所属类别	2021 年“华数杯”全国大学生数学建模竞赛	参赛编号
研究生组		CM210528

电动汽车目标客户销售策略研究

摘要

汽车产业是国民经济的重要支柱产业，而新能源汽车产业是战略性新兴产业。大力发展以电动汽车为代表的新能源汽车是解决能源环境问题的有效途径，市场前景广阔。但是，电动汽车毕竟是一个新兴的事物，与传统汽车相比，消费者在一些领域，如电池问题，还是存在着一些疑虑，其市场销售需要科学决策。

不同用户满意度比较（问题一）本文首先对数据库原始数据进行简单的统计描述，发现存在异常值和缺失值。进一步对数据清洗工作，利用箱线图法和标准差法检测异常值并用分位数补齐。修正异常值后利用相关变量与缺失变量的关系对缺失变量进行初步补齐，再用随机森林拟合算法对缺失数据最终补齐。对补齐后的数据做描述性统计分析，进一步利用熵值法对目标客户对于不同品牌汽车满意度做比较分析，发现用户对不同品牌汽车满意度存在差异。

不同品牌电动汽车销售影响因素（问题二）决定目标客户是否购买电动车的影响因素有很多，有电动汽车本身的因素，也有目标客户个人特征的因素。本文采用分类回归树、袋装技术和推进技术算法来获取影响电动汽车销售因素的客观特征重要性权重值，之后比较算法的预测精度等影响因素从而得到影响不同品牌销售的因素。

不同品牌电动汽车的客户挖掘模型（问题三）为更好对客户进行挖掘和预测以提高电动车的产量，本文采用随机森林（Random Forenst）、逻辑回归（Logistics Regression）、多层感知机（MLP）、袋装法（Bagging）和推进法（AdaBoost）等五种模型对 15 名目标客户购买情况进行预测，并对 5 种方法进行模型评价和预测准确度比较。综合考虑预测时间和精确度等因素，结果显示多层感知机（MLP）和推进法（AdaBoost）预测精度较好。运用该用模型对 15 目标名客户进行购买预测，发现有两名目标用户选择购买，对这两名目标用户做进一步分析，归纳出新能源汽车的目标客户人群以及新能源汽车需要着力改进的方向。

对目标客户实施销售策略（问题四）由于满意度的调整存在服务难度成本，企业需要知道调整不同品牌的哪些性能指标才能留住特定目标客户，实现针对性的性能升级和销售策略。通过调整不同品牌的三个预测未购买的特定客户的满意度得分百分点，来进行 AdaBoost 预测实验模拟，并进一步提出针对特定客户的营销策略。

根据研究结果提出建议（问题五）根据前文研究结果，本文对于三种不同品牌从电动汽车自身性能提升和目标客户群体选择两方面提出了相关建议。针对不同品牌电动汽车，从其相对应的主要影响因素着手，改善服务，提升用户满意度，从而增加最终成交量。

关键词：熵值法　推进法　随机森林算法　　逻辑回归　分类预测

一、前言

1.1 研究背景

汽车产业是国民经济的重要支柱产业，而新能源汽车产业是战略性新兴产业。大力发展以电动汽车为代表的新能源汽车是解决能源环境问题的有效途径，市场前景广阔。但是，电动汽车毕竟是一个新兴的事物，与传统汽车相比，消费者在一些领域，如电池问题，还是存在着一些疑虑，其市场销售需要科学决策。

1.2 问题重述

问题一：对数据集进行初步分析，采取合理的方法对异常值和缺失值进行处理。同时对数据集进行必要的数据清洗，以便于后续的分析。此外通过描述性统计，比较分析目标客户对合资品牌(1)、自主品牌(2)和新势力品牌(3)电动汽车在电池技术性能(a1)、舒适性(a2)、经济性(a3)、安全性(a4)、动力性(a5)、驾驶操控性(a6)、外观内饰(a7)和配置与质量品质(a8)等八方面的满意度。

问题二：根据目标客户对电动汽车电池技术性能(a1)、舒适性(a2)、经济性(a3)、安全性(a4)、动力性(a5)、驾驶操控性(a6)、外观内饰(a7)和配置与质量品质(a8)等八方面的满意度评分和目标客户的户口类型(b1)、本城市居住年限(b2)、居住区域(b3)、驾龄(b4)、家庭共同生活人数(b5)、婚姻家庭情况(b6)、拥有孩子数量(b7)、年龄(b8)、最高学历(b9)、工作年限(b10)、工作单位性质(b11)、职位(b12)、家庭收入(b13)、个人年收入(b14)、可支配年收入(b15)、全年房贷支出占家庭总收入的比例(b16)和全年车贷占家庭总收入比例(b17)等个人特征信息，结合目标客户最终是否购买电动汽车的结果，探讨不同品牌电动汽车的销售影响因素。

问题三：基于上述研究，建立不同品牌电动汽车的客户挖掘模型并对模型优良性进行评价，从而预测附件 3 中 15 名目标客户购买电动汽车的可能性。

问题四：销售部门认为，营销者通过改善服务有可能在短时间内提高 a1-a8 五个百分点，但服务难度与提高的满意度百分点成正比。基于上述研究结果，分别选取三个品牌各 1 名未购买电动汽车的客户进行分析，提出营销建议。

问题五：根据前面的研究结论，为销售部门写一封不超过 500 字的建议信。

二、模型假设

（1）假设附件中提供的目标客户满意度得分数据和用户个人特征数据真实有效；

（2）假设目标客户能够对所体验电动汽车进行公平公正地评价，所给出的满意度得分基于真实感受；

（3）假设目标客户能够独立地对所体验汽车的满意度进行评价；

（4）假设目标客户可以准确描述电动汽车的体验感，其满意度评分可以准确反映用户真实感受；

（5）假设电动汽车的销售情况只取决于附件提供的数据，不受其它因素影响

三、模型符号说明

变量符号说明

变量	变量解释
n	目标客户编号
i	品牌类型
a1	电池技术性能
a2	舒适性
a3	经济性
a4	安全性表现
a5	动力性表现
a6	驾驶操控性
a7	外观内饰
a8	配置与质量品质
b1	户口情况
b2	城市居住
b3	居住在以下哪个区域
b4	驾龄
b5	家庭成员个数
b6	婚姻家庭情况
b7	子女个数
b8	年龄
b9	最高学历
b10	工作年限
b11	工作单位性质
b12	职位
b13	家庭年收入
b14	个人年收入
b15	家庭可支配年收入
b16	房贷支出占比
b17	车贷支出占比
Willing	购买意愿

四、问题一的模型建立和求解

4.1 问题分析

问题一要求对数据集进行初步分析，采取合理的方法对异常值和缺失值进行处理。同时对数据集进行必要的数据清洗，以便于后续的分析。为对数据进行清洗，可以采用箱线图法和标准差法对异常值进行检测，采用 99 分位数和 1 分位数对异常值进行替换。对于缺失值可以采用随机森林算法进行补齐。预处理的变量利用相关热力图来分析变量之间的相关性。同时，利用雷达图、箱线图来比较不同品牌的满意度的差异。最后，为了更定量的比较品牌总体性能满意度差异，使用熵权法计算综合满意度，分品牌进行比较。图 1 是不同品牌满意度比较分析流程图。

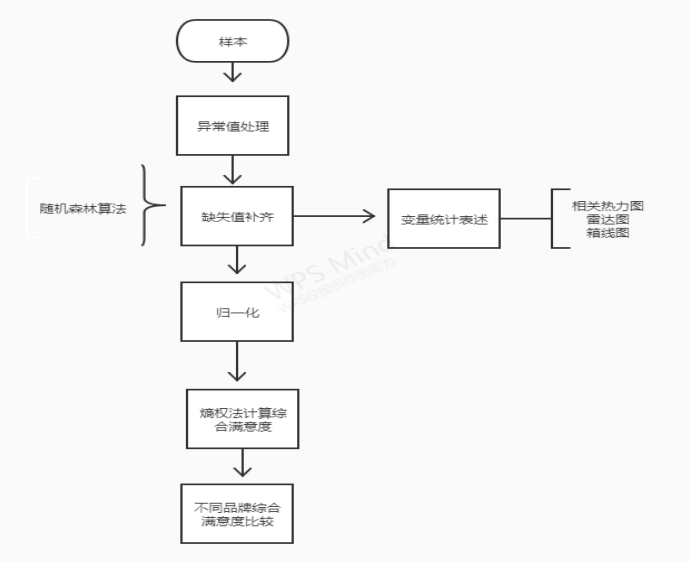


图 1

4.2 数据预处理

4.2.1 数据整体检测

采用箱线图法和标准差法对异常值进行检测，检测结果如图一所示。

箱线图法[1]：箱线图是数字数据通过其四分位数形成的图形化描述。这是一种非常简单但有效的可视化离群点的方法。其中，四分位距 IQR=Q3–Q1，下界 Lower Limit=Q1－1.5IQR，上界 Upper Limit=Q3+1.5IQR。把上下界作为数据分布的边界。任何高于上界或低于下界的数据点都可以认为是离群点或异常值。

标准差法：在统计学中，如果一个数据分布式近似正态分布，那么大约 68%的数据值在平均值的前后一个标准差范围内，大约 95%的数据值在平均值的前后两个标准差范围内，大约 99.7%的数据值在前后三个标准差的范围内。因此，出现在三个标准差范围外的数据点极有可能是异常值。

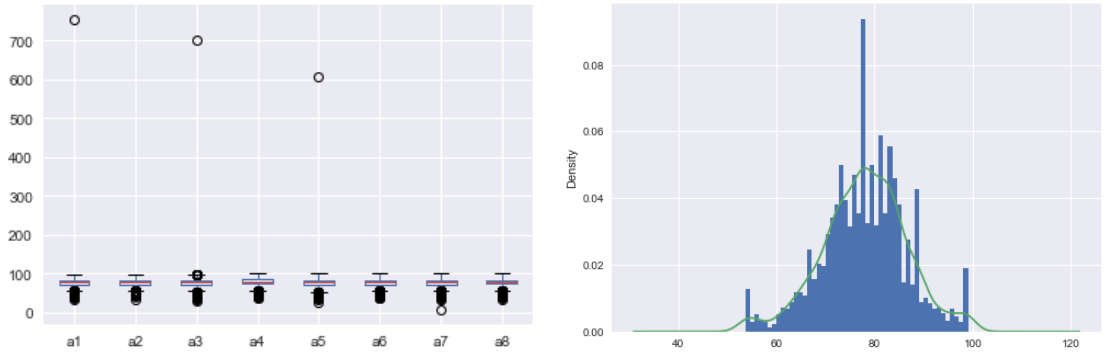


图 2

经检测发现原始数据有明显异常值，因此对异常数据进行处理。采用 99 分位数和 1 分位数对异常值进行替换，修正异常值。修正后的数据描述性统计如下所示。

表 1 数据描述统计表

变量	变量解释	变量个数	处理前 均值	处理前 方差	处理前 最小值	处理前 最大值	处理后 均值	处理后 方差	处理后 最小值	处理后 最大值
n	目标客户编号	1964	982.5	567.1	1	1964	982.5	567.1	1	1964
i	品牌类型	1964	1.786	0.553	1	3	1.786	0.553	1	3
a1	整体满意度 电池技术性能	1964	78.27	17.64	33.16	753.0	78.01	8.648	53.66	99.04
a2	整体满意度 舒适性	1964	78.13	9.058	35.77	99.03	78.18	8.893	51.86	99.03
a3	整体满意度 经济性	1964	76.22	17.59	29.59	703	75.95	10.33	44.45	99.03
a4	整体满意度 安全性表现	1964	78.84	9.145	37.48	99.98	78.90	8.928	52.28	99.98
a5	整体满意度 动力性表现	1964	77.43	15.23	25.23	605.6	77.24	9.258	49.39	99.98
a6	整体满意度 驾驶操控性	1964	77.88	9.370	39.15	99.99	77.94	9.174	52.48	99.99
a7	整体满意度 外观内饰	1964	78.02	9.235	7.880	99.99	78.11	8.892	53.20	99.99
a8	整体满意度 配置与质量品质	1964	77.58	9.588	33.32	99.98	77.64	9.393	51.44	99.98

4.2.2 变量缺失值补齐

由于子女个数的数据有部分缺失，故综合 b7 变量缺失的客户的其他特征信息对 b7 进行初步补齐。将 b7 之外的用户信息作为特征，b7 作为目标。将 b7 缺失的样本数据作为测试集，b7 没有缺失值的样本数据作为训练集，利用训练集对随机森林模型进行训练，利用训练好的随机森林模型对测试集进行预测，完成对 b7 进行补齐。

