中文图书分类号：0212.1
密　　级：公开
UDC：311
学　校　代　码：10005

硕士专业学位论文

PROFESSIONAL MASTER DISSERTATION

论　文　题　目：我国智能手环行业发展分析及销量预测研究

论　文　作　者：张碧佳
专业类别/领　域：应用统计
指　导　教　师：关丽　副教授

论文提交日期：2020 年 5 月

UDC：311　　学校代码： 10005
中文图书分类号：O212.1　　学　　号： S201806014
密　　级： 公开

北京工业大学硕士专业学位论文

（全日制）

题　　目： 我国智能手环行业发展分析及销量预测研究

英文题目： RESEARCH ON THE DEVELOPMENT ANALYSIS AND SALES FORECAST OF THE SMART BRACELET INDUSTRY IN CHINA

论 文 作 者： 张碧佳

专业类别/领域： 应用统计

研 究 方 向： 精算统计

申 请 学 位： 应用统计硕士专业学位

指 导 教 师： 关丽 副教授

所 在 单 位： 理学部

答 辩 日 期： 2020 年 5 月

授予学位单位： 北京工业大学

摘 要

在当今的社会背景下，科技发展迅速，人们的精神追求不断提高。智能手机、智能电脑、智能移动性可穿戴设备等逐步成为人们关注的对象。随着智能手机的普及，可穿戴智能设备已经成为制造商的宠儿，成为电子产品消费的新热点。特别是智能手环，不仅具有消息提醒、运动计步、心律检测、睡眠监测等功能，还拓展了 GPS 定位、支付宝离线支付、NFC 地铁公交支付等多项功能，极大的增强了用户体验，吸引了广大消费者的兴趣。因此，在当今背景下，对智能手环行业发展和销量进行分析研究，具有非常重要的意义。

近年来，机器学习算法与数据挖掘算法不断发展，在各个领域都得到了广泛的应用。不同于传统的线性回归模型，机器学习中的支持向量机、随机森林、XGBoost 等算法，不需要对数据分布进行假定，而且具有良好的预测效果，因此具有良好的应用意义。由于本文研究的是智能手环销量预测，底层数据较复杂，需要进行特征工程，而且对预测准确度有一定的要求，故本文采用机器学习及数据挖掘方法进行建模。

本文采用网络技术爬虫法，共爬取某电子商务网站 3128 条手环商品数据，包含价格、品牌、综合评分、功能用途、GPS 定位等 28 个特征变量。首先，对数据进行预处理，并使用方差选择、Lasso 算法、Boruta 方法对变量进行综合筛选，将贡献率较低的特征因素进行剔除，最终筛选出 21 个特征用于建模预测；其次，从品牌、功能、用户画像三个方面，对智能手环进行描述性统计分析，了解消费者的情况及关注点；再次，使用支持向量机、随机森林、XGBoost 三种算法对模型进行预测。为比较模型的预测效果，本文使用十折交叉验证的方法对模型效果进行对比，结果 XGBoost 算法的 NMSE 值最小，即预测效果最好；最后，对三种预测模型的误差进行分析，结果表明 XGBoost 算法在各个指标上的表现都是最优的，因此使用 XGBoost 算法建立智能手环的销量预测模型是值得推广的。

关键词：销量预测；特征选择；支持向量机；随机森林；XGBoost 算法

Abstract

In today's social background, the rapid development of science and technology, people's spiritual pursuit of continuous improvement. Smart phones, smart computers and smart mobile wearable devices have gradually become the focus of people's attention. With the popularity of smart phones, wearable smart devices have become the darling of manufacturers and the new hot spot for the consumption of electronic products. Especially, the smart bracelet not only has the functions of message reminder, movement step meter, heart rate detection, sleep monitoring, etc, but also expands the functions of GPS positioning, alipay offline payment, NFC subway and bus payment, etc, which greatly enhances the user experience and attracts the interest of consumers. Therefore, in today's background, it is of great significance to analyze and study the development and sales volume of the smart bracelet industry.

In recent years, machine learning algorithms and data mining algorithms have been developed and applied in many fields. Different from the traditional linear regression model, support vector machine, random forest, XGBoost and other algorithms in machine learning do not need to assume the data distribution, and have good prediction effect, so it has good application significance. Since this paper studies the sales prediction of smart bracelet, the underlying data is complex, the feature engineering is needed, and the prediction accuracy is required, so this paper adopts the methods of machine learning and data mining for modeling.

This paper adopts the crawler method of network technology to crawl a total of 3128 bracelet commodity data of an e-commerce website, including 28 characteristic variables such as price, brand, comprehensive score, functional purpose and GPS positioning. Firstly, the data was preprocessed, and the variance selection, Lasso algorithm, and Boruta method were used to comprehensively screen the variables, and the feature factors with low contribution rate were eliminated. Finally, 21 features were selected for modeling and prediction. Secondly, descriptive statistical analysis is carried out on the characteristics of the smart bracelet from three aspects: brand, function and user portrait, so as to understand the situation and concerns of consumers. Thirdly, support vector machine, random forest and XGBoost are used to predict the model. In order to compare the prediction effect of the model, this paper USES the method of ten-fold cross validation to compare the model effect. Finally, the errors of the three

prediction models are analyzed, and the results show that the performance of XGBoost algorithm in each indicator is optimal, so it is worth promoting to establish the sales prediction model of smart bracelet by using XGBoost algorithm.

Key words: Sales forecast，Feature selection，Support vector machine，Random forest，XGBoost algorithm

目 录

摘要..... I
Abstract..... III
第 1 章 绪论..... 1
1.1 研究背景与意义..... 1
1.2 国内外研究现状..... 2
1.2.1 国内研究现状..... 2
1.2.2 国外研究现状..... 3
1.3 研究框架及技术路线..... 4
1.3.1 研究框架..... 4
1.3.2 技术路线..... 5
第 2 章 相关理论介绍..... 7
2.1 特征选择方法..... 7
2.1.1 特征选择方法简介..... 7
2.1.2 方差特征选择法..... 7
2.1.3 Lasso 特征选择法..... 7
2.1.4 Boruta 特征选择法..... 8
2.2 支持向量机..... 9
2.3 随机森林..... 10
2.3.1 随机森林基本原理..... 10
2.3.2 随机森林算法简介..... 11
2.4 XGBoost 算法..... 12
2.4.1 XGBoost 基本原理..... 12
2.4.2 XGBoost 算法推导..... 12
2.4.3 XGBoost 的优点..... 15
2.5 本章小结..... 16
第 3 章 数据预处理与特征选择..... 17
3.1 数据来源及获取..... 17
3.2 候选指标选取..... 17
3.3 数据预处理..... 18
3.3.1 缺失值处理..... 18

3.3.2 Box-Cox 正态性变换.....19
3.3.3 归一化处理.....19
3.3.4 数据预处理结果.....20
3.4 指标体系特征选择.....22
3.4.1 方差特征选择法.....23
3.4.2 Lasso 特征选择法.....24
3.4.3 Boruta 特征选择法.....26
3.4.4 变量综合对比选择.....28
3.5 本章小结.....29
第 4 章 智能手环销量建模与预测.....31
4.1 分析思路.....31
4.2 智能手环描述性统计分析.....31
4.2.1 智能手环品牌及功能分析.....31
4.2.2 智能手环用户画像分析.....32
4.3 基于支持向量机的预测模型.....33
4.3.1 参数选择与模型建立.....33
4.3.2 模型拟合效果.....34
4.4 基于随机森林的预测模型.....35
4.4.1 确定参数 ntree.....35
4.4.2 确定参数 mtry.....36
4.4.3 变量重要性排序.....36
4.4.4 模型拟合效果.....38
4.5 基于 XGBoost 的预测模型.....39
4.5.1 参数说明.....39
4.5.2 参数调优与确定.....39
4.5.3 特征重要性排序.....41
4.5.4 模型拟合效果.....42
4.6 模型对比评估.....43
4.6.1 十折交叉验证.....43
4.6.2 预测误差评估.....44
4.7 本章小结.....45
结论与不足.....47
（一）结论与建议.....47

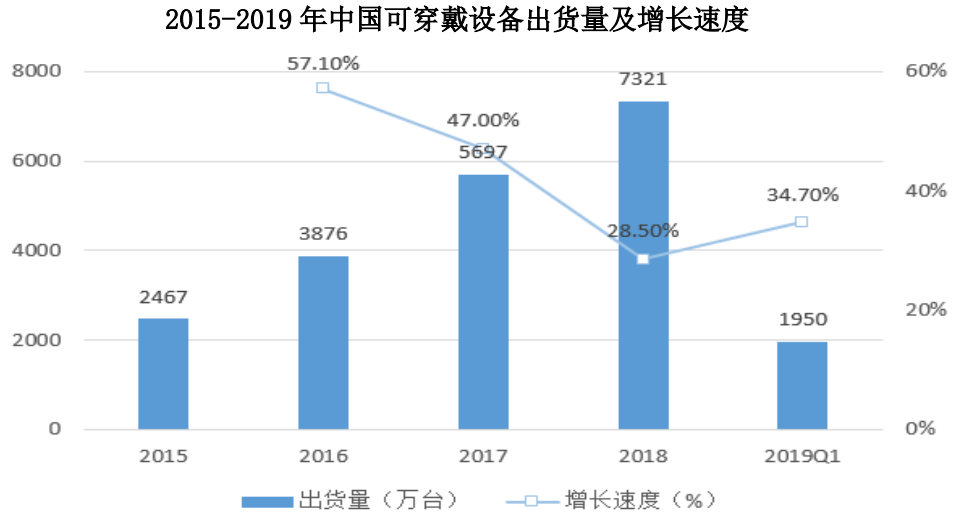
（二）存在的不足.....48
参考文献.....49
致谢.....53

第 1 章　绪论

1.1 研究背景与意义

随着我国科技的进步以及经济的发展，人们的生活品质也在不断提高，智能手机、平板电脑、智能手环等移动性可穿戴设备风靡全球。尽管从整体上看，近两年我国智能手环行业发展迅速，但是智能手环市场发展参差不齐，各品牌产品的宣传效果与自身实际性能存在差异，因此消费者存在一定的信息不对称性，使得消费者的选购具有一定的困难。比如现在市场上的小米、华为、fitbit、荣耀、乐心等占据了市场的半壁江山，各个品牌主打的功能也参差不齐。在这种背景下，用户应该关注哪些方面的功能，企业又该关注哪些销量影响因素，为客户提供更好的价值体验就尤为重要[1]。

根据 IDC 统计数据，2015-2018 年我国智能可穿戴设备呈现直线上升的趋势，2018 年智能可穿戴设备出货量达到 7321 万台，同比增长 28.5%。特别地，2019Q1 出货量达到 1950 万台，同比增长 34.7%。与不断下滑的智能手机市场相比，我国智能可穿戴设备的种类在不断增加，产品技术也不断成熟，用户体验及功能在不断完善，可穿戴设备的发展已经进入了井喷期，可穿戴设备市场保持着高速增长的趋势。具体数据如下图 1-1 所示。



来源：IDC 数据

图 1-1 2015-2019 年中国可穿戴设备出货量及增长速度图

Figure 1-1 Chart of wearable device shipments and growth rate in China from 2015 to 2019

随着人们对于健康生活的追求越来越高以及互联网技术的发展，健康类智能产品逐渐成为人们的关注对象。同时，我国近几年电商行业发展迅猛，网上购物也早已融入人们的生活。因此，本文以网上购物为平台，对智能手环进行销量预测及因素分析。从商家的角度来说，销量预测可以方便配置货物，最大效度的实现供应。而因素分析则可以为制造商或者销售商提供可靠的参考指标，通过对手环销量重要影响因素的分析，实现针对性的销售策略，从而提高自身的市场竞争力，达到企业利益最大化。

另一方面，若能对手环销量进行较好的建模预测及因素分析，便可以更好的掌握手环市场的发展方向，为用户提供更加优质的产品功能及服务。用户也可以更好地在众多品牌型号的手环中进行挑选与购买，提升用户体验价值。因此，本文对于智能手环的行业发展分析及销量建模预测有较好的应用价值。

1.2 国内外研究现状

随着社会经济的发展和科技的进步，智能设备的竞争愈加强烈，用户越来越注重智能产品的价值体验，企业也不断的挖掘影响销量的重要因素。近年来，国内外学者对用户购买愿意的原因进行分析，并对评论进行在线挖掘[2-5]，利用信度效度分析、BP 神经网络、文本挖掘和 Logistic 回归分析等方法进行深度分析，挖掘意愿购买的重要因素，在理论和实证方面取得很大的进展。

1.2.1 国内研究现状

孙晶晶[6]等通过对国内外不同品牌的智能手环进行研究，从外观设计、运动计步、血压监测、心率监测等方面进行横向对比研究，并以综合评分的形式对各项指标进行评判，最终全方面反映出不同品牌智能手环的综合性价比，可为消费者购买智能手环提供重要的参考。

罗步操[7]在学生对智能手环接受度方面进行了调查分析，调查内容包括学生对智能手环的使用情况、功能选择、了解程度、功能以外的要素需求、佩戴智能手环的顾虑情况、价格的接受区间以及学生对佩戴智能手环的主观意愿等，对于智能手环的销量预测起到很好的因素分析作用。

刘大为，蔡赛凤[8]对用户采纳模型进行研究，首先提出基于 UTAUT 模型的用户采纳假设，其次使用 SPSS 及 AMOS 软件对模型进行信度分析、回归分析及验证性因子分析，并且使用 KMO 及 Bartlett 球形检验对所有变量进行验证，最

后总结研究成果，对本文研究手环的影响因素有一定的作用。

吴江[9]等在可穿戴设备在线评论主题挖掘研究中，对华为、小米、三星、Fitbit 等目前最受关注的四大智能手环品牌进行分析，通过 R 语言利用 LDA 模型对评论文本进行挖掘，并且使用 LDA 模型和 Gibbs 抽样来进行主题聚类，探究不同手环品牌和不同时间阶段评论中的主题差异，对用户满意度进行深度分析。

陈华珍[10]等在传统 BP 神经网络算法的基础上，提出双适应 BP 神经网络算法，从学习率自适应调整法和自适应动量法两个方面进行分析。同时利用智能手环多传感器数据融合原理，对 BP 神经网络模型进行参数设计，进行实例验证分析。

王林[11]等对用户使用智能手环的影响因素进行研究，采用问卷调查的形式进行统计分析。借助 SPSS 工具，进行主成分分析、因子分析、KMO 与 Bartlett 检验。最后，对用户特征进行描述性统计，相关分析和回归分析。

朱振涛[12]等运用文本挖掘和 Logistic 回归分析研究了智能手环在线评论有用性的影响因素。利用 R 语言进行描述性统计，回归模型分析，并建立 Logistic 模型对品牌声誉和评论有效性的关系进行研究。

1.2.2 国外研究现状

Venkatesh[13]等通过整合不同模型提出 UTAUT 模型，构筑了四个核心概念，而性别、年龄、经验、实际行为则会受到直接影响，这些调节变量的引入增强了 UTAUT 的解释能力，有助于分析智能手环的影响因素。

Netzer[14]等采用了文本挖掘方法，并将其与语义网络分析工具相结合。对于市场结构，从基于用户的内容数据和基于传统的调查数据进行综合对比，以确定有效性并突出有意义的差异，对于智能手环的行业分析有很大的帮助。

Jung[15]等对智能可穿戴设备的影响因素进行调查分析，使用联合分析的方法，对用户所关注的特征进行因素分析，发现屏幕设计、通讯功能、价格、屏幕尺寸和品牌是影响用户使用意愿的重要因素，对于本文智能手环的影响因素研究具有重要意义。

Kim[16]等采用 AMOS 统计软件对采集的数据进行验证性因子分析(CFA)和结构方程建模(SEM)。研究结果表明，智能手表的亚文化吸引力和成本分别是用户态度和使用意愿的显著前因。通过对智能手表使用情况进行系统预测的首批学术尝试之一，对未来可穿戴技术的采用具有指导意义。

