DATA620007 数据挖掘 第 17 小组期末报告
小组成员：胡一航 22210980041，王宝琪 22210980075，阚舒耀 22210980046，姚闰翔 22210980125，梁溢笙 22210980124
2023 年 6 月 15 日

1 数据来源：本数据来自至 kaggle 上的一个比赛的真实数据。
https://www.kaggle.com/datasets/uciml/adult-census-income

2 数据集介绍
人口普查数据集从美国 1994 年人口普查数据库提取，共有 32561 条记录，分为训练数据集（26048 条）和测试数据集（6513 条）。我们整理出了一个变量表格，可以看到所有变量的含义、取值范围等信息。数据集中有 6 个连续型数值变量（年龄、权重、求学时长、资本收益、资本损失、每周工作时长），8 个类别变量（工作类型、教育水平、婚姻状况、职业、关系、种族、性别、国籍）。如图 1 示，数据共有 15 个列，其中表格前 14 列（年龄，工作类型、婚姻情况）自变量，最后一列收入为因变量，因变量已经区间化为大于 5 万美金和小于 5 万美金。我们希望通过对 14 个自变量的分析来预测人口的收入水平。

变数类型	变量名	变量含义	取值
自变量	Age	年龄	continuous
	workclass	工作类型	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
	fnlwgt	多少人有着类似的特征	continuous
	education	教育水平	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
	education.num	求学时常（年）	continuous
	marital.status	婚姻状态	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
	occupation	职业	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
	relationship	关系	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
	race	种族	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
	sex	性别	Female, Male
	capital.gain	资本收益	continuous
	capital.loss	资本损失	continuous
	hours.per.week	每周工时	continuous
	native.country	国籍	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
因变量	income	年收入	<=50K, >50K

图 1

3 数据预处理
在数据预处理过程中，首先发现数据集中存在很多缺失值，大概有 2000 個缺失值，数量较多。因此我们采用众数填补法，使用每个属性的众数来填补缺失值。

此外，通过变量表（圖 1）我们发现除了性别变量和因变量收入，其他类别变量

取值范围较大，且不存在明显的序数关系，不适合 one-hot 编码和标签（序数）编码，因此我们采用了目标编码方法，而对性别变量和因变量 income 我们采用了 01 编码。目标编码的方法是，将该列中的每个类别值都用该类别的平均目标值替代。例如，对于自变量教育中的值学士，我们用所有学士对应的标签的均值来代替值学士。之后，我们对所有数据进行标准化，以消除不同量纲的影响，方便后续处理。

最后，通过计算，我们发现收入大于 5 萬的人数与小于 5 萬的人数比大致为 1:3，数据明显出現不平衡。因此，我们采用 SMOTE 算法消除样本不均衡的影响。

4 描述性分析

首先我们建立一个热力图探索变量之间的相关系数矩阵如图 2 示，在图 2 中蓝色到红色代表相关系数从-1 到+1，可以看出自变量间 education（教育水平）和 education.num（教育时长）有着 0.9 的高度正相关，代表着教育水平越高，教育时长越长，这与我们的常识也符合，同时 marital.status 和 relationship 有 0.97 的高度正相关，因为婚姻关系其实与家庭关系两者高度相关。自变量与因变量之间婚姻状况和家庭关系与收入有较高的正相关性，其次是教育水平与教育时长。

