基于随机森林的商业性养老保险购买行为预测

李 强，陈衍姣

（贵州财经大学 大数据应用与经济学院，贵州省大数据统计分析重点实验室，贵阳 550025）

摘要：将随机森林应用到商业性养老保险购买行为预测过程中，对中国综合社会调查（CGSS）2017 年问卷调查数据进行分析。首先运用 SMOTE 过采样来平衡数据集，其次采用网格搜索确认模型输入参数，最后将改进后的随机森林模型进行分类预测，并与支持向量机模型对比。实例结果表明，SMOTE 过采样方法在处理非均衡数据方面表现良好，能够起到提高模型性能的效果，处理后的随机森林的分类效果优于支持向量机。

关键词：随机森林；客户识别；商业性养老保险

中图分类号：F842.6 **文献标志码**：A **文章编号**：1671－1807(2022)08－0271－05

从狭义的视角来看，商业性养老保险是为了满足个人和家庭养老风险保障、投资理财等需求而开发的产品。但是目前中国商业性养老保险的发展并不乐观。首先保费收入较少。根据银保监会公布的数据，2014 年商业性养老保险收益约 1.64 万亿元，仅占全年 GDP 的比重 2.6%。其次居民参保率低。调查发现，在调查样本中仅有 5.6%的采访者购买了商业性养老保险，同时中国商业性养老保险企业产品存在合同条款复杂、收益率低、缺乏创新性等问题，营销手段存在诈骗、被迫等倾向[1]。霍艾湘、赵常兴[2]认为，中国个税递延型商业保险存在着优惠设计偏离初衷，难以满足低收入人群的问题。

那么，如何解决商业性养老保险当前存在的问题，推动养老金制度体系的第三支柱发展呢？国外对于影响商业性养老保险购买行为因素研究较少，大多是针对寿险的研究，因为商业性养老保险是寿险的组成之一，因此本文在外文相关研究中主要借鉴对寿险的相关研究。西方的相关研究起源很早。Truett 等[3]通过对美国和墨西哥的实证数据的分析，认为年龄、收入水平和教育水平是影响寿险购买的主要因素。Browne 和 Kim[4]通过对全球 47 个国家的数据分析，认为通货膨胀、社会保障支出水平和国民收入等宏观因素也会影响寿险购买行为。国内相关研究成果也是基于社会数据分析的结论。陈其芳[5]运用 probit 模型实证证明，农村居民的年龄、受教育程度、家庭收入、对保险的理解、抚养子女和预防老年的态度以及政府宣传对农村居民商业养老保险购买行为有显著影响。张强、杨宜勇[6]通过构建商业养老保险参与影响因素逻辑回归模型，发现个人收入水平、教育程度、基本参保行为、家庭因素等都能够对参保行为产生显著影响。

当前正处于“互联网＋”时代，对于保险公司来说，合理利用大数据是一个巨大的机遇和挑战。而利用数据挖掘技术和机器学习算法，可以有效实现数据可视化，探索业务和数据的内在关联，提高工作效率。国外学者对于机器学习在保险研究中的应用较早。Yeo 等[7]采用 K-Means 对不同投保人风险分组后的理赔成本进行预测分析，提出了一个数据挖掘和非线性整数规划相结合的方法，来确定最佳保费。Kaveh 等[8]提出了一个两阶段聚类算法，用于预测客户的最佳保险范围。国内学者对于大数据技术的应用涉及的主要方法有逻辑回归、决策树、BP 神经网络、支持向量机等。倪泉[9]利用决策树和多元非线性回归的方法建立续期客户交费概率预测模型，对客户质量进行分类，运用聚类分析法，分析具有较高退保风险的客户。葛春燕[10]通过对国内保险公司实际业务分析，构建保险公司评估指标体系，运用 BP 神经网络模型对客户进行分类预测，达到为保险公司规避风险的目的。蔡桂全、陶建平[11]利用局部核函数和全局核函数的线性

收稿日期：2022-04-12

基金项目：国家社会科学基金（18XTJ004）。

作者简介：李强（1969—），男，河南焦作人，贵州财经大学大数据应用与经济学院，教授，博士，硕士研究生导师，研究方向为金融风险管理；陈衍姣（1998—），女，陕西商洛人，贵州财经大学大数据应用与经济学院，硕士研究生，研究方向为金融风险管理。

