·医疗保障·

# 基于大数据的基本医疗保险参保人欺诈风险评估

李　杰*　兰巧玲　马士豪
河北工业大学经济管理学院　天津　300401

【摘　要】目的:构造基本医疗保险参保人欺诈风险预测模型,发现欺诈行为的主要特征,进而建立风险评估指标体系,以期为医保基金智能监管提供决策支持。方法:利用 183 万多条我国基本医疗保险诊疗历史记录的大规模真实数据,应用 XGBoost 算法和 EasyEnsemble 方法构造基本医疗保险参保人欺诈风险评估集成模型。在此基础上,利用特征重要度计算进一步识别和量化欺诈行为人的潜在特征以构造欺诈风险评估指标体系。结果:模型预测结果的准确性为 83%;阳性与阴性预测值的加权平均值为 95%;参保人欺诈的可能性能够被正确评估的概率为 85%;其中,实际产生欺诈行为的所有参保人中,有 82% 的人员能通过本模型正确识别;各项费用发生金额、各阶段费用发生金额以及各类项目的数量等是区分欺诈与正常参保人的重要指标。结论:基于 XGBoost 集成模型构建的基本医疗保险参保人欺诈风险评估指标体系能够有效地用于识别潜在欺诈人员。建立健全的风险评估指标体系并开发基于医保大数据的智能化监控系统,对于提高医保管理服务水平,保障医保基金安全以及维护社会医保的公平性有重要作用。

【关键词】基本医疗保险; 保险欺诈; 风险评估指标 ; 数据挖掘
中图分类号: R197　文献标识码:A　doi:10. 3969/j. issn. 1674-2982. 2018. 10. 006

## Assessment on insurance fraud risk in basic medical insurance in the context of big data

*LI Jie*, *LAN Qiao-ling*, *MA Shi-hao*
*School of Economics and Management*, *Hebei University of Technology*, *Tianjin 300401*, *China*

【**Abstract**】Objectives:To construct a fraud risk prediction model for basic medical insurance holders, discover the main characteristics of fraud, and then establish a risk assessment index system to provide decision support for an apposite supervision of medical insurance funds. Methods: Using the large-scale real data including more than 183 million records of basic medical insurance diagnosis and treatment in China, the integrated risk assessment model for basic medical insurance holders is constructed using XGBoost algorithm and EasyEnsemble method. On this basis, this paper further identifies and quantifies the potential characteristics of fraud enforcement, and thus constructs a fraud risk assessment index system. Results: The proposed integrated model predicted the fraud risk with the accuracy of 83% , balance predictive value of 95% , and the balance sensitivity was 85% , respectively. Most importantly, the probability of the insured fraud being correctly evaluated was 82% in this fraud risk assessment model. Besides, the amount of various expenses incurred at each stage of assessment, and the number of various types of projects are important indicators to distinguish the fraud from the normal insurance holders. Conclusions:The fraud risk assessment index system constructed based on the XGBoost integrated model is effective for the identification of potential fraudsters among the basic medical insurance holders. Establishing a risk assessment index system and developing an apposite supervision system based on big data of medical insurance play an essential role in improving the level of medical insurance management services, which ensures the safety of medical insurance funds, and safeguards the social health insurance fairness.

【**Key words**】Basic medical insurance; Insurance fraud; Risk assessment index; Data mining

---

* 基金项目: 国家社会科学基金(16FGL014);河北省自然科学基金(G2019202350)
作者简介: 李杰,女(1973 年—),博士,教授,主要研究方向为医疗大数据与智能决策。E-mail:ljrsch@163. com

我国基本医疗保险主要包括城镇职工基本医疗保险、城镇居民基本医疗保险以及新型农村合作医疗,并分别对应城镇职工、城镇非就业居民和农村居民。[1]医疗保险基金是医疗保险运行的物质基础,然而近年来以各种欺诈骗保手段套取国家医保基金的案例层出不穷。根据我国主要城市的报告,医疗保险欺诈所造成的损失约占国内医疗费用的7% ~ 8%,高于发达国家的平均水平。[2]医疗保险欺诈是指个人或组织故意欺骗或歪曲事实以使本人或组织获得不法医疗保险资金的行为。[3]这种行为不仅对医疗保险基金安全构成了巨大威胁,还严重侵害了诚实投保人的合法权益,导致出现制度内的不公平,阻碍了我国社会医疗保险制度的有效运行。2016年,人社部明确指出"要适应信息化发展,大力挖掘和利用医保大数据,全面推广医保智能监控,强化医保经办机构能力建设,提升医保管理服务水平",并相继印发关于社会保险欺诈犯罪管理办法的通知。[4]可见,医疗保险欺诈风险的智能监控已成为社会医疗保险的重要课题之一,它也是制定医疗保险反欺诈政策的重要依据。

