

队伍编号	205506
题号	(D)

基于 ARIMA-SVM 和 Holt-Winters 的新零售精准销量模型

摘要

本文针对新零售精准预测销量问题，运用了随机森林算法、ARIMA-SVM 和 Holt-Winters 等理论或方法，构建了销量影响因素权重评估模型和销量精准预测模型，综合运用 MATLAB、SPSS、VSCode、RStudio、Jupyter notebook 等软件变成求解，得出了各销量影响因素的特征权值以及精准预测了铲平的销量。

针对问题一，我们首先进行数据预处理，使用 Hermit 插值法对缺失信息进行补全，将商品信息缺失的 skc 剔除，对有效数据进行累加选出占销总量前 50 的 skc。我们选取了折扣率 D、实际价格 rc、库存量 St、标签价 T、价格波动率 F、市场占有率 M 六个指标作为影响销量 (s) 的因素，采用单独考虑四个节假日和整体综合考虑四个节假日两种思想，使用 Pearson 相关系数计算各指标间的关联性，发现实际价格 rc、与标签价 T 相关性最大；利用随机森林算法计算各指标对销量影响的权重，综合考虑发现折扣率和市场占有率权重最大，标签价权重最小。

针对问题二，我们通过历史数据筛选出销量排名前 10 的小类作为研究对象，考虑目标小类下 skc 数目过多，我们根据总销量占比及选取占该小类比重较大的部分 skc 作为小类 skc 信息；我们首先以月为单位，使用灰色预测算法根据 19 年前九个月的月销量数据预测出 10 个小类的 3 个月销量，鉴于历史数据较少，预测后验差比值均高于 0.5，效果不理想；之后以天为单位，使用基于 LSTM 的神经网络算法，学习前九个月的售量数据，学习精确度达到 95%，然后再预测未来 90 天的日销量并计算出对应 MAPE 值，预测误差稳定于 2%。我们发现未来三个月的销售变化曲线在节假日时间内具有明显上升趋势，同时我们发现假期冲击对小类销量的影响具有周期性。

针对问题三，我们以问题二求出的 10 个目标小类为基础，根据 skc 售量占比，将占小类销量超过 82% 的多个 skc 作为研究对象。首先以天为单位，通过 ARIMA+SVM 模型，结合销售数据的线性和非线性特征，对历史约 270 天的 skc 进行学习，预测出对应 skc 未来 12 周的销量，准确率稳定于 82%，误差率达到 18%，虽然结果具有周期波动性，但其预测结果不能准确体现节日冲击对某一天的销售量的影响，预测效果不理想；然后以周为单位，我们通过对历史数据的分析可知销量存在的周期波动性，所以选用 Holt-Winters 加法模型预测未来 90 天的销量数据，计算每个 skc 12 个月的 MAPE 值。我们发现 HoltWinter 准确的把每个节日冲击对销售量的影响反映了出来。

针对问题四，我们将我们的模型以及预测结果呈现给了企业，并以 MAPE 值为依据证明我们的可靠性。同时，我们也提出模型的局限性，如没有考虑各种节假日对销量影响的差异等。

关键词：BP 新零售销量预测， 随机森林， R 型聚类分析， 时间序列

目录

一. 问题重述	4
二. 模型假设	4
三. 符号说明	5
四. 问题分析	5
五. 数据预处理	6
六. 问题一模型的建立与求解	7
6.1 分析和预处理数据	7
6.2 确定销量影响因素	7
6.2 算法分析	7
6.2.1 皮尔逊相关系数	7
6.2.2 线性相关性可视化分析	10
6.2.3 随机森林评估模型	11
6.3 结论	12
七. 问题二的建模与求解	14
7.1 数据预处理	14
7.2 算法分析	14
7.2.1 灰色预测模型	14
7.2.2 基于 LSTM 神经网络的销量预测模型	17
7.3 结论	19
八. 问题三的建模和求解	22
8.1 数据预处理	22
8.2 算法分析	22
8.2.1 ARIMA+SVM 模型预测	22
8.2.2 Holt-Winters 模型	25
九. 问题四——给企业的一封信	27

十. 模型的优缺点	30
十一. 总结	30
十. 参考文献	31

一. 问题重述

在如今的需求推动下，新零售企业的生产模式逐步向多品种、小批量迈进，这让商场内零售店铺里的饰品和玩具等种类变得更加琳琅满目，同时也给零售行业的库存管理增加了很大的难度。如何根据层级复杂，品类繁多的历史销售数据，以区域层级，小类层级乃至门店 skc(单款单色) 层级给出精准的需求预测，是当前大多数新零售企业需要重点关注并思考的问题。针对该现状，本次建模要求基于某企业提供给的产品销售数据解决以下问题，从 3 个方向为新零售企业解决“精准需求预测”问题贡献一份力量：

- 问题一：试分析 2018 年国庆节，双十一，双十二和元旦这四个节假日内各种相关因素对目标 skc 的销售量的影响，可考虑产品销售特征，库存信息，节假日折扣等因素。其中，目标 skc 为销售时间处于 2018 年 7 月 1 日至 2018 年 10 月 1 日内且累计销售额排名前 50 的 skc。
- 问题二：试结合上述分析结果，预测给定区域内目标小类在 2019 年 10 月 1 日后 3 个月中每个月的销售量，给出每个月预测值的 MAPE。其中，目标小类为历史销售时间处于 2019 年 6 月 1 日至 2019 年 10 月 1 日内且累计销售额排名前 10 的小类。
- 问题三：为了满足企业更加精准的营销需求，试着建立相关数学模型，在考虑小类预测结果的同时，预测目标小类内所有 skc 在 2019 年 10 月 1 日后 12 周内每周的周销量，并给出每周预测值的 MAPE(可以考虑 skc 销售曲线与小类销售曲线之间的差异)。
- 问题四：请给企业写一份推荐信，向企业推荐你的预测结果和方法，并说明你们的方案的合理性以及后续的优化方向。

二. 模型假设

- 不考虑经济政策变化导致的月度 CPI 和年度 CPI 波动对销量的影响。由于数据的时间较近，所以宏观经济环境的影响相对其他因素的可以忽略不计；
- 产品的信息如所属小类和标价在所研究的时间段内保持不变
- 在研究的范围内节假日对于商品的销售影响不考虑环境的干扰

三. 符号说明

符号	说明	符号	说明
S	销量	D	折扣率
S_t	库存	r_c	实际价格
T	标签价	l_r	价格波动率
M	市场占有率	n	天数
R_{xy}	皮尔逊相关系数	$Gini$	基尼系数
C	后验差比值	$MAPE$	平均绝对百分比误差
$x(n)$	灰色预测中原始数据列	$y(n)$	灰色预测中的修正数据列
$z(n)$	灰色预测中的预测数据列	p	ARIMA 模型序列本身的滞后数
d	ARIMA 模型的差分化的阶数	q	ARIMA 模型预测误差的滞后数
Y_t	SVM 模型中时间序列	L_t	SVM 模型时间序列中的线性部分
N_t	SVM 模型时间序列中的非线性残差	$K(x_i, x_j)$	SVM 模型的多项式核函数
α, β, γ	Holt-Winter 模型中平滑参数	l_t	Holt-Winter 模型中水平平滑方程
b_t	Holt-Winter 模型中趋势平滑方程	s_t	Holt-Winter 模型中季节平滑方程

四. 问题分析

问题一的分析：在问题一中，题目要求基于所给的 skc 的销售数据，以在 2018 年 7 月 1 日至 2018 年 10 月 1 日内且累计销售额排名前 50 的 skc 为研究对象，探究在同年国庆、双十一、双十二、元旦四个节假日中各种指标对销量的影响。该问属于相关性分析问题。我们首先对数据进行预处理，包括把信息残缺较多的 skc 项剔除、以及用分段三次埃尔米特 (Hermite) 插值法将少量信息残缺的 skc 项的数据补全。处理完数据后，我们从所给数据中分析得到实际价格 (r_c)、折扣率 (D)、库存量 (S_t)、标签价 (T)、价格波动率 (F)、市场占有率 (M) 六个指标，并通过皮尔逊相关系数来分析各指标之间的相关度，且作出相关系数热力图来可视化。为了得到各因素对销量的影响权重，我们选用了随机森林算法。考虑到每个节假日的促销力度以及其经济大环境等的区别，我们首先单独分析四个节假日的数据，得到每个节假日之内各指标权重值；而为了评估该六个指标在各个时间段的综合影响权重，我们又综合了所有节假日的数据计算了各指标在所有节假日中的综合权重值。最后，我们画出分别在四个节假日内以及综合所有节假日的数据的销量随各指标之间的变化趋势散点图来探究销量和各指标的线性相关性。

问题二的分析：在问题二中，研究对象为历史销售时间处于 2019 年 6 月 1 日至 2019 年 10 月 1 日内且累计销售额排名前 10 的小类。我们需要结合上一问的结论，预测这些目标对象同年 10 月 1 日后 3 个月中每个月的销售量，并计算对应的 MAPE 值。这是一个典型的时间序列预测问题。我们首先将市场看作灰色系统，选用灰色预测理论基于 10 月之前的月销量数据进行预测。由于灰色预测所需原始数据量较少，所以我们以月为单位。之后我们又运用了 LSTM 神经网络算法以日为单位来进行预测，并将这两者的误差大小进行比较。而预测的时间段包含了第一问所提到的国庆、双十一、双十二、元旦四个节假日，这些会对预测值产生非连续的影响。该影响的大小可以结合第一问的结论引入一个修正的系数来表

示，从而得到预测时间段中节假日的相关数据。

问题三分析：在问题三中，要求我们研究预测 10 个目标小类下所有 skc 在 2019 年 10 月 1 日后 12 周内每周的周销量，并给出每周预测值的 MAPE。这是在问题二中对 10 个目标小类销量预测的更深入探究；我们先选用 ARIMA 时间序列 +SVM 模型，以周为单位学习目标小类下对应 skc 的历史销量数据，利用 SVM 对预测值残差的修正来提高预测精度；之后通过 Holt-Winters 模型，利用其可捕获数据的周期性与季节性，对于国庆、双十一等节假日的敏感度较高，我们以天为基本都单位，对目标小类下的 skc 销量进行预测，再换算成周销量，计算出对应 MAPE 值。问题四分析：在问题四中，我们将我们的预测算法和结果呈现给了企业，并且分析了模型和算法的合理性和重要意义。同时，我们也指出了我们模型的不足和发展空间。

本论文的解题思路如下图所示

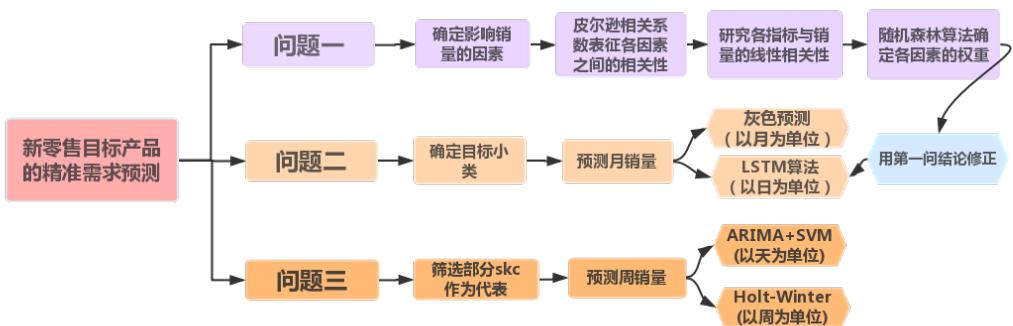


图 1: 思维导图

五. 数据分析和预处理

所提供的数据包含以下内容：

- 表格 1 为销售流水数据，包含有 2018 年 1 月 1 日至 2019 年 12 月 30 日的各 skc 的单日销售量 (s) 以及实际价格 ($real_cost$)，我们认为该实际价格即最后的成交价格；
- 表格 2 为产品信息表，包含 1 表中的部分 skc 在 2017 年记录的对应的小类 ($tiny_class$) 和标签价 (tag_price)，我们假设这部分数据在 2018 至 2019 年不变；
- 表格 3 为区域库存数据，其中有各 skc 在 2018 年 1 月 1 日至 2019 年 12 月 30 日的日库存量 (ie)；
- 表格 4 包含 2018 年和 2019 年各节假日和促销日对应的日期。

