中国大数据算法大赛-用户购买时间预测

队伍名称：Trident

演讲者：Dylan, M

2018.07.19

JDATA

目录

JDATA

团队介绍

Dylan

M

Trident

Lindada

初夏过道

JDATA

目录

JDATA

赛题分析

题目概述

1. 人群：近三个月内有目标品类 (30, 101) 购买行为的用户作为用户集合

2. 数据维度：从用户集合过去一年的订单、行为及评论表中提取特征

3. 目标：预测用户集合中下个月最有可能购买目标品类的用户（类似用户复购概率排序），与其首次目标品类购买日期

核心问题

1. 训练集与线下验证集时间段与人群的选取方法?

2. 特征的主要提取方式：时间滑窗？时间窗口的数量及大小？促销日处理?

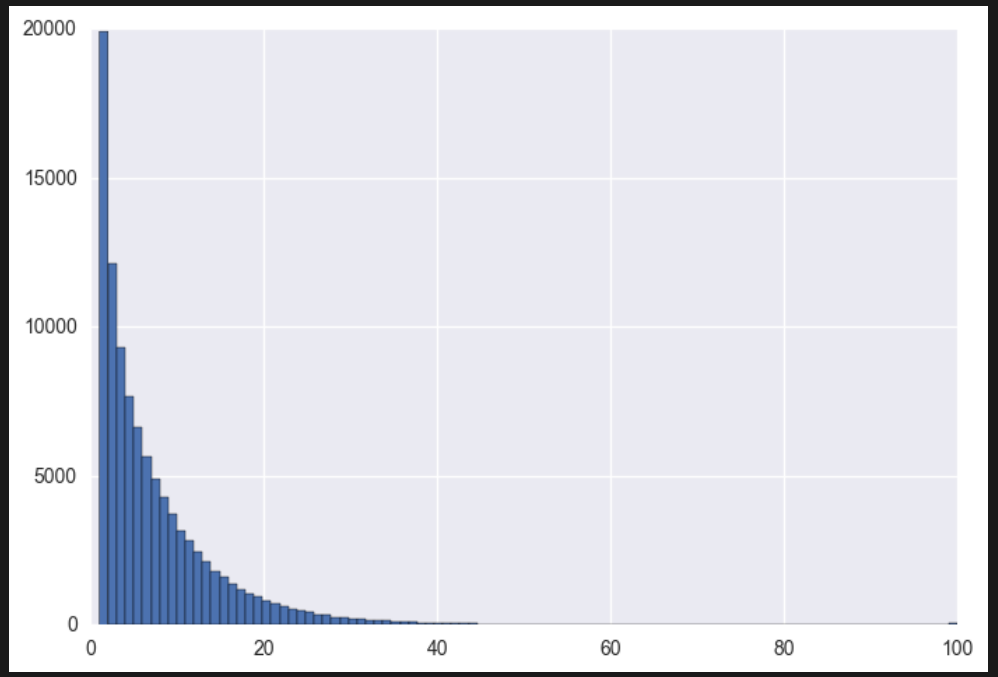
3. 30与101是否满足用户相似的需求？目标品类与其他品类之间是否可替代?

4. 模型评价函数的选取？S1、S2的target选取?

