Streaming Services

CSC 587/687

James McSweeney and Noah Jaccard

University of Miami

Dr. Aguiar-Pulido

April 22nd, 2022

Table of Contents

Introduction.

Living in the information age, consumers are left with many choices for watching shows and movies at home. With cable television's natural "phase-out," streaming services have overtaken and capitalized on cable TV's demise. New services require new questions. For example, which service has the most shows? Or which services are the most affordable? These questions are simple to answer, so we took it further. Using Netflix, Amazon Prime Video, Disney+, and Hulu's movie and tv series data, we propose complex questions that require statistical analysis to answer. 78% of consumers now use video-on-demand streaming. Using these classifications, providing valuable knowledge on which streaming service a consumer should buy, or where a producer should pitch to relying on the data. Roku TV could use such a model, suggesting on the homepage where a user wants to watch a show. This would be possible, let alone suggested in the future.

Our project has two different audiences: streaming service consumers and content creators. The streaming serivce consumers receive benefit from the dataset and genre distribution. This allows them to understand what kind of content is in each kind of service. For example, child friendliness varies highly amoung services. By analyzing our dataset, a consumer can make an informed decision on which service to purchase. The main target audience is the content creators. Our classification methods reveal which genres are important to each service. By analyzing what kind of content each service prefers over another, it would make pitching more successful.

These services were not picked out at random. These are some of the best streaming services on the market. Netflix, Amazon Prime, and Disney are the top 3 largest subscriber streaming platforms in the world. Hulu is also in the top 10. These companies are the largest platforms because of their large spending on content. Collectively they spend 67 billion each year on content. While this number increases every year. Breakdowns of subscribers count, and content spending are in the below table.

Service	Number of subscribers in millions	Content Spend in Billions
Hulu	45.3	12
Disney+	129.8	25
Amazon Prime	200	13
Netflix	221.8	17

Background

There are currently other streaming service recommendations programs out there. Including a Kaggle project that web scraped IMDB reviews for some content. We will be considering genre and child friendliness, which was not included in any of their models. This project also lacks and kind of regression or predictive elements, it is a collection of data visualizations.

Sharma, Aishwarya. “Which Streaming Service Should You Choose?” *Kaggle*, Kaggle, 20 June 2020, https://www.kaggle.com/code/aishwaryasharma1992/which-streaming-service-should-you-choose.

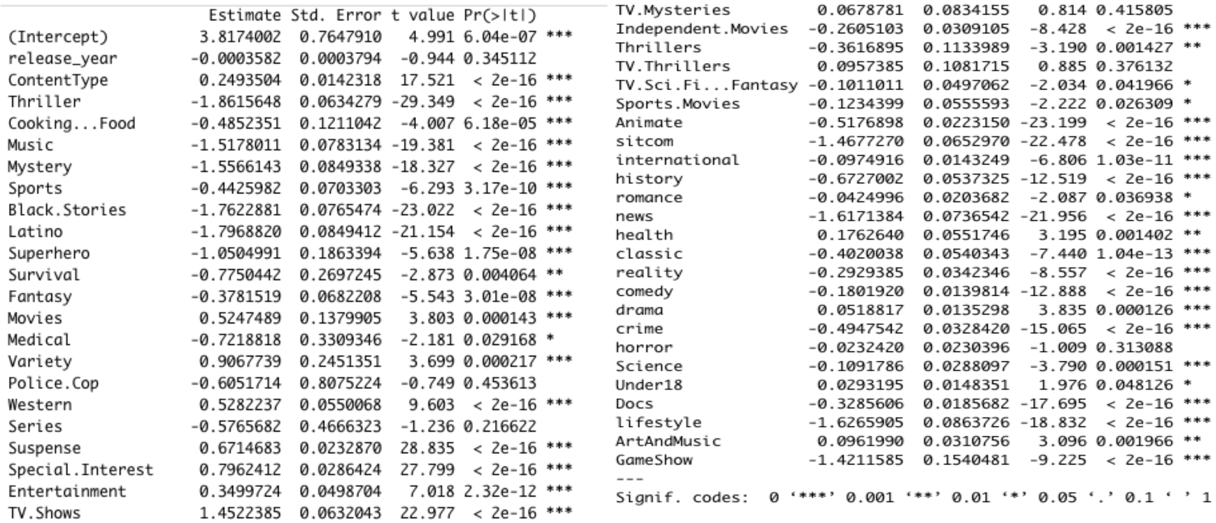
While it is not a publicly available project, it is highly likely all the services have internal models to understand what content should be acquired based on user data. These companies must bid on the right to stream content per country. These prices vary by country, which leads to countries having different content on the service. Therefore, these companies must calculate how much they should spend on titles based on how much their audience would enjoy them.

