Hans Journal of Data Mining 数据挖掘, 2021, 11(1), 1-10
Published Online January 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.111001

Hans汉斯

基于体检大数据的健康指数建模

谢昌锟[1]，赵明琪[2]，林世明[1*]

[1]厦门大学信息学院(国家示范性软件学院)，福建 厦门
[2]厦门大学数学科学学院，福建 厦门
Email: *xmulsm@xmu.edu.cn

收稿日期：2020年12月1日；录用日期：2020年12月31日；发布日期：2021年1月12日

摘 要

近年来，随着健康医疗大数据平台的快速发展，越来越多的体检数据整合到大数据平台上。如何挖掘并利用健康医疗海量数据提高医疗服务质量，提升医患沟通水平是一个全新的挑战。文中应用机器学习算法对45,374个体检用户，共3,529,829条体检数据进行分析数据的探索性分析和特征工程。在个人信用风险评分模型的基础上，将预测模型由梯度集成决策树改进为LASSO回归模型，增加评分卡的可解释性，同时结合体检的应用场景和输入数据，建立体检评分模型。实验结果表明在体检大数据集上，健康指数分数基本上服从正态分布，符合线性回归模型的先验假设。该评分模型同时具有稳健性和区分度的特点，可综合各项体检指标，较为客观地描述用户身体健康状况水平，降低体检用户同医生的沟通成本，督促用户更加关注身体整体健康状况水平。

关键词

机器学习，数据探索，LASSO回归，评分卡，健康指数

Health Score Model Based on Big Data of Physical Examination

Changkun Xie[1], Mingqi Zhao[2], Shiming Lin[1*]

[1]School of Informatics Xiamen University (National Demonstrative Software School), Xiamen Fujian
[2]School of Mathematical Sciences Xiamen University, Xiamen Fujian
Email: *xmulsm@xmu.edu.cn

Received: Dec. 1st, 2020; accepted: Dec. 31st, 2020; published: Jan. 12th, 2021

*通讯作者。

文章引用: 谢昌锟, 赵明琪, 林世明. 基于体检大数据的健康指数建模[J]. 数据挖掘, 2021, 11(1): 1-10.
DOI: 10.12677/hjdm.2021.111001

Abstract

In recent years, with the rapid development of health care big data platform, more and more physical examination data are integrated into the big data platform. A new challenge is how to improve the quality of medical services by using massive medical data. In this paper, we use machine learning algorithm to visually analyze 3,529,829 physical examination data of 45,374 physical examination users. On the basis of personal credit risk scoring model, the prediction model is improved from gradient integrated decision tree to lasso regression model, which increases the interpretability of scorecard. At the same time, combined with the application scenarios and input data of physical examination, we established the health score model. The health index score basically obeys normal distribution, which is consistent with the prior hypothesis of the linear regression model It can integrate various physical examination indicators, objectively describe the health status of users, reduce the communication cost between users and doctors, and urge users to pay more attention to the overall health status.

Keywords

Machine Learning, Data Exploration, Lasso Regression, Score Card, Health Score

1. 引言

《“健康中国 2030”规划纲要》中提出，健康中国建设需要从以“治已病”为中心向以“治未病”为中心转变[1]，同时随着体检机构的信息化建设的推进，健康体检向健康管理过渡成为一种必然趋势。2016 年，国务院印发《关于促进和规范健康医疗数据应用发展的指导意见》，文件指出健康医疗大数据是国家重要的基础性战略资源，其应用发展将带来健康医疗模式的深刻变化，有利于激发深化医药卫生体制改革的动力与活力，培育健康医疗大数据应用新业态[2]。近几年，随着人工智能、云计算、大数据、物联网等相关技术的发展，已初步建立区域级的健康医疗大数据平台，积累了一定的数据量，形成一个巨大的数据“矿产”。体检中心作为较为成熟的医疗机构，一方面数据格式相对统一，另一方面数据量巨大，如能利用机器学习、人工智能等技术手段挖掘数据背后的知识具有较大科研价值和社会效益。目前因数据隐私等问题体检大数据尚未得到有效利用，体检中心的体检信息缺乏智能算法的分析，只是简单堆积罗列的健康档案文档[3]，同时导致用户从体检到获得体检结果的时间周期很长。

