基于多信息源与文本挖掘的智能手机销量影响因素研究与决策建议

摘要：21 世纪是智能手机的时代，随着智能手机产业的快速发展和其市场的不断扩大，对手机销售数据影响因素的分析具有很大意义。它能够确定智能手机销量的主要影响因素。现今主要的研究方法多采用单一信息源，且未考虑不同变量间的联系。本文以 2014-2018 年有代表性的手机品牌型号为研究对象，采集硬件指标和手机舆论评价文本，考虑不同类型和多信息源的数据，通过一种考虑变量间相关性的回归方法，典型变量回归(Canonical Variate Regression,CVR)分析能够推动手机销售数据增长的因素；同时我们采用 LASSO 对结果进行比较，发现 CVR 有更好的预测能力。最后辅以实际问卷调研数据进行比较，得出了运行流畅度，手机的拍照功能和手机屏幕视觉效果是消费者选购手机的重点的结论。

关键词：多信息源；智能手机；文本挖掘；销量影响；典型变量回归(CVR)

一、 前言

(一) 研究背景与内容介绍

中国的手机市场从 2000 年左右就已经存在。自安卓系统问世以来，手机用户连年增加，市场对于手机的需求高速增长，国内厂商和国外厂商之间的相互竞争使得市场不断成型。2011 年 4G 网络技术普及，更是极大推动了智能手机性能的研发和提高。与此同时，各个厂商纷纷研发起了各自的高端机系列产品，手机市场的竞争逐渐偏向高端机市场高成本高利润的路线。其中，以苹果和三星公司为首的厂商主导了早期高端机市场，并从中获得了高额的利润。在利益驱动下，各大厂商纷纷选择高端机作为利润增长点，几年间国内的智能手机市场得到了强劲的发展。手机出货量因此不断增加，智能手机也因此不断普及，成为人们生活不可或缺的一部分。但近几年来，国内智能手机出货量出现了第一次同比下滑，高端智能手机市场趋于饱和。苹果和三星作为高端机代表品牌，二者在 2019 年财报中显示的收益都略低于市场预期，同时国产中高端机市场占比不断上升，挤占了一部分智能手机的市场，普通的高端机已经难以吸引用户的眼球，这也使高端机市场的竞争越发激烈，各个厂商必须拿出更明显的竞争优势。然由于利润可观，高端机仍将是市场主流。不过，在目前市场趋于饱和的环境下，厂商仍然需要新的营销策略或技术才能抢占更多的市场份额。因此，在面对国内市场时，如何尽可能缩短消费者智能手机更换，已然成为一个为各厂商销售部门为寻求新增长的新课题。目前主流的观点认为，可以通过两个渠道寻求新的增长点。

第一，通过对智能手机进行价格细分，针对更多层面的消费者设计手机来最大化企业利润；第二，通过显著的技术创新与软硬件迭代来吸引消费者更换更新、

更智能的手机。而这又引发了更深一层的思考：对于第一点，如何细分手机价位来个性化对待消费者，从而最大化企业利润？对于第二点，怎样的技术革新能够吸引消费者的购买？本文将从舆情信息（对手机品牌型号的评论）和手机硬件指标两个方面切入国内手机市场消费情况分析，预测智能手机销量情况，给予手机品牌商更具客观性与科学性的研发指导。

（二） 研究意义与创新性

从理论角度出发对智能手机品牌的营销进行战略探究是目前针对智能手机市场的主要研究方向。例如郭雪林等将小米的营销策略概括为精准的市场定位，成功的产品策略，巧妙的价格策略，网上营销的渠道策略以及有效的促销策略，并从这五个方面探究小米手机所占市场份额逐年上升的现象的原因；陈丽引用传统的 4P 理论，从 Product、Price、Place、Promotion 四个方面研究华为的营销策略，得出了华为擅长吸收国内外智能手机优点，高端机与中低端机并行，不断更改渠道以适应市场需求，建立花粉俱乐部与良好的口碑的结论；李智谈品牌忠诚度对小米品牌的重要性等。

但是，以上所述都是只含定性的方法，它们仅在概念层面上，从市场营销的角度，对扩大手机市场提供极为有限的建议和帮助，无法就实际的数字而言给出影响手机市场的数量解释；因此在实际情况中并不能起到很好的效用。而基于市场数据对智能手机销量的影响因素进行研究，能够为智能手机厂商扩展市场提供更有力的帮助。

因此，除了关注手机厂商本身的营销能力外，在研究智能手机销量的影响因素时，也有相当一部分论文利用市场数据探讨品牌的效用和网络口碑。定量的研究有刘丽娜等在其论文中通过测度在线声誉和品牌知名度，以手机销量排名作为因变量，通过在简单线性模型加入上述两变量发现 R-Square 更大而发现品牌竞争力能够加强对销售的影响，可以为电商平台的在线声誉系统提供更加客观而有益的补充；张馨悦等通过将手机的总评论数量、负面评论、追加评论和图片数量作为网络口碑的代表变量，同样以销量排名作为因变量，做简单的线性回归后进行假设检验，得出追加评论和差评对于手机销量有很大影响力的结论。

这些定量研究的出发点位于智能手机市场的各个方面，但是目前针对智能手机市场的研究领域仍然留有部分空白。第一，目前的大多数对智能手机市场进行的分析与仍然是通过单个信息源或是多个信息源的方式进行的。如网络上消费者对手机的评论，正负评论，图片数量等等。这些研究方面过于单一，没有交叉的信息源的处理论。这可能会失去一定的准确性与合理性。第二，目前还没有研究涉足智能手机性能本身，从智能手机硬件配置与客户需求角度出发对市场进行研究。目前对手机市场的研究大多集中于某一单独的方面，主要是品牌效应或者消

费者反馈，它们没有考虑到消费者对于智能手机的期望与需求。第三，这些研究采用的是简单线性模型，将每个变量当作完全独立的个体，不能很好地解释不同变量间拥有的复杂的内生关系，这样生成的模型可能会失去一定的准确性与合理性，无法对智能手机市场的发展趋势做出精准的评估。

对此，本项目有创新点如下：

1、以往对于手机销量的影响因素研究局限于单一信息源，例如网络口碑，价格促销折扣，营销渠道等；我们的研究将基于多信息源，从不同的数据源（手机性能、网络舆情）撷取数据，在建构多因素的综合评价体系的同时也得到更符合大数据时代特点的高关联高精度的具有现实意义的结论。

2、不同信息源包含的信息有差异性和相似性。差异性在于手机硬件指标大多是数值型数据，而网络舆论主要是文字性数据。对文字性指标的整合分析是一个难点，在这点上我们将文本进行分词，去停用词，特征提取成为词向量，最后降维标准化，使之具有数值型特点。

3、现有的研究多是将不同的变量直接建立简单线性回归模型，忽略了变量之间可能存在的强相关性；而我们使用的 CVR 模型能够很好地提取不同的变量的公共部分，提高模型的稳健性与预测能力。

4、建立了关于影响因素权重的手机发售后三个月的单个模型，能够就现有的数据，通过趋势分析得到手机销售三个月内的影响因素权重。这种具有及时性的信息将更好地帮助研发公司把控智能手机市场，获取更多利润。

5、将我们的模型结果与问卷调查得到的实际一手数据进行比较，可以发现高度的一致性，以此证明我们模型的有效性与科学性。

二、 数据采集与处理

（一） 数据采集与预处理

1、自变量采集与预处理

（1）手机硬件指标

本研究通过网络爬虫的手段对多信息源研究所需要的各方面数据进行爬取。首先，我们从中关村论坛[1]抓取了 2010 年到 2019 年 4 月期间发布的 631 款智能手机的硬件参数数据。共采集到了 17 个指标如下：价格，CPU，GPU，前置摄像头像素，后置摄像头像素，前置摄像头个数，后置摄像头个数，屏幕大小，屏幕分辨率，屏幕像素，厚度，手机内存，手机电池容量，重量，分辨率，边缘长度。我们对这些数据中的分类数据化为 0-1 指示变量，并把参数中一些不易于进行分析又难以化为指示变量的数据进行了一定处理，使之数值化或化为 0-1 指示变量。

[1] http://www.zol.com.cn/

这包括通过分别对每年发布的智能手机的CPU的能力进行评估，将CPU按能力分为强中弱三个等级；屏幕材质作为分类变量处理，主要包含以下种类：康宁第三代大猩猩，康宁大猩猩玻璃，TFT材质_IPS技术，Super_AMOLED，TFT材质，Retina_HD以及JDI；识别方式亦作为分类变量处理，包含以下：前置指纹识别，后置指纹识别，侧面指纹识别，面部识别；将某些变量数值化，如屏幕分辨率（800x480像素）取乘积形式结果作为新变量。

