附录 1：部分变量值设定的具体说明

本文选择的输入变量具体变量值根据 2021 年 CSS 调查问卷设立的选项加以确定，其中部分输入变量的补充说明如下：

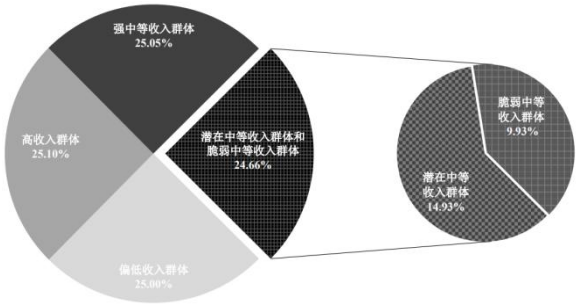
收入来源中，工资性收入包括劳动报酬收入（含工资、奖金、津贴、节假日福利等）；经营性收入包括农业经营纯收入与经商办厂净收入；财产性收入包括出售/出租房屋、土地收入以及家庭金融投资理财收入（债券、存款、放贷等的利息收入，股票投资收入及股息、红利收入等）；转移性收入包括家庭成员退休金、养老保险金、失业保险金、工伤保险金、生育保险金等社保收入，政府、工作单位和其他社会机构提供的社会救助收入（如最低生活保障、困难补助、疾病救助、灾害救助、学校奖学金/助学金、贫困学生救助等），政府提供的生产经营补贴、政策扶持收入（如农业补助、税费减免等），居委会、村委会提供的福利收入（如集体生产经营分红、非救助性补贴等）以及赡养收入（不在一起生活的亲属提供的赡养收入）；其他收入对应家庭收入中未被归为以上四类收入的部分。

支出结构中，食品支出对应饮食支出（包括外出饮食支出；自产食品估算其价格，并计算在内）；衣着支出对应衣着支出（衣服、鞋帽等）；居住支出包括缴纳房租的支出和电费、水费、燃气（煤炭）费、物业费、取暖费；生活用品及服务支出包括家用电器、家具、家用车辆等购置支出、家政服务支出（指的是受访者家庭的家政支出）；交通通讯支出包括通讯支持（如固定电话/手机的话费、上网费等）、交通支出（如上下班等交通费及家用车辆汽油费、保养费、路桥费等）；教育文化娱乐支出包括教育支出（如学费、杂费、文具费、课外辅导费、在校住宿费等，但在校的饮食支出计入家庭饮食支出）以及文化、娱乐、旅游支出；医疗保健支出对应医疗保健支出（如看病、住院、买药等的费用，扣除报销部分）；购房支出包括购房当年首付（非 2020 年首付不计）、分期偿还房贷的支出；其他支出包括赡养或者照料不在一起生活的亲属（如父母等老人）的支出、红白喜事支出：人情往来支出（如礼品、现金等）、保险支出（如个人缴纳的社会保险、人寿保险、重疾险、车险、商业医疗保险等）与其他支出。

外来务工人员指在当地居住满半年、户口登记在居住地所在县以外的农业户口居民。东部地区包括：北京、天津、河北、上海、江苏、浙江、福建、山东、广东和海南，中部地区包括：山西、安徽、江西、河南、湖北和湖南，西部地区包括：内蒙古、广西、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏和新疆，东北地区包括：辽宁、吉林和黑龙江。

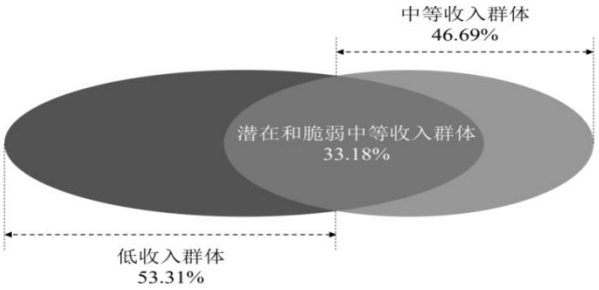
附录 2：数据预处理

初步处理后共得到 9399 条有效记录，其中低收入群体、中等收入群体与高收入群体占比分别为 39.93%、34.97%和 25.10%。在低收入群体中，有 25%的偏低收入群体和 14.93%的潜在中等收入群体；在中等收入群体中，有 25.05%的强中等收入群体和 9.93%的脆弱中等收入群体（见附图 1）。



