Modeling and Simulation 建模与仿真, 2023, 12(3), 2067-2075
Published Online May 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.123190

Hans汉斯

基于K-Means聚类的多元回归拟合定价模型

薛可人

上海理工大学机械工程学院，上海

收稿日期：2023年2月27日；录用日期：2023年5月5日；发布日期：2023年5月12日

摘　要

近年来，基于移动互联网的自助式劳务众包平台日渐火爆，用户可通过拍照做任务赚取酬金，其相比传统的市场调查方式可以大大节省调查成本，平台中的任务定价是其核心要素。若因为定价不合理，则用户人数减少，影响平台任务的正常运行，导致平台竞争力减弱。针对如何合理定价这一问题，本文采用“K-means聚类”作为主要研究手段，运用多元回归方程法，考虑多种因素的影响并归一化分析，取离散度、集中度最优的因素作为主要影响因子，建立多元回归方程，以利润为指标，找出各因素与价格最优的关系，评判本多元回归拟合定价模型的优劣。

关键词

K-Means聚类，多元回归分析，定价策略

Multiple Regression Fitting Pricing Model Based on K-Means Clustering

Keren Xue

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 27th, 2023; accepted: May 5th, 2023; published: May 12th, 2023

Abstract

In recent years, the self-service labor crowd-sourcing platform based on mobile internet has become increasingly popular. Users can earn remuneration by taking photos to do tasks. Compared with traditional market research methods, it can greatly save the cost of investigation. The task pricing in the platform is its core element. If the pricing is not reasonable, the number of users will decrease, which will affect the normal operation of the platform tasks and weaken the competitiveness of the platform. To solve the problem of how to reasonably price, this paper uses

文章引用: 薛可人. 基于 K-Means 聚类的多元回归拟合定价模型[J]. 建模与仿真, 2023, 12(3): 2067-2075.
DOI: 10.12677/mos.2023.123190

"K-means clustering" as the main research means, uses multiple regression equation method, considers the influence of multiple factors and normalizes the analysis, takes the factors with the best dispersion and concentration as the main influencing factors, establishes multiple regression equation, takes profit as the index, finds out the relationship between each factor and the best price, and judges the advantages and disadvantages of this multiple regression fitting pricing model.

Keywords

K-Means Clustering, Multiple Regression Analysis, Pricing Strategy

1. 引言

随着移动互联网以及智能手机的普及，“拍照赚钱”的服务模式迅速发展起来，用户可以下载应用程序，注册并接收需要他们拍照的任务(例如，检查超市的产品可用性)，用户完成任务后可得到一定报酬[1]。这一自助劳动众包平台为企业节省了业务检查和信息收集的费用，同时也确保了真实性并缩短了调查周期[2]。该应用程序对平台的运营至关重要，任务定价是核心要素。如果定价不合理，可能会忽略某些任务，导致检查失败。杨新平等[3]建立了基于随机森林方法的定价模型，该项研究为 APP 自助服务平台的新任务价格标定提供参考和借鉴。王妍等[4]运用数据可视化、层次分析法、模糊综合评价等方法，分别构建了任务定价模型、会员等级综合评价模型以及改进的综合定价模型。综合各位学者在定价模型的研究基础上，本文将关注点放在如何在完成尽可能多的任务情况下减少成本，即合理的定价成为一项急需解决的关键问题，本文考虑任务密度，会员位置，地区执行力等影响因素，将这些影响因素与价格拟合，得出多个多元线性方程式，然后将离散度、集中度合适的影响因素归一化处理，得出新的关系式，以此得到新的多元回归定价模型[5] [6] [7]。

2. 模型构建

2.1. K-Means 聚类简介

K-means 聚类算法是具有代表性的聚类方法，它在处理大规模数据集时可以保持良好的伸缩性和高效性[8]。该算法是一种基于距离的聚类算法，采用样本间的距离作为样本相似性度量，其主要思想是从数据集中随机选择 k 个样本对象，作为初始聚类中心，计算每个样本点到聚类中心的距离，通过迭代过程逐次更新聚类中心的位置，直到更新的聚类中心不再变化或变化小于某个阈值时，迭代算法终止，最终将整个数据集划分为 k 个不同聚类[9]。它适用于处理凸数据集或球形簇，具体的算法步骤如下：

