蔬菜定价与补货的预测与规划模型

吴韬略, 董宇坤, 刘兆宸

2023 年 10 月 5 日

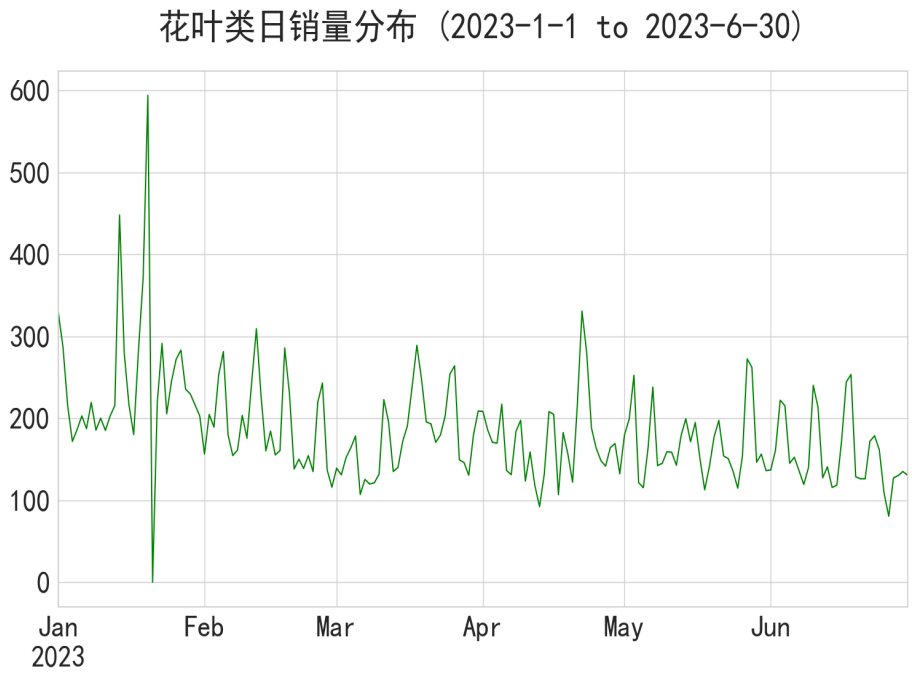
问题一

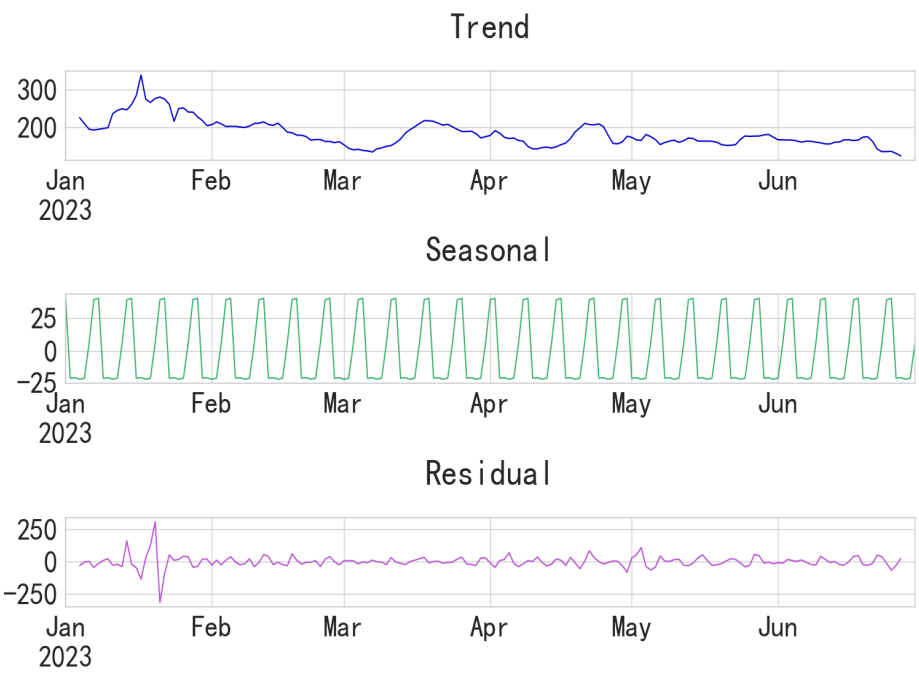
- 描述性统计与可视化；
- 灰色关联分析；
- 基于 FP-tree 的关联规则挖掘。

