所属类别	2021 年“华数杯”全国大学生数学建模竞赛	参赛编号
本科组		CM211553

基于 XGBoost 算法的电动汽车目标客户销售策略研究

摘　要

电动汽车市场具有新兴性，而目标客户的销售策略则是市场研究过程中一个重要的研究对象。基于此，针对提出的问题，给出了统计学与机器学习意义上的一系列方案，并进行了较为全面的挖掘与思考。

问题一要求填充缺失值处理异常值，这里异常值仅有 a1-a8 的问题中出现，采取小数点重新标定的方法更新；而缺失值为客户有几个孩子，这里先采用常数填充法补 0，再根据客户婚姻情况和年龄综合判断后完成。随即进行了客户对电动汽车各项指标的看法、客户自身生活状况的描述性统计分析并得到一些统计意义上的结论。

问题二要求筛选出影响客户购买行为的特征，连同问题一属于典型的**特征工程**问题，这里采用统计学方法和机器学习方法进行对比。对于统计学方法，使用配对样本 t 检验处理连续数据，卡方独立性检验处理离散数据；机器学习方法则采用 **XGBoost** 为每个特征分配重要性分数，最后筛选出客户对汽车的电池充电能力、经济性、动力性能、家庭年收入、个人年收入、可支配年收入、房贷占年收入比例和车贷占年收入比例八项指标，达到了很好的结果。

问题三对客户的挖掘，建立了两个层面的挖掘：一是对客户特征的挖掘构建的**因子分析-聚类模型**，根据问题二筛选的指标利用聚类算法对客户进行聚类，随后进行探索性因子分析找到背后的隐变量以便于降维和可视化，更直观地揭示客户群体的差异性；二是基于问题二结果构建**八属性 XGBoost 分类模型**，经过与其他算法的对比与测试得到 XGBoost 的最优结果，准确率可达 100%而 AUC 则可达到 97%，是较好的结果。

问题四的精准营销策略，我们根据问题三中 XGBoost 模型的树结构自底而上进行判断有哪些属性在我们干预的范畴之内，选择客户 5、8 和 15 为例，对数据进行分析，并归纳总结了生成精准营销策略的**多目标路径寻优算法**，能递归地寻找可以提升的属性以生成最优策略。

问题五则是对前面问题的归纳与总结。

模型结合了统计学与机器学习方法的优点，具有可靠性，并且结果具有良好的可解释性。从机器学习常用的衡量指标来看，模型的表现无疑是非常优秀的，能够高效率地进行客户购买行为的预测与背后特征原因探究，并生成最优策略，在实际商业应用中有一定价值。

关键词：特征工程, 因子分析-聚类模型, 八属性 XGBoost 分类,多目标路径寻优算法

一、 问题重述

1.1 问题背景

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。但是与传统汽车相比，消费者在一些领域还是存在着一些疑虑，其市场销售需要科学决策。基于此，对于给出的三款不同电动汽车及其客户的购买情况数据，包括客户对目标产品的感受和客户自身的一些特征，我们需要对目前的电动汽车销售情况进行解析与优化。

1.2 问题提出

我们需要解决的问题如下：

1.对数据进行清洗工作，给出异常值和缺失值的处理方法并对数据进行描述性统计分析。

2.特征筛选，分析哪些因素对客户购买行为影响最大。

3.结合问题一和问题二的结果进行客户挖掘，并预测附件 3 中目标客户的购买行为可能性。

4.销售部门认为，满意度是目标客户汽车体验的一种感觉，只要营销者加大服务力度，在短的时间内提高 a1-a8 五个百分点的满意度是有可能的，但服务难度与提高的满意度百分点是成正比的，即提高体验满意度 5%的服务难度是提高体验满意度 1%服务难度的 5 倍。基于这种思路和前面的研究成果，请你在附件 3 每个品牌中各挑选 1 名没有购买电动汽车的目标客户，实施销售策略。

5.根据前面的研究结论，请你给销售部门提出不超过 500 字的销售策略建议。

二、 问题分析

2.1 问题一的分析

问题一中我们认为 a1-a8 范围内只要数据在 0-100 之间都是正常值，而 B 系列问题中基本都是关于个人情况，经观察没有特征值。所以，异常值在 a1-a8 内，超过 100 的分数容易看出属于小数点标定错误导致，我们将其变为原有异常数值的 1/10；对于缺失值，我们发现缺失的是客户有没有子女，我们根据客户的年龄和婚配情况辅助判断是否有子女即可。随后就可以进行描述性统计分析。

2.2 问题二的分析

问题二的特征筛选可以使用统计学方法也可以用机器学习方法，这里采用先统计学假设检验初步筛选再机器学习的策略。对于离散性特征，需要进行卡方独立性检验；而对于连续性特征，则将客户按照是否购买作为特征分为两个样本，进行配对样本 t 检验。机器学习方法则采用 XGBoost 训练分类器，根据每个特征的重要性分数进行排序，观察使用排名前几的特征能够得到较好的效果。

2.3 问题三的分析

问题三较为开放，客户的挖掘可以是客户自身特征的挖掘，也可以是客户行为或属性的挖掘，还可以是客户产生购买行为的预测。这里通过聚类算法为客户进行画像，

再用因子分析归纳抽象问题二筛选的特征以分析各聚类簇的特征并可视化，最后根据筛选的属性建立部分属性的 XGBoost 分类器模型进行分类预测。

2.4 问题四的分析

问题四较为抽象，我们基于问题三中训练的部分属性 XGBoost 分类器，根据其基学习器生成树结构寻找哪些通路会使得叶子节点分数为正数，再根据通路中各节点的条件进行筛选和调整，观察当各项指标到达怎样的水平时会发生变化。根据评价过程，我们可以归纳营销策略的多目标路径寻优算法进行自动化测试。

2.5 问题五的分析

对于问题五，我们则根据前面四个问题中建立的模型进行优化后，根据实际情况来对销售部门提出最优的销售策略。

三、 模型假设

针对这些问题，我们的模型假设如下：

1.观察到购买和非购买人群比例悬殊较大，我们假设类别的失衡不会造成严重影响，或者说造成的影响在我们模型的误差范围之内。

2.假设 a1-a8 内超过 100 的数值是由于小数点标定错误导致的，且部分客户没有子女的情况。

3.实施精准营销策略时认为并非提升越多效果越好，只要提升量超过一定阈值使客户产生购买行为即为成功，且暂不考虑收益与营销成本的量化关系。

四、 符号说明

表 1　各变量符号说明及解释

符号	说明
$\overline{x}$	平均值
S^2	方差
n	数值个数
μ	数学期望
t	t 检验统计量
χ^2	卡方统计量
A	离散数据的实际列联表
T	卡方公式中指理想列联表，XGBoost 中指树深度
$p(x)$	x 的概率
$H(X)$	X 的信息熵
$H(X\|Y)$	按照 Y 划分的条件信息熵
$ID(X,Y)$	信息增益
$L(\Phi)$	XGBoost 的损失函数
$f_t(x)$	学习的基学习器

$\Omega(f_t(x))$	正则化项，与基学习器的深度与权值范数有关
λ	常数项
w	XGBoost 中的各项权值系数

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 模型的建立

在数据清洗工作中，首先是缺失数据的补充。我们发现，缺失数据只有 B7 一项。而对于买家的生子情况，其他非空值均为 1，2，3 等，却没有考虑到买家尚未生子的情况。所以，这里可以使用常数 0 进行填充，也较符合实际生活中此类客户没有孩子的情况。然后对于异常值的剔除，我们考虑到对车辆的评价（即 a1-a8）的评分只要在 0-100 分之间均可视为合理值（这代表客户个人对车辆的评价，不因为客户个人的评价过好或过差而认为其异常），一共有三个数据超过 100。我们认为这可能是小数点标定错误导致，将其重新输入为原有异常值的 1/10。而对于客户自身特征而言，可能出现异常的是有关时间的统计量。我们根据客户的出生年份计算客户的年龄，并认为，只要车龄或者居住时长不超过客户年龄就认为是合理的。事实上，对于客户自身特征，没有出现时间上的异常值；而对于客户年收入，不同客户的经济实力不同，所以只要是正实数我们一般不认为异常。

