知识付费用户参与度驱动因素差异研究

——基于社会资本视角

摘要：随着社会的快速发展，越来越多的社会化问答平台开始提供知识付费产品服务。本文以知乎live为例，通过Python爬取了6161条已结束的知乎live数据。首先基于知识供给者的行为信息，通过K-Means聚类将其划分为“专家型”、“中庸型”和“信息搜寻型”三组用户群体，其次构建lasso回归分析不同群体下影响付费用户参与度因素的差异，结果发现知识供给者累计的社会资本越多，更能激发他人知识付费行为；结构维度下的粉丝数量在“信息搜寻型”中对付费用户参与度的正向影响效应最大；认知维度下的两个变量在“专家型”群体中的正向影响效应最大；而关系维度的影响效应最弱，且只在“专家型”群体中有较小的促进作用。最后本文从知识供给者和平台两个角度为提高付费用户参与度提出建议。

关键词：知识付费行为；社会资本；K-Means聚类；lasso回归

**Research on Differences of Impact Factors of Users Paying for Knowledge**

——**Based on Social Capital Theory**

**Abstract:** With the rapid development of society, more and more socialized Q&A platforms have begun to pay for knowledge-based products. This article takes knowledge of live as an example, and through Python, it has crawled 6161 pieces of closed live data. Firstly, based on the behavior information of knowledge providers, K-Means clustering is divided into three groups: “expert type”, “moderate type” and “information search type”. Secondly, lasso regression analysis is used to analyze the affected users under different groups. The difference in participation factors shows that the more social capital accumulated by knowledge providers, the more they can stimulate others' knowledge payment behavior; the number of fans under the structural dimension has the greatest positive impact on the participation of paid users in the “information search type”. The two variables under the cognitive dimension have the greatest positive effect in the “expert” group; while the relationship dimension has the weakest effect and only has a small promotion effect in the “expert” group. Finally, this paper proposes suggestions for increasing the participation of paying users from the perspective of knowledge providers and platforms.

**Key words:** knowledge payment behavior; social capital; K-Means; Lasso

**目 录**

[1 绪论 1](#_Toc8067493)

[1.1 引言 1](#_Toc8067494)

[1.2 文献综述 1](#_Toc8067495)

[1.2.1 知识付费文献综述 1](#_Toc8067496)

[1.2.2 社会资本综述 2](#_Toc8067497)

[1.3 研究内容 3](#_Toc8067498)

[2 数据来源及知乎live基本现状 3](#_Toc8067499)

[2.1 数据抓取 3](#_Toc8067500)

[2.2 数据预处理 4](#_Toc8067501)

[2.3 主讲人用户群体分布 5](#_Toc8067502)

[2.4 知乎live现状分析 6](#_Toc8067503)

[2.5 相关性分析 7](#_Toc8067504)

[3 主讲人用户群体分组 8](#_Toc8067505)

[3.1 K-Means聚类分析算法介绍 8](#_Toc8067506)

[3.1.1 K-Means聚类分析原理 8](#_Toc8067507)

[3.1.2 评价聚类的指标 9](#_Toc8067508)

[3.2 分组主讲人用户基本特征描述 9](#_Toc8067509)

[4 用户参与度驱动因素差异性分析 12](#_Toc8067510)

[4.1 lasso回归介绍 12](#_Toc8067511)

[4.2 分组散点图 13](#_Toc8067512)

[4.3 模型结果分析 15](#_Toc8067513)

[5 结论建议 16](#_Toc8067514)

[5.1 结论 16](#_Toc8067515)

[5.2 建议 17](#_Toc8067516)

[5.2.1 针对知识供给者 17](#_Toc8067517)

[5.2.2 针对平台 18](#_Toc8067518)

[参考文献 19](#_Toc8067519)

[附录A 相关python代码 20](#_Toc8067520)

# 1 绪论

## 1.1 引言

移动网络的快速发展使得人们获取知识和信息共享的方式发生了巨大的改变。但是随着大众接受信息的渠道越来越多，信息的冗余让用户无法精准获取优质内容，为了更好地提升用户体验，内容市场开始走向差异化、多样化化和类型化的内容建设阶段。在此背景下，知识付费模式应运而生，并已成为人们知识获取及分享的一种新型方式。根据艾瑞咨询发布的市场报告，2017年我国知识付费产业规模约为49.1亿，同比增长近三倍[1]。如今，整个知识付费市场体系变得规范化，付费网络问答平台作为知识经济产品消费新形态，可帮助人们更高效地获得信息和知识，很多用户也愿意为髙质量问答买单。

虽然知识付费市场发展潜力巨大，但仍然还面对巨大的挑战。一方面，在社会化问答平台上，知识供给者可通过发表文章、知识问答等方式累计自己的社会资本，供给者用户群体的差异性也变得更加明显；另一方面，知识付费发展到如今阶段，用户的消费观念变得越来越理性，大部分产品都出现了用户流失和活跃度下降的现象。如何充分发挥网络上累计的用户行为数据，以促进整个知识付费行业的良性发展，会成为目前以及将来的研究方向。

在此背景下，本文以知乎live为例，爬取了知识供给者相关行为数据及live参与人数信息。由于网络用户群体差异明显，首先对知识供给者进行分组，并基于社会资本理论探讨在不同知识供给者群体下，付费用户参与度的影响因素及差异，旨在为更好为提高付费用户活跃度、洞察用户行为特征、促进付费平台健康发展提供参考和建议。

## 1.2 文献综述

### 1.2.1 知识付费文献综述

网络知识付费平台的出现开启了一种新型知识消费形态，大众全体也越来越接受这一消费模式。在此背景下，已有不少国内外学者对知识付费模式进行了研究。在发展趋势上，张杨燚（2017）基于知识差异化、去中心化、版权保护正向激励以及知识价值链线这四个角度，提出巩固知识付费模式核心竞争力的建议。罗星（2018）提出目前知识付费平台存在专栏内容的同质化现象、以及当前面临的精英壁垒和遇到的法律困境问题。

在用户付费意愿方面，Dou（2014）基于信任风险角度，通过实证分析发现，用户在付费行为决策过程中更多关注的是内容产品的价值以及平台的权威性；韩芳芳（2017）基于TAM模型、VAM模型及感知价值理论，发现用户更多关注的是付费问答平台所回馈问答知识的有用性、和使用过程的娱乐性，还有付费问答部分入驻用户的权威专业性。

