百分点 BAIFENDIAN.COM | 大数据践行者 BIG DATA PRACTITIONER
China · Beijing

基于机器学习方法对销售预测的研究

唐新春

百分点信息科技有限公司 数据科学家

ArchSummit 全球架构师峰会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ

2016年：1180亿元？

朕略懂 2016-10-21 10:46:30 支付宝 淘宝 阅读(1386) 评论(0)

2016年天猫双十一交易额净增加了268亿，相当于2013年天猫双十一的总量，2015年双十一销售额已经是2009年第一次双十一的1824.34倍，是 2013年的2.6倍。今年天猫双十一销售额将突破1000亿，或达1180亿元。 今年看张勇能否创造奇迹，我们拭目以待！

马云：2016“双十一”破1500亿，天猫淘宝双11玩法

花萌 2016-10-28 13:35:18 阅读(1574) 评论(0)

声明：本文由入驻搜狐公众平台的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。 举报

搜狐科技 > 互联网 > 天猫

原创 专家估计，2016年天猫双11销售量有望超1200亿

马继华 2016-11-10 13:34:55 天猫 双11 阅读(10432) 评论(1)

百分点 BAIFENDIAN.COM | 大数据践行者 BIG DATA PRACTITIONER
China · Beijing

基于机器学习方法对销售预测的研究

唐新春

百分点信息科技有限公司 数据科学家

ArchSummit 全球架构师峰会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ

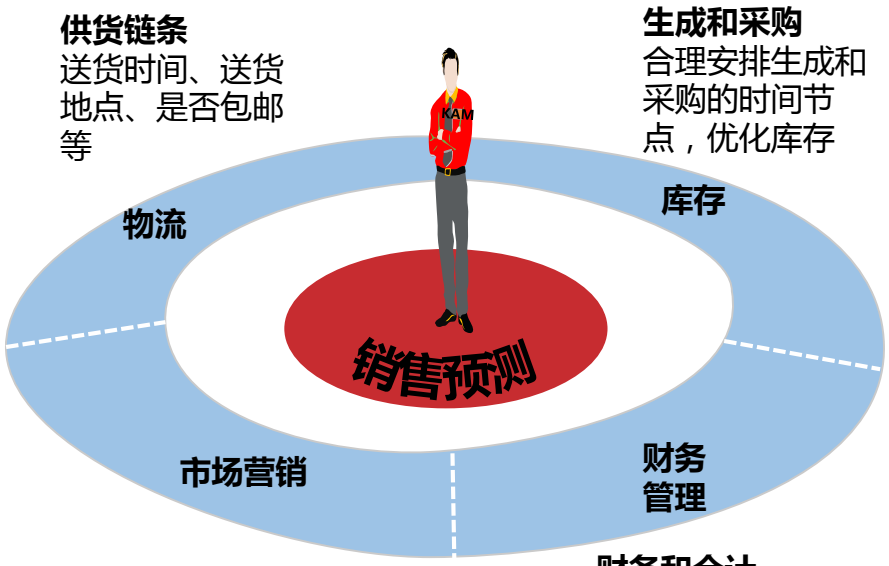
CONTENTS

销售预测的现状与痛点

销售预测是完善客户需求管理、指导运营、以提高企业利润为最终目的商业问题。

而**预测的精确性**是销售预测的核心痛点。

销售预测全景图



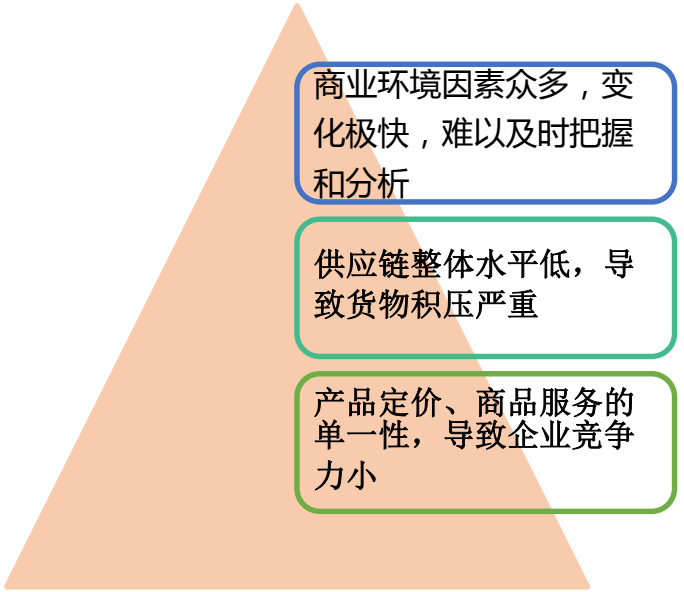
供货链条
送货时间、送货地点、是否包邮等

生成和采购
合理安排生成和采购的时间节点，优化库存

促销方案
商品减价、商品促销、组合销售等

财务和会计
实时反馈公司财务报表，监控公司资金流动管理

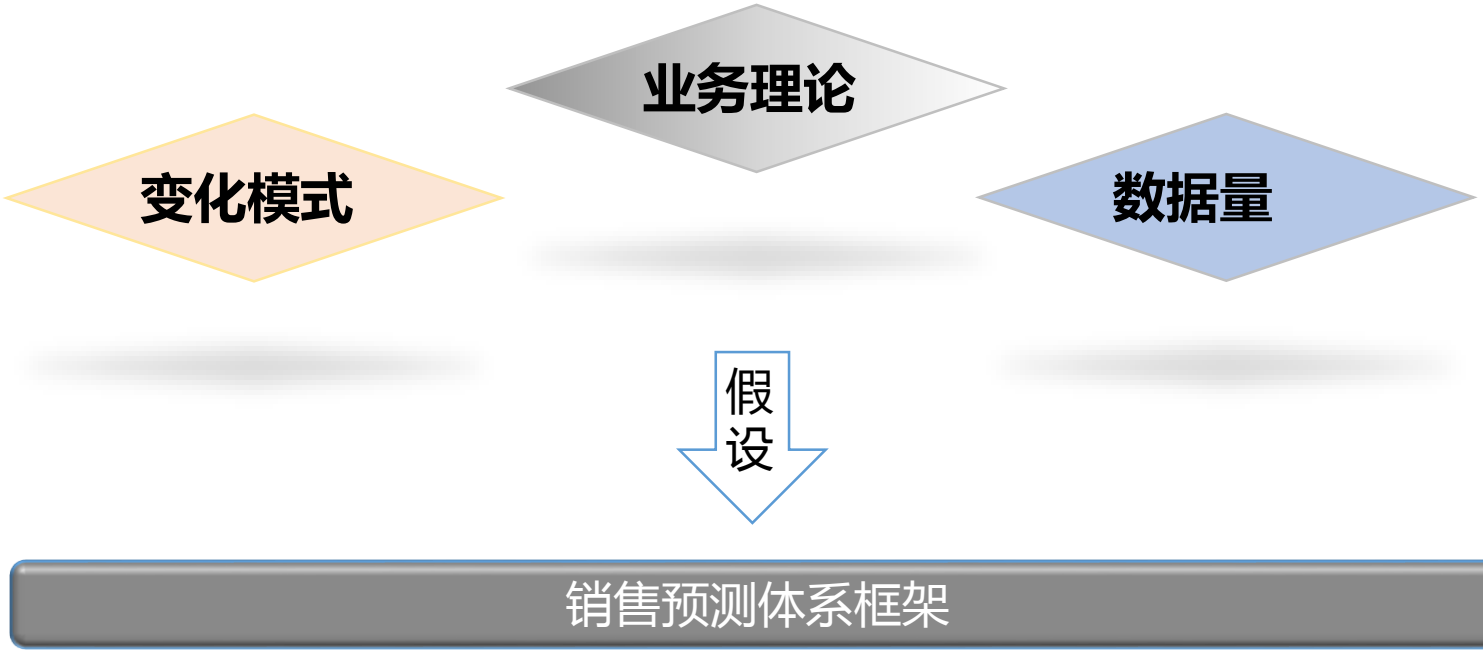
销售预测的痛点



预测的基本思想

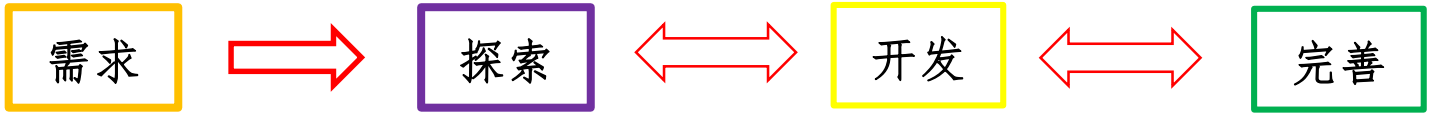
预测是通过历史数据或其他外部因素构建模型、学习其变化“模式”，利用该“模式”对未来事物进行预测的一个过程。

