**实验室报告**

- **基于LightGBM的结直肠癌风险预测模型**

**目的**

根据居民的历史就诊信息，预测患结直肠癌的概率。医疗管理部门可以筛选出预测风险最高的人群，以采取进一步的医疗措施。

**数据集以及基本分析**

本次研究共采集了萧山地区已建档的30周岁以上共1333323名居民的脱敏数据。在该数据中，共有5198（0.3899%）名居民确诊结直肠癌，1328125（99.6101%）名居民未确诊结直肠癌，基本分布如下图：

图表, 条形图

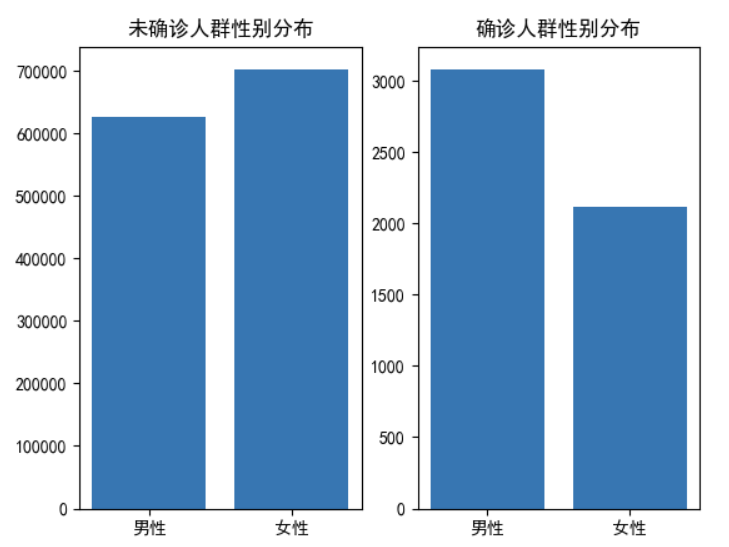
描述已自动生成

每份数据包含以下27条信息：

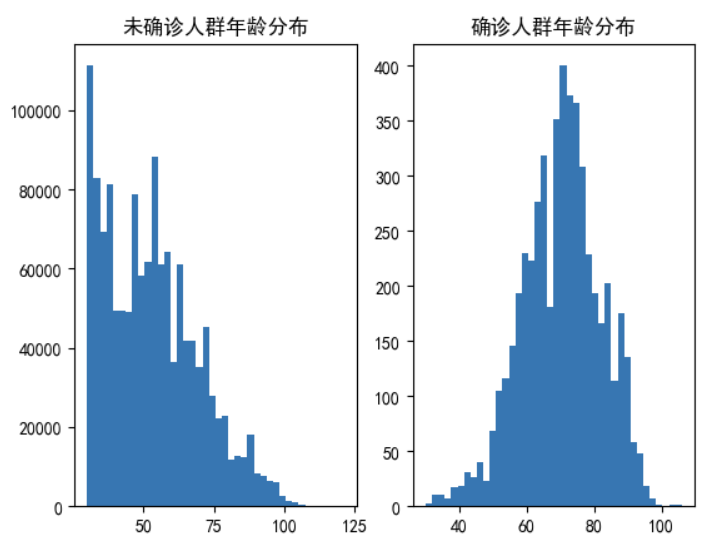
1. 年龄（30周岁以上）
2. 年龄分级
3. 性别
4. 一级亲属是否有结直肠癌史
5. 一级亲属是否有家族腺瘤息肉病病史
6. 本人是否有癌症史
7. 本人是否有肠道息肉史
8. 是否有慢性便秘
9. 是否有慢性腹泻
10. 是否有粘液血便
11. 是否有慢性阑尾炎
12. 是否有阑尾切除史
13. 是否有慢性胆道疾病史
14. 是否有胆囊切除史
15. 是否吸烟
16. BMI
17. 是否大量饮酒
18. 是否有溃疡性结肠炎
19. 是否有克罗恩病
20. 是否有糖尿病
21. 大便隐血是否阳性
22. 是否长期精神压抑
23. 癌胚抗原状态
24. 血红蛋白状态
25. 癌胚抗原值
26. 血红蛋白值
27. 是否确诊为结直肠癌

