STATISTICAL LEARNING: STREAMING SERVICE STATISTICS

By James McSweeney & Noah Jaccard

TABLE OF CONTENTS

01 INTRO

Introduction to the Dataset

What Providers are in the Tests

All datasets from Kaggle

Why Streaming Data?

There are many options a consumer could chose. We wanted to know which options are simply the best.

Data that we isolated among the services:

- Price
- Content Amount
- Genre
- Availability/Location
- Length of Movie/Series
- Et. cetera

Introduction

Potential Questions to Propose

- “Who should I pitch an action movie to based on all current selections?”
- “If I only can subscribe to two services, which two should I choose based on my preferences?”
- “How many action shows were released in India?”
- And really anymore relevant question would be possible to answer

Introduction

Potential Questions to Propose

- “Who should I pitch an action movie to based on all current selections?”
- “If I only can subscribe to two services, which two should I choose based on my preferences?”
- “How many action shows were released in India?”
- And really anymore relevant question would be possible to answer

State of the Art

Why make such a service?

- Pitching content to streaming services.
- Does not currently exist on the market.
- Would provide the capabilities to make an educated decision on whether or not it’s worth pitching a movie or tv show idea to a service.
    - This could be based on location, genre, amount of shows available, and more.

02 DATA PROCESSING

Data description, Preprocessing, and Variable Significance

Initial Data

- Four streaming services
- 23,000 TV Shows and Movies collectively.
- Each row has:
  - Title
  - Date Added
  - Year Released
  - Rating
  - Duration
  - Genres Listed In (up to three)

Data Pre-Preprocessing

- Needed an easy way to classify which service.
  - Added a “Service” column
  - Hulu = 1, Disney = 2, Netflix = 3, Prime = 4
- Needed to know TV Show or Movie.
  - Added a “ContentType” column
  - TV Show = 0, Movie = 1
- Mutated a “ChildFriendly” column based on rating.
  - Anything PG and lower is True, else False
- Created a column for each genre.
  - Binarily specified.

Data Preprocessing

- Binded the four datasets into one dataset
- Removed null rows
- Combined similar genres across services
- Fixed class imbalance through downsampling

Initial Genre #	Combined Genre #
119	39

Service	Before Downsampling	After Downsampling
Disney+	1450	1450
Netflix	8739	1450
Prime	9660	1450
Hulu	3072	1450

Significance of variables

	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	***

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

Data Processing

- 5 Fold Cross Validation
    - Train: 80% of selection
    - Test Holdout: 20% of selection
- Random Sampling
    - Instead of 80% of list, 20% of list
- Used for all classifications

X	title	date_added	release_year	rating	duration	listed_in	Service	ContentType	Cooking...Food	Music	Sport
1	Ricky Velez: Here's Everything	October 24, 2021	2021	TV-MA		Comedy, Stand Up	1	1	0	0	
2	Silent Night	October 23, 2021	2020		94 min	Crime, Drama, Thriller	1	1	0	0	
3	The Marksman	October 23, 2021	2021	PG-13	108 min	Action, Thriller	1	1	0	0	
4	Gaia	October 22, 2021	2021	R	97 min	Horror	1	1	0	0	
5	Settlers	October 22, 2021	2021		104 min	Science Fiction, Thriller	1	1	0	0	
6	The Halloween Candy Magic Pet	October 22, 2021	2021		1 Season	Family, Kids	1	0	0	0	
7	The Evil Next Door	October 21, 2021	2020		88 min	Horror, Thriller	1	1	0	0	
8	The Next Thing You Eat	October 21, 2021	2021		1 Season	Cooking & Food, Documentaries, Lifestyle & Culture	1	0	1	0	

03 Classifications

Classification trees, Random Forest, KNN, Naive Bayes, and SVM

Classification

- Classification groups data by similar characteristics
- Classification methods we used include Classification trees, Random Forest, K-Nearest Neighbors, Naive Bayes, and Support Vector Machines
- Performing multiclass classification to predict what service a piece of content is on based on its features
- Four different classes each representing a different service
  - Aforementioned Hulu = 1, Disney = 2, Netflix = 3, Prime = 4