国内外学者分别从不同视角展开对医保欺诈问题的研究,积累了大量颇具影响的理论成果。就国外研究而言,研究主体涉及医疗保险投保人、医疗服务提供者以及医疗保险承担者[5],其中以医疗服务提供者为主。研究方法多采用数据挖掘方法,大体上可分为三大类:有监督方法、无监督方法以及两者结合的方法。[6]由于我国医疗保障制度建立较晚,尚在不断发展与改善,因此对于医疗保险欺诈的研究大部分聚焦于医疗保险制度的完善以及欺诈的原因与防范措施等定性分析。在经济学方面,则主要集中于博弈论和信息不对称理论视角的研究。[7-8]而对于数据挖掘方法在医疗保险欺诈识别方面的应用研究尚处于起步阶段。

从现阶段研究成果来看,国际上相关研究为医保欺诈风险的智能评估提供了宝贵的理论与实践基础。但由于国内外医保制度、经济水平以及文化价值观等存在显著差异,国外的研究结论可能不符合我国的实际情况,因此有必要构建符合我国基本医疗保险欺诈特征的风险评估模型。此外,国内对于医疗保险欺诈的评估目前多是基于小样本的实证研究[9,10],其研究结论具有一定的局限性,而对于大数据样本的数据挖掘分析研究将具有更可靠、更普遍的意义。因此,本文旨在基于大规模现实数据,运用数据挖掘方法评估基本医疗保险参保人欺诈风险,进行欺诈预警,在此基础上进一步识别和量化欺诈行为人的潜在特征并构造基本医疗保险欺诈识别指标体系,从而推动医保基金智能监管,减少医疗保险欺诈行为,并为审核专家的后续处理与反欺诈措施的开展提供有效决策支持。

## 1 基本医疗保险欺诈风险评估指标构建

通过梳理国内外医疗保险欺诈的相关文献,对医疗保险欺诈风险评估的特征变量进行归类总结,其主要划分为医院信息、医生信息、患者信息和商业保险信息四个大类。然而,目前用于医疗保险欺诈风险评估的常用特征指标或多或少涉及到个人隐私问题,并且多为显性特征,没有充分考虑病人就诊历史信息中所隐含的潜在行为模式。此外,医保审核机构若要通过医疗机构和医生的相关信息进行参保人的欺诈风险评估,在数据获取方面存在一定的困难,在数据整合处理方面也面临较大挑战。事实上,不论参保人员通过何种手段套取医保基金,最终都会在医疗就诊费用记录上反映出来。因此,其申请报销的诊疗记录数据中必定会包含欺诈违规的相关信息。

鉴于此,本文在保障参保人隐私的前提下,主要基于其诊疗项目、诊疗费用和诊疗频率等大规模就诊历史记录,参考以往研究的评估指标,同时考虑现实中欺诈行为的表现形式,最终构造出27个基于就诊历史信息记录的基本医疗保险参保人欺诈评估指标。总体而言,可将其概括为两类,包括诊疗记录与保险报销指标(表1)。

## 2 数据来源与预处理

### 2.1 数据来源

本研究所用数据来源于2017年天池大数据竞赛中的全国社会保险大数据应用创新大赛"精准社保"赛题,该竞赛由中国社会保险学会主办,人社部信息中心、社保中心和医疗保险司指导,阿里云联合杭州数梦工场科技有限公司具体承办。数据样本是从我国部分地区以往年度的医疗保险就医结算记录中随机抽取的脱敏数据,主要包括2016年6月30日—12月31日期间20 000名参保人员的1 831 381条医疗费用记录及其在不同医保地址的6 533 889条消费金额明细与消费内容。此外,还包括经有关专家审核所得出的参保人员是否欺诈的数据标签(0 - 正常;1 - 欺诈),当中包含欺诈参保人1 000个,正常参保人19 000个。

