员工离职预测训练赛

## 选题背景及题目描述

员工离职是目前很多公司面临的十分头疼的问题之一，但员工的离职从来不是员工或公司哪一个方面的错，而是一个多种原因交错和相互影响的结果，这其中包括工作时长、是否加班薪资、雇员关系等等，结合当下被广泛讨论的“996”问题，我们将目光聚焦在了员工离职问题分析与预测上。

数据来源：DataCastle竞赛

<http://www.pkbigdata.com/common/cmpt/%E5%91%98%E5%B7%A5%E7%A6%BB%E8%81%8C%E9%A2%84%E6%B5%8B%E8%AE%AD%E7%BB%83%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html>

数据描述：所用到的数据取自于IBM Watson Analytics分析平台分享的样例数据。我们只选取了其中的子集，并对数据做了一些预处理。数据主要包括影响员工离职的各种因素（工资、出差、工作环境满意度、工作投入度、是否加班、是否升职、工资提升比例等）以及员工是否已经离职的对应记录。

数据分为训练数据和测试数据，分别保存在pfm\_train.csv和pfm\_test.csv两个文件中。

其中训练数据主要包括1100条记录，31个字段，主要字段说明如下：

（1）Age：员工年龄

（2）Attrition：员工是否已经离职，1表示已经离职，2表示未离职，这是目标预测值；

（3）BusinessTravel：商务差旅频率，Non-Travel表示不出差，Travel\_Rarely表示不经常出差，Travel\_Frequently表示经常出差；

（4）Department：员工所在部门，Sales表示销售部，Research & Development表示研发部，Human Resources表示人力资源部；

（5）DistanceFromHome：公司跟家庭住址的距离，从1到29，1表示最近，29表示最远；

（6）Education：员工的教育程度，从1到5，5表示教育程度最高；

（7）EducationField：员工所学习的专业领域，Life Sciences表示生命科学，Medical表示医疗，Marketing表示市场营销，Technical Degree表示技术学位，Human Resources表示人力资源，Other表示其他；

（8）EmployeeNumber：员工号码；

（9）EnvironmentSatisfaction：员工对于工作环境的满意程度，从1到4，1的满意程度最低，4的满意程度最高；

（10）Gender：员工性别，Male表示男性，Female表示女性；

（11）JobInvolvement：员工工作投入度，从1到4，1为投入度最低，4为投入度最高；

（12）JobLevel：职业级别，从1到5，1为最低级别，5为最高级别；

（13）JobRole：工作角色：Sales Executive是销售主管，Research Scientist是科学研究员，Laboratory Technician实验室技术员，Manufacturing Director是制造总监，Healthcare Representative是医疗代表，Manager是经理，Sales Representative是销售代表，Research Director是研究总监，Human Resources是人力资源；