通过文献检索发现关于“健康指数”相关研究目前还鲜有开展，本文创新性的提出一种基于体检大数据，并利用机器学习算法建立一个可以持续跟踪群体健康状况的量化指标——健康指数。该健康指数可以及时客观地反应用户的整体健康状况，以此描绘健康画像，结合历史数据预测健康走势。同时该模型通过用户各项体检指标的变量选择和参数估计，初步揭示群体身体状况与各种指标之间的相互联系，抓住影响健康的关键因素，为用户的健康管理提供参考，达到预防慢性非传染性疾病，提高人群生活质量，降低医疗支出的目的[4]。

2. 数据和分析方法

2.1. 数据来源

本数据来源某体检医院，包含两个数据表，分别是 MEDICAL_DIAG_EXPORT2010 和 MEDICAL_

INO_EXPORT2010，MEDICAL_INO_EXPORT2010 为体检项目记录表，含有数据记录 3,529,829 条，MEDICAL_DIAG_EXPORT2010 为体检诊断结果表，含有数据记录 202,203 条[5]。利用 Python 的 Pandas 模块提取并转换为 Dataframe 格式，数据基本情况如表 1、表 2 所示。

Table 1. Results of physical examination
表 1. 体检诊断结果表

列名	类型	非空条目数
病人 ID	Int64	202,203
体检号	Int64	202,203
性别	Object	202,203
年龄	Int64	202,203
诊断名称	Object	202,197
诊断类型	Object	202,203

Table 2. Items of physical examination
表 2. 体检项目表

列名	类型	非空条目数
病人 ID	int64	3,529,828
性别	Object	3,529,828
年龄	Int64	3,529,828
项目名称	Object	3,529,828
检查结果	Object	3,469,105
参考范围	Object	2,019,933
异常提示	Object	2,279,112
诊断标志	Float64	587,979

2.2. 数据预处理

2.2.1. 缺失值处理

体检信息数据中 45,375 例体检者原始数据进行初步筛选，设定阈值剔除缺失值超过体检者 2/3 的体检指标，剩余 102 项指标，同时加入“年龄”连续变量数值指标。针对体检诊断结果表格，将“诊断类型”数量为 0 的患者归为“健康”，用“0”编码；诊断类型数量大于 0 的为“非健康”，用“1”编码。数据显示，这种归类下健康的病例为 25,543 例，非健康的病例为 19,832 例。

2.2.2. 异常值处理

体检信息表诊断标志作为分类变量，存在小数的异常情况，取其整数部分。

连续变量中的年龄数值，如图 1 所示，存在 197 岁的异常值，使用中位数替换。同时 1 岁的体检数据最多，初步推断是婴幼儿体检，而[2, 16]这个年龄区间内数量很少，如果特征没有离散化，可能用户年龄增长一岁就会产生完全不同的输入，会给模型造成很大的干扰。为了让模型具有更好的鲁棒性，应用 kmeans 聚类方法进行连续变量离散化，分为 6 个区间：[0, 13.6]，[13.6, 31.1]，[31.1, 41.1]，[41.1, 52]，[41.1, 52]，[52, 65]，[65, 99]，对应年龄_0 到年龄_6，同时按照表 3 对区间独热(one-hot)编码化代替对数