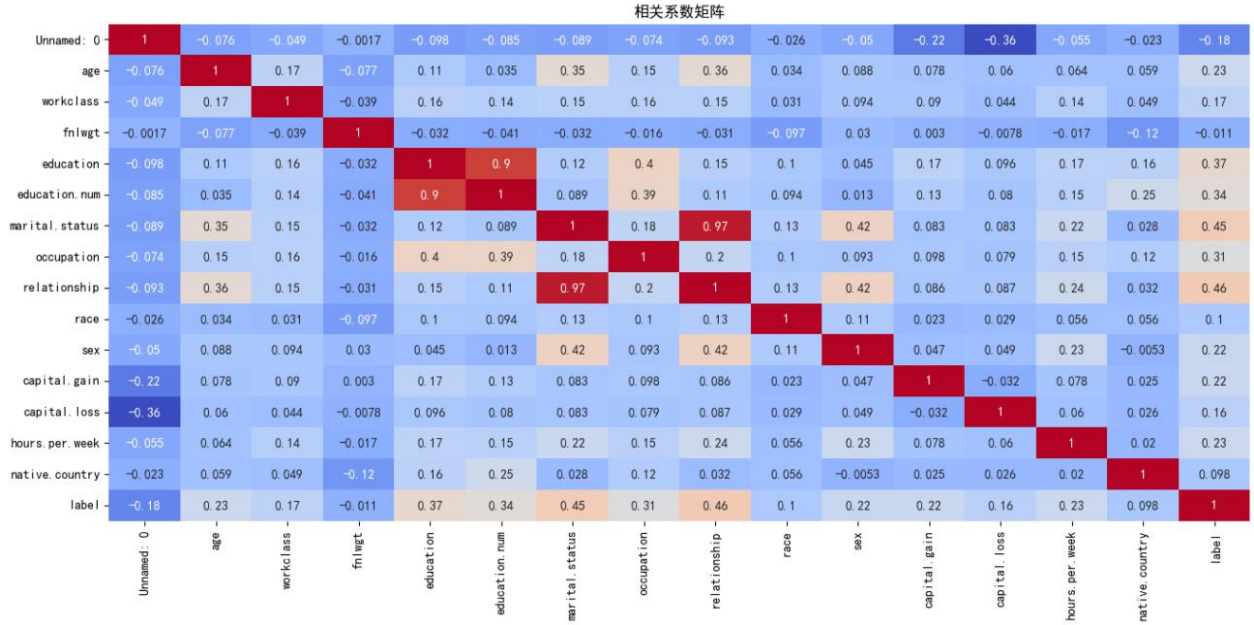


图 2

然后绘制不同连续性变量的分布直方图以观察它们的分布。观察图 3，可以发现年龄与 fnlwgt 呈现右偏分布，说明年龄处于 40 岁以下占大多数。教育时长在 8-10 年的人数最多，其次是 12-13 年，说明拥有高中学历和大学学历的人数较多。资本损失与收益方面大部分分布也较为集中，计算得出 75%的人都没有资本收益与支出，我们认为属于无关变量。每周工作时长绝大多数集中于 40 小时左右，这可能与西方国家的 8 小时工作制有关。