Step 1：随机初始化 k 个聚类中心点，并计算数据中每个点到 k 个聚类中心点的距离；

Step 2：将每个数据点分到距离聚类中心点最近的聚类中心中；

Step 3：针对每个类别重新计算聚类中心；

Step 4：重复上面的 Step2、3，直到达到预先设置的停止条件(迭代次数、最小误差变化等)。

K-means 聚类算法使用欧氏距离作为样本相似性度量，样本点的距离越近，表明这 2 个样本点的相

似性越大，样本点之间的欧式距离公式为

$$d\left(x_i,x_j\right)=\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2} \tag{1}$$

式中，$d\left(x_i,x_j\right)$表示样本点x_i和x_j之间的欧式距离；$x_i=\left(x_{i1},x_{i2},x_{i3},\cdots,x_{ip}\right)^{\mathrm{T}}$和$x_j=\left(x_{j1},x_{j2},x_{j3},\cdots,x_{jp}\right)^{\mathrm{T}}$均表示$p$维数据集中的两个样本点。

使用误差平方和作为K-means算法的准则函数，其表达式为：

$$SSE=\sum_{i=1}^{k}\sum_{x\in x_i}\left\|x-\mu_i\right\|^2 \tag{2}$$

式中，k为聚类数；x为聚类集x_i中的样本对象；μ_i为聚类中心；SSE为样本点的密集程度，理论上来说，SSE的值越小，聚类效果越好。

2.2. 多元回归定价模型

记因变量任务标价为y，自变量多个影响因素分别为$x_1,x_2,\cdots,x_n$，为了大致分析y与$x_1,x_2,\cdots,x_n$的关系，利用MATLAB做出任务标价y分别关于影响因素一x_1、影响因素二x_2、…影响因素x_n之间的多元回归模型为：

$$y=c_0+c_1x_1+c_2x_2+\cdots c_nx_n+\varepsilon \tag{3}$$

其中，c_0、c_1、$c_2\cdots c_n$为待估计的回归系数，ε表示随机误差。

回归系数矩阵的求解公式为：

$$A=\left(\sum x_ix_i^{\mathrm{T}}\right)^{-1}\left(\sum x_iy\right) \tag{4}$$

3. 模型求解与分析

本文根据文献[10]中的附件数据，结合当今人们的消费习惯和心理，分别考虑任务密度、会员密度和执行能力这三大影响因素对价格的影响，先找出各个因素对价格的拟合方程，最后进行多元回归，归一化处理，得到总的关系式，从而合理定价。

3.1. 影响因素一：任务密度

任务密度与价格很难找到直接的联系，本文采用“中心点媒介法”，运用MATLAB中K-means函数，寻找最优中心点，以中心点为媒介，构建两者的桥梁，然后逐渐向外扩展，找到单位圆环面积内价格的平均值，以及任务密度，以此拟合任务密度与价格的关系。位置信息可以看作样本点$\left\{x^{(1)},x^{(2)},\cdots,x^{(m)}\right\}$。基于K-means函数的聚类，找中心点的详细步骤如下：

