招商银行
CHINA MERCHANTS BANK

招商银行

信息技术助理实习汇报

《 在电话营销策略下的
存款客户预测问题研究 》

——以葡萄牙某银行机构为例

汇报人：康子浩

招商银行
CHINA MERCHANTS BANK

目 录
CONTENTS

招商銀行
CHINA MERCHANTS BANK

01 绪论

- ✓ 项目介绍
- ✓ 数据说明
- ✓ 描述性分析

招商银行
CHINA MERCHANTS BANK

项目介绍

电话营销模式在商业提升策略中占有极高的地位

问题：

过多的电话营销显然对客户更具侵入性

解决：

最大化客户生命周期价值来重新思考营销、评估可用信息和客户指标，从而根据业务需求建立更长，更紧密的关系。

招商银行
CHINA MERCHANTS BANK

数据说明

数据来源

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001

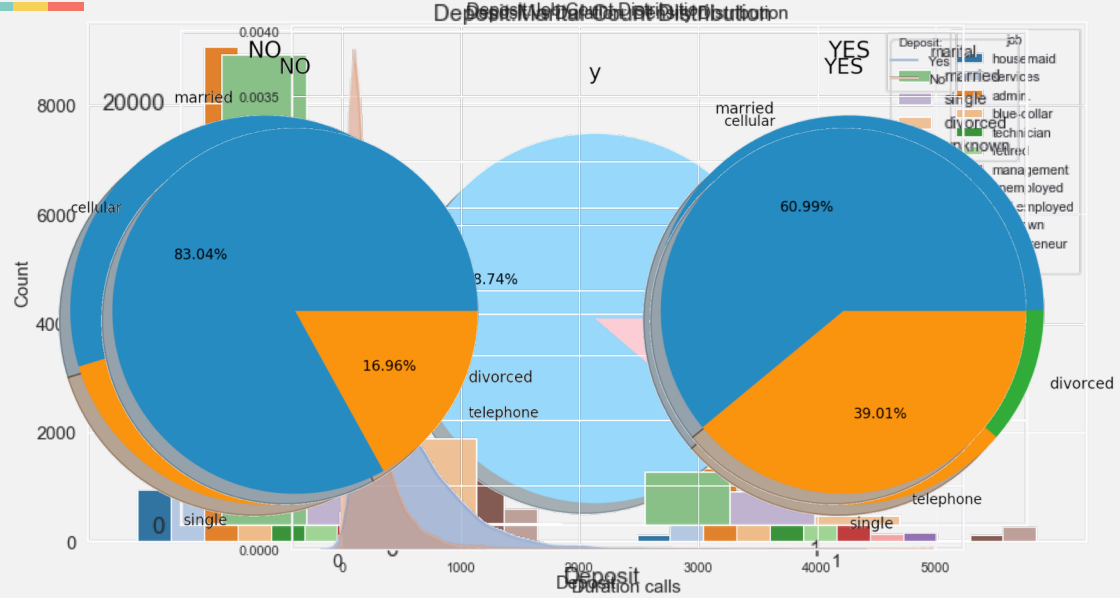
记录2008年至2013年的一份葡萄牙零售银行的数据（含金融危机）

数据说明 41189×21

bank client data:	related with the last contact of the current campaign:	other attributes:	social and economic context attributes:
1. age	1. contact	1. campaign	1. emp.var.rate
2. job	2. month	2. pdays	2. cons.price.idx
3. marital	3. day_of_week	3. previous	3. cons.conf.idx
4. education	4. duration	4. poutcome	4. euribor3m
5. default			5. nr.employed
6. housing			
7. loan			

招商銀行
CHINA MERCHANTS BANK

描述性分析



招商銀行
CHINA MERCHANTS BANK

02 数据预处理

- ✓ 缺失值处理
- ✓ 特征处理
- ✓ 数据无量纲化

招商银行
CHINA MERCHANTS BANK

缺失值处理

age	0	duration	0
job	330	campaign	0
marital	80	pdays	0
education	1731	previous	0
default	8597	poutcome	0
housing	990	emp.var.rate	0
loan	990	cons.price.idx	0
contact	0	cons.conf.idx	0
month	0	euribor3m	0
day_of_week	0	nr.employed	0

