**Capstone 1 Project Proposal: Box-office Prediction:**

**Amin Khoeini**

**Problem**

In the era of social media, Community-driven platforms such as Twitter and Facebook gave a voice to every user of a cultural commodity.  In the 1980’s,  people bought the Cahiers du Cinéma magazine to see if a film is acclaimed by critics and is worth watching. Today, the new generation uses websites such as IMDB and Rotten Tomatoes to see what ratings the movie earned. Each user of these websites can rate a movie and if they decide to initiate a twitter storm about it to hype or destroy its reputation, they have all the means to do so. In this era, the fate of a movie’s success  is in the hands of the users of those websites.

Squid Game is a South Korean TV show that struggled to find a producer for almost 10 years. Finally, it was picked up by Netflix in 2018, and during the first week of its release, was wildly unpopular. Then, there was a huge presence of the show on Twitter and other social media platforms and by the end of the second week, Squid Game was number one show on Netflix and ended up being the most viewed TV show on Netflix. Although a movie critic might call the show below average or even call it worthless, it currently has an 8/10 vote on IMDB with more than 380,000 votes. This is just one example about how ratings on websites impact the success of movies and TV shows in our era. This intrigues my interest to dig deeper to see how much the internet and social media impacts the box office of movies today.

**Goal and Utility**

The goal of this project is to predict the box-office of movies based on the rating from popular websites such as IMDB, Rotten Tomatoes and Metacritics. IMDB voting is only done by a user of the website who are regular movie viewers. Metacritics votes are done solely by critics and only normal users can see the rating. Rotten Tomatoes on the other hand has two rating systems; one for critics which is called Tomatometer and one for normal users which is called an audience score. First, I made a model using data to see how a combination of these votes can predict the box office and further to see different impacts of each type of voting (Audience and Critics) on the box-office prediction.

**Data**

The data set for this project consists of 3312 rating sets of movies. This dataset was gathered from 4 different sources:

* Box office mojo website which provides data about the box-office of movies. Data was manually scraped from the website using Beautiful Soup library.
* IMDB which can be found on Kaggle and consists of average vote, male vote, female vote, US vote and non-US vote accompanied by how many people voted for the movie.
* Rotten Tomatoes which is also found on Kaggle. This set has both a value for critic vote and audience vote.
* Metacritic which was obtained from the OMDB database api and consists only of critic votes for each film.

**Approach**

This is a supervised regression task. Each feature consists of votes and the number of people who voted in two columns and those will be used to create a single score for that feature. Then all the scores will be used as a predictor for the model. Movie box-office will be the dependent variable. During the first step, only one model will be made by using all scores as a predictor, to see how efficiently they can be used for the prediction purpose. Next I will attempt to make a separate model, one only using the critic score and a second one with the audience score. The comparison of performance of these two could be very informative about the different impact between critic and audience votes on the final box-office of the film.

**Deliverables**

The deliverables of this project will include the source code, data set, a slide deck, and a paper outlining the purpose, approach and result of the paper.