所属类别	2020 年“华数杯”全国大学生数学建模竞赛	参赛编号
研究生组		CM212376

基于 logistic 回归的 GA-XGBoost 电动汽车目标客户销售模型

摘要

电动汽车作为一个新兴事物，某些领域，如电池、配置问题，仍会使消费者产生疑虑，所以其市场销售急需科学决策。本文针对 1964 位目标客户对三款品牌电动汽车的满意度得分和个人特征的信息，首先进行数据清洗，并得出目标客户对于 3 个不同品牌汽车满意度的比较分析，然后建立基于 logistic 回归的 GA-XGBoost 模型从而判断目标客户购买电动车的可能性，再判断目标客户能否通过提升五个百分点的满意度从不购买变为购买，最后给出销售策略建议。

对于问题一，对数据进行清洗，并做描述性统计分析。首先将数据分为连续型和离散型数据，并根据数据类型设定不同的异常值和缺失数据的处理方法。然后结合集中趋势、离散程度做描述性统计分析，并得出目标客户对于 3 个不同品牌汽车满意度的比较分析。结果显示，B7 数据中存在 500 个空缺值，其中 496 个依情境补为 0，4 个用众数 1 补齐。对于异常值，有 4 个超出范围的连续型数据，用均值代替；B14 或 B15 大于 B13 的有 93 个客户，直接删除。对三种品牌汽车满意度进行比较分析，发现客户对品牌 1 汽车的总体满意度最高，对品牌 3 的最低，品牌 3 的满意度离散程度最大，品牌 2 的最小。

对于问题二，研究不同品牌电动汽车销售的可能影响因素。首先将影响因素分为连续型和离散型。然后根据数据类型采用不同的方法找出会对销售有影响的因素，对于 16 项连续型的影响因素采用 Z 检验，对于 9 项离散型的选用 χ^2 检验。最后将选出的因素汇总分析。本文设定 $P < 0.05$ 为有显著性影响，结果显示，对品牌 1、品牌 2、品牌 3 电动汽车销售有影响的因素分别有 13 项、17 项、9 项，具体结果见表。

对于问题三，建立不同品牌电动汽车的客户挖掘模型，并判断 15 名目标客户购买电动车的可能性。首先分别用最优子集法和 logistic 回归筛选主要影响因素。然后建立 XGBoost 模型进行预测，再应用遗传算法对参数进行调优，得出基于最优子集的 GA-XGBoost 模型和基于 logistic 回归的 GA-XGBoost 模型。之后用预测评价指标评价模型的优良性，通过分析比较，基于 logistic 回归的 GA-XGBoost 模型效果更好。最后运用该模型判断附件 3 中 15 名目标客户购买电动车的可能性，预测结果为 1 号顾客会购买品牌 1 汽车，6 号顾客会购买品牌 2 汽车，12 号顾客会购买品牌 3 汽车。

对于问题四，判断目标客户能否通过提升五个百分点的满意度从不购买变为购买。首先选出问题三中预测为 0 的客户数据，并分析三个品牌的基于 logistic 回归的 GA-XGBoost 模型，将被选入模型的满意度因素都提升五个百分点。然后将提升后的数据分别导入三个品牌的模型，看有哪些客户购买意愿从 0 变为 1。再将满意度提升点数逐渐降低，直至客户购买意愿不再从 0 变为 1，从而得到最低的满意度提升点数。最后从每个品牌中各挑选 1 名能通过提升五个百分点满意度从不购买变为购买的目标客户，实施销售策略。

关键字： 电动汽车　满意度　χ^2 检验　logistic 回归　GA-XGBoost

1. 前言

1.1 问题重述

某汽车公司最新推出三款品牌电动汽车，用 1、2、3 表示，销售部门邀请了 1964 位目标客户对其满意度打分，得到满意度 a1-a8，及目标客户的个人信息。通过建立数学模型解决以下问题：

1. 数据清洗，指出异常值、缺失值及处理方法，并分析比较目标客户对于不同品牌汽车的满意度。

2. 研究不同品牌电动汽车销售的可能影响因素。

3. 建立不同品牌电动汽车的客户挖掘模型，并运用其判断 15 名目标客户购买电动汽车的可能性。

4. 短时间内可以提高五个百分点的满意度，且服务难度与提高的满意度成正比，据此从每个品牌中选出 1 名没有购买电动车目标客户，对其实施销售策略。

5. 根据以上结论，给出销售部门不超过 500 字的销售建议。

1.2 问题分析

问题一分析

对于问题一，首先对数据进行分析，清楚其数据特点。为了更有效地进行数据预处理，将数据分为连续型和离散型数据，并根据数据类型设定不同的异常值和缺失数据的处理方法。然后结合集中趋势、离散程度做描述性统计分析，包括目标客户对于不同品牌汽车满意度的比较分析。

问题二分析

对于问题二，首先将影响因素分为连续型和离散型，再分别对三种不同品牌电动汽车根据数据类型采用不同的方法找出会对其销售有影响的因素。对于连续型的影响因素，因样本量比较大，可以采用 Z 检验，分析各因素对客户是否购买的影响。对于离散型的影响因素，因变量是二分类数据，所以选用 χ^2 检验。最后将选出的因素汇总分析。根据数据类型采用不同的方法找出会对其销售有影响的因素

问题三分析

对于问题三，首先分别用最优子集法和 logistic 回归筛选主要影响因素。然后建立 XGBoost 模型进行预测，再应用遗传算法对参数进行调优，得出基于最优子集的 GA-XGBoost 模型和基于 logistic 回归的 GA-XGBoost 模型。之后用预测评价指标评价模型的优良性，并选出效果较好的模型。最后运用该模型判断附件 3 中 15 名目标客户购买电动车的可能性。

问题四分析

对于问题四，首先选出问题三中预测为 0 的客户数据，并分析三个品牌的基于 logistic 回归的 GA-XGBoost 模型，将被选入模型的满意度因素都提升五个百分点。然后将提升后的数据分别导入三个品牌的模型，看有哪些客户购买意愿从 0 变为 1。再将满意度提升点数逐渐降低，直至客户购买意愿不再从 0 变为 1，从而得到最低的满意度提升点数。最后从每个品牌中各挑选 1 名能通过提升五个百分点满意度从不购买变为购买的目标客户，实施销售策略。