（2）手机舆情指标与词向量处理

我们通过对中关村论坛的智能手机评论进行抓取，得到了约10万条明确了的评论。由于论坛本身带有优缺点标签，我们所爬取的评论都是明确了正负情绪的评论。例如：

积极评价：

“非常满意，我相信这样的手机每个人都应该买一部。价格不高品质好。我是凭着产品和自己的良心点评的。”；

负面评价：

“吐槽下电池的续航能力实在有点不如意，充满电一直玩微信、QQ、微博这些，听半个多小时歌，到30%的时候才用了5、6个小时，然后睡觉6个小时起来发现电没怎么掉，25%左右。所以它是待机时间不错，但玩起来就耗得太快。”

可以看到，这些正负评论一般都较为直白，覆盖手机性能的各个方面，能够对手机的性价比作出全面评估。因此，本研究利用这些文本评论作为一个信息源来对智能手机进行分析。

方法介绍：

a)word2vec

word2vec是用来生成词向量的浅层双层神经网络模型，它通过目标词汇的相邻词汇的分析来对目标词汇的词向量进行估计。一个词汇的词向量可以成为其意思的表示。因此，词汇的意思近似程度可以表现在词向量的方向近似度上。近义词的词向量内积接近1，反义词的词向量内积接近-1。除此之外，我们可以通过词向量相加的方式得到新词的词向量。例如，在得到词汇‘女’与‘国王’的词向量后，我们可以通过词向量相加的方法，近似得到词汇‘王后’的词向量。词向量的这些性质能够帮助我们构建评论的句向量，从而利用评论所带有的特定情感进行智能手机舆情分析。

b)词向量构造句向量的WR算法

本文使用词向量带权相加的方法（WR算法）对一个句子（评论）的句向量进行计算。

设一个句子 s 的构成如下：

$$s = w_1 w_2 w_3 \ldots w_n.$$

其中，$w_i, i = 1,2,\ldots,n$ 是该句子分词后的第 i 个词汇。对于任意的词汇有一个通过 word2vec 模型训练出的词向量v_{w_i}。根据 WR 算法，我们可以得到句子 s 的句向量v_s:

$$v_s = \frac{1}{|s|}\sum_{i=1}^{n}\frac{a}{a + p(w_i)}v_{w_i}$$

其中，$|s|$为句子 s 的长度，即该句子分出的词汇的个数，a 为超参数，$p(w)$ 词汇 w 在整个评论样本中出现的频率，计算公式如下：

$$p(w) = \frac{n(w)}{\sum |s|}.$$

该算法的优势在于，它通过赋予不同词向量权重的途径，改进了取平均得到句向量方法可能导致的稀有词的词向量与助词的词向量带来的偏差。除此之外，该算法可以通过调整超参数的途径得到句子向量的最优解。

实现过程：

首先我们使用 jieba 中文分词的方法去除对我们爬取的评论进行分词，在分词的过程中我们将一些如“美图手机”、“双卡双待”等 jieba 无法识别的智能手机名词补充到了分词词典中，使得结果更加准确。例如，评论

“非常满意，我相信这样的手机每个人都应该买一部。价格不高品质好。我是凭着产品和自己的良心点评的。”

的分词结果如下：

“非常 满意 ， 我 相信 这样 的 手机 每个 人 都 应该 买 一部 。 价格 不 高品质 好 。 我 是 凭着 产品 和 自己 的 良心 点评 的 。”

其次，我们使用 word2vec 模型对这些分词后的句子进行训练，得到了得到了所有出现过的词汇的 30 维词向量。利用词向量构造句向量的 WR 算法，我们用训练出的词向量构造出了每一个评论所对应的句向量。接着，我们用 PCA 降维的方法，对句向量进行降维处理。经过多次尝试，我们认为将句向量降至 18 维时能在尽可能保留信息的同时减少高维度对模型带来的负面影响。最后，对于每一款手机，我们将其评论的句向量取平均值，得到关于这款智能手机的网络舆情指标。我们用这写指标来代表网络舆情信息源进行后续的分析。

2、响应变量采集与预处理

我们在数据监控网站 Talkingdata 抓取了近十年来各种型号的智能手机在国内市场的市场份额数据,该网站通过在各大社交媒体平台进行流量监测，抓取互联网用户的智能手机机型来推算各智能手机的实时市场占比。我们利用该网站

公开的市场占比数据作为我们研究所用的因变量，即市场份额。除此之外，由于631款智能手机中部分手机的销售数据无法得到，我们对这些手机进行了删除处理，最终留下405款智能手机。同时，由于对于单个手机品牌型号的市场份额数据的数据量是不固定的，例如该手机销量不高，因此可能只有1-2个月的市场份额数据；而有的手机品牌多达十几个月，因此我们采用迂回的思路，取单个手机品牌发布后三个月的市场份额作为研究对象。若缺失，则由其他手机品牌的数据通过线性回归补出缺失值。部分非正值被替换为0.

（二） 描述性统计

1、因变量基础描述性统计分析

为了了解智能手机市场份额的分布特征与数据的基本信息，我们对因变量进行描述性统计分析。

首先，我们计算出了三个月份市场份额的描述性统计量，如表1所示。除此之外，我们还绘制了市场份额分布图如下。需要注意的是，为了更好地展示大部分智能手机销量的分布情况，我们在绘制分布图时我们对大于0.02的离群值进行了剔除。

表1 因变量描述性统计分析

描述性统计量	市场份额1	市场份额2	市场份额3
平均值	0.002185	0.002681	0.003024
中位数	0.001678	0.002029	0.002313
标准偏差	0.001838	0.002240	0.002581
偏度	3.987226	3.472879	2.732807
峰度	19.58993	16.33694	9.621930
最小值	0	0	0
最大值	0.015365	0.019354	0.018178
第一分位数	0.001371	0.001620	0.001690
第三分位数	0.002171	0.002890	0.003398

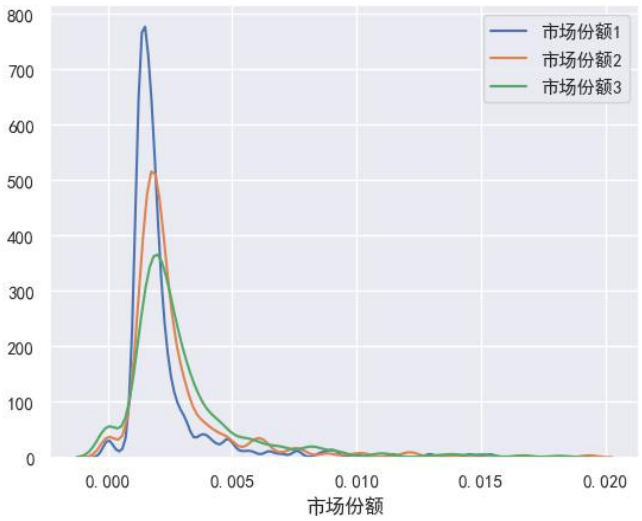

图 1 市场份额分布图

由图 1 和表 1 我们可以知道，智能手机的市场份额最大值与最小值相差较大，说明不同手机的市场份额差距有很大区别。

三个市场份额都呈右偏分布，平均值大于第三分位数，这说明存在极少部分的智能手机占有了很高的市场份额。

除此之外，三个市场份额的峰度值较高，这说明多数的智能手机市场份额都较低，仅有极少部分的智能手机市场份额销量极高。在三个市场份额之间，我们可以发现其峰度呈现下降趋势，这说明随着时间的推移，智能手机的销量与市场份额可能会由于口碑等后来因素而出现分化。

2、自变量描述性统计分析

（1）连续型

为了研究连续型自变量的特征，我们绘制了三个连续型变量的箱线图。由于硬件性能随时间而不断更新强化，为了削减这一因素的影响，在保持变量间基本关系的同时也只关注其本身，我们选择抽取了 2014 年份的智能手机进行绘图，并剔除了其中的销量离群值.

