西北农林科技大学

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

《中国农村经济》

齐秀琳　　汪心如
郑州大学商学院　　郑州大学商学院

分享人：王梓豪

2024-06-05

一、引言

引言

1. 背景

- 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**”。
- 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 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**。

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。

2. 以往的研究方法

- 相关性分析
- 因果推断
  - 倾向得分匹配法
  - 工具变晕法
  - 双重差分法

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**。

3. 本文方法

- 本文采用**多种机器学习方法**，通过**预测性建模**考察农业转移人口市民化水平的影响因素
- 相较解释性建模，本文运用预测性建模展开研究具有**三大优势**：

1. 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
2. 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
3. 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诉病的“黑箱”问题，还能够揭示解释性建模无法获取的关键信息
   - 沙普利加和解释 (SHapley Additive exPlanations, 简称 SHAP)
   - 偏依赖图 (partial dependence plot, 简称 PD)
   - 累积局部效应图 (accumulated local effects plot, 简称 ALE)

4. 机器学习 & 影响因素研究

[1] 刘岩,谢天.**跨国增长**实证研究的模型不确定性问题：机器学习的视角[J]. 中国工业经济, 2019, (12):5-22.
[2] 陆瑶,张叶青,黎波,等.**高管个人特征**与公司业绩——基于机器学习的经验证据[J]. 管理科学学报, 2020, 23(02):120-140.
[3] 肖争艳,陈衍,陈小亮,等.**通货膨胀影**响因素识别——基于机器学习方法的再检验[J]. 统计研究, 2022, 39(06):132-147.

5. 预测性建模 vs. 解释性建模

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) 。

6. 研究内容

- 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
- 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
- 采用 SHAP 值方法评估不同因素的**影响大小**，并通过 ALE **图分析**重要特征变量对农业转移人口市民化水平的**具体预测模式**

7. 可能的边际贡献

- **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
- 通过采用前沿的**集成学习**和**深度学习**方法有效规避**多元线性回归模型**设定上的局限性
- 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**

二、 模型与算法

机器学习模型

1. 多元线性回归模型 (Baseline)

$$citizenship_i = \alpha + \beta^\top X_i + \varepsilon_i \tag{1}$$

- 其中，
  - α: 截距项
  - $citizenship_i$: 个体i的市民化水平
  - X_i: 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
  - $\beta^\top$: 各影响因素的系数
  - ε_i: 误差项

2. 惩罚回归方法

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**

(a) LASSO 回归

公式 1 可以转化为优化问题：

$$\min_{\beta} \sum_{i=1}^{m} \left(citizenship_i - \beta^{\top} X_i - \alpha\right)^2 \quad (2)$$

在公式 2 中加入L_1正则项：

$$\min_{\beta} \sum_{i=1}^{m} \left(citizenship_i - \beta^{\top} X_i - \alpha\right)^2 + \lambda \|\beta\|_1 \quad (3)$$

其中，m: 样本总量
λ: 正则化参数
$\|\beta\|_1$: β的绝对值之和，即$\|\beta\|_1 = \sum|\beta|$

(b) 岭回归

在公式 3 中，将L_1换成L_2：

$$\min_{\beta} \sum_{i=1}^{m} \left(citizenship_i - \beta^{\top} X_i - \alpha\right)^2 + \lambda \|\beta\|_2^2$$

其中，λ: 正则化参数
$\|\beta\|_2 = \sqrt[2]{\sum \beta^2}$

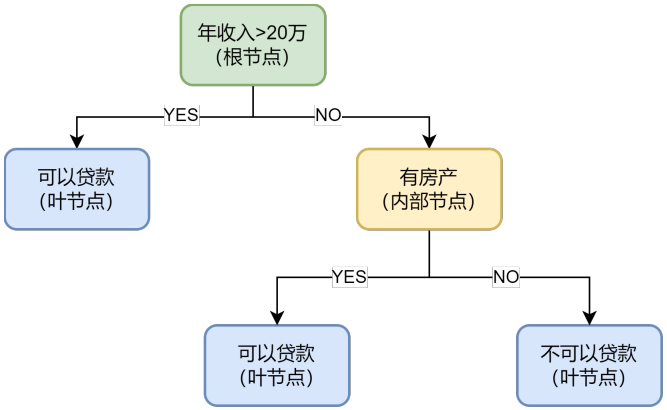
定义: L_p范数即为$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$

3. 集成学习方法

通过**装袋法(bagging)集成**多个预测效果一般的弱学习器，形成预测效果更优的强学习器

(a) 随机森林

- 随机森林 = 决策树 + bootstrap



例图 1　决策树示例

递归返回，情形(1).

