事件简介：

安然公司（**Enron Corporation**，台湾译**安隆**或**恩隆**；股票代码：NYSE：ENRNQ），曾是一家位于美国的得克萨斯州休斯敦市的能源类公司。在 2001 年宣告破产之前，安然拥有约 21000 名雇员，是世界上最大的电力、天然气以及电讯公司之一，2000 年披露的营业额达 1010 亿美元之巨。公司连续六年被《财富》杂志评选为“美国最具创新精神公司”，然而真正使安然公司在全世界声名大噪的，却是这个拥有上千亿资产的公司 2002 年在几周内破产，持续多年精心策划、乃至制度化、系统化的财务造假丑闻。安然欧洲分公司于 2001 年 11 月 30 日申请破产，美国本部于 2 日后同样申请破产保护。目前公司的留守人员主要进行资产清理、执行破产程序以及应对法律诉讼。从那时起，“安然”已经成为公司欺诈以及堕落的象征。

1. 【“数据探索”，“异常值调查”】

数据集包含 146 条记录，其中 18 个 poi，共有 21 个特征.

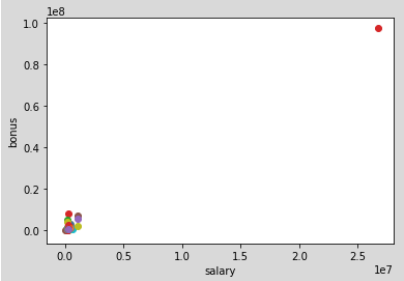
数据包含空值

特征	无效数据	有效数据	有效比
poi	0	146	1
salary	51	95	0.65
to_messages	60	86	0.59
deferral_payments	107	39	0.27
total_payments	21	125	0.86
loan_advances	142	4	0.03
bonus	64	82	0.56
email_address	35	111	0.76
restricted_stock_deferred	128	18	0.12
total_stock_value	20	126	0.86
shared_receipt_with_poi	60	86	0.59
long_term_incentive	80	66	0.45
exercised_stock_options	44	102	0.7
from_messages	60	86	0.59
other	53	93	0.64
from_poi_to_this_person	60	86	0.59
from_this_person_to_poi	60	86	0.59
deferred_income	97	49	0.34
expenses	51	95	0.65
restricted_stock	36	110	0.75
director_fees	129	17	0.12

数据中存在大量无效数据，对于无效数据含量过高的特征（有效比小于 0.2），应当剔除。

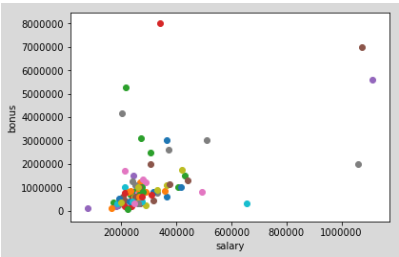
所以 restricted_stock_deferred 和 director_fees 不作为有效特征放入模

型。’email_address’作为文字类数据，不适合作为特征参与建模。也要移除。



对数据可视化后，发现有如下问题：

右上角出现偏离点，发现名为 TOTAL 的记录，显然不正常。



清除’TOTAL’后显示正常。
另外，名为“LOCKHART EUGENE E”的用户所有记录为 ”NaN”，也需要清除。
‘ THE TRAVEL AGENCY IN THE PARK’，非人名，删除。

2. 【“创建新特征”、“适当缩放特征”、“智能选择功能”】
创建特征：
from_poi_to_this_person 和 from_this_person_to_poi 分别反映了该雇员接收到嫌疑人员，以及发送给嫌疑人员的邮件数目。单纯的数字无法表现与 poi 的关系，所以使用该人员发送和接收自嫌疑人员与自身总发送接收邮件数目的比例。
新建 2 变量：

fraction_to_poi = from_poi_to_this_person / to_messages

fraction_from_poi = from_this_person_to_poi / from_messages

选择特征：
用 selectKBest 算法，算出除 poi 外得分最高的 2 个特征：'fraction_from_poi', 'fraction_to_poi'
通过 test_classifier 测试

```
['poi', 'fraction_from_poi']:                Accuracy: 0.70450  Precision: 0.15134   Recall: 0.03950 F1: 0.06265  F2: 0.04635
['poi', 'fraction_from_poi','fraction_to_poi']: Accuracy: 0.84137  Precision: 0.36980   Recall: 0.38200 F1: 0.37580  F2: 0.37950
```

算上 poi，3 个特征各项得分都高于 2 个特征，所以最终选择[‘poi’, ‘fraction_from_poi’,

'fraction_to_poi']3 个特征。

缩放特征：
特征缩放统一数据的范围值。贝叶斯和决策树对特征缩放不敏感但是用到的 SVM 算法，所以要统一缩放特征。

3. 【"选择算法"】
本项目中尝试用如下的算法，并使用 GridSearchCV 来调整算法的参数：
Gaussian Naïve-Bayes 朴素贝叶斯
Decision Tree Classifier 决策树
Support Vector Machines 支持向量机
RandomForest 随机森林
最后选用 Gaussian Naïve-Bayes 决策树算法，虽然 score 得分 SVM 最高，为 0.879，但 Gaussian Naïve-Bayes 在 precision 和 recall 及 F1 得分最高。Precision:0.45558 Recall: 0.3 F1: 0.36177
在性能上，Gaussian Naïve-Bayes 时间最短，而 RandomForest 耗时最长。

4. 【"调整算法"】
使用了 GridSearchCV()来调试各个算法的参数，将算法和相应的参数作为 GridSearchCV()的参数依次测试，得出最佳算法和参数组合。另外通过 test_classifier()测试了算法并给出了判断结果，经过调试最好的结果为 GaussianNB。不同的参数对最终的结果影响很大，某些情况下会造成过拟合。

5. 【"验证策略"】
验证是将训练出得模型，用测试数据进行评价的过程，验证中的典型错误是没有将数据分成训练和测试两部分，从而导致过拟合。验证使用 StratifiedShuffleSplit，对数据多次分割，确保训练集和测试集中 POI 与非 POI 的比例。

6. 【"评估度量的使用"】
评估度量使用准确率 Precision,召回率 Recall
关于精确率 Precision: 精确率(Precision)计算公式为： P = (TP) / (TP+FP)，表示被分为正例的示例中实际为正例的比例。在本项目中，精确率指的是模型预测出的 POI 中，真正为 POI 的比率。
关于召回率 Recall： 召回率是覆盖面的度量，度量有多个正例被分为正例，recall=TP/(TP+FN)=TP/P=sensitive,可以看到召回率与灵敏度是一样的。在本项目中，指的是所有真正的 POI 雇员中，有多少被真正的识别出来了。