Advances in Applied Mathematics 应用数学进展, 2022, 11(5), 2804-2813
Published Online May 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.115297

Hans汉斯

基于多项有序Logistic回归的汽车购买意愿影响因素研究

陈辛如

云南财经大学，云南 昆明

收稿日期：2022年4月23日；录用日期：2022年5月18日；发布日期：2022年5月25日

摘 要

本文以UCI机器学习数据库中的1728条汽车评价数据为研究对象，旨在探究影响消费者汽车购买意愿的因素以及有序Logistic模型在分类预测中的效果。数据集总共包括七个变量，其中消费者汽车购买意愿为因变量，购入费用、维修费用、车门数、座位、内部空间、安全程度为自变量，除车门数和座位外，其余变量均为分类型变量；本文借助于统计软件R，采用多项有序Logistic回归模型进行建模预测后发现：1) 购入费用和维修费用对消费者汽车购买意愿有显著的负向影响，车门数、座位数、内部空间、安全性对消费者汽车购买意愿有显著的正向影响；2) 安全性这个自变量对消费者汽车购买意愿的影响最大，其回归系数值为2.743，同时其优势比(OR值)为15.531，意味着安全性增加一个单位时，购买意愿的变化(增加)幅度为15.531倍；3) 利用构建的多项有序Logistic回归模型对测试集数据(后30%)进行预测时，对整体多项预测准确率达到了0.815，从因变量具体类别来看：Class: 0 (unacc)和Class: 3 (Vgood)的预测准确率最高，分别为0.830和0.829，Class: 2 (good)准确率最差仅为0.499。

关键词

多项有序Logistic回归，汽车购买意愿，影响因素，分类预测

Research on Influencing Factors of Automobile Purchase Intention Based on Multiple Ordered Logistic Regression

Xinru Chen

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Apr. 23rd, 2022; accepted: May 18th, 2022; published: May 25th, 2022

文章引用: 陈辛如. 基于多项有序 Logistic 回归的汽车购买意愿影响因素研究[J]. 应用数学进展, 2022, 11(5): 2804-2813. DOI: 10.12677/aam.2022.115297

Abstract

This paper takes 1728 automobile evaluation data in the UCI machine learning database as the research object, and aims to explore the factors influencing consumers' automobile purchase intention and the effect of the ordered Logistic model in classification prediction. The data set consists of seven variables in total, of which, the purchase intention of consumers is the dependent variable, while the purchase cost, maintenance cost, number of car doors, seats, interior space and safety degree are independent variables. Except for the number of car doors and seats, other variables are sub-type variables. With the help of statistical software R, this paper adopts multiple ordered Logistic regression model for modeling and prediction, and finds that: 1) Purchase cost and maintenance cost have a significant negative impact on consumers' purchase intention, and the number of doors, seats, interior space and safety have a significant positive impact on consumers' purchase intention; 2) The independent variable of safety has the greatest influence on consumers' intention to buy automobiles. Its regression coefficient value is 2.743, and its odds ratio (OR value) is 15.531, which means that when safety increases by one unit, the change (increase) range of purchase intention is 15.531 times. 3) When the constructed multiple ordered Logistic regression model was used to predict the test set data (the last 30%), the overall multiple prediction accuracy reached 0.815. From the specific category of dependent variables: The prediction accuracy of Class: 0 (unacc) and Class: 3 (Vgood) was 0.830 and 0.829, respectively, while the prediction accuracy of Class: 2 (good) was 0.499.

Keywords

Multiple Ordered Logistic Regression, Automobile Purchase Intention, Influencing Factors, Classification Prediction

1. 绪论

1.1. 研究背景和意义

1.1.1. 研究背景

我国的国民经济中有非常多的重要支柱产业，汽车是其中非常重要的产业之一，它承担了推动社会发展的重要责任，对于我国的经济发展同时也有着很强的带动作用。近几年来，我国经济进入稳步递增的阶段，汽车产销迅速发展，国内汽车保有量和销量得到非常大的提升，到 2018 年，我国汽车的生产量和销售量到已是连续 10 年赠联全球第一[1]。由于我国的汽车市场形态发生了非常大的变化，已经赶超国际汽车营销的步伐。当消费者购买汽车可以满足自己的需求，并且意识到汽车带来的利益，同时认为购买汽车是满足需要的理想途径，购买汽车感到满足之后，汽车企业可以获得成功。因此研究汽车购买决策并分析影响因素，能够帮助汽车企业作出明确的营销决策，制定正确的营销策略，同时也能帮助消费者自身作出更明智的购买决策。