特点：短期预测的精度要远远高于长期预测。



CONTENTS

预测的基本步骤



- 确定预测对象、预测目标和需求；
- 预测周期：短、中和长期
- 需求与预测精度的权衡
  准确性与可解释性的权衡

预测目标

- 收集数据、整理指标体系
- 数据描述与数据探索
- 数据预处理

收集数据

- 算法的选择
- 模型的训练与预测
- 过拟合问题的处理

建立建模

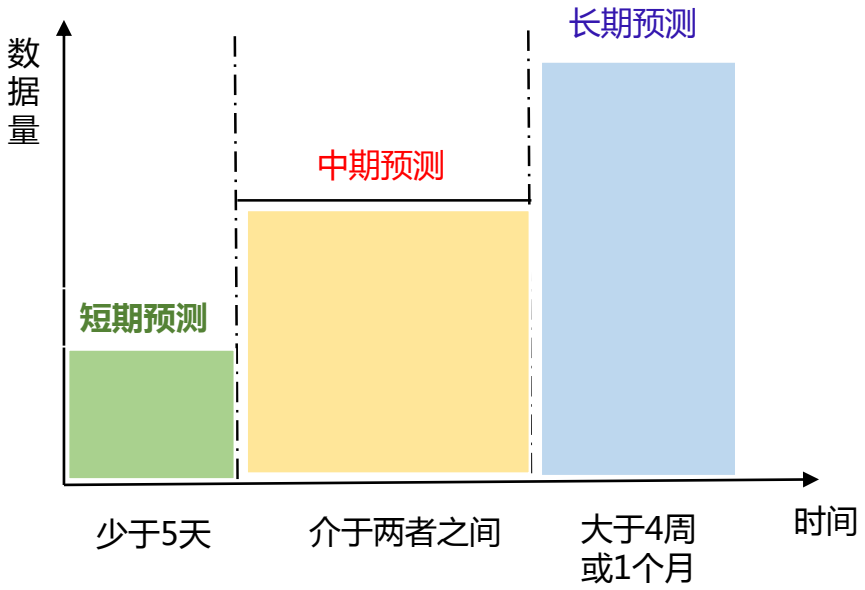
- RMSE、MAPE等定量评价指标
- AIC、BIC等模型评价指标

评价指标

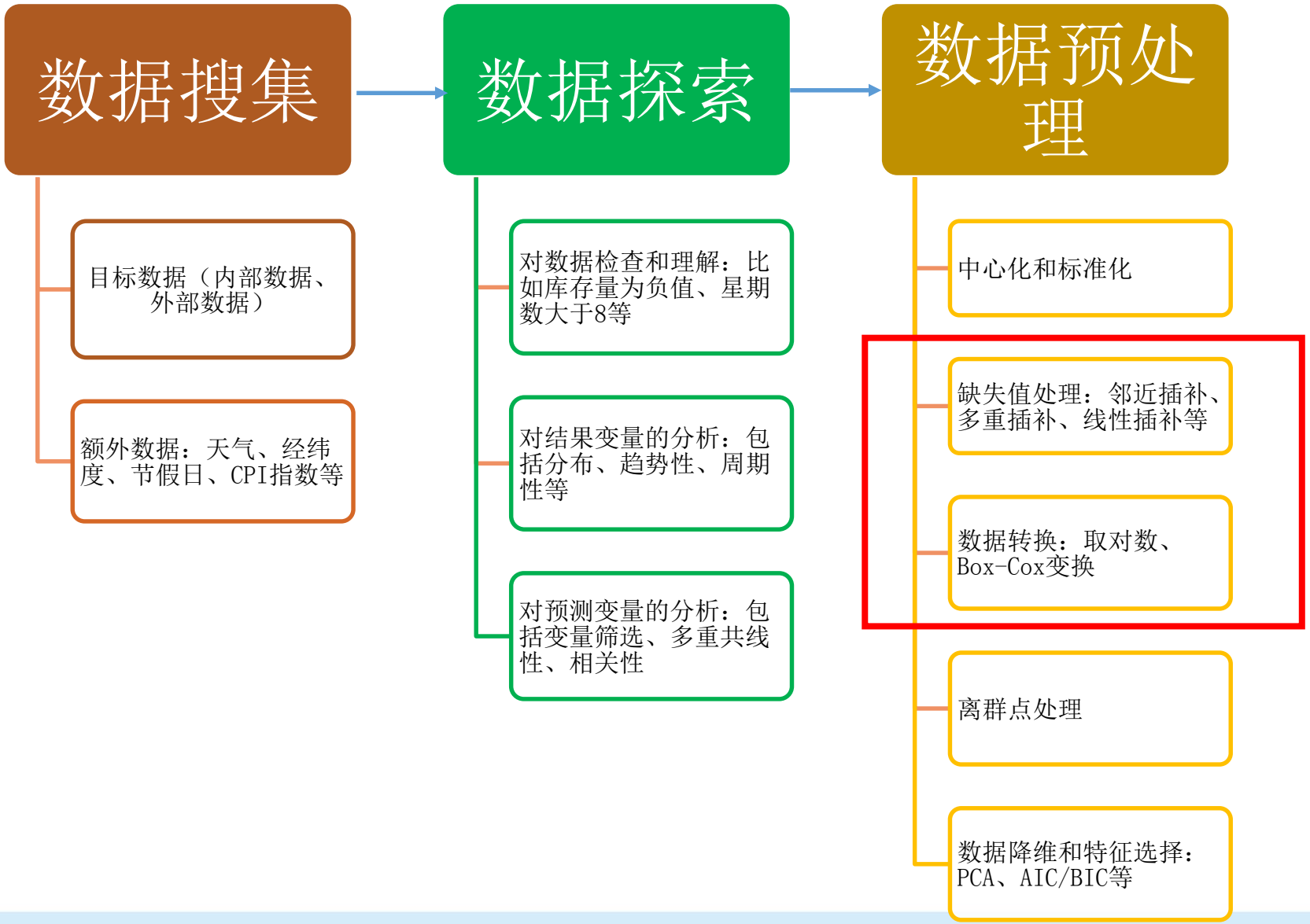
预测目标

预测对象：性质、结构、业务场景等
预测时间：短期预测、中期预测和长期预测等

业务目标：准确性和模型可解释性的匹配度



收集数据或理解数据


数据搜集
目标数据（内部数据、外部数据）
额外数据：天气、经纬度、节假日、CPI指数等
数据探索
对数据检查和理解：比如库存量为负值、星期数大于8等
对结果变量的分析：包括分布、趋势性、周期性等
对预测变量的分析：包括变量筛选、多重共线性、相关性
数据预处理
中心化和标准化
缺失值处理：邻近插补、多重插补、线性插补等
数据转换：取对数、Box-Cox变换
离群点处理
数据降维和特征选择：PCA、AIC/BIC等

收集数据或理解数据

数据预处理的缺失值部分

处理缺失值的两大类方法：

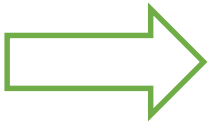
（1）直接删除缺失的预测变量

（2）利用不同的方法对预测变量的缺失值进行插补，插补方法有：均值插补、多重插补、随机插补、K近邻插补、线性插补等。

注意：一般对于带有时间戳的时序变量，考虑到变量的时效性和经济因素，通常采用邻近插补法或者线性插补。

时期	变量
2015-05-23	NA
2015-05-24	10.0
…	
2016-05-09	9.8
2016-05-10	NA

处理之前



时期	变量
2015-05-23	10.0
2015-05-24	10.0
…	
2016-05-09	9.8
2016-05-10	9.8

处理之后

收集数据或理解数据

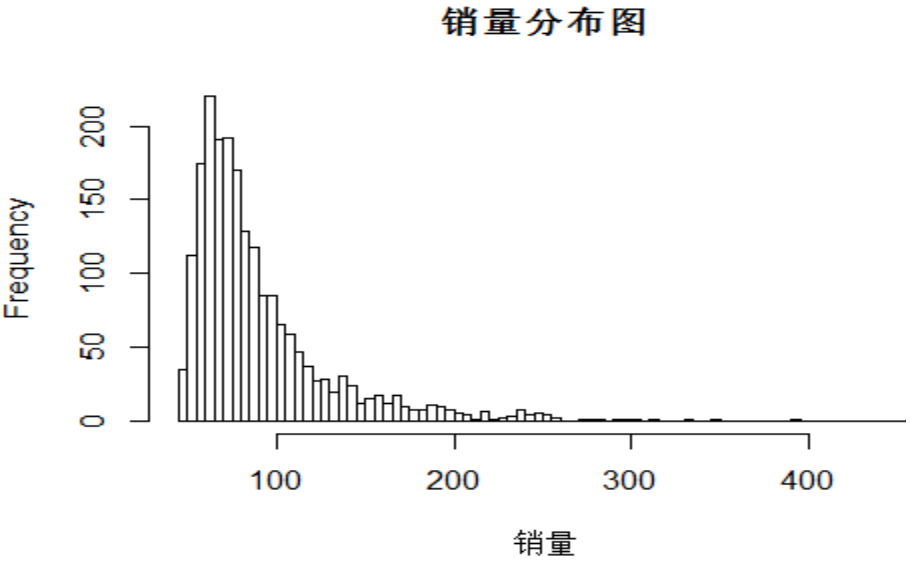
数据预处理的数据变换部分

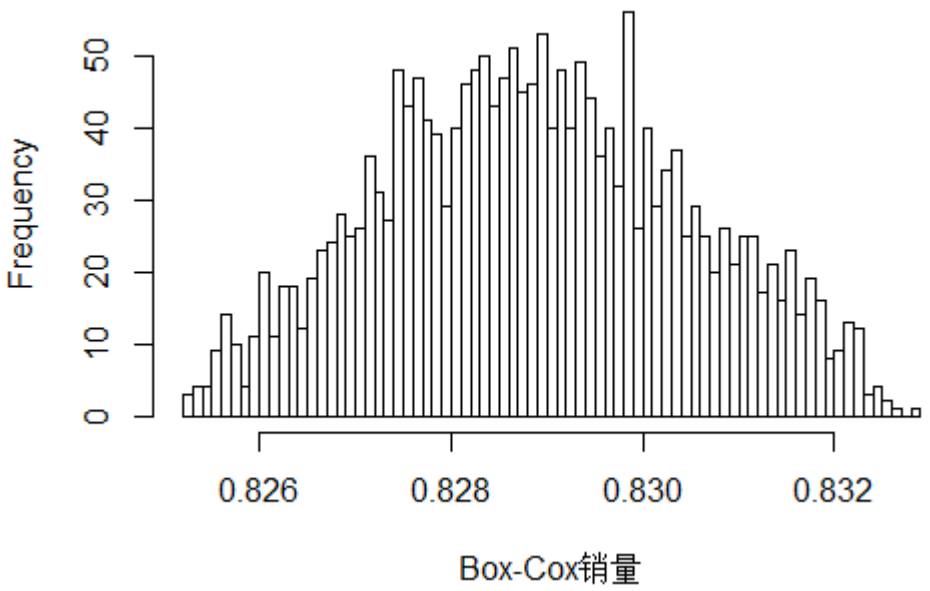
一个需要进行数据变换的原因是去除分布的偏度。一个无偏分布是大致对称的分布，这意味着随机变量落入分布均值两侧的概率大体一致。

