项目案例：保险行业数据分析

一、业务背景

1. 业务环境

- **宏观**

  中国是世界第二大保险市场，但在保险密度上与世界平均水平仍有明显差距。

- **业界**

  保险行业2018年保费规模为38万亿，同比增长不足4%，过去“短平快“的发展模式已经不能适应新时代的行业发展需求，行业及用户长期存在难以解决的痛点，限制了行业发展。

- **社会**

  互联网经济的发展，为保险行业带来了增量市场，同时随着网民规模的扩大，用户的行为习惯已发生转变，这些都需要互联网的方式进行触达。

  保险科技：当前沿科技不断应用于保险行业，互联网保险的概念将会与保险科技概念高度融合。

中国保险市场持续高速增长。根据保监会数据，2011~2018年，全国保费收入从1.4万亿增长至3.8万亿，年复合增长率17.2%。2014年，中国保费收入突破2万亿，成为全球仅次于美国、日本的第三大新兴保险市场市场；2016年，中国整体保费收入突破3万亿，超过日本，成为全球第二大保险市场；2019年，中国保费收入有望突破4万亿。

2. 发展现状

- **概览**

  受保险行业结构转型时期影响，互联网保险整体发展受阻，2018年行业保费收入为1889亿元， 较去年基本持平，不同险种发展呈现分化格局，其中健康险增长迅猛，2018年同比增长108%，主要由短期医疗险驱动。

- **格局**

  供给端专业互联网保险公司增长迅速，但过高的固定成本及渠道费用使得其盈利问题凸显，加上发展现状强，**自营渠道建设**及**科技输出**是未来的破局方法，**渠道端形成第三方平台为主，官网为辅的格局**，第三方平台逐渐发展出B2C、B2A、B2B2C等多种创新业务模式。

