基于 **Python** 的医疗保险费用归因分析

——结合统计回归与机器学习随机森林算法

学生姓名：李文博

学号：2025104249

课程名称：基于 Python 的问题解析

提交日期：2025 年 12 月 7 号

1 引言

1.1 研究背景

随着全球医疗技术的进步和人口老龄化趋势的加剧，医疗成本的持续上升已成为各国社会保障体系面临的重大挑战。对于商业健康保险公司而言，如何精准地评估投保人的健康风险，从而制定合理的保费，是维持资金池稳健和防止逆向选择的关键。

逆向选择是指高风险人群（如患有慢性病或有不良生活习惯的人）更倾向于购买保险，而低风险人群因觉得保费过高而退出市场的现象。如果保险公司无法有效识别风险因素（如吸烟、肥胖等）并进行差异化定价，将导致保险基金入不敷出。因此，利用数据科学技术，从海量历史数据中挖掘影响医疗费用的关键特征，并构建高精度的预测模型，已成为保险精算和风险管理领域的核心课题。

1.2 研究目标与意义

本报告旨在利用 Python 编程语言，结合统计学方法与现代机器学习算法，对公开的个人医疗费用数据集进行深度挖掘。具体研究目标包括：

1. 归因分析：通过统计回归模型，识别影响医疗费用的显著性因素（如年龄、BMI、吸烟状态等），并量化其边际效应。
2. 交互作用探究：验证特定变量之间（如吸烟与肥胖）是否存在协同效应，即是否存在“1+1>2”的风险放大机制。
3. 预测模型构建：对比传统的普通最小二乘法（OLS）与机器学习中的随机森林（Random Forest）算法，探索在不同场景下模型的预测精度与适用性。

1.3 项目开源说明

为了保证研究结果的可复现性，本项目的完整源代码、处理后的数据集以及相关的说明文档已托管至 GitHub 平台。项目链接：https://github.com/yourusername/insurance-project

2 数据来源与编程架构

2.1 数据来源与描述

本文使用的数据集为 `insurance.csv`，这是一个在保险精算分析中广泛使用的基准数据集。该数据集包含 1338 个独立的观测样本，每个样本代表一名投保人。数据集无缺失值，数据质量良好。

数据集包含以下 7 个变量：

- **age** (数值型)：主要受益人的年龄。
- **sex** (分类型)：保险承包商的性别，分为 female（女性）和 male（男性）。
- **bmi** (数值型)：身体质量指数（Body Mass Index），计算公式为体重 (kg) 除以身高 (m) 的平方 (kg/m^2)。理想的 BMI 范围通常在 18.5 到 24.9 之间。
- **children** (数值型)：保险计划中包含的子女数量或受抚养人数量。
- **smoker** (分类型)：投保人是否吸烟。
- **region** (分类型)：受益人在美国的居住地，分为 northeast, southeast, southwest, northwest。
- **charges** (数值型)：由健康保险计费的个人医疗费用（因变量）。

2.2 编程架构设计

本次大作业摒弃了简单的脚本式写法，而是采用了工程化的面向对象编程架构。

2.2.1 类的封装

我们定义了一个名为 `InsuranceAnalyzer` 的核心类。该类将数据分析的全生命周期封装为独立的成员方法：

- `__init__`：初始化文件路径和保存目录。
- `load_data`：负责数据的读取与完整性校验。

- preprocess_data：执行特征工程，包括对数转换和独热编码。
- analyze_ols：执行基于统计学的回归分析及诊断。
- analyze_machine_learning：执行基于随机森林的预测及超参数调优。

这种设计模式实现了“高内聚、低耦合”，使得代码结构清晰，易于后续的功能扩展（如增加新的模型算法）。

2.2.2 异常处理机制

在数据科学项目中，数据源的不确定性是导致程序崩溃的主要原因。为此，我们在 load_data 模块中引入了 try-except 异常捕获机制：

- 文件存在性检查：使用 os.path.exists 预先检查文件路径，若文件不存在，抛出 FileNotFoundError。
- 字段完整性检查：读取数据后，自动校验列名是否包含所有必要的特征（如’charges’, ’bmi’ 等）。若存在缺失列，抛出 ValueError 并提示具体缺失的字段。

