影响居民生活水平的城市化因素分析

第 3 组 殷明 闫函 夏玉 姜峰

一、 简介

拿破仑曾说:“中国是一只沉睡的雄狮，她一旦醒来，整个世界会为之颤抖”。

改革开放 40 年以来，中国经济取得了举世瞩目的成就，中国也成为了拥有全球最多超级城市的国家，但我们发现虽然总体上中国人民的生活得到了极大的改善，但城市之间居民生活水平的差异却日渐明显，因此本文致力于探讨不同的城市化因素，例如经济发展水平，多元化水平等等，对居民生活水平的影响状况，并进行深入的分析。

二、 非技术性阐述

随着中国经济的不断发展，城市居民生活水平也在不断地改善，但是我们观察到，由于城市化进程的发展方式和力度不同，不同城市之间居民生活水平存在着明显的差异。为了给地方政府制定政策，进行城市规划提供方向性建议，我们可以通过在历史数据中寻找成功的经验，因此我们要回答的问题是，哪些城市化因素最能影响居民的生活水平变化。

为了回答此问题，我们需要知道：有哪些因素能够反映城市化水平，有哪些因素能够反映居民生活水平，因素对居民生活水平影响的稳定性如何，哪些城市化因素最能影响居民的生活水平。通过研究和回答这些问题，我们得到了如下结论：

1. 影响居民生活水平最大的 4 个显著性因素是经济发展，社区服务，社区教育和健康质量。
2. 在现今社会，教育是影响一个人未来收入和生活质量的决定性因素，健康则是承受高强度工作的保证，而良好的经济发展则给予了一个受过良好教育，并且拥有健康体格的人得到高收入工作的机会。
3. 社区服务得分如此之高，表明该因素与居民的生活水平的相关性很高，但从常识来看，该因素更有可能是居民生活水平提高的结果而并非其解释因素。

三、 技术性阐述

3.1. 研究方法概述

首先，我们通过数据清洗，收集和构造一些城市化发展的重要因子和能够反映居民生活水平的基础特征，再利用 PCA 方法构建居民生活水平打分系统；接着，我们利用

一些可视化的方法，来展示不同地区不同年份居民生活水平的变化情况；

然后，对各个城市化发展因子进行因子分析，设置自定义指标，筛选出在时序分布上具有稳定影响的因子；最后，利用机器学习的方法，求得出各个因子的重要性排序，即得到哪些城市化因素最能影响居民的生活水平。

3.2. 数据清洗过程

总体来说，数据清洗包括缺失值处理，异常值处理和数据的归一化，在本次比赛中，我们用到的具体技术介绍如下。

3.2.1. 缺失值处理

1. Lasso 回归构造

对于相同 id 不同年份中，只存在某些年份的部分指标缺失的情况，我们采用 Lasso 回归的方式构造缺失数据。

Lasso 是在 OLS 的基础上，目标函数增加了一项关于回归系数绝对值的惩罚项，

$$\text{即：}\ \operatorname{argmin}\ (y-\omega x)^2+\alpha|\omega|$$

这样就很好地避免过度拟合带来的问题，尤其在训练样本相对于因子维度并不是很大情况下相对其他方法有卓越的表现。从目标函数的惩罚项形式可以看出我们需要对数据进行标准化以避免由于不同维度的x的量纲不同带来的系数差异过大。

以黑龙江省的城市化水平在不同方面的基本信息中缺失值处理为例，将所有属于黑龙江省的所有被调查社区的基本信息中同一方面按照不同年份取平均。然后将所有属于同一方面的不同年份数据进行减去均值再除以标准差的归一化处理，得到的数据如下图形式：

