基本特征

用户id(1931654) 1
问题id(1829900) 1
性别 1
盐值(95-890) 1
访问频率 1
邀请创建时间-day(3838-3867,3868-3874) 1
邀请创建时间-hour(0-23) 1
问题创建时间-day(3838-3867,3868-3874) 1
问题创建时间-hour(0-23) 1
用户二分类特征a 1
用户二分类特征b 1
用户二分类特征c 1
用户二分类特征d 1
用户二分类特征e 1
用户多分类特征a 1
用户多分类特征b 1
用户多分类特征c 1
用户多分类特征d 1
用户多分类特征e 1
邀请创建时间-weekday (1-7) 1
问题创建时间-weekday (1-7) 1

用户感兴趣/关注的话题文本embedding(average) 0
问题相关的话题文本embedding(average) 0

是否回答(10628265) 1

统计特征

用户id_count float64 1
问题id_count float64 1
性别_count float64 1

访问频率 count 1

邀请创建时间-hour_count 1
问题创建时间-day count 1
问题创建时间-hour_count 1
用户二分类特征a_count float64 1
用户二分类特征b_count float64 1
用户二分类特征c count float64 1
用户二分类特征d_count float64 1
用户二分类特征e count float64 1
用户多分类特征a_count float64 1
用户多分类特征b_count float64 1
用户多分类特征c_count float64 1
用户多分类特征d_count float64 1
用户多分类特征e_count float64 1
邀请创建时间-weekday count 1
问题创建时间-weekday count 1

邀请创建时间-Friday.Sat.Sun.计数 0

问题创建时间与邀请时间距离-day（最小/最大/平均：问题id、 1
问题创建时间与邀请时间距离-hour（最小/最大/平均：问题id 1

问题关联的邀请label（求和，均值，方差，总次数） 0
用户关联的邀请label（求和，均值，方差，总次数） 1
用户拒绝邀请次数 0

用户最感兴趣的话题 1
用户兴趣值的topic（最大，最小，均值，方差） 1
用户关注的topic的个数 1
用户感兴趣的话题 count 1
用户关注的话题 count 1
用户上次接受邀请距离当前时间-day 0

用户回答时间-hr_percent(用户id，hour) 0
用户回答时间-wk_percent(用户id，weekday) 0

注意不能把label的信息混到特征中！ 1

把用户根据回答历史情况拆开两个训练集，面向两个训练模型

按照日前向链法切割训练集验证集（交叉验证）进行训练 1

可以利用lightgbm分析特征的重要性！

归一化一下里面出现的所有连续数值特征

log变换/区间放缩

减少负样本，用采样的方法

如何处理缺失值？

归一化的时候，注意处理异常值

增加训练样本

学习residual

用dnn去学习lgb没学到的residual

user answer

用户上次回答距离当前时间-day（跟验证集切割时间点有 1
感谢数（最大，平均和求和） 1
点赞数（最大，平均和求和） 1
收藏数（最大，平均和求和） 1
取赞数（最大，平均和求和） 1
举报数（最大，平均和求和） 1
没有帮助数（最大，平均和求和） 1
反对数（最大，平均和求和） 1
评论数（最大，平均和求和） 1
被推荐次数（最大，平均和求和） 1
用户的回答被标优次数（最大，平均和求和） 1
用户的回答被收入圆桌次数（最大，平均和求和） 1
是否有视频（最大，平均和求和） 1
是否有图片（最大，平均和求和） 1
内容字数（最大，平均和求和） 1

用户回答时间与邀请时间距离-day（最小/最大/平均，用户 1
用户回答时间与邀请时间距离-hour（最小/最大/平均，用 1
用户真应的（问题创建时间和回答时间day距离）最大，平 1
用户回答过的问题的topic统计（出现最多的） 0
用户回答过的话题embedding(average) 0

用户习惯回答时间-wkday 1
用户习惯回答时间-hour 1

用户关联回答数（最近三周） 1
用户最近回答的数量（最近三天） 1
用户最近回答的数量（最近一周） 1
用户最近回答的数量（最近两周） 1

回答创建时间-day（3807-3867） 0

颜色说明

代表需要把时间线细分（最近三天，最近一周，最近两周）
需要新旧分类对待
代表证明有效的特征
代表无效，并且已经去除的特征
代表注意事项

question answer

问题关联回答数 1 0

问题上次回答时间（day） 1 0
问题最近一周回答比例 1 0

感谢数（最大，平均和求和） 1
点赞数（最大，平均和求和） 1
收藏数（最大，平均和求和） 1
取赞数（最大，平均和求和） 1
举报数（最大，平均和求和） 1
没有帮助数（最大，平均和求和） 1
反对数（最大，平均和求和） 1
评论数（最大，平均和求和） 1
被推荐次数（最大，平均和求和） 1
被标优次数（最大，平均和求和） 1
被收入圆桌次数（最大，平均和求和） 1
是否有视频（最大，平均和求和） 1
是否有图片（最大，平均和求和） 1
内容字数（最大，平均和求和） 1

user_question_topic_answer

每个话题对应回答的用户各种分类比例… 0
eg.每个话题对应回答的用户性别比例 0
问题的质量（根据话题的反馈数据（赞同数/收藏数）） 1
问题相关的topic文本内容和用户回答文本相关的topic内容相似度 0
问题相关的topic和用户回答相关topic的重叠数目 0

每个话题对应的邀请接受率，然后对应到问题（近3天/7天/14天：平均值，中位数） 1
每个话题对应的反馈数据，然后对应到问题（近3天/7天/14天：平均，最大，最小，方差， 1
每个话题对应的回答数量，然后对应到问题（近3天/7天/14天：平均，最大，最小，方差， 1

用户回答过的topic字典，也可以取最大值，平均值，和最大值对应的话题

用户对问题相关的topic擅长程度（根据在每个topic下的回答数目，回答字数，topic又跟问题 0
用户对问题相关的topic受欢迎程度（根据在每个topic下的赞数目，topic又跟问题重合，从问 0

用户回答话题数目的（mean/std）
用户回答过话题的数目
用户最擅长回答的话题id
用户回答最多的话题id

user_question

用户关注/和问题相关话题的重叠数目 1
用户感兴趣话题和问题相关话题的重叠数目 1
topic_intersection_values 0

在test集里出现，在train和回答集的并集里 没有出现的用户和问题