**华中科技大学**

**计算机科学与技术学院**

**《机器学习》结课报告**



专 业： 计算机科学与技术

班 级： 计算机7班

学 号： U202215561

姓 名： 瞿明睿

成 绩：

指导教师： 何琨

**完成日期：2024 年 5 月 16 日**

目录

[1. 实验要求 2](#_Toc166934056)

[2. 算法设计与实现 3](#_Toc166934057)

[3. 实验环境与平台 3](#_Toc166934058)

[4. 结果与分析 3](#_Toc166934059)

[5. 个人体会 4](#_Toc166934060)

# 实验要求

* 1. 实验题目

Binary Classification with a Bank Churn Dataset，即使用银行的数据集进行二元分类，预测客户是继续使用他们的账户还是流失。

评估方式提交在预测概率和观测目标之间的 [ROC 曲线下面积](http://en.wikipedia.org/wiki/Receiver_operating_characteristic)上进行评估，提交文件对于测试集中的每个变量，必须预测目标变量的概率。该文件应包含标头，并具有以下格式：id Exited。

* 1. 实验分析

本次不允许使用已经实现好的模型如sklearn中的决策树，逻辑回归等，亦或是torch中的隐藏层、激活函数搭建等模块，仅为实现我们需要的功能，通过我们手动实现而不是调用已经实现好的功能起到锻炼我们的能力，所以我们需要分析好我们的题目一步一步做好代码。

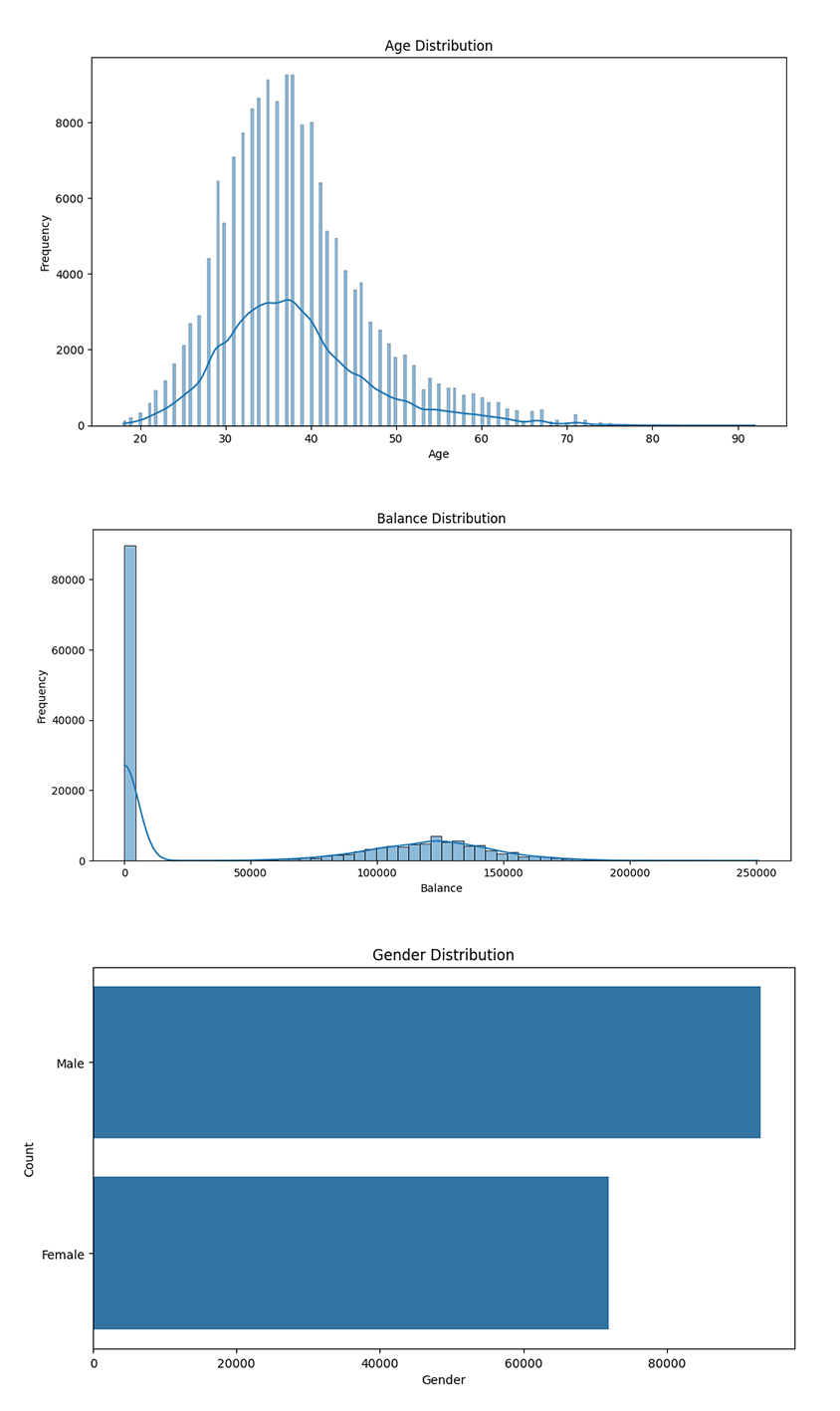
以下为初步分析的数据分布：

图1-1 训练集数据分布图

如图所示，我们可以关注到，是否退出和性别比例有明显的关系，并不是50%对半开。此外，我们发现balance是一种接近于正态分布的趋向，最后我们可以发现，年龄也有相当大的关系。所以在数据进行处理时，这些因素需要注意到。（补充：这家银行是国际银行，通过数据集可以分析出这样的结论，所以在实际操作时，我们也需要在处理数据时，将不同的国家进行处理赋值）

