

# Sales Num Prediction for Nature Month

update@2021-7-30

# 目录

- 预测目标说明
- 数据选取
- 难点
- 数据清洗与基本处理
- 极端值处理的几种方式
- 整体预测逻辑说明
- 以预测2021/2/1为例，用到3个lgb进行分类预测
- 以预测2021/2/1为例，对train data选取的说明
- 时序特征滑窗方式
- 特征工程
- 模型效果
- 从CommentNum推导到SalesNum的业务逻辑图
- 从CommentNum推导SalesNum的数学推导公式
- 数学推导公式中，取得weight/coef的方法及其优缺点
- CommentNum与SalesNum的效果展示
- 销量预测的常用应用
- 其它说明

# 预测目标说明

- (1) 预测JD平台上Remy数据的所有key的rolling 30的CommentNum



- (2) 预测JD平台上Remy数据的所有key的rolling 30的SalesNum



- (3) 预测JD平台上Remy数据的所有key的Nature Month的SalesNum <=> 利用数学公式, 从commentNum推导到salesNum <=> 预测JD平台上所有Remy数据的所有key的2021/2/1和2021/3/1的评论数

# 数据选取

- 方法一：选取每月中最后一次数据

特点：可用的数据量少，且部分月因为系统升级或搬迁，导致没有爬取到数据，即数据缺失；且，即使是每月的最后一次，也不一定是1/30或1/28这样的数据，有些是1/1或1/5或1/10爬取到；

- 方法二：选取每月中最早一次数据

特点：情况同方法一，但是相对而言，方法一比方法二的模型效果更佳；

- 方法三：选取所有数据（最终数据选取的方式）

特点：数据量多，但是脏数据也多，比如上下天数间隔只有1天的数据，出现较多冗余数据和脏数据；

# 难点

- 爬虫系统升级维护或者数据搬迁等，导致部分时间段没有数据；
- 数据值出现长尾现象，数据值分布不平衡；
- 数据爬取次数不平衡，有些历史仅1次记录，较难定位这些key的评论数的增加趋势；
- 爬取时间的不确定性，比如有些key是2021/3/5爬取到，然后2021/3/6又爬取到，造成数据冗余；
- Key的属性均为动态的，比如brand\_key, shop\_key, categ\_key，可能在三个月内是一致的，但是超过三个月可能就会有变动；