- **模式**

  互联网保险不仅仅局限于渠道创新，其核心优势同样体现在**产品设计**的创新和**服务体验**的提升。

3. 发展趋势

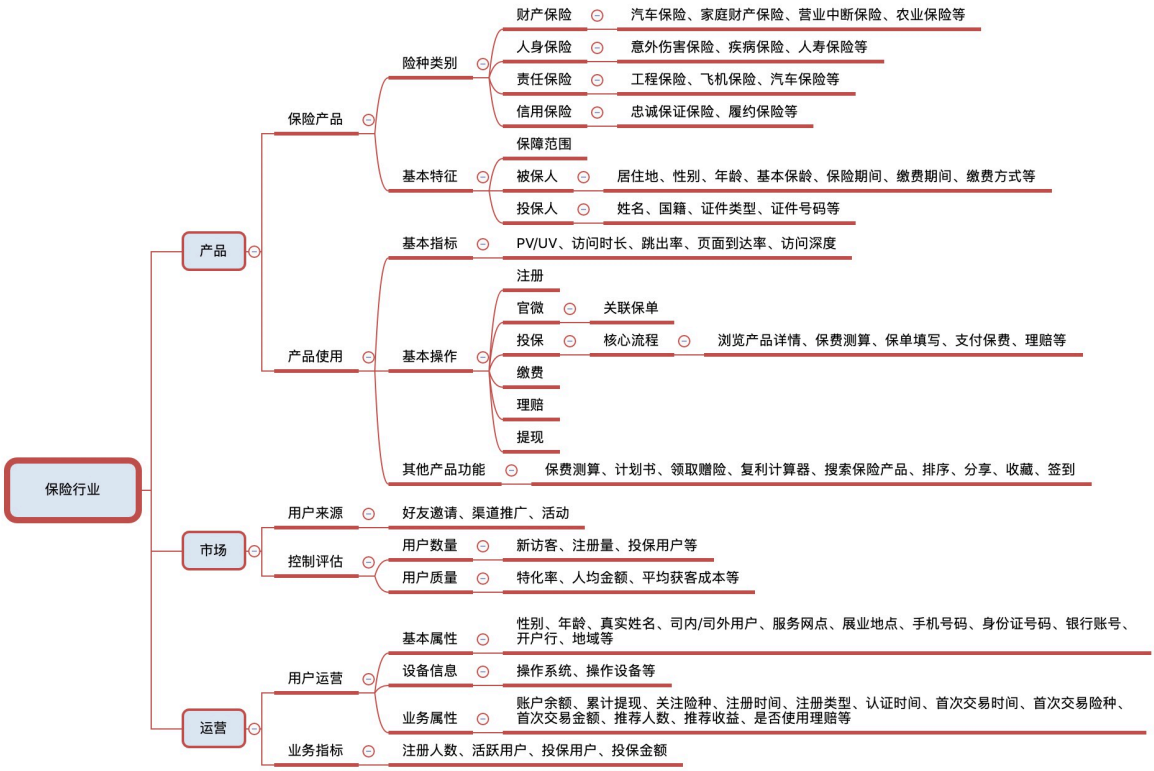
- **竞合格局**

  随着入局企业增多，流量争夺更加激烈，最终保险公司与第三方平台深度合作将成为常态。

- **保险科技**

  当前沿科技不断应用于保险行业，互联网保险的概念将会与保险科技概念高度融合。

4. 衡量指标



5. 业务目标

针对保险公司的健康险产品的用户，制作用户画像，然后进行精准保险营销。

二、案例数据

1. 数据来源

美国某保险公司，和本公司合作多年。现在该公司有一款新的医疗险产品准备上市。

2. 产品介绍

这一款新的医疗产品主要是针对65岁以上的人群推出的医疗附加险，销售渠道是直邮。

3. 商业目的

为保险公司某种健康险产品做用户画像，找出最具有购买倾向的人群以进行保险营销。

4. 数据介绍

本次案例数据共有76个字段，字段繁多，在处理数据时，需要先将数据按照类别进行归类，方便理解查看。

4.1 基本信息

变量名称	变量含义	备注
KBM_INDV_ID	用户ID	无意义特征
Resp_flag	用户是否购买保险	响应变量，也就是我们的目标变量
age	用户年龄	本产品为针对65岁以上人群保险
GEND	性别	

4.2 基本情况

变量名称	变量含义	备注
c210mys	学历	0-unknown; 1-初中；2-高中不到；3-高中毕业；4-大学未毕业；5-大专；6-本科；7-研究生；8-专业院校毕业；9-博士
POC19	是否有小孩	
CA00	小孩是否在0-2岁之间	
CA[XX]	小孩年龄	小孩年龄一共5个特征，通过编号进行区分

4.3 家庭成员

变量名称	变量含义	备注
NOC19	家庭小孩个数	
NAH19	家庭成年人个数	
NPH19	家庭成员人数量	
POEP	家庭是否有老人	

4.4 家庭成员情况

变量名称	变量含义	备注
U18	是否有家庭成员小于18岁	
N1819	是否有家庭成员在18-19岁之间	
N2029	是否有家庭成员在20-29岁之间	
N3039	是否有家庭成员在30-39岁之间	
N4049	是否有家庭成员在40-49岁之间	
N5059	是否有家庭成员在50-59岁之间	
N6064	是否有家庭成员在60-64岁之间	
N65P	是否有家庭成员在65岁以上	

4.5 疾病史

变量名称	变量含义	备注
AART	是否有关节炎	
ADBT	是否有糖尿病	
ADEP	是否有抑郁症	
AHBP	是否有高血压	
AHCH	胆固醇含量是否过高	
ARES	是否有呼吸疾病	
.....	.....	一共12个特征

4.6 金融信息

变量名称	变量含义	备注
BANK	是否有过破产记录	
FINI	是否用过保险服务	
INLI	是否投资过寿险	
INMEDI	是否购买过医疗险	
INVE	是否有投资	

4.7 个人习惯

变量名称	变量含义	备注
IOLP	是否网上购买过产品	
MOBPLUS	是否通过快递买过东西	M-通过多种快递渠道购买；P-或许通过多种快递读到购买；S-单一快递渠道购买；U-不知道
ONLA	是否上网	
SGFA	是否喜欢美术	
SGLL	是否经常有奢侈消费	
SGOE	是否经常户外活动	
SGSE	是否喜欢运动	
……	……	

4.8 家庭状况

变量名称	变量含义	备注
LIVEWELL	幸福指数	值越大，说明越幸福
HOMSTAT	是否有房子	Y:有房子；P:可能有房子；R: 租房； T,U:不确定
HINSUB	是否有医保补贴	A-C, 补贴依次增加
c210cip	收入所处排名	值越大，说明收入越高
c210ebi	普查家庭有效购买收入	值越大，说明有效购买收入越高
c210hmi	家庭收入	值越大，说明家庭收入越高
c210hva	家庭房屋价值	值越大，说明房屋价值越高
C210.....	家庭经济类数据	值越大，说明经济地位越高

