Advances in Applied Mathematics 应用数学进展, 2024, 13(8), 3814-3825
Published Online August 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.138363

Hans汉斯

基于机器学习的民航客运量预测方法研究

刘浩霖，赵子玉，谢文飞，吴念秋，唐甜甜*

中国民用航空飞行学院理学院，四川 广汉

收稿日期：2024年7月15日；录用日期：2024年8月9日；发布日期：2024年8月16日

摘　要

为提高民航客运量预测精准度，本文针对近18年的时间序列民航客运量数据，构建极限梯度提升树XGBoost预测模型，进行多特征分析，处理季节、节假日等主要因素，并与SVR模型进行对比。通过对比预测曲线图，反映出SVR模型在高维空间中可以找到最优超平面来拟合数据，XGBoost模型适用于复杂的非线性关系建模。实验结果表明，XGBoost预测模型相比于SVR向量回归模型、线性模型与随机森林模型，其精准度更高且对影响因素敏感；XGBoost模型有更高的R^2和更低的MSE，能够更有效提高民航客运量的预测精度和预测稳定性，为制定航空运输生产计划和发展航空运输业提供了重要参考。

关键词

机器学习，XGBoost模型，SVR模型，民航客运量，预测分析

Research on Civil Aviation Passenger Volume Forecasting Method Based on Machine Learning

Haolin Liu, Ziyu Zhao, Wenfei Xie, Nianqiu Wu, Tiantian Tang*

College of Science, Civil Aviation Flight University of China, Guanghan Sichuan

Received: Jul. 15th, 2024; accepted: Aug. 9th, 2024; published: Aug. 16th, 2024

Abstract

In order to improve the accuracy of civil aviation passenger traffic prediction, this paper, based on the civil aviation passenger traffic data of recent 18 years, builds the ultimate gradient lift tree XGBoost prediction model, conducts multi-feature analysis, processes major factors such as seasons and holidays, and compares it with the SVR model. By comparing the prediction curves, it

*通讯作者。

文章引用: 刘浩霖, 赵子玉, 谢文飞, 吴念秋, 唐甜甜. 基于机器学习的民航客运量预测方法研究[J]. 应用数学进展, 2024, 13(8): 3814-3825. DOI: 10.12677/aam.2024.138363

shows that SVR model can find the optimal hyperplane to fit the data in the high-dimensional space, and XGBoost model is suitable for complex nonlinear relationship modeling. The experimental results show that compared with SVR vector regression model, linear model and random forest model, XGBoost prediction model is more accurate and sensitive to influencing factors. XGBoost model has higher R^2 and lower MSE, which can improve the forecast accuracy and stability of civil aviation passenger volume more effectively, and provide an important reference for the development of air transport production plan and air transport industry.

Keywords

Machine Learning, XGBoost Model, SVR Model, Civil Aviation Passenger Volume, Predictive Analysis

1. 引言

航空运输作为当今全球化社会中不可或缺的重要组成部分，其运输效率和服务水平对社会经济发展有着重要的作用。航空公司需要可靠的工具对客运量进行预测，以便更好地调整航班计划、优化资源配置，满足旅客需求。

近年来，为了适应社会经济发展和科技进步的需求，通过预测和数据分析的方法来改进工作在各个领域发挥着重要作用[1]。在研究过程中，有学者通过信号处理方法将客运量时间序列中的线性和非线性成分分离，并探究疫情作为特征对民航客运量需求的影响[2]。也有学者对数据采用 EMD 进行处理后，结合 CNN-LSTM 模型进行了客运量的短期预测[3]。同时也有学者发现 XGBoost 高效，灵活和便携的特点[4]-[6]，提出了 GWO-XGBoost 算法来进行数据预测[7]。部分学者研究发现 SVR 模型短期网络有着较好的拟合效果[8]，通过对比 BP 神经网络来进行机器学习[9]-[11]。除此之外，学者使用长短期记忆模型(LSTM)、NARX 动态神经网络模型等先进算法对客运量进行预测，证实了有较高的预测正确性[12]。较新的研究表明许多学者通过独立使用 XGBoost 模型进行数据的预测有着良好的效果[13]。这些研究不仅为各行各业提供了重要的数据分析和预测方法，也为机器学习领域的应用提供了丰富的实践案例。

本文旨在利用 XGBoost 模型和 SVR 模型，结合时间序列分析方法，深入探讨民航客运量需求的预测方法，以期为相关行业的发展提供更为可靠和高效的预测方案，推动科技与社会经济的融合发展。

2. 模型原理

2.1. XGBoost 模型概述

XGBoost (Extreme Gradient Boosting)即极度梯度提升树模型。XGBoost 的目标函数由两部分组成，损失函数和正则项。

对于第 t 颗树，第 i 个样本的，模型的预测值：

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_{k(x_i)} = \hat{y}_i^{(t-1)} + f_{t(x_i)} \tag{1}$$

