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

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:

目标是基于公开的安然数据,通过构建一个有监督分类模型,识别出欺诈嫌疑人.

数据集总共有146条记录,一共20个属性,不包括 poi. 其中 poi 为 true 也就是嫌疑人人数是18人.

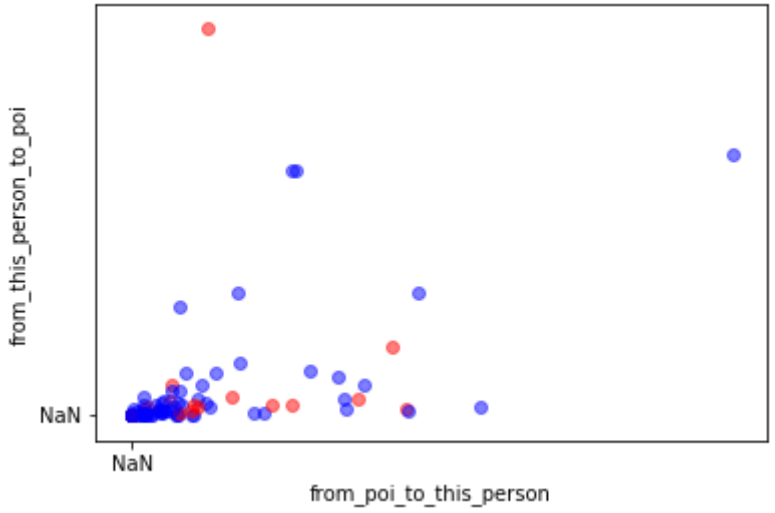
通过数据可视化,排除离群点,剔除了 key 为 TOTAL 的记录.

另外通过判断该条记录的 value 值有15个以上没有值,检索出了异常值,在通过肉眼观察剔除了THE TRAVEL AGENCY IN THE PARK 和 LOCKHART EUGENE E.前者为一个代理机构,后者所有特征全部是 NaN.

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:

当前数据集中, from_poi_to_this_person 和 from_this_person_to_poi 分别反映了改雇员接受到嫌疑人员以及以及发送给嫌疑人员的邮件数目。将这两个特征与poi标签可视化后如下图所示(红色表示POI雇员，蓝色表示非POI雇员)：

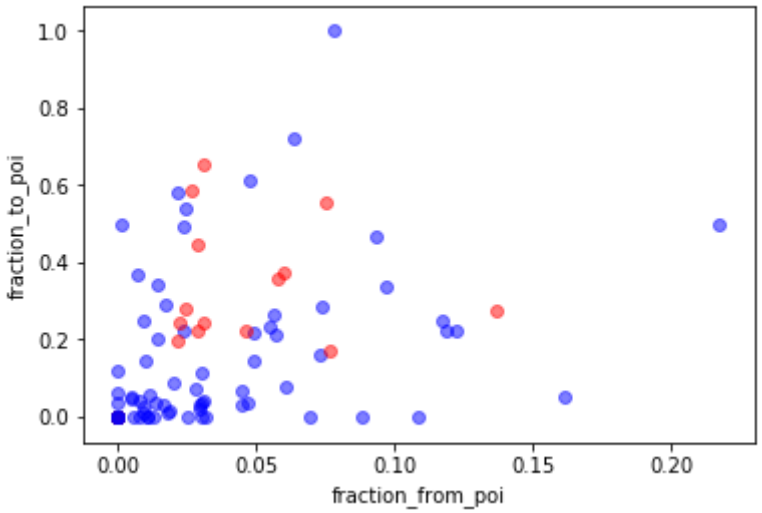


并没有发现明显规律,由于每个人处理邮件的数量不同,所以对邮件书做一个缩放:

fraction_from_poi = from_poi_to_this_person / to_messages

fraction_to_poi = from_this_person_to_poi / from_messages

将新标签重新可视化



可以看出,如果某个雇员的上述两个新特征低于某个值后,改雇员有欺诈嫌疑的概率就很小.

创建了新特征后.通过 SelectKBest算法,选出了最好的5个特征:

[('exercised_stock_options', 19.52327171588962), ('total_stock_value', 18.736172414610802), ('bonus', 15.209747974891794), ('salary', 12.305568237566376), ('fraction_to_poi', 11.587179422459965)]

上述5个特征中,前面4个都是收入,后面是发送给 POI 雇员的邮件占比.值的大小不在一个维度上,所以需要采用特征缩放,通过 MinMaxScaler 进行特征缩放.将值缩放到[0,1]之间.