province	survey_wa	urbanizati	populatio	diversity_s	economic_	quality_of	housing_c	market_cc	social_ser	transporta	communit	modern_r	sanitation_
Heilongjia	1997	-0.83869	-1.86916	-1.28608	-0.64885	0.400919	-0.33404	-0.56286		0.320503	0.047627	1.074062	-1.37608
Heilongjia	2000	-0.2953	-1.7638	-0.95846	-0.54092	-0.61032	-0.10251	0.58811	-1.12042	0.825566	0.410486	1.025037	-1.11824
Heilongjia	2004	-0.04218	-1.69356	-0.57842	0.141265	-0.19234	0.069291	0.46828	-0.95076	0.548984	0.437902	-1.44432	-0.73965
Heilongjia	2006	-0.45349	-1.65844	-0.31633	-0.09364	-3.06425	0.202166	-0.35156	-1.07164	-0.09637	0.590303	-2.38282	-0.2664
Heilongjia	2009	-0.1895	-1.55307	0.037505	0.448033	-0.49908	0.406215	-0.14485	-0.98884	0.705313	0.426613	-2.8809	-0.78534
Heilongjia	2011	-0.09335	-1.58819	0.116134	0.254963	-1.77324	0.673445	-0.44598	-0.34135	-0.93013	0.671744	-2.07978	-0.12606
Heilongjia	2015	0.365323	-2.31963	0.810689	0.295811	-0.59346	0.769955	-0.16016	0.319392	1.102147	1.48213	-1.2873	-0.46549

可以看到，只有 1997 年的 social_services_score 存在缺失，因此将所有其他年份的非 social_services_score 的指标和 social_services_score 制成 X_train、Y_train 作为训练样本，得到优化好的模型后将 1997 年的非数据放入训练，得到预测的值作为构造的样本。最后再将所有数据进行之前归一化处理的逆操作，即乘以对应的标准差再加上均值之后就得到想要的数据。

2. 三次样条插值构造

对于相同 id 的不同年份中，存在某一年的数据完全缺失，而该年份的前后几年的数据完全具备的情况，我们采用三次样条插值进行缺失数据的构造。

三次样条插值就是将原始长序列分割成若干段构造多个三次函数（每段一个），使得分段的衔接处具有二阶导数连续的性质（也就是光滑连接）。具备计算简单，局部控制性好的特点。

以辽宁省的城市化水平在不同方面的基本信息中缺失值处理为例，将所有属于黑龙江省的所有被调查社区的基本信息中同一方面按照不同年份取平均。然后将所有属于同一方面的不同年份数据进行减去均值再除以标准差的归一化处理，得到的数据如下图形

式：

province	survey_w	urbanizati	populatio	diversity_	economic_	quality_of	housing_c	market_cc	transporta	communit	modern_r	sanitation
Liaoning	1989	-0.87388	-0.04286	-1.63991	-1.22037	0.097548	-0.82582	-0.86449	-0.09637	-0.08139	0.08911	-1.39566
Liaoning	1991	-0.63848	-0.04286	-1.63991	-1.15356	-0.05077	-0.35199	-0.57025	0.733372	0.097621	0.310093	-0.87999
Liaoning	1993	-0.45073	-0.08409	-1.20176	-1.11369	0.012252	-0.29605	-0.44969	0.512908	0.111889	1.143514	-0.40348
Liaoning	1997											
Liaoning	2000	0.303517	0.433389	-0.01754	-0.54289	0.44784	0.634984	0.241768	0.67068	0.75796	1.613891	0.141701
Liaoning	2004	0.532789	0.736837	0.108271	0.483387	-0.6265	0.729589	-0.78376	1.290223	0.581465	-0.05453	0.502021
Liaoning	2006	0.953327	1.006568	0.49827	0.906726	0.44784	0.765762	-1.02527	2.175285	0.661197	-0.50277	0.865474
Liaoning	2009	0.994268	1.478599	0.951173	0.615619	2.043168	1.033086	-1.87889	1.732754	0.798987	0.067012	0.912472
Liaoning	2011	0.855879	0.972852	1.190205	0.982711	1.08856	1.021106	-1.91334	-0.12588	0.845433	-0.7828	0.599151
Liaoning	2015	1.380028	0.467105	1.404075	0.558384	2.382943	0.980373	1.501146	1.201717	1.589343	0.407325	0.586618

