Arvato Financial Services顾客分群

项目介绍

该项目分析了德国的一家邮购公司的顾客的人口统计数据，将它和一般的人口统计数据进行比较。使用非监督学习技术来实现顾客分类，识别出哪些人群是这家公司的基础核心用户。之后，通过Xgboost和RandomForest结合搭建的VotingRegressor来预测哪些人更可能成为该公司的顾客。本项目数据集由Bertelsmann Arvato Analytics公司提供。

该项目分为2个阶段：

1. 使用非监督学习技术来实现顾客画像分群，识别何种类型的人群契合该公司的基础核心用户画像，主要使用：PCA、KMeans来实现。
2. 使用监督学习技术来预测人群成为该公司潜在顾客的可能性，主要使用：RandomForest、Xgboost、VotingClassifier、VotingRegressor。

数据探索与清洗

主要分为以下4个步骤：

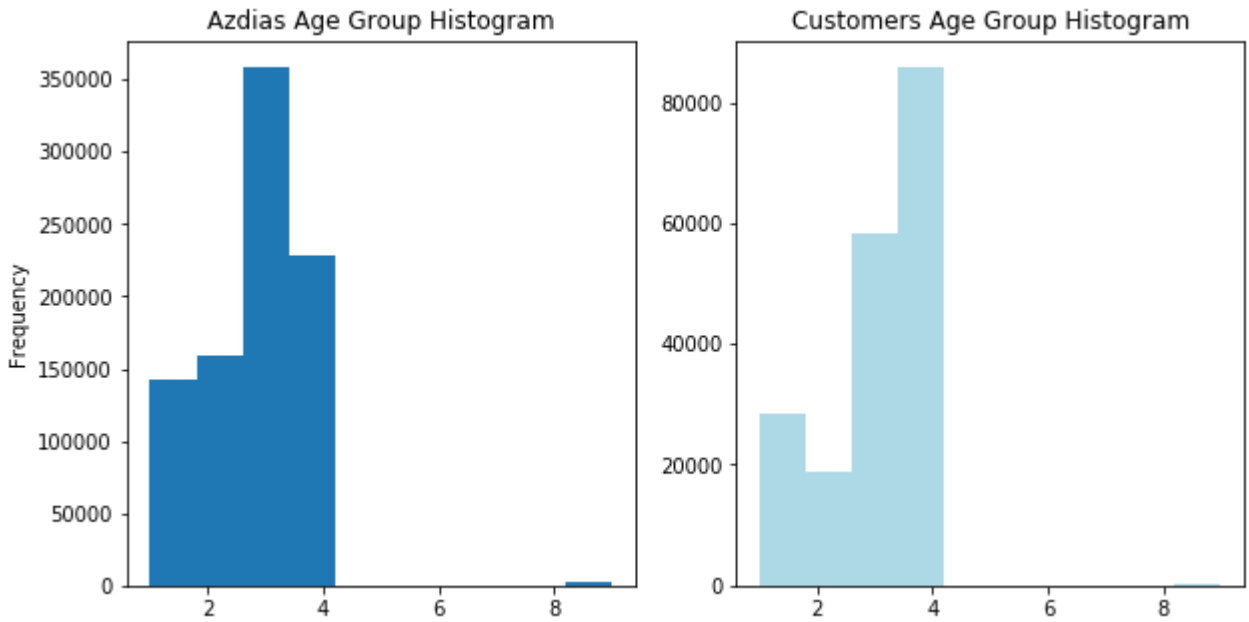
1.数据探索

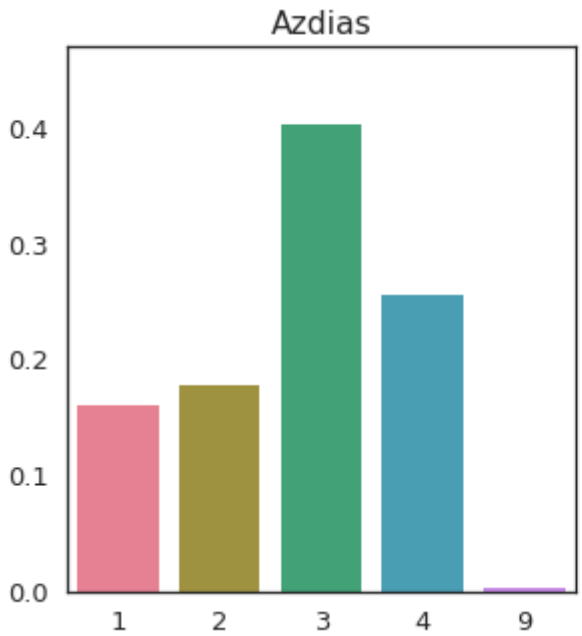
本项目使用的4个数据集由Bertelsmann Arvato Analycis公司提供，具体如下：