在非确诊人群和确诊人群中，相关特征分布如下所示：

1. 性别分布如下图



1. 年龄分布如下图：



1. BMI分布如下图：

图表, 直方图

描述已自动生成

1. 是否饮酒的人数分布如下图：

图表, 条形图, 直方图

描述已自动生成

1. 是否吸烟人数分布如下图：
2. 图表, 条形图, 直方图

   描述已自动生成
3. 癌胚抗原情况人数分布如下图：

图表, 条形图, 直方图

描述已自动生成

1. 血红蛋白情况人数分布如下图：

图表, 条形图

描述已自动生成

1. 大便隐血阳性情况人数分布如下图：

图表, 条形图, 直方图

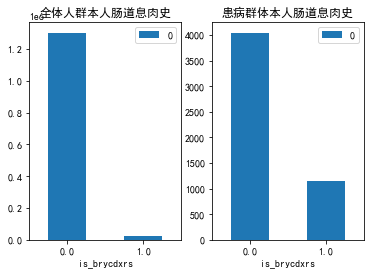
描述已自动生成

1. 是否有糖尿病人数对比如下图：

图表, 条形图, 直方图

描述已自动生成

1. 是否有肠道息肉史人数对比如下：



1. 一级亲属是否有结直肠癌史

图表, 条形图, 直方图

描述已自动生成

根据以上人数对比可以发现，每个风险因素在确诊与非确诊人群中的分布并不统一，存在较大差异，且个别风险因素在确诊人群和非确诊人群中的分布呈呈一致性，因此无法通过单一风险因素或组合风险因素来划分上述两类人群。为进一步验证以上观察，采用统计验证的方法判断每个风险因素在非确诊和确诊人群中是否存在统计学差异。统计验证结果如下：

| 字段 | P值 | 字段 | P值 |
| --- | --- | --- | --- |
| 年龄 | 0.000000e+00 | 是否吸烟 | 4.838373e-96 |
| 性别 | 1.393571e-69 | BMI | 0.000000e+00 |
| 一级亲属是否有结直肠癌史 | 4.396917e-04 | 是否大量饮酒 | 3.686322e-69 |
| 一级亲属是否有家族腺瘤息肉病病史 | nan | 是否有溃疡性结肠炎 | 3.564214e-34 |
| 本人是否有癌症史 | 0.000000e+00 | 是否有克罗恩病 | 2.467480e-07 |
| 本人是否有肠道息肉史 | 0.000000e+00 | 是否有糖尿病 | 2.496420e-221 |
| 是否有慢性便秘 | nan | 大便隐血是否阳性 | 0.000000e+00 |
| 是否有粘液血便 | 2.283622e-69 | 是否长期精神压抑 | 1.241835e-34 |
| 是否有阑尾切除史 | 3.337108e-07 | 癌胚抗原状态 | 0.000000e+00 |
| 是否有胆囊切除史 | 9.357556e-06 | 血红蛋白状态 | 7.838115e-67 |
| 是否有慢性胆道疾病史 | 0.000000e+00 | 是否有慢性阑尾炎 | 2.283622e-69 |
| 是否有慢性腹泻 | 4.322447e-73 |  |  |

除了是否有慢性便秘以及一级亲属是否有家族腺瘤息肉病病史外，其余特征P值均小于0.005，即不存在统计学上的显著性差异。

进一步进行相关性分析，得到如下图，与确诊人群相关系数排名前10的特征从高到低依次为本人有肠道息肉史、癌胚抗原值、是否慢性便秘、年龄、年龄分级、癌胚抗原状态、本人是否有癌症史、BMI、是否有糖尿病、是否有慢性胆道疾病史。

