Hans Journal of Data Mining 数据挖掘, 2018, 8(3), 121-130
Published Online July 2018 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2018.83014

Hans汉斯

The Analysis of Impact Factors on the Playback Amount of Himalaya's Paid Boutique

Liqing Guo

Xiamen University Tan Kah Kee College, Zhangzhou Fujian
Email: galinaguo@xujc.com

Received: Jul. 3rd, 2018; accepted: Jul. 16th, 2018; published: Jul. 23rd, 2018

Abstract

With the growth of GDP and the upgrade of consumption structure, various types of knowledge payment platforms have emerged. This paper selects the Himalayan knowledge payment platform with the highest number of launches per capita as the research object, and crawls the data of its paid boutique album courses in August 2017. The statistical analysis explores how the amount of play of courses changes under the impact of course category, pricing model, duration of courses, number of likes etc., establishes a corresponding regression model, and attempts to create a popular album. The data analysis shows that the market demand for education training and children's course categories is relatively high. The album name is mainly based on new ideas, and the number of likes, comments and tags is positively related to the number of play of courses. In the logarithmic linear regression model, the emotional life is 46.8% higher than the audio book with respect to the variable of album category.

Keywords

Knowledge Payment, Statistical Analysis, Regression Model

喜马拉雅付费精品播放量影响因素分析

郭丽清

厦门大学嘉庚学院，福建 漳州
Email: galinaguo@xujc.com

收稿日期：2018年7月3日；录用日期：2018年7月16日；发布日期：2018年7月23日

文章引用：郭丽清. 喜马拉雅付费精品播放量影响因素分析[J]. 数据挖掘, 2018, 8(3): 121-130.
DOI: 10.12677/hjdm.2018.83014

摘　要

随着GDP增长，消费结构升级，各类知识付费平台应运而生。本文选取人均启动次数最多的喜马拉雅知识付费平台为研究对象，爬取其2017年8月付费精品专辑课程的数据，统计分析探究课程类别、定价模式、课程时长、点赞次数等因素对课程播放量的影响，建立相应的回归模型，并尝试打造一个爆款专辑。数据分析表明，教育培训和儿童的课程类别市场需求度比较高，专辑起名主要靠新意，点赞数、评论数、标签个数与课程的播放量呈正相关。在对数线性回归模型中，相对于专辑类别这一变量，情感生活比有声书高46.8%。

关键词

知识付费，统计分析，回归模型

1. 引言

2016 年被称为“知识付费元年”。这与近几年来国内互联网技术、移动终端的发展普及，人均 GDP 增长、人们消费结构升级密切相关。以支付宝、微信支付为代表的移动支付工具对线上线下支付渠道的持续渗透，为知识付费提供便利；互联网信息爆炸，真正有价值信息的稀缺；中产阶级对消费品质的追求，对高质量的服务付费意愿不断增长。在这多重因素作用下，促发了知识付费的兴起。本轮知识付费的浪潮主要指有种有趣有料的知识网红将自己的专业技能，通过第三方平台出售给有相应需求的人，实现知识变现的过程[1] [2]。具体知识付费商业模式如下图 1 所示。

