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

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。

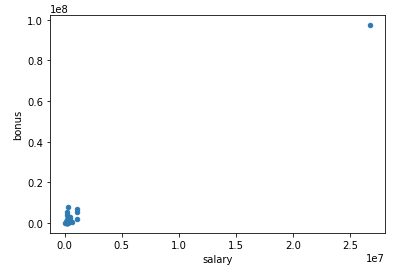
我们期待看到你的项目成果！

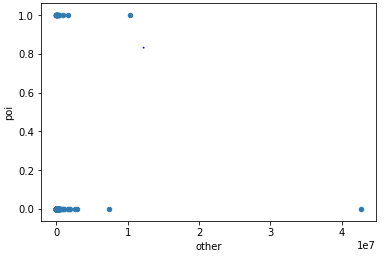
1. 向我们总结此项目的目标以及机器学习对于实现此目标有何帮助。作为答案的部分，提供一些数据集背景信息以及这些信息如何用于回答项目问题。你在获得数据时它们是否包含任何异常值，你是如何进行处理的？【相关标准项：“数据探索”，“异常值调查”】

本项目目标是根据公开的安然财务和电子邮件数据集，构建机器学习算法，找出有欺诈嫌疑的安然员工，我使用的数据是final\_project\_dataset.pkl，它是处理过的E+F数据，机器学习将帮助我们从已用的数据集里生成模型然后对新数据集做预测，得出谁是嫌疑人的结果，最后对结果进行评估。

该数据集有146条目标数据，有21个特征。选定其中的几个名称和结果相关性可能较高的特征做可视化异常值分析。

对bonus和salary做可视化观察发现一个TOTAL的异常值

  
通过对other和poi作可视化观察发现同样有个叫TOTAL的异常值



输出该目标查看各个特征

  
（图片可放大查看）  
再通过字面分析应该是指各特征的各自所有目标的总和，应该是处理数据的时候携带了电子表格所统计的值，所以应该要把这个TOTAL目标给去掉。

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”】  
     
   总体上说改数据集缺失的比较严重，除了最后的训练结果poi以外的所有特征没都含有缺失值。

我这里需要剔除的特征，缺失超过一半的特征我就先删除了，该数据集的总目标数是146也就是低于73的我就直接排除了。

缺失超过一般的特征由于数量不够而且部分数据比如restricted\_stock\_deferred的对象里完全没有poi，因此容易对最后的结果造成很大的误导故不选作为训练特征。

