Amazon Prime Video Dataset:

There are about 4000+ rows and 16 columns. Cvt\_per\_day is a measure on how much a movie is consumed by the audiences per day. Higher number means the movie is more popular on our platform.

The following are the descriptions of the columns:

|  |  |  |
| --- | --- | --- |
| Video\_id | : | A unique id for a movie |
| Cvt\_per\_day | : | Cumulated view time per day |
| weighted\_categorical\_position | : | Average vertical positions on the home page that the movie was placed |
| weighted\_horizontal\_poition | : | Average horizontal positions on the home page that the movie was placed |
| genres | : | genres of the movie |
| release\_year | : | the year the movie was released |
| imdb\_votes | : | the number of votes on IMDB, typically higher the votes the better |
| budget | : | budget of the movie production, typically the higher the better |
| boxoffice | : | gross box office in US as updated on IMDB, typically the higher the better |
| imdb\_rating | : | ratings on IMDB |
| duration\_in\_mins | : | How long is the content in minutes |
| mpaa | : | MPAA ratings |
| awards | : | TVPG ratings |
| Import\_id | : | Content partners |
| Metacritic Score | : | Metacritic score on IMDB page. Typically, the higher the better |
| Star\_cateogry | : | A score to measure how popular the actor/actress are associated with the movie |

# Questions:

1. Build a prediction model to predict whether a movie is going to perform well on our platform (cvt\_per\_day) based on the information in the dataset.
   1. What are the important features?
   2. How do you weigh in the position bias – When a movie is placed at the first position, it naturally has higher cvt because more viewers will click on it
2. We are curious to hear about (1) how you dealt with missing values, (2) how you selected, transformed, or engineered new features, (3) the type of model used, and why, and (4) any particular insights or other thoughts you have on this dataset.

# Instructions:

1. Please attach both your write-up and your code in one document that is easy to understand.
2. Including visualization in your write-up is highly recommended.
3. You are free to use whatever tools you are most comfortable with to work through the analysis.

You might want to start your investigation of missing data by finding out whether you have informative or non-informative missings. The first category is produced by random data loss; in this case, the observations with missing values are no different from the ones with complete data. As for informative missing data, this one tells you something about your observation. A simple example is a customer record with a missing contract cancellation date meaning that this customer's contract has not been cancelled so far. You usually don't want to fill in informative missings with a mean or a median, but you may want to generate a separate feature from them.