第一题的分析对后续的问题求解提供导向。

描述性统计与可视化

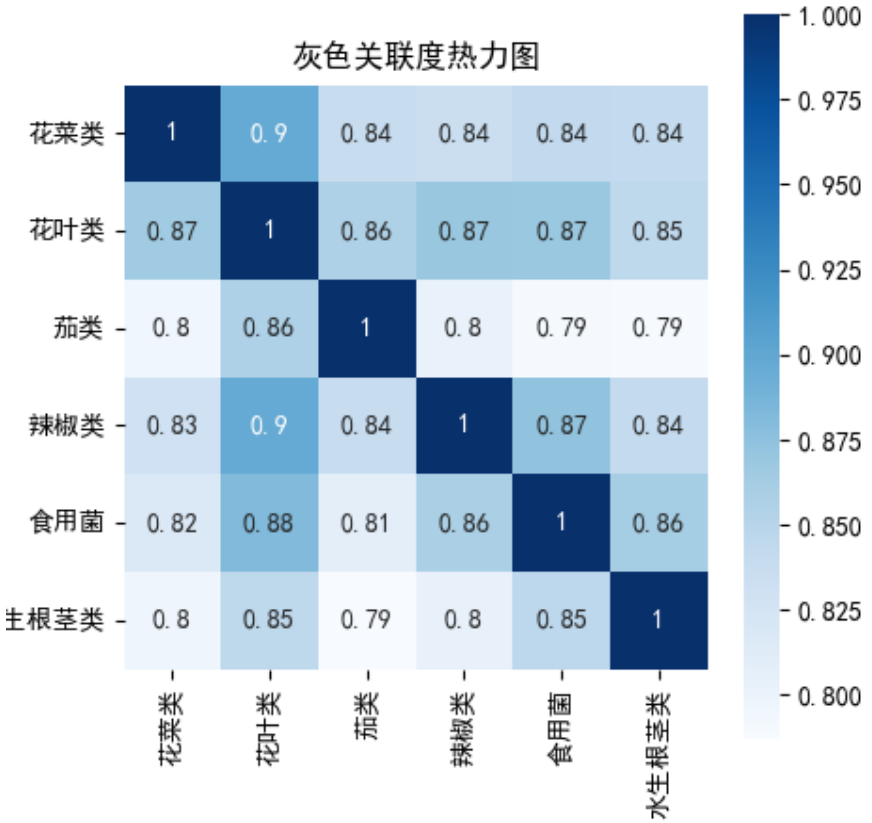
- 正偏度、极差高 etc.;
- 打折对于不同品类的倾向性;
- 各品类销量均以 7 天为周期;
- ...





关联性分析

本文主要使用灰色关联分析来分析品类之间的相关性。基于单品之间以及品类之间的关联性分析结果，会帮助在第三问时提升模型性能。



注：我们先使用 spearman 相关系数来初步探究，后面主要使用灰色关联分析，是因为 spearman 相关系数计算的时候只考虑了序关系，而灰色关联分析则在进行了标准化和归一化之后，考虑值大小的差距衡量曲线的相似度，更值得尝试。

关联规则挖掘

选用 FP 树（Frequent Pattern Tree）来进行数据挖掘。FP 树算法相对于传统的 A priori 算法，避免了生成候选项集的过程，减少了计算开销，并且树状的结构使得存储数据占用空间较小。
得到的具有强关联的组合如下：

