Modern Management 现代管理, 2024, 14(4), 830-844
Published Online April 2024 in Hans. https://www.hanspub.org/journal/mm
https://doi.org/10.12677/mm.2024.144095

Hans汉斯

预测中国GDP增长率：基于R语言和机器学习的分析

左晨睿

上海外国语大学国际金融贸易学院，上海

收稿日期：2024年3月8日；录用日期：2024年3月14日；发布日期：2024年4月30日

摘　要

本文旨在通过运用R语言和机器学习技术，包括多元线性回归和随机森林模型，对中国的GDP增长率进行预测。研究探讨了GDP增长率对中国宏观经济政策和商业策略的重要性，进一步探讨了选取合适的预测模型。多元回归模型旨在探究各经济指标对GDP变动的影响，而随机森林模型则用于捕捉数据间的复杂非线性关系，并通过构建多个决策树提高预测的准确性。多元回归模型显示固定资产和工业是影响中国GDP增长率的显著因素，而CPI和贸易平衡的影响不显著。随机森林模型则强调了固定资产和工业在预测中国GDP增长率中的重要性。最后，论文指出了固定资产和工业对GDP增长的显著影响，并讨论了研究中存在的问题和未来的研究方向。

关键词

中国GDP增长率，随机森林，多元回归模型，固定资产，机器学习模型

Forecasting China’s GDP Growth Rate: An Analysis Based on R Language and Machine Learning

Chenrui Zuo

School of Economics and Finance, Shanghai International Studies University, Shanghai

Received: Mar. 8th, 2024; accepted: Mar. 14th, 2024; published: Apr. 30th, 2024

Abstract

This paper aims to forecast China’s GDP growth rate by using R language and machine learning techniques, including multiple linear regression and random forest model. This paper discusses the

文章引用: 左晨睿. 预测中国 GDP 增长率: 基于 R 语言和机器学习的分析[J]. 现代管理, 2024, 14(4): 830-844.
DOI: 10.12677/mm.2024.144095

importance of GDP growth rate to China's macroeconomic policy and business strategy, and further discusses the selection of appropriate forecasting models. The multiple regression model aims to explore the impact of various economic indicators on GDP changes, while the random forest model is used to capture complex nonlinear relationships between data and improve the accuracy of prediction by constructing multiple decision trees. The multiple regression model shows that fixed assets and industry are significant factors affecting China's GDP growth rate, while CPI and trade balance have no significant effects. The random forest model emphasizes the importance of fixed assets and industry in predicting China's GDP growth rate. Finally, the paper points out the significant impact of fixed assets and industry on GDP growth, and discusses the existing problems and future research directions.

Keywords

China's GDP Growth Rate, Random Forest, Multiple Regression Model, Fixed Assets, Machine Learning Model

1. 研究背景及意义

中国 GDP 增长率作为衡量国家经济整体健康和发展速度的重要指标，对经济政策和商业策略产生深远影响。GDP 增长率不仅反映了国家经济的活力，也是政府制定宏观经济政策、企业制定商业战略的关键依据[1]。

首先，从宏观经济政策的角度来看，GDP 增长率直接影响政府的政策制定。当 GDP 增长放缓时，政府可能会实施刺激性的财政和货币政策，如增加政府支出、减税、降低利率等，以促进经济增长[2]。相反，如果 GDP 增长过快，可能带来通货膨胀等问题，政府则可能采取紧缩政策，如提高利率和减少政府支出，以避免经济过热。GDP 增长率还是政府评估其发展战略和政策效果的重要指标，有助于调整和优化长期发展规划[3]。

在商业策略方面，中国 GDP 增长率的变化对企业决策有着重要影响。经济增长往往伴随着消费能力的提升，为企业提供了扩大市场和增加销售的机会。在经济增长期，企业可能会增加投资、扩大生产规模、开发新产品或进入新市场。此外，GDP 增长率的变化也会影响消费者的信心和购买行为，进而影响企业的营销策略和产品定位[4]。

然而，GDP 增长放缓可能意味着市场需求减弱、投资回报率下降，企业可能需要调整策略，如削减成本、优化运营效率或转型升级。在这种情况下，企业可能更注重风险管理和现金流稳定，以应对潜在的经济不确定性和挑战[5]。

此外，中国作为全球第二大经济体，其 GDP 增长率的变化还会对全球经济产生影响。例如，中国经济增长放缓可能会减少对原材料和能源的需求，影响全球商品市场和贸易流。相反，强劲的 GDP 增长可以带动全球经济增长，为国际企业提供更多机会。

总之，中国 GDP 增长率的变化对国家的宏观经济政策和企业的商业策略都有着重要的指导意义。了解和预测 GDP 增长趋势有助于政府和企业更好地制定策略，应对经济变化带来的机遇和挑战。

本文的主要目标是运用 R 语言结合先进的机器学习技术来预测中国的 GDP 增长率。在经济全球化和国内外多重因素的共同作用下，中国经济呈现出复杂多变的趋势，对中国 GDP 增长率的准确预测对于宏

观经济政策制定、企业战略规划乃至国际投资决策具有重要的指导意义。通过多种经济指标和影响因素，本论文旨在构建一个相对高效、准确的预测模型，提供较为科学的数据支持和参考。

运用 R 语言进行数据处理和分析的选择是因为 R 语言在统计分析和图形表示方面的强大能力，以及其在学术和工业界的广泛应用。结合机器学习技术，特别是时间序列分析、回归模型和其他预测算法，本文将探索如何更有效地利用历史数据来预测 GDP 的未来走势[6]。

最终，希望通过本文提供一个具有实际应用价值的工具，不仅可以提高对中国经济发展趋势的认识，还可以为相关经济和商业决策提供数据驱动的见解。这不仅是一个技术挑战，更是对现代数据科学在经济领域应用的一次实践探索。

本文的预期成果是开发一个基于 R 语言和机器学习技术的较为高效、可靠的模型，用于预测中国 GDP 增长率。这个模型将能够较为准确地分析和预测中国经济的增长趋势，预期的分析结果将不仅能够提供未来一定时期内的 GDP 增长率预测，还能够揭示影响中国经济增长的关键因素和它们之间的相互作用。

这些成果在多个领域具有重要的应用价值。首先，对政策制定者来说，准确的 GDP 增长预测能够帮助政府更好地制定和调整宏观经济政策，如财政、货币政策和投资规划。其次，对于经济分析师和研究机构，这一模型提供了一个强大的工具，用于深入分析经济趋势和制定经济预测报告。对企业决策者而言，这一预测结果能够帮助他们在市场策略、投资规划和风险管理方面做出更为明智的决策。

此外，这一研究的成果还将对学术界有所贡献，特别是在应用数据科学技术进行经济预测的领域。通过实际案例分析，该研究将提供有关如何有效整合传统统计方法和现代机器学习技术进行经济数据分析的深刻见解。最终，我们希望这个研究能够成为连接数据科学和经济学的桥梁，为中国乃至全球的经济分析和预测实践带来新的视角和方法。