其中，$\hat{y}_i^{(t)}$ 是第 t 次迭代之后样本 i 的预测结果；$f_{t(x_i)}$ 是第 t 棵树的模型预测结果；$\hat{y}_i^{(t-1)}$ 是第 $t-1$ 棵树的预测结果；

对于第 i 个样本，最终的预测值为：

$$\hat{y}_i^{(T)} = \sum_{j=1}^{T} f_j(x_i) \tag{2}$$

最终得到原始目标函数：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{j=1}^{t} \Omega(f_j) \tag{3}$$

其中，第一项是损失函数；第二项是正则项，代表全部 t 棵树的复杂度。这两项是两个维度的问题，一个是针对所有样本，一个是针对所有树。

2.2. SVR 神经网络建模原理

SVR (Support Vector Regression)是一种支持向量机(SVM)的回归变体，它通过在线性函数两侧制造一个“间隔带”来处理回归问题。在 SVR 中，模型函数也为线性函数，形式为：

$$W^T x + b = \varepsilon \tag{4}$$

$$W^T x + b = 0 \tag{5}$$

$$W^T x + b = -\varepsilon \tag{6}$$

但与线性回归模型不同，SVR 的目标是找到一个超平面，使得所有落在间隔带内的样本点的损失为零，而间隔带外的样本点的损失则计入总损失。

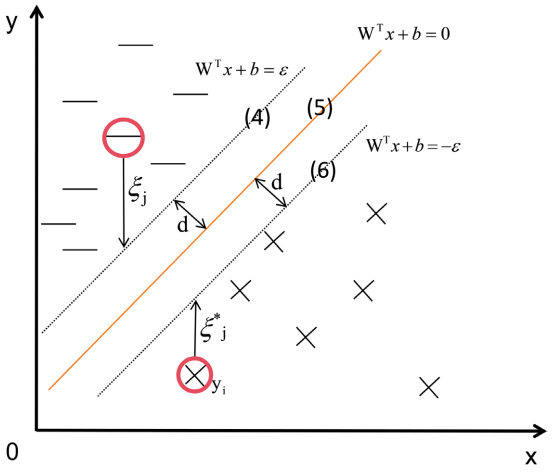

Figure 1. SVR model structure diagram
图 1. SVR 模型结构图

如图 1 所示，SVR 模型包含两个松弛变量 ξ_j 和 ξ_j^*，它们分别表示样本点 x_i 到超平面(间隔带的上下边缘)的距离与该距离与 y_i 值的差。当样本点落在间隔带内或边缘上时，ξ_j 和 ξ_j^* 为零；当样本点位于间隔带上边缘上方时，$\xi_j > 0$，$\xi_j^* = 0$；当样本点位于间隔带下边缘下方时，$\xi_j = 0$，$\xi_j^* = 0$。

SVR 在线性函数两侧制造了一个“间隔带”，间距为 d (也叫容忍偏差，是一个由人工设定的经验值)，对所有落入到间隔带内的样本不计算损失，当且仅当 $f(x)$ 与 y 之间的差距的绝对值大于 ξ 通过最大化间隔带的宽度与最小化总损失来优化模型。

SVR 的最优结果为：

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

定义误差为 *loss*，则 SVR 模型问题可表达为

$$\min_{w,b}\frac{1}{2}\|w\|^2+loss$$

此次时间序列民航客运量数据为二维样本 $T=\{(x_1,y_1)\cdots(x_n,y_n)\}$，模型预测值为 $f(x_i)$，误差为 y_i，SVR 模型中允许最大误差 d，当 $|f(x_i)-y_i|>d$ 时，开始计算误差，否则默认为准确预测值。

其中 *loss* 表达式为：

$$C\sum_{i=1}^{n}L_d\left(f(x_i)-y_i\right),$$

C 为惩罚系数，L_d 为损失函数。

3. 构建模型与预测分析

3.1. 数据来源

选取了中国民航总局官网中 2005 年 1 月至 2024 年 1 月的民航客运量数据作为研究对象，部分数据见表 1。

Table 1. Monthly data of domestic civil aviation passenger volume from 2005 to 2024
表 1. 国内民航客运量 2005~2024 年月度数据

月份	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
1	1500	1748	1941	2267	2567	2567	3058	3246	3736	4393	4647	4647	4401	3004	2942	3977
2	1502	1702	2025	2163	2349	2787	3109	3493	3898	4279	4843	5341	729	2386	3121	4320
3	1581	1800	2173	2282	2503	2878	3025	3670	3894	4431	5140	5341	1664	4768	1527	4569
4	1659	1883	2157	2437	2605	2849	3142	3579	4000	4402	5074	5341	2573	5091	1197	5000
5	1536	2264	2600	2758	3071	3476	3730	4165	4642	5046	5657	5930	4598	2227	2189	5170
6	1417	1771	2185	2325	2530	2865	3061	3404	3800	4374	4938	5341	3060	4116	2189	5312
7	1718	2103	2547	2660	2995	3266	3580	3915	4353	4860	5378	5930	3894	4899	3385	6243
8	1596	2264	2660	2758	3071	3476	3730	4165	4642	5046	5657	5930	4598	2227	3213	6396
9	1641	1931	2260	2501	2744	3049	3315	3672	4169	4655	5029	5930	4775	3599	1989	5349
10	1834	2171	2448	2638	2829	3165	3488	3855	4374	4883	5408	5930	5014	3875	1572	5605
11	1700	1958	2078	2371	2584	2851	3240	3498	3971	4646	5006	5930	4425	2142	1234	4899
12	1567	1875	2103	2300	2563	2793	3193	3500	4041	4666	5018	5930	4216	2698	1841	5059