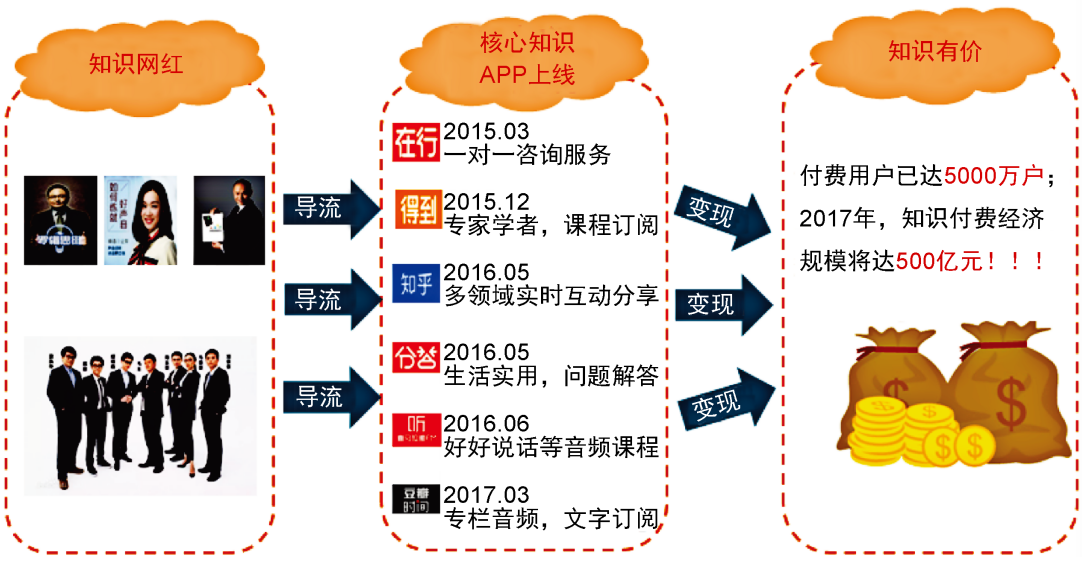

Figure 1. Business model of knowledge-paid
图 1. 知识付费商业模式

国内核心知识付费类 APP 起源于 2015 年 3 月果壳网推出一对一的咨询服务的“在行”。同年 12 月，罗辑思维团队出品“得到”，提倡碎片化学习方式，让用户在短时间内获得有效的知识。2016 年 4 月，知乎上线基于微信开发的付费产品“值乎”，朋友圈刮开可见完整答案。随后产品不断升级完善，推出了付费问题悬赏形态和付费语音问答形态。2016 年 5 月，由“在行”团队孵化的付费语音“分答”，旨在可以快速找到给自己提供帮助的人，用一分钟时间答疑解惑。2016 年 12 月 3 日，喜马拉雅 FM 发起国内首个内容消费节“123 知识狂欢节”，马东，叶武滨，吴晓波等“知识网红”大咖共同为其发声助威，当天销售额达 5088 万[3]。本次内容消费节号召付费是对知识最好的点赞，是知识付费一个重要的里程碑。进入 2017 年知识付费的趋势没有丝毫减速迹象，3 月豆瓣推出了“豆瓣时间”，腾讯官方表示微信公众号正在加快上线付费订阅中。

阿里应用分发发布的 2017 年 Q2 应用行业报告[4]显示(具体见图 2)，6 大主流知识付费平台中 5 家同比 2017 年 Q1 增长率达 50%以上。其中值乎(知乎 Live)增长率最快达 81%，喜马拉雅周人均启动次数最多为 5.8 次。报告指出，当前付费用户已达 5000 万户，且处于高速增长中，预计 2017 年知识付费经济规模将达 500 亿元。

那么如何开设用户喜欢的专辑，在知识付费的大蛋糕中分得一杯羹呢？具体哪些因素会影响专辑的播放量？课程类型，分享时长，还是定价方式？为了探究这一问题，本文爬取了喜马拉雅网站 2017 年 8 月付费精品专辑课程的相关数据，试图通探究影响专辑播放量的因素，并尝试在喜马拉雅上打造一个爆款专辑。

2. 变量说明

网站的数据丰富多彩，网页之间经常通过多层链接访问。喜马拉雅精品付费专辑(课程)包含两层网页链接，其中第一层是当前专辑信息，链接进入第二层可以获取对应专辑的各详细章节信息。本案例抓取了 1000 个第一层的专辑信息；因技术、时间等因素，并抓取其中 184 个专辑的第二层数据，共 12,826 条章节信息。每条数据包含专辑单位播放量(每个章专辑的章节个数各不相同，归一化处理统一量纲)、专辑名称、喜点、专辑类别等 12 个变量，各个变量的详细信息见表 1。

