食堂消费大数据可以精准识别贫困生吗？

——基于本科生行为数据、行政数据和问卷数据的实证研究

张存禄，马莉萍，陈晓宇

[摘　要] 基于某“双一流”建设大学本科生食堂消费大数据、助学金发放的行政数据以及家庭基本信息的调查数据，利用统计学习模型对基于食堂消费大数据识别家庭经济困难学生的精准性进行了估计。研究发现：食堂消费大数据对家庭经济困难学生的识别精准率仅能达到约 60%，采用更精细的消费时间序列数据可以将识别精准率提高到约 65%，进一步结合家庭基本信息的问卷调查数据则可以将识别精准率提高到约 92%。相比传统的逻辑回归，采用提升树和支持向量机等判别模型可以提高模型对家庭经济困难学生的识别能力。本文的研究发现说明，仅仅利用食堂消费数据作为补助依据的精准度仍有待提高，应将食堂消费数据与学生家庭基本信息数据相结合来提升资助精准度。

[关键词] 家庭经济困难学生；贫困生；学生资助；精准资助；校园大数据

一、研究背景

20 世纪 90 年代末，伴随高校学费上涨以及成本分担的施行，高校的学生资助体系随之建立，资助力度和资助范围也在逐年扩大。以面向家庭经济困难学生的助学金为例，2019 年各类助学金金额达 378.29 亿元，资助学生达到 1130.66 万人次(全国学生资助管理中心，2020)。如此巨额的助学金能否切实发挥作用，在很大程度上取决于资助对象是否为那些真正家庭经济困难的学生。有研究发现，很多发展中国家的学生资助体系均存在难以识别家

[收稿日期] 2021－10－21
[基金项目] 国家自然科学基金青年项目(72104010)。
[作者简介] 张存禄，北京大学教育学院，电子邮箱地址：zhangcl@pku. edu. cn；马莉萍(通讯作者)，北京大学教育经济研究所，电子邮箱地址：lpma@pku. edu. cn；陈晓宇，北京大学教育经济研究所，电子邮箱地址：xychen@gse. pku. edu. cn。

庭经济困难学生的问题(Johnstone，2003)。从中国的实际情况来看，由于缺少一套利用税收数据反映真实家庭收入的认定体系，学生资助的瞄准责任落在了高校，即由高校来收集学生的资助需求以及家庭收入的相关信息，以此作为是否给予资助的判断标准。然而，学生资助中城乡居民家庭经济状况复杂难以认定、地方行政部门对贫困家庭审核不严格、学生担心被“瞧不起”而不愿申请等问题一直存在(徐丽红，2015)。不同类型资助的精准程度也不尽相同，政府层面的学生资助瞄准效果较好，但是学校层面和社会捐赠类学生资助的瞄准效果则相对较差(Loyalka et al.，2012)。

2017年由财政部等四部门发布的《关于进一步落实高等教育学生资助政策的通知》明确提出，为进一步提高高校学生资助的精准度，高校等培养单位要逐步建立学生资助数据平台，融合校园卡等信息，为家庭经济困难学生认定提供支撑。2018年教育部等六部门发布的《关于做好家庭经济困难学生认定工作的指导意见》提出，认定家庭经济困难学生是实现精准资助的前提，是做好学生资助工作的基础。各地、各校要把家庭经济困难学生认定作为加强学生资助工作的重要任务，切实把好事做好、实事办实。学校可采取家访、个别访谈、大数据分析、信函索证、量化评估、民主评议等方式提高家庭经济困难学生认定精准度。

2019年《科技日报》发表了名为《用大数据发餐补“饱”暖学生心》的评论文章，文中介绍了某大学运用大数据计算识别家庭经济困难学生，并尝试将用餐补助直接发放到学生饭卡中。该校将每月在学校用餐60次以上、每天吃饭花销低于平均值8元作为认定家庭经济困难学生的统一标准(杨仑，2019)。该新闻一出立刻在高校乃至社会各界引发热议，褒贬不一，赞成者为高校以人为本实施学生资助的举措而鼓掌，反对者则为食堂消费数据的识别精准度而担忧。

相关研究表明，本科生的消费习惯受家庭背景等多方面因素影响，在校期间获得的经济资助仅对短期内的消费水平有提升作用，对长期的消费行为没有显著影响(张存禄等，2021)。因此，通过校内消费行为信息识别的家庭经济困难学生可能具备方法的可行性和结果的稳定性。但是，目前国内关于通过校园大数据识别家庭经济困难学生的实证研究较少，聚焦于实现方法和有效性的研究更加缺乏，根据校内消费行为信息是否可以有效地判定家庭经济困难学生，如何建立有效的识别模型，是值得研究的问题。

