### 分析流程 数据源： \_data\_processing2\_updated3\_归一化.csv 算法配置： 算法： 逻辑回归 变量： 变量Y：{ Finalist\_Weight }；变量X：{ Average Age mean，Average Net Result mean，Average Mark var，current\_match\_ranking mean，Years to Olympics，current\_match\_ranking min，Nation mean，Average Mark min }收起 分析结果： 逻辑回归是用于解决因变量为二分类变量的回归：模型的似然比卡方检验的结果显示，显著性P值0.000\*\*\*，水平上呈现显著性，拒绝原假设，因而模型是有效的。

### 分析步骤 1. 对分类因变量分布状况进行描述。 2. 对模型进行似然比卡方检验，分析似然检比卡方显著性，若拒绝原假设(P<0.05)，说明模型有效，反之模型不成立，若设计多个模型，可以结合其他分类评价或者信息准则（AIC、BIC值越小越好）进行综合分析。 3. 根据模型参数表，分析X是否呈显著性(P<0.05)，用于探究X对Y的影响关系。 4. 分析回归系数B与OR值（优势比），对比分析X对Y的影响程度。 5. 结合预测分类混淆矩阵与模型评价中的分类指标，分析模型预测。

### 详细结论

**输出结果1：二分类因变量基本汇总**

|  |  |  |  |
| --- | --- | --- | --- |
| 因变量 | 选项 | 频数 | 百分比(%) |
| Finalist\_Weight | 0.0 | 452 | 68.072 |
| 1.0 | 212 | 31.928 |
| 总计 | 664 | 100 |

**图表说明：**

上表展示了因变量各分组的分布情况。  
● 选项：当前字段数据下的去重类别  
● 频数：当前去重类别在数据中出现的次数  
● 百分比：当前去重类别的频数占比  
● 当因变量分类水平的数据量出现严重不平衡时，建议对数据进行过采样或者欠采样

**输出结果2：模型评价**

|  |  |  |  |
| --- | --- | --- | --- |
| 似然比卡方值 | P | AIC | BIC |
| 665.834 | 0.000\*\*\* | 683.834 | 724.319 |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | |

**图表说明：**

上表展示了模型评价指标，可用于对模型的表现进行评估或有效性进行验证，其包括似然比检验、P值、AIC值、BIC值。  
● 对P值进行分析，如果该值小于0.05，则说明模型有效；反之则说明模型无效。  
● AIC 值和BIC 值用于对比两个模型的优劣时使用，此两个值均为越小越好。

**智能分析：**

模型的似然比卡方检验的结果显示，显著性P值0.000\*\*\*，水平上呈现显著性，拒绝原假设，因而模型是有效的。

**输出结果3：二分类逻辑回归结果**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验组=1.0 | 回归系数 | 标准误差 | Wald | P | OR | OR值95%置信区间 | |
| 上限 | 下限 |
| 常数 | 31.315 | 19.497 | 2.58 | 0.108 | 39799022657617.1 | 0.001 | 1.5707818554187137e+30 |
| Average Age mean | 0.203 | 0.023 | 80.963 | 0.000\*\*\* | 1.225 | 1.172 | 1.28 |
| Average Net Result mean | -2.609 | 1.319 | 3.91 | 0.048\*\* | 0.074 | 0.006 | 0.977 |
| Average Mark var | -32.704 | 22.951 | 2.031 | 0.154 | 0 | 0 | 215271.095 |
| current\_match\_ranking mean | -0.126 | 0.124 | 1.026 | 0.311 | 0.882 | 0.692 | 1.125 |
| Years to Olympics | 0.227 | 0.08 | 8.138 | 0.004\*\*\* | 1.255 | 1.074 | 1.467 |
| current\_match\_ranking min | 0.52 | 0.17 | 9.303 | 0.002\*\*\* | 1.682 | 1.204 | 2.349 |
| Nation mean | 0.025 | 0.012 | 4.308 | 0.038\*\* | 1.025 | 1.001 | 1.049 |
| Average Mark min | -4.011 | 1.552 | 6.678 | 0.010\*\*\* | 0.018 | 0.001 | 0.38 |
| 因变量：Finalist\_Weight | | | | | | | |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | |

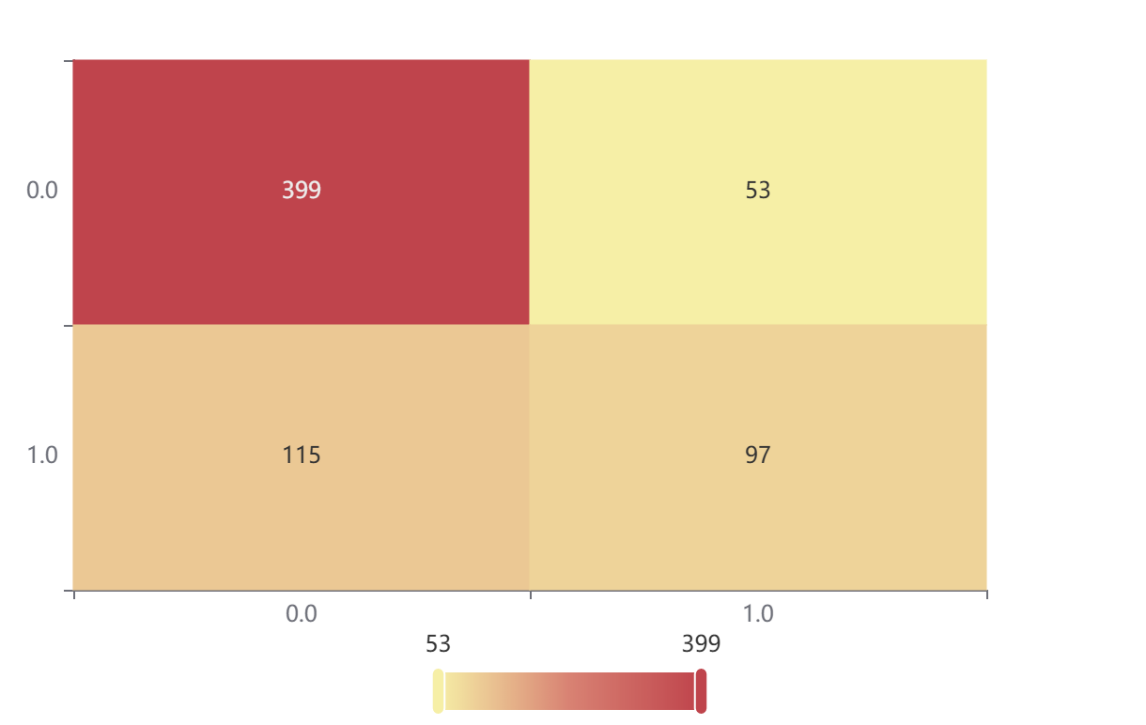
**图表说明：**

上表展示了模型的参数结果。包括模型的系数、标准误差、OR值、置信区间等用于分析模型的公式。  
● OR值（优势比）：为实验组的事件发生概率/对照组的事件发生概率。  
● 对于连续自变量的OR值的意义为：该变量每升高一个单位，发生实验组事件的概率比发生对照组事件的概率变化了（OR值-1）%。  
● 对于哑变量化的0-1分类自变量的OR值意义为：该变量每升高一个单位（即分类水平从0变为1），发生实验组事件的概率比发生对照组事件的概率变化了（OR值-1）%。

**智能分析：**

二分类逻辑回归的结果显示:  
字段常数显著性P值为0.108，水平上不呈现显著性，不能拒绝原假设，因此常数不会对Finalist\_Weight产生显著性影响。  
字段Average Age mean显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，因此Average Age mean会对Finalist\_Weight产生显著性影响，意味着Average Age mean每增加一个单位，Finalist\_Weight为1.0的概率比0.0的概率高了22.452%。  
字段Average Net Result mean显著性P值为0.048\*\*，水平上呈现显著性，拒绝原假设，因此Average Net Result mean会对Finalist\_Weight产生显著性影响，意味着Average Net Result mean每增加一个单位，Finalist\_Weight为1.0的概率比0.0的概率低了92.639%。  
字段Average Mark var显著性P值为0.154，水平上不呈现显著性，不能拒绝原假设，因此Average Mark var不会对Finalist\_Weight产生显著性影响。  
字段current\_match\_ranking mean显著性P值为0.311，水平上不呈现显著性，不能拒绝原假设，因此current\_match\_ranking mean不会对Finalist\_Weight产生显著性影响。  
字段Years to Olympics显著性P值为0.004\*\*\*，水平上呈现显著性，拒绝原假设，因此Years to Olympics会对Finalist\_Weight产生显著性影响，意味着Years to Olympics每增加一个单位，Finalist\_Weight为1.0的概率比0.0的概率高了25.506%。  
字段current\_match\_ranking min显著性P值为0.002\*\*\*，水平上呈现显著性，拒绝原假设，因此current\_match\_ranking min会对Finalist\_Weight产生显著性影响，意味着current\_match\_ranking min每增加一个单位，Finalist\_Weight为1.0的概率比0.0的概率高了68.174%。  
字段Nation mean显著性P值为0.038\*\*，水平上呈现显著性，拒绝原假设，因此Nation mean会对Finalist\_Weight产生显著性影响，意味着Nation mean每增加一个单位，Finalist\_Weight为1.0的概率比0.0的概率高了2.496%。  
字段Average Mark min显著性P值为0.010\*\*\*，水平上呈现显著性，拒绝原假设，因此Average Mark min会对Finalist\_Weight产生显著性影响，意味着Average Mark min每增加一个单位，Finalist\_Weight为1.0的概率比0.0的概率低了98.189%。