Yang[17]等关于用户是否愿意继续使用可穿戴设备的影响因素研究中，基于

感知价值建立了用户接受度模型，并使用 TAM 模型进行建模分析，得出感知娱乐性和主观规范性是影响用户意愿的重要因素，并且若有同类产品的使用经历，也将会对用户的使用意愿产生显著影响，这些对于智能手环影响因素的分析具有现实意义。

Deng[18-20]等对于用户使用满意度进行研究，使用调查问卷的方式获取用户对可穿戴设备的使用体验。研究发现，品牌与功能是用户最关注的方向，同时价格与评价也是影响用户使用意愿的重要因素，对于本文智能手环的因素分析具有较好的指导意义。

通过对比国内外研究学者有关智能手环的研究结果，可以发现大多数文章只是对于影响可穿戴产品的因素进行分析，而对于手环的影响因素及销量预测分析相对较少[21]。本文将着重分析智能手环的行业发展及销量预测问题，在国内外研究学者的理论基础上，更加深层次的挖掘影响手环销量的重要指标，建立多个预测模型并进行综合对比，确定最终影响因素及预测模型。

1.3 研究框架及技术路线

1.3.1 研究框架

第一章，绪论。对近两年我国智能手环行业发展的背景、目的及意义进行阐述，并简要介绍了国内外研究学者对智能手环行业发展及销量问题的研究成果，为本文智能手环的销量预测研究奠定良好的基础。

第二章，相关理论介绍。对于本文涉及的相关理论基本原理及算法进行简要介绍，包括 Box-Cox 变换、方差特征选择法、Lasso 特征选择法、Boruta 特征选择法、以及支持向量机回归算法、随机森林算法、XGBoost 算法，为实证分析做铺垫。

第三章，数据预处理与特征选择。本章首先对数据进行预处理，主要包括缺失值处理、Box-Cox 正态性变换、数据归一化处理和指标量化等方面；然后采用方差过滤法、Lasso 特征选择法和 Boruta 特征选择法进行特征选择，通过三种方法的综合对比，确定最终用于建模的特征。

第四章，智能手环销量建模与预测。本章首先对手环进行描述性统计分析，然后分别采用支持向量机回归算法，随机森林算法，以及 XGBoost 算法进行销量建模，并以十折交叉验证的方法对于三种模型的性能进行评估；最后对手环销量预测进行预测误差检验，将三种模型进行综合对比，选出最佳预测模型。

最后，基于上述分析进行全文总结。主要对本文的研究内容及成果进行总结，然后对制造商和企业提出合理化建议，最后指出本次研究的不足及有待提高的地方。

1.3.2 技术路线

为了更加直观的展示本文建模分析的整体思路，制作出技术路线图，如下图 1-2 所示。

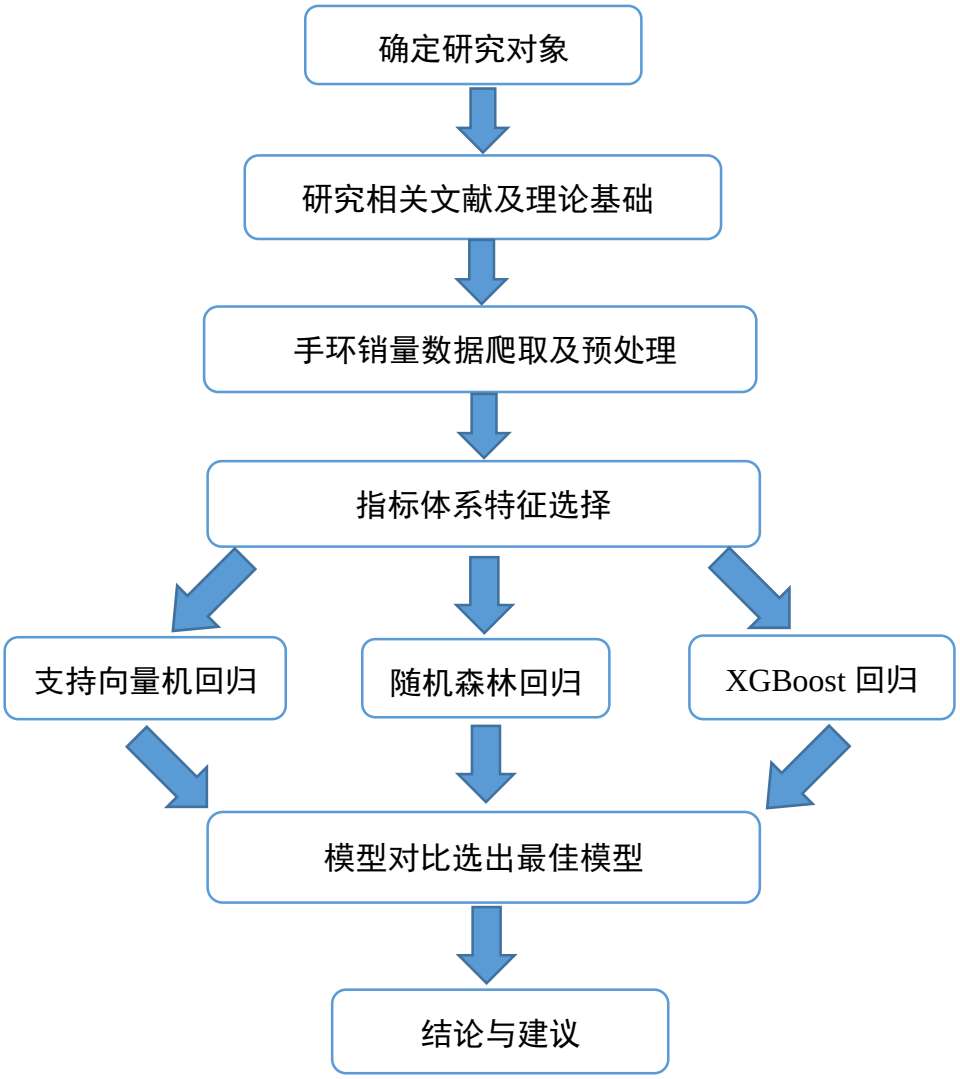


图 1-2 技术路线图

Figure 1-2 Technical road map

第 2 章 相关理论介绍

2.1 特征选择方法

2.1.1 特征选择方法简介

特征选择(Feature Selection)，是指从所有的特征中选择出有意义且对模型具有较好解释能力的特征子集[22]。特征选择可以有效的避免将所有特征都引入模型进行训练的情况，从而可以较好的提高模型预测的精确度。具体方法如下：

(1) 过滤法：过滤法的基本原理是使用发散性或相关性指标对每个特征进行指标评分，选择分数大于阈值的特征或者选择前 K 个分数最大的特征。

(2) 嵌入法：嵌入法是一种让算法自己决定使用哪些特征的方法，即特征选择和算法训练同时进行。嵌入法的应用，可以使用一些机器学习算法和模型进行训练以获得每个特征的权重系数。根据权重系数，我们从大到小选择特征，这些权重系数通常表示特征对模型的贡献或重要性程度。

(3) 包装法：包装法是一种基于贪婪搜索的算法，它会评估特征的所有可能组合，并为特定的机器学习算法选择能产生最佳结果的组合。该方法可以有效的选择出最佳特征组合，但是在特征集非常大的情况下，计算成本较高，速度较慢。

2.1.2 方差特征选择法

方差齐性检验(ANOVA)，又称 F 检验法，主要通过比较两组数据的方差，以确定他们的精密度是否有显著性差异。方差齐性检验的本质是寻找两组数据之间的线性关系，其原假设是“数据不存在显著的线性关系”。因此，方差齐性检验法是一种解释因变量与每个特征之间线性关系的过滤方法。方差齐性检验法既可以用来进行回归分析也可以做分类，因此包括 F 检验回归和 F 检验分类两种检验方法。其中 F 检验回归用于因变量是连续型变量的数据，而 F 检验分类用于因变量是离散型变量的数据。

2.1.3 Lasso 特征选择法

Lasso(Least Absolute Shrinkage and Selection Operator)方法属于嵌入法，它是

一种通过压缩系数进行变量选择的方法[23]，即通过加入惩罚项使得那些基本没有影响或影响较小的自变量系数趋近于 0，从而既能实现变量选择，又可以对模型进行参数估计[24]。Lasso 方法对于不显著的变量进行删除，可以显著降低模型的偏差，因此在多种共线性问题中有较好的应用。Lasso 方法不仅对于传统方法无法实现变量选择的缺点进行改进，而且在实现变量选择的同时使模型更具解释性[25]。

Lasso 方法由 Tibshirani 提出[26-28]，其基本算法如下：

假设 Y 与$x_1, x_2 \ldots, x_n$之间存在线性回归关系，则有：

$$Y = X\beta + \varepsilon, \tag{2-1}$$

其中$Y = (y_1, y_2, \ldots y_n)^T$,$X = (x_1, x_2, \ldots x_n)$,$x_1 = (x_{1t}, x_{2t}, \ldots x_{nt})^T, i = 1,2, \ldots p$，$n$ 为观测值个数，$\beta = (\beta_1, \beta_2, \ldots \beta_n)^T$为待估参数，$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$为误差向量，且满足$E(\varepsilon) = 0, Cov(\varepsilon) = \sigma^2 I$。

对上述线性模型中的参数β进行 Lasso 估计，确定β的估计值，公式如下：

$$\hat{\beta} = arg\min_{\beta}\left(\|Y - X\beta\|^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right), \tag{2-2}$$

其中，$\lambda\sum_{j=1}^{p}|\beta_j|$为惩罚项，$\lambda \geq 0$为调和参数，代表惩罚力度。

Lasso 方法的核心思想是：当$\lambda = 0$时，不对模型系数进行压缩处理。当λ足够大时，便会得到一个零模型，即所有系数的估计值都会被压缩到 0。因此，通过调节λ的值，Lasso 可以得到包含不同变量个数的模型，从而实现变量选择的目的。特别的，当一部分自变量是真实有效的且其他自变量系数非常小或者等于 0 时，Lasso 方法更为适用。

2.1.4 Boruta 特征选择法

Boruta 特征选择是以随机森林为基础的一种包装算法，因此 Boruta 算法与随机森林具有相同的思想。Boruta 算法通过向模型中引入随机性，并且在随机样本集合中进行建模，来实现降低相关性和随机波动产生误差的目的，从而得到特征重要性的排序结果[29]。Boruta 算法的具体实现过程如下[30]：

(1) 在原始数据集中对所有特征均引入随机成分(阴影属性)，得到一个扩展数据集，并将随机森林分类器方法应用于该拓展数据集，获得特征重要性的 Z 分数指标。

(2) 找出比最大 Z 分值(Z_{max})低的阴影属性，并将这些属性标记为+1，记为 hits。

(3) 当阴影属性的重要性无法确定时，使用 Z_{max} 对其进行双侧显著性检验，并使阴影属性分值小于 Z_{max} 的特征归为 0。

(4) 依据伯努利公式

$$P = \binom{n}{k} p^k (1-p)^{n-k}, \ p = \frac{1}{2}. \tag{2-3}$$

对步骤(3)中的 hits 属性进行计算，若概率值低于 0.01，则将其记为“重要”属性。反之，将其记为“不重要”属性，并从数据集中剔除。

(5) 重复上面的步骤，当属性全部被标记为“重要”或“不重要”，或者当算法达到随机森林预设的极限值时，停止运算。

2.2 支持向量机

支持向量机(Suport Vector Machine,常简称为 SVM)[31]，是一个监督式学习的方式。SVM 是一种广义的线性分类器，可以最小化经验误差，同时最大化几何边缘区域，因此也称为最大边缘区域分类器。SVM 可分为两类，若因变量为连续型变量，为支持向量回归机(SVR)；若因变量为分类型变量，则为支持向量分类机。

支持向量回归机(SVR)的基本原理如下[32-33]：

给定训练样本$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}, y_i \in \mathbb{R}$，其中$x_i$为输入值，$y_i$为对应的输出值。建立如下回归模型：$f(x) = \omega^T \cdot x + b$，其中$\omega$为自回归系数，$b$ 为误差值。SVR 假设我们能容忍 $f(x)$与 y 之间的最大偏差为ϵ，因此当 $f(x)$与 y 之间的距离大于ϵ时，才会进行模型损失的计算。如下图 2-1 所示，SVR 构建了以 $f(x)$为中心，2ϵ为宽度的一个间隔区域。如果模型样本点落在该区域中，则被认为预测是正确的。反之，预测是不正确的。

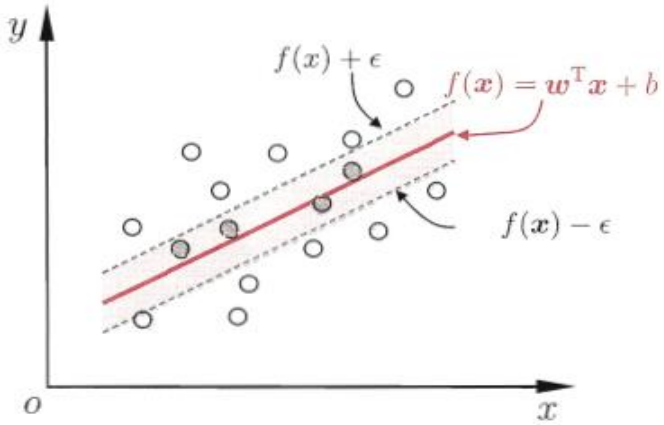


图 2-1 支持向量回归原理图

Figure 2-1 Support vector regression schematics

于是，SVR 的问题可形式化为

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\ell_\varepsilon(f(x_i)-y_i), \tag{2-4}$$

其中 C 为正则化常数，ℓ_ε是ϵ −不敏感损失$(\epsilon - insensitive)$函数

$$\ell_\epsilon(z)=\begin{cases}0 & ,\ |z|\le\epsilon;\\ |z|-\epsilon, & |z|>\epsilon.\end{cases}$$

引入松弛变量$\boldsymbol{\xi_i}$和$\hat{\boldsymbol{\xi}}_{\boldsymbol{i}}$，可将式(2-4)重写为

$$\min_{\omega,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|\omega\|^2 + C(\boldsymbol{\xi_i}+\hat{\boldsymbol{\xi}}_{\boldsymbol{i}}). \tag{2-5}$$

接下来，通过引入拉格朗日乘子$\mu_i \ge 0$，$\hat{\mu}_i \ge 0$，$\alpha_i \ge 0$，$\hat{\alpha}_i \ge 0$，使用拉格朗日乘子法，得到如下拉格朗日函数：

$$\begin{aligned}&L(\omega,\ b,\ \alpha,\ \hat{\alpha},\ \xi,\ \hat{\xi},\ \mu,\ \hat{\mu})\\ &=\frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i) - \sum_{i=1}^{m}u_i\,\xi_i - \sum_{i=1}^{m}\hat{\mu}_i\hat{\xi}_i\\ &+\sum_{i=1}^{m}\alpha_i\,(f(x_i)-y_i-\epsilon-\xi_i)\\ &+\sum_{i=1}^{m}\hat{\alpha}_i(y_i-f(x_i)-\epsilon-\hat{\xi}_i).\end{aligned} \tag{2-6}$$

再对上述对偶问题求偏导，即可得到 SVR 的对偶问题：

$$\max_{\alpha,\ \hat{\alpha}}\sum_{i=1}^{m}y_i\,(\hat{\alpha}_i-\alpha_i) - \epsilon(\hat{\alpha}_i+\alpha_i) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i-\alpha_i)\,(\hat{\alpha}_j-\alpha_j)x_i^T x_j$$

$$s.t.\ \ \sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)=0,\ 0\le\alpha_i,\ \hat{\alpha}_i\le C. \tag{2-7}$$

求得 SVR 的解形如(2-8)所示：

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\,x_i^T x + b. \tag{2-8}$$

2.3 随机森林

2.3.1 随机森林基本原理

随机森林(Random Forest，简称 RF)是 Bagging 的一个扩展变体，不仅能进行数据分类，还能进行数据回归[34]。RF 建立基于决策树的 Bagging 集成，将随机属性选择引入决策树进行模型训练。实际训练中，对划分属性进行选择时，传

