Music Box Project

Churn Label Definition

- 0 activities over certain window
  - 14 days (4/29-5/12)
  - No activities (play, down, search)
- Population
  - Include
    - 3/1-4/28 all active users
  - Exclusion
    - Outliers
    - Inactive users (3/1-4/28)
  - Enrich train data
    - Sliding window snapshot
    - Snapshot date
      - 14 days after: define label
      - Before: create feature

Down Sample, Train Data Prep

- Find Churn Users
    1. Find all active users in 3/1-4/28 (uid list)
        1. python
            1. Read file line by line, read uid
            2. Deduplicate
        2. Linux cmd
            1. Cut, uniq
    2. Find all active users in 4/29-5/12 (uid list)
        1. Similar
    3. Churn users = Uid missed in 2 from 1 (challenge?)
- Down Sampling Activity Data
    - Load churn uid set
    - Read file line by line, look up set
        - If in churn user set, keep
        - Else keep with a probability, output to a new file (append label)
    - Good user down sampled raw train set: uid, user activities, label
        - Sampling on uid level, not record level
            - Downsample uid (good users), keep good uid hash, look up for every row, in set keep.
        - 10x, 100x

Feature Creation – User Activities

- Count by time window (Velocity)
  - Activity type (3):
    - Play, down, search
  - Last x days (6)
    - 1, 3, 7, 14, 30 ,60 (inclusive/exlusive e.g. 1, 2-3,4-7 ….)
  - 18 features
- Ratios (de/acceleration)
  - Ratio of different time window: e.g.
    - play_1d_ov_7d,
    - play_1d_ov_down_1d
    - play_1d_ov_down_7d …
- Add granularity
  - Counts and Ratios + Song genre
    - E.g. (Rock)_play_1d_ov_down_1d

Feature Creation – User Profile

- User subscription time
- User preferred device
- User preferred genre: classic, new songs
- Gender
- User age
- ….

Train model

- Preliminary Model
  - Play * 6 time windows(1, 3, 7, 14, 30 ,60)
- Method
  - Logistic regression
  - Random Forest
  - GBDT