Predicting Movie Ratings

By Matthew Coulombe

September 29, 2023

Data Source:
https://www.kaggle.com/datasets/benjameeper/movie-violencesexprofanity-data?resource=download

Problem Statement

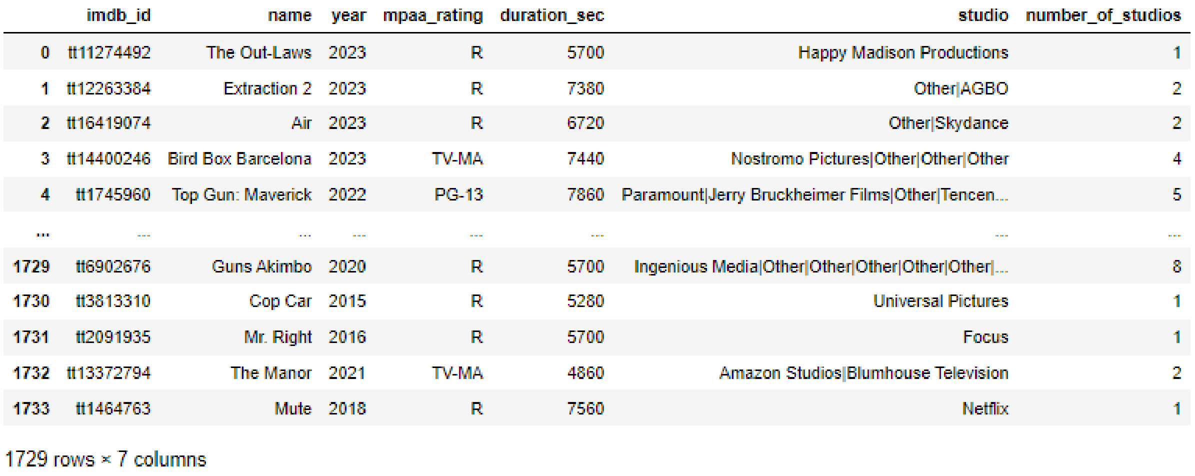
How can we take movie features such as language, violence, drugs, etc. and predict the MPAA movie rating? Some applications of this would be to allow movie review sites to make better initial evaluations for new movies and allow its users to make better decisions for what movies they should watch. Another application would be to find what the rating of a movie would be like if it has not yet been rated. Lastly, streaming platforms could use it as a parental control device to filter on specific content for movies.

Data Wrangling

Sourcing the data from a Kaggle user who has scraped content filtering movie data from the website VidAngel, we have a list of available movie tags, a list of movies with the ratings, duration, and studios, and a list of all the tags associated with these movies. From a cleaning perspective, we dropped the unnecessary columns and duplicate rows, handled missing values, and cleaned the studio column to create an “Other” category for any studio with 2 or less movies in our dataset. We also mapped studios under the same umbrella together so we could hopefully find some interesting insights into studios that have made a lot of movies. Lastly, we created a new feature for the number of studios associated with each movie as we found multiple studios were within the studio column.

After cleaning our data, here is what each of our DataFrames look like.

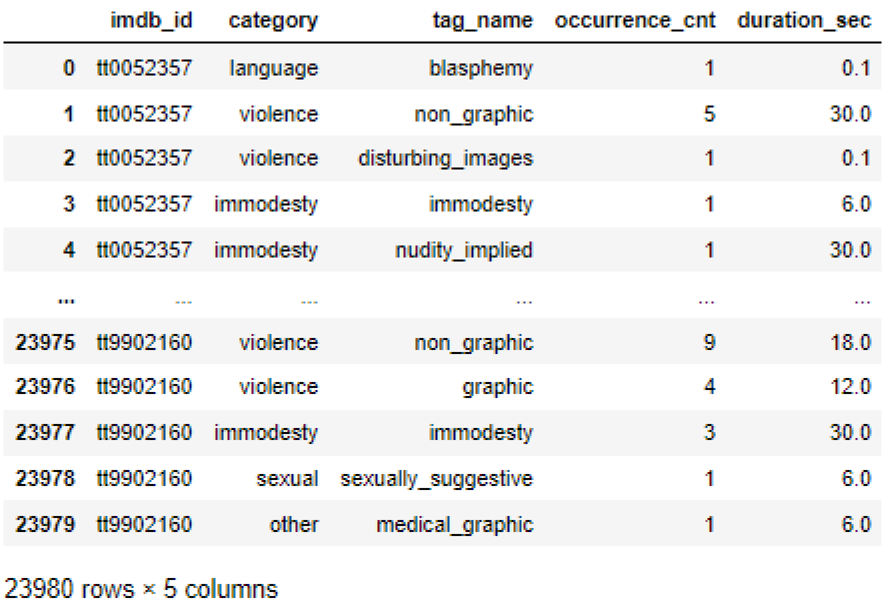
Movies



	imdb_id	name	year	mpaa_rating	duration_sec	studio	number_of_studios
0	tt11274492	The Out-Laws	2023	R	5700	Happy Madison Productions	1
1	tt12263384	Extraction 2	2023	R	7380	Other|AGBO	2
2	tt16419074	Air	2023	R	6720	Other|Skydance	2
3	tt14400246	Bird Box Barcelona	2023	TV-MA	7440	Nostromo Pictures|Other|Other|Other	4
4	tt1745960	Top Gun: Maverick	2022	PG-13	7860	Paramount|Jerry Bruckheimer Films|Other|Tencen...	5
...	...	...	...	...	...	...	...
1729	tt6902676	Guns Akimbo	2020	R	5700	Ingenious Media|Other|Other|Other|Other|Other|...	8
1730	tt3813310	Cop Car	2015	R	5280	Universal Pictures	1
1731	tt2091935	Mr. Right	2016	R	5700	Focus	1
1732	tt13372794	The Manor	2021	TV-MA	4860	Amazon Studios|Blumhouse Television	2
1733	tt1464763	Mute	2018	R	7560	Netflix	1

1729 rows × 7 columns

Movie Tags



	imdb_id	category	tag_name	occurrence_cnt	duration_sec
0	tt0052357	language	blasphemy	1	0.1
1	tt0052357	violence	non_graphic	5	30.0
2	tt0052357	violence	disturbing_images	1	0.1
3	tt0052357	immodesty	immodesty	1	6.0
4	tt0052357	immodesty	nudity_implied	1	30.0
...	...	...	...	...	...
23975	tt9902160	violence	non_graphic	9	18.0
23976	tt9902160	violence	graphic	4	12.0
23977	tt9902160	immodesty	immodesty	3	30.0
23978	tt9902160	sexual	sexually_suggestive	1	6.0
23979	tt9902160	other	medical_graphic	1	6.0

23980 rows × 5 columns

Exploratory Data Analysis

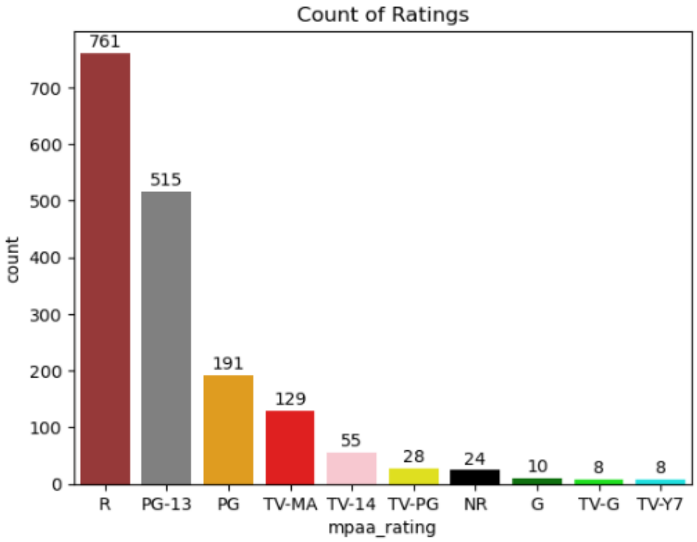
Once we cleaned our data, we then explored our data to see if we could find any interesting insights in the data.

Movies

Rating

After reviewing the data in our movies dataframe, we found that we have 10 total ratings for our movie ratings.

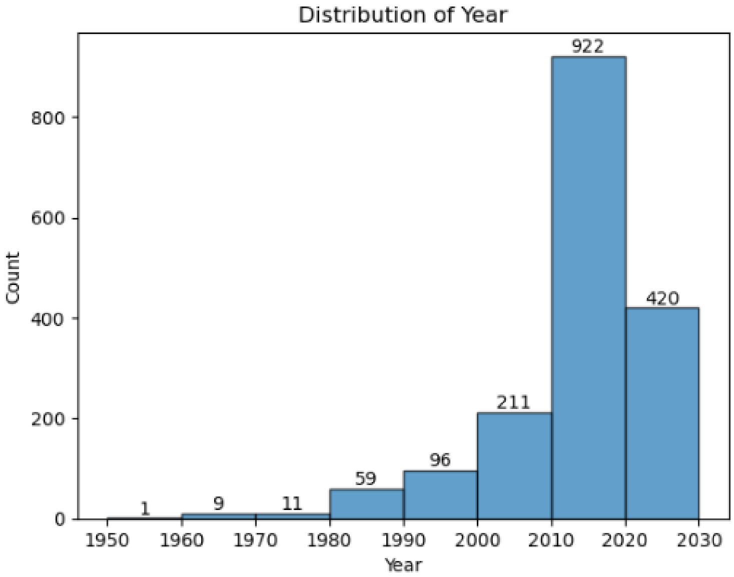
1. G
2. PG
3. PG-13
4. R
5. NR
6. TV-G
7. TV-Y7
8. TV-PG
9. TV-14
10. TV-MA



Of all of our ratings, we found that most of the TV ratings had very small counts compared to our total dataset. R is the most common rating followed by PG-13.

