安然提交开放式问题

email:waynewqh@foxmail.com

项目概述

安然曾是 2000 年美国最大的公司之一。2002 年，由于其存在大量的企业欺诈行为，这个昔日的大集团土崩瓦解。 在随后联邦进行的调查过程中，大量有代表性的保密信息进入了公众的视线，包括成千上万涉及高管的邮件和详细的财务数据。 你将在此项目中扮演侦探，运用你的新技能，根据安然丑闻中公开的财务和邮件数据来构建相关人士识别符。 为了协助你进行侦查工作，我们已将数据与手动整理出来的欺诈案涉案人员列表进行了合并， 这意味着被起诉的人员要么达成和解，要么向政府签署认罪协议，再或者出庭作证以获得免受起诉的豁免权。

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分析给定的安然员工财务数据及邮件数据，找出员工中有欺诈行为的嫌疑人。由于数据中已给定了欺诈嫌疑人的识别符（person of interest/POI），因此此项目可以通过机器学习中的监督学习算法，对数据集特征数据及标识符进行训练，找出特征数据和识别符之间的关系，从而有效识别出有欺诈嫌疑的员工。

数据探索

在进行项目前，我使用了pandas的dataframe来对原始数据进行初步分析。此数据集是一个二维的表格类型数据集，包含146个数据点样本以及21个变量。其中每个样本代表一个安然员工，而变量则包含了14个财务变量、6个邮件变量以及标识符’poi’。数据集中变量’poi’被标识为True的一共有18人，占比12.33%。

此数据集中的缺失值很多，有6个变量的缺失值超过数据点数量的一半。由于数据集不大缺失值又多，因此，大多数财务变量的分布并非近似正态分布，而且有着不少离群点，因此在训练时我认为填补缺失值的方法应该取用中位数。

异常值调查

在对数据集进行初步分析时，发现其中有着最多离群值的样本为’TOTAL’，这个数据点是对所有员工的各项财务数据值的汇总，可以从数据集中排除掉。其余有离群值变量的样本，经查证应是安然的高管，因此有着更高的收入，应予以保留。

此外，在分析时找到5个样本的变量有着超过18个缺失值，其中样本’THE TRAVEL AGENCY IN THE PARK’，经查证并非代表安然员工，应该排除。

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 SelectKBest），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”】

创建新特征

我创建了两类新特征，一类是把邮件变量标准化。我们希望邮件变量特征能体现与poi交往越频繁的人，本身是poi的可能性是否越大这一想法。但由于每个员工收发邮件（'from_messages'和'to_message'）的数量相差很大，因此与poi交往有关的变量（from_this_person_to_poi/from_poi_to_this_person）表现并不能单纯描述上述思想。在此建立新变量:
fraction_from_poi = from_poi_to_this_person / to_message
fraction_to_poi = from_this_person_to_poi / from_message
以区间[0, 1]的分数形式来体现每个员工与poi之间的交往密度。

另一类是把财务变量标准化。由于在进行数据分析时，发现财务数据并非近似正态分布，有一些高层员工的收入比大多数员工差了好几个数量级。由于高层员工中是poi的可能性很高，在进行回归时有可能会给有着极大离群值的变量给出更高的权重，从而可能会降低对一般员工的预测效果。因此我把财务变量中在样本里不含负数的变量都利用log函数进行转换，减少离群值对变量的影响。

由于选择尝试的所有模型中，用新的邮件特征对原始特征进行替换，都可以有效提升评价。
不过若在此之上，用新的财务特征替换原来的财务特征，支持向量机和决策树的评价会有所下降，而朴素贝叶斯和逻辑回归的评价会上升。由于最后选用的是朴素贝叶斯GaussianNB算法，下面利用tester中的test_classifier给出该算法下各特征选择的评价。

评估度量	原始特征	替换邮件特征	替换邮件特征+替换财务特征
Accuracy	0.84900	0.85053	0.87313
Precision	0.40596	0.42568	0.53745
Recall	0.28600	0.34650	0.34800
F1	0.33588	0.38203	0.42246

适当缩放特征

在尝试支持向量机SVM算法时，使用了MinMaxScaler对特征进行缩放，其他尝试的算法如朴素贝叶斯、决策树和逻辑回归等，并不需要进行特征缩放。

智能选择功能

由于最终选择的算法是朴素贝叶斯GaussianNB，因此在对算法调参时使用了SelectKBest来自动选择特征，并且使用PCA来对特征的主成分进行分析。最终的模型自动选取了11个特征，以及3个主成分，其特征得分如下：

