The analysis of China’s Wholesale and Retail Trade

Liu JinKua
Jinkual@kean.edu

Abstract

This research analyzes China's Wholesale and Retail trade data and give a general analysis as a background study which found that the retail trade may have the ability to reallocate the resource and encourage economic growth among the low economy provinces. Based on the analysis, the data from China Statistic Yearbook 2020 and the National Bureau of statistic from 2019 to 2022, researcher would implement and compare several mathematical models to predict the retail sales of post COVID-19 China's economy to assist the future strategy making to fill up the loss from the pandemic. The models include classification and regression tree (CART), multi-layer perceptron neural network (MLP),Long short-term memory (LSTM) and Support Vector Machine (SVM). Result indicates that CART has the most accurate prediction and the decision tree diagram could be used for strategy making.

Introduction

Recent years, China has gained numerous successes in economy. However, due to the COVID-19 pandemic, the situation has been changed, the global economy has been severely impacted as well as China. To understand current situation and recover the China economy, this paper conducts an analysis to pre and post China economy over wholesale and retail trade data. Massive research has investigated the pre and post pandemic China economy in various aspects from the entrepreneurship and private economy, film and drama industry and digital economy with comprehensive exploration. Despite these research, China wholesale and retail trade has yet to be exploration. Being second largest domestic product, the Wholesale and Retail Trade (WRT) occupied nearly 10%, shown in Fig 1. It may lead to some significant findings to analysis and resume China economy. Besides investigating the pre pandemic situation of China's Wholesale and retail trade market, an accurate prediction model would also assist governor to decision making. In recent years, machine learning has achieved state-of-art performance in mass area. Thus, researcher has selected and compared four different machine learning prediction models, DT-CART, SVM, NN-MLP and LSTM, to select a suitable model for economy prediction and assist to decision making. In brief, this research first conducted a background study centered on pre-COVID-19 China WRT data to explore the strategy to resume economy and then a machine learning models comparison on post-COVID-19 China WRT data to selecting the fittest model for post pandemic economy prediction assisting future decision making.

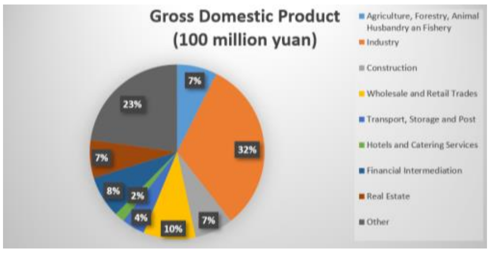


Figure 1: China's Tertiary Industry Seasonal value of add value

Literature Review

To have a fundamental understanding of China economy and popular machine learning models for prediction. Researcher has investigated several empirical studies over pre and post pandemic China economy and machine learning models for economy forecasting.

Before the COVID-19 outbreak, China's economy is actively maturing, and researchers has explored various aspects of its success comprehensively. Liu (2020) has studied the impact of reform and open-up structural transformation of China and conclude its terrific influence which led China economy to a new level. Li (2012) has investigated the flourish development of entrepreneurship and private economy in China. Based on stable support from the government, enhancing the quality of human capital, and an open-up strategy, Jin (2016) has an optimistic view of China's economic and predict that the role of China in the global economy would become essential. However, due to the outbreak of pandemic, the global economy has faced another disaster. The suspension of the supply chain, reduction of productivity (Ba,2020), and the delayed growth of the economy and globalization (Abodunrin, 2020), China has also been one of the victims of the pandemic. According to Hu (2021) pandemic has damaged the whole film and drama industry in China. However, benefiting from the effective prevention and control policy, China has recovered from the disaster firstly and promoted the global economy especially to the upper-middle level income country (Wang, 2020) Interestingly, the COVID-19 pandemic accelerates the digital economy in China that the pandemic has forced some traditional commercial, productivity and entertainment transform into a digital way and became the key to the recovery (Ba,2020). Researchers have conducted a detail and state-of-art analysis of pre and post China economy. However, rarely researcher has focus on China WRT market which the second largest domestic product and may lead to some significate clue for economy development. Thus, first part of this research would aim at analysis and exploring the pre-COVID China WRT market.

