多变量乘用车销量预测模型研究

段昊江　吴冰

（同济大学经济与管理学院，上海 201800）

【欢迎引用】段昊江，吴冰. 多变量乘用车销量预测模型研究[J].汽车文摘，2023(12): 55-62.

【Cite this paper】DUAN H J, WU B. Research on Multivariable Prediction Model of Passenger Car Sales[J]. Automotive Digest (Chinese), 2023(12): 55-62.

【摘要】在乘用车行业日渐繁荣的大环境下，准确把握行业发展方向，制定相适应的生产目标对各个车企来说非常重要。为了提高对乘用车销量的预测精度，选取历史销量、宏观经济指标和网络搜索关键词数据作为变量，对乘用车整体市场建立了多种销量预测模型，并经过对比分析得到综合考虑上述3种变量的梯度提升决策树模型效果最优，其平均绝对百分误差（MAPE）为10.35%，能够较好地预测销量变化。该模型可以帮助车企了解市场趋势，做出针对性的生产计划安排，同时为销量预测的研究提供一种新的参考模型。

关键词：乘用车市场；销量预测；网络搜索关键词；梯度提升决策树

中图分类号：F426.471　　**文献标识码：**A　　**DOI:** 10.19822/j.cnki.1671-6329.20220312

Research on Multivariable Prediction Model of Passenger Car Sales

Duan Haojiang, Wu Bing

（School of Economics and Management, Tongji University, Shanghai 201800）

【Abstract】In the environment that the passenger car industry is becoming increasingly prosperous, it is crucial for each automobile enterprise to accurately grasp the development direction of the industry and formulate suitable production goals. In this paper, historical sales, macroeconomic indicators and online search keyword data are selected as variables to establish a variety of sales prediction models for the overall passenger car market in order to improve the prediction accuracy of passenger car sales. Through comparative analysis, the Gradient Boosting Decision Tree (GDBT) algorithm model considering the above 3 variables has the best effect and its Mean Absolute Percentage Error (MAPE) is 10.35%. The model obtained in this paper can help automobile enterprises understand the development of market trends, make targeted production planning and provide a new reference model for the research of sales forecast.

Key words: Passenger car market, Sales forecast, Online search keyword, GBDT

缩略语

ARMA	Auto-Regressive Moving Average Model
RBF	Radial Basis Function
SARIMA	Seasonal Auto-Regressive Moving Average Model
BP	Back Propagation
GDP	Gross Domestic Product
VECM	Vector Error Correction Model
XGBOOST	eXtreme Gradient Boosting
VAR	Vector Autoregressive Model
CNN	Convolutional Neural Network
MSFM	Multivariate Sales Forecasting Model
CPI	Consumer Price Index
GDBT	Gradient Boosting Decision Tree
MAPE	Mean Absolute Percentage Error
SVR	Support Vector Regression

0　引言

近年来，随着国民经济的高速发展和居民可支配收入的不断提高，我国乘用车市场已经趋于饱和，乘用车市场消费主力萎缩，加之近几年我国城市公共交通系统的完善、高铁线路网的扩散、城市限行等因素的作用，导致了汽车销量的低迷。2020年，我国乘用车销量为2 006万辆，创近年新低。2021年，随着国内疫情逐步得到控制，乘用车行业逐渐回暖，全国乘用车销量达到2 175万辆。2022年，乘用车销量更是上涨到2 309万辆。在汽车行业日渐繁荣的大环境下，

无论是已经在市场中占据席位的车企还是打算进入汽车行业的“新人”，都应认真剖析市场现状格局和国家政策，准确把握行业发展方向，精准识别消费者需求，从而制定相适应的生产目标，实现按需生产。

在乘用车销量预测研究中，对于变量的选择，早期大部分学者只基于历史销量数据展开预测[1-2]，虽然预测模型与结果对市场长期性的变化有着比较好的效果体现，但是当市场出现波动时，时间序列模型的预测精度就会大幅度降低。有研究发现宏观经济因素会对乘用车的销量产生很大影响，是非常重要的外部因素，国内外学者开始考虑将宏观经济指标作为变量添加到模型中进行预测研究[3-5]。近年来，随着互联网和大数据的发展，消费者们几乎都喜欢在网络中搜索自己想购买的商品，并将网络中的相关信息和其他消费者的评价作为自己是否消费的参考。网络搜索数据为商品市场发展动向和趋势提供了一定程度的前瞻，越来越多的学者都开始应用时效性更好的网络搜索数据来进行乘用车的销量预测研究[6-10]。

但是，国内外学者们在进行乘用车销量预测的研究时，大多只选取一到两类影响变量进行研究，少有将上述变量类型综合起来同时进行考虑的研究，因此，本文在已有研究的基础上，以乘用车整体市场为研究对象，逐步添加历史销量、宏观经济指标、网络搜索关键词这些影响乘用车销量的变量，建立销量预测模型并进行效果比对分析。

