智源-看山杯 专家发现算法大赛 2019

队伍名：“救救菜鸡吧”团队
队员及机构：舒秀峰 西安电子科技大学
刘臣 杭州电子科技大学
孙睿 电子科技大学

【摘要】

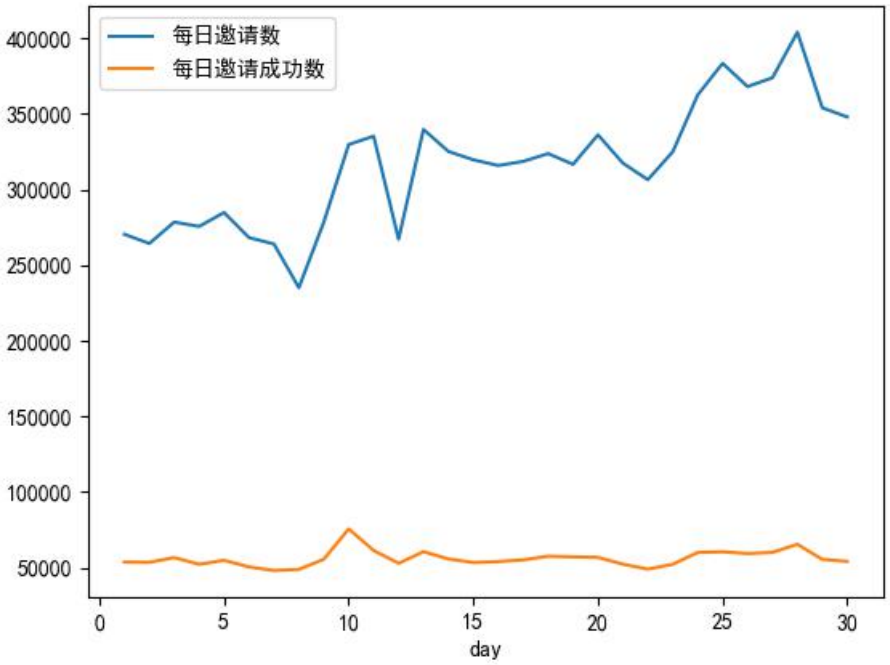
本次比赛结合了知乎的问题邀请与回答的相关场景和数据，需要选手们通过数据挖掘和建模来预测用户回答邀请问题的概率。结合相应的问题分析以及对数据集的探索，我们将该赛题等同为 ctr 预估的赛题。我们团队根据对数据集分布情况的探索，从多个角度构造特征，并验证特征的有效性，探索高效的训练模式。经过几个月时间对比赛的探索，我们构建了全局的统计计数特征，时间差相关的特征，用户和问题对应话题和标题信息的交互特征，以及模拟测试集分布构建了滑窗特征，总计 200 多种特征从不同角度，不同时间范围对特征进行了描述和表示。模型方面我们尝试了使用多种深度学习模型并加以改进和探索，也使用了多种竞赛中都取得了优异成绩的 lightgbm,经过了反复的尝试探索与分析，我们在最后一天使用了三个树模型从不同角度进行建模，最终取得了不错的线上成绩。

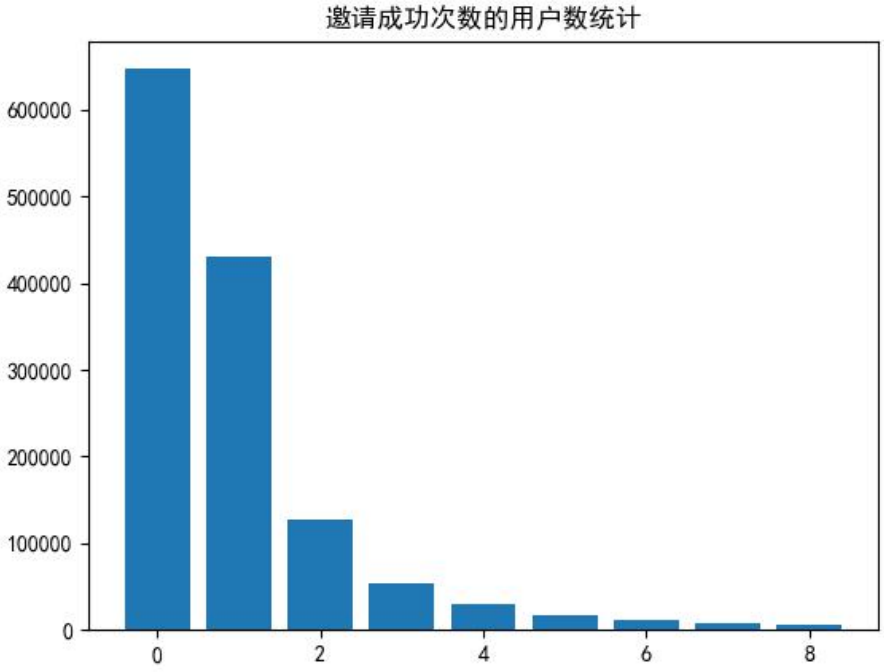
【关键词】

ctr 预估; 统计特征；滑窗特征；深度学习模型；lightgbm;

