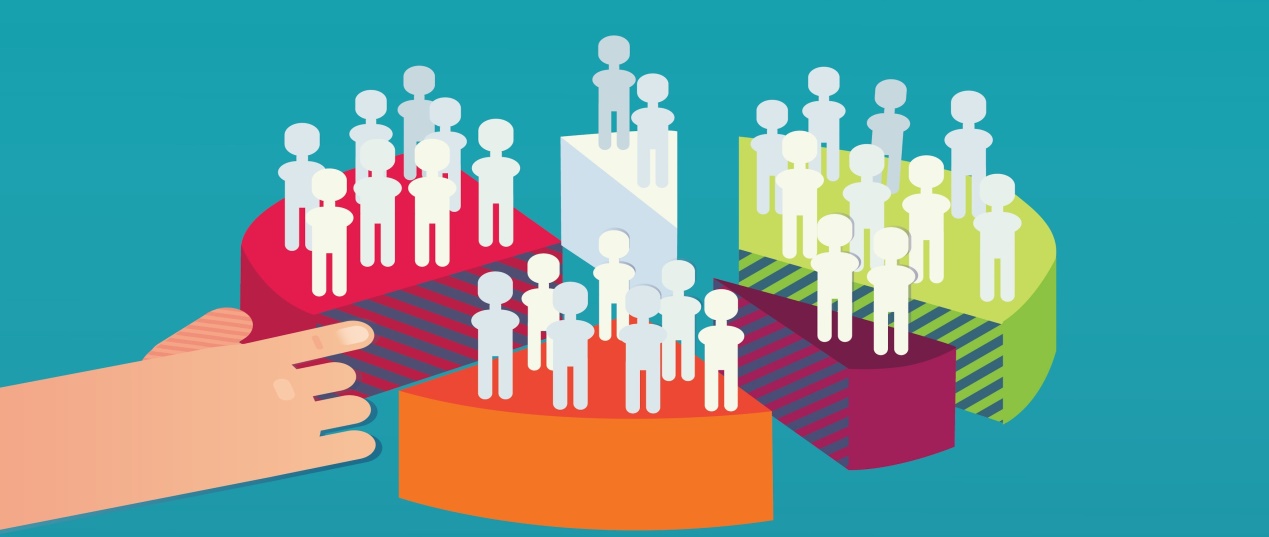
Bertelsmann Arvato 邮购公司客户细分和预测



# 介绍

## 项目概述

在商业直销活动中，使用标签化的用户模型对进行营销活动进行优化是实现精准营销的一个重要工具。本项目要求根据一家德国邮购公司已有客户的人口学特征数据，使用非监督学习技术来分析公司现有顾客的主要特征。然后利用监督学习技术搭建预测模型，在德国的总人口统计数据集上预测新客户。这篇文章是与贝塔斯曼Arvato Analytics合作提供的Udacity数据科学家纳米学位毕业项目的一部分。

本项目的数据和大纲由贝塔斯曼子公司Arvato Financial Solutions提供。

本项目该项目分为三个主要步骤：客户细分报告、监督学习模型和Kaggle竞争：

1. 现有客户细分报告：采用EDA、PCA和聚类分析的方法对总体和客户群进行分析，目的是能够描述公司的核心客户群。
2. 新客户的预测报告（监督学习）-使用在第一步中观察到的内容创建一个监督学习模型，用模型预测一个人是否会响应邮件营销活动。
3. Kaggle 比赛: 在营销活动的人口统计数据集上预测新客户

## 问题描述

根据Arvato Financial Solutions 的高级大客户经理Time Reis的说法，该项目要解决的基本商业问题是：如何让一家邮购公司更高效的获得新客户?

## 度量指标

根据问题的类型，我们将使用两种不同的度量：对非监督学习模型，要用PCA 和聚类来分析邮购公司现有客户的特征，并将其与德国一般人群的人口统计数据进行对比，我们将采用的策略是**肘部法则。 对**监督学习模型，需要对营销活动的对象（第三个数据集 MAILOUT）作预测，要判断一个人是否会对一个邮件营销活动有响应，该数据集有显著的类别不平衡问题（仅有约1.24%的人会对该活动有响应）。因此我们采用ROC - AUC (Receiver Operating Characteristic - Area Under the Curve)指标进行衡量.

