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。

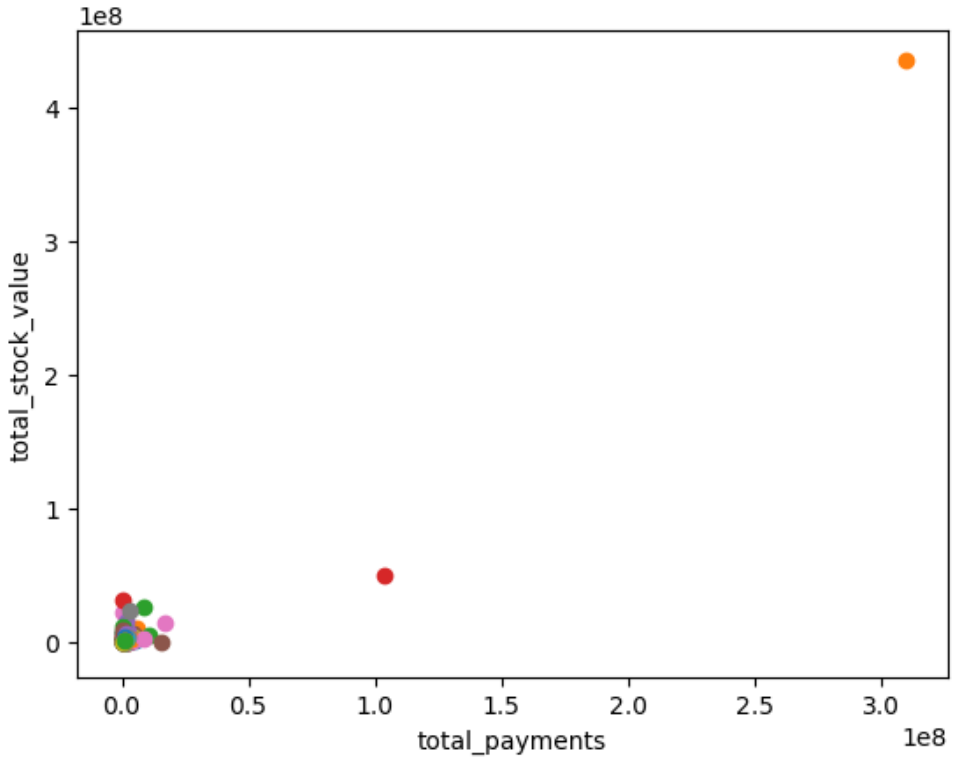
提交回答后，你的导师将查看并对你的一个或多个答案提出几个更有针对性的后续问题。

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”】

   答：此项目的目标是利用公开的安然财务和邮件数据集，构建一个机器学习算法，找到有欺诈嫌疑的安然雇员。机器学习可以对数据进行学习并做出预测，帮助我们找出区分是否属于 POI 的若干特征。
   安然公司曾被视为美国最具创新力的公司，有着千亿元的收入和两万名员工，由于管理体制的破败和受贿，包括人为制造能源危机致使加州大停电，安然公司最终面临破产，大批人员被判入狱。在调查过程中，美国联邦能源管理委员会收获了成千上万涉及高管的邮件和详细的财务数据，根据这些公开的财务和邮件数据，我们可以构建相关人士识别符，来识别有欺诈嫌疑的人员。
   加载数据集后，发现该数据属于字典结构，每个键代表一个人，值是个人的相关信息。通过查看数据，发现共有 146 条信息；POI 与非 POI 的人数分别是 18 和 128；除去标签‘poi’，总共包含了 20 个特征；某些特征缺失值较多，缺失比例一半以上的特征如下：
   'loan_advances': 142,
   'director_fees': 129,
   'restricted_stock_deferred': 128,
   'deferral_payments': 107,
   'deferred_income': 97,
   'long_term_incentive': 80
   除了'email_address'特征，其他特征均为数值，所以对除了'email_address'以外

特征的缺失值做置 0 处理。
选取"total_payments", "total_stock_value"两个特征做散点图，发现两个异常值，对应的名字分别是‘total’和‘LAY KENNETH L’，前者应该删除，而后者是安然的董事，总股票价值和总收入远远高于其他人属于正常现象，应该保留。



