ORACLE

ADW机器学习与案例

邹中凡

资深大数据解决方案架构师

议程

- 什么是机器学习
- ADW机器学习概述
- 原材料价格预测案例
- 零售精准营销案例
- Q&A

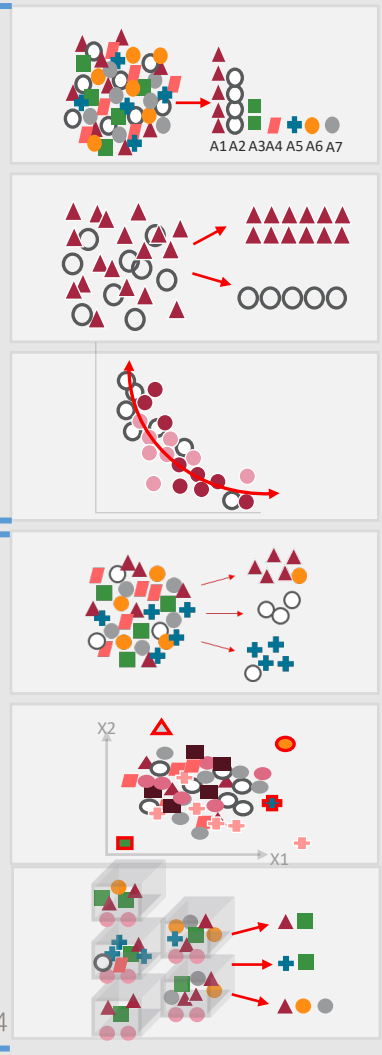
什么是机器学习

什么是机器学习?

算法自动学习大量数据，以发现隐藏的模式、新的洞察并做出预测

监式学习

非监督学习



确定最重要的影响因素(Attribute Importance)

预测客户行为 (Classification)

预测时间趋势 (Time Serials)

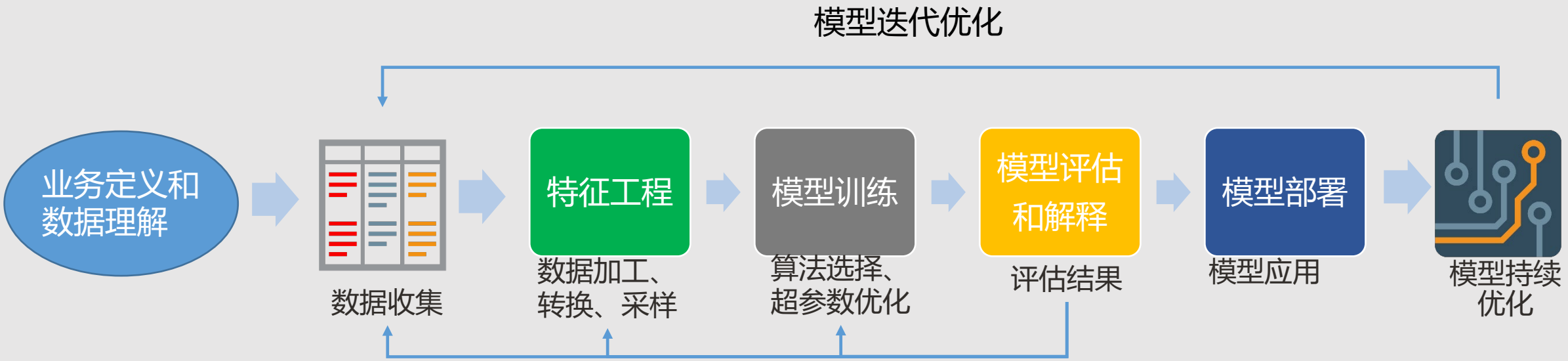
预测或估计数值(Regression)

人员分群(Clustering)

查找欺诈或“罕见事件”(Anomaly Detection)

确定"购物篮"中共同发生的项目(Associations)

机器学习的一般过程



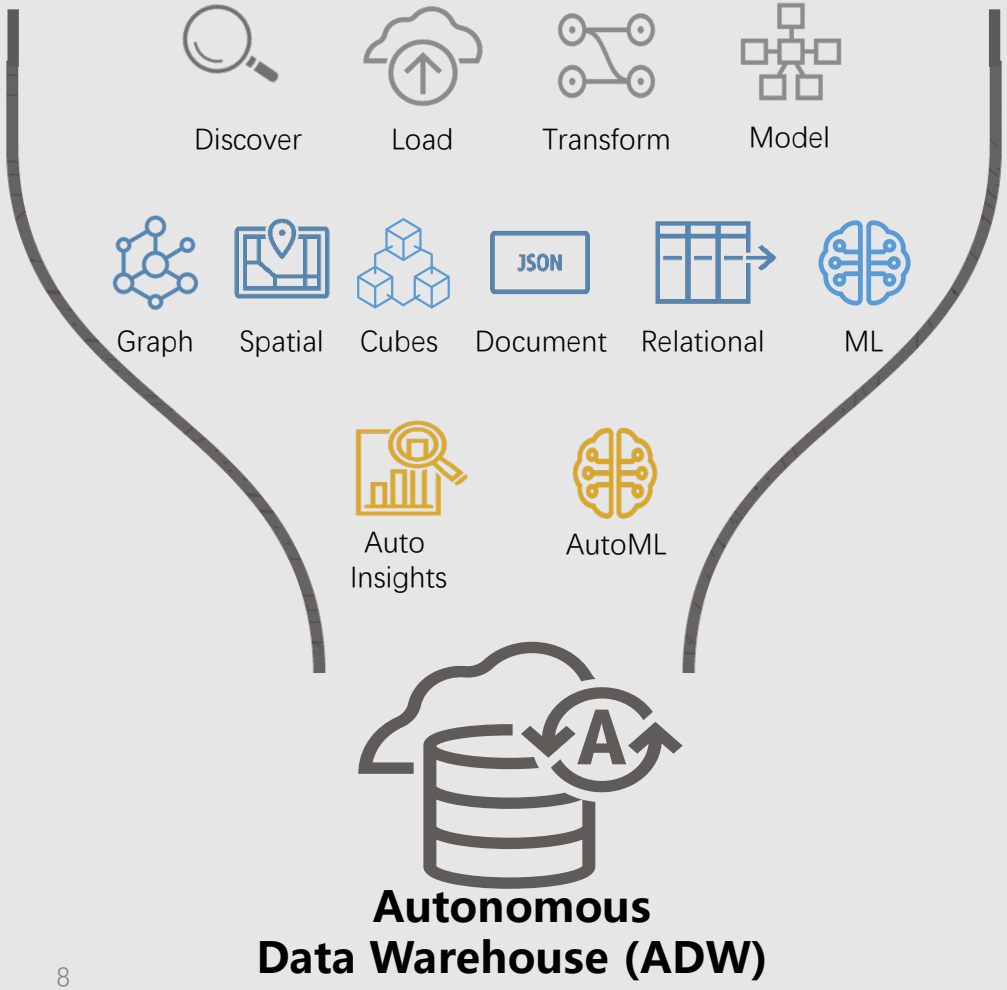
适合机器学习应用的业务场景

电信	消费品	金融服务	公共部门
客户分析 ▪ 客户细分、客户特征分析和行为分析 ▪ 产品组合分析（捆绑） ▪ 系统异常检测 ▪ 客户特征优化	**客户分析** ▪ 客户细分和特征分析 ▪ 客户交互优化 ▪ 购物篮和行为分析 ▪ 客户流失预警 ▪ 客户特征优化	**客户和企业分析** ▪ 收入预测 ▪ 购物篮分析 ▪ 信用风险/策略分析（账龄分析） ▪ 保险欺诈防范	**支出和公众分析** ▪ 支出预测 ▪ 欺诈识别 ▪ 安全/智能分析 ▪ 经济指标预测 ▪ 舆情分析
制造	**医疗保健和生命科学**	**媒体**	**旅游服务**
生产分析 ▪ 质量控制 ▪ 预测性资产维护 ▪ 成本费用分析 ▪ 采购分类 ▪ 人员流失分析	**患者和企业分析** ▪ 患者疗效分析 ▪ 药物开发 ▪ 转化研究（个性化医疗） ▪ 收入和需求预测 ▪ 分配和补货优化	**广告分析** ▪ 广告优化 ▪ 营销优惠优化 ▪ 收入预测 ▪ 价格优化 ▪ 实时优先级分配	**来宾分析** ▪ 细分、获取、保留、交互优化 ▪ 资产绩效分析 ▪ 娱乐场楼层布局优化 ▪ 收入预测和优化

