拍拍贷 ppdai.com 触手可及的金融

基础建模之菜鸟起飞

钟海/温舒

革故鼎新 | 追求卓越 | 简单靠谱 | 彼此成就

你能学到什么

什么是数据分析建模，可以用来干嘛

建模的基本套路，帮助大家实现起飞

案例：房价预测

讨论：猜猜这个小区的房价？

小区的房价：**40,000/平米**

所在城市：上海

所在区域：浦东新区唐镇

建造年代：2013年

离地铁站距离：2公里

生活配套：家乐福，购物中心，华夏公园，唐镇社区中心

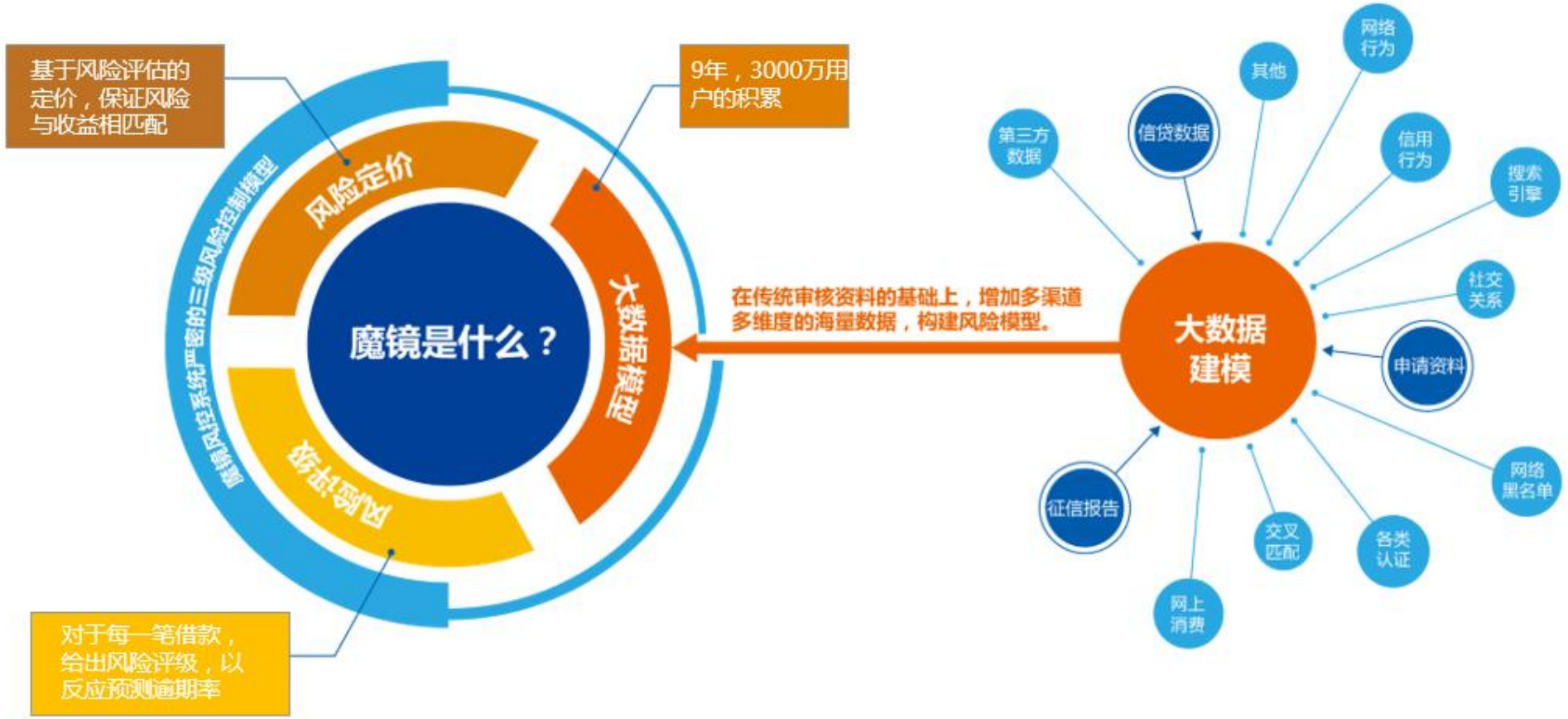
学区情况：唐镇小学

开发商：唐城投资公司

容积率：1.8

绿化率：36%

案例：魔镜系统


基于风险评估的定价，保证风险与收益相匹配
9年，3000万用户的积累
魔镜风控系统严密的三级风险控制模型
风险定价
大数据模型
风险评级
魔镜是什么？
对于每一笔借款，给出风险评级，以反应预测逾期率
在传统审核资料的基础上，增加多渠道多维度的海量数据，构建风险模型。
大数据建模
第三方数据
信贷数据
其他
网络行为
信用行为
搜索引擎
社交关系
申请资料
网络黑名单
各类认证
交叉匹配
网上消费
征信报告