# 探索数据和数据整理

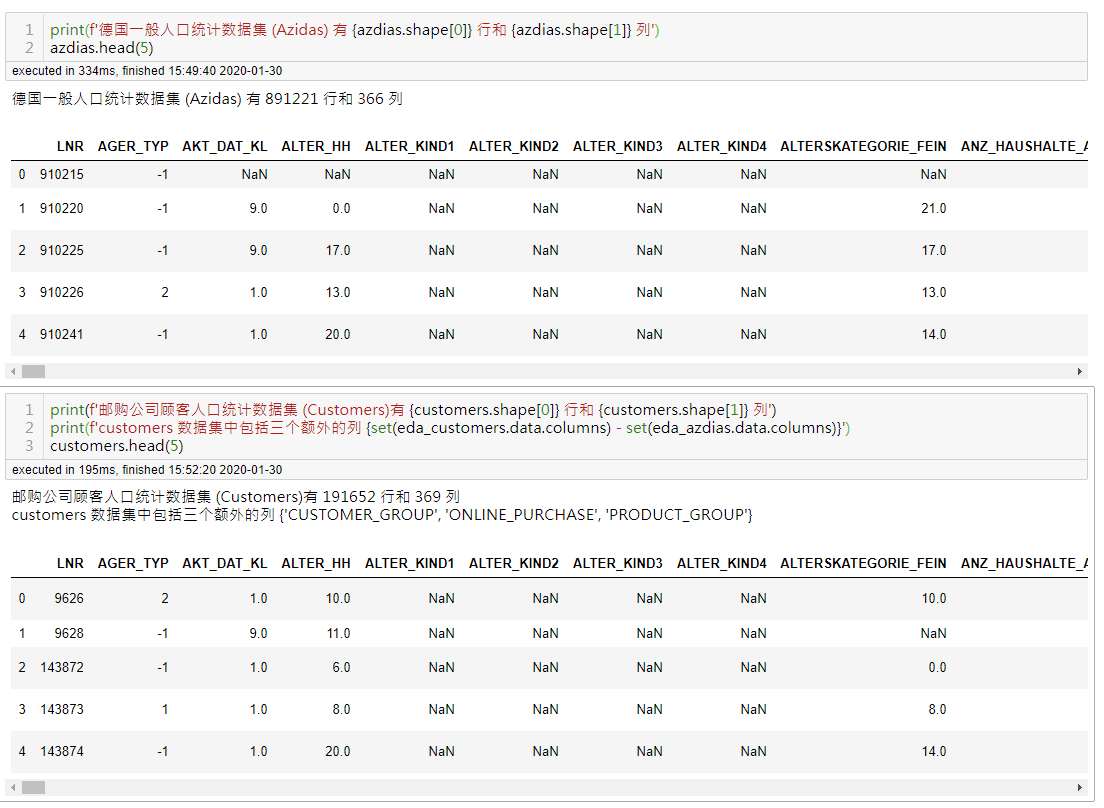
这个项目的数据包括四个数据文件和两个特征定义文件：

AZDIAS: 德国的一般人口统计数据

CUSTOMERS: 邮购公司顾客的人口统计数据

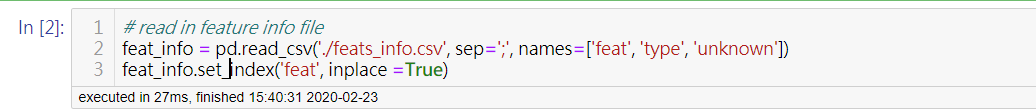
MAILOUT\_TRAIN 和 MAILOUT\_TEST: 营销活动的对象的人口统计数据

## 加载和浏览数据：

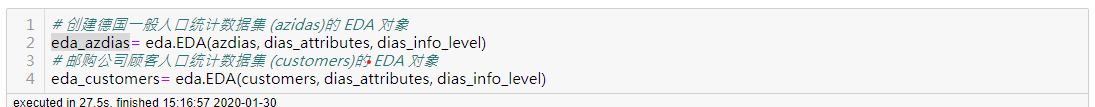


德国的一般人口统计数据集(Azdias)有891221行和366列，邮购公司顾客数据集(Customers)有191652 行369列。Customers数据集包括三个额外的列（'CUSTOMER\_GROUP'、'ONLINE\_PURCHASE' 和 'PRODUCT\_GROUP'），提供了文件中顾客的更多维度的信息。原始的 "MAILOUT" 包括一个额外的列 "RESPONSE"，表示每个收到邮件的人是否成为了公司的顾客。人口统计数据文件的每一行代表一个人，但也包含个人以外的信息，包括有关他们的家庭，建筑物和邻居的信息。

