WEEK 4:
PREDICT
IMDB SCORE HI/LO

6438169421 Pattaradanai Lakkananithiphan

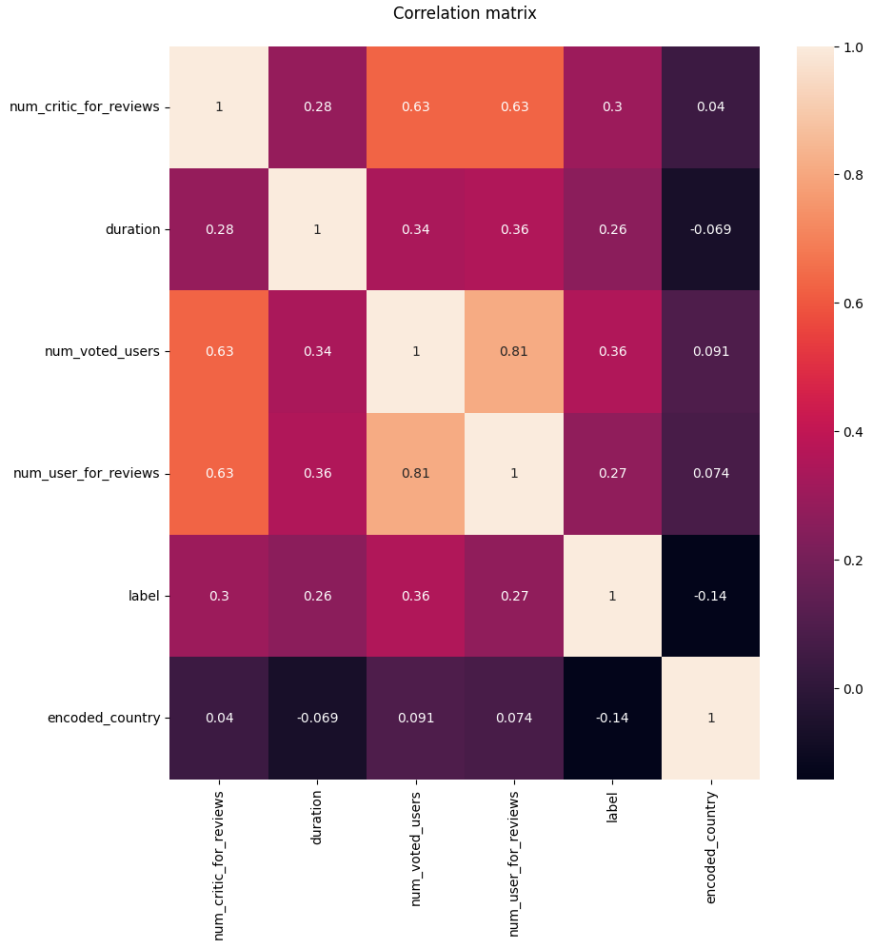
PREPROCESSING

- Drop the “imdb_score” column with values in the 40-60 percentile (20% of the data is dropped)
- Add a column of 0/1 according to the “imdb_score” with the condition for 1 being that it exceeds the median of the column -> name it “label”
- Drop the columns with a surplus amount of categories, unrelated to the problem, or low correlation with the output (I choose 0.25 as the correlation threshold)
- Encode the country column into 0/1 with the condition that if the film is from US it is 1 and this column is 0 elsewhere -> name it “encoded_country”
- Drop all NaN rows (Total of 54 rows were dropped)
- Remaining Features: "num_critic_for_reviews", "duration", "num_voted_users", "num_user_for_reviews", "encoded_country"

PREPROCESSING (CONT.)

