

用户人品预测大赛--getmax队--竞赛分享

加入收藏

•发布于 2016-03-24 14:49 •作者 [DataCastle运营 \(/user/5451\)](#) •29 次浏览 •来自 [微额借款用户人品预测大赛 \(/?tab=148\)](#)

参赛队队名：getmax

竞赛报告书

一、参赛作品概述

随着互联网的发展，互联网金融已成为当前最热门的话题，包括支付、理财、众筹、消费等功能在内的各类互联网金融产品和平台如雨后春笋般涌现。互联网金融是传统金融行业与互联网精神相结合的新兴领域，是对传统金融行业的有效补充，因此互联网金融的健康发展应遵循金融业的基本规律和内在要求，核心仍是风险控制。

传统金融的风险控制，主要是基于央行的征信数据及银行体系内的生态数据依靠人工审核完成。在国内的征信服务远远不够完善的情况下，互联网金额风险控制的真正核心在于可以依靠互联网获取的大数据。而机器学习将是大数据时代互联网金融企业构建自动化风控系统的利器。

从机器学习分类来看，大致可分成监督学习、无监督学习、半监督学习以及强化学习等。在企业数据的应用场景下，监督、无监督的使用非常普遍。近来来，半监督学习也越来越盛行。此次的互联网征信数据竞赛，我们普遍采用监督、半监督等相关技术。值得一提的是，在征信的数据环境，不难发现标签数据普遍较少，因此采用半监督技术可以有效改善模型的质量。

在机器学习领域，特征工程、模型融合以及参数优化是比较重要的技术环节。我们重点介绍我们在这些方面的做法：

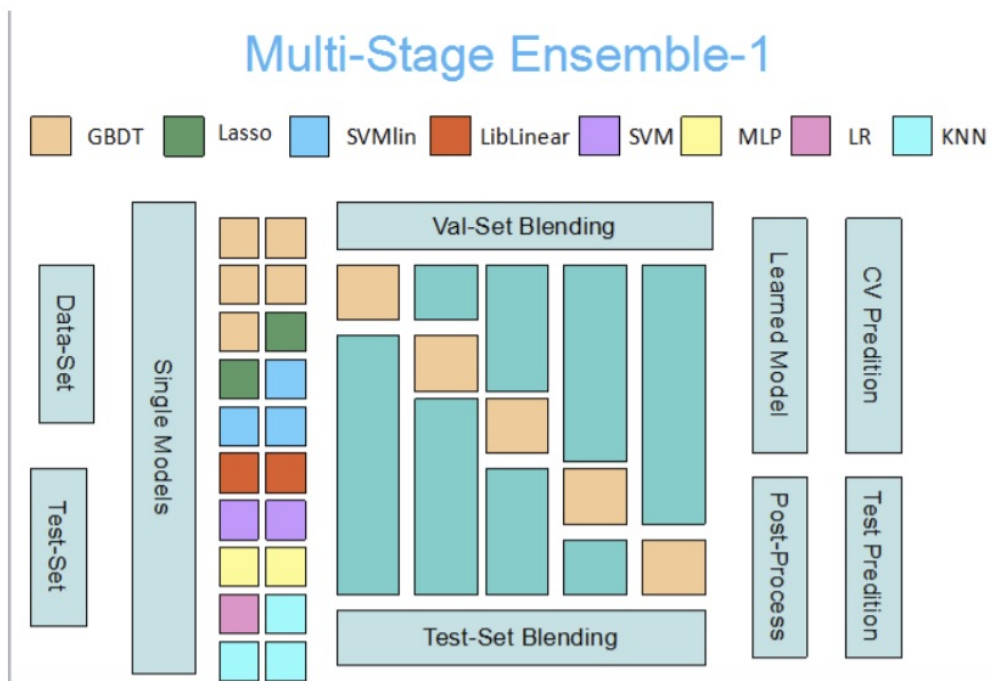
在特征工程方面，我们运用特征选择、抽样技术、特征组合来、异常点处理等方法进行数据加工。比如我们通过 T 检验、ANNOVA 检验等常用的统计手段选择统计显著的特征，也考虑分层抽样、Bagging 等技术划分训练集，也采用模型训练、遗传算法返回最优的特征子集。另外，我们也会深入数据，分析和研究数据特点，生成衍生变量等等。

