Group #6

Class: CAP4770

Sufia Rashid (64044256)

Andrea Allmon (43669862)

Emily Olsen (11135847)

Categorizing Movies on Letterboxd Based on Genre and Popularity

In this project, we will be utilizing Letterboxd data to categorize movies based on factors such as genre and popularity. We are hoping to draw conclusions from the categorization related to popular genre combinations and how well they are received by the public. In a real-life application, this analysis could be a helpful tool used by those in the entertainment industry to figure out what appeals most to the modern day movie-watcher. In order to organize and read the large data file of movie information, our team will use neural networks for the prediction of popularity amongst Letterboxd users and genre type. Dimensionality reduction will be an important factor in the preprocessing of our data as the dataset contains several other features that we will not need for our project. Some of the extra features include an url to a movie image, an imdb id and link, overview of the film, runtime and list of spoken languages in the film. Furthermore, we will reduce our dataset by only using those movies filmed in English. This step in preprocessing the data will be crucial to further organize the data according to genre and rating.

Details of the data set:

The dataset contains 10 different columns containing several different aspects of movies from Letterboxd, including titles, ratings, popularity, genre description, ID, etc. The team has chosen mostly to focus on ratings, titles, and genres for the project. Many of the other columns of the dataset are incomplete and thus would be less valuable. Furthermore, the team will make use of the column “movie language” to cut down the data. Column “year released” will be included as an additional factor with “movie title.”

Problem statement:

Our team has chosen to work on accomplishing the goal of organizing and separating data of movies based on the genre. Our problem consists of determining the optimal strategy to create categories for the movies while ensuring that the data in each category remains true to the genre. After categorization, we will be using a neural network to predict what categories could be popular in the future.

Evaluation metric:

One main evaluation metric the team will be using is accuracy. We will use this metric to classify how well our product is working the way it is supposed to and if it is ready to be considered a final product. We will also use precision to measure the accuracy of the positive categorizations of movies in each movie genre. The last evaluation metric will have to be a data-time-to-value metric as the team will have to ensure that the product is able to achieve the proper values within a reasonable time, especially when working with a fairly large dataset.

Baseline techniques to be used:

With neural networks we can find and predict correlations between different movie genres and their ratings, and use that to predict future popularity of categories. Attribute/Feature selection will be used to narrow our dataset to the genre, rating, title and year released sections. For the movies unassigned to a genre, imputation may be used to assign them to a non-categorized group or they can be excluded from the data via data reduction. We will also be using graphs to visualize the difference in popularity between the different genres, both present-day and potential future trends.

Improvements after the last presentation:

Two different analyses were taken on the dataset, one through a basic evaluation of the 3 most commonly watched movie genres popularity each year and the other through a neural network model of all movie genres popularity throughout the years. We changed our predictions from a five, ten and twenty year timeline to be more generalized just as future predictions as we had no way of specifying the five year mark.

Details on problem Statement:

There are many reasons why categorizing and organization of datasets is of great importance and a problem that requires a solution. The organization of the dataset allows for a logical objective as to what the data is representing, and in the case of the team project, what type of movies each category is presenting to the user. Furthermore, this would allow for Letterboxd or anyone to be able to analyze what movie genres are most popular, most common, or most sought after. Predicting what categories of movies will be the most popular in the future is also a major tool for those who produce movies, so they can accurately plan what kind of movies they will make.

The team has chosen to work in Python with many data-focused libraries to solve our problem. The team first separated into three different responsibility groups and took over different sections of the code. The various sections consisted of grouping the column title and year together, linking column genre with its movie info, and organizing the output into clean, easily readable sections. Furthermore, the team created a step-by-step plan to achieve the final product.

