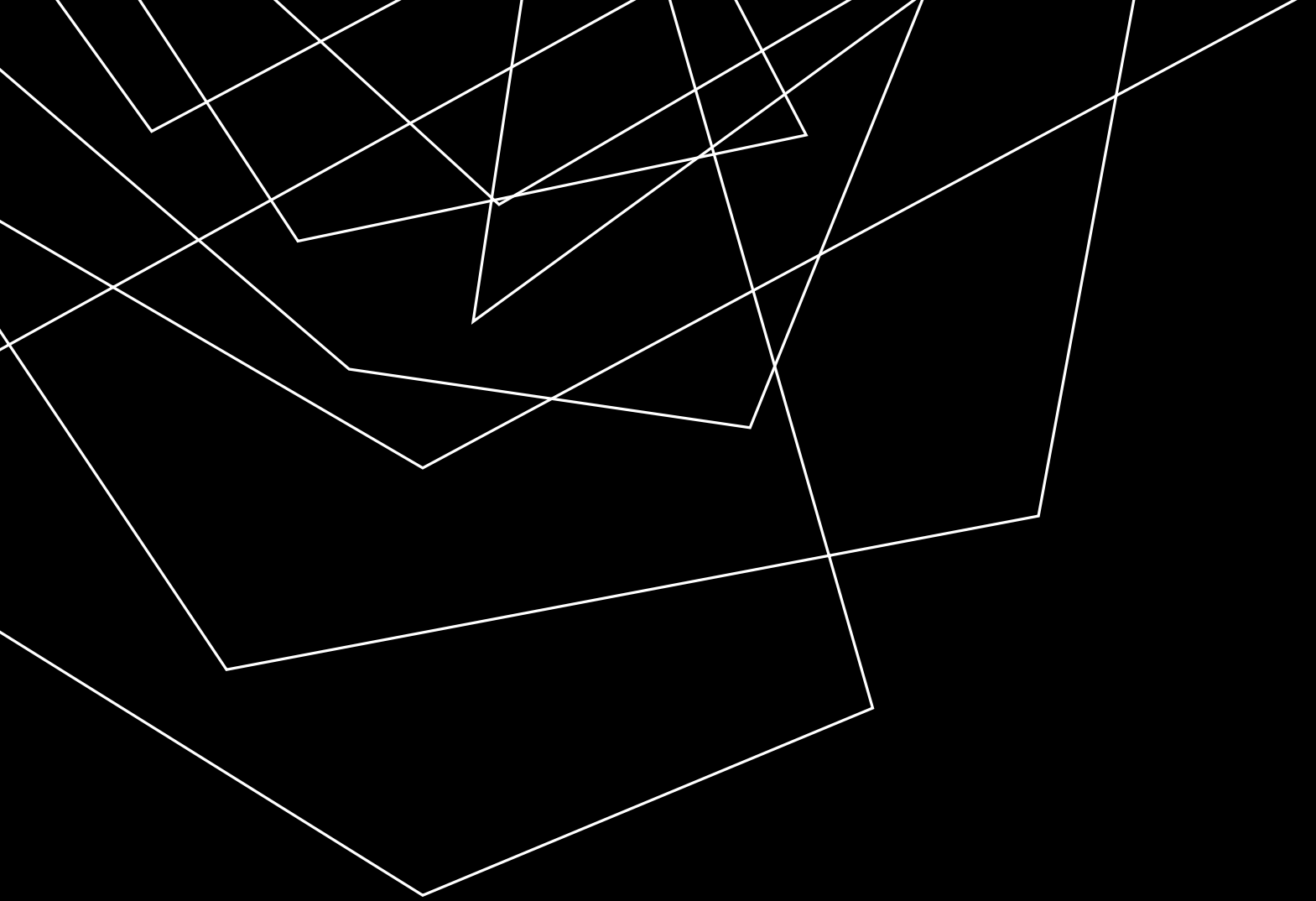


WEEK11
ENSEMBLE
6438169421
PATTARADANAI
LAKKANANITHIPHAN

PREPROCESSING

- Drop the rows with “imdb_score” column 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 lot of missing values
- Split the train and test set
- Drop the categorical columns as instructed
- Impute using median
- Scale the features
- **Select** rows with high correlation and low multicollinearity with the label

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   num_critic_for_reviews  2723 non-null   float64
 1   duration                2723 non-null   float64
 2   director_facebook_likes 2723 non-null   float64
 3   actor_3_facebook_likes  2723 non-null   float64
 4   actor_1_facebook_likes  2723 non-null   float64
 5   num_voted_users         2723 non-null   float64
 6   cast_total_facebook_likes 2723 non-null float64
