**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：低碳生活行为影响因素预测线性回归模型开发与测试

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）大学生低碳生活行为的影响因素数据集。

**2.**工作任务

在应对气候变化的背景下，了解和促进低碳生活行为变得越来越重要。现要求根据提供的“大学生低碳生活行为的影响因素数据集”，补全2.2.4.ipynb代码。选择合适的特征，开发一个预测大学生低碳生活行为的模型。利用测试工具对模型进行测试，并对测试结果进行分析，完成测试报告，并运用工具对错误原因进行纠正。  
（1）正确加载数据集，并显示前五行的数据。  
（2）请使用线性回归模型，要求设定自变量和因变量，并根据自变量特征进行模型训练，最终将训练好的模型以文件名2.2.4\_model.pkl保存到考生文件夹，结果文件以2.2.4\_results.txt保存到考生文件夹。  
（3）使用测试工具对模型进行测试，并记录测试结果，命名2.2.4\_report.txt，保存到考生文件夹。  
（4）对测试结果进行详细分析，并编写测试报告，包括模型性能评估、错误分析及改进建议，将答案写到答题卷文件中，答题卷文件命名为“2.2.4.docx”，保存到考生文件夹。  
（5）运用工具分析算法中错误案例产生的原因并进行纠正，重新得到模型训练结果，以文件名2.2.4\_results\_xg.txt保存到考生文件夹。  
（6）将以上代码以及运行结果，以html格式保存并命名为2.2.4.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) 能维护日常训练集与测试集。   
(2) 能使用工具对算法进行训练。   
(3) 能使用测试工具对人工智能产品的使用进行测试。   
(4) 能对测试结果进行分析，编写测试报告。   
(5) 能运用工具，分析算法中错误案例产生的原因并进行纠正。

**4.**质量指标

(1) 深入理解业务，训练符合业务需求的模型。   
(2) 数据预处理步骤完整，方法选择合理。   
(3) 代码实现正确，结果符合预期。   
(4) 测试结果分析全面，报告详细。

**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：低碳生活行为影响因素预测线性回归模型开发与测试

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 | 2 | 正确加载数据集得1分，显示前五行的数据得1分，总计得2分； | 根据数据 |  |  |
| M2 | 1 | 删除不必要的列得1分； | 根据数据 |  |  |
| M3 | 2 | 正确定义自变量得1分，因变量得1分，总计得2分； | 根据数据 |  |  |
| M4 | 1 | 训练集与测试集的准确划分得1分； | 根据数据 |  |  |
| M5 | 1 | 初始化线性回归模型得1分； | 根据数据 |  |  |
| M6 | 1 | 训练线性回归模型得1分； | 根据数据 |  |  |
| M7 | 1 | 正确保存线性回归模型文件得1分； | 根据数据 |  |  |
| M8 | 1 | 使用线性回归模型在测试集上进行结果预测得1分； | 根据数据 |  |  |
| M9 | 1 | 保存线性回归模型预测的结果得1分； | 根据数据 |  |  |
| M10 | 2 | 正确计算线性回归模型的均方误差得1分，决定系数得1分，总计得2分； | 根据数据 |  |  |
| M11 | 1 | 初始化XGBoost模型得1分； | 根据数据 |  |  |
| M12 | 1 | 训练XGBoost模型得1分； | 根据数据 |  |  |
| M13 | 1 | 使用XGBoost模型进行预测结果得1分； | 根据数据 |  |  |
| M14 | 1 | 正确计算XGBoost模型的均方误差和决定系数得1分； | 根据数据 |  |  |
| M15 | 3 | 正确填写2.2.4.docx测试报告：其中模型性能评估得1分、错误分析得1分，改进建议得1分，总计得3分； | 根据数据 |  |  |
| 合计配分 | 20 | 合计得分 | |  |  |

**参考答案**：

（1）2.2.4.ipynb代码

import pandas as pd

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

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

import joblib

from xgboost import XGBRegressor

# 加载数据集

data = pd.read\_excel( '大学生低碳生活行为的影响因素数据集.xlsx' )