Since the dataset had been taken from kaggle, part of the team implemented the code inside of a kaggle online notebook, which also allowed for integrating the large dataset file without taking up memory on the local computers. The dataset used for the project is linked below.

https://www.kaggle.com/datasets/samlearner/letterboxd-movie-ratings-data

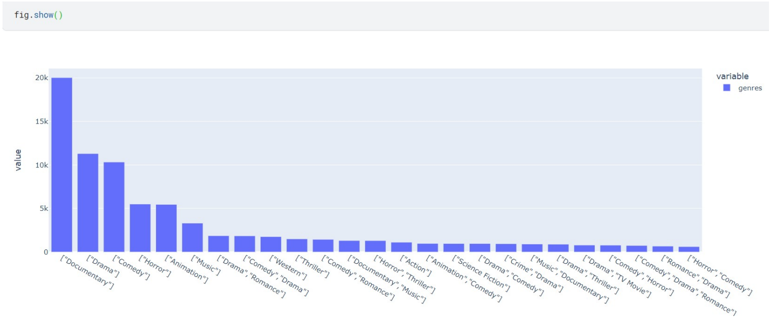
Approaches and programming tools:

One of the approaches the team used was keyword matching. We used the given list of keywords associated with each genre to analyze the movie title, year, popularity score of each movie. Furthermore, Python Pandas DataFrame drop() function was used to filter all noise and reduce the dimensionality of our dataset. The string.contains() method was used to ensure that only english language movies are taken in and the python library plotly.express was imported so that our team could show visualizations and plot out our data on charts.

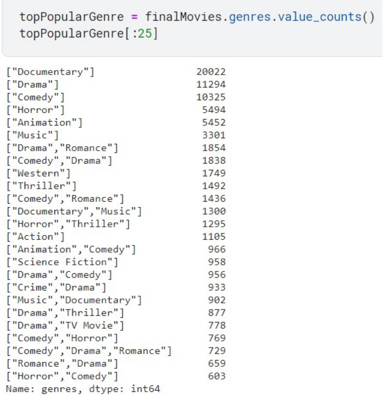
Lastly, our main approach was to use neural network models as our machine learning algorithm. To get the current predictions and correlations, first the genres and movie_titles columns were converted to numerical representations and the popularity scores were normalized. The team then worked to split the dataset into training and testing subsets, which ultimately allowed for the creation of our neural network model.

Evaluation setup:

There were two different evaluation methods used in our project. For the first evaluation method, we set up our data based on the results of the 25 most commonly watched/popular movie genres in order to gather information on the popularity of each genre. This allowed us to get an accurate visualization of what genres would best offer the most information.



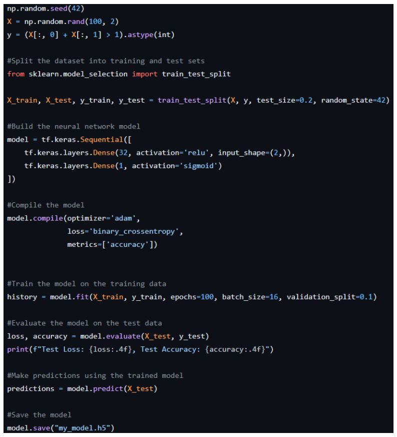
```
fig.show()
```



```
topPopularGenre = finalMovies.genres.value_counts()
topPopularGenre[:25]
```

```
["Documentary"]                20022
["Drama"]                      11294
["Comedy"]                     10325
["Horror"]                      5494
["Animation"]                   5452
["Music"]                       3301
["Drama","Romance"]             1854
["Comedy","Drama"]              1838
["Western"]                     1749
["Thriller"]                    1492
["Comedy","Romance"]            1436
["Documentary","Music"]         1300
["Horror","Thriller"]           1295
["Action"]                      1105
["Animation","Comedy"]           966
["Science Fiction"]              958
["Drama","Comedy"]               956
["Crime","Drama"]                933
["Music","Documentary"]          902
["Drama","Thriller"]             877
["Drama","TV Movie"]             778
["Comedy","Horror"]              769
["Comedy","Drama","Romance"]     729
["Romance","Drama"]              659
["Horror","Comedy"]              603
Name: genres, dtype: int64
```

