如何利用大数据分析技术预测员工离职？

作者：刘丽媛

**1.离职预测的意义**

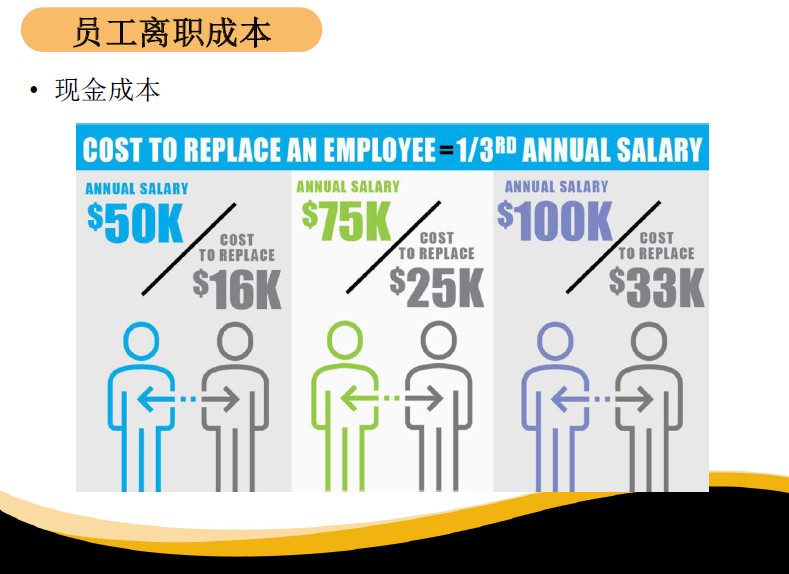
首先来看一组数据：2017我国平均的员工离职率在22%左右，对于一些一线的城市北上广可以达到23%。对于高科技行业，员工离职率更高，达到25%以上。所以，员工离职每个公司面临的问题。

再看国外数据：这是一家致力于员工评估的美国公司。他们的数据显示，1/4的员工是属于高风险人群，而这当中20%的都是高绩效员工。所以，如果我们准确预测员工的离职，就可以更好地帮助公司留住高绩效员工。

第二组数据显示：44%的员工表示他们如果能够在竞争公司中得到20%或者是更低的加薪，他们就会考虑跳槽。

第三组数据显示：超过70%的离职员工表示，如果他们为了寻求更好的职业发展，也会选择自愿离职。

我们来看看员工的离职到底对公司产生怎样的损失：**首先是公司现金的损失。**根据美国劳工部数据显示，如果一个员工离职，再重新招聘和培训一个员工取代他的位置，大概要消耗到相当于这名员工1/3年薪的现金。此外，还有其他的损失，比如说替换/选拔成本、时间、人力的损失等。



**其次是效率成本。**比如说一个中高级员工选择了离职，他手里的项目很有可能一段时间内效率放缓和停止。让新的员工去取代他的位置，也会造成一段时间项目效率的下降。

**还有文化流失成本。**如果我们公司当中有越来越多的员工离职，就会让现有员工心理上产生焦虑，对公司的企业文化来说也是一种流失。此外，还包括知识产权的损失，以及离职员工本身所带客户资源的损失。

**所以，这就是为什么我们要对员工离职进行预测，因为它会帮助公司减少员工离职成本的消耗。**

**2.数据收集**

HR可以做什么呢？我们可以根据现有的HR数据，对公司员工在职情况进行描述性分析。然后找出为什么员工会离职？哪些员工是离职的高风险群体。

网上曾经有文章介绍四大常见的离职原因：第一，员工觉得工作负担过重，生活和工作不能很好地平衡；第二，职业规划。他们希望寻求更通畅的竞争渠道，会跳槽寻找更好的机会；第三，为了得到更高的薪水而离职；第四，也许会有同事工作上关系的处理不当，导致员工选择离职。

**所以，HR数据分析要做的就是希望可以量化离职原因，找出哪个原因最重要，哪个原因没那么重要。**

当我们收集HR数据时，**首先最重要的第一点就是收集的数据一定是要以每一位员工为单位。**建立数据库的时候也要注意到一点，对于每一位员工都要有唯一的量化指标。比如说员工号码可以非常快地寻找到员工。然后再根据员工号码去收集员工的基本信息，比如说性别、离在职情况、工作状况、教育程度等。

其他待收集的员工信息包括薪资、社保、福利、绩效、升迁、培训发展、满意度等等。我们收集的信息越多、越准确，那么之后无论是进行描述型分析，还是预测型分析，都会为我们数据清洗的过程可以节省很多时间。

