基于机器学习的蔬菜类商品的自动定价及补货策略模型*

李鸿秀 唐子峰 黄纪 林宁**

南宁学院信息工程学院，南宁 530200

摘 要 为帮助商超获取更大的利润，本文针对蔬菜类商品销售量之间的关系、补货决策和定价决策的问题，使用了 Mann-Whiteny U 检验、Kruskal-Wallis 检验、Spearman 相关系数、XGBOOST 回归模型和随机森林回归模型[1]等一系列方法，并使用决策树-遗传算法、PSO 算法等，建立了决策树回归模型、决策树机器分类模型、规划求解模型、时序序列预测模型等，并利用 Spsspro 软件对问题进行了分析与求解。

关键字 Kruskal-Wallis 检验，Spearman 相关系数，XGBOOST 回归模型，PSO 算法

Machine learning-based Model for Automated Pricing and Replenishment Strategies for Vegetable Products

Li hongxiu Tang Zifeng Huang ji Lin Ning

School of Information Engineering
Nanning University, Nanning 530200 China

Abstract—In order to help supermarkets obtain greater profits, this paper uses a series of methods such as Mann-Whiteny U test, Kruskal-Wallis test, Spearman correlation coefficient, XGBOOST regression model and random forest regression model to solve the problems of the relationship between the sales volume of vegetable commodities, replenishment decision and pricing decision, and uses decision tree-genetic algorithm and PSO algorithm to establish a decision tree regression model, a decision tree machine classification model and a planning solution model, time series prediction model, etc., and the problem was analyzed and solved by using Spsspro software.

Key words—Kruskal-Wallis test; Spearman correlation coefficient; XGBOOST regression model; PSO algorithm

1 引 言

随着生活水平的日益提升，人民对于蔬菜的需求也逐渐提高，而生鲜商超中，蔬菜类商品的保鲜期短且品相易变，商超需要每天进行补货保证蔬菜的。蔬菜的定价一般采用成本加定价方法。市场需求分析对补货和定价决策很重要。

我们根据一些已知的数据分析求解问题一：蔬菜类商品不同品类或不同单品之间可能存在一定的关联关系，分析蔬菜各品类及单品销售量分布关系与相互关系。问题二：商超以品类为单位进行制作补货计划，分析各蔬菜品类的销售总量与成本加成定价的关系，从而制作出各蔬菜品类未来一周使商超收益最大的日补货总量和定价策略。问题三：蔬菜类商品的销售空间有限，可售单品总数控制在 27-33 个，且各单品订购量满足最小陈列量 2.5 千克和尽量满足市场对各品类蔬菜商品需求的前提下，制作出使得商超收益最大的单品的补货计划。问题四：为了更好地制定蔬菜商品的补货和定价决策，商超还需要采集哪些相关数据使得商超获得更多的利润。

对于本文关于蔬菜类商品的自动定价与补货决策的问题，本文做了以下几个假设：

假设数据期间相对稳定，不受突发事件影响。

假设市场竞争波动不大，竞争对手的价格策略和促销活动稳定。

假设数据完整无错误。

2 模型的建立与求解

2.1 问题一的模型汇总与求解

本文分别对蔬菜类商品不同品类和不同单品进行分析。Kruskal-Wallis 检验的原假设是各样本服从的

*基金资助：本文得到广西高等教育本科教学改革工程项目(2023JGB456)资助。
**通讯作者：林宁 bgy_2009@163.com

概率分布具有相同的中位数，原假设被拒绝意味着至少一个样本的概率分布的中位数不同于其他严格样本，从而检验出未识别出的差异发生在哪些个样本之间和差异大小[2][3]。使用 Kruskal-Wallis 检验分析不同品类之间销售量的差异，结果如表 1 所示。

表 1 Kruskal-Wallis 检验分析结果表

分析项	分组变量	样本量	中位数	标准差	统计量	P	Cohen's 值
销售量	1	1049	30.015	31.208	3777.736	0.000***	0.023
	2	1049	171.213	85.229			
	3	1049	33.858	22.775			
	4	1049	18.881	13.163			
	5	1049	72.168	53.475			
	6	1049	56.96	48.863			
	总计	6294	47.323	71.917			