1.1.2. 研究意义

本文切合当前社会发展要求，以社会热点汽车为研究对象，运用 UCI 机器学习数据库中的 1728 条

消费者对于汽车购买意愿的评价数据集，通过多类别 Logistic 回归方法建立模型来探究影响消费者对汽车购买意愿的诸多因素[2]，从而了解和掌握汽车潜在消费者的特征，以及得到影响消费者购买意愿的关键因素，另外本文还通过多类别 Logistic 回归方法建立模型对汽车前景进行分析和预测。本文通过分析验证得出消费者对汽车接受度以及影响消费者购买意愿关键因素的结论。从生产厂商以及销售者的角度，让生产厂商和销售者准确了解消费者关键诉求，从而有针对性地提升服务，提高产品质量，更准确得进行市场推广。使企业进一步认识到影响消费者决策行为的因素，为汽车企业今后在更加精确的市场细分中进行汽车设计、生产、定价、营销等活动，提供一定的借鉴。从消费者角度，为消费者提供适当的引导以及为有计划购买汽车的消费者提供参考。

1.2. 文章内容结构简介

本文共分为五个部分，具体内容如下：

第一部分为绪论，主要介绍文章的研究背景，叙述了论文在理论和实践方面的主要研究意义。

第二部分为基本方法与模型介绍，主要叙述了本文所用的主要方法和模型，包括广义线性模型、多类别 Logistic 模型以及其评估方法。

第三部分为数据说明。通过机器学习数据库找到汽车评价数据集，并对数据集中的指标作出说明和解释，为后续建立数学模型做准备。

第四部分为数据分析及实证结果。针对数据集进行实证分析，利用 Spearman 等级相关系数检验汽车的基本特征与消费者购买意愿的相关性，采用多项有序 Logistic 回归模型进行建模，并运用混淆矩阵对模型进行评估。

第五部分为研究结论和局限性。根据实证分析结果，得到影响购消费者买汽车的主要因素，并提出文章存在的不足之处。

2. 理论基础

2.1. 广义线性模型

一般的线性回归模型只能对数据进行线性关系的建模。为了更好得反映数据间的非线性关系，广义线性模型(GLM)得以提出[3]。它运用联结函数建立起输出变量的条件期望与输入变量的线性组合之间的非线性关系。通过变换联结函数的表达式，使得所建立的模型能够映现多种输出变量与输入变量间的非线性关系，这是线性模型在统计实践中发展的产物。假设 $Y_1,Y_2,\cdots,Y_n$ 是输出变量 $Y\in R$ 的 n 个独立观测，均服从指数族分布，即 Y_i 有概率密度函数：

$$f\left(y_i|\theta_i,\varphi\right)=\exp\left\{\frac{y_i\cdot\theta_i-b\left(\theta_i\right)}{\varphi}+c\left(y_i,\varphi\right)\right\},\left(1\le i\le n\right) \tag{1}$$

其中为为 Y_i 的自然参数，参数 φ 为尺度参数，与具体的 i 值无关，函数 $b\left(\theta\right)$ 是 $\theta_i\in R$ 参数 θ 的函数，$c\left(y_i,\varphi\right)$ 是 y_i 和 φ 的函数。$x_1,x_2,\cdots,x_n$ 为对应于 $Y_1,Y_2,\cdots,Y_n$ 的 p 维输入变量 $X\in R^P$ 的 n 个观测值。记 $\eta_i=x_i^{\mathrm{T}}\beta,1\le i\le n$，其中 $\beta\in R^P$ 为未知参数向量。假设 $\mu_i=E\left[Y_i|X=x_i\right]$，并且 μ_i 和 η_i 具有关系：

$$\eta_i=g\left(\mu_i\right),i=1,2,\cdots,n \tag{2}$$

称此定义的模型为广义线性模型。其中函数 $g\left(\bullet\right)$ 称为联结函数。

2.2. Logistic 回归模型

Logistic 回归是一种广义线性回归[4]，与多重线性回归分析有很多相同之处。它们的模型形式基本上

相同。Logistic 回归分析用于研究 X 对 Y 的影响，并且对 X 的数据类型没有要求，X 可以为定类数据，也可以为定量数据，但要求 Y 必须为定类数据，并且根据 Y 的选项数，使用相应的数据分析方法。一般可分为 3 类，分别是二元 Logistic 回归分析、多分类 Logistic 回归分析和有序 Logistic 回归分析。