数据变换一般有两种方法：

（1）对数据做变换，如取对数、平方根或倒数

（2）Box-Cox变换





CONTENTS

销售预测的基本方法

主观预测	时间序列	机器学习
专家法	指数平滑法	线性回归
	自回归移动模型	决策树
		随机森林
		xgboost
		神经网络
		支持向量回归

专家法

专家预测法： 由专家根据他们的经验和判断能力对待定产品的未来销售进行判断和预测，通常有三种不同的形式：

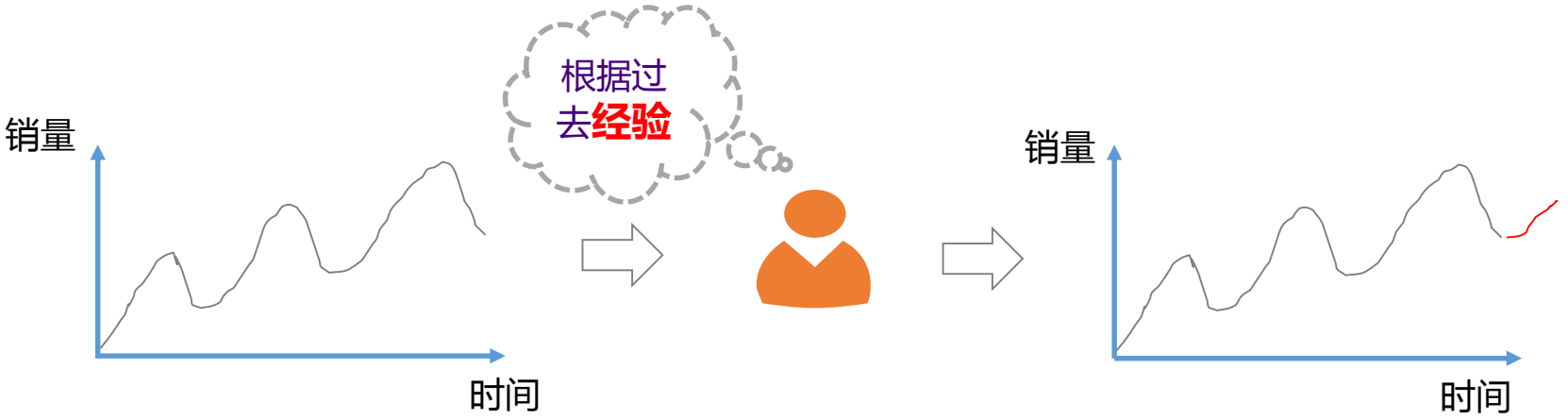
（1）个别专家意见汇集法

（2）专家小组法

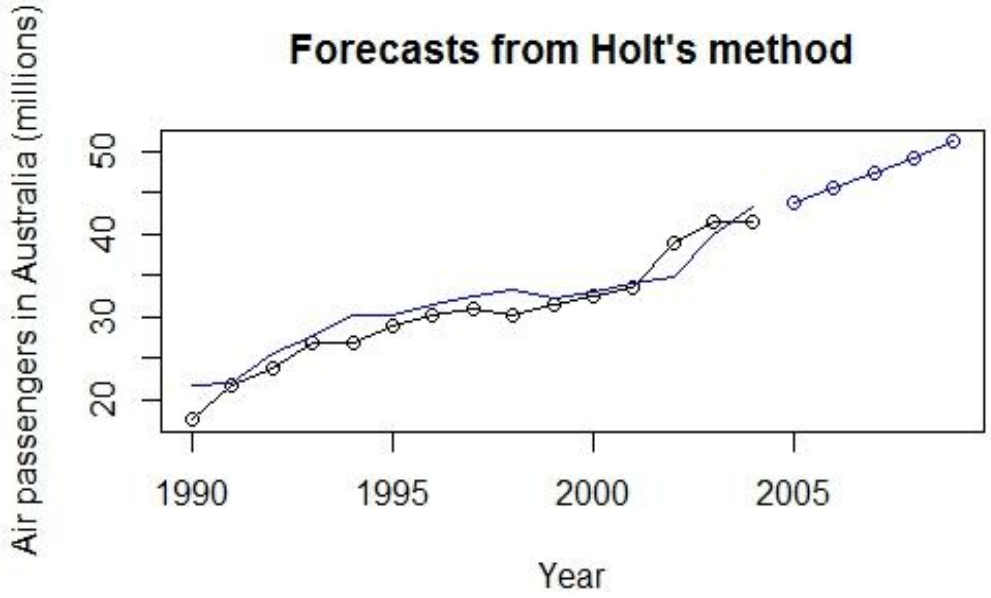
（3）德尔菲法

优点： 简单、快速

缺点： 准确率低、受人的主观影响大



指数平滑法



指数平滑遵循“重近轻远”原则，对全部历史数据采用逐步衰减的不等加权办法进行数据处理的一种预测方法。

基本公式：

$$y_{t+1} = a \cdot y_t + (1-a)y_{t-1}$$

其中，y_t是时间t的时间值；

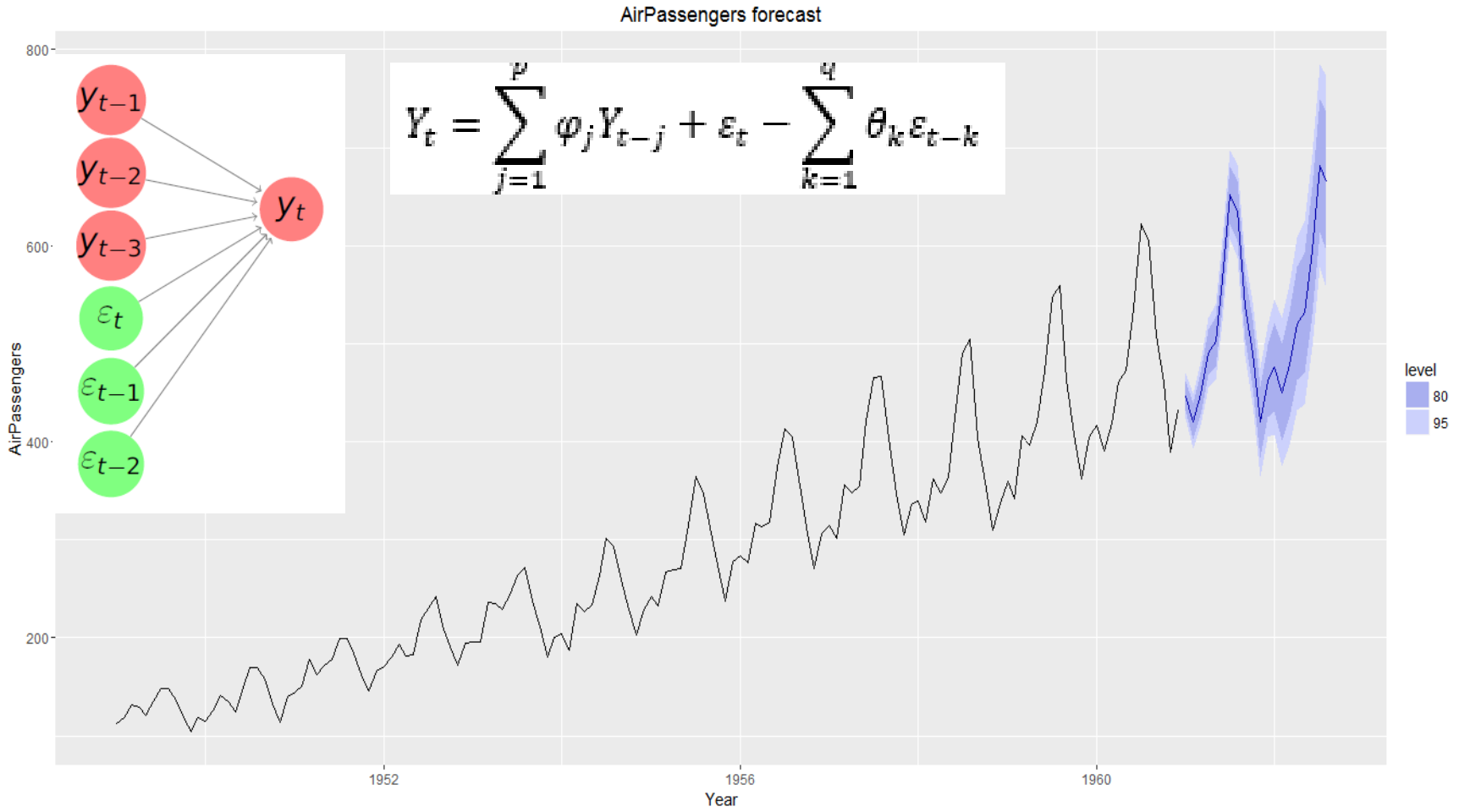
a是平滑常数，其取值范围为[0, 1]。

优点：简单、适合趋势预测、模糊预测

缺点： 准确率不高、需要趋势性较好的数据

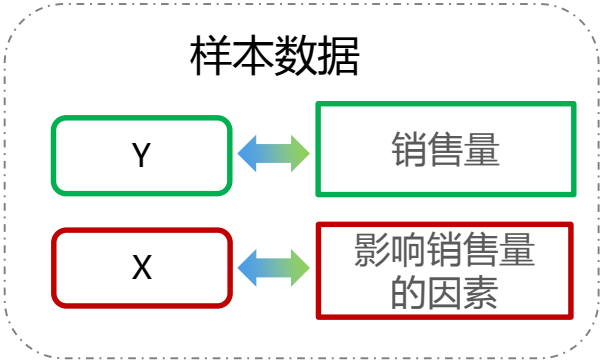
自回归移动模型（ARIMA）

ARIMA模型是指将非平稳时间序列转化为平稳时间序列，然后将结果变量做自回归（AR）和自平移（MA）。



机器学习的实现流程

使用机器学习的有监督学习对进行销量变化进行建模，
依据建模结果来预测未来销量值。其实现流程如下：



特征筛选

特征筛选是一类预测变量变换的方法，通过这种方式，能够用更具有信息量的变量来构建模型，排除无信息量的变量的噪声干扰，提高模型稳健性。

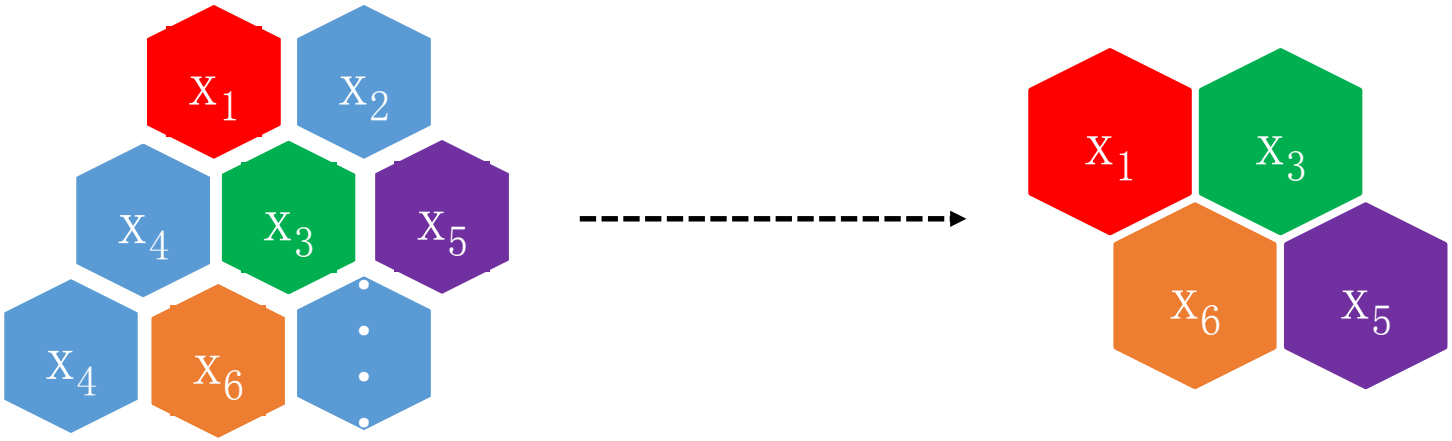
常用的特征筛选方法：

（1） 相关性等统计 [阈值过滤]

