Statistics and Application 统计学与应用, 2022, 11(1), 38-45
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111005

Hans汉斯

基于决策树可视化分析大学生购买烟酰胺护肤产品的影响因素

李晓燕，张　霖，左　航

成都师范学院，四川　成都

收稿日期：2022年1月8日；录用日期：2022年1月21日；发布日期：2022年2月8日

摘　要

近年，随着护肤成为社会各界一大研究热点，烟酰胺作为新兴美白成份被护肤产品行业大量使用。然则烟酰胺护肤产品的购买状况受诸多因素影响。本文通过实地调研获取相关数据，基于数据特征的不连续性，以CART算法构造出决策树模型，并将决策树进行可视化分析，得出大学生群体购买烟酰胺护肤产品的影响因素。为更多大学生消费群体提供更优的个性需求及选择。

关键词

决策树，可视化，大学生群体，烟酰胺

Visual Analysis of Influencing Factors of College Students' Purchase of Nicotinamide Skin Care Products Based on Decision Tree

Xiaoyan Li, Lin Zhang, Hang Zuo

Chengdu Normal University, Chengdu Sichuan

Received: Jan. 8th, 2022; accepted: Jan. 21st, 2022; published: Feb. 8th, 2022

Abstract

In recent years, as skin care has become a hot research topic in the society, nicotinamide, as a new whitening ingredient, has been widely used in the skin care product industry. However, the purchase of nicotinamide skin care products is affected by many factors. This paper obtains relevant data through field research. Based on the discontinuity of data characteristics, decision tree model

文章引用: 李晓燕, 张霖, 左航. 基于决策树可视化分析大学生购买烟酰胺护肤产品的影响因素[J]. 统计学与应用, 2022, 11(1): 38-45. DOI: 10.12677/sa.2022.111005

was constructed by CART algorithm, and visual analysis of decision tree was carried out to obtain the influencing factors of college students' purchasing nicotinamide skin care products, for more college students consumer groups to provide better individual needs and choices.

Keywords

Decision Tree, Visualization College Students Group, Niacinamide

1. 研究背景

据中商产业研究院统计显示，近年烟酰胺作为护肤界“明星”，备受护肤行业宠爱。曾一度在国内外掀起热潮。作为护肤产品购买主力军之一的大学生，是一个乐于接受新鲜事物，消费欲望强烈的群体[1]。为有效抑制大学生违法贷款，引导大学生理性消费[2]。同时，也为坚持实现新理科教育改革与建设要求，理论与实践相结合[3]。本文将采用决策树可视化分析，研究影响大学生购买烟酰胺护肤产品的各类因素。为有护肤需求但消费能力有限的大学生消费群体，提供更具有性价比的护肤选择。

2. 实地调研

为有效且准确地了解大学生消费群体的护肤消费状况以及影响大学生护肤消费观念的因素，本文对多所高校大学生的护肤消费情况采取了实地问卷和访谈调查。

2.1. 设置调研报告

实地问卷和访谈调查所设问题围绕大学生群体日常护肤消费情况与相关选择展开。站在大学生消费群体的角度，从多个维度考虑影响大学生群体护肤消费的因素，以便于保持调研的客观可用性与全面覆盖性。调研报告的设置严格从三个维度展开：信度(reliability) [4]方面，实地问卷与访谈调查相结合，调研报告信息源自在校大学生对问卷的如实填写和访谈的如实回答，搜集的数据具有真实性与实时性。调研报告所设题目精简凝练，问题数量适中、答案详尽，可信度高；效度(validity)方面，本次调研报告发放问卷共计 900 份，筛除 52 份信息不完整问卷，有效问卷共计 848 份，问卷的效度高达 93%；维度(dimensionality)方面，问卷发放选取的范围为省内多所理工、艺术高等院校，具有广泛性和普遍性，调查数据借鉴度高。调研报告问题设置涉及 8 个维度，分别从消费者性别、消费者年龄、生活费用值、周边环境、选择购买的方式、更换速度、价格结算、购买的款式对被调查者进行纸质问卷调查或实地访谈调查。