1) 随机选取k个图像作为处理的中心点作为随机质心，将位置数据一次赋值为聚类质心$\mu_1,\mu_2,\cdots,\mu_m\in R_n$；

2) 对剩余每个对象测到所选中心点的距离，重复过程直到收敛，并归为随机质心类。将与初始聚类中心距离最近的对象，分配利同一个聚类中；

3) 对于每一个样例i，计算其应该属于的类，如式(5)所示

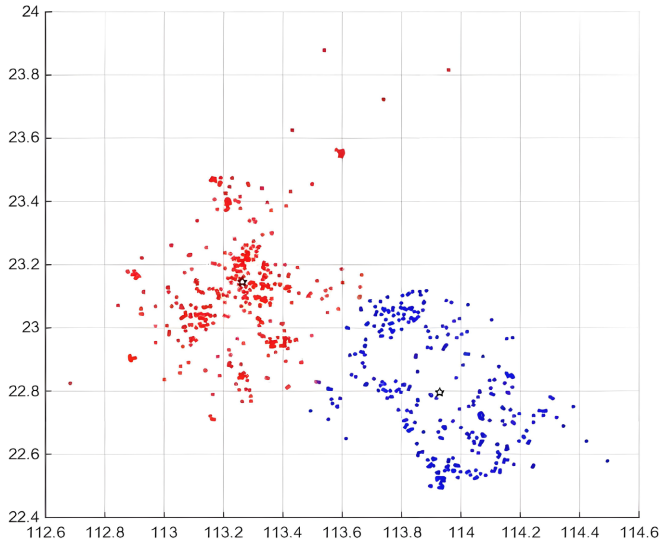
$$C^{(i)}=\arg\min\left\|x^{(i)}-\mu_c^{(i)}\right\|^2 \tag{5}$$

$$J\left(c,\mu\right)=\sum_{i=1}^{m}\left\|x^{(i)}-\mu_c^{(i)}\right\|^2 \tag{6}$$

4) 对每一个类 j，重新计算随机质心的经纬度。

$$\mu_j = \frac{\sum_{i=1}^{m} L\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{m} L\left\{c^{(i)} = j\right\}} \tag{7}$$

5) 计算每个聚类的总需求量，并对其进行降序排列。迭代 3 次，求合适的阈值。找出的中心点图 1 和表 1 所示：



注：黑色五角星处为中心点。

Figure 1. Cluster center point diagram
图 1. 聚类中心点图

Table 1. Coordinate of cluster center
表 1. 聚类中心点坐标

聚类	经度	纬度	附近任务点
1	22.726591	114.195465	175
2	21.570457	111.085959	3
3	22.993464	114.7285460	1
4	23.275771	113.611790	62
5	33.652050	116.970470	1
6	22.262783	112.797680	1

构建密度分布模型(软件构建)：

以选好的中心点为基础，向外扩展半径，上一步中 K-means 计算结果得出可行域 R=。在给定范围中均分为 2700 个小区，使 $i=1:2700$ 进行 2700 次求解，$r=0.0001*i$，每次以 0.001 为单位进行扩张，根据欧氏距离公式，找出距离中心点的参数点集。