（2） 信息增益、信息增益率、基尼系数等 [阈值过滤]

（3） 向前、向后和逐步选择法，如AIC/BIC准则 [最小值]

（4） 模型选择，如随机森林、LASSO等 [模型输出]



线性回归模型：广义线性模型

线性模型

随机分布

>指数分布、泊松分布等

线性回归

$$>Y = \theta_0 + \theta_1 * X_1 + \theta_2 * X_2 + \cdots + \theta_N * X_N$$

Y为销量值，X为预测变量，N为预测变量个数，θ为参数

激活函数

>激活函数：sigmod， log函数等

逻辑回归

原理简介：

- 通过结果变量与预测变量直接建立线性关系
- 数值型回归

优点：

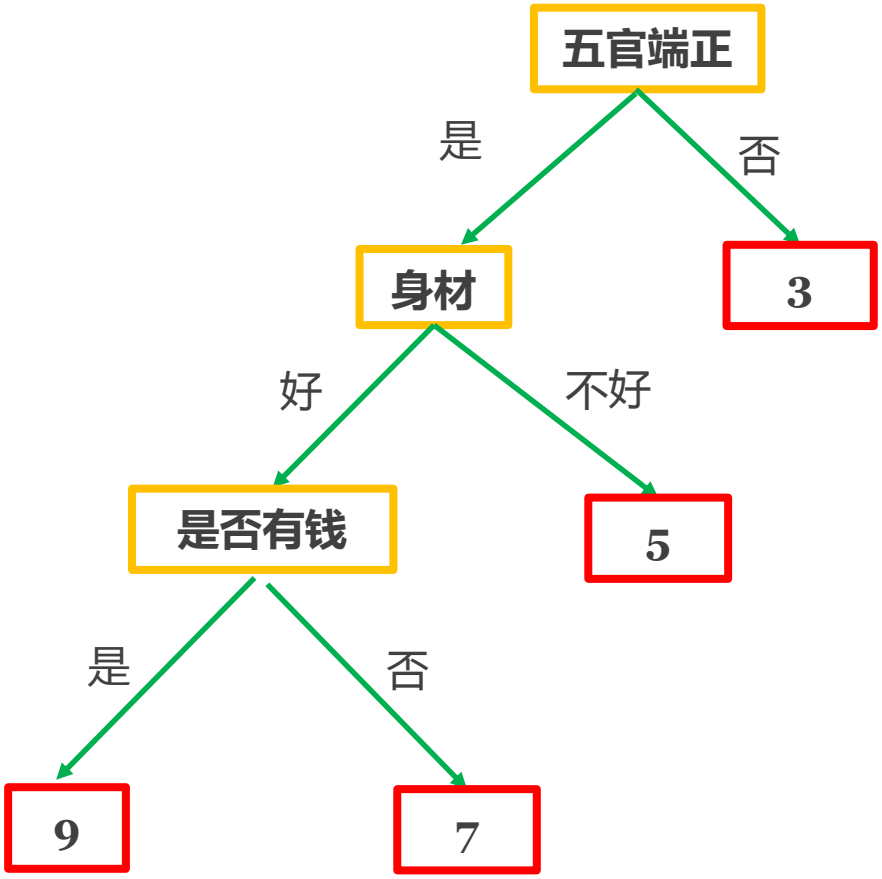
- 模型可解释性强

缺点：

- 只适用于线性规律

决策树（回归树）

妹纸评分：[0,10]



原理简介：

- 通过训练数据，形成if-then规则集合
- 由根节点到叶节点的每一条路径构成规则
- 对结果变量有主要解释作用的特征会先分裂形成规则
- 回归树用平方误差最小化准则，节结点为单元内数值的平均值

