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”】

Enron used to be one of the largest companies in the US, but it had collapsed decades ago due to corporate fraud. My work in this project is to use machine learning techniques to build a model to identify persons of interest in this scandal.

This dataset contains 146 data points and 21 features, among which only one outlier is found – “TOTAL”. Additionally, all features of an employee called LOCKHART EUGENE E are merely valueless “NaN”. As a result, I remove both of them.

I leave “outliers” such as LAY KENNETH L and SKILLING JEFFREY K in the dataset because they are POIs.

I also notice that some features have large numbers of ‘NaN’ values, such as ‘loan advances’(141 NaNs), ‘director fees’(128 NaNs), and ‘restricted stock deferred’(127 NaNs).

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”】

Intuitively I keep all the features except “email address”, and create three new features: “fraction from this person to poi”, “fraction to this person from poi”, “and fraction shared receipt with poi”. Because I feel that it’s more accurate to use fraction of emails sent to or from POIs rather than number of messages sent to or from POIs while building a predictive model.

I then use SelectKBest module to select top 10 features for the model. The features and their scores are listed below.

|  |  |
| --- | --- |
| FEATURES | SCORES |
| exercised\_stock\_options | 25.097541528735491 |
| total\_stock\_value | 24.467654047526398 |
| bonus | 21.060001707536571 |
| salary | 18.575703268041785 |
| fraction\_from\_this\_person\_to\_poi | 16.641707070468989 |
| deferred\_income | 11.595547659730601 |
| long\_term\_incentive | 10.072454529369441 |
| restricted\_stock | 9.3467007910514877 |
| total\_payments | 8.8667215371077717 |
| shared\_receipt\_with\_poi | 8.7464855321290802 |

Later I also use rank the importance of features while using Decision Tree, and the related importance of features are listed below.

|  |  |
| --- | --- |
| FEATURES | IMPORTANCE |
| exercised\_stock\_options | 0.226 |
| shared\_receipt\_with\_poi | 0.190641505474 |
| total\_payments | 0.132454050374 |
| fraction\_from\_this\_person\_to\_poi | 0.127535296381 |
| long\_term\_incentive | 0.124537409493 |
| total\_stock\_value | 0.105377808033 |
| restricted\_stock | 0.0800041285968 |
| bonus | 0.0134498016478 |
| salary | 0.0 |
| deferred\_income | 0.0 |

I did not use feature scaling because the algorithms I select are not affected by it.

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”】

I tried Naïve Bayes, but eventually selected Decision Tree as final algorithm.

Their performances are listed below.

|  |  |  |
| --- | --- | --- |
| ALGORITHMS | PRECISION SCORE | RECALL SCORE |
| Naïve Bayes | 0.38622 | 0.31400 |
| Decision Tree | 0.33923 | 0.32600 |

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”】

Tuning the parameters means finding the best parameters that help reduce the risk of variance of bias, and increase accuracy score, precision score, and recall score. Naïve Bayes does not have any parameters to tune, so I leave it and tune the parameters of Decision Tree. I use GridSearchCV() module to automatically find the best parameters.

The before and after tuning results of Decision Tree algorithm are listed below.

|  |  |  |
| --- | --- | --- |
| DECISION TREE | PRECISION SCORE | RECALL SCORE |
| Before tuning | 0.33923 | 0.32600 |
| After tuning | 0.45589 | 0.46250 |

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”】

Validation assesses the performance of a model, and a valid model is able to properly make predictions. Without validation a model could overfit to the training data, and fail to perform well on other data. I use StratifiedShuffleSplit() module to create an iterator for cross-validation.

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”】

I use precision score, and recall score to evaluate the model. A precision score of 0.45589 means that the classifier is able not to label 45.589% of the positive samples as negative. A recall score of 0.46250 means that the classifier is able to find 46.25% of the positive samples.

优达学城

2016年9月