WEEK10 ANN
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 categorical columns as instructed
- Drop all NaN rows of discrete numerical column and fill the rest with median
- **Select** rows with high correlation and low multicollinearity with the label
- Scale the features
- Split the train and test set

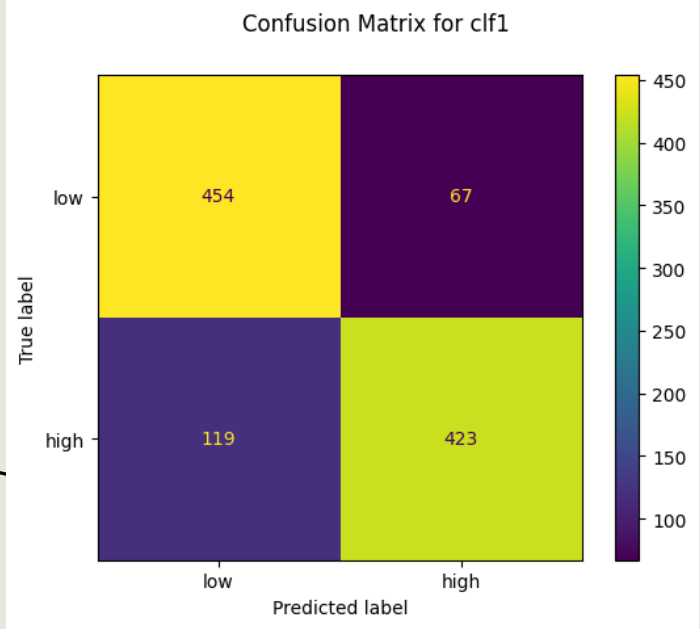
```
Data columns (total 15 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   num_critic_for_reviews     2479 non-null   float64
 1   duration                   2479 non-null   float64
 2   director_facebook_likes    2479 non-null   float64
 3   actor_3_facebook_likes     2479 non-null   float64
 4   actor_1_facebook_likes     2479 non-null   float64
 5   gross                      2479 non-null   float64
 6   num_voted_users            2479 non-null   float64
 7   cast_total_facebook_likes  2479 non-null   float64
 8   facenumber_in_poster       2479 non-null   float64
 9   num_user_for_reviews       2479 non-null   float64
 10  budget                     2479 non-null   float64
 11  title_year                 2479 non-null   float64
 12  actor_2_facebook_likes     2479 non-null   float64
 13  aspect_ratio               2479 non-null   float64
 14  movie_facebook_likes       2479 non-null   float64
```

Selected ->

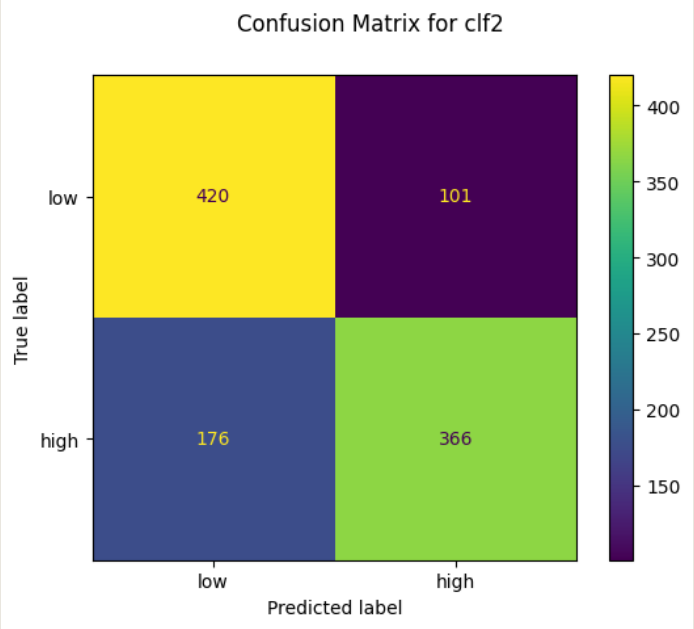
```
['num_critic_for_reviews',
 'duration',
 'num_voted_users',
 'title_year',
 'movie_facebook_likes']
```

THE MODELS

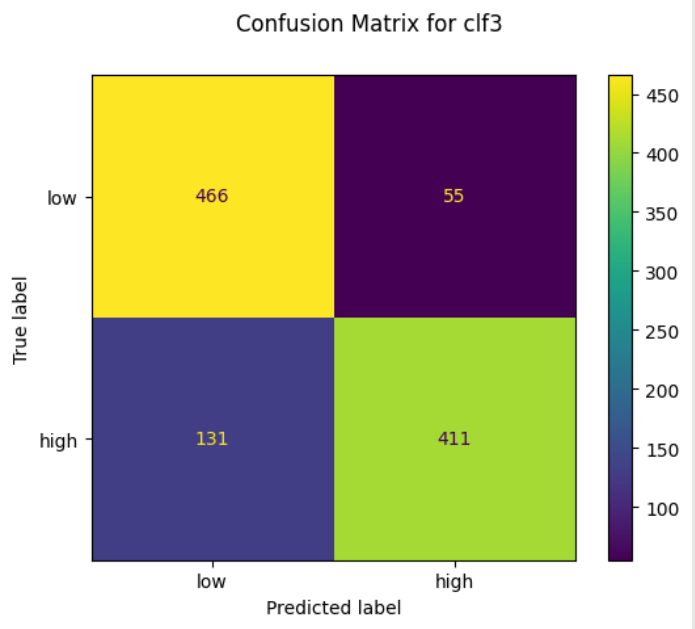
- All features
- (15,15) hidden layers
- Accuracy on train: 0.821
- Accuracy on test: 0.825



- Selected features
- (5,5) hidden layers
- Accuracy on train: 0.742
- Accuracy on test: 0.739



- All features + GridSearch
- (45,45) hidden layers
- Accuracy on train: 0.843
- Accuracy on test: 0.825



EVALUATION

- The models with more features perform better in comparison to the selected one
- The Gridsearched model does not perform much better than the original model (could be overfitting) while being much more complicated
- Choose **Model 1 : all features**