狗熊会精品案例系列

车险数据分析与商业化应用

关蓉，水妈，王汉生 - 2016 年 8 月

案例版权归狗熊会所有，如需转载，请联系小编

扫描二维码，获取更多精彩案例！

车险数据分析与商业化应用

摘要：本案例通过对车险数据进行统计分析，建立出险因素 0-1 回归模型，挖掘影响出险的重要变量。模型结果显示，车辆级别、车龄、所有者性质、驾驶人驾龄等因素显著影响出险。在商业应用层面，本案例建立的出险因素模型对于制定个性化车险产品、识别不同风险的驾驶人具有一定的指导作用。在未来，结合驾驶行为数据，可制定基于驾驶行为的 UBI 车险产品。

一、背景介绍

随着道路交通行业的持续发展，我国民用汽车保有量呈现逐年快速增长的趋势。截止 2014 年底，我国全国民用汽车保有量达到 15447 万辆，比 2013 年末增长了 12.4%[1]。

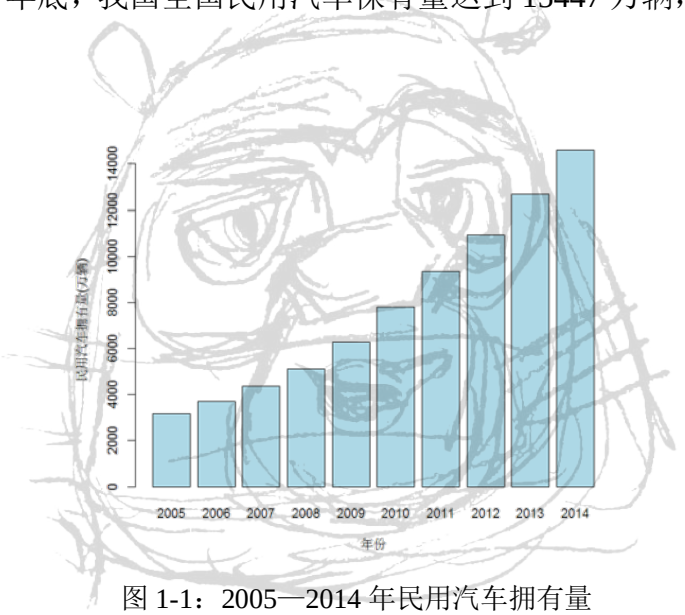

图 1-1：2005—2014 年民用汽车拥有量

汽车行业的繁荣为车险行业提供了蓬勃发展的平台，为车险产品带来了广阔的发展空间。车险产品主要通过车因素、人因素和环境因素三个方面衡量被保险人的风险水平，从而确定保费。其中，车因素包括车辆种类、型号、用途、车龄、行驶区域以及生产厂商等；人因素包括驾驶人年龄、性别、驾龄、婚姻状况、职业以及肇事记录等；环境因素包括气候、地貌、路况等地理环境风险因素以及治安、法制情况等社会环境风险因素。除了上述因素之外，司机的驾驶行为也是衡量风险的重要因素，对车险保费定价有指导作用。

随着互联网的发展与完善，车联网正在兴起，车联网数据可以实现实时采集，为基于司机驾驶行为的分析提供了数据支持。同时，保险大数据公司的诞生与车

[1] 数据来源：国家统计局

© 2016 狗熊会 版权所有

险费率改革制度的出台也推动着新的车险定价模式的诞生。这种新型车险就是UBI（Usage Based Insurance），即基于驾驶人行为的车险。UBI 模式车险是基于驾驶行为以及使用车辆相关数据相结合的个性化保险产品，其核心概念在于给予具有安全驾驶行为的驾驶人保费优惠。UBI 的推广不仅能够使保险公司强化车险定价能力，还可以产生良好的个人与社会效应，引导司机形成良好的驾驶习惯。

UBI模式车险业务在国际上的发展已经较为成熟。在美国保险市场，UBI业务代表公司为Progressive。该公司于1994年首度提出PAYD（pay as you drive）保险概念，之后又陆续推出autograph、tripsense等几代UBI产品。美国State Farm保险公司其后推出In-Drive保险产品，该产品根据司机驾驶行为提供了高达50%的车险费率折扣。在欧洲保险市场，英国保险合作社Cooperative Insurance Society、Insure The Box保险公司，德国、荷兰、意大利等国保险公司先后推出各类基于驾驶行为折扣保费的UBI保险产品。在亚洲保险市场，日本爱和谊日生同和保险公司于2005年与丰田合作在日本推出了类似UBI的车险产品PAYD，该产品仅针对累计里程折扣保费，而未加入其它驾驶行为数据。