附图 1　2021 年中国居民收入分配格局的结构特征

在不包括高收入群体的样本中（见附图 2），低收入群体、中等收入群体分别占 53.31%和 46.69%，并且有 33.18%的样本成员属于潜在中等收入群体和脆弱中等收入群体，其规模之高凸显了针对重点人群精准施策的必要性。

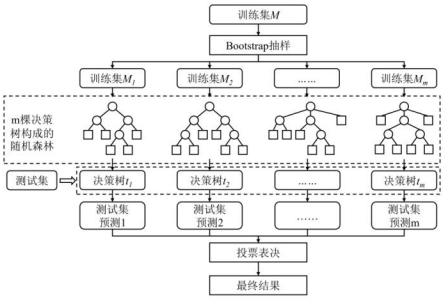


附图 2　样本数据集中各类收入群体比重韦恩图

附录 3：随机森林模型及评价指标介绍、模型预测效果对比

一、随机森林模型与相关评价指标

随机森林模型由 Breiman（2001）提出，是一种以决策树作为基模型的自举汇聚法（Bagging）集成分类器。其构建逻辑如下：首先，抽取样本形成训练集 M，利用 bootstrap 方法对 M 进行有放回的随机抽样并重复此过程 m 次[①]，得到 m 个随机训练集 M_{sample}；其次，针对每个随机训练集 M_{sample}，依据选择最大特征值的规则，在 K 个输入变量中随机选择 k 个（$k \leqslant K$）作为属性特征，并以最佳分割属性特征作为结点构建决策树；最后，将得到的 m 棵决策树进行组合，通过投票表决的方式确定最终的多分类模型系统（见附图 3）。由于随机森林模型在样本选取和特征选取环节遵循随机性原则，因此基模型之间保持了较低的相关性，从而增强了模型的稳健性和泛化能力，在避免过拟合问题与抗噪声干扰方面的能力优于其他算法（Li 等，2020；刘生龙等，2021）。



附图 3　随机森林模型的构建逻辑

随机森林模型的变量重要度得分衡量了在整个模型中每个特征对预测准确性的贡献程度，通过重要度评估功能，可以帮助我们排除无关或冗余的特征，而将注意力集中在对预测目标最有影响力的特征上。这一功能的实现方法有两类，分别以袋外错误率（out-of-bag error rate）或不纯度（impurity）为评估规则。以袋外错误率为评估规则的基本思想是用袋外样本测试已生成的随机森林模型性能，并随机改变袋外样本中某个输入变量的值，比较改变前后模型性能的变化。如果模型性能大幅下降，说明这个输入变量的重要度较高；若模型性能下降幅度不明显，则该输入变量的重要度排名较低（方匡南等，2010）。以不纯度为评估规则的基本思想则通过某个输入变量在随机森林所有决策树中结点分裂造成的不纯度变化计算该变量的特征重要性评分（*VIM*），其中不纯度的度量方式主要包括基尼系数（gini）和信息熵（entropy）两种。

以基尼系数为例[②]，结点 n 的不纯度定义为：

$$GI_n = 1 - \sum_{l=1}^{L} p_{nl}^2$$

* MERGEFORMAT (1)