组合作为权重，构造了多核支持向量机来预测农业保险需求，实证结果表明，该方法比基准支持向量机和 Logistic 回归更准确。

与现有文献相比，本文的创新点为：①创新性地尝试将随机森林算法应用于建立商业性养老保险购买行为预测模型；②能够考虑到商业性养老保险购买数据是一个典型的不均衡的数据，合理地应用处理不均衡数据的过采样方法，改进传统机器学习算法，提高分类准确性；③引入多种算法的对比，增加实证说服性。

1　实证方法应用

1.1　基于随机森林模型的商业性养老保险行为预测模型建立

基于随机森林模型的商业性养老保险行为预测模型如图 1 所示。第 1 阶段采用 SMOTE 算法处理不均衡样本。第 2 阶段网格搜索调节随机森林模型重要的几个输入参数。第 3 阶段运用第 2 阶段改进后的随机森林模型对第 1 阶段处理过的数据进行分类。

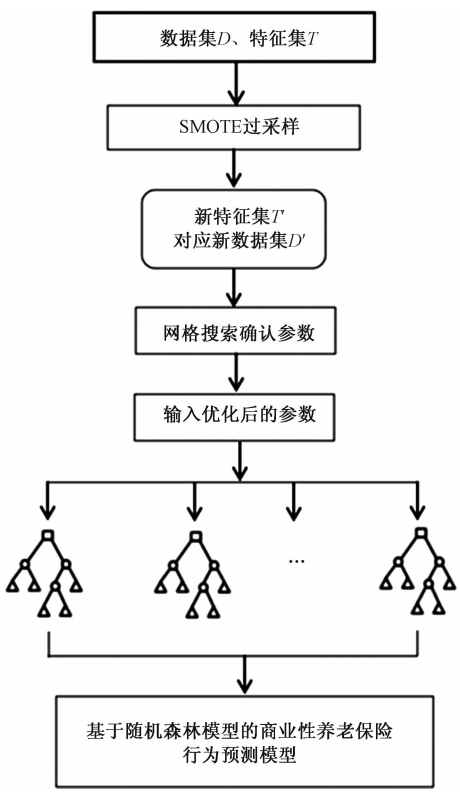

图 1　基于随机森林模型的商业性养老保险行为预测模型

SMOTE 算法流程如下：

1）对于少数类中的每个样本 x，以欧几里得距离为标准计算其到少数类样本集中所有样本的距离，并获得其 k 最近邻。

2）根据样本不平衡率设置采样率，以确定采样率 N，从每个样本 x 的 k 近邻中随机选择若干个样本，假设记为 x_n。

3）对于每个 x_n，根据以下的公式构建新的样本。

$$x_{\text{new}} = x + \text{rand}(0,1) \times |x - x_n| \qquad (1)$$

4）将合成的新样本加入原数据集形成平衡数据集。

随机森林（Random Forest，RF）是由 Leo Breiman 提出的包含多个决策树的组合分类器算法。随机森林在处理多维数据方面具有明显的优势，是目前最好的分类算法之一。随机森林分类（Random Forestforclassification）是采用 bootstrap 方法从原始训练样本集 N 抽取 k 个样本；其次，对 k 个抽取样本建立相应的决策树模型；最后，对得到的 k 种样本结果进行投票，根据少数服从多数的原则选择最终的分类结果。分类决策为

$$H(x) = \arg\max_y \sum_{i=1}^{k} I[h_i(x) = Y] \qquad (2)$$

式中：$H(x)$ 为组合分类模型；h_i 为决策分类模型；Y 为输出变量（目标变量）；$I[h_i(x) = Y]$ 为示性函数。

RF 模型含有许多重要参数，不同的参数组合可以产生不同的结果。为得到更好的预测精确度，采用网格搜索法对模型的重要参数进行调参操作。

实证主体是在 Python3.7 上配合一系列依赖库完成的。用到的最主要的库是 SciKit-learn（简称 Sklearn），是由数据学家 David Cournapeau 在 2007 年发起，专门为机器学习应用而开发的一款开源框架。

1.2　样本及评估指标说明

选用的数据来自中国综合社会调查（CGSS）2017 年调查问卷（居民问卷）的调查结果。中国综合社会调查涉及范围广、抽样方法科学、涵盖内容全面，能很好地反映影响商业性养老保险购买行为的个人因素指标和家庭因素指标，因此选用该调查结果作为研究数据是客观且具有代表性的。首先，利用 stata 将数据导为 Excel 格式，得到初始数据共 12 582 个。

