**金融服务公司实现一个顾客分类报告**

# 背景

该项目要求分析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。顾客数为191652，人口统计数为891221，自然情况下顾客的比例为21.5%，

分为两个步骤：

第一步，将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。

第二步，将识别的人群特征应用到第三个数据集上，该数据集是该公司的一场邮购活动的营销对象的人口统计数据，并且搭建的模型预测哪些人更可能成为该公司的顾客。

因为该数据集样本不平衡，因此评估采用ruc的值进行评估。

数据集的情况如下：

* Udacity\_AZDIAS\_052018.csv: 德国的一般人口统计数据；891211 人（行）x 366 个特征（列）
* Udacity\_CUSTOMERS\_052018.csv: 邮购公司顾客的人口统计数据；191652 人（行）x 369 个特征（列）
* Udacity\_MAILOUT\_052018\_TRAIN.csv: 营销活动的对象的人口统计数据；42982 人（行）x 367 个特征（列）
* Udacity\_MAILOUT\_052018\_TEST.csv: 营销活动的对象的人口统计数据；42833 人（行）x 366个特征（列）

# 变量分析

单变量分析的过程数据在excel文档《单变量分析过程中》<单变量分析的过程.xlsx>

## 人口属性

### 2.1.1 出生年份

结论如下：

* 出生在1920年-1939年的人70%以上成为了该公司的顾客
* 1965年以后出生的人成为该公司顾客的比例低于平均值较多
* 出生年份缺失值较多

### 2.2.2 金融类型：投资

结论如下：

* 金融投资非常高的人群顾客的比例比较高
* 金融投资从高到低顾客比例逐渐降低，非常低比例接近平均值

### 2.2.3 金融类型：低利率

结论如下：

* 金融类型：低利率非常低的人群顾客的比例比较高
* 金融类型：低利率从高到低顾客比例逐渐增加

### 2.2.4 贝隆最佳财务类型

结论如下：

* 省钱、投资者、准备好的几种类型成为该公司的顾客较多

### 2.2.5 度假习惯

结论如下：

* 追随自然环境、有黄金年龄度假习惯的用户顾客的占比高

### 2.2.6 亲和力量表：思想支配

结论如下：

* 中等偏高的认为亲和力是由思想支配的人群顾客占比高

### 2.2.7 青年时期主导运动等

结论如下：

* 三种类型的特别明显
* 另外未知的人群明显也容易成为顾客

### 2.2.8 其他结论

* HEALTH\_TYP未知的人群成为顾客比例高
* LP\_FAMILIE\_FEIN 中single和multi-generational household 顾客比例高。LP\_FAMILIE\_GROB包含的信息较LP\_FAMILIE\_FEIN少，所以删掉
* LP\_STATUS\_GROB生命周期来看，single high-income earner-couples，high-income earners of higher age from multiperson households顾客占比高
* LP\_STATUS\_FEIN中高收入，户主家庭顾客多占比高
* NATIONALITAET\_KZ德语的家庭顾客占比高

## 家庭属性

### 消费类型

结论如下：

* Versatile、Gourmet两个种消费类型的人群偏好该公司

### 家庭净收入

结论如下：

* 家庭的净收入越低顾客数越低

### 其他结论

* ANZ\_PERSONEN户口人数越多顾客比例越大，4人以上比较明显
* ANZ\_TITEL拥有学术头衔的顾客比例大
* D19\_GESAMT\_OFFLINE\_DATUM活跃的客户成为顾客比例大，交易金额越大，成为顾客的概率越高

## 客户类型

结论如下：

* CAMEO\_DEU\_2015中Successful Songwriter、Old Nobility、City Nobility三种顾客占比高
* CAMEO\_DEUG\_2015中upper class的顾客比率高

## 建筑物

结论如下：

* GEBAEUDETYP中residental building buildings without actually known household三种顾客占比高
* OST\_WEST\_KZ中East (GDR)的顾客比率低
* WOHNLAGE中very good neighbourhood顾客概率高very poor neighbourhood顾客概率 低