加载数据集特征的统计信息文件：



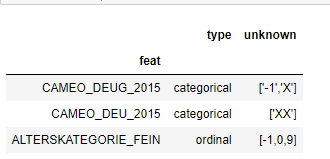
创建数据集 EDA 对象：



## 数据清洗

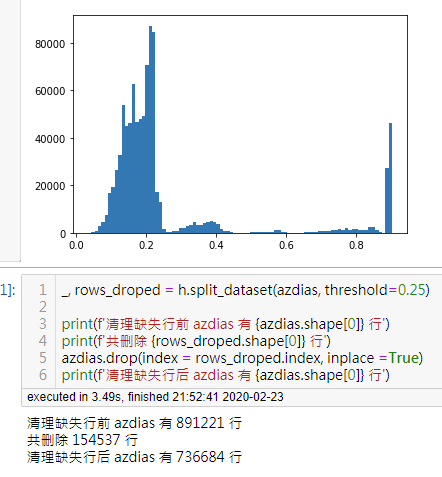
开始做饭的第一步都是把买回的菜洗干净，过一下开水。拿到的数据都需要先做清洗工作，排除异常值、空白值、无效值、重复值等。这项工作经常会占到整个数据分析过程将近一半的时间。

Step 1: 数据清洗的第一步是把数据集中的未知数据转换为NaN

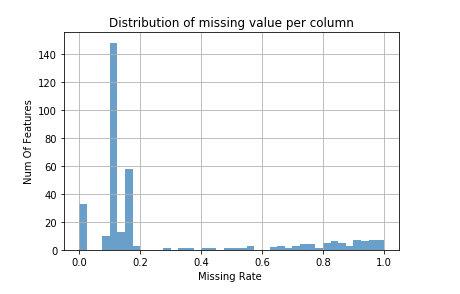


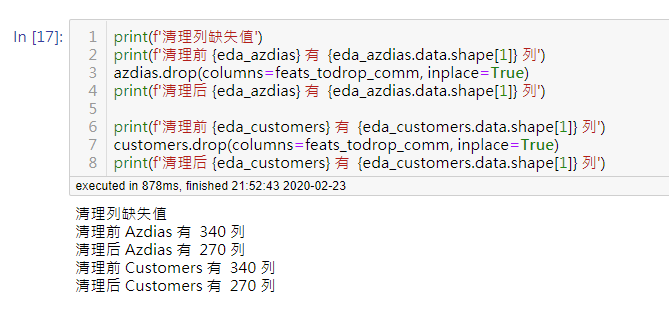
这个数据集把缺失数据定义为 unknown, 但如上图示例,具体到每个特征，unknown的标记有很多种。我们需要把这些缺失值和未知数据统一转换NaN. 在接下来我们将删除缺失值多的行和列。

Step 2: 删除缺失值多的行/列



上图显示，Azdias 数据集中，有154537行数据的缺失率超过0.25 ，这些行被删除。





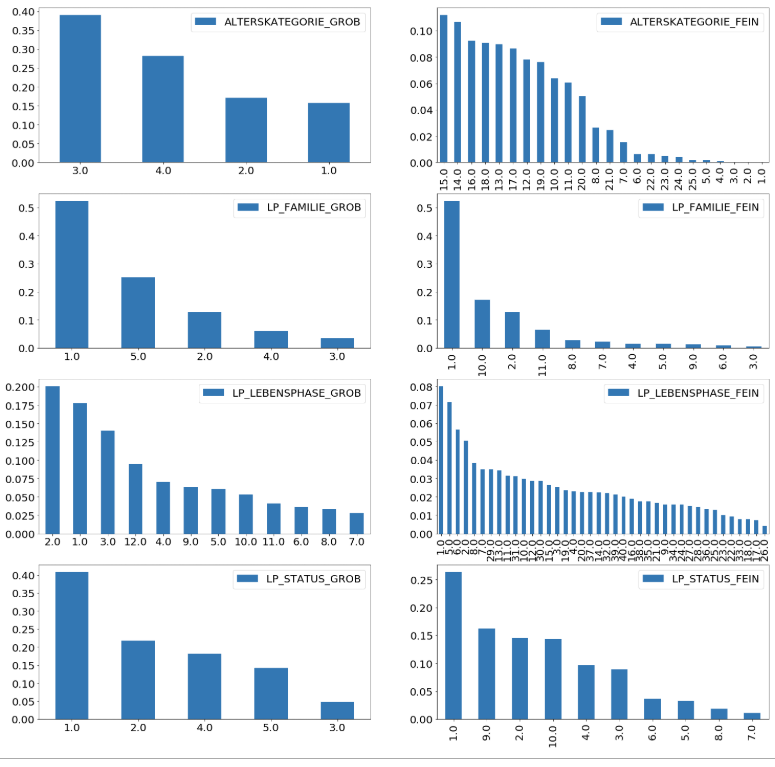
清理特征(列)时，缺失率阈值设置为0.37，如上图所示，Azdias和Customers数据集中有70列被删除。

