Team#2024060928277

2024 年第三届“钉钉杯”大学生

大数据挑战赛论文

香烟销量和销售额预测——

题　　目　　**基于多模型集成学习方法**

摘　要

在现代商业环境中，准确预测产品的销售额和销量对于企业的库存管理、市场策略制定及资源分配至关重要。本文旨在通过使用多种时间序列预测模型，对某品牌香烟的销售额和销量进行精准预测，并通过集成学习方法优化这些预测结果。为此，本文选用了四种主要的时间序列预测模型：自回归积分滑动平均模型（ARIMA）、Prophet 模型、长短期记忆网络（LSTM）和极端梯度提升模型（XGBoost）。每种模型从不同角度对时间序列数据进行建模，以捕捉其内在规律和特性。

首先，我们对香烟品牌 A1,A2,A3 和 A4 的历史销售数据进行了预处理，处理了异常值和缺失值，以确保数据质量。接着，分别应用 ARIMA 和 LSTM 模型对这些品牌的销售数据进行了预测，并对预测结果进行了性能评估。随后，我们使用 ARIMA、Prophet、LSTM、XGBoost 模型对香烟品牌 A5 的销售额和销量进行了预测，以进一步验证模型的适用性和预测能力。

为了提升预测精度，我们将上述四种模型的预测结果进行了集成。具体而言，采用线性回归模型作为元学习器，对各基础模型的预测结果进行组合，从而生成最终的预测结果。实验结果表明，集成学习模型在预测精度上显著优于任何单一模型，能够更有效地捕捉销售数据的复杂特性。

本文还展示了详细的预测结果及其可视化图表，提供了对各模型预测性能的深入分析。通过比较和集成多种时间序列预测模型，本文显著提高了香烟品牌销售额和销量的预测精度，展示了其在实际应用中的高实用价值。最后，本文提出了未来研究方向，包括进一步优化模型参数和探索更多的集成学习方法，以继续提升预测性能。

关键词　机器学习、时间序列预测、ARIMA、Prophet、LSTM、XGBoost、集成学习、

目录

一、问题回顾

1.1 问题背景

在现代商业环境中，准确预测产品的销售额和销量对于企业的库存管理、市场策略制定及资源分配至关重要。随着市场竞争的加剧，企业需要依靠数据驱动的决策来保持竞争力和市场份额。特别是在快速消费品行业，如香烟市场，销售数据的波动性和复杂性使得准确预测变得更加具有挑战性。因此，选择合适的时间序列预测模型来提高预测精度，对于企业的经营决策至关重要。

1.2 问题回顾

在现代商业环境中，对未来销量和销售金额的精准预测对于企业的战略决,和资源优化至关重要。为了对香烟品牌 A1、A2 的未来销量和品牌 A3、A4 的销售金额进行有效预测，我们首先采用了不同类型的时间序列预测模型，以捕捉数据中的复杂模式和趋势。通过这些模型，我们不仅可以了解每种模型的优缺点，还能够通过综合运用多种模型来优化最终的预测结果。

对于香烟品牌 A1 和 A2 的销量预测，我们选用了自回归积分滑动平均模型（ARIMA）和长短期记忆网络（LSTM）。ARIMA 模型是一种经典的时间序列预测工具，特别适合于处理平稳或经差分处理后的非平稳时间序列数据。我们首先对历史销售数据进行了差分处理，以实现数据的平稳性，然后通过自相关图和偏自相关图确定 ARIMA 模型的参数（p, d, q）。在对模型进行拟合之后，我们在验证集上评估了预测性能，主要通过均方根误差（RMSE）和平均绝对误差（MAE）来衡量预测的准确性。

我们也对 LSTM 模型进行了设计和训练，调整了隐藏层单元数、学习率等参数，通过交叉验证来确定最佳参数组合。经过训练和验证后，我们利用 LSTM 模型对 A1 和 A2 品牌的未来销量进行了预测。

在集成学习预测方面，我们还添加了了 Prophet 模型和极端梯度提升（XGBoost）模型。

1.3 相关文献

近年来，随着大数据和机器学习技术的发展，越来越多的研究者开始关注集成学习算法在各领域中的应用。

很多作者探讨了基于 SARIMAX 和 LSTM 模型的日照港货物吞吐量预测方法(徐浩帆等人，2024），通过细致调整模型超参数和引入贝叶斯优化，有效提高了预测准确度。有学者则研究了基于 Prophet 模型的民航商务旅客出行量预测（鲍斌等，2024），展示了该模型在预测精准度和可解释性方面的优势。学者将 PSO-Prophet 模型应用于农产品价格预测（刘合兵等，2024），通过融合消费者物价指数等影响因素，显著提高了预测精度。其他学者提出了一种结合序列分解和 Prophet 模型的时序预测方法（丁美荣和张迎春，2023），通过这种方法有效提升了模型的

整体预测精度和训练效率。有些研究者则通过改进型 SVR-SARIMAX 混合模型预测挖掘机销量（秦秋洪等，2022），展示了在考虑周期性和特殊因素的情况下，该模型相较传统方法有更优的预测性能。这些研究丰富了时间序列预测的理论，为不同行业提供了实用的预测工具，为未来的预测模型设计和应用提供了有益的启示。

此外，部分学者基于 XGBoost 算法对跨境电商备货进行预测（李融，2024），通过对比实验证明，与传统的线性回归、支持向量机等方法相比，XGBoost 算法具有更高的预测准确率。这一研究为跨境电商企业合理制定备货策略提供了有力支持。将时间序列和决策树模型应用于线上酒店销量预测中。虽然该研究并未直接使用 XGBoost 算法（路标，2023），但其提出的模型融合思路为后续研究提供了有益启示。我们可以借鉴这种思路，将 XGBoost 算法与其他模型相结合，进一步提高销量预测的准确性。有的学者则关注了新冠病毒肺炎疫情影响下的润滑油销量预测路（卢亚茹，2023）。该研究基于 XGBoost 算法构建了润滑油销量预测模型，并分析了疫情对润滑油销量的影响。这一研究不仅为润滑油企业提供了决策依据，也为我们展示了 XGBoost 算法在应对突发事件中的潜力。有人对集成学习算法在新能源汽车市场中的应用进行了分析和预测（刘凯迪，2022）。虽然该研究并未直接涉及 XGBoost 算法，但其对集成学习算法的讨论为我们理解 XGBoost 算法在新能源汽车市场中的表现提供了背景知识。个别学者的研究则将 LSTM 网络与 XGBoost 算法相结合（王细雨，2022），构建了电商商品短期销量预测模型。该研究通过实验证明，这种结合方法能够有效地提高销量预测的准确性。这一研究为我们展示了 XGBoost 算法与其他机器学习算法相结合的巨大潜力。

综上所述，SARIMAX，Prophet，LSTM，XGBoost 等模型在商业分析，销量预测，物流管理中的重要性日益增加。我们将综合使用 SARIMAX，Prophet，LSTM，XGBoost 等模型，实现对于销量和销售金额的预测。