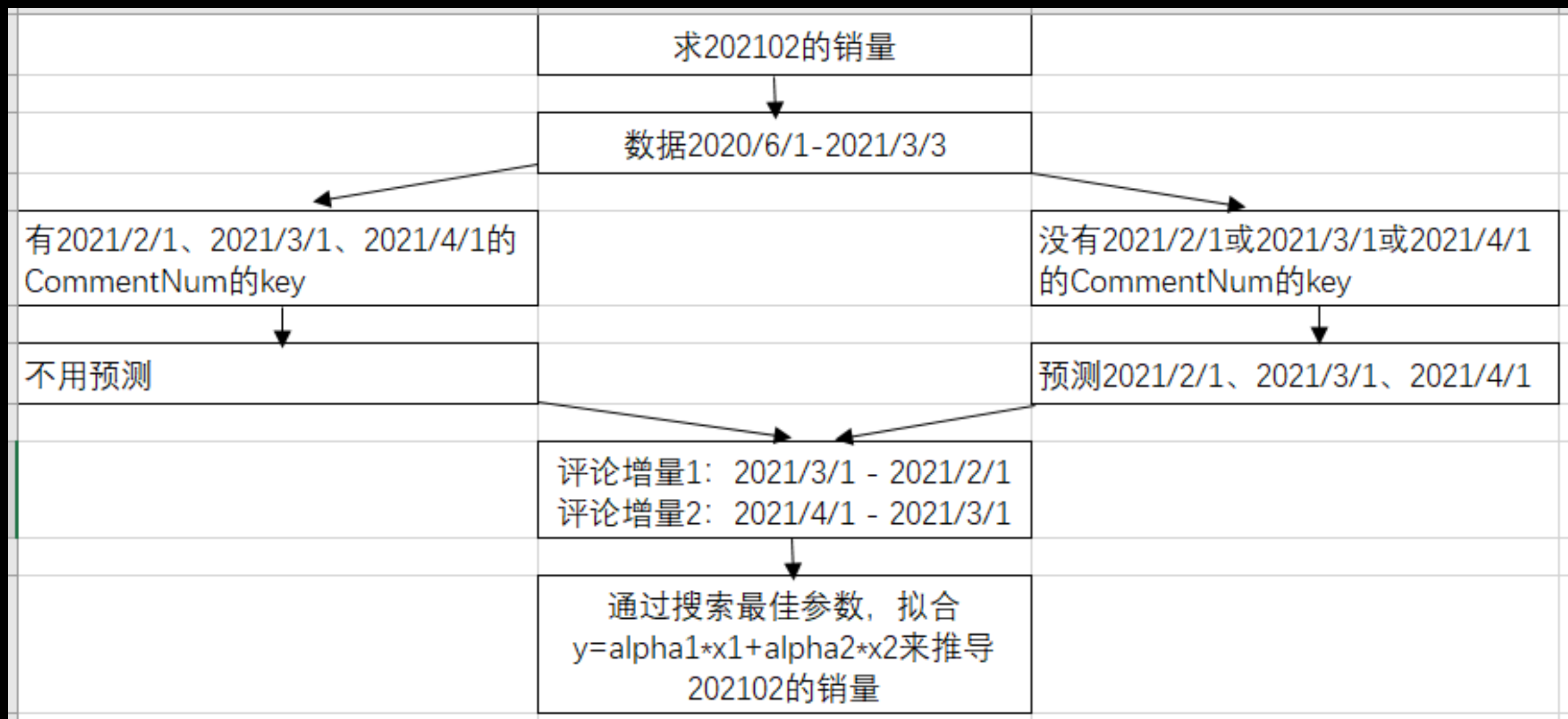
# 数据清洗与基本处理

- 去掉天数差 $<5$ 天, 且 $y\_diff=0$ 的数据
- 使用2轮, 去掉 $y\_diff<0$ 的数据
- 如果还出现 $y\_diff<0$ 的情况, 则去掉有这个情况的key的数据, 约200/60000; (该比例仅代表JD平台上的Remy的情况, 比如天猫2平台上的即饮饮料中的脏数据key占比是50000/83000)
- 对于brand\_key,shop\_key,categ\_key缺失的情况, 首先用其自身的众数进行补缺, 如果还有缺失值, 则用整体数据的众数进行补缺;

# 极端值处理的几种方式

- 方式一：删除，不做处理（我们最终选择的方式）
- 方式二：在模型训练的时候，加入对应的权重，比如对于极端评论数数据，给予的权重相应高，但效果提升不明显
- 方式三：加入一定的特征工程，去描述这类异常不平衡数据的趋势特征，对于通过一个key的预测的模型精度下降

# 整体预测逻辑展示





# 以预测2021/2/1为例， 用到3个lgb进行分类预测


针对于预测2021/2/1而言				
数据说明	模型名称	举例	滑窗window	说明
历史只有1次记录	lgb1	train: 历史记录cnt >= 2 test: 没有2021/2/1-2021/2/3的key,且历史记录cnt==1	1	对于重复预测的key, 经验证, 取预测偏大值更加
历史只有2-3次记录	lgb2	train: 历史记录cnt>=3 test: 没有2021/2/1-2021/2/3的key,且历史记录cnt>=2	2	
历史>=4次以上记录	lgb3	train: 历史记录cnt>=4 test: 没有2021/2/1-2021/2/3的key,且历史记录cnt>=3	3	