Classification Tree

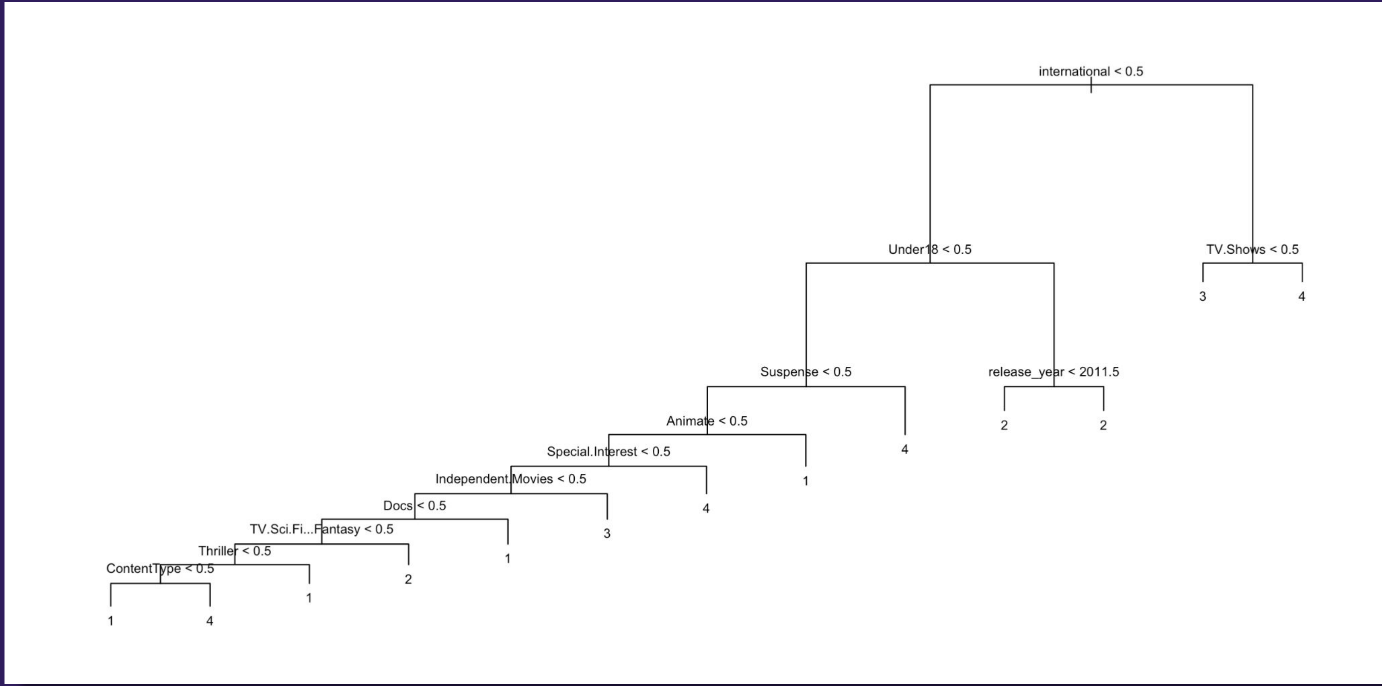
- Creates a tree through binary decisions
- By splitting it creates two branches
- The Leaf nodes are the represent the assigned classification
- Accuracy: 62.31%

Truth

Prediction

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

Tree Visualization



Random Forest

- Creates multiple classification trees
- Majority votes on the results using different classification trees
- Hyper parameter tuned with grid search
- Accuracy: 73.1%

Truth

Prediction

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

K Nearest Neighbors

- Useful for multiclass classification
  - We have four classes!
- Finds nearest relational neighbors and groups up the totals.
- Warning: If data is too similarly tied, it will not work.
  - Introduce slight noise to values
- k = 5
- Accuracy: 73.30

Truth

Prediction

	1	2	3	4
1	82	12	18	21
2	4	54	1	23
3	95	17	718	137
4	118	47	157	796

Naive Bayes

- Assumes independence between variables
- Assigns a probability to each feature
- Sums up probabilities to make a classification
- Accuracy: 33.64%

Truth

Prediction

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

SVM (Support Vector Machine)

- Advantages: High Dimensionality, Memory Efficient
- Disadvantages: Very Sensitive
- Accuracy: 74.15%

Truth

Prediction

	1	2	3	4
1	239	26	17	30
2	15	139	6	37
3	152	49	1498	307
4	194	81	0275	1535

Results

METHOD	ACCURACY
Classification Tree	62.21%
Random Forest	73.1%
Naive Bayes	33.64%
KNN	73.30%
SVM	74.15%

04 Conclusion

Conclusion and Future Work

Conclusion

- Imagine you are a writer/producer.
- You want to release a TV Show for adults listed under the genres Action, Sitcom, and Drama
- You format a table and predict it with SVM
- The model returns that Hulu would give you the best chance. You sell your show, you are now rich

Conclusion

- Highly possible to simulate and classify streaming service data.
- Random Forest and KNN are the best for reduced processing.
- SVM has the highest accuracy, but costly.
- Surprises:
    - Tied data for KNN: Noise

Conclusion

- Significant genres are the majority of the selections.
- A successful pitch would be a significant genre.
- Netflix is by far the biggest.

Future Work

- The Database is updated every 4 months
    - Could analysis new and deleted movie and tv show information
- Perform analysis on locations and actors
- Understanding the differences of each service in other countries
- Could try not condensing the genre’s and seeing what connections we could make from less generalized data.

Thank You!

Any Questions?