Operations Research and Fuzziology 运筹与模糊学, 2023, 13(3), 1735-1750
Published Online June 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.133174

Hans汉斯

基于Logistic回归的广告点击行为的影响因素分析

杨　旭

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2023年4月7日；录用日期：2023年6月9日；发布日期：2023年6月16日

摘　要

实时竞拍是现代互联网广告行业中非常流行的一种广告投放模式。本研究的主要目的是探索哪些因素是预测广告是否被点击的重要因素。为此，本研究考虑了以下7个指标：平台编码、竞拍低价、是否为全插屏广告、手机运营商、网络状况、设备制造商和广告展现时段。然后建立Logistic回归模型，通过AIC方法做出模型选择。结果显示：1) 竞拍底价越高，广告越不容易被点击。2) 设备制造商：相较于三星、小米、VIVO等设备制造商，苹果点击率较高。3) 全插屏广告，点击率较高。4) 广告展现时段：相较于上午，下午和晚上的点击率较高。

关键词

Logistic回归，实时竞拍，影响因素，全模型，AIC准则，BIC准则

Analysis of the Influencing Factors of Advertising Click Behavior Based on Logistic Regression Method

Xu Yang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Apr. 7th, 2023; accepted: Jun. 9th, 2023; published: Jun. 16th, 2023

Abstract

Real-time bidding is a very popular model of ad delivery in the modern Internet advertising industry. The main purpose of this study is to explore which factors are important in predicting

文章引用: 杨旭. 基于 Logistic 回归的广告点击行为的影响因素分析[J]. 运筹与模糊学, 2023, 13(3): 1735-1750.
DOI: 10.12677/orf.2023.133174

whether an ad is clicked or not. For this purpose, the following seven metrics were considered in this study: platform code, low bid price, whether it is a full insertion ad, cell phone operator, network condition, device manufacturer and ad presentation time slot. Then a logistic regression model is established and model selection is made by AIC method. The results show that: 1) the higher the bidding reserve price is, the less likely the ad is clicked. 2) Device manufacturers: Apple has a higher click rate compared to Samsung, Xiaomi, VIVO and other device manufacturers. 3) Full insertion ads with higher click-through rate. 4) Advertising display time: Compared to morning, afternoon and evening have higher click-through rates.

Keywords

Logistic Regression, Advertising Click Behavior, Influencing Factors, Full Model, AIC Guideline, BIC Guideline

1. 引言

随着互联网行业的不断发展，传统的线下广告行为模式逐渐过渡为在线广告形式，众多研究者对该类问题进行了研究。苗新等(2022)提出一种基于随机森林的在线广告点击购买预测模型，通过调整模型的相关参数进行优化，对在线广告点击购买行为进行精准预测，有效减少了广告方的营销成本，提高了用户的体验感[1]。万君等(2014)选取网络搜索用户的点击意愿为研究对象，提出了网络搜索用户对竞价广告点击意愿的影响因素模型假设，并结合结构方程模型思想进行了实证检验[2]。肖小玲等(2023)针对广告数据中海量数据信息特征处理特点，提出一种采用互信息特征选择后，基于 stacking 融合方式将 XGBoost 算法和 LightGBM 算法融合，构造 XGB_LGBM 预测模型，提升了广告点击率的预测效果[3]。李春红等(2016)在克服处理数据高维性和稀疏性方面不足的基础上，构建了一种基于 LASSO 变量选择方法的广告点击率预测模型，用数据进行模型验证，结果表明影响广告点击率的关键因素是广告关键词中的商标信息、地域信息和点击成本[4]。随着信息时代的发展和市场环境的变化，人们对准确的广告投放技术的需求与日俱增。实时竞拍(real time bidding, RTB)应运而生。Xie 等(2020)阐述了实时竞价模式与大数据的关系，以及具体的操作模式，对影响点击率的各种因素做了合理的分析，并提出了适当的改进建议和有效的广告策略[5]。Qin 等(2019)比较了佣金模式和两阶段转售两种模式下 DSP 的收入，然后在给定其他 DSP 的收入模式的条件下，探讨了一个 DSP 的最佳收入模式[6]。毛衡等(2016)改进了雅虎实验室对数据先分类再存储，然后使用梯度提升决策树和有限混合模型学习得到胜出竞价的分布模型框架，提出了修正的算法，得到了胜出竞价的概率密度函数[7]。

实时竞拍是现代互联网广告行业中非常流行的一种流量交易方式，主要发生于用户打开流量载体(APP 或网站)时，该广告位要展示的广告还未确定。此时，后台会发起一次广告竞价，当竞价结束后，竞价成功的广告才会被最终展示在页面上。此过程非常短，通常在 100 ms 内就可以完成，用户几乎没有感知。RTB 与大量购买投放频次不同，实时竞价规避了无效的受众到达，针对有意义的用户进行购买。目前越来越多的行业需要广告的投放，不管是小到个人，中到小型企业，还是大到上市公司，对于广告的需求量都是供不应求的。Logistic 回归适合二分类问题，不需要缩放输入特征。该方法相较于其他机器学习、深度学习类方法，Logistic 回归实施简单、计算量小、存储占用低，非常高效，可以在大数据场景

中使用，且该方法易理解，模型的可解释性非常好，从特征的权重可以看到不同的特征对最后结果的影响。因此，本研究通过分析实时竞拍数据，采用 Logistic 回归探索什么情况下广告更容易被点击，以此为公司增加盈利。

2. Logistic 回归与实时竞拍

2.1. Logistic 回归

Logistic 回归是一个概率型非线性模型，研究二分类观察结果与一些影响因素之间关系的多变量分析方法[8]。本文主要研究广告是否被点击的影响因素，其中自变量为广告基本特征因子指标值$(x_1, x_2, \cdots, x_m)$，因变量为二元变量广告是否被点击，分别为 1 和 0。m 个自变量作用下，广告被点击的概率为：

$$p = \frac{1}{1+\exp\left[-\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right)\right]} \tag{1}$$

其中，p 为广告被点击的概率；β_0 为截距；$\beta_1, \beta_2, \cdots, \beta_m$ 为回归系数。p 的取值范围为[0, 1]，上式两边取对数，可得：

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \tag{2}$$

然后通过极大似然估计便可得到 Logistic 回归模型的参数。

2.2. 实时竞拍

RTB 与大量购买投放频次不同，实时竞价规避了无效的受众到达，针对有意义的用户进行购买。它的核心是需求方平台(DSP)，RTB 对于媒体来说，可以带来更多的广告销量、实现销售过程自动化及减低各项费用的支出。而对于广告商和代理公司来说，最直接的好处就是提高了效果与投资回报率。与传统的互联网广告生态链相比，传统的互联网广告生态链一般最多只有三方，分别是广告主、广告代理商(即广告公司)以及互联网媒体。而在 RTB 广告交易模式中，原有的广告生态链发生了变化，整个生态链包括广告主、DSP、广告交易平台以及互联网媒体四个主体。广告主将自己的广告需求放到 DSP 平台上，互联网媒体将自己的广告流量资源放到广告交易平台，DSP 通过与广告交易平台的技术对接完成竞价购买。