表：关联规则较强的组别

先例组合	伴随组合	组合概率
(红灯笼椒 (1)，芜湖青椒 (1)，螺丝椒)	(红椒 (1))	0.91
(大白菜，枝江红菜苔，牛首油菜，西峡香菇 (1))	(泡泡椒 (精品))	0.91
(洪湖莲藕 (粉藕)，紫茄子 (2)，西兰花，金针菇 (盒))	(芜湖青椒 (1))	0.90
(净藕 (1)，小白菜，高瓜 (1))	(西兰花)	0.85

注：模型假设商超一分钟内的所有订单来自于同一个顾客。

问题二

- 分析得到目标日期附近各蔬菜品类的销售总量与成本加成定价的关系；
- 基于蔬菜类商品的历史批发价格对未来 7 天的批发价格进行预测；
- 建立优化模型，制定日间和日内的最佳定价策略，给出 7 日内最合理的补货量分布。

拟合销售总量与成本加成定价的关系

考虑到数据的时间特征不可忽略，我们将时间进行编码，结合该天的成本加成定价和打折率作为输入，以该天的销售总量作为标签建立 XGBoost 机器学习模型，得到销售总量与成本加成定价的映射关系。

为什么不选用别的模型？

我们试过决策树的做法，但是模型效果不如 XGBoost。而随机森林的性能与 XGBoost 差不多。

特征的表示与编码

时序信息的独热编码和正弦余弦编码

时序信息的特征表示为 (y,w,d,mcos,msin)。我们将日期拆解为年、月、日和星期。y 特征为该天年份减去 2022；w 特征为星期数；d 特征为该天的日除以 30。对于表征月份特征，考虑到月份在年的交界处有间断，且以月为单位更能体现问题的季节依赖性，因此这里将月份进行适量放缩后进行正余弦编码，进而得到 mcos,msin 的特征。

打折特征的表示

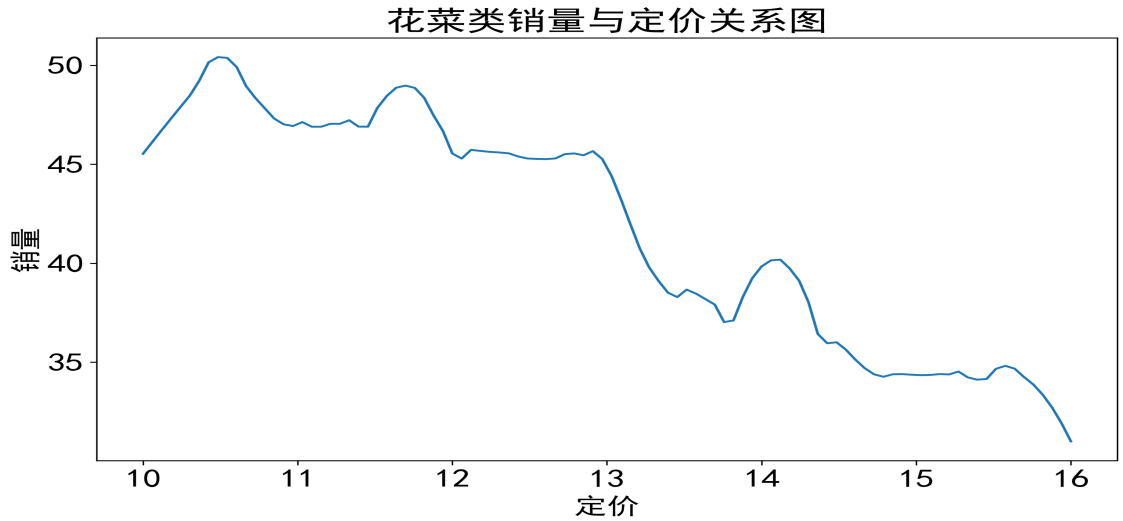
折扣信息的特征表示为 (dp,dr)，其中 dp 代表打折占比，其定义为该天该品类所有打折商品的销售量除以该天该品类的总销售量；dr 代表打折率，是该天该品类所有商品打折程度关于销售量的加权平均。

