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”】

美国安然公司曾是世界上最大的电力、天然气及电讯公司之一。2002年由于大规模的财务造假丑闻而破产。在接受调查后，安然公司的大量保密信息被公布，其中包括数万封员工电子邮寄及高层详细的财务数据。本项目提供的数据集包含23项特征和146条记录。其中数据特征主要有三类：财务特征、电子邮件特征和POI标签。数据中POI数量为18，非POI数量为128，数据不平衡，因此accuracy并不适合作为次项目评估度量。此外，数据集中有多个特征包含大量缺失值。其中loan\_advances 有142个缺失值，director\_fees有129个缺失值，restricted\_stock\_deferred有128个缺失值，deferral\_payments有107个缺失值，deferred\_income有97个缺失值，long\_term\_incentive有80个缺失值。基于此数据集，本项目的目标是建立一个有效的机器学习模型来从安然员工数据中自动识别进行过造假行为的“嫌疑人（POI）”。机器学习算法可以帮助我们找出POI和非POI的各特征取值有哪些不同，发现其中的规律，一些算法可以找出分类的决策边界，最终建立起能够自动识别POI的模型。

在146条数据记录中，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”。通过分析可以发现该条记录为其它记录的总计，被错误地作为一名员工的数据信息纳入数据集，另外两条为'TRAVEL AGENCY IN THE PARK’和'LOCKHART EUGENE E'，其数据全部为NA。因此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”】

最终用于训练模型的特征包括：exercised\_stock\_options, other, expenses, shared\_receipt\_with\_poi, fraction\_to\_poi五个特征

首先，根据所给变量创建两个新特征fraction\_from\_poi和fraction\_to\_poi。通过可视化可以发现，原数据集中from\_poi\_to\_this\_person和from\_this\_person\_to\_poi两个变量并未提供识别嫌疑人的有效信息。因此考虑对这两个变量进行比例缩放。具体方法为：

fraction\_from\_poi = from\_poi\_to\_this\_person / to\_messages

fraction\_to\_poi = from\_this\_person\_to\_poi / from\_messages

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: 0.23007，Recall: 0.21500，F1: 0.22228

加入新特征fraction\_from\_poi后，模型表现为：Precision: 0.22851，Recall: 0.21800， F1: 0.22313，此特征的重要性为：0.0

加入新特征fraction\_to\_poi后，模型表现为：Precision: 0.32391，Recall: 0.30950，F1: 0.31654，此特征的重要性为：0.1361

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有效信息。

添加以上两个新变量后，根据各特征的决策树特征重要性对数据集中所有特征进行手动选择。首先，定义get\_feature\_importances函数，使用StratifiedShuffleSplit将数据划分为测试集和训练集，划分次数为1000次，使用每次的训练集训练模型并得到特征的重要性及模型表现，计算出平均重要性及平均模型表现。

将包含全部变量的列表输入函数的结果为

Mean feature importances:

[(0.1476, 'fraction\_to\_poi'), (0.1453, 'exercised\_stock\_options'), (0.1173, 'expenses'), (0.1142, 'shared\_receipt\_with\_poi'), (0.1142, 'other'), (0.0867, 'bonus'), (0.0524, 'total\_stock\_value'), (0.0371, 'restricted\_stock'), (0.0343, 'deferred\_income'), (0.0331, 'total\_payments'), (0.0293, 'long\_term\_incentive'), (0.0204, 'salary'), (0.0174, 'from\_this\_person\_to\_poi'), (0.0162, 'from\_messages'), (0.0091, 'deferral\_payments'), (0.009, 'from\_poi\_to\_this\_person'), (0.0076, 'fraction\_from\_poi'), (0.0073, 'to\_messages'), (0.001, 'loan\_advances'), (0.0006, 'restricted\_stock\_deferred'), (0.0, 'director\_fees')]

precision: 0.301367, recall: 0.302500, f1\_score: 0.282979

然后每步删去一个重要性最低的特征，指导模型表现不再提升，最终选择了4个特征，其重要性列表为：

[(0.2892, 'exercised\_stock\_options'), (0.2621, 'expenses'), (0.2307, 'shared\_receipt\_with\_poi'), (0.218, 'other')]

此时模型表现为：precision: 0.390040, recall: 0.431000, f1\_score: 0.384433

此外，对选定特征没有进行任何特征缩放，因为选用的决策树算法在各节点仅根据单一特征分类，因此不要求特征缩放。

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”】

选择算法过程中尝试了三中不同的算法，各算法在默认参数下得到的模型性能如下：

随机森林: Precision: 0.42105, Recall: 0.16400, F1: 0.23606

朴素贝叶斯 : Precision: 0.36066, Recall: 0.17600, F1: 0.23656

决策树： Precision: 0.41051, Recall: 0.43350 , F1: 0.42169

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”】

参数调整是对算法所提供的可供调整的参数进行设置，以使模型的性能达到最佳。 如果不进行参数调整，可能会导致模型的过度拟合或欠拟合，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，我使用了手动调整，每次调整一个参数，依次对决策树算法的criterion、splitter、max\_feature、max\_depth、min\_samples\_split五个参数尝试不同的设置，并根据性能评估结果为每个参数选择最佳设置。

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。

我在项目中使用了StratifiedShuffleSplit交叉验证。

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 0.43009, 召回率:为0.42450, f1分数为0.42728。

在本项目中，模型的精确度表示正确预测为POI的数据点占所有预测为POI的数据点的比例。模型的召回率表示正确预测为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: 。

优达学城

2016年9月