# 以预测2021/2/1为例，对train data选取的说明

- 方法一：只要符合cnt要求即可放入train\_data（适合lgb2+lgb3,mape\_per提升近2-4个百分点）
- 方法二：既要符合cnt要求又要符合有一个历史记录在y日期内（适合lgb1,mape\_per提升近2个百分点）
- 方法三：放入2021/1/1的预测数据，做rolling 30的预测2021/2/1

# 时序特征滑窗方式

方法一： 平移滑窗法(最终选择的训练数据构建方式)

平移滑窗法				
举例				
key	date	y	data_set	desc
1	2021/2/1	11	train	
1	2021/1/22	10		
1	2021/1/16	10		
1	2021/1/2	9		
2	2021/1/21	0	test	求其2021/2/1的评论数
				
key	date_0	date_1	data_set	desc
1	2021/2/1	2021/1/22	train	特点： 直接上下次连接滑窗
1	2021/1/22	2021/1/16		
1	2021/1/16	2021/1/2		
2	2021/2/1	2021/1/21	test	

方法二：不固定规则滑窗法

不规则滑窗法				
举例				
key	date	y	data_set	desc
1	2021/2/1	11	train	
1	2021/1/22	10		
1	2021/1/16	10		
1	2021/1/2	9		
2	2021/1/21	0	test	求其2021/2/1的评论数
key	date_0	date_1	data_set	desc
1	2021/2/1	2021/1/22	train	特点：train中需要预测的日期均一致
1	2021/2/1	2021/1/16		
1	2021/2/1	2021/1/2		
2	2021/2/1	2021/1/21	test	

# 特征工程

特征工程	
1.自身特征群	
	月份、上下旬
	本身的基础评论数
	上下间隔天数、上下间隔评论增加数、上下间隔平均每天评论增加数
	历史每次间隔平均评论增加数、历史平均每天评论增加数的统计属性群
2.群体特征群	
	相同brand_key及评论基数的群体的上下间隔评论增加数量、上下间隔平均每天评论增加数的统计属性群
	相同shop_key及评论基数的群体的上下间隔评论增加数量、上下间隔平均每天评论增加数的统计属性群
	相同categ_key及评论基数的群体的上下间隔评论增加数量、上下间隔平均每天评论增加数的统计属性群
	相同月份及评论基数的群体的上下间隔评论增加数量、上下间隔平均每天评论增加数的统计属性群
3.消费趋势特征群	
	特殊节日下的消费趋势的统计属性

# 模型效果

train eval:  
0.9995341266481537 2337.69877719062 12.699698184448343  
R2 : 0.999534, MSE : 2337.734349, MAE : 12.670541

error_num	cnt	ratio
0	0-5	227348 0.746524
1	10以上	56097 0.184201
2	5-10	21097 0.069275

mape_per	cnt	ratio
0	0-0.01	183367 0.602107
1	0.01-0.02	47948 0.157443
2	0.02-0.03	21941 0.072046
3	0.03-0.04	12755 0.041883
4	0.04-0.05	7701 0.025287
5	0.05以上	30599 0.100475
6	999999	231 0.000759

train eval:  
0.9995643269815744 2235.953054935474 11.746805440601076  
valid eval:  
0.998951941360815 4446.813644480362 14.344523982479922  
R2 : 0.999698, MSE : 1573.984961, MAE : 8.940095

error_num	cnt	ratio
0	0-5	433013 0.784006
1	10以上	80993 0.146645
2	5-10	38302 0.069349

mape_per	cnt	ratio
0	0-0.01	373652 0.676528
1	0.01-0.02	67994 0.123109
2	0.02-0.03	32414 0.058688
3	0.03-0.04	20546 0.037200
4	0.04-0.05	13309 0.024097
5	0.05以上	44306 0.080220
6	999999	87 0.000158

train eval:  
0.9996928420469346 1793.9323856156655 11.732735392439746  
valid eval:  
0.9989871447071103 5610.193803041768 15.927185012561182  
R2 : 0.999770, MSE : 1364.979905, MAE : 9.010602

error_num	cnt	ratio
0	0-5	255195 0.768606
1	10以上	51016 0.153652
2	5-10	25812 0.077742

mape_per	cnt	ratio
0	0-0.01	230817 0.695184
1	0.01-0.02	45608 0.137364
2	0.02-0.03	19997 0.060228
3	0.03-0.04	11757 0.035410
4	0.04-0.05	7092 0.021360
5	0.05以上	16749 0.050445
6	999999	3 0.000009

