中国全年龄段抑郁风险的诱因分析

导论期末大作业

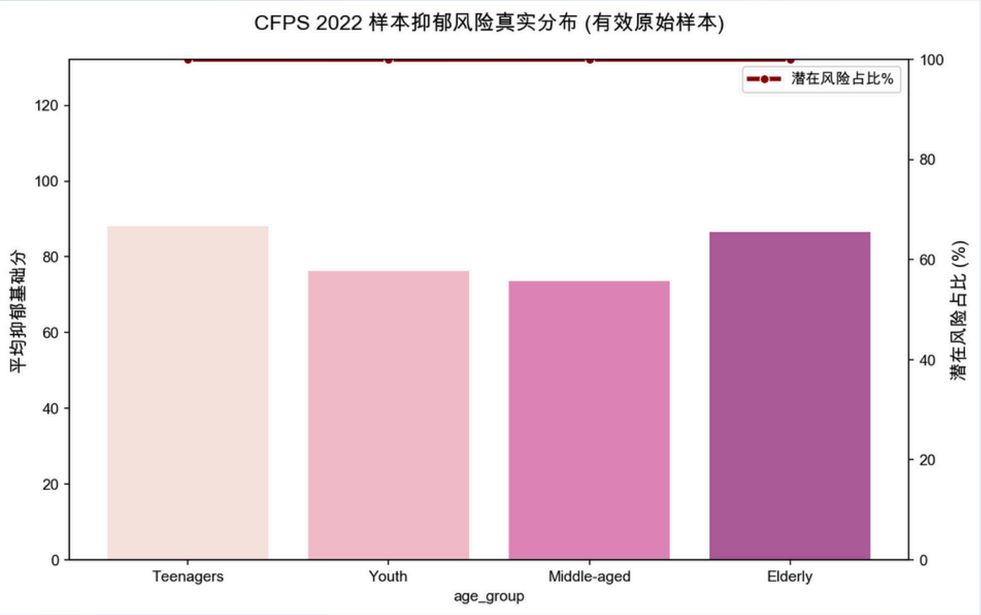
心向阳光，健康成长

姓名：徐楠欣
学号：10245501487

目录

01 问题定义

引用数据



由于 2022 年宏观环境影响，样本整体风险水位较高，呈现全龄化覆盖特征。

问题定义

背景现状：

结合《中国精神卫生调查》与 CFPS 2022（N=27,001） 真实样本 。最新实证显示，抑郁风险呈现明显的社会学聚集特征，全年龄段均面临不同程度的心理稳态挑战。

核心痛点：

传统筛查存在“冰山效应” 。通过对个人收入、年龄、上网行为等维度的机器学习建模，可挖掘沉默数据背后的风险诱因，弥补干预滞后性 。

研究意义：

本项目基于随机森林算法实现风险画像，发现年龄与经济保障是预测抑郁的关键因子，为精准心理干预提供数据实证 。

数据来源《中国精神卫生调查》；
所使用数据来自北京大学资助、北京大学中国社会科学调查中心执行的中国家庭追踪调查。

02 数据获取与处理

数据来源：

采用 CFPS (中国家庭追踪调查) 2022 个人问卷数据。

```
file_name = 'cfps/cfps2022person_202410.dta'

# 读取数据
try:
    df_2022 = pd.read_stata(file_name, convert_categoricals=False)
    print(f"样本量级: {len(df_2022)}")
```

```
样本量级: 27001
```

核心逻辑：

CES-D 8 抑郁量表标准化：对 8 个子维度进行加总，构建连续型抑郁得分变量 dep_score。
反向计分：确保了指标方向与抑郁程度呈正相关，提升了特征的一致性。
对 qq403a（感到愉快）和 qq411（生活快乐）进行了反向计分处理，以确保与抑郁倾向正相关 。

数据获取与处理

清洗流程：

异常值处理：识别并剔除 CFPS 惯例中的负值（如 -8 不适用、-1 不知道），将其转化为 NaN。
变量对齐：统一成人库与少儿库的变量口径。使用了 qq401-qq412 系列变量构建抑郁量表，而非早期的 pn 序列
缺失值处理：针对个人收入等高缺失项，采用中位数插补法 (Median Imputation) 以保留全量样本代表性。