1.3 技术路线图

本文的技术路线图如下：

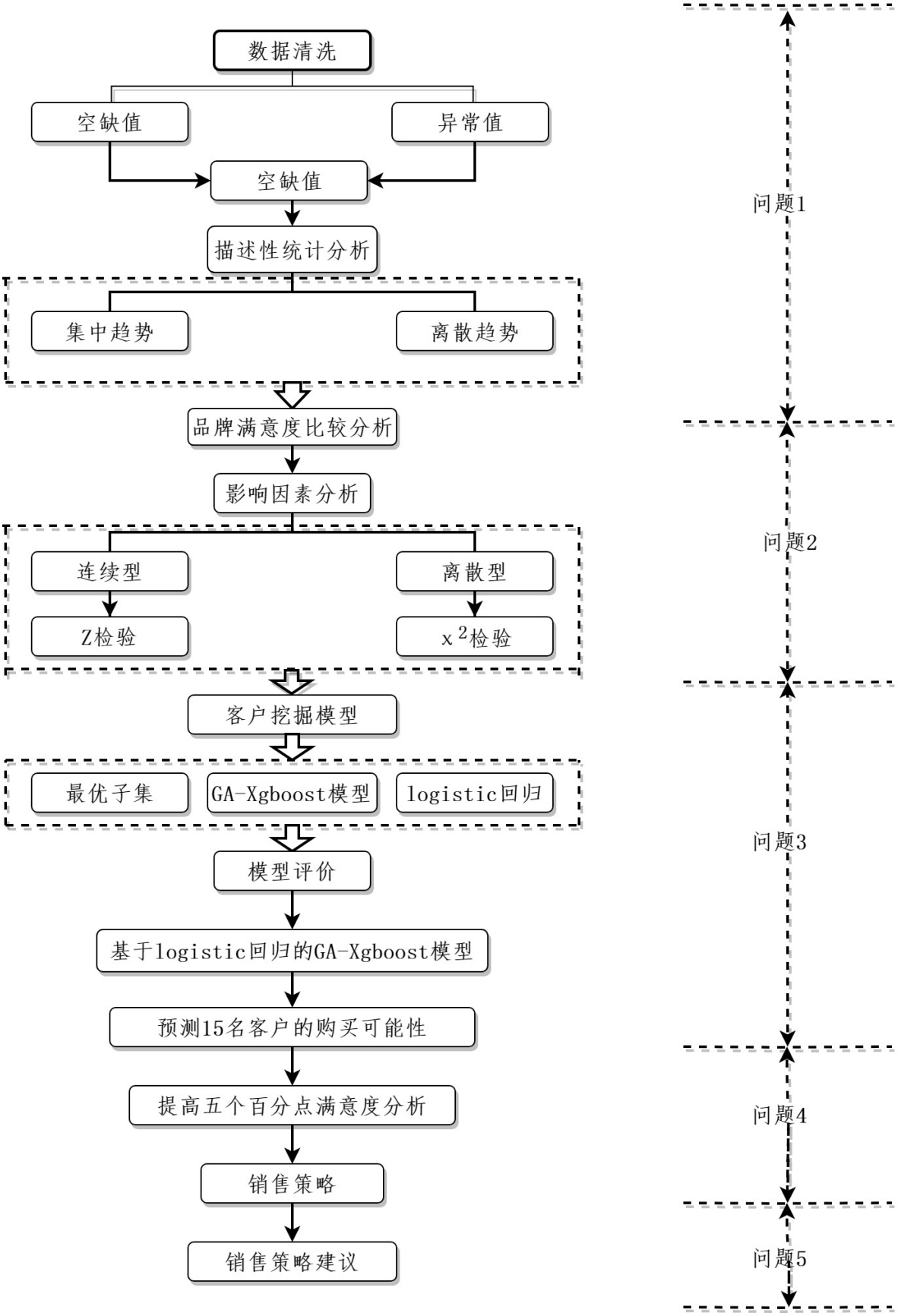
数据清洗
空缺值
异常值
空缺值
描述性统计分析
集中趋势
离散趋势
品牌满意度比较分析
影响因素分析
连续型
离散型
Z检验
x 2检验
客户挖掘模型
最优子集
GA-Xgboost模型
logistic回归
模型评价
基于logistic回归的GA-Xgboost模型
预测15名客户的购买可能性
提高五个百分点满意度分析
销售策略
销售策略建议
问题1
问题2
问题3
问题4
问题5

图 1　技术路线图

1.4 模型的假设

1. 假设目标客户体验者按真实感受打分；
2. 假设目标客户体验者填写的个人特征的信息真实可靠；
3. 假设其他因素对品牌电动汽车的销售没有显著影响；
4. 假设加大服务力度可以同时提高 a1-a8 满意度相同的百分点。

1.5 符号说明

表 1　符号说明

符号	意义
N	样本量
n	变量数
T	logistic 回归阳性数

2. 问题一：统计分析

2.1 数据清洗

对本文的数据进行分析，发现文中数据存在空缺值以及异常值。为了更好地进行数据分析，本文将数据 B6（出生年份）转化为了年龄段并用数字 1-5 编码，具体见表2：

表 2　B6 数据转化

出生年份	年龄段	编码
1991-2001	20-30	1
1986-1991	30-35	2
1976-1986	35-45	3
1966-1976	45-55	4
1966 以下	55 以上	5