3.2. 构建 XGBoost 模型

3.2.1. 建模步骤

图 2 所示为 XGBoost 模型的详细建模步骤：

1) 对时间序列数据进行预处理，根据模型的特征重要性进行特征选择；
2) 计算趋势特征(斜率、差分等)，捕捉时间序列数据中的趋势信息；
3) 将数据时间戳分解为年、月时间特征；
4) 创建滞后特征，将过去部分时间点的观测值作为特征输入到模型中；
5) 使用移动平均和移动总和特征来平滑数据并捕捉趋势；
6) 对时间序列数据进行滑动窗口交叉验证，确保模型对未来数据的泛化能力；
7) 通过交叉验证调整超参数，提升模型性能和泛化能力。

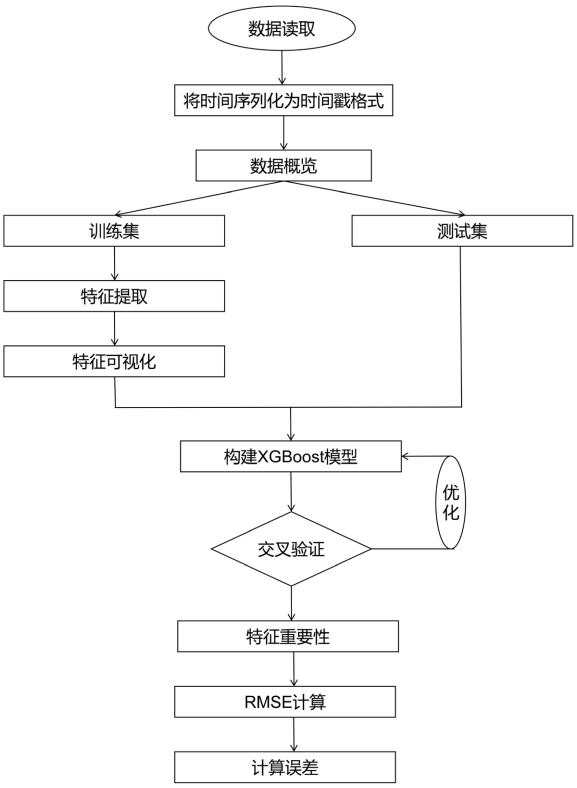

Figure 2. Flowchart for building the XGBoost model
图 2. 构建 XGBoost 模型的流程图

3.2.2. 数据展示

由图 3 可知数据在 2019 年之前呈现整体上升的趋势，但在 2019 年之后，因疫情影响，整体客运量数据波动幅度剧增并持续至最新数据。因此在数据使用时要考虑疫情期间的影响，必要时则需将其排除在外。

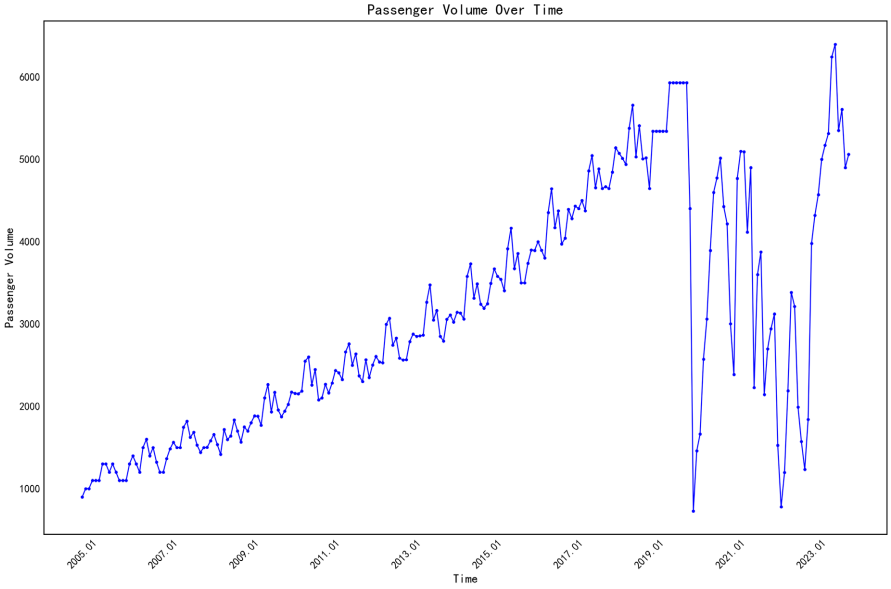

Figure 3. Time series chart of raw data on civil aviation passenger volume from 2005 to 2023
图 3. 民航客运量 2005~2023 年原始数据时序图

