

**专业实践课程论文**

|  |  |
| --- | --- |
| **题 目：** | 信用卡用户违约预测 |
|  |  |
| **学 院：** | 数据科学学院 |
| **专 业：** | 数据科学与大数据技术 |
| **年级班别：** | 2021级（1）班/ |
| **学 号：** | 421470102 |
| **学生姓名**： | 贝梓健 |

**提交日期：** 2024 **年** 12 月

# 

# 目录

[实用机器学习课程论文 1](#_Toc23408)

[目录 2](#_Toc3366)

[绪论 3](#_Toc27553)

[1. 认识信用卡客户数据集 3](#_Toc21961)

[2数据探索 4](#_Toc23251)

[2.1 快速检视数据 4](#_Toc21264)

[2.2 数据是否有缺失 6](#_Toc7788)

[2.3 数据样本平衡吗 6](#_Toc32516)

[3 数据预处理 ...................................................... 6](#_Toc26371)

[3.1 处理客户教育程度等分类变量................. 6](#_Toc26371)

[3.2 定义预测变量 7](#_Toc27231)

[3.3 划分训练集和测试集 8](#_Toc10543)

[4 模型建立和评估 9](#_Toc21444)

[4.1 kNN算法模型建立 9](#_Toc9325)

[4.2 决策数据算法模型建立 10](#_Toc28903)

[4.3 xgboost算法模型建立 10](#_Toc259)

[4.4 模型评估 11](#_Toc7491)

[5. 总结 12](#_Toc16354)

[得 分 12](#_Toc29232)

# 

# 绪论

本案例基于台湾地区信用卡客户数据，包括客户个体特征、信用卡额度、账单及还款情况等信息，建立多个模型对客户下个月是否会违约做出预测。通过学习这份案例，你将掌握对数据集的探索分析、数据预处理、利用sklearn构建和评估多种分类模型。