值变量的归一化[6]。

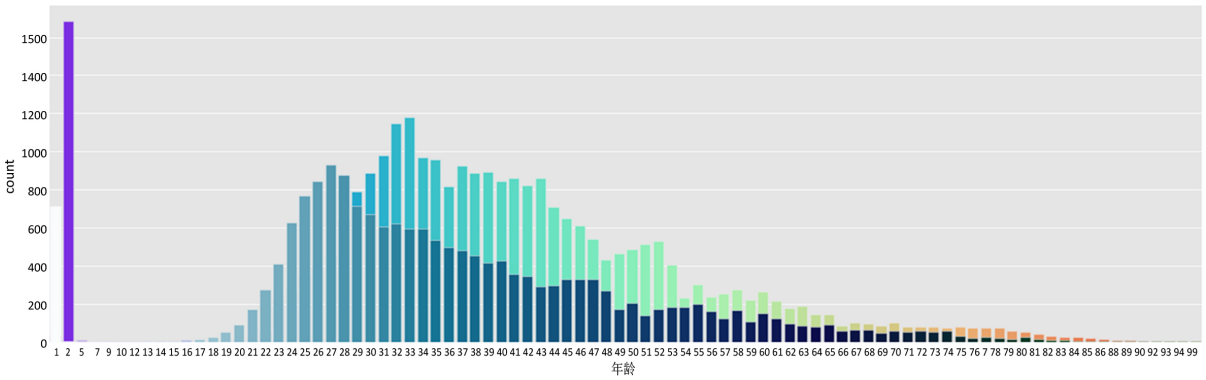

Figure 1. Distribution of ages image
图 1. 年龄分布图

Table 3. Binning and one-hot encoding of ages
表 3. 年龄值分箱以及对应独热编码

年龄区间	独热编码：[年龄_0，年龄_1，年龄_2，年龄 3，年龄_4，年龄_5]
(0,13.6]	[1, 0, 0, 0, 0, 0]
(13.6,31.1]	[0, 1, 0, 0, 0, 0]
(31.1,41.1]	[0, 0, 1, 0, 0, 0]
(41.1,52]	[0, 0, 0, 1, 0, 0]
(52,65]	[0, 0, 0, 0, 1, 0]
(65,99]	[0, 0, 0, 0, 0, 1]

采用独热编码，将一个很大权值管理一个特征，拆分成了许多小的权值管理这个特征多个表示，例如 x_1 表示原本的连续型特征(年龄)，离散化后拆分为 3 个特征 $\theta_1,\theta_2,\theta_3$，分别用权重参数 w_1,w_2,w_3 进行管理，见公式(1)，使得参数管理的更加精细，降低了特征值扰动对模型为稳定性影响。

$$p = w_1x_1 + b \rightarrow p = w_1\theta_1 + w_2\theta_2 + w_3\theta_3 + b \quad \#(1)$$

2.2.3. 非数值变量处理

某些体检指标的分类变量是离散的，比如“阴离子间隙”这一个指标中，0 代表正常，1 代表偏低，2 代表偏高，而数值大小和真实的数值大小无关，本质上属于非数字列值，同样需要对数据变量独热(one-hot)编码化处理。处理后的“阴离子间隙”这一列数据将扩充为 3 列，如表 4 所示分别是“阴离子间隙_正常”“阴离子间隙_低”“阴离子间隙_高”。

Table 4. One-hot encoding and expanded columns of taxonomic variable
表 4 分类变量独热编码扩增列

原列“阴离子间隙”分类数值	独热编码：[阴离子间隙_正常，阴离子间隙_低，阴离子间隙_高]
0	[1, 0, 0]
1	[0, 1, 0]
2	[0, 0, 1]

经过独热编码化加工处理后，提取的二元逻辑变量为 180 个。

2.3. LASSO 回归模型

LASSO (The Least Absolute Shrinkage and Selection Operator)回归是一种线性回归的缩减方法(也称正则化)，将回归系数收缩在一定的区域内[7]。LASSO 回归的特点是在拟合广义估计方程的同时进行变量筛选和复杂度调整，从而有效解决变量共线性问题并最终获得精简的统计模型。[8] Huang 等[9]利用 LASSO 从 150 多个临床指标中筛选出 24 个关键指标并以此开发并验证了影像组学联合 CT 和临床危险因素列线图模型，用于预测结直肠癌术前淋巴结转移的风险。LASSO 的主要思想是在残差平方和上添加惩罚函数来限制权重 w_1，以此来压缩和简化模型，也就是特征筛选。因此我们在 LASSO 回归中损失函数上添加一个惩罚项 $\sum_{i=1}^{n}|w_i|$，m 是样本个数，n 是特征个数，那么损失函数 $f(w)$ 是：

$$f(w)=\frac{1}{2m}\sum_{i=1}^{m}\left(y_i-\widehat{y_i}\right)^2+\lambda\sum_{i=1}^{n}|w_i| \quad \#(2)$$