**所以，我建议在数据收集阶段尽可能多地收集以下六个方面的数据。**

****

**3.数据清洗**

一个数据分析项目的流程是这样的：**首先是明确目标，其次是数据采集和清洗的工作。**在大部分数据工作时间中，数据采集和清洗占到70%的工作量。因为现实采集中，数据是非常大和混乱的。当我们把数据整合在一起需要花费很长的时间，这也是为什么之前提到，如果我们能够收集数据的时候就着眼于后面的分析目的，那后面采集和清洗的时间将大量减少。

**第三是描述性分析，主要是根据现有数据然后进行描述性分析**。例如描述离职员工和在职员工的平均薪资是怎样的，这些可以通过对比实现。建模这一环节我们称之为“预测型分析”**。**

**最后，我们还需要将所有结果以可视化的图表呈现。**无论是呈现给上级，还是分享给同事，数据的可视化都是非常重要的工具和技能。



对于员工的离职预测，我们的目标是什么呢？

**第一是我们想描述一下各类HR数据与员工离职之间的关系**，就是我之前提到的以六个板块为基础来分析每个板块和离职之间的关系。同时，6个板块互相之间也是有关系的，例如员工前一年的工资提升和员工后一年的绩效评估也可以拿出来一起进行比较，然后看看前一年工资增长是否可以导致后一年绩效提升。

**第二是预测员工的离职风险。**我们可以量化每一位员工离职的概率有多大。第三是明确员工离职的原因，这点也是大部分HR想知道的。

这里要多介绍一下数据的采集和清洗，这些工作是为了数据的描述和建模做准备的。可以每一位员工为单位进行数据的采集。数据清洗主要分为：数据的审查、描述逻辑、逻辑核查和人工核查。有些企业数据比较乱，我们需要人工删除不合格和重复的变量。

**数据清洗包括清洗错误值、异常值、缺失值和重新赋值。**比如员工年龄，收集时就有错误产生，有些员工年龄12岁，有些99岁，那我们可以在这些错误录入的数据全部找出来，然后以员工平均年龄作为替代。

让人头疼的一件事是缺失值。很多员工的数据是缺失的，比如员工的培训数据。最常见的方法是以平均值和中位值去代替缺失值、错误值和异常值。

此外还有数据转换，也就是将数据生成可供计算机分析的语言，比如说员工的教育程度，可能是高中、大专、本科、研究生这样的文字形式。这种文字在计算机建模时是不容易识别的。此时我们就要进行数据转化，比如将高中转化为“1”，大专转化为“2”，本科转化为“3”。



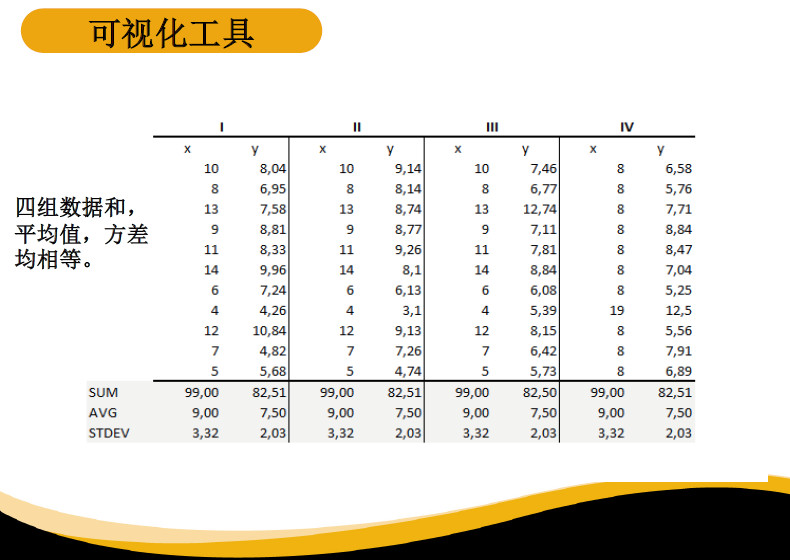
**最后是数据验证**，需要分析我们的数据是否可以建立正确的模型，以上是数据清洗和采集的基本流程。