因此，本文以某“双一流”A类建设高校作为研究对象，利用统计学习模型，分析学生的校园消费行为时间序列数据、历年新生入学前问卷调查数据、

贫困生认定和助学金发放的行政管理数据，尝试对基于消费行为数据认定家庭经济困难学生的精准性进行实证检验，以期为精准资助提供借鉴。

二、文献综述

精准资助依赖对资助对象的精确瞄准，应同时具备较低的遗漏率(应该资助的学生未获得资助)和泄漏率(不应资助的学生却获得资助)。对 100 所高校 6059 名本专科学生的调查发现，助学金在贫困生中的覆盖率仅为 47%，而非贫困生得到补助的比例却达到 57%，助学金的标的错误率高达 64%。即便实施了奖助学金后，仍然有 79%的贫困生没有脱贫，农村地区学生的贫困率仍然高达 25%(吴斌珍等，2011)。对首都高校研究生贫困资助政策的研究发现，贫困生获得贫困资助的可能性并不比非贫困生显著更高，成绩排名靠前、工科专业的贫困学生更有可能获得资助，这意味着贫困生资助成为了变相的奖优资助，会导致大量成绩不佳的贫困生无缘获得资助(杨钋和刘霄，2019)。

一些研究试图通过问卷调查的方法来确定识别家庭经济困难学生的变量。如利用安徽省 16 地市高校学生问卷调查数据的分析发现，家庭基本情况、家庭所在地、家庭年收入以及家庭负债是识别家庭经济困难学生的有效变量(宋俊秀，2017)。对上海市 5 所高校学生问卷调查数据的方差分析结果显示，家庭经济困难学生与其他学生在生源地、勤工助学时间、校内消费水平、娱乐消费、社交等方面存在显著差异(郑杰，2015)。对上海和山西 6 所高校学生问卷调查数据的分析发现，可以通过公共服务可得性、住房条件、父母职业类型、家庭固定资产、电子消费品等建立消费指数，并以此对家庭经济困难学生进行识别(田志磊和袁连生，2010)。

随着“大数据”时代的到来，涌现出一批利用大数据统计学习方法来提高贫困认定精准度的研究。如有研究基于已有贫困识别体系，通过优化构建出适用于相对贫困识别的指标体系，利用鄂西贫困县农户调查数据进行实证检验，减少“错进”和“漏评”的比例(陈志，2021)。基于提升树中的 XGBoost 算法，通过选取 14 维特征构建判别模型，利用贫困户建档立卡数据，对返贫人口进行识别，准确率达到 96.81%(李春雷等，2021)。基于 Lasso 稳健马田系统，通过构建相对贫困识别模型，对具备“不平衡性”和“比较性”的贫困数据进行识别，精准度高于马田系统和其他传统分类方法(陈闻鹤等，2022)。

同时，将数据挖掘技术用以分析大学生行为数据的研究也日益增多。例如，利用大学生课堂面部表情数据，使用愉悦、惊讶、难过、轻视等情绪的

出现频率，建模预测学生该学年的课程成绩，模型解释方差约76.4%(李德淇等，2018)。利用大学校园卡记录和学生成绩数据，使用校内消费时间、门禁刷卡时间和图书借阅数量等特征变量，建模预测学生该学年的课程成绩(刘譞，2017)。还有研究利用在线学习记录数据预测学生是否会中途辍学，模型ROC曲线下面积约0.617(Harrell and Bower，2011)。预测学生是否可以获得学习证书，模型准确率超过90%(蒋卓轩等，2015)。这类研究尽管使用的数据和预测的目标不尽相同，但多以经典统计学习算法作为研究基础，通过统计分析、聚类等方式提取包含主要有效信息的特征变量，建立识别模型，模型的预测准确率一般取决于特征变量对目标问题的解释度。

长时段的大学生食堂消费数据也属于“大数据”的范畴，学生的消费次数、消费金额等不仅反映了消费习惯，也在一定程度上反映了家庭经济水平。出于这一考虑，一些高校开始利用食堂消费数据作为给学生发放补助的依据。如对某理工类高校本科生食堂消费数据的分析发现，家庭经济困难学生在食堂用餐次数、平均用餐金额、用餐金额波动等方面与其他学生存在差异，并以此建立技术指标对家庭经济困难学生进行划分(吴朝文，2016)。但总体来说，这方面的研究还较少，且多采用家庭经济困难学生与其他学生在不同特征指标上的差异作为切入点，并以存在显著差异的特征指标为基础建立识别指标对家庭经济困难学生进行认定，很少对识别模型建立方法和有效性进行更加深入的探讨。本文尝试利用经典统计学习模型建立家庭经济困难学生识别模型，采用接受者操作特征曲线等技术手段对不同建模方式的有效性进行评价和比较，以期弥补现有研究的不足。