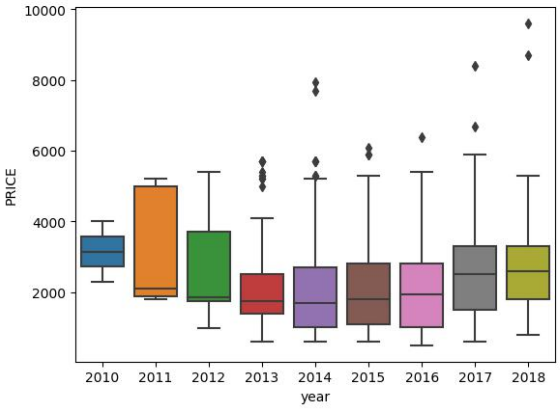

图 2-1 价格箱线图

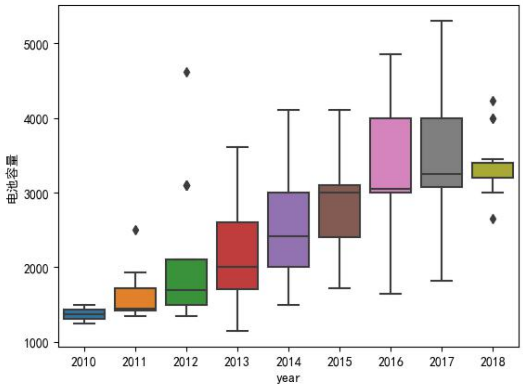

图 2-2 电池容量箱线图

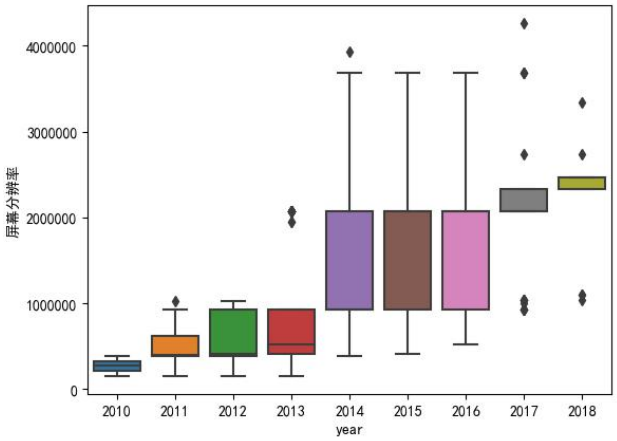

图 2-3 屏幕分辨率箱线图

由上述图可以发现，作为智能手机硬件指标的屏幕分辨率以及电池容量随着年份不断攀升，但是可以发现在最近几年其增长趋势都趋于平稳。且其上限值与下限值之差逐渐增大，说明智能手机的性能分化不断明显。除此之外，我们可以发现智能手机的平均价格在近几年逐渐上升，价格分布逐渐集中，而极少部分的智能手机价格走高至 10000 元人民币，这说明智能手机市场定价开始趋于稳定，常规智能手机与创新性的智能手机的价格差异开始出现。

（2）离散型自变量

由上文我们可以知道三个市场份额之间正相关且显著。因此，在进行离散性自变量的描述性统计时，我们只对其与第一个市场份额的关系进行分析。除此之外，为了避免硬件更新换代的影响，我们选取具有硬件指标更为丰富的发售于 2017 年的智能手机样本进行分析。

由图 2-1 我们发现，使用其他材质的屏幕的智能手机的销量平均值要高于其他的类型的屏幕。通过对属于该类的智能手机进行搜索，我们发现这部分手机的屏幕基本为国产屏幕（天马，京东方等品牌），并且手机品牌以小米居多。除此之外，屏幕材质为 Super AMOLED 的智能手机市场份额上限值较高，说明少部分使用这类屏幕的手机能够占有较高的市场份额，如 iPhone 等。

由图 2-2 可以看到，CPU 能力相对较差的智能手机的平均销量最高，而 CPU 能力强劲的智能手机平均销量最低。除此之外，可以看到 CPU 能力处于中等水平（Normal）于强劲水平的智能手机的市场份额上限比 CPU 能力较差的智能手机高，其中中等水平的上限值十分突出，说明 CPU 性能处于中间水平的智能手机在性价比的权衡方面较为优异。

由图 2-3 可以看到，无指纹解锁的手机的市场份额平均值最高，带有前置指纹识别的智能手机市场份额平均值最低。除此之外，可以看到后置指纹识别的上限值要高于其他两者，且离群值所属的智能手机使用的都是指纹识别解锁模式。这说明普通的密码解锁与其对应价格的性价比能够带来稳定且较高的市场份额，但是拥有指纹解锁的机型更有可能抢占更大的市场份额从而获取更多利润。

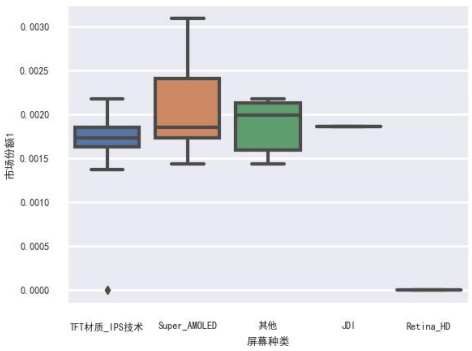

图 3-1 屏幕种类箱线图

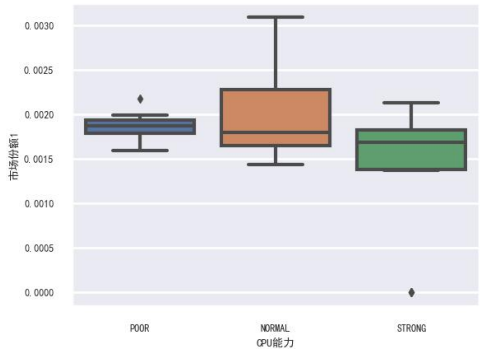

图 3-2 CPU 箱线图

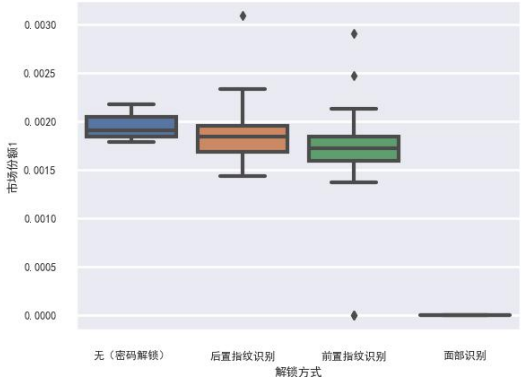

图 3-3 解锁方式箱线图

三、 模型建立与检验

（一） 模型变量说明

在我们这次模型的建立中，我们的响应变量分别为第一个月，第二个月，第三个月的手机市场份额。在这种研究情形下，我们更能清晰地看出单个手机品牌型号的市场份额随着时间的增长，在相同的影响因素下有如何的类时间序列的变化趋势。

模型中的自变量分为两部分：第一部分为手机硬件指标。

在这里我们进行了进一步的筛选和整理：选择是数值形式，且表示为非稀疏的向量的变量。综上，共有 13 个数值型指标，分别为 CPU 排名，价格，前置摄像头总数，后置摄像头总数，后置摄像头像素，前置摄像头像素，屏幕大小，边缘长度，内存，重量，电池容量，屏幕分辨率，摄像头总数。

第二部分为词向量，在前面已经介绍过了，每一个手机品牌型号的词向量是一个 17 维的向量。

（二） CVR

1、模型简介

CVR，即典型变量回归（Canonical Variate Regression），是在 CCA，即典型相关分析（Canonical Correlation Analysis）的基础上进行补充与改良的回归方法。在实际数据中，一方面我们需要探索多组实验的关联结构，二是建立一种预测未来结果的简约模型。为了同时解决这两个问题，CVR 提供了一个统一的规范变量回归框架。该框架将多个典型相关分析与预测建模相结合，平衡典型变量的关联强度及其对结果的联合预测能力。此外，所提出的准则同时寻找多个典型变量集，以便能够检查它们对结果的联合影响。

由于词向量由对手机的评论提取得到，其中手机评论基本都是针对手机硬件，因此我们考虑到词向量与手机硬件指标有着高度的相关性，所以我们使用 CVR 提取两部分的共同因子再进行回归，不仅提高准确度并且也更具有说服力。

CVR 的优化公式如下：

$$\min \eta \sum_{k<j} \frac{1}{2} \left\| B_k - B_j \right\|_F^2 + (1-\eta) \sum_{k=1}^{K} \ell(Y, Z_k) + \sum_{k=1}^{K} \rho_k(W_k, \lambda_k)$$

其中 $X_k W_k = B_k, B_k{}^T B_k = I_r, Z_k = 1\alpha^T + B_k \beta$.

