# 软件设计说明书

1. 软件名称

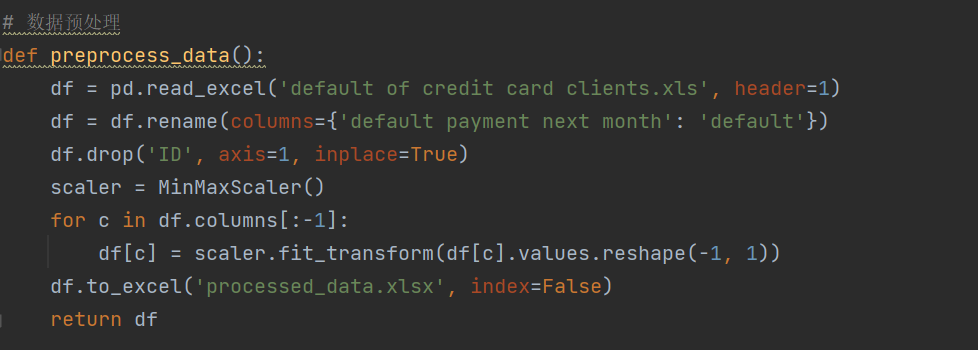
基于支持向量机的信用卡违约预测系统

2. 软件功能介绍

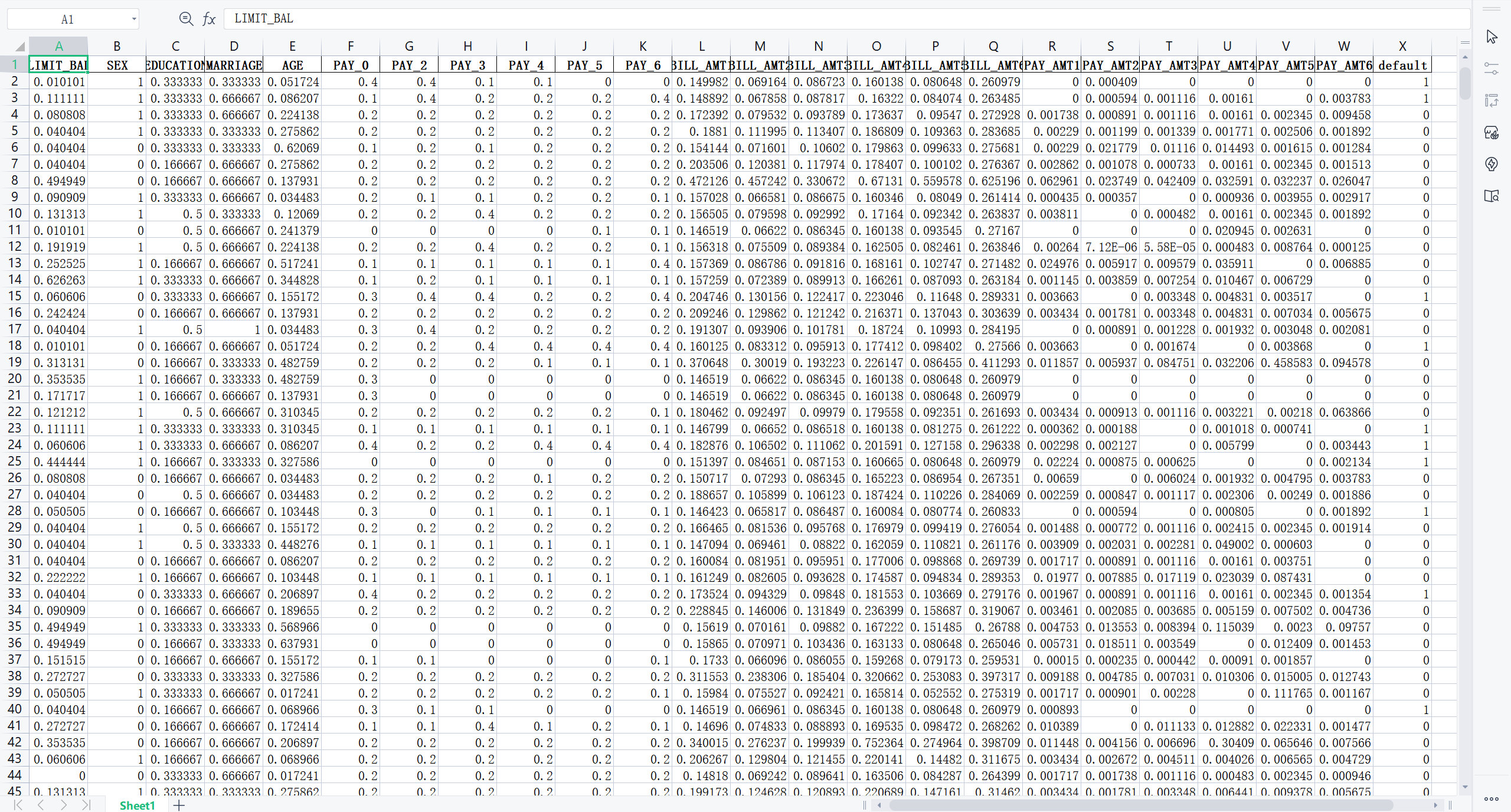
信用卡违约预测系统是一个基于随机森林模型的机器学习软件，主要功能包括数据预处理、模型训练、模型预测和模型评估。这款软件的目标是通过历史数据，预测客户在下个月是否会违约。

