马上 AI 全球挑战者大赛——违约用户风险预测

团队名称：我只吃两个剩下的都给你

一、方案概述

近年来，互联网金融已经是当今社会上的一个金融发展趋势。在金融领域，无论是投资理财还是借贷放款，风险控制永远是业务的核心基础。对于消费金融来说，其主要服务对象的特点是：额度小、人群大、周期短，这个特性导致其被公认为是风险最高的细分领域。

以借贷为例，相比于传统的金融行业需要用户自己提供的资产资料的较单一途径，互联网金融更能将用户线下的资产情况，以及线上的网络消费行为进行资料整合，来进行综合分析，以便为用户提供更好的服务体验，为金融商家提供用户更全面的了解和评估。

随着人工智能和大数据等技术不断渗透，依靠金融科技主动收集、分析、整理各类金融数据，为细分人群提供更为精准的风控服务，成为解决消费金融风控问题的有效途径。简言之，如何区别违约风险用户，成为金融领域提供更为精准的风控服务的关键。

基于本赛题，大数据金融的违约用户风险预测，本文解决方案具体包括以下步骤：

1.对用户的历史行为数据预处理操作；

2.根据历史行为划分训练集数据、验证集数据；

3.对用户历史数据进行特征工程操作；

4.对构建特征完成的样本集进行特征选择；

5.建立多个机器学习模型，并进行模型融合；

6.通过建立的模型，根据用户历史行为数据对用户在未来一个月是否会逾期还

款进行预测。

其中，图 1 展示了基于大数据金融的违约用户风险预测解决方案的流程图。

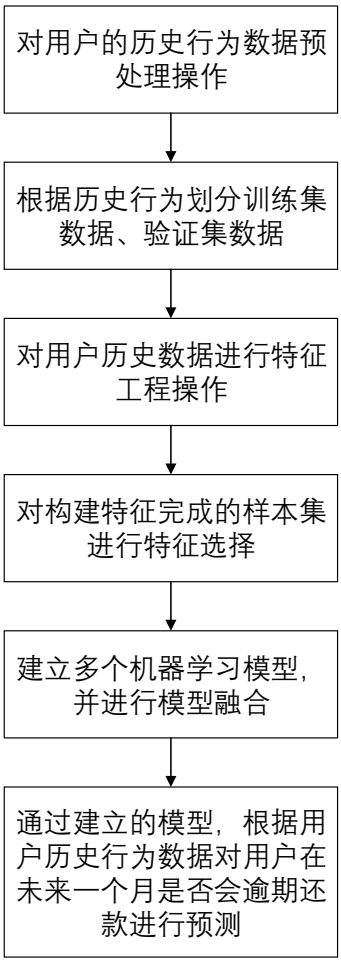


图 1 违约用户风险预测解决方案的流程图

二、数据洞察

2.1 数据预处理

1.异常值处理：针对数据中存在未知的异常值，采取直接过滤的方法进行处理会减少训练样本数量，从这里出发，将异常值用-1 或者其他有区别于特征正常值的数值进行填充；

2.缺失值的多维度处理：在征信领域，用户信息的完善程度可能会影响该用户的信用评级。一个信息完善程度为 100%的用户比起完善程度为 50%的用户，会更加容易审核通过并得到借款。从这一点出发，对缺失值进行了多维度的分析和处理。按列（属性）统计缺失

值个数，进一步得到各列的缺失比率，按对数据进行多维度处理，其中x_i为数据集中某属性列缺失值个数，Count为样本集总数，MissRate_i为数据集中该属性列缺失率：

$$\text{MissRate}_i = \frac{x_i}{\text{Count}}$$

3.其他处理：空格符处理，某些属性取值包含了空格字符，如“货到付款”和“货到付款 ”，它们明显是同一种取值，需要将空格符去除；城市名处理，包含有“重庆”、“重庆市”等取值，它们实际上是同一个城市，需要把字符中的“市”全部去掉。去掉类似于“市”的冗余之后，城市数目大大减少。