# 显示数据集的前五行

print(data.head())

# 删除不必要的列并处理分类变量

data\_cleaned = data.drop(columns=['序号', '所用时间']) # 删除不必要的列

data\_cleaned = pd.get\_dummies(data\_cleaned, drop\_first=True) # 将分类变量转换为哑变量/指示变量

# 定义目标变量和特征

target = '5.您进行过绿色低碳的相关生活方式吗?' # 确保这是目标变量

# 定义自变量因变量

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

y = data\_cleaned[target]

# 将数据拆分为训练集和测试集（测试集占20%）

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

# 初始化线性回归模型

model = LinearRegression()

# 训练线性回归模型

model.fit(X\_train, y\_train)

# 保存训练好的模型

model\_filename = '2.2.4\_model.pkl'

joblib.dump(model, model\_filename)

# 进行预测

y\_pred = model.predict(X\_test)

# 将结果保存到文本文件中

results = pd.DataFrame({'实际值': y\_test, '预测值': y\_pred})

results\_filename = '2.2.4\_results.txt'

results.to\_csv(results\_filename, index=False, sep='\t') # 使用制表符分隔值保存到文本文件

# 将测试结果保存到报告文件中

report\_filename = '2.2.4\_report.txt'

with open(report\_filename, 'w') as f:

f.write(f'均方误差: {mean\_squared\_error(y\_test, y\_pred)}\n')

f.write(f'决定系数: {r2\_score(y\_test, y\_pred)}\n')

# 分析并纠正错误（示例：使用XGBoost）

# 初始化XGBoost模型（设定树的数量为1000，学习率为0.05，每棵树的最大深度为5，）

xgb\_model = XGBRegressor(

n\_estimators=1000, # 增加树的数量

learning\_rate=0.05, # 降低学习率

max\_depth=5, # 调整树的深度

subsample=0.8, # 调整样本采样比例

colsample\_bytree=0.8 # 调整特征采样比例

)

# 训练XGBoost模型

xgb\_model.fit(X\_train, y\_train)

# 使用XGBoost模型进行预测

y\_pred\_xg = xgb\_model.predict(X\_test)

# 将XGBoost结果保存到文本文件中

results\_xg\_filename = '2.2.4\_results\_xg.txt'

results\_xg = pd.DataFrame({'实际值': y\_test, '预测值': y\_pred\_xg})

results\_xg.to\_csv(results\_xg\_filename, index=False, sep='\t') # 使用制表符分隔值保存到文本文件

# 将XGBoost测试结果保存到报告文件中

report\_filename\_xgb = '2.2.4\_report\_xgb.txt'

with open(report\_filename\_xgb, 'w') as f:

f.write(f'均方误差: {mean\_squared\_error(y\_test, y\_pred\_xg)}\n')

f.write(f'决定系数: {r2\_score(y\_test, y\_pred\_xg)}\n')

（5）对测试结果进行详细分析，并编写测试报告，包括模型性能评估、错误分析及改进建议，

请勿修改答题卷，在指定单元格内填写答案

**针对线性回归模型**

1、模型性能

|  |  |
| --- | --- |
| 均方误差（MSE） | 0.025 |
| 决定系数（R²） | 0.185 |

2. 错误分析

线性回归的局限性：线性模型无法捕捉非线性关系（如问卷选项的离散分布）。

特征工程不足：未对有序分类变量（如李克特量表）进行编码优化。

数据稀疏性：独热编码导致特征维度增加，部分特征重要性低。

3. 改进建议

a.优化特征处理：

对有序分类变量使用序数编码（如“从不”“偶尔”“经常”转为0,1,2）。

删除低方差特征（如95%样本取同一值的列）。

b. 模型调参：

对XGBoost使用网格搜索优化超参数（如max\_depth、learning\_rate）。

添加早停法（early\_stopping\_rounds=50）防止过拟合。

c.增强解释性：

使用SHAP值分析特征重要性。

可视化残差分布，检查异方差性。