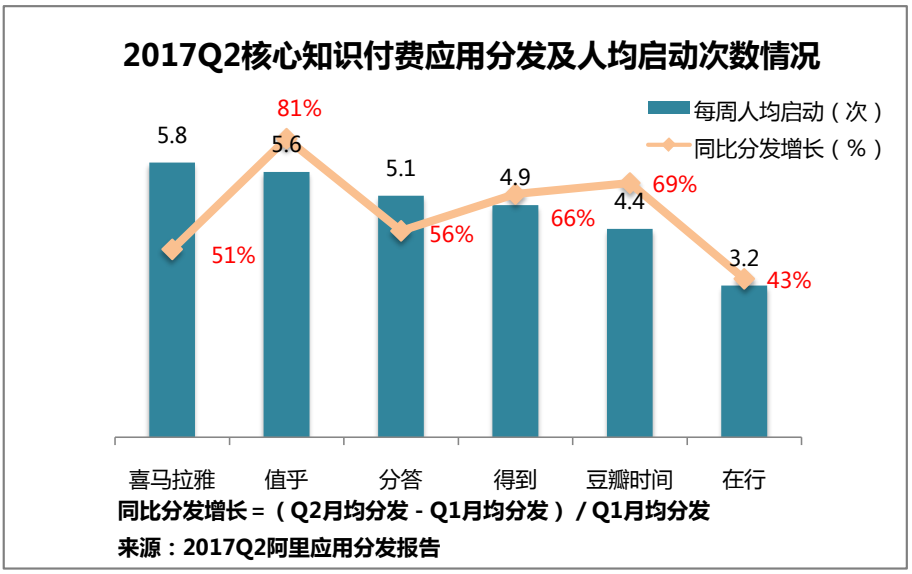

Figure 2. 2017 Q2 core knowledge paid application distribution and per capita startup times
图 2. 2017 Q2 核心知识付费应用分发及人均启动次数

Table 1. Variable description of Himalaya's paid boutique data
表 1. 喜马拉雅付费精品数据变量说明

变量类型		变量名	详细说明	备注
因变量		单位播放量	单位：次	专辑播放总量/当前专辑总章节数
自变量	专辑信息	专辑名称	文本类型	如：好好说话 • 康永来了
		喜点	[7.99~998]	类似 Q 币
		专辑类别	文本类型	如：人文、教育培训等(19 种)
		专辑标签	文本类型	如：好好说话等，其中没有标签的占 51.2%
		播放总量	[22, 4975 万]	
		当前总章节数	[1, 1418]	
	章节信息	章节名称	文本类型	如：026 周玄毅：男生学不会的吵架妙招
		内容时长	时间类型	转换为分钟，如 01:02:03，归一化为 1*60 + 2 + 3/60 = 62.05
		播放量	[0, 740.66 万]	章节播放次数
		点赞次数	[0, 9463]	
		评论次数	[0, 1.02 万]	

影响付费精品专辑播放量的自变量，包括两个部分：专辑信息和章节信息。专辑信息主要是对整体课程的介绍，包括专集名称、喜点(类似游戏中购买装备所花费的 Q 币)、专辑类别、专辑标签、播放总量和当前章节数。章节信息是专辑中每个小节的详细内容，包括章节名称、内容时长、播放量、点赞次数和评论次数。本文后续分析，涉及纯专辑数据，样本为 1000 个专辑信息；涉及到章节信息，样本则为 184 个专辑信息及其对应的 12,826 条章节信息。

3. 描述性分析

1) 单位播放量

不同专辑的章节数略有差异，为了更好刻画专辑的热门程序，本文关注专辑的单位播放量(单位播放量=专辑播放总量/当前专辑总数)。同时为了更好观测数据，我们对播放量取对数。从单位播放量对数柱形图(图 3)看，整体呈类正态偏左分布。具体地单位播放量均值为 9265 次，中位数为 1755 次。这一现象符合我们对日常的基本认知，20%的专辑占据了 80%的播放量，少数具有天量播放量的专辑拉高了播放量的平均水平。