1 文献综述

在乘用车销量预测方面，现有研究中考虑到的变量主要包括历史销量、宏观经济指标、网络搜索数据等，按照变量的选择与组合可以分为基于单类型变量的销量预测研究和基于双类型变量的销量预测研究。

1.1 基于单类型变量的销量预测

1.1.1 历史销量

早期关于销量预测的研究主要集中于使用时间序列分析方法，通过寻找销量数据过去的规律来预测其未来的发展趋势。李响等[1]通过自回归滑动平均模型（Auto-Regressive Moving Average Model，ARMA）与径向基函数（Radial Basis Function，RBF）神经网络相结合的混合模型来对销量进行了预测，并通过仿真试验验证了模型的有效性。王旭天等[2]考虑了季节性因素的影响，采用季节性差分自回归滑动平均模型（Seasonal Auto-Regressive Moving Average Model，SARIMA）并选取2004年到2015年的月度销量数据来进行销量预测。

1.1.2 宏观经济指标

在众多影响乘用车销量的因素中，宏观经济方面的因素是非常重要的外部因素。王栋等[3]通过灰色关联分析法筛选出国民总收入、人均国内生产总值（Gross Domestic Product，GDP）、进出口总额等7个影响汽车保有量的相关因子，将其加入反向传播（Back Propagation，BP）神经网络模型进行预测分析。Gao等[4]分析了中国汽车销量与经济变量之间的关系，建模结果表明钢铁产量和汽油价格为内生变量，并与中国汽车销量之间存在长期协整关系，因此建立了向量误差修正模型（Vector Error Correction Model，VECM）来定量分析内生变量对中国汽车销量的长期影响。

1.1.3 网络搜索数据

网络搜索数据能够体现用户的关注意向和需求变化，可以提高销量预测模型的精准度。袁庆玉等[6]首先将关键词进行合成，之后建立模型研究关键词合成指数与汽车销量之间的关系并进行预测。李忆等[7]运用文本挖掘技术确定网络搜索关键词库，构建固定效应模型来研究网络搜索数据与汽车销量的关系，最终发现两者间存在长期均衡关系。

1.2 基于双类型变量的销量预测研究

现有研究中也有不少学者结合两种类型的变量对汽车销量进行建模预测。王易等[5]结合历史销量数据与宏观经济数据，建立了极端梯度提升（eXtreme Gradient Boosting，XGBOOST）模型对汽车销量进行了预测研究。Yong等[8]提出预测电动汽车量化市场需求的模型，考虑5个经济指标和关键词“电动汽车”百度指数的外部影响，建立了多元向量自回归模型（Vector Autoregressive Model，VAR），使预测精度得到显著提高。刘吉华等[9]以大众汽车为研究对象，构建其网络搜索关键词库，并基于主成分分析将关键词进行合成，建立了结合历史销量数据与网络搜索数据的回归预测模型。刘吉华等[10]在以后的研究中还将深度学习的算法引入汽车销量预测，用卷积神经网络模型（Convolutional Neural Network，CNN）来对结合历史销量数据与网络搜索数据的大众汽车进行销量预测。

综上所述，可以发现目前对于汽车销量预测分析有很多效果良好的研究模型，但在变量的选择与组合方面，少有同时考虑历史销量、宏观经济指标和网络搜索数据作为自变量的研究，因此本文将在此基础上尝试逐步添加历史销量、宏观经济指标、网络搜索关键词，作为自变量建立多变量销量预测模型（Multivar-

iate Sales Forecasting Model，MSFM）并进行效果比对分析。

2 基于梯度提升决策树的销量预测模型构建

本文首先进行变量选择与数据获取，然后对各变量数据进行如主成分分析和标准化等预处理，最后将各变量数据逐步加入梯度提升决策树算法进行训练，本文所构建的销量预测模型的框架如图1所示。

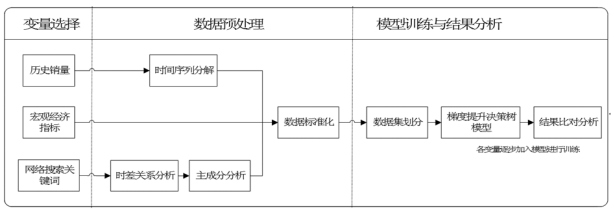


图1 销量预测模型构建框架

2.1 变量选择

本文将尝试在模型中逐步加入历史销量、宏观经济指标、网络搜索数据3种变量进行建模。下面首先对这3方面的变量进行介绍。

2.1.1 历史销量

在时间序列数据的构成要素中，季节变动和循环变动都是序列数据随着季节的变化或者固定型周期而发生的有规律的周期性变动[11]，因此研究历史销量数据的规律变动能够从一定程度上预测未来销量数据的变化趋势。

