数据挖掘分析报告

—— 美国个人收入状况分类

学年学期：2022-2023-2

课程名称：数据挖掘

学号：42023017

姓名：常远

专业：金融数学实验班

日期：2023 年 6 月 25 日

第一章 概述

1. 背景介绍

财富和收入不平衡是很多国家政府致力于解决的重要问题之一，贫富差距过大会造成犯罪率升高等社会潜在危害，对于社会的安全运行和健康发展是十分不利的。本项目基于美国1994年的人口普查数据，建立分类模型，正确判别某个人的年收入是否超过5万美元。

2. 数据描述

本项目数据集来源于美国人口调查局，包含了美国 1994 年人口普查的样本数据。此数据集包含了 16281 条数据，每条数据包含 9 个标称属性和 6 个数值属性共 15 个字段属性，其中“income”为二分类的目标属性，即某个人的年收入是否超过 5 万美元，具体属性信息见**表 1**。

表 1 数据集属性信息表

序号	属性名称	数据类型	属性	含义
1	age	int	数值属性	年龄
2	workclass	string	标称属性，9 类	工作状态
3	fnlwgt	float	数值属性	最终权重
4	education	string	标称属性，16 类	最高学历
5	education_num	int	数值属性	学习年数
6	marital_status	string	标称属性，7 类	婚姻状态
7	occupation	string	标称属性，15 类	职业类型
8	relationship	string	标称属性，6 类	家庭状态
9	race	string	标称属性，5 类	种族
10	sex	string	标称属性，2 类	性别
11	capital_gain	float	数值属性	资本收入
12	capital_loss	float	数值属性	资本支出
13	hours_per_week	int	数值属性	每周工作小时数
14	native_country	string	标称属性，41 类	原籍国籍
15	income (target)	string	标称属性，2 类	年收入是否 5 万美元

第二章 探索性数据分析

1. 数据质量分析

- 特殊字符

  通过部分数据的预览，数据集的列名和字段值存在空格、“ ？”等特殊字符。

- 缺失值

  数据集总体上缺失属性较少且缺失例较低，共有四个属性存在缺失，分别为“workclass”（940，5.77%），“occupation”（944，5.80%），“hours_per_week”（3，0.02%），“native_country”（298，1.83%）， 缺失值计数条形图见**图 1**。

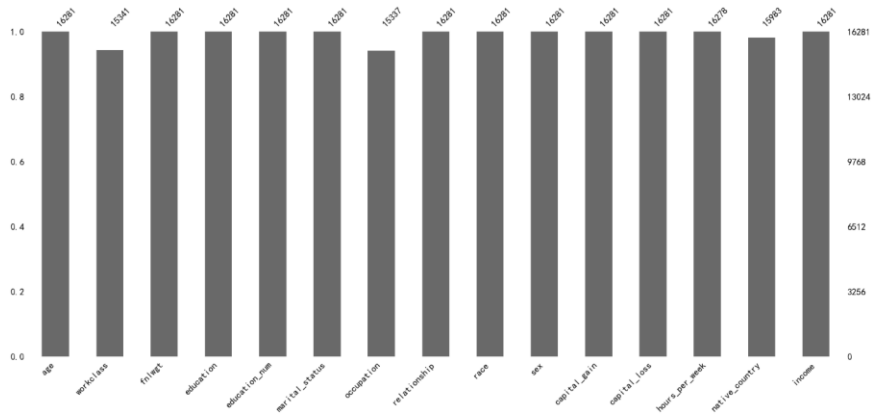

图 1 数据缺失值计数条形图

- 重复值：数据集共存在 5 条属性对应值完全相同的数据。
- 异常值

  数据集数值属性的描述性统计见**表 2** 和箱线图见**图 2**。总体上无明显异常值，其中数据集样本的年龄较为年轻，总体均值在 38 左右且第二三分位数间距较大，但均在合理范围之内；“fnlwgt”存在少数异常值在 1500000 在左右；资本收入存在较为明显两极分化，资本支出也存在两极分化；每周工作时间大多数在 50 左右，存在较为连续的异常值。

表 2 数据集描述性统计表

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
count	16281.000000	1.628100e+04	16281.000000	16281.000000	16281.000000	16278.000000
mean	38.667834	1.904670e+05	10.080646	1078.198268	82.928076	40.463878
std	13.655348	1.058821e+05	2.581534	7368.799360	392.992360	12.436656
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.179630e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.791370e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.369400e+05	13.000000	0.000000	0.000000	45.000000
max	90.000000	1.455435e+06	16.000000	99999.000000	4356.000000	99.000000

