用户人品预测大赛--数据大匠队--竞赛分享

•发布于 2016-03-24 14:20 •作者 DataCastle运营 (/user/5451) •54 次浏览 •来自 微额借款用户人品预测大赛 (/?tab=148)

参赛队队名：数据大匠

竞赛报告书

一、参赛作品概述

1） 赛题解读

本赛题的目标是利用给定的用户多维度行为数据来分析”小额微贷“申请借款用户的信用状况，以分析其是否逾期。其中带标签的训练集有 15000 个样本，无标签的训练集 5000 个样本,测试集为 5000 个样本。数据集提供了 1138 个用户特征，有 numeric（数值型）和 category（类别型），评价指标为 AUC

2） 建模思路

主要有四个部分：数据预处理、特征组合、模型训练预测、模型融合。在数据预处理上，对 label 数据进行 one-hot-encoder 处理、正负例倒置以及数值填充；特征组合按照特征重要性进行组合、随机组合，将不同组合进行单独训练并预测结果，再对结果进行融合。使用到的算法模型有 xgboost（不同版本）、RF，同一算法利用不同参数进行训练得到不同结果集，将其融合。此外，我们还利用模型对 unlabel 数据进行评分，选择评分低的样本作为负样本加入到模型训练中，能得到一定程度的提升。模型融合用了三种方法：1、简单的加权平均；2、用 score 进行排名得到 rank 值，再进行预先设置权重加权融合；3、将结果按照 score 排倒序，得到序号 rank 值，再按 1/rank 加权融合。

2. 算法原理

1） 数据预处理

由于 xgboost 的参数是设定正例权重 scale_pos_weight，本题的正例多于负例，正负例比为 8.7:1，为了让 scale_pos_weight 大于 0，对正负例进行倒置。同时特征值有数值型和 label 型，同时数值存在大量缺失值。针对数值型，对其进行模型归一化标准化，对于 label 型对其进行 one-hot-encoder 处理。对于缺失值，采用平均值和中位数填充，实际山上中位数填充效果要更好。

2） 特征分组

i. 用算法对特征进行评分，按重要性排序，以%7 为 index 分为 7 组，每组数据用 xgboost 和 RandomForest 分别训练输出结果进行融合

ii. 随机选取 400 个特征进行组合，同样用 xgboost 和 RandomForest 分别训练

3） 半监督

在本题数据中，负样本较少，可以根据半监督训练提取出更多的负例样本加入到训练集中。先采用较好的单模型 xgboost 来对 unlabeled 进行评价，根据 AUC 分数进行排名，取分数最低的 Top5000 作为训练样本负例保存。选择负例样本中分成 5 组，每次添加一组 1000 个训练样本添加到原本的训练集中作为新的训练集，采用新的 tune 好的 xgboost 进行线上测评，取线上效果最好的那组数据作为今后的训练数据集

4） xgboost

xgboost 的全称是 eXtreme Gradient Boosting。正如其名，它是 Gradient Boosting Machine 的一个 c++实现，作者为正在华盛顿大学研究机器学习的大牛 陈天奇 。他在研究中深感自己受制于现有库的计算速度和精度，因此在一年前开始着手搭建 xgboost 项目，并在去年夏天逐渐成型。xgboost 最大的特点在于，它能够自动利用 CPU 的多线程进行并行，同时在算法上加以改进提高了精度，在各类 Kaggle 比赛中非常活跃，也有人使用该算法斩获冠军。在本次比赛中 xgboost 单模型效果非常好，需要特别注意的 xgboost 调参部分，需要倒置正负例，设置 scale_pos_weight 大于 0，防止样本损失。同时高维稀

疏数据的特性，防止过拟合要增大正则项 L2 的系数 reg_lambda，为了让精度更高，牺牲时间为代价，调小学习率 eta 和增大 n_estimators 数，适当增加 max_depth。

5) RandomForest

Random forest， 随机森林，顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。建立每棵树的时候有 2 个采样的过程，包括特征采样和样本采样，这保证每次建树都有相当的差异性。按这种算法得到的随机森林中的每一棵都是很弱的，但是大家组合起来就很厉害了。可以这样比喻随机森林算法：每一棵决策树就是一个精通于某一个窄领域的专家（因为我们从 M 个 feature 中选择 m 让每一棵决策树进行学习），这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题（新的输入数据），可以用不同的角度去看待它，最终由各个专家，投票得到结果。