Nothing was found on Google Scholar relating to tailoring content for streaming services.

Methods

Linear Regression.

Linear Regression was performed to understand the significance of each variable. The asterisks at the far right of each variable represent the significance of the variable. With three asterisks being the highest significance.



```
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.8174002  0.7647910   4.991 6.04e-07 ***
release_year        -0.0003582  0.0003794  -0.944 0.345112    
ContentType          0.2493504  0.0142318  17.521  < 2e-16 ***
Thriller            -1.8615648  0.0634279 -29.349  < 2e-16 ***
Cooking...Food      -0.4852351  0.1211042  -4.007 6.18e-05 ***
Music               -1.5178011  0.0783134 -19.381  < 2e-16 ***
Mystery             -1.5566143  0.0849338 -18.327  < 2e-16 ***
Sports              -0.4425982  0.0703303  -6.293 3.17e-10 ***
Black.Stories       -1.7622881  0.0765474 -23.022  < 2e-16 ***
Latino              -1.7968820  0.0849412 -21.154  < 2e-16 ***
Superhero           -1.0504991  0.1863394  -5.638 1.75e-08 ***
Survival            -0.7750442  0.2697245  -2.873 0.004064 ** 
Fantasy             -0.3781519  0.0682208  -5.543 3.01e-08 ***
Movies               0.5247489  0.1379905   3.803 0.000143 ***
Medical             -0.7218818  0.3309346  -2.181 0.029168 *  
Variety              0.9067739  0.2451351   3.699 0.000217 ***
Police.Cop          -0.6051714  0.8075224  -0.749 0.453613    
Western              0.5282237  0.0550068   9.603  < 2e-16 ***
Series              -0.5765682  0.4666323  -1.236 0.216622    
Suspense             0.6714683  0.0232870  28.835  < 2e-16 ***
Special.Interest     0.7962412  0.0286424  27.799  < 2e-16 ***
Entertainment        0.3499724  0.0498704   7.018 2.32e-12 ***
TV.Shows             1.4522385  0.0632043  22.977  < 2e-16 ***
```

```
TV.Mysteries          0.0678781  0.0834155   0.814 0.415805    
Independent.Movies   -0.2605103  0.0309105  -8.428  < 2e-16 ***
Thrillers            -0.3616895  0.1133989  -3.190 0.001427 ** 
TV.Thrillers          0.0957385  0.1081715   0.885 0.376132    
TV.Sci.Fi...Fantasy  -0.1011011  0.0497062  -2.034 0.041966 *  
Sports.Movies        -0.1234399  0.0555593  -2.222 0.026309 *  
Animate              -0.5176898  0.0223150 -23.199  < 2e-16 ***
sitcom               -1.4677270  0.0652970 -22.478  < 2e-16 ***
international        -0.0974916  0.0143249  -6.806 1.03e-11 ***
history              -0.6727002  0.0537325 -12.519  < 2e-16 ***
romance              -0.0424996  0.0203682  -2.087 0.036938 *  
news                 -1.6171384  0.0736542 -21.956  < 2e-16 ***
health                0.1762640  0.0551746   3.195 0.001402 ** 
classic              -0.4020038  0.0540343  -7.440 1.04e-13 ***
reality              -0.2929385  0.0342346  -8.557  < 2e-16 ***
comedy               -0.1801920  0.0139814 -12.888  < 2e-16 ***
drama                 0.0518817  0.0135298   3.835 0.000126 ***
crime                -0.4947542  0.0328420 -15.065  < 2e-16 ***
horror               -0.0232420  0.0230396  -1.009 0.313088    
Science              -0.1091786  0.0288097  -3.790 0.000151 ***
Under18               0.0293195  0.0148351   1.976 0.048126 *  
Docs                 -0.3285606  0.0185682 -17.695  < 2e-16 ***
lifestyle            -1.6265905  0.0863726 -18.832  < 2e-16 ***
ArtAndMusic           0.0961990  0.0310756   3.096 0.001966 ** 
GameShow             -1.4211585  0.1540481  -9.225  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

All our variables with their significance codes. Most of the variables are specific types of genres