LASSO 回归模型在大规模数据变量模型中具有良好的变量选择性质，当残差平方 $\lambda^{\mathrm{T}}\lambda$ 很小的时候，一些自变量的系数会随着变为 0，这样不仅可以筛选出线性回归模型中具有多重共线性的特征，还能通过量化的权重来直观地展示体检指标对健康的影响情况。

2.4. 健康指数模型

与金融风险控制领域的个人信用风险评分类似，体检健康领域同样需要评分模型的稳健性和可解释性。但是在输入数据上，体检分类指标数据和历史信用记录数据相比，具有离散化的特点，更加适合结合 LASSO 回归模型的信用评分卡。因此在个人信用风险评分模型[10]的基础上，将预测模型由梯度集成决策树改进为 LASSO 回归模型，增加评分卡的可解释性，同时结合体检的应用场景和输入数据，建立体检评分模型。评分卡的分值刻度将分值 Score 表示为比率对数的线性表达式：

$$\text{Score}=A-B*\log(\text{odds}) \quad \#(3)$$

其中，A 为补偿，B 为刻度，都为常数。

$$\text{odds}=\frac{p}{1-p} \quad \#(4)$$

$$p=w_1X_1+w_2X_2+w_3X_3+\cdots+w_nX_n+b=W^{\mathrm{T}}X+b \quad \#(5)$$

其中 X_n 表示体检者各项检测指标，p 由 LASSO 线性回归模型得到，取值范围[0, 1]，表示非健康的概率。引入违约翻倍系数 PDO，即当 odds (非健康：健康比例)为两倍的时候，分数为 Score + PDO：

$$\text{Score}+\text{PDO}=A-B*\log(2*\text{odds}) \quad \#(6)$$

需要计算参数 A，B，解两个方程，得到：

$$B=\frac{\text{PDO}}{\ln(2)} \quad \#(7)$$

$$A=\text{Score}_0+B*\ln(\text{odds}_0) \quad \#(8)$$

odds_0 为基准坏/好比例，这里取非健康/健康比例，得到 $\text{odds}=\frac{19832}{25543}=0.776416$。为了将大部分的分数限制在 100 分以内，$\ln(\text{odds}_0)=-0.253066$ 设置基准分 $\text{Score}_0=70$，$\text{PDO}=20$。代入公式(7) (8)得到 $A=62.698, B=28.854$。

3. 结果

3.1. 训练结果

LASSO 回归模型的训练需要调整正则化参数 alpha。正则化参数越高，模型适应数据的复杂性能力越低，灵活程度越低，出现欠拟合的情况。当正则化参数越小时，模型过拟合。本文使用 scikit 模块的 Lasso CV (Cross Validation)，在 10 折交叉验证中找出最佳的 alpha = 0.0003433。

训练后的最优模型选取了 32 个体检指标，我们定义该 32 个体检指标为健康指数影响因子，同时淘汰了 150 个权重较低的体检指标，图 2 展示了模型权重中 10 个最重要的非健康特征和 10 个最重要的健康特征。