## 2.5 其他分类

Community各个指标顾客与人口统计属性差异不大

# 3. 模型构建

## 3.1 变量筛选

主要根据变量对顾客和人口统计整个属性中的差异程度和相关系数

### 3.1.1 顾客在人口统计中的差异程度

根据对比选择顾客数/人口统计数>2或者顾客数/人口统计数<0.5的变量

共174个标签，变量筛选分析过程.xls中<变量筛选分析过程.xlsx>详细记录了标签名，以及对应的顾客数/人口统计数的值

### 3.1.2 按照相关性系数删除部分变量

例如： 如下两组相关系数非常高，保留其中一组

A组：

KBA05\_HERST1

KBA05\_HERST2

KBA05\_HERST3

B组：

KBA05\_KRSHERST1

KBA05\_KRSHERST2

KBA05\_KRSHERST3

如下两组相关性也很高

C组：

D19\_GESAMT\_OFFLINE\_DATUM

D19\_GESAMT\_ONLINE\_QUOTE\_12

D组

D19\_VERSAND\_OFFLINE\_DATUM

D19\_VERSAND\_ONLINE\_QUOTE\_12

最终删除如下变量：

|  |
| --- |
| KBA13\_HALTER\_25 |
| KBA13\_HALTER\_30 |
| KBA13\_BMW |
| KBA13\_MERCEDES |
| PLZ8\_GBZ |
| PLZ8\_HHZ |
| KBA05\_HERST1 |
| KBA05\_HERST2 |
| KBA05\_HERST3 |
| KBA05\_KRSKLEIN |
| KBA05\_KRSOBER |
| KBA05\_KRSVAN |
| KBA05\_KRSZUL |
| KBA05\_MOD8 |
| D19\_GESAMT\_OFFLINE\_DATUM |
| D19\_GESAMT\_ONLINE\_QUOTE\_12 |
| FINANZ\_SPARER |
| FINANZ\_UNAUFFAELLIGER |

### 3.1.3 卡方检验



其中，A为实际值，T为理论值。

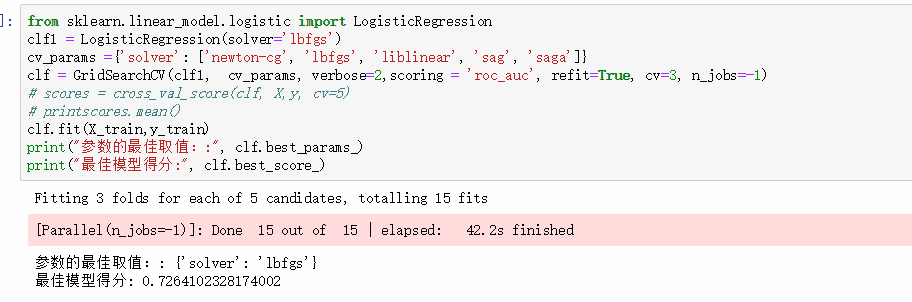
## 3.2数据预处理

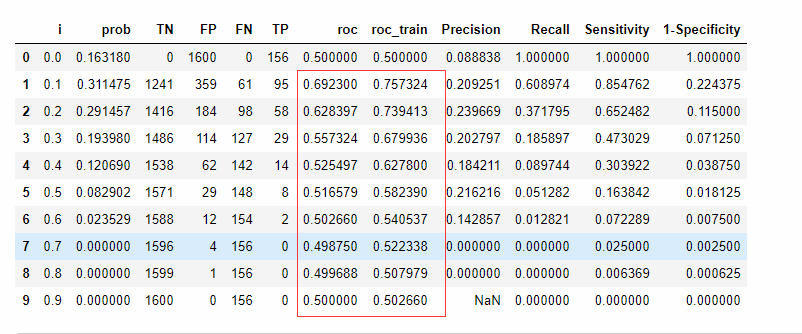
1. 在训练集中取全部正样本，共532条，负样本从42430中抽取5320条构成模型训练集
2. 将测试集也和模型训练集一起拼接，方便统一做one-hot编码
3. 离散变量做one-hot处理
4. EINGEFUEGT\_AM时间字段做拆分以及分段处理
5. 数值型变量合并
6. 空值用-1填充

