基于熵权–变异法对不同汽车品牌满意度影响因素评价研究

赵永刚[1]，于珍[2]

（1.广汽丰田汽车有限公司，广州 511458；2.中汽信息科技（天津）有限公司，天津 300300）

[摘要]新能源汽车产业是事关国民经济发展的战略性新兴产业，对其市场销售进行科学决策具有重要意义。本文建立数学模型，研究了电动汽车不同品牌的评价结果，并判断影响因素对市场销售进行科学决策。首先，对数据进行预处理，根据 KNN 算法计算数据集内部的距离，将偏离集群的数据作为异常剔除；然后，计算各个指标的均值、最值、标准差等有关统计量；其次，针对客户对于电动汽车的评价，采用了熵权法–变异系数法将八个指标通过相应的权重组合为客户对汽车的综合评价得分，发现最终的评价结果合资品牌好于自主品牌好于新势力品牌；最后，通过对用户自身特征分析，是否购买汽车高度相关的维度是家庭年收入、房贷和车贷的支出。

关键词：KNN；熵权法；变异系数；权重

Research on Evaluation of Factors Influencing Satisfaction of Different Automobile Brands based on Entropy Weight-Variation Method

Zhao Yonggang[1], Yu Zhen[2]

(1. GAC TOYOTA Motor Co., Ltd., Guangzhou 511458;
2. China Auto Information Technology (Tianjin) Co., Ltd., Tianjin 300300)

[Abstract] This paper establishes a mathematical model, studies the evaluation results of different brands of electric vehicles, and judges influencing factors to make scientific decisions on market sales. First, preprocess the data, calculate the distance within the data set according to the KNN algorithm, and eliminate the data that deviates from the cluster as anomalies; then, calculate the mean, maximum, standard deviation and other related statistics of each indicator;Secondly, for customers' evaluation of electric vehicles, the entropy method-coefficient of variation method was used to combine the eight indicators into the customer's comprehensive evaluation score for the vehicle through the corresponding weights. It was found that the final evaluation result of joint venture brands is better than independent brands, independent brands are better than new power brands;Finally, through the analysis of the user's own characteristics, the highly relevant dimensions of whether to buy a car are the family's annual income, housing loan and car loan expenditure.

Keywords: KNN, entropy weight, coefficient of variation, weight

0 引言

汽车产业是国民经济的重要支柱产业，发展新能源汽车是我国从汽车大国迈向汽车强国的必由之路，是应对气候变化、推动绿色发展的战略举措。经过 20 年研发和示范推广，我国新能源汽车产业已初具产业规模和技术优势。众多汽车企业积极发展新能源电动汽车，新能源汽车产销展现出勃勃生机。然而，电动汽车作为新兴事物，消费者在电池等领域仍对其存在一些疑虑，新能源汽车的营销成效一直以来都无法达到相关部门的预期。因此，需要建立合适的数学模型，探讨购买电动车的影响因素在电动车的自身产品特征以及用户自身的维度特征上，哪些特征是主要的购车影响，对其市场销售进行科学决策[1-4]。

1 国内外研究现状

熵权法–变异法方法经常被国内外学者用在评价研究项目上，如陈红[5]等人利用熵权法和变异系数法计算权重，得到的权重解决了单一客观权重分配

不合理的问题,使分配的权重具有协调性;吴艳霞[6]等人运用熵权法与变异系数组合方法,对西部地区高新技术产业产出能力、技术创新投入以及环境等要素进行评价;王建华[7]等人提出综合结构熵权法和变异系数法算出指标的综合权重,算出优异度,其模型的准确性和精度相比之前方法大幅提高;基于熵权法–变异法在评价领域的出色变现,因此本文提出了基于熵权法–变异法对不同汽车品牌的评价研究,给出企业相关部门提供科学的决策。

2 研究方法

本文数据来源于大约 5000 份调研数据,每一份数据都涉及用户个人特征的信息以及用户对新能源汽车 8 个技术指标进行评价(评价方式:1–10 分进行打分),同时每份问卷都有用户对三个不同品牌车型的打分(合资品牌、自主品牌以及新势力品牌),每个品牌所选的车型都是新能源紧凑型轿车,新能源汽车 8 个技术指标如下表 1 所示,个人特征维度如下表 2 所示。

表 1 汽车产品的 8 个指标

产品指标维度编码	产品指标维度名称
品牌 1	合资品牌
品牌 2	自主品牌
品牌 3	新势力品牌
a1	电池耐用耐用性
a2	舒适性满意度
a3	经济性满意度
a4	安全性满意度
a5	动力性满意度
a6	驾驶操满意度
a7	外观内饰满意度
a8	配置与质量品质满意度

