Cheah Jun Yitt (WQD180107)

Data Mining Project: Milestone 4 (Interpretation of Data) (Individual)

Recap

Previous milestones have established the following:

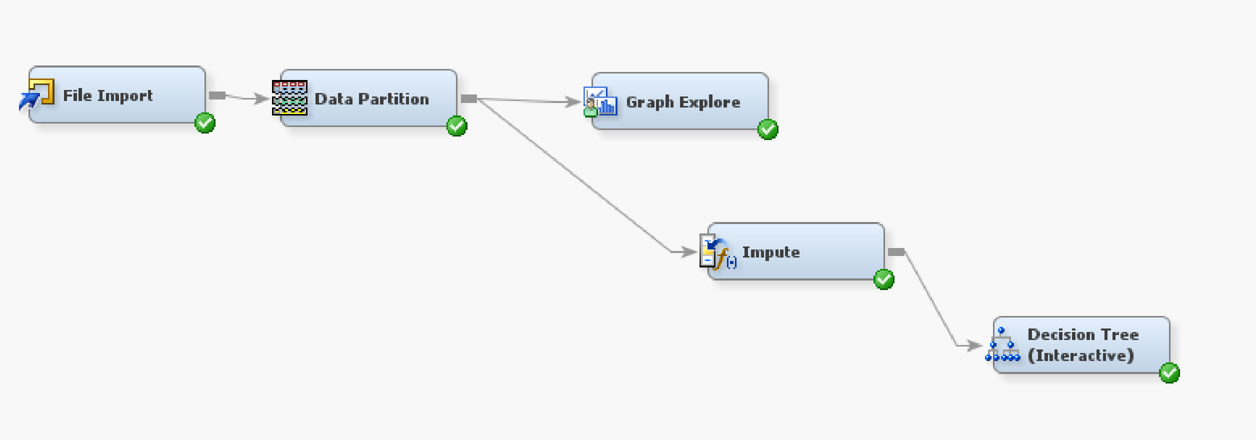
1) The need to have various sources of data, both structured and unstructured data. The data that we have collected are as shown below:
    1) Structured data: Movie information scraped from rottentomatoes.com
    2) Unstructured data (text data): Movie reviews scraped from rottentomatoes.com
    3) Structured data: Movie box office data scraped from boxofficemojo.com
2) Target variable (*audience_score_positive*), which indicates whether a movie is good or bad based on audience ratings on the movie.
3) Analysis Goal: To use sentiment score of user reviews on a movie, movie information and box office data to predict the user ratings of a movie.
4) SEMMA (Part 1: SEM):
    a. Sample – Import data from a CSV file
    b. Explore – Use histograms to identify missing values, inconsistencies or hidden patterns in the data.
    c. Modify – Impute missing values.

Interpretation of Data

In this milestone, the data will be first partitioned into 50% training set and 50% validation set, then modelled using decision tree model.

Data Partition

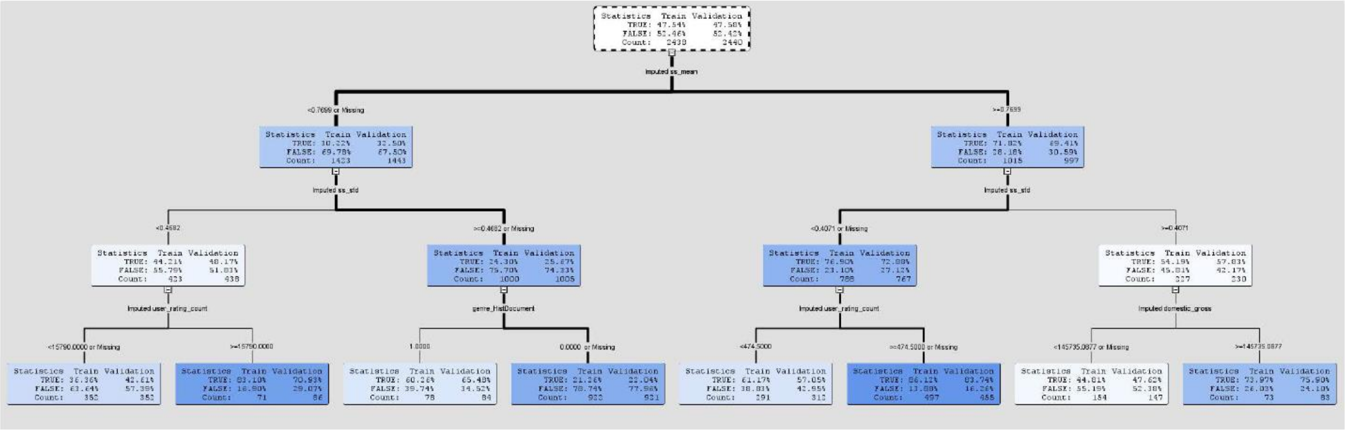
The data partition node is added after the file import node to split the data into 50% training set and 50% validation set as shown below.



Model

After imputing missing data (milestone 3), the data is modelled using decision tree. The decision tree is interactively built, where the node is split based on a few factors:

1) Information gain: Selecting attribute with high information gain is preferred.
2) Interpretability: Selecting attribute with a slightly lower information gain but can be easily understood and explained is preferred.
3) Diversity: Using attributes from various sources is preferred (as established in previous milestones), hence it is preferred to split the nodes with consideration of attributes from movie information, movie reviews (sentiment score) and movie box office performance.



