Advances in Applied Mathematics 应用数学进展, 2017, 6(9), 1236-1247
Published Online December 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.69149

Hans汉斯

Modeling and Prediction of Cross-Selling Problems in Big Data

Xingfang Huang[1], Xuelian Wang[2]

[1]Institute of Statistics and Data Science, Nanjing Audit University, Nanjing Jiangsu
[2]School of Mathematics, Southeast University, Nanjing Jiangsu
Email: xfhuang@nau.edu.cn, xuelianwang1991@163.com

Received: Dec. 1st, 2017; accepted: Dec. 22nd, 2017; published: Dec. 29th, 2017

Abstract

Multiple Logistic method and Two-stage Logistic method all have good advantages of dealing with large number of variables and big data. The main purpose for this paper is building a cross-selling model from Auto Insurance to Home Insurance and then predicting the customers' behavior. A famous American insurance company's cross-selling data in eleven months is used in this paper. Multiple Logistic method and Two-stage Logistic method are separately applied to build cross-selling models on California and Non-California area. The results for these models can predict which products the prospects are more likely to purchase. Finally, it makes a conclusion that Two-stage Logistic model performs better on both California and Non-California data.

Keywords

Cross-Selling, Multiple Logistic Model, Two-Stage Logistic Model

大数据下交叉销售问题的建模与预测

黄性芳[1]，王雪莲[2]

[1]南京审计大学统计科学与大数据研究院，江苏 南京
[2]东南大学数学学院，江苏 南京
Email: xfhuang@nau.edu.cn, xuelianwang1991@163.com

收稿日期：2017年12月1日；录用日期：2017年12月22日；发布日期：2017年12月29日

摘 要

多重Logistic回归和两阶段Logistic回归方法在处理多变量和大数据问题中具有较好的优越性。本文针对

文章引用: 黄性芳, 王雪莲. 大数据下交叉销售问题的建模与预测[J]. 应用数学进展, 2017, 6(9): 1236-1247.
DOI: 10.12677/aam.2017.69149

美国某保险公司一个保险项目11个月的交叉销售数据，分别在加利福尼亚州和非加利福尼亚州两个区域采用多重Logistic回归和两阶段Logistic回归方法，建立了由车险到家庭险的交叉销售模型，预测购买两种不同家庭险的潜在客户。实践表明，两阶段Logistic模型预测效果更佳。

关键词

交叉销售，多重Logistic模型，两阶段Logistic模型

1. 引言

“交叉销售”一词最早在 1965 年被国外银行业普遍使用[1]，35 年以后，交叉销售的理论和实践也得到了大规模的研究。Nash (1993) [2]和 Deighton 等学者(1994) [3]指出，交叉销售是指“鼓励一个已经购买了某公司 A 产品的顾客购买其 B 产品”。郭国庆(2003) [4]认为交叉销售的实质是：充分利用一切资源，服务市场、开展营销、赢得用户。其后，赵华、宋顺林(2007) [5]基于 ERMSW 算法，对已有的客户购买序列进行有维度约束，探究客户的消费趋势，预测匹配度满足一定条件的客户可能的购买行为。Li Chunqing 等(2010) [6]针对我国银行的实际情况，对 NPTB 模型的变量进行修正，并通过实证分析，论证了神经网络模型在交叉销售预测中的优越性。

目前国内外对交叉销售实证方面的研究，主要以识别交叉销售机会的方法为主，已有的方法有：潜在特质模型、获得模式、市场细分法和 NPTB 模型等，但是这些方法仅仅从各自的模型和理论出发，识别交叉销售机会，因此众说纷纭。特别是对于大数据，缺少相应的交叉销售实例研究。

本文以美国某知名保险公司(下记 A 公司)一个保险项目为例，交叉销售数据来自 2012 年 12 月到 2013 年 10 月共约 1.2 亿记录、230 个变量。将其分为加利福尼亚州(下记 CA)和非加利福尼亚州(下记 NCA)两个区域销售数据，分别建立交叉销售模型，从而在已成功购买车险的人群中预测客户购买两种不同家庭险的概率，一种是面向房屋拥有者的保险，一种是面向租客的保险，并对比两种方法优劣[7]。

该数据集为月数据，每个都有 230 个变量，约 1200 万条记录，其中性别、年龄、收入、住房所有权、居住类型、婚姻状况、教育水平、种族、子女等变量属于人口统计学数据；购买类型、客户服务获得、付款日期与数量、保险索赔或破产行为等变量属于行为数据；对风险的态度等属于心理数据[8]，详细数据说明参见王雪莲(2015) [7]。该数据集中的行为数据一般比其他类型数据在预测未来行为时效果好，但获取价格比较昂贵；人口统计学数据容易获取且比较稳定，可用于特征分析或预测；心理数据能提高模型的预测能力，但不易获取[8]。由于美国保险行业中 CA 区域保单政策明显不同于其它州，故本文模型中也将数据分为 CA 和 NCA 两个大区域分别建模。

