**模块C: 机器学习（40分）**

**竞赛题目：信用卡欺诈检测**

**背景介绍**

银行业是最早应用数据科学技术的领域之一，收集了大量结构化数据。如今，数据已经成为银行业最宝贵的资产，不仅可以帮助银行吸引更多的客户，提高现有客户的忠诚度，做出更有效的数据驱动的决策，还可以增强业务能力，提高运营效率，改善现有的服务，加强安全性，并通过所有这些行动获得更多的收入等等。

数据分析使银行业能够成功地执行众多任务，包括：

* 投资风险分析
* 客户终身价值预测
* 客户细分
* 客户流失率预测
* 个性化营销
* 客户情绪分析
* 虚拟助理和聊天机器人
* ……

欺诈检测是银行业中最常见的数据分析用例之一，除了银行业，欺诈活动还存在于许多领域。在政府、保险、公共部门、销售和医疗保健等领域，这都是一个具有挑战性的问题。也就是说，任何处理大量在线交易的企业都会面临欺诈风险。

金融犯罪的形式多种多样，包括欺诈性信用卡交易、伪造银行支票、逃税、网络攻击、客户账户盗窃、合成身份、诈骗等。

欺诈检测是为识别和防止欺诈活动和财务损失而采取的一种主动措施。欺诈检测系统的主要挑战是迅速适应不断变化的欺诈模式和欺诈者的策略，并及时发现新的和日益复杂的方案。

**数据描述**

数据集包含以下变量：

* 数值编码的变量 V1到 V28 是从PCA变换中获得的主分量。由于保密问题，未提供有关原始功能的背景信息。
* Amount 变量表示交易金额。
* Class 变量显示交易是否为欺诈（1）或非欺诈（0）。

注意：欺诈案件总是占少数，并且很好地隐藏在真实的交易中。

**评分标准**

**目标**

本次数据为信用卡交易数据集，目标为训练一个分类模型预测这些交易活动中的欺诈行为，并在测试数据集上尝试获得最优结果。

**评价指标**

F1分数: sklearn.metrics.f1\_score

F1分数：是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率。F1分数可以看作是模型精确率和召回率的一种调和平均，它的最大值是1，最小值是0。

**环境介绍**

在本竞赛模块环境中，已自动启动了 jupyter notebook 环境。环境中已安装 Python 中多种数据分析和机器学习库，包括：

* Numpy
* Scipy
* Pandas
* Matplotlib
* Seaborn
* Scikit-learn
* Xgboost
* Lightgbm
* 其他

**任务列表**

1. 数据获取
2. 数据探查
3. 数据预处理与特征工程
4. 模型训练
5. 模型预测
6. 结果保存

**任务启动**

**Step1. 数据获取(data/credit\_card\_fraud.csv,需自行划分)**

* 训练集路径：../input/train.csv.zip
* 测试集路径：../input/test.csv.zip

**Step2. 数据探查**

**Step3. 数据预处理与特征工程**

**Step4. 模型训练**

**Step5. 模型预测**

**Step6. 结果保存**

**任务结束**

恭喜完成本次比赛，祝您取得好成绩。