在模型融合方面，我们尝试 Linear Blending 以及 Non-Linear Blending 等方法提升单个分类器的准确度。在构建分类器时，我们也尝试了多种方法，比如树模型，逻辑回归，支持向量机，最近邻算法、神经网络等。在实验过程中，我们发现基于 BOOSTING 的 GBDT 模型的效果较好。

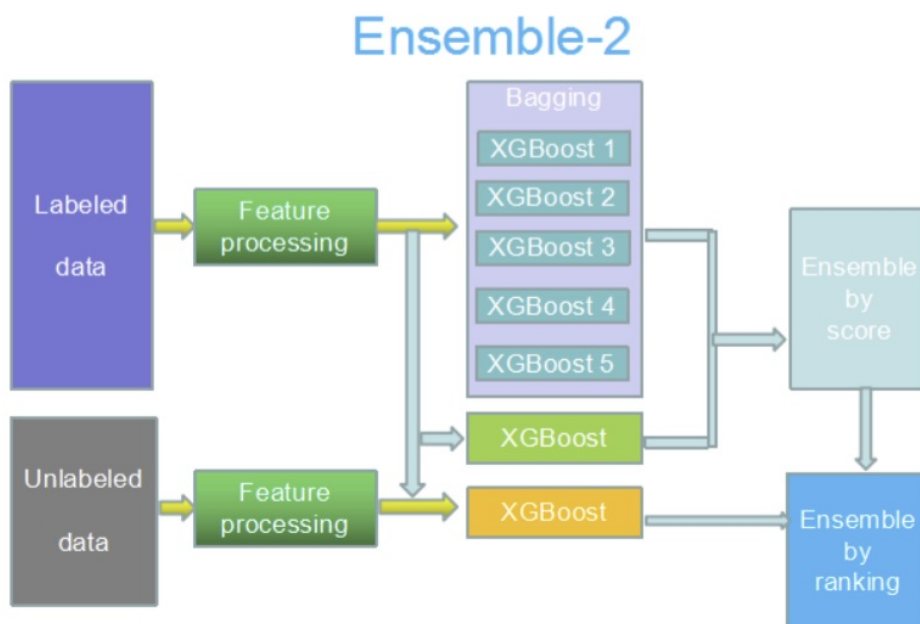
在参数优化方面，我们尝试对特征子集以及模型的超参数进行优化，主要通过网格搜索以及遗传算法。网格搜索和遗传算法都是多维空间搜索近似解的办法，其中遗传算法表现更出色，通过设置交叉和变异系数，可以得到理想的结果。

此外，我们也尝试通过半监督学习改进模型的表现。比如，我们通过 Self-Training、转导学习等相关技术扩充训练集的样本量，起到优化模型的目的。

二、参赛作品技术路线

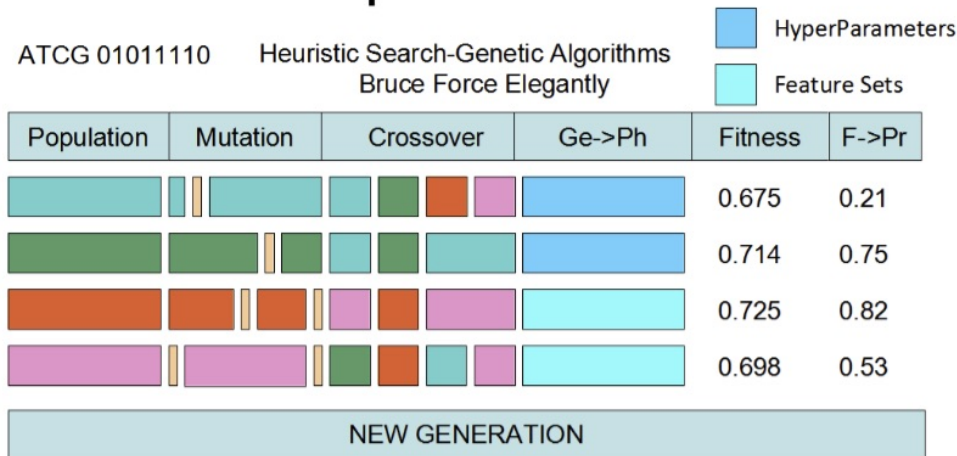


该图为参赛作品的 ensemble-1，我们采用 GBDT、Lasso、SVMLIN、LibLinear 等训练单个分类器 $g(x)$ ，再通过 Blending 技术实现多模型的融合 $G(x) = \sum w_i g_i(x)$ 达到提升模型精度的目的。



