

基于机器学习方法的农业转移人口 市民化水平影响因素研究*

齐秀琳 汪心如

摘要: 本文基于 2017 年中国流动人口动态监测调查数据,系统地运用多元线性回归、惩罚回归、集成学习和深度学习等多种机器学习方法,考察了农业转移人口市民化水平的影响因素。研究结果表明:集成学习方法在预测农业转移人口市民化水平方面明显优于多元线性回归模型,其中梯度提升回归树模型的预测效果最佳;在所有特征变量中,个体的受教育程度、性别、家庭规模、年龄和流动城市数量是影响农业转移人口市民化水平的最主要因素。此外,本文通过累积局部效应图展示了不同影响因素对农业转移人口市民化水平的具体预测模式,并发现年龄和流动城市数量对农业转移人口市民化水平有着明显的非线性影响。这些研究结论对政府进一步推进农业转移人口市民化具有重要参考价值。

关键词: 农业转移人口 市民化 机器学习

中图分类号: F323.6; C922 **文献标识码:** A

DOI:10.20077/j.cnki.11-1262/f.2024.05.007

一、引言

党的二十大报告明确提出要“推进以人为核心的新型城镇化,加快农业转移人口市民化”^①。农业转移人口市民化并不仅仅是户口性质的转变,还指从农村转移到城镇的人口能够获得城镇永久居住身份、平等享有城镇居民各项社会福利和政治权利,最终成为城镇居民并完全融入城镇社会的过程(魏后凯和苏红键,2013)。根据国家统计局数据,中国城镇常住人口从 2013 年的 7.31 亿人增加至 2023

*本文研究得到河南省哲学社会科学规划项目“河南数字农业发展赋能农业固碳减排的实现路径研究”(编号:2022BJJ095)、河南省高等学校哲学社会科学创新团队支持计划“数字经济与产业创新”(编号:2023-CXTD-01)、郑州大学人文社会科学优秀青年科研团队资助项目“乡村特色产业的培育机制与富农机理研究”(编号:2023-QNTD-02)的资助。感谢匿名评审专家提出的宝贵意见,但文责自负。本文通讯作者:汪心如。

^①习近平,2022:《高举中国特色社会主义伟大旗帜 为全面建设社会主义现代化国家而团结奋斗——在中国共产党第二十次全国代表大会上的报告》,北京:人民出版社,第 32 页。

年的 9.33 亿人，年均增长约 2020 万人，常住人口城镇化率增加了 12.43 个百分点^①。然而，受到多种因素限制，大量农业转移人口并未有效实现市民化（朱信凯，2005；刘金凤等，2023）。快速城镇化和低水平市民化之间的矛盾，不仅直接影响农业转移人口的切身利益，还抑制了消费增长，这在一定程度上阻碍了宏观经济发展（宁光杰和刘丽丽，2018；王青和刘烁，2020）。在此背景下，深入考察和厘清农业转移人口市民化水平的关键影响因素，具有十分重要的现实意义。

农业转移人口市民化是一个内涵丰富且复杂的转变过程，理论上受到多方面因素的影响（刘小年，2017；郭晓欣等，2023）。早期文献主要采用相关性分析方法展开研究（王桂新等，2008）。随着经济学经验研究中“可信性革命”的推进，对农业转移人口市民化水平影响因素的研究也经历了范式的深刻变革，即从考察相关关系转向识别因果关系（宁光杰和李瑞，2016；苏丽锋，2017）。在这一新研究范式下，倾向得分匹配法（祝仲坤和冷晨昕，2020）、工具变量法（祝仲坤，2021）和双重差分法（刘金凤等，2023）等计量经济学方法被广泛采用。虽然这一研究范式的转变极大地推动了学术界对农业转移人口市民化问题的理解，但从更广义的方法论角度来看，无论是相关性分析还是因果推断，本质上都属于解释性建模（陆瑶等，2020）。

与以往研究不同，本文采用多种机器学习方法，通过预测性建模来考察农业转移人口市民化水平的影响因素。相较于解释性建模，本文运用预测性建模展开研究具有三大优势：第一，预测性建模通过放弃估计系数的无偏性，能够更准确地捕捉到农业转移人口市民化水平的影响因素。在探寻变量间关系的建模中，研究者必须在偏差和方差之间进行权衡取舍（王芳等，2020）。解释性建模强调因果关系，致力于获得一致无偏的估计系数，即使这可能会导致方差变大（Athey，2019）。这意味着，虽然解释性建模的估计系数均值可能接近真实值，但单次回归系数的准确性无法得到保证，有时甚至会与真实值相差甚远。与此不同，预测性建模通过接受一定的偏差来换取更小的方差，从而提高了模型预测性能，并增强了模型捕捉关键影响因素的能力。第二，预测性建模不预设模型的具体形式，因此能够更好地刻画变量间的复杂关系。例如，农业转移人口市民化水平的影响因素极其复杂，不仅数量众多，而且不同影响因素之间，以及这些因素与农业转移人口市民化水平之间，均可能存在错综复杂的非线性关系。这种关系隐藏在数据中，且采用任何形式的预设模型都很难确保其形式的正确性，而集成学习和深度学习等机器学习方法由于并不预设具体模型形式，因此更能够挖掘出数据中隐藏的真实关系。这意味着，预测性建模不仅增强了模型的预测能力，也有助于探索解释性建模难以处理的复杂非线性关系。第三，机器学习可解释性方法的发展，不仅在一定程度上解决了机器学习模型过去常为人诟病的“黑箱”问题，还能够揭示解释性建模无法获取的关键信息。例如，沙普利加和解释（SHapley Additive exPlanations，简称 SHAP）等方法可基于拟合性能较高的机器学习模型，清晰地展现各特征变量对响应变量的影响大小，而传统解释性建模由于量纲差异和模型误设等问题，很难获取此类信息。再如，偏依赖图（partial dependence plot，简称 PD）和

^①资料来源：2013 年数据见 https://www.stats.gov.cn/sj/zxfb/202302/t20230203_1898455.html；2023 年数据见 https://www.stats.gov.cn/sj/zxfb/202402/t20240228_1947915.html。常住人口城镇化率=在城镇居住半年以上的常住人口/总人口×100%。

累积局部效应图（accumulated local effects plot，简称 ALE）等方法能够准确捕捉并直观展现数据中复杂的非线性关系，而传统解释性建模为了保持模型的“可解释性”，通常会限制变量高次项的引入，从而忽略了变量间复杂的非线性关系。

正是由于上述优势，机器学习方法已被广泛应用于探索各种重要变量的影响因素，包括经济增长（刘岩和谢天，2019）、公司业绩（陆瑶等，2020）和通货膨胀（肖争艳等，2022）等。本文也属于这一类研究，即利用机器学习方法，通过预测性建模的方式，为理解经典议题提供不同以往的视角和发现。值得指出的是，预测性建模与解释性建模在思路和方法上虽有差异，但并不相互排斥。本文的研究建立在众多基于解释性建模研究的基础之上，并非仅仅基于所谓的“数据驱动”。实际上，预测性建模与解释性建模之间具有一定的互补性：预测性建模不仅为评判解释性建模提供了新视角，而且其从数据中所发掘的新规律，也可以成为解释性建模的新起点（Shmueli，2010；郭峰和陶旭辉，2023）。

本文基于 2017 年中国流动人口动态监测调查数据（China Migrants Dynamic Survey，简称 CMDS），运用多元线性回归、惩罚回归、集成学习和深度学习等方法，系统探讨个体、家庭、迁移以及城市四个维度的特征变量对农业转移人口市民化水平的影响。在通过数据拟合选择性能最优的模型后，本文进一步采用 SHAP 值方法评估不同因素的影响大小，并通过 ALE 图分析受教育程度、家庭规模、年龄、流动城市数量和本地居留时长等特征变量对农业转移人口市民化水平的具体预测模式。本文可能的边际贡献在于：第一，首次综合性地运用多种机器学习方法研究农业转移人口市民化问题，提供与传统解释性建模分析不同的视角和结论，从而丰富相关研究；第二，通过采用前沿的集成学习和深度学习等方法，有效规避多元线性回归模型设定上的局限性，在探讨农业转移人口市民化水平影响因素时，更准确地揭示变量间的复杂关系；第三，利用机器学习中的可解释性方法，探讨不同影响因素对农业转移人口市民化水平的重要性，并分析受教育程度等重要影响因素的具体预测模式，这对政府进一步推动农业转移人口市民化具有重要的参考价值。

二、模型与算法

（一）机器学习模型

1. 多元线性回归模型。本文采用传统的多元线性回归模型作为其他机器学习方法分析结果的参照，回归模型如下：

$$citizenship_i = \alpha + \beta^T X_i + \varepsilon_i \quad (1)$$

（1）式中： α 为截距项， $citizenship_i$ 代表个体 i 的市民化水平； X_i 为个体特征、家庭特征、迁移特征和城市特征等一系列影响因素， β^T 为各影响因素的系数； ε_i 为误差项。为解决可能存在的异方差问题，本文在估计过程中采用了稳健标准误。

2. 惩罚回归方法。本文采用的惩罚回归方法包括 LASSO 回归和岭回归两种。惩罚回归的基本思想是通过引入正则项缓解多重共线性问题和过拟合问题。通常情况下，以平方误差项为损失函数，多元线性回归的优化目标可以表示为：

$$\min_{\beta} \sum_{i=1}^m (\text{citizenship}_i - \beta^T X_i - \alpha)^2 \quad (2)$$

(2) 式中: m 为样本总量。LASSO 回归在 (2) 式中加入一个 L_1 正则项 (参数绝对值之和), 这时优化目标变为:

$$\min_{\beta} \sum_{i=1}^m (\text{citizenship}_i - \beta^T X_i - \alpha)^2 + \lambda \|\beta\|_1 \quad (3)$$