JDATA

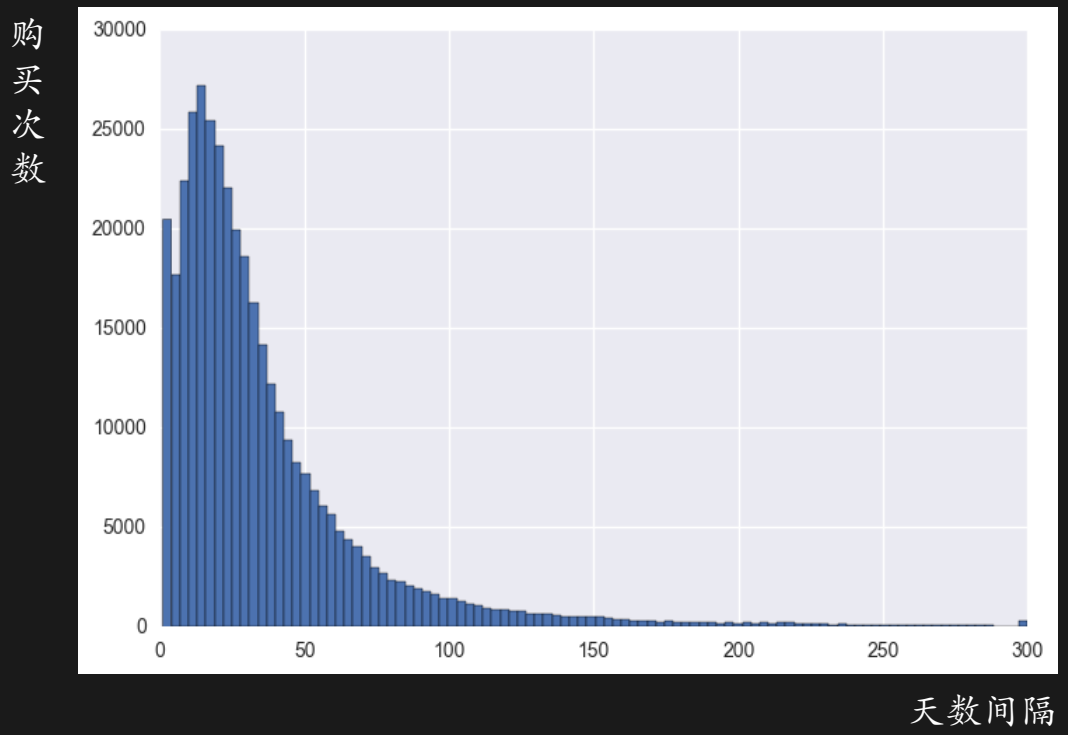
数据分析与探索

订单数的用户统计



Hot start

用户目标品类购买间隔



间隔较短，类似日用品

JDATA

数据分析与探索

关键发现

User ID: 27603

一些用户会在同一天出现多笔订单，
猜测可能是为了凑满减等促销方式；

构造某些特征时，需要进行预处理，
合并同一订单日来计算

user_id	sku_id	o_id	o_date	o_area	o_sku_num
27603	92841	430424	2017-06-17	20	3
27603	83023	525589	2017-06-17	20	3
27603	83023	393820	2017-06-17	20	2
27603	55837	328048	2017-06-17	2	1
27603	49127	328048	2017-06-17	2	3
27603	20458	57526	2017-06-18	10	1
27603	22422	620433	2017-06-18	10	1
27603	22422	602294	2017-06-18	10	1
27603	20458	2189	2017-06-18	10	1
27603	65733	566363	2017-06-18	20	1
27603	22422	262323	2017-06-18	10	1
27603	22422	330298	2017-06-18	10	1
27603	22422	483651	2017-06-18	10	1
27603	76439	580304	2017-06-18	10	5
27603	93317	421095	2017-06-18	20	1
27603	22422	495653	2017-06-18	10	1
27603	22422	74770	2017-06-18	10	1
27603	93317	437280	2017-06-18	20	1
27603	22422	139979	2017-06-18	10	1
27603	93317	416476	2017-06-18	20	1
27603	22422	470328	2017-06-18	10	1
27603	6847	191081	2017-06-18	20	7
27603	93317	553587	2017-06-18	20	1
27603	76439	233006	2017-06-18	10	5
27603	20458	607286	2017-06-18	10	1