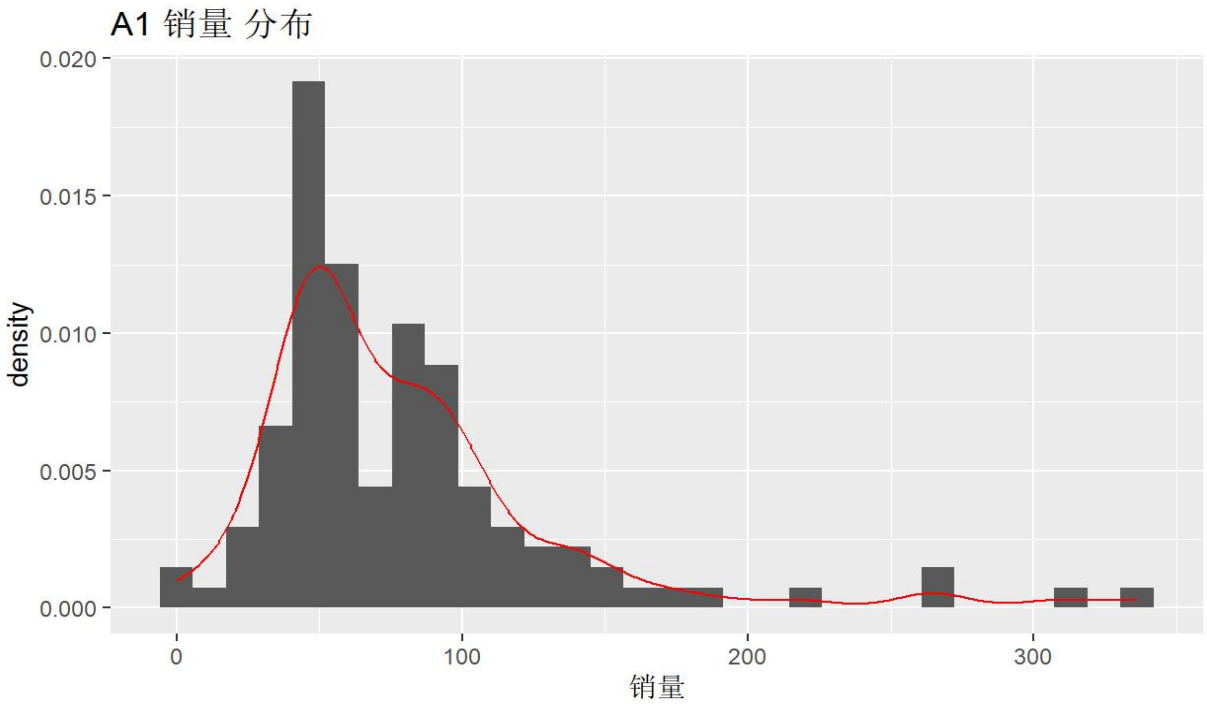
二、数据清洗与可视化

数据清洗是进行数据分析的前提。可视化的方式可以方便地查看时间序列的变化和数据的分布，为接下来的清洗和建模选择提供方向。通过对于数据的统计摘要和分布的查示，可以验证数据是否服从正态分布，修正数据分布，剔除极端值，避免影响预测。对于时间序列模型，采用稳定性检验、白噪声检验等方法，可以进一步确定模型选择。

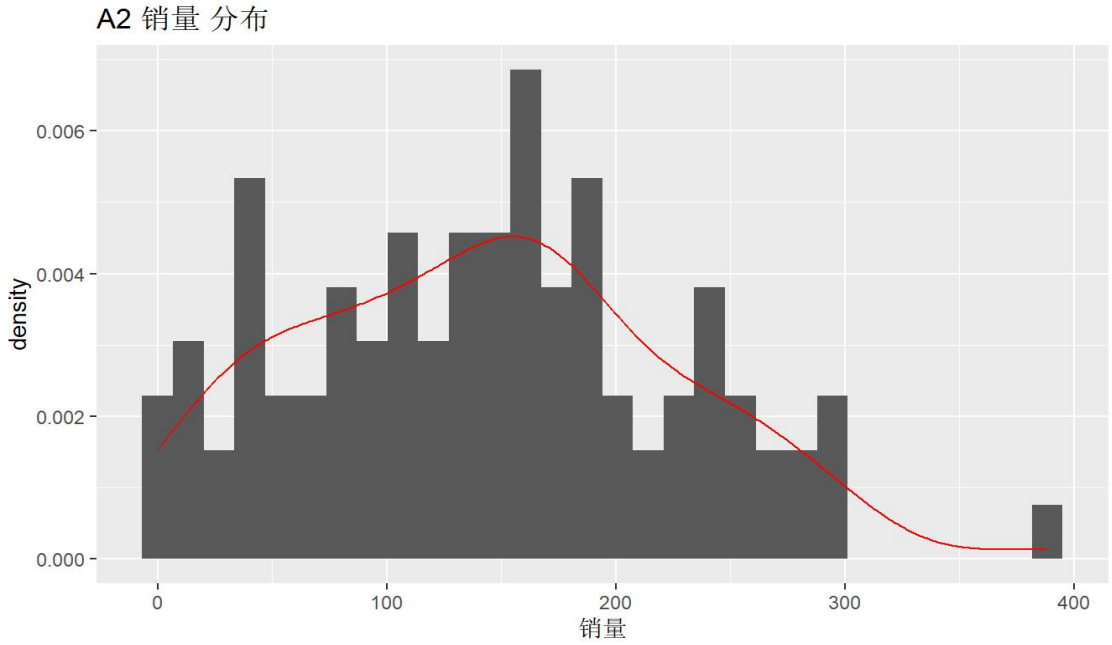
2.1 数据可视化

2.1.1 数据分布可视化

绘制出销量数据和金额数据的频数分布直方图和密度曲线，可以大致判定数据的分布规律，进而判定数据是否满足正态分布。示例如下：



图：A1 销量数据频数分布直方图和密度曲线



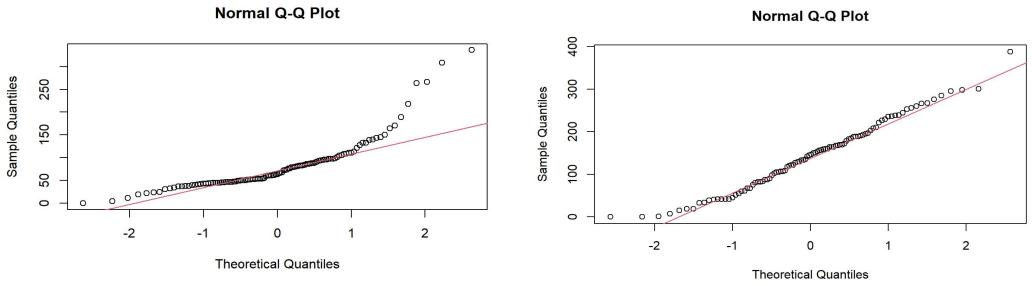
图：A2 销量数据频数分布直方图和密度曲线

其中，频数分布直方图的横坐标为连续变量的取值区间，纵坐标为对应区间频数，密度曲线的横坐标为连续变量，纵坐标为概率密度。二者结合使用便于衡量数据的分布形状，尤其是是否有偏，是否存在极端值等信息。

从上图中，可以发现 A1 销量数据存在明显的右偏，即数据分布的尾部在右侧延伸得比左侧更长，这表示在数据中存在一些比大多数观测值更大的值；与此相比，数据二分布更为对称。

正态分布对于数据的极端值判断和剔除乃至后续建模的选择都具有重要意义，由于正态分布具有对称性，可以认为 A1 销量数据不太可能满足正态分布，而 A2 销量数据则可能满足。

