第五届中国研究生人工智能创新大赛

基于多维时序模型的高糖发生次数监测

项目文档

3.0

2023.08.07

对不对对对队

华为赛题三

目录

记录更改历史

序号	更改原因	版本	作者	更改日期	备　注
1	无	1.0	对不对对对队	2023.7.21	Baseline
2	优化	2.0	对不对对对队	2023.7.27	提交版本
3	优化	3.0	对不对对对队	2023.8.6	最终提交版本

1 项目概况

1.1 背景和基础

本项目旨在利用智能穿戴技术监测糖尿病前期人群的血糖水平，以提供有效的控糖方案。根据 ADA 标准，我国成年人中糖尿病前期患病率超过 35%，而控糖对于降低 II 型糖尿病和心血管疾病的风险具有重大意义。然而，由于成本和体验等原因，糖尿病前期人群中通过有创设备监测控糖水平的人很少。因此，智能穿戴技术的进步为解决糖尿病前期人群的控糖水平监测提供了前所未有的机会。

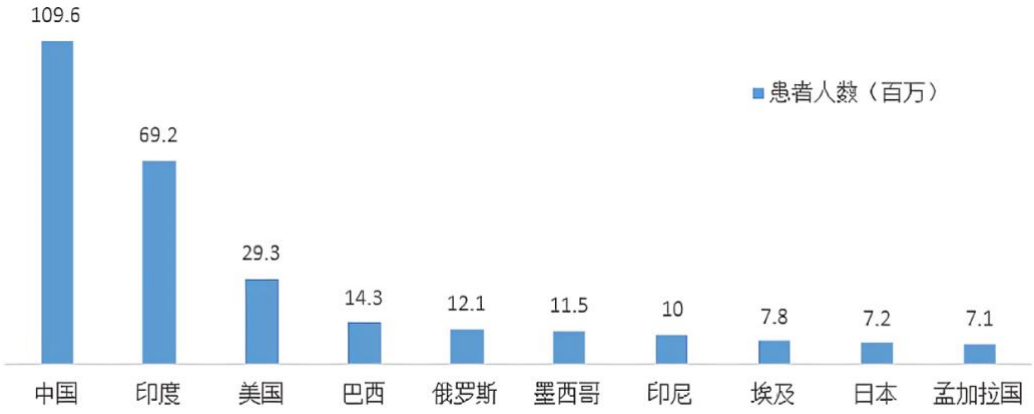

图 1-1 各国糖尿病人数统计图（2020 年）

该项目的灵感来自于对糖尿病前期人群的关注和对智能穿戴技术的发展。糖尿病前期人群面临着控制血糖水平的挑战，而传统的有创设备监测方法存在成本高、使用不便等问题。智能穿戴技术的快速发展为解决这一问题提供了新的可能性。通过结合智能穿戴设备和机器学习算法，我们可以实时监测糖尿病前期人群的血糖水平，并提供个性化的控糖方案，从而改善他们的生活质量和健康状况。

我们的团队由具有丰富软件工程经验的人员组成，团队成员包括四名计算机专业的学生。我们的团队具备深入了解糖尿病领域和机器学习技术的能力，以确保项目的成功实施。一名成员将负责开发智能穿戴设备的数据处理系统，两名成员将负责构建机器学习模型，一名同学将负责学习并提供专业的医学知识和指导。

1.2 场景和价值

本项目适用于糖尿病前期人群，通过智能穿戴设备监测血糖水平，提供个性化的控糖方案。该项目的应用场景包括但不限于以下几个方面：

1. 糖尿病前期管理：通过实时监测血糖水平，该项目可以帮助糖尿病前期人群更好地控制血糖，降低发展为 II 型糖尿病和心血管疾病的风险。

2. 健康管理：该项目可以为个人提供个性化的健康管理方案，帮助他们更好地了解自己的血糖水平，并采取相应的措施来改善生活方式和健康状况。

3. 医疗研究：通过收集大量的血糖数据，该项目可以为医疗机构提供有价值的数据，用于研究和改进糖尿病管理策略。这将有助于推动糖尿病领域的科学研究和医疗进步。

该项目的潜在社会价值巨大。首先，它可以帮助糖尿病前期人群更好地控制血糖水平，降低发展为 II 型糖尿病和心血管疾病的风险。其次，通过提供实时的血糖监测和个性化的建议，该技术可以改善糖尿病前期人群的生活质量。最后，这项技术还可以为医疗机构提供有价值的数据，用于研究和改进糖尿病管理策略。

在项目实施过程中，我们将进行市场调研和对比性分析，以评估项目的商业潜力和竞争优势。我们将研究市场上已有的智能穿戴设备和糖尿病管理解决方案，并分析它们的特点、优势和不足之处。这将帮助我们确定项目的定位和差异化竞争策略，以满足用户需求并获得市场份额。

1.3 所需支持

在项目实施过程中，我们需要以下支持：

1. 算力：为了训练和测试机器学习模型，我们需要足够的计算资源。这可能包括云计算服务或高性能计算集群。

2. 硬件：为了测试模型的实际效果，我们需要合适的硬件设备。这可能包括心率监测器、血容量脉冲传感器等。

3. 相关培训：为了确保团队成员具备足够的专业知识和技能，我们可能需要相关培训，包括糖尿病病理知识、机器学习算法等方面的培训。

4. 专业支持：我们可能需要与医疗专家合作，以确保我们的解决方案符合医学标准和最佳实践。

通过获得这些支持，我们将能够顺利实施项目，并为糖尿病前期人群提供更好的控糖方案，从而改善他们的生活质量和健康状况。

2 项目规划

2.1 整体目标

本项目的整体目标是希望通过智能穿戴技术和机器学习算法，实现对糖尿病前期人群高糖发生次数的实时监测和预测。通过展示模型的效果，我们将验证该技术在糖尿病管理领域的潜力，并为进一步的研究和开发奠定基础。具体目标如下：