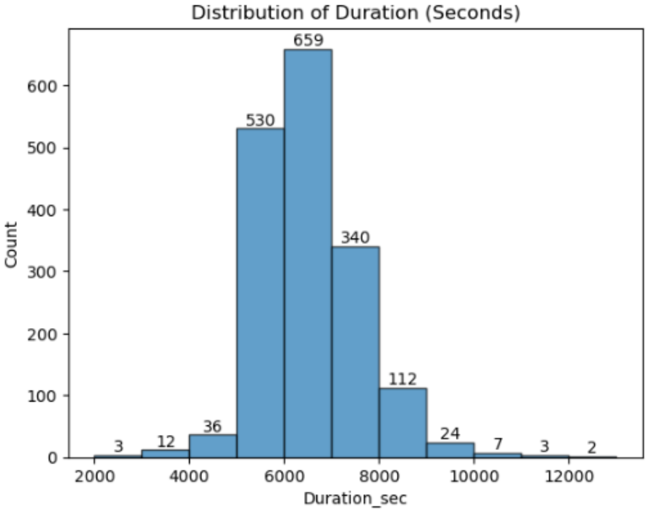
Year

Looking at the distribution for the years the movies in our dataset are from, we see that we skew toward more recent movies. However, it is still important that we are able to train our model with a mix of movies from different years as we want to be able to predict a movie's rating from any time period, not just for recent movies.



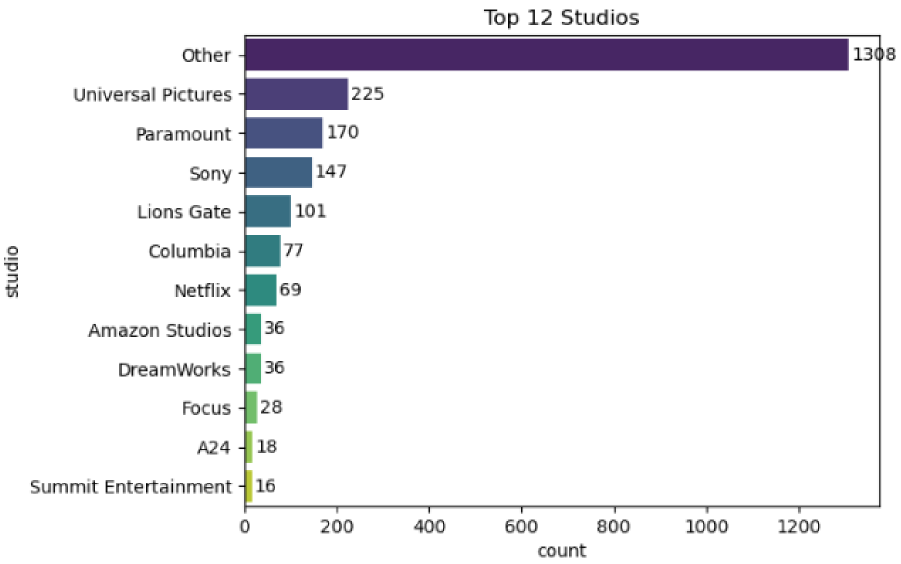
Duration

For duration of movies, we see the average is approximately between 6000-7000 seconds which corresponds to around 1 hour and 40 minutes - 2 hours.

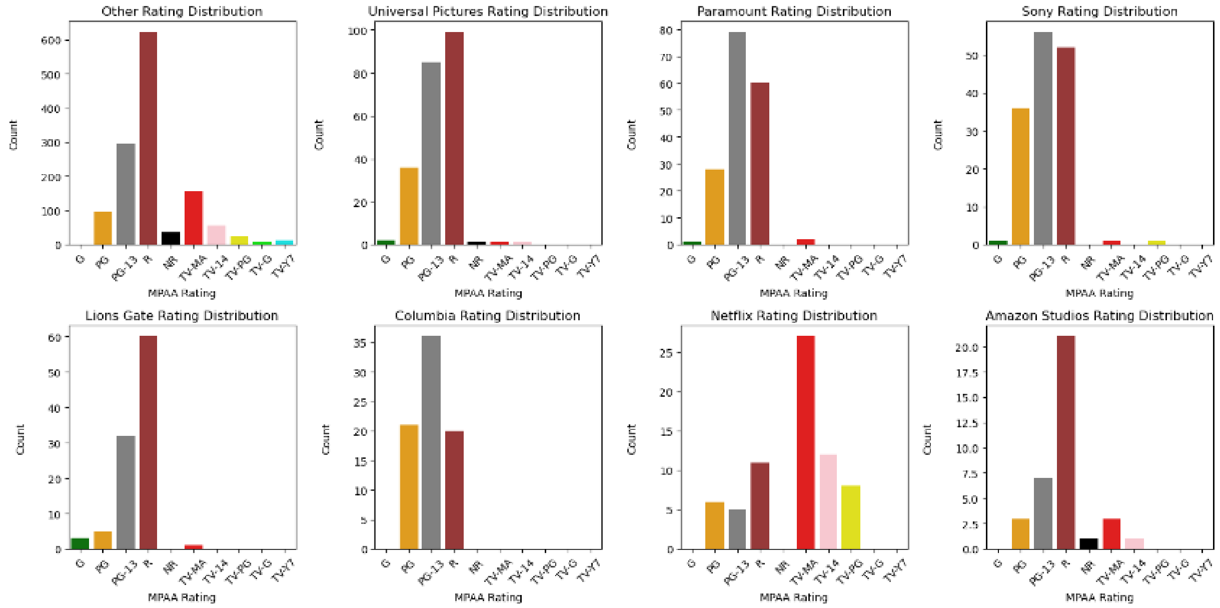


Studio

Looking at the top 12 studios, we see that Other is our most common studio which means we have a lot of studios in our data that have not made over 2 movies. This means for a majority of the movies we have, the studio will not be very helpful. However, for the ones that do have a decent amount of movies in our data (Universal, Paramount, Sony, Lions Gate, Columbia, and Netflix), our models may be able to gain some knowledge for those studios in particular.

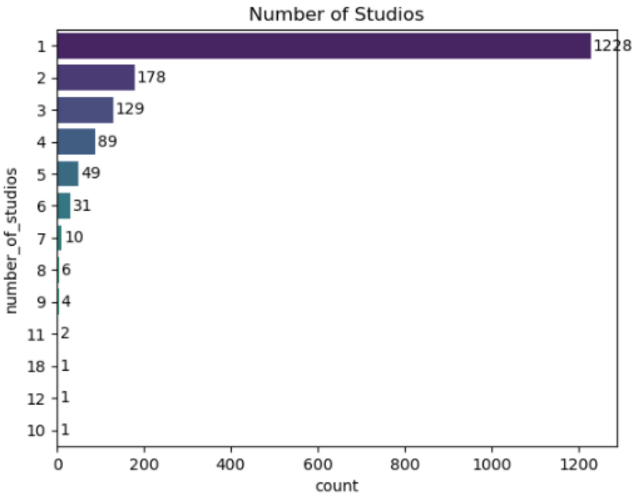


We do see when looking at the distribution of ratings for the top 8 studios, that Paramount, Sony, Columbia, and Netflix all have other ratings of movies that are higher than the R rating.

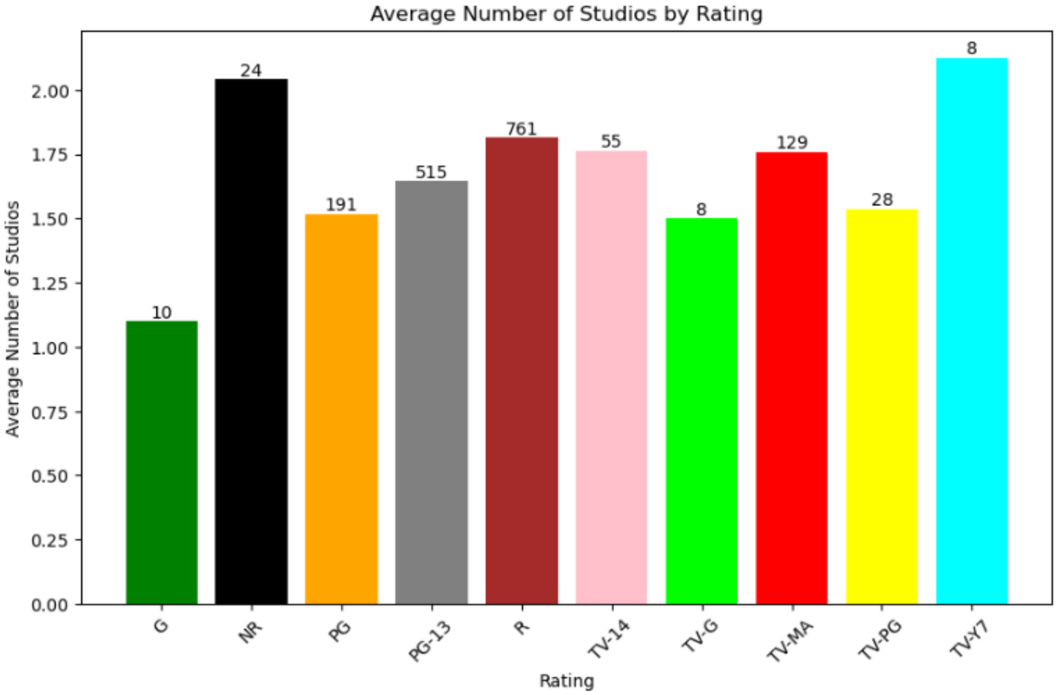


Number of Studios

For the number of studios related to a movie, we see that most of our movies have exactly 1 studio. The ones that have greater than 7 studios don't have as much data, so they will not be as very helpful for our analysis. Movies with 2-6 studios may give some information for our model.



When we review the relationship for the average number of studios per rating, we see there doesn't appear to be much of a trend. This shows that studios tend to collaborate on all different types of movies.

