找出“安然”事件嫌疑犯

项目目标

本项目的目标是通过公开的安然财务和邮件数据集，找出有欺诈嫌疑的安然雇员，属于监督学习当中的分类问题。利用机器学习可以从现有的数据集当中找到一种规则，一旦给定新的数据，便能根据其特征来判断 POI。

数据集概况

项目一共有 146 个数据点，其中 POI（嫌疑犯）有 18 人，POI 与非 POI 的比例为 14.4%。每个人的特征数为 20，且大多数特征存在缺失值，考虑到数据集实在很小，暂不进行处理。

异常值处理

通过对 salary 和 bonus 进行可视化，我发现了一个异常值，它的取值异常的高，通过查阅原始数据源，我发现它名为‘TOTAL’，这是在数据预处理中未将合计项去除导致的，故从数据中直接删除该异常值。另外，由于数据量很少，我又通过大致浏览来找异常，发现一项记录名为‘THE TRAVEL AGENCY IN THE PARK’，这显然不是人名，故直接删除。此外，LOCKHART EUGENE E 除了知道不是 POI，其他所有特征均为 NaN，不包含有用的信息，故将其删除。

创建新特征

我创建了一个名为'salary_bonus_ratio'的变量，旨在探索工资和奖金的比例能否帮助预测 POI，POI 可能有着与工资不匹配的奖金水平。经过测试发现，添加了该特征后模型的各项评分无任何变动，即对算法性能并无改进。

特征缩放

这里需要进行特征缩放，因为后面用到了 Knn 算法，分类器的实现根据点与点之间的欧几里得距离，如果一个特征比其它的特征有更大的范围值，那么距离将会被这个特征值所主导。因此每个特征应该被归一化，比如将取值范围处理为 0 到 1 之间。

特征选择

通过 SelectKBest，我选用了['salary'，'bonus'，'total_payments'，'exercised_stock_options'，'restricted_stock'，'total_stock_value'，'deferred_income'，'shared_receipt_with_poi']这 8 个特征，得分分别为 15.9、30.7、9、9.7、8.1、10.6、8.8、10.7。我只选了 8 个得分最高的特征，因为该项目的数据集本身就很小，如果特征过多容易导致过拟合，所以我需要带来尽量多信息的尽量少的特征。下表反映的是不同特征数下最终模型的评分情况，可以看到随着特征数增加，各项评分先升后降，当特征数为 7 或 8 的时候评分最好，而这个项目更关注找全 POI，所以选了 recall 更高的，即 8 个特征。

特征数	Precision	Recall	F1	F2
6	0.37458	0.22550	0.28152	0.24500
7	0.41082	0.30750	0.35173	0.32379
8	0.39409	0.31350	0.34921	0.32687
9	0.36639	0.31400	0.33818	0.32324
10	0.32550	0.31150	0.31834	0.31420

算法选择

最终使用了朴素贝叶斯算法，此外还尝试了 Knn 和 Decision Tree，这三个算法的 accuracy 很相近，但后二者的 Precision 和 recall 远远不如朴素贝叶斯，唯一满足 precision 和 recall 都大于 0.3 的只有朴素贝叶斯算法，具体情况如下（数据集本身的构造决定了 precision 和 recall 值不高）：

	Accuracy	Precision	Recall	F1	F2
GaussianNB	0.84420	0.39409	0.31350	0.34921	0.32687
KNN	0.83900	0.15589	0.04700	0.07222	0.05463
Decision Tree	0.85120	0.28986	0.08000	0.12539	0.09355

参数选择

调参就是通过调整算法的参数来提高模型的性能及防止过拟合。这里通过 GridSearchCV 来确定最佳效果参数，它的原理是在给定的参数范围内系统地遍历所有参数组合，通过对数据分组并交叉验证来自动选择最佳参数组合。

交叉验证

对于给定数据集，用大部分样本来训练模型，留小部分样本来防止模型过拟合以及测试模型好坏。未正确执行容易导致模型过拟合，对于新数据的泛化能力不强，也无法正确检验模型的好坏。

这里用 StratifiedShuffleSplit 对各模型进行交叉验证，该方法采用分层抽样，保证子集中 POI 与非 POI 比例一致，同时是取 1000 次验证的平均值，使得结果更具说服力和代表性。

结论解读

(1)recall（查全率）：0.31

即所正确判断出的 POI 占所有实际 POI 的比率，反映我们判断对的 POI 有多全。

(2)precision（查准率）：0.39

即预测为 POI 的人当中实际是 POI 的比率，反映我们判断出正确 POI 有多准。

这里不使用 accuracy 的原因是：accuracy 反映的是模型预测正确的比率，即正确判断出 POI 与非 POI 占数据总量的比例，因为非 POI 的人数远大于 POI，模型有很大概率能正确判断非 POI，而我们关注的是判断出多少 POI，所以 accuracy 对于项目的目的意义不大。