Sean McLean

ALY 6040

Professor Reilly

Module 3

Technique Practice

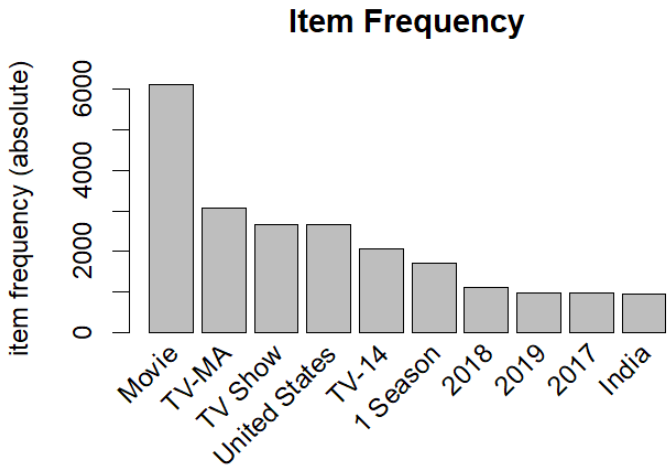
Introduction

The technique project in module 3 uses the final project dataset and one of the data mining methods to assess and visualize the data. The dataset being looked at pertains to a Netflix database of television shows and movies on the platform and the data mining technique implemented is association mining. The goal is to locate the most common patterns and frequencies when molding the dataset into transactions and item sets. This will help identify customer behaviors with how television shows and movies are selected and how recommendation systems are built based off these behaviors.

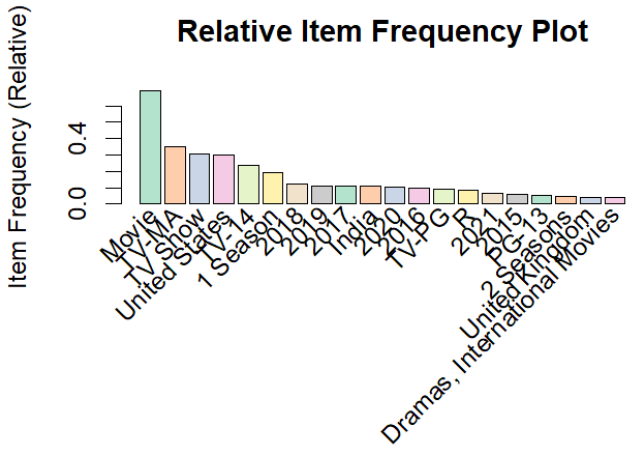
Code Walk Through

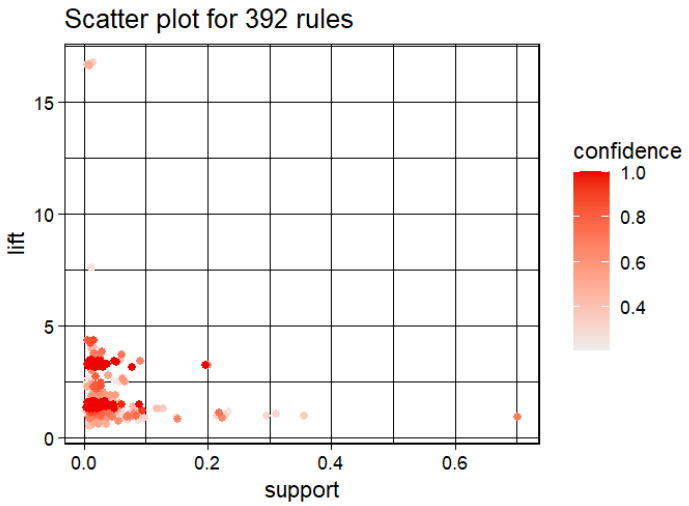
The Netflix dataset for the final project was uploaded into R after using the necessary 'arules' library function for downloading. A quick run through of the summary statistics of the Netflix dataset is observed for a refresher on the dimensions and its attributes. The dataset is then converted into a transaction object to allow for the proper analysis using association mining. The summary function is run to show the conversion and what are the most frequent items and the sizes of the transactions from all the rows in the dataset. The ‘apriori’ algorithm is then applied for the association rule mining that sets up the parameters and returns item sets and transactions. The first six rows of the dataset are then observed using the inspect function that shows the items on the left-handed side and right-handed side and their support values, confidence values, lift values, and the total transactions per association (ChatGPT, 2024).

Using the eclat function, the most frequent items are then calculated and displayed by showing their support and count values. These are visually presented with a bar chart with the item frequencies of the top items by total transaction counts.



The ‘apriori’ function is used again to create a new variable called ‘rules’ with the confidence metric being revised to 0.8. This is used to sort and inspect the transactions by confidence and by the highest lift. The length of the patterns is then adjusted to three items which then calculates them in total amount of rules and total amount of subset rules. An example is added to the rules function with director Rajiv Chilaka to show high confidence rules outputs and outputs of what customers watched if they watched a production with Rajiv Chilaka using high confidence rules. These outputs are visually displayed by first downloading the ‘arules’, ‘arulesViz’, and ‘RColorBrewer’ libraries, and the ‘apriori’ function to set up the parameters before creating the plots that show the most common item frequencies and level of confidence in the lift and support of the rules.