可以看到，只有 1997 年的数据存在缺失，因此将所有其他年份的每一项指标的数据进行三次样条插值，再用得到 1993~2000 这一段中的表达式计算 1997 年的对应数据作为构造的新样本。最后再将所有数据进行之前归一化处理的逆操作，即乘以对应的标准差再加上均值之后就得到想要的数据。

3. 直接删除

对于缺失数据量占比较少的数据就直接进行删除操作。在数据预处理过程中执行该操作的有 avg_monthly_wage, total_value_of_bonuses 等字段。

3.2.2. 数据标准化

本次比赛采用的标准化方法为对相同 id，不同调查年份的所有数据，采用减去它们平均值之后再处以标准差的方法进行归一化。

3.3. 数据结构

3.3.1 如何构造因子 X——城市化发展因素

这里我们选取的因子为城市化水平在不同方面的基本信息这一表格中的城市化水平、人口密度水平、多元化的水平、经济发展的水平、健康质量、住房发展水平、市场发展水平、社会服务、交通发展水平、社区教育、现代市场发展水平、卫生条件这几个因素。它们都和城市化水平呈正相关的关系。

具体在实施时采用先将所有数据按照不同省份不同年份做平均，之后舍去年份过少的省份数据，留下的省份包括湖北省、贵州省、山东省、江苏省、广西省、黑龙江省、辽宁省、河南省。接着根据具体情况采用不同的方式进行缺失值的构造，最后将所有数据按照不同的因子做成和年份、省份相关的表格以便后续的数据分析。

3.3.2 如何构造标签 y——居民生活水平

1. 收集基本指标

由于居民生活水平表现在许多方面，所以没有单一的指标能够综合性地反映居民的生活水平，因此我们收集一些基础指标，其中包括平均月工资，最高学历，总奖金，家庭净收入，家庭总收入，每月保费等，它们反映了居民生活水平的不同方面。

2. PCA 打分系统

当某一个事物存在多个维度的评价指标时，对各个指标设置不同的权重然后进行综合加权是个不错的处理方法，但是指标权重设置的合理性又成了新的衍生问题。由居民生活水平的基本指标可以看出，各个指标之间存在较大的相关性，所提供的信息存在不同程度的信息重叠。

因此，可以考虑采用主成分分析(Principle Component Analysis, PCA)的方法来构建一个居民生活水平的打分系统。具体思路是，我们首先利用 PCA 方法对原始指标进行转换，剔除重叠的信息；其次，主成分的特征值代表了其所解释的变差，即该成分所包含信息的大小，因此我们可以利用主成分的特征值来确定指标的权重，从而赋予信息贡献量大的指标以较大的权重，减轻信息含量较小的指标的权重。

关于主成分个数的确定，我们以累积方差贡献率是否超过 80%来作为主成分提取个数的标准。其中 PCA 确定指标权重的具体构造步骤如下所示：

(1) 对于原始指标 Y(Y 为 n*m，其中 n 代表样本个数，m 为上述基础指标的维数)，计算出 Y 的协方差矩阵，提取主成分对应的特征值λ和对应的特征向量矩阵 F

(2) 将原始指标进行降维：

$$F' = YF$$

F'为降维后的因子

(3) 将特征值λ进行归一化，得到λ'

