

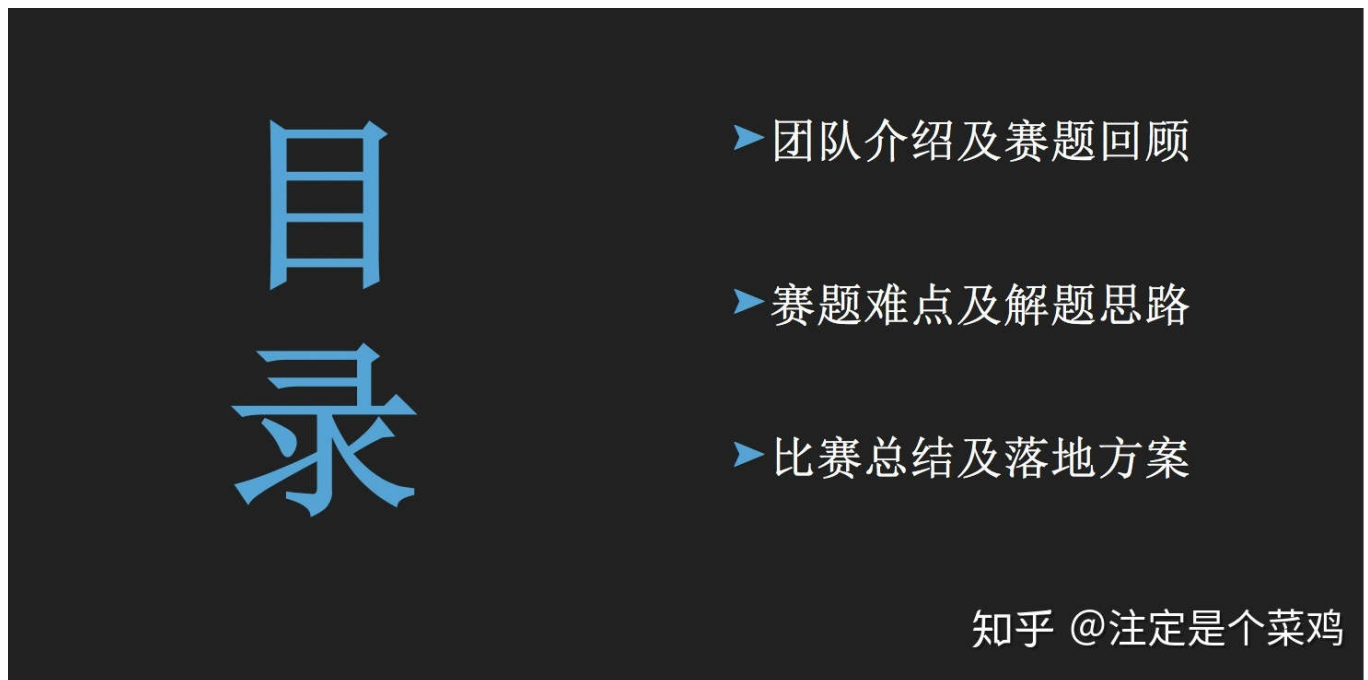
第三届融360天机-智能金融算法挑战赛拒绝推断赛题--TOP2方案



已关注

我们团队：‘一壶浊酒’，A榜第二，B榜也是第二真心遗憾，不尽人意。

拒绝推断赛题相对来说没有太大的难度，因为是匿名特征，更多的是常规操作，下面是我们团队的PPT，以及针对PPT的分享，和大家一起学习进步。



讲解分为三部分：1团队介绍及赛题回顾 2赛题难点及解题思路 3比赛总结及落地方案

团队介绍：‘注定是个菜鸡’，中科大研三鹏哥大佬，西南大学研二胡敏强力队友。

赛题回顾：

题目链接：<http://openresearch.rong360.com/#/question>

模型总是倾向于允许较好的客户进入模型，拒绝掉质量较差的客户，久而久之模型缺乏对于质量较差客户的特征的训练，导致学习不到坏客户的特征，可能会引入大量坏客户入模。针对这种现象，我们需要想办法让模型能够同时兼具对于好坏客户的判断能力。