2.1.2 宏观经济指标

乘用车行业深受经济因素变动的影响，经济因素大多体现一个国家的发展情况和人民的平均生活水平，一个国家经济水平越高，人民的生活水平越高，对乘用车的购买力就越强。本文选取了常用的消费者物价指数（Consumer Price Index, CPI）和92号汽油价格作为影响乘用车销量的宏观经济指标变量[4,8,12]。

2.1.3 网络搜索关键词

近年来，随着互联网和大数据的发展，消费者们在购买商品之前越来越倾向于在网络上通过搜索引擎查找或咨询商品的相关信息和评价，以此为参考来做出进一步的消费行为。所以，网络搜索数据为商品市场发展动向和趋势提供了一定程度的前瞻。消费者在进行网络搜索时，通常会选择自己关注商品的关键词进行搜索，因此现有的网络搜索数据大都是根据搜索引擎中用户对于关键词的搜索量为基础形成的。比如百度指数是指互联网用户对关键词搜索关注程度及持续变化情况，其是以网民在百度的搜索量为数据基础，以关键词为统计对象，科学分析并计算出各个关键词在百度网页搜索中搜索频次的加权[13]。本文也选取百度指数作为销量预测中的自变量之一。

本文销量预测模型中所涉及的变量如表1所示。

表1 销量预测模型涉及变量

变量类型	变量符号	变量描述
因变量	X_t	第t期销量
自变量	X_{t-b}	第$t-b$期销量（b是季节变化周期）
	CPI_t	第t期的消费者物价指数
	$92\#_t$	第t期的92号汽油价格
	$Search_{it}$	第t期的第i个百度指数

2.2 数据预处理

2.2.1 时间序列分解

一个时间序列通常由长期趋势、季节变动和不规则波动部分组成[14]。为了研究销量时间序列的规律变动，本文选择将时间序列进行分解来判断其季节变化周期。关于时间序列的分解模型主要分为加法模型和乘法模型，如公式（1）和公式（2）所示：

$$Value = Trend + Seasonal + Resid \quad (1)$$

$$Value = Trend \times Seasonal \times Resid \quad (2)$$

式中，*Value*为时间序列值；*Trend*为长期趋势值；*Seasonal*为季节变动值；*Resid*为不规则变动值。

加法模型中时间序列的各个组成成分是相互独立的，都有相同的量纲。而乘法模型中输出部分和趋势项有相同的量纲，季节项是比例数，不规则变动项为独立随机变量序列，服从正态分布。通过对时间序列进行分解能够快速找到时间序列的季节或周期变动规律。

2.2.2 时差关系分析

为了对关键词进行筛选，本文采取时差关系分析来研究关键词百度指数数据与乘用车销量之间的相关关系。时差关系分析是利用相关系数验证两组时间序列之间先行、一致或滞后关系的常用方法，其计算公式如公式（3）所示[15]：

$$r_l = \frac{\sum_{t=1}^{n}(x_{t-l}-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{n}(x_{t-l}-\bar{x})^2\sum_{t=1}^{n}(y_t-\bar{y})^2}}, l=0,\pm1,\cdots \quad (3)$$

式中，r_l是两组时间序列的相关系数；l表示先行或滞后阶数，l取负数时表示先行，取正数时表示滞后。在经济预测领域，一般只考虑具有预测作用的先行关键词。

对于r_l的取值，当$r_l>0$时表示两组时间序列呈正相关；当$r_l<0$时表示两组时间序列呈负相关。在

相关程度上，当$0<|r_l|\leqslant 0.3$时认为两组时间序列不存在相关关系或相关性极弱；当$0.3<|r_l|\leqslant 0.5$时认为两组时间序列为低度线性相关；当$0.5<|r_l|\leqslant 0.8$时认为两组时间序列为中度线性相关；当$|r_l|>0.8$时认为两组时间序列为高度线性相关[16]。

2.2.3 主成分分析

由于网络搜索关键词数据多而复杂，且多个相关关键词的百度指数数据之间一般存在多重共线性，于是本文采取主成分分析的方法来对关键词进行合成，以消除数据间的多重共线性。主成分分析是一种利用降维的思想，通过正交变换将一组可能存在相关性的变量转换为一组线性不相关变量的统计方法，转换后的这组变量叫主成分[9]。主成分分析的主要步骤如下：

（1）指标转换

假设进行主成分回归的指标变量有m个，分别为$X_1,X_2,\cdots,X_m$，共有n个样本数据，第i个样本的第j个指标的取值为a_{ij}。将各指标值a_{ij}转换成标准化指标值$\tilde{a}_{ij}$，如公式（4）所示：

$$\tilde{a}_{ij}=\frac{a_{ij}-u_j}{s_j},1\leqslant i\leqslant n,1\leqslant j\leqslant m \tag{4}$$