图片包含 图表

描述已自动生成

* **方法**

根据上述萧山地区的历史数据构建训练数据，以训练一个二分类的分类器。

* **训练过程**

1. 模型训练主要基于梯度提升决策树GBDT(graident boosting decesion tree)的分类模型LGBMClassifier。
2. 训练指标：由于我们最后是按照患病风险对病人从高到低进行排序，所以使用排序类别(ranking)的评价指标，即选取风险度由高到低排序的10000人（Top10000）来计算准确预测人数。
3. 训练、测试数据分割：70%的数据作为训练数据，30%的数据作为测试数据。且两份数据中正样本和负样本比例相当。
4. 模型训练：将上述训练数据以20%和80%的比例划分为验证集和训练集，模型在训练集上进行训练，在验证集上基于Top10000的准确预测人数来优化模型的训练参数。

* **评估方法**

评价方法是使用训练好的模型，预测测试数据中每个居民的结直肠癌风险度，并将该预测结果与实际是否结直肠癌确诊标记进行对比。效果评价指标主要使用前10000位、前50000位、前100000位、前150000位、前200000位的准确率。

在模型评估过程中使用以下数据集：

1. 上述地区数据集中除去训练数据以外的数据（测试集）

-**性能分析**

如下图为模型的排序类别评价指标，x轴为风险排序后的采样的top人数，左y轴为准确预测为患病的总人数，右侧y轴为相对于普筛规则的检出率。检出率达到90%、80%、70%、60%、50%需要采样人数分别为114000、67500、41350、27500、17500，符合模型的性能预期。

**图表, 条形图, 直方图

描述已自动生成**

-**评价结果**

上述地区数据集（共399997人，其中1594名结直肠癌确诊）

|  |  |  |
| --- | --- | --- |
| 预测风险最高的前 N 位居民 | 实际确诊人数 | 检出率  （实际确诊人数/总患病数） |
| 10000 | 644 | 40.40% |
| 50000 | 1185 | 74.34% |
| 100000 | 1390 | 87.20% |
| 150000 | 1502 | 94.23% |
| 200000 | 1561 | 97.93% |

与普筛模型对比

|  |  |  |
| --- | --- | --- |
| 筛选居民人数 | 实际确诊人数 | 检出率  （实际确诊人数/总患病数） |
| 普筛：220569 | 1528 | 95.86% |
| 算法：**162000** | 1528 | 95.86% |

总结：普筛需要对年龄大于等于50周岁的全部220569位居民进行相关检查，才能筛选出1528名结直肠癌患者。而使用模型预测，只需要对风险最高的大约前162000名居民进行检查，就可以定位普筛中的全部患者，相比于普筛规则减少了26.55%的筛查成本。

* **模型及预测结果相关解释性分析**

**模型解释性**

使用shap value模型对所训练的模型和得到的结果进行可解释性分析

下图为模型的权重，可以看到年龄、癌胚抗原状态、血红蛋白状态、是否大便隐血阳性、是否慢性便秘、BMI指数、是否吸烟、本人是否有肠道息肉史这9项指标权重较大，对模型的预测影响较大；后14项风险因素几乎对模型的输出几乎没有影响；

条形图

描述已自动生成

**表格

低可信度描述已自动生成**

进一步，我们分析每个特征对模型输出的影响的正负性，如下图；可以看到，较大的年龄、癌胚抗原值偏高、血红蛋白值偏低、大便隐血阳性（前5个特征右侧）会正向影响使模型输出，即模型以更高的概率预测其为患病；反之（前五个特征左侧），负向影响模型输出，即降低模型预测其为患病的概率；