① Bootstrap 方法是一种有放回的抽样。对于样本量为 N 的原始训练集 M，每个样本未被抽到的概率是$(1-1/N)^N$。当 N 足够大或趋于正无穷时，此概率值收敛于 $1/e$，约为 0.368。因此，训练集中平均有 37%左右的样本没有被抽取到 Bootstrap 样本中，这部分样本被称为袋外样本（Out of Bag，OOB）。由于袋外样本并未参与模型训练，基于此类样本计算的 OOB 得分可以作为随机森林模型预测效果的评估指标。本文绘制了三个随机森林模型的 OOB 得分与测试集精度曲线。结果显示，随着训练决策树的数量不断增加，三个模型的曲线均表现出迅速上升后逐渐趋于平稳的态势，取得了较高的训练得分。因篇幅所限，三个模型的 OOB 得分与测试集精度曲线并未在正文中展示。若有需要，可以联系作者索要具体计算结果。

② 基尼不纯度是指根据结点中样本分布情况对样本进行分类后，从中随机选择的样本被分错的概率。

其中，L 表示结点 n 中包含的样本类别，p_{nl} 表示结点 n 中 l 类样本所占比重。

在第 i 棵决策树中，变量 c_j 在结点 n 的重要度（VIM_{ijn}）是指结点 n 分支前后基尼不纯度的变化：

$$VIM_{ijn} = GI_n - GI_{left} - GI_{right}$$

* MERGEFORMAT (2)

其中，GI_{left} 和 GI_{right} 分别表示结点 n 分枝后左右结点的基尼不纯度。

进一步计算变量 c_j 在第 i 棵决策树的重要性评分（VIM_{ij}）为：

$$VIM_{ij} = \sum\nolimits_n VIM_{ijn}$$

* MERGEFORMAT (3)

将变量 c_j 在 m 棵决策树上的重要性评分加总后做归一化处理，得到最终的重要性评分：

$$VIM_j = \sum\nolimits_i VIM_{ij} / \sum\nolimits_j \sum\nolimits_i VIM_{ij}$$

* MERGEFORMAT (4)

为了全面合理地评价模型效果，本文基于混淆矩阵计算准确率、精准率、召回率、F1 得分和 AUC（Area Under Curve）得分等多项指标，以此考察随机森林模型的预测效果。对于二分类问题，混淆矩阵记录了测试集样本的真实类别与预测类别的四种组合，分别是真阳（True Positive，*TP*）、假阳（False Positive，*FP*）、真阴（True Negative，*TN*）、假阴（False Negative，*FN*）。其中，真阳是指实际类别为真、预测类别也为真的情况，真阴是指实际类别为假、预测类别也为假的情况。以上两种情况均属于模型预测正确，而假阳和假阴则分别对应“指假为真”和“指真为假”两类预测错误的情况。

准确率（*Accuracy*）是指预测正确的结果占总样本的比例，其计算公式为：

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$ * MERGEFORMAT (5)

精确率（*Precision*）是真阳样本在所有被预测为真的样本中所占比重，在本文中对应被准确预测为中等收入群体的样本占所有被预测为中等收入群体样本的比例，其计算公式为：

$$Precision = TP / (TP + FP)$$

* MERGEFORMAT (6)

召回率（*Recall*）是真阳样本在所有实际为真的样本中所占比重，在本文中对应被准确预测为中等收入群体的样本占实际中等收入群体样本的比例，其计算公式为：

$$Recall = TP / (TP + FN)$$

* MERGEFORMAT (7)

F1 得分（*F1_score*）综合考察精确率与召回率两个指标，本文参考刘云菁等（2022）的处理方法，以等权重法计算 F1 得分：

$$F1\_score = 2 \times Precision \times Recall / (Precision + Recall)$$ * MERGEFORMAT (8)