赛题回顾

- ▶ 拒绝推断这道赛题需要解决的问题是要在只有最优质的放款用户好坏标签的情况下，如何保证建模对所有放款用户和拒绝用户都有良好的排序能力。
 - ▶ 大赛以AUC为评测指标，带标签的用户训练数据约3万，无标签用户训练数据7万，不带标签的用户测试数据2万。
 - ▶ 团队主要从两个方向入手
 - ▶ 机器学习模型
 - ▶ 拒绝推断模型
 - ▶ 两个方向的模型融合，最终排名靠前，唯一一支AB榜同时进入前五的队伍
- PS：团队同时在赛题三取得了第三的成绩，这也证明团队模型的稳定性，算法的鲁棒性较好。

拒绝推断				
排名	预测评分	参赛队伍	所属单位	提交时间
1	0.8387	天线宝宝	滕条焖猪肉	2018-10-29
2	0.8387	一查造酒	数据库跑路	2018-11-08
3	0.8385	人民的儿子	夜游部门	2018-11-07

拒绝推断				
排名	预测评分	参赛队伍	所属单位	提交时间
1	0.8324	bee	null	2018-11-10
2	0.8322	一查造酒	数据库跑路	2018-11-10
3	0.8322	jph	上海财经大学	2018-11-10

知乎 @注定是个菜鸡

赛题难点：

- ▶ 针对大量匿名特征的特征选择和特征工程方法

f1	f2	f3	f4	f5
0.098	32.5	1	0.02139	4907
0.043	322	1	0.015625	2496
0.06	42.5	1	0.333333	1759
0.212	28	1	0	3845
0.089	57	1	0.031579	1931