本例中交叉销售成功是指已经购买车险的客户满 60 天后再次购买家庭险，其它情况则不属于交叉销售成功。以 60 天作为临界值来界定该保单是否属于交叉销售成功，主要有以下原因：(1) 经验值 60 选出的客户群购买行为最稳定；(2) 一份保单满 60 天后换到别家保险公司，能获得更多优惠政策。

建立交叉销售模型的预期目的为：(1) 建模过程首先对数据进行变量筛选，尽可能了解数据的来源与质量，有利于指导 A 公司获取数据的方向；(2) 有效探究易购买目标产品的潜在客户特征，便于对潜在客户进行相应的商业营销，进而提高销售利润、客户满意度和忠诚度等。

2. Logistic 模型介绍

在社会科学诸如人口学、心理学、社会学、经济学以及公共卫生学当中，大量的观测因变量是属性变量。Logistic 模型是研究二分类变量与其他影响因素之间关系的一种统计分析方法。

假设某一事件发生的概率为 P，其中 $0 \le P \le 1$，自变量与 P 之间的关系可用下式表示：

$$\ln\left(\frac{P_i}{1-P_i}\right)=\alpha+\beta x_i$$

将自变量与概率 P 之间的关系通过对数变换转为线性函数进行估计，因此可以使用线性回归的方法进行估计。当自变量个数为 k 时，上式可扩展为

$$\ln\left(\frac{P_i}{1-P_i}\right)=\alpha+\sum_{i=1}^{N}\beta_k x_{ki}$$

(一) 多重 Logistic 模型

当因变量有 k 个类别($k \ge 2$)，此时多重 Logistic 模型第 i 组样本因变量 y_i 取第 j 个类别的概率为：

$$\pi_{ij}=\frac{\exp\left(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip}\right)}{\exp\left(\beta_{01}+\beta_{11}x_{i1}+\cdots+\beta_{p1}x_{ip}\right)+\cdots+\exp\left(\beta_{0k}+\beta_{1k}x_{i1}+\cdots+\beta_{pk}x_{ip}\right)},$$
$$\left(i=1,2,\cdots,n,j=1,2,\cdots,k\right)$$

易知，上式中各回归系数同时加减一个常数后，π_{ij} 数值保持不变。不失一般性，我们把分母的第一项 $\exp\left(\beta_{01}+\beta_{11}x_{i1}+\cdots+\beta_{p1}x_{ip}\right)$ 中的系数取为 0，称为参照系数，得到新的回归函数的表达式：

$$\pi_{ij}=\frac{\exp\left(\beta_{0j}+\beta_{1j}x_{i1}+\cdots+\beta_{pj}x_{ip}\right)}{1+\exp\left(\beta_{02}+\beta_{12}x_{i1}+\cdots+\beta_{p2}x_{ip}\right)+\cdots+\exp\left(\beta_{0k}+\beta_{1k}x_{i1}+\cdots+\beta_{pk}x_{ip}\right)}$$

若有 n 个因变量，取其中一类作基本类(组)，其它 $n-1$ 类与该类分别建 $n-1$ 个 Logistic 模型，通过变量系数及其符号等指标来评判模型结果的优劣。

(二) 两阶段 Logistic 模型

首先将因变量的类别区分成两个大类进行 Logistic 建模，然后把模型结果中目标类数据细分成 2 个小类别再进行一次 Logistic 建模，最终得到两阶段 Logistic 模型。在本文中两阶段 Logistic 模型的第一阶段，以交叉销售家庭险的人群为目标变量，销售成功记为 1，其它为 0，即可建立第一个 Logistic 模型；第二阶段，定义购买房屋拥有者险的人群为记为 1，购买租客险的人群记为 0，由此得到第二个 Logistic 模型。

3. 模型的建立与结果分析

(一) 建模要求

建模的过程也是变量筛选的过程。根据一般要求，入选的变量模型响应率应大于 5%。其次，根据变量实际意义及保险行业模型要求不断调试模型：对意思相近变量，模型里只保留一个；与因变量相关性特别强也需要删掉；贡献率越大表示该变量越能将购买人群区分出来，贡献率太大也不好，通常最终会选入 7~20 个变量进行建模。对于变量选择更详细的要求参见文献王雪莲(2015) [7]。

(二) 评价模型的指标

对于模型的评价指标主要有基尼系数、C 值和整个模型分组得分，主要目的是筛选出更有可能购买交叉保险的客户群。

(1) 基尼系数

经济学上，基尼系数越大，表示潜在客户特征差距越大，重点对收入更高或经济实力更强的客户发营销广告，响应的人也会越多，这样更容易地找出更有可能购买保单的客户。因此，模型的基尼系数越大越好。