模型性能

我们训练的模型参数如下：

树的深度	树的个数	学习率	MAE
6	450	0.09	15.70
6	450	0.1	15.99

模型的最佳平均绝对误差 (MAE) 为 15.70，平均相对误差为 21.6%。下图展现了 2023-7-1 花菜类预测的销量和定价关系：

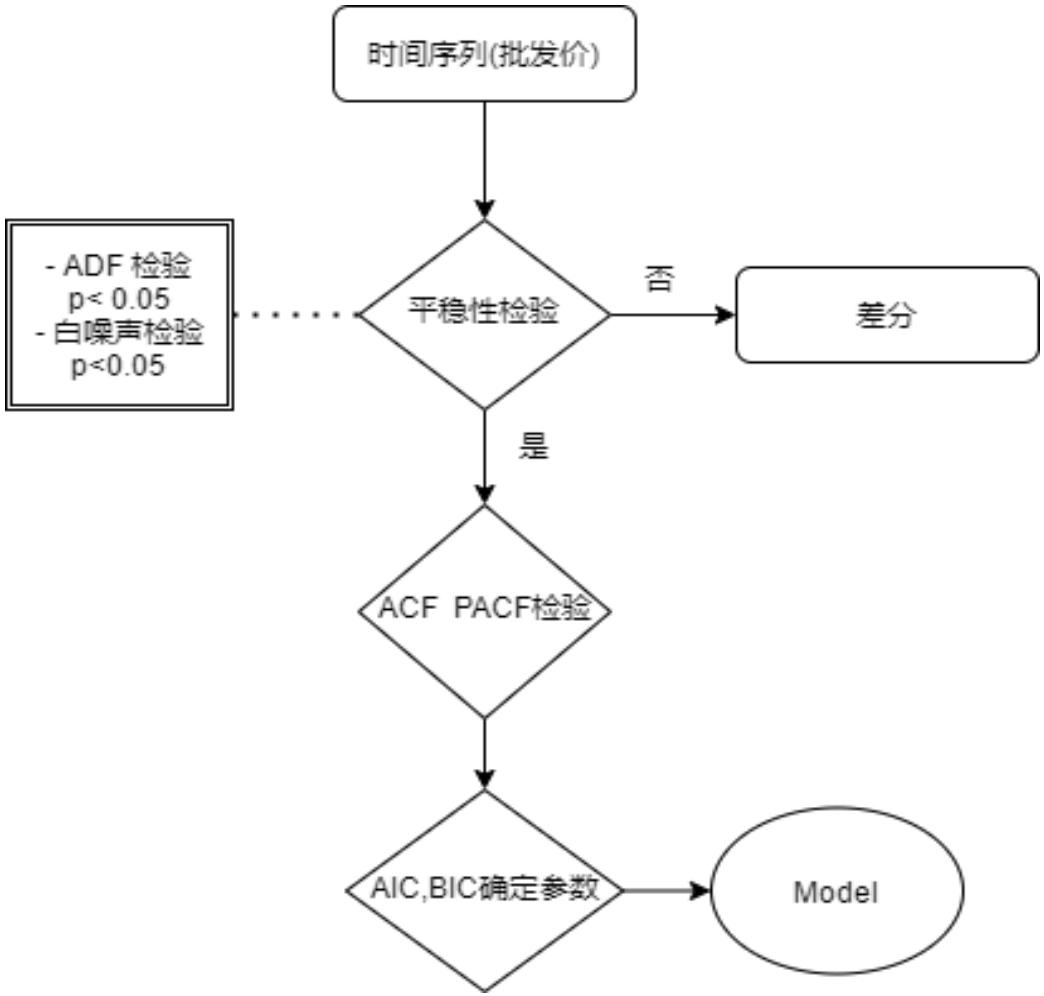


预测未来 7 天各品类的批发价格

这是时间序列的预测问题。
由于问题背景说明了商家在做补货决策的时候是不确切知道具体单品和进货价格的，在此进行预测是有必要的。
用附录 3 的数据关于销售量加权得到各品类的平均批发价格的时序数据，基于 ACF 图和 PACF 图对数据的时序特征的表征，选用合适的时间序列模型对于每个品类的平均批发价格进行预测。
