项目报告：从安然公司邮件中发现欺诈证据

项目概述

本项目的目标是准确预测给定数据集中有欺诈嫌疑的安然员工。通过对安然员工个人信息和邮件内容的探索，进行特征工程并运用监督学习算法，可以一定程度上预测有欺诈嫌疑的安然员工。机器学习是顺利进行项目的关键，因为它具有特征学习和预测目标的能力。

数据集概览

- 数据点总数：145。包含了143名安然员工和2个异常数据点（'TOTAL'/'THE TRAVEL AGENCY IN THE PARK'）。
- POI比例：15.3%。除异常数据点外，有19个POI和124个非POI。POI的比例较小说明测试算法性能时准确率（Accuracy）不是一个很好的指标。
- 使用的特征：
  - 财务特征：14个，均为整数（美元）；
  - 邮件特征：6个除"email_address"为字符串（邮箱地址），其他为整数（邮件数量）；由此可以判断，"email_address"不是一个合适的可用于机器学习的特征；
  - 具有大量缺失值的特征如下（超过100个缺失值），是选择特征时的重要参考依据：
    - 'deferral_payments': 106
    - 'loan_advances': 141
    - 'restricted_stock_deferred': 127
    - 'director_fees': 128
- 异常值：通过数据初步探索，发现了两个异常数据：
  - 'TOTAL'：该样本各数字特征的值为其他样本各数字特征值的和，它并不是真实存在的安然员工，因此对其做删除处理。
  - 'THE TRAVEL AGENCY IN THE PARK'：该样本同样不代表真实的安然员工，故删除。

特征选择和缩放

特征选择分为两步：

1. 设计新特征。新特征命名为'to_poi_fraction' 和 'from_poi_fraction'，分别表示收到和发送的邮件来自嫌疑人的比例。该特征的设计基于非常简单的想法：和嫌疑人有较多往来的人，自己本身就很有可能是嫌疑人。在之后的自动特征选择中，'to_poi_fraction'这一特征分数较高，显示出对算法的较大影响，因此纳入最终特征集。设计方法如下：

   - 'to_poi_fraction' = 'from_this_person_to_poi' / 'from_messages'
   - 'from_poi_fraction' = 'from_poi_to_this_person' / 'to_messages'
2. 自动选择特征。主要使用Sklearn中的SelectKBest方法，对所有特征（除了邮箱地址）进行筛选，保留特征分数最高的6个特征。特征及分数如下所示：

   - 'total_payments'(5.84)
   - 'exercised_stock_options'(18.85)
   - 'restricted_stock'(7.83)
   - 'from_poi_to_this_person'(4.42)
   - 'shared_receipt_with_poi'(6.21)
   - 'to_poi_fraction'(8.57)

本项目虽然尝试过特征缩放（参见代码注释），但由于最终采用的是不要求特征缩放的决策树算法，并且缩放后会降低算法性能（精确率和召回率均有所下降），因此选择不缩放特征。

算法选择

本项目共选择了朴素贝叶斯（基于高斯核函数）和决策树两种算法，测试代码结果如下：

- 朴素贝叶斯：Accuracy: 0.84 Precision: 0.35 Recall: 0.20 F1: 0.26 F2: 0.22
- 决策树算法：Accuracy: 0.82 Precision: 0.33 Recall: 0.35 F1: 0.34 F2: 0.34

由上可见，朴素贝叶斯算法仅在准确率（Accuracy）上有优势。综合来看，选择决策树算法更为适宜。

调整参数

调整算法的参数即改变算法的某些特征或执行方式，使得其能更好地适应特定的数据集，从而获得更好的预测效果。仅使用默认参数可能导致算法的效果不理想。

在本项目中，对决策树的max_depth参数进行不断调整测试，最终确定max_depth的值为10时算法性能较好。本项目尝试采用Sklearn的GridSearchCV方法。该方法可以自动为算法根据数据集选择最合适的参数。但对于决策树算法，该方法获得的参数值所确定的算法性能并不理想，故不采用。

验证

验证即将样本数据分为训练集和测试集，用训练集训练算法，用测试集测试算法性能并避免过拟合的方法。未正确验证的典型错误是没有打乱数据集的顺序，直接将有序的数据拆分。这样可能会导致算法准确率的测试结果过高或过低，无法反映算法的真实性能。

我的验证方法是采用Sklearn的StratifiedShuffleSplit方法，将70%的数据作为训练集，30%的数据作为测试集，共进行10次测试，取指标平均值以判断决策树算法的性能。由于POI所占比例较小，因此除了准确率以外，还引入精确率和召回率作为算法性能的评价指标，结果如下：

- 准确率： 0.83
- 精确率： 0.28
- 召回率： 0.28

评估度量

在该项目中，选择精确率（Precision）和召回率（Recall）作为算法的评估度量。本项目使用的决策树算法的各度量如下：

- 精确率: 0.33
- 召回率: 0.34

对两个度量的解读：

- 精确率：该指标越高，表示我们有更大的把握肯定，在测试集中标记的嫌疑人是真正的嫌疑人；其代价是，测试集中会有嫌疑人被算法遗漏，即未标记为嫌疑人。
- 召回率：该指标越高，表示在测试集中出现嫌疑人时，将其标出的可能性越大；其代价是，可能会错误地标注非嫌疑人；