其它应用： 自动驾驶，IT系统自动维护、图像/人脸识别、文字识别、NLP、语音识别等

ADW机器学习概述

什么是ADW-自治数据仓库



多模数据库

关系型、JSON、 XML、 图、 空间、 OLAP、 区块链

多种工作负载

在线交易、分析、机器学习、In-memory、 物联网、流媒体、 多租户、持久内存存储等

多种角色使用-开发人员和分析师

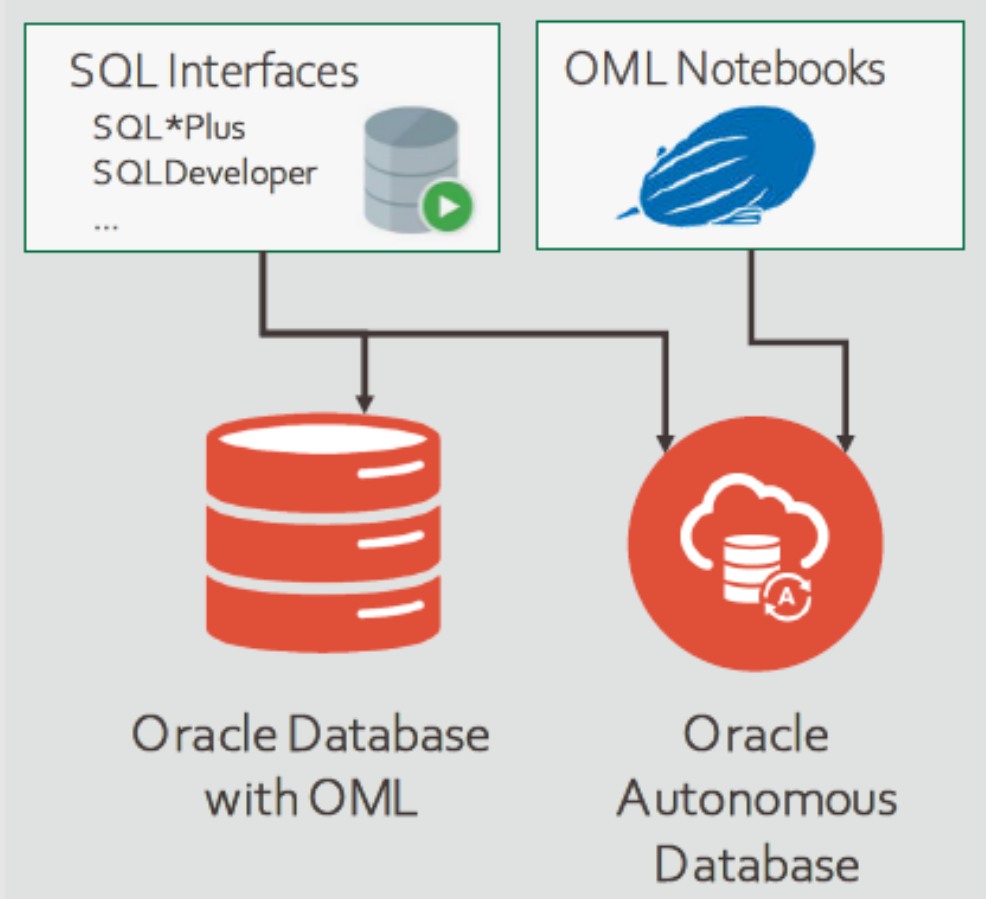
任何数据上的声明式SQL和事务、Java、 JavaScript、ML4SQL 、 ML4Python 、 AutoML、微服务、事件、CI/CD、 APEX

ADW-能做什么

自助式云数据仓库，有机器学习和高级分析的完整套件工具

ADW—机器学习

- 高性能
  - 库内并行的，分布式算法
  - 算法靠近数据，避免数据移动
  - 支持批量和实时评分
- 易用性
  - 提供AutoML自动机器学习
  - 写SQL即可构建机器学习模型
  - NoteBook可视化工具进行机器学习
- 多样性
  - 内置客户分类、异常检测、实时推荐、销售预测等30多种算法
  - 支持SQL、Python构建模型
  - 多种部署使用方式，SQL、Python、REST API等



ADW-自动机器学习

自动建立学习模型，为数据科学家和开发人员提供更快，更轻松的机器学习

- 自动模型选择
  - 识别每个工作负载的最佳预测算法
  - 比穷举搜索更快地找到最佳模型
- 自动特征选择
  - 通过确定最具预测性的特征来减少特征数量
  - 识别最佳预测结果的数据
  - 提高性能和准确性
- 自动优化超参数
  - 大大提高模型准确性
  - 避免使用手动或详尽的搜索技术