# [1. 认识信用卡客户数据集](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "1)

要认识一个信用卡客户数据集，首先需要对数据集中的各个字段进行理解。通常，信用卡客户数据集会包含以下类型的信息：

个体特征: 这些是描述客户的人口统计学特征，可能包括：

年龄 性别 婚姻状况 教育程度 职业

信用历史: 描述客户过去的信用行为和信用记录，例如：

信用额度 购买金额 提现金额 支付金额 账单金逾期次数 违约记录

还款情况: 反映客户的还款习惯和能力，如：

最小还款额 实际还款额 是否全额还款 还款频率

其他金融产品使用情况: 客户可能还拥有的其他金融产品，如储蓄账户、投资产品等。

违约标识: 目标变量，表明客户在下个账期是否违约（是/否）。

了解数据集之后，可以通过以下步骤进一步分析：

相关性分析: 计算特征之间的相关系数，了解哪些特征与违约行为有强相关性。

缺失值处理: 检查数据集中是否存在缺失值，并决定如何填补或处理这些缺失值。

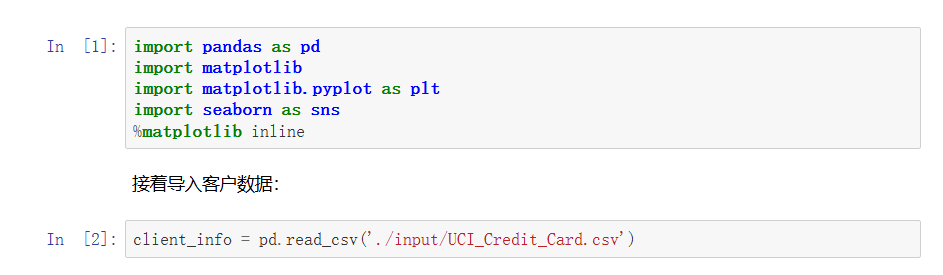
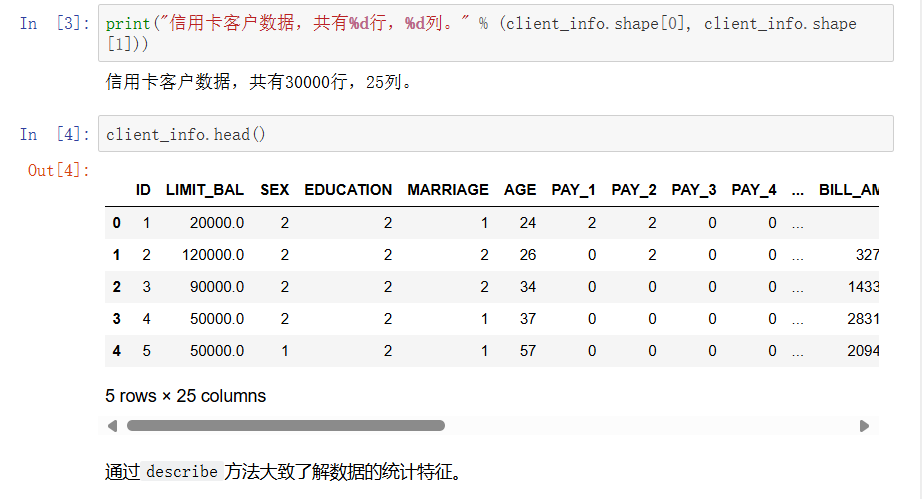
异常值检测: 查找并处理可能的异常值或离群点。

