ORACLE

ORACLE 21c机器学习

Minming Duan
SEHUB
2021.07.30

Difference between machine learning and AI:
If it is written in Python, it's probably machine learning.
If it is written in PowerPoint, it's probably AI.

—Mat Velloso

Agenda

01

简介

- 特征
- 目标
- 过程

02

OML4SQL

- 产品功能
- 数据模型
- 经典分析

03

算法

- 回归、分类
- 聚类、关联规则
- 预测

04

案例

电商
制造

01

机器学习简介

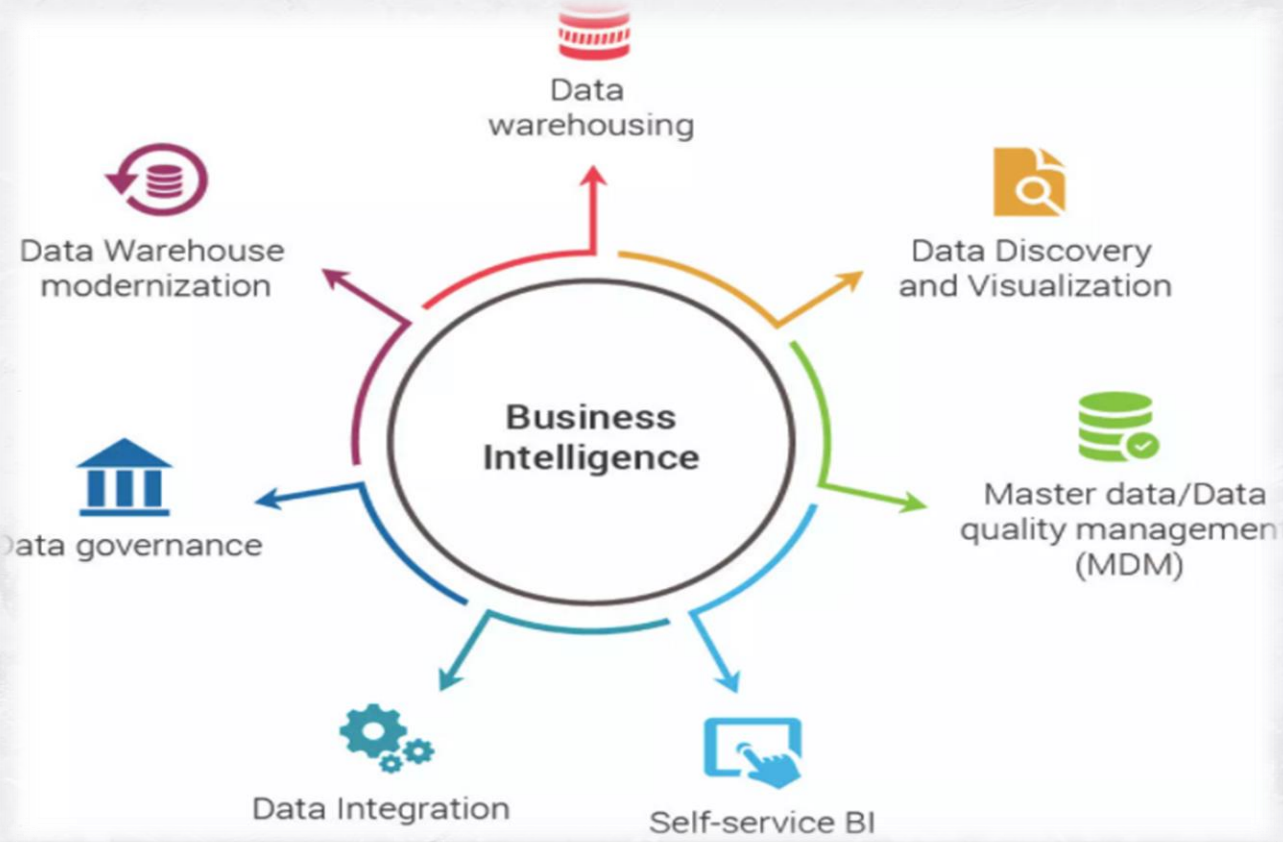
介绍

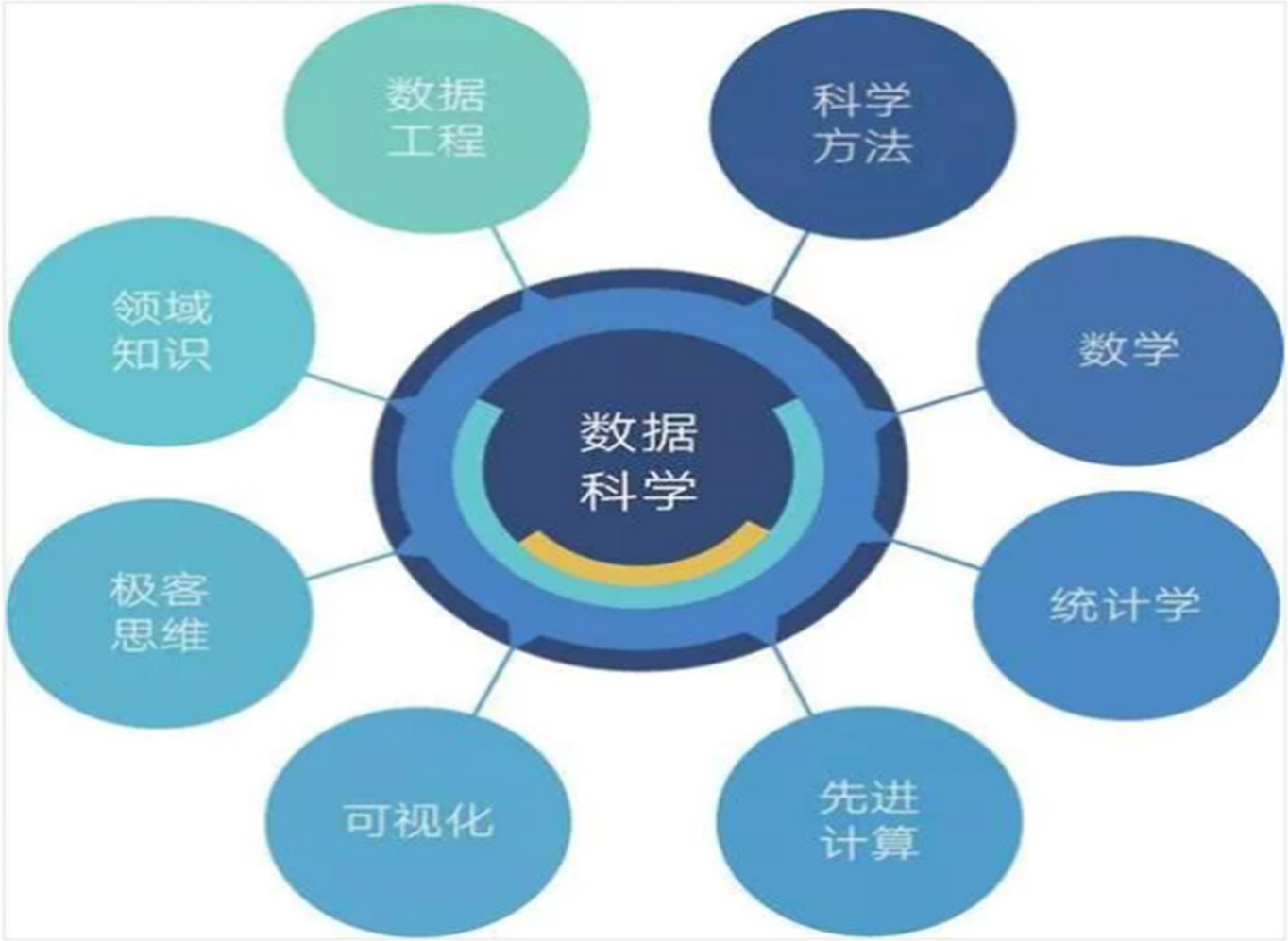
- 发现数据中未知关系的技术
- 解答传统演绎查询和报表技术无法解决的问题

目标

- 自动发现模式
- 预测可能的结果
- 发现可行性信息
- 分析潜在海量数据

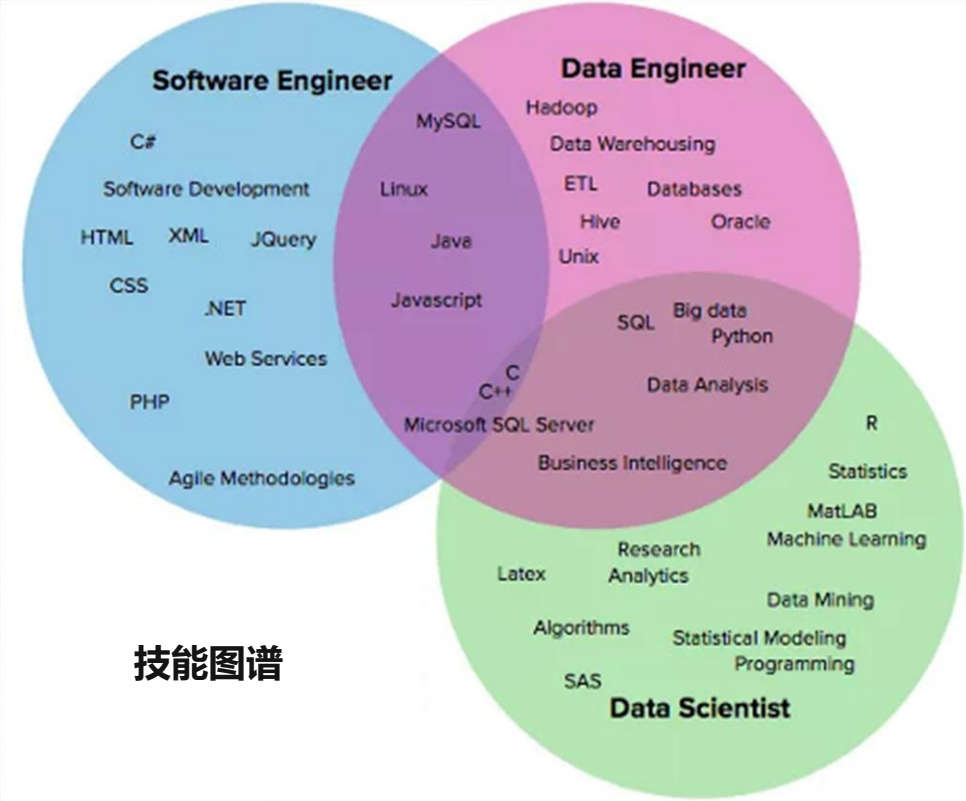
数据科学 VS 商业智能


Data warehousing
Data Warehouse modernization
Data Discovery and Visualization
Business Intelligence
ata governance