表 2 缺失值补齐表

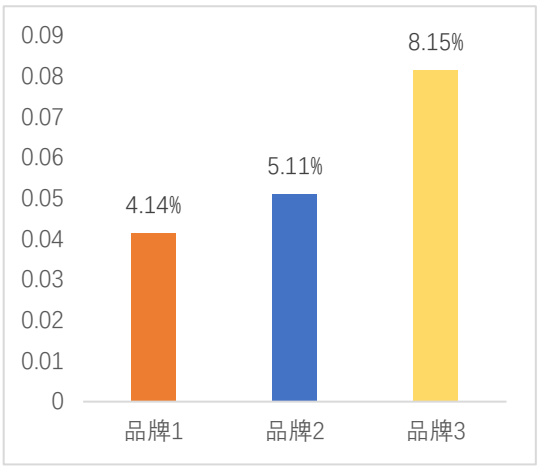
变量	变量解释	变量个数	均值	方差	最小值	最大值
b1	户口情况	1964	1.697	0.464	1	3
b2	城市居住	1964	21.38	11.44	1	60
b3	居住在以下哪个区域	1964	1.609	0.748	1	6
b4	驾龄	1964	7.636	4.128	1	30
b5	家庭成员个数	1964	3.456	1.081	1	6
b6	婚姻家庭情况	1964	4.531	1.417	1	8
b7（原始）	子女个数	1464	1.164	0.381	1	3
b7（补齐后）	子女个数	1964	0.871	0.604	0	3
b8	年龄	1964	34	5.005	61	21
b9	最高学历	1964	5.469	0.852	3	8
b10	工作年限	1964	10.08	4.907	1	38
b11	工作单位性质	1964	4.418	1.551	1	9
b12	职位	1964	4.878	2.526	1	11
b13	家庭年收入	1964	26.77	12.66	6	100
b14	个人年收入	1964	16.58	10.91	3	95
b15	家庭可支配年收入	1964	16.43	10.04	2	80
b16	房贷支出占比	1964	15.28	13.14	0	60
b17	车贷支出占比	1964	9.591	12.45	0	300
Willing	购买意愿	1964	0.0504	0.219	0	1

4.3 满意度分析

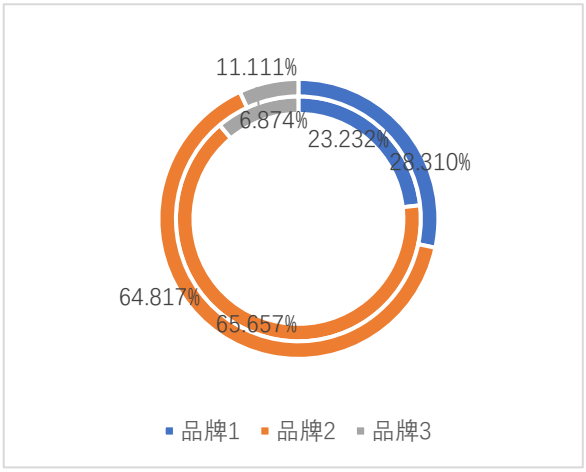
首先探索各满意度指标之间是否存在相关性，对数据标准化处理后做散点矩阵图。由附件图 1-1 可知各满意度指标之间存在显著相关性，如 a1 和 a2,a2 和 a5 之间均存在显著相关性，因此提升满意度不能只在单一指标发力，而应该从多方面着手。

然后分品牌进行比较分析。从图 3 可以看到品牌 3 体验用户的占比较少，但品牌 3

的意愿客户 11.11%的占比要大于其体验客户占比 6.87%，，品牌的客户购买体验比最高。



客户购买—体验比例分布



不同品牌客户目标比例分布图
（外圈：总客户 内圈：意愿客户）

图 3

图 4、图 5 展示了各品牌满意度指标得分情况和相关情况，可以看出各指标整体得分较为平稳，a3、a5 整体分数略低，表明电动汽车在经济性和动力性需要进一步提升。此外还可以看出 a3 和 a1 与其他指标相关性较低，因此电动汽车的电池技术性能和经济性是消费者会单独考虑的因素，不受其他因素影响。

以上是对原数据大致的分析，对于影响目标客户购买的关键因素还需要结合消费者的个人特征进一步分析判断。

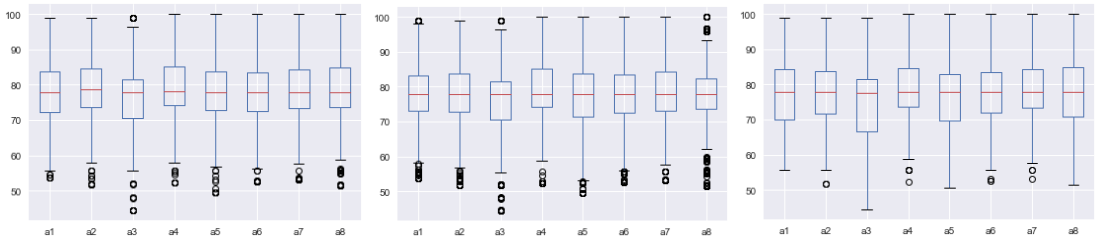


图 4

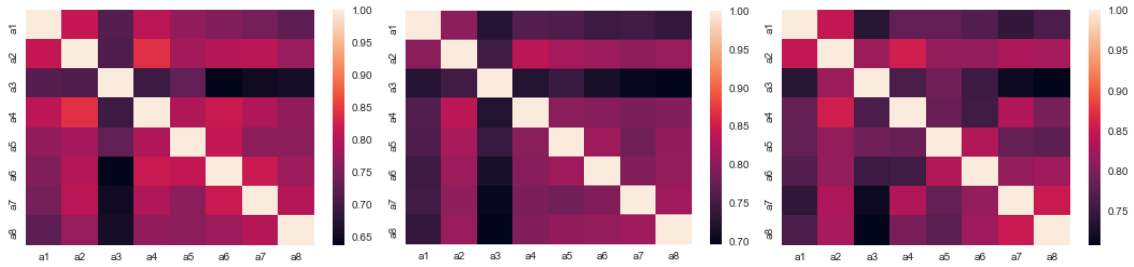


图 5

4.4 不同品牌汽车目标客户满意度比较分析

4.4.2 不同品牌汽车用户各方面满意度描述分析

分别对不同品牌电动汽车目标客户在电池技术性能、舒适性、经济性、安全性、动力性、驾驶操控性、外观内饰和配置与质量品质等八方面的体验满意度得分加权平均，利用 tableau 做对比雷达图，以分析不同品牌目标客户的满意度情况。

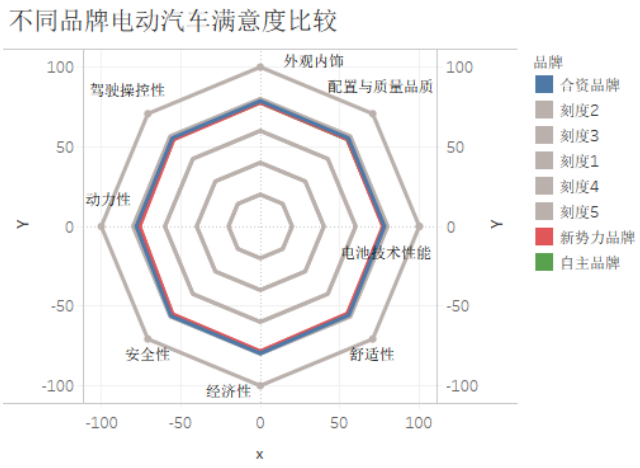


图 6

由上图可以看出不同品牌电动汽车客户满意度差别较小，且在电池技术性能、舒适性、经济性、安全性、动力性、驾驶操控性、外观内饰和配置与质量品质等八方面的体验满意度较为相近，基本集中在 80%左右。

4.4.1 不同品牌汽车综合满意度测度——熵权法

熵权法是综合评价中计算权重的一种方法。在信息论中，熵是对不确定性的一种度量。信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也越大。[2~3]

根据熵的特性，可以通过计算熵值来判断一个事件的随机性及无序程度，也可以用熵值来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的影响（权重）越大。比如样本数据在某指标下取值都相等，则该指标对总体评价的影响为 0，权值为 0。熵权法是一种客观赋权法，它仅依赖于数据本身的离散性。在本文的客户综合满意度测度中，目标客户对汽车的八种性能进行满意度打分，满意度分数差别比较大的汽车性能对综合评价影响大这是合理的。

（1）熵权法步骤

第一步：指标的归一化处理（异质指标同质化）：由于各项指标的计量单位并不统一，因此在用他们计算综合指标前，先要进行标准化处理，即把指标的绝对值转化为相对值，从而解决各项不同质指标值的同质化问题。

正向指标：

$$x'_{ij} = \frac{x_{ij} - \mathrm{m}\ \{x_{1j}, \dots, x_{nj}\}}{\mathrm{m}\ \{x_{1j}, \dots, x_{rj}\} - \mathrm{m}\ \{x_{1j}, \dots, x_{nj}\}} \tag{1}$$

负向指标：

$$x'_{ij} = \frac{\mathrm{m}\ \{x_{1j}, \dots, x_{nj}\} - x_{ij}}{\mathrm{m}\ \{x_{1j}, \dots, x_{rj}\} - \mathrm{m}\ \{x_{1j}, \dots, x_{nj}\}} \tag{2}$$

第二步：计算第 i 项指标下第 i 个样本值占该指标的比重

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

第三步：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 $k = 1/\ln(n) > 0$，满足 $e_j \geq 0$ 。

第四步：计算信息熵冗余度（差异）

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

第五步：计算各项指标的权重

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

第六步：计算各样本的综合得分

$$s_i = \sum_{j=1}^{m} w_j x_{ij},\ i = 1,\cdots,n \tag{7}$$

其中，x_{ij}为标准化后的数据。

（2）依据上述建立的熵权法模型，利用正向指标的标准化方法，本文利用体验客户的样本总数据来得到的不同方面满意度的权重，再利用权重计算用户对不同品牌电动汽车的综合满意度。满意度权重和编号前 10 的用户综合满意度情况如表 3，不同品牌的综合满意度情况如表 4。

表 3 编号前 10 客户满意度情况表

熵权法权重		a1	a2	a3	a4	a5	a6	a7	a8	熵权法综合满意度
		0.1086	0.1151	0.1666	0.1135	0.1283	0.1228	0.1147	0.1302	
目标客户编号	品牌类型	a1	a2	a3	a4	a5	a6	a7	a8	
1	2	71.68	72.41	74.05	69.92	80.84	75.23	73.2	77.34	74.4824
2	3	88.92	90.18	88.92	88.88	88.87	88.88	90.98	88.87	89.2789
3	3	67.26	67.1	66.69	70.34	69.73	69.74	73.52	66.66	68.7577
4	3	93.53	90.94	73.9	88.88	90.65	94.17	95.6	96.65	89.7856
5	3	89.65	93.95	73.9	96.3	88.87	96.92	95.6	95.88	90.5617
6	3	75.53	81.09	62.94	77.77	66.65	74.7	77.77	77.76	73.6317
7	3	81.3	90.59	81.56	78.66	71.67	69.74	73.26	85.2	79.0427
8	3	89.79	89.57	85.31	92.14	93.89	88.55	88.88	84.77	88.9006
9	3	86.35	86.89	88.92	88.88	88.87	88.88	86.6	84.77	87.5847
10	3	90.77	88.92	88.92	88.88	91.95	88.88	88.88	88.87	89.4893

从表 3 可以发现经济性整体满意度 a3 的权重最大为 0.1666，可能由于目标客户的家庭年收入等个人特征影响因素的方差比较大，导致不同客户对电动汽车的经济性满意度的评价差异比较大，权重赋予最大。而权重最低的是电池技术性能整体满意度 a1，权重为 0.1086。各满意度的权重比较接近。

表 4　基于熵权法不同品牌客户综合满意度统计描述

电动车品牌	变量	变量个数	均值	方差	最小值	最大值
1	熵权法满意度	556	77.61678	8.202259	50.72747	99.6126
2	熵权法满意度	1,273	77.52865	8.328923	50.72747	99.6126
3	熵权法满意度	135	78.83782	7.723352	56.49042	99.6126

表 4 可以说明品牌 3 的综合满意度平均水平比品牌 1 和品牌 2 要高为 78.8378 分，其综合满意度的趋势相对而言最集中，结合图 2 知道品牌 3 的客户体验—意愿购买占比