1. 实时监测和预测高糖发生次数：通过智能穿戴设备收集的生理数据，我们将实时监测糖尿病前期人群的高糖发生次数，并利用机器学习算法对数据进行分析和建模，实现对高糖发生次数的准确预测。我们将通过原型系统展示这一功能，并验证其准确性和实用性。

2. 验证技术在行业中的潜力：我们将与医疗专家和糖尿病管理机构合作，将模型应用于实际场景中，例如糖尿病前期人群的日常生活中。通过与行业专家的合作和反馈，我们将初步验证该技术在糖尿病管理领域的潜力，并收集用户的意见和建议，以进一步改进和优化原型系统。

3. 提供个性化的控糖建议：基于实时监测和预测的结果，我们将为糖尿病前期人群提供个性化的控糖建议。通过结合个体的生理数据和机器学习模型的预测结果，我们将为每个个体提供针对性的建议，帮助他们更好地控制高糖发生次数。

4. 推动糖尿病管理领域的创新：通过本项目的展示和验证，我们希望推动糖尿病管理领域的创新。我们将与行业专家和研究机构分享我们的研究成果，并积极参与相关学术会议和研讨会，以促进知识交流和合作，推动糖尿病管理领域的发展。

2.2 技术创新点

本项目的主要技术创新点包括以下几个方面：

1. 数据处理

针对本赛题给出的原始数据集较为混乱的问题，我们进行了清洗和分析处理。采取了空值填充和时序对齐等操作，以提高数据的可用性。通过增加高斯噪声对正样本进行扩充，解决了数据正负样本不均衡的问题，增强了模型的泛化能力。

2. 投票机制

对于 8 个训练集及对应的验证集，我们建立一个模型列表，包含八个模型：

$$List_{xgb} = \{XGB_1, XGB_2, \cdots, XGB_8\} \tag{2-1}$$

$$List_{lgb} = \{LGB_1, LGB_2, \cdots, LGB_8\} \tag{2-2}$$

每一个模型用不同的训练集和验证集训练：

$$XGR_i(D_i^{train}, D_i^{val}) or LGB_i(D_i^{train}, D_i^{val}) \tag{2-3}$$

$$i = \{1,2,\cdots,8\}$$

然后我们将模型通过加权求和，得到最终的大模型：

$$XGB = \sum_i^N \alpha_i * XGB_i \tag{2-4}$$

$$LGB = \sum_i^N \alpha_i * LGB_i \tag{2-5}$$

其中 α 的取值有多种选择，例如用样本个数占比来确定：

$$\alpha_i = \frac{D_i^{train} + D_i^{val}}{\sum_i^N (D_i^{train} + D_i^{val})}, N = 8 \tag{2-6}$$

最后计算整个测试集的结果：

$$Result = XGB(D_{test}) or LGB(D_{test}) \tag{2-7}$$

3. 损失函数类设计

在二分类问题中，样本不均衡是指不同类别的样本数量差异较大，这可能导致模型对数量较少的类别学习不足，从而影响模型性能。实验中发现，高糖数据在数据集中占比较小，导致模型过于偏向于预测低糖样本，从而影响了模型的准确性，为了解决这一问题，我们设计了一个损失函数类 WeightBCEWithLogitsLoss，该函数引入了样本权重因子 beta，用负样本数量除以正样本数量，使得损失计算过程中考虑样本不均衡的影响，让模型更关注高糖数据，并引入了参数 alpha 来调节样本权重因子 beta。计算回滚权重因子，用正样本数量除以总样本数量，这个因子将用于控制整体损失的缩放，保证正负样本对整体损失的贡献平衡。

经过实验验证，该损失函数成功地平衡了预测结果，提高了模型对高糖的预测能力，并保持了整体准确性，为解决样本不均衡问题提供了有效的解决方案。

3 实施方案

3.1 技术可行性分析

为了确保项目的顺利实施，我们进行了综合的技术可行性分析，主要考虑了以下几个方面：

1. 数据采集：

为实现通过穿戴设备监测高糖发生次数的目标，我们需要收集糖尿病前期人群的生理数据和血糖监测数据。数据集包含了 16 个糖尿病前期用户的多种生理指标，例如血糖浓度、血容量脉冲信号、皮肤电活动、皮肤温度和三轴加速度计等。我们将采用智能穿戴设备，如心率监测器、血糖监测仪等，实时记录用户的生理数据，并将其传输到数据采集系统中进行存储和分析。由于本赛题已经提供数据集，我们只需对数据进行必要的清洗和预处理。

2. 行业知识获取：

为了确保项目的成功实施，我们需要获取关于糖尿病前期人群管理和控糖水平监测的行业知识。这可以通过学习网络上的糖尿病相关知识、咨询医疗专家和糖尿病管理机构来实现。通过与行业专家交流，我们可以深入了解糖尿病前期人群的需求和挑战，了解当前行业中的最佳实践和最新研究成果。将行业知识融入项目设计和实施，可以更好地满足用户需求，提高项目的实用性和可操作性。

3. 算力和硬件支持：

项目进行中需要大量计算资源和存储空间，特别是在机器学习模型训练和数据分析阶段。我们将利用云计算平台或高性能计算集群来满足这些需求。通过合理规划和配置算力和硬件资源，我们可以保证项目的顺利进行，并提高数据处理和分析的效率，缩短项目周期。

4. 技术可行性评估：

在项目初期，我们将进行全面的技术可行性评估，以确定所选技术和方法的可行性。评估内容包括所选机器学习算法在给定数据集上的性能和准确性，以及所需的计算资源和时间成本。通过使用部分数据集进行实验和测试，我们可以及时发现潜在问题，优化方案，确保项目的可行性和成功实施。

综上所述，通过数据处理、获取行业知识、确保足够的算力和硬件支持，并进行技术可行性评估，我们可以确保项目的顺利进行。这些措施将充分发挥智能穿戴技术和机器学习算法的优势，实现对糖尿病前期人群高糖发生次数的准确监测和预测，为糖尿病前期人群提供更好的健康管理服务。这一技术在未来有望应用于糖尿病管理领域，为临床医生和糖尿病患