Due to the ability to process large, complex data and provide accurate prediction, various machine learning (ML) model has been implemented by more and more researchers in economic analysis. According to Ghoddusi in 2019, support vector machines (SVM), artificial neural networks (ANN), NN-MLP and genetic algorithms have been widely applied from 2005 to 2018 and mostly implemented to predict price and predict or model consumption in energy analysis. Nosratabadi et al. (2020) also has conducted a survey on ML and deep learning (DL), based on their finding, they forecast that ML and DL would be used to analyze more and more complex hybrid models. Garg (2021) has applied decision tree and random forest regressions to predict the impact of the COVID-19 pandemic on the Indian stock market which concluded that with the increasing number of people infected by COVID-19, the stock market would experience a downfall and influence people's income. Yoon (2021) has researched the use of ML by gradient boosting and random forest method to predict Japan's GDP and mentioned that gradient boosting has better performance to predict Japan's GDP. Magazzino (2020) has implemented the ML method to analyze the relationship between solar and wind energy and coal consumption, GDP, and carbon dioxide production. He (2022) predicted the oil price based on classification prediction with multi-model data feature which result in increasing accuracy around 16.8% and 17.6% and better than regression methods. Shahbazi (2022) forecasted the digital coin exchange rate and conclude that XGBoost has better performance comparing to CNN, Arima, LSTM and MLP. Aiming at selecting the fittest model for post pandemic economy prediction, this research compares the several models including SVM, ANN, NN-MLP and DT-CART.

Dataset

Data is collected from the China statistic yearbook 2020 and the National Bureau of statistics. China statistic yearbook 2020 collected China’s economic data from 2015 to 2019 and this research acquired the data 15th, Wholesale and Retail Trade. The data indicates the progress and trade of enterprises over the designated size of wholesale and retail trades, commodity circulation, consumption, market operation, and modernization of logistics. The major data contains the designated size, circulation of commodities, financial status, total retail sale of the product, turnover of large commodity transaction market, chain store of retail trade of China’s Wholesale and retail trade data. The data from the National Bureau of statistics contains the monthly data of the domestic trade and foreign economy from the April of 2019 to March of 2022. The used document is displayed in Tables 1 and 2.

Table 1 The main document implement in pre-COVID-19 analysis

Document
Main indicator of enterprises above Designed Size of Wholesale trade by status of Registration and Sector
Main indicators of Enterprises above Design Size of Wholesale Trade by Region
Main indicators of Enterprises above Design Size of Retail Trade by Region
Main indicator of Chain Retail Enterprises by Status of Registration
Main indicators of Chain Retail Enterprises by Sector and Business Categories

Table 2 The main document implement in Machine Learning model

Document
The total retail sales of social consumer goods
The retail value of textiles, clothing, shoes and hats
The retail value of gold, jewelry, jade
The retail value of furniture
The retail value of cars, motorcycles and spare parts

Methodology

To produce a landscape of the pre-pandemic China economy, several basic data analysis approaches are implemented. After the analysis, the highest time correlated data would be selected as the input of four models with 3 popular forecast models (NN-MLP, SVM, ANN) selected from Ghoddusi reviews and 1 unpopular model (DT-CART).

Commented [京都1]: Develop a little bit more for algorithm, and discuss the reason to choose the data.

Pre-COVID-19 data analysis:

The data visualization and analysis of 5 data sheets of China's Wholesale and Retail trade was conducted to determine a general view of China's economic condition and each province's economic condition. Python was used to handle the data. Standard Normalization, KDE plot, frequency bar, heatmap, linear regression, and data cluster were utilized to investigate the data and determine predictions and advice.

Machine learning models:

This research conducts a comparison of the prediction for each model based on the National Bureau of statistics with 70% of training data and 30% of testing data. Each model

Commented [京都2]: The reason why use ML and discuss a little bit more about the model and reason why.

would be evaluated with the Mean Absolute Error, Mean Squared Error, Root Mean Squared Error by comparing with the real data.