2. 文献综述

2.1. 中国 GDP 的预测意义及方法

GDP 预测一直是宏观经济研究的核心领域，因为 GDP 作为衡量一个国家总体经济活动的关键指标，对于政策制定、投资决策以及经济发展规划都有重要意义。在这方面，学术界进行了大量的研究，以期发展更为精确和可靠的预测模型。Williams (2023) [7]认为，GDP 预测对于政策制定者至关重要，因为它们能够指导货币政策、财政政策和社会发展规划。例如，准确的预测可以帮助政府制定刺激计划，以应对经济衰退，或者在经济过热时采取适当的冷却措施。对金融市场的参与者来说，Turner 等(2022) [8]认为，GDP 预测是制定投资策略的关键因素之一。它可以帮助投资者评估市场趋势、资产配置以及风险管理策略。

GDP 增长率作为经济发展的核心指标，对于政策制定者来说，它是衡量国家经济健康状况的重要工具。GDP 数据的增长与否可以反映出一个国家或地区经济扩张或收缩的情况，对于制定宏观经济政策至关重要。在货币政策方面，Bernanke (2005) [9]和 Yellen (2009) [10]指出，GDP 的增长情况直接影响中央银行的利率决策。如果 GDP 增长率超出预期，表明经济可能过热，政策制定者可能会提高利率以避免通货膨胀。相反，如果 GDP 增长缓慢或经济萎缩，中央银行可能会降低利率以刺激投资和消费，从而促进经济增长。财政政策方面，GDP 数据是政府预算编制和财政支出决策的依据。Keynes (1936) [11]认为，政府在面对 GDP 增长放缓时，可能会增加公共支出，通过基础设施投资和公共服务提升来刺激经济活动。同样，GDP 强劲增长可能促使政府采取减税措施，以避免经济过热并留出空间应对未来的经济下行风险。在社会福利政策制定上，GDP 增长率可以指示政府在教育、卫生和社会安全网等领域的投入是否足够。Stiglitz (2002) [12]认为，一个持续增长的 GDP 为政府提供了更大的财政空间来改善民生和提升社会整体的生活水平。

在对国家 GDP 增长率预测方法中，我们可以参考以下两位学者的研究：段树乔等(2023) [13]在其研究中提出，GDP 总量的单调增长型特性是国家社会和经济稳定的重要标志。张正华认为，非线性多项式是描述和拟合这种增长型特性的有效工具，并且可以通过这种方法构建出精确的未来 GDP 总量预测模型。通过运用非线性多项式拟合优化，张正华的研究强调了在预测模型中考虑 GDP 增长的非线性特征的重要性。李斌(2022) [14]的研究则采用了时间序列分析的理论框架，以新冠疫情期间中美两国的宏观经济为研究对象。李斌利用 ARIMA 模型对两国宏观经济单指标的联动和影响关系进行了深入分析，并进一步运用 VAR 模型对宏观经济的多指标联动进行了实证研究。李斌的研究不仅揭示了中美宏观经济指标间的因果关系，而且预测了两国未来宏观经济的发展趋势和联动关系，提供了对国家宏观经济策略制定的有价值的见解。这两项研究都表明，虽然预测方法和模型的选择可以多样化，但关键在于选取能够准确捕捉经济现象关键特征的模型。无论是张正华通过非线性多项式对 GDP 增长率的单调性进行建模，还是李斌运用时间序列模型深入分析宏观经济的动态关系，这些方法都为理解经济现象提供了重要的分析工具，并对未来的经济发展趋势提供了预测。这些研究成果对于宏观经济分析和政策制定具有重要的指导意义。

2.2. GDP 衡量标准

国内生产总值(GDP)作为衡量一个国家经济规模和经济活动的标准指标，其重要性在经济学和政策分析中已被广泛认可。GDP 衡量标准及其应用的文献综述涵盖了 GDP 的不同计算方法、它的局限性以及对经济发展的影响评估。

关于 GDP 的计算方法，Samuelson 和 Nordhaus (2010) [15]在文中提到经典的 GDP 衡量包括三种方法：生产方法、收入方法和支出方法。这些方法在不同国家和不同时期被不同地使用，并被不断修改以更准确地反映经济活动。陈国青等(2020) [16]指出，近年来，学者们也提出了新的方法，如使用大数据和实时数据来补充传统的 GDP 计算。

3. R 语言和机器学习基础

R 语言是一种广泛用于统计计算和图形展示的编程语言和软件环境。自 20 世纪 90 年代初由新西兰的 Robert Gentleman 和 Ross Ihaka 开发以来，它已经成为数据分析、统计建模、科学研究、可视化等领域的首选工具之一[17]。

机器学习，作为人工智能的一个关键分支，近年来在经济预测领域中的应用日益增多，成为一种强大的工具，用于解析复杂的经济数据和预测经济趋势。它通过从大量历史数据中学习，能够揭示数据之间的隐藏模式和关系，从而对未来的经济状况做出预测。

在预测中国 GDP 增长率的项目中，选择合适的机器学习模型至关重要。这些模型应能够处理时间序列数据的特性，如趋势、季节性和周期性，同时捕捉经济指标之间的复杂关系。以下是几种适用于预测 GDP 增长率的主要机器学习模型：

3.1. 时间序列分析模型

ARIMA (自回归积分滑动平均模型)：适用于分析和预测时间序列数据。ARIMA 模型能够处理时间序列的非平稳性，并且可以通过自回归和移动平均组件来模拟数据中的时间依赖性。

季节性 ARIMA (SARIMA)：在 ARIMA 的基础上增加了季节性组件，特别适用于具有明显季节性模式的经济数据。

3.2. 回归分析

多元线性回归：可以用来分析多个预测变量与 GDP 增长率之间的线性关系。这种模型简单直观，易

于解释，适用于初始分析[18]。

岭回归和套索回归：当预测变量数量很多且存在多重共线性时，这些正则化的线性回归方法可以提高模型的预测能力和稳定性。

3.3. 机器学习算法

随机森林[19]：作为一种集成学习方法，随机森林通过构建多个决策树并结合它们的预测结果来提高准确性，适合处理复杂的非线性关系[20]。

支持向量机(SVM)：SVM 能够有效处理高维数据，尤其适用于发现数据之间复杂的非线性关系。

3.4. 先进的机器学习模型

神经网络和深度学习：这些模型具有强大的数据拟合和特征学习能力，适合于捕捉经济数据中的复杂模式和动态关系。虽然这些模型通常需要大量的数据，但它们在提取数据深层次特征方面表现出色。

本文将通过多元线性回归、随机森林学习模型对中国 GDP 增长率进行简单预测。

4. 数据收集与预处理

4.1. 数据收集

影响中国 GDP 增长率的因素多样且复杂，涉及国内外经济、政治、社会等多个方面，包括政府政策、全球经济环境、国内消费和投资、技术创新和产业升级、人口和劳动力市场、环境因素和资源约束、国内外政治稳定、金融市场和货币政策等[21]。为了简化本文研究的模型，本文主要选取 3~4 个相对重要的经济指标进行估值预测中国 GDP 增长率，包括工业产出和制造业指数、消费者价格指数(CPI)、固定资产投资、贸易平衡四个指标进行测度。