Step 3: 删除未定义特征

删除缺少信息和无定义的特征,共计26个, 然后从 Customers 数据集中删除三个专有的列（'CUSTOMER\_GROUP'、'ONLINE\_PURCHASE' 和 'PRODUCT\_GROUP'）。因为后面比较Azdias 和 Customers两个数据集时，这些未定义特征和专有特征没有帮助。

Step 4: 删除信息重复的特征

数据集中有很多信息重复的特征，例如 LP\_FAMILIE\_FEIN, LP\_FAMILIE\_GROB， 都是表示家庭类型信息，LP\_FAMILIE\_FEIN表示细粒度信息，LP\_FAMILIE\_GROB表示粗粒度信息，二者只是信息粒度不同。



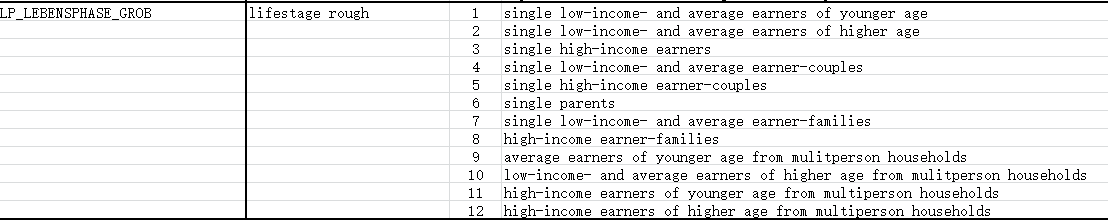
上图中同一行的两个特征都是表示相同的含义，左侧后缀是\_GROB的特征表示粗粒度信息，\_FEIN表示细粒度特征。我决定删除这几个信息重复特征，'ALTERSKATEGORIE\_FEIN', 'LP\_FAMILIE\_GROB', 'LP\_LEBENSPHASE\_FEIN', 'LP\_STATUS\_GROB','CAMEO\_DEU\_2015'。因为他们有些是没有定义，或者选择对应的\_GROB特征就够用了。

Step 5: 特征重编码

有很多特征需要重编码，例如 OST\_WEST\_KZ 特征的数值中有两个是字符 O, W. 由于要使用的非监督学习技术仅适用于以数字方式编码的数据，因此我们需要对其进行重编码。