注：***、**、*分别代表 1%、5%、10%的显著水平

由表 1 中我们可以得出结论基于变量销售量，检验结果 P 值小于 0.05，因此统计结果显著，说明不同品类在销售量上存在显著差异。其差异幅度 Cohen's f 值为：0.023，极小程度差异。这意味着本文可以拒绝原假设，认为不同品类的销售量之间存在统计学上的差异。

表 2 定量变量销售量的描述性统计和正态性检验的结果

变量名	样本量	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
销售量	6294	71.726	71.917	2.783	18.746	0.764(0.000***)	0.167(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著水平

Kolmogorov－Smirnov 检验，适用于大样本资料（样本量>5000）。若呈现显著性(P<0.05)，则说明拒绝原假设（数据符合正态分布），该数据不满足正态分布，反之则说明该数据满足正态分布[4]。定量变量销售量的描述性统计和正态性检验的结果如表 2 所示。数据正态性检验的结果如图 1 所示。由图 1 可以发现正态图基本上呈现出钟形，数据虽然不是绝对正态，但基本可接受为正态分布。

对于 Kruskal-Wallis 检验中发现的显著差异，需要进一步进行 Mann-Whitney U 检验事后多重分析。为了纠正多重比较可能导致的问题，本文将使用 Bonferroni 校正。检验结果如表 3 所示。

经过 Mann-Whitney U 检验，我们发现以上品类之间的销售量存在显著差异。通过引用 scipy 统计函数库，求出 Kendall's Tau 系数为:-0.010593253225296164，p 值为: 0.20775313691316966。因为 Kendall's Tau 系数小于 0，所以我们发现品类销售量整体之间相关性有是没有的。

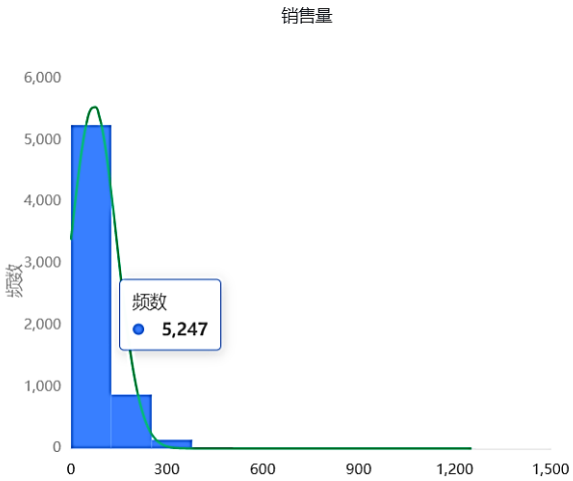

图 1 正态检验直方图

表 3 Mann-Whitney U 检验结果表

水生根茎类	茄类	0
水生根茎类	辣椒类	0
水生根茎类	食用菌	0
花叶类	花菜类	0
花叶类	茄类	0
花叶类	辣椒类	0
花叶类	食用菌	0
花菜类	茄类	6.18E-25
花菜类	辣椒类	0
花菜类	食用菌	0
茄类	辣椒类	0
茄类	食用菌	0
辣椒类	食用菌	1.53E-298

其次本文使用 Spearman 相关性[5]，来估计两个变量之间的相关性，其中变量间的相关性可以用单调函数来描述。如果两个变量取值的两个集合中均不存在相同的两个元素，那么，当其中一个变量可以表示为另一个变量的很好的单调函数时，两个变量之间的 ρ 可以达到+1 或-1[6]。使用 spssspro 分析得到图 2。

图 2 可以明显看出，不同个体之间的相关性是很差的，水生根茎类可能会跟食用菌类比较相关，相关度为 0.6，其他的相关性都是比较低的。

因为单品销售量之间的关系分析，由于单品种类太多，不适合用 Kruskal-Wallis 检验，所以这里只讨论相关性。通过 python 的统计函数 Statistical functions，求出 Kendall's Tau 系数为：-0.06234936890311886，p 值为: 0.0。可以发现品类销售量整体之间相关性有是没有的。使用 spssspro 进行 Spearman 相关性分析得到表 4(由于数据量庞大，这里只作相关系数表，且只截取部分数据)。

