# 基于机器学习的信用卡交易欺诈检测

**详细过程可以参考博客**[**【intel校企合作课程】基于机器学习的信用卡交易欺诈检测**](https://blog.csdn.net/Azreal233/article/details/134692737)

**AI 作业一：信用卡交易欺诈检测**

**问题描述：**

2021 年，与信用卡欺诈相关的损失超过 120 亿美元，同比增长近 11%。就重大财务损失、信任和信誉而言，这是银行、客户和商户面临的一个令人担忧的问题。

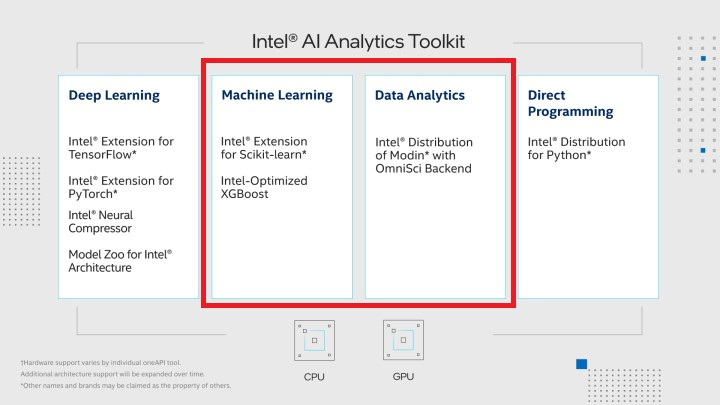
电子商务相关欺诈一直在以约 13% 的复合年增长率 (CAGR) 增加。由于欺诈性信用卡交易急剧增加，在交易时检测欺诈行为对于帮助消费者和银行非常重要。机器学习可以通过训练信用卡交易模型，然后使用这些模型更快、更准确地检测欺诈交易，在预测欺诈方面发挥至关重要的作用。

**预期解决方案：**

我们期待您参考英特尔的类似实现方案，基于提供的信用卡交易数据，训练一个或多个机器学习模型，有效预测信用卡交易是否为欺诈交易——这里推理时间和二分类准确度（F1分数）将作为评分的主要依据。

**要求：**

需要使用 英特尔®oneAPI AI分析工具包中数据分析和机器学习相关的优化库，基于参考资料的方案，进行信用卡交易欺诈检测的实现，建议跟官方开源的机器学习库运行性能进行分析对比。



**数据集 (内容相同，压缩格式及分发渠道不同)**

Zip格式 <https://filerepo.idzcn.com/dataset/assignment_1.zip>

Tar格式 <https://filerepo.idzcn.com/dataset/assignment_1.tar>

百度网盘：<https://pan.baidu.com/s/1KNdSIwQHiDrJLT-5K-sPmA>提取码：fly8

**数据集补充信息**

该数据集包含欧洲持卡人 2013 年 9 月通过信用卡进行的交易。

该数据集显示了两天内发生的交易，其中 284,807 笔交易中有 492 笔欺诈。数据集高度不平衡，正类（欺诈）占所有交易的 0.172%。

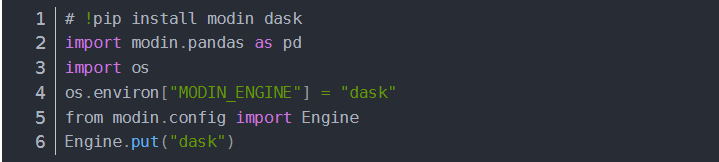
它仅包含PCA (Principal Component Analysis) 变换结果的数字输入变量。不幸的是，由于保密问题，我们无法提供有关数据的原始特征和更多背景信息。特征 V1、V2、…V28 是通过 PCA 获得的主要成分，唯一未通过 PCA 转换的特征是“时间”和“金额”。特征“时间”包含数据集中每个事务与第一个事务之间经过的秒数。特征“金额”是交易金额，该特征可用于示例相关的成本敏感学习。特征“类别”是响应变量，如果存在欺诈，则取值 1，否则取值 0。