(4) 由于特征值的大小就表示该主成分的重要性，因此最终进行加权即可得到综合的居民生活水平评价指标

$$y = \lambda' F'$$

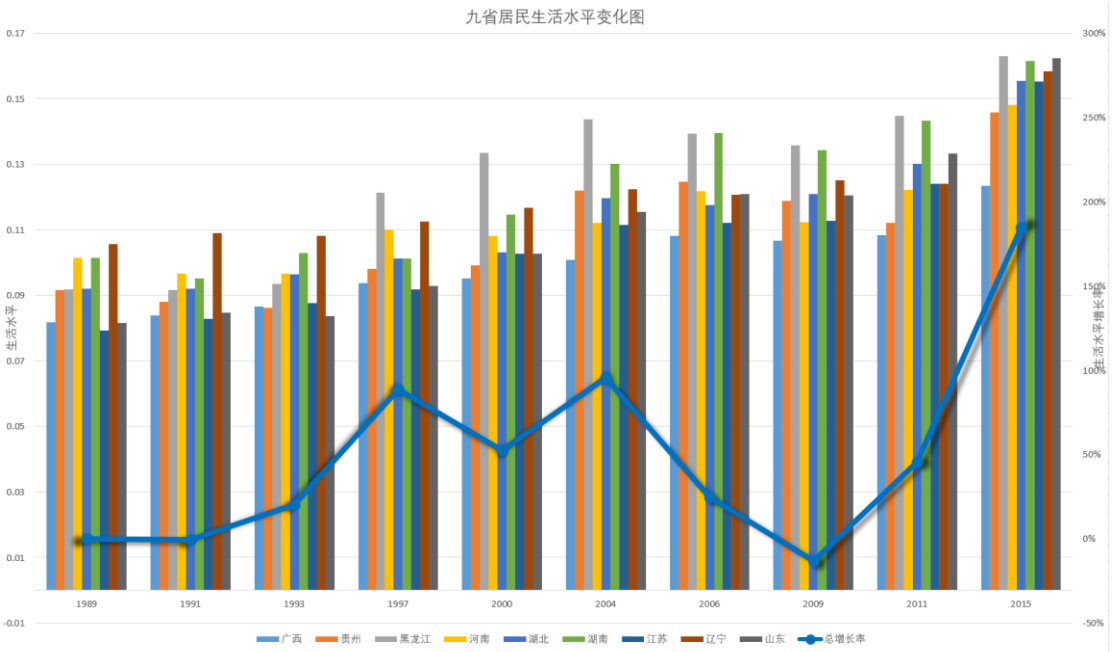
所得到的打分y为 1 维指标。

随后，对同一地区同一年份的个人指标进行平均，即可得到每个地区每个年份居民生活的平均水平 Y，如下图所示

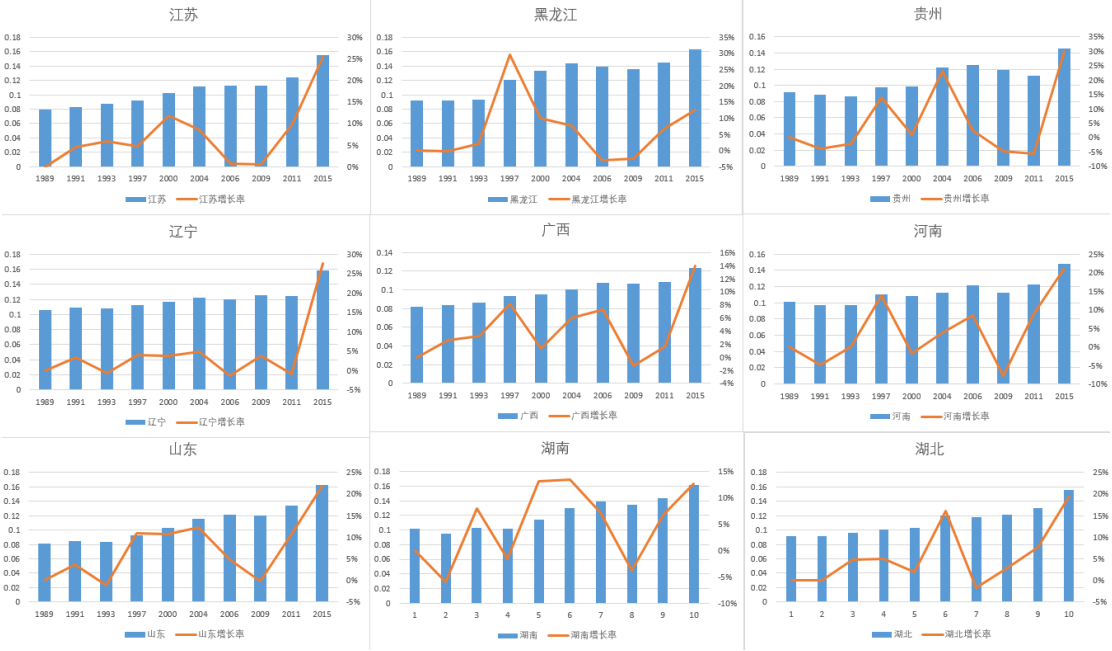
survey_wave	Guangxi	Guizhou	Heilongjiang	Henan	Hubei	Hunan	Jiangsu	Liaoning	Shandong
1991	0.026561	-0.038389	NaN	-0.047699	0.000401	-0.060339	0.045417	0.032349	0.037497
1993	0.032208	-0.022135	NaN	0.000393	0.048160	0.080211	0.058113	-0.007526	-0.011335
1997	0.082083	0.138650	NaN	0.138350	0.049739	-0.015679	0.046969	0.039960	0.108314
2000	0.013933	0.010740	0.099482	-0.017796	0.019645	0.132352	0.118398	0.038424	0.106862
2004	0.059559	0.229640	0.077171	0.037675	0.159816	0.135166	0.085554	0.047786	0.123370
2006	0.072781	0.023926	-0.030257	0.085088	-0.017626	0.071539	0.006904	-0.013983	0.047343
2009	-0.012854	-0.047976	-0.025990	-0.077291	0.028111	-0.037352	0.005943	0.037810	-0.002460
2011	0.015427	-0.055638	0.066076	0.088319	0.075892	0.067639	0.098744	-0.008768	0.106377
2015	0.139439	0.299958	0.125814	0.211000	0.194593	0.127155	0.252885	0.276850	0.217220