其中当定性因变量 Y 取两个类别时即为二元 Logistic 回归，二元逻辑斯蒂回归模型的表达式是一个条件概率分布 $P(Y|X)$，这里的 $X \in R^P$ 为输入变量，输出变量 $Y \in R$ 关于 X 的条件分布是伯努利分布，取值为 0 或者 1，我们通过如下的定义具体表达

$$\begin{aligned} p(C_1|x) &= \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1)+P(x|C_2)P(C_2)} \\ p(C_2|x) &= \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)+P(x|C_2)P(C_2)} \end{aligned} \tag{3}$$

当定性因变量 Y 取 k 个类别时，记为 $1,\cdots,k$，因变量 Y 取值于每个类别的概率与一自变量 $x_1,x_2,\cdots,x_p$ 有关，对于样本数据 $(x_{i1},x_{i2},\cdots,x_{ip};y_i)(i=1,2,\cdots,n)$， 多类别 Logistic 回归模型第 i 组样本的因变量 y_i 取第个类别的概率为

$$\frac{\exp(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip})}{1+\exp(\beta_{02}+\beta_{12}x_{i1}+\cdots+\beta_{p2}x_{ip})+\cdots+\exp(\beta_{0k}+\beta_{1k}x_{i1}+\cdots+\beta_{pk}x_{ip})} (i=1,2,\cdots,n;\ j=1,2,\cdots,k) \tag{4}$$

有序 Logistic 回归即定性因变量 Y 有多个选项，并且各选项之间可以对比大小。

2.3. 模型评估方法

2.3.1. 混淆矩阵

根据 Logistic 回归得到的模型其因变量是多元数据，通常在对模型进行评估时可以以混淆矩阵为基础来评价因变量估计结果的准确性。对二元数据，以 0.5 为分类阈值(预测结果小于 0.5，则 default 为 0，否则 default 为 1)，将预测值与训练集中的实际 default 进行比较，得到的混淆矩阵(见表 1)。

Table 1. Confusion matrix
表 1. 混淆矩阵

	Bad customer	good customer
Bad customer	True positive (TP)	False negative (FN)
good customer	False positive (FP)	True negative (TN)

采用预测精度，预测准确率，错误类型一以及错误类型二的方式对模型进行评级。其中：

$$\text{预测精度} = \frac{\text{TN}+\text{TP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}} \tag{5}$$

$$\text{预测准确率} = \frac{\text{TP}}{\text{TR}+\text{FP}} \tag{6}$$

2.3.2. 伪判定系数 R^2

伪判定系数 R^2 用于衡量回归模型相比于默认模型在解释数据时的效果，其值应该小于 1，若值大于 1，则得到的回归模型就不能被信任，数据集就不适合采用逻辑回归算法进行分类预测。

伪判定系数 R^2 定义为

$$R^2 = 1 - \frac{D_m}{D_n} \tag{7}$$

其中 D_n 是空模型的偏差， D_m 是回归模型的偏差。

3. 数据说明

本文数据是 UCI 机器学习数据库的汽车评价数据，数据资源来自于 http://archive.ics.uci.edu/ml/。该数据集总共包括七个变量，其中购买意愿为因变量，剩余六个自变量分别为购入费用，维修费用，内部空间，安全性，座位数；该数据总共包含 1728 条样本记录。本文所使用的数据分析工具主要是 R 语言[5]。首先，描述顾客对汽车购买意愿 Y 包含四种水平，不被接受(unacc)、偶有接受(acc)、较为接受(good)、很受欢迎(Vgood)，按程度由低到高依次记为 0、1、2、3。具体 6 个自变量见表 2。