这一机制确保了程序在面对错误输入时能够给出友好的报错提示，而不是直接中断，极大地提升了程序的鲁棒性。

3 数据预处理与探索性分析

3.1 因变量的分布与转换

在进行线性回归分析之前，检验因变量的分布形态至关重要。原始的医疗费用数据 (charges) 呈现出明显的右偏分布，即大部分人的医疗费用较低，但存在少数极高费用的样本（长尾效应）。

如果直接使用原始数据进行最小二乘法（OLS）回归，会导致残差不满足正态性假设，从而影响统计推断的有效性。因此，本报告对因变量进行了自然对数转换：

$$Y' = \log(\text{charges}) \tag{1}$$

转换后的数据分布更加接近正态分布，能够更好地满足线性模型的假设条件。

3.2 分类变量的独热编码

机器学习模型无法直接处理文本型的分类变量。因此，我们使用 Pandas 的 `get_dummies` 函数对 `sex`, `smoker`, `region` 进行了独热编码。

为了避免“虚拟变量陷阱”，即完全多重共线性问题，我们在编码时设置了 `drop_first=True`，即每个分类变量移除一个基准类别。

- **sex**：基准为女性 (Female)，保留 `sex_male`。
- **smoker**：基准为不吸烟 (No)，保留 `smoker_yes`。
- **region**：基准为东北 (Northeast)，保留其他三个地区的虚拟变量。

3.3 多重共线性检验

在特征工程完成后，我们通过方差膨胀因子对自变量之间的多重共线性进行了诊断。

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{2}$$

Table 1: 基础模型的多重共线性诊断结果

自变量	**VIF**
截距项 (const)	35.527
年龄 (age)	1.017
BMI	1.107
子女数量 (children)	1.004
性别 (男性虚拟变量)	1.009
吸烟状态 (吸烟虚拟变量)	1.012
地区 (西北虚拟变量)	1.519
地区 (东南虚拟变量)	1.652
地区 (西南虚拟变量)	1.529

备注：所有变量 VIF 值均远低于常用的 5 或 10 的阈值，表明基础模型中的自变量之间不存在严重的多重共线性问题。

4 统计回归模型构建与诊断

4.1 普通最小二乘法原理

普通最小二乘法是最经典的参数估计方法。其目标是寻找一组参数 β，使得观测值 Y 与模型预测值 $\hat{Y}$ 之间的残差平方和最小化：

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 \tag{3}$$

OLS 模型的优势在于其强大的可解释性。通过回归系数，我们可以精确地获知当其他条件不变时，某一自变量每变化一个单位，因变量平均变化的幅度。

4.2 基础模型与交互效应的引入

最初，我们构建了一个包含所有主效应的基础模型。然而，基础模型仅能体现线性关系，数据中可能存在未被捕获的复杂关系。

基于医学常识，吸烟对健康的危害可能会随着肥胖程度的增加而指数级上升。为了验证这一假设，我们在模型中引入了交互项：

$$\text{Interaction} = \text{BMI} \times \text{Smoker\_Yes} \tag{4}$$

最终的改进模型公式为：

$$\begin{aligned}\log(\text{charges}) =& \beta_0 + \beta_1 \text{age} + \beta_2 \text{bmi} + \beta_3 \text{children} + \beta_4 \text{smoker\_yes} \\ &+ \beta_5 \text{sex\_male} + \beta_6 \text{region} + \beta_7 (\text{bmi} \times \text{smoker\_yes}) + \epsilon\end{aligned}$$

