2023 CCF 大数据与计算智能大赛
《线上线下全场景生鲜超市库存履约一体化决策》

基于 LightGBM 的三段式供应链库存管理模型

是队伍名不是大名

陈雨荃

信息管理与信息系统 大三

南京大学

中国-南京

chenyuquan297@hotmail.com

刘怡鹏

计算机科学与技术 大四

重庆大学

中国-重庆

liuyipeng@cqu.edu.cn

团队简介

团队成员来自南京大学计算机科学与技术系 LANDS 实验室，研究方向包括端到端预测后优化、在线优化等。我们致力于使用机器学习技术解决实际生产中的难题，为产业发展贡献力量。我们追求在解决问题中学习，在学习中成长!

摘要

随着科技和零售的发展，电商和商超线上、线下独立经营的传统被打破，诞生了线上线下深度结合的全场景多业态业务模式。在中国拥有万亿市场规模的生鲜商品，在塑造超市品牌形象、影响消费者心智等方面有着重要作用。不同于可口可乐等标准化商品，生鲜商品货架期短、易损耗、价格变化频繁，同时多业态销售又增加了新的不确定性，对其进行库存决策变得十分困难。

我们设计了“预测、备货、分配”三段式策略来解决该问题。在销量预测中，我们构建 LightGBM 模型，通过使用特征工程、模型集成等方法取得了较好的预测精度。在备货中，我们针对数据特征设计了启发式的备货策略，达成了履约率和损耗成本的平衡。在库存分配中，我们构建运筹学模型，并使用贪心算法进行求解，取得了较好的履约效率。

关键词

数据挖掘，销量预测，供应链管理，库存分配，决策树

1 探索性数据分析

赛题提供了经脱敏和简化处理后的某电商平台生鲜品牌超市多家门店、过去两年线上和线下订单的销售数据、商品数据、库存数据、价格数据、促销数据，门店所在区域的天气数据，以及测评目标商品和日期，以及测评日期对应的商品价格、促销和天气信息。

在本部分，我们主要对影响建模的数据特征进行分析，对于其余无关数据分析不做展示。

如图 1 所示，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。

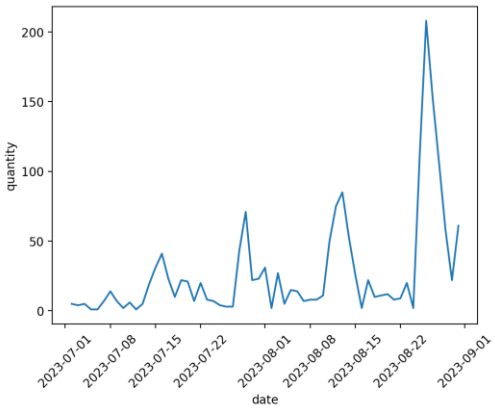

图 1：sku_id=11,store_id=1 的商品销量变化图

如图 2 所示（横轴代表商品开售时间距 2023-09-01 的天数，纵轴表示 sku 数量），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）。

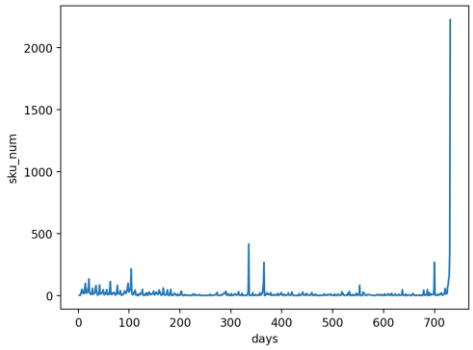

图 2：2023-09-01 距商品开售时间折线图

如图 3 所示，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。

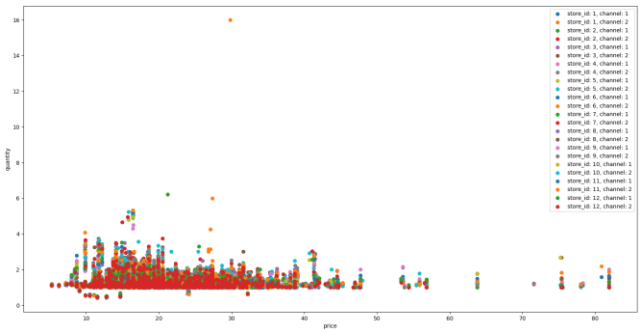

图 3：商品销量随价格变化散点图

2 销量预测

在本部分，我们首先进行特征工程得到有效的数据特征，然后训练回归模型拟合数据，最后使用模型预测销量。

2.1 特征工程

基于上述的探索性数据分析，我们对各种类型的特征进行融合，其中的方法包括数值差分、分类编码、数据展开等。

2.1.1 价格特征

价格对销量的影响很大，构造销量特征至关重要。首先我们需要为每个商店下的每个 sku 填充从开始日期到最大日期之间的所有日期，从而方便计算后续的价格变化率等时序特征。对于填充后缺失的值（价格、状态等）直接向前填充即可。