实时竞价的具体流程如下见图 1：

Figure 1. Real-time bidding flow chart
图 1. 实时竞价流程图

步骤 1：用户浏览访问媒体网站。

步骤 2：媒体网站向 ADX 或 SSP 发起广告请求，触发 ADX 或 SSP 发起竞价。

步骤 3：ADX 或 SSP 组织一次竞价，向多家 DSP 发送竞价请求。

步骤 4：DSP 根据用户 ID 向 DMP 查询该用户的相关信息。

步骤 5：DSP 根据自己的出价算法，分析了用户和手头代理的广告主需求的匹配程度后，决定好是否出价、出价多少。

步骤 6：由于用户和 DSP 手头代理的广告主需求十分匹配，大部分 DSP 纷纷选择出价，这些出价的 DSP 会将广告相关物料一起传给 ADX 或 SSP。没有出价的 DSP 坐等下一次合适的广告请求。

步骤 7：ADX 或 SSP 收到了多家 DSP 的出价，它挑选了其中一个出价最高的 DSP 作为本次竞价的赢家，并将竞赢的 DSP 的广告相关物料传给媒体网站。

步骤 8：媒体网站收到广告相关物料后，将广告渲染出来。用户看到广告，本次竞价流程结束。

3. 基于 Logistic 回归的广告是否被点击影响因素分析

3.1. 数据来源

数据来自于《商务数据分析与应用》模拟生成的 4000 条 RTB 数据。本研究的因变量是广告是否被点击的状态，一共有两种可能的状态：点击和不点击。自变量主要考虑了 7 个解释变量，见表 1：

Table 1. Description of data variables
表 1. 数据变量说明

变量类型	变量名	详细说明	取值范围	备注
因变量	是否点击	该条广告是否被点击	0 表示否；1 表示是	
自变量	平台编码	分类变量	3 表示 Inmobi；7 表示 Zplay； 8 表示 Baidu；13 表示 Iflytek	
	竞拍底价	数值型变量	0、3e−04、4e−04、0.00049809、 5e−04 等 27 个数值	
	是否为全插屏广告	分类型变量	0 表示否；1 表示是	
	手机运营商	分类变量	0 表示移动；1 表示联通； 2 表示电信；3 表示 AT&T	
	网络状况	分类变量	0 表示未知；1 表示 WIFI 2 表示 2G；3 表示 3G 4 表示 4G；5 表示 5G	
	设备制造商	分类变量	小米、华为、魅族、金立、 酷派、苹果、三星、OPPO、 VIVO、其他	
	广告展现时段	分类变量	上午、下午、晚上	相对时段

1) 平台编码：本文考虑四个平台编码：Inmobi，Zplay，Baidu 和 Iflytek。Inmobi 是全球最早起家的三大移动广告网络之一，Inmobi 一直专注于做移动广告品牌，帮助广告主将正确的信息在正确的时间推送至当今全世界的移动消费者。Zplay 专注于手机游戏发行和研发领域。Zplay 主要业务是移动终端游戏的代理、发行与自主研发以及游戏广告业务。Baidu 是流量三大巨头之一，国内大多数用户都在使用它的搜索引擎。而 Baidu 信息流则是在 Baidu 首页、Baidu 贴吧等版位中穿插展现在信息流中的原生广告。Baidu 信息流广告投放需要先筛选关键内容，使用符合用户习惯的内容投放。Iflytek 是亚太地区知名的智能语音和人工智能上市企业。讯飞 AI 营销是科大讯飞集团在数字广告领域发展的重要业务，基于深耕多年的人工智能技术和大数据积累，赋予营销智慧创新的大脑，以健全的产品矩阵和全方位的服务，帮助广告主用 AI 技术实现营销效能的全面提升，打造数字营销新生态。

2) 竞拍底价：竞拍底价是指拍卖价格的最低标准，如果竞拍过程中竞拍人所报出的最高价低于底价，则该拍卖不可成交。广告的竞价底价，就是在广告的竞价中给竞拍设定一个最低价。在真实的广告竞争市场中，很多时候广告位都无法得到充分竞争。因为广告位的位置会导致不充分竞争，同一广告位在一天中的不同时段的竞争程度也是不尽相同的。对于媒体平台来说，它们可以通过设置一个广告竞价的最低价格，也就是我们这里所说的竞拍底价来保护自己的广告位价值并且保证最低收益。当设置了底价以后，所有的广告竞价都不会低于这个价格，也就人为地抬高了广告位的竞争水准。竞拍底价会影响呈现给用户的广告数据信息，从而影响用户对该广告的点击状态。

3) 全插屏广告：全插屏广告是移动广告的一种常见形式，具有强烈的视觉冲击效果，是目前移动广告平台主流的广告形式之一，它是一种应用以全屏的形式弹出的广告形式，即在用户做出相应的操作(如开启、暂停、过关、跳转、切换、退出)后，弹出的以图片、动图、视频等为表现形式的全屏广告。当应用展示插屏广告时，用户既可以选择点击该广告，进而访问其目标网站，也可以将其关闭，并返回应用。全插屏广告尺寸大、视觉效果震撼。它是一种典型的模态窗口，在特定场景下由用户触发，也需要用户主动关闭后消失。在插屏广告展示时，用户需要点击或关闭才能进行下一步行动，常常引发用户误触。因此，一般来说，全插屏广告拥有较高的点击率，广告效果较好。

4) 手机运营商：运营商是指提供网络服务的供应商，如华为、中兴、诺基亚等这些通信设备的生产厂家叫生产商，而中国移动、中国联通、中国电信、中国广电这些公司叫运营商。国内四大运营商指的是中国移动、中国联通、中国电信，中国广电，不管是从通讯或者是地面网络角度去看这四家都是规模最大实力最雄厚的，当然他们彼此的侧重点不同，例如：中国移动主要运营 GSM、TD-SCDMA、TD-LTE 网络和固网。中国联通主要运营 GSM、WCDMA、FDD-LTE 和固网。中国电信主要运营 CDMA、CDMA2000、FDD-LTE、TD-LTE 网络和固网。本报考考虑国内三家运营商：中国移动、中国联通和中国电信；国外一家运营商：AT&T。美国电话电报公司(AT&T)是一家美国电信公司，创建于 1877 年，曾长期垄断美国长途和本地电话市场。AT&T 在近 20 年中，曾经过多次分拆和重组。目前，AT&T 是美国最大的本地和长途电话公司。

5) 网络状况：网络状态可以时时监护你的流量走向、网络速度，给用户准确的实时网络状态。在进行严格等值变换的条件下，实现了不同用户间由某一公用输电路径为其提供服务向各自由独立的输电路径为其提供服务的完全等值的解耦过程。本文考虑 6 种网络状况：未知，WIFI，2G，3G，4G，5G。WIFI，一般是家庭宽带，然后用无线路由器分享的无线网络。2G，3G，4G，5G 上网，则是流量上网，这个是需要按照标准收费的。一般现在如果没开通套餐是 1 元/m。从网速来看：WIFI 的网速，主要取决家里的宽带速度的好坏。另外，移动设备和无线路由器的好坏也会影响到 WIFI 的上网速度。而 2G，3G，4G，5G 需要根据信号的好坏来决定。从覆盖范围来看：一般在无线路由器很小的范围内才能有 WIFI 信号。