**图表, 箱线图

描述已自动生成**

**预测结果解释性**

1. **带有检验检查的患病样本分析**

我们随机选取了5位确诊为结直肠癌的患者，来观测模型对这5位患者的预测结果并利用shap value 对其进行解释；下表为选取的0-4号患病患者样本：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本 | 0 | 1 | 2 | 3 | 4 |
| 身份证（脱敏） | 92512178 | 134df788 | 6b714a66 | ad3b6b5d | e89983d6 |
| 年龄（30周岁以上） | 69 | 75 | 78 | 77 | 74 |
| 年龄分级 | 1 | 1 | 1 | 1 | 1 |
| 性别 | 女性 | 男性 | 女性 | 女性 | 女性 |
| 一级亲属是否有结直肠癌史 | NaN | NaN | NaN | NaN | NaN |
| 一级亲属是否有家族腺瘤息肉病病史 | NaN | NaN | NaN | NaN | NaN |
| 本人是否有癌症史 | NaN | NaN | NaN | NaN | NaN |
| 本人是否有肠道息肉史 | NaN | 1.0 | 1.0 | 1.0 | 1.0 |
| 是否有慢性便秘 | 1.0 | 1.0 | 1.0 | 1.0 | NaN |
| 是否有慢性腹泻 | NaN | NaN | NaN | NaN | NaN |
| 是否有粘液血便 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性阑尾炎 | NaN | NaN | NaN | NaN | NaN |
| 是否有阑尾切除史 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性胆道疾病史 | NaN | NaN | NaN | NaN | NaN |
| 是否有胆囊切除史 | NaN | NaN | NaN | NaN | NaN |
| 是否吸烟 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| BMI | 34.380001 | 24.030001 | 20.6 | 27.469999 | 21.52 |
| BMI等级 | 1 | 0 | 0 | 1 | 0 |
| 是否大量饮酒 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 是否有溃疡性结肠炎 | NaN | NaN | NaN | NaN | NaN |
| 是否有克罗恩病 | NaN | NaN | NaN | NaN | NaN |
| 是否有糖尿病 | NaN | 1.0 | NaN | NaN | NaN |
| 大便隐血是否阳性 | NaN | NaN | NaN | NaN | NaN |
| 是否长期精神压抑 | NaN | NaN | NaN | NaN | NaN |
| 癌胚抗原状态 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 血红蛋白状态 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 癌胚抗原值 | 4631.93 | 81.65 | 97.22 | 510.05 | 1892.0 |
| 血红蛋白值 | 51.0 | 87.0 | 83.0 | 78.0 | 93.0 |
| 是否确诊为结直肠癌 | 是 | 是 | 是 | 是 | 是 |
| 预测为患病的概率 | **0.997752** | **0.997757** | **0.997772** | **0.997779** | **0.997780** |

每个样本对应的shape value解释性图如下，条形越长风险因子所占比例越高：

0号确诊样本中，模型预测其患病的概率为0.997752，模型对此的依据为：癌胚抗原值4631.93所占比重最大，年龄69周岁、血红蛋白值51.0、有慢性便秘等也作为其患病的主要风险因素。

日程表

描述已自动生成

1号确诊样本中，模型预测其患病的概率为0.997757，模型对此的依据为：癌胚抗原值81.65所占比重最大，年龄75周岁、血红蛋白值87.0、有肠道息肉史等也作为其患病的主要风险因素。