根据对整个模型打分结果画出相应的洛伦茨曲线。其中，洛伦兹曲线是在一个总体(国家、地区)内，以“最贫穷的人口计算起一直到最富有人口”的人口百分比对应各个人口百分比的收入百分比的点组成的曲线。这里穷富靠分数高低来评判，如图 1 所示。

其中 S_A 表平均直线 OT 与洛伦茨曲线围成的面积，S_A+S_B 为绝对平均直线以下直角三角形 OPT 的面积，基尼系数的计算公式如下：$Gini=S_A/(S_A+S_B)$。

$Gini$ 的值在 0 与 1 之间，即 $0\le Gini\le 1$。计算基尼系数的方法有很多，如直接计算法、切块法、函数法、弓形面积法等。本文使用切块法。根据理论知识，先画出该方法基尼系数计算示意图，如图 2 所

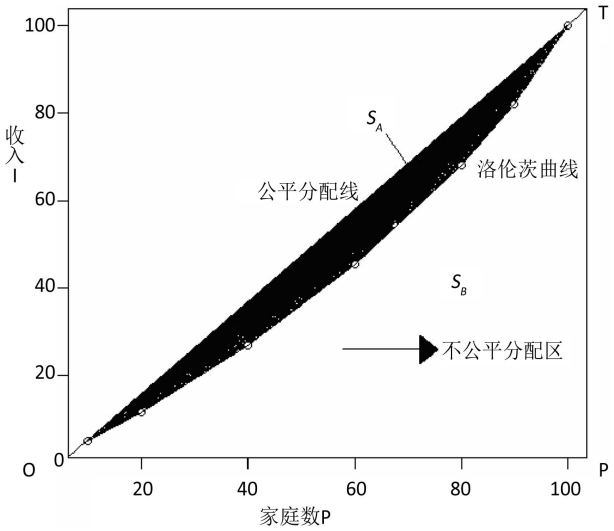

Figure 1. Gini score graph
图 1. Gini 打分分布图

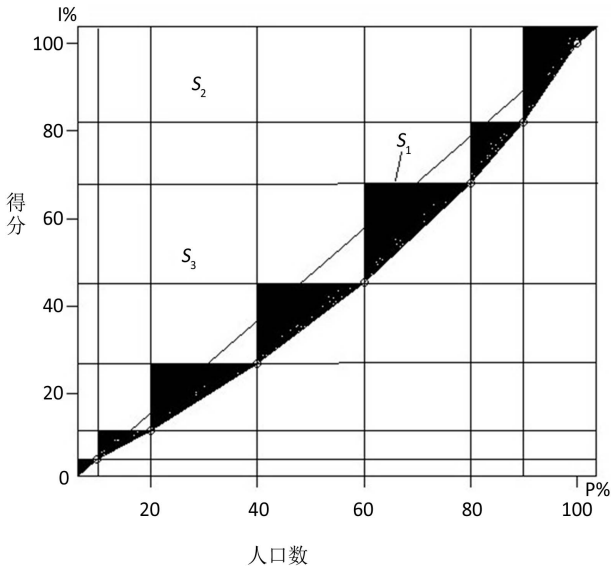

Figure 2. Gini coefficient obtained by jackknife method
图 2. 刀切法算基尼系数

示：设 P_i 为该记录对应的预测概率， I_i 为每条记录所占比例，即 $I_i = 1/n$ ，则基尼系数公式中 S_A 的面积可表述为下列三部分的代数和，其中 S_1 近似于图中所涂阴影部分面积和：

$$S_1 = \frac{1}{2}\left(P_1 I_1 + P_2 I_2 + \cdots + P_n I_n\right)$$

S_2 为洛伦茨曲线以上的面积中除去 S_1 的阴影面积的部分：

$$S_2 = P_1\left(I_2 + I_3 + \cdots + I_n\right) + P_2\left(I_3 + I_4 + \cdots + I_n\right) + P_{n-1} I_n$$

S_3 为单位正方形面积的一半， $S_3 = 1/2$ 。由于 $S_A = S_1 + S_2 - S_3$ ， $S_A + S_B = 1/2$ ， Q_i 为前 i 条记录的累积比例，故基尼系数基本公式为：

$$G = \sum_{i=1}^{n} I_i P_i + 2\left[P_1\left(1 - Q_1\right) + P_2\left(1 - Q_2\right) + \cdots + P_{n-1}\left(1 - Q_{n-1}\right)\right] - 1$$

