金融风控问题案例

寒小阳
七月在线　2017年1月22日

“魔镜杯”
风控算法大赛

背景介绍

- 比赛公开了国内网络借贷行业的贷款风险数据
  1. 包括信用违约标签（因变量）
  2. 建模所需的基础与加工字段（自变量）
  3. 相关用户的网络行为原始数据

本着保护借款人隐私以及拍拍贷知识产权的目的，数据字段已经过脱敏处理。

数据简介

- 数据编码为GBK。
- 初赛数据包括3万条训练集和2万条测试集。
- 复赛会增加新的3万条数据，供参赛团队优化模型，并新增1万条数据作为测试集。
- 所有训练集，测试集都包括3个csv文件。

数据信息

- Master(每一行代表一个成功成交借款样本，每个样本包含200多个各类字段。

  ❶ idx：每笔贷款的unique key，可与另外2个文件里的idx相匹配。
  ❷ UserInfo_*：借款人特征字段
  ❸ WeblogInfo_*：Info网络行为字段
  ❹ Education_Info*：学历学籍字段
  ❺ ThirdParty_Info_PeriodN_*：第三方数据时间段N字段
  ❻ SocialNetwork_*：社交网络字段
  ❼ LinstingInfo：借款成交时间
  ❽ Target：违约标签（1＝贷款违约，0＝正常还款）。测试集里不包含target字段。

数据信息

➢ Log_Info（借款人的登陆信息）

1. ListingInfo：借款成交时间
2. LogInfo1：操作代码
3. LogInfo2：操作类别
4. LogInfo3：登陆时间
5. idx：每一笔贷款的unique key

➢ Userupdate_Info（借款人修改信息）

1. ListingInfo1：借款成交时间
2. UserupdateInfo1：修改内容
3. UserupdateInfo2：修改时间
4. idx：每一笔贷款的unique key

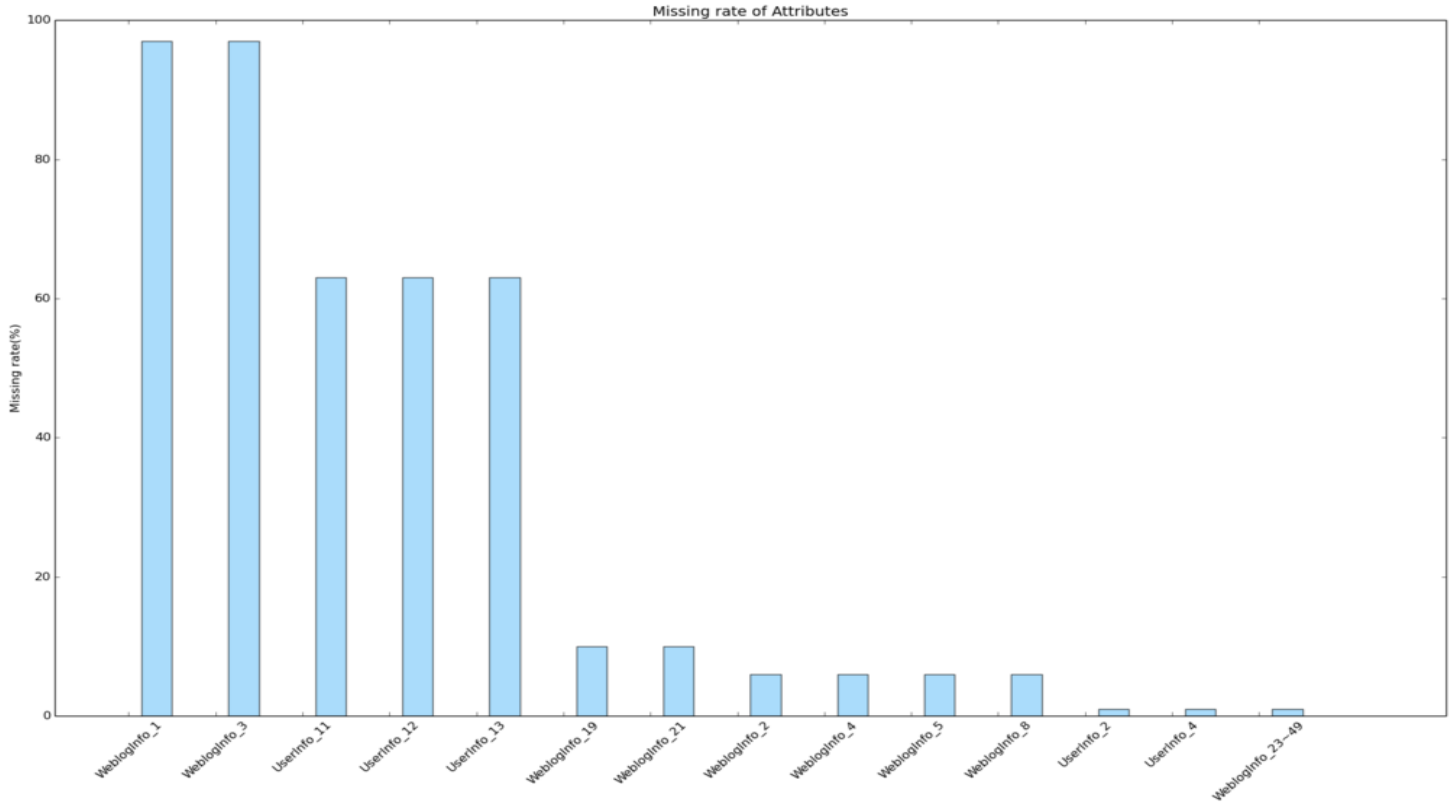
处理过程

1. 数据清洗
   - 对缺失值的多维度处理
   - 对离群点的剔除方法
   - 文本处理
2. 特征工程
   - 地理信息处理
   - 成交时间
   - 类别型编码
   - 组合特征
3. 特征选择
   - Xgboost重要度排序
4. 类别不平衡处理
   - 代价敏感学习 与 过采样
5. 模型设计与优化

