Bisola Olawuyi

Netflix report

03/22/2024

Objective - Hypothesis

This project's primary goal is to examine Netflix's content to identify genre-specific patterns and then use that information to improve the recommendation system. There's a hypothesis that trends in genre preferences can be found by looking at Netflix's programming. In principle, these patterns may be used to build predictive models that recommend TV shows or films based on users' past viewing preferences, increasing the effectiveness of Netflix's recommendation algorithm.

Data Description

To begin with, I imported the pandas library in order to manipulate the data for the analysis. Using pd.read_csv, the Netflix content data was loaded for a CSV file called "netflix_titles.csv." I then executed data information which printed out a table summary for the data frame. I proceeded to check for missing values , then made sure to remove rows with missing values in the 'country' column.The data was then grouped by both country and content type, in my opinion, as this makes it easier to examine how different countries deliver different types of content (movies or TV series).Following that, I used the values in the 'type' column of the original DataFrame content_by_country to construct two distinct data frames: movie_data, which contained only movies, and tv_data, which had only TV shows. Loading the Tv data was very crucial for me so I could see exactly what was in the data which contained the number of TV shows in different countries. I then proceeded in displaying the distribution of countries for the TV shows in the 'tv_data' data frame.The quantity of TV shows from each country is represented by the y-axis, while the x-axis indicates the countries. Using the same code, I produced a histogram that showed the distribution of the countries in your movie data. Countries are displayed on the x-axis, while the number of movies from each country is displayed on the y-axis. By contrast, I found that the movie distribution reaches a maximum of 30, while the TV show distribution has a maximum value of 10 on the y-axis. This implies that, in comparison to movies in this dataset, TV shows are typically less numerous per country.

My next goal was to determine which of the data's columns would need dummy variables in order to do machine learning tasks. Iterating through the column names in a list (column_list), it determines whether each column is present in the DataFrame. If a column exists, the function uses.unique() to extract its unique categories (distinct values). Finding unique values is still the main function; the final line transforms the result to a list. I then used 'type_cat', 'country_cat', ‘rating_cat’ and ‘listed_in’. I did this so that I could see the data in each of these segments to help me with executing my dummy variables which was the next step. For the dummy variables of ‘type’ , I created a dictionary to map text categories in the type column to numerical values (1 for "TV Show", 2 for "Movie").I then replaced the categorical values in the type column with their corresponding numerical values from the dummy dictionary. After incorporating the dummy variables I identified X and Y, then proceeded in splitting the dataset into training and test sets to be able to carry on with Logistic Regression,Knn and Decision Tree for the accuracy scores.

Process

I would say my biggest problem throughout this project was the dummy variables. I had to do them a lot of times and also research for more knowledge to make sure I was on the right track. I was finally able to execute all the dummy variables because I kept pushing and didn’t give up. Especially for the country listings I had to do the coding in a way that the commands were already integrated into the list to make it easier to attempt the dummy variable. I would say this technique definitely helped me and made the whole process move faster.

Learnings

Understanding what content is available in different countries:

I was able to analyze contents available in different countries using visualization and the content type had a lot of TV shows that were mostly international movies, drama, Horror, action and adventure. The years of these movies ran across from 2008- 2020.

Netflix has been increasingly focusing on TV rather than movies in recent years. Is this true? Answer using a visual.

I would say according to my visual this is not necessarily true given the fact that movies reaches a maximum of 30 on the chart ,while TV shows only reach a maximum of 10

From this data I drew that there are actually more movies than Tv shows across these countries and I realized this through the visualization I created. Another thing I noticed while doing this project was that there were multiple countries named ‘Argentina’ and they all had different numbers in relation to how many TV shows were present. Australia had the highest count of TV shows which was 46 while only Argentina had 16. So by looking at this I was able to come to the conclusion that Users preferred TV shows to movies in these areas.

Conclusion

In conclusion, the visuals and the coding tell us that users have specific shows they prefer and it it linked to the country they are in.