目录

[1. 数据概览 2](#_Toc166151662)

[2. 数据处理 2](#_Toc166151663)

[3. 特征探索与可视化 2](#_Toc166151664)

[4. 建模：Logistic 回归、决策树和 XGBoost 2](#_Toc166151665)

[5. 模型评估：混淆矩阵、ROC 曲线 2](#_Toc166151666)

[6. 特征重要性 4](#_Toc166151667)

[7. 最终模型选择 5](#_Toc166151668)

1. 数据概览

从数据集中导入了两个数据文件：**train.csv** 和 **test.csv**。训练集共包含 36 列，包括年龄、日工资、部门、教育程度等多个特征。数据总计 1176 条记录，测试集与训练集数据结构相同。目标变量是 **Attrition**，用于表示员工的离职状态。

2. 数据处理

* **缺失值**：数据集整体上没有发现缺失值。
* **重复值**：没有发现重复记录。
* **异常值**：通过箱线图和标准差检测法检测出部分异常值，已进行了过滤。

3. 特征探索与可视化

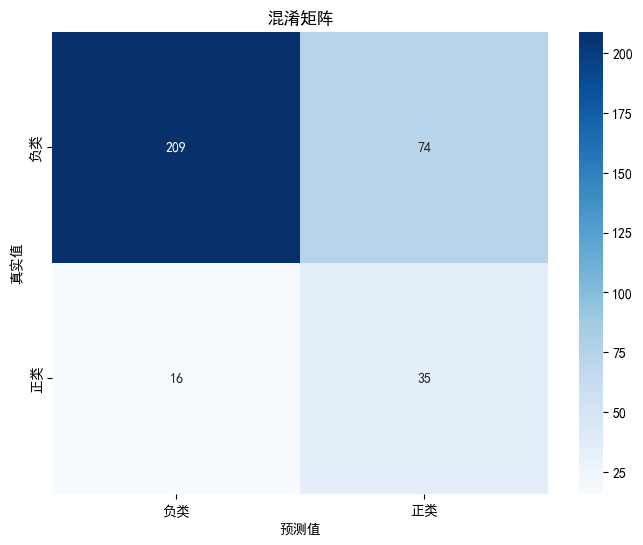
* **数值型特征**：特征包括年龄、月收入、工作年限等，通过分布图和箱线图可视化，离职员工通常工作年限较短。
* **分类特征**：如部门、教育背景、婚姻状况等。婚姻状态为“单身”、经常加班的员工离职率较高。

4. 建模：Logistic 回归、决策树和 XGBoost

* **Logistic 回归**：通过网格搜索调整了正则化参数和优化算法，并使用交叉验证确保模型的稳健性。最终模型在测试集上表现稳定。
* **决策树**：通过调整最大深度和分裂节点参数，在测试集中获得了较高的准确性，并具有较强的可解释性。
* **XGBoost**：调节学习率、树的深度、正则化参数等。该模型通过强大的特征选择和拟合能力，对数据集进行了深入挖掘，准确度最高。

5. 模型评估：混淆矩阵、ROC 曲线

* **混淆矩阵**：比较了各模型在离职预测的混淆矩阵，揭示了准确度和误报率的区别。
  + **Logistic 回归**：



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **指标** | 准确率 | 精确率 | 召回率 | F1 分数 |
| **值** | 0.7305 | 0.3211 | 0.6863 | 0.4375 |

* + **决策树**：

图表, 树状图

描述已自动生成

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **指标** | 准确率 | 精确率 | 召回率 | F1 分数 |
| **值** | 0.6796 | 0.2586 | 0.5882 | 0.3593 |

* + **XGBoost**：

图表

描述已自动生成

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 指标 | 准确率 | 精确率 | 召回率 | F1 分数 |
| 值 | 0.8323 | 0.439 | 0.3529 | 0.3913 |

* **ROC 曲线**：各模型的 AUC 分数如下：
  + **Logistic 回归**：约 0.75
  + **决策树**：约 0.66
  + **XGBoost**：约 0.77

6. 特征重要性

* **Logistic 回归**：通过回归系数评估特征重要性。overtime、JobRole\_Manufacturing Director、MaritalStatus\_Single等特征对离职影响较大。
* **决策树**：通过 **feature\_importances\_** 提取特征重要性。OverTime、YearsWithCurrManager、JobSatisfaction的影响显著。
* **XGBoost**：通过 **feature\_importances\_** 提取特征重要性。OverTime、MaritalStatus\_Single、YearsWithCurrManager等特征在模型中表现最为重要。



Figure 1: Top 10 important features

7. 最终模型选择

* **选择标准**：综合考虑模型的准确度、可解释性和执行效率，最终选择 XGBoost 模型进行预测。
* **原因**：XGBoost 模型在 AUC 和准确性方面表现最佳，并且通过调节参数具备更好的泛化能力。