在分析处理数据的过程中我们发现数据中有一些字段缺失，如 `season` 项。我们将这些缺失项剔除。销售量为前 50 中的一些 skc 的标价和对应的小类信息缺失。对于信息残缺较多的 skc 项，由于无法对其进行分析，我们将其剔除。对于只有少量信息残缺的 skc 项，我们使用分段三次埃尔米特 (Hermite) 插值法将其补全。

六. 问题一模型的建立与求解

6.1 数据预处理

当计算出每个商品在指定时间段内的总销售额并选出排名前 50 的 skc 后，我们发现一些目标 skc 对应的小类信息和标价信息缺失。对小类信息缺失的 skc，我们通过找到卖出日期信息与之相同或相似的小类，并认为该 skc 即归类为该小类。这个假设是合理的，因为 skc 当天有售出则其所属的小类也会有记录，所以 skc 有很大的概率属于卖出日期信息与之相同或相似的小类。如果找不到对应的小类，则剔除该 skc。对于标价信息缺失的 skc，由于无法分析计算其销量影响因素，我们将其剔除。

6.2 确定销量的影响因素

根据题目提示和文献调研，我们选择了折扣率 D 、实际价格 r_c 、库存量 S_t 、标签价 T 、价格波动率 F 、市场占有率 M 六个销量 (s) 的影响因素。其中，我们认为价格波动率 F 可以反应产品销售特征，如是否呈现季节性的大幅波动和变化等，而这个特征将会对促销日内的销量 (s) 造成影响^[5]。市场占有率的高低则综合体现了产品的信誉、品牌知名度、客户源以及它的垄断性，这些都会对销量 (s) 产生影响^[7]。我们还猜想库存 S_t 则通过影响商家贮存商品的成本从而影响商品的折扣率进而能够影响销量。

我们可以用数据中提供的 skc 的信息间接得到各目标 skc 的该六个指标的值。在所研究时间段内，价格 r_c 、库存量 S_t 、标签价 T 的含义分别各目标 skc 节假日内的 $real_cost$ 、 ie 和 tag_price 的值，折扣率 D 的定义为各 skc 的 $real_cost$ 除以 tag_price ，价格波动 F 为当年的 $real_cost$ 的标准差。其具体计算公式如下：

$$D = \frac{real_cost}{tag_price} \quad (1)$$

$$M = \frac{S_{skc}}{S_{tiny_class}} \quad (2)$$

$$F = \sqrt{\frac{1}{n} \sum_{t=2018/1/1}^{2018/12/1} (real_cost - \overline{rear_cost})^2} \quad (3)$$

其中 $\overline{rear_cost}$ 为年平均值。该六个指标的来源和与所提供的 skc 各信息的关系如下图所示：

6.3 算法分析

6.3.1 皮尔逊相关系数

在统计学中，皮尔逊相关系数是用于度量两个变量 X 和 Y 之间的相关性。其计算公式为

$$R_{xy} = \frac{\sum_{i=1}^{n=1} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

其中 \bar{x} 和 \bar{y} 分别为 x_i 和 y_i ($i = 1, 2, \dots, n$) 的算数平均值。

$|R_{xy}| \leq 1$ 。如果 R_{xy} 为正，则 y 与 x_i 正相关，若反之为负则负相关。且 $|R_{xy}| \leq 1$ 越接近 1，则说

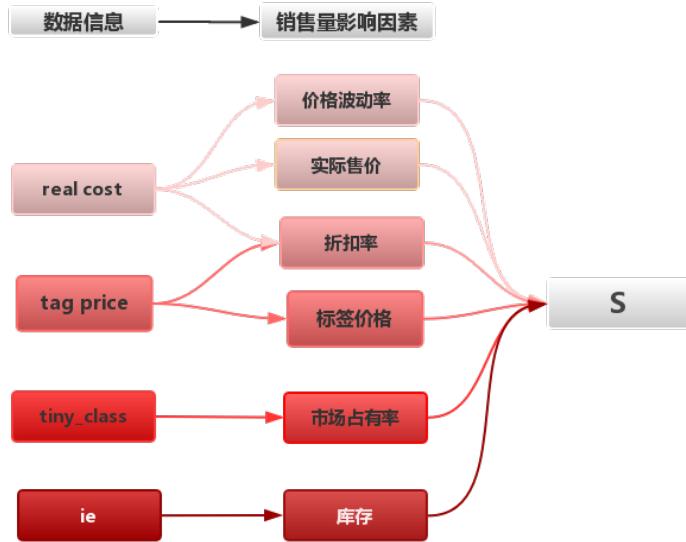


图 2: 六个影响因素与原始信息的关系

明变量 Y 与 X 之间的线性关系越强。

在该问中，我们分别计算了四个假期内以及综合四个假期六个指标之间的皮尔逊相关系数。其结果用热度图表示如下，其中颜色越浅表示 R_{xy} 越大。

图中可以看出四个节假日的相同之处为：

- 实际单价 (r_c) 和标签价 (T) 格均成相关性很高的正相关，这是符合常识的。
- 实际单价 (r_c) 和标签价格 (T) 与库存呈现相对较强的负相关，这可以解释为标价或实际价格 (r_c) 越高，对消费者的经济水平要求越高，故销量 (s) 越少从而导致库存较多。
- 标签价 (T) 或实际标价 (r_c) 与市场占有率为一定的负相关，可见价格越优惠市场占有率越高。
- 标签价 (T) 与折扣率均呈现一定的负相关，这体现了促销日的优惠政策特点，即标价越低的越优惠

四个节假日的不同之处为：

国庆节的标价与折扣率的相关系数最低，而元旦的最高，双十一与双十二也较低；与此同时，实际价格 (r_c) 与折扣率的相关指数中只有国庆为较大的负值，在其他节日实际价格 (r_c) 越高折扣率越低。这反应了国庆节价格高的商品优惠力度最低。

而四个节假日综合图可以进一步证明之前实际单价 (r_c) 和标签价 (T) 格均成相关性很高的正相关以及实际单价 (r_c) 和标签价格 (T) 与库存呈现相对较强的负相关的关系。同时还可以看到综合来看，折扣率 (D) 与市场占有率为正相关，说明折扣率 (D) 越高的商品客户源越多。至于库存

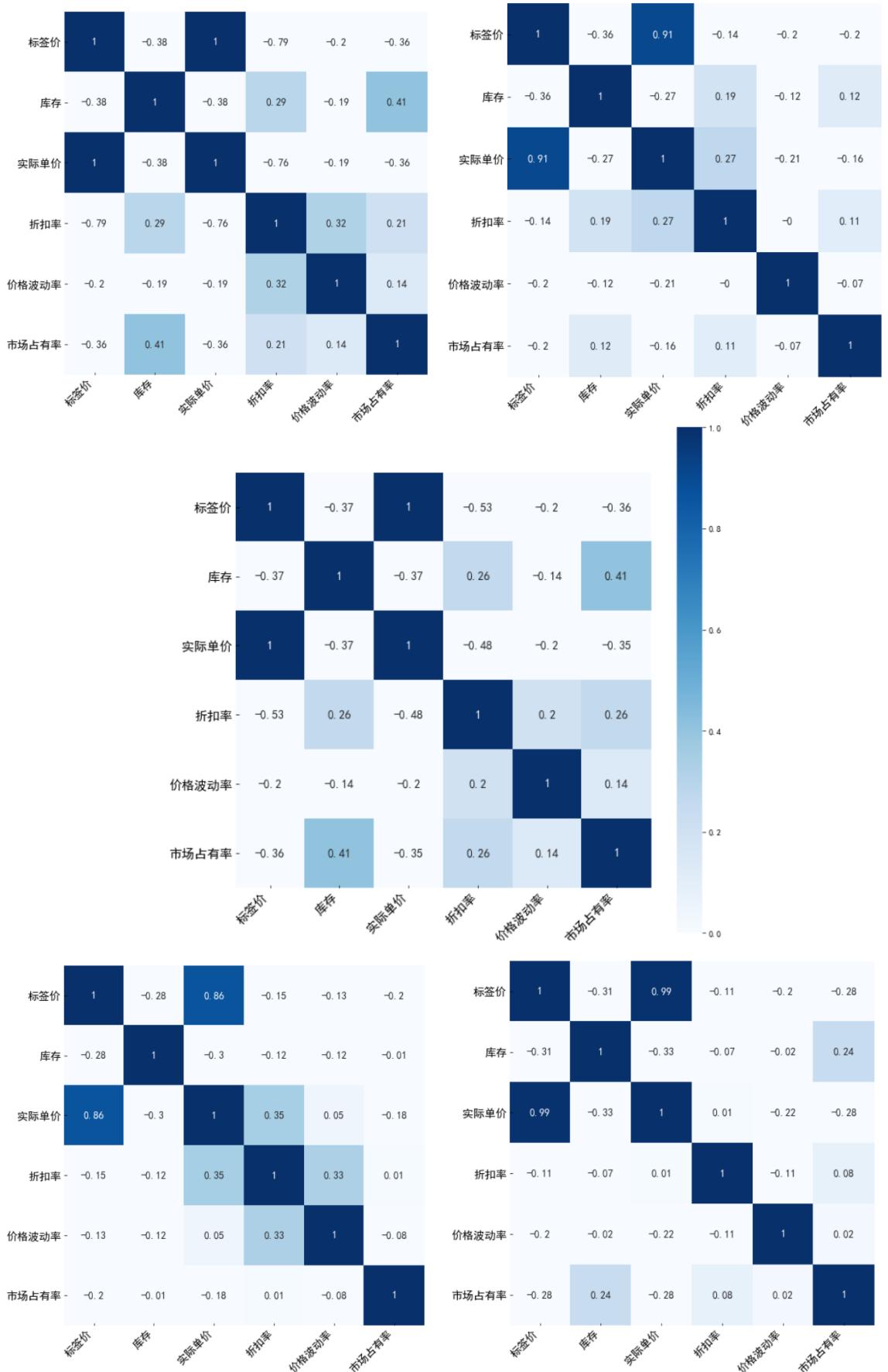


图 3: 各指标之间的皮尔逊相关系数分布热图 (a) 国庆; (b) 双十一; (c) 双十二; (d) 元旦

(S_t) 与市场占有率 (M) 正相关的关系则与我们的猜想相反，说明实际情况应也许是库存量 (S_t) 随客户需求而定，库存量 (S_t) 越高的客户需求越多从而销量越高、市场占有率 (M) 越高。