在产品属性层面，胡盼弟（2018）首先研究了在网络问答服务产品在生产与消费中的运营模式，并以长尾理论为起点，提出解决知识付费在实际运用过程中出现问题的方法；王佩文（2018）则根据专业化和垂直化细分市场在知识付费领域将会崛起的想法，设计了一款服务中国赴美攻读MBA学位的申请人，北美MBA项目组和企业单位多方的互联网产品雏型。

纵观学者的研究成果，从研究内容上看，目前的研究大都着重与知识付费产品属性、发展以及从产品角度的用户付费意愿，鲜有从主讲人用户群体维度展开的研究；在研究方法上，大多学者主要以理论分析或者调查问卷的方式，未能充分利用社交网络平台上累积的用户行为数据。而知乎用户群体庞大，也很少有学者对用户群体进行划分，再研究其付费用户活跃度驱动因素差异。因此本文通过网站上直接可以爬取的信息，基于主讲人的行为数据以量化社会资本，并对用户群体进行细分，基于主讲人不同社会资本条件下，洞悉付费用户参与度的驱动因素差异。

### 1.2.2 社会资本综述

社会资本指嵌入在社会网络中的可以被获取和利用的资源的总和，而个体能够利用这些资源达成相应的目标。社会学家布尔迪厄、科尔曼、帕特南等人提出并发展了社会资本理论。他们将存在于社会成员之间的信任、规范、网络等认为是一种可利用的资源，并且这种资源可作为生产要素进入生产领域，影响现实的经济生活。

近年来，知识共享模式的迅猛发展，使得越来越多的国内外学者将社会资本理论应用到该领域中。张璐、张鹏翼（2017），分析线下和线上社会资本用户的社区行为模式以及累计社会资本的特点，结果发现两者之间存在正向关联；Chung（2016）等基于知识供给者三个社会资本特征角度，发现网络外部性、社交互动和自我形象展现对用户参与知识共享意愿具有显著正向影响；G.Li（2014）通过问卷调查，设计指标体系量化校园聚合频道网站用户的结构性、关系性、认知性社会资本，发现早期用户的认知性社会资本与持续参与网站并贡献内容的意图有显著关联。

在社交问答平台上，社交关系的差异会导致个人社会资本特征的不同；因此本文将从社会领域中常用的三个维度即结构、关系和认知维度衡量知乎live主讲人用户的社会资本特征。结构维度表现为社交网络交互连接，在本研究中，根据毕莹莹（2015）提到的，结构维度指的是用户间的关注与被关注关系，故将主讲人的“粉丝数”和“关注人数”作为衡量结构维度的指标；认知维度通常指的是语言通用度和价值观相似度等，在社交网络中，两个人的价值观越相似，获得双方认可的可能性就越高。因此对于付费用户而言，在选择知识产品时，往往会选择与其有相同认知和兴趣的知识供给者，而这可以通过点赞收藏等方式体现，故一个人的结构维度可通过“赞同数”，“被收藏数”体现[15]；关系资本源于信任和认同，社群中成员间的信任以及对组织的共同目标的强烈认同，对该成员在该组织中的行为具有促进作用，而信任来自于个人能力。故本文将知识供给者发表的“文章数”和“参与公共编辑数”作为衡量关系维度的标准。

## 1.3 研究内容

本文基于爬取的数据，首先通过聚类分析对知识供给者群体进行细分，并识别相关群体特征，洞察其社会资本差异，以对不同类型的用户提出针对性建议；其次通过lasso回归探究不同社会资本条件特征下平均参与live人数的影响机理，寻找影响付费用户参与度的因素及影响程度，这对于不同类型下的主讲人用户如何维持和增加live参与数量可以起到非常好的指导作用。最后本文根据相关研究结论，为知识付费平台更好地提高用户活跃度，提升付费服务质量提供参考，最后研究思路如图1.1所示。

知识供给者基本

信息及行为指标

划分为不同

知识供给者群体

分析不同群体

社会资本特征

识别不同群体下付费用户

参与度驱动因素及其差异

为提升知识付费平台用户活跃度提出建议与支持

聚 类

lasso回归

图1.1 研究思路图

# 2 数据来源及知乎live基本现状

## 2.1 数据抓取

近年来，知识付费行业迈入了高速发展时期。因此，本文对用户规模庞大的网络知识分享社区——“知乎”进行实证研究。在2016年，知乎推出了实时问答互动产品——知乎“live”，它鼓励具有专业知识或者技能的用户通过语音方式在线分享知识、解答疑惑，而其他感兴趣的用户则可以通过付费来满足自己的知识需求。

本文通过Python在线爬取了已结束的6160条知乎live以及所对应主讲人用户信息的数据，每一条数据样本共提取了7个特征变量，具体如表2.1所示：

表2.1 变量说明

|  |  |  |  |
| --- | --- | --- | --- |
| 变量类型 | 变量名 | 代码 | 变量说明 |
| 因变量 | 用户参与度 | taken | 截止数据抓取日参与该live的人数 |
| 自变量 | 结构型社会资本 | follower\_count | 对应主讲人用户的粉丝数 |
| following\_count | 对应主讲人用户关注的人数 |
| 认知型社会资本 | favorited\_count | 主讲人累计被收藏的数量 |
| voteup\_count | 主讲人用户累计获得的赞同数 |
| 关系型社会资本 | articles\_count | 主讲人用户累计发表的文章数目 |
| logs\_count | 主讲人用户参与公共编辑的次数 |

## 2.2 数据预处理

一个知识供给者会主讲多个live主题，而每个主讲人用户会有自己的侧重领域，因此需要对参与live的人数进行加权平均，其中：

权重=每一个领域下主讲人用户举办的live数/总的live数量。

由于数据样本中存在极值的情况，分布差异较大，故对因变量和自变量都采用取对数的方式进行处理，对于0的情况，则将所有0只均替换为1，参与对数运算。

在对数据进行以上处理后发现，一些样本中0值占据的比重较大，故删除每一条样本中0值在4个以上的记录。

具体的数据预处理思路如图2.1所示：

对于每一个主讲人用户（共m场live）

**所属领域1**

举办的live主题1

举办的live主题2

……**m1**场live

参与总人数**n1**

权重1=**m1/m**

**所属领域2**

举办的live主题1

举办的live主题2

……**m2**场live

参与总人数**n2**

权重2=**m2/m**

平均参与live人数=∑（n1\*权重1+n2\*权重2+…）

**对数处理**

（0替换成1参与运算）

针对每一条样本，如果4个及以上的变量值为0，则在建模时将此记录删除

图2.1 数据预处理思路

## 2.3 主讲人用户群体分布

表2.2 连续性变量的的描述性统计分析

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变量 | 最大值 | 最小值 | 均值 | 标准差 | 偏度 | 峰度 |
| 平均参与live人数 | 4195352 | 5 | 7537 | 104889.6 | 33.12 | 1239.58 |
| 粉丝数 | 1957273 | 11 | 33912 | 90117.3 | 8.51 | 123.07 |
| 关注人数 | 6082 | 0 | 186 | 375.8 | 7.43 | 86.69 |
| 被收藏数 | 1707770 | 0 | 33501 | 90563.7 | 7.21 | 82.33 |
| 赞同数 | 5881744 | 0 | 45528 | 166574.7 | 21.74 | 712.07 |
| 发表文章数 | 1629 | 0 | 37 | 89.5 | 8.95 | 121.60 |
| 参与公共编辑数 | 111282 | 0 | 160 | 2442.5 | 44.13 | 2002.84 |

