**基于机器学习决策树模型的城区个人养老模式预测与分析**

作者1, 作者2

1. 工作单位
2. 工作单位

**摘要**

**目的：**通过对局部人群调查量表的计算分析，明确机器学习对于城区人群养老倾向性的指导意义。**方法：**建立养老相关资料调查量表，在市区内面对面发放并指导个人进行填写。资料收回汇总后规范化录入保存，应用Scikit Learn机器学习工具包进行分析。**结果：**调查采集2019年01月-2019年06月大连市区内个体共587人，年龄分为40周岁至80周岁以上共5个阶段，其中427人（72.7%）选择居家养老，160人（27.3%）选择养老院等家外养老方式。机器学习决策树模型基于全部资料按照f1评分标准CV score 0.76±0.10，预测家外养老precision 0.80，recall 0.84，f1 0.82，整体准确率0.88。结论：基于机器学习的城区家内外养老模型可以为城区家内外养老模式的制定提供参考。

**关键词：**机器学习，养老

养老是社会保障的一个种类，指对超过60岁以上的人口提供的闲居修养方式。养老意愿是指人们对自身养老所持有的愿望和打算，主要表现在对养老模式的不同选择上。中国的养老模式多样化，按照大体划分主要分为居家养老及家外养老。居家养老即老年人居住在家庭当中，其护理提供者可为家庭成员、保姆或者社区半专业护理人员，在许多国家被称之为informal care model；家外养老方式即老人寄宿于托老所、养老院等专业养老机构，其生活等护理由机构专业养老护理人员或者机构专业医务人员提供，一些国家称之为formal care model1。两种养老方式各有其利弊，其中居家养老方式满足了老年人不离家、不改变的归宿感，在减少老年人孤独感上略胜一筹，而家外养老机构能够担负起照顾老人的重担、及时为老年人解决生活、医疗等方面的困难，解决了儿女工作忙、不能兼顾老年人照护的难题，更是能及时为老年人的慢性病提供医疗、护理专业服务。相比之下，接受家庭成员、保姆、社区服务的居家老年人在获得专业医疗服务数量、质量上可能要低于机构照料2,3。事实上，不管哪种养老方式都具备各自的优势，如何根据不同区域、不同状况的人群养老需求，规划相关人力、物力、财力等配套制度，这是政府长期致力的一项重要工作内容。这些相关工作需要有科学、有效的政策制定依据，这依据之一就是提出“以人为本”为核心的理念4,5。如有针对性的对有社区居家养老需求的人群设立社区功能框架等；对有希望保姆提供护理的居家养老人群提供服务人员培养、人数需求预测等市场规划；针对有家外养老需求的老年人规划养老机构规模、数量；多年来，基于大规模的调查基础，并根据调查结果制定并实施相关政策不乏为一种可行的“以人为本”“按需供给”的科学方法，但各种研究均不可避免的具备其一定局限性，如何基于小规模的调查基础、实现群体的养老需求预测，反映民意又实现“有据可依”的科学政策制定方法，能够为推动养老监管框架、质量保证、政策制定、相关配套服务产业的设立做出微薄贡献，是本文立题的初衷。基于目前“互联网+大数据”背景下，机器学习方法也被越来越多的被应用到医疗保健、护理、老年人养老支持等各个领域的风险策略6-9。研究发现，个人的基本情况很大程度上决定了其养老倾向，如教育水平、是否独居、年龄等因素均是影响老年人是否接受寄宿式护理的社会经济因素10。单一的指标化难以预测个体养老意愿，常规的统计方法只能给出两组的差异性、养老意愿的相关因素等，在限定区域内利用机器学习方式，可以针对获得的城市人群养老意愿调查量表结果进行预测，使得探索内在规律预测个人养老意愿成为可能，借以将预测结果推广到区域化全民11。

## 材料与方法

**1.1 问卷设计**：自行设计调查问卷，并经过同行专家审订而成。问卷内容涉及调查者基本状况共10项，具体包括年龄、性别、职业、教育程度、自理状况、收入、共居状况、子女数、与子女亲密程度，未来养老模式的选择。其中养老模式包括居家养老及家外养老模式两大类：一是居家养老模式，包含由子女、亲人提供护理的或者由社区或者家政专业服务机构提供的居家养老；二是家外养老模式，主要是指养老院、托老所等寄宿式专业养老机构提供正式养老服务的养老方式。

**1.2 问卷发放：**调查于2019年1月至6月期间，调查对象为大连市内四区包括中山区、西岗区、沙河口区及甘井子区居民，为了保证抽样的真实性及代表性，访问员均进行严格培训，并亲自到市内四区进行问卷发放，一对一进行问卷调查和访谈，由调查员按照问卷项目逐项提问并记录。

**1.2 统计分析：**所有数据输入文件后常规统计，导入至Jupyter Notebook通过Python 3.8/Scikit Learn0.21机器学习工具包进行分析。拟选用便于理解的决策树模型，初筛对照使用的模型为LR(LogisticRegression)、LDA(LinearDiscriminantAnalysis)、QDA(QuadraticDiscriminantAnalysis)、SVM、SVC、KNN(KNeighborsClassifier)、NB(GaussianNB)、CART(DecisionTreeClassifier)、RFC(RandomForestClassifier)、ABC(AdaBoostClassifier)、GBC(GradientBoostingClassifier)、ETC(ExtraTreesClassifier)、MLPC(MLPClassifier)、XGB(XGBClassifier)、VC(VotingClassifier)。