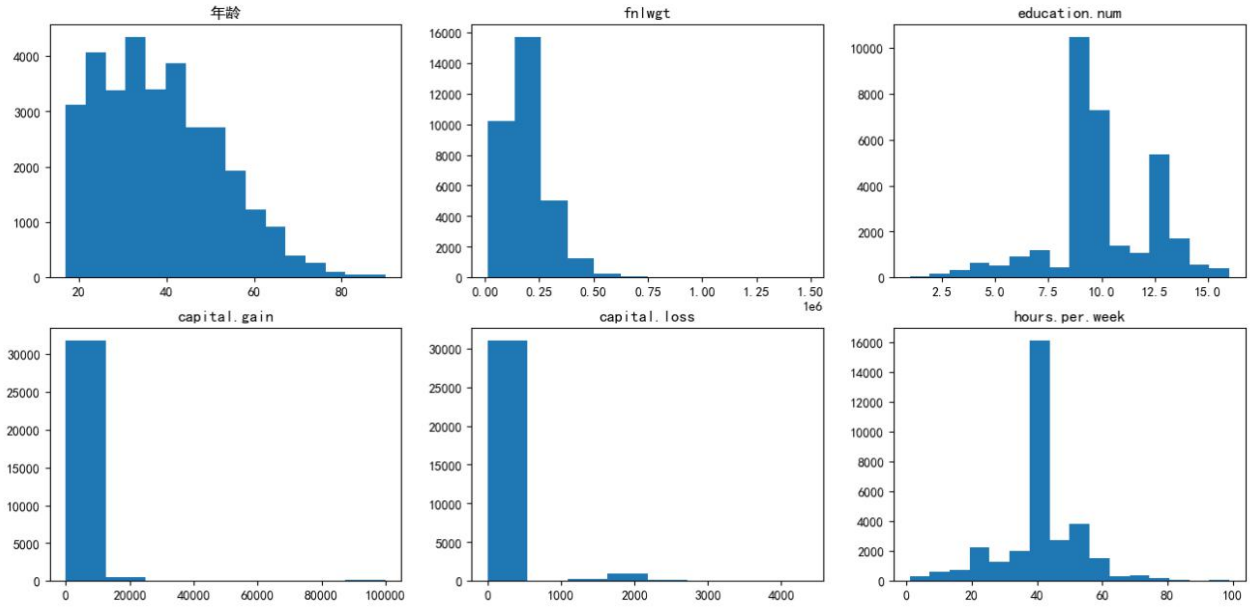
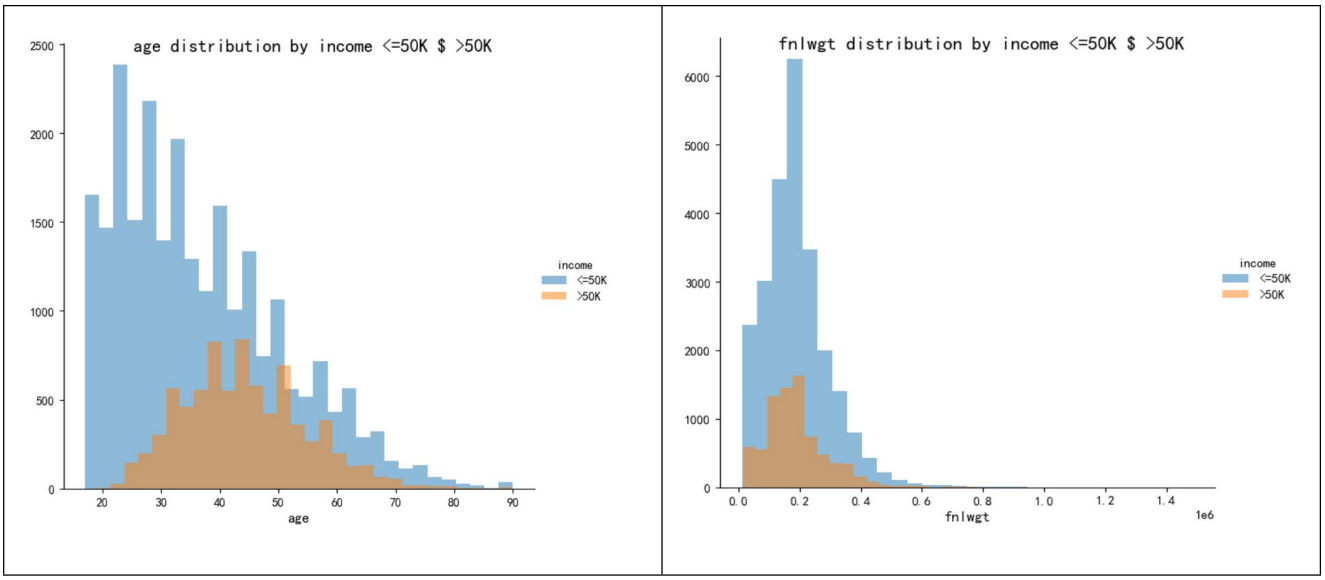


图 3

接著观察在收入分类下自变量的分布情况。我们对年龄、教育时长、权重、每周工作时长四个自变量就行了分析如图 4 示。首先是年龄，可以看出两个收入群体的年龄分布，明显不同。年龄对于收入小于 5 万的人群，年龄分布呈现右偏，收入大于 5 万的人群年龄分布比较对称，高收入群体年龄主要在 40-50 岁之间，低收入群体集中于在 40 岁以下。然后是权重 fnlwgt，可以看到低收入与高收入在 fnlwgt 这个变量趋势一致。接下来是教育时长，可以看出低收入与高收入在教育时长上分布大致一致，但可以观察到教育时长在 8 年以下几乎没有高收入者，随着教育经历的增长，高收入者比例整体上是提升的，这也说明了学习对提高收入的重要性。最后是每周工作时长，可以观察到每周工时在 38 小时以下几乎没有高收入者，数量非常之少，随着工作时长增加，高收入者比例总体上是提升的。



