基于在校数据挖掘的大学生心理抑郁预测模型以及分析

□林靖怡　吴平鑫　郑翊佳　刘玉盈　周燕（通讯作者）　华南农业大学数学与信息学院

【摘要】　当代大学生面临各方面压力，容易造成心理问题，为了让学生和学校做好预防心理抑郁的工作，本文提出了基于在校数据的学生抑郁预测模型。文用网络问卷的形式收集学生在校数据，包括基本信息、学业情况、消费行为、人际关系、运动情况、职业规划及睡眠情况等七个方面，并让学生完成 SCL-90 抑郁分量表作为判定学生心理状态的指标，分别利用随机森林 – 有序多分类 logistic 回归和 SMOTE-Xgboost 算法建立大学生抑郁预测模型。对比后发现 SMOTE-XGBoost 算法建立的模型效果最好，证实用在校数据来预测学生抑郁倾向是可行的，并针对影响抑郁程度的因素提供了相应的学生心理健康管理建议。

【关键词】　抑郁预测　在校数据　XGBoost　SMOTE

一、问题背景

随着抑郁症走进大众视野，大学生抑郁症的问题也逐渐披露。在上大学之前，大部分学生往往只关注学习，忽略了心理上的成长；上了大学之后，从学校走进一个小型社会，当学习从填鸭式变成自主式，生活从依赖到独立，不善处理的情感问题，需要处理宿舍人际关系的问题，后知后觉的就业压力，这些不可避免的处境都让大学生无所适从。对大学生短时间内在各方面快速成长的要求，让大学生的压力与日俱增，容易导致大学生焦虑甚至是抑郁。

本文着力于探索影响大学生心理抑郁的因素，预测出学生的抑郁倾向。鉴于大学生大部分时间都是在校内活动，因此本文通过调查学生在校情况，找出明显影响心理抑郁程度的因素，以供学生参考，关注自身心理变化，及时调整，保持健康的身心发展。

同时，学校也可以参考关注学生状态，以便尽早介入，及时调整学生心理状态，避免悲剧发生，培育更多有贡献力的人才。

二、数据介绍

本文通过网络问卷形式收集华南农业大学数学与信息学院学生的个人情况，最终收集问卷 301 份。问卷共设 43 个问题（其中包含 3 个可跳答问题），其中前 30 题涉及基本信息、学业情况、消费行为、人际关系、运动情况、职业规划及睡眠情况等七个方面；最后 13 个问题来自于症状自评量表 SCL-90 中抑郁分量表，作为最后评定学生的心理状态。抑郁分量表中每个项目采取 5 级评分制，包括没有、很轻、中度、偏重、严重等选项，最后，计算均值作为症状指数的分数。

如果症状指数在 3 分以上，表明被试抑郁程度较强；症状指数在 2 分以下，表明被试抑郁程度较弱。

三、基于数据挖掘算法的大学生抑郁预测建模

3.1 随机森林算法 -logistic 回归

为探究影响大学生心理抑郁程度的因素并进行相应的抑郁程度预测，本节利用随机森林及有序多分类 logistics 回归建立大学生抑郁预测模型。对最后抑郁分量表的症状指数进行划分，作为因变量 Y，建立三分类模型。其中 Y 变量的定义如下：

$$Y=\begin{cases}1\ \text{无抑郁症状（症状指数在 2 以下）}\\ 2\ \text{有抑郁倾向(症状指数在 2~3 以内)}\\ 3\ \text{有抑郁症状(症状指数在 3 以上)}\end{cases}$$

由于特征较多，样本量较少，需要先进行特征筛选，将前面 30 题与因变量 Y 进行随机森林的回归树建模，得到随机森林模型变量重要性结果，逐个对应的特征分别为：年级、日常运动强度、一般睡眠质量、失恋后需要多久时间恢复、消费类型、平时与同班同学的关系、是否会因不明确毕业后的方向而苦恼烦躁、生活费是否充裕。

在回归过程中，基于已进行筛选的特征，我们再剔除了不显著的特征，最后进行回归的

由回归结果可以得到：

$$\text{logit}[p(y\le 1)]=\text{logit}(p_1)$$
$$=-0.0619-0.21289*x_{13-2}+0.71364*x_{13-3}-0.47913*x_{29-2}$$
$$-1.28720*x_{29-3}+0.97144*x_{26-1}-0.65454*x_{26-2}+2.19724$$
$$*x_{16-2}+0.98112*x_{16-3}-0.08247*x_{16-4}-0.34367*x_{16-5}$$
$$\text{logit}[p(y\le 2)]=\text{logit}(p_1+p_2)$$
$$=2.4553-0.21289*x_{13-2}+0.71364*x_{13-3}-0.47913*x_{29-2}$$
$$-1.28720*x_{29-3}+0.97144*x_{26-1}-0.65454*x_{26-2}+2.19724$$
$$*x_{16-2}+0.98112*x_{16-3}-0.08247$$

表 1　回归预测结果

	预测结果			
实际结果	1	2	3	总计
1	84	42	3	129
2	35	91	2	128
3	7	29	8	44
总计	126	162	13	301

最后由预测结果表明，无抑郁症状的共有 129 个，预测正确的有 84 个，准确率为 65%；有抑郁倾向的共有 128 个，预测正确的有 91 个，准确率为 71%；有抑郁症状的共有 44 个，预测正确的有 8 个，准确率为 18%。考虑到样本中有抑郁症状的样本较少，因此对于有抑郁症状的预测结果准确率较低，