该图为 ensemble-2，其中仅仅采用了 XGBoost 模型，但是我们利用了 bagging 融合以及无标签数据，最后融合结果还是比较好的。

Optimization

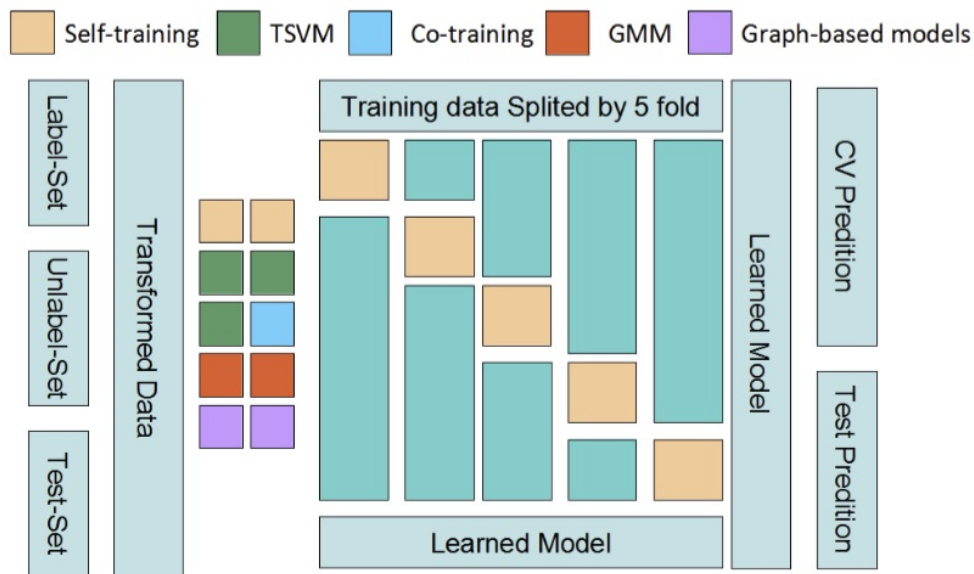


$$P_i = \frac{f_i(x)}{\sum f_i(x)}$$

The function of fitness can be defined as the mean auc of cv prediction.
And P_i means the probability of survival into next generation.

该图为参赛作品参数优化的架构图，我们运用在特征子集的选择、GBDT 种子的搜索、模型超参数的选择等场景。

Unlabel and Label Training



该图为参赛作品半监督学习的架构图，由于时间关系，我们只实现了 Self-training 以及 TSVM 部分。

实验结果：

online score

Model	score
ensemble-1	0.7243
ensemble-1	0.7254
ensemble-3	0.7266

三、作品总结

参赛作品的模型架构比较清晰，对于特征工程、模型融合以及参数优化等机器学习的典型问题给出了相应的解决方案。其中通过 Self-training 的方法构建半监督模型，使得模型融合的准确率有显著提升。


同时，我们也发现对特征工程的处理是不够的，特别是对征信数据的理解，衍生变量的生成等方面做得不够好。

Machine Learning in Credit Quality Prediction

By GetMax



Outline

- Team Members
 - Understanding and Analysis & Data Partition
 - Feature Engineering
 - Single Model Framework
 - Semi-Supervised Learning
 - Multi-Stage Ensemble
 - Public Score
- 

Understanding and Analysis

- Classification & Rank Problem
- The data is not much to learn the global classification
- Label imbalance and few data labeled
- Value Missing

