# Box Office Revenue

The main steps which I have used to solve this problem are:-

1. Loading the libraries and datasets
2. Analyzing train and test datasets
3. Exploratory Data Analysis (with visualization)
4. Json format columns to dictionary format
5. Feature Engineering
6. Modelling
7. Prediction

Here is the short description of each of the step which is enough to understand my code for the problem-

1. **Loading the libraries and datasets:-**

This step includes importing the libraries which I’ve used for the data analysis and the training and testing datasets.

1. **Analyzing train and test datasets:-**

In this step, basically I’ve found the type of features which are given in the datasets and the structure of dataset. The main point in this is to deal with the null values.

1. **Exploratory Data Analysis (EDA):-**

Exploratory data analysis (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

In this step, the relation of revenue with the other features has been seen.

1. **Json format columns to dictionary format:-**

There are some columns in the datasets which are in Json format. We can’t train our model using this format. So we have to modify our datasets. All this process is done in this step which gave some new columns and lastly I dropped the columns which were originally in Json format.

1. **Feature Engineering:-**

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Feature engineering was done in this step to transform raw datasets into a more informative datasets.

1. **Modelling:-**

The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. This whole process is called as Modelling.

Here I’ve used boosting algorithms to train the model as the number of features is too much.

I’ve tested the following three algorithms on the training dataset-

1. LightGBM
2. xgboost
3. The ensemble algorithm of the above two

From the above three algorithms, I got to know that the accuracies of xgboost and LightGBM are almost same but the time in training is very less for lgb (LightGBM).

1. **Prediction:-**

As the result of above step, I’ve used lgb algorithm for predicting the values of revenue for the test dataset. Final result is saved in the attached csv file ‘submission\_lgb’.