而 2G、3G、4G、5G，只有移动信号就能使用。

6) 设备制造商：设备制造商是指负责制造向用户提供服务的移动通信系统设备和终端，是利用自己掌握的关键的核心技术负责设计和开发新产品，控制销售渠道，具体的加工任务通过合同订购的方式委托同类产品的其他厂家生产，之后将所订产品低价买断，并直接贴上自己的品牌商标。市面上有各种各样的设备制造商，他们能够不断的产出十分畅销的产品。本文考虑了全球设备制造商：小米、华为、魅族、金立、酷派、苹果、三星、OPPO、VIVO 和其他设备制造商，不同的设备制造商展示的广告内容可能不一样，可能影响用户的点击状态。

7) 广告展现时段：在电波媒介的时间安排中，集中播放广告的一段时间，叫广告时段。广告时段一般放在两个节目之间；所占用的时间不长，一般为 5 分钟左右；在这个时段中，各种类型的广告一起播放，广告与节目内容没有必然联系。不同广告时段的广告费是不同的，通常黄金时段是广告收费最高的时段。本文考虑了三个展现时段，分别是：上午、下午和晚上。

3.2. 描述性分析

本研究涉及一个被解释变量，模拟生成的 4000 条样本量中，3245 个数据因变量取值为 0，占比 81.13%；755 个数据因变量取值为 1，占比 18.87%，见图 2 所示：

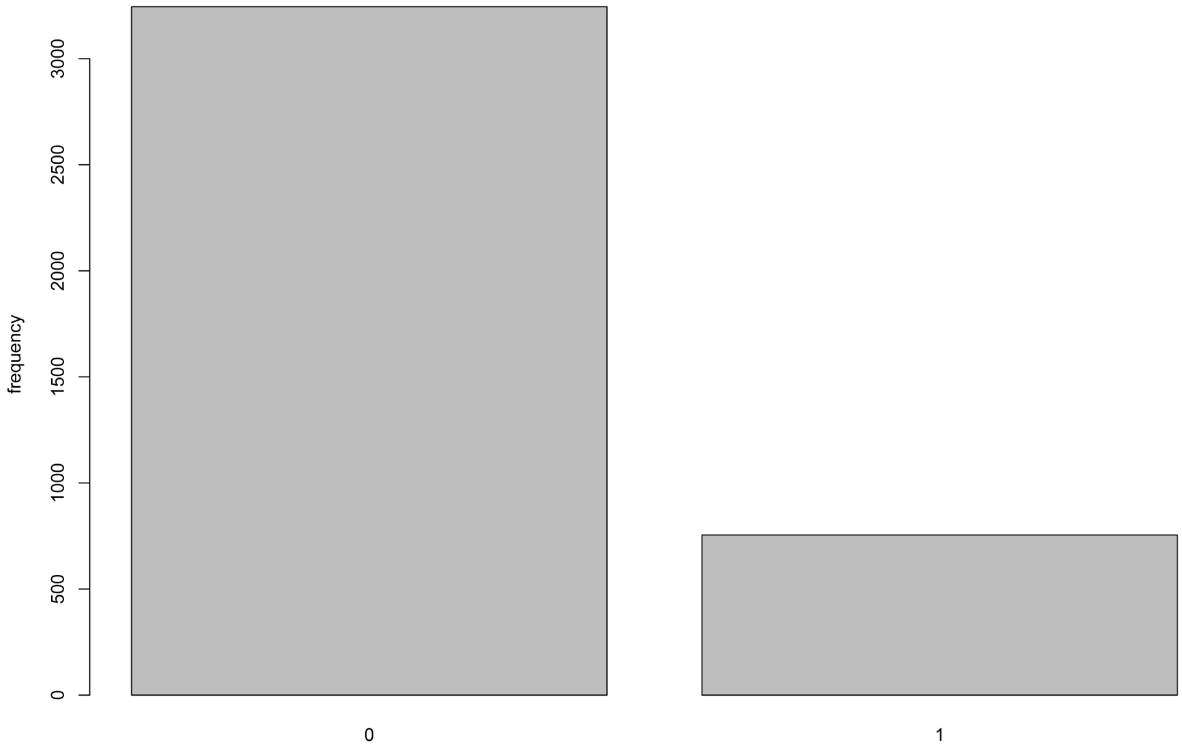

Figure 2. Histogram of whether the ad was clicked
图 2. 广告是否被点击直方图

本研究涉及 7 个解释变量：平台编码、竞拍底价、是否为全插屏广告、手机运营商、网络状况、设备制造商和广告展现时段，接下来本文进一步对解释变量进行分析，为后续建模做铺垫。

