

基于机器学习的良仓烘焙店销售需求预测模型研究

许嘉宝

江苏大学管理学院, 江苏 镇江

收稿日期: 2024年12月5日; 录用日期: 2024年12月19日; 发布日期: 2025年1月22日

摘要

随着消费者需求的波动性和市场竞争的加剧, 烘焙行业对销售需求预测的精确度提出了更高的要求。传统的销售预测方法, 如基于历史平均值或时间序列分析, 无法有效应对烘焙店销售中复杂的季节性、促销活动以及其他多因素影响。为此, 本文以良仓烘焙店为例, 探讨了烘焙产品的销售特性及影响销售需求的因素, 分析了传统预测方法的局限性。通过对机器学习算法的应用基础进行探讨, 构建了以历史销售数据为基础的销售需求预测模型, 以期为烘焙店优化经营策略提供参考。

关键词

机器学习, 销售需求预测, 烘焙店, 数据分析, 模型构建

Research on the Sales Demand Prediction Model of Liangcang Bakery Based on Machine Learning

Jiabao Xu

School of Management, Jiangsu University, Zhenjiang Jiangsu

Received: Dec. 5th, 2024; accepted: Dec. 19th, 2024; published: Jan. 22nd, 2025

Abstract

With the volatility of consumer demand and the intensification of market competition, the baking industry has put forward higher requirements for the accuracy of sales demand forecasting. Traditional sales forecasting methods, such as those based on historical averages or time series analysis, are unable to effectively address the complex seasonality, promotional activities, and other multi

factor influences in bakery sales. Therefore, this article takes Liangcang Bakery as an example to explore the sales characteristics of baked goods and the factors that affect sales demand, and analyzes the limitations of traditional forecasting methods. By exploring the application foundation of machine learning algorithms, a sales demand forecasting model based on historical sales data was constructed to provide reference for optimizing business strategies in bakeries.

Keywords

Machine Learning, Sales Demand Forecasting, Baking Shop, Data Analysis, Model Building

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

烘焙业的季节性和波动性都很明显,准确预测销售需求,对于库存管理和生产计划十分关键。在处理复杂多变的销售数据时,时间序列分析和线性回归等传统预测方法存在局限性,很多因素的综合影响很难捕捉到。利用机器学习模型来预测销售需求,已经成为大数据和机器学习技术发展的新趋势。本文以良仓烘焙店为例,研究基于机器学习的销售需求预测模型,以期提高预测精度、优化库存管理和生产计划、提升企业运营效率和市场竞争力。

2. 良仓烘焙店销售需求分析

2.1. 良仓烘焙产品销售需求特性分析

良仓烘焙产品的销售需求具有显著的季节性、周期性和多样性特征。首先,节假日和特定活动(如中秋、春节、情人节等)会推动特定产品(如月饼、蛋糕)的销量大幅增长。其次,日常需求呈现明显的周期性波动,例如在周末和下午时段,门店销售额通常高于工作日和早晨时段。此外,促销活动、会员积分兑换和天气因素(如温度、降雨)也对消费者的购买行为产生重要影响。例如,寒冷天气可能增加热饮和烘焙点心的销售。总体而言,良仓烘焙店的销售需求受多种因素交织影响,存在短期波动和长期趋势,需要精准预测来支持经营和库存管理。

2.2. 传统预测方法的局限性

传统预测方法在处理良仓烘焙店的销售需求时存在多方面的局限性:1) 时间序列分析方法(如ARIMA)假设数据具有稳定的趋势和季节性,但烘焙产品的销售受多种动态因素影响,如节假日、天气和促销活动,这些因素难以通过简单的统计模型捕捉;2) 线性回归模型假设输入变量与销售量之间存在线性关系,但实际销售数据往往包含复杂的非线性关系,导致预测精度不高;3) 传统方法通常依赖单一数据源,无法充分利用多源数据(如天气数据、节假日信息)的综合影响;4) 传统方法在处理大规模数据和实时更新方面能力有限,难以适应快速变化的市场需求[1]。这些局限性使得传统预测方法在烘焙行业的应用效果不佳,需要引入更先进的机器学习技术来提升预测的准确性和鲁棒性。

3. 机器学习在销售预测中的应用基础

3.1. 机器学习的基本概念

机器学习是一种人工智能技术,通过算法使计算机在无显式编程的情况下,从数据中学习并提高其

性能。通过训练数据集，机器学习模型学习输入和输出之间的关系，然后用于预测或分类新的数据。主要类型有监督、无监督、强化学习三种。监督学习适用于销售预测等任务，通过带有标签的数据培训模型；对聚类、降维使用的无标签数据进行无监督的学习处理；强化学习使模型在环境中决策持续优化，通过奖励机制。机器学习应用于销售预测，可以捕捉复杂的模式，提高预测的精确性和稳健性。

