todos

1. paper writing: background, models description(meaning, applications, problem settled and result) -> shining, baixue, daihuanfang (**deadline Saturday night**)
2. feature engineering -> liuhu(**deadline Friday noon**)
3. customer aggregation -> liuhongguang(**deadline Friday night**)
4. models and evaluation:
   - population regression -> liuhongguang(**deadline Saturday noon**)
   - BTYD package: Pareto/NBD, BG/NBD, BG/BB -> wuluyun(**deadline Saturday noon**)
   - GM11 gray model -> sunning(**deadline Saturday noon**)
   - other models -> liuhu(**deadline Saturday noon**)
   - ensemble -> liuhu(**deadline Saturday night**)
5. final paper writing -> huang hao(**deadline Monday night**)

feature engineering

population statistics(of each month, quater)

1. customer active ratio
2. purchase times
3. purchase unit of books
4. purchase value
5. purchase mean value

per-customer statistics

1. purchase units of books
2. purchase value
3. purchase mean value
4. last purchase time(amount of days, which month, which week)
5. last purchase value(and ratio of this customer)
6. last purchase amount(and ratio of this customer)

per-customer squential statistics(of each month, quater)

1. purchase units of books(and ratio of this customer)
2. purchase value(and ratio of this customer)
3. purchase mean value(and ratio of this customer)

models

1. BTYD package: Pareto/NBD, BG/NBD, BG/BB
2. GM11 gray model
3. Others: naive_bayes, knn, logistic regression, decision tree, random forest, svm, GBDT...

objectives

1. per-customer: weekly purchase frequency, purchase units of books, purchase value
2. whole customers set...
3. whole year or each of the 26 weeks

train, valid, test sets split

train:valid = 3:6

1. train: 1997 1-3 months, valid:1997 4-9 months, test:1998 1-6 months
2. train: 1997 4-6 months, valid:1997 7-12 months, test:1998 1-6 months

train:valid = 6:6

1. train: 1997 1-6 months, valid:1997 7-12 months, test:1998 1-6 months

train:valid = 3:3

1. train: 1997 1-3 months, valid:1997 4-6 months, test:1998 1-6 months
2. train: 1997 4-6 months, valid:1997 7-9 months, test:1998 1-6 months
3. train: 1997 6-9 months, valid:1997 9-12 months, test:1998 1-6 months

train:valid = 9:3

1. train: 1997 1-9 months, valid:1997 10-12 months, test:1998 1-6 months

train:valid = 10:2

1. train: 1997 1-10 months, valid:1997 11-12 months, test:1998 1-6 months

train:valid = 11:1

1. train: 1997 1-11 months, valid:1997 12 month, test:1998 1-6 months

git repository(scripts, reference scripts, reference projects...)

https://github.com/liuhu-bigeye/book-purchase-prediction.git

some useful tools

stata, spss, eviews, weka, orange