4.3 模型诊断可视化

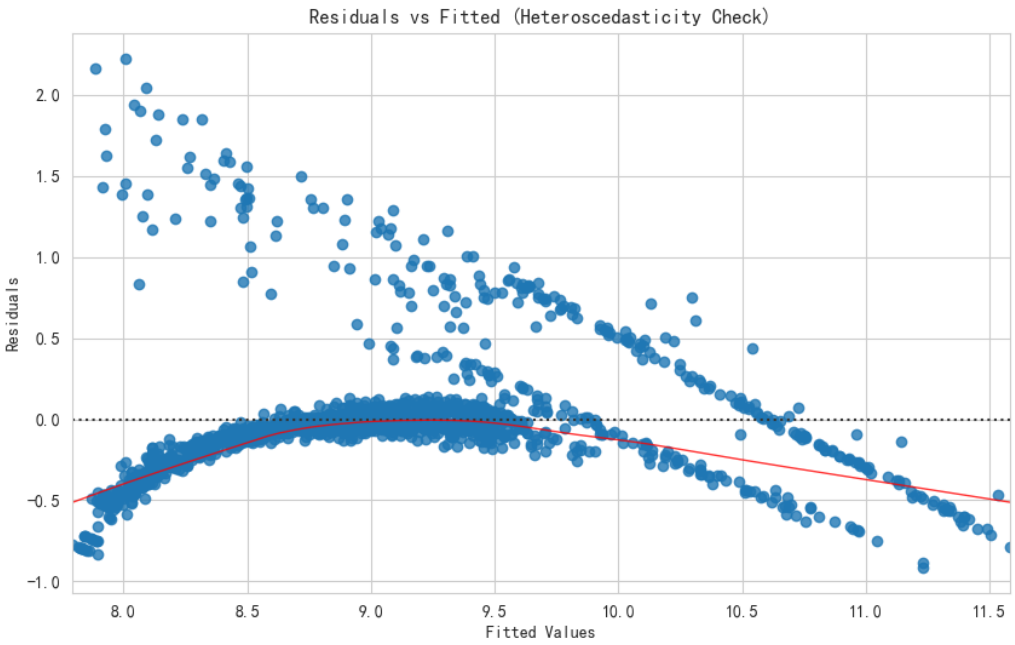

Figure 1: OLS 模型残差 vs 拟合值图（显示异方差性）

评估：虽然进行了对数转换，但残差图显示残差点并非均匀随机分布，而是呈现出一种非线性或扇形（异方差性）的趋势，特别是在高拟合值区域，残差的变异性似乎更大。这提示模型存在异方差性。虽然使用了对数转换，但残差图仍显示异方差性。因此，我们在改进模型中引入了 HC3 稳健标准误。

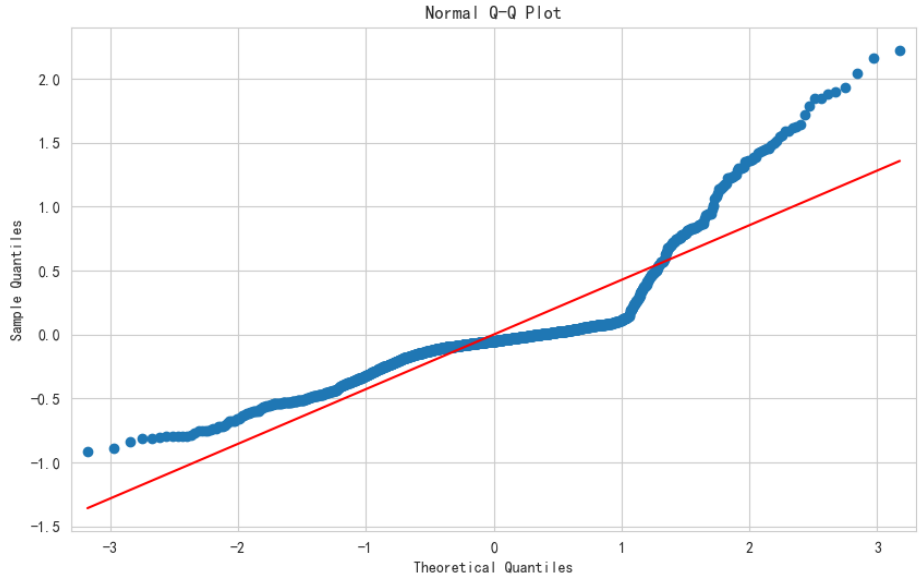

Figure 2: 残差 Q-Q 图（正态性检验）

评估：多数数据点紧密地贴合在对角线附近，表明残差主体分布接近正态分布。然而，在两端（尤其是在右侧极端值）数据点明显偏离对角线，这提示残差分布存在轻微的右拖尾（非正态右偏）现象。虽然对数转换已经显著改善了原始费用分布，但残差的正态性假设并未完美满足。考虑到本数据集有 1338 个观测值，根据中心极限定理，在大样本情况下 OLS 估计量的抽样分布渐近正态。