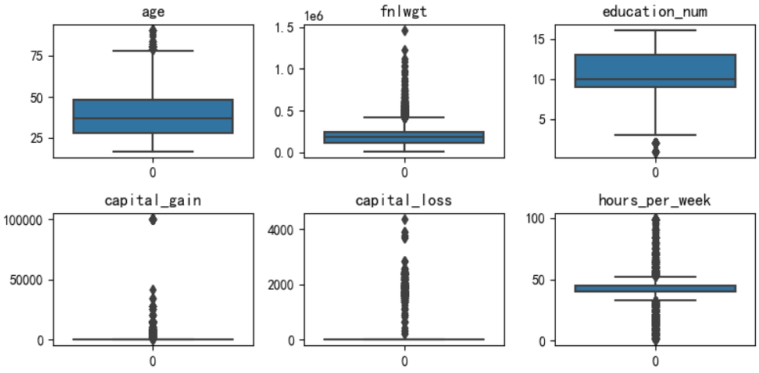

图 2 数值属性箱线图

2. 数据特征分析

2.1 目标特征分析

总体上数据的收入情况存在较为明显的不平衡，其中每年收入低于 5 万美金的占比高达 75.91%，但是基本上符合社会财富分布呈现右偏的规律，目标属性“income”的分布和结构图见**图 3**。

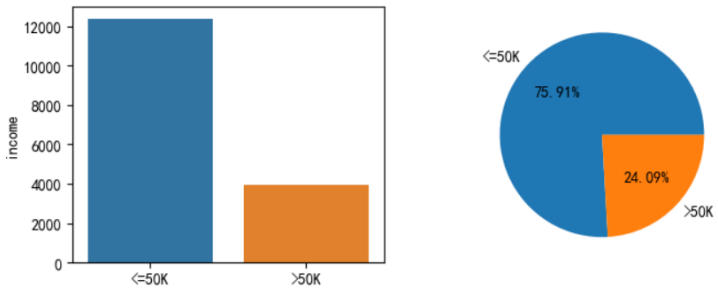

图 3 目标属性分布和结构图

2.2 变量特征分析

本项目将变量进行分类为连续变量和类别变量通过绘图各变量的分布，从而进行数据的特征分析。

- 年龄（age）

总体上数据集的年龄分布呈现右偏态，即人口普查样本较为年轻。按目标属性分类后的分布表明，年收入不超过 5 万美元集中在 40 岁以下，第一二分位数间距较小，分布呈现明显的右偏态，即年轻人表现出更明显的年收入低于 5 万美元；而年收入超过 5 万美元的样本分布较为匀称但存在明显峰度，即年收入超过的 5 万美元更集中均匀分布在 42 岁左右，年龄的总体、目标分类的分布和箱线图见**图 4**。

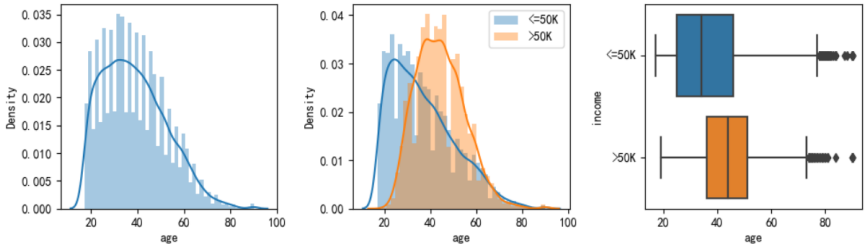

图 4 年龄（age）的分布和箱线图

• 最终权重（fnlwgt）

最终权重总体和按目标分类后的数据呈现出很明显的分布高度一致，分布差异很小，存在一定偏态，经对数变化后仍呈现出较为相似的分布[1]，故考虑直接删除此特征、标准化或分箱，最终权重的总体、目标分类的分布和箱线图见**图 5**。

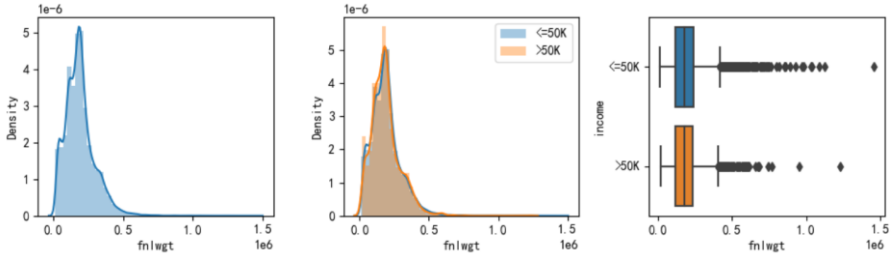

图 5 最终权重（fnlwgt）的分布和箱线图

• 学习年数（education_num）

总体上学习年数更长的群体的明显呈现出更好的收入情况，不同目标类别的学习年数总体上呈现出较明显的分布差异，学习年数的总体、目标分类的分布和箱线图见**图 6**。

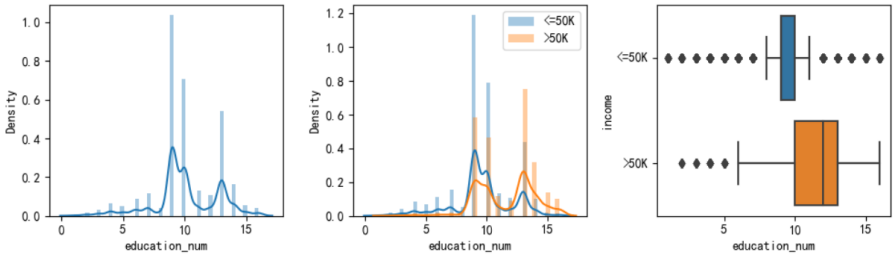

图 6 学习年数（education_num）分布和箱线图

• 资本收入 （capital_gain）

总体上资本收入分布存在极明显的峰度，大多数个体的资本收入为 0，且在不同群体中均有着较明显的两极分化，年收入不超过 5 万美元的人资本收入集中在 1 万美元以内，而年收入超过 5 万美元的人资本收入更集中分布在 2.5 万美元。考虑需要标准化处理或者分箱处理，资本收入的总体、目标分类的分布和箱线图见**图 7**。