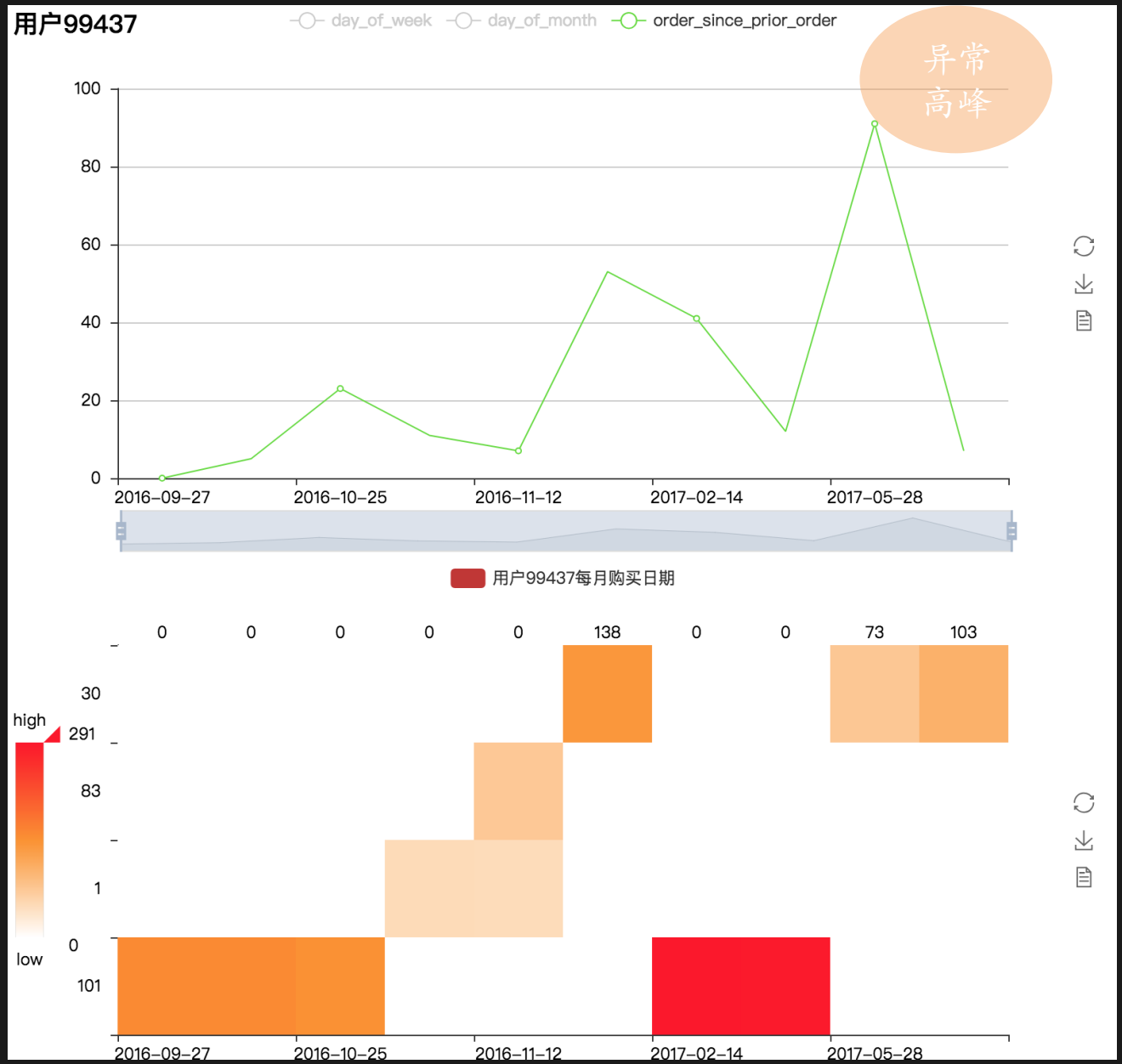
JDATA

数据分析与探索

关键发现

User ID: 99437

1. 购买para1的量可能与购买间隔有关
2. 猜测para1是商品容量规格类似的参数

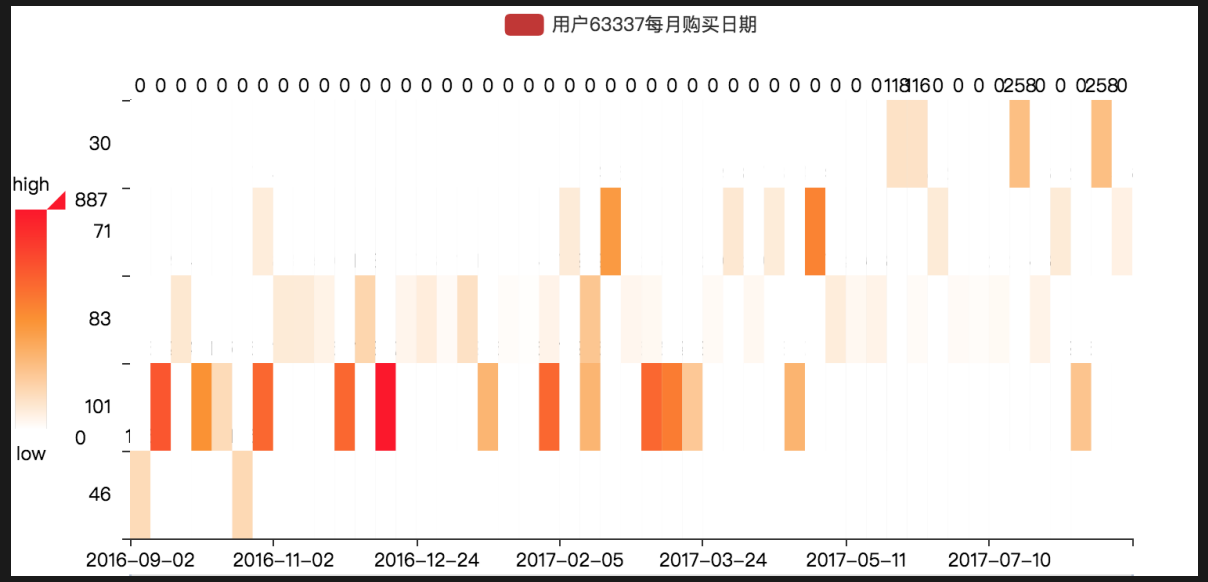


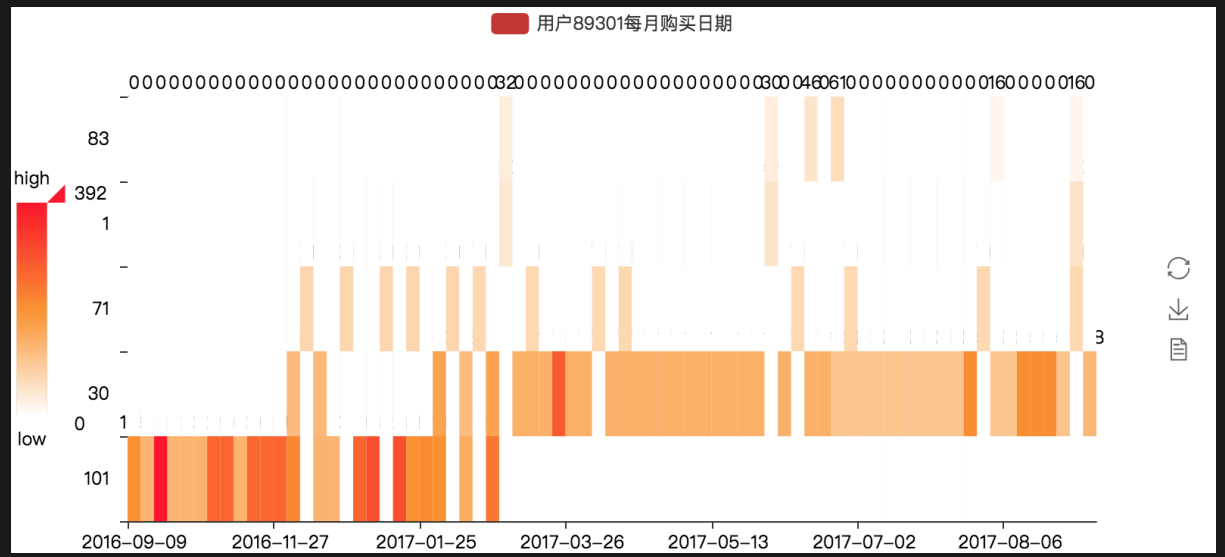
数据分析与探索

关键发现

User ID: 63337、89301

1. 101与30大致满足用户同一需求
2. 非目标品类的购买不影响用户目标品类间的购买间隔规律





JDATA

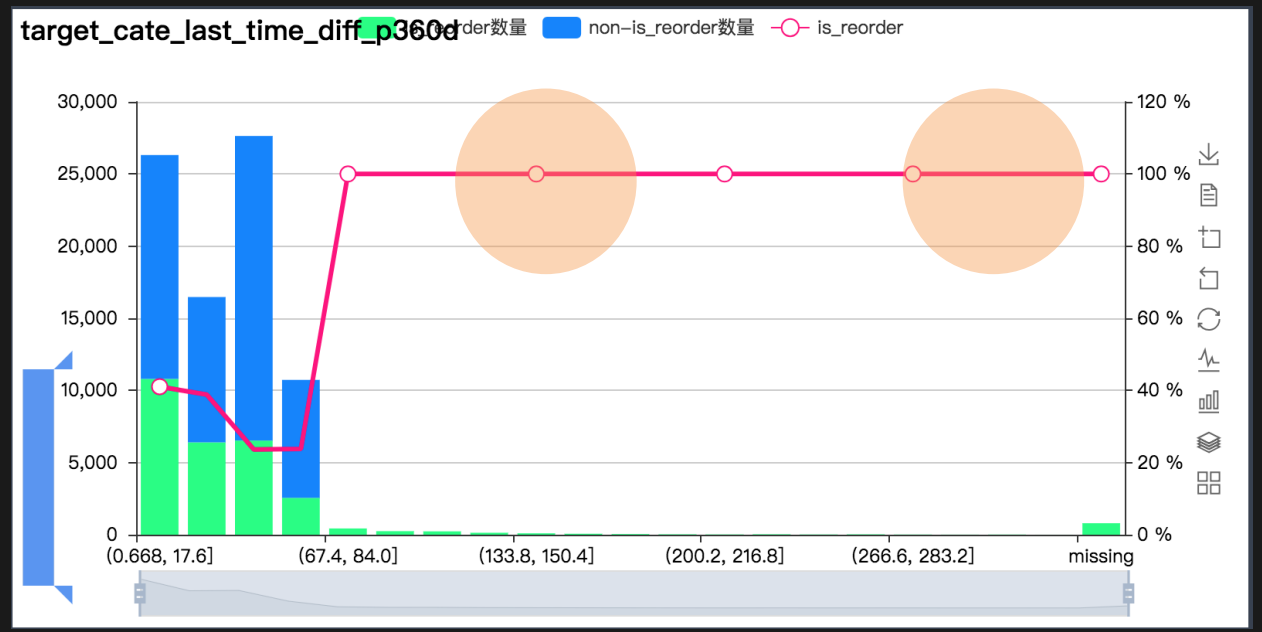
训练集与验证集划分 -- S1用户复购

选取6、7月份有购买目标品类的用户作为样本集，target为此批用户在8月是否有购买目标品类

线下验证方法为5Fold的交叉验证，同时对测试集做出预测取平均 → **线下、线上同方向，且涨幅基本相同**

若线下样本包含6、7月未购买过目标品类用户，会出现特征线下、线上分布不一致的情况

特征举例：上一次购买目标品类距今天（8月1日）的天数



由于用户在6、7、8一定会购买目标品类，线下当此距今天数大于两个月或缺失时，该部分用户必定会在8月购买，而线上该特征分布会截然不同

JDATA

训练集与验证集划分 -- S2 购买日期

S2 target y的选取方式考虑

方式	优点	缺点	结论
浏览至下单的间隔	可将用户最后一次浏览前的所有行为都包含在内	绝大部分用户浏览当天即下单	
目标品类订单间隔	拟合用户订单间隔规律	易过拟合“最后笔订单距今时间” 用户最后笔订单后的行为、无法使用	
首笔至考察时间间隔	与线上情况相同 可将行为都纳入考量	异受节假日、促销活动影响	

训练集与验证集划分 -- S2 购买日期

S2 样本窗口的选取

1. 选取前三个月有过购买行为的用户在接下来一个月内的购买情况
2. 适当滑窗来增加样本量；14天为大多数人的复购周期，半个月滑窗能使大部分的人有新的购买，使得样本不会过多重复。
3. 尽量避开节假日的影响，A榜中有春节和4.11的影响，B榜有6.18的影响，A、B榜滑窗3个刚好合适，同时在A榜中剔除了4.11节假日购买的样本。

