2018 CCF BDCI 6th
大数据与计算智能大赛
Big Data & Computational Intelligence Contest

《供应链需求预测》

团队名称： Miracccccccle
团队成员： 陈瑀 余薇 张浩 李航 闵子剑

CONTENTS

赛题分析

数据：

- 执御平台上沙特阿拉伯市场的历史数据积累，数据时间跨度为2017年3月1日至2018年3月16日数据。
- 商品属性数据，商品的市场表现数据，营销信息数据，也可爬取宗教和天气数据。

目标：

- ✓ 预测2018年5月1日起，未来5周的sku_id的销量值。

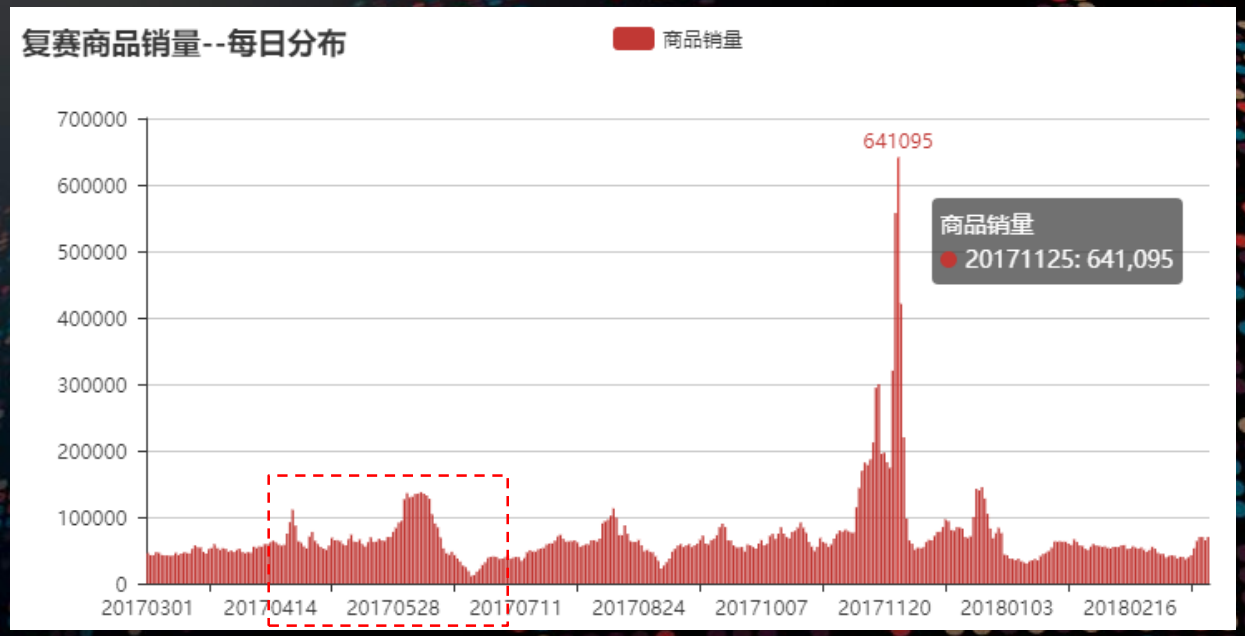
分析：

- 回归问题
- 单个goods_id包含多个sku_id
- 时序周期性

执古之道，以御今之有

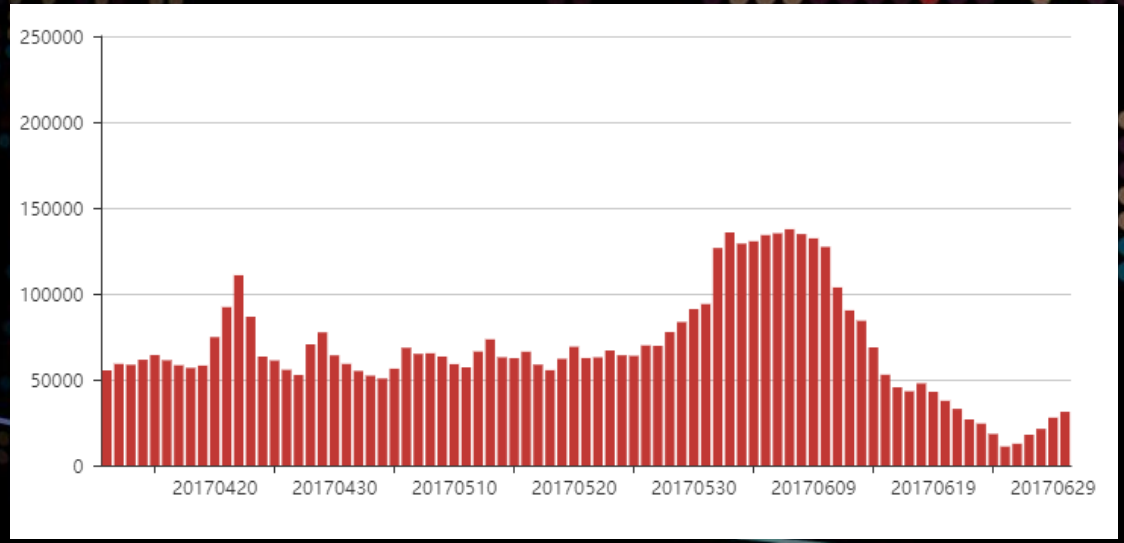
赛题分析

