BIBDR | 博雅大数据学院 BOYA BIG DATA COLLEGE | 数据酷客 COOKDATA

北京市大数据人才培训示范基地

数据科学“云实训”项目训练营

第四课：事实类与规则类标签构建

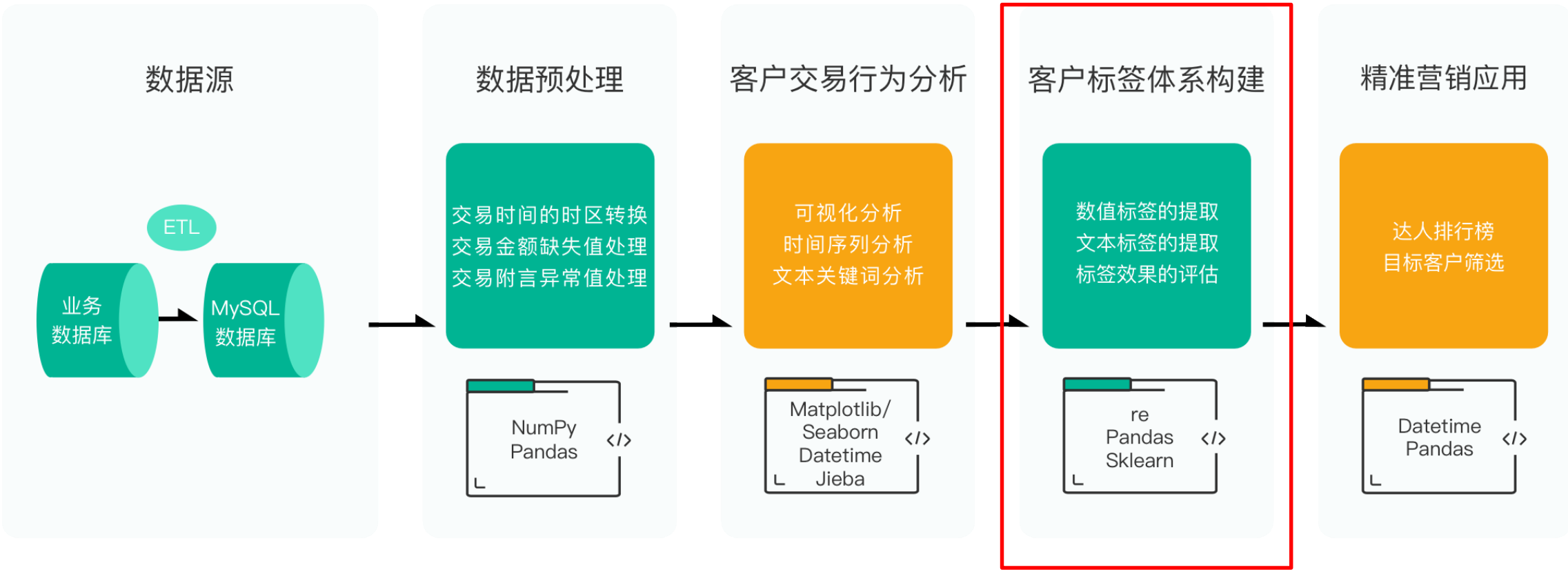
讲师：刘冬

数据酷客

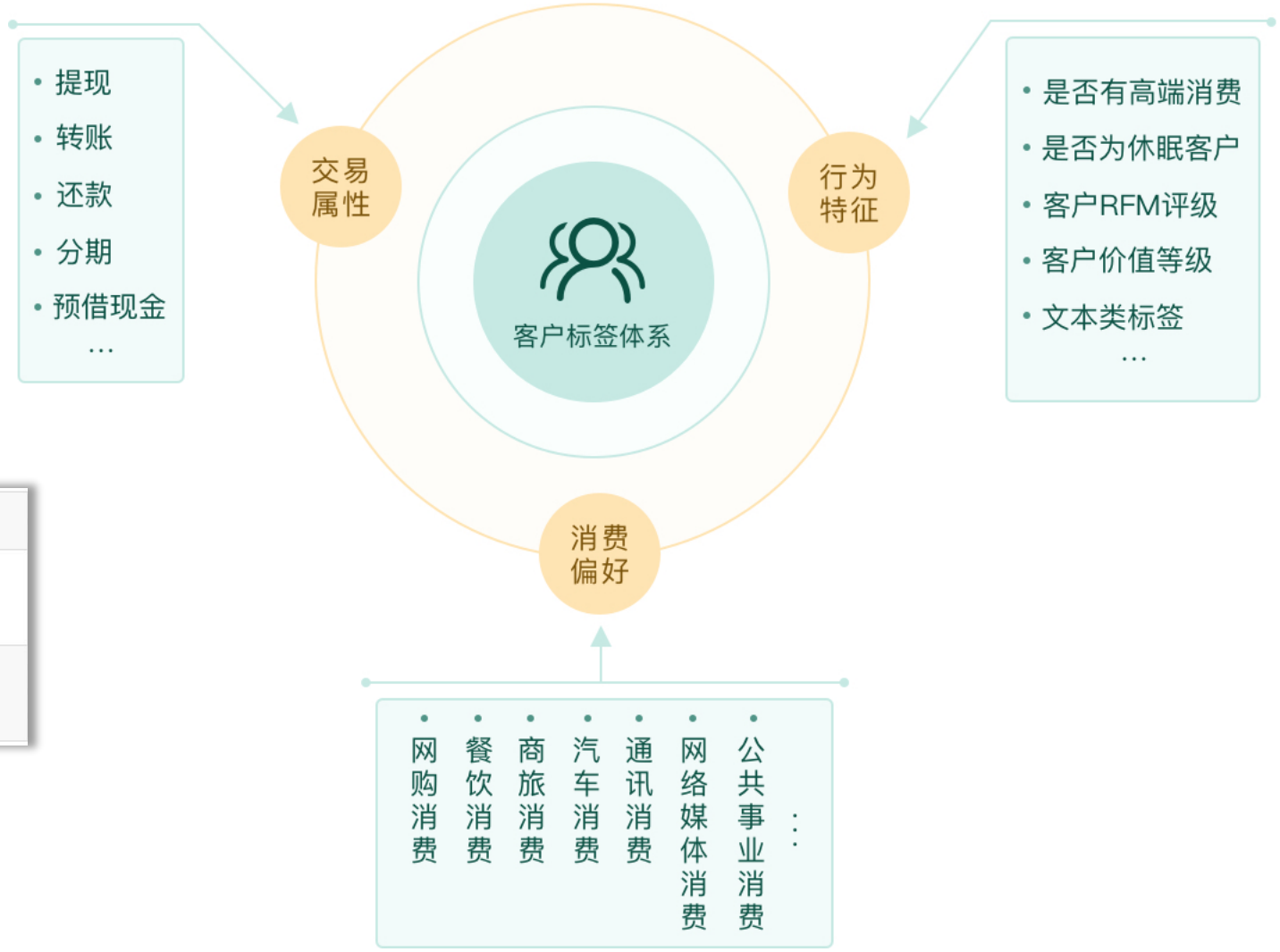
数据科学人工智能

项目流程

第四课开始


数据源
ETL
业务数据库
MySQL数据库
数据预处理
交易时间的时区转换
交易金额缺失值处理
交易附言异常值处理
NumPy
Pandas
客户交易行为分析
可视化分析
时间序列分析
文本关键词分析
Matplotlib/
Seaborn
Datetime
Jieba
客户标签体系构建
数值标签的提取
文本标签的提取
标签效果的评估
re
Pandas
Sklearn
精准营销应用
达人排行榜
目标客户筛选
Datetime
Pandas