软件主要包含以下四个功能：

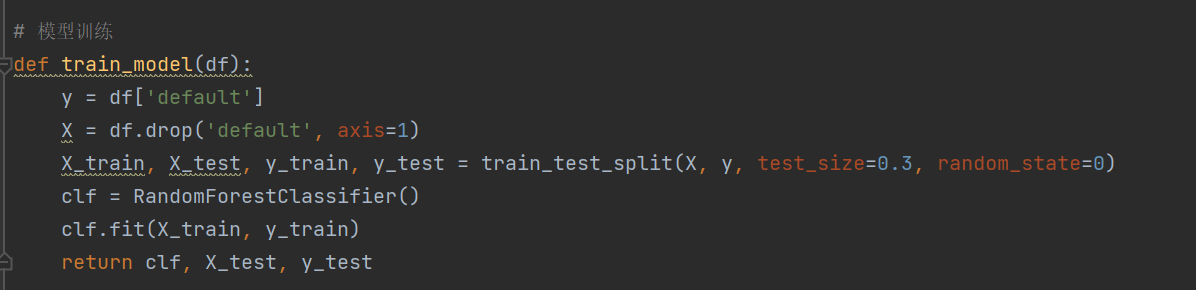
数据预处理： 软件会自动读取从UCI Machine Learning Repository下载Default of Credit Card Clients数据集。并对除了最后一列标签列外的所有数据进行最大最小值归一化。归一化是将所有特征缩放到相同的尺度，这是因为数据集中可能存在异常值或者极端值。处理后可以将这些值对结果的影响降低。