式中，u_j，s_j分别为第j个指标的样本均值和样本标准差。对应地称$\tilde{X}_j$为标准化指标变量。如公式（5）所示：

$$\tilde{X}_j=\frac{X_j-u_j}{s_j},1\leqslant j\leqslant m \tag{5}$$

（2）计算相关系数矩阵

相关系数矩阵$R=(r_{ij})_{m*m}$，式中对角线元素$r_{ii}=1$，$r_{ij}=r_{ji}$是第i个指标与第j个指标的相关系数，如公式（6）所示：

$$r_{ij}=\frac{\sum_{k=1}^{n}\tilde{a}_{ki}*\tilde{a}_{kj}}{n-1},1\leqslant i\leqslant m,1\leqslant j\leqslant m \tag{6}$$

（3）计算特征值与特征向量

计算相关系数矩阵R的特征值$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$时对应的特征向量$\mu_1,\mu_2,\cdots,\mu_m$，其中$\mu_j=[\mu_{1j},\mu_{2j},\cdots,\mu_{mj}]^T$，由特征向量组成$m$个新的指标变量，如公式（7）所示：

$$\begin{cases}Y_1=\mu_{11}\tilde{X}_1+\mu_{21}\tilde{X}_2+\cdots+\mu_{m1}\tilde{X}_m\\Y_2=\mu_{12}\tilde{X}_1+\mu_{22}\tilde{X}_2+\cdots+\mu_{m2}\tilde{X}_m\\\cdots\\Y_m=\mu_{1m}\tilde{X}_1+\mu_{2m}\tilde{X}_2+\cdots+\mu_{mm}\tilde{X}_m\end{cases} \tag{7}$$

式中，Y_1是第一个主成分；Y_2是第二主成分；…，Y_m是第m主成分。

（4）计算主成分贡献率及累计贡献率

根据公式（8）和公式（9），称b_j为主成分Y_j的信息贡献率，α_p为主成分$Y_1,Y_2,\cdots,Y_P$的累积贡献率，当α_p接近于1时，则选择前p个指标变量$Y_1,Y_2,\cdots,Y_P$作为p个主成分，代替原来m个指标变量，从而可对p个主成分进行综合分析（$p\leqslant m$）。

$$b_j=\frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k},1\leqslant j\leqslant m \tag{8}$$

$$\alpha_p=\frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k} \tag{9}$$

2.2.4 数据标准化

本文选取使用的梯度提升决策树（Gradient Boosting Decision Tree, GDBT）算法与梯度有关，为了避免预测结果向数值大的变量倾斜，需要对数据进行标准化处理，如公式（10）所示[17]：

$$x_{ij}^{'}=\frac{x_{ij}-\mu}{\sigma} \tag{10}$$

式中，$x_{ij}^{'}$是标准化后的数据；x_{ij}是原始数据；μ是第j个指标数据的平均值；σ是第j个指报数据的标准差。

2.3 模型训练与结果分析

2.3.1 数据集划分

机器学习算法在训练前一般需要将数据集划分成独立的3部分，即训练集、测试集和验证集。其中训练集用于训练模型，验证集用于评估模型，协助调整模型超参数，测试集则是用于检验最终选择的最优模型的性能表现。

2.3.2 梯度提升决策树

梯度提升决策树算法可以灵活处理各种类型的数据，且其模型一般具有较好的解释性和鲁棒性，因此本文选择梯度提升决策树算法为基础来进行乘用车销量预测研究。

梯度提升决策树（GDBT）是Friedman在2001年提出的一种迭代学习的决策树算法[18]。GDBT采用的是集成学习中Boosting算法的基本思想，即首先使用初始权重从训练集中训练出一个弱学习器，根据弱学习器的学习误差率来更新样本的权重，提高之前弱学习器学习率较高的训练样本点权重，使得这些误差率高的样本在后面的弱学习器中得到更多的重视。如此循环，直到得到指定数量的学习器，再结合策略进行整合，得到最终的强学习器。GBDT中弱分类器的形

