**关于2018 AJCAI大赛（阿里妈妈搜索广告转化预测）的回顾和反思**

近年的AJCAI大赛和去年的优惠券使用预估一样属于CTR类的问题。CTR问题一直时电商、社交以及搜索引擎领域的热门研究话题。因为这些企业通过广告的投放来实现商品销售和广告推广等方式获取利润。而对于CTR，哪怕是微小的提升，都可以给企业带来非常可观的业绩增长。

在过往的比赛中，已有多个比赛涉及CTR预估的问题，例如Kaggle已举办过两次CTR比赛，2017 AJCAI大赛的关于优惠券使用问题，也可以看作是CTR问题。

CTR问题可以简单描述为给定用户、商品、场景、时间等一系列的原始特征来预估用户是否会购买这个商品或者点击这个广告。其形式化的描述为一个条件概率问题，即P(is\_trade|features)的问题。

虽然本次参赛而未能进入复赛，但是在比赛过程中我也深入阅读了几篇相关论文，觉得对该类问题有了全新的认识，收货颇丰，因此写下此文来总结CTR问题的数据处理和建模方法。

1. **大赛简介（摘抄自大赛官网）**

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。

本次比赛以阿里电商广告为研究对象，提供了淘宝平台的海量真实交易数据，参赛选手通过人工智能技术构建预测模型预估用户的购买意向，即给定广告点击相关的用户（user）、广告商品（ad）、检索词（query）、上下文内容（context）、商店（shop）等信息的条件下预测广告产生购买行为的概率（pCVR），形式化定义为：pCVR=P(conversion=1 | query, user, ad, context, shop)。

结合淘宝平台的业务场景和不同的流量特点，我们定义了以下两类挑战：

（1）日常的转化率预估

（2）特殊日期的转化率预估

本次比赛为参赛选手提供了5类数据（基础数据、广告商品信息、用户信息、上下文信息和店铺信息）。基础数据表提供了搜索广告最基本的信息，以及“是否交易”的标记。广告商品信息、用户信息、上下文信息和店铺信息等4类数据，提供了对转化率预估可能有帮助的辅助信息。

1. CTR问题的建模方法

CTR问题的解决方法总的来说可以分为两种方法：一种是基于强特征的方法，这就是说需要建模人员通过特征提取、特征构造等特征工程方法来构建出强特征，然后使用简单的逻辑斯蒂回归的方法来进行建模；另一种是基于弱特征的方法（暂且这样命名），这种方法不需要构建出强特征，而是将原始数据简单处理后直接进行建模，让模型自动发现数据的内在规律和模式，这种方法的输入数据一般都是非常高维的且非常稀疏的矩阵，典型的算法包括Facebook工程师提出的GBDT+LR算法，以及台湾大学实现的FM、FFM算法等。此外，现在深度学习非常火热，但深度学习的方法也是属于弱特征方法，通过构建复杂神经网络来学习数据的内在知识，实现高精度的预测。

下面逐一介绍在CTR领域常用算法：

1. LR（Logistic Regression）

看过很多博客，LR算法一直是很多企业在进行二分类问题的常用方法。CTR也不例外。该方法简单，可解释性强。一旦特征工程做的好，该方法的准确率也非常高。

这是我本次参赛使用的模型，不过该算法的高准确性需要建立在强特征的基础上。因此，找出和目标变量强相关的特征是使用该算法成功的必备条件。

有学者提出了OWLQN算法，解决了逻辑斯蒂回归中损失函数在L1正则化非连续可到问题。该方法是一种Batch learning的方法，可以收敛到全局最优，且收敛速度快。

1. GBDT + LR

在LR中，我们说过，使用LR做为预测模型，需要有很好的特征工程的工作基础。但是往往在实际应用中，我们并不能很好地挖掘出非常有用的特征。而且，挖掘一个特征需要花费非常多的时间，包括特征构造，特征验证等等。最终，我们会发现，这样的工作往往吃力不讨好，模型的泛化能力和预期的并不一样。

Facebook的工程师在进行CTR预估的工作中，提出了GBDT+LR的融合模型。在这种模型中，我们并不需要人工构造特征，而是让模型自动发现特征。该方法结合了GBDT非线性模型能拟合非线性特征的特点，以及LR线性模型很好的扩展性和模型训练速度。

1. FM/FMM

FM算法（Factorization Machine）称为因子分解机， 是一种基于矩阵分解的机器学习算法。该模型自动拟合特征间的交互，处理高维稀疏问题，因此对稀疏的数据具有很好的学习能力时该算法的最大特点。

该算法的实现可以参看台湾大学的开源FM算法包。

和GBDT+LR的融合模型思路一样，由于FM算法的输入是高维稀疏矩阵，因此我们可以使用GBDT来自动学习特征，并将学习到的特征进行编码成高维矩阵。台湾大学的学生曾经就使用GBDT+FM算法夺得两次kaggleCTR预估大赛的大奖（一次冠军，一次亚军）。

针对FM和FFMs算法的详细算法细节，可以参看美团技术文章：[深入FFM原理与实践](https://tech.meituan.com/deep-understanding-of-ffm-principles-and-practices.html)

1. 神经网络
2. 我的特征工程及建模过程

CTR的数据和金融领域的用户欺诈数据很相似，都是正样本很少，而负样本很多，二者的比例常常可以达到1：50以上。但是CTR问题的数据一般不会像金融领域一样采用过采样或降采样的方法来处理数据集，而且对整个数据集完全保留地进行数据预处理，然后直接进行建。最后，通过评估指标（AUC、ROC）等来评估模型的好坏。

1. 方法改进及反思