目录

[摘要 1](#_Toc35761664)

[Abstract 2](#_Toc35761665)

[第一章 绪论 3](#_Toc35761666)

[1.1．课题背景与研究意义 3](#_Toc35761667)

[1.1.1课题背景 3](#_Toc35761668)

[1.1.2 研究意义 4](#_Toc35761669)

[1.2． 国内外研究现状与发展趋势 5](#_Toc35761670)

[1.2.1 国内外研究现状 5](#_Toc35761671)

[1.2.2 发展趋势 8](#_Toc35761672)

[第二章 数据处理 9](#_Toc35761673)

[2.1．探索性分析 9](#_Toc35761674)

[2.2．数据预处理 13](#_Toc35761675)

[第三章 特征工程 13](#_Toc35761676)

[3.1. 特征抽取 13](#_Toc35761677)

[3.2. 特征选择 14](#_Toc35761678)

[第四章 模型预测及优化 15](#_Toc35761679)

[4.1．模型介绍 15](#_Toc35761680)

[4.1.1逻辑回归 15](#_Toc35761681)

[4.1.2 随机森林 15](#_Toc35761682)

[4.1.3 GBDT 16](#_Toc35761683)

[4.1.4 LightGBM 16](#_Toc35761684)

[4.2．建模及优化 17](#_Toc35761685)

[4.3．预测结果 18](#_Toc35761686)

[第五章 结论与总结 19](#_Toc35761687)

[5.1. 研究结论 19](#_Toc35761688)

[5.2. 思考与总结 19](#_Toc35761689)

[致谢 20](#_Toc35761690)

[参考文献 21](#_Toc35761691)

[附录 22](#_Toc35761692)

# 摘要

信息时代的飞速发展使得在线广告成为众多互联网科技公司的支柱产业，并且随着分析与计算能力在大数据时代的进一步发展，如何将广告更加精确且更加高效的匹配到实际用户以提高自身利益将会是广告平台与广告投放商更加关注的热门问题。广告预测投放针对的主体就是客户，预测的关键就是针对用户所投的广告是否有效，即用户的点击率（CTR)，CTR是指推送给某个顾客的广告信息是否会被点击。无论从为了提高盈利能力，还是说收到更好的广告效应，都必须向客户推送点击率尽可能大的商品。

广告投放过程中将会产生极大规模的数据，所以广告用户点击率预测将会是一个拥有庞大数据的决策问题，所以在实际的操作过程中，要尽可能考虑如何处理拥有的数据信息，我们必须要做到既要关注预测结果的准确性，还需要考虑处理问题的效率。

CTR预测常常会用到逻辑回归，但以传统的批量算法无法有效地处理超大规模的数据集和在线数据流，同时逻辑回归在非线性问题时表现难以达到预期效果。本论文将采用梯度提升迭代决策树(GBDT)进行点击率的预测。在GBDT算法的工程实现时，因为LightGBM模型在以往的相关数据竞赛或案例中都有很好的表现，在保证较高的准确率的同时，也能够很好的应对大规模数据问题的处理，所以采LightGBM构建模型对用户点击率进行预测，并使用K折交叉验证的方法训练模型进一步优化。

本论文主要完成了一下任务：清洗理解庞大数据集，为模型训练做好数据准备；完成特征工程，选择出对于模型最有意义的特征输入；尝试多个模型，选择较为高效准确的模型，完成对于用户点击率预测的工作。

关键词：广告匹配；点击率预测；大数据处理；GBDT；LightGBM

# Abstract

The rapid development of the information age has made online advertising a pillar industry for many internet technology companies, and as analytics and computing power grow in the big data era, how to match the ads more accurately and efficiently to the actual users to improve their own interests will be a hot issue that advertising platforms and advertisers pay more attention to. The subject of advertising prediction is the customer. The key of prediction is whether the advertising targeted at the user is effective, that is, the user's Cross-Through Rate (CTR) refers to whether an advertising message pushed to a customer will be clicked. In order to increase profitability, or to receive better advertising effect, must be pushed to the customer click-through rate as large as possible products.

In the process of advertising, huge amounts of data will be generated, so the click-through rate prediction of advertising users will be a decision-making problem with huge amounts of data, so in the actual operation, to consider as much as possible how to handle the data we have, we must focus on both the accuracy of the predictions and the efficiency of the problem.

Logistic Regression is often used in CTR prediction. The traditional batch algorithm can not deal with large-scale data sets and online data streams effectively, but the performance of Logistic Regression in nonlinear problems is difficult to achieve the expected effect. In this paper, Gradient Boosting iterative Decision Tree(GBDT) is used to predict the click rate. In the engineering implementation of the GBDT Algorithm, because the LightGBM model has a good performance in previous relevant data competitions or cases, it can deal with large-scale data problems well while ensuring a high accuracy rate, therefore, LightGBM model is used to predict the CTR of users, and the training model is further optimized by k-fold cross-validation.

This paper mainly completes the following tasks: Cleaning and understanding large data sets, preparing data for model training; completing feature engineering, selecting the most meaningful feature input for the model; trying multiple models, select a more efficient and accurate model to complete the user click-through rate prediction.

Keywords: Ad Match; CTR Prediction; Big Data Processing; GBDT; LightGBM

# 第一章 绪论

## 1.1．课题背景与研究意义

### 1.1.1课题背景

互联网产业如今已是世界的一个支柱产业。互联网自出现以来，就一直在影响着人类生活的方方面面，如今更是渗透到了日常生活中的细微之处，我们的生活娱乐学习都与互联网有着密不可分的关系。这一次受到新型冠状肺炎COVID-19的影响，虽然我们国家为了人民的生命安全，为了尽早的打赢疫情防控阻击战，实行了严格的人员流动管制措施，对大家的工作学习都造成了巨大影响。防控疫情必须付出一定的代价，但是互联网的快速发展也为此次防疫做出了巨大贡献，学生们在网络上开展课堂学习，公司员工在家办公，尽量的将必要的工作运作起来。这次的疫情让大家更直观的感受到了互联网的实际意义，互联网的快速发展还在我们没有特别关注的地方有很多科学研究成果为我们的生活实现更加便捷，互联网产业容纳了很多传统行业，已经成为世界各个国家最关注的产业之一。

互联网产业有一个很突出的特点，就是追求使用比传统行业更少的资源，实现更加高效准确的工作效率。传统的广告行业主要是以广告位置，时间，面积，数量等来大致估算广告的收费，大多都是靠经验来设定的收费标准。但是自从进入互联网时代，不必再像传统的广告行业通过广而告之的方式以规模效应来招揽顾客，互联网广告追求将广告推送给潜在的客户群体，把消息通知到个体。用户点击率是指用户在看到广告后点击进行进一步深入了解的人数比例，用户点击率可以很明确的了解一条广告的实际效果如何。这对于广告主来说也是一个极其振奋的消息，这将意味着广告主可以将资金投入到效果更好的广告创意，实现收益最大。广告主愿意将资金投入到实际效果更好的方案，如果广告平台能够多接受广告效果好的提案，实现每条广告都是最可能被用户进行点击了解，那么收入也会随着增长。如此的影响之下，以用户点击率来进行收费的互联网计算广告越来越主流，如何预测每条广告创意用户点击率也就成为了互联网行业的热门问题。

计算广告[1]能够如此快速发展的原因主要有两个：计算能力的不断迭代增长与用户点击率预测模型的快速更新换代。最初的广告用户点击率预测仅仅只有逻辑回归一个模型可供使用，由于逻辑回归模型的缺点与限制，能够处理的数据类型与数据量都是有限的，预测的结果还不能够对实际生活有太大的影响意义，所以很多研究学者开始投入大量的精力改进原有模型或者开发新的模型，发展到如今模型数量在不断攀升，现在的用户点击率预测在选用模型上有充分的选择权，在不同的应用场景下选择不同的模型，以应对不同的数据特征，能够获得更好的效果。在模型发展的同时，计算能力也是很重要的支持，互联网刚起步时能处理的数据屈指可数，如今随着计算机运算能力有了质的提升，很多之前即使开发出来也难以验证的模型也成为现实。特别现在大数据时代，大规模数据通过优质模型运算预测的结果已经不可同日而语。

真实的商业环境下，关于在线广告的用户点击率预测实际是一个关于大规模非线性稀疏数据的机器学习的问题，在传统机器学习时代，由于传统机器学习的模型对数据处理的能力有限，所以一些隐藏的重要信息未能参与到决策。到了深度学习时代，越来越多在处理非线性的稀疏数据优秀的深度学习模型，可以将越来越多的隐藏信息转化为关键特征进行在线广告的用户点击率预测。