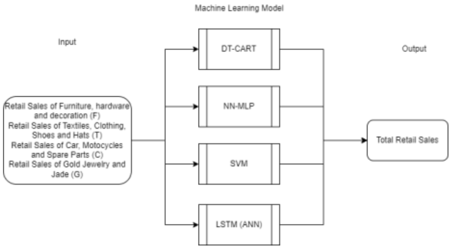

Machine Learning Model
Input
Retail Sales of Furniture, hardware and decoration (F)
Retail Sales of Textiles, Clothing, Shoes and Hats (T)
Retail Sales of Car, Motocycles and Spare Parts (C)
Retail Sales of Gold Jewelry and Jade (G)
DT-CART
NN-MLP
SVM
LSTM (ANN)
Output
Total Retail Sales

Figure 2: The diagram of the Machine Learning Model

DT-CART

Pre-COVID-19 data analysis

Pre-COVID-19 data analysis is conducted to have a primary background study of current WRT situation to find strategy against economy loss caused by pandemic and find the proper data for the machine learning prediction. The reason to analysis the pre-COVID data is due to its completeness comprehensiveness comparing post-COVID data.

Researcher first analyzed the main indicator of Wholesale and Retail trade to find the general tendency of China's economy. The Number of Corporate Enterprises (unit), Total Purchases Value (100 million yuan), Total Sales Value (100 million yuan), imports (100 million yuan), and Employed Persons at the Year-end (10000 people) of China’s Wholesale and Retail market is included in a series of bar charts to show the general tendency and a primary prediction is implemented by linear regression over those values to find the level of growth.

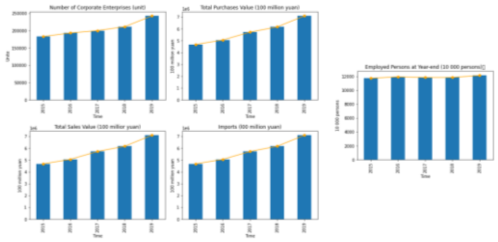
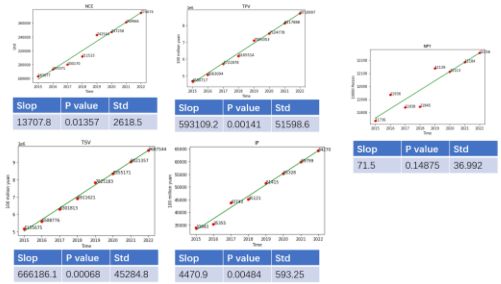


Figure 3: The general tendency of China's WRT market



Slop	P value	Std
13707.8	0.01357	2618.5

Slop	P value	Std
593109.2	0.00141	51598.6

Slop	P value	Std
71.5	0.14875	36.992

Slop	P value	Std
666186.1	0.00068	45284.8

Slop	P value	Std
4470.9	0.00484	593.25

Figure 4: The linear regression prediction of the pre-COVID-19 data

Those figures have inferred that China's pre-pandemic Wholesale and Retail market is developing and increasing rapidly. Except for the employed person at year-end indicates a mild increasement. From the prediction, Figure 6, shows that even though the employed person at the end of the year seems steady, it is still growing. And those predictions show that China's Wholesale and Retail Trade would keep on maturing generally.

Having the tendency of China WRT market, researcher examinates the WRT data of each province and processes the K means clustering in K=3 and the histogram to analyze the general distribution of the total profit in China.

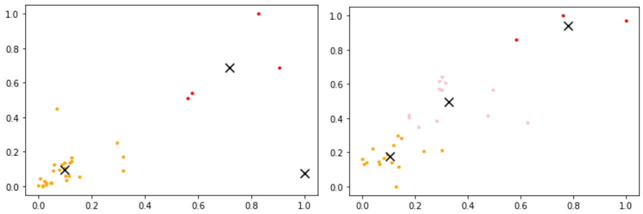


Figure 5: The cluster of each province by total profit and total asset in Wholesale data (right) and Retail data (left)