起始时间	截止时间	考察范围	特征计算截止时间
A榜			
2017.1.1	2017.4.1	4.1-4.30	2017.4.1
2016.12.15	2017.3.15	3.15-4.15	2017.3.15
2016.12.1	2017.3.1	3.1-3.31	2017.3.1
B榜			
2017.5.1	2017.8.1	8.1-8.31	2017.8.1
2017.4.15	2017.7.15	7.15-8.15	2017.7.15
2017.4.1	2017.7.1	7.1-7.31	2017.7.1

特征工程 – 特征组概览

1. 用户行为时间跨度很长，需要以滑动窗口的方式来构造各类特征，从而保证特征值间的公平性
2. 推算特征大致是从用户长期行为中找到规律，并扩展到最后一次的行为，推算出下一次的购买时间

原始特征	基础特征	复杂特征
1. 用户等级	1. 订单时间特征	1. 订单时间推算特征
2. 用户年龄	2. 订单计数特征	2. Para1需求强度推算特征
3. 用户性别	3. 订单金额特征	3. 最后笔订单后的行为特征
4. 用户区域	4. Sku参数特征	4. 浏览至下单间隔特征
	5. 行为时间特征	5. 浏览、下单比例特征
	6. 行为计数特征	6. S2输出结果特征

特征工程 – 特征组群（时间维度类）

特征类别	关键特征	行为描述
1 时间跨度	• 最后笔订单、目标品类订单距今时间 • 最后笔目标品类订单与最后笔订单时间差 • 第一笔订单、目标品类订单距今时间 • 第一笔与最后笔目标品类时间差 • 最后一次浏览、关注目标品类距今时间	• 统计用户订单账龄 • 用户是否可能已离开京东该平台
2 时间间隔	• 目标品类订单间的时间间隔均值 • 目标品类订单间的时间间隔最大值 • 目标品类订单间的时间间隔最小值 • 目标品类订单间的时间间隔标准差 • 目标品类行为间的时间间隔均值 • ……	• 用户目标品类购买间隔的大致规律

特征工程 – 特征组群（订单计数类）

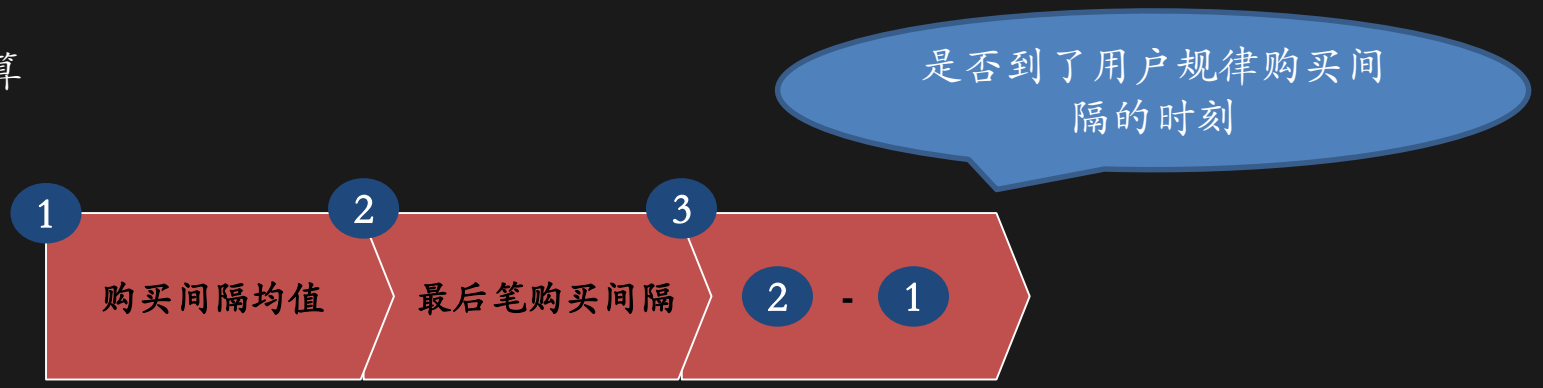
特征类别	关键特征	行为描述
1 订单数	• 订单、目标品类订单计数 • 目标品类订单数占比	• 订单维度刻画用户对于目标品类的需求程度
2 商品数	• 去重商品、目标品类商品计数 • 去重目标品类计数占比 • 总商品、目标品类商品计数 （乘数量） • 总目标品类商品计数占比 • 平均每笔目标品类商品数 • 每件目标品类商品平均购买次数	• 商品维度刻画用户对于目标品类的需求程度
3 天数	• 总品类、目标品类购买天数 • 目标品类购买天数占比 • 平均每天购买目标品类商品数统计	• 购买天数维度刻画用户对于目标品类的需求程度

