Project Proposal

2016010829 Hu Yiju, 2016010845 Wang Yuhan,
2016013327 Xiang Yutong, 2016010850 Xu Ming

Dec. 3rd. 2019

1 Problem Background

YouTube is the world-famous video sharing website, and it maintains a list of the top trending videos on the platform. To determine the year' s top-trending videos, YouTube uses a combination of factors including measuring users number of views, shares, comments and likes. This dataset is a daily record of the top trending YouTube videos.

Our project will walk through data exploration descriptive analysis to find some interesting conclusions drawn from the data. Like:

1. Different characteristics (channels, tags, etc.) of trending videos in different regions (Asia, European, America, Others)?
2. Most trending topics (channels, tags, clustered semantically) in different time periods?
3. Most widely used commentary words in trending videos?

Besides, based on this, we will try to use time series analysis methods and machine learning strategies such as regression to help predict trending videos, or other potentially useful work.

2 Evaluation Metrics

As our focus is binary-class, we choose the corresponding evaluation metrics to evaluate our model.

1. **Classification**:
   - **Confusion Matrix**: use Confusion Matrix combined with precision, recall, accuracy and error to evaluate the prediction results of this model.
   - F_β**-score**: Trade-off between precision and recall given a single threshold
   - **Receiver Operating Characteristic and Area Under the Curve**: Since we need to take True Negatives into account, we choose ROC rather than PRC and use AUC to compare the different algorithms.

2. **Regression Models**:

   - **Base Methods**: Squared Error, Absolute Error, Squared Logarithmic Error, Absolute Percentage Error.
   - **Adjusted-R2**: As the adjusted - R2 is normalized by sample size and feature size, we choose it to evaluate our regression model rather than the normal R2.

3 Dataset Specification

Collected by the YouTube API, this dataset(539.3 MB in total) comprises several months (and counting) of data on daily trending YouTube videos in different regions, including USA(US, 40949 entries), Great Britain(GB, 38916 entries), Germany(DE, 40840 entries), Canada(CA, 40881 entries), France(FR, 40724 entries), Russia(RU, 40739 entries), Mexico(MX, 40451 entries), South Korea(KR, 34567 entries), Japan(JP, 20523 entries) and India(IN, 37352 entries) respectively, with up to 200 listed trending videos per day.

Each region's data is in a separate csv file with 16 columns; each csv file includes a category_id field, which varies between regions and thus is stored in associated JSON files.

4 Preliminary Technical Solution

1. EDA strategies
2. Necessary Feature Engineering Methods
3. Time Series Analysis Methods (e.g. ARMA, ARIMA, AR) for prediction
4. Regression Models (e.g Random Forest, Adaboost, XGboost) for prediction
5. Evaluation Strategies as mentioned above