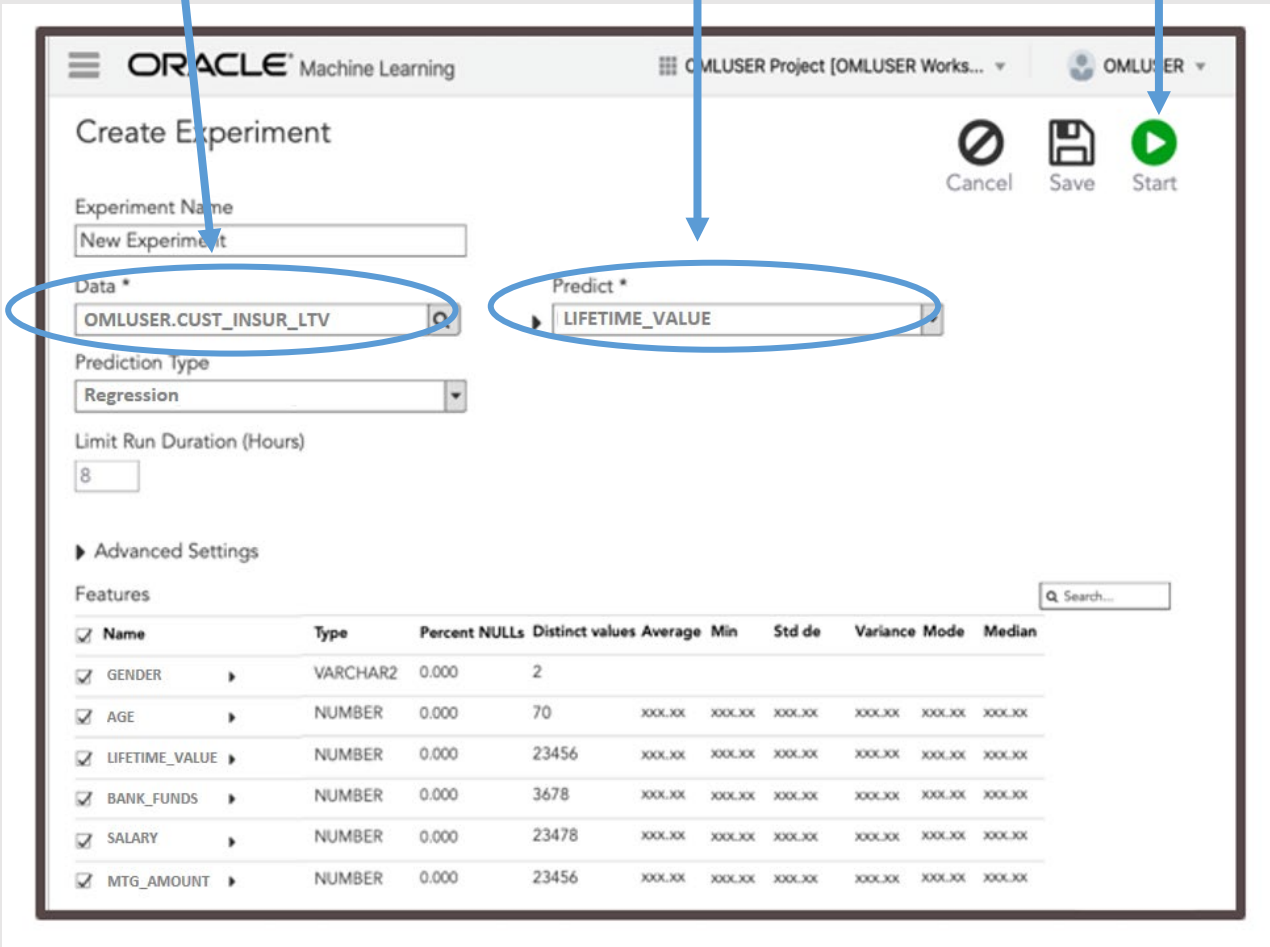
使非专业用户可以利用机器学习

ADW—自动机器学习界面

1.选择数据集

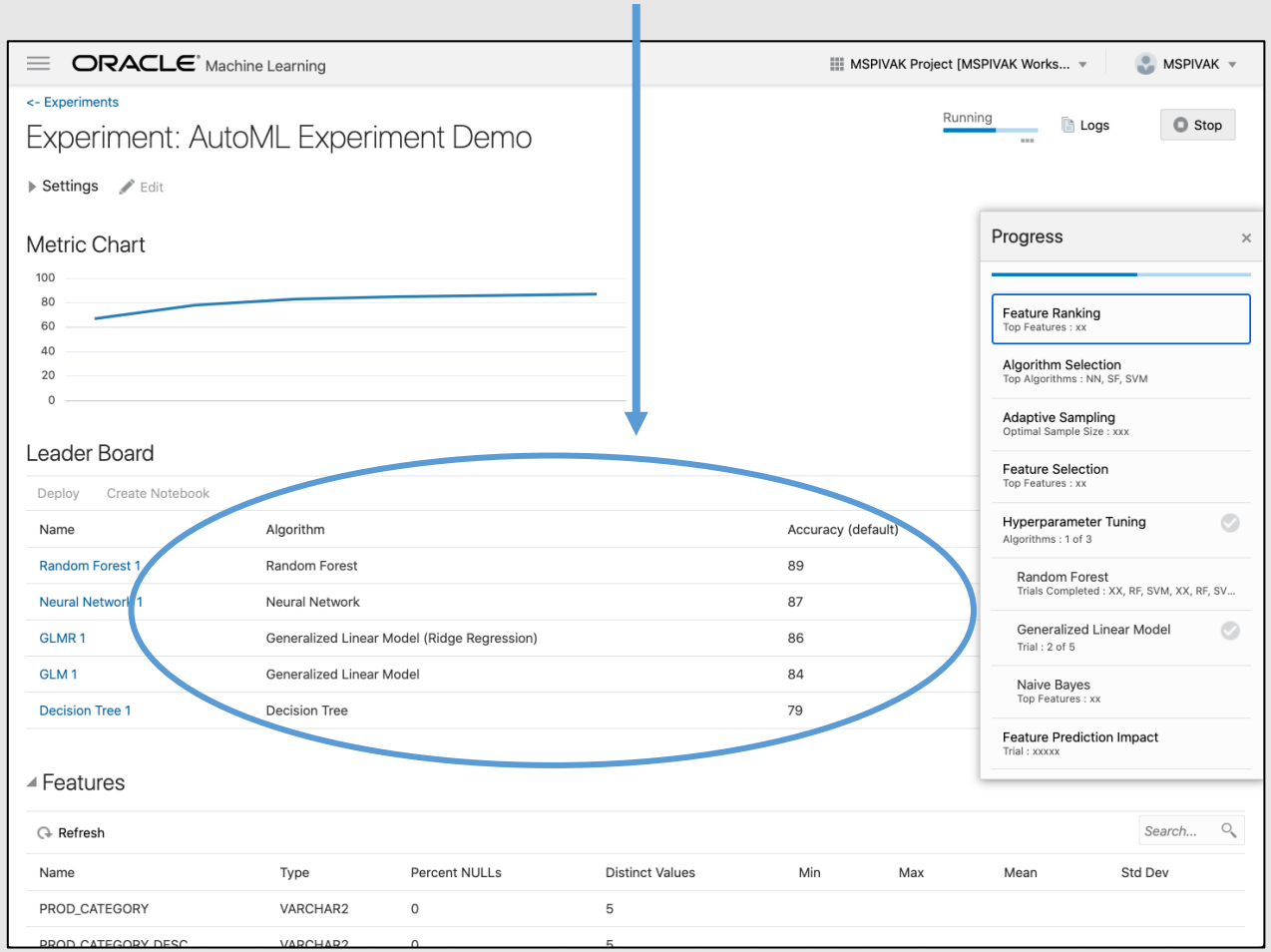
2.选择要预测的数据

3.按开始



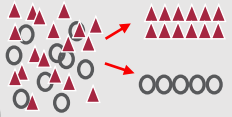
ADW—自动机器学习界面

多个预测算法比较最佳选择



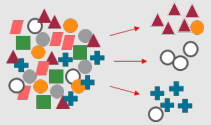
ADW—机器学习算法

分类



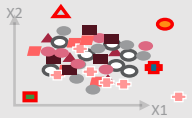
- Naïve Bayes
- Logistic Regression (GLM)
- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine
- Explicit Semantic Analysis
- XGBoost

聚类



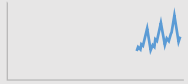
- Hierarchical K-Means
- Hierarchical O-Cluster
- Expectation Maximization (EM)

异常检测



- One-Class SVM
- *MSET-SPRT*

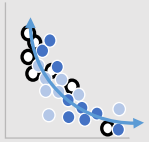
时间序列



- Forecasting - Exponential Smoothing
- Includes popular models

e.g. Holt-Winters with trends, seasonality, irregularity, missing data

回归

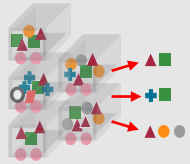


- Linear Model
- Generalized Linear Model
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- LASSO
- XGBoost

属性重要性

- Minimum Description Length
- Principal Comp Analysis (PCA)
- Unsupervised Pair-wise KL Div
- CUR decomposition for row & AI

关联规则



- A priori/ market basket

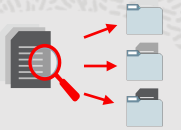
预测查询

- Predict, cluster, detect, features

SQL 分析

- SQL Windows
- SQL Patterns
- SQL Aggregates

特征提取

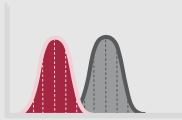


- Principal Comp Analysis (PCA)
- Non-negative Matrix Factorization
- Singular Value Decomposition (SVD)
- Explicit Semantic Analysis (ESA)

文本挖掘

- Algorithms support text
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA) for document similarity

统计分析



- Basic statistics: min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.

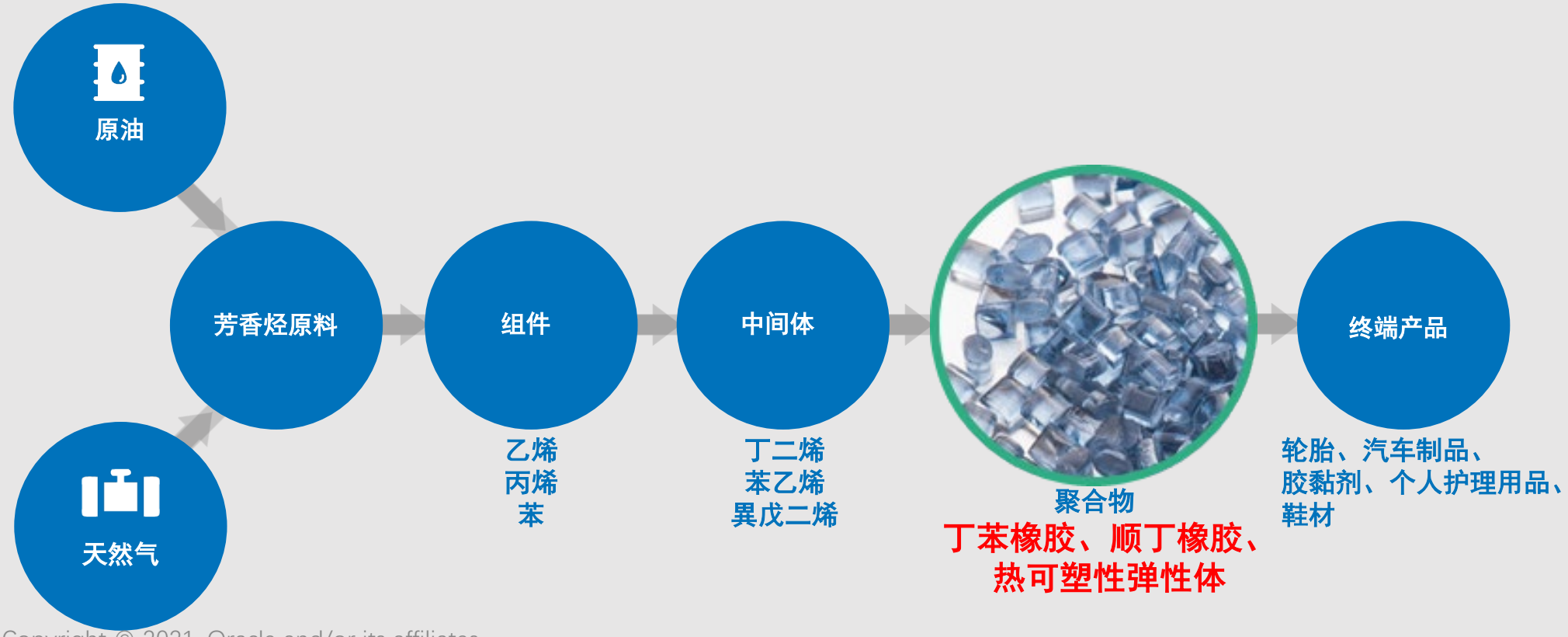
模型部署

- SQL—1st Class Objects
- Oracle RESTful API (ORDS)
- OML Microservices (for Apps)

