Hollywood Blockbuster Case Study - General Queries & Data Dictionary

Please submit the following:

- Code (via notebook, python/R scripts) **.csv file** output of the Scoring sheet with movie IDs and prediction category.
- Scoring sheet with predicted categories.
- Any documents/files that you think are relevant.

1, May I know what the field "total" & “category” means in this dataset?

➔ Total reflects the Total Gross Earnings (in Millions of $) & Category reflects the Category of Movie (based on Total).

2, Can I use scraping to get additional data about movies (like actors) and use them in the model?

➔ The model we evaluate you on should be without web scraping. That being said, if you would like to showcase your web scraping capabilities and how it could assist in model performance, you can submit a separate model with the web scraped data.

3, Would we be expecting a data pipeline of the model (Jupyter notebook) and a report summarizing the pipeline and results; or just the report emphasizing on the results?

➔ We would be expecting both.

4, However, for one of the variables "board_rating_reason", it contains a lot of text. Since its only purpose is to explain the rating, would it be ok for me to drop the variable? Or do you have specific requirements/expectations for that particular variable?

➔ There are no requirements, but we find most attempt to make use of this field.

5, How in depth do we want the data exploratory analysis to be?

➔ This is up to you... As long as you have done enough to grasp the structure of the data and engineer any features you think will improve your submission, you should be in good shape.

6, ‘production_budget’ is present in the scoring sheet but not in the training dataset, Similarly ‘total’ is present in the training dataset but not in the scoring sheet what should I do ?

➔ Make judicious use of variables. However, keep in mind that you are building a predictive model.

7, Importance of accuracy vs overall approach. Is the project evaluated on the accuracy of the results or the overall approach?

➔ Accuracy is important and so is the overall approach.

Data Dictionary

Training Dataset	
id	Unique Movie ID
name	Short Name (if exists)
display_name	Movie Name
production_year	Year of Production
movie_sequel	Is the movie a sequel or not (1=yes, 0=no)
creative_type	Creative category of the movie
source	Source of the movie script
production_method	Production Style (ex: Live Action, Animation)
genre	Genre of the movie
language	Original Movie Language (Audio)
board_rating_reason	Reason behind the movie rating assigned by the Motion Picture Association
movie_board_rating_display_name	Movie Rating
movie_release_pattern_display_name	The movie release level reflects how many theaters a movie is released in
total	Total Gross Earnings (in Millions of $)
Category	Category of Movie (based on Total)

Scoring Dataset	
id	Unique Move ID
name	Short Name (if exists)
display_name	Movie Name
production_budget	Budget allocated to the movie production
production_year	Year of Production
movie_sequel	Is the movie a sequel or not (1=yes, 0=no)
creative_type	Creative category of the movie
source	Source of the movie script
production_method	Production Style (ex: Live Action, Animation)
genre	Genre of the movie
language	Original Movie Language (Audio)
board_rating_reason	Reason behind the movie rating assigned by the Motion Picture Association
movie_board_rating_display_name	Movie Rating
movie_release_pattern_display_name	The movie release level reflects how many theaters a movie is released in