 7   facenumber_in_poster    2723 non-null   float64
 8   num_user_for_reviews    2723 non-null   float64
 9   title_year              2723 non-null   float64
 10  actor_2_facebook_likes  2723 non-null   float64
 11  aspect_ratio            2723 non-null   float64
 12  movie_facebook_likes    2723 non-null   float64
```

X for training

And y for training

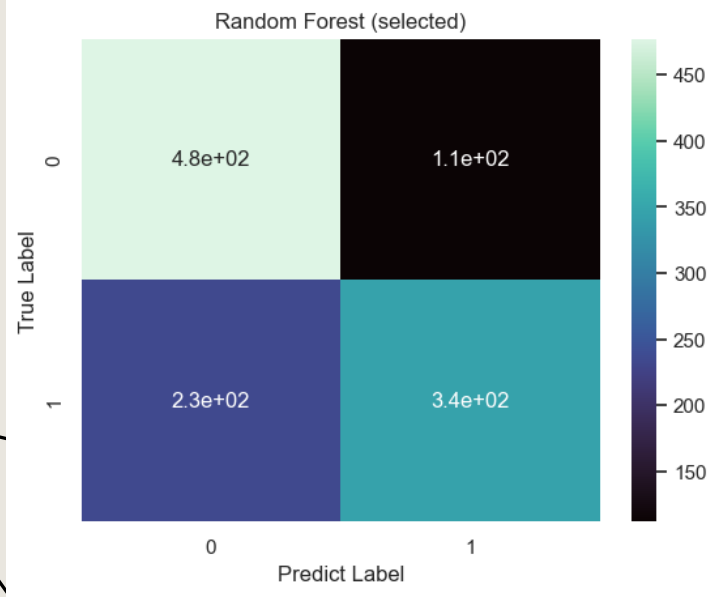
```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   label   2723 non-null   float64
```

Selected ->

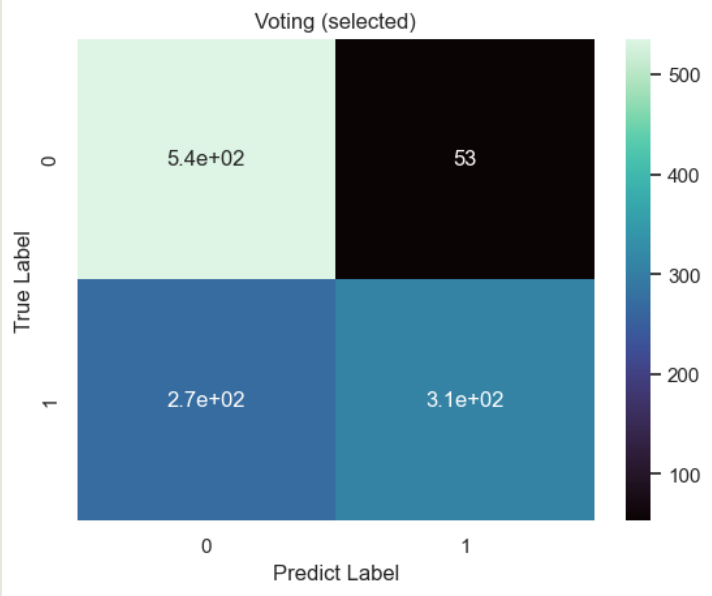
```
['num_critic_for_reviews',
 'duration',
 'num_voted_users',
 'title_year',
 'movie_facebook_likes']
```

THE MODELS (SELECTED FEATURES)

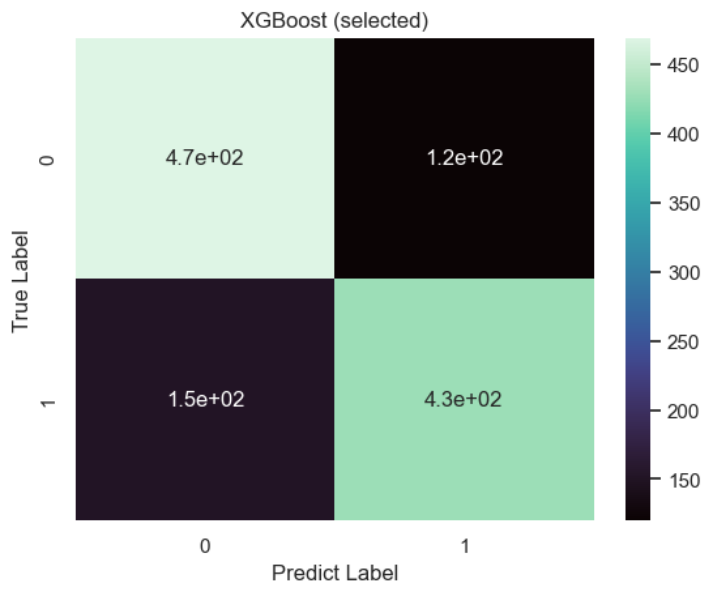
- Random Forest
- Max depth = 4
- Accuracy on train: 0.712
- Accuracy on test: 0.704



- Voting
- Lr,DT,KNN,SVC
- Accuracy on train: 0.726