表 1　基本医疗保险参保人欺诈风险评估指标构建

| 评估指标类别 | | 评估指标 | 指标描述 |
|---|---|---|---|
| 诊疗记录 | 就诊规律 | 就医时间间隔 | 患者各次就医时间间隔天数 |
| | | 各医院就医频率 | 患者到各个医院就诊的频率 |
| | | 就诊频率最高的医院 | 患者就诊次数最多的医院 |
| | | 就诊费用最高医院 | 患者就诊费用最高的医院 |
| | | 就诊过的医院数量 | 患者就诊过的医院的数量 |
| | 项目信息 | 各类项目的数量 | 各类统计项目的数量 |
| | | 各类项目各阶段数量 | 各类统计项目在不同阶段(上、中、下旬)的数量 |
| | | 各类项目数量各阶段增长比例 | 各类统计项目在不同阶段(上、中、下旬)的数量增长比例 |
| | 费用信息 | 总发生金额 | 该时间段内所有医疗项目的发生金额总和 |
| | | 各项费用发生金额 | 药品费、检查费、治疗费、手术费、床位费、医用材料费、输血费以及其他项目费用的发生金额 |
| | | 各阶段费用发生金额 | 各阶段(上旬、中旬、下旬)各项费用及总费用的发生金额 |
| | | 各阶段费用增长比例 | 各阶段(上旬、中旬、下旬)各项费用及总费用的增长比例 |
| | 账单数量 | 账单数 | 患者在该时间段内就诊产生的账单数量 |
| | | 各阶段账单增长比例 | 各阶段(上、中、下旬)的账单数量增长比例 |
| | | 各阶段账单数量 | 各阶段(上、中、下旬)发生的账单数量 |
| 保险报销 | 申报信息 | 总申报费用 | 患者在该时间段内申请报销的费用总额 |
| | | 总费用申报比例 | 患者申请报销的费用占总发生费用的比例 |
| | | 各项费用申报金额 | 各项医疗费用申请报销,由政府支付的金额 |
| | 审批信息 | 各项费用自费金额 | 各项医疗费用由患者自行承担的金额 |
| | | 各项补助金额 | 公务员、残疾军人以及民政救助等补助金额 |
| | | 审批金额 | 包括本次审批的总金额以及补助审批的金额 |
| | 保险标准 | 起付线标准 | 基本医疗保险起付线标准金额 |
| | | 限额标准 | 最高限额以上金额 |
| | 支付账户 | 可用账户报销金额 | 账户中可用的报销金额 |
| | | 统筹基金支付金额 | 费用总额中由统筹基金支付的金额 |
| | | 个人账户支付金额 | 费用总额中由个人账户支付的金额 |
| | | 非账户支付金额 | 并非通过账户支付(即现金支付)的金额 |

### 2.2 数据预处理

上述数据包含大量详细信息。其中,医疗费用记录表包含顺序号、个人编码、医院编码、药品费发生金额、药品费自费金额、药品费申报金额、检查费发生金额、起付线标准金额、基本医疗保险统筹基金支付金额、本次审批金额以及交易时间等共计 69 个特征变量。消费金额明细与消费内容表则包含顺序号、医院编码、服务项目、医院服务项目名称、单价、数量以及费用发生时间等共计 11 个特征变量。

原数据以每条费用记录为一条数据的形式存储,每人包含若干条记录,因此无法直接用于模型训练,需通过剔除无效变量、缺失值填充以及数据整合等预处理,将每名参保人的记录合并为一条数据,最终得到 20 000 名参保人的数据。随后参考表 1,构造基于不同维度下诊疗费用或项目数量的总和、均值及所占比例等统计量的特征变量。不难理解,这些特征变量分别表示这些评估指标所隐含的参保人行为规律。例如,单次就诊账单数的均值代表参保人员每次就诊所产生的账单数量的平均水平。最终得到 827 个特征变量,加上标签变量,与 20 000 个训练集样本构成维度为 20 000 ×828 的样本—特征矩阵。

## 3 模型建立

### 3.1 模型选择

医疗保险欺诈风险识别的实质就是区分医疗费用索赔账单是合法的还是欺诈或滥用的,是数据挖掘中典型的分类问题。决策树算法因其具有可解释性、分类速度快等优点,而在该类问题中被广泛应用。但其预测结果稳定性较低且容易出现过拟合,即在训练数据集中拟合效果很好而在新的数据集中预测效果不佳。而这个问题能够通过集成多棵决策树得以解决,即增强决策树(Tree Boosting)。

增强决策树算法中最常用的是 Adaboost(Adaptive Boosting)、GBDT(Gradient Boosting Decision Tree)和 XGBoost(eXtreme Gradient Boosting)。[11] 其中,XGBoost 算法,即极端梯度提升算法,是结合分类与回归树算法

(Classification and Regression Tree, CART)提出的梯度提升算法的变体,是一种适用于大规模数据的分布式集成学习算法。[12] XGBoost 因其运算速度快、预测准确以及不易过拟合等优点而被广泛应用。学术界也逐步开始尝试应用该算法解决分类与预测问题,并取得了不错的效果。[11,13] 因此,本文采用 XGBoost 构造医保欺诈风险评估的基模型。

