第二次上机实验报告

北京大学化学与分子工程学院 李梓烙 2101110396

摘要 本实验运用逻辑回归，对现有汽车用户购买保险的意愿进行拟合，并结合特征预提取、上采样与交叉验证，试图提升模型的预测能力，结果表明：（1）采用所有特征进行拟合，所得模型在训练集的准确率（accuracy）、精确度（precision）、召回率（recall）、F1分数、AUC分别为67.51%、25.26%、84.83%、0.3892、0.7496，在验证集则为67.43%、25.51%、84.60%、0.3920、0.7480，表明即使加入交叉验证，逻辑回归给出的预测结果准确率仍然较低，其原因在于模型倾向于把负样本预测成正样本，导致精确度显著偏低；（2）采用特征预提取得到的最显著的三个特征进行拟合，所得模型在训练集的准确率、精确度、召回率、F1分数、AUC分别为63.81%、24.93%、97.70%、0.3972、0.7840，在验证集则为63.81%、25.25%、97.75%、0.4013、0.7837，表明选取最显著的特征进行逻辑回归，可以大幅提升模型的召回率，但会牺牲一小部分的精确度，从而导致准确率略有下降。而F1分数与AUC的小幅上升，表明特征预提取使模型的预测能力略有改善，但改善的空间有限；（3）两次逻辑回归的结果表明，逻辑回归模型适合于潜在的汽车保险投保者的初步筛选，而不适合于对潜在投保者的进一步精选。

关键词 逻辑回归，分类器，上采样，交叉验证，保险购买，预测

1 引言

在日常生活中，我们不但需要预测某些变量的大小，还需要预测事物的分类，例如预测手机店的顾客喜欢的手机品牌，监测可能存在的传染病携带者等，这就是分类问题的范畴，用数学语言表述，就是用一个（或若干个）超平面将数据剖分为两部分（或多个部分），每一部分均含有一定的特征。

逻辑回归可视为一类分类算法，其思路为建立起线性函数

$$z(x_1,x_2,\dots,x_n)=w_0+w_1x_1+w_2x_2+\cdots+w_nx_n$$

并用Logistic函数$f(z)=\frac{1}{1+\mathrm{e}^{-z}}$进行转换，从而得到一个范围在0~1之间的数值。如果大于0.5，属于分类1，反之属于分类0。

由于逻辑回归的过程中涉及到线性函数的拟合，因此逻辑回归也会存在过拟合的情形，解决这一方案的办法与线性回归一致，仍是正则化，即通过添加惩罚项，抑制损失函数中过大或过小的参数，从而提高模型的泛化能力。本次实验采用二次惩罚项，此时按Scikit-learn的定义，损失函数为：

$$J(w,c)=\frac{1}{2}w^Tw+C\sum_{i=1}^{n}\left[\ln\mathrm{e}^{-y_i(x_i^Tw+c)}+1\right]$$

只要求解出相应的w与c，就可以求出最优模型。常用的求解方法为梯度下降法，其计算过程中不断计算函数的梯度，由此推算出下一步迭代时w与c的变化量，最终得到损失函数的最小值，以及相应的w与c。

2 实验部分

2.1 仪器

2.1.1 硬件

Surface Pro（第5代，处理器参数：Intel® Core™ i5-7300U CPU @ 2.60GHz，2.71 GHz，2个内核，4个逻辑处理器；内存容量：8.00 GB）

2.1.2 软件

操作系统：Windows 10家庭版，版本20H2

开发环境：Visual Studio 2019 Community，64位Anaconda 3（版本号2021.05，含64位Python 3.8.8、Conda 4.10.1、NumPy 1.20.1、Pandas 1.2.4、SciPy 1.6.2、Scikit-learn 0.24.1、Matplotlib 3.3.4、Imbalanced-learn 0.8.1）

2.1.3 训练和测试数据

训练数据：train_valid.csv，含381109条客户的信息，所含字段包括序号（id）、性别（Gender）、年龄（Age）、是否有驾照（Driving_License）、区域码（Region_Code）、之前是否有购买汽车保险（Previously_Insured）、汽车年限（Vehicle_Age）、汽车是否曾损坏（Vehicle_Damage）、年保险费（Annual_Premium）、销售渠道（Policy_Sales_Channel）、服务客户天数（Vintage）、客户是否对汽车保险感兴趣（Response）

测试数据：test.csv，含127037条客户的信息，所含字段除没有Response外，其余字段均与训练数据相同。

2.2 实验过程

2.2.1 训练和测试数据的读入

用Pandas的read_csv方法，读取train_valid.csv与test.csv的数据，作为最原始的数据备用。

2.2.2 根据部分特征的统计结果，绘制柱状图、扇形图、箱线图

首先，笔者按“客户是否对汽车保险感兴趣”这一特征，将train_valid.csv记录的客户划分为“拒绝保险”与“接受保险”两大类，随后在每一大类中，再按选取的特征，统计各类用户所占的比例，由此绘制出关于性别、汽车年限、之前是否有购买汽车保险、汽车是否曾损坏这四大要素的柱状图与扇形图。

另一边，笔者同样按“客户是否对汽车保险感兴趣”这一特征进行划分，然后统计划分后各组中，年龄与年保险费的分布情况，最终绘制出对应的箱线图。上述统计图将用于粗略分析各特征与保险购买意向的关系。

