

基于 Logistic 回归在二分类任务定价模型中的应用

宋佳莹, 葛亚平

(南通理工学院基础教学学院, 江苏 南通 226002)

[摘要] 随着大数据时代的发展, 互联网逐渐渗入到各个行业, 众包平台服务模式开启, 这是移动互联网下的一种自助式服务模式, 它将产品和消费者、厂商正确无误地联系在一起, 较之传统的市场调查方式, 更加时效并缩短了调查周期、减少成本和有利于数据的收集, 保证了调查数据的真实性. 本文主要针对众包任务定价方面进行分析, 利用多元统计方面的主成分分析、多元线性回归模型以及多元 Logistic 回归模型对相关任务进行分析、重新定价, 以使得在保证最低成本下, 任务最终能够尽可能被完成.

[关键词] 众包平台服务, 主成分分析, 多元线性回归模型, Logistic 回归模型, 重新定价

[中图分类号] F222.1 **[文献标志码]** A **[文章编号]** 1001-4616(2018)04-0033-06

The Application of Logistic Regression in Binary Task Pricing Model

Song Jiaying, Ge Yaping

(Institute of Basic Education, Nantong Polytechnic College, Nantong 226002, China)

Abstract: With the development of big data technology, the Internet has penetrated into every walk of life. Crowdsourcing platform service mode, a self-service mode with mobile Internet, has come into being. It accurately connects products with consumers and manufactures. Compared with traditional market research method, this mode is more time-effective. It not only decreases vestigation period but also reduces cost, which is conducive to data collection and guarantees the authenticity of collected data. This essay makes data analyses on the pricing of crowdsourcing tasks. Principal component analysis, multiple linear regression model as well as multiple logistic linear regression model are utilized to analyze the relevant tasks and to reprice so as to complete the relevant tasks with minimum cost as far as possible.

Key words: crowdsourcing platform service, principal component analysis, multiple linear regression model, Logistic regression model, repricing

近几年,随着现代市场竞争不断加剧,越来越多的企业开始尝试将具有一定技术性和创新性的工作任务,通过计算机互联网渠道委托给外部组织或个体完成,这种基于互联网的新兴开放式协作创新模式被称为众包. 如快递、外卖、拍照挣钱等行业. 用户通过从 APP 上领取要完成的任务(比如上超市去检查某种商品的上架情况),赚取 APP 对任务所标定的酬金. 而 APP 中任务定价则是其核心要素. 关于产品任务定价模型,在宏观经济学中,价格模型首先应在价格理论的指导下明确价格政策的社会经济目标,例如,整体经济效益最大化(或称资源分配最优化)、满足国民的基本需求等. 从合理分工、相互配合的原则出发,价格政策通常选整体经济效益最大化为目标.^[1]不同的目标引起最优定价的数值差异,也使得评价经济状况与政策效果的尺度有所不同. 在微观经济学中,价格确定主要取决于供求平衡关系,各种价格定价模型主要用来研究在不同的生产、销售和竞争条件下,如何按照厂商经营目标来确定最优价格.

对于众包平台服务的任务定价,如果价格和各项因素未平衡、定价不合理,就会导致会员对定价的不满意,有的任务就会无人问津,最终影响数据收集的时效性,从而导致商品检查的失败. 任务定价的高低很大程度上决定了任务能否被完成,定价是否合理直接影响任务完成情况. 本文主要针对任务定价及任务完成情况的诸多因素:任务所在位置、周边会员个数、周边会员可接受任务数量以及会员的信誉值等,进

收稿日期:2018-03-05.

基金项目:南通理工学院教研课题(2017NITJG11).

通讯联系人:宋佳莹,硕士,助教,研究方向:应用统计. E-mail:18751994127@126.com

行统计分析,首先对任务定价进行主成分分析,提取主成分,线性模拟,并对其重新定价.^[2]由于任务完成情况最终处于两种状态,完成亦或未完成,故是二分类任务,对其建立多元 Logistic 回归模型,新模型的建立,最终使得任务完成情况达到最优状态.^[3]

1 研究过程

本文主要选取众包平台服务模式下的“拍照挣钱”行业的数据,基于已有的相关数据,对其进行相关统计分析,为众包平台服务的任务定价寻求一定的定价方式,以促进任务尽可能被完成.