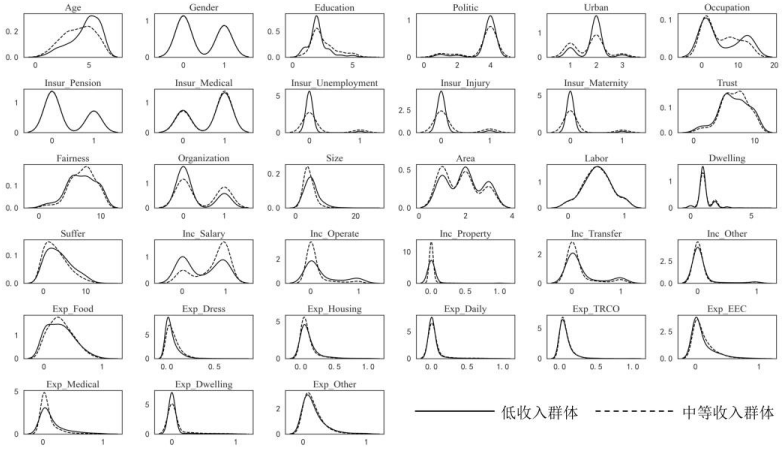
AUC 是受试者工作特征（ROC）曲线下的面积，这一曲线是通过遍历所有阈值绘制得

到的。若 ROC 曲线比较陡峭，AUC 的值较大，则模型的预测效果比较好[①]。

二、模型构建与效果评估补充说明

（一）输入变量的可视化描述

首先利用核密度曲线刻画 33 个输入变量在低收入群体与中等收入群体内的分布情况（见附图 4）。结果显示，2021 年我国低收入群体与中等收入群体在户主特征、家庭特征与家庭收支情况等方面均存在显著差异，这也印证了“扩中”的复杂性。如果只分析少数变量对中等收入群体规模的影响，可能会遗漏某些重要的扩中因素，为此借助机器学习方法对群体特征进行画像。从可视化描述的结果来看，本文选择的输入变量在不同收入群体间的分布具有明显的辨识度，具备构建随机森林模型的可行性，但不同输入变量在决定居民所属收入层级时的作用“孰重孰轻”，则需要基于建模结果进行分析。



附图 4　输入变量在低收入群体与中等收入群体间的分布差异

（二）样本划分和参数设定

本文共构建四个随机森林模型。基准模型得到训练集样本 10560 条，其中 5622 条属于低收入群体样本、4938 条属于中等收入群体样本；测试集样本 3520 条，其中 1884 条属于低收入群体样本、1636 条属于中等收入群体样本。经过 SMOTE（Synthetic Minority Over-sampling Technique）算法平衡化处理后，扩展模型一得到训练集样本 7050 条，其中 3499 条属于偏低收入群体样本，3551 条属于潜在中等收入群体样本；测试集样本 2350 条，其中 1201 条属于偏低收入群体样本，1149 条属于潜在中等收入群体样本。扩展模型二得到训练样本 7008 条，其中 4214 条属于潜在中等收入群体样本，2794 条属于脆弱中等收入群体样本；测试集样本 2336 条，其中 1398 条属于潜在中等收入群体样本，938 条属于脆弱中等收入群体样本。扩展模型三得到训练样本 7062 条，其中 3522 条属于脆弱中等收入群体样本，3540 条属于强中等收入群体样本；测试集样本 2354 条，其中 1186 条属于脆弱中等收入群体样本，1168 条属于强中等收入群体样本。

本文以准确率最大化为调参目标，采用网格搜索和 5 折交叉验证方法确定随机森林模型的最优参数组合[②]，需要调整的参数包括：n_estimators 控制树的数量即迭代的次数、

① ROC 曲线是以假阳率（FPR）为横轴、真阳率（TPR）为纵轴，在变化阈值的情况下绘制的曲线。其中，假阳率的计算公式为：$FPR = FP/(TN+FP)$，真阳率的计算公式为：$TPR = TP/(TP+FN)$。当 *TPR* 值较高、*FPR* 值较低时，ROC 曲线比较陡峭，此时可以认为模型的性能较好。

② 网格搜索与交叉验证常结合使用来作为机器学习的调参手段，以确定最佳的超参数组合并评估模型的性能。其中，网格搜索通过循环遍历尝试待选参数的全部组合，并从中选择表现最好的参数组合；交叉验证（*K* 折交叉验证）用于评估模型拟合