统决策树基于当前结点的属性集，选择最佳属性(假设有 d 个属性)；而在 RF 中，对于每一个基决策树的结点，随机地从结点的属性集中选择一个包含 k 个属性的子集，再从这个子集中选择一个最优属性进行划分。

随机森林的基本思想是：首先，通过使用 bootstrap 重复抽样方法，在训练集中抽取 k 个训练子集；其次，对于抽取的训练子集分别建立决策树模型，待测试样本便会在这 k 个决策树模型中进行预测，从而得到 k 个分类结果；最后，对 k 个决策树模型进行综合投票，根据投票结果确定最终的预测结果[35]。随机森林将 CART 决策树作为基学习器，但是与 CART 不同的是，随机森林在训练中会产生多棵决策树。预测变量基于建好的多棵决策树进行模型预测，当对预测变量进行分类预测时，随机森林选择将大多数决策树的分类结果作为最终预测结果；当对预测变量进行回归预测时，则返回所有决策树预测结果的平均值。

2.3.2 随机森林算法简介

假设 M 为样本特征个数，m 为小于 M 且大于 0 的整数。随机森林的具体算法过程图如下 2-2 所示：

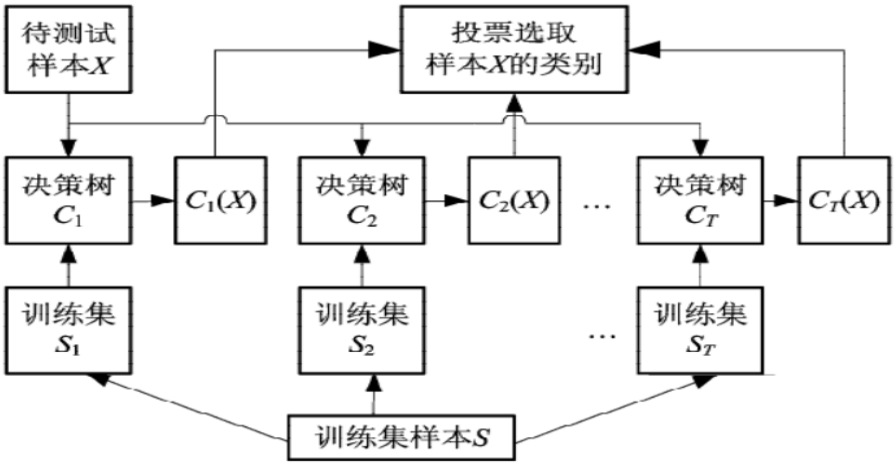


图 2-2 随机森林算法过程图

Figure 2-2 Process diagram of random forest algorithm

具体算法步骤如下[35]：

(1) 对训练集 S，使用 bootstrap 重复抽样方法进行抽样，随机抽取 T 个训练子集，记为 S_1，S_2，…，S_T；

(2) 以这 T 个训练子集为基础，分别建立 T 棵分类决策树 C_1，C_2，…，C_T，并且在非叶子节点选择特征之前，选取 $m(m=\sqrt{M}$或 $m=M/3)$个特征作为此节

点的分裂特征数。同时，以最小离差平方和的分枝形式对该节点进行分裂，生成最佳决策树。

(3) 将待测试样本 X 分别代入 T 棵生成好的决策树中进行模型预测，得到 T 个模型预测结果，记为 $C_1(X)$，$C_2(X)$，…，$C_T(X)$。

(4) 依据一定的方法准则，对(3)中的 T 棵决策树进行综合投票，最终确定测试样本的最佳预测结果。

2.4 XGBoost 算法

2.4.1 XGBoost 基本原理

XGBoost 是由 Chen 等[36]于 2015 年提出，全名叫做 eXtreme Gradient Boosting(也称极端梯度提升)。XGBoost 所应用的算法是 GBDT(Gradient Boosting Decision Tree)的改进，既可以用于分类也可以用于回归问题中。

XGBoost 的基本原理是：首先，利用二阶泰勒公式对损失函数进行形式变换；其次，选取树模型复杂度作为正则项，引入目标函数中；再次，在模型训练时，以随机森林算法的思想为基础，通过随机抽样的方式进行每次模型的迭代，并使用部分随机样本的部分特征进行模型训练；最后，在多核 CPU 并行运算的基础上进行建模与预测，可以很好的提高模型预测精度与算法运行速度。此外，XGBoost 算法还可以对特征的重要性进行排序，可以有效的展示模型的重要特征因素[37]。

2.4.2 XGBoost 算法推导

本文参考叶倩怡[38]、张培荣[39]等人的研究，主要从模型函数、目标函数、模型优化、树的生成四个方面来对 XGBoost 算法模型的推导过程进行介绍。

1、模型函数

对于一个给定的有 n 个样本的数据集$D = (x_i, y_i)$，树的集成如下式所示：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)，\ f_k \in F. \tag{2-9}$$

上式(2-9)中，$F = \{f(x) = \omega_{q(x)}\}(q: R^m \to T, \omega \in R^T)$为树的集合空间，$x_i$表示第 i 个变量，$\hat{y}_i$为样本x_i的预测结果，f_k为第 k 个叶子结点的预测分值，x 表示分类到某叶子节点，$q(x)$表示将样本在叶子节点上分类，T 表示叶子节点数量，

ω表示叶子节点的分数，故$\omega_{q(x)}$表示在叶子节点上的预测分值。

2、目标函数

目标函数包括两部分，公式如下所示：

$$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k). \tag{2-10}$$

上式(2-10)中，Θ=$\{\omega_j | j = 1,2,\dots,d\}$为线性模型系数。第一部分$l(y_i, \hat{y}_i)$为预测值$\hat{y}$与真实值$y_i$之间的训练误差；第二部分是每棵树的复杂度之和，即$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$，其中$\gamma$与$\lambda$均为参数。

3、模型优化(梯度树提升)

对于上式(2-9)目标函数的优化，采用加法训练(Additive Training)的学习方式，即通过向模型中加入新的函数来实现迭代，得到最终的预测模型[40]，迭代过程如下：

$$\hat{y}_i^{(0)} = 0,$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i),$$

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i),$$

...

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \tag{2-11}$$

上式(2-11)中，$\hat{y}_i^{(t)}$为第 t 轮的模型预测值，$\hat{y}_i^{(t-1)}$为第 t-1 轮的模型预测值，$f_t(x_i)$为第 t 轮的新增函数。每次迭代可以降低实际值与预测值之间的误差，从而提高模型精度，实现目标函数的最优化。进一步的，将式(2-11)代入式(2-10)中，目标函数改写如(2-12)所示：

$$\begin{aligned} Obj^t &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)\right) + \Omega(f_i) + C. \end{aligned} \tag{2-12}$$

对于一般情况下，通过泰勒展开可近似得到目标函数。泰勒展开公式：$f(x+\Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x^2$，将$f_i(x_i)$看作$\Delta x$，则原目标函数改写为：

$$Obj^t = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)\right) + \Omega(f_i) + C$$

$$\approx \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) f_t(x_i) + \frac{1}{2} \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) f_t(x_i)^2 \right] + \Omega(f_t) + C.$$

令 $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$，$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$，且对于第 t 棵树，$l(y_i, \hat{y}_i^{(t-1)})$为常数，同时去掉常数项 C，目标函数可改写为：

$$Obj^t \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t). \tag{2-13}$$

上式(2-13)中，$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda ||\omega||^2$表示树的复杂度，这里将复杂度写成 $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} w_j^2$,且由上文可知，$f(x) = \omega_{q(x)}$，此时目标函数如下所示：

$$Obj^t \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} w_j^2$$

$$= \sum_{i=1}^{n} \left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega^2_{\omega_{q(x_i)}} \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} w_j^2. \tag{2-14}$$

在上式（2-14）中，T 为第 t 棵树中总叶子节点的个数，定义每个叶子节点 j 上的样本集合为$I_j = \{i | q(x_i) = j\}$，则目标函数进一步改写为：

$$Obj^t = \sum_{i=1}^{n} \left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega^2_{\boldsymbol{\omega}_{\boldsymbol{q}(\boldsymbol{x}_i)}} \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} w_j^2$$

$$= \sum_{i=1}^{T} \left[(\sum_{i \in I} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I} h_i)\omega_j^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{t} w_j^2$$

$$= \sum_{i=1}^{T} \left[(\sum_{i \in I} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I} h_i + \lambda)\omega_j^2 \right] + \gamma T \ . \tag{2-15}$$

接下来，令$G_j = \sum_{i \in I} g_i$，$H_j = \sum_{i \in I} h_i$，则最终的目标函数为：

$$Obj^t = \sum_{i=1}^{T} \left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2 \right] + \gamma T. \tag{2-16}$$

上式(2-16)对ω_j求偏导，并使其导数为 0，求解得：

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}.$$

将ω_j^*带入目标函数，得到：

$$Obj^*(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i+\lambda} + \gamma T. \quad (2\text{-}17)$$

上式(2-17)得到的目标函数可以作为 XGBoost 的评分函数，且该函数的应用范围较广。通过此打分函数，可以计算得出最优的树，这是 XGBoost 预测准确率较高的一个重要原因。

4、树的生成

可以借助贪心算法来进行叶子节点的分裂，每次分裂都会得到一个增益值，通过选取最大增益值来确定最终的树结构。增益值的计算公式如下：

假设I_L和I_R是拆分后左右子树分裂点的总集合，拆分后的损失为：

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma. \quad (2\text{-}18)$$

上式(2-18)中，$\frac{G_L^2}{H_L+\lambda}$表示左子树的分数，$\frac{G_R^2}{H_R+\lambda}$表示右子树的分数，$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$表示不进行分割时的分数，$\gamma$表示新叶子节点加入时所带来的复杂度代价。

Gain 可以用来确定叶子节点是否进行分割，且 Gain 越大表示分割所获得的信息增益越大(即损失减少的越多)，故一般选择增益值较大的进行分割[39]。

2.4.3 XGBoost 的优点

(1) 高效性。借助于 OpenMP，XGBoost 自动将单个 CPU 的多个内核用于并行计算。 它结合了线性模型和树学习算法，这使 XGBoost 至少比现有的梯度上升实现至少 10 倍的提升。

(2) 准确性。与传统的 GBDT 相比，XGBoost 模型增加了对模型复杂度的控制，并且后期对其进行了修剪，从而使学习的模型更加不容易过拟合，更好的提高了模型的精度。

(3) 灵活性。XGBoost 可以自定义目标函数与评估函数，与传统的 CART 相比，XGBoost 不仅可以用来分类预测，还可以进行回归预测，具有较强的灵活性。

(4) 自主性。如果样本存在缺失值，XGBoost 可以进行自主学习，对数据的分裂方向进行自动预测，因此减少了数据的流失，提高了模型预测的精度[41]。

2.5 本章小结

本章对于论文中涉及的相关理论进行介绍，首先介绍了 Box-Cox 数据正态性变换方法；其次阐述了特征选择的三种方法，包括方差特征选择法、Lasso 特征选择法和 Boruta 特征选择法；最后对三种建模方法理论进行介绍，包括支持向量机回归算法、随机森林算法和 XGBoost 算法，为实证分析中手环的销量预测建模奠定良好的基础。

第 3 章 数据预处理与特征选择

3.1 数据来源及获取

近几年，电商行业发展迅速，人们的购物方式逐步从线下向线上进行转移。首先，网上购物方便快捷，节省了大量的时间与成本；其次，网上购物模式不断新颖化与多样化，消费者可以在网上购买任何自己喜欢的东西；再次，网上购物不断被国家及社会认可，品质及安全保证也逐步提高；最后，网上购物可以给消费者提供详细的商品信息，包括价格、品牌、功能用途等，同时消费者还可以看到商品评价、店铺评价等信息，对于用户购买商品具有很好的参考意义。

通过以上对于网上购物的分析，本文基于 python 工具，爬取某商城网站智能手环的相关信息，包括手环价格、销量、参数、店铺信息等，用于对智能手环的影响因素分析及销量预测建模。本次数据爬取，获取的是手环截面数据，避免了因为时间因素带来的手环销量及相关信息的变化。同时，也缩短了数据的获取时间，避免了因为时间问题带来的不必要错误。

为更好的对手环销量的影响因素进行分析，及销量进行建模预测。本文尽可能的爬取了每个手环商品的全面相关信息，但是信息的爬取是一个比较复杂的过程，各大电商网站也设有反爬技术，故手环的爬取及数据的整理耗费了很大一部分时间。因此，本文对于手环数据的采集工作量相对较大，最终爬取 3128 条智能手环的商品数据。

3.2 候选指标选取

本文通过深度分析智能手环商城网页商品信息的可获取性、适用性及数据量化的难易程度，确定用于建模的基本指标。同时，综合国内外参考文献及相关书籍涉及的手环影响因素，确定最终的侯选指标体系。因为本文研究的是手环的销量预测，故将手环销量作为因变量，并将候选指标体系大致分为五大类，共 28 个指标，如下表 3-1 所示。

表 3-1 侯选指标

Table 3-1 Candidate indicators

变量类型	变量名称
基本参数	价格、品牌、适用人群、好评度、评价详情、续航时间
功能参数	NFC 支付、GPS 定位、泳姿识别、运动模式识别、功能用途、防水等级、蓝牙通话、娱乐功能
外观屏幕	屏幕尺寸、触控方式、屏幕显示、自动调节亮度
店铺情况	综合评分、商品评分、物流评分、售后评分、店铺类型
其他	广告、优选服务、重量、连接方式、腕带材料

3.3 数据预处理

数据预处理是从数据中找出不准确或者不适合用于建模的数据，这是进行建模至关重要的一步。由于本文中智能手环的商品数据是通过商城网页爬取的，因此可能存在数据缺失、数据异常、数据量纲不同、数据偏态等情况。这些数据都会对建模产生影响，显著降低预测的准确性。因此，为提高模型预测精度，需要进行数据预处理。

3.3.1 缺失值处理

在建模过程中，由于数据可能通过爬虫等多种方式获得，因此记录中可能存在部分特征的部分数据缺失的情况。当数据量较大时，可以通过舍弃部分不重要的数据来进行建模；但是当数据量较小时，或者当缺失数据是较为重要的特征时，往往不能通过简单删除来操作。因此，缺失值处理在数据预处理中至关重要。下面对常用的缺失值处方法进行介绍：

(1) 删除法

删除法是处理缺失值最常用且最简单的方法。删除法通常包括两个方面：删除特征或者删除样本。删除特征适用于该特征不重要或者该特征的缺失数据较为严重的情况，而删除样本则适用于数据量较大，缺失数据所占比例较小的情况。

(2) 替换法

替换法是除删除法外，较为常用的一种缺失值处理方法。替换法适用于特征较为重要，不可以直接删除的情况。对于分类特征来说，一般使用特征的众数来代替；而对于数值型特征来说，经常使用数据的平均数或者中位数来代替。

(3) 插补法

删除法与替换法虽然比较常用且简单易操作，但是可能会使数据信息缺失甚至产生有偏的情况。插补法便很好的解决了以上的问题，使数据可以充分被利用，回归插补与多重插补是插补法最常用的两种形式。回归插补是指将缺失值所在的特征作为因变量，其他特征作为自变量，通过建立回归模型，来对缺失值进行预测。多重插补指利用一系列可能的值去代替缺失值，然后进行多次替代产生若干个数据集，最后将多个数据集进行综合，获得整体参数的估计值。

3.3.2 Box-Cox 正态性变换

Box-Cox 正态性变换是一种经典的数据偏态处理方法。当连续型因变量或自变量不满足正态分布时，可以使用 Box-Cox 进行正态化处理。Box-Cox 变换之后，可以一定程度上减小残差和预测变量的相关性。