5.1.2 模型的求解

由上述方法可以对缺失值和异常值进行初步处理，下面我们结合处理后的数据来进行描述性统计分析。

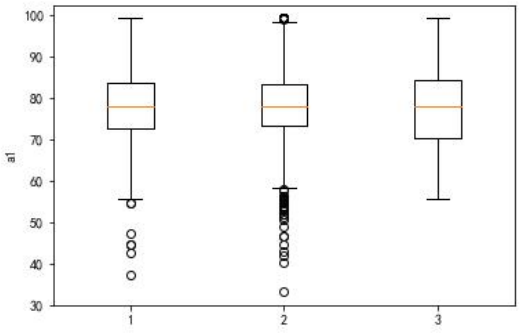
针对客户购车情况，合资品牌（以下简称类型 1）的客户有 28.3%，自主品牌（以下简称类型 2）有 64.8%，新势力品牌（以下简称类型 3）有 6.9%，有 5.04%的客户选择购买，这说明购买的客户群体非常少，而 2 型车吸引客户人群较多，可能是由于自主品牌电动汽车性价比较高或者品牌效应明显。新势力品牌由于产品尚未完善或者营销宣传力度不够，客户群体占比较少。

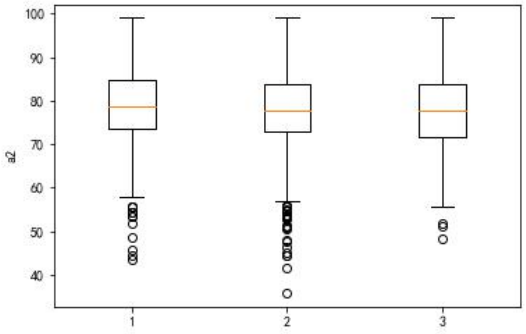
另外，针对三种不同品牌的汽车客户分别对其 a1-a8 的评价，我们做出了如表 2 所示的统计指标表格，并绘制了如图 1 所示的箱线图，以便更精确直观地展示统计特征，图表分别如下所示

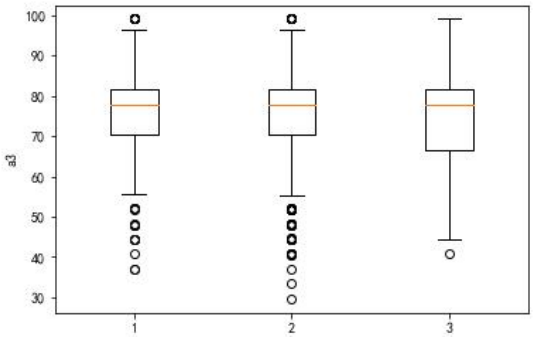
表 2　客户对三种不同品牌汽车评价结果的统计特性

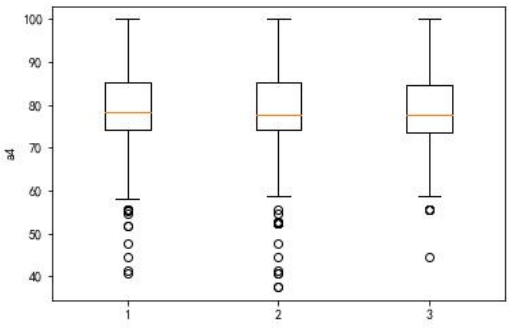
品牌	满意度	平均值	标准差	最大值	最小值
1	a1	77.99840926	9.377710318	37.04094194	99.0368881
	a2	78.69596965	9.099138763	43.40194575	99.0301718
	a3	76.03458468	10.49565222	37.0937021	99.0325502
	a4	79.4253961	9.275956614	40.69870252	99.98334246
	a5	77.90169693	9.511590177	38.11361401	99.97749794
	a6	78.25649529	9.511960265	41.56989073	99.99186723
	a7	78.65582135	9.455310839	39.86464126	99.9927132
	a8	78.29836889	10.06085818	40.32828249	99.98036497

2	a1	77.96769168	8.595643418	33.15990933	99.0368881
	a2	77.95365913	8.980535659	35.76983339	99.0301718
	a3	75.96752657	10.40539242	29.5879959	99.0325502
	a4	78.68020794	8.971889249	37.48352896	99.98334246
	a5	76.96131552	9.428455939	25.23083537	99.97749794
	a6	77.79064181	9.235644388	39.14919422	99.99186723
	a7	77.78002438	9.095785615	7.88434139	99.9927132
	a8	77.29162903	9.276785526	33.32345564	99.98036497
3	a1	77.202421	9.629213242	55.58049501	99.0368881
	a2	77.45731383	9.562869763	48.14953772	99.0301718
	a3	74.68118796	10.75313514	40.70161223	99.0325502
	a4	77.96038252	10.10074175	44.43259739	99.98334246
	a5	75.93347861	9.661407999	50.52738456	99.97749794
	a6	77.15873722	10.02608945	50.58954409	99.99186723
	a7	77.68131563	9.579442094	40.05872663	99.9927132
	a8	77.3497235	10.37559092	44.43127419	99.98036497









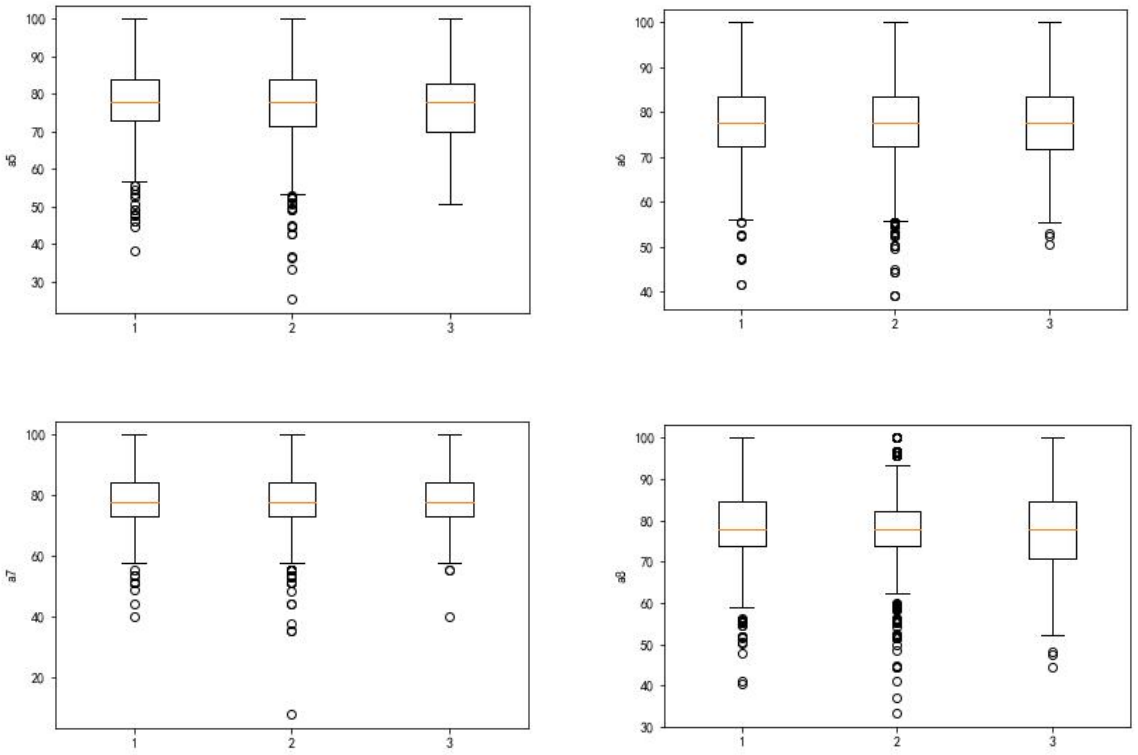


图 1　客户对三种不同品牌的汽车的评价

从表 2 和图 1 中可以很直观地看到，三者在各问题上的均值相近，说明整体上三者差异并不算太大，但在一些问题例如 a1、a3、a8 等品牌 3 的标准差较大，说明不同人群对品牌 3 的体验也有一定程度上的不同，具有主观差异性。而品牌 2 箱体最窄，但离群点也最多，且多为低于均值的离群点，说明买家对品牌 2 普遍不太看好。