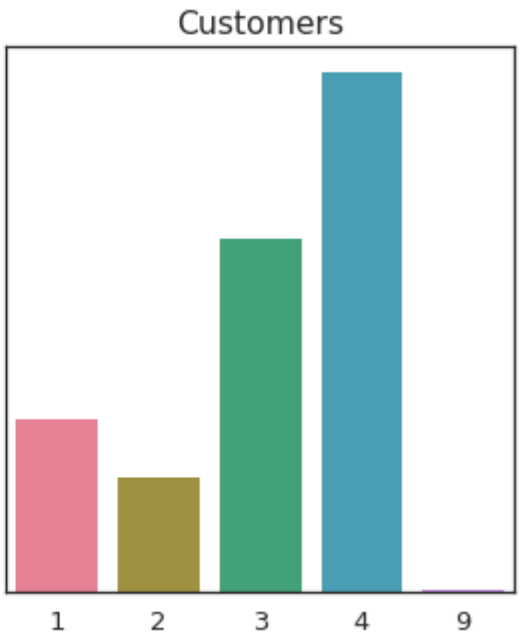
- **AZDIAS**：德国的一般人口统计数据；
- **CUSTOMERS**：邮购公司顾客的人口统计数据；
- **MAILOUT_TRAIN**：训练用营销活动的对象的人口统计数据；
- **MAILOUT_TEST**：预测用营销活动的对象的人口统计数据。

首先了解AZDIAS和CUSTOMERS数据集的年龄和性别分布

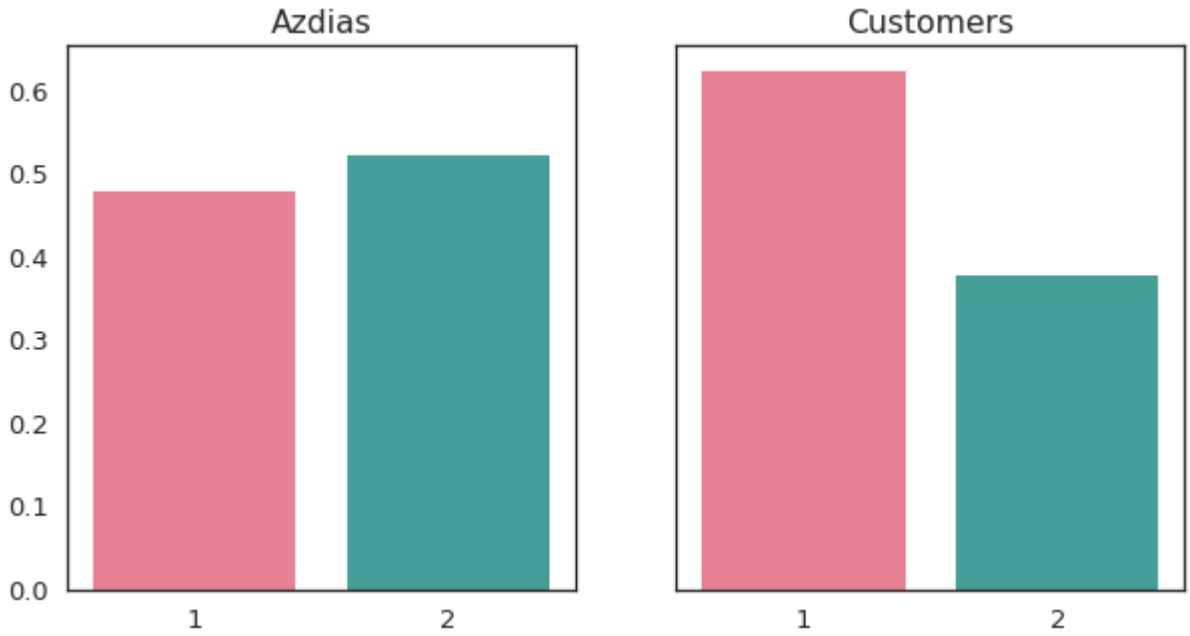
1. azdias中年龄组3（46-60岁）占比最高，customers中年龄组4（>60岁）占比最高。
2. customers中年龄组3和4的占比之和超过75%，即46岁以上的顾客占比超过75%。