2.2 发现时序关系

根据用户历史数据，统计违约数量和未违约数量跟时间周期的关系，可视化实现如下图所示：

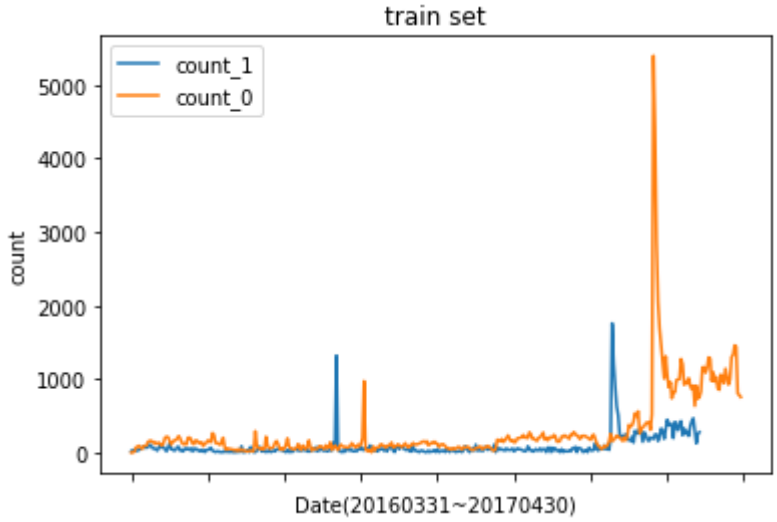


图 2 违约数量和未违约数量跟时间周期的关系图

可以看出，时间对用户是否违约是成一定周期性的，且 2017 年明显比 2016 年的数量增加了很多，因此本文解决方案涉及很多时序特征。

2.3 划分训练集、验证集

对违约用户风险预测是一个长期且累积的过程 ,采取传统的按训练和测试集对应时间段滑窗法划分数据集并不是最佳方案，从这里出发，将历史用户数据全部用于训练集，更好的训练用户行为习惯，其中，验证集的构建采取交叉验证的方式，交叉验证如下图所示：

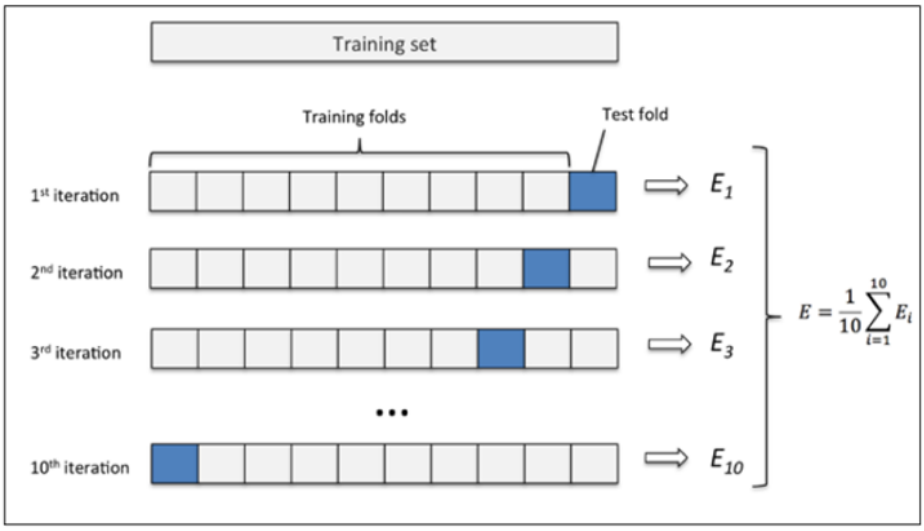


图 3 交叉验证示意图

三、特征工程

3.1 0-1 特征

主要基于 auth、credit、user 表提取，这三张表的 id 没有重复。

（1）标记 auth 表的 Id_card、auth_time、phone 是否为空；标记 credit 表的 credit_score、overdraft、quota 是否为空；标记 user 表的 sex、birthday、hobby、merriage、income、id_card、degree、industry、qq_bound、wechat_bound、account_grade 是否为空。