3.2.3. 特征可视化

观察图 4，可以看出在滞后一天和两天的曲线 lag1 和 lag2 对模型特征重要可视化影响较小，mouth 即月份对模型特征重要性可视化的影响也较小，但 year 即年份对模型特征重要性可视化的影响较大。

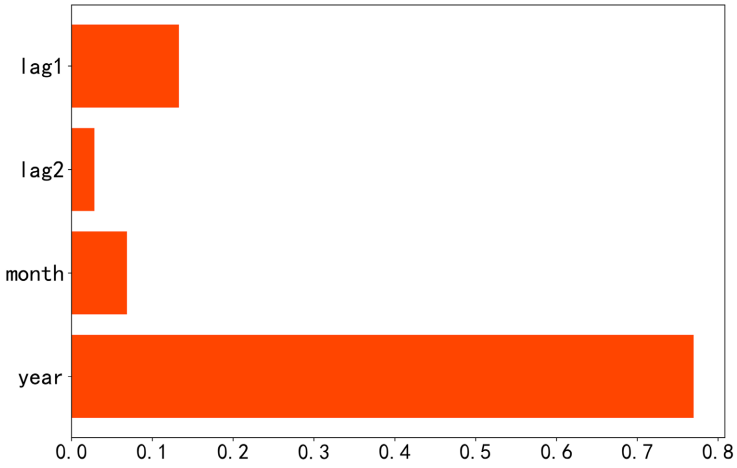

Figure 4. Visualization of feature importance
图 4. 特征重要性可视化

观察图 5，可以看出模型训练集曲线的纵坐标接近于 1，代表模型的学习能力很好，但测试集曲线的纵坐标只在 0.85 以下，表示模型的泛化能力较弱，并且测试集与训练集相差较大，代表模型过拟合，重新调整训练模型。

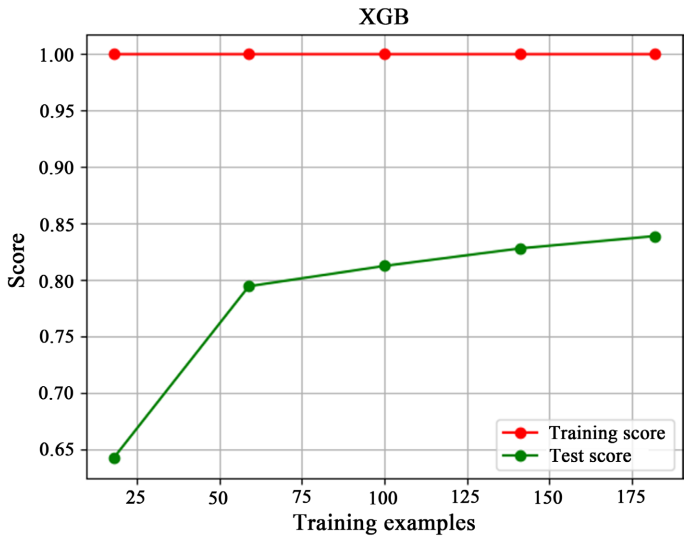

Figure 5. Cross validation
图 5. 交叉验证

观察图 6，可以看出参数学习曲线在横坐标大约 50~100 时曲线大幅上升，但在超过 100 之后曲线逐渐恢复平稳，可以得出树的数量对模型表现是有极限的，因此树的棵树为 100 时模型能力最佳。

