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#!/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”】

**项目目标**：

运用机器学习技能构建一个算法，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。

**机器学习的作用**：

基于机器学习算法对大量数据进行处理，可以轻松获得强大的预测能力。

**数据集信息**：

数据点总数：146

类之间的分配（POI/非 POI）：18/128

特征总数量：20

是否有哪些特征有很多缺失值：有很多特征都有缺失值，比如salary（51个缺失值）、email\_address（35个缺失值）

**异常值检测和处理**：

首先包含一个明显的异常值，是源自表格录入时的bug（将TOTAL这一行写入了数据），因此选择将其删除。

另外还有几个异常值，其"salary"和"bonus"的数值明显高出其他绝大多数人很多。但这些异常值是正常而有效的数据，因为这几个人确实有理由获得明显高于他人的工资和奖金。所以选择对其不做处理。

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”】

**创建新特征**：

由安然欺诈案背景信息可知，安然高管为公司设计的架构主要依靠股价的上涨维持，而各高

管也都从股价的上涨中获得了巨大的利益，因此有理由相信，从股票中获得的总收益可能包

含较多的嫌疑人信息，因此基于现有与股票收入相关的特征设计新特征:

stock\_yield\_fraction=（restricted\_stock+exercised\_stock\_options）/total\_payments

根据后面的步骤，该创建的新特征因重要值较低没有被选中为最终特征，因此对最终算法的

性能无影响

**单变量特征选择**：

用决策树的feature\_importances\_ 属性获得各特征的重要性，取重要性大于0.1的特征：

'total\_payments': 0.112874779541

'bonus':0.109451528031

'expenses':0.205554989874

'exercised\_stock\_options':0.2

'restricted\_stock': 0.120480432567

**特征缩放**：

因为财务数据的数值范围相差较大，因此要进行特征缩放。

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”】

共尝试朴素贝叶斯、k近邻、决策树三种算法，三者的性能表现各有优缺点，具体见poi\_id文件，最终选择调整参数后的决策树，可以满足项目要求的性能标准。

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”】

**调整参数**：调参就是改变模型的参数。调整参数可能会对算法的性能起到很大提升作用，特征的处理和调参直接影响最终结果的好坏。

调整参数的方法有手动调参和网格搜索GridSearchCV等方法，在用决策树分类器时使用了网格搜索方法，并最终获得了符合项目性能要求的结果。

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”】

**验证及其重要性**：通过验证可以评估算法的性能，并且检查是否会发生过拟合问题。

**验证策略**：对train\_test\_split方法和交叉验证方法两种验证策略都有尝试，具体代码见poi\_id文件，因为数据集较小，应优先选择交叉验证。

工作原理：将原始数据一部分分为train\_data，一部分分为test\_data。train\_data用于训练，test\_data用于测试，重复多次进行划分，可以比较好的衡量算法的性能。

用StratifiedShuffleSplit将数据随机分成100组train/test。

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”】

精确度：算法识别出的poi是判断正确的比率。

召回率：算法正确识别出的poi占所有poi的比率。

优达学城

2016年9月