6.3.2 线性相关性可视化分析

我们首先评估六个指标与销量 (s) 的线性相关性，即画出各时间段内销量 (s) 随六个指标的函数关系图。结果如所示。

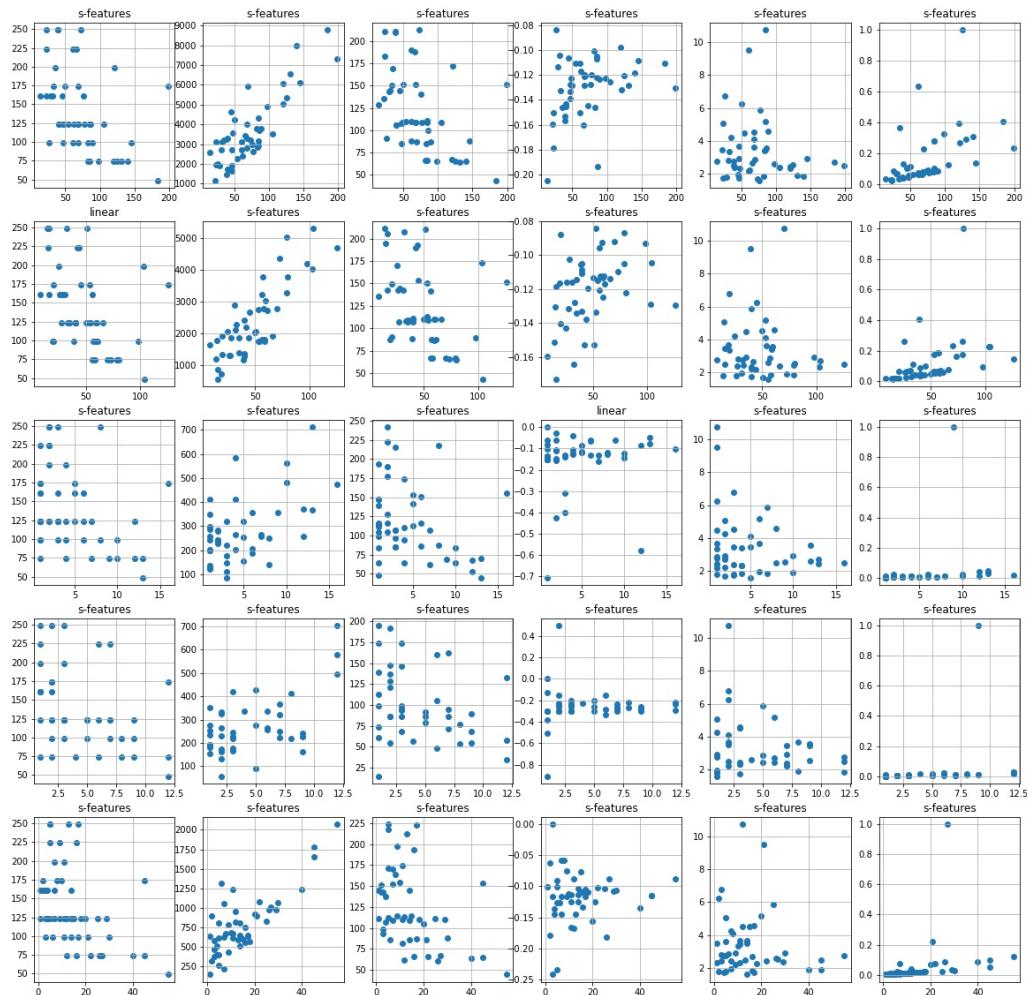


图 4: 各影响因素分别在四个节假日内与销售量的线性相关度

从该图可以看出，除了库存量 (S_t) 外所有的指标与销量 (s) 都没有明显的线性关系。而对于库存量 (S_t)，从图 4 a) 来看与销量 (s) 为正的线性相关。这与实际相符合，一般库存量 (S_t) 与销量 (s) 的比值

在一个范围内波动。

实际价格 (r_c)、标签价格 (T) 与价格波动与销量 (s) 关系图的数据点主要集中在左下角，呈现一定的负相关性。

6.3.3 随机森林评估模型

6.3.3.1 随机森林算法原理

随机森林是一种利用多棵决策树对样本进行训练并预测的一种分类器，其原理由 Breiman 等人提出^[1]，其原理图如下所示：

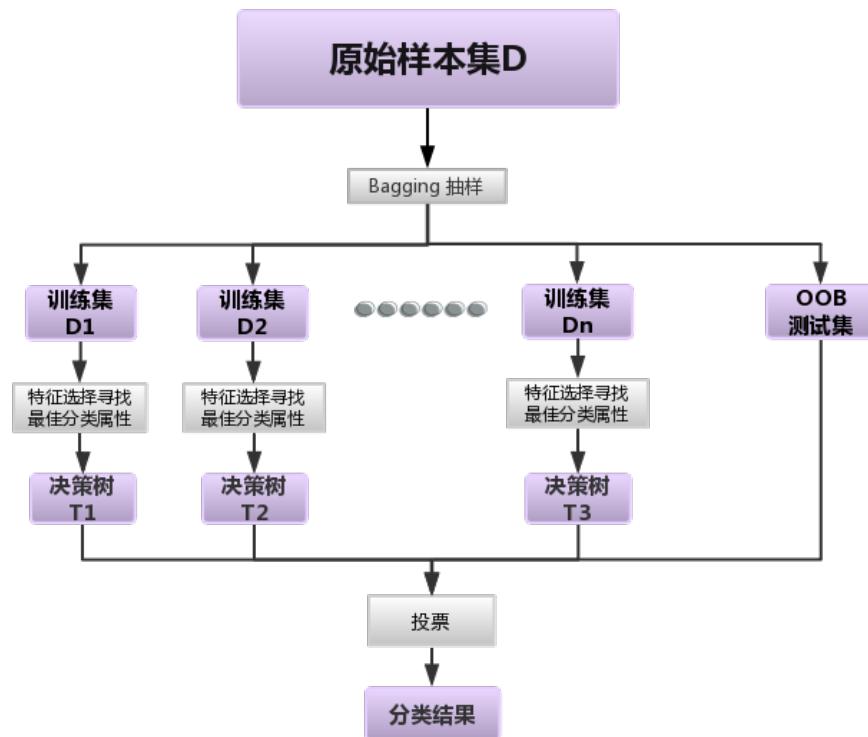


图 5：随机森林流程图

随机森林使用的 CART 决策树是基于基尼系数进行特征选择的，所以我们主要基于讨论基于基尼系数的特征选择。

在决策树中，设总共 K 类，一个样本是第 k 类的概率表示为 p_k ，此时基尼系数表示为：

$$Gini(p) = \sum_{k=1}^{K-1} p_k(1-p_k) = 1 - \sum_{k=1}^{K-1} (p_k)^2 \quad (5)$$

基尼指数越大代表不确定性越大。遍历每个特征的每个分割点时计算其基尼系数。在训练随机森林中，不断遍历 CART 决策树的特征子集的所有可能的分割建立最终决策树，将 Gini 系数最小的特征作为分割点，此时数据集划分为两部分。该过程进行至一直到满足退出停止条件。最后可以得到各因素的重要性评分。

由于国庆、双十一、双十二、元旦四个节假日中商家宣传力度、优惠政策、物价水平以及人们的消费行为会有所差异，我们首先分别研究四个节假日内六个指标的重要性重要性，即分别用四个假期的数据集进行随机森林算法处理得到各指标的权重。然后为了看到在各时间段和大环境下六个指标的综合重要性，我们用所有假期的数据集得到六个指标的综合权重。结果如下图所示：

6.3.3.2 随机森林算法计算权重

在该问中，我们使用随机森林来对进行以上六个影响因子的重要性进行评估，算出其权重。

在各影响因素分别在四个节假日的权重分布图中，可以看到国庆节的分布特征与其他三个节日内的差别较大，其各因素中价格标准差 (F)、库存 (S_t)、市场占有率为 (M)、实际单价 (r_c)、折扣率 (D) 的权重相近，而标签价 (T) 的权重很小。与其他三个节假日相比，其市场占有率为 (M)、标签价 (T) 的权重最低，而实际单价 (r_c) 和价格标准差 (D) 的影响最高。结合实际，这差异这反应了国庆节的促销力度较低，与皮尔逊相关系数分析中得到的结论一致。

对双十一、双十二和元旦三个节日来说，市场占有率为 (M) 对销量 (s) 的影响权重普遍较高；尤其是双十二而言，市场占有率为 (M) 权重最高。可以进一步发现，双十二与元旦的趋势相近，而双十一中折扣率的权重非常大，再结合实际可以知道，双十一的优惠力度在所有促销节日中最大。

综合四个节假日来看，折扣率的影响因素最高，其次是实际价格 (r_c)、市场占有率为 (M) 和价格变化率。由于 (T) 与实际单价 (r_c) 和折扣率有关，所有不能从其权重较低得出结论。

6.3 结论

通过对数据进行处理、确定并计算得到六个销量影响因素之后，我们对各因素之间进行皮尔逊相关系数分析，再利用随机森林算法得到各个因素对销量影响因素的权值。最后可视化各因素与销量之间的线性相关关系。值得注意的是，我们将四个节假日分开考虑，即我们没有包括时间该因素。并且我们还综合四个节假日的数据得到各指标的权重和与销量的线性相关性来观察其综合影响

从以上分析中，我们可以得到商家的一些销售特征。例如可以从皮尔逊相关系数和权重分析得到双十一的优惠力度最大，而国庆最小。

同时，我们还能发现在对销量的影响中，综合四个节日的信息来看折扣率是影响最大的，尤其是在双十一促销和优惠力度最大的时候。而对于促销力度不大的时候，如国庆节，标签价 (T) 和市场占有率为 (M) 的影响最小，而实际单价 (r_c) 的影响最高，反应了此时顾客关注的主要实际的成交价格。这些信息将可以指导零售商的销售活动。

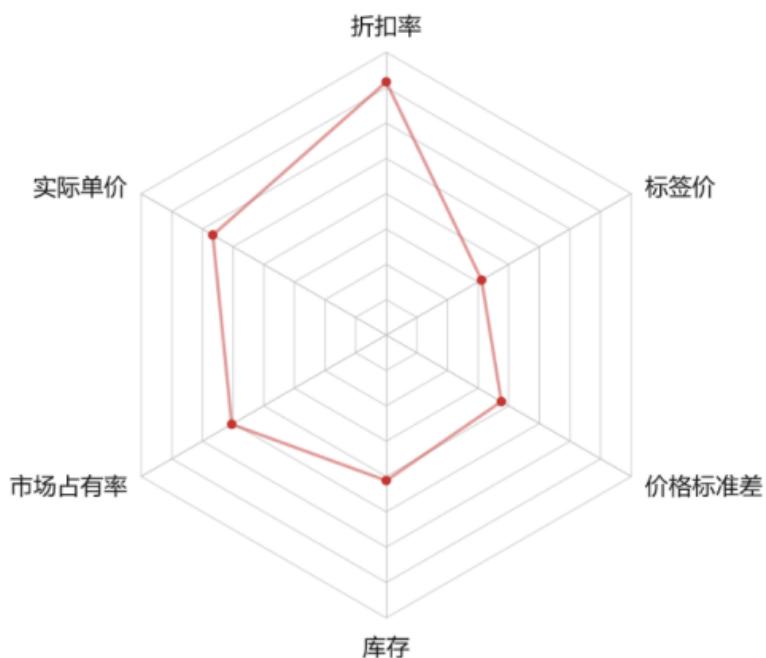
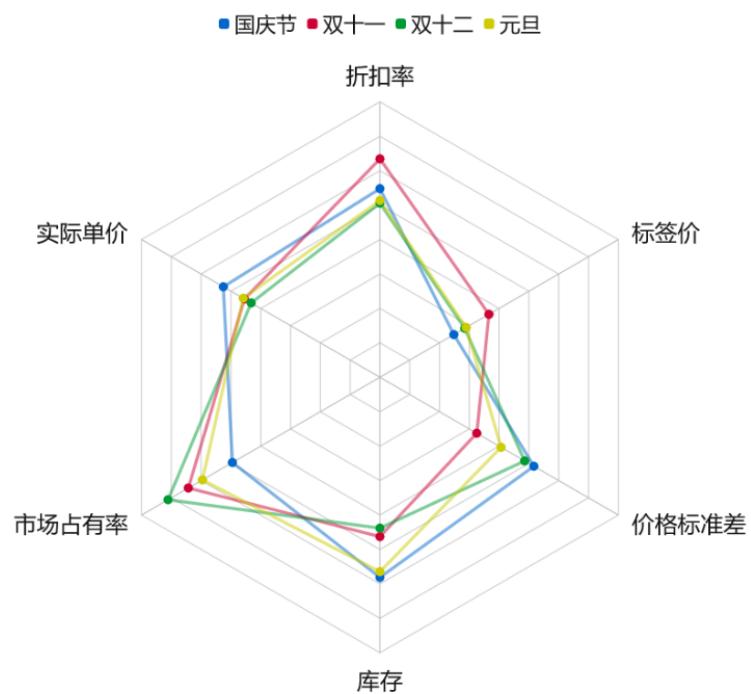


图 6: 各销量 (s) 影响因素权重分布图 (a) 分别在四个节假日; (b) 综合所有节假日

七. 问题二的建模与求解

7.1 数据预处理

该问要求以历史销售时间处于 2019 年 6 月 1 日至 2019 年 10 月 1 日时累计销量前十的小类为研究对象，预测它们在之后 10 月、11 月、12 月中的月销量。