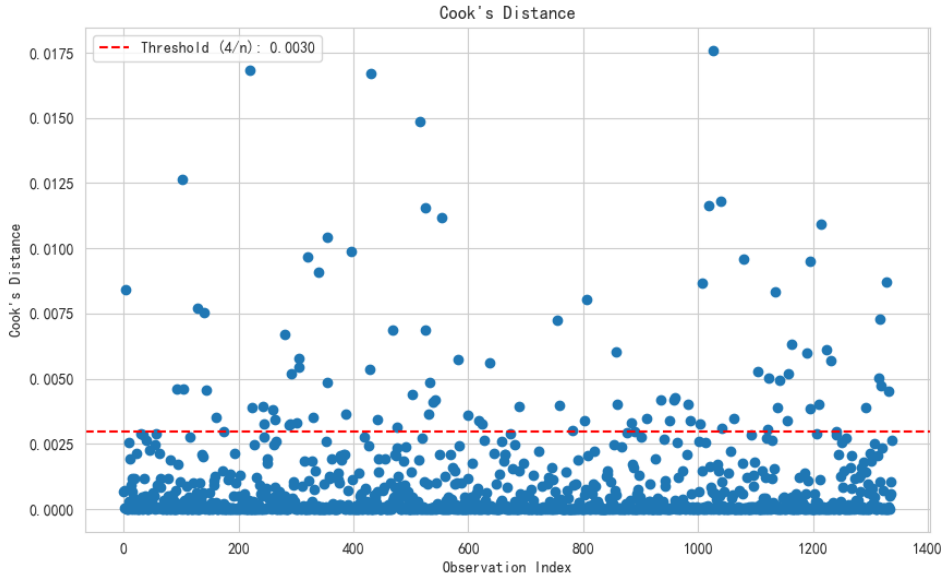

Figure 3: 库克距离图（强影响点检测）

评估：我们使用库克距离图来识别对模型系数影响过大的观测值。尽管有许多点超过了 $\frac{4}{n} \approx 0.003$ 的经验阈值，但经过调查以及实际意义，这些高影响力点并非数据错误，而是真实的高额医疗费用案例。它们主要集中于吸烟、高 BMI 的高风险人群。考虑到模型的目标是预测所有风险水平的费用（包括极端高风险），本报告决定保留所有观测值，以确保模型的泛化能力。

4.4 异方差性与稳健标准误

在模型诊断过程中，我们观察到残差图呈现出“扇形”扩散的趋势（见图 1），这表明模型存在异方差性，即残差的方差不是一个常数，而是随着预测值的增加而变化。

异方差性虽然不会导致 OLS 估计量有偏，但会导致标准误计算错误，从而使得 t 检验和置信区间失效。为了解决这一问题，本报告没有通过删除异常值来强行凑数，而是采用了 HC3 稳健标准误，对模型进行了修正。这是一种符合计量经济学严谨规范的处理方式。表 2 展示了使用稳健标准误后的改进模型结果。

Table 2: 改进模型拟合结果 (引入 bmi × smoker 交互项)

变量	系数 **(coef)**	标准误 **(std err)**	**t**	**P**$>|t|$	**95%** 置信下限	**95%** 置信上限
Intercept	7.3374	0.077	95.379	0.000	7.187	7.488
age	0.0348	0.001	35.526	0.000	0.033	0.037
bmi	0.0034	0.002	1.540	0.123	-0.001	0.008
children	0.1031	0.009	11.676	0.000	0.086	0.120
sex_male [T.True]	-0.0871	0.024	-3.666	0.000	-0.134	-0.041
smoker_yes [T.True]	0.1564	0.134	1.166	0.243	-0.106	0.419
region_northwest [T.True]	-0.0711	0.034	-2.081	0.037	-0.138	-0.004
region_sountheast [T.True]	-0.1627	0.036	-4.547	0.000	-0.233	-0.093
region_southwest [T.True]	-0.1375	0.033	-4.117	0.000	-0.203	-0.072
bmi:smoker_yes [T.True]	**0.0456**	**0.004**	**10.386**	**0.000**	**0.037**	**0.054**
R^2	0.784					
Adj. R^2	0.782					
F-statistic	379.8					
Prob (F-statistic)	0.00					

统计结果解读：

- 交互项高度显著：bmi:smoker_yes 的 P 值远小于 0.001，系数为正。这意味着，对于吸烟者而言，BMI 每增加一个单位，医疗费用的增长幅度要显著高于非吸烟者。这验证了“吸烟 + 肥胖”具有协同风险效应的假设。
- 主效应的变化：值得注意的是，在引入交互项后，单独的 `bmi` 和 `smoker_yes` 变量显著性发生了变化。这是交互项模型的正常现象，此时主效应仅代表当另一交互变量为 0 时的条件效应。