国内的UBI模式车险业务尚处于市场探索阶段，但已具备推出基础。保险大数据公司的成立与车险费率改革制度的发布为中国UBI车险业务提供了政策支持，大数据、云计算、车联网技术的成熟为该业务提供了技术支持，前装与后装市场以及智能APP的出现则为该业务提供了丰富的数据基础，从而为中国UBI业务的启动创造了空间。

二、数据来源与说明

本案例所使用的数据来自某保险公司提供的车险数据，共 4233 条记录。数据共包含 11 个变量。其中，因变量为某年度的车险理赔金额，当理赔金额为 0 时，代表当年没有出险；当理赔金额大于 0 时，代表实际的出险金额。本案例将因变量处理成 0-1 变量，即某年度是否出险，通过后续建模挖掘影响出险行为发生与否的重要因素。自变量即为相关影响因素，可分为汽车因素和驾驶人因素两类。具体变量说明如表 2-1 所示。

三、描述性分析

在对出险的影响因素进行建模之前，首先进行描述性分析，初步判断出险行为与各潜在影响因素之间的关联，为后续建模研究做铺垫。

© 2016 狗熊会 版权所有

表 2-1：数据变量说明表

变量类型		变量名	详细说明	取值范围	备注
因变量		是否出险	定性变量（2 水平）	1 代表出险；0 代表未出险	出险占比 28.46%
自变量	驾驶人因素	驾驶人年龄	单位：岁	21~66	只取整数
		驾驶人驾龄	单位：年	0~20	只取整数
		驾驶人性别	定性变量（2 水平）	男/女	男性占比 89.18%
		驾驶人婚姻状况	定性变量（2 水平）	已婚/未婚	已婚占比 95.15%
	汽车因素	汽车车龄	单位：年	1~10	只取整数 建模时离散化
		发动机引擎大小	单位：升	1~3	建模时离散化
		是否进口	定性变量（2 水平）	是/否	国产车占比 70.16%
		所有者性质	定性变量（3 水平）	公司/政府/私人	私人车占比 71.50%
		固定车位	定性变量（2 水平）	有/无固定车位	有车位占比 83.77%
		防盗装置	定性变量（2 水平）	有/无防盗装置	无防盗装置占比 77.60%

（一）自变量：驾驶人因素

驾驶人因素共包含四个变量：驾驶人年龄、驾驶人驾龄、驾驶人性别和驾驶人婚姻状况。通过图 3-1，能够得到以下结论：

- ✧ 驾驶人年龄：从 3-1(a)的箱线图中可以看出，出险和未出险驾驶人年龄的平均水平（中位数）和波动水平的差异并不明显。
- ✧ 驾驶人驾龄：从 3-1(b)的箱线图中可以看出，出险驾驶人驾龄的平均水平（中位数）要明显低于未出险驾驶人，说明新手司机更有可能出险。
- ✧ 驾驶人性别和婚姻状况：从 3-1(c)和 3-1(d)的棘状图可以看出：女性驾驶人的出险率更高，但样本量远小于男性驾驶人；未婚驾驶人出险率略高，但样本量远小于已婚驾驶人。初步的结论是：驾驶人的性别和婚姻状况可能对出险行为有影响。然而，这种影响也可能是由于数据本身的样本量差异形成的。