什么是分析建模

建模是指把具体问题抽象成为某一类问题并用数学模型表示，是应用于工程、科学等各方面的通用方法，是一种对现实世界的抽象总结。

特征

X

所在区域，生活配套，
交通情况，等

预测目标

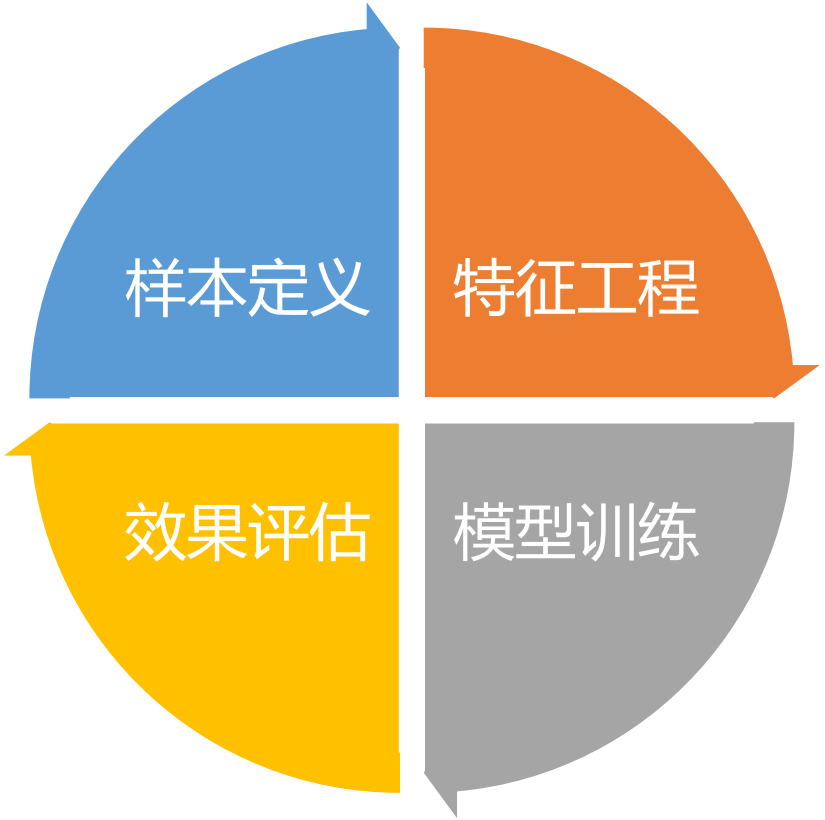
Y

小区房价

建模四大步骤

1

训练样本如何选取，Target如何定义



2

生成模型需要的特征

把数据整理好喂给模型进行训练

3

评估模型效果，根据效果进行优化

4

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，即是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。

Choose a dev set and test set to reflect data you expert to get in the future and consider important to do well on.