从单位播放量 Top10 的专辑(表 2)看，排名前 10 的单位播放量均超过 15 万次，时间管理达到 24.5 万次，看来对时间的管理非常重视。“教育培训”和“儿童”占前 10 的 6 个席位，这类专辑市场需求度较高。从当前专辑总数看，80%在 100 集以内；喜点定定价最低 12，最高 199。

2) 专辑信息

本案例中将自变量归为两类，一类是专辑信息，包括专辑的名称、定价(喜点)、类型、标签等。为了研究专辑名称对播放量的影响，我们对这个变量进行分词，挑选出高频词根。从图 4 看，大家喜欢使用演播、如何、科学、故事、有声书这类的词语，创业、管理、儿童、亲子也比较受市场喜欢。然而专辑名称中出现高频词汇的个数对播放量呈负相关，名称中出现的高频词汇个数越多，播放量反而相对越少。可能的原因，高频词汇主要是一些常见的中规中矩的描述，无法吸引用户的眼球，起名主要靠创意。

专辑类别包含有声书，儿童，教育培训，人文等 19 种，前 5 种主流类型(占比 78.2%)；其中，教育培训和人文比较受欢迎。箱线图的“胖瘦”代表样本量大小(文中其他箱线图也是同样的效果)，在专辑类

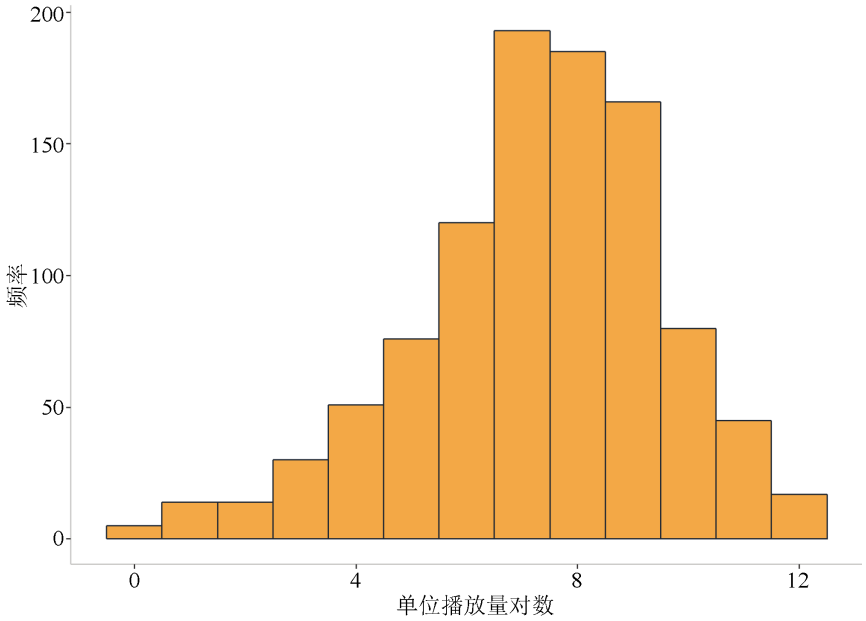

Figure 3. Histogram of logarithmic of unit playback amount
图 3. 单位播放量对数柱形图

Table 2. Top 10 albums of unit playback amount
表 2. 单位播放量 Top10 的专辑

专辑名称	类别	单位播放量	专辑总数	喜点
叶武滨时间管理十堂课——易效能	教育培训	245,000	21	198
舒克贝塔历险记(上部)	儿童	234,248	101	99
央视老炮的戒烟妙招	健康养生	210,650	20	49
余世维家庭教育课	儿童	192,761	88	199
马东携奇葩天团亲授“好好说话”	教育培训	179,603	277	198
蒙曼品最美唐诗	人文	172,111	81	199
《新黑猫警长》第一部：时空奇案	儿童	169,250	12	12
郭德纲超清经典相声集	相声评书	167,000	20	19.9
声音教练徐洁：如何练就好声音	教育培训	157,123	57	118
湖畔三板斧：马云首次公开创业心法	商业财经	151,509	59	99