2.2. 统计调研数据

经初步数据筛查，筛除存在问题与疑似存在问题问卷，对剩余 848 份有效问卷数据进行处理。利用 excel 数据处理功能[5]，将所得数据进行准确录入、分类整理与归类统计，再使用 excel 工具的作图功能[6]，将所得数据转化为饼状图[7]，得出大学生护肤消费情况(见图 1)。

对图 1 进行初步判断，发现不同影响条件下，大学生护肤消费特征各不相同。并对其可能原因进行了初步推测，结果如下。

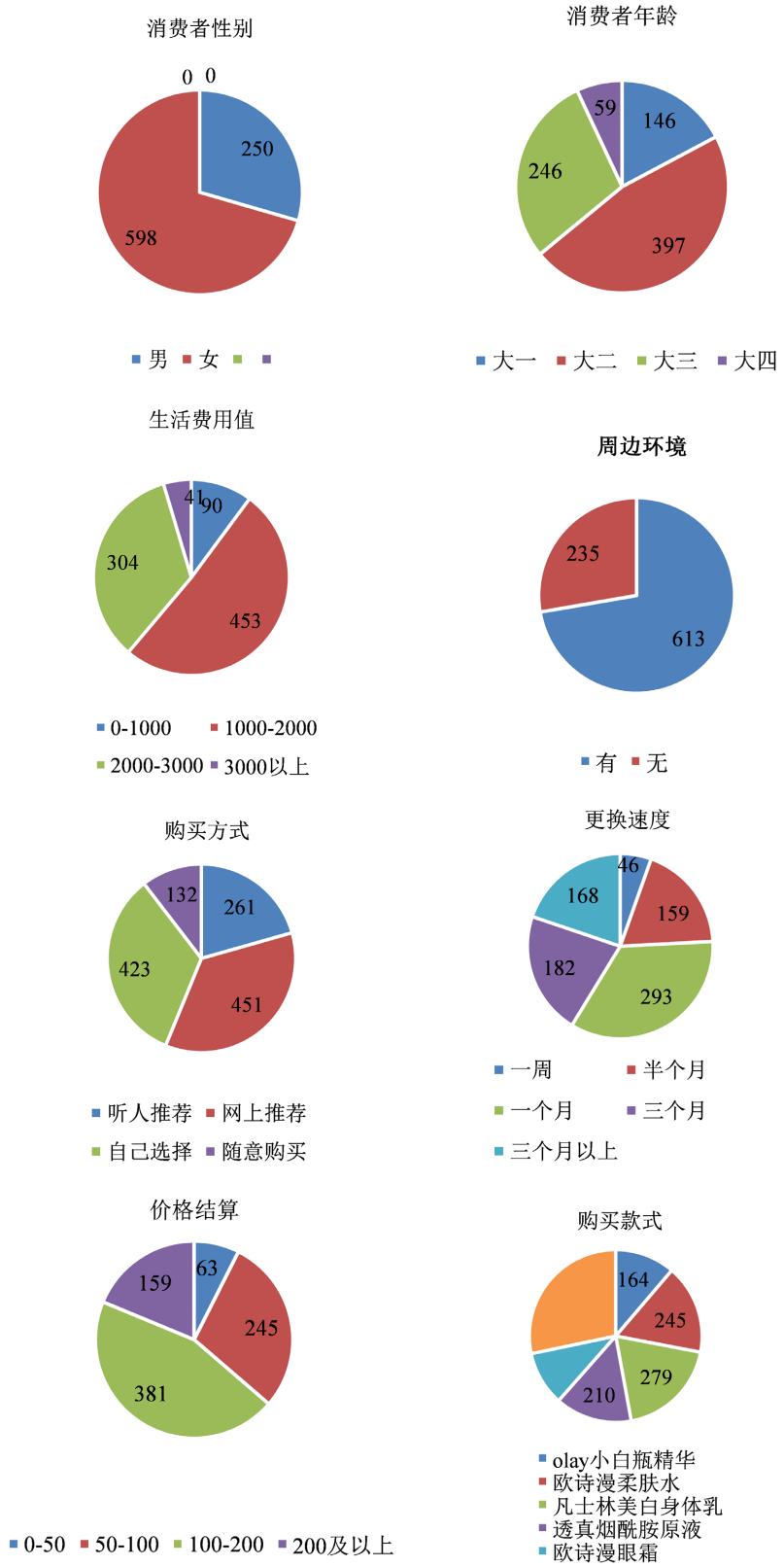


Figure 1. Consumption of skin care among college students
图 1. 大学生护肤消费情况

大学生护肤消费受性别影响程度大。女性大学生护肤消费群体约为男性大学生消费群体的 2.5 倍。这可能与东方传统审美观念有关。东方传统观念下“一白遮百丑”可能是刺激女性大学生护肤消费的一大原因；处于大二年龄阶段的大学生对护肤消费需求较高，处于大四年龄阶段的大学生对护肤消费需求反而较低。可能因为大二阶段大学生空闲时间多，然大四阶段大学生面临就业、考研等问题，日常护肤时间少，以致护肤消费降低；生活费用值在$(1000,2000]$区间的大学生护肤消费群体远大于生活费用值在$(2000,3000]$、$(3000,+\infty)$的大学生护肤消费群体。最合理的解释是大学生群体的生活费用值众数在$(1000,2000]$间；周边环境中有喜欢使用护肤产品的大学生约为无喜欢使用护肤产品的大学生的 3 倍。说明大学生护肤消费存在严重从众现象，极大程度不利于大学生群体理性消费；大学生消费群体购买护肤产品方式多样。网上推荐占比大，说明各大社交、美妆 app 会对大学生群体购买护肤产品产生较大影响；多数大学生一月更换一次护肤产品。说明大学生群体对护肤产品的需求量较大；大学生消费群体购买护肤产品主要价格结算区间为$(100,200]$。因此，大多中低端护肤产品更受大学生消费群体青睐；大学生购买护肤产品的款式多元化，众多护肤产品中有五类最受大学生群体喜爱。