刀切法算基尼系数的优缺点分别为：

优点：洛伦茨曲线的曲线越平缓，即曲率越小，越接近对角线，估计的精度越高。

缺点：在计算过程中省略了弓形面积，所以其结果较真实基尼系数值偏小。

(2) C 值：一个衡量 Logistic 模型预测准确程度的统计值。

比如有一个数据，共 10,000 条记录。其中 1000 条表响应的人(目标变量 = 1)，另外 9000 条是不响应的人(目标变量 = 0)。通过预测这 10,000 个人每个人响应的概率，得到所有预测正确的记录所占的百分比。

(3) 整个模型分组得分

调试模型完成后，将每组数据值代入模型，可估计出 y 值，称为得分数值。将这些得分按降序进行排序，将其均分为若干组，得到所在组的秩。

为了避免模型过拟合现象，我们采用交叉验证方法进行建模。首先用训练集建立模型，得到参数估计和模型表达式，例如 $y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b$ ，这里的 y 为属性变量，表示买或者不买该产品， x 是自变量。用此模型得到测试集的 y 值，从高到低排列分成 10 组，得到每一个 y 的秩。理想的模型是这些秩依次递减、不跳跃，而且第一组指标值越大越好。最后，根据经验取一个得分分割点作为临界值，对分值大于临界值对应的潜在客户进行营销，其他不营销。

综合评价变量和模型的统计量及其作用，总结如表 1 所示。

(三) 多重 Logistic 模型

多重 Logistic 模型的因变量有 3 个类别，定义购买家庭险下的面向房屋拥有者的人群为目标变量 1，定义购买家庭险下的面向租客保险的人群为 2，其它人群为 3。回归函数以因变量 3 作为基准，回归系数取做 0。对其它两类别，每个类别都和其它人群一起建立一个 Logistic 函数，因此每个自变量都有两个回归系数，自由度为 2。CA 上的模型结果见表 2。

其中，响应模式 1 是房屋拥有者和其它人群对比建立的模型，模式 2 是租客与其它人群建立的模型。只有两对变量符号不同，其它变量符号都一样，而且这两对符号不同的变量在模型中的贡献率之和低于 10%。从表中结果来看，多重 Logistic 模型存在一些局限性。以“花在车险上的钱”这个变量为例说明，在模式 1 中，它为正号，表明资金较多的客户更容易购买面向房屋拥有者的保险，而在模式 2 中，它也是正号，表示资金越多客户越容易购买面向租客的保险，这样的结果说明同一个潜在客户资金越多，两种家庭险都容易购买，这与实际情况不吻合。由此说明从变量角度，很难区分这两部分人群。经过实证分析，发现不管是 CA 还是 NCA，多重 Logistic 模型都存在以下问题：

1、选出的变量一样，且系数相差不大。

2、相应的变量符号一样，即使有符号不同的，但它们在整个模型里贡献率之和还未达到 10%。

Table 1. Evaluation and effect of statistic for Variables and models
表 1. 评价变量和模型的统计量及其作用

统计量	趋势	模型拟合	作用
P 值	越小	越好	拒绝 H0 所犯第一类错误的概率
卡方值	越大	越好	对因变量的解释能力
AIC	越小	越好	度量类残差平方和
SC	越小	越好	修正的 AIC
基尼系数	越大	越好	评价人群贫富差距
C 值	越大	越好	度量观测值和预测概率个体的一致性
模型打分	越大	越好	评价模型潜在人群特征

Table 2. Results of multiple Logistic model for California
表 2. 加利福尼亚州多重 Logistic 模型结果

	响应模型 1		响应模型 2		数据类型
变量	系数	贡献率	系数	贡献率	
截距项	−5.9396		−4.4353		
保单在超过 60 天时取消了	1.0356	26.6%	0.9962	24.3%	B
对与 A 公司关系密切团体的保单打折	0.6529	23.8%	0.4787	13.3%	A
该车险已经投保时间	−0.2485	17.3%	−0.1982	12.7%	A
双保险每人限额	0.2751	14.7%	0.1697	6.0%	A
最新车的车龄	0.1866	5.6%	−0.1306	3.2%	A
****汽车数量	0.3690	5.4%	−0.1737	1.3%	A
防锁制动系统	0.3119	2.5%	0.1851	1.2%	A
最大客户的年龄	−0.0059	1.9%	−0.0214	26.3%	A
****倾向于每月支付两次险费	0.1839	1.9%	−0.3820	7.0%	B
花在车险上的钱	0.0581	0.4%	0.1829	4.9%	A

综合以上原因，可认为多重 Logistic 模型并不能有效地区分出两种家庭险的客户，即不能有效地达到预期效果，于是我们尝试使用两阶段 Logistic 模型。

(四) 两阶段 Logistic 模型

两阶段 Logistic 模型中，对于其目标变量的定义如下：第一阶段将成功交叉销售家庭险的人群定义为 1，其它为 0。第二阶段将购买家庭险下的面向房屋拥有者保险的人群定义为目标变量 1，购买家庭险下的面向租客保险的人群定义为 0。