(3) 式中: λ 为正则化参数, $\lambda \|\beta\|_1$ 为惩罚项。若将 L_1 正则项换为 L_2 正则项 (参数平方之和), 则 LASSO 回归变为岭回归。

3. 集成学习方法。本文采用的集成学习方法包括随机森林 (Random Forest, 简称 RF)、梯度提升回归树 (Gradient Boosting Regression Tree, 简称 GBRT) 和极端梯度提升 (eXtreme Gradient Boosting, 简称 XGBoost)。集成学习方法通过组合多个预测效果一般的弱学习器, 以适当的方式形成预测性能更优的强学习器。主要的集成算法包括装袋法、提升法和堆叠法。

(1) 随机森林。随机森林是基于装袋法的一种经典模型。装袋法以决策树作为基本构成单元, 通过自助采样法从训练集中有放回地抽取样本并进行模型训练, 最终通过计算多棵树的预测结果的均值来进行结果整合 (陈强, 2021)。具体步骤如下:

首先, 在训练集中进行有放回地抽样, 得到 B 个自助样本。第 b 个自助样本为:

$$\{x_i^{*b}, y_i^{*b}\}_{i=1}^n, b=1, \dots, B \quad (4)$$

(4) 式中: x_i^* 和 y_i^* 分别为特征变量与响应变量, n 为自助样本的样本容量。

其次, 利用自助样本估计 B 棵不同的决策树, 估计过程中不进行修枝。记第 b 棵树的预测结果为:

$$\{\bar{f}^{*b}(x)\}, b=1, \dots, B \quad (5)$$

最后, 将 B 棵决策树的预测结果取平均处理后, 得到最终预测结果 $\bar{f}_{bag}(x)$:

$$\bar{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \bar{f}^{*b}(x) \quad (6)$$

随机森林建立在上述算法的基础上, 其独特性在于: 在决策树的每个节点分裂时, 它随机选择部分变量 (而非全部变量) 作为候选的分裂变量 (Breiman, 2001)。这种做法显著降低了决策树之间的相关性, 从而减少了样本均值的方差。

(2) GBRT 和 XGBoost。基于提升法, GBRT 和 XGBoost 通过迭代方式来减小模型的预测误差, 逐步提升模型性能。具体地, GBRT 的算法步骤如下 (Friedman, 2001; 陆瑶和施函青, 2022):

首先, 设定初始回归函数为:

$$f_0(x) = \arg \min_{\rho} \sum_{i=1}^m L(y_i, \rho) \quad (7)$$

(7) 式中: $L(\cdot)$ 为损失函数, 当参数 $\rho = f_0(x)$ 时, $\sum_{i=1}^m L(y_i, \rho)$ 取值最小。

其次，循环进行以下 D 次迭代过程 ($d = 1, \dots, D$) :

第一步，计算损失函数负梯度 $\psi_{i,d} = -\frac{\partial L(y_i, f(x))}{\partial x}$ ，其中， $f(x) = f_{d-1}(x)$ ；

第二步，以 $\psi_{i,d}$ 为残差近似值拟合出新回归树 $g_d(x) = E(\psi | x)$ ；

第三步，选择使误差最小的梯度下降幅度 $\tau = \arg \min_{\tau} \sum_{i=1}^N L(y_i, f_{d-1}(x) + \tau g_d(x))$ ；

第四步，计算新的预测函数 $f_d(x) = f_{d-1}(x) + \tau g_d(x)$ 。其中， ν 为收缩参数。

最后，经过 D 次迭代后的模型 $f_D(x)$ 即为最终预测模型。

XGBoost 在 GBRT 的基础上提高了训练效率，并在拟合过程中利用了二阶导数信息。具体地，XGBoost 在损失函数中加入了 $L(y_i, f_{d-1}(x) + \tau g_d(x))$ 对 $f_{d-1}(x)$ 的一阶导数和二阶导数。此外，XGBoost 还加入了针对回归树 $g_b(x)$ 中节点的 L_1 和 L_2 范数惩罚项，以控制模型复杂度。

4. 深度学习方法。本文采用的深度学习方法为前馈神经网络 (Feedforward Neural Network, 简称 FNN) 模型。FNN 模型是一种由人工神经元和权重连接构成的模型，在计算过程中，每一层的神经元都能接收到前一层神经元的信号，并向下一层传递信号。在结构上，FNN 模型包含一个输入层、多个隐藏层以及一个输出层。信号在这些层间单向传播，隐藏层利用非线性变换对输入数据进行特征转换，使模型能够有效学习并预测复杂模式。FNN 模型的基本结构如图 1 所示 (陈强, 2021)：

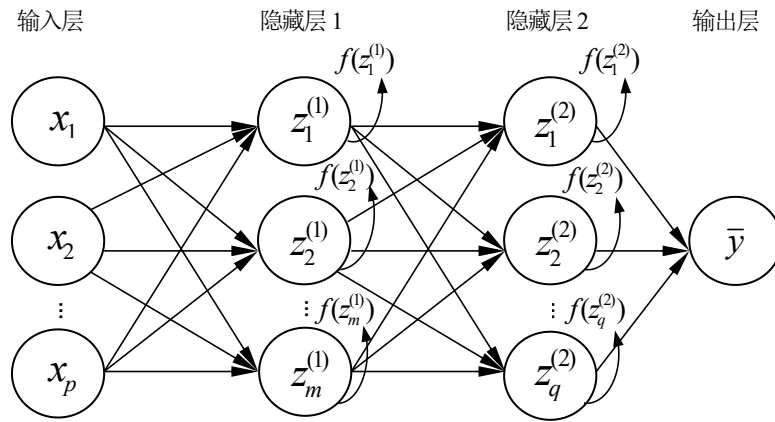


图1 前馈神经网络的基本结构

注： $f(\cdot)$ 为激活函数。 z 为 x 在施加激活函数之前的加总值。

(二) 模型的参数调整与选择

1. 模型的参数调整。包括集成学习和深度学习在内的机器学习方法在捕捉变量间的复杂关系时具有优势，但所构建的模型可能存在过拟合问题。鉴于此，本文采用 K 折交叉验证来确定最优参数 (刘景江等, 2023)。具体而言，首先，通过随机抽样的方式将数据分为训练集和测试集，比例为 8 : 2；其次，在训练集上执行 K 折交叉验证，通常 K 值取 2 至 10 之间。考虑到本文样本量较小，本文选择 5 折交叉验证以平衡模型训练的稳定性、准确性和计算资源消耗。在此过程中，训练集被均匀划分为

5 个子集，每次用 4 个子集进行训练，剩下的 1 个子集用于验证，这一过程重复 5 次；最后，根据交叉验证结果确定各机器学习模型的最优参数。

经过上述调参过程，对不同模型参数设定如下：①惩罚系数（ α ）：LASSO 回归为 0.01，岭回归为 10；②决策树数量（ $n_estimators$ ）：随机森林为 120，GBRT 为 155，XGBoost 为 150；③决策树深度（ max_depth ）：随机森林为 10，GBRT 和 XGBoost 为 3；④学习率（ $learning_rate$ ）：GBRT 和 XGBoost 为 0.1；⑤FNN 的隐藏层数为 3，每层神经元个数分别为 9、9、6。

2. 模型选择。为选出最优模型，本文参考陆瑶等（2020）和陈运森等（2023）的研究，利用样本内拟合优度 R_{IS}^2 、样本外拟合优度 R_{OOS}^2 、可解释方差 EVS_{OOS} 、均方误差 MSE_{OOS} 、平均绝对误差 MAE_{OOS} 以及绝对中位差 $MedAE_{OOS}$ 六个指标对模型进行评价。各指标含义与计算公式如下：

$$R_{IS}^2(R_{OOS}^2) = 1 - \frac{\sum_{i=1}^n (y_i - y^p)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

$$EVS_{OOS} = 1 - \frac{\sum_{i=1}^n (y^p - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

$$MSE_{OOS} = \frac{1}{n} \sum_{i=1}^n (y^p - y_i)^2 \quad (10)$$

$$MAE_{OOS} = \frac{1}{n} \sum_{i=1}^n |y^p - y_i| \quad (11)$$

$$MedAE_{OOS} = \text{median}|y_i - y^p| \quad (12)$$

(8) ~ (12) 式中： y_i 为样本真实值， $\bar{y}$ 为所有真实值的平均值， y^p 为预测值。值得说明的是，样本内拟合优度 R_{IS}^2 和样本外拟合优度 R_{OOS}^2 的计算公式相同，差别在于前者利用训练集数据计算得出，后者利用测试集数据计算得出。这些指标是本文模型选择的基础。首先， R_{IS}^2 反映了模型预测值与训练集实际观测值的拟合程度，该指标越大，模型对训练集样本的解释力越强。其次， R_{OOS}^2 和 EVS_{OOS} 分别反映了模型预测值与测试集实际观测值的拟合程度，以及模型预测值变异程度与测试集实际观测值变异程度的拟合程度。这两个指标越大，模型的泛化能力越好。最后， MSE_{OOS} 、 MAE_{OOS} 和 $MedAE_{OOS}$ 分别代表了样本外预测值与实际观测值之差的平方期望值、绝对值的期望和绝对值的中位数。这三个指标越小，模型预测的准确性越高。

（三）机器学习的可解释性方法

1. SHAP 值方法。SHAP 值方法相较于其他测度变量重要性的方法具有两大优点：一是其理论基础更为坚实，该方法提供了一个统一的框架来对影响因素的重要性进行排序（Lundberg and Lee, 2017）；二是在进行贡献度评价时，该方法满足了一致性条件（周卫华等，2022）。所谓一致性，是指当改变某个特征边际贡献度时，其他特征的边际贡献度不会改变。因此，为了测度并比较农业转移人口市民