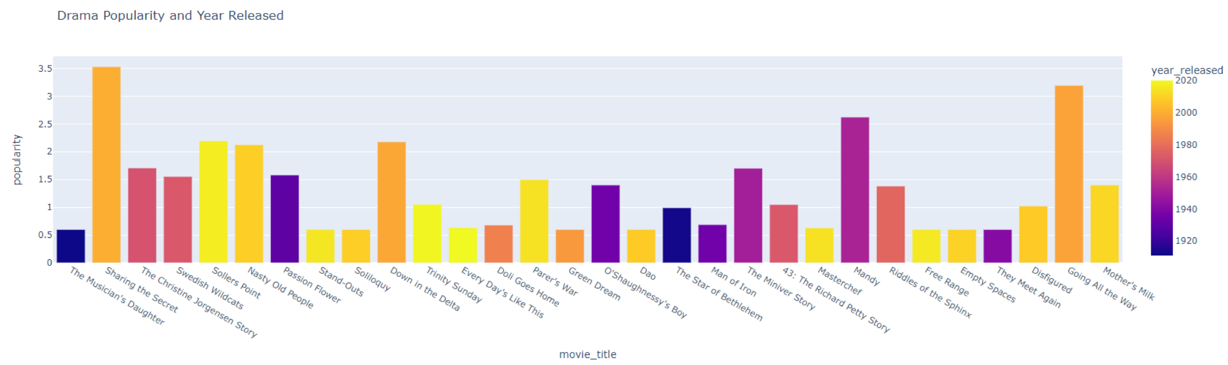
The second method of evaluation (seen through a neural network model) was set up by first creating a test model to ensure our program was running correctly. This test model provided us with a basic understanding of how our team should go about training data for our final model for finding the future popularity results of genres.



```
np.random.seed(42)
X = np.random.rand(100, 2)
y = (X[:, 0] + X[:, 1] > 1).astype(int)

#Split the dataset into training and test sets
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

#Build the neural network model
model = tf.keras.Sequential([
    tf.keras.layers.Dense(32, activation='relu', input_shape=(2,)),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

#Compile the model
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])


#Train the model on the training data
history = model.fit(X_train, y_train, epochs=100, batch_size=16, validation_split=0.1)

#Evaluate the model on the test data
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Loss: {loss:.4f}, Test Accuracy: {accuracy:.4f}")

#Make predictions using the trained model
predictions = model.predict(X_test)

#Save the model
model.save("my_model.h5")
```

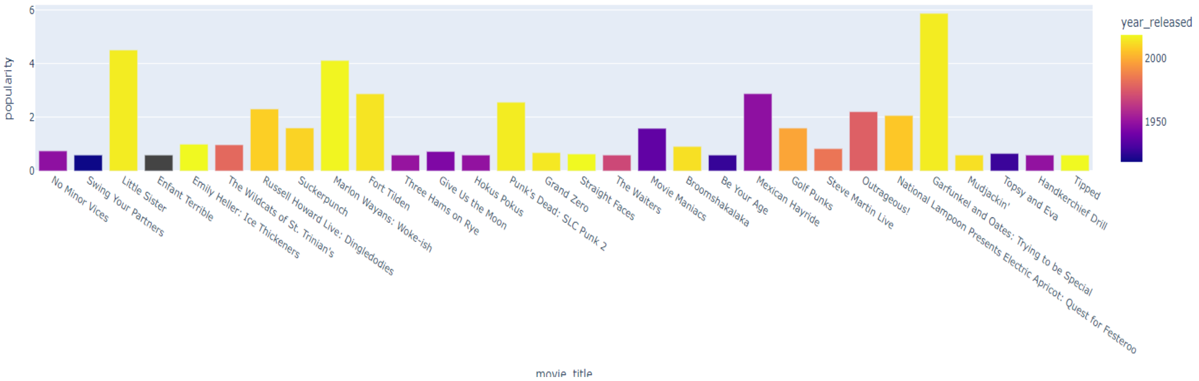
Analysis of results and comparisons:

After cleaning and preparing the data, the data set consisted of 163,610 movies and their respective genres, language, release year and popularity rating. From there, we constructed bar plots of the three most popular genre’s top 30 movies comparing their popularity and using a color scale to identify patterns within the years released. The final data set consisted of 11,294 “Drama” movies, 20,022 “Documentaries” and 10,325 “Comedies” amongst other genres such as “Romance”, “Animation” and “Horror”.