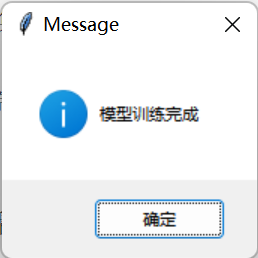




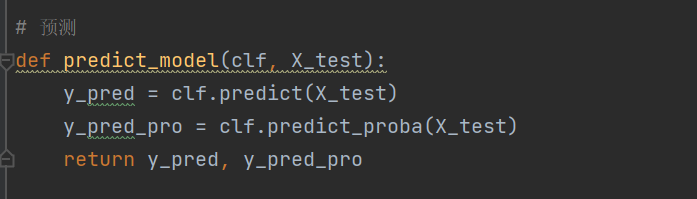


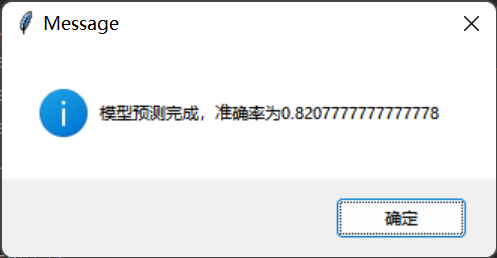
模型训练： 使用随机森林模型，对数据集进行训练拟合。



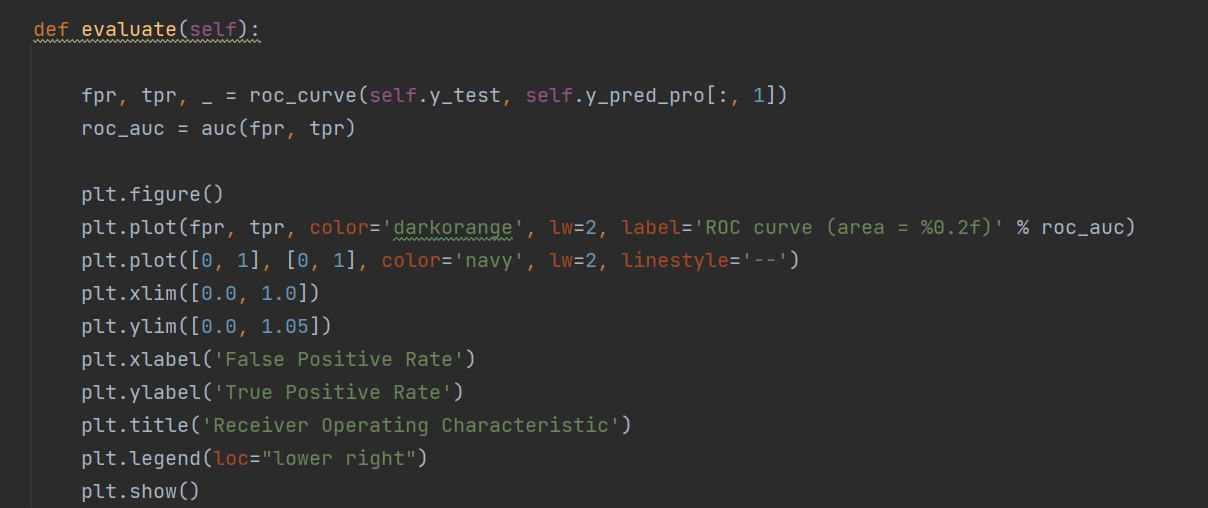


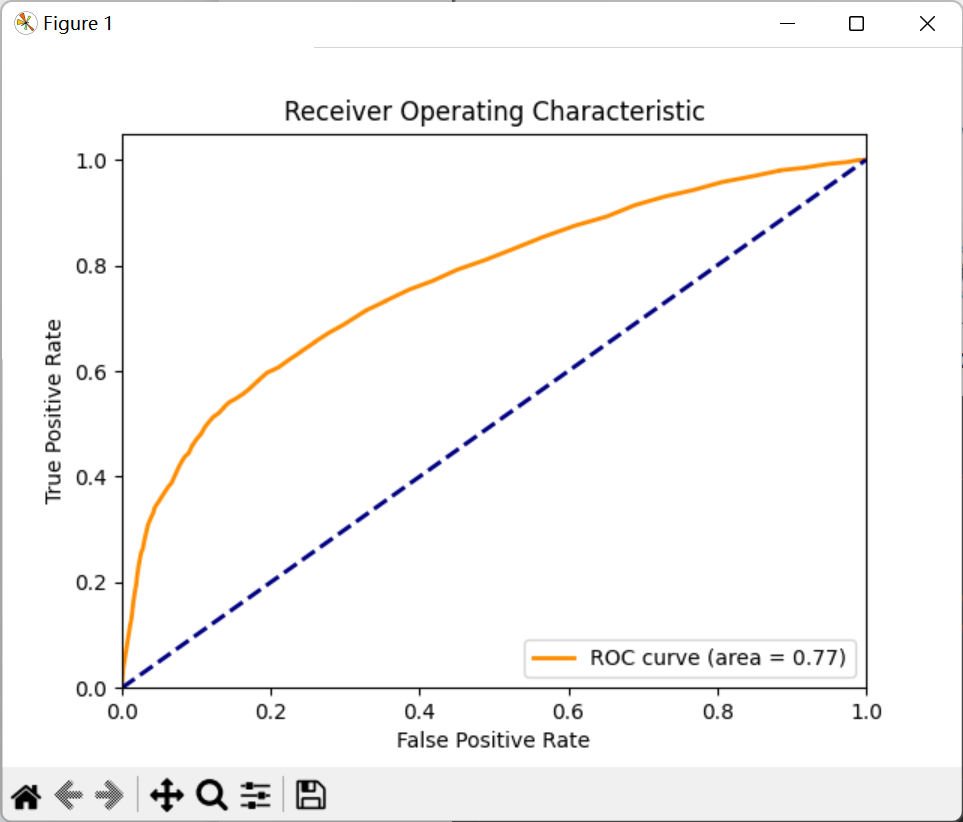
模型预测： 通过训练好的模型对测试集进行预测并展示预测的准确率。





模型评估： 根据AUC评估模型，并绘制ROC曲线。





3.软件使用说明：

使用此软件，用户只需要按照以下步骤进行：

点击"数据预处理"按钮，软件将自动读取并预处理数据集。

点击"模型训练"按钮，软件将自动对预处理后的数据进行训练，并创建一个随机森林模型。

点击"模型预测"按钮，软件将使用训练好的模型对测试集进行预测。

点击"模型评估"按钮，软件将评估模型的性能，计算准确率，并绘制ROC曲线。

1. 算法原理：

随机森林原理：

随机森林是一种集成学习算法，主要用于分类和回归问题。它通过构建多个决策树并组合它们的预测结果来工作。每个决策树都在对数据集的一个随机子集上进行训练，这种子集采样是有放回的，称为bootstrap样本。在进行分裂时，随机森林不是使用所有特征，而是在所有特征中随机选择一部分特征。这两个随机性的引入可以增加模型的多样性，从而降低过拟合的风险。

随机森林的预测结果是基于所有决策树的预测结果的。在分类问题中，通常采用多数投票的方式；在回归问题中，通常采用平均预测值的方式。

ROC曲线原理：

ROC曲线是一种用于评估二元分类模型在所有可能的阈值下的性能的工具。ROC曲线的横轴是假阳性率（False Positive Rate, FPR），纵轴是真阳性率（True Positive Rate, TPR）。

TPR，也称为召回率或灵敏度，表示的是所有真实为阳性的样本中，模型预测为阳性的样本的比例。

FPR表示的是所有真实为阴性的样本中，模型错误地预测为阳性的样本的比例。

ROC曲线下的面积（Area Under the Curve，AUC）可以量化模型的总体性能：AUC越接近1.0，模型的性能越好；AUC越接近0.5，模型的性能越接近随机猜测。