原材料价格预测案例

案例背景

客户是台湾一家专注于橡胶产品制造企业，希望借助机器学习方法进行丁二烯原材料价格的预测。

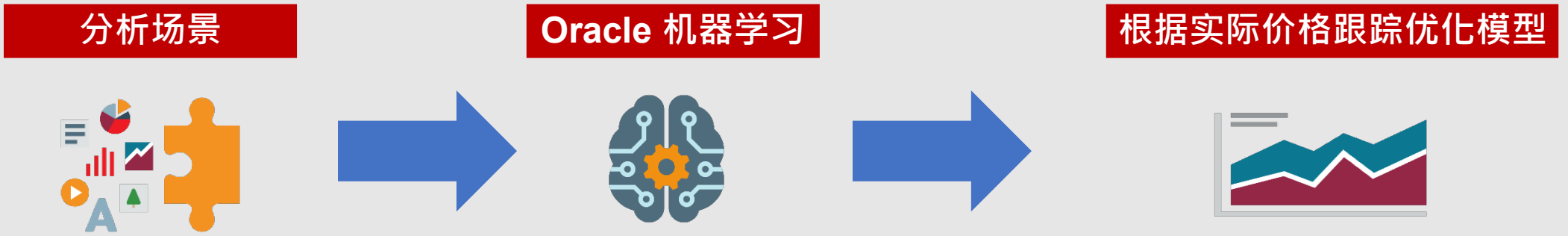


现状—人工预测

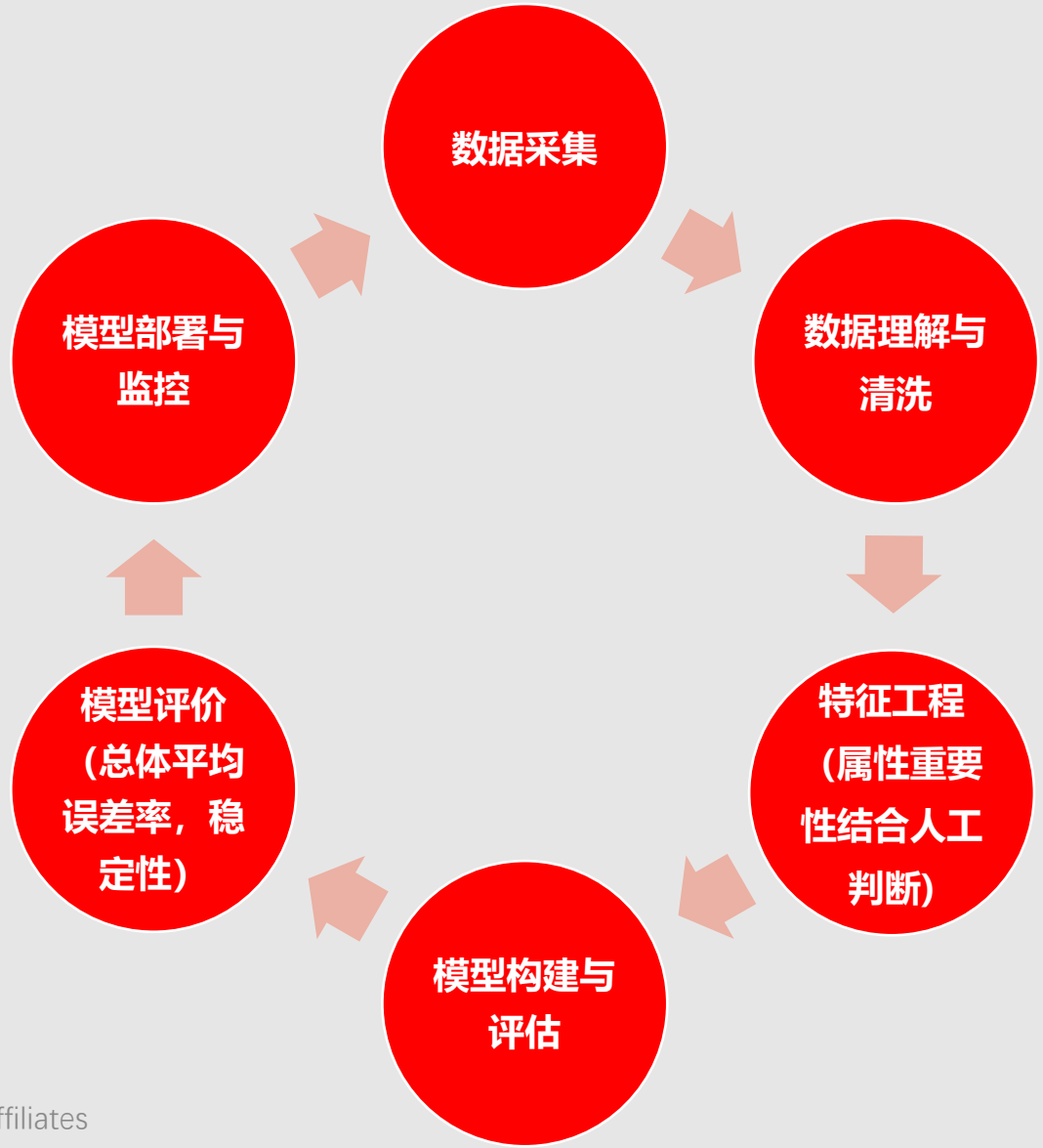
- 采购人员通过**历史资料过往经验**，估算各原物料价格的涨跌变化与价格区间
- 目前原物料价格没有存入数据库，以及原物料价格的涨跌变化较大，**难以稳定**预测原材料涨跌趋势

业务场景—算法预测

- 通过机器学习算法，找到过往历史数据的**关键因子**，借助识别的关键因子，来预测原材料价格的涨跌变化
- 保存预测结果数据，与实际原材料实际价格进行比较，**追踪成效**以及**修正预测模型**结果



建模方法



数据采集—原始数据样例

栏位	数据说明
原始数据提供方式	Excel
原始数据类型	Platts普氏价格指数，反应不同原材料在不同地区的报价，ICIS价格指数，原油价格指数等
数据粒度	每周一条
数据问题	有些周数据缺失，有些列数据缺失

数据样例

Sources	PLATTS			
Product	Benzene			
	FOB Korea			
Unit	$/mt			
Date	下限	上限	平均	下
2010/1/15	1,028	1,029	1,028	1,04
2010/1/22	1,048	1,049	1,048	1,06
2010/1/29	931	932	932	94
2010/2/5	932	933	933	94
2010/2/12	897	898	898	91
2010/2/19	957	958	958	97
2010/2/26	908	909	909	92
2010/3/5	874	875	874	89
2010/3/12	920	921	920	92
2010/3/19	958	959	959	96
2010/3/26	977	978	977	97
2010/4/2	984	985	985	98
2010/4/9	996	997	997	1,00
2010/4/16	1,003	1,004	1,003	1,00

Sources	Platts	Nymex	ICIS			ICIS			ICIS			
Product		Crude Oil	Butadiene			Butadiene			Butadiene			
	EUR/USD	WTI近月	FD NWE CP			FOB ARA SP			FD NWE INLAND SP			
Unit			EUR/TONNE		EUR/TONNE	USD/TONNE			EUR/TONNE		EUR/TON	USD/TON
Date			下限	上限	平均	下限	上限	平均	下限	上限	平均	平均
2013/3/8	0.74	917.68	1,415	1,415	1,415	1,600	1,650	1,625	1,365	1,415	1,390	
2013/3/15	0.74	917.77	1,415	1,415	1,415	1,600	1,650	1,625	1,370	1,450	1,410	
2013/3/22	0.74	911.89	1,415	1,415	1,415	1,600	1,650	1,625	1,415	1,470	1,443	
2013/3/29	0.74	910.29	1,415	1,415	1,415	1,600	1,650	1,625	1,415	1,480	1,448	
2013/4/5	0.74	847.28	1,415	1,415	1,415	1,600	1,650	1,625	1,415	1,480	1,448	
2013/4/12	0.74	827.88	1,415	1,415	1,415	1,600	1,650	1,625	1,415	1,480	1,448	
2013/4/19	0.74	815.86	1,415	1,415	1,415	1,600	1,650	1,625	1,415	1,480	1,448	
2013/4/26	0.74	819.25	1,415	1,415	1,415	1,600	1,650	1,625	1,415	1,480	1,448	
2013/5/3	0.74	813.28	1,340	1,340	1,340	1,600	1,650	1,625	1,340	1,400	1,370	
2013/5/10	0.74	834.38	1,340	1,340	1,340	1,600	1,650	1,625	1,340	1,350	1,345	
2013/5/17	0.74	839.18	1,340	1,340	1,340	1,500	1,550	1,525	1,340	1,350	1,345	

数据理解与清洗

92个分析因子

559笔数据（每周一条）

目标
丁二烯价格

PLATTS
Butadiene
CFR TWN SP

PLATTS Asia 影响因子

PLATTS	PLATTS	PLATTS	PLATTS
Propylene	Propylene	Propylene	Butadiene
CFR SE Asia	FOB Japan	CFR China	FOB KOR SP

训练数据区间
2010-01-08 ~
2016-12-16
361 笔数据
验证数据区间
2017-01-06~
2020-10-30
198笔数据

ICIS 影响因子

ICIS	ICIS	ICIS	ICIS	ICIS	ICIS
Butadiene	Butadiene	Butadiene	Butadiene	Butadiene	Butadiene
TPC, Lyondell, Shell CP	ExxonMobil CP	CIF USG SP	YNCC CP	FPCC CP	CFR NE Asia SP

原油 影响因子

Nymex	
Crude Oil	Europe Naphtha

特征工程—属性重要性

应用属性重要性算法，选择如下字段作为训练模型的因子：

Nymex Crude Oil
Europe Naphtha
Crude C4 FOB NWE SP
Crude C4 CIF USG SP
Crude C4 FOB USG FORMULA
BLOCK COPOL CFR SE Asia
PTA CFR China HIPS CFR SE Asia
HDPE YARN CFR SE Asia
PVC SUSP CFR SE Asia
Benzene CFR Taiwan
LLDPE CFR SE Asia
HDPE INJ CFR SE Asia
ABS INJ CFR SE Asia
PP INJ CFR SE Asia
OX CFR SE Asia
LLDPE Metallocene C6
CFR SE Asia

