CCF BDCI CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST

2023 CCF 大数据与计算智能大赛 11th

线上线下全场景生鲜超市库存履约一体化决策

是队伍名不是大名

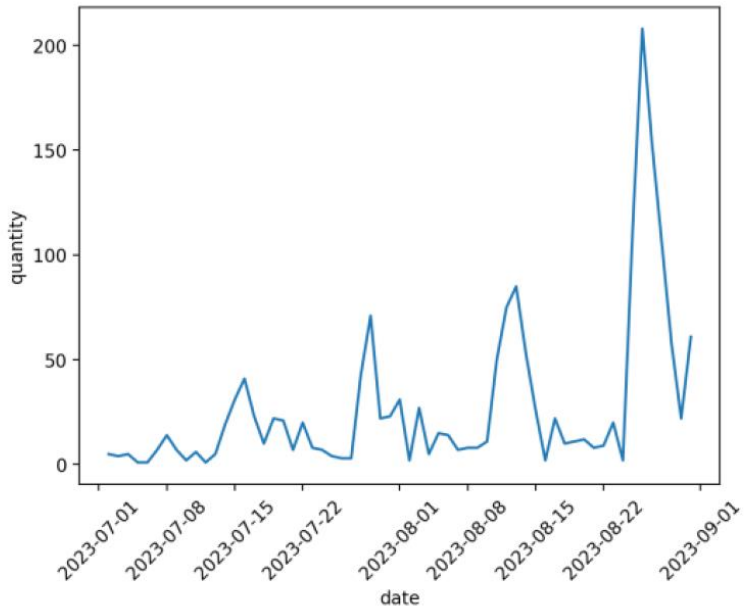
陈雨荃 刘怡鹏

目录

CONTENT

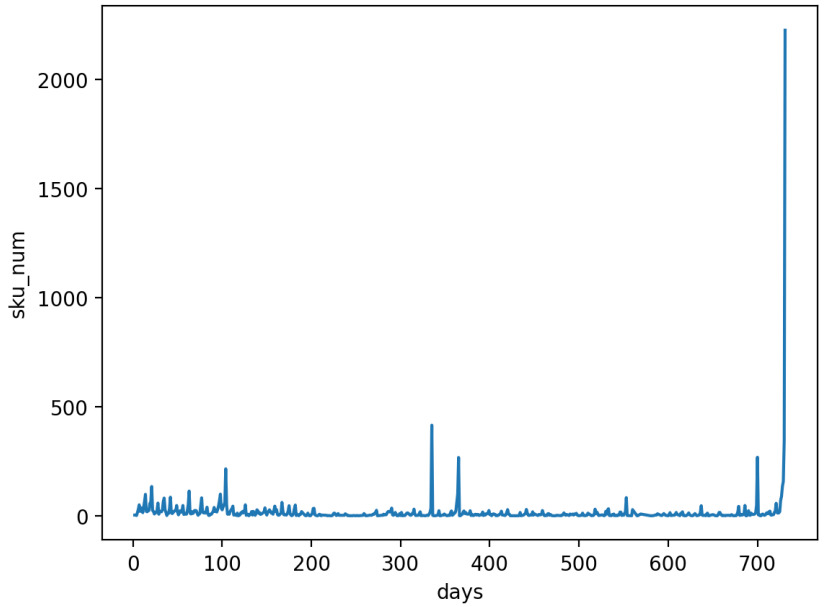
探索性数据分析

- 某些商品常常具有周期性，因此在后续的特征工程中我们需要融入年月周等信息。



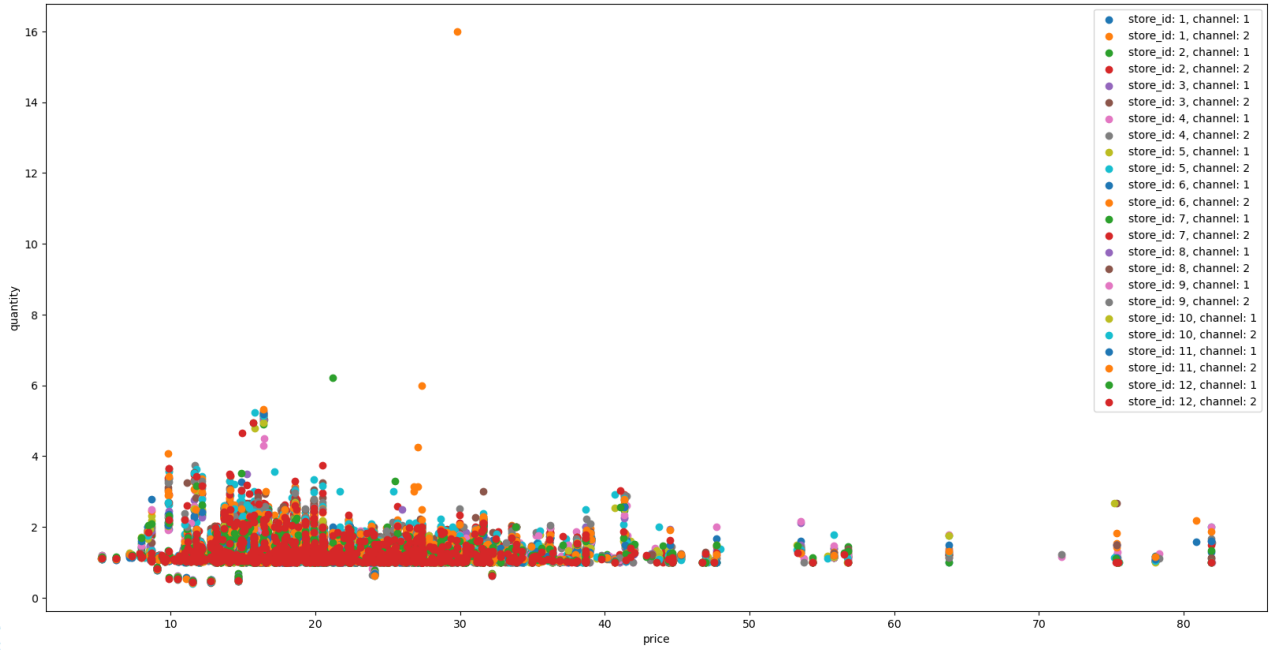
探索性数据分析

- sku的开售日期并不是固定的，某些sku可能最近才刚引入商店，因此会存在不同的sku的数据量不一样的情况（即长尾分布问题）

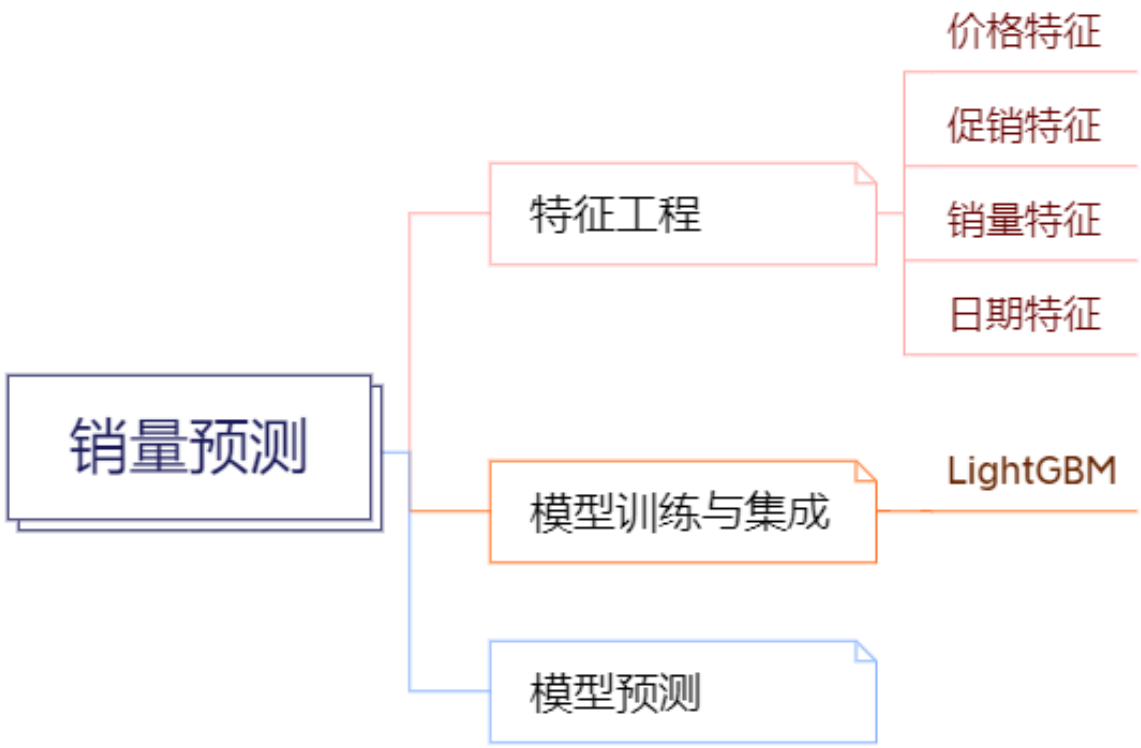


探索性数据分析

- 商品的价格会一定程度影响销量，价格越低，销量越高的概率越大。



销量预测



销量预测——特征工程——价格特征

- 对相同(store_id, sku_id)的商品填充从开始日期到最大日期之间的所有日期
- 填充后的缺失值：向前填充
- 特征构建
  - 原始价格最大值、最小值、平均值、标准差、中位数、偏度
  - 原始价格分位数（25%、75%、50%、5%、95%、10%、90%）
  - 原始价格归一化（当前值除最大值）
  - 商品原始价格数量（商品有多少种原始价格）
  - 相同原始价格的商品数量
  - 相对前一天的原始价格变化率