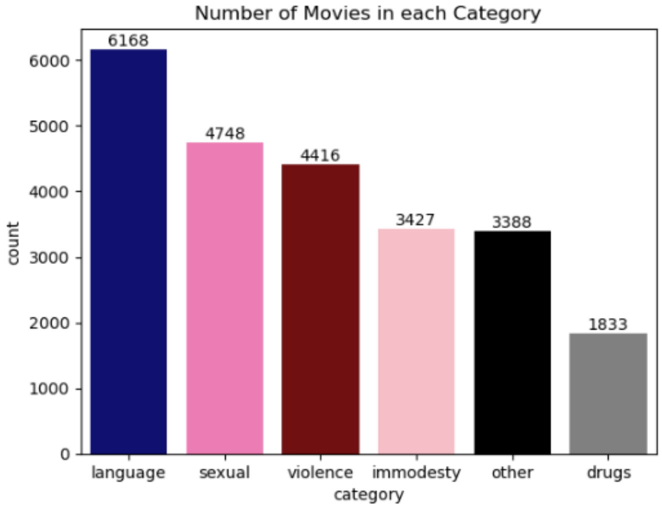


Movie Tags

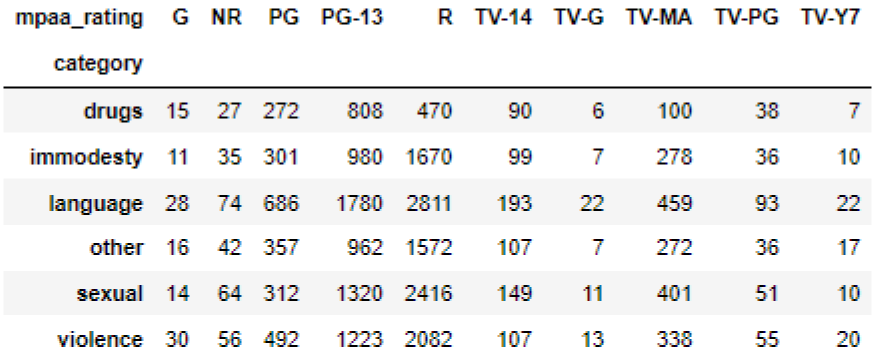
Category

After reviewing the data in our movie_tags dataframe, we found there are 6 movie tag categories.

1. Language
2. Sexual
3. Violence
4. Immodesty
5. Other
6. Drugs



The language tag is the most common, and the drugs category is the least common. Let's see the distribution of these categories for each rating.

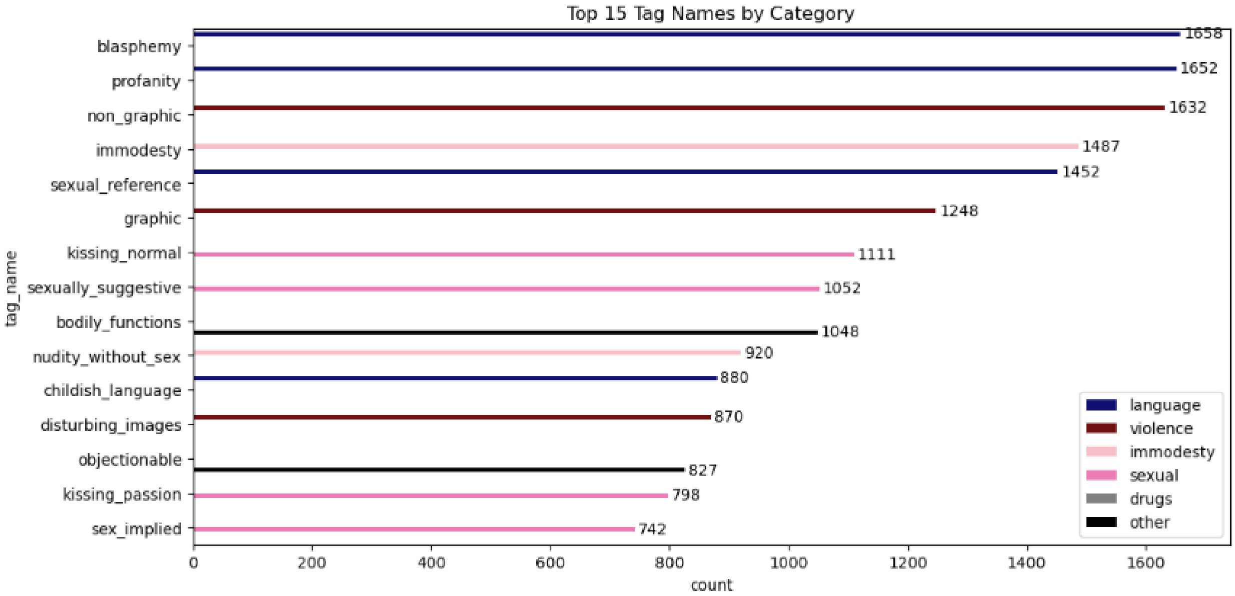


mpaa_rating	G	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
category										
drugs	15	27	272	808	470	90	6	100	38	7
immodesty	11	35	301	980	1670	99	7	278	36	10
language	28	74	686	1780	2811	193	22	459	93	22
other	16	42	357	962	1572	107	7	272	36	17
sexual	14	64	312	1320	2416	149	11	401	51	10
violence	30	56	492	1223	2082	107	13	338	55	20

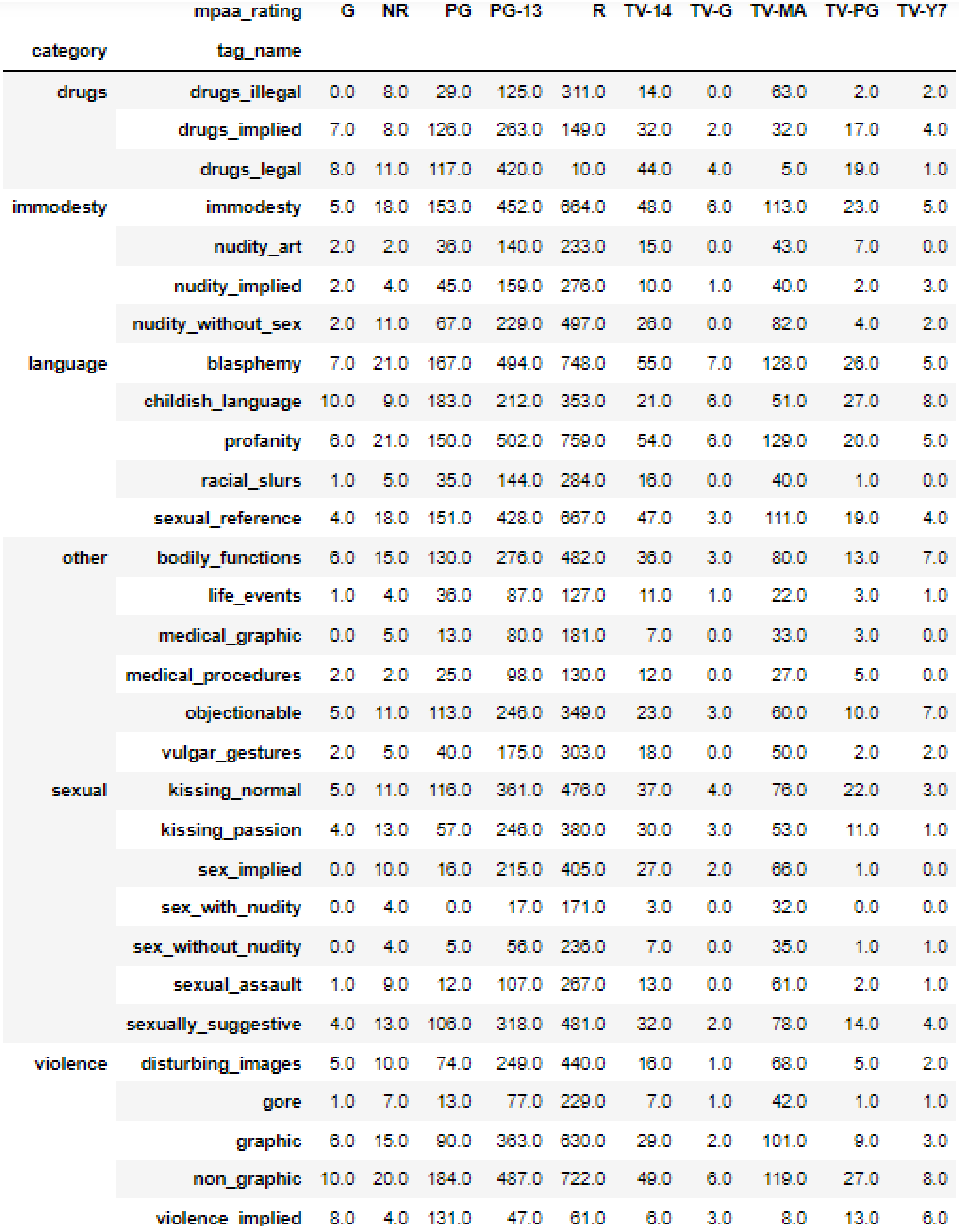
We see that language, which has the most movies generally, is also the highest for each rating. This is why language is the most common tag category that we have. On the other hand, for the drugs category, we see that it is highest in the PG-13 rating which is an interesting correlation and may be something our model can pick up on.

Tag Name

Looking at our top 15 Tag Names, we see that each tag name belongs to only 1 category. We also see an even mix for all the categories except for Drugs, which makes sense as that is the lowest category in terms of volume of tags.



We can also look at the distribution for each tag name in each category and what rating they correspond to.