### 3.2 XGBoost 的基评估模型

定义包含 $n$ 个基本医疗保险参保人和 $m$ 个特征属性的数据集 $D=\{(X_i,y_i)\}$ $(|D|=n, X_i\in Rm, y_i\in\{0,1\})$,其中 $X_i$ 表示参保人 $i$ 的特征向量,$y_i$ 表示参保人 $i$ 是否欺诈(0 - 正常,1 - 欺诈)的分类标签。单棵 CART 能够为每一个样本训练出对应的预测分数 $f(X_i)$,XGBoost 则是集成多个 CART 的预测结果所得到的加法模型,如式(1)所示。

$$\widehat{y_i}=\phi(X_i)=\sum_{k=1}^{K}f_k(X_i), f_k\in F \tag{1}$$

其中,$F=\{f(x)=\omega_{q(x)}\}(q:R^m\to T,\omega\in R^T)$ 表示 CART 的空间,$q$ 表示每棵树的结构,即将样本映射到叶子节点的索引,$T$ 表示树上的叶子数量,$\omega$ 表示叶子节点的权重(分数),每一棵 $f_k$ 对应一个独立的树结构 $q$ 和叶子权重 $\omega$。对于一个给定的参保人样本,通过运用 $K$ 棵 CART 的决策规则,将其映射到对应的叶子节点,并将各个叶子节点的映射分数相加则可得到该样本最终的分类预测得分。

模型的目标函数为:

$$\begin{aligned}Obj(\Theta)&=L(\Theta)+\Omega(\Theta)\\&=\sum_{i=1}^{n}l(y_i,\widehat{y_i})+\sum_{k=1}^{K}\Omega(f_k)\end{aligned} \tag{2}$$

由上式可知,目标函数包括两部分,第一部分 $\sum_{i=1}^{n}l(y_i,\widehat{y_i})$ 表示预测值 $\widehat{y_i}$ 与真实值 $y_i$ 之间的训练误差。它衡量模型是否符合训练数据规律,旨在通过其优化促使训练数据接近其真实的潜在分布。第二部分 $\sum_{k=1}^{K}\Omega(f_k)$ 表示模型复杂度的惩罚项(正则化函数)。通过控制惩罚项有助于实现叶子权重平滑,其目的是鼓励训练出相对简单的模型,以期在未来的预测中减小方差,从而避免模型过拟合,使预测结果更加稳定。

由于该模型的参数中包含函数,即树的结构,而不是数值向量,无法采用传统的欧几里得空间求解方法。因此,采用迭代的方式进行模型求解,从常数预测开始,每轮迭代新增加一个函数到模型中,即:

$$\begin{aligned}\widehat{y_i}^{(0)}&=0\\\widehat{y_i}^{(1)}&=f_1(x_i)=\widehat{y_i}^{(0)}+f_1(x_i)\\\widehat{y_i}^{(2)}&=f_1(x_i)+f_2(x_i)=\widehat{y_i}^{(1)}+f_2(x_i)\\&\cdots\\\widehat{y_i}^{(t)}&=\sum_{k=1}^{t}f_k(x_i)=\widehat{y_i}^{(t-1)}+f_t(x_i)\end{aligned} \tag{3}$$

其中:$\widehat{y_i}^{(t)}$ 表示第 $i$ 个样本在第 $t$ 次迭代中的预测分数。每一轮迭代中的 $f_t(x_i)$ 由目标函数优化得到,即:

$$\begin{aligned}&Obj^{(t)}+\sum_{i=1}^{t}\Omega(f_i)\\&=\sum_{i=1}^{n}l(y_i,\widehat{y_i}^{(t-1)}+f_t(x_i))+\Omega(f_t)\end{aligned} \tag{4}$$

利用泰勒公式对目标函数进行二阶展开可实现快速优化。因此,定义 $g_i=\partial_{\widehat{y_i}(t-1)}l(y_i,\widehat{y_i}(t-1))$,$h_i=\partial^2_{\widehat{y_i}(t-1)}l(y_i,\widehat{y_i}^{(t-1)})$,则可得到近似目标函数:

$$\begin{aligned}Obj^{(t)}&\cong\sum_{i=1}^{n}[l(y_i,\widehat{y_i}^{(t)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]\\&\quad+\Omega(f_t)\end{aligned} \tag{5}$$

移除上式中的常数项,即可得到第 $t$ 次迭代的简化目标函数:

$$\begin{aligned}\widetilde{Obj}(t)&\cong\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]\\&\quad+\Omega(f_t)\end{aligned} \tag{6}$$

将 $f_t$ 表示成树的结构形式,用叶子的权重来定义树,即:$f_t(x)=\omega_{q(X)}(q:R^m\to T,\omega\in R^T)$。根据叶子节点数量 $T$ 和叶子节点权重 $\omega_j$ 的 $L_2$ 范数定义复杂度惩罚项,即:$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$,其中 $\gamma$ 和 $\lambda$ 分别为两者的惩罚系数,用以控制正则化的程度。最后定义每个叶子中包含的样本集合为:$I_j=\{i\mid q(X_i)=j\}$,则目标函数可按叶子节点改写成:

$$\begin{aligned}Obj^{(t)}&=\sum_{j=1}^{T}[(\sum\nolimits_{i\in I_j}g_i)\omega_j+\frac{1}{2}(\sum\nolimits_{i\in I_j}h_i+\lambda)\omega_j^2]\\&\quad+\gamma T\end{aligned} \tag{7}$$

