# Predicting Movie Success using Machine Learning Algorithms

DATS6103 Project Proposal

Amna Gul
Madhuri Yadav
Hemanth Koganti

## Introduction:

Watching movies is a major source of entertainment for most of us. Film industry is one of the top grossing industries in the world. Millions of dollars are invested in making of movies expecting high margin of profit. Some of the top production houses are 20th Century Fox, Metro Goldwyn Mayer, Warner Bros etc. These production houses make movies of different genres which involve different cast and crew. Irrespective of budget, movies can earn high profit, or make bare minimum/loss. The purpose of the project is to classify a movie into one of the following categories : Blockbuster or Flop, based on various features from our dataset.

## Dataset, Modelling, Performance Evaluation Metrics:

For analysis purposes we are going to work on an ensemble of data from Kaggle (source of origin being TMDB and GroupLens). Here is the link for it. This dataset involves a collection of metadata for over 45,000 movies released on or before July 2017. Browsing through the csv files we found some anomalies, for example, there are a lot of missing values. Some of the columns e.g. “Budget” and “Revenue” contain zeros instead of an actual figure. Some columns need to be dropped because they are irrelevant etc. Data is distributed in separate csv files so it might require some integration too. So to summarize, this dataset will require pre-processing before applying a model to it.

Since this is a classification problem, we plan to use Decision Trees and Random Forest technique in their standard format for analysing success of the movies. We are going to implement our source code in Pycharm using different Python packages like Scikit-Learn.

After modelling is complete, for estimating performance measures, we would use Confusion Matrix and F-1 scores as metrics to judge the accuracy of our results.

## Tentative Schedule:

## Future References:

1. “Predicting Movie Success using Machine Learning Techniques” (2017) Retrieved from http://www.diva-portal.org/smash/get/diva2:1106715/FULLTEXT01.pdf