化水平不同影响因素的重要性,本文采用了 SHAP 值方法。该方法源于合作博弈论中的 Shapley 值概念,其核心思想是通过计算各影响因素加入模型时的平均边际贡献,来衡量它们对模型预测结果的贡献程度 (Lundberg and Lee, 2017)。具体地,第 j 个影响因素的 SHAP 值计算公式如下:

$$SHAP_*^j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} (v_*(S \cup \{j\}) - v_*(S)) \quad (13)$$

(13) 式中: F 为所有影响因素的全集, $|F|$ 为 F 中元素个数; S 为不包含第 j 个影响因素的集合,其元素个数为 $|S|$ 。 $(v_*(S \cup \{j\}) - v_*(S))$ 为当影响因素组合为 S 时,第 j 个影响因素对预测值的期望的影响大小。对于利用集合 S 所得期望 $v_*(S)$,本文采用 Aas et al. (2021) 的方法进行计算以克服变量相关性对结果的影响。

2. ALE 图。本文采用 ALE 图来描述各影响因素对农业转移人口市民化水平的具体预测模式。以往文献常用偏依赖图刻画影响因素的预测模式,但该方法的使用需要假设各影响因素之间相互独立。相比之下, ALE 图通过计算局部效应消除了变量相关性的影响。鉴于农业转移人口市民化水平影响因素之间可能存在复杂的相互关系,所以 ALE 图更适合本研究。在具体操作中,首先,将特征变量的取值范围划分为若干区间,确保每个区间内具有相同数量的数据点。其次,计算每个区间内的局部效应,以此得出特征变量对模型预测的累积局部效应值。相关计算公式如下:

$$ALE^*(x_j) = \sum_{k=1}^{k(x_j)} \frac{1}{n_j(k)} \sum_{i: x_j^{(i)} \in N_j(k)} [\hat{f}(z_{k,j}, x_{-j}^{(i)}) - \hat{f}(z_{k-1,j}, x_{-j}^{(i)})] \quad (14)$$

(14) 式中: $ALE^*(x_j)$ 表示特征变量 x_j 未经中心化处理的累积局部效应值; k 代表划分的区间个数; $n_j(k)$ 代表第 j 个特征在第 k 个区间内的样本数量; $z_{k-1,j}$ 和 $z_{k,j}$ 分别是第 $k-1$ 和第 k 个区间的网格值,代表特征 x_j 在第 $k-1$ 和第 k 个区间的边界值; $N_j(k)$ 是第 k 个区间内数据点索引集合; $\hat{f}(z_{k,j}, x_{-j}^{(i)})$ 是特征 x_j 达到边界值 $z_{k,j}$ 时的模型预测值, $x_{-j}^{(i)}$ 代表将第 i 个样本中的 x_j 特征排除后的特征集合。

本文将 (14) 式计算出的累积局部效应值进行中心化处理后,可获得特征变量 x_j 的累积局部效应值。计算公式见 (15) 式:

$$ALE(x_j) = ALE^*(x_j) - \frac{1}{n} \sum_{i=1}^n ALE^*(x_j) \quad (15)$$

三、变量说明和数据来源

(一) 变量说明

1. 响应变量。本文的响应变量为农业转移人口市民化水平。在社会发展的不同阶段,农业转移人口市民化的含义有所不同 (蔡昉, 2013; 魏后凯和苏红键, 2013; 洪银兴等, 2021)。早期,学者普遍认为,农业转移人口市民化的关键在于获得城市户籍 (刘传江和徐建玲, 2007)。然而,随着社会

福利和公共服务等与户籍逐步脱钩，城市户籍对农民的吸引力下降，“落户”不再直接等同于市民化（邹一南，2021）。此外，考虑到农业转移人口实现市民化的实际障碍，在新一轮户籍制度改革的推动下，农业转移人口在城市落户的门槛已显著降低。目前，农业转移人口市民化水平不高主要体现在就业不稳定、发展能力弱、价值观念与城市不同步以及归属感缺失等方面（Chen et al., 2018; 李爱民等，2022）。因此，在现今背景下，农业转移人口市民化不仅仅是户籍身份的转变，更是包括生产和生活方式以及价值观念的全面转型。

关于农业转移人口市民化过程的理解，以往研究通常分为二段论（刘传江和徐建玲，2007）、三段论（王桂新等，2008）和四段论（王春超和蔡文鑫，2021）。本文基于四段论对农业转移人口市民化的内涵进行界定：一是经济市民化，即农业转移人口拥有负担城市生活成本的经济能力（刘传江和程建林，2009）。这种经济能力主要涵盖就业状况、消费能力和收入水平等方面（何军，2011；苏丽锋，2017）。目前农业转移人口在这些方面与城市居民相比仍存在较大差距（王春超和蔡文鑫，2021）。二是公共服务市民化，即农业转移人口应享受与本地城市居民同等水平的公共医疗和教育资源（李爱民等，2022；陈锡文，2023）。研究显示，更好的医疗保障有助于农业转移人口更好地融入城市（秦立建和陈波，2014）。同时，子女是否能够获得流入地的教育资源，也是衡量农业转移人口市民化水平的重要标准（佟大建等，2022）。三是社会市民化，它反映了农业转移人口市民化过程中非常重要的非物质维度（王桂新等，2008），包括个体对社会组织参与程度（涂尔干，2000）。农业转移人口从农民到市民的转变，不仅是生活空间的改变，还包括适应并习得城市居民的交往规则和社会规范，积极参与城市社会和政治活动，实现对城市社会的“嵌入”（梁波和王海英，2010；魏后凯和苏红键，2013）。四是观念市民化，即农业转移人口在观念上认同自己是城市居民。一方面，市民化意愿是农业转移人口实现身份转变的前提（苏丽锋，2017）；另一方面，被本地市民接纳的自我感知和对本地人身份的自我认同，也体现了农业转移人口观念市民化的水平（辛宝英，2016）。

根据以上分析，本文从经济市民化、公共服务市民化、社会市民化和观念市民化四个维度选择变量，建立指标体系来衡量农业转移人口的市民化水平。为了避免由流入地差异导致的衡量偏差，本文参考以往文献的做法，采用比值法测度进入指标体系的各变量（魏后凯和苏红键，2013；王春超和蔡文鑫，2021；佟大建等，2022）。例如，如果农业转移人口的月收入为4000元，他们在一些中小城市（如洛阳）可能表现出较高的经济市民化水平，而在一线大城市（如上海）则可能较低。这表明，只有通过与流入地居民水平的比较，才能合理地衡量市民化水平。具体地，本文首先计算每个变量在流入地城市居民中的中位数，然后将农业转移人口在该变量上的取值与中位数相比，作为衡量农业转移人口市民化水平的基础。在对指标进行赋权加总时，考虑到熵值法等基于变量变异程度的赋权方式可能会掩盖或扭曲某些关键变量的贡献，本文借鉴以往文献（宁光杰和李瑞，2016；苏丽锋，2017），使用等权重法来对构成总指标的子指标进行赋权。具体地，同一上层指标下的所有下层指标被赋予相同的权重。

农业转移人口市民化水平测度指标体系及说明见表1。

表 1 农业转移人口市民化水平测度指标体系		
一级指标	二级指标	三级指标
经济市民化	就业状况	周工作小时数（小时） 签订劳动合同情况：有固定期限=6，无固定期限=5，完成一次性工作任务=4，试用期=3，未签订劳动合同=2，不清楚=1 单位性质：国有和集体=4，私企=3，个体=2，无单位=1
	消费能力	家庭人均月消费支出（元）
	收入水平	月劳动收入（元）
公共服务市民化	基础医疗可及性	身体不适时选择去哪里看病：本地综合或专科医院=6，本地社区卫生站=5，本地个体诊所=4，本地药店=3，老家或除本地和老家以外的其他地方=2，没治疗=1
	医疗保险	参加医疗保险情况 ^a ：参加公费医疗=4，参加城镇居民医疗保险或城镇职工医疗保险=3，参加城乡居民合作医疗保险=2，参加新型农村合作医疗保险=1
	子女教育	目前在本地没有子女上学的困难：是=1，否=0
	居住证	是否办理居住证：是=1，否=0
社会市民化	在流入地社会活动参与	公益活动参与频繁程度：经常=4，有时=3，偶尔=2，没有=1 参与社区管理频繁程度：经常=4，有时=3，偶尔=2，没有=1
	在流入地政治活动参与	向政府部门提出政策建议频繁程度：经常=4，有时=3，偶尔=2，没有=1 参与党务活动频繁程度：经常=4，有时=3，偶尔=2，没有=1
观念市民化	融入意愿	愿意融入本地人当中：完全同意=4，基本同意=3，不同意=2，完全不同意=1
	被接纳程度	本地人愿意接纳我：完全同意=4，基本同意=3，不同意=2，完全不同意=1
	身份认同	自我感觉是本地人：完全同意=4，基本同意=3，不同意=2，完全不同意=1

注：a 根据本文所用样本计算，就四类保险中农业转移人口在流入地的参保比例而言（流入地参保人数/流入地和户籍地的参保总人数），新型农村合作医疗保险最低（2.94%），城乡居民合作医疗保险次之（41.36%），城镇居民医疗保险或城镇职工医疗保险最高（96.34%）。这意味着，对不同医疗保险的参与在一定程度上能够反映农业转移人口享受本地公共服务的水平。