© 2016 狗熊会 版权所有

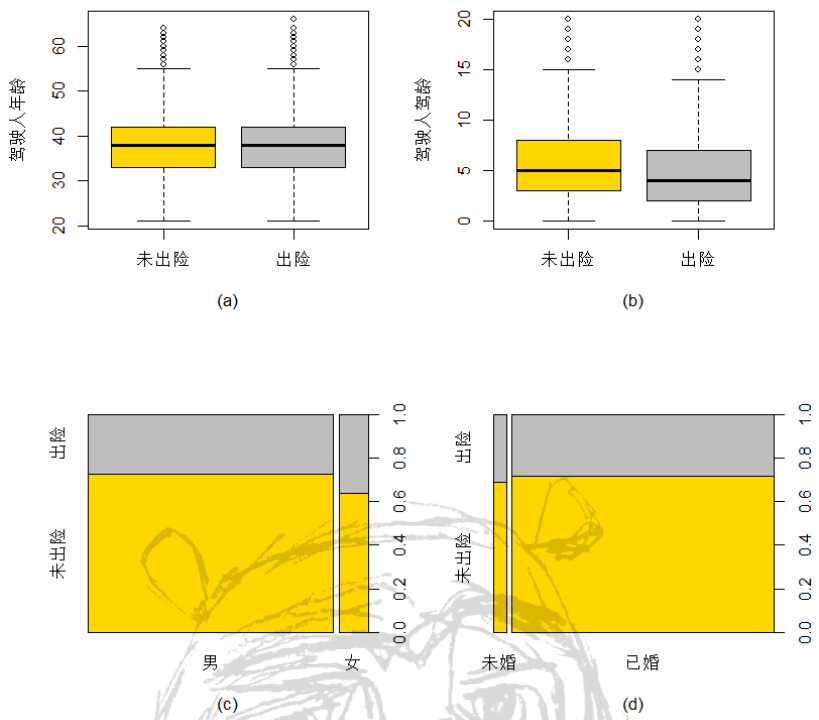

图 3-1：驾驶人因素描述统计图汇总

注：(a) 驾驶人年龄分组箱线图；(b) 驾驶人驾龄分组箱线图；(c) 驾驶人性别棘状图；(d) 驾驶人婚姻状况棘状图。

（二）自变量：汽车因素

案例数据中汽车因素包括六个变量：汽车车龄、发动机引擎大小、是否进口车、所有者性质、是否有固定车位和是否有防盗装置。

首先将车龄变量和引擎大小变量进行离散化处理，即将车龄为 1 年的看作是新车，车龄大于 1 年的看作是旧车；将引擎小于等于 1.6 升的车看作是普通级，引擎大于 1.6 升的看作是中高级。从图 3-2 可以看出，新车出险率更高，普通级车辆出险率更高。因此可以初步判定汽车车龄和车辆级别会影响出险行为。

从图 3-3 可以看出，有防盗装置、有固定车位、进口车以及私人车的出险率略高。值得注意的是，样本量在有无防盗装置、有无固定车位、是否进口车和所有者性质的不同水平之间，分配并不均匀。因此，这种差异是否显著，需要借助后续建模结果进行判断。