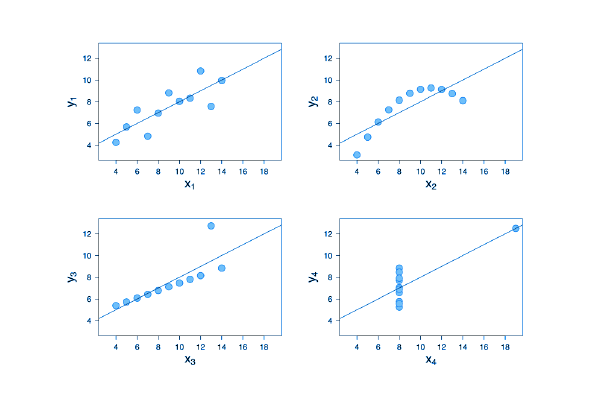
**4.描述性分析**

描述型分析可以通过数据可视化的方法来描述出数据之间的一些关联。

**首先，我们可以通过描述型分析得到数据的频数分析。**比如说每一年公司到底有多少百分比员工离职、离职员工和在职员工的绩效是否一样、他们工资是否一样，等等。建议通过可视化数据工具来帮助描述分析。

我们通过一个案例来看看为什么数据的可视化非常重要。这是四组不同数据，每一组数据都有X和Y，各组数据的平均数和方差都是相等的。大家可能会觉得这四组数据反映的内容是相同的。





但是，如果我们以数据可视化去看这四组数据，就可以看到四组数据的差异，所表达的信息也完全不同的。所以，如果大家做描述型分析，或者是想向上级呈现数据，做数据的可视化呈现更能反映出数据的特点。

**5.数据拆分和平衡类别**

**预测型分析就是建模的分析，我们希望建立机器学习的模型去预测未来可能发生的事情。**

我们采集的数据分为两部分：**第一部分是目标变量。**在员工预测离职当中，目标变量就是员工离职或在职状态。通过数据转换的方式，将员工的离职转为“1”，在职转为“0”。其他自变量就是之前提到的6个方面，或者是大家遵循企业自身的特点扩展自变量，比如绩效、工龄、工作满意度、晋升、培训等数据。

**建模的目标就是希望找到一个定量的系数，在每一个自变量和目标变量之间建立联系。**到底绩效越高员工越容易离职，还是绩效越低员工越容易离职；到底是工龄越长员工更容易离职，还是工龄越短员工更易离职。

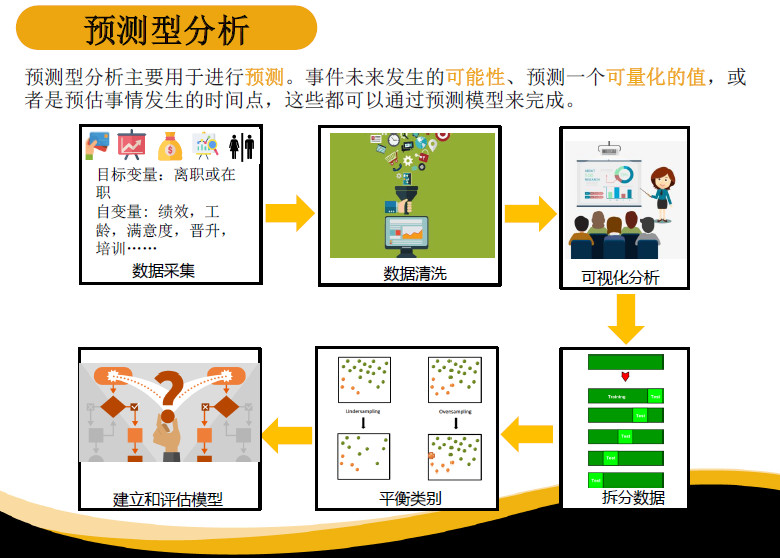
当我们的模型建立以后，每个自变量前面都有一个系数，这个系数就可以反映它和员工的离职是正相关还是负相关的关系。

在完成了数据收集、清洗和可视化分析之后，就到了数据拆分。

为什么要拆分数据？因为建模是通过采集的数据进行的。当模型建好之后，我们并不能100%确定模型是准确的。这时候我们就会把之前所收集到的**数据进行拆分**。比如，我们将数据的70%作为训练模型的数据，再去把系数代到剩下的30%数据当中，得到的结果和原有离职结果进行比较，这样可以求出一个偏差。**偏差就是模型的准确率**，这样可以初步判断模型是好还是不好。

**下一个步骤是平衡类别**，这是对于特殊的数据才会应用到的技术。这里是指员工离职和在职数量不一致的时候，我们会做一个平衡类别。比如公司有一千名员工数据，其中有900名在职员工，100名离职员工。如果我们瞎猜，猜所有的员工都是在职的，那我就猜对了900个，模型的准确率是90%。然而这是没有任何意义的，因为它没有准确地去分析出哪些员工离职。