我们首先筛选出该目标小类并作出其前 9 月的月销量随时间变化图。如所示，各小类在前九个月销量均基本呈现上升趋势，且增速和变化趋势较为接近。

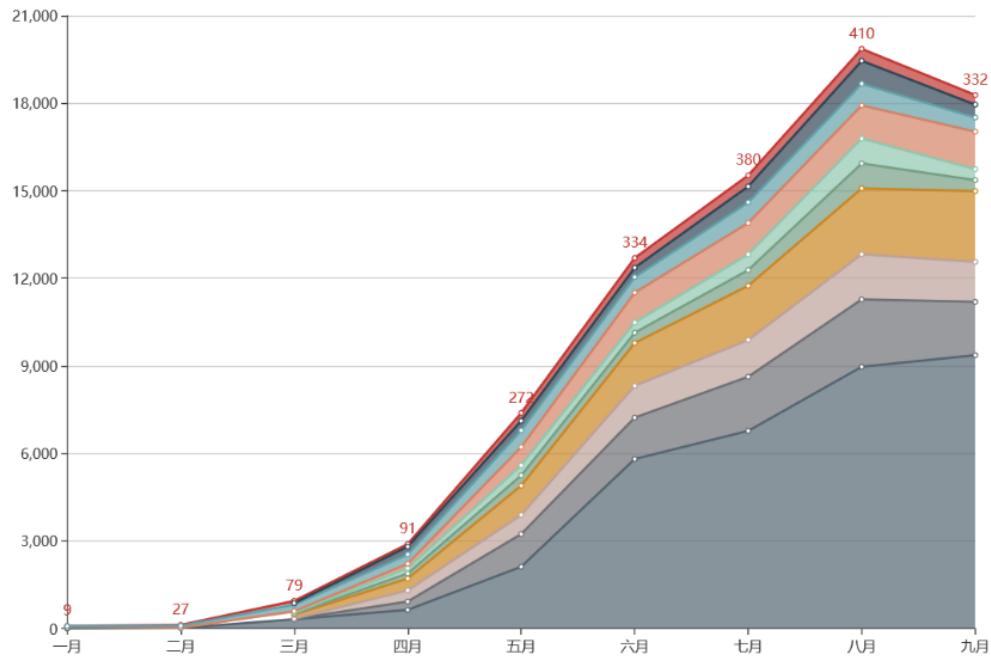


图 7: 2019 年前九个月目标小类的月销量变化趋势图

7.2 算法分析

7.2.1 灰色预测模型

7.2.1.1 灰色预测原理

灰色预测是对既含有已知信息又含有不确定信息的系统（即灰色系统）进行预测^[8]。它通过鉴别系统因素之间发展趋势的相异程度，即进行关联分析，并对原始数据进行生成处理来寻找系统变动的规律，生成有较强规律性的数据序列，然后建立相应的微分方程模型，从而预测事物未来发展走势的状况。灰色预测的主要特点是模型使用的不是原始数据序列，而是生成的数据序列。其核心体系是灰色模型（Grey Model，简称 GM），即对原始数据作累加生成（或其它方法生成）得到近似的指数规律再进行建模的方法。灰色预测模型对于不同问题采用不同模型，GM（1，1）模型主要解决生成序列是有指数变化规律，只能描述单调的变化过程。其主要步骤是将原始数据列中的数据按照某种要求作数据处理称为灰色生成。

对原始数据的生成就是企图从杂乱无章的现象中去发现内在规律。其原理和步骤如下^[13]:

1. 数据的检验预处理

首先，为了保证建模方法的可行性，需要对已知数列做必要的检验处理。

设参考数据为 $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n))$ ，计算数列的级比:

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, k = 2, 3, \dots, n \quad (6)$$

如果所有的级比 $\lambda(k)$ 都落在可容覆盖 $\Theta = (e^{-\frac{2}{n+1}}, e^{\frac{2}{n+2}})$ 内，则数列 $x^{(0)}$ 可以作为模型 GM(1,1) 的数据进行灰色预测。否则，需要对数列 $x^{(0)}$ 做必要的变换处理，使其落入可容覆盖内。即取适当的常数 c ，作平移变换

$$y^{(0)}(k) = x^{(0)}(k) + c, k = 1, 2, \dots, n \quad (7)$$

则使数列 $y^{(0)} = (y^{(0)}(1), y^{(0)}(2), \dots, y^{(0)}(n))$ 的级比

$$\lambda_y(k) = \frac{y^{(0)}(k-1)}{y^{(0)}(k)} \in \Theta, k = 2, 3, \dots, n \quad (8)$$

2. 建立模型

将处理后的数列 $y^{(0)} = (y^{(0)}(1), y^{(0)}(2), \dots, y^{(0)}(n))$ ，做一次累加 (AGO) 生成数列

$$\begin{aligned} y^{(1)} &= (y^{(1)}(1), y^{(1)}(2), \dots, y^{(1)}(n)) \\ &= (y^{(1)}(1), y^{(1)}(1) + y^{(0)}(2), \dots, y^{(1)}(n-1) + y^{(0)}(n)) \end{aligned} \quad (9)$$

其中 $y^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i)$ ($k = 1, 2, \dots, n$)。求均值数列:

$$z^{(1)}(k) = 0.5y^{(1)}(k) + (2), 0.5y^{(1)}(k-1), k = 2, 3, \dots, n \quad (10)$$

则 $z^{(1)}(k) = (z^{(1)}(1), z^{(1)}(1) + z^{(0)}(2), \dots, z^{(1)}(n-1) + y^{(0)}(n))$ 。于是建立灰微分方程为

$$x^{(0)}(k) + az^{(1)}(k) = b, k = 2, 3, \dots, n \quad (11)$$

相应的白化微分方程为:

$$\frac{dx^{(1)}}{dt} + ax^{(1)}(t) = b \quad (12)$$

记 $u = (a, b)^T$, $Y = (y^{(0)}(2), y^{(0)}(3), \dots, y^{(0)}(n))^T$, $B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(2) & 1 \\ \dots & \dots \\ -z^{(1)}(n) & 1 \end{bmatrix}$, 则由最小二乘法, 求得使 $J(\hat{u}) = (Y - B\hat{u})^T(Y - B\hat{u})$ 达到最小值的 $\hat{u} = (a, b)^T = (B^T B)^{-1} B^T Y$ 。于是求解方程 (1) 得到预测值

$$y^{(1)}(k+1) = y^{(0)}(1) - \frac{b}{a} e^{-ak} + \frac{b}{a}, k = 0, 1, \dots, n-1, \dots \quad (13)$$

3. 检验预测值

为检验按灰色模型预测的可信性, 需要进行后验差检验。其原理和计算步骤如下:

原始数据列的实际数据的平均值 \bar{y} 与方差 S_1^2 分别为:

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X^{(0)}(k) \quad S_1^2 = \frac{1}{n} \sum_{k=1}^n (X^{(0)}(k) - \bar{X})^2 \quad (14)$$

把第 k 项数据的原始数据值 $X^{(0)}(k)$ 与估计值之差 $X^{(0)}(k)$ 称为残差:

$$q(k) = X^{(0)}(k) - \hat{X}^{(0)}(k) \quad (15)$$

则整个数据列所有数据项的残差的平均值 \bar{q} 和方差 S_2^2 分别为:

$$\bar{q} = \frac{1}{n} \sum_{k=1}^n q(k) \quad S_2^2 = \frac{1}{n} \sum_{k=1}^n (q(k) - \bar{q})^2 \quad (16)$$

通过计算后验差比值 C 来进行后验差检验

定义: 后验差比值为 $C = \frac{S_2^2}{S_1^2}$

后验差比值 C 越小越好, C 越小按灰色模型计算的估计值与实际值越接近

7.2.1.2 灰色预测销量

本题中由于销量 (s) 的影响因素繁多, 市场可以看作一个灰色系统, 故我们首先采用灰色预测中的 GM(1,1) 模型来预测各目标小类在 10 月之后三个月的销量 (s) 的值。

由于灰色预测所需的数据量较少, 我们决定原始数据列以月为单位。我们计算并提取出目标小类于 2019 年前 9 个月的日销量 (s) 数据并取每月的平均值作为原始数据来进行预测 2019 年。预测的前四类的结果如图 8 所示:

可以看到, 用灰色预测得到的结果与真实值相差较大。对结果进行后验差检验, 发现三个的值均高于 0.5, 预测精度勉强, 误差较大。其原因可能为灰色预测输入的数据以月为单位、精度较低, 且无法预测双十一、双十二等节假日中销量突增的现象。

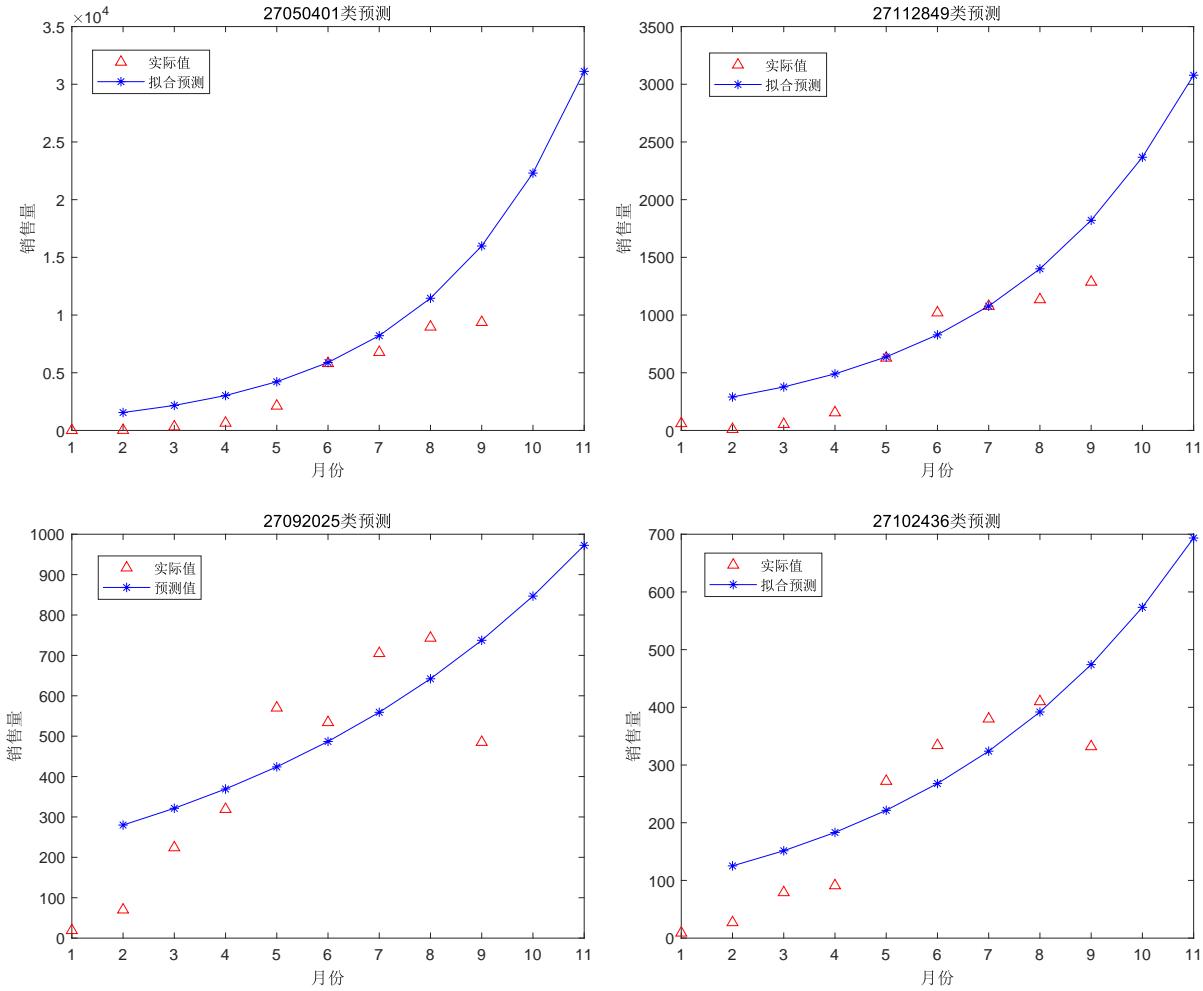


图 8: 灰色预测 4 个目标小类销量结果图

7.2.2 基于 LSTM 神经网络的销量预测模型

7.2.2.1 LSTM 神经网络算法原理

由于灰色预测方法数据量低、误差较大，我们又采用了长短期记忆人工神经网络 (Long-Short Term Memory, LSTM) 方法预测日销售量。

时间序列模型最常用最强大的的工具就是递归神经网络 (recurrent neural network, RNN)^[?]。而长短期记忆人工神经网络 (Long-Short Term Memory, LSTM) 是一种改进的 RNN，比一般的 RNN 能够记住更长周期上的信息模式，在解决很多问题上都取得了成功，例如自然语言处理、中文文本分类研究、机器翻译等^[?]。LSTM 模型避免了长期依赖问题，采用特殊隐式单元，在继承了大部分 RNN 模型特性的同时解决了梯度反传过程中由于逐步缩减而产生的 Vanishing Gradient 问题，适用于非线性回归变量，可以解决多个输入变量的问题，模型准确度高，训练速度快，并行处理能力强^[?]。LSTM 更适合用于处理与短期时间序列高度相关的问题，在 n 个示例批次中不断迭代，能够快速和准确地对大量短期时间序列数据进行处理，是解决时间序列预测问题最常用的工具。其原理图如下所示：