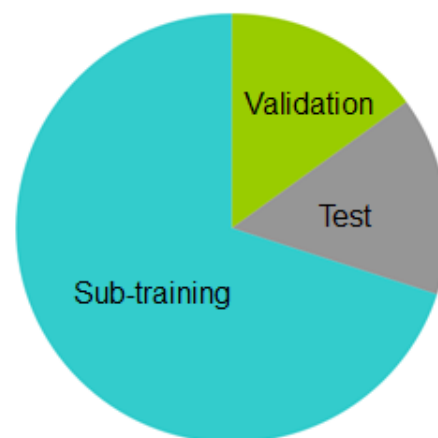
Data Partition

Sub-Training:Validation:Test=7:1.5:1.5

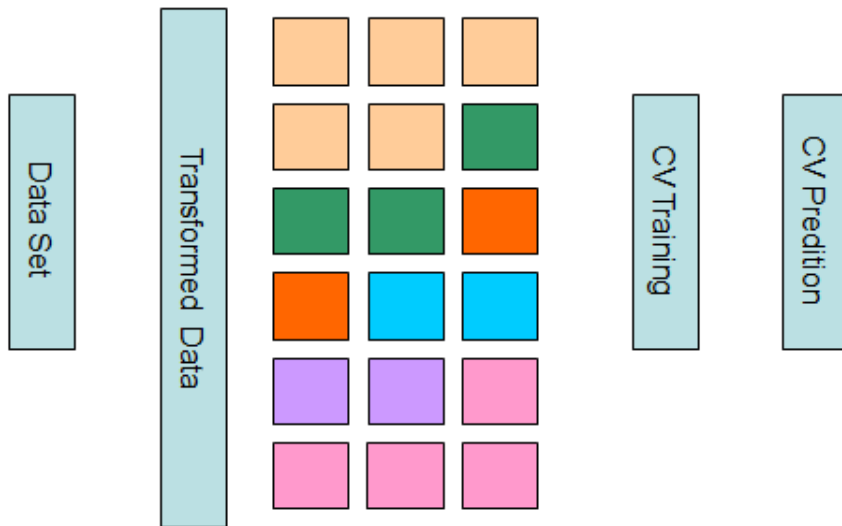
Sub-training : used to learn the parameters of models.

Validation : used to train the weight of several models in the blending stage.

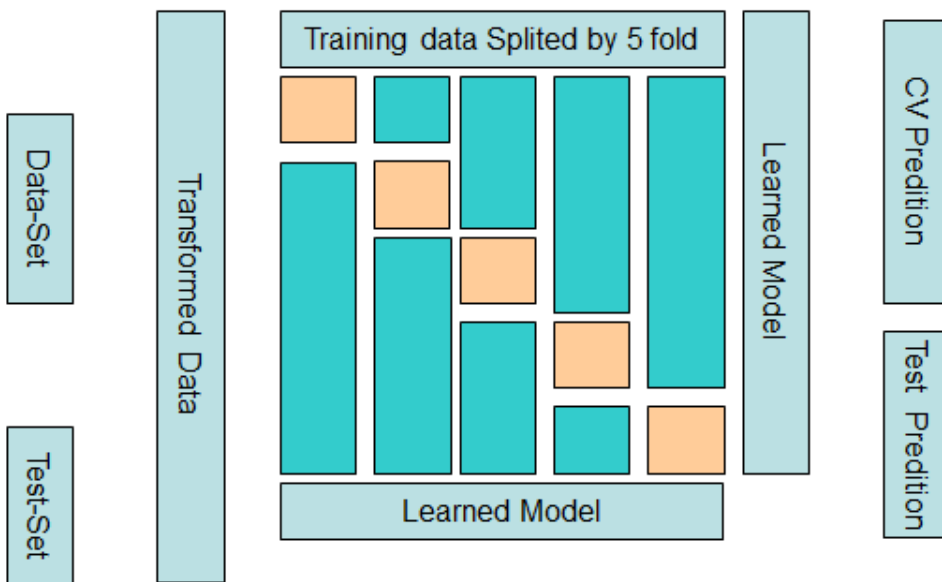
Test : used to learn the hyper-parameters of models.