递归返回，情形(2).

我们将在下一节讨论如何获得最优划分属性.

递归返回，情形(3).

从 A 中去掉 a_*.

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
　　属性集 $A = \{a_1, a_2, \ldots, a_d\}$.
过程： 函数 TreeGenerate(D, A)
1: 生成结点 node;
2: **if** D 中样本全属于同一类别 C **then**
3:　将 node 标记为 C 类叶结点; **return**
4: **end if**
5: **if** $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
6:　将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; **return**
7: **end if**
8: 从 A 中选择最优划分属性 a_*;
9: **for** a_* 的每一个值 a_*^v **do**
10:　为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
11:　**if** D_v 为空 **then**
12:　　将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; **return**
13:　**else**
14:　　以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
15:　**end if**
16: **end for**
输出： 以 node 为根结点的一棵决策树

算法 1　决策树算法[1]

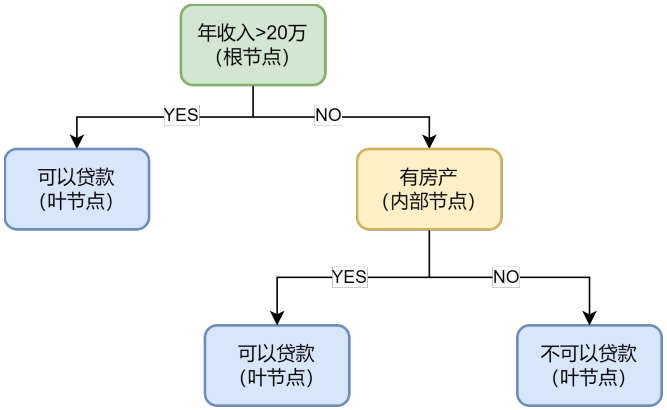
[1]周志华，机器学习，清华大学出版社，2016

3. 集成学习方法

通过**装袋法(bagging)集成**多个预测效果一般的弱学习器，形成预测效果更优的强学习器

(a) 随机森林

- 随机森林 = 决策树 + bootstrap



例图 1　决策树示例

- 信息增益[1]
  - 假设样本集合D中第k类样本所占比例为p_k，则D的信息熵为 $\mathrm{Ent}(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$.
  - 假定特征a有V个可能的取值$\{a^1, a^2, ..., a^V\}$,若使用a来对样本集D进行划分，则会产生V个分支结点，其中第v个分支结点包含了D中所有在特征a上取值为a^v的样本，记为D^v.我们可计算出D^v的信息熵，再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重$\frac{|D^v|}{|D|}$,即样本数越多的分支结点的影响越大，于是可计算出用特征 a 对样本集 D 进行划分所获得的“信息增益”,

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

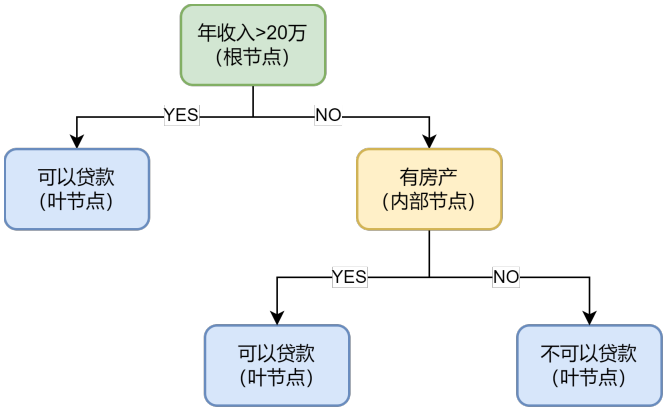
  - 选择信息增益最大的：$a_* = \arg\max_{a \in A} \mathrm{Gain}(D, a)$

3. 集成学习方法

通过**装袋法(bagging)集成**多个预测效果一般的弱学习器，形成预测效果更优的强学习器

(a) 随机森林

- 随机森林 = 决策树 + bootstrap



例图 1　决策树示例

- 随机森林算法

Step 1　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：

$$\left\{\boldsymbol{x}_1^{*b}, y_1^{*b}, \boldsymbol{x}_2^{*b}, y_2^{*b}, ..., \boldsymbol{x}_n^{*b}, y_n^{*b}\right\}, \quad b \in \{1, ..., B\} \tag{4}$$