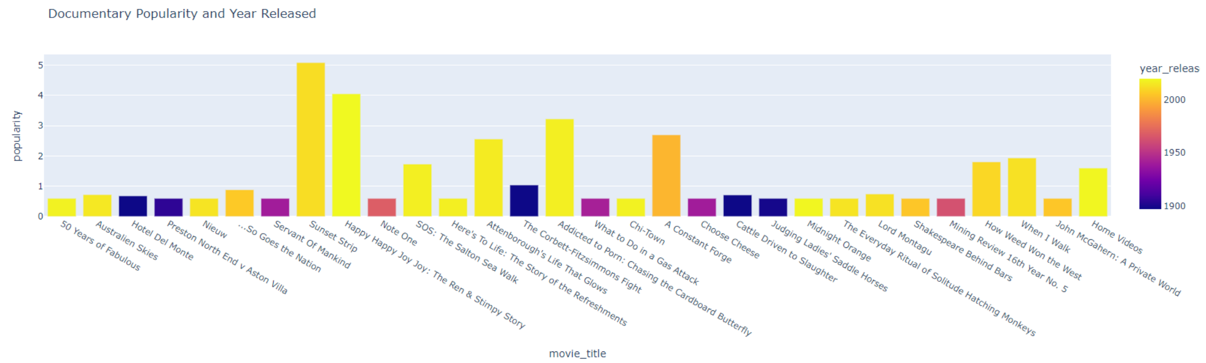


For the drama genre, only about 26 percent of the movies displayed were released between 1920 through 1960 and had below median popularity scores. It can also be noted that the two movies with highest popularity indices were released around the year 2000.

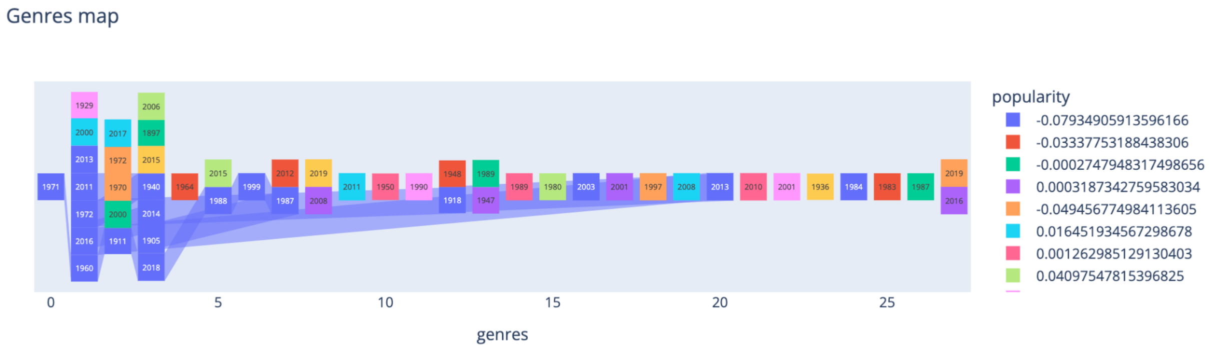
Comedy Popularity and Year Released



The comedy movies showed impressive peaks post 2000 release dates; of the five comedy titles with the highest popularity indices, 4 were released around 2020. Titles released before 1960 were greater numerically at 36 percent but still staying low in popularity with only one title surpassing the 2 index.



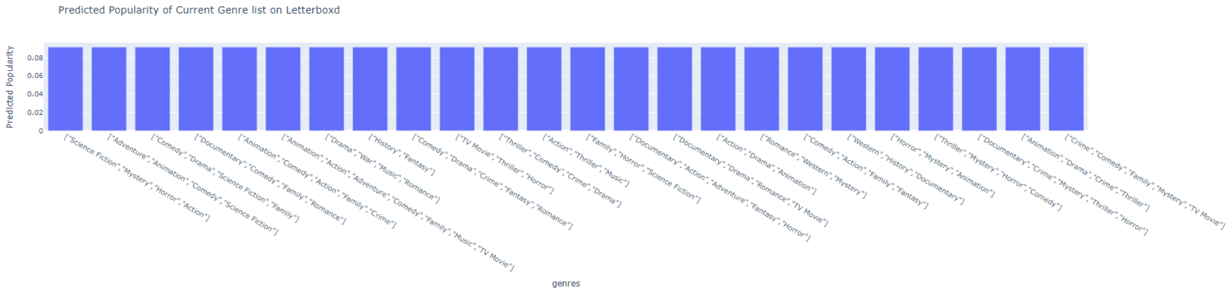
Documentaries had the most titles in the final movie data set, with close to 9000 more titles than its runner-up, drama. From the genres analyzed it seemed to have the most yellow bars (meaning a more recent release date). This could indicate an increase in documentary production and popularity after the year 2000. Documentaries from 1900 to 1950 hold about 26 percent of the 30 top documentaries and all but one had a popularity index less than 1.