根据阅读文献以及问卷的实际情况，共选取两大类数据，即个人因素和家庭因素。个人因素选取的指标包含年龄、性别、婚姻状况、政治面貌、身体状况、是否购买基本医疗保险、基本养老保险、个人去年总收入、工作性质和单位性质；家庭因素选取

的指标包含去年家庭总收入、子女个数、拥有几处房产、是否有小汽车、是否从事投资活动。然后,对离散特征进行赋值处理,构建的商业性养老保险购买行为预测指标体系见表1。

选择是否购买商业性养老保险为响应变量,商业性养老保险购买行为预测是一个典型二分类问题,购买记为1,否则为0。根据多次试验结果,本文随机从样本中按比例选取30%为测试集,剩余70%为训练集,将回答不明确以及拒绝回答的样本剔除,删除有缺失值的样本,最后保留3 859条数据。

表1　预测指标体系

指标分类	变量名称	赋值
个人因素指标	年龄 X_1	
	性别 X_2	1=男,0=女
	婚姻状况 X_3	1=已婚,0=未婚
	政治面貌 X_4	1=群众,2=共青团员,3=明主党派,4=共产党员
	身体状况 X_5	1=不健康(很不健康、比较不健康)2=一般,3=健康(比较健康,很健康)
	受教育程度 X_6	1=初中及以下,2=职业高中、普通高中、中专、技校,3=大学专科(成人、正规高等教育)、大学本科(成人、正规高度教育),4=研究生及以上
	是否购买基本医疗保险 X_7	1=是,0=否
	是否购买基本养老保险 X_8	1=是,0=否
	个人去年总收入 X_9	
	工作性质 X_{10}	1=全职,0=非全职
	单位性质 X_{11}	1=党政机关,2=企业,3=事业单位,4=社会团体、居/村委会,5=无单位/自顾(包括个体户),6=其他
家庭因素指标	去年家庭总收入 X_{12}	
	子女个数 X_{13}	
	总共拥有几处房产(包括与他人共同拥有)X_{14}	
	是否有小汽车 X_{15}	1=是,0=否
	是否从事投资活动 X_{16}	1=是,0=否

2　实证计算过程及结果分析

基于随机森林算法模型的商业性养老保险购买行为预测模型的实证计算主要在Python语言环境下完成。主要过程如下。

2.1　SMOTE算法过采样

从数据样本容量可以看出,讨论商业性养老保险购买行为,不难发现,与不购买商业性养老保险相比,选择购买是一个明显的小样本事件。这也是金融数据常常会出现的问题,就是数据不均衡。数据不均衡为主流机器学习模型的分类效果带来严峻挑战,稀有事件和噪声发生混淆,少数特征被扭曲,使得模型学习力不足,导致模型预测效果不理想。因此,首先采用一个典型的过采样方法SMOTE进行数据处理。

2.2　参数优化

在模型训练过程中,模型的输入参数的设置对于模型评估时的准确度有着决定性作用。采用网格优化法对5个超参数——森林中树的数目n_estimators、单个决策树使用特征的最大数量max_features、树的最大深度max_depth、叶子节点最少样本数min_samples_leaf、拆分内部节点所需的最小样本数min_samples_split进行寻优,其中选用十折交叉验证法,最终得到输出参数分别为70、12、15、10、40。

2.3　基于ROC曲线的模型性能比较

讨论商业性养老保险购买行为预测问题,不难发现,这是一个典型的二分类问题。考虑一个二分问题,会出现4种情况,即真正类(turepositive)、假正类(falsepositive)、真负类(truenegative)和假负类(falsenegative)。ROC曲线是通用的检验二分类模型性能的方法。

在用ROC曲线评价模型性能时,一般通过对比ROC曲线下的面积AUC来衡量,曲线下面积AUC的值越大,可认为效果越好。对SMOTE过采样前的模型与采样后的模型及采用默认参数的支持向量机(SVC)模型的性能进行ROC曲线分析,其ROC曲线及比较结果如图2所示。

将图2(a)、(b)、(c)放在同一坐标轴进行比较,如图2(d)所示,可以看出明显的差别:3种模型的AUC从大到小排序为:过采样前的RF＜SVC≪过采样后的RF。即数据SMOTE过采样方法能很好地提高模型性能,采用过采样后的随机森林(RF)进行分类的效果要远远优于采用支持向量机(SVC)。基于上述分析,可以得出SMOTE过采样方法在处