© 2016 狗熊会 版权所有

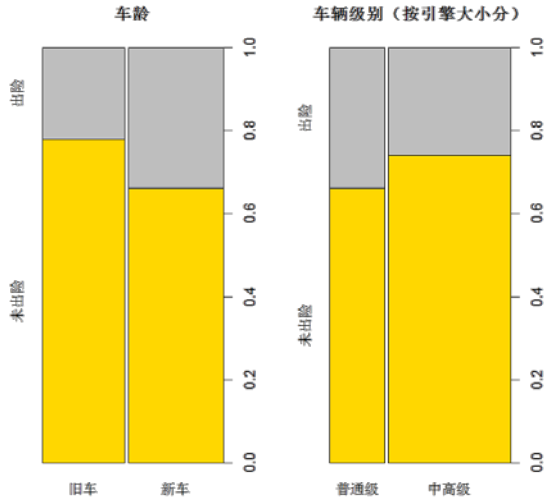

图 3-2：汽车车龄、级别与出险行为棘状图

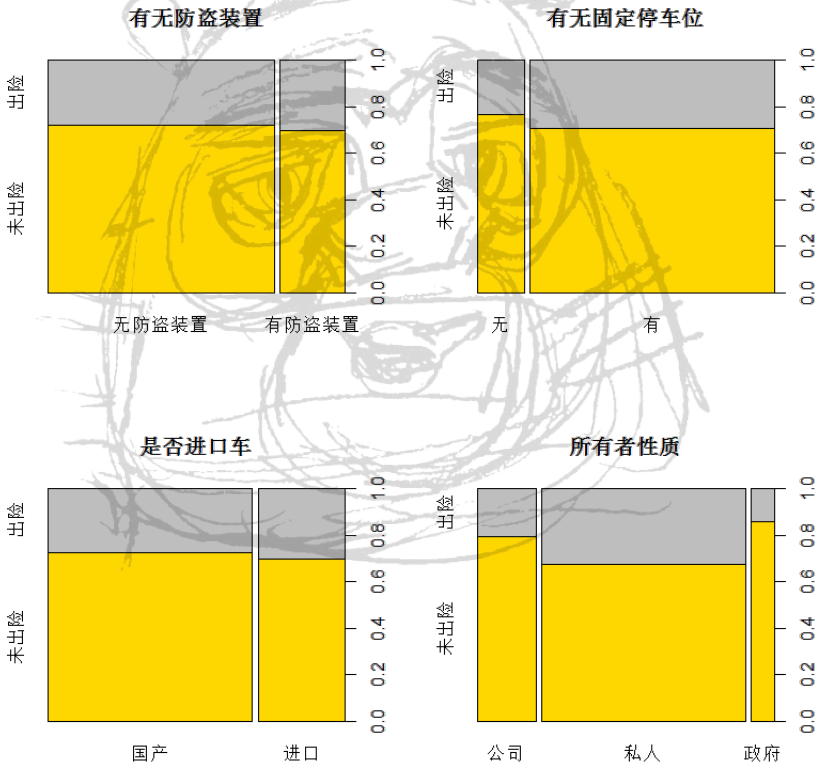

图 3-3：其他汽车因素与出险行为棘状图

通过对数据的描述性分析，本案例认为汽车本身的属性特征、驾驶人的特征都很可能会影响出险行为的发生与否。

四、出险因素统计模型

为了深入挖掘影响出险的显著因素，本案例将建立出险因素的 0-1 回归模型（logistic 模型）。考虑到模型涉及诸多自变量，本案例试图建立模型选择的 AIC

© 2016 狗熊会 版权所有

和 BIC 标准，并综合模型的复杂程度和预测精度，选择最利于刻画出险行为影响因素的统计模型。

（一）模型建立

首先，对所有变量建立 0-1 回归全模型，结果如表 4-1 所示。

表 4-1：0-1 回归全模型结果

变量	回归系数	显著性	备注
截距项	-1.173	***	
中高级车	-0.310	***	
新车	0.359	***	
有防盗装置	0.086		
有固定车位	0.222	*	
进口车	0.149	.	
所有者性质-私人	0.356	***	基准组：企业
所有者性质-政府	-0.338	.	
驾驶人年龄	-0.004		
驾驶人驾龄	-0.027	**	
女司机	0.169		
驾驶人已婚	0.044		
全模型似然比检验	p 值<0.001		

注：***0.001 显著；**0.01 显著；*0.05 显著；. 0.1 显著

从表 4-1 可以看出，显著影响出险的因素有（在 0.05 的显著性水平下）：汽车级别、是否新车、是否有固定车位、所有者性质、驾驶人驾龄。然而全模型还包含很多不显著的因素，因此考虑根据 AIC 准则和 BIC 准则，选择更加简洁的模型，其结果如表 4-2 所示。

表 4-2：AIC 回归模型和 BIC 回归模型结果

变量	AIC 回归系数	显著性	BIC 回归系数	显著性	备注
截距项	-1.252	***	-1.065	***	
中高级车	-0.304	***	-0.261	***	
新车	0.364	***	0.409	***	
有防盗装置					
有固定车位	0.212	*			
进口车	0.158	*			
所有者性质-私人	0.351	***	0.366	***	基准组：
所有者性质-政府	-0.332	.	-0.346	.	企业
驾驶人年龄					
驾驶人驾龄	-0.028	**	-0.029	***	
女司机	0.174				
驾驶人已婚					
模型似然比检验	p 值<0.001		p 值<0.001		

注：***代表 0.001 显著；**代表 0.01 显著；*代表 0.05 显著；.代表 0.1 显著

© 2016 狗熊会 版权所有

从表 4-2 可以看出，AIC 模型和 BIC 模型得出了不一样的结论：在 0.05 的显著性水平下，AIC 模型保留了 7 个变量，而 BIC 模型只保留了 4 个变量。全模型，AIC 模型和 BIC 模型在变量选择和回归系数的估计上都存在差异，那么究竟应该保留哪一个模型？接下来本案例将引入多个概念来进行分析。