KNN classification

KNN classification is a non-parametric supervised learning method. It classifies points based on the majority voting of points based on its K closest points. The distance measurement used to find the nearest neighbors is the Euclidean distance. The value of K is a hyperparameter that can be changed to increase or decrease the number of points that vote on the classification.

	1	2	3	4
1	155	37	20	25
2	64	212	48	56
3	20	0	184	15
4	42	35	44	203

Confusion matrix for KNN. The columns are the truth values while the rows are the prediction values.

Classification Trees

Classification Tree is a predictive model that uses branches and leaves to classify new points. The variables that matter the most to classification are at the top. After a series of binary tests, the point travels down the branches. Once it reaches an end node it is classified under the respective valued

	1	2	3	4
1	250	67	24	98
2	87	321	1	27
3	66	65	244	63
4	55	70	37	276

Confusion matrix for Classification trees. The columns are the truth values while the rows are the prediction values.

Random Forest

Random Forest is an ensemble method of layering multiple trees. We used 500 trees in the model. The classification value will be voted on by most of the results of each tree. Hyperparameter tunning was also used on MTRY. MTRY is the number of variables that should be considered at each split in the trees. Grid search was the method used to tune. This process is when the model is run with different values for MTRY until the optimal value is found.

	1	2	3	4
1	302	50	60	50
2	46	365	34	45
3	33	1	296	26
4	58	20	48	317

Confusion matrix for Random Forest. The columns are the truth values while the rows are the prediction values.

Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes Theorem, with the assumption that every feature is independent of each other. Bayes Theorem is P(A|B) =(P(A)*P(B|A)) / P(B). Where P(A) is the probality of event A occuring. P(B) is the probality of event A occuring. P(B|A) is the probality of event B given event A is true. Similarly, P(A|B) is the probability of event A given event B is true.

	1	2	3	4
1	0	0	381	58
2	0	18	256	162
3	0	0	422	16
4	0	0	289	149

Confusion matrix for Naive Bayes. The columns are the truth values while the rows are the prediction values.

Support Vector Machine (SVM)

Support vector machine is a supervised learning model. More explicitly it is a non-probabilistic binary linear classifier. This means that it uses a decision boundry to classify points on each side. The goal of SVM is to have a hyperplane that can classify the points with the maximal margin. SVM will also be applied as a nonlinear classifier by using the kernel attribute.

	1	2	3	4
1	155	37	20	25
2	64	212	48	56
3	20	0	184	15
4	42	35	44	203

Confusion matrix for Support Vector Machine. The columns are the truth values while the rows are the prediction values.

Dataset

We will be using four different datasets for our project. They are all content data from the largest streaming services. The streaming services that are going to be analyzed are Netflix, Amazon Prime, Disney+, and Hulu. Our data is sourced from kaggle.com.

Bansal, Shivam. "Netflix Movies and TV Shows." *Kaggle*, 27 Sept. 2021, https://www.kaggle.com/shivamb/netflix-shows.

Bansal, Shivam. "Hulu Movies and TV Shows." *Kaggle*, 25 Oct. 2021, https://www.kaggle.com/shivamb/hulu-movies-and-tv-shows.

Bansal, Shivam. "Amazon Prime Movies and TV Shows." *Kaggle*, 12 Oct. 2021, https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows.

