***Business Objective***

In today’s world movies make an estimated $42 billion, the film industry is getting more popular than ever before. The aim of the project was to forecast the overall worldwide box office collection of a movie, and the influence of factors like director, actor, budget on the movie’s revenue. Given, the metadata on over 7000 films from The Movie Database. Data points include cast, crew, plot, keywords, budget, posters, release dates, languages, production companies, and countries.

***Methodology***

**Data Exploration**

* 80% of the fields were text columns among that majority being in JSON format.
* More than 50% null values in columns like ‘id’, ‘belongs to collection’, ‘homepage’, ‘tagline’, ‘imdb id’, ‘poster path’. Therefore, dropping these columns.

**Data Preprocessing**

* Python library “Abstract Syntax Trees” was used to extract useful data from JSON formatted columns.
* To condense multilabel columns like cast, crew, keywords, spoken languages, production countries, production companies, genre, a list of frequent categories of a column was created and mapped to each movie entry.

**Feature Engineering**

* MultilabelBinarizer was used to create dummy variables for multilabel columns
* TFIDF was formulated for overview column which contained summary of a movie.
* OneHotEncoding of release month.

**Modelling**

* Adaboost, Random forest regressor and xgboost were ran.
* RMSE was the model selection criteria with Random forest regressor giving the best result (RMSE = 1.58).

**Result and Visualization**

* Predicted Box office busters - Avengers, The Hobbit, The Dark Knight, John Carter.
* Visualized feature importance bar graph with budget being the highest contributor to revenue.