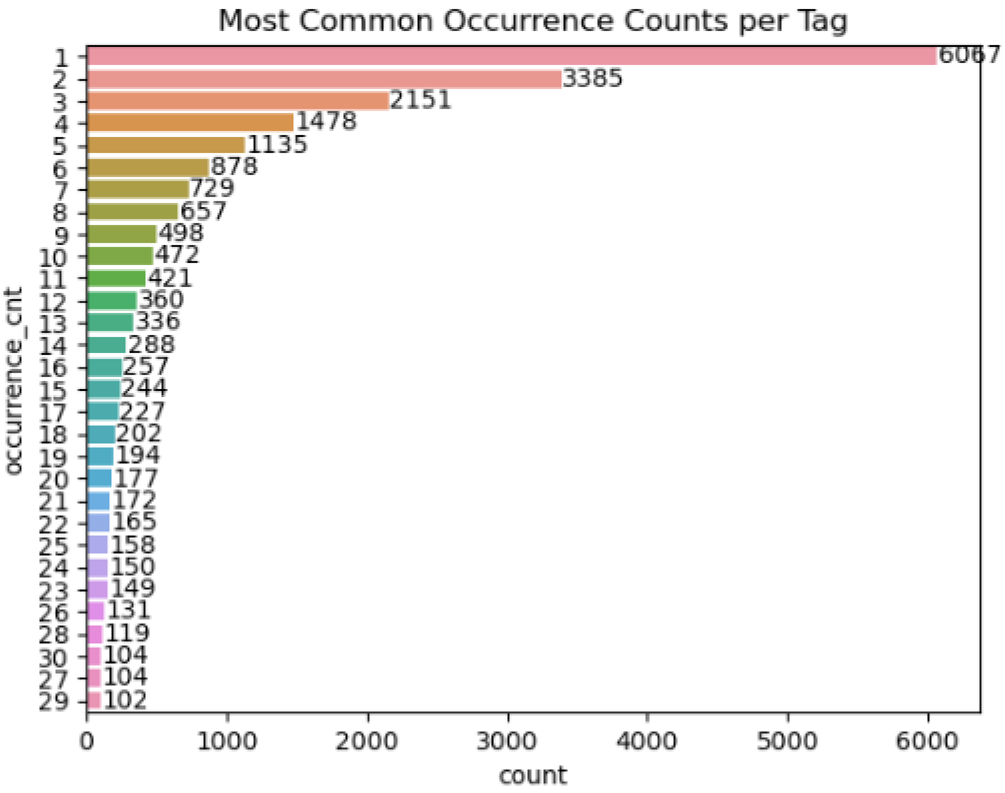


	mpaa_rating	G	NR	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
category	tag_name										
drugs	drugs_illegal	0.0	8.0	29.0	125.0	311.0	14.0	0.0	63.0	2.0	2.0
	drugs_implied	7.0	8.0	126.0	263.0	149.0	32.0	2.0	32.0	17.0	4.0
	drugs_legal	8.0	11.0	117.0	420.0	10.0	44.0	4.0	5.0	19.0	1.0
immodesty	immodesty	5.0	18.0	153.0	452.0	664.0	48.0	6.0	113.0	23.0	5.0
	nudity_art	2.0	2.0	36.0	140.0	233.0	15.0	0.0	43.0	7.0	0.0
	nudity_implied	2.0	4.0	45.0	159.0	276.0	10.0	1.0	40.0	2.0	3.0
	nudity_without_sex	2.0	11.0	67.0	229.0	497.0	26.0	0.0	82.0	4.0	2.0
language	blasphemy	7.0	21.0	167.0	494.0	748.0	55.0	7.0	128.0	26.0	5.0
	childish_language	10.0	9.0	183.0	212.0	353.0	21.0	6.0	51.0	27.0	8.0
	profanity	6.0	21.0	150.0	502.0	759.0	54.0	6.0	129.0	20.0	5.0
	racial_slurs	1.0	5.0	35.0	144.0	284.0	16.0	0.0	40.0	1.0	0.0
	sexual_reference	4.0	18.0	151.0	428.0	667.0	47.0	3.0	111.0	19.0	4.0
other	bodily_functions	6.0	15.0	130.0	276.0	482.0	36.0	3.0	80.0	13.0	7.0
	life_events	1.0	4.0	36.0	87.0	127.0	11.0	1.0	22.0	3.0	1.0
	medical_graphic	0.0	5.0	13.0	80.0	181.0	7.0	0.0	33.0	3.0	0.0
	medical_procedures	2.0	2.0	25.0	98.0	130.0	12.0	0.0	27.0	5.0	0.0
	objectionable	5.0	11.0	113.0	246.0	349.0	23.0	3.0	60.0	10.0	7.0
	vulgar_gestures	2.0	5.0	40.0	175.0	303.0	18.0	0.0	50.0	2.0	2.0
sexual	kissing_normal	5.0	11.0	116.0	361.0	476.0	37.0	4.0	76.0	22.0	3.0
	kissing_passion	4.0	13.0	57.0	246.0	380.0	30.0	3.0	53.0	11.0	1.0
	sex_implied	0.0	10.0	16.0	215.0	405.0	27.0	2.0	66.0	1.0	0.0
	sex_with_nudity	0.0	4.0	0.0	17.0	171.0	3.0	0.0	32.0	0.0	0.0
	sex_without_nudity	0.0	4.0	5.0	56.0	236.0	7.0	0.0	35.0	1.0	1.0
	sexual_assault	1.0	9.0	12.0	107.0	267.0	13.0	0.0	61.0	2.0	1.0
	sexually_suggestive	4.0	13.0	106.0	318.0	481.0	32.0	2.0	78.0	14.0	4.0
violence	disturbing_images	5.0	10.0	74.0	249.0	440.0	16.0	1.0	68.0	5.0	2.0
	gore	1.0	7.0	13.0	77.0	229.0	7.0	1.0	42.0	1.0	1.0
	graphic	6.0	15.0	90.0	363.0	630.0	29.0	2.0	101.0	9.0	3.0
	non_graphic	10.0	20.0	184.0	487.0	722.0	49.0	6.0	119.0	27.0	8.0
	violence_implied	8.0	4.0	131.0	47.0	61.0	6.0	3.0	8.0	13.0	6.0

All 6 of the language tag names have very high counts which contribute to that total being high. We also see for the drugs category, the drugs_legal tag_name is very low in the R rating compared to PG-13 and PG movies. The R rating does have a high count for drugs_illegal so these may be good indicators for our model when determining the ratings of these movies.

Occurrence Count

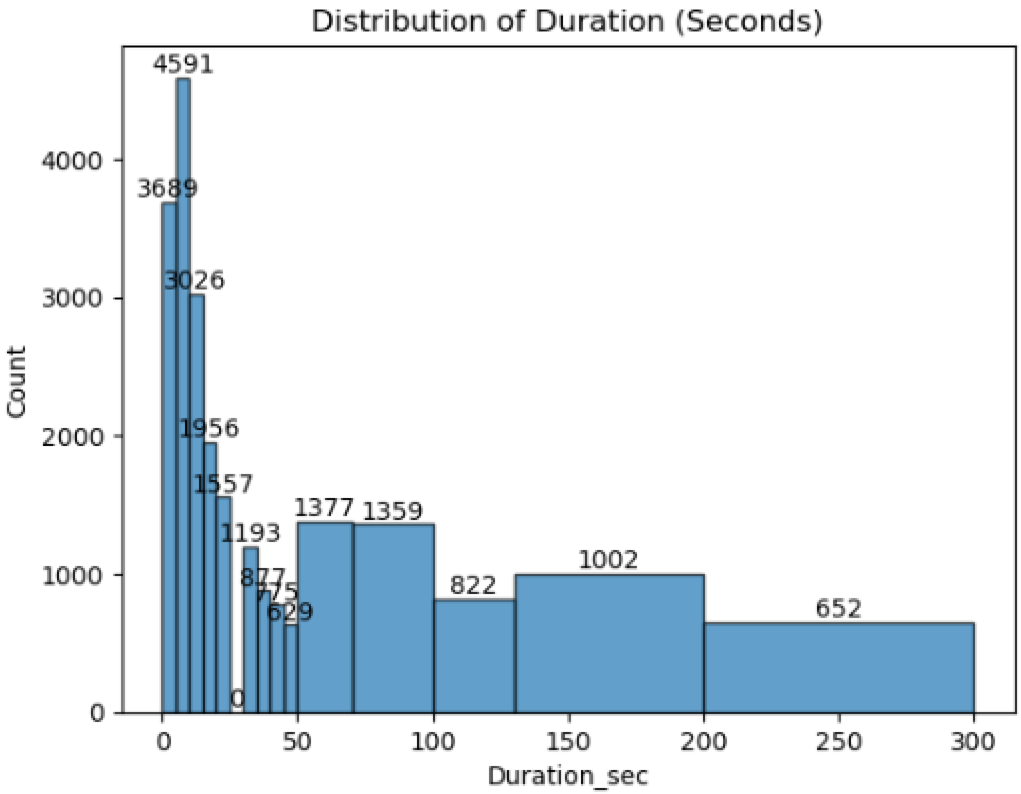
We see that most of the occurrences of these tags occur only a few times as the most common counts are 1-10.



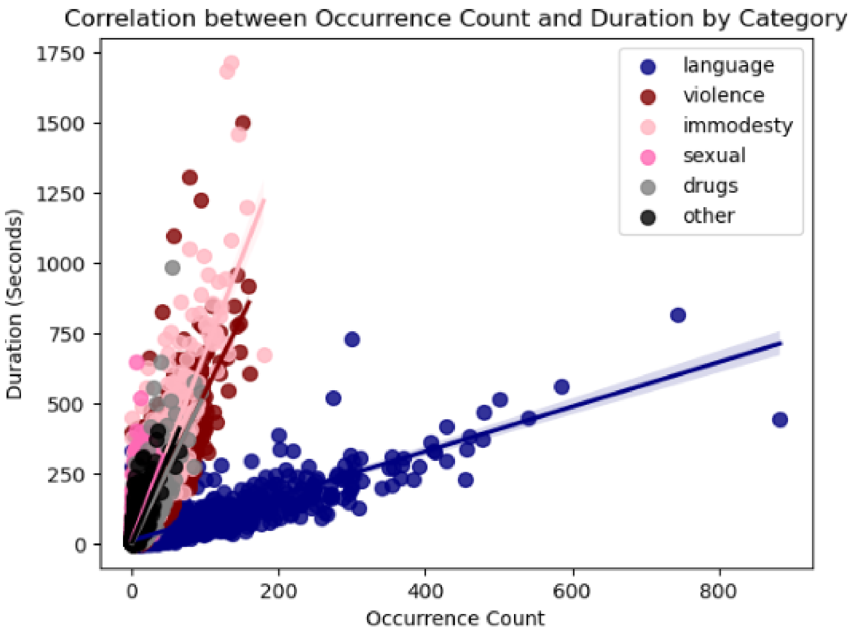
We will take a look at the Duration next and then we can see the relationship between the Occurrences and Duration for each category.

Duration (Sec)

For the duration of each tag, we plotted a histogram to see the distribution in seconds for the durations of these tags.



There is a big cluster of short durations and then as the durations get longer, there are less and less tags. Let's see what the relationship is between the Occurrence Count and Duration for each category.