$X_k W_k$ 是典型变量因子其中 W_k 是对应 X_k 的权重，不为 0 的行对应 X_k 中有影响作用（被提取以作为公共基部分）的列；β，α 是拟合后的系数，η 用于调节增强典型变量间的紧密程度和提高模型预测能力之间的相对权重，λ 用以调整 $X_k W_k$ 的稀疏程度。在模型中会首先得到对应各个 X_k 的 W_k，然后再次拟合得到 β，α。

得到拟合参数后可通过计算 $\alpha + X_k W_k \beta$ 得到预测的 $\hat{Y}$。

2、**模型建立**

在将词向量和手机硬件指标变量进行匹配的过程中我们损耗了多个手机样本，因此进入模型的共有有效手机品牌型号 244 个，这即是我们的样本容量。

首先我们选择总数据量的百分之三十作为测试集，百分之七十作为训练集。由此，测试集共有 74 个数据，训练集共有 170 个数据。

其中 X_1 是词向量组成的变量矩阵，X_2 是硬件指标组成的变量矩阵。$X_1 W_1$ 以及 $X_2 W_2$ 分别代表词向量典型变量因子和硬件指标典型变量因子。

我们对训练集调用 R 中 CVR 包中的 CVR 函数，参数分别为 Y，响应变量；Xlist，将两部分自变量指标合并在一起的自变量矩阵；rankseq 是 CVR 通过 Cross Validation 将从其中选出的最佳秩，由于我们的第一部分列个数为 13 小于第二部分列个数 17，因此我们设置其为包含 1 至 13 整数的整数向量；neta 是 η 的个数，默认为 5；nlam 是λ的个数，默认为 25；family 是响应变量的类型，设置为"g"，即服从高斯分布；nfold 是进行交叉验证时所取的验证集与训练集集合个数总和，默认为 10。

经过 20 次 CVR 的拟合，由于交叉验证的随机性，我们得到模型拟合的秩分别为：13、2、2、3、3、4、5、2、4、3、8、13、6、4、3、4，出现频率最高的分别是 3，2，4，13.由于 13 在第一次拟合中结果中的 MAPE 过大，再加上模型结构过于复杂，我们选取 2，3，4 作为预设的秩代入模型中训练，并得到结果。

结果输出一下部分：$X_1 W_1$ 和 $X_2 W_2$ 的相关系数矩阵以及拟合的 W_1、W_2，β，α。

每次用拟合好的各个系数，利用测试集的数据通过 CVR 模型准则计算出 $\alpha + X_k W_k \beta$，即是 $\hat{Y}$。通过比较 Y 和 $\hat{Y}$ 计算出平均百分比误差，即 $\sum_{i=1}^{n}\left|\frac{\hat{Y}-Y}{Y}\right|\times\frac{100}{n}$ (MAPE)，作为我们衡量模型好坏的指标。

3、**模型结果展示及比较**

由于输出结果非常详细，我们谨在此列出关键结果。其余结果可在附录一中查看。

（1）**首月手机市场份额**

当秩为 2 时，X_1W_1 和 X_2W_2 的相关系数矩阵为 $\begin{bmatrix} -0.071 & 0.073 \\ 0.193 & -0.167 \end{bmatrix}$，

回归方程为

$$\begin{bmatrix} X_1W_1 & X_2W_2 \end{bmatrix}\begin{bmatrix} 2.500\text{e-}03 \\ 8.753\text{e-}03 \\ -1.614\text{e-}04 \\ -1.568\text{e-}05 \end{bmatrix} + 0.000554$$

此时 R-Square 为 0.9941272，MAPE 为 0.4674679。
同时输出了 W_1、W_2。其中 W_1 如下：

表 2 W_1 数值矩阵

0	0
0.004182942	0.007666354
0	0
-0.004210232	0.03381805
0	0
0	0
-0.065938803	0.029688712
0	0
0	0
0	0
-0.032692315	-0.062193904
0	0
0	0
0	0

可以看见该矩阵非常稀疏，CVR 起到了很好的降维和选择相关变量的效果。
当秩为 3 时，X_1W_1 和 X_2W_2 的相关系数矩阵为

$$\begin{bmatrix} 0.312 & -0.313 & -0.312 \\ 0.200 & -0.202 & -0.201 \\ 0.176 & -0.177 & -0.177 \end{bmatrix}$$

回归方程为

$$[X_1W_1 \quad X_2W_2]\begin{bmatrix} 2.252e-03 \\ 8.904e-03 \\ -2.783e-03 \\ 1.242e-06 \\ 3.530e-05 \\ -1.324e-05 \end{bmatrix} + 0.00260$$

此时 R-Square 为 0.994132，MAPE 为 0.466385。

当秩为 4 时，X_1W_1 和 X_2W_2 的相关系数矩阵为

$$\begin{bmatrix} -0.108 & -0.109 & -0.109 & 0.109 \\ -0.011 & -0.013 & -0.013 & 0.013 \\ 0.055 & 0.054 & 0.054 & -0.053 \\ -0.133 & -0.131 & -0.131 & 0.131 \end{bmatrix}$$

回归方程为

$$[X_1W_1 \quad X_2W_2]\begin{bmatrix} 0.00258 \\ 0.00445 \\ -0.000922 \\ 0.00122 \\ -0.0424 \\ 0.102 \\ -0.0566 \\ 0.0468 \end{bmatrix} + 0.0400$$

此时 R-Square 为 0.9941406，MAPE 为 0.4733892。

（2）次月手机市场份额

秩为 2 时，X_1W_1 和 X_2W_2 的相关系数矩阵为 $\begin{bmatrix} 0.421 & 0.418 \\ 0.263 & 0.256 \end{bmatrix}$，回归方程为

$$[X_1W_1 \quad X_2W_2]\begin{bmatrix} 5.510e-03 \\ 4.371e-03 \\ -3.826e-04 \\ 4.459e-05 \end{bmatrix} + 0.00291$$

此时 R-Square 为 0.9941362，MAPE 为 0.4481619。

秩为 3 时，X_1W_1 和 X_2W_2 的相关系数矩阵为

$$\begin{bmatrix} 0.401 & 0.401 & -0.401 \\ 0.108 & 0.108 & -0.109 \\ 0.244 & 0.243 & -0.244 \end{bmatrix}$$

回归方程为

$$[X_1W_1 \quad X_2W_2]\begin{bmatrix} 5.872e-03 \\ 9.820e-03 \\ 4.186e-03 \\ -2.638e-07 \\ 3.775e-05 \\ 1.892e-05 \end{bmatrix} + 0.00255$$

此时 R-Square 为 0.994134，MAPE 为 0.4469069。

秩为 4 时，X_1W_1和X_2W_2的相关系数矩阵为

$$\begin{bmatrix} 0.059 & 0.059 & 0.059 & 0.059 \\ -0.063 & -0.062 & -0.063 & -0.062 \\ 0.113 & 0.113 & 0.113 & 0.113 \\ 0.109 & 0.109 & 0.109 & 0.109 \end{bmatrix}$$

回归方程为

$$[X_1W_1 \quad X_2W_2]\begin{bmatrix} 0.00404 \\ -0.001 \\ -0.00456 \\ 0.00477 \\ -0.113 \\ 0.0356 \\ 0.0876 \\ -0.0312 \end{bmatrix} - 0.00901$$

此时 R-Square 为 0.9941302，MAPE 为 0.4463717。

（3）第三个月手机市场份额

秩为 2 时，X_1W_1和X_2W_2的相关系数矩阵为

$$\begin{bmatrix} 0.171 & 0.171 \\ -0.181 & -0.181 \end{bmatrix}$$

回归方程为

$$[X_1W_1 \quad X_2W_2]\begin{bmatrix} 2.630e-03 \\ -6.777e-04 \\ -6.805e-05 \\ 7.607e-05 \end{bmatrix} + 0.00495$$

此时 R-Square 为 0.9941374，MAPE 为 0.4765387。

秩为 3 时，X_1W_1和X_2W_2的相关系数矩阵为

$$\begin{bmatrix} 0.029 & -0.031 & -0.031 \\ 0.001 & -0.080 & -0.080 \\ 0.114 & 0.081 & 0.081 \end{bmatrix}$$