Feature Engineering



Single Model Framework



Supervised Learning

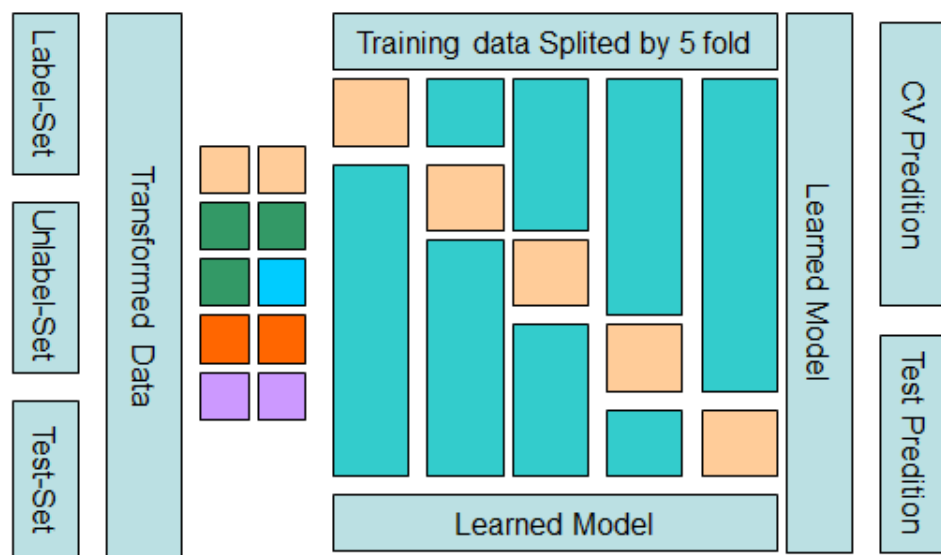
Abbreviation	Algorithms
GBDT	Gradient Boosting Decision Tree
Lasso	Least Absolute Shrinkage and Selection Operator
SVMlin	Semi-supervised learning
LibLinear	A Library for Large Linear Classification
SVM	Supported Vector Machine
KNN	K-Nearest Neighbor
MLP	Neural Network-MLP with 3-Layer
LR	Logistic Regression with L1 norm

Semi-Supervised Learning

Method	Decription
Self-training	Give labels to unlabeled data until convergence
Co-training	Give labels to unlabeled data until convergence
Generative models	EM with generative mixture models
Transductive SVM	Maximize the unlabeled data margin
Graph-based models	Construct a graph based on labeled and unlabeled data, propagate labels along the paths

Unlabel and Label Training

■ Self-training
 ■ TSVM
 ■ Co-training
 ■ GMM
 ■ Graph-based models



Optimization

ATCG 01011110

Heuristic Search-Genetic Algorithms
Bruce Force Elegantly

■ HyperParameters
■ Feature Sets

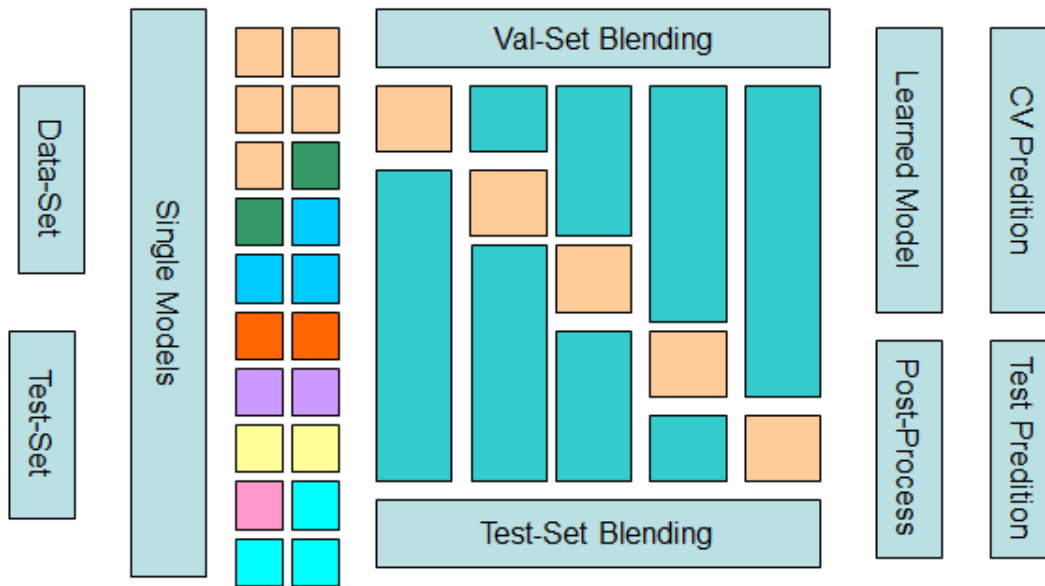
Population	Mutation	Crossover	Ge->Ph	Fitness	F->Pr
■	■ ■	■ ■ ■ ■	■	0.675	0.21
■	■ ■ ■	■ ■ ■	■	0.714	0.75
■	■ ■ ■ ■	■ ■ ■	■	0.725	0.82
■	■ ■ ■ ■ ■ ■ ■	■	■	0.698	0.53
NEW GENERATION					