# 从CommentNum推导到SalesNum的业务逻辑图

202102销量						
train: 2020/9/1-2021/3/3			C30	weight	占比	当月评论增量
2021/2/3-2021/3/2的评论新增 +16	2020/12/3-2021/1/2的销量	系统评论	14		87.50%	(1)
	2021/1/3-2021/2/2的销量	自动评论	2		12.50%	
	2021/2/3-2021/3/2的销量	自动评论		w1		
2021/3/3-2021/4/2的评论新增 +18	2021/1/3-2021/2/2的销量	系统评论	16		88.89%	(2)
	2021/2/3-2021/3/2的销量	自动评论	2	w2	11.11%	
	2021/3/3-2021/4/2的销量	自动评论				
2021/4/3-2021/5/2的评论新增 +26	2021/2/3-2021/3/2的销量	系统评论	22	w3	84.62%	(3)
	2021/3/3-2021/4/2的销量	自动评论	4		15.38%	
	2021/4/3-2021/5/2的销量	自动评论				
理论数学公式:	y=w1*(1) + w2*(2) + w3*(3)					

# 从CommentNum推导SalesNum的数学推导公式

理论数学公式:	$y = w1*(1) + w2*(2) + w3*(3)$
使用2个月的推导公式 相比单月推导，效果更佳	$\begin{aligned} y &= w1*(1) + w2*(2) + w3*(3) \\ &= w1*(1) + w2*(2) + w3*ratio2*(1) \\ &= (w1 + w3*ratio2) * (1) + w2*(2) \end{aligned}$ <p>又因为，w1,w2,w3,ratio1,ratio2都是常数 所以，令 <math>w1 + w3*ratio2 = \alpha1</math> , <math>w2 = \alpha2</math> 所以，<math>y = \alpha1 * (1) + \alpha2 * (2)</math>，所以我们的目标就转换成了找到最佳alpha1和alpha2</p>
使用1个月的推导公式	$\begin{aligned} y &= w1*(1) + w2*(2) + w3*(3) \\ &= w1*(1) + w2*ratio1*(1) + w3*ratio2*(1) \\ &= (w1 + w2 * ratio1 + w3*ratio2) * (1) \end{aligned}$ <p>又因为，w1,w2,w3,ratio1,ratio2都是常数 所以，令 <math>w1 + w2 * ratio1 + w3*ratio2 = \alpha</math> 所以，<math>y = \alpha * (1)</math>，所以我们的目标就转换成了找到最佳alpha</p>



# 数学推导公式中，取得weight/coef的方法及其优缺点

- 方式一：从预测的comment\_num与实际的sales\_num中搜索最佳参数  
(最终选择放入框架的)

优点：提取最佳参数比较方便

缺点：不通用，且有真实的sales\_num的数据量太少，仅100/60000，所以不可以用100个去通用到60000个key上，但是可以从60000个key上的结果通用到100个上；

- 方式二：从mv\_sim中的c30提取

优点：不用依赖用户的真实sales num表单

缺点：如果使用 $y=w1*x1+w2*x2+w3*x3$ ，则时间受到限制，即2021/3/1的时候，是无法获得2021/5/1的c30的数据的，且c30不能完全代表w1，即w1是c30中的一部分；

- 方式三：使用LinearRegression等算法做从comment\_num到sales\_num的回归预测

优点：方便，通用

缺点：同方式一，真实的sales num数据量太少，仅100/60000，无法可靠训练；

- 方式四：使用迁移学习，将taobao平台上即饮饮料的key从comment num推导到sales num的weight/coef，迁移到，JD平台上的Remy品类的key的从comment num推导到sales num的weight/coef；(后续验证发现com\_goods\_statistics\_monthly\_v2与f\_platform\_goods\_his不match，且f\_platform\_goods\_his中平台3的即饮饮料的date\_key截止到2021-03-08之后就没了数据,详见下附表)