观察图 7，其中两条红色虚线表示方差线，中间黑色曲线表示 R^2 线，可以看出树的数量在超过 100 之后曲线一直保持平稳。

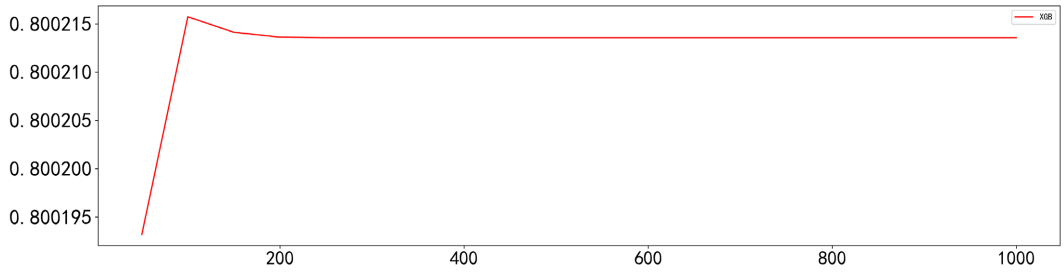

Figure 6. Parameter learning curve
图 6. 参数学习曲线

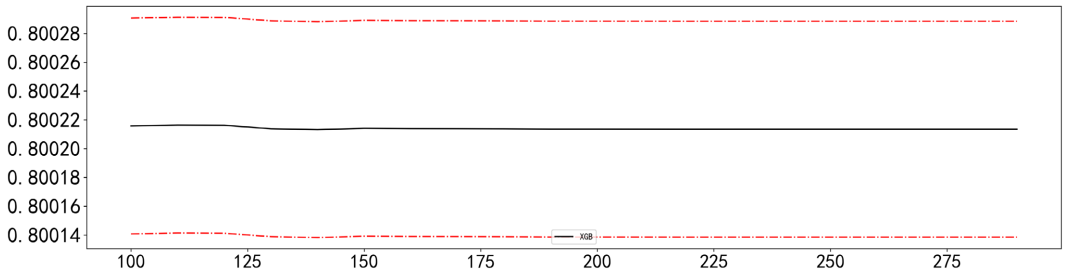

Figure 7. Variance and generalization error
图 7. 方差与泛化误差

3.2.4 模型的预测

在图 8 中，蓝色、橘色曲线分别表示真实、预测数据曲线。从数据的拟合情况可以看出，大部分预测数据和真实数据相差不大，有些月份预测误差较大，但整体预测效果较好，在几个跳跃点上模型也能够及时作出正确的判断，预测曲线能够较好地拟合原数据曲线。

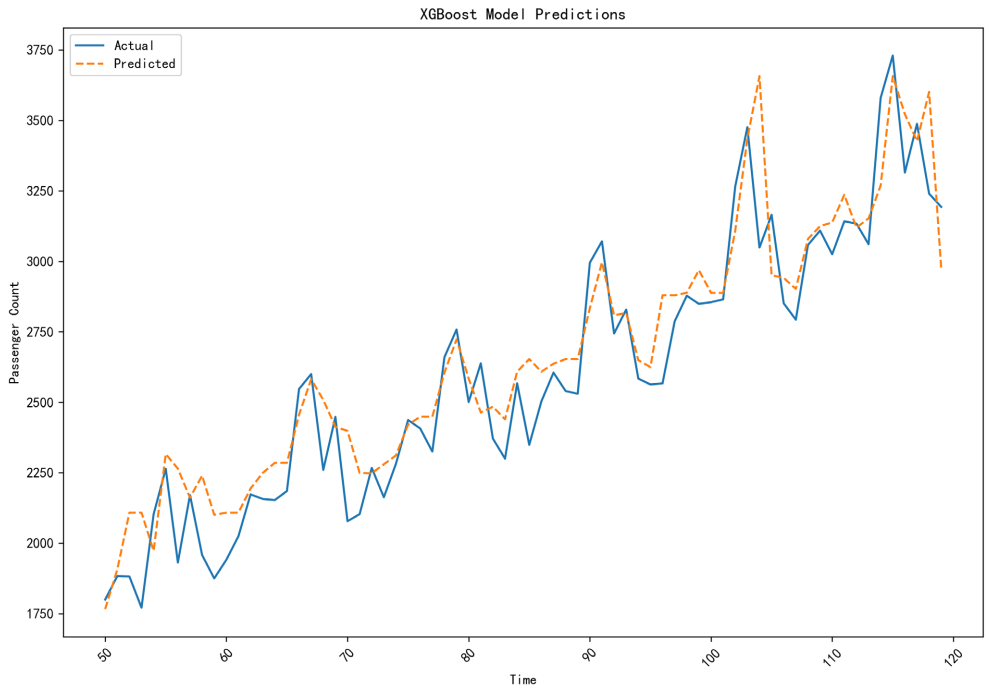

Figure 8. XGBoost model prediction diagram
图 8. XGBoost 模型预测图

3.3. 构建 SVR 模型

3.3.1. 建模步骤

图 9 所示为 SVR 模型的构建流程，首先进行原始数据平稳性检验，其次进行特征提取并划分训练集和测试集，然后对模型参数进行训练调优，最后得到模型结果。

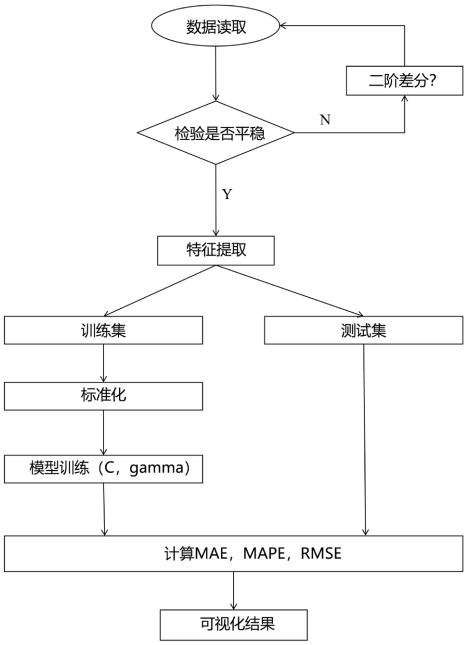

Figure 9. SVR model construction process diagram
图 9. SVR 模型构建流程图

3.3.2. 数据预处理

由图 10 可知，原始数据时序图存在不平稳性和周期性，差分法可以剔除周期性因素。数据进行一阶差分之后观察发现时序还存在不平稳性，从而需要进行二阶差分处理，最终二阶差分处理后的数据呈现周期性，说明数据已经具有平稳性。