（14）JobSatisfaction：工作满意度，从1到4，1代表满意程度最低，4代表满意程度最高；

（15）MaritalStatus：员工婚姻状况，Single代表单身，Married代表已婚，Divorced代表离婚；

（16）MonthlyIncome：员工月收入，范围在1009到19999之间；

（17）NumCompaniesWorked：员工曾经工作过的公司数；

（18）Over18：年龄是否超过18岁；

（19）OverTime：是否加班，Yes表示加班，No表示不加班；

（20）PercentSalaryHike：工资提高的百分比；

（21）PerformanceRating：绩效评估；

（22）RelationshipSatisfaction：关系满意度，从1到4，1表示满意度最低，4表示满意度最高；

（23）StandardHours：标准工时；

（24）StockOptionLevel：股票期权水平；

（25）TotalWorkingYears：总工龄；

（26）TrainingTimesLastYear：上一年的培训时长，从0到6，0表示没有培训，6表示培训时间最长；

（27）WorkLifeBalance：工作与生活平衡程度，从1到4，1表示平衡程度最低，4表示平衡程度最高；

（28）YearsAtCompany：在目前公司工作年数；

（29）YearsInCurrentRole：在目前工作职责的工作年数

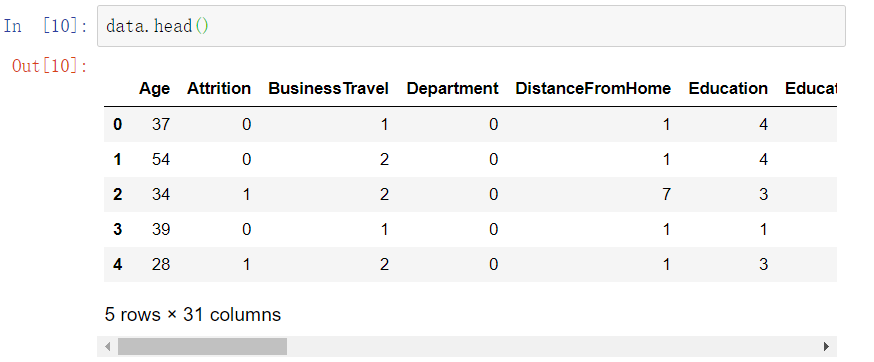
（30）YearsSinceLastPromotion：距离上次升职时长