互联网产业的成果往往不是单一的，计算广告的研究也是与很多其他相关研究所结合，相辅相成，例如计算广告和推荐系统就是密切相关的两个项目。另一方面，计算广告是一个涉及大规模的文本与数据收集，特征获取，分类模型优化等等的混合问题，其中的一些研究的结论也可以推广至其他的应用。计算广告能够如此快速的发展，也有很多其他方面的辅助，从一开始逻辑回归在其他分类问题上的成熟运用才能被采用在最初的用户点击率预测问题，如今深度学习的快速发展也给计算广告问题带来了大量便利。

### 1.1.2 研究意义

关于计算广告的研究，首先最为关注的就是广告最后的收益问题。虽然现在随着模型与计算能力发展，用户点击率预测已经取得了不错的效果，但是还有很大的提升空间。并且由于现在的规模比较庞大，每一次的微小提升都可以为广告平台带来不容小觑的收益加成，特别是现如今互联网用户激增，大多数公司都愿意将资金投到互联网的广告平台，互联网公司的广告收入占据了公司财政的很重要一环。所以从经济利益方面来说，优秀的用户点击率预测将会使得相关产业的利润收入提升。

互联网资源利用率一直都在追求更加极致的使用，对于广告资源来说也是。传统的以广告量来收取费用的方式是以一定的规模去覆盖用户，其中大部分都不属于广告的面向群体，并且在符合的群体中，也还存在一大部分不属于潜在客户，这对于广告资源的利用来说是一种极大的浪费，如果能够将广告主，平台与用户相匹配，这就可以有针对性的减少广告的投入成本。无论对于广告主与广告平台来说，精准的广告投放所节约的资源都是共享的，对于用户点击率的深入研究是提升整个行业的资源利用率，实现的效益最大化的有效措施。

互联网广告平台是以自身大量的用户数量来吸引到广告主的广告投放方案，但是过多的用户所不期望的广告投放可能会造成用户的流失，如何在投放广告的同时留住甚至增加用户，也是互联网公司所思考的。提升广告的质量是解决这问题的根本，广告的质量是指针对不同用户的喜好所提供的不同广告，满足更多人的需要才是投放的原则，所以在预测更加准确的用户点击率后，广告更加符合用户的使用喜好，更加满足用户的人性化需求，可以提升用户的使用满意度，留存用户数量。

如前面所提到，计算广告是一项复杂的工程，计算广告的发展得益于其他互联网产业的发展，也在直接或是间接的为其他研究起着促进的积极作用。在提升计算广告的精确度与效率的同时，也是在为整个互联网的发展增砖添瓦，很多细节的内容都是互联网不断发展的基石，特别文本分析，数据处理，特征工程都是机器学习等内容的关键步骤。

综上所述，计算广告关于用户点击率的不断深入研究，具有多方面长久的现实意义，既能提高社会的经济效益，也是互联网科学技术不断进步的组成。

## 1.2． 国内外研究现状与发展趋势

### 1.2.1 国内外研究现状

关于CTR的发展大致可以分为两个部分，即传统机器学习时代与深度学习时代两个部分，对于前者是从最基本的逻辑回归(Logistic Regression, LR)开始发展，2010后开始不断的进化出了因子分解机(Factorization Machine, FM),梯度提升树(Gradient Boosting Decision Tree, GBDT)等效果更好，训练速度更快的CTR模型；后来进入深度学习时代之后，模型发展更是日新月异，更多在各类数据集，各种应用场景都有很好表现的模型不断涌现，今天关于计算广告用户点击率预测的模型训练速度与表达能力都已经发展到全新的水平。

传统机器学习时代最开始是以逻辑回归模型开始，在2010年以前针对用户点击率预测都是使用了逻辑回归。逻辑回归在之前是比较好的模型，都是也有很多的弊端，逻辑回归主要是针对线性数据有很好的表现，在其他一些方面不是特别适合，但是由于计算能力的限制以及当时数据量还不是特别大的缘故，逻辑回归还是能够较好的完成预测任务。随着计算能力的不断发展，很多之前使用时可能训练比较复杂导致能以得出结果的模型也开始发挥作用，同时数据量大幅增加，在大数据时代，逻辑回归已经注定要被取代，虽然逻辑回归模型只是最简单的模型，但是它的重要性还是不可忽视的，逻辑回归依旧是CTR模型的重要核心与研究基础；随着数据的不断增加，逻辑回归模型已经不再能够胜任关于CTR预测的问题，随即FM在2010年提出，FM的提出比较好的解决了在数据处理之后数据稀疏性的问题，并且再后来基于FM提出了FFM(Field-aware Factorization Machine )，FFM比FM的进步在于更多的引进对于模型更有帮助的信息。FFM的提出在当时的CTR预估赛事中大红大紫，一举斩获了多个冠军，后来美团等将这一算法在公司内部CTR中使用；国内的阿里巴巴曾有一个模型LS-PLM(Large Scale Piece-wise Linear Model)，也会被成为被称为M逻辑回归(Mixed Logistic Regression)，这个模型被应用于各类场景的。从后一个名称可以看出，这个一个对于逻辑回归进一步的推广，M逻辑回归是对于线性回归基础上加入了聚类的思想，M逻辑回归采取了划分-训练的模式 因为有了分而治之的思想，所以对逻辑回归只具有线性能力这一缺点进行改进，M逻辑回归模型可以从大量稀疏数据中捕捉非线性模式，从而避免了繁重的特征工程的工作。自2012年以来，MLR模型已经成为阿里巴巴在线展示广告系统的主要点击率预测模型，每天为数亿用户服务，；2014年，相比于阿里巴巴对于逻辑回归的改进使用，Facebook推出了一个采用GBDT加上逻辑回归混合的模型解决方案，这一方案相对于其他模型，机器高效的完成特征选择与组合的问题，所以开启了特征工程模型化的进程，不过这一方案也是有它的缺点，比较容易发生过拟合的情况，之后到了深度学习时代的各类网络结构，以及Embedding技术的应用都是特征工程化趋势的一种延续发展；后来Google基于一种是不是更新更快的模型能够更好的适应变化，模型的效果是不是更加显著的思考，又推出了名为FTRL(Follow-the-regularized-Leader)的 实时在线模型训练方法，在2010年提出了FTRL的基本思路。三年后，FTRL成为现实，虽然当时Google只是针对逻辑回归做出的工程实现，但由于FTRL是一个在线的模型训练方法，事实上FTRL是可以针对任何梯度下降训练的模型进行训练的，后来快速的成为了在线学习的主流。如图（Figure 1.1）,展示了CTR预测模型在前深度学习时代，模型的演化发展，可以更加明显的看出，在深度学习到来之前，虽然模型有了不断的进步发展，但是一直是以逻辑回归为模型发展的基础，各类模型通过改进或发展逻辑回归，适应不同的应用场景。LS-PLM模型主要以分而治之的思想去弥补逻辑回归的难以应对大规模的非线性稀疏数据的缺点；FM与FFM两个模型在面对这一问题时，则采用了特征交叉的方法；GBDT模型与逻辑回归融合的案例则是发挥不同模型优点去改进模型存在的缺点；FTRL模型同时兼顾稀疏性与实时性，成为当时在线学习的主流模型。