3.2. 机器学习在销售预测中的应用优势

机器学习在销售预测方面表现出了多方面的优势：1) 机器学习模型通过对这些因素的综合分析，可以处理历史销售数据、天气信息、节假日及促销活动等大量多源数据，提高预测的准确性和全面性；2) 机器学习算法克服了传统统计方法的限制，尤其是在应对季节性强、波动性大的销售数据时，能够捕捉数据中复杂的非线性关系，表现更为优异；3) 具备高灵活性和可扩展性的机器学习模型，能够快速变化的市场，并随着新数据的加入而不断优化更新；4) 机器学习可以使预测过程自动化，减少人为干预，提高预测的准确性和一致性[2]。

4. 数据收集与预处理

4.1. 数据来源与类型

在良仓烘焙店的销售需求预测中，数据来源多样，主要包括历史销售数据、天气数据、节假日信息和促销活动记录。历史销售数据包括每日销售量、产品种类和销售额等，提供了销售趋势的基础信息。天气数据涵盖温度、湿度和降雨量等，影响消费者的购买行为。节假日信息包括法定假日和特殊节日，对销售量有显著影响。促销活动记录包括折扣、赠品和会员优惠等，反映营销策略的效果。

4.2. 数据清洗与整理

数据清洗与整理是确保销售需求预测模型准确性的关键步骤，具体流程如图 1 所示。

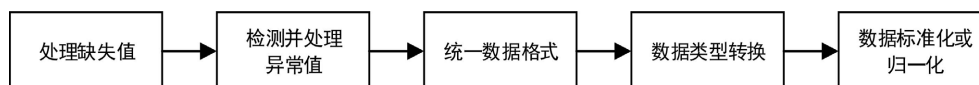


Figure 1. Data cleaning and organization process

图 1. 数据清洗与整理流程

识别数据集中的遗漏数据，对于数值类型的数据，可以用均值、中值或者是中值来进行填列，这样的数值类型可以用中值来填列；对于类别型的数据，可以用众数来填充，也可以用具体的默认值来填充。若遗漏值比重过高，可考虑将有关记录删除。如历史销售数据中的遗漏值，可以采用前补齐，也可以采用后补齐的方法来进行分析：通过统计方法识别异常值(如 Z-SCORE 或 IQR)。对于 Z-SCORE 方法来说，超过阈值(例如 3)的数值被视为异常值，计算每个数据点的标准分数。对于 IQR 方法，计算出四分位范围，超过 IQR1.5 倍的值就算是异常值了。可以用邻近值代替，也可以用删除记录的方法处理异常值。保证各数据板块格式一致。例如日期格式统一为 YYYY-MM-DD，时间格式统一为 YYYY-MM-DD。对于文字数据，如去掉多余的空格、特殊的字符等，都要进行规范的处理[3]。将数据转换成合适的类型，例如将日期字符串转换成日期对象，将类别型数据转换成值数据(例如独热编码)。对于数值型的数据，进行标准化处理或归一化处理，使其在同一个范围内，数值型的数据可以归一化，也可以归一化。常用的方法包括最小 - 最大归一化和 Z-score 标准化。

4.3. 特征选择与提取

特征选择和提取是提升销售需求预测模型性能的关键步骤，通过相关性分析和特征重要性评估来选

择历史销售数据、天气状况、节假日和促销活动等与销量高度相关的特征。进行特征工程，以捕捉更多影响销售的隐含信息，创建诸如假日标志、周末标志以及天气指数等新的衍生特征。例如，移动平均线可以被计算来平滑历史销售数据，公式为：

$$MA_t = \frac{1}{n} \sum_{i=0}^{n-1} S_{t-i}$$

其中， MA_t 表示在时间 t 的移动平均值， n 是窗口大小， S_{t-i} 是过去 n 天的销售量。通过这些手段，能够降低噪音，提高解读能力，提高模型的预测准确率。