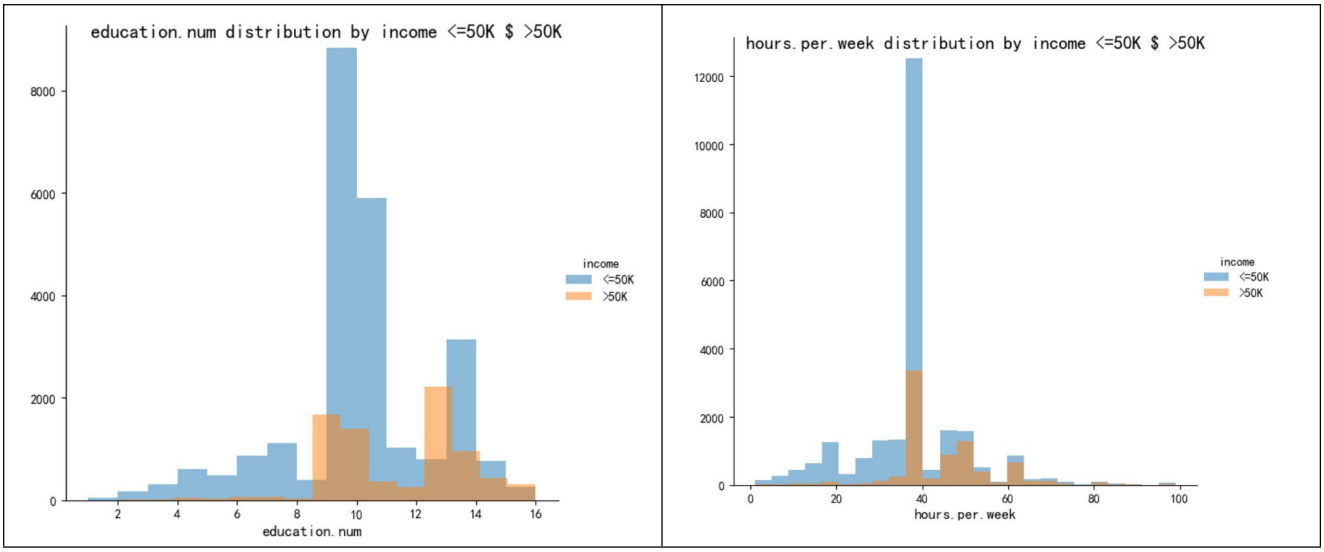


图 4

最后，我们使用饼图来描绘各个变量的相对比例，包括工作类型、婚姻状况、种族、职业、性别和收入等。这有助于我们快速了解数据的整体分布情况，如图 5 所示。在工作类型中，约 70%的人从事私人机构工作。在种族方面，约 85%的人是白人，10%的人是黑人，其余为亚裔和其他种族。性别中，有 75％的数据是来自男性，而 25％来自女性。