回归方程为

$$\begin{bmatrix} X_1W_1 & X_2W_2 \end{bmatrix} \begin{bmatrix} 0.00157 \\ 0.00571 \\ -0.000983 \\ -0.0548 \\ -0.0236 \\ 0.0318 \end{bmatrix} - 0.00532$$

此时 R-Square 为 0.9941398，MAPE 为 0.5254488。

秩为 4 时，X_1W_1 和 X_2W_2 的相关系数矩阵为

$$\begin{bmatrix} -0.037 & -0.037 & -0.037 & -0.041 \\ 0.012 & 0.012 & 0.011 & 0.018 \\ 0.119 & 0.118 & 0.118 & 0.123 \\ -0.115 & -0.115 & -0.115 & -0.117 \end{bmatrix}$$

回归方程为

$$\begin{bmatrix} X_1W_1 & X_2W_2 \end{bmatrix} \begin{bmatrix} 0.00246 \\ 0.00121 \\ -0.00223 \\ 0 \\ -0.00603 \\ 0.00995 \\ -0.00796 \\ 0 \end{bmatrix} + 0.0174$$

此时 R-Square 为 0.9941353，MAPE 为 0.4851382。

通过比较，得出以下总表。

表 3 CVR 不同秩情况下不同月份 R-square 与 MAPE

第一个月	R=2	R=3	R=4
R-square	0.9941272	0.994132	0.9941406
MAPE	0.4674679	0.466385	0.4733892
第二个月	R=2	R=3	R=4
R-square	0.9941362	0.994134	0.9941302
MAPE	0.4481619	0.4469069	0.4463717
第三个月	R=2	R=3	R=4
R-square	0.9941376	0.9941398	0.9941353
MAPE	0.4765387	0.5254488	0.4851382

首先我们以 MAPE 为主要指标；同时再比较 X_1W_1 和 X_2W_2 的相关系数矩阵，相关系数矩阵的对角元素越大，代表 CVR 提取出的两部分自变量的共同部分关联性越强，也代表 CVR 模型的作用效果越好，模型表现力更好；最后通过秩大小考虑模型复杂度。因此我们在第一个月时选取秩为 3 的模型，在第二个月考虑秩为 2 的模型，在第三个月选取秩为 2 的模型。

4、模型解释

通过模型给出的拟合结果，我们可以直接给出对应的预测公式 $\alpha + X_k W_k \beta$. 其次我们可以根据 W_k 观察到被提取到公共部分的即重要部——W_k 不为 0 的行即对应了重要指标。

就词向量指标来看，手机第一个月的市场份额的关键指标为第 5，10，12，13，15，17 维；第二个月的市场份额的关键指标为第 3，5，8，10，12，13，14，17 维；第三个月的市场份额的关键指标为第 5，10，11，12，13 维。

就手机硬件指标来看，手机第一个月的市场份额的关键指标为价格，前置摄像头像素，边缘长度，屏幕分辨率；第二个月的市场份额的关键指标为价格，后置摄像头像素，前置摄像头像素，屏幕大小，边缘长度和屏幕分辨率；第三个月的市场份额的关键指标为价格，后置摄像头总数，后置摄像头像素，前置摄像头像素，边缘长度，屏幕分辨率和摄像头总数。

（三） LASSO

1、方法介绍

为了和 CVR 模型效果进行对比，我们另采取 lasso 方法进行建模。Lasso 是一种压缩估计，在 RSS 最小化的计算中加入一个范数作为罚约束。它通过构造一个惩罚函数得到一个较为精练的模型，使得它压缩一些回归系数，即强制系数绝对值之和小于某个固定值；同时设定一些回归系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

2、模型建立

首先对参数数据和词向量标准化，按照手机型号匹配参数数据和评论词向量后删掉不可避免的缺失值，得到有效的 244 条数据（含缺失值的条数为 40 条）。其中 $X_1 - X_{17}$ 为词向量指标，销售额占比为响应变量，其他指标为参数指标。我们随机取样 70%作为训练集，剩下 30%作为测试集。分别对第一月、第二月和第三月销售额占比进行 lasso 回归并筛选变量。

为了检验词向量和参数指标是否都对销售额占比有所影响，我们将分别对词向量和参数指标进行 lasso 回归查看效果。对三个月的销售额占比进行回归的过程十分类似，此处第一个月的销售额占比为例。

（1）同时放入两种指标

对第一月的两个指标同时进行 lasso 回归结果：

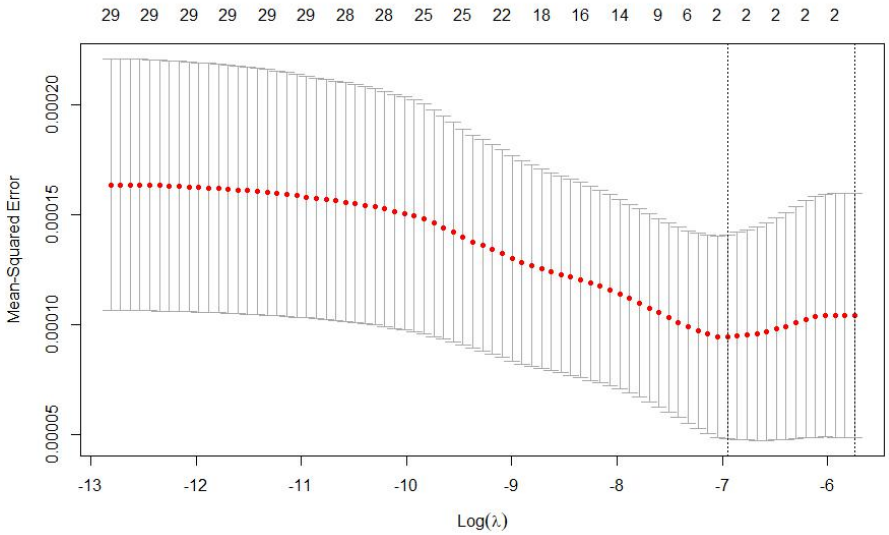

图 4-1 两指标模型 Log(λ)-MSE 效果图

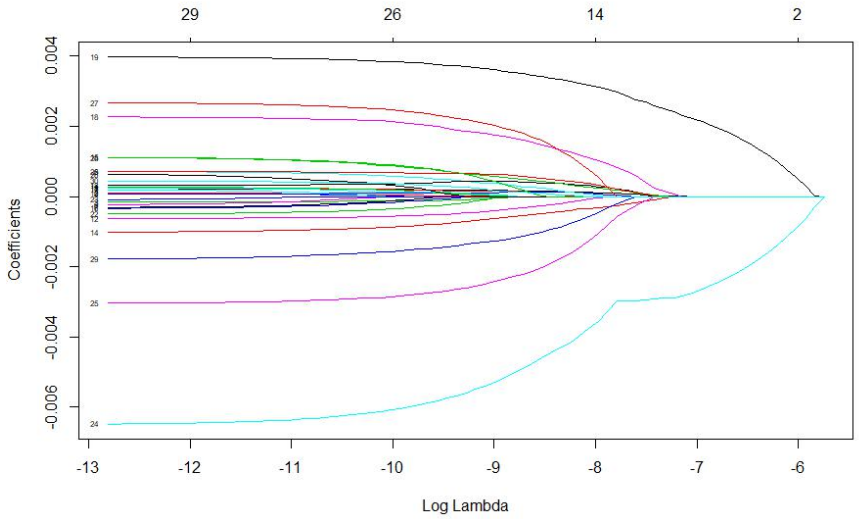

图 4-2 两指标模型 Log(λ)-Coefficients 效果图

从图 4-1 可以看见 $Log(\lambda)$ 为-7 左右时，模型的 MSE 达到最小，但从图 4-2 中可以看到各个系数的收缩效果非常明显，对应于 $Log(\lambda)$ 为-7 时的只有两个指标的系数不为 0，这使得我们难以分析各个指标对于销售额占比的影响，因此我们手动筛选指标数为 10 个左右，选择 λ=0.0005，模型筛选的变量如表，通过计算，此时的模型方程为

3.542315e03+5.707113e05 X_1 +1.786125e04 X_{14} +1.025316e04 X_{15} +5.873540e04 CPU 性能排名+2.762891e03PRICE+1.285486e04 前置摄像头总数+2.977064e03 屏幕大小+2.564768e04 边缘长度

MAPE 为 1.785575。

表 4-1 LASSO 两指标模型系数表格

自变量	系数	自变量	系数
(Intercept)	3.542315e03	CPU 性能排名	5.873540e04
X1	5.707113e05	PRICE	2.762891e03
X13	1.148784e04	前置摄像头总数	1.285486e04
X14	1.786125e04	屏幕大小	2.977064e03
X15	1.025316e04	边缘长度	2.564768e04