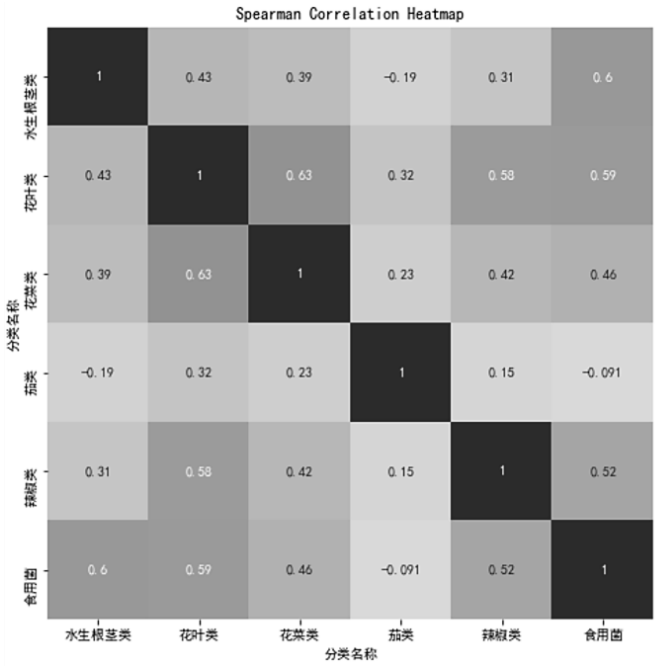

图 2 品类 Spearman 相关系数热力分布

图 2 可以明显看出，不同个体之间的相关性是很差的，水生根茎类可能会跟食用菌类比较相关，相关度为 0.6，其他的相关性都是比较低的。

因为单品销售量之间的关系分析，由于单品种类太多，不适合用 Kruskal-Wallis 检验，所以这里只讨论相关性。通过 python 的统计函数 Statistical functions，求出 Kendall's Tau 系数为：-0.06234936890311886，p 值为：0.0。可以发现品类销售量整体之间相关性有是没有的。使用 spsspro 进行 Spearman 相关性分析得到表 4(由于数据量庞大，这里只作相关系数表，且只截取部分数据)。

由以上分析可见，蔬菜单品销售量之间的相关性是要比蔬菜品类销售量之间的相关性更低的。对于蔬菜品类销售量之间关系，经过 Kruskal-Wallis 检验和 Mann-Whitney U 检验结果显示不同品类在销售量上存在显著差异，其 Kendall's Tau 系数小于 0，也可以发现品类销售量整体之间相关性有是没有的。

Spearman 相关系数分析结果也显示品类销售量之间的相关性很差。这意味着本文可以拒绝原假设，认为不同品类的销售量之间存在统计学上的差异。而对于蔬菜单品销售量之间的关系，经过计算 Kendall's Tau 系数，发现该系数也小于 0，蔬菜单品销售量整体之间相关性有是没有的，由 Spearman 相关系数热力分布图表也可以直观的看出蔬菜单品销售量之间的相关性要比蔬菜品类销售量之间的相关性更差。

2.2 问题二的模型汇总与求解

机器学习可以自动处理大量数据，提高处理效率。因为数据以蔬菜品类为单位，通过使用机器学习技术处理数据，避免了人工处理数据存在的错误。挑选一种品类进行分析，利用批发价格和销售之间的关系构建 XGBOOST 回归模型，从而会得到一个预测模型预测出销量，并以此类推构建出回归模型；还可以构建规划求解模型并结合启发式算法得出该品类每日收益最大情况下对应的销量以及定价，最后依此类推，将每个品类进行代入即可求出达到最高收益对应的日补货量与定价策略。模型求解思路如图 3 所示。