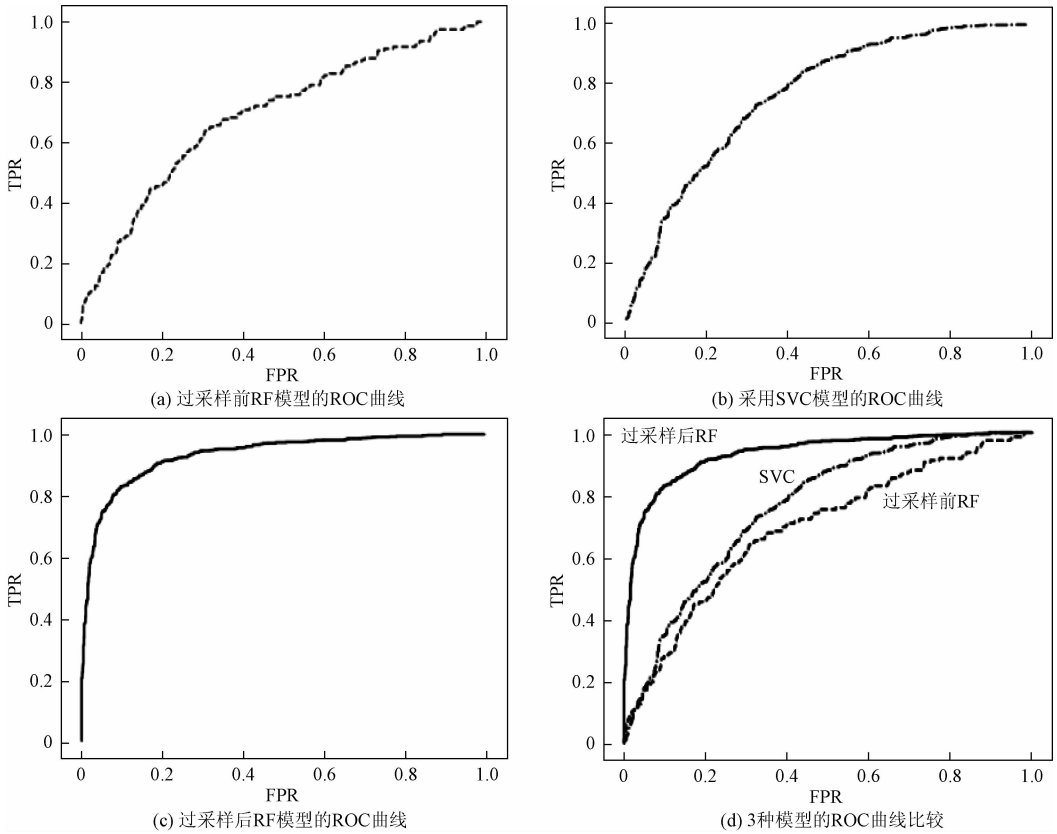

图 2　3 种模型的 ROC 曲线及比较

理非均衡数据方面表现良好，能够起到提高模型性能的效果；与支持向量机模型（SVC）对比，过采样后的随机森林模型（RF）用于对商业性养老保险购买行为预测，具有一定的优势，该模型是可行有效的。

2.4　指标重要性分析

在大数据兴起的时代，各个保险公司汇聚着大量的业务数据信息，若保险公司能够有效利用已有数据，对客户实现精准营销，不仅可以提高工作效率，还可以节省成本提高收益。那么哪些因素最能影响商业性养老保险购买行为成为重点关注的问题。随机森林模型可以计算各变量对因变量的重要性程度。对数据进行过采样和预处理以及对随机森林参数进行选择确认之后，得出各特征对商业性养老保险购买行为的重要程度排序，如图 3 所示。

从指标的重要性程度可以看出，对商业性养老保险购买行为影响最大的是个人去年总收入（X_9）和家庭去年总收入（X_{12}），这与常识一致，收入越高，可支配资金也就越充分，也就更愿意消费和投资，更具有购买商业性养老保险的积极性。排名第 3 的是总共拥有几处房产（包括与他人共同拥有）（X_{14}），有研究表明房产总值对家庭消费的影响显著为正，房产数量能在一定程度上反映房产总值，那么拥有房产越多的人更具有消费欲望，购买可能性也更大。排名第 4 的是受教育程度（X_6），商业养老保险产品往往越来越复杂。过度专业化的保险条款可能会对教育程度较低的居民的理解造成一些障碍。因此，这些居民不太可能购买商业养老保险，教育程度越高，往往越倾向于购买商业性养老保险。其后的