此外，我们还制作了 a1-a8 八个指标之间的相关性图像，并绘制在图 2 的热力图中。图中颜色越浅则代表相关性越高，而相关性最低也有 0.7，说明各指标之间有着强烈的相关性。而 a3 与其它指标关联性最弱，a3 代表经济性。相关性相对要更强一些的是 a2 和 a4，代表舒适性与安全性之间关联还是比较大的。强关联的八项指标可以说构成了客户对车型评价的统一整体。

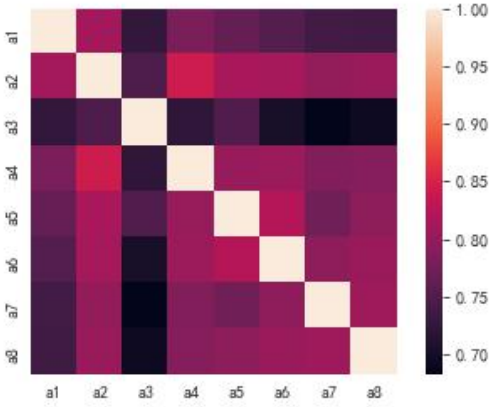


图 2　a1-a8 评价之间的相关性

然后是对客户自身的一些特性分析。我们将问卷中涉及到的问题按照连续和离散做了区分，将离散的数据绘制为如图 3 所示的扇形图，从左到右从上到下分别为买家户口、居住地、学历、单位性质、生活情况和职业属性。

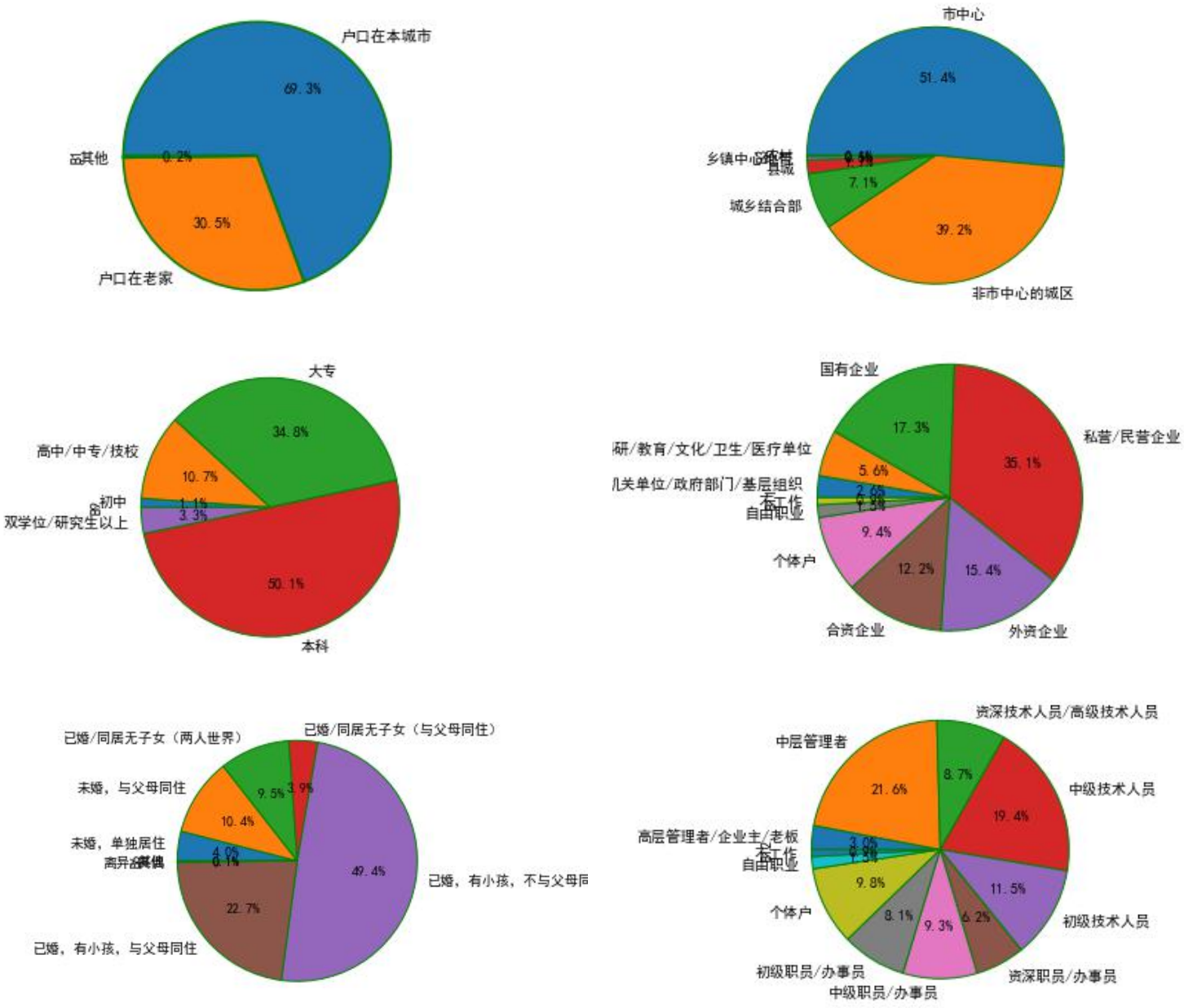


图 3　客户自身工作生活情况统计

可以看到买家大都是城市人口，多居住于市中心或非市中心的城区，二者加起来已经达到了 90.6%的比例，说明生活水平不算太差。从学历的角度来看，没有未接受正规教育或小学学历者，初中学历也仅有 1.1%。过半客户已经获得了本科学历，并且 3.3%客户甚至是双学位或研究生以上，学历水平较高。而买家们多就业于各类企业，包括国有企业（17.3%）、私营民营企业（35.1%）、外资企业（15.4%）、合资企业（12.2%），其中，大部分客户职业为各级技术人员或职员、办事员，超过 20%的客

户为中层管理者，少部分是企业高层。这说明客户的生活水平大都为小康水平，具备一定经济实力但很少有特别富有或特别困难的。

就生活状况来看，客户大都已婚，比例达到 85.5%。而 72.1%的客户是有孩子的，少部分客户没有孩子。问题 B7 孩子数量为 0 的比例基本与其一致，从侧面证实了我们的缺失值处理方法是合理的。