模型构建与模型评估

模型构建：

1，构建时间序列
- ✓ Exponential Smoothing

2，构建线性回归
- ✓ GLM
- ✓ SVM

3，构建非线性回归
- ✓ GLM，打开特征选择和特征生成
- ✓ SVM，引入核函数

模型评估：
- ✓ 基于测试集数据应用各个模型，并计算各个模型的R方，找到最优的模型

4，运行ADW AutoML

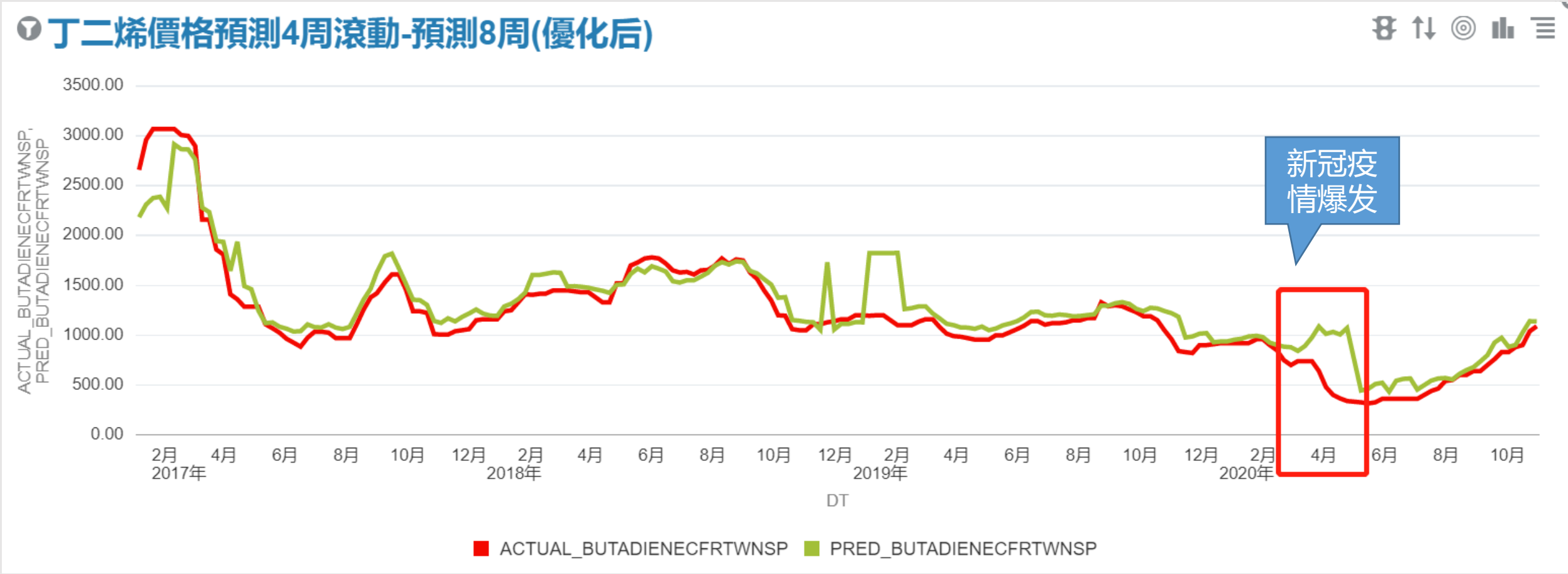
排行榜

部署　建立記事本　度量

演算法	模型名稱	R2
支援向量機 (高斯)	svmg_ac1a34afa8	0.9658
神經網路	nn_739a034159	0.9602
廣義線性模型 (脊迴歸)	glmr_5faf1323e3	0.7279
廣義線性模型	glm_429fb3f931	0.7196
支援向量機 (線性)	svml_bf9bf35980	0.6626

模型结果—丁二烯价格预测價格預測 (4周滚动-预测8周)

从2017年1月开始每四周滚动预测未来8周的价格，一共预测了25个批次



注：红色是实际值，军绿色是预测值

模型部署与监控

总结

1，我们基于ADW，预测模型在准确率、稳定性等方面优于台湾本地专业做数据分析的厂商。我们得到客户的认可，并赢单。

2，在机器学习的整个过程，包括数据集成、清洗、模型建立、评估、部署上线都是在ADW中完成。

3，构建一个性能好、稳定的模型需要不停地迭代尝试。

4，生产上线后，需要持续监控模型的运行情况，观察模型的准确率和稳定性，并持续优化。

零售精准营销案例

案例背景

客户

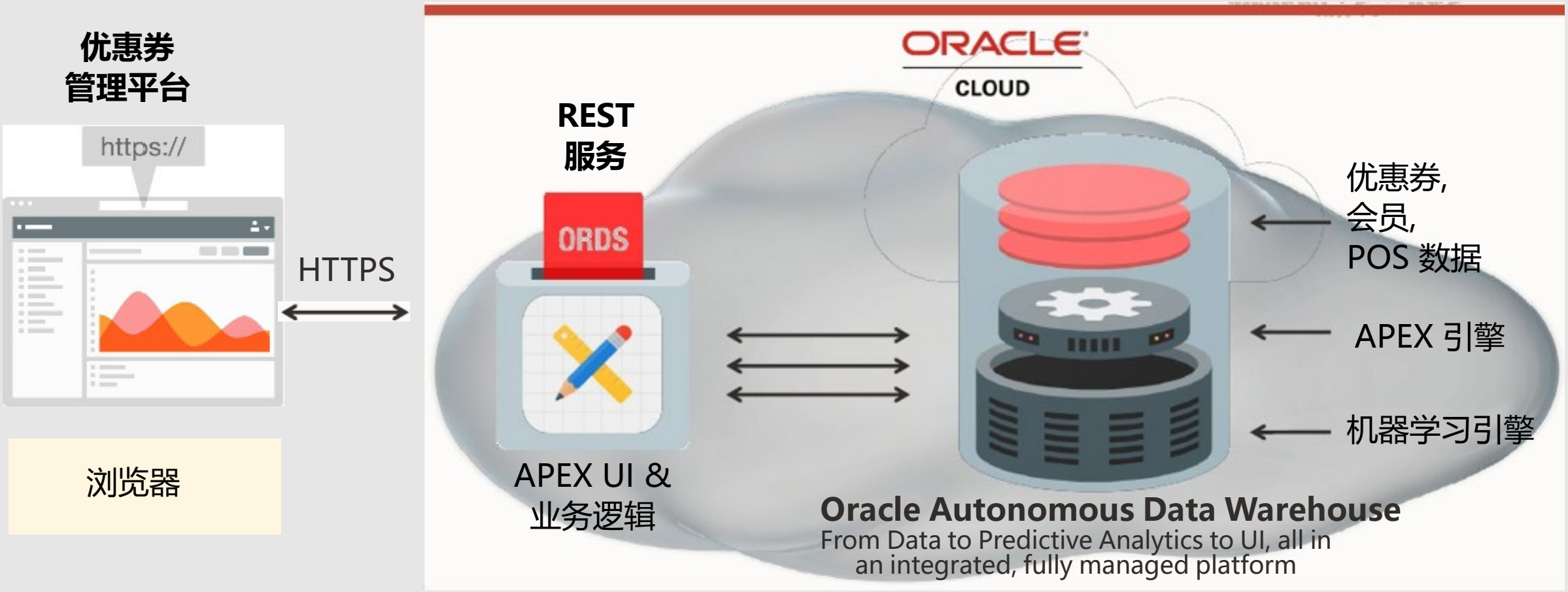
客户是全球大型跨国连锁餐厅，超过3.2万家分店，主营业务是售卖汉堡包，以及薯条、炸鸡、汽水、冰品、水果等快餐食品。

业务目标

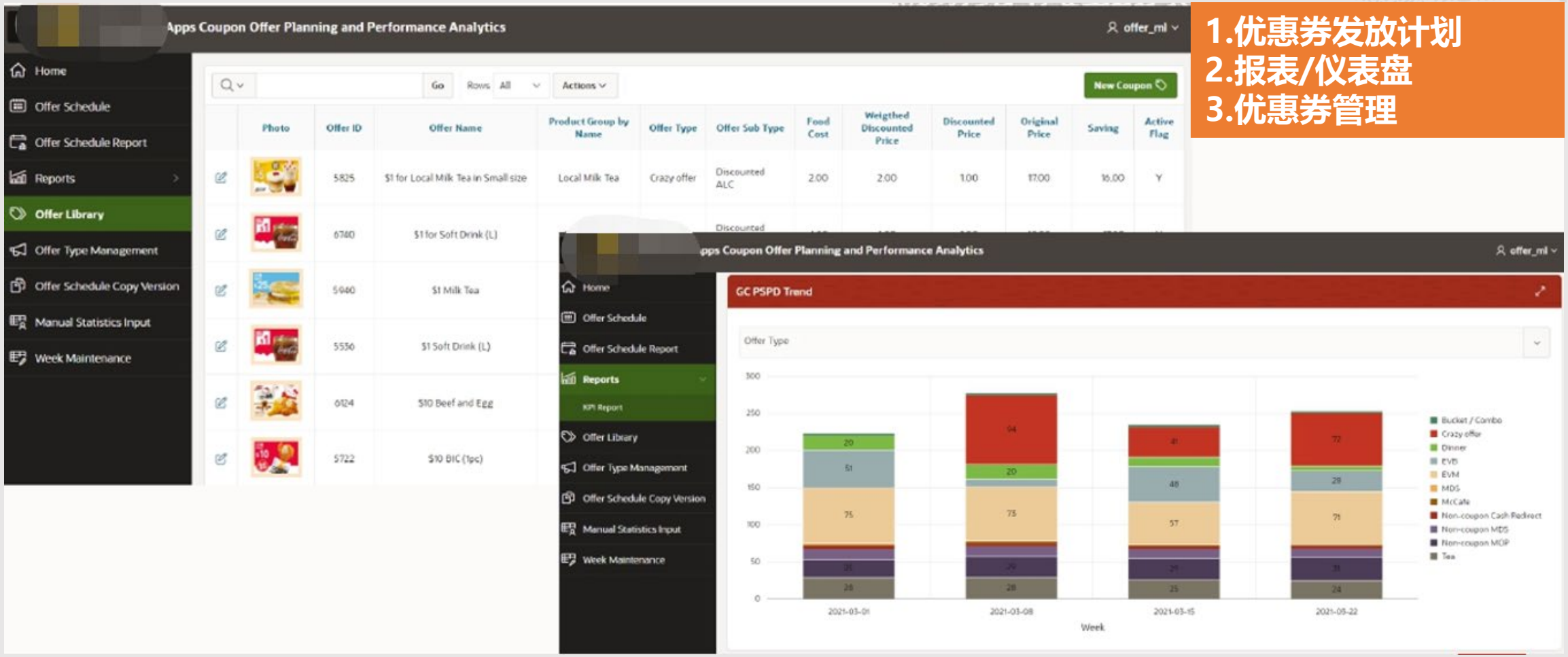
1. **机器学习模型——预测未来一周优惠券的使用量**
   - ✓ 基于优惠券的历史数据预测未来一周优惠券的使用量
2. **机器学习模型——客户分群**
   - ✓ 洞察客户，量身定制客户的促销和优惠
3. **基于APEX开发——优惠券管理**
   - ✓ 管理优惠券的发放计划
   - ✓ KPI仪表盘展示/报表

优惠券预测分析——架构

数据存储、分析、机器学习和应用的多合一平台



基于APEX开发——优惠券管理



1.优惠券发放计划
2.报表/仪表盘
3.优惠券管理

机器学习模型——预测未来一周优惠券的使用量

原始数据

栏位	数据说明
原始数据提供方式	ADW数据库的2个schema（dw,offer_ml）
原始数据表	dw.M_CUSTOMER(客户基本信息表，1.8千万个)
	dw.F_TICKET(客户订单交易明细，1.5亿笔)
	dw.F_ITEM(客户订单Item级明细，6.6亿笔)
	dw.M_ITEM(Item明细，3万个)
	M_CAMPAIGN(促销活动表,2300个)
	M_OFFER_LIBRARY(优惠券,177个)