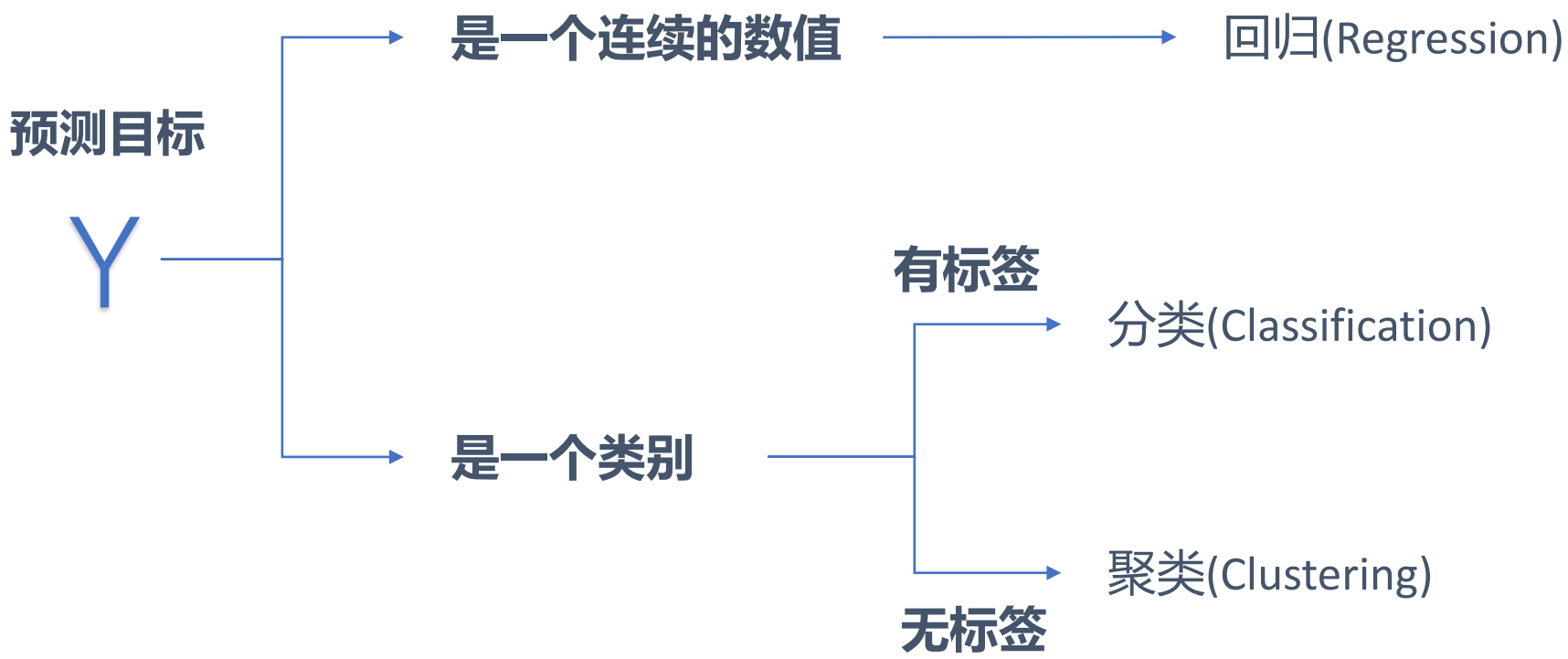
--- Andrew Ng

分层抽样

是从一个可以分成不同子总体（或称为层）的总体中，按规定的比例从不同层中随机抽取样品（个体）的方法。

优点：样本的代表性比较好，抽样误差比较小。

预测目标Y


预测目标
Y
是一个连续的数值
回归(Regression)
是一个类别
有标签
分类(Classification)
聚类(Clustering)
无标签

选择合适的预测目标

Y的定义：业务模型的Y多数都跟时间相关，所以时间窗要明确。

与时间无关，有客观定义：人脸识别

与时间有关，有客观定义：某段时间的购买

与时间有关，无客观定义：客户流失

2 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，而模型和算法只是逼近这个上限而已。”

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work.

from Wikipedia

特征主题分类

交易/持有/收益表现
Transaction/Position/Performance

接触信息
Contact

浏览信息
Visit

用户权益
User Rights

市场环境
Market

用户属性
User Attributes

设备信息
Device Info

三方数据
Third-party

强业务数据

弱业务数据

业务无关数据

生成特征

特征主题	业务行为	生成特征
接触信息 Contact	拨打客服电话	过去1个月拨打客服电话次数
	留言咨询客服	是否在论坛发过帖吐槽拍拍贷
	在论坛上发帖/回帖	
主动接触		
被动接触	营销电话	过去3个月是否收到过营销电话或短信
	营销短信	注册后被成功营销总次数
	营销优惠券	

衍生特征

原始特征：年龄

衍生特征：年龄是否大于30岁，年龄是否大于40岁，年龄除以10再取整， log(年龄)

衍生特征的其他例子：

14天内购买和	吐槽拍拍贷次数	注册渠道WOE
30天最大登录时间	是否是高净值用户	职业WOE
首付款比例	短信分	省份WOE
相邻月的购买差		

WOE的全称是“Weight of Evidence”，即证据权重。WOE是对原始自变量的一种编码形式。表示的实际上是“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”的差异。

数据处理

整个分析建模套路中，特征工程应该是最耗时的一步，而数据处理则是特征工程中最基础的一步。

实用类	功能	简介
StandardScaler	标准化	均值-标准差化数据标准化
MinMaxScaler	标准化	极值化法数据标准化
Normalizer	归一化	行记录单位化
Binarizer	二值化	连续变量离散化
OneHotEncoder	分类编码	将定性数据编码为定量数据
Imputer	缺失值插补	缺失值插补
PolynomialFeatures	多项式变换	多项式数据变换

特征选择

单变量分析：

我们只关注1个变量（特征），看看这个变量在不同取值上的 target rate 区分度怎么样。

性别	数量	百分比	购买	购买率
男	500	50%	250	50%
女	400	40%	100	25%
缺失	100	10%	20	20%
总计	1,000	100%	370	37%