考虑到类别不平衡率，我们建议使用精确率-召回率曲线下面积 (Area Under Precision-Recall (PR) Curve即AUPRC) 来测量准确度。混淆矩阵精度对于不平衡分类没有意义。

**项目介绍**

通过提供的信用卡交易诈骗数据集，对数据首先进行数据探索、数据预处理、利用机器学习建立模型，并进行欺诈数据的检测。  
1、**数据探索**：查看数据集规模、数据类型、缺失值情况以及统计性描述。  
2、**数据预处理**：处理缺失值、平衡数据样本  
3、**利用机器学习建立模型**：支持向量机分类、集成学习

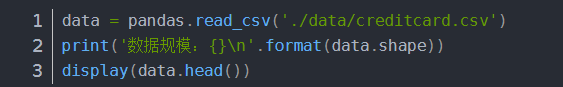
**数据探索**

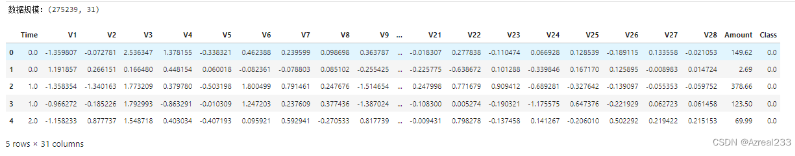


**Modin**是一个Python第三方库，可以通过并行来处理大数据集。其中语法与pandas相近，拥有出色的性能弥补了pandas处理大型数据集的缺陷。

而**Dask** 是一个用于分析计算的灵活的并行计算库，实现大型多维数据集分析的更快执行以及加速和扩展数据科学制作流程或工作流程的强大工具。

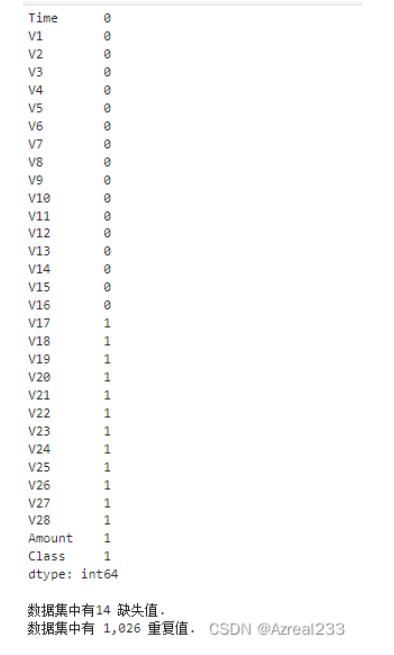
**查看数据集**



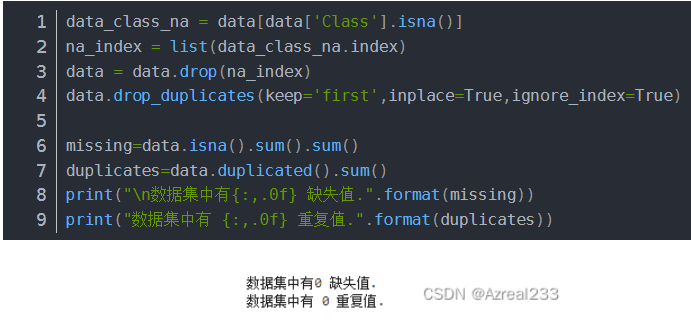


可以看到数据集已经进行**PCA (Principal Component Analysis)** 变换，并且由于数据集进行了脱敏处理，没有必要再进行降维选择特征，我们可以选择**归一化或者标准化**消除不同属性值区间带来的误差。