最高，故测度结果具有一定的合理性 。品牌 1 和品牌 2 的综合满意度比较接近，总体上品牌 1 的满意度更高为 77.617 分。

五、问题二模型建立和求解

5.1 问题分析

根据任务要求，需要找到那些因素不同品牌的电动汽车销售有影响。考虑到决定目标客户是否购买电动车的影响因素有很多，有电动汽车本身的因素，这主要体现在体验客户的满意度打分上；也有目标客户个人特征的因素。分析数据信息，客户对满意度的打分等变量为数值类数据，同时也存在部分分类数据如户口情况、家庭婚姻情况、地区分布等。为找到哪些因素影响客户的购买意愿，本文利用决策树算法来计算特征权重，将样本按照品牌分为三类，分别找到影响不同品牌因素的特征权重。

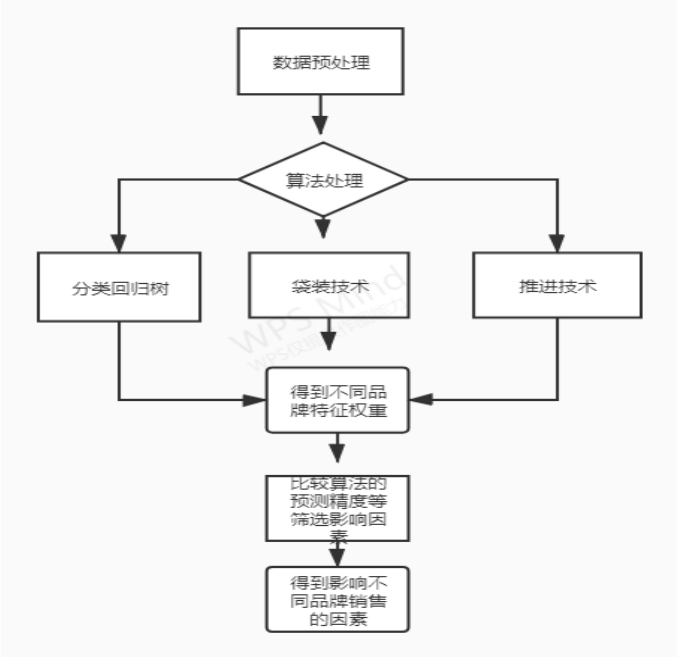


图 7 筛选营销品牌销售因素流程图

5.2 数据预处理

分类变量处理：对于分类变量，通过 stata 生成分类变量的虚拟变量，如户口情况 b1 的值 1、2、3 三类，为避免虚拟变量陷阱，将生成 2 个虚拟变量 b1_1、b1_2，并将原来的 b1 变量删除。

数值变量处理：对数值变量用最小—最大规范化方法去掉量纲。预处理后共生成 56 个变量（除去 n、i）。

表 5 分类变量说明

分类变量	变量说明	生成虚拟变量个数
b1	户口情况	2
b3	居住在以下哪个区域	5
b6	婚姻家庭情况	6
b9	最高学历	4
b11	工作单位性质	8
b12	职位	10

5.3 分类回归树和装袋技术

（1）分类回归树算法[4][5]：

Step 1：计算各输入变量的信息增益率，以信息增益率最大的变量为最佳分组变量。若分组变量为 k 类分类变量，则形成 k 个分枝。若为数值型，则用分箱法进行处理后再分枝。

Step 2：计算各节点误差，若子节点误差大于其父节点误差，则进行剪枝。

（2）袋装技术建立组合分类树[6][7][8]：分类回归树具有不稳定性，模型会随着训练样本的变化而剧烈变化。为提高稳健性，本文选择组合预测模型——袋装技术，其核心是重抽样自举法。具体步骤如下：

Step1:采用 Bootstraping 方法从训练集中随机 进行 k 次抽取,每次抽取的训练集的样本个数与本 文设置的原始训练集样本数相同。

Step2:使用 k 个弱分类器对 k 个训练集分别进 行训练,可以得到 k 个模型。此处的分类器可以是 个或多个分类算法。

Step3: 对 k 个模型采用投票方式即可得到分类 结果,在投票过程中,每个模型给与的权重相同。图 8 显示了 Bagging 算法的具体过程。

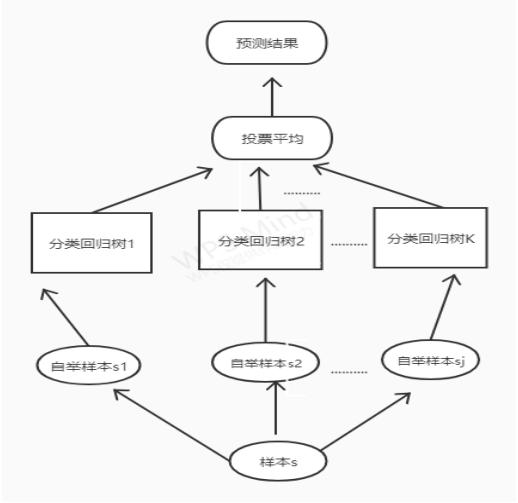


图 8 袋装技术示意图

5.4 推进技术分类问题原理

由于袋装技术的自举样本是完全随机的，多个模型在预测投票中的地位是相同的，未考虑不同模型的预测精度差异性。本文选择推进技术进行了调整。其包括两个阶段：第一，建模阶段；第二，预测阶段。

5.4.1 AdaBoost 分类建模

对于二分类问题，用 AdaBoost[9][10][12]实现的具体过程如下：

(1) 给定一个样本数据集 $(x_{i1}, x_{i2,\ldots,}x_{im}, y_i), i = 1,2,\ldots,n$。 其中 $X = (x_1, x_{2,\ldots,}x_m)$ 是样本特征向量，$y_i \in Y = \{-1, +1\}, Y$ 为样本类别集合。

(2) 样本权重初始化，设定每个样品相同的权重，即 $\omega_i^{(0)} = 1/n$ 。选择合适的基础分类器，在加权样本的基础上训练分类模型，预测样本的类别。根据分类结果更新权 重，分类错误的样本权重将被进一步提升。

(3) 样品权重的更新，记 t 为迭代轮数，$t = 1,2,\ldots,T$ 。 AdaB oost 算法使用参数 $\alpha(t)$ 来表示衡量弱的分类器 $f^{(i)}$ 在组合分类器中的权重,算法 提出的时候 $\alpha(t)$ 有下面的定义:

$$\alpha^{(t)} = \frac{1}{2}\ln\left(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}}\right) \tag{8}$$

$$\varepsilon^{(t)} \quad = Pr_{i\sim w_t}\left[f^{(t)}(x_i) \neq y_i\right] = \sum_{i:f^{(t)}(x_i)\neq y_i} w^t(i) \tag{9}$$

(4) 样品权重 $\omega_i^{(t)}$ 的更新:

$$w_i^{(t+1)}(i) = \frac{w_i^{(t)}\exp\left(-\alpha^{(t)}y_i f^{(t)}(x_i)\right)}{z^{(t)}} \tag{10}$$

$$z^{(t)} = \sum_i w_i^{(t)}\exp\left(-\alpha^{(t)}y_i f^{(t)}(x_i)\right) \tag{11}$$

此处 $z(t)$ 是正则化因子，使得 $\sum_i w_i^{(t+1)} = \mathbf{1}$ 。

(5) 输出最终的预测模型

$$F(x_i) = \text{sign}\left[\sum_{t=1}^{T}\alpha^{(t)}f^{(t)}(x_i)\right] \tag{12}$$

$F(x_i)$ 的值域为样本类别的子集，$F(x_i)$ 的值即为第 i 个样本最终的类别。由于分类器的权重 $\alpha^{(t)}$ 为 $(0,1)$ 上 $\varepsilon^{(t)}$ 的单调递减函数，且 $\varepsilon^{(t)} = 0.5$ 时 $\alpha^{(t)}$ 值为 0，所以弱分类器在最终的预测模型分类器中的权重与其分类误差成反比关系。符合预测效果越准的分类器越重要的常识。

对于二分类问题，依据预测类别分别计算权重的总和，权重总和最高的类别即为观测样本的最终预测类别。

5.5 模型实现

5.5.1 模型参数说明

本文分类回归树利用 python 的 sklearn 选取总样本的 0.2 作为训练集，80%作为测试集。通过 python 实现模型。装袋技术和推进技术使用 R 语言实现，都将节点的最小样本量设置为 20，指定进行交叉验证剪枝时的交叉折数设置为 10，按变量重要性排序，输出当前分组变量的前 4 个候选变量。将指定最小代价复杂度剪枝中的复杂度参数设置为 0.01。推进技术中的重复次数设置为 20，使用“Breiman”定义权重。具体的参数设置如下表 6。

表 6 组合预测技术算法 R 语言函数参数设置一览表

函数	袋装技术	函数	推进技术
rpart.control	minsplit=20 maxcompete=4 maxdepth=30 cp=0.01 xval=10	rpart.control	minsplit=20 maxcompete=4 maxdepth=30 cp=0.01 xval=10
bagging	mfinal=25	boosting	boos=TRUE mfinal=25 coeflearn="Breiman"

5.5.2 决策树不同品牌特征权重获得

通过决策树对样本进行预测，预测精度为 0.0504，总体的预测精度较高，但倾向于预测客户不购买。故在本文下一节，将对模型进行改进。下图是 CP 参数为 0.01 时分类回归树的决策树。

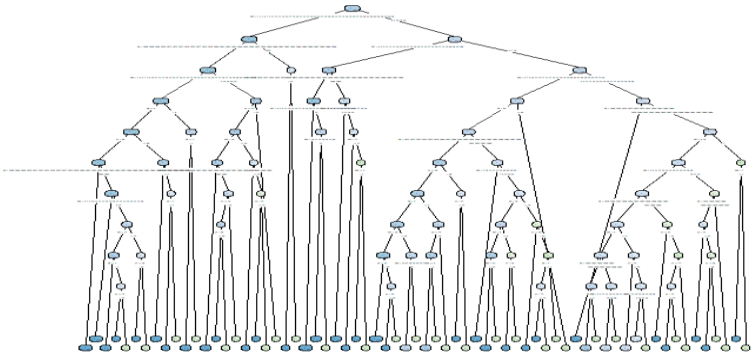


图 9 CP 参数为 0.01 时的决策树

通过附录表 2-1 和图 2-1 可以得到分类回归树得到的电动汽车各品牌的销售影响因素，发现影响总品牌的销售首要的使电池性能，其权重到达 0.1144，考虑到体验产品是电动汽车，电池性能是客户的一大关注点比较正常。同时，电动车本身的属性特征如舒适性 a2 和经济性 a3 和驾驶操控性 a6 等，也是影响电动汽车行业销售市场的重要因素。用户的个人特征如城市居住年份 B2 的权重达到 0.102 房贷支出占比 B16 权重达到 0.083，考虑到电动汽车并不适合跨省出行，在城市常住的客户更倾向于够不够买汽车。其次，其他一些个人特征，在城市家庭年收入 B13、家庭可支配年收入 B15、年龄 B8 也是比较重要的影响因素。 在所有影响因素中，动力性性能 a5、学历 b9 不是影响客户购买意愿的因素，并未出现在这分类回归树中。

在品牌 1 中，消费者首要考虑的是品牌的外观内饰 a7，其权重是 0.125。其次是舒适性 a2 和动力性表现 a5,客户特征比较重要的是工作单位是不是合资企业，职业是不是资深的职员。品牌 1 汽车产品可以在外观上别出心裁，并且可以对资深职员制定差异化营销策略。个人年收入和城市居住年份也是重要影响因素，这与总体市场的客户特征类似。舒适性 a2、安全性能表现 a4 对品牌 1 的销售影响作用不大。同时，客户驾龄 b4、子女个数 b7、年龄 b8、和房贷支出占比 B16 也不影响客户的购买意愿。

对于品牌 2，客户比较在意品牌产品的一些性能如外观内饰 a7、舒适性 a2 等。年龄 b8 的权重为 0.09，家庭年收入 b13 和房贷支出占比 b16 以及城市居住年数 b2 也是比较重要的影响因素。驾龄 b4 不影响客户对品牌 2 的购买意愿。可以看到 b12_2 出现在决策树中，考虑品牌 2 的消费群体主要是中层管理者。

对于品牌 3 电动汽车，由之前的描述可知其体验客户群体不多，但体验客户对其评价颇高。其经济性 a3 和舒适性 a2 是影响其购买的重要因素，是否是初级职员 b12_8 以及年龄 b8 是影响用户购买的用户特征。品牌 3 客户比较注重经济性，对产品除经济舒适性之外的其他因素要求不高，品牌 3 的主要目标客户可以面向刚入职场的年轻人。