总体来看，该数据集非常的不平衡，poi 类别占比很低，所以不适合用 accuracy 来评估，precision 和 recall 能提供更好的评估性能。同时考虑使用 Stratified Shuffle Split 来做交叉验证，以降低数据不平衡性产生的影响。样本数量不算很多，可以使用 GridSearchCV 来进行参数调整。

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”】

答：考虑到嫌疑人之间互通邮件的频率可能会比嫌疑人和非嫌疑人之间频率更高，我决定建立两个新特征，分别为 fraction_from_poi(收到从嫌疑人发来的邮件数占总收件数的比列)和 fraction_to_poi(发送给嫌疑人的邮件数占总发件数的比例)。并且我感觉那些缺失值比例较高的特征可能会影响到最终的评估结果，所以决定删除缺失比例大于 50%的特征，分别是：
'loan_advances': 142, 'director_fees': 129, 'restricted_stock_deferred': 128,

'deferral_payments': 107, 'deferred_income': 97, 'long_term_incentive': 80
保留下来进行训练的特征为：
['poi','salary','total_payments','bonus','total_stock_value','expenses','exercised_stock_options','other','restricted_stock','to_messages','from_poi_to_this_person','from_messages','from_this_person_to_poi','shared_receipt_with_poi','fraction_from_poi','fraction_to_poi']
接下来我打印了重要性大于 0.1 的 feature_importances_： ['poi', 'salary', 'total_payments']。用这个列表进行测试，但是得分不理想，我转而使用 SelectKBest 进行特征选择，手工尝试了多个 K 值，结果发现当 K=4 时，Precision 和 Recall 的得分均最高。Featrure_list: ['poi', 'total_payments', 'bonus', 'expenses']

K	Precision	Recall
4	0.36320	0.38100
5	0.32975	0.33700
6	0.30496	0.32600
7	0.28041	0.29850
8	0.28747	0.30400
9	0.27099	0.26950

最后我进行了参数调整，最终测试得分为：Precision: 0.35055 ,Recall: 0.56000 与之前的分数相比，Recall 分数得到了大幅提高。

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”】
答：我选择朴素贝叶斯、svm、决策树三种算法进行测试。因为选择的测试算法中包含 SVM，所以我首先对原始数据集进行了特征缩放，但是默认特征保持不变。三种算法都是采用默认参数。测试结果如下：
GaussianNB : Precision: 0.28298 Recall: 0.19200 F1: 0.22878
SVC: Precision: 0.75000 Recall: 0.00300 F1: 0.00598
DecisionTreeClassifier: Precision: 0.33797 Recall: 0.31600 F1: 0.32661
根据测试结果，我选择了决策树算法进一步进行特征选择及参数调整。

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”】
答：算法调整是优化机器学习算法的参数以实现给定数据集上的最大性能的过程，它可以优化分类器的性能，并解决过拟合和欠拟合等问题。我选择的算法是决策树，使用 GridSearchCV 进行参数调整，选择调整的参数是 min_samples_split（分割内部节点所需的最小样本数），max_depth（树的最大深度）和 class_weight(样

本各类别的的权重)，最终得出最优的参数组合是：{'min_samples_split': 20, 'max_depth': 3, 'class_weight': 'balanced'}。参数调整前后最大的区别是 Recall 得分，调整前为 0.38，调整后为 0.56。

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”】

   答：验证就是将测试数据输入训练好的模型，将得出的结果与实际结果进行对比，评估模型的性能。未正确执行验证的典型错误是无法得知模型的泛化能力，可能只是在训练数据集上表现良好。我使用项目提供的 test.py 脚本进行验证，脚本所使用的验证方式为交叉验证，使用到了分层抽样（StratifiedShuffleSplit）来确保每次取样的数据中包括所有的类别数据，折数为 1000，使算法的评估更加准确。

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”】

   答：当分类非常不平衡时，Precision 和 Recall 是比较有效的指标，而项目中的 POI 与非 POI 的比例也是非常的不平衡，所以很适合用这两个指标来评估。Precision 就是检索出来的条目中有多少是准确的，Recall 就是所有准确的条目有多少被检索出来了。最终算法的准确率与召回率分别是 Precision: 0.35，Recall: 0.56。

优达学城
2017 年 12 月