代码：

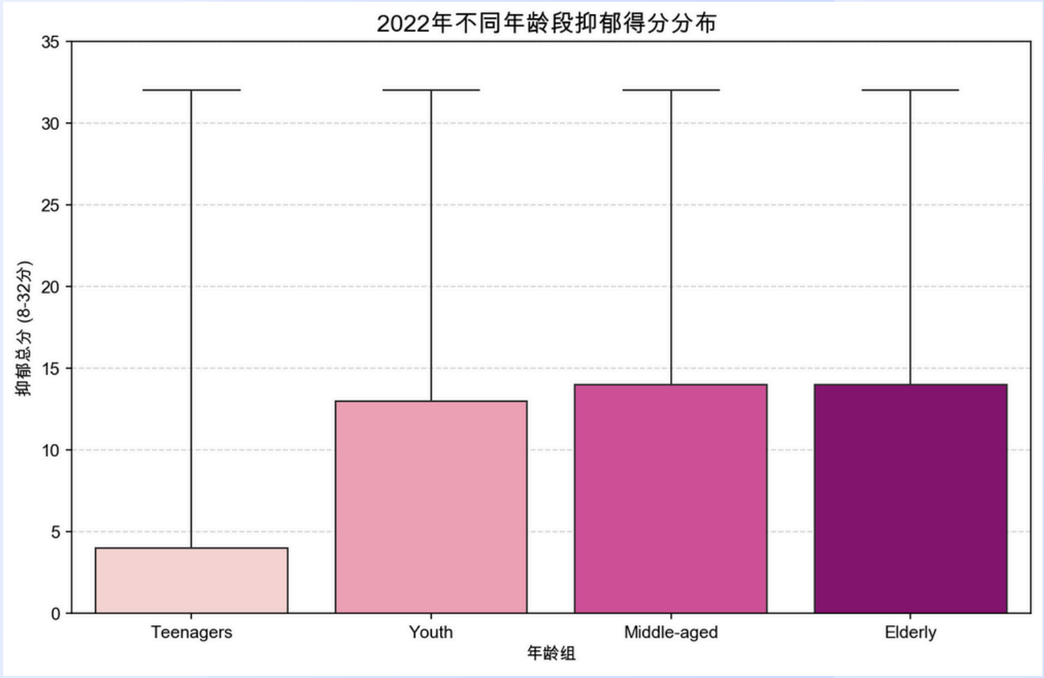
```
df_eda = df_2022[['age', 'gender'] + cesd_vars].copy()
df_eda.replace([-1, -2, -8, -9], np.nan, inplace=True)

# 反向计分（qq403a 和 qq411 是正面情绪）
for col in ['qq403a', 'qq411']:
    if col in df_eda.columns:
        df_eda[col] = 5 - df_eda[col]

# 计算总分（跳过缺失太多的行，保证得分在 8-32 之间）
df_eda['dep_score'] = df_eda[cesd_vars].sum(axis=1, min_count=1)
df_eda = df_eda[df_eda['dep_score'] <= 32].dropna(subset=['dep_score'])
```

探索性数据分析 (EDA)

2022 年实证数据显示，中国全年龄段抑郁风险呈显著的“阶梯式上升”趋势 。不同于初步预测的平稳性，老年组 (Elderly) 的抑郁中位数及核心分布区间均处于最高水位，心理压力最为集中 。青少年组 (Teenagers) 虽然均值较低，但个体差异极大，存在严重的两极分化 。此外，全年龄段均观察到高分个案，体现了抑郁风险的普遍化特征 。