2.2.3 划分训练集与验证集，并分析训练集和验证集的部分特征

用Scikit-learn的train-test-split方法，将train_valid.csv的数据按训练集: 验证集=3: 1的比例，划分为训练集与验证集，并分别存储为train.csv与valid.csv。之后，利用Pandas附带的各种统计函数，如value_count、min、max、mean、median等，对制定的特征进行统计分析，包括：（a）男性和女性客户的比例；（b）年龄的最小值、最大值、平均值、中位数；（c）有驾照的比例；（d）之前有购买汽车保险的比例；（e）汽车年限分别在1年以下、1~2年、2年以上的比例；（f）汽车曾经损坏的比例；（g）年保险费的最小值、最大值、平均值、中位数。训练集与验证集的特征统计结果输出至statistical_properties_of_dataset_train.dat与statistical_properties_of_dataset_valid.dat

2.2.4 训练集和验证集特征的映射

划分后的训练集和验证集既有连续变量（年龄、保险费、服务时长等），也有离散变量（性别、是否有驾照、汽车年限、汽车是否曾损坏等），考虑到本实验解决的是分类问题，为了之后的变量相关性分析与拟合的方便，需要将连续变量分组，并赋予相应的数值；而离散变量中，有部分变量采用字符串表示，而非数值表示，因此也要做特征映射。

笔者的程序中，内置了continuous_var_segmentation与discrete_var_mapping两个函数。前一个函数主要针对连续变量，其作用是将连续变量按等间距划分为指定的段数，并根据所处区间映射为从0开始递增的数值，若设置溢出限，则将超出溢出限的数据额外划分为溢出域（算在指定的段数内）。利用这一函数，笔者将年龄及区域码各自等分为8组等长区间，将销售渠道与服务时长各自等分为18组等长区间，而由于年保险费固有的长尾分布特性，笔者为年保险费划定了90000的溢出限，溢出限以下者，笔者将其九等分，溢出限以上者，笔者为其专门划出溢出域。

后一个函数主要针对非数值型离散变量，其作用则是将含有字符串的数据列按给定的字典映射为数字。结合2.2.2的统计结果，笔者设定如下映射规则：（a）女性设为0，男性设为1；（b）汽车年限小于1年者设为0，大于或等于1年者设为1；（c）车辆未曾损坏者设为0，曾损坏者设为1。

2.2.5 映射后训练集和验证集特征的相关性分析

将训练集与验证集映射成合适的数值后，便可以进行各特征的相关性分析。此处笔者采用了三种相关性衡量指标：其一是Pearson相关系数，它主要应用于连续变量的线性关系检验，其定义为

$$r_{xy}=\frac{\langle xy\rangle-\langle x\rangle\langle y\rangle}{\langle x^2\rangle-\langle x\rangle^2}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

其二是Spearman相关系数，它既可以用于连续变量，也可以用于离散变量，只需两组变量保证单调递增或递减即可，其定义为

$$\rho=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}$$

上式中n为样本数目，d_i为两个变量分别排序后成对的变量位置（等级）差，对于连续变量而言，其排序后的位置即为d_i的值，而对于离散变量而言，若排序后前后位置均有相同取值，则应将这些相同取值的位置取平均后作为该取值对应的d_i的值。

其三是调整互信息，它利用信息熵计算两变量之间的相关程度，与前两种衡量指标相比，调整互信息还能反映出两变量间非线性的相关关系，但调整互信息取值为$[0,1]$，而非Pearson相关系数或Spearman相关系数的$[-1,1]$，因此难以判断两变量之间究竟为正相关还是负相关。

记信息熵为$S_X=-\sum_{x\in\mathcal{X}}P(x)\log_2P(x)$，它代表变量X取值的统一程度，或曰信息的丰富（混乱）程度，同时定义相对信息熵为：假定变量X的概率分布可用Q(X)拟合，而实际分布为P(X)时，交叉熵与真实分布的信息熵之差，即

$$D(P||Q)=\sum_{x\in\mathcal{X}}\big[P(x)\log_2P(x)-P(x)\log_2Q(x)\big]=\sum_{x\in\mathcal{X}}P(x)\log_2\frac{P(x)}{Q(x)}$$

则互信息可视为用$P(x)P(y)$模拟联合分布$P(x,y)$时的相对信息熵：

$$\begin{aligned}I(X;Y)&=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)}=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}P(x,y)\big[\log_2\frac{P(x,y)}{P(x)}-\log_2P(y)\big]\\&=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}P(x,y)\log_2\frac{P(x,y)}{P(x)}-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}P(x,y)\log_2P(y)\\&=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}P(x,y)\log_2\frac{P(x,y)}{P(x)}-\sum_{y\in\mathcal{Y}}P(y)\log_2P(y)\equiv S_{Y|X}-S_X\end{aligned}$$

为消除样本数量的影响，还需要通过下式转换为调整互信息，其中E(I(X;Y))是互信息的期望值，由超几何分布随机模型推定：

$$I_{\text{adj}}(X;Y)=\frac{I(X;Y)-E(I(X;Y))}{\max\{S_X,S_Y\}-E(I(X;Y))}$$

对训练集与测试集的所有特征（包含作为因变量的“Response”）两两配对，并计算出上述三种指标后，用热力图绘制出特征间的相关矩阵。

2.2.6 训练集的上采样处理

机器学习总会遇见如下情形：训练集中阴性样本非常多，而阳性样本极少，这会导致模型的预测结果更容易偏向于阴性（反之亦然）。为了解决样本结果分布不均的问题，通常会采取如下措施：（1）上采样（过采样），即通过程序生成占少数的样本，达到平衡训练集的目的；（2）下采样（欠采样），即通过删除占多数的样本，达到平衡训练集的目的；（3）对样本重新分布权重，使得占少数的样本更容易