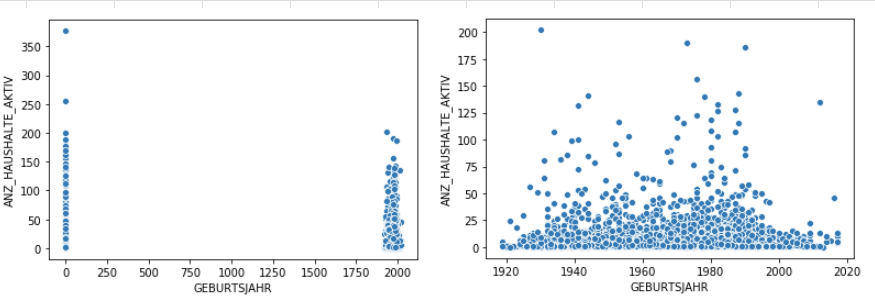
****

还有一些特征是混合型分类特征，需要拆分为多个单一类型特征。 例如下图的



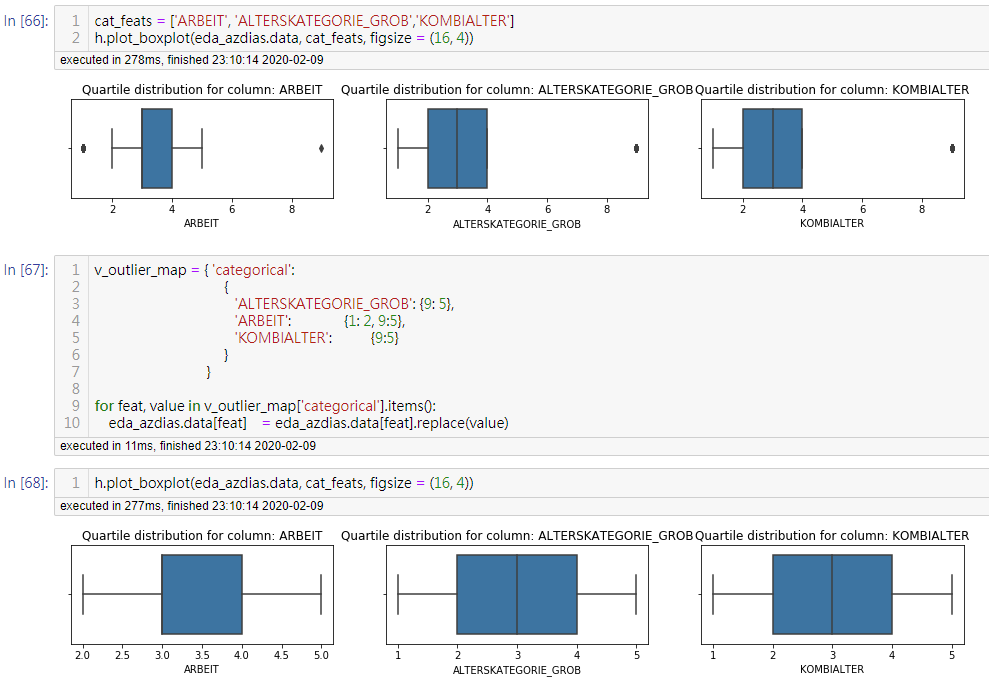
混合型特征LP\_LEBENSPHASE\_GROB - lifestage rough. 被拆分为三个包含年龄、家庭和收入的新特征（LP\_LEBENSPHASE\_GROB\_SPLIT\_FAMILY, LP\_LEBENSPHASE\_GROB\_SPLIT\_AGE, LP\_LEBENSPHASE\_GROB\_SPLIT\_INCOME）。

Step 6: 处理离群值



上面左侧的散点图中，特征GEBURTSJAHR ( year of birth) 中有很多零值，出生日期是0明显不合理，需要转换为NaN. 右侧是处理后的结果。

对特征ARBEIT, ALTERSKATEGORIE\_GROB,KOMBIALTER 中的离群值进行处理。



总结一下数据清理过程，我主要做了以下操作：

* 将未知数据还原为NaN，并删除缺失率百分比较高的行和列
* 清理了信息重复的特征
* 特征重编码和拆分混合特征
* 用归一化进行特征缩放,并用“mean”填充缺少的值

数据清理是一个迭代的过程，也是最耗时的工序，本项目中，数据清理过程的消耗时间占总时间的80%以上。

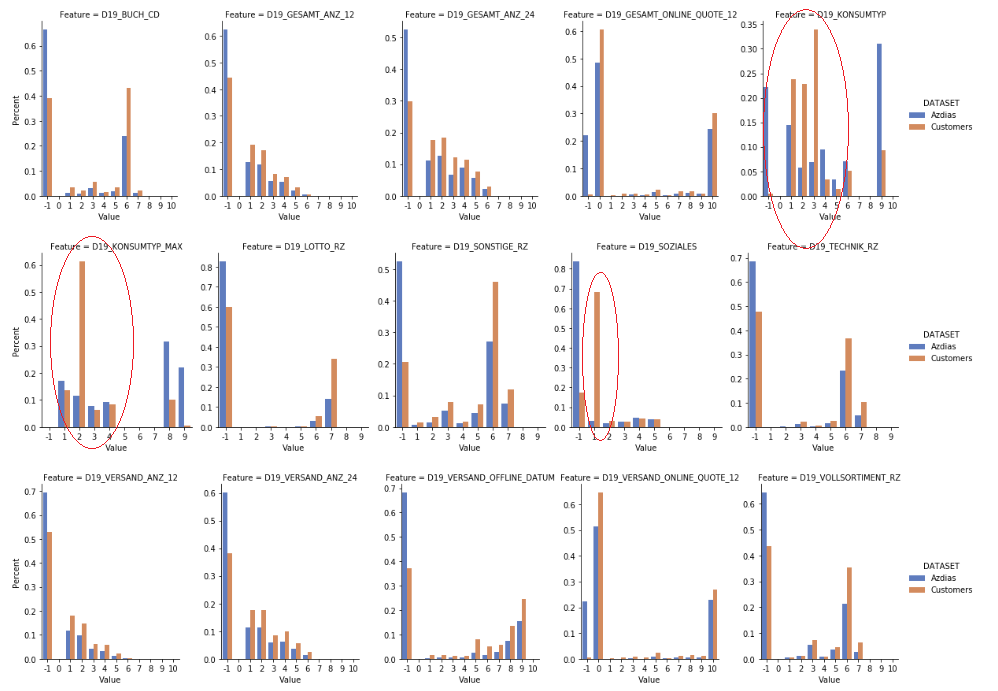
# 客户细分报告

在本节中，我将使用非监督学习技术来刻画邮购公司已有顾客和德国一般人群的人口统计数据的关系，发现一般人群中的哪一类人更可能是邮购公司的主要核心顾客，哪些人则很可能不是。接下来，我将使用两种常用的非监督机器学习方法,降维和聚类进行客户细分。