无抑郁症状与有抑郁倾向的结果预测准确率不是很高，综合考虑模型的总体预测效果一般。

3.2 SMOTE-Xgboost 算法

为探究影响大学生心理抑郁程度的因素并进行相应的抑郁程度预测，本节利用 smote 和 Xgboost 算法建立大学生抑郁预测模型。

特征来自前 30 题的信息，对最后抑郁分量表的症状指数进行划分，作为因变量 Y，建立二分类模型。其中 Y 变量的定义如下：

$$Y=\begin{cases}1 \text{ 有抑郁症状（症状指数在 3 分以上）}\\0 \text{ 无抑郁症状（症状指数在 3 分以下）}\end{cases}$$

在收集的 301 份问卷中，正负类样本的比例是 44:257，样本分布不均匀，因此使用 SMOTE 算法对少数类样本进行并根据少数类样本人工合成新样本添加到数据集中。使用 SMOTE 处理后数据点增加到 354，正负类样本比例为 1:1。

通过 30 个特征值与因变量 Y（有无抑郁症状）用 Xgboost 算法进行二分类建模。首先将数据分为训练集和测试集，比例为 7:3，得到结果为：准确率 94.7%，召回率 91.5%，F1 值 93.1%，模型整体有效。从特征重要性图（这里仅展示前 10 个）可以看到，对大学生抑郁程度影响最大的特征是日常运动频率（X23）和日常运动强度（X24），其次是 X2、失恋后需要多久时间恢复（X19）和是否有过自残行为（X22），剩下依次是平时有无暴饮暴食（X21）、一般睡眠时长（X28）、消费类型（X13）、平时有无节食（X20）。

四、 结论与建议

本文利用随机森林 – 有序多分类 logistic 回归建立的抑郁预测模型效果一般，但也给学生和校方提供了预防抑郁需要注意的方面：消费类型、平时与同班同学之间的关系、因为毕业后方向不明确而产生的负面情绪、平时睡眠质量。

本文通过 smote-Xgboost 算法建立的大学生抑郁预测模型整体有效，准确率高达 94%，其中对因变量（有无抑郁症状）影响较大的因素有运动习惯、情绪调节能力、饮食习惯、睡眠时长及消费类型等等。

对于大学生来说，如果发现自己这几个方面情况糟糕，应该要考虑去就医检查，及早发现及早治疗。相应地，大学生自己也要注意预防抑郁症，应该养成良好的运动习惯、饮食习惯和消费习惯，适当运动有助于调节心情，恢复身体精神，合理饮食方能维持身体正常状态；注意培养自己的情绪调节能力，找到合适的情绪发泄方式，及时调整心态，保证足够的睡眠时长。

对于学校来说，班主任要多注意学生情况，提供正确引导。学校的干预从预防开始会更好，举办一些教导学生心理调节的文体活动，拥有大量学生关注的官方公众号也可以推送一些相关科普内容，给学生打预防针。

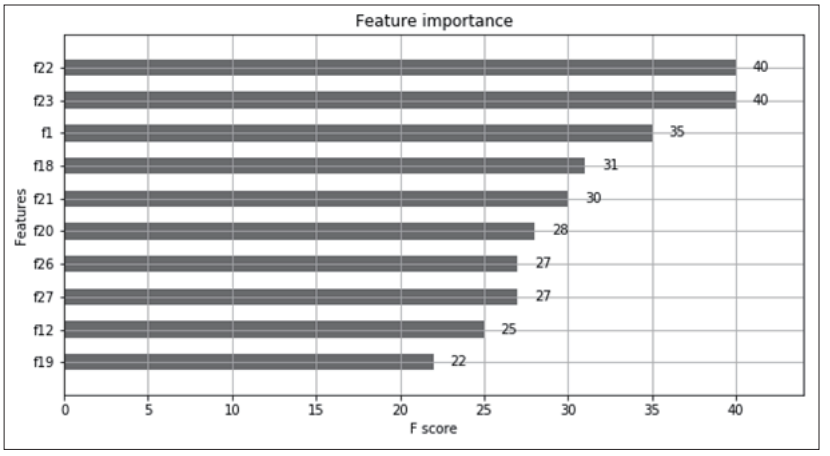

图 1　特征重要性图

参 考 文 献

[1] 陈树林 , 李凌江 .SCL-90 信度效度检验和常模的再比较 [J]. 中国神经精神疾病杂志 , 2003(05):323-327.

[2] 韦杰 , 曾萍 . 基于 R 的有序分类资料 logistic 回归分析 [J]. 软件 ,2014,35(06):56-57+61.

[3] 何颖 , 季浏 . 不同的体育锻炼类型对大学生抑郁水平的影响及其心理中介变量 (Body-esteem) 的研究 [J]. 体育科学 ,2004(05):32-35+52.

[4] 萧文泽 , 周辉 , 夏阳 , 袁晶 , 严青 . 大学生抑郁状况及其危险因素的研究 [J]. 中国行为医学科学 ,2006(07):647-649.

2019 年省级大学生创新训练项目《基于决策树与 5Level 算法的社交数据挖掘的心理健康预警建模及学生管理研究》项目编号 201910564128

林靖怡 (1999)，女，广东高要，本科，主要研究方向为数据挖掘、文本挖掘。

吴平鑫 (2000)，男，山西运城，本科，主要研究方向为 Python 语言应用研究。

郑翊佳 (1998)，男，广东广州，本科，主要研究方向为数据分析、数据挖掘。

刘玉盈 (2000)，女，广东中山，本科，主要研究方向为运筹学、数据挖掘。

周燕（1980-），女，汉，广西桂林硕士研究生，华南农业大学数学与信息学院数学系教师，讲师，主要研究方向金融统计与数据挖掘。