阿里妈妈搜索广告转化预测

黄焯恒　2120171017　毕丽阳　2120160978
李博　2120171028　闫斌　2120171086

源码：https://github.com/Cena21/DataMining_Project

1　引言

搜索广告是一种常见的互联网营销方式，商家（广告主）根据商品特点自主购买特定的关键词，当用户输入这些关键词时相应的广告商品就会展示在用户看到的页面中。随着互联网的快速发展，搜索广告和电商广告在互联网广告中的占比越来越高，成为互联网行业最主要的商业模式之一。与此同时，搜索广告以其巨大的商业价值和研究价值吸引了大量的专家学者，在学术界得到了广泛的研究。

搜索广告的转化率，作为衡量广告转化效果的指标，从广告创意、商品品质、商店质量等多个角度综合刻画用户对广告商品的购买意向，即广告商品被用户点击后产生购买行为的概率。举例来说，用户在淘宝搜索栏输入“女装”并点击，相关的女装列表将会展现给用户，用户点击感兴趣的女装进入详情页，通过查看商品介绍、店家信誉、用户评论等信息综合决定是否购买，如果有 M 个用户进入同一商品详情页，其中 N 个购买了该商品，那么该商品的转化率为成交总数和点击总数的比值（N/M）。在这个过程中，如果能够将转化率高的商品返回给用户，那么用户看到的商品正好就是想要购买的商品，这样用户将会更快速地找到喜欢的商品，从而提高用户体验；另一方面，广告每被用户点击一次商家都要付出一定的成本，如果广告被点击却没有成交，广告主将白白付出成本，而如果展现给用户且被点击的广告商品都产生了购买，那么商家虽然付出成本但还是能从成交中获得收益。总结来说，准确预估转化率，能够使得广告主匹配到最可能购买自家商品的用户，提升广告主的投入产出比（ROI）；另一方面，也能让用户快速找到购买意愿最强的商品，从而提升在电商平台中的用户体验。

阿里巴巴（淘宝、天猫）是中国最大的电子商务平台，为数亿用户提供了便捷优质的交易服务，也积累了海量的交易数据。阿里妈妈作为阿里巴巴广告业务部门，在过去几年利用这些数据采用深度学习、在线学习、强化学习等人工智能技术来高效准确地预测用户的购买意向，有效提高了用户的购物体验和广告主的 ROI。然而，作为一个复杂的生态系统，电商平台中的用户行为偏好、商品长尾分布、热点事件营销等因素依然给转化率预估带来了巨大

挑战。比如，在双十一购物狂欢节期间，商家和平台的促销活动会导致流量分布变化剧烈，在正常流量上训练的模型无法很好地匹配这些特殊流量。如何更好地利用海量的交易数据来高效准确地预测用户的购买意向，是人工智能和大数据在电子商务场景中需要继续解决的技术难题。

2 模型介绍

2.1 LigthGBM

xgboost 的出现，让人们告别了传统的机器学习算法。而 lightGBM 是微软推出的一个新的 boosting 框架，它在一定程度上超越了 xgboost 的性能。顾名思义，lightGBM 包含两个关键点：light 即轻量级，GBM 梯度提升机。lightGBM 在保证和 XBGoost 精度相当的前期下，提升了速度。LightGBM 是一个梯度 boosting 框架，使用基于学习算法的决策树。它是分布式的，高效的，有以下优势：