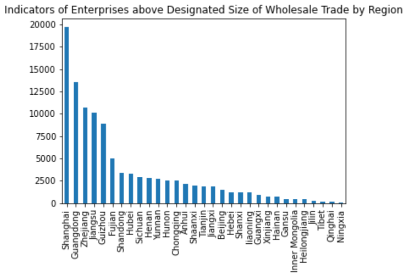


Figure 7: The distribution of total profit of each province in Wholesale data

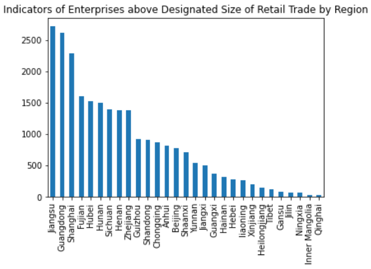


Figure 8: The distribution of total profit of each province in retail data

Provinces are divided into three stages based on their profit and total assets. Fig 5 reveals that there are still some provinces in low profit and low asset situation, Moreover, the result shows that majority of China provinces are in the second and third stage of the total profit and assets. Fig 7&8 display the distribution of the total profit of the provinces in Wholesale and Retail trade data. And fig 7&8 indicate that the problem of the imbalance in economy still exists among provinces in China. Based on fig 7&8, researcher infer that the retail trade with variance 584646.59, standardized, may have the ability to reallocate the market resource equally compared to the wholesale trade with 19913273.04, standardized. Moreover, potential risk may exist in the structure of the wholesale trade since the resources is centralized. With the impact of the top profit province, the whole China economy may be influenced. So, encouraging the development of the retail trade may be a way to promote the economic growth of the second and third stage provinces and ensure the stability of the supply chain and productivity of China. Besides the analysis of the exist data, an accurate and efficient prediction model is in demand to assist the future strategy making. In the later sections, researcher has produced an analysis to select the suitable dataset for prediction and a performance comparison of different prediction models.

An analysis is conducted to select a suitable database for prediction. The WRT market could be divided into two submarket, general market and special market. According to Fig 11, special market in MRT has more turnover than the general market. Thus, researcher decide to set the special market as the analysis target and defined two variables, Per Booth's average Profit (PBP) and Per market average Profit (PMP) shown in formulas 2 and 3. The PBP and PMP of the general market and special market are shown in fig 9,10,11. The vertical line is the mean of each dataset.

$$PMP = \frac{Turnover(100M\ yuan)}{NumBooth\ (Unit)} \qquad (2)$$

$$PBP = \frac{Turnover(100M\ yuan)}{NumMarket\ (Unit)} \qquad (3)$$

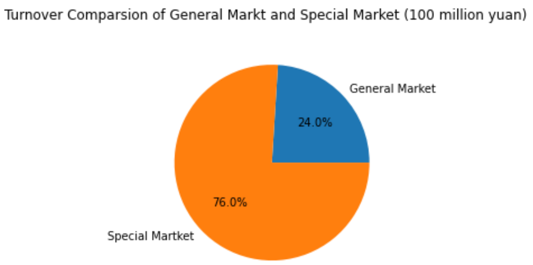


Figure 9: Turnover comparison of General Market and Special Market

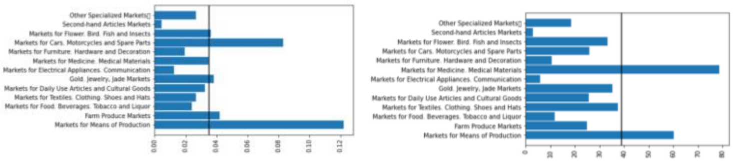


Figure 10: PBP & PMP of special market

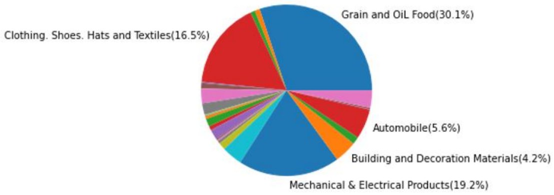
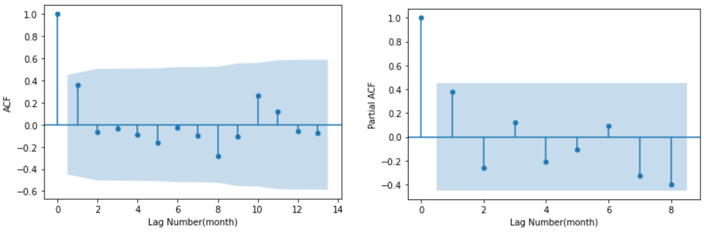