Master data/Data quality managemen (MDM)
Data Integration
Self-service BI


数据
工程
科学
方法
领域
知识
数学
数据
科学
极客
思维
统计学
可视化
先进
计算

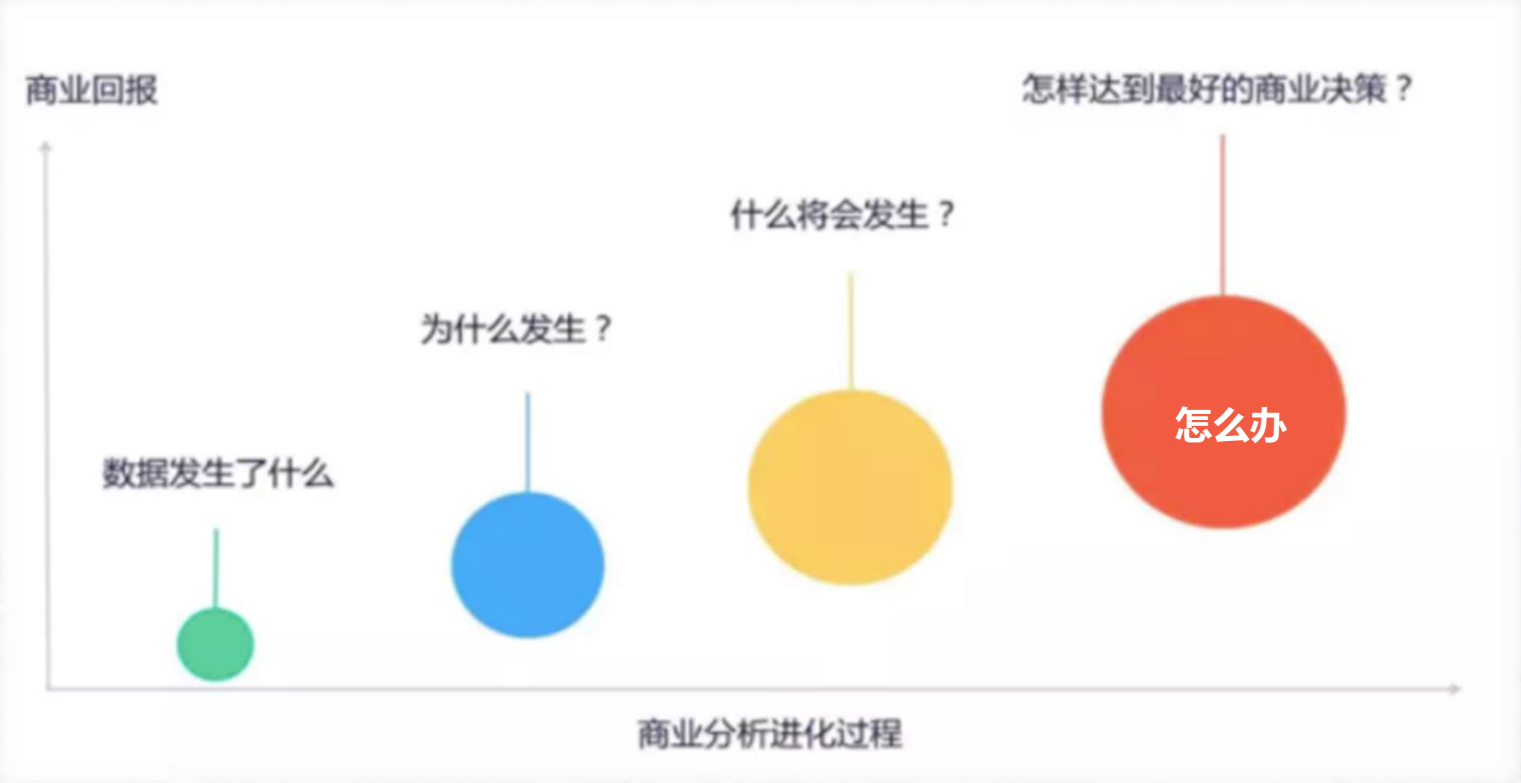
	数据科学家	数据分析师
SKILLSET	• Data Modelling • Predictive Analytics • Advanced Statistics • Engineering/Programming	• BI Tools • Intermediate Statistics • Solid Programming Skills • Regular Expression (SQL)
SCOPE	Macro	Micro
EXPLORATION	• Search Engine Exploration • Machine Learning • Artificial Intelligence • Big data - Often Unstructured	• Data Visualization Techniques • Designing Principles • Big Data - Mostly Structured
GOALS	Discover New Questions to Drive Innovation	Use Existing Information to Uncover Actionable Data