第一,针对任务定价问题,首先分析 674 个已完成任务的定价方案,根据所给任务位置及会员的经纬度计算每个任务及会员之间的距离,并找出距每个任务 10 km 内会员个数(X_1)、10 km 内会员到任务点的平均距离(X_2)、平均信誉值(X_3)、以及平均预定任务限额(X_4),此处选择 10 km 作为分界线,主要是假设超过 10 km 时会员即认为距离太远,拒绝接受任务. 根据经济学原理,任务的定价(Y_1)与 X_1 、 X_2 、 X_3 以及 X_4 4 个变量有着密切联系. 首先分析了任务定价 Y_1 的 4 个影响变量之间的相关性,由于存在较高的相关性,故而使用主成分分析对多个指标变量进行降维,对降维后的数据与定价进行多元线性模拟,并对任务进行合理的重新定价.

第二,在以上重新定价的基础上,对影响任务完成情况的 5 个因素: X_1 、 X_2 、 X_3 、 X_4 以及任务的重新定价(X_5),首先进行多指标变量的降维处理,使用主成分分析方法,选取特征值大于 1 的两个主成分(FF_1 、 FF_2)与任务完成情况(Y_2)进行 Logistic 回归分析,计算每个任务完成概率,并与原定价方案进行比较,在任务定价总额和任务完成率上明显优于原方案,即模型拟合达到最优解.

2 数据来源及数据分析

本文选取的数据主要是分布在深圳及周边地区已完成的 674 个拍照任务,以及可以接收任务的共 1 877 位会员相关信息数据,已知信息整理如表 1.

表 1 相关信息表

Table 1 Related information sheat

已完成任务信息	会员信息	已完成任务信息	会员信息
任务号码	会员编号	任务标价	预定任务限额
任务 gps 纬度	会员 gps 纬度	任务执行情况	信誉值
任务 gps 经度	会员 gps 经度		

本文数据分析主要使用的指标变量如表 2.

表 2 相关指标变量

Table 2 Correlation indicator variable

指标变量名称	符号	指标变量名称	符号
任务标价	Y_1	任务的重新定价	X_5
任务执行情况	Y_2	任务定价影响因素提取的第一主成分	F_1
距离每个任务 10 km 内的会员数	X_1	任务定价影响因素提取的第二主成分	F_2
距离每个任务 10 km 内会员的平均距离	X_2	任务执行情况影响因素提取的第一主成分	FF_1
距离每个任务 10 km 内会员的平均信誉值	X_3	任务执行情况影响因素提取的第二主成分	FF_2
距离每个任务 10 km 内会员的平均预定任务限额	X_4		

2.1 数据处理

根据选取的 674 个任务以及 1877 个会员位置的经纬度数据,运用两个经纬度间距离公式计算出每个会员到每个任务的距离,两个经纬度间距离公式如下:

$$d=111.12\cos\frac{1}{\sin A_1\sin B_1+\cos A_1\cos B_1\cos(B_2-A_2)},$$

式中: A_1 为第一个任务点纬度; A_2 为第一个任务点经度; B_1 为第二个任务点纬度; B_2 为第二个任务点经度.

利用 Excel 表格筛选出距离每个任务 10 km 内的会员数(X_1),由于数据量比较大,利用 Matlab 软件计

算出距离每个任务 10 km 内所有会员的平均距离(X_2)、距离每个任务 10 km 内所有会员的平均信誉值(X_3)、距离每个任务 10 km 内所有会员的平均预定任务限额(X_4)。

2.2 数据分析

2.2.1 任务标价与任务所处位置分析

利用 Matlab 对任务标价、GPS 经纬度位置做三维散点图如下(图 1),一方面,可以看出任务的标价主要集中在 65~75 元之间,标价在 75 元以上的任务相对较少,任务的位置则是相对比较分散,一些集中的任务主要处于经度 114,纬度 22.8 附近,同样,不同的位置均有不同数量的“拍照”任务;另一方面,根据任务的经纬度,第一,以商业中心附近任务为基准,每个项目任务的标价为 65 元,向四周扩散,根据离商业中心距离的远近对任务标价相应增加,第二,根据任务的密集程度,在密集程度高的地方,任务标价相对较低,密集程度低的地方价格定得较高。因而需分析任务标价与任务位置等指标变量之间的关系,以便对任务进行重新定价。

