**【参考答案】**

做描述性统计分析，有缺失值（2分），需要处理，有部分异常值（2分），考虑真实情况，不予处理，量纲不一致，暂不做处理（2分）

连续变量的描述性统计（2分），数据都有一定的偏差，暂不做处理

样本较均衡较差（>4:1）暂不处理，先看结果再考虑进一步处理（3分）









标签编码：（2分）

将2分类类别型字段进行标签编码处理，包括联系方式和是否购买2个字段

独热编码：（2分）

将多分类文本型字段进行独热编码（或哑变量）处理，最好选择哑变量，或者是独热编码后去掉最后一个字段进入模型，包括婚姻状况、信用卡是否有违约、是否有房贷、是否有其他贷款、上一次联系是周几、之前活动结果6个字段



**（2）**判断婚姻状况、年龄与是否购买有无关联

原假设：婚姻状况和是否购买没有显著关联；

备择假设：婚姻状况和是否购买有显著关联。

卡方检验分析结果：



P值为1.0053107489917588e-10，小于0.05，拒绝原假设。即婚姻状况和是否购买有显著关联。

（原假设备择假设合理4分，p值2分，结果分析合理并且结论正确2分）

原假设：年龄和是否购买没有显著关联；

备择假设：年龄和是否购买有显著关联。

单因素方差分析结果：



P值为0.0002734643957663127，小于0.05，拒绝原假设。年龄和是否购买有显著关联。

（原假设备择假设合理4分，p值2分，结果分析合理并且结论正确2分）



**（3）**

以是否购买作为因变量，其他变量作为自变量，构造分类模型。

逻辑回归（参数默认）：（选择一种分类算法2分+模型结果展示2分+模型结果分析4分）

训练集：



逻辑回归结果欠拟合

决策树：深度为6，其他默认（选择一种分类算法2分+模型结果展示2分+模型结果分析4分）

训练集：



决策树结果也欠拟合

综合来看，逻辑回归和决策树的结果都不太令人满意，需要进一步优化模型

优先选择调整样本均衡性，选择过采样（SMOTE抽样，欠采样也得分），重新进行分类模型构建（2分）

结果如下：

逻辑回归：





比样本均衡性之前泛化能力有所提升

决策树：





比样本均衡性之前泛化能力有所提升

调整好的模型泛化能力得到了极大地提升，两种方法在解释力上决策树略占优势，逻辑回归可通过回归系数进行解释，而决策树可以通过树的形状来解释分类规则。（4分）



**（4）**

解释是否购买的影响因素：（2分）



系数为正代表，该变量取值越高，越倾向于购买，系数为负，则该变量取值越高，越不会购买

建议：（3分）

年龄越高，购买的可能性越高

已婚人士购买的可能性较低

信用卡由违约的客户购买的可能性较高

无房贷的客户购买的可能性较高

之前营销活动成功的客户购买的可能性更高

结合上述结论，给出以下建议：

多元市场细分：除了关注信用卡违约客户外，还可以考虑其他客户群体。例如，根据年龄、婚姻状况、就业情况等因素对市场进行细分。

定制化营销策略：对于不同细分市场，可以设计定制化的营销策略。例如，对于年轻人群体，可能更倾向于使用数字渠道和社交媒体进行推广。

产品创新：考虑创新金融产品以满足不同客户群体的需求。例如，为已婚人士提供家庭财务规划服务，或为高就业变动率行业的人士提供灵活的信贷产品。

客户关系管理：加强客户关系管理，提高客户满意度和忠诚度。例如，通过定期沟通和个性化服务来维护客户关系。

风险控制：在推广金融产品的同时，注意风险控制，特别是对于信用记录不良的客户。

数据驱动的决策：继续使用数据分析来指导营销策略，包括A/B测试和客户行为分析。