3.4. 数据可视化

在得到各地区居民生活水平打分之后，可以做出各地区生活水平变化的走势图如下所示：



上图中，柱状图表示各省市居民的生活水平，折线表示总体的居民生活水平增长率。从总体变化情况可以看到，改革开放以来，各个省份居民平均生活水平随着时间在不断地改善，且总体增长率只有在经济危机后的 2009 年出现负增长。各个地区的居民平均生活水平变化图如下所示：



3.5. 稳定性因子筛选

对于每个因子，我们得出每一年相对应之前年份因子的增长率，作为模型的输入 X，X 即代表了当地政府对该项城市化因素的投入和建设程度；并得到地区居民生活水

平相对于上一年的增长率，作为模型的标签 Y。

在横截面上，利用线性回归模型，将居民生活水平增长率 Y 和城市化建设程度 X 进行线性回归，得到当期的回归系数β。这样，每个因子就能得到一个β序列，如下图所示：

survey_wave	community_education_category	diversity_score	economic_component_score	housing_component_score	market_component_score	modern_market
1991.0	-0.405926	0.218865	0.113717	0.032838	-0.250983	
1993.0	-0.020752	0.593919	0.074709	0.066407	-0.050815	
1997.0	0.066622	0.084921	-0.002542	-0.095318	-0.205165	
2000.0	-0.273792	-0.718469	0.042921	-0.852382	0.348242	
2004.0	-0.346560	-0.107505	-0.110440	0.191182	-0.104187	
2006.0	-0.196642	0.504975	0.137310	0.089129	0.009630	
2009.0	0.139886	-0.265153	-0.276059	-0.346659	0.098747	
2011.0	0.267125	-0.478404	-0.219904	0.017412	-0.231008	
2015.0	0.183624	-0.060932	-0.299436	-1.266714	0.060870	