主要原因如下：第一，工业产出和制造业指数作为经济活动的重要组成部分，工业产出和制造业指数直接反映了国家的生产能力和工业活力。这些指标对 GDP 增长有显著影响，尤其是在以制造业为主导的经济体中。第二，消费者价格指数(CPI)是衡量通货膨胀水平的关键指标，反映了消费者购买力和市场需求的变化。GDP 增长与消费者购买力和消费信心密切相关。第三，固定资产投资反映了国内基础设施建设和资本投入的规模，是推动经济增长的关键因素。在中国，政府和企业的投资项目对经济增长起到了重要的推动作用。第四，贸易平衡(出口与进口差额)。中国作为全球贸易的主要参与者，其出口和进口的状况能够反映国际市场需求和全球经济状况对中国经济的影响。贸易平衡的变化直接关系到外部需求对 GDP 的贡献。

因此，对于上述的数据本文进行了收集，数据主要来源于 Wind 数据库和国际货币基金组织(IMF)。

4.2. 数据清洗和预处理方法

本文选取了 1995~2019 年的数据，80 年代之前我国的市场经济发展得并不完善，因此进行研究无实际意义，2020 至 2022 年处于疫情时代，经济增长波动幅度较大，尤其是进出口受到非常大的影响，不利于研究结论的得出。如图 1 所示，1995 年前我国的汇率变化幅度较大，因此不利于研究[22]。

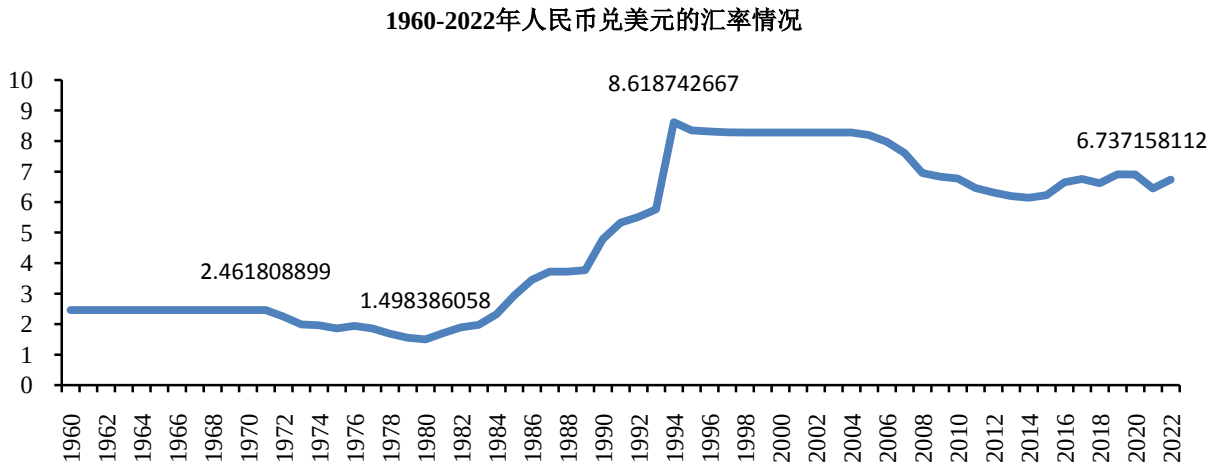
因此，本文对数据在 R 中进行了清洗和预处理，步骤如下：

4.2.1. 数据导入

使用 read.csv、read.table、readxl 等函数根据数据格式导入数据。

数据检查：

使用 head()、tail()、summary()、str()等函数查看数据的基本结构和摘要。



数据来源：世界银行。

Figure 1. Changes in Yuan-Dollar exchange rate from 1960 to 2022
图 1. 1960~2022 年人民币兑美元汇率变化情况

4.2.2. 处理缺失值

使用 is.na()识别缺失值。

根据情况选择填充缺失值(例如使用均值、中位数、或前后值)，或删除含有缺失值的行/列。

4.2.3. 数据类型转换

确保所有变量的数据类型正确(如数字、字符、日期等)，使用 as.numeric()、as.factor()、as.Date()等函数进行转换。

4.2.4. 去除异常值

根据数据的统计摘要或可视化分析，识别并处理异常值。

4.2.5. 数据标准化/归一化(如果需要)

对于某些机器学习模型，可能需要将数据标准化或归一化，可以使用 scale()函数。

4.2.6. 创建时间序列对象(针对时间序列分析)

如果进行时间序列分析，使用 ts()创建时间序列对象。

4.2.7. 创建与修改

根据需要创建新的变量或修改现有变量，例如，构建滞后变量、季节性变量等。

4.2.8. 数据分割

将数据集分割为训练集和测试集，用于模型训练和评估。

5. 模型建立和实施

根据之前的介绍，本文旨在通过应用多元线性回归、随机森林统计机器学习模型，对中国的 GDP 增长率进行较为详细的分析和预测。

5.1. 多元线性回归

首先采用多元回归模型作为起点，对中国 GDP 增长率进行初步预测分析，旨在探究各经济指标对 GDP 变动的影响。

5.1.1. 选取原因

选择多元回归模型对中国 GDP 增长率进行预测有几个原因[23]：首先，多元回归是建立在统计学和经济学理论基础上的，能够处理多个自变量对因变量的影响，这对于分析和理解影响 GDP 增长率的多个经济因素特别有用。其次，多元回归模型的结果容易解释。它提供了关于各个自变量如何影响因变量的量化信息，这对于制定和调整经济政策至关重要。再者，多元回归分析在经济学中广泛应用于预测和趋势分析，它提供了一个熟悉和广泛接受的工具来研究 GDP 增长。另外，与更复杂的模型相比，多元回归对数据的质量和数量的要求相对较低，这使得在有限或不完全数据的情况下成为一种实用的选择。最后，多元回归常常作为分析的起点，识别哪些变量值得进一步探究。它为后续使用更复杂模型(如随机森林或神经网络)提供了基础和方向。

5.1.2. 模型建立

$$GDP\_Growth = \beta_0 + \beta_1 Fixed\_Assets + \beta_2\, CPI + \beta_3 Industry + \beta_4 Trade\_Balance + \xi$$

其中，Fixed_Assets 表示中国：全社会固定资产投资完成额，CPI 表示中国居民消费指数，Industry 表示中国：GDP 指数：第二产业：工业，Trade_Balance 表示中国:进出口差额:一般贸易，β_0、β_1、β_2、β_3、β_4 分别为 Fixed_Assets、CPI、Industry、Trade_Balance 的系数，ξ 为扰动项。