数据清洗

❶ 缺失值的多维度处理

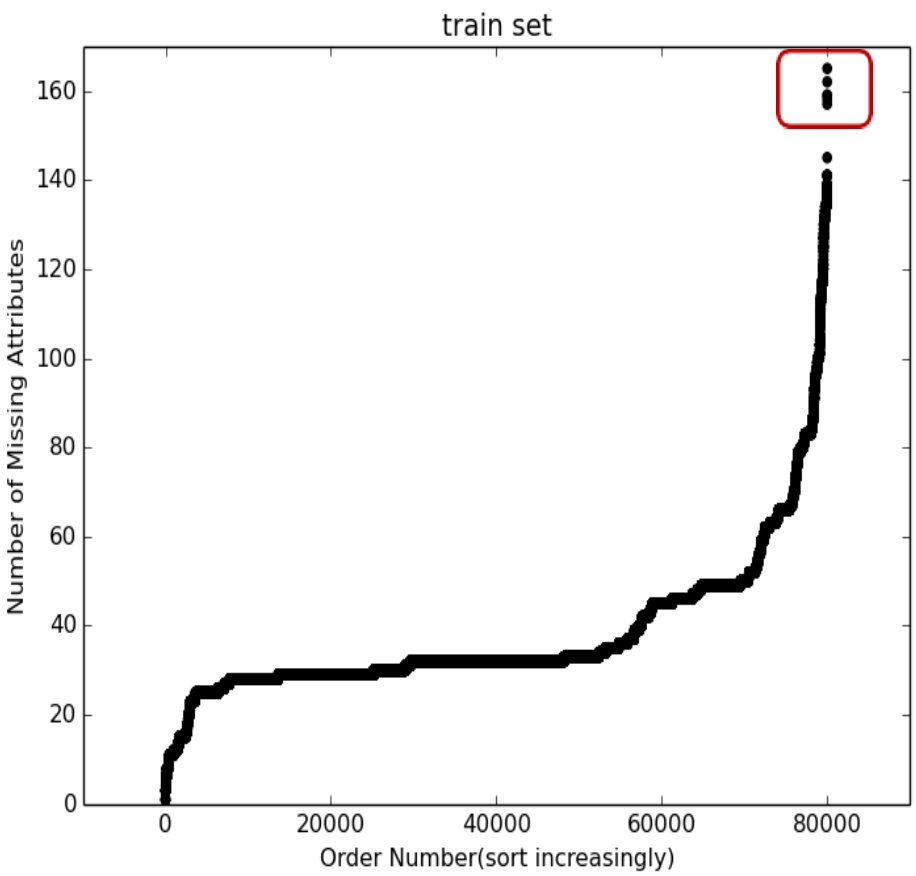
- 按列（属性）统计缺失值个数，进一步得到各列的缺失比率

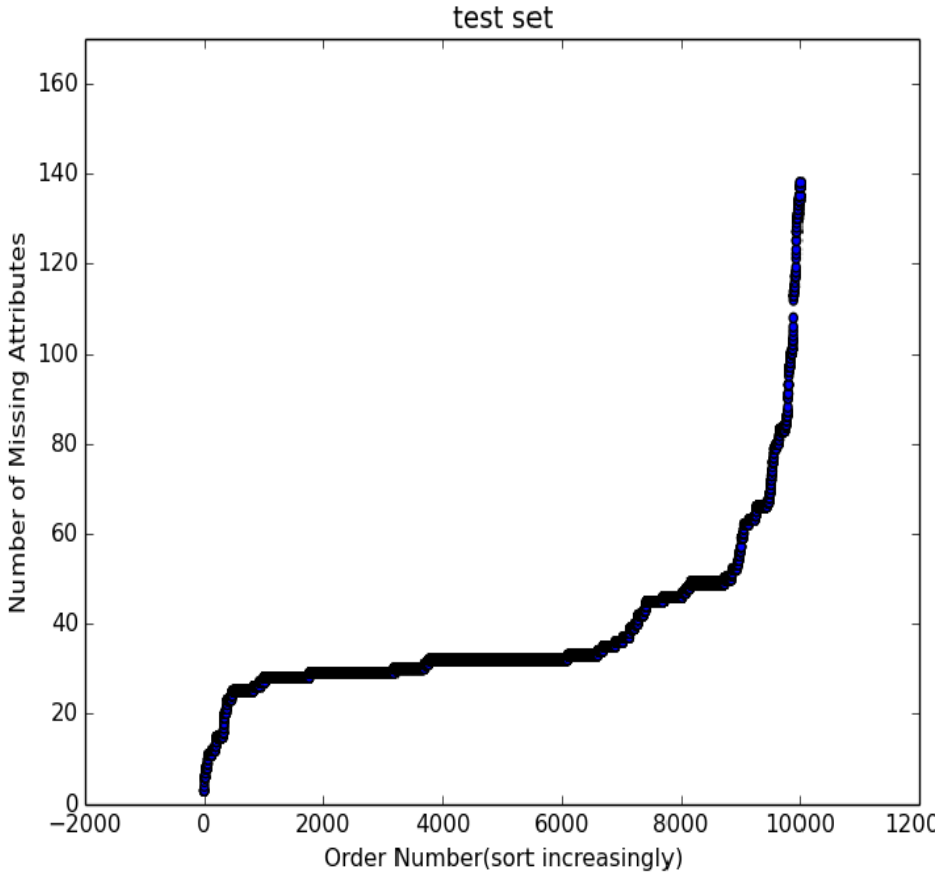


数据清洗

❶ 缺失值的多维度处理

- 按行统计每个样本的属性缺失值个数，将缺失值个数从小到大排序





数据清洗

❷ 剔除常变量

原始数据中有 190 维数值型特征，通过计算每个数值型特征的标准差，剔除部分变化很
小的特征，下表列出的 15 个特征是标准差接近于 0 的，剔除这 15 维特征。