表 2 用户个人特征属性

数据维度编码	数据维度名称
B1	户口
B2	居住年限
B3	居住区域
B4	驾龄
B5	家庭人口数
B6	婚姻
B7	拥有孩子数量
B8	出生年份
B9	学历
B10	工作年限
B11	单位性质
B12	职位
B13	家庭年收入
B14	个人年收入
B15	可支配年收入
B16	房贷费用
B17	车贷费用

2.1 数据预处理

数据预处理在这里指的是对异常值的处理,对异常值的处理采用基于距离的异常检测方法——KNN 算法来计算数据集内部的距离,并将偏离集群的数据作为异常值剔除。KNN 异常点检测方法计算步骤如下:

(1)计算各样本点两两之间的距离矩阵;

(2)依据距离矩阵建立 k 邻近列表;

(3)计算每个点与 k 个最邻近距离之和 Sum_KNN;

(4)检测异常点,确定阈值 t,只要 Sum_KNN> t,则该样本点被认为是异常点,并对该异常点进行标记;

(5)删除异常数据;

在文中,我们选择 k=10,并以最邻近距离之和 Sum_KNN 的均值的 2 倍作为阈值 t, 若 Sum_KNN> t,则将该样本点作为异常点进行标记并剔除。异常点检测结果如下图 1 所示。其中,红色标记的 k 个最邻近距离之和超出阈值, 将其剔除后再进行后续的数据处理。

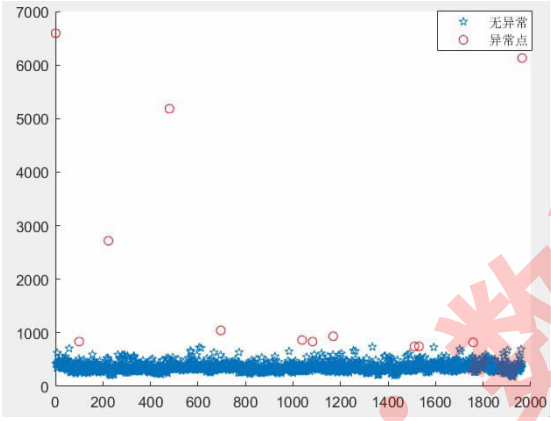

图 1 异常点检测情况

2.2 数据统计分析

首先根据原始数据, 对客户对不同品牌的满意度数据进行简单统计量的计算和求解。每个品牌 a1–a8 各个满意度的平均值和方差刻画了用户对于 8 种电动汽车性能指标满意度的平均水平和波动程度, 满意度的最大值和最小值刻画了用户对于该项性能的认可度的最高与最低水平。每个品牌总体满意度的平均值代表了用户对于该品牌汽车的平均满意水平,各项统计量计算结果如下,其中 3 个品牌的每个指标满意度平均值如下表 3,3 个品牌的每个指标满意度方差值如下表 4,3 个品牌的每个指标满意度得分最高值如下表 5,3 个品牌的每个指标满意度得分最低值如下表 6。

表 3 满意度平均值

满意度得分平均值	a1	a2	a3	a4	a5	a7	a8	总体平均值
合资品牌	78.14	78.82	76.16	79.55	78.04	78.79	78.44	78.29
自主品牌	78.07	78.04	76.06	78.77	77.10	77.95	77.41	77.66
新势力品牌	77.20	77.46	74.68	77.96	75.93	77.68	77.35	76.93

表 4 满意度方差

满意度得分方差	a1	a2	a3	a4	a5	a6	a7	a8
合资品牌	84.23	79.76	107.08	82.41	86.36	86.68	85.48	97.35
自主品牌	70.41	77.64	105.41	76.54	83.53	81.35	74.55	81.89
新势力品牌	92.72	91.45	115.63	102.02	93.34	100.52	91.77	107.65

表 5 满意度得分最高值

满意度得分最高值	a1	a2	a3	a4	a5	a6	a7	a8	总体最高值
合资品牌	99.04	99.03	99.03	99.98	99.98	99.99	99.99	99.98	99.99
自主品牌	99.04	99.03	99.03	99.98	99.98	99.99	99.99	99.98	99.99
新势力品牌	99.04	99.03	99.03	99.98	99.98	99.99	99.99	99.98	99.99

