Analysis of Movie Dataset

Introduction

The task at hand want to analyze, visualize, and make prediction of the Movie Dataset. The dataset consists of 28 rows of data for each movie present. Some features might not be co-related or not required to perform analysis. We will go ahead by picking out the necessary features for model training.

Data Pre-Processing

- Initial Analysis is done by using the head() and describe() methods from the pandas library.
- Later we proceed by plotting histograms of the features, have a look at the range of values and then go on to the co-relation matrix.
- We check for "NA" values among the features and set them to mean of the feature.
- I have defined a helper function to set a user defined threshold and filter the features based on the co-relation value from the co-relation matrix.
- Now, we have our dataset filtered with the necessary features required for model fitting.

Feature Engineering

- We need to work with the features that have object datatype. We can convert these to numerical values using LabelEncoder().
- We add the label to the dataframe consisting of object features and check the updated co-relation matrix. Then add the required features to our final dataframe.
- We now develop our X(features) and Y(label)
- As we now have all the required parameters for model training, the last step is to scale these values. We can use the MinMaxScaler() method to perform this operation.
- Now, we split our data into training and test set and it's ready for model training.

Model Training

- The task at hand is regression as the target variable is continuous.
- We can check which model will suit the data best using the lazypredict package. This package performs a check on all the model and provides us a list of best fit models.
- From the results obtained, I have decided to check the performance for RandomForestRegressor, LGBMRegressor, GradientBoostingRegressor and NN with 3 layers.

Results

Model	Training MAE	Test MAE
RandomForestRegressor	0.744907102	0.74544494
LGBMRegressor	0.498070598	0.581876975
GradientBoostingRegressor	0.562650404	0.601599649
Neural Network	0.649151898	0.671434384

As seen from the results we can proceed with LGBMRegressor for production deployment among the models as it has low mean absolute error and is stable on Test Set as well.