4.9 居住城市

变量名称	变量含义	备注
STATE_NAME	所处的省份	
c210apvt	贫穷以上人的比例	值越大，说明比例越高
c210b200	所处地区有多少居住小区在2000年及以后建立	值越大，说明比例越高
c210blu	所处地区蓝领所占百分比	值越大，说明比例越高
c210bpvt	贫穷以下人的比例	值越大，说明比例越高
c210mob	所处地区mobile home的比例	值越大，说明比例越高
c210pdv	离婚或者分居人群所占比例	值越大，说明比例越高
.....	居住地统计数据	

5. 分析思路

- 根据经验，我们可以大概判别哪些特征很可能和用户是否购买保险会有相关关系。
- 结合我们的业务经验，以及数据可视化，特征工程方法，先行探索这些特征中哪些特征更重要。
- 建模之后，再回顾我们这里认为比较重要或不重要的特征，看一下判断是否准确。

三、案例分析

1. 分析流程

- 导入数据，观察数据
  - 了解数据样本和特征个数、数据类型、基本信息等
  - 统计数据基本信息、统计空值数量
  - 检查数据中是否有重复值需要删除（将用户ID删除后再检测一次）

- 探索数据及数据可视化分析
  - 探索样本分类是否平衡
  - 用户年龄分布情况
  - 探索用户年龄和购买商业医疗保险之间的关系
  - 探索用户性别，以及性别和购买保险之间的关系
  - 探索用户学习，以及学历购买之间的情况
  - 根据业务理解对数据进行探索分析

- 空值填充
  - 探索数据中有哪些特征含有空值
  - 探索空值的个数和比例是多少
  - 分析这些空值的特点，确定填充策略
  - 根据我们的策略编写函数进行空值填充

- 变量编码
  - 将无效特征删除
  - 思考不同的特征应该采用什么方法编码
  - 根据分类变量的分类水平个数进行数值化编码
  - 将编码后的数据进行保存

- 数据建模
  - 切分数据集
  - 查看模型基础效果
  - 模型调参

2. Python代码实现

【详见课堂代码】

3. 输出结果的业务应用

我们来看一下购买比例最高的两类客户的特征是什么：

- **第一类**
    - 处于医疗险覆盖率比例较低区域
    - 居住年限小于7年
    - 65-72岁群体

那么我们对业务人员进行建议的时候就是，建议他们在医疗险覆盖率比例较低的区域进行宣传推广，然后重点关注那些刚到该区域且年龄65岁以上的老人，向这些人群进行保险营销，成功率应该会更高。

- **第二类**
    - 处于医疗险覆盖率比例较低区域
    - 居住年限大于7年
    - 居住房屋价值较高

这一类人群，是区域内常住的高端小区的用户。这些人群也同样是我们需要重点进行保险营销的对象。

除此之外，我们还可以做什么呢?

- **了解客户需求**

  我们需要了解客户的需求，并根据客户的需求举行保险营销。PIOS数据：向客户推荐产品，并利用个人的数据（个人特征）向客户推荐保险产品。旅行者：根据他们自己的数据（家庭数据），生活阶段信息推荐的是财务保险、人寿保险、保险、旧保险和用户教育保险。外部数据、资产保险和人寿保险都提供给高层人士，利用外部数据，我们可以改进保险产品的管理，增加投资的收益和收益。

- **开发新的保险产品**

  保险公司还应协助外部渠道开发适合不同商业环境的保险产品，例如新的保险类型，如飞行延误保险、旅行时间保险和电话盗窃保险。目的是提供其他保险产品，而不是从这些保险中受益，而是寻找潜在的客户。此外，保险公司将通过数据分析与客户联系，了解客户。外部因素将降低保险的营销成本，并直接提高投资回报率。

四、【补充】特征降维-PCA

sklearn中降维算法都被包括在模块decomposition中，这个模块本质是一个矩阵分解模块。在过去的十年中，如果要讨论算法进步的先锋，矩阵分解可以说是独树一帜。矩阵分解可以用在降维，深度学习，聚类分析，数据预处理，低纬度特征学习，推荐系统，大数据分析等领域。