客户标签体系



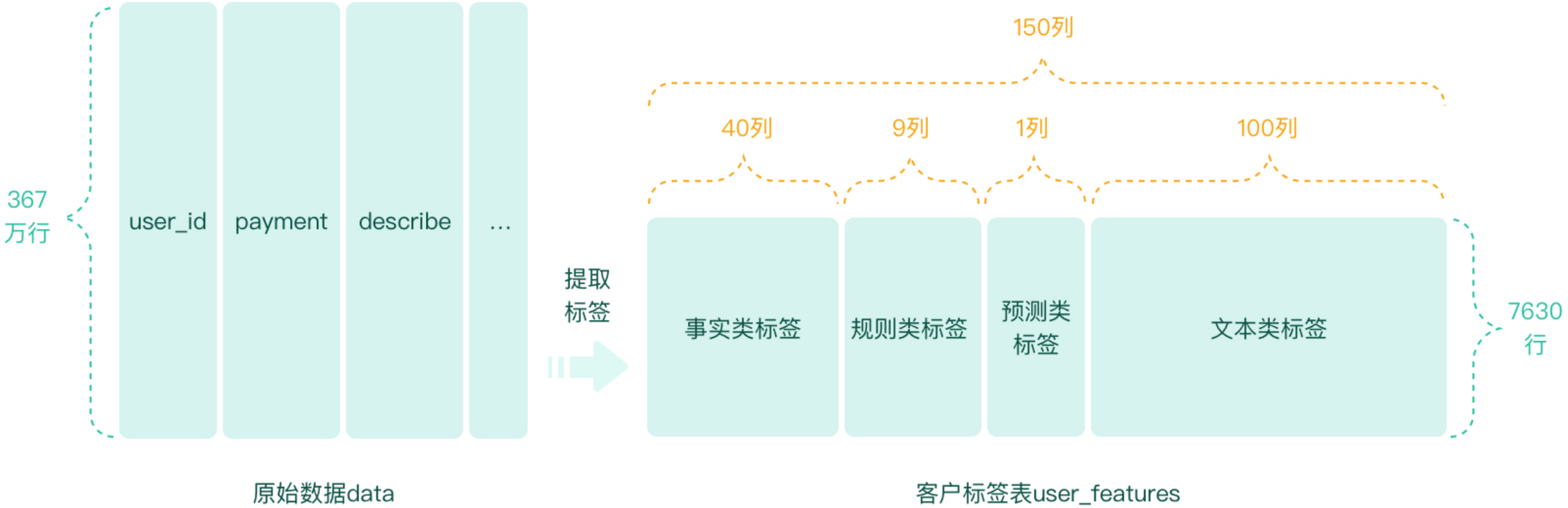
user_id	payment	describe	unix_time	pay_time
22171955	65.00	湖州天虹百货有限公司	1509379200	2017-10-31 00:00:00
22171955	224.50	湖州市星火服装有限公司	1509379200	2017-10-31 00:00:00

标签构建流程

第四课

- **事实类标签**：可以直接从客户交易记录中进行统计和计算的标签例如网购消费、餐饮消费、商旅消费等
- **规则类标签**：规则类标签是在事实类标签的基础上，结合人工经验，对客户的某项指标进行的计算或归类例如RFM标签、是否休眠客户、是否有高端消费等
- **预测类标签**：原始数据中不能直接提取，需要借助模型进行预测的标签例如客户价值等级
- **文本类标签**：从客户交易记录的文本中提取的关键词也可用于描述客户偏好，将此部分关键词作为文本类标签例如彩票、儿童、孕妇、基金等

标签构建流程



user_id	交易次数	交易总额	有无高端消费
109464	629	180107	0
115043	783	270203	0
125322	538	2009921	1
131673	77	26839	0
136544	278	90871	1

.....

事实类标签构建

40个事实类标签提取方法：

- 关键词匹配
- 分组聚合

英文	中文	英文	中文
max_consume_amt	单次最大消费金额	return_cnt	退货订单数
consume_order_ratio	消费订单比例	public_pay_amt	公共事业缴费金额
mon_consume_frq	月均消费频度	internet_media_cnt	网络媒体类消费次数
consumption_channel	最常用支付工具	internet_media_amt	网络媒体类消费总金额
online_cnt	网购订单次数	phone_fee_cnt	话费通讯类消费次数
online_amt	网购订单总金额	phone_fee_amt	话费通讯类消费总金额
online_avg_amt	网购订单平均金额	is_installment	有无分期
mon_online_frq	月均网购频度	cash_advance_cnt	预借现金次数
online_buy_first_date	网购首单时间	cash_advance_amt	预借现金总金额
online_buy_last_date	网购尾单时间	total_transactions_amt	交易总金额
dining_cnt	餐饮订单次数	total_transactions_cnt	交易次数
dining_amt	餐饮订单总金额	withdraw_cnt	提现次数
dining_avg_amt	餐饮订单平均金额	withdraw_amt	提现总金额
business_travel_cnt	商旅次数	total_deposit	ATM存款总金额
business_travel_amt	商旅消费金额	total_withdraw	ATM取款总金额
business_travel_avg_amt	商旅消费平均金额	transfer_cnt	转账次数
mon_business_travel_frq	月均旅行频次	transfer_amt	转账总金额
car_cnt	汽车消费次数	transfer_mean	转账平均金额
car_amt	汽车消费总金额	credit_card_repay_cnt	信用卡还款次数
payroll	有无代发	credit_card_repay_amt	信用卡还款总金额

