证券公司客户流失预测

一、任务背景

一家证券公司的业务经理面临着客户流失的问题。他们希望分析已有的数据，探寻背后的原因，并利用这些数据构建一个分类模型，以预测未来可能会流失的客户。

二、数据简介

每个样本有一个唯一标识 ID，特征维度为 22，特征类型包括离散、连续与目录型。部分特征简介如下：

- Income_Categorysort：账户持有人的年收入类别
- Credit_Limit：客户信用额度
- Education_Level：客户学历等级

样本的标签（第 Attrition_Flag 列）：1 表示帐户已关闭，也即客户已流失；0 表示帐户仍开通，也即客户未流失。

总共给定三个数据集：

- 训练集（train.csv），包含了 7626 条样本，样本的标签给定，但存在 10%（760 个）的错误，也即原来 10%的样本被随机选择后，其标签取反（1→0 / 0→1）。且其中一半的错误标签对应的样本 ID 给出（存放在 noisy ID1.xlsx），另外一半的没有给出；
- 测试集 1（test1.xlsx），包含了 500 个样本，样本的标签给定；
- 测试集 2（test2.xlsx），包含了 1000 个样本，样本的标签未给定。

三、具体任务

3.1 任务一：数据分析与预处理

- ✧ 缺失值处理：数据集在 Gender、Education_Level 等特征上有不同数量的缺失值，请使用适当的方法进行处理。需要介绍所使用的方法并论述合理性；（10 分）
- ✧ 特征选择：可能存在着无关或者冗余的特征，请使用适当的方法进行特征选择。需给出挑选出的无关/冗余特征维度的名字，介绍所使用的方

法并论述合理性；（10 分）

- ✧ 其它处理：可使用其它的有利于后续任务的数据分析方法；需要给出结果、介绍所使用的方法并论述合理性。（10 分）

3.2 任务二：错误标签检测

- ✧ 训练集中有 10%样本的标签是错误的，其中有一半（也即 380 个）的错误标签所在的样本 ID 是给出的（存放在 noisy ID1.xlsx）。需要构建有效的错误标签检测算法，找出另外一半（也即 381 个）的错误标签对应的样本 ID。请返回一个样本 ID 列表（Excel 格式），包含所检测的 760 个错误标签的样本 ID；（15 分）
- ✧ 请计算已知的 380 个错误标签样本 ID 的检测准确率；（5 分）
- ✧ 请分析所使用的错误标签检测算法的计算复杂度与合理性。（10 分）

3.3 任务三：分类模型构建

- ✧ 请使用训练集以及错误标签的检测结果来训练三个或以上的分类模型，并报告与对比这些分类模型在测试集 1 上的准确率 Accuracy；（20 分）
- ✧ 对分类算法的选取、训练过程以及超参数选择的策略进行介绍，需论述分类算法选取的合理性；（10 分）
- ✧ 返回一个 Excel 文件，记录在测试集 2 上的预测结果；（10 分）
- ✧ 鼓励自己实现一个分类算法（KNN\Logistic Regression\SVM\Random forest\Adaboost\Decision tree）。自己编写的算法程序并经一对一在线验证后，若 F1>70%，可根据算法的复杂性和效果额外加 10 到 15 分（与平时成绩之和不高于 100 分）。

四、 需提交的文件

- ● 三个任务的程序代码（matlab、python 均可，需可执行）及运行说明；
- ● 一个大作业报告（pdf 格式），包括每一个任务的分析、介绍、解决方法与结果等；
- ● 两个结果列表（Excel，后缀为.csv）：第一个记录算法所发现的 760 个错误标签的样本 ID，请按列放置；第二个记录测试集 2 上的标签预测结果，请按列放置；
- ● 上述文件请放到一个文件夹并转成压缩包（zip/rar），命名为“学号+姓

名+大作业”。

五、注意事项

- 如发现不可忽视比例的雷同文字、代码，将询问涉及的同学，一旦判定存在抄袭，将按照学校规定处理；
- 请注意文档与代码的书写规范，这也是一个评分点；
- 结果的准确性只是重要参考，将主要根据文档与代码来判定成绩；
- 自行实现的分类算法，需要一对一在线讲解代码细节并现场给出运行结果；
- 不遵循上述文件提交格式要求的，对应的得分点为 0 分。