删除

此属性会严重影响输出目标（例如，如果持续时间= 0，则y =“否”）。然而，在执行呼叫之前不知道持续时间。此外，在通话结束后，显然已知y。

大多缺失数据在个人信息项目中

处理异常值

特征中的duration这一项目，根据实际情况分析对其进行删除

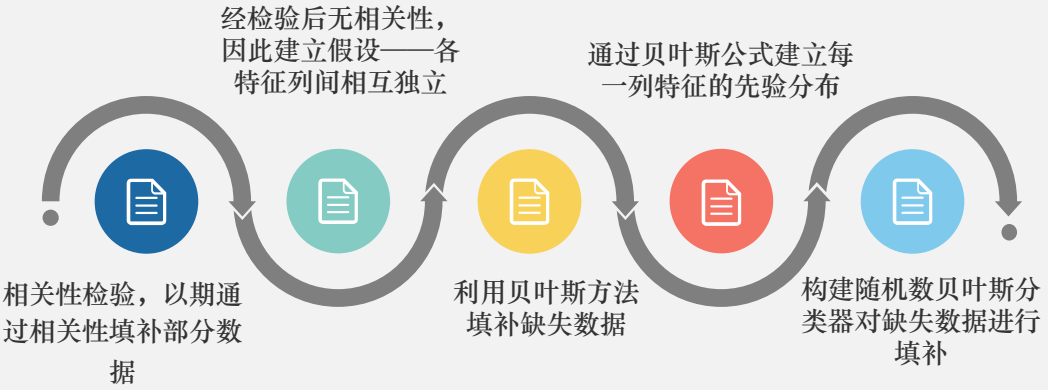
缺失值处理

- 直接删除
- 缺失值填补

1. 对缺失三项及以上的数据直接删除
2. 对仅缺少至多两项的数据进行填补

缺失值填补

$$P(A_i|B) = \frac{P(B|A_i)P(A_j)}{\sum_{j=1}^{n} P(B|A_i)P(A_j)}$$



利用了相关性不高的特点，建立了相互独立的假设，进而保证了各个特征在这样填补缺失值之后，各个类所占的比例保持不变。

招商银行
CHINA MERCHANTS BANK

特征处理

映射编码	独热编码
education	marital
有序变量	名义变量

illiterate	0
basic.4y	1
basic.6y	2
basic.9y	3
high.school	4
university.degree	5
professional.course	6

哑变量

divorced	1	0	0
married	0	1	0
single	0	0	1

招商银行
CHINA MERCHANTS BANK

数据无量纲化

当数据按照最小值中心化后,再按极差(最大值—最小值)缩放,数据移动了最小值个单位,并且会被收敛到[0,1]之间,而这个过程,就叫做数据归一化，归一化之后的数据服从正态分布,公式如下:

$$x^* = \frac{x-min}{max-min}$$

这个公式可以理解上面是中心化,下面就是缩放。

避免某个取值范围特别大的特征对之后的模型距离计算造成影响。因此采用数据归一化，将数据都压缩到[0,1]的区间内。

招商銀行
CHINA MERCHANTS BANK

03 模型处理

- ✓ 模型选择、调整
- ✓ 结果分析

招商银行
CHINA MERCHANTS BANK

模型选择、调整

- logistics
- KNN
- SVM支持向量机
- 决策树
- 随机森林
- 朴素贝叶斯
- XGBOOST
- GBDT 梯度提升法

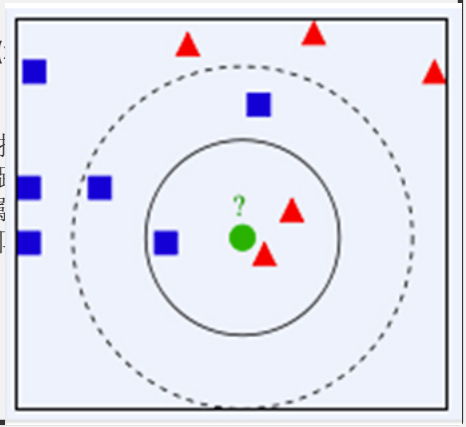
在分割数据为训练集和测试集之后，利用了各种模型对数据进行训练预测分类

KNN

Input:
A[n]为N个训练样本的分类特征；
k为近邻个数；

Initialize：
选择A[1]至A[k]作为x的初始近邻；
计算初始近邻与测试样本x间的距离d(x, A[i]), i=1,2,...k；
按d(x, A[i])从小到大排序；
计算最远样本与x间的距离D，即max{d(x, A[j]) | j=1,2...k}；
for(i=k+1; i<n+1; i++)
计算A[i]与x间的距离d(
if (d(x, A[i]) < D) then
用A[i]代替最远样本；
按照d(x, A[i])从小到大
计算最远样本与x间的
计算前k个样本A[i]所属
具有最大概率的类别即
end for