被抽中而用以训练。本次实验主要采用方法（1）与方法（3）。

SMOTE（Synthetic Minority Oversampling Technique）是最常用的上采样处理算法之一，它的运行思路为：对每一个少数类样本，采用K-近邻算法找出所有近邻，然后根据少数类样本占原样本的比例，确定采样比例，并从近邻中按采样比例随机挑选，最后在挑选样本与原样本间取点，以生成新样本。本次实验便是采用SMOTE，对划分后的训练集进行上采样。

2.2.7 采用全特征的模型训练、预测与结果评价

逻辑回归中带有惩罚项，其参数C决定了对系数的惩罚力度。为了更高效地挑选C值，实现模型契合程度与泛化能力的平衡，笔者采用带交叉验证的逻辑回归LogisticRegressionCV，对经过上采样并除去“id”一栏的训练集进行拟合，其中交叉验证折数cv为10，C值选取范围为$[10^{-5}, 10^{5}]$内的101个实数，样本权重为“平衡采样”（“balanced”）。此外，为使代价函数变得更加平缓，从而加快训练速度，训练时所有数据都进行正则化处理，即把任意特征X映射为$\frac{X-X_{\min}}{X_{\max}-X_{\min}}$，。

训练完成后，对训练集、验证集、测试集分别进行预测，并将预测结果分别输出至train_predict.csv、valid_predict.csv、test_predict.csv。对于含“Response”真实结果的训练集与验证集，笔者还计算了预测结果相对于真实结果的准确率（accuracy）、精确度（precision）、召回率（recall）、F1分数、AUC，并将上述衡量指标，与模型的惩罚项参数C，一并输出至文件evaluation_of_train_logistic_reg_all_traits.dat与evaluation_of_valid_logistic_reg_all_traits.dat。

2.2.8 采用预选取特征的模型训练、预测与结果评价

仿照2.2.7节的流程，笔者进行另一个逻辑回归模型的训练、预测与结果评价，与2.2.7节不同的是，训练样本不再采用全特征，而是采用2.2.5节分析得到的相关程度最高的三个特征进行训练。训练完成后，同样对训练集、验证集、测试集分别进行预测，并将预测结果分别输出至train_predict_trait_sele.csv、valid_predict_trait_sele.csv、test_predict_trait_sele.csv。对于含“Response”真实结果的训练集与验证集，笔者也计算了预测结果相对于真实结果的准确率、精确度、召回率、F1分数、AUC，并将上述衡量指标，与模型的惩罚项参数C，一并输出至文件evaluation_of_train_logistic_reg_selected_traits.dat与evaluation_of_valid_logistic_reg_selected_traits.dat

3 实验结果与数据

3.1 部分特征统计结果的柱状图、扇形图、箱线图

按照既定的实验流程，笔者首先统计了拒买保险与接受保险的客户中，性别、汽车年限、曾购买汽车保险、车辆损坏这四项特征的占比，由此作出对应的柱状图与扇形图，如图1至图4所示。就性别而言，拒买保险的客户中男性仅略过半数，而接受保险的客户中男性占比六成，看起来男性似乎更愿意购买保险。但考虑到拒买保险与接受保险的人数有可能不均衡，仅仅靠此图断定男性客户更愿意接受保险，未必合适，更好的处理方法是统计不同性别中接受保险的占比，或计算性别与购买保险态度的相关系数，这在稍后的章节中会加以阐述。

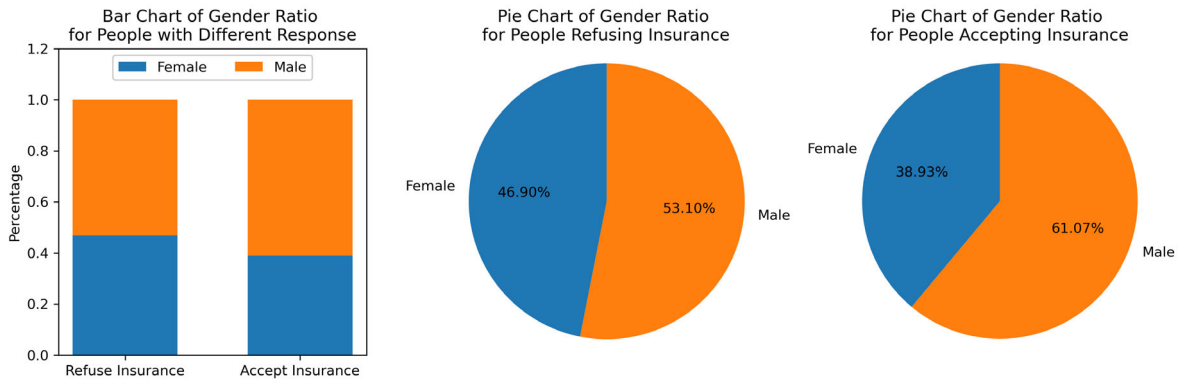

图1 训练-验证集中关于性别占比的柱状图与扇形图

类似的，就汽车年限而言，拒买保险的客户中汽车年限小于1年的客户，与汽车年限大于或等于1年的客户相当，而接受保险的客户中，汽车年限大于或等于1年的客户总计超过八成，故猜想汽车年限越大，客户越有可能购买保险。