特征工程 – 特征组群（复杂特征类）

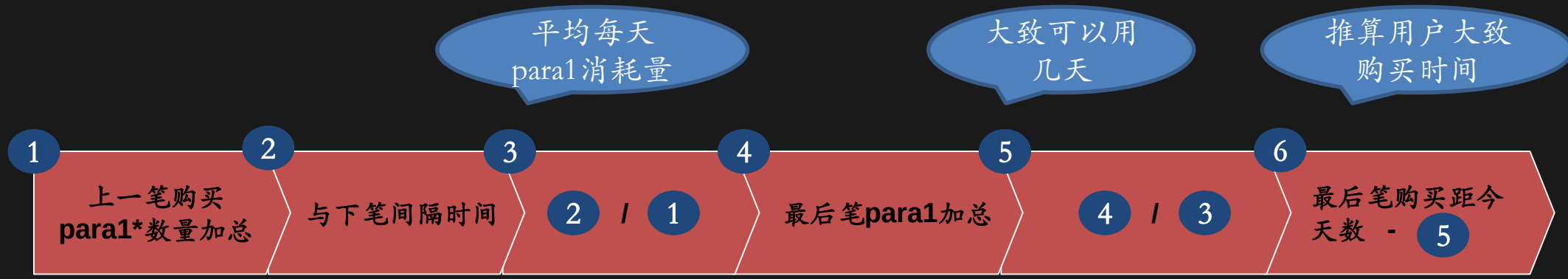
特征类别	关键特征	行为描述
1 参数	• para1的平均、求和等统计 • para2的平均、求和等统计 • para3的平均、求和等统计	• 描述用户对于目标品类不同参数的偏好
2 订单、行为结合	• 最后笔订单后的行为计数 • 最后笔订单后的行为距今时间统计 • 目标品类浏览至下单的平均间隔 • 目标品类浏览/下单的占比	• 描述用户最后笔订单后的附加行为 • 刻画用户下单的决策过程
3 S2输出	• S2预测输出（距离8月1日、9月1日天数）作为S1输入	