- Scale the features using the StandardScaler
- Split the train/test sets using the ratio of 30:70

Final Correlation Matrix ->

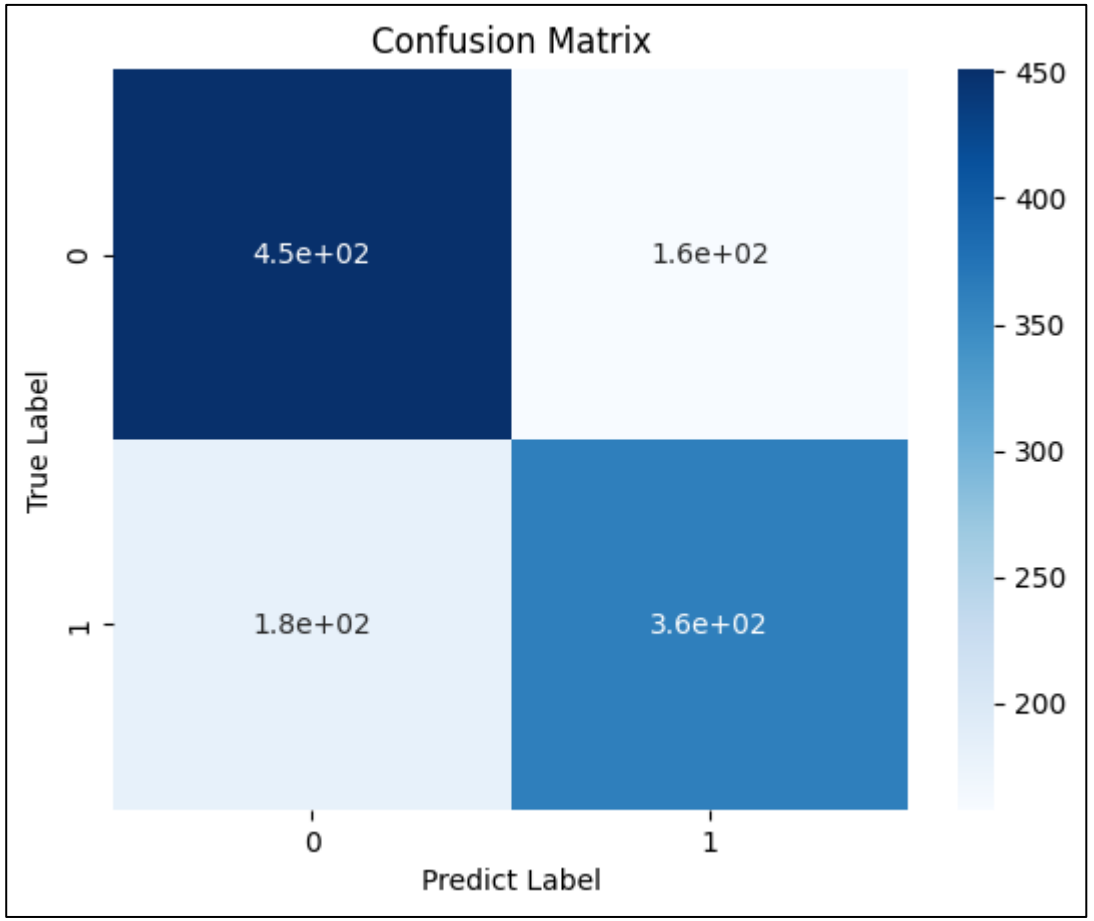


```
Accuracy on train:   0.706

Accuracy on test:  0.706
```

MODEL 1: LOGISTIC REGRESSION

```
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.74      0.73       609
           1       0.70      0.67      0.68       543

    accuracy                           0.71      1152
   macro avg       0.70      0.70      0.70      1152
weighted avg       0.71      0.71      0.71      1152
```

Features	Coef/Int
scaled(num_critic_for_reviews)	-0.106916
scaled(duration)	0.332843
scaled(num_voted_users)	2.591016
scaled(num_user_for_reviews)	-0.603571
scaled(encoded_country)	-1.051137
Intercept	1.219118