6) 模型融合

i. 均值融合。不同的预测结果直接取平均的 score

ii. rank 加权融合。将不同的结果按 score 排正序，得到每个样本的 rank，再按此 rank 加权融合得到新的 score

iii. 1/rank 加权融合。将不同结果按 score 排倒序，得到每个样本的 rank，再按 1/rank 加权融合得到新的 score

3. 算法实验结果

封榜时我们的结果是 0.734，排名第 2 名，以下是我们在比赛过程中各个模型以及融合的得分情况

模型	参数	得分
M1：特征分组训练+随机抽取特征训练按均值融合	**RF:** **树的棵树：300~500** **树深度：10~15** **xgboost：** **objective: 'binary:logistic',** **early_stopping_rounds:100,** **scale_pos_weight:1400.0/13458.0,** **eval_metric: 'auc',** **gamma:0.1,**	**0.726**

	max_depth:8, lambda:550, subsample:0.7, colsample_bytree:0.4, min_child_weight:3, eta: 0.02, seed:1225	
M2：xgboost(陈天奇版本)	n_estimators: 8000, scale_pos_weight: 8.0, max_depth: 5.0, objective: 'binary:logistic', learning_rate: 0.02, gamma: 0.48, min_child_weight: 4, reg_lambda: 2300, subsample: 0.655, colsample_bytree: 0.4	0.7245
M3：xgboost(graphlib 版本)	max_iterations:1500, max_depth:8, min_child_weight:4.6333144, row_subsample:0.747, min_loss_reduction:2.8913, column_subsample:0.78, step_size:0.027492	0.7159
M4：xgboost(不同参数，数据集)	objective:'binary:logistic' eta:0.03 max_depth:8 eval_metric:'auc' silent:1 min_child_weight:10 subsample:0.7 colsample_bytree:0.3 gamma: 0.1 reg_lambda:100 seed:1250 scale_pos_weight:1400.0/13458.0	0.719
M5： 0.5*M2+0.28*M3+0.22*M4 按照权重对 rank 进行加权融合	—	0.7284
M6： xgboost(加半监督学习负例)	n_estimators: n_estimators, scale_pos_weight: 5.0, max_depth: 6.0, objective: 'binary:logistic', learning_rate: 0.02,	0.7258

	gamma: 0.3, **min_child_weight: 2,** **reg_lambda: 3010,** **subsample: 0.7,** **colsample_bytree: 0.3,**	
M7：M1+M5+M6 按 1/rank 加权融合，其中 rank 为 score 排倒序		**0.734**

三、作品总结

1、算法优势

a) 不过度依赖于调参，利用模型间差异性融合优势大。rank 融合方式效果比一般的均值融合在提升 auc 评价效果更好；

b) 算法框架能集成多种模型，Xgboost 和 RandomForest，效果好于最佳单模型；

c) 相同模型运用于不同数据集产生不同 meta-feature 能加以融合；

d) 半监督学习能有效增加负例数从而进一步改进模型效果。

2、可能的改进方向

a) 对特征进行更详细的分析，如果赛方能提供更具体的业务知识，结合这些知识进行相关特征工程工作，进一步提升单模型效果。

b) 由于 AUC 评价涉及相对排序问题，是个排序优化问题，可以尝试 Learning to Rank 算法的 LambdaMart 框架进行排序学习。

c) 由于数据特征高维、稀疏且含有缺失值，可以考虑数据特征降维，诸如 SVD、PCA 等算法进行降维处理，或许也能采用深度学习 DNN 的处理方式进行特征降维选择。