美国住房情况具有以下特点，大多数美国人对拥有自己房屋的美梦并不难实现，例如不需全部现款购屋的交易方式，已大大提高一般人民的购买能力，再加上一些减税的政策，使一般房屋拥有者的负担得到减轻，所以说在美国买房比住房划算[9]。但是在美国住房由工作单位或政府分配的情况很少，所以住出租房也是相当普遍的现象。另一方面，经济条件好的家庭一般会优选买房自住，经济条件不好的家庭会优选租房住，且在城市里中租房的人较多，而农村住房拥有者更多。对于美国车险费率制度，其基本信息见表 3。

了解这些基本情况后，本文分别在 CA 和 NCA 建立两阶段 Logistic 模型，各变量、系数及其模型贡献率结果见表 4 和表 6。

Table 3. Basic information about the U.S. auto insurance rate system
表 3. 美国车险费率制度基本信息

人群特征	车险保险费
投保人驾龄越大	越低
教育程度越高	越低
居住在大都市	越低
汽车型号越高端	越高
汽车数量越多	越低
婚姻状况	已婚人士相对低
年龄结构	中年人最低、青年人最高

Table 4. Results of two-stage Logistic model for California
表 4. 加利福尼亚州两阶段 Logistic 模型结果

变量	系数	贡献率
截距项	−4.3145	
更倾向于按月付款	−0.9755	50.4%
最大驾驶员的车龄	0.8962	16.1%
汽车数目	0.5581	12.7%
居住在高家庭收入地区	−0.0014	8.6%
过去一年索赔额	−0.1829	5.8%
双保险每次事故限额	0.1184	3.0%
最新车的车龄	−0.0208	2.4%
拥有宝马、奔驰等高端车	0.1188	0.5%
客户占有率	−0.1647	0.4%

从表 4 结果知，以上变量及其在模型里的符号，与实际情况相符，且第一个变量的贡献率约 50%，对变量一一作如下分析：(1) 倾向于按月付保额的人，说明其不稳定性较大，相对来说是房屋拥有者的可能性较小，可认为是租客；(2) 一个家庭里年龄较大的驾驶员，越老越有可能是房屋拥有者，因老年人更倾向于住在自己的房子里；(3) 家庭拥有的汽车数目越多，说明这个家庭有较好的经济状况，则越不可能是租客；(4) 居住地区的平均家庭收入越高，一般为城市，因而租客居多，较不可能购买保单；(5) 过去一年索赔额越多，越有可能是租客。(6) 每次双保险事故限额越高，说明该客户经济收入越稳定，越有可能是房屋拥有者；(7) 最新的车越老，则越不可能是房屋拥有者；(8) 拥有保时捷、宝马 Mini 汽车、斯巴鲁、沃尔沃、奥迪、雷克萨斯、奔驰、宝马和捷豹等高端车的人越有可能是房屋拥有者；(9) 家庭居住在客户保险公司的客户占有率越大的地区，越有可能是房屋拥有者。

以上显示为模型变量的信息，表 5 给出了模型各项评价指标值。

其中 C 值处在保险行业规定的 0.6~0.8 之间，基尼系数也相对较高。此时总模型表示为：

$$Y = -4.3145 - 0.9755 * X_1 + 0.8962 * \ln(X_2) + 0.5581 * \ln(X_3) - 0.0014 * X_4$$
$$- 0.1829 * X_5 + 0.1184 * \ln(X_6) - 0.0208 * X_7 + 0.1188 * X_8 - 0.1647 * \ln(X_9)$$

上式中 X_1 表示更倾向于按月付款，X_2 表示最大驾驶员年龄，X_3 表示汽车数目，X_4 表示居住在高收入

家庭区，X_5 表示过去一年索赔额，X_6 表示双保险每次事故限额，X_7 表示最新车的车龄，X_8 表示拥有保时捷、奔驰、宝马等高端车，X_9 表示客户占有率。