## 数据比较Customers vs. Azdias

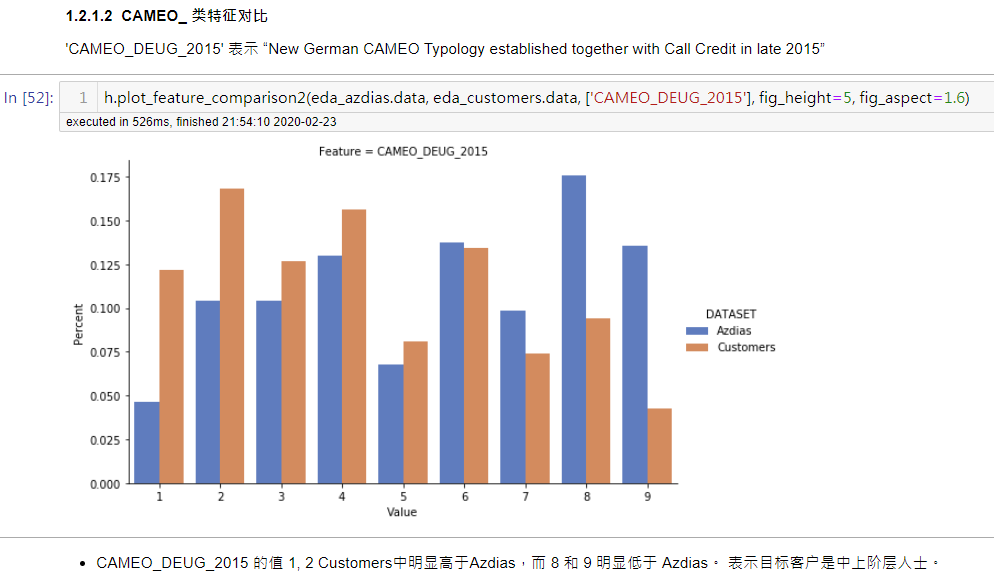
消费者邮购活动的相关特征 (以D19\_开头) 对比：

D19: This data contains unique data regarding the mail-order activity of consumers, differentiated 根据字典定义，以D19\_开头的特征是包含有关消费者邮购活动的唯一数据。



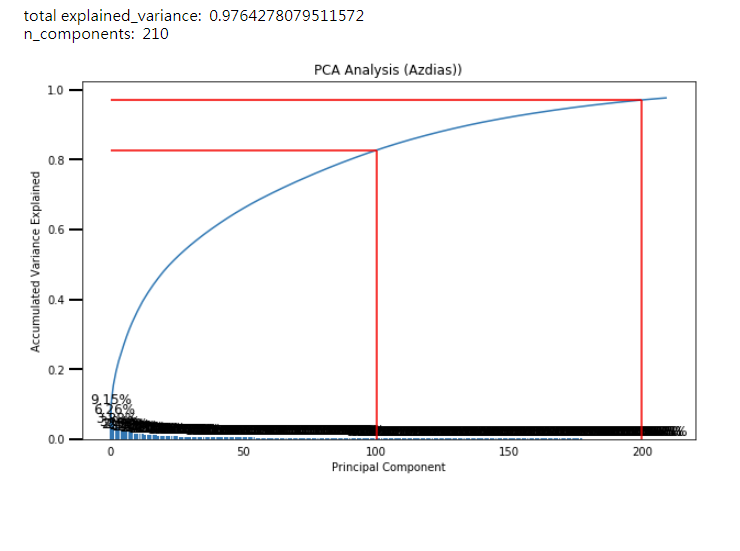
D19\_SOZIALES 的值 1 在Customers中显著高于Azdias，虽然特征字典没有给出D19\_SOZIALES的定义。但是可以猜到它表示的是一种商品，而值 1表示在过去一年中有多次购买 (Multibuyer 0-12 months). D19\_SOZIALES 的值 1 在Customers中显著高于Azdias， 虽然特征字典没有给出D19\_SOZIALES的定义。但是可以猜到它表示的是一种商品，而值 1表示在过去一年中有多次购买 (Multibuyer 0-12 months). D19\_KONSUMTYP\_MAX 的值 2 显著高于Azdias, 8 和 9 明显低于Azdias. 但字典没有给出 D19\_KONSUMTYP\_MAX 的定义。