三、研究设计

(一)研究对象

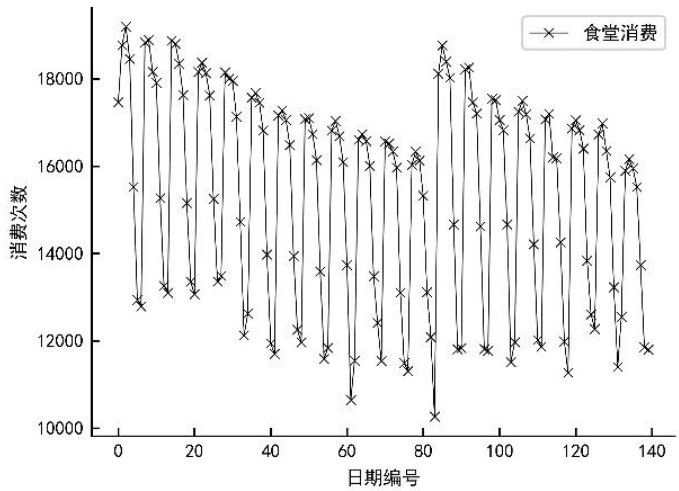
本文以某所“双一流”建设A类高校四届共10976名本科学生作为研究对象，其中包含1670名被学校认定的家庭经济困难学生，约占学生总数的15%。该校本地生源比例不高，绝大多数学生住校，加之对这些学生消费数据的采集是在5年前，外卖行业对校园的渗透远不如今，因此食堂仍然是他们餐饮消费的主要组成部分，仅有极少数家庭经济条件较好的学生会选择校外饭店就餐作为补充。为避免假期部分学生返乡对消费行为的干扰，研究选取了不含假期的20周作为采样区间，共计2143272条食堂消费数据。

在数据分析过程中发现，学生存在一次就餐多次刷卡的情况，为了避免

多次刷卡导致的就餐次数和平均就餐金额偏差，本文将 4：00－10：30、10：30－15：30、15：30－20：30、20：30－24：00 分别定义为早餐、午餐、晚餐、夜宵四个就餐区间，并将学生同一天中同一就餐区间内的消费记录合并为一次消费行为。①

为了刻画在校本科生的就餐行为特征，本文选取了三个维度的统计指标，并希望以此对家庭经济困难学生进行有效识别：一是“在校就餐比例”指标。一般来说，由于校内食堂的消费水平远低于校外餐厅，因此家庭经济困难学生会更倾向于在校内食堂就餐。二是“餐均消费金额”指标。本科生的消费习惯受多年家庭环境的影响，控制性别信息后，平均消费金额对家庭经济困难学生有识别作用。三是“经济型食堂就餐比例”指标。校内有部分食堂消费相对廉价，家庭经济困难学生选择在这类食堂就餐的可能性相对更高，根据食堂消费数据中关于消费位置的信息，可以对其进行统计。

图 1 展示了 140 天采样区间内该校本科生每日食堂消费次数和平均消费金额分布情况。② 可以看出，周一至周四的消费情况较为接近，因此近似地将这四天的消费次数、平均消费金额等指标合并为一个统计量。类似地，周六、日的消费情况也可以一起统计，周五的消费情况介于二者之间，这里为尽量保留有效信息单独进行了统计。



① 10：30、15：30、20：30 为该校食堂就餐人数分布中，两餐消费高峰间的消费低谷时间，可以较合理地对餐次进行划分。

② 日期编号 84 是秋季学期最后一个采样点，后面直接拼接了春季学期的采样点，因此该处存在阶跃变化。

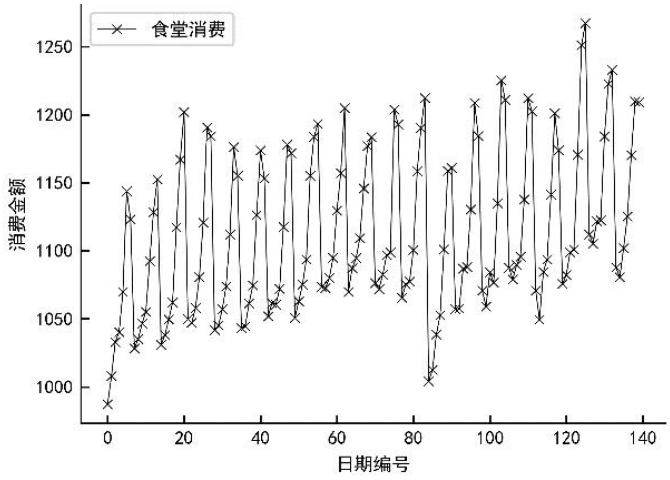


图1 采样区间内食堂消费次数和金额分布

表1进一步展示了不同类别学生的消费指标均值。① 对比家庭经济困难学生和其他学生的消费指标不难发现，家庭经济困难学生在校就餐比例更高、平均消费金额更低、在经济型食堂就餐的比例也更高，同时各项指标在工作日、周五和休息日的取值也存在较大差异，研究中选取的指标变量包含了较丰富的信息。

