# 5G 用户预测分析报告

## 任务介绍

本次作业的目标是通过使用 Python 编程环境实现相应的人工智能解决方案，对 5G 用户进行预测。任务的核心是基于用户的基本信息和通信相关数据，如用户话费信息、流量、活跃行为、套餐类型、区域信息等特征字段，通过训练数据集训练模型，预测测试集中每个样本是否为 5G 用户。本次任务评估的标准是 AUC（Area Under the Curve），即分数越高，效果越好。

## 任务的思考

5G 用户预测问题本质上是一个二分类问题，即我们需要预测某个用户是否为 5G 用户。具体来说，目标是利用已知的用户特征来构建一个分类模型，从而对新用户进行分类预测。在实际应用中，准确预测潜在的 5G 用户对于通信运营商制定营销策略、优化资源配置具有重要意义。因此，我们需要通过数据预处理、特征工程、模型选择与评估等步骤，构建并优化我们的分类模型。

## 数据分析

数据集包含 60 个字段，其中 target 字段为预测目标，主要特征字段分为两类：离散型特征（cat）和数值型特征（num）。数据文件名为 **train.csv**，包含以下主要字段：

* **id**: 样本标识
* **cat\_0** ～ **cat\_19**: 离散型特征
* **num\_0** ～ **num\_37**: 数值型特征
* **target**: 目标字段，是否为 5G 用户

数据分析的主要步骤如下：

1. **数据加载与探索**：使用 **pandas** 加载数据，初步查看数据的基本信息，包括数据类型、缺失值情况、基本统计量等。
2. **缺失值处理**：对缺失值进行处理，对于数值型特征，采用均值填充的方法；对于离散型特征，采用众数填充的方法。
3. **特征编码**：对离散型特征进行编码转换，将其转换为数值型数据。我们采用的是类别编码方法。
4. **特征标准化**：对数值型特征进行标准化处理，以确保所有特征具有相似的尺度。

通过上述步骤，我们可以确保数据的完整性和一致性，为后续的模型训练打下良好的基础。

## 模型选择的原因

为了实现 5G 用户预测，我们选择了两种常见的机器学习算法：逻辑回归和随机森林。

1. **逻辑回归**：逻辑回归是一种广泛使用的线性分类模型，用于预测分类变量的二元结果。它使用逻辑函数将线性回归的输出映射到0到1之间的概率值，表示一个样本属于某一类别的概率。

**逻辑回归模型有以下优点：**

* + **简单且易于解释**：逻辑回归模型是线性模型，具有较好的可解释性，便于理解和解释模型的预测结果。
  + **适合二分类问题**：逻辑回归是典型的二分类模型，适用于当前的 5G 用户预测任务。
  + **训练速度快**：逻辑回归模型训练速度较快，适合处理大规模数据。

1. **随机森林**：随机森林是一种集成学习方法，通过构建多个决策树并结合它们的输出来提高分类精度。每个决策树在训练过程中会从数据集中随机抽取样本和特征，形成一个“森林”。最终的预测结果通过多数投票（分类问题）或平均（回归问题）来决定。随机森林模型具有良好的泛化能力和抗过拟合特性。

**随机森林模型有以下优点：**

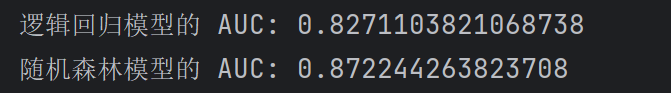
* + **强大的集成学习方法**：随机森林通过集成多棵决策树，具有较高的准确性和鲁棒性，能够处理复杂的非线性关系。
  + **处理高维数据**：随机森林模型能够处理大量特征，具有较好的特征选择能力，能够自动评估特征的重要性。
  + **抗过拟合能力强**：通过引入随机性，随机森林模型具有较强的抗过拟合能力，能够在训练数据上表现良好。