We see that for the really high Occurrence counts, they are all Language categories. As the only category that directly corresponds to the audio component of a movie, this makes sense as each time bad language is used, it can count as an Occurrence. The durations tend to be shorter since dialogue usually has a lot of breaks between characters talking and different types of scenes.

Meanwhile, all the rest of the categories tend to have lower Occurrence counts, but longer durations. This makes sense as the rest of the categories would be more visual as opposed to audio, so those would stretch for entire scenes as opposed to a quick sound in dialogue.

Model Preprocessing

Before we started preprocessing our data, one thing we needed to do was remove any movies rated NR from the dataset. NR stands for Not Rated which means the movie was never submitted for a rating. As a result, the movie's content could range across any of the actual ratings and would only hurt our model when we trained it since those ratings are unknown. Once we removed those rows, we ended up with 1705 total movies in our dataset to work with.

Now that our dataset was cleaned, we next needed to merge our two dataframes so we could have one dataframe with all the data that we want our model to use. Since we have the imdb_id column that was the unique id for each movie in both our movies and movie_tags dataframes, we could combine them with a pivot table. We used the category and tag_name columns from movie_tags as that was a unique combination and then aggregated for each unique tag the occurrence count and duration in seconds for that tag. This allowed us to successfully merge the two dataframes into one where each row was a movie that included all the duration in seconds and occurrence counts for each tag a movie had.

Once we had merged our dataframes, we still needed to split out the studio column as there were multiple studios listed in that column that were separated by |'s. We created dummy columns for each studio where it had a 1 if the studio existed for that movie, and a 0 if it did not.

After that, we were left with all numeric columns except for our identifier columns (imdb_id and name), our target variable (mpaa_rating), and the studio column which we had just split out into dummy columns. We were able to drop those remaining non numeric rows to create our X variable, and created dummy columns for our target variable as our y which created 9 rows, one for each rating (we originally had 10 ratings but dropped the NR to get to 9).

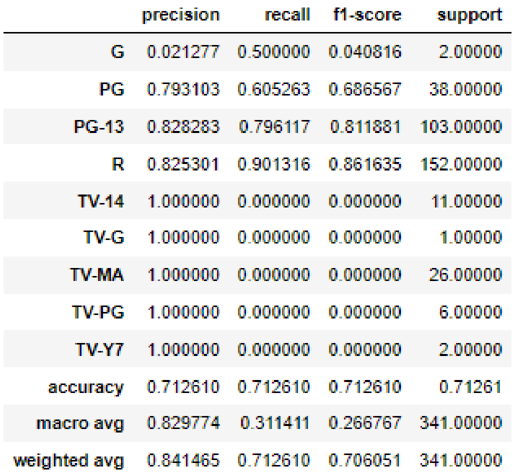
Once our X and y variables were created, we then split them into training and testing samples with a split of 80% for the training data and 20% for the testing. Then we scaled our X_train and X_test data using the X_Train data to normalize the data. This concluded our preprocessing step so we were now ready to create some models to run our data through.

Modeling

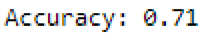
We created 4 different models to see how the data performed with each model. Our main method to compare them was the accuracy of the model along with the confusion matrix that was created for the classes.

Random Forest

After performing some hyperparameter tuning for our Random Forest model, we settled on 100 for n_estimators and 40 for max_depth. After running our model, we created the following classification report to review the performance of our model.

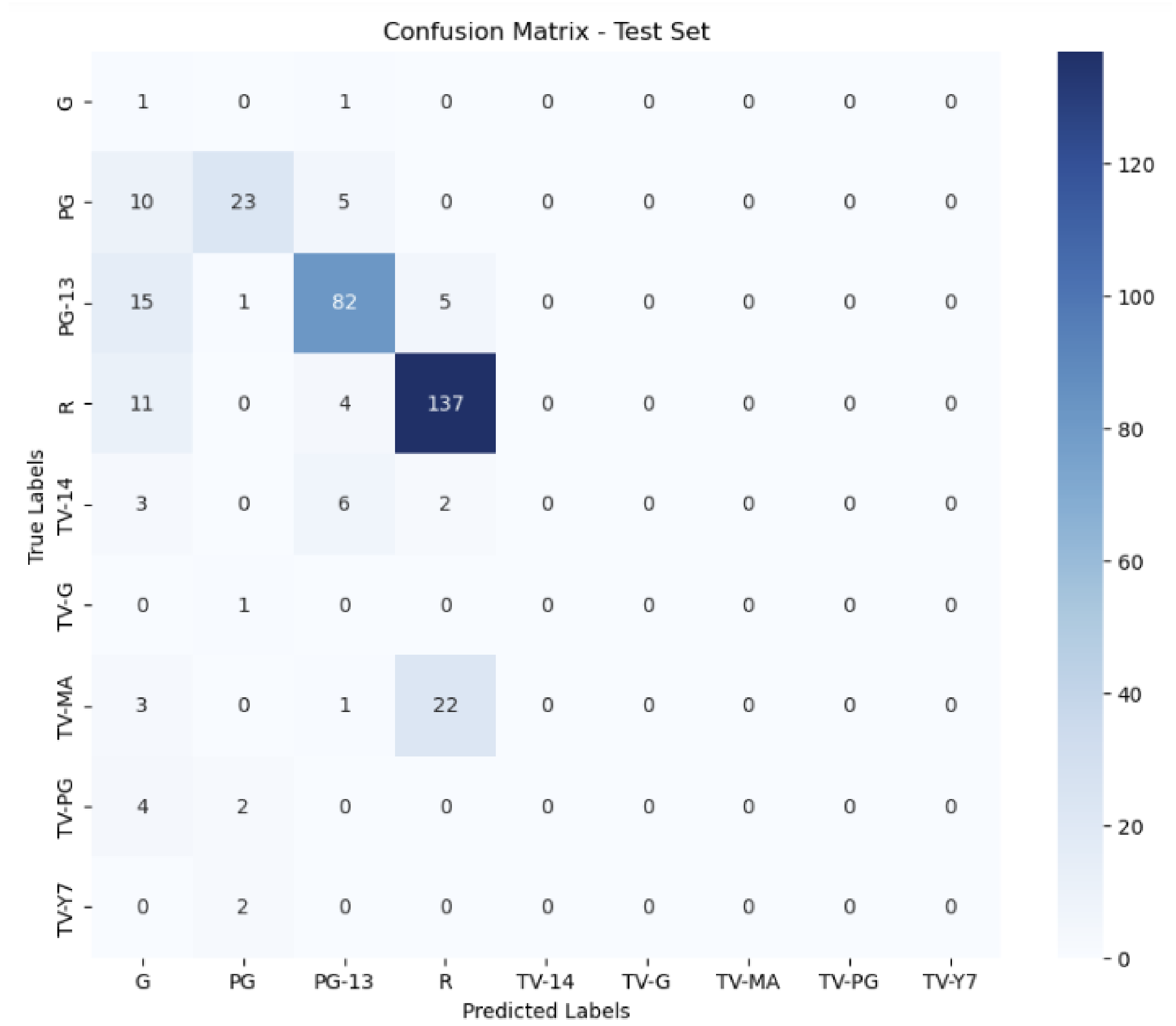


	precision	recall	f1-score	support
G	0.021277	0.500000	0.040816	2.00000
PG	0.793103	0.605263	0.686567	38.00000
PG-13	0.828283	0.796117	0.811881	103.00000
R	0.825301	0.901316	0.861635	152.00000
TV-14	1.000000	0.000000	0.000000	11.00000
TV-G	1.000000	0.000000	0.000000	1.00000
TV-MA	1.000000	0.000000	0.000000	26.00000
TV-PG	1.000000	0.000000	0.000000	6.00000
TV-Y7	1.000000	0.000000	0.000000	2.00000
accuracy	0.712610	0.712610	0.712610	0.71261
macro avg	0.829774	0.311411	0.266767	341.00000
weighted avg	0.841465	0.712610	0.706051	341.00000



```
Accuracy: 0.71
```

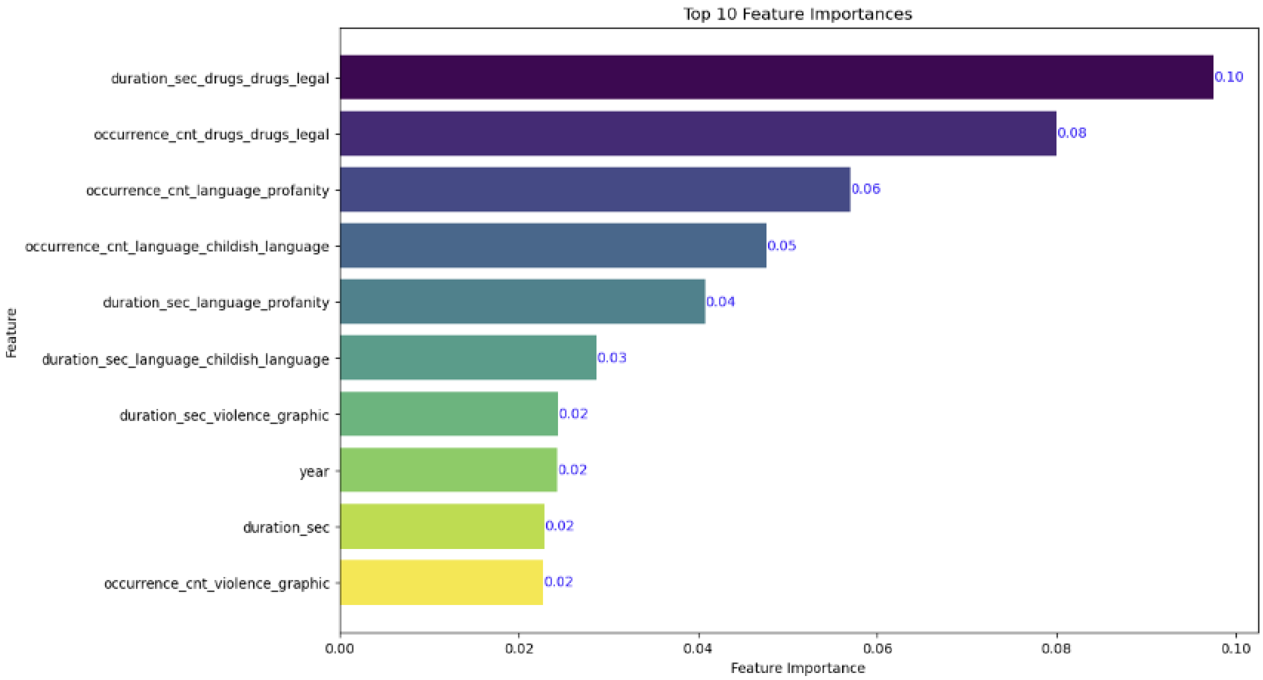
We see our accuracy for our model is around 71% which is pretty good. When plotting the confusion matrix, we can view which ones the model was correctly and incorrectly predicting.



