智源-看山杯专家发现算法大赛

队伍名：救救菜鸡吧
舒秀峰　　刘臣　　孙睿

摘要

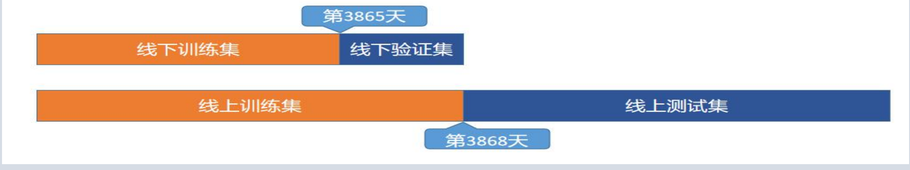
本次比赛结合了知乎的问题邀请与回答的相关场景和数据，需要选手们通过数据挖掘和建模来预测用户回答邀请问题的概率。结合相应的问题分析以及对数据集的探索，我们将该赛题等同为ctr预估的赛题。

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，探索有效的训练模式。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，我们在比赛的最后一天使用了三个lightgbm模型从不同角度进行建模，最终取得了不错的线上成绩。

特征工程

全局统计特征	向量特征	时间统计特征	历时滑窗特征
对问题，用户相关属性，话题构造各类统计特征，如话题的出现次数，问题的出现次数，用户的日出现次数，问题的描述词长，用户关注的话题数目，话题下的问题数等等多个方面构造统计特征。	充分利用问题与话题的词向量，并利用向量间的预选相似度特征，结合用户的历史回答，构造了问题与用户历史回答问题的相似度，用户关注话题与问题话题的相似度等特征。	充分利用了数据中的时间信息，构造了各类时间差的特征。如：问题邀请时间距离问题创建时间的差值，用户上一次回答问题的时间差，两次邀请的时间间隔等特征。	通过滑动窗口的方式构造特征，保证了训练集，验证集，测试集中的特征分布与误差保证一致，并大大减少了训练所需要的数据量，同时也构造出了一些可以反应用户不同长度时间窗内的回答习惯。

模型与训练



为了保证线上线下分数一致性，我们线下尝试了多种策略，主要有如下三种：

1. 随机划分验证集。用该时刻之前的所有数据构造特征，随机划分验证集进行线下验证。
2. 后三天作为验证集。用该时刻之前的所有数据构造特征，去后三天的数据作为线下验证集。
3. 特征提取区间一致，后三天作为验证集。测试集第7天的样本前六天的label信息都缺少，为了保证训练集、验证集、测试集特征分布的一致性，故训练集取[3858, 3864]天，特征提取区间为[3840, 3857]，验证集取[3865, 3867]天，特征提取区间为[3846, 3863]。

经过不断尝试，第3中方案的线上线下分数差别最小，并且保持一致性。

总结

纸上得来终觉浅，要知此事要躬行！通过本次比赛，不仅将理论知识进行了实践，提升了动手能力，更让我们体会了数据科学的魅力。最后衷心感谢举办方提供的这次机会，也希望平台越办越好。