别中商业财经类专辑数量最多，其次为有声书。有标签的播放量明显高于没有标签。整体呈向上趋势并逐步趋于稳定，2 个以上标签明显高于 1 个。标签个数为 3 播放量均值稍有下降，这可能与样本量相对较少有一定关系。具体如图 5 和图 6 所示。

Figure 4. High-frequency roots of the album name
图 4. 专辑名称包含的高频词根

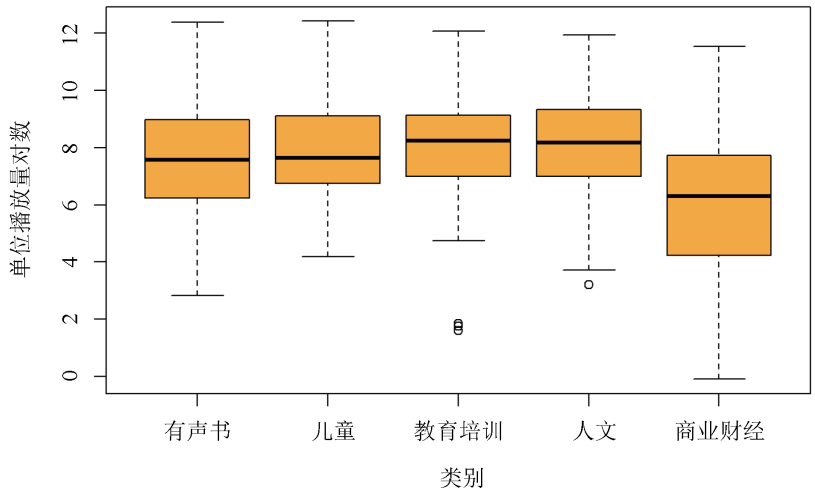

Figure 5. Boxplot of "album category-logarithm of playback amount"
图 5. 专辑类别－播放量对数箱线图

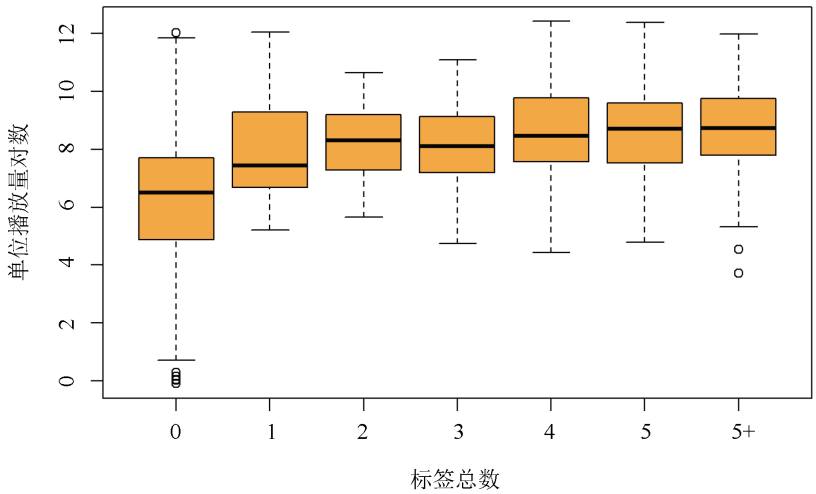

Figure 6. Boxplot of "number of labels-logarithm of playback amount"
图 6. 标签总数－播放量对数箱线图