- 1、历史商品销量统计中，发现11月25日出现了波峰，销量总计为**641,095**。且前后销量均增大，存在异常**推测为节假日**。猜测为白色星期五，平台购物节。



区域缩放

- 2、通过可视化的区域缩放，观测了去年（2017年）5月-6月的销量趋势，猜测5周的销量并不是稳定分布的，**存在销量波动**，猜测待预测的5周销量也是呈现一定波动性。



算法模型

数据划分(亮点一：寻找最优分布)

- 根据线上规则分布在历史区间寻找最优分布区域

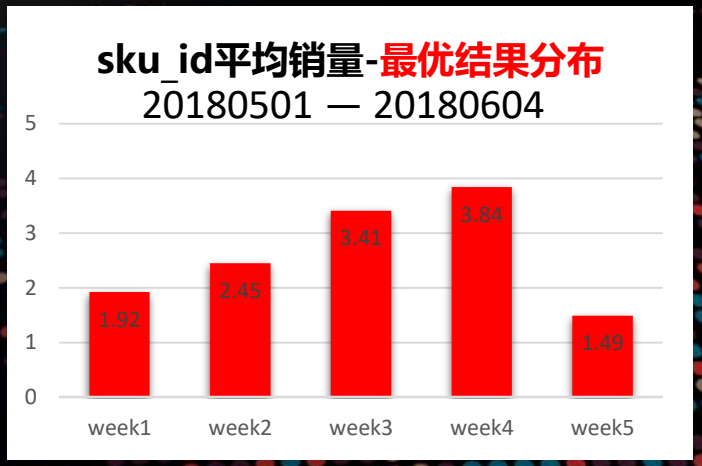
规则尝试：

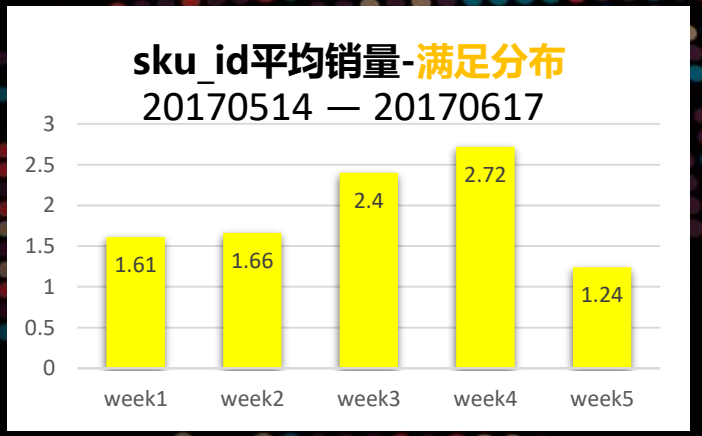
1、goodsale表中，data_date区间为20180214~20180316；
2、计算sku_id**销量均值乘以不同的比例**赋予week1~week5；
3、纯规则初赛A榜可达0.095，复赛A榜可达0.078的成绩。