**查看缺失值和重复值情况**

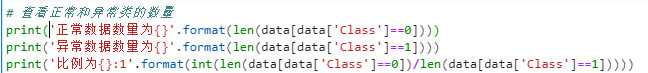


**处理缺失值**



可以根据数据集情况选择不同的缺失值处理方式，本文经过查看数据集发现缺失数据存在于一行，直接删除缺失数据。

**查看数据类别分布情况**





数据集正类和负类极度不平衡，比例达到了567:1，对数据建模前需要进行数据平衡。

**平衡数据**

**主要方法**

**欠采样（Undersampling）**：从多数类别中随机选择一部分样本，使得多数类别的样本数量与少数类别的样本数量相近。这种方法的优点是简单快捷，但可能会丢失一些有用信息。

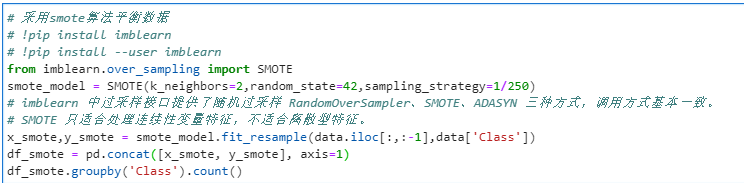
**过采样（Oversampling）**：从少数类别中随机复制一些样本，使得少数类别的样本数量与多数类别的样本数量相近。这种方法的优点是可以充分利用数据集，但可能会导致过拟合。

**SMOTE（Synthetic Minority Over-sampling Technique）算法**：是一种常用的过采样方法，它通过对少数类别样本进行插值生成新的样本来扩充数据集。这种方法可以有效地避免过拟合问题。

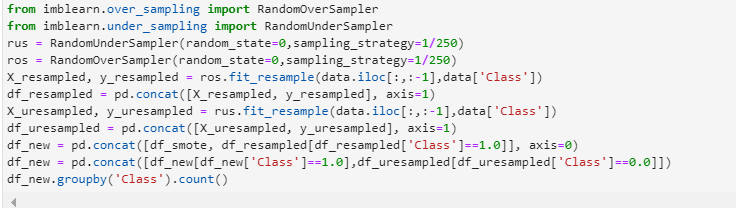
**混合采样（Mixed Sampling）**：结合欠采样和过采样的优点，既可以减少数据量，又可以充分利用数据集。可以先进行欠采样，然后再对欠采样后的数据进行过采样。

