**人工智能训练师（三级）操作技能考核**

**试题单**

准考证号：

试题代码：

试题名称：信用评分模型数据清洗和标注流程设计

考核时间：20min

**1.**场地设备要求

（1）人工智能训练师主机 1 台；   
（2）Python 编译环境；   
（3）Finance数据集。

**2.**工作任务

互联网金融飞速发展，使得个人金融理财变得越来越容易。而其中信用评分技术是一种对贷款申请人（信用卡申请人）做风险评估分值的统计模型，可以根据客户提供的资料、客户的历史数据、第三方平台数据（芝麻分、京东、微信等），对客户的信用进行评估。现要求根据提供的Finance数据集，选择合适的特征，开发一个申请的评分模型，对未来一段时间内借贷人出现违约的概率进行预测，对客户信用进行评估打分。提供的数据集样本数据一共15000条，10个自变量，1个因变量（SeriousDlqin2yrs）。在开发评分模型之前，首先要对数据进行数据清洗，请补全2.1.3.ipynb代码完成下面的数据预处理任务，并设计一套标注流程规范：  
（1）正确加载数据集，并显示前五行的数据；  
（2）检查数据集中的异常值并处理异常值，使用箱线图检测异常值，使用IQR方法处理异常值；  
设置图像的尺寸为12英寸宽和8英寸高；  
将画布分成3行4列，总共可以容纳12个子图；  
（3）检查数据集中的重复值并删除所有重复值，并记录删除的行数；  
（4）对数据进行归一化处理；  
（5）创建新的特征IncomeToDebtRatio，MonthlyIncome，并添加到数据集中；  
（6）将SeriousDlqin2yrs设为目标变量并标注；  
（7）对数据进行划分；  
（8）保存处理后的数据，并命名为：2.1.3\_cleaned\_data.csv，保存到考生文件夹；  
（9）制定数据清洗和特征工程规范，将答案写到答题卷文件中，答题卷文件命名为“2.1.3.docx”，保存到考生文件夹；  
（10）将以上代码以及运行结果，以html格式保存并命名为2.1.3.html，保存到考生文件夹，考生文件夹命名为“准考证号+身份证后6位”。

**3.**技能要求

（1）能进行数据清洗和特征工程，包括缺失值处理、异常值处理、数据标准化和特征创建；   
（2）能使用Python编程实现上述数据预处理和特征工程步骤。

**4.**质量指标

（1）数据预处理步骤完整，方法选择合理。   
（2）代码实现正确，结果符合预期

**人工智能训练师（三级）操作技能考核**

**试题评分表**

准考证号：

试题代码：

试题名称：信用评分模型数据清洗和标注流程设计

考核时间：20min

测量分评分表

| 细则编号 | 配分 | 评分细则描述 | 规定或  标称值 | 结果或  实际值 | 得分 |
| --- | --- | --- | --- | --- | --- |
| M1 | 1 | 数据集正确加载，显示前五行的数据得1分；报错或无显示不得分； | 根据数据 |  |  |
| M2 | 2 | 使用IQR方法处理和移除异常值正确得2分； | 根据数据 |  |  |
| M3 | 1 | 正确处理重复值得1分； | 根据数据 |  |  |
| M4 | 1 | 对数据进行归一化处理得1分； | 根据数据 |  |  |
| M5 | 1 | 设定目标变量正确得1分； | 根据数据 |  |  |
| M6 | 2 | 定义特征正确得1分，标注目标变量正确得1分，总计得2分； | 根据数据 |  |  |
| M7 | 1 | 数据划分正确得1分； | 根据数据 |  |  |
| M8 | 1 | 保存处理后的数据得1分； | 根据数据 |  |  |
| M9 | 3 | 回答数据清洗规范：每回答正确1个规范点，得1分，最高得3分； | 根据数据 |  |  |
| M10 | 2 | 回答特征工程规范：每回答正确1个规范点，得1分，最高得2分； | 根据数据 |  |  |
| 合计配分 | 15 | 合计得分 | |  |  |

**参考答案**：

（1）2.1.3.ipynb：

import pandas as pd

# 加载数据

data = pd.read\_csv('finance数据集.csv')

# 显示前五行的数据

data.head()

import matplotlib.pyplot as plt

import seaborn as sns

# 设置图像尺寸

plt.figure(figsize=(12, 8))

# 识别数值列用于箱线图

numeric\_cols = data.select\_dtypes(include=['float64', 'int64']).columns

# 创建箱线图

for i, col in enumerate(numeric\_cols, 1):

plt.subplot(3, 4, i)

sns.boxplot(x=data[col])

plt.title(col)

plt.tight\_layout()

plt.show()

# 使用IQR处理异常值

Q1 = data[numeric\_cols].quantile(0.25)

Q3 = data[numeric\_cols].quantile(0.75)

IQR = Q3 - Q1

# 移除异常值

data\_cleaned = data[~((data[numeric\_cols] < (Q1 - 1.5 \* IQR)) | (data[numeric\_cols] > (Q3 + 1.5 \* IQR))).any(axis=1)]

# 检查重复值

duplicates = data\_cleaned.duplicated()

num\_duplicates = duplicates.sum()

data\_cleaned = data\_cleaned[~duplicates]

print(f'删除的重复行数: {num\_duplicates}')

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

data\_cleaned[numeric\_cols] = scaler.fit\_transform(data\_cleaned[numeric\_cols])

# 将SeriousDlqin2yrs设为目标变量

target\_variable = 'SeriousDlqin2yrs'

from sklearn.model\_selection import train\_test\_split

# 定义特征和目标

X = data\_cleaned.drop(columns=[target\_variable])

y = data\_cleaned[target\_variable]

# 划分数据

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# 显示划分后的数据形状

print(f'训练数据形状: {X\_train.shape}')

print(f'测试数据形状: {X\_test.shape}')

# 保存清洗后的数据到CSV

cleaned\_file\_path = '2.1.3\_cleaned\_data.csv'

data\_cleaned.to\_csv(cleaned\_file\_path, index=False)

（2）制定数据清洗和特征工程规范  
**数据清洗规范（答对3点即可）**

1. 数据加载：使用 pandas 库加载数据集，检查数据的基本结构和类型。

2. 异常值处理： 使用箱线图可视化数值型变量，基于IQR（四分位距）法过滤异常值（范围：Q1-1.5IQR 至 Q3+1.5IQR）。

3. 处理重复行：删除重复的行数

4. 数据归一化: 对数值型变量使用Min-Max归一化（范围0-1）。

5. 保存清洗后的数据: 将经过清洗和处理后的数据保存为新的 CSV 文件，以便后续使用。  
**数据标注规范（答对3点即可）**

1. 数据来源：标注数据的来源，包括数据集的名称、获取日期和数据提供者。

2. 数据描述：提供详细的数据描述，包括每列数据的含义、单位和可能的取值范围。

3. 特征选择: 确定对目标变量预测最有用的特征。

4. 目标变量设定: 保留所有原始10个自变量 + 新建特征IncomeToDebtRatio。目标变量：SeriousDlqin2yrs，需确保标签分布平衡（如不平衡需采用过采样/欠采样）

5. 数据划分: 将数据分为训练集和测试集，通常采用 80/20 的比例，以便于模型的训练和评估。

6. 保存处理后的数据：保存处理后的数据，并记录保存文件的路径和文件名。

7. 数据清洗和标注规范文档