Box-Cox 变换相比普通的数据变换方式坚持正态性假设，通过各种数据转换函数将非正态数据转换为正态，常用的变换方式有以下几种：(1)倒数变换；(2)平方根变换；(3)对数变换；(4)平方根后取倒数；(5)平方根后再取反正弦；(6)幂变换。

Box-Cox 变换的公式如下[42]：

假设 y 为正的随机变量，Box-Cox 变换为：

$$y \to y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0; \\ ln(y), & \lambda = 0\ . \end{cases} \tag{3-1}$$

上式(3-1)中，λ为待定变换参数。当响应变量和自变量联合服从正态分布时，线性模型假设成立。基于此，Box-Cox 变换的基本思想是变换后的响应变量 y 在自变量 x 给定下的条件分布为正态分布：

$$y^{(\lambda)}|x \sim Normal.$$

Box-Cox 的变换形式由λ决定，一般可以使用最大似然估计或者贝叶斯方法来估计λ的值。

3.3.3 归一化处理

数据归一化是对“无量纲化”处理的常用工具，它可以消除不同特征的量纲，使得特征可以以一致、规范的形态进入模型，在降低偏差的同时，提高模型的精度。同时，在确保数据完整的条件下，数据归一化可以将数据进行最大程度的缩小，不仅可以降低内存和提高效率，一定程度上还可以提高模型预测的精度。

下面对归一化方法进行介绍，包括最值法与标准化法[43]：

(1) 最值法

最值法是指数据按照最小值进行中心化，然后按照极差进行缩放，数据移动了最小值个单位且被缩放到[0,1]之间。公式如下：

$$\hat{X}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad , \tag{3-2}$$

其中$\hat{X}_i$为归一化后的值，x_i为初始值，x_{max} 、x_{min}分别是初始数据的最大值和最小值。

(2) 标准化法

标准化法是指将数据按照均值进行中心化，然后按照标准差进行缩放，最终得到均值为 0，方差为 1 的标准正态分布。具体公式如下：

$$z = \frac{x-u}{\sigma} \quad , \tag{3-3}$$

其中，u 为均值，σ为标准差。

一般情况下，标准化法比最值法的应用更为广泛，因为最值法对异常值较为敏感。当模型中不涉及距离度量，协方差计算或者数据需被压缩到特定空间时，最值法有较好的应用。

3.3.4 数据预处理结果

1、缺失值处理结果

由于本文智能手环数据为网络爬虫得到，因此存在大量的缺失值。首先，在保证信息完整度的情况下，对于缺失值较多的样本进行整条删除，最终得到 2546 条数据。其次，对于数值型样本数据，用缺失值所在变量的均值代替，如物流评分、售后评分等。最后，对于分类型样本数据，使用缺失值所在分类变量的众数代替，如广告、泳姿识别等。

2、正态性处理结果

本文的因变量手环销量数据存在一定的偏差性，因此通过 Box-Cox 变换对因变量进行正态性变换，有效的避免了数据的有偏性，可有效降低数据预测误差，提高预测精度。如下图 3-1 所示，分别为 Box-Cox 变换前后因变量的 QQ 正态图。

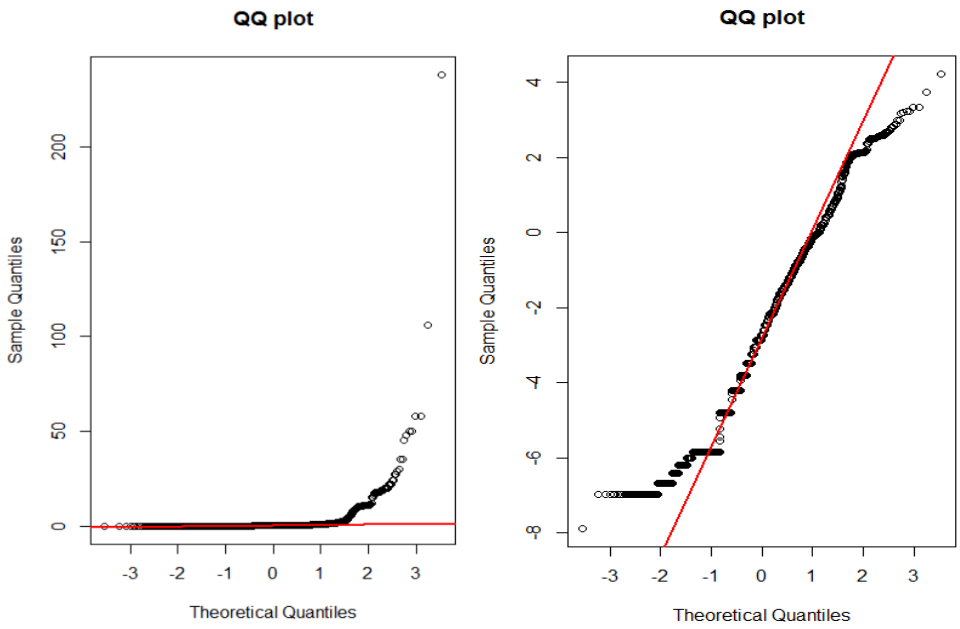


图 3-1 QQ 正态图

Figure 3-1 QQ plot

3、归一化处理结果

本文采用最值法进行归一化处理，将数据进行中心化及缩放，最终范围为[0,1]。数据归一化处理后，所有特征都为同一量纲，为后续建模提供良好的数据准备。

4、指标问题处理结果

由于本文手环销量的影响指标除数值型指标外，还有很多分类型指标，因此我们需要将其量化，这样可以很好的进行数据建模。数据预处理后，共得到 2546 条手环商品数据，28 个特征变量。其中数值型变量包括：价格(万元)、综合评分、商品评分、物流评分、售后评分、好评度、重量(kg)。分类型变量及其量化结果如表 3-2 所示。

表 3-2 分类型变量量化结果

Table 3-2 Quantitative criteria for classified variables

变量名称	量化标志
店铺类型	品牌官方旗舰店-1、京东自营旗舰店-2、其他-3
广告	是-1、否-0
品牌	小米-1、荣耀-2、华为-3、乐心-4、博之轮-5、VOSSTR-6、穆奇-7、欧易-8、Dido-9、奥迪斯-10、其他-11
适用人群	男士-1、女士-2、老人-3、儿童-4、通用-5
优先服务	30 天无忧退-1、一年质保-2、两年质保-3、三年质保-4
评价详情	功能丰富-1、简单方便-2、高端大气-3、精准度高-4、其他-5
NFC 支付	是-1、否-0
GPS 定位	是-1、否-0
泳姿识别	是-1、否-0
运动模式识别	一种模式-1、两种模式-2、三种模式-3、四种模式-5、五种模式-5、六种模式-6、无-7
功能用途	一种功能-1、两种功能-2、三种功能-3、四种功能-4、五种功能-5
防水等级	不防水-1、生活防水-2、30 米防水-3、50 米防水-4、50 米以上防水-5
续航时间	7 天以下-1、7~10 天-2、10~12 天-3、12~14 天-4、14~16 天-5、16~18 天-6、18~20 天-7、20~24 天-8、24 天以上-9
屏幕尺寸	0.8 英寸以下-1、0.8~1.0 英寸-2、1.0~1.2 英寸-3、1.2 英寸以上-4
蓝牙通话	支持-1、不支持-0
连接方式	蓝牙-1、WIFI-0
触控方式	单点触控-1、多点触控-2、全屏触控-3、非触控屏-4
腕带材质	硅胶-1、TPU-2、金属-3、皮革-4、其他-5
娱乐功能	音乐播放-1、防丢失-2、拍照-3、音乐播放+拍照-4、防丢失+拍照-5、音乐播放+防丢失+拍照-6、无-7
屏幕显示	彩屏-1、黑白屏-2、无屏幕-3
自动调节亮度	支持-1、不支持-0

3.4 指标体系特征选择

经过数据清洗与预处理后，下面将通过三种特征选择的方法进行综合对比，选出影响智能手环销量的显著因素。

3.4.1 方差特征选择法

方差特征选择法的本质是寻找两组数据之间的线性关系，其原假设是“数据不存在显著的线性关系”。它返回 F 值和 p 值两个统计量。我们希望选取 p 值小于 0.05 或 0.01 的特征，这些特征与标签是显著线性相关的，而 p 值大于 0.05 或 0.01 的特征则被我们认为是和标签没有显著线性关系的特征，应该被删除。

通过调用 R 语言的 aov()函数，对数据进行方差分析，结果如下表 3-3 所示。其中第一列为方差源，即待筛选的特征；第二列 Sum Sq 是偏差平方和，第三列 Mean Sq 是均方和，第四列 F value 是 F 值，即偏差平方和与残差平方和之比；第五列是 P 值，即假设检验是否通过的标准，若特征的 P 值<0.05，则特征与因变量是相关的；最后一列是每个特征的显著性表示，星号个数越多显著性越强。

表 3-3 方差分析表

Table 3-3 Analysis of variance table

变量名称	Sum Sq	Mean Sq	F value	Pr(>F)	
价格	2.7312	2.7312	99.8221	<0.0001	***
广告	3.7869	3.7869	138.4068	<0.0001	***
综合评分	3.9025	3.9025	142.6292	<0.0001	***
商品评分	0.2591	0.2591	9.4701	0.0021	**
物流评分	4.0017	4.0017	146.2572	<0.0001	***
售后评分	0.0066	0.0066	0.2423	0.6226	
店铺类型	0.5739	0.5739	20.9753	<0.0001	***
品牌	3.7619	3.7619	137.4916	<0.0001	***
适用人群	0.6096	0.6096	22.2788	<0.0001	***
优选服务	5.7216	5.7216	209.1175	<0.0001	***
好评度	1.9993	1.9993	73.0705	<0.0001	***
重量	0.1043	0.1043	3.8103	0.0511	.
评价详情	1.3504	1.3504	49.3535	<0.0001	***
NFC 支付	0.2760	0.2760	10.0890	0.0015	**
GPS 定位	0.6271	0.6271	22.9195	<0.0001	***
泳姿识别	0.1845	0.1845	6.7414	0.0095	**
运动模式识别	1.1248	1.1248	41.1108	<0.0001	***
功能用途	0.9069	0.9069	33.1453	<0.0001	***
防水等级	0.4359	0.4359	15.9311	<0.0001	***
续航时间	0.0637	0.0637	2.3269	0.1273	

屏幕尺寸	0.0506	0.0506	1.8511	0.1738	
蓝牙通话	0.0036	0.0036	0.1309	0.7175	
连接方式	0.1117	0.1117	4.0832	0.0434	*
触控方式	0.9168	0.9168	33.5066	<0.0001	***
腕带材料	0.0109	0.0109	0.3968	0.5288	
娱乐功能	0.7383	0.7383	26.9832	<0.0001	***
屏幕显示	0.0914	0.0914	3.3398	0.0677	.
自动调节亮度	1.0314	1.0314	37.6956	<0.0001	***

由上表可以看出，使用方差特征选择法最终选出的变量有 21 个，排除掉 7 个。

3.4.2 Lasso 特征选择法

Lasso 回归复杂度调整的程度由参数λ来控制，λ越大模型复杂度的惩罚力度越大，即惩罚项具有将其中某些系数的估计值强制设定为 0 的作用。因此，Lasso 可以根据不同的λ取值，得到包含不同变量个数的模型。

使用 R 语言中的 glmnet()函数，进行基于 Lasso 的变量选择。其中，设置 alpha=1(默认值)，family="Gaussian"，λ的范围为$[10^{-2},10^{10}]$。从系数图像 3-2 可知，随着λ的增大，有些特征的系数被压缩为 0。

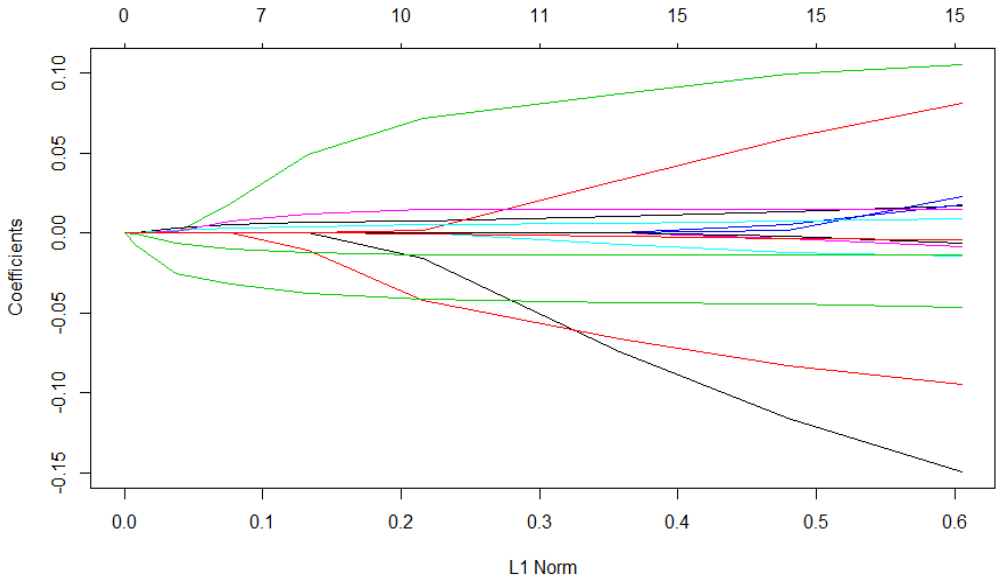


图 3-2 Lasso 系数压缩图

Figure 3-2 Lasso coefficient compression diagram

接下来，为了确定一个最佳λ值，我们通过 cv.glmnet()函数对模型进行十折交叉验证。结果如下图 3-3 所示，当log(λ)在-6 左右时，模型均方误差达到最小。

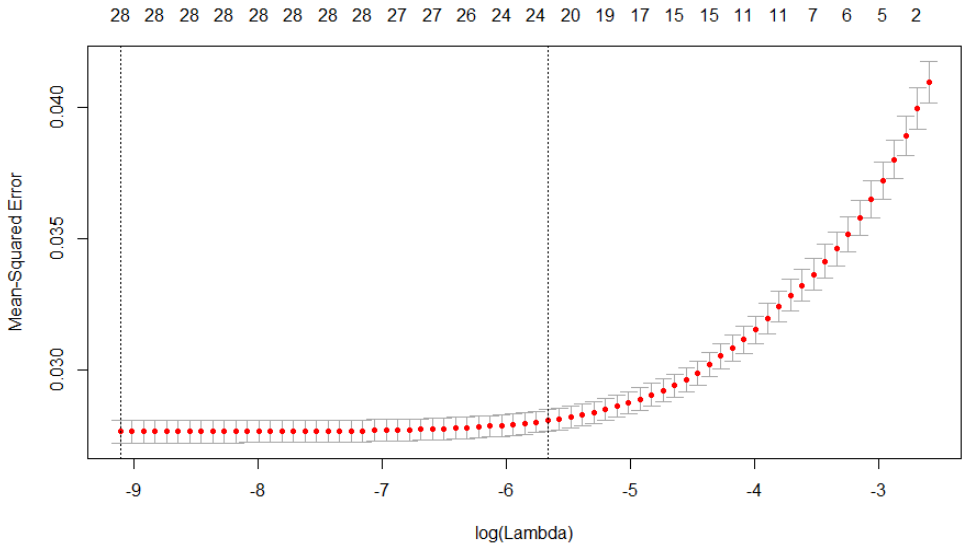


图 3-3 MSE 与log λ拟合图

Figure 3-3 The MSE and log λ lambda figure fitting

最后，将最佳λ=0.0001 代入模型，不重要的特征系数已经被压缩为 0，输出结果如下表 3-4 所示。

表 3-4 Lasso 模型结果汇总表

Table 3-4 Summary of lasso model results

变量名称	系数	变量名称	系数
价格	-0.1491	GPS 定位	0.0175
广告	-0.0944	泳姿识别	0.0000
综合评分	0.0000	运动模式识别	0.0094
商品评分	0.0000	功能用途	0.0147
物流评分	0.1053	防水等级	0.0000
售后评分	0.0230	续航时间	0.0000
店铺类型	-0.0144	屏幕尺寸	0.0000
品牌	0.0000	蓝牙通话	0.0000
适用人群	-0.0083	连接方式	0.0000
优选服务	0.0172	触控方式	-0.0060
好评度	0.0816	腕带材料	0.0000
重量	0.0000	娱乐功能	-0.0043
评价详情	-0.01342	屏幕显示	0.0000
NFC 支付	0.0000	自动调节亮度	-0.0463