3. 建立数学模型

为避免因选取的数学模型的不合适而造成问题处理不当、得出结论不可靠，进而无法较为准确的为大学生护肤群体提供具有可取性与可靠性的护肤消费建议，笔者将先对问卷调查所得数据进行分析与处理[8]。以便于准确的选取合适并且有效的数学模型。

3.1. 分析与处理数据

对图 1 所示数据进行初步分析，预计将此类影响烟酰胺护肤产品消费的实际问题转化为数学模型，再对相关数学问题进行分析与研究。由于大学生消费群体具有明确的消费目标和多种消费选择，并且大学生群体护肤消费情况的基本数据相互不连续且有明确的边界，属于典型的离散型数据[9]。又因决策树模型可以生成能够理解的相关规则，相关数据计算量相较不是很大，并且可以清晰的判断出各个影响因素间的重要程度。所以，本文决定选取决策树可视化模型[10]对影响大学生群体护肤消费的因素进行处理。于是据图 1 所示的统计数据，对所得的数据进行预处理[11]，分析出影响大学生护肤消费的因素(见图 2)。

编号	1	2	3	4	5	6	7	8
变量名	消费者性别	消费者年龄	生活费用值	周边环境	选择购买方式	更换速度	价格结算	购买款式

Figure 2. Factors affecting skin care consumption of college students
图 2. 影响大学生护肤消费的因素

由于变量的多元化及选择的多样性，决定采用 0-1 对大部分只有两个结果值的数据进行处理，采用自然整数对结果多样的数据或极个别只有两个结果值的数据进行处理。以数字代表各种变量的各种选择，以达到简化、量化数据的效果。处理如下：