本题选取食用菌作为示例进行分析，以销售日期来整合，求取均值，结果如表 5 所示。

根据以下公式，对数据进一步处理：

$$G = x \times (1 - a) \times (b - c) \quad (1)$$

其中 G 是指商超收益，x 为销量，a 是指损耗率，b 是销售单价，c 是指批发价格。

表 4 部分单品 Spearman 相关系数表

	七彩椒(1)	七彩椒(2)	七彩椒(份)	上海青	上海青(份)	门口小白	丝瓜尖
七彩椒(1)	1	-0.3 131	-0.42 939	0.007 301	0.41 052	0.193 849	-0.08 405
七彩椒(2)	-0.3 131	1	0.245 171	0.062 683	-0.40 145	-0.06 817	0.184 782
七彩椒(份)	-0.42 939	0.245 171	1	0.2 682	-0.27 661	-0.3 153	0.236 983
上海青	0.007 301	0.062 683	0.2 682	1	-0.03 504	0.122887	0.263 701
上海青(份)	0.41 052	-0.40 145	-0.27 661	-0.03 504	1	0.072 986	-0.24 285
门口小白	0.193 849	-0.06 817	-0.3 153	0.122 887	0.072 986	1	-0.0 098
丝瓜尖	-0.08 405	0.18 4782	0.236 983	0.263 701	-0.24 285	-0.0 098	1

表 5 食用菌类以销售日期整合的均值表

销售日期	单品编码	分类编码	销量（千克）	销售单价（元/千克）	批发价格（元/千克）	损耗率（%）
2020/7/1	1.029 000E+14	1.011 011E+09	0.294708	11.866 667	7.314 917	7.832 333
2020/7/2	1.029 000E+14	1.011 011E+09	0.303187	13.63875	7.871937	8.830 063
2020/7/3	1.029 000E+14	1.011 011E+09	0.284846	14.32349	8.120671	9.164 832
…	…	…	…	…	…	…
2023/6/30	1.040 170E+14	1.011 011E+09	0.573 507	13.986 957	9.240 145	6.138 551

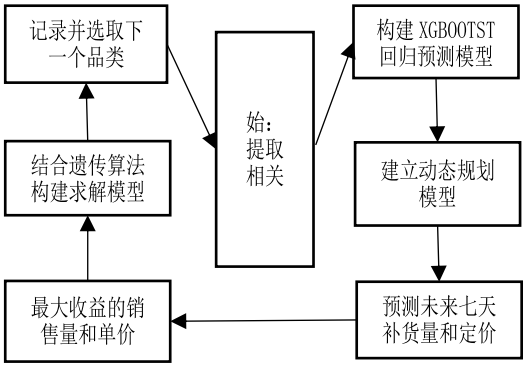

图 3 问题二求解方法

$$z = (b - c) \div c \tag{2}$$

其中 z 是平均每千克差价占比批发价格比例，b 是指销售单价，c 是指批发价格。

调用均值函数 mean 函数求出损耗率平均值作为整体的损耗率，平均损耗率为 8.758815129904535。

首先明确出计算目标，建立 K 个回归树，使得树群的预测值尽量接近真实值（准确率）而且有尽量大的泛化能力（更为本质的东西），从数学角度看这是一个泛函最优化，多目标，观察目标函数：

$$L(\phi) = \sum_i l(\hat{y}_i - y_i) + \sum_k \Omega(f_k) \tag{3}$$

其中 i 表示第 i 个样本，$l(\hat{y}_i - y_i)$表示第 i 个样本的预测误差，误差越小越好。后面$\sum_k \Omega(f_k)$表示树的复杂度的函数，越小复杂度越低，泛化能力越强。表达式为:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

其中，误差/损失函数鼓励此模型尽可能的去拟合训练数据，使得最后的模型会有比较少的 bias。而正则化项则鼓励更加简单的模型。因为当模型简单之后，有限数据拟合出来结果的随机性比较小，不容易过拟合，使得最后模型的预测更加稳定。

运行下述代码（片段），

```
print("\nXGBOOST回归:")
xgb_mo=xgb.XGBRegressor()
xgb_mo.fit(X,Y)
y_pd = xgb_mo.predict(X)
print(" 测试集平均绝对百分比误差:{:.3f}".format(mape(Y,y_pd)))
print("测试集r2_sce",r2_score(Y,y_pd))
```

得出 XGBOOST 回归结果:

测试集平均绝对百分比误差:16.972

拟合程度：0.9390063934411754

此时会得到一个预测模型，这个模型可以根据销售单价和批发价格预测出当日的销量，本次选取了 2023 年 7 月 1 日的销售单价和批发价格作为测试，预测得销量为 36.88491，真实值为 35.365。