- ▶ 针对大量无标签数据的处理方法

知乎 @注定是个菜鸡

data mining的处理点

► 模型构建

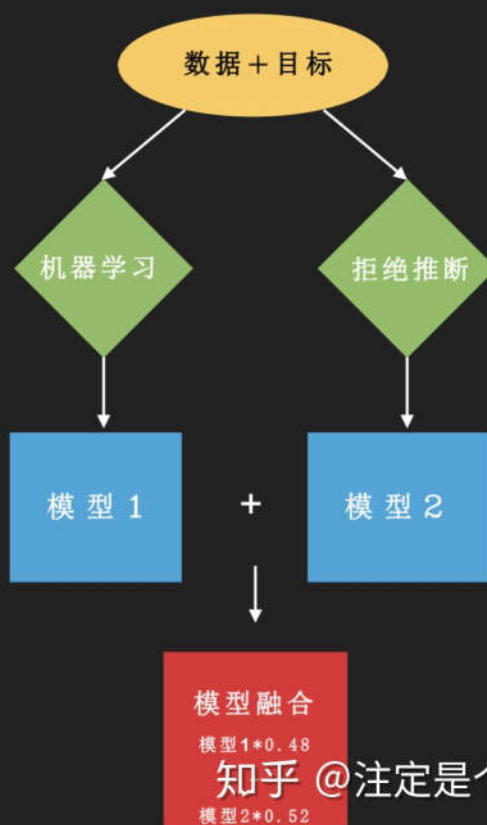
- 预处理
- 特征
- 样本
- 模型
- 参数
- 模型融合

思考方向

知乎 @注定是个菜鸡

解题思路：针对两难点，打造准而不同的两个模型。下面从上面的6项要点逐一介绍，并着重介绍两大难点。

- 不同的思考模式，保证模型的差异性大
- 简单的模型，保证模型解释度高
- 简单的融合方法
- 最好的效果

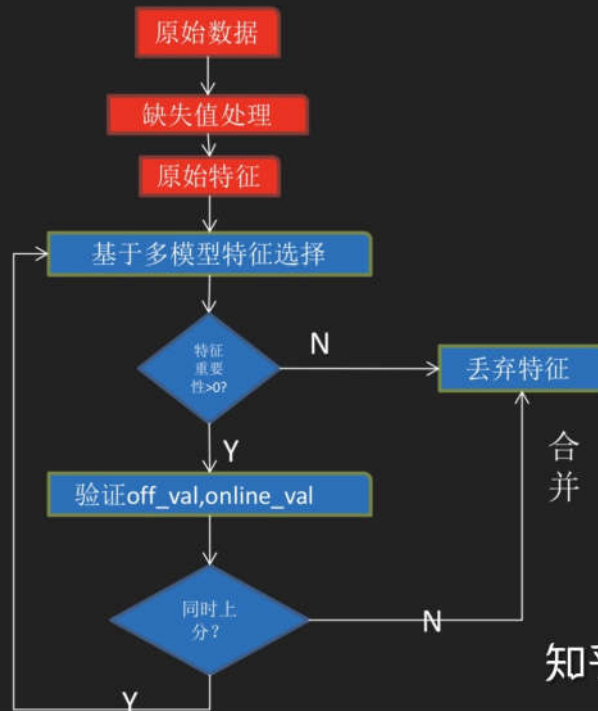


知乎 @注定是个菜鸡

首先特征选择，由于是海量匿名变量，我们需要确定数据探索的顺序，那么我们可以根据feature_importance的顺序来探索。

► 多模型筛选最优特征

由于树模型筛变量将相关性高的特征只输出一个高的分数，所以团队采用多模型多次筛选特征的方法。



知乎 @注定是个菜鸡

简单浏览数据的大致分布

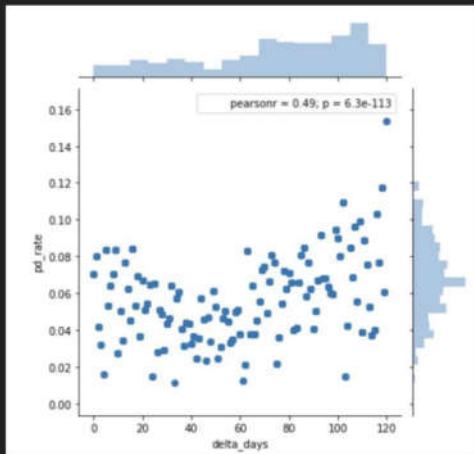
匿名特征的信息分布

	mean	std	min	25%	50%	75%	max
f1	28.224	5.483	0	24	27	31	98
f2	0.444	2.171	0	0	0	0	16
f3	3.325	0.720	1	3	3	4	4
f4	51762.428	5338536.859	0	5000	10000	20000	999999999
f5	2.646	0.985	1	2	2	4	5
f6	0.506	1.279	0	0	0	0	5

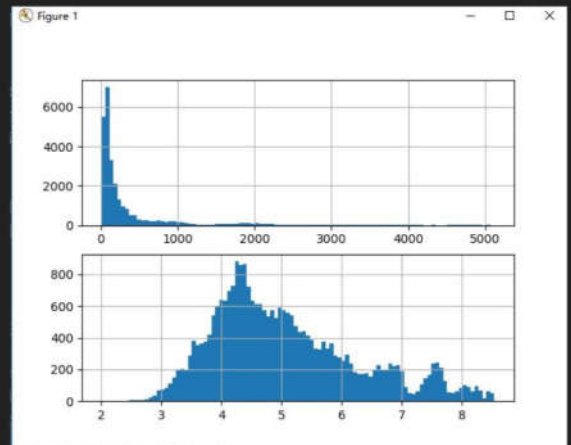
知乎 @注定是个菜鸡

按照特征重要性，依次进行离散特征的去噪或连续变量的平滑等处理。对时间的观察也十分重要，通常来讲我们需要考虑在时间轴上是否存在日，月，年，星期或月初月末等周期性问题，以及考虑节日所带来的影响。本题中我们发现样本数量和逾期率都成U型，猜测是策略所致。时间在金融风控中经常属于不稳定因素，经常受到市场和政策变化而受到影响，我们在本次比赛中考虑到商业应用，没有入模。

数据探索



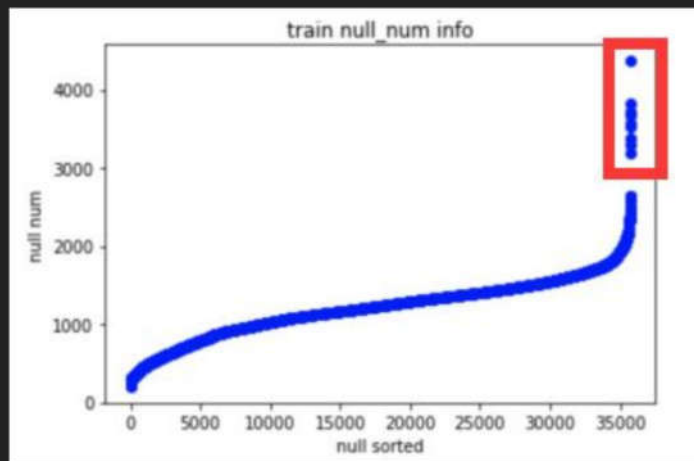
图一 时间与违约概率图



图二 数据长尾分布图

我们发现无论如何做特征选择，统计样本缺失数量的时候，一直都存在一些噪声，去除噪声后，效果有明显提升。同时我们针对缺失值给予一个没有的数，当做一种类别来处理。尝试离散变量和连续变量分开填充可能效果会更好。