（1）更快的训练效率

（2）低内存使用

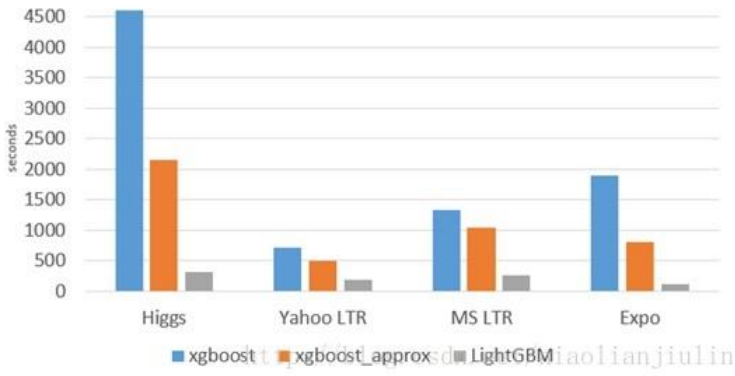
（3）更高的准确率

（4）支持并行化学习

（5）可处理大规模数据

（5）与常用的机器学习算法进行比较：速度很快

下图为 LigthGBM 与其他算法性能的比较图：



概括来说，lightGBM 主要有以下特点：

（1）基于直方图的决策树算法。

（2）带深度限制的 Leaf-wise 的叶子生长策略。

（3）直方图做差加速。

（4）直接支持类别特征(Categorical Feature)。

（4）Cache 命中率优化。

（5）基于直方图的稀疏特征优化。

（6）多线程优化。

其中，基于直方图的决策树算法以及带深度限制的 Leaf-wise 的叶子生长策略是人们关注的中心。直方图算法的基本思想是先把连续的浮点特征值离散化成 k 个整数，同时构造一个宽度为 k 的直方图。遍历数据时，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。带深度限制的 Leaf-wise 的叶子生长策略为 Level-wise 过一次数据可以同时分裂同一层的叶子，容易进行多线程优化，也好控制模型复杂度，不容易过拟合。但实际上 Level-wise 是一种低效算法，因为它不加区分的对待同一层的叶子，带来了很多没必要的开销，因为实际上很多叶子的分裂增益较低，没必要进行搜索和分裂。Leaf-wise 是一种更为高效的策略：每次从当前所有叶子中，找到分裂增益最大的一个叶子，然后分裂，如此循环。因此同 Level-wise 相比，在分裂次数相同的情况下，Leaf-wise 可以降低更多的误差，得到更好的精度。Leaf-wise 的缺点是可能会长出比较深的决策树，产生过拟合。因此 LightGBM 在 Leaf-wise 之上增加了一个最大深度限制，在保证高效率的同时防止过拟合。

2.2 CatBoost

CatBoost 具有提供分类列索引的灵活性，这样就可以使用 one_hot_max_size 将其编码为独热编码（对于所有具有小于或等于给定参数值的特征使用独热编码进行编码）。如果在 cat_features 引数中传递任何内容，CatBoost 将把所有的列都视为数值变量。如果没有在 cat_features 中提供具有字符串值的列，CatBoost 会显示错误。另外，默认整型的列会被默认当作数值处理，因此必须在 cat_features 中对它进行详细说明以使算法将其视为分类。对于剩下的那些具有罕见的比 one_hot_max_size 更大的若干类别的分类列，CatBoost 使用一种有效的编码方法，它与平均数编码相似，但减少了过度拟合。