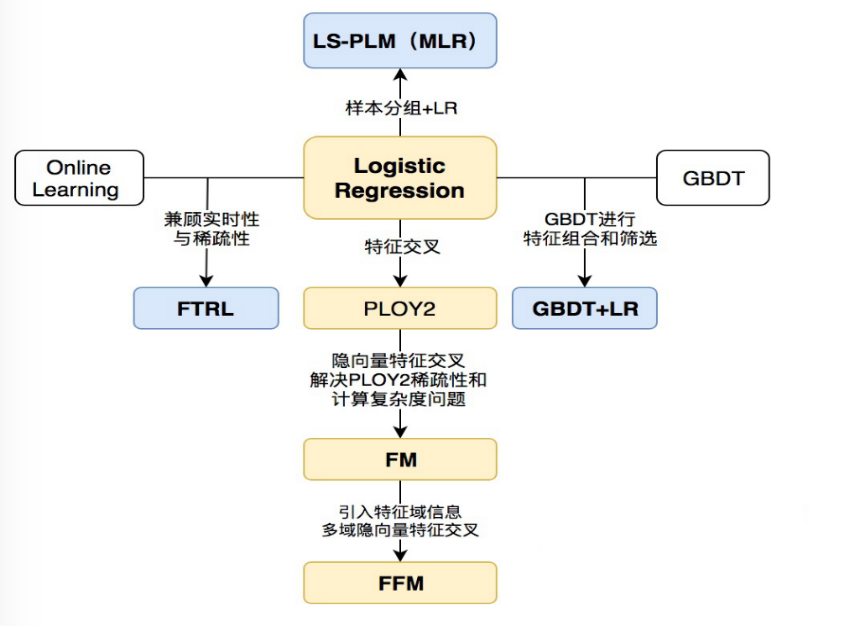


Figure 1.1

深度学习时代的到来，使得CTR的发展速度加快的同时，也带来也种类繁多，思路不一的优秀深度学习CTR预测模型，如图（Figure 1.2）展示的模型发展与之前传统的CTR模型不同，深度学习时代发展更加多样化，模型发展种类更多，能够更好的应对不同的数据场景，以深度神经网络(Deep Neural Network,DNN)为基础发展起来了众多模型，依靠神经网络的强大学习能力，基于深度神经网络发展起来的模型基本都具有了对特征组合的自动学习能力，但是不是所有的特征组合对于所研究的课题的最后决策都是有用的，所以学习效率不会太高，而且由于隐式的特征组合学习方式也带来了结果不可解释性的缺点，模型在深度神经网络基础上发展过程中也在不断解决这些不足。Deep Crossing模型可以说是深度学习CTR模型的最典型和基础性的模型，因为它涵盖了深度学习CTR模型预测的最典型的要素，通过加入Embedding Layer将稀疏特征转化为低维稠密特征，用Stacking Layer，或被称做Concat Layer将分段的特征向量连接起来，再通过多层神经网络完成特征组合以及转换等操作，最终用Scoring Layer完成计算。Deep Crossing模型是由微软公司与2016年提出， 开启了计算广告的深度学习时代；后来FNN(Factorization-machine supported Neural Networks)直接在FM上接入若干全连接层,使用FM的隐层向量作为user和item的Embedding，相比于Deep Crossing模型，FNN避免了完全采取随机状态去训练Embedding，有效降低了深度学习模型复杂度和训练不稳定性；2016年上海交通大学(SJTU)提出了PNN(Product-based Neural Network)模型，这一模型的关键在于在Embedding Layer和Fully Connected Layer之间加入了Product Layer,PNN不同于传统DNN(Deep Neural Networks)直接通过多层Fully Connected Layer完成特征的交叉和组合,PNN更加具有针对性，能在不同的特征域进行交叉组合，增强了表达不同数据模式的能力。

Google同时也在2016年发表了Wide & Deep的深度学习框架，从模型名字即可直观的了解到，这是一个模型具有两部分，其中Wide部分擅长大量稀疏特征，使得模型具有记忆性(Memorization),另外的Deep部分则是挖掘特征背后隐藏的数据模式，使得最终模型具有泛化性(Generalization), 两者统一在一起就是最终的模型,Wide & Deep的组合就是为了平衡Wide模型和Deep模型的记忆能力和泛化能力。在Wide & Deep模型中，Wide和Deep部分的输出通过以加权方式合并到一起，并通过损失函数进行最终输出；之后一年，Google再次推出Deep & Cross的混合模型，将Wide部分替换为了Cross，增强了特征的交互能力。Cross网络的基本动机是为了增加特征之间的交互力度，使用多层Cross Layer对输入向量进行特征交叉，特征交叉可以增强模型对于特征的表达能力，Wide部分被替换为Cross的实质也就是对Wide & Deep模型中 Wide部分表达能力不足的问题进行改进。并且在Cross网络中每一层都会进行特征交叉，能够自行高效学习组合特征。后续又发展出同样对Wide部分改进的新模型DeepFM,将其中的逻辑回归改为FM；另外还有对Deep部分进行改进的两个模型，分别是AFM(Attentional Factorization Machines)和NFM(Neural Factorization Machines),其中AFM引入了在自然语言处理常用的注意力机制（Attention Mechanism）。为了训练Attention权重，加入了Attention Net，利用Attention Net训练好Attention权重后，再反馈于FM二阶交叉特征之上，使FM获得根据样本调来整特征权重的能力，NFM对于Deep部分的改进则是将FM的特征交叉部分替换为一个带Bi-interaction Pooling层的DNN模型。

2018年，阿里巴巴提出DIN(Deep Interest Network)，时隔一年又再次升级模型为DIEN(Deep Interest Evolution Network),虽然前后两个模型看似同根同源，但是绝对不是简单的替换升级。DIN模型主要是加入了用户兴趣的特征因素，使得模型能够根据候选商品的不同，调整不同特征的权重，但是DIN模型还有问题就在于用户的兴趣是在不断的变化发展，而且从用户的行为获取的兴趣不一定是准确的。在此基础上DIEN为解决上述问题，在Embedding Layer和Concatenate Layer之间加入了生成兴趣的Interest Extractor Layer和模拟兴趣演化的Interest Evolving layer，通过引入序列模型 AUGRU模拟了用户兴趣进化的过程，最后时刻又将兴趣与其他特征一起进入模型进行CTR的预测。

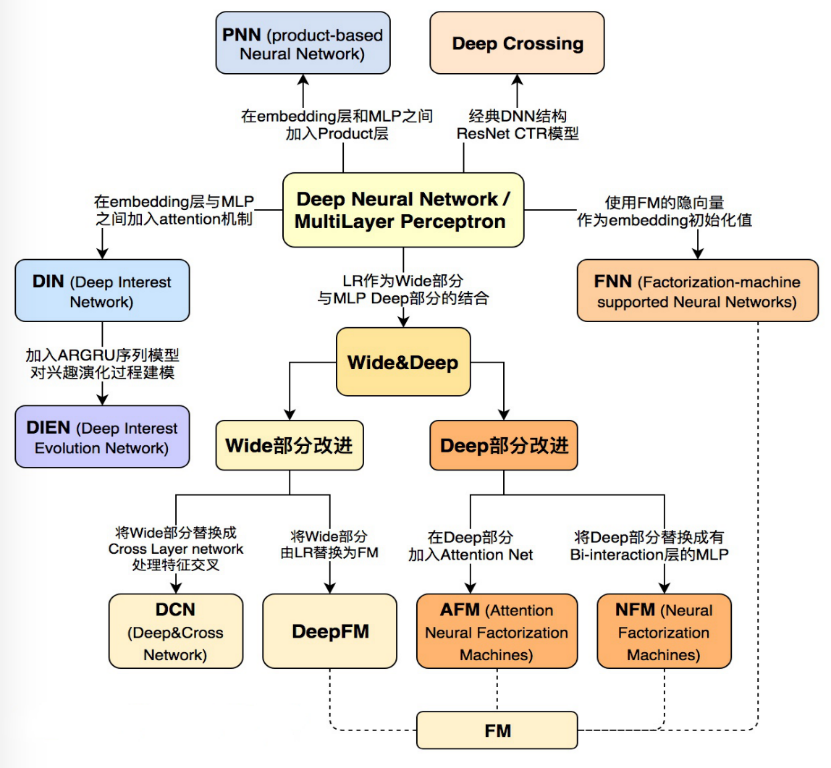


Figure 1.2

### 1.2.2 发展趋势

根据《2018中国互联网广告发展报告》显示，仅2018年互联网广告总收入3,694亿元，年增长率为24.2%，保持了较快的增长速度。由此可以看出中国网络广告市场相当庞大，并且还将在一段时间内会持续的增长。由于目前移动端市场增长迅猛，移动端广告市场体量已经不可忽视，并且由于移动端接入流量的大幅度提高，未来互联网广告向移动端倾斜更加明显，可能在将来成为主要的互联网广告市场。

目前众多的互联网科技公司在利益驱动科技创新的大趋势下，不断依靠自己的科研能力在计算广告这一领域取得新的成果，虽然近一年发布的研究成果还未能实现大规模的商业应用，但是这些成果所提供技术创新和所展示的未来发展趋势觉有重要的意义。这些公司依靠创新的业务模式、产品及技术优势，将为互联网广告健康增长注入了新动力，新生力量的崛起使互联网广告市场的竞争更加激烈，也会倒逼传统广告媒体做出更多技术创新与发展。