数据探索



图三 特征缺失分布图

在数据探索的过程中发现，赛题数据中大部分样本都有缺失值，且缺失值个数较多，有的样本甚至有上千个缺失值。我们可以观察发现有一部分数据缺失值过多，删掉此部分噪声数据，防止干扰到训练模型效果。

模型、参数：我们团队尝试了多种模型，和两种使用参数的方案，最终来确定最优方案。

模型

尝试lightGBM, XGBoost, CatBoost, DNN等多个模型

参数

使用最优定参和浮动参数两种方式同时保证精准度和鲁棒性，
增强模型的鲁棒性

知乎 @注定是个菜鸡

特征：我们团队针对匿名特征工程，提出了启发式的二次特征工程方案，其中包括统计类特征，交叉特征，描述类特征。如果在有业务意义的情况下，建立特征之间的联系会更好的提升模型效果。

机器学习模型

▶ 2级特征构建

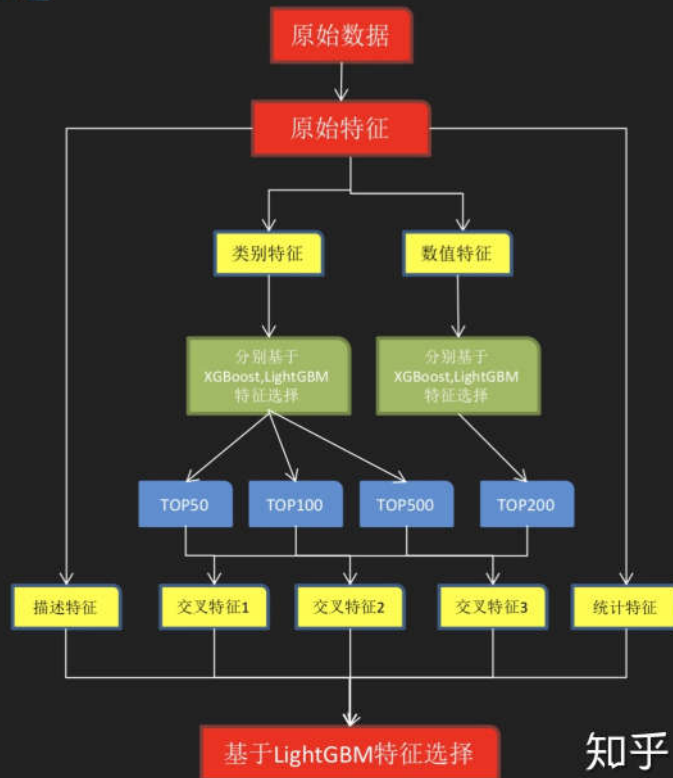
▶ 原始特征扩展

	统计类特征	交叉特征	描述类特征
f1	类别特征占比情况	特征相乘	数值特征最大值
f2	类别特征出现次数	特征相除	数值特征最小值
f3	类别特征唯一取值	特征相加	数值特征取对数
f4	类别特征风险率	特征相减	