  - 相对全年的价格变化率
  - 相对全月的价格变化率

销量预测——特征工程——价格特征

- 分类编码：对相同的键，对原始价格进行求均值和标准差，这样可以得出不同类别类内的统计特征。

```
cols = [
    ['store_id'],
    ['sku_id'],
    ['channel'],
    ['store_id', 'sku_id'],
    ['store_id', 'channel'],
    ['sku_id', 'channel'],
    ['store_id', 'sku_id', 'channel'],
    ['item_first_cate_cd'],
    ['item_first_cate_cd', 'item_second_cate_cd'],
    ['item_first_cate_cd', 'item_second_cate_cd', 'item_third_cate_cd'],
    ['brand_code'],
    ['item_first_cate_cd', 'item_second_cate_cd', 'item_third_cate_cd',
'brand_code']
]
```

销量预测——特征工程——促销特征

- EDA的发现：在相同 (store_id, sku_id, date, channel) 下，促销最多有3个
- 对单个商品的所有促销展开，nan填充为0

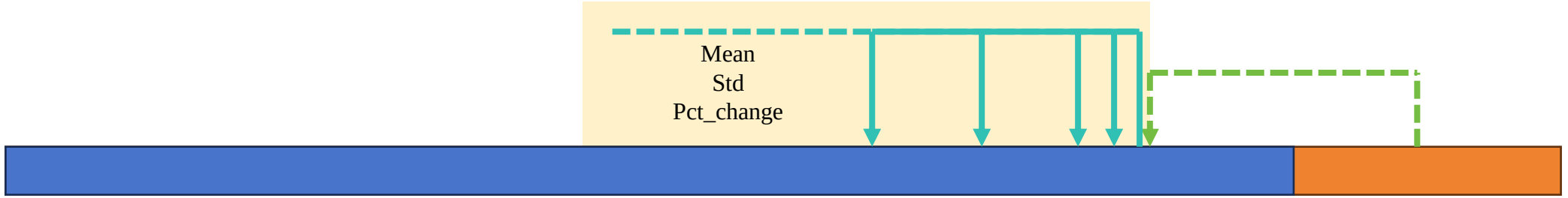
	store_id	sku_id	date	promotion_id	curr_day	total_days	promotion_type	threshold	discount_off	channel
151251	1	434	2021-08-31	96	5	14	2	5.0	0.14	1
1432693	1	434	2021-08-31	1092	6	15	5	1.0	0.27	1

	store_id	sku_id	date	channel	promotion_id_0	promotion_id_1	promotion_id_2	curr_day_0	curr_day_1	curr_day_2	...
108649	1	434	2021-08-31	1	96	1092	0	5	6	0	...

```
Index(['store_id', 'sku_id', 'date', 'channel', 'promotion_id_0',
       'promotion_id_1', 'promotion_id_2', 'curr_day_0', 'curr_day_1',
       'curr_day_2', 'total_days_0', 'total_days_1', 'total_days_2',
       'promotion_type_0', 'promotion_type_1', 'promotion_type_2',
       'threshold_0', 'threshold_1', 'threshold_2', 'discount_off_0',
       'discount_off_1', 'discount_off_2'],
      dtype='object')
```