者提供有力的支持和帮助。同时，技术可行性分析也为项目的顺利实施奠定了坚实的基础，为论文的成果和影响力打下了良好的基础。。

3.2 技术细节

3.2.1 数据处理

分析题目中给出的数据集可知: 此数据集由 16 个糖前（或者接近糖前）用户组成，血糖监测设备记录用户的葡萄糖浓度（mg/dl）,穿戴记录了血容量脉冲（BVP）信号、皮肤电活动（EDA）、皮肤温度和三轴加速度计。数据集包含有 001-016 共 16 个子文件夹，其中每个文件夹有 8 个文件数据。此处首先针对 001 的数据做初步的分析。

下表是对数据的简单说明：

表 3-1 数据集分文件初步分析表

序	文件名	数据类型	数据说明	数据属性	数据示例
1	ACC.csv	三轴加速度计数据	数据包括“时间戳”作为日期时间值，加速度计数据将用于“X”、“Y”、“Z”方向	datetime, acc_x, acc_y, acc_z	2020-02-13 15:28:50.000000, -34.0,17.0, 55.0
2	BVP.csv	血容量脉搏数据	数据将包括“时间戳”作为日期时间值，“血容量脉搏值”作为当时记录的测量值。	datetime, bvp	2020-02-13 15:28:50.000000, -0.0
3	Dexcom.csv	间质葡萄糖浓度数据	数据将包括“时间戳”作为日期时间值和“间质葡萄糖浓度值”作为当时记录的测量值。（Glucose Value 作为监督数据）	Index, Timestamp, … Glucose Value , …	13,2020-02-13 17:23:32, EGV,,,,iPhone G6,61.0,,,,, 11101.0
4	EDA.csv:	皮肤电活动数据	数据将包括“时间戳”作为日期时间值和“皮肤电活动值”作为当时记录的测量值。	datetime, eda	2020-02-13 15:28:50.250, 0.001281

5	Food_Log.csv	参与者在整个研究过程中消耗的食物日志	数据将包括“日期”作为日期值，…，“total_fat”作为数值	date, time, … total_fat	2020-02-13,18:00:00, 2020-02-13 18:00:00…
6	HR.csv	心率数据	数据将包括“时间戳”作为日期时间值和“指心率值”作为当时记录的测量值。	datetime, hr	2020-07-12 15:29:00,94.0
7	IBI.csv	心跳间隔数据	数据将包括“时间戳”作为日期时间值和“指心跳间隔值”作为当时记录的测量值。	datetime, ibi	2020-02-13 15:33:22.059328, 0.828162999999999
8	TEMP.csv	皮肤温度数据	数据将包括“时间戳”作为日期时间值和“皮肤温度值”作为当时记录的测量值。	datetime, temp	2020-02-13 15:28:50.000, 30.21

空值填充

同时，经过统计，发现该数据集的部分文件存在含有缺失值的属性，具体数据量和缺失情况如下表所示:

表 3-2 001 数据文件缺失情况表

文件名	时间长度	数据形状	空值情况
ACC	1/32 s	(20296428, 4)	无空值
BVP	1/64 s	(40592838, 2)	无空值
Dexcom	5 min	(2573, 13)	大量空值
EDA 1/4 s	1/4 s	(2537046, 2)	无空值
Food_Log	时间段不固定	(61, 14)	部分空值
HR	1 s	(634188, 2)	无空值
IBI	时间段不固定	(266366, 2)	无空值
TEMP 1/4	1/4 s	(2537040, 2)	无空值

由表中可以看出，001 文件中只有 Dexcom 和 Food_Log 两个子文件有空值，需要单独对这两个文件进行处理，主要采用的方法是删除一些空值较多且对目标值没有依赖关系的属性，并采用一些空值填充的方法（如插值）进行填充。

时序对齐

由于各文件收集数据时的时间长度不同，需要进行时序对齐。由于最终需要以 Dexcom 文件中的 Glucose Value 作为金标，所以要以 Dexcom 文件的时间粒度为标准进行时序对齐。

表 3-3 001 分文件时序分析表

文件	时间区间	是否包含 Dexcom 所在区间
ACC	2020-02-13 15:28:50--2020-02-22 17:56:03	是
BVP	2020-02-13 15:28:50--2020-02-22 17:56:03	是
Dexcom	2020-02-13 17:23:32--2020-02-22 17:53:23	/
EDA	2020-02-13 15:28:50--2020-02-22 17:56:03	是
Food_Log	2020-02-13 18:00:00--2020-02-22 15:10:00	相隔不远，可以近似处理
HR	2020-02-13 15:29:00--2020-02-22 17:56:03	是
IBI	2020-02-13 15:33:22--2020-02-22 17:51:35	相隔不远，可以近似处理
TEMP	2020-02-13 15:28:50--2020-02-22 17:56:03	是

经过对上表的分析可知，以处理后的 Dexcom 文件为准，可以对齐 Dexcom 和各子文件的时间。下表是最终处理完成的数据示例：

Timestamp	acc_x	acc_y	acc_z	bvp	eda	calorie	total_carb	sugar	protein	hr	ibi	temp	Glucose Value
2020/2/13 17:23	18.1296992	1.16502193	34.3895894	0.00021997	0.26220702	456	85	83	16	73.376233	0.86949175	33.3987167	61
2020/2/13 17:28	1.0940625	53.3115104	33.4972917	-0.0008885	0.70173193	456	85	83	16	58.145	0.74203631	32.9641667	59
2020/2/13 17:33	11.1063542	41.8947917	20.9789583	-0.0022432	1.0784631	456	85	83	16	58.375	0.79443333	33.3022333	58
2020/2/13 17:38	30.0026607	26.0711603	10.6370513	0.00510799	1.45635476	456	85	83	16	58.58	0.77924753	33.2745987	59
2020/2/13 17:43	35.7888147	30.8278381	-9.08702	-0.001881	2.76300999	456	85	83	16	58.75	0.7954909	33.5300833	63
2020/2/13 17:48	30.2038021	38.0750521	-36.16474	-0.0062302	5.12487362	456	85	83	16	58.885	0.76974672	33.9847167	67
2020/2/13 17:53	26.1789583	23.8836458	-35.078542	0.00084844	4.52801531	456	85	83	16	58.975	0.78547404	34.573925	68
2020/2/13 17:58	37.9075499	7.64512271	-13.487937	0.00639845	2.16480046	456	85	83	16	59.015	0.75315936	34.818397	63
2020/2/13 18:03	33.2955491	9.3765079	24.4486793	-0.0170274	1.63687878	488	2	0	63	59.05	0.80201983	34.76675	59
2020/2/13 18:08	19.0579167	15.3453646	45.6076042	-0.0031755	1.08706405	488	2	0	63	59.085	0.79349333	34.9256	60