日程表

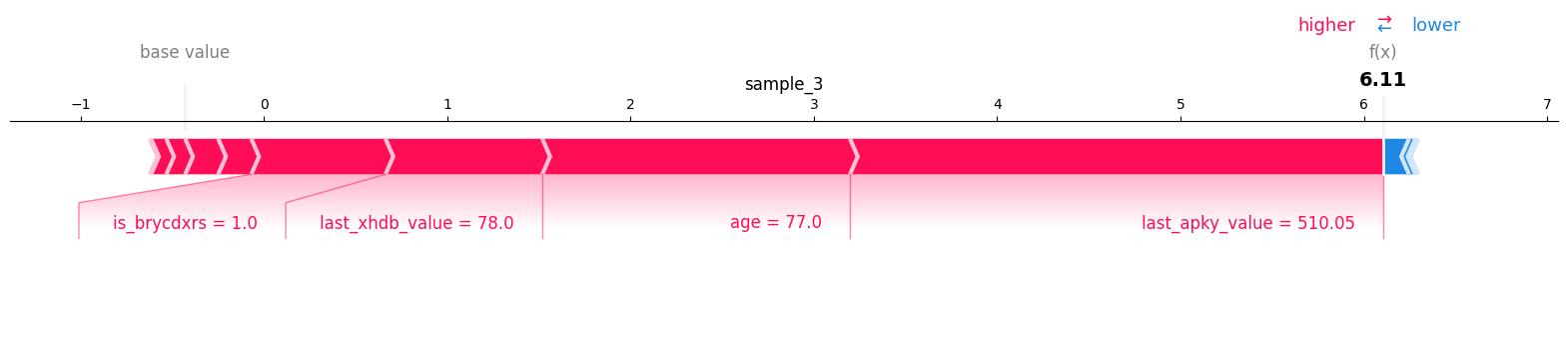
描述已自动生成

2号确诊样本中，模型预测其患病的概率为0.997772，模型对此的依据为：癌胚抗原值97.22所占比重最大，年龄78周岁、有肠道息肉史、血红蛋白值83.0、有慢性便秘等也作为其患病的主要风险因素。

日程表

描述已自动生成

3号确诊样本中，模型预测其患病的概率为0.997779，模型对此的依据为：癌胚抗原值510.05所占比重最大，年龄77周岁、有肠道息肉史、血红蛋白值78.0等也作为其患病的主要风险因素。



4号确诊样本中，模型预测其患病的概率为0.997780，模型对此的依据为：癌胚抗原值1892.0所占比重最大，年龄74周岁、有肠道息肉史、血红蛋白值93.0等也作为其患病的主要风险因素。

日程表

描述已自动生成

可以看到，模型对于这些有癌胚抗原值异常的样本，将其预测为患病的风险度较高，说明模型可以筛选出癌胚抗原值异常人群的能力；

1. 患病但无检验检查数据患者分析

我们随机选取了5位无检验检查数据但确诊为结直肠癌的患者，来观测模型对这5位患者的预测结果并利用shap value 对其进行解释；下表为选取的0-4号患病患者样本：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本 | 0 | 1 | 2 | 3 | 4 |
| 身份证（脱敏） | 77bc4eef | 9955e2b6 | 342b6cc9 | 779391c0 | f48f12c2 |
| 年龄（30周岁以上） | 84 | 65 | 87 | 74 | 74 |
| 年龄分级 | 1 | 1 | 1 | 1 | 1 |
| 性别 | 女性 | 男性 | 女性 | 女性 | 女性 |
| 一级亲属是否有结直肠癌史 | NaN | NaN | NaN | NaN | NaN |
| 一级亲属是否有家族腺瘤息肉病病史 | NaN | NaN | NaN | NaN | NaN |
| 本人是否有癌症史 | NaN | NaN | NaN | NaN | NaN |
| 本人是否有肠道息肉史 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性便秘 | 1.0 | 1.0 | NaN | 1.0 | 1.0 |
| 是否有慢性腹泻 | NaN | 1.0 | 1.0 | NaN | NaN |
| 是否有粘液血便 | 1.0 | NaN | NaN | 1.0 | 1.0 |
| 是否有慢性阑尾炎 | 1.0 | NaN | NaN | NaN | NaN |
| 是否有阑尾切除史 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性胆道疾病史 | NaN | NaN | NaN | NaN | NaN |
| 是否有胆囊切除史 | NaN | NaN | NaN | NaN | NaN |
| 是否吸烟 | NaN | NaN | NaN | NaN | NaN |
| BMI | NaN | NaN | NaN | NaN | NaN |
| BMI等级 | 0 | 1 | 1 | 0 | 0 |
| 是否大量饮酒 | NaN | NaN | NaN | 0.0 | 0.0 |
| 是否有溃疡性结肠炎 | NaN | NaN | NaN | NaN | NaN |
| 是否有克罗恩病 | NaN | NaN | NaN | NaN | NaN |
| 是否有糖尿病 | NaN | 1.0 | NaN | 1.0 | NaN |
| 大便隐血是否阳性 | NaN | NaN | NaN | NaN | NaN |
| 是否长期精神压抑 | NaN | NaN | NaN | NaN | NaN |
| 癌胚抗原状态 | NaN | NaN | NaN | NaN | NaN |
| 血红蛋白状态 | NaN | NaN | NaN | NaN | NaN |
| 癌胚抗原值 | NaN | NaN | NaN | NaN | NaN |
| 血红蛋白值 | NaN | NaN | NaN | NaN | NaN |
| 是否确诊为结直肠癌 | 是 | 是 | 是 | 是 | 是 |
| 预测为患病的概率 | **0.985764** | **0.985093** | **0.984073** | **0.98381** | **0.982312** |