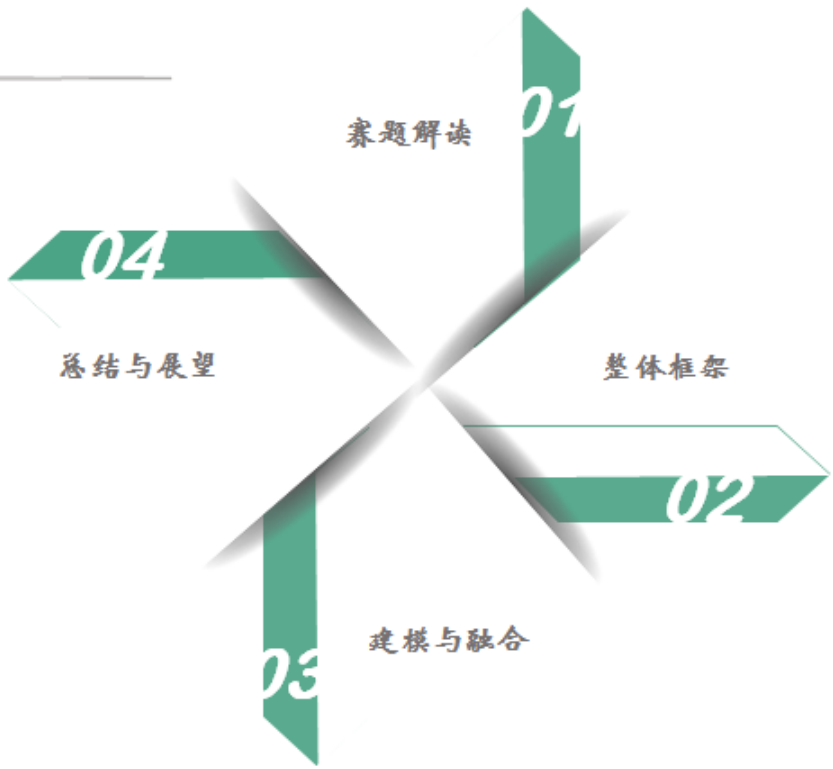
d) 尝试使用蚂蚁金服内部广为使用的 bilinear 模型，能够大大提高分数。

答辩PPT

用户人品预测大赛答辩

团　　队：　数据大匠

目录



赛题解读

赛题描述：根据给定的训练集预测测试集中用户的人品，其中1代表用户人品杠杠滴，0则代表人品堪忧。预测结果提交线上评价，评价指标为AUC，按得分进行排名。

AUC(Area Under Curve) 等价于Wilcoxon-Mann-Witney Test公式：

$$AUC = \frac{\sum_i S_i}{|P| * |N|} \qquad S_i = \begin{cases} 1 & score_{i\_p} > score_{i\_n} \\ 0.5 & score_{i\_p} = score_{i\_n} \\ 0 & score_{i\_p} < score_{i\_n} \end{cases}$$

