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”**】

此项目的目标是通过关于嫌疑人信息的数据集，找出安然公司中的嫌疑人。人为的对数据集进行处理和查找异常复杂，机器学习可以帮助我们从复杂的数据中找出嫌疑人的“patterns”，帮助我们更好更快地进行判断。

经过探索发现该数据集中数据总数有146个，其中POI个数: 18，非POI个数128个。因此该数据集非常不平衡，将所有员工全部预测为非POI也能得到85%的准确率，所以在此项目中，不采用accuray\_score作为检验数据集好坏的标准，而是采用Precision,recall和F1值来评估模型。

通过对数据集中的数据，采用箱线图进行可视化，发现数据集中的异常点-Total，并删除。

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”】

通过查找每个Feature的NaN值发现，NaN最多的元素为：loan\_advances、director\_fees、restricted\_stock\_deferred、deferral\_payments、deferred\_income、long\_term\_incentive，缺省值分别为：142，129，128，107，97，80考虑到数据中一共仅有146个样本，因此删除这些特征。并且在考虑与POI之间收发邮件数量的绝对数没有与POI之间收发邮件占比有效，设计新特征to\_poi\_ratio和from\_poi\_ratio,这两个特征是该员工总收发送邮件中发给POI的比例，计算公式为 ：给POI的收（发）邮件数量/该员工总收（发）邮件数量（例如，一个人可能给POI发过100封邮件，但是他总共发送了1000000封，那么其实比例就很小了）

**对于决策树：**

首先采用所有已选择的Features: 'salary', 'bonus', 'exercised\_stock\_options', 'expenses', 'to\_poi\_ratio', 'from\_poi\_ratio', 'restricted\_stock', 'shared\_receipt\_with\_poi', 'total\_payments', 'total\_stock\_value'。发现salary,expense, from\_poi\_ratio重要性都为0，因此删去。重复该过程，删除重要性太低的特征后，最终得到的特征为：'bonus'，'exercised\_stock\_options'，'to\_poi\_ratio'，和'shared\_receipt\_with\_poi'。特征重要性分别为0.29524807 0.22776736 0.11023286 0.3667517。并且得到了较高的F1值。

经过测试发现，在不使用to\_poi\_ratio而单独使用'from\_messages'','from\_this\_person\_to\_poi'时，precesion值和recall值分别约为0.36和0.34，但是在使使用to\_poi\_ratio这一新特征后，这两个评估值均提高到了约4.0左右。

**对于逻辑回归：**

先删除NaN太多的Feature，然后进行逻辑回归，逻辑回归里设置了L1 Penalty，因此删除所有系数为0的变量，但是最终，逻辑回归并没有带来很好的结果。

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”】

尝试了逻辑回归，逻辑回归得到了较高的Precision值，但是recall值得非常低，说明逻辑回归在此处很难检查出POI，只是基本上把所有员工都推断为非PO I。经过测试，认为逻辑回归不如决策树。

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”】

调整算法的参数是为了使得模型的结果更好，如果不调整参数，容易出现过拟合或者欠拟合。在决策树分类器中，因为POI的样本比例很小，所以我认为讲决策树的min\_samples\_split设置成2较为合适，虽然有过拟合的危险，但是能更准确的检验出POI，在寻找嫌疑人的模型中，我们宁可查错一个，也不愿漏查一个。

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”】

验证是为了提高算法的泛化能力，未正确执行下的典型错误是过拟合。

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”】

使用StratifiedShuffleSplit将数据进行类失衡处理，分为测试集和训练集，每次从数据集中抽取14个数据进行测试，并且迭代1000次，如此以来，可以避免因为POI样本太小而导致的一系列问题。

Precision：检索出来的条目有多少是准确的

Recall：所有准确的条目有多少被检索出来了，

F1值综合了这两者，计算公式为 正确率 \* 召回率 \* 2 / (正确率 + 召回率)。

优达学城

2016年9月