图 3-1 001 数据文件处理后前 10 条数据示例

数据扩充（高糖样本）

我们加的是高斯白噪声。高斯白噪声定义如下：

定义一：如果一个噪声，它的瞬时值服从高斯分布，而它的功率谱密度又是均匀分布的，则称它为高斯白噪声。

定义二：在一般的通信系统的工作频率范围内热噪声的频谱是均匀分布的，好像白光的频谱在可见光的频谱范围内均匀分布那样，所以热噪声又常称为白噪声。由于热噪声是由大量自由电子的运动产生的，其统计特性服从高斯分布，故常将热噪声称为高斯白噪声。

信噪比公式如下：

$$SNR = 10\log_{10}\frac{P_{singal}}{P_{noise}} = 10\log_{10}\frac{\sum x^2}{\sum n^2} \quad (3\text{-}1)$$

本项目中我们按照信噪比 SNR 大小来给原始信号中添加白噪声。

相关性分析

为了便于在调试模型时进行参数调整，提前使用热力图做了相关性分析，使我们对于各个属性与目标值的相关性有一个大致的把握。结果如下图所示：

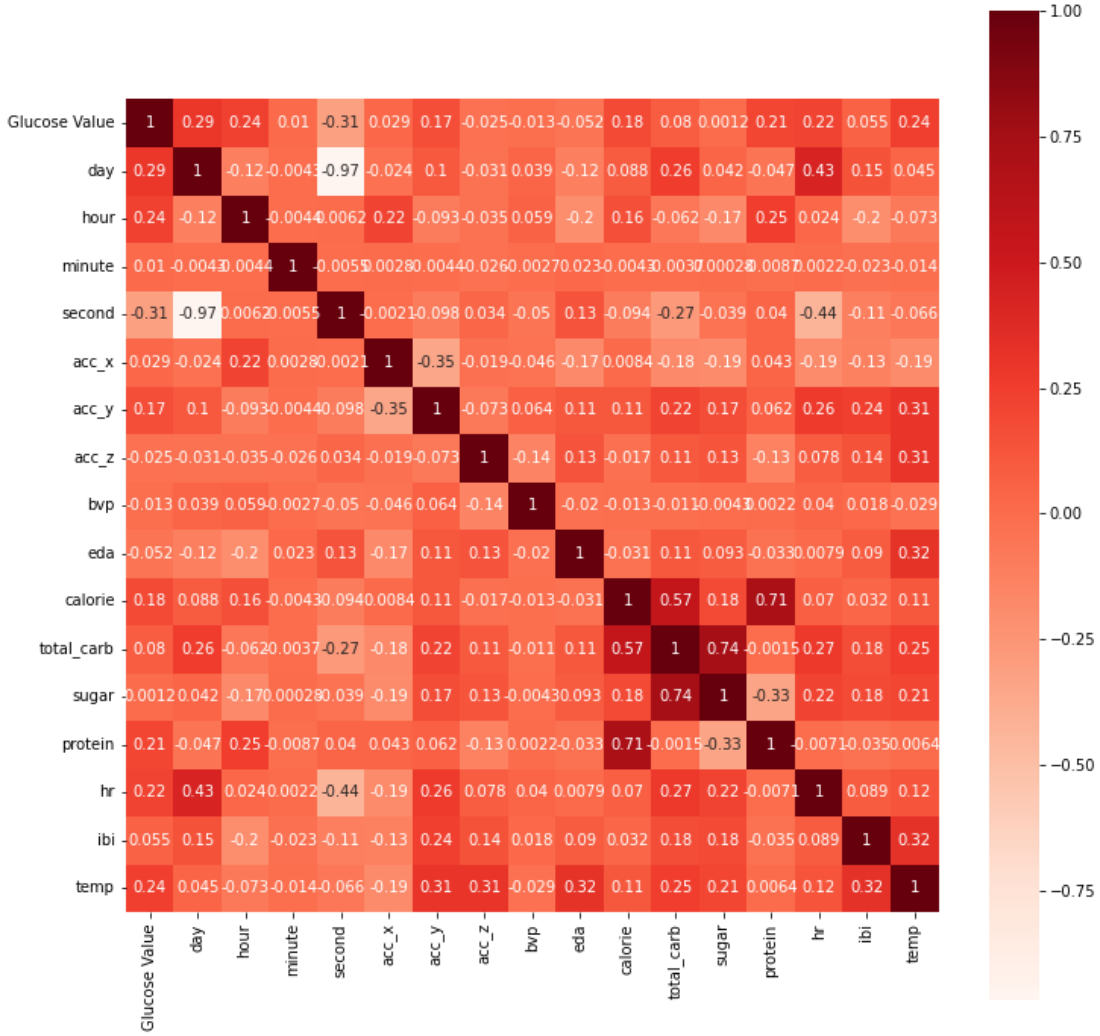

图 3-2 属性相关性分析热力图

由热力图可知，day、hour、protein、hr、temp 五个属性与目标值 Glucose Value 的相关

性较高，均超过了 0.2。

3.2.2 机器学习模型

本文选择的机器学习模型是$XGBoost$和 $Lightgbm$。

XGBoost 模型

$XGBoost$是提升方法中的一个可扩展的机器学习系统.其原理如下:

$XGBoost$ 的优化目标为:

$$\begin{cases} L(\phi) = \sum_i l(\hat{y}_i, y) + \sum_k \Omega(f_k) \\ where \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \end{cases} \tag{3-2}$$