5.5.3 组合预测模型获得权重——bagging 技术和 AdaBoost 技术

(1)预测精度比较。利用 R 语言建立 Baggig 技术组合分类树，对总数据的预测准确率为 0.9516，较之前的分类回归树的预测精度有一定的提高。使用 AdaBoost 方法建立预测模型，预测的精度为 1，预测的准确率相比 Bagging 技术大大提高。下图展示了两种预测技术预测编号前 10 的用户的预测结果。可以看到推进技术预测的概率精度更高，例如 13 号客户，Bagging 技术预测其完全不会有销售意愿，但 AdaBoost 技术预测其购买的概率高达 0.68，预测完全准确。故在下面主要分析 AdaBoost 获得的影响因素权重，将 Bagging 技术得到的权重仅仅作为参考。

n	i	willing	pre_willir	pro0	pro1
1	2	0	0	1	0
2	3	0	0	1	0
3	3	0	0	0.96	0.04
4	3	0	0	1	0
5	3	0	0	0.92	0.08
6	3	0	0	1	0
7	3	0	0	1	0
8	3	0	0	1	0
9	3	0	0	1	0
10	3	0	0	1	0
11	3	0	0	1	0
12	3	0	0	1	0
13	3	1	0	1	0
14	3	0	0	1	0
15	3	0	0	0.8	0.2
16	3	0	0	0.92	0.08
17	3	0	0	0.84	0.16
18	3	0	0	0.96	0.04
19	3	0	0	1	0

Bagging 技术预测

n	i	willing	willing_pr	pro0	pro1
1	2	0	0	0.9577281	0.0422719
2	3	0	0	0.9661563	0.0338437
3	3	0	0	0.6994998	0.3005002
4	3	0	0	0.8829853	0.1170147
5	3	0	0	0.7732261	0.2267739
6	3	0	0	0.9577281	0.0422719
7	3	0	0	0.8024416	0.1975584
8	3	0	0	0.8566955	0.1433045
9	3	0	0	0.8848512	0.1151488
10	3	0	0	0.8157734	0.1842266
11	3	0	0	0.9577281	0.0422719
12	3	0	0	0.9252113	0.0747887
13	3	1	1	0.321194	0.678806
14	3	0	0	0.8044591	0.1955409
15	3	0	0	0.7645116	0.2354884
16	3	0	0	0.8583134	0.1416866
17	3	0	0	0.7967578	0.2032422
18	3	0	0	0.8406685	0.1593315
19	3	0	0	0.9671982	0.0328018
20	3	0	0	0.7704223	0.2295777

AdaBoost 技术预测

图 10 用户编号前 20 预测结果对比图

（1）如附录表 2-2 影响所有品牌电动汽车销售的重要因素。推进技术预测的首要因素是家庭年收入，其次是房贷支出占比，Bagging 技术得到的权重排名前四的皆为用户特征。对于产品性能，体验客户比较关注外观内饰|、舒适性和动力性表现。户口情况、子女个数以及居住的区域对客户购买意愿几乎不造成影响，三者的权重都低于 0.015。

（2）如附录表 2-2 影响品牌 1 的主要因素性能因素有电池技术性能、配置与质量品质，电池性能因素的权重达到 0.102，明显高于其他因素。客户本身的意愿影响因素有家庭可支配年收入、车贷支出占比以及家庭年收入和城市居住年数，这些影响因素都是比较客观的影响因素。户口情况、最高学历、居住区域对用户购买不造成影响。

（3）如附录表 2-3 影响品牌 2 首要因素是房贷支出占比达到 0.8221，其次一些影响客户购买的个人因素有家庭年收入和个人年收入。性能方面，体验客户比较在意电池技术性能、舒适性以及配置与质量品质。户口情况、子女个数以及婚姻家庭情况对购买意愿影响不大。

（4）品牌 3 的体验客户最关注的电动汽车的电池技术性能，a1 的权重高达 0.1374，其次比较关注车动力性表现，对其他车的性能要求不是很高。除了常规收入影响因素外，驾龄也是影响品牌 3 客户购买意愿的重要因素。户口情况以及婚姻家庭情况对以及子女个数对客户的购买意愿影响不大。

5.4 影响因素剔除

结合上面分析，可以将不影响客户购买意愿的因素进行剔除，同时可以将分类变量生成更少的虚拟变量。户口情况 b1 以及婚姻家庭情况 b6 以及子女个数 b7、最高学历 b9、居住区域 b3 对客户的购买意愿影响不大，本文考虑剔除。关于职业 b12，注意到权重比较高的分别是 b12_3、b12_9、b12_8、b12_2,其分别代表“资深技术人员”、“中层管理者”、“个体户/小型公司业主”“初级职员”，本文考虑将职业只分为 5 类，即为只生成 4 个虚拟变量。

六．问题三模型建立和求解

为预测附件 3 中 15 名目标用户的购买情况，本文建立了不同品牌电动汽车的随机森林（Random Forenst）、逻辑回归（Logistics Regression）、多层感知机（MLP）、袋装法（Bagging）和推进法（AdaBoost）等五种模型对 15 名目标客户购买情况进行预测，

并对 5 种方法进行模型评价和预测准确度比较。

6.1 方法简介及评价

6.1.1 随机森林

1.模型介绍

随机森林算法[13]（RandomForest,RF）是一种组成式的有监督学习方法。它通过 Bagging 集成学习的思想组合多个决策树，最终结果通过投票法或取均值法取得，使模型整体的性能得以提升。随机森林中的决策树在分裂过程中先从所有的待选特征中随机选取一个包含多个特征的子集，然后根据特征划分准则从随机选取的特征中选择最优的特征划分当前节点，提升模型的分类能力。算法原理流程图如图 12 所示。

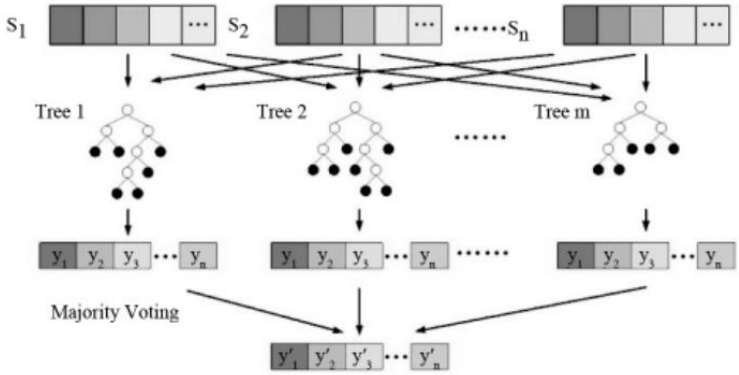


图 11

输入原始样本 S1，S2．．．，Sn 后，首先从中随机选取一部分构成新的训练集 X 并取代原始训练样本，以降低分类树之间的相关性，提高每棵分类树的精度，进而提高随机森林方法的分类精度。然后从 X 中随机选取一部分样本作为 Bootstrap(进化树) 训练集，构成与该训练集一一对应的回归树 $\{h(x, \Theta_k)，k=1，2，3，．．．\}$ (其中 $\{\Theta_k\}$ 是一组独立且具有相同分布的随机向量)，即分类树。每棵分类树 Tree Predictor(树预测器) 根据一组与输入样本有关的随机向量 $\{\Theta_k\}$ 进行分裂生长，最终众多分类树构成一个随机森林。每棵分类树根据训练集 X 进行分类，获得各自的分类结果，用多数投票法将所有分类树的分类结果进行综合，从而得到最终结果。

2. 模型求解与参数调优

（1）试验参数设置

为增强实验的可靠性和实用性，对随机森林参数：决策树个数 n_estimators、构建决策树最优模型时考虑的最大特征 max_features、决策树最大深度 max_depth 、叶子节点含有的最少样本 min_samples_leaf 、节点可分的最小样本数 min_samples_split 以及是否使用袋外样本评估模型好坏。参数设置如表 7。

表 7 实验参数范围

实验参数	取值范围
n_estimators	9,11, 13, 15
max_features	0.2,0.4, 0.5
max_depth	2, 3, 4, 5, 6, 7, 8
min_samples_	4, 8, 12, 16, 20, 24, 28

（2）网格搜索（Grid Search）寻找最优参数

用网格搜索（Grid Search）寻找最优参数。网格搜索法是指定参数值的一种穷举搜索方法，其核心原理是先设置好要搜索的参数区域，然后将该区域划分成网格，而网格中所有的交叉点就是要搜索的所有参数组合通过网格搜索法，得到模型训练数据如下表8。

表 8 最优实验参数表

实验参数	最优取值
n_estimators	9
max_features	0.4
max_depth	2
min_samples_	4

（3）模型求解与评估

对调整参数后的模型进行评估，评估得到准确率、召回率、F1 分数如表 9 所示。

表 9 分类结果评估分数表

Category	Precision	Recall	F1	Accuracy
Total	0.93	0.96	0.95	0.96
0	0.96	1.00	0.98	-
1	0.00	0.00	0.00	-

从表 9 可看出，模型对购买客户和未购买客户的预测准确性相差不大，在所有判定为未购买的客户中有 96%是真实未购买的，模型总体的精确度为 93%，召回率为 96%，F1 值为 95%，准确度为 96%。

此外，在训练过程中发现由品牌一训练得出的模型，在对测试集进行预测时，吸引了一名品牌三的目标客户进行购买，因此对品牌一提出建议，扩大目标客户范围，吸引更多客户购买品牌一的新能源汽车。

3．模型评价

随机森林增加了属性扰动的多样性，增加了多学习器的泛化能力，预测准确率较高；同时，通过随机性的引入增强了模型的抗噪声能力，不容易出现“过拟合”现象。但当随机森林中的决策树个数较多时，训练所需的空间和时间较大，运行速度较慢，不适于实时性要求很高的情况。

5.1.2 逻辑回归（Logistics Regression）

1.模型介绍

逻辑回归[11]是一种广义线性回归模型，常用于解决分类问题。不同于多元线性模型，逻辑回归模型的形式为：$p = L(W'X + b)$，L 表示逻辑回归函数，根据 p 的值确定因变量的值。自变量既可以是连续的，也可以是分类的。逻辑回归的因变量可以是二分类的也可以是多分类的。本文采用二分类逻辑回归模型，因变量为购买或不购买。

2．模型建立与求解

本文以目标客户作为研究对象，目标客户是否购买体验的电动汽车为因变量（购买为 1，不购买为 0），客户满意度 a1~a8 和用户特征 b1~b17 为自变量，并对于户口类型(b1)、居住区域(b3)、婚姻家庭情况(b6)、最高学历(b9)、工作单位性质(b11)和职位(b12)这些分类变量引入虚拟变量，建立 logistic 模型。其形式为

$$\ln\frac{P}{1-P} = c + \beta 1a1 + \ldots \beta 8a8 + \beta 9b1\_1 + \ldots + \beta 55b17$$

其中 P 指目标客户购买电动汽车的概率，c 为截距项。根据附件 1 中目标客户的相关信息，利用向前极大似然法进行参数估计，建立不同品牌电动汽车的二元 logistic 回归模型，据此对附件 3 中 15 名目标用户的购买结果进行预测。

（1）模型假定

①假设残差和因变量服从二项分布；

②各观测对象相互独立；

③自变量和 logistic 概率是线性关系。

（2）模型结果

表 10 各品牌销售重要影响因素表

	品牌 1			品牌 2			品牌 3		
变量	B(系数)	EXP(B)	显著性	B(系数)	EXP(B)	显著性	B(系数)	EXP(B)	显著性
b10	—	—	—	—	—	—	0.154	0.857	0.043
b13	—	—	—	0.029	1.029	0.003	—	—	—
b16	-0.111	0.895	0.002	-0.1	0.905	0	0.011	0.909	0.011
b17	-0.155	0.856	0.004	-0.086	0.918	0.001	—	—	—
b9_1	—	—	—	2.318	10.15	0.007	—	—	—
b9_3	—	—	—	—	—	—	1.742	5.707	0.02
b9_4	1.039	2.828	0.034	—	—	—	—	—	—
b11_4	—	—	—	-0.742	0.476	0.02	—	—	—
b11_5	-18.66	0	0.996	—	—	—	—	—	—
b12_3	1.415	4.115	0.032	—	—	—	—	—	—
b12_5	—	—	—	-18.18	0	0.995	—	—	—
常量	-2.161	0.115	0	-2.164	0.115	0	-0.376	0.687	0.663
内戈尔科 R 方		0.35			0.295			0.223	
样本数		556			1273			135	