CatBoost 过程如下：

（1）对输入的观察值的集合进行随机排列，生成多个随机排列。

（2）将标签值从浮点或类别转换为整数

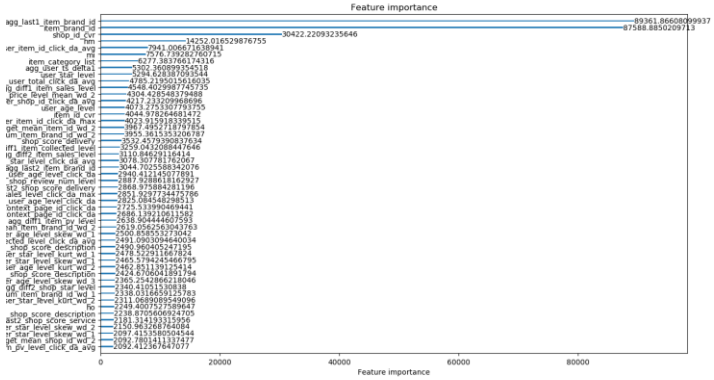
（3）使用如下公式将所有的分类特征值转换为数值：

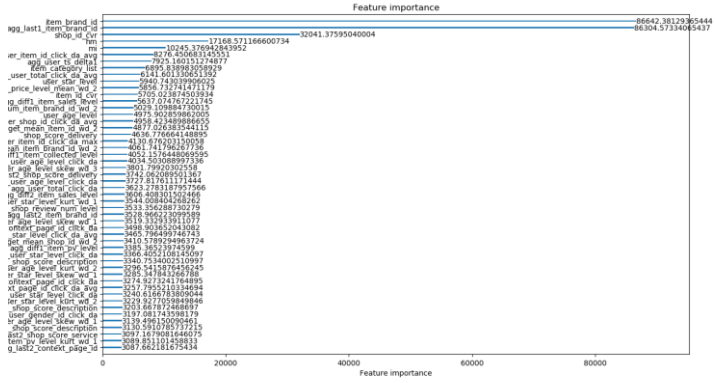
Avg_target=(countInClass+prior)/(totalCount+1)

其中，CountInClass 是指，对于那些具有当前分类特征值的对象来说，其标签值等于"1"的次数。Prior 是分子的初始值。它由起始参数决定。TotalCount 是对象的总数（直到当前的一个），它具有与当前的对象相匹配的分类特征值。

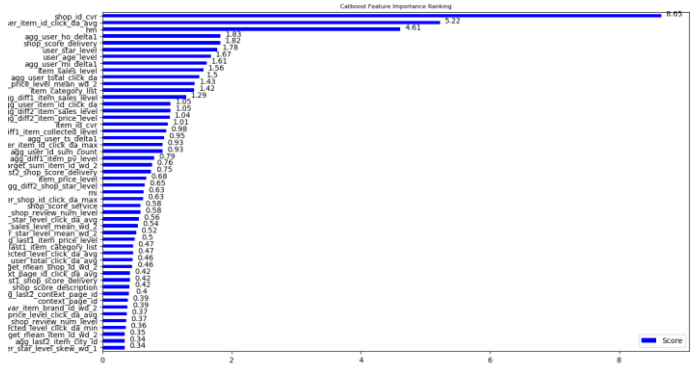
3 实验结果及分析

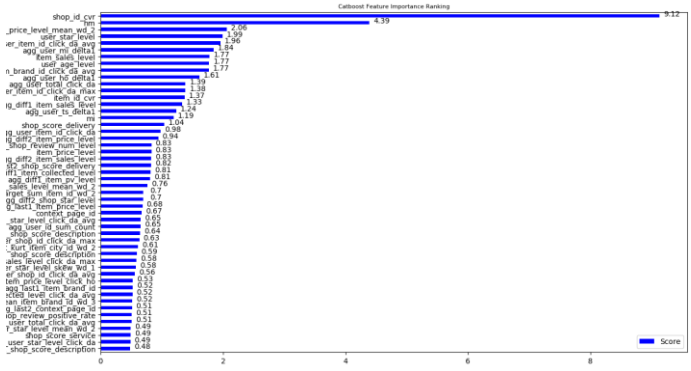
下图为 LightGBM 模型特征重要程度的输出结果





下图为 Catboost 的特征重要程度实验结果：





从 LightGBM 模型输出结果可以看出，agg_last1_item_brand_id 以及 item_brand_id 这两个特征占有很重要的地位，从 Catboost 模型输出结果中可以看出 shop_id_cvr、hm、er_item_id_click_da_avg 特征占有重要地位。

最终，经过分析预测得到如下结果：横坐标表示广告转化率，纵坐标表示某个概率区间的广告数统计。可以看出，广告的一般转化率比较低，这是符合实际情况的。