知乎 @注定是个菜鸡

下面是我们在二次特征工程的方案，离散变量尝试特征重要性TOP50，100，500三种情况，连续变量使用特征重要性TOP200，与离散变量做交互。

► 组合特征的筛选

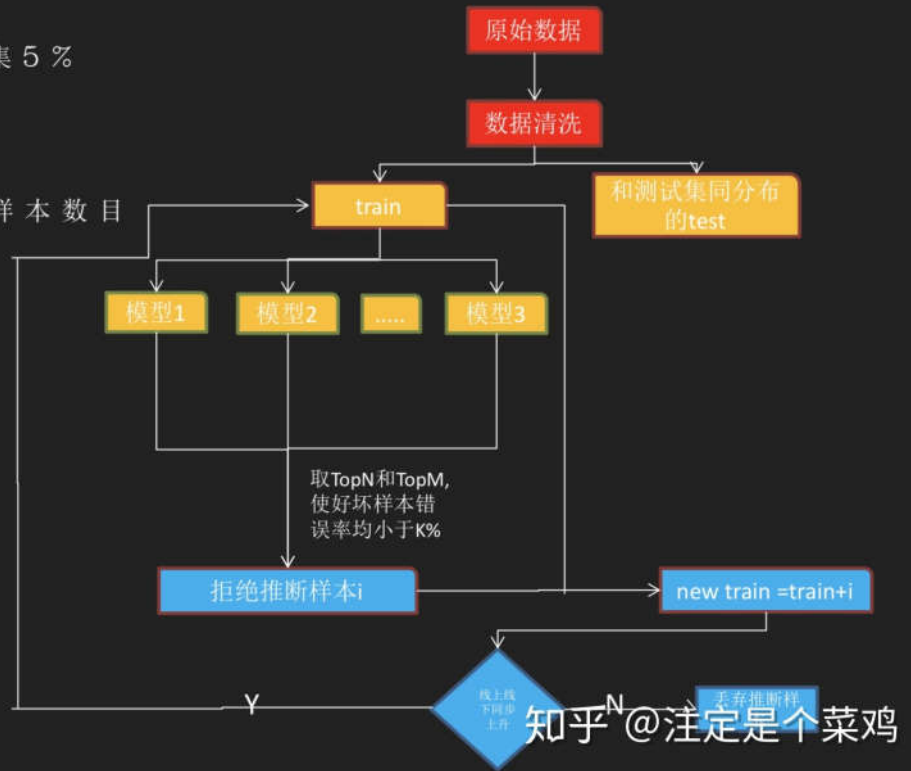


知乎 @注定是个菜鸡

样本：如何使用大量的无标签样本也是本题的重中之重，这里提出recursive sample selection方法，在构造的与A榜同分布的offline_val中，通过多模型确定offline_val的最优阈值，我们团队在本题中确保预测错误率在1%以下。最后用构造出的复合模型预测unlabel，来保证unlabel的最优预测。以此方案反复循环，不断将最新样本入模保证使用最优的复合模型来预测unlabel，知道达到最优样本。(在此过程中，很可能会出现一批样本针对某参数有极好的预测能力，但我们希望是一批稳定预测的样本，所以在多组参数下依然稳定的样本，才可以被确定为最终入模的unlabel)

拒绝推断法

- 多模型投票取交集 5 %
- 两轮拒绝推断
- 按比例添加好坏样本数目



融合：高差异性，准而不同的两模型根据线上表现加权融合。我们在排序中看差异性经常更多的关注斯皮尔曼。

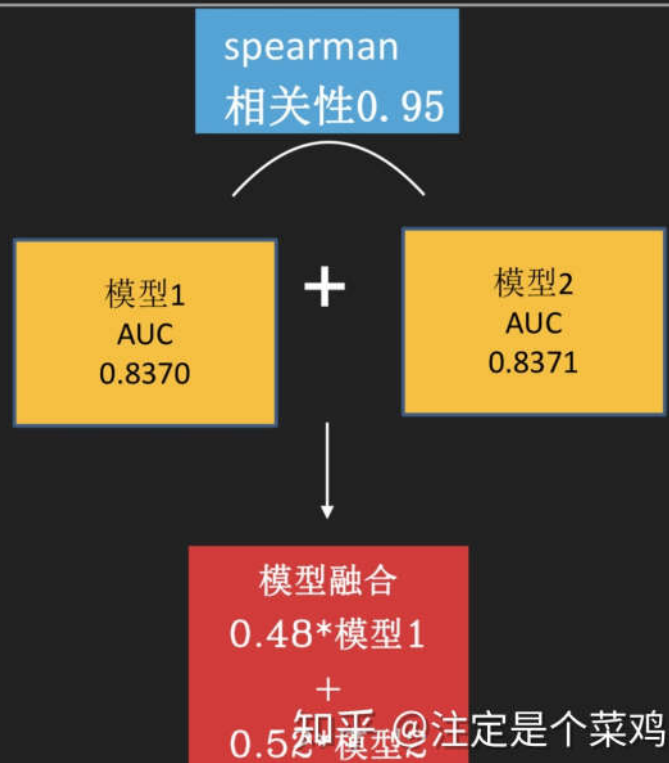
加权平均

- 依据线上模型成绩
- 0.8370
- 0.8371

采用 rank 加权融合

- 0.48
- 0.52

进行融合，最终成绩
0.8387



创新点及落地方案：

► 创新点

- 启发式特征，对数据的二次特征
- 提出递归样本选择方案，选取错误率最小的无标签样本集进入到训练集。

► 落地改进方案

- 调整P次K折，降低模型部署难度和时间复杂度
- 在匿名特征转化为已知业务含义的特征时根据真实情况使用一部分的二次特征，减少引入的噪声，提高模型稳定性和精准备性