# 算法设计与实现

* 1. 算法选择以及比较

在选择算法的过程中，由于需要手动实现，我选择了几个方向：svm，逻辑回归，以及多层感知机。

在初次尝试的过程中，svm性能表现很差，在训练集上花费的时间非常多，通常十分钟都无法进行一次训练，这使得使用cuda成为一种必须，但是这花费太大，故放弃。

第二次尝试使用逻辑回归，在实现的初步既可以得到一定的效果，但是在寻找最佳参数的时候，在不使用任何优化方法的情况下，正确率一直在0.65徘徊无法提升，给出贝叶斯寻找1000次找到的可能的最佳的参数：Learning Rate = 0.1478, Iterations = 1409, Validation AUC = 0.6512。所以我选择了逻辑回归并且决定在寻找一种方式使得能够达到更高的准确度。

第三次尝试我选择了多层感知机，因为室友的思路分享启发了我，我也可以通过几个技术方向解决这个问题，我也想知道不同的处理对结果影响究竟有多大，在初步尝试中，我得到了相当高的准确度，某种意义上薄纱了逻辑回归，这也是符合这样的问题，因为逻辑回归其实并不太适合这样的具有一定的概率分布的数据，所以其上限是不如svm等路线的，但是我还是对其进行了优化。

* 1. 具体的算法实现

在不同的算法实现中，我们使用了近似的数据处理方法。在这里先给出训练集的数据样id, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited；0,15674932,Okwudilichukwu,668,France,Male,33.0,3,0.0,2,1.0,0.0,181449.97,0

可以看出，有些数据是本身就是数据化的，有些数据是需要处理的，我们的处理如下，提取客户的id以在讲预测的结果和id相关联，如何我们需要将性别也转换为数值型，将男性表为1，女性表为0，再其次将地理位置也类似于性别处理，比如将法国设置为0，德国设置为1，西班牙设置为2。最后再将无用的客户id以及字符串姓名去除得到我们需要的数据。

再其次是将我们的将特征和标签分离，X是特征，y是标签（即流失情况）

# 实验环境与平台

系统环境windows10专业版 操作系统内部版本 19045.4412

处理器 11th Gen Intel(R) Core(TM) i9-11900KF

系统类型 64 位操作系统, 基于 x64 的处理器

Gpu: 2080ti 11g(cuda12.4)

Python 3.12.3 64bits

所用库以及版本：numpy 1.26.4

bayesian-optimization 1.4.3

pandas 2.2.2

scikit-learn 1.4.2

# 结果与分析

在本次实验中，我们对数据进行了详细的预处理。首先，我们将数据中的类别变量 Geography 和 Gender 转换为数值表示，并且应用了独热编码（One-Hot Encoding）以避免数值映射带来的潜在影响。独热编码将类别变量扩展为多维特征向量，从而使模型能够更好地处理类别特征。

原始特征：

- Geography: France, Germany, Spain

- Gender: Male, Female

独热编码后的特征：

- Geography\_France, Geography\_Germany, Geography\_Spain

- Gender\_Male, Gender\_Female

为了提高模型的训练效果，我们对所有数值特征进行了标准化处理（Standardization）。标准化处理将特征的均值调整为0，标准差调整为1，从而使特征具有相同的尺度。这有助于加速模型的收敛并提高模型的性能。

在验证集上，我们通过预测概率来评估模型的AUC值：

在对测试集进行预测后，我们生成了最终的提交文件。模型对测试集进行了预测，并输出了每个样本的预测概率

通过以上步骤，我们实现了一个从数据预处理、特征工程到模型训练和评估的完整机器学习工作流程。使用逻辑回归模型，我们在验证集上达到了的AUC值，证明了模型的高效性和稳定性。独热编码和标准化处理确保了特征处理的一致性，为模型提供了良好的基础数据。

在未来的工作中，我们可以尝试更多的特征工程和模型调优方法，进一步提高模型的预测准确性和稳定性。同时，可以考虑引入更多的数据增强技术和更复杂的模型（如神经网络）来处理更复杂的特征关系。

# 5. 个人体会

**个人第一次接触python处理这样的问题，从学习pip，py的语法以及各种各样的模型，打包好的数据处理方法都让我觉得十分新奇，也觉得人工智能是一个非常有意思的方向**