Figure 11: Commodity Exchange Market with 100M trade Turnover

Based on the highest dataset above and existed corresponding special market data provided by the National Bureau of statistics, research has selected four datasets from the special market, Market for Car Motorcycles and Spare Parts (C) and Gold Jewelry and Jade Markets (G) by PBP and Markets for Furniture (F) and Markets for Textiles Clothing Shoes and Hats (T) from the Commodity Exchange Market with 100M trade Turnover. After that researcher has studied the time correlation between the markets and the retail sales data. ACF, PACF, and CCF diagrams are applied to find the time correlation. Fig 12 display the ACF and PACF diagram of the retail sale data. Fig 13 displays the CCF diagram between retail sales and each market. Table 3 displays the statistical properties of the data.

Table 3 The statistical properties of the data in Machine Learning Model

Statistical Index	T(100M yuan)	G(100M yuan)	C(100M yuan)	F(100M yuan)	R(100M yuan)
Minimum	689.0	147.2	2609.0	105.1	9984.3
Maximum	1526	289.1	4871.1	212.3	26300.8
Mean	1110.8	214.4	3567.0	157.4	12836.5
Standard deviation	238.0	37.1	520.20	28.4	1617.6
Skewness coefficient	0.37961	0.12913	0.68554	0.01116	0.38474
Variation coefficient	0.20857	0.16839	0.14120	0.17569	0.12266



Figures 12 : ACF and PACF of Retail sale data

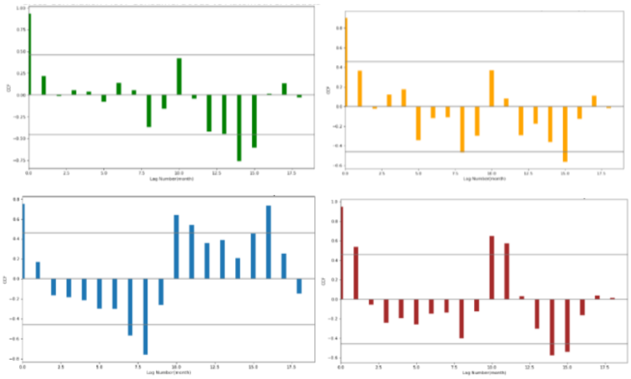


Figure 13: CCP of each selected dataset

From the diagram above, researcher finds the time correlation of each market. Table 4 display the time correlation. And due to the less correlation of furniture data with the retail sale, furniture data has been excluded.

Table 4 The time correlation of each market data

Market	Interval
Retail Sale	{11},{7}
Cars Motorcycles and Spare Parts	{11}
Gold Jewelry, Jade Markets	{11}
Furniture	{17}
Textiles, Clothing, Shoes and Hats	{11}

Machine Learning Model

In this section, researcher has implemented four prediction methods and compared the result of each of them to find the most suitable method for post-COVID prediction. Fig 14 and 15 have shown the result of the DT-CART and NN-MLP. Fig 16 and 17 has shown the result of the SVM and LSTM, one of the ANN. models

Commented [京都3]: Also need to developed a little bit from the result point of view discuss why some model is perform better than other model from technical view.

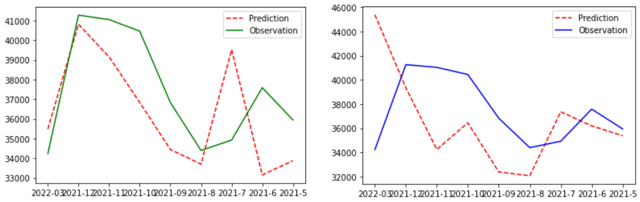


Figure 14&15: Diagram of WRT retail sale prediction for DT-CART and NN MLP

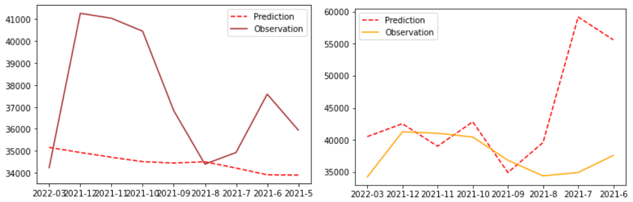


Figure 16&17: Diagram of WRT retail sale prediction for SVM and LSTM (ANN)

To analyze the result, the researcher has applied the R^2, MAE, MSE, and RMSE to the CART-DT and MLP-NN shown in Tables 5.