表1 不同类别学生消费行为特征

指标	类别	全体学生	家庭经济困难学生	其他学生
在校就餐比例	工作日	61.01%	66.43%	60.04%
	周五	49.72%	56.06%	48.58%
	休息日	43.95%	51.43%	42.60%
餐均消费金额(元)	工作日	12.01	11.23	12.16
	周五	12.87	12.14	13.02
	休息日	13.32	12.59	13.48
经济型食堂就餐比例	工作日	26.45%	30.63%	25.70%
	周五	22.08%	27.10%	21.17%
	休息日	20.33%	25.96%	19.32%
学生人数	—	10976	1670	9306

国内通过校园大数据识别家庭经济困难学生的相关研究一般仅分析学生

① 限于篇幅，这里仅展示了午餐和晚餐的平均消费情况，午餐和晚餐的消费情况接近，同时覆盖了在校本科生主要的食堂消费。

的校内消费行为数据，而完全不将认定家庭经济困难学生的主要依据——家庭基本经济情况信息考虑在内。这种方式固然具备数据的客观性、连续性、实时性等优势，并可以动态地对资助对象进行追踪和调整，但也可能存在自变量对因变量解释度不高、识别准确率较低等问题，原因是学生的校内消费行为可能只是其家庭经济状况的一个侧面写照。

为了解决这一问题，本文通过该校本科生新生入学调查问卷，获取到独生子女、户口类型、父母工作职位、家庭年收入等家庭基本信息。该校调查是由独立于学生资助部门的教育研究机构独立开展的调查，旨在了解被录取新生在入学前的学习和生活经历，其中包含的家庭社会经济背景相关变量与贫困生认定、助学金发放没有任何关联，具有较高的真实性和客观性。由于面向该校新生的大规模问卷调查开始较晚，因此这里仅关联了两届学生的家庭基本信息与校内消费行为信息，共包含 2503 名学生的问卷调查数据。学生的主要基本特征分布如表 2 所示，[①] 可以看出，家庭经济困难学生和非困难学生在各项家庭特征维度上的分布存在较大差异，经济困难学生所在家庭的独生子女比例更低、农村户籍比例更高、家庭居住在乡镇农村的比例更高、父亲受教育程度更低、父亲职业为农民或农民工比例更高、家庭年收入在 5 万以下的比例更高，说明使用这些特征变量识别家庭经济困难学生具有一定的合理性。

表 2　子样本学生的主要基本特征(%)

	类别	全体学生	家庭经济困难学生	未受资助的学生
性别	男生	58.85	53.91	59.75
	女生	41.15	46.09	40.25
独生子女	是	78.91	35.68	86.74
	否	21.09	64.32	13.26
户籍	城镇	84.86	43.23	92.40
	农村	15.14	56.77	7.60
城镇类型	直辖市、省会	36.96	5.47	42.66
	地级市	26.41	13.80	28.69
	县级市	23.53	27.08	22.89
	乡镇农村	13.10	53.65	5.76

① 限于篇幅，这里仅展示了父亲的受教育程度和工作类型，在具体模型构建中同时考虑了父亲和母亲的信息。

续表

	类别	全体学生	家庭经济困难学生	未受资助的学生
地区	东部	50.38	21.89	55.55
	中部	30.56	48.18	27.37
	西部	19.06	29.95	17.08
父亲受教育程度	初中或以下	14.82	54.95	7.55
	高中或中专	18.34	27.60	16.66
	大学或以上	66.84	17.45	75.79
父亲工作类型	高级管理人员	23.05	3.13	26.66
	专业技术人员	43.95	15.36	49.13
	技术辅助人员	20.89	27.60	19.68
	无业人员	3.56	10.94	2.22
	农民工、农民	8.55	42.97	2.31
家庭年收入	1万元以下	3.96	19.27	1.18
	1～5万	23.17	67.71	15.10
	5～10万	25.81	9.89	28.69
	10万元以上	47.06	3.13	55.03
样本量	—	100.00	15.34	84.66

(二)研究方法

本文的研究主要分为以下三步：第一步，使用食堂消费统计信息识别家庭经济困难学生。表3包含了建模中涉及的指标变量，分别涉及工作日、周五、休息日的早餐、午餐、晚餐、夜宵对应的在校就餐比例、餐均消费金额、经济型食堂就餐比例3类，总计36个指标，另外添加了性别信息，以此控制男女就餐行为差异对识别结果的影响。为了提升模型对数据的挖掘能力，本文在传统逻辑回归模型的基础上加入了提升树、[①] 支持向量机[②]两个经典的统计学习模型(李航，2019)，以期得到更好的识别效果。