其中，l 是可导且凸的损失函数，衡量 y 和 $\hat{y}_i$ 的相似程度，Ω 是正则项，T 表示叶子结 点的个数，γ 越大，T 越小；最后一部分是 $L2$ 正则项，对叶子结点的权值 w 进行惩罚。

上式整体上难以优化，因此采用贪心策略，每一步增加一个基分类器，使得每步损 失变小，得到以下评价当前基学习器的函数：

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3-3}$$

将上式在f_t=0 处二阶泰勒展开并删去常数项，得到：

$$L^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{3-4}$$

上式等价于：

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3-5}$$

对 w_j 求偏导并将其置为零，得到最优权值：

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} (h_i + \lambda)} \tag{3-6}$$

将上式回代到目标函数中即可得到评分函数：

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{i=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} (h_i + \lambda)} + \gamma T \tag{3-7}$$

算法基本思想是:遍历所有特征的所有分割方法，选择损失最小的，得到两个子树，然后继续遍历.为限制树的深度，可以设定阈值，分裂后增益大于阈值才继续分裂。

L*ightgbm*模型

相对于 $XGBoost$ 算法，该算法大大的降低了运行的速度.主要的原因就是传统的 $boosting$ 算法需要对每一个特征都要扫描所有的样本点来选择最好的切分点，这非常的耗时.为了解决这种在大样本高纬度数据的环境下耗时的问题，$Lightgbm$ 使用了如下两种解决办法:一是 $GOSS(Gradient-basedOne-SideSampling$, 基于梯度的单边 采样)，不是使用所用的样本点来计算梯度，而是对样本进行采样来计算梯度;二是 $EFB(ExclusiveFeatureBundling$，互斥特征捆绑) ，这里不是使用所有特征来进行扫描获得最佳切分点，而是将某些特征捆绑在一起来降低特征的维度，从而减少寻找最佳切分点的时间消耗。

1. $GOSS$ 算法

在 Adaboost 算法中，每次迭代时会更加关注上一次错分的样本点，即上一次错分的样本点的权重会增大。而在 GBDT 中，并没有类似 Adaboost 中的本地样本权重调整的过程，因此 Adaboost 中提出的采样模型不能直接应用在 GBDT 中。然而，每个样本的梯度却提供了非常有用的信息。如果一个样本点的梯度较小，那么该样本点的训练误差就较小，并且该样本点已经经过了较好的训练。直接抛弃梯度较小的样本点可能会改变数据的分布并降低模型的精度。

为了解决这两个问题，GBDT 中提出了 GOSS（Gradient-based One-Side Sampling，基于梯度的单边采样）算法。GOSS 充分利用了样本点的梯度信息来进行样本采样。具体而言，GOSS 在样本采样时保留了所有梯度较大的样本点，而对梯度较小的样本点进行采样。这样一来，可以保留那些训练误差较大且需要更多关注的样本，同时也减少了对那些训练误差较小且已经较好训练的样本的处理，从而提高了算法的效率。。

Algorithm : Gradient-based One-Side Sampling

Input: I: training data, d: iterations
Input: a: sampling ratio of large gradient data
Input: b: sampling ratio of small gradient data
Input: $loss$: loss function, L: weak learner
models ← {}, fact ← $\frac{1-a}{b}$
topN ← a × len(I) , randN ← b × len(I)
for i = 1 ***to*** d **do**
  preds ← models.predict(I)
  g ← $loss$(I, preds), w ← {1,1,...}
  sorted ← GetSortedIndices(abs(g))
  topSet ← sorted[1:topN]
  randSet ← RandomPick(sorted[topN:len(I)], randN)
  usedSet ← topSet + randSet
  w[randSet] × = fact ▷ Assign weight $fact$ to the small gradient data.
  newModel ← L(I[usedSet], − g[usedSet], w[usedSet])
  models.append(newModel)

图 3-3 $GOSS$算法流程图

输入训练数据，迭代步数 d，大梯度数据的采样率 a，小梯度数据的采样率 b，损失函数和弱学习器的类型，根据样本点的梯度绝对值对其进行排序，取前 $a * 100\%$的样本生成一个大梯度样本点的子集，对剩下的样本集合随机选取 $b * (1 - a) * 100\%$ 个样本点，生成一个小梯度样本点的集合，将大梯度样本集合和采样的小梯度样本集合合并，并将小梯度样本集合乘上一个权值系数 $(1 - a)/b$，使用上述采样样本学习一个新弱学习器，不断重复直到达到规定的迭代次数或者收敛为止。

2. EFB算法

L$ightgbm$算法在实现中不仅进行了数据采样，还采用了一种特殊的特征抽样方法，以进一步降低模型的训练速度。这种特征抽样与传统的特征抽样方法有所不同，它将互斥特征绑定在一起，从而减少特征维度。

该方法的主要思想是针对高维度稀疏数据，我们可以设计一种几乎无损的方式来减少有效特征的数量。在稀疏特征空间中，许多特征往往是互斥的，即很少同时出现非 0 值。因此，我们可以安全地将这些互斥特征绑定在一起，形成一个新的特征，从而减少特征的数量。

然而，将特征划分为更小的互斥绑定是一个 NP-hard 问题，即在多项式时间内无法找到准确的解决方案。为了应对这个问题，L$ightgbm$算法采用了一种近似解决方法。它允许特征之间存在少量非互斥的样本点，即某些对应的样本点之间不同时为非 0 值。这种允许小部分冲突的策略可以获得更小的特征绑定数量，从而进一步提高计算的有效性。

在下图中，展示了互斥特征绑定算法的流程图，这个算法能够在实际应用中有效地进行特征抽样，进一步优化模型的训练过程。

通过 L$ightgbm$算法中的数据采样和特征抽样，我们成功降低了模型训练的复杂度和计算成本，同时保持了较好的性能。这种创新性的方法为高维稀疏数据的处理提供了一种有效的解决方案，为模型的实际应用带来了显著的优势。

