Assignment 3 report

Part 1:

Performance of our model:

After cleaning the data and representing it in a way that our regression model can input , I achieved a **mean squared error of 6565598521735442.0** and a **Pearson correlation coefficient of 0.44,** these metrics were used to evaluate the performance of the test set. The same metrics were used to evaluate the performance on the training set, to see how well our model is performing on the train set. Different features were included and excluded on trial and intuitive methods to reach a point where our model did not have a problem overfitting or underfitting.parameters such as number of top actors, genres ,keywords, directors, producers were also tuned to fit the model. Different regression models were also used , and found out random forest regression was the best performing model.

Problems faced:

Data such as title , website link etc were clearly useless for training, so we removed them in the first place.

I cannot obviously input Json columns to machine learning algorithms, hence those json were converted into list and then only the top occuring members in the list were represented using one hot encoding , and only the top colouring keywords were considered, this will cause missing out on data andone hot encoding will also increase the dimensionality of our input.

Release date was decomposed into year and month columns and year .

Some of the revenue columns had a revenue below 200 which is highly unlikely or it must be represented in millions unit , hence i removed rows with very low revenue for training, Some of the keywords and other information where missing from the dataset, which will cause our model to not perform well, so i removed those rows before training our model.

To scale the input , I had to normalize our input to the same range.

Part 2:

Performance of the model:

After cleaning our data and representing it in a way that our classifier model can input , I achieved a **precision:0.68 ,recall:0.66 and accuracy:0.73**
These metrics were used to evaluate the performance of the test set. The same metrics were . Different features were included and excluded on trial and intuitive methods to reach a point where our model did not have a problem overfitting or underfitting.parameters such as number of top actors, genres ,keywords, directors, producers were also tuned to fit the model. Different classifier models were also used , and I found out that knn classifier was the best performing model.

Problems faced:

Data such as title , website link etc were clearly useless for training, so we removed them in the first place.

I cannot obviously input Json columns to machine learning algorithms, hence those json were converted into list and then only the top occuring members in the list were represented using one hot encoding , and only the top colouring keywords were considered, these will cause missing out on data , this will also increase the dimensionality of our input.

And most of the cleansing is similar to the cleansing done in the first part.

While using knn classifier , i had to tune the hyperparameter number of neighbours considered for classification , on further investigation **gradient boosting classifie**r performed better than knn.