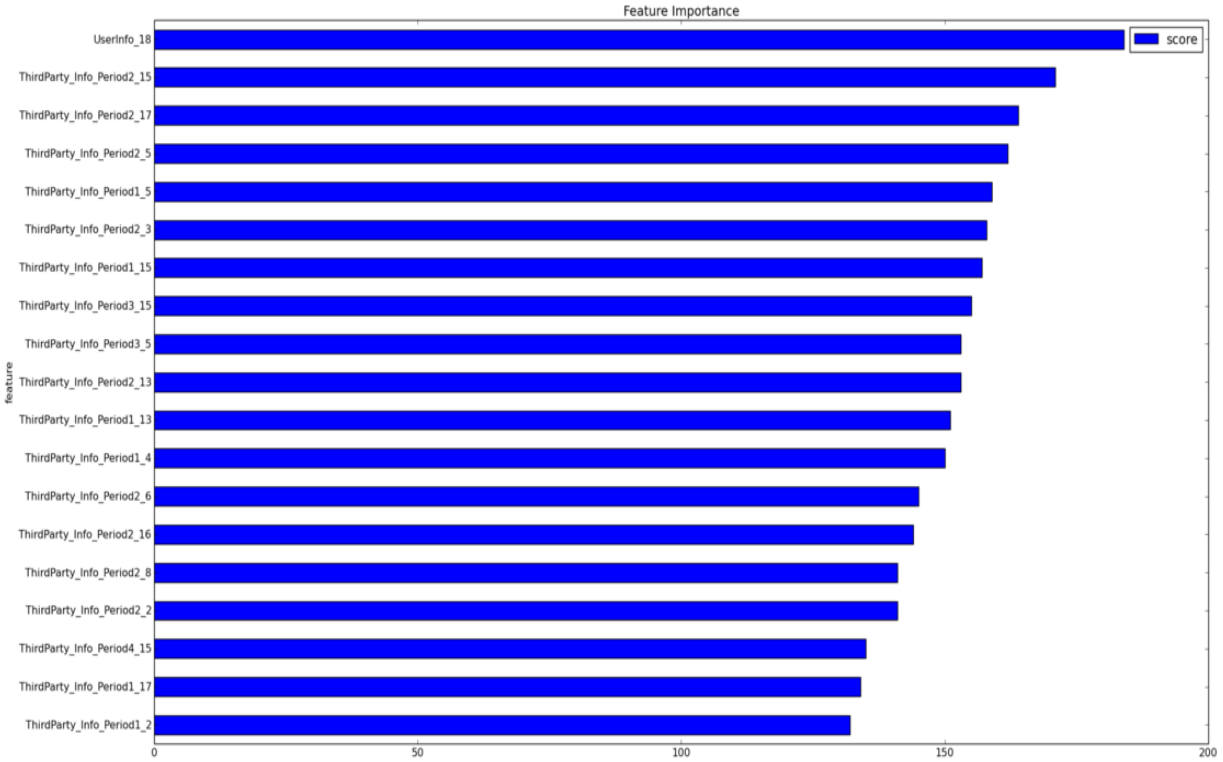
表 1.剔除数值特征标准差

属性	标准差	属性	标准差	属性	标准差
WeblogInfo_10	0.0707	WeblogInfo_41	0.0212	WeblogInfo_49	0.0071
WeblogInfo_23	0.0939	WeblogInfo_43	0.0372	WeblogInfo_52	0.0512
WeblogInfo_31	0.0828	WeblogInfo_44	0.0166	WeblogInfo_54	0.0946
WeblogInfo_32	0.0834	WeblogInfo_46	0.0290	WeblogInfo_55	0.0331
WeblogInfo_40	0.0666	WeblogInfo_47	0.0401	WeblogInfo_58	0.0609

数据清洗

❸ 离群点剔除

在原始数据上训练 xgboost，用得到的 xgb 模型输出特征的重要性，取最重要的前 20 个特征（如图 3 所示），统计每个样本在这20 个特征上的缺失值个数，将缺失值个数大于 10 的样本作为离群点。



数据清洗

❹ 其余处理

(1) 字符大小写转换

Userupdate_Info 表中的 UserupdateInfo1 字段，属性取值为英文字符，包含了大小写，如“_QQ”和“_qQ”，很明显是同一种取值，我们将所有字符统一转换为小写。