$$d(x,y)=\sqrt{\sum_{k=1}^{3}(x_k-y_k)^2} \tag{8}$$

进行 for 循环求这个点到中心点的距离，判定该点是否在目标范围之内再通过累加器求取密度分布。结果图 2 如下：

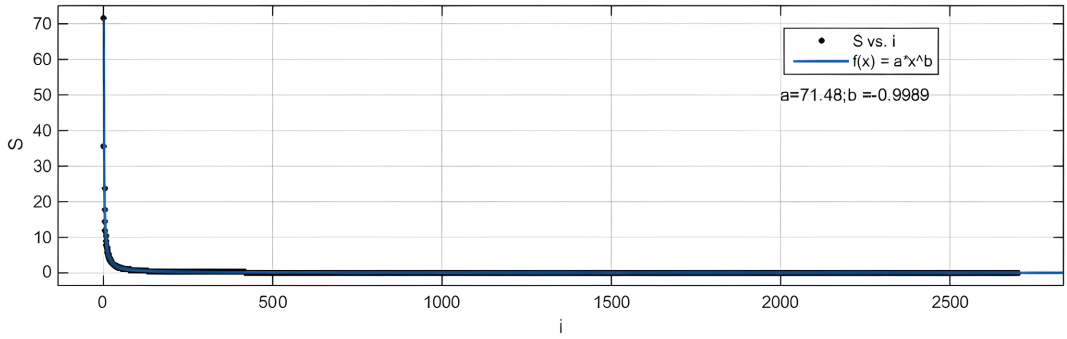


Figure 2. Density and radius
图 2. 密度与半径

最后将任务密度与价格进行模型建立，结合数据得出最终关系式：

$$f(x)=0.0007972*x^4-0.02218*x^3+0.1899*x^2+0.5023*x+0.1532 \tag{9}$$

可以看出用 K-means 聚类算法，拟合决定系数 $R^2=0.9932$，接近于 1，精确的拟合了任务密度与价格之间的关系。

3.2. 影响因素二：会员密度

会员密度与任务密度两者和价格的关系，分析模式有相似之处，在此不再赘述重复的建模过程，利用聚类找中点方法与上一部分类似，不再重述。得到会员分布图 3 如下：

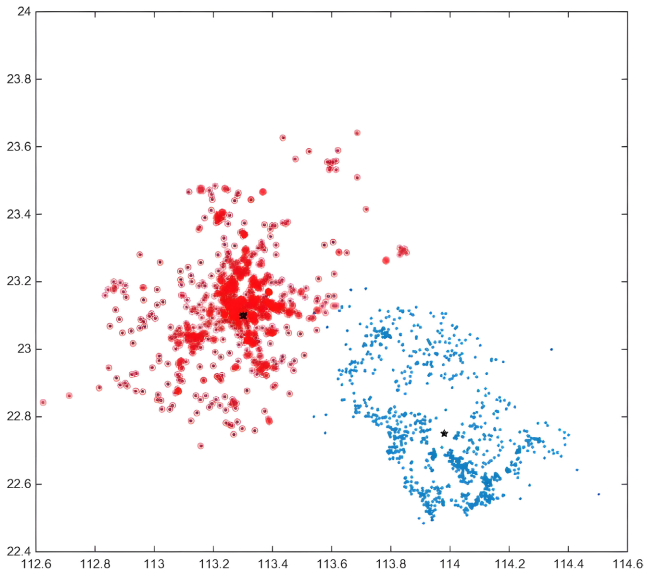


Figure 3. Distribution of member centers
图 3. 会员中心分布

构建会员密度与任务密度的区别在于，任务密度可以直接的对应着价格，而会员没有所对应的价格，所以还要通过会员所在经纬度位置与价格连上关系，转接会员与价格的关系。最后得出结果如图 4 所示：