注：根据模型假设，评估每一品类单日成本加成定价时，按该品类下所有单品售价关于销量的加权平均估计。

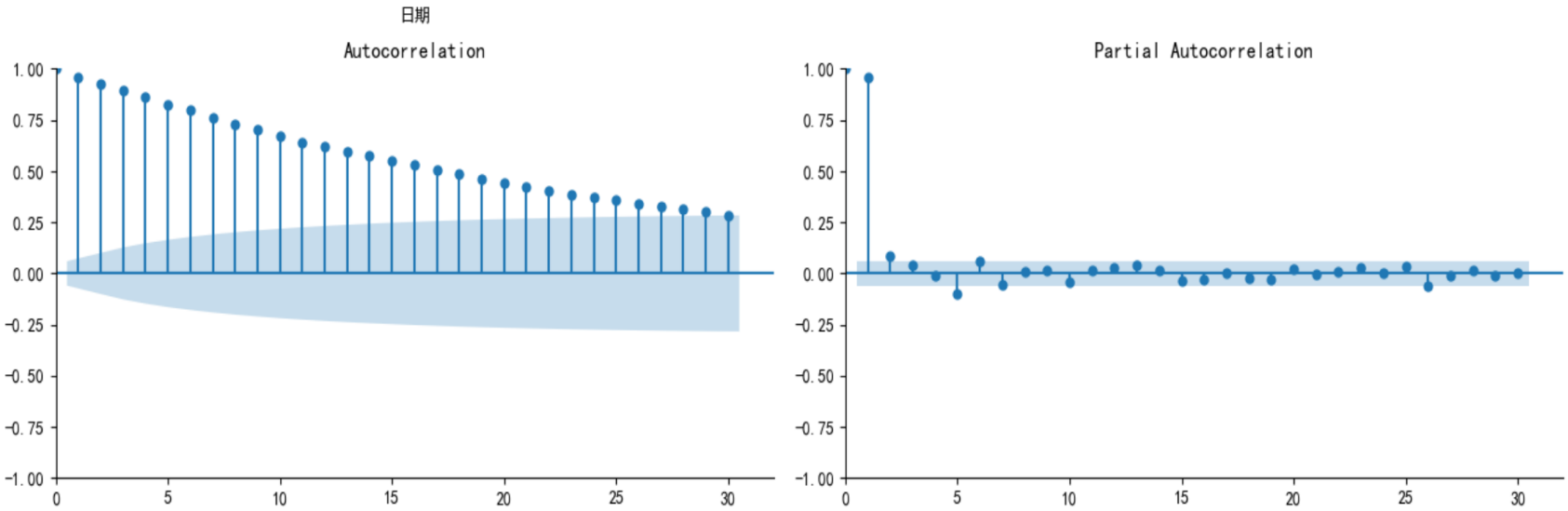
时序分析的步骤与结果

下图展现了研究该时间序列的流程图：



时序分析的步骤

本题中，ADF 检验认为序列是平稳的，白噪声检验认为序列的残差存在自相关性。
花菜类批发价的序列数据的 ACF 图和 PACF 图如下：



观察 ACF 呈拖尾型，PACF 呈断尾型。根据时序模型的参数准则 [1] 和 BIC （贝叶斯信息准则），我们选用自回归（AR）模型，并且选定模型阶数 $p = 2$。