（2）标记 auth 表的 Id_card、auth_time、phone 是否正常（不为空）；标记 credit 表的 credit_score、overdraft、quota 是否正常(不为空);标记 user 表的 sex、birthday、

hobby、merriage、income、id_card、degree、industry、qq_bound、wechat_bound、account_grade 是否正常（不为空）。

3.2 信息完整度特征

主要基于 auth、credit、user 表提取，标记这三张表每条样本的信息完整度，定义为该条样本非空的属性数目/总属性数目。

3.3 one-hot 特征

主要基于 user 表提取。

One-hot 离散 user 表的 sex、merriage、income、degree、qq_bound、wechat_bound、account_grade 属性。

3.4 业务特征

基于业务逻辑提取的特征，最有效的特征，主要基于 credit、auth、bankcard、order 表提取。

（1）用户贷款提交时间（applsbm_time）和认证时间（auth_time）之差

（2）用户贷款提交时间（applsbm_time）和生日（birthday）之差

（3）信用评分（credit_score）反序

（4）信用额度未使用值（quota 减 overdraft）

（5）信用额度使用比率（overdraft 除以 quota）

（6）信用额度使用值是否超过信用额度（overdraft 是否大于 quota）

（7）银行卡（bankname）数目

（8）不同银行的银行卡（bankname）数目

（9）不同银行卡类型（card_type）数目

（10）不同银行卡预留电话（phone）数目

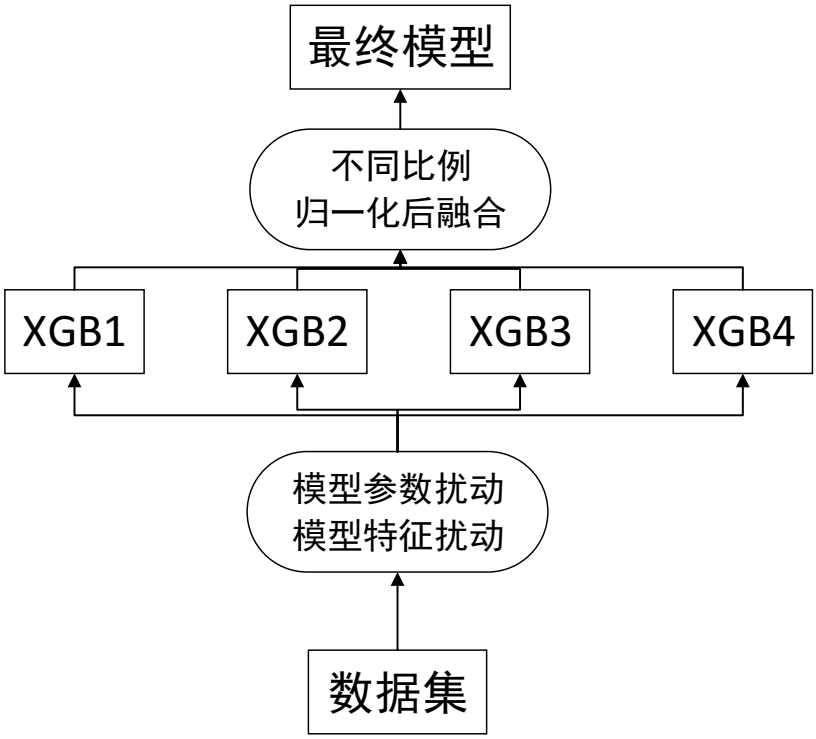
（11）提取 order 表的 amt_order 次数、type_pay_在线支付、type_pay——货到付款、sts_order_已完成次数，按 id 对 order 表去重，保留 id 重复的第一条样本

四、特征筛选

特征工程部分，构建了一系列基础特征、时序特征、业务特征、组合特征和离散特征等，所有特征加起来高达数百维，高维特征一方面可能会导致维数灾难，另一方面很容易导致模型过拟合。从这一点出发，通过特征选择来降低特征维度。比较高效的是基于学习模型的特征排序方法，可以达到目的：模型学习的过程和特征选择的过程是同时进行的，因此我们采用这种方法，基于 xgboost 来做特征选择， xgboost 模型训练完成后可以输出特征的重要性（见图 2），据此我们可以保留 top n 个特征，从而达到特征选择的目的。