关键词匹配

对某一列Series对象进行关键词匹配，使用contains()函数带入关键词进行匹配

Series.str.contains(pattern, case=True)

- pattern：待匹配的关键词
- case：区分大小写，默认为True

```
1  s = pd.Series(['苹果','葡萄','香蕉','西瓜','橙子'])
2  s
```

```
0    苹果
1    葡萄
2    香蕉
3    西瓜
4    橙子
dtype: object
```

```
1  s.str.contains('西瓜')
```

```
0    False
1    False
2    False
3     True
4    False
dtype: bool
```

```
1  s[s.str.contains('西瓜')]
```

```
3    西瓜
dtype: object
```

分组聚合

DataFrame中的groupby()函数将数据依照某一个属性进行分组

DataFrame.groupby(by=None)

by用来表示确定groupby()函数的分组依据

groupby()之后返回GroupBy对象，可接着使用聚合函数进行运算

- size()，查看各组的个数
- sum()，求各组之和
- mean()，求各组的平均值

	name	value
0	A	1
1	A	1
2	B	7
3	B	1
4	C	2
5	C	3
6	C	1

```
1  data.groupby('name').size()
```

```
name
A    2
B    2
C    3
dtype: int64
```

```
1  data.groupby('name')['value'].sum()
```

```
name
A    2
B    8
C    6
Name: value, dtype: int64
```

```
1  data.groupby('name')['value'].mean()
```

```
name
A    1
B    4
C    2
Name: value, dtype: int64
```

规则类标签构建

规则类标签提取方法：

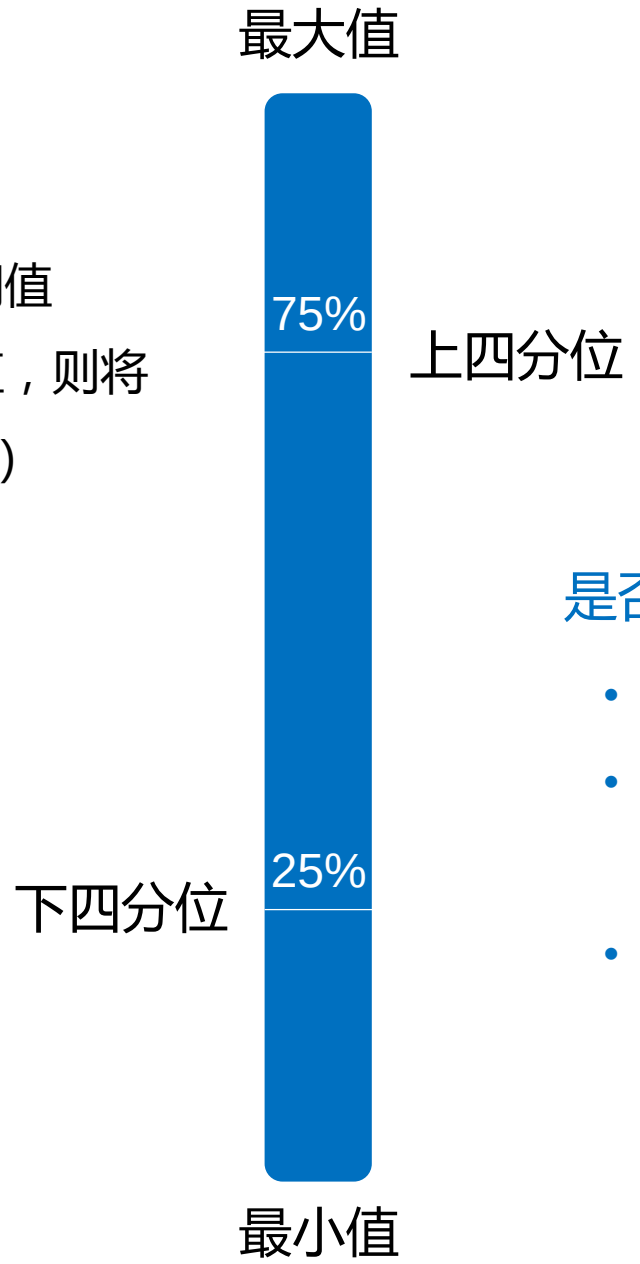
- 从事实类标签进行延伸
- 根据RFM模型进行计算

英文	中文
high_consumption	有无高端消费
sleep_customers	是否休眠客户
recency	近度
frequency	频度
monetary	值度
R_score	近度得分
F_score	频度得分
M_score	值度得分
Total_Score	RFM总得分