本文对数据中空缺值及异常值的整理思路如图2：

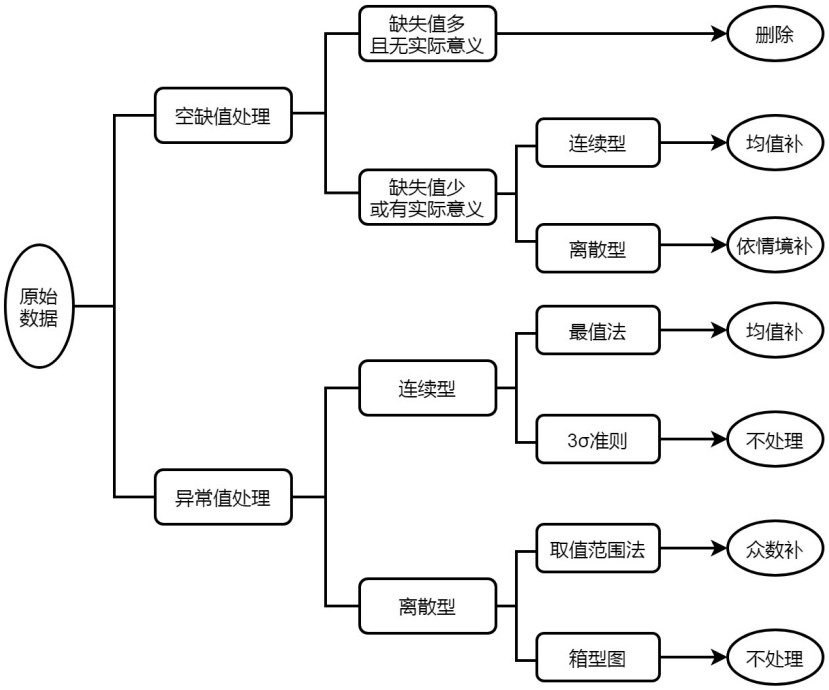

图 2　样本确定方法

2.1.1 空缺值处理

本文将空缺值分为两个类。第一类是缺失值多并且无实际意义的缺失值，这类缺失数据对总体样本影响不大，故直接删除。第二类为缺失值少或者有实际意义的缺失值，这类缺失数据若是连续型数据，则采用均值替补；若为离散型数据，则依情境补充。例如，问题 B7 的数据存在较多的空缺值，若问题 B6 的回答是 1、2、3、4、8 这几个选项，则问题 B7 空缺值以 0 补，否则用众数补。

2.1.2 异常值处理

我们通常将明显偏离它们样本的数据称为异常值，本文将数据分为连续型和离散型，分别用不同的方法对其异常值进行处理。

对于连续型数据，我们首先采用最值法，例如数据 a1，对汽车的舒适性（环保与空间座椅）整体表现满意度打分，总分为 100 分，那么超出 100 分的一定为异常值，这类异常值我们用均值替补。但是 B13、B14、B15 无法用这一范围表示，分析附录 2 中的问卷调查得，B13、B14、B15 数据必然大于 0，且 B14、B15 数据一定小于 B13，否则数据不符合现实情况，故本文将 B13、B14、B15 大于 0 且 B14、B15 小于 B13 作为其范围。由于数据 B13、B14、B15 关联性较强，用均值替补不适合，故直接删除。然后再用 3σ 准则检测异常值情况，发现这类异常值仍然存在重要信息，故不作处理。

对于离散型数据，首先通过取值范围进行判断，比如问题 B1：您的户口情况？只有 1、2、3 三个选项，所以若 B1 出现其他值那么则认为是异常值，这类异常值采用众数替补；然后再用箱型图检验异常值情况，分析认为这类异常值仍然存在重要信息，故也不做处理。

数据的具体范围，见表3：

表 3　变量的取值范围

数据类型	变量	范围
连续型	a1，a2，...,a8 B2,B4,B10,B16,B17	0-100
	B13,B14,B15	>0 且 B14,B15≤B13
离散型	B1	int,1-3
	B3	int,1-6
	B5	int,1-20
	B6	int,1-8
	B7	int,0-20
	B8	int,1-5
	B9	int,1-8
	B11	int,1-9
	B12	int,1-11

3σ 准则

在处理数据时，有时为了提高数据的准确性，需要对某些异常值进行剔除，在这种情况下，可以考虑 3σ 准则。

假设一组测量变量只包含随机误差，设这些变量分别为 $x_1, x_2, ..., x_n$, 对它们求均值记为 x, 计算出剩余误差 $v_i = x_i - x, i = 1, 2, ..., n$, 利用贝塞尔公式算出标准误差 σ，若某个测量值 x_i 的剩余误差 v_i，满足 $|v_i| = |x_i - x| > 3\sigma$, 就认为此剩余误差属于粗大误差，应该剔除。贝塞尔公式如下：

$$\sigma = [\frac{1}{n-1}\sum_{i=1}^{n} V_i^2]^{1/2} = \left\{ \left[\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^{\ 2} /n \right] /(n-1) \right\}^{1/2}$$

箱型图

其中箱型图又称盒式图，是美国杰出学家 John Tukey 于 1977 年发明。箱型图不仅可以反映出原始数据的分布特征，且能够直接看出数据的异常值。如图3：

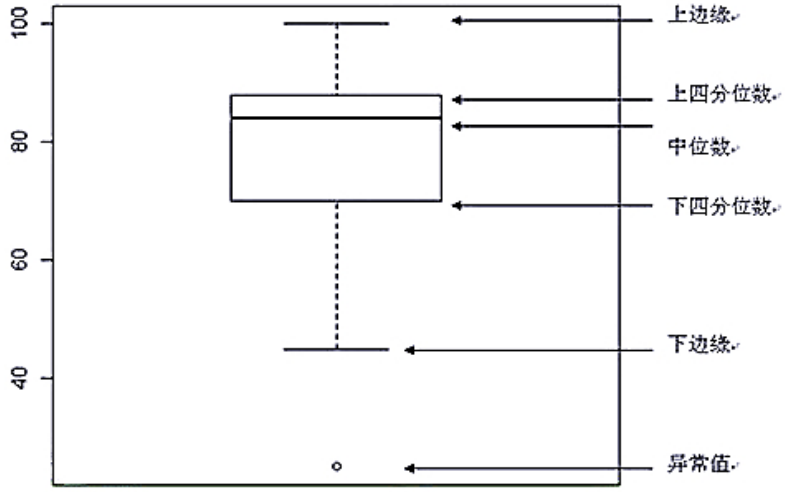

图 3　箱型图图示

绘制箱型图的具体步骤如下：

1. 将数据由小到大排序，找出中位数，上四分位数和下四分位数以及数据的上边缘和下边缘；

2. 然后连接上四分位数和下四分位数构成箱体；

3. 将上下边缘与箱体相连，中位数位于箱体中间位置。

在这里我们记上四分位数为 Q_3, 下四分位数 Q_1，IQR 为统计内距，即 $IQR = Q_3 - Q_1$，上边缘为 $Q_3 + 1.5IQR$, 下边缘为 $Q_1 - 1.5IQR$，其中上边缘和下边缘又称为异常值截断点，处于上边缘或下边缘之外的点就为数据的异常点。

