### 分析流程 数据源： predata.xlsx 算法配置： 算法： 随机森林分类 变量： 变量X:{12、以下因素对您选择报名考研机构的影响程度是？—学科基础，学习氛围，线下地点距离，交通是否方便，师资力量，培训费用，自主学习能力，规划能力，他人影响，上岸率，口碑}；变量Y:{7、如果打算考研，您认为是否有必要报考研培训机构？}收起 分析结果： 随机森林分类基于准确率、召回率、精确率、F1指标对模型进行评价，请看详细结论。

### 分析步骤 1. 通过训练集数据来建立随机森林分类模型。 2. 通过建立的随机森林来计算特征重要性。 3. 将建立的随机森林分类模型应用到训练、测试数据，得到模型的分类评估结果。 4. 由于随机森林中具有随机性，每次运算的结果不一样，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算分类。 5. 注：随机森林无法像传统模型一样得到确定的方程，通常通过测试数据分类效果来对模型进行评价。

### 详细结论

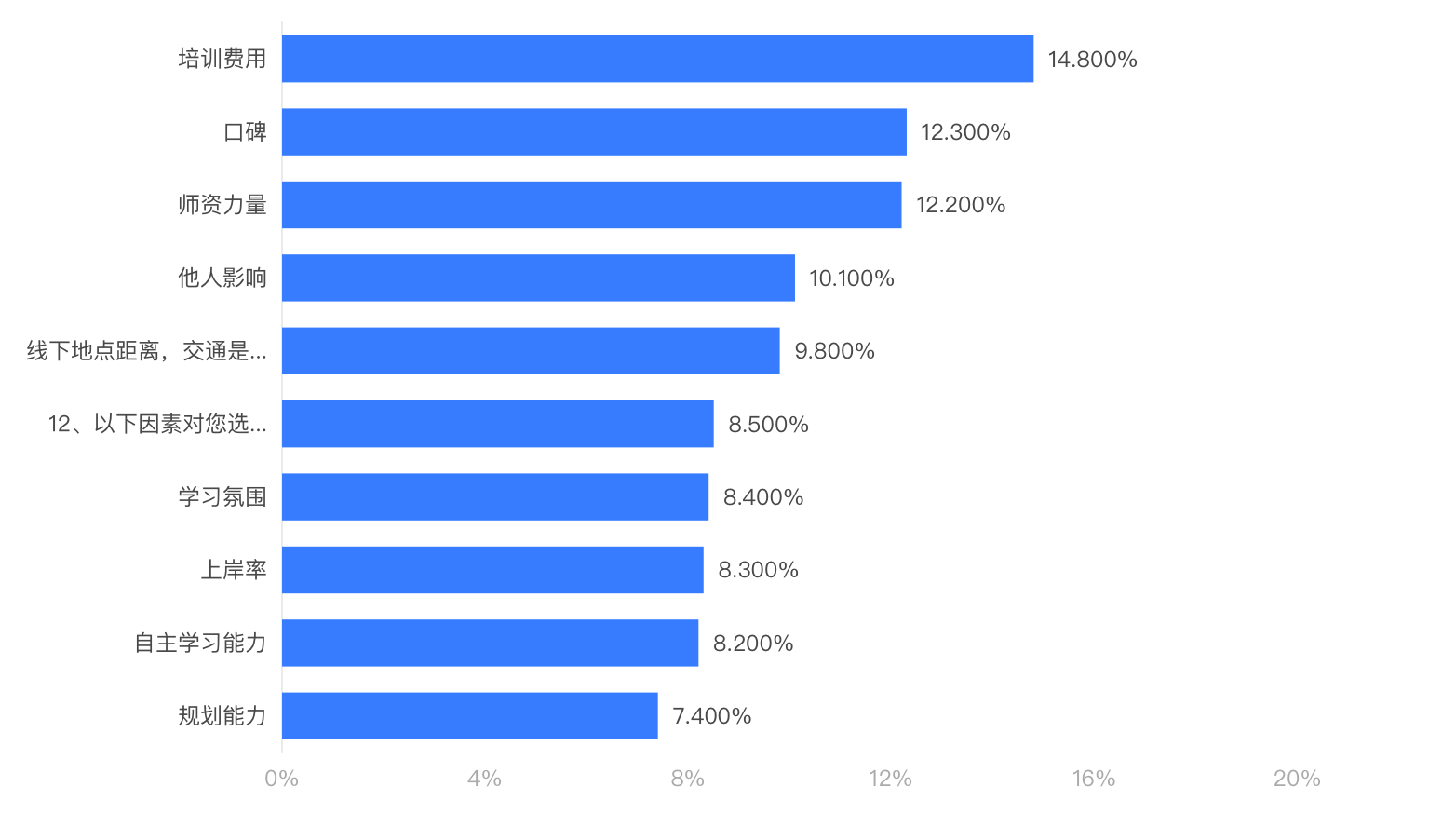
**输出结果1：模型参数**

|  |  |
| --- | --- |
| 参数名 | 参数值 |
| 训练用时 | 0.186s |
| 数据切分 | 0.7 |
| 数据洗牌 | 否 |
| 交叉验证 | 否 |
| 节点分裂评价准则 | gini |
| 决策树数量 | 100 |
| 有放回采样 | true |
| 袋外数据测试 | false |
| 划分时考虑的最大特征比例 | auto |
| 内部节点分裂的最小样本数 | 2 |
| 叶子节点的最小样本数 | 1 |
| 叶子节点中样本的最小权重 | 0 |
| 树的最大深度 | 10 |
| 叶子节点的最大数量 | 50 |
| 节点划分不纯度的阀值 | 0 |