数据科学VS数据VS软件工程师

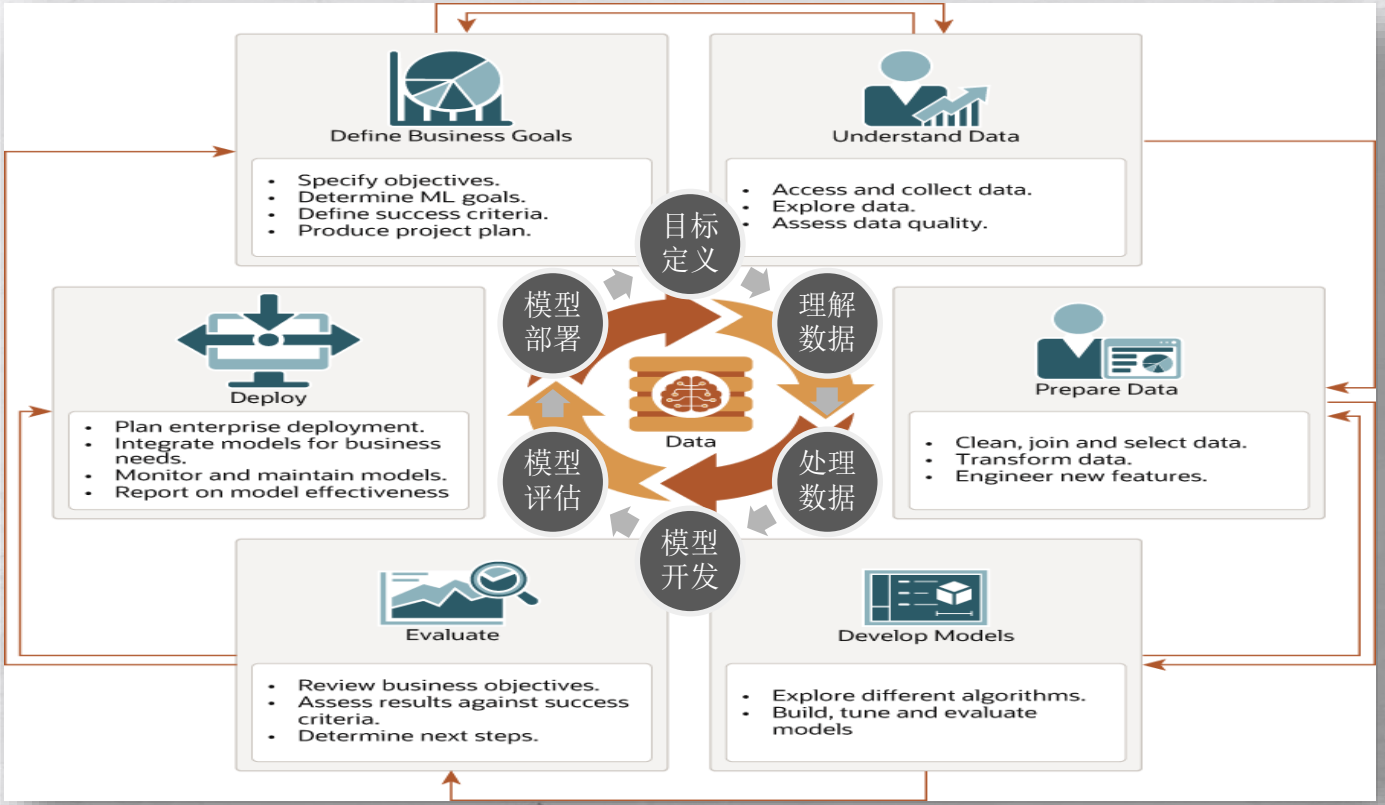


技能图谱

为什么要分析?



机器学习过程


Define Business Goals
Specify objectives.
Determine ML goals.
Define success criteria.
Produce project plan.
Understand Data
Access and collect data.
Explore data.
Assess data quality.
目标定义
理解数据
处理数据
模型开发
模型评估
模型部署
Data
Deploy
Plan enterprise deployment.
Integrate models for business needs.
Monitor and maintain models.
Report on model effectiveness
Prepare Data
Clean, join and select data.
Transform data.
Engineer new features.
Evaluate
Review business objectives.
Assess results against success criteria.
Determine next steps.
Develop Models
Explore different algorithms.
Build, tune and evaluate models

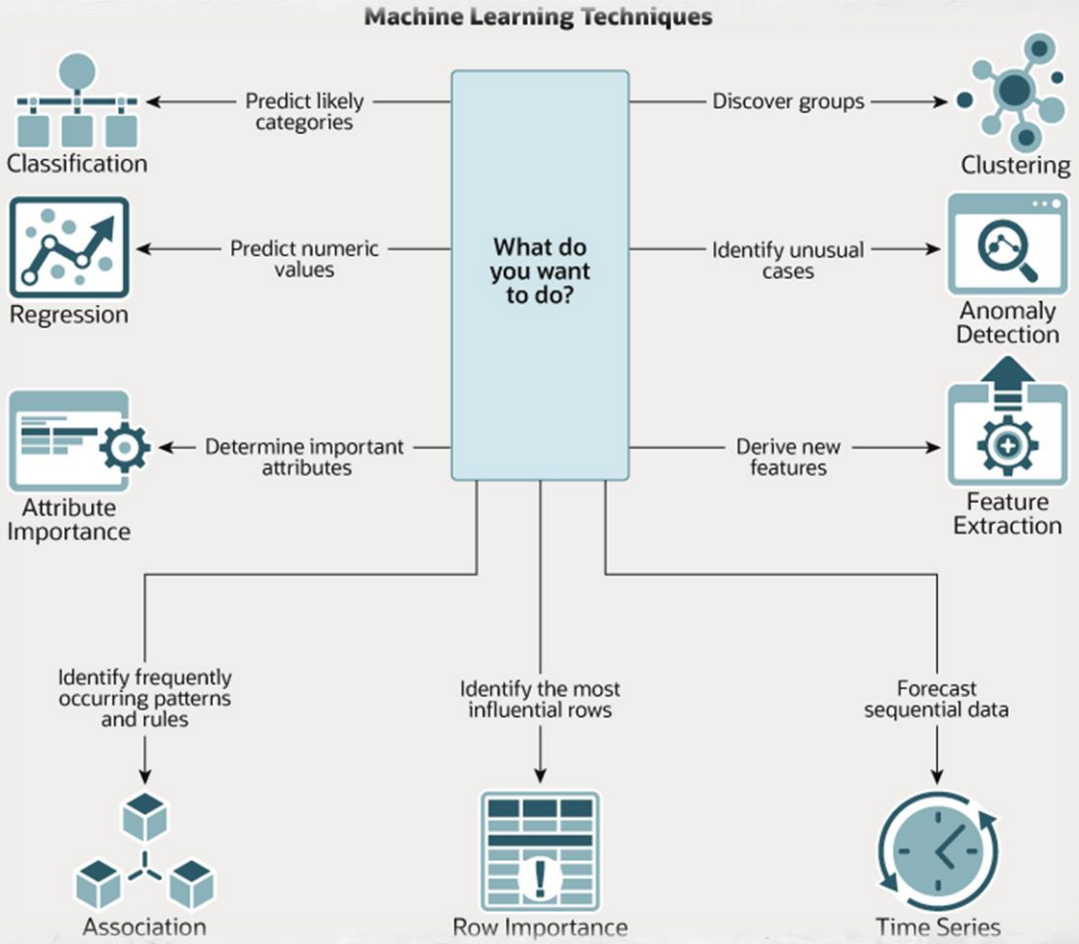
分析思维模型-5W2H


谁？由谁来做？
WHO
何时？什么时间做？什么时机最适宜？
WHEN
WHERE
何处？在哪里做？
是什么？目的是什么？做什么工作？
WHAT
HOW
何怎么做？如何提高效率？如何实施？方法是什么？
为什么要做？可不可以不做？有没有替代方案？
WHY
HOW MUCH
多少？什么程度？数量？质量水平？费用产出？

02

产品介绍

OML4SQL 功能


Machine Learning Techniques
What do you want to do?
Predict likely categories
Classification
Discover groups
Clustering
Predict numeric values
Regression
Identify unusual cases
Anomaly Detection
Determine important attributes
Attribute Importance
Derive new features
Feature Extraction
Identify frequently occurring patterns and rules
Association
Identify the most influential rows
Row Importance
Forecast sequential data
Time Series

模型参数

参数值	描述
ASSOCIATION	描述性的机器学习，识别物体关联关系及其在数据集中出现的可能性（关联规则）。
	关联模型--Apriori算法。
ATTRIBUTE_IMPORTANCE	预测性机器学习，识别属性在预测给定结果中的相对重要性。
	最小描述长度算法和CUR矩阵分解。
CLASSIFICATION	预测性机器学习，用历史数据来预测未知分类。
	朴素贝叶斯，神经网络，决策树，逻辑回归，随机森林，支持向量机，显式语义分析或XGBoost。默认值为朴素贝叶斯。
	异常检测方法：单类SVM模型和多元状态估计技术-顺序概率比测试模型指定分类机器学习。
CLUSTERING	描述性的机器学习，识别数据集中的自然分组。
	聚类模型可以使用k -Means，O-Cluster或Expectation Maximization。默认值为k -Means。
FEATURE_EXTRACTION	描述性的机器学习，创建一组优化的属性。
	非负矩阵分解，奇异值分解（也可以用于主成分分析）或显式语义分析。默认值为非负矩阵分解。
REGRESSION	预测性机器学习，使用历史数据来预测数字型目标变量。
	支持向量机，GLM回归或XGBoost。默认值为支持向量机。
TIME_SERIES	预测性机器学习，预测用户指定的时间窗口内按时间顺序排列的历史数值数据序列的未来值。时间序列模型使用指数平滑算法（ETS）。默认值为指数平滑。