阿里巴巴作为国内的互联网巨头，一直在互联网科技创新的前沿，在2019年度SIGKDD的国际顶级数据挖掘会议上，阿里巴巴提出了DSTN(Deep Spatio-Temporal Neural Networks)模型应用于用户点击率的预测，DSTN模型能够考虑到更多关于空间域与时域的辅助信息，DSTN模型关键在于能够更好的了解每种类型的数据与目标广告相互之间的影响作用，强调那些更加重要的隐藏信息，目前阿里巴巴已经将性能最佳的DSTN模型部署在神马搜索，并进行其他一系列的离线实验都表明大幅由于以往的DeepFM和GRU等应用比较多的模型；同样是国内的互联网公司华为在WWW 2019上提出基于卷积神经网络的CTR特征生成方法FGCNN(Feature Generation by Convolutional Neural Network)模型，包括两个部分即特征生成与深度分类器，这样一来就可以将任意两个模型组合起来，利用不同模型的特点优化预测过程；在国外，Google一直以来都站在科技的前沿，在用户点击率预测问题上，Google多次提出引领科技发展趋势的模型与理论，2019年Google公司连续在WSDM 2019和IJCAI 2019两个顶级会议上发表了关于强化学习应用在将来的用户点击率问题的论文，虽然强化学习暂时还没有足够成熟以用于大规模的商业实践，但是这两篇论文给大家的研究指明了比较清晰的方向，并且Google声称已经在Youtube推荐排序层的线上实验中相对线上已有的深度学习模型获得了显著的收益，这更是给研究者们注入更大的动力。首先，在WSDM 2019会议上发布了Top-K Off-Policy修正方案的算法，算法根据用户日志反馈的信息进行模型训练，在Youtube的实验中，该算法取得了近两年最大的reward增长的成绩。之后，在IJCAI 2019会议中又发布了强化学习(Reinforcement Learning)应用于基于候选列推荐系统(Salte-based Recommender Systems),该推荐系统还使用名为Salte-Q的Q-Learning算法进行候选值优化.

基于目前顶级会议的展示研究成果，我们可以看出，未来在对于用户点击率预测问题的研究中需要对以往不能或难以处理的数据信息进行更多的挖掘，找出更多的重要信息，同时一些在其他应用中表现优异的模型也可在用户点击率预测中使用。未来关于用户点击率的一个研究热点可能是强化学习，特别是Google这两篇论文及它在Youtube的成功实验，一方面给强化学习的研究发展指明方向，同时也增强强化学习未来在实际操作中大规模应用的信心。

# 第二章 数据处理

## 2.1．探索性分析

针对于分类预测这类问题，必须对数据有较好的理解才能更好的完成研究目的，所以探索性分析[2]的目的就是为了了解数据的组成成分，数据结构与相互间的影响关系，并对数据做一些基本的可视化，为后续对数据的处理和预测模型的选用提供辅助。

本文的数据来自于2018年科大讯飞AI营销算大赛的初赛以及复赛数据，本次大赛提供了科大讯飞AI营销云的海量的广告数据和用户数据， 本论文将结合此数据讨论如何有效利用这些数据去通过人工智能技术构建预测模型预估用户的广告点击概率， 即给定广告点击相关的广告、媒体、用户、上下文内容等信息的条件下预测广告点击概率。

2018年科大讯飞AI营销大赛的数据分为初赛与复赛两个部分，首先进行数据的融合作为后续的研究数据，将两部分的训练集融合作为实验的训练，以复赛的测试集作为后续测试集使用。其中训练集包含了3000000条数据，测试集含有80276条记录，数据是由34个特征信息与1个关于用户点击的标签组成，数据组成如表1所示，其中分为五类信息，即基本数据，用户信息，媒体信息，广告信息，上下文信息。所有信息数据包含五种数据类型，其中整数与浮点数类型的数据可在后续由64位改为16位，节省内存空间。

整个数据集中user\_tags，make，model，osv，app\_cate\_id，f\_channel，app\_id这7字段含有缺失值。creative\_is\_js，creative\_is\_voicead和app\_paid这三个字段只具有唯一的特征值，对于这些只具有单一特征值的字段在后续操作中对于预测结果没有过多影响，不适合作为影响用户是都i单机的预测因素，在数据处理的过程中可以考虑将其删除。

|  |  |
| --- | --- |
| 表1 数据信息表 | |
| 基本数据 | instance\_id(int64) click(int64) |
| 用户信息 | user\_tags (object) |
| 媒体信息 | app\_cate\_id(float64)f\_channel(object)app\_id(float64)  inner\_slot\_id(object)app\_paid(bool) |
| 广告信息 | adid(int64)advert\_id(int64)orderid(int64) advert\_industry\_inner(object)advert\_name(object) campain\_id(int64)creative\_id(int64)creative\_type(int64) creative\_tp\_dnf(int64)creatice\_has\_deeplink(bool) creatice\_is\_jump(bool)creative\_is\_jump(bool) creative\_is\_download(bool)creative\_is\_js(bool) creative\_is\_voicead(bool)creative\_width(int64)  creative\_height(int64) |
| 上下文信息 | city(int64)carrier(int64)time(int64)provice(int64) nnt(int64)devtype(int64)os\_name(object)osv(object) os(int64)make(object)model(object) |

观察整个训练集，其中正负样本的比例约为1 ：4，如图所示（Figure 2.1）：通过做出用户点击的标签信息的饼图，我们可以看出超过百分之八十的广告投放之后，用户未进行点击。这也从一个侧面反应出目前广告投放之后的效率不够高，有很大的提升空间，所以对于广告用户点击率的预测研究前景还很广阔。

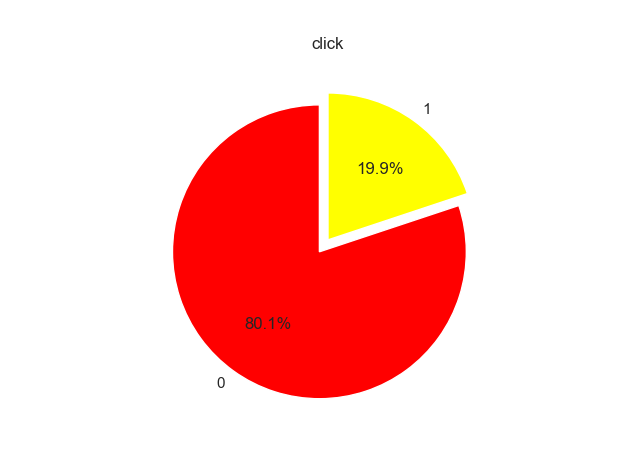
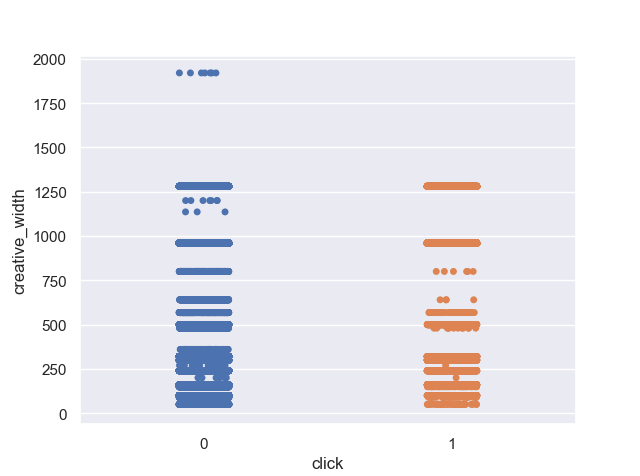
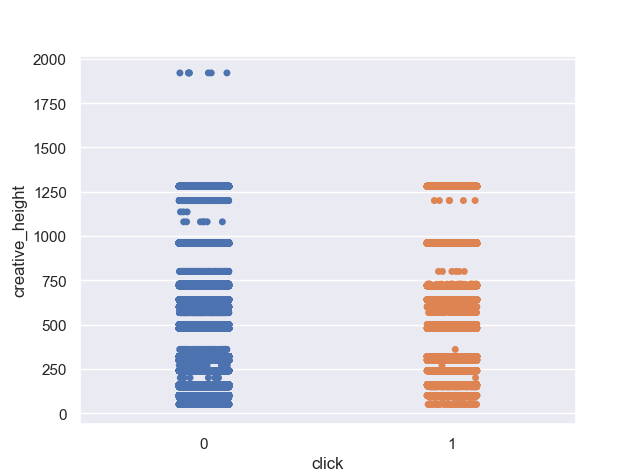


Figure2.1

广告的大小会直接会影响用户的反应，相对于较大的广告可能更加容易引起

用户的注意，通过观察广告的高度与宽度在正负样本中的分布（Figure2.2(a)(b)）可以看出，无论是广告的长度的分布还是说广告宽度的分布在正负样本中都是具有明显差异的，为了更加准确直接的判断出广告创意的高度与宽度是否对用户点击具有积极的意义，对于creative\_width和creative\_height两个类似的信息进行卡方检验，结果显示两者检验的P值小于0.5，说明了广告创意的高度与宽度是明显具有统计学意义的，对于用户点击率具有积极影响，所以两个字段均可