## 3.2 模型选择

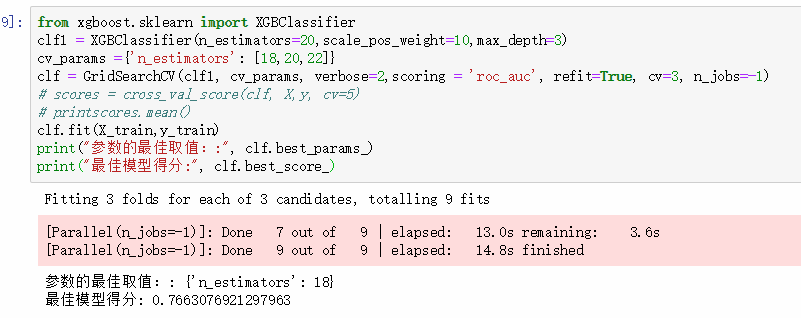
### 3.2.1逻辑回归模型

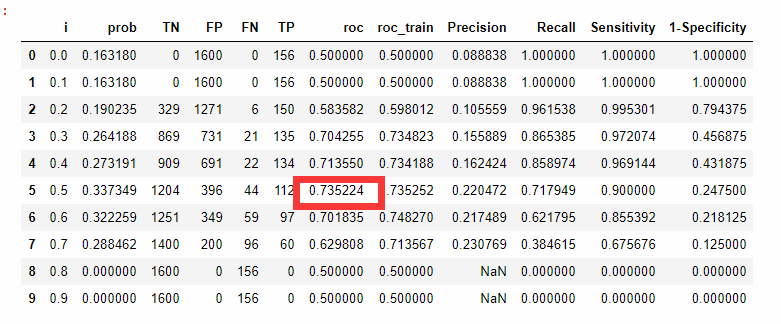
模型在训练集和测试集上面差异比较大，过拟合情况比较严重，变量比较多的情况下不适合用回归模型





### 3.2.2 XGBOOST模型





Xgboost是集成算法，明显比决策树的算法效果要好

## 3.3 模型调优

分别针对n\_estimators，max\_depth进行了调优，最终结合测试集合上面的roc的值选择n\_estimators=18，max\_depth=3