2.1.3 数据清洗结果

利用以上数据清洗方法，用 Python 软件来预处理附录 1 中的数据，结果显示，B7 数据中存在 500 个空缺值，其中 B6 为 1-4，即“没有小孩”的有 495 个，B6 为 8，即“其他”的有 1 个，这些空缺值直接依情境补为 0；B6 为 5，6，7 的有 4 个，这些空缺值用众数补齐，经计算众数为 1，故补为 1。对于异常值，首先数据 a1，a3，a5，B17 超出范围的各一个，用均值代替；B14 大于 B13 或 B15 大于 B13 的一共有 93 个客户，这些数据直接删除；其他通过 3σ 准则及箱型图发现的异常值均蕴含有用信息，故不作处理。

其中部分箱型图如图4：

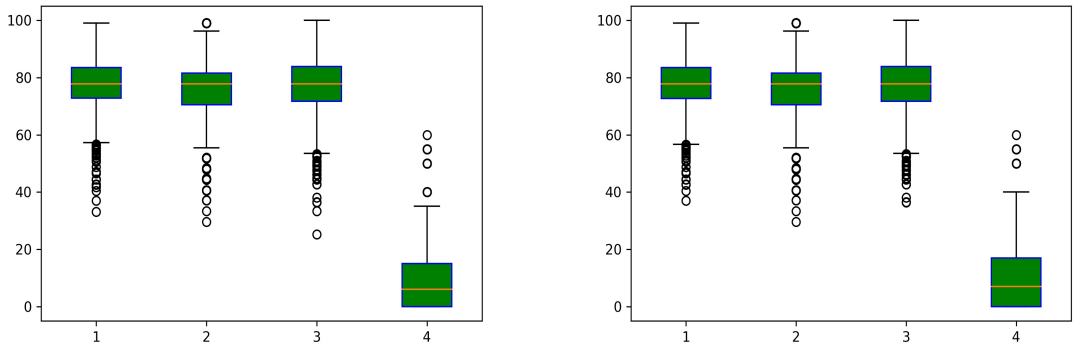

图 4　部分箱型图结果

2.2 数据的描述性统计分析

描述性统计分析主要包括数据的集中趋势分析、离散程度分析。

数据的集中趋势

通常将数据一组数据向某一中心值靠拢的程度，其度量方法包括：均值、中位数、众数。由于它们的定义、特点和使用范围不同，故衡量这些度量方法的代表准确性时，要依据不同情况具体分析。

查阅资料，得到不同类型数据最适合的度量方法见表4：

表 4　数据类型与集中趋势测度值

数据类型	定类数据	定序数据	定距数据	定比数据
适应的测量值	众数	中位数	均值	均值

对附件中的数据进行分析，发现 a1-a8、B2、B4、B10、B13、B14、B15、B16、B17 为定比数据，故用均值代表集中趋势。其他为定类数据，故用众数代表集中趋势。

计算得定比数据的均值，具体见表5；定类数据 B1、B3、B5、B6、B7、B8、B9、B11、B12 的众数分别为：2，即“户口在本城市”；1，即“居住在市中心”；3，即“家中有 3 口人”；5，即“已婚，有小孩且不与父母同住”；1，即“家中有 1 个孩子”；3，即“年龄在 30-35”；6，即“本科学历”；4，即“在私营或民营企业工作”；2，即“职位是中层管理者”。

数据的离散趋势

数据的离散趋势是指变量值远离其中心值的程度，本文分别从四分位差、标准差两个层面衡量数据的离散趋势。

对连续型数据的集中趋势及离散趋势的统计描述具体情况见表5。

2.3 不同品牌汽车满意度的比较分析

通过对三种品牌汽车满意度进行比较分析，发现数据 a2、a4、a5、a6、a7、a8 满意度均值按照降序排序都为 1、2、3；而 a1、a3 满意度均值按照降序排序为 2、1、3。这说明从均值层面分析，客户对品牌 1 汽车（合资品牌）的总体满意度最高，对品牌 3 汽车（新势力品牌）的总体满意度最低。满意度均值具体比较情况，见图5。

满意度标准差比较，发现数据 a1-a8 满意度全距按照降序排列全为 3、1、2。说明品牌 3 的满意度离散程度较大，说明有些客户对品牌 3 较为满意，有些客户对品牌 3 非常不满意；品牌 2 的满意度离散程度较小，说明客户对品牌 2 的满意度差距不大。满意度标准差具体比较情况，见图5。

表 5　连续型数据描述统计一览表

变量	最大值	最小值	均值	四分位差	标准差
a1	99.04	37.04	78.26	10.78	8.84
a2	99.03	43.4	78.13	11.03	8.98
a3	99.03	29.59	76.22	11.11	10.47
a4	99.98	40.7	78.84	11.11	9.04
a5	99.98	36.4	77.43	12.18	9.37
a6	99.99	39.15	77.88	11.11	9.31
a7	99.99	7.88	78.02	11.11	9.17
a8	99.98	40.33	77.58	11.11	9.5
B2	60	1	21	1	11.49
B4	30	1	7.6	21	4.16
B10	38	1	10	1	4.94
B13	100	7	26.77	5	12.51
B14	95	3	16.6	1	10.74
B15	80	2	16.43	1	9.96
B16	60	0	15.28	0	13.11
B17	60	0	9.59	6	10.65