竞拍底价取 27 个不同取值，见图 3，可以发现 3460 个样本竞拍底价为 0，占比为 86.5%；540 个样本竞拍底价大于 0，占比为 13.5%。

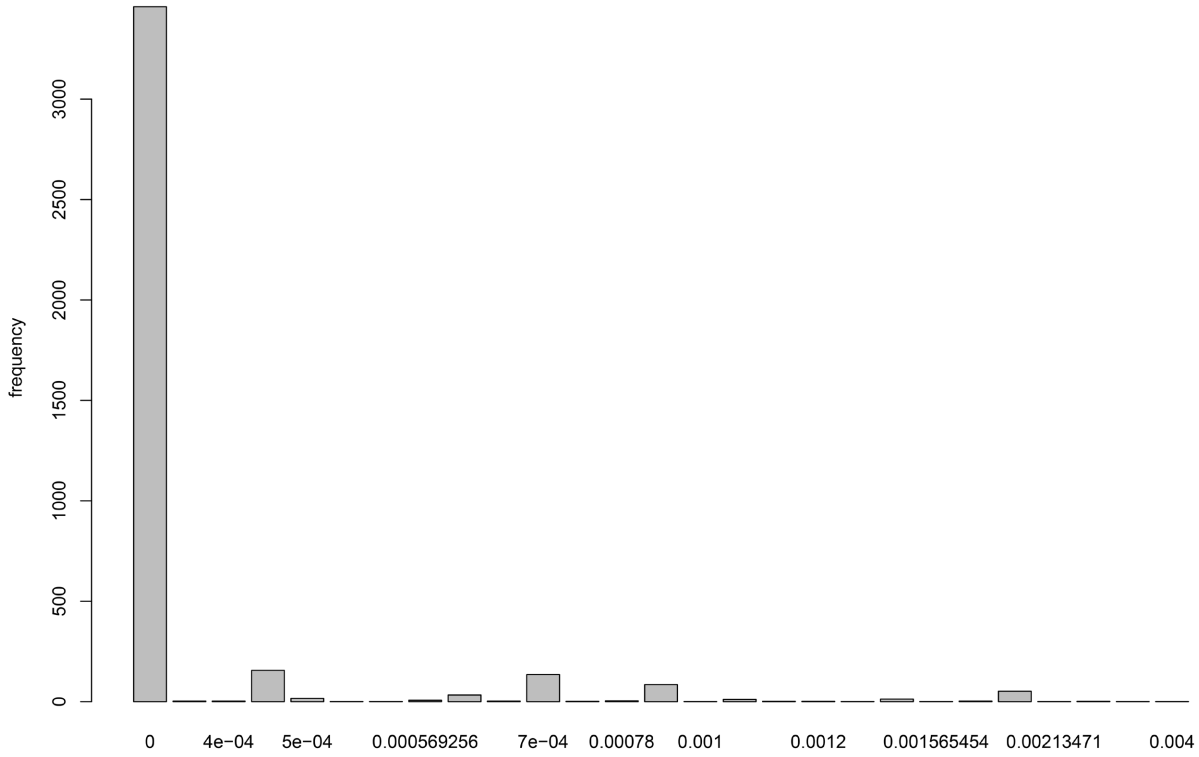

Figure 3. Histogram of reserve bids
图 3. 竞拍底价直方图

本研究考虑了全球设备制造商：小米、华为、魅族、金立、酷派、苹果、三星、OPPO、VIVO 和其他设备制造商，其中 4000 个数据中，各个设备制造商占比数据见表 2 和图 4：

Table 2. Equipment manufacturers
表 2. 设备制造商

设备制造商	个数	占比
OPPO	696	17.4%
VIVO	492	11.8%
华为	592	14.8%
金立	97	2.43%
酷派	121	3.03%
魅族	194	4.85%
苹果	136	3.4%
三星	373	9.33%
小米	597	14.93%
其他	702	17.55%

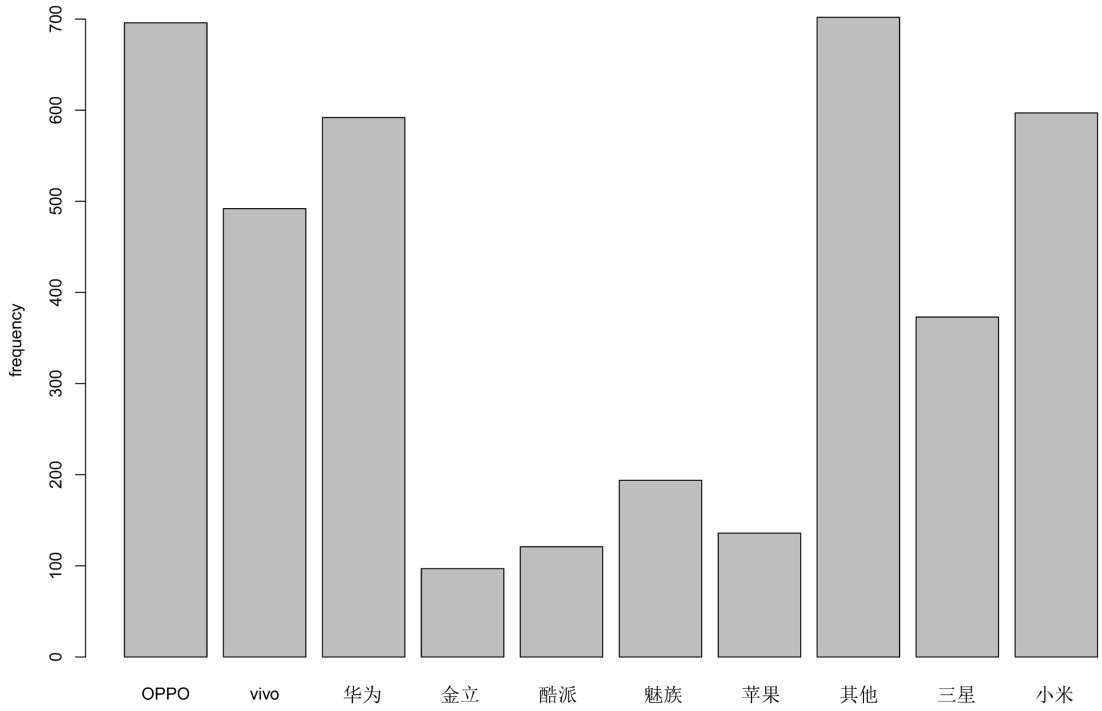

Figure 4. Histogram of equipment manufacturers
图 4. 设备制造商直方图

本文选取 5 个水平的网络状况：0 表示未知，1 表示 WIFI，2 表示 2G，3 表示 3G，4 表示 4G，5 表示 5G。样本量中，该解释变量网络状况水平用户使用 WIFI 较多，频数为 3164，占比 79.1%，见图 5。

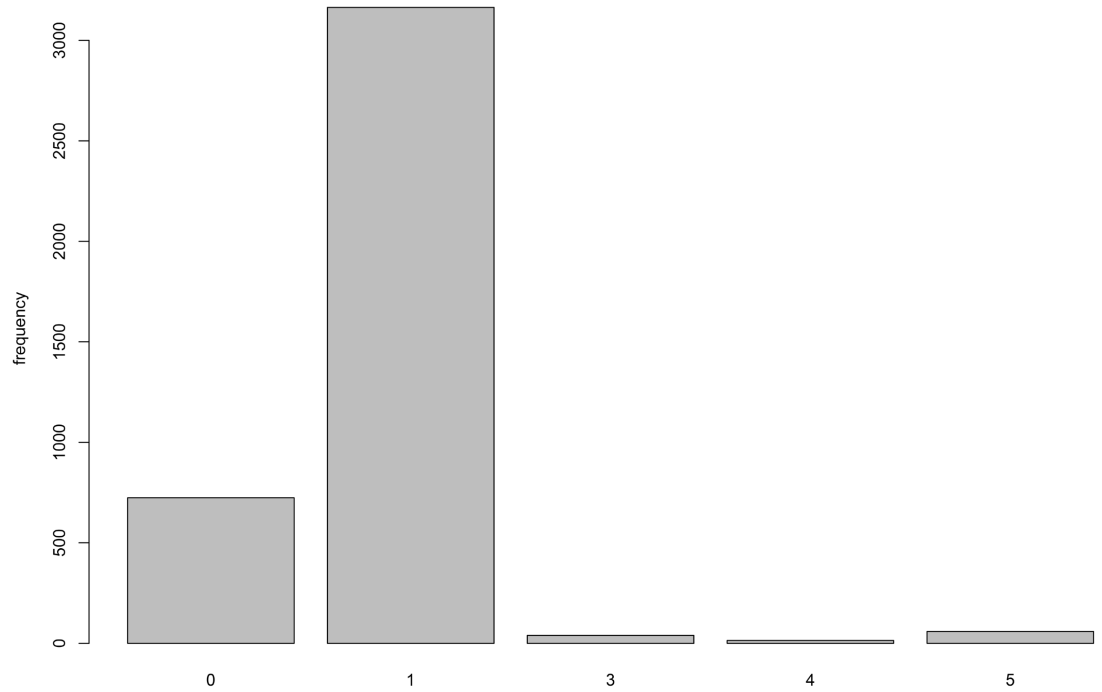

Figure 5. Histogram of network status
图 5. 网络状况直方图

其他 4 个变量直方图见图 6，可以发现解释变量平台编码中，用户使用 Baidu 平台较多，频数为 3460，占比 86.5%；全插屏广告较少，仅 197 个样本中为全插屏广告，占比 4.93%；手机运营商中，联通较多，其频数为 2530，占比为 63.25%；广告展现时段中，上午、下午和晚上三个时段比较均衡，相差不大，占比分别为 30.05%、36%和 3.95%。