Table 2. Variable description
表 2. 变量说明

自变量	自变量含义及单位	说明
X_1	buying (购入费用)	1 表示 low；2 表示 med (适中)；3 表示 high；4 表示 vhigh (非常高)
X_2	maint (维修费用)	1 表示 low；2 表示 med (适中)；3 表示 high；4 表示 vhigh (非常高)
X_3	Doors (车门数)	5 表示 more，代表 5 个及以上
X_4	Persons (座位数)	6 表示 more，代表 6 个及以上
X_5	lug_boot (内部空间)	1 表示 small 为；2 表示 med；3 表示 big
X_6	Safety (安全性)	1 表示 low；2 表示 med；3 表示 high

4. 实证分析

4.1. 数据分析

4.1.1. 描述性统计

将汽车评价数据作为基础数据集(见表 3)，总样本数为 1728 个。对于因变量 Y (购买意愿)而言，分类属性为 0 即不被接受的频数为 1210，占比 70.02%；分类属性为 1 即偶有接受的频数为 384，占比 22.22%；分类属性为 2 即较为接受的频数为 69，占比 3.99%；分类属性为 3 即很受欢迎的频数为 65，占比仅为 3.76%，数据分布不太均匀。

Table 3. Frequency distribution of dependent variables analyzed by ordered Logistic regression
表 3. 有序 Logistic 回归分析因变量频数分布

名称	选项	频数	百分比
Y (购买意愿)	0 (unacc)	1210	70.02%
	1 (acc)	384	22.22%
	2 (good)	69	3.99%
	3 (Vgood)	65	3.76%
	总计	1728	100.0

4.1.2. 变量相关性分析

首先对变量间相关关系进行探究，由于较多变量为分类型变量，因此采用 Spearman 相关系数，调用程序包 corrplot 进行分析得到相关系数图(见图 1)，通过图 1 我们可以看到因变量 Y 即购车意愿与 X_1 (购入费用)、X_2 (维修费用)的 Spearman 相关系数分别为−0.24 和−0.21，均为负相关；与其余自变量的 Spearman 相关系数均为正数，同时与 X_6 (安全性)的相关系数最高，达到了 0.47。

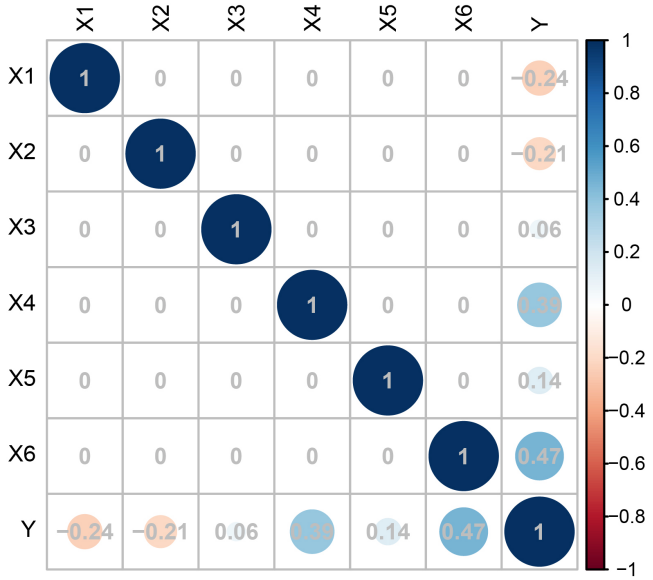

Figure 1. Spearman correlation coefficient diagram of variables
图 1. 变量的 Spearman 相关系数图

另外，回归分析一般假设建模变量之间都是独立不相关的，如果有任何两个变量存在强相关性，则只需保留其中一个而删除其余变量。通过将建模指标中存在明显相关的指标进行筛选剔除，使得所有自变量都不显著相关，避免自变量之间存在很强的共线性而导致模型效果不好。自变量间的 Kappa 系数均小于 100，存在较弱的多重共线性，表明 6 个评级指标都适合作为消费者对汽车接受程度的自变量来建立数学模型。