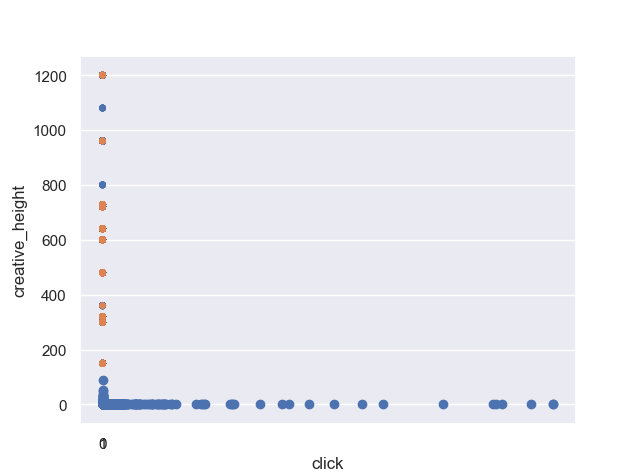
作为预测模型的特征。

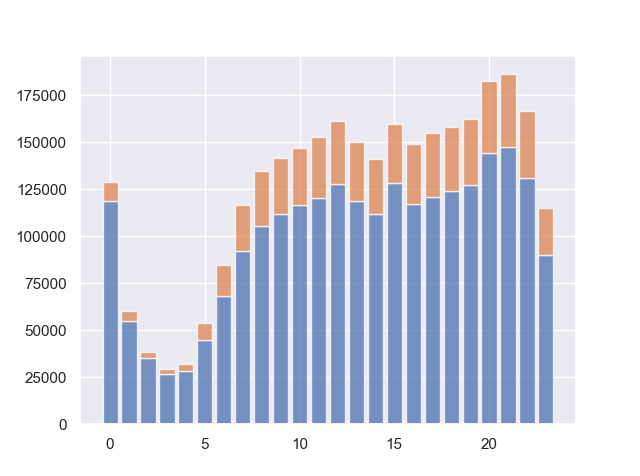
 （a）

(b)

Figure 2.2

再针对其他字段进行分析处理。首先针对广告id进行处理，训练集内删去重复值后有2187个广告id，共有3000000条关于adid的记录，最多一个广告id有124653条记录，最低1条，平均每个广告id含有1372条记录。通过对广告记录数对应广告量的散点图（figure2.3）分析，每个广告得点击记录水平虽有不同，但是大部分广告还是集中在点击量比较少的区域，目前广告投放的准确性还有很大的进步空间；训练集拥有39个广告主id的3000000条记录，分析发现广告主的点击率之间相差较大，但是不同广告主的曝光次数也差异比较大，这就不可确定差异是否与之有关，为此进行卡方检验，可得P值小于0.05，有显著的统计学意义，即广告主对用户点击率有积极的影响作用；训练集时间字段共有2187，重复做多的字段占用其中62个，但是时间字段与点击率相关的主要为

hour部分，所以进行更细致的切分，通过曝光量与点击量的时间(figure2.4)，可以看出与每天时间变化存在一定的关系，后续将一天分为四个时间段进行再次作图分析（Figure 2.5），可以明显看出广告的曝光度与点击度与时间段有着密切的联系。

 Figure 2.3

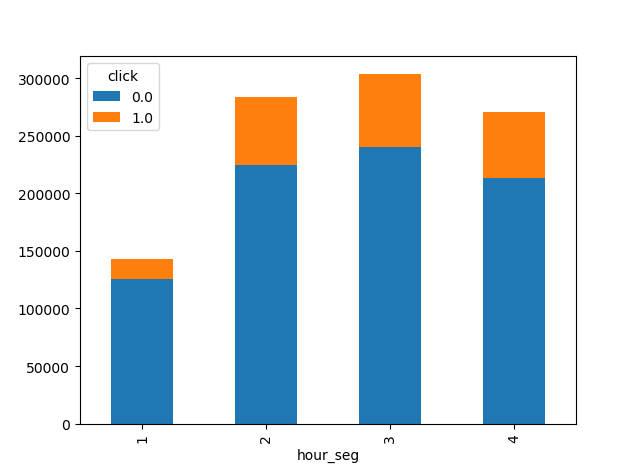
Figure 2.4

Figure 2.5

在分析数值类型的字段时，都应用Numpy第三方库中的unique函数进行探索，这既可以看出一些内容是否需要进一步的细分处理或是合并清洗，同时也可以进行初步的判断每个字段大致的取值范围。由于f\_channel的下一个频道对应于app\_id,初步的实验表明f\_channel缺失值过多，后续考虑选择实验app\_id，剔除f\_channel；在关于model和make的探索中，都有比较高的非缺失率，数据质量很好，都是其中很多的相关点击率都是100%，这一现象主要是由于基数造成的，不具有泛化能力，同时手机的品牌太多，特别模型品牌下属的子品牌可以考虑在后续合并，粒度太细在后续模型预测中也容易发生过拟合的情况；类似的在操作系统及其版本的数据处理中也是一样的，操作系统与后续的版本等重复的数据太多，需要统一进行清洗，可以考虑将os与os\_name去除其中一个，版本也可进行分类合并。

完成对数值类型的字段的初步处理之后，进一步进行数据的可视化工作。首先针对广告id进行可视化分析，首先在广告id曝光次数前十名的统计图（Figure2.6）中，可以看出adid为1537069的广告是整个训练集中曝光次数最多的广告，超过了12万次，曝光前五名次数相差不太大，之后曝光次数梯度下降，可以简单看出曝光次数相对集中。在广告的点击率前十名图（Figure2.7）上可以看出adid为1547080的广告鹤立鸡群，点击率最高几乎达到0.5，后续在0.2至0.3的区间。对用户的设备品牌做类似的分析，做出了品牌数量前七统计图（Figure2.8）与品牌点击率前七的统计图（Figure2.9），其中oppo与vivo既是数量占据榜一榜二，同时也是点击率最高的两个品牌，其他的五个品牌虽然在用户数量上有明显差距，但是品牌的用户点击率没有太大差距。apple与xiaomi两个品牌的数量旗鼓相当，但是apple的用户点击率更高。

## 2.2．数据预处理

样本数据一般不能直接用投入模型进行训练，因为数据一般数据缺失，数据噪声，数据不一致和数据冗余等问题，数据集需要进行预先处理。数据分析的整个过程数据预处理占据着很重要的地位，数据预处理是工作量最大也最复杂的地方，一般分为四个步骤：数据集成、数据清洗、数据变换、数据归约。数据集成是为了将不同的数据源合并；数据清洗主要是为了对数据缺失值，重复数据和异常值进行处理；数据转换和数据归约的操作实际等同于特征工程的部分相关步骤，前者包括数据的归一化，标准化处理，后者则是由于数据集中的部分特征对于模型的预测没有影响，所以对于多余的特征进行删除。

数据预处理首先是进行数据去重，去掉合并后数据中的重复记录，再以time字段开始对时间进行更加细化的划分，以此增加了day和hour两个新的特征，并且将提取出的小时字段内容分为四个时间段，时间的进一步细分对于后续时间序列的分析有很好的提升；关于媒体广告位不重要和不清楚具体含义的后缀删除，留下可以更好使用的部分；广告主所在的行业信息可以分为两级，分为一级行业与二级行业两个新的字段信息，能够在准确描述广告主行业的同时，还能在增加对于预测更有帮助的特征；然后进行机型与品牌的清洗，品牌的清洗过程主要是对用户使用的设备品牌进行合并，例如redmi或mi等可以合并为xiaomi。关于机型的清洗则是对机型中各类符号进行删除；实验数据中的操作系统与版本太过臃肿，其中的操作系统的名称过多，有大小写不一致，表述方式不同等问题，将同样的版本进行合并，并且对于操作系统的版本细分，以三个等级分出新的特征值供特征选择使用。对于操作吸引与操作系统中名称相同的进行合并；在探索性分析时看出，数据中有6个bool类型的字段，将这些bool类型的值转化为能够被直接使用的数值类型，并把用户标签构造为索引的形式；最后，因为数据中有部分缺失值，将缺失值补充完整，数据预处理就完成。

# 第三章 特征工程

## 3.1. 特征抽取

探索性分析时已经提到，有三个字段的特征值为一个定值，这类字段由于没有与用户是否点击发生互动，自然就对于未来的预测没有实质性的帮助，所以首先将只有单个值的特征剔除。