(2) 空格符处理

Master 表中 UserInfo_9 字段的取值包含了空格字符，如“中国移动”和“中国移动 ”，它们是同一种取值，需要将空格符去除。

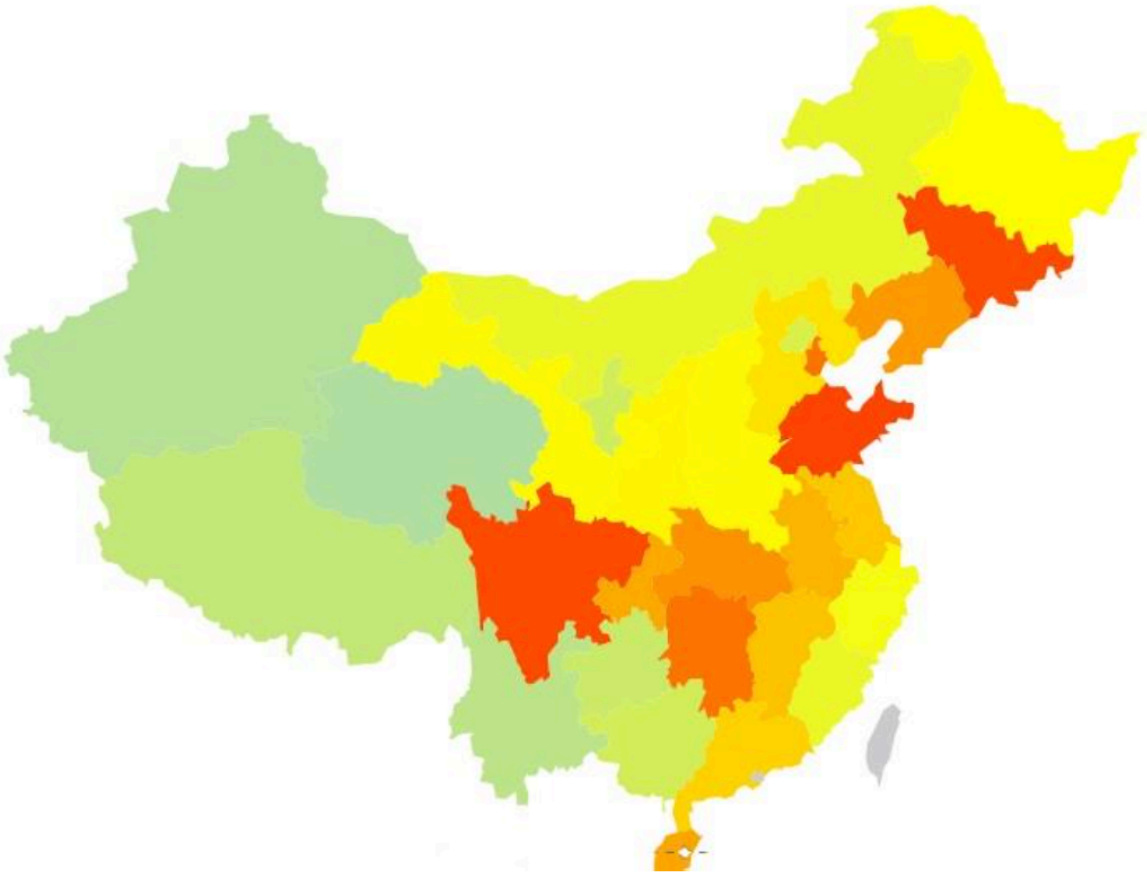
(3) 城市名处理

UserInfo_8 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“市”之后，城市数由 600 多下降到 400 多。

特征工程

❶ 地理位置的处理

UserInfo_7 和 UserInfo_19 是省份信息，其余为城市信息。统计每个省份和城市的违约率，以 UserInfo_7 为例



特征工程

❶ 地理位置的处理

违约率最大的几个省 份或直辖市为四川、湖南、湖北、吉林、天津、山东

构建 6 个二值特征:“是否为四川省”、“是否为湖南省”....“是否为 山东省”,其取值为 0 或 1

特征工程

❶ 地理位置的处理

按城市等级合并

类别型特征取值个数太多时，独热编码后得到太高维的稀疏特征。
除了采用上面提到的特征选择方法外，还可以使用了合并变量的方法。按照城市等级，将类别变量合并，例如一 线城市北京、上海、广州、深圳合并，赋值为 1，同样地，二线城市合并为 2，三线城市合 并为 3。