变量IV值:0.33

变量的IV值

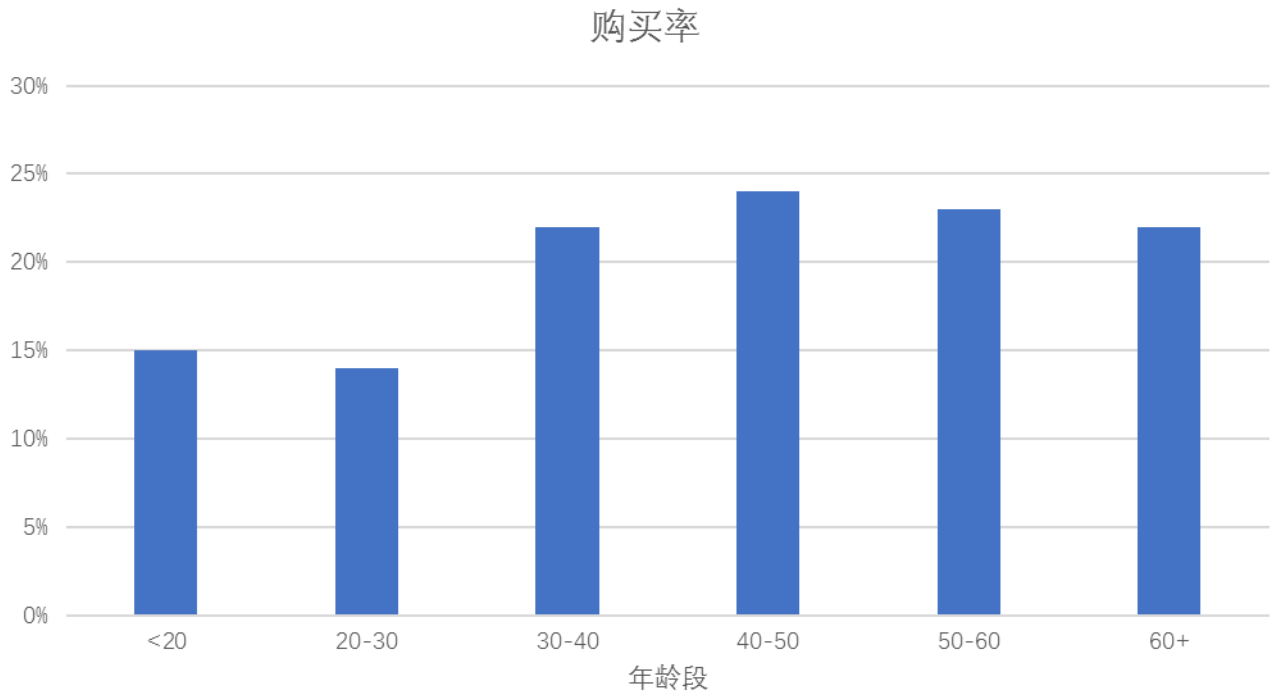
IV值（Information Value），即信息价值指标，可以体现变量对于一个二分类问题的重要程度，一般用做特征选择。

IV值	预测能力
< 0.02	无预测能力
0.02 ~ 0.1	较弱的预测能力
** 0.1 ~ 0.3 **	** 预测能力一般 **
** 0.3 ~0.5 **	** 较强的预测能力 **
> 0.5	可疑

特征选择的例子

原始特征：年龄

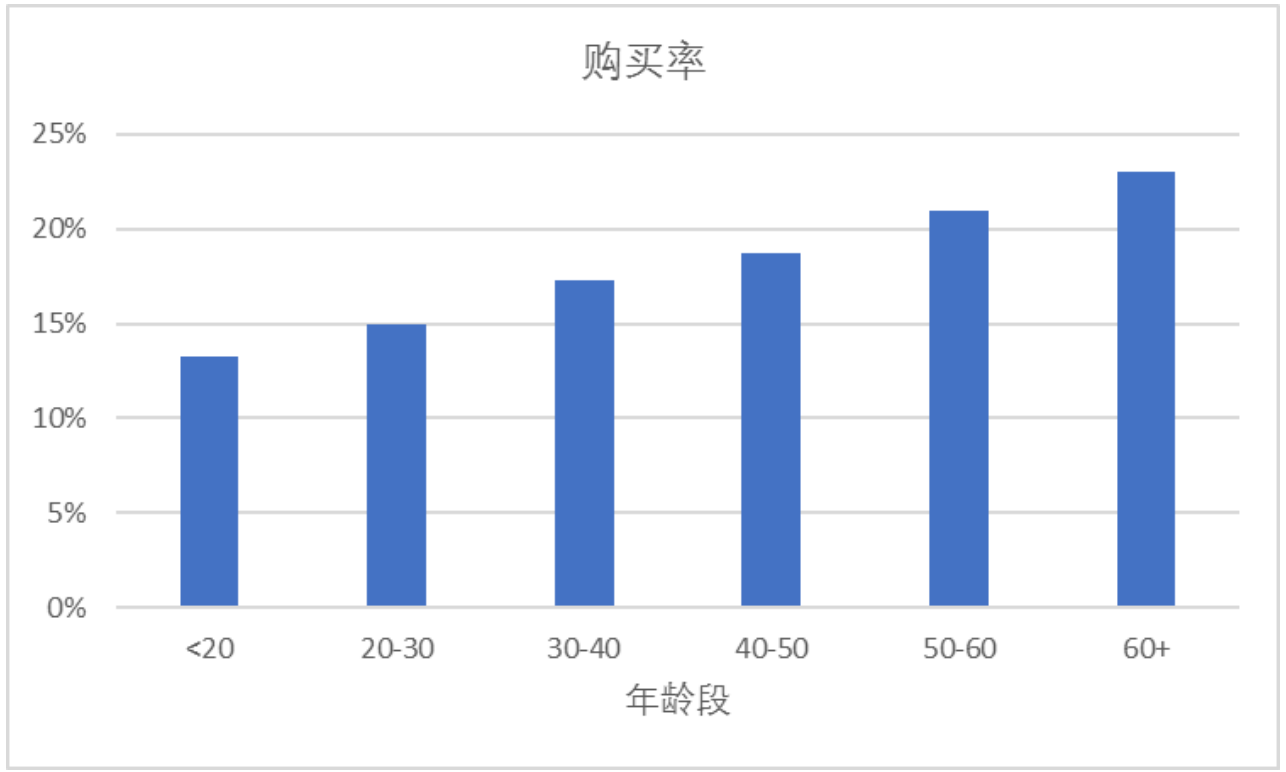
衍生特征：年龄是否大于30岁，年龄是否大于40岁，年龄除以10再取整， log(年龄)