由上表可以看出，不同品牌电动汽车各项满意度得分对最终是否购买汽车的影响较不显著，目标客户的个人特征对最终是否购买电动汽车有显著影响，且不同品牌电动汽车的影响因素有所差异，但全年房贷支出占家庭总收入的比例对三种品牌电动汽车的购买均有影响。

2. 模型评价

逻辑回归模型具有操作简单、训练速度较快、内存占用小、成本较低、适用二分类数据和可解释性较强等优点，但由于模型在何处停止纳入变量是通过似然比检验决定的，所以能够改善似然比检验但系数不显著的变量也会纳入模型中。此外，逻辑回归模型不能解决非线性问题，且准确率相对较低，因此应与其他方法相结合进行预测。

5.1.3 多层感知机（MLP）

1.原理

感知机由两层神经元组成，只有输出层神经元进行激活函数处理，学习能力有限，不能解决异或这种非线性可分问题。为解决这一问题，多层感知机（MLP，Multilayer Perceptron）也叫人工神经网络（ANN，Artificial Neural Network）应运而生。

2.模型步骤

多层感知机包括输入、输出层和二者中间的多个隐层。输入层神经元接收外界输入，隐层与输出层神经元对信号进行加工，最终结果由输出层神经元输出。最简单的 MLP 只含一个隐层，即三层的结构，如下图：

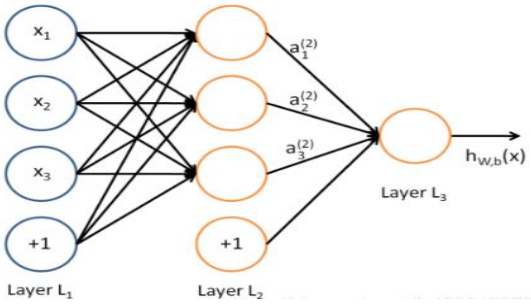


图 12

由上图可以看到，多层感知机层与层之间是全连接的。多层感知机最底层是输入层，中间是隐藏层，最后是输出层。假设向输入层输入向量 X，则隐藏层的输出为 f(W1X+b1)，W1 是权重（也叫连接系数），b1 是偏置，激励函数 f 一般为常用的 sigmoid 函数或者 tanh 函数，隐藏层根据阈值与激励函数进行权重的调整，直到达到允许迭代的最大数量或可接受的错误率，最终由输出层输出结果。

经预测，MLP 预测准确率如下图所示。

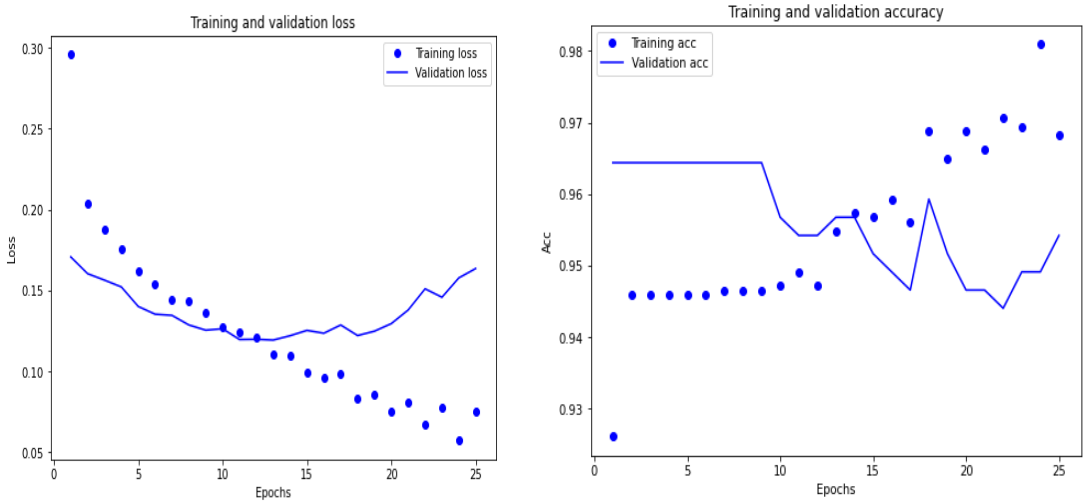


图 13　MLP 预测准确率图

3.模型评价

多层向量机模型可以很好的训练复杂的非线性数据，学习能力较强，但存在计算复杂度与网络复杂度成正比、容易出现过拟合问题以及模型解释力不强等缺陷，此外有隐藏层的多层向量机包含一个非凸性损失函数和阈值，不同的随机初始权重可能得到不同的验证精确度，而且多层向量机对特征缩放比较敏感。

6.1.4 推进法（AdaBoost）和袋装法（Bagging）

推进法与袋装法原理前文已经叙述，此处不再赘述。两者的区别在于 Boosting 方法的弱学习器之间有强的依赖关系，各个弱学习器之间串行生成；Bagging 方法的弱学习器之间没有依赖关系，各个学习器可以并行生成，最终通过某种策略进行集成得到强学习器，加快计算速度。将 Bagging 与 AdaBoost 预测结果进行对比，结果如下图所示。

n	i	willing	pre_willing	pro0	pro1
230	2	1	1	0.48	0.52
266	2	1	1	0.48	0.52
1553	1	1	1	0.44	0.56
1601	3	1	1	0.48	0.52

Bagging 技术

n	i	willing	pre_willing	pro0	pro1
230	2	1	1	0.277513702	0.722486298
266	2	1	1	0.2848076	0.7151924
1553	1	1	1	0.296242358	0.703757642
1601	3	1	1	0.320517678	0.679482322

Adaboos 技术

图 14 两种方法共同预测正确意愿客户概率比较

6.2 五种方法对比

6.2.1 预测精度比较

为避免单一模型存在缺陷，本文采用五种不同方法进行预测。由结果可以看出，多层感知机（MLP）和推进法（AdaBoost）预测精度较好。

表 11 各品牌在不同方法下的预测情况

品牌	方法		RF	MLP	AdaBoost	Bagging	LR
1	预测时间		2	1	1	1	0.1
	预测精度		0.960	0.974	1.000	0.960	0.959
	预测情况	0-1	0	1	0	0	0
		1-0	6	2	0	23	23
		1-1	0	0	23	1	0
		0-0	161	113	533	533	533
	样本数		167	116	556	556	556
2	预测时间		3	1	1	1	0.1
	预测精度		0.950	0.920	1.000	0.950	0.948
	预测情况	0-1	0	10	0	0	2
		1-0	19	10	0	65	64
		1-1	0	0	65	2	1
		0-0	363	230	1208	1207	1206
	样本数		382	250	1273	1273	1273
3	预测时间		2	1	1	1	0.1
	预测精度		0.930	0.926	1.000	0.926	0.910
	预测情况	0-1	0	0	0	10	1
		1-0	3	2	0	10	11
		1-1	0	0	11	1	0
		0-0	38	25	124	124	123
	样本数		41	27	135	135	135

6.2.2 购买预测情况

经过预测，结果显示 1 号和 12 号目标用户有可能购买新能源汽车。

表 12 目标用户购买预测情况

用户	RF	MLP	AdaBoost		Bagging		LR	
	pre	pre	pre	pro	pre	pro	pre	pro
1	0	0	1	0.41	0	0.80	0	0.75
2	0	0	0	0.61	0	0.92	0	0.99
3	0	0	0	0.96	0	1.00	0	1.00
4	0	0	0	0.97	0	1.00	0	1.00
5	0	0	0	0.73	0	0.96	0	1.00
6	0	0	0	0.77	0	1.00	0	0.84
7	0	0	0	0.69	0	0.88	0	1.00

8	0	0	0	0.97	0	1.00	0	1.00
9	0	0	0	0.96	0	1.00	0	1.00
10	0	0	0	0.87	0	1.00	0	1.00
11	0	0	0	0.77	0	0.92	0	0.95
12	0	1	1	0.38	0	0.84	0	0.81
13	0	0	0	0.77	0	1.00	0	0.98
14	0	0	0	0.68	0	0.96	0	0.92
15	0	0	0	0.72	0	1.00	0	0.92

观察这两位预测购买目标客户的满意度和个人特征。经过观察附录表 3-1 目标用户满意度特征，发现两位目标客户共同打分较高的几项为 a1（有电池技术性能）、 a3（经济性）、a7（外观内饰）、 a8（配置与质量品质）；共同打分较低的为 a4（安全性表现）、a5（动力性表现）；a2（舒适性）、a6（驾驶操控性表现）表现一般。

然后分析用户个人特征。由附录表 3-2 目标用户共同特征可以看出，家住城区，驾龄较短，家中人口多，工作稳定，职位处于中间阶层，收入处于中间水平且还没有车的客户倾向于购买新能源汽车。

七．问题四模型建立和求解

7.1 问题分析

用户的特征是企业不能改变的，企业只可以根据不同的消费者制定不同的策略。企业还可以提高产品的质量来提高 a1-a8 五个百分点的满意度，但服务难度与提高的满意度百分点是成正比的。为在提高满意度的服务难度最小的前提下，追求企业的利益最大化，即追求消费者购买意愿概率提升。本文进一步提出如下假设：

假设 1:提高 a1-a8 的服务难度是一样的,企业最多提升所有的满意度 5 个百分点，不考虑满意度指标调整之间的相互影响对调整难度的影响。

假设 2：企业每提升用户满意度 5 个百分点，服务难度会增加 25 个百分点，设开始的服务难度为 1。

假设 3：各个品牌市场客户的满意度指标权值为共同知识，但企业不清楚具体客户的满意度指标对购买意愿影响的权重。

假设 4：企业按照市场的满意度权值大小，依次对满意度指标进行调整，每次增加满意度指标 5 个百分点。

7.2 解题流程设计

7.2.1 客户筛选

本文根据前文 AdaBoost 技术预测得到的 15 名体验客户，选择每个购买意愿为 0，但购买意愿最接近 0.5 的用户作为模拟调整客户。因此，品牌 1 选出的模拟调整客户是编号为 2 的客户 k2，其购买概率是 0.392；品牌 2 选出的客户是编号为 7 的客户 k2，品牌 3 中选出的客户是编号为 14 的客户 k3。被选客户的购买意愿预测概率分布情况如下表 18。

表 13 客户的购买意愿预测概率情况情况

客户编号	体验品牌	预测购买意愿	预测购买概率
1	1	0	0.588149807
2	1	0	0.392641609
3	1	0	0.042271877
4	1	0	0.033467821

5	1	0	0.269491533
6	2	0	0.232065326
7	2	0	0.305032854
8	2	0	0.033467821
9	2	0	0.038807033
10	2	0	0.127820212
11	3	1	0.228863425
12	3	0	0.621051228
13	3	0	0.23103099
14	3	0	0.321523398
15	3	0	0.277659422

7.2.2 客户满意度调整流程

本文首先导入所有品牌样本的数据作为训练集，利用 AdaBoost 技术对训练集进行学习和预测，得到训练好的 AdaBoost 预测模型。根据前文假设，企业会按照先验的满意度影响权重（问题二中 AdaBoost 技术得到的权重）选择调整满意度指标，如企业对客户 k3 的满意度指标调整依次是 a1、a2、a8、a3、a7、a5、a4。将客户 k1 的数据代入模型，得到初始概率p_0，再将权重排名第一的满意度指标提高五个百分点导入模型进行模拟得到概率p_1，若概率$p_1 > p_0$，继续提高下一排名满意度指标五个百分点；若$p_1 < p_0$，将之前调整的满意度调整为原来大小，再进一步调整下一满意度指标。以此循环，直到调整完所有的满意度指标，在对下一客户进行模拟。满意度调整的流程图如图 15 。