从事实类标签进行延伸

有无高端消费(high_consumption)

- 取最大消费金额的上四分位数作为阈值
- 如果客户的最大消费金额大于该阈值，则将该客户定义为有高端消费（取值为1）
- 反之则无高端消费（取值为0）

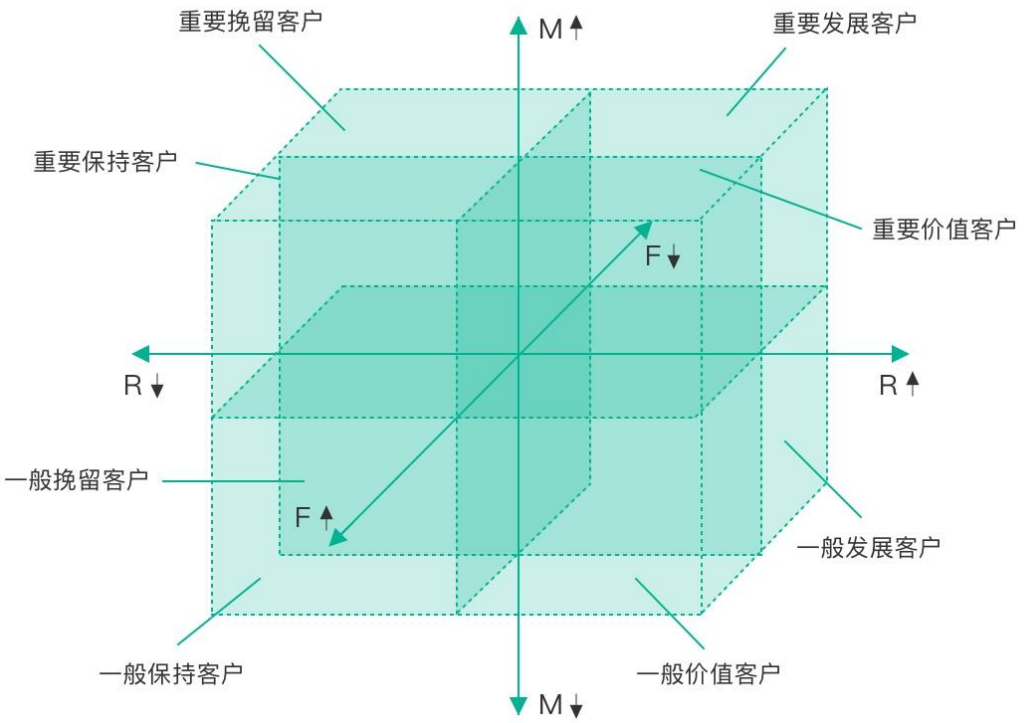


是否休眠客户(sleep_customers)

- 设定交易次数的下四分位数为阈值
- 交易次数小于阈值的客户则视为休眠客户（取值为1）
- 交易次数大于等于阈值的客户则视为活跃客户（取值为0)