Commented [京都4]: Based on the observation also mention the trend of the current China situation. Also get the graph of the COVID cases in China.

Commented [京都5R4]: Number of COIVD & China economy.

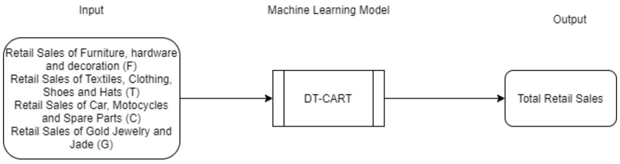
Result

Table 5 The evaluation of the Machine Learning Model (CART & MLP)

	DT-CART	NN-MLP	SVM	LSTM
MAE	2378.25	3896.09	3164.01	8430.52
MSE	7740185.49	24906201.55	15610068.61	116807715.53
RMSE	2782.12	4990.61	3950.96	10807.76

As the graph and table shown above, DT-CART would be the best-fit machine learning model in this research which has lower MSE and RMSE compared to rest of the models and a better tendency in the prediction data compared with NN-MLP, SVM and the LSTM. LSTM requires larger data to provide an accurate prediction which is not suitable for post-COVID-19 data and SVM is not suitable to predict time series data. Ultimately, the prediction model is shown below.

Commented [京都6]: Discuss from the recommendation



Input
Machine Learning Model
Output
Retail Sales of Furniture, hardware and decoration (F)
Retail Sales of Textiles, Clothing, Shoes and Hats (T)
Retail Sales of Car, Motocycles and Spare Parts (C)
Retail Sales of Gold Jewelry and Jade (G)
DT-CART
Total Retail Sales

Figure 18: The prediction model for prediction of China's retail sale

Besides DT-CART also provide governor a strategy to maximize the total profit. As graph shown in Fig 19, governor could encourage the development of Gold Jewelry, Jade Markets (G) and Textiles, Clothing, Shoes and Hats (T) and constrain the development of Cars Motorcycles and Spare Parts (C) to get the maximum profit by comparing the data of current month (T: 1487.2, G: 289.4, C: 4651.3) and the path to the maximum profit node of the DT-CART shown in Table 6.

Commented [京都7]: Put today's value to the training set and see whether it follow the prediction of the maximum

Commented [京都8R7]: Validate it.

Commented [京都9R7]: recommendation & validation

Commented [京都10R7]: Finished but result seem not good.

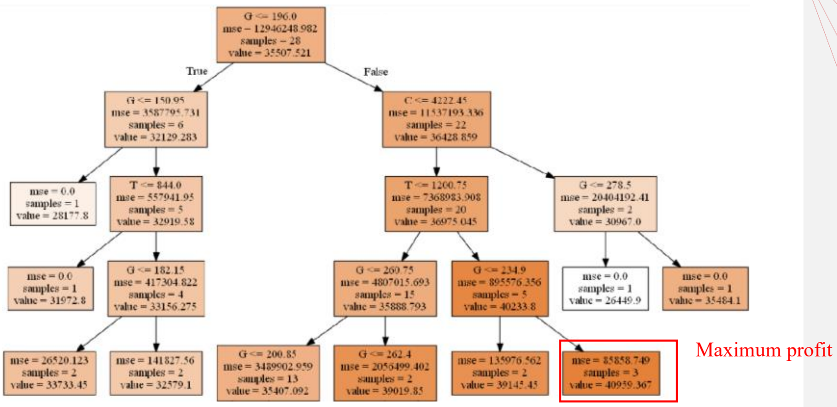
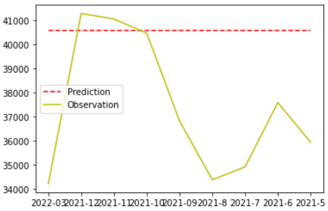


Fig 19: Part of DT-CART diagram with maximum profit

Table 6 The strategy based on DT-CART

T	G	C	Total Profit
>1200.75	> 234.9	<=4222.45	40959.367

However, DT-CART predicts based on the previous data which indicates that the maximum predicted values would not exceed the maximum training value. As a result, DT-CART may only suitable for short-term prediction and decision making.