① 提升树以决策树为基本分类单元，通过改变正/误分类样本的权重，学习多个分类单元，并将这些分类单元进行线性组合，被认为是统计学习中最有效的方法之一。

② 支持向量机的基本思想是求解能够正确划分训练数据集并且对于最难区分的正负样本分离间隔最大的分离超平面，被认为是统计学习中最有效的方法之一。

表 3　识别家庭经济困难学生所使用的统计指标

	在校就餐比例	餐均消费金额	经济型食堂比例
早餐	工作日早餐	工作日早餐	工作日早餐
	周五早餐	周五早餐	周五早餐
	休息日早餐	休息日早餐	休息日早餐
午餐	工作日午餐	工作日午餐	工作日午餐
	周五午餐	周五午餐	周五午餐
	休息日午餐	休息日午餐	休息日午餐
晚餐	工作日晚餐	工作日晚餐	工作日晚餐
	周五晚餐	周五晚餐	周五晚餐
	休息日晚餐	休息日晚餐	休息日晚餐
夜宵	工作日夜宵	工作日夜宵	工作日夜宵
	周五夜宵	周五夜宵	周五夜宵
	休息日夜宵	休息日夜宵	休息日夜宵

第二步，使用消费的时间序列数据识别家庭经济困难学生。使用统计指标进行建模固然可以减少指标变量的数量，降低模型的复杂度，减低偶然因素带来的奇异值对结果的影响，但也会造成有效信息的丢失。本文选用是否在食堂用餐、用餐金额、是否在廉价食堂用餐 3 个变量对学生每次用餐的行为进行描述，采样区间的 140 天内共有 560 餐次，总计 1680 个指标变量，另外添加了性别信息，以此控制男女就餐行为差异对识别结果的影响。通过对有效信息足够精细的描述，可以帮助我们判断使用校内消费行为识别家庭经济困难学生的有效性。

第三步，结合家庭基本信息识别家庭经济困难学生。家庭基本情况信息是目前认定家庭经济困难学生的主要依据，但据此针对每一位在校本科生进行人工识别工作量极大，本文将结合 2503 名在校本科生的校内消费行为信息①与家庭基本信息，进一步验证将家庭基本信息与校内消费行为信息相结合进行建模是否具有更高的准确率。所使用的家庭基本信息如表 4 所示。

① 由于样本量较小，这里选用第一步中食堂消费统计指标对校内消费行为进行描述，如实际工作中可以获得全校范围的家庭基本信息，考虑选用时间序列数据对校内消费行为进行描述，提高模型的识别能力。

表4 家庭基本信息变量定义和描述性统计

类别	均值	标准差	性质	说明
年级	0.53	0.50	分类	“2016级”=0，“2017级”=1
独生子女	0.79	0.41	分类	“否”=0，“是”=1
户籍类型	0.85	0.36	分类	“农村”=0，“城镇”=1
城镇类型	1.13	1.06	分类	“直辖市、省会”=0，“地级市”=1，“县级市”=2，“乡镇农村”=3
地区	0.69	0.77	分类	“东部”=0，“中部”=1，“西部”=2
父亲学历	14.23	2.66	连续	“初中或以下”=9，“高中或中专”=12，“大学或以上”=16
父亲工作类型	1.31	1.12	分类	“高级管理人员”=0，“专业技术人员”=1，“技术辅助人员”=2，“无业人员”=3，“农民工、农民”=4
母亲学历	13.87	2.78	连续	“初中或以下”=9，“高中或中专”=12，“大学或以上”=16
母亲工作类型	1.59	1.08	分类	“高级管理人员”=0，“专业技术人员”=1，“技术辅助人员”=2，“无业人员”=3，“农民工、农民”=4
家庭年收入	2.16	0.91	分类	“1万元以下”=0，“1～5万元”=1，“5～10万元”=2，“10万元以上”=3
当地相对收入	0.94	0.47	分类	“贫困或非常贫困”=0，“中等”=1，“富裕或非常富裕”=2

四、实证结果

(一)使用食堂消费统计信息识别家庭经济困难学生

图2展示了使用食堂消费统计信息识别家庭经济困难学生的效果，采用接受者操作特征曲线(ROC)[①]判断模型优劣，横坐标越大代表误判的家庭经济困难学生占其他学生总数的比例越高；纵坐标越大代表识别出的家庭经济

① 接受者操作特征曲线(ROC)，以虚惊概率为横轴，以击中概率为纵轴，用来检验分类模型的好坏，AUC代表ROC曲线下方的面积，数值越大代表模型效果越好。

困难学生占家庭经济困难学生总数的比例越高，曲线下方的面积越大则可以用较小的误判代价识别出越多的家庭经济困难学生；图中虚线代表随机挑选。直观来看，使用食堂消费统计信息识别家庭经济困难学生是有区分效果的，但识别效果并不足够令人满意。