- Accuracy on test: 0.723



- XGBoost
- Max depth = 4
- Accuracy on train: 0.916
- Accuracy on test: 0.768

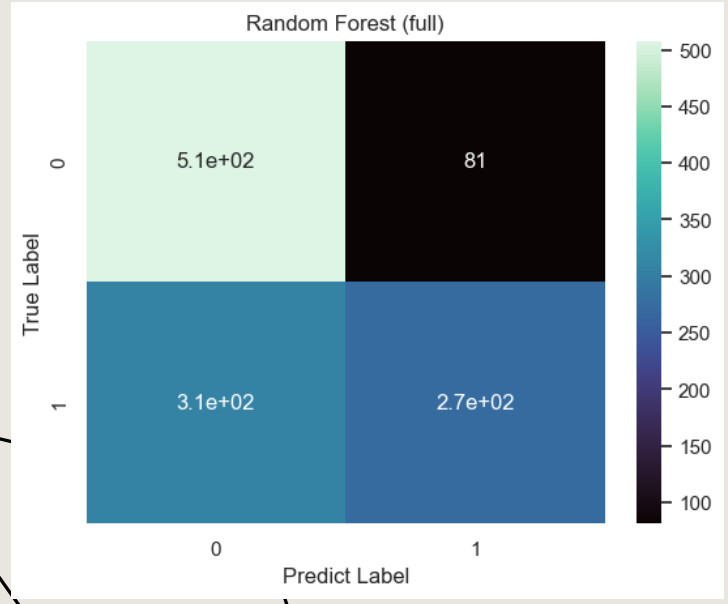


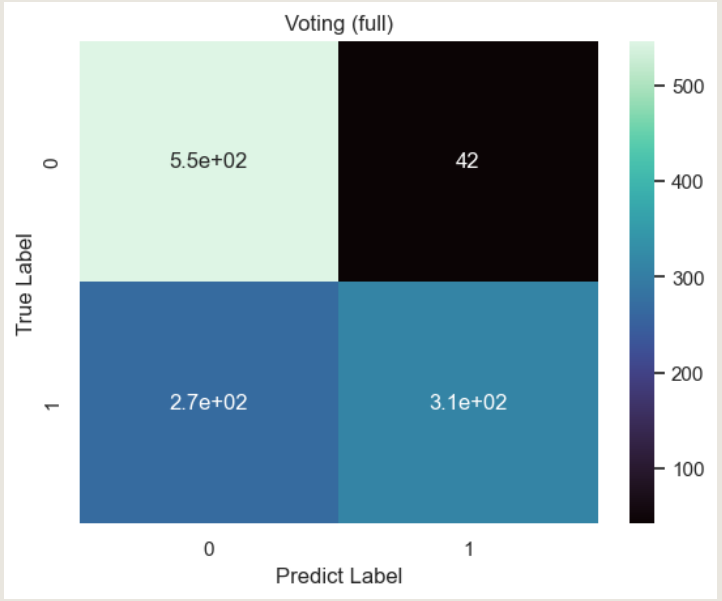
THE MODELS (ALL FEATURES)

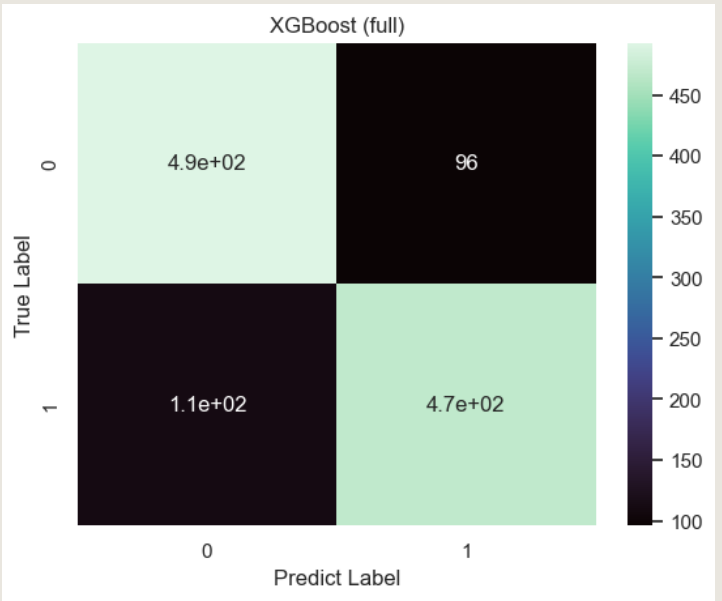
- Random Forest
- Max depth = 4
- Accuracy on train: 0.668
- Accuracy on test: 0.668

- Voting
- Lr,DT,KNN,SVC (Parameter from Gridsearch of previous weeks)
- Accuracy on train: 0.741
- Accuracy on test: 0.735

- XGBoost
- Max depth = 4
- Accuracy on train: 0.968
- Accuracy on test: 0.824







EVALUATION

- The models with more features perform better in comparison to the selected one when it comes to the XGBoost and Voting
- The XGBoost model has the best accuracy score (Although since the training accuracy is very high this could be overfitting)
- Choose **XGBoost (FULL) for its accuracy 82.4%**