接下来还需要构建一个时序序列预测，未来七天的批发价格是未知的，我此时需要根据第一天的批发价格来预测第二天的批发价格，根据第二天的预测第三天的，以此类推，构建一个回归模型。

通过随机森林回归来拟合，拟合优度为 0.8518795038770092。

从图 4 可观察出，拟合效果较好。其次，把全部数据投进去训练，并预测出未来七天的批发价如表 6 所示。

为了更进一步完善吗，还需要构建一个规划求解模型，利用遗传算法来求解。

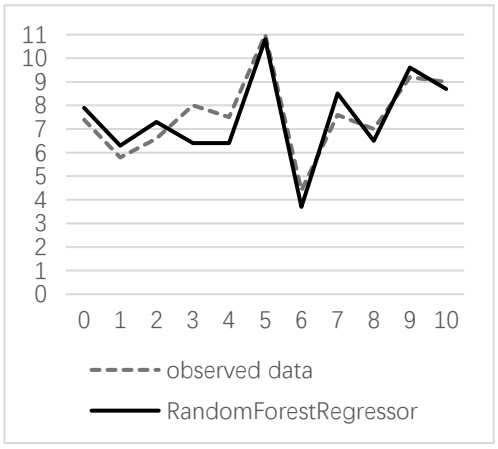

图 4 随机森林回归拟合图

表 6 食用菌未来七天批发价表

日期	批发价
7 月 1 日	8.411 665 843 100 95
7 月 2 日	8.960 400 386 043 59
7 月 3 日	9.215 648 220 289 57
7 月 4 日	8.971 439 743 052 33
7 月 5 日	7.876 682 258 094 11
7 月 6 日	7.424 529 265 669 04
7 月 7 日	7.043 501 862 916 41

其中 c 为批发价格，a 为损耗率，x 代表销售量，b 为销售单价，reshape(1,-1)代表将二维数组重整为一个一行的二维数组：

$$c：8.411665843100959$$

$$a：8.758815129904537$$

$$\mathrm{x} = \text{XGBOOST 回归预测模型}(\mathrm{b},\mathrm{c}).\mathrm{reshape}(1,-1) \tag{5}$$

目标函数：

$$\max \quad x \times (1 - a) \times (b - c) \tag{6}$$

约束：

$$b > c \tag{7}$$

基于以上的预测做迭代，以实现动态规划。

表 7 是由 XGBOOST 回归模型、遗传算法等方法对食用菌未来七天做出的补货总量和定价策略。基于该模型，我们即可对辣椒类、水生根茎类、花叶类、花菜类和茄类的未来七天的补货量及定价策略做出预测。

表 7 食用菌日补货量和定价

未来一周	销售单价	补货量	损耗率%	销售数量	商超收益	品类
第一天	11.9 272	419.401	8.75 882	120.86	387.677	食用菌
第二天	12.4231	193.8	8.75 882	99.3186	313.79	食用菌
第三天	16.7 894	194.47	8.75 882	85.2499	589.108	食用菌
第四天	11.9 253	227.19	8.75 882	114.467	308.5	食用菌
第五天	11.9 286	158.131	8.75 882	142.831	528.057	食用菌
第六天	9.93 921	340.182	8.75 882	211.655	485.626	食用菌
第七天	11.2 878	480.026	8.75 882	383.227	1484.05	食用菌
商超 7 天总收益					4096.81	

2.3 问题三的模型汇总与求解

问题三的解决方法和问题二的类似，都是依据已出的销售数据建立机器学习模型来预测后面的销售数据。基于问题二的回归预测模型，分析关系建立出目标规划模型。由于销售单品的数量要求，从而需要对单品销售数据进行去重，然后排序后选择出前 33 个单品，其次通过均值函数 mean()计算以及模型测试，再通过决策树回归模型算出结果。本题也需要再构建一个时序序列预测模型以及随机森林回归拟合得出批发价格，最后编写规划求解模型并选择 PSO 算法求解即可得出。