4.2. 数据建模

4.2.1. 模型似然比检验

从表 4 可知：此处模型检验的原定假设为：是否放入自变量(X_1, X_2, X_3, X_4, X_5, X_6)两种情况时模型质量均一样；分析显示拒绝原假设(chi = 1335.270, p = 0.000 < 0.05)，即说明本次构建模型时，放入的自变量具有有效性，本次模型构建有意义。

Table 4. Likelihood ratio test of ordered Logistic regression model
表 4. 有序 Logistic 回归模型似然比检验

模型	−2 倍对数似然值	卡方值	*df*	*p*	AIC 值	BIC 值
仅截距	2888.373					
最终模型	1553.102	1335.270	6	0.000	1571.102	1620.195

4.2.2. 建模结果

因为数据中的因变量购买意愿为四项有序的分类变量，因此我对数据构建多项有序 logistic 回归进行分析，利用 R 软件中的 MASS 包中的“polr”函数进行建模[6]，代码如下：

model1=polr(as.factor(Class.Values)~buying+maint+doors+persons+lug_boot+safety, method='logistic', Hess=T, data=CARdata)

summary(model1)

由于 R 软件[7]中“polr”函数构建的多项有序 logistic 回归模型没有变量系数的显著性检验项，因此利用如下代码“p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2”计算系数项的 P 值，并利用如下代码“exp(coef(model1))”计算自变量的优势比(OR 值)。

最终分析结果见表 5：

Table 5. Results of multiple ordered Logistic regression
表 5. 多项有序 Logistic 回归结果

自变量	回归系数	标准误差	t 值	P 值
X_1	−1.208	0.076	−15.937	0.00
X_2	−1.019	0.072	−14.158	0.00
X_3	0.276	0.062	4.415	0.00
X_4	1.078	0.058	18.63	0.00
X_5	0.917	0.091	10.101	0.00
X_6	2.743	0.132	20.724	0.00
因变量	回归系数	标准误差	t 值	P 值
0	9.108	0.553	16.468	0.00
1	12.527	0.633	19.78	0.00
2	13.903	0.663	20.959	0.00
AIC 值	1571.102			
McFadden R 方：	0.462			
Cox 和 Snell R^2：	0.538			
Nagelkerke R^2：	0.538			

注：0 代表因变量中不被接受(unacc)；1 代表因变量中偶有接受(acc)；2 代表因变量中较为接受(good)。

Table 6. Odds Ratio of each variable (OR value)
表 6. 各自变量优势比(OR 值)

自变量	OR 值	OR 值 95% CI
X_1	0.299	0.257~0.347
X_2	0.361	0.313~0.416
X_3	1.317	1.166~1.489
X_4	2.94	2.625~3.294
X_5	2.501	2.093~2.987
X_6	15.531	11.983~20.131

从表 5 我们可以看出，模型伪 R 平方值(McFadden R 方)为 0.462，意味着 X_1，X_2，X_3，X_4，X_5，X_6 可以解释 Y 的 46.2%变化原因；以及模型公式如下：

$$\begin{aligned}&\text{logit}\left[P(Y\le 0.0)/(1-P(Y\le 0.0))\right]\\&=9.108-1.208*X_1-1.019*X_2+0.276*X_3+1.078*X_4+0.917*X_5+2.743*X_6\\&\text{logit}\left[P(Y\le 1.0)/(1-P(Y\le 1.0))\right]\\&=12.527-1.208*X_1-1.019*X_2+0.276*X_3+1.078*X_4+0.917*X_5+2.743*X_6\\&\text{logit}\left[P(Y\le 2.0)/(1-P(Y\le 2.0))\right]\\&=13.903-1.208*X_1-1.019*X_2+0.276*X_3+1.078*X_4+0.917*X_5+2.743*X_6\end{aligned}$$

结合表 5 各变量优势比值(见表 6)，我们可以得出，X_1 的回归系数值为−1.208，并且呈现出 0.01 水平的显著性($z=-15.937$, $p=0.000<0.01$)，意味着 X_1 会对 Y 产生显著的负向影响关系。以及优势比(OR 值)为 0.299，意味着 X_1 增加一个单位时，Y 的变化(减少)幅度为 0.299 倍；