Algorithm : Greedy Bundling

Input: F: features, K: max conflict count
Construct graph G
searchOrder ← G.sortByDegree()
bundles ← {}, bundlesConflict ← {}
for *i* ***in*** *searchOrder* **do**
    needNew ← True
    for $j = 1$ ***to*** *len(bundles)* **do**
        cnt ← ConflictCnt(bundles[j],F[i])
        if *cnt + bundlesConflict[i]* $\leq K$ **then**
            bundles[j].add(F[i]), needNew ← False
            break
    if *needNew* **then**
        Add $F[i]$ as a new bundle to *bundles*
Output: *bundles*

图 3-4 EFB 算法流程图

输入特征 F, 最大冲突数 K，图 G, 构造一个边带有权重的图，其权值对应于特征之间的总冲突，通过特征在图中的度来降低排序特征，检查有序列表中的每个特征，并将其分配给具有小冲突的现有特征集合。

$Lightgbm$ 关于互斥特征的合并用到了直方图 $Histogram$ 算法。直方图算法的基本思想是先把连续的特征值离散化成 k 个整数，同时构造一个宽度为 k 的直方图。在遍历数据的时候，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后根据直方图的离散值，遍历寻找最优的分割点。

在训练模型时，本文使用的 $Lightgbm$ 的具体参数设置如下表所示。

3.2.3 深度学习模型

本文选择的深度学习模型是 LSTM 和 Transformer。

LSTM 模型

LSTM（Long Short-Term Memory，长短期记忆）是一种用于处理序列数据的深度学习模型，可以很好地解决长期依赖问题，被广泛应用于自然语言处理、语音识别、时间序列预测等领域。LSTM 模型通过引入记忆单元和三个门控机制来解决长期依赖问题。记忆单元负责存储序列信息，三个门控机制（遗忘门、输入门、输出门）则可以控制记忆单元的读写和保留程度，从而允许模型选择性地忽略或重要的信息。在 LSTM 模型中，每个时间步骤都有一个输入和一个输出，输入是当前的输入数据和上一个时间步骤的隐藏状态，输出是当前时间步骤的输出和当前时间步骤的隐藏状态。模型的参数在所有时间步骤上是共享的，因此可以处理任意长度的输入序列。

LSTM 模型的优点包括：

1. 可以处理长期依赖问题：由于引入了记忆单元和门控机制，LSTM 模型可以很好地处理长期依赖问题，避免了传统的 RNN 模型中的梯度消失或爆炸问题。

2. 高度可扩展性：LSTM 模型可以很容易地堆叠多个 LSTM 层，以处理更复杂的序列数据。

3. 适用于多种序列任务：LSTM 模型可以应用于各种序列任务，如文本分类、情感分析、机器翻译、时间序列预测等。

4. 可解释性：LSTM 模型的门控机制使其具有可解释性，可以帮助我们理解模型对序列数据进行的处理。

LSTM 就是一种门控 RNN，其单一节点的结构如下图所示。LSTM 的巧妙之处在于通过增加输入门控，遗忘门控和输出门控，使得自循环的权重是变化的，这样一来在模型参数固定的情况下，不同时刻的积分尺度可以动态改变，从而避免了梯度消失或者梯度膨胀的问题。

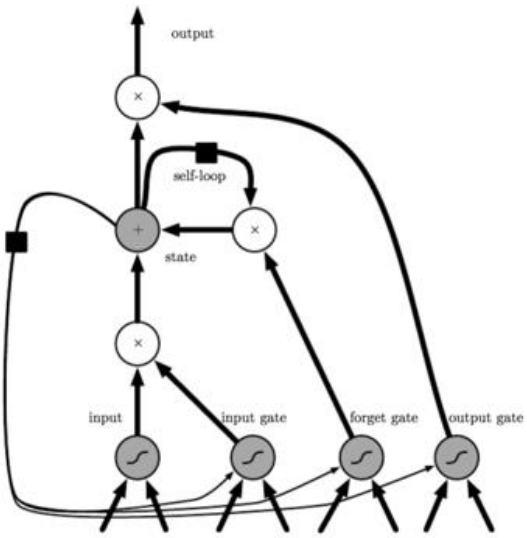

图 3-5 LSTM 单节点示意图

Transformer 模型

Transformer 模型是一种用于自然语言处理和其他序列到序列任务的深度学习模型。它在 2017 年由 Google 的研究人员提出，并在机器翻译任务中取得了显著的突破。传统的序列模型，如循环神经网络（RNN）和卷积神经网络（CNN），在处理长序列和捕捉长期依赖关系时存在一些限制。而 Transformer 模型通过引入自注意力机制（self-attention）来解决这些问题。Transformer 模型由编码器（Encoder）和解码器（Decoder）组成。编码器将输入序列映射为一系列的隐藏表示，而解码器则根据编码器的输出和之前的预测，逐步生成目标序列。Encoder-Decoder 的结构如下图所示：

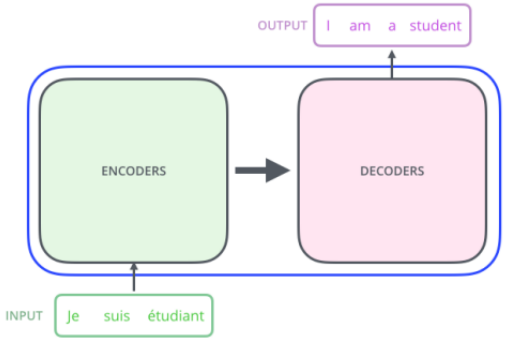

图 3-6 Encoder-Decoder 结构示意图

Transformer 是一个基于多头注意力机制的模型，自注意力机制是 Transformer 模型的核心。它允许模型在生成隐藏表示时，根据输入序列中的不同位置之间的关系进行加权。这样，模型可以更好地捕捉到序列中的重要信息，而不受序列长度的限制。除了自注意力机制，Transformer 模型还引入了位置编码（Positional Encoding）来提供序列中单词的位置信息。这样，模型可以更好地理解单词在序列中的顺序。Transformer 的模型结构如下图所示：