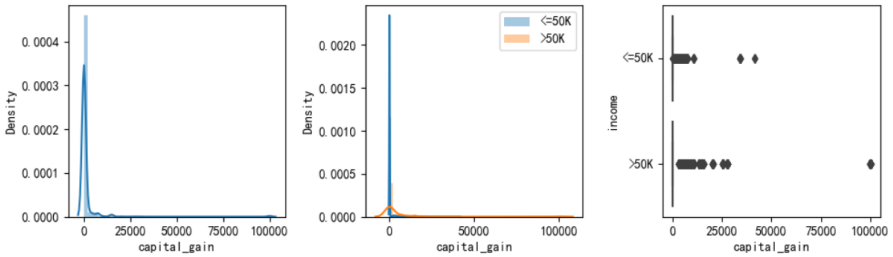

图 7 资本收入（capital_gain）分布和箱线图

• 资本支出 （capital_loss）

同资本收入类似，总体上资本支出分布存在极明显的峰度，不同点在于年收入不超过 5 万美元的人有着更大范围的高资本支出。考虑需

[1] 对数变化后分布图见代码部分。

要标准化处理，资本支出的总体、目标分类的分布和箱线图见**图 8**。

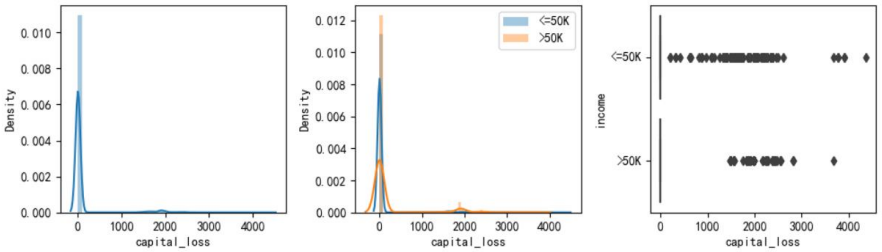

图 8 资本支出（capital_loss）分布和箱线图

• 每周工作小时数 （hours_per_week）

总体上，收入较高的个人每周有着更长时间的工作时间，即收入情况基本与工作时间正比，每周工作小时数的总体、目标分类的分布和箱线图见**图 9**。

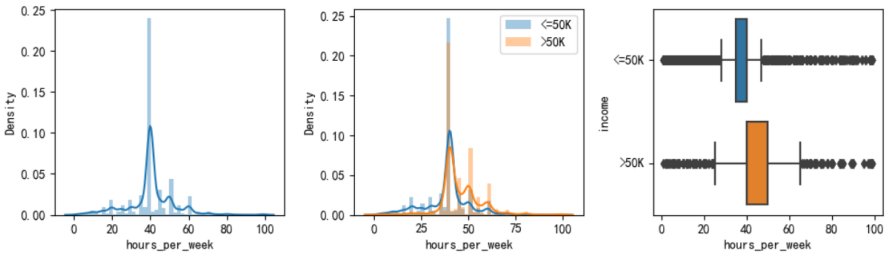

图 9 每周工作小时数（hours_per_week）分布和箱线图

• 工作状态（workclass）

总体上，工作状态为私密的个人最多，除没有工作和无报酬外分布较为均衡，特别的是一些职业类型之间可能存在某种关系例如州政府、当地政府和联邦政府等，考虑将相似类别进行合并再分类，工作状态分布见**图 10**。

• 教育（education）

总体上，年收入超过 5 万美元的人教育水平主要在高中毕业、上过大学和获得学士学位，与年收入未超过5万美元的人集中分布类似，但是收入水平较好的人教育较少人低于高中毕业以下水平，也即是收入情况较好的人往往有着较高的教育水平，考虑将高中及以下的教育水平进行合并，教育水平分布见**图 10**。

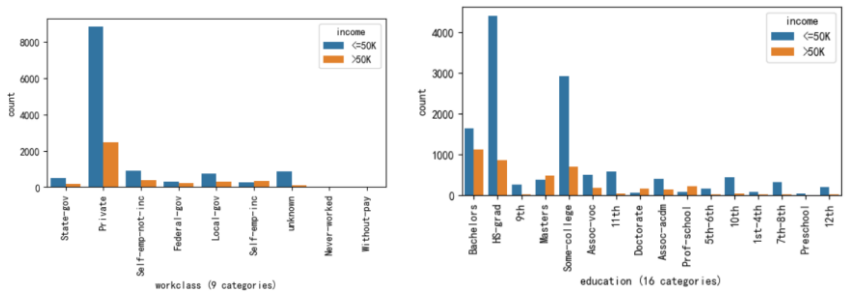

图 10 工作状态（workclass）和教育（education）分布图

• 婚姻状态（marital_status）

总体上，年收入超过 5 万美元的绝大多数是已婚公民配偶，少量为从未结婚和离婚人士，分布较为集中；而年收入未超过5万美元的主

要为从未结婚、已婚公民配偶和离婚人士，少数为分居等。考虑将部分类别进行合并，婚姻状态分布见**图 11**。

• 职业类型（occupation）

总体上，收入情况较好的个体主要集中在管理相关、教授、工艺维修、销售类型职位，相较于收入情况较差的个体，其分布更集中，处在清洁工等可替代性较强的职位较少，职业类型的分布见**图 11**。