原始特征可以直接进行使用的，直接将这类特征提取出来，包括原有的广告信息，媒体信息，用户信息和上下文信息中可以直接进行使用的特征；再构造统计特征，如计数特征、比例特征和唯一值个数相关特征，最后效果都还不错。比如类别变量的唯一值个数特征(如广告主id有多少个不同的广告id);对于计数特征，一次曝光对应的特征对于点击与否的影响，如果仅仅通过一次曝光与特征的对应关系不够全面，我们可以进一步去挖掘特征的相关特征和属性。构建特征中不同变量对应的曝光次数，提取曝光次数，我们就有了特征对应的属性特征。

特征使用LabelEncoder函数进行硬编码，LabelEncoder的作用是对离散型的分类特征赋以0到n\_class-1的编码，完成编码之后再调用fit\_transform()拟合数据，并转化为标准形式。对除用户标签以外的类别特征进行Labelencoder编码，数据中的分类变量编码为数值变量，由于我们最终使用的树模型，所以数值变量没有距离上的差异。

对于一些属于类别偏好的特征建立比例特征，如每个广告主id针对某一广告id的投放比例，比例特征能够挖掘出广告主的投放相关的隐藏信息。用户标签与其他字段将组成一些平均特征，如针对于不同性别或是其他用户标签信息，广告id对于其的投放比例。

对于用户点击率预测的问题，主导的因素还是用户，所以关于用户的信息重要性是比广告信息更加重要，用户信息值得更加深入的挖掘。user\_tags中包含用户的属性信息，对于结果影响较大，用CountVectorizer()提取用户文本属性，以便更好的表达其中有效属性，助力模型预测的发挥。

## 3.2. 特征选择

特征选择(Feature Selection)注意分为四个过程，分别是产生过程(Generation Procedure)，评价函数(Evaluation Function)，停止准则(Stopping Criterion)和验证过(Validation Procedure)。其中，产生过程就是搜集特征，产生特征子集的过程，搜索方法主要是完全搜索，启发式探索以及随机的搜索；评价函数当然是对一个特征子集进行评价，因为特征选择就是为了选择优秀的有利于最后模型决策的特征，对于没有帮助或是收效甚微的特征就放弃或是减少比重，防止因为特征个数过多，分析特征和训练模型所需的时间太长，甚至引起维度灾难；搜索是一个需要在有限时间内完成的任务，所以停止准则就是一个阈值，当通过评价函数后的值达到这个阈值后就停止搜索；最后不能直接就结束特征选择的过程，因为还需要验证是否已经选择出自己需要的，并且是有效的特征子集。

具体的特征选择方法一般分为三类：过滤法(Filter),包装法(Wrapper)及嵌入法(Embedded)。首先，皮尔森()相关系数，卡方检验和距离相关系数等是属于过滤法分类的，过滤法按照发散性或是相关性对各个特征进行评价，设定阈值或者待选择阈值的个数，选择特征子集；包装法有前向搜索和后向搜索，包装法是使用不同的特征组合来不断测试学习算法进行特征的选择，通常情况下会选用普遍效果较好的算法，例如随机森林(Random Forest，RF)，支持向量机(Support Vector Machine, SVM)，K最近邻(kNN，k-NearestNeighbor)等;嵌入法分别是基于惩罚项特征选择和基于学习模型的特征排序，先使用某些机器学习的算法和模型进行训练，得到每一个特征的权值系数，根据系数从大到小排序选择特征，思路上是类似于过滤法，在波士顿房价数据集上使用sklearn的随机森林回归时，就可以采用基于学习的特征排序进行单变量选择。

本论文主要采用的特征选择的方法为结合特征重要性的F-score。针对于本论文采用的模型，在未对模型进行细致的参数调优之前，计算出不同的特征对于模型的贡献，也就是每个特征所具有的重要性，重要性也可作为是特征对于模型的准确率，这一重要性将是后续选择特征的重要条件；根据特征的重要性，按照降序的方式进行排序，根据重要性及其他的参数以F-score的方法做出特征选择，选择出前k项的特征值。

# 第四章 模型预测及优化

## 4.1．模型介绍

### 4.1.1逻辑回归

逻辑回归[3]是CTR预测问题最基本的模型，因为逻辑回归有很强的解释性，而且训练方法简单，所以首先使用逻辑回归进行模型训练预测。逻辑回归并不单单应用于CTR问题，在很多工业模型中都采用逻辑回归，逻辑回归在概率预测方面的应用场景较多，针对于用户点击率预测这一类的二分类问题，逻辑回归是比较好的首选。

首先尝试模型是否适应于所以采用了初赛的数据经过简单处理进行预测，首先以NULL进行缺失值填充，并将在探索性分析时提到的只具有单一特征值的三个字段删除；数据中含有一部分是不属于数值类型的字段，所以将这一类的字段编码，转化为模型可以利用的数据类型，编码完成之后便分为训练集与测试集，开始训练模型。

### 4.1.2 随机森林

由于在之前的泰坦尼克号乘客生存预测中，使用过随机森林模型，所以在尝试过逻辑回归以后，以相同的数据特征处理方式准备数据，再以随机森林模型进行参数训练，保持与逻辑回归一致，不进行参数调优，保存最终结果。

随机森林也是传统机器学习中应用比较广泛的分类模型，随机森林即代表是由多棵决策书组成的决策森林。随机森林在应对众多的操作场景都可以产生高准确率的分类器。随机森林由两大特点就是随机选择数据以及特征选择的随机性。数据的随机选择可以使用相同的数据集产生不同的数据子集，相同的特征选择随机性还能使相同数据子集特征不一致。这两大特性使得随机森林有足够的样本产生分类器，并保证分类器具有较好的准确率。

### 4.1.3 GBDT

本论文首先采用较为传统的模型进行训练，但是由于逻辑回归和随机森林的局限，最终更改了模型。最初采用了曾经在泰坦尼克号事件相关预测所使用过的随机森林模型进行尝试，然后有使用的CTR预测的基础模型逻辑回归进行模型训练及预测。两次尝试最终的效果都比较差，与预期有较大的差距，并且处理较大规模数据速度太慢，在CTR预测中不太适用。

根据尝试结果及其他模型的学习结论，综合考虑较好的数据处理效率和准确性，决定采用在大量预测点击率和搜索排序都表现不错的梯度提升迭代决策树(GradientBoostingDecisionTree,GBDT) 模型进行训练，GBDT采用Boosting思想每次分类都会迭代产生更强的分类器，同时相对于其他模型，GBDT模型还具有比较好的记忆能力和特征筛选能力，在众多数据竞赛中都有不错的表现，在本次CTR预测中经过测试效果显著，所以决定采用GBDT模型。

GBDT模型目前最经典的应用案例为美国Facebook公司于2014年提出的GBDT加上逻辑回归形成的用户点击率预测模型。这属于是一个混合模型的案例，由于LR不能很好的应对非线性的数据，所以以其中的GBDT作为自动筛选特征的模型对样本数据的训练集提取特征作为新的训练数据输入到下一模型，逻辑回归模型作为新训练输入数据的分类器。GBDT模型特征选择的原则是优先选择整体上有区分度的特征，再选用针对部分样本有区分度的特征。逻辑回归是广义线性模型， 逻辑回归将函数值映射到0至1区间，映射后的函数就是CTR的预估值。同时逻辑回归模型十分适合并行化，因此对于大数据的训练有效。但是线性模型的学习能力是有限的，因此需要提前做好大量的特征工程预先分析出有效的特征或者是特征组合，从而去间接的增强逻辑回归的非线性学习能力。

### 4.1.4 LightGBM

GBDT模型目前主要的工程实现方式包括极端梯度提升(eXtreme Gradient Boosting,XGBoost)[5]和LightGBM(Light Gradient Boosting Machine)，其中LightGBM是一个由微软亚洲研究院分布式机器学习工具包（DMTK）团队开源的梯度提升框架，是一个分布式，高效的基于决策树的算法。