式就是各棵决策树。总体上，GBDT是通过采用加法模型（即基函数的线性组合）以及不断减小训练过程产生的残差来达到将数据分类或者回归的算法。

2.3.3 结果比对分析

为了验证各个变量对销量预测的影响，本文在GDBT算法基础上，分别构建只考虑历史销量的模型1，同时考虑历史销量和宏观经济指标的模型2，同时考虑历史销量、宏观经济指标和网络搜索关键词的模型3来进行训练（如式（11）至式（13）所示）并进行结果比对分析。

$$\text{模型1：}X_t=f(X_{t-b}) \tag{11}$$

$$\text{模型2：}X_t=f(X_{t-b},CPI_t,92\#_t) \tag{12}$$

$$\text{模型3：}X_t=f(X_{t-b},CPI_t,92\#_t,Se\,\text{arch}_{it}) \tag{13}$$

在模型评价指标的选择上，由于平均绝对百分误差（Mean Absolute Percentage Error，MAPE）数值是百分比的形式，更能形象地表达出误差值，易于解释，因此本文选取MAPE来对不同模型进行效果比较[19]，其计算公式如公式（14）所示：

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{14}$$

式中，$\hat{y}_i$是t时刻的预测值；y_i是t时刻的真实值。理论上，MAPE的值越小，说明预测模型拟合效果越好，具有更好的精确度。

3 实证研究与结果分析

3.1 数据获取与预处理

3.1.1 销量数据

本文从车主之家官网获取了乘用车市场2013年1月至2021年12月共72个月的销量数据，其时序图如图2所示。

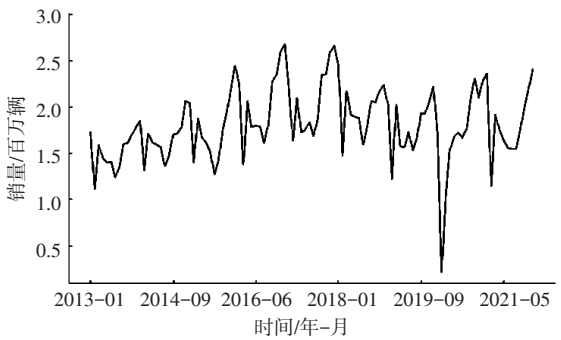
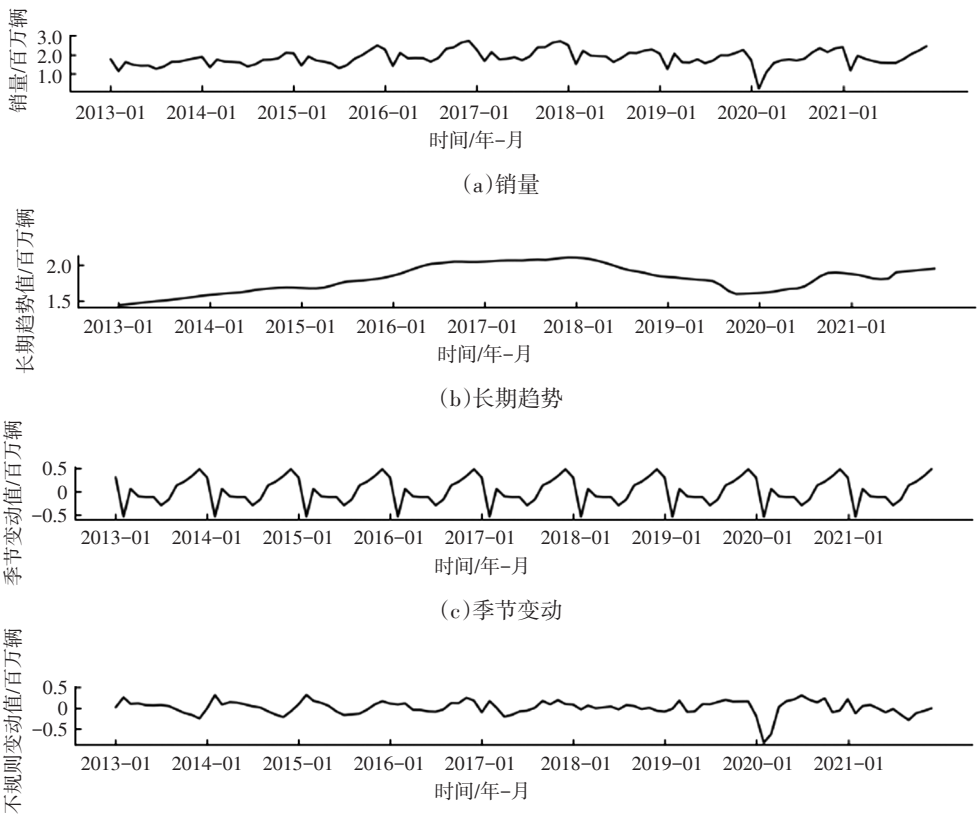


图2 乘用车销量时序图

由图2所展示的时序图可以初步认为乘用车的月度销量数据有一定的周期性变化，接下来对其进行分解来判断其季节变化周期。本文采用Python的statesmodel库中的seasonal_decompose函数来对乘用车的销量数据进行时间序列分解，加法模型和乘法模型的结果分别如图3、图4所示。