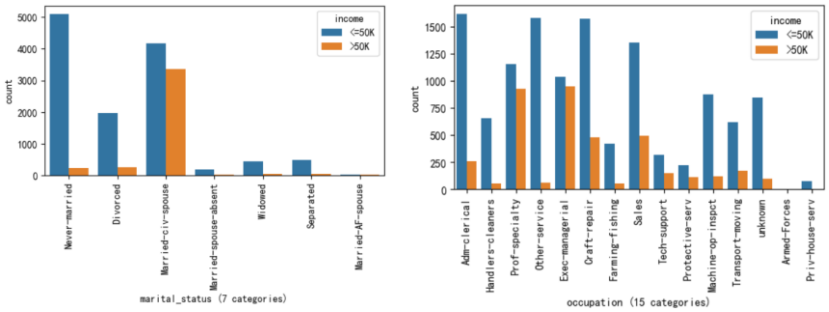

图 11 婚姻状态（marital_status）和职业类型（occupation）分布图

• 家庭状态（relationship）

家庭状态分布见**图 12**，结合婚姻状态分布图表明，大多数已婚家庭有着明显较好的收入情况，同时男性往往是多数家收入支柱；而收入情况欠佳总体上分布在较为广泛，最多的是未结婚和已婚公民配偶。

• 种族（race）

总体上，白人在各类收入情况中都占据压倒性优势，但是相较于年收入未超过 5 万美元，年收入超过 5 万美元分布呈现更集中，这有可能与白人在美国所处的历史社会地位有关，种族分布见**图 12**。

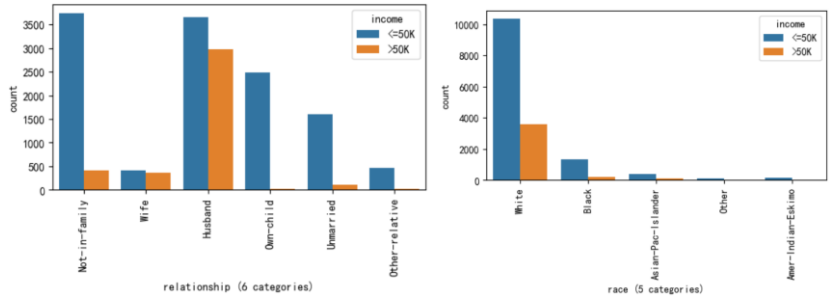

图 12 家庭状态（relationship）和种族（race）分布图

• 性别（sex）

男性在两类收入情况中均有着较高的占比，原因在于数据集本身男女性别存在不平衡的情况，此外发现数据集中女性总体上收入情况较男性差，性别分布图**见 13**。

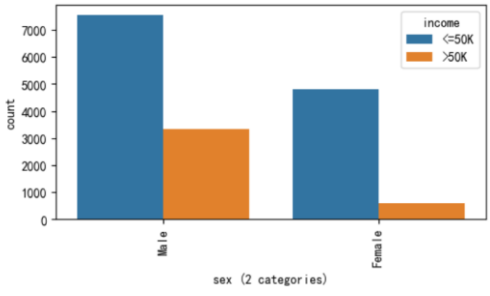

图 13 性别（sex）分布图

• 原籍国籍（native_country）

总体上数据集的原籍国籍压倒性分布在美国，其原因主要是数据集来源于美国人口普查，其类别数较多且剩下分布几乎很少，故考虑将美国以外的国籍进行合并后再分类，原始国籍分布见**图 14**。

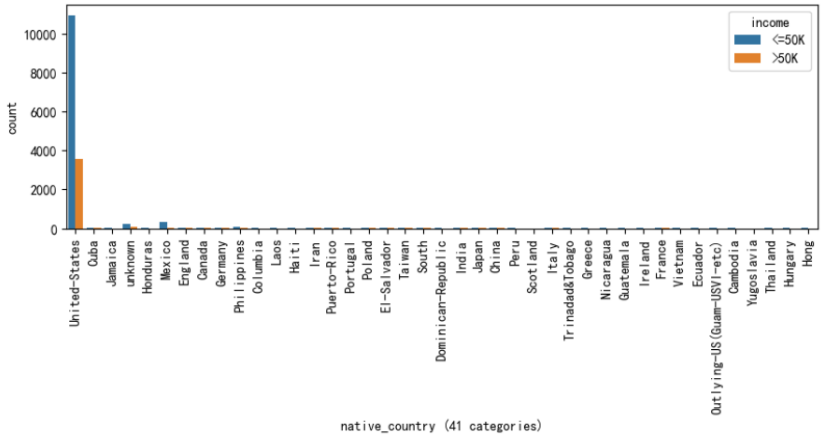

图 14 原籍国籍（native_country）分布图

2.3 相关性分析

本项目对于目标属性进行了二元编码，计算了变量间和变量与目标间相关系数，结果表明总体上有着更长学习年数、每周更长的工作时间和更高的资本收入的人更有可能有着更好的收入情况，具体结果如下，相关系数热力图见**图 15**。