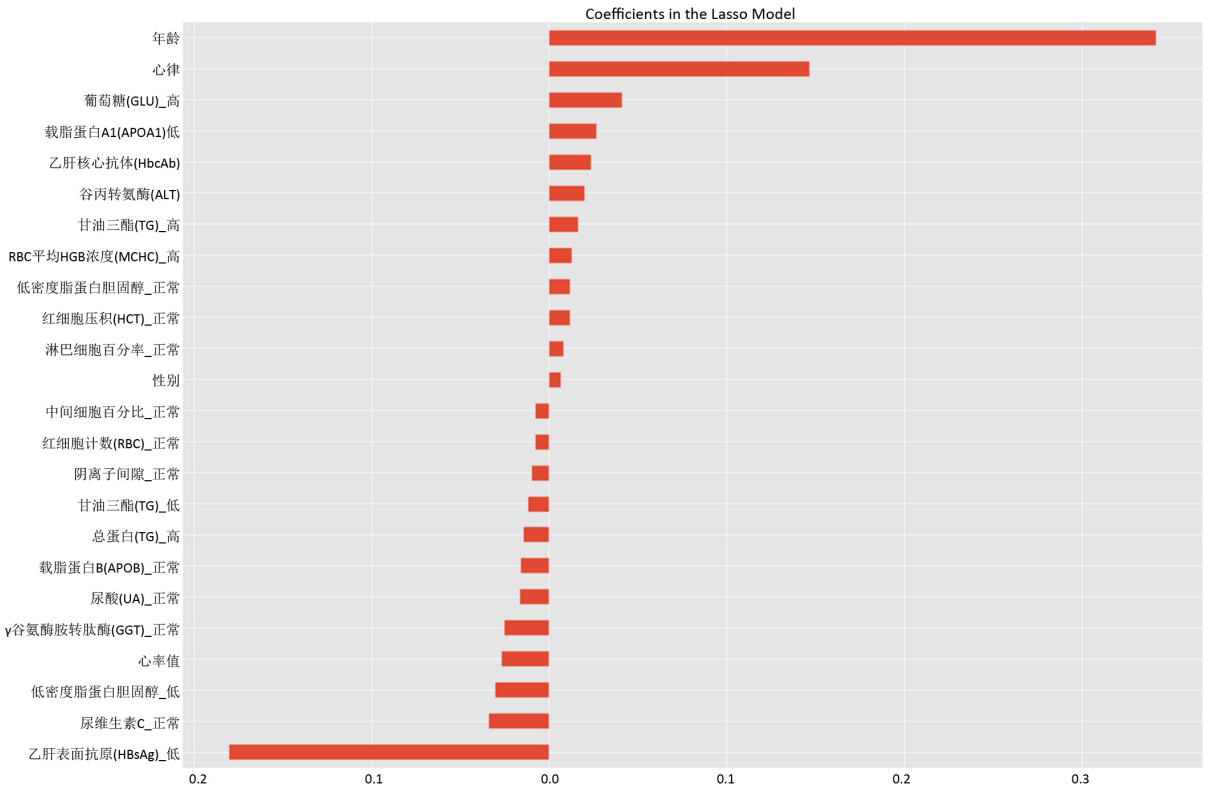

Figure 2. Coefficients histogram of Lasso model
图 2. 体检指标权重直方图

图 2 中非健康特征中影响最大的是年龄的影响，年龄越高，评分越低。其它的非健康特征基本属于异常指标，比如“葡萄糖_高”“载脂蛋白 A1_低”“谷丙转氨酶”和“甘油三酯”，这些指标和常见慢性病有关[11]。表 5 中举 3 个用户的例子，为了简化表格，每组对比用户中未列出的体检项目指标变量相同。

Table 5. Physical examination item record and health score prediction of sample users
表 5. 示例用户体检项目记录和健康分预测表

用户	性别	年龄	γ 谷氨酰胺转肽酶(GGT)	中性粒细胞绝对值	低密度脂蛋白胆固醇	平均 RBC 血红蛋白量	总胆固醇(CHOL)	淋巴细胞百分率	甘油三酯	碱性磷酸酶(ALP)	载脂蛋白B(APOB)	阴离子间隙	健康分
A	男	45	阳性	阴性	高	高	高	阴性	高	阴性	阳性	阳性	53.5
B	女	47	阴性	阳性	低	高	正常	阳性	低	阴性	阴性	阴性	72.8
C	女	48	阴性	阳性	正常	正常	正常	阴性	正常	阳性	阴性	阴性	67.0

Table 6. Coefficients of three physical examination items
表 6. 三个体检项目的指标权重

体检指标	权重
γ 谷氨酰胺转肽酶(GGT)_阴性	−0.0358
γ 谷氨酰胺转肽酶(GGT)_阳性	0
低密度脂蛋白胆固醇_正常	−0.0022
低密度脂蛋白胆固醇_低	0.0036
低密度脂蛋白胆固醇_高	0
甘油三酯(TG)_正常	0
甘油三酯(TG)_低	−0.0312
甘油三酯(TG)_高	0.0153

从表 5 中看出，年龄相近的体检用户之间，由于体检项目指标的差别，预测健康分在相邻的区间中，体现了模型的稳健性。表 6 是三个体检项目中不同指标的权重，正值代表非健康影响影响因素，负值代表健康影响因素，绝对值越大影响越大。通过量化的指标权重，揭示不同的体检指标对健康的正负面影响，具体来看，用户 A 的三项异常指标导致影响健康分较低，同时表 7 的疾病诊断结果为“脂肪肝”也印证了三项指标和健康之间的联系。