（a）销量

（b）长期趋势

（c）季节变动

（d）不规则变动

图3 乘用车销量序列分解（加法模型）

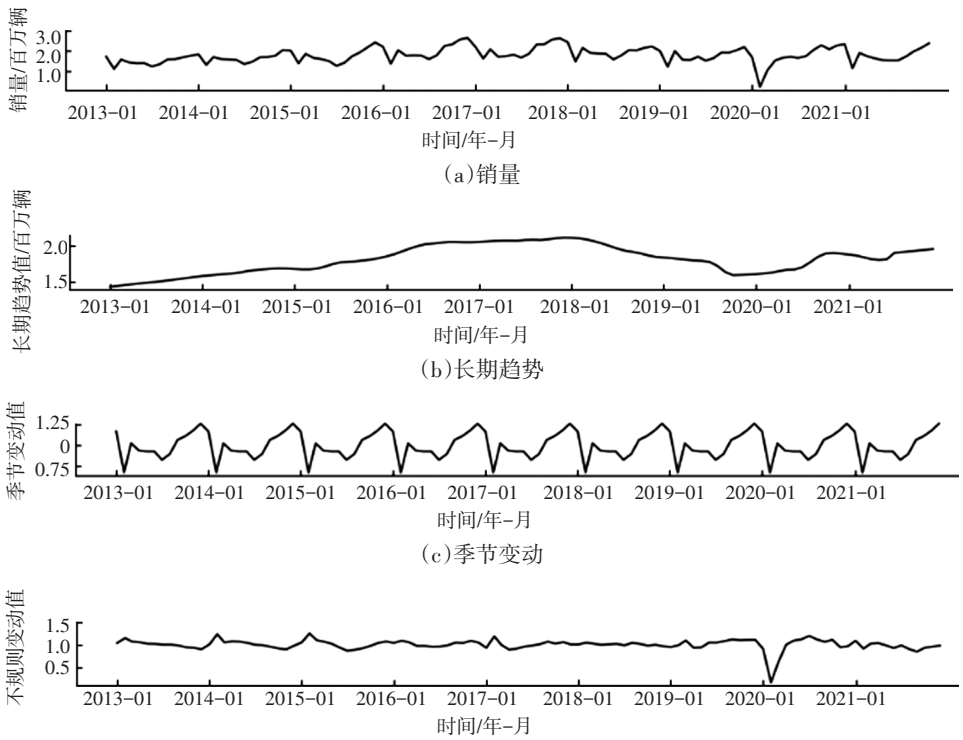


图4　乘用车销量序列分解(乘法模型)

由图3和图4所展示的时间序列分解结果可知,乘用车销量数据的季节变化周期均为12个月,故选择将去年同期的销量数据作为模型中的变量之一。

3.1.2　宏观经济数据

本文从东方财富网上通过网络爬虫获取了2014年1月至2021年12月的宏观经济数据,包括消费者物价指数(CPI)和92号汽油价格,其部分数据如表2所示。

表2　宏观经济样本数据表(部分数据)

时间	CPI	92号汽油价格/元·L^{-1}
2021年1月	99.7	5.95
2021年2月	99.8	6.16
2021年3月	100.4	6.69
2021年4月	100.9	6.6
2021年5月	101.3	6.76
2021年6月	101.1	6.9

注:根据东方财富网整理

3.1.3　网络搜索关键词数据

由于受到不同消费者的语言习惯、搜索风格主观因素的影响,网络搜索关键词库变得庞大而复杂,因此构建合适的关键词组就成了研究的关键。

(1)关键词的获取

本文首先分别以“乘用车”作为初始核心关键词,然后利用百度指数需求图谱与相关词热度模块来进行两轮关键词的拓展,同时去除搜索指数较低以及明显与乘用车无显著关系的关键词,最终得到的关键词为86个,之后再根据选取的关键词在百度指数官网上通过网络爬虫获取到百度指数数值。采集到的部分数据如表3所示。

表3　网络搜索关键词百度指数数据表(部分数据)

时间	乘用车	商用车	二手汽车市场	汽车油耗
2021年1月	9 211	9 721	547	7 504
2021年2月	6 502	7 892	678	6 767
2021年3月	9 111	10 551	1 810	8 292
2021年4月	8 703	9 560	1 392	8 608
2021年5月	8 525	9 343	2 607	7 428
2021年6月	7 953	8 864	1 992	6 605

注:根据百度指数整理

(2)关键词的筛选

本文采取时差关系分析来研究关键词百度指数数据与乘用车销量之间的相关关系,从而对关键词进行筛选。首先取$l\in[-6,0)$(即先行阶数取1至6),在每个关键词百度指数数据中选择相关程度为低度相关及以上(即相关系数$|r_l|>0.3$)且相关系数值最大的先行阶数[16],最终筛选得到的关键词为23个,部分结果如表4所示。