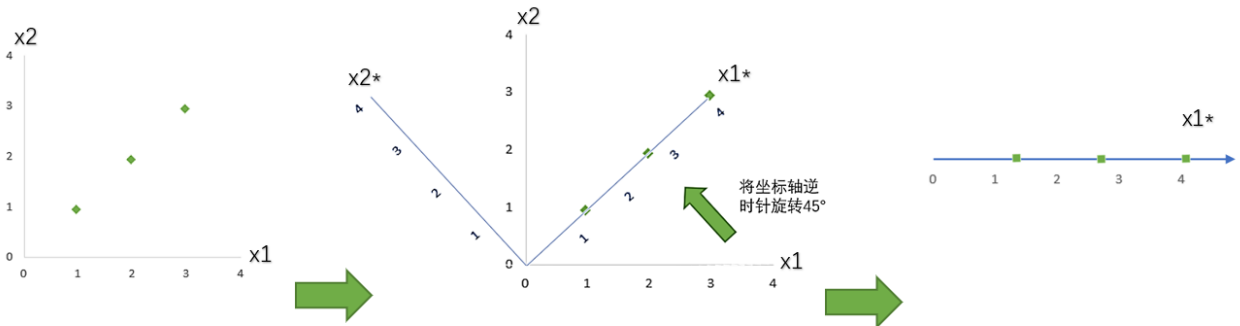
其中主成分分析PCA属于矩阵分解算法中的入门算法，通过分解特征矩阵来进行降维。

在降维过程中，我们会减少特征的数量，这意味着删除数据。数据量变少则表示模型可以获取的信息会变少，模型的表现可能会因此受影响。同时，在高维数据中，必然有一些特征是不带有有效的信息的（比如噪音），或者有一些特征带有的信息和其他一些特征是重复的（比如一些特征可能会线性相关）。我们希望能够找出一种办法来帮助我们衡量特征上所带的信息量，让我们在降维的过程中，能够**既减少特征的数量，又保留大部分有效信息**——将那些带有重复信息的特征合并，并删除那些带无效信息的特征等等——逐渐创造出能够代表原特征矩阵大部分信息的，特征更少的，新特征矩阵。

1. 降维究竟是怎样实现的?

class* `sklearn.decomposition.PCA`(*n_components=None, copy=True, whiten=False, svd_solver='auto', tol=0.0, iterated_power='auto', random_state=None*)

PCA作为矩阵分解算法的核心算法，其实没有太多参数，但不幸的是每个参数的意义和运用都很难，因为几乎每个参数都涉及到高深的数学原理。为了参数的运用和意义变得明朗，我们来看一组简单的二维数据的降维。



x1	x2
1	1
2	2
3	3

x1*	x2*
$\sqrt{2}$	0
$2\sqrt{2}$	0
$3\sqrt{2}$	0

x1*
$\sqrt{2}$
$2\sqrt{2}$
$3\sqrt{2}$

我们现在有一组简单的数据，有特征x1和x2，三个样本数据的坐标点分别为(1,1)，(2,2)，(3,3)。我们可以让x1和x2分别作为两个特征向量，很轻松地用一个二维平面来描述这组数据。这组数据现在每个特征的均值都为2，方差则等于：

$$x1\_var = x2\_var = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{2} = 1$$

每个特征的数据一模一样，因此方差也都为1，数据的方差总和是2。

现在我们的目标是：只用一个特征向量来描述这组数据，即将二维数据降为一维数据，并且尽可能地保留信息量，即让数据的总方差尽量靠近2。于是，我们将原本的直角坐标系逆时针旋转45°，形成了新的特征向量x1*和x2*组成的新平面，在这个新平面中，三个样本数据的坐标点可以表示为$(\sqrt{2},0)$，$(2\sqrt{2},0)$，$(3\sqrt{2},0)$。可以注意到，x2*上的数值此时都变成了0，因此x2*明显不带有任何有效信息了（此时x2*的方差为0了）。此时，x1*特征上的数据均值是$2\sqrt{2}$，而方差则可表示成：

$$x1^*\_var = \frac{(\sqrt{2}-2\sqrt{2})^2+(2\sqrt{2}-2\sqrt{2})^2+(3\sqrt{2}-2\sqrt{2})^2}{2} = 2$$

x2*上的数据均值为0，方差也为0。

此时，我们根据信息含量的排序，取信息含量最大的一个特征，因为我们想要的是一维数据。所以我们可以将x2*删除，同时也删除图中的x2*特征向量，剩下的x1*就代表了曾经需要两个特征来代表的三个样本点。通过旋转原有特征向量组成的坐标轴来找到新特征向量和新坐标平面，我们将三个样本点的信息压缩到了一条直线上，实现了二维变一维，并且尽量保留原始数据的信息。一个成功的降维，就实现了。