特征选择的例子

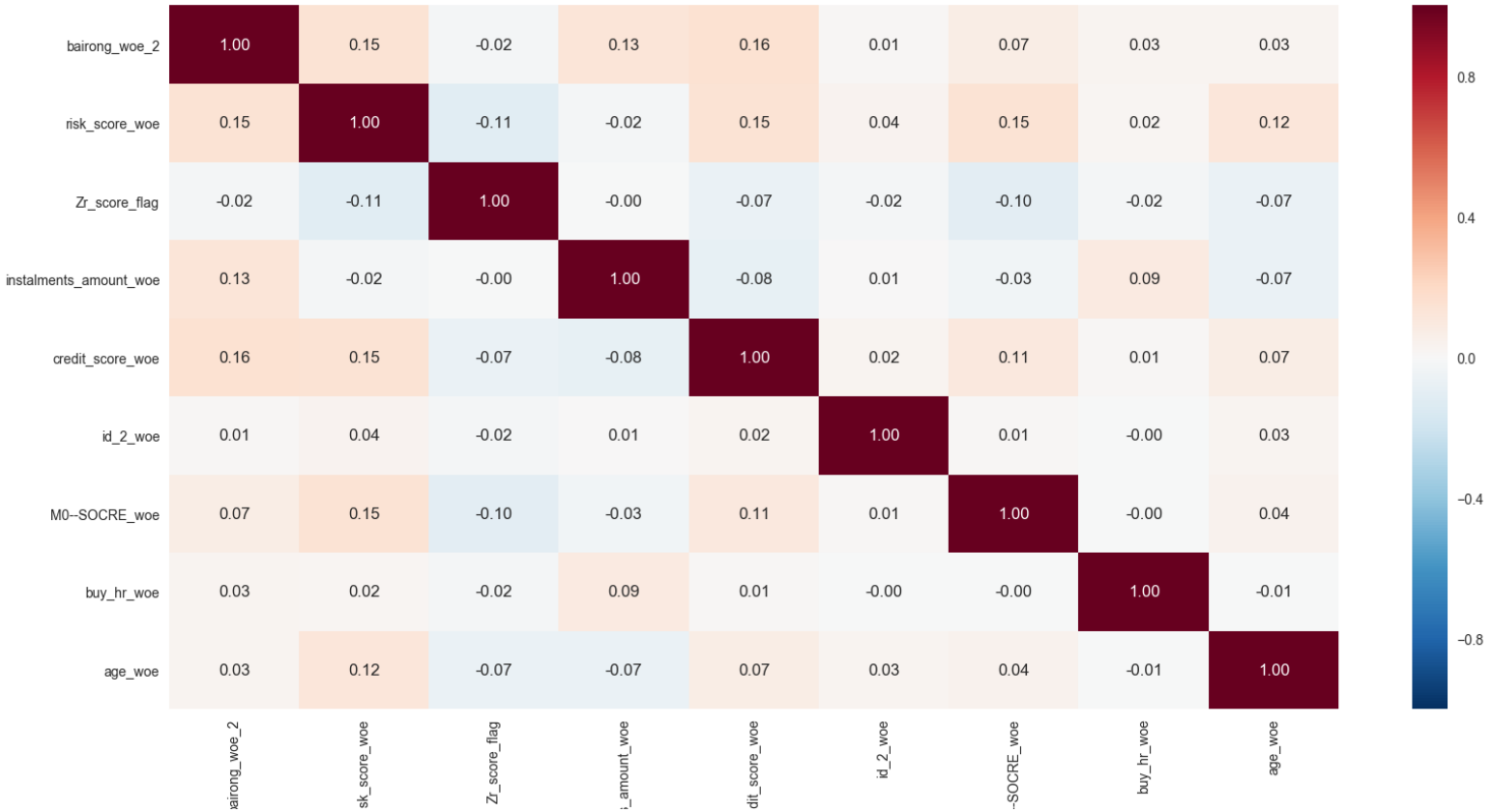
原始特征：年龄

衍生特征：年龄是否大于30岁，年龄是否大于40岁，年龄除以10再取整， log(年龄)



相关性分析

对两两变量计算相关系数，系数高的变量对需要去掉1个。



建议：

1. 相关系数最好不要超过0.5
2. 去掉哪个？
   a) IV值小的
   b) 业务解释不合理的
   c) 计算复杂度大的

最后一步

生成一张大表，
扔给模型。

样本粒度　　特征们　　标签

样本

ID	Time	X1	X2	…	Y
1	20180302	男	22	…	1
2	20180305	女	24	…	0
3	20180401	Na	25	…	0
4	20180402	女	Na	…	1
5	20180402	男	43	…	0
6	20180501	男	26	…	1
7	20180502	男	Na	…	0
8	20180606	男	23	…	0
…	…	…	…	…	…

3 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**，需要对模型样本进行切分。

训练集（training dataset）—— 用于训练模型

验证集（validation dataset）—— 数据对象与训练集相同，用于模型效果评估和调优

测试集（test dataset）—— 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：找X（特征）和Y（预测目标）的关系，即 $y = f(x)$