**图表说明：**

上表展示了模型各项参数配置以及模型训练时长。

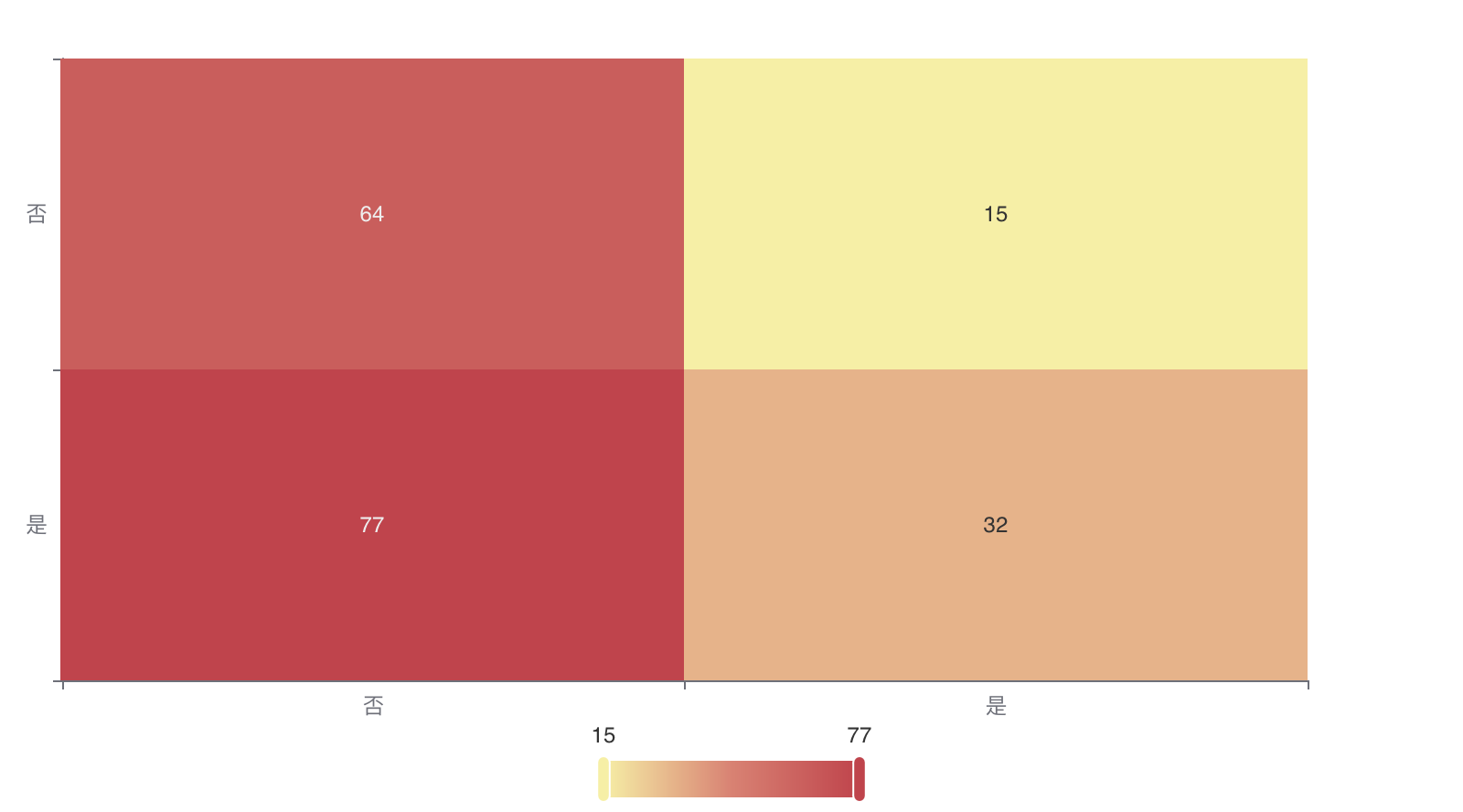
**输出结果2：特征重要性**



**图表说明：**

上柱形图或表格展示了各特征（自变量）的重要性比例。

**输出结果3：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4：模型评估结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 准确率 | 召回率 | 精确率 | F1 |
| 训练集 | 0.817 | 0.817 | 0.831 | 0.817 |
| 测试集 | 0.511 | 0.511 | 0.585 | 0.482 |

**图表说明：**

上表中展示了训练集和测试集的分类评价指标，通过量化指标来衡量随机森林对训练、测试数据的分类效果。 ● 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。 ● oob\_score：对于分类问题，oob\_score是袋外数据的准确率。若在建立树过程中选择有放回抽样时，大约1/3的记录没有被抽取。没有被抽取的自然形成一个对照数据集，可用于模型的验证。所以随机森林不需要另外预留部分数据做交叉验证，其本身的算法类似交叉验证，而且袋外误差是对预测误差的无偏估计（当算法参数选择了“袋外测试数据”后，才会通过oob\_score来检验模型的泛化能力）。

**输出结果5：预测结果 下载**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 预测结果Y | 7、如果打算考研，您认为是否有必要报考研培训机构？ | 预测测试结果概率\_否 | 预测测试结果概率\_是 | 12、以下因素对您选择报名考研机构的影响程度是？—学科基础 | 学习氛围 | 线下地点距离，交通是否方便 | 师资力量 | 培训费用 | 自主学习能力 | 规划能力 | 他人影响 | 上岸率 | 口碑 |
| 是 | 否 | 0.4255522469321988 | 0.5744477530678012 | 2 | 3 | 3 | 4 | 3 | 4 | 4 | 3 | 3 | 4 |
| 否 | 否 | 0.9653476678136412 | 0.034652332186358534 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 1 | 5 | 5 |