销量预测——特征工程——销量特征

- 使用(store_id, sku_id, date, channel)对quantity求和统计销量
- 计算预测窗口前x天的销量均值、标准差、均值变化，最后对填充后的nan值填充为0




```
PREDICT_WINDOW = 14

for day in [1,3,5,7,14,21,30,60,90]:
    print("processing day: ", day)
    data_merged[f'sales_{day}_pw_mean'] = data_merged.groupby(['store_id', 'sku_id', 'channel'])['quantity'].transform(lambda x: x.shift(PREDICT_WINDOW).rolling(day).mean())
    data_merged[f'sales_{day}_pw_std'] = data_merged.groupby(['store_id', 'sku_id', 'channel'])['quantity'].transform(lambda x: x.shift(PREDICT_WINDOW).rolling(day).std())
    data_merged[f'sales_{day}_pw_mean_change'] = data_merged.groupby(['store_id', 'sku_id', 'channel'])['quantity'].transform(lambda x: x.shift(PREDICT_WINDOW).rolling(day).mean(

data_merged.fillna(0, inplace=True)
data_merged
```

销量预测——特征工程——日期特征

- 年份、月份、月内第几天、周内第几天、年内第几天、季度
- 是否月初、是否月末、是否季初、是否季末、是否年初、是否年末、是否闰年