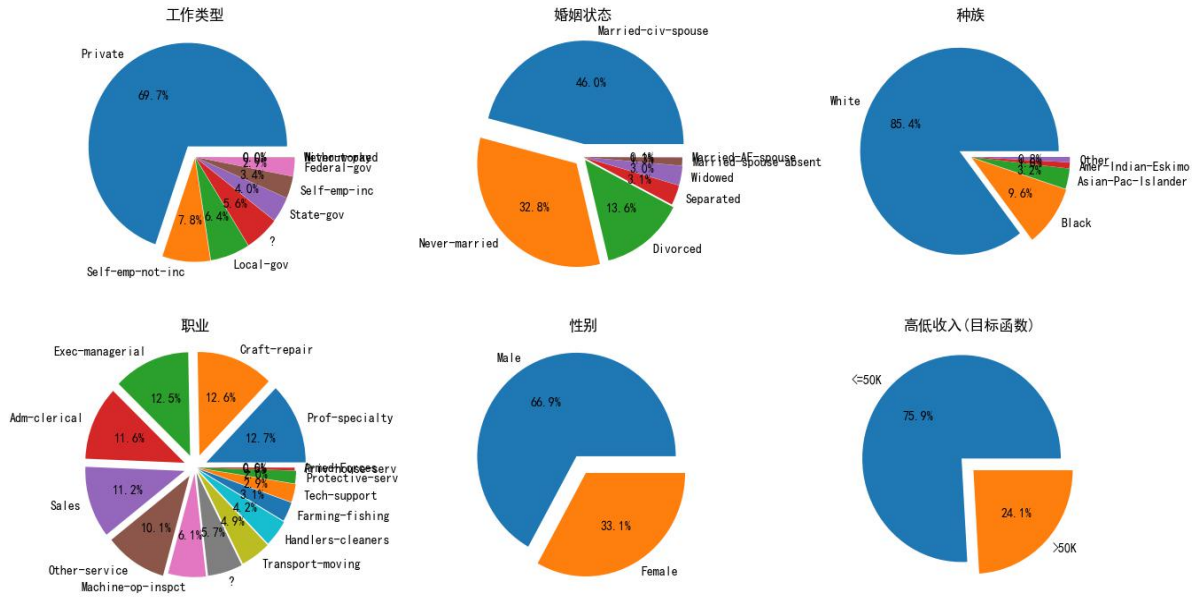


图 5

5 建模分析

预测收入本质是一个监督学习的二分类问题，我们使用了 8 中模型来对收入情况进行预测，分别是逻辑回归、KNN 模型、决策树模型支持向量机模型、朴素贝叶斯模型还有集成算法中的随机森林 Adaboost 模型、GBDT 模型。其次再对数据进行 smote 处理，再比较 smote 处理前后的模型表现。这里面大部分算法（如逻辑回归、决策树等）我们先用网格搜索方法确定最佳参数，然后用最佳参数确定的模型去计算测试集上面的准确率与 ROC 曲线等指标，表 1 和表 2 分别为 smote 前后的数据建模选择的变量相同。观察表 1，在 smote 处理前表现最好的模型分别依次为 GBDT，随机森林，Adaboost，朴素贝叶斯，逻辑回归，SVM，决策树，KNN。从表中可见，经 smote 处理后的数据，大多数模型的 AUC 得到了明显提升，当中随机森林,Adaboost,GBDT 的 AUC 都从 88％上升到 93％左右。准确率方面，有部分的算法准确率提升，部分算法准确率下降。

原数据准确率比较

模型	训练集准确率	测试集准确率	AUC
逻辑回归	82.9%	82.4%	0.877
KNN	85.1%	82.1%	0.858
决策树	84.2%	83.6%	0.873
SVM	82.9%	82.4%	0.876
随机森林	85.2%	83.8%	0.885
Adaboost	83.6%	83.3%	0.883
GBDT	84.4%	83.7%	0.886
朴素贝叶斯	82.9%	82.9%	0.880

表 1

使用 smote 后准确率比较

模型	训练集准确率	测试集准确率	AUC
逻辑回归	80.2%	80.6%	0.885
KNN	87.5%	82.5%	0.894
决策树	86.8%	85.1%	0.921
SVM	79.0%	79.3%	0.881
随机森林	86.6%	86.0%	0.937
Adaboost	84.2%	84.3%	0.924
GBDT	86.0%	86.1%	0.941
朴素贝叶斯	80.7%	80.9%	0.892

表 2

6 总结及建议

根据对人口普查数据集的探索性分析和建模分析，我们得出以下结论：

通过数据可视化方法，我们可以发现数据集中的因变量存在着不平衡的问题，同时数据集中有部分变量存在较多的缺失值。为了解决这些问题，我们采用了 smote 算法处理样本不均衡和众数填补法填补缺失值。

在建模分析中，我们使用了 8 种不同的模型，包括逻辑回归、KNN、决策树、支持向量机、朴素贝叶斯以及集成算法中的随机森林、Adaboost 和 GBDT 模。

利用 AUC 表现，我们发现经过 SMOTE 处理的数据集在大多数模型中表现更好，尤其是随机森林、Adaboost 和 GBDT 模型的 AUC 得分有了明显提升。

综合分析结果，我们可以得出结论：在人口普查数据集中，婚姻状况、家庭关系、教育水平、教育时长、年龄、每周工作时长等变量与收入水平有着一定的关系。同时，在进行建模分析时，我们应该注意样本不平衡和缺失值的问题，并采用适当的方法进行处理。在选择模型时，我们可以考虑使用集成学习算法，如随机森林、Adaboost 和 GBDT 等，以提高预测的准确性。