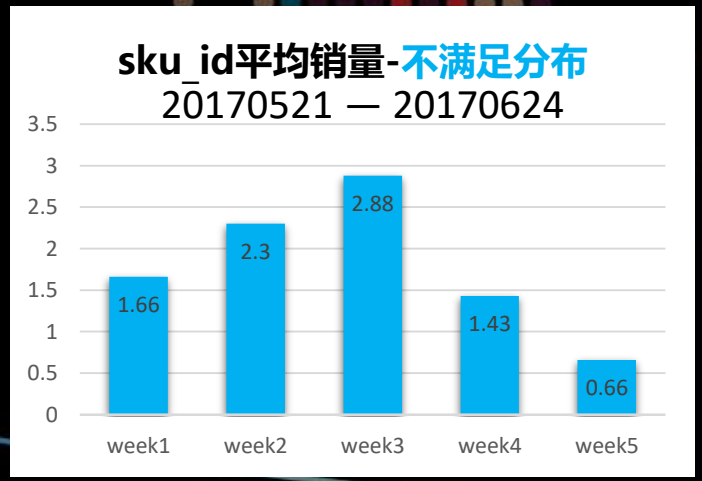
寻找最优分布：

根据规则已经可达较优预测结果，我们根据这个线上的每周均值分布，在历史区间中，寻找最满足分布的训练集的标签区间。

最符合分布的标签区间	
20170512 — 20170615	20171129 — 20180102
20170513 — 20170616	20171130 — 20180103
20170514 — 20170617	20171201 — 20180104
20171127 — 20171231	20171202 — 20180105
20171128 — 20180101	20171203 — 20180106







算法模型

特征工程(常规特征)

主要按照特征群进行提取：基础统计特征，离散特征，时序相关特征

基础统计特征

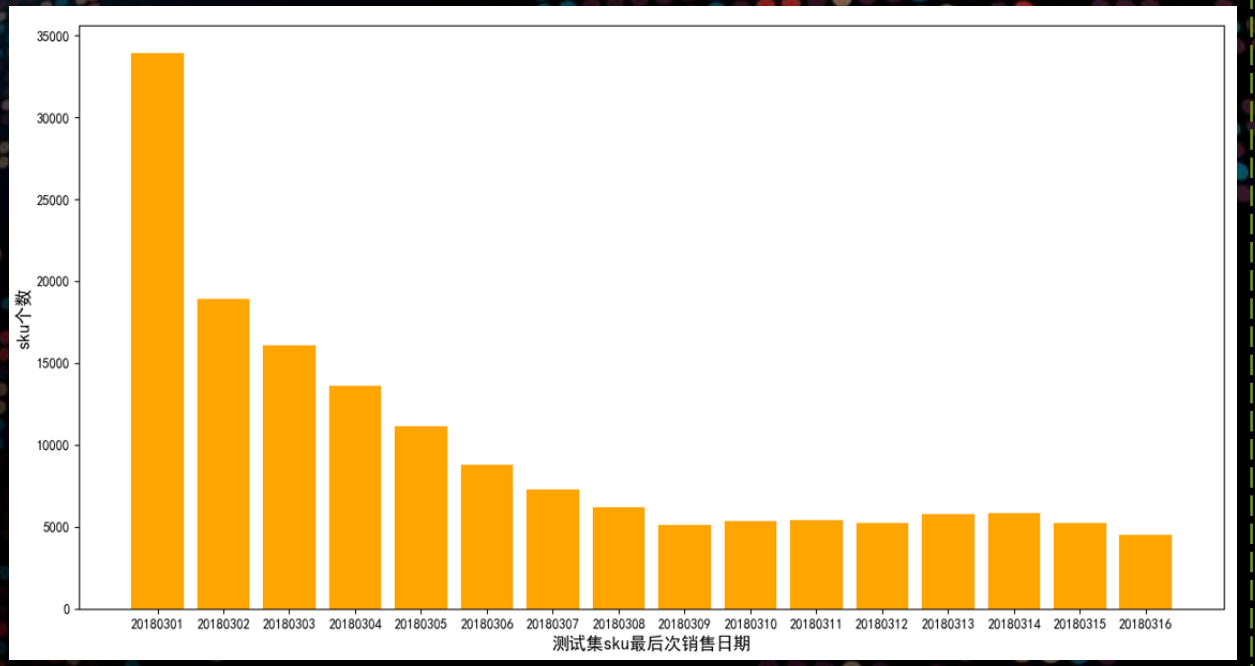
- goods点击次数最值均值等
- goods加购次数最值均值等
- goods收藏次数最值均值等
- goods购买次数最值均值等
- ……