表2.2展示了主讲人用户群体问答行为的基本情况，从各个变量的最大值和最小值来看，中间跨度很大，可见各个主讲人用户之间还是存在较大差异的，因此有必要对知识供给者进行分组；从均值来看，所爬取数据样本的平均粉丝人数为33912，高于一般水平，其中最主要的原因是本研究爬取的都是主持过live的群体信息，在一定程度上可以帮助累计他们的社会资本，因此他们在社交网络中的影响力要大于大多数人，平均水平也就相对较高。

## 2.4 知乎live现状分析

将各个领域下的live数量和平均参与人数进行统计，得到图2.2：

图2.2 不同领域下live个数及平均参与人数分布

从live个数来看，数据样本中有关教育和职业领域的live数量较多，分别占了16.40%和12.76%，但是平均参与人数处于较低水平，可见社交网络也是传播教育理念的一个重要途径；而随着当代社会竞争压力的加大，很多用户也喜欢在知乎上分享自己的职业生涯体会；互联网领域的live数和平均参与人数均处于较高的水平，这很大程度上也是目前社会发展带来的效应；值得注意的是阅读与写作领域，尽管该领域的live数量处于中下水平，但是平均参与人数是远远高于其他领域，可见该领域主讲人用户的社会资本条件非常强大，对阅读写作方面有深刻的间接，能吸引广泛用户群体为其付费。

根据词频统计，表2.3列出了各个领域下排名前5的关键词：

表2.3 词频统计结果

|  |  |
| --- | --- |
| 领域 | 排名前5的关键词 |
| 教育 | 考研、英语、留学、高考、专业 |
| 职业 | 职场、面试、工作、职业、求职 |
| 互联网 | 数据、产品经理、运营、创业、程序员 |
| 金融经济 | 投资、入门、财务、交易、理财 |
| 生活方式 | 摄影、选购、减肥、攻略、装修 |
| 音影游 | 音乐、电影、设计、摄影、钢琴 |
| 科学技术 | 数据、人工智能、宇宙、工程师、机器人 |
| 医学健康 | 治疗、健康、牙齿、误区、预防 |
| 艺术 | 摄影、艺术、欣赏、建筑、技巧 |
| 阅读写作 | 阅读、故事、写作、通识、文学 |
| 心理学 | 孩子、心理学、沟通、应对、焦虑 |
| 设计 | 设计、建筑、装修、PPT、体验 |
| 体育 | 健身、游泳、运动、NBA、体育 |
| 法律 | 法律、律师、离婚、买房、法官 |
| 商业 | 创业、营销、投资、商业、谈判 |
| 旅行 | 旅行、日本、攻略、旅游、规划 |
| 美食 | 西餐、课程、美食、厨房、地道 |

表2.3的词频统计结果也反映了当代人生活的状态，例如在教育领域，考研、留学出现的频率较高，近年来大学生中选择继续升造的比例大幅度上升，对于这方面的需求也会较大；互联网领域中，技术、产品和运营作为三大主力岗位，有关这方面的live数量也占据较大比例；在心理学领域中，孩子出现的频率最高，可见现在大多数父母都很注重于孩子的沟通，同时焦虑也出现在前五的词频中，说明当今生活节奏的加快使得越来越多的人面临压力大的问题。

## 2.5 相关性分析

表2.4 相关性分析结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 参与人数 | 粉丝数 | 关注人数 | 被收藏数 | 赞同数 | 文章数量 | 参与公共编辑数 |
| 参与人数 | 1.00 | 0.54 | 0.15 | 0.40 | 0.37 | 0.21 | 0.13 |
| 粉丝数 | 0.54 | 1.00 | 0.46 | 0.87 | 0.88 | 0.57 | 0.57 |
| 关注人数 | 0.15 | 0.46 | 1.00 | 0.50 | 0.51 | 0.35 | 0.53 |
| 被收藏数 | 0.40 | 0.87 | 0.50 | 1.00 | 0.53 | 0.63 | 0.66 |
| 赞同数 | 0.37 | 0.88 | 0.51 | 0.53 | 1.00 | 0.60 | 0.71 |
| 文章数量 | 0.21 | 0.57 | 0.35 | 0.63 | 0.60 | 1.00 | 0.59 |
| 参与公共编辑数 | 0.13 | 0.57 | 0.53 | 0.66 | 0.71 | 0.59 | 1.00 |

分析结果显示主讲人用户粉丝数、被收藏数量和赞同数与平均参与人数存在一定的相关性，而关注人数、文章数量和回答问题数量与因变量的关系偏弱，同时可以发现自变量之间存在高度相关性，可见主讲人用户一个社会资本维度的提升，会带动其他两个维度。同时如果后续建立传统的线性回归研究参与度差异，会出现多重共线性问题，故采用lasso回归进行删选变量。

# 3 主讲人用户群体分组

## 3.1 K-Means聚类分析算法介绍

### 3.1.1 K-Means聚类分析原理

聚类分析是一种无监督的算法，它基于点与点距离的相似性，来对样本进行分类。K-means是聚类中最常用的方法之一，它主要通过不断的迭代最终收敛到中心点到组内距离最小。