经纬度特征的引入

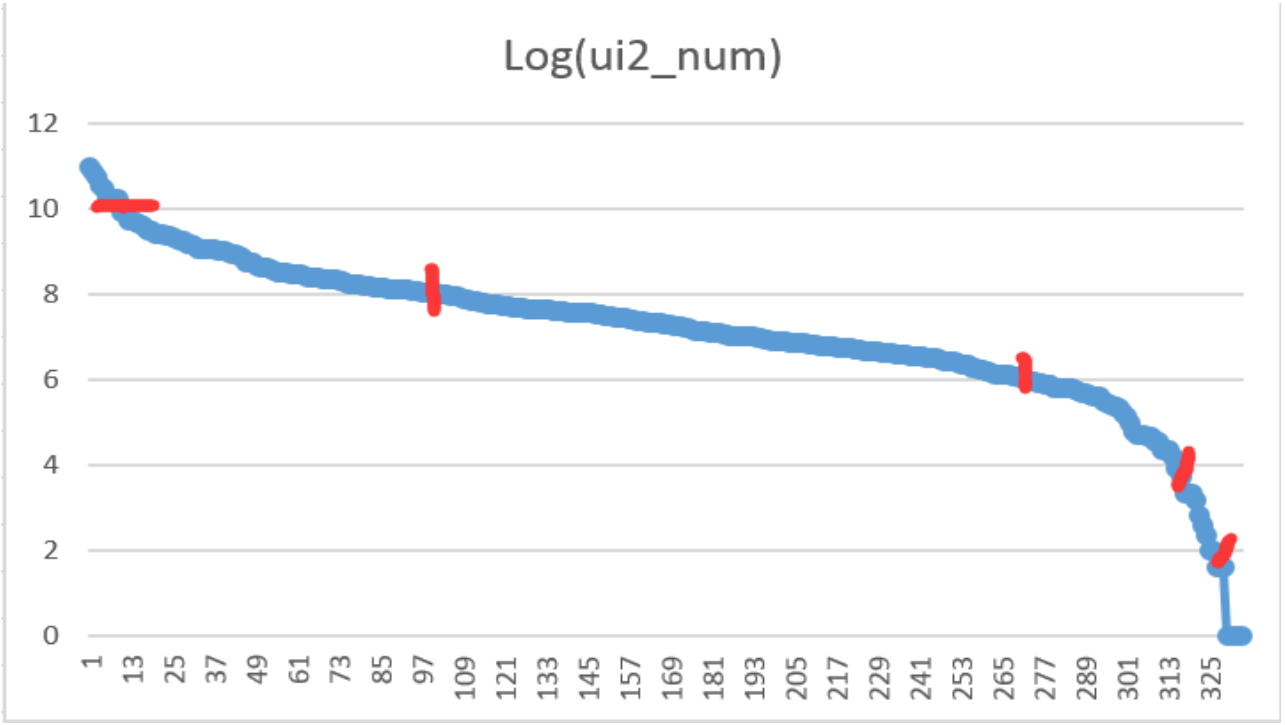
以上对地理位置信息的处理，都是基于类别型的，收集各个城市的经纬度，将城市名用经纬度替换，这样就可以将类别型的变量转化为数值型的变量，比如北京市，用 经纬度（39.92,116.46）替换，得到北纬和东经两个数值型特征。加入经纬度后，线下的 cross validation 有千分位的提升。

特征工程

❶ 地理位置的处理

城市特征向量化

将城市特征里的城市计数，并取 Log，然后等值离散化到 6~10 个区间内。如下图，将 UserInfo_2 这个特征里面的 325 个城市离散为一个 6 维向量。向量“100000”表 示该城市位于第一个区间。



特征工程

❶ 地理位置的处理

地理位置差异特征

如下图所示，1,2,4,6 列都是城市。我们构建一个城市差异的特征，
比如 diff_12 表示 1，2 列的城市是否相同。
如此构建 diff_12,diff_14,diff_16,diff_24,diff_26,diff_46 这 6 个城市差异的特征。

Idx	UserInfo_2	UserInfo_4	UserInfo_7	UserInfo_8	UserInfo_19	UserInfo_20
10005	广州	韶关	广东	广州	辽宁省	锦州市
10013	郴州	广州	广东	广州	湖南省	郴州市
10020	惠州	惠州	广东	惠州	四川省	广安市
10033	枣庄	枣庄	山东	枣庄	山东省	枣庄市
10035	深圳	南平	福建	南平	福建省	不详
10038	济宁	济宁	山东	济宁	山东省	济宁市
1004	连云港	连云港	江苏	连云港	江苏省	连云港市
10042	德州	德州	山东	滨州	山东省	德州市
10043	青岛	聊城	不详	不详	山东省	聊城市
10046	深圳	汕尾	广东	汕尾	广东省	汕尾市
1005	新乡	新乡	河南	新乡	河南省	新乡市