X_2 的回归系数值为−1.019，并且呈现出 0.01 水平的显著性($z=-14.158$, $p=0.000<0.01$)，意味着 X_2 会对 Y 产生显著的负向影响关系。以及优势比(OR 值)为 0.361，意味着 X_2 增加一个单位时，Y 的变化(减少)幅度为 0.361 倍；

X_3 的回归系数值为 0.276，并且呈现出 0.01 水平的显著性($z=4.415$, $p=0.000<0.01$)，意味着 X_3 会对 Y 产生显著的正向影响关系。以及优势比(OR 值)为 1.317，意味着 X_3 增加一个单位时，Y 的变化(增加)幅度为 1.317 倍；

X_4 的回归系数值为 1.078，并且呈现出 0.01 水平的显著性($z=18.630$, $p=0.000<0.01$)，意味着 X_4 会对 Y 产生显著的正向影响关系。以及优势比(OR 值)为 2.940，意味着 X_4 增加一个单位时，Y 的变化(增加)幅度为 2.940 倍；

X_5 的回归系数值为 0.917，并且呈现出 0.01 水平的显著性($z=10.101$, $p=0.000<0.01$)，意味着 X_5 会对 Y 产生显著的正向影响关系。以及优势比(OR 值)为 2.501，意味着 X_5 增加一个单位时，Y 的变化(增加)幅度为 2.501 倍；

X_6 的回归系数值为 2.743，并且呈现出 0.01 水平的显著性($z=20.724$, $p=0.000<0.01$)，意味着 X_6 会对 Y 产生显著的正向影响关系。以及优势比(OR 值)为 15.531，意味着 X_6 增加一个单位时，Y 的变化(增加)幅度为 15.531 倍；总结分析可知：X_3，X_4，X_5，X_6 共 4 项会对 Y 产生显著的正向影响关系，以及 X_1，X_2 共 2 项会对 Y 产生显著的负向影响关系。

4.3. 模型预测与评估

将数据集的前 70%的数据划分为训练集，后 30%的数据划分为测试集，利用前面所构建的有序 logistic 模型进行预测，预测代码为“mean(predicted.classes == test_data$Y)”，得到训练集上总的平均预测准确度为 0.815，通过 R 里面的 caret 程序包中 confusionMatrix 函数分别构建混淆矩阵用于模型评估[8]。

我们可以计算得出在混淆矩阵中测试集因变量总的预测准确度和四个类别各自的预测准确度(见表 7)。

测试集总的预测准确度：(344 + 65 + 0 + 14)/519 = 0.815

0 (unacc)的预测准确度：344/(344 + 33) = 0.912

1 (acc)：的预测准确度：65/(36 + 65 + 3) = 0.625

2 (good)：的预测准确度：0/17 = 0.000

3 (Vgood)：的预测准确度：14/(7 + 14) = 0.667

可以得出在整个测试集的预测精度还比较好，但是对于类别 2 (good)由于它的样本数量较少，预测准确度较低。

Table 7. Confusion matrix of ordered Logistic regression model
表 7. 有序 Logistic 回归模型的混淆矩阵

		实际			
		0 (unacc)	1 (acc)	2 (good)	3 (Vgood)
预测	0 (unacc)	344	36	0	0
	1 (acc)	33	65	15	7
	2 (good)	0	1	0	0
	3 (Vgood)	0	2	2	14
准确度	0.815	0.912	0.625	0.000	0.667
95% CI	(0.7789, 0.8475)	/	/	/	/

通过表 8 可以得出，整体的预测准确率为 0.815，95%置信区间为(0.779, 0.848)，同时麦克尼马检验的 p 值为 0 [9]，表明在统计上是显著的，另外 R 软件中 confusion Matrix 函数对于测试集上通过七种统计量给出因变量四种类别的预测准确度加以对比，不同的预测统计量给出的预测准确率有所不同，但是可以发现在类别为 Class: 0 (unacc)和 Class: 3 (Vgood)的各项准确率较高，其余两项类别中准确率较差；最终的平均预测率中，Class: 0 (unacc)和 Class: 3 (Vgood)的预测准确率分别为 0.830 和 0.829，Class: 1 (acc)为 0.746，Class: 2 (good)准确率最差仅为 0.499。

