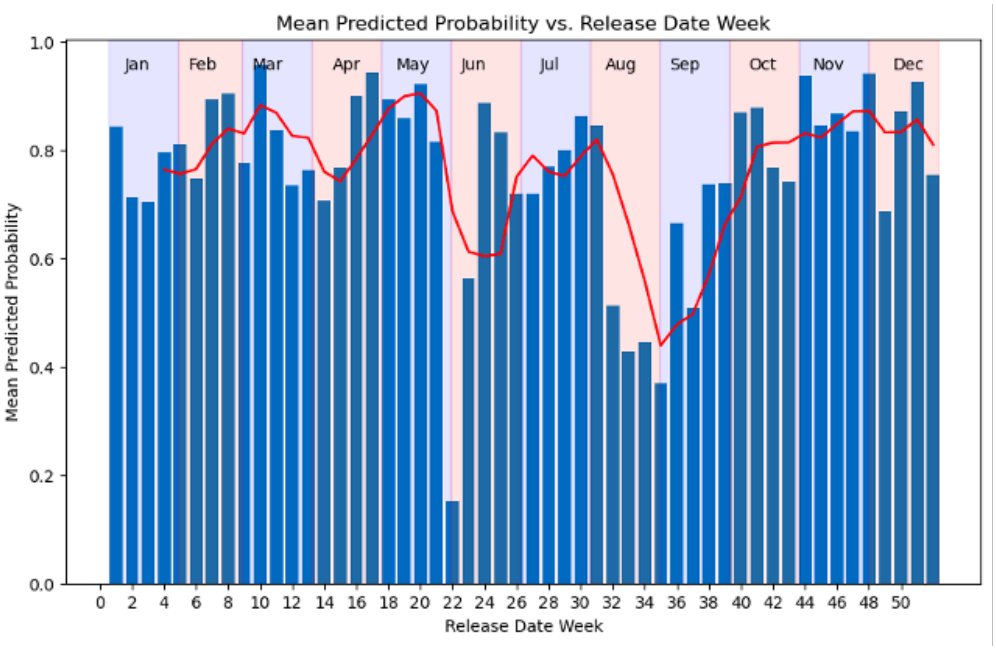

Figure 10: TV Shows - Test set mean predicted probability of hitting the top 10 vs. release week (4-week moving average

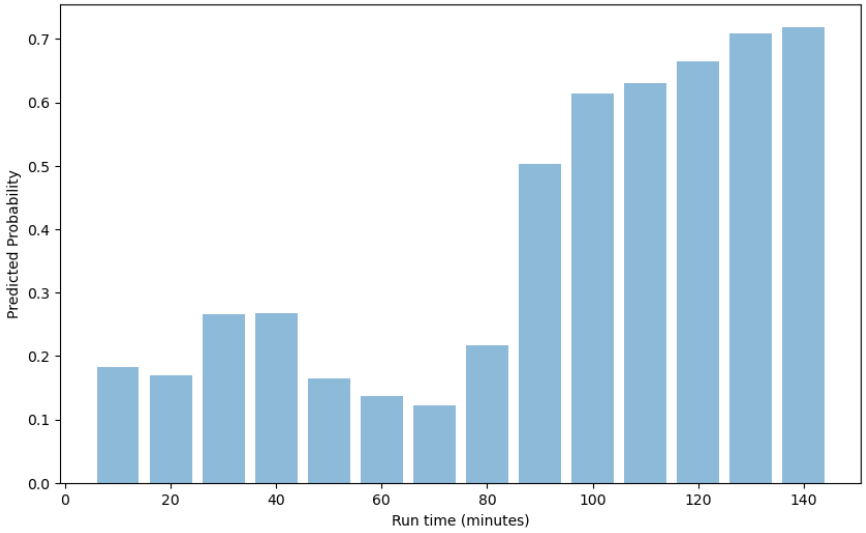

Figure 11: Movies - Test set mean predicted probability vs. run time (minutes)

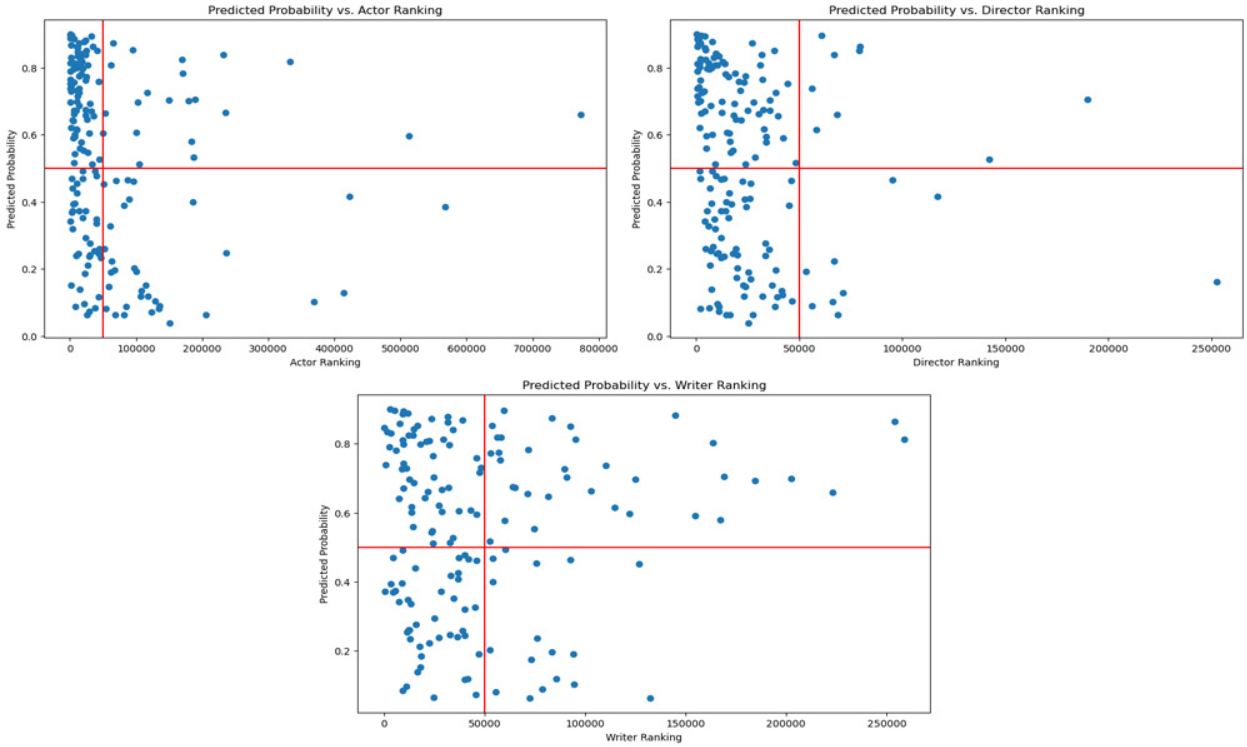

Figure 12: Movies - Test set predicted probability of hitting the top 10 vs. Actor / Director / Writer ranking