| 否 | 是 | 0.5039493285007733 | 0.4960506714992266 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 否 | 否 | 0.8482462767773213 | 0.15175372322267802 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 否 | 是 | 0.576506158877324 | 0.423493841122676 | 5 | 5 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 否 | 是 | 0.7434136286526167 | 0.25658637134738316 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 5 | 5 |
| 否 | 否 | 0.7927850642467541 | 0.20721493575324587 | 4 | 3 | 2 | 4 | 4 | 4 | 3 | 1 | 4 | 4 |
| 否 | 是 | 0.5644954341100096 | 0.4355045658899903 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 |
| 否 | 否 | 0.8482462767773213 | 0.15175372322267802 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 否 | 否 | 0.9741713154092722 | 0.02582868459072764 | 5 | 4 | 5 | 5 | 5 | 5 | 3 | 3 | 5 | 5 |
| 否 | 否 | 0.8482462767773213 | 0.15175372322267802 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 否 | 是 | 0.8220012606502068 | 0.17799873934979316 | 4 | 4 | 3 | 4 | 5 | 4 | 5 | 4 | 5 | 5 |
| 否 | 否 | 0.5644954341100096 | 0.4355045658899903 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 |
| 否 | 否 | 0.5039493285007733 | 0.4960506714992266 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 否 | 否 | 0.769988910279831 | 0.230011089720169 | 5 | 3 | 5 | 5 | 4 | 4 | 4 | 2 | 4 | 3 |

**图表说明：**

上表格为预览结果，只显示部分数据，全部数据请点击下载按钮导出。 上表展示了随机森林模型对测试数据的分类结果，分类结果值是拥有最大预测概率的分类组别。

### 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] 周志华. 机器学习[M]. 清华大学出版社, 2016.