## 2、结果

**2.1 基本情况：**本次调查共发放问卷587份，获得有效问卷587份，有效回收率为100%。基本情况如表1。

**表1 调查人群构成特征**

|  |  |  |
| --- | --- | --- |
| 特征 | *n* | 百分比 |
| 年龄 |  |  |
| 40--55岁 | 196 | 33.4 |
| 55—60岁 | 38 | 6.5 |
| 60—70岁 | 215 | 36.6 |
| 70—80岁 | 126 | 21.5 |
| 80岁以上 | 12 | 2.0 |
| 性别 |  |  |
| 男 | 212 | 36.1 |
| 女 | 375 | 63.9 |
| 教育程度 |  |  |
| 小学 | 209 | 35.6 |
| 初中 | 98 | 16.7 |
| 高中 | 66 | 11.2 |
| 专科 | 56 | 9.5 |
| 本科及以上 | 158 | 26.9 |
| 职业 |  |  |
| 企业单位 | 255 | 43.4 |
| 事业单位 | 212 | 36.1 |
| 机关政府部门 | 42 | 7.2 |
| 个体 | 40 | 6.8 |
| 无正式工作 | 38 | 6.5 |
| 月收入水平 |  |  |
| 0-2000 | 155 | 26.4 |
| 2000-4999 | 151 | 25.7 |
| 5000-7999 | 162 | 27.6 |
| 8000-10000 | 64 | 10.9 |
| 10000以上 | 55 | 9.4 |
| 共居状况 |  |  |
| 老伴 | 318 | 54.2 |
| 子女 | 58 | 9.9 |
| 老伴及子女 | 96 | 16.4 |
| 独居 | 91 | 15.5 |
| 其他 | 24 | 4.1 |
| 自理状况 |  |  |
| 完全自理 | 364 | 62.0 |
| 基本自理 | 155 | 26.4 |
| 半自理 | 53 | 9.0 |
| 不能自理 | 15 | 2.6 |
| 子女数 |  |  |
| 0 | 34 | 5.8 |
| 1 | 188 | 32.0 |
| 2 | 295 | 50.3 |
| 3个及以上 | 70 | 11.9 |
| 与子女亲密程度 |  |  |
| 无 | 34 | 5.8 |
| 非常亲密 | 269 | 45.8 |
| 比较亲密 | 140 | 23.9 |
| 一般 | 115 | 19.6 |
| 疏远 | 29 | 4.9 |

**2.2 养老意愿的分布**：在有效样本分布中，427人选择居家养老方式，占总人数比例的72.7%，在这427人中，其中278人选择由子女、直系亲属进行护理的居家养老方式，占总人数的47.4%，149人选择由社区、保姆非专业机构提供护理的居家养老方式，占总人数25.4%；160人选择家外养老方式，占总人数比例的27.3%，见表2。

表2 调查对象养老方式选择百分比

|  |  |  |
| --- | --- | --- |
| 养老方式 | n | 百分比 |
| 居家养老  子女、亲属进行  社区、保姆进行 | 427  278  149 | 72.7  47.4  25.4 |
| 家外养老 | 160 | 27.3 |

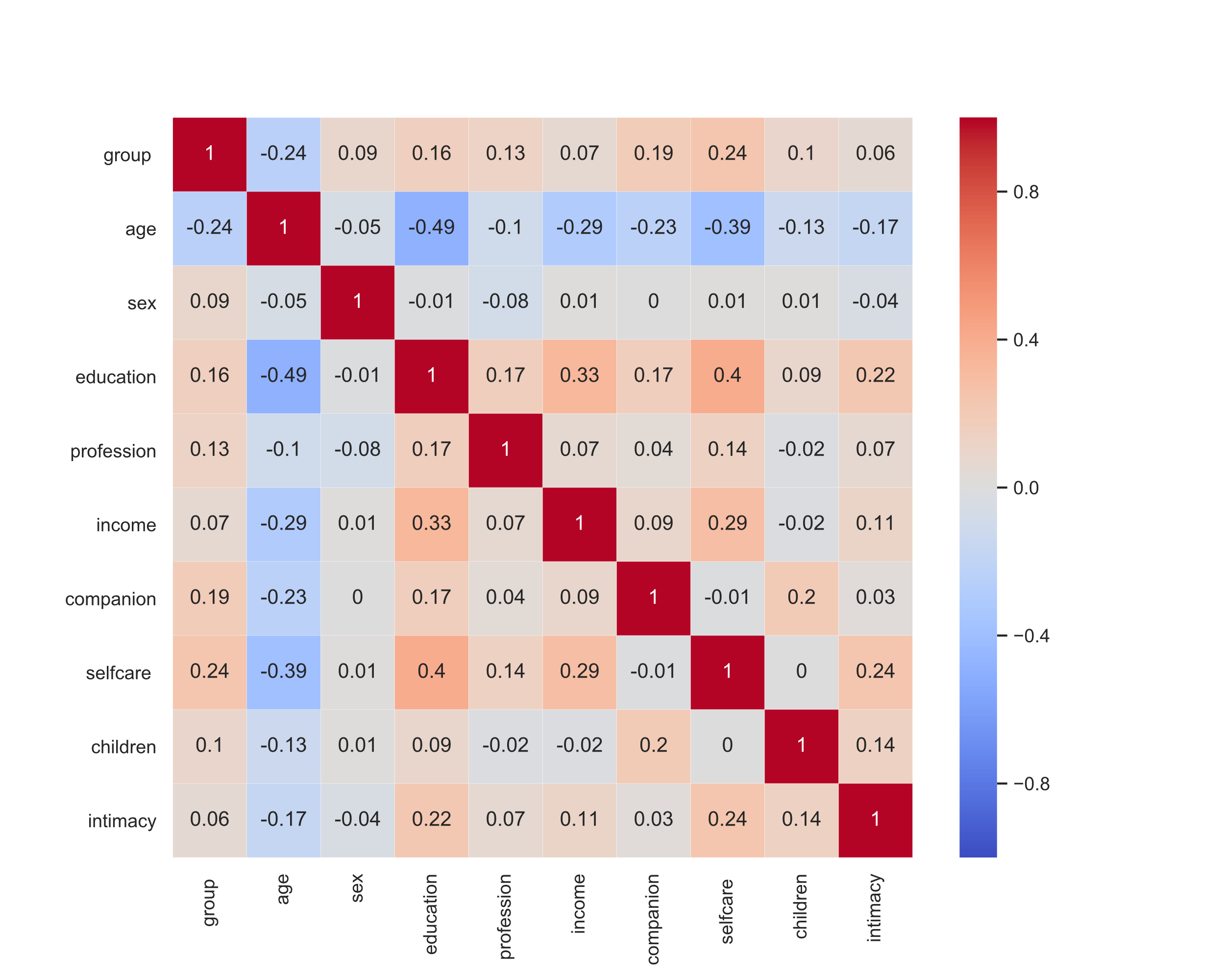
**2.3 养老意愿的影响因素：**在居家、家外养老选择方式上，年龄、性别等9项客观因素不同的中老年人，其养老方式均选择不同，见表3。