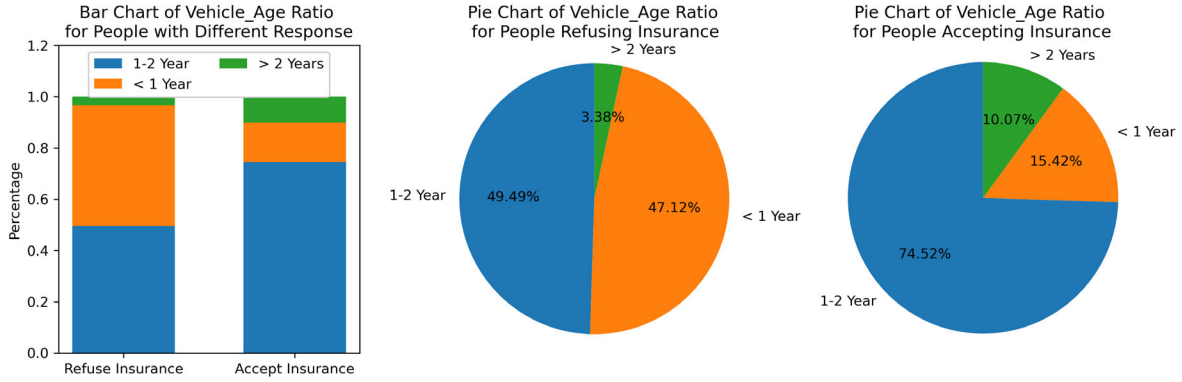

图2 训练-验证集中关于汽车年限占比的柱状图与扇形图

而对“是否有购买汽车保险”与“汽车是否曾损坏”这两个特征的分析，则得出了较为有趣的结果：在拒买保险的客户中，“是否有购买汽车保险”与“汽车是否曾损坏”各自的两个选项的占比都在五成上下浮动，而在接受保险的客户中，之前未曾购买汽车保险的用户，以及车辆曾损坏过的用户，均为压倒性多数，占比均达到95%以上。以上数据意味着，“是否有购买汽车保险”与“汽车是否曾损坏”很有可能也是影响保险购买的重要因素。

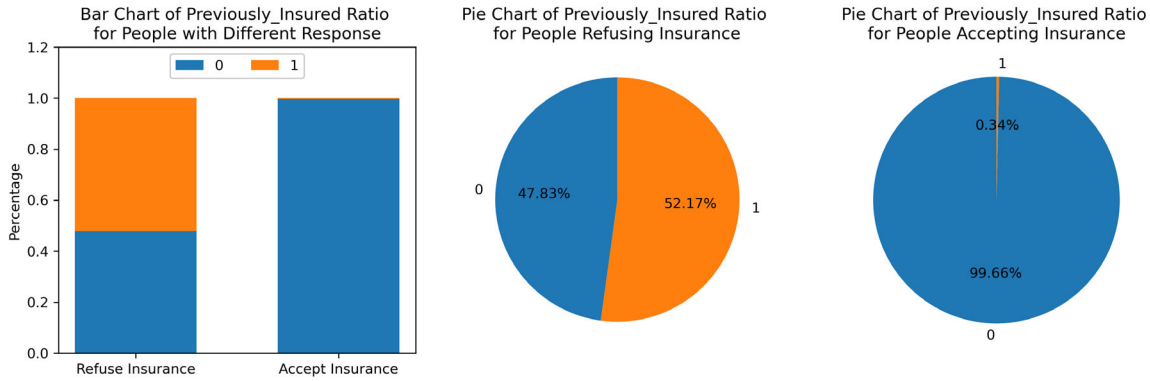

图3 训练-验证集中关于曾购买汽车保险占比的柱状图与扇形图

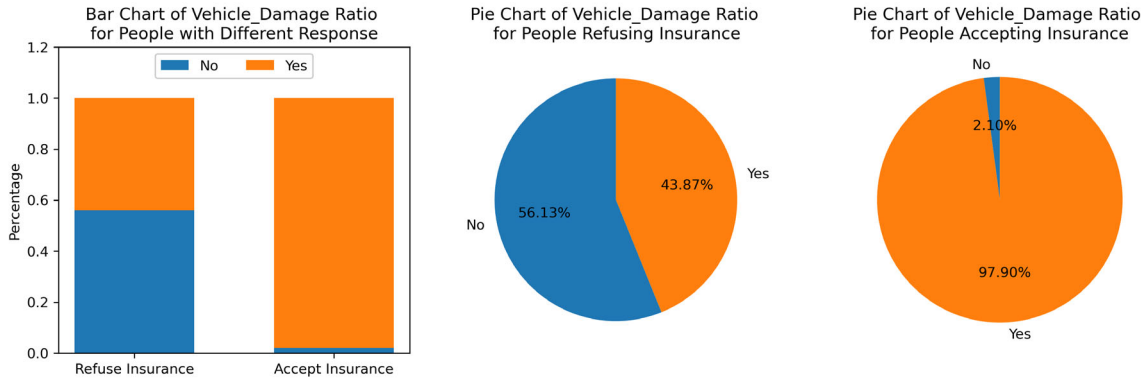

图4 训练-验证集中关于车辆损坏占比的柱状图与扇形图

接下来，笔者绘制了年龄与年保险金分布的箱线图，以此分析拒买保险与接受保险的客户中，年龄与年保险金的分布情况，如图5、图6所示。对年龄的分析表明，拒买保险者年龄的中位数，低于接受保险者年龄的中位数，同时，拒买保险者年龄的前25%~75%人群，其年龄跨度大于接受保险者的年龄跨度。由此笔者猜测，年龄较小的客户，比较有可能倾向于拒绝投保。