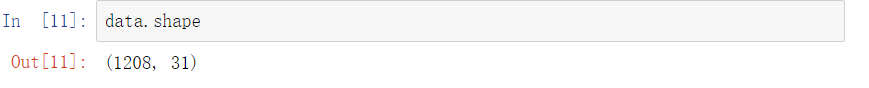
（31）YearsWithCurrManager：跟目前的管理者共事年数；

## 数据初步分析

1. 加载数据

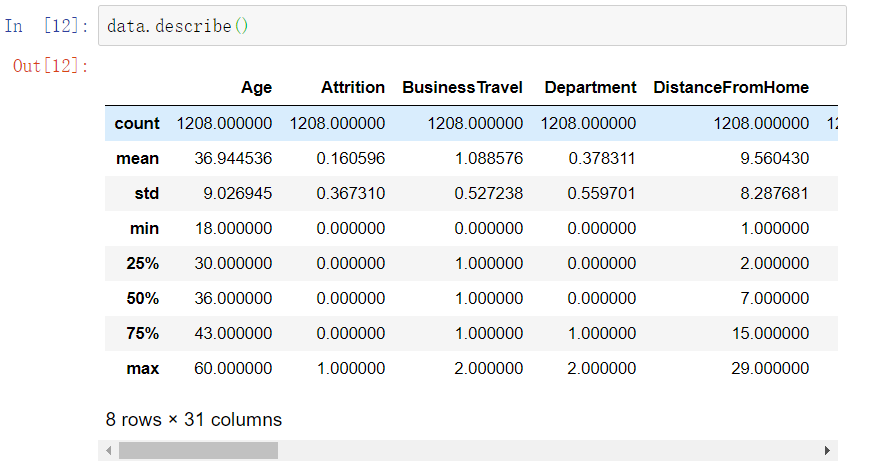






可以看到，测试集数据为1208x31

查看数据的基本统计描述：



员工年龄：18-60岁，平均年龄37岁

离家距离：1-29km，平均9.5km

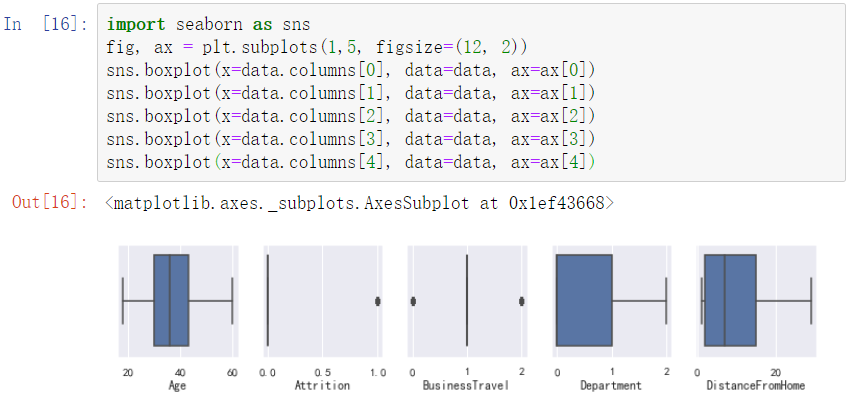
工作环境满意度：1-4，平均2.72

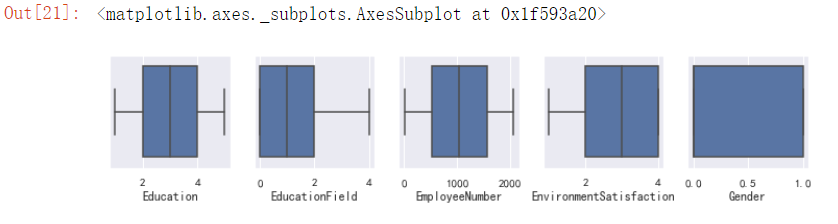
同事关系满意度：1-4，平均2.68

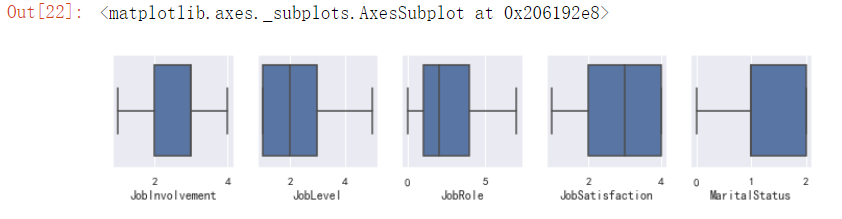
工龄：0-40年，平均11年

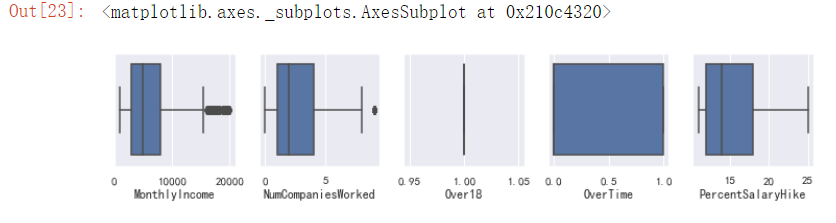
该数据集中，员工的离职率为16%

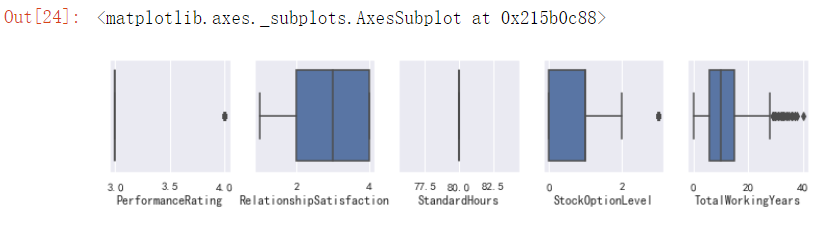
1. 数据预处理
   1. 异常值处理

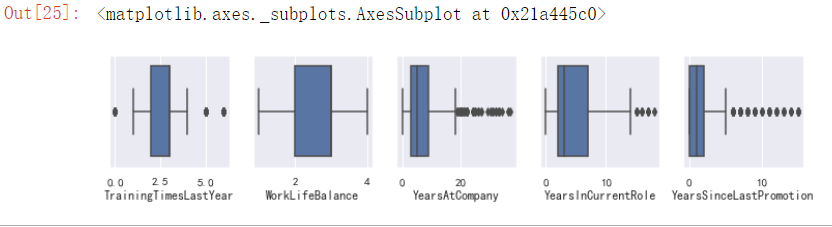








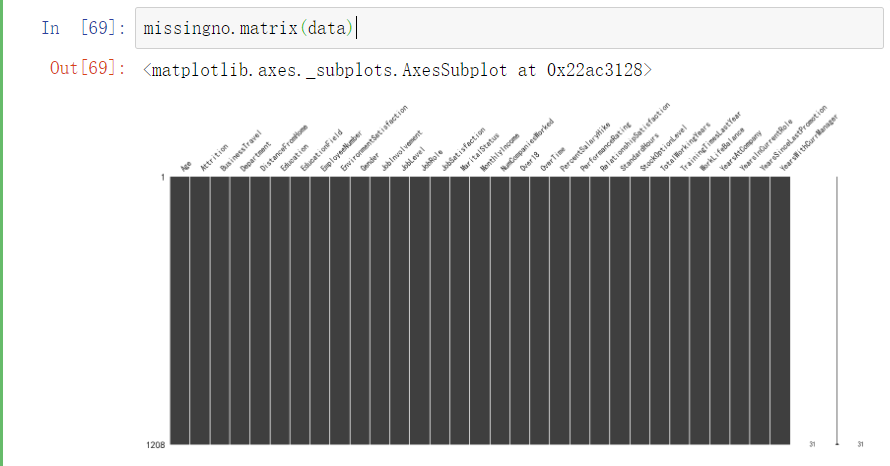




可见，存在异常值的因素主要在工资、工作年限、在公司时间和晋升频率上，一般来讲，此类异常值需要进行处理，但结合实际考虑，对于已经比较成熟的大型企业，存在薪资、工作时间上的“异常值”，反而是十分正常的现象，故此处不再进行异常值的处理。