进一步，使用 Quantile-Quantile plot 判断数据是否满足正态分布，其横轴为标准正态分布的理论分位数$\phi^{-1}(p)$，即标准正态分布函数在概率 p 处的反函数。纵轴为第 i 个样本的分位数，$\frac{i-0.5}{n}$，其中 n 为样本数。通过观察 QQ 图中的点的分布模式来判断数据集是否符合理论分布的假设。如果数据集与正态分布一致，QQ 图中的点将近似落在一条直线上偏离直线的情况可能表明数据不符合正态分布，可能存在偏斜或其他非正态特征。示例如下：



图：A1 和 A2 销量数据 QQ 图

如图，A1 的销量数据在右端距离直线有明显偏离，说明不呈现正态分布，而 A2 的销量基本沿直线分布，说明可能呈现正态分布。

2.1.2 正态分布检验

在数据可视化后，使用统计检验的方式进一步验证数据的正态性。Shapiro-Wilk normality test 是常用的正态统计检验方式，其方式为将样本从小到大排序为$x_1, x_2, \ldots\ldots, x_n$,然后计算出 W 统计量，表达式为：

$$\frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

如果 W 统计量小于显著性水平对应的临界值，则不拒绝原假设，认为数据来自于正态总体，也可以通过比较 p 值和显著性水平来进行判断，如果 p 值小于显著性水平，则认为数据来自于正态总体。取显著性水平为 0.05，判断 A1、A2 销量数据，A3、A4 销售金额数据，以及 A5 销量及销售金额数据的显著性，结果如下：

数据来源	Shapiro-Wilk 检验 p 值	是否拒绝原假设
A1 销量数据	<0.01	拒绝
A2 销量数据	0.1687	不拒绝
A3 销售金额数据	<0.01	拒绝
A4 销售金额数据	0.3628	不拒绝
A5 销量数据	<0.01	拒绝
A5 销售金额数据	<0.01	拒绝

2.1.3 数据清洗

查示数据可以发现，存在极端值，在预测前应当予以剔除和修正，避免影响后续预测的结果。对于正态分布的数据，可以采用3σ原则识别异常值，即大于$\mu+3\sigma$或小于$\mu-3\sigma$的数据为异常数据，其中μ和σ分别为数据的均值和标准差。对于非正态分布的数据，使用 boxcox 函数使其正态化。boxcox 函数如下:

$$y(\lambda)=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda} & x>0\\ ln(y) & x=0.\end{cases}$$

对于非正态数据，可以寻找到合适的λ，用于将非正态数据 y 转换为正态数据。使用程序进行遍历，并根据 Shapiro-Wilk 统计值 W 选择最优参数，将非正态数据正态化后，进行异常值筛选，剔除。此外，通过查示数据，发现存在部分月份缺失。对于剔除数据和缺失数据，使用线性插值法进行填充，原理如下：

$$y=\frac{y_2-y_1}{x_2-x_1}(x-x_1)$$

使用线性插值，可以通过缺失数据两端的数据，对其进行预测。这种方法可以处理少量缺失数据。但是通过查示发现，A5 数据的缺失较为严重（缺失从 2018 年 9 月至 2022 年 12 月的数据），不适合使用线性插值的方式，因此这一部分使用集成学习的方式进行填补，可以较好还原数据的波动特性。

3.1.4 时间序列检验

对于时间序列数据，需要进行稳定性检验，用于判定是否存在趋势（trend）、白噪声检验，用于判定序列是否为纯随机波动、季节性检验，用于判断是否存在明显的季节性。使用 Augmented Dickey-Fuller 检验（ADF 检验）用于判定时间序列数据是否具有单位根，从而判定

稳定性，其核心是通过构建一个回归模型来检验序列的单位根。通常考虑的模型是一个自回归过程（AR），其中包含序列的滞后项。ADF 检验的一般形式如下：

$$\Delta y_t = a + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} \delta y_{t-1} + \epsilon_t$$

其中，Δy_t 是一阶差分（即当前值与上一个值的差），t 是时间趋势项，a 是截距，β 是时间趋势的系数，γ 是序列的滞后项系数，δ_i 是差分项的系数，ϵ_t 是白噪声误差项。检验滞后系数 γ 是否显著来判断序列是否平稳，通过比较检验 p 值与显著性水平实现。如果 p 值小于显著性水平，则拒绝原假设，认为序列是平稳的。

使用 Box-Pierce 检验用于判断序列是否为白噪声序列。白噪声序列不存在自相关，序列的波动由随机性带来，因此研究价值低。Box-Pierce 检验的原理如下：

$$bQ = n \sum_{k=1}^{\mathrm{h}} \hat{\rho}_k^2$$

其中，$\hat{\rho}_k^2$ 是样本自相关系数，n 是时间序列长度，h 是滞后阶数。Q 的显著大于临界值，或 q 小于显著性水平，则拒绝原假设，认为序列存在自相关，就可以拒绝原假设，认为时间序列存在自相关性，从而序列为非白噪声序列。检验结果如下：

表 时间序列检验

	ADF 检验 p 值	Box-Pierce 检验 p 值
A1 销量数据	0.1108	<0.01
A2 销量数据	0.01	<0.01
A3 销售金额数据	0.2394	0.0295
A4 销售金额数据	0.01	<0.01
A5 销量数据	0.2242	<0.01
A5 销售金额数据	0.2803	<0.01

从上述结果可以判断，A2 销量数据和 A4 销售金额数据可以视为平稳序列，其余数据均均存在趋势；此外，所有的数据都不是白噪声序列，具有分析的价值。

三、模型的建立与求解

3.1 问题一模型和问题二模型的建立与求解

3.1.1 A1、A2 和 A3、A4 品牌品牌销量预测模型的构建与求解

1.自回归积分滑动平均模型（ARIMA）

我们使用自回归积分滑动平均模型（ARIMA），它是一种广泛应用于时间序列预测的统计模型，它结合了自回归（AR）、差分（I）和滑动平均（MA）三个部分。ARIMA 模型的设计目的是通过捕捉时间序列数据中的线性关系来进行未来值的预测。下面将详细介绍 ARIMA 模型的数学分析和公式。

（1）ARIMA 模型概述

ARIMA 模型通过将时间序列数据分解为三个部分来进行建模：自回归部分（AR）、差分部分（I）和滑动平均部分（MA）。其基本形式可以表示为 ARIMA(p, d, q)模型，其中：

p：自回归部分的阶数，表示模型中包含的滞后值的数量。

d：差分的阶数，表示数据在建模之前需要进行多少次差分以实现平稳性。

q：滑动平均部分的阶数，表示模型中包含的滞后预测误差项的数量。

（2）自回归（AR）部分

自回归部分表示时间序列值与其自身滞后值之间的关系。AR(p)模型的数学表达式为：

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \epsilon_t$$

其中，$\phi_1,\phi_2.....\phi_t$是自回归系数，ϵ_t 是白噪声（即随机误差项）。

（3）差分（I）部分

差分部分用于将非平稳时间序列转化为平稳序列。通过对时间序列数据进行差分操作，可以消除趋势和季节性影响。d 阶差分的定义为：

$$\Delta^d X_t = (1-B)^d X_t$$

其中，B 是滞后算子，定义为 $B^k X_t = X_{t-k}$例如，1 阶差分可以表示为：

$$\Delta X_t = X_t - X_{t-1}$$

（4）滑动平均（MA）部分

滑动平均部分表示时间序列值与过去的预测误差项之间的关系。MA(q)模型的数学表达式为：

$$X_t = \mu + \epsilon_t + \theta_1 X_{t-1} + \theta_2 X_{t-2} + ... + \theta_q X_{t-q}$$

