基础建模

案例：房价预测——找到特征（小区地点，配套，交通）；预测目标（房价）

案例：魔镜系统

**什么是分析建模：**现实问题抽象化用数学模型表示

**建模的四大步骤：**

样本定义：训练样本如何选取，target如何定义

特征工程：生成模型需要的特征

模型训练：把数据整理好喂给模型训练

效果评估：根据结果进行优化

1. 样本定义

* 需要反映预测数据集

首先明确模型的使用场景，在什么样本上进行预测

训练样本需要按照预测样本来定义，尽量保持一致

（eg：预测新进来的app新客风险情况，用历史的app新客数据进行训练是不可以的，因为历史数据只有部分通过用户的数据，未通过的用户抓不到他们的数据，因此不能反映全量）

* 分层抽样
* 预测目标

连续的数值（房价）——回归

类别（魔镜评分）——有标签（分类）；无标签（聚类）

* 选择合适的预测目标

y的定义：业务模型的y多与事件相关，所以时间窗要明确

eg：与时间无关，有客观定义：人脸识别

与时间有关，有客观定义：某段时间用户购买意愿

与时间有关，无客观定义：客户流失

1. 特征工程

* 特征工程是什么：

数据和特征决定了机器学习的上限，模型和算法只是逼近这个上限

* 特征主题分类：

强业务数据：交易/持有/收益表现

弱业务数据：接触信息（电话）/浏览信息（app登陆）/用户权益/市场环境

业务无关数据：用户属性/设备信息/三方数据

* 生成特征：

**特征主题——业务行为——生成特征（量化）**

主动接触——拨打客服电话——xx时间拨打客服电话次数/频率/时长/内容

被动接触——营销电话——xx时间收到营销电话次数/转化率

* 衍生特征：

原始特征：年龄

衍生特征：年龄是否大于30岁，log（年龄）等等

eg：14天内购买和，30天最大登陆时间，短信分，注册渠道；职业；省份WOE*（WOE：当前分组中响应客户占所有响应客户的比例和当前分组中没有响应的客户占所有没有响应的客户的比例）***（类别型变量变成数字型）**

* 数据处理：

标准化，归一化，二值化，分类编码，缺失值插补，多项式变换

* 特征选择：

**单变量分析：**看这个变量在不同取值上的target rate区分度怎么样

用变量IV值（信息价值指标）衡量（越高说明贡献度越高）

0.1-0.3（预测能力一般）；0.3-0.5（预测能力强）；>0.5（可疑）

eg1：看性别的特征对最后模型的预测强不强

表格中展示的性别男女区分度比较大，但不能只看购买率，还要看样本分布的百分比，要算变量的IV值=0.28

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 性别 | 数量 | 百分比 | 购买 | 购买率 |
| 男 | 600 | 60% | 300 | 50% |
| 女 | 300 | 30% | 100 | 33% |
| 缺失 | 100 | 10% | 20 | 20% |
| 总计 | 1000 | 100% | 420 | 42% |