（2）只放入词向量指标

对词向量单独进行 lasso 回归，结果为

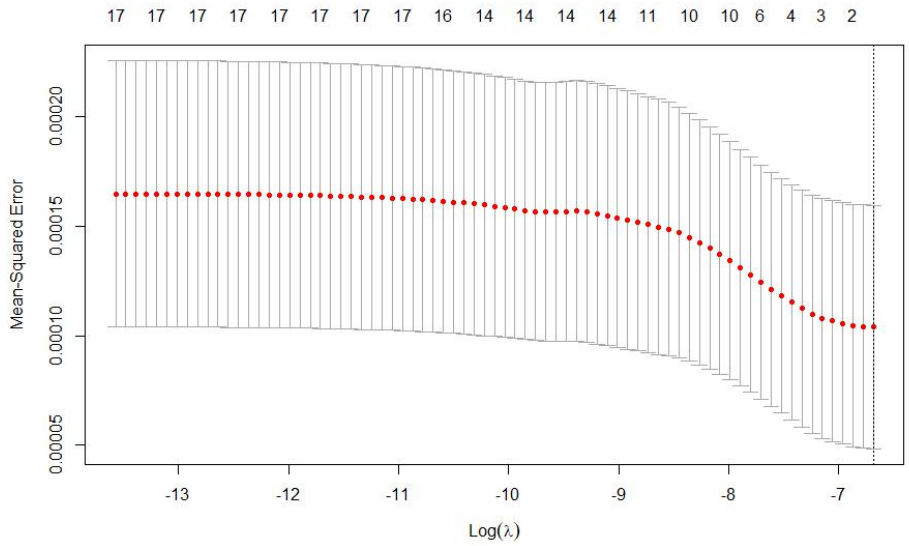

图 4-3 词向量模型 Log(λ)-MSE 效果图

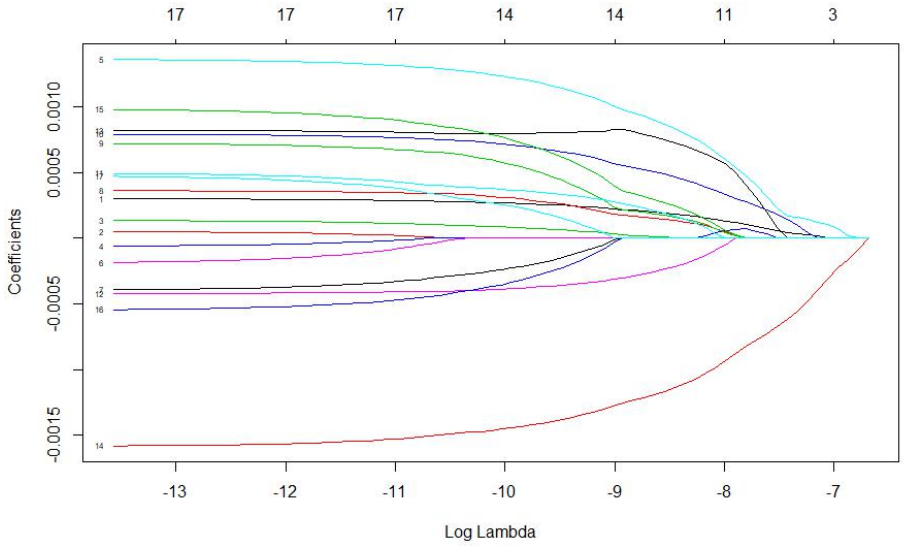

图 4-4 词向量模型 Log(λ)-Coefficients 效果图

从图 4-3 可以看见 $Log(\lambda)$ 为-7 左右时，模型的 MSE 达到最小，但从图 4-4 中可以看到各个系数的收缩效果非常明显，对应于 $Log(\lambda)$ 为-7 时的只有两个指标的系数不为 0，这使得我们难以分析各个指标对于销售额占比的影响，因此我们手动筛选指标数为 10 个左右，选择 λ=4e-04，模型筛选的变量如表，此时模型方程为

3.528399e-03+1.096143e-04 X_1 +7.258018e-05 X_4 +4.883745e-04 X_5 +6.972440e-06 X_8 +4.466316e-06 X_9 +2.838687e-04 X_{10} +4.277038e-04 X_{13} +-8.357623e-04 X_{14} +6.102150e-06 X_{15}

此时 MAPE=1.236344。

表 4-2 LASSO 词向量模型系数表格

自变量	系数	自变量	系数
(Intercept)	3.528399e-03	X9	4.466316e-06
X1	1.096143e-04	X10	2.838687e-04
X4	7.258018e-05	X13	4.277038e-04
X5	4.883745e-04	X14	-8.357623e-04
X8	6.972440e-06	X15	6.102150e-06

（3）只放入参数指标

对参数数据单独进行 lasso 回归，结果为

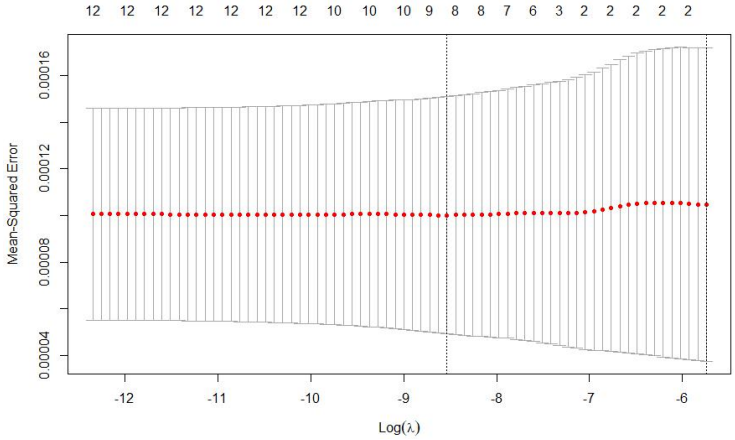

图 4-5 硬件指标模型 Log(λ)-MSE 效果图

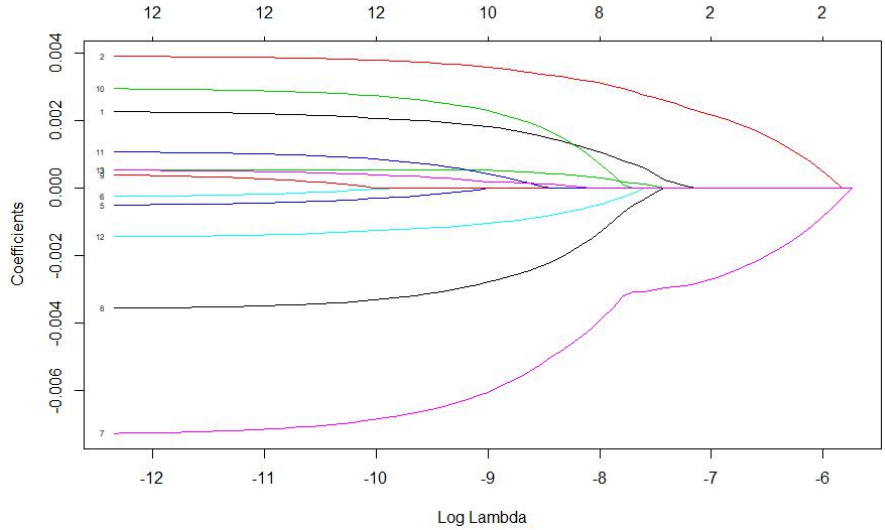

图 4-6 硬件指标模型 Log(λ)-Coefficients 效果图

从图 4-5 可以看见 $Log(\lambda)$ 为-7 左右时，模型的 MSE 达到最小，但从图 4-6 中可以看到各个系数的收缩效果非常明显，对应于 $Log(\lambda)$ 为-7 时的只有两个指标的系数不为 0，这使得我们难以分析各个指标对于销售额占比的影响，因此我们手动筛选指标数为 10 个左右，选择 λ=0.00012，筛选的变量如表，此时模型方程为

3.604155e03+1.842528e03CPU 性能排名+3.592942e03PRICE+5.365265e04 前置摄像头总数+1.930287e05 后置摄像头像素总和+2.804766e03 边缘长度
+2.315172e03 重量+4.399085e04 电池容量+1.056420e03 屏幕分辨率
+2.042557e04 摄像头总数