其中，$\theta_1,\theta_2...\theta_q$是滑动平均系数，$\mu$是常数项，$\epsilon_t$是白噪声。

（5）ARIMA 模型的综合表达

将自回归、差分和滑动平均部分结合起来，ARIMA(p, d, q)模型的数学公式为：

$$\Phi(B)(1-B)^d X_t = \Theta(B)\epsilon_t$$

其中：

$\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p$是自回归部分的多项式。

$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 - ... + \theta_q B^q$是滑动平均部分的多项式。

2.SARIMA（季节性自回归积分滑动平均）

SARIMA（季节性自回归积分滑动平均）模型扩展了 ARIMA 模型，以便处理季节性数据。其数学表达式可以表示为：

$$(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t$$

这些部分协同工作，使 SARIMA 模型能够捕捉时间序列中的趋势、周期性和季节性特征，从而成为时间序列预测的一个强大工具。最终，模型的预测结果如下所示。

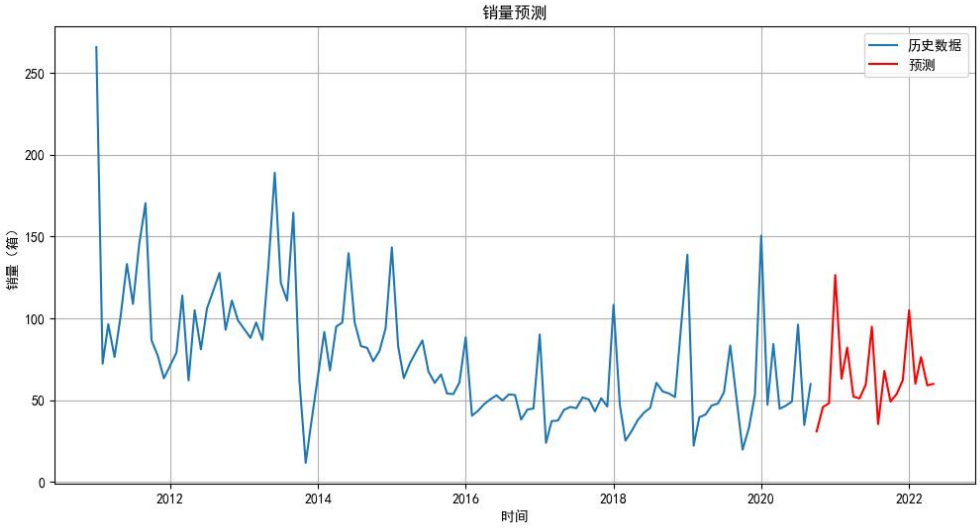


图 A1 预测数据

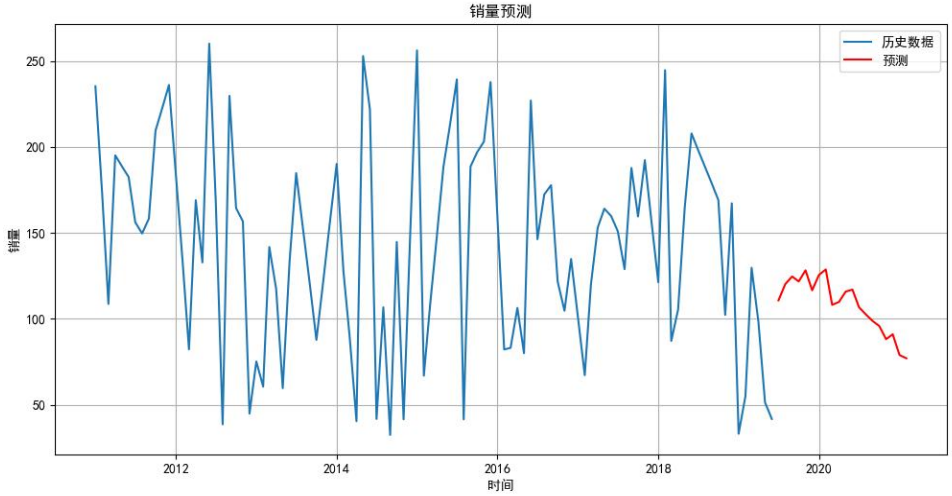


图 A2 预测数据

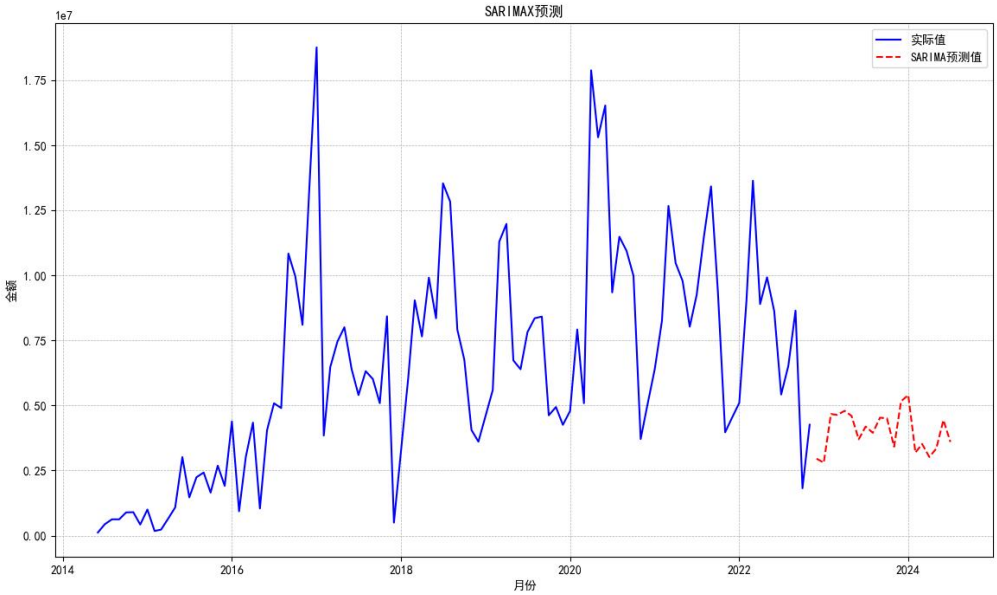


图 A3 预测数据

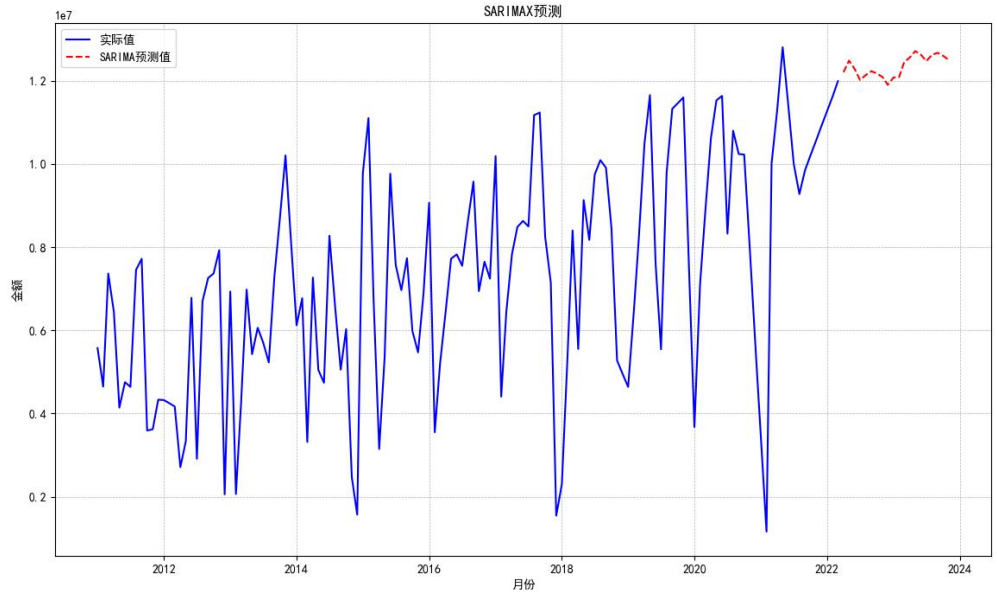


图 A4 预测数据

3.LSTM（长短期记忆网络）

我们将“销量”数据归一化到 0 和 1 之间，以适应 LSTM 模型的输入要求。具体而言，我们使用一个函数来创建数据集，其中 X 为前 12 个月的销售数据，y 为第 13 个月的销售数据。LSTM 单元由遗忘门、输入门和输出门组成。以下公式描述了这些门的工作原理：

遗忘门（Forget Gate）：决定需要遗忘的信息。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

输入门（Input Gate）：决定更新哪些信息。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

记忆单元状态更新：综合遗忘门和输入门的信息，更新记忆单元状态。

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

输出门（Output Gate）：决定输出哪些信息。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \cdot tanh(C_t)$$

其中，各个符号符号的含义为：

符号	含义
f_t	遗忘门的输出
i_t	输入门的输出
o_t	输出门的输出
C_t	当前记忆单元状态
$\tilde{C}_t$	候选记忆单元状态
h_t	当前时刻的隐藏状态
h_{t-1}	前一时刻的隐藏状态
x_t	当前输入
σ	sigmoid 激活函数
$tanh$	双曲正切激活函数
W_f, W_i, W_C, W_o	权重矩阵
b_f, b_i, b_C, b_o	偏置项

我们构建了一个神经网络模型，该模型包括两层 LSTM 和一层全连接层。我们使用前 80%的数据进行训练，并将后 20%的数据用于测试。在训练完成后，我们利用最后 12 个月的数据来预测未来 20 个月的销售量。为了实现连续预测，我们逐步将每个预测值作为下一个时间步的输入。最后，我们将预测结果与实际销售量进行比较，并绘制结果图表以展示模型的预测效果。具体如下所示。