通过选择这两种模型，我们可以比较它们在 5G 用户预测任务中的表现，并选择最佳模型进行最终预测。

## 不同模型的比较

为了评估不同模型的性能，我们将数据集拆分为训练集和测试集，分别训练逻辑回归模型和随机森林模型，并使用测试集进行评估。我们使用 AUC 分数作为评估指标，结果如下：

* **逻辑回归模型的 AUC: 0.8271103821068738**
* **随机森林模型的 AUC: 0.872244263823708**



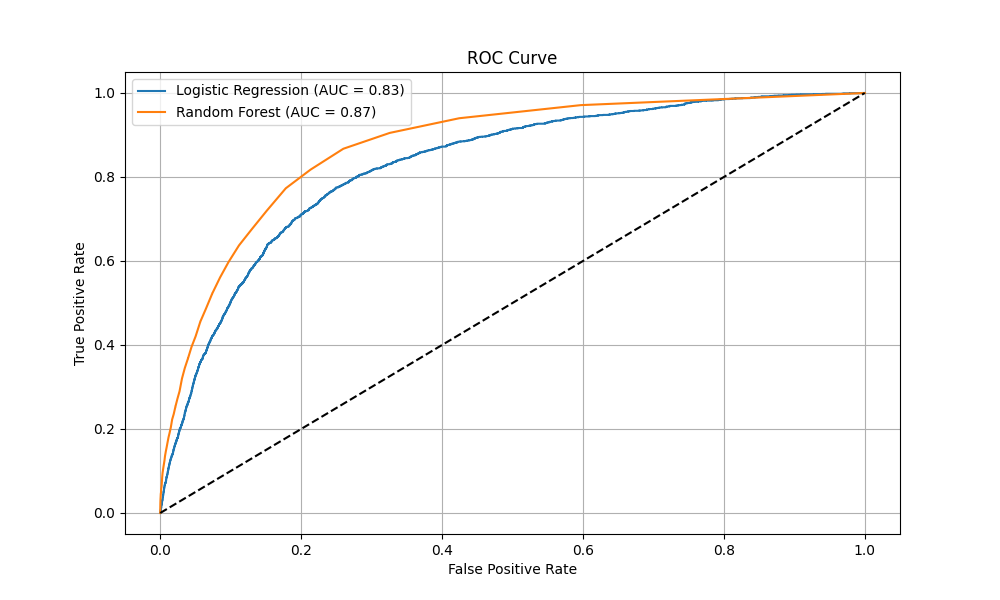
虽然逻辑回归模型的AUC得分较低，但仍然在0.83左右，表明其在正负样本之间的分类性能良好。而随机森林模型的AUC得分略高于逻辑回归模型，达到了0.87，进一步强调了其在预测5G用户方面的优越性。