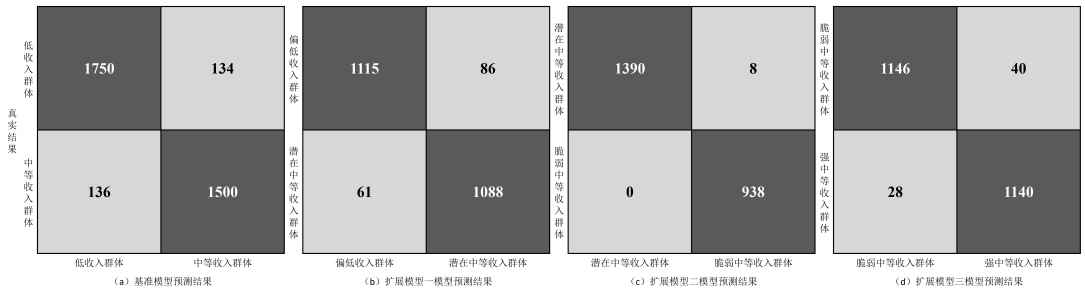
max_depth 控制树的最大深度、class_weight 控制每个类别的权重平衡方式、criterion 控制分类时使用的度量标准、max_features 决定选择最大特征数量的方式。参考既有文献确定具体的调参范围[①]（Gao 等，2018；Du 等，2021），经过网格搜索和交叉验证进行逐一尝试后确定三个模型效果达到最优时的参数组合。具体结果见附表 1 所示。

附表 1　　随机森林模型训练参数

参数	基准模型	扩展模型一	扩展模型二	扩展模型三
n_estimators	280	180	60	180
max_depth	28	24	18	26
class_weight	balanced	balanced_subsample	balanced	balanced
criterion	entropy	entropy	gini	entropy
max_features	sqrt	sqrt	sqrt	sqrt

（三）预测效果评估

本文使用训练后的随机森林模型对测试集进行预测，并得出四个模型的混淆矩阵（见附图 5）。结果显示，四个混淆矩阵的主对角线元素值大幅领先于副对角线元素值，表明相应的随机森林模型在预测准确性方面表现较好。其中，基准模型的预测准确率达到 92.33%，三个扩展模型的预测准确率更是达到 93.00%以上，能够比较准确地判别样本家庭所属的真实收入层级。为了系统评估随机森林模型的预测效果，本文计算了不同评估指标得分，并与传统逻辑回归模型[②]与决策树模型的预测效果进行比较[③]（见附表 2）。结果显示，相比广义线性模型和单分类器模型，随机森林模型的预测效果在各项评价指标下的表现都更优秀，具有很高的收入层级识别能力，能够准确完成不同收入群体的特征画像，具有较好的泛化性。



附图 5　四个随机森林模型的混淆矩阵

附表 2　　不同预测模型的效果评估　　（%）

分类目标	模型	Accuracy	Precision	Recall	F1_score	AUC
低收入群体	逻辑回归	69.12	66.89	66.44	66.67	76.28
	决策树	69.49	67.85	65.28	66.54	76.28