First layer:

The interactively-built decision tree is shown above. First, the tree is split based on "*Imputed ss_mean*" which represents the mean sentiment score of all the reviews of a movie, it has the 3rd highest information gain (88.8877) and easier to understand (mean sentiment score is easier to understand than 25th percentile of sentiment score). The resulting nodes show that the if the mean sentiment score is higher than 0.7699, then about 69% of the time, the movie is good and should be recommended to the users.

Second layer:

After examining a movie's mean sentiment score, the tree is further split by the "*Imputed ss_std*", which represents the standard deviation of the sentiment score of reviews of a movie. A low standard deviation implies that the reviews have a common sentiment, while a high standard deviation implies that the reviews are mixed in sentiment (some people thinks the movie is good, some do not). Based on the tree above, we can derive a few insights:

1) Movie with a low mean sentiment score (<0.7699), and a high standard deviation in sentiment score (>=0.4682), it is very likely to be a bad movie (74.33%). Therefore, a bad movie typically has mixed sentiment reviews, and the sentiment tends to be negative.

2) Movie with a high mean sentiment score (>=0.7699), and a low standard deviation in sentiment score (<0.4071), it is very likely to be a good movie (72.88%). Therefore, a good movie typically has similar sentiment reviews that are positive.

Third layer:

The tree is further split based on the following paths:

Path 1: Low mean sentiment score (<0.7699), low standard deviation in sentiment score (<0.4682)

The node is split based on "*Imputed user_rating_count*" attribute, which represents the number of user rating on the movie. A high number of user rating (>=15790) imply that a lot of people have watched the movie and are compelled to make a rating. Hence, this generally shows that the movie is good (70.93%).

Path 2: Low mean sentiment score (<0.7699), high standard deviation in sentiment score (>=0.4682)

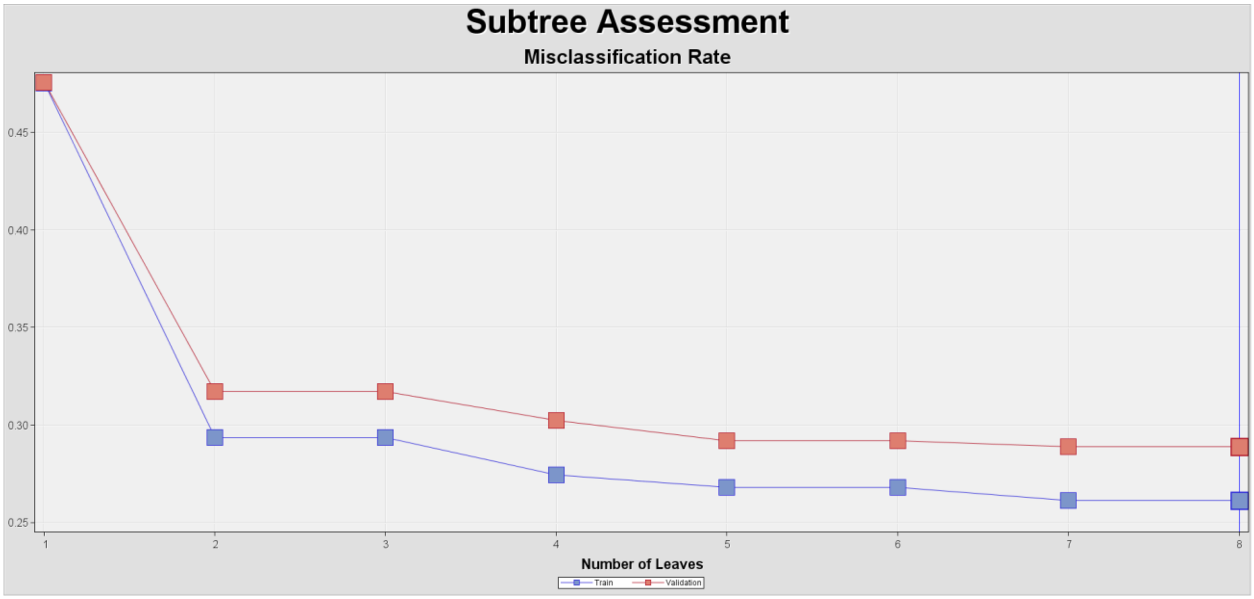
The node is split based on "*genre_HistDocument*", which indicates whether a movie is a historical documentary or not. A historical documentary movie with a slightly lower mean sentiment score and highly debated (high standard deviation) shows that the movie presented a fascinating plot that left the audience to critically review on it. Hence, generally this type of historical documentary movie is considered a good movie (65.48%) worth recommending to users.

Path 3: High mean sentiment score (>=0.7699), low standard deviation in sentiment score (<0.4071)

The node is split based on "*Imputed user_rating_count*". A high number of user rating implies that the movie attracted a lot of users to watch and rate it. Hence, the movie is generally good (86.12%).

Path 4: High mean sentiment score (>=0.7699), high standard deviation in sentiment score (>=0.4071)

The node is split based on "*Imputed domestic_gross*", which represents the total domestic gross amount of the movie throughout its screening on theatre. A higher domestic gross amount (>=145735.0877) implies a lot of people have paid to watch the movie on theatre, hence the movie is considered good in general (75.90%).



Based on the subtree assessment plot shown above, having about 3 leaves are enough to achieve a low misclassification rate. At the same time, having not too little or not too many leaves allow the decision tree to be interpreted and understood easily for decision making.