通过与 A 公司负责人深入探讨，公司反馈回来的信息显示这部分人确实更容易发生购买行为。

非加利福尼亚州的各变量，系数及其模型贡献率如表 6 所示。同样，模型里的变量及其符号都符合实际，且第一个变量的贡献率明显未超过 60%，对变量一一作如下分析：(1) 家庭里驾驶员数目越多，越有可能是房屋拥有者；(2) 信用评价越差，越不可能是房屋拥有者；(3) 倾向于按月付保险款的人，说明他不稳定性较大，越不可能是房屋拥有者，这与 CA 情况相同；(4) 一个家庭里最年轻驾驶员的年龄越大，说明财产收入等越稳定，越可能是房屋拥有者；(5) 最新的车越老，说明该家庭越没有钱，则越不可能是房屋拥有者，这也与 CA 情况相同；(6) 每次双保险事故限额越多，说明该客户的经济收入越稳定，越有可能是房屋拥有者，这也与 CA 情况相同；(7) 人口密度越大的地区，越有可能是城市，那么生活在该地区的人越有可能是租客；(8) 保险公司为了吸引更多的稳定客户，会对经济能力强的客户在他进行保单转移时有折扣，那么这些人也更有可能是房屋拥有者；(9) 对与保险公司关系密切的团体进行保单打折，这些人也更有可能是房屋拥有者。

以上显示为模型变量的信息，模型的各项评价指标值见表 7。

Table 5. Model evaluation index for California
表 5. 加利福尼亚州模型评价指标

基尼系数	0.394
C 值	0.697
首组打分值	323

Table 6. Results of two-stage Logistic model for non-California
表 6. 非加利福尼亚州两阶段 Logistic 模型结果

变量	系数	贡献率
截距项	−2.2996	
驾驶员数目	1.1104	23.9%
信用评分	−0.0611	23.7%
更倾向于按月付款	−0.7215	15.0%
最年轻驾驶员的年龄	0.0187	9.6%
最新车的车龄	−0.0499	9.5%
双保险每次事故限额	0.2835	9.4%
人口密度	−0.1031	4.1%
保单转移有折扣	0.3361	4.0%
对与 A 公司关系密切团体的保单打折	0.2484	0.8%

Table 7. Model evaluation index
表 7. 非加利福尼亚州模型评价指标

基尼系数	0.558
C 值	0.779
首组打分值	399

其中 C 值处在保险行业规定的 0.6~0.8 之间，基尼系数也相对较高。总模型为：

$$Y=-2.2996+1.1104*\ln(X_1)-0.0611*X_2-0.7215*X_3+0.0187*X_4$$
$$-0.0499*X_5+0.2835*\ln(X_6)-0.1031*\ln(X_7)+0.3361*X_8+0.2484*X_9$$

上式中 X_1 表示驾驶员数，X_2 表示信用评分，X_3 表示更倾向于按月付款，X_4 表示最年轻驾驶员的年龄，X_5 表示最新车的车龄，X_6 表示双保险每次事故限额，X_7 表示人口密度，X_8 表示保单转移有折扣，X_9 表示对与保险公司关系密切的团体进行打折。

同理，实际结果也显示这部分人更容易发生购买行为。

4. 方法分析比较

从上面的结果可知，多重 Logistic 模型结果无法合理地解释选出的变量及其符号意义，而两阶段 Logistic 模型能够与实际情况相符。多重 Logistic 模型在加利福尼亚州的得分情况见表 8，两阶段 Logistic 模型在加利福尼亚州的得分情况见表 9，两种模型的累计交叉销售成功数对比见图 3。多重 Logistic 模型在非加利福尼亚州的得分情况见表 10，两阶段 Logistic 模型在非加利福尼亚州的得分情况见表 11，两种模型累计交叉销售成功数对比见图 4。

(1) 加利福尼亚州。见表 8，表 9，图 3。

(2) 非加利福尼亚州。见表 10，表 11，图 4。

对比上面两种建模方法，在房屋拥有者保险方面，可发现两阶段 Logistic 模型在各个组别上的累积得分都高于多重 Logistic 模型的累积得分，其中第一组别累积得分为 323，高于多重 Logistic 模型第一组别的累积得分 312，又因为两阶段 Logistic 模型的各个变量及其符号都与实际情况相符合，所以两阶段 Logistic 模型在 CA 的预测效果更好。

5. 结论与展望

对比上面两种建模方法，对房屋拥有者保险而言，在各个组别上的累积成功交叉销售指标值，其中

Table 8. Scores of multiple Logistic model for California
表 8. 多重 Logistic 模型在加利福尼亚州的得分表

多重 Logistic 模型								
等级	人数	交叉销售成功	房屋拥有者	租客	房屋拥有者比例	房屋拥有者指标	累计房屋拥有者比例	累计房屋拥有者指标
1	27,626	722	354	368	1.28%	312	1.28%	312
2	27,626	408	199	209	0.72%	175	1.00%	244
3	27,626	328	144	184	0.52%	127	0.84%	205
4	27,626	270	114	155	0.41%	101	0.73%	179
5	27,626	208	83	125	0.30%	73	0.65%	158
6	27,626	182	80	102	0.29%	71	0.59%	143
7	27,626	145	62	83	0.23%	55	0.54%	131
8	27,626	106	46	60	0.17%	40	0.49%	119
9	27,626	81	32	49	0.12%	28	0.45%	109
10	27,626	40	19	21	0.07%	17	0.41%	100
总数	276,260	2490	1133	1356	4.11%	-	6.98%	-