离散特征

- 商品季节属性
- 平台活动类型
- 商品类目属性
- 节假日特征
- ……

时序相关特征

- 按时间粒度统计goods销售量最大值、最小值、均值、求和排名
- 按时间粒度统计sku销售量最大值、最小值、均值、中位数、求和排名
- ……



算法模型

特征工程(亮点二：爬取外部数据)

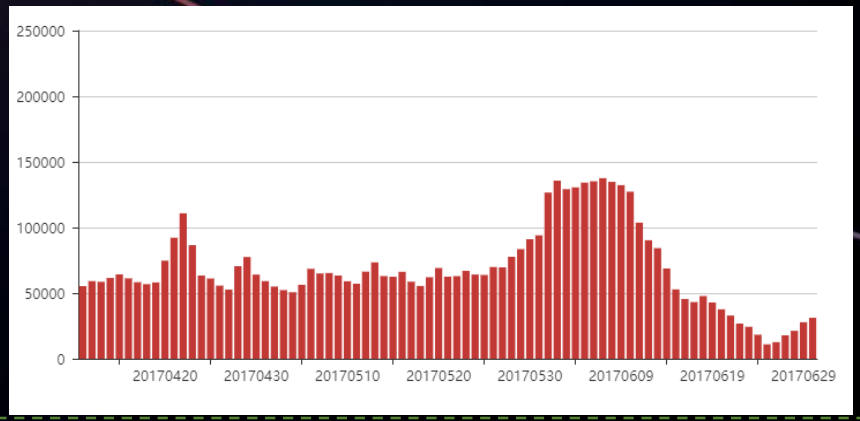
➢ 斋月

时间： 2017年： 5月26日—6月26日

2018年： 5月16日—6月16日

特点： 1、商家白天不营业，跨境电商销量大幅增长。

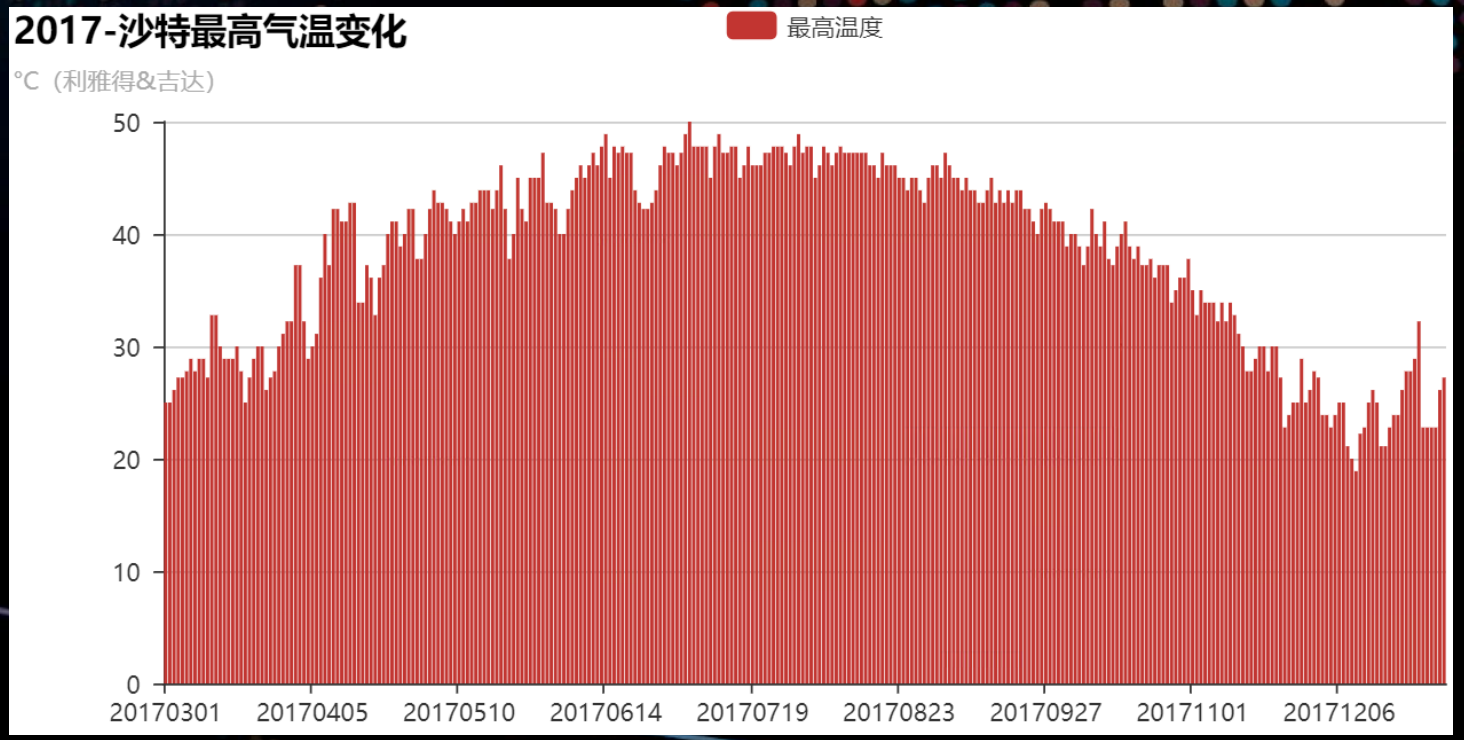
2、人们为过节做准备，并且第14天为沙特儿童节。



➢ 气候

特点： 1、气温高，有世界“热极”之称。

2、沙特降水极少，年降雨量不足200毫米。



算法模型

特征工程(亮点三：排名特征)

- 排名特征1（sku局部排名）：体现出用户更喜欢同一goods下哪一个sku，反映**用户偏好**。
- 排名特征2（sku全局排名）：体现出sku的销量竞争力和购买程度，反映**商品热度**。

goods_id	sku_id	sku销量	sku全局排名	sku局部排名
1	1_1	200	4	2
1	1_2	100	5	3
1	1_3	100	5	3
1	1_4	300	3	1
2	2_1	400	2	2
2	2_2	200	4	3
2	2_3	500	1	1

