优达学城数据分析师纳米学位项目 P5

安然提交开放式问题

说明：[你可以在这里下载此文档的英文版本](https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/Enron+Submission+Free-Response+Questions.pdf)。

机器学习的一个重要部分就是明确你的分析过程，并有效地传达给他人。下面的问题将帮助我们理解你的决策过程及为你的项目提供反馈。请回答每个问题；每个问题的答案长度应为大概 1 到 2 段文字。如果你发现自己的答案过长，请看看是否可加以精简！

当评估员审查你的回答时，他或她将使用特定标准项清单来评估你的答案。下面是该标准的链接：[评估准则](https://review.udacity.com/?&_ga=1.206413269.797580181.1464528540" \l "!/rubrics/310/view)。每个问题有一或多个关联的特定标准项，因此在提交答案前，请先查阅标准的相应部分。如果你的回答未满足所有标准点的期望，你将需要修改和重新提交项目。确保你的回答有足够的详细信息，使评估员能够理解你在进行数据分析时采取的每个步骤和思考过程。

提交回答后，你的导师将查看并对你的一个或多个答案提出几个更有针对性的后续问题。

我们期待看到你的项目成果！

1. 向我们总结此项目的目标以及机器学习对于实现此目标有何帮助。作为答案的部分，提供一些数据集背景信息以及这些信息如何用于回答项目问题。你在获得数据时它们是否包含任何异常值，你是如何进行处理的？【相关标准项：“数据探索”，“异常值调查”】

数据中的数据点总数为146，其中POI为18，非POI为128。可以使用的特征数量为20个。其中非缺失较少的特征为：工资(salary)数据为95，非缺失的email地址为111，total\_payments的缺失值为21(约占总数的14.4%)，其中POI的total\_payments不存在缺失值；deferral\_payments的非缺失值仅为39，loan\_advances的非缺失值为4，这样的变量由于包含的信息过少，因此不适合作为纳入模型的特征变量。

通过对salary和bonus两个变量绘制散点图，发现三条异常记录。其中一条键名为“TOTAL”，分析财务文档后发现该条记录总计，并非员工的名字。另外两条为'TRAVEL AGENCY IN THE PARK’和'LOCKHART EUGENE E'，其数据全部为缺失值。在数据清理过程中将这三条记录删除。

1. 你最终在你的 POI 标识符中使用了什么特征，你使用了什么筛选过程来挑选它们？你是否需要进行任何缩放？为什么？作为任务的一部分，你应该尝试设计自己的特征，而非使用数据集中现成的——解释你尝试创建的特征及其基本原理。（你不一定要在最后的分析中使用它，而只设计并测试它）。在你的特征选择步骤，如果你使用了算法（如决策树），请也给出所使用特征的特征重要性；如果你使用了自动特征选择函数（如 SelectBest），请报告特征得分及你所选的参数值的原因。【相关标准项：“创建新特征”、“适当缩放特征”、“智能选择功能”】

**新特征的创建**

每一个样本与POI的邮件往来数量是区分一个人是否是POI的有效特征。然而不同职位的员工所发邮件的总数不同，直接采用from\_poi\_to\_this\_person和from\_this\_person\_to\_poi并不能够提供识别嫌疑人的有效信息。可以通过邮件总数变量对这两个变量进行比例缩放。

生成两个新的变量：

propotion\_to\_poi = from\_poi\_to\_this\_person/to\_messages

propotion\_from\_poi=from\_this\_person\_to\_poi/from\_messages

**新特征性能的测试**

在未加入新特征时，使用所有其他特征训练的决策树模型表现为： Precision: 0.25，Recall: 0.2，F1: 0.22222

加入新特征propotion\_from\_poi后，模型表现为：Precision: 0.25，Recall: 0.2， F1: 0.2222

加入新特征propotion\_to\_poi后，模型表现为：Precision: 0.333333，Recall: 0.4，F1: 0.363636

可以发现，fraction\_to\_poi对模型表现提升较为明显，而fraction\_from\_poi没有为模型提供明显的提升。

**特征的筛选**

采用SelectKBest对所有的特征进行自动筛选，并根据各个特征的得分进行排序。通过GridSearchCV自动筛选模型表现最佳的的k值，最终筛选出得分前10的特征值以降低维度。报告结果如下：

exercised\_stock\_options 24.8150797332

total\_stock\_value 24.1828986786

bonus 20.7922520472

salary 18.2896840434

propotion\_to\_poi 16.409712548

deferred\_income 11.4584765793

long\_term\_incentive 9.92218601319

restricted\_stock 9.21281062198

total\_payments 8.77277773009

shared\_receipt\_with\_poi 8.58942073168

因此采用以上10个特征作为最终的训练特征。

由于采用了决策树算法，在各个节点仅根据单一特征进行分类，因此不需要进行特征缩放。

1. 你最终使用了什么算法？你还尝试了其他什么算法？不同算法之间的模型性能有何差异？【相关标准项：“选择算法”】

由于数据中POI数量为18，非POI数量为128，是一个分布不平衡的数据集，会导致accuracy值特别高。因此accuracy并不是一个合适的度量指标。

选择了三种不同的算法，各算法在默认参数下得到的模型性能如下：

朴素贝叶斯 : Precision: 0.25, Recall: 0.2, F1: 0.22

决策树： Precision: 0.375, Recall: 0.6 , F1: 0.4615

随机森林: Precision: 0.33, Recall: 0.2, F1: 0.25

最终选择了表现最好的决策树算法。

1. 调整算法的参数是什么意思，如果你不这样做会发生什么？你是如何调整特定算法的参数的？（一些算法没有需要调整的参数 – 如果你选择的算法是这种情况，指明并简要解释对于你最终未选择的模型或需要参数调整的不同模型，例如决策树分类器，你会怎么做）。【相关标准项：“调整算法”】

参数调整是对算法相应的可调整参数进行调试，使得模型的性能达到最佳。如果不进行参数调整，可能会导致模型过度拟合或欠拟合，或者模型的训练和预测速度过慢，从而导致模型性能无法达到最佳。

在本项目的算法调整中，使用GridSearchCV对参数进行优化调整，对SelectKBest的选择特征数量k、以及决策树的criterion、min\_samples\_leaf、min\_samples\_split、max\_depth进行调整。并选择了报告的最优参数设置。

1. 什么是验证，未正确执行情况下的典型错误是什么？你是如何验证你的分析的？【相关标准项：“验证策略”】

验证是使用训练集之外的独立测试集对训练好的模型进行验证，可以检验模型的泛化能力，防止模型过度拟合。未正确执行情况下的典型错误是没有正确使用训练集或测试集，可能导致使用相同的数据集对模型进行训练和验证，而使验证无效。

由于本项目中POI的比例较小，为了确保类之间的比例，在本文分析中使用StratifiedShuffleSplit交叉验证，进行多次洗牌和分割。

1. 给出至少 2 个评估度量并说明每个的平均性能。解释对用简单的语言表明算法性能的度量的解读。【相关标准项：“评估度量的使用”】

最终模型的精确度为 0.47006, 召回率:为0.46700, f1分数为0.46852。

在本项目中，Precision表示正确预测为POI的数据点占所有预测为POI的数据点的比例。模型的Recall表示正确预测为POI的数据点占所有实际为POI的数据点的比例。F1分数可以解释为精确率和召回率的一种加权平均，计算方法为: 。