4.4. 数据集划分

将数据集分为训练集、验证集和测试集，以保证模型训练和评价的有效性。用于模型训练的训练集，用于超参数调优和模型选型的验证集，以及用于对模型进行泛化能力最终评估的测试集。常见的划分比例为训练集的 70%，验证集的 15%，测试集的 15%。对于时序数据，应依时序将数据集分割，以免日后数据外泄至训练集中。比如最早的数据可以拿来训练，中间的数据可以拿来验证，最近的数据可以拿来考。这样就能保证模型在实景中有更可靠的表现。随机分割方法可以用于非时间序列数据。利用随机种子保证每一分的连贯性，避免评估中随意性带来的偏差。例如，随机划分使用 Python 的“train_test_split”函数。可以采用交叉验证的方法，进一步提高模型的坚固性。K 折式交叉校验将数据集分为 K 个子集，每次以 K - 1 子集为单位进行训练，其余 1 个子集为单位进行校验，重复 K 次，取平均成绩作为最后的评估成绩[4]。集合划分的合理性和模型评估的精确性，为后续的模式培训与优化提供了可靠的依据。

5. 模型构建与训练

5.1. 机器学习模型选择

在良仓烘焙店的销售需求预测中，选择了线性回归、决策树、随机森林和支持向量机等多种机器学习模型进行比较。每个模型都有独特的优点和适用场景，但由于在预测精确性和稳健性上都有最好的表现，最终选择了随机森林模型。随机森林是通过多个决策树的构建和对其预测结果的总结来提高模型的精确性和稳定性的一种综合学习方法。随机林可以处理高维度的数据与非线性关系，对于过拟合的抗性较佳。随机森林通过从原始数据集中随机抽取若干个子样本，每个子样本都被用来训练一棵决策树。在各个节点上，为了减少各特征之间的关联性，随机选取一部分特征进行拆分。建立多棵决策树，以子样本和随机选取的特点为基础。对于回归任务，随机森林最终预测值是通过对所有决策树的预测结果进行平均计算得出的。随机森林的预测公式为：

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T h_t(x)$$

其中， \hat{y} 是最终的预测销售量， T 是随机森林中决策树的数量， $h_t(x)$ 是第 t 棵决策树的预测结果。

5.2. 模型训练过程与参数优化

在良仓烘焙店的销售需求预测中，采用了随机森林模型进行训练，将预处理后的数据集划分为训练集、验证集和测试集，确保数据的时间顺序性和一致性。训练集用于模型训练，验证集用于超参数调优，测试集用于最终评估模型性能。初始化随机森林模型，设置初始参数，如决策树的数量(“n_estimators”)、最大深度(“max_depth”)、最小样本分裂数(“min_samples_split”)和最小样本叶子数(“min_samples_leaf”)等。使用训练集数据对随机森林模型进行训练。在训练过程中，随机森林为了减少过度拟合的风险，以随机抽样的子样本和特征为基础，通过构建多棵决策树来训练每棵树。参数优化采用网格搜索和交叉验

证两种方式。网格搜索通过对预定义的参数组合进行遍历，从而找到最佳的参数组合。交叉验证通过进一步将训练集分成若干子集，对模型进行多次训练和验证，从而保证参数选择的稳定性和概括能力。利用验证集合评估优化后的模型，主要评估指标包括均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)和决定系数(R^2) [5]。通过这些指标，各模型的预测表现可以得到综合评价。性能最优的模型参数根据验证集的评测结果进行选择。

5.3. 模型验证

在模型构建与训练过程中，模型验证是确保预测准确性的重要步骤，采用交叉验证方法，以提高模型的泛化能力和预测稳定性，具体如图 2 所示。

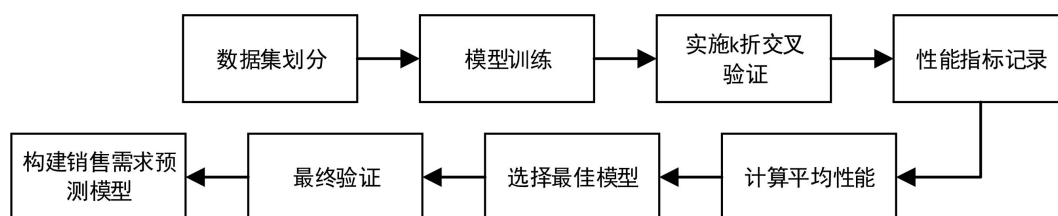


Figure 2. Model validation methods

图 2. 模型验证方法

将数据集划分为训练集和测试，划分比例为 70% 训练集、15% 验证集和 15% 测试集。在训练过程中，模型通过输入特征学习数据中的模式，优化参数以减少预测误差。实施 k 折交叉验证。首先将训练集分成 k 个大小相近的子集，通常取 k 为 5 或 10。在每个交叉验证的迭代中，选择其中一个子集作为验证集，剩余的 $k-1$ 个子集用于模型训练。这样，模型将进行 k 次训练与验证，每次验证都使用不同的子集，确保每个子集都能参与验证，减少了模型对特定训练集的过拟合风险。每次迭代后，记录模型在验证集上的性能指标，如均方根误差(RMSE)和平均绝对误差(MAE)。完成 k 次训练后，计算所有迭代的平均性能指标，以获得模型的整体表现。这种方法不仅可以提供对模型的稳健评估，还能帮助选择最优超参数配置。经过交叉验证后，选定最佳模型并用测试集进行最终验证，以确保模型在未见数据上的有效性与预测能力。