程度，将训练集的所有数据平均划分成 K 份（本文选择 K=5），取其中一份作为验证集，余下的 K-1 份作为交叉验证的训练集，从而提升模型的性能。

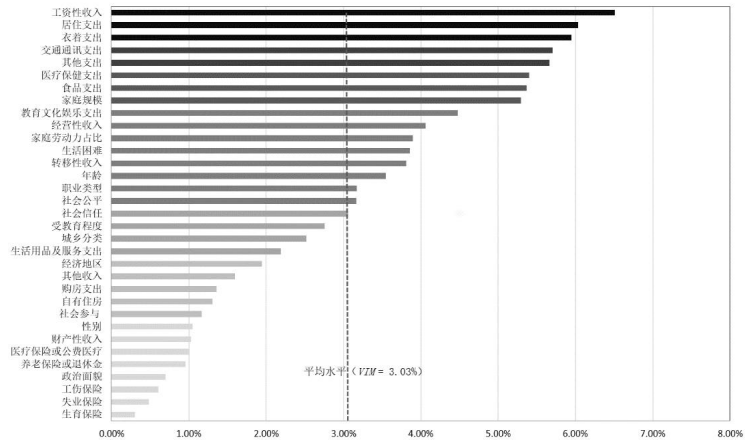
① 各参数的调参范围如下：n_estimators 范围是 20~300（搜索步长为 20），max_depth 范围是 8~30（搜索步长为 2），class_weight 选择 balanced 或 balanced_subsample 模式，criterion 取基尼系数（gini）或信息熵（entropy），max_features 取 sqrt、log2、auto 或 None，按间距调整直到训练的模型准确率达到最大为止。

② 逻辑回归模型是一种广义线性回归模型，常被用于研究二分类响应变量与诸多自变量间的相互关系。本文通过交叉验证的方式构建用于预测家庭收入层级的逻辑回归模型，基于 L1 范数确定最佳正则化强度倒数值（Cs）。经验证，当四个模型的参数 Cs 分别为 1.50、0.10、0.05 和 10.00 时，逻辑回归模型的效果达到最优。

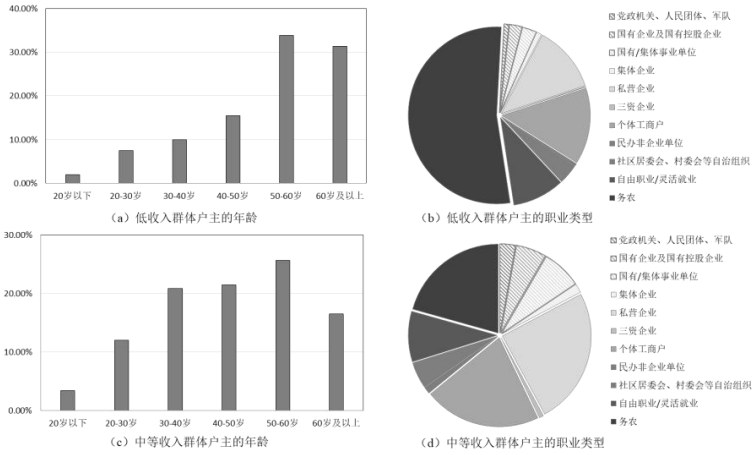
③ 决策树模型是一种广泛应用的归纳推理算法，该算法把事例从决策树的根结点分步排列到叶子结点来对其进行分类，所处的叶子结点即为所属分类。本文首先使用默认参数构建用于预测家庭收入层级的决策树模型，并针对存在的过拟合问题对叶子结点数与树的深度进行参数搜索，确定模型效果最优时的参数组合为：基准模型 max_depth 为 9、max_leaf_nodes 为 43，扩展模型一 max_depth 为 9、max_leaf_nodes 为 49，扩展模型二 max_depth 为 10、max_leaf_nodes 为 49，扩展模型三 max_depth 为 8、max_leaf_nodes 为 47。

/ 中等收入群体	随机森林	92.33	91.80	91.69	91.74	98.04
偏低收入群体	逻辑回归	67.06	64.50	72.58	68.30	71.92
/	决策树	72.72	68.59	81.55	74.51	78.90
潜在中等收入群体	随机森林	93.74	92.67	94.69	93.67	98.75
潜在中等收入群体	逻辑回归	61.04	54.29	18.87	28.01	59.97
/	决策树	65.50	59.76	43.07	50.06	67.68
脆弱中等收入群体	随机森林	99.66	99.15	99.99	99.58	99.99
脆弱中等收入群体	逻辑回归	63.85	63.29	64.64	63.96	68.56
/	决策树	70.60	72.71	65.24	68.77	78.48
强中等收入群体	随机森林	97.11	96.61	97.60	97.10	99.67