不难注意到，在这个降维过程中，有几个重要的步骤：

步骤	2维特征矩阵	n维特征矩阵
1	输入原数据，结构为 (3,2) 找出原本的2个特征对应的直角坐标系，本质是找出这2个特征构成的2维平面	输入原数据，结构为 (m,n) 找出原本的n个特征向量构成的n维空间V
2	决定降维后的特征数量：1	决定降维后的特征数量：k
3	旋转，找出一个新坐标系 本质是找出2个新的特征向量，以及它们构成的新2维平面 新特征向量让数据能够被压缩到少数特征上，并且总信息量不损失太多	通过某种变化，找出n个新的特征向量，以及它们构成的新n维空间V
4	找出数据点在新坐标系上，2个新坐标轴上的坐标	找出原始数据在新特征空间V中的n个新特征向量上对应的值，即“将数据映射到新空间中”
5	选取第1个方差最大的特征向量，删掉没有被选中的特征，成功将2维平面降为1维	选取前k个信息量最大的特征，删掉没有被选中的特征，成功将n维空间V降为k维

思考：PCA和特征选择技术都是特征工程的一部分，它们有什么不同?

特征工程中有三种方式：特征提取，特征选择和特征创造。仔细观察上面的降维例子和上周我们讲解过的特征选择，你发现有什么不同了吗?

特征选择是从已存在的特征中选取携带信息最多的，选完之后的特征依然具有可解释性，我们依然知道这个特征在原数据的哪个位置，代表着原数据上的什么含义。

而PCA，是将已存在的特征进行压缩，降维完毕后的特征不是原本的特征矩阵中的任何一个特征，而是通过某些方式组合起来的新特征。通常来说，**在新的特征矩阵生成之前，我们无法知晓PCA都建立了怎样的新特征向量，新特征矩阵生成之后也不具有可读性**，我们无法判断新特征矩阵的特征是从原数据中的什么特征组合而来，新特征虽然带有原始数据的信息，却已经不是原数据上代表着的含义了。以PCA为代表的降维算法因此是特征创造（feature creation，或feature construction）的一种。

可以想见，PCA一般不适用于探索特征和标签之间的关系的模型（如线性回归），因为无法解释的新特征和标签之间的关系不具有意义。在线性回归模型中，我们使用特征选择。

2. 重要参数n_components

n_components是我们降维后需要的维度，即降维后需要保留的特征数量，降维流程中第二步里需要确认的k值，一般输入[0, min(X.shape)]范围中的整数。一说到K，大家可能都会想到，类似于KNN中的K和随机森林中的n_estimators，这是一个需要我们人为去确认的超参数，并且我们设定的数字会影响到模型的表现。如果留下的特征太多，就达不到降维的效果，如果留下的特征太少，那新特征向量可能无法容纳原始数据集中的大部分信息，因此，n_components既不能太大也不能太小。那怎么办呢?

可以先从我们的降维目标说起：如果我们希望可视化一组数据来观察数据分布，我们往往将数据降到三维以下，很多时候是二维，即n_components的取值为2。

2.1 鸢尾花数据集的可视化

```
#导入模块和包
from sklearn.datasets import load_iris   #导入sklearn中内置的鸢尾花数据集
from sklearn.decomposition import PCA    #导入PCA降维算法
import matplotlib.pyplot as plt          #绘图包

#提取数据集
iris = load_iris()
y = iris.target
X = iris.data

#调用PCA建模
pca = PCA(n_components=2)          #实例化
pca = pca.fit(X)                   #拟合模型
X_dr = pca.transform(X)            #获取新矩阵

#绘制可视化图形
#要将三种鸢尾花的数据分布显示在二维平面坐标系中，对应的两个坐标（两个特征向量）应该是三种鸢尾花降维后的x1和x2，怎样才能取出三种鸢尾花下不同的x1和x2呢?

X_dr[y == 0, 0] #这里是布尔索引，看出来了么?

#要展示三中分类的分布，需要对三种鸢尾花分别绘图
#可以写成三行代码，也可以写成for循环

"""
plt.figure()
plt.scatter(X_dr[y==0, 0], X_dr[y==0, 1], c="red", label=iris.target_names[0])
plt.scatter(X_dr[y==1, 0], X_dr[y==1, 1], c="black",
label=iris.target_names[1])
plt.scatter(X_dr[y==2, 0], X_dr[y==2, 1], c="orange",
label=iris.target_names[2])
plt.legend()
plt.title('PCA of IRIS dataset')
plt.show()
```

```
"""

colors = ['red', 'black', 'orange']
iris.target_names

plt.figure()
for i in [0, 1, 2]:
    plt.scatter(X_dr[y == i, 0]
                ,X_dr[y == i, 1]
                ,alpha=.7
                ,c=colors[i]
                ,label=iris.target_names[i]
               )
plt.legend()
plt.title('PCA of IRIS dataset')
plt.show()
```