由上表 3-4 可知，使用 Lasso 特征选择法，将系数为 0 的变量进行剔除，最终选出 15 个变量，排除掉 13 个变量。

3.4.3 Boruta 特征选择法

使用 R 语言中的 Boruta()函数进行特征选择，28 个变量中有 16 个变量被确认为重要(接受)，包括 NFC 支付、店铺类型、防水等级、好评度、价格等；6 个变量被确认为不重要(拒绝)，包括广告、连接方式、评价详情、屏幕尺寸、腕带材料等；6 个变量被待定，包括 GPS 定位、触控方式、功能用途、蓝牙通话、适用人群等，结果如下图 3-4 所示：

```
> print(boruta.train)
Boruta performed 99 iterations in 5.506173 mins.
 16 attributes confirmed important: NFC支付, 店铺类型, 防水等级, 好评度, 价格 and 11
more;
 6 attributes confirmed unimportant: 广告, 连接方式, 评价详情, 屏幕尺寸, 腕带材料 and 1
more;
 6 tentative attributes left: GPS定位, 触控方式, 功能用途, 蓝牙通话, 适用人群 and 1 more;
```

图 3-4 特征选择输出结果

Figure 3-4 Output results of feature selection

Boruta 特征选择法所确定的变量选择结果，可以用可视化盒状图来展示，如下图 3-5 所示。红色、黄色、绿色盒装图分别代表拒绝、待定以及接受的 Z 分数，蓝色盒装图从左到右分别为最小、平均以及最大 Z 分数。其中 Z 分数用平均损失除以标准差的计算方法作为重要度，因为它考虑了森林中树木之间平均准确度损失的波动，取值越大表明特征的重要性越高。

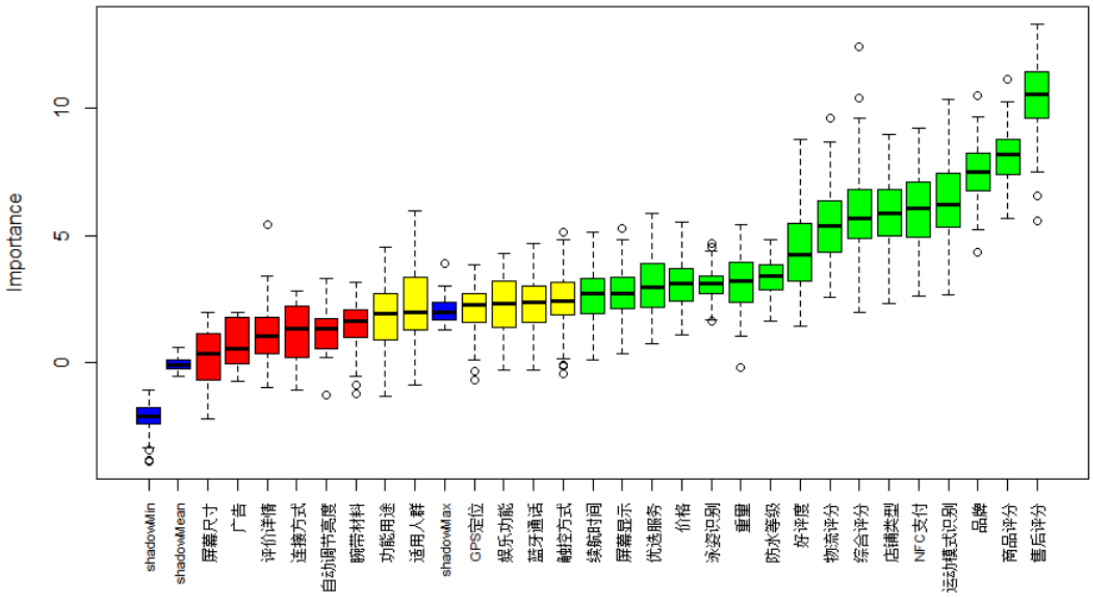


图 3-5 特征选择盒状图

Figure 3-5 Feature selection box

根据 Boruta 特征选择盒状图，我们将特征选择结果进行整理，如下表 3-5 所示：

表 3-5 特征选择结果

Table 3-5 Results of feature selection

选择结果	特征名称
接受	售后评分、商品评分、品牌、运动模式识别、NFC 支付、店铺类型、综合评分、物流评分、好评度、防水等级、重量、泳姿识别、价格、优选服务、屏幕显示、续航时间
待定	触控方式、蓝牙通话、娱乐功能、GPS 定位、适用人群、功能用途
拒绝	腕带材料、自动调节亮度、连接方式、评价详情、广告、屏幕尺寸

最后，通过将待定 Z 分数与最佳阴影 Z 分数的中位数进行比较，最终将待定 Z 分数归类到接受属性或者拒绝属性，如下表 3-6 所示。

表 3-6 特征选择最终结果

Table 3-6 Feature selection final result

变量名称	meanImp	medianImp	minImp	maxImp	normHits	decision
价格	3.1102	3.0917	1.0994	5.5455	0.8182	Confirmed
广告	0.7041	0.5353	-0.7330	1.9889	0.0000	Rejected
综合评分	5.8779	5.6769	1.9588	12.444	0.9899	Confirmed
商品评分	8.1415	8.1937	5.6821	11.159	1.0000	Confirmed
物流评分	5.4580	5.4030	2.5409	9.6087	1.0000	Confirmed
售后评分	10.4261	10.5459	5.5560	13.320	1.0000	Confirmed
店铺类型	5.8635	5.8521	2.3363	8.9829	0.9899	Confirmed
品牌	7.4618	7.5162	4.3324	10.532	1.0000	Confirmed
适用人群	2.2621	1.9702	-0.9136	5.9538	0.5758	Confirmed
优选服务	2.9886	2.9489	0.7444	5.8710	0.7677	Confirmed
好评度	4.4309	4.2490	1.4439	8.8098	0.9394	Confirmed
重量	3.1961	3.2307	-0.1869	5.4356	0.8384	Confirmed
评价详情	1.2597	1.0498	-0.9826	5.4330	0.0707	Rejected
NFC 支付	6.0303	6.0744	2.6333	9.2089	1.0000	Confirmed
GPS 定位	2.1715	2.2602	-0.7079	3.8690	0.5758	Confirmed
泳姿识别	3.0763	3.1321	1.6100	4.6754	0.8889	Confirmed
运动模式识别	6.3035	6.1983	2.6750	10.361	1.0000	Confirmed
功能用途	1.8303	1.9049	-1.3249	4.5293	0.5152	Confirmed
防水等级	3.3838	3.3847	1.6439	4.8331	0.9394	Confirmed
续航时间	2.6545	2.6952	0.1153	5.1108	0.6869	Confirmed
屏幕尺寸	0.1397	0.3474	-2.2320	1.9851	0.0101	Rejected
蓝牙通话	2.2933	2.3707	-0.3216	4.6930	0.5960	Rejected
连接方式	1.1599	1.3269	-1.0739	2.8323	0.1010	Rejected
触控方式	2.3587	2.4122	-0.4470	5.1256	0.6465	Confirmed
腕带材料	1.4324	1.6358	-1.2469	3.1479	0.2525	Rejected
娱乐功能	2.2754	2.2941	-0.2786	4.2967	0.5859	Confirmed
屏幕显示	2.7479	2.7039	0.3519	5.2956	0.7879	Confirmed
自动调节亮度	1.1930	1.3278	-1.3049	3.3272	0.0101	Rejected

由上表 3-6 所示，使用 Boruta 特征选择法最终从 28 个特征中选出 21 个变量，排除掉 7 个变量。

3.4.4 变量综合对比选择

我们通过方差过滤法、Lasso 特征选择法和 Boruta 特征选择法三种方法，对特征进行综合对比，当变量至少有两种方法都选择时，我们将其纳入指标体系，最终筛选出 21 个变量，结果如下表 3-7 所示。

表 3-7 三种方法的变量选择对比

Table 3-7 Variable selection comparison of three methods

变量名称	方差	Lasso	Boruta	最终选择
价格	√	√	√	√
广告	√	√		√
综合评分	√		√	√
商品评分	√		√	√
物流评分	√	√	√	√
售后评分		√	√	√
店铺类型	√	√	√	√
品牌	√		√	√
适用人群	√	√	√	√
优选服务	√	√	√	√
好评度	√	√	√	√
重量			√	
评价详情	√	√		√
NFC 支付	√		√	√
GPS 定位	√	√	√	√
泳姿识别	√		√	√
运动模式识别	√	√	√	√
功能用途	√	√	√	√
防水等级	√		√	√
续航时间			√	
屏幕尺寸				
蓝牙通话				
连接方式	√			
触控方式	√	√	√	√
腕带材料				
娱乐功能	√	√	√	√
屏幕显示			√	
自动调节亮度	√	√		√

3.5 本章小结

本章通过对网络爬虫的 3128 条商品信息进行数据加工及预处理，得到一份适用于模型建立的数据。首先对数据进行缺失值处理，采用替换法及删除法，最终得到 2546 条商品数据；其次对连续型数据进行 Box-Cox 正态性变换，使得数据具有良好的正态性，有利于降低误差，提高预测准确率；再次对数据进行归一化处理，使用最小-最大规范化方法消除数据的量纲；最后，采用方差过滤法、Lasso 特征选择法和 Boruta 特征选择法对变量进行筛选，最终从 28 个变量中确定 21 个重要特征。

第 4 章 智能手环销量建模与预测

4.1 分析思路

本章对数据预处理后的 21 个变量、2546 条商品数据进行模型建立与预测，主要分析思路如下：

1、智能手环描述性统计分析。从品牌、功能及用户画像三个方面对智能手环进行描述性统计分析，了解消费者的情况及关注点。

2、模型建立与预测。将预处理后的数据集按 7:3 的比例划分为训练集和测试集，在训练集上分别采用支持向量机、随机森林和 XGBoost 三种模型进行预测，通过拟合三种模型在测试集上预测值与真实值的对比图，来初步判定模型的预测效果。

3、模型对比与评估。模型评估是体现一个模型好坏最直观的方法。首先，本文将 NMSE 作为评价的一个指标，分别求出三种模型在训练集和测试集上的 NMSE。然后，对三种模型在测试集上的预测误差通过 RMSE 等多个指标进行对比评估。最终，得出手环最佳预测模型。

4.2 智能手环描述性统计分析

4.2.1 智能手环品牌及功能分析

对于智能手环品牌的分析，主要使用本文爬取的 3128 条商品数据，作出品牌的词云图。品牌的字体型号越大，说明其销量越高。由下图 4-1 可以看出小米、华为、荣耀、乐心等品牌的销量较高。据调查显示，我国小米及华为手环一如既往的领跑市场，占比 47.6%，接近半数之多。

对于手环的功能分析，取品牌销量 TOP30 的商品数据，作出手环功能的雷达图。如下图 4-1 所示，近年来，健康监测占比 27%，是用户的主要关注点。其次是身份识别，消息提醒等方面的功能。因此，智能手环的发展不仅局限于基本的运动计步，而是逐渐向智能健康领域发展，高龄人群逐步成为智能手环发展的潜在市场。

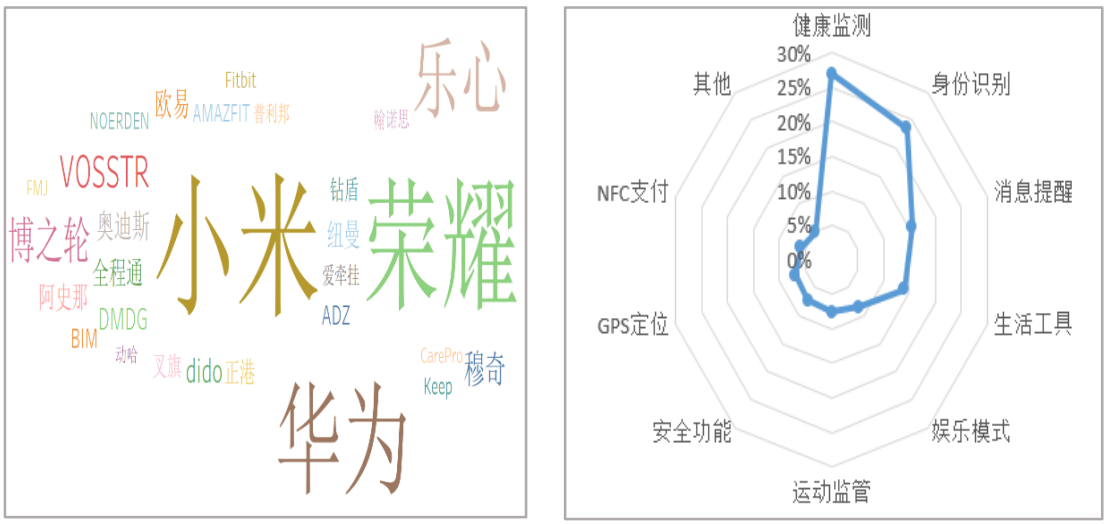


图 4-1 品牌词云图及功能雷达图

Figure 4-1 Brand word cloud chart and function radar chart

4.2.2 智能手环用户画像分析

对智能手环的性别及年龄进行分析，如下图4-2所示。在智能手环的用户中，20～39 岁用户占比 83.7%，成为手环用户的主要群体，智能手环呈现出年轻化趋势。另外从性别上看，男士占比 77%，而女士仅有 23%，说明女士手环市场仍待挖掘。因此，20～39 岁的男性用户几乎占据了整个智能手环的市场，与此同时，高龄人群将会是下一个值得挖掘的市场，除基本的运动计步功能之外，健康领域的研发也是一项重要的探索方向。

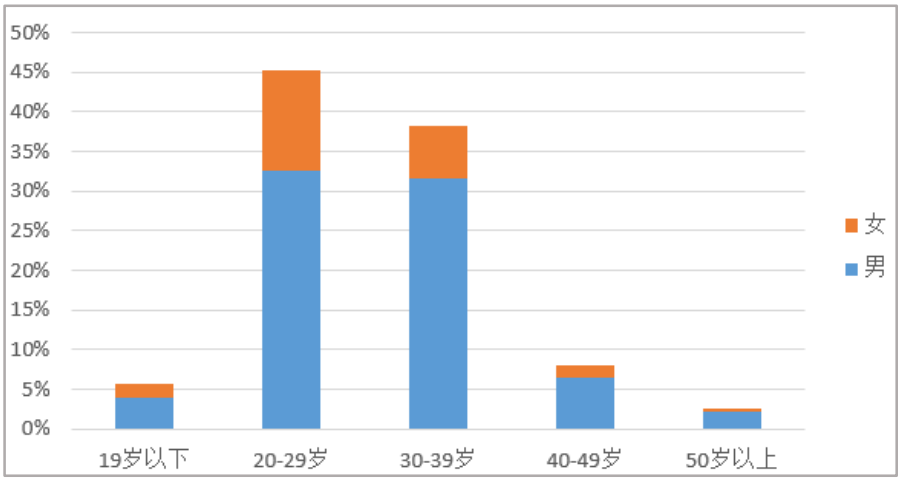


图 4-2 用户年龄及性别分布图

Figure 4-2 User age and gender profile

综合以上对数据的描述性分析，可以了解到用户对智能手环品牌和功能的偏好，以及手环用户市场可挖掘的方向，对智能手环行业发展分析及建模预测有一定的意义。

4.3 基于支持向量机的预测模型

4.3.1 参数选择与模型建立

首先，将方差特征选择、Lasso 特征选择、Boruta 特征选择最终选出的 21 个特征作为模型的自变量，手环的销量作为模型因变量，建立 SVR 模型。本文 SVR 模型的建立使用的是 svm( )函数，格式如下所示：

svm(formula,data=NULL,type=NULL,kernel="radial",gamma=1,cost=1,…)