知乎 @注定是个菜鸡

针对这次拒绝推断赛题，我们团队在常规处理的基础上，针对赛题两难点分别做了处理，在某一方面上实现了风险对冲，所以模型稳定性较强，也是赛题一中唯一——一个AB榜同时进入前十的队伍。欢迎点赞，谢谢各位大佬

赛中分析

Hallo大家好，我是‘干不过各位大佬’，前几天参加了融360举办的第三届天机杯的风控比赛，由于时间不是很充分，只参加了第一题，目前单模成绩8365，暂时排在第二，考虑到时间成本和模型易部署性，一直只用一套特征和一个lightgbm上分，最近又有好多天池大佬入场，简直是慌的一批，赶紧出个思路和大家一起学习下，然后融融模型就准备弃赛了。说指导的话完全谈不上，只是希望可以做个分享和大家互相学习，共同进步。

接下来想说一下，之后本专栏会专门针对比赛做一些分享或给出baseline，算是一种个人思路的整理，也算是一种回馈吧，主要写手是我同学们天池中经常前排的一拨人(隔壁深度学习写手是从半年多前进入比赛届后三连冠，从不做大佬，几天后会有几个深度学习比赛分享，大家之后可以多关注下深度学习大佬的文章。)当然如果侥幸我这种菜鸡也不小心拿了前排自然也会分享出来，希望和我们一样的奋斗者们尽快的也能在这个领域有所斩获，尽快成功。这样做的原因其实有很多，其一是对自己思路的整理同时也能帮助到一部分和我们一样的人，我认为这是一件非常开心并且有成就感的事情，其二是我们几个小伙伴其实也是受益于各种开源的代码和思路，才逐渐成长，虽然不是完整的一个系统和套路，但零零散散的收获也是要感谢乐于分享的这些大佬们。所以也欢迎小伙伴们关注。

那么接下来废话少叙了，说一下这道题

融360中的第一道比赛题目是拒绝推断，首先先了解下题目业务内容和数据情况：

题目链接：<http://openresearch.rong360.com/#/question>

赛题的数据可以在题目链接中获取，训练样本10000W（由33465有标签样本和66535的灰样本组成），A榜测试集数据2W，B榜测试集数据2W

2016年1月，机构A通过自建风控模型开始放贷，初期获得了良好的收益。随着时间的推移，机构A发现在样本通过率5%不变的前提下，机构逾期率由2016年1月的5%逐步升至2017年7月的15%，大量坏账导致机构A由盈利

陷入亏损境地。公司模型人员仔细检查模型，发现其在训练集和测试集上都表现很好，并没有任何异常，百思不得其解。

在金融信贷场景中，放款机构会通过模型评分筛选用户，评分较好的用户可以获得放款，评分较差的用户直接被拒绝，机构只能获得放款用户样本的好坏标签，对于大量拒绝用户的还款情况无法获得。随着时间的推移，机构手中的训练样本都是“评分较好”的通过用户，而没有“评分较差”的拒绝用户，由此训练的模型在“评分较好”用户中表现越来越好，在“评分较差”用户中却无法得到任何验证。但是，金融风控模型真实面对的客群却包括了“评分较差”的用户，模型在“评分较差”用户中无法得到验证，导致训练的模型越来越偏离实际情况，甚至通过了大量应该被拒绝的坏用户，致使大量坏账出现，直接带来巨大经济损失。因此，在只有最优质的放款用户好坏标签的情况下，如何保证建模对所有放款用户和拒绝用户都有良好的排序能力，是金融风控模型需要解决的重要问题。解决该问题可以是传统的拒绝推断技术，也可以尝试其他机器学习技术，参赛者可自行选择。

1. 具体内容

训练样本：从2018.1.1到2018.5.1放款用户样本，信用评分top30%的样本给出每个样本是否逾期，后70%样本只有3000个给是否逾期。（约10万样本） 验证样本：从2018.1.1到2018.5.1放款用户样本，验证集不提供样本是否逾期，参赛选手自行完成是否逾期预测后，可以提交至比赛平台评估结果。（约2万样本）

