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。

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。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”】

   本项目中数据源是安然公司的员工邮件信息数据和员工工资和股票数据。其中包括被标为犯罪嫌疑人的数据，该项目的目标是使用机器学习算法得到能够识别犯罪嫌疑人的模型。机器学习算法可以从这些数据中找出犯罪嫌疑人存在的规律，从而提高犯罪嫌疑人识别率。

   数据中 salary 和 bonus 特征包含异常值，直接剔除异常值所在行。员工 LOCKHART EUGENE E 的所有字段均为空，删除该条数据。处理异常值之前，共 146 条数据，其中 128 条数据 poi 为 False，18 条数据 poi 为 True；处理异常值后，剩余 143 条数据，其中 126 条数据 poi 为 0，17 条数据 poi 为 1。

   deferral_payments、loan_advances、restricted_stock_deferred、director_fees 四个变量空值较多；email_address 这个变量为字符型变量，无法进行计算，也可忽略。将其他变量的空值用 0 替换。

2. 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”】

最终使用了 15 个特征，包括：['salary', 'total_payments', 'bonus', 'deferred_income', 'total_stock_value','expenses', 'exercised_stock_options','long_term_incentive', 'restricted_stock', 'to_messages', 'from_poi_to_this_person','from_this_person_to_poi', 'shared_receipt_with_poi', 'from_poi_ratio', 'to_poi_ratio']

全部变量
['salary', 'total_payments', 'bonus', 'deferred_income', 'total_stock_value', 'expenses', 'exercised_stock_options', 'other', 'long_term_incentive', 'restricted_stock', 'to_messages', 'from_poi_to_this_person', 'from_messages', 'from_this_person_to_poi', 'shared_receipt_with_poi', 'from_poi_ratio', 'to_poi_ratio']进入 SelectKBest，得分如下：
[1.23815616e+01 1.74217130e+00 1.39542854e+01 1.18497814e+01
1.63125076e+01 4.97016273e+00 1.68003518e+01 6.53156972e-03
5.23275955e+00 3.31485544e+00 1.05638708e+00 4.42830698e+00
1.31732482e-01 2.61649395e+00 7.08626103e+00 3.21442217e+00
1.40800024e+01]
剔除 1 分以下的变量后未最终使用变量。从得分可以看出产生的两个新特征 from_poi_ratio 和 to_poi_ratio 得分较高，说明这两个变量有较强的预测性。

对特征使用标准化缩放，因为在使用 PCA 时数据范围不同会对降维结果产生影响。

创建了 to_poi_ratio 和 from_poi_ratio 特征，计算发给 poi 邮件数占所有发邮件数量占比，从 poi 接收的邮件占所有收到邮件数量占比。

3. 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”】

最终使用 GaussianNB 算法，也尝试了 LogisticRegression 和 SVC 算法，GaussianNB 算法效果最佳。

	Accuracy	Precision	Recall	F1
GaussianNB	0.83647	0.37239	0.33050	0.35020
LogisticRegression	0.86400	0.08333	0.00200	0.00391
SVC	0.84047	0.30978	0.16000	0.21101

由于数据存在不平衡性，Accuracy 不能准确描述算法性能，GaussianNB 算法的 Precision、Recall 和 F1 得分最高，所以该算法性能最好。

4. 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”】

算法中有一些根据实际情况调整的参数，调整算法可以改变算法性能，如果不调整，算法使用默认参数，不能使结果达到最优。本项目中使用 pipeline 将数据处理过程与机器学习算法串联起来，使用 GridSearchCV 选择处理过程和算法中的最优参数组合。

5. 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”】

验证是使用测试数据集检验模型性能。由于本数据集存在严重的数据不均衡问题，如果未正确使用验证将大大影响模型最终预测结果。
使用 StratifiedShuffleSplit 将数据集进行多次分层抽样，使用训练集训练模型，使用测试集检验模型预测性能。

6. 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”】

		预测值		
		0	1	
真实值	0	True Negative(TN)	False Positive(FP)	TN+FP
	1	False Negative(FN)	True Positive(TP)	FN+TP
		TN+FN	FP+TP	TN+FP+FN+TP

Accuracy=(TN+TP)/(TN+FP+FN+TP)
Recall=TP/(TP+FN)评价模型从 0 中挑选出 1 的性能
Precision=TP/(TP+FP)
F1=2*Recall*Precision/(Recall+Precision)
由于样本存在偏差，F1 是用来评价模型的整体性能。

本项目中最终结果为

	Accuracy	Precision	Recall	F1
GaussianNB	0.83647	0.37239	0.33050	0.35020

说明在预测正确的数据中，有 33.05%的人可能为犯罪分子，在真实的犯罪分子中正确找出了 37.239%的犯罪分子。

优达学城
2018 年 1 月