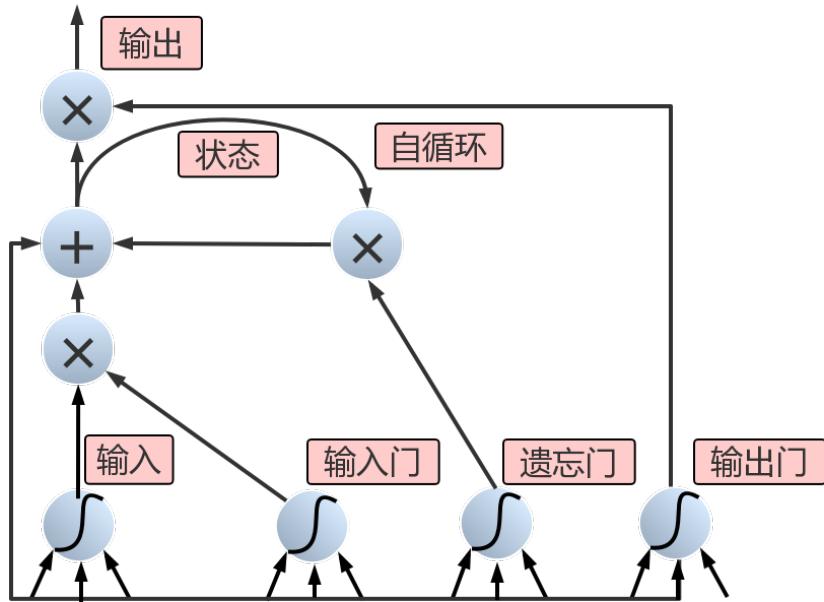


图 9: LSTM 结构图^[?]

如原理图所示，LSTM 设计两个门控制记忆单元状态 c 的信息量：一个是遗忘门 (forget gate)。所谓的“遗忘”，也就是“记忆的残缺”。它决定了上一时刻的单元状态有多少“记忆”可以保留到当前时刻；另一个是输入门 (input gate)，它决定了当前时刻的输入有多少保存到单元状态。实际上，为了表述方便，很多文献还添加了一个门，叫候选门 (Candidate gate)，它控制着以多大比例融合“历史”信息和“当下”刺激。最后，LSTM 还设计了-个输出门 (output gate)，来控制单元状态有多少信息输出。下面对这 4 个门分别进行详细介绍。

$$\text{遗忘门模型: } f_t = \sigma(W_f^T \times s_{t-1} + (U_f)^T \times x_t + b_f)$$

$$\text{输入门模型: } i_t = \sigma(W_i^T \times s_{t-1} + (U_i)^T \times x_t + b_i)$$

$$\text{候选门模型: } C'_t = \tanh(W_c^T \times s_{t-1} + U_c^T \times x_t + b_c)$$

$$\text{于是记忆单元的模型函数就是: } C_t = f_t \times C_{t-1} + i_t \times C'_t$$

$$\text{输出门的模型是: } O_t = \sigma(W_o^T \times s_{t-1} + U_o^T \times x_t + b_o) \text{ 最终的时间序列上的输出量是: } s_t = O_t \times \tanh(C_t)$$

7.2.2.2 LSTM 预测销量

我们将经过剔除和插值处理后的 2019 年 1 月到 9 月的 *tiny_class* 日销量的数据集分割为训练集、测试集和验证集。它们的占比分别为是 70%,15%,15%。在训练过程中，我们对参数的设置为： $epochs = 5$, $batch_size = 1$, $verbose = 2$ 。训练过程的损失函数如下图所示：

如图所示，经过 50 次迭代之后，测试的准确度达到了 0.82 而且测试损失率降低 0.18，这说明该神经网络已经训练得具有较高的准确度。我们用它来预测 2019 年 10 月 1 日至年底的销量，结果如图 17 所示。

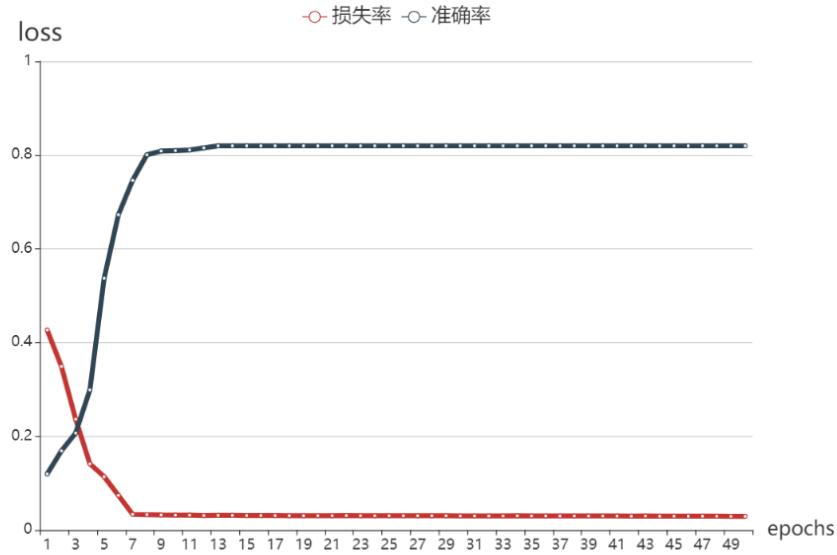


图 10: 训练过程

我们发现预测结果误差较大，其原因可能为学习时间过短，只有 2018 年前九个月的数据，所以很难预测双十一、双十二等节假日的冲击。

为解决该问题，我们利用了第一问的结论，假定 2019 年的节假日对销量的影响与 2018 年相同，我们便可以在 2019 年预测模型中加上 2018 年的历史数据中双十一、双十二和元旦对销量的冲击，来弥补机器学习时间过短无法预测节假日中销量突变的问题。具体操作是对 2019 年 LSTM 预测模型进行参数调整，例如使用 softsign 激活函数替代 tanh。

由于从图中可以得到每个小类的销量增减趋势相同，我们选取了销量第一的小类调整前与调整后的预测结果为例来体现所有小类调整后预测的效果。

为评估预测的准确性，我们继而计算了其 MAPE 的值。MAPE 的计算公式如下：

$$MAPE = \sum_{i=1}^n \frac{y_i - \hat{y}_i}{n * y_i} = \sum_{i=1}^n \frac{1}{n} * \frac{1}{n} \sum_{i=1}^n * \frac{y_i - \hat{y}_i}{y_i} \quad (17)$$

其结果如下表所示，可以看到 MAPE 值基本均小于 0.1，说明预测结果较为准确。

该问详细完整的预测结果见附录文件

7.3 结论

本问对目标小类的销量预测我们首先尝试了灰色预测，由于灰色预测所需数据量以月为单位，得到的灰色预测精度较低，尤其是无法预测节假日销量的突变。于是我们使用 LSTM 算法以天为单位预测销量，并且通过结合第一问的结论加上节假日对销量的影响弥补时间机器学习时间过短的问题，使得预测的准确性大大提高。