定义 $G_j=\sum_{i\in I_j}g_i$,$H_j=\sum_{i\in I_j}h_i$,则有:

$$Obj^{(t)}=\sum_{j=1}^{T}[G_j+\frac{1}{2}(H_j+\lambda)\omega_j^2]+\gamma T \tag{8}$$

对于给定的第 $t$ 棵树的结构 $q(X)$,可由 $\partial_{\omega_j}Obj^{(t)}=0$ 求得叶子的最优权重 $\omega_j*=-\frac{G_j}{H_j+\lambda}$,相应的最优目标函数为:

$$Obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j 2}{H_j+\lambda} + \gamma T \qquad (9)$$

上式可以作为衡量一棵树分类效果好坏的标准,其值越小越好。对于包含大量特征变量的模型,列举所有可能的树结构几乎是不可能实现的。因此,对于树结构的求解采用贪婪算法,即从树深度为 0 开始,对于每一个叶子节点迭代地添加一个特征进行分裂。然后计算分裂前后的叶子分数以求得信息增益(Gain),即添加该特征能够使数据集 D 的分类不确定性减少的程度。信息增益的计算公式如下:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{(H_L+H_R+\lambda)}\right] - \gamma \qquad (10)$$

其中:$\frac{G_L^2}{H_L+\lambda}$表示分裂后的左侧叶子分数,$\frac{G_R^2}{H_R+\lambda}$表示分裂后的右侧叶子分数,$\frac{(G_L+G_R)^2}{(H_L+H_R+\lambda)}$表示分裂前的叶子分数,$\gamma$ 则表示引入额外叶子的复杂度成本。选择信息增益最大的特征及其最佳分裂点进行树的分割,并在信息增益≤0 或迭代次数 $t$ 达到规定的阈值时停止分裂,最终即可获得“不纯度”最小的分类树结构。

### 3.3 基于 EasyEnsemble 的医保欺诈风险评估集成模型

在构建分类模型时,需要为数据样本划分训练集和测试集。其中,训练集的数据用以拟合模型,挖掘其特征向量背后隐藏的规律。测试集则用来测试模型对新样本的判别能力,即模型用于预测新的参保人员是否欺诈的风险评估效能。原始数据中包含 20 000 名参保人员,其中欺诈人员 1 000 名,正常人员 19 000 名。由此可见,这是一个数据类别严重不平衡的分类问题。为保留数据分布特征,在划分训练集与测试集时应保证大类样本和小类样本的比例不变。并且为了保证模型的泛化能力,训练集和测试集样本应该尽可能地互斥。因此,本文采用分层抽样的方法,以 7:3 的比例随机划分训练集和测试集。

两类样本的比例高达 1:19,具有极度不平衡的特点,这也是现实数据的真实特征。而大多数分类学习算法的基本假设为不同类别训练样本的数目相当,若差别较大则会对学习过程造成干扰,从而影响模型的预测效能。因此,本文借鉴 EasyEnsemble 方法[14],利用集成学习机制,将训练集中大类样本通过欠抽样划分为若干子集,再分别与小类样本组合构成不同的训练集样本,并应用上述 XGBoost 算法训练出多个基评估模型。最后进行模型集成,将所有基评估模型对于测试集样本欺诈可能性概率的预测结果求均值,即可得到测试集样本的最终风险评估得分。基于 XGBoost 算法和 EasyEnsemble 方法的欺诈风险评估集成模型构建思路如图 1 所示。