从定价模式看，专辑定价 84.5%，章节定价 15.5%。专辑定价单位播放量明显高于章节定价，可能原因是打包销售定价较为优惠。播放量随着喜点的个数呈先上升后有所回落。定价 100~150 喜点的专辑单位播放量次数最多。具体如图 7 和图 8 所示。

3) 章节信息

另一类自变量是章节信息，包括章节的内容时长，章节点赞数和评论数等。此处，我们同样对章节播放量取对数处理。

从图 9 的箱线图中可以看出：① 30 分钟以内的章节播放量没有明显差异，15~30 分整体效果最好；30 分钟以上呈下降趋势。② 点赞或评论次数越多，章节播放量越大。

4. 多元线性回归模型

为了更好地量化各因素对专辑播放量的影响，本文建立了专辑单位播放量 y，对单位点赞数、单位评论数、单位内容时长、是否专辑定价、专辑类别、标签总数、专辑名称中的高频词个数 7 个自变量 $(x_1, x_2, \cdots, x_7)$ 的对数－多元线性回归模型。

$$\ln y = \alpha_0 + \sum_{i=1}^{7} \alpha_i x_i + \varepsilon$$

本模型中因变量 y 取播放量的对数，α_0 是常数项，α_i 代表各自变量的回归系数，ε 为随机误差项。回归结果如表 3 所示，其中 P 值小于 0.001，说明存在显著相关的变量；调整 R 方为 0.5004，结果尚可接受。

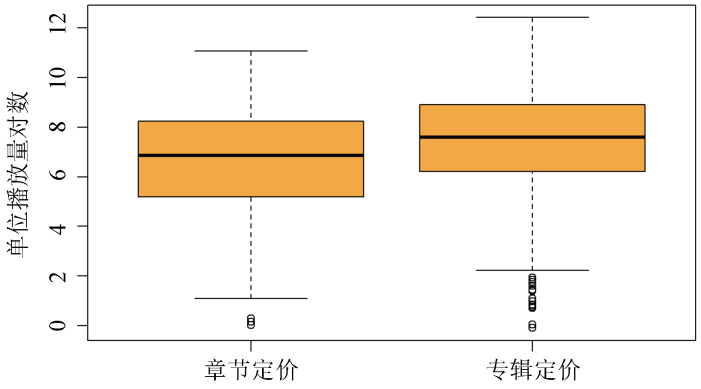

Figure 7. Boxplot of "pricing mode-logarithm of playback amount"
图 7. 定价方式－播放量对数箱线图