（二）模型选择

在对全模型，AIC 模型和 BIC 模型进行比较之前，先引入几个相关概念。首先，表 4-3 给出了一个**混淆矩阵**的概念示意表。值得说明的是表中的预测值是根据某个给定的模型，以及一个特定的阈值（如 0.5）计算得到的。具体的预测规则是：当模型的预测概率大于等于阈值时，预测成出险（+）；当模型的预测概率小于阈值时，预测成未出险（-）。注意，阈值必须是 0 到 1 之间的数。

表 4-3：混淆矩阵

		预测值		总计
		未出险（-）	出险（+）	
真实值	未出险（-）	TN	FP	N
	出险（+）	FN	TP	P
总计		N*	P*	N+P

为了理解后续的分析，有以下几组概念需要强调：

概念 1：TPR（True Positive Rate）：TPR=TP/P，TPR 表示的是“抓住坏蛋的概率”，在本案例中表示的是：成功预测出出险的概率。

概念 2：FPR（False Positive Rate）：FPR=FP/N，FPR 表示的是“冤枉好人”的概率，本案例中表示的是：错把未出险预测为出险的概率。

概念 3：ROC 曲线（Receiver Operating Characteristic Curve）：其横坐标为 Specificity，即 1-FPR，纵坐标为 Sensitivity，即 TPR。ROC 曲线是一条向上凸起的曲线。

概念 4：AUC（Area Under Curve）：ROC 曲线下方的面积，反映的是模型的预测能力。AUC 取值越大，模型的预测能力越强。

接下来，分别绘制全模型、AIC 模型和 BIC 模型的 ROC 曲线并进行比较。如图 4-1 所示，全模型和 AIC 模型的 ROC 曲线非常接近，而 BIC 模型的 ROC 曲线相对而言比较靠下。经计算，全模型的 AUC 为 0.6253，AIC 模型的 AUC 为 0.6241，BIC 模型的 AUC 为 0.6177。综合考虑模型的预测精度和模型的复杂程度后，本案例选择 AIC 模型作为出险因素模型。

© 2016 狗熊会 版权所有

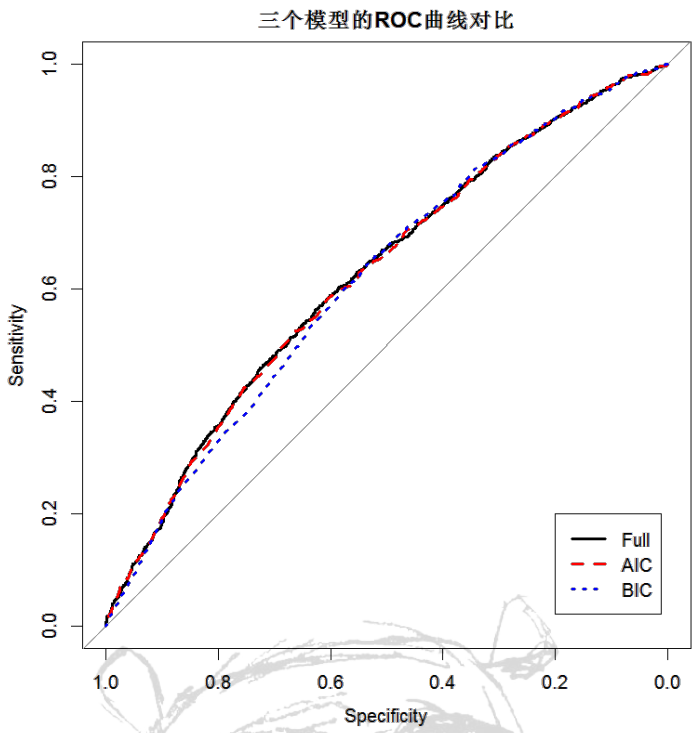

图 4-1：三个模型的 ROC 曲线

（三）模型解读

以下通过表 4-2 对 0-1 回归的 AIC 模型进行解读。本案例首先关注通过了系数显著性检验的变量，它们是与出险行为发生与否高度相关的因素：

- 汽车因素：车辆级别、车龄、有无固定车位、是否进口车和所有者性质
- 驾驶人因素：驾龄

需要特别注意的是，本案例关注的是发生出险行为的可能性，而非发生事故的可能性。

在解读 0-1 回归模型的时候，系数估计值的正负号往往更值得关注，代表了发生出险行为的相对风险（系数为正，则风险较大；反之，风险较小）。对于 AIC 模型的 7 个显著变量而言，若控制住其他影响因素不变，有：