特征工程 – 特征组群（推算特征类）

目标品类时间类推算

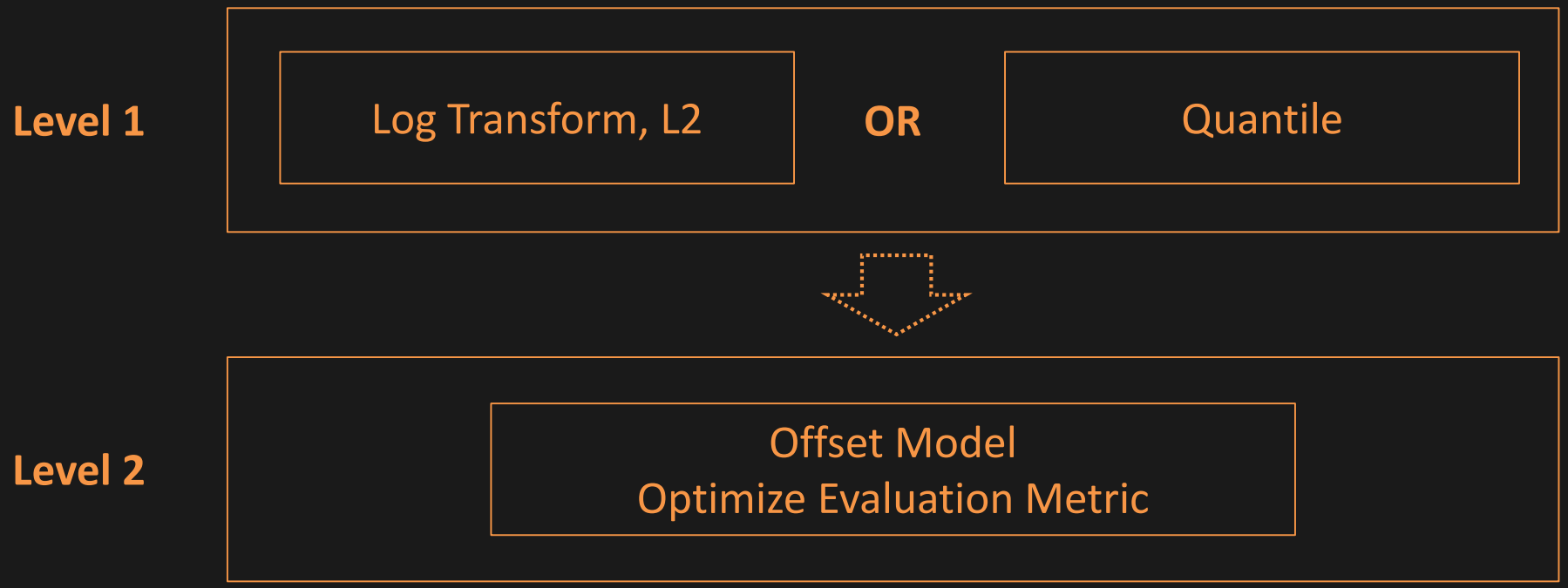


目标品类Para1容量规格推算



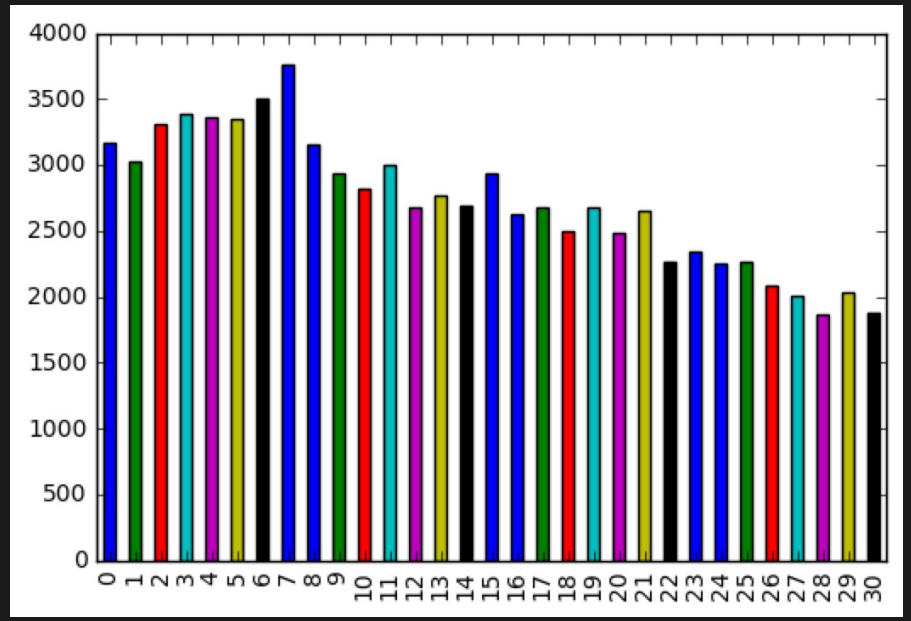
JDATA

评价指标优化 - S2


Level 1
Log Transform, L2
OR
Quantile
Level 2
Offset Model
Optimize Evaluation Metric

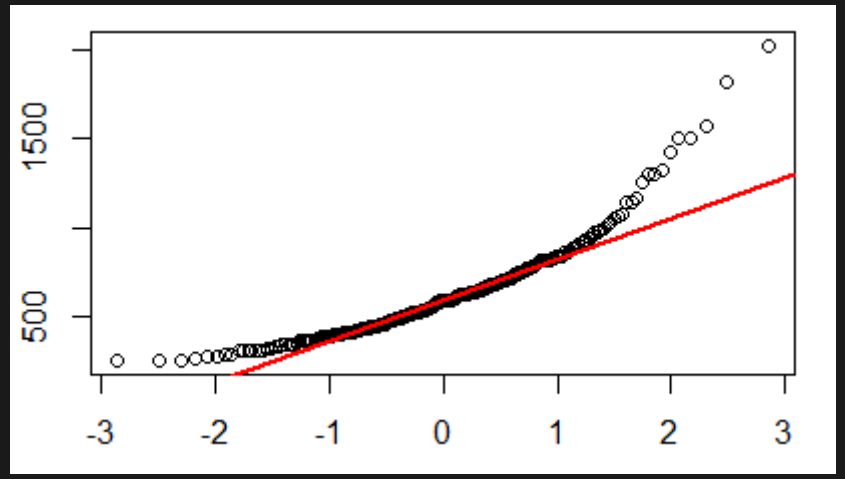
Level 1: Log Transform

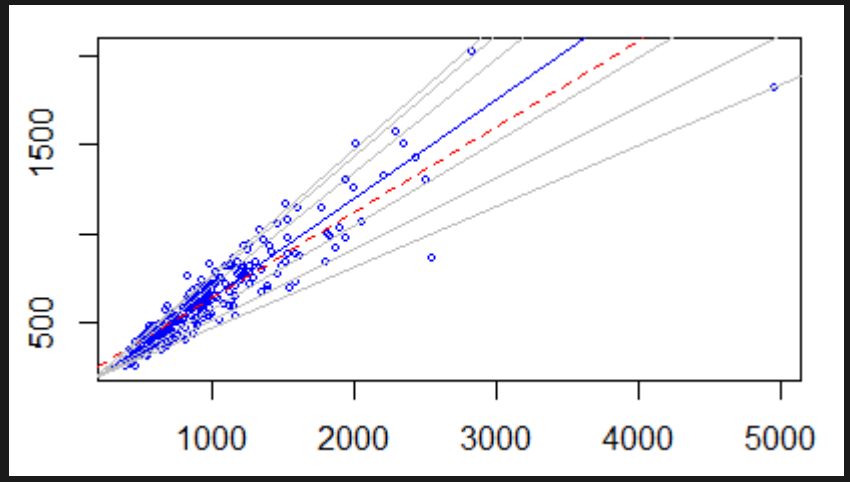
Y的分布



Y取值分布更靠近小的值，尽可能转换到normal distribution
所以对y做了log1p的转换，同时下降函数使用了S2的评价函数

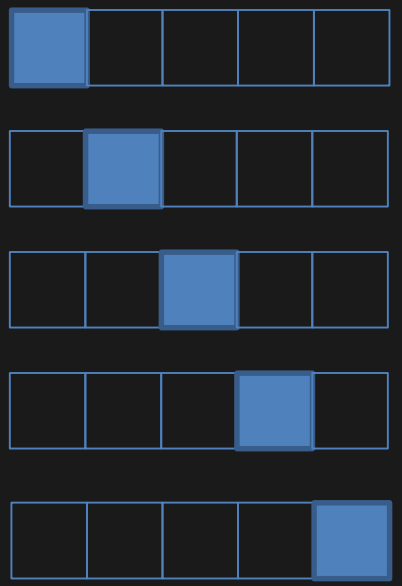
Level 1: Quantile Regression





- 因变量y明显不服从正态分布，但是呢，分位数回归不要求y服从正态分布，不仅如此，而且分位数回归还对异常值点不敏感
- 分位数回归可以拟合出多条直线，这个对于我们数据分布比较复杂的时候，很有用处，每条线反应了不同档次下，自变量与因变量的关系。实际上这个只是分位数回归的一小部分应用，得到不同分位点下的数据，我们还可以进行概率密度估计，得到相应的概率密度预测。