数据模型

有监督学习

分类、时间序列、回归、属性重要性等

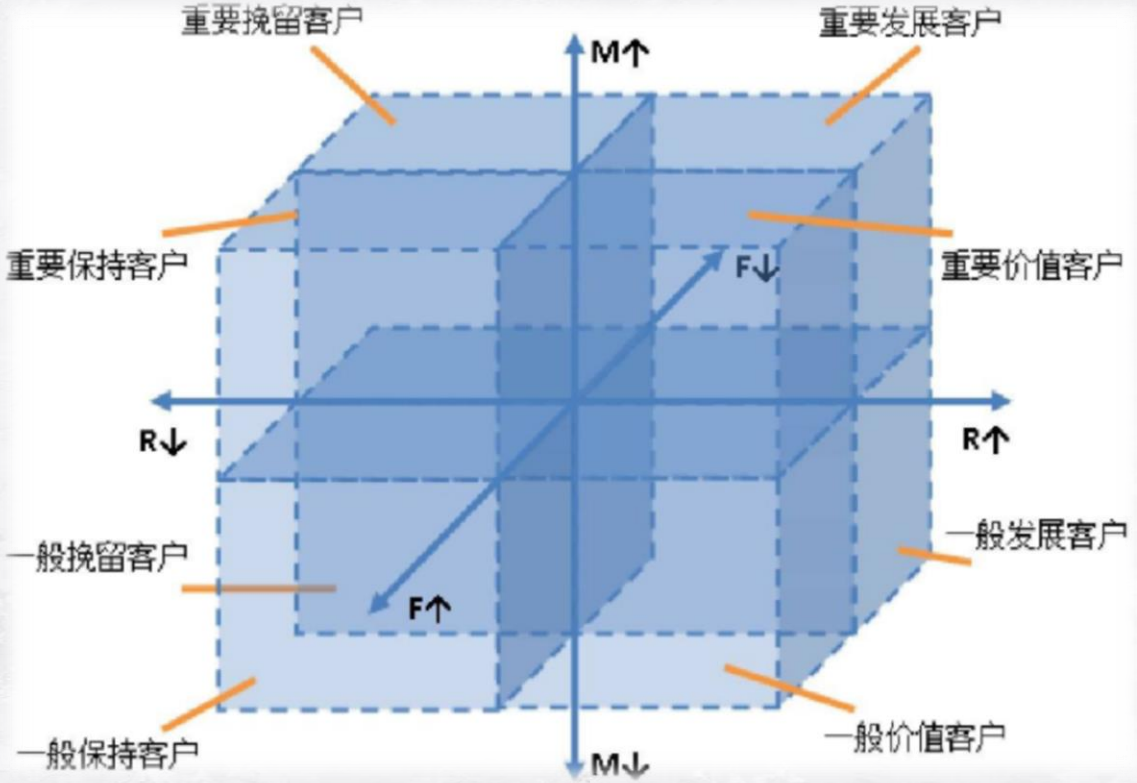
无监督学习

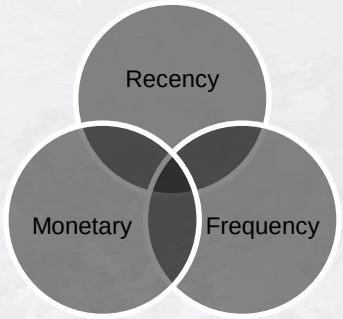
关联规则、聚类、特征提取、因子分析等

MSET

异常检测等

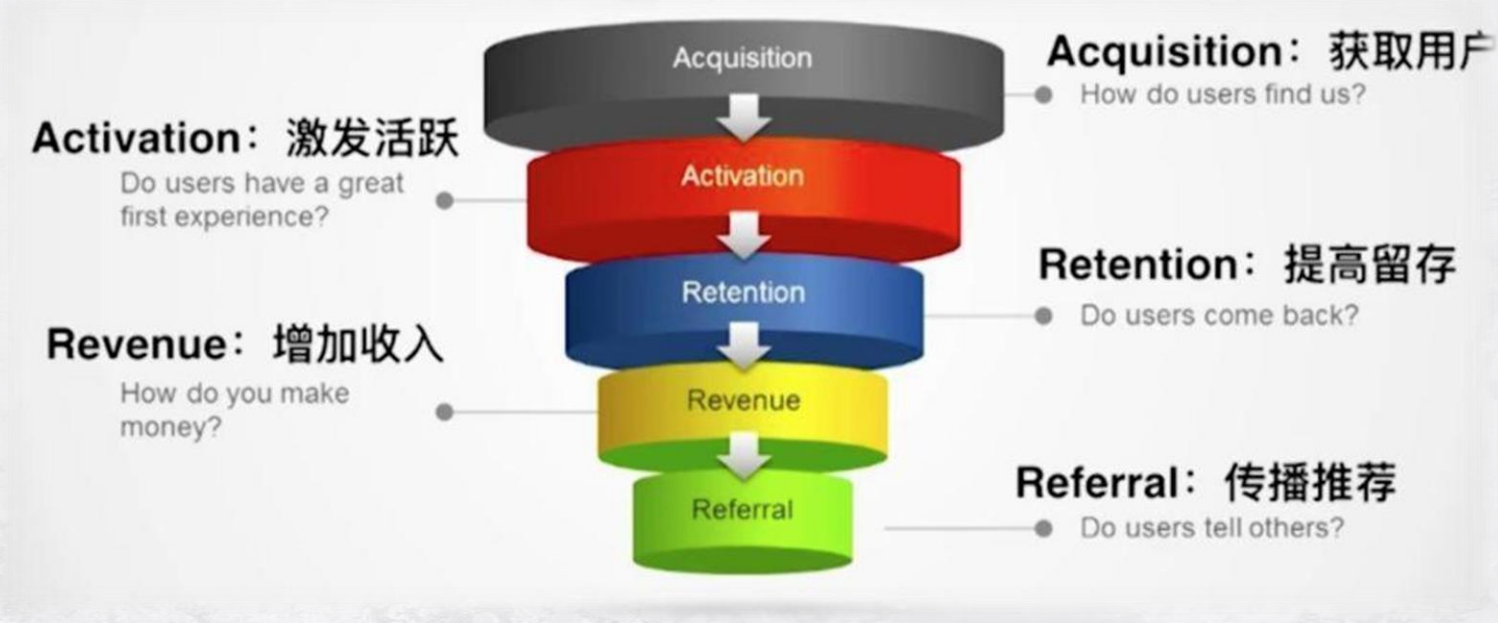
RFM模型





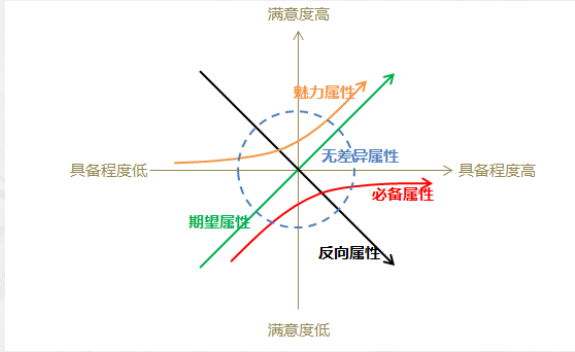
最近一次消费 (Recency)
消费频率 (Frequency)
消费金额 (Monetary)

客户生命周期模型


Acquisition
Activation
Retention
Revenue
Referral
Acquisition：获取用户
How do users find us?
Activation：激发活跃
Do users have a great first experience?
Retention：提高留存
Do users come back?
Revenue：增加收入
How do you make money?
Referral：传播推荐
Do users tell others?