LightGBM不仅仅可以处理大规模数据，甚至还有更快的训练速度和效率，对于内存的使用也比较小，最吸引人的是最终模型的准确率也是有保障，同时还支持并行化的学习。相对于XGBoost，LightGBM模型训练速度接近于XGBoost的25倍。

lightGBM主要是进行了重心位于模型训练速度的优化。首先，lightGBM采用单边梯度采样算法(Grandient-based One-Side Sampling，GOSS)，LightGBM使用GOSS算法对训练样本采样过程进行优化。在GBDT的基本算法中是没有关于样本的权重，所以LightGBM采用了基于每一个样本的梯度进行训练样本的优化，数据拥有较大梯度的时候对计算信息增益的贡献比较大。当一个样本点的梯度很比较小的时候，就说明该样本的训练误差其实很小，即该样本已经被充分训练。而如果单纯在计算过程中，仅仅保留梯度较大的样本，抛弃梯度较小样本，这样的做法会改变样本的分布并且降低学习的精度。GOSS算法的提出就是针对这个问题，GOSS算法的基本思想是根据梯度首先对训练集数据排序，并且预设一个恰当的比例，保留在所有样本中梯度高于的数据样本；梯度低于该比例的数据样本也不会被直接丢弃，而是设置一个采样比例，从梯度较小的样本中按比例抽取样本。同时为了弥补对样本分布造成的影响，GOSS算法在计算信息增益时，会对较小梯度的数据集乘以一个系数，用来放大。这样，在计算信息增益时，算法就会去更加注意还未被充分训练的样本数据。

除此之外，lightGBN还有一个很重要的算法Exclusive Feature Bundling算法(EFB)。lightGBM不仅优化了采集样本的过程，还进行特征抽取，特征抽取与特征提取不同，特征抽取并不减少训练时数据特征向量的维度，而是通过将互斥特征通过一定的算法绑定在一起，从而减少特征维度。数据中互斥特征被绑定在一起后，就会形成低维的特征集合，这使得模型能够有效的避免那些对0值特征的不必要计算。

实际上，在算法中，这就可以有效的降低创建直方图的时间复杂度，从而快速过渡到下一个优化算法直方图算法（Histogram算法),相对于极端梯度提升（eXtreme Gradient Boosting，XGBoost）采用的需要进行提前排序的exact算法，Histogram算法能减少内存消耗，还能做减差加速作用。lightGBM中一个子节点的直方图可以通过父节点的直方图减去兄弟节点的直方图即可得到，从而实现加速，并且实际的数据集上表明，离散化的分裂点对最终学习的精度影响并不大，甚至会更好一些。因为这里的决策树本身就是弱学习器，采用Histogram离散化特征值反而会起到正则化的效果，从而提高算法的泛化能力；lightGBM采用的分支策略为按叶子生长（leaf-wise）的策略，相对于大多数按层生长（level-wise)的策略，这种策略更加高效，该策略每次从当前决策树所有的叶子节点中，找到分裂增益最大的一个叶子节点，然后分裂，如此循环往复。这样的机制下，减少了对增益较低的叶子节点的分裂计算，减少不必要的开销。与level-wise的策略相比，在分裂次数相同的情况下，leaf-wise对误差处理效果更好，能有效降低误差得到更好的精度。Leaf-wise算法有一个缺点就是可能会生成较深的决策树，因此，LightGBM在Leaf-wise上增加了限制最大深度的参数，既能保证算法高效的同时，还能防止过拟合。

本次研究过程中在通过对于用户点击率预测的准确性和高效性的综合考虑之后，决定采用LightGBM的方法实现使用GBDT模型对广告用户点击率的预测。

## 4.2．建模及优化

为了尽量减少内存的使用，提高运行的效率，首先对数据类型进行调整，降低存储内存，例如将64位或是32位的数据降低位16位，即没有改变原有数据的内容，还有利于训练模型。

各种交叉统计的特征占据了太多的内存，并且特征的维度也因此在不断的增加，给训练模型增加了困难，并且还会出现设备内存不足造成各种卡慢甚至死机的情况，为了改善此类情况，采用Stacking的融合模型获取新的特征，并且减少了内存与维度的使用，Stacking的方式在减少特征维度的同时能够实现最大限度的保留所有特征的区分度信息，减少内存的同时可以保留完整特征的信息。如图（Figure 4.1）所示，所有数据集被分为训练集和测试集，以Model 1为基模型在训练集训练出预测结果，通过基模型产生五个预测结果会产生新的特征，在五个基模型中使用的训练集通过平均形成新的测试集。在基模型训练的基础上，再利用Model 2的次级学习器进行训练。Stacking的利用有效的提升了机器学习模型的效果。

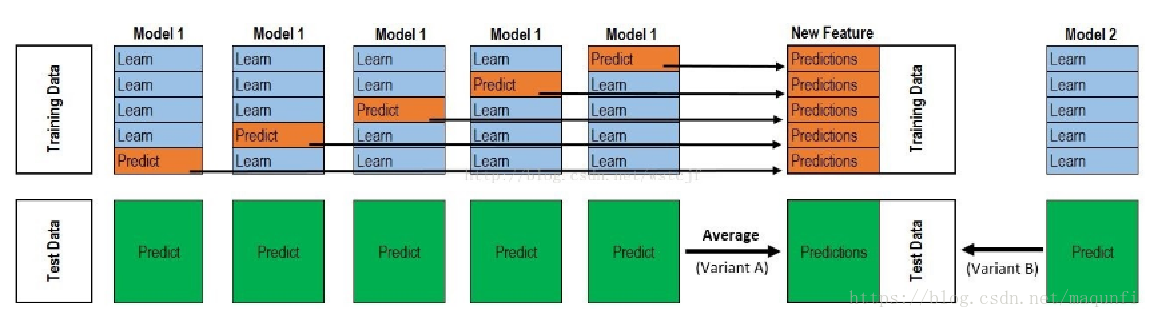


Figure 4.1

Stacking的方式虽然能够使得机器学习模型的训练效果大幅提升，但是在实现的过程中容易发生数据泄露，数据泄露即将样本数据既作为训练集又作为测试集，数据泄露的结果就是最终的预测结果发生过拟合，为解决这一问题，在对模型进一步调优方面使用K折交叉检验，通过自己对十折交叉验证和五折交叉验证两个取值的分别检验，五折交叉验证有着更好的表现，以五折交叉检验得到了五个基模型及对应的预测结果，预测的结果形成新的特征，测试集通过平均形成新的测试集，再以次级模型进行训练得到最终的预测结果，最终再将结果加权平均或者调和平均即可完成。

参数调优的方法使用网格搜索，即通过使用GridSearchCV进行高效调整参数，根据所使用的模型自动进行交叉验证，通过调节每一个参数来跟踪评分结果，该过程代替了进行参数搜索时的循环过程。

## 4.3．预测结果

预测结果使用对数损失函数(Logarithmic Loss Function,Logloss)[4]来进行评价，对于Logloss来说，损失函数值越小，模型就是越优秀。在追求函数值越来越低的过程中需要注意的一个问题就是过拟合效应。由于监督学习目标就是追求经验风险与结构风险最小化，但是经验风险最小化在样本容量足够大的时候能够发挥很好的作用，当样本容量不足时，经验风险最小化就是过拟合的原因，所以增加结构风险以应对过拟合，只有在实现经验风险与结构风险最优化的时候才能实现模型最优，即对数损失函数值越低。

本次使用的三种模型，首先通过未进行参数优化的基础模型以初赛数据进行训练，其中以逻辑回归为模型进行训练的结果得到0.3333，以随机森林模型得到的训练结果是0.2222，以LightGBM实现梯度提升迭代决策树的基础模型得到的结果为0.1111，所以选第三种方案进行进一步的优化提升，最终以整个数据集进行训练，本论文最终预测的结果为0.51515。

# 第五章 结论与总结

## 5.1. 研究结论

本次课题使用了三种模型，分别是随机森林，逻辑回归和梯度提升迭代决策树。前两种模型在未调优之前，效果相似，梯度提升迭代决策树模型在使用LightGBM实现的方案明显优于前两者，所以最终选择梯度提升迭代决策树模型作为本次研究所使用的最终模型。

LightGBM在样本数据集的预测结果为损失函数值0.41，属于符合预期的水平，具有用户点击率预测的实际意义。目前最前沿的模型能够做到优于本次使用 的模型，可以将结果提升至0.40以下，例如阿里巴巴提出的DSTN(Deep Spatio-Temporal Neural Networks)模型在论文中有数据显示预测效果大幅度超过以往的模型，损失函数值达到了0.3993.