表 6 满意度得分最低值

满意度得分最低值	a1	a2	a3	a4	a5	a6	a7	a8	总体最低值
合资品牌	37.04	43.40	37.09	40.70	44.43	41.57	44.44	41.10	37.04
自主品牌	40.36	44.45	29.59	40.75	36.40	39.15	44.44	44.43	29.59
新势力品牌	55.58	48.15	40.70	44.43	50.53	50.59	40.06	44.43	40.06

从上述简单统计量结果分析可知，用户对于三种品牌的满意度差距不大，总体满意度平均值分别为 78.29,77.6,76.93,合资品牌的满意度略高于自主品牌，略高于新势力品牌。对于合资品牌,用户对于其安全性(79.55 分)、舒适性(78.82 分)、外观内饰(78.79 分)整体表现的平均满意度最高,其中安全性和舒适性满意度的波动水平最小。对于自主品牌,用户对于其安全性 (78.77 分)、电池技术性能(78.07 分)、舒适性(78.04 分)整体表现的平均满意度最高,其中电池技术性能的波动水平最小。对于新势力品牌,用户对于其安全性(77.96 分)、外观内饰(77.68 分)、舒适性(77.46 分)整体表现的平均满意度最高,其中舒适性和外观内饰的满意度波动水平最小。

2.3 不同品牌汽车满意度比较分析

针对客户对于不同品牌的电动汽车的满意度比较分析,可以采用熵权法/变异系数法将八个指标通过相应的权重组合为客户对汽车的综合评价。

2.3.1 变异系数法确定指标权重

变异系数法是常用的客观赋权方法之一。变异系数法所确定的各指标权重是由通过衡量其观测值变动程度得到的。在处理数据的过程中,变异系数法最大限度地保留了原始数据信息,这也正是该方法的最大优点。对于本文所研究满意度指标权重的计算,利用变异系数法的实际操作步骤如下:

首先计算指标的变异系数 v_j，计算公式如下公式(1)所示:

$$v_j=\frac{\sigma_j}{\overline{x}_j} \tag{1}$$

其中第 j 项满意度指标各观测值的平均数记作 $\overline{x}_j$,第 j 项满意度指标各观测值的标准差记作 σ_j。对变异系数进行归一化处理，进而得到各指标的权重 w_j,计算公式如下公式(2):

$$w_j=\frac{v_j}{\sum v_j} \tag{2}$$

通过变异系数法对各指标赋权，得到各指标的权重，即各项满意度指标对用户总体满意度影响的重要程度,如下表 7 变异系数法得到的权重。

表 7 变异系数法得到的权重

指标	a1	a2	a3	a4	a5	a6	a7	a8
权重	0.118	0.120	0.144	0.120	0.126	0.124	0.120	0.128

2.3.2 熵权法确定指标权重

熵权法是一种客观赋权方法。其依据的原理为,指标的变异程度越小,所反映的信息量也越少,其对应的权值也应该越低。利用熵权法的实际操作步骤如下:

Step1:选取 n=3 个品牌,每个品牌选取上述 m=8 个指标作为变量，以 x_{ij} 表示第 i 个品牌的第 j 个指标的数值;

Step2: 计算第 j 项指标下第 i 个样本值占该指标的比重;

$$p_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}},i=1\cdots n,j=1\cdots m \tag{3}$$

Step3:计算第 j 项指标的熵值;

$$e_j=-k\sum_{i=1}^{n}p_{ij}\ln(p_{ij}),j=1\cdots m \tag{4}$$

其中,$k=\frac{1}{\ln(n)}>0$,满足 $e_j\geqslant 0$

Step4:计算信息熵冗余度;

$$d_j=1-e_j,j=1\cdots m \tag{5}$$

Step5:计算各项指标的权值;

$$w_j=\frac{d_j}{\sum_{j=1}^{m}d_j},j=1,2\cdots m \tag{6}$$

综上所述,我们采用 MATLAB 代入公式进行求解,结果如下表 8。

表 8 熵权法得到的权重

指标	a1	a2	a3	a4	a5	a6	a7	a8
权重	0.110	0.115	0.166	0.114	0.127	0.123	0.114	0.130

对比变异系数法和熵权法所得权重，可以发现a3、a5、a8的权重最大，即经济性(耗能与保值率)整体满意度、动力性表现(爬坡和加速)整体满意度、配置与质量品质整体满意度对客户整体满意度贡献了较大信息，对整体满意度其起较大影响。本文以变异系数法所得权重，作为综合评价指标的权重进行计算。