时序分析的步骤

在此基础上，进一步地对所得模型结果残差进行白噪声检验残差序列是否为白噪声序列，若是，残差已经无信息可提取。
经编程求解残差白噪声检验得到 p 值为：0.938>0.05，无法拒绝原假设，接受残差序列是白噪声序列这一假设，这意味着自相关信息提取较为充分。

使用的模型

这里我们选用遗传算法求解非线性规划模型。
在我们的模型中，最棘手的是 XGBoost 的模型响应过于复杂，这不是一个解析的函数关系。而遗传算法则可以处理此类多变量的非简单函数的优化问题，并且在并不明确解空间的情况下，防止陷入局部最优解。

目标函数

考虑如下的目标函数：

$$\mathrm{W}=\sum_{i=1}^{6}\left[\left(\alpha_i\left(1-\varepsilon_i\right)+\alpha_i\beta_i\varepsilon_i\right)\min\left\{s_i,k_i\right\}\left(1+\zeta_i\right)-c_ik_i\right]$$

中括号内的第一项是总收入，乘积三项分别是：

- 该品类的平均销售单价。这里根据模型假设，损坏的商品打折出售，折扣率为 β_i，而损坏率 ε_i 是品类 i 近 7 天依据附件 4 的各单品损耗率关于销量的加权平均。
- 销量。取 min 是考虑到了进货量不够的情况，而 s_i 是通过训练好的 XGBoost 模型关于当天日期、折扣率 β_i 和定价 α_i 的响应。
- 噪声项。模拟销量的涨落以增强模型的稳健性。

减去的一项为进货价 c_i 乘以补货量 k_i。此模型待优化的变量为 $\alpha_i,\beta_i,k_i(i=1,\cdots,6)$，并且使得 W 最大化。