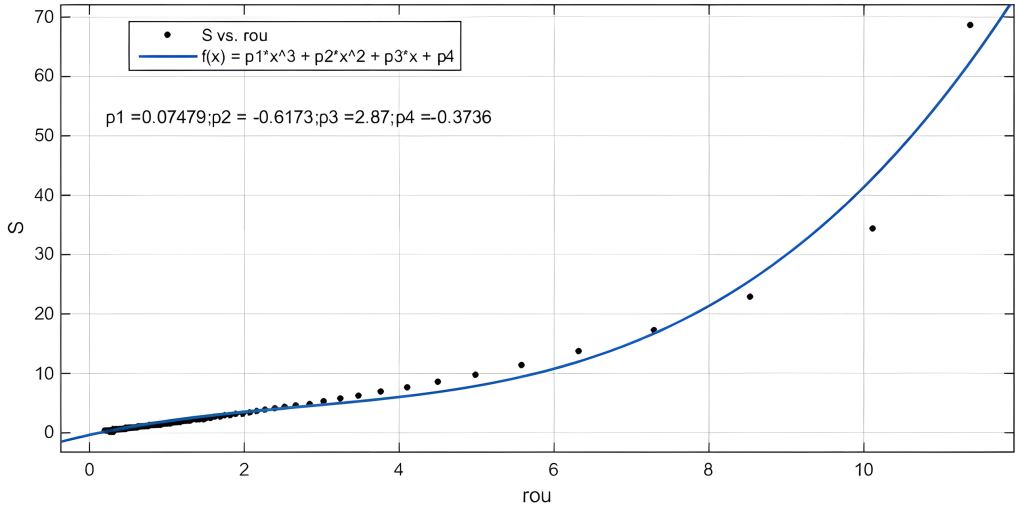


Figure 4. Membership density and price
图 4. 会员密度与价格

编写程序，拟合多元线性方程，得出最终方程式：

$$f(x) = 0.07479 * x^3 - 0.6173 * x^2 + 2.87 * x - 0.3736 \tag{10}$$

可以看出用 K-means 聚类算法，拟合决定系数 $R^2 = 0.9771$，接近于 1，精确的拟合了会员密度与价格之间的关系。

3.3. 影响因素三：执行能力

执行能力指：某一个地区或者某一个会员对所给任务的完成性。执行能力可以反映出对任务的亲和度，所以，我们运用了一种“中心点扩散法”算法，通过信誉度和任务限额两个数据，得出每个会员的执行能力，如图 5 所示。设执行能力 $= y$，任务限额 $= a$，归一化信誉度 $= c$，最大信誉度 $= c_{\max}$，最小信誉度 $= c_{\min}$，则：

$$y = a * c \tag{11}$$

$$c = \frac{x - c_{\max}}{c_{\max} - c_{\min}} \tag{12}$$

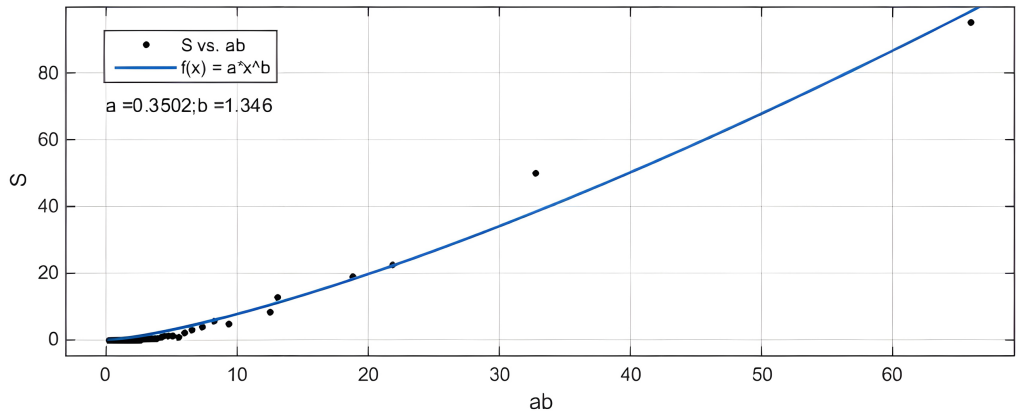


Figure 5. Execution ability and price
图 5. 执行能力与价格

本文首先考虑的是个人执行能力与价格的关系，但是经过回归拟合分析后，拟合精度不足 0.9，结果较差，不能作为有力的说明论据。故经改进后，选择考虑地区的执行能力，利用“中心点扩散法”将地域能力值与价格构成联系。去除噪点后，将能力值与经纬度做拟合曲线，如下图 6 所示：