Bansal, Shivam. "Disney+ Movies and TV Shows." *Kaggle*, 29 Nov. 2021, https://www.kaggle.com/shivamb/disney-movies-and-tv-shows.

Description of datasets

Service	Disney	Amazon	Netflix	Hulu

Unique Values	1450	9668	8807	3073
Percentage of values	6.3%	42.03%	38.8%	13.36 %
Values after down sampling	1450	1450	1450	1450
Percentage of values after down sampling	25%	25%	25%	25%
Average Release Year	2003	2008	2014	2012
Price	15.49	7.99	15.99	12.99

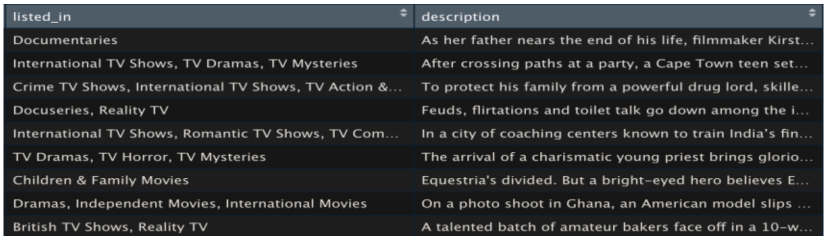
Data preprocessing.

The datasets all have the same fields: title, director, cast, country, data_added, release year, rating, and duration. All these variables are categorical besides duration. Duration's fields vary between amounts of seasons for tv shows and length in minutes for movies. Most of our questions will be solved with classification since there are categorical variables. To be able to answer more complex research questions we split up each genre into its own column. This caused an issue because different services have different names for the same genres. For example, TV comedy, Comedy shows, Comedies, and Comedy Movies, were all considered different. So, we combined similar columns into groupings which reduced the amount of genre columns from 119 to 60.

We used four related datasets. They were all sourced from Kaggle.com from the user "Shivam Bansal." The actor and director columns had been completely empty in the Hulu dataset. To standardize all the data, those columns were removed from each dataset.

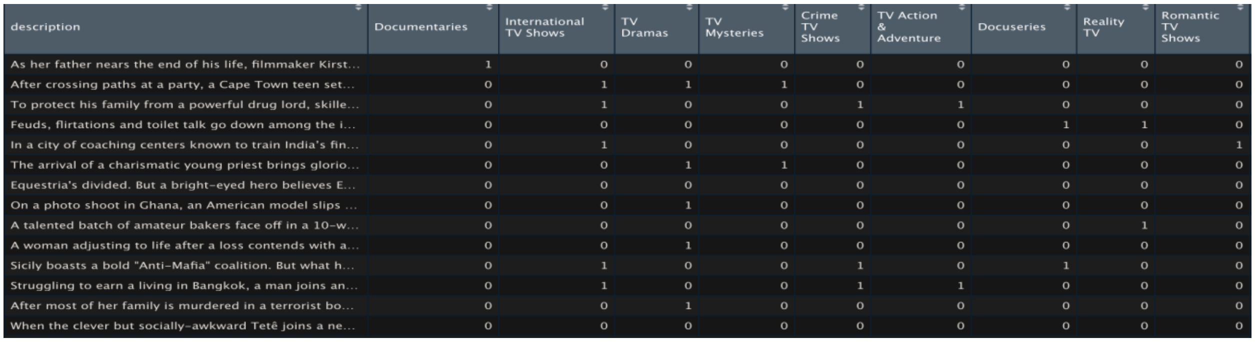
The data was generated through web scraping of the respective sites. The creator of the dataset reruns this web scraping to update the datasets every quarter to keep up with the latest content or the removal of old content.

To do that, we combined the tables of all streaming services data. From there, each row of data had a comma separated list of genres. Genres interested us to regress on, so we had to separate each genre into their own vectors.



listed_in	description
Documentaries	As her father nears the end of his life, filmmaker Kirst…
International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town teen set…
Crime TV Shows, International TV Shows, TV Action &…	To protect his family from a powerful drug lord, skille…
Docuseries, Reality TV	Feuds, flirtations and toilet talk go down among the i…
International TV Shows, Romantic TV Shows, TV Com…	In a city of coaching centers known to train India's fin…
TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brings glorio…
Children & Family Movies	Equestria's divided. But a bright-eyed hero believes E…
Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model slips …
British TV Shows, Reality TV	A talented batch of amateur bakers face off in a 10-w…

After that, we added columns of genres filled with 0s. The 0 meaning this movie or show was not of this genre, 1 meaning it was. We added said columns to the original data table and ran a program that ran through each vector of genres per row and switched each genre they were apart of to factor 1. From there we combined genres like "Anime" and "Adult Cartoon," and so on.



description	Documentaries	International TV Shows	TV Dramas	TV Mysteries	Crime TV Shows	TV Action & Adventure	Docuseries	Reality TV	Romantic TV Shows
As her father nears the end of his life, filmmaker Kirst...	1	0	0	0	0	0	0	0	0
After crossing paths at a party, a Cape Town teen set...	0	1	1	1	0	0	0	0	0
To protect his family from a powerful drug lord, skille...	0	1	0	0	1	1	0	0	0