表3 调查对象的基本状况对养老方式选择比较分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 特征 | 居家养老 | 家外养老 | *χ2* | *df* | *P* |
| 年龄 |  |  | 282.719 | 4 | 0.000 |
| 40-55岁 | 104 | 92 |  |  |  |
| 55--60 | 26 | 12 |  |  |  |
| 60--70 | 188 | 27 |  |  |  |
| 70--80 | 104 | 22 |  |  |  |
| 80岁以上 | 5 | 7 |  |  |  |
| 性别 |  |  | 45.262 | 1 | 0.000 |
| 男 | 166 | 46 |  |  |  |
| 女 | 261 | 114 |  |  |  |
| 文化  小学  初中  高中  专科  本科及以上  职业 | 173  70  49  32  103 | 36  28  17  24  55 | 143.332  390.658 | 4  4 | 0.000  0.000 |
| 企业单位 | 201 | 54 |  |  |  |
| 事业单位 | 148 | 64 |  |  |  |
| 机关政府单位 | 38 | 4 |  |  |  |
| 个体 | 24 | 16 |  |  |  |
| 无正式职业 | 16 | 22 |  |  |  |
| 月收入水平 |  |  | 97.898 | 4 | 0.000 |
| 10000以上 | 35 | 20 |  |  |  |
| 8000-10000 | 42 | 22 |  |  |  |
| 5000-7999 | 127 | 35 |  |  |  |
| 2000-4999 | 110 | 41 |  |  |  |
| 0-2000 | 113 | 42 |  |  |  |
| 共居状况 |  |  | 456.961 | 4 | 0.000 |
| 老伴 | 261 | 57 |  |  |  |
| 子女 | 31 | 27 |  |  |  |
| 老伴及子女 | 60 | 36 |  |  |  |
| 独居 | 67 | 24 |  |  |  |
| 其他 | 8 | 16 |  |  |  |
| 自理状况 |  |  | 500.257 | 3 | 0.000 |
| 完全自理 | 232 | 132 |  |  |  |
| 基本自理 | 135 | 20 |  |  |  |
| 半自理 | 48 | 5 |  |  |  |
| 不能自理 | 12 | 3 |  |  |  |
| 子女数 |  |  | 288.128 | 3 | 0.000 |
| 0 | 15 | 5 |  |  |  |
| 1 | 110 | 38 |  |  |  |
| 2 | 128 | 97 |  |  |  |
| 3个及以上 | 25 | 9 |  |  |  |
| 与子女亲密程度 |  |  | 325.973 | 4 | 0.000 |
| 非常亲密 | 187 | 82 |  |  |  |
| 比较亲密 | 99 | 41 |  |  |  |
| 一般 | 98 | 17 |  |  |  |
| 疏远 | 23 | 115 |  |  |  |

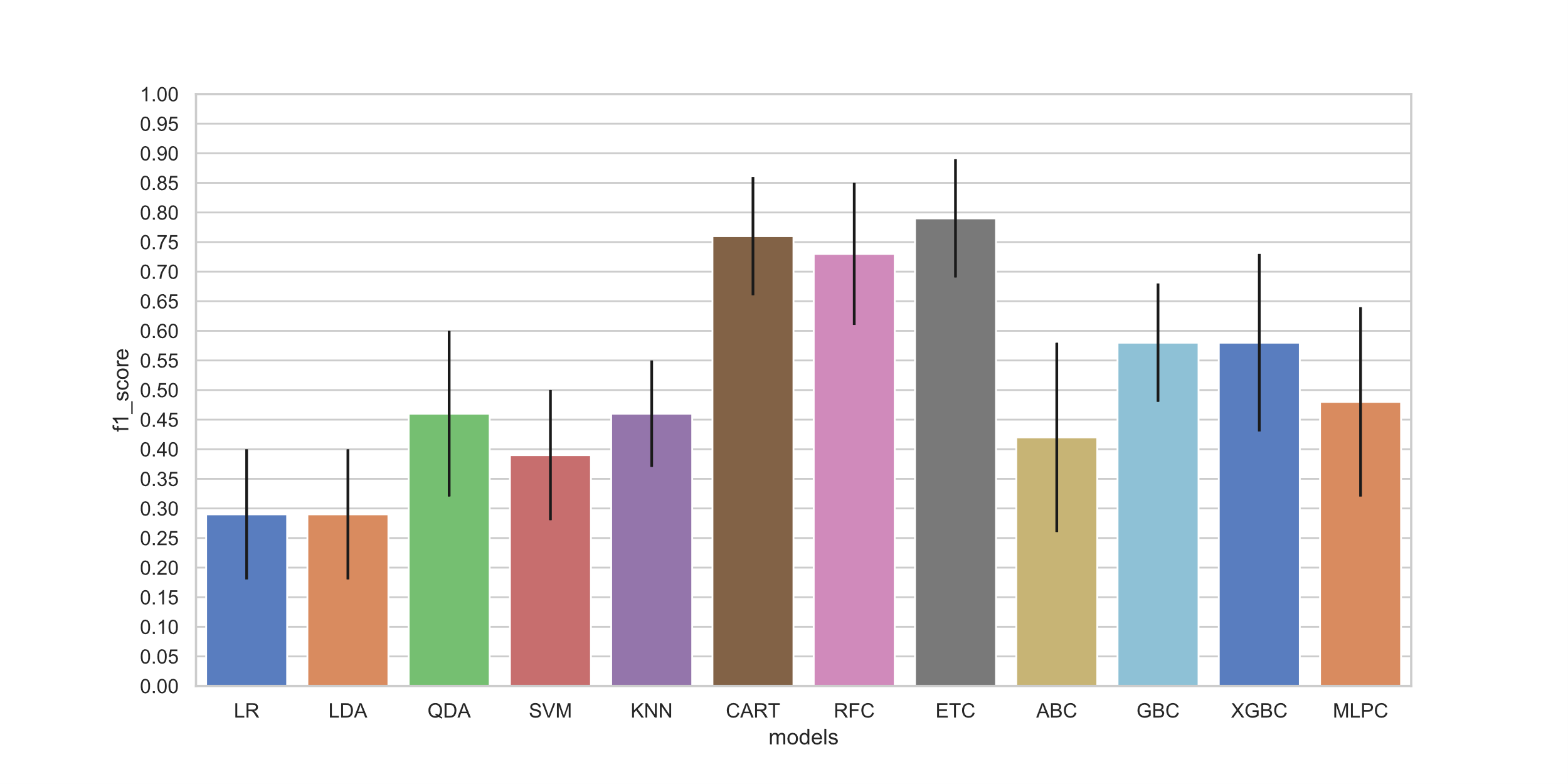
**2.4 相关分析：**将各项因素按Kendall法做相关分析，如图1。可见教育与年龄呈现负相关，由于历史原因的影响，年长者受教育水平、学历水平普遍较年轻者偏低。影响养老意愿因素中，其中sex,children存在正相关问卷顺序：年龄、文化递增顺序；收入、自理状况递减顺序。（已修改，5收入、7自理、14子女亲密度变成递增顺序）