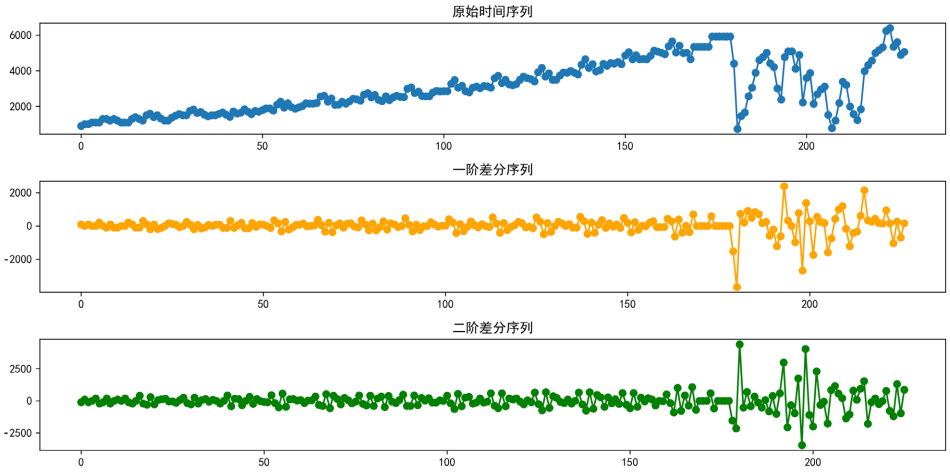

Figure 10. Stability test of passenger volume
图 10. 客运量平稳检验

3.3.3. 模型的训练

C 参数控制了正则化的强度，它的值越大，正则化效果越弱。选择合适的 *C* 可以避免模型过拟合或欠拟合，这里通过交叉验证来调整 *C* 的值，从图 11 可以看出 *C* 取值为 1 时，预测效果最佳。

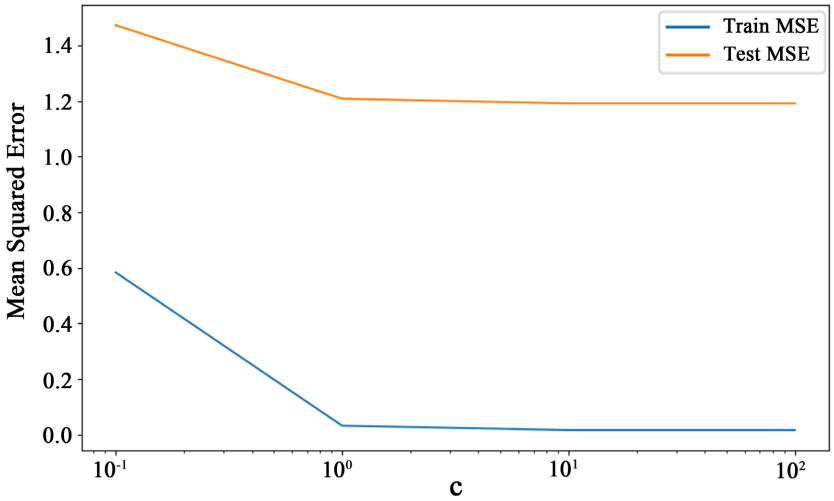

Figure 11. Observation of SVR model parameter *C*
图 11. SVR 模型参数 *C* 观察

Gamma 参数是 RBF 核函数的一个参数，它控制了数据点的影响范围。较小 *gamma* 值表示影响范围较大，较大的 *gamma* 值表示影响范围较小，这里通过交叉验证来调整，从图 12 中可以看出 *gamma* 取值为 1 时模型预测效果最佳。

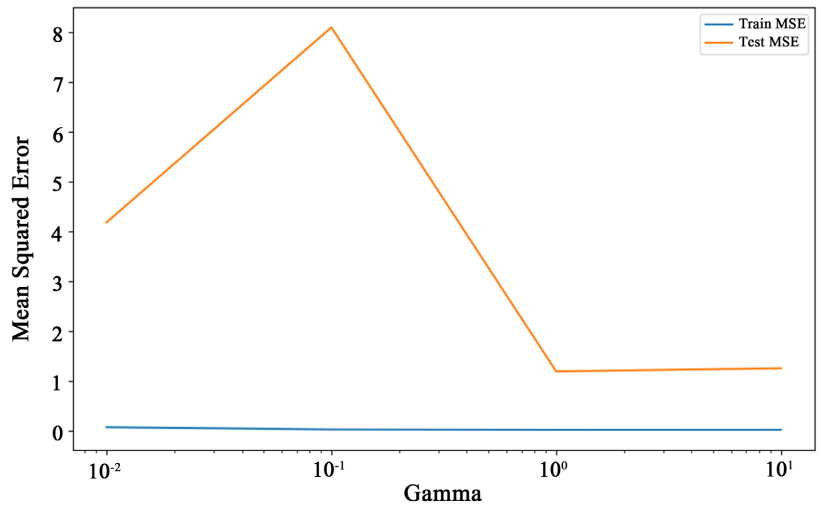

Figure 12. Observation of *gamma* parameters in SVR model
图 12. SVR 模型参数 *gamma* 观察