• 学习年数与收入情况较弱的正相关关系，相关系数为 0.33。

• 年龄与收入情况较弱的正相关关系，相关系数为 0.23。

• 资本收入与收入情况较弱的正相关关系，相关系数为 0.22。

• 每周工作小时数与收入情况较弱的正相关关系，相关系数为 0.22。

• 变量间总体呈现不明显的相关性关系，其中年龄和每周工作小时数与学习年数存在微弱的正相关关系。

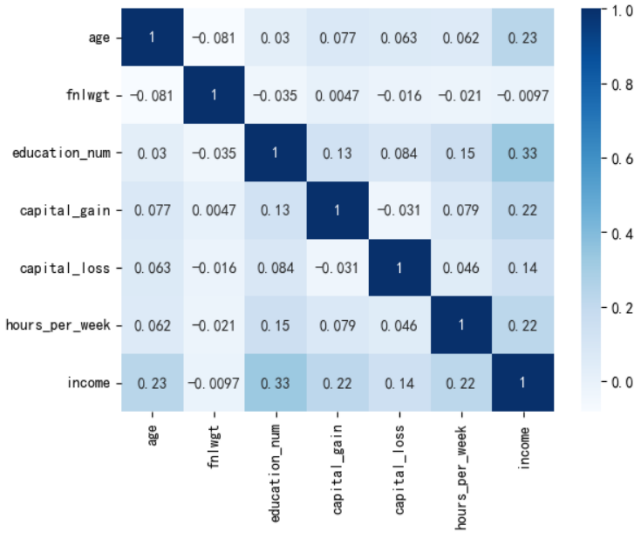

图 15 相关系数热力图

2.4 其他探索性发现

- 教育（education）与学习年数（education_num）

在初步探索中发现，教育属性和学习年数属性均有着 16 个不同属性值，进一步地按照教育属性对数据集进行分组聚合[2]，发现教育属性和学习年数属性有着一一对应的映射关系，在相关性分析中也发现学习年数与收入有着最强的正相关关系，故考虑删除教育属性。

第三章 数据预处理

1. 数据清洗

- 特殊字符

本项目将属性值为“ ?”替换为空值，并清除了数据集列名和属性值为字符串的前后空格，从而进行特殊字符的清洗。

- 缺失值

经特殊字符清洗后，数据集总体缺失比例较低，共有四个属性存在缺失，分别是“workclass” (940, 5.77%), “occupation” (944, 5.80%)，“hours_per_week”(3, 0.02%), “native_country” (298, 1.83%)。相较之下缺失最少的每周工作小时采用均值填充，其余三个类别特征采用“unknown”填充。

- 重复值

数据集共存在 5 条所有属性对应值完全相同的数据，重复数量较少且重复数据的特征值完全一致，可能是数据的重复录入，故直接将重复数据进行删除。

- 异常值

在数据质量分析中未发现不合理的异常值，故在此不做处理。

2. 特征变换

2.1 标称属性

- 收入（income）：本项目对收入目标数值变量进行二元属性编码，其中年收入<=50K 美元的值映射为 0，年收入收入>50K 美元的值映射为 1。
- 工作状态（workclass）：据工作状态的属性值分析，本文中通过属性值类别重构，即将州政府（“State-gov”），联邦政府（“Federal-gov”）和当地政府（“Local-gov”）属性值映射为“gov”，将私密（“Private”）映射为“unknown”，将从未工作过（“Never-worked”）和无报酬（“Without-pay”）映射为“not_work”，进一步地，对其进行独热编码变化为数值属性。
- 婚姻状态（marital_status）：本文对婚姻状态进行独热编码。
- 职业类型（occupation）：本文对职业类型进行独热编码。
- 家庭状态（relationship）：本文对家庭状态进行独热编码。

[2] 分组聚合结果表见代码部分。

- 种族（race）：本文对种族进行独热编码。
- 性别（sex）：本文对性别进行独热编码。
- 原籍国籍（native_country）：根据原籍国籍属性的特征分布，对统计个数前二的其余属性值进行映射，即将“United-States”和“Mexico”保留为原值，其他国籍映射为“unknown”，进一步地对其进行独热编码转化为数值属性。

2.2 连续数值属性

对于很多分类算法，连续属性离散化有助于处理异常值和高度偏斜的变量，从而提高模型鲁棒性，与此同时起到了简化逻辑回归、决策树等模型的复杂性，从而降低了模型过拟合的风险，常见离散化的方法包括分箱方法，基于聚类分析的方法。在本项目中对于部分连续属性分别采用了不同的离散化方法，最后使用独热编码进行数值变化。