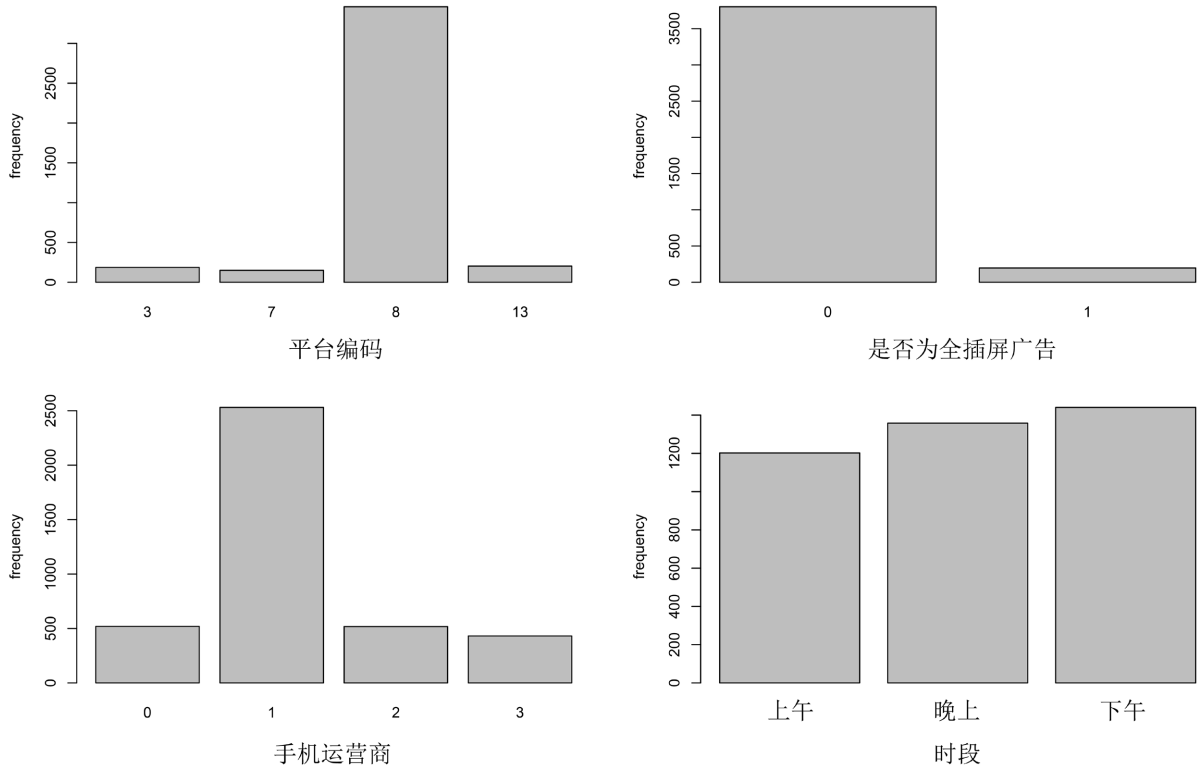

Figure 6. Histograms of the other 4 variables
图 6. 其他 4 个变量直方图

接下来，对广告是否被点击和各个解释变量的相关关系做简单描述。由于因变量广告是否被点击是一个 0~1 变量，可以绘制堆积条形图观察各个解释变量和被解释变量的关系，分别见图 7~13：

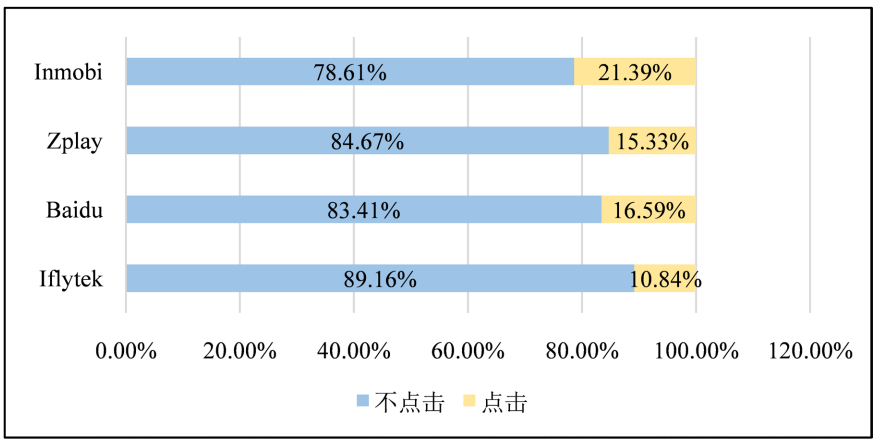

Figure 7. Stacked bar chart of platform code and dependent variable
图 7. 平台编码与因变量的堆积条形图

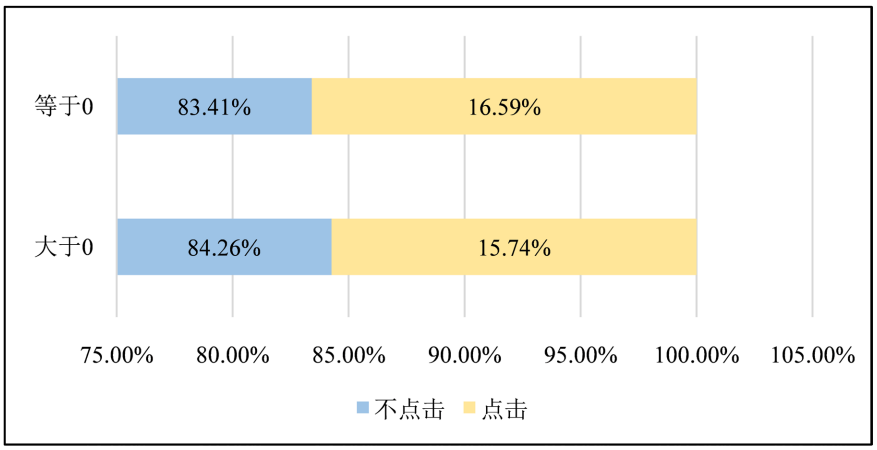

Figure 8. Stacked bar chart of bidding reserve price and dependent variable
图 8. 竞拍底价与因变量的堆积条形图