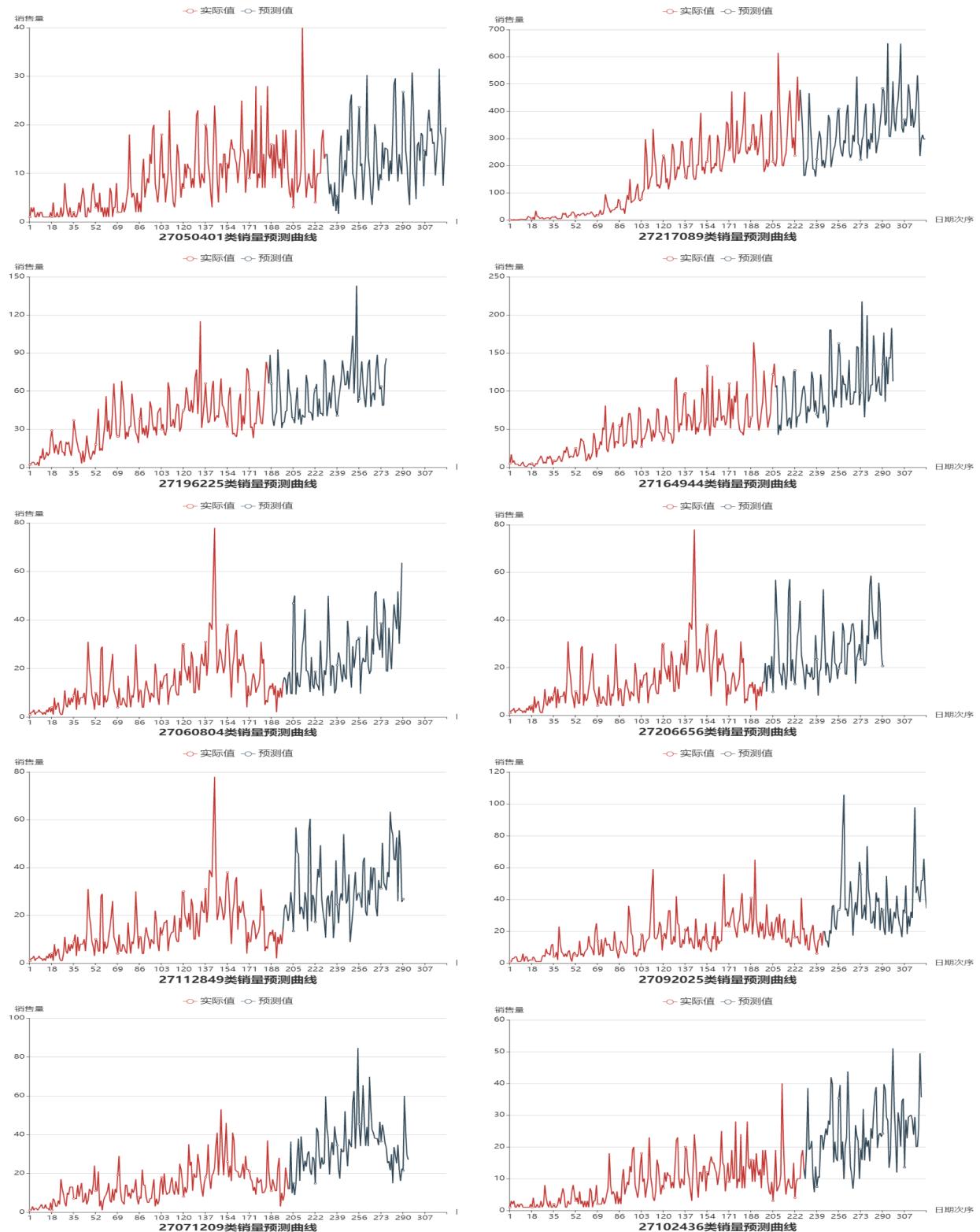


图 11: LSTM 预测 10 个目标小类销量结果图

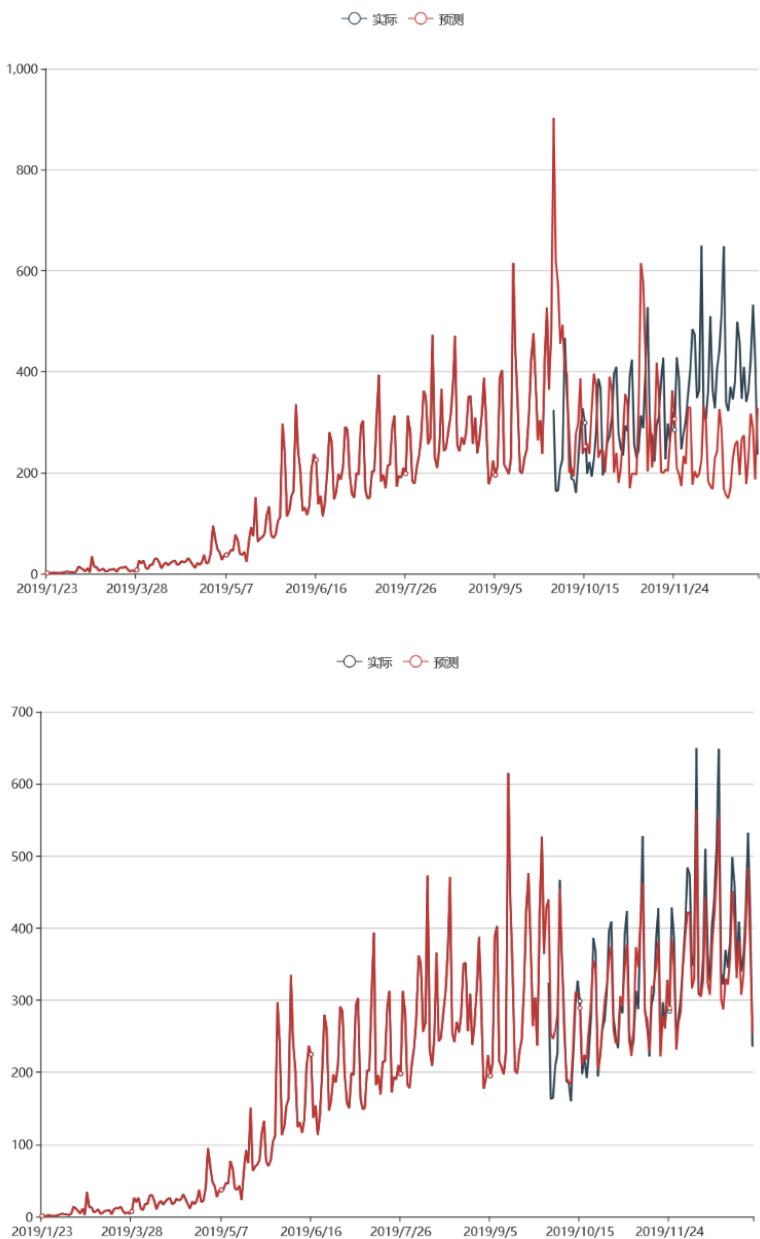


图 12: LSTM 预测第一个小类销量参数调整前后的预测图

小类	1号	2号	3号	4号	5号	6号	7号	8号	9号	10号
10月	0.0856	0.0471	0.0836	0.0970	0.0192	0.0255	0.0919	0.0169	0.0294	0.0791
11月	0.0492	0.1007	0.0989	0.0688	0.0294	0.0565	0.0302	0.0861	0.0806	0.0688
12月	0.0549	0.0459	0.1002	0.0292	0.0741	0.0335	0.0234	0.0745	0.0212	0.0466

表 13: LSTM 预测结果 MAPE 值

八. 问题三的建模和求解

8.1 数据预处理

第三问要求预测第二问中目标小类内所有 skc 在 2019 年 10 月 1 日后 12 周内每周的周销量。而于第二问中的目标小类所含的 skc 的数目太多，且有一些 skc 日销量的数据不全，不适合机器学习。因此我们选取每个目标小类销售量排名前十的子类 skc 作为研究对象。由于我们发现部分 skc 的销售量和整体的销售量满足 $s_{tiny_class} = K \cdot s_{skc} + b$ 线性关系，且所选用的 skc 占其所属小类的比重较大如下表所示，所以我们可以认为该部分 skc 的销量变化特征可以反映整个目标小类的销量变化特征。

目标小类	skc数目	选用skc销量占比
27050401	700	76.20%
27217089	89	84.30%
27196225	94	86.40%
27164944	169	79.60%
27060804	354	72.90%
27206656	40	81.30%
27112849	273	91.40%
27092025	112	88.70%
27071209	103	84.50%
27102436	72	82.60%

表 14: 选取的 skc 在各小类中销量的比重

8.2 算法分析

8.2.1 ARIMA 时间序列 +SVM 模型预测销量模型

8.2.1.1 ARIMA 和 SVM 的原理

ARIMA 模型 (Autoregressive Integrated Moving Average model)，即差分整合移动平均自回归模型，又称整合移动平均自回归模型，是时间序列预测分析方法之一。ARIMA 是在平稳的时间序列基础上建立起来的，因此时间序列的平稳性是建模的重要前提。如果时间序列不稳定，也可以通过一些操作去使得时间序列稳定（比如取对数，差分），然后进行 ARIMA 模型预测，得到稳定的时间序列的预测结果，最后对预测结果进行之前使序列稳定的操作的逆操作（取指数，差分的逆操作），就可以得到原始数据的预测结果。其具体过程如下^[12]：

ARIMA(p,d,q) 模型:

$$(y')_t = \alpha_0 + \sum_p^{i=1} \alpha_i (y')_{t-i} + \epsilon_t + \sum_{i=1}^q \beta_i \epsilon_{t-i} \quad (18)$$

$$y'_t = \delta^d y_t = (1 - L)^d y_t \quad (19)$$

$$(1 - \sum_{i=1}^p \alpha_i L^i)(1 - L)^d y_t = \alpha_0 + (1 + \sum_{i=1}^q \beta_i L^i) \epsilon_t \quad (20)$$

其中:

p 代表预测模型中采用的时序数据本身的滞后数 (lags), 也叫做 AR/Auto-Regressive 项;

d 代表时序数据达到稳定的差分化的阶数, 也叫 Integrated 项。

q 代表预测模型中采用的预测误差的滞后数 (lags), 也叫做 MA/Moving Average 项

ARIMA 优点为模型十分简单, 只需要内生变量而不需要借助其他外生变量。但是缺点也很显著: 1. 它要求时序数据是稳定的 (stationary), 或者是通过差分化 (differencing) 后是稳定的。如果不稳定的数据, 是无法捕捉到规律的。比如股票数据用 ARIMA 无法预测的原因就是股票数据是非稳定的, 常常受政策和新闻的影响而波动。2. 本质上只能捕捉线性关系, 而不能捕捉非线性关系。而 SVM(Support Vector Machine) 又称为支持向量机, 是一种分类模型^[3]。支持向量机可以分为线性核非线性两大类。其主要思想为找到空间中的一个能够将所有数据样本划开的超平面, 并且使得本本集中所有数据到这个超平面的距离最短。

SVM 向量模型该模型的提出解决了灰色系统误差较大问题, 避免了神经网络存在的模型结构难确定、精度难保证及泛化能力不高的缺点, 为非线性滑坡预测模型在寻优过程中易陷入局部最优的情况提供了新的思路。其模型步骤为^[?]:

设样本集 $T = (x_1, y_1), \dots, (x_l, y_l), \dots (x_l, y_l) (i = 1, 2, \dots, l)$, 其中 $x_i \in R^d$ 为 d 维输入向量, $y_i \in R$ 为系统输出。

考虑用线性函数 $f(x) = \omega \cdot x + b$ 拟合数据, 采用 ε 误差不敏感函数, 并允许拟合误差 $\varepsilon_i^{(*)}$ 存在的软间隔支持向量回归及 (ϵ -SVR) 可描述如下:

$$\min_{\omega, b, \varepsilon^*} \tau(\omega, b, \xi^{(*)}) = \frac{1}{2} (\|\omega\|)^2 + C \sum_{i=1}^l (\varepsilon_i + (\varepsilon_i)^*) \quad (21)$$

$$s.t. \begin{cases} y_i - (\omega \cdot x_i + b) \leq \epsilon + \varepsilon_i \\ (\omega \cdot x_i + b) - y_i \leq \epsilon + \varepsilon_i^* \epsilon^{(*)} \geq 0 \end{cases} \quad (22)$$

其中 $\epsilon > 0$ 为控制拟合精度的参数, $C > 0$ 为惩罚系数, $\varepsilon^{(*)} = (\varepsilon_1, \varepsilon_1^*, \dots, \varepsilon_i, \varepsilon_i^*, \dots, \varepsilon_l, \varepsilon_l^*)^T$ 为松弛变量。利用 Langrange 优化方法, 可求得 24 中的对偶优化问题:

$$\min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i, x_j) - \sum_{i=1}^l (\alpha_i^* - \alpha_i) y_i + \sum_{i=1}^l (\alpha_i^* - \alpha_i) \epsilon \quad (23)$$

$$s.t. \begin{cases} \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0 \\ \alpha_i^*, \alpha_i \in [0, C], i = 1, \dots, l \end{cases} \quad (24)$$

其中 $K(x_i, x_j) = (F(x_i) \cdot F(x_j))$ 是核函数, 它等价于将向量 x 映射到空间 $F(x)$ 后的内积。目前

比较常用的核函数包括：多项式核函数 $K(x_i, x_j) = (x_i \cdot x_j + c)^d$ (其中 $c \geq 0, d$ 为任意整数)；高斯径向基核函数 (RBF 核) $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2})$ ；Sigmoid 核函数 $K(x_i, x_j) = \tanh(\kappa(x_i, x_j) + v)$ (其中 $\kappa \geq 0, v \leq 0$) 等。通过求解方程 (2) 得解向量 $(\alpha, (\alpha^*))$, 对新输入的 x , 构造回归预测函数

$$f(x) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) k(x_i, x) + b \quad (25)$$

SVM 方法的理论基础是非线性映射, 即利用内积核函数代替向高维空间的非线性映射^[?]。所以 SVM 在解决非线性问题更有优势^{[14][?]}

而它的缺点是 SVM 算法对大规模训练样本难以实施且用 SVM 解决多分类问题存在困难^[6]

如前所述, ARIMA 与 SVM 模型各有优缺点, 但由于分别对线性模型及非线性模型处理具有优势, 他们之间存在优势互补, 因此, 二者组合起来进行价格预测, 可能会收到较好结果^[10]。假设时间序列 Y , 可视为线性自相关部分 L 与非线性残差 N , 两部分的组合, 即: $Y_t = L_t + N_t$, 本文拟采取如下步骤构建组合预测模型:

(1) 利用 ARIMA 模型对线性部分建模, 设预测结果为 L , 序列汇的残差为 N_t, N_t , 中包含了序列 Y , 的非线性关系。

(2) 对上步得到的 N_t , 序列进行重构得到 SVM 样本集, 利用 SVM 对残差进行预测, 得到预测结果 \hat{N}_t 。

(3) 将线性预测得到的 L 与非线性集合得到的 \hat{N}_t 组合, 得到预测结果 $Y_t = L_t + \hat{N}_t$ ^[2]。组合预测原理如下图:

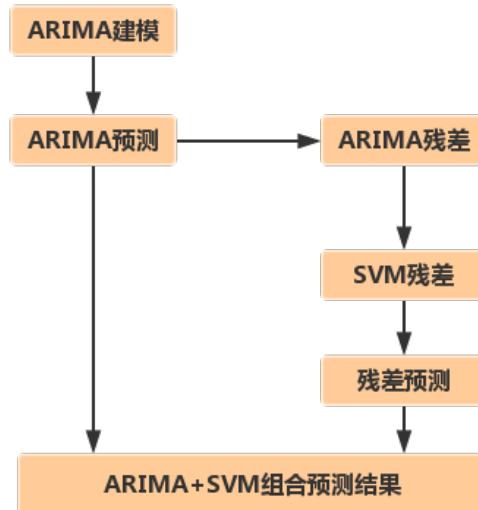


图 15: ARIMA+SVM 算法结构框图

8.2.1.2 ARIMA 时间序列 +SVM 预测

在本问中，我们即借鉴该思路来预测 skc 的销量。我们以天为单位代入各目标小类销量前 10 个 skc 的前九个月的数据，预测 2019 年 10 月 1 日至 12 月 31 日的日销售量，最后再换算成周销售量。