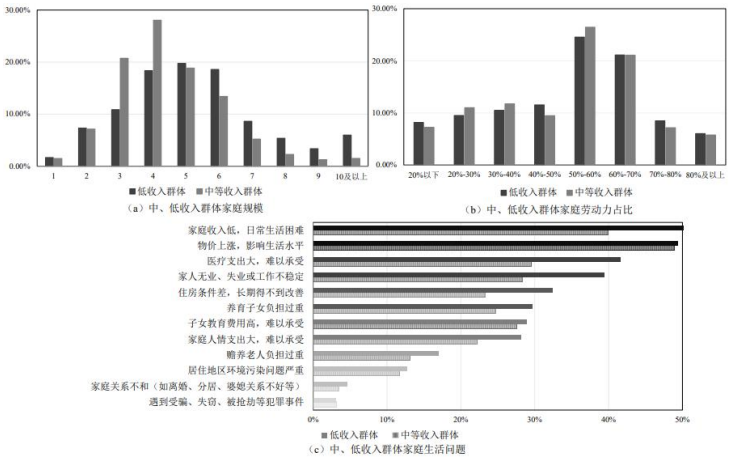
附录 4：实证结果补充



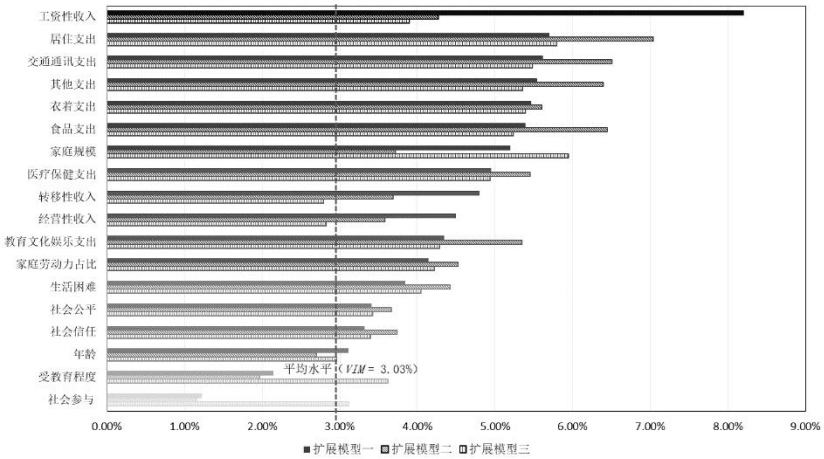
附图 6　基准模型输入变量的特征重要性评分



附图 7　低收入群体与中等收入群体的户主年龄与职业类型分布情况



附图 8　低收入群体与中等收入群体的家庭规模、家庭劳动力占比与生活问题分布情况



附图 9　扩展模型输入变量的特征重要性评分

参考文献

[1] 方匡南, 吴见彬, 朱建平, 等. 信贷信息不对称下的信用卡信用风险研究[J]. 经济研究, 2010, 45(S1): 97-107.

[2] 刘生龙, 张晓明, 杨竺松. 互联网使用对农村居民收入的影响[J]. 数量经济技术经济研究, 2021, 38(04): 103-119.

[3] 刘云菁, 伍彬, 张敏. 上市公司财务舞弊识别模型设计及其应用研究——基于新兴机器学习算法[J]. 数量经济技术经济研究, 2022, 39(07): 152-175.

[4] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[5] Du X, Xu H, Zhu F. Understanding the Effect of Hyperparameter Optimization on Machine Learning Models for Structure Design Problems[J]. Computer-Aided Design, 2021, 135(5): 103013.

[6] Gao J, Nuyttens D, Lootens P, et al. Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery[J]. Biosystems Engineering, 2018, 170: 39-50.

[7] Li Y S, Chi H, Shao X Y, et al. A novel random forest approach for imbalance problem in crime linkage[J]. Knowledge-Based Systems, 2020, 195: 105738.