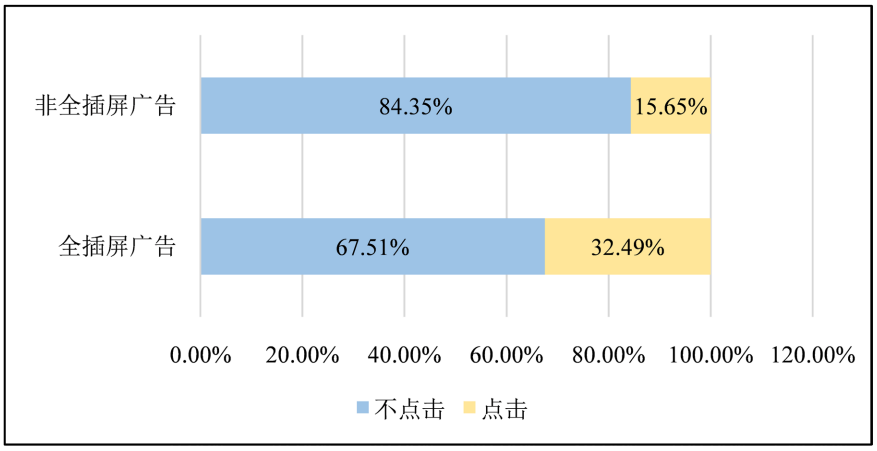

Figure 9. Whether or not it is a stacked bar chart of full inset ads with dependent variables
图 9. 是否为全插屏广告与因变量的堆积条形图

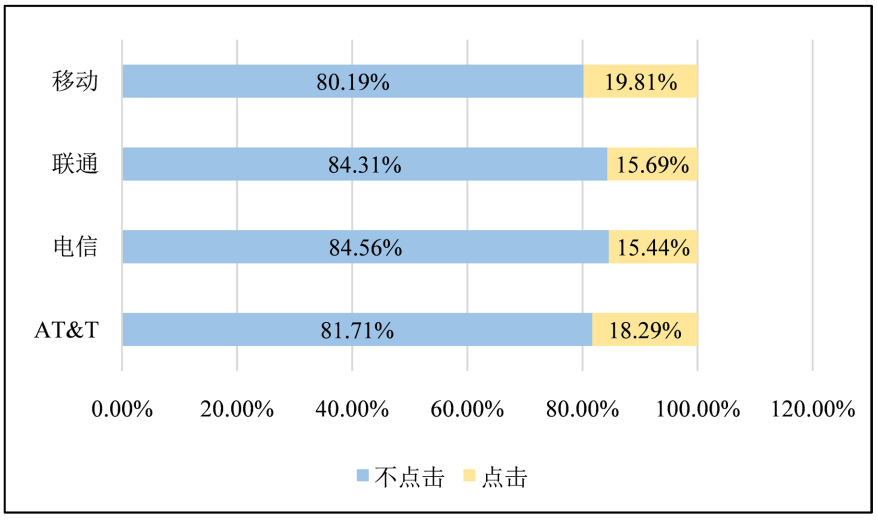

Figure 10. Stacked bar chart of cell phone operators and dependent variables
图 10. 手机运营商与因变量的堆积条形图

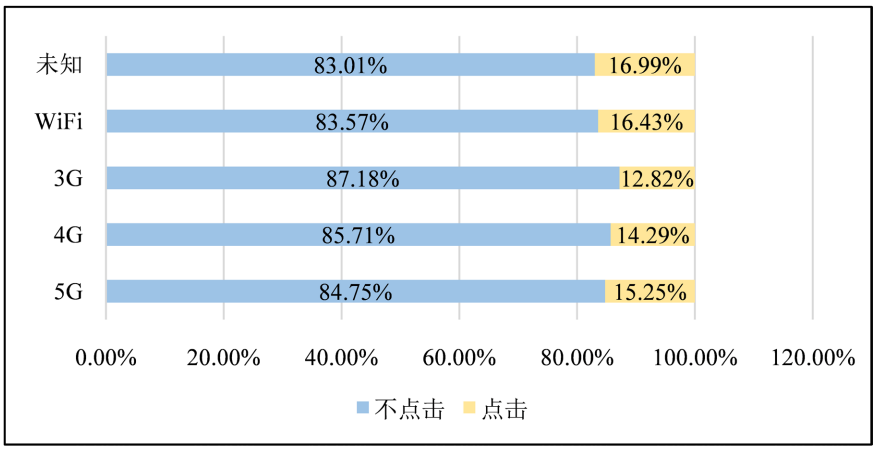

Figure 11. Stacked bar chart of network conditions and dependent variables
图 11. 网络状况与因变量的堆积条形图

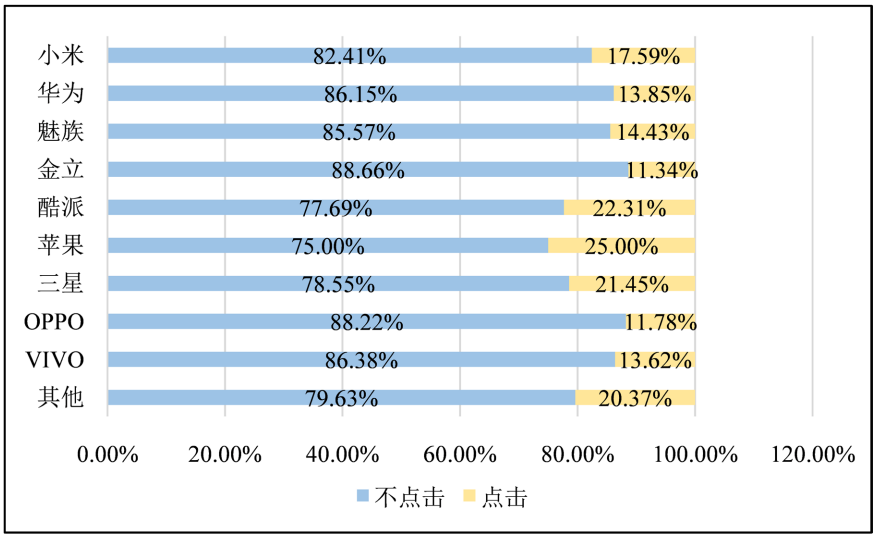

Figure 12. Stacked bar chart of equipment manufacturers and dependent variables
图 12. 设备制造商与因变量的堆积条形图

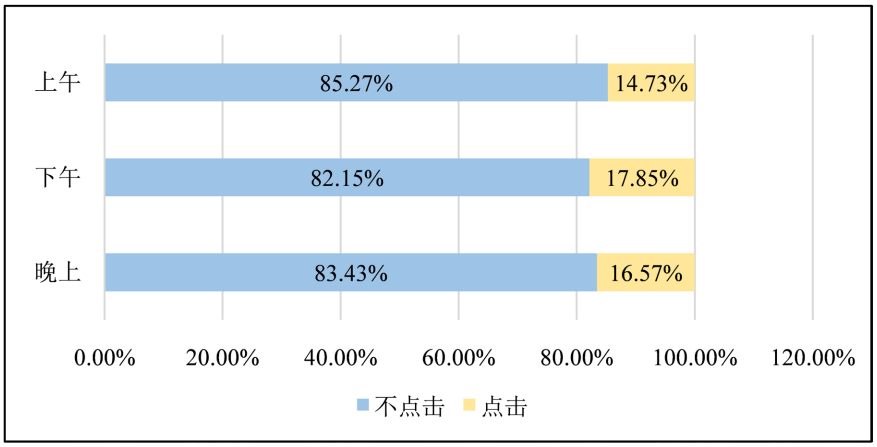

Figure 13. Stacked bar graphs of ad display periods and dependent variables
图 13. 广告展现时段与因变量的堆积条形图