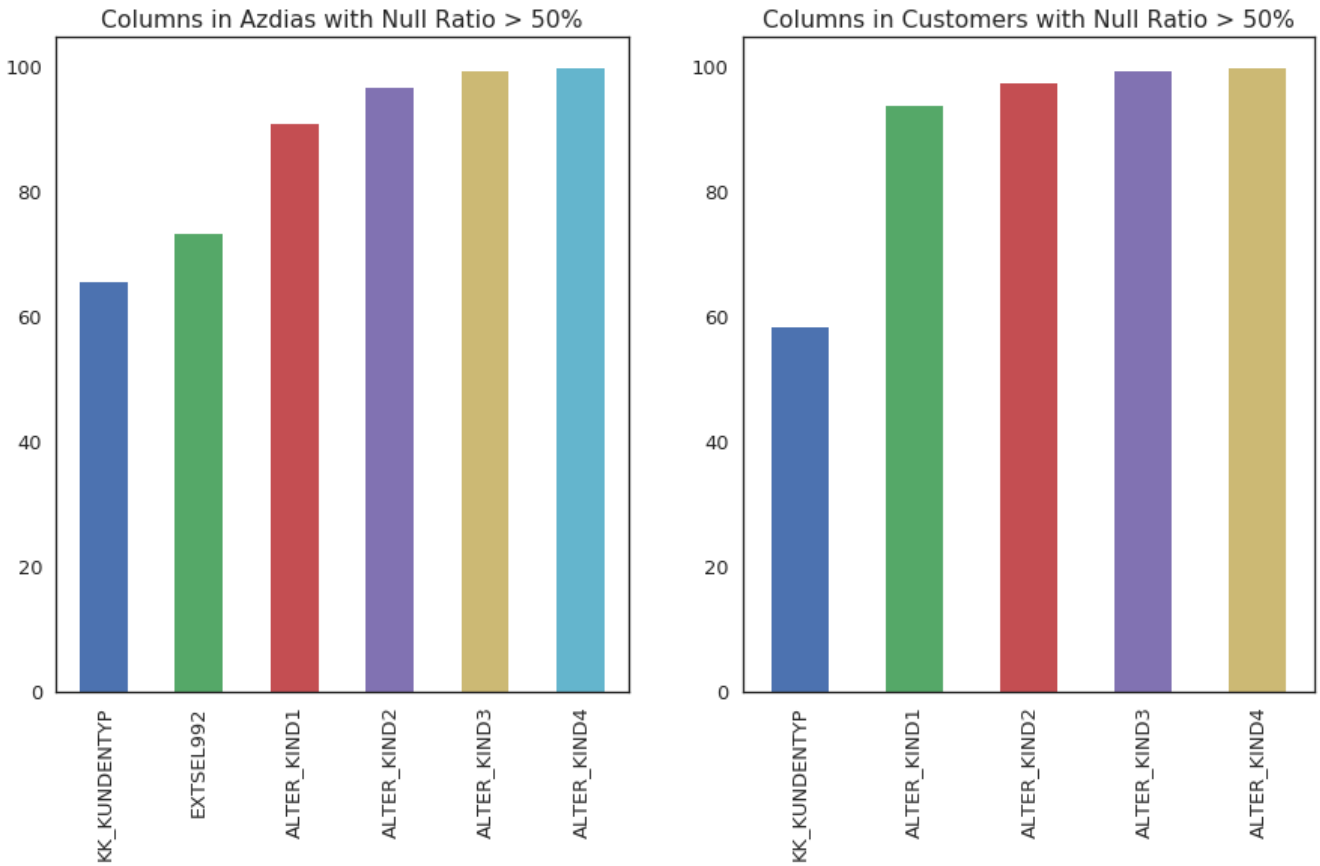


3. customers数据集中男性的比例超过60%，高于azdias（1表示男性，0表示女性）。

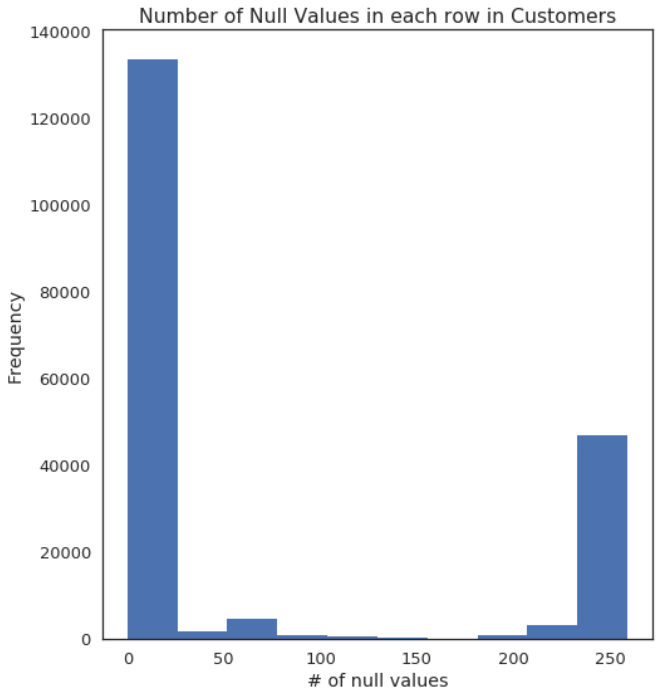