Step 2　利用自助样本估计B棵不同的决策树，估计过程中不进行剪枝。记第b棵树的预测结果为：

$$\overline{f}^{*b}(\boldsymbol{x}), \quad b \in \{1, ..., B\} \tag{5}$$

Step 3　将B棵决策树的预测结果取平均处理后，得最终预测结果：

$$\overline{f}_{bag}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\overline{f}^{*b}(\boldsymbol{x}) \tag{6}$$

3. 集成学习方法

通过**提升法**(Boosting)**集成**多个预测效果一般的弱学习器，形成预测效果更优的强学习器

(b) GBRT

- **梯度提升**回归树
- Gradient Boosting Regression Tree
- 简称 GBRT

(c) XGBoost

- 极端**梯度提升**
- eXtreme Gradient Boosting
- 简称 XGBoost

- GBRT 算法

设定初始回归(预测)函数，其中$l(\cdot)$为损失函数：

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

For $d = 1 : D$ do

计算损失函数负梯度$\Psi_{i,d} = -\frac{\partial L(y_i, \hat{y}_i \leftarrow f_{d-1}(\boldsymbol{x}))}{\partial \boldsymbol{x}}$

以$\Psi_{i,d}$为残差近似值拟合出新回归树$g_d(\boldsymbol{x}) = \mathbb{E}(\Psi|\boldsymbol{x})$

选择使误差最小的梯度下降幅度

$$\tau = \arg\min_{\tau} \sum_{i=1}^{N} l(y_i, f_{d-1}(\boldsymbol{x}) + \tau g_d(\boldsymbol{x}))$$

计算新的回归(预测)函数$\hat{y}_i \leftarrow f_d(\boldsymbol{x}) = f_{d-1}(\boldsymbol{x}) + v\tau g_d(\boldsymbol{x})$

End For

得到最终的回归(预测)函数$\hat{y}_i \leftarrow f_D(\boldsymbol{x})$

3. 集成学习方法

通过**提升法**(Boosting)**集成**多个预测效果一般的弱学习器，形成预测效果更优的强学习器

(b) GBRT

- **梯度提升**回归树
- Gradient Boosting Regression Tree
- 简称 GBRT

(c) XGBoost

- 极端**梯度提升**
- eXtreme Gradient Boosting
- 简称 XGBoost

- XGBoost 算法

目标方程：$\min \sum_{i=1}^{N} l(y_i, \hat{y}_i) + L_.(\hat{y}_i)$

其中，

$\hat{y}_i$包含$l(\cdot)$的一阶和二阶偏导

$L_.$包含L_1和L_2惩罚项

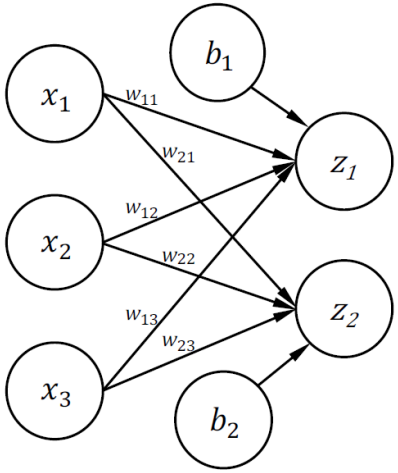
4. 深度学习方法

本文采用的深度学习方法为前馈**神经网络**模型，某种意义上这是**堆叠法**(stacking)

前馈：

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$$

$$\begin{cases} z_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1 \\ z_1 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2 \end{cases}$$



例图 2　一个有三个输入和两个输出的神经元的神经网络

反向传播、梯度下降(GD)：

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \alpha\frac{\partial l}{\partial \boldsymbol{W}}$$

$$\boldsymbol{b} \leftarrow \boldsymbol{b} - \alpha\frac{\partial l}{\partial \boldsymbol{b}}$$

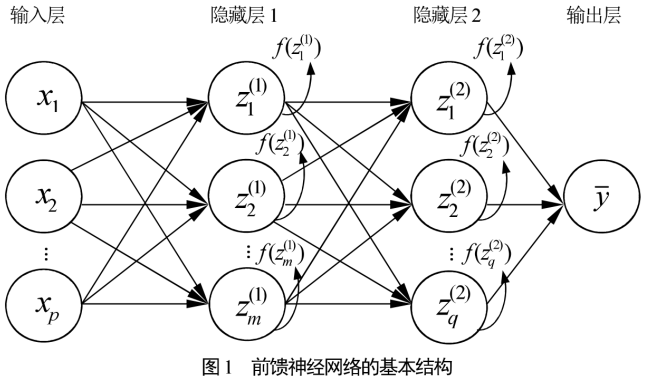


图 1　前馈神经网络的基本结构

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

图 1　一个具有两个隐藏层和一个输出层的多层感知器

模型的参数调整与选择

1. 模型的参数调整

- 训练集:测试集 = 8:2
- K折交叉验证
- 选择各种模型的超参数

2. 评价指标

$$R_{\mathrm{IS}}^2\left(R_{\mathrm{OOS}}^2\right)=1-\frac{\sum_{i=1}^n\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^n\left(y_i-\overline{y}_i\right)^2} \tag{8}$$