特征	特征得分
other_log	14.9023451972
total_stock_value	14.6913086526
expenses_log	12.4800258499
bonus_log	9.3011642725
fraction_to_poi	8.24301643826
salary_log	7.87538537929
restricted_stock	6.57694907676
shared_receipt_with_poi	5.49532137993
total_payments_log	5.44463781522
deferred_income	5.30405931551
fraction_from_poi	2.34933760922

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”】

选择算法

我尝试了朴素贝叶斯（sklearn.naive_bayes.GaussianNB）、支持向量机（sklearn.svm.SVC）、决策树（sklearn.tree.DecisionTreeClassifier）以及逻辑回归（sklearn.linear_model.LogisticRegression）。最后决定使用朴素贝叶斯算法，一个原因是朴素贝叶斯最终训练结果的评价在四种算法之中表现最好，其次是朴素贝叶斯算法本身并不需要调参。以下给出经过调参的四种算法对替换了新特征的数据进行训练后的评价。

评估度量	朴素贝叶斯	支持向量机	决策树	逻辑回归
Accuracy	0.87313	0.83067	0.78907	0.87033
Precision	0.53745	0.35638	0.20457	0.52436
Recall	0.34800	0.33500	0.20150	0.29600
F1	0.42246	0.34536	0.20302	0.37840

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”】

调整算法

调整算法的参数是指通过遍历算法中参数的组合，使得模型输出和实际观测数据之间达到较佳的拟合程度。由于算法和数据的复杂性，若只用默认参数而不进行调整，难以让算法适应训练数据。

我在一开始尝试SVC算法对原始特征进行训练时，并没有对参数进行调整，只得出了’Accuracy: 0.86627，Precision: 0.36364，Recall: 0.00400，F1: 0.00791’的评价。通过查找文档，我认为对该算法影响较大的几个参数是内核kernel、参数C和gamma，利用GridSearchCV查找对评价度量’f1’最优的模型，对参数进行一定范围的遍历，最终得出的最优模型为’kernel = ‘sigmoid’， C = 10.0， gamma = 100’，评价为’Accuracy: 0.82667，Precision: 0.33749，Recall: 0.31150，F1: 0.32397’。调参后输出的模型在Recall这一评估度量上大大提升。

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”】

验证策略

验证是指把原始数据集分为不相交的训练集和测试集，利用训练集来训练模型，用测试集来评估模型。如未对数据进行独立测试集验证，把所有数据都进行训练，出现的一个典型的错误是模型可能是对原始数据集过度拟合，以至于模型不利于推广至新的数据集。另一个典型错误是，在进行数据集拆分为训练集和测试集时，特别是在有序的数据集上，未做到随机分配，可能会导致同一类型的标识符过于集中在训练集或测试集上，使训练出来的模型评价可能会失真。

在对样本数较少的数据集进行验证时，可采用交叉验证法。我在验证时通常使用tester中的方法即StratifiedShuffleSplit，此方法能够以固定比例对数据集随机划分，同时大幅增加预测次数，对小样本的评价效果较好。

此外我还尝试了十折交叉验证法（KFold）进行验证，即把数据集随机平均分成十份，每次用其中的一份进行测试集，其余九份作为训练集训练模型，一共进行十次训练和预测直至全部十份数据集都作为过测试集，最后把十次验证的各个评估度量取均值。在最终选择的模型中，得出评价为’Accuracy: 0.880，Precision: 0.45，Recall: 0.4167，F1: 0.4067’。

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”】

评估度量的使用

常用的评估度量有准确率accuracy、查准率precision、查全率eecall和f1。以上评估度量是一个模型预测和真实情况的二分类问题，其中准确率指模型预测正确的样本数占样本总数的比例，查准率

是指模型预测的真正例占所有预测正例的比例，查全率是指模型预测的真正例占所有真正例总数的比例，f1是对查准率和查全率的综合评价。

如下面矩阵所示：

	预测正例	**预测反例**
真实正例	TP（真正例）	FN（假反例）
真实反例	FP（假正例）	TN（真反例）

accuracy = (TP + TN) / (TP + FN + FP + TN)
precision = TP / (TP + FP)
recall = TP / (TP + FN)
f1 = 2 * precision * recall / (precision + recall) = 2 * TP / (2 * TP + FN +FP)

对于本次项目的数据集，由于反例（不是poi）占样本总数的比例很大，模型很容易就对反例预测正确，因此准确率accuracy稳定在一个偏大的区间里。即使对正例预测正确与否，也不会对准确率影响太多，因此准确率对本次项目并不是一个好的评估度量。此项目的目标是希望把真实的嫌疑人尽量多地找出来，因此查全率recall很重要，此外我们也不希望浪费太多资源在调查没有嫌疑的员工身上，所以查准率precision同样重要。在模型评估上，应尽量对这两个度量进行综合评价，因此我在选取最优模型时，以查找最优的f1为标准。