数据科学

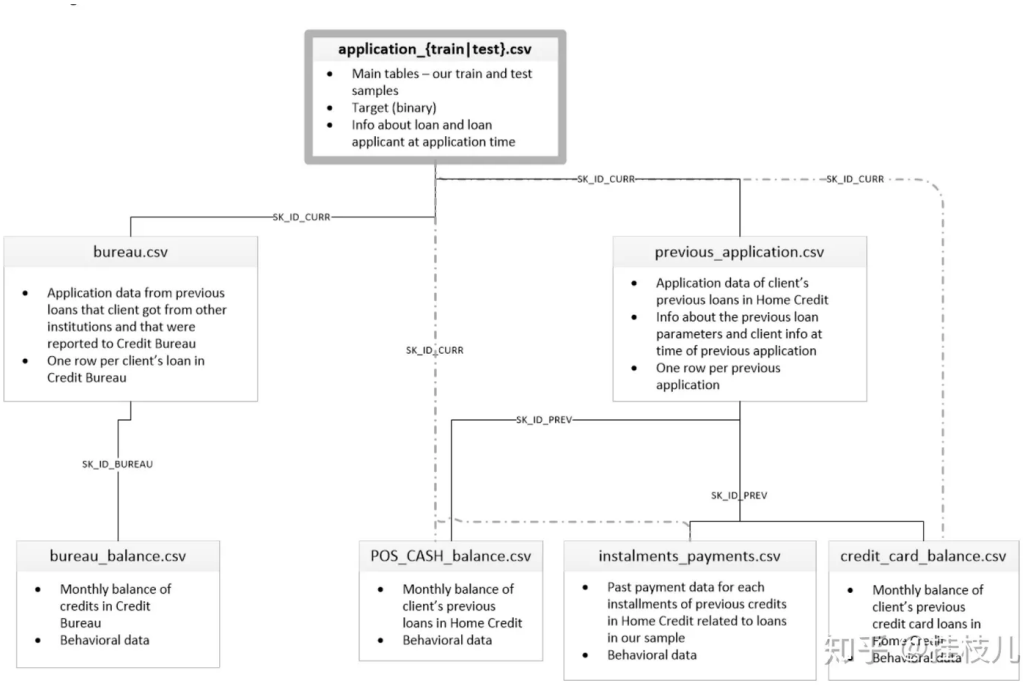
目前选题：Home Credit Default Risk | Kaggle

问题描述

❓ 你能预测每个申请人偿还贷款的能力吗？由于信用记录不足或不存在，许多人难以获得贷款。而且，不幸的是，这些人经常被不可靠的贷方利用，例如高利贷，校园贷。

捷信努力为没有银行账户的人群扩大金融包容性。为了确保这些服务不足的人群获得积极的贷款体验，捷信利用各种替代数据（包括电信和交易信息）来预测客户的还款能力。

Home Credit捷信目前正在使用各种统计和机器学习方法进行这些预测，以帮助他们释放数据的全部潜力。这样做将确保有能力还款的客户不会被拒绝，并且贷款的本金、到期日和还款日历将使他们的客户获得成功。