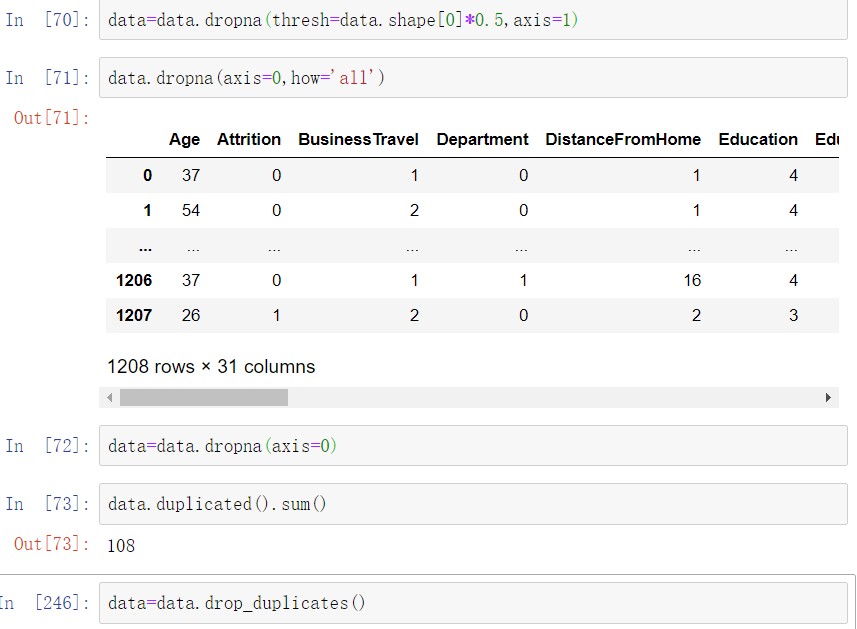
* 1. 空值处理

查看数据中的空值



可见，数据集中不存在空值，故无需进行空值处理

* 1. 重复值处理



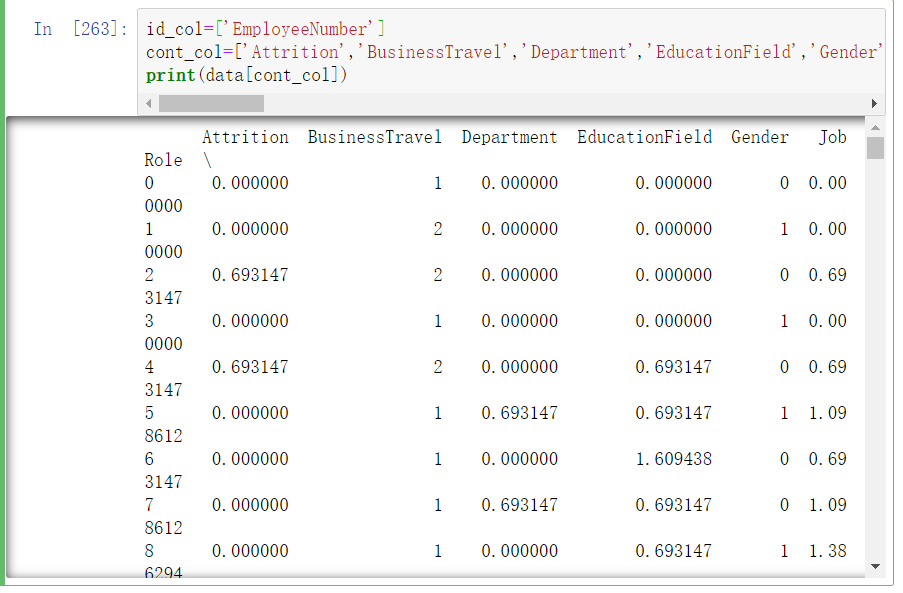
数据集中存在108个重复条，将其舍弃

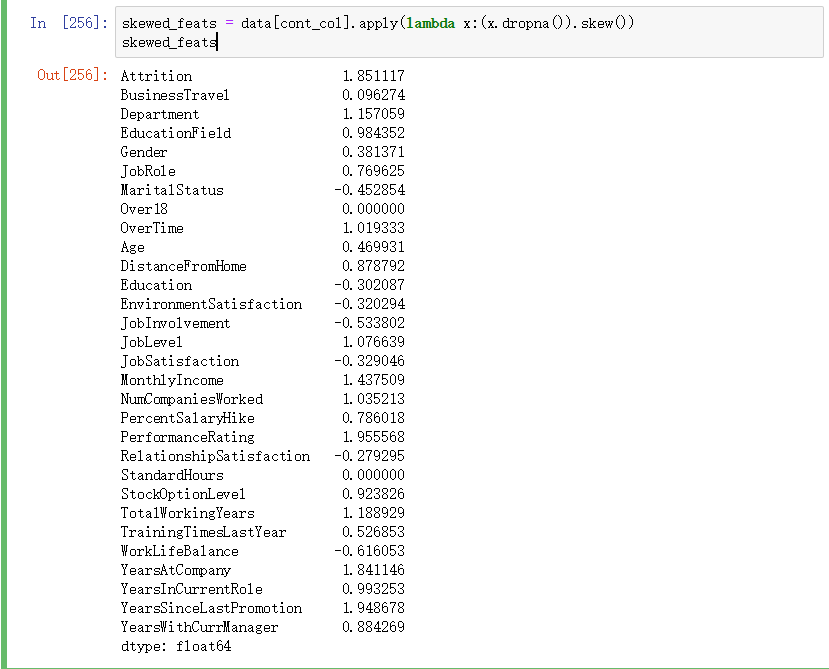
## 特征工程

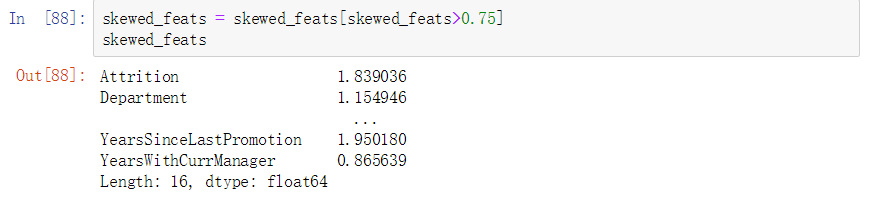
1. 离散值定序与定类

这里，我们在处理数据时，将不同的离散值赋予具体数值进行处理，而没有采用Pandas的get\_dummies方法，原因是该数据集中离散值数量较多，且有部分存在有序性，而无序性部分方差不大，故将其全部转换为数值进行处理

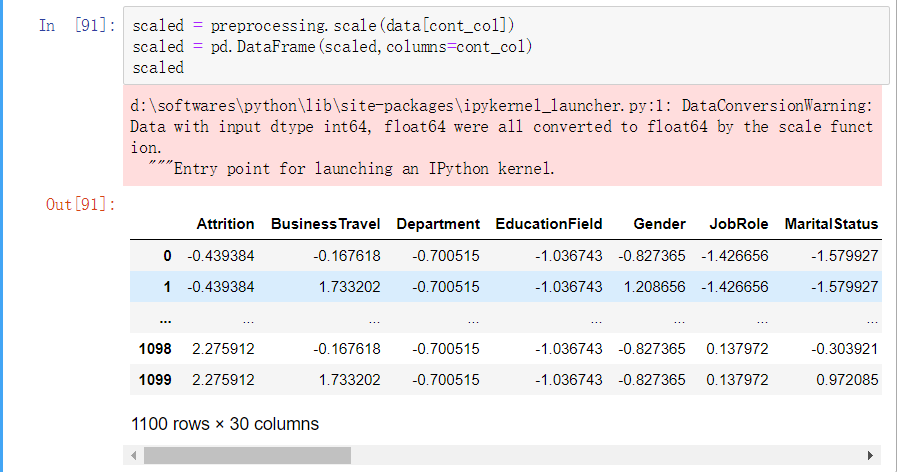
1. 连续型数值处理

 将偏正态分布的数据进行处理

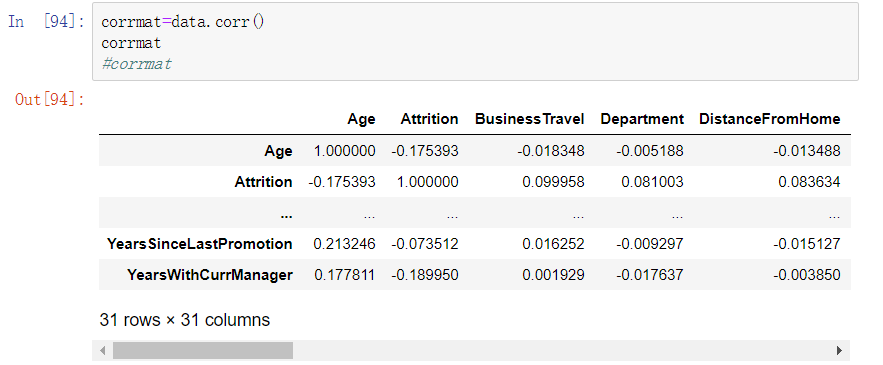
 对于连续型数据，看偏度，一般大于0.75的数值做一个log转化，使之尽量符合正态分布，因为很多模型的假设数据是服从正态分布的