2.3.3 综合评价模型

根据上述方法得出的指标权重，对目标研究对象进行综合评价并排序，得到三种品牌满意度的综合评价结果，计算公式如下公式(7)：

$$c_i=\sum_{j=1}^{8}w_j\times\bar{x}_{ij},i=1,2,3 \tag{7}$$

其中，C_i 表示第 i 个品牌的综合评价值，w_j 表示第 j 项满意度指标的权重，$\bar{x}_{ij}$ 表示第 i 个品牌第 j 种满意度指标的平均值。带入数据求解，得到合资品牌(品牌 1)、自主品牌(品牌 2)和新势力品牌(品牌 3)的综合评价得分分别为 78.18，77.56，76.80，即最终的评价结果为品牌 1 略好于 2 略好于 3。

2.4 用户自身因素分析

2.4.1 变量共线性检验

通过观察数据发现，客户特征信息中存在关联性较大的信息，故推测变量间可能存在多重共线性。鉴于变量的分布未知，可以使用数据的排序等级来反映变量间是否具有相同或相反的趋势来反映它们之间的关联性。因此本文采用斯皮尔曼相关系数分析客户自身条件的因素相关性(共线性)，斯皮尔曼相关系数对变量的总体分布形态和样本容量大小不做要求，能很好地满足本题中共线性诊断需求。

斯皮尔曼相关系数使用排序的方法消除量纲，在相关性分析中，用数据大小的排序代替原始的数据，起到了消除量纲的作用。斯皮尔曼相关系数的公式(8)：

$$r_{sp}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \tag{8}$$

其中 n 是样本的数量，d_i 代表数据样本之间的等级差。在 MATLAB 中，我们直接使用 corr 指令计算斯皮尔曼相关系数。

在斯皮尔曼相关系数的计算中，由于样本带有随机性，得到了数值以后也无法知晓到底相关系数多大才是相关性强，多小才是相关性弱，为了表明强弱关系，我们需要引入假设检验的方法。设定原假设为 H0：研究的总体之间无相关，备择假设为 H1：研究的总体之间有相关。在文中样本数量远大于 30，属于大样本的情况，可以通过公式(9)服从 $t(n-2)$ 的 t 分布：

$$t=r_{sp}\sqrt{\frac{n-2}{1-r_{sp}^2}} \tag{9}$$

采用双侧 t 检验，在得到的 p 值中，如果 p 的值大于 0.05，则没有显著性差异，也就是说没有理由认为显著性差异存在，即没有相关性。如果 p 值小于 0.05 的话，我们可以认为存在显著性的差异。

将客户特征信息数据代入计算，利用 MATLAB 函数计算斯皮尔曼相关系数。结果表明，客户数据中 B5、B6、B7；B4、B8、B10；而 B13、B14、B15 间 p 值大于 0.05，即这些变量间存在较强的共线性。

2.4.2 Lasso 回归模型的建立

在多元线性回归模型中由于存在异方差和多重共线性对模型的影响，变量过多会导致多重共线性问题造成的回归系数的不显著。Lasso 是另一种数据降维方法，在多元线性回归模型的损失函数上加上了惩罚项，基于惩罚方法对样本数据进行变量选择，通过对原本的系数进行压缩，将原本很小的系数直接压缩至 0，从而将这部分系数所对应的变量视为非显著性变量，将不显著的变量直接舍弃[8-11]。

Lasso 方法是在普通线性模型中增加 L2 惩罚项，对于普通线性模型 Lasso 估计如下公式(10)：

$$\hat{\beta}_{Lasso}=\arg\min_{\beta\in R^d}\left(\left\|Y-X\beta\right\|^2+\lambda\sum_{j=1}^{d}\left|\beta_j\right|\right) \tag{10}$$

其中，t 与 λ 一一对应，为调节系数。

令 $t_0=\sum_{j=1}^{d}|\hat{\beta}_j(OLS)|$，当 $t<t_0$ 时，一部分系数就会被压缩至 0，从而降低 X 的维度，达到减小模型复杂度的目的。

利用 MATLAB 的 Lasso 函数库代入用户数据进行求解,得到筛选结果。得到 MSE 指标变化曲线和 MSE 关于自变量取值数目变化曲线如下图 2 和图 3 所示。