此时 MAPE 为 2.25068。

表 4-3 LASSO 硬件指标模型系数表格

自变量	系数	自变量	系数
(Intercept)	3.604155e03	边缘长度	2.804766e03
CPU 性能排名	1.842528e03	重量	2.315172e03
PRICE	3.592942e03	电池容量	4.399085e04
前置摄像头总数	5.365265e04	屏幕分辨率	1.056420e03
后置摄像头像素总和	1.930287e05	摄像头总数	2.042557e04

（4）第二月和第三月

同理，可以得到第二月、第三月的 MAPE 结果如表

表 5 LASSO 不同月份不同模型 MAPE 表格

MAPE	词向量&参数数据	词向量	参数数据
第一月	1.785575	1.236344	2.25068
第二月	1.482916	1.41039	1.958688
第三月	1.82381	1.637647	2.231098

由此我们可以知道将两个指标同时放入模型时具有一定的可靠性。同时也说明了词向量确实有着优秀的泛化能力，同时在我们的模型中作为舆情指标具有非常好的解释能力。

3、模型解释

根据模型结果，在词向量中，对第一个月市场份额影响关键的变量为第 1，13，14，15 维；对第二个月市场份额影响关键的变量为第 1，4，13，14，15 维；对第三个月市场份额影响关键的变量为第 1，4，13，14，15 维。

在参数数据中，对第一个月市场份额影响关键的变量为 CPU 性能排名，价格，前置摄像头总数，屏幕大小，边缘长度；对第二个月市场份额影响关键的变量为 CPU 性能排名，价格，前置摄像头总数，屏幕大小，边缘长度，屏幕分辨率；对第三个月市场份额影响关键的变量为价格，前置摄像头总数，屏幕大小，边缘长度，屏幕分辨率。

四、 调研实证数据分析

为了验证我们的模型不仅在面板数据上具有极高的准确性，我们亦采取问卷调查的形式，拟将我们模型得到的结果和实际消费者的购买倾向进行比较，验证。

问卷调查共有 6 个部分：

1、您现在使用的手机品牌/手机型号；

2、您现在使用的手机价位？

3、理想的手机价位？

4、您看重的手机硬件设置？

5、手机舆论评价对你的购买选择影响大吗？

6、您的年龄？

我们共收到实际填写有效问卷数 336 份，其中年龄误填/漏填（小于 10 或大于 100）共 10 份，在统计年龄时被去掉。问卷调查结果如下：

（一）消费者智能手机价位占比：

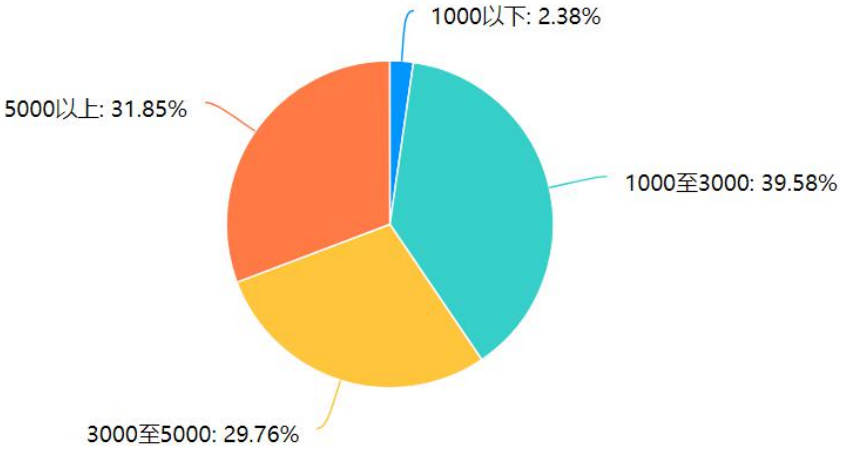

图 5-1 消费者使用手机价位饼图

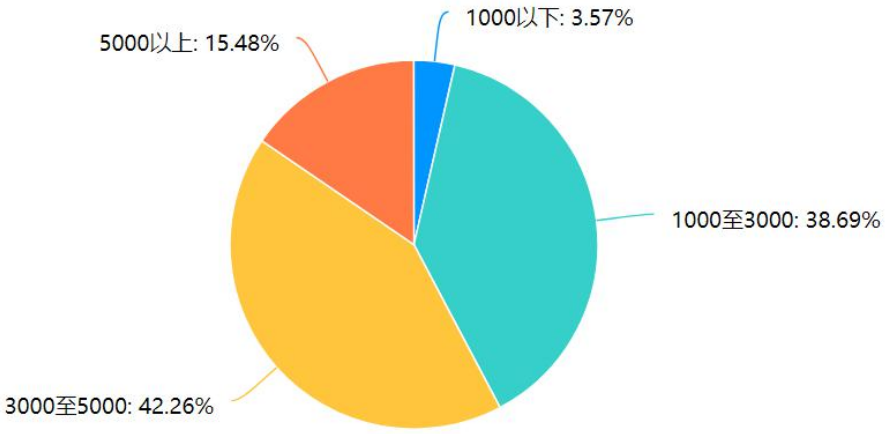

图 5-2 消费者理想手机价位饼图

由此可见，消费者现使用的手机从 1000 至 5000 多元的价位不等，看起来比较平均；而理想的手机价位主要集中在 1000-5000 间，其中支持 3000 至 5000 元的手机价位的消费者占多数。

（二）网络舆情对消费者的影响程度统计：

根据问卷调查的问题 5，我们制作了以下图表：

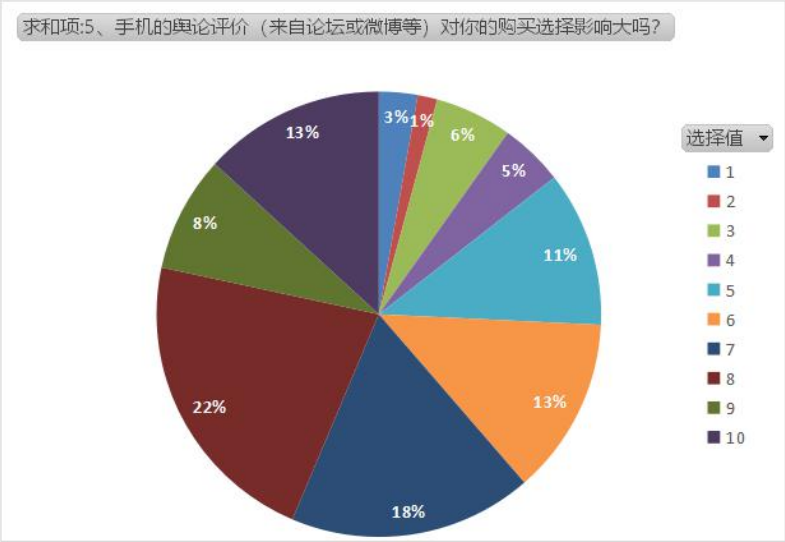

图 5-3 舆论评价对消费者的影响能力饼图

由问卷调查统计得知，舆论评价对消费者的影响能力平均值为 5.4。除此之外，我们可以发现有 75%以上的消费者选择了五分以上，这说明网络舆情对于一款智能手机发布后的后续销量可能会产生很大影响。因为大多数会基于已经购买的消费者对智能手机做出的评价来决定是否购买这部手机。因此，本模型对网络舆情的考虑具有一定的合理性。