5.2. 随机森林

5.2.1. 选取原因

选择随机森林模型对中国 GDP 增长率进行预测有几个原因：首先，随机森林能够捕捉数据间的复杂非线性关系，这对于预测经济指标如 GDP 增长率，尤其是在其受多种因素和潜在相互作用影响的情况下，非常有用。其次，随机森林通过构建多个决策树并对它们的结果进行平均，降低了模型过拟合的风险，使预测更加稳健和可靠。再者，随机森林能够提供关于各个特征对预测结果影响大小的量化评估，这有助于识别哪些经济指标对中国 GDP 增长率影响最大，从而为政策制定和优先排序提供依据。除此之外，随机森林模型可以处理各种类型的数据(包括分类和连续变量)，并且能够轻松地扩展以包含更多的变量和数据，这使得模型能够适应复杂和不断变化的经济环境。另外，随机森林通常在各种预测任务中表现出色，提供高精度的结果。在经济预测中，准确性尤为重要，以确保可靠和实用的洞察。最后，与多元回归等传统统计方法相比，随机森林不需要对数据的分布做出严格假设，使其在处理现实世界复杂数据时更加灵活[24]。

5.2.2. 模型建立

randomForest (GDP_Growth ~ Fixed_Assets + CPI + Industry + Trade_Balance, ntree = 500, importance = TRUE)

6. 结果分析与解读

6.1. 多元回归模型

模型评估：展示模型的准确性和可靠性评估结果。

结果解释：分析预测结果及其对未来经济政策和市场的潜在影响。

结果可视化：使用 R 中的图形工具展示预测结果。

6.1.1. 结果展示

Call:

lm(formula = GDP_Growth ~ Fixed_Assets + CPI + Industry + Trade_Balance,

data = filtered1_data_selected)

Residuals:

Min	1Q	Median	3Q	Max
−3.8059	−0.6862	0.1329	0.8396	2.2541

Coefficients:

Table 1. Result of the multiple regression analysis
表 1. 多元回归分析结果

	Estimate	Std. Error	t-value	Pr (>|t|)
(Intercept)	−1.432e+01	1.330e+01	−1.077	0.294170
Fixed_Assets	−6.222e−05	1.576e−05	−3.947	0.000796***
CPI	1.953e−01	1.327e−01	1.473	0.156443
Industry	5.321e−03	1.539e−03	3.459	0.002481**
Trade_Balance	3.811e−05	5.895e−05	0.646	0.525342

注：显著性代码：0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1；残差标准误差：1.415，自由度为 20；多重 R 平方：0.6993，调整后的 R 平方：0.6391；F 统计量：11.63，自由度为 4 (分子)和 20 (分母)，p 值：4.834e−05。

6.1.2. 模型评估

根据表 1 输出结果进行分析，

1) 残差分析：

最小值、第一四分位数、中位数、第三四分位数、最大值：残差范围从−3.8059 到 2.2541。残差的中位数接近 0 (0.1329)，表明模型预测相对准确。

2) 系数：

Intercept (截距)：表明当所有自变量为 0 时，GDP 增长的预测值为−14.32。但实际上这可能没有经济意义，因为自变量不可能都为 0。

Fixed_Assets (固定资产)：系数为−0.00006222，表明固定资产每增加一个单位，GDP 增长率预期将减少约 0.00006222 个单位，这个效应是统计上显著的(p-value < 0.001)。

CPI：系数为 0.1953，意味着 CPI 每增加一个单位，GDP 增长率预期将增加约 0.1953 个单位，但这个效应不是统计上显著的(p-value > 0.05)。

Industry (工业)：系数为 0.005321，表明工业增加对 GDP 增长有正面影响，且这个效应是统计上显著的(p-value < 0.01)。

Trade_Balance (贸易平衡)：系数为 0.00003811，意味着贸易平衡每增加一个单位，GDP 增长率预期将增加约 0.00003811 个单位，但这个效应不是统计上显著的。

3) 模型拟合优度：

R 方：0.6993，表明约 69.93%的 GDP 增长率变异可以由模型解释。

调整后的 R 方：0.6391，考虑到自变量的数量后，约 63.91%的变异被模型解释。这个值比多重 R 方略低，表明模型可能包括了一些不显著的解释变量。

残差标准误差：1.415，给出了模型预测值与实际值之间差异的标准度量。

F 统计量和 p 值：表明模型整体上是统计显著的(p-value = 4.834e−05)，意味着至少有一个预测变量对 GDP 增长率有显著影响[25]。

结论：

这个模型显示固定资产和工业是影响中国 GDP 增长率的显著因素。然而，CPI 和贸易平衡的影响不显著。模型解释了相当一部分的 GDP 增长率变异，但调整 R 方表明可能存在过度拟合的问题。

6.1.3. 结果解释

1) 预测结果分析

固定资产的影响：模型显示固定资产的增加与 GDP 增长率的下降显著相关。这可能表明过度依赖固定资产投资可能不会持续带来经济增长，或者可能表明在数据的特定时间段内，固定资产投资的效率有所下降。

工业的影响：工业增长对 GDP 增长有正面显著影响。这强调了工业部门在推动经济增长中的重要性，并可能指向提高工业生产效率和创新作为增长策略的潜力。

CPI 和贸易平衡的非显著性：CPI 和贸易平衡对 GDP 增长的影响在统计上不显著，这可能意味着在模型的时间框架内，这些因素与 GDP 增长的关系可能较为复杂或被其他变量所掩盖。

2) 对未来经济及市场的影响

政策制定者可能需要重新评估对固定资产投资的依赖，探索如何提高投资效率或促进其他类型的投资，如技术创新和人力资本发展。投资者可能会根据对固定资产和工业部门的分析调整其投资组合，寻求更有效的资产配置。鉴于工业对 GDP 增长的重要性，政府可能会考虑采取措施支持工业部门的现代化和高技术转型，以促进更高效和可持续的增长。

然而，还需要考虑其他经济指标：CPI 和贸易平衡的非显著性可能促使我们寻找其他更具预测性的经济指标，来探究这些指标与 GDP 增长之间更复杂的关系。

6.1.4. 可视化预测

预测模型解读：蓝线表示实际 GDP 增长率，而红线表示模型预测的 GDP 增长率。时间指数可能代表数据集中的不同时间点。图 2 显示，模型的预测值在大多数点上都紧跟实际值，尽管在某些点上有明显偏差。横轴为模型预测的 GDP 增长率，纵轴为残差(实际值与预测值之差)。理想情况下，这些点应该随机分布在横轴附近，没有明显的模式。图中残差似乎随机分布，没有明显的系统性偏差，这表明模型对于不同水平的预测值拟合得相对均匀。横轴为预测的 GDP 增长率，纵轴为标准化残差。图 3 用于检测是否有离群值，即标准化残差的绝对值特别大的点。从图中看，大部分数据点的残差都位于−2 到 2 之间，这通常被视为正常范围，没有明显的离群值。横轴为理论分位数，纵轴为样本分位数。Q-Q 图用于评估数据的分布是否近似正态分布。图中点大体上跟随了红线，这意味着残差接近正态分布。然而，尾部的点稍微偏离直线，表明可能存在轻微的正态性偏差[26]。

总体来看，预测模型表现良好，残差没有表现出明显的非随机模式，且残差分布大致符合正态分布，这表明模型的假设可能是恰当的。

6.2. 随机森林

6.2.1. 结果展示与解读

从表 2 和图 4 可以看出，%IncMSE 表示当从随机森林中的每棵树去除该变量时，模型误差(均方误差)的平均增加百分比。这个指标衡量了变量被随机置换后对模型预测性能的影响。如果置换一个变量导致误差显著增加，则该变量被认为是重要的。在该模型中，固定资产(Fixed_Assets)和工业(Industry)的%IncMSE 值相对较高，这表明这两个变量在模型中的重要性较高。移除这些变量会导致模型性能的显著下降。