- 年龄（age）：本项目基于常见的年龄划分标准将年龄划分为 0-20, 20-30，30-40，40-60 和>60 这五个区间范围。
- 最终权重（fnlwgt）：该特征采用了一种有监督的分箱方法——决策树分箱，最终划分为了<40000，40000-80000，80000-190000 ， 190000-210000 ， 210000-280000 和>280000 这五个区间范围。
- 资本收入（capital_gain）：同最终权重属性方法一致，资本收入划分为<5000 和 5000-100000 这两个区间范围。
- 净资本（net_ capital）：同最终权重属性方法一致，净资本划分为<5000 和 5000-100000 这两个区间范围。
- 每周小时数（hours_per_week）：同最终权重属性方法一致，每周小时数被划分为<33，33-40，40-42 和 42-100 这四个区间范围。

以上连续特征进行分箱后并对其分别进行了独热编码从而完成类别变化，而对于资本支出（capital_loss）属性，在数据探索性分析中发现其存在较为明显的峰度，由于后续某些模型需要进行距离计算等，本项目中对其进行了标准化处理；对于学习年数（education_num）有序数值属性，分布较为均匀极差也较小，故未作处理。

3. 特征构造

通过对比分析变量间的关系，通过某些特征构造了如下新特征。

- 净资本（capital_net）

  capital_net = capital_gain – capital_loss

4. 特征筛选

常见特征筛选方法主要包含过滤法、包装法和嵌入式，相较于过滤法和包装法，嵌入式特征选择能够与数据挖掘模型的构建完全地融合在一起，同时不需耗时的迭代搜索过程而具有更高的效率，在本项目中分

别尝试了多种方法进行比较，最终采用了基于随机森林模型嵌入式方法对特征进行选择。最终共筛选出 18 个特征进行模型训练，筛选后的特征结果见**表 3**。

表 3 筛选后特征表

属性	属性名称
education_num	学习年数
capital_loss	资本支出
workclass:unknown	工作状态：未知（私密）
marital_status:Married-civ-spouse	婚姻状态：已婚公民配偶
marital_status:Never-married	婚姻状态：从未结婚过
occupation:Exec-managerial	职业类型：执行管理
occupation:Prof-specialty	职业类型：教授
relationship:Husband	家庭状态：丈夫
age:20-30	年龄：20-30
age:40-60	年龄：40-60
fnlwgt:80000-190000	最终权重：80000-190000
fnlwgt:210000-280000	最终权重：210000-280000
fnlwgt:>280000	最终权重：大于 280000
capital_net:<5000	净资本：小于 5000
capital_net:5000-100000	净资本：5000-100000
capital_gain:<5000	资本收入：小于 5000
capital_gain:5000-100000	资本收入：5000-100000
hours_per_week:42-100	每周工作小时数：42-100

第四章 模型训练

本项目训练集测试集，分别尝试使用了多种分类器对模型进行训练并进行比较，同时利用堆叠模型对模型进行了集成。

1. Logistical Regression

由于在数据预处理过程中引入了大量的哑变量，其中存在一定程度相关性较高的变量[3]例如男性和女性等，与朴素贝叶斯的条件独立性假设要求相比，逻辑回归不需要考虑样本的相关性，同时逻辑回归在模型结果上有着较好的解释性。本项目中训练逻辑回归模型参数设置如下：

- penalty：正则项参数默认 L2 范式。
- max_iter：在 10、100、1000、10000 中基于准确率进行网格搜索，最终选择最大迭代次数 1000 次。

2. Decision Tree Classifier

相较于其他模型，决策树模型有着很强的可解释性，同时决策树的

[3] 预处理后所有变量的相关系数热力图见代码部分。

学习过程往往有着较好的效果，但也存在着较大过拟合风险。本项目中训练决策树分类模型参数设置如下：

- criterion：以准确率作为指标，在信息熵和基尼系数进行搜索，最终选择基尼系数作为特征选择的超参数。
- max_depth：以准确率作为指标，在 1-10 之间进行网格搜索，最终选择 9 为树最大深度。
- max_leaf_nodes: 以准确率作为指标，在 $10\text{-}10^5$ 之间进行网格搜索，最终选择 100 作为树最大叶子节点数量。

3. KNN Classifier

KNN 算法思想简单，对数据本身也没有朴素贝叶斯较强的假设要求，同时对于异常点也较不敏感，但其解释性与决策树和逻辑回归相比较差，本项目中训练 k 近邻分类模型参数设置如下：

- K：以误差率作为指标，在 1-40 的范围内进行网格搜索，结合训练模型的复杂性权衡考虑选择了 20 作为邻居数量的超参数。

4. Support Vector Classifier

支持向量机的学习策略即最大化间隔，其最终决策边界只有少数支持向量决定，泛化性能较好且具有一定的鲁棒性，但其对核函数、参数较为敏感而存在一定局限性。本项目支持向量分类模型参数设置如下：

- kernel：基于准确率指标的网格搜索结果表明高斯核（“rbf”）拟合效果最佳，但是可能存在较大过拟合的风险。考虑到本项目中输入特征经筛选后共 18 个，特征维度仍稍高，样本倾向于线性可分，同时与训练时间资源的权衡，尝试使用线性核（“linear”）进行模型训练模型，结果表明线性核也表现出较好的性能。因此在最终模型中考虑简化模型的复杂度和提高模型泛化能力，最终选择了线性核（“linear”）。

5. ANNs (Multi-Layer Perceptron Classifier)

神经网络相较传统机器学习算法有着很强的学习能力和构建非线性的复杂关系的能力，同时其对输入变量不加任何限制，但存在明显解释性很差，训练耗费时间长等问题。本项目中训练多层感知分类模型参数设置如下：