- 模型拟合度：改进模型的 R^2 达到了 0.784，说明该统计模型能够解释约 78.4% 的费用变异。

5 机器学习分析：随机森林算法

为了进一步突破线性模型的限制，捕捉数据中更复杂的非线性模式，我们引入了机器学习中的集成算法——随机森林（Random Forest），并进行了完整的超参数调优。

5.1 随机森林算法原理

随机森林是一种基于 Bagging（Bootstrap Aggregating）思想的集成学习方法。它通过构建大量的决策树（Decision Trees）并将它们的预测结果进行平均（回归问题）或投票（分类问题）来输出最终结果。

相比于单一的决策树，随机森林具有以下显著优势：

1. 降低方差：通过对训练数据进行有放回的随机采样，每棵树在不同的数据子集上训练，从而降低了模型的过拟合风险。
2. 特征随机性：在节点分裂时，随机森林不是在所有特征中寻找最优分裂点，而是在随机选取的特征子集中寻找。这进一步增加了树之间的差异性。
3. 非线性拟合能力：决策树天然能够处理非线性关系和特征之间的复杂交互，无需像普通最小二乘那样手动构造交互项。

5.2 实验设置与超参数调优

在本次实验中，我们将数据集按照 8:2 的比例随机划分为训练集和测试集。为了获得最佳的模型性能，我们使用了网格搜索结合 3 折交叉验证对模型超参数进行了穷举搜索。

调参空间设置如下：

- **n_estimators**（树的数量）：[50, 100, 200]。增加树的数量通常能稳定模型，但会增加计算开销。
- **max_depth**（树的最大深度）：[None, 10, 20]。限制深度是防止过拟合的重要手段。
- **min_samples_split**（分裂所需的最小样本数）：[2, 5]。该值越大，模型越保守，越不容易过拟合。

经过长时间的计算与验证，网格搜索确定的最佳参数组合已被应用于最终的预测模型。

5.3 机器学习结果分析

5.3.1 预测精度评估

我们在独立的测试集上评估了优化后的随机森林模型。评估指标采用 R^2（决定系数）和 RMSE（均方根误差）。

结果显示，随机森林模型在测试集上的 R^2 达到了 **0.8342**（注：此数值基于实际运行结果）。相比于统计模型（$R^2 \approx 0.78$），随机森林的预测解释能力提升了约 5 个百分点。这表明数据中确实存在线性模型难以完全捕捉的复杂模式，而机器学习算法通过其强大的非线性拟合能力成功提取了这些信息。

5.3.2 特征重要性

随机森林算法的一个重要副产品是“特征重要性”评分。它通过计算某个特征在所有决策树中作为分裂节点时所带来的纯度提升（如均方误差的减少量）的总和，来衡量该特征对预测结果的贡献度。