如表 5 所示，使用逻辑回归模型可以识别出 59.4%的家庭经济困难学生，被筛选出的学生中有 21.0%确实为家庭经济困难学生，测试集准确率约 58.8%；使用支持向量机模型可以识别出 65.7%的家庭经济困难学生，被筛选出的学生中有 21.9%确实为家庭经济困难学生，测试集准确率约 59.3%。相较于逻辑斯特回归，采用提升树和支持向量机等判别模型，可以提高模型对家庭经济困难学生的识别能力。

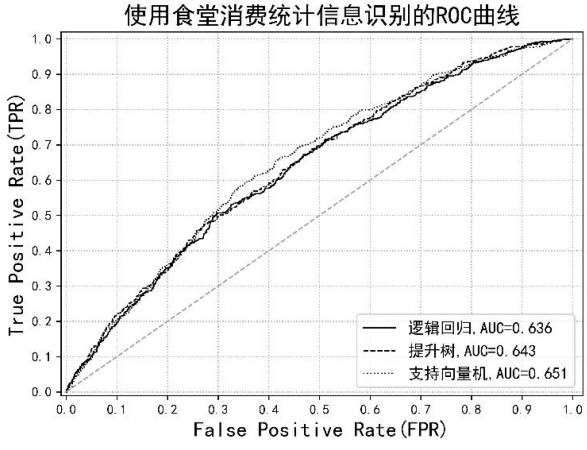


图 2　使用食堂消费统计信息识别家庭经济困难学生的 ROC 曲线

表 5　识别模型的评价指标

	AUC	精确率	召回率	F 值	训练集准确率	测试集①准确率
逻辑回归	0.636	0.210	0.594	0.310	0.635	0.588
提升树	0.643	0.206	0.630	0.310	0.677	0.577
支持向量机	0.651	0.219	0.657	0.329	0.598	0.593

(二)使用食堂消费时间序列数据识别家庭经济困难学生

图 3 展示了使用食堂消费时间序列数据识别家庭经济困难学生的效果，

① 建模过程中，从总样本随机抽取 70%的样本作为训练集拟合模型参数，剩余 30%的样本作为测试集检验模型的识别效果。

以逻辑回归模型为例，相比于统计指标，使用更加精细的时间序列数据具有更好的识别效果，如表6所示，测试集的识别准确率从58.8%提高到64.6%，但识别效果依旧不够令人满意。鉴于时间序列数据已经对本科生的校内消费行为信息进行了相当精细的描述，如果要进一步提升模型的识别效果，需要引入其他维度的有效信息。

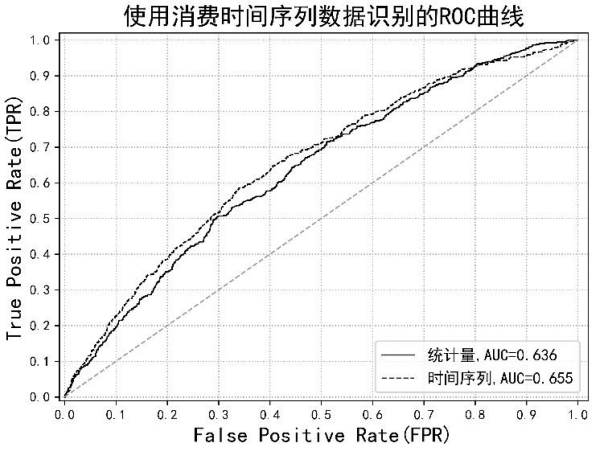


图3 使用食堂消费的时间序列数据识别家庭经济困难学生的ROC曲线

表6 认定模型的评价指标

	AUC	精确率	召回率	F值	训练集准确率	测试集准确率
统计量数据	0.636	0.210	0.594	0.310	0.635	0.588
时间序列数据	0.655	0.228	0.588	0.328	0.750	0.646

(三)结合家庭基本情况信息识别家庭经济困难学生

图4展示了结合家庭基本情况信息识别家庭经济困难学生的效果，逻辑回归、提升树、支持向量机三种模型的识别效果均有较大幅度的提升。如表7所示，使用逻辑回归模型可以识别出86.0%的家庭经济困难学生，被筛选出的学生中有47.8%确实为家庭经济困难学生，测试集准确率约85.6%，已经具备了较高的识别准确度；使用支持向量机模型可以识别出87.2%的家庭经济困难学生，被筛选出的学生中有68.0%确实为家庭经济困难学生，测试集准确率约91.6%，识别出绝大多数家庭经济困难学生的同时，误判率较低，模型可以较好地对家庭经济困难学生进行识别。