K-Means算法需设定一个初始值k，即表示需要将现有数据集分成k个组，不论这种分类是否合理，或者是否有意义。算法需要最小化平方误差：

 （3.1）

其中x是每个组Ci的均值向量，其中代表每个样本点到均值点的距离。

距离度量最常用的就是闵可夫斯基距离（亦即p范数），即

 （3.2）

当p=2的时候，该距离即为欧氏距离（2范数）；当p=1的时候，该距离即为曼哈顿距离。

K-Means聚类的基本过程为：

（1）那个样本中随机地抽取k个点作为每个组的中心，并计算初始的均值向量{μ1,μ2,⋯,μk}

（2）循环以下几步直到达到停止条件（收敛条件）：

I 令Ci=∅(1≤i≤k)；

II 计算每一个样本点到k个均值向量的距离，取其中最短的距离，即将样本点汇入到与其最近的组；

III 对每一个簇重新计算他们新的均值向量，并与原始的均值向量进行比较，如果有变化就更新，将其作为新的簇中心，如果没有变化则保持原来的均值向量，然后继续重复上述步骤。

### 3.1.2 评价聚类的指标

（1）inertias：是K均值模型对象的属性，他表示在没有真实的分类结果下，样本到最近的聚类中心的距离总和。该值越小说明样本在类间的分布越集中。

（2）兰德指数：通过比较每个样本的实际类别信息和聚类标签结果，a表示两者类别相同的元素对数，b表示在不同类别的元素对数，则兰德指数为：

 （3.3）

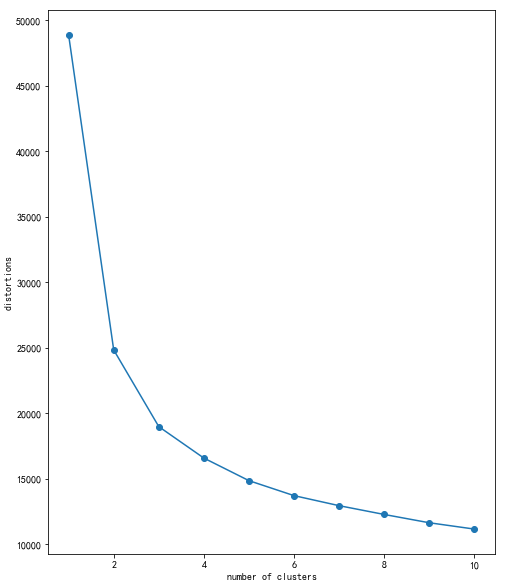
其中为数据集中可以组成的总元素对数。

在该值的取值范围在0~1之间，值越大说明最终聚类的标签结果与实际情况越接近。

（3）同质化得分（Homogeneity）：如果所有的聚类都只包含属于单个类的成员的数据点，则聚类结果满足同质性。取值范围[0,1]，值越大意味着聚类结果与真实情况越符合[17]。

## 3.2 分组主讲人用户基本特征描述

为了更直观地分析在不同社会资本特征条件下，用户参与live人数的驱动因素差异，首先对主讲人用户进行分组，其中用到的特征工程为6个自变量，并采用log处理后的数据。本文采取K-Means算法，首先基于簇内误差平方和，使用肘方法确定簇的最佳数量。肘方法的基本理念就是找出聚类偏差骤增是的k值，通过画出不同k值对应的聚类偏差图[18]，具体如图3.1所示：

图3.1 聚类偏差图

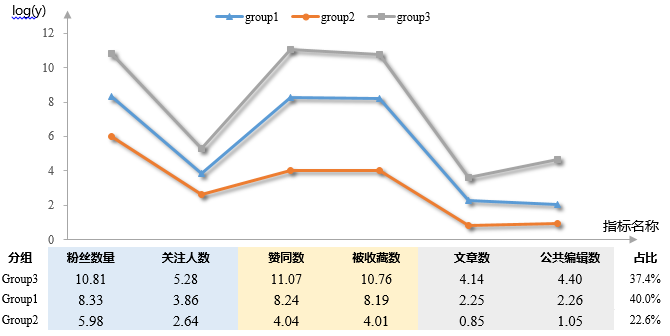
从图3.1中可以得到，在3个聚类类别之后，下降趋势相对变得缓慢，故将主讲人用户样本分成3个组比较合适。对分离出来的3组样本进行描述性统计分析，根据每组样本指标的层中心绘制出3个类别主讲人用户的基本特征描述图，观察内部社会资本特征差异。具体如图3.2所示：

图3.2 三组主讲人用户基本特征描述

总体来看，三组用户在粉丝数量、赞同数、文章数等指标均有较明显差异，且结构维度水平较高的用户群体，其他两个维度水平也会较高；同时在认知维度上，三组用户的差异最为明显，认知维度指的是个体交流之间语言通用度和价值观相似度，即双方之间要产生共鸣，而共同语言的形成更多需要的是彼此之间有相似的专业背景，因此他对主讲人用户的技能专业等方面要求更高，也更容易将三组用户区分开来。

其中，第三组用户整体的社会资本水平较高，人数占比为37.4%，说明该组主讲人群体有相对庞大的用户群体网络，并能通过回答问题，发表文章等方式提升自己的价值，同时获得他人的认可程度也较高，可称为“专家型用户”；第一组用户整体社会资本情况处于中间水平，可称为“中庸型用户”；最后二组用户回答问题能力相对偏弱，获得别人认可程度较低，猜测这类群体更多的是以检索信息为主，故称为“信息搜索型用户”。

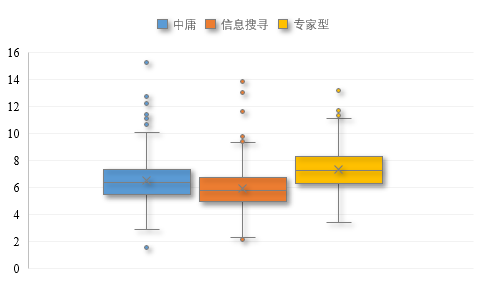
对每一组主讲人用户下的参与人数绘制箱型图，如图3.3所示：

图3.3 分组箱型图

从图3.3中可以得到，在取对数后，三组用户参与人数波动程度相当，且都没有呈现明显的偏态分布，其中“专家型”用户的中位数水平都要明显高于其他两组供给者用户，初步说明了用户累计的社会资本越多，越能激发其他人知识付费行为。