Table 8. Statistical results
表 8. 统计结果

Overall Statistics				
Accuracy:	0.815			
95% CI:	(0.779, 0.848)			
No Information Rate:	0.726			
P-Value [Acc > NIR]:	0.000			
Kappa:	0.56			
Mcnemar's Test P-Value:	0.000			
Statistics by Class:				
	Class: 0	Class: 1	Class: 2	Class: 3
Sensitivity	0.913	0.625	0.000	0.667
Specificity	0.747	0.868	0.998	0.992
PosPred Value	0.905	0.542	0.000	0.778
NegPred Value	0.763	0.902	0.967	0.986
Prevalence	0.726	0.200	0.033	0.040
Detection Rate	0.663	0.125	0.000	0.027
Detection Prevalence	0.732	0.231	0.002	0.035
Balanced Accuracy	0.830	0.746	0.499	0.829

5. 研究结论与局限性

5.1. 研究结论

通过构建四分类有序 Logistic 回归模型进行建模预测后发现：首先购入费用和维修费用对消费者汽车购买意愿有显著的负向影响，车门数、座位数、内部空间、安全性对消费者汽车购买意愿有显著的正向影响；购入费用的优势比(OR 值)为 0.299，意味着它增加一个单位时，购买意愿的变化(减少)幅度为 0.299 倍；维修费用优势比(OR 值)为 0.361，意味着它增加一个单位时，购买意愿的变化(减少)幅度为 0.361 倍。而安全性这个自变量对消费者汽车购买意愿的影响最大，同时其优势比(OR 值)为 15.531，意味着安全性增加一个单位时，购买意愿的变化(增加)幅度为 15.531 倍，这表明样本数据中消费者更加注重安全性，其实车企制造商要注重汽车安全性能，同时可以提高购车优惠；其次，利用构建的有序 Logistic 回归模型对测试集数据(后 30%)进行预测时，对整体预测准确率达到了 0.815，在现实生活中，在对具体类别进行预测时部分效果不尽人意。

本文通过构建多项有序 Logistic 回归进行研究，研究探究了与汽车直接相关的影响因素对于购买决策的重要性，证实了汽车安全性成为消费者在进行家用汽车购买决策中的重要的衡量指标。这就要求对于汽车生产企业在汽车制造上面需要更加注重安全性。

5.2. 研究的局限性

本文存在如下几个不足之处：一、由于数据的限制，在对购车意愿的影响因素进行探究时只挑选了数据集中仅有的六个变量，在实际生活中购车意愿可能受其他重要因素影响，因此可能遗漏了重要变量；二、数据集的数量仅有一千余条，数据量偏少，在划分测试集后，部分类别样本数仅为十几条，这可能是某些类别最后的预测准确度不够理想的部分原因；三、在做分类预测时更好的做法是采用几种不同的分类模型加以对比，以看出有序 Logistic 回归模型在本次分类中是否为最优选择。

参考文献

[1] 徐国虎, 许芳. 新能源汽车购买决策的影响因素研究[J]. 中国人口资源与环境, 2010, 20(11): 91-95.

[2] 葛君. 基于 Logistic 模型的信用卡信用风险研究[J]. 中国信用卡, 2010(24): 26-32.

[3] 李星星. 广义线性模型的若干估计及比较[D]: [硕士学位论文]. 扬州: 扬州大学, 2017.

[4] 赵红. Logistic 曲线参数估计方法及应用研究[D]: [硕士学位论文]. 长春: 吉林农业大学, 2015.

[5] 石永东, 胡树华. 汽车购买行为模型及其评价[J]. 汽车工业研究, 2003(2): 7-10.

[6] 李倩星. R 语言与大数据编程实战[M]. 北京: 电子工业出版社, 2017: 230-250.

[7] 张良均, 云伟标, 王路. R 语言数据分析与挖掘实战[M]. 北京: 机械工业出版社, 2015: 66-88.

[8] 许可. 关于在我国推行绿色汽车保险的可行性分析——基于调研数据的 Logistic 模型研究[J]. 应用概率统计, 2012(4): 334.

[9] Yee, T.W. (2015) Vector Generalized Linear and Additive Models (in Preparation). Springer, New York. https://doi.org/10.1007/978-1-4939-2818-7