EXECUTIVE SUMMARY:

The primary aim of the project is to build a machine learning model that will predict the revenue of a movie given such features as cast, crew, keywords, budget, release dates, languages, production companies, and countries. 3000 movies data collected from The Movie Database (TMDB) as part of a kaggle.com Box Office Prediction Competition.

Hypothesis Generation is the first step involved in a building a machine learning model. This involves analyzing the problem with its possible factors that can impact the outcome.

Exploratory Data Analysis is done to know deeply about the data. It involves studying each features its correlation to the outcome, how its correlated with other features, finding the missing values and outliers, fix the issues in data and creating new variables. Libraries like ‘pandas profiling’ makes the EDA process very easy and it saves a lot of time. It gives an insight to do the further procedures for the project.

Data Cleaning is the lengthiest process in any machine learning problem. Its very unlikely to get a clean data for analysis. Raw dataset contains outliers, missing values, invalid string data, data having high cardinality and so on. Using proper library like ‘re’, ‘json’ etc. makes the cleaning easier.

Feature Engineering is the process of manipulating the features to get the best result. The best features give the best model.

The final step is building model. Three models where deployed

1. Linear Regression
2. Random Forest Regressor
3. Light BGM.

Comparing the scores Light BGM performed the best.