Table 7. Diagnostic results for the sample users
表 7. 示例用户的诊断结果表

用户	疾病诊断
A	脂肪肝
B	无
C	无

表 8 的中两个不同年龄段的用户体现评分模型的区分度，其中 E 用户处于中老年，同时如表 9 所示，该用户体检指标中的异常指标：“乙肝核心抗体(HbcAb)”“尿维生素 C”和“阴离子间隙”属于权重绝对值较大指标，因此健康分低于该年龄区间的平均分数 29.5 分。同时 D 用户的异常指标：“低密度脂蛋白胆固醇”和“淋巴细胞百分率”在模型中的权重绝对值相对较小，因此健康分高于同一年龄段的平均健康分 78.9。所以健康分模型不仅在同一个年龄段之间，在不同年龄段之间也可以客观地反映出用户的健康状况。

Table 8. Physical examination item record and health score prediction of sample users
表 8. 示例用户体检指标及健康分

用户	性别	年龄	乙肝核心抗体(HbcAb)	低密度脂蛋白胆固醇	尿维生素 C	平均 RBC 血红蛋白量	淋巴细胞百分率	阴离子间隙	健康分
D	女	26	阴性	低	正常	正常	阳性	阴性	85.9
E	女	73	阳性	正常	异常	高	阴性	阳性	23.6

Table 9. Coefficients of physical examination items
表 9. 指标权重

体检指标	权重
乙肝核心抗体(HbcAb)_阳性	0.0283
尿维生素 C_异常	0.0364
阴离子间隙_阳性	0.0201
低密度脂蛋白胆固醇_低	0.0036
淋巴细胞百分率_阳性	0

3.2. 模型评估和比较

3.2.1. 模型评估

根据图 1 我们发现年龄分布不符合正态分布，显然年龄也不符合正态分布，我们对其余 30 个健康指数影响因子逐个作图分析发现他们基本上符合正态分布。为解决多随机因素作用的问题，法国数学家棣莫弗和拉普拉斯首先提出了中心极限定理并给出了证明[12] [13] [14]，该定理表明所研究的随机变量如果是有大量独立的而且均匀的随机变量相加而成，那么它的分布将近似于正态分布。下图展示健康指数模型输出的用户评分分布直方图。线性回归模型中的假设前提是因变量 p 服从正态分布，而在图 3 中，分数特别低(低于 30)和特别高(高于 100)的人占比都较少，大多数健康分数中等(75 左右)，整体上基本符合正态分布，符合线性回归模型的先验假设。

表 10 显示健康分数分值越高，非健康率越低，体现了体检者可以通过分数的高低进行健康状况的评估。

3.2.2 模型比较

对比运用 PCA (主成分分析) [15]的统计方法，因获得的主成分公共因子是实际自变量因子的线性组合，所以其难以与分析健康因子的实际问题相对应。特征变量较少的样本适合使用 PCA 进行降维，因其对公共因子有更好的解释性，本文的体检指标数量较多(182 个)，不适合使用 PCA 分析。