2.特征变量。据以往研究，农业转移人口市民化水平的影响因素既包括微观层面的个体特征、家庭特征和迁移特征，也包括宏观层面的流入地城市特征。具体如下：

①个体特征。研究表明，性别、年龄、政治面貌、婚姻状况、健康状况及受教育程度等个体差异对农业转移人口市民化水平有显著影响（张斐，2011；王孝莹和王目文，2020；朱纪广等，2020；刘金凤等，2023）。②家庭特征。根据新迁移经济学理论，迁移决策通常并非由个体独自做出，而是由更多相关人员组成的一个更大单位（如家庭）集体决定的（Stark，1991）。这意味着，家庭特征会显著影响农业转移人口的市民化水平。以往研究表明，家庭结构性特征（如家庭规模和子女数量）和家庭财产性特征（如老家是否有困难、老家耕地承包情况、老家宅基地情况和老家村集体分红情况）都是影响农业转移人口市民化水平的关键因素（王孝莹和王目文，2020；李国正，2020）。③迁移特征。首先是迁移方式，有研究表明，当有家属随迁时，农业转移人口可以更好地融入城市（李强和龙文进，2009）；其次是流动距离，有研究发现，省内流动的农业转移人口市民化意愿和能力都高于跨省流动

的样本（宁光杰和李瑞，2016）；最后是迁移时长，有学者指出，农业转移人口在本地居留时长和离开家（户籍地）时长越久，其市民化水平越高（Hu et al., 2011；王成利和王洪娜，2020）。④流入地城市特征。除了上述微观层面的特征外，流入地所具有的城市特征也影响着农业转移人口市民化水平（张占斌等，2013；刘静和张锦华，2021）。一方面，城市发展水平和公共服务状况是影响农业转移人口市民化水平的关键因素（吴业苗，2012；张继良和马洪福，2015；苏丽锋，2017；祝仲坤，2021）；另一方面，户籍等正式制度和以社会包容性为主要表现的非正式制度也对农业转移人口市民化水平产生了显著影响（王桂新等，2008；徐延辉和龚紫钰，2019）。

在前述分析的基础上，本文从个体特征、家庭特征、迁移特征和城市特征四个维度出发，选择了总计 37 个特征变量。除了以往文献中常见的性别、年龄等变量之外，本文还引入了流动城市数量、方言多样性以及用于衡量城乡差距的变量（徐现祥等，2015）。这些变量涵盖了城乡收入差距以及城乡公共服务差距。后者包括城乡教育水平差距、城乡医疗服务水平差距和城乡社会保障水平差距，均以城乡均等化水平衡量（黄寿峰和赵岩，2023）。具体测算方法为：首先，选择衡量不同公共服务水平的指标，其中，教育水平包括每万人拥有普通小学学校数、每万人拥有普通中学学校数、普通小学师生比和普通中学师生比，医疗服务水平包括每万人拥有医院卫生院机构数、每万人拥有医院卫生院床位数和每万人拥有执业或助理医师数，社会保障水平包括基本养老保险参保率、基本医疗保险参保率和失业保险参保率；其次，利用熵值法分别测算各地农村和城市公共服务供给水平；最后，用农村与城市公共服务供给水平之比表示均等化程度。本文主要变量的说明和描述性统计见表 2。

表 2 变量说明与描述性统计

变量分类	变量名称	变量说明	均值	标准差	最小值	最大值
响应变量	市民化水平	依照农业转移人口市民化水平测度指标体系（表 1）进行测算	0.996	0.184	0.559	2.157
特征变量	个体特征					
	性别	受访者性别：男=1，女=0	0.489	0.500	0	1
	年龄	受访者年龄（岁）	33.930	9.164	15	60
	婚姻状况	受访者婚姻状况：未婚=1，已婚=0	0.168	0.374	0	1
	民族	受访者民族：汉族=1，其他=0	0.972	0.164	0	1
	健康状况	受访者健康状况：健康=4，基本健康=3，不健康但生活能自理=2，生活不能自理=1	3.853	0.391	1	4
	受教育程度	受访者受教育程度：研究生=7，大学本科=6，大学专科=5，高中/中专=4，初中=3，小学=2，未上过小学=1	3.501	1.021	1	7
	政治面貌	受访者政治面貌：中共党员=1，其他=0	0.034	0.180	0	1
	家庭特征					
	家庭规模	受访者家庭成员人数（人）	3.043	1.081	1	5
	子女数量	受访者家庭中 16 岁以下人口数（人）	0.858	0.800	0	4

表2 (续)

特征变量	老家耕地承包情况	受访者在老家(户籍所在地)是否有承包耕地: 有=1, 没有/不清楚=0	0.528	0.499	0	1
	老家村集体分红情况	受访者在老家(户籍所在地)是否有集体分红: 有集体分红=1, 没有/不清楚=0	0.035	0.184	0	1
	老家宅基地情况	受访者在老家(户籍所在地)是否有宅基地: 有宅基地=1, 没有/不清楚=0	0.726	0.446	0	1
	老家是否有困难	受访者在老家(户籍所在地)是否有困难(包括老人赡养、子女照看、子女教育费用、配偶生活孤单、家人有病缺钱治、土地耕种等缺劳动力, 以及其他困难): 有困难=1, 没有困难=0	0.613	0.487	0	1
	迁移特征					
	家属随迁	受访者是否有家属随迁: 有=1, 没有=0	0.383	0.486	0	1
	流动距离	受访者相对户籍所在地的流动范围: 跨省=3, 省内跨市=2, 市内跨县=1	2.386	0.597	1	3
	流动城市数量	受访者总共流动(跨区县1个月及以上, 以工作、生活等为目的)的城市数量(个)	1.929	1.037	1	5
	本地居留时长	受访者自进入当前流入地以来居留时长(年)	5.097	3.503	1	15
	离开家(户籍地)时长	受访者多长时间没有回过老家(户籍所在地)(年)	1.222	0.982	1	24
	城市特征					
	地区生产总值	地区生产总值(亿元)	8779.942	4568.113	2707.529	21503.151
	产业结构	第三产业增加值占地区生产总值的比重(%)	59.507	8.117	50.810	71.750
	普通小学师生比	普通小学专任教师数(人)/普通小学在校学生数(人)	0.055	0.004	0.051	0.061
	普通中学师生比	普通中学专任教师数(人)/普通中学在校学生数(人)	0.083	0.009	0.071	0.098
	每万人拥有医院卫生院机构数	医院卫生院数(个)/年末总人口数(万人)	0.376	0.126	0.235	0.586
	每万人拥有医院卫生院床位数	医院卫生院床位数(张)/年末总人口数(万人)	84.061	24.230	47.634	128.248
	每万人拥有执业或助理医师数	执业或助理医师数(人)/年末总人口数(万人)	48.510	12.396	22.433	68.410
	每万人拥有公共汽车数量	年末实有公共营运汽电车(辆)/年末总人口(万人)	12.265	4.805	2.707	19.739
	城镇职工基本养老保险参保情况	城镇职工基本养老保险参保人数(人)/年末总人口数(人)	0.598	0.205	0.278	0.844
	城镇职工失业保险参保情况	城镇职工失业保险参保人数(人)/年末总人口数(人)	0.403	0.189	0.144	0.656

表 2 (续)

特征 变量	生活垃圾无害化处理率	生活垃圾无害化处理量与生活垃圾产生量之比	0.999	0.001	0.997	1.000
	污水处理厂集中处理率	污水处理厂处理污水量与污水排放总量之比	0.956	0.025	0.910	0.987
	最低工资	城市每小时最低工资与社会在岗职工平均每小时工资之比	0.823	0.053	0.764	0.915
	户籍开放度	户籍开放度越高，表示户籍管制对劳动力流动的阻碍作用越小	0.663	0.214	0.287	0.865
	方言多样性	城市方言数量（种）	1.557	0.646	1	3
	城乡收入差距	城镇居民人均可支配收入（元）/农村居民人均可支配收入（元）	2.143	0.287	1.716	2.547
	城乡教育水平差距	城乡教育均等化水平	0.847	0.100	0.735	1
	城乡医疗服务水平差距	城乡医疗服务均等化水平	0.777	0.199	0.351	1
	城乡社会保障水平差距	城乡社会保障均等化水平	0.723	0.352	0.002	1

注：最低工资和户籍开放度的具体计算方法参考中国人民大学国家发展与战略研究院发布的《中国劳动力市场化指数编制》；城乡教育水平差距、城乡医疗服务水平差距和城乡社会保障水平差距的计算方式见上文；为了使描述性统计有意义，表中对地区生产总值的描述性统计结果为其原值的统计值，在后文分析中则对其进行取对数处理。

（二）数据来源

本文用于测量农业转移人口市民化水平的数据源自 2017 年 CMDS 数据。CMDS 采用分层、多阶段、按规模大小成比例的概率抽样方法进行抽样。相较于其他年份的数据，CMDS（2017）问卷中不仅包含了大量有关农业转移人口市民化的问题，而且提供了流入地城市居民的相关信息，非常适合本研究的需求^①。CMDS（2017）（C 卷）的访谈对象为城市中的流动人口。为了筛选出合适的研究样本，首先，本文选择具有农业户口的个体；其次，删除在重要变量上有缺失值或异常值的样本。经过筛选，最终的观测样本总数为 7619 个。为了精准测算农业转移人口市民化水平，本文还使用了 CMDS（2017）（D 卷）中报告的城市居民信息。本文中，农业转移人口的个体、家庭和迁移特征变量数据来自 CMDS（2017），城市特征变量中最低工资和户籍开放度数据来自中国人民大学国家发展与战略研究院 2019 年 3 月发布的《中国劳动力市场化指数编制》，方言多样性数据来自《汉语方言大词典》，其他城市特征变量数据来自 2018 年《中国城市统计年鉴》和各省市统计年鉴。

四、实证结果和分析

（一）多元线性回归结果

表 3 报告了多元线性回归结果。从回归结果来看，在其他条件不变的情况下，个体特征中的男性身份、受教育程度和党员身份显著提高了农业转移人口市民化水平；家庭特征中的家庭规模对农业转