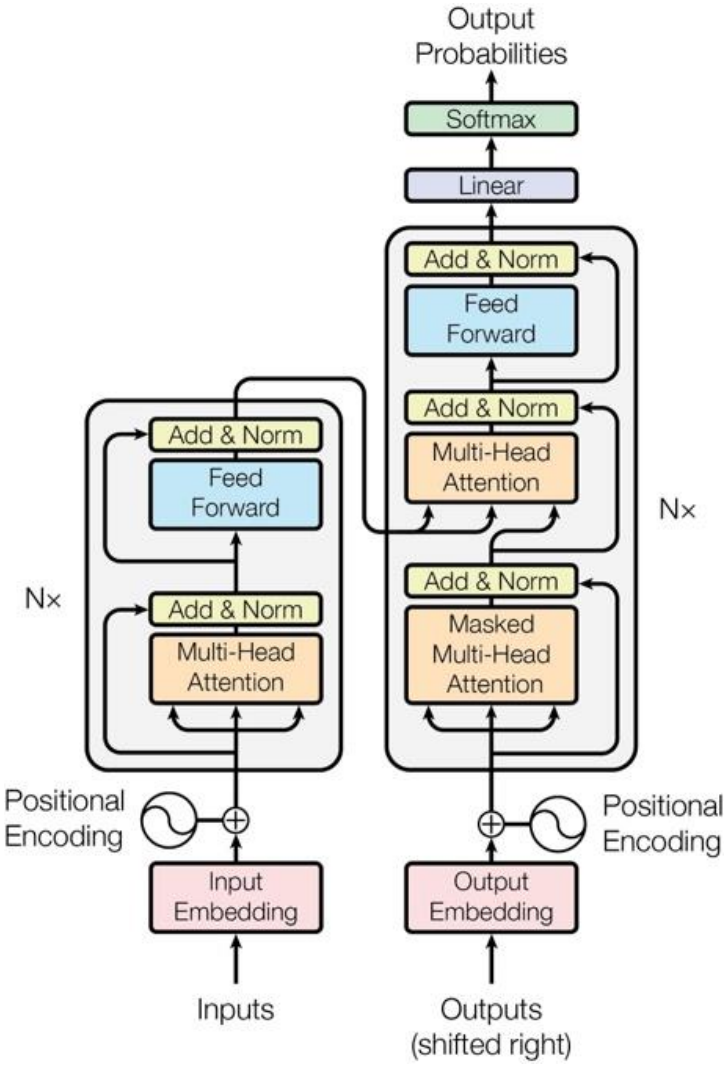

图 3-7 Transformer 模型结构示意图

Transformer 模型的优点包括并行计算能力强、能够处理长序列、能够捕捉全局依赖关系等。它在机器翻译、文本生成、语言理解等任务中取得了很好的效果，并成为了自然语言处理领域的重要模型之一。本项目中血糖浓度受时间影响较大，数据也是时间序列数据，使用 Transformer 模型能更好的提取数据的特征。

3.2.4 模型训练与调参

五折交叉验证

本文使用五折交叉验证方法对模型进行训练和评估。在进行训练时，将整个训练集分为三个部分：训练集（train_set）、评估集（valid_set）和测试集（test_set）。其中，测试集仅用于观测模型在数据集上的测试效果，不参与实际训练过程。

在实际训练中，我们发现模型对训练集的拟合程度通常较好（可能受初始条件影响），但对于训练集之外的数据的拟合程度通常不尽如人意。为了更客观地评估模型的性能，我们不将所有数据用于训练，而是将一部分数据留出来（不参与训练），用于测试训练集生成的参数在训练集之外数据上的拟合程度。这就是交叉验证（Cross Validation）的思想。

简而言之，交叉验证是多次进行 train_test_split 划分的过程。每次划分时，在不同的数据集上进行训练和测试评估，得出一个评价结果。本文采用五折交叉验证，即在原始数据集上进行五次划分，每次划分进行一次训练和评估。最后，将这五次划分的评估结果取平均，得到最终的评分。这样做的目的是更好地评估模型的性能，并减少因数据划分不同而引起的评估结果波动。

总之，五折交叉验证方法在训练和评估过程中充分利用数据集，对模型性能进行客观、准确的评估，帮助我们更好地理解模型在训练集之外数据上的表现。下图是五折交叉验证的示意图：。五折交叉验证的示意图如下图所示。

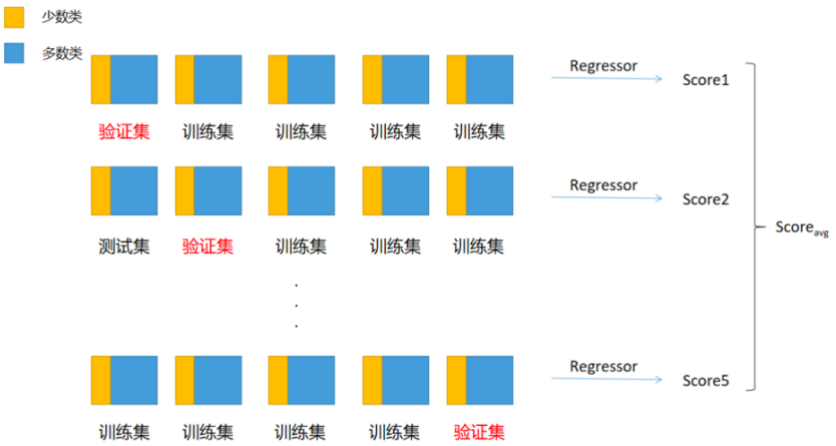

图 3-8 五折交叉验证示意图

在模型调优时，使用 K(K=5) 折交叉验证用于模型调优，找到使得模型泛化性能最优的超参值，找到后，在全部训练集上重新训练模型，并使用独立测试集对模型的性能做出最终评价。K 折交叉验证使用了无重复抽样技术的好处：每次迭代过程中每个样本只有一次被划入训练集或测试集。

