ORIGINAL CASE STUDY: MOVIE STREAMING PROVIDERS

Carlos Palomo

Purpose of the Project

With ever increasing rise in popularity and demand of video streaming services, trends in available content are becoming increasingly important to the entertainment sector for both distributors and viewers. This has become an even larger point of interest following the shutdown of traditional movie theaters due to the COVID-19 pandemic. Where following the Covid-19 pandemic internet users are spending 32% of their time on streaming devices and platforms and 54% of internet users are watching more shows and films on streaming services[1]. Furthermore, the usage of streaming services has greatly impacted how new movies being released as well how older movies are being accessible to the public. This has led to the problem of which service should consumers use as the number of services being offered is ever increasing.

Goal of the Project

To shed more light on this topic, I will investigate current movie offerings of the four of the most popular streaming services: Netflix, Hulu, Prime Video, and Disney+. Using the dataset[2] I will attempt to develop a predictive model that will help determine which of the four major streaming providers a movie will be most likely to be offered through or combination of, as there is potential that it can be offered in multiple services. This will be based off of what is already being provided by each of the streaming providers. I will perform exploratory data analysis to better understand the relationship both as a whole and for each specific provider. The final step will be to create a multinomial logistic regression for predictions.

Methodology

The data cleanup process began by removing the variables that were not going to be used for this analysis. Then removing any record with missing data which left a total 7,046 records with 18 variables, which was then split between a testing and training dataset. The decision to remove the records with missing values was due to the variety of the data missing. For example, the potential filming locations realistically would have varied based on movie genre, year of production, and potentially having multiple filming locations, all of which, or a subset, could be missing in addition to location for records.

The dependent variable, or target variable, was added based off of four columns, each denoting if the movie was currently being available through a specific provider. The main independent variables focused on this analysis are year of release, runtime, IMDb rating, genre, country of filming, language, director, and targeted age. In order to analysis this data genre, country, language, director, and age was encoded.

---

[1] Statistics found from marketing data company market.us; https://market.us/statistics/online-video-and-streaming-sites/

[2] Dataset created by Ruchi Bhatia found on Kaggle. This is the second version of the data set; https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

Through heatmaps I gained a high-level idea how these key variables are correlated to one another (Figure 1). Which looking at the figure it appears that all variables have some correlation to one another, though directors has the least correlation compared to the other variables. To explore the streaming providers further exlusively, I ran similar heatmaps for only records where the movies was available for that provider. The results were spread out though for each provider exclusive there was correlation with amongst the target variables. With Disney+ appearing to have the weakest correlation and Hulu having the strongest. I then proceded to use ExtraTreeClassifer (Figure 2) and Principal Component Analysis (PCA) to verify that each variable and linear combination of varibles should be used to create the model respectively. Both test showed some importance to each of the variables.

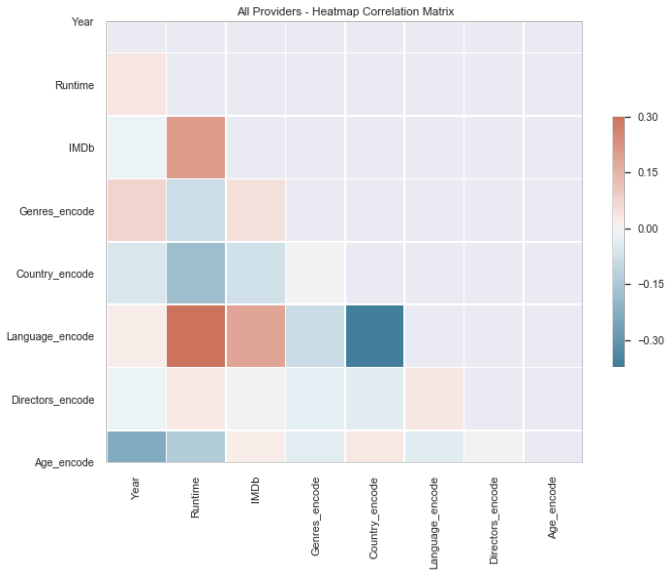

Figure 1: Correlation Heatmap of Independent Variables

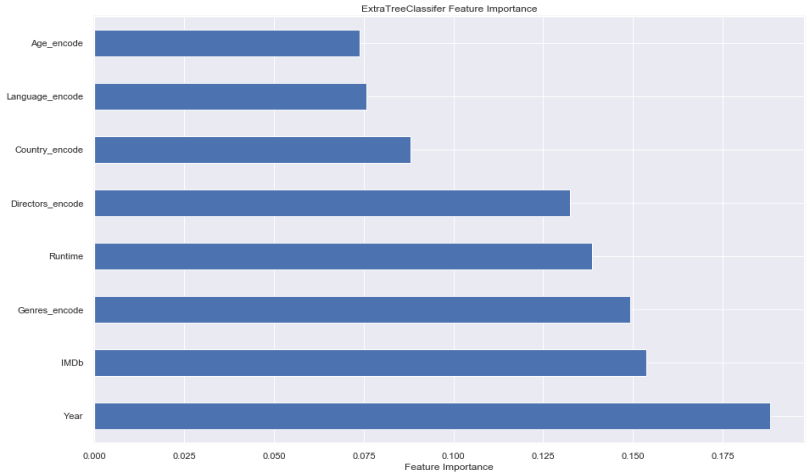

Figure 2: ExtraTreeClassifer Feature Importance of Independent Variables

Results

Overall the results were fair. Though the model did appear to have more trouble when a movie was available through multiple providers. Using a K-fold of 5 folds the model scored a cross validation score ranging from 0.680 to 0.709. Furthermore looking at the classifaction report, there are large range across the precision, recall, and f1-score (Figure 3), with model working best at predicting Prime Videos.

	precision	recall	f1-score	support
Disney+	0.70	0.55	0.62	154
Hulu	0.22	0.02	0.03	111
Hulu, Disney+	0.00	0.00	0.00	1
Hulu, Prime Video	0.00	0.00	0.00	54
Hulu, Prime Video, Disney+	0.00	0.00	0.00	1
Netflix	0.55	0.42	0.47	435
Netflix, Disney+	0.00	0.00	0.00	2
Netflix, Hulu	0.00	0.00	0.00	3
Netflix, Hulu, Disney+	0.00	0.00	0.00	1
Netflix, Hulu, Prime Video	0.00	0.00	0.00	1
Netflix, Prime Video	0.33	0.02	0.04	52
Prime Video	0.72	0.92	0.81	1299
Prime Video, Disney+	0.00	0.00	0.00	0

Figure 3: Classifaction Report of the Model

Conclusion

The model is not developed enough to properly predict which streaming services a movie will likely be in. Despite this result the model, the projects accomplishs showing that there is value in this type of machine learning model. To further develop the model severals thing can be done to imporve and make it more robust. The first imporvement would be to design the model to predict only one provider, as the multiple providers might be causing extra noise to the model and data. The second improvement would be to prefrom further feature reductions to help streamline the data. Finally the inclusion of another movie data source can help include a feature that is currently missing, such as production company and box office revenue. By improving the model it possible to predict which streaming provider will most likely carry a specific type of movie, which hold value for the customer as well as giving a better understanding demographics that each provider targets.