^①在目前公开的可用于研究农业转移人口市民化水平影响因素的大型微观数据库中，仅有 CMDS（2017）调查并报告了流动人口所在城市居民的相关信息。

移人口市民化水平具有显著负向影响，而子女数量具有显著正向影响；迁移特征中家属随迁、流动城市数量和本地居留时长对农业转移人口市民化水平具有显著正向影响。

上述结果在一定程度上印证了以往文献的发现（张斐，2011；刘金凤等，2023），但也存在一些局限。一方面，尽管以往文献在进行“假设一检验”时多将模型设定为线性形式，但通过偏残差图分析，本文发现年龄等变量与农业转移市民化水平之间并非线性关系。当然，这时可以通过加入高阶项或交互项来捕捉这些非线性关系，但该做法并不能确保模型形式正确匹配数据结构。另一方面，本文虽然采用了稳健标准误来处理潜在的异方差问题，但在纳入大量解释变量后，仍面临难以克服的多重共线性问题（VIF 最大值为 21.73）。虽然可以通过筛选变量的方法来缓解多重共线性问题，但这可能会损害影响因素分析的全面性和系统性。鉴于此，本文通过引入惩罚回归、集成学习和深度学习等方法，以期更全面和深入地分析农业转移人口市民化水平的影响因素。

表 3 农业转移人口市民化水平影响因素的多元线性回归结果

	被解释变量：农业转移人口市民化水平						
	性别	年龄	婚姻状况	民族	健康状况	受教育程度	政治面貌
系数	0.0383**	-0.0001	-0.0021	-0.0226	-0.0146	0.0367***	0.0949***
稳健标准误	(0.0126)	(0.0002)	(0.0109)	(0.0150)	(0.0089)	(0.0042)	(0.0145)
	家庭规模	子女数量	老家耕地承包情况	老家村集体分红情况	老家宅基地情况	老家是否有困难	
	-0.0260***	0.0091*	0.0106	0.0568	0.0037	0.0036	
系数							
稳健标准误	(0.0026)	(0.0036)	(0.0062)	(0.0292)	(0.0088)	(0.0038)	
	家属随迁	流动距离	流动城市数量	本地居留时长	离开家（户籍地）时长		
	0.0139*	0.0033	0.0144**	0.0043***	0.0031		
系数							
稳健标准误	(0.0057)	(0.0066)	(0.0045)	(0.0009)	(0.0022)		
	地区生产总值	产业结构	普通小学师生比	户籍开放度	方言多样性		
	0.6070	0.0047	-5.0760	8.3150	2.9403		
系数							
稳健标准误	(0.5544)	(0.0041)	(4.7788)	(7.9092)	(2.7611)		

注：①***、**和*分别表示 1%、5%和 10%的显著性水平；②部分城市特征变量因多重共线性问题在回归中被剔除。

（二）基于不同方法的模型预测效果评价

表 4 展示了基于不同方法的模型对农业转移人口市民化水平的预测效果。首先，惩罚回归中的 LASSO 回归预测效果明显差于多元线性回归，岭回归与多元线性回归预测效果基本相同；其次，尽管前馈神经网络在拟合能力上并不优于多元线性回归，但其在泛化能力方面表现更佳；最后，集成学习方法在预测效果上表现最为优秀。以表现最好的 GBRT 模型为例，其三个正向指标 R^2_{IS} 、 R^2_{OOS} 和 EVS_{OOS} 相较于多元线性回归模型分别提升了 56.90%、69.10%和 73.26%，而三个负向指标 MSE_{OOS} 、 MAE_{OOS} 和 $MedAE_{OOS}$ 相较于多元线性回归分别下降了 21.86%、10.83%和 7.14%。

集成学习方法之所以在预测效果上优于多元线性回归模型，主要是因为农业转移人口市民化水平的影响因素极其复杂。无论是特征变量之间的相互作用，还是特征变量与响应变量之间的非线性关系，传统的多元线性回归模型都难以精确捕捉（Mullainathan and Spiess，2017）。这充分表现为样本内拟合优度小、样本外泛化能力弱以及各类型的误差大等问题。相较之下，集成学习并不预设任何模型形式，而是直接基于数据集来训练模型，从而可以更好地处理各种非线性关系。因此，下文将主要基于预测效果最优的 GBRT 模型进行分析。

表 4 基于不同方法的模型预测效果评价

模型	(1) R^2_{IS}	(2) R^2_{OOS}	(3) EVS_{OOS}	(4) MSE_{OOS}	(5) MAE_{OOS}	(6) $MedAE_{OOS}$
多元线性回归	0.1413	0.1149	0.1152	0.0311	0.1404	0.1162
LASSO 回归	0.0958	0.0827	0.0831	0.0322	0.1422	0.1187
岭回归	0.1413	0.1150	0.1153	0.0311	0.1404	0.1163
随机森林	0.3381	0.1940	0.1941	0.0285	0.1338	0.1139
GBRT	0.2217	0.1943	0.1996	0.0243	0.1252	0.1079
XGBoost	0.2201	0.1674	0.1675	0.0313	0.1364	0.1148
前馈神经网络	0.1375	0.1195	0.1202	0.0309	0.1400	0.1172

（三）不同特征变量的重要性排序

为进一步探求不同特征变量对农业转移人口市民化水平的影响差异，本文利用 SHAP 值方法测度并比较不同特征变量对农业转移人口市民化水平的预测能力。表 5 展示了特征变量重要性排序的前十位。根据 GBRT 模型和随机森林的结果，农业转移人口的受教育程度、性别、家庭规模、年龄和流动城市数量位列前五，表明这些因素对农业转移人口市民化水平的预测效果最好。同时，表 5 显示这五个特征变量在两种模型中的排名基本相同，证明了结论的稳健性。此外，表 5 的结果既验证了以往研究的发现，即受教育程度等因素对农业转移人口市民化水平具有重要影响（刘金凤等，2023；郭晓欣等，2023），也提供了新的见解：由于存在量纲和模型设定等问题，传统计量方法难以准确比较不同因素的影响力度，而 SHAP 值方法可以解决该问题。

表 5 基于 SHAP 值方法的特征变量重要性排序

排名	GBRT		随机森林	
	变量名称	SHAP 均值	变量名称	SHAP 均值
1	受教育程度	0.0265	受教育程度	0.0259
2	性别	0.0205	性别	0.0191
3	家庭规模	0.0176	家庭规模	0.0099
4	年龄	0.0144	流动城市数量	0.0094
5	流动城市数量	0.0130	年龄	0.0081
6	本地居留时长	0.0125	本地居留时长	0.0062
7	户籍开放度	0.0086	户籍开放度	0.0055
8	政治面貌	0.0047	政治面貌	0.0051

表 5 (续)

9	子女数量	0.0046	每万人拥有医院卫生院机构数	0.0041
10	老家宅基地情况	0.0046	污水处理厂集中处理率	0.0034

(四) 主要特征变量对农业转移人口市民化水平的预测模式

本文接下来利用 ALE 图来探讨最重要的几个特征变量对农业转移人口市民化水平的具体预测模式。考虑到性别是一个二值变量, 本文选择了在特征变量重要性排序中位列第六位的本地居留时长作为替代, 分别绘制了受教育程度、家庭规模、年龄、流动城市数量和本地居留时长等特征变量的 ALE 图。其中, 纵轴代表累积局部效应的大小, 即特征值变化对模型预测输出的平均影响程度; 黑色实线的斜率表示特征变量的边际效应, 正值意味着特征变量与响应变量正相关, 负值则表示负相关。ALE 图中的虚线代表蒙特卡洛样本的边际分布, 反映了模型在特定特征值下预测的不确定性。虚线越接近实线, 表明在该特征值附近蒙特卡洛样本的预测值越一致, 模型的不确定性也就越低。

1. 受教育程度。图 2 展示了农业转移人口受教育程度与市民化水平之间的 ALE 图。由图 2 可知, 随着受教育程度的提高, 农业转移人口的市民化水平也相应提升。农业转移人口市民化的一个基本前提是其拥有在城市中生存和生活的能力, 而个体受教育水平在这一过程中起到了至关重要的作用。拥有更高的受教育程度, 意味着个体在就业市场上更有竞争力, 更有机会获得更为优越的工作机会以及更高的收入水平, 这极大地促进了其市民化水平的提升(单菁菁, 2014)。此外, 图 2 显示受教育程度与农业转移人口市民化水平之间存在近似的线性关系。这种关系虽然可以通过解释性模型的拟合方法得到类似的线性图形来呈现, 但两者在本质上并不相同: 解释性建模来自对模型形式的设定; 而 ALE 图揭示的是来自数据的、受教育程度与市民化水平之间“真实”的近似线性的关系。换言之, ALE 图所展现的变量关系即使与基于解释性模型的研究结论相符, 也不是对以往研究结论的简单重复。