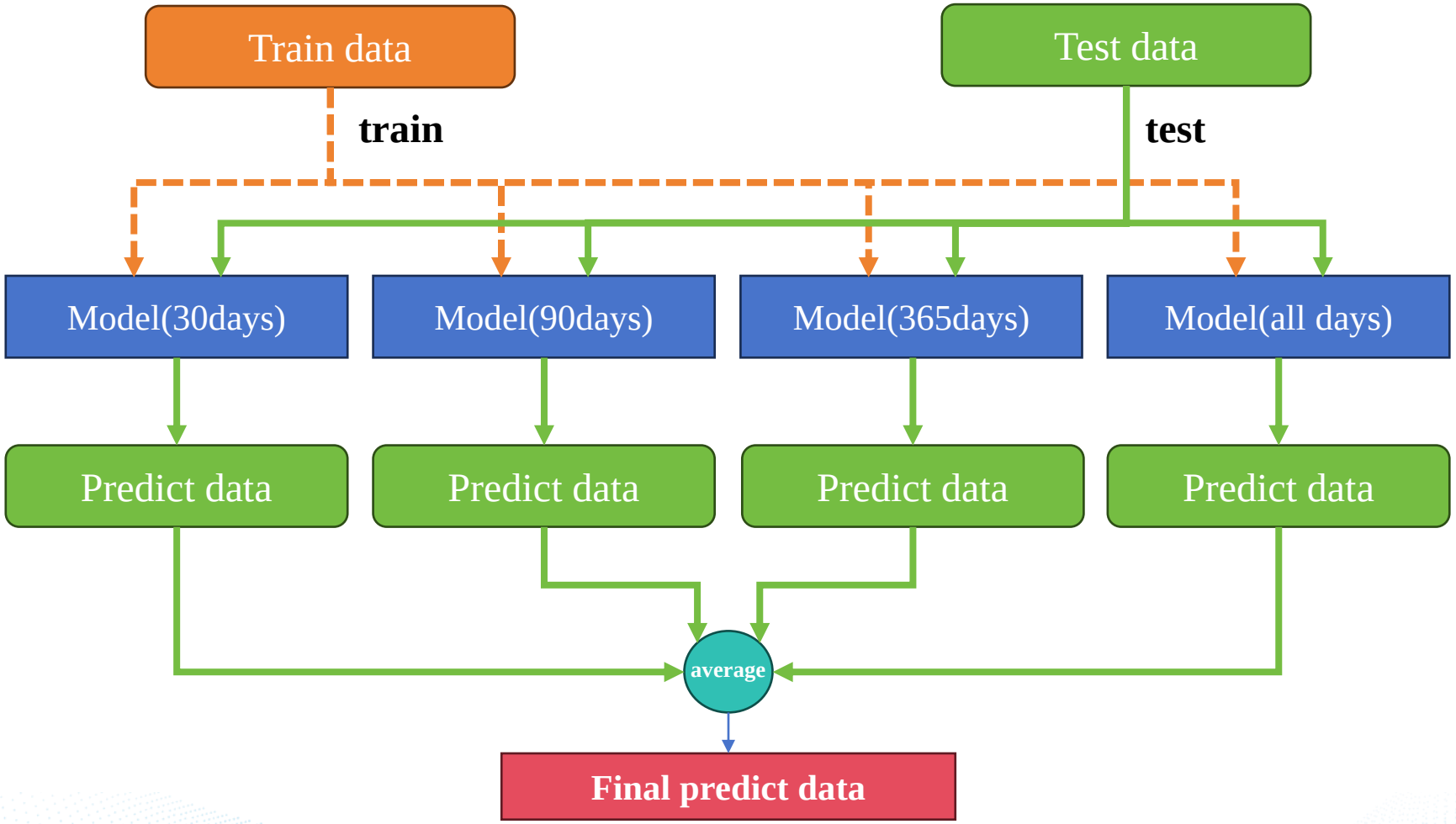
销量预测——模型训练与集成

- LightGBM tweedie回归
  - Baseline：Kaggle M5竞赛中的Top1解决方案
  - X: 除了date和quantity之外的所有列
  - y: quantity

```
lgb_params = {
    'boosting_type': 'gbdt',
    'objective': 'tweedie',
    'tweedie_variance_power': 1.1,
    'metric': 'rmse',
    'subsample': 0.5,
    'subsample_freq': 1,
    'learning_rate': 0.015,
    'num_leaves': 2**11-1,
    'min_data_in_leaf': 2**12-1,
    'feature_fraction': 0.5,
    'max_bin': 100,
    'n_estimators': 3000,
    'boost_from_average': False,
    'verbose': -1,
    # 'device': 'gpu'
}
```

销量预测——模型训练与集成

- 问题：**商品开售时间不同**，存在时间序列的不一致性、特征的不平衡、过拟合等各种问题
- 方案：使用不同时间段的数据来训练多个模型然后进行集成
  - 2023-09-01前30、90、365、所有天，共四份数据
  - 使用同一模型结构训练
  - 对四个不同模型进行集成（平均处理）


Train data
Test data
train
test
Model(30days)
Model(90days)
Model(365days)
Model(all days)
Predict data
Predict data
Predict data
Predict data
average
Final predict data