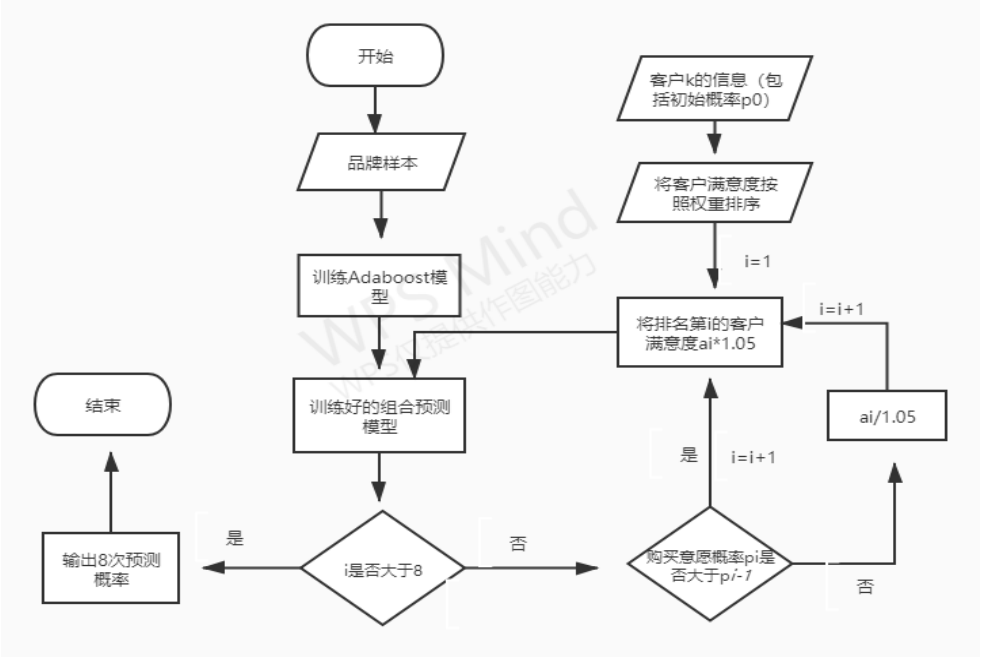


图 15　满意度调整试验模拟流程

7.3 模拟试验结果分析

由于假设提高所有满意度指标的百分点的服务难度一样，故可以用提高满意度得到的购买意愿概率P_{a_i}提高百分点来表示客户 k 满意度指标a_{ki}重要系数$W_{a_{ki}}$，其中 i 代表满意度指标调整的顺序。具体计算如下：

$$W_{a_{ki}} = \Delta P_{a_i} = P_{a_i} - Max\left(\left\{P_{a_{k1}}, P_{a_{k2}}, \dots\dots P_{a_{k,i-1}}\right\}\right) \quad i \le 8$$

服务难度百分点计算如下：

$$D_k = (1 + 0.25)^m$$

其中，m 是影响客户 k 购买概率的满意度指标个数。

7.3.1 客户 k1 模拟试验结果

表是客户 k1 的模拟结果，刚开始品牌 1 企业将电池技术性能 a1 提高 5 个百分点，客户 k1 的该买概率将上升 0.0422，服务难度提升 25 个百分点。进一步将 a2 即为舒适性提高 5 个百分点，此时客户的购买意愿几乎没有变化。故系统默认将 a2 的值退回原来的值，服务难度不提升。继续将经济性满意度指标 a3 提升 5 个百分点，可以从表 14 和图 16 中发现，购买意愿概率提升了 0.0694，概率明显提升，服务难度提升至了 0.3125（由（1+0.25）×0.25 获得）。可以发现客户 k1 比较注重电动汽车的经济性。继续提高 a4 五个满意度，概率下降，系统返回原来的值。提高动力性性能满意度指标 a5，概率提升 0.0368。进一步提高满意度相关性能指标 a7、a6，发现对概率提升作用不大，系统返回原来值。最终提升配置与质量品质性能 a8 五个百分点，客户购买意愿从 0.4484 提升到了 0.4912，提升 0.0428 购买意愿概率。最终目标客户有较大的概率会购买。综上，影响客户购买意愿的满意度产品性能因素有电池技术性能 a1、经济性 a3、动力性性能 a5、配置与质量品质性能 a8，其中经济性对客户 k1 购买意愿影响最大。

表 14 客户满意度指标调整购买意愿变化

	购买概率	提高 5%	原始	品牌 1 权重	重要性系数	服务难度
原始数据	0.3001	-	-	-	-	1
a1	0.3422	89.292	85.04	10.2191	0.0421	1.25
a2	0.3422	92.0325	87.65	4.9537	0.0000	-
a3	0.4116	85.638	81.56	2.5090	0.0694	1.5625
a4	0.3777	93.324	88.88	3.9212	-0.0339	-
a5	0.4484	89.9115	85.63	3.5890	0.0368	1.9531
a6	0.4156	90.447	86.14	2.0721	-0.0328	-
a7	0.4077	88.5255	84.31	4.2032	-0.0080	-
a8	0.4912	85.512	81.44	5.0770	0.0428	2.4414

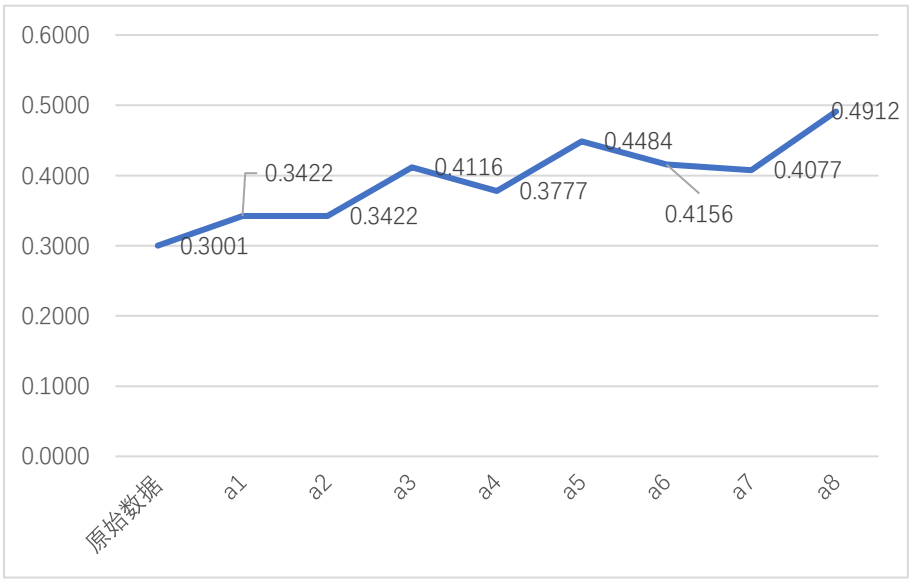


图 16　客户 k1 意愿购买概率变化

7.3.2 客户 k2 模拟试验结果

通过表 15 和图 17，发现对客户 k2 购买意愿概率提升有影响的指标只有经济性满意度指标 a3 和动力性性能指标 a5，两指标对客户 k3 的权重数值接近，客户最看重的是产品的经济性。总体的服务难度将从 1 提升至 1.5625。通过试验模拟可以发现，提高 a2、a8 等指标对客户的意愿概率提升并不大。通过模拟试验，企业提升性能更具有针对性和方向性。

表 15 客户满意度指标调整购买意愿变化

	购买概率	提高 5%	原始	品牌 2 权重	重要性系数	服务难度
原始数据	0.2128	-	-	-	-	1
a1	0.2128	89.292	87.07	7.9571	0.0000	-
a2	0.2128	92.0325	87.65	6.4650	0.0000	-
a8	0.2128	85.638	81.56	5.3637	0.0000	-
a3	0.2488	88.6578	85.19	4.8867	1.1692	1.25
a7	0.2488	89.9115	82.61	4.6813	0.0000	-
a6	0.2488	85.92465	83.6	3.8642	0.0000	-
a5	0.2898	84.099225	84.49	3.5869	1.1647	1.5625
a4	0.2898	85.512	84.77	3.1199	0.0000	-

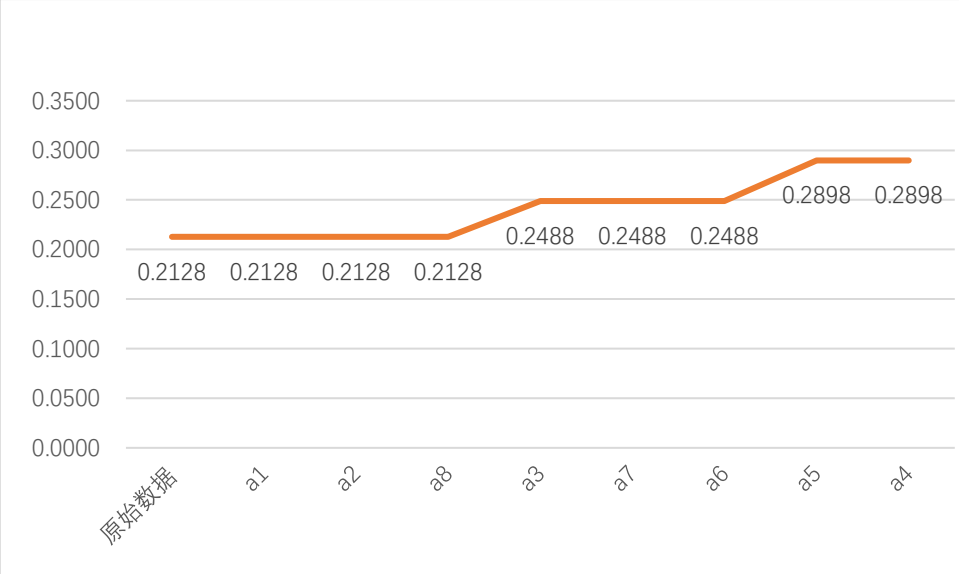


图 17 客户 k2 意愿购买概率变化

7.3.3 客户 k3 模拟试验结果

从表 16 和图 18，对客户 k3 购买意愿影响的产品性能指标只有 a8，即为产品的配置与质量品质。客户对产品的外观内饰属性并不决定他的购买意愿。企业的总体服务难度需要提升至 1.25，才能使 k3 的购买意愿概率上升至 0.3476。

表 16 客户满意度指标调整购买意愿变化

	购买概率	提高 5%	原始	品牌 3 权重	重要性系数	服务难度
原始数据	0.3110	-	-	-	-	1
a1	0.3110	80.7345	76.89	13.7394	0.0000	-
a5	0.2771	77.8	77.8	5.0263	-0.0339	-
a3	0.3110	77.7525	74.05	3.4171	0.0000	-
a6	0.2782	77.77	77.77	3.2738	-0.0328	-
a7	0.2422	74.69	74.69	3.0755	-0.0688	-

a2	0.2788	81.6585	77.77	2.7380	-0.0322	-
a4	0.2681	77.77	77.77	2.4217	-0.0429	-
a8	0.3476	81.648	77.76	1.4717	0.0366	1.25

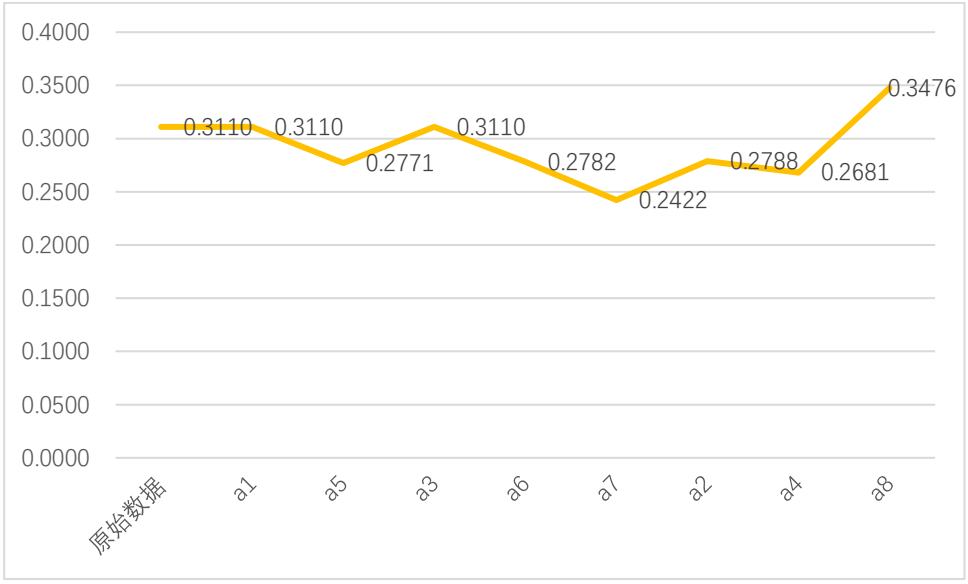


图 18　客户 k3 意愿购买概率变化

7.4 销售策略建议

从表 17 和表 18 中，对于客户 k1，其属于青年，职位为个体户（或小型公司业主），家庭年收入比较客观，每年的房贷占比为 10%，通过前文的试验模拟可以发现通过客户对品牌的经济性比较注重，企业可以考虑对品牌 1 实施价格差异化的营销策略。同时企业也要对品牌 1 号汽车的配置和质量品质、电池技术性能进行一定的性能该进，才更有可能留住 k1 客户。