模型参数表

各个模型的参数表如下：

表 3-4 $XGBoost$模型参数表

参数	值
n_estimators	120
learning_rate	0.1
gamma	0
subsample	0.8
colsample_bytree	0.9
max_depth	6
verbose	30
early_stopping_rounds	15

表 3-5 $Lightgbm$模型参数表

参数	值
learning_rate	0.0596
boosting_type	gbdt
objective	regression_l1
n_estimators	1000
min_child_samples	168
min_child_weight	0.11
num_leaves	60
max_depth	60
n_jobs	-1
reg_alpha	1
reg_lambda	1
verbose	30
early_stopping_rounds	15

表 3-6 Transformer 模型参数表

参数	值
TRAIN_NUM	16
input_size	12
hidden_size	32
output_size	1
epochs	20
lr	0.01
time_step	10

表 3-7 T LSTM 模型参数表

参数	值
TRAIN_NUM	16
input_size	12
hidden_size	32
output_size	1
epochs	10
lr	0.01

3.2.5 模型测试结果

机器学习模型

1. $XGBoost$模型

首先使用$XGBoost$模型，将数据集按照个体随机划分为 8+8 的训练集和测试集，训练集再根据 8:2 划分为训练数据和验证数据，做五折交叉验证。同时利用验证数据将模型做早停（输入模型参数即可）。模型结果如下表所示：（选取 5 次实验）

表 3-7 $XGBoost$模型结果测试表

	acc	累计金标差异次数	平均每天金标差异次数
1	0.804919	3589	65.25
2	0.808851	2888	52.51
3	0.830056	3079	55.98
4	0.723712	2079	37.80
5	0.730126	2510	45.64

由上表可以得知，$XGBoost$模型预测的测试集准确率较高，但累计金标差异天数比较大，即出现了较多高糖样本并未准确分类。

2. $Lightgbm$模型

接着使用 $Lightgbm$模型，同样将数据集按照个体随机划分为 8+8 的训练集和测试集，训练集再根据 8:2 划分为训练数据和验证数据，做五折交叉验证。同时利用验证数据将模型做早停（输入模型参数即可）。模型结果如下表所示：（选取 5 次实验）

表 3-8 $Lightgbm$模型结果测试表

	acc	累计金标差异次数	平均每天金标差异次数
1	0.832261	3005	54.63
2	0.693199	2302	41.85
3	0.813807	3260	59.27
4	0.821684	3332	60.58
5	0.801634	3194	58.07

由上表可以得知，$Lightgbm$模型与$XGBoost$模型相比效果基本相同，均是具有较高的准确率，但累计金标差异天数较大，说明高糖样本的检测准确率较低。

深度学习模型

1. LSTM 模型

使用 LSTM 模型，首先采用均值标准差进行数据标准化，然后定义 LSTM 模型和损失函数，接着定义训练函数并进行 5 折交叉验证，模型最终结果如下表所示：（选取 5 次实验）

表 3-9 LSTM 模型结果测试表

序号	test Average Loss	test rate	实际高糖发生次数	预测高糖发生次数	累计金标差异次数	平均每天金标差异次数
1	3.404205e-11	0.777713	2292	2569	429	7.8
2	2.450838e-11	0.775038	2457	2185	400	7.27
3	3.739548e-11	0.769216	2616	1475	1143	20.78
4	2.670791e-11	0.8003818	2220	1363	857	15.58
5	2.629416e-11	0.745771	2883	2260	683	12.42

由上表可以得出，LSTM 模型最终的累计金标差异天数下降了很多，即说明该模型对于高糖样本（正样本）的分类具有较好的效果。同时也说明我们设计的损失函数起到了很好的作用。

2. Transformer 模型

然后使用 Transformer 模型，采取与 LSTM 模型相同的处理方式，最终结果如下表所示：（选取 5 次实验）

表 3-10 Transformer 模型结果测试表

序号	test Average Loss	test rate	实际高糖发生次数	预测高糖发生次数	累计金标差异次数	平均每天金标差异次数
1	4.789279e-11	0.710542	3023	3321	400	7.27
2	4.390203e-11	0.844994	2319	397	1922	34.95
3	4.173727e-11	0.743062	2431	2892	565	10.27
4	4.119208e-11	0.789782	1996	2121	293	5.32
5	4.084695e-11	0.823598	2475	571	1904	34.62

由此结果可知，Transformer 模型的效果也不错，但与 LSTM 模型相比，稳定性稍弱，

偶尔容易出现累计金标差异天数过大的情况。

综上所述，进过对比，我们选取 LSTM 模型作为最终的模型。

3.3 计划和分工

3.3.1 整体计划

时间	事项
7.7-7.10	项目相关调研
7.11-7.15	清洗数据
7.16-7.20	数据分析并处理
7.20-7.23	构建 xgb 模型作为 baseline
7.24-7.27	训练模型并测试
7.28-8.5	模型优化，确定结果

3.3.2 团队分工

姓名	分工
张邱德	数据处理，模型训练
周俊丞	数据处理，模型训练
张志浩	数据处理，文档撰写
潘洁	项目调研，模型测试

4 参考资料

[1]汪涛. 基于集成学习的融合模型在血糖值预测中的应用研究[D].兰州理工大学,2021.DOI:10.27206/d.cnki.ggsgu.2021.000724.

[2]王延年,雍永强,贾晓灿等.融合 ARIMA 和 RBFNN 的血糖预测[J].计算机工程与设计,2015,36(06):1652-1656.DOI:10.16208/j.issn1000-7024.2015.06.046.

[3]谭金祥. 基于机器学习的胰岛素评估模型与算法研究[D].桂林电子科技大学,2022.DOI:10.27049/d.cnki.ggldc.2022.000582.

[4]王朱宇. 基于机器学习的血糖光谱数据挖掘研究[D].长春理工大学,2021.DOI:10.26977/d.cnki.gccgc.2021.000525.

[5]靳帅杰. 基于 AT-LSTM-GRU 的血糖预测模型研究[D].郑州大学,2022.DOI:10.27466/d.cnki.gzzdu.2022.003094.