**Figure1. Correlation**



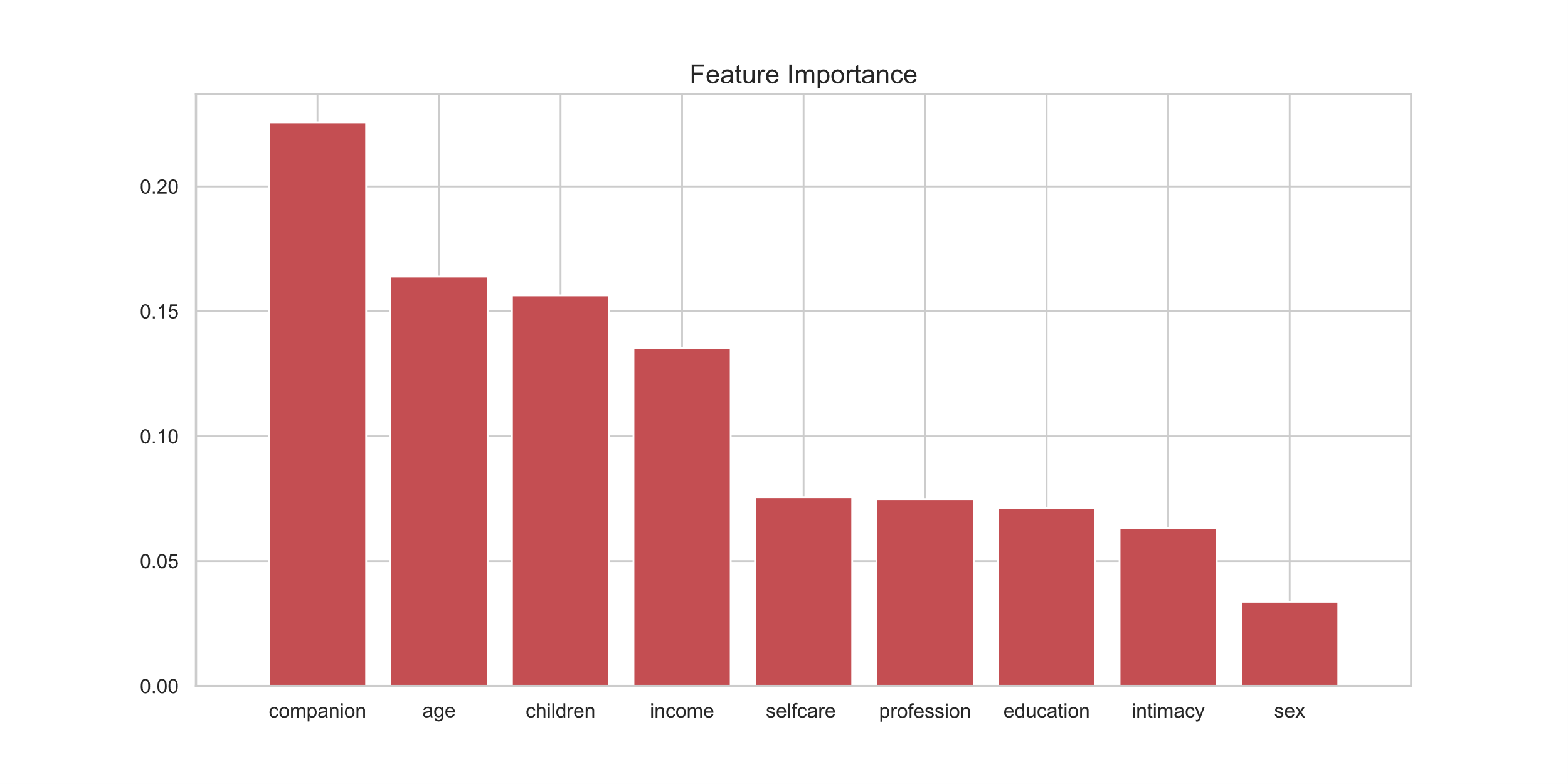
**2.5 模型选择：**筛选时以交叉验证的f1\_score均值作为评分标准，各分类器按照默认参数进行计算。决策树模型达到0.76的交叉验证评分，后续数据以决策树模型进行分析，见图2.