图1-1导入数据

图1-2统计特征 通过上述步骤，可以对信用卡客户数据集有一个全面的认识，为后续的数据预处理和模型建立打下基础。

# [2数据探索](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "3)

## [2.1 快速检视数据](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "3.1)

快速检视数据通常是指在不进行深入分析的情况下，迅速了解数据集的基本情况。这有助于确定数据的完整性、一致性和结构。以下是一些建议的步骤：

**查看数据集的大小**：了解数据集包含多少行和列，以及是否有缺失值。

检查数据类型：查看每个列的数据类型，例如数值、字符串或日期。

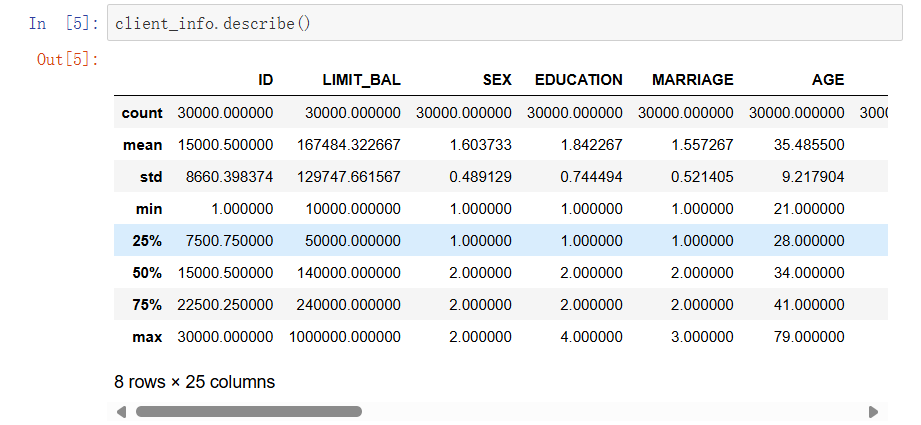
**统计概要**：计算各列的平均值、中位数、最小值和最大值等统计量。

**分布情况**：查看数据的分布情况，例如正态分布、偏态分布等。

**类别变量的频率**：对于分类变量，查看各个类别的频率分布。

**相关性分析**：检查各个变量之间的相关性，以了解它们之间的关系。

通过以上步骤，可以快速了解数据集的基本特征，为后续的数据分析和建模工作奠定基础。

 图2-1数据统计

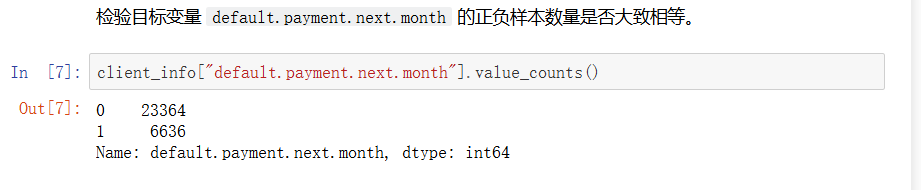
通过上面的统计结果，我们知道数据集中一共有30,000个信用卡客户。从信用额度看，平均信用额度是167,484台币，最小的是1万台币，最大的达到了100万台币，波动非常大；从年龄来看，平均年龄是35.4岁。 default.payment.next.month 的平均值为0.221，由于下个月不违约记为0，违约记为1，所以平均值0.221意味着有22.1%的客户下个月将会违约（会在接下来的分析中验证）。

## [2.2 数据是否有缺失](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "3.2)

图2-2检查数据

数据非常完整，所有的变量都不存在缺失。

## [2.3 数据样本平衡吗](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "3.3)

图2-3检验数据

有6636个客户下个月将会违约，占比22%，远小于未违约用户数量，所以数据中存在着明显的正负样本不平衡问题。

[3 数据预处理](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "4)

[3.1 处理客户教育程度等分类变量](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "4.1)

在处理客户数据时，教育程度这类分类变量的处理方式至关重要。以下将详细探讨如何有效处理包括教育程度在内的分类变量：

分类变量的转换

直接用数值表示：虽然可以直接用数值表示教育程度，但这种方法忽略了不同教育程度之间的实际差异和对目标变量影响的非线性特性。

创建虚拟变量：为每个教育程度创建一个指示变量（除了一个类别以避免共线性），这样可以在统计分析中更精确地衡量每个教育程度对因变量的影响。

转换为教育年限：将教育程度转换为对应的受教育年数，虽然简化了模型，但可能会丢失某些重要的分类特异信息。

## 特定处理方法的应用

One-hot编码：通过One-hot编码，可以将分类变量转换为一种更适合机器学习模型处理的格式。这种方法为每个类别创建一列，只有是或否的标记，从而避免了错误的距离假设。

WOE编码：特别是在处理具有二元目标变量的分类问题时，WOE编码可以有效地将分类变量转换为连续变量，这有助于捕捉不同类别对目标变量的不同影响强度。

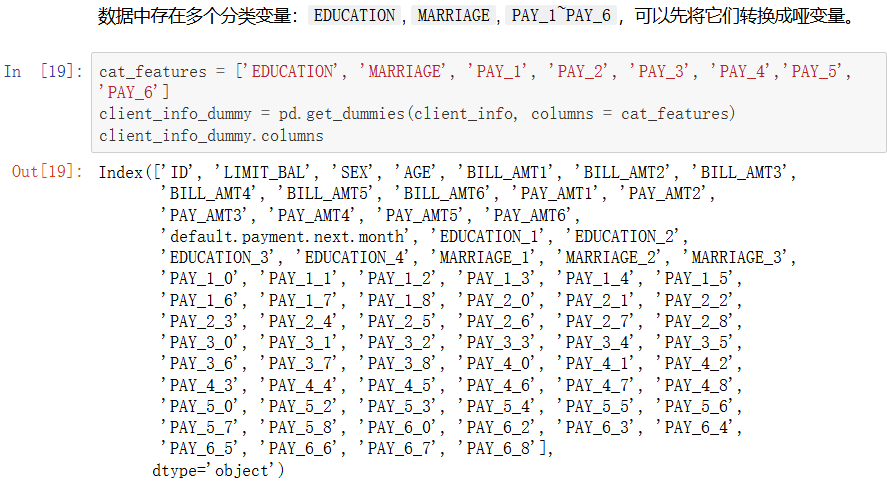
聚类分析：在某些情况下，可能需要根据客户的多个特征（包括教育程度）进行聚类，以识别具有相似特征的客户群体，进而实现更精细化的市场划分。