3.3.4. 模型的预测

选取 SVR 模型对预处理的民航客运量数据进行部分月份的预测分析。从图 13 中可以看出，个别点民航客运量是增长还是降低存在误判的情况，但是总体的趋势预判较为合理。其中，蓝色曲线表示原始数据曲线，橙色曲线表示 SVR 模型预测的曲线。

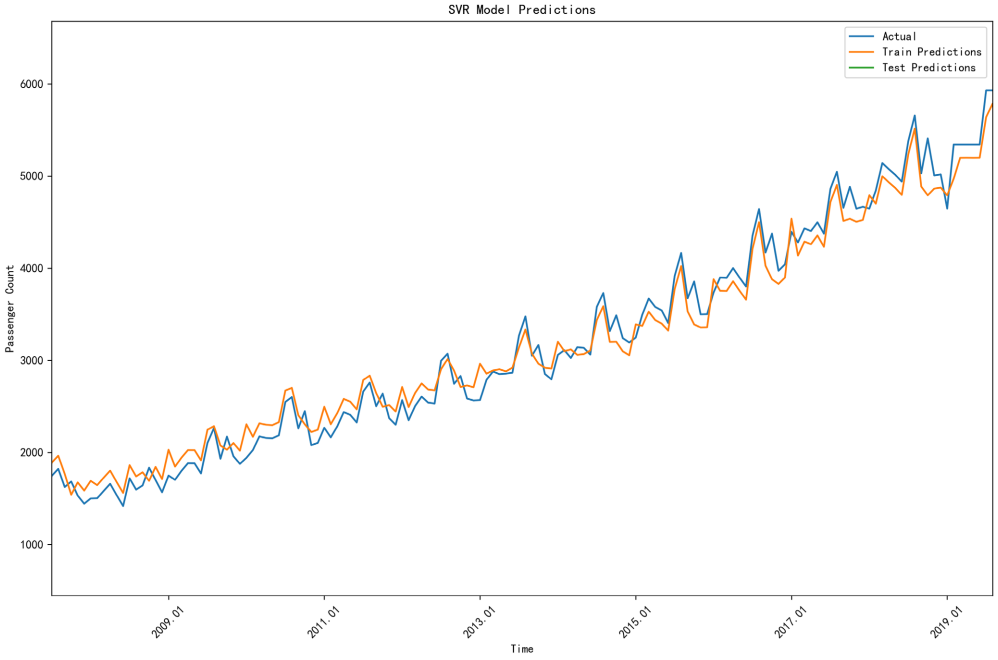

Figure 13. SVR model prediction diagram
图 13. SVR 模型预测图

3.4. 预测结果对比分析

为了验证模型的有效性，将构建的 XGBoost 模型与 SVR 模型的实验预测结果进行对比。虽然它们很有用，而且被广泛用于比较同一数据集上的不同方法，但在这里，表达相对于我们试图预测的时间序列的大小的误差会更有用。为了更加清楚直观地对比，对两个模型进行均方根误差 RMSE、平均绝对百分比误差 MAPE 和平均绝对误差 MAE 的对比分析。各指标数据对比见表 2。评判指标的计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2} \tag{7}$$

RMSE 均方根误差，范围$[0,+\infty)$，当预测值与真实值的差值越接近 0，表示模型越完美；误差越大，该值越大。

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{8}$$

MAPE 平均绝对百分比误差，范围$[0,+\infty)$，MAPE 的值越接近 0%表示模型越完美，MAPE 大于 100%则表示劣质模型。

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{9}$$

MAE 平均绝对误差，是绝对误差的平均值，平均绝对误差能更好地反映预测值误差的实际情况。范围$[0,+\infty)$，当预测值与真实值完全吻合时等于 0，即完美模型；误差越大，该值越大。

图 14 所示为四种模型的预测结果，从表 2 中不难看出 XGBoost 模型，SVR 模型预测结果均比较理想，但是 XGBoost 模型的各项指标均优于 SVR 模型。SVR 预测模型虽然也可以较好地对时序数据进行预测，但拟合效果不如 XGBoost 模型表现出色。

Table 2. Comparison of model prediction error results
表 2. 模型预测结果误差对比

预测模型	RMSE	MAE	MAPE(%)
XGBoost 模型	0.97	0.89	8.36%
SVR 模型	0.99	0.97	8.74%
随机森林模型	1.06	1.39	11.73%
线性回归模型	1.36	1.49	25.74%