**Figure2. Model Selection**



**2.6 特征重要性分析：**对养老模式选择影响最大的是Companion、age、children、income这四项，性别特征影响最弱，见图3。

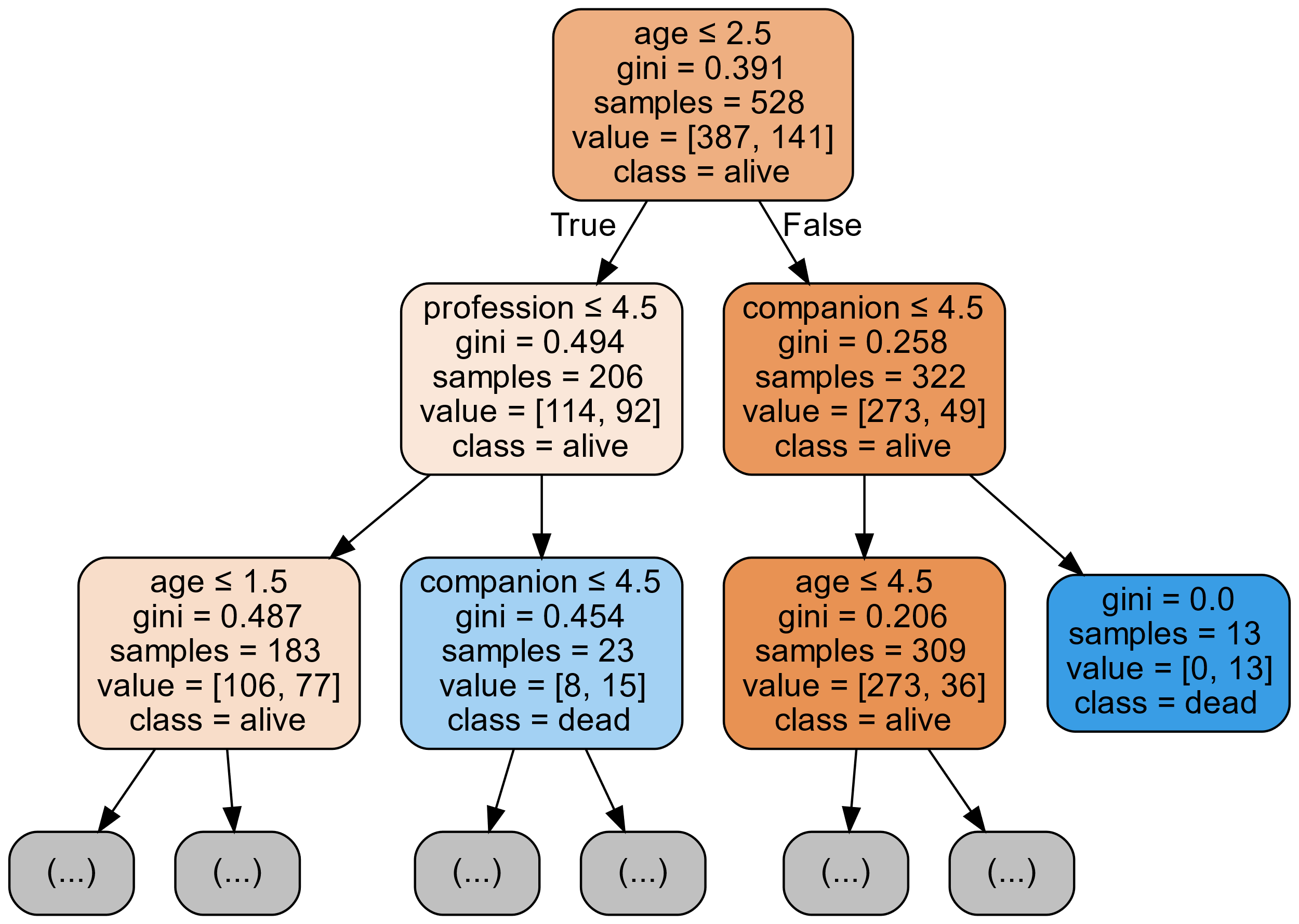
**Figure3. Feature Importance**



**2.7 模型调参:** 通过GridSearchCV最终选择参数为DecisionTreeClassifier(ccp\_alpha=0.0, class\_weight=None, criterion='gini', max\_depth=17, max\_features=8, max\_leaf\_nodes=None, min\_impurity\_decrease=0.0, min\_impurity\_split=None, min\_samples\_leaf=1, min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0, presort='deprecated', random\_state=42, splitter='best')

