Group 3 Project Summary - BUAN 6356.003 - Business Analytics with R - S21

Alixander Allen Ian Clarkson Kimberly Cury Duane Harber Samundra Kharel Rob Lindsay

Business Scenario: We operate a consulting firm that facilitates through two vertical channels:

Distributor Channel: Provide predictive analysis of most likely streaming platform for both major and independent production houses to enable them to identify the most likely streaming service on which their movie will be distributed, effectively allowing them to focus their efforts on signing a deal with the platform they would land on.

End User Channel: Provide a service to budget driven consumers who want to make sure that the streaming service they select offers the most relevant content to their preferences.

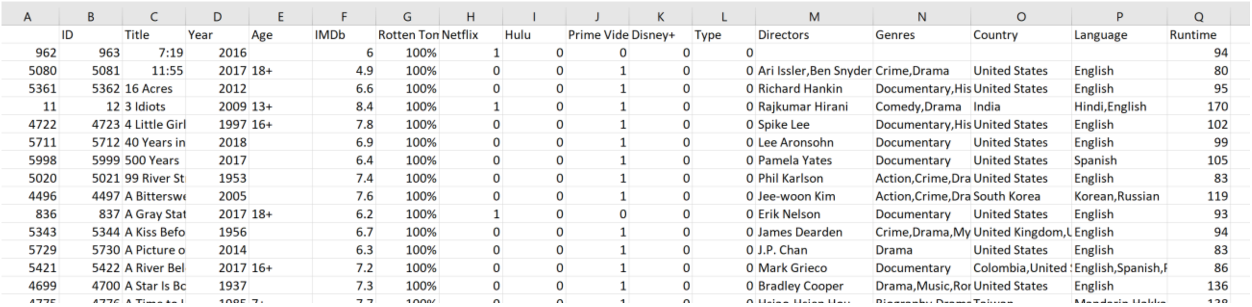
Data Characterization:

Source: Kaggle

Summary: Contains the nearly 17,000 movies offered on Amazon Prime, Disney+, Hulu, and Netflix

Descriptive Elements: Title, Year, Age, IMDB rating, Rotten Tomatoes Rating, Service Platform, Directors, Genres, Country, Language, Runtime

Sample of Data



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	ID	Title	Year	Age	IMDb	Rotten Ton	Netflix	Hulu	Prime Vide	Disney+	Type	Directors	Genres	Country	Language	Runtime
962	963	7:19	2016		6	100%	1	0	0	0	0					94
5080	5081	11:55	2017	18+	4.9	100%	0	0	1	0	0	Ari Issler,Ben Snyder	Crime,Drama	United States	English	80
5361	5362	16 Acres	2012		6.6	100%	0	0	1	0	0	Richard Hankin	Documentary,His	United States	English	95
11	12	3 Idiots	2009	13+	8.4	100%	1	0	1	0	0	Rajkumar Hirani	Comedy,Drama	India	Hindi,English	170
4722	4723	4 Little Girl	1997	16+	7.8	100%	0	0	1	0	0	Spike Lee	Documentary,His	United States	English	102
5711	5712	40 Years in	2018		6.9	100%	0	0	1	0	0	Lee Aronsohn	Documentary	United States	English	99
5998	5999	500 Years	2017		6.4	100%	0	0	1	0	0	Pamela Yates	Documentary	United States	Spanish	105
5020	5021	99 River St	1953		7.4	100%	0	0	1	0	0	Phil Karlson	Action,Crime,Dra	United States	English	83
4496	4497	A Bitterswe	2005		7.6	100%	0	0	1	0	0	Jee-woon Kim	Action,Crime,Dra	South Korea	Korean,Russian	119
836	837	A Gray Stat	2017	18+	6.2	100%	1	0	0	0	0	Erik Nelson	Documentary	United States	English	93
5343	5344	A Kiss Befo	1956		6.7	100%	0	0	1	0	0	James Dearden	Crime,Drama,My	United Kingdom,U	English	94
5729	5730	A Picture o	2014		6.3	100%	0	0	1	0	0	J.P. Chan	Drama	United States	English	83
5421	5422	A River Bel	2017	16+	7.2	100%	0	0	1	0	0	Mark Grieco	Documentary	Colombia,United S	English,Spanish,F	86
4699	4700	A Star Is Bo	1937		7.3	100%	0	0	1	0	0	Bradley Cooper	Drama,Music,Ror	United States	English	136
4775	4776	A Time to L	1985	7+	7.7	100%	0	0	1	0	0	Hsiao-Hsien Hou	Biography,Drama	Taiwan	Mandarin,Hakka	138

Challenges In Data: Our first realization was that we would not be able to use any of the platform column information as this would give an unfair insight into the prediction. If a production company was coming to us, we would have no way of knowing the data for any of the platforms so we could not incorporate it into our predictive analysis.