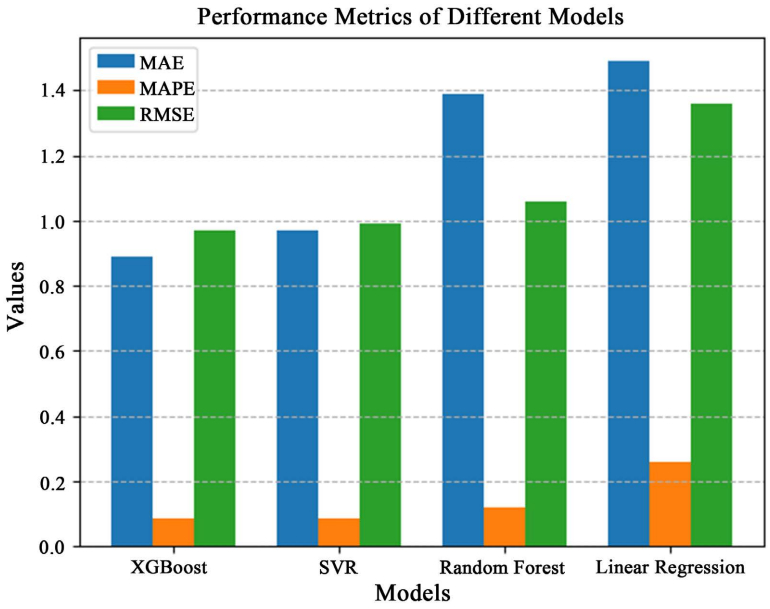

Figure 14. Error in prediction results of each model
图 14. 各模型预测结果误差

本文主要使用 XGBoost 模型描述时间序列数据的变化，SVR 模型模拟数据的非线性规律，可以很好的处理非线性问题和不确定性问题。

4. 结论

机器学习技术在时间序列预测方面取得了较大的发展与成功应用。选取民航客运量时间序列数据，对其进行分析和处理后，构建 XGBoost 模型，SVR 模型，提取时间序列特征对民航航线的客运量进行预测。与其他模型相比，XGBoost 模型和 SVR 模型具有运算速度快、精度高、泛化能力强等优点。民航客运量预测工作数据量大，分布规律呈现非线性函数关系，因此采用 XGBoost 模型和 SVR 模型解决该问题。将决策树模型与向量回归模型进行对比分析，试图找出更符合实际情况的模型。通过对基于时间序列民航客运量的预测，对比两类方法，得出结论：相较于 SVR 模型，XGBoost 模型均能表现出更加优秀的预测精度，XGBoost 模型在学习与预测时间序列能力方面更为优秀，实验数据的预测效果更好。

基金项目

学生科技基金(XSB2024-009) + 青年基金项目(QJ2023-037)。

参考文献

[1] 果泽泉, 何波, 何强, 等. 集中供热热力站短期热负荷预测模型对比研究[J]. 区域供热, 2024(1): 14-15.

[2] 王景荣. 民航与铁路客运需求预测以及疫情的影响[D]: [硕士学位论文]. 南昌: 江西财经大学, 2021.

[3] 孟琪琳, 窦燕. 基于 EMD-CNN-LSTM 模型的铁路客运量短期预测研究[J]. 铁道运输与经济, 2023, 45(12): 65-73.

[4] 刘芳, 李士伟, 卢熹, 等. 基于 PSO-CNN-XGBoost 水下柱形装药峰值超压预测[J]. 兵工学报,2024, 45(5): 1602-1612.

[5] 赵兵朝, 张晴, 王京滨, 等. 基于 SSA-XGBoost 模型的地表下沉系数预测研究[J]. 矿业研究与开发, 2024, 44(2): 89-95.

[6] 曹缘, 王振华, 张继红, 等. 基于 WOA-XGBoost 的膜下滴灌棉花蒸散量预测模型[J]. 排灌机械工程学报, 2024(1): 1-8.

[7] 牛景辉. 基于 GWO-XGBoost 的工业污水水质关键数据预测算法[J]. 工业水处理, 2024, 44(1): 184-190.

[8] 梁亚玲, 陈英伟, 刘思佳. 基于 SSA-SVR 模型的国内新能源汽车销量预测研究[J]. 现代工业经济和信息化, 2023, 13(9): 290-293.

[9] 魏棕凯, 王晓兰, 刘洋成, 等. 基于 SVR 与 BP 神经网络的水电机组瓦温预测[J]. 水电与新能源, 2024, 38(1): 71-74.

[10] 杨赟, 张丽丽. 基于 ISOA-SVR 模型的短期网络舆情预测[J]. 计算机工程与设计, 2024, 45(1): 168-176.

[11] 李轩, 周新苗, 吴晓峰. 基于 HW-EEMD-SVM 模型的民航客运量预测[J]. 数量经济研究, 2023, 14(2): 189-204.

[12] 赵烜. 一种基于 ARIMA 和 LSTM 的民航旅客订座组合预测模型[J]. 计算机与现代化, 2020(11): 65-69+76.

[13] 樊智勇, 王振良, 刘哲旭. 基于 XGBoost 的民航飞机发动机性能参数预测模型[J]. 计算机测量与控制, 2023, 31(6): 46-52.