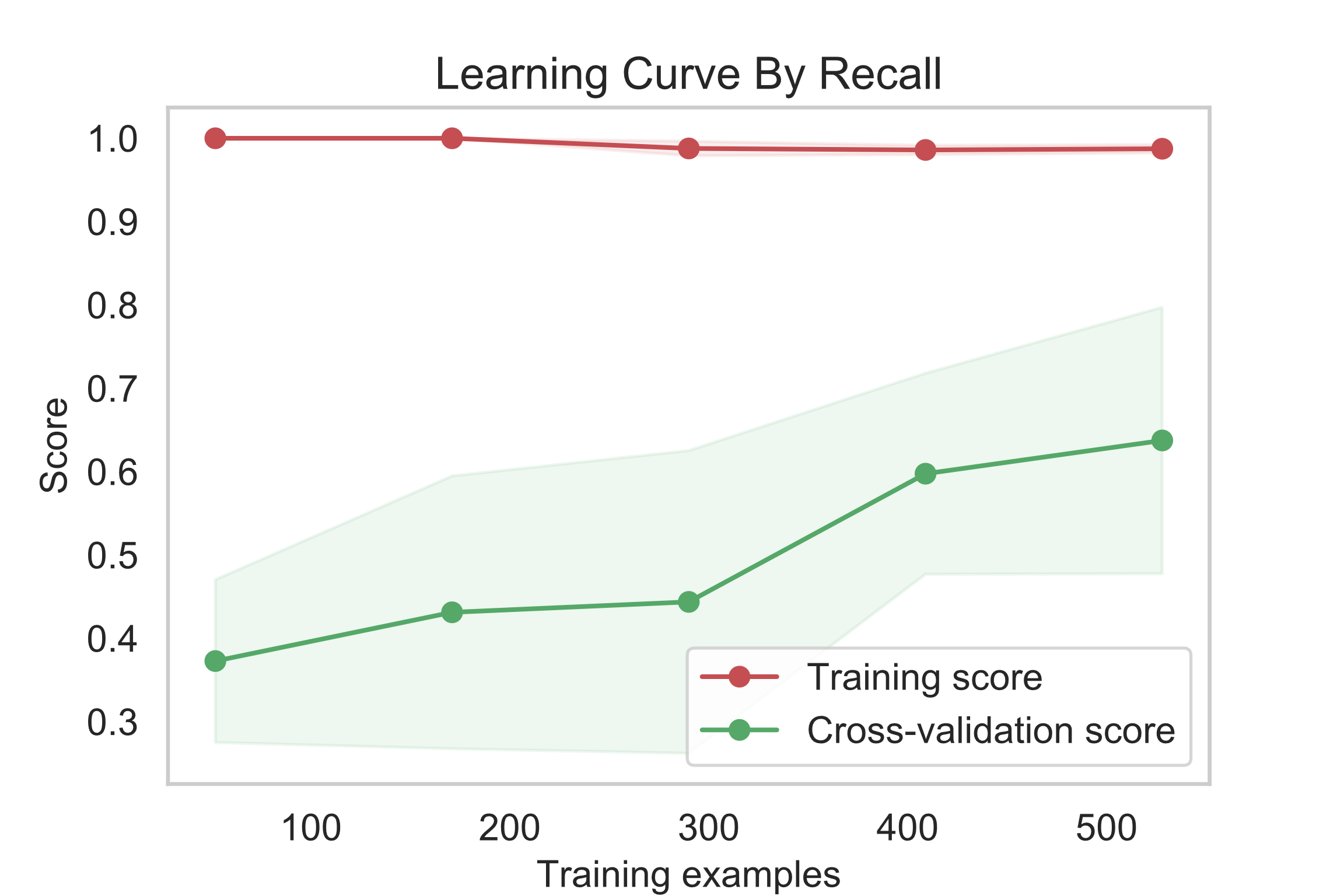
**2.8 决策树图示：见图4。**

**Figure4. Decision Tree**



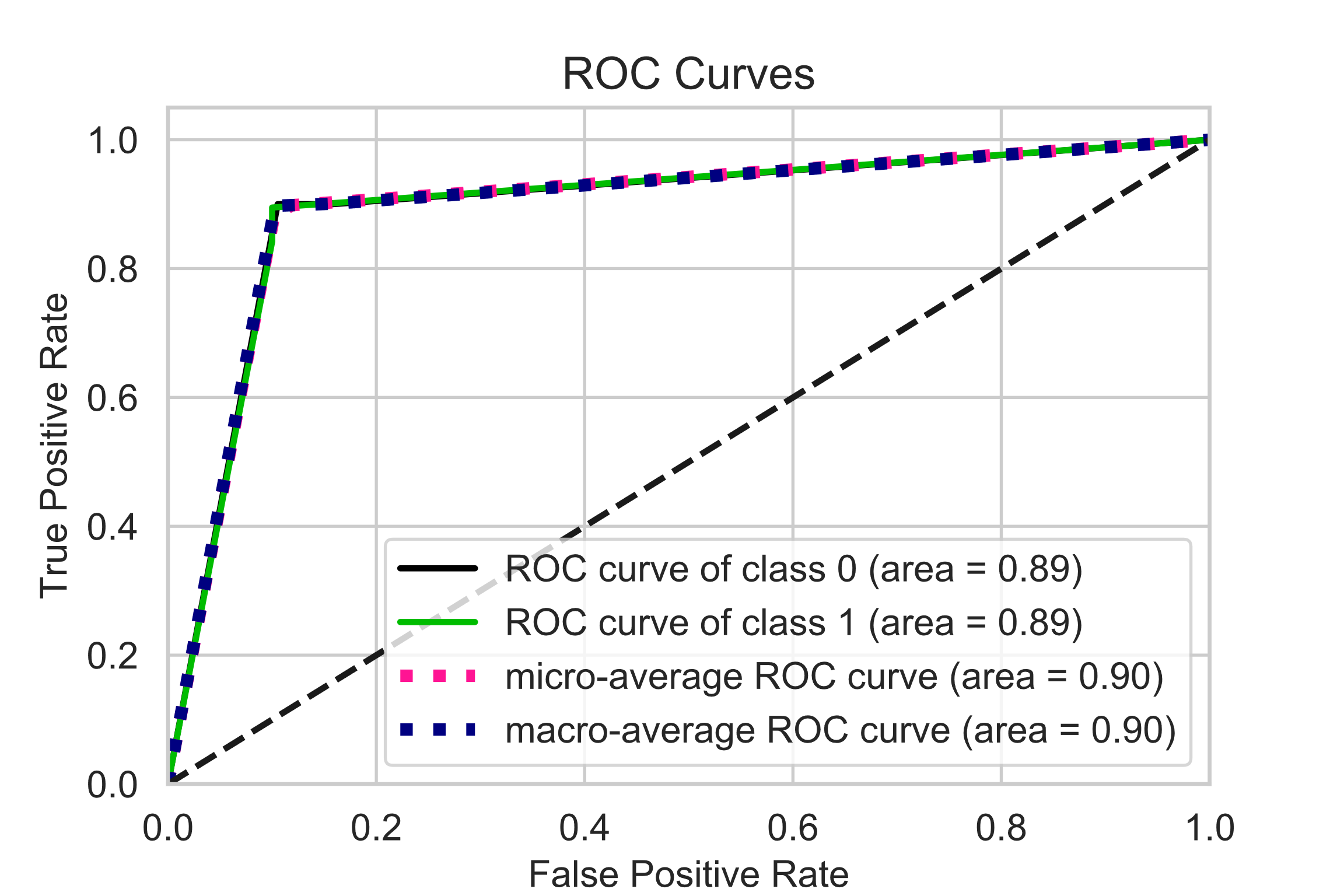
**2.9 学习曲线：**样本例数对模型的样性，如果加大样本例数还可以进一步提到准确率，见图5。