回归系数β的大小代表了当期因子对居民生活水平增长率 Y 的影响程度的大小。当β的绝对数值较大时，则认为当期因子对 Y 具有显著影响，我们希望找到在过去长时间内对 Y 具有稳定影响的因子，为此，我们构造了一个β的稳定性指标，其思路借鉴自夏普比率(sharpe ratio)，定义如下：

$$\text{stability} = \frac{E(abs(\beta_t))}{std(\beta_t)}, \qquad t = 1991, 1993, \ldots, 2015$$

该值越大，说明该因子在过去较长时间内对 Y 具有稳定的影响，因此我们选取出该值较大的因子。通过对比图形和设定阈值(0.8)，我们筛选出具有在时序分布上具有稳定影响的因子，如下所示：

factors	scores
social_services_score	1.011659
transportation_component_score	0.988663
urbanization_index	0.909783
community_education_category	0.905378
economic_component_score	0.881352
quality_of_health_score	0.851802
modern_markets_component_score	0.839542
market_component_score	0.834443
diversity_score	0.828914
population_density_score	0.797273
sanitation_score	0.767205
housing_component_score	0.698915

分别为：社会服务、交通发展水平、城市化水平、社区教育、经济发展、健康质量、现代市场发展水平、市场发展水平、多元化水平。

3.6. 因子重要性分析

3.6.1 建模原理

由以上过程我们筛选出了一些效果相对稳定的因子，这些因子与我们关注的居民生活水平之间可能会存在线性或者非线性的关系，为了衡量这种相关关系的大小，我们需

要对因子的重要性进行评分。为此，我们决定采用决策树类的方法，因为决策树类方法对线性和非线性关系具有较好的拟合效果，而且可以通过构建决策树的过程来评价因子的分类能力。

在模型的选择上，我们分别采用了随机森林和 Xgboost 模型，两者皆为决策树的集成方法(ensemble methods)，具有较好的拟合效果。其中随机森林是并行算法模型，具有较低的方差(variance)，但同时具有较低的准确度(precision)；Xgboost 是串行算法模型，具有较高的精准度，但可能会导致过拟合，从而有较高的方差。因此，我们将两者对因子重要性的评分结果进行结合对比，可以得到更为可信的分析结论。

3.6.2 建模方法

决策树的学习过程，以特征的信息增益程度作为对特征重要性的打分。以 CART(分类与回归树)生成过程为例，对于分类树采用 Gini 系数最小化准则进行特征选择，生成二叉树。

其中 Gini 系数的定义为，假设有 K 个类，样本点属于第 k 类的概率为$\mathrm{p_k}$，则概率分布的基尼系数定义为

$$\mathrm{Gini(p)} = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

给定样本集合 D，其 Gini 指数为：

$$\mathrm{Gini(D)} = 1 - \sum_{k=1}^{K} \left(\frac{C_k}{D}\right)^2$$

这里，C_k是 D 中属于第 k 类的样本子集，K 是类的个数。

在特征 A 的条件下，集合 D 的 Gini 系数定义为：

$$\mathrm{Gini(D,A)} = \frac{|D_1|}{|D|}\mathrm{Gini(D_1)} + \frac{|D_2|}{|D|}\mathrm{Gini(D_2)}$$

Gini(D)表示集合 D 的不确定性，Gini(D,A)表示经特征 A 分割后集合 D 的不确定性。

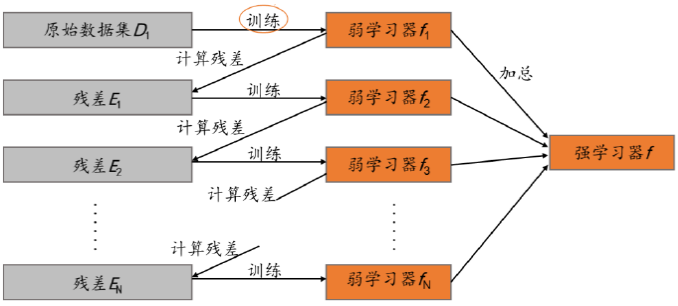
则特征重要性的计算需要借助于节点分裂时的Gini系数：

$$I_i(\mathrm{A}) = \mathrm{Gini(D_i)} - \mathrm{Gini(D_i}, A)$$

$$S(\mathrm{A}) = \sum_i I_i(A)$$

其中，$I_i(\mathrm{A})$表示结点i根据特征A分裂为两个子结点后，Gini 指数相对于母结点分裂前的下降值。故而可定义特征$\boldsymbol{A}$的绝对重要性$S(\mathrm{A})$为所有按特征A分裂的结点处的$I_i(\mathrm{A})$之和。