从表 17 和表 18 中，对于客户 k2，其也属于青年，职位为初级技术人员，收入状况属于中低水平，同时背负有房债，经济压力较大。通过前文的试验模拟可以发现通过客户对品牌的经济性非常注重。在问题二中，品牌 2 主要面向的消费群体是收入比较可观的中层管理者，而品牌 3 主要面向中低收入人群。故企业最大程度调整品牌 2 的产品满意度，最终 k2 的购买意愿概率也并没有明显的上升。可以考虑建议客户 k2 去体验品牌 3 产品。

表 17　被选客户特征表

客户特征	变量解释	客户 k1	客户 k2	客户 k3
n	客户编号	2	7	14
i	品牌	1	2	3
a1	电池技术性能	85.04	87.07	76.89
a2	舒适性	87.65	87.65	77.8
a3	经济性	81.56	81.56	74.05
a4	安全性表现	88.88	85.19	77.77
a5	动力性表现	85.63	82.61	74.69
a6	驾驶操控性	86.14	83.6	77.77
a7	外观内饰	84.31	84.49	77.77

a8	配置与质量品质	81.44	84.77	77.76
B2	城市居住年份	13	32	48
B8	年龄	35	33	50
B12	职位	9	5	3
B13	家庭年收入	60	25	70
B14	个人年收入	55	15	50
B16	房贷支出占比	10	10	0
pro1	原始购买概率	0.3926	0.3050	0.3215

对于客户 k3，从表 18 得知客户原始的购买概率较高，客户 k3 对品牌 3 的产品满意度总体都属于较低的水平。由表 18 得知客户 k3 最关注的是配置和质量品质性能，而由前文可知，品牌 3 大部分群体最关注的是经济性。这可能因为客户 k3 收入较为可观，没有房贷压力，属于中年人年龄段，经济压力小。企业可以推荐客户 k3 去体验品牌 2 和品牌 1 的电动汽车。

表 18 客户满意度指标排名

客户 K1 满意度指标权重排名	客户 K2 满意度指标权重排名	客户 K3 满意度指标权重排名
经济性	经济性	配置与质量品质性能
配置与质量品质	动力性性能	-
电池技术性能	-	-
动力性表现	-	-

八．问题五建议

通过不同品牌电动汽车的销售情况进行统计分析，我们发现电动汽车的购买体验比相对较低。基于此我们对目标客户的相关信息进行深入分析并提出以下建议：

第一，对于合资品牌汽车，目标客户更关注外观内饰、经济性、动力性、配置与质量品质、驾驶操控性和电池技术性能；对于自主品牌汽车，目标客户更在意外观内饰、舒适性、电池技术性能和动力性；对于新势力品牌汽车，经济性和舒适性是客户最先考虑的因素。因此，针对不同品牌电动汽车，应从其主要影响因素方面入手，改善相关服务，提升用户满意度，增加最终成交量。

第二，合资品牌电动汽车的主要客户群体为在合资企业工作、本城市居住时间较长、个人年收入较高、车贷支出占家庭总收入比例较高、工作年限较长的资深职员；具备家庭收入较高、房贷支出占全年收入比例较高、年龄较大、工作年限较长和可支配收入较高特征的用户更倾向于购买自主品牌电动汽车；而驾龄较长、年龄较大、个人年收入较高、居住在市中心、车贷支出占家庭总收入比例较高的初级职员则更青睐于新势力品牌电动汽车，应针对性地挖掘不同品牌电动汽车的潜在客户，增加销售量。

参考文献

[1] 方安然,李旦,张建秋.异常值和未知观测噪声鲁棒的非线性滤波器[J].航空学报,2021,42(07):532-545.

[2] 郭伟,姚加林.基于 IAHP-熵权法和 Vague 集的高铁客运枢纽离站换乘评价[J/OL].工业工程与管理:1-12[2021-08-08].

[3] 薛薇.R 语言数据挖掘方法及应用[M].北京:电子工业出版社,2016:142-165.

[4] 吴东鹏,王峥,童薇,叶枫,宋楚翘.基于决策树-逻辑回归模型精确识别僵尸企业[J].应用科学学报,2021,39(04):569-580.

[5] 韩家琪,毛克彪,葛非凡,郭晶鹏,黎玲萍.分类回归树算法在土壤水分估算中的应用[J].遥感信息,2018,33(03):46-53.

[6] 周成骥.基于机器学习的商品购买行为预测模型设计[C].广州大学,2018.

[7] 王宵宇,谢然红,毛治国,张斌,刘若彤,邵亮,王堂宇.基于集成学习的烃源岩总有机碳含量测井评价方法研究[J/OL].地球物理学进展:1-13[2021-08-08].

[8] 周志华.机器学习[M].北京:清华大学出版社,2016:173-180.

[9] 常满想.AdaBoost 在基因表达数据分类中的应用[C].大连理工大学,2017.

[10]马鸣宇.基于机器学习的 P2P 网络借贷风险预测[C].华中科技大学,2018.

[11]李晨,张杨,陈长生.Logistic 回归应用的常见问题及其注意事项[J].中国儿童保健杂志,2020,28(03):358-360.

[12]Xu Yuan et al. A novel AdaBoost ensemble model based on the reconstruction of local tangent space alignment and its application to multiple faults recognition[J]. Journal of Process Control, 2021, 104 : 158-167.

[13]Dimitrios Gounaridis and Sotirios Koukoulas. Urban land cover thematic disaggregation, employing datasets from multiple sources and RandomForests modeling[J]. International Journal of Applied Earth Observations and Geoinformation, 2016, 51 : 1-10.

附录 1 正文图表补充

虚拟变量设定表

变量	定义		
b1:户口类型	b1_1:户口是否在老家	1=是	0=否
	b1_2:户口是否在本城市	1=是	0=否
b3:居住区域	b3_1:是否居住在市中心	1=是	0=否
	b3_2:是否居住在非市中心城区	1=是	0=否
	b3_3:是否居住在城乡结合部	1=是	0=否
	b3_4:是否居住在县城	1=是	0=否
	b3_5:是否居住在乡镇中心地带	1=是	0=否
b6:婚姻家庭情况	b6_1:是否为未婚，单独居住	1=是	0=否
	b6_2:是否为未婚，与父母同住	1=是	0=否
	b6_3:是否为已婚/同居无子女，单独居住	1=是	0=否
	b6_4:是否为已婚/同居无子女，与父母同住	1=是	0=否
	b6_5:是否为已婚，有小孩，单独居住	1=是	0=否
	b6_6:是否为已婚，有小孩，与父母同住	1=是	0=否
	b6_7:是否为离异/丧偶	1=是	0=否
b9:最高学历	b9_1:是否为初中及以下	1=是	0=否
	b9_2:是否为高中/中专/技校	1=是	0=否
	b9_3:是否为大专	1=是	0=否
	b9_4:是否为本科	1=是	0=否
b11:工作单位性质	b11_1:工作单位是否为机关单位/政府部门/基层组织	1=是	0=否
	b11_2:工作单位是否为事业单位	1=是	0=否
	b11_3:工作单位是否为国有企业	1=是	0=否
	b11_4:工作单位是否为私营/民营企业	1=是	0=否
	b11_5:工作单位是否为外资企业	1=是	0=否
	b11_6:工作单位是否为合资企业	1=是	0=否
	b11_7:工作单位是否为个体户/小型公司	1=是	0=否
	b11_8:是否为自由职业者	1=是	0=否
b12:职位	b12_1:是否为高层管理者/企业主/老板	1=是	0=否
	b12_2:是否为中层管理者	1=是	0=否
	b12_3:是否为资深技术人员/高级技术人员	1=是	0=否
	b12_4:是否为中级技术人员	1=是	0=否
	b12_5:是否为初级技术人员	1=是	0=否
	b12_6:是否为资深职员/办事员	1=是	0=否
	b12_7:是否为中级职员/办事员	1=是	0=否
	b12_8:是否为初级职员/办事员	1=是	0=否
	b12_9:是否为个体户/小型公司业主	1=是	0=否
	b12_10:是否为自由职业者	1=是	0=否

问题一图表

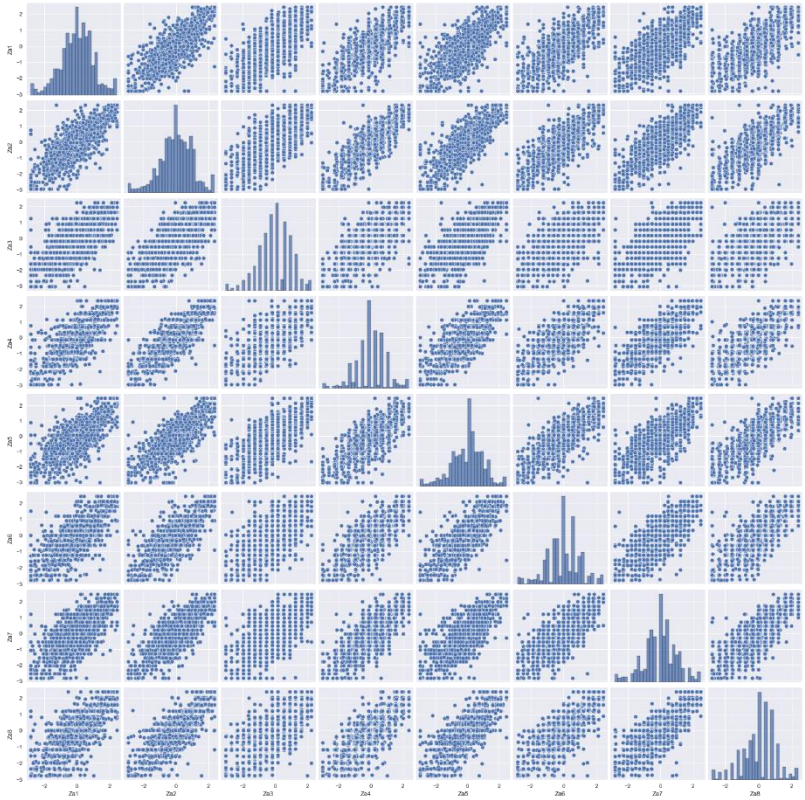


图 1-1

问题二图表

表 2-1 分类回归树排名前 10 影响因素权重

变量	总品牌重	变量	品牌 1 权重	变量	品牌 2 权重	变量	品牌 3 权重
a1	0.1144	a7	0.1254	B13	0.1051	a3	0.1840
B2	0.1022	a3	0.1121	B16	0.0973	B4	0.1818
B16	0.0835	a5	0.0945	a7	0.0954	a2	0.1028
a2	0.0717	b11_6	0.0823	B8	0.0917	B8	0.1006
B13	0.0675	a8	0.0816	B10	0.0661	B14	0.0967
B4	0.0533	B2	0.0734	a2	0.0555	b12_8	0.0909
a6	0.0475	B14	0.0605	B15	0.0513	b3_1	0.0606
a3	0.0423	a6	0.0581	B2	0.0437	B17	0.0574
B15	0.0419	B17	0.0547	a1	0.0351	b9_3	0.0553
B8	0.0391	b12_6	0.0526	b12_2	0.0337	B10	0.0418
b12_3	0.0336	B10	0.0502	a5	0.0330	b12_9	0.0281
B14	0.0323	a1	0.0487	B14	0.0330	a1	0.0000
a7	0.0317	b6_6	0.0476	b6_6	0.0299	a4	0.0000
B17	0.0303	B13	0.0314	b9_1	0.0296	a5	0.0000
B10	0.0281	b12_9	0.0173	b11_6	0.0282	a6	0.0000