数据清洗

训练数据集说明——交易基本信息

CUSTOMERID_ADJ	POSTRANSACTIONTIME_SCORE	OfferID	Transaction Details	Customer Profile before POSTRANSACTIONTIME_SCORE	Available Offer
12345678	27-MAR-2110.43.07.000000AM	6374			
12345678	06-MAR-2110.30.59.000000AM	0			
12345678	28-FEB-2109.58.16.000000AM	6374			
12345678	20-FEB-2108.30.11.000000AM	6374			

Field	Description	Remark
CUSTOMERID_ADJ	Customer ID after linked other Tenders	Not for modeling
POSTRANSACTIONTIME_SCORE	Transaction Datetime	Not for modeling
DAYPART_SCORE	Breakfast, Lunch, Tea, Dinner	Input variable
WEEKNAME_SCORE	Weekday/Weekend	Input variable
BUSINESSWEEK_SCORE	Week of the Transaction date, e.g. 22-MAR-21	Not for modeling
OFFER_ID_SCORE	Used Offer for the transaction, e.g. 6374	Output variable
OFFER_NAME_SCORE	Used Offer Name for the transaction, e.g. $3 off EVB	Not for modeling
PRODUCT_GROUP_SCORE	Used Offer product group for the transaction, EVB	Not for modeling
OFFER_SUBTYPE_SCORE	Used Offer sub type for the transaction, $ off EVB	Not for modeling
OFFER_TYPE_SCORE	Used Offer type for the transaction, EVB	Not for modeling
OFFER_FLAG_SCORE	1 = Using offer, 0 = Without using offer	Output variable
BRANDEXTENSION_SCORE	Brand Extension of the transaction, e.g. Main, Dessert, MDS, Mccafe	Not for modeling
SERVICETYPE_SCORE	Service Type of the transaction, e.g. Take Out, Eat In, Delivery	Not for modeling
PROMOTIONID_SCORE	Promotion Code of the transaction	Not for modeling

数据清洗

训练数据集说明——客户基本信息

CUSTOMERID_ADJ	POSTRANSACTIONTIME_SCORE	OfferID	Transaction Details	Customer Profile before POSTRANSACTIONTIME_SCORE	Available Offer
12345678	27-MAR-2110.43.07.000000AM	6374			
12345678	06-MAR-2110.30.59.000000AM	0			
12345678	28-FEB-2109.58.16.000000AM	6374			
12345678	20-FEB-2108.30.11.000000AM	6374			

Field	Grouping	Description
MEMBERSHIPLENGTH	Customer Basic Information	Days of the customer joined as a GMA user
BIRTHYEAR		Birth Year
GENDERID		Gender
WITHKIDS		With children flag
DOESACCEPTPROMOTION	App setting	Accept Promotion in the app
OPTIONPREFERENCE		Number of option preference opened in the app
APPNOTIFICATION		Number of app notification opened in the app
EMAILNOTIFICATION		Number of email notification opened in the app

数据清洗

训练数据集说明——RFM

Field Field	Grouping Grouping	Description Description
ROFFER	Recency of using Offer	Days from last transaction with using offer

Field	Grouping	Description
FREQUENCY	Dec-20 – Mar-21	Total transactions of the customer
FOFFER_7DAY	No. of transaction in last 7 days with using offer	Transactions in last 7 days with using Offer
FOFFER_7DAY_B		Breakfast Transactions in last 7 days with using Offer
FOFFER_7DAY_L		Lunch time Transactions in last 7 days with using Offer
FOFFER_7DAY_T		Teatime Transactions in last 7 days with using Offer
FOFFER_7DAY_D		Dinner Transactions in last 7 days with using Offer
FOFFER_7DAY_EXP	Average No. of transaction in last 7 days with using offer	Average of Transactions in last 7 days with using Offer
FOFFER_7DAY_B_EXP		
FOFFER_7DAY_L_EXP		
FOFFER_7DAY_T_EXP		
FOFFER_7DAY_D_EXP		
TOTAL_PROD_SALES	Other profile	Total product sales of the customer
PROMRATE		Rate of using offer
LATEST_OFFER		Last used Offer
TOP_STORE_LVL3		Most frequent purchase location (e.g. Causeway Bay, Fo Tan)
OFFERNUM_B		Average offer used in the breakfast
OFFERNUM_L		Average offer used in the lunch time
OFFERNUM_T		Average offer used in the teatime
OFFERNUM_D		Average offer used in the dinner
ROFFER_DESSERT_EXP	Average Recency of using Offer by Brand Extension	
ROFFER_MDS_EXP		
ROFFER_MCCAFE_EXP		
ROFFER_DELIVERY_EXP	Average Recency of using Offer by Service Type	
ROFFER_EATIN_EXP		
ROFFER_TAKEOUT_EXP		

数据清洗

训练数据集说明——Daypart & Week Name

CUSTOMERID_ADJ	POSTRANSACTIONTIME_SCORE	OfferID	Transaction Details	Customer Profile before POSTRANSACTIONTIME_SCORE	Available Offer
12345678	27-MAR-2110.43.07.000000AM	6374			
12345678	06-MAR-2110.30.59.000000AM	0			
12345678	28-FEB-2109.58.16.000000AM	6374			
12345678	20-FEB-2108.30.11.000000AM	6374			

Field	Grouping	Description
DOWDP_WEEKDAY	Proportion of transactions by Week Name with using Offer	Proportion of transactions in weekday with using Offer
DOWDP_WEEKEND		Proportion of transactions in weekend with using Offer
DOWDP_BREAKFAST	Proportion of transactions by daypart with using Offer	Proportion of transactions at breakfast with using Offer
DOWDP_LUNCH		Proportion of transactions at lunch time with using Offer
DOWDP_TEA		Proportion of transactions at teatime with using Offer
DOWDP_DINNER		Proportion of transactions at dinner with using Offer
DOWDP_WD_B	Proportion of transactions by daypart in weekday with using Offer	Proportion of transactions at weekday breakfast with using Offer
DOWDP_WD_L		Proportion of transactions at weekday lunch time with using Offer
DOWDP_WD_T		Proportion of transactions at weekday teatime with using Offer
DOWDP_WD_D		Proportion of transactions at weekday dinner with using Offer
DOWDP_WE_B	Proportion of transactions by daypart in weekend with using Offer	Proportion of transactions at weekend breakfast with using Offer
DOWDP_WE_L		Proportion of transactions at weekend lunch time with using Offer
DOWDP_WE_T		Proportion of transactions at weekend teatime with using Offer
DOWDP_WE_D		Proportion of transactions at weekend dinner with using Offer

数据清洗

训练数据集说明——Brand Extension

Field	Grouping	Description
BEX_B_MAIN	Proportion of transactions by Brand Extension at breakfast with using Offer	Proportion of transactions from Main at breakfast with using Offer
BEX_B_MCCAFE		Proportion of transactions from McCafe at breakfast with using Offer
BEX_B_DESSERT		Proportion of transactions from Dessert at breakfast with using Offer
BEX_B_MDS		Proportion of transactions from MDS at breakfast with using Offer
BEX_L_MAIN	Proportion of transactions by Brand Extension at lunch time with using Offer	Proportion of transactions from Main at lunch time with using Offer
BEX_L_MCCAFE		Proportion of transactions from McCafe at lunch time with using Offer
BEX_L_DESSERT		Proportion of transactions from Dessert at lunch time with using Offer
BEX_L_MDS		Proportion of transactions from MDS at lunch time with using Offer
BEX_T_MAIN	Proportion of transactions by Brand Extension at teatime with using Offer	Proportion of transactions from Main at teatime with using Offer
BEX_T_MCCAFE		Proportion of transactions from McCafe at teatime with using Offer
BEX_T_DESSERT		Proportion of transactions from Dessert at teatime with using Offer
BEX_T_MDS		Proportion of transactions from MDS at teatime with using Offer
BEX_D_MAIN	Proportion of transactions by Brand Extension at dinner with using Offer	Proportion of transactions from Main at dinner with using Offer
BEX_D_MCCAFE		Proportion of transactions from McCafe at dinner with using Offer
BEX_D_DESSERT		Proportion of transactions from Dessert at dinner with using Offer
BEX_D_MDS		Proportion of transactions from MDS at dinner with using Offer
BEX_WE_MAIN	Proportion of transactions by Brand Extension in weekend with using Offer	Proportion of transactions in weekend from Main with using Offer
BEX_WE_MCCAFE		Proportion of transactions in weekend from McCafe with using Offer
BEX_WE_DESSERT		Proportion of transactions in weekend from Dessert with using Offer
BEX_WE_MDS		Proportion of transactions in weekend from MDS with using Offer

