DATA WRANGLING

Data wrangling is the process of converting data from its raw form to the tidy form ready for analysis. Data wrangling is an important step in data pre-processing like data importing, data cleaning, data structuring, string processing, HTML parsing, handling dates and times, handling missing data and text mining.

This section describes the various data cleaning and data wrangling methods applied on the Movielens datasets to make it more suitable for further analysis. The following sections are divided based on the procedures followed:

Cleaning:

The Movielens data are contained in six files, genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv and tags.csv. It contains 20000263 ratings and 465564 tag applications across 27278 movies. These data were created by 138493 users between January 09, 1995 and March 31, 2015. This dataset was generated on October 17, 2016.

The dataset had a lot of features which had 0s for values it did not possess. These values were converted to NaN. These NaN values are filled in two ways according to the type of the data

1. Categorical type was filled with the mode values.
2. Numerical type was filled with mean values.

Removing Unnecessary Features:

This process was done in different ways

Number of missing values:

If a particular feature has more than 50% of missing values in it then most of the times that particular feature will not play any significant role in learning the model. But in my dataset I am not at all having a single feature with more than 50 % of missing values. But there are some features with fewer NaNs. Those features are tag, year and tmdbId but I am not removing those features.

In this Movielens dataset I am having almost 12 features. But I am not using all those 12 features whichever the features are important to me for solving the recommendation problem, I am using those features only.

Significance:

MovieLens dataset bases its recommendations on input provided by users of the website, such as movie ratings. For each user, MovieLens predicts how the user will rate any given movie on the website. Based on these predicted ratings, the system recommends movies that the user is likely to rate highly.

Outliers:

For features like Ratings, whose rating-value is > 5 or negative values, those values are called as Outliers and removing those values will help a lot in predicting the accurate output. And here in this case there are no outliers present in this dataset.