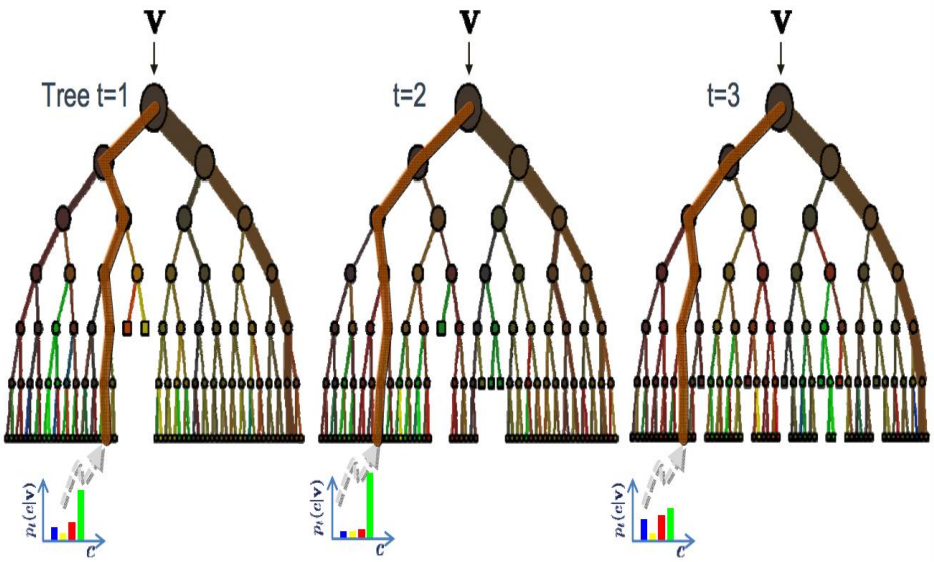
优点：

- 可拟合非线性规律，计算复杂度较低

缺点：

- 容易出现过拟合

随机森林



原理简介：

- 是包含多个回归树的组合器
- 输出的数值是由个别树输出的数值的平均而定

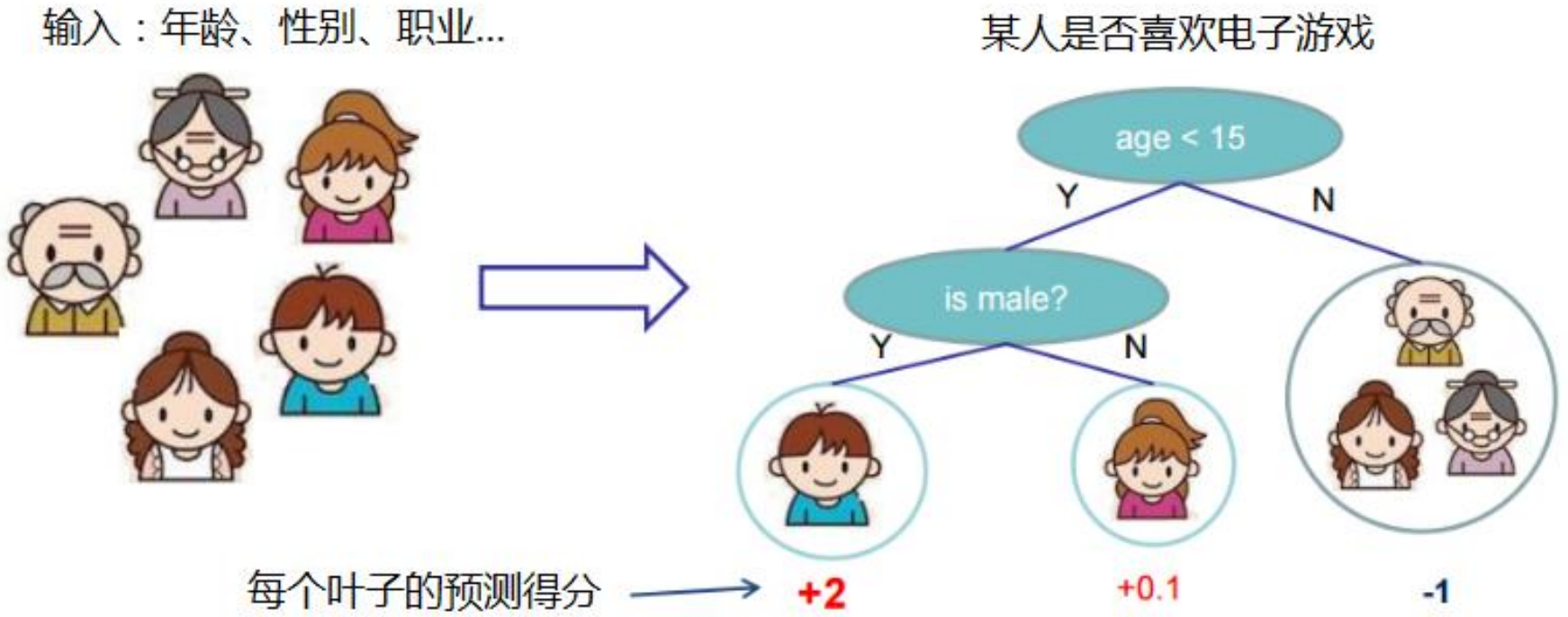
优点：

- 准确度高
- 训练速度快
- 容易做出并行算法
- 可处理大量变量并评估变量重要性

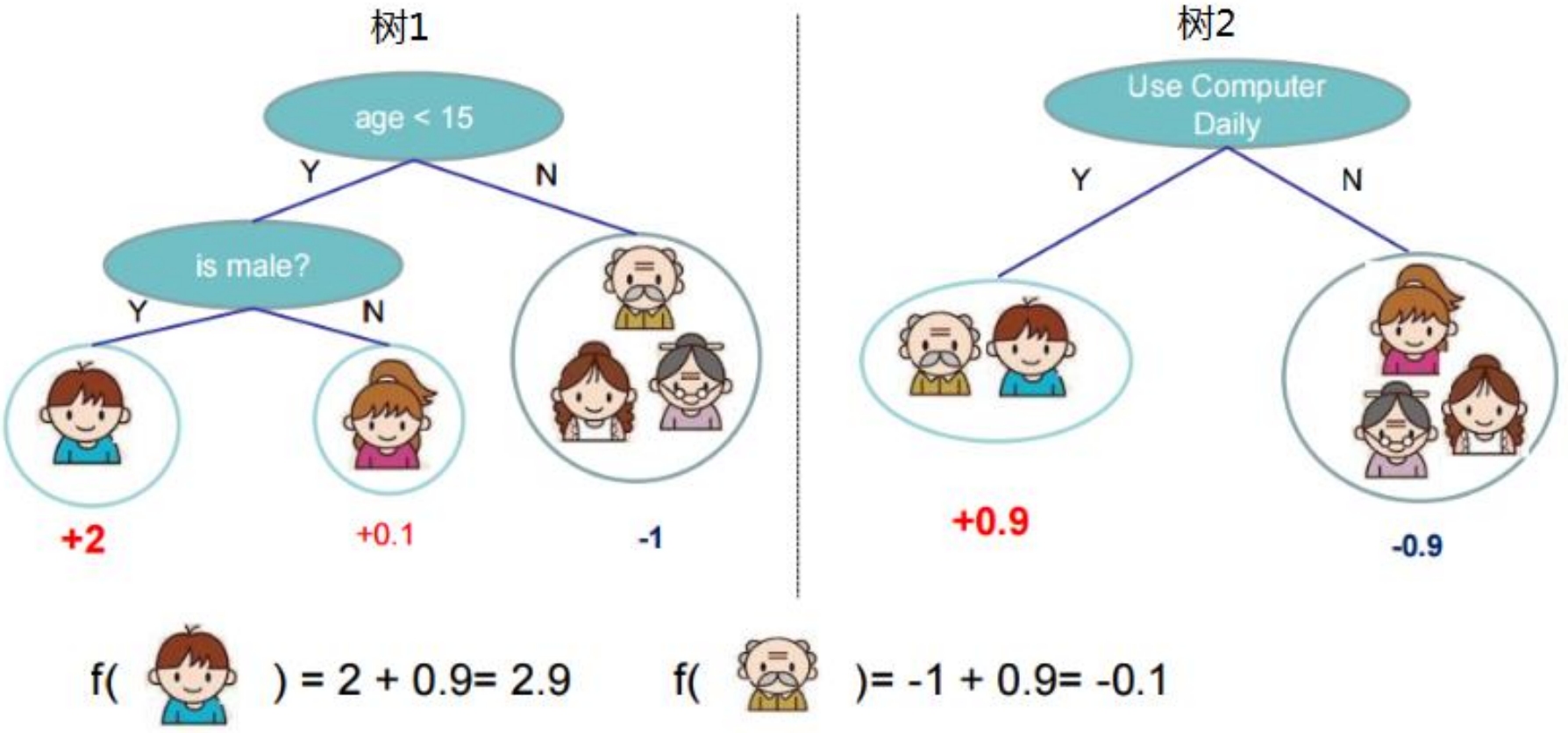
缺点：

- 在噪声较大的数据上会有过拟合问题

xgboost



xgboost



f( ) = 2 + 0.9= 2.9

f( )= -1 + 0.9= -0.1

xgboost

目标函数去掉常数项：

$$Obj^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)$$

找到那颗树f_t，使得目标函数达到最优即可。

原理简介：

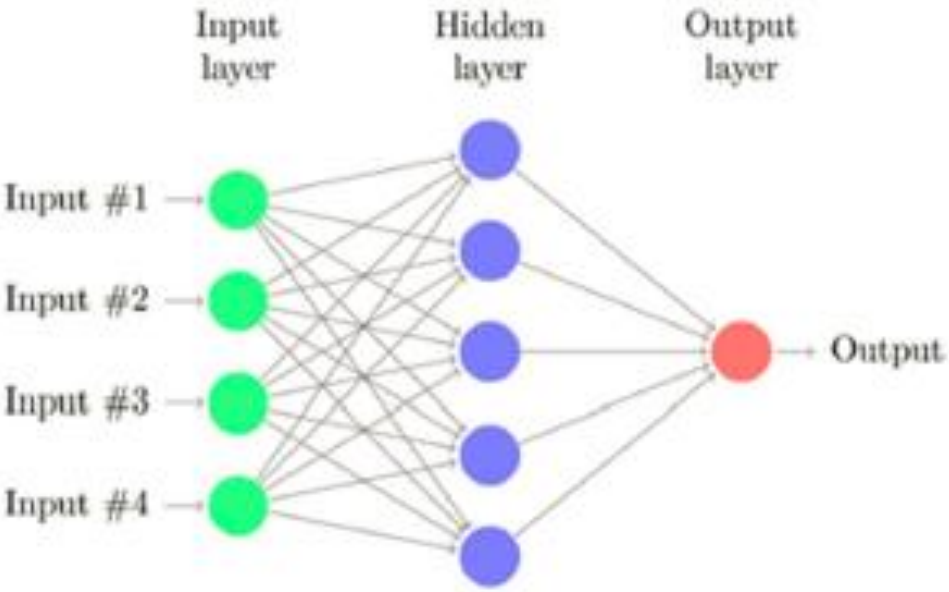
- 是基于传统的GBDT上做了一些优化的开源工具包，目前有python,R,Java版。

优点：

- 高速准确
- 可移植，可以自己定义假设函数
- 可容错

神经网络

$$Y = f(X)$$（非线性映射）



原理简介：

- 是利用一系列非线性回归，将预测变量映射到结果变量的一种方法。

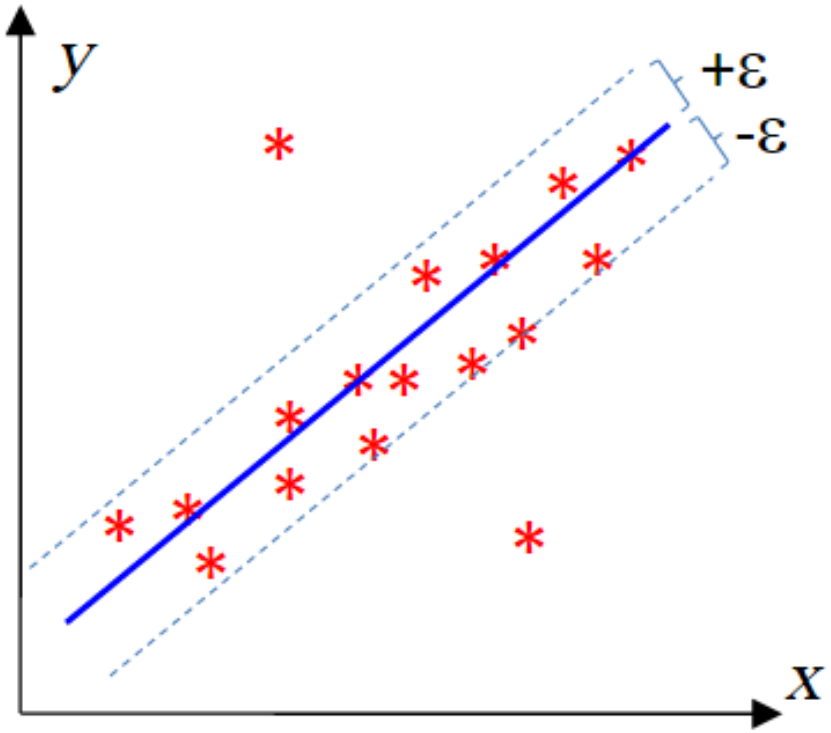
优点：

- 准确度高
- 训练速度快
- 并行处理能力强

缺点：

- 需要大量的参数
- 不能观察学习的过程，对结果难以解释

支持向量回归（SVR）



原理简介：

- 是通过寻求结构化风险最小来提高学习泛化能力，实现经验风险和置信范围最小化，从而达到获得良好统计规律的目的

优点：

- 可以解决小样本情况下的机器学习问题
- 可以解决高维、非线性问题

缺点：

- 对非线性问题没有通用解决方案，对核函数的选择非常敏感

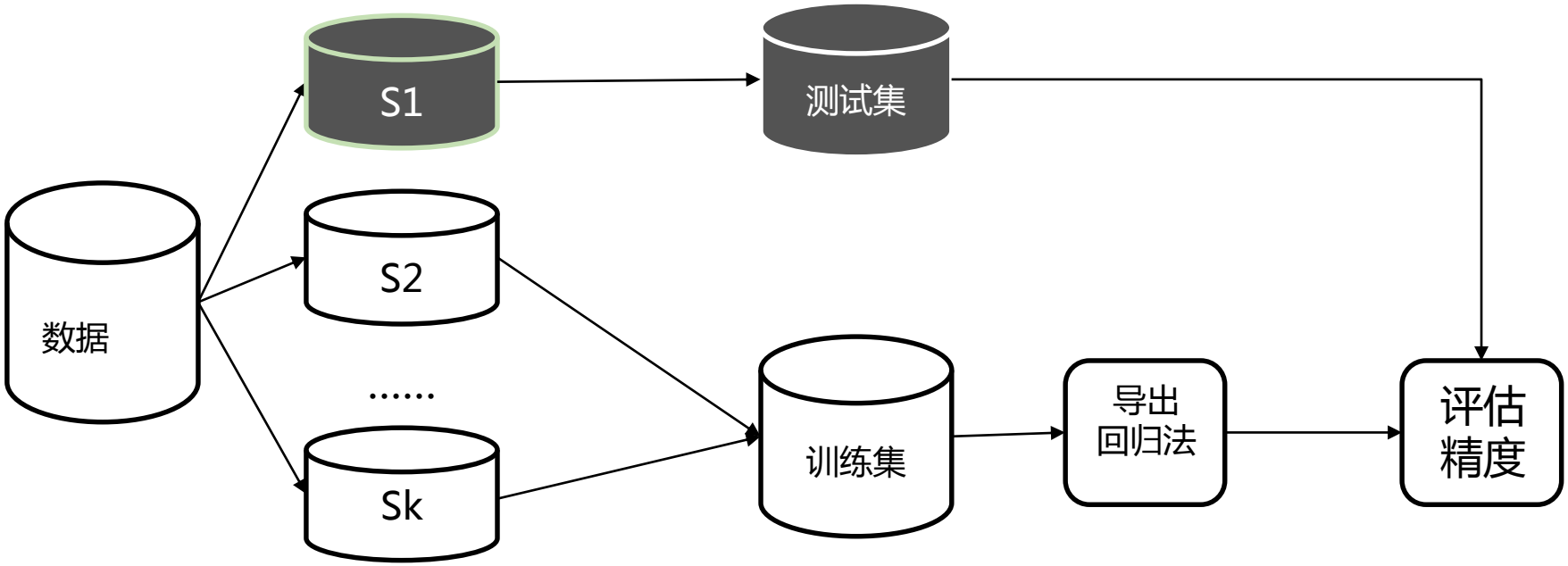
SVR最本质与SVM类似，都有一个margin，只不过SVM的margin是把两种类型分开，而SVR的margin是指里面的数据不会对回归有任何帮助。