数据清洗

训练数据集说明——Service Type

Field	Grouping	Description
SERTYPE_DELIVERY	Proportion of transactions by daypart with using Offer	Proportion of delivery transactions with using Offer
SERTYPE_EATIN		Proportion of dine in transactions with using Offer
SERTYPE_TAKEOUT		Proportion of take away transactions with using Offer
SERTYPE_WD_DELIVERY	Proportion of transactions by daypart in weekday with using Offer	Proportion of delivery transactions in weekday with using Offer
SERTYPE_WD_EATIN		Proportion of dine in transactions in weekday with using Offer
SERTYPE_WD_TAKEOUT		Proportion of take away transactions in weekday with using Offer
SERTYPE_WE_DELIVERY	Proportion of transactions by daypart in weekend with using Offer	Proportion of delivery transactions in weekend with using Offer
SERTYPE_WE_EATIN		Proportion of dine in transactions in weekend with using Offer
SERTYPE_WE_TAKEOUT		Proportion of take away transactions in weekend with using Offer
SERTYPE_B_DELIVERY	Proportion of transactions by daypart at breakfast with using Offer	Proportion of delivery transactions at breakfast with using Offer
SERTYPE_B_EATIN		Proportion of dine in transactions at breakfast with using Offer
SERTYPE_B_TAKEOUT		Proportion of take away transactions at breakfast with using Offer
SERTYPE_L_DELIVERY	Proportion of transactions by daypart at lunch with using Offer	Proportion of delivery transactions at lunch time with using Offer
SERTYPE_L_EATIN		Proportion of dine in transactions at lunch time with using Offer
SERTYPE_L_TAKEOUT		Proportion of take away transactions at lunch time with using Offer
SERTYPE_T_DELIVERY	Proportion of transactions by daypart at teatime with using Offer	Proportion of delivery transactions at teatime with using Offer
SERTYPE_T_EATIN		Proportion of dine in transactions at teatime with using Offer
SERTYPE_T_TAKEOUT		Proportion of take away transactions at teatime with using Offer
SERTYPE_D_DELIVERY	Proportion of transactions by daypart at dinner with using Offer	Proportion of delivery transactions at dinner with using Offer
SERTYPE_D_EATIN		Proportion of dine in transactions at dinner with using Offer
SERTYPE_D_TAKEOUT		Proportion of take away transactions at dinner with using Offer

数据清洗

训练数据集说明——Top Item Purchased

Field	Grouping	Description
ITEM_FOOD_OFFER_T_RANK1	Top/Second Tea Food Item Purchased	Top Tea Food Item Purchased with Offer
ITEM_FOOD_OFFER_T_RANK2		Second Tea Food Item Purchased with Offer
ITEM_FOOD_T_RANK1		Top Tea Food Item Purchased
ITEM_FOOD_T_RANK2		Second Tea Food Item Purchased
ITEM_DRINK_OFFER_T_RANK1	Top/Second Tea Drink Item Purchased	Top Tea Drink Item Purchased with Offer
ITEM_DRINK_OFFER_T_RANK2		Second Tea Drink Item Purchased with Offer
ITEM_DRINK_T_RANK1		Top Tea Drink Item Purchased
ITEM_DRINK_T_RANK2		Second Tea Drink Item Purchased
ITEM_FOOD_OFFER_D_RANK1	Top/Second Dinner Food Item Purchased	Top Dinner Food Item Purchased with Offer
ITEM_FOOD_OFFER_D_RANK2		Second Dinner Food Item Purchased with Offer
ITEM_FOOD_D_RANK1		Top Dinner Food Item Purchased
ITEM_FOOD_D_RANK2		Second Dinner Food Item Purchased
ITEM_DRINK_OFFER_D_RANK1	Top/Second Dinner Drink Item Purchased	Top Dinner Drink Item Purchased with Offer
ITEM_DRINK_OFFER_D_RANK2		Second Dinner Drink Item Purchased with Offer
ITEM_DRINK_D_RANK1		Top Dinner Drink Item Purchased
ITEM_DRINK_D_RANK2		Second Dinner Drink Item Purchased

数据清洗

训练数据集说明——Top Item Subcategory Purchased

Field	Grouping	Description
ITEMSUBCAT_FOOD_OFFER_T_RANK1	Top/Second Tea Food Item Subcategory Purchased	Top Tea Food Item Subcategory Purchased with Offer
ITEMSUBCAT_FOOD_OFFER_T_RANK2		Second Tea Food Item Subcategory Purchased with Offer
ITEMSUBCAT_FOOD_T_RANK1		Top Tea Food Item Subcategory Purchased
ITEMSUBCAT_FOOD_T_RANK2		Second Tea Food Item Subcategory Purchased
ITEMSUBCAT_DRINK_OFFER_T_RANK1	Top/Second Tea Drink Item Subcategory Purchased	Top Tea Drink Item Subcategory Purchased with Offer
ITEMSUBCAT_DRINK_OFFER_T_RANK2		Second Tea Drink Item Subcategory Purchased with Offer
ITEMSUBCAT_DRINK_T_RANK1		Top Tea Drink Item Subcategory Purchased
ITEMSUBCAT_DRINK_T_RANK2		Second Tea Drink Item Subcategory Purchased
ITEMSUBCAT_FOOD_OFFER_D_RANK1	Top/Second Dinner Food Item Subcategory Purchased	Top Dinner Food Item Subcategory Purchased with Offer
ITEMSUBCAT_FOOD_OFFER_D_RANK2		Second Dinner Food Item Subcategory Purchased with Offer
ITEMSUBCAT_FOOD_D_RANK1		Top Dinner Food Item Subcategory Purchased
ITEMSUBCAT_FOOD_D_RANK2		Second Dinner Food Item Subcategory Purchased
ITEMSUBCAT_DRINK_OFFER_D_RANK1	Top/Second Dinner Drink Item Subcategory Purchased	Top Dinner Drink Item Subcategory Purchased with Offer
ITEMSUBCAT_DRINK_OFFER_D_RANK2		Second Dinner Drink Item Subcategory Purchased with Offer
ITEMSUBCAT_DRINK_D_RANK1		Top Dinner Drink Item Subcategory Purchased
ITEMSUBCAT_DRINK_D_RANK2		Second Dinner Drink Item Subcategory Purchased

数据清洗

训练数据集说明——Top Offer Used

Field	Grouping	Description
OFFER_PG_B_RANK1	Top/Second Offer Product Group used by daypart	Top Offer Product Group used at breakfast
OFFER_PG_B_RANK2		Second Offer Product Group used at breakfast
OFFER_PG_L_RANK1		Top Offer Product Group used at lunch
OFFER_PG_L_RANK2		Second Offer Product Group used at lunch
OFFER_PG_T_RANK1		Top Offer Product Group used at tea
OFFER_PG_T_RANK2		Second Offer Product Group used at tea
OFFER_PG_D_RANK1		Top Offer Product Group used at dinner
OFFER_PG_D_RANK2		Second Offer Product Group used at dinner
OFFER_ID_B_RANK1	Top/Second Offer ID used by daypart	Top Offer ID used at breakfast
OFFER_ID_B_RANK2		Second Offer ID used at breakfast
OFFER_ID_L_RANK1		Top Offer ID used at lunch
OFFER_ID_L_RANK2		Second Offer ID used at lunch
OFFER_ID_T_RANK1		Top Offer ID used at tea
OFFER_ID_T_RANK2		Second Offer ID used at tea
OFFER_ID_D_RANK1		Top Offer ID used at dinner
OFFER_ID_D_RANK2		Second Offer ID used at dinner

数据清洗

训练数据集说明

CUSTOMERID_ADJ	POSTRANSACTIONTIME_SCORE	OfferID	Transaction Details	Customer Profile before POSTRANSACTIONTIME_SCORE	Available Offer
12345678	27-MAR-2110.43.07.000000AM	6374			
12345678	06-MAR-2110.30.59.000000AM	0			
12345678	28-FEB-2109.58.16.000000AM	6374			
12345678	20-FEB-2108.30.11.000000AM	6374			

Available Offer Product Group (0/1)
AVAILOPG_(Angus)
AVAILOPG_(Apple pie)
AVAILOPG_(BIC)
AVAILOPG_(Bacon Big Mac)
AVAILOPG_(Big Mac)
AVAILOPG_(Breakfast combo)
AVAILOPG_(Dinner)
AVAILOPG_(Dinner Combo)
AVAILOPG_(Double FOF)
AVAILOPG_(EVB)
AVAILOPG_(EVM)
AVAILOPG_(FOF)
AVAILOPG_(Family Combo)
AVAILOPG_(GCB)
AVAILOPG_(Ham & Cheese)
AVAILOPG_(LTO pie)
AVAILOPG_(Local Milk Tea)
AVAILOPG_(McCafe Combo)
AVAILOPG_(McCafe Drinks)
...

Available Offer Subtype (0/1)
AVAILOST_($ off Combo)
AVAILOST_($ off Drinks)
AVAILOST_($ off EVB)
AVAILOST_($ off EVM)
AVAILOST_(Add on)
AVAILOST_(BIC Bucket)
AVAILOST_(BIC EVM)
AVAILOST_(Combo for 2)
AVAILOST_(Discounted ALC)
AVAILOST_(Discounted Combo for 1)
AVAILOST_(Discounted Combo for 2)
AVAILOST_(Discounted Combo for 3)
AVAILOST_(Discounted EVB)
AVAILOST_(Discounted EVM)
AVAILOST_(Food + Drinks)
AVAILOST_(Food + Snack)
AVAILOST_(Free item)
AVAILOST_(Free product)
AVAILOST_(LTO EVB)
...

Available Offer Type (0/1)
AVAILOST_($ off Combo)
AVAILOST_($ off Drinks)
AVAILOST_($ off EVB)
AVAILOST_($ off EVM)
AVAILOST_(Add on)
AVAILOST_(BIC Bucket)
AVAILOST_(BIC EVM)
AVAILOST_(Combo for 2)
AVAILOST_(Discounted ALC)
AVAILOST_(Discounted Combo for 1)
AVAILOST_(Discounted Combo for 2)
AVAILOST_(Discounted Combo for 3)
AVAILOST_(Discounted EVB)
AVAILOST_(Discounted EVM)
AVAILOST_(Food + Drinks)
AVAILOST_(Food + Snack)
AVAILOST_(Free item)
AVAILOST_(Free product)
AVAILOST_(LTO EVB)
...

Available Offer ID (0/1)
AVAILO_(5525)
AVAILO_(5531)
AVAILO_(5535)
AVAILO_(5751)
AVAILO_(5753)
AVAILO_(5825)
AVAILO_(5876)
AVAILO_(5929)
...

