Project Proposal

Group Members: Abhimanyu Swaroop (as6434), Karveandhan Palaisamy (kp2941), Wanru Hu(wh2483), Xiaorui Qin (xq2209), Binghong Yu (by2325)

1. Background and context to the problem statement

With the development of machine learning, machine learning models are gradually applied to all aspects of our life. From a business point view, predicting the revenue of movies is a very exciting topic for the film industry. If we could predict the box office of a movie before it is released, filmmakers would adjust the cast, directors, budget, and propaganda plan of movies in time, and fans would also decide their expectations of movies. With this background, we choose The Movie Database Box Office Prediction dataset (TMDB) for further research. In this project, we will use a variety of features (e.g., budgets, crew, cast, plot keywords, posters, release dates, and production companies) to predict the worldwide revenue for 4398 movies.

2. Description of Dataset

Link to the dataset: https://www.kaggle.com/c/tmdb-box-office-prediction/data
The training dataset contains 7398 movies with 22 features some of which are cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. Some movies share the same titles while having unique ID thus should be treated as separate movies. This happens due to different movies having the same name or old movies having a remake.
The dataset contains few missing values, and text data which might require appropriate pre processing. Few of the features have been described below.

Feature Name	Data Type	Description
id	Int	Basic key denoting each data
budget	Int	Cost of production
genres	Dictionary	Genres of the film
key_words	Dictionary	Specific terms relating to the film apart from genres (Ex: sequel, bollywood, baseball, serbia)
Production Companies	Dictionary	Records which company the movie is produced by
cast	Dictionary	Information about all cast members
crew	Dictionary	Structure of the crew team
overview	Text	Plot description
popularity	Float	Popularity of the movie
runtime	Int	Duration of the film

3. Proposed ML techniques you are proposing on applying to solve the problem

Since we are trying to solve a regression problem, the goal is to reduce the Root Mean Squared Logarithmic Error of our prediction. We take a log to make sure that blockbuster movie revenues are not overweighted. To do so we are planning on using models such as linear regression (with and without regularization), if a linear relationship between variables can be observed. In the event of there being no linear relationship, we will attempt to use tree based models such as Decision Trees and Random Forests. Eventually we will create an ensemble to minimize the risk of our model overfitting by using boosting models such as XGBoost, LightGBM, CATBoost and/or an ensemble of them!