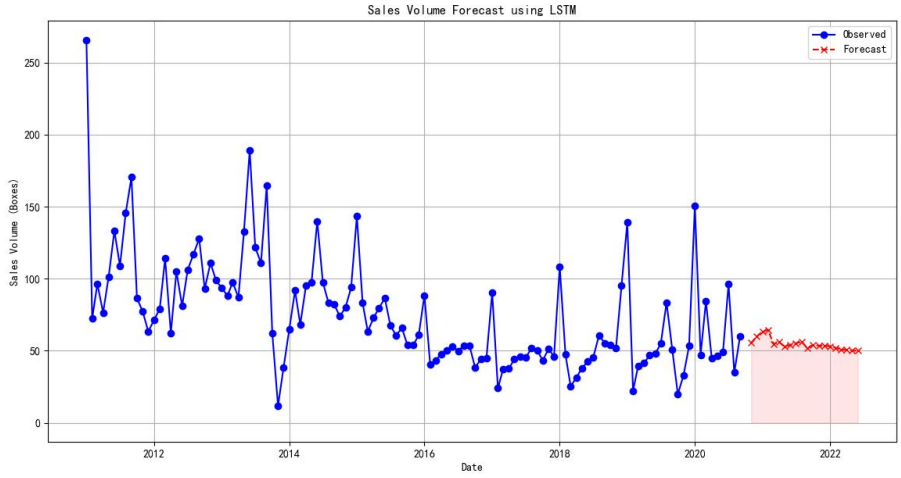


图 A1 预测数据

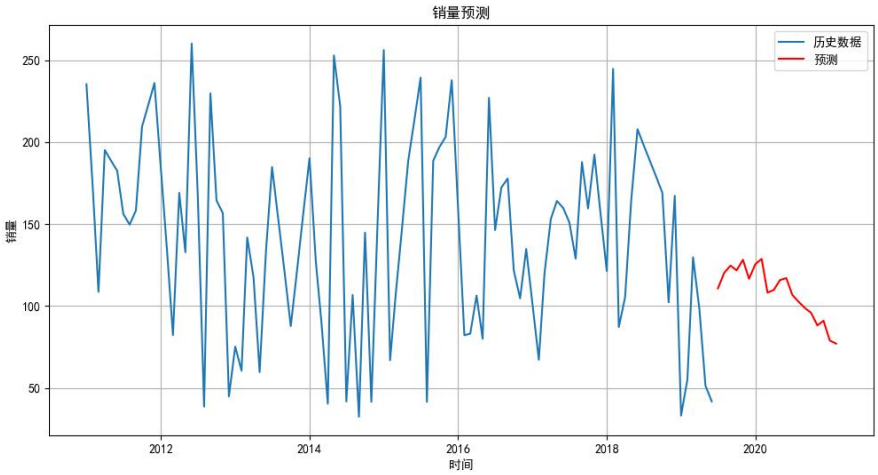


图 A2 预测数据

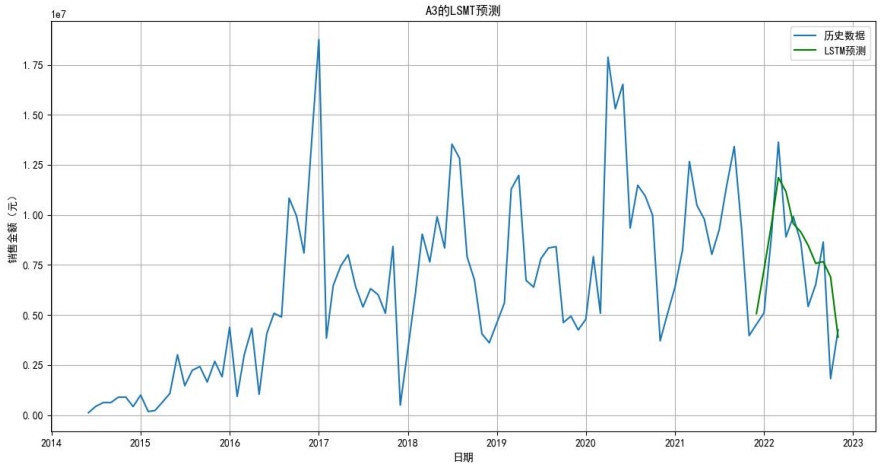


图 A3 预测数据

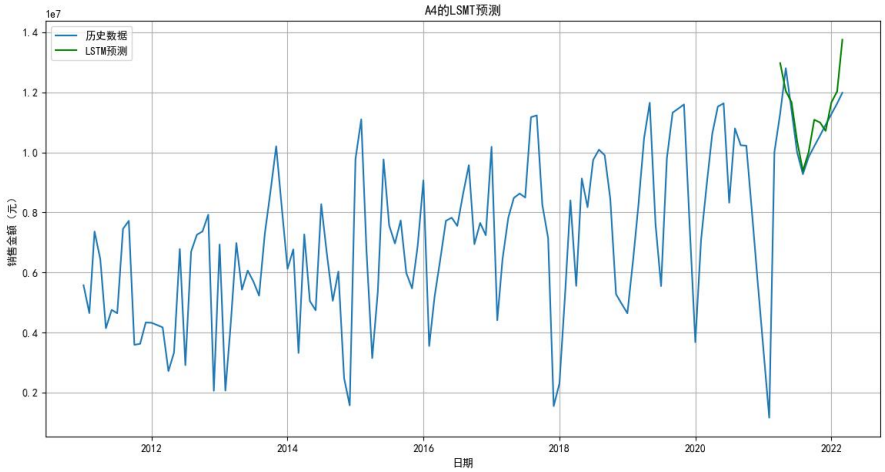


图 A4 预测数据

表 A1-A4 预测结果

A（绿和）销量数据	A(硬)	A(硬蓝爱你)	A(长征)
35.006975	174.5246	10006450	9301122
42.570354	143.5025	9621563	10247880
43.325461	74.77258	10874140	5777941
41.422282	156.1129	9256566	7472386
44.373272	107.0074	8521695	5899391
53.362573	133.2612	9028344	5887891

58.974787	107.9793	9041881	7490138
62.148507	171.1375	8982472	8998165
59.427934	141.1125	8269241	11703650
62.254705	141.7722	7765531	10715730
63.584819			11856750
61.775652			11237760
65.606167			
60.094284			
63.153559			
55.889744			
67.935638			
6.61E+01			
65.189952			
55.982875			

图：LSTM 的预测结果

4.3 问题三模型的建立与求解

4.3.1 各种模型预测

ARIMA 模型:遍历不同的参数，找到最好的 ARIMA 模型和参数，然后使用最佳参数拟合模型，并进行预测。

Prophet 模型:通过将趋势部分（Trend） 、季节性部分（Seasonality）和节假日部分（Holidays）相加，再加上噪声部分（Error） ，来构建完整的时间序列预测模型。公式可以表示为：

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t 。$$

这种分解方法使得模型既可以捕捉到长期的趋势变化，也能适应短期的周期性波动和特殊事件的影响，从而提供准确且灵活的时间序列预测。

LSTM 模型:使用 Keras 库构建和训练 LSTM 模型，再根据历史数据进行未来的预测。

XGBoost 模型:

xgboost（eXtreme Gradient Boosting）是一种高效的梯度提升算法，它通过集成多个决策树来构建一个强大的预测模型。在时间序列分析中，xgboost 可以用来预测未来的数值趋势，例如在金融、天气预报、股票市场等领域的应用。

xgboost 模型的目标是最小化一个目标函数，该函数包括损失函数和正则化项。损失函数衡量模型与数据的契合程度，而正则化项则控制模型的复杂度，以防止过拟合。具体来说，xgboost 的目标函数可以表示为：

$$obj(t) = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{t-1} + f_t(x_i)] + \frac{1}{2} h_i f_t^2(x_i) + \Omega(f_t)$$