RFM模型

Frequency（频度）：客户在一段时间内消费的次数，通常来说最常消费的客户，忠诚度相对高于其它客户



Recency（近度）：最近一次消费距离观察点的天数，上一次消费时间越近的客户应该是比较好的客户，对提供即时的商品或是服务最有可能响应

Monetary（值度）：客户在一段时间内消费的总金额，消费金额越高的客户越重要，消费金额的意义不言而喻

RFM总得分

- 每个维度排序后等频划分为四组，每个组计算得分

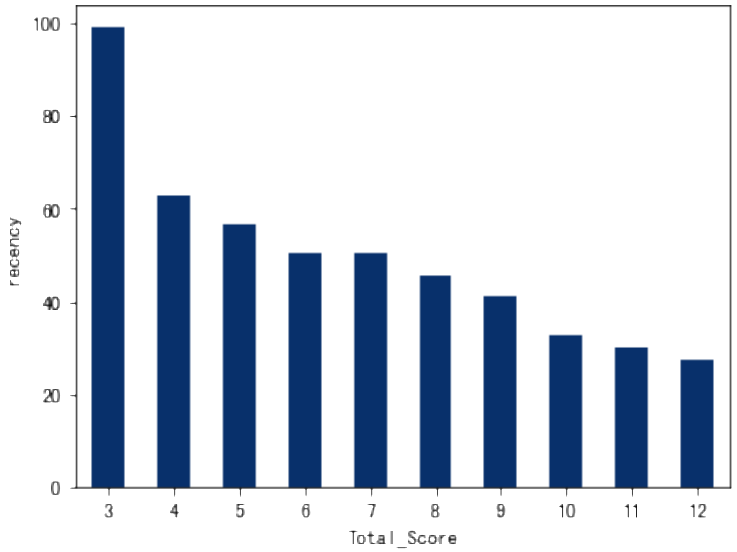
最近一次消费时间距离	得分标记
75%-100%	1
50%-75%	2
25%-50%	3
0%-25%	4

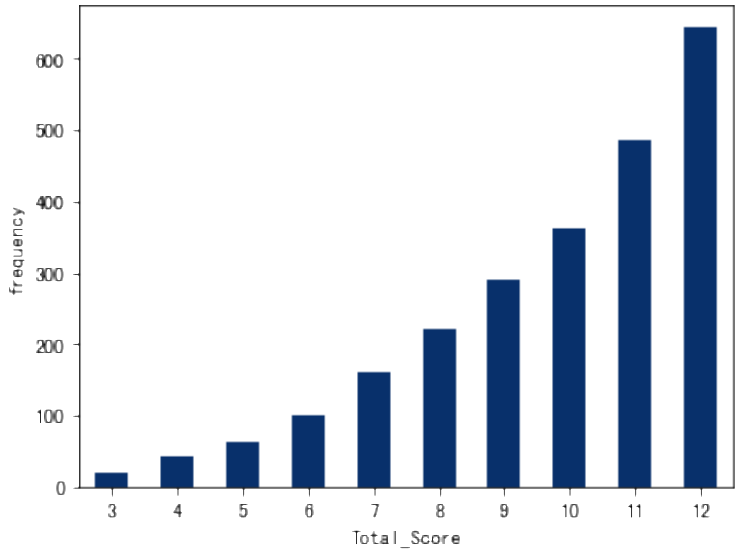
- 最近一次消费(R)按时间得分：取值越小得分越高
- 消费频次(F)及消费金额(M)取值越大得分越高

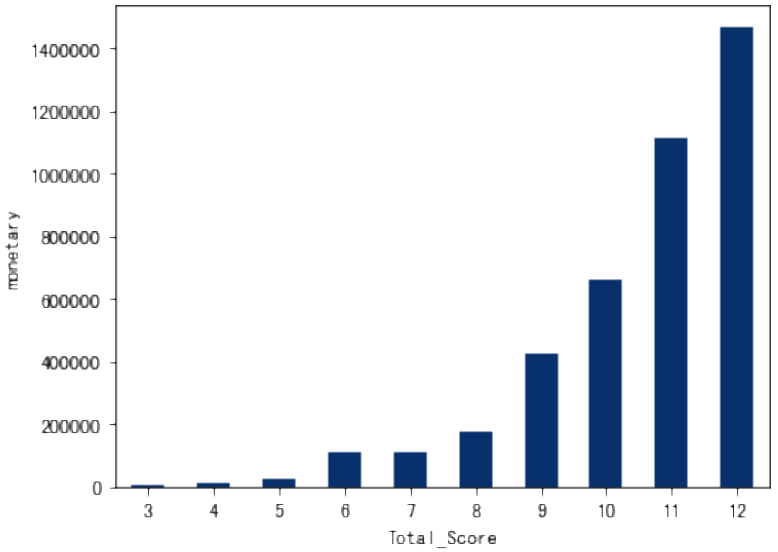
消费频次/消费金额	得分标记
75%-100%	4
50%-75%	3
25%-50%	2
0%-25%	1

- 将每个客户对应的三个评分标记相加，作为客户RFM的总得分(Total_Score)

RFM可视化分析







- 随着Total_Score的增大，recency的平均值逐渐减小，这也印证了较优质的客户群体，最近一次消费普遍较近
- 随着Total_Score的增大，frequency、monetary的平均值逐渐增大，这也印证了越优质的客户群体，消费频率和消费金额普遍越高

数据酷客官网

数据科学人工智能

加入数据酷客交流群