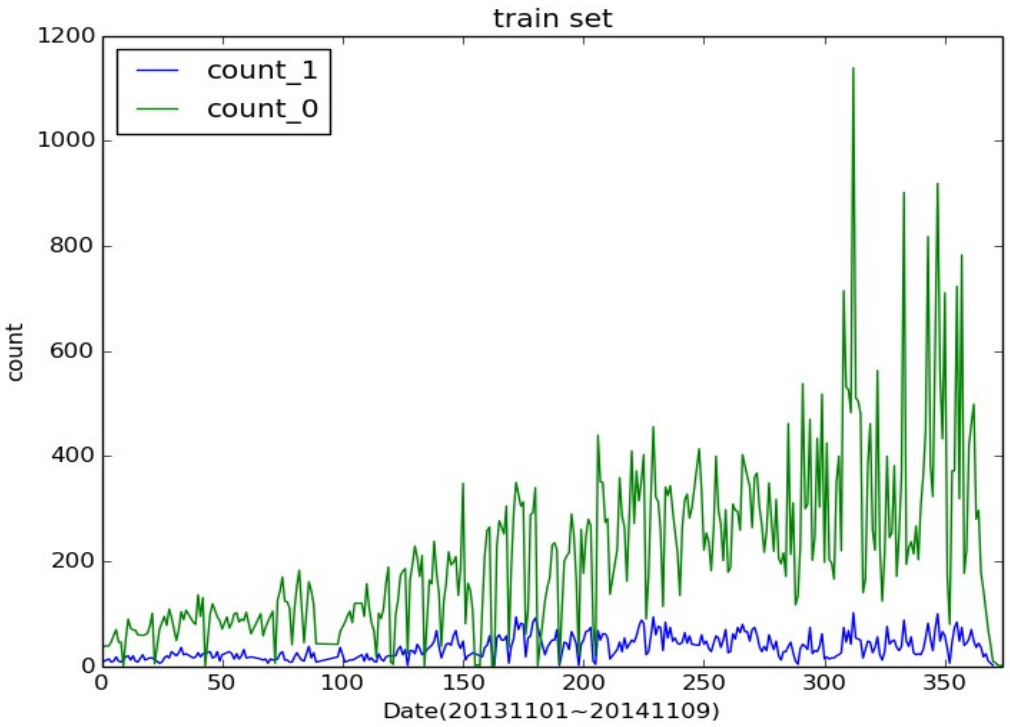
特征工程

❷ 成交时间

将成交时间的字段 Listinginfo 做几种处理

一种 是直接将其当做连续值特征，另一种是离散化处理

每 10 天作 为一个区间，也就是将日期 0~10 离散化为1，日期 11~20 离散化为2



每日借贷量统计

特征工程

❸ 类别型特征

除掉上述特殊生成的特征，其余都做独热编码

❹ 组合特征

Xgboost的训练完成后可以输出特征的重要性，发现第三方数据特征“ThirdParty_Info_Period_XX”的featurescore比较大
于是用这部分特征构建了组合特征：
将特征两两相除得到7000+个特征，然后使用xgboost对这7000多个特征单独训练模型，训练完成后得到特征重要性的排序，取其中top500个特征线下cv能达到0.73+的AUC值。将这500个特征添加到原始特征体系中，线下cv的AUC值从0.777提高到0.7833。另外，也组合了乘法特征（取对数）：log(x*y)，刷选出其中的270多维，加入到原始特征体系中，单模型cv又提高到了0.785左右。
特殊生成的特征，其余都做独热编码

特征工程

❺ UpadteInfo 表特征

根据提供的修改信息表，从中抽取了用户的修改信息特征，比如：修改信息次数，修改信息时间到成交时间的跨度，每种信息的修改次数等等特征。

❻ LogInfo 表特征

类似地，从登录信息表里提取了用户的登录信息特征，比如登录天数，平均登录间 隔以及每种操作代码的次数等。

❼ 排序特征

对原始特征中 190 维数值型特征按数值从小到大进行排序，得到 190 维排序特征。排序 特征对异常数据有更强的鲁棒性，使得模型更加稳定，降低过拟合的风险。

特征选择

除了采用降维算法之外，也可以通过特征选择来降低特征维度。

特征选择的方法很多：

- 最大信息系数（MIC）
- 皮尔森相关系数（衡量变量间的线性相关性）
- 正则化方法（L1，L2）
- 基于模型的特征排序方法

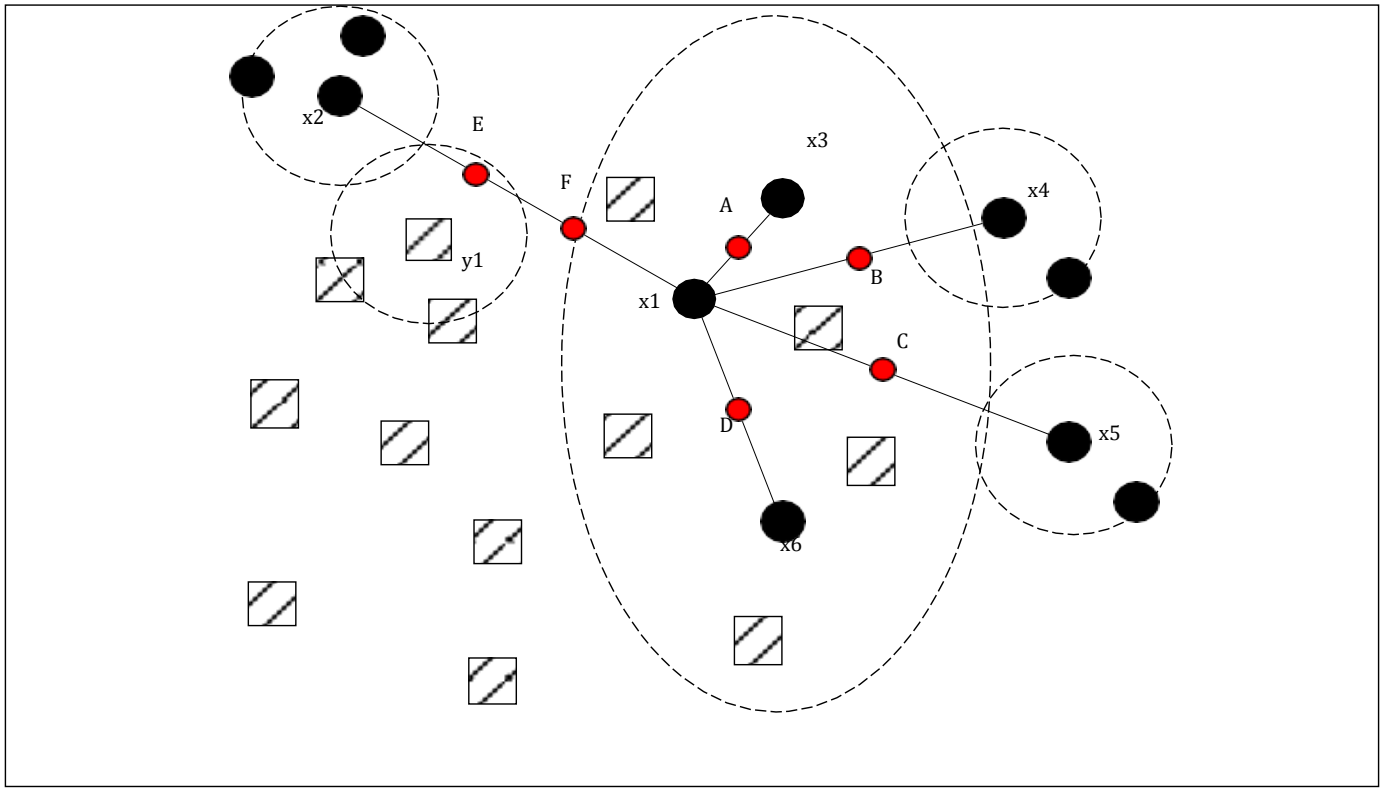
比较高效的是最后一种，即基于学习模型的特征排序方法，这种方法有一个好处：模型学习的过程和特征选择的过程是同时进行的，因此采用这种方法，基于 xgboost 来做特征选择，xgboost 模型训练完成后可以输出特征的重要性，据此可以保留 Top N 个特征，从而达到特征选择的目的。