每个样本对应的shape value解释性图如下：

0号确诊样本中，模型预测其患病的概率为0.985764，模型对此的依据为：年龄84周岁所占比重最大，本人有癌症史、本人有慢性便秘等也作为其患病的主要风险因素。

日程表

描述已自动生成

1号确诊样本中，模型预测其患病的概率为0.985093，模型对此的依据为：年龄65周岁所占比重最大，本人有癌症史、本人有肠道息肉史等也作为其患病的主要风险因素。

日程表

描述已自动生成

2号确诊样本中，模型预测其患病的概率为0.984073，模型对此的依据为：年龄65周岁所占比重最大，本人有肠道息肉史等也作为其患病的主要风险因素。

图形用户界面, 应用程序, Word

描述已自动生成

3号确诊样本中，模型预测其患病的概率为0.98381，模型对此的依据为：年龄74周岁所占比重最大，本人有癌症史、有慢性便秘等也作为其患病的主要风险因素。

日程表

描述已自动生成

4号确诊样本中，模型预测其患病的概率为0.982312，模型对此的依据为：年龄74周岁所占比重最大，本人有癌症史、有慢性便秘等也作为其患病的主要风险因素。

日程表

描述已自动生成

可以看到，虽然上述样本无确切的检验检查数据，模型也可以依据其他既往史等风险因素来筛选这一批风险人群。

1. 未患病但有检验检查数据样本分析

我们随机选取了5位有检验检查数据但并未被确诊为结直肠癌的疑似患者，来观测模型对这5位患者的预测结果并利用shap value 对其进行解释；下表为选取的0-4号患病患者样本：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 样本 | 0 | 1 | 2 | 3 | 4 |
| 身份证（脱敏） | 9115db5b | 9d4935d2 | 1ae5aef0 | abf507f3 | 3b918f5d |
| 年龄（30周岁以上） | 68 | 67 | 68 | 70 | 68 |
| 年龄分级 | 1 | 1 | 1 | 1 | 1 |
| 性别 | 女性 | 男性 | 男性 | 男性 | 女性 |
| 一级亲属是否有结直肠癌史 | NaN | NaN | NaN | NaN | NaN |
| 一级亲属是否有家族腺瘤息肉病病史 | NaN | NaN | NaN | NaN | NaN |
| 本人是否有癌症史 | NaN | NaN | NaN | NaN | NaN |
| 本人是否有肠道息肉史 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性便秘 | NaN | NaN | 1.0 | 1.0 | 1.0 |
| 是否有慢性腹泻 | 1.0 | NaN | NaN | NaN | NaN |
| 是否有粘液血便 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性阑尾炎 | NaN | NaN | NaN | NaN | NaN |
| 是否有阑尾切除史 | NaN | NaN | NaN | NaN | NaN |
| 是否有慢性胆道疾病史 | NaN | NaN | NaN | NaN | NaN |
| 是否有胆囊切除史 | NaN | NaN | NaN | NaN | NaN |
| 是否吸烟 | NaN | NaN | 0.0 | 1.0 | NaN |
| BMI | NaN | NaN | 21.01 | NaN | NaN |
| BMI等级 | 0 | 0 | 0 | 0 | 0 |
| 是否大量饮酒 | NaN | NaN | 0.0 | 0.0 | NaN |