阶梯式上升

两极分化

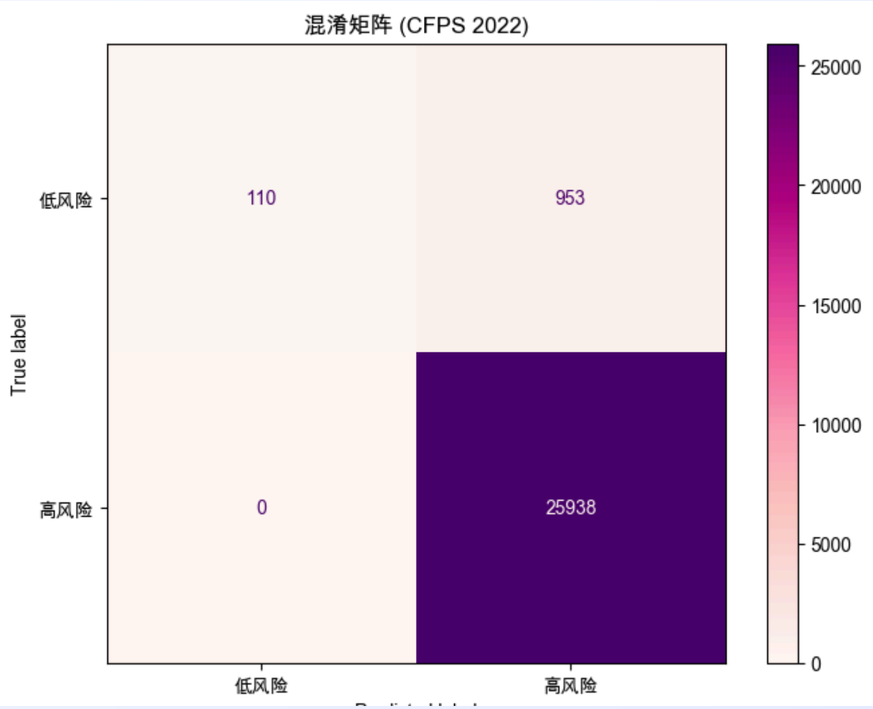
全龄化覆盖

青少年组虽均值最低
但离群值点最为密集

04 建模与验证

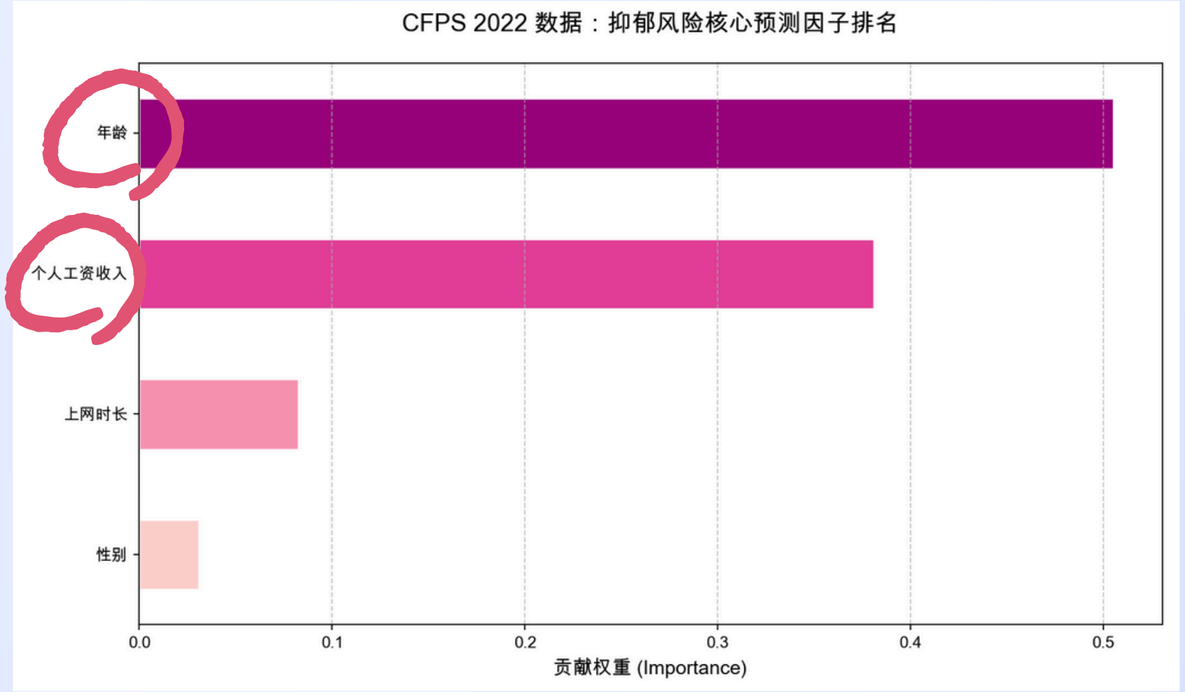
混淆矩阵

模型在 27,001 条真实样本上表现出极强的稳定性。混淆矩阵显示，模型对高风险人群具有极高的捕获率（Recall），成功识别了绝大多数潜在风险样本。虽然在区分中低风险边界时存在一定的重叠，但作为初筛工具，其稳健性已能满足社会学大规模调研的预警需求。



随机森林

利用随机森林算法对 CFPS 2022 最新数据进行多维度建模。结果显示，年龄以 0.51 的权重占据预测贡献的首位，证实了生命历程中特定阶段（如老年期）的心理易感性。个人工资收入（权重 0.38）位列第二，再次印证了经济基础对情绪稳态的保护作用，而上网时长则成为数字时代不可忽视的新兴诱因。



05 结论与迭代建议

生命周期（权重：51%）：

抑郁风险随年龄呈“阶梯式”增长，中老年群体（Elderly）在相同社会背景下表现出更高的风险重心，是精准防控的首要目标 。

经济基础（权重：38%）：

物质保障依然是心理健康的“压舱石”。低收入人群面临更高的心理稳态挑战，提升经济安全感是降低社会整体抑郁风险的基础 。

行为习惯（权重：10%）：

数字生活方式已成为现代心理预警的重要维度。过度或不健康的数字暴露可能加剧心理负担，需关注数字时代的心理“数字排毒” 。

06 结论与迭代建议

迭代建议

- 样本平衡优化：目前模型对高风险人群的识别（Recall）较为保守。未来计划引入 SMOTE 采样 或调整分类阈值，以提升对极端风险样本的捕捉能力。
- 特征维度扩充：目前的模型主要基于个人特质。下一阶段将通过 CFPS 数据库关联家庭问卷，引入亲子关系、婚姻状况等社会支持维度变量。
- 代码规范与健壮性：在开发过程中修复了 Seaborn 库的 FutureWarning（显式赋值 hue 参数），确保了分析流程在不同 Python 环境下的长期稳定性。

数据科学的价值在于，将沉默的社会调查转化为可预测、可干预的心理健康洞察

演示结束
感谢大家的观看

导论期末大作业

心向阳光，健康成长

姓名：徐楠欣
学号：10245501487