

课程设计报告

**中文题目**

**糖尿病预测数据建模项目**

English Title

**Diabetes Prediction Data Modeling Project**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程名称 | 大数据概论 | 学院 | 软件学院 软件工程专业 |
| 小组成员 | （首位组长）杨丰竹  张旭旸  张韶卿  王炳澈  曲翰林 | | |

**北京交通大学**

2025年6月

项目成员姓名： 曲翰林23301073

摘要：

Abstract:

## 研究内容

本项目旨在通过机器学习模型对糖尿病进行预测。研究的核心问题是如何利用 BRFSS（行为风险因素监测系统）调查数据，构建一个有效预测糖尿病发病风险的模型。通过数据预处理、样本平衡、特征标准化与特征重要性分析，最终选取合适的模型并评估其性能。

## 设计方案

项目整体设计流程如下：

1. **数据准备与预处理**
   * 加载 BRFSS 数据集。
   * 处理缺失值、异常值。
   * 类别变量编码 + 数值标准化
   * SMOTE 合成样本以缓解类别不平衡
2. **样本平衡**
   * 由于患病和未患病样本比例严重不平衡，采用上采样方法对少数类样本进行扩增。
3. **模型选择与训练**
   * 使用 XGBoost、LightGBM、Random Forest 构建基模型。
4. 基于 StackingClassifier 构建集成模型进行最终预测。
5. **模型评估**
   * 采用准确率、召回率、ROC 曲线、混淆矩阵等多维指标进行评估。
   * 使用特征重要性图解释模型。

## 设计实现

主要步骤在 model.ipynb 中实现：

* **数据划分与标准化**：

scaler = MinMaxScaler()

X\_scaled = scaler.fit\_transform(X)

* **类别不平衡处理（上采样）**：

df\_majority = df[df['diabetes'] == 0]

df\_minority = df[df['diabetes'] == 1]

df\_minority\_upsampled = resample(df\_minority, replace=True,

n\_samples=len(df\_majority), random\_state=123)

df\_upsampled = pd.concat([df\_majority, df\_minority\_upsampled])

* **模型训练（以 XGBoost 为例）**：

model = XGBClassifier(use\_label\_encoder=False, eval\_metric='logloss')

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

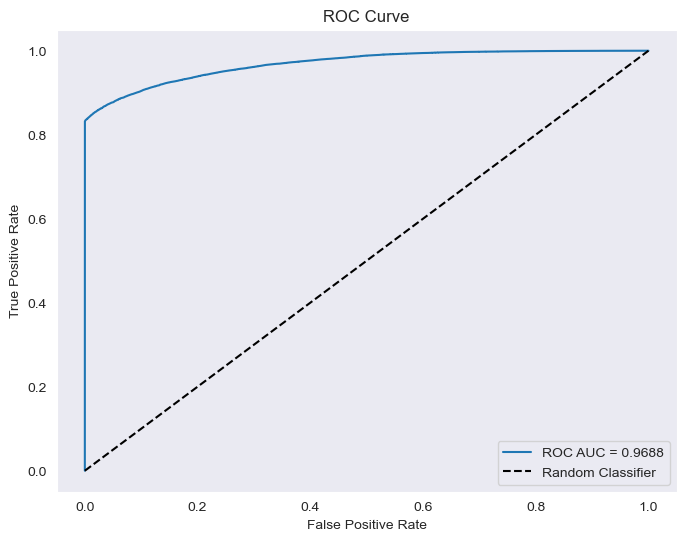
* **性能评估**：

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

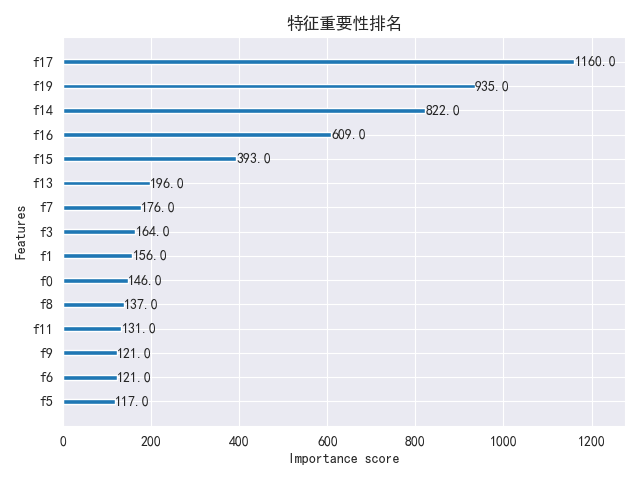
print(classification\_report(y\_test, y\_pred))

## 结果与可视化

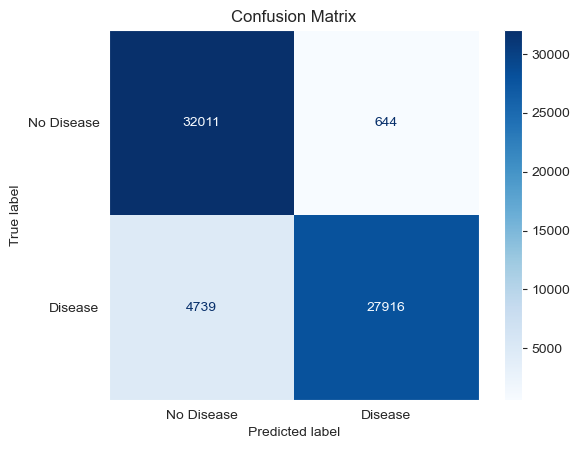
ROC 曲线：



特征重要性图：



**混淆矩阵：**



## 总结与展望

本项目成功实现了基于 BRFSS 数据的糖尿病预测模型，具备如下特点：

* 模型具备较高的精度和召回率，能有效识别高风险人群；
* XGBoost 在特征表达能力上优于传统模型；
* 结果具备一定解释性，便于公共健康政策调整参考。

**未来可拓展方向：**

* 引入更多医疗指标（如血糖、胰岛素水平）；
* 尝试深度学习模型进一步提升预测性能；
* 构建 Web 接口用于实际部署，服务于社区健康平台

## 参考文献

 Friedman, J.H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". Annals of Statistics.

 Breiman, L. (2001). "Random Forests". Machine Learning.

 Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system". Proceedings of the 22nd ACM SIGKDD.

 BRFSS: Behavioral Risk Factor Surveillance System — [CDC 官网](https://www.cdc.gov/brfss/index.html)

 sklearn 官方文档: <https://scikit-learn.org/>

LightGBM 文档：<https://lightgbm.readthedocs.io/>