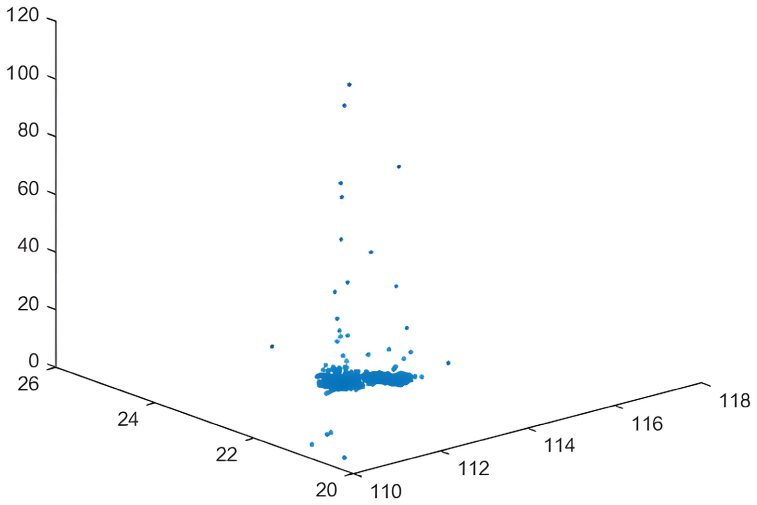


Figure 6. Execution ability and longitude and latitude
图 6. 执行能力与经纬度

得出最终方程：

$$f(x)=0.3502*x^{1.346} \tag{13}$$

可以看出用 K-means 聚类算法结合“中心点扩散法”，拟合决定系数 $R^2=0.9814$，接近于 1，精确的拟合了执行能力与价格之间的关系。

4. 多元回归模型分析

整理上述各因素的拟合关系式如下表 2：

Table 2. Function fitting formula and characteristics of three major influencing factors
表 2. 三大影响因素函数拟合式及特点

影响因素	函数拟合关系	函数模型	特点
任务密度	$f(x)=0.0007972*x^4-0.02218*x^3+0.1899*x^2+0.5023*x+0.1532$	四阶线性拟合	$x\in[0,320]$ 时曲线上升
会员密度	$f(x)=0.07479*x^3-0.6173*x^2+2.87*x-0.3736$	三阶线性拟合	$x\in[0,12]$ 时曲线上升
执行能力	$f(x)=0.3502*x^{1.346}$	幂函数拟合	$x\in[0,70]$ 时曲线上升

经过这三个主要因素分别与价格的关系，为了综合各方面的因素，得出最优方案，将这三个拟合方程做归一化处理，记任务定价为 y，任务密度为 x_1，会员密度为 x_2，执行能力为 x_3。基于以上的分析，我们利用 x_1，x_2，x_3 建立 y 的定价方案模型。

为了方便分析 y 与 x_1， x_2， x_3 的关系，将 y 与 x_1 之间的四阶线性模型记做模型一，y 与 x_2 之间的三阶线性模型记做模型二， y 与 x_3 之间的幂函数模型记做模型三。

模型一： $Y = p_0 * x^4 + p_1 * x^3 + p_2 * x^2 + p_3 * x + p_4$；

模型二： $Y = P_0 * x^3 + P_1 * x^2 + P_2 * x + P_3$；

模型三： $Y = a * x^b$。

综合以上分析，结合模型一、模型二及模型三，建立如下的回归模型：

$$y_i = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij} + \cdots \beta_n x_{i\gamma} + \varepsilon_i \tag{14}$$

其中， β_0、 β_1、 $\beta_2 \cdots \beta_n$ 为待估计的回归系数， ε_i 表示随机误差，实际样本量为 γ，第 i 次观测值为 $x_{i1}, x_{i2}, x_{i3}, \cdots, x_{i\gamma}, y_i \ (i = 1, 2, \cdots, n)$。

联立成多元线性方程组，其 n 次的观察值可写为如下形式：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots \beta_\gamma x_{1\gamma} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots \beta_\gamma x_{2\gamma} + \varepsilon_2 \\ \cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots \beta_\gamma x_{n\gamma} + \varepsilon_n \end{cases} \tag{15}$$

和一元线性回归分析一样，我们假定是相互独立且服从同一正态分布 $N(0, \sigma)$的随机变量。由于残差平方和

$$Q = \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{m} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots b_r x_{ir} \right) \right]^2 \tag{16}$$