注：假设销量的预测值是当天市场对该单品的需求量。

模型求解

对于此优化问题，基于遗传算法的优化目标如下：

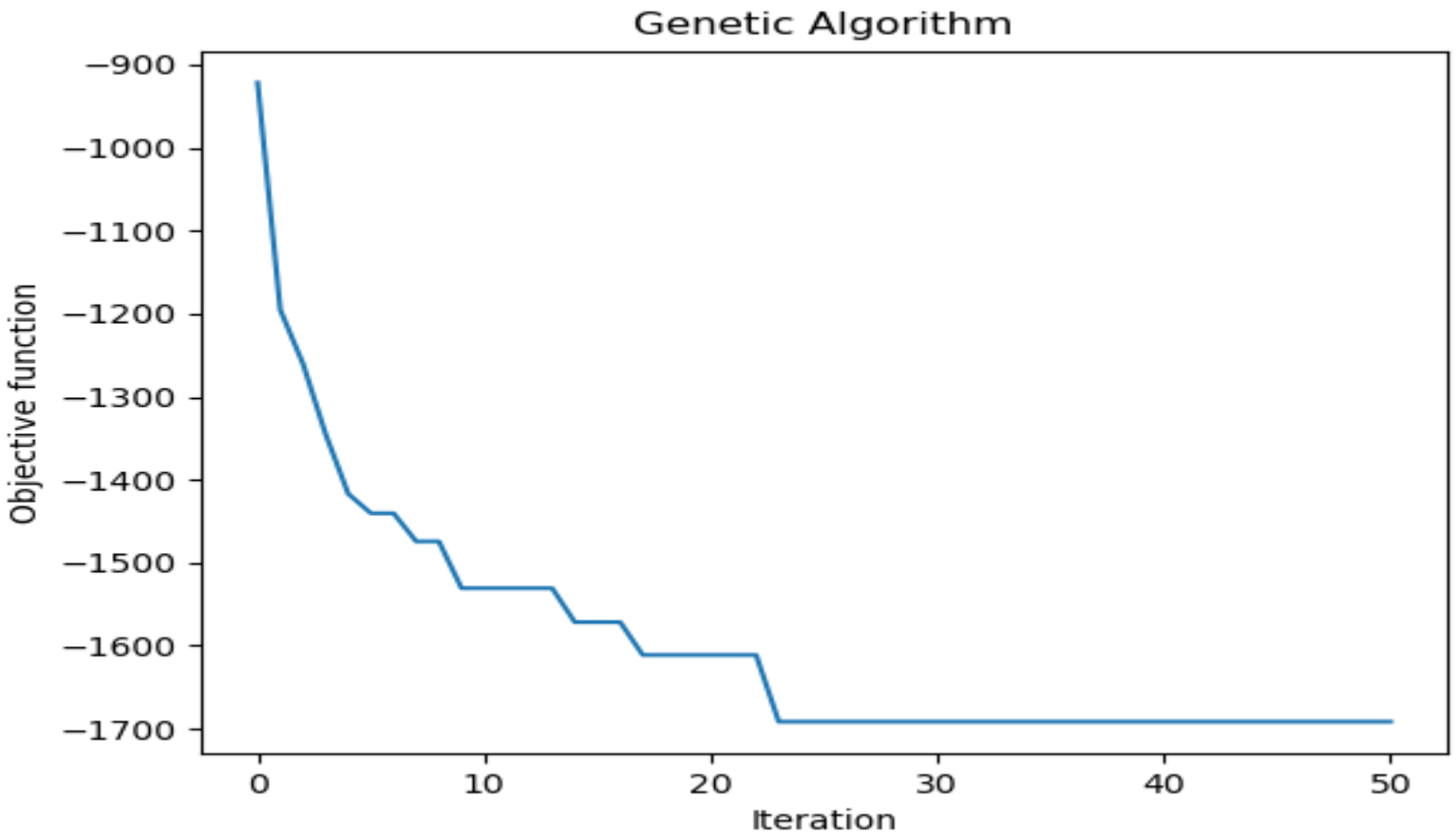
$$\max_{\alpha_i,\beta_i,k_i} \mathrm{W} = \sum_{i=1}^{6} \left[\left(\alpha_i \left(1 - \varepsilon_i \right) + \alpha_i \beta_i \varepsilon_i \right) \min \left\{ s_i, k_i \right\} \left(1 + \zeta_i \right) - c_i k_i \right]$$

$$s.t. \begin{cases} \alpha_i^1 \leq \alpha_i \leq \alpha_i^2 \\ k_i^1 \leq k_i \leq k_i^2 \\ 0.75 \leq \beta_i \leq 1 \end{cases}$$

我们对于待优化的变量进行了一定的范围限制，α_i^1 和 α_i^2 分别是近一周该品类定价范围加权定价最小值的 0.75 倍和最大值的 1.25 倍，k_i^1 和 k_i^2 的定义同理。这样做是考虑到 XGBoost 模型的训练集输入的范围有限，一旦离开训练集输入范围过远，模型可能会失效；此外，考虑到现实因素，优化的解不会离近期的情况过于远，如价格翻一番显然不是很现实。

模型性能

下图展示了模型在 7 月 1 日目标函数的收敛情况：



图中显示，每一轮迭代都有一定的优化，最终找到了最优解。

日内附加策略：时间衰减折扣模型

根据文献 [2]，假设品类 i 的新鲜度 K 的时间指数衰减模型如下（初始新鲜度为 1）：

$$K = \exp(-\lambda t)$$

t 为距离早上 3 点的分钟数。这里认为对于每个品类的衰减率不一样，但是因为新鲜度衰减开始打折的阈值一样，均为默认值 0.6。下面考察近一个月各个品类最早的晚于 16:00 的打折出售时间，拟合出对应的 λ 值，再根据 $0.6 = \exp(-\lambda t)$ 得到每个品类在新鲜度指数衰减模型意义下的平均开始打折时间，这对于制定各品类开始促销的时间点有一定参考价值。