# 4 用户参与度驱动因素差异性分析

## 4.1 lasso回归介绍

设随机变量*y*与变量*x*1，*x*2，…，*x*p的线性回归模型为：

 （4.1）

如果有*n*组观察数据（*x*i1,*x*i2,…*x*ip;*y*i）（*i*=1,2..,*n*），传统的回归模型是用最小二乘法求解系数：

 （4.2）

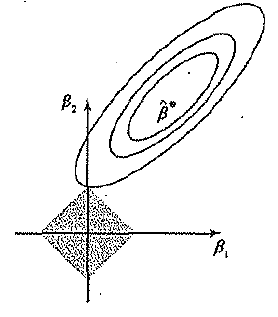
在建立模型之初，通常会建立全模型，避免缺少重要的变量。然而变量之间会存在多重共线性等问题，因此需要对变量进行筛选。在建模过程中会有一系列的指标，例如R2、AUC等，通过这些指标来判定变量是否应该加入到模型中，最终找到对因变量解释性最强的自变量组合。如今也有很多模型和算法能对变量进行选择，例如逐步回归、岭回归等。而Lasso算法则是一种能够实现指标集合精简的估计方法[19]。

Lasso方法是一种压缩估计。它通过构造一个惩罚函数，压缩系数，同时会设定一些不重要的变量系数为0，进而得到一个较为精炼的模型。因此该方法保留了子集收缩的优点，可以很好地解决多重共线性问题，但系数是为有偏估计，即：

 （4.3）

其中约束条件为：。

可见Lasso的基本思想是设定一个常数阈值s，并且在回归系数的绝对值之和小于该值的约束条件下，使残差平方和最小化，进而将一些回归系数压缩至0，最终得到可以解释的模型[20]。

图4.1 变量选择特征

如图4.1所示，菱形代表lasso中的限制区域，如果s足够大，那么限制区域将包含，并且系数将和最小二乘一致。以为中心的每个椭圆代表了某一个 RSS数值，也就是说，在椭圆边上的每个点所代表的RSS是相同的，当椭圆与最小二乘系数估计相差越来越大，RSS也逐渐增大。lasso系数估计就是条件区域（灰色区域）与椭圆的相交点，而当椭圆在坐标轴上与条件区域相交时，其中一个系数就会为零[21]。

在选择的时候，通常采取*k*拆交叉验证法，先将数据随机地分成大小基本相同的k个组，选择k-1组数据作为训练集拟合模型，其次将拟合得到的模型去预测剩下的1组数据，得到均方误差MSE1。将这个步骤重复k次，每一次把不同的观测组作为验证集，最终会计算得到k个均方误差，MSEl，MSE2，…MSEk，k折CV估计由这些值求平均计算得到：

 （4.4）

## 4.2 分组散点图

图4.2为每个知识供给者用户群体下，各个自变量与平均参与人数的散点图：



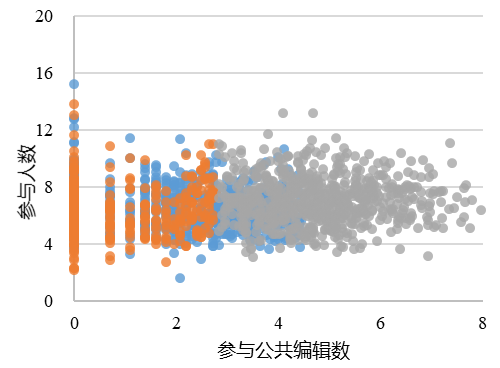
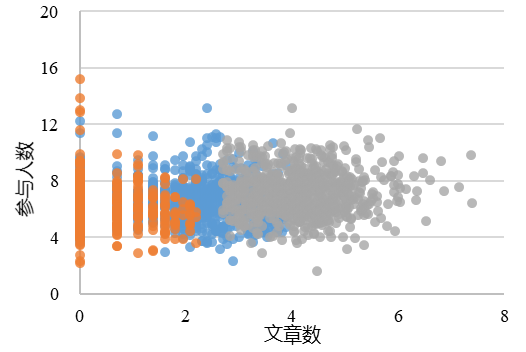
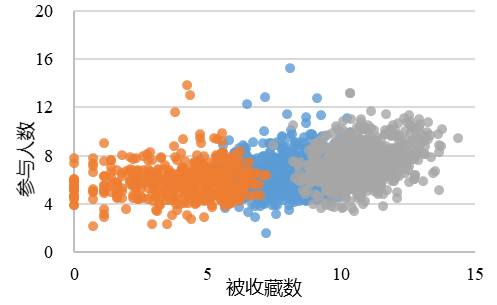
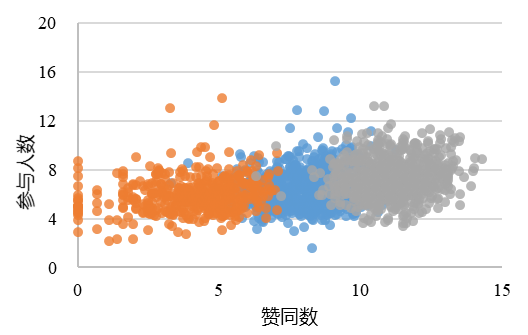
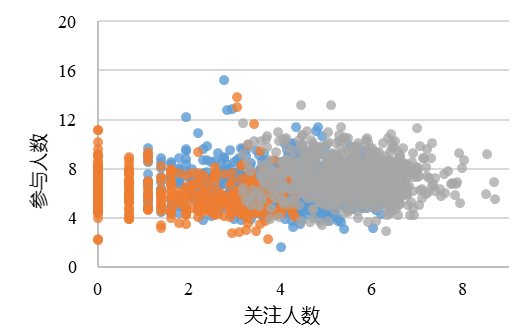
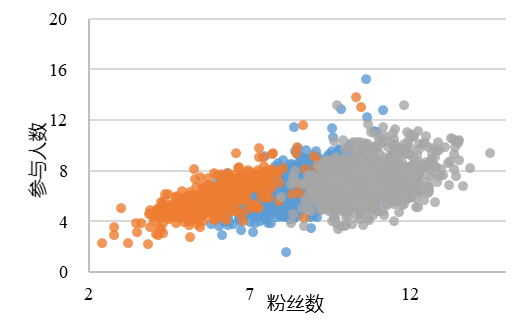


图4.2 分组散点图

从分组散点图中可以看到，供给者的粉丝数、赞同数和被收藏数与用户的参与度有较为明显的线性关系，其余的散点图分布较为分散。其中粉丝数与参与人数的关系图中，可以发现随着粉丝数的增长，参与人数增长趋势变得越来越缓慢，及影响效应是递减的；在被收藏数与参与人数散点图中，“专家型”用户组的增长趋势明显要大于其他两组，可见被收藏数在该组的影响效应最大。