根据时差关系分析结果,将关键词按照先行阶数与销量数据进行对应,得到关键词筛选后的百度指数数据。

表4　关键词百度指数数据时差关系分析结果（部分数据）

关键词	相关系数	先行阶数/阶	关键词	相关系数	先行阶数/阶
乘用车	0.381 222	6	商用车	0.330 239	1
汽车保险费用	0.386 819	1	汽车油耗	0.389 124	1
宝腾	0.408 190	5	电动汽车	0.319 744	1
一汽大众公司	0.388 478	1	家用车推荐	0.355 112	6
二手汽车市场	0.387 466	6	汽车工业协会	0.300 430	1

（3）关键词的合成

将经过预处理的百度指数数据运用Python的sklearn库中的PCA()函数进行主成分分析，结果如表5所示。

表5　百度指数数据主成分分析结果

	特征方差占比	累积占比
第1主成分	0.685 430	0.685 430
第2主成分	0.260 600	0.946 030
第3主成分	0.033 656	0.979 686
……	……	……

由表5的结果可以得出，经过主成分分析降维后，前3个主成分的特征方差累计占比超过95%，包含了原数据的大部分信息，因此考虑将百度指数数据降维到3维。

3.1.4　数据标准化

本文使用Python的sklearn库中的StandardScaler()函数对以上获取并预处理过的销量数据、宏观经济数据和百度指数数据进行标准化处理。

3.2　模型训练

对于数据集的划分，本文选取2014年1月至2020年12月的数据作为训练集和验证集，选取2021年1月至12月的数据作为测试集。

模型训练方面，本文使用Python的sklearn库中的GradientBoostingRegressor()函数进行GDBT算法的建模。对于模型中的超参数选择，选择用网格搜索的方法进行确定，即在指定的参数范围内，按步长依次调整参数，利用调整的参数训练学习器，从所有的参数中找到在验证集上精度最高的参数[20]。同时考虑到不同验证集对模型结果的影响不同，对训练集和验证集采取7折交叉验证的方式进行建模。本文使用sklearn库中的GridSearchCV()函数遍历多种超参数组合并通过交叉验证确定最优效果参数。

3.3　模型结果比对分析

在GDBT算法中逐步加入历史销量数据、宏观经济数据、网络搜索关键词数据所得到的各模型对乘用车2021年1月至12月的预测结果MAPE指标如表6所示。

表6　GDBT模型预测的MAPE指标结果

模型	使用变量	MAPE/%
模型1	历史销量	14.37
模型2	历史销量、宏观经济指标	12.71
模型3	历史销量、宏观经济指标和网络搜索关键词	10.35

由表6所展示的预测结果可知，只使用历史销量数据进行建模时，模型MAPE值为14.37%。添加了宏观经济指标进行建模后，MAPE值有所降低，为12.71%，说明宏观经济指标对销量有一定的影响。当在模型中同时考虑历史销量、宏观经济指标和网络搜索关键词变量时，MAPE值进一步降低，模型效果进一步提升，表明了网络搜索数据有助于乘用车销量的预测。

在时间序列的相关预测问题中，季节性差分自回归滑动平均模型（Seasonal Auto-Regressive Moving Average Model，SARIMA）、支持向量机回归（Support Vector Regression，SVR）模型也得到了广泛应用[2,21]，为了证明本文所使用模型的有效性，进一步对比了同时考虑3个变量时GDBT模型与SARIMA模型、SVR模型的预测结果MAPE指标，结果如表7所示。

表7　不同模型预测的MAPE对比　%

	SARIMA模型	SVR模型	GDBT模型
MAPE	17.72	16.07	10.35

由表7可知，GDBT模型的MAPE指标比SARIMA模型、SVR模型都低，表明GDBT模型的拟合效果最优。

综合上述分析可以得出，同时使用历史销量、宏观经济指标、网络关键词搜索数据变量会使销量预测模型的精度提高，本文中得到的最优模型是GDBT模型，其MAPE值结果为10.35%，能够较好的预测销量变化。

4　结论与展望

4.1　研究结论

本文以乘用车整体市场为研究对象，在现有研究的基础上分别建立了只考虑历史销量的预测模型，同时考虑历史销量和宏观经济指标的预测模型，同时考虑历史销量、宏观经济指标和网络搜索关键词的预测模型，最终将建立的模型进行对比分析，发现同时考虑了3种变量的梯度提升决策树模型的MAPE值最小，说明其拟合效果最好，也表明了宏观经济指标和网络搜索数据有助于乘用车销量预测。