$$\mathrm{EVS}_{\mathrm{OOS}}^2=1-\frac{\sum_{i=1}^n\left(\hat{y}_i-\overline{y}_i\right)^2}{\sum_{i=1}^n\left(y_i-\overline{y}_i\right)^2} \tag{9}$$

$$\mathrm{MSE}_{\mathrm{OOS}}=\frac{1}{n}\sum_{i=1}^n\left(\hat{y}_i-\overline{y}_i\right)^2 \tag{10}$$

$$\mathrm{MAE}_{\mathrm{OOS}}=\frac{1}{n}\sum_{i=1}^n\left|\hat{y}_i-\overline{y}_i\right| \tag{11}$$

$$\mathrm{MedAE}_{\mathrm{OOS}}=\operatorname{median}\left|\hat{y}_i-\overline{y}_i\right| \tag{12}$$

- R_{IS}^2 越大，训练的越好
- R_{OOS}^2、$\mathrm{EVS}_{\mathrm{OOS}}^2$ 越大，泛化能力越好
- $\mathrm{MSE}_{\mathrm{OOS}}$、$\mathrm{MAE}_{\mathrm{OOS}}$、$\mathrm{MedAE}_{\mathrm{OOS}}$ 越小，预测准确率越高

机器学习的可解释性方法

1. 沙普利加和解释 (SHAP)

- 两大优点：
  - 影响因素的**重要性排序**
  - 在进行贡献度评价时，该方法满足了**一致性**条件
    - 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。
- 第j个影响因素的 SHAP 值计算公式如下

$$\mathrm{SHAP}_*^j = \sum_{S \subseteq F \setminus \{j\}} = \frac{|S|!(|F| - |S| - 1)!}{|F|!}(v_*(S \cup \{j\}) - v_*(S)) \tag{13}$$

其中，

- S: 不包含第j个影响因素的集合
- F: 所有影响因素的全集
- $|X|$: 集合X的元素个数
- $\frac{|S|!(|F| - |S| - 1)!}{|F|!} \approx \frac{1}{C(|F|,|S|)}$: 集合$F$中随机选出集合$S$的概率
- $v_*(S)$: 利用S通过算法得到预测值的期望，
- $(v_*(S \cup \{j\}) - v_*(S))$: 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

2. 累积局部效应图 (ALE)

- 以往文献常用**偏依赖图**刻画预测模式，但需要各影响因素之间**相互独立**
- ALE 图通过**计算局部效应**消除了变量相关性的影响

Step 1 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
Step 2 算每个区间内的**局部效应**:

- 特征变量x_j未经中心化处理的累积局部效应值:

$$ALE^*\left(x^j\right)=\sum_{k=1}^{k\left(x^j\right)}\frac{1}{\left|N_k^j\right|}\sum_{i:x_i^j\in N_k^j}\left[f\left(x_i^j=z_k^j,\boldsymbol{x}_i^{\setminus j}\right)-f\left(x_i^j=z_{k-1}^j,\boldsymbol{x}_i^{\setminus j}\right)\right] \tag{14}$$