购买款式中 olay 小白瓶精华、欧诗漫柔肤水、凡士林美白身体乳、透真烟酰胺原液、欧诗漫眼霜、其他分别以 0 代表不购买该产品，以 1 代表购买该产品。

消费者性别中男性、女性分别以 1、2 表示。

消费者年龄中大一、大二、大三、大四分别以 1、2、3、4 表示。

生活费用值$(0,50]$、$(50,100]$、$(100,200]$、$(200,+\infty)$分别以 1、2、3、4 代表。

周边环境无喜欢使用肤护品的用 0 代表，有喜欢使用肤护品的用 1 代表。

选择购买方式中听人推荐、网上推荐、自己选择、随意购买分别以 0 代表不选取该种方式购买，1 代表选取该种方式购买。

更换速度中一周、半月、一月、三月、三月以上分别以 1、2、3、4、5 代表。

量化数据后，将进行决策树以及可视化的程序编写。采取 cart 二叉树算法[12]。

3.2. 构建决策树

1) 确定训练数据集 D

大学生购买护肤产品价格各异，但聚集在$(50,100]$、$(100,200]$间较多。将大学生购买护肤产品价格结算区间分为：$(0,50]$、$(50,100]$、$(100,200]$、$(200,+\infty)$四段。以大学生购买护肤产品费用作为训练数据集 D。计算大学生购买护肤产品的价格结算区间的基尼指数：

基尼指数公式：

$$\mathrm{Gini}(D)=1-\sum_i P_i^2$$

由图 1 有：接受调查大学生人数为 848。其中，购买护肤产品价格结算在$(0,50]$、$(50,100]$、$(100,200]$、$(200,+\infty)$区间的大学生人数分别为 63、245、381、159。于是，令 D 为大学生购买护肤产品的价格结算区间，P_1、P_2、P_3、P_4 分别为购买护肤产品价格结算在$(0,50]$、$(50,100]$、$(100,200]$、$(200,+\infty)$区间的大学生的比例，则有：

$$P_1=\frac{63}{848},\quad P_2=\frac{245}{848},\quad P_3=\frac{381}{848},\quad P_4=\frac{159}{848},$$

$$\mathrm{Gini}(D)=1-\sum_{i=1}^{4}P_i^2=1-\left[\left(\frac{63}{848}\right)^2+\left(\frac{245}{848}\right)^2+\left(\frac{381}{848}\right)^2+\left(\frac{159}{848}\right)^2\right]=0.674$$

2) 将其他条件特征看成一个个节点。

将其余七个特征：消费者性别、消费者年龄、生活费用值、周边环境、选择购买的方式、更换速度、购买款式将作为指导之后众多节点分裂的依据。

3) 穷尽前特征的每种分割方式，找到最佳分割点。依次将各个数据划分成不同子节点，每次划分后计算所有子节点特征的基尼指数。

特征 a 下的基尼指数公式：

$$\mathrm{Gini}(D,a)=\sum_v\frac{|D_v|}{|D|}\mathrm{Gini}(D_v)$$

由图 1 有：

$$\begin{aligned}\mathrm{Gini}(D,\text{性别})&=\frac{63}{848}\times\left[1-\left(\frac{32}{63}\right)^2-\left(\frac{31}{63}\right)^2\right]+\frac{245}{848}\times\left[1-\left(\frac{79}{245}\right)^2-\left(\frac{166}{245}\right)^2\right]\\&\quad+\frac{381}{848}\times\left[1-\left(\frac{97}{381}\right)^2-\left(\frac{284}{381}\right)^2\right]+\frac{159}{848}\times\left[1-\left(\frac{42}{159}\right)^2-\left(\frac{117}{159}\right)^2\right]\\&=0.407\end{aligned}$$

$$\begin{aligned}\mathrm{Gini}(D,\text{周边环境})&=\frac{63}{848}\times\left[1-\left(\frac{38}{63}\right)^2-\left(\frac{25}{63}\right)^2\right]+\frac{245}{848}\times\left[1-\left(\frac{166}{245}\right)^2-\left(\frac{79}{245}\right)^2\right]\\&\quad+\frac{381}{848}\times\left[1-\left(\frac{283}{381}\right)^2-\left(\frac{98}{381}\right)^2\right]+\frac{159}{848}\times\left[1-\left(\frac{126}{159}\right)^2-\left(\frac{33}{159}\right)^2\right]\\&=0.395\end{aligned}$$