类别不平衡处理

赛题数据的类别比例接近13:1，采用两种解决类别不平衡问题的方法

1）在训练模型时设置类别权重，即代价敏感学习

2）过采样

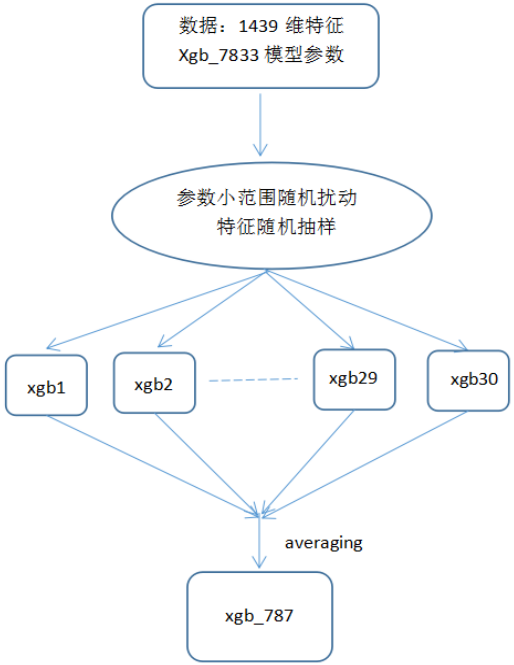


建模与模型融合

❶ Logistic regression + L1正则化

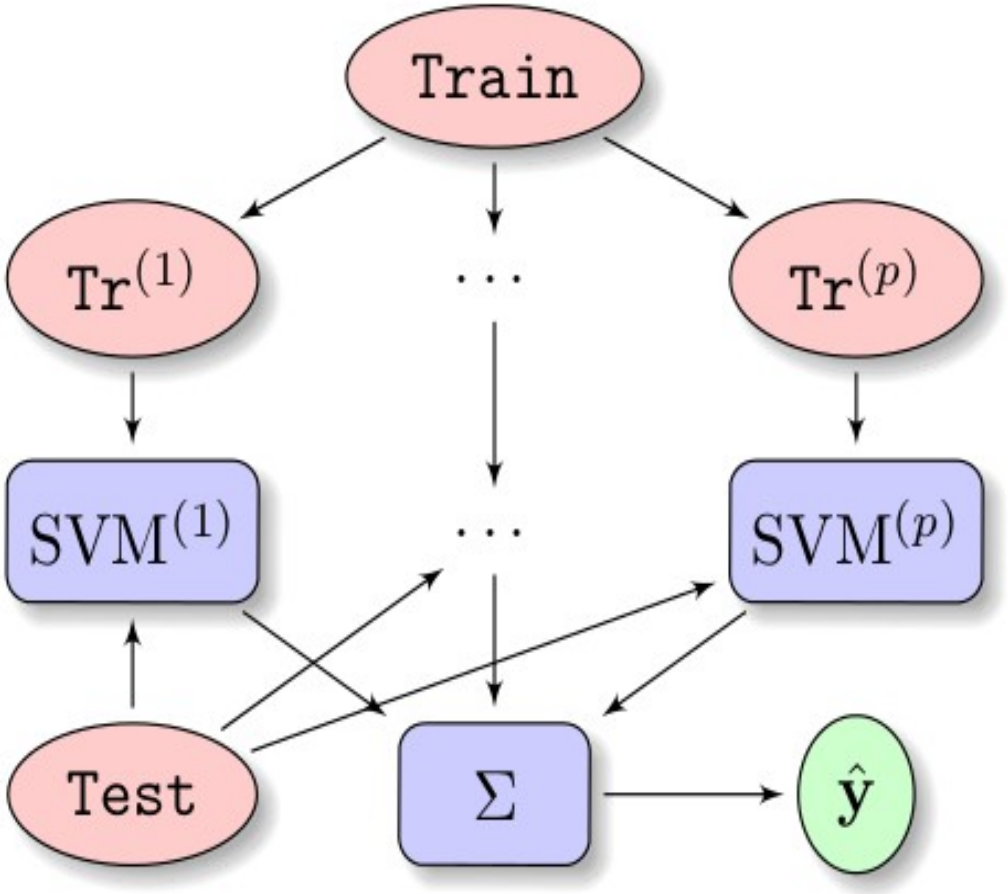
1700 维特征上训练模型，单模型验证集上得到的 AUC 值为 0.772 左右