L2: Offset Model



Regression Model

label	pred
1	0.7
2	1.1
10	13.9

scipy.optimize.fmin_powell

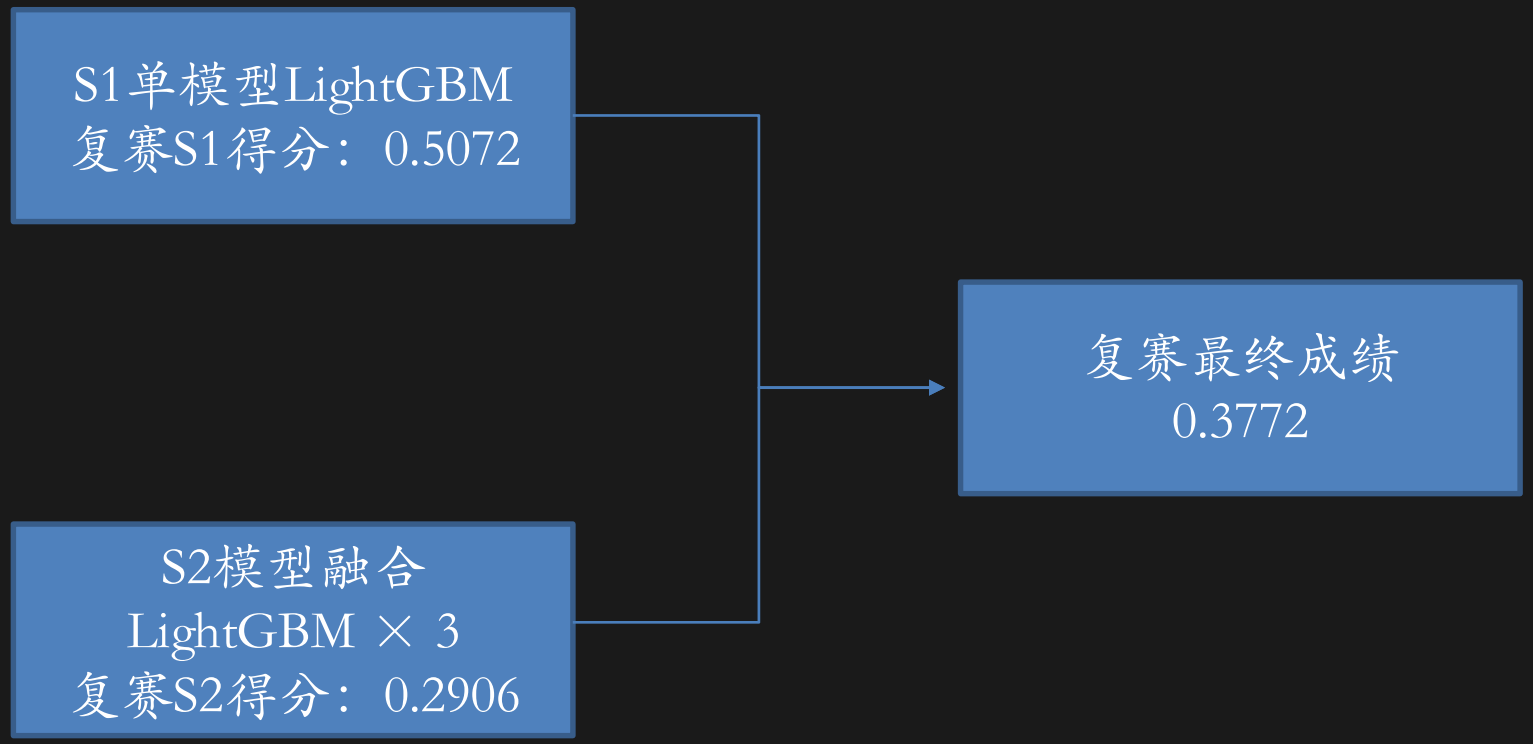
Offset Model

Offset 字典
1: 0.5
2: -0.5
...
10: 1.1
...
31: -5.2

pred_1	pred_2	pred_3	pred_4	pred_5	pred	pred offset
					1.1	2
					2.7	2
					10.1	11

JDATA

模型与融合


S1单模型LightGBM
复赛S1得分：0.5072
S2模型融合
LightGBM × 3
复赛S2得分：0.2906
复赛最终成绩
0.3772

目录

JDATA

经验总结

1. 对于时间跨度大类似于时序的问题，近期数据会比历史久远的数据相对更重要些
2. 构建稳定可靠的线下验证集，若在此之前就开始一味堆积特征维度，是没有意义的
3. 不需要从比赛初期就关注排行榜，细致做好每一个维度的特征
4. 从业务出发，换位思考用户行为的真实场景来构造特征，易于找到强特，且不过拟合
5. 针对评估指标（业务指标）对模型进行优化
6. **TRUST YOUR LOCAL CV ！**

感谢您的时间

Thanks

JDATA