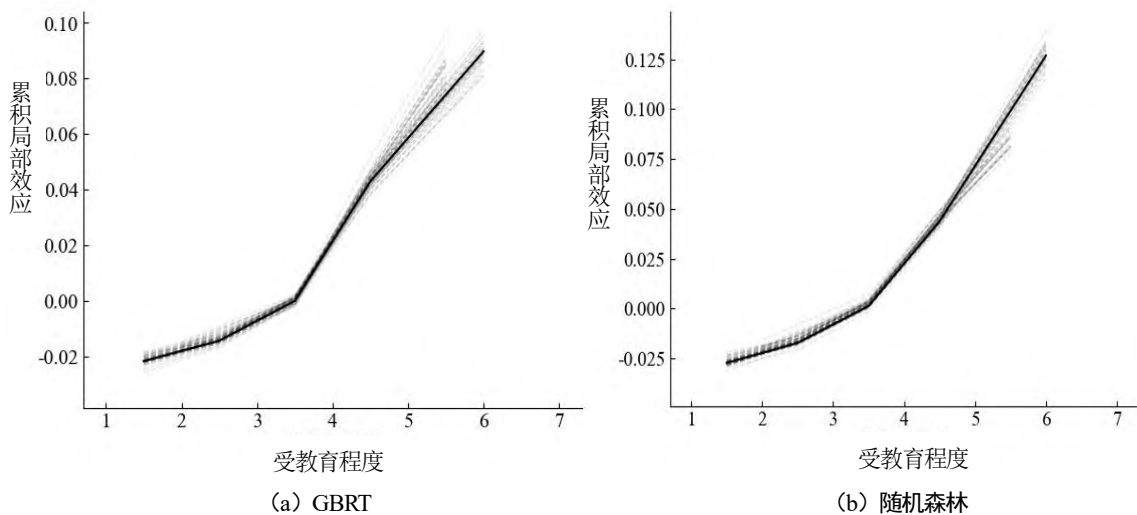


图 2 农业转移人口受教育程度与市民化水平的 ALE 图

2. 家庭规模。图 3 展示了农业转移人口家庭规模与市民化水平之间的 ALE 图。由图 3 可知, 当家庭规模扩大时, 农业转移人口市民化水平会先迅速下降, 后下降速度逐渐减缓。可能的原因在于: 一

方面,家庭规模扩大会带来更高的生活压力,从而削弱农业转移人口市民化的能力(苏丽锋,2017),因此,家庭规模扩大降低了农业转移人口的市民化水平;另一方面,随着家庭规模的进一步扩大,家庭成员共同分担市民化成本的能力也在提高(孙战文和杨学成,2013),这将在一定程度上减缓农业转移人口市民化水平的下降速度。

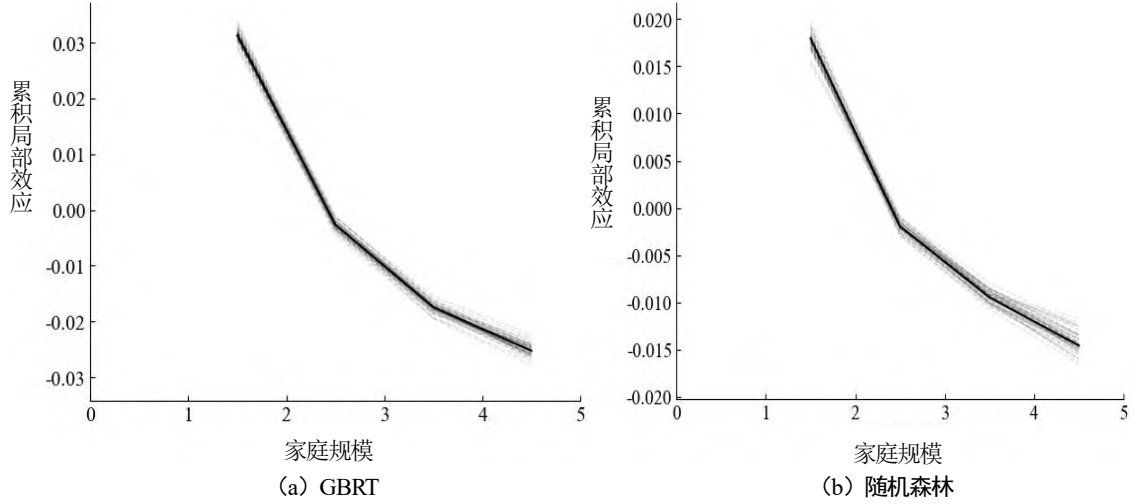


图3 农业转移人口家庭规模与市民化水平的 ALE 图

3.年龄。图4展示了农业转移人口年龄与市民化水平之间的 ALE 图。由图4可知,当个体年龄在33岁以下时,农业转移人口市民化水平随着年龄增加而提高;个体年龄超过33岁,农业转移人口市民化水平则随着年龄增加而下降。图4表明农业转移人口的年龄与市民化水平之间呈先上升后下降的倒U型关系,与认为市民化水平与年龄正相关的研究有所差异(苏丽锋,2017)。可能的原因在于:一方面,较年轻农业转移人口的人力资本和社会资本等会随着年龄增加而快速积累,从而对其市民化过程产生积极影响(王桂新等,2010),但随着年龄渐长,这些资本的积累速度和影响效应都会减弱;另一方面,相对于年龄更大的群体,更年轻的农业转移人口融入城市社会的意愿更强,且在思想观念和行为方式上更容易实现从农村居民到城市居民的转变(单菁菁,2014;佟大建等,2022)。

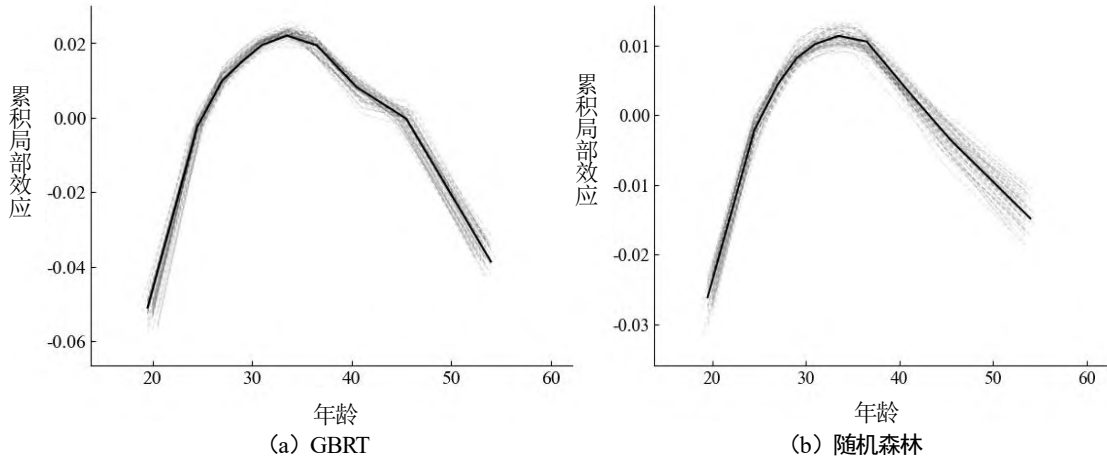


图4 农业转移人口年龄与市民化水平的 ALE 图

4.流动城市数量。图5展示了农业转移人口流动城市数量与市民化水平之间的ALE图。由图5可知,随着流动城市数量的增加,农业转移人口市民化水平呈上升趋势,但在流动城市数量超过2个时,这种上升趋势会明显放缓。可能的原因在于:一方面,农业转移人口在流动过程中会随着城市相关政策,以及自身或家庭需求的变化而重新选择流动城市,这在一定程度上反映他们融入城市的意愿和能力(王洁晶和张沐华,2023);另一方面,农业转移人口流动城市越多,则越不易于在城市中形成稳定的社会关系网络,从而不利于其实现市民化。

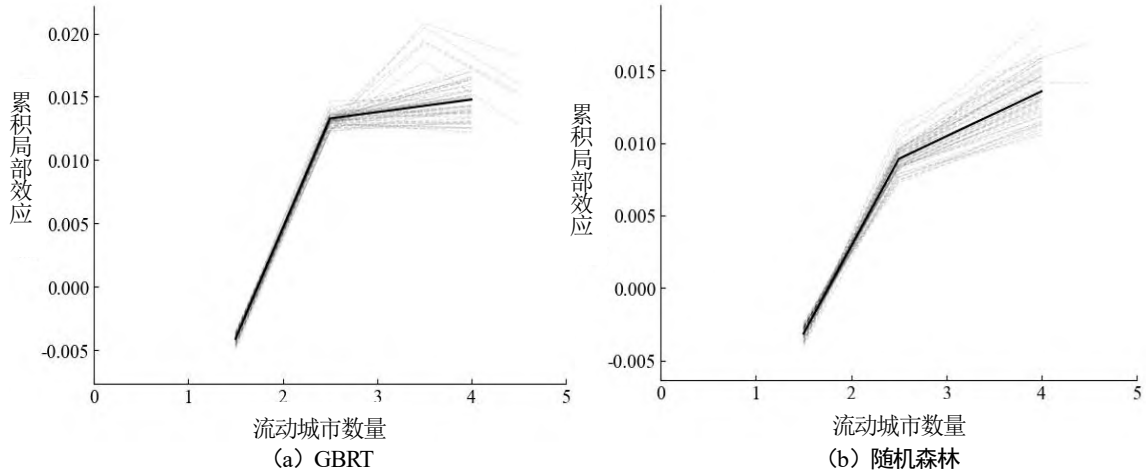


图5 农业转移人口流动城市数量与市民化水平的ALE图

5.本地居留时长。图6展示了农业转移人口本地居留时长与市民化水平之间的ALE图。由图6可知,随着在本地居留时长的增加,农业转移人口市民化水平呈上升趋势。可能的原因在于:一方面,农业转移人口在流入地的居留时间越长就越能形成稳定的生活方式,对流入地生活的适应性与认同感也相应提高(王桂新等,2010;朱纪广等,2020);另一方面,中国近年来相关政策的颁布,使得城市就业机会逐渐增多(何可等,2023)。长期的城市生活有助于农业转移人口的社会资本积累,增强了其就业稳定性以及在劳动力市场上的竞争力,进而提高市民化的意愿与能力(杨雪和魏洪英,2017)。