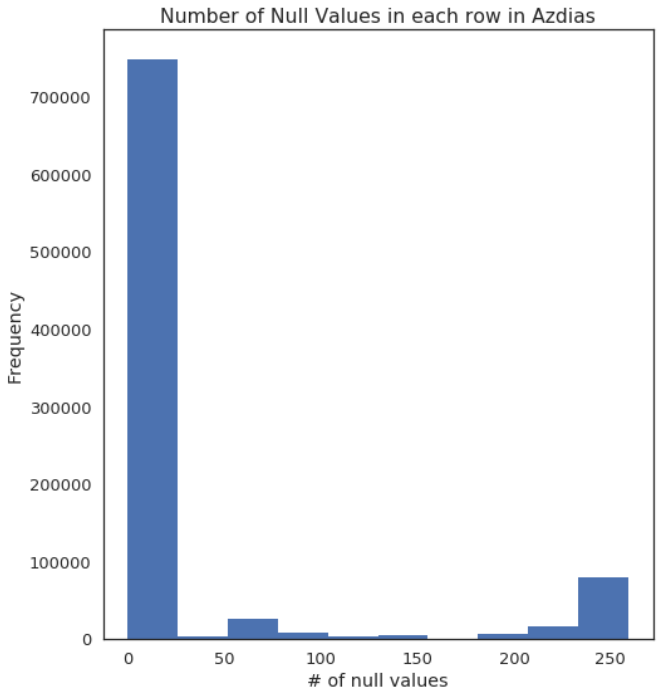
分析数据集中的缺失值，缺失值对机器学习算法的影响很大，azdias和customers数据集有超过300个特征列，数十万行的数据，需要对缺失值做特殊的处理。