Output：
x所属的类别。



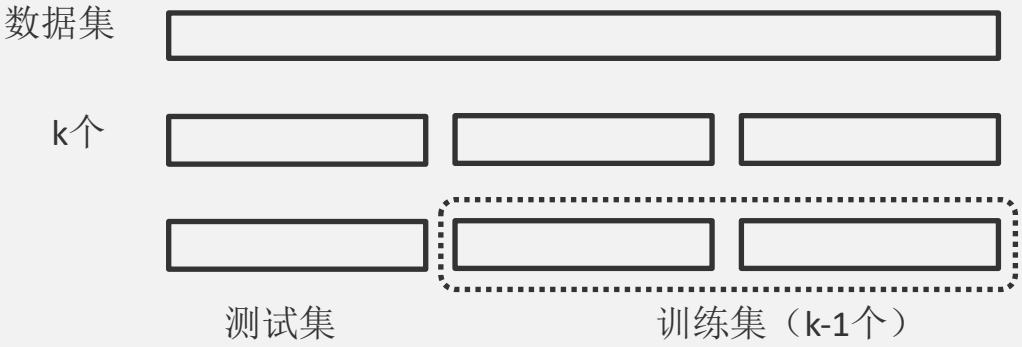
招商银行
CHINA MERCHANTS BANK

模型选择、调整

进行参数调整，改变训练集中标签权重，并写出各个模型的混淆矩阵

混淆矩阵		预测	
		0	1
实际	0	TN	FP
	1	FN	TP

利用k-fold法进行交叉验证



招商银行
CHINA MERCHANTS BANK

结果分析

混淆矩阵		预测	
		0	1
实际	0	TN	FP
	1	FN	TP

精确率、精度（Precision）

表示被分为正例的示例中实际为正例的比例。

$$P = \frac{TP}{TP + FP}$$

召回率（Recall）

召回率是覆盖面的度量，度量有多个正例被分为正例。

$$R = \frac{TP}{TP + FN}$$

综合评价指标（F-Score）

F-Score是Precision和Recall加权调和平均

$$F_\beta = \frac{(\beta^2 + 1)P \times R}{\beta^2(P + R)}$$

综合评价指标（F3-Score）

对Recall赋予3的权重，即$\beta = 3$，以提高指标对Recall的要求

招商銀行
CHINA MERCHANTS BANK

结果分析

各个模型的F_3得分如下

Models	Score
XGBoost	0.607845
Logistic Model	0.607659
Support Vector Machine	0.571648
Gausian NB	0.450851
Decision Tree Classifier	0.361905
Random Forest Classifier	0.323551
K-Near Neighbors	0.315939
Gradient Boosting	0.276854

招商銀行
CHINA MERCHANTS BANK

04 模型改进

- ✓ 改进理论
- ✓ 改进结果

改进理论

训练集中标签为1的样本过于少，训练效果不足，
因此此处我们采用了新的采样方法——SMOTEENN

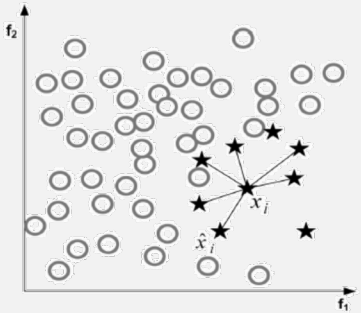
过采样SMOTE

(1)对于少数类中每一个样本x，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其k近邻。

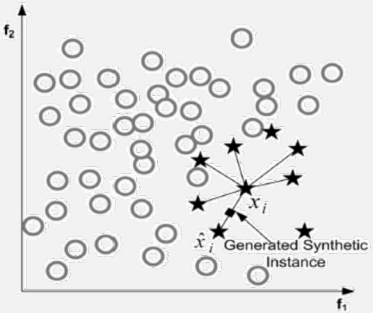
(2)根据样本不平衡比例设置一个采样比例以确定采样倍率N，对于每一个少数类样本x，从其k近邻中随机选择若干个样本，假设选择的近邻为x_n。

(3)对于每一个随机选出的近邻x_n，分别与原样本按照如下的公式构建新的样本。

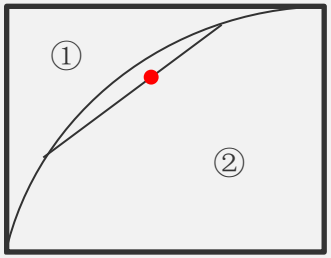
$$x_{new} = x + \text{rand}(0,1) \times (\tilde{x} - x)$$



(a)



(b)



!

招商銀行
CHINA MERCHANTS BANK

改进理论

因此要对该过采样进行组合改进
SMOTEEN就是SMOTE方法与EditedNearestNeighbours方法的组合

欠采样EditedNearestNeighbours

应用最近邻算法来编辑数据集, 找出那些与邻居不太友好的样本然后移除

筛选出的数据他们的绝大多数或者全部的近邻样本都属于同一个类

SMOTEENN

(1)利用SMOTE方法进行过采样
(2)利用EditedNearestNeighbours方法对过采样后的样本进行欠采样筛选

有效解决不平衡样本的问题

招商银行
CHINA MERCHANTS BANK

改进结果

经过重新采样后的训练集及测试集应用于模型中，对比所得F_3值有所提高

Models	Score
K-Near Neighbors	0.966060
Random Forest Classifier	0.932717
XGBoost	0.930494
Decision Tree Classifier	0.905915
Gradient Boosting	0.899433
Logistic Model	0.878718
Support Vector Machine	0.800988
Gausian NB	0.740555

招商銀行
CHINA MERCHANTS BANK

05 实习总结

招商银行
CHINA MERCHANTS BANK

实习总结

合作与分工

发现与解决问题

规划与完成目标

开阔眼界

学习、了解业务
知识及实用技能

招商銀行
CHINA MERCHANTS BANK

TAHNK YOU FOR WATCHING

汇报人：康子浩