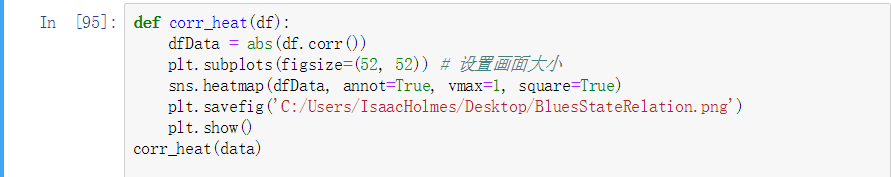
将连续性数据标准化

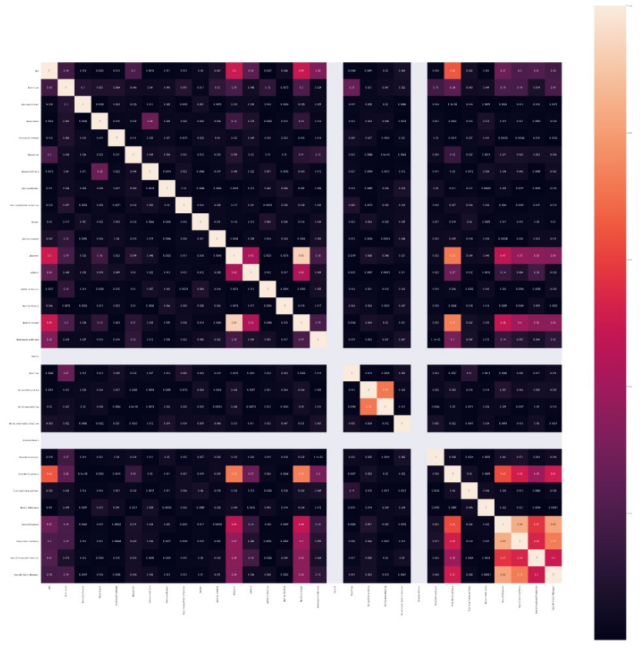


查看变量之间的相关性

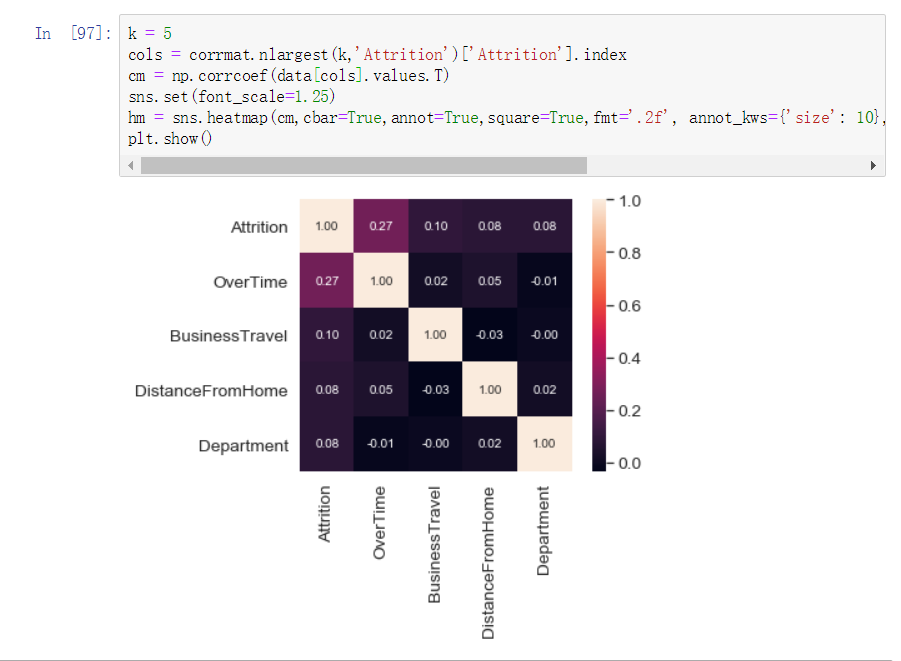


通过热力图进行可视化查看





取与是否离职相关性最大的五个变量：



至此，数据清洗与特征工程的工作已经完成，为了方便数据集及已清洗数据的调用，以下所用的已清洗数据，均是调用sklearn的PCA库函数得到的数据。

## 预测模型

1. 决策树模型

决策树(DTs)是一种用于分类和回归的非参数监督学习方法。目标是创建一个模型，通过从数据特性中推导出简单的决策规则来预测目标变量的值。

决策树的一些优势是:

便于说明和理解，树可以可视化表达；

需要很少的数据准备。其他技术通常需要数据标准化，需要创建虚拟变量，并删除空白值。

使用树的成本是用于对树进行训练的数据点的对数。

能够处理数值和分类数据。其他技术通常是专门分析只有一种变量的数据集。

能够处理多输出问题。

使用白盒模型。如果一个给定的情况在模型中可以观察到，那么这个条件的解释很容易用布尔逻辑来解释。相比之下，在黑盒模型中(例如:在人工神经网络中，结果可能更难解释。

可以使用统计测试验证模型。这样就可以解释模型的可靠性。

即使它的假设在某种程度上违反了生成数据的真实模型，也会表现得很好。

决策树的缺点包括:

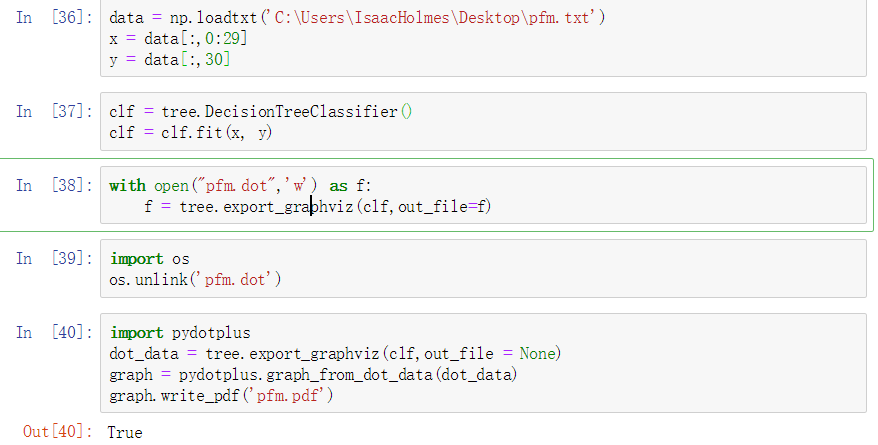
决策树学习可能会生成过于复杂的树，不能代表普遍的规则，这就称为过拟合。修剪(目前不支持的)机制，设置叶片节点所需的最小样本数目或设置树的最大深度是避免此问题的必要条件。

决策树可能不稳定，因为数据中的小变化可能导致生成完全不同的树。这个问题通过在一个集合中使用多个决策树来减轻。

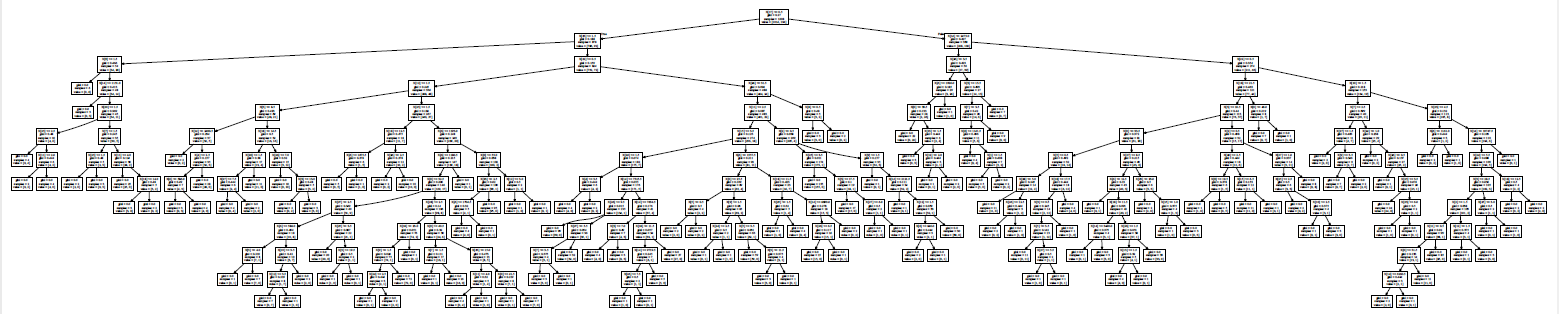
实际的决策树学习算法是基于启发式算法的，例如在每个节点上进行局部最优决策的贪婪算法。这种算法不能保证返回全局最优决策树。通过在集合学习者中训练多个树，可以减少这种情况，在这里，特征和样本是随机抽取的。

