**Problem Statement:**

1. Movie production and Streaming companies put enormous effort to increase revenue and make streaming profitable. The competition between the streaming companies is also high these days and increasing the revenue is very challenging.
2. Media Streaming subscribers unable to choose right movies during their well-deserved relaxation time. Some movies are bad and people ending up wasting their time picking a wrong one.
3. Parents sometimes are confused in picking right movie with good age certification Genre and rating for their kids to watch.

**Business and Financial Impact:**

1. Due to competition, the stock price for the streaming companies going down. Please refer the graphs shown below.
2. Waste of time watching bad movie causing frustration in people sometimes.
3. Parents end up picking wrong movie with inappropriate age certification and Genres.

**Product proposed:**

1. A Model that helps increase revenue for the Movie Production/Streaming Companies by highlighting the features that directly influence the increase of Gross Revenue. Per my analysis, the movie rating, votes and meta score in IMDB are directly influencing the Gross Revenue. Hence by increasing the movie rating, the gross revenue can be increased.
2. Rating is fully influenced by votes and meta score. My model will help identify the features influencing the Rating.
3. Age certification and Genre also influence Ratings by some means and the same can be helpful picking the right age Certified good rated movie with necessary Genre.

**Data Source:**

IMDB movie rating website.

**Web Scraping:**

I am planning to use web scraping extension that Google chrome provides that can be installed on my machine. It’s a very easy and quick scrapping tool helps me extract data from any website.

**Data Set:**

<https://www.imdb.com/list/ls047884000/?st_dt=&mode=detail&page=1&sort=list_order,asc>

**Data Clean up:**

After scraping the data, I have to do data clean up, identify features and targets needed for my product development.

**Techniques:**

Various Supervised, Unsupervised and deep learning models.

**Challenge:**

1. Web scraping extension that Chrome provides need to be carefully used while selecting data from websites, as it is sometimes picking up wrong fields.
2. Cleaning up the data is the biggest challenge, as the Chrome extension web scrapper gives all the data in single column. I need to use various string split and extraction techniques to bring it to usable dataset.
3. Due to time constraints, I am limiting the movie data for the countries US and Canada only, as expanding it to other countries data may be quiet challenging at this moment.

**Note:** I tried multiple websites and tried scraping public data. I spent 4 days on the same. The free data is not very useful. I even scraped and extracted Zillow house pricing data but it is not very useful. Some websites ask for payment or credit card information. I am not very comfortable proceeding with those, as I am not very sure if they are reliable websites. I ended up with this simple solution, as I am already running behind my schedule.

**Expected Results/use cases:**

1. Movie production companies can decide which Genre to choose while producing their next movie.
2. Movie production companies can decide where to stream their movies.
3. Streaming companies / Movie production companies can assess how age certification impacts the votes and metascore which in turn impacts the Rating and GrossRevenue.
4. Media streaming subscribers and parents can wisely pick the good rated , age certified and needed Genre movies that also have good votes and meta score. The same will make them feel good about their free time well spent.

**Scraped Data:**

[**https://drive.google.com/file/d/1FFrP4v9IqluPXJEtEUDxs3fUPIcWrIYn/view?usp=sharing**](https://drive.google.com/file/d/1FFrP4v9IqluPXJEtEUDxs3fUPIcWrIYn/view?usp=sharing)

[**https://drive.google.com/file/d/1jPcUzNOoMfUcLUgj0rl3Lcr7BmAAjEND/view?usp=sharing**](https://drive.google.com/file/d/1jPcUzNOoMfUcLUgj0rl3Lcr7BmAAjEND/view?usp=sharing)

[**https://drive.google.com/file/d/1eZLLO28thSBxQ6oVJmn582fvXc5KG2GW/view?usp=sharing**](https://drive.google.com/file/d/1eZLLO28thSBxQ6oVJmn582fvXc5KG2GW/view?usp=sharing)

[**https://drive.google.com/file/d/1EikZxIK7S19ZByfJDbUHY7t0p9YklAvI/view?usp=sharing**](https://drive.google.com/file/d/1EikZxIK7S19ZByfJDbUHY7t0p9YklAvI/view?usp=sharing)

**Jupyter Notebook [To extract and clean up data]:**

[**https://colab.research.google.com/drive/1R\_hUYrsVysPJyrbsNPRQmxGN9XJj3Oa9?usp=sharing**](https://colab.research.google.com/drive/1R_hUYrsVysPJyrbsNPRQmxGN9XJj3Oa9?usp=sharing)

**Reference:**

<https://www.statista.com/statistics/187193/box-office-gross-of-film-studios-in-north-america-2010/>

<https://www.businessofapps.com/data/video-streaming-app-market/>