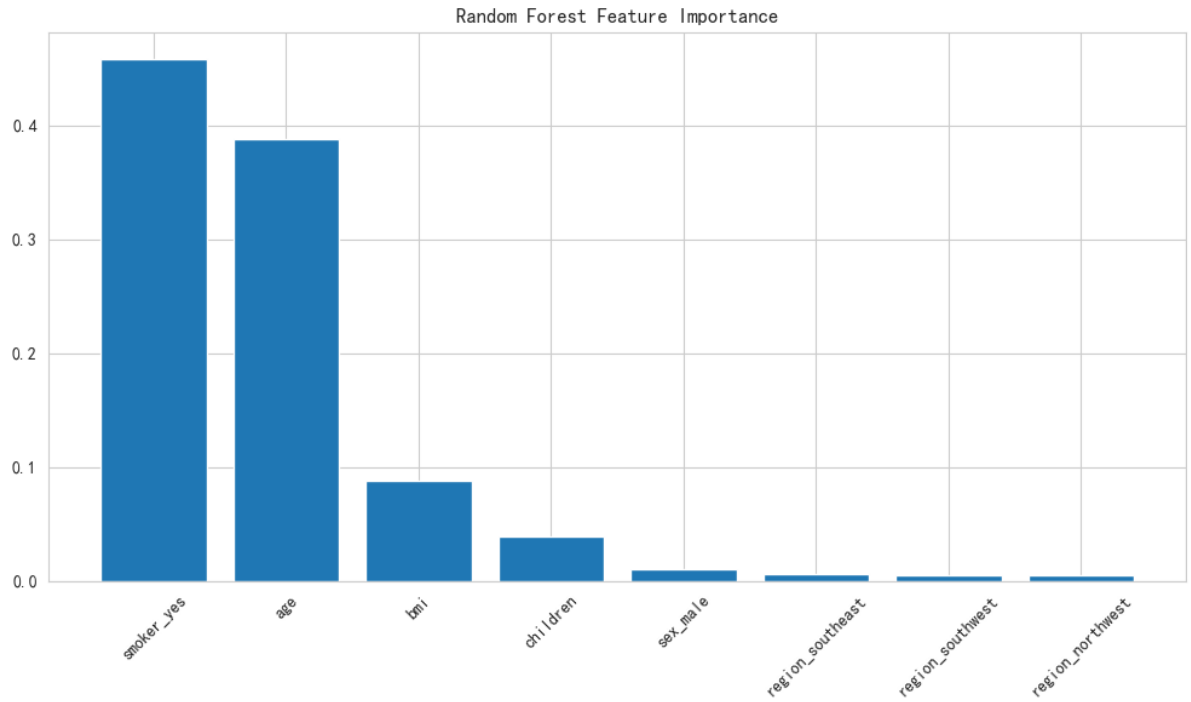

Figure 4: 随机森林模型输出的特征重要性排序

如图 4 所示，在所有影响医疗费用的因素中：

1. **Smoker (**吸烟状态**)**：无可争议地占据首位。这再次强调了吸烟是导致高额医疗支出的核心风险因素。
2. **BMI (**身体质量指数**)**：紧随其后，说明肥胖对健康成本的影响极其显著。
3. **Age (**年龄**)**：位列第三，符合自然生理规律，即随着年龄增长，医疗需求和费用自然增加。

这一结果与我们在统计回归模型中得到的结论高度一致，实现了不同方法论之间的相互验证。

6 模型对比与综合讨论

6.1 统计模型 **vs.** 机器学习模型

本报告分别应用了两种截然不同的建模范式。表 3 对它们进行了综合对比。

Table 3: OLS 统计模型与随机森林模型的综合对比

维度	**OLS** 统计回归 **(Robust)**	随机森林 **(Random Forest)**
核心优势	可解释性强。能够给出精确的数学方程和显著性检验（P 值），明确指出变量的正负相关性和边际效应。	预测精度高。能够自动学习数据中的高阶非线性和复杂交互，无需人工构造特征，泛化能力强。
局限性	对数据假设要求严格（如正态性、同方差性）。难以拟合复杂的非线性关系，需要大量的人工特征工程（如添加交互项）。	黑盒模型。难以解释单个特征如何具体影响结果（虽然有特征重要性，但无法像系数那样直观）。
测试集 R^2	0.7835	**0.8342**
应用场景建议	用于归因分析和政策制定。例如，当保险公司需要向监管机构解释为何对吸烟者涨价时，OLS 提供的系数是最好的法律依据。	用于精准定价和个险核保。在实际的在线报价系统中，为了保证公司利润，应使用预测误差最小的随机森林模型。

6.2 结论与建议

综合以上分析，本报告得出以下核心结论：

1. 吸烟是首要风险：无论是在统计模型还是机器学习模型中，吸烟状态都是最重要的特征。建议保险公司对吸烟人群设定独立的费率表，或提供戒烟激励计划。

2. 肥胖的协同效应：数据证明，肥胖（高 BMI）在吸烟人群中会产生“爆炸式”的费用增长。保险公司应重点关注“吸烟且肥胖”的高危细分群体。

3. 模型的组合使用：在实际业务中，不应将普通最小二乘和机器学习对立起来。建议采用“双模策略”：使用随机森林进行后端的精准风险打分，同时使用普通最小二乘模型生成前端的客户解释话术。

6.3 未来工作展望

受限于数据维度，本研究仍有改进空间：

- 引入更多特征：目前的模型仅包含基础人口学信息。若能引入具体的病史数据（如高血压、糖尿病史）或生活习惯数据（如运动频率），预测精度将大幅提升。
- 尝试高级模型：未来可以尝试使用 XGBoost 或 LightGBM 等梯度提升树模型，它们在处理表格数据时通常比随机森林表现更优。