其中，formula 为模型形式；data 为训练数据；type 为模型类别，包括 C-classification、nu-classification、one-classification、eps-regression 和 nu-regression 五种，前三种用于分类模型，后两种用于回归模型[44]；kernel 为核函数，包括 linear、polynomial、radial 和 sigmoid 四种； gamma 和 cost 分别表示惩罚因子和核函数的参数。下面进行模型参数的具体确定。

(1) 确定 type 和 kernel 参数

由于智能手环销量为连续型数值型变量，因此 type 可取 eps-regression 和 nu-regression 两种；kernel 参数取默认的四种即可，然后将两个参数进行组合，共建立 8 次预测模型，计算每次模型下的均方误差，公式定义如下：

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{N} \quad , \tag{4-1}$$

其中，Y_i为真实值，$\hat{Y}_i$为预测值。

表 4-1 type 与 kernel 组合下的均方误差

Table 4-1 Mean square error under type and kernel combination

	linear	polynomial	radial	sigmoid
eps-regression	0.0267	0.0188	0.0162	2.2493
nu-regression	0.0259	0.0186	0.0160	1.5760

由上表 4-1 可知，当 SVR 模型的 type 为 nu-regression，kernel 为 radial 时，均方误差最小，为 0.0160。参考相关的文献也表明，当 kernel 为 radial(径向基函数)时，模型性能较好。

(2) 确定 gamma 和 cost 参数

gamma 参数的默认值为特征个数的倒数，本文用于建模的特征共有 21 个，故 gamma 默认为 0.0476。但为了模型能拟合出最优效果，我们将 gamma 的值设定为 0.001、0.01、0.0476(默认值)、0.1 四个值；对于惩罚因子 cost，为防止模型过拟合，取 0.1、1(默认值)、10 三个值。然后将两个参数进行组合，共建立 12 次

预测模型，并计算不同组合下的模型预测误差。

表 4-2 cost 与 gamma 组合下的均方误差

Table 4-2 Mean square error in combination of cost and gamma

	0.001	0.01	0.0476	0.1
0.1	0.0308	0.0253	0.0217	0.0218
1	0.0263	0.0210	0.0160	0.0143
10	0.0241	0.0180	0.0139	0.0142

由上表 4-2 可知，当 cost=10，gamma=0.0476(默认)时，均方误差最小，为 0.0139。因此本文 SVR 模型选择惩罚因子 cost=10，gamma=0.0476(默认)。

最后，将确定好的参数带入 SVR 模型，使用 summary()函数得到模型的回归结果，如下图 4-3 所示，在 2546 条商品数据中，找到 1420 个支持向量。

```
Call:
svm(formula = 销量 ~ 价格 + 广告 + 综合评分 + 商品评分 + 物流评分 + 售后评分 + 店铺类型 +
    品牌 + 适用人群 + 优选服务 + 好评度 + 评价详情 + NFC支付 + GPS定位 + 泳姿识别 + 运动模式识别 +
    功能用途 + 防水等级 + 触控方式 + 娱乐功能 + 自动调节亮度, data = train_data, type = "nu-regression",
    kernel = "radial", cost = 10, gamma = 0.0476)


Parameters:
   SVM-Type:  nu-regression
 SVM-Kernel:  radial
       cost:  10
         nu:  0.5

Number of Support Vectors:  1420
```

图 4-3 支持向量机输出结果

Figure 4-3 Output results of support vector machine

4.3.2 模型拟合效果

为了更加直观的展示 SVR 模型的拟合效果，建立预测值和真实值的回归模型，并绘制模型的拟合效果图。

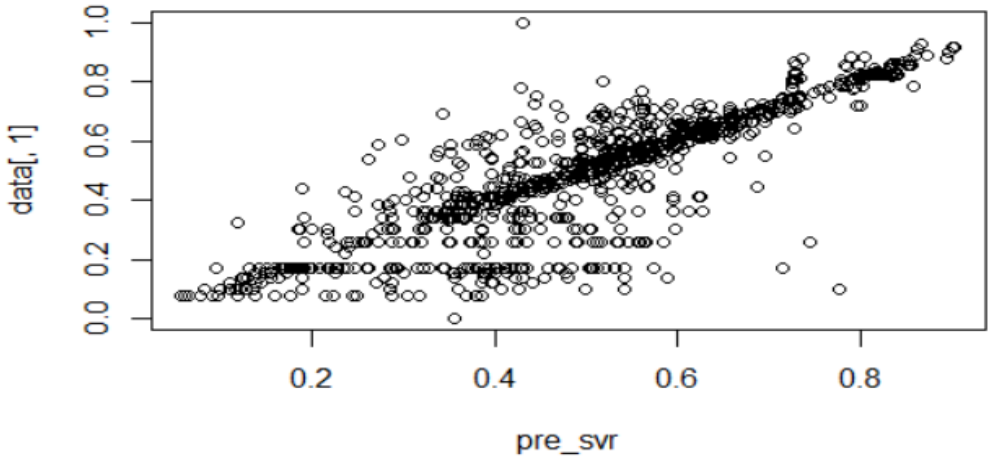


图 4-4 SVR 拟合效果图

Figure 4-4 SVR fitting effect drawing

如上图 4-4 所示，预测值与真实值组成线性趋势，但是有相当多一部分数据落在直线外面，说明还是有一定的误差存在。总体来说，SVR 模型还是有一定的拟合效果，接下来尝试用随机森林进行建模。

4.4 基于随机森林的预测模型

4.4.1 确定参数 ntree

ntree 表示决策树的数量，ntree 参数的确定需要先指定一个 mtry 值。若设模型中特征个数为 p，在回归模型中，mtry 参数的一般设定值为 p/3，在判别模型中则设定为$\sqrt{p}$。因此，本文随机模型 mtry 参数设定为 21/3=7。然后，调用 R 语言的 randomForest( )函数，设定 mtry=7 不变，作出预测误差随决策树数量变化的趋势图。

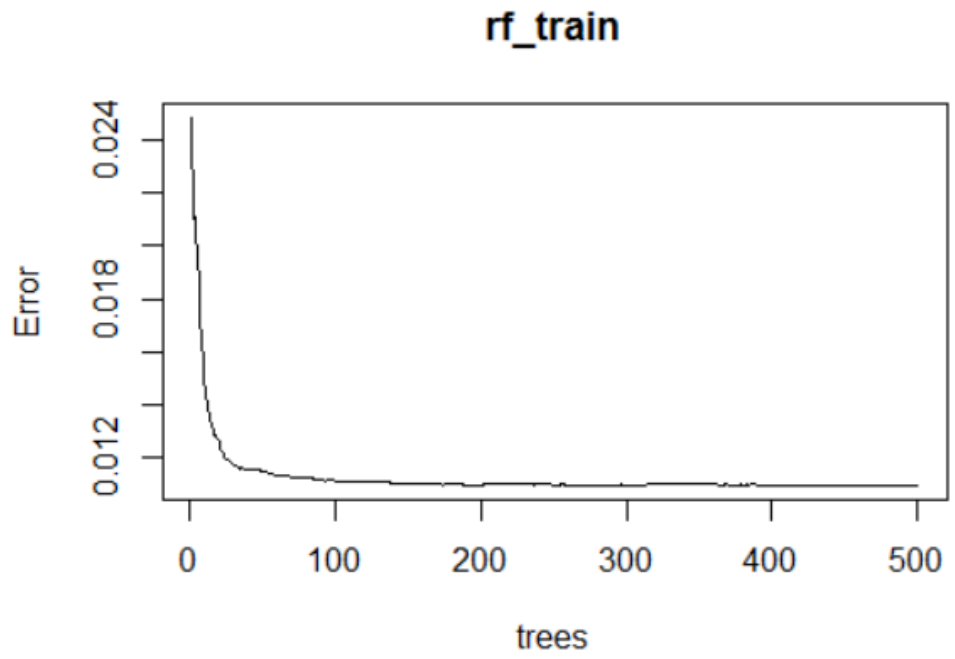


图 4-5 不同 ntree 下的预测误差趋势图

Figure 4-5 Prediction error trend graph under different ntrees

由上图 4-5 可知，预测误差随着决策树的增加而减少，并逐渐趋于平稳。当 ntree=500 时，模型误差处于非常稳定的状态，故本文中取 ntree=500，即随机森林中包含 500 棵决策树。

4.4.2 确定参数 mtry

参数 mtry 表示随机森林的各决策树在分枝时所选择的特征个数。由上文可知，我们将 ntree 参数设定为 500，然后拟合出不同 mtry 下目标变量的方差解释度和残差平方和。设定 mtry 的值从 3 开始，逐次取到，得到解释度和均方误差的对比图，从而选出最佳 mtry，如下图 4-6 所示。

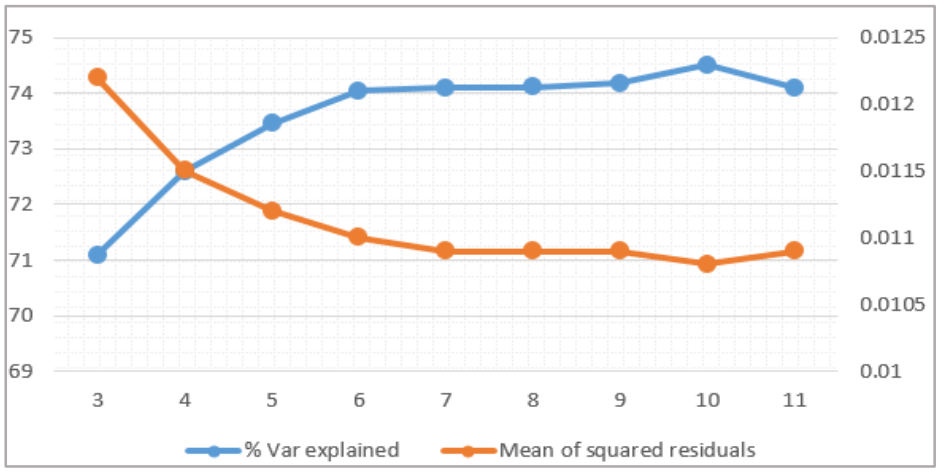


图 4-6 不同 mtry 下解释度与均方误差对比图

Figure 4-6 Comparison of interpretation degree and mean square error under different mtry

由上图 4-6 可知，当 mtree 小于 10 时，解释度不断上升，均方误差不断下降。当 mtry=10 时，目标变量解释度达到最高 74.45%，均方差也达到最低 0.0108。因此，当 ntree 为 500，mtry 为 10 时，随机森林预测模型达到最优。

4.4.3 变量重要性排序

使用随机森林建模预测的同时，还可以给出特征的重要性排序，直观的看到影响智能手环销量的重要因素。特征重要性排序结果如下表 4-3 所示：

表 4-3 特征重要性结果

Table 4-3 Feature importance result

变量名称	%IncMSE	IncNodePurity
价格	0.8146	7.0566
广告	0.0349	0.5445
综合评分	0.5889	2.3342
商品评分	1.1884	6.2835
物流评分	2.0533	9.2139
售后评分	1.4163	7.2045
店铺类型	1.1540	3.3543
品牌	1.4306	5.4906

适用人群	0.0880	0.78412
优选服务	1.0706	5.5503
好评度	0.4484	2.9929
评价详情	0.7560	6.3124
NFC 支付	0.0933	0.5409
GPS 定位	0.0664	0.4277
泳姿识别	0.1617	0.7156
运动模式识别	0.6114	4.0329
功能用途	0.0020	1.5900
防水等级	0.2233	1.1144
触控方式	0.2849	1.2726
娱乐功能	0.2396	1.2906
自动调节亮度	0.7262	3.6557

为了更直观的显示特征重要性结果，可将其可视化，结果如下图 4-7:

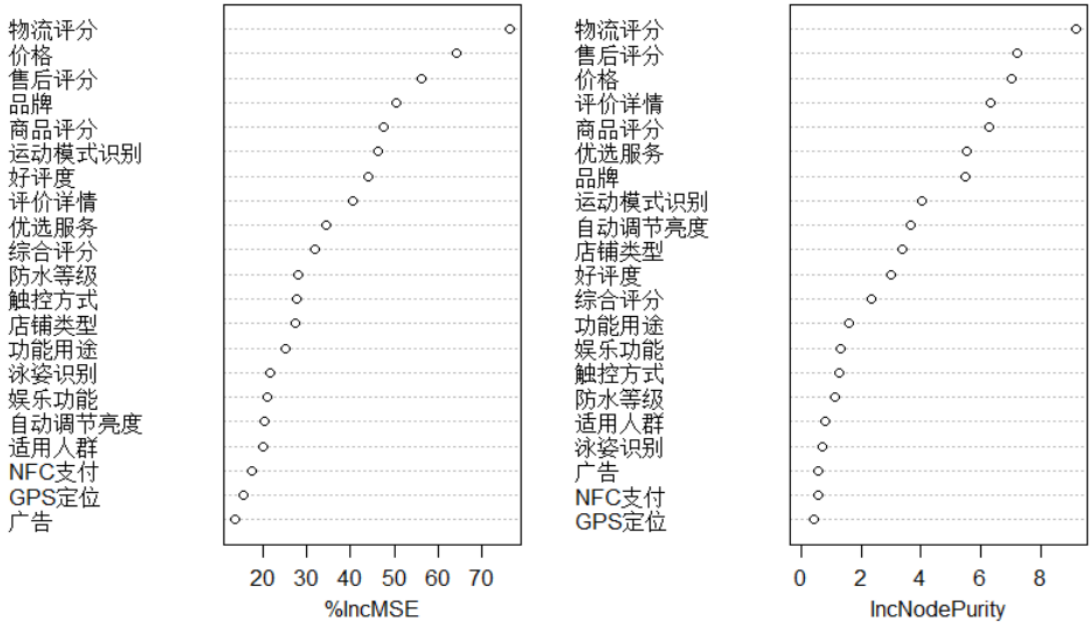


图 4-7 特征重要性趋势图

Figure 4-7 Feature importance trend chart

%IncMSE 表示模型误差的增加程度，即模型预测精度的减少程度。IncNodePurity 与%IncMSE 原理相同，两者的值越大，特征重要性越高。

由特征重要性趋势图 4-7 可知，手环销量的重要影响特征有：物流评分、价格、售后评分、品牌、商品评分、运动模式识别、好评度及评价详情。可见除了平时需要关注的价格、品牌及功能外，用户对于物流评分及评价有很大的关注度，这也与当前快节奏、高质量的生活息息相关。因此，商家可以提高物流速度，同时指导已购用户及时反馈产品体验，提升店铺好评度等。相反，重要性较低的特征有广告、GPS 定位、适用人群、自动调节亮度等指标，这些特征对手环的影响

程度较低。

4.4.4 模型拟合效果

将 mtry=10，ntree=500 代入模型中，可以得到随机森林模型的拟合效果，如下图 4-8 所示：

```
Call:
 randomForest(formula = 销量 ~ 价格 + 广告 + 综合评分 + 商品评分 +       物流评分 + 售后
评分 + 店铺类型 + 品牌 + 适用人群 + 优选服务 +       好评度 + 评价详情 + NFC支付 + GPS定位
+ 泳姿识别 + 运动模式识别 +       功能用途 + 防水等级 + 触控方式 + 娱乐功能 + 自动调节亮度,
     data = train_data, ntree = 500, mtry = 10)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 0.01077381
                    % Var explained: 74.51
```

图 4-8 随机森林预测效果图

Figure 4-8 Random forest prediction renderings

由上面的随机森林预测效果图可知，当 ntree=500，mtry=10 时，建立随机森林回归模型，得到的残差平方均值(Mean of squared resisuanls)为 0.0108，模型因变量解释度(% var explained)为 74.51%。随机森林模型的残差平方均值较小，同时也有较好的模型解释度，因此使用随机森林对手环销量进行建模具有良好的效果。为了更直观的看到模型在测试集上的预测效果，下面作出随机森林真实值与预测值的拟合图。