→ 相对排序问题

赛题解读

数据集：有1138个特征，特征值分数值型和类别型，都进行脱敏处理。

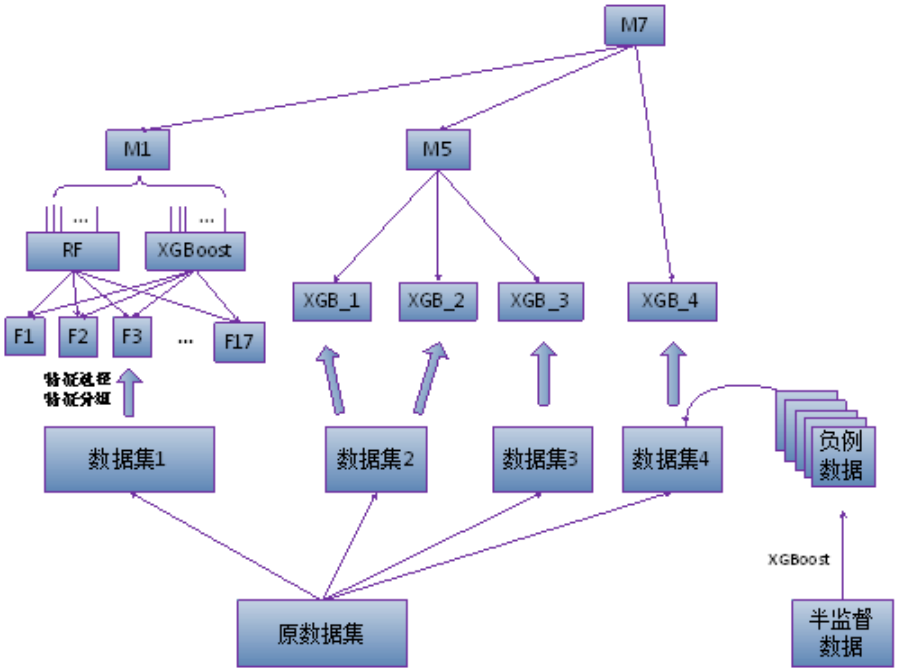
- 训练集（带标签）
- 测试集
- 半监督数据集（无标签）

难点：

- 未给出特征具体含义，很难结合业务知识
- 数据维度高维稀疏且含有缺失数据
- 正负例不均衡（8.7:1）
- 线上评测与线下CV评测不一致（推测测试集与训练集不一致）