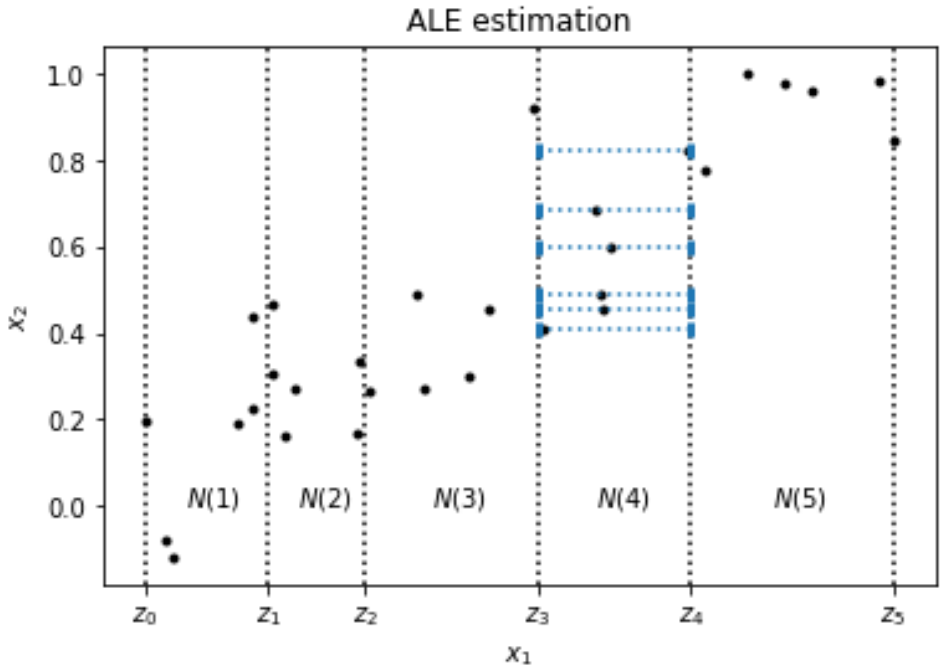
其中，
- $k\left(x^j\right)$: x^j所在区间的索引
- N_k^j: 特征j的第k个区间的样本集合
- z_k^j: 特征j的第k个区间的**右**边界值
- $x_i^{\setminus j}$: N_k^j中第i个样本，剔除特征j，剩余的特征

- 中心化处理后的累积局部效应值:

$$ALE\left(x^j\right)=ALE^*\left(x^j\right)-\frac{1}{n}\sum_{i=1}^{n}ALE^*\left(x_i^j\right) \tag{15}$$

$$ALE^*(x^j) = \sum_{k=1}^{k(x^j)} \frac{1}{|N_k^j|} \sum_{i:x_i^j \in N_k^j} \left[f\left(x_i^j = z_k^j, \boldsymbol{x}_i^{\setminus j}\right) - f\left(x_i^j = z_{k-1}^j, \boldsymbol{x}_i^{\setminus j}\right) \right]$$

$$ALE\left(x^j\right) = ALE^*\left(x^j\right) - \frac{1}{n} \sum_{i=1}^{n} ALE^*\left(x_i^j\right)$$



例图 3　与特征x_2相关的特征x_1的 ALE 计算示意图[1]

[1]https://docs.seldon.io/projects/alibi/en/latest/methods/ALE.html

三、 变量说明和数据来源

变量说明

1. 响应变量：农业转移人口市民化水平

- 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。

- 基于**四段论**(王春超和蔡文鑫, 2021)对农业转移人口市民化的内涵进行界定：
  a. **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
  b. **公共服务市民化**，即农业转移人口应享与本地城市居民同等水平的公共医疗和教育资源
  c. **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
  d. **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

- 为了避免由流入地差异导致的衡量偏差，本文采用**比值法**测度进入指标体系的各变量。
  - 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**

- 考虑到嫡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，本文使用**等权重法**赋权

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。

变量说明

2. 特征变量

①个体特征

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

变量说明

2. 特征变量

②家庭特征

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
特征变量	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
	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

变量说明

2. 特征变量

③迁移特征

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
特征变量	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

变量说明

2. 特征变量

③迁移特征

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
特征变量	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
	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

- 城乡公共服务差距*：
  - 城乡收入差距
  - 城乡教育水平差距
  - 城乡医疗服务水平差距
  - 城乡社会保障水平差距

* 均等化程度：农城之比

数据来源

- 农业转移人口**市民化水平**
  - 2017 年中国流动人口动态监测调查数据（CMDS(2017), C 卷, D 卷）
- 农业转移人口的**个体**、**家庭**和**迁移**特征变量
  - CMDS(2017)
- **城市**特征变量：

  最低工资和户籍开放度
  - 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
  - 《汉语方言大词典》

  其他城市特征变量
  - 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

多元线性回归结果

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	
系数	−0.0260***	0.0091*	0.0106	0.0568	0.0037	0.0036	
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		
系数	0.0139*	0.0033	0.0144**	0.0043***	0.0031		
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		
系数	0.6070	0.0047	−5.0760	8.3150	2.9403		
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。

多元线性回归结果

虽然印证的之前研究的一些发现，但是，

1. **模型是线性的**
   - 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。
     - 虽然可以加入高阶项，但是加入后的模型与实际数据匹配吗？
2. 虽然采用了**稳健标准误**来处理潜在的**异方差问题**
   - 但是x太多了$\Rightarrow$**多重共线性问题** (VIF 最大值为 21.73)
     - 可以筛选变量，但是影响因素分析的全面性和系统性？