数据集中缺失值比例超过50%的特征列



行缺失值数量分布





2.处理缺失值：

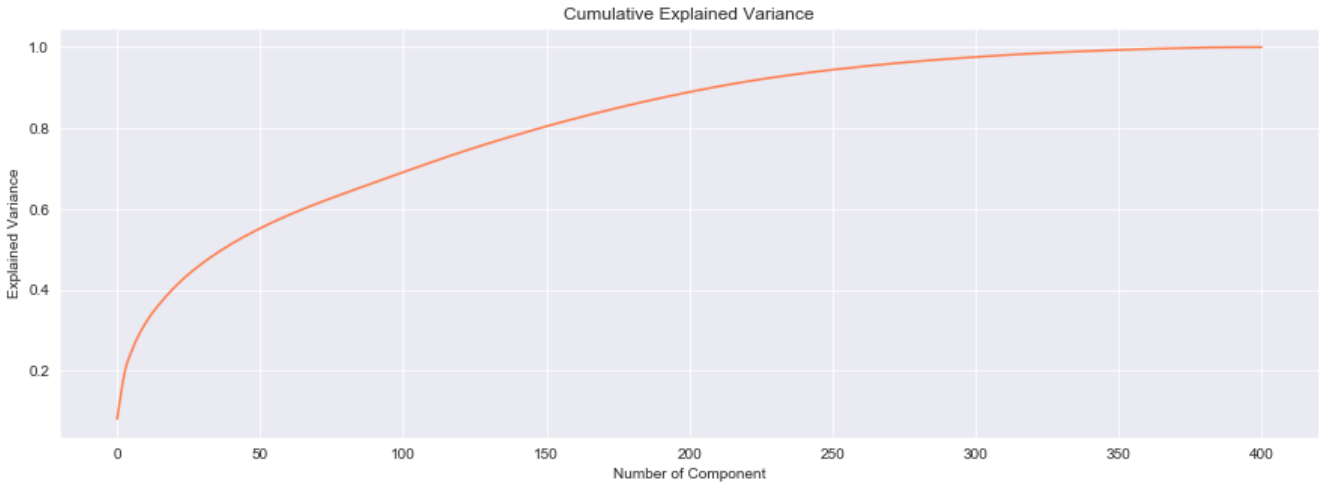
1. 通过数据探索了解到数据集中存在部分特征列的缺失值比例高，部分行的缺失值比例高，对于缺失值比例高于70%的特征列直接舍弃。
2. 通过数据属性 - 值解释文件（DIAS Attributes - Values），发现部分特征存在缺失值的同时又会使用0或-1或9来表示Unknown数据，这些值实际意义为缺失，统一处理。
3. 使用随机森林算法对数据集中的缺失值进行补全，方法如下：
    - 数据集中类型为object的纯数字特征列转化为Float，非数字的object类型特征进行独热编码。
    - 将特征列按缺失值数量从少到多排序。
    - 使用缺失值比例为0的特征列和RandomForest算法对含缺失值的特征进行补全（numeric类型使用RandomForestRegressor，object类型使用RandomForestClassifier）。