客户满意度模型-KANO

属性名称	属性特征
魅力属性A	超出用户预期的功能/服务，该功能/服务完善程度高，用户满意度会明显上升，如果没有该功能/服务时，用户满意度下降不明显
期望属性O	有某功能/服务会提升满意度，没有会使满意度下降
必备属性M	有某功能/服务不会提升满意度，但没有会使满意度下降
无差异属性I	有和没有某功能/服务均不影响满意度
反向属性R	没有某功能/服务满意度会更高
可疑结果Q	用户没有很好理解某问项或误答



KANO 模型评价结果分类对照表

功能/服务		负问题 不喜欢（1分）	能忍受（2分）	无所谓（3分）	理应如此（4分）	喜欢（5分）
正问题	不喜欢（1分）	Q	R	R	R	R
	能忍受（2分）	M	I	I	I	R
	无所谓（3分）	M	I	I	I	R
	理应如此（4分）	M	I	I	I	R
	喜欢（5分）	O	A	A	A	Q

A：魅力属性，O：期望属性，M：必备属性，I：无差异属性，R：反向属性，Q：可疑属性

正向问题	关于“收到新留言有小红点提示”您的态度是？ □A满意 □B理所应当 □C无所谓 □ D可以忍受 □ E不满意
负向问题	关于“取消收到新留言小红点提示”您的态度是？ □A满意 □B理所应当 □C无所谓 □ D可以忍受 □ E不满意

客户画像

自然属性

性别，年龄

产品订购

订购产品、浏览和偏好

使用行为

业务使用和历史等

社会、职业

教育背景、职业等

交互

投诉、语音交互等

动态属性

兴趣转移、属性更新

03

算法功能

OML算法

回归

- Generalized Linear Model
- Neural Network
- Support Vector Machine
- XGBoost

分类

- Decision Tree
- Explicit Semantic Analysis
- Generalized Linear Model
- Mset
- Naive Bayes
- Random Forest
- Support Vector Machine
- XGBoost

预测

- Exponential Smoothing
- Mset

聚类、规则

- Apriori
- Expectation Maximization
- k-Means
- O-Cluster

算法--当程序员不想解释他们做了什么时，用算法一言以蔽之

OML有监督学习算法

算法	功能	描述
DT	分类	以人类可理解的规则的形式提取预测信息，规则是if-then-else表达式：解释预测规则。
Explicit Semantic Analysis	分类	ESA对文本数据进行预测，可以处理具有数十万个类，归类为特征提取算法。
ETS	时间序列	提供时间序列数据的预测，在用户指定的预测窗口内针对每个时间段进行预测。ESM总共提供14种不同的时间序列模型，包括所有最流行的趋势和季节影响估计，标准时空状态模型。
GLM	分类和回归	数字型目标分类的逻辑回归和连续目标的线性回归，支持预测概率的置信区间和成本估计。
Minimum Description Length	属性重要性	信息理论模型的选择原则，假定最简单，最紧凑的数据表示形式是数据的最佳和最可能的解释。
Naive Bayes	分类	使用贝叶斯定理预测，从数据中观察到的基础证据中得出预测的可能性。
Neural Network	分类和回归	类似生物神经网络人工算法，用于估计或近似依赖于大量通常未知的输入的函数。
RandomForest	分类	强大的机器学习算法，建立了许多决策树模型，并使用树的集合进行预测。
SVM	分类和回归	使用不同的内核函数来处理不同类型的数据集，支持线性和高斯（非线性）内核。 尝试以最大的余量来分离目标类别。 回归试图找到一个连续函数，使最大数据点数位于围绕它的ε单位圆内。
XGBoost	分类与回归	XGBoost用于回归和分类的机器学习算法，调用XGBoost，建立并保持模型并将模型应用于预测。

OML无监督学习算法

算法	功能	描述
Apriori	关联	识别集合中的同时出现的项目（频繁项目集）来执行关联分析，查找支持大于指定最小值的规则支持和置信度大于指定的最小置信度。
CUR Matrix Decomposition	属性重要性	支持向量机（SVM）和主成分分析（PCA）的替代方法，也是进行探索性数据分析的重要工具。算法执行分析处理并挑选出重要的列和行。
EM	聚类	概率聚类的密度估计算法，目标是构建一个密度函数，以捕获给定总体的分布方式，基于观察数据代表样本的密度概率估计。
		支持概率集群和数据频率估计、Expectation Maximization的其他应用程序。
Explicit Semantic Analytics	特征提取	属性向量代表每个特征或概念，ESA创建一个反向索引，该索引将每个属性映射到知识库概念或概念-属性关联向量值。
k-means	聚类	基于距离的聚类算法，可将数据划分为预定数量的聚类。每个簇都有一个质心（重心）。聚类中的个案（种群中的个体）尽可能靠近质心。
		支持k -Means的增强版本，超越了经典的实现集群的层次化父子关系。
MSET	异常检测	非线性的，非参数的异常检测机器学习技术，应用在监控关键过程和质量管控。可检测到细微的异常，将错误警报降至最低。
NMF	特征提取	使用原始属性的线性组合生成新属性，NMF模型会将原始数据映射到该模型发现的一组新的属性（功能）中。
One-Class SVM	异常检测	单类SVM建立一个类的分类模型，识别出与该概要件有所不同的案例。
O-cluster	聚类	基于网格的分层聚类模型，定义了属性空间中的密集区域。灵敏度参数定义基线密度水平。
SVD & PCA	特征提取	奇异值分解（SVD）和主成分分析（PCA）是正交线性变换，捕获数据的基础方差，可降低高维数据的维数和方便数据可视化。
		除了降维之外，SVD和PCA还具有许多其他重要的应用程序，例如数据降噪（平滑），数据压缩，矩阵求逆和求解线性方程组。

OML算法参数

ALGO_NAME Value	Algorithm	Default?	Machine Learning Model Function
ALGO_AI_MDL	Minimum Description Length	—	Attribute importance
ALGO_APRIORI_ASSOCIATION_RULES	Apriori	—	Association
ALGO_CUR_DECOMPOSITION	CUR Matrix Decomposition	—	Attribute importance
ALGO_DECISION_TREE	Decision Tree	—	Classification
ALGO_EXPECTATION_MAXIMIZATION	Expectation Maximization	—	Clustering
ALGO_EXPLICIT_SEMANTIC_ANALYS	Explicit Semantic Analysis	—	Feature extraction and classification
ALGO_EXPONENTIAL_SMOOTHING	Exponential Smoothing	—	Time series
ALGO_EXTENSIBLE_LANG	Language used for an extensible algorithm	—	All machine learning functions are supported
ALGO_GENERALIZED_LINEAR_MODEL	Generalized Linear Model	—	Classification and regression
ALGO_KMEANS	*k*-Means	yes	Clustering
ALGO_MSET_SPRT	Multivariate State Estimation Technique - Sequential Probability Ratio Test	—	Anomaly detection (classification with no target)
ALGO_NAIVE_BAYES	Naive Bayes	yes	Classification
ALGO_NEURAL_NETWORK	Neural Network	—	Classification
ALGO_NONNEGATIVE_MATRIX_FACTOR	Non-Negative Matrix Factorization	yes	Feature extraction
ALGO_O_CLUSTER	O-Cluster	—	Clustering
ALGO_RANDOM_FOREST	Random Forest	—	Classification
ALGO_SINGULAR_VALUE_DECOMP	Singular Value Decomposition (can also be used for Principal Component Analysis)	—	Feature extraction
ALGO_SUPPORT_VECTOR_MACHINES	Support Vector Machine	yes	Default regression algorithm; regression, classification, and anomaly detection (classification with no target)
ALGO_XGBOOST	XGBoost	—	Classification and regression

04

案例分析

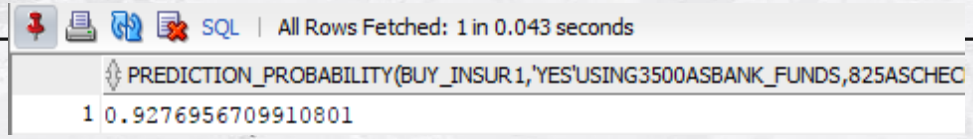
OML建模和预测演示-分类

Model build (PL/SQL)

```
BEGIN
   DBMS_DATA_MINING.CREATE_MODEL(
     model_name           => 'BUY_INSUR1',
     mining_function      => dbms_data_mining.classification,
     data_table_name      => 'CUST_INSUR_LTV',
     case_id_column_name  => 'CUST_ID',
     target_column_name   => 'BUY_INSURANCE',
     settings_table_name  => 'CUST_INSUR_LTV_SET');
END;
```