Table 9. Scores of two-stage Logistic model for California
表 9. 两阶段 Logistic 模型在加利福尼亚州的得分表

两阶段 Logistic 模型								
等级	人数	交叉销售成功	房屋拥有者	租客	房屋拥有者比例	房屋拥有者指标	累计房屋拥有者比例	累计房屋拥有者指标
1	27,626	627	366	261	1.32%	323	1.32%	323
2	27,626	382	198	184	0.72%	175	1.02%	249
3	27,626	316	146	171	0.53%	128	0.86%	209
4	27,626	264	116	148	0.42%	102	0.75%	182
5	27,626	224	92	133	0.33%	81	0.66%	162
6	27,626	191	73	118	0.26%	64	0.60%	146
7	27,626	158	53	105	0.19%	46	0.54%	131
8	27,626	139	41	98	0.15%	36	0.49%	119
9	27,626	112	33	79	0.12%	29	0.45%	109
10	27,626	76	17	59	0.06%	15	0.41%	100
总数	276,260	2489	1135	1356	4.10%	-	7.10%	-

Table 10. Scores of multiple Logistic model for non-California
表 10. 多重 Logistic 模型在非加利福尼亚州的得分表

多重 Logistic 模型								
等级	人数	交叉销售成功	房屋拥有者	租客	房屋拥有者比例	房屋拥有者指标	累计房屋拥有者比例	累计房屋拥有者指标
1	98,176	1951	1155	797	1.18%	351	1.18%	351
2	98,176	1125	511	614	0.52%	155	0.85%	253
3	98,176	871	381	490	0.39%	116	0.69%	207
4	98,176	677	306	371	0.31%	93	0.60%	179
5	98,176	560	253	306	0.26%	77	0.53%	158
6	98,176	458	210	248	0.21%	64	0.48%	143
7	98,176	354	165	189	0.17%	50	0.43%	129
8	98,176	292	139	153	0.14%	42	0.40%	119
9	98,176	224	113	111	0.12%	34	0.37%	109
10	98,176	117	56	61	0.06%	17	0.34%	100
总数	981,760	6629	3289	3340	3.36%	-	5.87%	-

Table 11. Scores of two-stage Logistic model for non-California
表 11. 两阶段 Logistic 模型在非加利福尼亚州的得分表

			两阶段 Logistic 模型					
等级	人数	交叉销售成功	房屋拥有者	租客	房屋拥有者比例	房屋拥有者指标	累计房屋拥有者比例	累计房屋拥有者指标
1	98,176	1834	1314	520	1.34%	399	1.34%	351
2	98,176	949	536	413	0.55%	163	0.94%	253
3	98,176	749	358	391	0.36%	109	0.75%	207
4	98,176	657	298	359	0.30%	90	0.64%	179
5	98,176	554	224	331	0.23%	68	0.56%	158
6	98,176	492	177	315	0.18%	54	0.49%	143
7	98,176	447	150	297	0.15%	46	0.44%	129
8	98,176	377	117	259	0.12%	36	0.40%	119
9	98,176	340	78	262	0.08%	24	0.37%	109
10	98,176	231	38	194	0.04%	11	0.34%	100
总数	981,760	6630	3290	3341	3.35%	-	6.27%	-

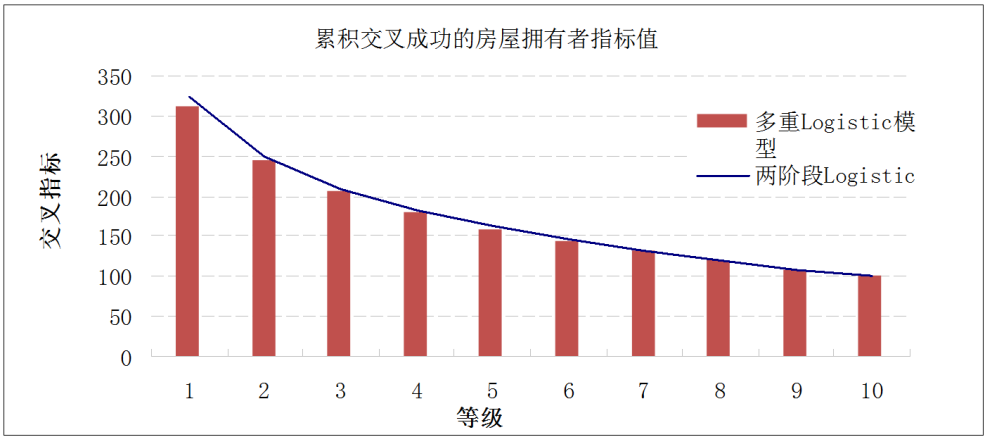

Figure 3. Cumulative score comparison of two models for California
图 3. 加利福尼亚州两种模型累积得分对比图