不同的f就代表了不同的模型

模型参数，模型需要算出来的东西，为了使预测值与真实值尽量一致

$$y = f(x, w, w_h)$$

预测值

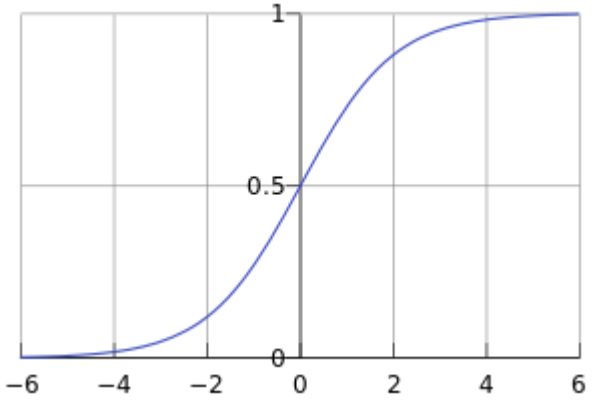
样本的特征们

模型超参数，是模型本身的东西，与样本无关，需要自己指定和调优

模型例子：逻辑回归

一个非常经典的针对二分类问题的模型，使用场景广泛。

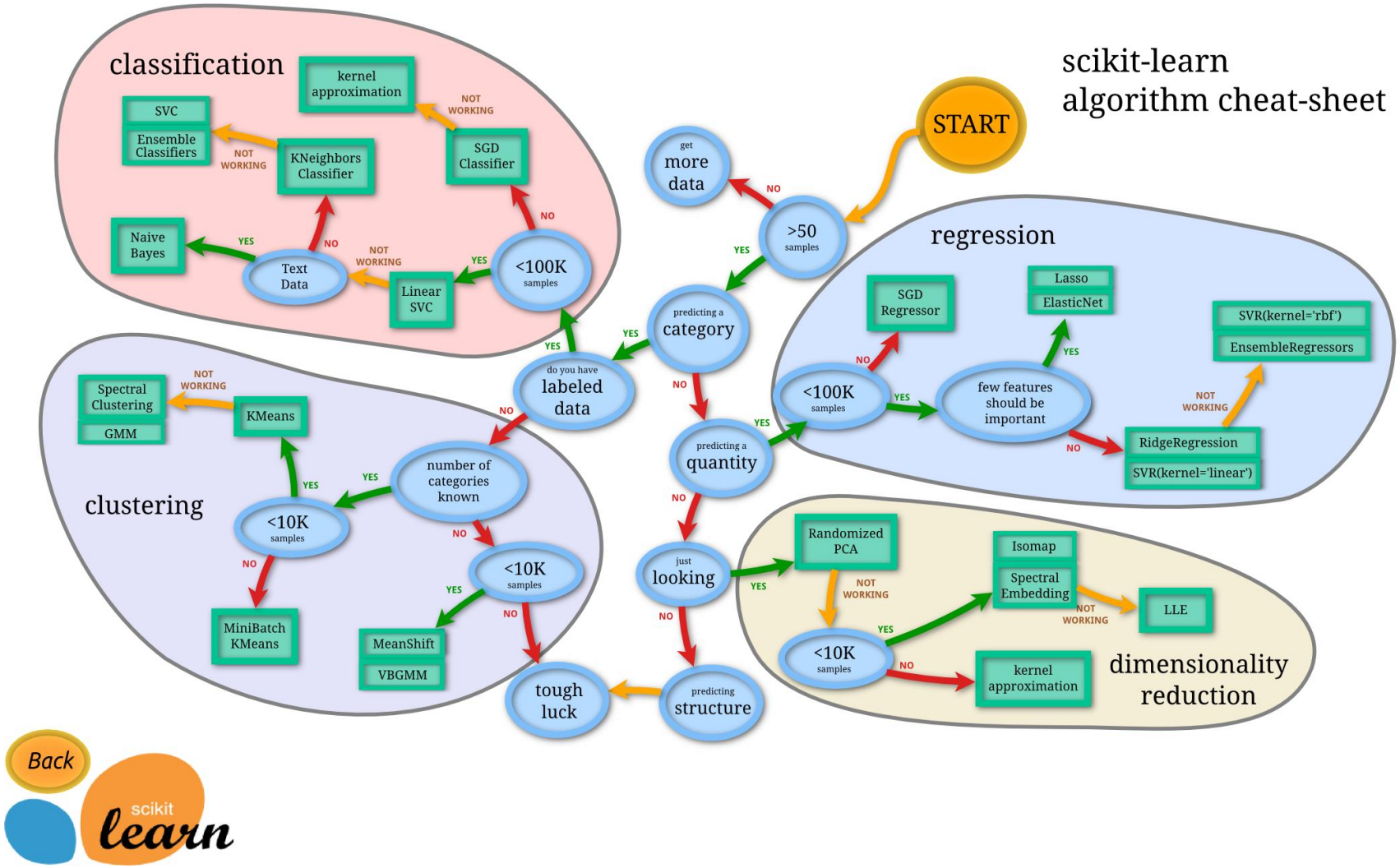
$$y = f(x, w, w_h) = \frac{1}{1 + e^{-w^T x}}$$



预测值 y 的取值范围为0到1。

假设我们问题是抓坏人，越接近1代表我们预测这个人越像坏人。

模型选择


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