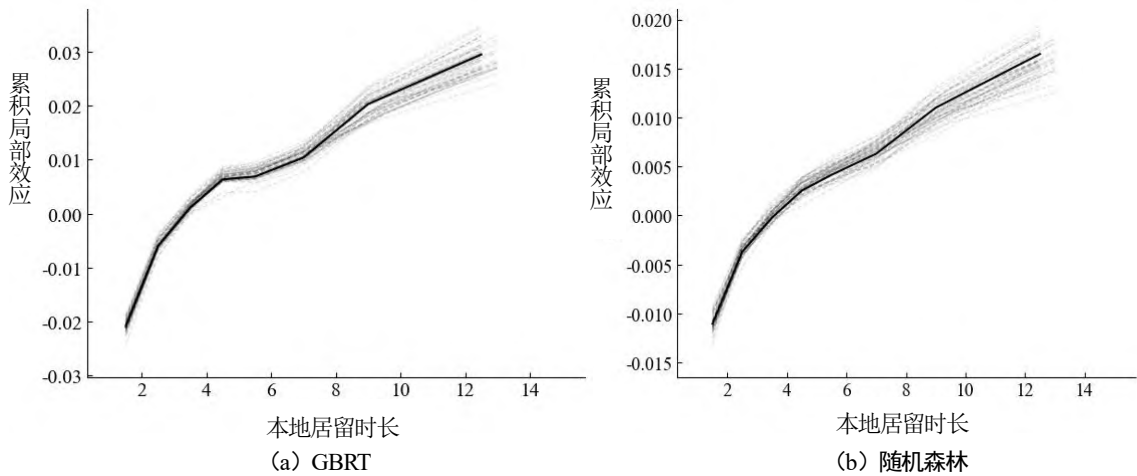


图6 农业转移人口本地居留时长与市民化水平的ALE图

（五）稳健性检验

1. 更换响应变量测度方式。前文在测量农业转移人口市民化水平时，为避免流入地差异带来的偏差，计算了农业转移人口与流入地城市居民各指标中位数的比值。本文将使用均值替代中位数来重新测量响应变量，并在此基础上应用各种模型进行拟合。相较于中位数，均值可以更充分地利用全部数据的信息。表 6 展示了基于新响应变量的各模型预测效果，结果显示集成学习方法的预测效果依然优于多元线性回归模型，这进一步证明了本文结论的稳健性。

表 6 基于新响应变量的模型预测效果评价

模型	(1) R^2_{IS}	(2) R^2_{OOS}	(3) EVS_{OOS}	(4) MSE_{OOS}	(5) MAE_{OOS}	(6) $MedAE_{OOS}$
多元线性回归	0.0990	0.0648	0.0654	0.0395	0.1557	0.1282
LASSO 回归	0.0088	0.0096	0.0111	0.0416	0.1606	0.1288
岭回归	0.0949	0.1018	0.1019	0.0382	0.1556	0.1349
随机森林	0.5252	0.1313	0.1313	0.0362	0.1541	0.1313
GBRT	0.1687	0.1202	0.1205	0.0383	0.1554	0.1296
XGBoost	0.1562	0.1250	0.1251	0.0372	0.1540	0.1299
前馈神经网络	0.0843	0.0764	0.0764	0.0348	0.1497	0.1282

2. 利用特征变量重要性方法对变量排序。鉴于 SHAP 值方法在理论上相较于其他方法具有优势，上文主要采用 SHAP 值方法对特征变量的重要性进行了测度和排序。接下来，本文利用以往文献中常用的特征变量重要性方法重新衡量特征变量的重要性。不同于 SHAP 值方法通过计算各个影响因素的平均边际贡献衡量其贡献程度的做法，特征变量重要性方法根据决策树分裂前后的信息增益衡量变量重要性。如表 7 所示，基于特征变量重要性方法计算的特征变量重要性排序在前五位的是受教育程度、年龄、性别、家庭规模和本地居留时长，与基于 SHAP 值方法计算的结果基本一致，进一步证明了本文结论的稳健性。

表 7 基于特征变量重要性方法的特征变量重要性排序（前五位）

排名	GBRT		随机森林	
	变量名称	相对重要性 (%)	变量名称	相对重要性 (%)
1	受教育程度	28.987	受教育程度	17.250
2	年龄	10.629	年龄	13.297
3	性别	9.884	本地居留时长	9.736
4	家庭规模	8.698	家庭规模	6.716
5	本地居留时长	7.030	性别	6.503

五、结论

本文基于 2017 年 CMDS 数据，运用多元线性回归、惩罚回归、集成学习和深度学习等多种方法，

系统探讨了个体、家庭、迁移以及城市四个不同维度的特征变量对农业转移人口市民化水平的影响。研究表明,集成学习方法在预测效果上明显优于多元线性回归模型,尤其是 GBRT 模型表现最佳。通过 SHAP 值方法的分析发现,个体受教育程度、性别、家庭规模、年龄和流动城市数量是影响农业转移人口市民化水平的重要因素。此外,本文还通过 ALE 图揭示了不同影响因素对农业转移人口市民化水平的具体预测模式,结果表明年龄和流动城市数量对农业转移人口市民化水平有着显著的非线性影响。

本文研究得到的政策启示如下:首先,应以系统性的政策思路持续推进农业转移人口市民化进程。本文研究结果表明,农业转移人口市民化水平受多种因素影响,且这些因素之间存在复杂的非线性关系。因此,一方面,政府应明确农业转移人口市民化的中长期目标和阶段性任务,在法规上保障农业转移人口在住房、就业、教育和医疗等方面的合法权益;另一方面,政府应通过制度与组织设计,使不同部门形成合力,出台各种配套措施,确保政府、企业、个人三方在推动农业转移人口市民化进程中的角色和作用得到有效协调。其次,应采取差异化的政策措施,加快农业转移人口市民化进程。本文研究结果表明,不同特征变量对农业转移人口市民化水平的预测效果各异,且某些变量的影响呈现明显非线性特征。因此,在制定与实施政策时,应充分考虑农业转移人口的个体、家庭和迁移特征,结合流入城市的特点,分阶段、分层次地推动农业转移人口市民化。例如,对于初到城市的农业转移人口,重点应放在提供必要的职业技能培训和就业指导上,帮助他们尽快融入城市劳动力市场。对于已经在城市稳定就业的农业转移人口,政府应更多关注其社会保障和住房等方面需求的满足。同时,政府还应着力提升与农业转移人口密切相关的公共服务水平,制定并实施城乡公共服务均等化政策,切实降低农业转移人口市民化的成本。最后,以提升人力资本为抓手,助力农业转移人口市民化进程。本文基于各种机器学习方法分析得出,受教育程度是影响农业转移人口市民化水平的最重要因素。因此,要进一步改善农村教育质量,推动城乡教育均衡发展,整合各类资源,提升农业转移人口人力资本,使其更有能力实现市民化。具体措施包括加强农村教育基础设施建设、优化对农村教师的激励机制、推动现代教育技术在农村教育中的应用,以及建立城乡教育资源共享机制,促进教育资源均衡分配等。

本文利用机器学习方法系统考察了农业转移人口市民化水平的影响因素,并通过 SHAP 值方法和 ALE 图等可解释性方法展现了各特征变量的预测能力和模式,从而得出了以往基于解释性建模的分析无法得到的结论。但与其他以预测能力为出发点的研究相同,本文的结论并不具备因果性。未来研究应进一步整合机器学习方法与因果推断框架,主要包括:其一,利用机器学习方法在预测方面的优势,结合双重差分等方法,构建更准确的反事实状态,以识别各影响因素与农业转移人口市民化水平之间的因果关系;其二,以预测性建模发掘变量间关系,以解释性建模构建和检验因果机制,充分发挥两者在目标和功能上的互补性,以深化和拓展对农业转移人口市民化水平影响因素的理解。

参考文献

1. 蔡昉, 2013: 《以农民工市民化推进城镇化》, 《经济研究》第 3 期, 第 6-8 页。

- 2.陈强, 2021: 《机器学习及 Python 应用》, 北京: 高等教育出版社, 第 343 页、第 455 页。
- 3.陈锡文, 2023: 《当前农业农村的若干重要问题》, 《中国农村经济》第 8 期, 第 2-17 页。
- 4.陈运森、周金泳、黄健峤, 2023: 《企业的慷慨因何而来? ——基于机器学习的证据》, 《财经研究》第 6 期, 第 153-169 页。
- 5.郭峰、陶旭辉, 2023: 《机器学习与社会科学中的因果关系: 一个文献综述》, 《经济学(季刊)》第 1 期, 第 1-17 页。
- 6.郭晓欣、钟世虎、李子健, 2023: 《农村流动人口城市化的影响机制——基于社会网络视角的发现》, 《中国人口科学》第 4 期, 第 51-66 页。
- 7.何军, 2011: 《代际差异视角下农民工城市融入的影响因素分析——基于分位数回归方法》, 《中国农村经济》第 6 期, 第 15-25 页。
- 8.何可、朱信凯、李凡略, 2023: 《聚“碳”成“能”: 碳交易政策如何缓解农村能源贫困? 》, 《管理世界》第 12 期, 第 122-144 页。
- 9.洪银兴、杨玉珍、王荣, 2021: 《城镇化新阶段: 农业转移人口和农民市民化》, 《经济理论与经济管理》第 1 期, 第 4-16 页。
- 10.黄寿峰、赵岩, 2023: 《政府信息化能否促进城乡基本公共服务均等化——来自信息惠民国家试点政策的证据》, 《吉林大学社会科学学报》第 6 期, 第 105-119 页。
- 11.李爱民、年猛、戴明锋, 2022: 《我国农业转移人口深度市民化研究》, 《中国软科学》第 8 期, 第 67-78 页。
- 12.李国正, 2020: 《农地权益保障与农业转移人口市民化》, 《中国土地科学》第 10 期, 第 117-124 页。
- 13.李强、龙文进, 2009: 《农民工留城与返乡意愿的影响因素分析》, 《中国农村经济》第 2 期, 第 46-54 页。
- 14.梁波、王海英, 2010: 《城市融入: 外来农民工的市民化——对已有研究的综述》, 《人口与发展》第 4 期, 第 73-85 页。
- 15.刘传江、程建林, 2009: 《双重“户籍墙”对农民工市民化的影响》, 《经济学家》第 10 期, 第 66-72 页。
- 16.刘传江、徐建玲, 2007: 《第二代农民工及其市民化研究》, 《中国人口·资源与环境》第 1 期, 第 6-10 页。
- 17.刘金凤、刘瑞明、石阳, 2023: 《从“半城市化”到“城市化”: 农业转移人口市民化进程中的教育推动机制研究》, 《数量经济技术经济研究》第 9 期, 第 138-156 页。
- 18.刘静、张锦华, 2021: 《城市异质影响下的农民工市民化程度——基于需求可识别双变量 Probit 和 HLM 模型的测度与分析》, 《浙江社会科学》第 10 期, 第 63-71 页。
- 19.刘景江、郑畅然、洪永淼, 2023: 《机器学习如何赋能管理学研究? ——国内外前沿综述和未来展望》, 《管理世界》第 9 期, 第 191-216 页。
- 20.刘小年, 2017: 《农民工市民化的影响因素: 文献述评、理论建构与政策建议》, 《农业经济问题》第 1 期, 第 66-74 页。
- 21.刘岩、谢天, 2019: 《跨国增长实证研究的模型不确定性问题: 机器学习的视角》, 《中国工业经济》第 12 期, 第 5-22 页。