优点：方便，通用，可靠

缺点：需要先验证迁移属于正向还是负向，若是正向迁移，则可用，否则不可用

表1			表2						
com_goods_statistics_monthly_v2 上的true sales num			f_platform_goods_his上的true comment num			如果表1和表2是match的话，应该是如下 这样的数据			
key	qty_m_current	month	key	total_comment_num	date_key	key	total_comment_num	date_key	
44347462	1076	202106	44347462	15	20210307	44347462		20210307	
44347462	7	202105	44347462	14	20210227	44347462	11946	20210227	
44347462	1751	202104	44347462	14	20210220	44347462		20210220	
44347462	1459	202103	44347462	13	20210215	44347462		20210215	
44347462	1359	202102	44347462	13	20210207	44347462		20210207	
44347462	2424	202101	44347462	13	20210131	44347462	10587	20210131	
44347462	5967	202012	44347462	12	20210125	44347462		20210125	
44347462	2196	202011	44347462	10	20210117	44347462		20210117	
			44347462	8	20210110	44347462		20210110	
			44347462	7	20210105	44347462		20210105	
			44347462	6	20201226	44347462	8160	20201226	
			44347462	5	20201219	44347462		20201219	
			44347462	3	20201212	44347462		20201212	
			44347462	3	20201205	44347462	2200	20201205	
			44347462	0	20201128	44347462	0	20201128	

# CommentNum与SalesNum的效果展示

-202101销量

A	B	C	D	E	F	G	H	I
求的月份	使用月份	alpha1	alpha2	alpha3	r2_score	mse	mae	有效key的数量
求202101	用连续三个月	0.9	0.9	nan	0.951892	1146.916	16.96154	65
求202101	用单月202101	1.1	nan	nan	0.932527	1721.517	14.95692	65

# CommentNum与SalesNum的效果展示

## -202012销量

求的月份	使用月份	alpha1	alpha2	alpha3	r2_score	mse	mae	有效key的数量
求202012	用连续三个月	0.3	0.3	0	0.723879	1474.163	16.596	50
求202012	用单月202012	0.8	nan	nan	0.645365	1762.438	18.236	50
求202012	用单月202101	0.5	nan	nan	0.642826	1810.368	18.16176	68
求202012	用单月202102	2	nan	nan	0.346736	3218.882	28.91176	68

# CommentNum与SalesNum的效果展示

## -202011销量

求的月份	使用月份	alpha1	alpha2	alpha3	r2_score	mse	mae	有效key的数量
求202011	用连续三个月	1.7	0.4	0	0.905237	605.4082	12.53725	51
求202011	用单月202012	1	nan	nan	0.835313	1238.192	18.38462	52
求202011	用单月202011	2	nan	nan	0.662888	1025.843	12.27451	51
求202011	用单月202101	0.7	nan	nan	0.583656	4176.831	26.63043	69

# 销量预测的常用应用

(1) 本身店铺中的爆款等商品的分析;

(2) 竞争店铺的定位;

(3) 网店营销策略数据参考, 如做促销, 提供优惠券, 送小样, 或者结合比价系统做促销价定位等;

(4) 库存管理;

(5) 采购管理等;

# 其它说明

- 因为我们使用的数据是CommentNumber数据，所以可以保证CommentNumber预测的准确率；
- 因为remy用户，一个月仅提供了100个以内的key样本的真实sales num数据，所以推导的alpha参数并不能通用到整个Remy上60000个key，所以不能保证从CommentNumber推导到SalesNumber的准确度；
- 一般情况下，如果需要做销量预测，就需要提供对应的一定量的销量数据，而不是评论数据；



谢 谢