而对于一些连续属性而言，我们将客户的年龄、收入分配等绘制到如图 4 所示的频率分布直方图中，并对某些属性用正态分布曲线进行拟合，曲线图如下

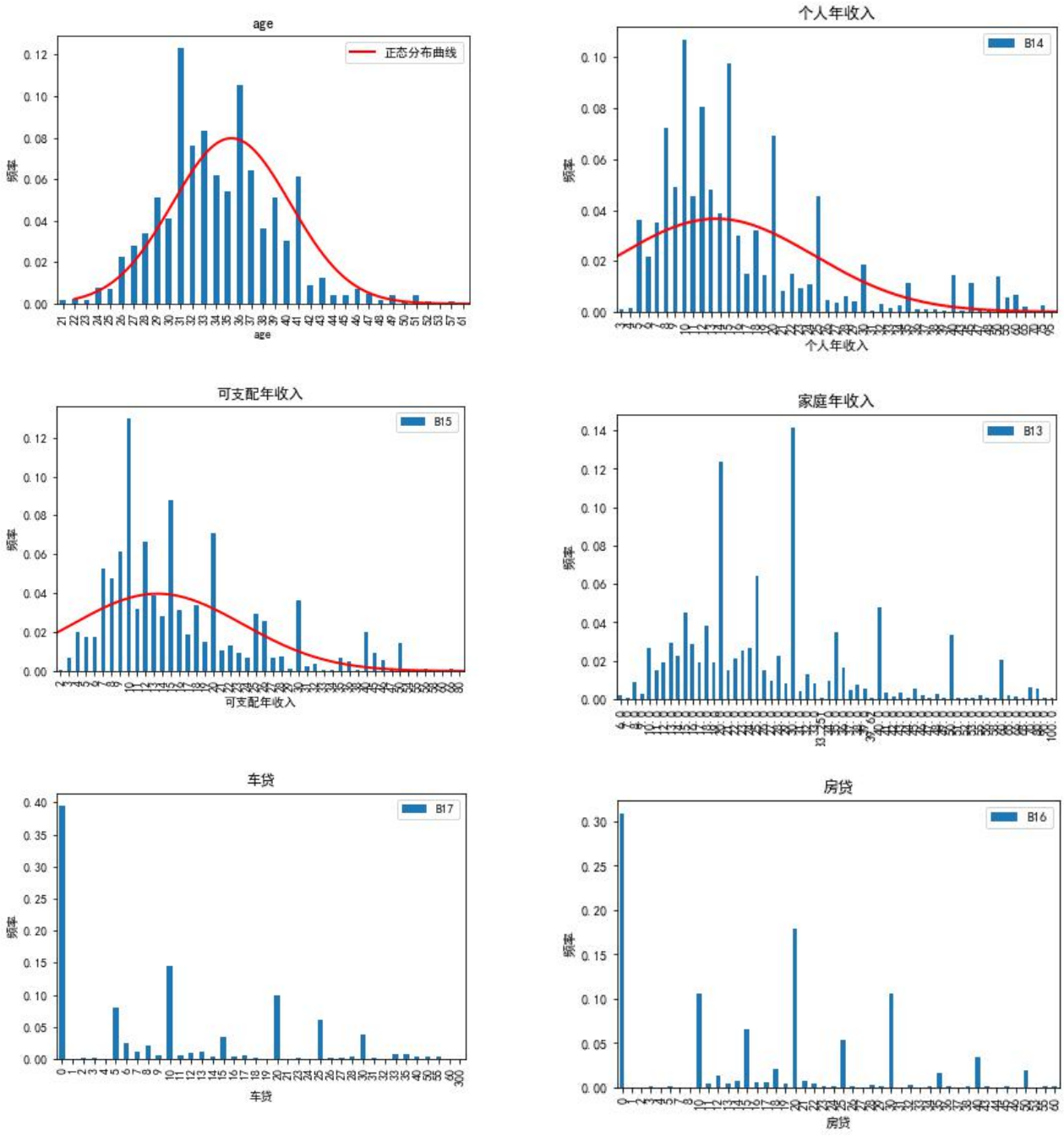


图 4　客户的年龄与收入分配

可以看到，客户的年龄集中在 30-40 岁之间，分布近似服从正态分布，说明目标群体以 30-40 的成年人为主。这些人的家庭收入在 20-30 万之间人数居多，而可支配

年收入大约在 10-16 万元之间，占年收入的将近 50%，弹性大，自由度高。大部分客户的房贷支出不超过 20%，车贷支出不超过 10%，相对来说经济条件略显充裕。

5.2 问题二模型的建立与求解

5.2.1 模型的建立

基于假设检验的统计学模型

问题二需要我们找出影响客户购买行为的影响因素，是一类典型的特征工程问题，这一问题可以建立统计学模型，采用假设检验的方法进行求解，也可以建立机器学习模型自动求解。

t 检验也用来判断样本均值和总体均值的显著性差异。很多地方 t 检验和 Z 检验类似，但是最大的区别在于总体的理论方差是未知的，t 分布只能用样本数据估计。独立样本 t 检验分析定类数据与定量数据之间的差异，配对样本 t 检验用来揭示定量数据的对比关系，样本先后的顺序要一一对应[1]。配对 t 检验的统计量定义为：

$$t=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \tag{1}$$

t 检验的原假设为统计量在两样本中无显著差异，备择假设为两样本中这一统计量有显著性差异。通常若概率小于 0.05，则接受备择假设，显著度 0.05。

卡方检验是一种用途很广泛的假设检验方法。根本思想就在于比较对应的理论频数和实际频数的拟合优度。而卡方独立性检验则常被用于检验两个离散统计量之间的独立性，原假设为两个量独立，若概率值小于 0.05 则接受备择假设认为两个量有一定程度的相关性，显著度 0.05[2]。统计量定义为：

$$\chi^2=\sum\frac{(A-T)^2}{T} \tag{2}$$

基于机器学习方法的特征工程

对于是否购买这一行为，可以将其视作一个二分类问题。对于分类问题的特征筛选，可以采用机器学习的策略进行。

特征工程分为数据预处理、特征抽取、特征构造和特征选择等几个步骤[3]，使用机器学习算法自动进行特征筛选的一类重要模型是树模型，这里选用 XGBoost 算法进行测试。XGBoost 算法是一类基于树结构的算法，而树结构为每个特征分配重要性分数的核心在于信息增益的计算。信息增益是指在按照某一条件划分以后的信息熵相较划分前信息熵的增量，能够衡量划分的效果，此外，信息增益率和基尼指数也可以衡量划分能力[4]。

$$IG(X,Y)=H(Y)-H(Y|X)=\sum_{x,y}p(x)\cdot p(y|x)\cdot\log p(y|x)-\sum_{y}p(y)\cdot\log p(y) \tag{3}$$

使用信息熵我们就能够计算出，根据哪一特征划分信息增益最大，然后按照信息增益从大到小排序即可得到最重要的特征。

5.2.2 模型的求解

根据式(1)、(2)，我们将数据集中连续属性用配对样本 t 检验，而离散属性用卡方独立性检验，并将品牌作为一个独立特征，得到统计检验结果分别如表 3 和表 4 所示：

表 3　配对样本 t 检验结果

满意度	平均值 1	平均值 2	检验统计量	概率值	显著度 0.05
a1	87.22872288	77.42985004	11.00384105	2.27E-27	是
a2	86.52311285	77.68413748	9.681982855	1.09E-21	是
a3	85.819737	75.37141921	9.927481645	1.07E-22	是
a4	86.64555328	78.42743528	8.885136976	1.41E-18	是
a5	85.60489742	76.70843588	9.29654399	3.74E-20	是
a6	85.3760682	77.4811245	8.309990915	1.76E-16	是
a7	85.55914104	77.62103504	8.482236407	4.28E-17	是
a8	84.92170592	77.19093868	7.940063397	3.37E-15	是
B2	24.57575758	21.21394102	2.853356466	0.004371374	是
B4	7.96969697	7.618766756	0.824196669	0.409928	否
B5	3.464646465	3.455227882	0.084470731	0.932690791	否
B7	0.888888889	0.866487936	0.35915445	0.719518197	否
B8	34.08080808	34.31581769	-0.455193771	0.649020207	否
B10	9.666666667	10.10348525	-0.863150273	0.388160347	否
B13	31.27272727	26.52915335	3.644532214	0.000274845	是
B14	19.17171717	16.44718499	2.42519336	0.015389957	是
B15	20.83838384	16.19302949	4.509567429	6.88E-06	是
B16	3.717171717	15.89383378	-9.174430345	1.12E-19	是
B17	2.292929293	9.978552279	-6.040665	1.83E-09	是

表 4　卡方独立性检验结果

满意度	检验统计量	概率值	0.05 显著度	0.01 显著度	0.001 显著度
B1	2.178295885	0.902592	否	否	否
B3	14.79505183	0.252836333	否	否	否
B6	5.807122075	0.99004826	否	否	否
B9	7.421281106	0.68516416	否	否	否
B11	18.59782964	0.416971242	否	否	否
B12	19.52358344	0.612813496	否	否	否
品牌差异	36.8402086	0.00719347086	是	是	否