有些概念很难学习，因为决策树无法很容易地表达它们，例如XOR、奇偶性或多路复用问题。

在本问题中，我们直接调用sklearn的tree库函数，对已经清洗好的txt格式训练集文件进行分类，结果保存为pdf文件，代码如下：



生成的决策树如下：



1. 朴素贝叶斯分类器（Naïve Bayes）

贝叶斯分类器的思想就是计算属性和分类之间的条件概率，选择使得条件概率最大的分类作为最终的分类结果，是一种基于统计的分类方法。

朴素贝叶斯分类器基于贝叶斯理论：

X代表特征向量，C代表分类，目标即为找到使得后验概率最大的类X

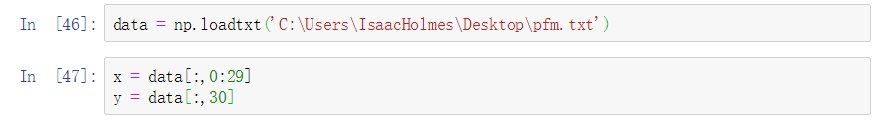
在scikit-learn中，一共有三个朴素贝叶斯分类算法类，分别是GaussianNB，MultinomialNB和BernoulliNB。其中GaussianNB就是先验为高斯分布的朴素贝叶斯，MultinomialNB就是先验为多项式分布的朴素贝叶斯，而BernoulliNB就是先验为伯努利分布的朴素贝叶斯

* 1. 高斯朴素贝叶斯

GaussianNB 实现了运用于分类的高斯朴素贝叶斯算法。特征的可能性(即概率)假设为高斯分布:

参数和使用最大似然法估计。

代码如下：

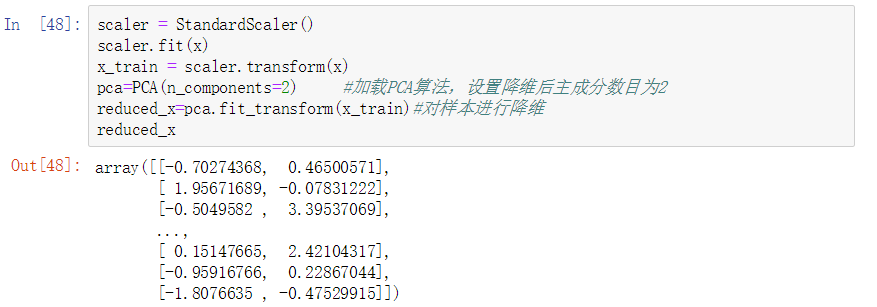


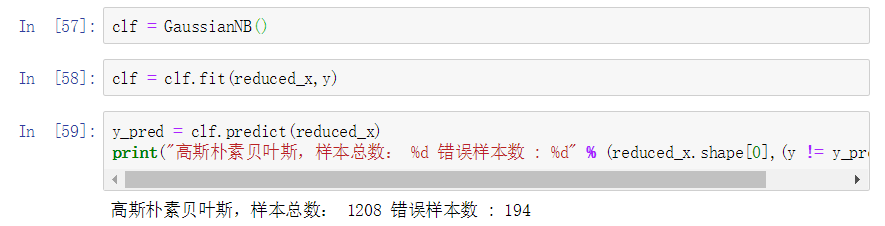
加载数据集，并划分特征向量和类

调用GaussianNB进行高斯朴素贝叶斯分类：



可以发现，错误样本数有254个，尝试将特征向量X降维至二维：



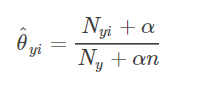


可见，错误样本数降低为194个，分类准确度提高

* 1. 多项式分布朴素贝叶斯

MultinomialNB 实现了服从多项分布数据的朴素贝叶斯算法，也是用于文本分类(这个领域中数据往往以词向量表示，尽管在实践中 tf-idf 向量在预测时表现良好)的两大经典朴素贝叶斯算法之一。 分布参数由每类 y 的向量决定，式中n时特征的数量，时样本中属于类y中特征i概率