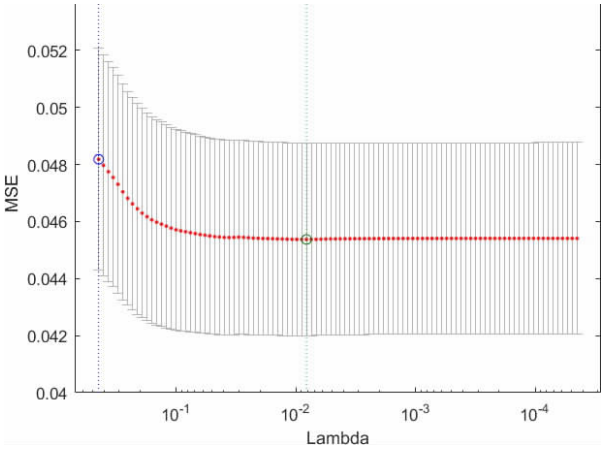

图 2 MSE 指标变化曲线

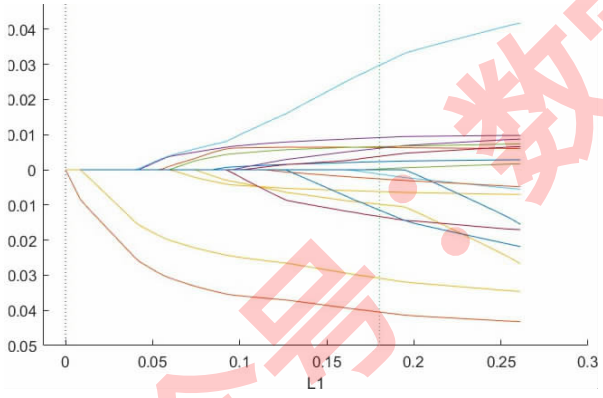

图 3 MSE 关于自变量取值数目变化曲线

结论:除 B2、B11、B13、B16、B17 外,其他项的系数均被压缩为 0,即被舍弃。得到最终与客户是否购买汽车高度相关的客户信息特征因素为 B2、B11、B13、B16、B17。

3 结束语

根据 a1–a8 这 8 个指标可以得出各种车品牌的综合评价得分,占比最主要的几个成分为 a3、a8、a5,即客户对于新能源电动汽车的实用性、经济性性能最为关注。因此销售部可以针对经济性、车辆性能进行推销,大力宣扬电动汽车相比传统汽汽油车的优点,主要体现在电动汽车有不输于传统汽油车的加速、巡航以及其优秀的节能特点;可以在允许范围内购车赠送小礼品、小额车险、免费保养等变相减少购车成本,从而影响客户对于电动汽车的评价结果进而影响购买意愿。

根据 B1–B17 的数据分析发现对于潜在客户自身情况的挖掘也是销售部需要做的事情之一,即建立目标客户群体的画像:如有机动车通勤需求(城市居住因素)、个人或家庭购车需求的,较年轻(对于电动汽车的接受度较高)、有一定存款负债率较低的群体。同时,销售部门可以在目标客户群体中投放广告,增加品牌知名度,吸引潜在客户购买电动汽车。

参考文献

[1] 牛丽薇. 新能源汽车购买意愿的影响因素及引导政策研究[D]. 徐州: 中国矿业大学, 2015.
[2] 赵斌. 比亚迪新能源汽车消费的影响因素研究[D]. 长沙: 中南大学, 2010.
[3] 张婕. 中国自主品牌新能源汽车购买意愿影响因素分析[D]. 北京: 北京交通大学,2019.
[4] 谭慧. 消费者购买新能源汽车偏好及影响因素研究[D]. 镇江: 江苏科技大学, 2014.
[5] 王月辉, 王青. 北京居民新能源汽车购买意向影响因素——基于 TAM 和 TPB 整合模型的研究[J]. 中国管理科学, 2013, 21(S2): 691–698.
[6] 陈红光, 李晓宁, 李晨洋. 基于变异系数熵权法的水资源系统恢复力评价——以黑龙江省 2007–2016 年水资源情况为例[J]. 生态经济, 2021, 37(1): 179–184.
[7] 吴艳霞, 周春光. 西部地区高新技术产业技术创新能力评价研究——基于熵权与变异系数组合赋权法的综合评价模型[J]. 新疆农垦经济, 2018(12): 83–88.
[8] 王建华, 杨静. 基于结构熵权法和改进 TOPSIS 法的可持续供应链绩效评价模型与算法[J]. 中国市场, 2013(26): 15–20.
[9] 方匡南, 章贵军, 张惠颖. 基于 Lasso–logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济研究, 2014, 31(2):125–136.
[10] 李宝东, 宋瀚涛. 数据挖掘在客户关系管理(CRM)中的应用[J]. 计算机应用研究, 2002(10): 71–74.
[11] 朱浩刚, 孙煜鸥, 戴伟辉. 基于数据挖掘的移动通讯业客户流失管理[J]. 计算机工程与应用, 2004(1): 215–219.