Analysis

After the dataset has been converted for association mining analysis, the summary shows that the most common items pertain to movies and television shows, the ratings TV-MA and TV-14, and productions from the United States. This would indicate that the most frequent viewers of Netflix are watching national productions for adult viewing. The association rules imposed on the dataset show the correlations between item sets with the ‘other’ item and ‘movie’ item relate to high values in the support, confidence, coverage, and lift measures. Whatever the other item is means that leads to an occurrence of the movie item which could be a rating, a director, cast member, or some other attribute. This could be a part of how a recommendation system is built for a user looking for a movie to watch.

An evaluation of the frequencies of the item relationships by support measure and count shows that movies and television shows are still quite common. The ratings of the content, the year the content was released, and the location of the production is always prevalent. The release dates are all recent, so viewers are seeking newer content over older movies and television shows. The only other country besides the United States is India which indicates that there is content from there that has a popular viewership following on Netflix. The top three items in terms of support measure values are movies, the rating TV-MA, and television movies, so adults are likely to be looking for those items with that rating over other demographics and preferences on the platform.

Observing the top rules in terms of high confidence, the item sets that pertain to television shows are reality TV, 4 seasons, and kids television shows and television comedies. From their high levels of confidence that equal one, we can be sure that users are watching these types of shows on Netflix. Shows with at least four seasons suggest that binge watching certain shows is still immensely popular on the platform. The one item that is connected to movies is the run times of movies which are 110 minutes, 107 minutes, and 87 minutes. These run times which are roughly between an hour and a half and just under two hours would indicate that this is a popular run time for viewers and could be looked at by Netflix for bringing in movies in the future that are around that time range.

The content that has the highest lift values are all television programming that are international shows and also comedies, reality shows, and kids programming. The large lift ratios of support and confidence shows that there are relationships between certain types of content with international shows on the platform. When incorporating the director Rajiv Chilaka into the rules with high confidence as an example, it shows what people that watch his productions that he directed will also watch in the future. The items with the highest confidence values are other kid's programs like movies and television shows and also programming with particular actors in the content. This correlation shows that viewers will most likely watch content that is like the work of Chilaka which makes sense with kids watching only kids rated programming.

The most frequent items in terms of the association rule measures are visually presented in two plots shown above. The most common items relate to ratings, release dates, and where the content takes place geographically. This shows that Netflix has content that is relevant outside the United States and that people are watching newer shows with a certain rating. The scatterplot that has 392 rules is clumped together with a lift value below five and a support value below 0.1. There are some outliers in the plot that could be looked at more deeply with how they relate to the dataset overall.

Interpretations & Recommendations

The focal point of the database with Netflix is still movies and television shows and it is prevalent among the associations of all the items. Some of the features of what is available on the platform include newer content which means that from past historical data that viewers want what is new more and they should continue to stay up to date on what is trending. From the

correlations there is also popular content from outside the United States that people are tuning into. I recommend that Netflix continues to look beyond the national landscape for quality content because from the data there is a market for quality television and movies from other countries like India and South Korea where the association mining results show that viewership is high from shows that are released from those parts of the globe.

An interesting aspect of the results is that the features of movies that rated highly from the data mining were the run time, ratings, and the release dates. This gives the platform the opportunity to release an array of different genres for adults that are from the last five years and are also within the runtime that is most common to users. I would recommend the platform brings as many hit movies to the platform that fits this mold because the genre is not as important as the other popular traits that the mining indicated. This is also supported from high confidence values that predict that these shows together with their characteristics have high confidence for future occurrences.

The characteristics of the television shows center more on certain genres like kids programming, comedies, and reality shows. The results also would suggest that there are shows that viewers are binge watching that has been a staple of streaming services the last few years. I would recommend bringing in more shows of this caliber where you have audiences of all age ranges that can binge watch different genres of shows that ran for multiple seasons. This could be ideal for families with children who can watch their favorite shows while the parents can watch shows from genres like comedies, international content, or reality shows. This is a demographic that the platform that should pursue or continue to pursue to increase its total subscribers. These results are also backed up by high lift and confidence scores that predict that they will provide these results in the future. This can also be used to build stronger recommendation systems for all Netflix viewers based on the data mining results.

Conclusion

The association mining technique is an effective tool to be used when predicting what users will watch based off its past viewing history. The content on Netflix which is primarily movies and television shows is the primary force in its relationships with other items in the dataset. It is also a valuable way to see how items are linked together and from those correlations how patterns and trends can be identified. The platform can use these data analysis results to select material that is

based off what has been popular and what kind of programming will most likely be watched in the future.

References

ChatGPT. (2024, February 28th). Default (GPT 3.5). https://chat.openai.com/

Kabacoff, R. (2015). *R In Action. Data analysis and graphics with R. Second Edition.* Manning.