预测优惠券使用量—回归

1. 业务目标

- ✓ 基于优惠券的历史数据预测未来一周优惠券的使用量

2. 回归模型构建方法

- ✓ 使用截止到最近的星期天之前的数据预测下一周（星期一-星期天）优惠券的每天使用量。
- ✓ 滚动预测每一天的优惠券使用量
- ✓ 应用自动机器学习功能进行训练，找到Top 50个特征
- ✓ 结合对数据进行统计分析，并反复进行训练模型，测试模型，确定最终喂入模型的特征

3. 数据集

- ✓ 汇总每个优惠券每天的使用量
- ✓ 数据范围是：2020-12-01到2021-05-31，一共3726条记录，25个字段
- ✓ 字段：OFFER_ID, OFFER_NAME, PRODUCT_GROUP, OFFER_SUBTYPE, OFFER_TYPE等25个字段

预测优惠券使用量—回归

特征工程

方法	字段名称	说明	备注
1	OFFER_ID	优惠券ID	
2	OFFER_NAME	优惠券名称	
3	PRODUCT_GROUP	产品组	
4	OFFER_SUBTYPE	产品小类	
5	OFFER_TYPE	产品大类	
6	SEGMENT_ID		
7	WEIGTHED_DISCOUNTED_PRICE	价格	
8	BUSINESSDATE	交易日期	
9	BUSINESSWEEK	交易星期	
10	DAY_OF_WEEK	星期几	
11	WEEKNAME	工作日/双休日	
12	PROMOTION_OFFER_FLAG	是否促销	
13	REDEMPTION_UNIT		
14	PRODUCTION_COST	产品成本	

预测优惠券使用量—回归

特征工程

方法	字段名称	说明	备注
15	MEDIA_COST	包装成本	
16	TOTAL_COST	总成本	
17	HK_PUBLIC_HOLIDAY	香港假期（圣诞节，元旦等）	
18	PREPOST_HOLIDAY_IMPACT	香港假期(Y|N|H)	
19	DOUBLE_UP_PERIOD		
20	OUTLIER_RATIO	每个优惠券的使用率/平均优惠券的使用率	
21	OFFER_APPEARED_DAYS	从20年12月1日到21年5月31日，优惠券可用的周数。	让模型捕捉一些优惠券的下跌趋势
22	OFFER_APPEARED_DAYS_NORM	对OFFER_APPEARED_DAYS 计算最小-最大归一化	每张优惠券将有不同的下降率。需要标准化进行比较
23	MAVG_W	每个优惠券在T-1周和T-2周的平均使用量	加强趋势预测
24	MAVG_D	每个优惠券的7天前使用量	加强趋势预测
25	SUBSTITUTE_OFFER_TYPE	同一周内相同Offer Type优惠券的数量	同一个优惠类型的优惠券有相互竞争关系

预测优惠券使用量—回归

特征工程之后，示例数据

OFFER_ID	WEEK_SKEY	OFFER_SKEY	OFFER_NAME	PRODUCT_GROU...	OFFER_SUBTYPE...	OFFER_TYPE	SEGMENT_ID	WEIGTHED_DISCOUNTED_PRICE
6024	20201207	104	[Tea] $17 for food & drink	Tea	Food + Drinks	Tea		17.3
6024	20201207	104	[Tea] $17 for food & drink	Tea	Food + Drinks	Tea		17.3

FOOD_COST	BUSINESSDATE	BUSINESSWEEK	DAY_OF_WEEK	WEEKNAME	PROMOTION_OFFER_FLA...	REDEMPTION_UNIT	OUTLIER_RATIO	PRODUCTION_COST
5.3	2020-12-07 00:00:00	2020-12-07 00:00:00	MON	WEEKDAY	N	3157	0.963028	20
5.3	2020-12-08 00:00:00	2020-12-07 00:00:00	TUE	WEEKDAY	N	3131	0.955097	20

MEDIA_COST	TOTAL_COST	HK_PUBLIC_HOLIDA...	PREPOST_HOLIDAY_IMPACT	DOUBLE_UP_PERIOD ...	OFFER_APPEARED_DAYS	OFFER_APPEARED_DAYS_NORM
25	45	Normal Day	N	N	7	0.2068
25	45	Normal Day	N	N	7	0.2068

SUBSTITUTE_OFFER_TYPE...	MAVG_W	MAVG_D	CREATION_DATE	UPDATE_DATE
1	11730		2021-09-24 11:13:15	2021-09-24 11:13:15
1	11730	2797	2021-09-24 11:13:15	2021-09-24 11:13:15
1	11730	2803	2021-09-24 11:13:15	2021-09-24 11:13:15
1	11730	3075	2021-09-24 11:13:15	2021-09-24 11:13:15
1	11730	3055	2021-09-24 11:13:15	2021-09-24 11:13:15
1	11730		2021-09-24 11:13:15	2021-09-24 11:13:15
1	11730		2021-09-24 11:13:15	2021-09-24 11:13:15
1	13565	3157	2021-09-24 11:13:13	2021-09-24 11:13:13
1	13565	3131	2021-09-24 11:13:13	2021-09-24 11:13:13

预测优惠券使用量—回归

回归模型-SVM（高斯）

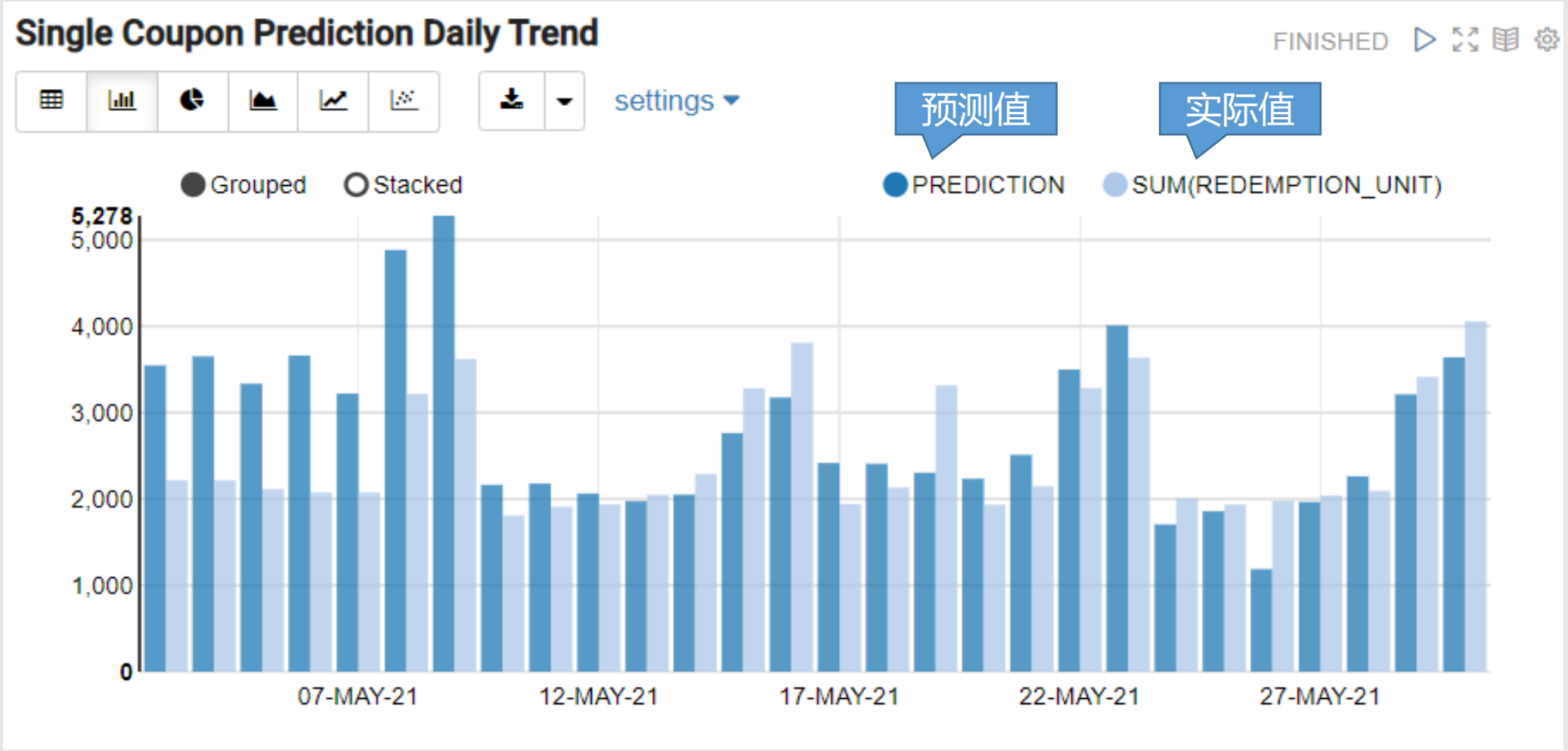
根据3类的Offer Type分别构建回归模型。

预测结果：

- ✓ 6024 ([Tea] $17 for food & drink) (4.7% error%)
- ✓ 6057 ($38 for Honey BBQ BIC EVM) (6% error%)
- ✓ 6374 ($3 off EVB) (7.5% error%)

预测优惠券使用量—回归

查看某一个优惠券预测结果



总结

高效的数据处理

- 借助SQL对数据进行清洗理
- ADW弹性的计算能力

All-In-SQL模型训练

- 模型训练、应用、测试都在数据库中完成
- 不涉及数据移动

机器学习和高级分析能力

- 内置30多种算法适用各种场景
- 借助Auto ML简化机器学习建模

Q&A

谢谢聆听


ORACLE