1. **硬件设施关注度：**

我们基于问题 4 的反馈得到了下图的结果：

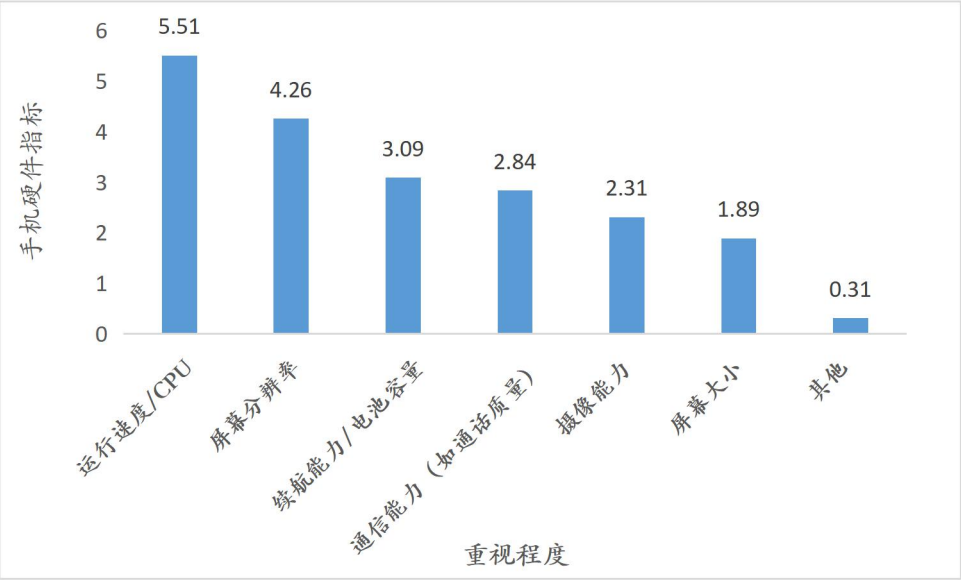

图 5-4 硬件设施关注度条状图

由该图我们可以发现，手机消费者对手机硬件的关注度从大到小分别是：CPU，电池容量，摄像能力，屏幕分辨率，屏幕大小以及通话和其它功能（外观/系统/内存/屏幕刷新率等）。

我们的模型与实证调查相近的结果在于对 CPU，摄像能力和屏幕分辨率，屏幕大小对消费者具有吸引力占有一定的地位具有一定的契合度，但我们的模型并没有对电池容量有突出作用有一定的说明。

我们考虑情况如下：随着手机的逐步发展，手机的电池容量越来越大，从老牌智能机中兴的早期智能机 U793 到现今最受人青睐的华为 P30 Pro，电池容量从 1150mAh 飞跃到了 4500mAh，几乎是三倍的数量关系，因此消费者自然是关注手机的电池容量的。

但我们同时发现手机的电池容量又不是消费者在实际购买时关注的重点，第一是市面上同时期的手机电池容量与其价格拥有显著相关性，且不同智能手机的电池容量几乎相同，而不像 CPU 和像素等其他硬件一样性能参差不齐。因此，目前不同手机的电池容量无法给予消费者一定的区分度。第二，手机的电池容量仍然很难赶上消费者对手机更频繁和更高强度的使用需求。随着移动物联网的不断发展，消费者在智能手机上的生产生活越来越复杂而频繁，对手机电量的消耗也因此不断提升。我们从商场内的充电宝租借服务的火爆可以看出目前的智能手机电池容量仍然是难以满足需求的，消费者的充电频率仍然在上升；再加上手机的电池有一定的使用寿命，使用时间越长损耗速度越快，进一步使得当下的电池容量无法满足消费者的需求。

除此之外，我们发现舆论对消费者购买选择的影响力度是 5.4 左右（满分为 10），这并不是代表影响力弱，而恰恰相反，因为该数字是在描述程度，当大于 5 时，我们便有自信认为舆论有一定且较大的影响。而不是特别大的原因在于：1、有部分中老年消费者了解信息媒介有限；2、如 5-5 图，根据影响程度<=5 的消费者所使用的手机品牌，我们发现使用华为、iPhone 和小米的消费者共占有 83%；而这三个品牌的品牌效应强，品牌黏度大，舆论评价对忠实的品牌拥护者自然影响不大。因此我们可以得到结论：消费者对舆论是有一定的敏感度的。

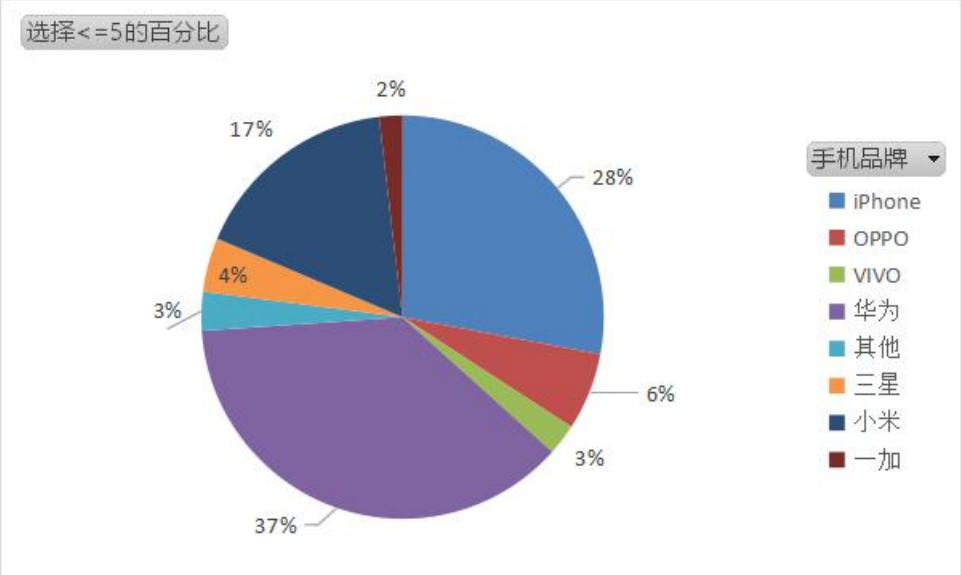

图 5-5 影响程度<=5 的消费者所使用的手机品牌饼图

综上，本模型中对智能手机的消费者评论，即舆论效应的考虑，以及对各硬件设置对智能手机销量的影响能力的评估，具有一定的科学性和合理性。

五、 总结

根据 CVR 和 LASSO 给出的结果，我们有如下表格。

表 6 CVR 和 LASSO 结果 MAPE 表格

MAPE	CVR	LASSO
第一月	0.466385	1.785575
第二月	0.4481619	1.482916
第三月	0.4765387	1.82381

经过 CVR 和 LASSO 的变量筛选，我们发现选出来的词向量变量略有差异，而手机硬件指标变量几乎相同，这说明我们的模型达到了一定的作用，体现了相同数据下结论相同的一致性。

同时我们发现 CVR 模型的 MAPE 远小于 LASSO 给出的 MAPE，相对大小比分别为 282%，231%，283%；并且选出更少的变量，给出更为简洁的预测方程，因此我们更加验证了 CVR 模型对具有一定关联强度的典型变量的分析更有稳健性。

综上我们可以得出如下结论：

后置摄像头像素，前置摄像头像素和屏幕分辨率在三期的手机市场份额里都占有重要地位的变量，因此我们认为这三个变量是消费者非常关心的三个手机硬件指标；

CPU，屏幕大小也占有重要地位；

价格作为购买时的必要条件也时消费者考虑的重点之一；

随着手机推出到市场后的时间逐日增长，后置摄像头像素和屏幕大小，边缘长度和摄像头总数的重要性逐渐体现；

消费者对手机的功能性的要求并主要体现在手机的运行流畅度，手机的拍照功能和追求使用手机良好的视觉感触（边缘长度，屏幕大小和屏幕分辨率）上，对手机的内存、续航等的要求没有这么强烈。

因此我们建议手机厂家在拍照功能，手机视觉效果和手机运行流畅度上加大资金投入，从而以消费者为导向进行智能手机的开发与销售，提高利润，寻求智能手机行业利润增长模式的转型。

参考文献

[1]饶东宁,邓福栋,蒋志华. 基于多信息源的股价趋势预测[J]. 计算机科学,2017,44(10):193-202.
[2]侯长海. 2016年上半年智能手机市场分析[J]. 互联网天地,2016(10):74-76.
[3]Chongliang Luo, Kun Chen.Canonical variate regression[J].Biostatistics,2016,17(3):468-483.
[4]Sanjeev Arora, Yingyu Liang, Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings：ICLR 2000：proceedings of Conference TrackInternational Conference on Learning RepresentationsToulon, France, April 24 - 26[C].
[5]Robert, Tibshirani. Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996.
[6]张馨悦,张亿辰,周世宁. 网络口碑对手机销量影响的实证研究[J]. 现代商业,2017(24):15-16.
[7]郭雪林,卢黎莉. 小米手机的营销策略[J]. 中国商论,2018(29):59-60.
[8]陈丽. 华为手机营销策略分析[J]. 科技经济市场,2018(09):106-107.
[9]刘丽娜,齐佳音,张镇平,曾丹. 品牌对商品在线销量的影响——基于海量商品评论的在线声誉和品牌知名度的调节作用研究[J]. 数据分析与知识发现,2018,2(09):10-21.
[10]李智. 浅谈手机品牌忠实度的塑造[J]. 财经界(学术版),2014(07):30.