## 4.3 模型结果分析

因为数据的波动幅度较大，故对变量对数处理后进行建模，并对每一份数据样本采用交叉验证选择Lasso回归最优正则化项参数，表4.1给出了最终的模型估计结果：

表4.1 模型系数估计结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量 | 中庸组 | 专家 | 信息搜寻 | all |
| 截距 | 2.276 | 0.921 | 3.368 | 2.550 |
| 粉丝数量 | 0.650 | 0.584 | 0.955 | 0.735 |
| 关注人数 |  | 0.027 |  |  |
| 赞同数 | 0.193 | 0.203 | 0.105 | 0.201 |
| 被收藏数量 | 0.190 | 0.244 | 0.109 | 0.215 |
| 文章数量 | -0.009 | 0.037 |  | 0.056 |
| 公共编辑数 |  | 0.034 |  | 0.029 |

首先从总体来看，各个用户群体中最终保留的变量是有差异的，其中“中庸型”主讲人用户样本剔除了关注人数和公共编辑数变量，而在和“信息搜索型”用户中，剔除了关注人数、参与公共编辑数和文章数量这三个变量；如果没有对主讲人用户进行分组，剔除的则是关注人数。接下来分析在三个社会资本维度下各个变量对用户参与度的影响差异：

**（1）结构维度对用户参与度的影响差异**

在结构维度中可以明显看到，知乎live主讲人用户“关注人数”并未对用户参与知识付费动机产生显著影响，其中尽管在“专家型用户”中，该变量没有被剔除，但是系数很小，可见，关注他人虽然可累计自己的社会资源，还是难以对用户参与live决策产生影响；“粉丝数量”则对用户参与度起到了积极的促进作用，而获得他人认可度是促使用户开展付费交易的一个重要因素，可见知识供给者拥有的粉丝数量与获得其他用户认可度的可能性之间存在正向关系，因为粉丝数量在很大程度上可反映一个人的知名程度，数量越多，越容易刺激他人开展付费交易。供给者关注的人数更多地是从自身角度，期望从其他用户那里获得知识和信息，因此该变量无法直接体现出个人的知识储备，进而很难对他人付费决策行为产生显著影响。

从三个知识供给者用户群体分开来看，在“专家型用户”群体中，“粉丝数量”对用户知乎live参与度的影响要小于其他两个用户群体，系数为0.584。在上述分析中我们可以发现，“专家型用户”的粉丝数量是处于最高水平的，可见知识供给者用户从粉丝数量每增长单位中所吸引的付费用户是越来越少的，这也符合经济学上的边际效用理论。

**（2）认知维度对用户参与度的影响差异**

在不对知识供给者用户群体进行分组的情况下，获得的赞同数和被收藏数对参与live人数影响系数分别为0.201、0.215，且对三组主讲人用户下的平均参与live人数均有正向作用。对于专业的社会问答平台，用户往往是有针对性地搜寻信息，寻找和自己有共同语言的群体，并通过点赞收藏等方式表达自己的认同感。研究结果也说明用户更倾向于选择获得广泛认同的供给者进行付费，从而彼此之间更好地开展知识交流，付费用户也可以保障知识获取质量。

其中可以注意到，认知维度下的两个变量在“专家”用户群体的影响效应要大于其他两组的，在中间阶段，主讲人群体一方面在知乎上搜寻信息，另一方面自己也会参与知识共享，随着时间的推移，这部分用户信息内化程度是最高的，基于互惠利他等动机，他们提供知识的水平程度会大幅度提高，获得他人赞同的可能性也越大，因此更容易驱动他们付费行为。

**（3）关系维度对用户参与度的影响差异**

关系维度下的变量影响效应要小于其他变量，且在“信息搜寻”用户中，剔除了“文章数量”和“公共编辑数”这两个变量，从图4.2的散点图中也可以看到，该组用户的散点图分布较为分散，难以找到规律；在“中庸型”供给者用户中，“文章数量”前面的系数为负，为-0.009，并剔除了“公共编辑数”和“关注人数”这两个变量。用户在选择付费对象时，供给者的专业能力是付费行为决策过程中考量的一个重要因素。在同等条件下，供给者的专业知识背景越强，公众对其的感知信任程度就会越高，即公众相信自己所花费的金钱成本能换来高质量的知识服务；而供给者的专业能力可以通过其发表文章等方式体现，对于“中庸”供给者用户群体来说，随着他们关注的范围变大，精力会受限，导致他们对于知识的了解程度也会较浅，在能力受限的情况下，他们发表的文章质量没有那么高，当他人期望落空时，产生付费行为动机的可能性就会减小。而当供给者达到专家型水平时，他个人的知识和专业技能可以通过他发表的文章和参与公共编辑等方式体现出来，因此他们发表文章的数量越多，就越会激发他人付费行为。

# 5 结论建议

## 5.1 结论

本文基于社会化问答平台上不同群体的知识供给者社会资本特征，探索了付费用户参与影响因素差异，得到以下结论：

**（1）从知乎live现状来看，各个主题领域下live主题数和平均参与人数存在明显差异。**其中教育领域的live数量最多，且考研在该领域的词频统计中占据榜首；阅读与写作领域平均参与人数最多，达到了18892，远远高于其他领域；

**（2）知识供给者累计的社会资本对付费用户的参与度有正向作用。**根据社会资本下的三个维度特征，将知识供给者分为了“专家型”、“中庸型”用户和“信息搜寻性”，其中三组用户的特征如下：I “专家型”群体整体累计的社会资本最多；II 在认知维度上，三组用户的差异最为明显；III “专家型”群体下，付费用户的参与度要明显高于其他两组群体。可见知识供给制者累计的社会资本越多，越能获得他人的信赖，进而激发他人知识付费行为，同时关系维度下的两个变量的影响效应要小于其余的两个维度。

**（3）结构维度中的粉丝数量对付费用户参与度的影响效应为：“信息搜寻型”>“中庸型”>“专家型”。**根据lasso回归系数显示，供给者的关注人数没有对用户付费行为起到明显促进作用，但是粉丝数目在三个群体的影响系数分别为0.955，0.650，0.584，即粉丝数量越多，带来付费用户增长效应是递减的。

**（4）认知维度下的赞同数和被收藏数在“专家”用户群体的影响效应要大于其他两组。**这部分知识供给者群体整体专业水平是最高的，因此他们提供知识的水平程度会高于其他两组，获得他人赞同的可能性也越大，因此更容易驱动用户付费行为。

**（5）关系维度下的文章数和参与公共编辑数影响效应最弱，且只在“专家型”用户中有较小的促进作用。**在“信息搜寻”用户中，剔除了“文章数量”和“参与公共编辑”这两个变量；在“中庸型”供给者用户中，“文章数量”前面的系数为负，为-0.003，并剔除了“公共编辑数”这个变量。而关系维度可以体现一个人的专业水平和能力，只有当知识供给者的水平达到一定程度，符合他人期望时，文章数和参与编辑数对付费用户的参与度才能产生一定正向影响作用。