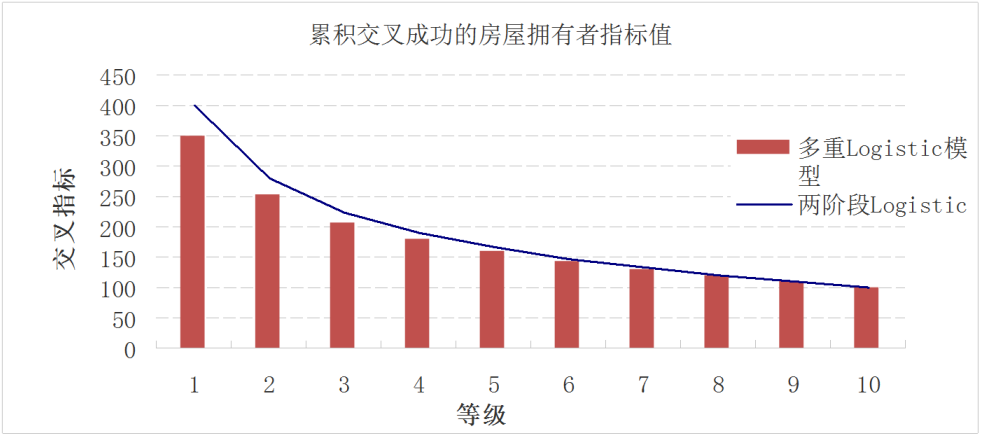

Figure 4. Cumulative score comparison of two models for non-California
图 4. 非加利福尼亚州两种模型累积得分对比图

第一组别累积得分为 399，明显高于第一组别的多重 Logistic 模型累积得分 351，可发现两阶段 Logistic 模型在各个组别上的累积得分也都高于多重 Logistic 模型的累积得分，又因为两阶段 Logistic 模型的各个变量及其符号也都与实际情况相符合，所以两阶段 Logistic 模型在 NCA 的预测效果也是更好。再次印证了两阶段模型表现更优的结论。

综上所述，在加利福尼亚州和非加利福尼亚州，不管是从对模型变量及其符号的解释上，还是从模型得分的对比上，两阶段 Logistic 模型预测效果均优于多重 Logistic 模型。通过本文的努力，模型选出的变量更有利于指导 A 公司获取数据的方向，有助于探究易购买目标保险的潜在客户特征，能够达到交叉销售模型的预期目标。

致　谢

本工作受到国家自然科学基金和教育部人文社科基金青年项目资助。感谢匿名审稿人提出的宝贵意见，对论文的修缮工作起到重要作用。感谢美库尔商务咨询(南京)有限公司、南京审计大学统计科学与大数据研究院提供数据及编程环境、多次参与论文的讨论并给出专业意见。

基金项目

本文获得国家自然科学基金 11401094、11571073，教育部人文社科基金 13YJC910006 及江苏省高校优势学科 PAPD 项目资助。

参考文献 (References)

[1] 汪涛, 崔楠. 国外交叉销售研究综述[J]. 外国经济与管理, 2005, 27(4): 43-49.

[2] Nash, E.L. (1993) Database Marketing: The Ultimate Marketing Tool. McGraw-Hill, New York.

[3] Deighton, J., Peppers, D. and Rogers, M. (1994) Consumer Transaction Databases: Present Status and Prospects. In: Blattberg, Glazer and Little, Eds., *The Marketing Information Revolution*, Harvard Business School Press, Boston, 58-79.

[4] 郭国庆. CRM 与交叉销售在美国金融业的应用及其启示[J]. 山东大学学报, 2003(5): 79-84.

[5] 赵华, 宋顺林. 改进的序列模式挖掘算法在交叉营销中的应用[J]. 计算机工程与设计, 2007(5): 1219-1222.

[6] Li, C.Q., Qin, C.L. and Li, G. (2010) The Implication of Logistic Regression and Neural Nets in Cross-Selling of Bank's Individual Customer. *Proceedings of the Ninth Wuhan International Conference on E-Business Interface*, Alfred University, USA, 2560-2564.

[7] 王雪莲. 针对保险交叉销售问题的 Logistic 建模与预测[D]: [硕士学位论文]. 南京: 东南大学, 2015.

[8] 王新军, 胡曼. 寿险交叉销售的聚类技术实务分析[J]. 保险研究, 2012(1): 86-95.

[9] 李玲瑶. 浅谈美国的住房问题[J]. 科技导报, 1986, 4(2): 44-47.