多元线性回归结果

虽然印证的之前研究的一些发现，但是，

1. **模型是线性的**
   - 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。
     - 虽然可以加入高阶项，但是加入后的模型与实际数据匹配吗？
2. 虽然采用了**稳健标准误**来处理潜在的**异方差问题**
   - 但是x太多了$\Rightarrow$**多重共线性问题** (VIF 最大值为 21.73)
     - 可以筛选变量，但是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。

基于不同方法的模型预测效果评价

- LASSO > 多元线性回归
- 岭回归 ≈ 多元线性回归
- 前馈神经网络拟合能力 < 多元线性回归，但泛化能力反之
- 集成学习（随机森林、GBRT、XGBoost）最优秀
  - GBRT 最好，下文主要以 GBRT 为例分析

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

不同特征变量的重要性排序

- GBRT 和随机森林前 5 个因素排名基本相同，说明**结果稳健**
- **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)
- 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
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

主要特征变量对农业转移人口市民化水平的预测模式

1. 受教育程度

- 由图 2 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~~
  [对农业转移人口的市民化水平的影响越大]
- 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**？
  - 虽然解释性模型也能拟合出类似的图形，但是解释性模型是基于**模型假设设定**的
  - 而 ALE 展示的是预测性模型“**真实**”的线性关系。即使结论一致，但也不是简单重复