参数使用平滑过的最大似然估计法来估计，及相对频率计数

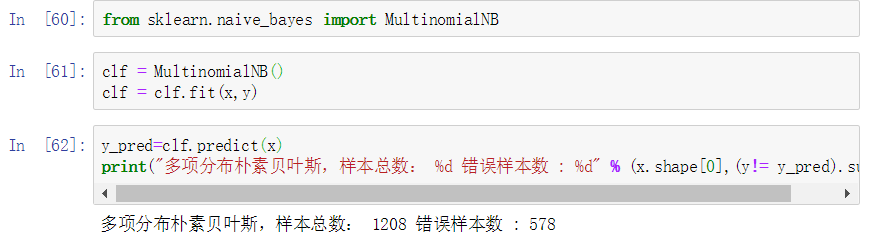


式中是训练集T中特征i在类y中出现的次数

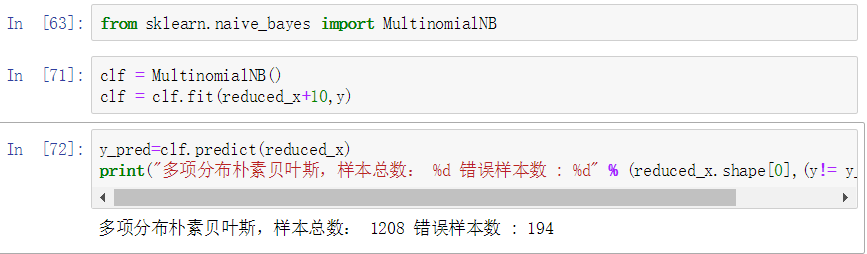
是类y中出现的所有特征的计数总和

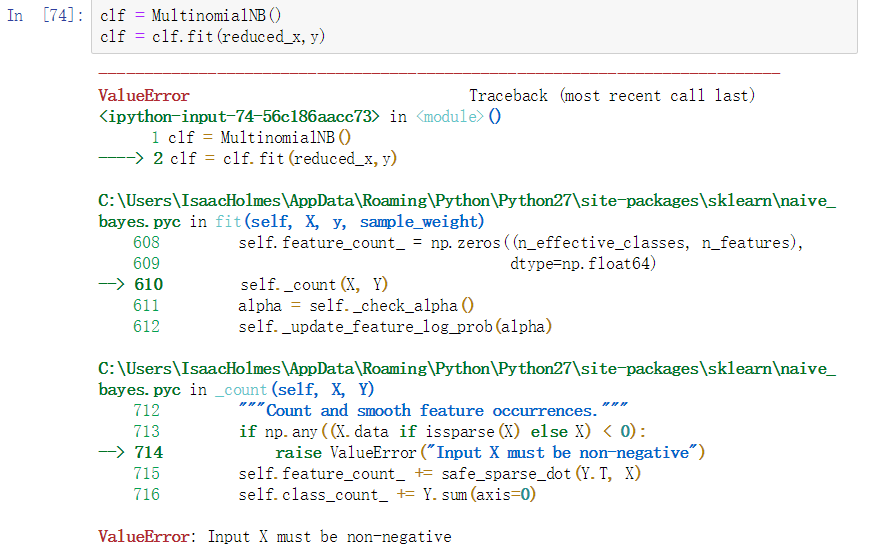
先验平滑因子α≥0应用于学习样本中没有出现的特征，以防在将来的计算中出现0概率输出。 把 α=1 被称为拉普拉斯平滑(Lapalce smoothing)，而α<1被称为利德斯通平滑(Lidstone smoothing)。

代码如下：



采用原始的特征向量进行多项分布朴素贝叶斯分类结果很不理想，错误样本数高达578个，同样，采取降维：



注意到此处采用了并非原始降维后特征向量reduced\_x，而是在此基础上+10，原因是杜象是分布朴素贝叶斯的输入X一定是非负的，而降维后特征向量含有负值，将整体+10，不影响分类结果。  


可见，多项式分布朴素贝叶斯算法在降维和对降维后的特征向量进行简单处理后，分类正确率也达到了较高水平

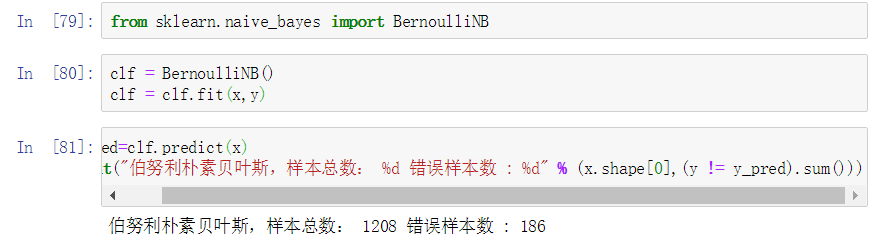
* 1. 伯努利朴素贝叶斯

BernoulliNB 实现了用于多重伯努利分布数据的朴素贝叶斯训练和分类算法，即有多个特征，但每个特征 都假设是一个二元 (Bernoulli, boolean) 变量。 因此，这类算法要求样本以二元值特征向量表示；如果样本含有其他类型的数据， 一个 BernoulliNB 实例会将其二值化(取决于 binarize 参数)。

伯努利朴素贝叶斯的决策规则基于

与多项分布朴素贝叶斯的规则不同 伯努利朴素贝叶斯明确地惩罚类 y 中没有出现作为预测因子的特征 i ，而多项分布分布朴素贝叶斯只是简单地忽略没出现的特征。

代码如下：



可以看到，在没有对原始特征向量降维的情况下，伯努利朴素贝叶斯分类方法的正确率已经很高，但进行降维后的分类准确率不升反降：



从这里可以直观看出，PCA的优点与不足之处，即在大部分情况下，进行主元分析，对特征值进行降维可以起到优化算法、提高算法与运行速度的效果，但在一些情况下，主元分析所没有涵盖到的特征向量，将会对分类或预测精度造成影响。

* 1. 朴素贝叶斯分类器小结

本实例中，我们仅对朴素贝叶斯算法的原理进行了解，对三种先验条件下的贝叶斯分类器的原理做了一点了解，主要调用三种分类器对我们的数据集进行分类，对其优劣和适用领域还只限于有一些感觉而并不清楚原理的境地，通过三种分类器的适用，了解了对原始数据进行主元分析的好处与不足之处，也是意外之喜。

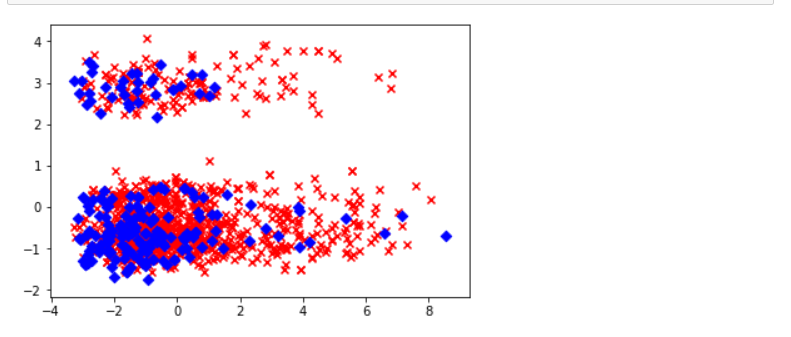
1. 支持向量机

支持向量机是一种监督机器学习算法，它在训练样本空间中构造超平面来对样本进行分类，优点是对高维度不敏感，但效率比较低，实施起来较为复杂，SVM是将低维无需杂乱的数据通过核函数那户（RBF，poly，linear，sigmoid）映射到高维空间，通过超平面将其分开。