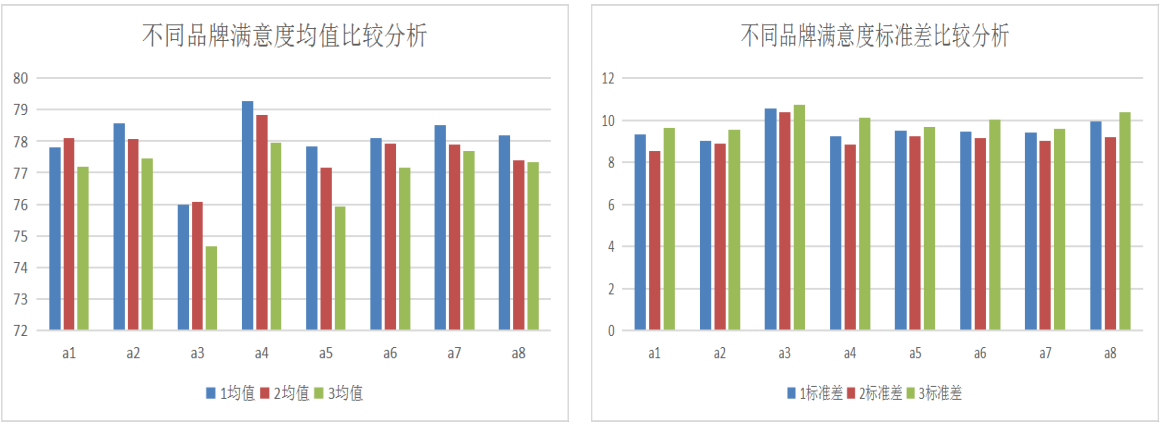

图 5　不同品牌汽车满意度比较分析

3. 问题二：因素筛选

本文对 16 项连续型因素进行 Z 检验，对 9 项离散型因素进行 χ^2 检验。

对于品牌 1，在 16 项连续型因素的 Z 检验结果中有 12 项因素的 P 值 <0.05（结果见表6）。对 9 项离散型因素的 χ^2 检验结果中只有 B3（居住区域）这个指标的 $P < 0.05$。故

我们认为影响品牌 1 销售的因素有：a1：电池技术性能（电池耐用和充电方便）、a2：舒适性（环保与空间座椅）、a3：经济性（耗能与保值率）、a4: 安全性表现（刹车和行车视野）、a5: 动力性表现（爬坡和加速）、a6: 驾驶操控性表现（转弯和高速的稳定性）、a7: 外观内饰、a8: 配置与质量品质、B2: 在本城市居住年数、B15: 家庭可支配年收入、B16: 全年房贷的支出占家庭年总收入的比例、B17: 全年车贷的支出占家庭年总收入的比例、B3：居住区域。

对于品牌 2，在 16 项连续型因素的 Z 检验结果中有 14 项因素的 P 值 <0.05，分别是 a1-a8、B2、B14、B15、B16、B17（结果见表6）。对 9 项分类指标的 χ^2 检验结果中有 3 项指标的 P 值 <0.05，分别是 B9、B11、B12。因此我们认为影响品牌 2 汽车销售的因素有：a1：电池技术性能（电池耐用和充电方便）、a2：舒适性（环保与空间座椅）、a3：经济性（耗能与保值率）、a4: 安全性表现（刹车和行车视野）、a5: 动力性表现（爬坡和加速）、a6: 驾驶操控性表现（转弯和高速的稳定性）、a7: 外观内饰、a8: 配置与质量品质、B2: 在本城市居住年数、B14：个人年收入、B15: 家庭可支配年收入、B16: 全年房贷的支出占家庭年总收入的比例、B17: 全年车贷的支出占家庭年总收入的比例、、B9：学历、B11：所在单位的性质、B12：职位。

对于品牌 3，在 16 项连续型因素的 Z 检验中有 9 项指标的 P 值 <0.05，分别是 a1-a8、B16（结果见表6）。在 9 项离散型因素的 χ^2 检验结果中没有指标的 P 值 <0.05。故我们认为影响品牌 3 汽车销售的因素为：a1：电池技术性能（电池耐用和充电方便）、a2：舒适性（环保与空间座椅）、a3：经济性（耗能与保值率）、a4: 安全性表现（刹车和行车视野）、a5: 动力性表现（爬坡和加速）、a6: 驾驶操控性表现（转弯和高速的稳定性）、a7: 外观内饰、a8: 配置与质量品质、B16: 全年房贷的支出占家庭年总收入的比例。

表 6　连续型因素 Z 检验结果

	品牌 1		品牌 2		品牌 3	
	因素	p 值	因素	p 值	因素	p 值
a1	-7.795	.000	-18.257	.000	-4.587	.000
a2	-7.789	.000	-14.781	.000	-4.712	.000
a3	-9.396	.000	-14.957	.000	-7.15	.000
a4	-3.631	.000	-14.197	.000	-2.984	.003
a5	-3.631	.000	-15.945	.000	-4.168	.000
a6	-3.729	.000	-12.277	.000	-3.431	.001
a7	-6.098	.000	-14.264	.000	-2.64	.009
a8	-3.275	.001	-11.751	.000	-3.071	.003
B2	-2.064	.040	\	\	\	\
B13	\	\	-3.71	.000	\	\
B14	\	\	-2.971	0.003	\	\
B15	-2.046	.041	-4.471	.000	\	\
B16	9.481	.000	16.17	.000	2.077	.040
B17	7.136	.000	13.099	.000	\	\

4. 问题三：基于 logistic 回归的 GA-XGBoost 模型

首先分别用最优子集法和 logistic 回归筛选主要影响因素。然后建立 XGBoost 模型进行预测，再应用遗传算法对参数进行调优，得出基于最优子集的 GA-XGBoost 模型和基于 logistic 回归的 GA-XGBoost 模型。之后用预测评价指标评价模型的优良性，并选出效果较好的模型。