顾客分类

降维：

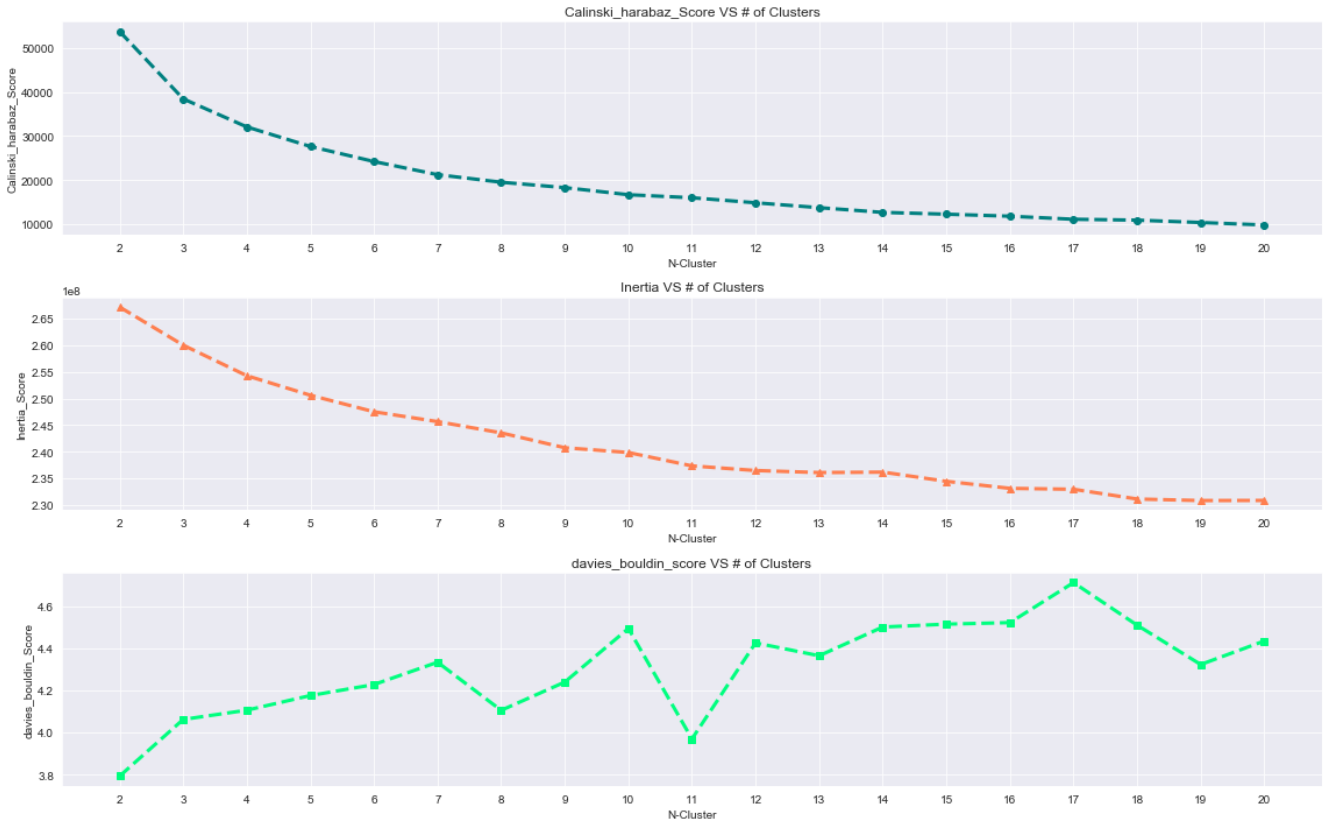
在使用PCA降维前，先使用StandardScaler将数据做归一化处理。

确定PCA降维后需要保留的维度数，如下图所示，208个主成分特征可以解释90%的方差。



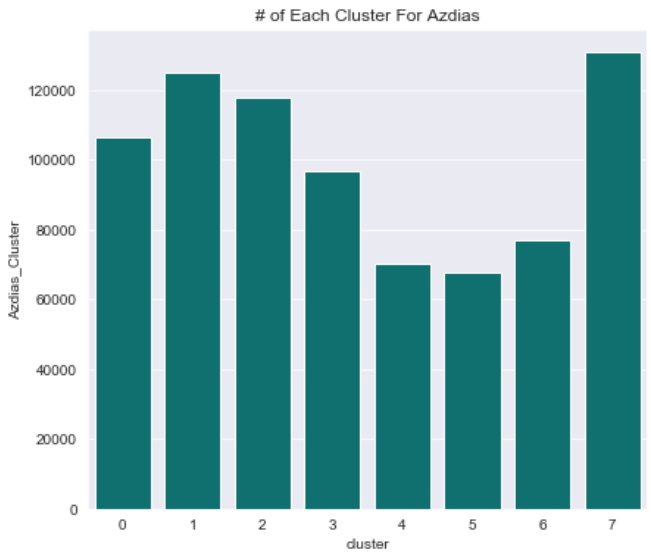
聚类:

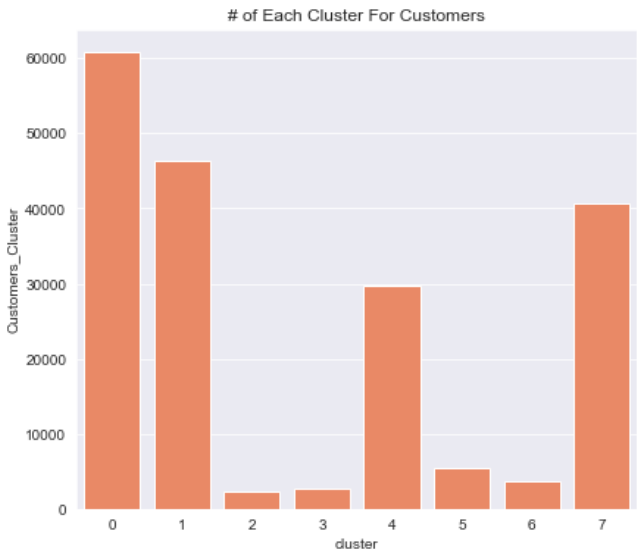
为确定合适的聚类类别数目，先使用MiniBatchKmeans算法快速计算各类别数的Calinski_Harabaz分数、Inertia分数和Davies_Bouldin分数，如下图所示，8或者11聚类效果相对较好。