Real-time scoring (SQL query)

```
SELECT prediction_probability(BUY_INSUR1, 'Yes'
  USING 3500 as bank_funds, 825 as checking_amount, 400 as credit_balance, 22 as age,
  'Married' as marital_status, 93 as MONEY_MONTLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```

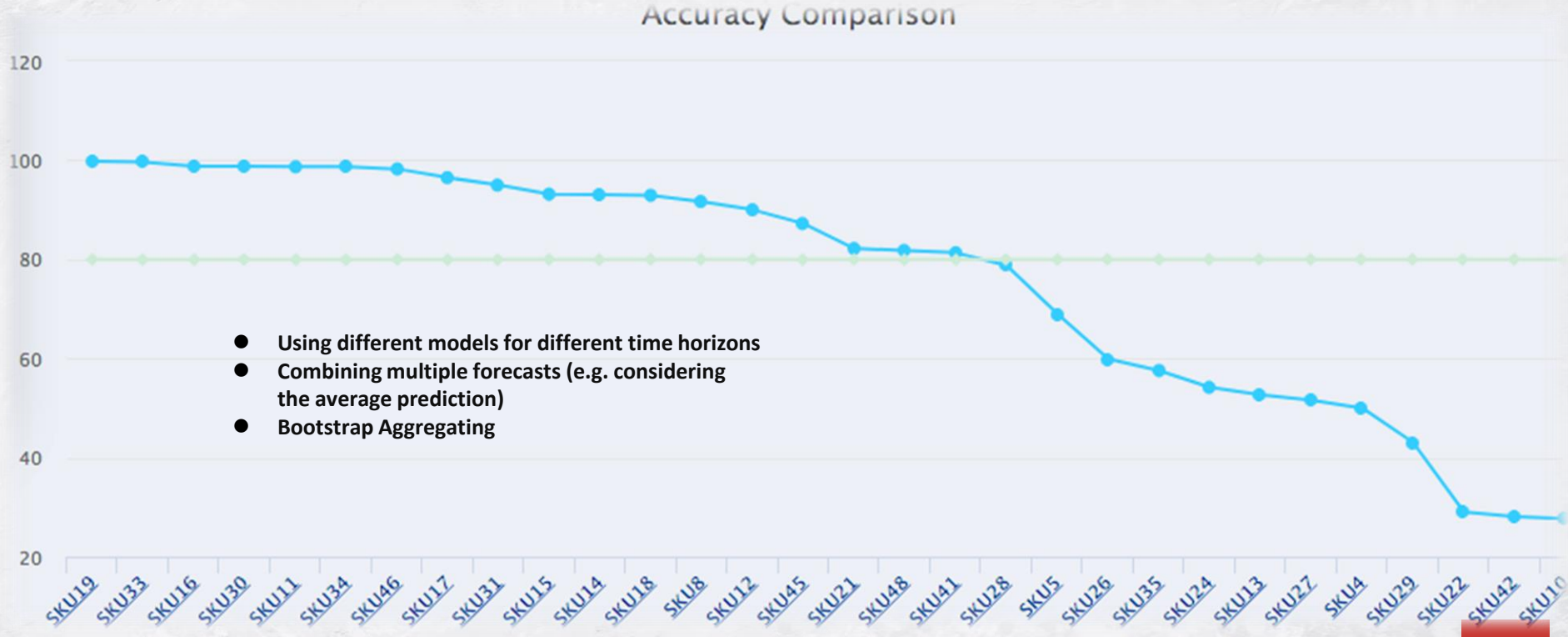


销量预测

Time series forecasting models

Regression	Advanced Method	TBATS	Exponential Smoothing	ARIMA
Dynamic Linear	NNETAR, BAGGED MODEL	BATS, BOX-COX, TBATS	ETS, Holt-winters	AR, MA, ARMA, SARIMA

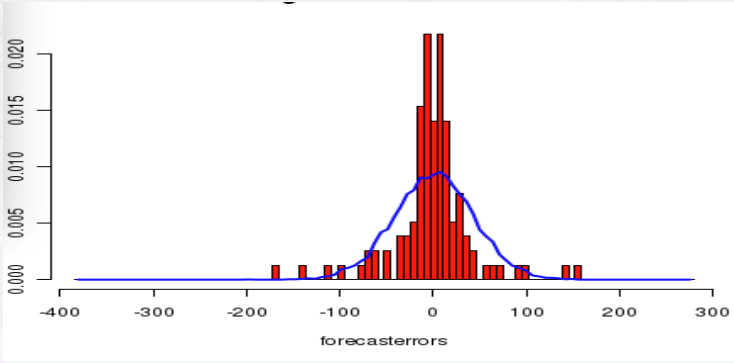
预测结果

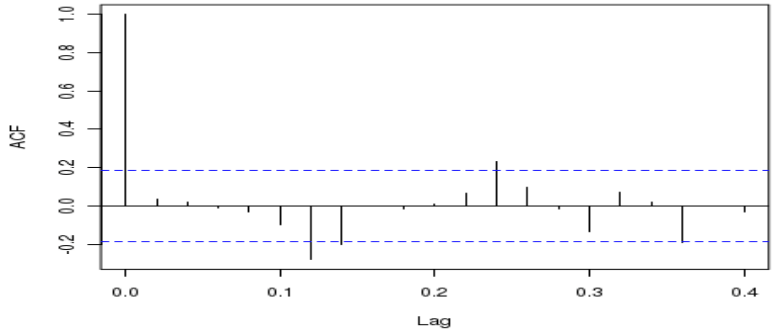


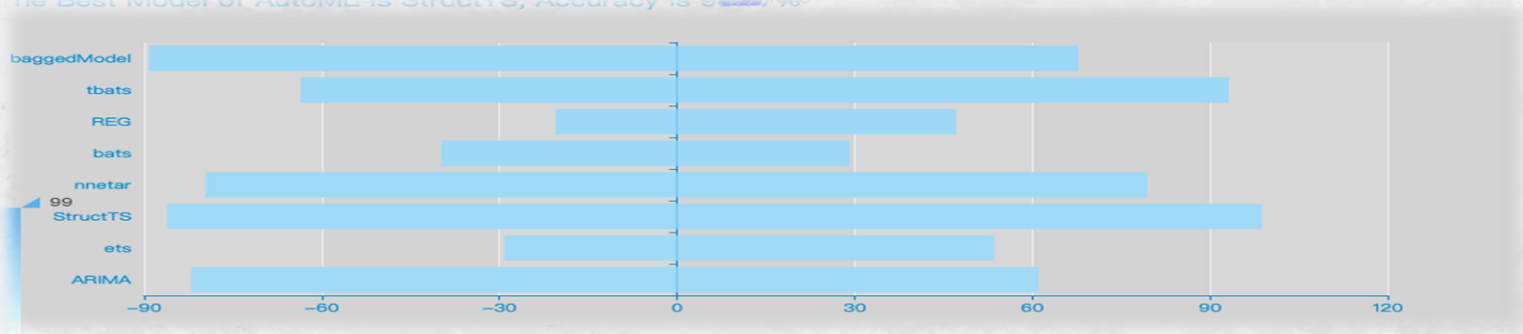
- Using different models for different time horizons
- Combining multiple forecasts (e.g. considering the average prediction)
- Bootstrap Aggregating