### 3.3.1 learning\_rate的调参

* 尝试了0.1，0.2，0.5



* 尝试了0.1，0.01，0.001



* 尝试了0.02,0.05,0.08,0.12，0.15 
* 尝试了0.06,0.07,0.08



### 3.3.2 n\_estimators的调参

在固定学习率的情况下对n\_estimators调参

* 尝试了10,20,50



* 尝试了15,18,20,22,25



### 3.3.3 max\_depth的调参

在固定了学习率于树个数情况下，对树深度调参

* 尝试了5,10,15



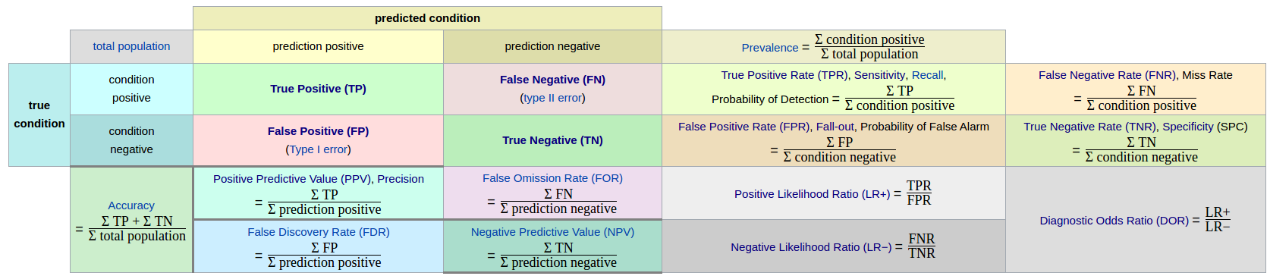
* 尝试了1,2,3,4



## 3.4 模型评估

### 3.4.1模型评估的理论

首先从混淆矩阵开始理解评估的各项指标如下图：



#### 3.4.1 Accuracy(准确率)

**准确率=(TP+TN)/(TP+TN+FP+FN)**

**当正负样本平衡，并且对于0，1判断正确与否都比较关心，那么可以用精准率。**

**当正负样本不平衡，尤其是正样本占比非常低且正样本的判断才是关注重点的时候，该方法不合适。**

**举个极端的例子，比如，正样本占比比较低，全部样本都标识为0，精准率还是比较高**

#### 3.4.2 Recall（召回率）

**召回率=TP/(TP+FN)**

**召回率表示的是预测正确的正样本占所有正确样本的比例，**反映**预测集中正确样本的覆盖所有正样本数情况。**

**该指标不受正负样本比例影响。**

#### 3.4.3 Precision（精准率）

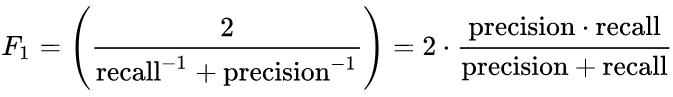
**精准率=TP/(TP+FP)**

精准率表示的是预测正确的正样本占预测中的所有样本的比例，反映的是预测集的正确率。

**该值会受到正样本占比的影响，过采样后会影响该结果。**

#### 3.4.4 F1 Score（F1值）

因为精准率和召回率都是跷跷板的两头，按照评分分档统计中，精准率提高，召回率就会下降。所以为了综合评判不同精准率和召回率情况下的模型效果，F1值是合适的选择。



因为F1值是由精准率和召回率综合的评分，过采样或者正负样本比例会影响值的高低。

#### 3.4.5 ROC曲线

ROC曲线中的主要两个指标就是真正率和假正率。其中横坐标为假正率（FPR），纵坐标为真正率（TPR）

* 真正率（TPR） = 灵敏度 = TP/(TP+FN)
* 假正率（FPR） = 1- 特异度 = FP/(FP+TN)