- hidden_layer_sizes：（100，100），即两层 100 个神经元的隐藏层。
- activation：激活函数选择 sigmoid 函数即“logistic”。
- solver：由于数据集规模较小，故选择“lbfgs”作为权重优化方法参数，其相较有着更快的收敛速度和更好的表现。
- early_stopping：为减缓过拟合问题采用早停法，故设置“True”。
- learning_rate：学习率更新方法选择适应性更新方法“adaptive”。

6. Stacking

模型集成通过构造多个模型的组合，往往能够建立一个性能更有的分类器。本项目中最终尝试以上模型使用堆叠进行集成，其部分参数设

置如下：

- final_estimator：本项目中选择逻辑回归模型作为基模型，即 LogisticRegression()。
- cv：模型选择 10 折交叉验证。

第五章 模型讨论

1. 模型比较

本项目中每个模型分别计算 10 折交叉验证的平均准确率，由于在第二章数据探索性分析中发现数据集存在较为明显的收入分类分布不平衡，故绘制了受试者工作特征曲线（ROC）并计算 AUC 进行模型性能比较，其中模型单次训练的混淆矩阵见代码部分，模型的 10 折交叉验证平均准确率见**表 4**，ROC 比较曲线见**图 16**。

表 4 10 折交叉验证结果

模型	Mean Accuracy
Logistic Regression	0.852
Decision Tree Classifier	0.848
KNN Classifier	0.844
Support Vector Classifier	0.853
Multi-Layer Perceptron Classifier	0.852
Stacking	0.854

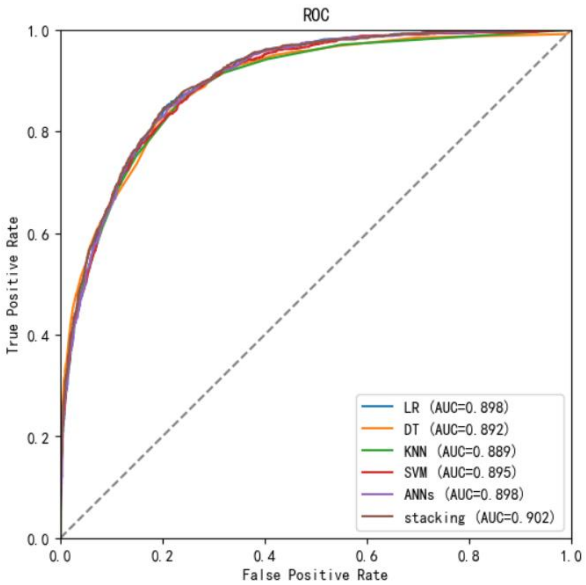

图 16 受试者工作特征曲线图（ROC）

10 折交叉验证的平均准确率结果表明，单个模型总体上支持向量机分类模型、逻辑回归模型和多层感知机分类模型性能最佳，平均准确率高达 0.852 左右，其次是决策树和 K 近邻分类模型稍差。经堆叠模型集成后的模型平均准确率最高，比任何一个单模的性能表现都更佳。ROC 曲线和 AUC 值也表明，总体上模型性能均表现较佳，其中堆叠后的模型