predicting a category
YES
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
RidgeRegression
SVR(kernel='linear')
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

4 效果评估

评估指标（一）

模型效果评估：**预测值** VS **真实值**

准确率（Accuracy）： 预测正确的样本量 / 总样本量

评估指标（二）

精确率（Precision）

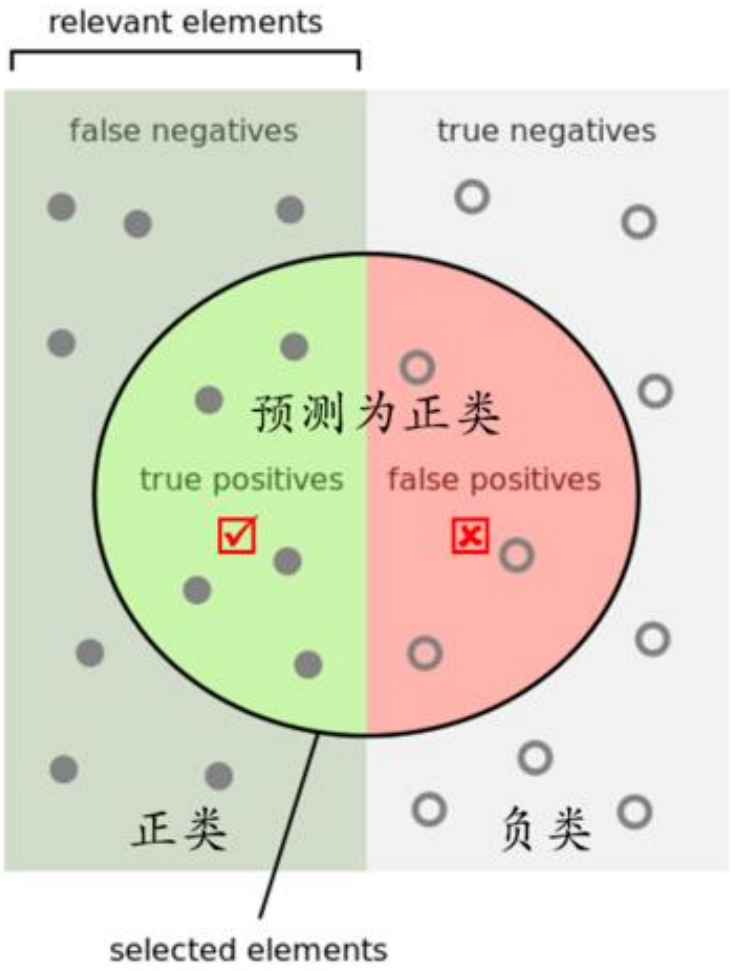
召回率（Recall）

How many selected items are relevant?