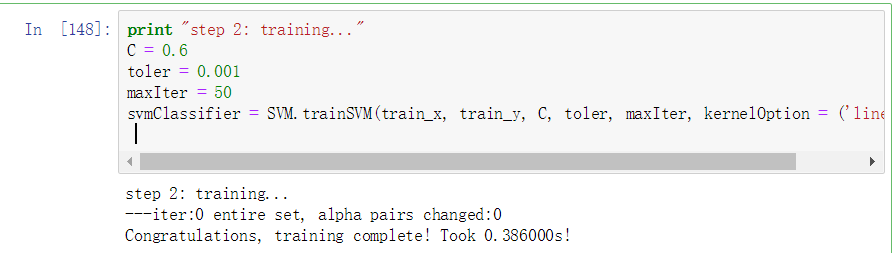
代码部分如下：

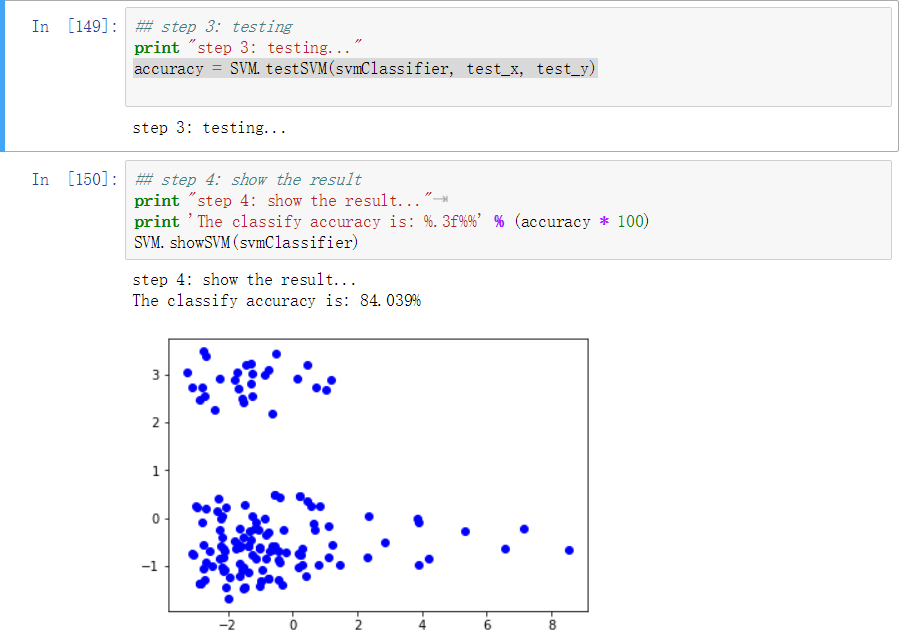
数据可视化：





此案例中，我们直接调用了SVM库，并按照相应格式对源数据集进行训练集与测试集的划分，在此，我们没有使用比赛中的测试集，是为了在训练成果显示时更加方便。





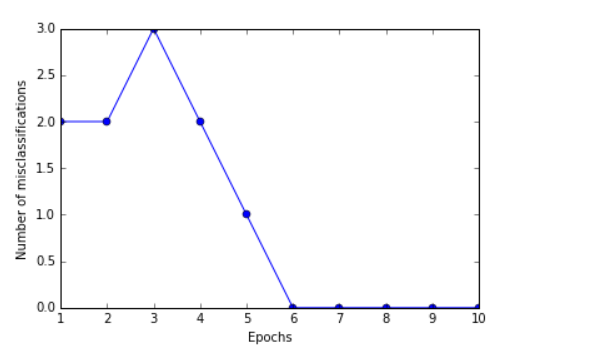
可以看到，使用SVM分类器，分类精度可达到84%

1. 基于神经网络的分类器

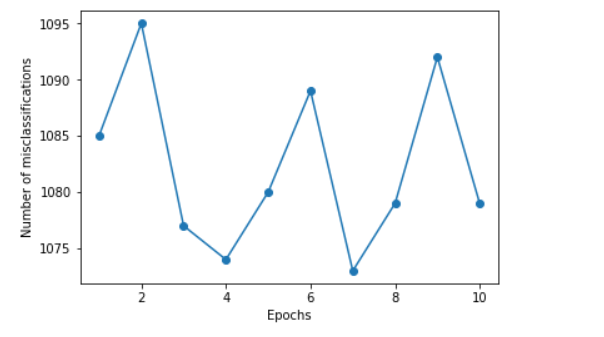
此问题是一个二分类问题，将0与1作为类别标签，激活函数可以表示为：

这里我们没有对神经网络模型的实现方法进行深入探究，而由于神经网络模型的特殊性，很多神经网络模型存在过拟合问题，在其他分类问题上表现并不如意。

例如我们用到了一个花萼类型判断的神经网络模型，在它原有的数据集上，在经过六代学习后，分类准确度可达到100%：



在代入我们的数据集后，分类精度机器糟糕，在世代学习后，仍然绝大部分都分类错误



小组时间精力不足，便未在此问题上深入探究。

## 总结

在此次数据挖掘实战中，我们选取了员工离职预测这一问题进行研究，利用pandas库的强大功能，进行了数据清洗，使其更好地满足模型训练要求，在模型训练过程中，对决策树模型，朴素贝叶斯分类器，基于支持向量机的分类器和基于神经网络的分类器进行了不同深度的学习了解与应用，并在此过程中，对分类问题的本质、主元分析的优点与不足之处、各种分类器模型的原理等多方面知识有了更加深入的理解，在应用之于实战中时，感受到了十足的充实感和愉悦体验。

参考链接：

<https://blog.csdn.net/eric_doug/article/details/51769644>

<https://blog.csdn.net/luanpeng825485697/article/details/78967139>

<https://blog.csdn.net/m0_37269455/article/details/74201689>