性能表现最佳，单模中逻辑回归模型和多层感知机分类模型性能最好，K 近邻性能表现最差，主要与参数 K 选择上模型复杂性权衡有关。

2. Logistical Regression

逻辑回归模型具有很强的可解释性，其变量系数反映了样本x作为正例的相对可能性。本项目对逻辑回归模型的变量系数按降序排序后绘制了特征重要性条形图见**图 17**。

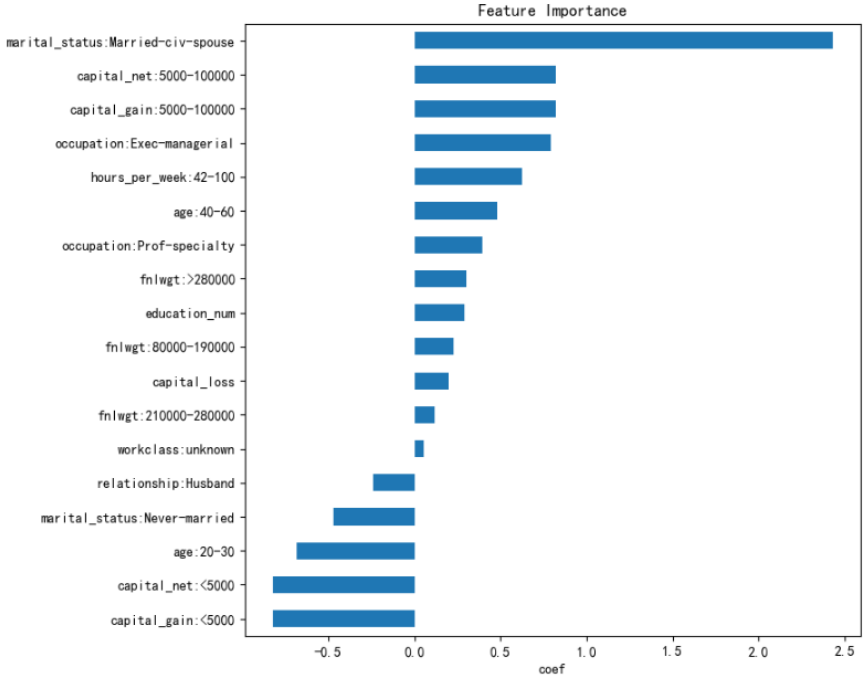

图 17 逻辑回归模型特征重要性条形图（排序表见代码部分）

逻辑规格模型的变量系数取绝对值较高的特征均有着较为合理的解释意义。婚姻状态为已婚公民配偶（marital_status: married-civ-spouse）与个人年收入是否超过 5 万美元有着很显著的正相关关系，资本收入在 5000-10000 范围（capital_gain:5000-100000）、净资本在 5000-10000 范围（capital_net:5000-100000）、职业类型为执行管理（occupation:Exec-managerial）、每周工作时长大于 42 小时（hours_per_week:42-100）和年龄段为中年（age:40-60）与个人较好收入情况有着明显的正相关关系，此外资本收入小于 5000（capital_gain:<5000）、净资本小于 5000（capital_net:<5000）、年龄阶段处在早期青年（age:20-30）和婚姻状态为从未结过婚（marital_status: Never-married）与个人年收入是否超过 5 万美元有着明显的负相关关系。

3. Decision Tree Classifier

决策树模型其可以看作 if-then 规则的集合，路径上内部结点的特征对应着规则的条件，叶节点的类对应于规则的结论，通过可视化决策树可以很好的解释决策树模型分类规则，本项目决策树模型部分可视化见**图 18**。决策树通过特征选择的标准计算特征重要性，本项目中决策树特征重要性条形图见**图 19**。

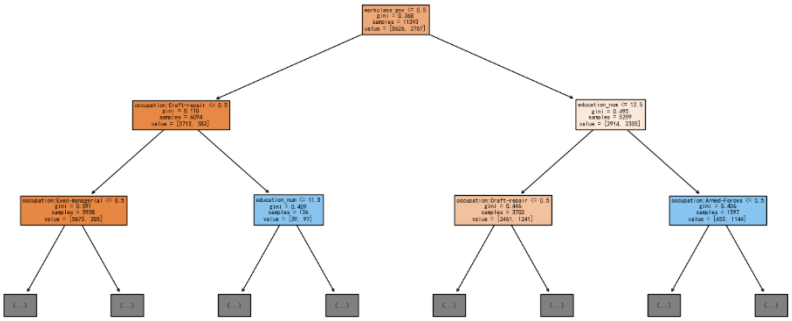

图 18 决策树模型可视化图（部分，详见代码部分）

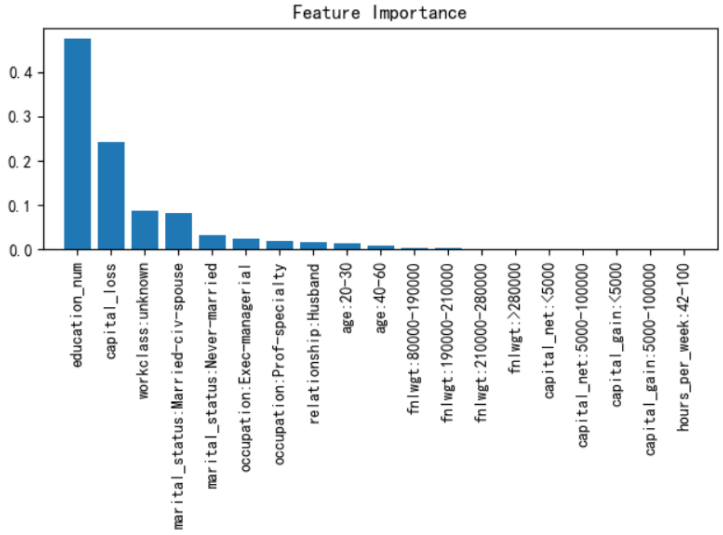

图 19 决策树模型特征重要性条形图（排序表见代码部分）

决策树模型结果表明，学习年数（education_num）、资本支出（capital_loss）、工作状态为未知（workclass:unknown）、婚姻状态为已婚公民配偶（marital_status: married-civ-spouse）、婚姻状态为从未结过婚（marital_status: Never-married）和年龄的两个阶段（age:20-30, age:40-60）对模型有着较大的影响，与逻辑回归模型解释有着一定程度重叠，总体上其解释基本较为合理。此外总体上存在一定数量的特征重要性为 0 的特征，即说明重要性靠前的特征已经能够较好的划分个人收入情况了，在预处理部分未来可以进行深入实验比较特征处理和选择。

第六章 总结

本项目基于美国 1994 年人口普查数据进行了数据挖掘，实验过程中尝试了多种数据预处理方法，比较了数据集在多种机器学习分类模型下性能表现，最后对可解释性较强的模型进行了模型解释和讨论。总体上项目中模型均表现出较佳的性能，但是在数据预处理的特征处理部分仍存在一些地方值得去深入学习探索，例如连续特征分箱后采用独热编码忽略了数值间有序的信息，标称属性进行再分类时分类标准的选择等。