**相关性分析：**对两变量计算相关系数，系数高的变量去掉一个（最好不要超过0.5；去掉IV值小的/业务解释不合理的/计算复杂度大的）

* 生成表（样本），扔给模型

样本粒度：建在标上的/人上的模型

特征：x1，x2

标签

1. 模型训练

* 样本切分：为了同时提高模型的准确率和泛化能力

训练集（IS）：用于训练模型

验证集（OOS）：数据对象与训练集相同，用于模型效果评估和调优

测试集（OOT）：数据对象与训练集不同，仅用于模型效果评估

* 模型是怎么学习的：

模型训练的目的：

：预测值

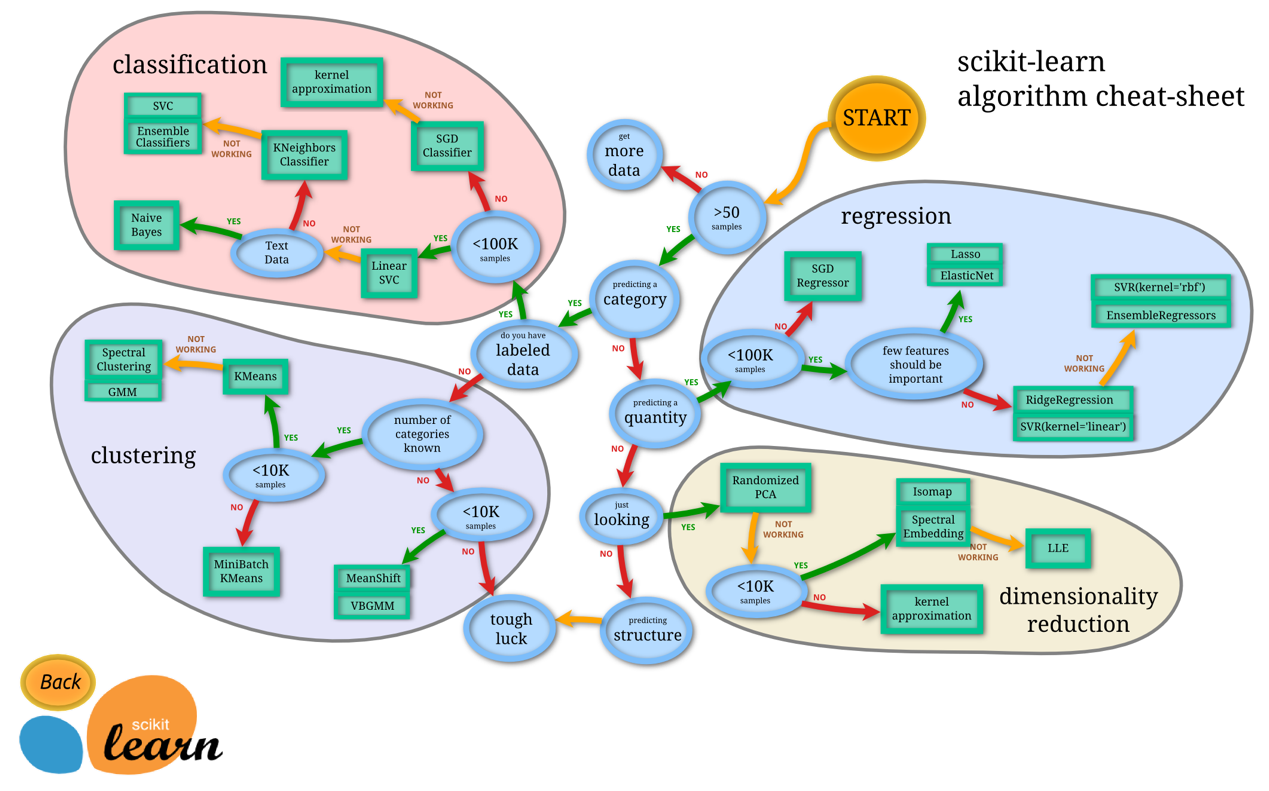
：不同模型

：样本的特征们

：模型参数，模型需要算出来的东西

：模型超参数，模型本身的东西，需要自己指定和调优

* 模型选择：



1. 效果评估

* 评估指标：

模型效果评估：预测值 VS 真实值

准确率：预测正确的样本量/总样本量（不怎么看）

精确率（precision）：预测对的/预测的总体

召回率（recall）：预测对的/所有样本中对的

（追求召回率高，精确率就会低，反之亦然）

设定明确的阈值：eg模型分切50%阈值，再计算precision和recall

AUC：area under the ROC curve （0.5-1，0.5为极差）

对应AUC更大的分类器效果更好

KS：Kolmogorov-Smirnov （>0.2可认为预测能力好）

值越大，表示模型能够将正、负客户区分开的程度越大

* 过拟合：

需要在训练集/验证集/测试集上进行评估，比较效果

|  |  |  |  |
| --- | --- | --- | --- |
|  | IS | OOS | OOT |
| AUC | 0.85 | 0.72 | 0.37 |
| KS | 0.54 | 0.39 | 0.23 |

说明模型有衰减，泛化能力差

解决方法：丰富样本；减少选取特征的数量；模型调参

一般15万样本，10-15个特征

正负样本不平衡的问题：10%-20%左右可以直接训练；如果非常不平衡1%，把0的样本人为丢掉一些；在模型里设置损失函数（lost），真实值-预测值的平方和越小越好，如果正样本判错给惩罚