Proposal: Movie Box-Office Success Prediction

1. **What I'm Doing**
   I will try to predict a movie's success with a model that is based on various pre-release features like budget, country of origin, certification, movie studio, genre(s), main actors, director etc. I will also try to consider features like YouTube trailer views (for more recent movies) and other unconventional attributes for which I may need to augment the existing dataset that I use.
   Success of a movie can be measured in multiple ways but I will just be considering its revenue as a multiple of its budget. This does not account for movies released directly on streaming or highly rated films that make minimal amounts of money.
   Based on the revenue-over-budget multiplier, I will classify the movie into various classes ranging from big flop to big hit; I will decide on the exact number of classes as I get started on the project.

2. **How I Am Doing It**
   I will use multiple methods and contrast and compare their results along with trying to understand why a particular method works best and why another method doesn't.
   I cannot specify the final details of the number of methods and which methods I will be using since there may be some time constraints and resources that might crop up, along with whether I can sufficiently understand all the methods I will be applying.
   For the success classification problem I will try to use logistic regression, SVMs, decision trees, random forests, naive bayes and some other methods depending on the feasibility and my own understanding. I will also explore boosting and other ensemble methods.
   If it is feasible, I could also try exploring regression methods to get an exact revenue and roughly contrast the accuracy with the classification methods.

   Dataset
   I was looking through various datasets online on Kaggle and other sites but couldn't find one that would have all the features that I wanted. Even with the existing datasets, I could not choose one particular one that I would fix for the project. I may need to combine features from various datasets and merge them for each movie.
   Some datasets I am considering are TMDB, this, and this.

3. **What I can show as proof**
   Apart from the project writeup, I can show the outcomes of some specific test cases of movies or show the model's prediction for a movie given by the course staff, subject to the condition that the relevant data exists for the model to make a prediction for that particular movie and that the movie was released after the movies used in the training data.