7 数据链接

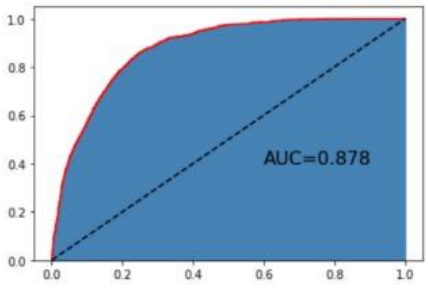
https://pan.baidu.com/s/19Ph2aP2EVL4NUvKSEoFe_g?pwd=cloz

8 小组分工

胡一航：数据集预处理
王宝琪：建模分析和模型解释
姚闰翔：数据描述性分析
阚舒耀：制作 ppt 和现场演示
梁溢笙：内容整理和撰写报告

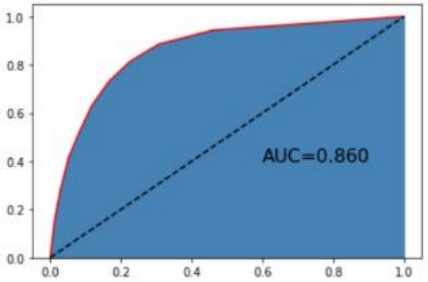
9 附錄

1.逻辑回归



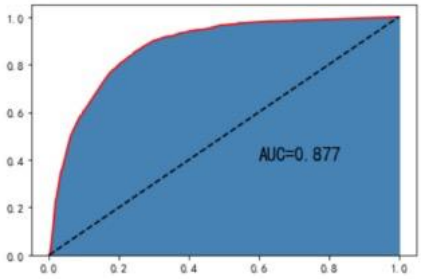
training accury：0.829
testing accury：0.824

2.KNN模型



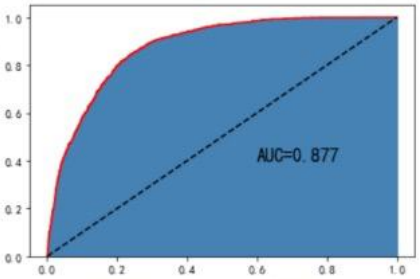
training accury：0.855
testing accury：0.824

3.决策树



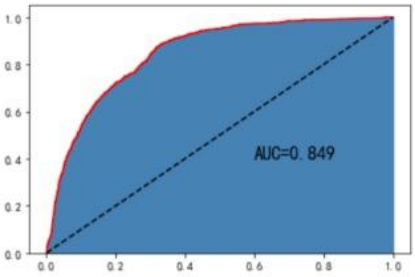
training accury：0.842
testing accury：0.834

4.支持向量机



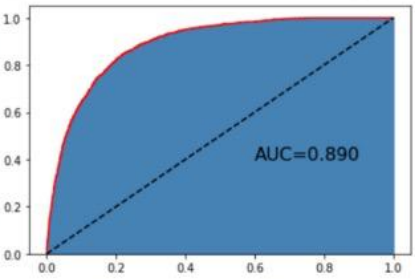
training accury：0.830
testing accury：0.824

5.朴素贝叶斯



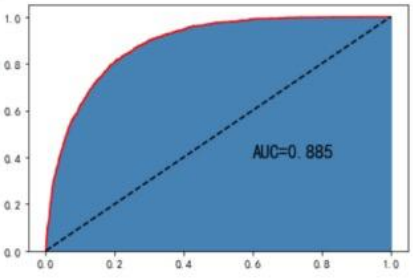
training accury：0.760
testing accury：0.761

6.随机森林



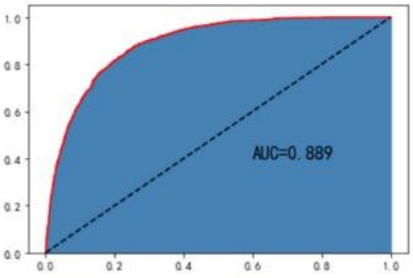
training accury：0.855
testing accury：0.842

7.Adaboost模型



training accury：0.839
testing accury：0.837

8.GBDT模型



training accury：0.847
testing accury：0.840