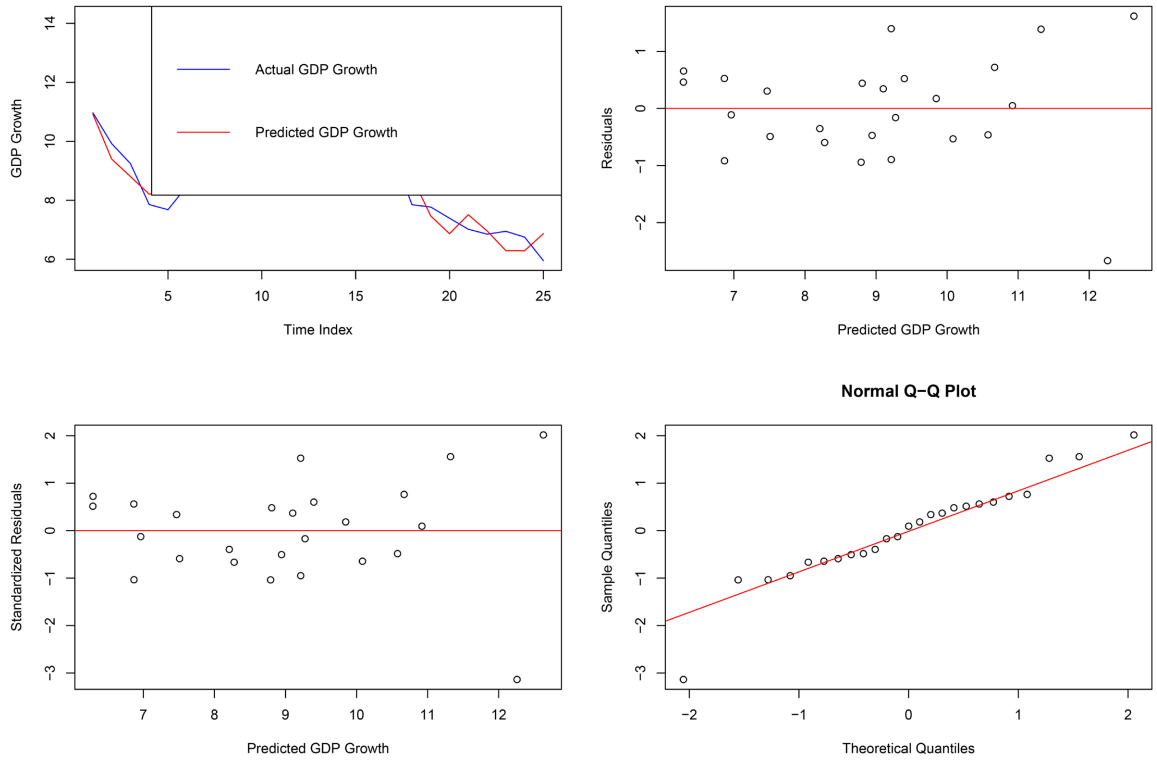


Figure 2. Prediction of the multiple regression model
图 2. 多元回归模型预测

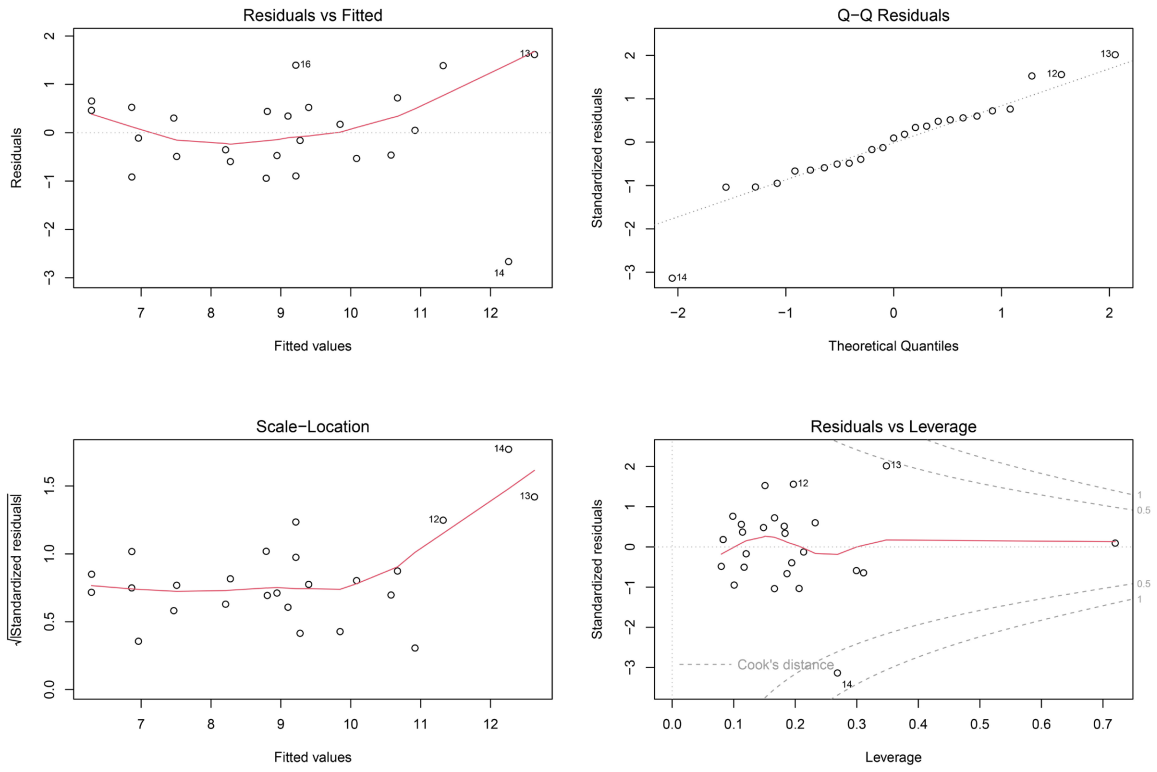


Figure 3. Diagnosis of the multiple regression model prediction
图 3. 多元回归模型预测诊断

Table 2. Results of random forest model analysis
表 2. 随机森林模型分析结果

	%IncMSE	IncNodePurity
Fixed_Assets	13.123273	33.96031
CPI	6.647998	20.82371
Industry	13.903986	35.38221
Trade_Balance	4.244237	28.99617