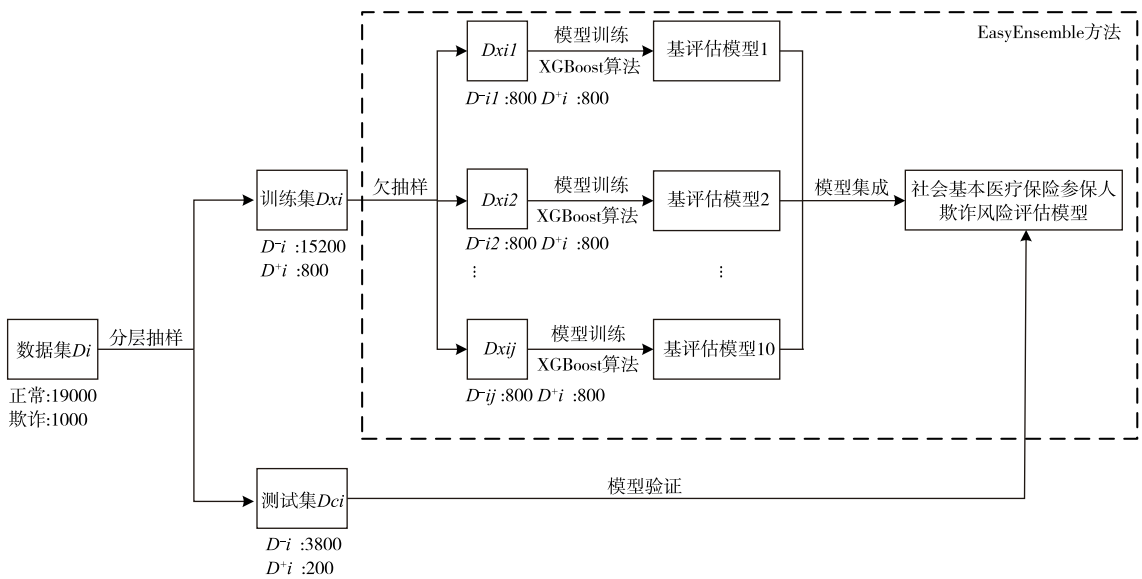


图 1　基本医疗保险参保人欺诈风险评估集成模型

将该模型应用于测试集的 6 000 个样本中进行欺诈风险预测(图 1),最终得到这 6 000 个参保人员的欺诈风险预测概率。其中,欺诈风险概率大于 0.5 则判定为欺诈(“1”),反之则判定为正常(“0”)。对于判定为欺诈的样本,发出预警提醒审核专家展开审核,进而实施相应的预防、警告与惩罚措施。

## 4 结果

### 4.1 模型性能评估

为了评估基本医疗保险参保人欺诈风险，对于模型效果评价需关注的指标为：准确性 ACC(Accuracy)、平衡预测值 BPV(Balance Predictive Value)、平衡敏感性 BS(Balance Sensitivity)以及 AUC 值(Area Under ROC Curve)四个指标。根据数据样本真实类别与模型预测类别的结果组合定义“混淆矩阵”，(表 2)。

表 2　混淆矩阵

| 真实值 | 预测值 | |
|---|---|---|
| | 欺诈(“1”) | 正常(“0”) |
| 欺诈(“1”) | TP(真阳性) | FN(假阴性) |
| 正常(“0”) | FP(假阳性) | TN(真阴性) |

则阳性预测值($PV_1$)、阴性预测值($PV_0$)、敏感性($S_1$)、特异性($S_0$)分别为：

$$PV_1 = \frac{TP}{TP + FP}, PV_0 = \frac{TN}{TN + FN} \tag{11}$$

$$S_1 = \frac{TP}{TP + FN}, S_0 = \frac{TN}{TN + FP} \tag{12}$$

其中，阳性与阴性预测值表示测试集中两类样本被正确预测的比例，即预测值与真实值相符的样本数占该类别总样本数的比例。敏感性与特异性表示测试集中两类样本被正确“召回”的比例，即实际为欺诈(或正常)的样本中被正确预测为欺诈(或正常)的样本比例。一般情况下，在数据挖掘中采用平均预测值和平均敏感性作为评估指标。但由于研究数据的样本类别存在极度不平衡的特点，因此应根据样本比例求平衡预测值和平衡敏感性。即：

$$BPV = \omega_1 PV_1 + \omega_0 PV_0 \tag{13}$$

$$BS = \omega_1 S_1 + \omega_0 S_0 \tag{14}$$

最终得到基于 XGBoost 算法的基本医疗保险参保人欺诈风险评估集成模型总体准确率为 0.83，平衡预测值为 0.95，平衡敏感性为 0.85。即预测结果与真实结果相符的样本比例为 95%，测试样本能够被正确预测的比例为 85%。此外，为了减少欺诈行为所导致的损失，应当尽可能多地识别出有可能产生欺诈行为的参保人，因此还应重点关注欺诈类样本的敏感性，其结果为 0.82，即实际产生欺诈行为的参保人中，有 82% 的人员能通过本模型有效识别。模型的 AUC 值，即受试工作者曲线 ROC(Receiver Operating Characteristic Curve)下的面积为 0.91，说明该模型对于两类样本的区分效果较好。综上所述，本文所采用的基于 XGBoost 算法的风险评估集成模型能够有效地预测基本医疗保险参保人的欺诈风险，从而实现快速有效的智能化风险监管。因此，模型中的重要特征亦能够用于基本医疗保险参保人欺诈风险评估指标体系的构建。

### 4.2 欺诈风险评估指标体系

XGBoost 是一个基于增强树的算法模型，每棵树的构造均基于特征重要度分数，从而亦表明了每个特征对于欺诈风险评估的重要性。特征越多地被用于增强树构造的关键决策，则该特征的重要度得分就越高。具体而言，该算法通过信息增益来计算特征重要度得分，即本文 3.2 部分中所提及的衡量树分类纯度的主要参考指标。本文所采用的是基于 EasyEnsemble 方法的 XGBoost 集成模型，因此，首先对其所有子分类模型的特征取并集，并计算每个特征的重要度得分均值，即可得到集成模型的 328 个重要特征变量，及其对应的特征重要度得分。随后，将这些变量依照表 1 的风险评估指标分类汇总并对各类指标的特征重要度得分求总和。最后，为了实现各类指标重要度的有效对比，按照式(15)求取相对重要度得分，即可构建出基本医疗保险参保人欺诈风险评估指标体系。