真正率和假正率是两个占比，并且分母为实际正样本和实际负样本，两个公式相对独立，所以不会随着正负样本比例变化而变化，对于样本不平衡问题，过采样样本数据仍然有效。

#### 3.4.6 RUC值

RUC为ROC曲线下的面积,对于样本不平衡问题没有影响

**AUC的一般判断标准**

**0.5 - 0.7：**效果较低，但用于预测股票已经很不错了

**0.7 - 0.85：**效果一般

**0.85 - 0.95：**效果很好

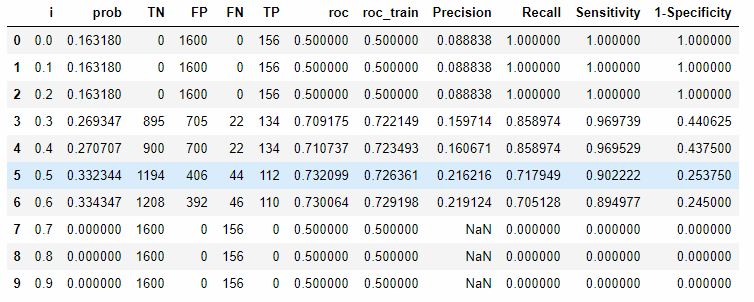
**0.95 - 1：**效果非常好，但一般不太可能

### 3.4.2模型评估的应用结论

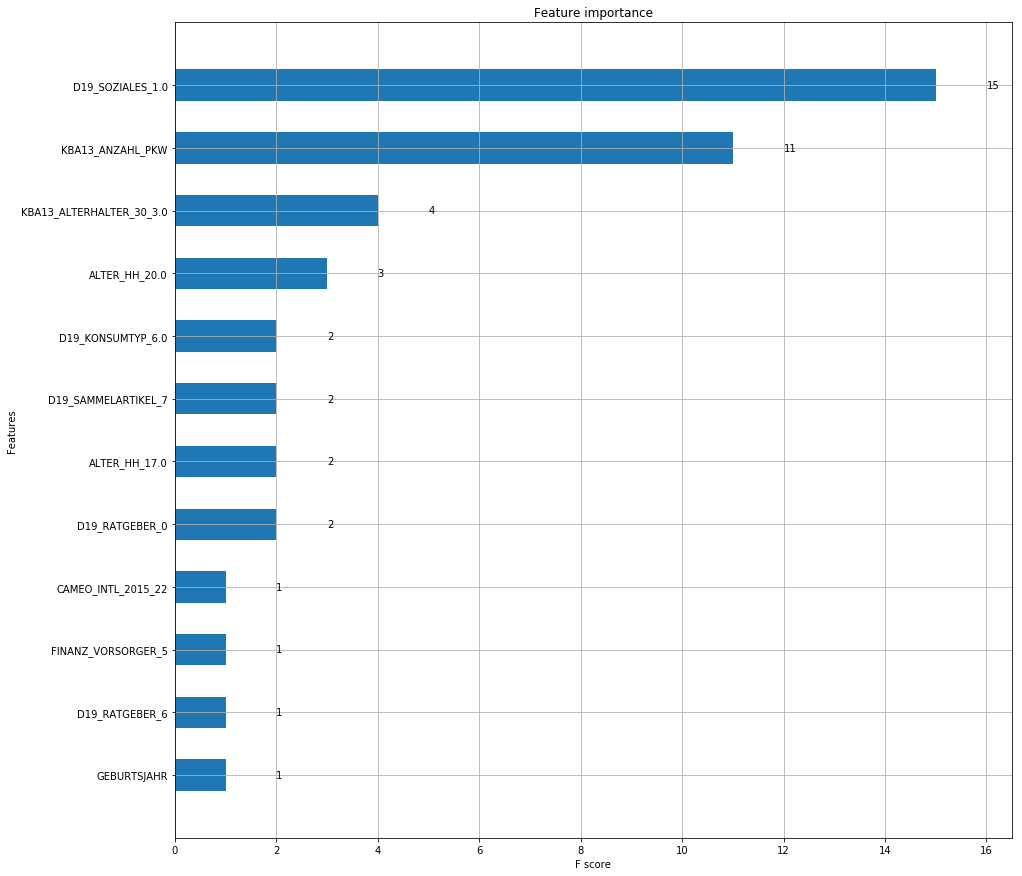
因为模型的样本不平衡且为了保证用过采样后的样本建立模型指标，可以估计在测试集上的评估指标，因此使用roc的曲线和RUC值进行评估。

当概率值大于0.5，ruc为0.73，模型效果一般。

在训练集和测试集的ruc基本一致，模型稳定性强



### 3.4.3 重要变量



# 模型预测

使用测试集进行预测，但是因为kaggle无法上传，所以看不到测试的结果

# 总结

## 4.1 过程总结反思

* 抽样分析

刚开始分析数据的时候量比较大，跑一晚上都跑不出来顾客与人口统计的结果。后来想数据量比较大，抽样应该对于比较相对结果没有太大影响，因此先进行了抽样。

后面分析的时候，只对重点做单变量可视化的变量全部数据统计

* 相关性分析

因为有价值的变量还是比较多的，为了删减变量，用相关性系统方法，相关性高的变量选择其中一个进入模型。

## 4.2 有趣的地方

* 变量非常多可以用的方法比较多，可以不断调优，精进模型开发的能力。
* 最重要的变量D19\_SOZIALES和网格属性有关。

## 4.3 几个优化的方向

1. 使用神经网络算法拟合
2. 按照顾客数据集进行聚类，形成的聚类群作为新变量纳入到预测模型中
3. 单变量的分析用模型的重要变量

# 参考文献

1. 用pandas进行数据分析实战

<https://mp.weixin.qq.com/s/RcrQmqty1FHEDbQfxv2XTQ>

1. 【集成学习】sklearn中xgboost模块的XGBClassifier函数

<https://www.cnblogs.com/wanglei5205/p/8579244.html>

1. <https://scikit-learn.org/>
2. ROC和RUC的使用

<https://segmentfault.com/a/1190000016686335>

<https://www.deeplearn.me/1522.html>