CAMEO\_ 类特征对比：



## 主成分分析 (PCA)

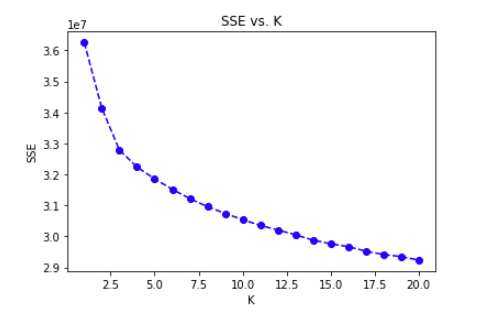
降维是将大量的特征压缩为（通常是）更小的特征集。主成分分析 (PCA)是一种数据转换技术，它消繁就简，删除次要信息，只保留主要特征而允许你保留数据中的主要信息。我们将使用scikit-learn 中的 PCA 来降低数据集的维数



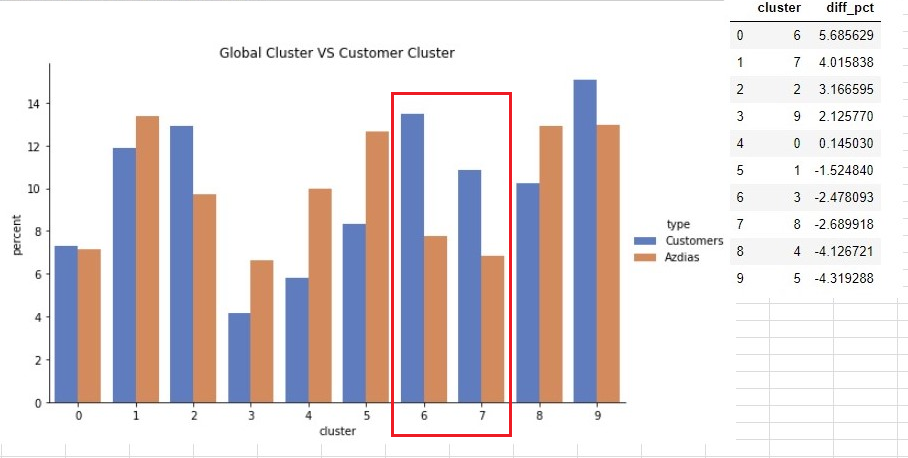
最后，我选择保留210个主成分，其累积方差是0.97%， 我认为其已经能够保留数据中的主要信息。

## 聚类

聚类是基于相似性对数据进行分组。在选择好主成分后，我们希望通过聚集这些主成分来帮助细分客户。我使用scikit-learn的K-Means来聚类主成分。



根据肘部法则，我选择聚类数量为 10.