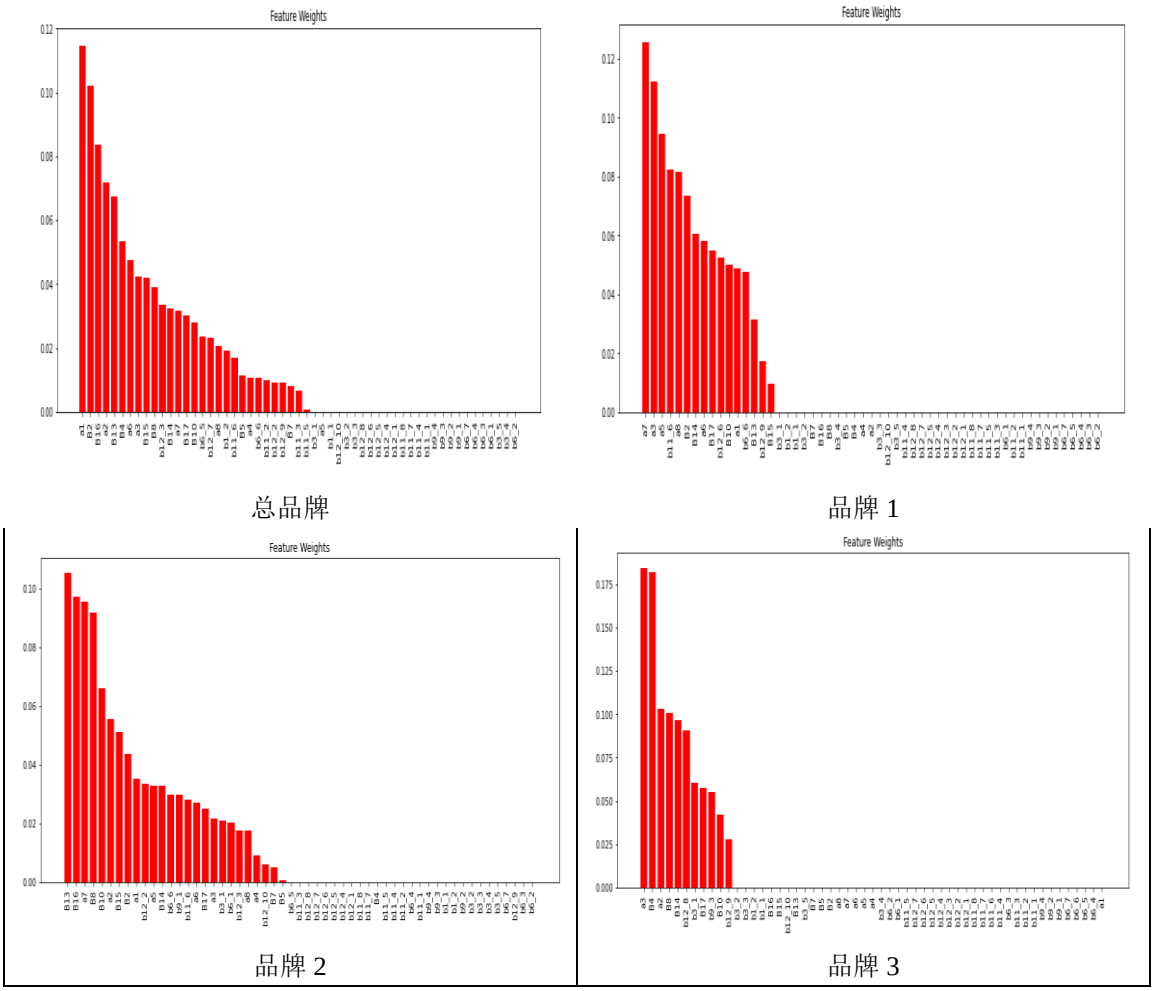


图 2-1

表 2-2　组合模型预测技术获得总品牌和品牌 1 影响因素权重

Bagging 技术二分类方法				AdaBoost 技术二分类方法			
变量	所有品牌	品牌 1	变量	变量	所有品牌	品牌 1	变量
B16	15.2687	B17	15.6285	B13	7.8188	a1	10.2191
B13	7.7847	a2	10.0712	B16	7.8103	B15	7.0515
B17	6.7844	B16	7.2782	a7	6.4587	B17	6.9301
B15	6.0471	B13	6.9482	a2	6.2551	B13	6.8110
a5	5.8525	B14	5.8000	a5	5.7872	B14	5.9266
a1	5.6617	a5	5.5842	a1	5.0318	B2	5.5898
a2	5.1856	B15	5.4751	a4	4.9546	B16	5.5767
a7	4.4714	a3	4.8274	B2	4.9488	a8	5.0770
a6	4.4234	a1	4.6170	B14	4.9365	B11	5.0185
a8	4.3112	b9_4	3.7969	B8	4.8987	a2	4.9537
B4	4.0635	B2	3.5191	a6	4.6430	B8	4.3251
B14	3.9544	B5	3.1010	a8	4.6269	a7	4.2032
a3	3.0028	b11_6	3.0283	B17	4.2222	B4	3.9391
B2	2.6955	a7	2.8721	a3	4.0478	a4	3.9212

表 2-3　组合模型预测技术获得总品牌 2 和品牌 3 影响因素权重

Bagging 技术二分类方法				AdaBoost 二分类方法			
品牌 2	变量	品牌 3	变量	品牌 2	变量	品牌 3	变量
B16	15.4760	a2	16.2893	B16	8.2211	a1	13.7394
a5	6.9067	B14	10.1956	a1	7.9571	B16	10.3274
B15	6.7335	a4	9.2930	a2	6.4650	B14	9.2358
a7	6.7029	a5	8.6191	B13	6.0186	B4	7.7972
B17	6.3737	a6	6.6144	a8	5.3637	B10	7.3795
a2	6.1630	a3	6.2704	B14	5.3371	B13	6.7479
B13	6.0444	B16	6.2289	B2	5.1869	B8	6.0050
B2	4.9431	B8	5.4575	B15	4.9644	a5	5.0263
a1	4.7137	b9_3	4.3571	a3	4.8867	B9	4.0032
a4	4.4697	b3_3	4.3182	B8	4.7270	B5	3.6442
B10	4.0182	B4	3.6760	a7	4.6813	a3	3.4171
b11_6	3.2687	b12_9	3.5175	a6	3.8642	a6	3.2738
a6	2.9069	a1	3.2710	B4	3.8585	a7	3.0755
B8	2.8866	B2	2.3468	B17	3.7805	B2	3.0307

问题三图表

表 3-1 目标用户满意度特征

用户	a1	a2	a3	a4	a5	a6	a7	a8
1 号	89.84	87.65	88.92	88.88	88.87	91.63	99.99	99.98
12 号	84.67	85.63	85.17	81.45	84.08	83.6	84.44	85.54

表 3-2 目标用户共同特征

个人特征	1 号	12 号	备注
B1	1	2	
B2	8	30	
B3	2	1	市中心和城区
B4	5	3	驾龄较短
B5	3	4	家庭人口 3 人及以上
B6	5	5	已婚有小孩
B7	1	2	
B8	1989	1990	年龄 30 岁左右
B9	6	6	本科学历
B10	9	7	有一定工作年限
B11	4	5	
B12	7	4	中间阶层
B13	36	37	家庭年收入 35 万左右
B14	20	23	个人年收入 20 万左右
B15	24	34	
B16	0	0	无车
B17	0	0	

附录 2　问题一熵权法计算综合满意度

```
#满意度熵权法
a=read.csv("data_buqi1.csv",header=TRUE);head(a)
library(forecast)
library(XLConnect)
b=a[,3:10];head(b)
#normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x)))}#0_1 标准化
#data1=as.data.frame(lapply(b,normalize))#lapply 是批量处理的高效函数
#head(data1)
data1=b
#求出所有样本对指标 Xj 的贡献总量
first1 <- function(data)
{
   x <- c(data)
   for(i in 1:length(data))
      x[i] = data[i]/sum(data[])
   return(x)
}
dataframe <- apply(data1,2,first1)
#将上步生成的矩阵每个元素变成每个元素与该 ln（元素）的积并计算信息熵。
first2 <- function(data)
{
   x <- c(data)
   for(i in 1:length(data)){
      if(data[i] == 0){
         x[i] = 0
      }else{
         x[i] = data[i] * log(data[i])
      }
   }
   return(x)
}
dataframe1 <- apply(dataframe,2,first2)

k <- 1/log(length(dataframe1[,1]))
d <- -k * colSums(dataframe1);d <- 1-d #计算冗余度
w <- d/sum(d);w
write.csv(w,"熵值法权重 1.csv")
sc=as.matrix(data1)%*%w;sc
write.csv(sc,"熵值法满意度 1.csv")
#
a$sc=sc
head(sc)
```

附录 3　问题四代码试验模拟

```
#对客户的满意度进行模拟，以客户实例 k3##########
k3_a1=data.frame()
##
i=1
#12837654
a.pred=predict.boosting(Boot,newdata = k3)
a.pred$prob
#15367248
k3[1,1]=1.05*k3[1,1];k3
k3[1,5]=k3[1,5]/1.05
k3[1,3]=1.05*k3[1,3]
k3[1,6]=k3[1,6]/1.05
k3[1,7]=k3[1,7]/1.05
k3[1,2]=1.05*k3[1,2]
k3[1,4]=k3[1,4]/1.05
k3[1,8]=1.05*k3[1,8]
k3
for ( i in range(1)) {
   j=8    #1-8
   a.pred=predict.boosting(Boot,newdata = k3)
   boos_test=a.pred$prob
   k3_a1[j,1]=boos_test[,2]
   k3_a1[j,2]=k3[1,8]
   i=i+1
}
k3_a1
write.csv(k3_a1,"k3.csv")
```

附录 4　问题二、三计算权重

```
#袋装技术计算权重和预测
BAg<-bagging(willing~.,data=a1,control=Ct1,mfinal=25)
BAg$importance#权重得到
a.pred<-predict.bagging(BAg,n1)
boos_test=a.pred$prob
boos_test0=data.frame(boos_test);boos_test0
boos_test0$class=a.pred$class;head(boos_test0)
n$pro0=boos_test0[,1]
n$pro1=boos_test0[,2]
n$class=boos_test0[,3]
write.csv(n,"bagging_test0.csv")

#推进技术
#数据预处理
n=read.csv("need_test1.csv",header=TRUE);head(n)#导入试验样本
n1<-n[,-1:-2]#去掉前两列
a=read.csv("data_buqi1.csv",header=TRUE);head(a)
test=as.data.frame(a);head(test)
a1<-a[,-1:-2]
a1$willing=as.factor(a1$willing)#将输入变量设置为
b1=subset(a,i==1);head(b1)
p1<-b1[,-1:-2]
p1$willing=as.factor(p1$willing)#将输入变量设置为
head(p1)
b2=subset(a,i==2);head(b2)
p2<-b2[,-1:-2]
p2$willing=as.factor(p2$willing)#将输入变量设置为
b3=subset(a,i==3);head(b3)
p3<-b3[,-1:-2]
p3$willing=as.factor(p3$willing)#将输入变量设置为
library("adabag")
#总数据
Boot<-
boosting(willing~.,data=a1,control=Ct1,boos=TRUE,mfinal=25,coeflearn="Breiman")
write.csv(Boot$importance,"boos_impor0.csv")
a.pred=predict.boosting(Boot,newdata = a1)
a.pred$class
boos_test=a.pred$prob
boos_test0=data.frame(boos_test);boos_test0
boos_test0$class=a.pred$class;head(boos_test0)
test$pro0=boos_test0[,1]
test$pro1=boos_test0[,2]
test$class=boos_test0[,3]
write.csv(test,"boos_data0.csv")
#拿总数据预测来测试待测试
k1=n1[2,];k1
a.pred=predict.boosting(Boot,newdata = k1)
#boos_test=
```

```
a.pred$prob
boos_test0=data.frame(boos_test);boos_test0
boos_test0$class=a.pred$class;head(boos_test0)
n$pro0=boos_test0[,1]
n$pro1=boos_test0[,2]
n$class=boos_test0[,3]
write.csv(n,"boos_test0.csv")
```

附录 5　问题三、问题二 MLP 二分类预测和分类回归树得到权重

```
#MLP 进行二分类
导入筛选和清洁后的数据
import numpy as np
import pandas as pd
from sklearn import preprocessing
%matplotlib inline
from matplotlib import pyplot as plt
data = pd.read_csv("data_xuni.csv",engine='python')

data1=data[data["i"]==1]
data2=data[data["i"]==2]
data3=data[data["i"]==3]
from keras.layers import Dense,LSTM,Dropout
from keras.models import Sequential
from keras import optimizers
from sklearn import metrics
#测试集训练集获得详见代码
model = Sequential()
model.add(Dense(55, input_dim = 55, activation = 'relu'))
model.add(Dropout(0.2))
model.add(Dense(128, activation = 'relu'))
model.add(Dropout(0.2))
model.add(Dense(64, activation = 'relu'))
model.add(Dense(1, activation = 'sigmoid'))
model.compile(optimizer = optimizers.Adam(lr = 0.001),loss = 'binary_crossentropy',metrics = ['accuracy'])
#决策树得到权重
history = model.fit(X_train, y_train,epochs = 25,  batch_size = 50, validation_data = (X_test, y_test))
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()  # 初始化
dtc.fit(X_train, y_train)  # 训练
# 获取特征权重值
weights = dtc.feature_importances_
print('>>>特征权重值\n', weights)

a=pd.DataFrame(weights)
b=pd.DataFrame(train.iloc[:, 2:57].columns)
#b
b["weights"]=a
#b.to_csv("pinpai_weight.csv")
```