表：部分品类平均开始促销时刻

品类	时间
水生根茎类	19:33
花叶类	18:47
花菜类	20:02

问题三

- 使用相关性聚类编码单品名和所属大类名，并增加其作为特征输入 XGBoost 以提升 R^2 得分;
- 在问题二的基础上增加示性变量为决策变量，并且增加约束;
- 仿照问题二利用遗传算法求解。

相关性聚类编码

相较于第二问数据，我们对单品数据额外增加了品类特征训练，训练的研究对象由 6 种变为了 49 种，因此寻找一种高效的编码手段是有必要的。我们在问题一的灰度分析的基础上引入了相关性聚类编码。品类的编码参考灰色关联的热力图矩阵，相关性大的就近编码。单品的编码按照其所在的品类聚类后，再按照顺序就近编码。按这样的方法编码，相关性大的属性之间距离更小，更容易分类。
最终模型 MAE=4.27，拟合较好。并且引入了相关性聚类编码之后，模型性能有所提升。

确定优化目标和约束条件

引入示性变量 δ_j 表示第 j 个单品是否选入购买。因此我们得到修正的目标函数如下：

$$\mathrm{W} = \sum_{j=1}^{49} \delta_j \left[\left(\alpha_j \left(1 - \varepsilon_j \right) + \alpha_j \beta_j \varepsilon_j \right) s_j \left(1 + \zeta_j \right) - c_j k_j \right]$$

这个目标函数相比问题二的新增了示性变量，并且删掉了取销量、补货量最小值的函数，这一点在约束里面会有体现。在本情形中，有如下四种约束：

- 与问题二类似的现实约束，即优化后的值不能离常规值太远;
- 最小陈列量为 2.5kg;
- 可售单品总数控制在 27-33 个;
- 补货量不小于销量。

确定优化目标和约束条件

综上所述，该优化模型可以表达为下面形式：

$$\max_{\alpha_j,\beta_j,k_j,\delta_j} \mathrm{W} = \sum_{j=1}^{49} \delta_j \left[\left(\alpha_j \left(1-\varepsilon_j \right) + \alpha_j \beta_j \varepsilon_j \right) s_j \left(1+\zeta_j \right) - c_j k_j \right]$$

$$s.t. \begin{cases} x_j^1 \le x_j \le x_j^2, x = \alpha, \beta, k \\ k_i \ge 2.5 \\ 27 \le \sum \delta_j \le 33 \\ s_i \le k_i \end{cases}$$

模型求解同问题二。

问题四

针对问题四，我们从以下的角度来分析与讨论：

- 季节/天气性数据;
- 顾客流量数据;
- 重大意外事件;
- 蔬菜供应链数据;
- 附近商超的销售政策流量数据;
- 节假日数据;
- 消费者数据。

模型优点

- 对于时间序列，我们对其时间平稳性进行了充分检验，并通过 BIC 准则自动调整模型参数；
- 对于机器学习模型的数据，我们根据数据特征进行特定的处理，包括时间数据的分离与正余弦化处理，品种编码按照灰色关联度相近的优先顺序进行排序，提高了模型的性能；
- 建模充分考虑了随机情况的影响对优化目标的影响，即突发情况对销量的可能影响，加入高斯噪声提高了模型的准确性；
- 打折在两个模型中都有参与，充分考虑打折促销的影响。

模型缺点

- 即使使用了 XGBoost 后，预测的销量仍然偏高，可见特征构建以及模型建立仍然欠妥；
- 采用遗传算法求解非线性规划可能没有得到全局最优值；
- 机器学习的模型精度还有提升空间。

参考文献

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and GM Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

汪晓彤, 基于 *DIT* 的连锁超市生鲜果蔬产品库存控制策略研究, 2020.

谢谢！