从以上 7 个堆积条形图可以看出：各个解释变量似乎都与广告是否被点击有一定的关系。比如：平台编号为 Inmobi 时，点击广告的可能性更高；竞拍底价为 0 时，广告被点击的概率更高；全插屏广告的点击率更高；手机运营商为移动的用户点击广告的可能性更高；网络状况为 WIFI 时，广告被点击的概率更高；手机设备商为苹果时，广告被点击的概率更高；广告下午被点击的可能性更高。

3.3. 模型建立

为了深入挖掘影响广告是否被点击的显著因素，接下来将建立 0~1 回归模型。首先，对所有变量建立 0~1 回归全模型，结果见表 3 所示：

Table 3. 0~1 regression full model results
表 3. 0~1 回归全模型结果

变量名称	回归系数	标准误	P-值
截距项	−1.894	0.274	<0.001
平台编码	−0.031	0.028	0.270
竞拍底价	−236.7	159.7	0.138
是否为全插屏广告	1.000	0.157	<0.001
手机运营商	0.076	0.052	0.145
网络状况	−0.002	0.064	0.972
设备制造商 VIVO	0.167	0.167	0.317
设备制造商华为	0.182	0.159	0.253
设备制造商金立	0.005	0.313	0.987
设备制造商酷派	0.787	0.238	0.001
设备制造商魅族	0.339	0.220	0.124
设备制造商苹果	1.161	0.289	<0.001
设备制造商其他	0.561	0.145	<0.001
设备制造商三星	0.715	0.165	<0.001
设备制造商小米	0.487	0.152	0.001
广告展现时段晚上	0.183	0.106	0.084
广告展现时段下午	0.279	0.103	0.007
模型全局检验		P-值 < 0.001	

从表中可以看到，全模型的似然比检验高度显著(P-值 < 0.001)，这表示在考虑的所有因素中，至少有一个是对广告是否被点击有显著影响的。在 5%的显著水平下，是否为全插屏广告，设备制造商苹果，设备制造商酷派，设备制造商其他，设备制造商三星，设备制造商小米，广告展现时段下午是显著的，

且这些变量回归系数项为正值，这表示全插屏广告更容易被点击，插屏广告展现尺寸更大，视觉效果更好，可以让用户看到更多的信息，更容易被用户点击。通过对模型的分析，可以发现设备品牌对广告是否被点击有影响，三星、苹果、小米等设备上的广告更容易被点击，其点击率更高。广告展现时段下午广告更容易被点击，上午一般是人们忙于工作的时间，打开手机 APP 点击广告的概率较低，而在下午和晚上，人们使用手机 APP 休闲娱乐的时间可能较长，相应地，广告被点击的概率也较大。

通过对全模型的分析，可以发现全模型还包含很多不显著的因素，因此考虑根据 AIC 准则和 BIC 准则，选择更加简洁的模型，其结果见表 4 和表 5 所示：

Table 4. AIC regression model results
表 4. AIC 回归模型结果

变量名称	回归系数	标准误	P-值
截距项	−2.048	0.129	<0.001
竞拍底价	−290.905	149.048	0.051
是否为全插屏广告	1.004	0.157	<0.001
设备制造商 VIVO	0.169	0.167	0.312
设备制造商华为	0.189	0.159	0.236
设备制造商金立	−0.011	0.312	0.970
设备制造商酷派	0.787	0.238	0.001
设备制造商魅族	0.333	0.220	0.130
设备制造商苹果	1.197	0.268	<0.001
设备制造商其他	0.555	0.144	<0.001
设备制造商三星	0.727	0.165	<0.001
设备制造商小米	0.483	0.152	0.001
广告展现时段晚上	0.181	0.105	0.086
广告展现时段下午	0.282	0.103	0.006
模型全局检验		P-值 < 0.001	

Table 5. BIC regression model results
表 5. BIC 回归模型结果

变量名称	回归系数	标准误	P-值
截距项	−1.524	0.042	<0.001
是否为全插屏广告	1.038	0.153	<0.001
模型全局检验		P-值 < 0.001	

AIC 结果表明：从假设检验的结果看，除设备制造商 VIVO、华为、金立、酷派在 10%的水平下显著以外，AIC 准则选择出的其他的变量都显著。相较而言，BIC 仅选择了一个变量，BIC 准则认为只有是否为全插屏广告一个变量是重要的。

最后，对三个不同的模型(全模型、AIC 模型、BIC 模型)计算内样本的 AUC，以评价他们的预测能力。由于该数据是模拟生成的，通过对比发现，三个模型的 AUC 取值分别为 60.83%，60.66%，53.09%，取值都不是很高，可能是由于模拟生成的数据质量较低。全模型和 AIC 模型的 ROC 曲线非常相似，而 BIC 的 ROC 曲线向左上方凸起的程度相对较差，AUC 较差。AIC 考虑的变量较全模型来说，考虑的变量较少，较简单，因此考虑 AIC 模型。