对模型有**6个万分位**的涨幅。

算法模型

模型构建

回归问题

模型

linear regression

V

xgboost

V

lightgbm

线性模型：最后n天滑窗统计、天气宗教信息

树模型：全部特征、调整最后n天滑窗统计、排名特征

规则　统计学的方法寻找数据分布规律

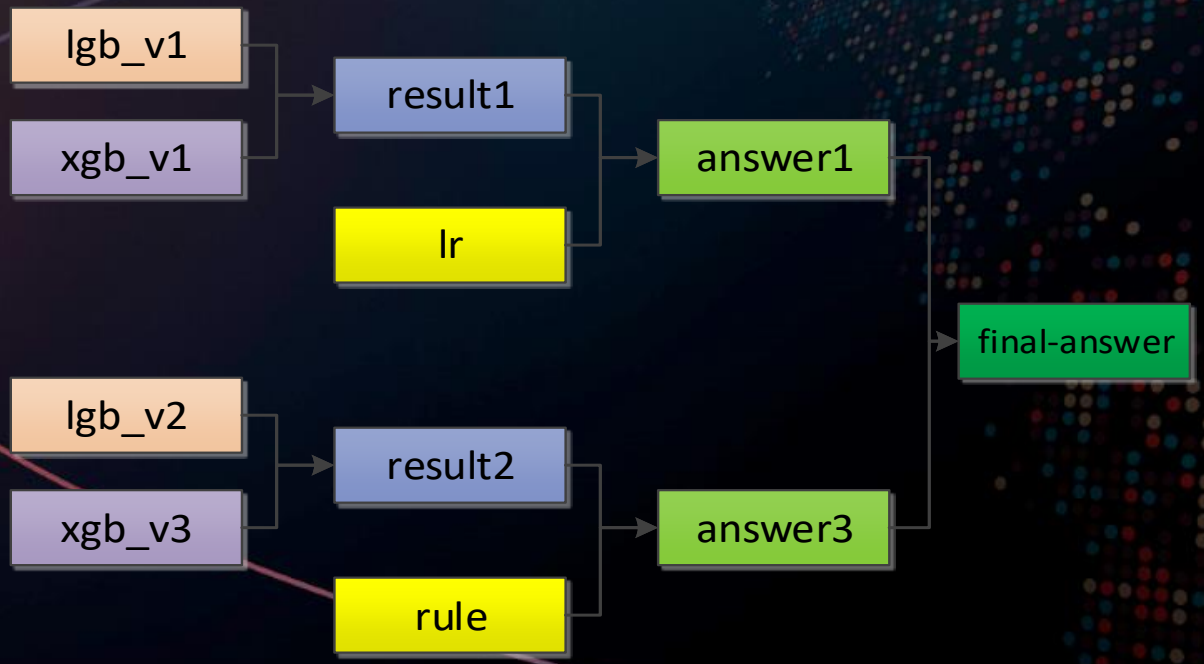
如何构建模型差异性?

- 1、训练集扰动：构建不同训练集（初赛、复赛数据拼接），数据滑窗的不同
- 2、参数扰动：为模型构建不同的参数设置
- 3、特征扰动：构建不同特征群（抽样特征、构建有区分度的特征）

算法模型

模型集成

- 基于“树形结构” 的融合



多模型线性加权融合

↓

基于树形结构的由低到高线性加权融合

↓

权重更好确定，模型得分涨幅明显

方案潜力

- 稳定性：大量的业务分析与数据分析可视化分析，确保模型的稳定性。
- 可扩展性：结合用户信息，可扩展到用户购买预测、商家广告精准投放、优惠信息及时推送等场景下。
- 泛化能力：具有高可解释性的特征工程以及外部信息的引入，使得对于非特殊日期亦有较好的预测效果。

总结提升

提升：

- 加入商品促销信息
- 爬取石油、货币汇率等数据进行特征提取

总结：

- 多从业务层面思考问题
- 对数据进行深入的统计分析
- 不到最后一刻，永不放弃

致谢

感谢组织单位全体相关工作人员的辛勤付出

感谢DataFountain为我们提供良好的比赛环境

感谢浙江执御信息技术有限公司给予的宝贵数据

感谢王进老师的悉心指导

感谢Miracccccccle团队全员一直以来坚持不懈的努力

感 谢 观 看　请 您 提 问