Testing Accuracy = 1- |sum(Forecast)/sum(Sale) -1|

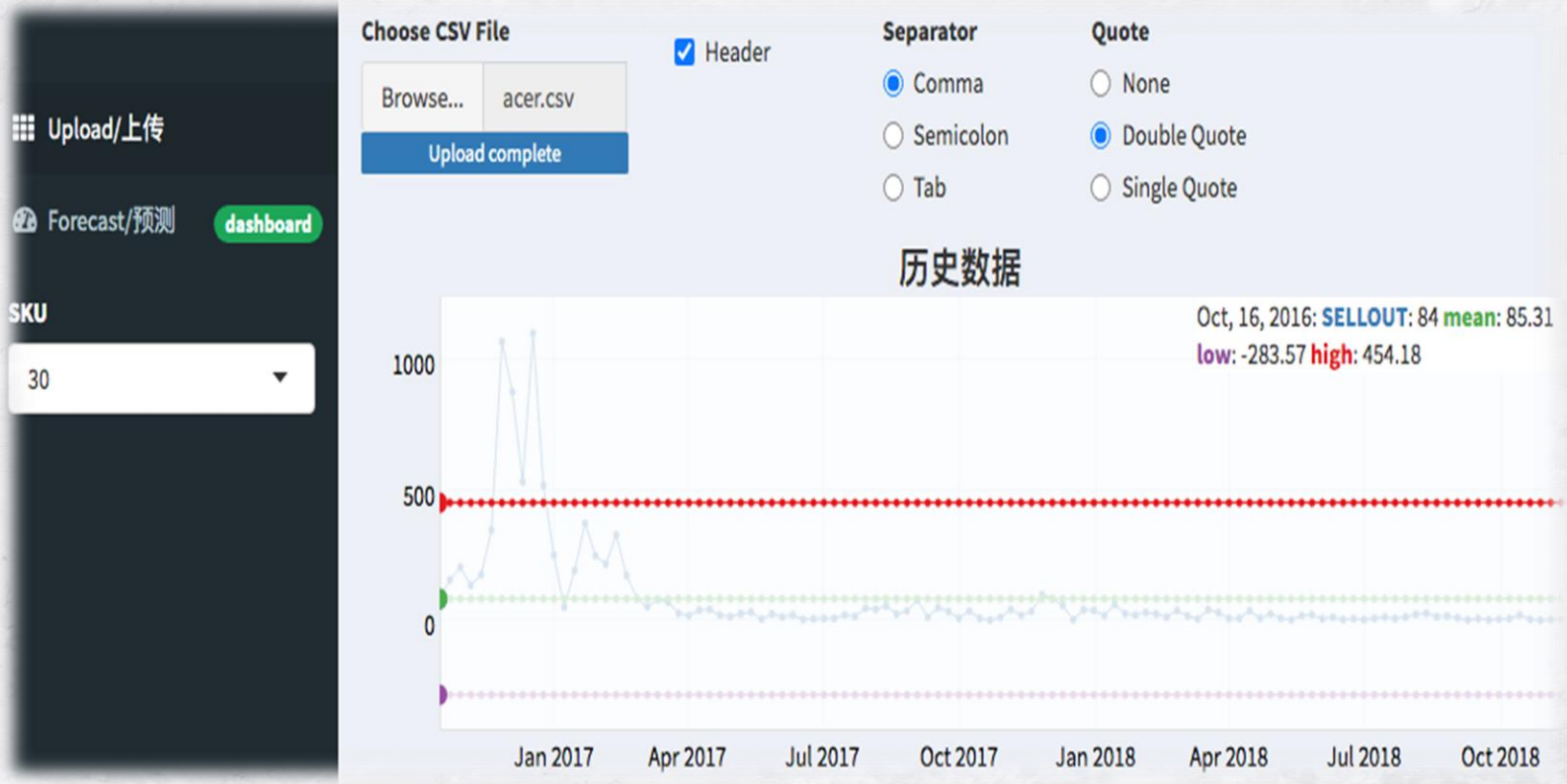
模型评估


0.020
0.015
0.010
0.005
0.000
-400
-300
-200
-100
0
100
200
300
forecasterrors


ACF
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0.0
0.1
0.2
0.3
0.4
Lag

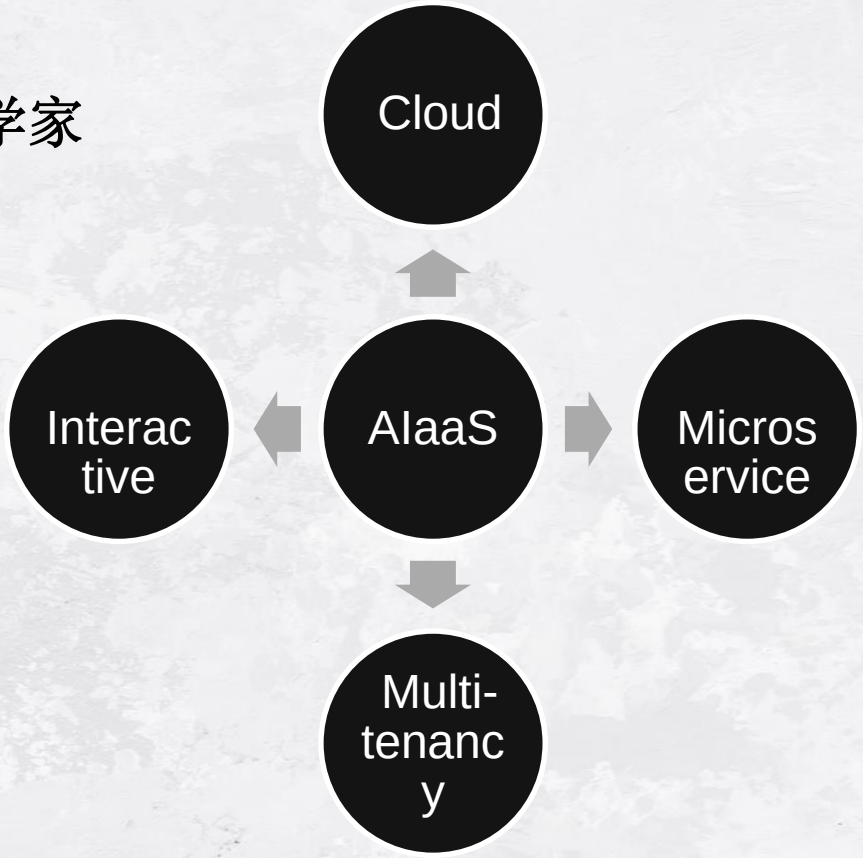

baggedModel
tbats
REG
bats
nnetar
99
StructTS
ets
ARIMA
−90
−60
−30
0
30
60
90
120

Demo

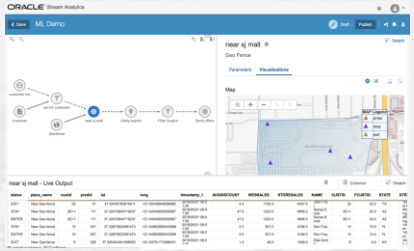

Upload/上传
Forecast/预测
dashboard
SKU
30
Choose CSV File
Browse...
acer.csv
Upload complete
Header
Separator
Comma
Semicolon
Tab
Quote
None
Double Quote
Single Quote
历史数据
Oct, 16, 2016: SELLOUT: 84 mean: 85.31
low: -283.57 high: 454.18
1000
500
0
Jan 2017
Apr 2017
Jul 2017
Oct 2017
Jan 2018
Apr 2018
Jul 2018
Oct 2018

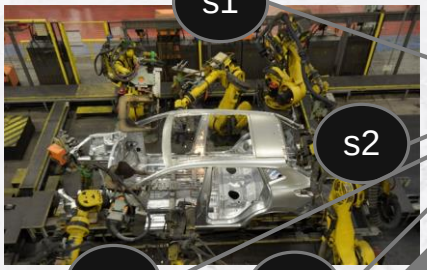
预测方案总结

人人都是数据科学家