可以看到，除却品牌差异以外，客户自身的离散特征与购买行为之间并没有太大关联，而连续属性中 a1-a8 都可以通过配对样本 t 检验，未购买的群体各项打分的平均值都是要低于购买人群的。另外，通过显著度为 0.05 的 t 检验的客户自身属性有 B2、B13、B14、B15、B16、B17，说明购买和未购买人群中客户的经济实力情况存在一定程度的差异。购买人群的经济状况相对更宽裕，而且房贷车贷压力更小。

而使用机器学习算法得到的排名前十的特征如图 5 所示

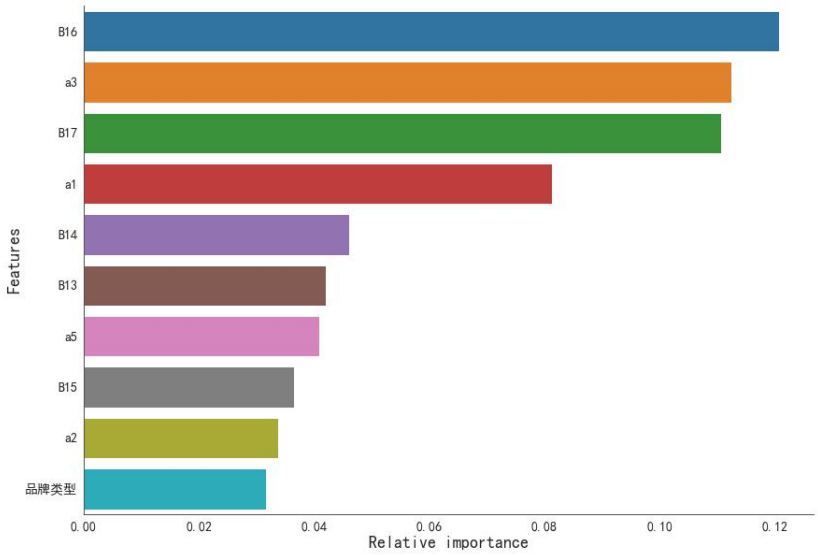


图 5　排名前十的特征

经检验，当选择到排名前八的特征时(B16,a3,B17,a1,B14,B13,a5,B15)模型的训练准确率和 F1 分数均可达到 100%不再上升，故而选择这八项特征参与后续训练。

B13 为家庭年收入，B14 为个人年收入，B15 为可支配年收入，B16 为房贷占家庭年收入的比例，B17 为车贷占家庭年收入的比例，反映了客户自身的经济情况；a1 为电池耐用性和充电能力，a3 为耗能和保值率，反映了汽车的性价比与经济性，a5 为动力性表现，也是性能的一个指标。可以看到，客户在购买行为时考虑更多的还是自身经济状况以及车辆本身的性价。

很显然，客户自身经济实力水平会极大程度上影响购买行为，因为如果客户自身经济状况并不太宽裕，或者可支配的资金不多，那么电动汽车有可能会超出客户承受范围。影响客户购车行为的自身条件首要的必然是经济条件。而对于车体而言，电动汽车最核心的两条性能就是电力与动力，而经济性与保值则可以代表客户购买电动汽车的风险，有充分的可解释性。

5.3 问题三模型的建立与求解

5.3.1 模型的建立

基于聚类方法和统计学分析的客户画像模型：因子分析-聚类模型

为了进行客户挖掘，需要对客户的特征进行区分聚类与画像描述。针对问题二中筛选出的八项指标，我们进行了聚类分析。这里我们选用 K-Means 算法和层次聚类两种方法进行对比。

K-Means 算法是基于距离的无监督聚类算法，其主要思想是:在给定 K 值和 K 个初始类簇中心点的情况下，把每个点(亦即数据记录)分到离其最近的类簇中心点所代表的类簇中，所有点分配完毕之后，根据一个类簇内的所有点重新计算该类簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数[3]。它实现起来比较简单，聚类效果也不错，因此应用很广泛。

层次聚类是基于层次的聚类算法，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点[4]。创建聚类树有自下而上合并和自上而下分裂两种方法，并且可以可视化展示，直观清晰。

silhouette 分数又名轮廓系数，通过计算样本 i 到同簇其他样本的平均距离 a，a 越小，说明样本 i 越应该被聚类到该簇。将 a 成为样本 i 的簇内不相似度；再通过计算样本 i 到其他某簇 C 的所有样本的平均距离 b，称为样本 i 与簇 c 的不相似度。定义式如(3)所示，当分数越接近 1 则聚类越合理[5]。

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{4}$$

采用 silhouette 分数作为评价指标，探索聚类簇数目从 2 到 9 变化时 silhouette 分数的变化趋势如图 6 所示

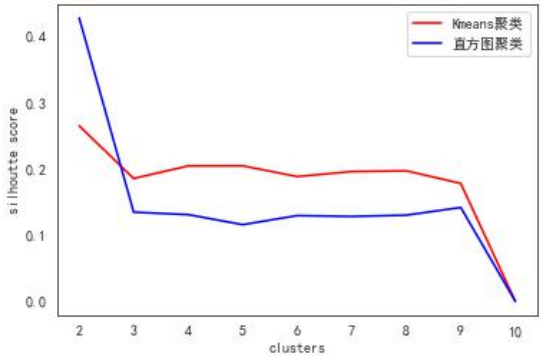


图 6　聚类数量簇与轮廓系数的关系

可以看到，无论是基于距离的 K-Means 聚类还是层次聚类，在聚类簇数量为 2 时效果最佳。并且，一个有趣的现象是，两种聚类方法得到的聚类结果基本一致。我们将层次聚类得到的层次树绘制如图 7 所示

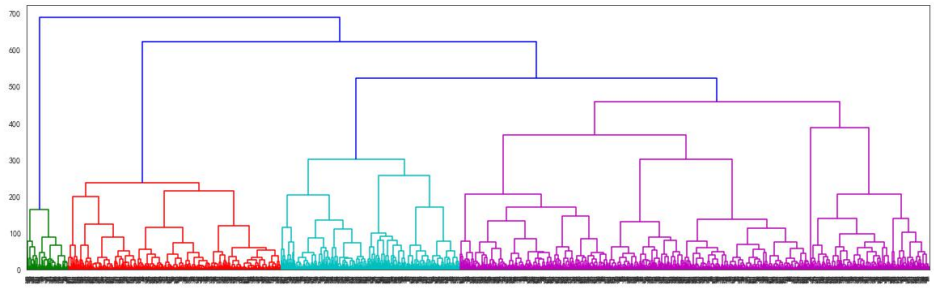


图 7　层次聚类的层次树

为了将用户的特征进行可视化，当数据维度过高时通常需要使用降维算法将其降到二维或三维欧几里得空间内。这里可以选用因子分析法探索变量背后的原因，构造隐变量达到降维的目的。

基于机器学习方法的客户购买行为预测模型：八属性 XGBoost 分类模型

这一模型则要求我们根据客户的自身特征进行预测。我们从问题二筛选出的八属性模型为根据，分别在不同的机器学习算法上进行测试。所对比的算法有：逻辑回归，

决策树，随机森林和 XGBoost。

决策树是一类基于树结构的监督学习算法，可以进行回归也可以用来分类。在决策树的算法中引入了信息论的方法，用熵来衡量非叶节点的信息量的大小，决策树中的非叶节点表示属性，叶子节点表示样本实例所属类别[6]。通过输入一组带有类别标记的数据，输出一棵二叉或多叉的树，优点是能够直观展示其结构。

随机森林(RF)是一种统计学习理论,它是利用 bootsrap 采样方法从原始样本中抽取多个样本, 对每个样本进行决策树建模, 然后综合多棵决策树的结果,通过投票得出最终预测结果，具有很高的预测准确率,对异常值和噪声具有很好的容忍度,且不容易出现过拟合[7]。

由 Chen 等人提出的 XGBoost 是对梯度提升树框架(Gradient Boost Decision Tree Framework)的一种具体实现，以 CART 决策树作为基学习器，既可以用于分类问题也可以用于回归问题[8]。XGBoost 的基本形式为

$$\begin{cases} L(\phi)=\sum_i l(\hat{y}_i,y_i)+\sum_k \Omega(f_k) \\ \Omega(f)=\gamma T+\frac{1}{2}\lambda\ ||\omega||^2 \end{cases} \tag{5}$$