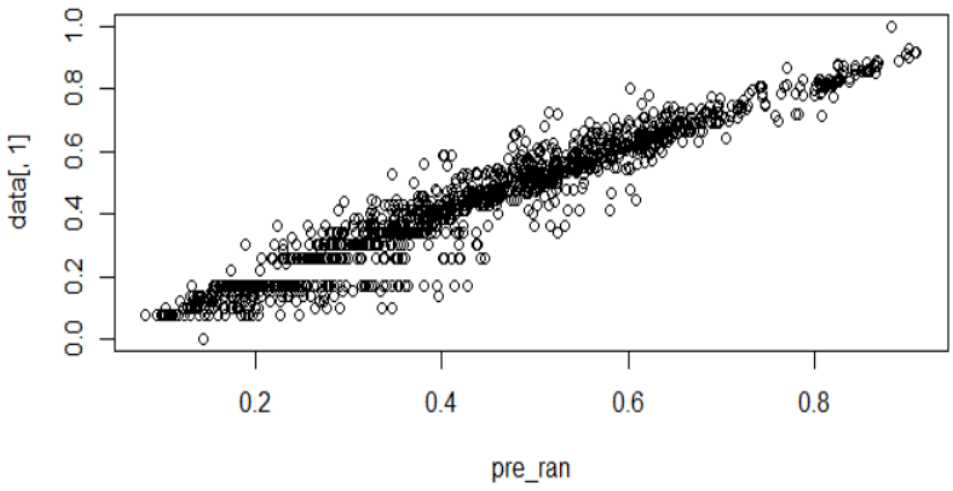


图 4-9 随机森林拟合效果图

Figure 4-9 Random forest fitting effect diagram

上图 4-9 横坐标为随机森林预测值，纵坐标为手环销量真实值。可以看到较支持向量机来说，模型拟合的线性趋势性更强，说明随机森林效果良好。

4.5 基于 XGBoost 的预测模型

4.5.1 参数说明

XGBoost 模型的参数通常有 3 种类型：通用参数、Booster 参数与学习任务参数。通用参数用于确定模型的宏观功能，包括模型的类型选择等；Booster 参数是基于通用参数的模型类型，进行精细的参数控制；学习任务参数是用来对学习任务及目标进行定义的参数[38]。XGBoost 模型涉及的参数较多，下面对比较常用的参数进行简单介绍：

(1) booster：设置模型每次迭代的模型类型，包括 gbtree(树模型)和 gblinear(线性模型)，默认为 gbtree。

(2) nrounds：最大迭代次数，即最大生成树的个数，通常与学习率配合使用。

(3) eta：模型步长的收缩程度，即学习率。学习率通常需要乘以权重，因为决策树叶子节点在更新的过程中容易出现过拟合的现象，往往较小的学习率可以有效的避免过拟合现象。通常取值范围为[0.01,0.2]。

(4) max_depth：模型中树的最大深度，该值越大，模型的复杂度就越大，可以有效避免模型过拟合的现象。通常取值范围为[3,10]。

(5) min_child_weight：模型叶子节点最小样本数。该值通常是判断该节点是否继续分裂的依据。该值越大，算法越保守，可以有效避免过拟合现象。通常取值范围为[0,10]。

(6) subsample：指每棵树在训练的时候，训练样本集占总体样本集的比例。该值越小越容易欠拟合。通常取值范围为[0.5,1]。

(7) colsample_bytree：指每棵树在训练的时候，随机抽取的特征数占总特征数的比例。该值越小越容易欠拟合。通常取值范围为[0.5,1]。

(8) gamma：为加入新叶子节点所带来的复杂度代价，该值越大，越不容易过拟合。通常取值范围为[0,0.2]。

(9) objective：该参数用来对损失函数的类型进行定义，包括 reg:linear(线性回归)，reg:logistic(逻辑回归)等，默认为 reg:linear。

4.5.2 参数调优与确定

XGBoost 模型的参数比较多，但是模型的主要影响参数为通用参数和 Booster 参数。下面，对于 XGBoost 的主要影响参数进行调优。首先，对 12 个

XGBoost 参数进行初始值的设定，如下表 4-4 所示：

表 4-4 参数初始化取值

Table 4-4 Parameter initializes the value

参数	值	参数	值
objective	reg:linear	eta	0.1
nrounds	100	max_depth	5
min_child_weight	1	gamma	0
subsample	0.8	colsample_bytree	0.8
reg_alpha	0	reg_lambda	1
booster	gbtree	seed	0

其次，设定好初始值后，按照不同参数对模型影响的重要性程度，进行分步骤参数调优，具体可按如下步骤进行：

(1) 确定最佳迭代次数（nrounds）

(2) 确定最大深度（max_depth）及最小叶子节点数（min_child_weight）

(3) 确定最小损失函数下降（gamma）

(4) 确定每棵树子样本比例（subsample）及子特征比例（colsample_bytree）

(5) 确定正则化系数 reg_alpha 和 reg_lambda

(6) 确定学习速率 eta

下面对于重要参数的调优过程进行说明，即(1)、(2)步，具体过程如下：

(1) 确定最佳迭代次数

nrounds 参数表示生成最大树的个数，即最大迭代次数,是影响模型误差的重要因素。设 nrounds 的初始值为 100，范围为 100-800，间隔为 50，绘制均方误差随 nrounds 变化的趋势图。

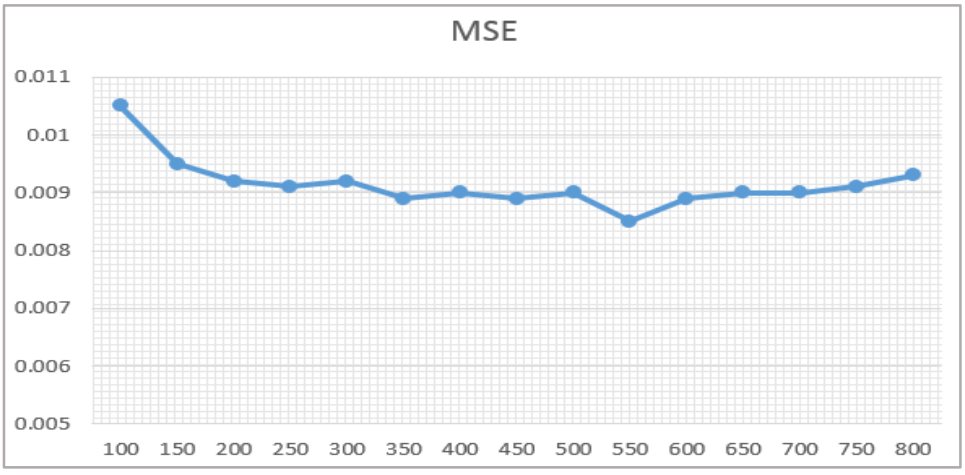


图 4-10 迭代次数的均方误差趋势图

Figure 4-10 Trend chart of mean square error of iteration times

由上图 4-10 可知，当 nrounds 参数为 550 时，模型误差达到最小，基本保持稳定，因此，取 nrounds 为 550。

(2) 确定最大深度及最小叶子节点数

对于 max_depth 参数，取值范围为 3-9；对于 min_child_weight 参数，取值范围为 1-3。两个参数进行组合，结果如下表 4-5：

表 4-5 max_depth 与 min_child_weight 组合下的误差

Table 4-5 Error under max_depth and min_child_weight combination

	3	4	5	6	7	8	9
1	0.0102	0.0091	0.0089	0.0090	0.0085	0.0083	0.0086
2	0.0107	0.0095	0.0089	0.0085	0.0088	0.0088	0.0087
3	0.0109	0.0094	0.0089	0.0085	0.0085	0.0088	0.0090

可知，当 max_depth=8，min_child_weight=1 时，均方误差最小为 0.0083。

最后，通过以上六步参数调优，可以得到 XGBoost 模型下的最佳参数组合，均方误差达到 0.0079，如下表 4-6 所示：

表 4-6 XGBoost 最佳参数组合

Table 4-6 XGBoost best combination of parameters

参数	值	参数	值
objective	reg:linear	eta	0.05
nrounds	550	max_depth	8
min_child_weight	1	gamma	0
subsample	0.8	colsample_bytree	0.7
reg_alpha	0.05	reg_lambda	1
booster	gbtree	seed	1314

4.5.3 特征重要性排序

调用 R 语言中的 xgb.plot.importance()函数，可以对 XGBoost 模型的特征进行重要性排序，直观的反应智能手环的重要影响因素，如下图 4-11 所示。

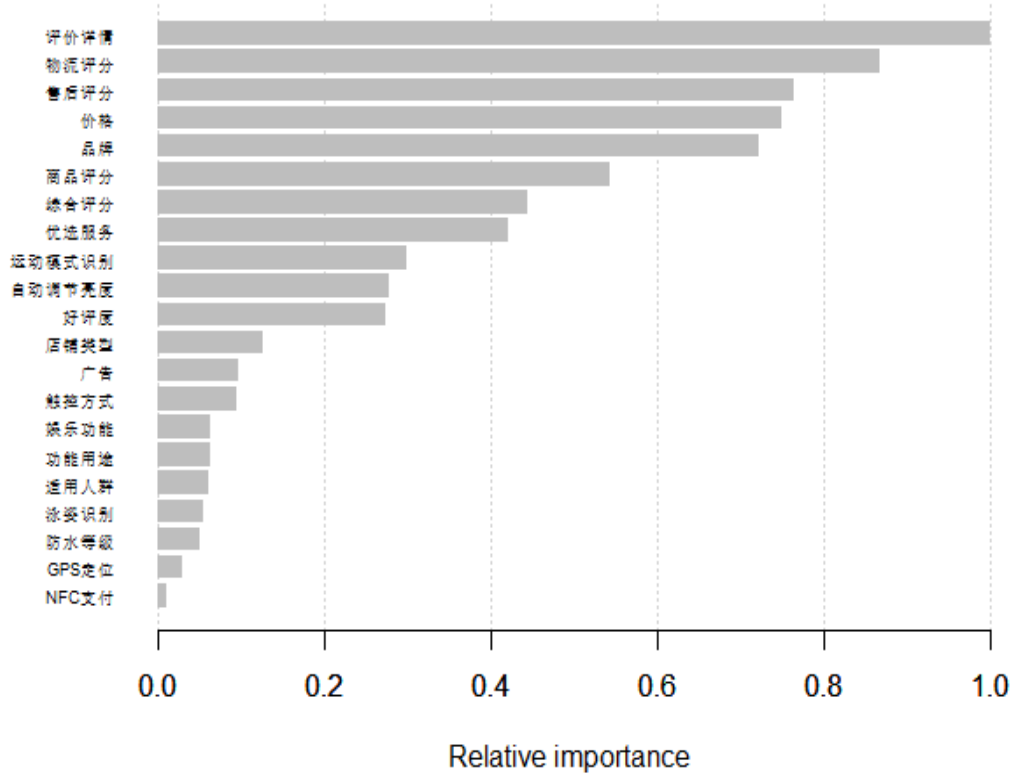


图 4-11 XGBoost 特征重要性结果

Figure 4-11 XGBoost feature importance results

由上图 4-11，可以清楚的看到 XGBoost 模型下各变量的重要性程度。特别需要关注的是，评价详情和物流评分的排名靠前，在随机森林模型中这两个特征排名也靠前，说明物流速度及用户评价是影响手环销量的重要因素，商家要特别考虑。其次，价格和品牌的排名也比较靠前，说明手环否品牌效应还是很高的，这也是显而易见的。NFC 支付、GPS 定位、防水等级的排名较靠后，说明用户对于手环的高级功能可能不太考虑，这些特征对手环销量的影响也就稍弱。

4.5.4 模型拟合效果

为了更加直观的展示 XGBoost 模型的拟合效果，建立预测值和真实值的回归模型，并绘制模型的拟合效果图，如下图 4-12 所示。

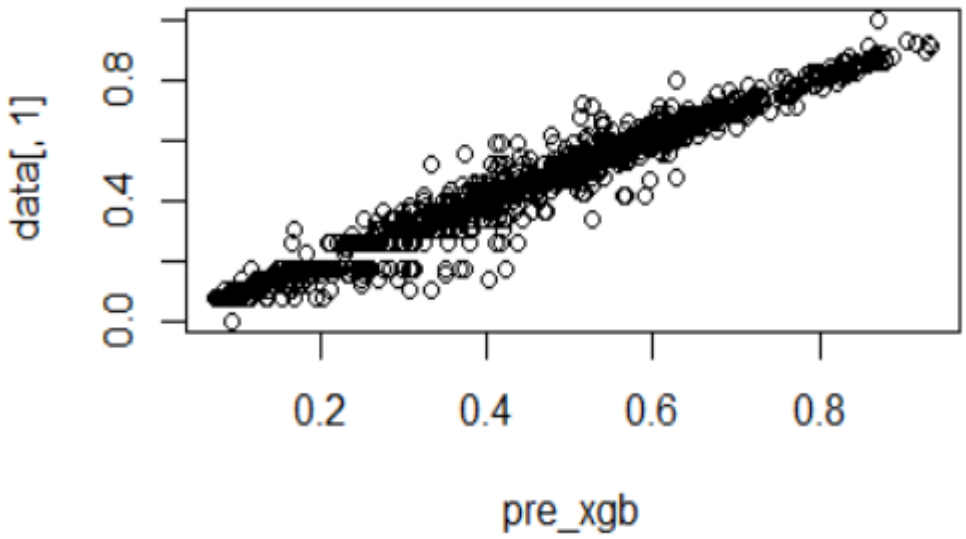


图 4-12 XGBoost 预测值与真实值拟合图

Figure 4-12 XGBoost predicted values fit into the true values

可以看到与支持向量回归模型和随机森林模型相比，XGBoost 模型拟合的线性趋势性更强，说明 XGBoost 的效果更好。为了更加准确的对三种模型的预测效果进行对比，下面进行准确的模型评估。

4.6 模型对比评估

4.6.1 十折交叉验证

交叉验证法[45](也称循环估计)是统计学中常用的模型评估方法。十折交叉验证的基本思想是：将数据集随机等分为 10 份，其中 1 份作为测试集，其余 9 份作为训练集，建立 10 个模型进行综合预测打分。本文采用 NMSE(标准化均方误差)作为十折交叉验证的评价指标，公式如下所示。

$$NMSE = \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \quad , \tag{4-2}$$

其中，$\bar{y}$为因变量的均值，$\hat{y}$为因变量的预测值。

对于评价指标 NMSE，分子表示真实值与预测值的距离平方和，分母表示真实值与其均值的距离平方和。因此，对于给定的手环销量数据，当真实值与预测值的差异越小，即 NMSE 越小时，模型的预测误差就越小。并且当 NMSE 显著小于 1 时，模型具有较好的预测精度。

本文采用十折交叉验证的方法对三种模型进行评估，分别得到三种模型在训练集和测试集上的标准均方误差，结果如下表 4-7 所示。

表 4-7 十折交叉验证下的 NMSE 结果

Table 4-7 NMSE results under ten fold cross validation

模型	训练集 NMSE	测试集 NMSE
支持向量机	0.2016	0.3670
随机森林	0.1586	0.2309
XGBoost	0.1384	0.2146

由上表 4-7 可以看出，SVR 模型在训练集和测试集上的 NMSE 均较大，随机森林与 XGBoost 的结果比较相近，但 XGBoost 的效果更好，模型性能更优。

4.6.2 预测误差评估

预测误差是进行模型评估的关键性指标，也是机器学习中常用的评估方法。本文对智能手环销量数据进行建模分析，找出影响手环销量的重要因素，同时通过模型对手环销量进行预测。因此，判断模型预测的好坏是至关重要的。模型的预测精度越高，对于手环销量的建模就越可靠，参考价值就越大。对于模型预测误差评估的主要内容为：首先，将数据集按 7:3 的比例进行划分，即训练集与测试集的比例为 7:3；其次，在训练集上建立支持向量机、随机森林以及 XGBoost 模型，最后，在测试集上使用以下三种指标进行评估。