$$\begin{aligned}\text{Gini}\left(D,\text{生活费用值}\le 1000\right)&=\frac{63}{848}\times\left[1-\left(\frac{19}{63}\right)^2-\left(\frac{44}{63}\right)^2\right]+\frac{245}{848}\times\left[1-\left(\frac{30}{245}\right)^2-\left(\frac{215}{245}\right)^2\right]\\&\quad+\frac{381}{848}\times\left[1-\left(\frac{30}{381}\right)^2-\left(\frac{351}{381}\right)^2\right]+\frac{159}{848}\times\left[1-\left(\frac{17}{159}\right)^2-\left(\frac{142}{159}\right)^2\right]\\&=0.194\end{aligned}$$

$$\begin{aligned}\text{Gini}\left(D,1000<\text{生活费用值}\le 2000\right)&=\frac{63}{848}\times\left[1-\left(\frac{40}{63}\right)^2-\left(\frac{23}{63}\right)^2\right]+\frac{245}{848}\times\left[1-\left(\frac{113}{245}\right)^2-\left(\frac{132}{245}\right)^2\right]\\&\quad+\frac{381}{848}\times\left[1-\left(\frac{210}{381}\right)^2-\left(\frac{171}{381}\right)^2\right]+\frac{159}{848}\times\left[1-\left(\frac{70}{159}\right)^2-\left(\frac{152}{159}\right)^2\right]\\&=0.380\end{aligned}$$

$$\vdots$$

以此类推，可以得到 $\text{Gini}\left(D,2000<\text{生活费用值}\le 3000\right)$，$\text{Gini}\left(D,3000<\text{生活费用值}\right)$，$\text{Gini}\left(D,\text{听购买人推荐}\right)$，$\text{Gini}\left(D,\text{网上购买}\right)$，…

4) 从第三步穷尽的所有特征中，选出最佳特征及该特征最佳划分的方式，得出最终子节点。

基尼指数(Gini)是特征属性量化后的纯度值的表现方式之一。使用 Gini 来判断出当前数据集分割的特征属性：若基尼指数值越小，则该特征属性纯度越大，那么此属性就作为决策树的上层存在。因此，只需比较步骤 3 中所计算的所有的 $\text{Gini}\left(D,a\right)$，得到 $\min\left\{\text{Gini}\left(D,a\right)\right\}$。将 $\min\left\{\text{Gini}\left(D,a\right)\right\}$ 中的 a 作为当前节点的特征。

5) 对子节点重复进行 3~4 步，至每个最终子节点达到足够“纯”的地步。

以第二层节点到第三层节点为例：在步骤 3~4 中，通过 $\text{Gini}\left(D,a\right)=\sum_v\frac{\left|D_v\right|}{\left|D\right|}\text{Gini}\left(D_v\right)$ 确定第二层节点特征，进入第三层节点计算。计算方式同 3~4 步骤。通过比较各个特征基尼指数值大小，确定第三层节点特征。

接下来，第三层节点进行分裂，分裂依据同 3~4 步操作。

6) 当子节点达到一定“纯”度后，给定停止条件，终止决策树。

当 value 值不超过 2 时，认为子节点足够“纯”。强制终止决策树。

3.3. 可视化决策树

通过 3.2 得到决策树的拓扑结构，编写相应代码，依次录入调查所得数据。利用 python 的代码执行功能，键入代码与对应数据，并运行。再利用 graphviz 的画图功能，便可以得到可视化决策树模型[13] (见图 3)。对所得结果进行分析，从而得出消费情况与影响因素的相关关系，进而为大学生护肤群体提供实用性建议与意见。