填充结束后，我们计算以下统计量并作为特征：

- 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
- 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

除此之外，我们还进行了分类编码：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。我们基于数据集中的列名定义设计了以下键：['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']。

2.1.2 促销特征

促销对销量的影响也非常大。对于每个商店中单个sku在某天不同渠道来说，促销不是单一的。从实际业务来看，商品参与多种促销活动是很正常的现象。因此，如何处理促销以使其对齐，从而能方便训练是非常重要的。

我们通过探索性数据分析发现，在相同(store_id, sku_id, date, channel)下，促销最多有3个，因此我们将单个个体所有促销进行展开，展开为3大列，每一大列包括promotion_id、curr_day、total_days、promotion_type、threshold、discount_off六小列。对于促销小于3个的，展开后会出现nan值，我们直接对其填充0（因为0并未对应任何促销）。

2.1.3 销量特征

商品之前的销量对当天的销量会有一定影响。我们将所有的订单使用(store_id, sku_id, date, channel)对quantity求和来统计销量。

因为递归预测容易出现误差累积的情况，而且预测相对耗时，所有我们不使用递归预测。由于预测周期是两周，我们可以对每一天的数据向前推两周，然后再计算前x(x=1,3,5,7,14,21,30,60,90)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。

2.1.4 日期特征

部分商品的销量呈现周期性趋势，通过日期提取时间特征至关重要。我们使用日期计算出了以下特征，从而能捕获时序特征：

- 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

2.2 模型训练及预测

考虑到数据特征的复杂性，我们使用LightGBM[1]进行对数回归，参数设置参考M5竞赛Top1方案。实际上，我们尝试了使用MQRNN等深度学习模型进行建模，但最终效果不是很好，我们分析主要是因为深度学习模型对数据的利用效率较低，导致没有很好地拟合数据。

在探索性数据分析时（如图2所示），我们发现每个商品的开售时间有所不同，如果直接将数据一股脑喂入模型训练可能会导致模型无法很好的拟合（因为这会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）。因此，我们希望使用不同时间段的数据来训练多个模型然后进行集成，这样就能综合捕获长短期特征，使得模型的精确度得到提升。

具体来说，我们将2023-09-01前30、90、365、所有天共四份数据分别使用同一模型结构训练出四个不同的模型。集成时直接对模型输出取均值。实验及线上评测表明，这样的方式可以有效提升预测精度。

3 备货

预测模型总是做不到100%准确，因此我们需要考虑到误差的影响，对商品进行备货。备货策略有很多，我们尝试了如下库存模型：

$$\tilde{y} = \lceil \alpha\hat{y} + \beta \rceil \tag{1}$$

其中，$\tilde{y}$表示备货量，$\hat{y}$表示预测量，$\alpha(\alpha \geq 1)$为魔法系数，$\beta(\beta \geq 0)$为安全库存。

我们可以通过网格搜索得出最优的α和β。通过实验及线上评测发现，当$\alpha = 1.51, \beta = 0$时，可以比预测即备货的方式（即$\alpha = 1, \beta = 0$）的结果好，即利润更高。

分析可以发现，这样的方式对于所有商品都是用同样的魔法系数和安全库存，那么当$\alpha > 1$时，如果预测结果很大，那么备货量就会相对更大，而实际上根本不需要准备如此多的额外库存，这么多的额外库存只会导致损耗增多。同时，我们发现模型对大销量的商品预测准确度显著高于小销量的商品。从业务角度来看，这也是很正常的，因为对大销量的商品，需求相对稳定，预测准确率自然会相对较高。

综合以上两点，我们认为，小销量的商品应该有更多的额外库存，而大销量的商品则更少。因此，我们设计出如下模型：

$$\tilde{y} = \left\lceil \left(e^{-\alpha\hat{y}} + \beta\right)\hat{y} \right\rceil \tag{2}$$

其中，$\tilde{y}$表示备货量，$\hat{y}$表示预测量，$\alpha(\alpha \geq 0), \beta(\beta \geq 0)$为超参数。这样就能使得结果满足上述要求。我们可以通过

验证集来对参数使用梯度下降法进行最优化搜索，最优化目标为利润：

$$\begin{aligned} &\max S(\tilde{y}) \\ &s.t. \alpha \geq 0 \\ &\quad \beta \geq 0 \end{aligned} \tag{3}$$