以单品上海青作为举例，其他单品按照下述模型进行计算。对上海青这个单品数据进行提取计算，得到上海青这一单品的历史销量数据，包括商超收益以及平均每千克差价占比批发价格比例：

$$G = x \times (1 - a) \times (b - c) \tag{8}$$

$$z = (b - c) \div c \tag{9}$$

G 表示商超收益，x 表示销量，a 表示损耗率，b 表示销售单价，c 表示批发价格，z 代表平均每千克差价占比批发价格比例。

通过均值函数 mean()求得上海青这一单品的平均损耗率为：9.251765602741036。

根据模型测试，在问题三使用决策树回归模型得出的结果是要优于 XGBOOST 回归模型的。使用 XGBOOST 回归模型得到的结果为：数据集平均绝对百分比误差为 35.315%，拟合程度 0.9265608411890007。由于百分比误差较大，所以本题将使用决策树回归模型来进行求解。

我们运行下述决策树回归模型代码（片段），

```
print("\n决策树回归:")
tree=DecisionTreeRegressor(max_depth=50,random_state=0)
tree.fit(X,Y)
y_pred = tree.predict(X)
print(" 测 试 集 平 均 绝 对 百 分 比 误 差:{:.3f}".format(mape(Y,y_pred)))
print("测试集r2_score",r2_score(Y,y_pred))
```

得到的结果为：

数据集平均绝对百分比误差为 0.093%，拟合程度为 0.999996194594405。决策树回归模型可以根据销售单价和批发价格预测当日销量，预测结果为：4.953 千克。

同样的，本题也需要构建一个时序序列预测模型（表 8），用来预测批发价格。

表 8 时序序列表

	Y	shiftX_0
0	5.171 429	5.413 333
1	7.801 667	5.171 429
2	4.736 000	7.801 667
3	6.171 538	4.736 000
…	…	…
1071	2.982 000	4.791 250

通过随机森林回归来拟合，拟合优度为 0.8272854245276333。拟合效果见图 5 所示。最后预测得 2023 年 7 月 1 日的批发价格为 2.939775757575757。

调用 max 函数求出平均每千克差价占比批发价格比例为：2.2653061224489797。

到这里还需要编写一个规划求解模型，本题选择使用 PSO 算法进行求解。

其中 c 为批发价格，a 为损耗率，x 代表销售量，b 为销售单价，reshape(1,-1)代表二维数组重整为一个一行的二维数组。

$$c = 2.93977576$$

$$a = 9.251765602741026$$

$$x = 决策树回归预测([b, c]).reshape(1, -1) \tag{10}$$

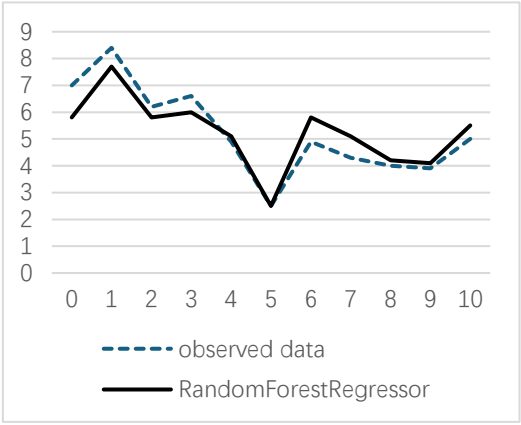

图 5 随机森林回归拟合效果图

目标函数为:

$$\max \quad x \times (1-a) \times (b-c) \tag{11}$$

约束条件为:

$$b > c \tag{12}$$

$$x > 2.5 \tag{13}$$

PSO 预测结果如下:

最优的销售单为:7.402460306319234，最优的销售数量为: 36.447，最大的商超收益为:147.60333149247444。

补货量等于销售量加上损耗，则日补货量为:40.16276461025243984。

图 6 是求优迭代图，可以从此图看出，使用 PSO 算法在迭代 300 次左右结果就趋于平稳了。

基于上述模型，对蔬菜单品进行随机森林回归拟合，再用 PSO 算法进行迭代，即可求解出其余 32 个蔬菜单品日补货量和定价策略，表 9 是部分对蔬菜单品的预测。

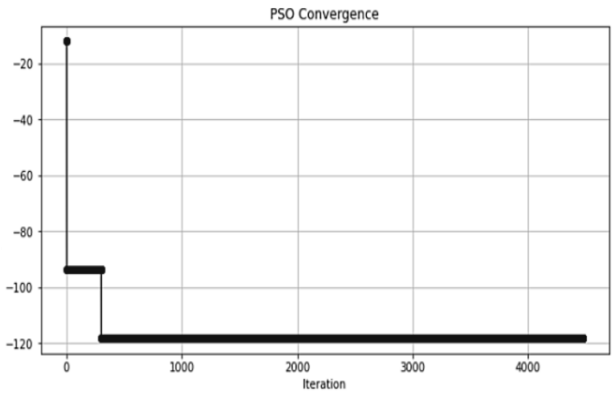

图 6 求优迭代图

2.4 问题四的意见与理由

季节性因素。收集各种蔬菜的在不同季节的需求量、销量和价格。某些蔬菜在某些季节可能更受欢迎。了解这些季节性变化可以帮助商家调整补货策略。

表 9 部分单品日补货量和定价

单品编号	销售单价	日补货量	损耗率%	销售数量	商超收益
102900011016701	6.72 737	50.315	5.7	47.447	146.737 348
102900011030059	10.0 513	103.787	9.43	94	540.059 007 179
102900005115786	5.05 359	38.2762	13.62	33.063	82.267 940
102900005116714	11.9052	53.9354	9.26	48.941	.206 651
102900005115250	21.9	23.9966	10.8	21.405	120.435280
102900011031100	9.92 787	38.6 441	9.43	35	244.702 003

天气和节假日数据。收集未来的天气预报和节假日数据。天气状况，如暴雨、台风等极端天气，可能会影响消费者的购物行为。节假日则通常伴随特定的消费习惯和促销活动。

供应链和物流数据。了解供应商的配送频率、时间和可靠性，监测供应链的风险。有助于预测未来的库存状况和调整补货策略，避免供应链中断从而影响销售。

综合考虑这些数据可以帮助商超制定更精确的补货策略和定价策略，提高效益并满足顾客的需求。另外，定期地去监测这些数据是很有必要的，这可以令商超快速适应市场的变化和新趋势。

3 结束语

在解决问题的过程中，本文使用了 Mann-Whiteny U 检验、Kruskal-Wallis 检验、Spearman 相关系数、XGBOOST 回归模型和随机森林回归模型等一系列方法，并使用决策树-遗传算法、PSO 算法等，建立了决策树回归模型、决策树机器分类模型、规划求解模型、时序序列预测模型，从而分析出蔬菜各品类及单品销售量分布关系与相互关系，以及对商超未来最大收益的补货计划做出了预测。可以利用 Spsspro 软件，根据已知的历史销售数据，分别对品类和单品之间进行分析预测。本文得出的部分分析预测结果说明了本文所提出的模型具有可靠性。

如需要提高模型的可靠性，则要综合考虑更多的影响因素，例如，天气和节假日数据、供应链和物流数据、季节性因素等。

参 考 文 献

[1] 李静波,张莹,盖荣丽.基于机器学习的星载短波红外 CO 2 柱浓度估算[J].中国环境科学,2023,43(04):1499-1509.DOI:10.19674/j.cnki.issn1000-6923.20221117.025.

[2] 朱芮,刘布楼,刘艺语,等.基于文本注意力的推荐系统可解释性研究[J].信息安全学报,2021,6(05):128-143.DOI:10.19363/J.cnki.cn10-1380/tn.2021.09.10.

[3] 白娟. 基于行业差异和卷积神经网络的上市公司财务预警模型研究及应用[D].桂林电子科技大学,2023.DOI:10.27049/d.cnki.ggldc.2022.000018.

[4] 郑晓将. SMT 锡膏检测阈值参数优化研究[D].重庆大学,2022.DOI:10.27670/d.cnki.gcqdu.2021.001004.

[5] 朱慧明,董丹,郭鹏.基于 Copula 函数的国际原油价格与股票市场收益的相关性研究[J].财经理论与实践,2016,37(02):32-37.DOI:10.16339/j.cnki.hdxbcjb.2016.02.006.

[6] 孙昀灿. 绿色公共建筑热舒适与空调供暖末端运行策略研究[D].天津大学,2020.DOI:10.27356/d.cnki.gtjdu.2018.000829.