**Figure5. Learning Curve By Recall**

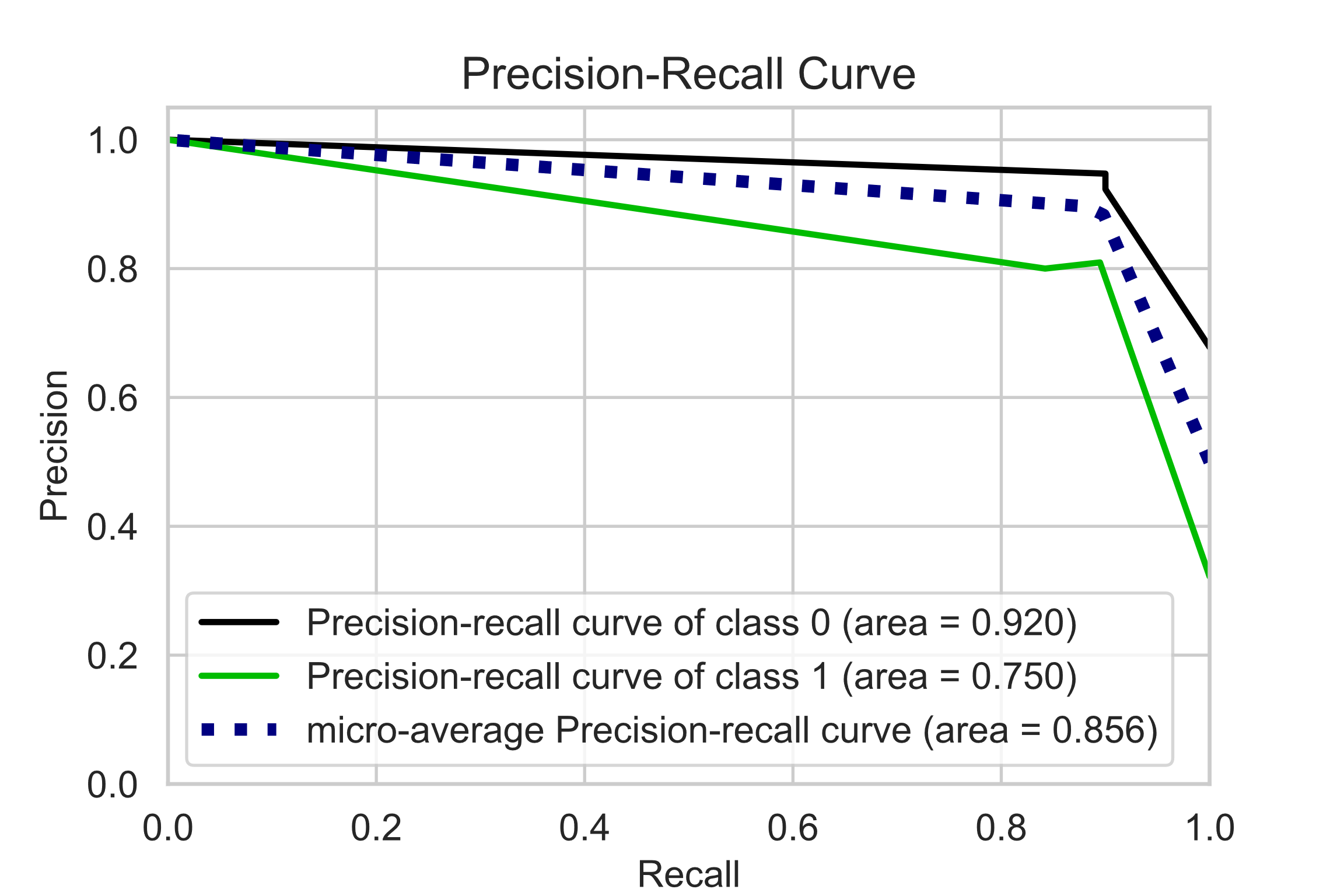


**2.10 ROC曲线：**受试者操作特性曲线(ROC)下面积评估模型class 1 0.89，诊断准确性尚可，见图6。

**Figure6. ROC Curves**

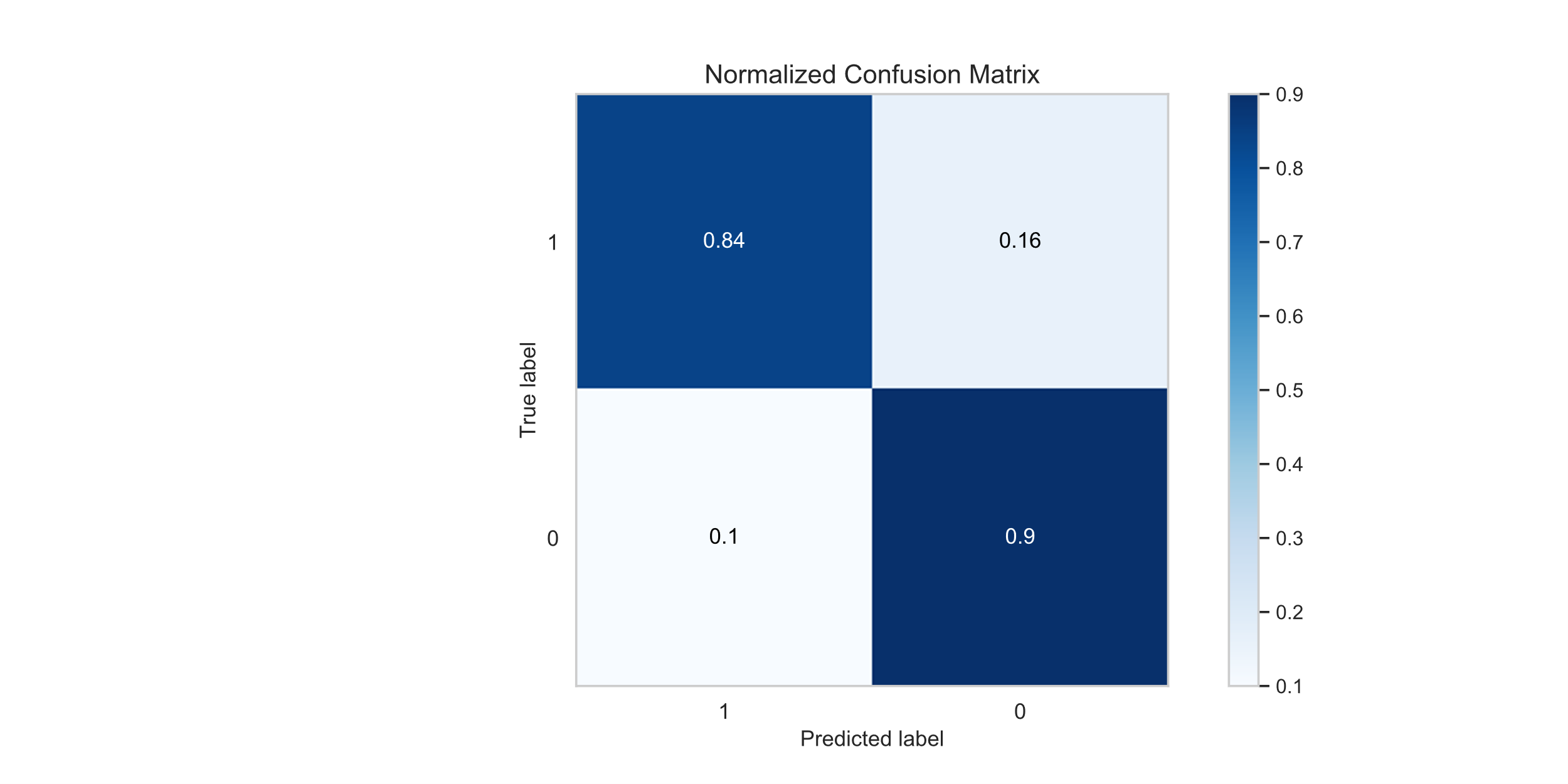


**2.11 PR曲线：**0.8的召回率以后会明显下降，见图7。



**2.12 混淆矩阵：**测试数据结果的标准化混淆矩阵家外养老precision 0.80，recall 0.84， f1-score 0.82，总体accuracy 0.88，见图8。

**Figure8. Normalized ConFusion Matrix**



**3、讨论**

3.1 调查结果显示，绝大多数人倾向于居家养老，这是与目前世界各地的养老意愿、趋势所相符的14，对于大多数老年人以及即将步入老年行列的中青年人来说，其仍然以居家养老作为首要的养老方式选择15-17。在养老服务支持政策中，家庭护理所需要的相关费用支出、社会支持政策方面，仍然是中国政府与大部分国家政府所承担的重任15。 社区-居家养老模式中，这种模式即老年人居家，由社区提供护理的居家养老模式、寄宿式养老模式对于政策制定者以及接受照顾的老年人来说都是一个可行的选择。从市场角度，它承诺了更高的成本效益，从养老意愿角度，它尊重了越来越多的人的偏好，让老年人留在自己的家里，18越来越多的人需要居家的社区服务，这种态势下，政府需要按照不同客观状况提供社区护理人员数目、对应护理项目、津贴补助，规划寄宿式养老机构的规模、床位数、辅助器具的配备等，均可利用小规模调查基础数据结果，结合机器学习预测方法，实现达到预测并加以应用到的效果，科学的制定相关政策、制度，合理规划资源配置，避免资源不足或者浪费。