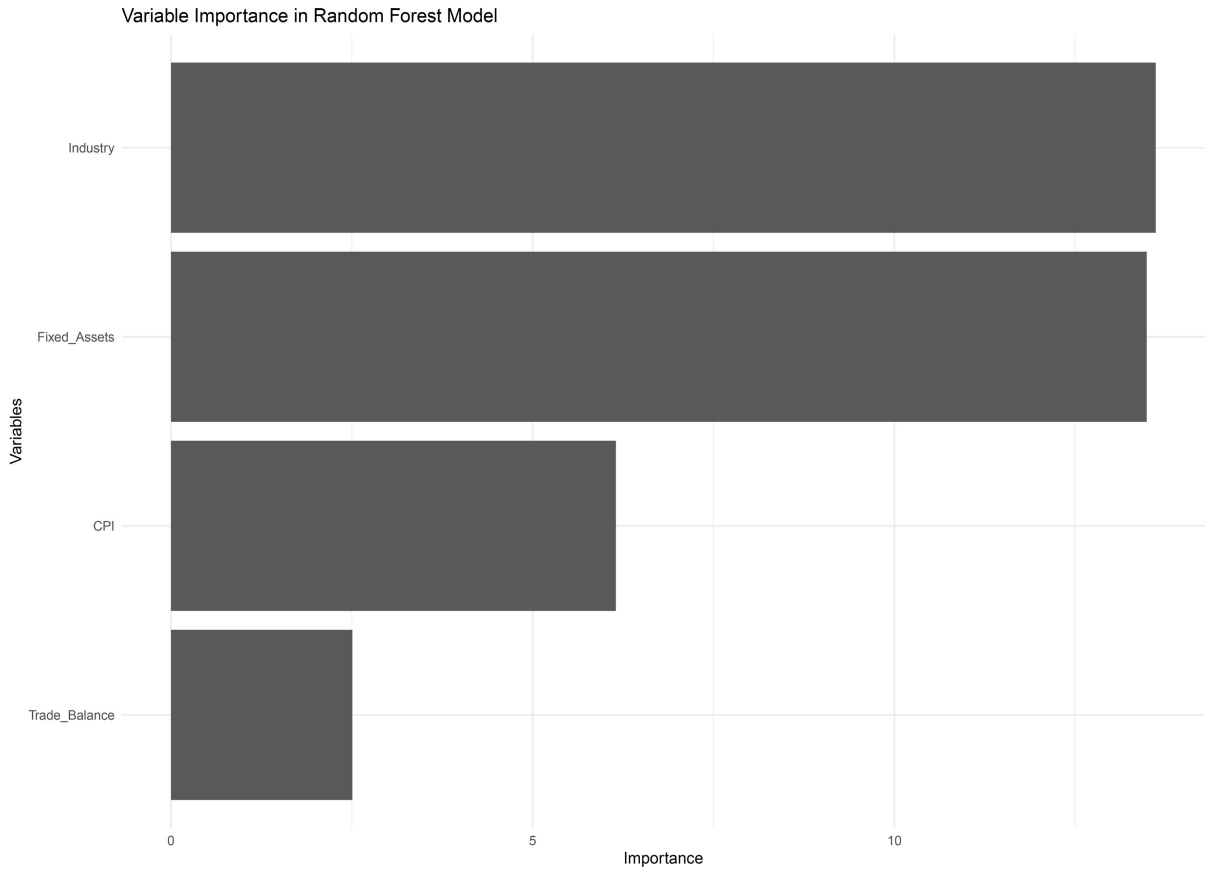


Figure 4. Importance diagram of varieties
图 4. 变量重要性图

IncNodePurity 衡量的是该变量在树节点划分中对提高节点纯度的贡献。这通常是通过计算变量在所有树中所有节点划分中对减少不纯度的总贡献来衡量的。更高的 IncNodePurity 值意味着该变量在分割节点时更能提高预测的准确性。根据该模型，工业(Industry)和固定资产(Fixed_Assets)的 IncNodePurity 值相对较高，这表明它们在树节点划分中起到了重要的作用。

固定资产(Fixed_Assets)和工业(Industry)在两个指标上都显示出较高的重要性，这意味着它们对预测中国 GDP 增长率具有显著影响。特别是在模型误差和节点纯度方面，这些变量的变动会显著影响模型的预测性能。CPI 的重要性在两个指标上都相对较低，表明它对模型预测 GDP 增长的贡献小于其他变量。贸易平衡(Trade_Balance)的%IncMSE 值最低，但其 IncNodePurity 值相对较高，这可能意味着贸易平衡在

某些特定的节点划分中有较大影响，但总体上对模型误差的影响较小。

6.2.2. 模型评估

图 5 显示了一个随机森林模型的误差率随树的数量增加而变化的情况[27]。在这张图中，横轴表示树的数量，纵轴表示误差率。这里的误差率通常是指 Out-of-Bag (OOB)误差。

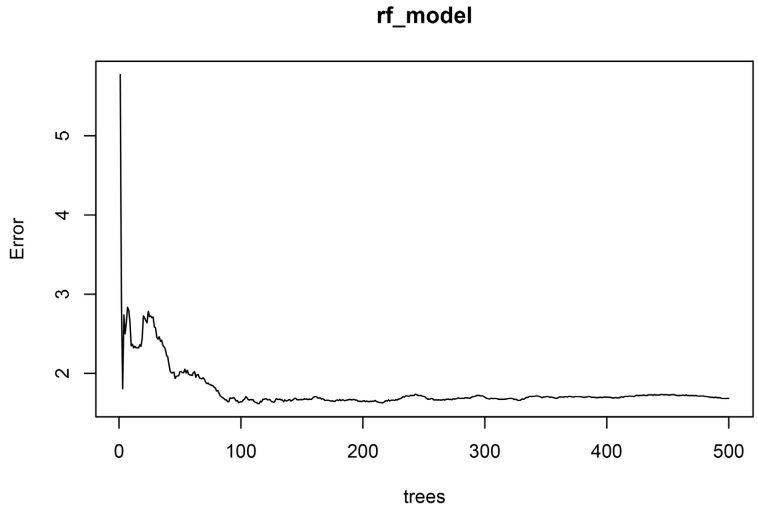


Figure 5. Analysis diagram of error rate on the random forest model
图 5. 随机森林模型误差率分析图