Xgboost 为决策树的串行集成算法，每次迭代训练一个弱学习器$\mathrm{f_i}$，学习第 k-1 个分类器得到的残差$\mathrm{E_{i-1}}$，最终通过加权各个弱分类器得到强分类器f. 并且在学习的过程中按照上述方法对特征的重要性进行记录，得到最终的总特征重要性评分。



而随机森林对特征重要性评分的思路是基于排列测试(permutation test)：如果一个特征非常重要，那么假设将该特征进行扰动，则会显著影响预测的准确性。对于特征x_i，将该特征进行随机的排列，计算随机排列后袋外错误率(out-of-bag error)的增加。如果袋外错误率的增加较大，说明该特征较重要，因此使用该值作为特征的重要性分值。

3.6.3 模型实验

在训练过程中，Xgboost 和随机森林的主要参数如下表所示：

参 数	说 明
max_depth	树的最大深度，较深的树容易过拟合，较浅的树容易欠拟合。
learning_rate	每个弱学习器的权重缩减系数，也称作步长，取值范围为[0, 1]。对于同样的训练集拟合效果，较小的学习率意味着需要更多的弱学习器的迭代次数。
n_estimators	最大的弱学习器的个数。与学习率 learning_rate 一同考虑。
early_stopping_rounds	提前停止训练。给定一个验证集，该模型将开始训练，直到验证得分停止提高为止，防止过拟合。

算法的具体流程如下：

1. 将数据集划分为训练集 Xtrain 和验证集 Xtest
2. 通过网格搜索(GridSearchCV)的方法在训练集样本 Xtrain 上寻找最优参数，同时将 Xtest 作为训练过程中的验证集来控制是否提前停止训练
3. 对步骤 2 得到的最优参数，在全样本上构造 XgBoost 和随机森林并拟合
4. 通过训练好的分类器对因子进行重要性打分，得到因子的重要性

首先，我们将原始数据划分为训练集和验证集，划分比例为 0.75 和 0.25。在训练集上，寻找模型的最优参数，待选参数表如下所示：

参 数	取值范围	适用模型
max_depth	[3, 5]	RF, Xgboost
learning_rate	[0.01, 0.1, 0.3]	Xgboost
n_estimators	[10, 20]	RF, Xgboost
early_stopping_rounds	50	Xgboost

3.6.4 模型结果

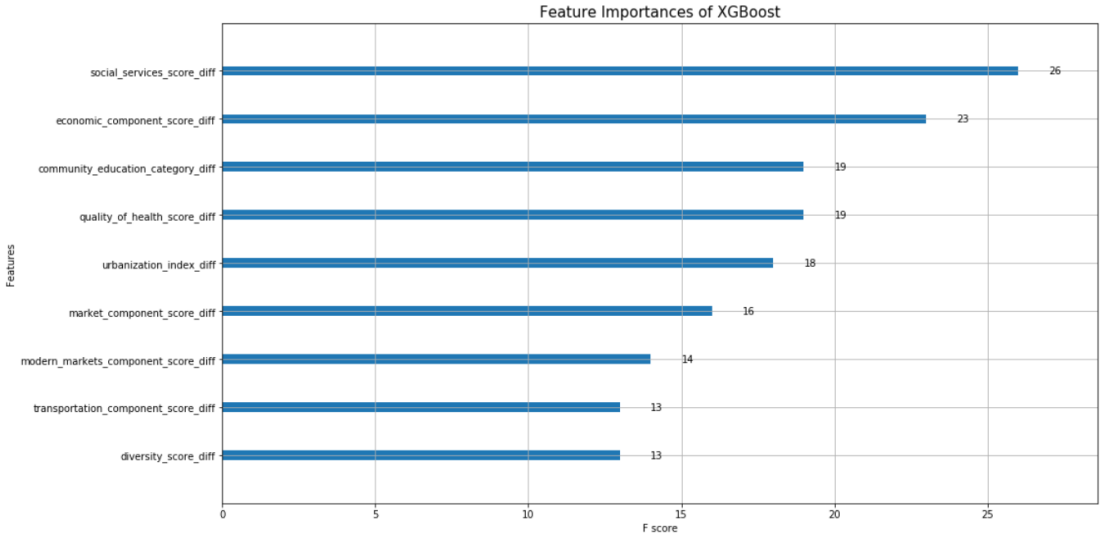
通过网格搜索，选择在验证集中平均误差最小的模型，Xgboost 对应的参数即为最优参数：