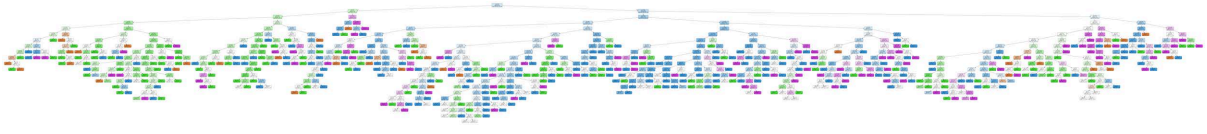


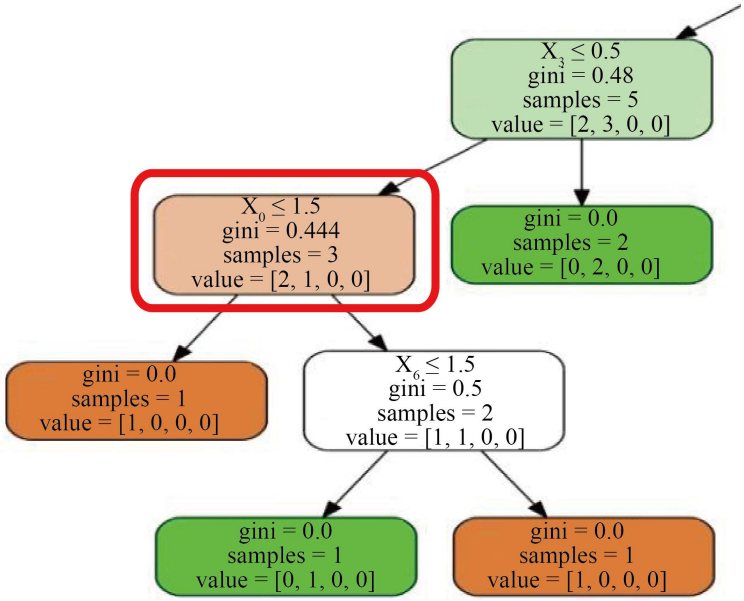
Figure 3. College students' skin care consumption decision tree
图 3. 大学生护肤消费决策树

4. 结论分析

决策树可视化分析

由于决策树分支过多，所以采用举例说明。

如图 4，x_i 表示诸多特征。以图 4 中 x_0 为例，“$x_0 \le 1.5$”，即某一特征 x_0 小于等于 1.5。某一特征即指例如“性别”或者“周边环境”等。不妨以性别为例。由于“男性大学生”在处理数据时以“1”表示，“女性大学生”在处理数据时以“2”表示。所以，“$x_0 \le 1.5$”表示的特征即男性大学生。当前节点以下的样本全是满足“男性大学生”这一特征的样本。当前节点之所以以男性大学生为分裂特征，是因为在计算本次分裂特征时“男性大学生”的 gini 指数等于 0.444，是本次分裂中基尼指数最小的特征。“samples = 3”，指在满足当前层次分裂的诸多分裂特征条件下，同时也满足当前层次分裂特征的人数有 3 人。若在此枝节中，将起始节点记为第 1 次分裂，此处“$x_0 \le 1.5$”的节点处记作第 n 次分裂。则“samples = 3”代表在被调查的 848 名大学生中满足前 n 个条件的大学生数目为 3。value = [2,1,0,0] 代表满足前 n 个条件的 3 名大学生中有 2 名大学生可接受护肤产品价格结算在 $(0,50]$，有 1 名大学生可接受护肤产品价格结算在 $(50,100]$，有 0 名大学生可接受护肤产品价格结算在 $(100,200]$，有 0 名大学生可接受护肤产品价格结算在 $(200,+\infty)$。由此可以看出，只需要根据可视化末端的 value 结果就可得出受各种特征影响的各个价格结算区间的大学生人数。又所有枝节由上到下的特征影响度是依次递减的。所以，可找出每个价格结算区间的大学生的护肤消费受到哪些特征的影响较大，哪些特征影响较小。根据这两个信息，便可统计分析出大学生群体购买烟酰胺护肤产品的影响因素。从而给大学生消费群体提供更具性价比的护肤消费建议。



注：1. gini—基尼指数。基尼指数最早被应用于经济学中，是用来衡量收入分配公平度的手段之一。Cart 算法的决策树中经常用基尼指数去度量待研究数据的不确定性或不纯度。基尼指数值越小，表明数据纯度越高。
2. samples—当前分流时的样本总数
3. value—class buys computer 属性的各个类型所包含的具体样本数。

Figure 4. Local diagram of college students' skin care consumption decision tree
图 4. 大学生护肤消费决策树局部图

