Anime Rating Prediction – Objective & Deliverables

# Problem description:

## The objective here is to perform a Linear regression analysis to arrive at a model that can be used to predict the ***Ratings*** received by the enlisted anime releases (Movie/Web series), such that, in future, the anime production studios can develop their strategies which can improve the ratings.

**Data Dictionary:**

**About the dataset:** This dataset comprises the scrapped information about anime releases (Movie/Web series/etc.) from anime-planet (founded in 2001), which is the first anime & manga recommendation database. It comprises the anime & manga release logistics (Title, Description, Episodes, Duration, etc.) along with the viewer’s response behaviour statistics (Watched, Want to watch, Watching, Votes) records from the year 2005 to 15th June, 2020.

**rating:** Average user rating given by the viewers for the anime releases.

**title:** Name of the anime releases.

## **mediaType:** Format of publication of the anime releases (Web/DVD special/Movie/TV special/TV).

**eps:** Number of episodes (movies are considered 1 episode).

**duration:** Duration of each episode (in minutes).

## **Ongoing:** Whether the anime is ongoing or not (Yes/No).

**sznOfRelease:** The season of release of the anime (Winter/Spring/Fall/Summer).

**description:** Synopsis of plot of the anime.

## **studios:** Studios responsible for the creation of different anime.

**tags:** Tags, genres, etc. of different anime.

**contentWarn:** Content warning provided for the different anime.

## **watched:** The number of users who completed watching it.

**watching:** The number of users who are watching it.

**wantWatch:** The number of users who want to watch it.

## **dropped:** The number of users who dropped it before completion.

**votes:** The number of votes that contribute to the ratings received by different anime.

**Recommended Project Steps & Guidelines**:

* 1. Identify the Problem Statement - what are you trying to solve?
  2. Import the dataset and identify the Target variable in the data.
  3. **Identifying the type of variables:** Identifying the nature of different columns (Continuous/Categorical/Qualitative), removing garbage columns (if any) and conversion of categorical variables to factors if they are not in factors.
  4. **Data pre-processing:** Checking and treating the missing values with appropriate measures. Checking the presence of outliers by creating boxplots and treating the outliers (if any).
  5. **Univariate and Bivariate Analysis:** Explore each "Potential" predictor for distribution (visual analysis – histogram/barplot) and also explore their relationship with the target variable (visual analysis – scatterplot/boxplot and statistical tests – correlation/ANOVA).
  6. **Feature selection:** Finalize the set of potential predictors to be used in the linear regression algorithm.
  7. **Splitting the data into train & test:** Divide the data into two parts: Train sample (70%) and test sample (30%). The machine learning algorithm will be applied on the Train set and the model will be validated on the test set.
  8. **Model Building:** Form the multiple linear regression model with the set of potential predictors identified from Exploratory data analysis and obtain the significant predictors.
  9. **Multicollinearity check:** Check the presence of Multicollinearity in your final model and remove the variables with high multicollinearity one by one from your final model to arrive at the model which will be used to generate predictions on the test data.
  10. **Model accuracy:** Using the final model, generate the predictions on the test data. Calculate the Mean Absolute Percentage Error and Median APE. Obtain the mean accuracy and median accuracy of the Linear regression model.
  11. **Improving model performance:** If the goodness of fit/accuracy is turning out to be low – you can make transformations of certain variables (for example - log transformations) in the data and use them in modeling to see if the goodness of fit/accuracy is improving or not.