决策树是针对每个特征分别进行判断，比如通过年龄进行判断时，如果单位是月份，5岁以上就是60月以上，以年为单位，5岁以上就是5年，虽然判断单位变了，但是最终判断的结果不变。

线性回归，回归方程式为 y = ax1 + bx2 + c*x3 + m ，如果x1缩小一半，那对应系数a就会增加一倍，所以整体没有变化，故特征缩放对其没有影响。

但是SVN和k均值聚类，都是要计算特征点到线之前的距离，如果特征值发生缩放，则距离值发生变化，最终分类也会有变化。

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:

最终使用了 GaussianNB算法.
总共用了5总算法:
Gaussian Naïve-Bayes朴素贝叶斯
Decision Tree Classifier决策树
Support Vector Machines支持向量机
RandomForest随机森林

- 时间消耗如下:
- GaussianNB score time: 0.023 s

- Decision Tree Classifier: 1.842 s
- Decision Tree Classifier: 0.071 s
- RandomForest 47.853 s
-

再加上对算法的评分结果,综合判断Precision,Recall,F1.最终选定的GaussianNB

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:

因为数据样本较少,所以使用GridSearchCV 来进行参数调参.通过通过 test_classifier() 测试了算法并给出了判断结果,通过调试得出最好的结果是的GaussianNB

对参数的调试是机器学习中很重要的一环,因为不同的算法函数和初始设定会对最终结果产生很多影响。在某些情况下，为算法选择了错误的参数，会造成过拟合。

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”】

答:

验证是将训练出得模型，用测试数据进行评价的过程，验证中的典型错误是没有将数据分成训练和测试两部分，从而导致过拟合。

在交叉验证的时候，因为数据的不平衡性，选用StratifiedShuffleSplit，StratifiedShuffleSplit是一种交叉验证的方式，通过对数据进行多次洗牌和分割，能够确保训练集和测试集中POI与非POI的比例，比较适合于该数据。 最终得到评价结果如下：

	GaussianNB	Decision Tree	SVM	RandomForest
Accuracy	0.83685	0.85000	0.85000	0.81392
Precision	0.45225	0.55297	0.94643	0.33696
Recall	0.28650	0.13050	0.02650	0.21650

F1	0.35078	0.21117	0.05156	0.26362
F2	0.30916	0.15404	0.03289	0.23317
Total Predictions	13000	13000	13000	13000
True positives	573	261	53	433
False positives	694	211	3	852
False negatives	1427	1739	1947	1567
True negatives	10306	10789	10997	10148

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:

最后四个参数的意义:

True Positive(真正，TP)：将正类预测为正类数
True Negative(真负，TN)：将负类预测为负类数
False Positive(假正，FP)：将负类预测为正类数误报 (Type I error)
False Negative(假负，FN)：将正类预测为负类数→漏报 (Type II error)

对应的预测类别表:

	yes	no	总计
yes	TP	FN	P(实际为 Yes)
no	FP	TN	N(实际为 No)
总计	P'(被分为 yes)	N'(被分为 No)	P+N

关于精确率Precision: 精确率(Precision)计算公式为： P = (TP) / (TP+FP)，表示被分为正例的示例中实际为正例的比例。在本项目中，精确率指的是模型预测出的POI中，真正为POI的比率。

关于召回率Recall：召回率是覆盖面的度量，度量有多个正例被分为正例，recall=TP/(TP+FN)=TP/P=sensitive，可以看到召回率与灵敏度是一样的。在本项目中，指的

是所有真正的POI雇员中，有多少被真正的识别出来了。

关于综合评价指标F1：Precision与Recall有时候会出现矛盾，如上表所示，这时就需要综合考虑他们，最常见的方法是F-Measure,F-Measure的计算公式如下：

```
F = (a**2 + 1)*P*R / a**2(P+R)
```

F是Precision与Recall的加权调和评价，当a=1时，就是最常见的F1，公式为 F1 = 2*P*R / (P+R)，可见F1综合了P和R的结果，当F1较高时则能说明该模型效果不错，上表中也显示GaussianNB的F1为其中的最高值，为0.35078。

优达学城
2016年9月