所以，**平衡类别是通过一些机器学习的手段，根据现有离职员工的数据，去虚拟地产生一些离职员工数据。**虽然这些数据是虚拟的，但它可以反映出离职员工的特点。可以将我们的数据变成离职员工和在职员工1：1或者3：1的比例。通过这样的比例再去做预测模型，就会比之前准确度要高出很多。

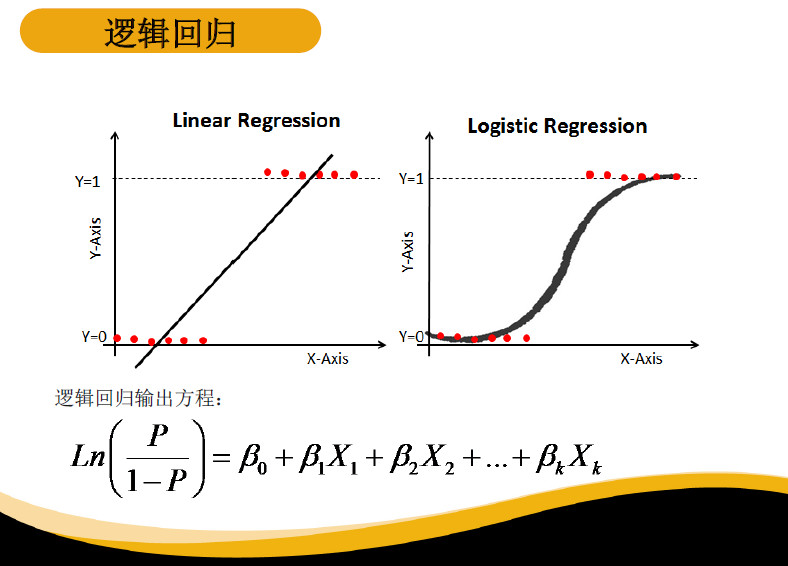


**6.建模**

建模是最简单的一步，也是耗时最少的一步。

对于现在的机器学习而言，很多的预测性模型，例如决策指数、逻辑回归、深度学习等都可以预测未来所可能发生的事情。今天我们主要介绍逻辑回归，也就是我们的目标变量只有0和1只有两个变量。

比如，今天下雨和不下雨、银行贷款者违约还是不违约，都可以用0和1的表现形式，而逻辑回归也是用于0和1问题中一个常见的模型。



这就是逻辑回归的输出方程，P代表员工离职的概率，X1、X2、X3代表输入的自变量，我们需要找到X的系数，这样可以求出员工离职的概率。

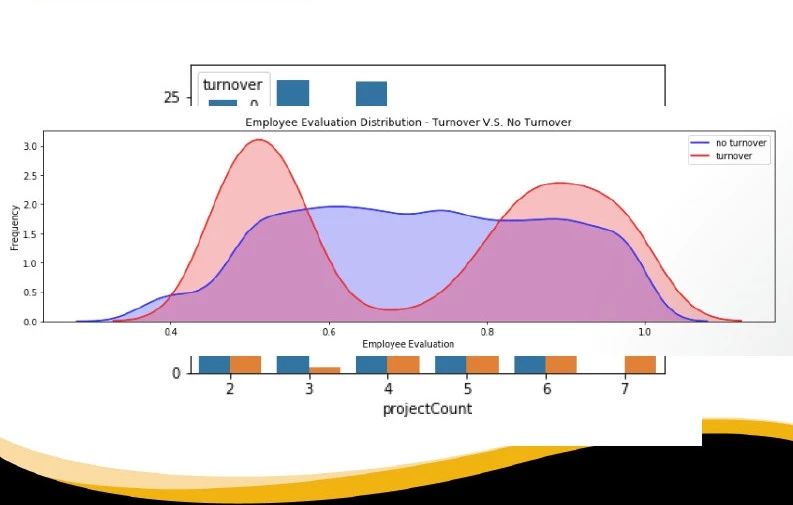
**混淆矩阵是用来检验模型质量的，分为真实值和预测值两部分。**真实值就是训练数据集的目标值，关于一个员工事实是在职还是离职的；预测值是我们输入系数以后，结果预测这位员工是离职还是在职。

如果一名员工真离职了，预测也是离职，我们就将它最左上方的小方格叫TP；如果员工离职了但我们却预测在职，就是FN；如果员工在职，我们预测也是在职，就是TN。用TP和TN之和去除以所有的变量，就可以得出模型的准确率。