测试样本：与验证样本来源相同且同分布。测试集不提供样本是否逾期，参赛选手只能在比赛最后的评比阶段将预测结果提交至比赛平台评估，且只能提交一次。（约2万样本）

2. 我们需要你们完成

对用户各类信用相关数据进行分析处理，挖掘数据价值。根据验证样本和测试样本的样本特点，从训练样本中选取合适的训练集，完成建模，保证模型在验证样本和测试样本上的效果。

于是看到题目和数据你想到了哪些？是不是有一个框架和思路在脑海里，看到题目我想到了以下一些问题，如果有缺漏或错误欢迎补充和纠正哈。

1.数据分布看了没。训练样本中大部分都是打分靠前的，只有3000个是打分靠后并混合了拒绝样本，那么他们和验证集和unlabel之间的分布情况看了么？分布一致么？那么应该如何选取train和offline val呢？这题K折offline val很飘，那你如何保证offline val和online val的一致性来验证呢？还有我们平时看分布要注意的是为了保证特征的稳定性，是否将训练集和验证集一起看了？

2.特征这么多，是不是需要降维。我们常常通过降维来降低过拟合风险，使模型泛化能力更强，增强模型对特征之间关系的理解，但如何把降维后的特征做好非常重要，如何才能使重要信息丢失最少呢？

3.数据挖掘最重要的一步EDA做的怎么样。在做EDA的时候这么多变量应该从哪个变量开始要有自己的想法，根据EDA的结果做怎样的特征？比如有没有周期性的问题，有没有噪声，有没有可做的规则。

4.数据零散杂乱无章，噪声去了么？缺失值又是如何填充的呢，分类和连续变量填充方法是否一样，如果想分开填充如何区分分类和连续变量的呢？如果没有办法很好的区分开是不是要想其他办法？去噪声和缺失值填充都可上分

5.模型参数如何调整的？可以参考下模型理论知识，应该可以调整到较优参数，但是参数的调优不可全信，A榜高不一定B榜也高,已经深深感到换数据后要被吊打了，瑟瑟发抖在角落。。

6.模型怎么融合的都试了下么？总有一款是适合的。什么时候简单融合什么时候加权融合，stacking, blending还是排序后标准化再融，这些不同方法的特点是不同的，需要多去思考和尝试。模型融合一般我们希望准而不同，如何能做到准而不同，又十分稳定呢。防止B榜崩盘，个人不是非常建议使用高但不稳定的模型，参数变动一点点整体auc浮动很大的话，或样本或特征一定是有问题的，我们重在解决问题，尤其金融领域，稳定胜于一切，这类模型往往B榜或是实践中可能会出现各种意外情况

7.这个时候模型的预测能力很强了吧？灰样本利用了么？我们利用灰样本要尽可能的准确了解预测情况才好利用，而不是3%，5%，10%的去蒙。灰样本的利用找不好的话会很飘，这里如果可以找到合适的样本集对整体的预测会非常有利，我们的训练样本希望是能够在模型中稳定表现的。所以如何才能找到最准确而合适的灰样本也是本题的重点。

哈哈，记得一位大佬说过其实数据挖掘就是面向数据编程，我们要根据数据的状态与变化去针对性解决问题，当然套路是有的，但套路不是一层不变的，所有人都可以拿大佬们的开源框架，但成绩却有高有低，比如这道拒绝推断题，拒绝推断的方案有N多种，选择哪一种方案是更优的是需要不断的思考与验证的。所以说技术绝对不是做好数据挖掘应用的瓶颈，解决问题的能力才是做好一件事的关键。多用脑思考多动手验证，更多的思考和动手才能更好的解决问题。

以上这些个人的思考过程和想法，很多想法没有全部都验证完，如果全部都做完猜测应该会至少再高几个千，比赛结束后，会将具体实现方案分享出来，希望可以和大家一起学习讨论下，本来是写了不少自己的实现方案的，但是这两天发现群里眯着几位大佬，大佬面前就不献丑了，估计大佬出手肯定是要分分钟被干掉了。大家比赛玩的愉快咯哈哈 一起加油各位奋斗的小伙伴们，愿奋斗路上，我们一起成长。

编辑于 2018-12-13