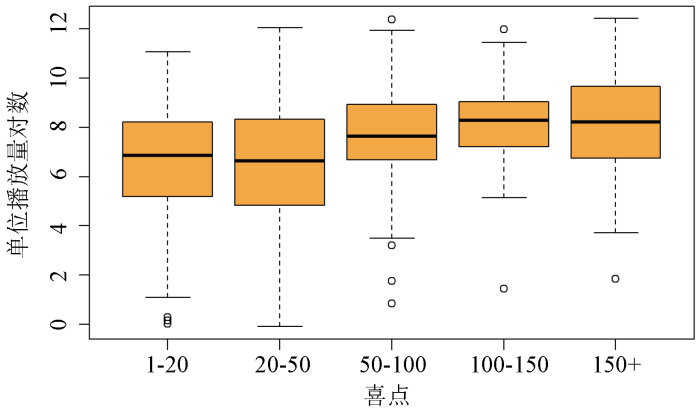

Figure 8. Boxplot of "Xi Dian-logarithm of playback amount"
图 8. 喜点－播放量对数箱线图

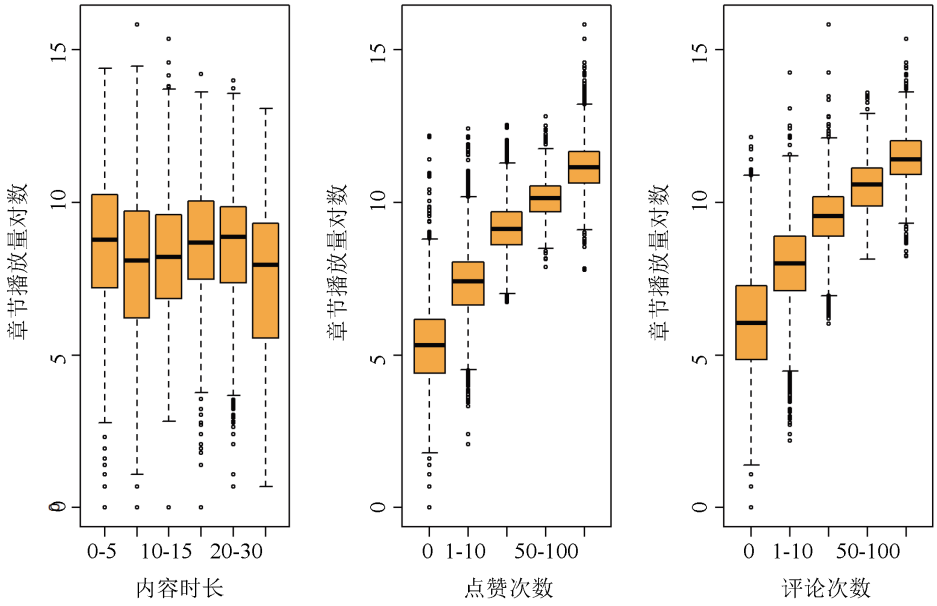

Figure 9. Boxplot of "chapter factors-logarithm of chapter playback amount"
图 9. 章节因素－章节播放量对数箱线图

Table 3. Estimation results of regression model
表 3. 回归模型估计结果

变量名	估计	P 值	备注
常数项	7.835	<0.01	
单位点赞数	0.004	<0.01	专辑点赞总数/总章节数
单位评论数	0.011	<0.01	专辑评论总数/总章节数
单位内容时长	0.018	<0.01	专辑评论总时长/总章节数
专辑定价	−0.322	<0.01	基准组：章节定价
专辑类别——儿童	0.034	0.541	
专辑类别——教育培训	−0.188	0.0016	
专辑类别——人文	−0.305	<0.01	
专辑类别——商业财经	−0.22	<0.01	基准组：有声书
专辑类别——情感生活	0.468	<0.01	
专辑类别——外语	−0.904	<0.01	
专辑类别——其他	0.226	<0.01	
标签总数	0.143	<0.01	
专辑名称高频词个数	−0.198	<0.01	
F 检验	**P 值 < 0.001**	**调整 R 方**	**0.5004**

对数－线性的回归模型解读为“增长率”[5]。在控制其他变量不变时：点赞数、评论数、内容时长和标签总数与播放量显著正相关，专辑名称的高频词汇个数呈负相关，与描述性分析专辑起名主要靠新意的想象吻合。对于定价方式这一变量，专辑定价的播放量比章节定价的播放量低于 32.22%。对于专辑类别这一变量，情感生活的播放量比有声书的播放量提升 46.8%，外语的播放量比有声书低于 90.36%。其他详细解读如图 10：回归分析模型解读。