讨论前的需要了解的内容

1’ 了解项目目的

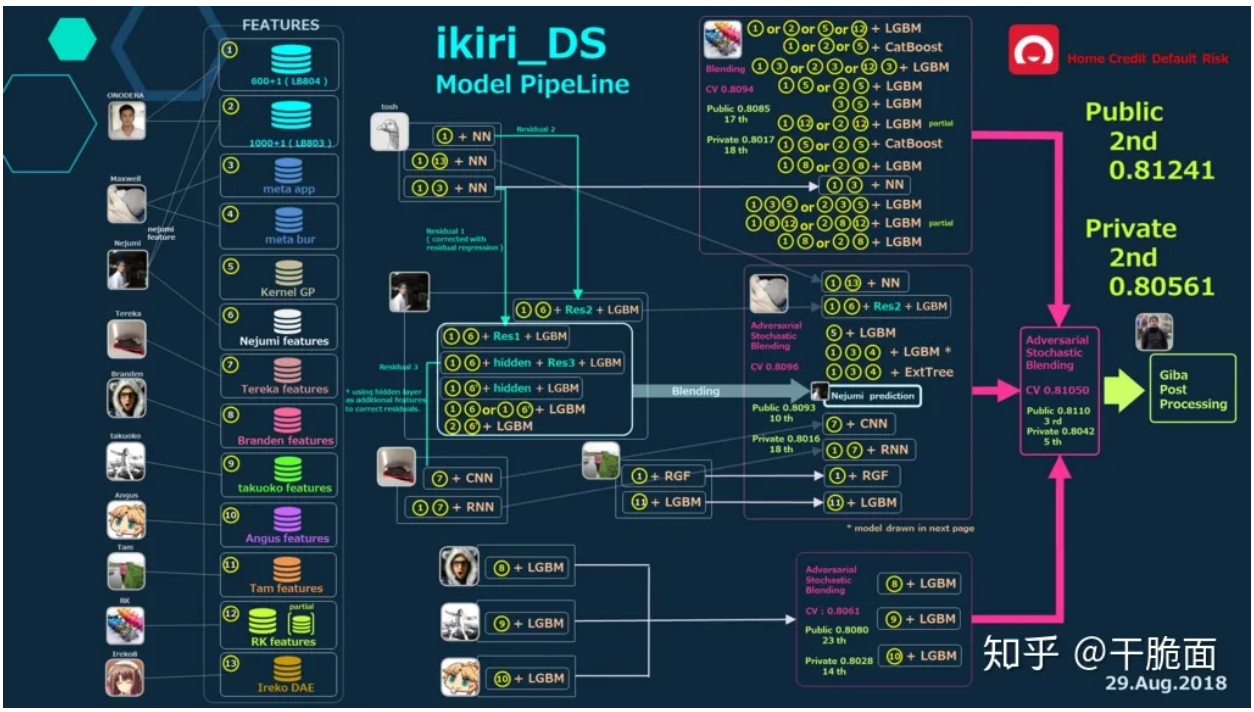
2‘ 了解数据内容、特征

3’ 大致了解这个数据集中的常用方法

本次希望达成的共识

4‘ 梳理本次数据科学项目的框架（至少是初步的）

已经实现的方案汇总



一个总述性的小结，里面汇总了决赛前几名的方案：Kaggle竞赛-Home Credit Default Risk小结 - 知乎 (zhihu.com)

数据特征描述：Home Credit Default Risk 违约风险预测，kaggle比赛，初级篇，LB 0.749_home credit用户信贷违约预测_Li Kang的博客-CSDN博客

国外大神的方案（翻译版）：Home Credit Default Risk - 1 之基础篇 - 知乎 (zhihu.com)

Kaggle比赛——Home Credit Default Risk | 码农家园 (codenong.com)

home credit default risk（捷信违约风险）机器学习模型复现(论文_毕业设计_作业) - 哔哩哔哩 (bilibili.com)

▼ 如何快速达成比赛

优秀的可扩展的代码框架

kaggle是一个相对于国内更OPEN的社区平台，每次比赛都有kaggler提供优秀精巧的开源代码，因此在无任何业务经验的大前提下，有一个优秀的可扩展的代码框架，无疑可以达到事半功倍的效果，本次比赛我是从一个开源的PB成绩0.774的代码框架开始做的，这份代码让我了解到了以下几件事情：

1、哪些特征的重要性非常高，比如说EXT_SOURCE_1~EXT_SOURCE_3、AMT_ANNUITY、AMT_CREDIT等，因此下意识的工作自然是对于这些TOP特征做一些交互的操作，比如说比值，乘积等，看是否能交互出更优秀的特征。

2、接下来的工作，我该从哪里出发，开源的代码提炼的特征虽然不多，但是框架搭的特别棒，一些开源代码未做的特征统计，可以在这个基础上进一步尝试，看线上反馈带来的效果。

通过这些尝试，你的模型在评价效果上将会有一个大幅度的提升。

更多维度的尝试

时间维度的尝试：HOME CREDIT中期的一个提升，就来源于在时间维度的尝试，这个其实也非常容易理解，对于客户的近期行为和远期行为，对于客户违约逾期肯定有着不同的影响，在各张表上都进行了30天、90天、120天、365天不同时间段的数据统计。

行为次数维度的尝试：部分表在行为次数维度上进行了尝试，比如最近5次，10次，15次的行为数据的数据统计，2种维度的统计方法进行了混用，最终选择了一套在PB

上表现更优秀的划分策略。

但是随之而来也引入了一个问题，就是特征的共线性问题，通过对特征的共线性统计，存在大量的相关系数为1的特征，虽然LIGHTGBM对于共线性特征不是特别的敏感，但是去掉共线性特征后，PB有一定程度的提升。

数据驱动的模型，这应该是一个必然的过程，因为对特征字段的不了解，盲目暴力的引入特征，导致特征共线性严重，这样的特征在LR或者SVM模型下，表现必然是一塌糊涂，好在干部（XGB,LGB）系列的模型拯救了我们。

更细致的调整参数

通过测试发现，开源模型的参数和自己调优的模型参数，在本地CV上存在相当大的差异，可乐大佬提供的模型参数，CV表现上比原始参数高4个千分点，以往的比赛中往往更注重特征的提取，但是在KAGGLE这样竞争异常激烈的比赛中，显然只怼特征是远远不够的，这也从侧门反映了团队作战的好处。

特征字段的交互

对于一些多个表中都存在的字段，我尝试进行了一些交互操作，因为并不真正了解字段的含义，因此也是采用了暴力测试的手段，因此一不小心又整了一大堆特征出来，最后从中选择了一些特征重要性比较高的特征，但是自己也无法理解暴力交互出来特征的业务含义。最终特征数量在3200个左右，基本是自己这台联想启天商务机的极限了。