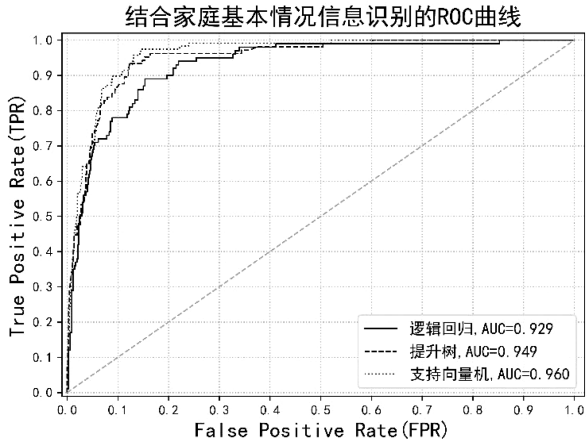


图 4　结合家庭基本情况信息识别家庭经济困难学生的 ROC 曲线

表 7　认定模型的评价指标

	AUC	精确率	召回率	F 值	训练集准确率	测试集准确率
逻辑回归	0.929	0.478	0.860	0.614	0.924	0.856
提升树	0.949	0.560	0.895	0.689	0.887	0.887
支持向量机	0.960	0.680	0.872	0.764	0.958	0.916

五、结论与讨论

本文以某“双一流”A 类建设高校多届本科生作为研究对象，利用食堂消费数据、贫困生认定和助学金发放数据、家庭基本信息调查数据，分析通过校园消费大数据对家庭经济困难学生进行认定的可行性和实现方法。研究发现，校园消费行为数据具备对家庭经济困难学生进行识别的作用，但识别效果不够理想。其中，使用食堂消费统计信息进行识别的准确率约为 58.8%，使用消费时间序列数据进行识别的准确率约为 64.6%。若将食堂消费数据与入学前家庭基本情况调查信息相结合后，可将识别的准确率提升到约 85.6%，采用更加有效的统计学习模型后识别准确率可提升到 91.6%。由此可见，引入大数据技术对食堂消费数据进行分析建模，并与学生自述报告中的家庭经济特征相结合，可以非常准确地实现对家庭经济困难学生的认定和瞄准，是实现精准资助的重要基础和发展方向。

借助统计学习模型对家庭经济困难学生进行认定，可以提高识别能力。

但也正因为这类模型的拟合能力强，更容易导致过拟合的情况发生，在实践中应当利用相应的技术手段减少过拟合发生的概率。首先，选用带有复杂度约束的模型，如研究中选用的支持向量机模型和提升树中的XGBoost模型(Chen and Guestrin，2016)，可以弱化自变量过多且样本量不足导致的过拟合，避免模型过于复杂导致的泛化能力下降。其次，保证训练集和测试集的准确率在同一水平，如果发现训练集的准确率显著高于测试集，应适当降低模型复杂度，或在选择超参数的时候使用交叉验证。最后，尽可能增大样本量，本研究中因条件所限，结合家庭基本信息－消费信息的综合数据样本量仅有2503，未来可以定期对贫困生的相关数据进行整理和存档，为建模识别贫困生提供坚实的数据基础。

本研究中的样本标签通过校方认定的贫困生名单确定，而标签本身的确可能存在泄漏率(不应资助的学生却获得资助)和遗漏率(应该资助的学生未获得资助)的问题。在实践中，泄漏率问题可以通过对“接受资助学生”的辅导员跟踪反馈和同学调查进行校正，将泄漏的“贫困生”样本剔除。同时，可以参考信号博弈的相关研究，通过提高伪装成本与期望处罚之和，从政策角度降低泄漏率(彭桥等，2020)。解决遗漏率问题是利用消费数据识别贫困生的初衷，本研究仅包含了学生在食堂就餐的消费数据，在实际操作中可以扩大消费数据的采集范围，进而增大消费信息相对于家庭基本信息的权重。同时，在使用模型预测贫困生时，可以适当降低识别阈值，增大召回率，对模型认定的“贫困生”进行人工确认，降低遗漏率的同时保证泄漏率不过分升高，将遗漏的“贫困生”样本找回。

本研究具有以下政策启示。首先，食堂消费数据不仅反映了学生的家庭经济特征，还反映了学生的消费习惯，利用食堂消费数据来识别贫困生可以识别那些贫困但没有申请资助的学生，即降低遗漏率。但是，仅仅利用消费数据的识别准确率仍然不够理想，不宜作为识别家庭经济困难学生的唯一标准。其次，将食堂消费数据与学生基本家庭信息相结合来构建预测模型可以大大提升预测准确度。学生基本家庭信息的准确性非常关键，如果高校获取的学生家庭基本情况的途径和目的与学生资助挂钩，则可能会引入一些人为偏差，这时食堂消费数据的客观性、连续性、实时性等优势便可以发挥更大的作用，在一定程度上起到矫正作用，同时可以实现对资助对象的动态追踪和调整。最后，利用校园食堂消费大数据进行建模可以提升资助瞄准程度，但是在实际应用中建议使用更长时段的数据，并排除一些极端时间或事件的影响，以保证模型的预测精度。