在迭代过程中的损失函数表达式如下

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+\Omega(f_t) \\ &\simeq \sum_{i=1}^{n}\left[l\left(y_i,\hat{y}^{(t-1)}\right)+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \end{aligned} \tag{6}$$

不同于普通的 GBDT，XGBoost 在 L 后加入了与决策树深度 T 和分类器分数 w 有关的正则化项，使得精度本就出众的模型更具有鲁棒性。此外从某种意义上讲，它本身还具有降维的效果，并且对类别失衡的数据效果较好。

5.3.2 模型的求解

基于聚类方法和统计学分析的客户画像模型：因子分析-聚类模型

为了更好地可视化，并且从更高层级对八项指标进行抽象，我们使用因子分析的方法对指标进行进一步降维，找到能够描述这八项指标的隐藏变量。这八项指标的因子载荷矩阵绘制为如图 8 所示的热力图

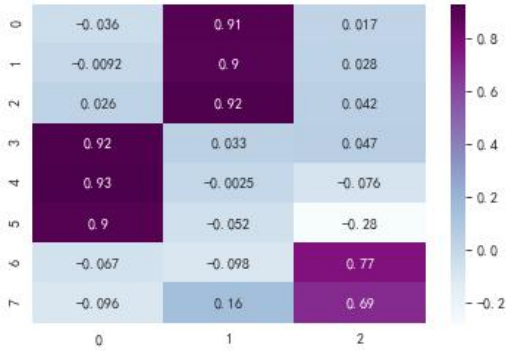


图 8　原始数据的因子载荷矩阵

经计算，当因子数量为 3 时可获得 0.95 的累计方差贡献率。图 8 中纵轴从 0 到 7 分别代表 a1,a3,a5,B13,B14,B15,B16,B17。因子 0 在 B13、B14、B15 的载荷最高，将这三者抽象为因子：“经济能力”；因子 1 在 a1、a2、a3 上载荷最高，这将被抽象为因子：“车辆自身属性”；而因子 2 在 B16、B17 上载荷最高，代表客户的“贷款压力”。

降维以后，我们将客户的聚类图绘制在如图 9 所示的三维坐标系内，并将其在因子 0 和因子 1 构成平面上的投影绘制在二维平面内（横轴为经济能力，纵轴为车辆自身属性）来构造我们的因子分析-聚类模型

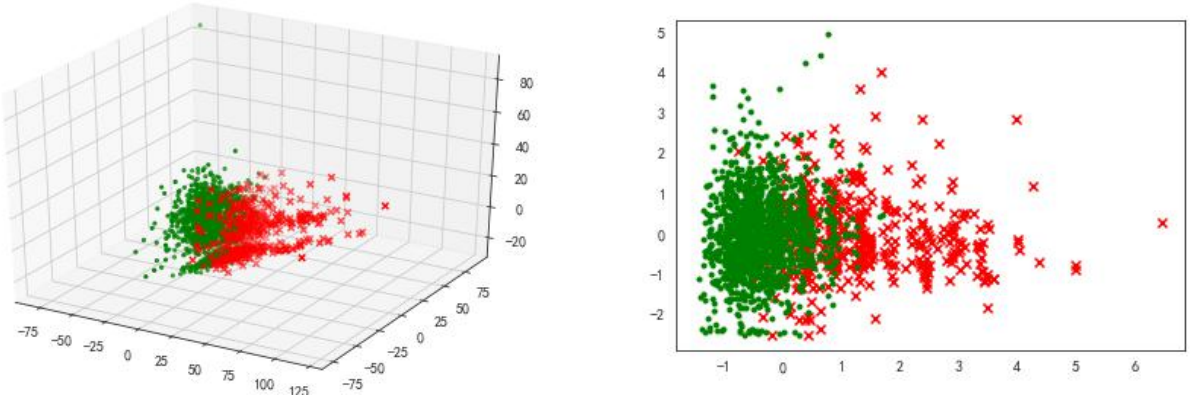


图 9　因子分析降维后客户聚类情况的可视化

可以看到，客户在这一几何空间内有一定程度上的区分度。我们尤其要考虑客户自身情况来进行精准营销策略控制。从两种客户密集区域的中心点与原点形成的向量方向来看，红色代表的客户群体，具有经济实力更强并且房贷车贷压力相对不算太强的特征，车的质量和性价比在其承受范围之内；而绿色的群体则认为汽车质量更高，自身经济实力更弱并且房贷车贷压力更大（z 坐标值普遍比红色区域高）。

这也描述了两类目标群体个体属性与对车体评价的关联性与差异性，尽管某些客户会认为目标商品是很好的，对其各项指标打分也比较高，但并不意味着他们会产生购买行为，还要考虑到客户自身的经济实力。而客户有经济实力的情况下，目标商品在他们的承受范围内，却不一定认为商品是他们的理想商品。那么尤其是针对客户最感兴趣的 a1、a3、a5 三个属性，从电池性能、经济保值与动力系统等维度进行精准营销。

基于机器学习方法的客户购买行为预测模型：八属性 XGBoost 分类模型

将数据集按照 7：3 的比例随机打乱后切分为训练集和测试集进行算法性能的对比。经测试，几种算法在训练集和测试集上的 AUC 曲线如图 10 所示。可以看到，XGBoost 无论是在 AUC 还是准确率分数上都表现得优于其它算法。对于这一算法，用 AUC、F1 分数、查准率、召回率和准确率来描述它。