6. 实验分析

6.1. 实验条件设置

为确保实验的有效性，将在相同的环境条件下进行销售需求预测模型的实验。选择一个为期三个月的历史销售数据作为基础数据集，包括各类烘焙产品的日销售量、促销活动、天气情况等影响因素。设置固定的机器学习模型参数，如学习率、迭代次数和批次大小，以便于结果的可比性。确保使用相同的硬件平台和软件环境，以消除系统性能波动对实验结果的影响。实施 k 折交叉验证，确保每次实验都能在多个验证集上进行，增强模型的可靠性和准确性。

6.2. 实验方法

为了确保实验的有效性和可比性，采用了为期三个月的历史销售数据集，这一时间跨度不仅能够捕捉到销售数据的周期性波动，同时也保证了数据集的规模适中，既不过大导致计算资源浪费，也不过小影响模型训练效果。数据集按照 70% 训练集、15% 验证集和 15% 测试集的比例进行划分，这样的比例设置旨在平衡模型训练的需求与评估模型性能的必要性。70% 的数据用作训练集，确保模型能够从足够多

的样本中学习到数据的内在规律；15%的验证集用于模型调优，避免过度拟合；而剩下的 15%作为测试集，则用来最终评估模型的泛化能力，确保其在未见过的数据上的表现。

在数据预处理阶段，对数据进行了清洗，包括处理缺失值和异常值，确保数据的质量。随后，根据业务逻辑和相关性分析，选取了日销售量、促销活动、天气情况等作为特征变量，这些特征对于销售预测具有重要的指示意义。接下来，使用随机森林模型进行训练，并设置了固定的模型参数，以减少随机因素对实验结果的影响。通过实施 5 折交叉验证，进一步提高了模型性能评估的可靠性和稳定性。模型性能的评估指标包括均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)和决定系数(R^2)，这些指标从不同角度全面反映了模型的预测精度和稳定性。最后通过对比不同模型的预测结果，分析了模型的准确性和鲁棒性，确保实验结果既可靠又具有良好的解释性。

6.3. 实验结果分析

随机森林模型在良仓烘焙店销售需求预测中表现出色，具体如表 1 所示。

Table 1. Experimental results
表 1. 实验结果

数据集	MSE	RMSE	MAE	R^2
训练集	12.56	3.54	2.87	0.92
验证集	15.89	3.99	3.21	0.88
测试集	16.23	4.03	3.25	0.87

训练集上的 MSE 为 12.56，RMSE 为 3.54，MAE 为 2.87， R^2 为 0.92，表明模型在训练数据上具有很高的拟合度。验证集上的 MSE 为 15.89，RMSE 为 3.99，MAE 为 3.21， R^2 为 0.88，验证了模型的泛化能力。测试集上的 MSE 为 16.23，RMSE 为 4.03，MAE 为 3.25， R^2 为 0.87，进一步确认了模型在未见过的数据上的稳定性和准确性。总体来看，随机森林模型在预测精度和鲁棒性方面表现最佳，适合应用于实际的销售需求预测中。

7. 结语

综上所述，通过对销售需求的特征分析和影响因素的识别，基于机器学习的良仓烘焙店销售需求预测模型，不仅提高了预测的准确性，还增强了模型的适应性与鲁棒性。所构建的模型能够有效捕捉到销售模式，指导库存管理和促销策略的制定，从而优化经营效率。未来的研究可进一步结合实时数据分析与深度学习技术，以提升预测能力和决策支持，为烘焙行业的可持续发展提供更强的助力。

参考文献

- [1] 周雨, 段永瑞. 基于聚类与机器学习的零售商品销量预测[J]. 计算机系统应用, 2021, 30(11): 188-194.
- [2] 罗蓉. 基于机器学习的连锁企业销售预测模型研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2022.
- [3] 韩亚娟, 高欣. 基于机器学习组合模型的电商商品销量预测[J]. 计算机系统应用, 2022, 31(1): 315-321.
- [4] 许姗姗. 基于机器学习的商品销售预测的研究[J]. 统计与管理, 2019(4): 49-52.
- [5] 徐琪, 张阳奎. 基于 ARIMA 模型与随机森林模型的零售服装动态销售预测模型探析[J]. 中国管理信息化, 2022, 25(6): 100-104.