本文先对异常样本采用了SMOTE算法进行插值生成新的样本来扩充数据集，然后对插值后的异常样本进行了过拟合，对正常样本进行了欠拟合。

本文采用混合采样，先采用**smote算法**对异常数据进行插值，然后对插值后异常数据进行过拟合然后在单独对正常数据进行欠拟合。



**过拟合和欠拟合**



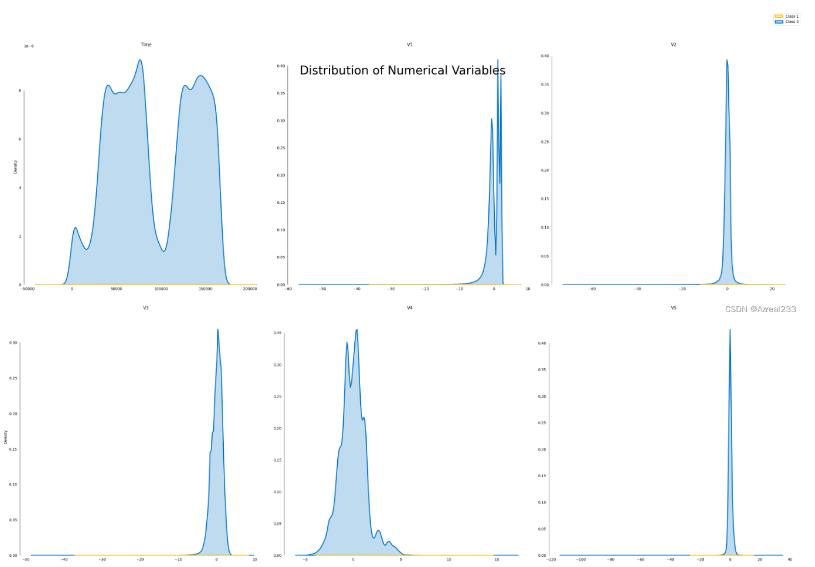
**数据可视化**

同学们可以选择不同的可视化方法对数据进行探索，为后续数据建模提供依据和前提。

**密度图**

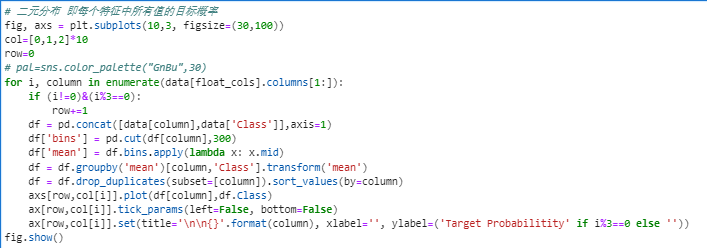
密度图的目的是提供数据潜在分布的视觉呈现。它可以帮助您理解数据的形状和分布，并识别任何异常值或离群值。

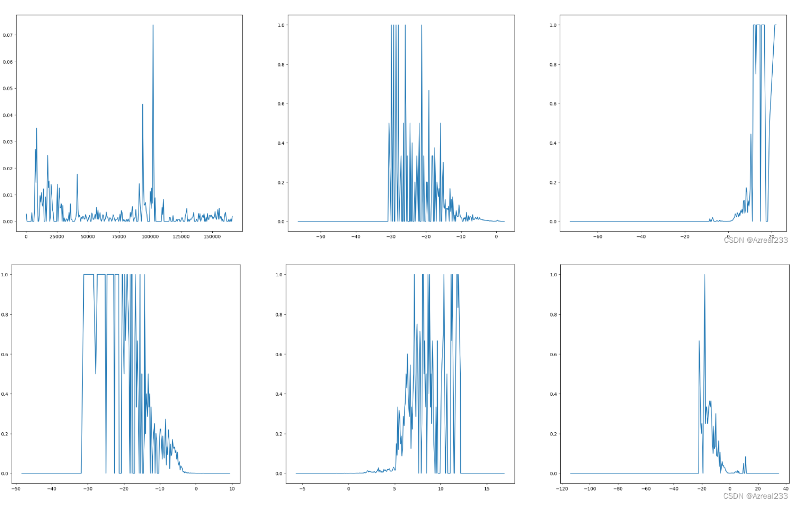




**二元分布图**

通过二元分布图查看每个特征中所有值的目标概率





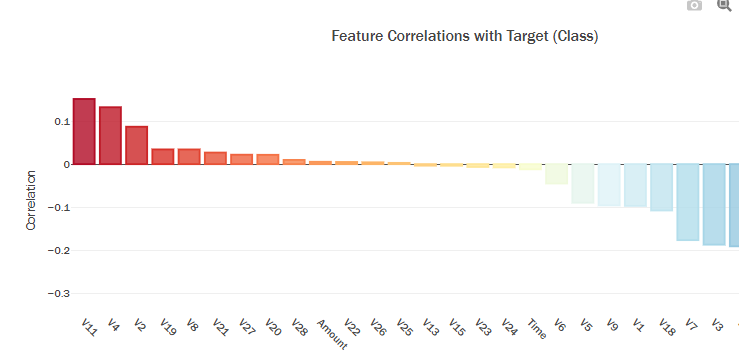
**热力图**

通过热力图直观反映属性两两之间的相关性。



**查看各属性与类别相关性**



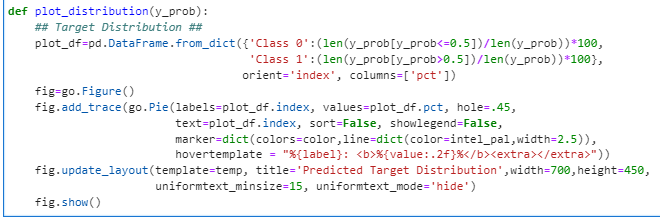


**模型拟合**

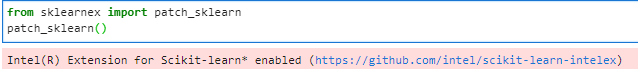
首先定义了三个函数，划分数据集、绘制绘制ROC/PR曲线和预测目标分布、绘制预测数据分布