图5 训练-验证集中关于年龄分布的箱线图

而对年保险金的分析表明，无论是拒买保险还是接受保险，客户年保险金的金额分布范围、前25%~75%人群分布范围、中位数均无明显差异，且集中在50000以下，表明拒买保险与接受保险的客户，其年保险金的分布没有太大差异，可认为两者相关性极弱；而超过50000的离群点占比虽少，但在庞大的样本数据的放大下，离群点数量也非常可观，表明年保险金具有长尾分布的特点，如果要分箱映射，一定要设定溢出值与溢出域。

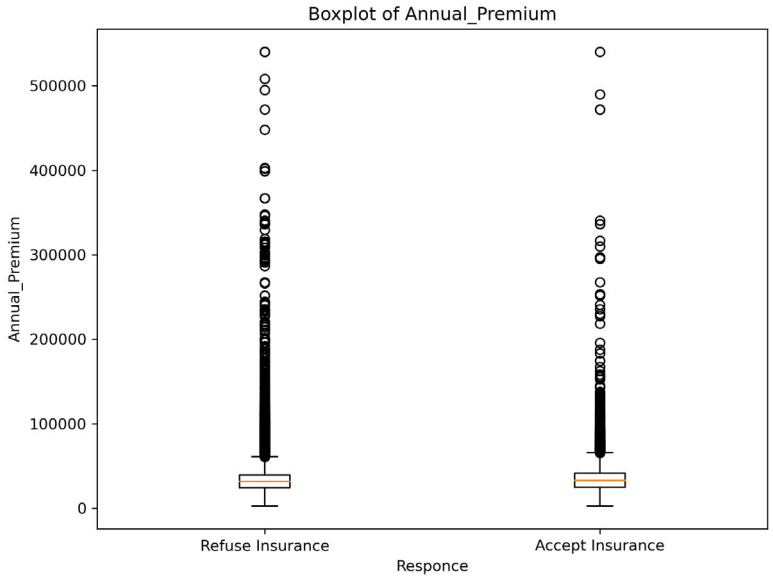

图6 训练-验证集中关于年保险金分布的箱线图

3.2 划分后训练集和验证集的部分特征统计结果

运用Scikit-learn的train_test_split，笔者得到了样本数比为3: 1的训练集与验证集。对训练集部分特征的占比分布的统计（表1）表明，性别、是否曾购买汽车保险、汽车曾损坏这三个特征，在训练集中大致均等分布；汽车年限中，大于2年的客户占比很少，但考虑到在3.1中，1~2年的客户与大于2年的客户均表现出对汽车保险的接受倾向，故可将这两组用户合并，这样汽车年限的数据可认为勉强平衡。相比之下，在"持有驾照"一栏里，超过99%的用户持有驾照，而未持有驾照的客户寥寥无几，这意味着该特征出现了极其有偏的分布，如果采用该特征进行训练，模型的有偏程度也会迅速提升。

表1 训练集中部分特征的占比分布表

性别	持有驾照	曾购买汽车保险	汽车年限	汽车曾损坏
男：54.05%	无：0.21%	是：54.12%	小于1年：43.14%	无：49.58%
女：45.95%	有：99.79%	否：45.88%	1~2年：52.63%	有：50.42%
			大于2年：4.22%	

而年龄与年保险费的统计指标（表2），则用不同于箱线图的方式，反映了两者的数据分布情况。年龄分布中，尽管最大值与最小值有一定跨度（65岁），但平均数与中位数均在35至40岁这一区间，小于最大值与最小值的平均值52.5岁，表明客户年龄主要集中在20至40岁，而在高龄区域有一定的“拖尾”。年保险费分布中，最大值与最小值相差远超50万，即使是平均数或中位数，其与最大值的差距仍有50万，表明年保险费属于典型的长尾分布，这与3.1的结论一致。

表2 训练集中部分特征的最小值、最大值、平均数、中位数

	最小值	最大值	平均数	中位数
年龄	20.00	85.00	38.86	36.00
年保险费	2630.00	540165.00	30554.48	31654.00

为确保验证集的统计结果是否与训练集相同，笔者又统计了验证集对应的特征，结果如表3、表4所示，与表1、表2对比，发现两者数值相差不大，也即从训练集得到的统计特性，对验证集也适用。

表3 验证集中部分特征的占比分布表

性别	持有驾照	曾购买汽车保险	汽车年限	汽车曾损坏
男：54.15%	无：0.21%	是：54.35%	小于1年：43.53%	无：49.31%
女：45.85%	有：99.79%	否：45.65%	1~2年：52.34%	有：50.69%
			大于2年：4.13%	

表4 验证集中部分特征的最小值、最大值、平均数、中位数

	最小值	最大值	平均数	中位数
年龄	20.00	85.00	38.72	36.00
年保险费	2630.00	472042.00	30594.12	31710.00

3.3 映射后训练集和验证集特征的相关性分析

按2.2.4至2.2.5节所述，将训练集和验证集各数据映射为离散变量，随后用Pearson相关系数、Spearman相关系数、调整互信息，对训练集和验证集特征进行两两配对与分析，结果如图7、图8表示。比照图7与图8可知：

（1）训练集与验证集的相关矩阵几乎没有差别，表明训练集与测试集的随机划分没有影响各特征的统计分布；

（2）“曾购买汽车保险”与“是否对汽车保险感兴趣”呈较强的负相关，“销售渠道”取值与“是否对汽车保险感兴趣”呈较弱的负相关，而“汽车年限”及“汽车是否曾损坏”，则与“是否对汽车保险感兴趣”呈较弱的正相关，且“汽车年限”的正相关程度更弱，其余特征基本没有表现出可区分的相关性；