**7.案例分析**

一家公司一共有14999名员工，员工的离职以“1”代表，“0”代表员工未离职。自变量包括员工对公司的满意程度、绩效评估结果、所参加过的项目数目、平均每个月的工作时长、工作是否发生过差错、5年内晋升情况、所属部门以及薪资水平等。数据已经通过清洗转化成了可识别的数字语言。





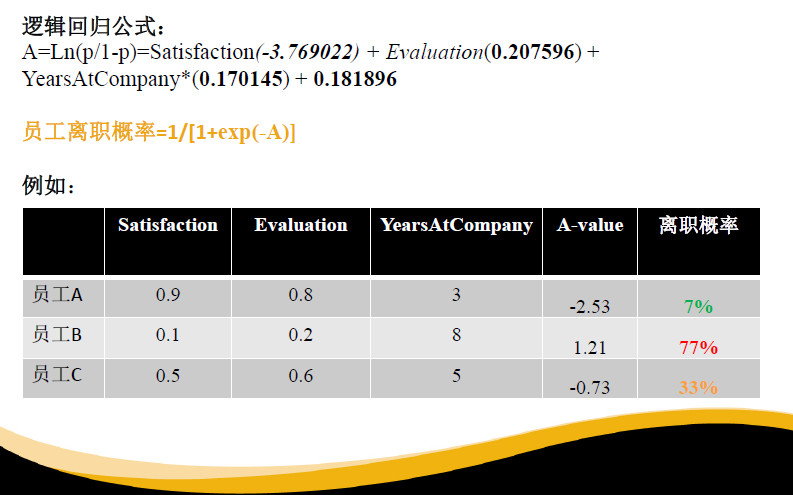
用这张描述性分析的图来反映员工的评估和离在职的情况：蓝色区域代表的是员工在职，红色区域代表的是员工离职。可以看到红色区域有两个峰值，说明员工评估过低或过高时都会导致员工离职。

我们选取了三个最重要的变量，包括员工满意度、绩效评估结果和在公司的在职年数，做一个逻辑回归的模型。根据模型可以看到员工对工作的满意度与离职是呈负相关，系数是负3.77。也就是说员工满意度越高，越不容易离职。

第二个变量是员工的绩效评估，系数是0.20，可以看到它和离职是正相关的关系。也就是说员工的评估结果越高，越容易离职。

第三个变量是员工在公司的司龄，也是呈正相关的关系。员工在公司的时间越长越容易离职。

当然以上数据所反映出来的情况不一定对每家公司适用，所以我们要通过每个公司的特定数据去找出特定原因。



假如我们给出了ABC三位员工的案例，他们可以作为未来我们想要分析的新数据。因为已经有这样的模型了，**所以对于新的员工数据，我们可以很快用上面的三个系数去计算出员工的离职概率。**以上三个员工通过模型求出的A的离职概率是7%，B是77%，C是33%。

所以，我们可以预测到员工B是高风险离职人群。假如还有更多数据，我们可以测算出企业当中到底哪部分员工是高风险离职人群，公司可以对这部分员工采取一些员工关怀、升职加薪等政策，这样更好地留住他们。

**8.总结**

大家在做人力资源数据分析时，**一定要做到以数据讲故事**。很多时候如果把数字呈现在别人面前，有可能并不太直观。用数据的可视化去讲数据内部的联系时，效果会比直接呈现数字好很多。

**其次是要寻找Actionable Insights。**比如，假如我们分析得出男员工比女员工更容易离职，这个发现对公司而言，不太容易采取什么措施来改变。不能说公司在招聘的时候多招聘女员工，少招聘男员工。更重要的是我们要找到一些公司可以对此采取措施的发现。

比如，我们发现前一年的工资增长不能很好体现下一年绩效的增长，那么我们可能就要调整一下，看是否我们前一年的工资增长存在问题。

还有就是要收集有效数据，**一定要尽可能地多收集干净数据**。我曾经遇到过这样的情况，有公司它可能有一万多名员工，但是由于数据收集的缺失，当我们整理多方数据时，发现最后表里只有两千多名员工。相当于八千多名员工的数据是丢失的。

所以在数据收集的过程中，如果我们可以整齐地收集每一位员工的数据，那么对于之后数据分析是有帮助的。

**最后分享一下做数据分析的软件，Weka**是基于Java语言的软件，**Orange**是根据Python语言的软件，这两个软件非常好而且免费。使用它们不需要任何编程，只需通过简单的拖拽就可以实现分析。

其他的工具包括**Excel**也是非常广泛应用于数据分析，还有需要编程基础的Python，大家有兴趣都可以去了解。