| 是否有溃疡性结肠炎 | NaN | NaN | NaN | NaN | NaN |
| 是否有克罗恩病 | NaN | NaN | NaN | NaN | NaN |
| 是否有糖尿病 | NaN | NaN | NaN | 1.0 | 1.0 |
| 大便隐血是否阳性 | NaN | NaN | NaN | NaN | 0.0 |
| 是否长期精神压抑 | NaN | NaN | NaN | NaN | NaN |
| 癌胚抗原状态 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 血红蛋白状态 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 癌胚抗原值 | 29.67 | 30.58 | 32.68 | 309.62 | 29.17 |
| 血红蛋白值 | 86.0 | 65.0 | 102.0 | 108.0 | 83.0 |
| 是否确诊为结直肠癌 | 否 | 否 | 否 | 否 | 否 |
| 预测为患病的概率 | **0.997668** | **0.997657** | **0.997642** | **0.997589** | **0.997452** |

每个样本对应的shape value解释性图如下：

0号未确诊样本中，模型预测其患病的概率为0.997668，模型对此的依据为：癌胚抗原值29.67所占比重最大，年龄68周岁、血红蛋白值86也作为其患病的主要风险因素。

**日程表

描述已自动生成**

1号未确诊样本中，模型预测其患病的概率为0.997657，模型对此的依据为：癌胚抗原值30.58所占比重最大，年龄67周岁、血红蛋白值65也作为其患病的主要风险因素。

**日程表

描述已自动生成**

2号未确诊样本中，模型预测其患病的概率为0.997642，模型对此的依据为：癌胚抗原值32.68所占比重最大，年龄68周岁、血红蛋白值102、有慢性便秘也作为其患病的主要风险因素。

**日程表

描述已自动生成**

3号未确诊样本中，模型预测其患病的概率为0.997589，模型对此的依据为：癌胚抗原值309.62所占比重最大，年龄70周岁、血红蛋白值108、有慢性便秘也作为其患病的主要风险因素。

**日程表

描述已自动生成**

4号未确诊样本中，模型预测其患病的概率为0.997452，模型对此的依据为：癌胚抗原值29.17所占比重最大，年龄68周岁、血红蛋白值83、有慢性便秘也作为其患病的主要风险因素。

**日程表

描述已自动生成**

可以看到，虽然上述样本在建档初期并未未被诊断为确诊，但模型也可以依据其相关的风险因素来准确筛选出这一批未诊漏诊的风险人群以此来重点关注；

**结论和讨论**

* **从以上结论来看，模型效果要比传统的普筛更为准确，能够一定程度地减少筛选范围。**
* **模型输出的风险度与每个居民风险因素的多少、以及这些因素的权重正相关。也就是一个居民的风险因素越多、所占的权重越高、单一风险因子越离群，那么该居民的患有该并的风险度就越高，模型预测其为高风险（例如风险度高达90%以上）。目前尚未确诊的人群中，部分模型预测为高风险的人群可以通过筛选来重点关注。**
* **模型输出的前20万名高风险因素居民，并没有涵括所有的确诊患者。这可能是由于数据缺失导致。后续应该改进数据质量，例如减少空值或者直接去除空值较多的特征。**
* **模型部署的时候，可以根据部署区域的实际数据对模型进行进一步微调，以改善模型性能。**

**古珀研究院 何嘉华**

**2023.03.21**