CONTENTS

模型评估方法：k折交叉验证法

K折交叉验证法

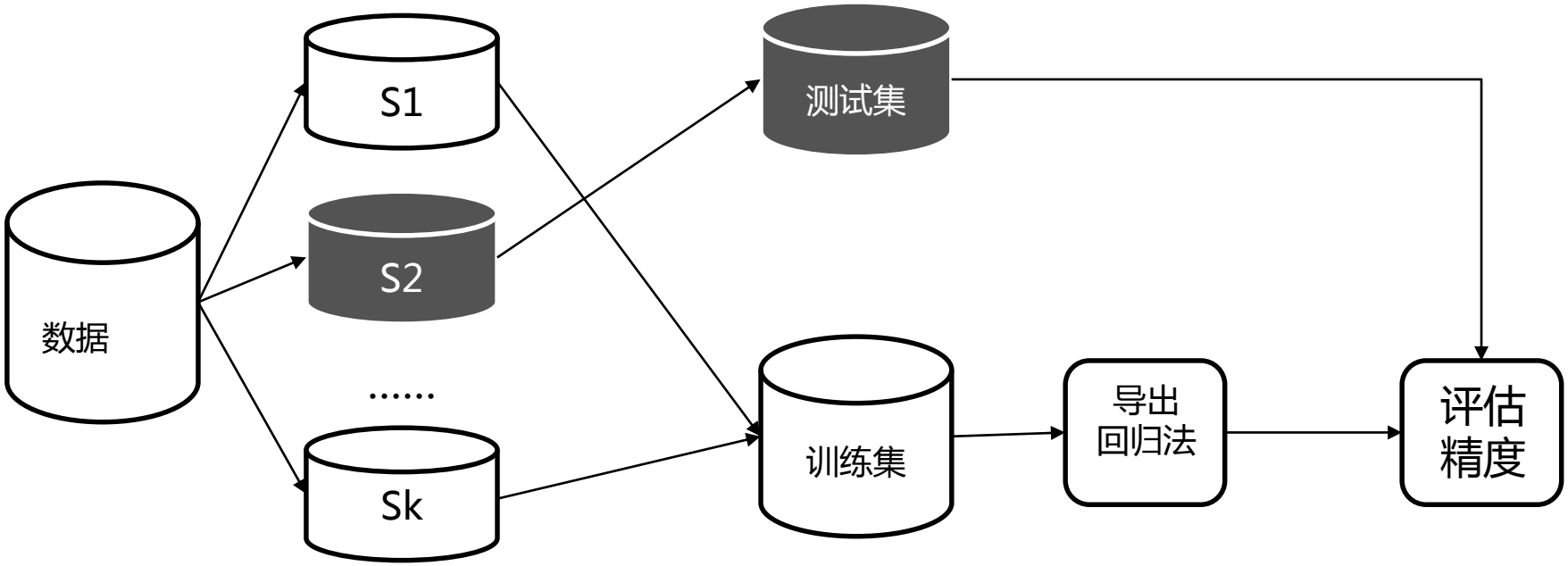
- 在k-折交叉验证中，初试数据被划分成k个互不相交的子集或“折”，每个折的大小大致相等。训练和测试k次。在第i次迭代中，第i折用作测试集，其余的子集都用于训练分类法。
- 准确率估计是k次迭代正确分类数除以初始数据中的样本总数。



模型评估方法：k折交叉验证法

K折交叉验证法

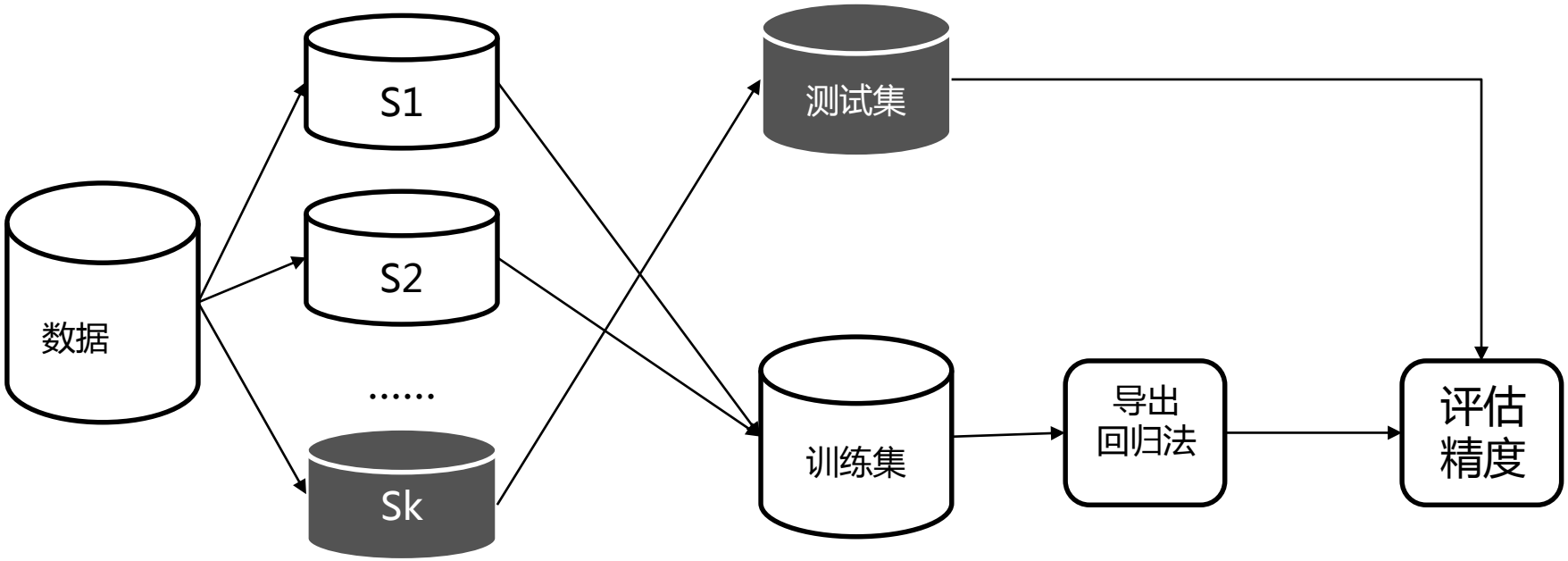
- 在k-折交叉验证中，初试数据被划分成k个互不相交的子集或“折”，每个折的大小大致相等。训练和测试k次。在第i次迭代中，第i折用作测试集，其余的子集都用于训练分类法。
- 准确率估计是k次迭代正确分类数除以初始数据中的样本总数。



模型评估方法：k折交叉验证法

K折交叉验证法

- 在k-折交叉验证中，初试数据被划分成k个互不相交的子集或“折”，每个折的大小大致相等。训练和测试k次。在第i次迭代中，第i折用作测试集，其余的子集都用于训练分类法。
- 准确率估计是k次迭代正确分类数除以初始数据中的样本总数。



模型评估指标： RMSE（均方根误差）

– RMSE

– 与分类模型不同，回归模型是对连续的因变量进行预测，因此判断回归模型的准确率需要考虑的是预测值与真实值之间差异的大小。

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}}$$

– 其中，y_i为第i个样本的真实值，$\widehat{y_i}$为第i个样本的预测值，n为样本量。

– 有时也用$MSE = \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}$来评估回归模型的准确率，与RMSE效果相同。

模型评估指标： AIC & BIC

AIC准则是评估统计模型的复杂度和衡量统计模型拟合优度的一种标准:

- $$AIC = -2\ln(L) + 2p$$
- 其中L是在相应模型下的最大似然估计值，p是模型的变量个数。
- 增加变量的数目提高了拟合的优良性，但可能造成过度拟合的情况。AIC鼓励数据拟合的优良性但是尽量避免出现过度拟合（overfitting）的情况。
- AIC值越小，模型越好。AIC准则是寻找可以最好地解释数据但包含最少自由参数的模型。

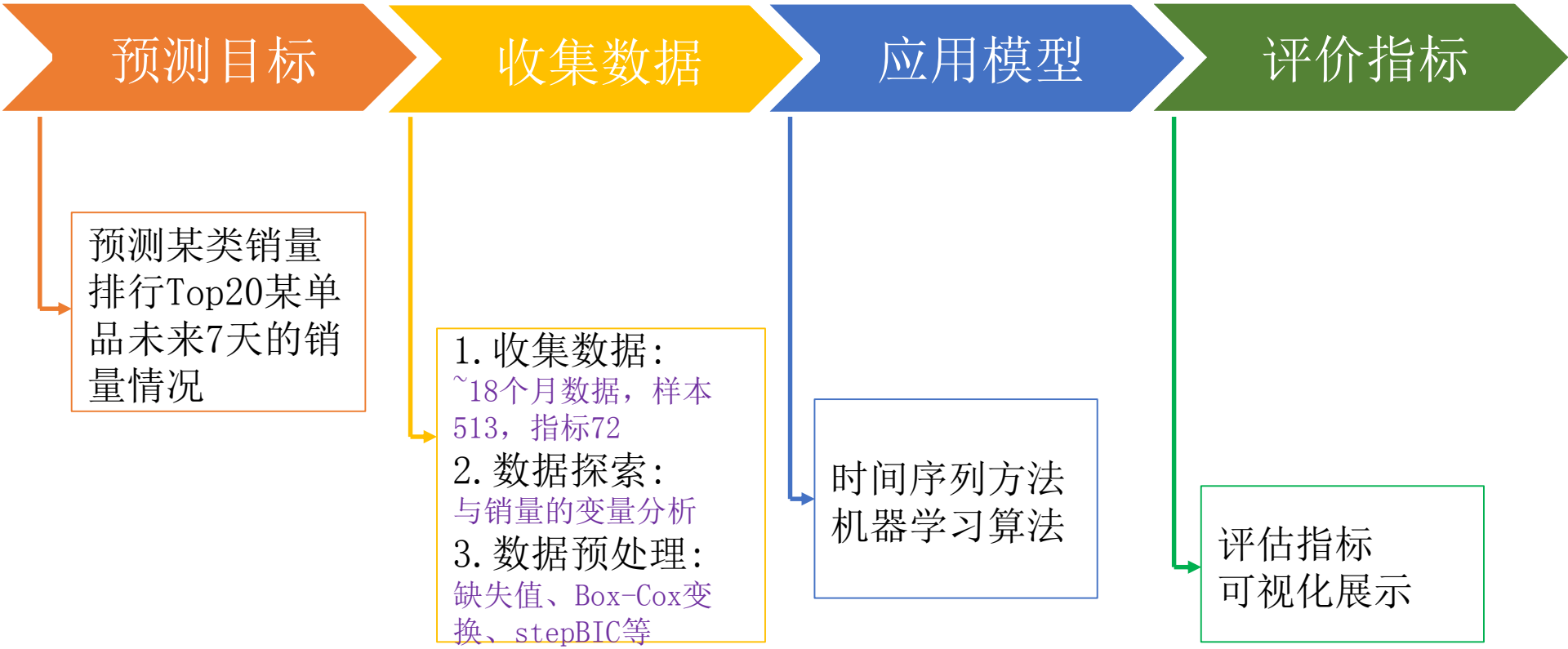
- BIC准则是依贝叶斯理论提出的一种模型选择准则。
- $$BIC = -2\ln(L) + \ln(n)p$$
- 其中L是在相应模型下的最大似然估计值，n是样本量，p是模型的变量个数。
- BIC值越小，模型越好。
- AIC准则倾向于过拟合，BIC准则倾向于欠拟合，BIC选出的模型相对于AIC的更为精简。