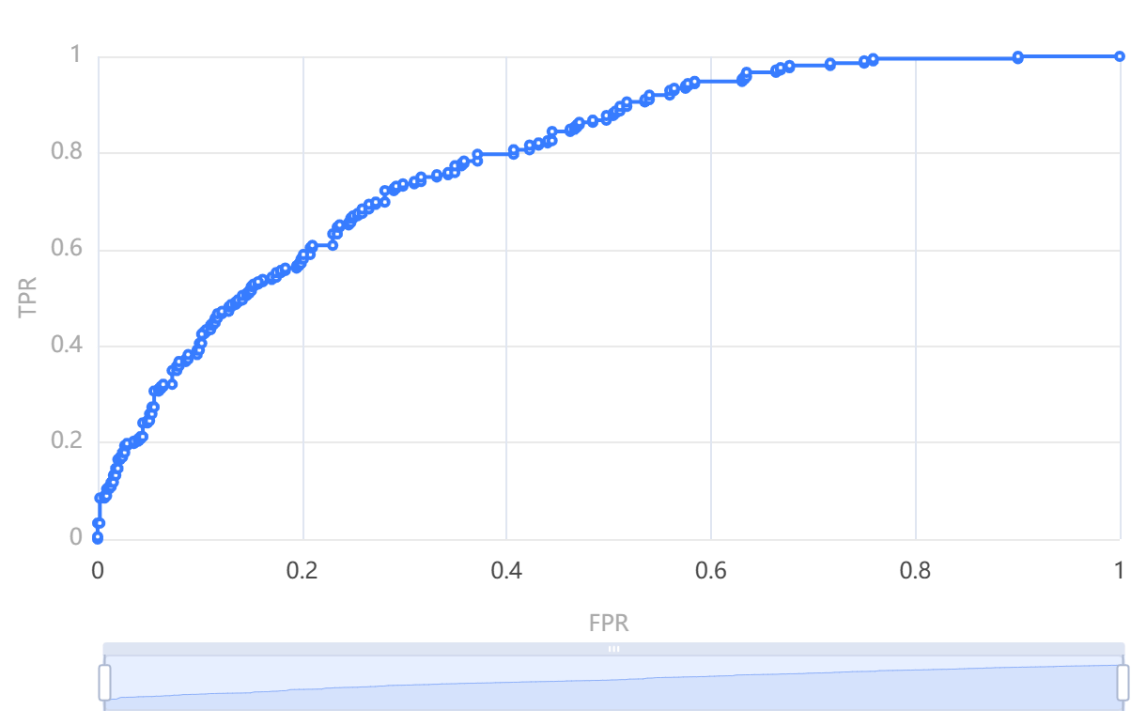
**输出结果4：混淆矩阵热力图**



**图表说明：**

上表以热力图的形式展示了混淆矩阵。

**输出结果5：ROC曲线**



**图表说明：**

上图展示了ROC曲线图，用于衡量逻辑回归的分类效果。  
● ROC曲线图把灵敏度（TPR）和特异度（FPR）结合，可以同时衡量两者关系。理想情况下，TPR应该接近1，FPR应该接近0。  
● 灵敏度：实际为正样本的结果中，预测为正样本的比例。  
● 特异度：实际为负样本的结果中，预测为正样本的比例。

**输出结果6：分类评价指标**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 准确率 | 召回率 | 精确率 | F1 | AUC |
| 0.747 | 0.747 | 0.735 | 0.733 | 0.79 |

**图表说明：**

上表中展示了分类评价指标，进一步通过量化指标来衡量逻辑回归的分类效果。  
● 准确率：预测正确样本占总样本的比例，准确率越大越好。  
● 召回率：实际为正样本的结果中，预测为正样本的比例，召回率越大越好。  
● 精确率：预测出来为正样本的结果中，实际为正样本的比例，精确率越大越好。  
● F1：精确率和召回率的调和平均，精确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是精确率高、召回率就低，或者召回率低、但精确率高。若需要兼顾两者，那么就可以用F1指标。  
● AUC：AUC值越接近1说明分类效果越好。  
注：精确率、召回率、F1值、AUC值是分别度量分类器对某一类别预测结果的评价指标值，后续再对所有类别根据样本进行加权平均后得到整体评价指标值。

### 参考文献 [1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com. [2] 袁翔宇,张蓬鹤,熊素琴,等. 基于逻辑回归算法的异常用电辨识方法研究[J]. 电测与仪表,2021,58(12):81-87.