## 案例分析与应用

决策树模型：在应用决策树模型进行客户分类分析时，教育程度可以通过上述方法之一进行编码，作为重要的特征输入模型，帮助提高分类的准确性和解释性。

信用风险评估：在信用评分模型中，教育程度的处理尤为关键，采用WOE编码等方法可以有效反映不同教育程度对信用风险的预测能力。

综上所述，可以看到处理如教育程度这样的分类变量有多种方法可供选择。每种方法都有其适用场景和优缺点，选择合适的处理方法需要基于具体的业务需求、数据特性以及最终的模型类型。在实践中，尝试多种方法并比较它们的表现往往是寻找最佳解决方案的有效途径。此外，随着数据科学领域的发展，新的方法和技术不断涌现，持续学习和实验是提高数据分析和建模技能的关键。

图3-1数据变量

[3.2 定义预测变量](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "4.2)

在统计学和机器学习领域，预测变量的正确定义与处理对于建立有效的预测模型至关重要。预测变量，也称为自变量，是用于在回归分析中预测因变量（目标变量）的变量。简单线性回归分析是最常见的一种形式，它涉及一个自变量和一个因变量，两者通过一条直线关系进行模拟。以下是关于如何定义和处理预测变量的详细指南：

**确定预测变量**：确定哪些变量可以作为预测变量，这通常基于对业务问题的理解以及数据探索的结果。例如，如果任务是预测房价，那么房屋的大小、位置、年龄等都可能成为潜在的预测变量。

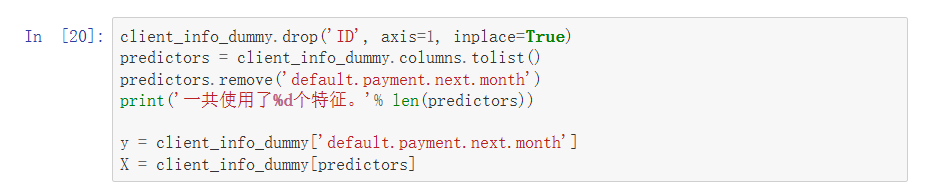
**选择预测变量**：在选择预测变量时，需要确保这些变量与因变量之间存在某种统计上的相关性或因果关系。可以通过绘制散点图来初步观察变量间的关系，使用相关系数来衡量变量间的线性关系强度。

**处理预测变量**：一旦选定了预测变量，接下来需要对这些变量进行适当的数据处理，以确保它们适合用于建模。这包括处理缺失值、异常值，以及必要时对类别变量进行编码处理。

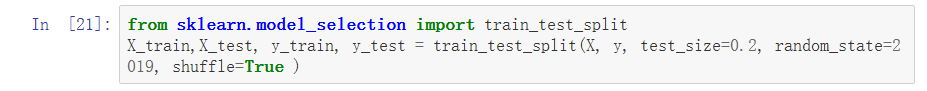
**验证回归模型：**模型建立后，需要对其进行验证和测试，以评估其预测能力。这可以通过交叉验证、计算模型的R方值（表示模型解释的变异性的百分比），或是使用t检验和F检验来评估模型参数的显著性来完成。

**优化预测变量**：在某些情况下，可能需要对模型进行优化，比如通过添加交互项、多项式项或转换变量来改善模型的拟合度和预测准确性。

通过以上步骤，可以有效地定义和使用预测变量来构建预测模型。需要注意的是，虽然这里以简单线性回归为例，但在实际中可能需要考虑更多的因素，如多重共线性、变量之间的非线性关系等。在实际应用中，应结合具体问题和数据情况灵活选择和应用这些方法。

图3-2

[3.3 划分训练集和测试集](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "4.3)

图3-3

将数据集划分为训练集和测试集，分别用于模型训练和评估，比例是4:1，所以将test\_size设为0.2，通过设置shuffle为True将数据随机打乱。并将 random\_state（随机数种子）设置为固定的值，保证每次随机的结果都是一样的。

# 4 模型建立和评估

## 4.1 kNN算法模型建立

kNN是一种基于实例的学习方法，它根据特征空间中相邻点的类别进行分类。其原理是通过测量不同特征之间的距离，找到离待分类样本最近的k个训练样本，通过多数表决来确定待分类样本所属的类别。

