客户流失

客户流失分析方法

按业务规则定义，经验模型

从数据中提取，机器学习

数据获取整理

数据分析与建模

制定挽留策略

实施挽留策略

应用

通信，商业，金融，服务

机器学习

监督学习

学习样本中有结果标记

无监督学习

学习样本中无结果标记

分类模型

决策树，有监督的学习

决策树就像是在回答一系列的问题，不同答案导向不同的分支，最终得到一个分类结果。

如何构建决策树

准备工作，明确自变量和因变量，确定信息度量的方式，确定终止条件

明确信息度量方式：信息增益

熵

·信息熵

描述混乱度

取值0~1

计算公式 -pi \*logPi

基尼系数

信息增益

信息是确定性的增加

从一个状态到另一个状态信息的变化

信息增益越大，对确定性贡献越大

明确分支种植条件

纯度

记录条数

循环次数

选择特征，计算所有特征信息度量，得到最佳分类特征

特征选择的时候选择信息增益量更大的一个

计算整体的熵

计算基于不同特征的熵

计算差选择信息增益更大的。

创建分支

是否终止

生成结果

决策树总结

ID3 迭代树3代

核心是信息熵

存在的问题

信息度量不合理：倾向于选择取值多的字段

输入类型单一：离散型

不做剪枝，容易过拟合

C4.5和ID3的相比

用信息增益率代替信息增益

能对连续属性进行离散化，对不完整数据进行处理

C50：c4.5

使用了Boosting

前修剪和后修剪

CART（Classification and Regression Tree)

核心是基尼系数

分类是二叉树

支持连续值和离散值

后剪枝进行修建

支持回归，可以预测连续值

集成学习:针对同一数据集，训练多种学习器，来解决同一问题

Bagging

有放回抽样构建多个子集

训练多个分类器

最终结果由各分类器结果投票得出

实现非常简单

Boosting

重复使用一类学习器来修改训练集

每次训练后根据结果调整样本的权重

每次学习器加权后的线性组合即为最终结果

Stacking

由两级组成，第一级为初级学习器，第二级学习器、

第一级的输出作为第二学习器的输入

随机森林(RandomForest)

由许多决策树组成，树生成采用了随机的方法

Smart Bagging

生成步骤

随机采用，生成多个样本集

对每个样本集构建决策树

优点

可以处理多分类

不会过拟合

容易实现并行

对数据集容错能力强

特征工程

列

数据预处理：标准化，缩放，缺失值，变换，编码

特征产生：结合业务数据，派生新的特征

特征选择：通过各种统计量，模型评分等，筛选合适的特征

降维：PCA ,LDA等减少特征。

模型评估：解释，泛化

换数据集后，表现非常差，就叫泛化性差，过拟合

CRISP-DM流程

商业理解

数据理解

数据准备

建立模型

模型评估

方案实施