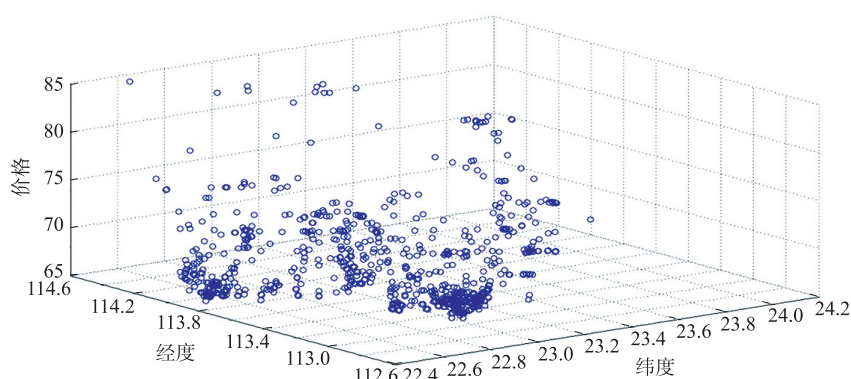


图 1 任务标价及 GPS 三维散点图

Fig. 1 Mission pricing and GPS three dimensional scatter diagram

2.2.2 任务执行情况与任务 GPS 位置分析

利用原始数据分析每个任务位置及完成情况,本文利用 Excel 表格画出任务执行情况所对应的每个任务所处位置,如图 2,可以看出任务执行情况与位置有着密切关系,未完成任务也相对集中于如下两个位置,分别是深圳偏远地区,及佛山相对不发达区域,通过对未完成任务的位置、周边会员情况、以及任务标价等的分析易知,未完成任务的位置多较为偏僻、会员个数相对较少,又由于任务标价没有显著性,导致了任务未能被完成。针对这些地理位置,可以采取相对抬高任务标价,以吸引会员完成任务,亦或者采取对附近任务进行打包,将未完成任务适当打包在已完成任务一起,有利于会员选择多个任务一起以使任务能够被完成。

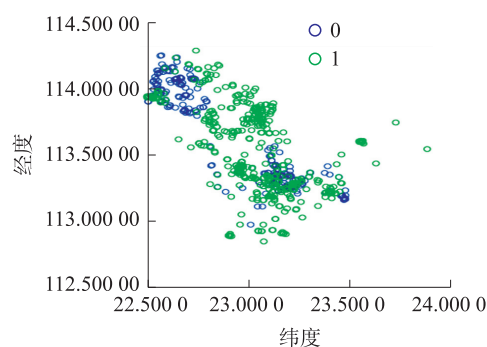


图 2 任务 GPS 位置散点图

Fig. 2 Mission GPS location scatter diagram

3 模型构建

3.1 标价与影响因素模型建立

3.1.1 数据标准化

根据实际情况,可知距离每个任务 10 km 内所有会员的平均距离(X_2)、距离每个任务 10 km 内所有会员的平均信誉值(X_3)、距离每个任务 10 km 内所有会员的平均预定任务限额(X_4),4 个因素是任务标价的主要影响因素。由于指标的量纲往往不同,所以在数据分析之前应先消除量纲的影响,消除数据的量纲有很多方法,常用方法是将原始数据标准化,即做如下数据变换:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, p,$$

式中: $\bar{x}_j=\frac{1}{n}\sum_{i=1}^nx_{ij},s_j^2=\frac{1}{n-1}\sum_{i=1}^n(x_{ij}-\bar{x}_j)^2$.

3.1.2 影响因素的主成分分析

根据数学公式可知,任何随机变量对其作标准化变换后,其协方差与其相关系数相同,即标准化后的变量协方差矩阵就是其相关系数矩阵.

(1)利用标准化后的数据计算变量之间的协方差矩阵,即可得原变量之间的相关系数矩阵: $\sum=(s_{ij})_{4\times 4}$,其中: $s_{ij}=\frac{1}{n-1}\sum_{k=1}^n(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j),\quad i,j=1,2,3,4$.

分析计算如表 3,由于相关分析的显著性水平 $\alpha<0.05$,故均通过了检验,可得出 4 个变量之间存在一定的相关关系.