使用KNeighborsClassifier类建立kNN模型。首先，导入KNeighborsClassifier类，然后创建一个实例，可以设定k值和其他参数，最后使用fit()方法拟合训练数据。

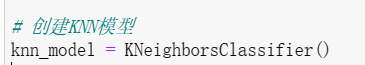


图4-1 KNN模型构建代码

对KNN模型进行交叉验证，搜索不同邻居数量（n\_neighbors）下的最佳性能。在搜索过程中，会尝试不同的邻居数量，从1到20，并绘制出每个邻居数量下的平均准确率曲线图。

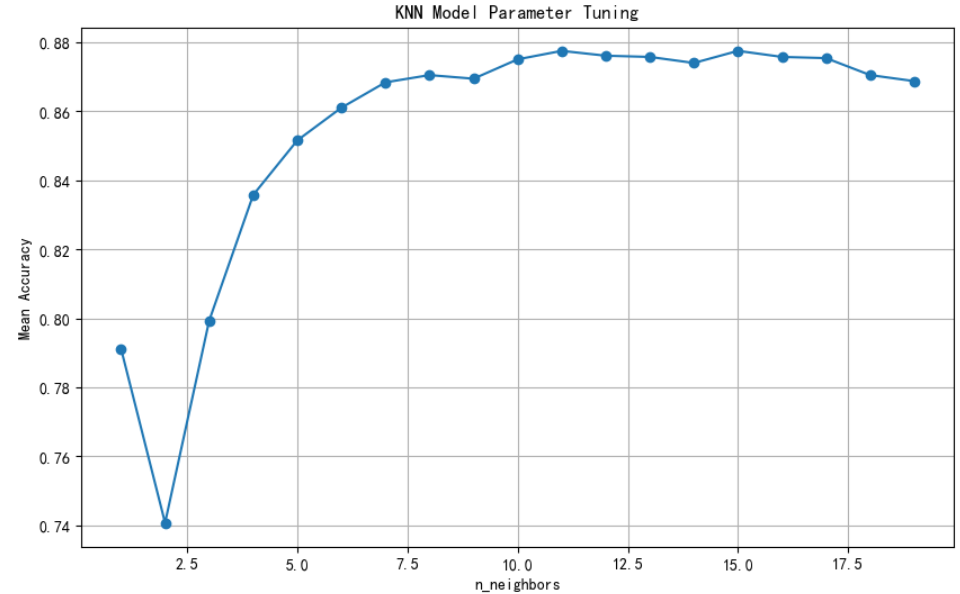


图4-2 KNN模型参数调优曲线图

根据曲线图选择最优的邻居数量15。

## 4.2 决策数据算法模型建立

决策树是一种基于树结构的分类模型，通过对特征进行递归划分来实现分类。它通过选择最优特征对数据进行划分，构建一个树状结构，在每个节点根据特征的值进行分支，直到叶子节点表示数据的最终分类。

使用Scikit-learn库中的DecisionTreeClassifier类可以建立决策树模型。导入类后，创建一个实例，可以设定树的深度、分裂标准等参数，然后使用fit()方法拟合训练数据。

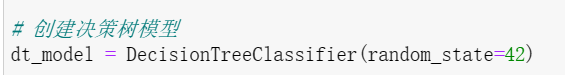


图4-3决策数模型构建代码

决策树算法时，常用的参数包括树的深度、最小分割样本数、最小叶节点样本数等。使用交叉验证来搜索最佳的参数组合，进而提高模型性能。



图4-4决策数模型参数调优结果

## 4.3 xgboost算法模型建立

XGBoost是一种梯度提升树（Gradient Boosting Decision Tree）的算法。 XGBoost是一个强大的梯度提升库，可以使用XGBClassifier类建立分类模型。首先导入该类，创建一个实例，可以设定树的数量、学习率等参数，然后使用fit()方法拟合训练数据。