五、模型训练

本文共计四个 xgb 模型，分别进行参数扰动、特征扰动，单模型效果均通过调参和特征选择，保证单模型最优，按四个模型不同比例融合，最终生成模型结果。



六、重要特征

通过 XGBOOST 模型输出特征重要性，降序排序，选取 top20，可视化如下：

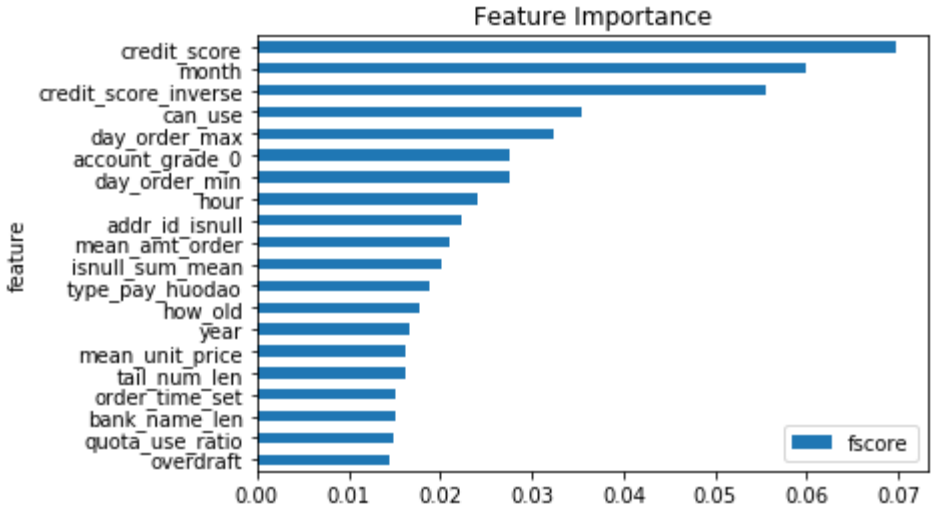


图 4 特征重要性排序

列出模型所选的重要特征的前 20 个：表格样式如下：

特征名称	特征释义	特征重要性排名
credit_score	网购平台信用评分	1

month	当前样本月份	2
credit_score_inverse	网购平台信用评分反序	3
can_use	剩余使用贷款额	4
day_order_max	当前样本时间-订单最大时间	5
account_grade_0	会员级别 0 类型的离散值	6
day_order_min	当前样本时间-订单最小时间	7
hour	当前样本小时数	8
addr_id_isnull	地址信息是否为 null	9
mean_amt_order	订单金额的均值	10
isnull_sum_mean	缺失值总数的均值	11
type_pay_huodao	货到付款类型的数量	12
how_old	用户年龄	13
year	当前样本的年份(2016/2017)	14
mean_unit_price	商品单价均值	15
tail_num_len	银行卡号码长度	16
order_time_set	用户下单时间不同的次数	17
bank_name_len	银行卡长度	18
quota_use_ratio	用户贷款额使用率	19
overdraft	网购平台信用额度使用值	20

七、创新点

7.1 特征

原始数据集很多属性比较乱，清洗了例如日期这样的属性方便特征提取；加入了信息完整度特征，很好地利用到了含有空值的样本；对于 order 这个 id 含有重复的样本，尝试了提取特征后按时间去重和按第一条和最后一条去重，发现按第一条去重效果是最好的，很好地使用到了 order 的信息；通过特征的重要性排序筛选了特征，也发现了提取的业务相关的特征是最重要的。

7.2 模型

模型的创新点主要体现在模型融合上。考察指标为 AUC，侧重于答案的排序。在进行加权融合时，先对每个模型的结果进行了归一化，融合效果很好。

八、赛题思考

清洗数据非常重要，像时间这样的属性非常乱，处理起来也比较麻烦，我们只是简单地进行了处理，如果能够更细致的处理效果应该更好；某些属性，例如 hobby，内容太复杂没有使用到，但这个属性肯定包含了许多有价值的信息，但遗憾没有发现一个好的处理方式。