$$\omega'_k = \frac{\omega_k}{\sum_{k=1}^{n} \omega_k} \tag{15}$$

其中，$\sum_{k=1}^{n} \omega_k = 1$，$n$ 表示最终评估指标的数量，$\omega_k$ 表示第 $k$ 个指标的重要度得分，$\omega'_k$ 表示第 $k$ 个指标的最终得分，即相对重要度得分。最终得到由 $n = 23$ 个指标构成的基本医疗保险参保人欺诈风险评估指标体系(表 3)。

结合表 1 和表 3 的结果进行整体分析：(1)就诊疗记录而言，费用信息类的指标重要度均排名靠前，表明诊疗费用是能够反映基本医疗保险参保人是否存在欺诈行为最重要的一类指标，在进行欺诈风险评估时应重点参照；各类项目的数量和账单数量类次之；最不重要的是就医时间间隔等就诊规律类指标。(2)总申报费用、限额标准、各类项目各阶段数量以及各类项目数量各阶段增长比例这四个指标对于基本医疗保险参保人的欺诈风险评估效能几乎没有影响，因此在进行欺诈风险评估时可以不纳入考

虑范围。(3)就保险报销记录而言,进行欺诈风险评估时应重点关注各项费用申报金额、各项费用自费金额以及支付账户类指标,而对于总费用申报比例和各项补助金额则无需过多考虑。

表 3　基本医疗保险参保人欺诈风险评估指标体系

| 排序 | 评估指标 | 重要度 | 指标类别 | 排序 | 评估指标 | 重要度 | 指标类别 |
|---|---|---|---|---|---|---|---|
| 1 | 各项费用发生金额 | 0. 241 | 费用信息 | 13 | 审批金额 | 0. 025 | 审批信息 |
| 2 | 各阶段费用发生金额 | 0. 123 | 费用信息 | 14 | 账单数 | 0. 025 | 账单数量 |
| 3 | 各类项目的数量 | 0. 113 | 项目信息 | 15 | 各阶段账单数量 | 0. 021 | 账单数量 |
| 4 | 各项费用申报金额 | 0. 081 | 申报信息 | 16 | 各医院就医频率 | 0. 015 | 就诊规律 |
| 5 | 各阶段费用增长比例 | 0. 061 | 费用信息 | 17 | 各阶段账单增长比例 | 0. 014 | 账单数量 |
| 6 | 统筹基金支付金额 | 0. 048 | 支付账户 | 18 | 就诊过的医院数量 | 0. 009 | 就诊规律 |
| 7 | 各项费用自费金额 | 0. 041 | 审批信息 | 19 | 总费用申报比例 | 0. 008 | 申报信息 |
| 8 | 总发生金额 | 0. 041 | 费用信息 | 20 | 各项补助金额 | 0. 008 | 审批信息 |
| 9 | 起付线标准 | 0. 033 | 保险标准 | 21 | 就诊费用最高医院 | 0. 004 | 就诊规律 |
| 10 | 非账户支付金额 | 0. 031 | 支付账户 | 22 | 就诊频率最高的医院 | 0. 003 | 就诊规律 |
| 11 | 个人账户支付金额 | 0. 027 | 支付账户 | 23 | 就医时间间隔 | 0. 003 | 就诊规律 |
| 12 | 可用账户报销金额 | 0. 026 | 支付账户 | | | | |

结合表 3 中评估指标的现实含义来看,各项费用发生金额指标最重要,其背后所代表的含义是基本医疗保险参保人员中的欺诈者有可能大量地进行某几类项目的诊疗。例如一定时间内超量购买某些医保药品,以供他人使用或高价转卖给“黄牛”谋取利益。与之相对的各类项目的数量指标亦是同样的道理。各阶段费用发生金额指标,即上旬、中旬和下旬的费用发生金额,表明欺诈人员有可能集中于每个月的固定时期进行相对规律的医保项目消费。因此,有可能存在被他人定期使用保险证/卡非法申领保险金以及主动要求医院开具本人不必要的药品由他人代用等情况。甚至可能存在诈骗团伙诱使参保人出借尚余报销额度的医保卡,从而在某一时段(如月末)集中非法开药,倒卖医保药品的问题。统筹基金支付金额与各项费用自费金额指标,则表明欺诈人员实施欺诈行为时会综合参考统筹基金支付与自付的相关费用,即有可能倾向于参与统筹基金支付比例较高的项目。