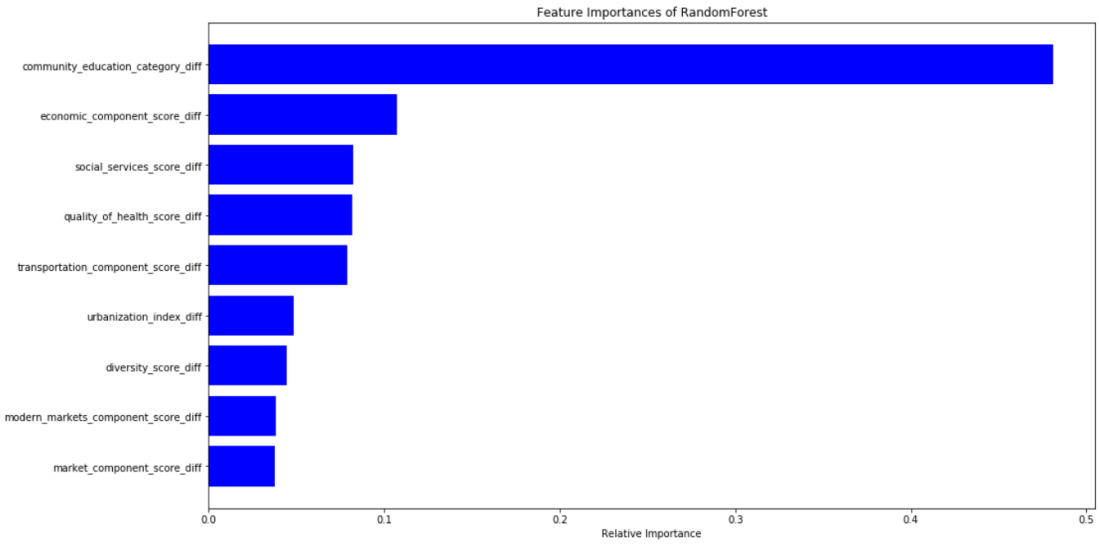
参 数	最优值
max_depth	3
learning_rate	0.3
n_estimators	20

随机森林对应的最优参数为：

参 数	最优值
max_depth	3
n_estimators	10

随机森林和 Xgboost 对因子的重要性评分结果如下所示：





可以看到，Xgboost 模型的结果中，社会服务，经济发展水平，社区教育和健康质量排名靠前，说明政府对这四项城市化建设对居民生活水平的提高影响最大；由随机森林模型得到的结果可以看到，社会服务，经济发展水平，社区教育和健康质量虽然排名顺序不同，但是其重要性同样排名靠前(1-4 名)。

因此我们可以得出结论，社会服务，经济发展水平，社区教育和健康质量对居民生活水平最有影响。该结论也说明了在现今社会，教育是影响一个人未来收入和生活质量的决定性因素，健康则是承受高强度工作的保证，而良好的经济发展则给予了一个受过良好教育，并且拥有健康体格的人得到高收入工作的机会。