本文在理论层面以乘用车整体市场为研究对象，使用历史销量、宏观经济指标和网络搜索关键词数据作为影响乘用车销量的变量，得出同时使用上述3种变量的梯度提升决策树模型为最佳模型，为销量预测的研究提供一种新的参考模型。在实践层面，对于乘用车市场中已有一定渗透率的企业，可以基于本文得出的最优销量模型对市场趋势发展情况做参考，从而进行针对性的生产计划安排与优化以及市场营销战略的制定。

4.2 研究不足与展望

（1）本文在销量预测中所考虑的影响乘用车销量的变量仍然不完备，后续研究中可以继续发掘其他影响因素进行综合考虑，例如宏观经济环境、政策因素、零部件供给端相关因素和消费者评价因素，来进一步提高模型的预测精确度。

（2）本文在销量预测中仅应用了单一模型，后续研究中可以进一步将各类预测模型进行融合来建模并分析比对效果，得到表现更优的模型。

（3）本文是以乘用车整体市场为研究对象进行预测，未来对新能源汽车的销量预测也是重要的研究方向。

参 考 文 献

[1] 李响, 宗群, 童玲. 汽车销售混合预测方法研究[J]. 天津大学学报(社会科学版), 2006(3): 175-178.

[2] 王旭天, 李政远, 舒慧生. 基于SARIMA的我国汽车销量预测分析[J]. 中国市场, 2016(1): 71-74.

[3] 王栋. 基于灰色关联和BP神经网络的汽车保有量预测[J].计算技术与自动化, 2015, 34(1): 29-33.

[4] GAO J J, XIE Y N, CUI X M, et al. Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model[J]. Advances in Mechanical Engineering, 2018, 10(2): 168781401774932.

[5] 王易, 赵佳, 张帆. 基于多源异构和XGBOOST模型的销量预测[J]. 中国汽车, 2021(1): 9-15.

[6] 袁庆玉, 彭赓, 刘颖, 等. 基于网络关键词搜索数据的汽车销量预测研究[J]. 管理学家(学术版), 2011(1): 12-24.

[7] 李忆, 文瑞, 杨立成. 网络搜索指数与汽车销量关系研究——基于文本挖掘的关键词获取[J]. 现代情报, 2016, 36(8): 131-136+177.

[8] ZHANG Y, ZHONG M E, GENG N N, et al. Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China.[J]. PLOS ONE, 2017, 12(5): e0176729.

[9] 刘吉华, 张梦迪. 基于百度指数的大众汽车销量预测研究[J]. 统计与管理, 2020, 35(10): 23-31.

[10] 刘吉华, 张梦迪, 彭红霞, 等. 基于卷积神经网络的汽车销量预测模型[J]. 计算机科学, 2021, 48(S1): 178-183+189.

[11] 赖慧慧. 大数据背景下基于ARMA模型的增值税销项税额预测[J]. 税务研究, 2019(2): 41-46.

[12] FANTAZZINI D, TOKTAMYSOVA Z. Forecasting German car sales using Google data and multivariate models[J]. International Journal of Production Economics, 2015, 170(12): 97-135.

[13] 周荣庭, 何同亮, 李佳艺, 等. 中国区块链发展的控制路径[J]. 科技管理研究, 2021, 41(24): 27-34.

[14] 赵家波, 游晓明, 刘升. 结合价格波动策略与动态回溯机制的蚁群算法[J]. 计算机科学与探索, 2022, 16(6): 1390-1404.

[15] 张玲玲, 张笑, 崔怡雯. 基于聚类方法的百度搜索指数关键词优化及客流量预测研究[J]. 管理评论, 2018, 30(8): 126-137.

[16] 刘晨阳. 基于百度指数的汽车销量预测研究[D]. 武汉：湖北大学, 2018.

[17] 朱道平, 张灿凤. 考虑不同平台评论情绪的电商产品销量预测研究[J]. 市场周刊, 2021, 34(3): 91-93.

[18] FRIEDMAN J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

[19] 刘业政, 章旭, 王锦坤. 考虑品牌情感的汽车销量预测模型[J]. 合肥工业大学学报(自然科学版), 2017, 40(9): 1276-1282.

[20] 杜帆, 李立国.中国博士生教育规模增长预测分析——基于1996—2018年省际面板数据的实证研究[J]. 学位与研究生教育, 2020(6): 55-63.

[21] 陈荣, 梁昌勇, 谢福伟. 基于SVR的非线性时间序列预测方法应用综述[J]. 合肥工业大学学报(自然科学版), 2013, 36(3): 369-374.

【作者简介】

段昊江，同济大学经济与管理学院，硕士研究生，研究方向：商务智能。

E-mail：2230398@tongji.edu.cn

吴冰，同济大学经济与管理学院，副教授，硕士生导师，研究方向：商务智能、社交媒体。

E-mail：wubingsem@tongji.edu.cn