4.1 最优子集法

最优子集选择的思路很容易理解，就是把所有自变量的组合都拟合一遍，比较一下哪个模型更好，选出最优模型。

给定一组观测值 $(x_{i1}, x_{i2}, ...x_{ip}, y_i(i = 1, 2, ...n))$，$p$ 个因素有 $2^p - 1$ 个模型，那么哪个是最优的呢？根据一些比较标准我们就能得到最优模型。以下有两个常用比较标准：

- Bayesian Information Criterion（BIC）

$$BIC = k\ln(n) - 2\ln SSE(p')$$

BIC 较小的模型最优；

- Mallows 的 $C(p)$ 统计量：

$$C(p) = \frac{SSE(p')}{MSE(p')} - [n - 2(p' + 1)]$$

$C(p)$ 的值较小并且与 p' 较接近的模型最优。

其中:p' 是用来建立回归方程的因子个数；$SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ 为残差平方和，$y_i, \hat{y_i}$ 分别为观测值和预测值；$MSE = \frac{SSE(p)}{n-p-1}$ 为残差均方和。上述评判标准会同时指向同一模型，如果不是，选择得到指向最多的模型。

4.2 logistic 回归

第二问通过 Z 检验、χ^2 检验找出了影响不同品牌汽车销售的可能影响因素，为了提高建立模型的精确度，还需要筛选出影响不同品牌汽车销售的主要因素。

查阅文献资料将自变量筛选常见的方法整理见表7：

表 7　自变量筛选常见方法

样本情况	严进	宽出
N<10n， T<10n	所有候选变量先进行单因素筛选， P 值 <0.05 的纳入	（1）P<0.05 的纳入因素不多， 则采用一次性将 P<0.05 的因素纳入模型 （2）P<0.05 的纳入因素较多， 则需要采用逐步回归法进行变量筛选
10n<N<20n， 10n<T<20n	所有候选变量可以（但不是必须） 先进行单因素筛选，P<0.05 纳入	（1）采用全变量模型， 一次性纳入 P<0.05 的变量开展多因素回归 （2）如果存在着多重共线性，导致异常结果出现， 则需要采用逐步回归法对 P<0.05 的变量进行筛选
N>20n, T>20n	所有候选变量全部纳入， 不必进行单因素筛选	（1）采用全变量模型， 所有变量一次性纳入多因素回归 （2）如果存在多重共线性，导致异常结果出现， 则需要采用逐步回归法进行变量筛选
注：N: 样本量，n：变量数，T:logistic 回归阳性数		

分析数据得，品牌 1、2 中 $N > 20n$，按照表7中的方法应该选择第三类情况，但是它的 $T < 10n$, 故选择第一类的方法更合适。经分析，发现 $P < 0.05$ 的因素很少，因此采用一次性将 $P < 0.05$ 的因素全部纳入模型中。

对于品牌 3，它的 $N < 10n$,$T < 10n$, 故需要先进行单因素筛选，发现 $P < 0.05$ 不多，所以也采用一次性将 $P < 0.05$ 的因素全部纳入模型中。

logistic 回归实质上是把不同类别的样本区分开来。对于二分类问题 $y \in \{0, 1\}$,1 表示正例，0 表示负例。logistic 回归是在线性函数 $\theta^T x$ 输出预测实际值的基础上，寻找一个假设函数 $h_\theta(x) = g\left(\theta^T x\right)$, 将实际值映射到 0，1 之间，如果 $h_\theta(x) \geq 0.5$，则预测 $y = 1$，及 y 属于正例；如果 $h_\theta(x) < 0.5$, 则预测 $y = 0$，即 y 属于负例。

4.3 XGBoost 模型

XGBoost 模型的算法具有快速、高效、泛华能力强、适应于处理大规模数据等优点，故本文建立 XGBoost 模型对附件 3 中 15 名目标客户购买电动车的可能性进行预测。

对于 n 条 m 维的数据集：

$$\mathrm{D} = \{(\mathrm{x_i}, \quad \mathrm{y_i})\} \left(\mathrm{x_i} \in \mathrm{R^m}, \mathrm{y_i} \in \mathrm{R}, \mathrm{i} = 1, 2, \ldots \mathrm{n}\right)$$

XGBoost 模型表示为：

$$\widehat{y_l} = \sum_{k=1}^{K} f_k\left(x_i\right), f_k \in F(i = 1, 2, \ldots n)$$

上式中，K 代表树的颗数，x_i 表示第 i 个数据点的特征向量，f_k 表示一颗具体的 CART 决策树，F 是所有 CART 决策树的结构集合。

目标函数包含两部分：

$$Obj = \sum_{i=1}^{n} l\left(y_i, \widehat{y_l}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)$$

上式中：l 代表训练中存在的误差，是预测值和目标值之间按照特定标准计算的差异程度。Ω 则代表所有 CART 树的复杂度之和。

由于 Xboost 模型不是一个显式函数，无法用传统的方法进行优化，Xboost 使用加性训练进行优化，也就是每次优化一棵树，直到所有树都得到优化。程序如下：

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

$$\widehat{y}_i^{(1)} = \mathrm{f}_1\left(x_i\right) = \hat{y}_i^{(0)} + \mathrm{f}_1\left(x_i\right)$$

$$\widehat{y}_i^{(2)} = \mathrm{f}_1\left(x_i\right) + \mathrm{f}_2\left(x_i\right) = \hat{y}_i^{(1)} + \mathrm{f}_2\left(x_i\right)$$

$$\cdots\cdots$$

$$\hat{y}_i^{(t)} = \sum_{\mathrm{k}=1}^{\mathrm{t}} \mathrm{f}_{\mathrm{k}}\left(x_i\right) = \hat{y}_{\mathrm{i}}^{(\mathrm{t}-1)} + \mathrm{f}_{\mathrm{t}}\left(x_i\right)$$

其中，$\hat{y}_i^{(t)}$ 为第 t 次模型的预测值，$\mathrm{f}_{\mathrm{t}}\left(x_i\right)$ 为第 t 次加入的新函数。每一轮加入新函数是为了尽可能让目标函数值最大程度的减小。将 $\hat{y}_i^{(t)}$ 带入公式 $Obj = \sum_{i=1}^{n} l\left(y_i, \widehat{y_l}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right)$ 得到第 t 次模型的目标函数：