## 5.2 建议

根据上述结论，本文提出以下建议：

### 5.2.1 针对知识供给者

（1）知识供给者应充分发挥社交问答平台的优势，并提供高质量的知识服务。本文的研究结论发现关系维度对用户参与度的促进作用没有那么明显，可能原因是知识供给者发表的文章没有得到其他用户认可。因此，对于知识供给者而言，为了尽可能体现其知识水平和专业技能，不仅要保证发表文章的数量，更要注重其质量。同时对于“信息搜寻型”知识供给者群体来说，其“粉丝数量”的效应是最为明显的，应在知乎平台上拓宽人际圈、通过回答问题等方式维持与他人的联系，开始多元化的知识交流，吸引他人关注，从而提高自己在社交领域的影响力，进而产生更多的经济效益。

（2）知识供给者应注重认知型社会资本的积累，积极获取他人信任。研究结论发现三组用户在认知型社会资本上差异作为明显，且其在“专家型”群体中影响效应最为明显。因此用户在发表文章等行为的同时，也可从多渠道共享自己的知识，以尽可能获得他人认可和收藏，并标注自己擅长的领域，让他人能全方面了解自己，进而找到专业、知识水平上志同道合的对象，开展交流和知识付费交易。

### 5.2.2 针对平台

（1）牵手“专家型用户”内容生产者，在建立多元化内容前提下，强化知识供求两方的社交关系。本研究发现知识供给者累计的社会资本越多，付费用户的参与度会越高。而“专家型”用户在知乎平台上具有较强的影响力，一方面，这部分群体知识共享内容能够得到保证，用户对其也比较信任；另一方面，他们能为平台带来更多的流量。因此充分发挥社交平台先天性的优势，将“专家型用户”与其他用户群体连接起来，建立多元化的社交关系，维持付费用户的活跃度。

（2）对付费平台上知识供给者建立完善的评价体系，科学评估其专业能力和主讲内容。从上述研究结论可以发现，尽管有些主题领域下live数量很多，但是平均参与人数却处于较低水平。为了提高用户付费行为意愿，应尽可能地较低用户的搜索成本，精准地推荐给他们真正想付费的知识产品。因此平台可基于知识供给者的用户行为数据，合理量化其社会资本，并构建完善的指标体系和排序策略；而其他用户可根据自己的需求，选择相应的测度指标，平台可为其精准推荐符合要求的知识供给者，以进一步提高知识付费服务效率。

# 参考文献

[1] 李武.在线知识付费平台:何去何从?[J].图书情报知识,2018,40(3):1-1.

[2] 张杨燚.在线平台知识付费研究综述[J].情报探索,2018(08):129-134.

[3] 罗星.知识付费平台：内容价值还是社群经济[D].安徽大学,2018.

[4] Dou W.Will Internet Users Pay for Online Content?[J] Journal of Advertising Research,2014,44(4):349-359

[5] 韩芳芳.付费网络问答平台使用意愿影响因素研究[D].华中师范大学,2017.

[6] 王大庆.在线知识付费产品用户付费意愿的影响因素研究[D].武汉大学,2018.

[7] 胡盼弟.长尾理论视域下知识付费类自媒体经营策略探究[D].山东大学,2018.

[8] 王佩文.中国互联网知识付费产业商业模式分析与设计[D].台湾大学,2018.

[9] 王镇.社会资本——关于社会结构与行动的理论[J].青年与社会(上),2013.44(7):77-77.

[10] 李东旭“社会资本”概念的缘起与界定[J].学术交流,2012.15(8):124-126.

[11] Chung N, Nam K，Koo C. Examining Information Sharing in Social Networking Communities: Applying Theories of Social Capital and Attachment[J]. Telematics&Informatics,2016,33 (1):77-91.

[12] 张璐,张鹏翼.线上线下社会资本与社会化问答行为的关系研究——以知乎医学和健康话题为例[J].图书情报工作,2017,61(17):84-90.

[13] L1 G, YANG X, 11UANG S. L,ffects of social capital end community support on online community members' intention to create user generated content[J].Journal of electronic commerce research,2014,15(3):190-199.

[14] 毕莹莹.微博用户的社会资本度量方法[D].华南理工大学,2015.

[15] Chiu C M H, Meng H W, Eric T G. Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capitaland Social Cognitive Theories[J]. Decision Support Systems,2007,42(3):1872-1888.

[16] 王贇芝.自媒体用户信息内容消费意愿影响因素研究[D].安徽大学,2017.

[17] 龙泽海,杨毅.基于lasso方法的银行对中小企业贷款供给意愿研究[J].金融与经济,2017,14(3):58-65.

[18] 曾五一.统计学导论[M].武汉:华中师范大学出版社,2014.8 .

# 附录A 相关python代码

import pandas as pd

import numpy as np

import wordcloud

import math

from sklearn.cluster import KMeans

from sklearn import preprocessing

from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt

from scipy import stats

import matplotlib

from sklearn.linear\_model import Lasso,LassoCV

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

from sklearn.metrics import r2\_score

from sklearn.linear\_model import LinearRegression

from statsmodels.stats.anova import anova\_lm

from statsmodels.formula.api import ols

import statsmodels.formula.api as smf

import seaborn as sns

import jieba

from wordcloud import WordCloud

from collections import Counter

data = pd.read\_csv("alldata.csv",encoding="gbk")

data\_group=data.groupby(['user\_id'])[['taken','answer\_count','articles\_count', 'columns\_count','favorite\_count','favorited\_count','follower\_count', 'following\_columns\_count','following\_count','following\_favlists\_count', 'following\_question\_count','following\_topic\_count','hosted\_live\_count','logs\_count','marked\_answers\_count','participated\_live\_count','pins\_count','question\_count', 'thanked\_count',\'voteup\_count']].mean().reset\_index(drop=False)

# 数据预处理（选择列，以及剔除异常值）

## 添加一列，知乎收录的回答数/回答问题数量

data\_group['ratio']=data\_group['marked\_answers\_count']/data\_group['answer\_count']

data\_group['ratio\_type']=data\_group['ratio'].apply(lambda x : 'low' if x==0 or pd.isnull(x) else ('medium' if x<=0.15 and x>0 else 'high'))

data\_group['column\_type']=data\_group['columns\_count'].apply(lambda x : 'low' if x==0 else ('medium' if x<=2 and x>0 else 'high'))

data\_group.groupby(['column\_type']).count() #计数

## 选择列(简单平均)

columns=['user\_id','taken','follower\_count','following\_count','voteup\_count','favorited\_count','articles\_count','logs\_count','column\_type']

data\_group1 = data\_group[columns]