其中，l 是损失函数，y_i 是真实值$\hat{y}_{t-1}^{i}$是模型的前一步预测值，$f_t(x_i)$是新添加的决策树，g_i和 h_i分别是损失函数关于预测值的一阶和二阶导数，$\Omega(f_t)$ 是正则化项。

为了找到最优的决策树 f_t，xgboost 采用贪婪算法遍历所有可能的树结构，选择使目标函数最小化的结构。每棵树的结构由其叶子节点的划分和对应的预测值 w 决定。对于每个节点，计算它的得分，并据此确定最佳的切分点。树的复杂度由叶子节点的数量T和节点权重的L_2范数决定。

在构建时间序列预测模型时，我们采用了滑动窗口技术来准备适合监督学习的数据集。具体操作是，选取连续的 12 个月销售金额作为特征输入，紧随其后的第 13 个月销售金额作为预测目标。这一步骤将时间序列问题转化为了监督学习框架。为了提高模型的训练效率和效果，我们对特征数据和目标值进行了标准化处理，使用 StandardScaler 来消除不同量级带来的影响，确保所有特征都在相同的尺度上参与模型训练。

随后，我们利用预先训练好的 XGBoost 模型对处理过的数据进行预测。完成预测后，我们将得到的标准化预测结果逆转换到原始尺度，使其能够与实际的销售金额直接对比。

为了直观展示模型的预测性能，我们使用 matplotlib 库绘制了包含实际销售金额和预测销售金额的对比图。这些图表帮助我们形象地看到模型预测的准确性和趋势。

1.在集成学习阶段：

集成学习技术通过整合多个预测模型的输出，并采用线性回归作为最终的决策模型，以此提升预测的准确性。在集成学习的 Stacking 方法中，它通过汇集多个基础预测模型的预测值来构建一个高层模型，旨在进一步提高预测的整体性能。在本案例中，我们运用了四种不同的时间序列预测模型——ARIMA、Prophet、LSTM 和 XGBoost——来独立预测销量和销售额。接着，将这些模型的预测结果作为输入特征，使用线性回归模型作为顶层模型，从而产生最终的预测结果。

2.在模型训练阶段：

ARIMA 模型：运用自回归积分滑动平均模型（ARIMA）对时间序列数据进行训练，从而获得预测结果。

Prophet 模型：采用 Facebook Prophet 工具实施时间序列的预测分析。

LSTM 模型：通过应用长短期记忆网络（LSTM）来进行时间序列的预测任务。

XGBoost 模型：利用 XGBoost 回归方法对时间序列进行预测。基础模型预测结果生成：

3.训练各个模型，并针对每个模型产生未来的预测数据。构建训练数据集：

将各基础模型产出的预测结果汇集为一个新的特征集，其中每列特征表示一个基础模型的预测值。

4.元学习器训练：

以线性回归作为高层学习器，通过训练该线性回归模型来汇聚所有基础模型的预测数据，以得出最终的预测数值。

5.最终预测生成：

借助已训练的元学习器，融合所有基础模型的预测结果，以此来产生最终的预测值。公式表述

元学习器的训练公式，其核心在于如何通过基础模型的输出特征训练元学习器以生成最终的预测结果。这个过程可以被视为一个嵌套优化问题，外部循环优化元学习器参数，而内部循环则针对每个任务优化基础模型参数。

在时间序列预测的背景下，假设有 N 个基础模型，每个模型针对未来某时间段的预测可以表示为一个 T 维向量。设第 i 个基础模型的预测结果为$f_i = [f_{i1}, f_{i2}, \ldots, f_{iT}]$。所有基础模型的预测结果可以组合成一个$N \times T$维的特征矩阵 F，其中每一行对应一个基础模型的预测结果。