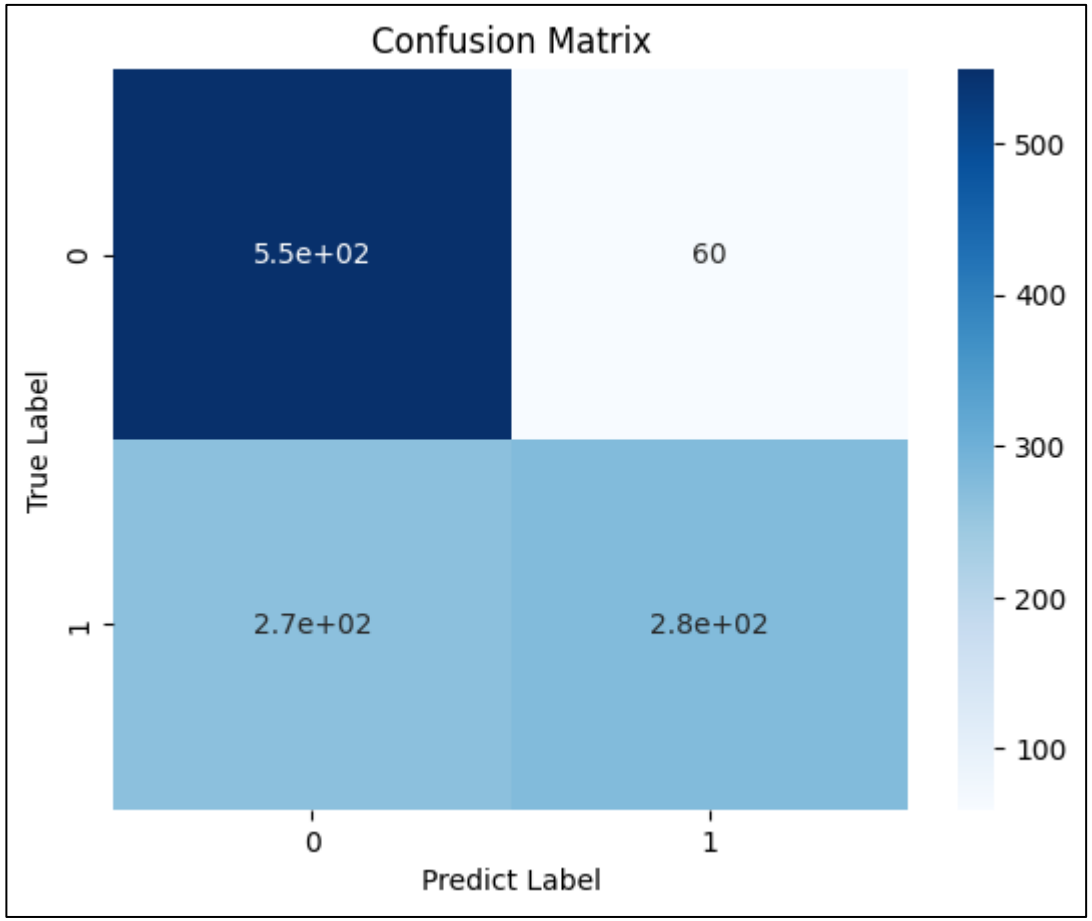


MODEL 2: GAUSSIAN NAÏVE BAYES

```
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.90      0.77       609
           1       0.82      0.51      0.63       543

    accuracy                           0.72      1152
   macro avg       0.75      0.71      0.70      1152
weighted avg       0.74      0.72      0.70      1152
```

This model only uses only numerical features. The generated “encoded_country” is not part of this model training data set

```
Accuracy on train:   0.68

Accuracy on test:  0.717
```



CONCLUSION

METRIC: ACCURACY

Since the categories of the label are evenly distributed and there are no major costs for false negatives or false positives

CHOOSE THE GAUSSIAN NB MODEL

The accuracy scores are pretty much equal as well as the other metrics we did not choose as our primary focus, but this model uses fewer features

GAUSSIAN NB: ACCURACY = 71.7%