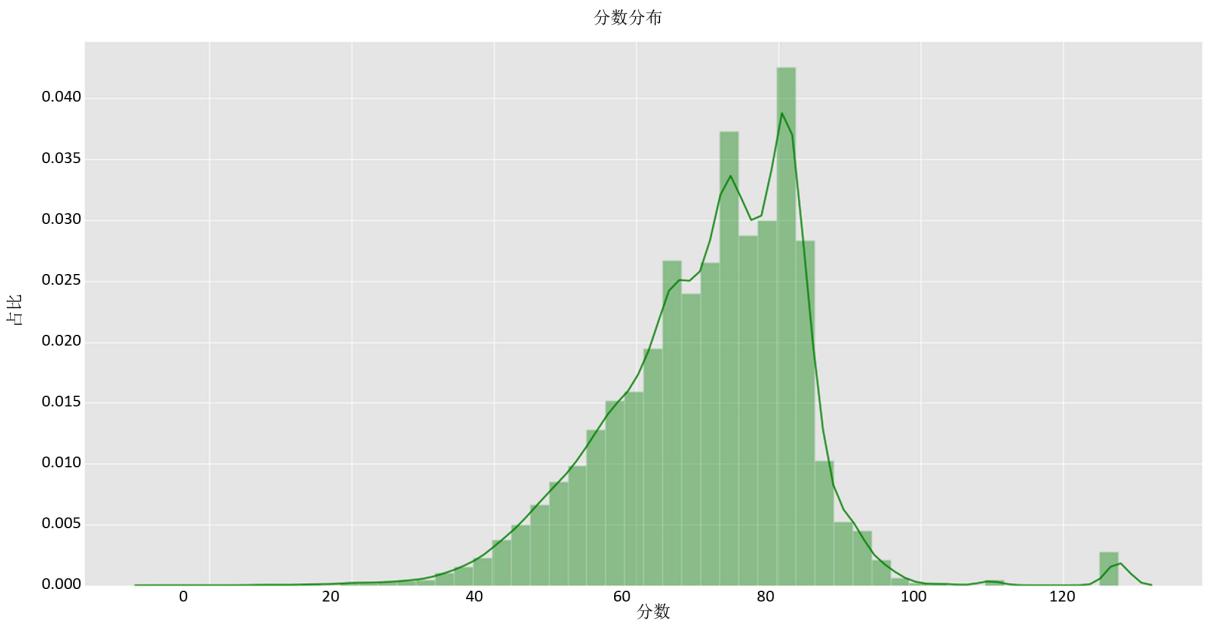

Figure 3. Histogram of score distribution
图 3. 分数分布直方图

Table 10. Interval statistics of health scores
表 10. 健康分区间统计表

分数分段	分段人数	分段人数百分比	累计人数百分比
0~10	32	0.07%	0.07%
10~20	135	0.30%	0.37%
20~30	516	1.14%	1.51%
30~40	1109	2.44%	3.95%
40~50	1947	4.29%	8.24%
50~60	4627	10.20%	18.44%
60~70	10398	22.92%	41.35%
70~80	16340	36.01%	77.36%
80~90	9420	20.76%	98.12%
90~100	382	0.84%	98.97%
100~110	2	0.00%	98.97%

对比应用 XGBoost 中的回归模型，模型评价指标为预测值的均方误差(MSE)，其中 n 为样本的个数，y_i 是真实数据，$\widehat{y_i}$ 是预测值。MSE 越小，说明预测模型描述实验数据具有更好的精确度。

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}w_i\left(y_i-\widehat{y_i}\right)^2 \quad \#(9)$$

经过 10 折交叉验证，最后得到不同机器学习模型使用均方误差的评分如表 11 所示。

Table 11. Scores for different machine learning models
表 11. 不同机器学习模型评分

模型	均方误差
LASSO 回归模型	0.23480
岭回归模型	0.23496
梯度提升决策树	0.23501
随机森林回归	0.26548
支持向量回归	0.29774

其中 LASSO 回归模型在数据集中的准确率略优于其他模型，与之相近的梯度提升决策树只能对变量进行重要性排序[16]，并不能输出非健康概率的最终计算权重系数，而 LASSO 回归中模型的实际系数即代表健康/非健康权重，可解释性更好。

4. 结论

本文基于体检数据，在个人信用评分卡模型基础上提出了一种基于 LASSO 回归的健康指数模型。通过模型的对比，本文使用的 LASSO 模型的变量压缩效果可以很好地筛选自变量，兼具变量子集选择和岭回归的优点，因此可以兼顾解释性和预测准确率，在个人信用评分中已经得到广泛应用。在和随机森林[17]等更加复杂的机器学习模型的对比中，LASSO 回归模型的预测精确率在体检大数据集上表现的更好，同时该模型可以更好地解释体检指标特征对健康评分的影响。实验结果表明该健康指数模型大体上呈现

正态分布，符合线性回归模型的先验假设。通过该健康指数，体检用户可以从宏观整体上直观感知个人身体健康状况水平，为长期健康管理提供一个可以参考的量化指标，降低体检用户同医生的沟通成本，督促用户更加关注身体整体健康状况水平。

本文提出的健康评分指数建模中体现了区分度和稳健性，但是本文仍存在几方面问题：一是先验假设中的体检健康标准科学性问题，是否可以由疾病诊断分类中得到更加权威的体检健康标准来改进二值化设定；二是健康影响因子中各因子独立性问题，需要进一步分析加以筛选。如何更加深入地将数据科学和体检健康科学结合起来优化评分模型，是下一步研究的重点。

基金项目

国家级大创项目基金支持，编号 202010384213.

