IMDB Movie Rating Predication

Submitted By – Pradnya Gorhe

Rang Technology analysis of movie dataset

Problem Statement

In order to predict movie rating these features are taken from movie data which consist imdb_score, gross , budget , rating ,facebook likes, director , genres , number of votes , critic reviews. IMDB movie reviews are generally aggregated score of user reviews that builds and becomes more accurate over period of time as there are less number of rating for newly released movies. In order to better predication, I have run machine learning models.

Dataset

We data of about 5000 movies with 28 columns which include: color, director_name', 'num_critic_for_reviews', 'duration', 'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name', 'actor_1_facebook_likes', 'gross', 'genres', 'actor_1_name', 'movie_title', 'num_voted_users', 'cast_total_facebook_likes', 'actor_3_name', 'facenumber_in_poster', 'plot_keywords', 'movie_imdb_link', 'num_user_for_reviews', 'language', 'country', 'content_rating', 'budget', 'title_year', 'actor_2_facebook_likes', 'imdb_score', 'aspect_ratio', 'movie_facebook_likes',gross"

Feature Processing

Columns such as Color, Language and Country were removed as they did not capture much variance. See EDA plots for detailed explanations. Missing numerical values and duplicates values has been were removed. Movie_imdb_link columns removed which is not require for analysis. Rename few columns and identified correlation with Pearson method.

Modeling Technique

In order to build machine learning models I have consider numerical variables where I have used Random Forest , KNN Classifier and decision tree models. Split the data into training and testing where training is 80% and testing is 20% The confusion metrics are used to optimize this model is F1 Score , Recall , Precision in order to evaluate decision tree and random forest model.

How Confusion Metrics used to Evaluate Model Performance:

Accuracy - Accuracy is the most important measure in any model or metric which simply a ratio of correctly predicted observation to the total observations.
One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

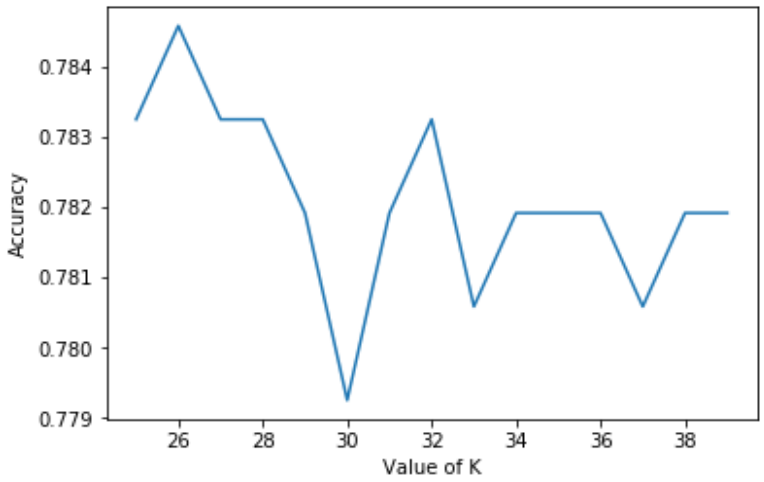
Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.
Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

KNN CLASSIFER

```
Maximum Accuracy for KNN: 0.784574
```



Decision Tree Model

```
             precision    recall  f1-score   support

          A       0.49      0.52      0.50       114
          B       0.82      0.82      0.82       568
          C       0.27      0.25      0.26        63
          D       0.14      0.14      0.14         7

avg / total       0.72      0.72      0.72       752

[[ 59  52   3   0]
 [ 59 465  38   6]
 [  2  45  16   0]
 [  0   4   2   1]]
Accuracy of prediction: 0.697
```

For decision tree we got 0.72 precision and recall which is pretty good. Recall should be above 0.5 with decision tree its giving 0.72 which is pretty good for this model. Accuracy is low as compared to other models.

Random Forest

```
             precision    recall  f1-score   support

          A       0.49      0.52      0.50       114
          B       0.82      0.82      0.82       568
          C       0.27      0.25      0.26        63
          D       0.14      0.14      0.14         7

avg / total       0.72      0.72      0.72       752

[[ 50  64   0   0]
 [ 17 551   0   0]
 [  0  58   5   0]
 [  0   7   0   0]]
Accuracy of prediction: 0.799
```

For decision tree we got 0.72 precision and recall which is pretty good. Recall should be above 0.5 with decision tree its giving 0.72 which is pretty good for this model. Accuracy is low as compared to other models. The highest accuracy with random forest model is .79% which pretty good as compare to KNN and Decision Tree.

Final Conclusions:

• Get More data Sources to better the explain the target variable get higher accuracy

• Try more Nonlinear model

• Random Forest would be the model of choice for given data set but we can also consider KNN classifier as its giving 78% accuracy which is pretty good and we can improve model performance.