5. 购买建议

于是，对图 3 进行分析，就可以得出不同条件下购买烟酰胺护肤产品的大学生比例。例如：价格结算在$(0,50]$、$(50,100]$、$(100,200]$、$(200,+\infty)$区间的人数比例分别为 7.43%，28.89%，44.93%，18.75%。价格结算在$(100,200]$区间条件下的男性大学生占比 24.45%，价格结算在$(100,200]$区间条件下的女性大学生占比 75.55%。价格结算在$(100,200]$区间条件下且生活费用值在$(1000,2000]$区间的女性大学生，占比 8.80%……根据比例给出大学生护肤消费建议：

1) 生活费用值在众多影响大学生护肤消费的因素中占据较高重要度。所以，大学生护肤消费群体在决定价格结算区间时最好优先考虑自己生活费用值区间。在所能承受范围内购买必须护肤产品。这在一定程度上利于减少大学生违法贷款现象。生活费用值较低的大学生可选择性购买中低端烟酰胺护肤产品。生活费用值较高的大学生可选择性购买中高端护肤产品，可以购买烟酰胺护肤产品，也可购买非烟酰胺护肤产品。

2) 周边环境在众多影响大学生护肤消费的因素中占据较为重要的位置。周边环境属于外在因素，却占据较为重要的位置，可见大学生心智不成熟，喜欢从众。因此，大学生应该理性消费，认准自己的肤质，精心挑选合适自己的护肤产品，切记盲目从众。

3) 消费者性别在众多影响大学生护肤消费的因素中也较为重要。女性大学生护肤消费群体数目远大于男性大学生护肤消费群体。可以看出，女性大学生更注重日常护肤。并且较大部分的女性大学生乐意购买具有美白功效的烟酰胺护肤产品。

4) 购买途径在众多影响大学生护肤消费的因素中占据较低的地位。因此建议大学生在购买护肤产品时少纠结于购买途径。

5) 在大学生购买的烟酰胺护肤产品中，烟酰胺含量为 2.5%的护肤产品更受大学生的喜爱。建议有购买烟酰胺护肤产品倾向的大学生购买烟酰胺含量为 2.5%的产品。

基金项目

1. 2021 年教育部产学合作协同育人项目：产教融合背景下理工科类管理专业人才培养体系建设(项目编号：202102155030)。

2. 成都师范学院 2020 年省级大学生创新创业训练计划项目：美白产品中烟酰胺用量的测量数据与功效价值的研究(项目编号：S202014389144)。

参考文献

[1] 赵书虹. 大学生群体消费行为特征及营销策略分析[J]. 学术论坛, 2014, 37(12): 57-61.

[2] 李嘉明, 康早早, 李希, 刘凤霞. 大学生生活费分配方式调查研究[J]. 劳动保障世界, 2019(3): 78+80.

[3] 张永明. 基于新教改的数学教学探讨[J]. 数理化学习, 2010(10): 43-44.

[4] 屈芳, 马旭玲, 罗林明. 调查问卷的信度分析及其影响因素研究[J]. 继续教育, 2015, 29(1): 32-34.

[5] 周文玉. 数据处理中 Excel 的应用[J]. 电子技术与软件工程, 2017(19): 166.

[6] 张桥珍. 浅论 Excel 图表制作四步曲[J]. 数字技术与应用, 2019, 37(1): 228-229.

[7] 刘敬伟. Excel 数据处理常用方法分类解析[J]. 电脑知识与技术, 2014, 10(2): 426-427+433.

[8] 李根, 邹国华, 张新雨. 高维模型选择方法综述[J]. 数理统计与管理, 2012, 31(4): 640-658.

[9] 王君. 离散型分布的发散思维探究性教学设计[J]. 高等数学研究, 2015, 18(1): 115-117.

[10] 王秀岩. 决策树算法及其应用[J]. 电子技术与软件工程, 2014(5): 189.

[11] 黄雄, 宋中山, 刘少英. 决策树的数据预处理[J]. 软件导刊, 2009, 8(10): 32-35.

[12] 叶萌. 决策树学习研究综述[J]. 黑龙江科技信息, 2011(34): 22+156.

[13] 李乐茹. 决策树可视化系统模型研究[J]. 长江大学学报(自然科学版)理工卷, 2009, 6(2): 265-267.