题目 6. 银行客户流失预测

问题背景

假设你是金融学研究者，在研究银行客户关系的问题上，了解是否某种特定的年龄、性别、地域、姓氏，会让客户更容易注销银行账号，离开银行（客户流失）。当然，离开银行也有其他原因，例如工资入不敷出、信用较差等。为定量分析这个问题，你获得了一份银行的客户表，里面记载了客户的年龄、性别、所处国家、姓氏，以及是否离开了银行。如何在分类模型的基础上，研究某种年龄段、性别、所处国家、姓氏是不是离开银行的显著原因？请对此展开研究。

数据集和资料

数据集 `q_6_churn.csv`

在线性回归中, 研究某些变量是否对回归有影响时, 应先用全部变量作回归 (Unrestricted, UR), 然后去掉想要研究的变量,再作回归 (Restricted, R). 计算 UR 和 R 模型的残差平方和 (Sum-squared Error, SSE), 作 F 检验得到的 p 值即为想要研究的变量的显著性。

线性回归中加入冗余变量，会引起多重共线性；虽然每一项的系数变得没有意义，但 SSE 本身不会因为加入新变量而增加。但在机器学习模型中，即使新加入的变量不包含任何信息，甚至包含一些有用的信息，数据维度增加也会使得 SSE 增加。在研究变量的贡献时，不应该直接删除维度，而是在样本层面将这一个维度的信息随机打乱，破坏样本和这一列数值的对应关系，即 bootstrap 方法。

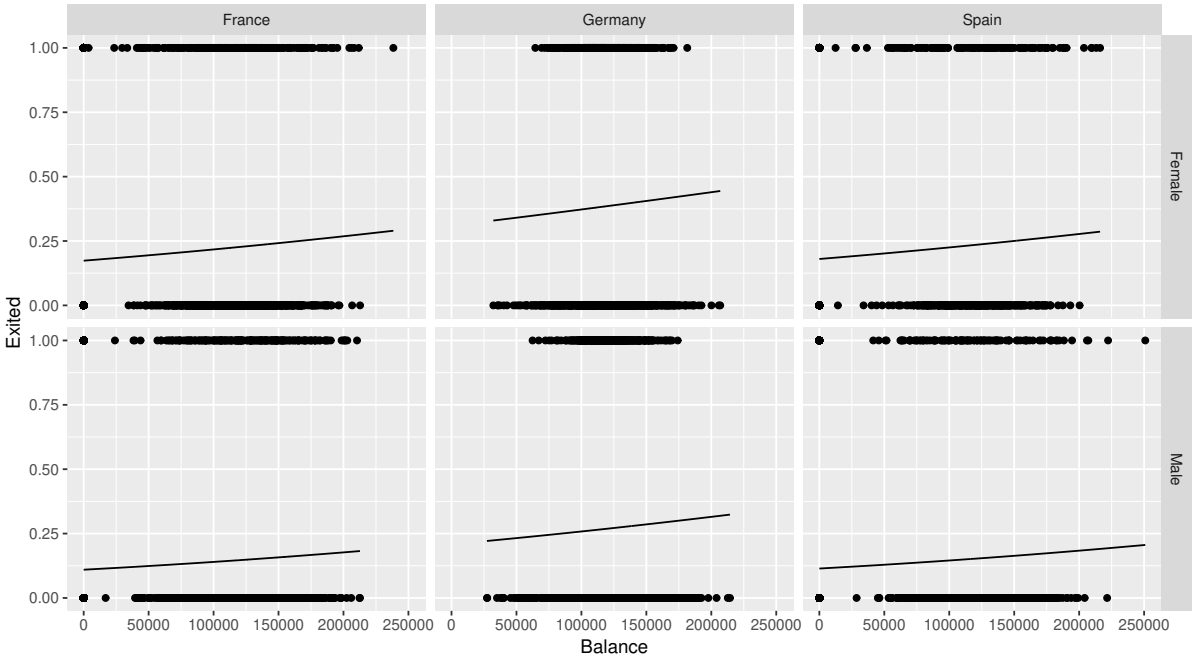

图 1: Logistic 回归在性别和地区之间的差异

年龄、性别、所处国家、姓氏分属 4 个变量，它们有可能单独对结果有影响，也有可能几种变量的组合才会有显著的不同。为了解决这个问题，一方面可以用 one-hot 编码将离散变量拆分成多个二元变量，另一方面也可以做一些可视化。图 1 展示了 `Exited ~ Balance + Gender + Geography` 的

Logistic 回归在性别和地区之间的差异。这里只选择了 Balance 作为连续自变量，在完成研究任务时，请结合其他的连续自变量一起考虑。

研究任务

问题 6-1. 使用 Logistic 回归预测客户流失，将样本拆分为 80% 的训练集和 20% 的测试集，在训练集上启用 5 折交叉验证。检验年龄、性别、所处国家、姓氏系数分别的显著性（T 检验）和总体的显著性（F 检验）。

问题 6-2. 使用随机森林模型重复问题 6-1，在交叉验证的验证集上搜索最优的超参数，计算年龄、性别、所处国家、姓氏的特征重要性。比较 Logistic 回归 T 检验显著性与随机森林的特征重要性，阐述二者的关系。

问题 6-3. 在随机森林模型上进行 F 检验，评价年龄、性别、所处国家、姓氏对随机森林回归结果的影响。随机森林模型没有系数的概念，如何评价这样的影响是正向的还是负向的？例如年龄较大的客户相比年龄较小的，更容易还是更不容易流失？