Conclusion

This study has shown the importance of the Wholesale and Retail trade in China's economy which indicates the retail trade in China may reallocate the resource and encourage the growth of the economy in second and third stage provinces and analyze prediction models resulted fittest model for post-COVID prediction, DT-CART. However, DT-CART is only suitable for short-term prediction and decision making. This research is still a primary study of China's Wholesale and Retail Trade data with limited time and resources. In the future with enough time and dataset, the researcher would conduct more comprehensive analysis of China post COVID-19 Wholesale and Retail Trade and better models for prediction.

Reference

Abodunrin, O., Oloye, G., & Adesola, B. (2020). Coronavirus pandemic and its implication on global economy. International journal of arts, languages and business studies, 4.

Ba, S., & Bai, H. (2020). Covid-19 pandemic as an accelerator of economic transition and financial innovation in China. Journal of Chinese Economic and Business Studies, 18(4), 341-348.

Garg, K. D., Gupta, M., & Kumar, M. (2021). The impact of Covid-19 epidemic on indian economy unleashed by machine learning. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012085). IOP Publishing.

Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. Energy Economics, 81, 709-727.

He, H., Sun, M., Li, X., & Mensah, I. A. (2022). A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features. Energy, 244, 122706.

Hu, J., Yue, X. G., Teresiene, D., & Ullah, I. (2021). How COVID19 pandemic affect film and drama industry in China: an evidence of nonlinear empirical analysis. Economic Research-Ekonomska Istraživanja, 1-19.

Jiang, X. (2020). Digital economy in the post-pandemic era. Journal of Chinese Economic and Business Studies, 18(4), 333-339.

Jin, X., Li, D. D., & Wu, S. (2016). How will China shape the world economy?. China Economic Review, 40, 272-280.

Li, H., Yang, Z., Yao, X., Zhang, H., & Zhang, J. (2012). Entrepreneurship, private economy and growth: Evidence from China. China Economic Review, 23(4), 948-961.

Magazzino, C., Mele, M., & Schneider, N. (2021). A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. Renewable Energy, 167, 99-115.

Nation Bureau of statistics. Monthly Data. Retrieve from: Naional Data: https://data.stats.gov.cn/easyquery.htm?cn=A01

National Bureau of statistics of China. (2020). China statistical yearbook 2020. Retrieve from: http://www.stats.gov.cn/tjsj/ndsj/2020/indexeh.htm

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., ... & Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. Mathematics, 8(10), 1799.

Roman Timofeev, Dr. Wolfgang H¨ardle. (2004). Classification and Regression Trees (CART) Theory and Applications. Retrieve from: ACADEMIA: https://www.academia.edu/13700196/Classification_and_Regression_Trees_CART_Theory_and_Applications

Shahbazi, Z., & Byun, Y. C. (2022). Knowledge Discovery on Cryptocurrency Exchange Rate Prediction Using Machine Learning Pipelines. Sensors, 22(5), 1740.

Spotle.ai. (9/12/2019). Classification and Regression Tree (CART) with Python. Retrieve from: Spotle.ai:https://spotle.ai/feeddetails/Classification-and-Regression-Tree-CART-with-Python/3644

Wang, Q., & Zhang, F. (2021). What does the China’s economic recovery after COVID-19 pandemic mean for the economic growth and energy consumption of other countries?. Journal of Cleaner Production, 295, 126265.

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. Computational Economics, 57(1), 247-265.

Zounemat-Kermani Mohammad, Ramezani-Charmahineh Abdollah, Reza, R., Meysam, A., & Ouarda Taha, B. M. J. (2020). Machine learning and water economy: A new approach to predicting dams water sales revenue. Water Resources Management, 34(6), 1893-1911. doi:http://dx.doi.org.kean.idm.oclc.org/10.1007/s11269-020-02529-0

驻卡拉奇总领事馆经济商务处 . 3/26/2020. COVID-19 将对中国经济产生什么影响. Retrieve from: MINISTRY OF COMMERCE PEOPLE'S REPUBLIC OF CHINA: http://www.mofcom.gov.cn/article/i/jyjl/j/202003/20200302947229.shtml