销量预测——模型预测

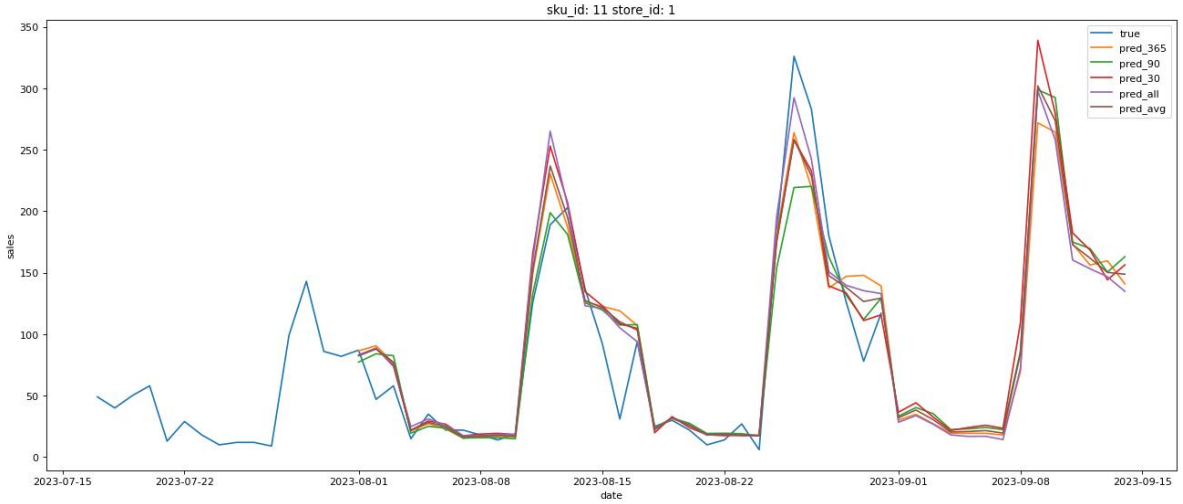
- 输出示例

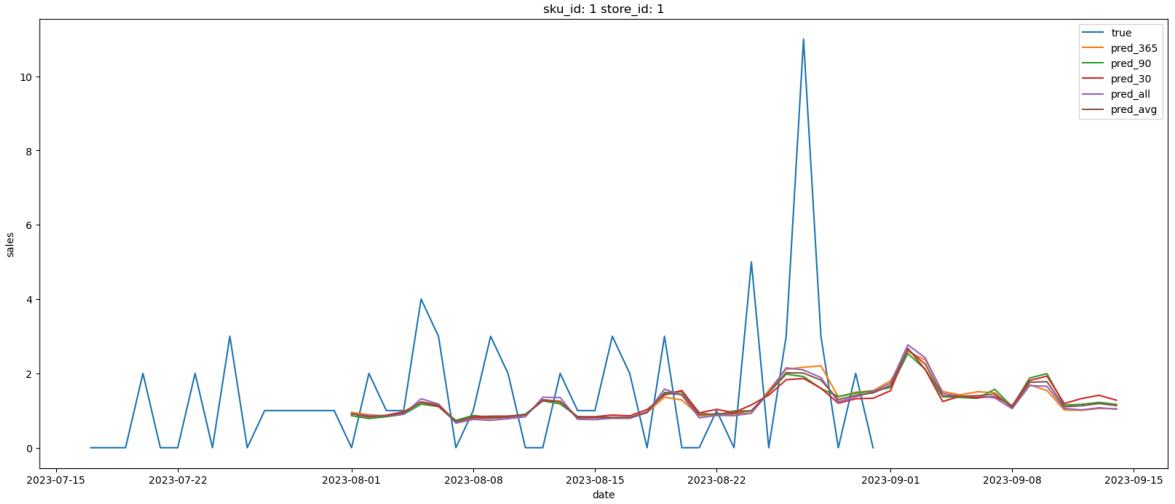
	store_id	sku_id	date	channel	quantity
0	1	1	2023-09-01	1	2.0
1	1	1	2023-09-01	2	1.0
...	...	...	...	...	...

备货策略

$$\tilde{y} = \lceil \alpha \hat{y} + \beta \rceil$$

- 问题：当$\alpha > 1$时，对 $\hat{y}$ 较大的结果，$\tilde{y}$ 会更大，造成冗余库存和利润损耗
- 观察：模型对大销量的商品预测准确度显著高于小销量的商品



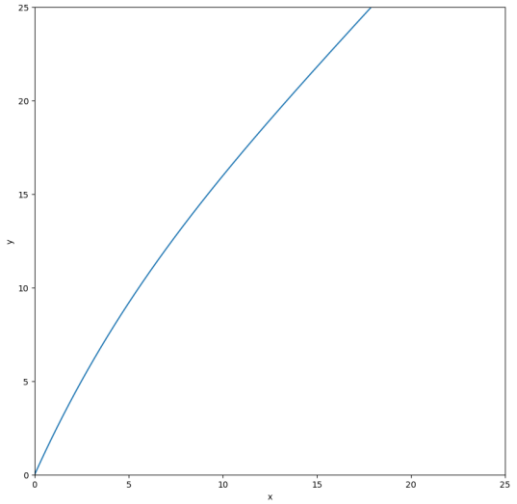


备货策略

$$\tilde{y} = \lceil \alpha\hat{y} + \beta \rceil$$

- 问题：当$\alpha > 1$时，对 $\hat{y}$ 较大的结果，$\tilde{y}$ 会更大，造成冗余库存和利润损耗
- 观察：模型对大销量的商品预测准确度显著高于小销量的商品

$$\tilde{y} = \left\lceil \left(e^{-\alpha\hat{y}} + \beta\right)\hat{y} \right\rceil$$