聚类比较 Customers vs. Azdias 显示聚类6和7，Customers明显高于Azdias。指示这里有特定人群对公司产品感兴趣。

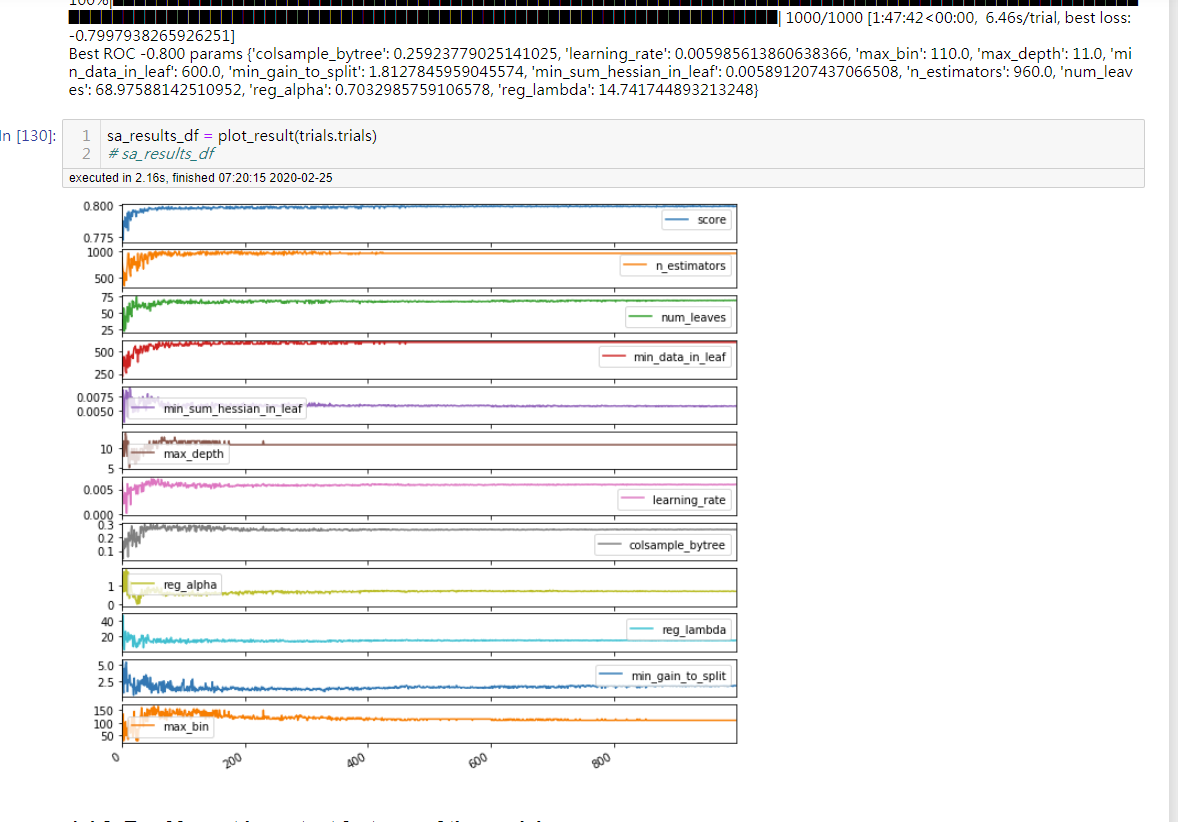
# 监督学习模型和Kaggle比赛

现在开始搭建预测模型，"MAILOUT"数据文件的的每一行表示一个邮购活动的潜在顾客，我们将使用MAILOUT\_TRAIN数据集训练这个模型，然后再MAILOUT\_TEST数据集上做预测，来决定是否该把他作为该活动的营销对象。

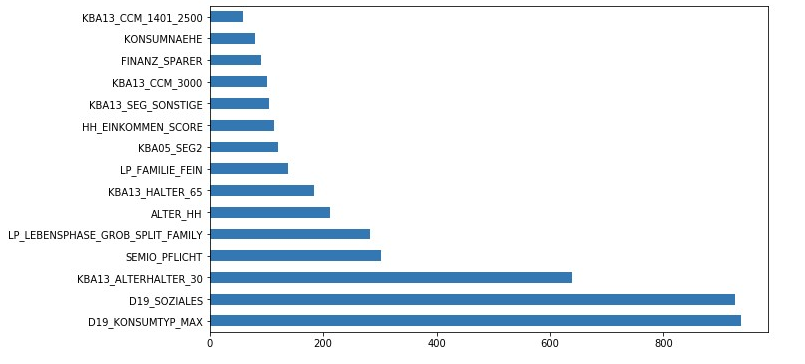
浏览训练数据集(MAILOUT)，我发现此训练数据集存在类别高度不平衡问题，即在42962人中，只有532人对邮递活动做出了回应。正例数远小于负例数量。因此在训练模型时将使用 sklearn库的分层k折交叉验证工具“StratifiedKFold”来分割训练数据和验证数据，以保证充分利用有限的正例数据，同时在创建模型是设置参数 class\_weight='balanced'，

我们要解决的基本商业问题是：如何让一家邮购公司更高效的获得新客户? 要解决这个问题我们需要一些工具，这些工具就是各种模型。可供选择的模型有很多, 比如有线性或逻辑回归，BaggingClassifier和决策树。我使用 LGBMClassifier，并使用10折分层折交叉验证方法，参数优化使用了 hyperopt。

## 模型训练结果



Top 15 主要特征：



## Kaggle比赛

我把结果提交到 Kaggle上后，roc\_auc\_score(Receiver Operating Characteristic Curve) 分数是 0.76682。



# 总结

在本博客中，我介绍了一个实际生产中的一个机器学习项，该项目是优达学城数据科学家纳米学位的一个毕业项目，数据由贝塔斯曼的子公司Arvato Financial Solutions 提供，在本项目中，我作的主要事情如下：

探索了德国的一般人口统计数据和邮购公司顾客的人口统计数据集

清洗数据，包括排除异常值、空白值、无效值、重复值等等

通过数据对比可视化和非监督学习技术，即PCA和KMeans，将群体分割（分成不同的集群），为公司推荐潜在客户

利用监督学习来预测公司的新客户

分析监督学习模型中的重要特征

最后，我要感谢Udacity、Arvato Bertelsmann和collaborate 公司提供如此精彩的数据集，让我学习和实践我的机器学习和数据科学技能。谢谢你的时间，希望你喜欢这个博客。