CONTENTS

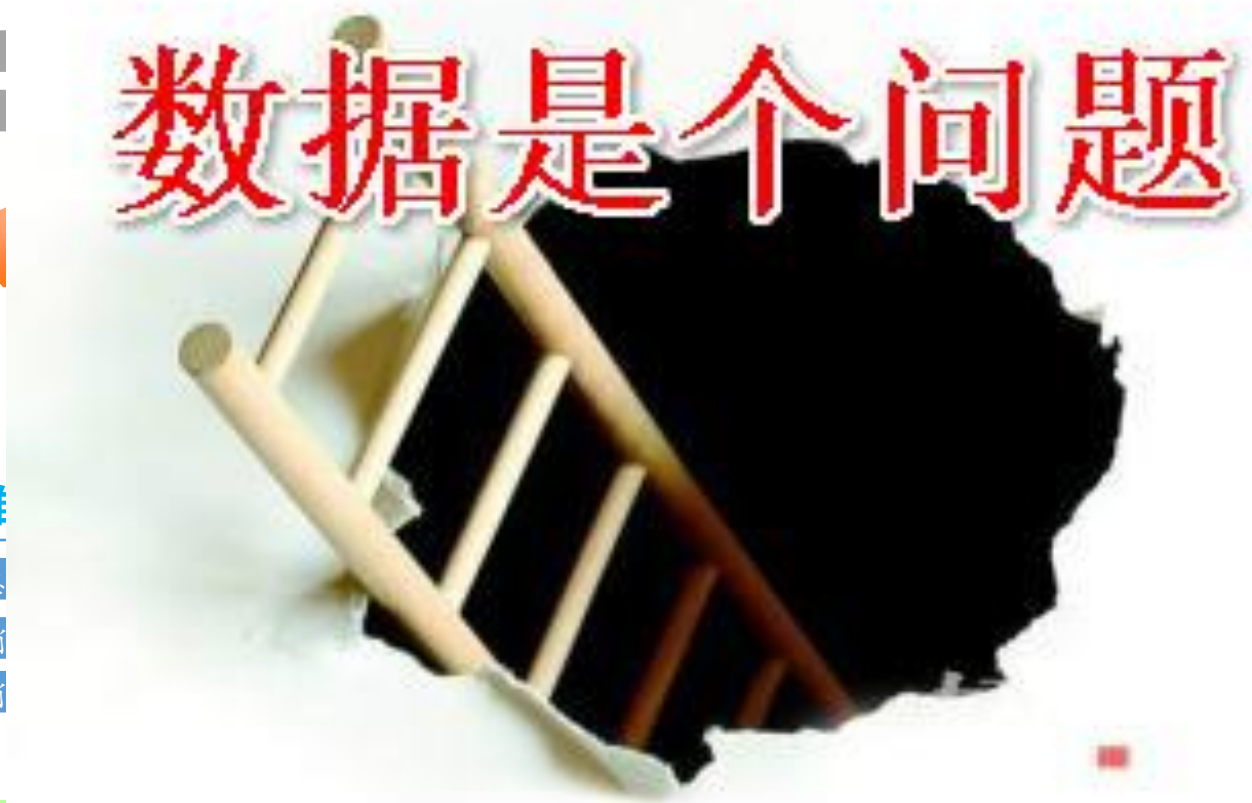
销售预测的企业案例分析

项目背景

某电商平台主营海外代购业务，由于海外代购物流时间长、发货时间慢等因素导致 该电商平台存在大量库存积压情况，想通过销售预测模型改善安排进货、提高发货速度以及优化库存。