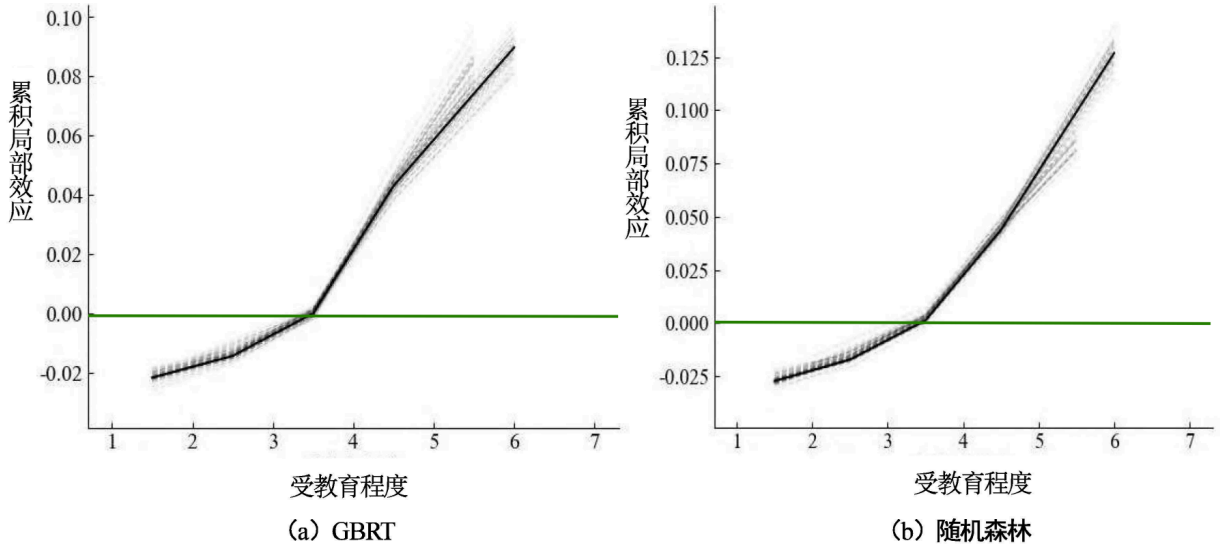


图 2　农业转移人口受教育程度与市民化水平的 ALE 图

主要特征变量对农业转移人口市民化水平的预测模式

2. 家庭规模

- 由图 3 可知，当家庭规模扩大时，[对]农业转移人口市民化水平[的影响]会先迅速下降，后下降速度逐渐减缓负向增加。

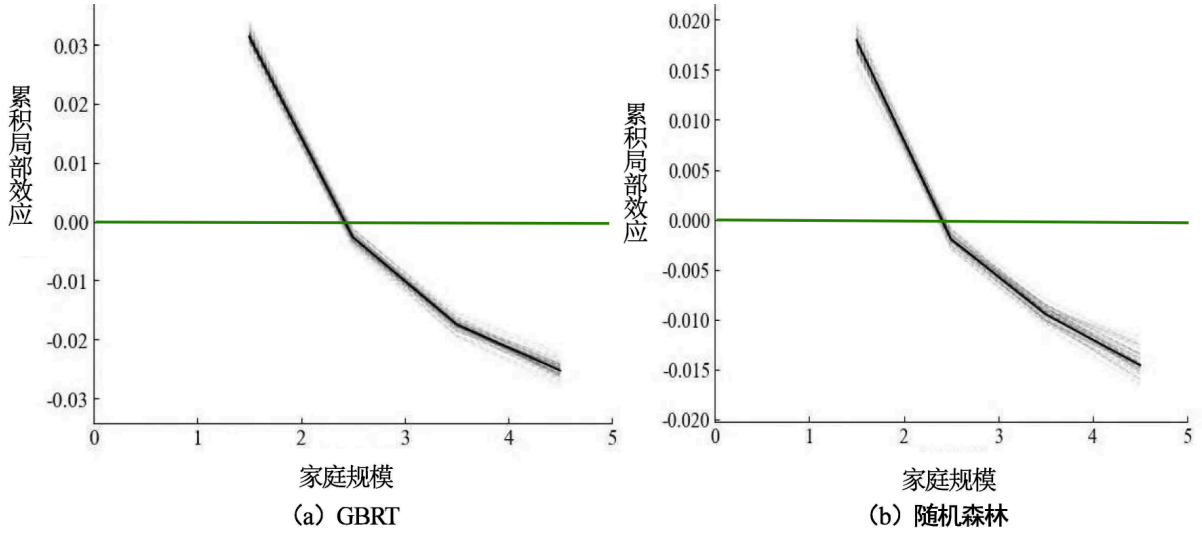


图 3　农业转移人口家庭规模与市民化水平的 ALE 图

主要特征变量对农业转移人口市民化水平的预测模式

3. 年龄

- **倒 U 型**关系
- 当个体年龄在 33 岁以下时，农业转移人口市民化水平[的变化]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[的变化]则随着年龄增加而下降。
- 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)

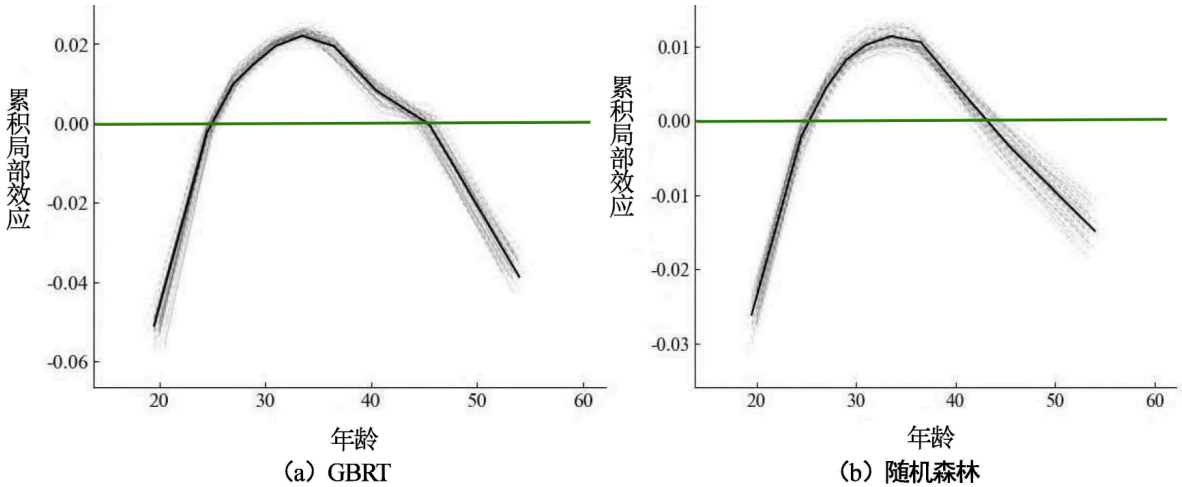


图 4　农业转移人口年龄与市民化水平的 ALE 图

主要特征变量对农业转移人口市民化水平的预测模式

4. 流动城市数量

- 随着流动城市数量的增加，[对]农业转移人口市民化水平[的影响]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。