表 3 相关性矩阵表
Table 3 Correlation matrix table

变量		X_1	X_2	X_3	X_4
相关性	X_1	1	-0.282	-0.371	-0.288
	X_2	-0.282	1	0.111	0.144
	X_3	-0.371	0.111	1	0.931
	X_4	-0.288	0.144	0.931	1
显著性(单尾)	X_1		0	0	0
	X_2	0		0.002	0.028
	X_3	0	0.002		0
	X_4	0	0.028	0	

(2)根据巴特利特球形检验的显著性水平 $\alpha<0.05$,因而该组数据适合做主成分分析.并求出 \sum 的特征值 λ_i 及相应的正交化单位特征向量 a_i , \sum 的前 m 个较大的特征值 $\lambda_1\geq\lambda_2\geq\cdots>0$,就是前 m 个主成分对应的方差, λ_i 对应的单位特征向量 a_i 就是主成分 F_i 的关于原变量的系数,则原变量的第 i 个主成分表达式 F_i 为:

$$F_i=a_i'X.$$

在计算特征向量前,首先计算主成分的方差(信息)贡献率,用来反映信息量的大小, a_i 为:

$$a_i=\lambda_i/\sum_{i=1}^m\lambda_i,$$

最终主成分个数的确定,即 F_1,F_2,\cdots,F_m 中 m 的确定是通过方差(信息)累计贡献率 $G(m)$ 来确定:

$$G(m)=\sum_{i=1}^m\lambda_i/\sum_{k=1}^p\lambda_k.$$

当累积贡献率达到 80%以上时,就认为所选主成分能足够反映原来变量的信息,对应的 m 就是抽取的前 m 个主成分.计算特征值相关数据如表 4,可得出特征值大于 1 的两个方差贡献率分别为 51.712%, 29.176%,累计方差贡献率为 80.888%,因而选取两个主成分用于反应存在相关关系的 4 个影响因素,用 F_1,F_2 来表示两个提取出的主成分.

表 4 总方差解释
Table 4 Total variance interpretation

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积/%	总计	方差百分比	累积/%
1	2.068	51.712	51.712	2.068	51.712	51.712
2	1.167	29.176	80.888	1.167	29.176	80.888
3	0.71	17.74	98.628			
4	0.055	1.372	100			

(3) 计算主成分载荷

主成分载荷是反映主成分 F_i 与原变量 X_j 之间的相互关联程度,原变量 $X_j(j=1,2,\cdots,p)$ 在诸主成分 $F_i(i=1,2,\cdots,m)$ 上的荷载 $l_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,p)$

$$l(Z_i, X_j) = \sqrt{\lambda_i} a_{ij} (i=1,2,\cdots,m; j=1,2,\cdots,p),$$

可得出主成分 F_i 对应的特征向量 a_i ,并根据主成分分析原理,得到如下形式的主成分表达式:

$$\begin{cases} F_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p, \\ F_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p, \\ \vdots \\ F_m = a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mp}X_p. \end{cases}$$

“成分矩阵”反应的就是主成分载荷矩阵. 根据成分矩阵数据计算对应主成分表达式,并作为影响任务标价的两个新变量:

$$\begin{aligned} F_1 &= -0.330\ 2X_1 + 0.181\ 543X_2 + 0.648\ 358X_3 + 0.661\ 726X_4, \\ F_2 &= -0.575\ 024X_1 - 0.739\ 42X_2 + 0.255\ 803X_3 + 0.239\ 158X_4. \end{aligned}$$

3.1.3 任务标价的多元线性回归

根据主成分因子 F_1, F_2 计算结果,对任务标价 Y_1 与 F_1, F_2 做多元线性回归模型,并对任务标价这一因变量进行重新定价,由于多元线性回归模型中方差分析表对应的显著性水平 $\alpha < 0.05$,因而该组数据适合进行多元线性回归,根据分析结果得出以下公式:

$$Y_1 = 64.527\ 41 + 0.000\ 646F_1 - 0.000\ 589F_2.$$

根据公式对任务标价重新进行定价,下文用 X_5 来表示任务的重新标价.

3.2 任务执行情况与影响因素的模型建立

3.2.1 影响因子的主成分分析

影响任务执行情况的因素主要有: X_1, X_2, X_3, X_4, X_5 . 对这 5 个影响因子进行数据标准化,并进行主成分分析,提取特征值大于 1 的两个主成分,结果如下:

$$\begin{aligned} FF_1 &= -0.250\ 83X_1 + 0.247\ 71X_2 + 0.534\ 64X_3 + 0.523\ 01X_4 + 0.562\ 43X_5, \\ FF_2 &= 0.458\ 101X_1 - 0.740\ 19X_2 + 0.337\ 53X_3 + 0.341\ 53X_4 - 0.108\ 14X_5. \end{aligned}$$

3.2.2 Logistic 回归模型

根据两个提取的主成分因子表达式计算出两个主成分对应的新变量,并将其作为影响任务执行情况 (Y_2) 的因素,进行 Logistic 回归模型建立.