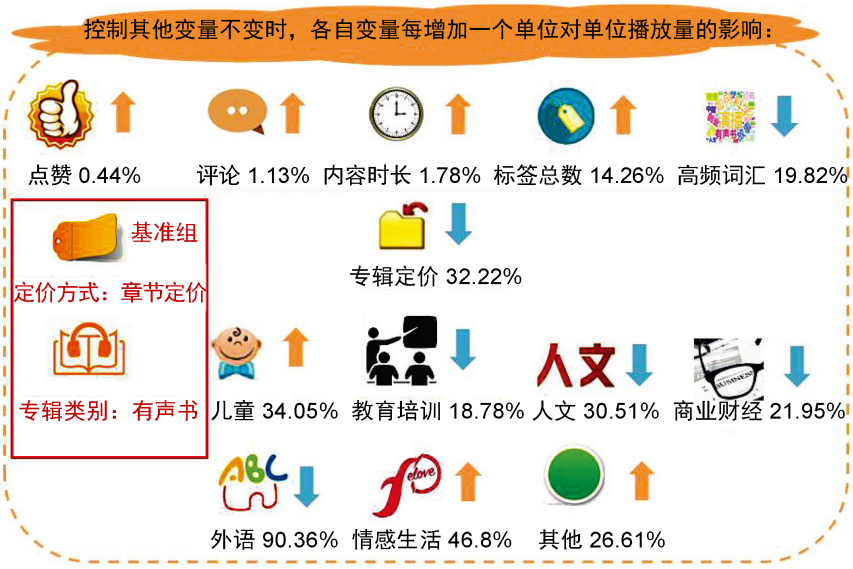

Figure 10. Interpretation of regression analysis model
图 10. 回归分析模型解读

5. 打造爆款

通过喜马拉雅付费精品专辑数据的播放量因素统计模型，我们可以对喜马拉雅新出专辑的播放量进行初步地预测，并随着专辑上线时间评论数、点赞数等变化动态预测未来的专辑总播放量。为此，我们尝试打造一个爆款专辑。假设在喜马拉雅上开设一个“情感生活”类型的专辑，每次课时为 30 分钟，采用章节定价方式，专辑名称中高频词汇为 0 个，标签个数为 4 个；在初期评论数和点赞数为 0 的情况下，根据模型预测出来将有 12,170 次播放量，高于平均水平(9265 次) 31.35%。随着时间推移，用户的口碑效应，当该专辑的单位点赞数达到 300、单位评论数达到 200，模型预测结果将达 44.02 万次单位播放量。

6. 总结及展望

国家统计局数据，2017 年教育文化娱乐支出占人均消费支出的比重 11.4%，比上年增长 8.9% [6]。随着 GDP 的增长，国民的消费结构在不断更新升级，知识付费已成为一种必然的趋势。本文借助喜马拉雅的付费精品专辑数据，探讨了课程(专辑)播放量的影响因素。研究发现：专辑名称的高频词对于提高播放量没有实质帮助，甚至可能降低专辑的播放量，起名主要靠新意；专辑需要设置一定的标签数，便于消费者理解课程的相关内容，建议 3~4 个为好；每个章节(小节)时长不宜过长，30 分钟以内为好。

同时，本案例中模型存在一定的不足，结果与描述性分析结论略有偏差，如标签总数，箱线图显示 4 个以上基本趋于平稳；回归系数表明越大越好。模型诊断的残差图存在异方差现象，模型中可能存在变量的交互影响，后续考虑尝试其他模型进行优化。此外，在未来的研究中可以加入评论内容，进行文本挖掘。

基金项目

国家自然科学基金项目青年科学基金项目，基于模型辨识的社会网络行为传播机制与行为预测的研究(批准号：61703355)，2018.01~2020.12。

参考文献

[1] 极光大数据. 极光大数据: 知识付费行业研究报告[J]. 信息与电脑(理论版), 2017(7): 21-23.

[2] 吕尧. 国内外付费问答社区研究综述[J]. 情报探索, 2018, 247(5): 129-134.

[3] 科技响铃说. 喜马拉雅内容消费节大卖 5000 万持平首年双十[Z/OL]. http://www.sohu.com/a/120676162_491065, 2016-12-05.

[4] 阿里应用分发大数据中心. 2017 年 Q2 阿里应用分发行业数据报告[EB/OL]. http://www.199it.com/archives/621344.html, 2017-08-09.

[5] 杰弗里·M·伍德里奇. 计量经济学导论[M]. 北京: 清华大学出版社, 2014.

[6] 国家统计局. 2017 年居民收入和消费支出情况[EB/OL]. http://www.stats.gov.cn/tjsj/zxfb/201801/t20180118_1574931.html, 2018-01-18.