2022年首届钉钉杯大学生大数据挑战赛初赛题目

初赛A：银行卡电信诈骗危险预测

一、问题背景：

数字支付正在发展，但网络犯罪也在发展。电信诈骗案件持续高发，消费者受损比例持续走高。报告显示，64%的被调查者曾使用手机号码同时注册多个账户，包括金融类账户、社交类账户和消费类账户等，其中遭遇过电信诈骗并发生损失的比例过半。用手机同时注册金融类账户及其他账户，如发生信息泄露，犯罪分子更易接管金融支付账户盗取资金。

随着移动支付产品创新加快，各类移动支付在消费群体中呈现分化趋势，第三方支付的手机应用丰富的场景受到年轻人群偏爱，支付方式变多也导致个人信息也极易被不法分子盗取。根据数据泄露指数，每天有超过 500 万条记录被盗，这一令人担忧的统计数据表明 - 对于有卡支付和无卡支付类型的支付，欺诈仍然非常普遍。

在今天的数字世界，每天有数万亿的银行卡交易发生，检测欺诈行为的发生是一个严峻挑战。

二、数据描述：

该数据来自一些匿名的数据采集机构，数据共有七个特征和一列类标签。下面对数据特征进行一些简单的解释（每列的含义对我们来说并不重要，但对于机器学习来说，它可以很容易地发现含义。它有点抽象，但并不需要真正了解每个功能的真正含义。只需了解如何使用它以便您的模型可以学习。许多数据集，尤其是金融领域的数据集，通常会隐藏一条数据所代表的内容，因为它是敏感信息。数据所有者不想让他人知道，并且数据开发人员从法律上讲也无权知道）

- distance_from_home：银行卡交易地点与家的距离；
- distance_from_last_transaction：与上次交易发生的距离；
- ratio_to_median_purchase_price：近一次交易与以往交易价格中位数的比率；
- repeat_retailer：交易是否发生在同一个商户；
- used_chip：是通过芯片（银行卡）进行的交易；

➢ used_pin_number：交易时是否使用了 PIN 码；

➢ online_order：是否是在线交易订单；

➢ fraud：诈骗行为（分类标签）；

三、解决问题：

1）使用多种用于数据挖掘的机器学习模型对给定数据集进行建模；

2）对样本数据进一步挖掘分析，通过交叉验证、网格调优对不同模型的参数进行调整，寻找最优解，将多个最优模型进行进一步比较；

3）通过对 precision（预测精度）、recall（召回率）、f1-score（F1 分数值）进行计算，给出选择某一种预测模型的理由；

4）将模型性能评价通过多种作图方式进行可视化