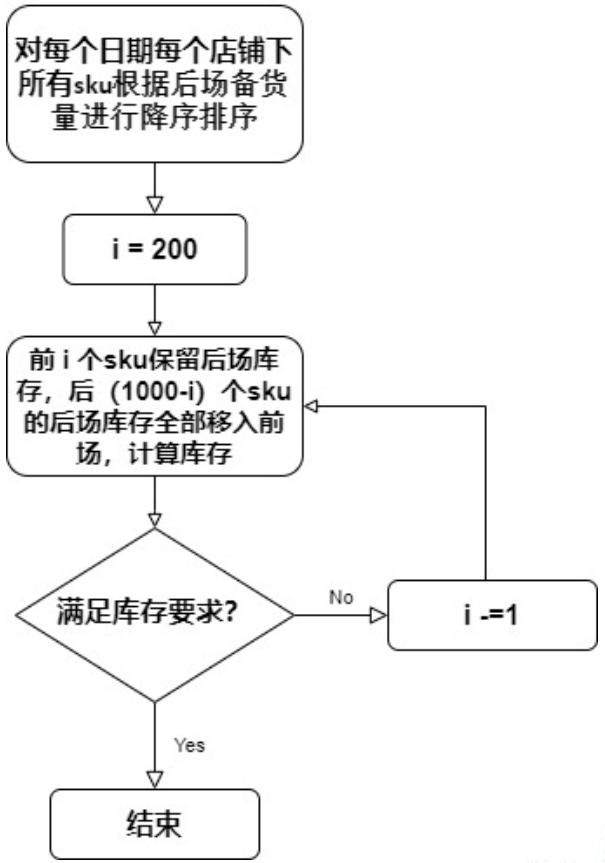


- $\tilde{y}$表示备货量， $\hat{y}$表示预测量，$\alpha(\alpha \geq 0), \beta(\beta \geq 0)$为超参数
- 超参数的确定：启发式网格搜索
- $\alpha = 0.1, \beta = 1.23$

库存分配

依据备货量，根据贪心算法求得满足题目要求的库存分配

$$
\begin{aligned}
&\max S(x_{ijt}^k, x_{ijt}^m) \\
&\text{s.t.} x_{ijt}^k \geq 0 \\
&\quad x_{ijt}^m \geq 0 \\
&\quad x_{ijt}^k + x_{ijt}^m = \tilde{y}_{ijt}^k + \tilde{y}_{ijt}^m \\
&\quad x_{ijt}^k > 0 \text{ if } x_{ijt}^k + x_{ijt}^m > 0 \\
&\quad \frac{\sum_{i=1}^{I} \mathbb{I}(x_{ijt}^m > 0)}{\sum_{i=1}^{I} \mathbb{I}(x_{ijt}^k + x_{ijt}^m > 0)} \leq 0.2 \\
&\quad \frac{\sum_{i=1}^{I} x_{ijt}^m}{\sum_{i=1}^{I} x_{ijt}^k + x_{ijt}^m} \leq 0.4
\end{aligned}
$$



项目总结——应用效果

A 榜　B 榜

我的成绩

到目前为止，您的最好成绩为 **2769717.40000000** 分，第 **3** 名，在本阶段中，您已超越 **55** 支队伍。

排名	排名变化	队伍名称	有效提交次数	最高分提交时间	最高得分	履约效率	平均履约率
1	-	default132643...	51	2023-11-27 22:59	2865198.01000000	0.50000000	0.87000000
2	-	default7586194	13	2023-11-27 17:53	2813182.19000000	0.48000000	0.87000000
3	-	是队伍名不是...	49	2023-11-26 23:57	2769717.40000000	0.48000000	0.86000000
4	↓ 1	default132624...	27	2023-11-27 20:50	2740684.41000000	0.47000000	0.86000000

A 榜　B 榜

我的成绩

到目前为止，您的最好成绩为 **3543262.29000000** 分，第 **2** 名，在本阶段中，您已超越 **20** 支队伍。

排名	排名变化	队伍名称	有效提交次数	最高分提交时间	最高得分	履约效率	平均履约率
1	-	default132643...	1	2023-11-29 00:00	3761645.63000000	0.50000000	0.87000000
2	-	是队伍名不是...	1	2023-11-29 10:20	3543262.29000000	0.48000000	0.86000000
3	-	default132624...	1	2023-11-29 12:28	3476761.92000000	0.45000000	0.86000000

- 算法应用效果良好且稳定。
- 测试执行效率高（生成答案一次仅需2min）。

```
[4]  pred_365 = test('partial_365_final.txt')
     ✓ 22.0s

[5]  pred_all = test('all_iteration_1400.txt')
     ✓ 1m 11.8s

[6]  pred_90 = test('partial_90_final.txt')
     ✓ 5.6s

[7]  pred_30 = test('partial_30_final.txt')
     ✓ 11.6s
```

项目总结——价值潜力

- 使用多阶段方法
  - 销量预测
  - 备货策略（考虑预测准确度进行备货）
  - 库存分配

CCF BDCI | CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST

2023 CCF 大数据与计算智能大赛 11th

感谢垂听，敬请批评指正！

是队伍名不是大名

陈雨荃 刘怡鹏