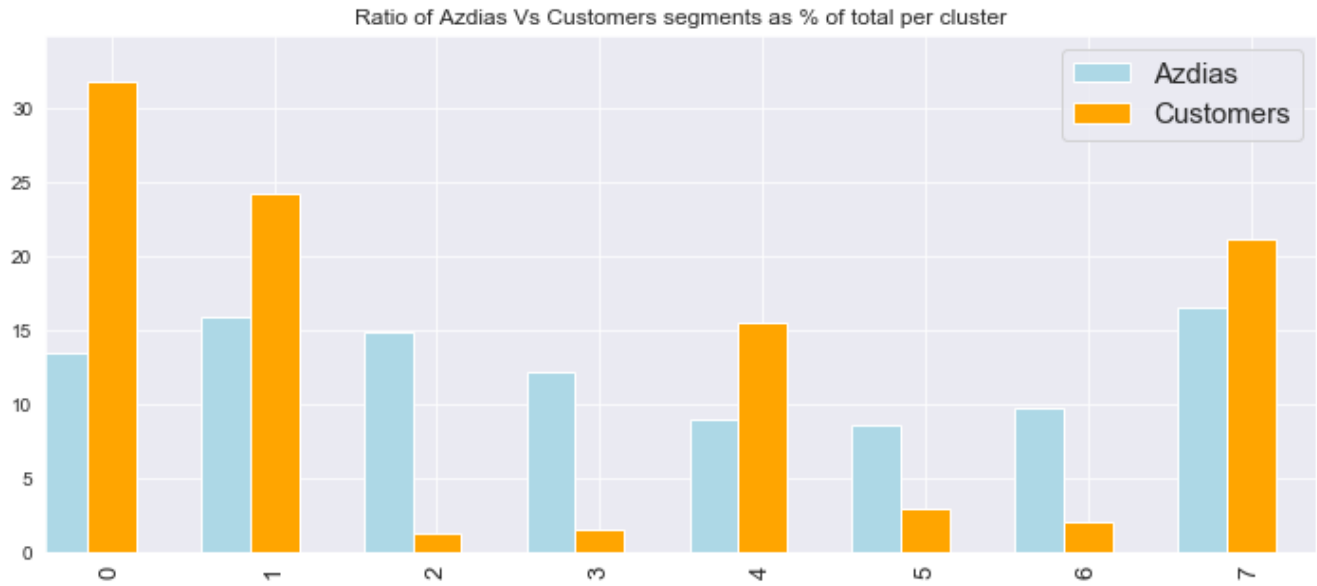


确定最终类别数目，使用Kmeans算法重新计算类别数为8和11的上述3个指标分数，最终确定类别数目为8。

使用上述StandardScaler、PCA和Kmeans方法拟合customers数据集。从下图可以看出customers数据集主要集中在类别0、1、4、7，且比例和azdias显著差异。