对于元学习器，一个线性回归模型可以通过最小化以下损失函数来训练参数：

$$L(w,b) = \sum_{t=1}^{T} [y_t - (w_T f_t + b)]^2 + \lambda ||w||^2$$

其中，y_t 是时刻t的真实值，w和b分别是线性回归模型的权重和偏置项，是$\lambda \| w \|^2$正则化项用于防止过拟合，λ是正则化系数。

通过求解上述损失函数的最小值问题，可以得到最优的 w 和 b，从而确定线性回归模型的最终形式。这一过程通常通过梯度下降等优化算法实现。最终，元学习器的输出可以表示为所有基础模型预测结果的加权和加上偏置项：

$$y^t = w_T f t + b$$

这种层级化的学习策略使得模型能够捕捉到不同模型预测结果之间的相关性，进而提升整体预测性能。

通过上述步骤，我们能够有效地整合来自多个基础模型的优势。

采用集成学习(Stacking)术来增强时间序列预测的精确度。最终的结果展示如下：

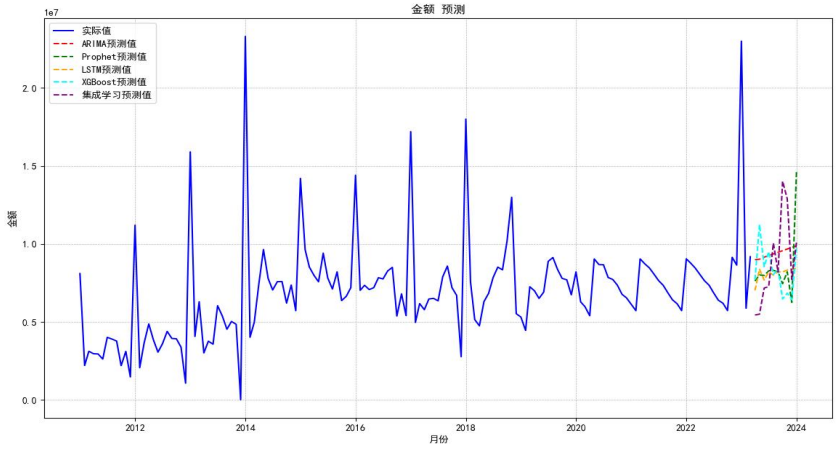


图 集成模型销量预测

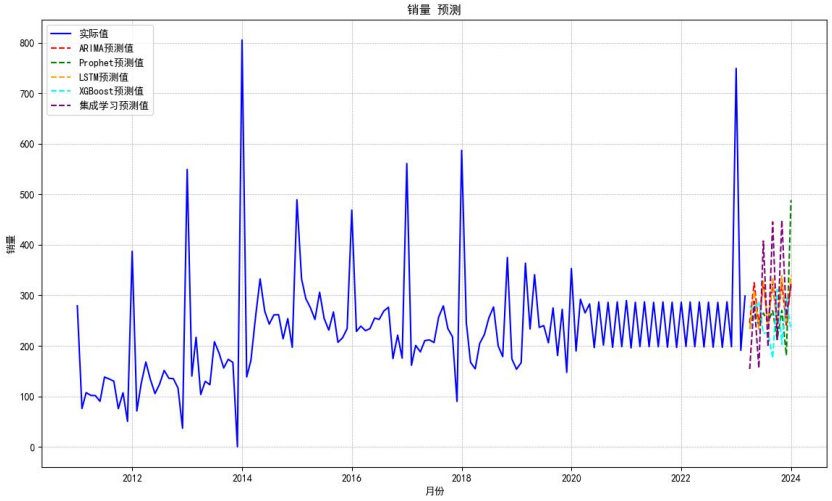


图 集成模型销售金额预测

在首张图表中，呈现了 A5 品牌销售金额的预测结果。实际的历史销售额由图中的蓝色实线表示。红色虚线、绿色虚线、橙色虚线以及青色虚线分别代表经由 ARIMA、Prophet、LSTM 和 XGBoost 模型所进行的预测。紫色虚线则展示了集成学习模型的预测结果。从图中可以观察到，进入 2022 年后，各个单一模型的预测呈现出差异，而集成学习模型则融合了这些模型的预测，呈现出一个综合的预测视图。该集成模型的预测显得更为稳定，并且在某种程度上融合了各独立模型的优点。

在第二张图中，呈现了品牌 A5 的销量预测情况。这里，蓝色实线依旧表示实际销量的历史数据。红色虚线、绿色虚线、橙色虚线和青色虚线分别展示了通过 ARIMA、Prophet、LSTM 和 XGBoost 模型得到的预测结果。紫色虚线是集成学习模型的预测展示。观察图表可以发现，在 2022 年之后，各模型给出的预测存在差异，例如 LSTM 模型的预测值相对偏低，而 XGBoost

模型的预测值则相对偏高。与此相比，集成学习模型汲取了这些单一模型的预测信息，并给出了一个更为平滑且更贴近历史走势的预测。这种集成方法使得学习模型在综合不同模型特点的基础上，进一步提升了预测的精确度和稳定性。A5 最终预测数据如下：

A(新版)销量数据	A(新版)销销售金额数据
251.5377	6959988
341.5732	4470901
167.1394	8234375
376.6811	6937412
197.905	9362375
458.7198	8981598
169.4864	11066279
427.1476	11073754
197.7125	10493851
300.7536	10767688

五、模型总结

5.1 模型优点

5.1.1 多模型融合增强预测精准度：

本文通过运用 ARIMA、Prophet、LSTM 和 XGBoost 等多种模型来进行销售预测，并借助集成学习（Stacking）方法将这些模型的预测结果进行加权融合，从而显著提升了预测的精确度和健壮性。各模型均具备其独特的优势，例如 ARIMA 模型极适于线性及季节性强的时间序列分析，Prophet 模型在处理季节性变化和节假日影响方面表现突出，LSTM 模型则在应对非线性挑战和长期依赖关系上有着出色的预测能力，而 XGBoost 模型在处理特征密集型数据时展现出卓越的性能。

5.1.2 高度适应性：

通过细致的模型参数调优和采用集成学习方法，本研究方案展现出对不同数据特性的强

大适应性。ARIMA 模型依靠遍历式的参数优化来寻找最佳组合，Prophet 模型能自动适应节假日及季节性的变化，LSTM 模型擅长捕捉复杂的时间依赖关系，XGBoost 模型则利用其决策树的优势实现高效率的预测。这种灵活的适应性确保了模型在多种实际应用环境中均能发挥出色的预测性能。

5.1.3 复杂时间序列数据处理能力：

在本研究中，LSTM 和 XGBoost 模型在处理复杂时间序列数据方面显示出特别的能力。LSTM 模型利用其递归神经网络的特性以捕捉长期的依赖关系，而 XGBoost 模型则通过其提升算法以及决策树集成的方法，增强了模型的泛化能力。这些特性使得模型能更有效地识别并预测数据中的复杂模式和趋势。

5.1.4 集成学习方法的优点：

采用集成学习方法能有效减少依赖单一模型可能带来的偏差和方差，显著提升预测结果的稳定性和可信度。通过使用线性回归作为元学习器来确定各个基学习器的权重，最终的集成模型能够充分利用各基学习器的独特优势，从而实现更准确的预测。

5.2 模型缺点

5.2.1 数据预处理环节具有较高的要求：

不同模型对数据的预处理需求各不相同，涉及到多次的数据转换、归一化以及差分处理等步骤，这无疑增加了数据处理的复杂性。同时，针对缺失值的处理方式在不同模型间也有所差异，需要在数据预处理阶段进行额外的操作和调整，以确保数据的一致性和模型的有效性得到保障。

