Milestone\_Report\_3\_Team\_E

Project Goal:

1. Build regression model to predict Youtube average daily view counts.

2. Rank the importance of feature that contributes to the daily counts.

3. Find out the top 10 rule that contribute to high views.

Current Progress & Challenge:

1. We used exhausted feature selection to find the weights of each feature and rank the importance of feature to the view counts prediction according to the weights. We used stepwise model selection method and decision tree to find the importance of features for linear regression.
2. However, the features of daily likes, daily dislikes, and daily comment counts are highly correlated to the results. This causes a dilemma that we cannot tell whether those features contribute the view counts or they are also the results. Therefore, we decided not to use them neither as feature for prediction nor for association rule.
3. After deleted the three aforementioned features, the number of left features are low, and therefore we added more features.
4. Also, as suggested by professor, we found some our features are highly correlated, such as channel frequency high and channel frequency low. Hence, we checked the correlation between each pair of feature and deleted one feature in pairs that are highly correlated to each other. (by using Pearson correlation coefficient)
5. To find out words that are frequently shown in the title and tags, we did bag of words analysis. However, the matrix for the result is too large, we decided to select some high-frequency words as features to analyze.

Next Steps:

1. Feature engineer by pearson correlation coefficient

2. Bag of words(select important words from 20000)

3. Try linear regression(model selection, regularization) again and plot scatter of prediction point(view, view\_predict) and calculate R square

4. Try decision tree again and get the importance

5. Try association rule

6. Try Gradient boost