保留下的特征有

email\_address 111 non-null object

exercised\_stock\_options 102 non-null float64

expenses 95 non-null float64

from\_messages 86 non-null float64

from\_poi\_to\_this\_person 86 non-null float64

from\_this\_person\_to\_poi 86 non-null float64

other 93 non-null float64

restricted\_stock 110 non-null float64

salary 95 non-null float64

shared\_receipt\_with\_poi 86 non-null float64

to\_messages 86 non-null float64

total\_payments 125 non-null float64

total\_stock\_value 126 non-null float64

接着移除email\_address原因是与结果无关， 剔除的特征有

email\_address 111 non-null object

deferral\_payments 39 non-null float64

deferred\_income 49 non-null float64

director\_fees 17 non-null float64

loan\_advances 4 non-null float64

long\_term\_incentive 66 non-null float64

restricted\_stock\_deferred 18 non-null float64

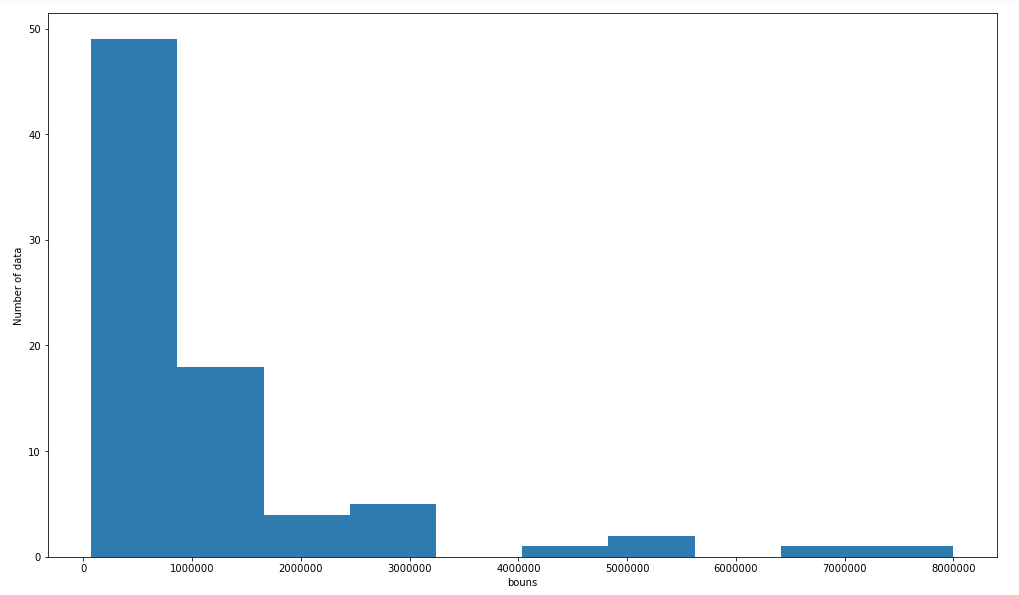
接下来我再在这些保留的特征里进行建立新的特征和选择原有的特征作为训练特征

新建特征，观察邮件的首发情况将'from\_this\_person\_to\_poi','from\_messages','from\_poi\_to\_this\_person','to\_messages'这四个数量的特征转化为相关的与poi接收和发送率即from\_ratio和to\_ratio。将没有值的目标用中位数来补充，最后去掉这四个特征。

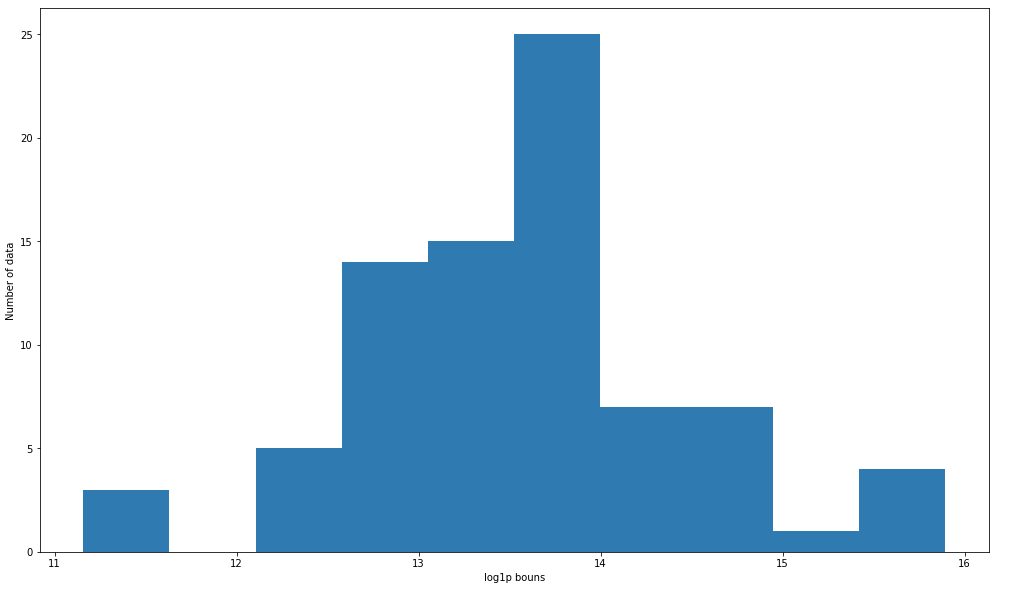
特征缩放，观察其余的各个留存的特征均为数值型，但是数值之间的差距巨大很多不在同一范围内，我将逐一的来查看各个特征

bonus斜态分布，且数值相差巨大大部分都集中在1000000以下，我选择做一个对数转化同时将其转化为正态分布，这样的好处是，它可以额加快梯度下降求最优解的速 度，还可以提高计算的精度

之前的图形



使用对数之后的图形



将其做对数处理后赋值并对其缺失值补充中位数

类似的特征还有exercised\_stock\_options，expenses，other，shared\_receipt\_with\_poi，total\_payments， salary都一并采用类似的方式处理。

restricted\_stock特征很奇怪有一个很大的负数值这个特征一般不会是负数，有可能是失误，我把它处理为正然后用对数处理赋值。

total\_stock\_value 特征也有一个负数，stock\_value不可能为负数，所以做绝对值处理。然后再用对数做处理。

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”】
2. 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”】
3. 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”】
4. 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”】