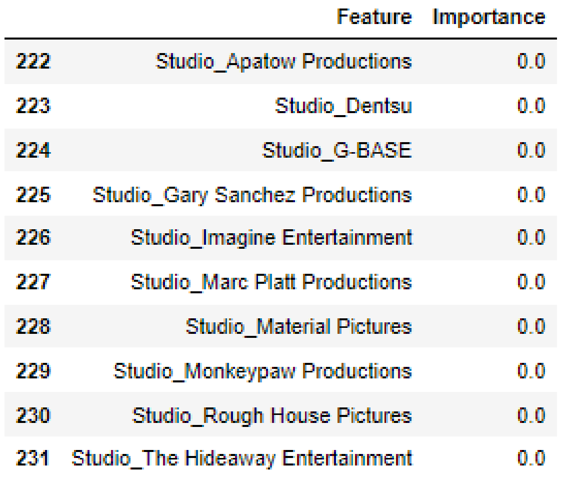
Confusion Matrix - Test Set

	G	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
G	1	0	1	0	0	0	0	0	0
PG	10	23	5	0	0	0	0	0	0
PG-13	15	1	82	5	0	0	0	0	0
R	11	0	4	137	0	0	0	0	0
TV-14	3	0	6	2	0	0	0	0	0
TV-G	0	1	0	0	0	0	0	0	0
TV-MA	3	0	1	22	0	0	0	0	0
TV-PG	4	2	0	0	0	0	0	0	0
TV-Y7	0	2	0	0	0	0	0	0	0

We see that our model only predicted on the non TV ratings. However, it was guessing G for all different types of movies including R rated movies which is not particularly close. Other than that though, our model performed very well, especially when predicting the R rating.



Looking at our top 10 most important features, we see the duration_sec and occurence_cnt for drugs_legal was the most important factor. That confirms what we were seeing in our Exploratory Data Analysis where most of the movies with that tag were actually PG-13 movies. The next four important features were all related to the language in the movie as well which seems to be a contributing factor to most of these ratings. We saw that language was our most common category, so it's good to see there is some importance with those tags.



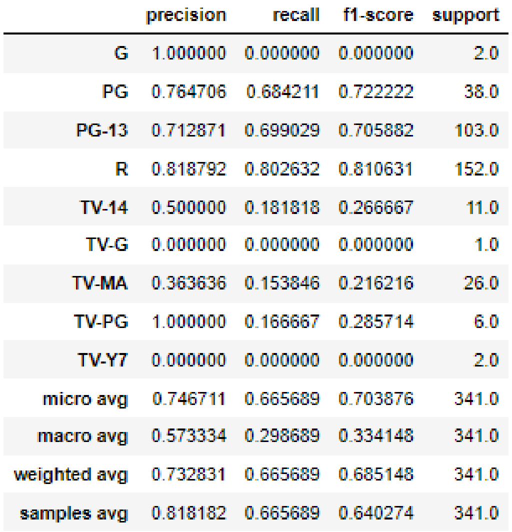
	Feature	Importance
222	Studio_Apatow Productions	0.0
223	Studio_Dentsu	0.0
224	Studio_G-BASE	0.0
225	Studio_Gary Sanchez Productions	0.0
226	Studio_Imagine Entertainment	0.0
227	Studio_Marc Platt Productions	0.0
228	Studio_Material Pictures	0.0
229	Studio_Monkeypaw Productions	0.0
230	Studio_Rough House Pictures	0.0
231	Studio_The Hideaway Entertainment	0.0

Looking at the least important features (figure above), we see that they are all studios. These look to be studios that didn't have a lot of movies in our data, so it makes sense they wouldn't have a lot of information with them.

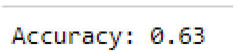
Overall, the Random Forest model performed very well on our dataset. Let's take a look at the next model.

Support Vector Machine

For our Support Vector Machine (SVM) model, we settled on a linear kernel and a C of 0.1 for our hyperparameters. Here is the classification report and the accuracy for this model.

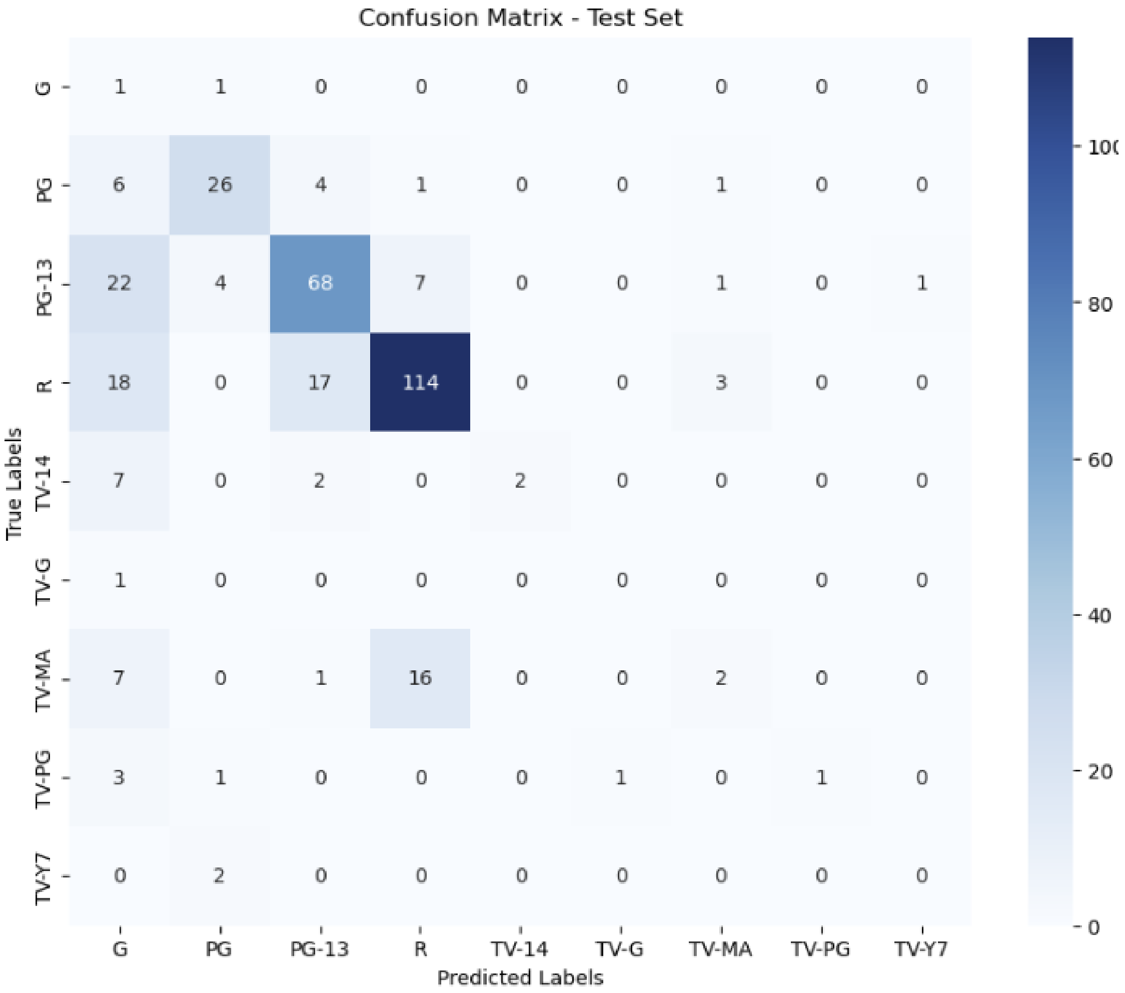


	precision	recall	f1-score	support
G	1.000000	0.000000	0.000000	2.0
PG	0.764706	0.684211	0.722222	38.0
PG-13	0.712871	0.699029	0.705882	103.0
R	0.818792	0.802632	0.810631	152.0
TV-14	0.500000	0.181818	0.266667	11.0
TV-G	0.000000	0.000000	0.000000	1.0
TV-MA	0.363636	0.153846	0.216216	26.0
TV-PG	1.000000	0.166667	0.285714	6.0
TV-Y7	0.000000	0.000000	0.000000	2.0
micro avg	0.746711	0.665689	0.703876	341.0
macro avg	0.573334	0.298689	0.334148	341.0
weighted avg	0.732831	0.665689	0.685148	341.0
samples avg	0.818182	0.665689	0.640274	341.0



```
Accuracy: 0.63
```

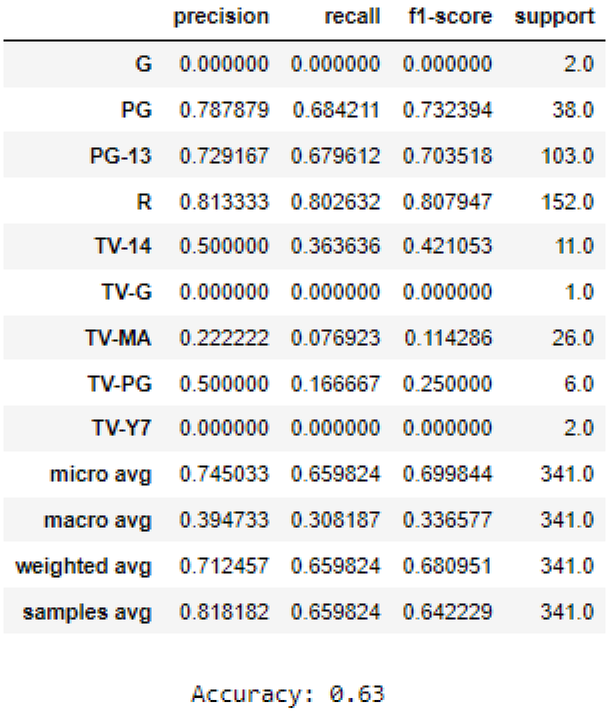
We see an accuracy of 0.63 which is not nearly as high as our Random Forest model.