$$P_i = \frac{f_i(x)}{\sum f_i(x)}$$

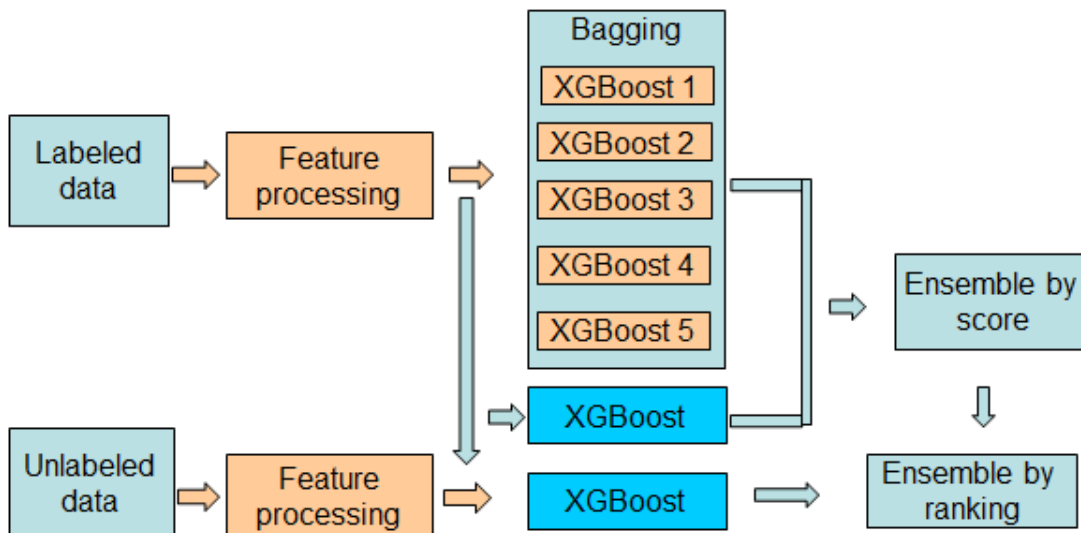
The function of fitness can be defined as the mean auc of cv prediction. And P_i means the probability of survival into next generation.

Multi-Stage Ensemble-1

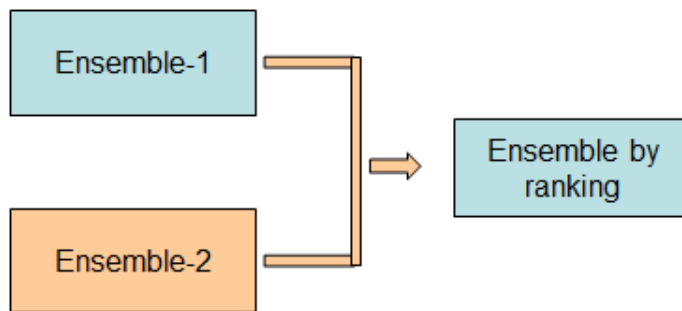
GBDT Lasso SVMlin LibLinear SVM MLP LR KNN



Ensemble-2



Ensemble-3



Public Score

Model	Public Scores
XGBOOST	0.7236
LASSO	0.6720
MLP	0.6635
....
Ensemble-1	0.7243
Ensemble-2	0.7254
Ensemble-3	0.7266

Thank you for your patients
and congrats to all winners!



0 回复

添加回复 注:回复会奖励1点DC币，但被管理员删除回复，将扣除作者2DC币;可以使用@符号回复其他人

</>

B

U

I

S

“

f v

Tl v

H v

>_

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--火星队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

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

关于我们

服务条款
(http://www.pkbigdata.com/page/html/common/tos.html)

隐私协议
(http://www.pkbigdata.com/page/html/common/privacy.html)

商务合作

联系人：周莹
电话：18300526662
邮箱：ying.zhou@hirebigdata.cn

DC QQ群

名称：DataCastle
群号：423732457

DC 微博

名称：DataCastle

我们的朋友

wangEditor
(http://wangeditor.git)



DC微信公众号

我们的客户

(http://www.pkbigdata.com/page/html/user/clients.html)

联系我们

(http://www.pkbigdata.com/page/html/message/contactUs.html)