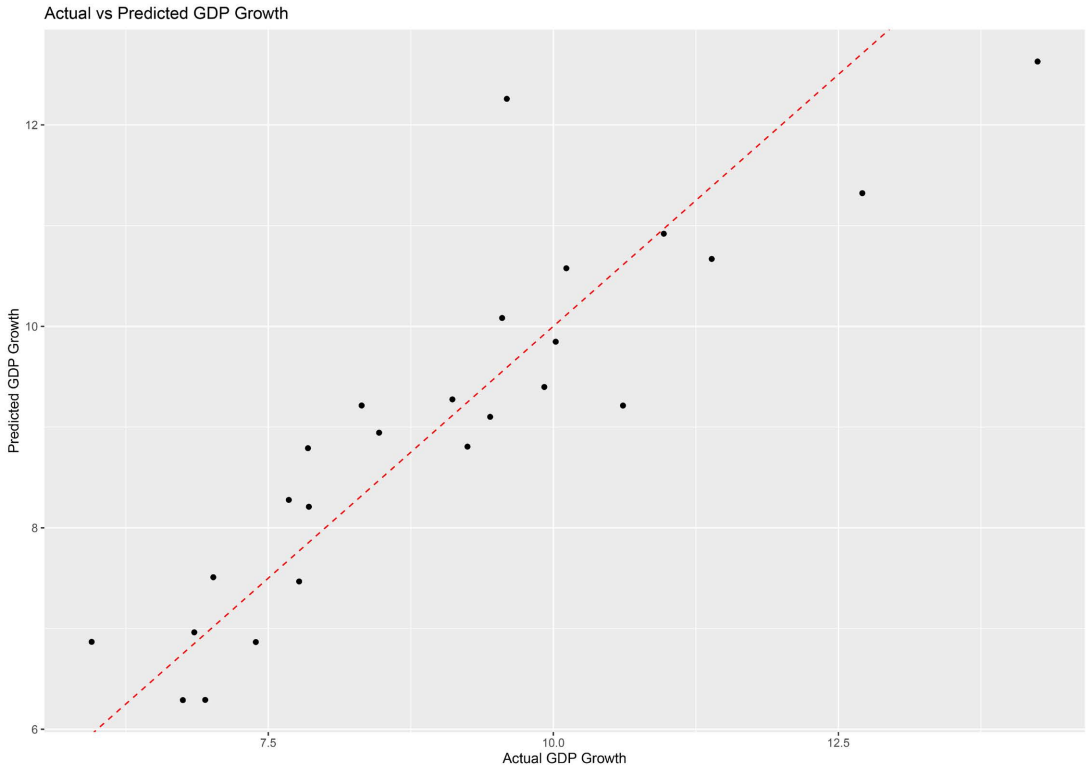


Figure 6. Predicting GDP by the random forest model
图 6. 利用随机森林模型预测 GDP

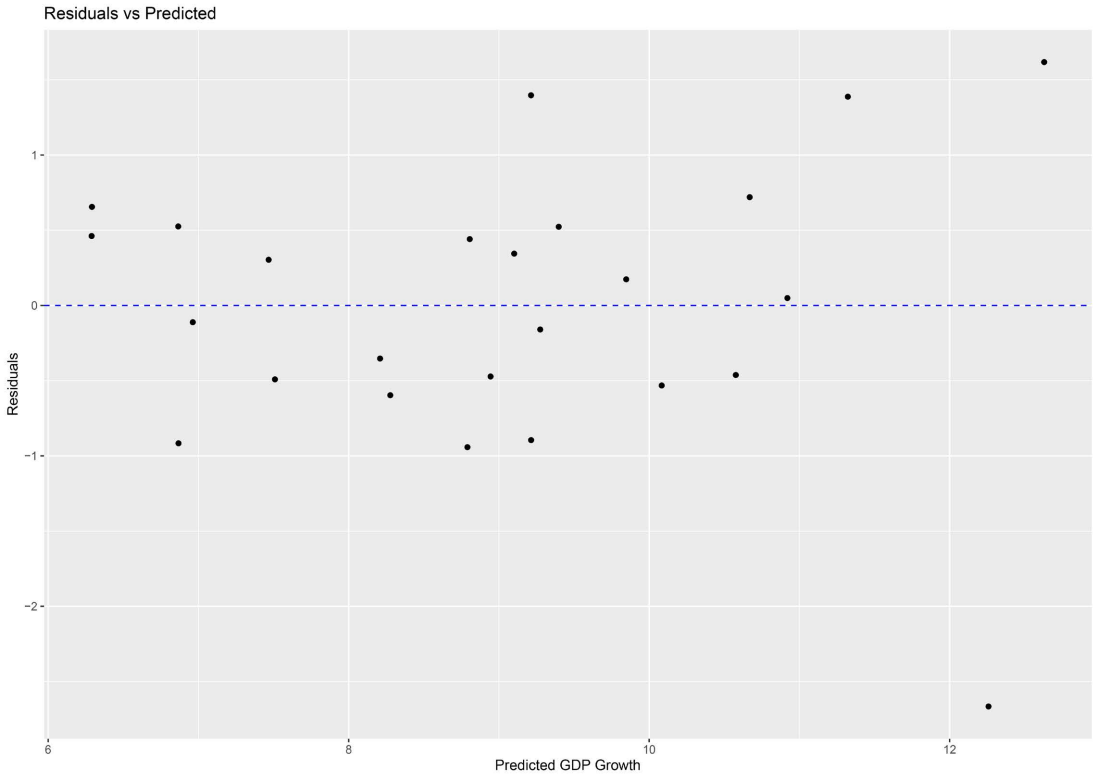


Figure 7. Residual error diagram
图 7. 残差图

从图 5 中可以看出，误差率的初始值相对较高，但随着树的增加，误差率迅速下降并趋于稳定，稳定后的误差率大约在 2.0 左右。在一些情况下，2.0 的误差率可能是可以接受的。

在应用这些见解于政策制定或经济分析时，应考虑固定资产和工业部门的发展对经济增长的潜在影响，同时监控 CPI 和贸易平衡的变化，以进一步了解它们在不同经济环境下的作用。这些结果提供了宝贵的信息，有助于政策制定者和分析师理解经济增长的驱动因素，并制定相应的经济策略。

6.2.3. 可视化预测

图 6 展示了实际 GDP 增长与预测 GDP 增长之间的关系[28]，其中：横轴(X 轴)代表实际 GDP 增长；纵轴(Y 轴)代表预测 GDP 增长；红色虚线表示理想情况下的完美预测，即预测值等于实际值。

从图中我们可以观察到：大多数点沿着红色虚线集中，这表明预测值和实际值通常非常接近，模型的预测性能相对良好。虽然大部分点接近红线，但也有些点偏离这条线，表明在这些点上，模型的预测与实际值有较大差异。根据图 7，没有系统性的偏差，即没有明显的模式表明预测值系统地高估或低估了实际值。这是一个良好的预测模型的迹象。图表中似乎没有离群点远离其他点集中的区域，这表明模型在各个级别的 GDP 增长上都保持了一致的预测准确性[28]。

7. 结论和未来展望

本文主要通过两种方法，多元回归模型和随机森林学习模型对中国的 GDP 进行简单预测。主要探讨了工业产出和制造业指数、消费者价格指数(CPI)、固定资产投资、贸易平衡四个指标对中国 GDP 的影响。研究发现，固定资产对中国 GDP 的影响为负(−0.00006222)，且统计显著，表明随着固定资产投资的增加，GDP 增长率降低。工业对中国的 GDP 的影响为正(0.005321)，且统计显著，说明了工业生产在推

动经济增长中的积极作用。然而，贸易平衡(0.00003811)和 CPI 系数(0.1953)虽然为正，但不具有统计显著性。这表明在模型中，贸易平衡对 GDP 增长的预测贡献不确定，其经济影响可能被其他更强的经济力量所掩盖，或在不同的时间范围内变动较大。CPI 的增加与 GDP 增长的正相关关系不够稳定，无法在 95%的置信水平上确定其对 GDP 增长有显著影响。在政策制定时，可能需要探究其他因素或在更长的时间序列上分析 CPI 的影响。