收集数据：采集影响商品销量指标维度



3 营销推

商品优惠

商品促销

商品分销

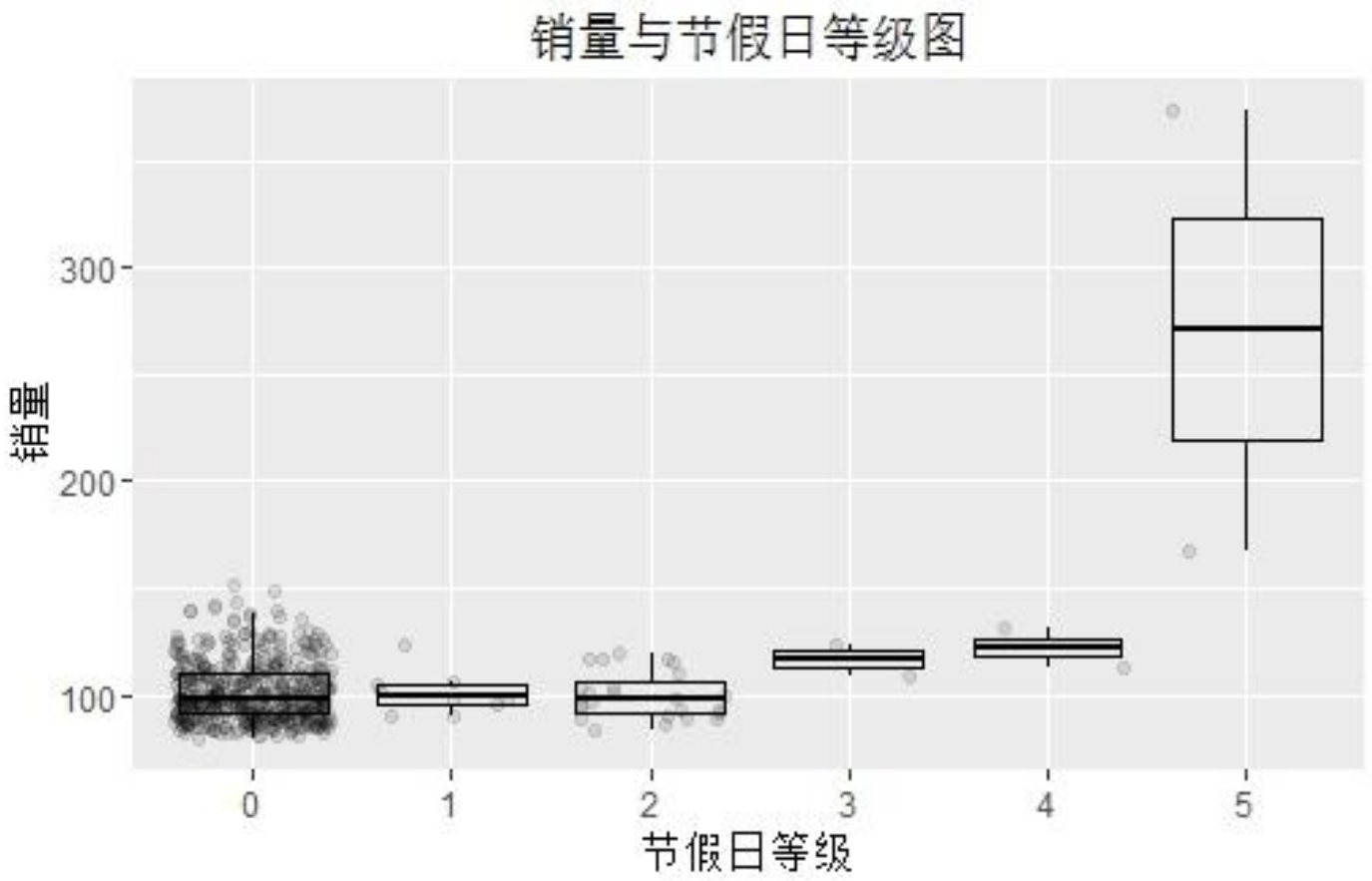
6

车行为

单行为

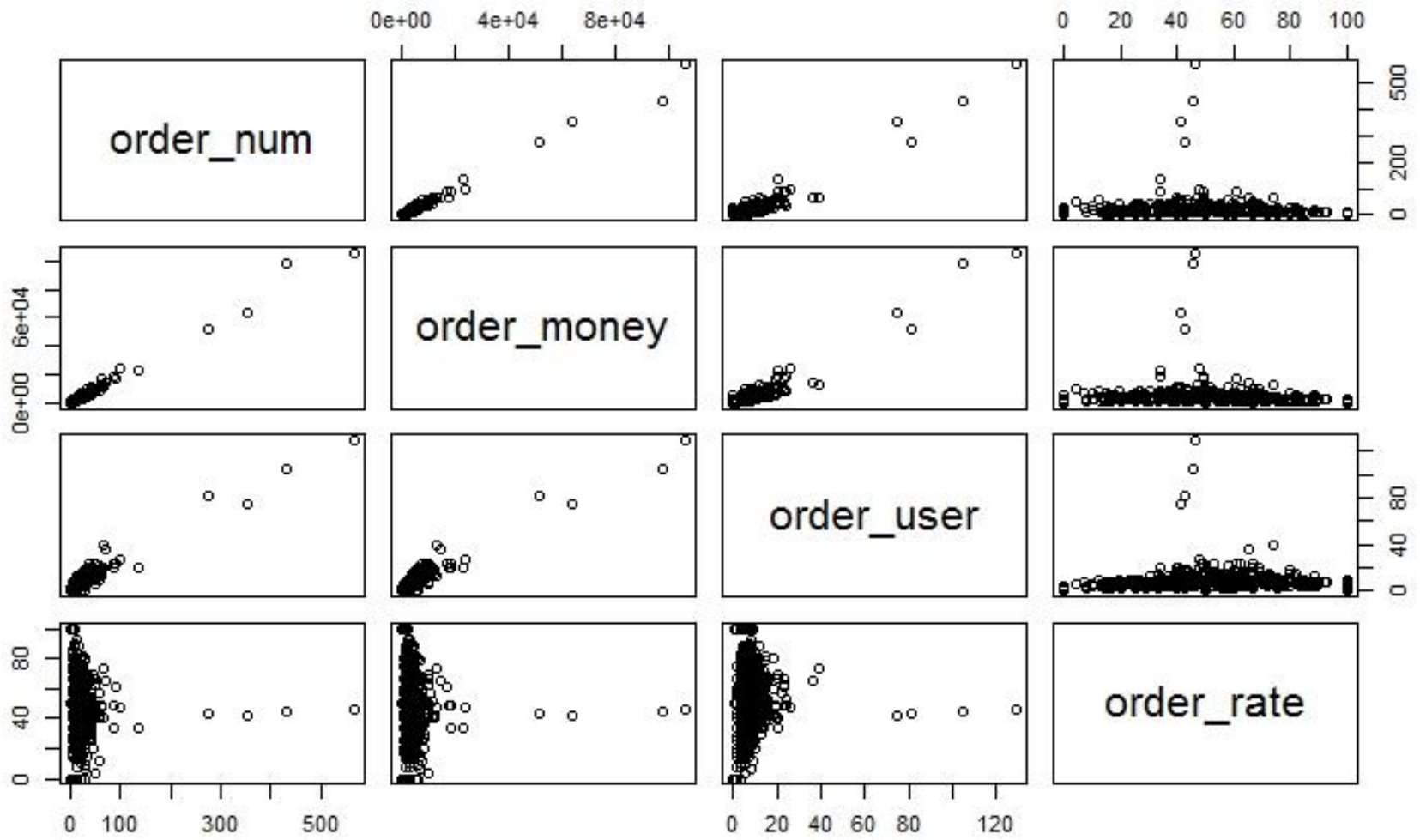
寸行为

数据探索：节假日与销量关系



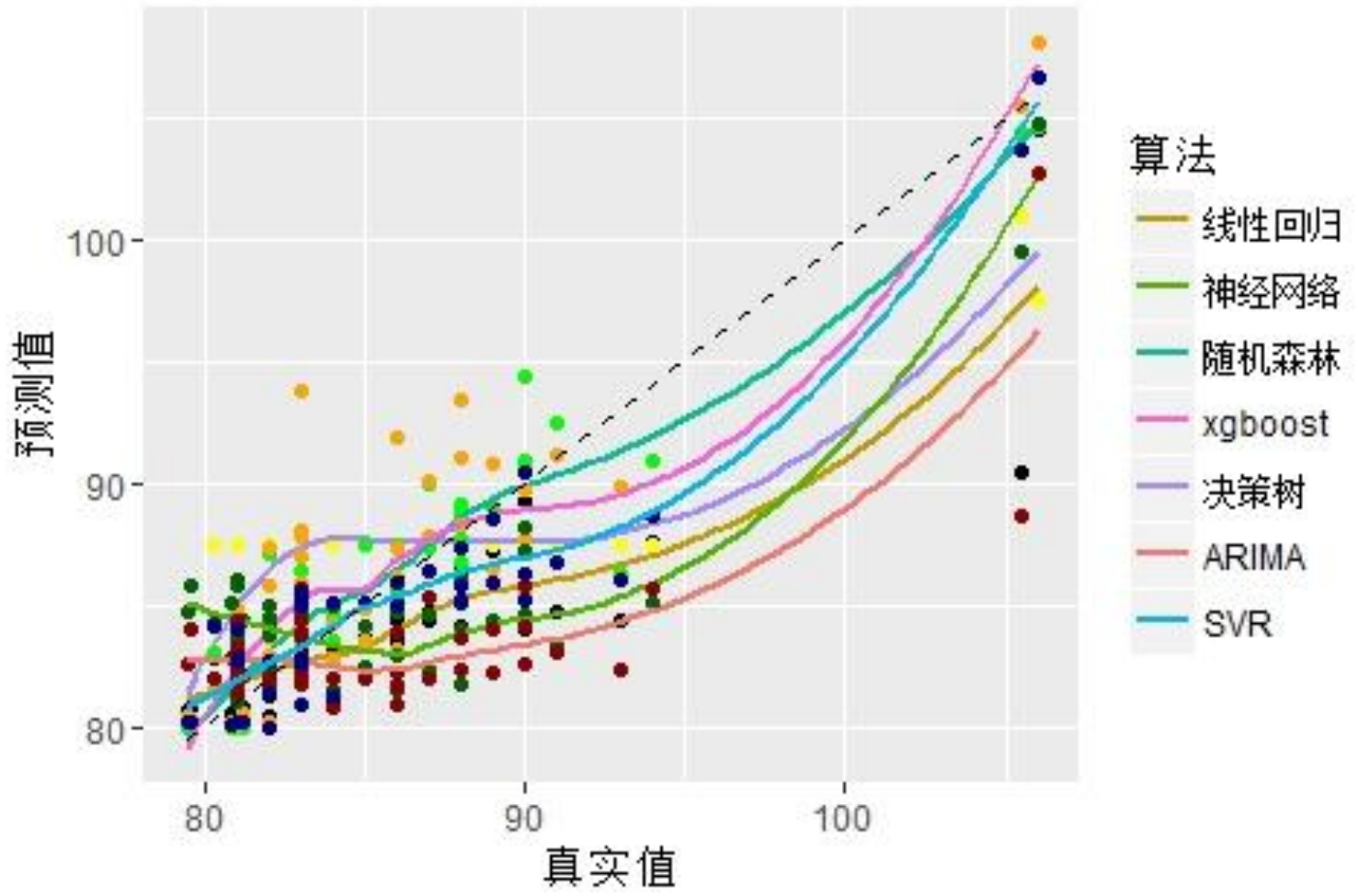
节假日变量对销量的影响明显

数据探索：预测变量之间的多重共线性



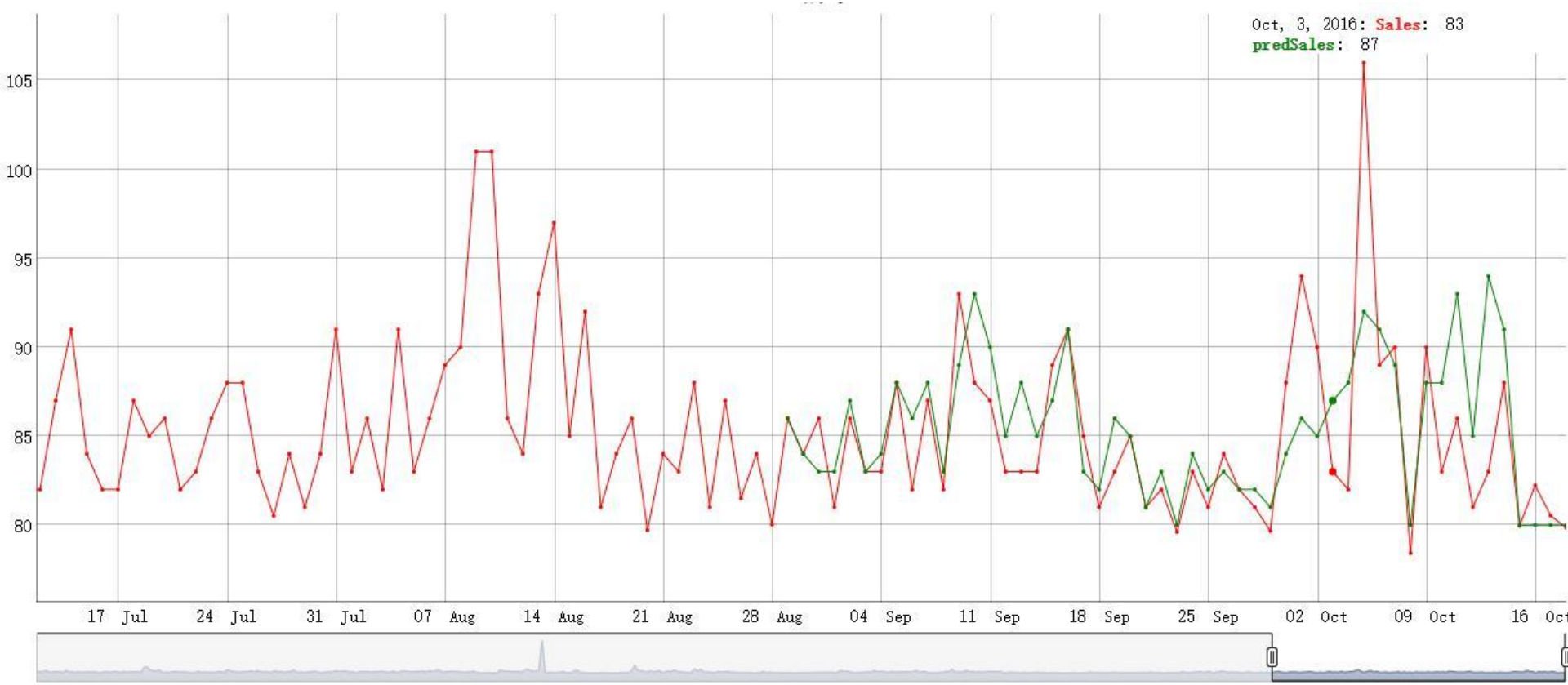
结论：订单数、订单金额和订单用户数之间相关性较高！需要过滤多重共线性！

模型效果：真实值与预测值对比图



xgboost和随机森林的预测效果较佳，线性回归和ARIMA效果较差

应用模型：xgboost



红色点线：某商品的真实销量
绿色点线： 某商品的预测销量
RMSE（均方根误差）：3.68837

模型评估：RMSE

算法名称	RMSE	性能
ARIMA	5.32	速度较慢，2.5min
线性回归	4.28	速度快，< 1min
决策树	5.02	速度快， < 1 min
随机森林	2.85	速度快， < 1 min
xgboost	3.68	速度适中，1.5min
神经网络	4.99	速度快， < 1 min
支持向量回归	3.27	速度快， < 1 min

◇**随机森林：**采取的是重抽样，具有自动选择重要特征的功能，无需做特征筛选，在一定程度上避免了过拟合

◇ **xgboost:** kaggle比赛上表现卓越的算法之一，从本质上分析是一个集成的决策树，但是可以让弱回归树集成成强回归树

因此，它们可以在本案例中能够取得的好预测效果

◇**而其他算法：**处于数据的局限性或模型的参数未达到最优，会存在一定的过拟合，导致预测效果相对较差

模型总结

算法名称	预处理	变量选择	可解释性	准确性
ARIMA	缺失值/变量筛选	stepBIC	低	低
线性回归				低
决策树				低
随机森林				高
xgboost				高
神经网络				低
支持向量回归	缺失值/标准化/变量筛选	stepBIC	低	高

基于机器学习的销售预测总结

机器学习

是场景局限性，机器学习**不是万能的**；

研究的是**相关关系**，而不是因果关系。

数据

是核心，**无数据或数据质量低**，会影响模型预测效果；

是模型选择的先决条件，**先数据，后模型**。

效果

评估需要参考业务对接、预测精度、模型可解释性和产业链整体能力等因素**综合考虑**；

不能简单作为企业利润增加的**唯一标准**。

业务

对建模提供**业务理论基础**；

算法问题要回归到**业务问题**。

基于机器学习的销售预测展望

- 可以尝试使用更复杂的模型来做销售预测，如**HMM**，**深度学习（Long Short-Term Memory网络）**等，同时，也需要考虑到模型的**可解释性**、模型的**可落地性**和**可扩展性**、避免**“黑箱”预测**；
- 可以尝试采用**混合**的机器学习模型，比如GLM+SVR，ARIMA + NNET等；
- 销售预测几乎是商业智能研究的**终极问题**，要解决终极问题还有一段路要走。

践行于大数据最前沿

ArchSummit
全 球 架 构 师 峰 会 2016

[北京站]

主办方 Geekbang. 极客邦科技 InfoQ