Precision = 

How many relevant items are selected?

Recall = 



评估指标（三）

分类模型产生的结果通常是一个概率值不是直接的0/1分类（比如逻辑回归）。

如果能设定明确的阈值，可以使用Precision和Recall进行模型评估，否则需要新的评估指标。

AUC (Area Under the ROC Curve)

AUC取值一般在0.5到1之间，值越大模型效果越好。AUC=0.5说明模型的预测能力与随机结果没有差别。

KS (Kolmogorov-Smirnov)

KS值表示了模型将正负样本区分开来的能力。值越大，模型的预测准确性越好。一般来讲，KS>0.2即可认为模型有比较好的预测准确性。

过拟合（ overfitting ）

过拟合（overfitting）

过拟合（overfitting）

需要在训练集/验证集/测试集上进行评估，比较效果。

METRIC	IS	OOS	OOT
AUC	0.85	0.72	**0.37**
KS	0.54	0.39	**0.23**

METRIC	IS	OOS	OOT
AUC	0.78	0.76	**0.72**
KS	0.45	0.44	**0.41**

解决方法举例：

丰富样本；减少选取特征的数量；模型调参（超参数w_h）。

第一届魔镜杯排行榜

初赛排行榜

排名	团队	最终分数	提交次数	最后提交时间
1	Deadshot	0.780529	14	16-03-31 23:39
2	涌泉	0.780442	9	16-03-30 12:40
3	三湖连江数据分析团队	0.779833	7	16-03-24 22:51
4	全民狙击	0.776913	12	16-03-31 16:18
5	秦晏观殊	0.775682	9	16-03-31 02:05
6	世属三	0.775304	16	16-03-31 09:35
7	我这么纯洁根本听不懂	0.774523	7	16-03-31 20:42
8	nemo	0.774514	16	16-03-31 00:41
9	liuxp	0.773645	4	16-03-31 23:16
10	GoDown	0.773497	13	16-03-31 10:43

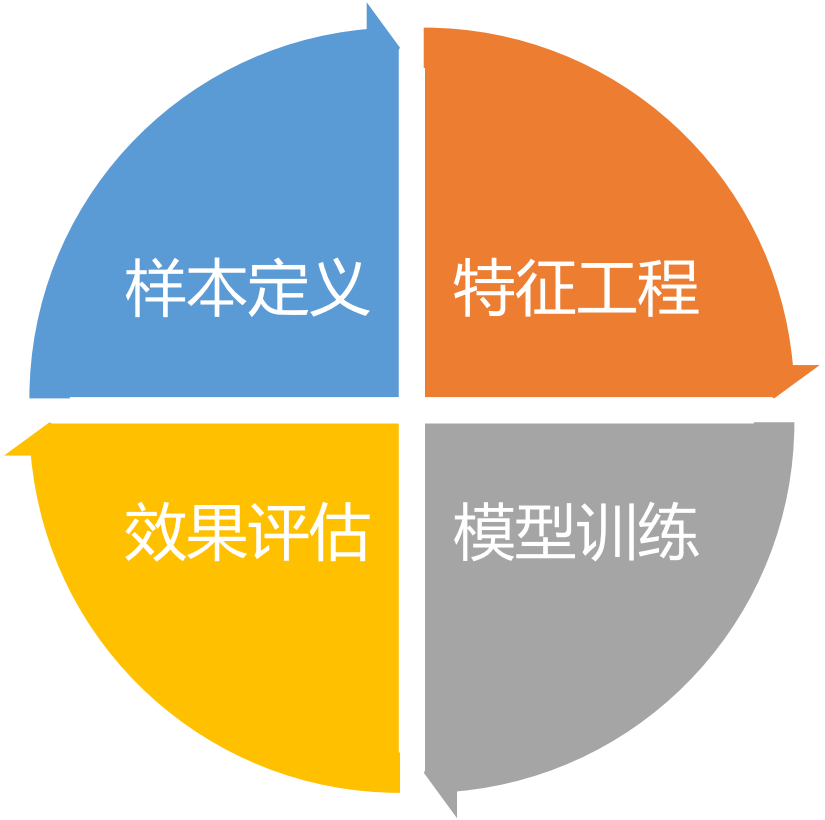
课程内容回顾

1

A. 需要反映预测数据集
B. 分层抽样
C. 选择合适的预测目标

2

A. 主题分类
B. 衍生特征
C. 特征选择



A. 5种评估指标
B. 过拟合

4

A. 样本切分
B. 模型是如何学习的
C. 逻辑回归

3