预测结果在经济分析和决策中的潜在应用可以广泛地涉及以下几个方面：

从政策制定的角度来看，首先，政策制定者可能需要重新评估固定资产投资的效益。负面的系数指出增加投资并不总是能带来 GDP 的增长，这可能暗示投资的效率和质量需要提高，或者在固定资产投资之外寻找其他增长动力。其次，工业生产的积极影响表明政策制定者应继续支持工业部门的现代化和创新，可能包括技术升级、工业自动化和劳动力培训。再者，政府可以根据对 GDP 增长的影响来优先分配财政预算和资源。

从投资的角度来看，首先，投资者和企业可以使用这些预测来规划长期投资战略，尤其是在工业生产和固定资产领域。其次，虽然贸易平衡对 GDP 增长的影响不明确，但这个指标的不确定性可能对企业和投资者的外贸策略产生影响，要求他们对国际贸易环境的变化保持警惕。

从宏观经济研究的角度来看，首先，消费者价格指数(CPI)对 GDP 增长的影响不显著可能影响消费者的信心和消费行为，特别是在通货膨胀预期管理和个人财务规划方面。其次，该模型结果可以作为进一步研究的基础，激发对影响 GDP 增长的其他潜在因素的探索，包括社会政策、环境变化和国际经济形势。

研究中国 GDP 增长率的过程中存在的问题，及其对未来研究方向的影响，是分析经济数据时常见的挑战。这些问题在量化模型和预测分析中尤为重要，它们涉及模型选择、数据质量和外部因素的综合考虑。

1) 变量选择问题

在进行经济预测时，选择正确的解释变量是至关重要的。本文选取的四个变量虽有其合理性，但可能未能充分覆盖影响 GDP 增长的所有关键因素。例如，消费支出、政府支出、科技创新、教育水平、劳动力市场状况、资源利用效率和国际贸易政策等因素，都可能对 GDP 产生重要影响。未来的研究可以通过增加这些宏观经济指标来提高模型的解释力。

2) 数据质量和处理

数据选取的质量和数量会直接影响模型的可信度。缺失值的处理方法多种多样，如插值、利用预测模型填补或者使用更复杂的数据清洗技术。简单地删除缺失值可能会导致样本偏差和信息丢失。在后续的研究中，可以采用多种缺失值处理方法，并评估这些方法对模型预测能力的影响，或者尝试获取更完整的数据集。

3) 未考虑的其他变量

金融市场是影响国家 GDP 增长的重要因素，它通过影响投资、消费和信贷等渠道间接影响经济。此外，全球经济形势、地缘政治事件、国内政策变动、环境变化等因素也可能对 GDP 增长产生重大影响。未来的研究可以考虑引入这些因素，或者运用更加复杂的经济模型来探索这些因素的影响。

接下来，可对本文的方法论进行改进。除了多元回归和随机森林模型外，还有其他高级的统计和机器学习方法，如向量自回归(VAR)、主成分分析(PCA)、深度学习等，它们能够处理更多变量和复杂的非线性关系。这些方法可以提供更深入的理解，并可能提高预测的精确度。

总体而言，未来研究需要在模型和数据选择上进行更细致的工作，以确保预测模型能够准确反映经济现实。这包括使用更广泛的数据集、考虑更多的潜在解释变量、应用更先进的统计方法，并且持续跟踪和纳入新出现的经济趋势和变化。通过这些努力，可以期待未来的预测模型在经济分析和政策决策中发挥更大的作用。

参考文献

[1] 任昱衡. 数据挖掘: 你必须知道的 32 个经典案例[M]. 北京: 电子工业出版社, 2018.

[2] 袁志刚, 余宇新. 中国经济长期增长趋势与短期波动[J]. 学术月刊, 2012, 44(7): 62-72.

[3] 王朝霞. 数据挖掘[M]. 北京: 电子工业出版社, 2018.

[4] 余宇新, 史建明. 洞穿变局: 数字时代与新商业文明[M]. 上海: 上海人民出版社出版, 2022.

[5] 芮雪莹. 基于随机森林的中国上市公司财务舞弊预测模型[D]: [硕士学位论文]. 上海: 上海财经大学, 2020.

[6] 叶莉. 基于 Shapley 回归框架的中国 GDP 预测研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2021.

[7] Williams, J. (2023) I Don't Have a Recession in My Forecast. I Have Pretty Slow Growth. Financial Times, Nikkei Inc., London.

[8] Turner, K., Alabi, O., Katris, A., *et al*. (2022) The Importance of Labour Market Responses, Competitiveness Impacts, and Revenue Recycling in Determining the Political Economy Costs of Broad Carbon Taxation in the UK. *Energy Economics*, **116**, Article ID: 106393. https://doi.org/10.1016/j.eneco.2022.106393

[9] Bernanke, B.S. (2005) The Logic of Monetary Policy. *Vital Speeches of the Day*, **71.6**, 165.

[10] Yellen, J.L. (2009) A View of the Economic Crisis and the Federal Reserve's Response. FRBSF Economic Letter, Federal Reserve Bank of San Francisco, San Francisco.

[11] Keynes, J.M. (1936) The General Theory of Employment, Interest and Money. Palgrave Macmillan, London.

[12] Stiglitz, J.E. (2002) Globalization and Its Discontents. W.W. Norton & Company, New York.

[13] 段树乔, 张正华. 疫后中国经济走向及未来 GDP 预测[J]. 当代经济, 2023, 40(6): 10-17.

[14] 李斌. 新冠疫情背景下中美宏观经济的影响与联动分析[D]: [硕士学位论文]. 北京: 对外经济贸易大学, 2022.

[15] Nordhaus, W.D. and Samuelson, P.A. (2010) Economics. 19th Edition, McGraw Hill, New York.

[16] 陈国青, 曾大军, 卫强, 等. 大数据环境下的决策范式转变与使能创新[J]. 管理世界, 2020, 36(2): 95-105.

[17] Gentleman, R. and Ihaka, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314. https://doi.org/10.1080/10618600.1996.10474713

[18] Stock, J.H. and Watson, M.W. (2015) Introduction to Econometrics. Pearson, Upper Saddle River.

[19] Athey, S. (2018) The Impact of Machine Learning on Economics. In: Agrawal, A., Gans, J. and Goldfarb, A., Eds., *The Economics of Artificial Intelligence*: *An Agenda*, University of Chicago Press, Chicago, 507-552. https://doi.org/10.7208/chicago/9780226613475.003.0021

[20] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[21] 何强. 利用混频大数据预测中国季度 GDP 增速研究[J]. 调研世界, 2018(7): 7-12.

[22] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning. Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7

[23] Varian, H.R. (2014) Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, **28**, 3-28. https://doi.org/10.1257/jep.28.2.3

[24] 戴雅榕. 随机森林模型能够预测中国债券违约吗? [D]: [硕士学位论文]. 厦门: 厦门大学, 2020.

[25] Friedman, J., Hastie, T. and Tibshirani, R. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

[26] Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting: Principles and Practice. OTexts, Melbourne.

[27] Wickham, H. (2009) Ggplot2: Elegant Graphics for Data Analysis. Springer, New York. https://doi.org/10.1007/978-0-387-98141-3

[28] CSDN. https://blog.csdn.net