Feuds, flirtations and toilet talk go down among the i...	0	0	0	0	0	0	1	1	0
In a city of coaching centers known to train India's fin...	0	1	0	0	0	0	0	0	1
The arrival of a charismatic young priest brings glorio...	0	0	1	1	0	0	0	0	0
Equestria's divided. But a bright-eyed hero believes E...	0	0	0	0	0	0	0	0	0
On a photo shoot in Ghana, an American model slips ...	0	0	1	0	0	0	0	0	0
A talented batch of amateur bakers face off in a 10-w...	0	0	0	0	0	0	0	1	0
A woman adjusting to life after a loss contends with a...	0	0	1	0	0	0	0	0	0
Sicily boasts a bold "Anti-Mafia" coalition. But what h...	0	1	0	0	1	0	1	0	0
Struggling to earn a living in Bangkok, a man joins an...	0	1	0	0	1	1	0	0	0
After most of her family is murdered in a terrorist bo...	0	0	1	0	0	0	0	0	0
When the clever but socially-awkward Tetê joins a ne...	0	0	0	0	0	0	0	0	0

The data had to be down sampled to 1450 to prevent class imbalance. If the data were kept in its original state, the classification methods would favor the two largest services which together are 80% of the dataset. It is especially important to down sample the data if there are drastic differences in size.

For KNN we had to implement noise on our functions. We created a function that chose about 10% of the data to give the slightest bit of noise. The values would be between positive 0.001 and 0.01 additions. This caused all the predicted values to have some slight value of positive noise on top of the values categorized. This looked like "3.011" if it were Netflix. We truncated the values into integers (now 3.011 would be 3). Then, compared those truncated integers to the test categorical factors.

Techniques applied, describing the techniques applied to the dataset. For the presentation listing them is enough.

Evaluation

K-fold Cross validation is a method for splitting training and testing sets. The dataset is split into K divisions. The model is then run K times with different test and training sets. The purpose of K fold is to ensure that every point has been included in a training and testing set. All our Cross validation was 10-fold.

Hyperparameter tunning was used to get the optimal model for three of our classification methods. The methods and values that were tuned include K in K Nearest Neighbors, Cost in SVM, MTRY in Random Forest.

Results

METHOD	ACCURACY
Classification Tree	62.21%

Random Forest	73.1%
Naive Bayes	33.64%
KNN	65.01%
SVM	68.62%

Conclusions

The dataset is updated every quarter, so it would be interesting to see what kinds of content each service decides to add or remove. This would provide some insight into what kind of content each service believes to be more successful on their platform. Since starting this project the author of the datasets has uploaded data for other streaming services such as HBO Max. Adding more services to this project would increase the number of use cases since it would enable content creators to pitch to more services.

Now that we have isolated access to data relating to the genres, we are much more comfortable isolating the data of cast members, and locations for new regressions.

This project we set out to understand the trends in each streaming service to help consumers and creators find streaming services. We succeeded in this task using classification methods. This was tested using a fake new sitcom drama action tv show. It was found that it is a good fit for the Hulu catalog. This models a real-life situation for content creators who are trying to find a service to pitch to.