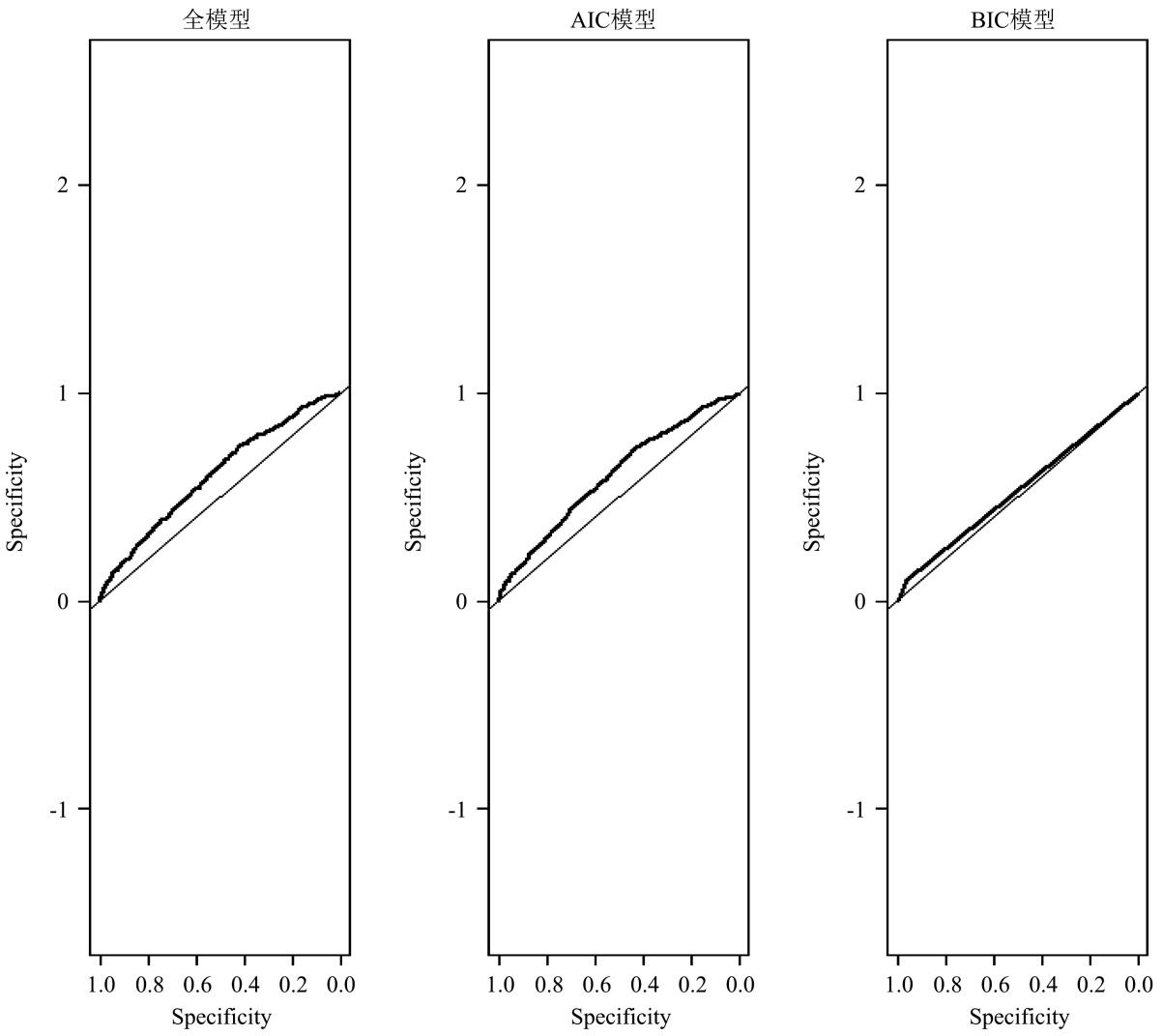

Figure 14. Comparison of ROC curves of three models
图 14. 三个模型的 ROC 曲线比较

见图 14，从 AIC 模型结果可以发现：在 10%的显著性水平下，AIC 模型有 4 个显著变量：竞拍底价、是否为全插屏广告、设备制造商和广告展现时段，设备制造商考虑 5 个水平：苹果、酷派、三星、小米、其他，广告展现时段下午和晚上两个水平。竞拍底价越高，广告越不容易被点击；设备制造商：相较于

三星、小米，苹果点击率较高；全插屏广告，点击率较高；广告展现时段：相较于上午，下午和晚上的点击率较高。

4. 模型应用

本模型可以帮助广告商预判广告被点击的可能性以及相关原因。考虑某条广告平台编码为 Baidu，竞拍底价为 0，全插屏广告，手机运营商为移动，网络状况为 4G，设备制造商为苹果，广告展现时段为下午。综合所有信息后，用该模型进行预测，可以发现该条广告被点击的概率为 65.92%。该广告被点击的可能性较高，可能是因为该条广告的竞拍底价为 0，竞拍底价较低；广告设计为全插屏广告，从广告类型的比较可以看出，非全插屏广告的点击率明显低于全插屏广告；该用户使用苹果设备，可能属于较高的收入群体，更有可能被广告吸引并点击它，或者可能是因为苹果设备上的广告更美观，对用户更具吸引力；广告投放时间段为下午，相较于上午来看，该用户有更多空闲的时间休闲娱乐，点击广告。

5. 结论与展望

本研究基于《商务数据分析与应用》模拟产生的数据，以广告是否被点击为因变量，构造描述广告特征的指标作为自变量，包括平台编码、竞拍底价、是否为全插屏广告、手机运营商、网络状况、设备制造商、广告展现时段 7 个自变量。建立对广告是否被点击的状态具有一定预测能力的 Logistic 回归模型。本文探究了哪些是影响广告点击状态的重要因素，主要结论可以归纳如下：

1) 竞拍底价越高，广告越不容易被点击；

2) 设备制造商：相较于三星、小米等设备制造商，苹果点击率较高；

3) 全插屏广告，点击率较高；

4) 广告展现时段：相较于上午，下午和晚上的点击率较高。

本研究存在一些不足还需要进一步完善，本文选取的自变量中，6 个都是分类变量，1 个数值型变量，考虑的变量类型有限，未来可以考虑加入连续型变量，比如：广告呈现时间秒数等自变量。我们还可以研究分析每周各天点击率，分析各天各时间段广告点击率，分析各年龄层次人群广告点击率，分析各个消费等级人群点击率，使广告在合适的时间被精准投放给感兴趣的用户。因为用户和广告匹配度越高，越不容易引起用户反感，广告被点击的可能性越高。

Logistic 回归方法虽然可以解决二分类问题，但该方法准确率并不是很高，因为形式类似线性模型，非常简单，很难去拟合数据的真实分布，且容易欠拟合，所以回归效果相较于机器学习、深度学习模型可能会不太理想，所以未来可以进一步考虑使用决策树、支持向量机、神经网络等精度较高的模型对该问题进行解决。

参考文献

[1] 苗新, 王倚天, 刘爽. 基于随机森林的在线广告点击购买预测[J]. 信息与电脑(理论版), 2022, 34(12): 54-56.

[2] 万君, 吴迪, 赵宏霞. 网络搜索用户对竞价广告的点击意愿研究[J]. 现代情报, 2014, 34(12): 7-11.

[3] 肖小玲, 李香君, 刘天赐. 一种改进的广告点击率预估模型研究[J/OL]. 长江大学学报(自然科学版): 1-7. https://doi.org/10.16772/j.cnki.1673-1409.20220530.001, 2023-04-06.

[4] 李春红, 吴英, 覃朝勇. 基于 LASSO 变量选择方法的网络广告点击率预测模型研究[J]. 数理统计与管理, 2016, 35(5): 803-809. https://doi.org/10.13860/j.cnki.sltj.20160922-022

[5] Xiang, L.J. and Xiao, X.L. (2020) Visualization Analysis of Real-Time Bidding Data of Online Advertising Based on Hadoop and Python. *IETI Transactions on Engineering Research and Practice*, **4**, 1-9.

[6] Qin, R., Yuan, Y. and Wang, F.Y. (2019) Exploring Optimal Revenue Models for DSPs in Real Time Bidding Advertising. In 2019 *IEEE International Conference on Service Operations and Logistics, and Informatics* (*SOLI*), Hangzhou,

17-19 October 2019, 181-185. https://doi.org/10.1109/SOLI48380.2019.8955015

[7] 毛衡, 胡宁, 陈蔚, 等. 实时广告竞拍平台中的海量数据分析和竞价预测[J]. 应用数学与计算数学学报, 2016, 30(1): 1-15.

[8] 张侠. 基于 SVM 和逻辑回归的糖尿病数据分析与研究[J]. 沧州师范学院学报, 2023, 39(1): 19-23+84. https://doi.org/10.13834/j.cnki.czsfxyxb.2023.01.018