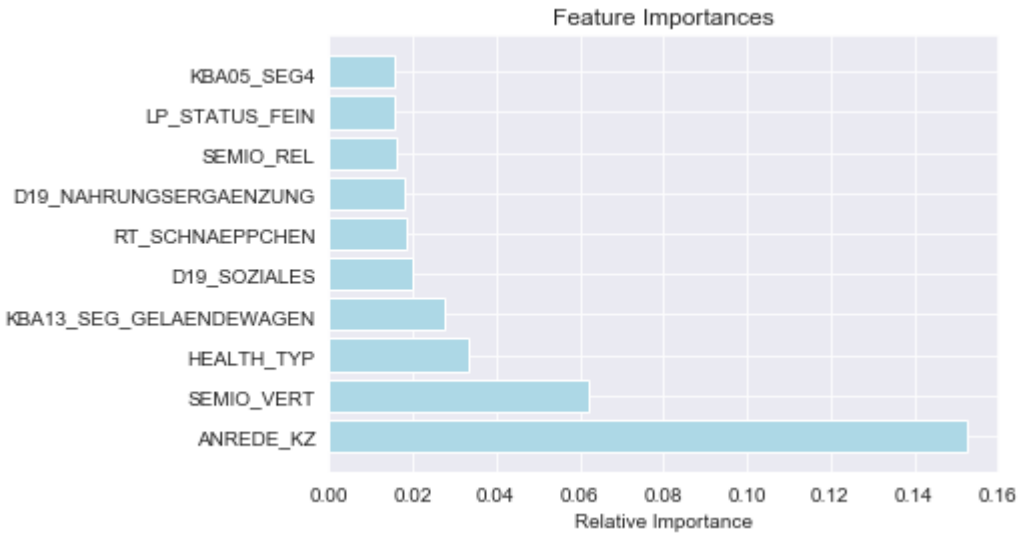


监督学习模型预测

完成顾客画像分群和分析之后，来到该项目的最后一部分，分析处理MAILOUT_TRAIN数据集，挖掘合适的算法来预测MAILOUT_TEST中的顾客是否会成为该公司的潜在顾客。在初次使用xgboost和randomforest模型对MAILOUT_TEST进行测试发现，该数据集存在明显的样本不平衡，正样本远小于负样本，三种解决方法1）使用Regressor回归预测；2）使用Classifier分类预测，输出概率值；3）使用Borderline-SMOTE算法over-sampling（过采样）来补全正样本，简易步骤如下。

1. 首先使用第一部分的随机森林缺失值填充法处理MAILOUT_TRAIN。
2. 复制数据，分为原始数据和使用Borderline-SMOTE算法过采样的数据。
3. 使用Xgboost和Randomforest（regressor和classifier），并对步骤2的两种数据集拟合。
4. Cross Validation评价、GridSearch寻找最佳参数。
5. 得到最好的模型，预测MAILOUT_TEST数据集。

xgboost得到的排名前10的特征：



模型比较和选择过程：

- 首先使用了Xgboost和RandomForest对清洗后的数据集进行了拟合，发现算法的性能较差；
- 使用交叉验证而不是直接将数据集分为training和testing，验证了算法的可行性，发现了样本存在较为严重的不均衡问题；
- 使用了BorderlineSMOTE和ADASYN两种过采样算法来平衡样本；
- 应问题本身为分类问题，所以尝试使用Classifier预测分类，发现算法在过采样后的训练数据集上过拟合，在测试集上表现严重低下；
- 改为使用Regressor回归器，预测成为潜在顾客的分数值，并运用了Voting的策略思想，结合Xgboost和RandomForest两种算法来提高泛化性能；
- 对Xgboost和RandomForest的树的数目、树的深度进行限制，xgboost中添加L1和L2正则化项，提高模型的泛化性能；
- 将最终模型运用在三种测试集上，1. BorderlineSMOTE算法过采样的数据；2. ADASYN算法过采样的数据；3. 未经采样的数据集。
- 采用auc分数来评价模型的性能，最初模型的性能几乎不好于纯随机预测，在不断迭代和试验后，模型性能得到了一个显著的改善。

Kaggle比赛

在整个模型使用和算法调参过程中，经过尝试，最终使用的模型为VotingRegressor包含了XgboostRegressor和RandomForestRegressor，相比于初始使用的VotingClassifier，预测能力有了显著的提升。

MAILOUT_TEST的预测结果上传至Kaggle，得到auc分数为0.73556。

结果讨论：

- 模型的预测分数不高，潜在原因可能是由于本数据集的特征数量较多，需要更深入地对各个特征做清洗、特征工程；
- 模型存在一定的过拟合，尤其是使用过采样后训练数据集的模型，通过优化过采样算法，或是直接使用原数据集而限制模型的复杂度，来降低过拟合问题，提高模型的泛化能力；
- 训练数据集的类别存在严重的不平衡问题，正样本严重过少，所以如果使用accuracy来评价模型会导致任何模型（纯随机预测）都会有高accuracy，因此recall、auc分数更适合来评价模型的性能；

未来方向：

- 进一步对数据集进行特征工程，本项目初步清洗了数据集，使用随机森林算法填充了缺失值，但因为特征较多，过多特征影响模型性能，参考xgboost和randomforest的特征权重排序来选择合适的特征；
- 使用集成学习的思想，在voting模型里尝试加入更多算法如svg、logisticregression等；
- 改善样本不平衡问题，尝试其他的过采样技术来解决样本不平衡问题。

结论

通过对Arvato Financial Services提供的德国一般人口数据和核心基础顾客数据的分析，完成了顾客画像的建立、分群，帮助公司能够识别顾客群；通过搭建监督学习模型，预测了潜在人群成为顾客的可能性。

项目趣点：

真实的数据集和完整的数据科学流程实践，在数据预处理和清洗环节，首次面对特征大、数据量大的数据集，一步步探索数据，挖掘数据中信息的过程是本项目的有趣点亦是挑战点。

代码改进：

优化代码结构，将代码写成函数或类

十分感谢由Arvato提供的真实数据集和Udacity提供的项目平台。