5.2.2 模型调优的难度相对较大：

每个模型都有多个参数需要进行优化调整，特别是 ARIMA 模型的参数选择、LSTM 模型的结构设计和训练参数设置、XGBoost 模型的参数调整等，都需要大量的实验和验证工作，从而增加了模型调优的难度和工作量。此外，在集成学习方法中，权重参数的优化也需要通过线性回归来实现，进一步增加了调优的复杂性。

5.2.3 高度依赖历史数据的质量：

时间序列预测模型在很大程度上依赖于历史数据的质量和完整性。如果历史数据中存在较多的噪声、缺失值或异常值，将直接影响模型的训练效果和预测准确性。特别是对于 LSTM 和 XGBoost 模型来说，对数据质量的要求更为严格，因此在数据预处理过程中需要进行严格的质量控制和异常值处理。

5.2.4 对长时间趋势变化的适应性有限：

虽然 LSTM 和 XGBoost 模型能够处理复杂的时间依赖性，但在面对长期趋势变化较大或突发事件影响较强的时间序列时，模型的适应性仍然有限。特别是对于 ARIMA 和 Prophet 模型来说，在处理非线性和非平稳数据时，预测效果可能会有所下降，此时可能需要引入更多的外部特征或进行模型改进来提高预测效果。

5.3 模型推广

本文所探讨的 ARIMA、Prophet、LSTM 和 XGBoost 模型的集成方法，不仅在销售预测领域表现出色，还具有在其他领域广泛推广应用的潜力。这种多模型融合的策略能够显著提升预测精度和健壮性，是解决复杂数据问题的强大工具。

在金融市场预测方面，这些模型可以用来分析和预测股票价格、利率以及其他金融指标的走势。ARIMA 模型擅长处理线性和季节性数据，因此在预测长期趋势和周期性变化方面表现出色。Prophet 模型能够自动识别并适应节假日和季节性变化，对于金融市场中因节假日或季节性因素引起的波动具有重要作用。LSTM 模型则通过捕捉时间序列中的长期依赖关系，可以处理股票价格等金融数据中的非线性波动和复杂模式。XGBoost 模型在处理高维特征数据时具有卓越性能，可以有效应对金融数据中的多变量分析。

在电力负荷预测领域，这些模型同样具有重要应用。ARIMA 模型可用于分析电力消耗的周期性变化，Prophet 模型可以调整因节假日或特殊事件导致的用电量波动，LSTM 模型通过其记忆功能捕捉长期依赖关系，从而精准预测未来的电力需求，XGBoost 模型则能利用各类影响因素进行高效预测，帮助电力公司优化电力生产和分配，提高资源利用效率。

气象预报也是一个重要的应用领域。ARIMA 和 Prophet 模型能够处理气象数据中的季节性和趋势性变化，LSTM 模型擅长捕捉复杂的时间依赖关系，能够处理气候数据中的非线性特征，XGBoost 模型则能通过综合各种气象因素进行精确预测，从而为天气预报提供更加可靠的支持。

在医疗健康领域，这些模型可以用于患者健康指标的监测和预测。通过分析患者的历史健康数据，ARIMA 和 Prophet 模型可以预测长期健康趋势，LSTM 模型能够捕捉患者健康指标的非线性变化，XGBoost 模型则可以利用大量的健康数据特征进行综合分析，从而为医疗决策提供支持，提高诊断和治疗的精准度。

总的来说，ARIMA、Prophet、LSTM 和 XGBoost 模型的多模型融合方法在金融、能源、气

象、医疗等多个领域均具有广泛的应用前景。这些模型通过各自的优势互补，形成了一个强大而灵活的预测体系，能够处理各种复杂的时间序列数据，提升预测的准确性和稳定性，为各行各业的预测和决策提供有力支持。

六、结论

在本文中，通过运用 ARIMA、Prophet、LSTM 和 XGBoost 等多种模型进行销售预测，并采用集成学习（Stacking）方法将这些模型的预测结果进行加权融合，显著提升了预测的精确度和健壮性。各模型各具其独特优势，形成了多模型融合的强大体系。具体而言，ARIMA 模型擅长处理线性及季节性强的时间序列分析，Prophet 模型在应对季节性变化和节假日影响方面表现突出，LSTM 模型在处理非线性挑战和长期依赖关系方面具有出色的预测能力，而 XGBoost 模型在处理特征密集型数据时展现出卓越性能。通过这种多模型融合，集成学习不仅增强了整体预测的精准度，还提升了预测结果的稳定性和可信度。

本研究方案通过细致的模型参数调优和集成学习方法，展现出对不同数据特性的高度适应性。ARIMA 模型依靠遍历式的参数优化寻找最佳组合，Prophet 模型能够自动适应节假日及季节性的变化，LSTM 模型擅长捕捉复杂的时间依赖关系，而 XGBoost 模型则利用决策树的优势实现高效率的预测。这种灵活的适应性，确保了模型在多种实际应用环境中均能发挥出色的预测性能。此外，LSTM 和 XGBoost 模型在处理复杂时间序列数据方面表现尤为突出。LSTM 模型利用递归神经网络的特性，能够捕捉长期的依赖关系，而 XGBoost 模型则通过提升算法和决策树集成的方法，增强了模型的泛化能力，使其能够更有效地识别和预测数据中的复杂模式和趋势。

参考文献：

[1]李融.基于 XGBoost 算法的跨境电商备货预测研究[J].太原城市职业技术学院学报,2024,(01):29-31.
[2]路标.时间序列和决策树模型在线上酒店销量预测中的应用[D].导师：陈涛.南昌大学,2023.
[3]卢亚茹.新冠病毒肺炎疫情影响下基于 XGBoost 的润滑油销量预测[J].石油化工管理干部学院学报,2023,25(04):35-40.
[4]刘凯迪.集成学习算法在新能源汽车市场中的分析和预测[D].导师梁鑫.广西师范大学,2022.
[5]王细雨.基于 LSTM-XGBoost 的电商商品短期销量预测[D].导师：刘洪;温荣泉.中南财经政法大学,2022.
[6]刘辰阳.J 社区团购平台基于 XGBOOST 的快消品销量预测方法[D].导师：胡祥培.大连理工大学,2022.
[7]徐浩帆. 贝叶斯优化下 SARIMAX 和 LSTM 模型在日照港货物吞吐量预测中的应用[J].物流工程与管理, 2024, 46 (04): 24-28.
[8]鲍斌,云雄,甘国操,谢佳,刘辉.基于 Prophet 的民航商务旅客出行量预测研究[J]. 航空计算技术,2024,54(02):79-82+87.

[9]刘合兵,王一飞,王垒,席磊,尚俊平.基于融合影响因素 PSO-Prophet 模型的农产品价格预测[J]. 湖北农业科学, 2024, 63 (01):185-189.
[10]丁美荣,张迎春.融合序列分解与 Prophet 模型的时序预测[J].计算机系统应用,2023,32(11):294-301.
[11]秦秋洪,陈羽,向哲,黄小英.一种挖掘机的改进型 SVR-SARIMAX 混合销量预测模型方法[J].工程机械,2022,53(09):149-155+13.