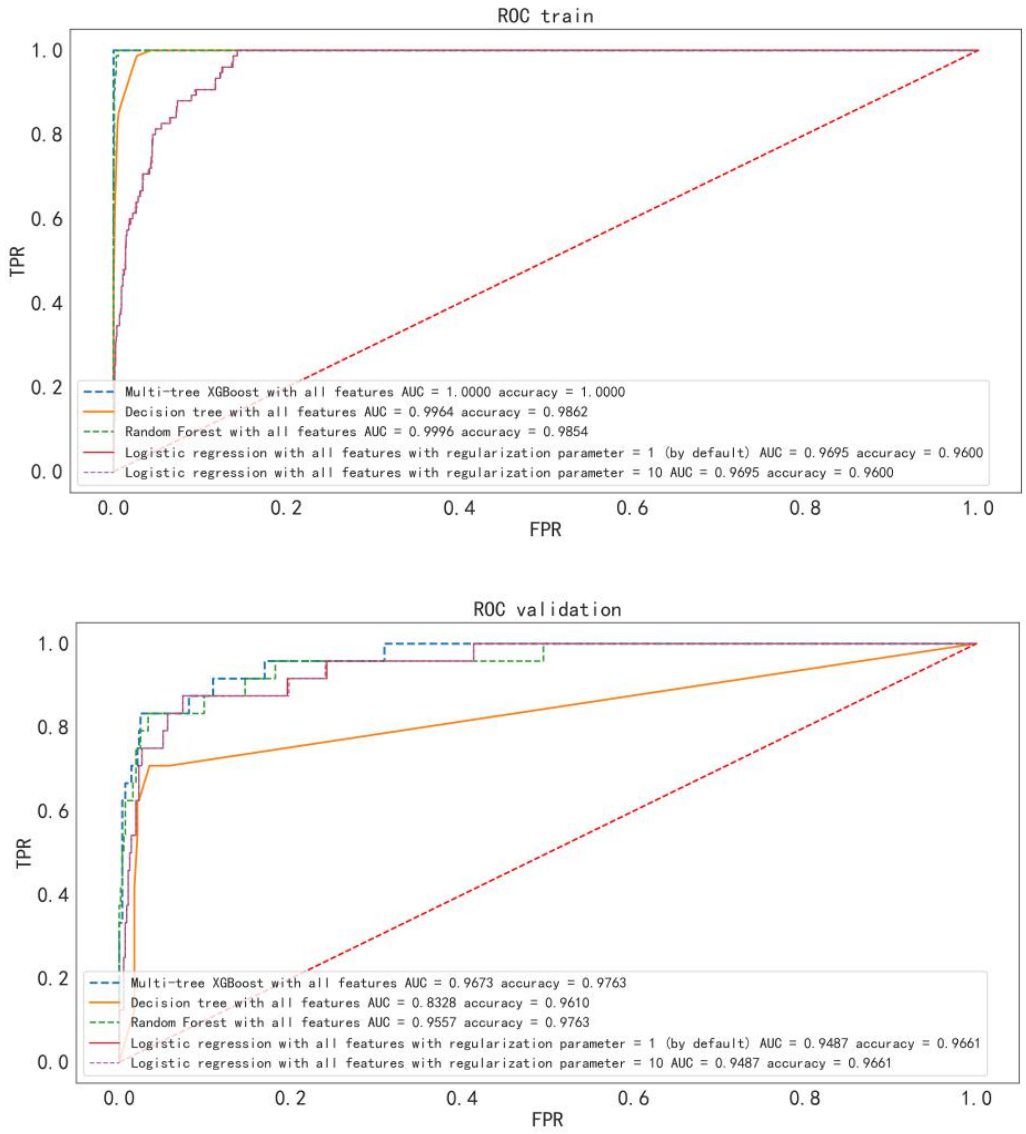


图 10　几种不同机器学习算法的测试性能对比

对于二分类问题，机器学习中有诸多指标评价分类器的分类效果，这里我们主要比较 AUC、F1 分数和准确率。若用混淆矩阵表示，准确率即分类结果与实际相同的样本在总样本中占比，是对样本分类整体精度的描述，其能够直接反映总样本中被正确分类的样本的占比；

精准率和召回率是对样本分类中细节精度的描述，而 F1 分数则是对分类问题中细节的权衡考量。F1 分数为查准率和查全率的调和平均数，在一定程度上可以反映模型预测的鲁棒性[9]；将预测为正类的样本中正例率和反例率绘制到图像中构成ROC曲线，ROC 曲线与横轴围成的面积即为 AUC 值[10]。AUC 值和 F1 分数、准确率一样，越接近 1 说明效果越佳。

八属性的 XGBoost 模型在随机划分的测试数据集上测试性能如表 5 所示

表 5　XGBoost 模型的测试效果

类别	precision	recall	F1-score	support
0	0.99	0.99	0.99	566
1	0.73	0.67	0.70	24
accuracy			0.98	590
Macro-avg	0.86	0.83	0.84	590
Micro-avg	0.98	0.98	0.98	590

可以看到，模型的准确率达到了 0.98，在测试集上的 AUC 也可以达到 0.9673，属于一个相当好的结果，说明模型在一定程度上确实可以达到想要的效果。

将这一模型迁移到待分类的 15 个客户上，预判结果为客户 1、2、6、7、11、12 六位客户会选择购买，而另外九位客户仍然不会购买。

5.4 模型四的建立与求解

5.4.1 模型的建立

对于问题四，我们可以通过销售策略提升 a1、a3 和 a5 的分数，至于另外五项指标，这属于客户自身的属性，销售者无法干预。那么也就是说，在前面建立的因子分析模型中，我们可以改变因子 1“车辆自身属性”来使客户获得更好的印象。

一个可行的想法是，通过观察分类器中树的结构判断当对应属性到达何种水平时能够对问题有一定提升。分类器的结构可视化如图 11 所示

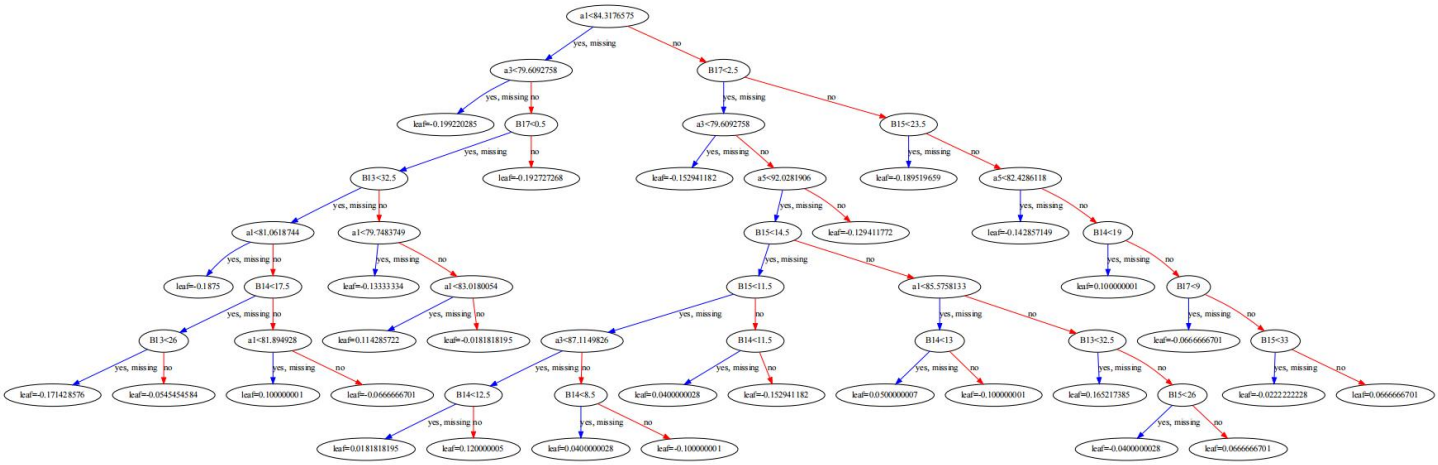


图 11　八属性 XGBoost 模型可视化

5.4.2 模型的求解

由图 11 可以看到，a1 处于最顶端，a3 和 a5 的层级基本相同。所以，对 a1 的改进力度要比 a3 和 a5 高。考虑到客户评分很高时会因为个人因素无法购买，而评分很低时提升难度很大，我们以客户 5、8 和 15 为例，对数据进行分析。

分析策略结合生成树采取自底而上的策略，根据叶子节点分数正负值寻找到根节点通路，以及通路中每个中间节点的条件，再结合所有可行路径判断客户原本按照哪

条路径进行分类，若改变某些量又应该经过哪条路径才能分类为购买。

对于客户 5 而言，客户经济实力较充足，但车贷压力较大。所以在营销时要注意经济效益方面的影响。但根据图 10 所绘制的决策树来看，客户自身的一些经济实力因素对购买行为可能有着更大的影响，营销策略只可能尽力使其改变想法。从根节点上看，客户 5 的 a1 值是超过 84.32 的，所以在电池技术性能上不需要做太大力度的营销宣传；同样的，对于第三层的 a3，客户 5 的 a3 值也超过了阈值。重点在于第 4 层的 a5，使其突破阈值只需提升不到三个百分点，也就是说在动力性能上进行更大力度的宣传。

对于客户 8 而言，观察根节点左半边的生成树，仅有 1 条通路使得分类结果为“购买”，但这一通路中 B17 节点无法干预，故考虑右半边生成树。若想要 8 号客户进入右半边生成树，首先 a1 起码提高 1.52%；随后，唯一的可行路径需要 a3 提升至 79.61%，也就是上升 1.8 个百分点。这意味着要在充电性能和经济性上加强宣传。

对于客户 15 而言，首先根节点左半边的生成树，已经没有通路，必须使 a1 突破阈值，也就是需要提升 1.73%。随后可以找到一条唯一通路符合条件无需进行其它更改，也就是说只需要针对 a1 的电池充电性能实施精准营销即可。