There was a considerable number of blank entries within our data. For example, 9,390 Movies contained no Age reference while there were 11586 blank Rotten Tomatoes entries as well. We had to take these blanks into consideration during the different stages of our predictive analysis.

There were also 4 columns that contained comma separated lists for several of the entries. These attributes fell under “Directors”, “Genres”, “Country”, and “Language” characterizations. We did not see the value in separating out the Directors or Countries, but we did count the number that were contained in each list. For Genres we decided to separate them out and apply the information to our end user channel recommendation. For Languages, we initially identified the top 11 languages within the United States and classified all other languages as “Other Languages”. During the project, we found that this was a fluid grouping that could change depending on the task at hand.

Finally, the data is heavily unbalanced. There are considerably more Amazon Prime movies as they make up almost 74% of the data. There are fewer that 700 movies that can be found on multiple platforms which we ended up taking into consideration during our final recommendation.

Amazon Prime	Disney+	Hulu	Netflix
12354	564	903	3560
73.8%	3.4%	5.4%	21.3%

Distribution Service Recommendation:

Initially, we began crafting our code going after one platform, Netflix. We felt like it was prevalent enough to enable for us to balance the data and create an effective predictive tool. Wanting to see how much balancing the data effected our recommendation, we created both unbalanced and balanced predictions using Decision Tree, Post Pruning, Bagging, Boosting, Logit, and Naïve Bayes. We then went a step further and created ensemble packages. We first tried a simple ensemble package that just took the average recommendation of the initial predictors (separated by balanced and unbalanced). We then applied the same prediction techniques we used previously to create multiple ensemble techniques. Having several predictions formed for Netflix, we stored each confusion matrix into a data frame that allowed us to compare each of them through a numerical matrix. Our primary gages were Accuracy, Specificity, Sensitivity, and Balanced Accuracy. With several predictive tools performing in the 70% range across our 4 metrics, we mistakenly felt good about our predictive tools.

As we completed the code for our Netflix prediction, we began to realize that rewriting all of the code for each service platform would be extremely time consuming. So, we decided to try to standardize the variables to allow for easy Copy -> Paste -> Find -> Replace by using PRM, NFX, DSY and HUL for the platform and intuitive descriptions for other parts of the variables. Once we felt that the code was ready to be duplicated, we set our sights on Amazon Prime.

Our confidence grew as Amazon Prime provided similar positive feedback through our metrics. Moving on to Disney+ and Hulu seemed to indicate the same story. But as we looked at the actual confusion matrix as opposed to the conglomerated numerical matrix, we realized that our chosen metrics were not providing a particularly good indication of what was really taking place as our unbalanced data was hiding the true story. We realized that we needed to be focusing on Pos Predicted Value and Negative Predicted Value. Once we added these values to our numerical matrix, we began to realize that we only had one really good “Yes” predictor in Amazon Prime with an amazing 95% Positive Predicted Value for several of the prediction tools and 3 really good “No” predictors in Netflix, Hulu, and Disney+. Despite playing with Minsplits, CP Values, and different variables, we could not derive results that would make a big impact.

At this point, we began to come to terms that our predictive analytics wouldn’t serve as a good tool in a real business scenario. Only providing a yes for one platform, even if it represented a majority of the movies offered through the four platforms, was not something a business would likely be able to hang their hat on. Prepared to share our disappointment, we had an epiphany. We had correctly eliminated the knowledge of the 4 platforms during our tool creation. But what if we were to introduce our predictions into the equation? Could our phenomenal predictive performer for Amazon Prime serve as the base to drive our predictions for the other platforms?

Yes, at least to a point where we are more confident in making a recommendation to the client. So how did we go about it? We decided to take our strongest performer from the Amazon Prime tools and make a prediction on all movies within our data set. Since this was created with balanced data, we felt confident that it was not too affected by the original test data. We then decided on a rule that was driven from our observation of the overall data. Very few movies were on multiple platforms. This is knowledge that was not built into our predictive tools. So, we introduced a waterfall scenario where anything that was not already predicted onto a previous platform was eligible for prediction on the current platform test. We decided to use the logit function for each of the Netflix, Hulu, and Disney+ platforms because the amount of code needed to be written would be minimal and would tell us if this was a viable approach.