（3）“曾购买汽车保险”与“汽车是否曾损坏”呈非常强的负相关，与“年龄”、“汽车年限”呈一般的负相关，考虑到未购买汽车保险的用户，有可能是刚买车的用户，而这样的汽车不太容易出故障或发生事故，故呈现出这样的关系可以理解；“销售渠道”取值与“年龄”及“汽车年限”呈较强的负相关，而“年龄”及“汽车年限”呈较强的正相关，这可能是由于年轻客户与年老客户拥有车辆时长不同，且与保险公司联系的方式也不同。

（4）Pearson相关系数、Spearman相关系数、调整互信息给出的，具有显著相关性的变量一致，即三种衡量指标均能较好地筛选出强相关变量。

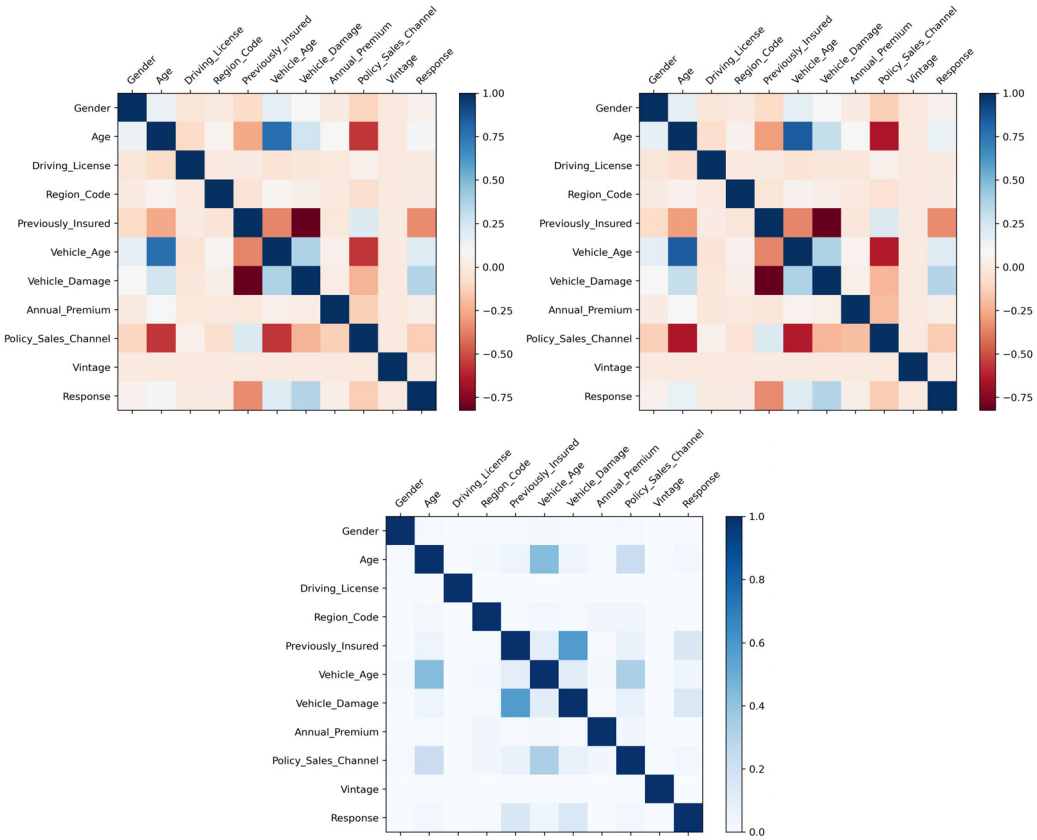

图7 训练集相关性矩阵的热力图表示。左：Pearson相关系数矩阵；中：Spearman相关系数矩阵；右：调整互信息矩阵

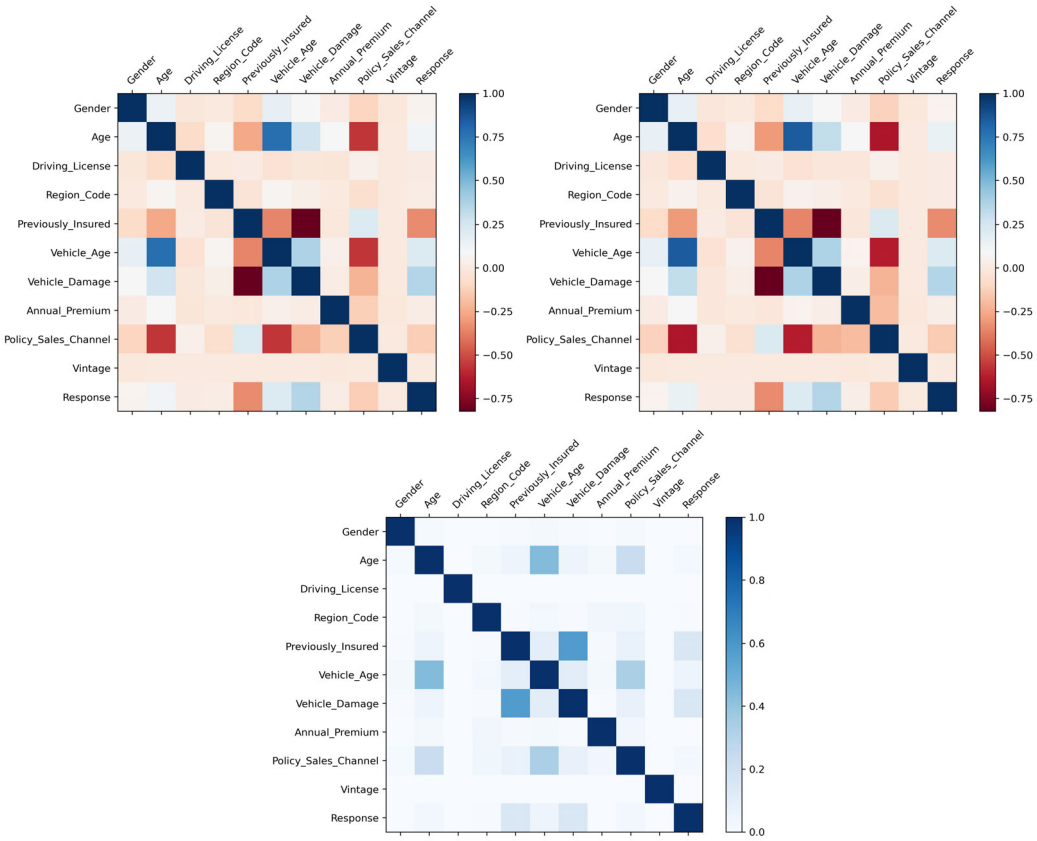

图8 测试集相关性矩阵的热力图表示。左：Pearson相关系数矩阵；中：Spearman相关系数矩阵；右：调整互信息矩阵

3.4 上采样对各特征相关性的影响