参考文献

[1] 中共中央 国务院印发《“健康中国 2030”规划纲要》[J]. 中华人民共和国国务院公报, 2016(32): 5-20.

[2] 国务院办公厅. 关于促进和规范健康医疗大数据应用发展的指导意见(国办发[2016]47 号) [Z]. 2016.

[3] 叶荔姗, 赵飞, 陈坚, 徐秋实, 许志坚. 基于智能电子健康档案平台的大数据应用研究与实践[J]. 中国卫生信息管理杂志, 2019, 16(6): 672-676.

[4] 熊辉, 何振峰. 基于 R 平台的体检数据分析研究[J]. 福建电脑, 2017, 33(11): 73-75.

[5] Wang, L., Wang, Y., Chen, Y., Liu, C. and Fan, X. (2017) Prediction of Lymphocytosis Using Machine Learning Algorithm Based on Checkup Data. 2017 4*th International Conference on Systems and Informatics* (*ICSAI*), Hangzhou, 11-13 November 2017, 649-654. https://doi.org/10.1109/ICSAI.2017.8248369

[6] 余秋燕, 赵莹, 孙继佳, 邵建华. 典型机器学习算法在脂肪肝分类预测研究中的实现与比较[J]. 数理医药学杂志, 2019, 32(1): 1-3.

[7] 方匡南, 章贵军, 张惠颖. 基于 Lasso-Logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济研究, 2014, 31(2): 125-136.

[8] 李阳, 陈晓泓, 王一梅, 胡家昌, 沈子妍, 沈波, 林静, 丁小强. 基于 LASSO 变量选择联合贝叶斯网络构建恶性肿瘤相关急性肾损伤(AKI)风险预测模型[J]. 复旦学报(医学版), 2020, 47(4): 521-530.

[9] Huang, Y.Q., Liang, C.H., He, L., *et al.* (2016) Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. *Journal of Clinical Oncology*, **34**, 2157-2164. https://doi.org/10.1200/JCO.2015.65.9128

[10] 韩修龙. 基于 XGBOOST 的用户信用评分建模[J]. 电脑知识与技术, 2018, 14(5): 7-8.

[11] 贾瑞珍, 杜兵. 健康体检的深层价值探讨(附 1300 例体检结果分析) [J]. 中国全科医学, 2007(1): 58-59.

[12] 孟祥飞, 王瑛, 李超, 亓尧, 孙赟. 独立不同分布不确定变量中心极限定理证明及其应用[J]. 上海交通大学学报, 2019, 53(10): 1230-1237.

[13] Dolgopyat, D. and Goldsheid, I. (2018) Central Limit Theorem for Recurrent Random Walks on a Strip with Bounded Potential. *Nonlinearity*, **31**, 3381. https://doi.org/10.1088/1361-6544/aab89b

[14] Benoist, Y. and Quint, J.-F. (2016) Central Limit Theorem for Linear Groups. *The Annals of Probability*, **44**, No. 2. https://doi.org/10.1214/15-AOP1002

[15] 缪柏其, 宁静, 肖婕. 主成分分析和因子分析在体检数据分析中的应用——中国科技大学高级知识分子健康状况及影响因素分析[J]. 数理统计与管理, 2000(6): 16-19.

[16] 王小强. 基于随机森林的亚健康状态预测与特征选择方法研究[J]. 计算机应用与软件, 2014, 31(1): 296-298, 307.

[17] 张占林, 孙勇, 妥小青, 叶勒丹·马汉, 龚政, 田恬, 陈珍, 古丽斯亚·海力力, 戴江红, 姚华. 随机森林算法对体检人群糖尿病患病风险的预测价值研究[J]. 中国全科医学, 2019, 22(9): 1021-1026.