Our second method of evaluation was using a neural network model to predict future genre popularities. We utilized tensorflow and sklearn.model_selection to create, train and test our neural networks. Our bar plots display the predicted popularities based on neural network predictions. We enumerated the genres for training and testing purposes and used a reverse mapper to change the values back to genre names for plotting purposes. The first bar plot shows an overall prediction for the genres. There is a large range in predictions amongst the first three genres to be noted. Also to be noted is that the majority of the genres will increase in popularity and many converged to around the same number. The neural network predicts increased popularity in multi-genre films and a decrease in popularity to those conforming to only one genre; this can be seen best in the third bar plot.

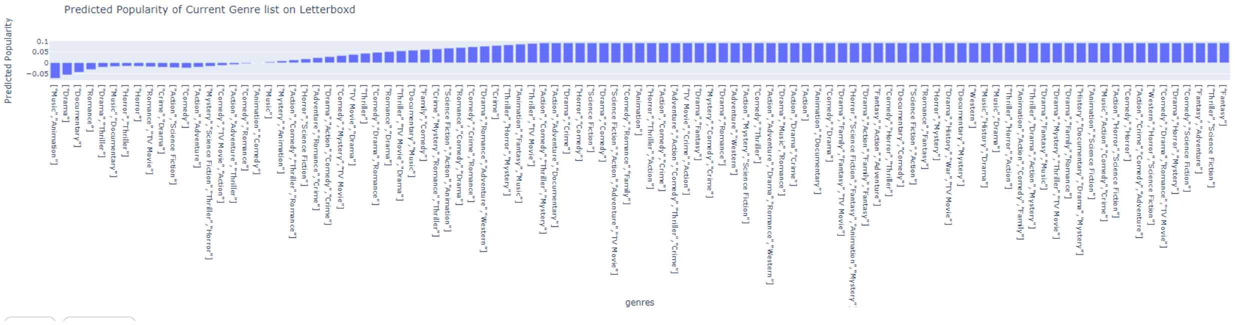
Predicted Popularity of Current Genre list on Letterboxd
Predicted Popularity
0.1
0.05
0
−0.05
genres

Another bar plot displays the 25 most popular predicted genres, notice none of them conform to solely one genre. The prediction seems to conform to the idea that users of Letterboxd rate movies with multiple genres higher, as movies with one genre might lack certain dimensionality.



This chart shows the 100 least popular genres. As can be seen, even though genres such as documentary and romance were most common on the genre list, they still came in with negative popularity, indicating that they will overtime become less and less popular. 17 genres were below

0 popularity, their popularity ranging from 0 to -0.08.



Conclusion - Challenges and lessons learned:

There are two main challenges that our team faced while working through our project. Firstly, because of the expansive nature of our dataset, we suffered from the "Curse of Dimensionality." The large number of dimensions in our dataset might have added some valuable information to our findings but overall increased the redundancy and noise of our analysis. Furthermore, without dimensionality reduction being applied, computational efforts needed to process our given datasets and analysis were greatly increased. The lesson our team learned from this challenge is the importance of dimensionality reduction and that more information does not always equal better or more accurate results.

Another challenge our team faced is working with neural networks and figuring out the best way to apply this to our project. Neural networks can be quite data hungry and thus, if you do not have a lot of data to train them, tend to overfit easily. They also tend to be computationally intensive which is difficult since our team used regular laptops to run the network.

The lesson our team learned from this experience is the correct way of training data with neural networks and also making sure that enough data is being used to train a neural network model. Furthermore, through our success in dimensionality reduction as well as creating a neural network model, the greatest lesson we learned was that there is a reasonable solution to every problem as long as time and effort are put into solving it.