Testing all three platforms we applied the knowledge of our Amazon prediction and only allowed a positive prediction if no positive Amazon prediction had been made. We then compared the three platforms and saw that Netflix had nearly doubled its Positive Predicted Value and was near 50%. This was the best of the three platforms. We then applied the same knowledge to Disney+ and Hulu with knowledge of both Amazon and Netflix trumping Disney+ and Hulu predictions. Disney was now in the 44% range while Hulu still struggled. Even after we applied knowledge of all three predictions and despite tripling its initial Positive Predicted Value, Hulu remained near the 10%-12% range.

So, we felt much better about our predictive tooling. These new tools provide lower bounds for “Yes” values and further analysis would be done in a real-world environment to see how we could maximize our Positive Predicted Value to provide even more confidence. As for Hulu, we are satisfied with the ability to confidently tell someone “No” as our Specificity currently sits above 96% but we would not be able to make a “Yes”.

Consumer Channel Recommendations:

Our second business vertical is a service for consumers of movie content. Movie watchers on a budget are often unsure of which streaming platform (Netflix, Prime Video, Hulu, or Disney+) will have movies that are relevant to their preferences. Our inexpensive recommendation service uses the same movie data from our first business vertical to help consumers determine which streaming platform they should subscribe to.

First, we compared the platforms by movies offered in English and Non-English Languages. All platforms had more movies offered in English than Non-English Languages. Disney+ offered the fewest number of movies in Non-English Languages. If viewing movie content in a language other than English is a priority for a customer, we would recommend Prime Video which has over 2,500 movies in Non-English Languages, followed by Netflix with over 1,500 movies offered in languages other than English.

Next, we compared movie options on each of the platforms by IMDb rating scores. When visualizing IMDb scores for each platform compared against IMDb score distribution for the entire data set, we observed that scores were similarly distributed regardless of platform. Therefore, IMDb rating scores are not a useful variable to help customers decide on a streaming service.

Age groupings in movies were an interesting endeavor to tackle. When examining the overall age groupings of the data, we found that most of the series and movies offered were in fact not rated into any age group. The ratings showed that IMDb and Rotten Tomatoes both showed a similarity in their grading across content, and that when compared to age groups you will find no dominant categories in certain ratings brackets. When you group the content offered by each streaming service into age brackets, you find that Prime and Disney contain more family-oriented content, while Hulu and Netflix cater more to the adolescent to adult demographics. As far as movies go by age category, Netflix and Prime remain the top movie providers, but we now know which age grouping the content can be applied too. IMDb showed a more comprehensive number of ratings across age groupings and tended to be tighter in their ratings. Rotten Tomatoes lacked in the series age rating, but typically rated adult series content regularly.
Rotten Tomatoes was more condensed across the 16+ age range, while IMDb had a more well-rounded rating.

Finally, we compared the streaming platforms by genres offered (i.e., Action, Adventure, Thriller, etc.). We counted the number of movies found in each genre in each platform. Not surprisingly, Prime Video, which has the most movies in the data set, had the highest number of movies in every genre category, followed by Netflix in all but one category. (Disney+ had the second greatest number of Family movies available.) We also assessed the streaming platforms by the proportion of movies in each genre they offer. Netflix, Prime, and Hulu had very similar proportions of movies by genre with only slight variations. For example, Netflix has a slightly higher proportion of comedy while Prime has somewhat higher proportion of drama movies compared to the other platforms. Disney+ demonstrated highest proportions of musicals, fantasy, family, and adventure movies-consistent with family-friendly content.

Based on our data analysis, if a customer is looking for sheer volume of movies, Prime Video is the obvious choice. For customers who are seeking content in languages other than English, Prime Video remains the top candidate. For customers looking a large variety of movie content but who don't want to be overwhelmed by options, we would recommend Netflix. If a customer is looking for content specific to a younger age range or with family-friendly content in mind, we would also recommend Disney+.

Limitations in the data:

Many of the movies in our data set were unrated for various descriptive elements which complicates interpretation. Where possible we attempted to omit or acknowledge the missing data.

Our data set was static in nature which is not consistent with actual content flow on streaming platforms. Streaming services update movie content often on a monthly basis. Because our data set was uploaded a year ago, it cannot reflect the current state of these streaming services.

Provided this were a real business venture that we were heading, we would want to continue to expand on our analysis for both verticals. Adding the visual representations to the predictive analysis may add value to our first vertical. Including television series and original platform content data into our analysis would allow for more complete picture of platform preferences. For example, Hulu often offers television shows the day after they air live on TV. If a customer wants to "keep up" with a television series, then Hulu may be a better choice for that customer despite Hulu's relatively low number of movies.