挑战：对半监督数据集进行半监督学习

整体框架



Ensemble技术的关键是
模型差异性

数据预处理

1. 对类别特征one-hot编码

类别特征 → one-hot-encoder → 特征扩充

f1 : {"1","2","3"}

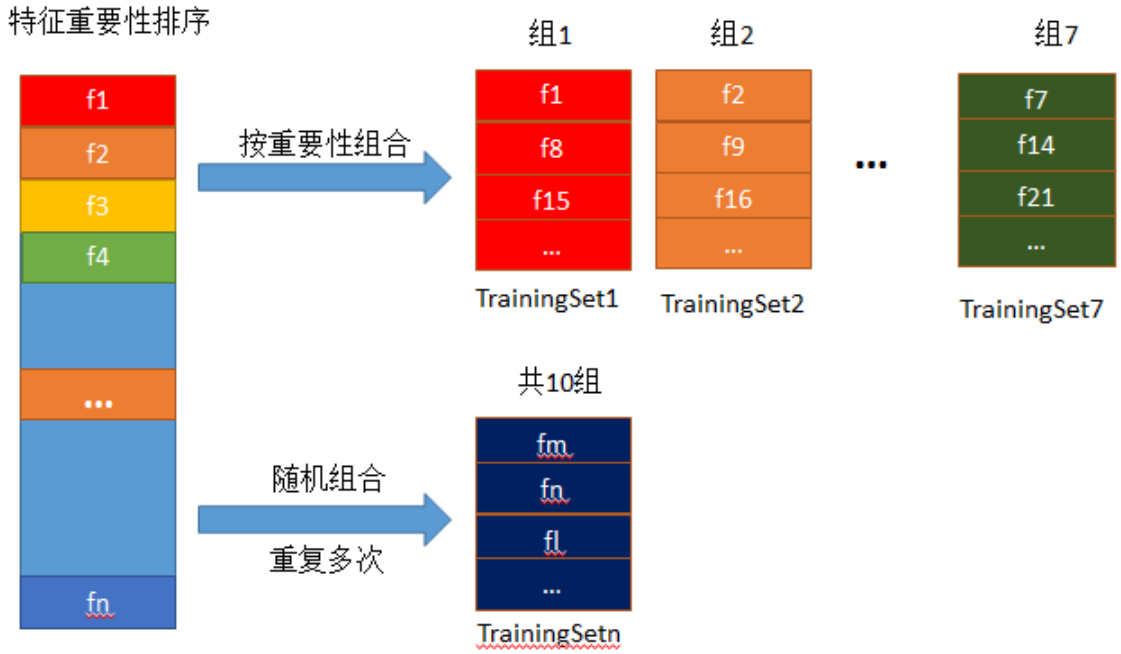
f11: {0,0,1}
f12: {0,1,0}
f13: {1,0,0}

2. 对数值型特征进行归一化以及标准化处理

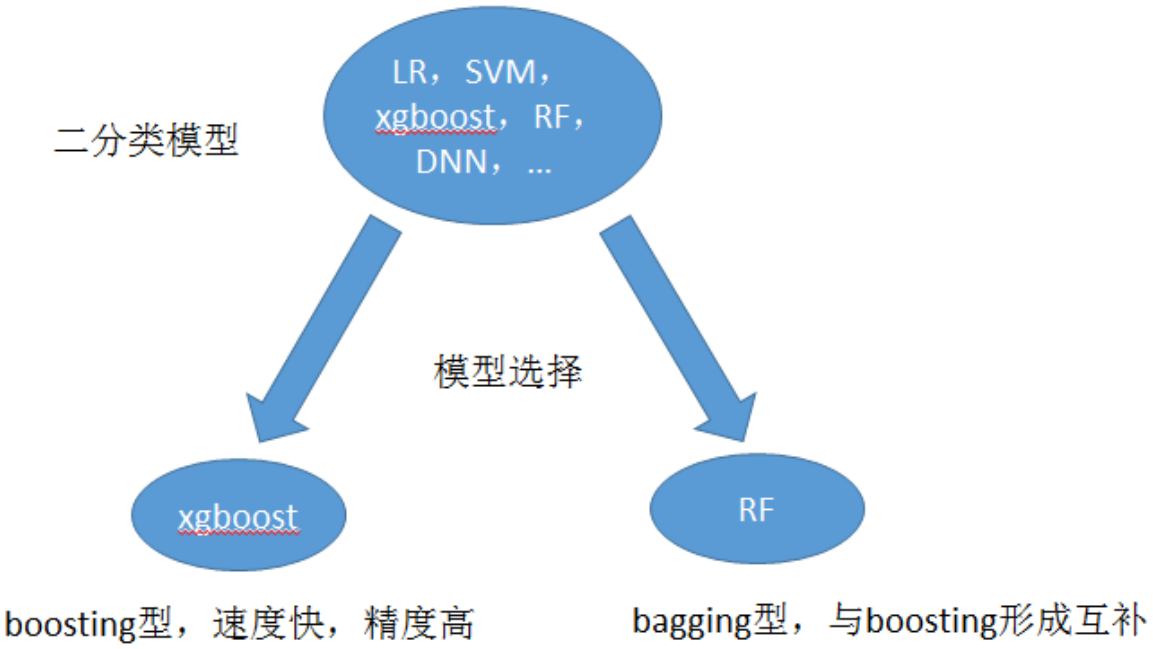
3. 缺失值填充，采用中位数

4. 正负样本比控制，正负例倒置

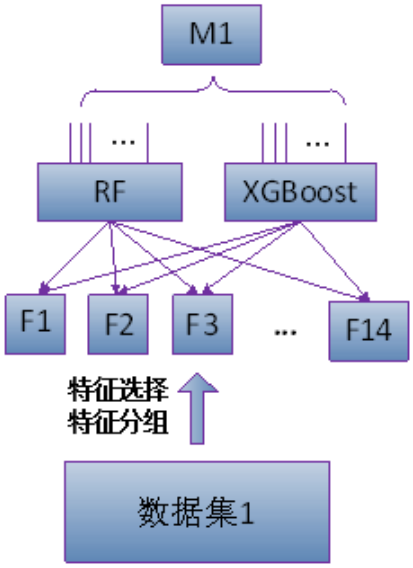
特征分组


特征重要性排序
f1
f2
f3
f4
...
fn
按重要性组合
组1
f1
f8
f15
...
TrainingSet1
组2
f2
f9
f16
...
TrainingSet2
...
组7
f7
f14
f21
...
TrainingSet7
随机组合
重复多次
共10组
fm
fn
fl
...
TrainingSetn

模型选择


二分类模型
LR，SVM，
xgboost，RF，
DNN，...
模型选择
xgboost
RF
boosting型，速度快，精度高
bagging型，与boosting形成互补

