**Problem:**

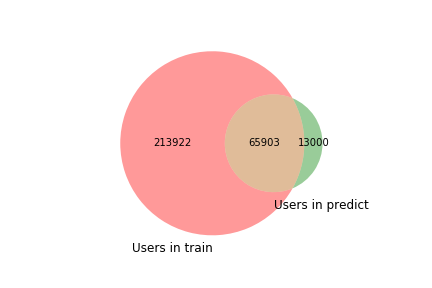
Purchase behavior prediction after receiving the vouchers

To predict whether buyers will use the voucher and continue to purchase

**Data source**

|  |  |  |
| --- | --- | --- |
| Source | Attributes | Time span |
| Training (710078,9) | userid, promotionid\_recieved (voucher\_code), voucher\_received\_time,used?,repurchased\_x (15d,30d,60d,90d) | 2017.1-8 (sparse) |
| Predict (78903,4) | userid, promotionid\_recieved (voucher\_code), voucher\_received\_time | 2017.8.16 |
| Transactions | Userid,shopid,total\_price,order\_time,promotionid\_used | 2015 Q2- 2017 Q3 |
| User profiles | Userid,registration\_time,is\_seller,gender,… |  |
| Likes | Userid,voucher\_received\_date,status,ctime |  |
| view\_log\_0-30 | Userid,promotionid\_received,,date,event\_name,count | 2017.4-2017.8 |
| Voucher\_distribution\_active\_date | Userid,promotionid\_received,time,date,active sessions counts |  |
| Voucher\_mechanics | Promotionid,discount,max\_value |  |

**Data probe**



279803 (out of 279825 users in train) can be found in transactions table

78895 (out of 78903 users in predict) can be found in transactions table

Totally 94 unique promotion id. 92 in train data, 4 in predict data

**Feature generation**

Three super groups: user, voucher, date

13 subgroups

All features are listed in naodong.docx

**Model construction**

1. Split data into training set, validation set and test set
2. Feature extraction
3. Normalization and add missing values
4. Feature reduction
5. Train model

**Result analysis**

Model 1: Random Forest

Model 2: GBDT

Model 3: xgboost

Model 4: LR

Model 5: blending