我们将十个小类的销售量以及其对应的十个 skc 代表的销量结果绘制成以下的十张图。同样，我们计算预测结果的 MAPE 值来评估其准确性。计算结果显示 MAPE 值较大。我们选取了部分 skc 销量预测值的 MAPE 为代表呈现于下表，表中数值普遍大于 0.2 或接近 0.2。可知该预测结果误差较大。原因可能为学习的数据量太少。虽然以天为单位进行数据学习，可以使得预测结果大致反映历史特征，但是由于数据时间跨度太小、无法反应异常值，即无法预测时间序列上的突发畸变，所以 10 月到 12 月中节假日的冲击带来的销量的突增无法预测到。

8.2.2 Holt-Winters 模型

基于 ARIMA 和 SVM 无法感知节假日冲击的弱点，我们选用了能体现季节特征的 Holt-Winters 模型，并一周为单位进行预测。由于历史数据中四月、五月、六月和九月均存在一些节日对销量呈现一定周期性的影响，且影响的趋势相同，而运用 Holt-Winter 能够更学习这些周期特征，故而能够更准确预测 10 月至 12 月的销量、反应在其中四个节假日的增减趋势^[9]。

8.2.2.1 Holt-Winters 算法原理

Holt 和 Winters 扩展了 Holt 的方法来捕获季节性。Holt-Winters 季节性方法包括预测方程和三个平滑方程，一个用于水平 l_T ，一个用于趋势 b_T ，一个用于季节性成分 s_T ，具有相应的平滑参数 α , β^* 和 γ 。我们用 m 表示季节性的频率，即一年中的季节数^[7]。

此方法有两种变化形式，它们在季节性成分的性质上有所不同。当季节性变化在整个系列中大致恒定时，首选加法；而当季节性变化与系列的水平成比例时，则采用乘法法。通过加法，季节分量以观测序列的比例尺中的绝对值表示，而在水平方程中，通过减去季节分量来季节性调整序列。在每年内，季节性成分总计约为零。使用乘法方法时，季节成分以相对术语（百分比）表示，并且序列除以季节成分可对季节进行调整^[11]。

加法的组成形为：

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (26)$$

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1} - 1) \quad (\text{水平平滑方程}) \quad (27)$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1} \quad (\text{趋势平滑方程}) \quad (28)$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m} \quad (\text{季节平滑方程}) \quad (29)$$

$$(30)$$

其中 k 是 $\frac{h-1}{m}$ 的整数部分，可确保用于预测的季节性指数的估计来自样本的最后一年。级别方程式显示了经过季节性调整的观测值 $(y_t - s_{t-m})$ 与非季节性预测 $(l_{t-1} + b_{t-1})$ 之间的加权平均值时间。趋势方程与 Holt 线性方法相同。季节性方程式显示当前季节指数 $(y_t - l_{t-1} - b_{t-1})$ 和去年同一季节的季节指数之间的加权平均值。季节性分量的等式通常表示为：

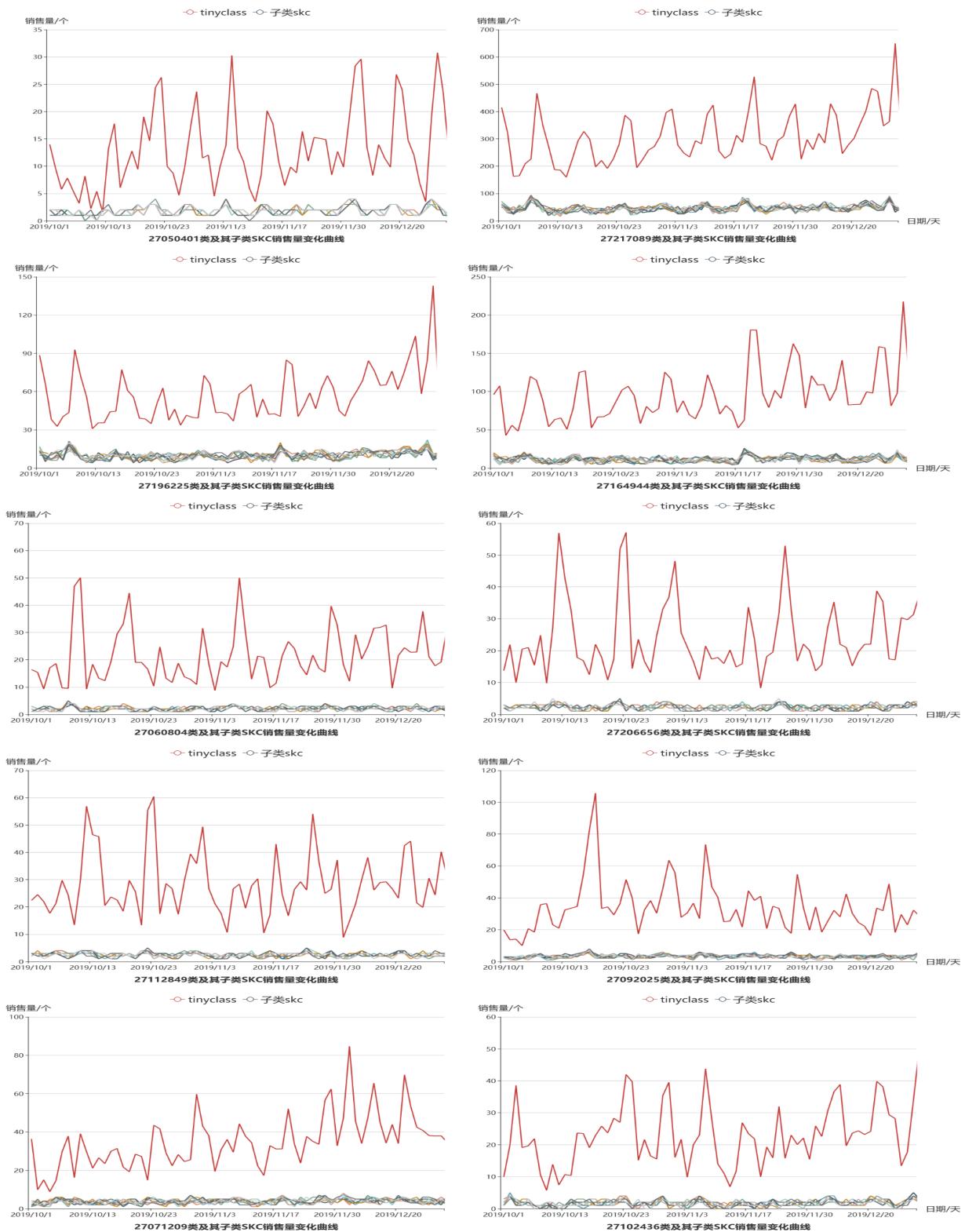


图 16: ARIMA+SVM 预测结果图

$$s_t = \gamma^*(y_t - l_t) + (1 - \gamma^*)s_{t-m} \quad (31)$$

如果用平滑方程式中的 l_t 代替上述分量形式的水平，则得到

$$s_t = \gamma^*(1 - \alpha) + (y_t - l_{t-1} - b_{t-1}) + [1 - \gamma * (1 - \alpha)]s_{t-m} \quad (32)$$

与我们在此处指定的季节性分量的平滑方程相同，带有 $\gamma = \gamma^*(1 - \alpha)$ 。通常的参数限制是 $0 \leq \gamma^* \leq 1$ ，它转换为 $0 \leq \gamma \leq 1 - \alpha$ 。

8.2.2.2 Holt-Winters 预测销量

我们用 Holt-Winter 加入季节性、以及节日的周期，结果如图所示：

我们对预测结果的 MAPE 进行计算。我们选取能代表平均水平的某一个小类的 skc 呈现如下：

相比 ARIMA+SVM 模型的预测结果，Holt-Winter 的 MAPE 值减小了很多，可知其预测准确度相比 ARIMA+SVM 模型提高了很多。详细完整的预测结果和 MAPE 值见附件。

九. 问题四——给企业的一封信

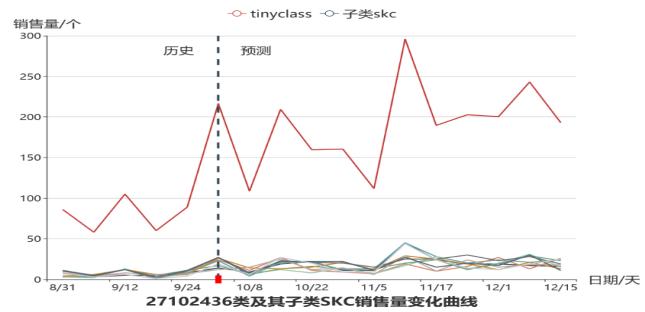
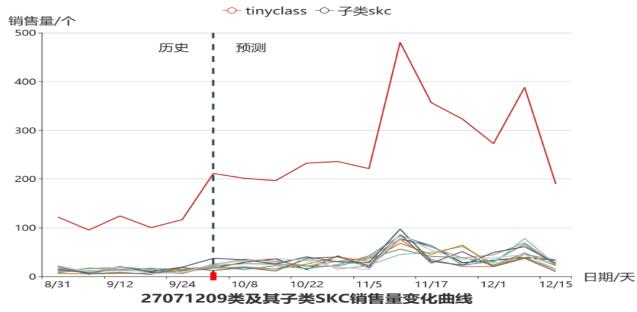
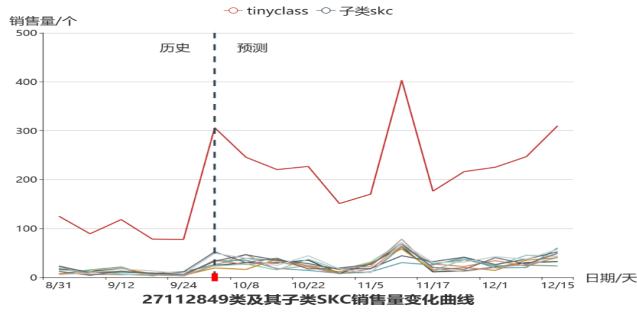
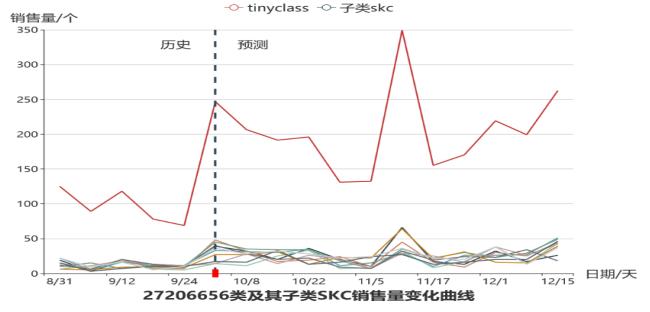
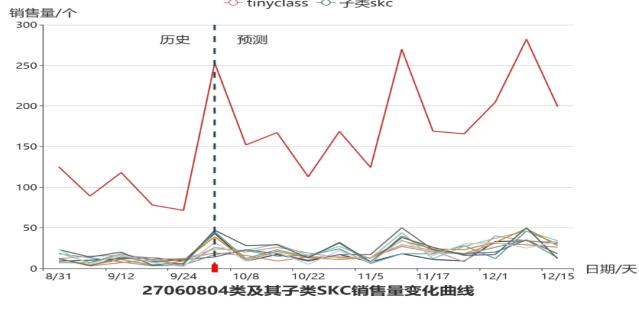
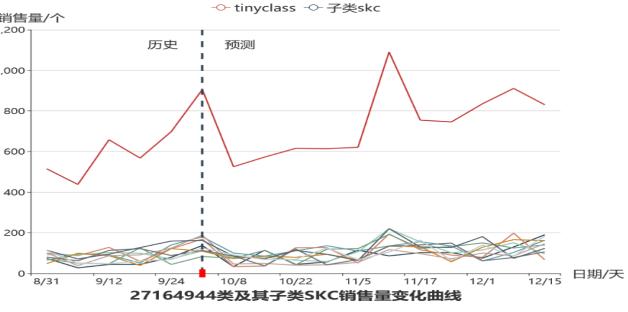
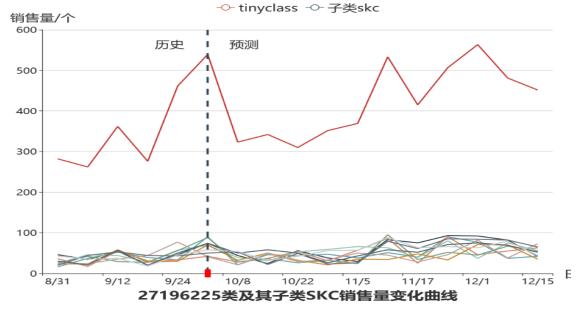
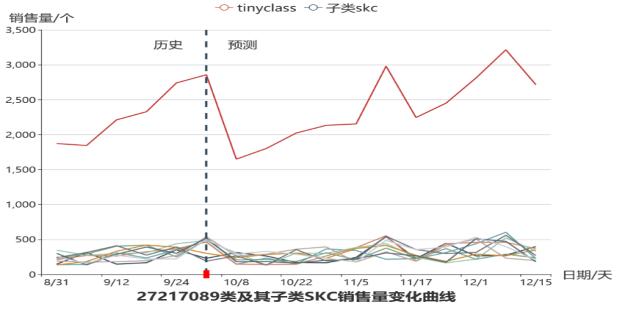
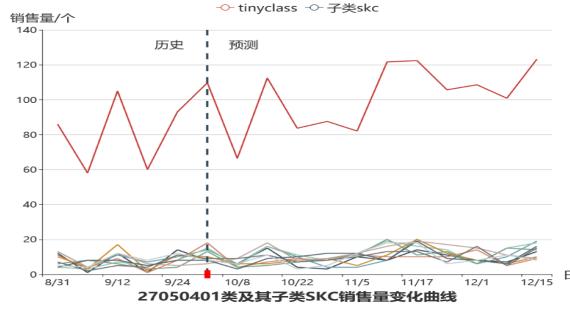
尊敬的企业经理：