For our confusion matrix, we see overall worse scores in every prediction category. The only category this model performed better in is for PG movies, but they were still very close when compared to the Random Forest model for predicting that rating. We also see the same thing where the G value is picked for a big range of different ratings. Overall, this model did not perform nearly as well as the Random Forest, so at the moment, our Random Forest model is our best model.

Logistic Regression

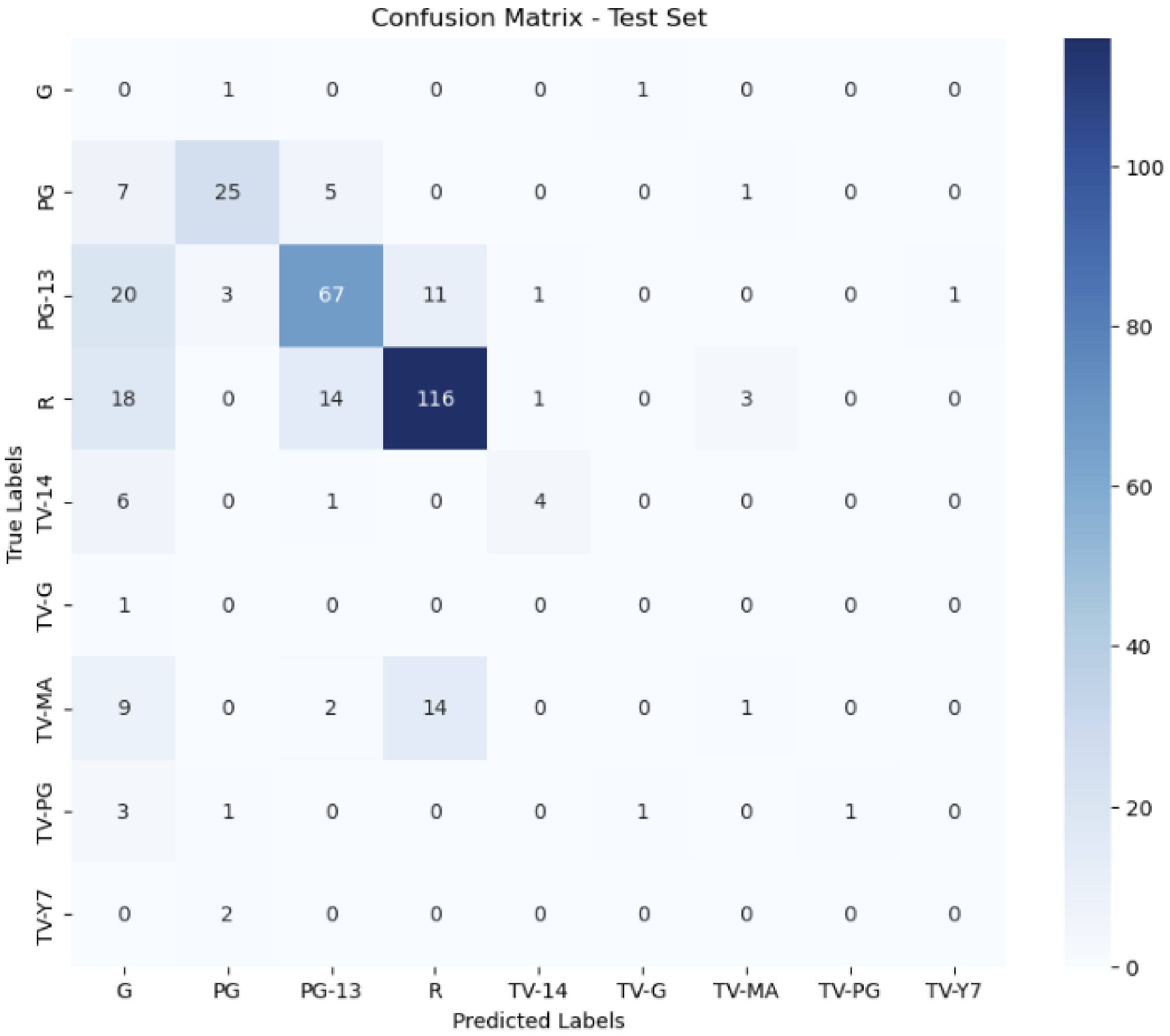
For our Logistic Regression model, we used a C value of 0.1 for our hyperparameter. Here is the classification report and the accuracy for this model.



	precision	recall	f1-score	support
G	0.000000	0.000000	0.000000	2.0
PG	0.787879	0.684211	0.732394	38.0
PG-13	0.729167	0.679612	0.703518	103.0
R	0.813333	0.802632	0.807947	152.0
TV-14	0.500000	0.363636	0.421053	11.0
TV-G	0.000000	0.000000	0.000000	1.0
TV-MA	0.222222	0.076923	0.114286	26.0
TV-PG	0.500000	0.166667	0.250000	6.0
TV-Y7	0.000000	0.000000	0.000000	2.0
micro avg	0.745033	0.659824	0.699844	341.0
macro avg	0.394733	0.308187	0.336577	341.0
weighted avg	0.712457	0.659824	0.680951	341.0
samples avg	0.818182	0.659824	0.642229	341.0

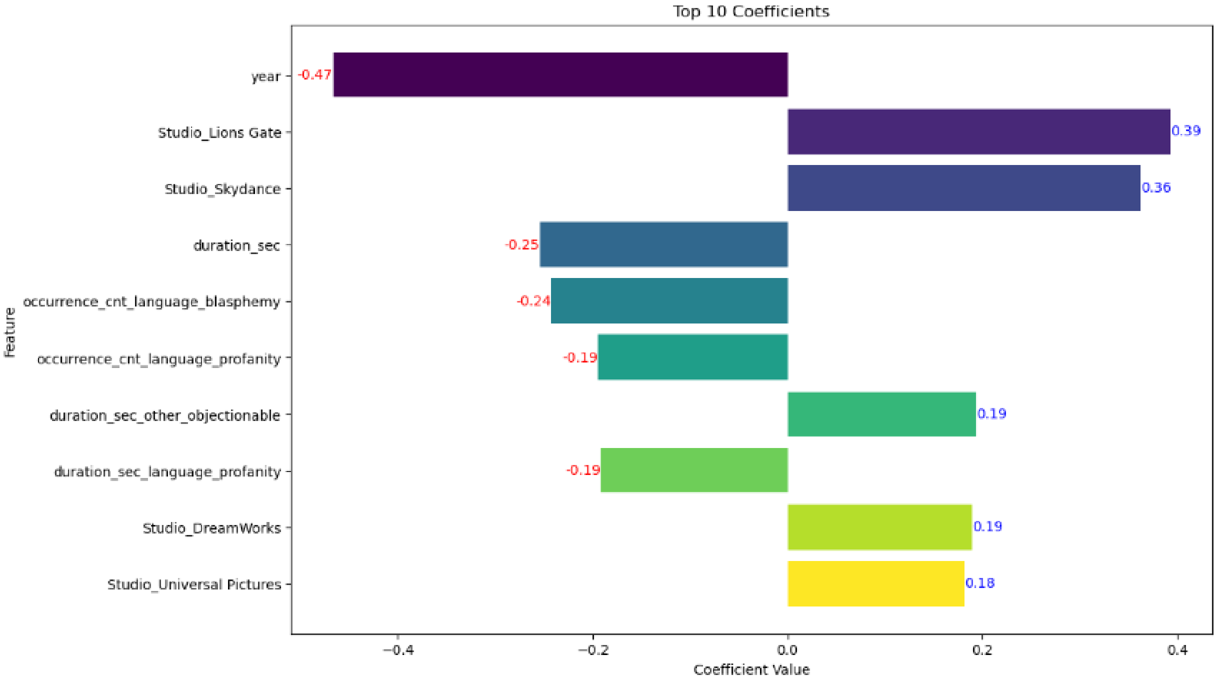
```
Accuracy: 0.63
```

We see the Logistic Regression model has an accuracy of 63% which is around the level of the SVM model.

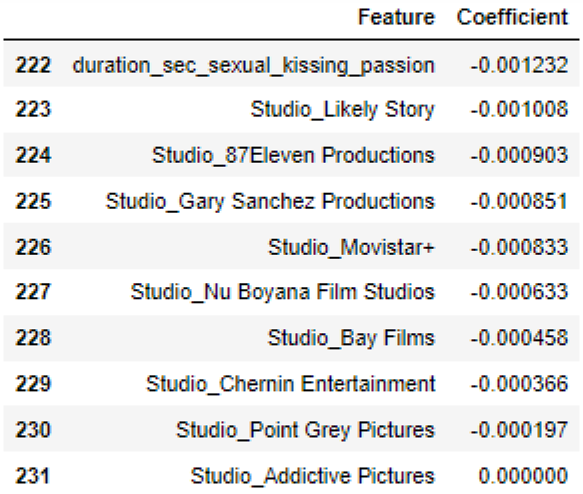


	G	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
G	0	1	0	0	0	1	0	0	0
PG	7	25	5	0	0	0	1	0	0
PG-13	20	3	67	11	1	0	0	0	1
R	18	0	14	116	1	0	3	0	0
TV-14	6	0	1	0	4	0	0	0	0
TV-G	1	0	0	0	0	0	0	0	0
TV-MA	9	0	2	14	0	0	1	0	0
TV-PG	3	1	0	0	0	1	0	1	0
TV-Y7	0	2	0	0	0	0	0	0	0

Looking at the confusion matrix, we see very similar results to our SVM model as well, with the same overall problems where it performs worse than the Random Forest model for essentially every category.



Looking at the largest coefficients the Logistic Regression model was using, we see that there are a couple of our top studios that made an impact on the rating the model predicted. We also see the year was an important factor as well which also was something our Random Forest model used as one of its most important features.