【正文】

一. 数据探索





通过数据探索我们发现，训练数据的每天的邀请次数和邀请成功次数保持平稳，并无很大的波动，说明数据分布没有收到外界因素的影响。大部分用户的邀请成功次数在6次以下，很大一部分用户没有接受邀请。

二. 数据集划分

由于线上提交的次数有限，因此，保证线上线下分数一致尤为重要。为了验证特征的有效性，为了减少线上成绩的运气成分，同时为了保证模型的鲁棒性，我们认真的进行了线下测试。为了保证线下验证集和线上测试集的新用户和新问题比例一致，我们选取了训练集的第3858到第3864这7天为训练集，第3865到第3867为线下验证集。为了构造的特征分布一致，我们还保证了相关特征的提取区间的时间段长度相同，即都使用前一个月的数据构造对应的特征。通过这种数据集划分方式，我们保证了线上线下分数的强一致性。

三. 特征工程

1. 基础特征:

刚接受比赛，通过对可用数据集的了解，我们很容易先使用一些比较基础的特征，比如用户的基

本特征，问题的title的长度，用户关注的话题的个数，用户的喜好评分等等这些，这些特征虽然不是强特，但是其它特征的构造也是从这些特征中衍生出来的，因此我们保留了一部分加入了我们的模型

2. 统计类特征:

这部分统计类的特征其实是非常有帮助的，可以从计数，均值，排序，比例等多个角度对用户信息，问题信息，话题信息进行统计。例如，可以统计用户历时受邀请的个数，用户当日出现的次数，问题的邀请次数，问题首日的邀请次数等等。这类特征构造思路简单，但确实可以反映出相关问题的回答热度以及用户的回答习惯，这部分特征对我们的分数带来了较大的提升

3. 滑窗特征：

由于验证集有7天，这部分数据由于没有标签，很难像训练集那样从多个角度构造特征。为了模拟这种情况，我们在有标签的数据集中选择了7天的数据作为训练集，最后3天的作为验证集，然后对训练集，验证集，测试集在它们所在时间段之前时间长度相等的邀请表区间，回答表区间构造了很多特征，这样使得训练集、测试集、验证集可以构造具有相同分布，相同误差的特征，并且大大减少了训练所需要的数量。经过我们实验，我们发现这样训练虽然减少了训练数据的数量，但是反而取得了更好的预测结果。

4. 时间特征与嫁接学习：

有关时间的特征我们发现都是强特，比如问题创建距离邀请过去的时间，问题邀请的时间间隔等等。此外，为了构造更多时间性相关的特征，我们利用模型预测的标签代替测试集中的标签，然后这样可以在测试集中的7天区段内构造更多的特征，并且使训练集和验证集构造这部分的标签也用模型预测的概率替代，这样就保证了误差的同步。

四. 模型选择和训练

模型方面，我们前期主要集中在尝试用深度学习方法对问题进行建模，我们一共尝试了DeepFM,NFM,Deep&Cross,xDeepFM等这几种模型，并探索了模型参数的选择和特征的输入方式。众所周知，这几个模型通过对类别特征进行embedding的表示，以及对特征进行自动交叉，可以达到不错的效果并在一定程度上可以代替人工提取特征。对于类别特征，我们直接为每一个特征建立了它所对应的embedding向量，让网络自动学习，而对于数值特征，我们的队员之间有着不同的尝试和使用方法，第一种方式，我们对数值特征进行归一化，然后为每一类数值特征维护一个embedding,将归一化的结果和embedding相乘送入模型。第二种方式，我们通过分析特征的分布，通过分箱的方法将数值特征转化为类别特征送入模型，不同的数值特征我们会依据不同的数据分布选择不同的分箱方法，如对于在数据集中呈现长尾分布的特征，我们使用了取对数再取整的方式进行分箱，而对于其它特征，我们通过观察分布选择等值分箱或者手动阈值分箱。参数方面，我们通过对比运行，选择了合适的初始化方式和固定的学习率和batchsize。此外，我们积极的应用了官方的预训练权重来对话题和问题描述进行描述，并尝试了多种方式将这些特征加入网络，如多值特征平均，水平拼接，attention加权等等，最终在模型的表现上有不错的提升。

然而，受比赛激烈程度的影响，以及复赛只有一天时间处理新的测试集。在验证了特征占主导因素的情况下，模型之间的差异不是特别大，我们选择了lightgbm作为主要的预测模型，并且可以高效的验证特征，训练结果，最终通过对学习率，特征抽取等参数的调试，我们取得了不错且快速的预测结果。