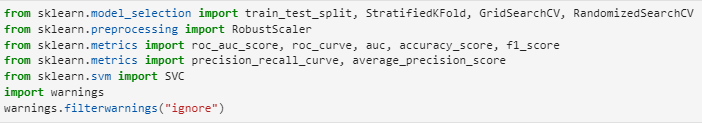






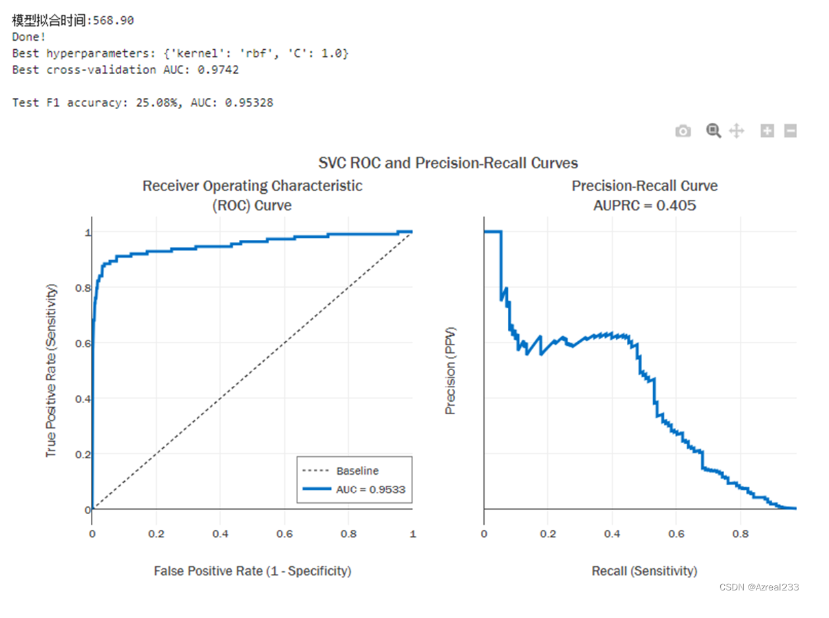
**引入Intel(R) Extension for Scikit-learn**





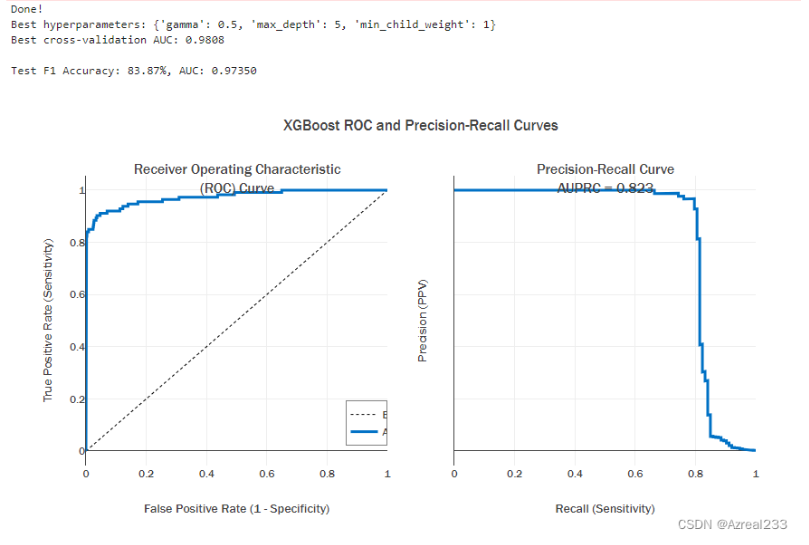
利用Intel AI Analytics Toolkit中Machine Learning模块中的intel extension for Scikit-learn进行模型拟合。主要是使用了支持向量机分类来进行拟合。





**XGBClassifier**





可见采用XGBClassifier效果更好，能够获得不错的F1分数和AUC值。