(1) 逻辑回归 (Logistic regression) 属于概率型非线性回归,是研究二分类观察结果与一些影响因素之间关系的一种多变量分析方法. 假设响应变量 Y 是二分变量,令 $p = P(Y=1)$,影响 Y 的因素有 k 个 x_1, x_2, \cdots, x_k ,则称:

$$\ln \frac{p}{1-p} = g(x_1, x_2, \cdots, x_k)$$

为二分数数据的逻辑斯蒂 (Logistic) 回归模型,简称逻辑斯蒂回归模型. 其中的 k 个因素称为逻辑斯蒂回归模型的协变量. 最重要的逻辑斯蒂回归模型是 Logistic 线性回归模型,多元 Logistic 模型的形式为:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

式中, $\beta_0, \beta_1, \cdots, \beta_k$ 是待估参数. 根据上式可以得到优势的值:

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k},$$

参数 β_i 是控制其它 x 使 x_i 每增加一个单位对优势产生的乘积效应. 概率 p 的值为:

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}.$$

在模型计算出概率 p 值以后,根据实际问题,选择概率 p 取值范围,判断事件发生或是不发生.

(2)对应模型系数的 Omnibus 检验,显著性水平均小于 0.05,霍斯默-莱梅肖检验的显著性水平小于 0.05,可得出该数据适合建立 Logistic 回归模型.

根据软件分析结果得出 Logistic 回归模型的表达式为:

$$\ln \frac{p}{1-p} = 0.343\ 205 + 0.000\ 354FF_1 + 0.000\ 098FF_2.$$

为了计算最终任务执行情况的概率值,将上述表达式变形如下:

$$p = \frac{e^{0.343\ 205 + 0.000\ 354FF_1 + 0.000\ 098FF_2}}{1 + e^{0.343\ 205 + 0.000\ 354FF_1 + 0.000\ 098FF_2}}.$$

根据重新定价以后的数据建立模型并计算每一个任务执行情况的概率值,并设定概率值大于 0.61 的任务为最终可以完成任务,可以得出任务完成情况达到 97.7%,初始定价任务完成百分比约为 63.65%,易发现新定价模型所得结果远好于初始定价的任务完成情况. 因而认为模型建立有效.

4 总结与展望

通过对众包平台服务模式中任务定价进行重新定价及分析,可以得出以下结论:影响任务定价的最主要因素是任务周边会员的个数,可想而知,任务周边会员个数越多,会员会争相抢任务,任务的标价便可以定得较低,从而达到效益最高化. 当然任务周围会员的信誉值、会员任务承受能力均对任务标价以及任务最终能否完成产生一定影响. 在任务定价方面,最终目的都是为了达到利益最大化. 本文主要针对已有的任务进行价格模拟,利用统计数据分析,根据数据的相关规律进行重新定价,在费用保持一定的前提下,使得任务完成率增加,即达到逐渐最优化状态. 对于本文的模型,首先,可以使用在对新任务进行定价中,预测任务的执行情况,并及时进行改进,使得任务能够尽可能的都被会员接收完成. 其次,本文模型还可以推广到其他类似问题中,比如,近几年,餐饮业外卖任务的定价及利润预测以及物流行业的快递费用定价及预测等.^[4]其次,本文模型也可以有效地降低平台运营成本,在任务完成率和会员活跃度方面,均得以提高. 对会员而言,提高了会员获益的广度,同时未增加平台所付出的成本,达到双赢,所以无论从会员角度还是平台角度出发,该模型都是值得推广的模型. 再次,本文将统计模型与经济学相关问题相结合,为经济学方面的问题解决开辟了一条宽阔的道路,同时也使经济学从定性研究向定量方面研究转化,更加具有理性和发散思维.

[参考文献]

- [1] 范海峰. 产品定价决策影响因素及方法探讨[J]. 财会通讯(理财版),2007(8):16-17.
- [2] 张博,戴薇. 基于主成分分析的我国整车上市公司财务状况评价[J]. 西北农林科技大学学报(社会科学版),2012,12(5):102-106.
- [3] 董纯洁. 基于实例与逻辑回归的多标签分类模型[D]. 南京:南京大学,2013.
- [4] 唐海龙. 第三方物流企业成本分析及定价模式研究[D]. 大连:大连海事大学,2005.

[责任编辑:陆炳新]