根据上面的分析方法，我们提出精准营销策略路径的生成算法，由于生成树中有多个叶子节点分数都为正数，所以这可以视作一个递归的多目标路径寻优。

5.4.3 算法的提出

我们提出的递归多目标路径寻优算法的 Python 伪代码如下：

算法 1. 多目标路径寻优算法

```
'''
Input: target_customer, XGBTree
'''
paths=[]
Cost={}
for leaf in XGBTree.leaves:
    if leaf.score > 0:
        path = connect_nodes_recursively(leaf)
        # path[-1]=leaf, path[i] is the father node of path[i+1]
        paths.append(path)
for path in paths:
    if f(B) in path:
        if f(target_customer.B[i]) == False:
            # if a customer's B is unsatisfied then remove this path
            paths.remove(path)
        elif f(a)-target_customer.a[i] > 5:
            # there f(a) means threshold, different from f(B) above
            # more than 5 means it exceeds our ability
            paths.remove(path)
        else:
            pass
path_0 = Generate_path(target_customer, XGBTree)
for node in path_0[::-1]:
    #from bottom to top
```

```
        if set(path_0.change(node)) & set(paths):
            # recursively change nodes and generate new paths
            # if there exists a path in new paths generated and original paths
            # then this is a strategy
            new_path = set(path_0.change(node)) & set(paths)
            for path in new_path:
                cost = calculate_total_cost(path_0, path)
                Cost[to_string(path)] = cost
    if Cost:
        sort_by_value(Cost)
        print(Cost[0])
    else:
        print("No best strategy")
    '''
    Output: path(string), least cost
    '''
```

算法开始根据输入 XGBoost 树结构中所有分数为正的叶子节点逐步回溯至根节点来找到所有可行路径，随后针对每一条路径，如果其中有有关于 B 属性的条件且目标客户的 B 属性不满足条件则删除该路径，如果 a 属性的条件超过了我们的最大干预能力即提升 5 个百分点，则同样删除。留下的可行路径中，先观察目标客户在树结构中会沿着哪一条实际路径被分类为不购买，随后针对这一路径从叶子节点出发，自底而上递归式回溯，观察改变 a 属性节点以后的所有路径集合是否与可行路径集交集非空，若存在则将所有的路径与需要改变的代价保存起来，按照代价大小排序输出最低代价；而若交集为空集则认为找不到最优策略。

经程序设计验证，该方法切实有效。

5.5 模型五的建立与求解

以下为我们的建议信：

==========

尊敬的销售部门领导：

您好！

我们对电动汽车目标客户销售策略进行了相关分析，通过合理地对异常值和缺失值进行处理，并进行客户对电动汽车各项指标的态度、客户自身生活状况的描述性统计分析，得到关于目标客户对不同品牌汽车满意度等统计意义上的结论；并将统计学方法和机器学习方法进行结合，筛选出包括经济性等在内的对不同品牌的汽车销售有影响的八项指标;运用因子分析-聚类模型和八属性 XGBoost 分类模型，建立了不同品牌电动汽车的客户挖掘模型，并在测试集上达到了很好的效果；根据问题三中模型结构，提出了生成精准营销策略的多目标路径寻优算法。

根据我们的研究成果，给出了一些关于电动汽车目标客户销售策略的建议：

（1）根据购买群体的特征，推出个性化的精准服务。购买人群在居住环境、学历、经济实力等方面具有较为明显的特征，应当根据相应特征，推荐能满足客户个性化需求的汽车。

（2）重视客户的经济状况与电动汽车的性价比。客户在购买电动汽车时，除了根据自身经济条件进行选择外，还会特别关注汽车电池、动力性能和经济保值性。因

此销售经济性与性价比相统一的汽车，成交的可能性更大。

（3）提升车辆自身属性，使客户获得更好的印象。家庭年收入等指标，属于消费者自身的属性，销售者无法干预，但可以通过提升电池性能等车辆属性来提升消费者的满意度。

以上建议仅供参考，还希望领导能够予以考虑。

此致

敬礼！

六、 模型的评价、改进与推广

6.1 模型的优点

该模型从统计学角度出发，结合机器学习算法，保证了结果的可靠性，随后我们根据一些模型评价指标评估，认为模型在本数据集上取得了较好的效果。我们认为，我们在这一系列问题中提出的模型有以下优点：

1.使用统计学与机器学习算法，结果准确可靠，模型表现好。

2.利用 XGBoost 进行了特征工程与特征筛选，有力降低维度选择最核心的指标。

3.通过聚类算法和因子分析对客户特征进行区分与画像，能够描述不同群体的独立特征，有助于精准营销。

4.在问题四的营销策略中根据 XGBoost 的生成树提出了营销策略生成算法，自底而上对可行策略进行递归式的变化分析，能高效准确地分析营销策略。

6.2 模型的缺点

尽管模型整体表现良好，但我们认为还存在一些不足之处，例如：

1.对于客户数据的画像，未能将两类客户在画像的同时进行区分，可以尝试在降维算法中加入聚类的效果进行测试。

2.营销策略的生成仍然存在一定的问题，算法时间复杂度较高并且严重依赖生成树，而且没能充分利用到营销成本与提升百分点的关系。可以考虑从结构方程的角度入手进行进一步分析。

6.3 模型的改进

该模型除了可以使用 XGBoost 作为分类器以外，同样属于 GBDT 框架下的 LightGBM 作为 XGBoost 的一种改进算法也可以用于尝试。另外，对于末尾提出的多目标路径寻优算法中出现有多次循环，加之本身就存在递归寻路的操作使得算法的时间复杂度较大，实际上这一系列循环是可以通过更改算法结构简省的。通过将循环合并或利用 numpy 等加速工具能够提升算法的工作效率。

6.4 模型的推广

该问题的求解方法不仅可以用于电动汽车的销售领域，结合问题实际的简单数值处理原理纯粹但有高度的可解释性和效率；统计学与机器学习结合的特征筛选与特征工程能高效筛选出强影响特征；聚类和以因子分析为代表的降维算法结合揭示了不同用户的特征而 XGBoost 也是一种很好的分类模型；最后，我们提出了基于寻路的最优销售策略生成算法。这一系列模型具有高度的可移植性。

七、 参考文献

[1] 李龙.配对样本t检验在实验室分析质量控制中的应用[J].上海计量测试,2020,47(05):32-34+37.

[2] 吕世杰,许茂发,任佳,姚荣,卫智军.卡方独立性检验的实践与可操作性研究[J].统计与管理,2015(05):41-44.

[3] Dong G, Liu H. Feature Engineering for Machine Learning and Data Analytics[M],2018

[4] Bicici Ufuk Can,Akarun Lale. Conditional information gain networks as sparse mixture of experts[J]. Pattern Recognition,2021,120:

[5] Mitchell T M . Machine Learning[M]. McGraw-Hill, 2003.

[6] Wu Cong,Li Hongxin,Ren Jiajia. Research on hierarchical clustering method based on partially-ordered Hasse graph[J]. Future Generation Computer Systems,2021(prepublish):

[7] Godwin Ogbuabor,Ugwoke F. N. Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value[J]. International Journal of Computer Science and Information Technology,2018,10(2):

[8] 杜丽英.基于数据挖掘的决策树算法分析[J].吉林建筑工程学院学报,2014,31(05):48-50.

[9] 方匡南,吴见彬,朱建平,谢邦昌.随机森林方法研究综述[J].统计与信息论坛,2011,26(03):32-38.

[10] Chen T , Tong H , Benesty M . xgboost: Extreme Gradient Boosting[J]. 2016.

[11] 王照国,张红云,苗夺谦.基于F1值的非极大值抑制阈值自动选取方法[J].智能系统学报,2020,15(05):1006-1012.

[12] 王彦光,朱鸿斌,徐维超.AUC统计特性概述[J].电子世界,2021(13):107-109.

附录

环境：OS: Windows 10; CPU: Intel i7; GPU: NVIDIA GEFORCE 1650
Language:
Python 3.8.2 Jupyter notebook
文件列表：
图片：
Boxplot-a1.png
Boxplot-a2.png
Boxplot-a3.png
Boxplot-a4.png
Boxplot-a5.png
Boxplot-a6.png
Boxplot-a7.png
Boxplot-a8.png
FA 聚类图.png
FA 三维聚类图.png
Single_tree.png
车贷.png
房贷.png
个人年收入.png
户口.png
个人职业.png
属性相关性.png
家庭年收入.png
居住区域.png
可支配年收入.png
年龄.png
生活情况.png
学历.png
购买比例.png
代码：
华数杯.ipynb
配置要求：有 sklearn，安装 xgboost 即可
代码见下页