- ✧ 对于车辆级别而言，普通级车辆（引擎小于等于 1.6 升）比中高级车辆（引擎大于 1.6 升）更可能出险；
- ✧ 对于车龄而言，新车更可能出险（车龄为 1 年）；
- ✧ 对于有无固定车位而言，有固定车位的车辆更可能出险；
- ✧ 对于是否为进口车而言，进口车比国产车更可能出险；
- ✧ 对于所有者性质而言，私人车最可能出险，其次是公司的车，最不可能出险的是政府车辆；

© 2016 狗熊会 版权所有

✧ 对于驾驶人驾龄而言，驾龄越大，越不可能出险（相对于老司机，新手司机更可能出险）。

综上所述，较可能出险的车辆具有如下特征：新手司机、进口车、私家车、有固定停车位、新车（车龄为 1 年）、普通级车（排量小于等于 1.6 升）。

（四）模型预测

使用 0-1 回归模型可以预测出险行为发生的概率，进而判断是否会发生出险行为。接下来，本案例将使用 0-1 回归的 AIC 模型对出险行为进行预测。

使用 AIC 模型可以得到每辆车出险概率的预测值。当预测概率大于等于最佳阈值时，预测为出险；当预测概率小于最佳阈值时，预测为未出险。选取不同的阈值，得到的混淆矩阵以及对应的预测准确率也会不同，因此需要选取一个最佳的阈值。

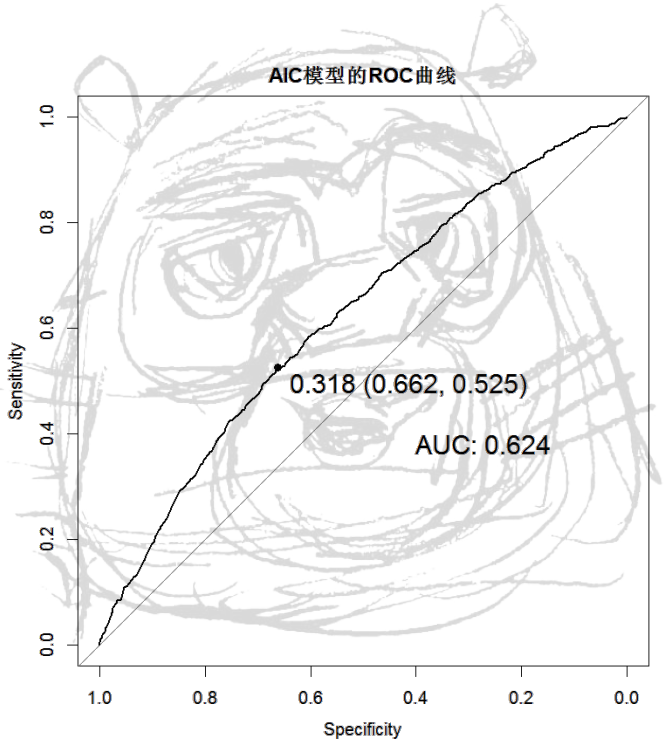

图 4-2：AIC 模型的 ROC 曲线及最佳阈值

最佳阈值的选择标准一般是平衡 TPR 和 FPR。图 4-2 为 AIC 模型的 ROC 曲线、AUC 值以及最佳阈值的选取值（0.318）。对于某一被保险人及其车辆，当 0-1 回归的 AIC 模型得到的预测概率大于等于 0.318 时，可预测为出险。

表 4-4 展示了根据最佳阈值计算得到的混淆矩阵，经过计算，整体错判率为 (1024+572)/4233=37.7%，TPR：633/1205=52.5%（抓住坏蛋的概率），FPR：1024/3028=33.8%（冤枉好人的概率），Sensitivity = TPR =52.5%，Specificity = 1-FPR = 66.2%。

© 2016 狗熊会 版权所有

表 4-4：混淆矩阵（AIC 模型）

		预测值		总计
		未出险（-）	出险（+）	
真实值	未出险（-）	2004	1024	3028
	出险（+）	572	633	1205
总计		2576	1657	4233