3.2 在选择居家及家外两种不同模式养老方式中，性别、年龄、子女数、共居者状况均在一定程度上影响老年人的养老意愿，而养老机构规模、床位数的设置是否与当前的人群养老意愿所相符，利用机器学习方法分析人群的客观状况，将家外养老意愿的相关人群做以预测，可以更科学、合理的规划养老机构相关配备、避免浪费。

3.3 当然除了以上的相关益处，机器学习是否能够应用于养老的其他领域、例如根据养老需求的调查来预测相关部门的护理提供项目、老年护理部门质量监测、国家养老资源正确利用与否等，这尚需要长时间的探索及实践，以及其他方法的科学利用，毕竟在目前，对于在养老相关部门以及老年护理部门实施治理、制裁的风险还是存在的。19

当政府以及相关养老机构对养老院规模进行规划时，需要对现有以及未来10-20年的家内外养老人群需求人群进行评估。城市人群的养老意愿可以对调查量表运用机器学习的方法进行预测，个人的基本情况很大程度决定了其养老倾向，这种方式有助于政府、养老相关养老部门对局部人群养老的政策进行更科学的规划。

3.4 当政府对现有养老院规模进行评价，评价是否能够满足现有以及未来10-20年内家外养老人群需求时，以及制定相应养老制度时，需要对模型中分类1即家外养老人群进行充分评估。当分类为1时，就是政府评估时尽量将所有为1的分类，就是家外养老的人挑选出来进行计算和决策，所以recall要尽量大一些，同时保证precision更高些，这样f1指标就很重要。

**结论：**

机器学习决策树模型可以为个人养老模型进行预测，有益于政府养老规划的制定。

**References**

**1.** Temple JB, Jukic M, Dow B. Informal care relationships and residential aged care recommendations: evidence from administrative data. *BMC Geriatr.* Dec 19 2017;17(1):289.

**2.** Su SW, Wang D. Health-related quality of life and related factors among elderly persons under different aged care models in Guangzhou, China: a cross-sectional study. *Qual Life Res.* May 2019;28(5):1293-1303.

**3.** Visvanathan R, Amare AT, Wesselingh S, et al. Prolonged Wait Time Prior to Entry to Home Care Packages Increases the Risk of Mortality and Transition to Permanent Residential Aged Care Services: Findings from the Registry of Older South Australians (ROSA). *J Nutr Health Aging.* 2019;23(3):271-280.

**4.** Jorgensen M, Siette J, Georgiou A, Warland A, Westbrook J. Modeling the Association Between Home Care Service Use and Entry Into Residential Aged Care: A Cohort Study Using Routinely Collected Data. *J Am Med Dir Assoc.* Feb 2018;19(2):117-121 e113.

**5.** Edvardsson D, Fetherstonhaugh D, Nay R, Gibson S. Development and initial testing of the Person-centered Care Assessment Tool (P-CAT). *Int Psychogeriatr.* Feb 2010;22(1):101-108.

**6.** Chen R, Lu A, Wang J, et al. Using machine learning to predict one-year cardiovascular events in patients with severe dilated cardiomyopathy. *Eur J Radiol.* Aug 2019;117:178-183.

**7.** Lo Y, Lynch SF, Urbanowicz RJ, et al. Using Machine Learning on Home Health Care Assessments to Predict Fall Risk. *Stud Health Technol Inform.* Aug 21 2019;264:684-688.

**8.** Antonelli M, Johnston EW, Dikaios N, et al. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *Eur Radiol.* Sep 2019;29(9):4754-4764.

**9.** Hung M, Voss MW, Rosales MN, et al. Application of machine learning for diagnostic prediction of root caries. *Gerodontology.* Dec 2019;36(4):395-404.

**10.** Yu S, Byles J. Waiting times in aged care: What matters? *Australas J Ageing.* May 3 2019.

**11.** Pleschberger S, Wosko P. [Informal non-kin support for elderly people living alone and end of life care. Literature review]. *Zeitschrift fur Gerontologie und Geriatrie.* Jul 2015;48(5):457-464.

**12.** Wu M, Zhong X, Peng Q, et al. Prediction of molecular subtypes of breast cancer using BI-RADS features based on a "white box" machine learning approach in a multi-modal imaging setting. *Eur J Radiol.* May 2019;114:175-184.

**13.** Patterson BW, Engstrom CJ, Sah V, et al. Training and Interpreting Machine Learning Algorithms to Evaluate Fall Risk After Emergency Department Visits. *Med Care.* Jul 2019;57(7):560-566.

**14.** Walker H, Paliadelis P. Older peoples' experiences of living in a residential aged care facility in Australia. *Australas J Ageing.* Sep 2016;35(3):E6-e10.

**15.** Feng Z. Global Convergence: Aging and Long-Term Care Policy Challenges in the Developing World. *J Aging Soc Policy.* Jul-Sep 2019;31(4):291-297.

**16.** Swain N. Distress in informal carers of the elderly in New Zealand. *N Z Med J.* Nov 9 2018;131(1485):60-66.

**17.** Zamanzadeh V, Rahmani A, Pakpour V, Chenoweth LL, Mohammadi E. Psychosocial changes following transition to an aged care home: qualitative findings from Iran. *Int J Older People Nurs.* Jun 2017;12(2).

**18.** Palesy D, Jakimowicz S, Saunders C, Lewis J. Home care in Australia: an integrative review. *Home Health Care Serv Q.* Apr-Jun 2018;37(2):113-139.

**19.** Ellis JM, Howe A. The role of sanctions in Australia's residential aged care quality assurance system. *Int J Qual Health Care.* Dec 2010;22(6):452-460.