按2.2.6节所述，采用SMOTE进行上采样，然后按2.2.5节计算上采样后训练集的Pearson相关系数、Spearman相关系数、调整互信息，并绘制相应的相关矩阵热力图，结果如图9所示。对照图7发现，“曾购买汽车保险”、“销售渠道”与“是否对汽车保险感兴趣”的负相关程度，以及“汽车是否曾损坏”与“是否对汽车保险感兴趣”的正相关程度，均有较大幅度的提升，而“汽车年限”与“是否对汽车保险感兴趣”的正相关程度提升不大。上述现象说明，采用上采样，可以放大部分特征与因变量的关系，减小正负样本不均匀对拟合结果的影响。

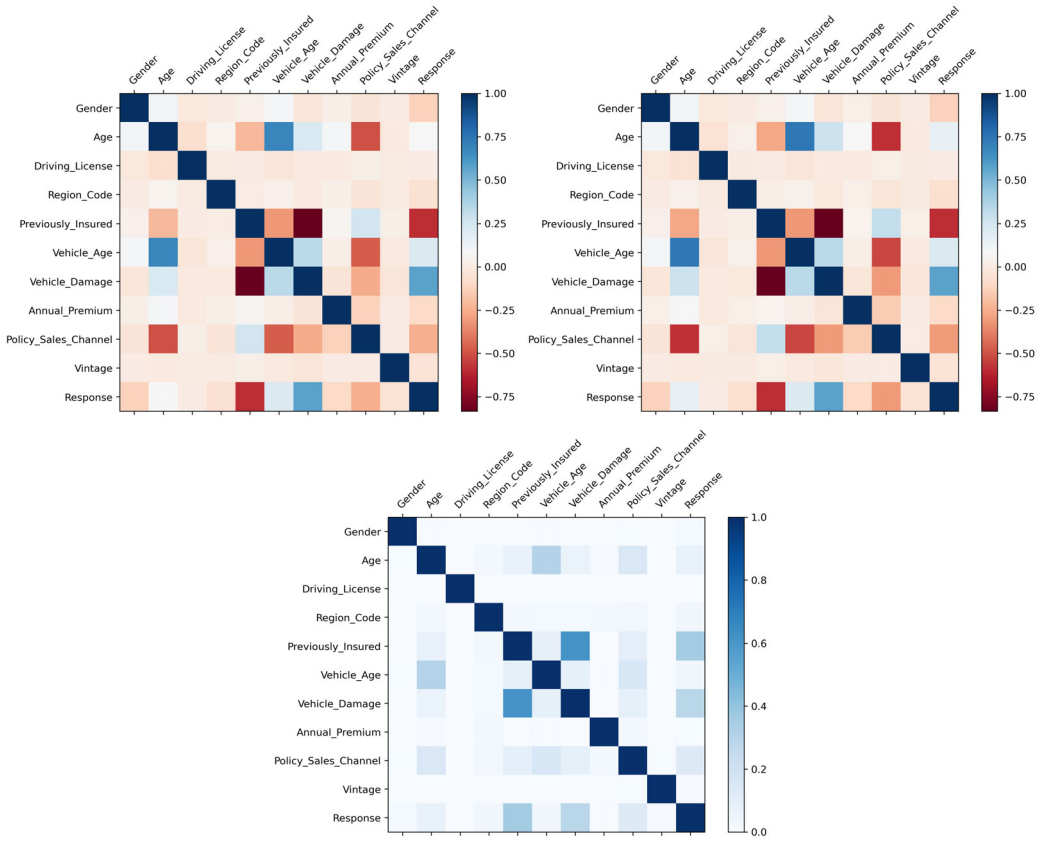

图9 上采样后训练集相关性矩阵的热力图表示。左：Pearson相关系数矩阵；中：Spearman相关系数矩阵；右：调整互信息矩阵

3.5 采用全特征的模型预测结果

按2.2.7节所述，采用全部特征，对上采样后的训练集数据进行逻辑回归拟合，并预测训练集、验证集、测试集中客户对汽车保险的态度。对含有“是否对汽车保险感兴趣”真实数据的训练集与验证集，笔者分别计算了模型在训练集与验证集的表现参数，结果如表5所示。