您好。

我们是 Mathorcup 的建模团队。我们基于贵公司产品的销售数据建立了数学模型以预测产品的销量。我们的主要工作和结论如下：

首先，我们以在 2018 年 7 月 1 日至 2018 年 10 月 1 日内累计销售额排名前 50 的 skc 为研究对象，探究了双十一，双十二和元旦这四个节假日内的折扣率、实际价格、库存量、标签价、价格波动率、市场占有率为六个因素对销量的影响，其中，价格波动率我们定义为一年内产品的实际价格方差。我们选用随机森林算法计算出了各种因素分别在四个节假日内对销量影响的权重，并且还综合了四个节假日的数据来体现各一 i 尿素的综合影响。结果显示，当节假日的优惠力度较低的时候，折扣率对销量的影响因素最大，如双十一的折扣率的影响权重高达 0.9。同时，综合来看折扣率的权重同样是最高的，实际单价和市场占有率为对销量有着不可轻视的影响。

结合以上的结论，我们又以 2019 年 6 月 1 日至 2019 年 10 月 1 日内累计销售额排名前 10 的小类为研究对象，预测其在 2019 年 10 月 1 日后 3 个月中每个月的销售量。经过尝试，我们最后选用了 LSTM 神经网络算法来进行预测，它是时间序列模型最常用最强大的的工具，模型准确度高，训练速度快，并行处理能力强。在上一个研究的基础上，我们对该模型进行参数，从而使其得以在短期历史数据学习的基础上能够预知节日的冲击。我们对预测出来的结果计算平均绝对百分比误差 (MAPE) 以检验其预测的有效性和准确性，结果显示 MAPE 均在合理范围内，表明预测很成功，部分预测数据如下所示。为精准预测以指导营销，我们又探究以上目标小类内所有 skc 在 2019 年 10 月 1 日后 12 周内每周的周销量。经过探索，我们最终选用了 Holt-Winters 模型。由于我们用以学习和预测的历史数据中，四月、五月、六月和九月均存在一些节日对销量呈现一定周期性的影响，且影响的趋势相同，而运用 Holt-Winters



MAPE	第1周	第2周	第3周	第4周	第5周	第6周	第7周	第8周	第9周	第10周	第11周	第12周
999572122903	0.1854	0.3330	0.2380	0.1556	0.3855	0.2191	0.2635	0.3714	0.1884	0.1970	0.2713	0.2988
996572122716	0.2226	0.2424	0.2919	0.1079	0.2753	0.3926	0.2942	0.3225	0.1167	0.1501	0.1318	0.3075
996572122672	0.3823	0.3079	0.3759	0.2605	0.2037	0.3468	0.3368	0.3568	0.3311	0.2452	0.1013	0.2572
996572122650	0.1678	0.2183	0.2492	0.3406	0.3169	0.1784	0.1270	0.1261	0.2471	0.3040	0.3757	0.3524
996572122617	0.1397	0.2029	0.3268	0.3684	0.1391	0.1062	0.2126	0.2578	0.3588	0.3383	0.1150	0.2062

图 18: Holt-Winter 模型预测结果 MAPE 值

目标小类 月销售量	1号	2号	3号	4号	5号	6号	7号	8号	9号	10号
10月	10321	2530	1656	3422	762	620	1832	1024	933	653
11月	7896	1864	1495	2983	624	645	1656	563	635	365
12月	8500	1698	1364	2684	564	563	1654	361	564	450

图 19:

能够更学习这些周期特征，故而能够更准确预测 10 月至 12 月的销量、反应在其中四个节假日的增减趋势。我们对预测结果进行 (MAPE) 检测，结果非常的理想，均小于 0.1，说明预测结果和实际值拟合得较好，表明我们预测模型的成功性。部分结果如下图所示。

以上就是我们所作的工作。总而言之，我们的模型能够做到精准预测小类以及小类下的各 skc 在

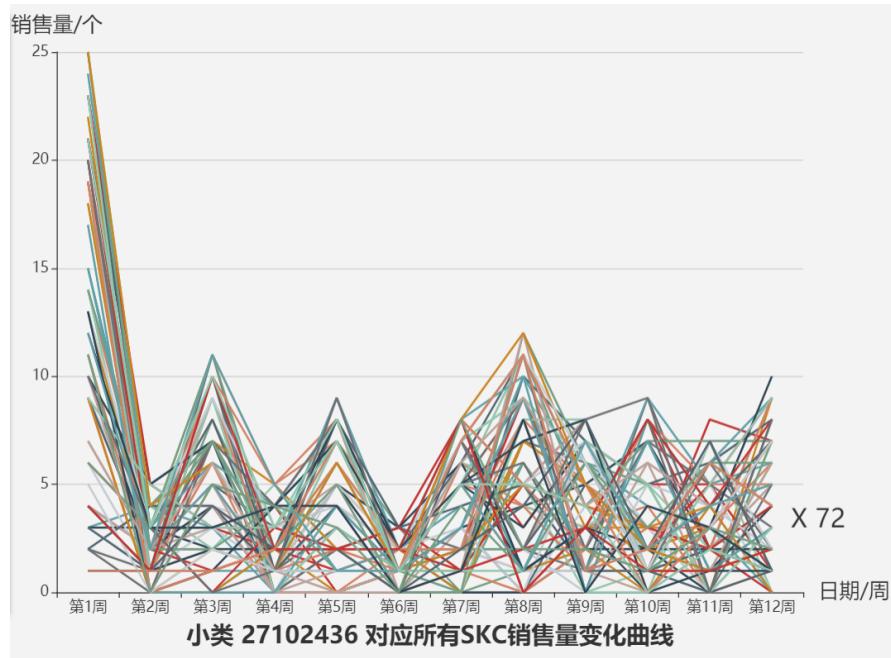


图 20: Holt-Winters 销量预测图

短期内的销量，从而能够很好的指导营销和生产，为大数据时代下新零售的发展作出贡献。但是我们的模型也存在一定的问题，例如没有对不同节假日对销量的影响做进一步区分，如国庆之类的传统节日和双十一等电商促销节日，以及偏重线上促销的节日和线下促销的节日区别。这些问题有待进一步完善。希望

我们的工作能够对贵公司的生产营销有所帮助。

orcup

十. 模型的优缺点

优点：

- 灰色预测可以处理不确定量，使之量化。可以充分利用已知信息寻求系统的运动规律，且能处理贫信息系统。
- 随机森林模型的学习速度是非常迅速的，能运用在检测普通基准误差模型上且在产生一定误差时，可以继续根据内部调整完成不偏差的结果输出。
- 模型十分简单，只需要内生变量而不需要借助其他外生变量。
- SVM 模型的提出解决了灰色系统误差较大问题，避免了神经网络存在的模型结构难确定、精度难保证及泛化能力不高的缺点，为非线性滑坡预测模型在寻优过程中易陷入局部最优的情况提供了新的思路。

缺点：

- 我们忽略了经济大环境如物价水平等对销量的影响
- ARIMA 模型本质上只能捕捉线性关系，而不能捕捉非线性关系。且预测时序数据，必须是稳定的，如果不稳定的数据，是无法捕捉到规律的。
- 随机森林模型在某些误差较大，识别不明显的分类或者问题上会产生过拟行为。

十一. 总结

面对新零售企业的产品销售量预测问题，我们对大量历史数据进行处理，化大样本数据为小样本数据，选取典型的目标小类与 skc 进行研究；采用了基本的算法 Pearson 相关系数、灰色预测；综合考虑数据的特性，利用 ARIMA 时间序列 +SVM 模型的组合算法；尝试了现在流行的机器学习算法——LSTM 神经网络，通过不同算法的预测结果对比，得出预测精确度较高的销量值；探究了新零售企业在节假日、营销日里的销量冲击，对于企业合理高效的进行仓库物品的调配，追求资源最大化、利益最大化作出了我们的贡献。我们从中也感受到数学建模给生活、生产带来的重大意义，让我们更加沉浸在数学建模的热爱之中，在学习中联系实践活动，用建模思维解决实际问题！

十. 参考文献

- [1] Xiao Han. Method and apparatus for learning-enhanced atlas-based auto-segmentation. 2015.
- [2] Manish Kumar and M. Thenmozhi. Forecasting stock index returns using arima-svm, arima-ann, and arima-random forest hybrid models. International Journal of Banking Accounting & Finance, 5(3):284, 2014.
- [3] 刘俊娥, 慕梓咛, and 刘丙午. 固有模态 svm 预测模型在零售销量预测中的应用. 物流技术, 032(011):76–78,97, 2013.
- [4] 刘胜. 基于 ARIMA 与 SVM 组合模型的国内旅游市场预测研究. PhD thesis, 东华理工大学, 2017.
- [5] 古再努尔·伊斯马伊力. 实体连锁零售企业的新零售模式转型. 全国流通经济, (21):4–5, 2017.
- [6] 吴广伟. 基于 ls-svm 的非线性预测控制研究. 2008.
- [7] 朱紫玉. 新零售模式下零售企业的战略转型研究. 北方经贸, 000(8):32–33, 2017.
- [8] 杨华龙, 刘金霞, and 郑斌. 灰色预测 gm(1,1) 模型的改进及应用. 数学的实践与认识, (23):41–48, 2011.
- [9] 汪鹏, 彭颖, and 杨小兵. Arima 模型与 holt-winters 指数平滑模型在武汉市流感样病例预测中的应用. 现代预防医学, (3), 2018.
- [10] 王佳敏 and 张红燕. 基于 arima-svm 组合模型的移动通信用户数预测. 计算机时代, 000(9):12–15,17, 2014.
- [11] 王庆荣. 基于神经网络与 holt-winters 模型的铁路货运量组合预测. 兰州交通大学学报, (4):128–131, 2010.
- [12] 胡彦君. Arima 模型在汽车销量预测中的应用及 sas 实现. 河北企业, 000(4):11–12, 2012.
- [13] 胡挺峰 and 钱雪忠. 基于 matlab 的多变量灰色模型在商业企业销售量预测中的应用. 信息与电脑: 理论版, 000(008):P.136–137, 2009.
- [14] 陈涛. 基于 de-svm 非线性组合预测模型的研究. 计算机工程与应用, 047(13):33–36, 2011.