Stacking的使用能够有效的降低数据的维度，降低数据占用的内存，最大限度的保留原始信息的完整度，同时通过模型的融合，并不需要过多的调参和特征选择，还可以通过增加正则项对抗LightGBM最容易发生的过拟合现象，高效的提升预测模型的效果，Stacking的数据泄露问题可以通过K折交叉检验的方式解决。

## 5.2. 思考与总结

本次研究数据中用户id信息了解不足。用户点击率预测中的主体依旧是用户，针对于以用户为主导的研究来说，用户id都是一个极其重要的信息，所以都必须进行深层次的挖掘，但是由于本次数据关于用户id这一项内容的展示不够清晰，在后续的数据处理及模型预测都没有太大的实际意义，所以只能舍弃。

本次数据规模不是太大，所以采用GBDT模型就能够比较好的处理，预测的结果也符合预期的设想，最终只是使用了单一的模型。如果数据增加，需要优化模型结构的情况下，可以在实现GBDT时，采取优化策略首先使用LightGBM作为处理数据的基模型，发挥其处理数据较强的优势，快速完成对基本数据的处理，提取出数据中对于模型决策有实际一意义的特征信息，数据处理后在使用XGBoost模型训练，XGBoost模型有数据处理的效率不高的问题，但是训练结果更加优秀，如此的二级模型在更大数据集中能提升用户点击率预测的效率与准确性。

由于是竞赛的数据，并没有说明是何种应用场景，所以只能仅凭数据和少量的数据信息预测，如果有应用场景的数据可以在数据处理过程和特征提取过程中更有针对性，可以对与数据的重要性进行区分，提高模型的预测准确率。

由于数据量较大，内存占用过大，直接进行模型训练容易出现存储不足的问题，故采用Stacking的方式进行特征提取，Stacking减少了维度与内存的使用，并且能够完整保存数据特征的信息。但是使用过程中需要注意数据泄露的问题，数据会造成过拟合效应，K折交叉检验可以解决关于数据泄露的问题。

通过这次研究不仅关于点击率预测相关知识有了进一步的了解，还有很多包括模型的发展，构建以及优化的各方面有了进一步的提升，也会明白有很多还没学习过的知识，对目前互联网前沿的一些技术有了更进一步的了解。这一次的研究效果虽然不错，但是还有一些方面需要进一步的提升：（1）本次的数据量大，但是相对于实际商业环境下的数据规模还是具有明显的差距，所以本次的研究过程相比在实际商业活动中的操作会显得比较简单.（2）本次样本数据略有遗憾，由于信息经过了脱敏处理，其中有一些比较重要的用户信息未有足够的信息说明，所以不能够充分的使用全部的信息。（3）特征选择的方法还有很多没有使用的方法，在以后进一步的研究中，应该尝试更多方法，实现特征信息在模型中的效果更优

# 致谢

2020年是一个特殊的年份，不仅仅是因为这一年我将离开上海大学，这一年一开始就爆发了如同SARS一般席卷全国，甚至全球的新型冠状肺炎疫情。刚开始我和大多数人都是一样的心态，这应该不会太严重吧，或许也就是小范围的一场传染病，现实马上狠狠的打了我一巴掌，看着日益剧增的感染人数，我知道这将是一次严重的卫生安全突发事件，这将是对我们国家，对全国人民的一次考验，这是新时代的战争。毫无疑问，无论是社会经济，还是人民生活都受到了巨大影响，作为这一年的毕业生，我们注定要度过一个难忘的毕业季。

从一开始，每天睁开眼的第一件事就是打开手机，点到疫情地图，心中默念不要再增加了，但是事与愿违的结果就是每天不仅在增加，而且增加幅度还比前一天更加的剧烈。没过多久开始有了死亡病例。北野武有一句话：“灾难并不是死了两万人这样一件事，而是死了一个人这件事，发生了两万次”。对一件事不同的表达更加真实的反映了疫情的影响，当某一个夜晚，微博中浏览到一个视频，还处在疑似阶段的老母亲就被病魔夺去了生命，具有传染性的尸体自然需要尽快的处理，殡仪馆的车到了，带着母亲的尸体开始缓缓向前，这时的女儿还不能接受母亲的离开，用尽痛哭之后仅剩的一点点力气喊“妈，妈……”，那一刻我真的受不了了，看着渐渐远去的灵车，留下的是孤孤单单的一个刚刚失去母亲的孩子。她悲伤，无助，这一刻她就如同死去的几千人后面的那些家庭的一个缩影。

世界上能打倒自己的从来就只有自己，当疫情来临，有的人退却了，有的人冲了上去；有的人借此发财，有的人捐出所有；有的人找到借口虚度时光，有的人却毫无怨言寻找生活的意义。捌号仓库，这是一家店的名字，店主不是什么特别出名的人，只是一个普通市民，每天只睡四个小时，全家人齐上阵为周围的医护人员做盒饭。当有人去采访她，为了什么，只是淡淡的笑笑，说着看着医护人员的朋友圈感觉难受，只想做点什么，不为什么。当问道是不是党员，她说担心自己学历各方面条件不够。每天关注疫情持续严重，心里还是很焦急担心的，但看还有许多如此这般的人，我觉得只要怀揣着理想，就没有什么过不去。如同在这店里的话一般，人总要仰望点什么，想着高远，支撑着生命和灵魂。毕业设计的过程中，也遇到不少觉得很烦躁甚至想赌气放弃的时候就会想起，每个人都有选择生活的权利，选择不一定轻松的方式，或许累完应该会微微一笑，能够深切感受到活出了生命的意义。

冬天的时候疫情来了，疫情如寒风在一点点侵蚀，但是再寒冷的冬也不能打破春的希望。当通过屏幕看到火神山、雷神山一点点建设起来的时候，我们同时也看到了控制疫情的希望。而这次疫情中央视新闻不仅直播了两所医院建设过程给我们，还直播了一个新生儿——小石榴，她的爸爸妈妈都因为疫情的原因隔离了，小石榴一生下来就不能与父母见面，每天有众多的云爸爸云妈妈来看她，可爱的小石榴给大家带来了生机也带来了希望。2020年，中国在奔小康最后一点路上遇到了一个绊脚石，这给大家高涨的热情交了冷水，但是不应该放弃希望，“冬天都到了，春天还会远吗？”。

岁月静好，只因有人负重前行。早早写好请战书，剪了短发，匆匆收好行李这便上了抗疫的战场，和我一般大的90后冲锋在前，这让我感到十分自豪，也让我心中增了一丝紧迫感，一张张青春的脸庞因为长时间戴口罩出现了印记，没有人愿意憋在防护服几个小时不喝水不上厕所，这些种种不愿意他们却义无反顾，逆行而上。还有很多这样的人，头戴国徽的人民警察，为了学生坐做起直播的人民教师，他们都诠释了一种叫责任的信念。他们都是最可爱的人，都有着如雷锋般的精神。我们毕业生是一个很大的群体，也牵动了许许多多的关注，老师们也一直不辞辛劳的为我们服务，在此感谢四年来那么多照顾过我的老师们。

父母是我最亲最敬最爱的人，虽然他们已经不知道我所学习的专业知识，也不能帮助我完成自己的任务，但是他们一直用自己的方式在帮助我爱护我。如果没有这次疫情，我应该早早回到学校，开始完成毕业设计，但是被迫留在家中，但是因为在家爸爸妈妈给了我很大的能量完成大学最后的任务，他们给我缓解焦虑，他们让我坚持到了最后，不善言语的他们却对我做了最长情的告白。

这次的疫情可怕，随时会带走身边的亲人朋友，影响我们正常的工作学习，完全打乱了大家希望能够放松娱乐的假期生活。但是任何事物都是具有两面性，疫情使我们亲人阴阳两隔，也给医护警察增加了工作，但是也会让大家发现人世间更多的闪光点。平凡人也会因为自己为信仰所做的点滴努力令人敬佩，小小的一点希望也可以给大家以巨大的精神动力，肩上的责任成就了人间天使与英雄，隔离病毒隔离不了爱，反而让我们看得清爱是多么细致入微。感谢此次战疫中负重前行的所有人，因为他们我所收获的也比看到的多。

四年一晃而过，再次感谢老师同学朋友家人。一辈子很短，如白驹过隙，转瞬即逝；一辈子很长，如高山大川，连绵不绝，因为遇到的人值得珍惜，经历的事历久弥新。

# 参考文献

1. 刘鹏，王超.计算广告[M].第二版.北京:人民邮电出版社,2019.4-5.
2. Ouyang W , Zhang X , Li L , et al. Deep Spatio-Temporal Neural Networks for Click-Through Rate Prediction[J]. 2019.
3. DavidCHoaglin,FrederickMosteller,JohnWTuke.《探索性数据分析》[J].数理统计与管理,1999(06):53.
4. 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012:77-79.
5. Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.
6. Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.
7. LightGBM
8. Stacking
9. 胡局新, 张功杰. 基于K折交叉验证的选择性集成分类算法[J]. 科技通报, 2013(12):123-125.
10. 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012:6-13.

# 附录