高层数据流和诊断分析





s1

s2

s3

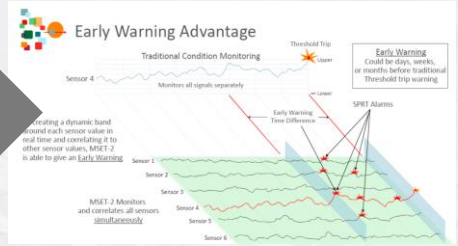
s4

传感器

IoT

MSET

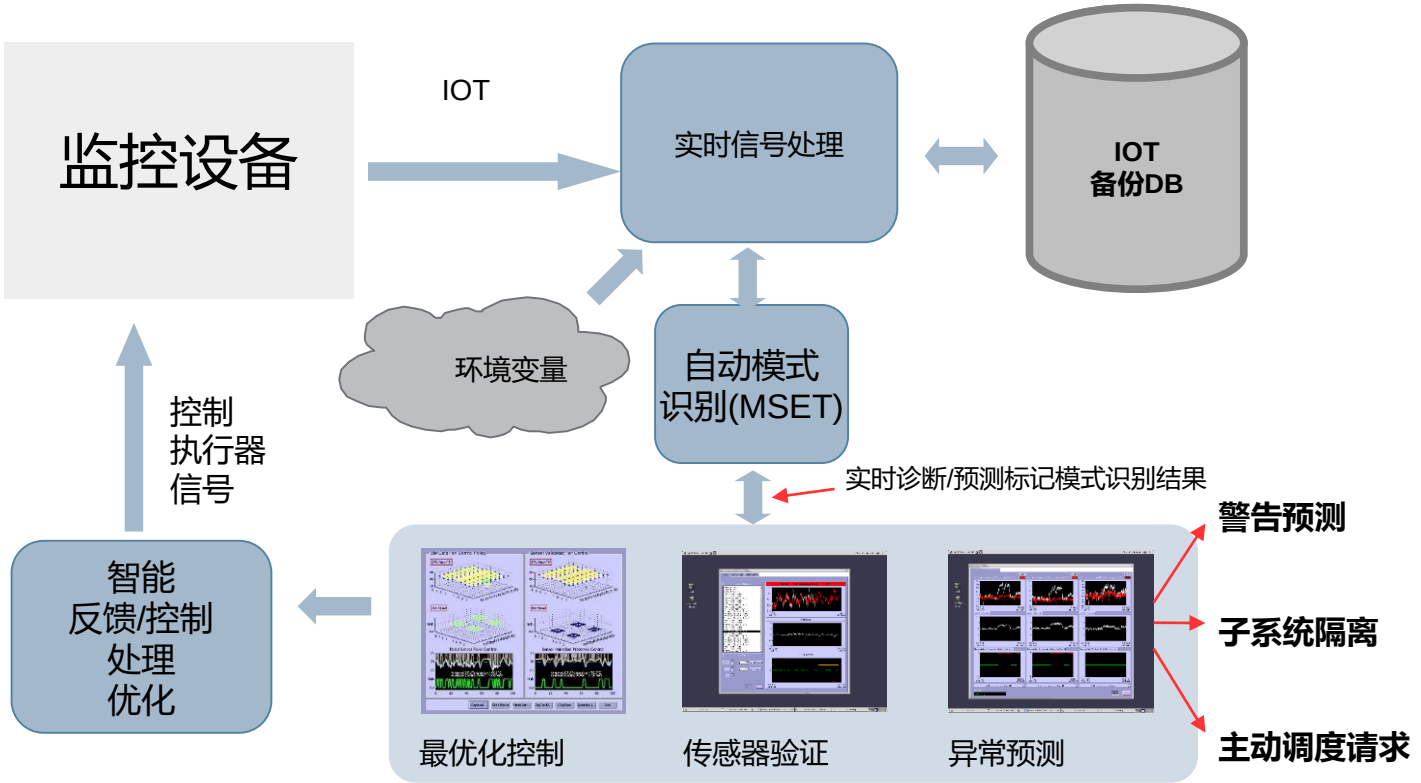
诊断



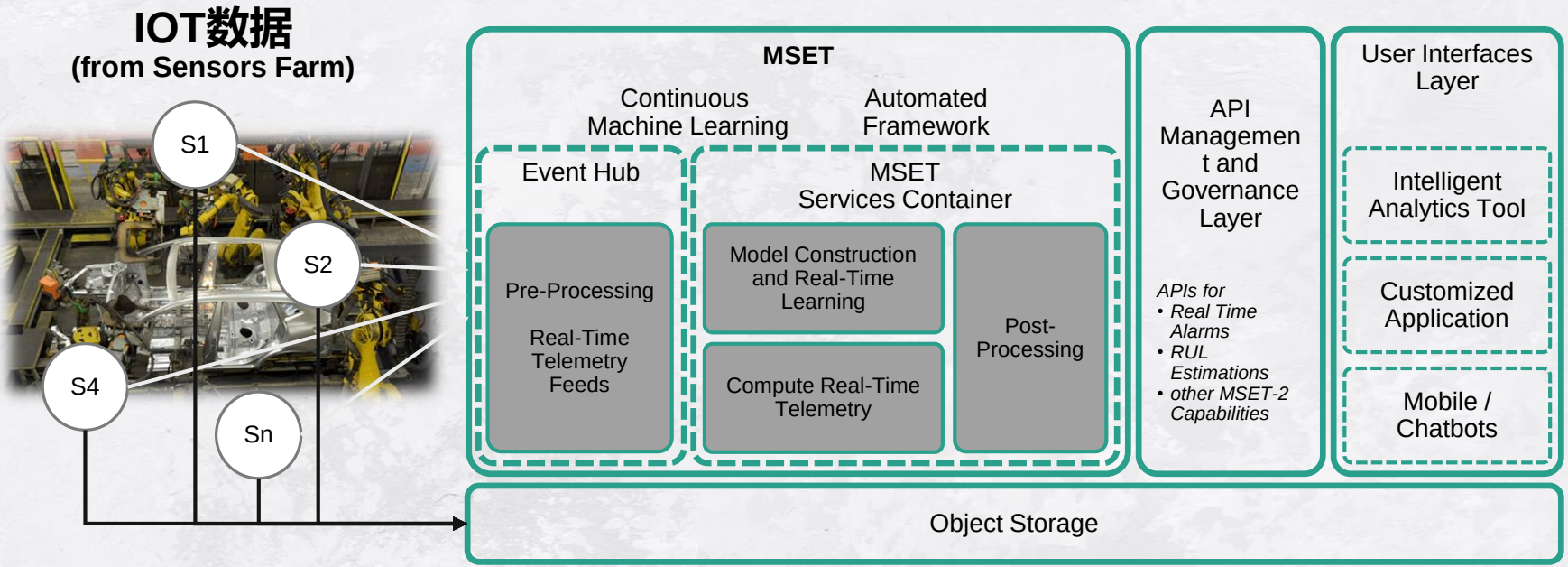
行动← 决策 ← 分析/洞察

自动化运营和预测性运维

模式识别应用

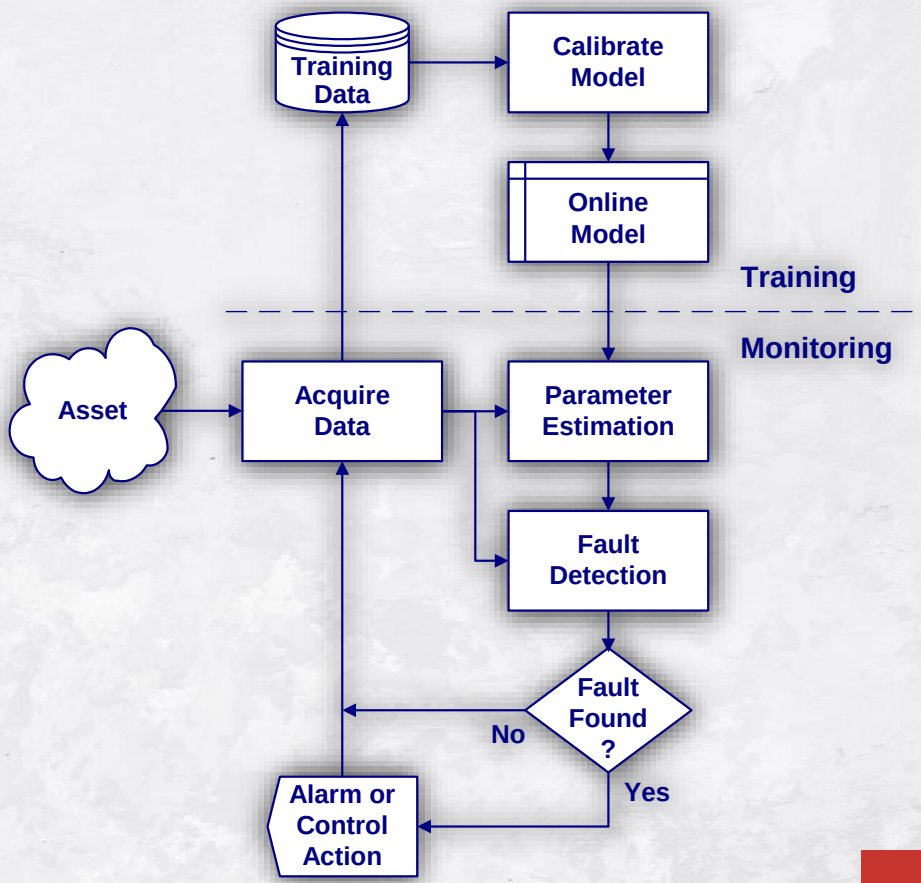

监控设备
IOT
实时信号处理
IOT
备份DB
环境变量
自动模式
识别(MSET)
控制
执行器
信号
实时诊断/预测标记模式识别结果
智能
反馈/控制
处理
优化
警告预测
子系统隔离
主动调度请求
最优化控制
传感器验证
异常预测

制造业高级分析逻辑架构

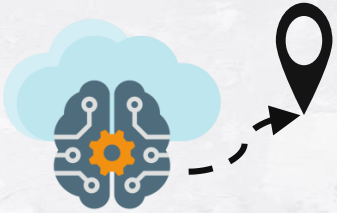


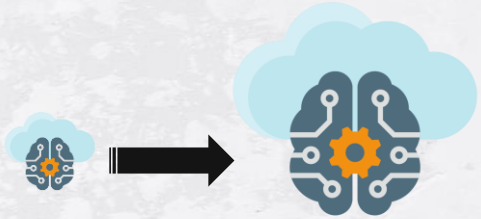
MSET实时应用

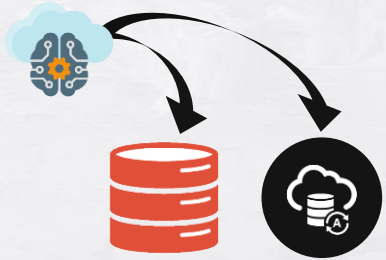

从在线数据中
建立分析模型
实时检测新数
据异常情况
MSET 实时
预警


Training Data
Calibrate Model
Online Model
Training
Monitoring
Asset
Acquire Data
Parameter Estimation
Fault Detection
Fault Found ?
No
Yes
Alarm or Control Action

OML特性







自动

高拓展

高效部署

提高生产力，实现企业目标，更多创新

Thanks!

Do you have any question?