表5 全特征模型在训练集与验证集的表现参数

	准确率	精确度	召回率	F1分数	AUC
训练集	67.51%	25.26%	84.83%	0.3892	0.7496
验证集	67.43%	25.51%	84.60%	0.3920	0.7480

表5显示，模型在训练集与验证集的预测准确率在67%左右，仅比抛硬币随机决定略优。查看该模型的精确度与召回率，发现其召回率达84%，即该模型可以将大部分有购买保险需求的客户囊括进潜在的客户群，而精确度仅为25%，相当于该模型采取“广撒网”策略，只要稍微符合条件，就把待预测样本分类为正样本，导致部分负样本被认为是正样本而出错。

3.6 采用预选取特征的模型预测结果

根据3.3、3.4节的分析，“曾购买汽车保险”、“汽车年限”、“汽车是否曾损坏”、“销售渠道”这四个变量，与“是否对汽车保险感兴趣”具有较显著的相关性，而“销售渠道”则可以用“汽车年限”解释（年轻客户拥车不久，且恰好偏好某个销售渠道，年老客户拥车较久，同时偏好另一个销售渠道）。因此，笔者预选取“曾购买汽车保险”、“汽车年限”、“汽车是否曾损坏”这三个变量，接着按2.2.8所述，对上采样后的训练集数据进行逻辑回归拟合，并预测训练集、验证集、测试集中客户对汽车保险的态度。对含有“是否对汽车保险感兴趣”真实数据的训练集与验证集，笔者分别计算了模型在训练集与验证集的表现参数，结果如表6所示。

表6 预选取特征模型在训练集与验证集的表现参数

	准确率	精确度	召回率	F1分数	AUC
训练集	63.81%	24.93%	97.70%	0.3972	0.7840
验证集	63.81%	25.25%	97.75%	0.4013	0.7837

将表6与表5对比，可以发现模型的召回率呈显著上升态势，达97-98%，即预选取特征后，训练出的模型能将更多有购买保险需求的客户囊括进潜在的客户群，而召回率的提升，也带动了F1分数和AUC的提升，尤其是AUC，从全特征模型的0.75左右，上升至预选取特征模型的0.78，表明预选取特征能在一定程度上改进预测结果。然而，这种改善不是无偿的：与全特征模型相比，预选取特征模型的精确度略有下降，连带着准确率从67%左右，下降至64%，即这种改善相当于把网撒得更大，“宁抓一千，不放一个”。因此，根据逻辑回归模型的预测表现，以及体现出的可能的预测思路，笔者认为，逻辑回归适合于潜在的汽车保险投保者的初步筛选，而不适合于对潜在投保者的进一步精选。

4 分析与讨论

4.1 为何要采用上采样/类加权

笔者曾尝试不采用上采样，同时将采样权重设为“不加权”（“None”），再进行训练，所得模型在验证集上有80%以上的准确率，但精确度与召回率均为0%，这是由于样本中正样本远少于负样本，导致形如逻辑回归那样的线性模型，会倾向于把样本预测为负样本而非正样本，因此准确率看起来高，但精确度与召回率一塌糊涂。上采样/类加权相当于增大模型选择占比偏少的样本，避免数据的有偏性对模型训练结果造成影响。

4.2 模型准确率不高的原因，以及可能的改进方向

即使是采用预选取特征，逻辑回归模型的预测准确率仍然不高，这一现象的直接原因，是模型倾向于把负样本预测成正样本，导致精确度显著偏低；而若进一步挖掘数据的分布，我们会发现，即使用上采样/类加权的方式，让正负样本得以平衡，样本内部各个特征的分布仍然是不均衡的。例如“是否持有驾照”一栏，便有超过99%的人持有驾照，这就导致我们难以知晓未持有驾照者对购买保险的意见；又如“车辆是否损坏”一栏，在接受保险的客户中，绝大多数都有车辆损坏的经历，只有少数客户尚未遇上车辆损坏。此外，训练模型是维度较低的逻辑回归模型，这种模型在提取高维信息用于分类时较为乏力。

倘若有进一步改善模型预测指标的机会，可能的改进方向有：（1）对统计频率较少的特征，应增大采集力度，以便获得更多具有此特征的数据；（2）换用具有更高维度的分类模型，如神经网络、决策树、支持向量机等。

4.3 实验结论

（1）采用所有特征进行拟合，所得模型在训练集的准确率（accuracy）、精确度（precision）、召回率（recall）、F1分数、AUC分别为67.51%、25.26%、84.83%、0.3892、0.7496，在验证集则为67.43%、25.51%、84.60%、0.3920、0.7480，表明即使加入交叉验证，逻辑回归给出的预测结果准确率仍然较低，其原因在于模型倾向于把负样本预测成正样本，导致精确度显著偏低；

（2）采用特征预提取得到的最显著的三个特征进行拟合，所得模型在训练集的准确率、精确度、召回率、F1分数、AUC分别为63.81%、24.93%、97.70%、0.3972、0.7840，在验证集则为63.81%、25.25%、97.75%、0.4013、0.7837，表明选取最显著的特征进行逻辑回归，可以大幅提升模型的召回率，但会牺牲一小部分的精确度，从而导致准确率略有下降。而F1分数与AUC的小幅上升，表明特征预提取使模型的预测能力略有改善，但改善的空间有限；

（3）两次逻辑回归的结果表明，逻辑回归模型适合于潜在的汽车保险投保者的初步筛选，而不适合于对潜在投保者的进一步精选。