需要注意的是，在实际数据分析中，也可以考虑使用样本的出险率作为阈值。

五、商业应用

通过车险数据的出险因素统计模型，可以得到一些十分具有应用前景的信息。

（一）个性化车险定制

前面提到，近年来国外保险公司产生一种新的车险费率厘定模式，即 UBI 驾驶人行为保险。UBI 的理论基础是驾驶习惯良好的驾驶员应获得保费优惠，保费取决于实际驾驶时间、具体驾驶方式等指标的综合考量。保险公司可以直接检测和评估驾驶行为，当车辆发生事故时，车载设备记录下的事故速度以及相关信息会使得理赔评估和处理更有效率。

本案例的出险因素模型即可应用于 UBI，制定个性化车险产品。根据影响出险的显著因素（如车龄、驾龄），对其“出险概率”进行预测，并根据预测结果及驾驶人驾驶特征制定适当的保费标准。不仅如此，还可以进一步结合驾驶行为数据，制定基于驾驶行为的 UBI 车险产品，如对具有“良好”驾驶行为特征的驾驶人给予保费优惠，对具有“不良”驾驶行为习惯的驾驶人适当提高保费。

（二）人群细分

上述的出险因素模型还有一个十分有价值的应用领域：出险人群细分。大致做法是：首先按照 AIC 模型的预测出险概率进行从高到低排序，然后将排序后的驾驶人等分成 5 份，代表从高到底 5 种不同风险人群。将人群进行了细分之后，可以计算这 5 种人群的实际出险概率（如图 5-1 所示）。

可以看出，根据 AIC 模型识别出的低风险人群占总人群的 20%，而其实际出险率只有 17%，比样本的整体出险率 28%低了 11%。模型比较好地识别出了不易出险地人群。对于人群风险等级的划分，也可以应用到 UBI 车险产品中，对高风险人群收取高保费，对低风险人群适当减少保费。

© 2016 狗熊会 版权所有

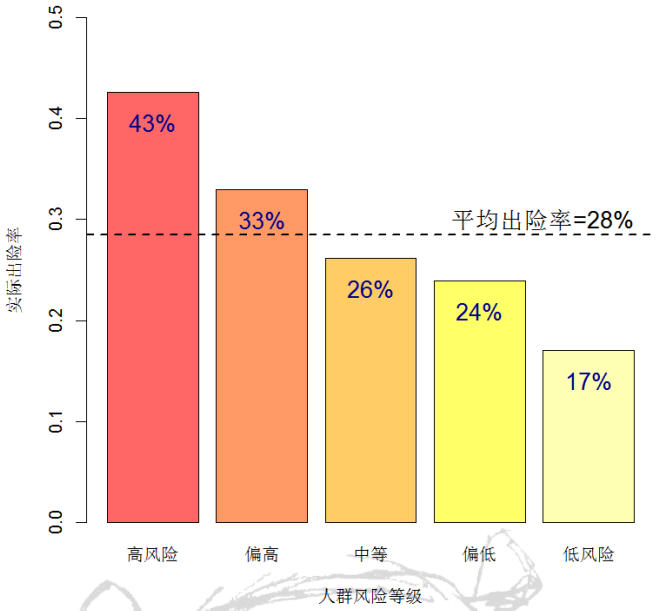

图 5-1：不同风险人群实际出险概率

六、结论与展望

通过以上对车辆出险行为影响因素的分析，可以得到如下结论：

1. 汽车出险行为的发生与否和汽车本身的属性以及驾驶人的特征有关联，出险概率较高的汽车具有如下特征：普通车（发动机引擎小于等于 1.6 升）、新车、有固定车位、进口车、私家车、新手司机。

2. 按照上述特征，0-1 回归模型可测算出险概率、预测出险行为。

3. 根据出险概率对驾驶人进行排序和人群划分，能够有效识别高风险驾驶人、低风险驾驶人，并且有一定的商业价值。

本案例探究的出险因素模型在样本量和变量个数上都较小，因此仍存在一些不足、需要完善。另外，在未来探索中还可以开展更多的研究方向。

1. 本案例的预测结论是基于内样本得到的，这样会高估模型预测精度；未来获取到更多样本时可以对外样本进行预测，这样得到的结论更加有说服力。

2. 模型中的变量较少，未来可以考虑驾驶行为数据来完善模型，制定基于驾驶行为的 UBI 车险产品。

3. 本案例中把出险行为简单地定义为是否出险，在未来可以结合出险险种等数据，预测出险金额。

© 2016 狗熊会 版权所有