总之,本文所构建的基本医疗保险参保人欺诈风险评估指标体系能够很好地挖掘欺诈人员的潜在行为特征与行为规律。但值得注意的是,该指标体系不能直接作为审核标准,而是在模型发出欺诈可能性预警后为专家提供评估思路与方向,实现决策支持。例如,费用信息类指标最能反映欺诈人员的行为特征,则在评估参保人欺诈风险时可构造其各项费用发生金额、各阶段费用发生金额与各阶段费用增长比例等指标的均值、标准差、最大值或最小值等,并与历史记录中正常参保人的平均水平相比较,从而得出评估结论。此外,亦有助于针对这些欺诈行为特征制定合理的反欺诈政策。

## 5 结论及建议

基本医疗保险制度的持续、有效运行对于保障与改善民生至关重要,合理评估基本医疗保险参保人的欺诈风险并构建风险评估指标体系是社会保险反欺诈获得成功的前提条件。本文基于我国基本医疗保险诊疗历史记录的大规模真实数据,运用数据挖掘中的 XGBoost 算法构造基本医疗保险参保人欺诈风险评估集成模型,从而预测参保人的欺诈风险概率,进行参保人欺诈预警,并根据模型中的重要特征构造基本医疗保险参保人欺诈风险评估指标体系。

研究结果表明,运用该模型对基本医疗保险参保人进行欺诈风险评估,预测结果与真实结果相符的正确率为 95%,参保人的欺诈可能性能够被正确评估的概率为 85%。其中,实际产生欺诈行为的参保人中,有 82% 的欺诈者能通过本模型有效识别。因此,本文所构建的基本医疗保险参保人欺诈风险评估指标体系能够很好地区分欺诈人员与正常人员。进一步在此基础上开发出基本医疗保险参保人欺诈风险智能评估系统,就能保证对参保人行为的及时监控,从而实现医保基金更加智能的监管。

为维护医保基金安全,保障医保体系有效运行,结合基本医疗保险参保人欺诈风险评估指标体系的

重要指标,即应从开展诚信宣传教育,加强医疗服务规范与医疗欺诈行为监管,构建大数据智能化监控系统,完善反欺诈法律法规入手进行政策构建。

**作者声明本文无实际或潜在的利益冲突。**

## 参 考 文 献


[1] 姚奕, 陈仪, 陈聿良. 我国基本医疗保险住院服务受益公平性研究[J]. 中国卫生政策研究, 2017, 10(3): 40-46.

[2] Peng H, You M. The Health Care Fraud Detection Using the Pharmacopoeia Spectrum Tree and Neural Network Analytic Contribution Hierarchy Process[C]//Trustcom/BigDataSE/ISPA, 2016 IEEE. IEEE, 2016: 2006-2011.

[3] Ortega P A, Figueroa C J, Ruz G A. A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile[J]. DMIN, 2006(6): 26-29.

[4] 人社部. 人力资源社会保障部关于积极推动医疗、医保、医药联动改革的指导意见[EB/OL]. (2016-06-29)[2018-01-03]. http://www.mohrss.gov.cn/SYrlzyhshbzb/shehuibaozhang/zcwj/yiliao/201607/t20160705_242949.html

[5] Faseela V S, Thangam P. A Review on Health Insurance Claim Fraud Detection[J]. International Journal of Engineering Research Science, 2015, 1:47-49.

[6] Pranali P, Namrata G. Fraud Detection in Health Insurance using Random Forest Algorithm[J]. International Journal of Informative & Futuristic Research, 2017, 4(8): 7301-7308.

[7] 曹传帅, 胡西厚, 王雪蝶. 基于系统动力学的流动人口医保关系转接问题博弈分析[J]. 中国卫生事业管理, 2017, 34(7): 505-507.

[8] 胡思洋. 大病医疗保险中医保机构的道德风险问题研究[J]. 西安财经学院学报, 2017, 30(1): 91-96.

[9] 李连友, 林源. 新型农村合作医疗保险欺诈风险度量实证研究[J]. 中国软科学, 2011(9): 84-93.

[10] 林源. 基于BP神经网络的新农合欺诈识别实证研究——以定点医疗机构欺诈滥用为中心[J]. 云南师范大学学报(哲学社会科学版), 2015(3): 117-128.

[11] 张钰, 陈珺, 王晓峰, 等. Xgboost在滚动轴承故障诊断中的应用[J]. 噪声与振动控制, 2017, 37(4): 166-170.

[12] Chen T, Guestrin C. Xgboost: A scalable Tree Boosting System[C]//Proceedings of the 22nd AcmSigkdd International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[13] Torlay L, Perrone-Bertolotti M, Thomas E, et al. Machine Learning-XGBoost Analysis of Language Networks to Classify Patients with Epilepsy[J]. Brain Informatics, 2017, 4(3): 159-169.

[14] Liu X Y, Wu J, Zhou Z H. Exploratory Undersampling for Class-imbalance Learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(2): 539-550.


[收稿日期: 2018-04-09 修回日期: 2018-07-23]
(编辑 薛云)