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

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

采用泰勒展开近似定义误差函数 l，得到目标函数为：

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, \widehat{y}_i^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \right]$$

$$+\Omega\left(f_t\right) + \ constant$$

其中，$g_i = \partial_{\widehat{y}_l(t-1)} l\left(y_i, , \widehat{y}_l^{(t-1)}\right)$，$h_i = \partial^2_{\widehat{y}_l^{(t-1)}} l\left(y_i, \hat{y}_l^{(t-1)}\right)$，即误差函数 l 的一阶导数和二阶导数。把不影响优化的常数项移除后，得到最终目标函数：

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

XGBoost 模型的伪代码如表8:

表 8　XGBoost 模型的伪代码

GA-XGBooat 伪代码	
输入：	种群数量 P, 迭代次数 T，参数数量 N，优秀个体数量 M
1.	for i←1 to P do
2.	初始化各种群的参数组 (（$\theta_{i1}, \theta_{i2}, \cdots\cdots, \theta_{iN}$)
3.	end for
4.	while(不满足停止准则)
5.	训练集训练 XGBoost 模型
6.	对验证集进行预测，计算适应度
7.	根据适应度由高到低保留 M 组优秀个体
8.	GA(P,T,N,M)
9.	产生新参数
10.	end while
输出：	最佳参数组合

4.4 遗传算法模型

遗传算法（GA）又叫基因算法或进化算法，是一种启发式搜索算法。遗传算法中每个个体都是独立的一个解，通过选择、交叉、变异操作模拟生物的进化过程，从而产生一群更适应环境的个体，重复该过程不断繁衍进化，最后得到一群最适应环境的解。

遗传算法的伪代码如表9:

表 9　遗传算法伪代码

遗传算法伪代码	
输入：	种群规模 P, 迭代次数 T，交叉概率 PC，变异概率 PM 等参数
1.	遗传代数 t=0
2.	初始化种群 P(t）
3.	计算 P(t）适应度
4.	while（不满足停止准则）do
5.	t=t+1
6.	从 P(t-1）中选择 P(t）
7.	按照 PC 进行交叉
8.	按照 PM 进行变异
9.	产生新的种群 P(t）
10.	计算 P(t）适应度
11.	end while
输出：	最佳个体

4.5 模型评价

1、准确率（accuracy) 是预测正确的样本占全部样本的比例，公式为：

$$accuracy(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\hat{y}_i = y_i\right)$$

2、召回率（recall）是针对原始样本而言的指标，表示真正样本中有多少预测对了，公式为：

$$\frac{TP}{TP + FN}$$

3、F1 分数 (F1-score) 是 precision 和 recall 的调和平均数（倒数平均数），将它们看成同等重要。通常 precision 和 recall 不能兼得，查准率高了查全率可能会偏低，查全率高了查准率可能会偏低，使用 F1 分数可以综合考虑它们二者。

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 \times precision \times recall}{precision + recall}$$

4、ROC 曲线, 绘制 ROC 曲线需要计算真正例率（TPR) 和假正例率（FPR), 其定义为：$TPR = \frac{TP}{TP+FN}$，$FPR = \frac{FP}{TN+FP}$。ROC 曲线以图形方式组合 TPR 和 FPR，可以直观反映分类器的真正例率和假正利率的关系。但是在实际应用中 ROC 曲线可能出现交叉情况，无法直观判断孰好孰坏，所以通常依据 ROC 曲线下方的面积 AUC 来判断，AUC 越靠近 1，说明模型性能越好。

本文分别通过准确率、召回率、F1 分数及 ROC 曲线这四方面对模型进行评价，具体结果见图6。由图6可以明显看出基于 logistic 回归的模型各项得分都比较高，故本文采用基于 logistic 回归的预测模型。

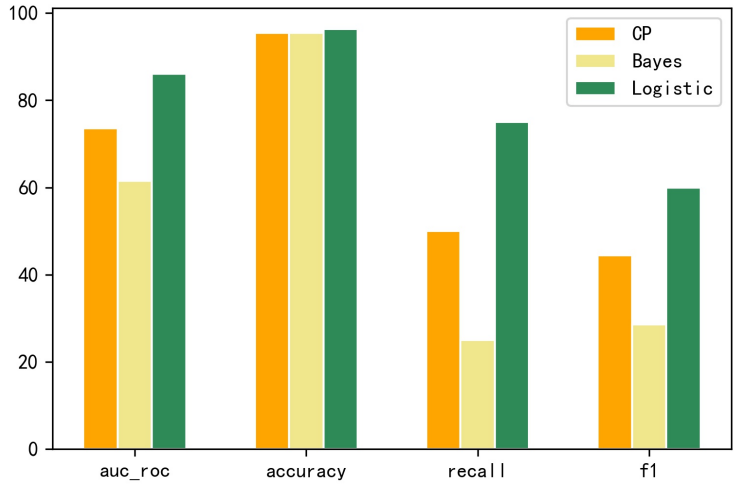

图 6　基于最优子集回归和 logistic 回归的模型评价

4.6 附录 3 的预测

预测前，需要对附录 3 的数据按照问题一中的数据清洗方法进行数据预处理。但是附录 3 中出现的空缺值或异常值都是非主要影响因素，在 logistic 回归中已经被剔除，所以对附录 3 中的数据不作处理。

对附录 3 中的数据进行基于 logistic 回归模型的预测，预测结果为：1 号顾客会购买品牌 1 汽车，6 号顾客会购买品牌 2 汽车，12 号顾客会购买品牌 3 汽车。