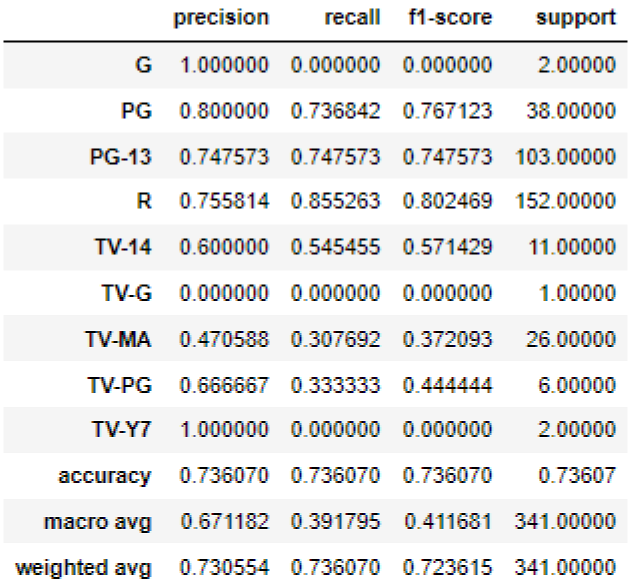
	Feature	Coefficient
222	duration_sec_sexual_kissing_passion	-0.001232
223	Studio_Likely Story	-0.001008
224	Studio_87Eleven Productions	-0.000903
225	Studio_Gary Sanchez Productions	-0.000851
226	Studio_Movistar+	-0.000833
227	Studio_Nu Boyana Film Studios	-0.000633
228	Studio_Bay Films	-0.000458
229	Studio_Chernin Entertainment	-0.000366
230	Studio_Point Grey Pictures	-0.000197
231	Studio_Addictive Pictures	0.000000

Looking at the least important coefficients (see figure above), we see that most of them are studios as well so the studios with low counts didn't make much of an impact. This is also something our Random Forest model picked up on. One interesting thing to note here is this model did not see the duration of passionate kissing as an important factor.

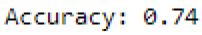
Overall, this model does not perform better than the Random Forest model, but we have one more model to review.

Deep Learning

For our Deep Learning model, we used one hidden layer with ReLu activation. We then compiled the model with the Adam optimizer and Categorical Crossentropy for the loss function. We then fit the model with 50 epochs, a batch size of 32, and a validation split of 0.2. Here is the classification report and the accuracy obtained for this model.

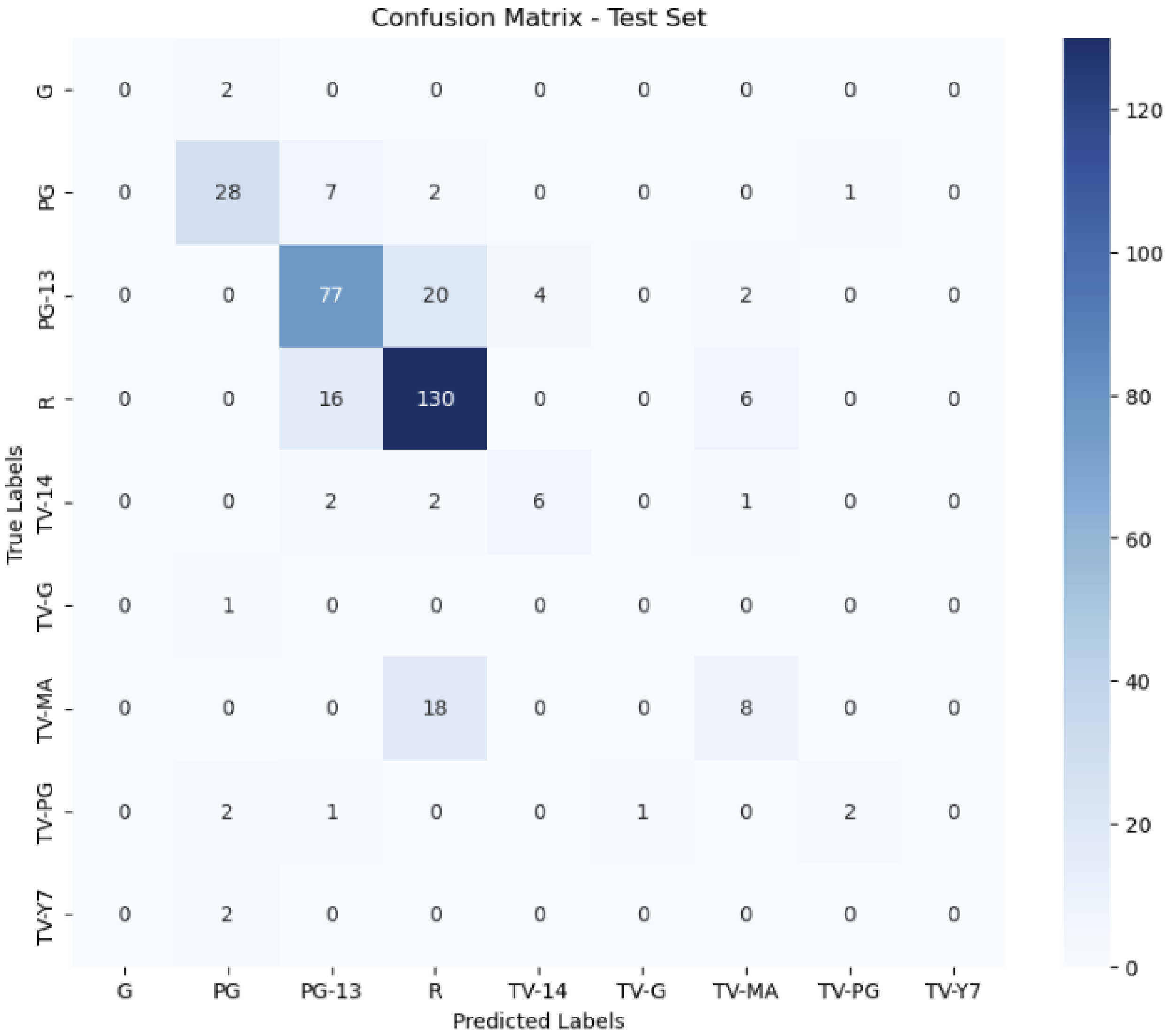


	precision	recall	f1-score	support
G	1.000000	0.000000	0.000000	2.00000
PG	0.800000	0.736842	0.767123	38.00000
PG-13	0.747573	0.747573	0.747573	103.00000
R	0.755814	0.855263	0.802469	152.00000
TV-14	0.600000	0.545455	0.571429	11.00000
TV-G	0.000000	0.000000	0.000000	1.00000
TV-MA	0.470588	0.307692	0.372093	26.00000
TV-PG	0.666667	0.333333	0.444444	6.00000
TV-Y7	1.000000	0.000000	0.000000	2.00000
accuracy	0.736070	0.736070	0.736070	0.73607
macro avg	0.671182	0.391795	0.411681	341.00000
weighted avg	0.730554	0.736070	0.723615	341.00000



```
Accuracy: 0.74
```

We see an accuracy of 74% which is slightly better than our Random Forest model.



Confusion Matrix - Test Set

True Labels \ Predicted Labels	G	PG	PG-13	R	TV-14	TV-G	TV-MA	TV-PG	TV-Y7
G	0	2	0	0	0	0	0	0	0
PG	0	28	7	2	0	0	0	1	0
PG-13	0	0	77	20	4	0	2	0	0
R	0	0	16	130	0	0	6	0	0
TV-14	0	0	2	2	6	0	1	0	0
TV-G	0	1	0	0	0	0	0	0	0
TV-MA	0	0	0	18	0	0	8	0	0
TV-PG	0	2	1	0	0	1	0	2	0
TV-Y7	0	2	0	0	0	0	0	0	0

Looking at our confusion matrix, we see much better predictions and higher quality predictions. We see that if a guess is incorrect, it's usually one rating away or the same rating on the TV side. We see similar high performances in all the categories, and for this model, we aren't predicting G for a diverse range of movies. This is what makes this model stand out from the rest in that it did not randomly guess G for a subset of movies. Because the accuracy is higher and the overall quality of the predictions are better, this looks to be the best model we have found.

Recommendation and Conclusion

From exploring our data, we found that most of our movies have an R rating. We also saw that most of the movies in the data are from 2010 or later which means the data is skewed toward newer movies. The average length of a movie in our data is between 1 hour and 40 minutes to 2 hours. 4 of the top 7 non-Other movie studios have something other than R as the most common rating. The number of studios does not appear to be related to the rating in any way. Language is the most common category and drugs is the least common category. The Drugs category is more common in PG-13 movies than in R movies. The drugs_legal tag_name is very low in R ratings when compared to PG and PG-13 movies. R has a high count for drugs_illegal.

Most tags only occur a small amount of times (count between 1-3) and have a short duration (under 50 seconds). The language category typically has high occurrences and short duration. The other non auditory categories tend to have lower occurrences but higher duration.

Some things we found in our models are studios with lower counts aren't a good predictor in the models, as well as when a model guesses incorrectly, it tends to be relatively close to the rating by one scale, or on the TV side of the rating. Exception to this was the G rating which was guessed poorly by all the models except deep learning.

After reviewing our four models, we found that the best model for predicting the rating of a movie given specific characteristics about the movie is our Deep Learning model. It has the best accuracy and highest quality predictions across all the different ratings. Using this model, we can predict the rating on new movies, unrated movies, as well as older movies to determine what the likely rating would be.

Future Scope of Work

Now that we have our model created, some next steps would be to feed in the Not Rated movies from our original data and see what it would predict. We can then do some research on those movies to see how accurate they would be based on the actual content of the movie. It also might make sense to gather some more data, specifically on the other ratings that we don’t have a lot of information on so we can help our model learn those ratings better and be able to predict them even better. We could also create further features for our model based on the names of the movies so the model could see if different types of language in the movie title can contribute to the rating as well.

Some other things we could do is further classify our predictions into what the recommended age group for a particular movie would be. This would allow people to make better decisions for the types of content they are watching and allowing their kids to watch. Lastly, we could develop an app to connect our input data for our model to an API for movie information so we can have an easy way to look up a particular movie and quickly determine everything we need to know about it.