其中S表示使用备货量计算利润的整个过程（由于过程非常复杂，可以考虑使用代理函数进行代替）。

但上述方法相对难实现且比较耗时，且不一定能收敛到最优解，在赛题时间紧张情况下，我们采用了启发式网格搜索的方式来确定参数，当$\alpha = 0.1, \beta = 1.23$的时候，利润可以达到一个不错的结果。

4 库存分配

前后场备货量确定后，我们需要调整备货方案以使其满足条件，并得到较好的利润及履约效率。建立模型如下：

$$\begin{aligned} &\max S\left(x_{ijt}^{k}, x_{ijt}^{m}\right) \\ &s.t. x_{ijt}^{k} \geq 0, x_{ijt}^{m} \geq 0 \\ &\quad x_{ijt}^{k} + x_{ijt}^{m} = \tilde{y}_{ijt}^{k} + \tilde{y}_{ijt}^{m} \\ &\quad x_{ijt}^{k} > 0 \text{ if } x_{ijt}^{k} + x_{ijt}^{m} > 0 \\ &\quad \frac{\sum_{i=1}^{N} I\left(x_{ijt}^{m} > 0\right)}{\sum_{i=1}^{N} I\left(x_{ijt}^{k} + x_{ijt}^{m} > 0\right)} \leq 0.2 \\ &\quad \frac{\sum_{i=1}^{N} x_{ijt}^{m}}{\sum_{i=1}^{N} x_{ijt}^{k} + x_{ijt}^{m}} \leq 0.4 \end{aligned} \tag{4}$$

其中所有的符号定义来自上文及题目，在此不再赘述。

同样的，由于模型的利润确实较难去计算，要使用运筹模型进行优化时间复杂度过高，恐超过时间限制。因此，我们采用贪心方法得到一个次优解。

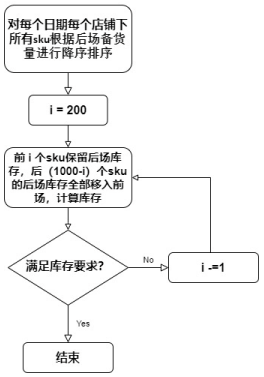

图 4：贪心算法求解库存分配问题流程图

如图 4 所示，我们直接对每个日期每个店铺下所有 sku 根据后场备货量进行降序排序，取前 200 个 sku 保留后场库存，而后 800 个 sku 将后场库存全部放到前场中。如果前 200 个 sku 的后场库存之和没有满足条件，那么就依次取前 199、198、...、0 个 sku 作为保留后场库存的 sku，直至满足条件。

使用这样的算法一定程度上保证了解的有效性，同时计算十分迅速，可以轻松应对大数据量。实验和线上评测表明，这样的方法能取得较好的履约效率。

5 项目总结

在本项目中我们主要是针对数据特征基于 LightGBM 设计了“预测、备货、分配”的一体化决策方案，取得了优秀的业务指标（利润及履约效率）。除提交的解决方案外，我们还尝试了时序预测（如 Prophet[2]）、深度学习（MQRNN[3]）等方法，但结果并不理想，我们认为可能是这些方法在较小数据量下存在泛化能力差、容易过拟合等问题。

当前解决与赛题类似的供应链问题正越来越重要，其中端到端的解决方案表现出良好的结果。[4]中提出使用强化学习的方法构建库存模型，[5]中提出使用深度学习解决库存管理问题，他们的模型在京东的供应链场景下都有不错的表现。迫于时间限制及我们对该领域的了解还尚浅，在实现中遇到了很多问题，期待能在后续研究中逐步攻克难题，设计模型做出更科学的库存履约一体化决策。

致谢

感谢竞赛主办单位、承办单位、合作单位和竞赛平台提供的机会和支持。感谢京东提供的赛题和数据集。感谢工作人员耐心地解答问题。感谢团队所有成员两个月来的合作和付出。最后希望 CCF BDCI 能越办越好，后续有机会继续参加比赛。

参考

[1] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).
[2] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." The American Statistician 72.1 (2018): 37-45.
[3] Wen, Ruofeng, et al. "A multi-horizon quantile recurrent forecaster." arXiv preprint arXiv:1711.11053 (2017).
[4] Liu, Xiaotian, et al. "Deep Reinforcement Learning for Large-Scale Inventory Management." Available at SSRN 4490327 (2023).
[5] Qi, Meng, et al. "A practical end-to-end inventory management model with deep learning." Management Science 69.2 (2023): 759-773.