- 22.陆瑶、施函青, 2022:《我国科技企业融资的决定因素研究——基于科创板企业的机器学习分析》,《金融研究》第9期,第132-151页。
- 23.陆瑶、张叶青、黎波、赵浩宇, 2020:《高管个人特征与公司业绩——基于机器学习的经验证据》,《管理科学学报》第2期,第120-140页。
- 24.宁光杰、李瑞, 2016:《城乡一体化进程中农民工流动范围与市民化差异》,《中国人口科学》第4期,第37-47页。
- 25.宁光杰、刘丽丽, 2018:《市民化意愿对农业转移人口消费行为的影响研究》,《中国人口科学》第6期,第55-68页。
- 26.秦立建、陈波, 2014:《医疗保险对农民工城市融入的影响分析》,《管理世界》第10期,第91-99页。
- 27.单菁菁, 2014:《农民工市民化研究综述:回顾、评析与展望》,《城市发展研究》第1期,第18-21页。
- 28.苏丽锋, 2017:《中国流动人口市民化水平测算及影响因素研究》,《中国人口科学》第2期,第12-24页。
- 29.孙战文、杨学成, 2013:《农民工家庭成员市民化的影响因素分析——基于山东省1334个城乡户调查数据的Logistic分析》,《中国农村观察》第1期,第59-68页。
- 30.佟大建、金玉婷、宋亮, 2022:《农民工市民化:测度、现状与提升路径——基本公共服务均等化视角》,《经济学家》第4期,第118-128页。
- 31.涂尔干, 2000:《社会分工论》,渠东译,北京:生活·读书·新知三联书店,第183-186页。
- 32.王成利、王洪娜, 2020:《城市长期居留流动人口的落户意愿及影响因素——基于差别化落户政策》,《中南财经政法大学学报》第5期,第64-72页。
- 33.王春超、蔡文鑫, 2021:《流动人口市民化与推进路径测算研究——基于同质化水平测度的视角》,《经济社会体制比较》第5期,第161-173页。
- 34.王芳、王宣艺、陈硕, 2020:《经济学研究中的机器学习:回顾与展望》,《数量经济技术经济研究》第4期,第146-164页。
- 35.王桂新、陈冠春、魏星, 2010:《城市农民工市民化意愿影响因素考察——以上海市为例》,《人口与发展》第2期,第2-11页。
- 36.王桂新、沈建法、刘建波, 2008:《中国城市农民工市民化研究——以上海为例》,《人口与发展》第1期,第3-23页。
- 37.王洁晶、张沐华, 2023:《中国流动人口市民化:理论、概念、影响因素与制度创新》,《西北师大学报(社会科学版)》第5期,第135-144页。
- 38.王青、刘烁, 2020:《进城农民工多维贫困测度及不平等程度分析——基于社会融合视角》,《数量经济技术经济研究》第1期,第83-101页。
- 39.王孝莹、王目文, 2020:《新生代农民工市民化的微观影响因素及其结构——基于人力资本因素的中介效应分析》,《人口与经济》第1期,第113-126页。
- 40.魏后凯、苏红键, 2013:《中国农业转移人口市民化进程研究》,《中国人口科学》第5期,第21-29页。
- 41.吴业苗, 2012:《城郊农民市民化的困境与应对:一个公共服务视角的研究》,《中国农村观察》第3期,第71-77页。

- 42.肖争艳、陈衍、陈小亮、陈彦斌, 2022:《通货膨胀影响因素识别——基于机器学习方法的再检验》,《统计研究》第6期,第132-147页。
- 43.辛宝英, 2016:《农业转移人口市民化程度测评指标体系研究》,《经济社会体制比较》第4期,第156-165页。
- 44.徐现祥、刘毓芸、肖泽凯, 2015:《方言与经济增长》,《经济学报》第2期,第1-32页。
- 45.徐延辉、龚紫钰, 2019:《社会质量与农民工的市民化》,《经济学家》第7期,第90-100页。
- 46.杨雪、魏洪英, 2017:《流动人口长期居留意愿的新特征及影响机制》,《人口研究》第5期,第63-73页。
- 47.张斐, 2011:《新生代农民工市民化现状及影响因素分析》,《人口研究》第6期,第100-109页。
- 48.张继良、马洪福, 2015:《江苏外来农民工市民化成本测算及分摊》,《中国农村观察》第2期,第44-56页。
- 49.张占斌、冯俏彬、黄锬, 2013:《我国农村转移人口市民化的财政支出测算与时空分布研究》,《中央财经大学学报》第10期,第1-7页。
- 50.中国人民大学国家发展与战略研究院劳动力市场研究中心课题组, 2019:《中国劳动力市场化指数编制》, http://nads.ruc.edu.cn/upfile/file/20190329142808_855000_87895.pdf。
- 51.周卫华、翟晓风、谭皓威, 2022:《基于XGBoost的上市公司财务舞弊预测模型研究》,《数量经济技术经济研究》第7期,第176-196页。
- 52.朱纪广、张佳琪、李小建、孟德友、杨慧敏, 2020:《中国农民工市民化意愿及影响因素》,《经济地理》第8期,第145-152页。
- 53.朱信凯, 2005:《农民市民化的国际经验及对我国农民工问题的启示》,《中国软科学》第1期,第28-34页。
- 54.祝仲坤, 2021:《公共卫生服务如何影响农民工留城意愿——基于中国流动人口动态监测调查的分析》,《中国农村经济》第10期,第125-144页。
- 55.祝仲坤、冷晨昕, 2020:《自雇行为如何影响农民工的市民化状态——来自中国流动人口动态监测调查的经验证据》,《南开经济研究》第5期,第109-129页。
- 56.邹一南, 2021:《农民工落户悖论与市民化政策转型》,《中国农村经济》第6期,第15-27页。
- 57.Aas, K., M. Jullum, and A. Løland, 2021, “Explaining Individual Predictions when Features are Dependent: More Accurate Approximations to Shapley Values”, *Artificial Intelligence*, Vol.298: 1-24.
- 58.Athey, S., 2019, “The Impact of Machine Learning on Economics”, in A. Agrawal, J. Gans, and A. Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda*, Chicago: University of Chicago Press, 507-547.
- 59.Breiman, L., 2001, “Random Forests”, *Machine Learning*, Vol.45: 5-32.
- 60.Chen, Y., L. Wang, and M. Zhang, 2018, “Informal Search, Bad Search?: The Effects of Job Search Method on Wages among Rural Migrants in Urban China”, *Journal of Population Economics*, 31(3): 837-876.
- 61.Friedman, J. H., 2001, “Greedy Function Approximation: A Gradient Boosting Machine”, *Annals of Statistics*, Vol.29: 1189-1232.
- 62.Hu, F., Z. Y. Xu, and Y. Y. Chen, 2011, “Circular Migration, or Permanent Stay? Evidence from China’s Rural-Urban Migration”, *China Economic Review*, 22(1): 64-74.

63.Lundberg, S. M., and S. Lee, 2017, "A Unified Approach to Interpreting Model Predictions", Conference Paper, 2017 Neural Information Processing Systems, Long Beach: State of California, https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

64.Mullainathan, S., and J. Spiess, 2017, "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, 31(2): 87-106.

65.Shmueli, G., 2010, "To Explain or To Predict?", *Statistical Science*, 25(3): 289-310.

66.Stark, O., 1991, *The Migration of Labour*, Oxford: Basil Blackwell Press, 216-235.

(作者单位: 郑州大学商学院)

(责任编辑: 何 可)

Factors Influencing the Citizenization of Rural-Urban Migrants: A Study Based on Machine Learning

QI Xiulin WANG Xinru

Abstract: Based on the China Migrants Dynamic Survey (CMDS) data in 2017, this paper systematically investigates the influencing factors of the citizenization of rural-urban migrants using the multiple linear regression, penalized regression, ensemble learning, deep learning, and other machine learning methods. The results show that the ensemble learning method outperforms the multiple linear regression method in predicting the level of citizenization of rural-urban migrants, among which the gradient boosting regression tree (GBRT) model performs the best prediction. Moreover, among all the characteristic variables, the individual's education level, gender, family size, age, and number of mobile cities are the most important influencing factors for the citizenization of rural-urban migrants. In addition, we use the accumulated local effects (ALE) plot to show the specific prediction patterns of different influencing factors, and find that factors such as age and number of mobile cities have non-linear characteristics in their influence on the citizenization of rural-urban migrants. The conclusions of this paper are informative for the design of policies aimed at accelerating the citizenization of rural-urban migrants.

Keywords: Rural-Urban Migrants; Citizenization; Machine Learning