鸢尾花的分布被展现在我们眼前了，明显这是一个分簇的分布，并且每个簇之间的分布相对比较明显，也许versicolor和virginia这两种花之间会有一些分类错误，但setosa肯定不会被分错。这样的数据很容易分类，可以遇见，KNN，决策树，朴素贝叶斯等分类器在鸢尾花数据集上，未调整的时候都可以有95%上下的准确率。

- **探索降维后的数据**

```
#属性explained_variance_，查看降维后每个新特征向量上所带的信息量大小（可解释性方差的大小）
pca.explained_variance_

#属性explained_variance_ratio，查看降维后每个新特征向量所占的信息量占原始数据总信息量的百分比
#又叫做可解释方差贡献率
pca.explained_variance_ratio_
#大部分信息都被有效地集中在了第一个特征上

pca.explained_variance_ratio_.sum()
```

2.2 选择最好的n_components

2.2.1 累计可解释方差贡献率曲线

当参数n_components中不填写任何值，则默认返回min(X.shape)个特征，一般来说，样本量都会大于特征数目，所以什么都不填就相当于转换了新特征空间，但没有减少特征的个数。一般来说，不会使用这种输入方式。但我们却可以使用这种输入方式来画出累计可解释方差贡献率曲线，以此选择最好的n_components的整数取值。

累计可解释方差贡献率曲线是一条以降维后保留的特征个数为横坐标，降维后新特征矩阵捕捉到的可解释方差贡献率为纵坐标的曲线，能够帮助我们决定n_components最好的取值。

```
import numpy as np
pca_line = PCA().fit(X)
plt.plot([1,2,3,4],np.cumsum(pca_line.explained_variance_ratio_))
plt.xticks([1,2,3,4]) #这是为了限制坐标轴显示为整数
plt.xlabel("number of components after dimension reduction")
plt.ylabel("cumulative explained variance ratio")
plt.show()
```

2.2.2 最大似然估计自选超参数

除了输入整数，n_components还有哪些选择呢？之前我们提到过，矩阵分解的理论发展在业界独树一帜，勤奋智慧的数学大神Minka, T.P.在麻省理工学院媒体实验室做研究时找出了让PCA用最大似然估计(maximum likelihood estimation)自选超参数的方法，输入“mle”作为n_components的参数输入，就可以调用这种方法。

```
pca_mle = PCA(n_components="mle")
pca_mle = pca_mle.fit(X)
X_mle = pca_mle.transform(X)

X_mle
#可以发现，mle为我们自动选择了3个特征

pca_mle.explained_variance_ratio_.sum()
#得到了比设定2个特征时更高的信息含量，对于鸢尾花这个很小的数据集来说，3个特征对应这么高的信息含量，并不需要去纠结于只保留2个特征，毕竟三个特征也可以可视化
```

2.2.3 按信息量占比选超参数

输入[0,1]之间的浮点数，并且让参数svd_solver =='full'，表示希望降维后的总解释性方差占比大于n_components指定的百分比，即是说，希望保留百分之多少的信息量。比如说，如果我们希望保留97%的信息量，就可以输入n_components = 0.97，PCA会自动选出能够让保留的信息量超过97%的特征数量。

```
pca_f = PCA(n_components=0.97,svd_solver="full")
pca_f = pca_f.fit(X)
X_f = pca_f.transform(X)

pca_f.explained_variance_ratio_
```

特征工程非常深奥，虽然我们日常可能用到不多，但其实它非常美妙。多读多看多试多想，技术逐渐会成为你的囊中之物~