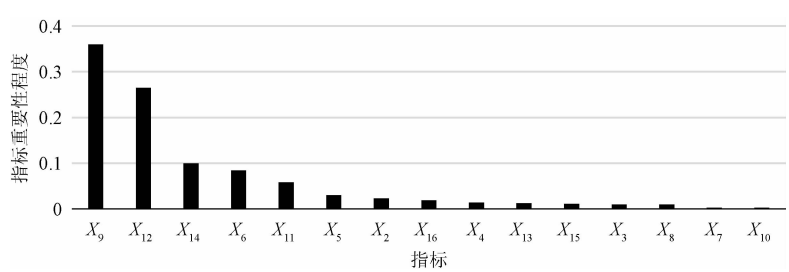

图 3　指标重要性排序

相对重要的特征按顺序依次是单位性质(X_{11})、身体状况(X_5)、性别(X_2)、是否从事投资活动(X_{16})。

3　结论

以2017年中国社会综合调查问卷数据为研究对象,采用SMOTE过抽样算法和随机森林算法,建立了基于随机森林的商业养老保险购买行为预测模型。得到以下结论:

1)大数据带来了丰富的数据信息,智能时代机器学习的引入,提升了数据分析的可视化和智能化,为挖掘数据、运用数据提供了巨大的便利。实例证明随机森林算法的评估模型与目前运用较为广泛的支持向量机算法的评估模型相比,具有一定优势。由于金融数据往往是不均衡数据,采用SMOTE过采样法能较好地解决该问题,提升模型准确性。

2)通过随机森林对指标重要性排序可以看出,对商业性养老保险购买行为影响最大的是收入因素,包括家庭和个人收入。其次是拥有的房产数量,再者是受教育程度。所以保险公司在选择客户群体时,要重点关注这几个指标,依据指标进行客户筛选,采用合理的营销手段,进行针对性的推销。

参考文献

[1] 中国财政学会招标课题"应对人口老龄化财政政策研究"课题组.2021年我国商业养老保险市场调研分析[J].财政科学,2021(8):78-85.

[2] 霍艾湘,赵常兴.个税递延型商业养老保险:实践困境与优化建议[J].西南金融,2021(3):15-27.

[3] TRUETT D B, TRUETT L J. The demand for life insurance in Mexico and the United States: a comparative study[J]. Journal of Risk and Insurance, 1990, 57(2): 321-328.

[4] BROWNE M J, KIM K. An international analysis of life insurance demand[J]. Journal of Riskand Insurance, 1993, 60: 616-634.

[5] 陈其芳.农村居民购买商业养老保险意愿的影响因素分析[J].财经理论与实践,2016,37(1):59-63,109.

[6] 张强,杨宜勇.商业养老保险参与的影响因素分析[J].华中农业大学学报(社会科学版),2017(5):138-143,150.

[7] YEO A C, SMITH K A, BROOKS R. A mathematical programming approach to optimise insurance premium pricing within a data mining framework[J]. Journal of the Operational Research Society, 2002, 53(11): 1197-1203.

[8] KAVEH K D, FARSHID A, SHAGHAVEGH A. Solving customer insurance coverage recommendation problem using a two-stage clustering-classification model[J]. International Journal of Management Science & Engineering Management, 2019, 14(1): 9-19.

[9] 倪泉.基于数据挖掘技术的保险续期催交方法研究[D].上海:华东师范大学,2006.

[10] 葛春燕.数据挖掘技术在保险公司客户评估中的应用研究[J].软件,2013,34(1):116-118.

[11] 蔡桂全,陶建平.基于主成分分析和多核支持向量机的农业保险需求预测[J].济南大学学报(自然科学版),2021,35(2):138-143.

Prediction of Commercial Endowment Insurance Purchasing Behavior Based on Random Forest

LI Qiang, CHEN Yanjiao

(College of Big Data Applications and Economics, Guizhou University of Finance and Economics, Guizhou Province Big Data Statistical Analysis Key Laboratory, Guiyang 550025, China)

Abstract: The applications of random forests is used to predict commercial endowment insurance purchasing behavior. China's general social survey (CGSS) questionnaire survey data in 2017 is analyzed. SMOTE sampling is used to balance data set, then grid search is used to confirm mode input parameters. Finally the improved random forest model predictions is classified. And it is compared with support vector machine model. The results show that SMOTE oversampling method has a good performance in treating disequilibrium data, can improve the model performance, and the classification effect of stochastic forest after treatment is better than that of SVM.

Keywords: random forest; customer identification; commercial endowment insurance