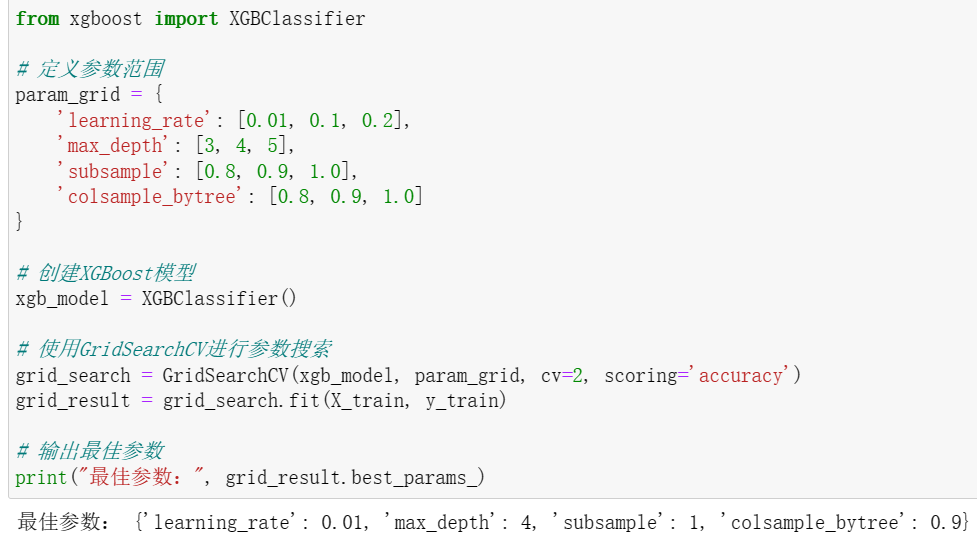


图4-5XGBoost模型参数调优结果

## 4.4 模型评估

计算每个模型的模型的准确率、精度、召回率和F1值。结果如下图所示。

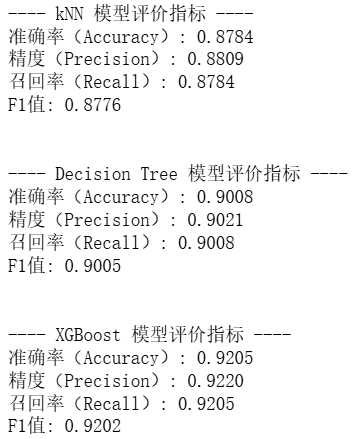


图4-6 三种模型的评估结果

从评估结果来看，XGBoost模型在所有指标上都表现最佳，F1值为0.9202，显著优于kNN和决策树模型。决策树模型的表现次之，准确率为0.9008，精度为0.9021，召回率为0.9008，F1值为0.9005，虽然不如XGBoost，但也表现出较强的分类能力。决策树模型能够通过递归划分特征空间构建树状结构，实现较为精细的分类。而kNN模型的表现相对较弱，虽然其准确率达到了0.8784，但在精度、召回率和F1值上均低于其他两个模型。总体而言，XGBoost模型在本次分类任务中表现最佳，是最终推荐的模型。

# [5. 总结](http://cookdata.cn/media/note_html/1183_56e6e47f9b96de78f141c85b4b3d407d_show.html?time=1718457311.82" \l "5.%E6%80%BB%E7%BB%93)

本研究以机器学习为基础，旨在对信用卡用户违约预测，分别建立了kNN、决策树和XGBoost三种分类模型，并通过交叉验证和参数调优来优化每个模型的性能。最终，通过准确率、精度、召回率和F1值对各模型进行了评估。

研究结果显示，XGBoost模型在所有评估指标上均表现最佳，其准确率达到了0.9205，显著优于kNN和决策树模型。这表明XGBoost模型在处理本研究中的图像分类任务时，具备更强的分类能力和泛化能力。通过本次研究，我们不仅掌握了图像预处理和特征提取的技术方法，还深入理解了不同机器学习算法在图像分类任务中的适用性和优缺点。

# 

# 