## 加权计算

data\_group2=data.groupby(['user\_id','short\_name'],as\_index=False).agg({'taken':['sum','count']})

data\_group2.columns = ['user\_id','short\_name','taken','count']

data\_group2 = data\_group2.groupby(['user\_id']).apply(lambda x :np.average(x['taken'],\

weights=x['count'])).reset\_index(drop=False).\

rename(columns={0:'taken'})

data\_group2 = pd.merge(data\_group2,data\_group.drop(['taken'],axis=1),on=['user\_id'])

data\_group2 =data\_group2[columns]

# 描述性统计分析

## 词频统计

count=pd.DataFrame()

for i in set(data['short\_name']):

subject=data[data['short\_name']==i].reset\_index(drop=True)['subject']

su=subject[0]

print(i)

for j in subject.index:

su=su+subject[j]

su=su.replace('如何','')

su\_cut = " ".join(jieba.cut(su))

dd=pd.DataFrame(dict(Counter(su\_cut.split(' '))),index=['count']).T.reset\_index(drop=False)

dd['shot\_name']=i

count=count.append(dd)

count=count.sort\_values(by=['shot\_name','count'],ascending=False,)

## 主讲人用户群体分布

de = data\_group2.describe()

stats.skew(data\_group2.drop(['user\_id'],axis=1)) #偏度

stats.kurtosis(data\_group2.drop(['user\_id'],axis=1)) #偏度

## 去除ratio为空的情况 data\_group1 = data\_group1.dropna()

## 每一列和每一行有多少0

(data\_group1==0).sum(axis=0)

pd.value\_counts((data\_group2==0).sum(axis=1))

## 剔除每一行中0的值大于4的记录

data\_group3 = data\_group2[(data\_group2==0).sum(axis=1)<=4].reset\_index(drop=True)

# 数据预处理（标准化数据，log处理）

data\_scaled1 = data\_group3.drop(['user\_id','column\_type'],axis=1).applymap(lambda x :math.log(x+1) if x==0 else math.log(x))

data\_scaled2 = data\_scaled1.drop(['taken'],axis=1)

# 聚类

## 确定最佳聚类个数

d=[]

for i in range(1,11):

km=KMeans(n\_clusters=i,init='k-means++',n\_init=10,max\_iter=300,random\_state=0)

km.fit(data\_scaled2)

d.append(km.inertia\_) #inertia簇内误差平方和

plt.plot(range(1,11),d,marker='o')

plt.xlabel('number of clusters')

plt.ylabel('distortions')

plt.show()

## 确定聚类个数为3

interation=500

kmodel=KMeans(n\_clusters=3) #n\_job等于并行数，一般等于cpu比较好

kmodel.fit(data\_scaled2)

r1=pd.Series(kmodel.labels\_).value\_counts()

r\_data1=pd.DataFrame(kmodel.cluster\_centers\_) #层中心

r\_data1.columns = ['粉丝数量','关注人数','赞同数','被收藏数','文章数','参与公共编辑数']

finaldata1=pd.DataFrame({'group':kmodel.labels\_.tolist(),

'user\_id':data\_group3['user\_id'].tolist(),

'type':data\_group3['column\_type'].tolist()})

finaldata2=pd.merge(finaldata1,data\_scaled1,left\_index=True,right\_index=True)

# 建模

## 分组数据

def dummies(num):

group = finaldata2.loc[finaldata2['group']==num].drop(['group'],axis=1)

group = pd.merge(finaldata2.loc[finaldata2['group']==num].drop(['group'],axis=1),\

pd.get\_dummies(group['type'], prefix='type'),left\_index=True,right\_index=True)

return group

group1 = dummies(0) #中

group2 = dummies(1) #低

group3 = dummies(2) #高

finaldata2 = pd.merge(finaldata2, pd.get\_dummies(finaldata2['type'], prefix='type'),left\_index=True,right\_index=True)

## 每一组的相关性分析

a1=group1[columns[1:8]].corr()

a2=group2[columns[1:8]].corr()

a3=group3[columns[1:8]].corr()

a=data\_scaled1.corr()

## 此处 alpha 为通常值 #fit 把数据套进模型里跑,通过10折交叉验证选出最佳alpha

alphas = 10\*\*np.linspace(-5, 10, 500)

dropcolumns = ['taken','user\_id','type','type\_high','type\_low','type\_medium']

indexnames=['intercept','follower\_count','following\_count','voteup\_count','articles\_count','logs\_count']

def model(data,dropcolumn,indexname,renamecolumn):

lasso=LassoCV(alphas = alphas,cv=10).fit(data.drop(dropcolumn,axis=1), data['taken'])

model\_lasso=Lasso(alpha=lasso.alpha\_).fit(data.drop(dropcolumn,axis=1),data['taken'])

coef=pd.DataFrame([model\_lasso.intercept\_]+model\_lasso.coef\_.tolist(),index=indexname).rename(columns={0:renamecolumn})

return model\_lasso,coef

model\_lasso1 = model(group1, dropcolumns, indexnames, 'group1')[0]

coef1 = model(group1, dropcolumns, indexnames, 'group1')[1]

model\_lasso2 = model(group2, dropcolumns, indexnames, 'group2')[0]

coef2 = model(group1, dropcolumns, indexnames, 'group2')[1]

model\_lasso3 = model(group3, dropcolumns, indexnames, 'group3')[0]

coef3 = model(group1, dropcolumns, indexnames, 'group3')[1]

model\_lasso = model(finaldata2, dropcolumns+['group'], indexnames,'all')[0]

coef = model(finaldata2, dropcolumns+['group'], indexnames,'all')[1]

tot\_coef=pd.DataFrame({'变量':['截距','粉丝数量','关注人数','赞同数','文章数量','参与公共编辑数'],

'group1':[model\_lasso1.intercept\_]+model\_lasso1.coef\_.tolist(),\

'group2':[model\_lasso2.intercept\_]+model\_lasso2.coef\_.tolist(),\

'group3':[model\_lasso3.intercept\_]+model\_lasso3.coef\_.tolist(),\

'all':[model\_lasso.intercept\_]+model\_lasso.coef\_.tolist()})