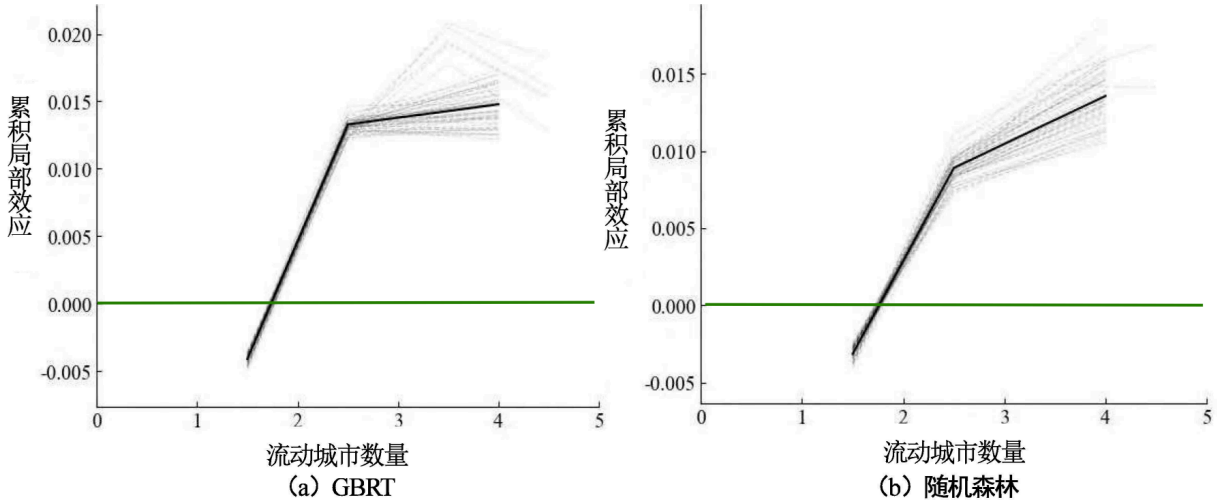


图 5　农业转移人口流动城市数量与市民化水平的 ALE 图

主要特征变量对农业转移人口市民化水平的预测模式

5. 本地居留时长

- 随着在本地居留时长的增加，[对]农业转移人口市民化水平[的影响]呈上升趋势。

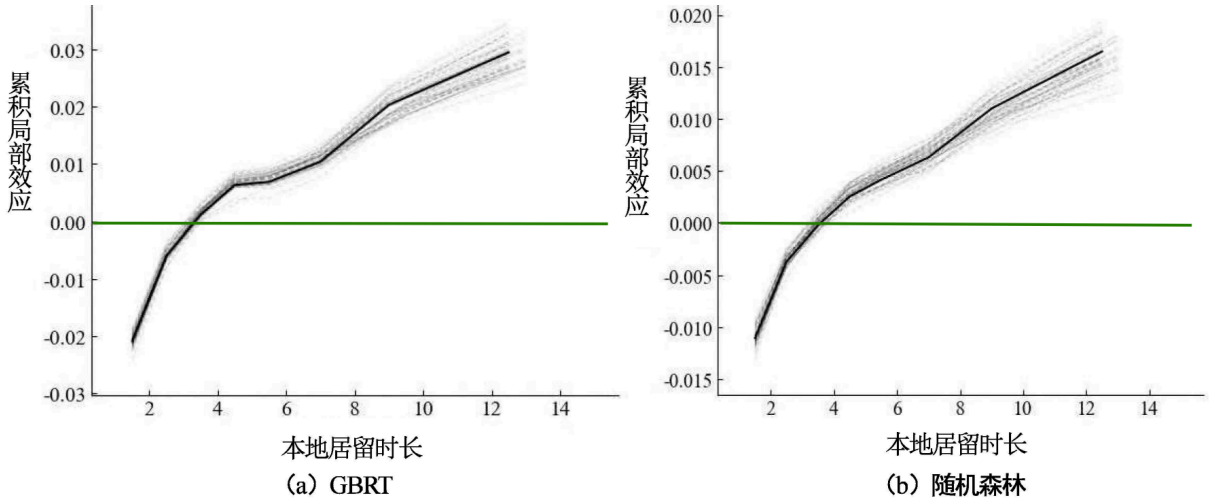


图 6　农业转移人口本地居留时长与市民化水平的 ALE 图

稳健性检验

1. 更换响应变量测度方法

- 用**均值**替代**中位数**来重新测量响应变量
  - 依然是集成学习的效果比多元回归更好，证明稳健

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

稳健性检验

2. 利用特征变量重要性方法对变量排序

- 决策树分裂前后的**信息增益**衡量变晕重要性
  - 前五位除了流动城市数量换成了本地居留时长(表 5 第六位)，其余一样，证明稳健

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

结论

1. 结果

- 略

2. 政策启示

- 应以系统性的政策思路待续推进农业转移人口市民化进程
- 应采取差异化的政策措施，加快农业转移人口市民化进程
- 以提升人力资本为抓手， 助力农业转移人口市民化进程(受教育程度是最重要的影响因素)

3. 局限与展望

- 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**
- 结合双重差分等方法，构建更准确的反事实状态，识别因果
- 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

*本汇报的幻灯片由 typst 编写，源文件可访问该连接获得.

Thanks!