❷ Xgboost + bagging



建模与模型融合

❸ Large-Scale SVM



建模与模型融合

❹ 多模型blending

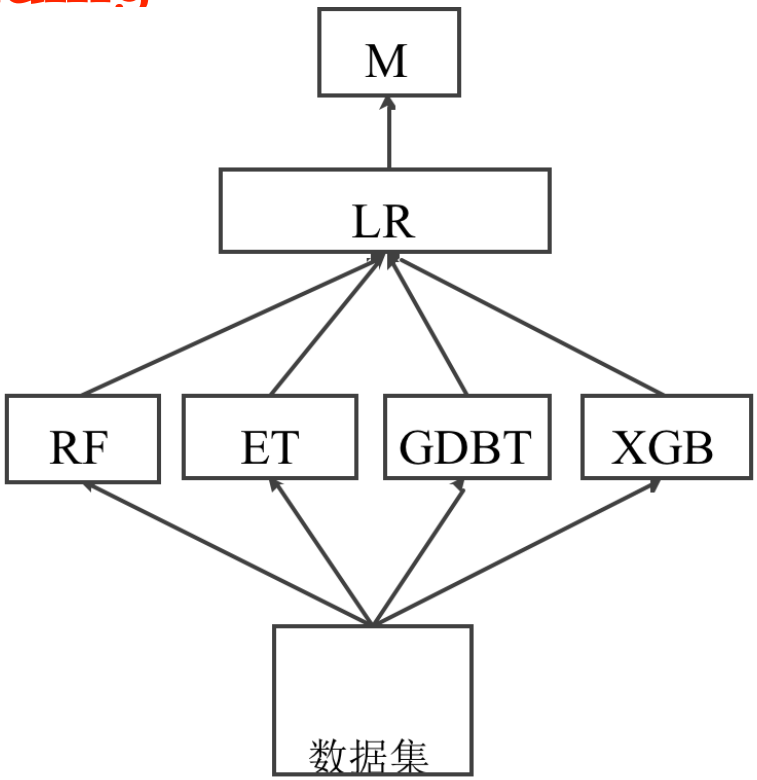
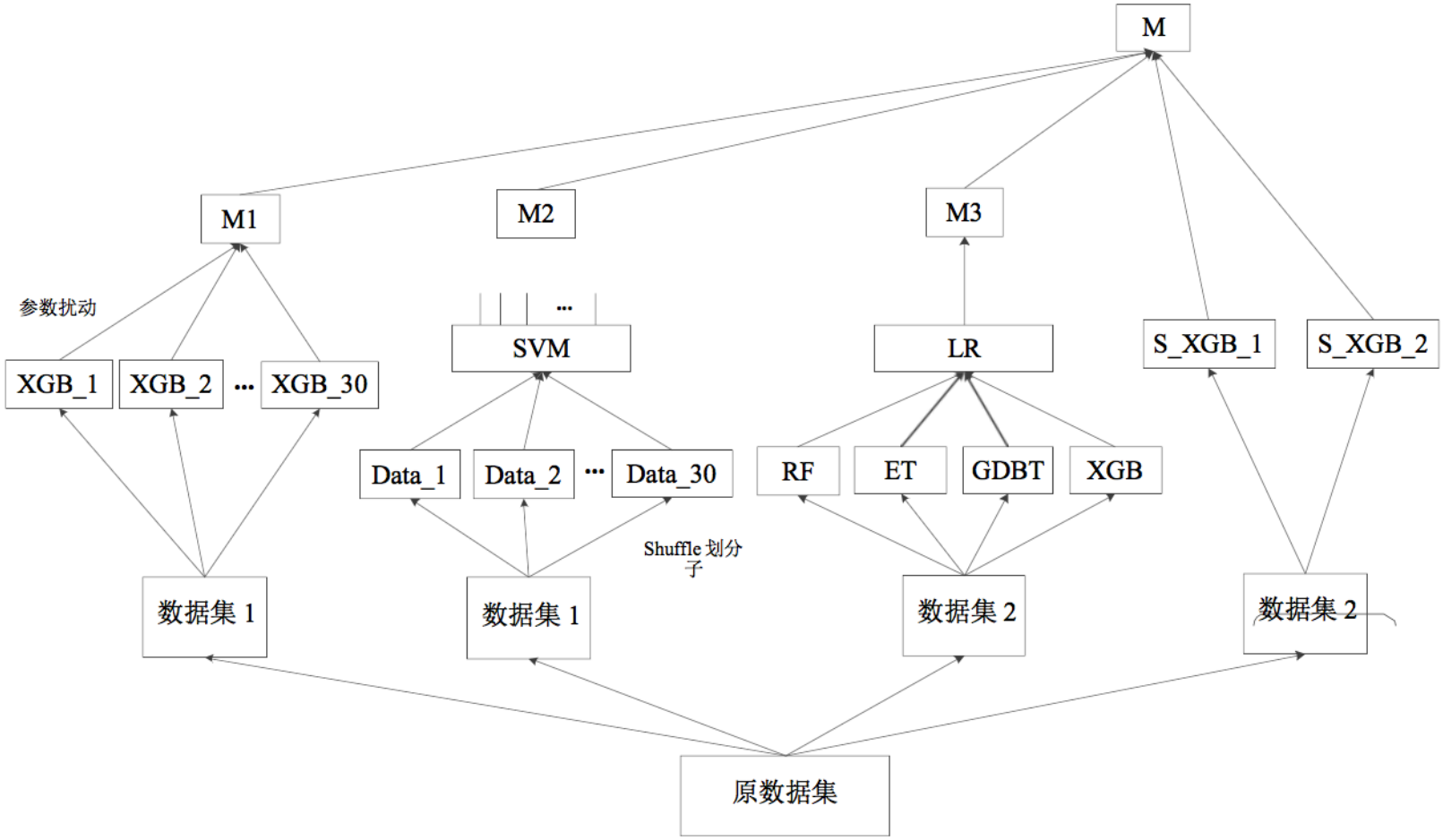


图13. 多模型blendin ensemble

建模与模型融合

❺ 模型融合



感谢大家！

恳请大家批评指正！