[参考文献]

陈闻鹤、程龙生、常志朋、周涵婷，2022：《基于 Lasso 稳健马田系统的相对贫困识别方法》，《系统工程理论与实践》第 2 期。

陈志、丁士军、吴海涛，2021：《贫困识别评估指标优化及实证》，《统计与决策》第 17 期。

蒋卓轩、张岩、李晓明，2015：《基于 MOOC 数据的学习行为分析与预测》，《计算机研究与发展》第 3 期。

李春雷、王文生、郭雷风、陈桂鹏，2021：《基于集成学习算法的返贫人口识别模型——以 H 省 F 县贫困户建档立卡数据为例》，《江苏农业科学》第 17 期。

李德淇、黄南雄、张凯悦、朱郑州，2018：《基于学习者情绪的成绩预测研究》，《中国教育网络》第 11 期。

李航，2019：《统计学习方法》，北京：清华大学出版社。

刘譞，2017：《基于学生行为的成绩预测模型的研究与应用》，西安电子科技大学博士学位论文。

彭桥、肖尧、陈浩，2020：《精准扶贫与扶贫对象识别——基于信号博弈分析框架》，《兰州学刊》第 12 期。

全国学生资助管理中心，2020：《2019 年中国学生资助发展报告》，《人民日报》，2020 年 5 月 21 日。

宋俊秀，2017：《家庭经济困难学生精准识别的指标体系构建与实施途径》，《教育财会研究》第 5 期。

田志磊、袁连生，2010：《采用非收入变量认定高校家庭经济困难学生的实证研究》，《北京大学教育评论》第 2 期。

吴斌珍、李宏彬、孟岭生、施新政，2011：《大学生贫困及奖助学金的政策效果》，《金融研究》第 12 期。

吴朝文、代劲、孙延楠，2016：《大数据环境下高校贫困生精准资助》，《黑龙江高教研究》第 12 期。

徐丽红，2015：《贫困认定：高校资助工作的“阿喀琉斯之踵”》，《高教探索》第 7 期。

杨仑，2019：《用大数据发餐补“饱”暖学生心》，《科技日报》，2019 年 9 月 26 日。

杨钋、刘霄，2019：《研究生收费前贫困资助政策的瞄准和减贫效果分析——以首都高校研究生为例》，《教育与经济》第 2 期。

张存禄、马莉萍、陈晓宇，2021：《贫困生资助对大学生消费行为的影响：基于校园卡消费大数据和问卷调查数据的研究》，《教育与经济》第 3 期。

郑杰，2015：《上海市高校家庭经济困难学生识别系统构建研究》，华东师范大学博士学位论文。

Chen，T. and C. Guestrin，2016，“XGBoost：A Scalable Tree Boosting System”，*In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*.

Harrell，I. L. and B. L. Bower，2011，“Student Characteristics That Predict Persistence in

Community College Online Courses", *American Journal of Distance Education*, 25(3): 178－191.

Johnstone, B. D, 2003, "Cost Sharing in Higher Education: Tuition, Financial Assistance, and Accessibility in a Comparative Perspective", *Czech Sociological Review*, 39(3): 351－374.

Loyalka, P., Y. Song and J. Wei, 2012, "The Distribution of Financial Aid in China: Is Aid Reaching Poor Students", *China Economic Review*, 23(4): 898－917.

Can Canteen Consumption Big Data Accurately Identify Students with Financial Difficulties? ——An Empirical Study Based on College Students' Behavioral Data, Administrative Data and Survey Data

ZHANG Cun-lu[1], MA Li-ping[2], CHEN Xiao-yu[2]

(1. Graduate School of Education, Peking University;
2. Institute of Education Economics, Peking University)

Abstract: Based on the big data of undergraduates' canteen consumption, the administrative data of financial aid and the survey data of socioeconomics information at a double first-class university in China, the paper examines the accuracy of identifying students with financial difficulties by using the statistical learning models and draws the following conclusions: the accuracy of identifying those undergraduates with financial difficulties is 60% by using big data of canteen consumption; and the accuracy can be improved to about 65% by using more refined consumption time series data; the accuracy can be further improved to about 92% by combining consumption data with questionnaire survey data of socioeconomics information. Compared with traditional logistic regressions the accuracy of discriminant models such as Lifting Tree and Support Vector Machine is significantly higher. These findings indicate that the accuracy of identifying students with financial difficulties needs to be improved by combining canteen consumption data with students' socioeconomics information.

Key words: students with financial difficulties; poverty students; financial aid; accuracy of identifying students; big data of undergraduates' canteen consumption

(责任编辑：郑磊　责任校对：郑磊　胡咏梅)