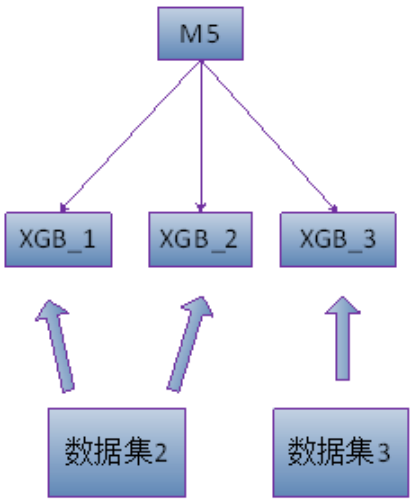
模型融合



均值融合：

$$score = \sum_{i=1}^{n} \frac{\text{score}_i}{n}$$

模型融合

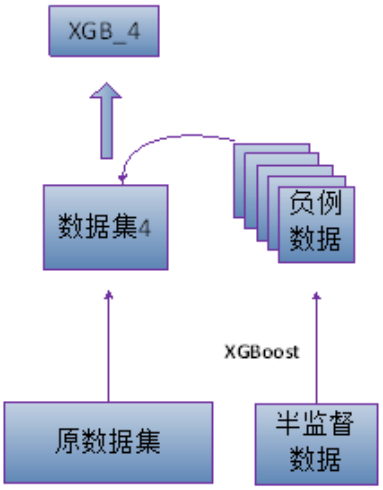


rank加权融合：

$$score = \sum_{i=1}^{n} w_i * \text{rank}_i$$

$$M5 = 0.5 * XGB\_1 + 0.28 * XGB\_2 + 0.22 * \text{XGB\_3}$$

半监督学习



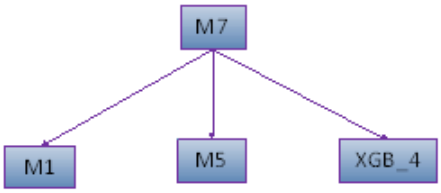
题目中的负例数据较少，解决这个问题：

- 通过重抽样方式
- 通过半监督学习添加负例数据

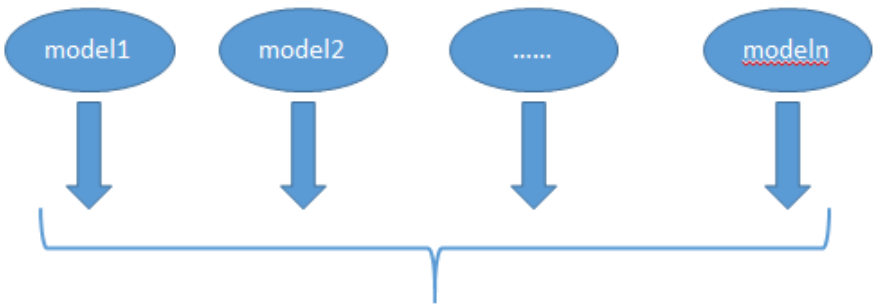
步骤：

1. 用最好单模型xgboost对半监督进行评价，取AUC分数最低的Top5000作为训练样本负例保存。
2. 将负例样本分成5组，每次添加一组数据到原本的训练集中组成新的训练集
3. 采用新的tune好的xgboost进行训练和测评，取线上效果最好的那组数据作为今后的训练数据集

模型融合



1/rank加权融合（按score降序）：



$$score = \sum_{i=1}^{n} \frac{w_i}{\mathrm{rank}_i}$$

总结与展望

总结：

1. 算法框架能集成多种模型，效果好于最佳单模型；
2. 利用模型间差异性，rank融合比均值融合在提升AUC效果更好；
3. 通过特征分组，运用不同模型产生不同meta-feature能加以融合；
4. 半监督学习通过增加负例数进一步改进模型效果。

展望：

1. 需要对特征进行更详细的分析，如果赛方能提供更具体的业务知识，结合这些知识进行相关特征工程工作，进一步提升单模型效果。
2. AUC评价涉及相对排序优化问题，可以考虑采用Learning to Rank算法中LambdaMart与XGBoost结合进行排序优化。
3. 数据特征高维、稀疏且含有缺失值，可以考虑采用目前流行的深度学习DNN中的稀疏自编码技术进行特征选择降维。
4. 尝试使用蚂蚁金服内部广为使用的bilinear模型，能够大大提高分数。

致谢

现金巴士 | CashBUS
微额速达（上海）金融信息服务有限公司

回复

作者

(/user/5451) DataCastle运营 (/user/5451)

DC币: 428

无人回复话题

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)

用户人品预测大赛--挖掘业务队--竞赛分享 (/topic/17416447cdab4bd5ad6a4bc00053f91e.html)

作者其他话题

用户人品预测大赛获奖团队分享 (/topic/58870500b2f84ddb9cbd4f6a45f180df.html)

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)