5. 问题四：提升满意度销售策略

对于品牌 1，只有 1 号客户会购买，所以需要加大服务力度，提高 2、3、4、5 号顾客对品牌 1 汽车的满意度，至其愿意购买。由以上分析结论得，品牌 1 汽车销售的主要影响因素是 a1、a3，我们首先将这些因素的满意度全部提高百分之五，发现客户购买品牌 1 汽车的意愿并没有提高。经分析发现 2、3、4 号顾客车贷占全年收入的一定比例，则表明 2、3、4 号顾客近期已经购买车辆足以家庭的日常出入需求，故不再需要额外购买车辆；而 5 号顾客尽管 B16 为 0，即没有车贷，但是 B17 高达 30，表明 5 号顾客近期房贷压力比较大，在金钱方面可能不足以再购买一辆车了。

对于品牌 2，只有 5 号客户会购买，故需要合理提高 7、8、9、10 号顾客对品牌 2 汽车的满意度，至其愿意购买。品牌 2 汽车销售的主要影响因素是 a1、a3、a5。首先将这些因素的满意度全部提高百分之五，发现 7 号顾客的购买意愿由 0 转为 1，接着逐一降低各因素的百分比，发现当百分比降为百分之一的时候 7 号顾客的购买意愿不改变，故我们认为至少要将各主要影响因素满意度提高百分之二，才能使 7 号顾客由以前不买车，变为买车。也就是说，我们要加大 a1：舒适度、a3：安全性、a5: 驾驶操控性表现，这三个方面的服务力度，将这 3 个因素的满意度提高百分之二，从而让 7 号顾客购买品牌 2 汽车。

对于品牌 3，只有 12 号顾客会购买，因此需要合理提高 11、13、14、15 号顾客对品牌 3 汽车的满意度，至其愿意购买。影响品牌 3 汽车销售的主要因素是 a2。将 a2 提高百

分之五，发现 11 号顾客的购买意愿被改变，再逐一降低 a2 的满意度，发现至少需要将 a2 满意度提高百分之三，才能使 11 号顾客改变购买意愿。也就是说，我们要加大服务力度，将 a2：经济性（耗能与保值率）的满意度提高百分之三，从而使 11 号顾客购买品牌 3 汽车。

6. 问题五：销售策略建议

首先销售人员要锁定目标客户群，精准把握产品的市场定位，制定合理的销售计划。其次对客户的信息记录要力求完整、有效，不可丢失重要信息，从而根据客户信息提供精准服务，确保服务目标有效完成。

通过模型训练及给出的预测结果，从品牌 1、品牌 2、品牌 3 整体来看，客户特征中 B16(全年房贷的支出占家庭年总收入的比例) 和 B17(全年车贷的支出占家庭年总收入的比例) 两项对于消费者的购车决策起到了决定性作用，通常这两项贷款支出较高的客户不会有购车意愿，故销售人员应将目标群体锁定为两项贷款支出较低的客户。

进一步通过第四问的结果可以看出，在满足上述要求的客户中，当其主观满意度低于 70 分时，一般较难通过提高服务力度改变客户的购车决策，而对于满意度高于 80 分以上时，则更有可能实现销售策略，故销售人员需要对客户的信息进行仔细分析，挑选满意度较高的客户群体实施销售策略更有可能成功。

7. 模型评价及展望

模型优点：

本文在特征筛选时通过最优子集与逻辑回归两种方法，所筛选出的特征可靠性较高，拟合得到的模型经度高，效果好。

本文将筛选出的特征输入到 XGBoost 集成模型中，再设置 ROC_AUC_score 为目标函数值用遗传算法对其超参数进行优化得到 ROC_AUC、accuracy 、recall、F1 四个分数均较高的模型，有效的避免了人为调整超参数的繁琐性以及不准确性，得到了性能优良的模型。

模型缺点：

由于时间问题，本文在处理第四问提高服务力度时采取同步提升满意度而不是将满意度提升分配在输入模型的几个特征中。

展望：

在本文的基础上，可以更加有效的探索将满意度提升分配给几个特征时对客户购车决策的影响，提出针对某一个满意度更加精确的销售建议。

参考文献

[1] Tiago Martins,Rui Neves. Stock Exchange Trading Using Grid Pattern Optimized by A Genetic Algorithm with Speciation[M].:2021-07-15.

[2] Min wook Kang,Paul Schonfeld. Artificial Intelligence In Highway Location And Alignment Optimization: Applications Of Genetic Algorithms In Searching, Evaluating, And Optimizing Highway Location And Alignments[M].World Scientific Publishing Company:2020-08-14.

[3] 张春富. 基于参数优化的机器学习算法在糖尿病预测中的应用 [D]. 西南科技大学,2020.

[4] Jili Tao,Ridong Zhang,Yong Zhu. DNA Computing Based Genetic Algorithm[M].Springer, Singapore:2020-01-01.

[5] Mo Jamshidi,Renato A. Krohling,Leandro dos S. Coelho,Peter J. Fleming. Robust Control Systems with Genetic Algorithms[M].CRC Press:2018-10-03.

[6] John J. Grefenstette. Genetic Algorithms and their Applications[M].Taylor and Francis:2013-08-21.
[7] John Mantas,Arie Hasman,George Anastassopoulos,Adam Adamopoulos,Dimitrios Galiatsatos,Georgios Drosos. Feature Extraction of Osteoporosis Risk Factors using Artificial Neural Networks and Genetic Algorithms[M].IOS Press:2013-06-15.
[8] Hosmer David W.,Lemeshow Stanley,Sturdivant Rodney X. Applied Logistic Regression[M].John Wiley & Sons, Inc.:2013-03-29.
[9] Joseph M. Hilbe. Logistic Regression Models[M].CRC Press:2009-05-11.
[10] Hyeoun-Ae Park,Peter Murray,Connie Delaney,Huey-Ming Guo,Yea-Ing Lotus Shyu,Her-Kun Chang. Combining Logistic Regression with Classification and Regression Tree to Predict Quality of Care in a Home Health Nursing Data Set[M].IOS Press:2006-06-15.
[11] B. Cesnik,A.T. McCray,J.-R. Scherrer,Lucila Ohno-Machado,Donald Bialek. Diagnosing Breast Cancer from FNAs: Variable Relevance in Neural Network and Logistic Regression Models[M].IOS Press:1998-06-15.