**差异分析**：

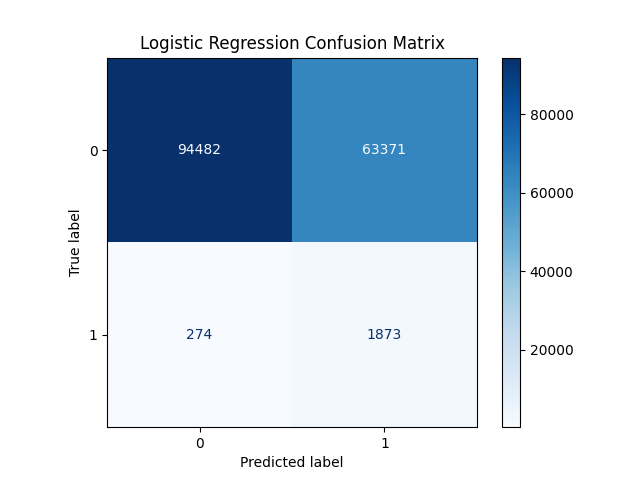
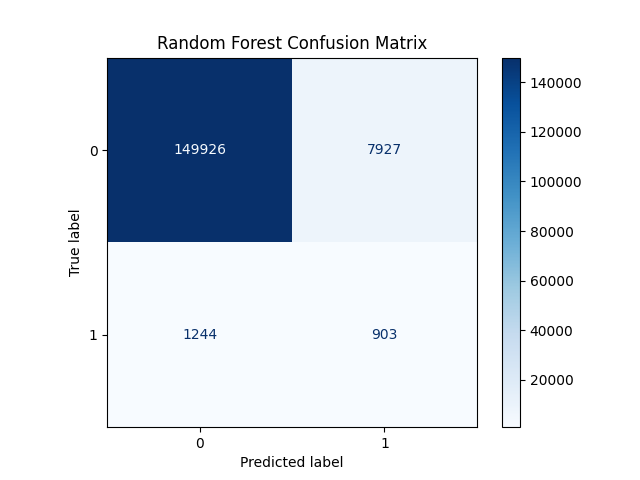
* 结果显示随机森林模型的 AUC（0.87）略高于逻辑回归模型的 AUC（0.83）。
* 这可能是因为随机森林模型具有更高的灵活性和非线性建模能力，能够更好地捕捉数据中的复杂关系。
* 另外，随机森林在处理不平衡数据时表现更加出色，因为它可以通过调整类别权重或使用过采样技术（如SMOTE）来处理类别不平衡问题，而逻辑回归在这方面的表现相对较弱。

## 结果的分析

在结果分析过程中，绘制了 ROC 曲线和混淆矩阵，以进一步了解模型的性能。ROC 曲线可以帮助我们直观地了解模型在不同阈值下的性能，而混淆矩阵可以帮助我们了解模型的误分类情况。



1. **ROC 曲线**：
   * ROC 曲线显示了在不同的分类阈值下，模型的真阳性率（True Positive Rate）与假阳性率（False Positive Rate）之间的关系。曲线越靠近左上角，表示模型性能越好。
   * 逻辑回归模型的 ROC 曲线下面积（AUC）为 0.83，而随机森林模型的 AUC 为 0.87。这表明随机森林模型在整体上略优于逻辑回归模型。



1. **混淆矩阵**：
   * 混淆矩阵显示了模型在测试集上的预测结果与真实标签之间的关系。主要包括真正例（True Positives）、真负例（True Negatives）、假正例（False Positives）和假负例（False Negatives）。
   * 对于逻辑回归模型和随机森林模型，混淆矩阵可以帮助评估模型在不同类别下的分类准确度。

## 改进思路

尽管两个模型的AUC得分已经不错，但我们仍有一些可以改进的方向：

1. **特征工程**：可以尝试生成更多的特征，或者通过特征选择方法筛选出更有意义的特征。例如，我们可以考虑使用主成分分析（PCA）或特征重要性分析来选择最有价值的特征。
2. **超参数优化**：使用网格搜索或随机搜索来优化模型的超参数，以进一步提高模型性能。例如，对于随机森林模型，我们可以调优树的数量、最大深度、最小样本分裂数等参数。
3. **数据增强**：如果数据集较小，可以通过数据增强技术增加数据量，以提高模型的泛化能力。例如，我们可以通过生成合成数据或采样技术来增加训练数据的多样性。
4. **模型融合**：尝试使用不同模型的组合（如集成方法）来提高预测效果。例如，我们可以使用投票分类器、堆叠模型等技术来融合多个模型的预测结果。
5. **使用更复杂的模型**：尝试使用更复杂的模型，如 XGBoost 或深度学习模型，来进一步提升预测准确性。例如，XGBoost 是一种基于梯度提升的树模型，具有较高的性能和灵活性，而深度学习模型则可以通过深度神经网络捕捉复杂的非线性关系。

通过以上改进，我们可以期望进一步提高模型的预测性能，提升对 5G 用户的预测准确性。