所以最小值一定存在，根据极值原理，当 Q 取得极值时，b 应满足： $\frac{\partial Q}{\partial b} = 0$，将三个影响因素方程式列为方程组：

$$\begin{cases} \sum_{i=1}^{m} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots b_r x_{ir} \right) \right] = 0 \\ \sum_{i=1}^{m} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots b_r x_{ir} \right) \right] * x_{i1} = 0 \\ \sum_{i=1}^{m} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots b_r x_{ir} \right) \right] * x_{ij} = 0 \\ \sum_{i=1}^{m} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots b_r x_{ir} \right) \right] * x_{ip} = 0 \end{cases} \tag{17}$$

根据上述式(14)~(17)推导，计算得出最后多元线性回归结果：

$$f(x) = 61.18 - 5.87 A_i - 0.44 R_i + 15.16 D_i \tag{18}$$

其中， A_i、 R_i 和 D_i 分别为可能存在的各方面影响因素，该模型不光局限于本文提到的三个影响因素，其具有普遍使用性。

5. 模型优劣评价

为验证任务定价模型的有效性，随机从竞赛 B 题附件 1 中选取 7 组任务的价格数据，将其与所建立模型结果进行对比，比较的结果如表 3 所示，采用相对误差作为评价指标，结果表明 7 组数据中相对误差最大的 1.47%，可以说本文研究的定价模型能够较好的反映真实数据，具有良好的可靠性。

Table 3. Comparison between the predicted value of the model and the actual value
表 3. 模型预测值与真实值对比

任务编号	A0179	A0317	A0344	A0554	A0680	A0740	A0830
实际值	67.0	65.5	66.0	67.5	68.0	80.0	85.0
计算值	67.5	65.6	66.8	67.1	69	79.2	85.3
相对误差	0.75%	0.15%	1.21%	0.59%	1.47%	1.00%	0.35%

6. 小结

本文考虑了任务密度、会员密度和执行能力三方面因素，采用数据计算结果与原方案相比：1) 降低了东莞的定价，保证了较高的完成率，又提高了公司利润。2) 深圳任务标价适当提高，使会员尽可能多的接受任务，提高了任务完成度。3) 对于广州，佛山本身原方案完成率不算很低，通过价格与完成率得出的利润可以接受，对其价格进行微调，提高了利润。

综上所述，本模型具有高度统计学意义，故该定价模型具有普遍适用性，对于不同的任务项目，可用该模型做出合适的任务标记，具有实际意义。

参考文献

[1] 崔艺馨, 沈梓崴, 张旭. 利用“拍照赚钱”新型商业模式研究[J]. 现代商业, 2019(36): 13-16.

[2] 覃珂槟, 赵庚升, 邢敏. 基于智能 APP 自助服务平台的商业模式探索[J]. 中国集体经济, 2013, 389(21): 34-35.

[3] 杨新平, 杨云源, 黄建飞, 等. 基于随机森林的“拍照赚钱”定价方案分析[J]. 楚雄师范学院学报, 2019, 34(3): 133-138.

[4] 王妍, 朱家明, 王欣宇, 等. 基于定性数据分析对“拍照赚钱”任务定价的研究[J]. 焦作大学学报, 2019, 33(1): 68-73.

[5] 娄思佳, 王书博, 王奇. 基于多元线性回归的任务定价规律模型[J]. 中国高新区, 2018(4): 40.

[6] 王书博. 基于改进梯度下降法求解多元线性回归方程[J]. 数学的实践与认识, 2022, 52(10): 167-172.

[7] 王华丽. 多元线性回归分析实例分析[J]. 科技资讯, 2014, 12(29): 22+24.

[8] 周世兵, 徐振源, 唐旭清. K-means 算法最佳聚类数确定方法[J]. 计算机应用, 2010, 30(8): 1995-1998.

[9] 张嘉龙. 一种新的选取 K-means 初始聚类中心算法[J]. 现代计算机, 2021, 726(18): 56-59.

[10] 全国大学生数学建模竞赛组委会. 2017 高教社杯全国大学生数学建模竞赛(CUMCM)题目 B 题[EB/OL]. 全国大学生数学建模竞赛网站. https://max.book118.com/html/2017/0903/131501127.shtm, 2017-09-14.