(1) 相对误差

相对误差是模型预测误差评估的一项基本指标，它是模型绝对误差与真实值的比值，其计算公式如下：

$$\delta = \frac{Y_i - \hat{Y}_i}{Y_i} \times 100\%, \tag{4-3}$$

其中Y_i为真实值，$\hat{Y}_i$为预测值，$Y_i - \hat{Y}_i$表示绝对误差。

相对误差的取值范围在[-10%,10%]之间较好，但是由于本文智能手环数据范围较广且影响因素较多，因此将误差范围控制在[-25%,25%]之间。

(2) 平均绝对百分比误差

平均绝对百分比误差（MAPE）也是预测误差评估常用的指标之一，计算公式如下：

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|(Y_i - \hat{Y}_i)|}{Y_i}, \tag{4-4}$$

其中Y_i为真实值，$\hat{Y}_i$为预测值。

该指标是绝对误差的绝对值与真实值比值的平均值。一般取值范围为[0,+∞)，若MAPE为 0，则该模型为完美模型；若MAPE大于 1，则该模型为劣质模型。因

此，MAPE的值越小，模型预测效果越好。

(3) 均方根误差

均方根误差（RMSE）在机器学中应用十分广泛，具体公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}\ . \tag{4-5}$$

均方根误差是指真实值与预测值偏差的平方与观测次数 n 的比值的平方根，常用来衡量预测值与真实值之间的偏差。同时均方根误差对于异常值较敏感，因此可以很好的反应数据的紧密度。均方根误差越小，模型的预测效果越好。

下面对于支持向量机、随机森林、XGBoost 三种回归模型，在测试集上进行预测，并计算出相对误差、MAPE 及 RMSE 的值，计算结果如表 4-8 所示。

表 4-8 评价指标计算结果

Table 4-8 Evaluation index calculation results

模型	相对误差	MAPE	RMSE
支持向量机	72.35%	29.44%	0.1180
随机森林	76.02%	24.66%	0.0936
XGBoost	77.72%	20.87%	0.0891

由上表 4-8 可知，在支持向量机、随机森林、XGBoost 三种回归模型中，XGBoost 模型在三个指标上的效果均达到最好，其相对误差在 25%范围内的数据为 77.72%，MAPE 为 20.87%，RMSE 为 0.0891。因此，通过三种模型的综合对比，可知 XGBoost 模型对于智能手环销量预测是最准确，且最值得推广与应用的。

4.7 本章小结

本章首先对于智能手环数据进行描述性统计分析，对于智能手环行业及用户研究具有较好的意义；其次，将手环数据按 7:3 的比例进行拆分，在训练集上分别建立支持向量回归模型、随机森林回归模型和 XGBoost 回归模型，并在测试集上进行预测；最后，通过 NMSE、相对误差、MAPE 及 RMSE 等评价指标的计算，综合对比出三种模型预测精度的高低。

结论与不足

（一）结论与建议

本文以智能手环为研究对象，使用 python 爬虫技术获得某商城 3128 条商品数据，包括品牌、价格、功能用途、好评度等 28 个特征。首先进行数据预处理，得到可以用于建模的样本数据 2546 条；然后使用三种特征选择方法，对变量进行特征选择，得到用于建模的 21 个特征；然后从品牌、功能、用户画像三个方面对手环的行业发展进行分析，为建模奠定基础；接着对智能手环销量数据进行建模分析，建立支持向量机、随机森林、XGBoost 三个模型，并对手环拟合效果进行分析；最后使用交叉验证和预测误差评估方法，对三种模型的预测效果进行对比评估，结论如下：

(1) 对模型的特征选择，可以有效的反映影响智能手环销量的重要影响因素。方差特征选择法筛选出 21 个变量，Lasso 特征选择法筛选出 15 个变量，Boruta 特征选择法筛选出 21 个变量。将三种方法进行对比分析，当变量至少有两种方法都选择时，我们将其纳入指标体系，最终从 28 个变量中筛选出 21 个变量。同时可以看到价格、店铺类型、运动模式识别、功能用途、好评度等特征被三种方法同时接受，说明用户对手环的价格、功能、评分等比较看重。

(2) 对模型的描述性统计分析中，从手环品牌、功能、用户画像三个角度进行研究。可以看出用户对小米、华为、荣耀等手环品牌的专注度较高，对于健康监测的功能需求较大。同时手环的用户画像表明，20~39 岁的男士占据手环 80%的市场，说明制造商及销售商应该向老人及女士的用户市场进行拓展，以获得最大的市场收益。

(3) 对于支持向量机模型，当模型的 type 为 nu-regression，kernel 为 radial，cost=10，gamma=0.0476(默认)时，模型的均方误差最小，为 0.0139。同时在 2546 条数据中，有 1420 个支持向量机，说明模型的拟合效果比较好，对于智能手环销量预测有一定的意义。

(4) 对于随机森林模型，当 ntree 为 500，mtry 为 10 时，因变量解释度达到最高为 74.45%，均方误差为 0.0108。与支持向量机相比，随机森林有更低的均方误差，拟合的模型更加准确。同时随机森林还给出了变量的重要性排序，可以看出，用户对于物流评分、价格、售后评分、品牌、商品评分、运动模式识别、好评度及评价详情的关注度较高。值得关注的是，物流评分是首要影响因素，因

此商家可以通过提高物流速度来获取更多的客户，从而实现更大的价值。

(5) 对于 XGBoost 模型，通过对 12 个参数进行调优，最终得到最优模型，均方误差为 0.0079。XGBoost 较支持向量机与随机森林有更低的均方误差。同时，对于特征的重要性分析中，评价详情、物流评分、售后评分、价格与品牌是手环销量的重要影响因素。说明除用户平时关注的品牌与价格，商品详情、物流及售后评分是十分重要的。

(6) 在模型评估与对比中，首先使用交叉验证的方法对模型进行评估，计算三种模型的 NMSE 值，结果发现 XGBoost 模型要优于支持向量机与随机森林模型。然后对模型的预测误差进行评估，通过相对误差、MAPE 和 RMSE 三个指标进行模型评估，结果显示 XGBoost 在三个指标上的预测误差均达到最优。因此，XGBoost 模型对于手环销量具有较好的预测功能，值得进行推广与应用。

(7) 本文通过对智能手环行业发展及销量建模进行分析，可以对制造商及销售商的手环销售提供一定的借鉴意义。对于手环销量的影响因素中，除基本的价格与品牌，用户更加倾向于关注手环的物流评价、评价详情、好评度、功能用途等因素。因此，销售商应该多采取措施来提高商品的配送速度及好评度。其次智能手环的用户群体可以向女性用户进行拓展，设计出美观且更加符合女性功能的智能手环。最后老人用户也是值得挖掘的市场，如今人们更加关心健康与养老，因此健康功能的研发是提高手环销量的重要举措。

（二）存在的不足

本文通过参考国内外文献，运用机器学习与数据挖掘的方法，对智能手环商品数据进行建模，得出以上结论与建议。但同时也有一定的不足之处：

(1) 在数据获取方面，本文智能手环商品数据通过 python 网络爬虫获得，因此数据具有较大的不确定性。对于指标的选取，通过国内外文献及商城页面获取，因此指标可能不够全面，不能详细的反应智能手环的影响因素。同时，手环销售商众多，本文只选取了某个商城进行数据分析，具有一定的局限性。

(2) 在建模方面，本文使用支持向量机、随机森林和 XGBoost 三种方法进行建模。但是数据挖掘方法众多，可能会有更加适合手环销量数据的模型，同时模型参数的调整可能也会存在可以改进的地方。因此，在建模方面还有很多需要改进的地方。

因此，本文对于手环行业发展分析及销量建模预测具有较好的研究价值，同时针对以上问题，我会在今后的学习中逐步优化与完善。

参考文献

[1] 候长海. 2015 年上半年可穿戴智能设备发展状况分析[J]. 互联网天地, 2015(08): 81-84.

[2] Adapa A. Factors influencing the adoption of smart wearable devices[D]. Rolla: Missouri University of Science and Technology, 2016: 30-37.

[3] 李金海, 何有世, 马云蕾. 大数据时代基于在线评论挖掘的企业网络口碑危机预警研究[J]. 情报杂志, 2015(2): 53-58.

[4] 涂海丽, 唐晓波, 谢力. 基于在线评论的用户需求挖掘模型研究[J]. 情报学报, 2015，34(10): 1088-1097.

[5] 房文敏, 张宁, 韩雁雁. 在线评论信息挖掘研究综述[J]. 信息资源管理学报, 2016(1):4-11.

[6] 孙晶晶, 田波, 何曙. 智能手环横向测评实验研究[J]. 日用电器, 2018(07): 11-14.

[7] 罗步操. “物联网+智能手环”智慧生态校园建设的探索[J]. 教育信息技术, 2018(Z2): 135-139.

[8] 刘大为, 蔡赛凤. 基于 iWatch 与小米手环对比分析的可穿戴设备用户采纳行为影响因素研究[J]. 生产力研究, 2016(11): 68-73.

[9] 吴江, 周露莎, 刘冠君. 基于 LDA 的可穿戴设备在线评论主题挖掘研究[J]. 信息资源管理学报, 2017, 7(03): 24-33.

[10] 陈华珍, 夏国清, 宗建华. 双自适应 BP 算法在智能手环中的应用[J]. 单片机与嵌入式系统应用, 2018, 18(08): 5-10.

[11] 王林, 胡梦迪, 朱文静. 运动社交平台对用户使用智能手环行为的影响研究[J]. 信息资源管理学报, 2017, 7(03): 5-14.

[12] 朱振涛, 李娜, 陈星光. 在线评论信息越完备越有用?——基于智能手环在线评论的实证数据[J]. 辽东学院学报(社会科学版), 2017, 19(05): 55-63.

[13] Venkatesh V, Morris M G, Davis G B , et al. User acceptance of information technology: Toward a unified view[J]. MIS Quarterly, 2003: 425-478.

[14] Netzer O, Feldman R, Goldenberg J, et al. Mine your own business: Market-structure surveillance through text mining[J]. Marketing Science, 2012, 31(3): 521-543.

[15] Jung Y, Kim S, Choi B. Consumer valuation of the wearables: The case of smartwatches[J]. Computers in Human Behavior, 2016, 63(1): 899-905.

[16] Kim K J, Shin D. An acceptance model for smart watches: Implications for the adoption of future wearable technology[J]. Internet Research, 2015, 25(4): 527-541.

[17] Yang H, Yu J, Zo H, et al. User acceptance of wearable devices: An extended perspective of perceived value[J]. Telematics and Informatics, 2016, 33(2): 256-269.

[18] Deng L, Turner D E, Gehling R, et al. User experience, satisfaction, and continual usage intention of IT[J]. European Journal of Information Systems, 2010, 19(1): 60-75.

[19] Wu L H, Wu L C, Chang S C. Exploring consumers'intention to accept smartwatch[J]. Computers in Human Behavior, 2016, 64(1):383-392.

[20] Chuah H W, Rauschnabel P A, Krey N, et al. Wearable technologies: The role of usefulness and visibility in smartwatch adoption[J]. Computers in Human Behavior, 2016, 65(1): 276-284.

[21] Erdem T, Keane M P. Decision-Making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets[J]. Marketing Science, 1996, 15(1): 1-20.

[22] Guyon I, Elisseeff A, Kaelbling L P. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3(6): 1157-1182.

[23] 邵玉娥, 王暕来, 周生华. 基于 LASSO 的雷达脉压压缩方法[J]. 电子科技, 2020(11): 1-6.

[24] Tibshirani R. Regression shrinkage and selection via the Lasso[J]. Journal of the Royal Statistical Society, 1996(58): 267-288.

[25] 钟金花. 基于 Lasso 方法的上海经济增长影响因素实证研究[J]. 统计与决策, 2013, 1(1): 154-156.

[26] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society, 2006, 68(1): 49-67.

[27] Wang H, Leng C. A note on adaptive group lasso[J]. Computational Statistics & Data Analysis, 2008, 52(12): 5277-5286.

[28] 罗昊. 基于自适应 LASSO 变量选择的 Logistic 信用评分模型研究[D]. 南京: 东南大学, 2016.

[29] Kursa M B, Rudnicki W R. Feature selection with the boruta package[J]. Journal of Statistical Software, 2010, 36(11): 1-13.

[30] 郭海山, 高波涌, 陆慧娟. 基于 Boruta-PSO-SVM 的股票收益率研究[J]. 传感器与微系统, 2018, 37(03): 51-53.

[31] Vapnik V. The Nature of Statistical Learning Theory[M]. Springer, 1995.

[32] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[33] Gareth James, Daniela Witte. 统计学习导论-基 R 应用[M]. 机械工业出版社, 2016.

[34] Breima L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[35] 张裕禄, 毕红葵, 叶泽浩. 基于随机森林的 HRGV 滑翔段飞行状态识别[J]. 战术导弹技术, 2020(02): 1-8.

[36] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 789-794.

[37] 张爱武, 董喆, 康孝岩. 基于 XGBoost 的机载激光雷达与高光谱影像结合的特征选择算法[J]. 中国激光, 2019, 46(04): 150-158.

[38] 叶倩怡. 基于 XGBoost 方法的实体零售业销售额预测研究[D]. 南昌大学, 2016.

[39] 张培荣. 基于 XGBoost 模型的企业财务危机预警研究[J]. 财会通讯, 2019(35): 109-112.

[40] Bauer E. An empirical comparison of voting classification algorithms: Bagging Boosting and Variants[J]. Kluwer Academic Publishers, 1999, 36: 105-139.

[41] 黄日康. 基于 XGBoost 算法的个人信用评分方案策划[D]. 上海师范大学, 2019.

[42] 王小宁, 黄俊文. R 语言实战[M]. 人民邮电出版社, 2016.

[43] 张良均, 云伟标, 王路. R 语言数据分析与挖掘实战[M]. 机械工业出版社, 2015.

[44] 赵萌. 基于网络搜索数据的游客量预测模型研究[D]. 西安理工大学, 2018.

[45] Shao J. Linear model selection by cross-validation[J]. Publications of the American Statistical Association, 1993, 88(422): 486-494.

致谢

在北京工业大学读研的两年生活即将结束，心中充满了不舍与感恩。不舍学校的图书馆与操场，感恩教导我的各位老师和我亲爱的同学们。两年的校园生活，学会了很多专业知识与技能，学会了很多为人处世的方法，受益终生。

首先，由衷的感谢我的导师关丽副教授。关老师无论在学习中还是生活中，都给予了我很大的帮助。在教学中，关老师上课认真，条理清晰，擅于带动课堂气氛，学习到很多专业知识，受益匪浅；在生活中，关老师就像我们的好朋友一样，愿意倾听我的难处，给出切实可行的指导建议；本次论文中，关老师更是非常负责，从开题报告，论文初稿，再到最后的定稿。关老师都非常有耐心，指出我在论文中存在的问题，提出可行的解决方案。特别感谢关老师的付出，谢谢您！

其次，感谢的我的任课老师及及本次论文参考文献的学者们。各位老师对我们的学习都非常负责，在两年的学习生活中教会我们统计学的基础理论，数据分析软件等，这些知识可以帮助我用统计学思维去思考问题，对今后的工作也有巨大帮助。感谢本次参考文献的学者们，通过你们的研究成果，我对于本次论文的写作有了较强的理论方法，使我能够运用丰富的知识进行建模分析。

最后，感谢我的父母与朋友们。感谢父母的培育之恩，是父母让我懂得要学会感恩，是你们给我创造了良好的学习环境，教导我做一个坚持不懈、奋发向上的人。感谢我身边的同学和朋友们，是你们陪我度过了难忘的研究生生活，在我伤心难过时，是你们对我进行无微不至的关怀，感谢有你们一直陪在我身边！感谢你们，在以后的学习及工作中，我一定会做的更好！